Etude de la perturbation précoce des marques épigénétiques dans le cerveau fœtal exposé à l'alcool et de l'implication des voies de réponse au stress

Agathe Duchateau

Thèse de doctorat en **biologie moléculaire - épigénétique**
Thèse de l'université Sorbonne Paris Cité
préparée à l'Université de Paris

Ecole Doctorale BioSPC - ED562

UMR7216 - Epigénétique et Destin Cellulaire
Equipe « Interface entre Développement et Environnement »

Etude de la perturbation précoce des marques épigénétiques dans le cerveau fœtal exposé à l'alcool et de l'implication des voies de réponse au stress

Agathe DUCHATEAU

Thèse de doctorat de **biologie moléculaire - épigénétique**
présentée et soutenue publiquement à Paris le 18 décembre 2019

JURY

Président du jury	Pr. Reiner Veitia, Institut Jacques Monod
Rapporteur	Pr. Claire Vourc'h, Université Grenoble Alpes
Rapporteur	Dr. Jean-François Deleuze, Centre national de recherche en génomique humaine - CEA
Examinateur	Pr. Dennis Thiele, Duke University School of medecine
Examinateur	Dr. Maxim Greenberg, Institut Jacques Monod
Directeur de thèse	Dr. Délara Sabéran Djoneidi, UMR7216

Titre : **Etude de la perturbation précoce des marques épigénétiques dans le cerveau fœtal exposé à l'alcool et de l'implication des voies de réponse au stress**

Résumé : Le cerveau fœtal est vulnérable aux stress **environnementaux comme l'exposition prénatale à l'alcool** (EPA), première cause non génétique de retard mental. Ce stress induit un large spectre de défauts neurodéveloppementaux souvent diagnostiqués tardivement. Mieux comprendre les mécanismes moléculaires à l'origine de ces défauts permettrait d'élaborer des outils diagnostics fiables, pour une prise en charge précoce des sujets à risques.

Le facteur de transcription HSF2 est un acteur majeur de la réponse à l'EPA dans le cerveau en développement. Bien que nécessaire au développement physiologique du cortex, il conduit, en contexte de stress, à des anomalies du développement cérébral, en changeant de cibles génomiques. De plus, dans le cortex embryonnaire, dans un contexte physiologique ou suite à une EPA, HSF2 lie DNMT3A, protéine responsable de la méthylation *de novo* de l'ADN.

Comme, il a été montré que des adultes ayant subi une EPA présentent une perturbation de leur méthylome, il était concevable que l'interaction DNMT3A/HSF2 dans un contexte de stress, puisse, en partie, être à l'origine de cette modification du méthylome. Pour tester cette hypothèse, une alcoolisation de type *binge drinking* dans des cortex embryonnaires murins a été réalisée et trois études de **séquençage à haut débit** (**NGS**) ont été menées en parallèle, puis intégrées.

Un *ChIP-seq* ciblant HSF2 dans le cortex cérébral fœtal murin a permis de cartographier les sites de fixation de ce facteur et d'identifier 280 cibles de HSF2 après cette EPA. La plupart de ces cibles sont des gènes impliqués dans le développement cérébral, dans la réponse au stress et/ou associées dans la littérature à des effets d'une EPA.

L'identification de 432 **régions différentiellement méthylées** (**DMRs**), immédiatement après EPA entre des cortex fœtaux témoins (traités au PBS) ou traités à l'alcool a été possible, par une capture du méthylome. Cette analyse a nécessité la mise au point d'un outil et d'une approche bio-informatiques spécifiques. Ces DMRs se situent majoritairement au niveau d'*enhancers* actifs du cortex adulte. Une forte proportion des gènes affectés correspond à des gènes soumis à l'empreinte ou des gènes codant des protocadhérines (impliquées dans le neurodéveloppement ou des fonctions cérébrales), connus comme étant perturbés chez l'adulte à la suite d'une EPA. Ces gènes sont donc porteurs de marques épigénétiques anormales, déposées dès la fin de l'EPA, et constituent de potentiels biomarqueurs d'exposition. Ainsi, des cicatrices épigénétiques sont mises en place rapidement après l'EPA et suggèrent, à la lumière de la littérature, que certaines persistent chez l'adulte.

Pour estimer les conséquences fonctionnelles de l'EPA sur le cerveau en développement, une étude de l'accessibilité de la chromatine et de l'expression des gènes sur la période encadrant le stress a été réalisée en contexte physiologique, en analysant des données publiques (ENCODE) d'*ATAC-seq* et de *RNA-seq* provenant de cortex préfrontaux murins non stressés. Ce *data mining*, a permis de dénombrer et d'identifier les régions chromatiniennes différentiellement ouvertes ou fermées, ainsi que les gènes activés ou réprimés entre les stades embryonnaires E13 et E16 dans le cortex en développement. Les DMR sont associées à des régions chromatiniennes dont l'accessibilité varie au moment du stress, mais aussi au niveau de gènes dont l'expression augmente au cours du développement, suggérant une sensibilité particulière de ces régions dynamiques du génome.
L'analyse intégrée des différents jeux de données NGS, n'a pas permis de mettre en évidence une corrélation entre les sites fixés par HSF2 et les DMR. En revanche, les sites de fixation de HSF2 étant souvent associés à des sites de liaison de *readers* du méthylome ou des *remodellers* de la chromatine, il est possible que HSF2 soit impliqué dans les conséquences fonctionnelles des perturbations du méthylome dues à l'EPA, plutôt que dans la mise en place de ces défauts.

Mots clefs : Exposition prénatale à l'alcool, méthylation de l'ADN, HSF2, cerveau, neuro-développement, épigénétique, réponse au stress, séquençage à haut débit, bio-informatique

Title: **Analysis of early epigenetic mark disruptions caused by prenatal alcohol exposure in the mouse developing brain, and involvement of stress-response pathways**

Abstract: Fetal brain is vulnerable to environmental stress such as prenatal alcohol exposure (PAE), the leading non-genetic cause of mental retardation. This stress induces a wide spectrum of neurodevelopmental defects that are often lately diagnosed. A better understanding of molecular mechanisms underlying these defects would help to develop reliable diagnostic tools for the early care of high-risk subjects.

HSF2 transcription factor is a major actor of PAE response in the developing brain. Necessary for cortical physiological development, it also leads, in the context of chronic PAE stress, to abnormalities in brain development by changing its genomic targets. In addition, V. Mezger's team has demonstrated, in the embryonic cortex, that HSF2 binds DNMT3A - in a physiological context or following a PAE, DNMT3A being responsible for *de novo* DNA methylation.

Since it has been shown that the methylome profile of adults that were exposed to a PAE stress, is often perturbed, it was conceivable that the DNMT3A/HSF2 interaction, in a stressful context, may, in part, be responsible for this methylome profile modifications. To test this hypothesis, three integrative high-throughput sequencing (NGS) studies were conducted, using cerebral cortices of mouse embryos exposed, or not, to PAE corresponding to a binge drinking alcoholization.
ChIP-seq experiments, targeting HSF2 in alcohol-exposed fetal cortices, allowed us to map its binding sites in the genome, and identify 280 HSF2 targets. Most of these target sites are associated with genes involved in brain development or in stress response. Some of these genes are also linked, in the literature, with PAE effects.

Few hours after PAE, 432 differentially methylated regions (DMRs) were identified between control (PBS treated) or alcohol-treated fetal cortices, using a methylome capture protocol. This analysis required the development of specific bioinformatics tools and approaches. These DMRs are mainly localized in active enhancers of the adult cortex. A high proportion of their associated genes correspond to imprinted genes or genes encoding clustered Protocadherins, both involved in neurodevelopment or brain function, and known to be impacted in adults prenatally exposed to an alcoholic stress. Because their deposition is linked to PAE *per se* and show some persistence in the postnatal/adut period, this strongly reinforces their potential as biomarkers of exposition. These results indicate that epigenetic ‘scars‘ are deposited very quickly after PAE and suggest, based on the literature, that some of them persist in adults.

To estimate the functional consequences of PAE on the developing brain, a study of chromatin accessibility and gene expression over the stress period was conducted in a physiological context, analyzing public data (ENCODE) of ATAC-seq and RNA-seq from unstressed murine prefrontal cortices. This data mining study allowed us counting and identifying the chromatin regions that are differentially opened or closed, as well as the genes that are activated or repressed between the embryonic stages E13 and E16 in the developing cortex. Of note, a proportion of DMRs are significantly associated with chromatin regions whose accessibility varies - under physiological conditions - during the stress period, but also with genes whose expression increases during development, suggesting a particular vulnerability at these dynamic regions of the genome to stress.
Our integrative analysis of the different NGS datasets did not reveal any correlation between HSF2 binding sites and the DMRs. However, since HSF2 target sequences contain often binding sites of methylome *readers* or chromatin *remodellers*, HSF2 might be involved in functional consequences of PAE-induced methylome disturbances, rather than in the establishment of these defects.

Keywords: Prenatal alcohol exposure, DNA methylation, HSF2, brain, epigenetics, stress response, neurodevelopment, NGS, bioinformatics

A mon "bon vieux Bon Papa",

parti avant de connaitre

la fin de cette histoire.

Remerciements

Je tiens tout d'abord à remercier les membres de mon jury, d'avoir accepté d'évaluer mon travail de thèse, notamment Claire Vourc'h et Jean-François Deleuze pour avoir pris le temps de lire mon manuscrit dans des délais courts. Je remercie également Dennis Thiele et Maxim Greenberg qui ont accepté d'être examinateurs, et Reiner Veitia pour avoir présidé le jury.

Je souhaite vivement remercier l'ensemble de l'équipe « Interface entre Développement et Environnement », qui m'a chaleureusement accueillie au sein de l'UMR7216 et m'a soutenue tout au long de ma thèse. Je voudrais porter une attention particulière à Délara, pour sa patience et son encadrement sans faille. Tu as été extrêmement disponible pour moi, bien au-delà de tes fonctions de directrice de thèse. Tes encouragements et ton regard critique m'ont permis de réaliser ce travail avec plus d'assurance et de fiabilité. Avec toi, j'ai un peu poursuivi mes « fouilles archéologiques », non pas dans la boue, mais plutôt dans la masse de données sur disques durs.

Un grand merci également à Valérie, qui m'a fait confiance en me proposant ce sujet et en m'accueillant au sein de son équipe. Ton avis constructif a été d'une grande aide pour la réalisation de ce projet.

Je tiens aussi à remercier Véronique, Aurélie, Anne et Myriame (girls power), pour leur soutien et leurs conseils précieux tout au long de la thèse. Ce fut un vrai plaisir d'apprendre à vos côtés, mais aussi de discuter de tout et de rien. Merci aussi pour les intentions sucrées que vous avez apportées régulièrement.

Merci aussi aux anciens membres de l'équipe, Federico Miozzo, Rachid El Fatimy et Anne Laure Schang. Malgré vos nouvelles occupations, vous avez pris le temps de répondre à mes questions, ce qui m'a permis de poursuivre le projet dans la bonne direction.

Une petite pensée également pour les stagiaires présents pour de (trop) courte durée, qui ont apporté une bonne ambiance de travailet pour certains, leur aide dans mon projet : merci Oriane, Yoann, Aurélie, Camille B., Camille A. et Clara.

Un énorme merci à l'ensemble de l'UMR7216, pour l'ambiance de travail extrêmement plaisante de l'unité. Beaucoup de délires au moment des repas, mais aussi lors des activités scientifiques et les extra. Je garderai en mémoire de bons souvenirs de la retraite d'unité, des pique-niques, Christmas Party, flashball, paint ball, soirée Wii,… sans oublier les petites danses de couloir. Votre bonne humeur (quasi) quotidienne et les échanges scientifiques et « technico-techniques » que j'ai pu avoir vous tous ont contribués à l'avancement de mon projet. Je tiens particulièrement à remercier Olivier Kirsh, qui m'a épaulé en bioinformatique dès le début de ma thèse, et ce jusqu'à la fin même quand je râlais à cause de problèmes informatiques. Je ne ramènerais jamais assez de kilos de chocolats pour te remercier. Merci également à Laure, Laurence, Guillaume V. et Giacomo pour leurs nombreux conseils (et matériaux fournis) sur la mise au point du protocole *ChIP*. Merci aussi à Kévin, Jean François et Costas pour leur aide en bioinformatique et à Ingrid, grande « cheffe gestion » mais aussi cheffe cuisinière.

Je souhaite également remercier Maxim Greenberg et Slimane Ait Si Ali, membres de mon comité de suivi de thèse, qui ont été de bons conseils, rassurants et enthousiastes sur mon travail et la poursuite de ma thèse.

Durant ma thèse, j'ai pu tester mes « talents » en enseignement. Je tiens de ce fait, à remercier particulièrement le corps enseignant qui m'a épaulé et m'a permis de remplir cette fonction avec assurance. Merci à Dominique Buffard, Mélanie Franco et Vinciane Régnier. Merci aussi à Délara, Olivier, Olga Rospopoff et Maryline Moulin, j'ai pris beaucoup de plaisir à enseigner à vos côtés. Merci également aux techniciens qui ont pris le soin de préparer les salles de TP et à Elodie Rousseau, d'avoir toujours veillée au bon déroulement des TP (et pour son soutien moral, quand j'ai fait une partie du TP à la place des étudiants qui devaient choisir entre aller en examen ou en TP…).

Parce que la thèse a aussi été l'occasion pour moi de me former en bioinformatique, je voudrais remercier vivement les formateurs du DU omiques que j'ai suivie, Gaëlle Lelandais, Pierre Poulain et Bertrand Cosson, ainsi que toute la promotion « test ». C'était un plaisir de vous voir tous les mois pour deux jours d'apprentissage intensifs et de franches rigolades. Je me dois de finir ce paragraphe par « point point ». Merci en particulier à Gaëlle pour le temps que tu m'as consacré afin de répondre à mes problématiques statistiques en lien avec mon projet..

Un grand merci à ma famille et à mes amis, qui ont toujours été présents, dans les moments joyeux mais également dans les périodes plus difficiles. Vous avez su me réconforter et me *rebooster* par votre présence chaleureuse. En particulier, merci à Nicolas pour son cran et son soutien sans faille malgré mes râleries. Merci à Hélène, Ghina et Marion, qui m'ont fait (re)découvrir des artistes hors norme (GAJ <3). Merci à Krys pour la dégustation de Pulparindo, qui m'a permis de survivre au froid strasbourgeois et à Nada qui, m'a fait découvrir sa culture à travers ses talents de cuisinière ce qui n'a pas contribué à mon régime (brick, couscous, muffin, cheesecake, café turc… <3), et m'a appris l'espagnol avec Julia (Jirafa Jirafa Jirafa). Merci à Leïla pour les soirées confidences et le fou rire mémorable qu'on a eu récemment. Merci à Valentin qui m'a toujours soutenu malgré des moments délicats.

I also want to thank a lot, collaborators of the team who helped me a lot. Thank you Sascha for your bioinformatics and statistical advices. Without your help, I would still be wondering how to analyze the methylome capture dataset.

Léa Sistonen and Jenny Joutsen are acknowledged for sharing the very valuable anti-HSF2 antibodies. Without this material, *ChIP-seq* experiment would not have worked.

Table des matières

Liste des abréviations

5hmC	5-hydroxy-méthyle-cytosine
5mC	5-méthyle-cytosine
ADN	acide désoxyribonucléique
ARNm	ARN messagers
ATAC-seq	*Assay for Transposase-Accessible Chromatin with highthroughput sequencing*
CGi	*CpG island*
ChIP	*Chromatin Immuno-precipitation*
ChIP-seq	*Chromatin Immuno-precipitation with high throughput sequencing*
CP	*Cortical Plate* ; Plaque Corticale
CpG	dinucléotide Cytosine-Guanine
CTCF	*CCCTC-binding factor*
DEG	*Differentially methylated Gene(s)* ; gène(s) différentiellement exprimé(s)
DMC	*Differentially methylated cytosine* ; cytosine différentiellement méthylée
DMR	*Differentially methylated region* ; région différentiellement méthylée
DNMT	*DNA methyl-transferase* ; ADN méthyl-transférase
DOCR	*Differentially Opened or Closed Region(s) ;* régions différentiellement ouverte(s) ou fermée(s)
DOHaD	*Developmental origins of health and disease*
FC	*fold change*
FDR	*false discovery rate*
E(16.5)	(16.5^{ème}) jour embryonnaire
EPA	exposition prénatale à l'alcool
ES	cellules souches embryonnaires
H3K27ac	acétylation de la lysine 27 de l'histone H3
H3K4me1	mono-méthylation de la lysine 4 de l'histone H3
HET	hétérozygote
HSE	*Heat Shock Element*
HSF2	*Heat Shock Factor 2*
Hsf2$^{-/-}$	souris invalidéé pour *Hsf2*
Hsf2$^{-/+}$	souris hétérozygote *Hsf2*
Hsf2$^{+/+}$	souris sauvage (*wild-type*) pour *Hsf2*
IG	*imprinted genes* ; gènes soumis à l'empreinte
IP	Immuno-précipitation
KO	*knock-out*
MAPs	*Microtubules Associated Proteins*
P0	*Postnatal day 0* ; naissance
Pcdh	protocadhérine
pval	*p-value ;* valeur-p
qval	*q-value* ; valeur-q
RNA-seq	*RNA-sequencing*
SAF	syndrome d'alcoolisation fœtale
SAM	S-adénosyl-methionine
SNC	système nerveux central
TCAF / FASD	troubles causés par l'alcoolisation fœtale / *fetal alcohol syndrome disorders*
TET	*ten eleven translocation protein*
TSS	*Transcription Start Site ;* Site d'initiation de la transcription
VZ	*ventricular zone* ; zone ventriculaire
WT	*wild type*

Table des illustrations

INTRODUCTION

Revue

OBJECTIF DE LA THESE

MATERIEL ET METHODES

RESULTATS

Papier en préparation

Table des tableaux

Chapitre 1 : Introduction

Introduction

Le cerveau en développement est vulnérable à divers stress environnementaux. Parmi ces stress, l'exposition prénatale à l'alcool (EPA), représente une des principales causes de retard mental non génétique dans les pays développés. En fonction de nombreux paramètres tels que la durée d'exposition, le trimestre de grossesse concerné ou encore la quantité d'alcool consommée, les désordres neuro-développementaux causées par ce stress, associées ou non à des défauts phénotypiques faciaux, peuvent être plus ou moins marqués chez l'enfant exposé. Ainsi, l'EPA provoque un large spectre de défauts neuro-développementaux, dont certains sont difficilement perceptibles avant la scolarisation de l'enfant, rendant le diagnostic des sujets à risques parfois délicat. Or, une identification le plus tôt possible des personnes à risques de développer des troubles neuro-développementaux est indispensable et critique pour permettre une intervention précoce, qui limiterait les effets néfastes de l'alcool chez ces individus.

Au vu de la prévalence de ce stress alcoolique subi *in utero* et de l'impact sociétal et budgétaire que ce stress représente, il serait intéressant de mieux comprendre les mécanismes moléculaires à l'origine de ces défauts qu'il engendre dans le cerveau en développement. En effet, bien que les dommages provoqués par une EPA soient bien documentés, les mécanismes moléculaires sous-jacents, à l'origine de ces troubles ne sont pas clairement identifiés. Une des pistes prometteuses, étudiée depuis quelques années, s'oriente vers une possible dérégulation par l'EPA, de mécanismes épigénétiques, qui modulent l'expression des gènes sans modifier la séquence nucléotidique de l'ADN. Parmi les marques épigénétiques concernées, la méthylation de l'ADN semble particulièrement affectée par ce stress, puisque son altération a été observée dans le cerveau d'individus adultes ayant subi une EPA, sans pour autant que les mécanismes à l'origine de ces perturbations soient identifiés.

Outre les altérations de marques épigénétiques causées par l'EPA, la voie de réponse aux stress associée aux facteurs de transcription *Heat Shock Factors* (HSFs), semble jouer un rôle important dans la médiation des effets néfastes de l'EPA dans le cerveau en développement. En particulier, le facteur HSF2 semble avoir un rôle multi-facettes, en étant à la fois impliqué dans la réponse au stress, mais aussi dans le développement physiologique du cerveau.

C'est dans ce contexte que s'inscrit ce travail de thèse, qui vise à identifier si le cerveau en développement présente, tout comme le cerveau adulte, des défauts de méthylation de l'ADN suite à l'EPA. Ce projet vise également à préciser le rôle du facteur de réponse au stress HSF2 dans la réponse à l'EPA, et à proposer et tester un mécanisme moléculaire qui pourrait être à l'origine de défauts du méthylome causés par ce stress.

Dans cette introduction, nous aborderons ainsi brièvement les thématiques suivantes : les mécanismes épigénétiques (en se focalisant plus précisément sur la méthylation de l'ADN), le développement physiologique du cerveau, l'exposition prénatale à l'alcool et ses conséquences sur le cerveau, et enfin, le rôle des HSF (principalement de HSF2) dans le développement du cerveau et dans la réponse au stress alcoolique.

Partie 1 : Un mécanisme épigénétique particulier : la méthylation de l'ADN

1. Qu'est-ce que l'épigénétique ?

1.1. Définitions

Dans une cellule, les chromosomes sont constitués de molécules d'ADN de plusieurs cm de long, totalisant pour une cellule de mammifère, une longueur d'environ deux mètres. L'ADN est donc compacté dans un noyau de quelques microns chez les eucaryotes. Cette compaction, qui compromet l'accessibilité à l'ensemble de l'information génétique contenue dans l'ADN, est contrôlée par des mécanismes épigénétiques.

L'enchainement des bases nucléotidiques (A - adénines, C - cytosines, T - thymines et G - guanines) le long de la séquence de l'ADN constitue cette information génétique. En modifiant biochimiquement les acteurs de la compaction de l'ADN, sans remanier la séquence nucléotidique de l'ADN, ces mécanismes épigénétiques procurent un niveau crucial de contrôle de l'expression des gènes. En effet, ils permettent ou non aux machineries cellulaires, d'accéder à l'information génétique, en modulant son accessibilité (Bird, 2007). Les mécanismes épigénétiques supervisent donc la transcription de gènes codant ou non des protéines.

La compaction de l'ADN dans le noyau est assurée par son enroulement autour d'un ensemble protéique, composé d'histones. Ce complexe ADN-protéines, appelé nucléosome, constitue l'unité de base de la chromatine (Figure 1.1). Ces nucléosomes, ainsi que des ordres supérieurs de compaction, limitent l'accessibilité à l'information génétique contenue dans cette portion de l'ADN (Dulac, 2010; Probst et al., 2009; Schang et al., 2017).

1.2. Les mécanismes épigénétiques majeurs

Plusieurs mécanismes épigénétiques agissent à différents niveaux sur cette compaction et donc sur l'accessibilité de la chromatine. Certains d'entre eux engendrent des modifications chimiques de résidus cytosines de l'ADN (Pour plus de détails, cf partie *méthylation de l'ADN*). D'autres sont responsables de modifications post-traductionnelles des histones, protéines qui composent les nucléosomes (Figure 1.1). Ces variations biochimiques sont très diverses : acétylation, méthylation, phosphorylation, ubiquitination, *etc*, pour ne citer que les plus étudiées. Combinées les unes aux autres, ces modifications chimiques (ou marques épigénétiques) constituent un véritable « code histone » (Lacoste and Côté, 2003) interprété par des protéines spécifiques, en analogie au code

génétique porté par la séquence d'ADN. Ce « code histone » induit soit une stabilisation, soit une déstabilisation des nucléosomes, en fonction (**1**) des propriétés biochimiques conférées par ces modifications (l'affinité des histones pour l'ADN, chargé négativement, peut être renforcée ou au contraire affectée, selon les charges électriques des histones, qui sont plus ou moins neutralisées par les modifications chimiques post-traductionnelles présentes au niveau de ces histones ; Hsieh and Gage, 2004), mais aussi (**2**) en recrutant des complexes répresseurs ou activateurs de l'expression génique.

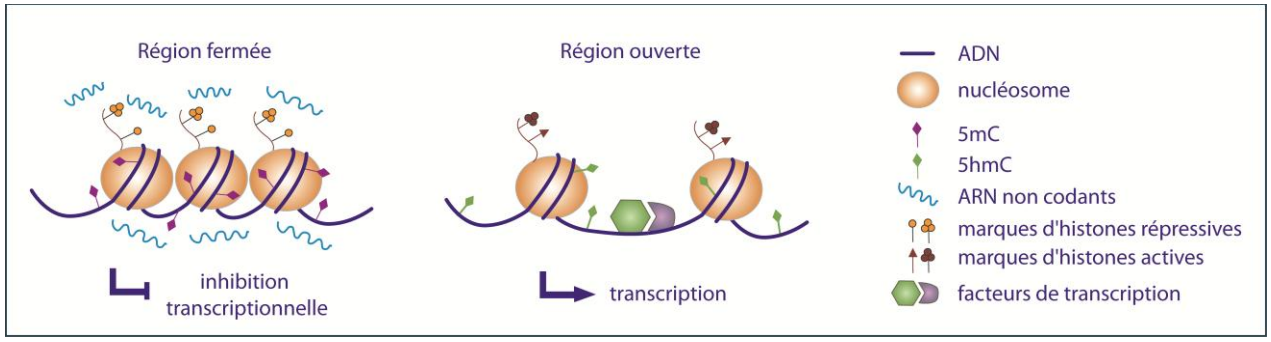


Figure 1.1 : Des mécanismes épigénétiques combinés contrôlent l'expression des gènes en modulant l'accessibilité de la chromatine.
Au sein d'une cellule, certaines régions chromatiniennes sont inaccessibles à la machinerie transcriptionnelle (régions fermées où les nucléosomes sont compactés), alors que d'autres sont ouvertes (les nucléosomes y sont moins denses). Dans ce dernier cas, l'ADN est accessible aux facteurs de transcription, facilitant ainsi l'expression de gènes spécifiques à un type cellulaire donné. La combinaison de marques épigénétiques définit l'état de la chromatine et son maintien. Bien qu'il existe quelques exceptions à ce schéma général, les régions chromatiniennes fermées portent des marques épigénétiques répressives : méthylation de l'ADN au niveau de cytosines (5mC), marques d'histones répressives et ARN non codants. Dans les régions chromatiniennes ouvertes, l'ADN possède majoritairement des cytosines non méthylées ou hydroxyméthylées (5hmC) et les histones portent des marques actives. Voir la section suivante sur la méthylation de l'ADN pour une description plus précise des différents contextes.

Ces complexes contiennent des acteurs épigénétiques (« *writers*/écrivains »), responsables du dépôt de ces marques épigénétiques (Probst et al., 2009). Ils sont constitués également de molécules «lectrices» (« *readers* ») qui interprètent ces marques épigénétiques et provoquent, soit un déroulement de l'ADN par suppression de nucléosomes, soit une augmentation de l'enroulement de l'ADN *via* l'intervention de «remodeleurs» de la chromatine déposant des nucléosomes additionnels. On aboutit alors respectivement, à une conformation dite fermée ou ouverte de la chromatine (Figure 1.2).

Les ARNs non-codants, qui par définition ne codent pas de protéines, participent également à la régulation épigénétique de l'expression des gènes. Ces ARNs peuvent être distingués selon leur taille

(long *versus* petits ARNs, incluant les micro-ARNs). Les longs ARNs non codants, ont la capacité de favoriser l'interaction de protéines et peuvent ainsi, par exemple, maintenir les acteurs épigénétiques à proximité du site à remodeler. Au contraire, ils peuvent aussi séquestrer des protéines ou des micro-ARNs en agissant comme des pièges ou « éponges » (Briggs et al., 2015). Les petits ARNs non-codants, dont les micro-ARNs, régulent le niveau de traduction en protéines des ARN messagers (ARNm), grâce à des homologies de séquences entre ces différents ARNs. Un miARN peut réguler la traduction d'ARNm de plusieurs gènes et, l'expression d'un gène est régulée par plusieurs miARNs (Yi and Fuchs, 2011).

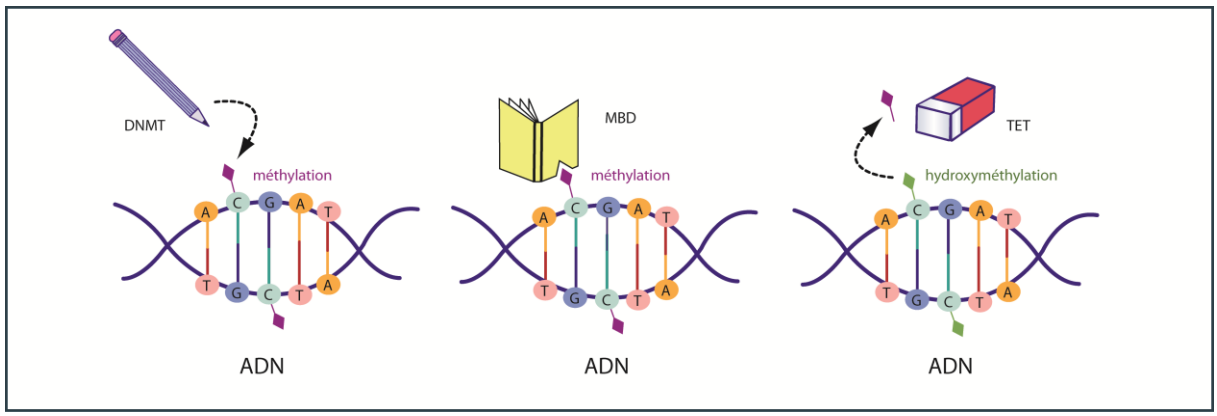


Figure 1.2 : « Ecrivains, lecteurs et modificateurs » : les acteurs moléculaires de l'épigénétique.
Plusieurs acteurs moléculaires interviennent dans les mécanismes épigénétiques : des enzymes dites « écrivains » ou *writers* apposent des marques épigénétiques, d'autres identifient et fixent ces marques puis recrutent d'autres molécules régulatrices de la transcription. Les marques épigénétiques étant réversibles, il existe également des facteurs *erasers* capables de supprimer ou modifier ces marques. Dans l'exemple présenté ici, les ADN méthyl-tranférases (DNMT) permettent la méthylation des cytosines de l'ADN, qui sont reconnues et liées par des protéines ayant un domaine MBD (*methyl-binding domain*). Les enzymes de translocation *Ten-Eleven* (TET), modifient chimiquement les cytosines méthylées en hydroxy-méthyl-cytosines.

1.3. Marques épigénétiques : persistantes, mais modulables selon l'environnement

Déjà dans les années 1940, alors que le support de l'information génétique n'était pas encore démontré, il avait été postulé que des caractéristiques cellulaires pouvaient être héritées, sans qu'il y ait de modifications de l'information génétique (Waddington 1942 dans Waddington, 2012). Lors de l'embryogenèse, la spécification des destins cellulaires se met en place grâce à des mécanismes épigénétiques, déposant, au niveau de la chromatine, des marques spécifiques de l'identité cellulaire (Probst et al., 2009). Ainsi, ces mécanismes sont impliqués dans la différenciation cellulaire, et le maintien de l'identité cellulaire. Comme les marques épigénétiques sont héritables d'une division cellulaire à l'autre, l'épigénome est stable et durable, ce qui permet de maintenir un état

transcriptomique donné, et ainsi maintenir l'identité cellulaire (Probst et al., 2009). Ces mécanismes épigénétiques sont donc cruciaux pour la régulation des fonctions cellulaires.

Toutefois, l'environnement des cellules peut être amené à changer : les cellules s'adaptent alors à ces changements environnementaux. Des marques épigénétiques peuvent ainsi être acquises *de novo*, supprimées (par des protéines « *erasers* », Figure 1.2) ou modifiées en réponse à des stimuli environnementaux (Bale, 2015) nécessitant une adaptation des fonctions cellulaires à plus ou moins long terme. Les mécanismes épigénétiques s'inscrivent donc dans un véritable paradoxe entre stabilité de l'identité cellulaire et adaptation/plasticité à l'environnement (Meaney, 2010). Ces modifications des marques épigénétiques en réponse à l'environnement, est un mécanisme physiologique d'adaptation indispensable aux cellules. Cependant, il arrive que certains changements environnementaux, soient plutôt perçus comme des stress délétères, qui perturbent les mécanismes épigénétiques de manière plus ou moins durables, ce qui engendrent une « maladaptation » des cellules à ce nouvel environnement, avec des conséquences à plus ou moins long terme. C'est ainsi que certains mécanismes pathologiques se mettent en place, ou perdurent dans le temps. Les marques épigénétiques étant réversibles, il est possible, dans certains cas de corriger ces défauts de l'épigénome.

Ainsi, les mécanismes épigénétiques interviennent dans la régulation des gènes, à plusieurs niveaux possibles. Les ARN non codants, la méthylation de l'ADN ou encore les modifications des histones régulent la transcription des gènes. En plus de leur régulation au niveau transcriptionnel, Les ARN non codants sont aussi impliqués dans la régulation post-transcriptionnelle.

2. Méthylation de l'ADN

Parmi les marques épigénétiques, la méthylation de l'ADN est particulièrement étudiée. Cette modification covalente de l'ADN correspond à l'ajout d'un groupement méthyle (-CH3). Chez les mammifères, la méthylation a lieu au niveau des cytosines (Penn et al., 1972; Zemach et al., 2010, (Figure 1.3) et joue un rôle dans la régulation de l'expression de nombreux gènes. La méthylation de l'ADN est impliquée dans divers processus biologiques essentiels, tels que le développement, la régulation de la transcription, l'inactivation du chromosome X, la répression des éléments transposables ou l'empreinte génomique (Li and Zhang, 2014). Sa dérégulation est associée à de nombreuses pathologies (Robertson, 2005).

2.1. Méthylation en contextes CpG et non CpG

Le contexte nucléotidique dans lequel se trouve une cytosine impacte sur la régulation de sa méthylation. Chez les mammifères, la méthylation de l'ADN est présente majoritairement au niveau de cytosines en **contexte dinucléotides Cytosine-Guanine** (**CpG**, Guibert and Weber, 2013). Ce motif étant palindromique et les deux brins de l'ADN présentant généralement le même état de méthylation (Law and Jacobsen, 2010), la conservation du schéma de méthylation pendant la réplication de l'ADN et les divisions cellulaires est facilitée. La méthylation dans ce contexte est donc propagée de façon stable, ce qui suggère que cette marque épigénétique, dans ce contexte, peut relayer des modifications de la régulation de l'expression des gènes, même quand le signal déclencheur, à l'origine du dépôt de ces méthylations, n'est plus présent (Guibert and Weber, 2013).

La méthylation asymétrique de l'ADN (*i.e.* hors contexte CpG, parfois appelée contexte CHG ou CHH, H faisant référence aux bases A, T ou G) a aussi été décrite chez les mammifères, dans certains types cellulaires (Haines et al., 2001; Lister et al., 2009). Contrairement au contexte CpG, la méthylation des cytosines hors contextes CpG doit être (r)établie *de novo* après chaque division cellulaire. Cette méthylation hors contexte CpG, très abondante dans les cellules souches embryonnaires humaines, est perdue au moment de la différenciation cellulaire (Lister et al., 2009). La méthylation en contexte **non CpGs** est limitée dans les tissus différenciés (Pelizzola and Ecker, 2011). Les méthylations hors contextes CpG sont toutefois nombreuses dans le cerveau postnatal, notamment dans les neurones, et peuvent avoir des effets plus ou moins marqués sur la régulation de l'expression des gènes (Guo et al., 2014; Lister et al., 2013; Schultz et al., 2015)

2.2. Les protéines responsables de la méthylation de l'ADN

Des protéines particulières, appelés les ADN méthyltransférases (*DNMTs – DNA methyltransferase*), sont responsables de la méthylation de l'ADN. Elles utilisent, pour cela, la S-adénosine-méthionine (SAM), donneuse de groupement méthyle, comme substrat.

Plusieurs DNMTs ont été répertoriées dans le passé : DNMT1, DNMT2, DNMT3A, DNMT3B et DNMT3L. Toutefois, la protéine DNMT2, identifiée par homologie de séquence, méthyle plutôt l'ARN que l'ADN (Goll et al., 2006) et n'est donc plus considérée actuellement comme une DNMT.

Après la réplication, la méthylation de l'ADN en contexte CpG est maintenue par **DNMT1** (Bestor et al., 1988), qui reconnait préférentiellement les hémi-méthylation des cytosines du CpG (Bestor, 1992). **DNMT3A et DNMT3B** sont responsables de la méthylation *de novo* de l'ADN (Okano et al., 1998), et introduisent ainsi des méthylations en contextes CpG ou non (Gowher and Jeltsch, 2018).

Bien que ce clivage de fonctions entre les DNMTs (maintenance/*de novo*) soit couramment admis, des études récentes montrent qu'il est simpliste, puisqu'il semble que chacune des DNMTs puissent être impliquées dans ces deux processus de méthylation de l'ADN (Gowher and Jeltsch, 2018; Jeltsch and Jurkowska, 2014). Le facteur **DNMT3L** (pour *DNMT3-like*) ne peut pas méthyler l'ADN car il ne possède pas de domaine catalytique, mais il participe au processus de méthylation en activant DNMT3A et DNMT3B, en régulant leur oligomérisation et leur localisation nucléaire (Jeltsch and Jurkowska, 2016).

Concernant la méthylation en contextes CpG dans le cerveau postnatal, Lister et collaborateurs ont montré, que sa mise en place coïncidait avec l'augmentation, de la quantité d'ARN et de protéines DNMT3A, ce qui suggère l'implication de cette protéine en particulier, dans la mise en place *de novo* de cette marque dans cet organe (Lister et al., 2013).

Les protéines DNMTs peuvent être recrutées de manières spécifiques ou non sur des *loci* particulier de l'ADN, à travers l'interaction avec d'autres protéines (*e.g.* EZH2, HP1), mais les acteurs de ce recrutement ne sont pas tous bien définis (Hervouet et al., 2018). Par ailleurs, leur activité catalytique est régulée par plusieurs mécanismes, notamment des modifications post traductionnelles, des modifications de conformations allostériques, et des partenaires moléculaires répresseurs ou activateurs (Jeltsch and Jurkowska, 2016).

2.3. Répartition génomique non homogène de la méthylation : les îlots CpG

La méthylation de l'ADN est une marque épigénétiques mutagène - une cytosine méthylée peut être modifiée spontanément en thymine par déamination (Bird, 1986) - mais elle est limitée dans le génome (Bird, 1980; Coulondre et al., 1978). Chez les mammifères, environ 3.5 à 4.5% des cytosines de l'ADN sont méthylées, dans les tissus adultes en fonction des types cellulaires (Gowher and Jeltsch, 2018).

Les **dinucléotides CpG** sont sous-représentés dans le génome et ne sont pas répartis de façon homogène au sein du génome. Ainsi, il existe des régions où les CpG sont regroupés, formant ainsi des séquences génomiques riches en CpG appelées **îlots CpG** (en anglais, *CGi* pour ***CpG islands***, Bird, 1986; Gardiner-Garden and Frommer, 1987). Ces îlots CpG, souvent associés à des régions régulatrices en amont (régions promotrices) de gènes régulés par méthylation de l'ADN (Saxonov et al., 2006), sont pour la plupart, non méthylés. La présence de cette marque au niveau de ces îlots induit la répression de l'expression des gènes concernés.

De même, **la méthylation de l'ADN** n'est pas répartie de façon homogène au sein du génome. Il existe des régions pauvres en méthylation de l'ADN, notamment les îlots CpGs, les *enhancers* ou encore des *insulators* (Schlesinger et al., 2013). La méthylation hors contexte CpG se situe

principalement dans le corps des gènes, surtout au niveau de gènes très transcrits et des gènes présentant un fort taux de pré-ARNm, ce qui suggère que cette marque épigénétique, dans ces contextes, pourraient être impliquée dans des mécanismes associés à l'épissage (Lister et al., 2009).

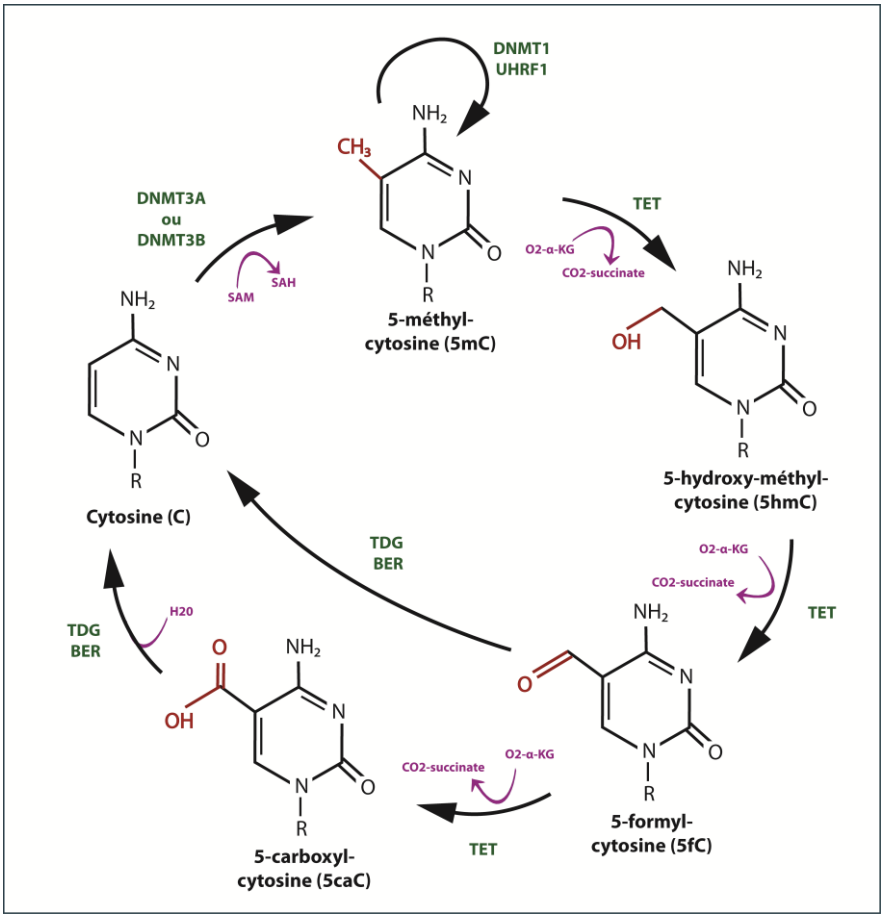


Figure 1.3 : Représentation simplifiée des modifications chimiques portées par la cytosine au niveau de l'ADN.

Les cytosines de l'ADN portent des modifications chimiques, qui n'affectent pas la séquence de l'ADN (l'ordre des nucléotides) mais qui sont informatives pour la cellule. Ainsi, une cytosine peut être successivement modifiée par les ADN méthyl-transférases (DNMT3A ou DNMT3B) en 5-méthyl-cytosine (5mC), puis par les enzymes de translocation *Ten-Eleven* (TET) en 5-hydroxy-méthyl-cytosine (5hmC), 5-formyl-cytosine (5fC) et 5-carboxyl-cytosine (5caC). La forme 5caC peut être remplacée par une cytosine non modifiée chimiquement, *via* des étapes non détaillées ici, qui impliquent les enzymes *TDG* et *BER*. Pour plus de détails, voir la revue de Wu et Zhang (2014). Ainsi, une déméthylation active des cytosines est assurée par les enzymes *TET*, tandis qu'une déméthylation passive, par dilution de la méthylation au cours de la réplication de l'ADN, est également possible. Toutefois, l'enzyme DNMT1, couplée à UHRF1 (*ubiquitin-like plant homeodomain and RING finger domain 1*), permet de maintenir la méthylation au cours de la réplication de l'ADN, au niveau de la plupart des cytosines en contexte CpG, devant rester méthylées. **Abréviations** : **BER** : *base excision repair*, **KG** : kétoglutarate, **SAH** : S-adénosyl-homocystéine, **SAM** : S-adenosyl-méthionine, **TDG** : *thymine DNA glycosylase*.

2.4. Méthylation de l'ADN et expression de gènes : activation ou répression ?

La méthylation de l'ADN est associée, à la **régulation positive ou négative de l'expression** des gènes, selon le *loci* concerné.

L'absence de méthylation de l'ADN est généralement associée à une chromatine accessible, permettant aux facteurs de transcription de se fixer à l'ADN (Thurman et al., 2012) même si des études récentes suggèrent des situations plus complexes.

La méthylation de l'ADN au niveau de régions promotrices riches en CpGs est fortement corrélée à de la répression transcriptomique (Keshet et al., 1986). En revanche, dans les régions génomiques pauvres en CpGs, généralement situés dans le corps du gène ou en régions intergéniques, son effet sur la transcription est plus variable et dépend notamment du contexte génomique (Jones, 2012).

2.5. Implication de la méthylation de l'ADN dans divers processus biologiques

Elle peut directement moduler la fixation de facteurs de transcription, au niveau de séquences régulatrices, comme les promoteurs ou les *enhancers*, régulant ainsi l'**expression de certains gènes** (Tate and Bird, 1993). La méthylation de l'ADN peut également moduler cette fixation dans le corps du gène, jouant alors un rôle dans l'**épissage alternatif** : sa présence, au niveau de certaines régions génomiques correspondant à des exons alternatifs, peut empêcher la fixation de facteur de transcription (*e.g.* CTCF, *CCCTC-Binding Factor*) qui favorisent l'inclusion de cet exon dans l'ARNm mature (Maunakea et al., 2010, 2013; Shukla et al., 2011b).

Ainsi, la méthylation de l'ADN est impliquée dans divers **processus biologiques**. Des études menées sur des souris mutantes pour *Dnmt1, Dnmt3a* ou *Dnmt3b* ont mis en évidence le rôle essentiel de ces facteurs dans le **développement de l'embryon** (Li et al., 1992; Okano et al., 1999). Parmi les autres processus biologiques impliquant la méthylation de l'ADN, se trouvent la **différentiation** des cellules souches (Meissner, 2010), l'**expression des gènes liés à l'empreinte** (voir ci-dessous, Barlow, 2011), l'**inactivation du chromosome X** (Robertson, 2005) ou encore de la **répression des éléments transposables** (Bestor, 1998).

La méthylation de l'ADN est également importante pour le développement du cerveau et son fonctionnement (voir Introduction, Partie 2 - Le cerveau en développement).

Ainsi, la dérégulation de la méthylation de l'ADN est associée à de nombreuses pathologies, comme par exemple des **cancers**, le **syndrome** d'immunodéficience **ICF** (syndrome immuno-déficience combinée (I), instabilité de l'hétérochromatine paracentromérique (*C*) et dysmorphie faciale (*F*)), ou encore des maladies avec des atteintes neurologiques, tels que le **syndrome d'Angelman** ou le **syndrome de l'X fragile** (Baylin and Jones, 2011; Robertson, 2005).

Au vu du nombre important de processus biologiques impliquant la méthylation de l'ADN, Il y a un intérêt particulier à mieux comprendre les mécanismes impliqués dans l'établissement, le maintien et l'interprétation de la méthylation de l'ADN au sein des cellules.

2.6. Méthylation allèle-spécifique : les gènes soumis à l'empreinte

Chez les organismes diploïdes, certaines régions génomiques peuvent posséder un profil de méthylation de l'ADN différents au niveau de leurs deux allèles (Bock, 2012; Song et al., 2013). Ce différentiel peut dépendre de l'origine parentale de l'allèle. C'est le cas des **gènes liés à l'empreinte** (*IG*, pour ***imprinting genes***, Tremblay et al., 1995). L'empreinte génomique se définit comme un mécanisme épigénétique qui conduit à l'expression différentielle entre les deux allèles, selon leur origine maternelle ou paternelle (Perez et al., 2016). Les gènes soumis à l'empreintes sont grandement impliqués dans la développement (Perez et al., 2016). Ils interviennent notamment dans de nombreux processus neuro-développementaux (neurogenèse, migration neuronale, apoptose, développement des axones et dendrites, voir Introduction, Partie 2 - Le cerveau en développement Perez et al., 2016). Ils sont également impliqués dans diverses fonctions cérébrales chez l'adulte, tels que l'apprentissage et la mémoire, la transmission synaptique, et les comportements cognitifs et émotionnels (Perez et al., 2016).

Ces gènes soumis à l'empreinte sont régulés par méthylation différentielle de l'ADN sur chacun des allèles parentaux. Cette méthylation différentielle se met en place dans les cellules germinales, et se maintient durant le développement, à l'exception de quelques régions pouvant être reprogrammées au cours du temps (Reik and Walter, 2001).

La méthylation différentielle a lieu au niveau de régions génomiques spécifiques, appelées régions de contrôle de l'empreinte (*ICRs*, *imprinting control regions*), contrôlant l'expression des gènes d'un domaine génomique. La protéine insulatrice CTCF (*i.e.* protéine qui sépare des régions génomiques en différents domaines fonctionnels), joue un rôle important dans la régulation des gènes soumis à l'empreinte, en régulant l'accès des promoteurs aux *enhancers* (Klenova et al., 2002). Cette protéine permet l'expression d'un seul allèle parental, en se fixant uniquement à l'allèle non méthylée.

2.7. Hydroxyméthylation et déméthylation de l'ADN

Bien que la méthylation de l'ADN soit une marque épigénétique plutôt stable, faisant de cette marque un biomarqueur fiable, le profil de méthylation des cellules peut être modifié, dans le cadre de fonctions biologiques bien particulières (*e.g.* plasticité neuronale, voir Introduction, Partie 2 - Le cerveau en développement) ou en réponse à l'environnement (Lister et al., 2013; Pelizzola and Ecker, 2011).

Il existe donc des mécanismes impliquant les « *writers* » DNMTs permettant de méthyler l'ADN (voir section sur les DNMTs), mais aussi des mécanismes permettant d'effacer cette marque. Cet effacement peut avoir lieu de manière passive, par dilution de la marque au cours des divisions cellulaires, lorsque DNMT1 ne la maintient pas, ou de manière active ce qui implique l'intervention des protéines «*erasers* » (Figure 1.2, Wu and Zhang, 2014). Les protéines TET (*Ten Eleven Translocation proteins*) peuvent oxyder la méthyl-cytosine et ainsi la transformer successivement en 5-hydroxy-méthyl-cytosine (5hmC), 5-formyl-cytosine (5fC) et 5-carboxyl-cytosine (5caC). Cette dernière peut être excisée de l'ADN et être remplacée par une cytosine non méthylée (voir détails dans la revue Wu and Zhang, 2014). Contrairement à la méthylation (5mC), l'hydroxy-méthylation (5hmC) est peu reconnue par DNMT1, ce qui peut engendrer une déméthylation passive de l'ADN (Tahiliani et al., 2009).

Il est important de noter que les différentes formes de cytosines (5hmC, 5fC et 5caC) ne sont pas nécessairement des éléments intermédiaires du cycle de déméthylation active de l'ADN, et peuvent jouer un rôle biologique. Notamment, l'hydroxyméthylation des cytosines est maintenant prise en compte comme une marque épigénétique *per* se, car elle peut moduler l'expression de gènes, en recrutant divers modeleurs de la chromatine (Hahn et al., 2014; Sadakierska-Chudy et al., 2015). Cette marque est particulièrement présente dans le cerveau adulte et en développement, et est activée au cours du développement (Guo et al., 2014; Kriaucionis and Heintz, 2009; Lister et al., 2013). Chen et collaborateurs (2013), ont par exemple montré, dans l'hippocampe en développement, que le profil de méthylation, et plus particulièrement l'équilibre entre méthylation et hydroxy-méthylation, régule la différenciation neuronale et la maturation spatiotemporelle des neurones (Chen et al., 2013).

3. Résumé

Les mécanismes épigénétiques sont impliqués dans des rôles moléculaires primordiaux. Ils modulent l'expression des gènes de façon héritable, sans modifier la séquence de l'ADN, mais en changeant l'accessibilité de la chromatine. Ces marques épigénétiques stables, sont ainsi responsables de l'établissement et du maintien du destin cellulaire.

Parmi les marques épigénétiques, la méthylation de l'ADN est particulièrement importante et très étudiée. Cette méthylation est impliquée dans de nombreux processus biologiques essentiels, tels que l'inactivation du chromosome X, l'empreinte génomique, le développement - notamment dans le neuro-développement – ou encore des fonctions cérébrales variées.

Bien que stables, les marques épigénétiques peuvent évoluer au cours du temps, notamment en réponses à un changement d'environnement et permettre à la cellule de s'adapter à son

nouveau milieu. Toutefois, certains changements d'environnements perturbent de façon notable les mécanismes épigénétiques, engendrant une « mal-adaptation » des cellules, pouvant aboutir à la mise en place de pathologies. L'exposition prénatale à l'alcool fait partie des stress prénataux altérant les mécanismes épigénétiques, et notamment la méthylation de l'ADN, ce qui peut avoir des conséquences plus ou moins graves sur le développement de l'embryon. Avant de présenter ce stress prénatal, nous allons introduire quelques notions au sujet du développement du cerveau, organe principalement affecté par ce stress.

Partie 2 : Le cerveau en développement

1. Composition cellulaire du cerveau et fonctions associées

Le cerveau des vertébrés est un tissu constitué de progéniteurs neuraux, de neurones, et de diverses populations de cellules gliales (astrocytes, oligodendrocytes, microglie).

1.1. Les cellules gliales

Les cellules gliales formant le tissu conjonctif nerveux, représentent une proportion non négligeable du cerveau des mammifères, puisqu'elle constitue 33 à 66% de la masse du cerveau, selon l'espèce considérée (Herculano-Houzel, 2014). Elles sont impliquées dans diverses fonctions biologiques du cerveau, bien au-delà de leur rôle de soutien aux neurones (névroglie, ou « glue nerveuse », d'où le nom donné à ces cellules, Jäkel and Dimou, 2017).

Les astrocytes sont les cellules gliales les plus abondantes dans le cerveau adulte (Kettenmann and Ransom, 2004) et possèdent une multitude de fonctions biologiques. Ils sont impliqués notamment dans le maintien de la barrière hémato-encéphalique, qui protège activement le cerveau (Daneman, 2012), mais aussi dans le maintien de l'homéostasie (Jäkel and Dimou, 2017).

Les oligodendrocytes constituent la myéline, constituant qui isole les axones et permet une transmission du signal nerveux rapide entre neurones (Nave, 2010). Cette myéline sert également de support nutritif aux neurones (Nave, 2010).

Les cellules de la microglie sont des macrophages résidents du système nerveux central (SNC), impliquées dans le système immunitaire et le maintien du SNC (Kettenmann et al., 2011 ; Ginhoux and Prinz, 2015). Contrairement aux autres cellules du cerveau, qui ont une origine neuro-ectodermique, la microglie provient de progéniteurs situés dans le sac vitellin, qui colonisent le cerveau au cours du développement (Kettenmann et al., 2011 ; Ginhoux and Prinz, 2015). La microglie joue un rôle de sentinelle, permettant de détecter les signes d'une invasion par un pathogène. Lors d'une inflammation, ces cellules favorisent la réparation et le remodelage du tissu endommagé (Ginhoux and Prinz, 2015). Par ailleurs, la microglie joue un rôle important aussi dans l'élagage synaptique pendant le développement (cf ci-dessous, étapes clefs du développement du cerveau, Wu et al., 2015), ainsi que dans la modulation synaptique en conditions basales ou pathologiques (Bessis et al., 2007 ; Hong et al., 2016).

1.2. Les neurones

Les neurones sont les cellules nerveuses spécifiques du SNC, qui transmettent des informations à l'ensemble du corps, à travers des signaux électriques. Ils sont constitués d'un corps cellulaire, de dentrites et d'un axone, ces derniers étant des extensions cytoplasmiques spécialisées (Lodish et al., 1999, Figure 1.4).

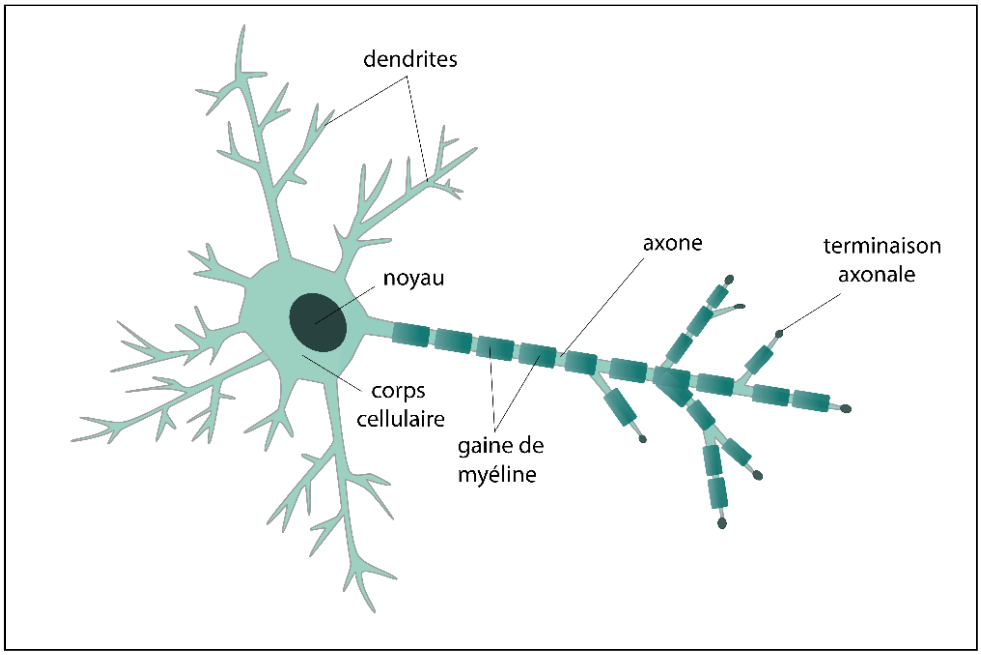


Figure 1.4: Structure schématique d'un neurone.

Le corps cellulaires contenant le noyau, est le lieu de synthèse et de dégradation des protéines, à l'exception de quelques protéines, produites dans les dendrites (Lodish et al., 1999). Le transport des protéines le long de l'axone ou des dendrites est assuré par un réseau organisé de microtubules (Lodish et al., 1999). L'axone, très souvent unique, est une structure spécialisée dans la conduction du potentiel d'action, impulsion électrique allant du corps cellulaire vers l'extrémité terminale de l'axone (Lodish et al., 1999). Des potentiels d'actions successifs stimulent le neurone et permettent le relargage de **neurotransmetteurs** (*e.g.* dopamine, sérotonine, glutamate, adrénaline, noradrénaline, acide γ-aminobutyrique - GABA), molécules chimiques excitatrices ou inhibitrices, permettant la transmission du signal aux neurones environnants (Valenzuela et al., 2011, Figure 1.6). Ces neurotransmetteurs se fixent à leurs récepteurs spécifiques situés au niveau des dendrites des neurones adjacents (Valenzuela et al., 2011, Figure 1.6). Ces échanges moléculaires ont lieu au niveau d'une synapse : l'axone d'un neurone donné peut y interagir simultanément avec des dendrites de différents neurones, ce qui permet d'induire des réponses moléculaires simultanées au sein de ses neurones (Lodish et al., 1999).

Les dendrites, nombreuses et ramifiées, permettent, comme nous venons de le voir, de réceptionner l'information électrique transmise par un neurone, par l'intermédiaire des neurotransmetteurs. Ces derniers, une fois relargués au niveau de la synapse, se fixent à leurs récepteurs spécifiques, situés aux niveaux des épines dendritiques du neurone post-synaptique. Ces épines permettent d'augmenter la surface d'échanges entre les neurones. Cette fixation modifie le potentiel électrique de la membrane de ce neurone, en changeant sa perméabilité aux ions. Ainsi, les dendrites convertissent l'information transmise par les neurotransmetteurs, en un signal électrique, qui se propage de l'épine dendritique, vers le corps cellulaire du neurone (Lodish et al., 1999).

Les neurones du SNC ont de longs axones et dendrites, ce qui leur permet de recevoir et transmettre l'information, à la fois sur de longues distances et, avec de nombreux neurones différents.

Dans les organismes complexes, plusieurs types de neurones se distinguent et forment des circuits neuronaux, répartis dans diverses structures du cerveau (cortex, hippocampe, …). Des cellules interneuronales, permettent de faire un lien entre deux neurones ne pouvant pas l'être, ce qui favorise la transmission du signal, et permet de coordonner des processus biologiques complexes (Lodish et al., 1999).

Cette organisation complexe du cerveau, nécessaire à son bon fonctionnement, résulte en grande partie de nombreux évènements biologiques ayant lieu lors du développement cérébral. En effet, même si les connexions interneuronales peuvent être remodelées au cours de la vie de l'individu, certains défauts susceptibles de se mettre en place pendant le neuro-développement peuvent être irréversibles.

2. Les étapes clefs du développement du cortex cérébral

Le cerveau est un organe qui se développe très tôt, dès le début du premier semestre de grossesse, et qui continue à se développer *in utero*, mais aussi après la naissance, durant l'enfance et l'adolescence (Ayala et al., 2007). Plusieurs évènements indispensables, et finement régulés, se succèdent et assurent le bon développement du cerveau. Les grandes étapes du développement du cortex cérébral, sont plutôt bien conservées entre espèces. De ce fait, les observations faites à partir d'un modèle murin peuvent généralement être appliquées à l'Homme (Kriegstein et al., 2006).

Au début du de la grossesse, le **tube neural** se ferme (Ayala et al., 2007) et, forme par la suite le cerveau, la moelle épinière et les vertèbres du fœtus (Figure 1.5 et Figure 1.6). Plus précisément, la partie la plus antérieure du tube neural forme une protubérance, appelée prosencéphale (Ayala et

al., 2007). En se développant, ce dernier forme en partie antérieur le télencéphale, qui donne naissance au cerveau antérieur, constitué des hémisphères cérébraux, des bulbes olfactifs et des ganglions de la base (Kriegstein et al., 2006).

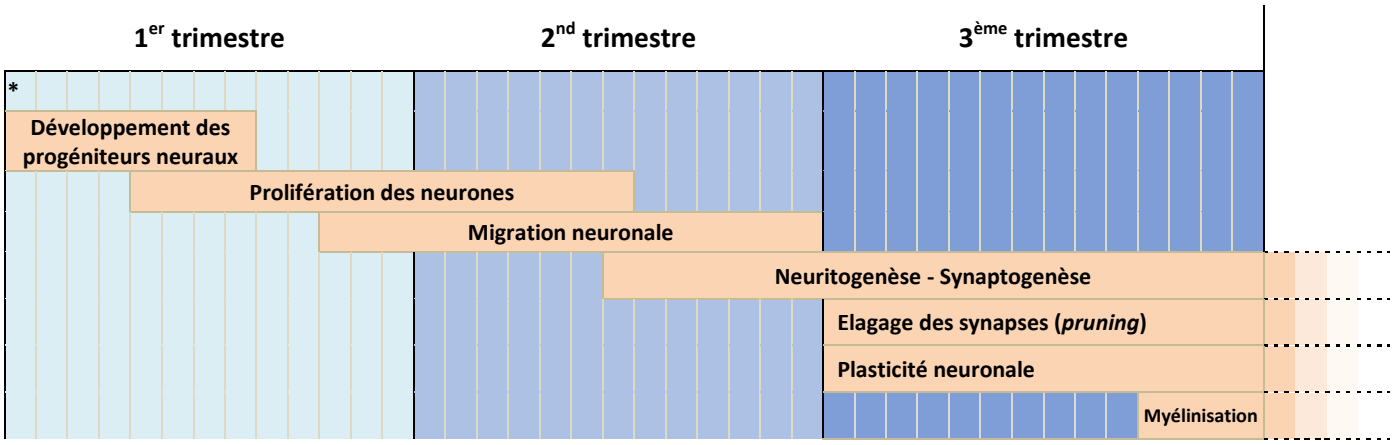


Figure 1.5 : Etapes clés, ayant lieu au cours de la grossesse, permettant la formation du cerveau. L'astérisque symbolise la fermeture du tube neural, au tout début de grossesse. Les traits horizontaux délimitent les semaines de grossesse. Certains évènements, comme la synaptogenèse, ont lieu également après la naissance (en pointillé). Les limites temporelles de chaque évènement sont fournies ici à titre indicatif, étant donné que la durée et la période de chaque étape peut être légèrement variables, d'une structure cérébrale à une autre. Cette figure est inspirée de la figure proposée par Zieff and Schwartz-Bloom, 2008.

Durant les dix premières semaines de grossesse, les **progéniteurs neuraux** des cellules gliales et des neurones, se forment et **prolifèrent** massivement (Kriegstein et al., 2006 ; Figure 1.5 et Figure 1.6). Ces proliférations sont localisées au niveau de la **zone ventriculaire** (**VZ**), qui, après développement du cerveau, constituera la couche la plus interne du cortex cérébral. Avant que la neurogenèse ne commence, les cellules progénitrices se divisent de façon symétrique, ce qui signifie que les deux cellules filles résultantes sont des progéniteurs neuraux (Ayala et al., 2007; Kriegstein et al., 2006, Figure 1.6). Ces divisions cellulaires successives permettent ainsi d'augmenter le nombre de progéniteurs. Dans le cortex cérébral, ces cellules progénitrices sont des cellules neuroépithéliales se présentant sous la forme de cellules radiales (Noctor et al., 2004). Après cette étape de prolifération cellulaire, ces progéniteurs se divisent de manière asymétrique : une des deux cellules filles est un progéniteur neural, alors que l'autre est un jeune neurone (Kriegstein et al., 2006 ; Ayala et al., 2007). Vers la fin de la neurogenèse, les progéniteurs neuraux, qui ont contribué à former les neurones, se divisent à nouveau de manière symétrique, et donnent naissance, non plus à des neurones, mais à des astrocytes (Qian et al., 2000). Le programme de différenciation cellulaire de ces cellules progénitrices est donc modifié au cours du développement.

Les neurones, formés au niveau de la zone ventriculaire, migrent ensuite hors de cette zone, et forment le cortex cérébral (Figure 1.5 et Figure 1.6), constitué, *in fine*, de six couches de neurones interconnectés, chez l'Homme ou la souris (Kriegstein et al., 2006). Une première vague de neurones migrent autour du 11ème jour de gestation chez la souris (*embryonic day* E11 ; Ayala et al., 2007). Cette période est équivalente à la fin du premier trimestre de grossesse pour l'humain. Ces neurones en migration forment la pré-plaque, nouvelle couche corticale située juste au-dessus de la VZ (Kriegstein et al., 2006 ; Ayala et al., 2007). Une zone sub-ventriculaire (SVZ) se forme ensuite entre la pré-plaque et la VZ, au niveau du cortex dorsal. Cette région constitue aussi une zone prolifération, où des progéniteurs neuronaux se divisent de façon symétriques pour donner naissance à deux neurones (Kriegstein et al., 2006). Une seconde vague de migration, qui a lieu vers E13, sépare la pré-plaque en deux nouvelles couches : (**1**) une couche extérieure, appelée zone marginale (**MZ**), composée des cellules de Cajal-Retzius, neurones différenciées provenant de la première vague de migration, (**2**) et une couche plus profonde, la sous-plaque (Ayala et al., 2007). En parallèle de la neurogenèse corticale, plusieurs vagues de **migrations de cellules neurales** se succèdent ensuite tout au long du second trimestre, positionnant les cellules, et notamment les neurones, au sein des différentes couches corticales, qui se mettent en place entre la zone marginale et la sous-plaque (Figure 1.5 et Figure 1.6, Kriegstein et al., 2006). Ainsi, lors du développement, chaque neurone est positionné dans une région bien définie du cerveau. La position laminaire d'un neurone est guidée par de nombreux facteurs (voir ci-dessous), mais son positionnement semble déterminé par le jour où il a été généré. En effet, les neurones générés le même jour vont se retrouver dans la même couche corticale (Kriegstein et al., 2006). Il est intéressant de noter que les couches corticales les plus récentes sont les plus externes (Marín and Rubenstein, 2003). Les neurones proliférant au niveau de la zone ventriculaire (zone interne), migrent donc au travers des couches déjà mises en place pour rejoindre leur destination. Chaque couche corticale étant impliquée dans des fonctions spécifiques, le positionnement correct des neurones est essentiel au bon fonctionnement du cerveau (Ayala et al., 2007).

Deux types de migrations majeures existent : la **migration radiale** des neurones, où les cellules migrent le long des cellules gliales, de la zone prolifération (zone ventriculaire) vers la surface extérieure, et la **migration tangentielle**, suivant une direction orthogonale par rapport à la migration radiale (Figure 1.6, Marín and Rubenstein, 2003). La migration radiale est à l'origine de l'organisation laminaire précédemment décrite, au niveau du SNC (cortex cérébral, cortex cérébelleux) et d'autres structures cérébrales (*e.g.* striatum et thalamus, Ayala et al., 2007). La migration tangentielle ne concerne qu'une sous-population de cellules (*e.g.* oligodendrocytes corticaux, interneurones exprimant l'acide γ-aminobutyrique - interneurones GABAergique).

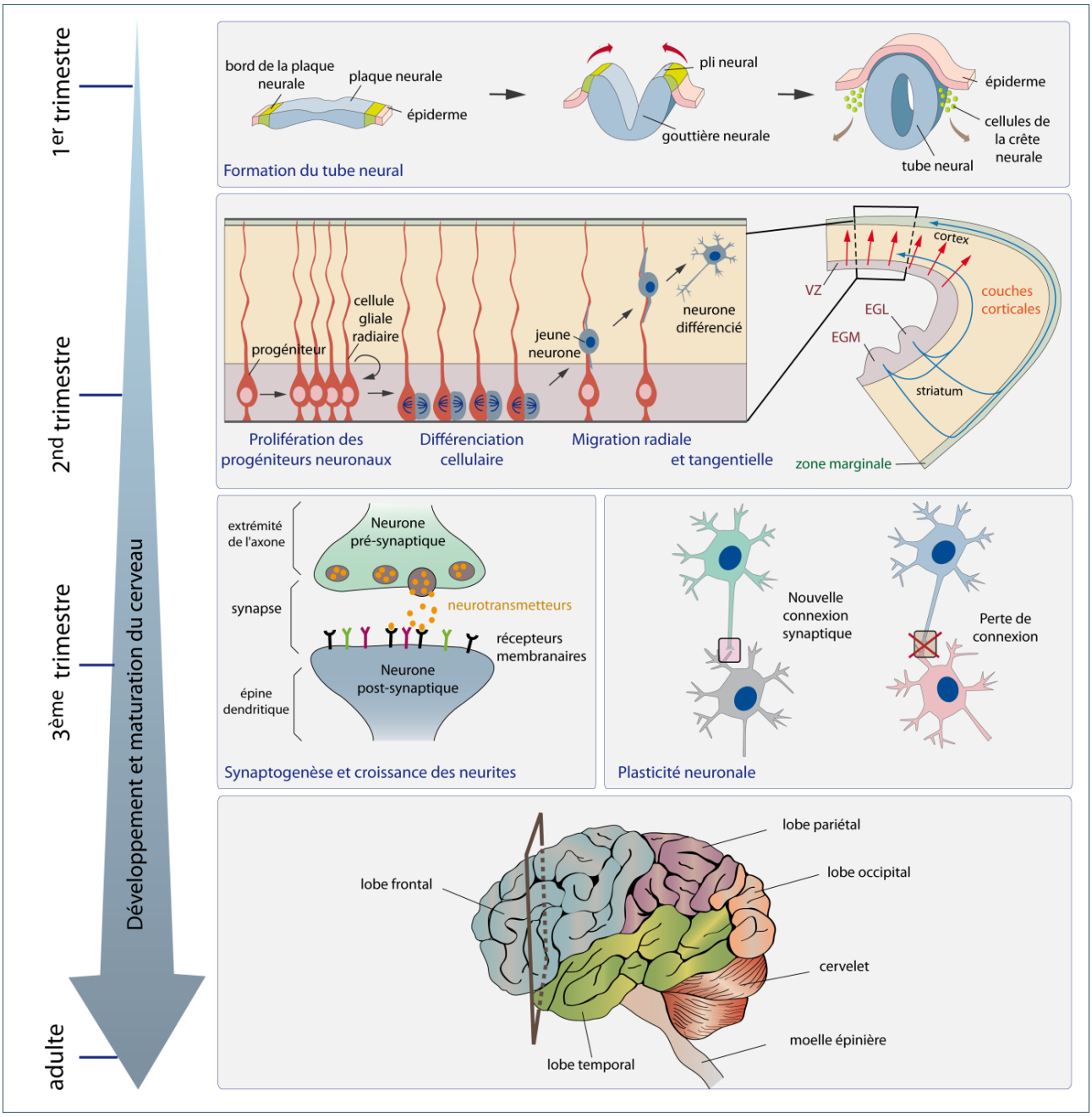


Figure 1.6 : Principales étapes du développement du cortex cérébral.

Au début du premier trimestre de grossesse, le tube neural se ferme et forme par la suite, le cerveau, la moelle épinière et les vertèbres du fœtus. Les progéniteurs neuraux prolifèrent ensuite dans le cortex cérébral, générant de jeunes neurones. Ces derniers migrent de manière tangentielle (mouvement parallèle aux couches corticales) et radiale (de la surface interne vers la surface extérieure) le long de cellules de la glie radiaire. Ces migrations radiales, ont lieu principalement pendant le second trimestre de grossesse, mettant en place un cortex laminaire formé de six couches corticales, aux fonctions et morphologies spécifiques. Lors du troisième trimestre de grossesse, la croissance des neurites et la synaptogenèse permettent la formation des connexions inter-neuronales. Ces connexions évoluent continuellement, au cours de la vie, en fonction de l'activité cérébrale : on parle alors de plasticité cérébrale. Comme le cerveau se développe tout au long de la grossesse, toute exposition prénatale à l'alcool peut altérer le cerveau et avoir des conséquences à long terme. Le plan indiqué sur le cerveau indique la position de la coupe transversale schématisée ci-dessus, qui montre les migrations neuronales. **Abréviations** : **EGL** : éminence ganglionnaire latérale, **EGM** : éminence ganglionnaire médiane, **VZ** : zone ventriculaire.

Deux voies majeures de migration tangentielle sont répertoriées : une migration qui part de l'éminence ganglionnaire médiane (EGM) vers le cortex et l'hippocampe, et une migration qui persiste jusqu'à l'adolescence, allant de l'éminence ganglionnaire latérale (EGL), vers le bulbe olfactif (Figure 1.6, Kriegstein and Noctor, 2004 ; Marín and Rubenstein, 2003). Ces différentes migrations sont très régulées spatio-temporellement (Ayala et al., 2007), par de nombreux mécanismes moléculaires.

La **neuritogenèse** débute au milieu du second trimestre et se poursuit au-delà de la naissance (Zieff and Schwartz-Bloom, 2008, Figure 1.5 et Figure 1.6). Cette étape consiste en la formation des neurites, c'est-à-dire les dendrites et l'axone du neurone, les deux types de prolongements cytoplasmiques neuronaux indispensables à la transmission de l'influx nerveux. La **synaptogenèse** se met alors en place, et permet la communication inter-neuronale, médiée par les neurotransmetteurs, au niveau des épines dendritiques, d'une part, et des terminaisons axonales d'autre part. Ces connexions synaptiques forment un véritable réseau dans lequel les informations sensorielles circulent et sont interprétées.

Les neurones sont également recouverts d'une gaine de myéline, générée par les oligodendrocytes, ce qui permet une transmission du signal électrique inter-neuronal rapide et à grande distance (Williamson and Lyons, 2018). Cette **myélinisation** débute tardivement, pendant le troisième trimestre de grossesse, et se prolonge au-delà de la naissance (Zieff and Schwartz-Bloom, 2008).

Lorsque la plupart des connexions sont générées, une étape d'**élagage synaptique** est nécessaire. Cet évènement, qui a lieu à partir du troisième trimestre et se poursuit après la naissance (Figure 1.5), permet d'éliminer les neurones anormaux par **apoptose** (Zieff and Schwartz-Bloom, 2008). Cette étape est cruciale pour maintenir une transmission du signal efficace.

Cette étape d'élagage synaptique a lieu de façon simultanée à la **plasticité neuronale** : une fois mis en place, le réseau neuronal n'est pas figé, au contraire il n'a de cesse que d'évoluer. A chaque apprentissage et nouvelle mémorisation, ce réseau se modifie et intègre ces nouvelles informations. Ainsi, des connexions cérébrales se créent, d'autres s'amplifient ou encore se perdent (Figure 1.6). Cette réorganisation du réseau cérébral - ou plasticité neuronale - a lieu tout au long de la vie et permet l'efficacité cérébrale dans la transmission des signaux.

Ainsi, le cerveau se développe selon un schéma bien défini d'un point de vue spatio-temporel, où de nombreuses étapes clés se succèdent ou coexistent. Chacune de ces étapes de développement et de maturation du cerveau, qui sollicite l'expression de gènes spécifiques, doit donc être impérativement réalisée dans une fenêtre de temps bien précise pour assurer le neuro-développement (Watson, 2015). Ces étapes clefs, n'ont pas lieu de manière homogène dans les différentes structures constituant le cerveau (Zieff and Schwartz-Bloom, 2008). Ainsi, certaines régions sont matures plus rapidement que d'autres. De nombreux mécanismes biologiques sont nécessaires à la régulation de l'expression de ces gènes au cours du temps, et permettent ainsi le bon déroulement du développement du cerveau et son intégrité.

3. Mécanismes de régulation

Comme cette thèse porte sur l'étude d'un stade de développement murin équivalent au second semestre de grossesse chez l'humain, je me suis intéressée aux mécanismes régulant les évènements ayant lieu durant cette période de développement, et en particulier à la migration neuronale.

3.1. Régulation de la prolifération neuronale

Des études ont montré que l'expression de certains gènes pouvait avoir un impact sur la quantité de neurones présents dans le cerveau. Par exemple, le gène suppresseur de tumeur *Pten* semble réguler négativement la prolifération neuronale, puisque une mutation au niveau de ce gène est associée à une prolifération prolongée des cellules souches neurales (Groszer et al., 2006). A l'inverse, une mutation au niveau des gènes *MCPH* (microcéphaline) ou *ASPM* (*Abnormal Spindle-like Microcephaly-Associated gene*), engendrent des microcéphalies, caractérisées par une diminution drastique de la taille du cortex (Bond et al., 2002; Woods et al., 2005). Ce dernier gène pourrait réguler la prolifération et la différentiation des progéniteurs neuraux (Fujimori et al., 2014). Une étude récente menée chez le furet suggère que le gène ASPM est impliquée dans l'affinité des cellules gliales avec la surface ventriculaire (Johnson et al., 2018).

3.2. Régulation de la migration neuronale

La migration neuronale, une des étapes clefs du développement du cerveau, est finement régulée à plusieurs niveaux. Le mouvement des cellules se produit grâce aux réarrangements des constituants du cytosquelette des neurones (*e.g.* les faisceaux de **microtubules**, filaments d'**actine**), en réponses à des signaux extracellulaires (Ayala et al., 2007). Ces signaux de guidage sont reconnus par la cellule en mouvement, grâce à des récepteurs membranaires présents à sa surface, qui relaient l'information en activant des voies de signalisation capables de moduler le cytosquelette

(Ayala et al., 2007). Ainsi, l'actine, les microtubules et leurs protéines associées (*e.g.* filamine 1*,* qui interagit avec l'actine, *MAP1A, MAP1B, MAP2* - ***Microtubules associated proteins*** *-, Lis1 (Lissencephaly 1) ou encore DCX (doublecortin), associées aux* microtubules). Des mutations au niveau de ces protéines associées à l'actine ou aux microtubules sont associées à des défauts de la migration neuronale, engendrant des désordres congénitaux (Ayala et al., 2007).

La migration neuronale dépend aussi de **signaux extracellulaires**. Ces signaux sont nombreux et peuvent agir à plus ou moins grandes distances. L'un des signaux « longue distance » les plus connus, est la rééline (Rice and Curran, 2001). Cette glycoprotéine est impliquée dans la mise en place de l'organisation du cortex cérébral. En se fixant à ses récepteurs, elle induit un transducteur de signal particulier, appelé Dab1 (*DAB1, Reelin Adaptor Protein*).

De nombreuses **kinases** régulent également la migration neuronale, en phosphorylant des acteurs clefs de la migration. Pour ne donner qu'un exemple, il a été montré que des souris ne possédant plus la kinase *Cdk5* (Cyclin-dependent kinase 5) ou un de ses régulateurs (P35 ou P39), présentaient de nombreux défauts de positionnement des neurones, dans l'hippocampe, le cervelet ou le cortex (Dhavan and Tsai, 2001). Cette kinase phosphoryle de nombreux substrats, notamment des *MAPs* (Dhavan and Tsai, 2001).

Il est intéressant de noter que des facteurs impliqués dans la machinerie de régulation de la migration, sont également régulateurs de la neurogenèse. C'est le cas, par exemple de *DCLK* (Ayala et al., 2007) *et Nde1* (Feng and Walsh, 2004). Cela suggère que le positionnement correct des neurones dans le cerveau dépend aussi de la neurogenèse.

3.3. Régulation épigénétique du cortex embryonnaire et adulte

Le développement du cerveau et son fonctionnement à l'âge adulte sont également étroitement contrôlés par des **mécanismes épigénétiques** (Bird, 2007). Ces mécanismes épigénétiques sont à l'œuvre tout au long des étapes qui régissent la construction du cerveau, son intégrité et ses fonctions. Ils sont également impliqués dans la plasticité neuronale, qui débute au dernier trimestre de la grossesse mais persiste bien au-delà de la naissance.

Des mécanismes épigénétiques modulent notamment l'expression de certains gènes, contribuant ainsi à la détermination du destin cellulaire des progéniteurs neuraux dans le cerveau en développement (Hsieh and Gage, 2004). La différenciation des cellules peut être gouvernée par des signaux extracellulaires, qui sont perçus par les cellules et relayés par transduction du signal. Cette transduction aboutit à l'activation de facteurs de transcription activateurs ou répresseurs, qui se lient au niveau de leurs gènes cibles et modulent leur expression. Parmi les mécanismes épigénétiques, les

modifications d'histones ou les changements de la méthylation de l'ADN peuvent contrôler la transduction du signal (Hsieh and Gage, 2004). Nous avons vu précédemment qu'au cours du développement, les mêmes cellules progénitrices permettaient la formation des neurones dans un premier temps, puis des astrocytes par la suite. Ces cellules progénitrices étant des cellules gliales, les gènes spécifiques à la lignée gliale sont réprimés temporairement, et orientent la différenciation cellulaire vers la formation des neurones. La répression de certains de ces gènes est gouvernée par la méthylation de l'ADN (Takizawa et al., 2001), notamment au niveau du gène GFAP (*glial fibrillary acidic protein*), marqueur de la différenciation des astrocytes, qui s'exprime logiquement dans les astrocytes, mais pas dans les neurones. L'état de méthylation du site de fixation de STAT3, facteur de transcription activateur de GFAP, rend possible ou non la fixation de ce facteur. Ainsi, dans les cellules neuroépithéliales précoces (progéniteurs formant les neurones), le site de fixation de STAT3 est méthylé, empêchant sa fixation et l'expression de GFAP. Au contraire, ce site n'est plus méthylé dans les progéniteurs à des stades plus avancées du développement, permettant les recrutement de STAT3, ce qui active l'expression de GFAP et permet la différenciation des cellules progénitrices en astrocytes (Takizawa et al., 2001).

Par ailleurs, une dynamique très particulière de méthylation dans des contextes non-CpG a été identifiée dans le cortex préfrontal humain et murin, après la naissance (Lister et al., 2013). Absente du cerveau fœtal, elle est caractérisée par une augmentation brutale autour de la naissance ou la petite enfance, et atteint son maximum à la fin de l'adolescence. Cette forme de méthylation est très proéminente dans les neurones en particulier (He and Ecker, 2015; Lister et al., 2013; Schultz et al., 2015). Chez la souris, elle coïncide avec une augmentation de l'ADN méthyl-transférase DNMT3A, responsable de la méthylation *de novo* des cytosines (Lister et al., 2013). Elle coïncide aussi, chez l'homme et la souris avec une augmentation de la densité de synapses durant l'enfance, puis avec l'élagage synaptique durant l'adolescence, ce qui souligne le rôle proéminent de la méthylation de l'ADN dans le développement et la fonction cérébrale. De façon notable, la méthylation des cytosines dans un contexte non-CpG n'est pas lue par la cellule selon les mêmes grilles d'interprétation que la méthylation CpG, avec des impacts opposés sur l'expression des gènes, dans les cellules souches et le cerveau mature (Lister et al., 2013).

La méthylation de l'ADN est impliquée dans la plasticité neuronale : dans les processus d'apprentissage et dans la mémoire à long terme (Yu et al., 2011). Il est intéressant de noter que l'activité neuronale elle-même, modifie l'épigénome, c'est-à-dire qu'elle remodèle les marques épigénétiques et la conformation de la chromatine à l'échelle du génome (Su et al., 2017; Yu et al., 2011). Ainsi, certaines marques épigénétiques présentes chez le fœtus, peuvent ne plus l'être chez

l'adolescent ou chez l'adulte. Au contraire, de nouvelles marques épigénétiques peuvent se mettre en place au cours du temps.

3.4. Perturbations des mécanismes épigénétiques et pathologies neuro-développementales

En résumé, le développement du cerveau a lieu tout au long de la grossesse et fait intervenir de nombreux évènements biologiques spécifiques, indispensables au bon fonctionnement du cerveau, chez le fœtus, mais aussi à l'âge adulte. Chaque étape nécessite l'expression d'un ensemble de gènes particuliers. L'expression de ces gènes est finement régulée d'un point de vue spatio-temporel, grâce à divers mécanismes, dont des mécanismes épigénétiques. L'importance de ce contrôle épigénétique est soulignée par le grand nombre de maladies neuro-développementales et neuropsychiatriques associées à des mutations ou variants de gènes codant des acteurs épigénétiques (*e.g.* syndrome de Rett, syndrome de Rubinstein-Taybi, autisme, schizophrénie ; Bourgeron, 2015; Gräff et al., 2011; LaSalle et al., 2013; Liu et al., 2016).

Le cerveau en développement est particulièrement vulnérable aux stress environnementaux. Ces stress prénataux sont également associés à des modifications épigénétiques, qui sont parfois délétères pour le développement et le fonctionnement du cerveau (Bale, 2015, voir Partie 3 – exposition prénatale à l'alcool). Ces stress sont susceptibles de perturber le neuro-développement et l'organisation du cortex cérébral, de manière plus ou moins prononcée et persistante. En effet, il est désormais admis qu'un certain nombre de troubles psychiatriques provient d'une combinaison de facteurs génétiques et d'agressions environnementales survenant dans la période neuro-développementale.

Il est intéressant de constater que des gènes dont l'expression est nécessaire au neurodéveloppement, sont également en jeu lors de l'activité du cerveau adulte. Il existe donc un continuum entre le développement embryonnaire et la plasticité cérébrale. Dès lors, si l'expression de ces gènes est perturbée de façon durable sous l'effet d'un stress périnatal, la plasticité cérébrale peut être affectée à l'âge adulte. Cette notion selon laquelle certaines pathologies neuropsychiatriques ont une origine neurodéveloppementale (Faa et al., 2016), a été formalisée sous le concept de **DOHaD** (*Developmental Origins of Health and Disease*). Il distingue les pathologies trouvant leur origine dans des défauts développementaux (*e.g.* défaut de migration neuronale conduisant à une structuration cérébrale altérée), de celles induites par des stress prénataux,

conduisant à une altération de la régulation de l'expression génique sans affecter l'organisation macroscopique du cerveau.

Dans le partie suivante, nous nous intéresserons à l'exposition prénatale à l'alcool (EPA), considérée comme une cause majeure d'anomalies neuro-développementales (Popova et al., 2012) ayant des conséquences fonctionnelles délétères, plus ou moins persistantes au cours du temps.

Partie 3 : L'exposition prénatale à l'alcool

Le cerveau en développement est vulnérable à divers stress environnementaux, tels que la malnutrition ou la consommation de drogues, de tabac et d'alcool (Perera and Herbstman, 2011). Parmi ces stress, **l'exposition prénatale à l'alcool** (**EPA**) est l'une des principales causes non génétiques de désordres neuro-développementaux (Jones and Smith, 1973; Popova et al., 2012; Ehrhart et al., 2019), engendrant un éventail de défauts, regroupé sous le terme de troubles causés par l'alcoolisation fœtale (**TCAF**), à l'origine de troubles neuropsychologiques et de troubles comportementaux (Mandal et al., 2015). La forme la plus sévère des TCAF est appelée syndrome d'alcoolisation fœtale (**SAF**). Les dommages neuro-développementaux causés par l'exposition *in utero* à l'alcool, s'accompagnent parfois, mais de façon non systématique, de dysmorphies cranio-faciales et d'un retard de croissance. Cette absence de traits faciaux caractéristiques chez certains nouveau-nés rend le diagnostic des sujets atteints de TCAF compliqués, alors qu'ils peuvent souffrir de défauts du SNC, aussi marqués que ceux des sujets atteints du SAF (Clarke and Gibbard, 2003). Ainsi, ces individus de prime abord «parfaitement sains » souffrent d'un **handicap masqué** qui engendre des difficultés d'apprentissage et un comportement parfois inattendu et incompris (Zieff and Schwartz-Bloom, 2008). Un **diagnostic précoce** permettrait une meilleure prise en charge de ces individus (Burd et al., 2003; Lussier et al., 2018; Ehrhart et al., 2019). Cependant, comme les mécanismes à l'origine de ses défauts ne sont pas précisément identifiés, les outils diagnostics sont rares. Mieux comprendre ces mécanismes est donc nécessaire, pour définir, à terme, des biomarqueurs moléculaires pouvant servir d'outils diagnostiques fiables des troubles du spectre de l'alcoolisation fœtale.

1. Historique

Même si la toxicité de l'alcool sur le fœtus est connue depuis bien longtemps, les premières descriptions de troubles neuro-développementaux et physiques causés par l'exposition prénatale à l'alcool ont été réalisées en **1968** par le pédiatre nantais **Paul Lemoine et ses collaborateurs**, grâce à l'étude d'une cohorte de 127 enfants exposés à ce stress (Lemoine et al., 1968). Cette étude a été complétée quelques années après par une étude pionnière américaine en **1973**, dans laquelle le terme de **syndrome d'alcoolisation fœtale** (**SAF**) a été utilisé pour la première fois pour décrire les troubles partagés par les enfants exposés *in utero* à l'alcool (**Jones and Smith**, 1973). Depuis, plusieurs recherches ont menées visant à explorer les mécanismes fondamentaux à l'origine des défauts causés par l'exposition *in utero* à l'alcool, et à estimer leurs conséquences à long terme (Clarke and Gibbard, 2003).

2. Epidémiologie

La prévalence du SAF et TCAF varie considérablement selon les sources (Burd et al., 2003). En effet, l'estimation de la prévalence du SAF varie de quelque cas pour 10 000 naissances vivantes, à 19% des naissances[1] (Burd et al., 2003). Le SAF, toucherait entre 0.5 et 3 pour 1000 naissances vivantes en France, suivant une disparité régionale (Houchi et al., 2015). Le plus couramment, on estime que le **SAF affecterait 1 à 1,5 ‰ naissances vivantes** (Burd et al., 2003), ce qui rejoint les données épidémiologiques officielles de la Haute Autorité de Santé, qui indique que l'incidence du SAF en France serait de l'ordre de 1,3 ‰ naissances vivantes (Haute Autorité de Santé, 2013). Dans son spectre large (**TCAF**), L'EPA affecterait 6 à 8 fois plus de naissances dans les pays développés (soit **≈ 9 ‰ des naissances**, Burd et al., 2003).

Aux Etats Unis, 6% des mortalités sont spécifiquement dues aux TCAF (Burd et al., 2003), ce qui reflète l'importance de continuer et de renforcer les campagnes de prévention et de sensibilisation à ce sujet, tel que celle mis en place par le Ministère de la santé en 2016, auprès des personnels médicaux (Ministère de la santé, 2016). Malgré ces campagnes, il a été rapporté que 20% des femmes aux Etats Unis, continuent de consommer de l'alcool en sachant qu'elles sont enceintes (Zieff and Schwartz-Bloom, 2008).

Ainsi, bien que ce stress prénatal soit totalement évitable, de nombreux nouveau-nés sont atteints de troubles en lien avec une exposition *in utero* à l'alcool. Comme les campagnes de prévention ne semblent que partiellement suivies ou encore trop peu relayées, l'étude des mécanismes moléculaires à l'origine de ces troubles apparait comme un enjeu majeur de santé publique, dans le but de proposer des solutions adaptées à la prise en charge des individus atteints de TCAF.

3. Description des TCAF

L'EPA est une cause majeure d'anomalies neuro-développementales (Jones and Smith, 1973; Popova et al., 2012; Ehrhart et al., 2019). Outre provoquer des avortements spontanés et des morts subites inexpliquées du nourrisson, l'EPA engendre un large spectre de déficits dont la sévérité est variable. Ce spectre s'étend de défauts structurels et macroscopiques de diverses régions cérébrales à des incapacités neurocomportementales plus subtiles et est répertorié sous le terme **TCAF** (ou TSAF pour troubles du spectre de l'alcoolisation fœtale - *FASD* pour *fetal alcohol spectrum disorders* ; Streissguth and O'Malley, 2000).

[1] Cette grande différence de proportion peut dépendre en partie du fait qu'un dépistage systématique de la population a lieu dans certains cas, mais pas dans d'autres.

3.1. Défauts physiques associés au TCAF

Certains enfants exposés *in utero* à l'alcool présentent des défauts phénotypiques visibles particuliers. Les enfants atteints de **SAF** présentent notamment un **retard de croissance** et des **malformations cranio-faciales caractéristiques** (Figure 1.7).

Des anomalies musculo-squelettiques (Myrie and Pinder, 2018), ainsi que des malformations congénitales, à l'origine de cardiopathies (Burd et al., 2007) sont également observées chez des individus ayant subi une EPA.

Les défauts phénotypiques visibles observées lors d'un SAF s'atténuent avec l'âge et disparaissent même chez l'adulte, à l'exception notable de la microcéphalie, de la lèvre supérieure fine et de la petite stature des individus (revue dans Moore and Riley, 2015).

En plus de ces anomalie, les individus exposés *in utero* à l'alcool présentent souvent des défauts du contrôle moteur, des troubles du langage et une réduction de la vision causée par une malformation de la rétine (Burd et al., 2003).

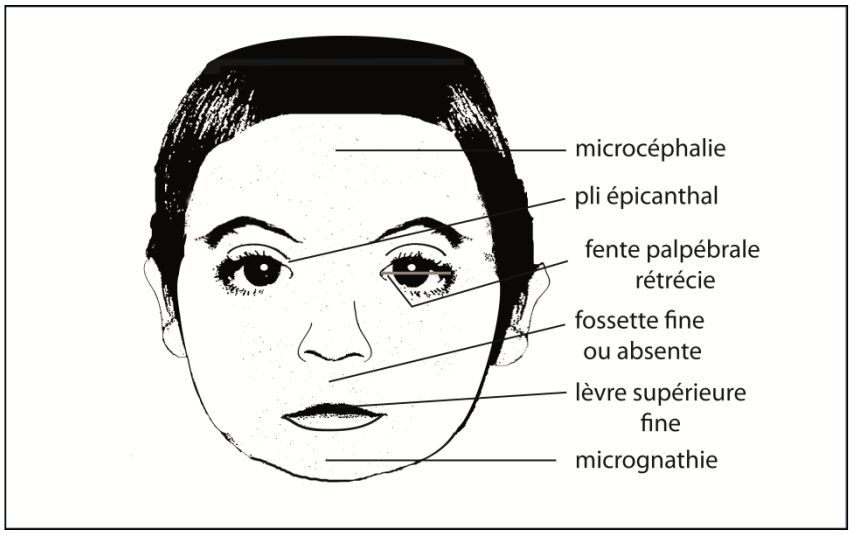


Figure 1.7: Phénotype facial caractéristique des enfants atteints du syndrome d'alcoolisation fœtale.

3.2. Déficits cognitifs et comportementaux associés au TCAF

L'EPA affecte particulièrement le cerveau de l'enfant à naître. De ce fait, un large spectre de troubles neuro-comportementaux a été décrit et peut être observé chez des individus exposés *in utero* à l'alcool (Table 1.1). Certains de ces déficits sont proches de traits observés dans le cadre de l'autisme (Mattson et al., 2011a; Varadinova and Boyadjieva, 2015). Dans le cas où des anomalies structurales du cerveau sont décelées en imagerie cérébrale, on parle de Syndrome d'alcoolisation fœtale (SAF).

Table 1.1 : Principaux troubles cognitifs et comportementaux associés à l'EPA

	A	L	A	R	M
Troubles de la concentration et de l'attention	■		■		
Troubles de la mémoire à court-terme, mais pas à long terme					■
Difficultés à traiter l'information :					
• comprendre la notion de cause à effet	■			■	
• considérer les conséquences d'une action	■			■	
• apprendre du passé, adapter son comportement aux situations	■				
• résoudre un problème	■	■	■	■	■
frustration, mensonges, familiarité auprès de personnes étrangères					
Difficultés à comprendre les concepts abstraits :					
• généraliser une information	■			■	
• comprendre les similarités et les différences	■			■	■
• planifier des actions car difficulté à comprendre le concept de temps	■			■	
• comprendre le concept de monnaie (50 centimes < 1euro, or 50 > 1).	■	■	■	■	■
• difficultés en mathématiques (calculs, nombres et symboles abstraits)	■	■	■	■	■
obstination, entêtement, frustration, faible estime de soi, retard, vols et mensonges, isolement social, personne risquant d'être escroquée par des individus mal attentionnés.					
Troubles de l'apprentissage :					
• **Lecture** : bon lecteur mais difficulté de compréhension (métaphores, humour/second degré…) et d'interprétations (identification du sujet abordé, conclusions…)	■	■		■	■
• **Ecriture** : difficultés à organiser les idées, à comprendre la ponctuation (abstrait)	■	■		■	■
• **Mathématique** (cf ci-dessus)	■	■	■	■	■
obstination, frustration, faible estime de soi *difficultés à communiquer menant à l'isolement social*					
Troubles des fonctions exécutives (*i.e.* ensemble de processus impliqués dans la capacité à planifier et à adapter son comportement, pour atteindre un objectif, de façon efficace) :					
• Difficulté à passer d'une action à une autre	■		■	■	
• Difficultés à organiser, planifier des actions			■	■	■
• Troubles de la mémoire					■
• Manque de discernement	■			■	
• Temps de réaction important				■	
anxiété intermittente, appréhension des évènements à venir, impulsivité					
Déficit intellectuel : quotient intellectuel (QI) de retard mental à QI normal	■	■	■	■	■

Les comportements associés aux troubles cognitifs causés par l'EPA sont indiqués en italique. Le lien avec l'outil diagnostic mnémotechnique « **ALARM** » est également précisé (Clarke and Gibbard, 2003). Dans l'ordre : **A :** *Adaptive functioning* ; **L :** *Language/Learning* ; A : **Attention ; R :** *Reasoning* ; **M** : Memory.

Les **dysfonctionnements primaires** du cerveau affectant plus particulièrement le comportement et la cognition sont référencés sous l'acronyme anglais « **ALARM** » (Clarke and Gibbard, 2003) pour :

- ***A**daptive functioning* (aptitudes sociales, maturité émotionnelle, compréhension),
- ***L**anguage/**L**earning* (langage, apprentissage),
- ***A**ttention* (impulsivité, hyperactivité et difficultés d'attention, taux élevé de troubles de l'attention/hyperactivité – TDAH)
- ***R**easoning* (raisonnement et fonctions exécutives),
- ***M**emory* (troubles de la mémoire verbale et non verbale)

Certains individus présentent également des capacités intellectuelles diminuées, mais ce n'est pas un déficit systématiquement constaté chez les sujets atteints de TCAF (Zieff and Schwartz-Bloom, 2008).

3.3. Le non-diagnostic du TCAF engendre des désordres secondaires plus ou moins graves

Les troubles de l'apprentissage ou de la mémorisation se manifestent généralement pendant la petite enfance ou lors de la scolarisation. Bien que certains enfants atteints de TCAF ne souffrent que de légers problèmes d'apprentissage perturbant peu leur scolarité, d'autres au contraire, requièrent un encadrement éducatif spécifique pour faire face à leurs difficultés et éviter le retard ou l'échec scolaire. Des stratégies pédagogiques simples et efficaces peuvent être mise en place pour maximiser la réussite de ces enfants (*e.g.* salle de classe bien configurée/structurée pour faciliter la concentration, organisation quotidienne routinière pour aider l'enfant à anticiper les évènements, présentations courtes au moyen de supports variés, *etc*… ; Zieff and Schwartz-Bloom, 2008).

Très souvent, lorsque des enfants atteints de TCAF ne sont pas pris en charge précocement et guidés vers un programme éducatif adapté, des **désordres secondaires** qui, reliés à des difficultés d'insertion sociale, sont susceptibles de se mettre en place à l'adolescence ou à l'âge adulte, rendant ces individus plus vulnérables. Parmi ces désordres, se trouvent :

- des **maladies neuropsychiatriques** notamment des troubles de l'anxiété, troubles de la personnalité, dépression majeure et une vulnérabilité aux addictions (O'Connor and Paley, 2009).
- un manque de discernement, engendrant des **comportements à risque ou inappropriés**, et parfois des démêlés avec la justice (vols, agressions sexuelles, agressivité ; (Burd et al., 2003; Clarke and Gibbard, 2003).

En résumé, l'EPA cause un large spectre de désordres neuro-développementaux, accompagnés ou non d'anomalies phénotypiques visibles. Les individus exposés *in utero* à l'alcool ne présentent donc pas tous le même profil, rendant délicat le diagnostic, notamment celui des sujets peu atteints physiquement. Pourtant un diagnostic et une prise en charge précoce des TCAF est indispensable pour accompagner les individus exposés *in utero* à l'alcool, et limiter le risque d'émergence de désordres secondaires (Ehrhart et al., 2019).

4. Facteurs pouvant moduler l'effet de l'alcool

Les enfants exposés *in utero* à l'alcool ne sont pas tous affectés de la même manière par ce stress. Certains ne présenteront jamais de symptôme de ce stress alors que d'autres seront sévèrement touchés (Charness et al., 2016; Zieff and Schwartz-Bloom, 2008). Il est estimé que parmi les femmes alcooliques enceintes exposant l'enfant à naitre à de grande quantité d'alcool pendant leur grossesse, seulement 5% d'entre auront un enfant avec le SAF complet (Burd et al., 2003). En revanche, la probabilité, qu'une mère donne naissance à un enfant atteint de SAF est de 75% lorsque son premier enfant l'est également (Burd et al., 2003). Ainsi, il n'est pas possible d'expliquer la sévérité des défauts causés par l'alcool uniquement par la quantité d'alcool auquel a été exposé le fœtus durant son développement. La gravité de l'effet de l'alcool sur développement du cerveau et son intégrité est modulée par de nombreux facteurs (Pollard, 2007). Ces facteurs multiples expliquent les défauts neuro-développementaux polymorphes associés à l'EPA. Dans cette section, quelques facteurs seront présentés à titre d'exemple.

4.1. Période d'exposition

Comme le développement du cerveau se déroule tout au long de la grossesse, toute EPA peut altérer le développement du cerveau ou ses fonctions à l'âge adulte (Kleiber et al., 2013; Thompson et al., 2009, voir Introduction - Partie 2 - le cerveau en développement, Figure 1.5 et Figure 1.6). De ce fait, les effets de l'alcool sur le neuro-développement dépendent particulièrement du moment d'exposition.

Bien que l'ensemble du cerveau soit exposé à l'alcool, certaines régions peuvent être plus affectées que d'autres (Roebuck et al., 1998). Cela dépend notamment des processus neuro-développementaux qui se déroulent au moment de l'exposition à l'alcool, qui ne sont pas synchronisés dans l'ensemble des régions cérébrales (Guerri et al., 2009; Zieff and Schwartz-Bloom, 2008).

Lors du **premier trimestre** de grossesse se déroule l'organogenèse, dont la formation du cerveau. Une EPA pendant ce trimestre, où la femme ignore parfois qu'elle est enceinte, peut engendrer des

défauts neuro-développementaux importants. En effet, la dysmorphie faciale, caractérisant le SAF, semble n'apparaître que lorsque la consommation d'alcool a lieu pendant la gastrulation - qui débute vers E6.5 chez la souris ; Ott, 1996 - Sulik, 2005). De plus, une exposition à l'alcool pendant la gastrulation engendre une diminution de la quantité de progéniteurs neuraux (Rubert et al., 2006), impactant à long terme le cortex préfrontal (Ashwell and Zhang, 1996).

Lors du **second trimestre** de grossesse, L'alcool peut affecter la migration neuronale dans le cortex cérébral en développement, et ainsi altérés l'organisation des couches corticales organisées aux fonctions spécialisées (Figure 1.6, El Fatimy et al., 2014). Le fonctionnement du cortex peut donc être perturbé par l'EPA.

Lors des **deux derniers trimestres** de grossesse, les neurones développent leurs ramifications cytoplasmiques (dendrites et axones) et établissent des connexions inter-neuronales (Zieff et Schwartz-Bloom, 2008). Une consommation d'alcool à ce moment du développement peut causer des désordres neurologiques sévères, engendrant des troubles du comportement et des difficultés d'apprentissage (Zieff et Schwartz-Bloom, 2008).

Ainsi, en fonction du trimestre de grossesse, la consommation d'alcool peut avoir des effets variés, corrélés aux étapes de développement. Si la consommation d'alcool est épisodique (de type *binge drinking*), les effets néfastes vont être plus sélectifs - comparés à une consommation chronique - et dépendants des pics de croissance existants lors du développement.

4.2. Les modes d'alcoolisation : consommation chronique *versus* binge drinking

Le mode d'alcoolisation pendant la grossesse est un facteur important pour estimer si l'enfant à naître va souffrir de troubles neurologiques, associés ou non à des défauts physiques (Zieff and Schwartz-Bloom, 2008). Deux modes d'alcoolisation principaux existent : on distingue l'alcoolisation chronique, où la mère consomme de l'alcool tout au long de la grossesse, de l'alcoolisation aiguë (aussi appelé *binge drinking* dans les pays anglo-saxons), définie par une consommation élevée d'alcool en peu de temps.

Le plus souvent, une consommation d'alcool chronique durant la grossesse engendrera des défauts variés, à la fois neurologiques et physiques. Une consommation aiguë, quant à elle, engendrera des défauts neurologiques et physiques, si la prise d'alcool à lieu au début de la grossesse, ou uniquement neurologiques, si la prise d'alcool a lieu en fin de grossesse (Sulik, 2005). Il semblerait qu'une forte consommation d'alcool, sur un temps court, serait un facteur de risque aggravant (Burd et al., 2003).

Toutefois, des doses faibles ou modérés d'alcool peuvent tout autant engendrer des altérations neurologiques du fœtus (Houchi et al., 2015) et les individus atteints par cette exposition, présentent

des défauts à vie, étant donné qu'aucun traitement n'existe pour soigner les altérations causées par l'alcool.

4.3. Autres facteurs

D'autres facteurs (listes non exhaustives) peuvent également moduler les effets de l'EPA sur le développement du cerveau du fœtus, mais aussi sur les risques de développer des désordres secondaires :

- **consommation d'autres substances psycho actives**, tels que le tabac, les drogues (Burd et al., 2003) ou **exposition à d'autres stress environnementaux**
- **environnement socio-économiques** défavorable (Burd et al., 2003)
- **malnutrition** (Burd et al., 2003)
- **facteur génétique** : au moins chez la souris, des différences de sensibilité à l'alcool ont pu être observés selon le fond génétique des souris étudiés (Xu et al., 2019)
- **âge élevé de la mère** (Burd et al., 2003)
- **sexe de l'enfant à naitre** : les garçons et les filles ayant subi une EPA, présentent des phénotypes différents, au niveau de troubles cognitifs, comportementaux, de la susceptibilité au stress, ou encore de désordres mentaux (Bale, 2015; Hellemans et al., 2008; Oldehinkel and Bouma, 2011).

Ainsi, de nombreux facteurs peuvent moduler l'effet de l'alcool sur le développement du cerveau de l'enfant à naitre, rendant impossible la prédictibilité individuelle des atteintes allant être causées par l'EPA. Comme il est difficile d'estimer une dose d'alcool maximale qu'il serait possible de consommer durant la grossesse, sans présenter de risque pour le développement du fœtus (Charness et al., 2016), le slogan «Par précaution, zéro alcool pendant la grossesse » reste valable encore aujourd'hui.

5. Mécanismes épigénétiques à l'origine des défauts neuro-développementaux causés par l'EPA

Bien que les défauts provoqués par une exposition *in utero* à l'alcool soient bien documentés, les mécanismes moléculaires sous-jacents, à l'origine de ces troubles ne sont pas encore bien élucidés. Toutefois, de nombreuses études suggèrent que des mécanismes épigénétiques pourraient être impliqués dans la mise en place des défauts causés par l'EPA (Lussier et al., 2017). Parmi ces mécanismes épigénétiques altérés par l'alcool, des ARNs non codants, des modifications post

traductionnelles particulières des histones, ou des défauts de méthylation de l'ADN sont répertoriés. Dans le cadre de cette thèse, nous nous intéresserons principalement aux études portant sur les profils aberrants de la méthylation de l'ADN causés par l'EPA. Ces profils aberrants suggèrent que les DNMTs sont redistribuées sur d'autres régions que leurs sites physiologiques, mais les modes de recrutement ne sont pas encore identifiés.

5.1. EPA, TCAF et perturbations épigénétiques : modèles et protocoles variés

Bien que des cohortes d'enfants atteints de TCAF ou des échantillons fœtaux également impactés par l'EPA ont été étudiés et sont informatives (Frey et al., 2018; Laufer et al., 2015; Lussier et al., 2018), elles sont constituées de données hétérogènes, qui sont délicates à interpréter au vu des multiples facteurs pouvant influencer les effets de l'EPA sur les différents individus (cf section ci-dessus, période et temps d'exposition à l'alcool, génétique, addiction de la mère à d'autres substances psychotropes, dépression chez la mère, problèmes économiques et éducationnels, etc…).

Ainsi, l'étude des modèles cellulaires *in vitro*, mais surtout de modèles animaux (principalement rongeurs, poulets et primates), a été et reste déterminante dans la compréhension des effets de l'EPA sur l'épigénome (Lussier et al., 2017), puisque, contrairement aux contextes cliniques, les sources de variabilité peuvent être plus facilement limitées ou contrôlées pour une analyse donnée.

Toutefois, la comparaison des résultats des différentes études reste difficile, étant donné d'une part, la diversité des modèles et d'autre part, la multitude de protocoles d'EPA testés :

- Type d'alcoolisation (chronique, aiguë, intermittente)
- Type d'administration (voie orale (nourriture liquide, ou semi-liquide), gavage, injection intra-péritonéale…)
- Doses d'alcool
- Durée du traitement alcoolique
- Moment de l'EPA par rapport au stade de développement du cerveau
- Temps écoulés entre l'EPA et le moment où les échantillons sont collectés et analysés (études des effets à court, moyen ou long-terme)
- Pour des modèles, comme la souris : fond génétique.
- Echantillons analysés : cerveau entier, cortex préfrontal, etc…
- Méthodes d'analyses : par exemple, pour l'étude du méthylome, les méthodes sont variées (Schang et al., 2017) et permettre d'observer des informations distinctes.

De nombreuses études s'intéressent au cerveau entier, ce qui introduit un biais. En effet, les différentes régions du cerveau ne présentent pas les mêmes proportions de types cellulaires. De plus, ces diverses régions n'ont pas non plus le même profil de méthylation, étant donné que ce profil dépend du type cellulaire (Lussier et al., 2018).

Au vu de l'ensemble de ces observations, il n'est donc pas possible de définir une liste des régions du génome différentiellement marquées épigénétiquement par l'alcool. Néanmoins, ces études permettent de dégager des tendances, permettant d'orienter les recherches vers de potentiels futurs biomarqueurs.

5.2. Altération générale (globale, non spécifique) de la méthylation de l'ADN par le stress alcoolique

Des études reposant sur divers modèles, ont permis une évaluation générale (non ciblée et non spécifique) du taux de méthylation de l'ADN ou de l'activité des DNMTs suite à une EPA ou à un stress alcoolique (*e.g.* Chen et al., 2013; Mukhopadhyay et al., 2013; Otero et al., 2012; Perkins et al., 2013, revue dans Lussier et al., 2017).

Ces études montrent que ce stress altère globalement le méthylome et/ou l'expression des DNMTs. Les premières observations en ce sens ont été obtenues par l'analyse de fœtus murins entiers de 11 jours (E11, *embryonic day* 11) ayant été exposés à l'alcool par gavage, pendant les jours embryonnaires E9 à E11. Cette étude a montré que cette exposition accrue à l'alcool engendrait une diminution globale de la méthylation de l'ADN, probablement due à une diminution de l'activité de DNMT1 (Garro et al., 1991). A la suite de cette étude, d'autres ont montré que le profil de méthylation de l'ADN est globalement perturbé par le stress alcoolique, dans l'hippocampe, le cortex préfrontal, ou des cellules souches neurales. Ces perturbations sont toutefois variables d'un système à l'autre - réduction globale de la méthylation de l'ADN observée par Chen et al., 2013; Nagre et al., 2015, augmentation globale mise en évidence par Liyanage et al., 2015; Otero et al., 2012; Perkins et al., 2013 - et leur interprétation reste limitée puisqu'il est difficile de relier ces observations générales à un processus biologique particulier, qui pourrait expliquer les défauts neuro-développementaux causés par l'EPA.

5.3. Etudes spécifiques (gènes candidats ou à grande échelle) des perturbations de la méthylation de l'ADN par l'EPA

Des études portant sur des gènes candidats (*e.g* Downing et al., 2011; Kaminen-Ahola et al., 2010; Liyanage et al., 2015; Maier et al., 1999) ou des expériences à grande échelle (génome entier ou « capture » du génome ; *e.g.* Khalid et al., 2014; Liu et al., 2009; Portales-Casamar et al., 2016),

ont complétées les études globales non spécifiques, présentées ci-dessus, concernant les perturbations de la méthylation de l'ADN suite au stress alcoolique.

En plus des facteurs de variabilité liés au modèle d'étude et plan expérimental, les méthodes bioinformatiques et biostatistiques utilisées pour l'identification des régions différentiellement méthylées (DMRs) après le stress alcoolique varient. Il est donc impossible d'établir de manière rigoureuse une liste pertinente des gènes ou de régions concernées dans les perturbations épigénétiques induites par l'EPA. Toutefois, des tendances se dégagent qui sont en faveur de la détection de biomarqueurs d'EPA et de leur utilisation future.

Contrairement aux analyses globales non spécifiques qui montraient une baisse ou une augmentation générale de la méthylation de l'ADN, ces nouvelles études « gènes candidats » ont montré à la fois des régions hyper ou hypométhylées suite au stress, au sein d'un même échantillon (Liu et al., 2009; Marjonen et al., 2015; Portales-Casamar et al., 2016), et les études « à grande échelle » ont mis en évidence une redistribution globale de la méthylation à l'échelle génomique (*e.g.* Laufer et al., 2015; Liu et al., 2009; Portales-Casamar et al., 2016, revues de Kleiber et al., 2014a (chez l'adulte) et de Lussier et al., 2017). Ces résultats suggèrent que les DNMTs sont redistribuées sur d'autres régions que leurs sites physiologiques par des modes de recrutement qui restent encore à éclaircir. De façon notable, cette redistribution de la méthylation de l'ADN a été observée principalement à distance temporelle de l'EPA, les études s'intéressant aux effets tardifs du stress alcoolique sur l'organisme.

Malgré les différences de modèles étudiés et de méthodologies d'analyses employées, certains éléments génomiques présentant des défauts de méthylation de l'ADN semblent identifiés dans des études distinctes. C'est le cas des gènes soumis à l'empreinte, mis en évidence dans plusieurs études (revue de Laufer et al., 2017). De même, des gènes de clusters de protocadhérines semblent présenter des défauts de l'ADN, à la fois dans un modèle murin, mais aussi dans des cellules d'épithéliales buccales d'enfants atteints de SAF (Portales-Casamar et al., 2016).

5.4. Disponibilité des métabolites nécessaires aux acteurs épigénétiques.

L'EPA perturbe la disponibilité et la synthèse de molécules clés du métabolisme cellulaire ayant un impact sur l'activité des enzymes responsables du dépôt de marques épigénétiques.

Le cycle de la méthionine, dans lequel la molécule SAM (S-adénoysl-méthionine, molécule donneuse de groupement méthyle) est synthétisée, est notamment altérée par l'EPA, ce qui peut avoir un effet sur la méthylation de l'ADN. Par exemple, l'EPA altère le transport, de la mère vers l'embryon, de folate (Hutson et al., 2012), permettant de produire la méthionine, formant la molécule SAM après réaction chimique. Un traitement avec des métabolites par complément alimentaire améliore le métabolisme des individus atteints de TCAF. La supplémentation en choline,

donneur de groupement méthyle est associée à une augmentation du niveau de méthylation de l'ADN de l'hippocampe et du cortex préfrontal de rats (âgés de 2 à 20 jours) ayant subis une EPA lors d'une période équivalent au troisième trimestre de grossesse chez l'humain (Otero et al., 2012). Cette augmentation est corrélée au rétablissement partiel des comportements altérés par l'EPA. Des résultats équivalents sont obtenus en traitant des cellules souches neurales au 5-azacytidine, agent chimique inhibiteur de la méthylation de l'ADN (Zhou et al., 2011), suggérant que les défauts de méthylation de l'ADN peuvent contribuer à reprogrammer l'identité des cellules neurales, perturbant ainsi leurs fonctions biologiques.

Par ailleurs, le stress oxydatif provoqué par l'exposition à l'éthanol engendre (**1**) une diminution de la molécule SAM et (**2**) une augmentation d'acétyl-CoA (précurseur de groupement acétyle), utilisé par les molécules HATs pour acétyler les histones (revues de Chater-Diehl et al., 2017; Kleiber et al., 2014a). L'apport en choline, précurseur de S-adénosyl-homocystéine et de SAM, peut donc atténuer les marques épigénétiques de méthylation aberrantes causés par l'EPA (Monk et al., 2012).

5.5. Autres mécanismes épigénétiques

D'autres mécanismes épigénétiques, semblent également impliqués dans la mise en place et le maintien des défauts causés par l'EPA, notamment des mécanismes associés à des microARN non codants spécifiques ou encore des modulateurs de modifications des histones, modulant l'accessibilité de la chromatine. Ces mécanismes sont détaillés dans les revues de Lussier et collaborateurs ainsi que Chater Diehl et collègues (Chater-Diehl et al., 2017; Lussier et al., 2017), et seront pas abordés plus en détail dans ce manuscrit.

6. Défauts transcriptomiques associés aux défauts neuro-développementaux causés par l'EPA

Plusieurs études ont mis en évidence des altérations de l'expression des gènes, suite à une EPA, associées ou non à des défauts de mécanismes épigénétiques (Kleiber et al., 2013, 2014; Marjonen et al., 2015). Notamment, des altérations au niveau du cerveau de rongeurs fœtaux, néonataux, ou adultes ont été décrites (Downing et al., 2011; Hard et al., 2005; Kleiber et al., 2013; Zhou et al., 2011).

Outre des défauts de mécanismes épigénétiques, il est également possible que l'EPA modifie la combinatoire de facteurs de transcription responsables de l'expression des gènes, ce qui engendrerait des défauts transcriptionnels.

Plusieurs études ont reportés des défauts transcriptomiques, à **grande distance temporelle** de l'EPA (Kleiber et al., 2013). De façon remarquable, il existe une « mémoire » de l'EPA très précise,

puisqu'elle se traduit par des modifications de l'expression génique qui affecte programmes de transcription spécifiquement actifs au moment de l'exposition (Kleiber et al., 2013) :

- les programmes de **prolifération cellulaire** chez la souris adulte sont affectés de manière proéminente lorsque l'EPA s'est produite au cours de l'équivalent du **premier trimestre** de grossesse, période pendant laquelle les progéniteurs neuraux se divisent activement (Figure 1.6).
- des atteintes de programmes gouvernant la **migration cellulaire** (et donc neuronale) sont observées chez l'adulte quand la période d'exposition à l'alcool correspond au **second trimestre**, au moment de la migration de jeunes neurones permettant de former les couches du cortex (Figure 1.6).
- Enfin, lorsque l'EPA s'effectue en période postnatale chez la souris, correspondant à la formation **d'axones et de dendrites et à la synaptogenèse** (équivalent du **3ème trimestre** de grossesse chez l'humain), on observe une dérégulation majeure de gènes impliqués dans la formation des synapses, le remaniement des réseaux neuronaux et la plasticité cellulaire chez l'adulte.

7. Conclusion

L'ensemble de ces données (méthylomiques et transcriptomiques) montrent les conséquences à long terme de l'EPA sur le profil de méthylation de l'ADN et l'expression génétique, dans le cerveau postnatal et adulte des modèles de rongeurs et des cohortes humaines. Malgré ces résultats, les mécanismes à l'origine de ces défauts transcriptomiques et épigénétiques, observées à distance du stress, sont encore peu connus. Notamment, comme peu d'études sont menées à proximité du stress - et, à notre connaissance, aucune étude des effets rapides d'une EPA sur le cortex en développement n'a encore été réalisée -, il est difficile de savoir si les défauts observés chez l'adulte ayant subi une EPA sont le résultat d'effet direct ou indirect de l'alcool (voir introduction du papier - Chapitre 4 - partie 1.2 pour plus de détails). Or, mieux comprendre l'origine de ces altérations, pourrait permettre d'identifier des biomarqueurs pertinents, en vue d'un diagnostic précoce des enfants atteints de TCAF, qui aiderait à une prise en charge plus rapide et donc, plus efficace de ces individus.

Dans la prochaine partie, nous nous intéresserons au facteur HSF2, protéine de réponse au stress, dont l'EPA, jouant un rôle important dans le neurodéveloppement. Cette protéine pourrait éventuellement être impliquée dans des mécanismes moléculaires à l'origine de défauts causés par l'EPA.

Partie 4 : HSF2, un facteur essentiel au neuro-développement et à la réponse à l'EPA

Les **protéines HSFs** (*Heat Shock Factors*) constituent une famille de facteurs de transcription particulières. Initialement découverts comme des facteurs de réponse au stress thermique (Parker and Topol, 1984; Wu, 1984), les HSFs sont en réalité impliqués dans la **réponse à divers stress cellulaires protéotoxiques** (stress thermique, stress oxydatif, hypoxie, métaux lourds, revue Gomez-Pastor et al., 2018). Suite à ce type de stress altérant les protéines, les HSFs induisent notamment la synthèse de protéines HSPs (*Heat Shock Proteins*), codant majoritairement des protéines chaperonnes nécessaires au bon repliement des protéines (Fujimoto and Nakai, 2010).

L'activité des HSFs ne se limite pas à la réponse aux stress protéotoxiques. En effet, ces facteurs de transcription sont impliqués dans **divers processus physiologiques**, notamment dans le **développement**, où leur contribution est parfois majeure (Abane and Mezger, 2010; Barna et al., 2018; Joutsen and Sistonen, 2019). Ainsi, la dérégulation de ces facteurs HSFs est associée à **diverses pathologies** (maladies neurodégénératives, cancer, infections, désordres métaboliques, inflammation, revue Gomez-Pastor et al., 2018) . De plus, puisque les facteurs HSFs interagissent avec de nombreux acteurs épigénétiques, ils participent à divers mécanismes épigénétiques, notamment au remodelage du paysage chromatinien (Miozzo et al., 2015; de Thonel et al., 2018).

Dans le cadre de cette thèse, nous nous sommes particulièrement intéressés au facteur HSF2, qui tient un rôle multi-facette dans le cortex cérébral en développement : nécessaire au neurodéveloppement (Chang et al., 2006; Kallio et al., 2002), il est également un acteur principal de la réponse à une EPA de type chronique (El Fatimy et al., 2014).

1. Rôle de HSF2 dans le développement du cerveau : revue Duchateau et al., en révision

Lors de ma thèse, j'ai contribué à la rédaction d'une revue concernant l'implication des facteurs HSFs (*Heat Shock Factors*) dans le développement du cerveau, en conditions physiologiques ou pathologiques. Cette revue invitée, en cours de révision, sera prochainement publiée dans la revue *Neuroscience Letters*.

En particulier, cette revue aborde en détail les points principaux suivants :

- Les facteurs de transcription HSFs constituent une famille de protéines impliquées dans la réponse à divers stress.
- Ces facteurs sont également impliqués dans le développement physiologique du cortex, et interviennent dans ces processus biologiques à plusieurs niveaux.
- Des régulations moléculaires particulières sous-tendent la complexité de leur rôle dans le neuro-développement.
- La dérégulation de ces facteurs engendre des défauts neuro-développementaux divers.

Cette revue aborde également d'autres notions importantes dans le cadre de ma thèse, qui seront aussi brièvement discutés, à la suite de la revue.

The “HSF connection”: pleiotropic regulation and activities of Heat Shock Factors shape pathophysiological brain development

Agathe DUCHATEAU[1,2,3,§], Aurélie de THONEL[1,2,§], Rachid EL FATIMY[1,2,*], Véronique DUBREUIL[1,2], Valérie MEZGER[1,2,&]

[1] Université de Paris, Epigenetics and Cell Fate, CNRS, F-75013 Paris, France
[2]Département Hospitalo-Universitaire DHU PROTECT, Paris, France.
[3]ED 562 BioSPC, Université Paris Diderot Paris 7, F-75205 Paris Cedex 13, France

[§] These authors equally contributed to the study
[&] corresponding author : valerie.meger@univ-paris-diderot.fr
* present address: Department of Neurology, Ann Romney Center for Neurologic Diseases, Brigham and Women's Hospital and Harvard Medical School, 60 Fenwood Rd, 9006, Boston, MA, 02115, USA.

Abstract

The Heat shock factors (HSFs) have been historically identified as a family of transcription factors that are activated and work in a stress-responsive manner, after exposure to a large variety of stimuli. However, they are also critical in normal conditions, in a lifelong manner, in a number of physiological processes that encompass gametogenesis, embryonic development and the integrity of adult organs and organisms. The importance of such roles is emphasized by the devastating impact of their deregulation on health, ranging from reproductive failure, neurodevelopmental disorders, cancer, and aging pathologies, including neurodegenerative disorders.

Here, we provide an overview of the delicate choreography of the regulation of HSFs during neurodevelopment, at prenatal and postnatal stages. The regulation of HSFs acts at multiple layers and steps, and comprises the control of (i)HSF mRNA and protein levels, (ii)HSF activity in terms of DNA-binding and transcription, (iii)HSF homo- and hetero-oligomerization capacities, and (iv)HSF combinatory set of post-translational modifications. We also describe how these regulatory mechanisms operate in the normal developing brain and how their perturbation impact neurodevelopment under prenatal or perinatal stress conditions. In addition, we put into perspective the possible role of HSFs in the evolution of the vertebrate brains and the importance of the HSF pathway in a large variety of neurodevelopmental disorders.

Keywords

Heat shock transcription factors, HSFs; neurodevelopment; prenatal stress; transcription; vertebrates

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

References

[1] C.S. Parker, J. Topol, A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp 70 gene, Cell. 37 (1984) 273–283.

[2] C. Wu, Activating protein factor binds in vitro to upstream control sequences in heat shock gene chromatin, Nature. 311 (1984) 81–84. doi:10.1038/311081a0.

[3] M. Åkerfelt, R.I. Morimoto, L. Sistonen, Heat shock factors: integrators of cell stress, development and lifespan, Nat. Rev. Mol. Cell Biol. 11 (2010) 545–555. doi:10.1038/nrm2938.

[4] R. Abane, V. Mezger, Roles of heat shock factors in gametogenesis and development: Role of the HSF family in development, FEBS J. 277 (2010) 4150–4172. doi:10.1111/j.1742-4658.2010.07830.x.

[5] R. Gomez-Pastor, E.T. Burchfiel, D.J. Thiele, Regulation of heat shock transcription factors and their roles in physiology and disease, Nat. Rev. Mol. Cell Biol. 19 (2018) 4–19. doi:10.1038/nrm.2017.73.

[6] F. Miozzo, D. Sabéran-Djoneidi, V. Mezger, HSFs, Stress Sensors and Sculptors of Transcription Compartments and Epigenetic Landscapes, J. Mol. Biol. 427 (2015) 3793–3816. doi:10.1016/j.jmb.2015.10.007.

[7] R. Gomez-Pastor, E.T. Burchfiel, D.W. Neef, A.M. Jaeger, E. Cabiscol, S.U. McKinstry, A. Doss, A. Aballay, D.C. Lo, S.S. Akimov, C.A. Ross, C. Eroglu, D.J. Thiele, Abnormal degradation of the neuronal stress-protective transcription factor HSF1 in Huntington's disease, Nat. Commun. 8 (2017) 14405. doi:10.1038/ncomms14405.

[8] M. Fujimoto, A. Nakai, The heat shock factor family and adaptation to proteotoxic stress: Evolution and function of the HSF family, FEBS J. 277 (2010) 4112–4125. doi:10.1111/j.1742-4658.2010.07827.x.

[9] E. Takaki, M. Fujimoto, K. Sugahara, T. Nakahari, S. Yonemura, Y. Tanaka, N. Hayashida, S. Inouye, T. Takemoto, H. Yamashita, A. Nakai, Maintenance of olfactory neurogenesis requires HSF1, a major heat shock transcription factor in mice, J. Biol. Chem. 281 (2006) 4931–4937. doi:10.1074/jbc.M506911200.

[10] M. Akerfelt, A. Vihervaara, A. Laiho, A. Conter, E.S. Christians, L. Sistonen, E. Henriksson, Heat Shock Transcription Factor 1 Localizes to Sex Chromatin during Meiotic Repression, J. Biol. Chem. 285 (2010) 34469–34476. doi:10.1074/jbc.M110.157552.

[11] T. Shinkawa, K. Tan, M. Fujimoto, N. Hayashida, K. Yamamoto, E. Takaki, R. Takii, R. Prakasam, S. Inouye, V. Mezger, others, Heat shock factor 2 is required for maintaining proteostasis against febrile-range thermal stress and polyglutamine aggregation, Mol. Biol. Cell. 22 (2011) 3571–3583. doi:10.1091/mbc.E11-04-0330.

[12] R. El Fatimy, F. Miozzo, A. Le Mouel, R. Abane, L. Schwendimann, D. Saberan-Djoneidi, A. de Thonel, I. Massaoudi, L. Paslaru, K. Hashimoto-Torii, E. Christians, P. Rakic, P. Gressens, V. Mezger, Heat shock factor 2 is a stress-responsive mediator of neuronal migration defects in models of fetal alcohol syndrome, EMBO Mol. Med. 6 (2014) 1043–1061. doi:10.15252/emmm.201303311.

[13] F. Miozzo, H. Arnoux, A. De Thonel, A.L. Schang, D. Saberan-Djoneidi, A. Baudry, B. Schneider, V. Mezger, Alcohol exposure promotes DNA methyltransferase DNMT3A up-regulation through reactive oxygen species-dependent mechanisms, Cell Stress Chaperones. (2018).

[14] S. Lecomte, F. Desmots, F. Le Masson, P. Le Goff, D. Michel, E.S. Christians, Y. Le Dréan, Roles of heat shock factor 1 and 2 in response to proteasome inhibition: consequence on p53 stability, Oncogene. 29 (2010) 4216–4224. doi:10.1038/onc.2010.171.

[15] A. Rossi, A. Riccio, M. Coccia, E. Trotta, S. La Frazia, M.G. Santoro, The proteasome inhibitor bortezomib is a potent inducer of zinc finger an1-type domain 2a gene expression: role of heat shock factor 1 (HSF1)-heat shock factor 2 (HSF2) heterocomplexes, J. Biol. Chem. 289 (2014) 12705–12715. doi:10.1074/jbc.M113.513242.

[16] J. Joutsen, A.J.D. Silva, M.A. Budzynski, J.C. Luoto, A. de Thonel, J.-P. Concordet, V. Mezger, D. Saberan-Djoneidi, E. Henriksson, L. Sistonen, HSF2 protects against proteotoxicity by maintaining cell-cell adhesion, BioRxiv. (2018) 506881. doi:10.1101/506881.

[17] M. Tanabe, A. Nakai, Y. Kawazoe, K. Nagata, Different thresholds in the responses of two heat shock transcription factors, HSF1 and HSF3, J. Biol. Chem. 272 (1997) 15389–15395.

[18] M. Tanabe, Y. Kawazoe, S. Takeda, R.I. Morimoto, K. Nagata, A. Nakai, Disruption of the HSF3 gene results in the severe reduction of heat shock gene expression and loss of thermotolerance, EMBO J. 17 (1998) 1750–1758. doi:10.1093/emboj/17.6.1750.

[19] Y. Kawazoe, M. Tanabe, N. Sasai, K. Nagata, A. Nakai, HSF3 is a major heat shock responsive factor during chicken embryonic development, Eur. J. Biochem. 265 (1999) 688–697.

[20] A. Nakai, R.I. Morimoto, Characterization of a novel chicken heat shock transcription factor, heat shock factor 3, suggests a new regulatory pathway, Mol. Cell. Biol. 13 (1993) 1983–1997.

[21] M. Fujimoto, S. Inouye, A. Nakai, Physiological roles of heat shock transcription factors, Seikagaku. 76 (2004) 419–428.

[22] J.-N. Min, Y. Zhang, D. Moskophidis, N.F. Mivechi, Unique contribution of heat shock transcription factor 4 in ocular lens development and fiber cell differentiation, Genesis. 40 (2004) 205–217. doi:10.1002/gene.20087.

[23] L. Bu, Y. Jin, Y. Shi, R. Chu, A. Ban, H. Eiberg, L. Andres, H. Jiang, G. Zheng, M. Qian, B. Cui, Y. Xia, J. Liu, L. Hu, G. Zhao, M.R. Hayden, X. Kong, Mutant DNA-binding domain of HSF4 is associated with autosomal dominant lamellar and Marner cataract, Nat. Genet. 31 (2002) 276–278. doi:10.1038/ng921.

[24] A. Vihervaara, C. Sergelius, J. Vasara, M.A. Blom, A.N. Elsing, P. Roos-Mattjus, L. Sistonen, Transcriptional response to stress in the dynamic chromatin environment of cycling and mitotic cells, Proc. Natl. Acad. Sci. 110 (2013) E3388–E3397. doi:10.1073/pnas.1305275110.

[25] M.L. Mendillo, S. Santagata, M. Koeva, G.W. Bell, R. Hu, R.M. Tamimi, E. Fraenkel, T.A. Ince, L. Whitesell, S. Lindquist, HSF1 Drives a Transcriptional Program Distinct from Heat Shock to Support Highly Malignant Human Cancers, Cell. 150 (2012) 549–562. doi:10.1016/j.cell.2012.06.031.

[26] S.K. Rabindran, R.I. Haroun, J. Clos, J. Wisniewski, C. Wu, Regulation of heat shock factor trimer formation: role of a conserved leucine zipper, Science. 259 (1993) 230–234. doi:10.1126/science.8421783.

[27] L. Pirkkala, P. Nykänen, L. Sistonen, Roles of the heat shock transcription factors in regulation of the heat shock response and beyond, FASEB J. 15 (2001) 1118–1131.

[28] C. Wu, Heat shock transcription factors: structure and regulation, Annu. Rev. Cell Dev. Biol. 11 (1995) 441–469. doi:10.1146/annurev.cb.11.110195.002301.

[29] R.I. Morimoto, Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators, Genes Dev. 12 (1998) 3788–3796. doi:10.1101/gad.12.24.3788.

[30] N.D. Trinklein, J.I. Murray, S.J. Hartman, D. Botstein, R.M. Myers, The role of heat shock transcription factor 1 in the genome-wide regulation of the mammalian heat shock response, Mol. Biol. Cell. 15 (2004) 1254–1261. doi:10.1091/mbc.E03-10-0738.

[31] A. de Thonel, A. Le Mouël, V. Mezger, Transcriptional regulation of small HSP-HSF1 and beyond, Int. J. Biochem. Cell Biol. 44 (2012) 1593–1612. doi:10.1016/j.biocel.2012.06.012.

[32] P. Ostling, J.K. Bjork, P. Roos-Mattjus, V. Mezger, L. Sistonen, Heat Shock Factor 2 (HSF2) Contributes to Inducible Expression of hsp Genes through Interplay with HSF1, J. Biol. Chem. 282 (2007) 7077–7086. doi:10.1074/jbc.M607556200.

[33] A. Sandqvist, J.K. Björk, M. Åkerfelt, Z. Chitikova, A. Grichine, C. Vourc'h, C. Jolly, T.A. Salminen, Y. Nymalm, L. Sistonen, Heterotrimerization of Heat-Shock Factors 1 and 2 Provides a Transcriptional Switch in Response to Distinct Stimuli, Mol. Biol. Cell. 20 (2009) 1340–1347. doi:10.1091/mbc.E08-08-0864.

[34] Y. Shi, P.E. Kroeger, R.I. Morimoto, The carboxyl-terminal transactivation domain of heat shock factor 1 is negatively regulated and stress responsive, Mol. Cell. Biol. 15 (1995) 4309–4318. doi:10.1128/mcb.15.8.4309.

[35] J. Zuo, D. Rungger, R. Voellmy, Multiple layers of regulation of human heat shock transcription factor 1., Mol. Cell. Biol. 15 (1995) 4319–4330.

[36] E.K. Sullivan, C.S. Weirich, J.R. Guyon, S. Sif, R.E. Kingston, Transcriptional Activation Domains of Human Heat Shock Factor 1 Recruit Human SWI/SNF, Mol. Cell. Biol. 21 (2001) 5826–5837. doi:10.1128/MCB.21.17.5826-5837.2001.

[37] J. Joutsen, L. Sistonen, Tailoring of Proteostasis Networks with Heat Shock Factors, Cold Spring Harb. Perspect. Biol. 11 (2019) a034066. doi:10.1101/cshperspect.a034066.

[38] J.O. Hensold, C.R. Hunt, S.K. Calderwood, D.E. Housman, R.E. Kingston, DNA binding of heat shock factor to the heat shock element is insufficient for transcriptional activation in murine erythroleukemia cells., Mol. Cell. Biol. 10 (1990) 1600–1608. doi:10.1128/MCB.10.4.1600.

[39] N. Hentze, L. Le Breton, J. Wiesner, G. Kempf, M.P. Mayer, Molecular mechanism of thermosensory function of human heat shock transcription factor Hsf1, ELife. 5 (2016) e11576. doi:10.7554/eLife.11576.

[40] L. Sistonen, K.D. Sarge, R.I. Morimoto, Human heat shock factors 1 and 2 are differentially activated and can synergistically induce hsp70 gene transcription, Mol. Cell. Biol. 14 (1994) 2087–2099.

[41] L.A. Sheldon, R.E. Kingston, Hydrophobic coiled-coil domains regulate the subcellular localization of human heat shock factor 2, Genes Dev. 7 (1993) 1549–1558.

[42] M. Vujanac, A. Fenaroli, V. Zimarino, Constitutive nuclear import and stress-regulated nucleocytoplasmic shuttling of mammalian heat-shock factor 1, Traffic. 6 (2005) 214–229. doi:10.1111/j.1600-0854.2005.00266.x.

[43] A. Nakai, T. Ishikawa, Cell cycle transition under stress conditions controlled by vertebrate heat shock factors, EMBO J. 20 (2001) 2885–2895. doi:10.1093/emboj/20.11.2885.

[44] R. Takii, M. Fujimoto, Y. Matsuura, F. Wu, N. Oshibe, E. Takaki, A. Katiyar, H. Akashi, T. Makino, M. Kawata, A. Nakai, HSF1 and HSF3 cooperatively regulate the heat shock response in lizards, PLOS ONE. 12 (2017) e0180776. doi:10.1371/journal.pone.0180776.

[45] T. Somasundaram, S.P. Bhat, Canonical heat shock element in the alpha B-crystallin gene shows tissue-specific and developmentally controlled interactions with heat shock factor, J. Biol. Chem. 275 (2000) 17154–17159. doi:10.1074/jbc.M000304200.

[46] J. Li, L. Chauve, G. Phelps, R.M. Brielmann, R.I. Morimoto, E2F coregulates an essential HSF developmental program that is distinct from the heat-shock response, Genes Dev. 30 (2016) 2062–2075. doi:10.1101/gad.283317.116.

[47] H. Xiao, O. Perisic, J.T. Lis, Cooperative binding of Drosophila heat shock factor to arrays of a conserved 5 bp unit, Cell. 64 (1991) 585–593.

[48] P.E. Kroeger, R.I. Morimoto, Selection of new HSF1 and HSF2 DNA-binding sites reveals difference in trimer cooperativity, Mol. Cell. Biol. 14 (1994) 7592–7603.

[49] R.S. Hilgarth, Y. Hong, O.-K. Park-Sarge, K.D. Sarge, Insights into the regulation of heat shock transcription factor 1 SUMO-1 modification, Biochem. Biophys. Res. Commun. 303 (2003) 196–200. doi:10.1016/S0006-291X(03)00312-7.

[50] T. Guettouche, F. Boellmann, W.S. Lane, R. Voellmy, Analysis of phosphorylation of human heat shock factor 1 in cells experiencing a stress, BMC Biochem. 6 (2005) 4.

[51] S. Raychaudhuri, C. Loew, R. Körner, S. Pinkert, M. Theis, M. Hayer-Hartl, F. Buchholz, F.U. Hartl, Interplay of acetyltransferase EP300 and the proteasome system in regulating heat shock transcription factor 1, Cell. 156 (2014) 975–985. doi:10.1016/j.cell.2014.01.055.

[52] N. Kourtis, R.S. Moubarak, B. Aranda-Orgilles, K. Lui, I.T. Aydin, T. Trimarchi, F. Darvishian, C. Salvaggio, J. Zhong, K. Bhatt, E.I. Chen, J.T. Celebi, C. Lazaris, A. Tsirigos, I. Osman, E. Hernando, I. Aifantis, FBXW7 modulates cellular stress response and metastatic potential through HSF1 post-translational modification, Nat. Cell Biol. 17 (2015) 322–332. doi:10.1038/ncb3121.

[53] Z. Tang, S. Dai, Y. He, R.A. Doty, L.D. Shultz, S.B. Sampson, C. Dai, MEK Guards Proteome Stability and Inhibits Tumor-Suppressive Amyloidogenesis via HSF1, Cell. 160 (2015) 729–744. doi:10.1016/j.cell.2015.01.028.

[54] N. Hashikawa, N. Yamamoto, H. Sakurai, Different Mechanisms Are Involved in the Transcriptional Activation by Yeast Heat Shock Transcription Factor through Two Different Types of Heat Shock Elements, J. Biol. Chem. 282 (2007) 10333–10340. doi:10.1074/jbc.M609708200.

[55] M.A. Budzyński, M.C. Puustinen, J. Joutsen, L. Sistonen, Uncoupling Stress-Inducible Phosphorylation of Heat Shock Factor 1 from Its Activation, Mol. Cell. Biol. 35 (2015) 2530–2540. doi:10.1128/MCB.00816-14.

[56] A. Murshid, S.-D. Chou, T. Prince, Y. Zhang, A. Bharti, S.K. Calderwood, Protein kinase A binds and activates heat shock factor 1, PloS One. 5 (2010) e13830. doi:10.1371/journal.pone.0013830.

[57] Y. Zhang, A. Murshid, T. Prince, S.K. Calderwood, Protein kinase A regulates molecular chaperone transcription and protein aggregation, PloS One. 6 (2011) e28950. doi:10.1371/journal.pone.0028950.

[58] S.D. Westerheide, J. Anckar, S.M. Stevens, L. Sistonen, R.I. Morimoto, Stress-inducible regulation of heat shock factor 1 by the deacetylase SIRT1, Science. 323 (2009) 1063–1066. doi:10.1126/science.1165946.

[59] R. Raynes, K.M. Pombier, K. Nguyen, J. Brunquell, J.E. Mendez, S.D. Westerheide, The SIRT1 modulators AROS and DBC1 regulate HSF1 activity and the heat shock response, PloS One. 8 (2013) e54364. doi:10.1371/journal.pone.0054364.

[60] E. Zelin, B.C. Freeman, Lysine deacetylases regulate the heat shock response including the age-associated impairment of HSF1, J. Mol. Biol. 427 (2015) 1644–1654. doi:10.1016/j.jmb.2015.02.010.

[61] V. Hietakangas, J. Anckar, H.A. Blomster, M. Fujimoto, J.J. Palvimo, A. Nakai, L. Sistonen, PDSM, a motif for phosphorylation-dependent SUMO modification, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 45–50. doi:10.1073/pnas.0503698102.

[62] J. Anckar, L. Sistonen, Heat shock factor 1 as a coordinator of stress and developmental pathways, in: Mol. Asp. Stress Response Chaperones Membr. Netw., Springer, 2007: pp. 78–88. http://link.springer.com/chapter/10.1007/978-0-387-39975-1_8 (accessed March 25, 2017).

[63] J. Anckar, V. Hietakangas, K. Denessiouk, D.J. Thiele, M.S. Johnson, L. Sistonen, Inhibition of DNA Binding by Differential Sumoylation of Heat Shock Factors, Mol. Cell. Biol. 26 (2006) 955–964. doi:10.1128/MCB.26.3.955-964.2006.

[64] Y. Tateishi, M. Ariyoshi, R. Igarashi, H. Hara, K. Mizuguchi, A. Seto, A. Nakai, T. Kokubo, H. Tochio, M. Shirakawa, Molecular Basis for SUMOylation-dependent Regulation of DNA Binding Activity of Heat Shock Factor 2, J. Biol. Chem. 284 (2008) 2435–2447. doi:10.1074/jbc.M806392200.

[65] A. Mathew, S.K. Mathur, R.I. Morimoto, Heat shock response and protein degradation: regulation of HSF2 by the ubiquitin-proteasome pathway, Mol. Cell. Biol. 18 (1998) 5091–5098.

[66] A.N. Elsing, C. Aspelin, J.K. Björk, H.A. Bergman, S.V. Himanen, M.J. Kallio, P. Roos-Mattjus, L. Sistonen, Expression of HSF2 decreases in mitosis to enable stress-inducible transcription and cell survival, J. Cell Biol. 206 (2014) 735–749. doi:10.1083/jcb.201402002.

[67] J.K. Ahlskog, J.K. Björk, A.N. Elsing, C. Aspelin, M. Kallio, P. Roos-Mattjus, L. Sistonen, Anaphase-promoting complex/cyclosome participates in the acute response to protein-damaging stress, Mol. Cell. Biol. 30 (2010) 5608–5620. doi:10.1128/MCB.01506-09.

[68] A. de Thonel, J.K. Ahlskog, R. Abane, G. Pires, V. Dubreuil, J. Bertelet, A.L. Aalto, S. Naceri, M. Cordonnier, C. Benasolo, M. Sanial, A. Duchateau, A. Vihervaara, M.C. Puustinen, F. Miozzo, M. Henry, D. Bouvier, J.-P. Concordet, P. Fergelot, E. Lebigot, A. Verloes, P. Gressens, D. Lacombe, J. Gobbo, C. Garrido, S.D. Westerheide, M. Petitjean, O. Taboureau, F. Rodrigues-Lima, M. Lancaster, S. Passemart, D.S. Djoneidi, L. Sistonen, V. Mezger, CBP/EP300-dependent acetylation and stabilization of HSF2 are compromised in the rare disorder, Rubinstein-Taybi syndrome, BioRxiv. (2018) 481457. doi:10.1101/481457.

[69] F.-L. Yeh, L.-Y. Hsu, B.-A. Lin, C.-F. Chen, I.-C. Li, S.-H. Tsai, T. Hsu, Cloning of zebrafish (Danio rerio) heat shock factor 2 (HSF2) and similar patterns of HSF2 and HSF1 mRNA expression in brain tissues, Biochimie. 88 (2006) 1983–1988. doi:10.1016/j.biochi.2006.07.005.

[70] M. Rallu, M.T. Loones, Y. Lallemand, R. Morimoto, M. Morange, V. Mezger, Function and regulation of heat shock factor 2 during mouse embryogenesis, Proc. Natl. Acad. Sci. 94 (1997) 2392–2397.

[71] M. Kallio, Y. Chang, M. Manuel, T.-P. Alastalo, M. Rallu, Y. Gitton, L. Pirkkala, M.-T. Loones, L. Paslaru, S. Larney, S. Hiard, M. Morange, L. Sistonen, V. Mezger, Brain abnormalities, defective meiotic chromosome synapsis and female subfertility in HSF2 null mice, EMBO J. 21 (2002) 2591–2601. doi:10.1093/emboj/21.11.2591.

[72] Y. Chang, P. Östling, M. \AAkerfelt, D. Trouillet, M. Rallu, Y. Gitton, R. El Fatimy, V. Fardeau, S. Le Crom, M. Morange, others, Role of heat-shock factor 2 in cerebral cortex formation and as a regulatorof p35 expression, Genes Dev. 20 (2006) 836–847.

[73] D. Walsh, Z. Li, Y. Wu, K. Nagata, Heat shock and the role of the HSPs during neural plate induction in early mammalian CNS and brain development, Cell. Mol. Life Sci. 53 (1997) 198–211.

[74] M. Manuel, J. Sage, M.-G. Mattéi, M. Morange, V. Mezger, Genomic structure and chromosomal localization of the mouse Hsf2 gene and promoter sequences, Gene. 232 (1999) 115–124.

[75] P. Nykänen, T.-P. Alastalo, J. Ahlskog, N. Horelli-Kuitunen, L. Pirkkala, L. Sistonen, Genomic organization and promoter analysis of the human heat shock factor 2 gene, Cell Stress Chaperones. 6 (2001) 377–385.

[76] S.-S. Lee, S.-H. Kwon, J.-S. Sung, M.-Y. Han, Y.-M. Park, Cloning and characterization of the rat Hsf2 promoter: a critical role of proximal E-box element and USF protein in Hsf2 regulation in different compartments of the brain, Biochim. Biophys. Acta. 1625 (2003) 52–63.

[77] K.D. Sarge, S.P. Murphy, R.I. Morimoto, Activation of heat shock gene transcription by heat shock factor 1 involves oligomerization, acquisition of DNA-binding activity, and nuclear localization and can occur in the absence of stress, Mol. Cell. Biol. 13 (1993) 1392–1407.

[78] Y. Kawazoe, A. Nakai, M. Tanabe, K. Nagata, Proteasome inhibition leads to the activation of all members of the heat-shock-factor family, Eur. J. Biochem. 255 (1998) 356–362.

[79] I.R. Brown, S.J. Rush, Cellular localization of the heat shock transcription factors HSF1 and HSF2 in the rat brain during postnatal development and following hyperthermia, Brain Res. 821 (1999) 333–340.

[80] A.J. Morrison, S.J. Rush, I.R. Brown, Heat shock transcription factors and the hsp70 induction response in brain and kidney of the hyperthermic rat during postnatal development, J. Neurochem. 75 (2000) 363–372.

[81] M.A. Lancaster, M. Renner, C.-A. Martin, D. Wenzel, L.S. Bicknell, M.E. Hurles, T. Homfray, J.M. Penninger, A.P. Jackson, J.A. Knoblich, Cerebral organoids model human brain development and microcephaly, Nature. 501 (2013) 373–379. doi:10.1038/nature12517.

[82] S. Velasco, A.J. Kedaigle, S.K. Simmons, A. Nash, M. Rocha, G. Quadrato, B. Paulsen, L. Nguyen, X. Adiconis, A. Regev, J.Z. Levin, P. Arlotta, Individual brain organoids reproducibly form cell diversity of the human cerebral cortex, Nature. (2019). doi:10.1038/s41586-019-1289-x.

[83] J.G. Camp, F. Badsha, M. Florio, S. Kanton, T. Gerber, M. Wilsch-Bräuninger, E. Lewitus, A. Sykes, W. Hevers, M. Lancaster, J.A. Knoblich, R. Lachmann, S. Pääbo, W.B. Huttner, B. Treutlein, Human cerebral organoids recapitulate gene expression programs of fetal neocortex development, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) 15672–15677. doi:10.1073/pnas.1520760112.

[84] K. Hashimoto-Torii, M. Torii, M. Fujimoto, A. Nakai, R. El Fatimy, V. Mezger, M.J. Ju, S. Ishii, S. Chao, K.J. Brennand, F.H. Gage, P. Rakic, Roles of Heat Shock Factor 1 in Neuronal Response to Fetal Environmental Risks and Its Relevance to Brain Disorders, Cell Neuron. 82 (2014) 560–572. doi:10.1016/j.neuron.2014.03.002.

[85] Y. Hu, N.F. Mivechi, Association and Regulation of Heat Shock Transcription Factor 4b with both Extracellular Signal-Regulated Kinase Mitogen-Activated Protein Kinase and Dual-Specificity Tyrosine Phosphatase DUSP26, Mol. Cell. Biol. 26 (2006) 3282–3294. doi:10.1128/MCB.26.8.3282-3294.2006.

[86] S. Uchida, K. Hara, A. Kobayashi, M. Fujimoto, K. Otsuki, H. Yamagata, T. Hobara, N. Abe, F. Higuchi, T. Shibata, S. Hasegawa, S. Kida, A. Nakai, Y. Watanabe, Impaired hippocampal spinogenesis and neurogenesis and altered affective behavior in mice lacking heat shock factor 1, Proc. Natl. Acad. Sci. 108 (2011) 1681–1686. doi:10.1073/pnas.1016424108.

[87] M. Tanabe, N. Sasai, K. Nagata, X.D. Liu, P.C. Liu, D.J. Thiele, A. Nakai, The mammalian HSF4 gene generates both an activator and a repressor of heat shock genes by alternative splicing, J. Biol. Chem. 274 (1999) 27845–27856.

[88] K. Hashimoto-Torii, Y.I. Kawasawa, A. Kuhn, P. Rakic, Combined transcriptome analysis of fetal human and mouse cerebral cortex exposed to alcohol, Proc. Natl. Acad. Sci. 108 (2011) 4212–4217. doi:10.1073/pnas.1100903108.

[89] G. Wang, J. Zhang, D. Moskophidis, N.F. Mivechi, Targeted disruption of the heat shock transcription factor (hsf)-2 gene results in increased embryonic lethality, neuronal defects, and reduced spermatogenesis, Genesis. 36 (2003) 48–61. doi:10.1002/gene.10200.

[90] D.R. McMillan, E. Christians, M. Forster, X. Xiao, P. Connell, J.-C. Plumier, X. Zuo, J. Richardson, S. Morgan, I.J. Benjamin, Heat Shock Transcription Factor 2 Is Not Essential for Embryonic Development, Fertility, or Adult Cognitive and Psychomotor Function in Mice, Mol. Cell. Biol. 22 (2002) 8005–8014. doi:10.1128/MCB.22.22.8005-8014.2002.

[91] R. Ayala, T. Shu, L.-H. Tsai, Trekking across the brain: the journey of neuronal migration, Cell. 128 (2007) 29–43. doi:10.1016/j.cell.2006.12.021.

[92] T. Sapir, M. Frotscher, T. Levy, E.-M. Mandelkow, O. Reiner, Tau's role in the developing brain: implications for intellectual disability, Hum. Mol. Genet. 21 (2012) 1681–1692. doi:10.1093/hmg/ddr603.

[93] X.-D. Ju, Y. Guo, N.-N. Wang, Y. Huang, M.-M. Lai, Y.-H. Zhai, Y.-G. Guo, J.-H. Zhang, R.-J. Cao, H.-L. Yu, L. Cui, Y.-T. Li, X.-Z. Wang, Y.-Q. Ding, X.-J. Zhu, Both Myosin-10 isoforms are required for radial neuronal migration in the developing cerebral cortex, Cereb. Cortex N. Y. N 1991. 24 (2014) 1259–1268. doi:10.1093/cercor/bhs407.

[94] J. Ko, S. Humbert, R.T. Bronson, S. Takahashi, A.B. Kulkarni, E. Li, L.H. Tsai, p35 and p39 are essential for cyclin-dependent kinase 5 function during neurodevelopment, J. Neurosci. Off. J. Soc. Neurosci. 21 (2001) 6758–6771.

[95] Y. Tanaka, M. Kameoka, A. Itaya, K. Ota, K. Yoshihara, Regulation of HSF1-responsive gene expression by N-terminal truncated form of p73α, Biochem. Biophys. Res. Commun. 317 (2004) 865–872. doi:10.1016/j.bbrc.2004.03.124.

[96] N.E. Faulkner, D.L. Dujardin, C.Y. Tai, K.T. Vaughan, C.B. O'Connell, Y. Wang, R.B. Vallee, A role for the lissencephaly gene LIS1 in mitosis and cytoplasmic dynein function, Nat. Cell Biol. 2 (2000) 784–791. doi:10.1038/35041020.

[97] Y. Feng, C.A. Walsh, Mitotic spindle regulation by Nde1 controls cerebral cortical size, Neuron. 44 (2004) 279–293.

[98] J. Yingling, Y.H. Youn, D. Darling, K. Toyo-Oka, T. Pramparo, S. Hirotsune, A. Wynshaw-Boris, Neuroepithelial stem cell proliferation requires LIS1 for precise spindle orientation and symmetric division, Cell. 132 (2008) 474–486. doi:10.1016/j.cell.2008.01.026.

[99] G.P. Demyanenko, M. Schachner, E. Anton, R. Schmid, G. Feng, J. Sanes, P.F. Maness, Close homolog of L1 modulates area-specific neuronal positioning and dendrite orientation in the cerebral cortex, Neuron. 44 (2004) 423–437. doi:10.1016/j.neuron.2004.10.016.

[100] Y. Matsunaga, M. Noda, H. Murakawa, K. Hayashi, A. Nagasaka, S. Inoue, T. Miyata, T. Miura, K.-I. Kubo, K. Nakajima, Reelin transiently promotes N-cadherin-dependent neuronal adhesion during mouse cortical development, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) 2048–2053. doi:10.1073/pnas.1615215114.

[101] S. Inoue, K. Hayashi, K. Fujita, K. Tagawa, H. Okazawa, K.-I. Kubo, K. Nakajima, Drebrin-like (Dbnl) Controls Neuronal Migration via Regulating N-Cadherin Expression in the Developing Cerebral Cortex, J. Neurosci. Off. J. Soc. Neurosci. 39 (2019) 678–691. doi:10.1523/JNEUROSCI.1634-18.2018.

[102] H. Jinnou, M. Sawada, K. Kawase, N. Kaneko, V. Herranz-Pérez, T. Miyamoto, T. Kawaue, T. Miyata, Y. Tabata, T. Akaike, J.M. García-Verdugo, I. Ajioka, S. Saitoh, K. Sawamoto, Radial Glial Fibers Promote Neuronal Migration and Functional Recovery after Neonatal Brain Injury, Cell Stem Cell. 22 (2018) 128-137.e9. doi:10.1016/j.stem.2017.11.005.

[103] M.J. Molumby, R.M. Anderson, D.J. Newbold, N.K. Koblesky, A.M. Garrett, D. Schreiner, J.J. Radley, J.A. Weiner, γ-Protocadherins Interact with Neuroligin-1 and Negatively Regulate Dendritic Spine Morphogenesis, Cell Rep. 18 (2017) 2702–2714. doi:10.1016/j.celrep.2017.02.060.

[104] M. Yamagata, X. Duan, J.R. Sanes, Cadherins Interact With Synaptic Organizers to Promote Synaptic Differentiation, Front. Mol. Neurosci. 11 (2018) 142. doi:10.3389/fnmol.2018.00142.

[105] S.D. Santos, M.J. Saraiva, Enlarged ventricles, astrogliosis and neurodegeneration in heat shock factor 1 null mouse brain, Neuroscience. 126 (2004) 657–663. doi:10.1016/j.neuroscience.2004.03.023.

[106] S. Homma, X. Jin, G. Wang, N. Tu, J. Min, N. Yanasak, N.F. Mivechi, Demyelination, Astrogliosis, and Accumulation of Ubiquitinated Proteins, Hallmarks of CNS Disease in hsf1-Deficient Mice, J. Neurosci. 27 (2007) 7974–7986. doi:10.1523/JNEUROSCI.0006-07.2007.

[107] E. Takaki, M. Fujimoto, T. Nakahari, S. Yonemura, Y. Miyata, N. Hayashida, K. Yamamoto, R.B. Vallee, T. Mikuriya, K. Sugahara, H. Yamashita, S. Inouye, A. Nakai, Heat Shock Transcription Factor 1 Is Required for Maintenance of Ciliary Beating in Mice, J. Biol. Chem. 282 (2007) 37285–37292. doi:10.1074/jbc.M704562200.

[108] X. Cui, J. Zhang, R. Du, L. Wang, S. Archacki, Y. Zhang, M. Yuan, T. Ke, H. Li, D. Li, C. Li, D.W.-C. Li, Z. Tang, Z. Yin, M. Liu, HSF4 is involved in DNA damage repair through regulation of Rad51, Biochim. Biophys. Acta. 1822 (2012) 1308–1315. doi:10.1016/j.bbadis.2012.05.005.

[109] X. Cui, H. Liu, J. Li, K. Guo, W. Han, Y. Dong, S. Wan, X. Wang, P. Jia, S. Li, Y. Ma, J. Zhang, H. Mu, Y. Hu, Heat shock factor 4 regulates lens epithelial cell homeostasis by working with lysosome and anti-apoptosis pathways, Int. J. Biochem. Cell Biol. 79 (2016) 118–127. doi:10.1016/j.biocel.2016.08.022.

[110] M. Fujimoto, R. Takii, E. Takaki, A. Katiyar, R. Nakato, K. Shirahige, A. Nakai, The HSF1-PARP13-PARP1 complex facilitates DNA repair and promotes mammary tumorigenesis, Nat. Commun. 8 (2017) 1638. doi:10.1038/s41467-017-01807-7.

[111] X. Shi, B. Cui, Z. Wang, L. Weng, Z. Xu, J. Ma, G. Xu, X. Kong, L. Hu, Removal of Hsf4 leads to cataract development in mice through down-regulation of γS-crystallin and Bfsp expression, BMC Mol. Biol. 10 (2009) 10. doi:10.1186/1471-2199-10-10.

[112] H. Izu, S. Inouye, M. Fujimoto, K. Shiraishi, K. Naito, A. Nakai, Heat Shock Transcription Factor 1 Is Involved in Quality-Control Mechanisms in Male Germ Cells, Biol. Reprod. 70 (2004) 18–24. doi:10.1095/biolreprod.103.020065.

[113] L. Pignataro, A.N. Miller, L. Ma, S. Midha, P. Protiva, D.G. Herrera, N.L. Harrison, Alcohol regulates gene expression in neurons via activation of heat shock factor 1, J. Neurosci. Off. J. Soc. Neurosci. 27 (2007) 12957–12966. doi:10.1523/JNEUROSCI.4142-07.2007.

[114] L. Pignataro, F.P. Varodayan, L.E. Tannenholz, P. Protiva, N.L. Harrison, Brief alcohol exposure alters transcription in astrocytes via the heat shock pathway, Brain Behav. 3 (2013) 114–133. doi:10.1002/brb3.125.

[115] S. Ishii, M. Torii, A.I. Son, M. Rajendraprasad, Y.M. Morozov, Y.I. Kawasawa, A.C. Salzberg, M. Fujimoto, K. Brennand, A. Nakai, V. Mezger, F.H. Gage, P. Rakic, K. Hashimoto-Torii, Variations in brain defects result from cellular mosaicism in the activation of heat shock signalling, Nat. Commun. 8 (2017) 15157. doi:10.1038/ncomms15157.

[116] A.L. Schang, D. Sabéran-Djoneidi, V. Mezger, The impact of genomic and epigenomics NGS approaches on our understanding of neuropsychiatric disorders, Clin. Genet. (2017).

[117] Y. Hakak, J.R. Walker, C. Li, W.H. Wong, K.L. Davis, J.D. Buxbaum, V. Haroutunian, A.A. Fienberg, Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia, Proc. Natl. Acad. Sci. 98 (2001) 4746–4751. doi:10.1073/pnas.081071198.

[118] D. Arion, T. Unger, D.A. Lewis, P. Levitt, K. Mirnics, Molecular evidence for increased expression of genes related to immune and chaperone function in the prefrontal cortex in schizophrenia, Biol. Psychiatry. 62 (2007) 711–721. doi:10.1016/j.biopsych.2006.12.021.

[119] W. Hennah, P. Thomson, A. McQuillin, N. Bass, A. Loukola, A. Anjorin, D. Blackwood, D. Curtis, I.J. Deary, S.E. Harris, E.T. Isometsä, J. Lawrence, J. Lönnqvist, W. Muir, A. Palotie, T. Partonen, T. Paunio, E. Pylkkö, M. Robinson, P. Soronen, K. Suominen, J. Suvisaari, S. Thirumalai, D. St Clair, H. Gurling, L. Peltonen, D. Porteous, DISC1 association, heterogeneity and interplay in schizophrenia and bipolar disorder, Mol. Psychiatry. 14 (2009) 865–873. doi:10.1038/mp.2008.22.

[120] P.F. Sullivan, The genetics of schizophrenia, PLoS Med. 2 (2005) e212. doi:10.1371/journal.pmed.0020212.

[121] H. Hagberg, P. Gressens, C. Mallard, Inflammation during fetal and neonatal life: implications for neurologic and neuropsychiatric disease in children and adults, Ann. Neurol. 71 (2012) 444–457. doi:10.1002/ana.22620.

[122] M. Lin, D. Zhao, A. Hrabovsky, E. Pedrosa, D. Zheng, H.M. Lachman, Heat shock alters the expression of schizophrenia and autism candidate genes in an induced pluripotent stem cell model of the human telencephalon, PloS One. 9 (2014) e94968. doi:10.1371/journal.pone.0094968.

[123] R.M. Fame, C. Dehay, H. Kennedy, J.D. Macklis, Subtype-Specific Genes that Characterize Subpopulations of Callosal Projection Neurons in Mouse Identify Molecularly Homologous Populations in Macaque Cortex, Cereb. Cortex N. Y. N 1991. 27 (2017) 1817–1830. doi:10.1093/cercor/bhw023.

[124] F. García-Moreno, Z. Molnár, Subset of early radial glial progenitors that contribute to the development of callosal neurons is absent from avian brain, Proc. Natl. Acad. Sci. U. S. A. 112 (2015) E5058-5067. doi:10.1073/pnas.1506377112.

[125] S. Aivazidis, C.M. Coughlan, A.K. Rauniyar, H. Jiang, L.A. Liggett, K.N. Maclean, J.R. Roede, The burden of trisomy 21 disrupts the proteostasis network in Down syndrome, PloS One. 12 (2017) e0176307. doi:10.1371/journal.pone.0176307.

[126] R.A. Veitia, S. Bottani, J.A. Birchler, Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects, Trends Genet. TIG. 24 (2008) 390–397. doi:10.1016/j.tig.2008.05.005.

[127] A. Duchon, Y. Herault, DYRK1A, a Dosage-Sensitive Gene Involved in Neurodevelopmental Disorders, Is a Target for Drug Development in Down Syndrome, Front. Behav. Neurosci. 10 (2016) 104. doi:10.3389/fnbeh.2016.00104.

[128] T. Gidalevitz, A. Ben-Zvi, K.H. Ho, H.R. Brignull, R.I. Morimoto, Progressive disruption of cellular protein folding in models of polyglutamine diseases, Science. 311 (2006) 1471–1474. doi:10.1126/science.1124514.

[129] T. Gidalevitz, E.A. Kikis, R.I. Morimoto, A cellular perspective on conformational disease: the role of genetic background and proteostasis networks, Curr. Opin. Struct. Biol. 20 (2010) 23–32. doi:10.1016/j.sbi.2009.11.001.

[130] S.L. Rutherford, S. Lindquist, Hsp90 as a capacitor for morphological evolution, Nature. 396 (1998) 336–342. doi:10.1038/24550.

[131] D.F. Jarosz, S. Lindquist, Hsp90 and environmental stress transform the adaptive value of natural genetic variation, Science. 330 (2010) 1820–1824. doi:10.1126/science.1195487.

[132] A.M. Jaeger, C.W. Pemble, L. Sistonen, D.J. Thiele, Structures of HSF2 reveal mechanisms for differential regulation of human heat-shock factors, Nat. Struct. Mol. Biol. 23 (2016) 147–154. doi:10.1038/nsmb.3150.

[133] T. Neudegger, J. Verghese, M. Hayer-Hartl, F.U. Hartl, A. Bracher, Structure of human heat-shock transcription factor 1 in complex with DNA, Nat. Struct. Mol. Biol. 23 (2016) 140–146. doi:10.1038/nsmb.3149.

[134] V. Mezger, O. Bensaude, M. Morange, Unusual levels of heat shock element-binding activity in embryonal carcinoma cells, Mol. Cell. Biol. 9 (1989) 3888–3896. doi:10.1128/mcb.9.9.3888.

[135] M. Fujimoto, H. Izu, K. Seki, K. Fukuda, T. Nishida, S. Yamada, K. Kato, S. Yonemura, S. Inouye, A. Nakai. HSF4 is required for normal cell growth and differentiation during mouse lens development. EMBO J. 2004 Oct 27;23(21):4297-306. Epub 2004 Oct 14. doi: 10.1038/sj.emboj.7600435

2. HSF2 et EPA : notions complémentaires à la revue

Cette revue montre l'importance du facteur HSF2 :

- dans le **neuro-développement**, et notamment son implication physiologique dans la migration neuronale au niveau du cortex cérébral (Chang et al., 2006; Kallio et al., 2002).
- dans la **réponse à l'EPA** dans le cortex cérébral embryonnaire murin :

- Suite à ce stress, HSF2 perd sa liaison avec certaines de ses cibles physiologiques, au profit d'autres régions génomiques (El Fatimy et al., 2014). Parmi les cibles physiologiques affectées par cette perte de fixation de HSF2, se trouvent les gènes *MAPs*, essentiel à la migration neuronale. Les défauts de fixation de HSF2 peuvent donc être à l'origine de troubles neuro-développementaux tels que ceux associés aux troubles causés par l'alcoolisation fœtale.

Il semble donc que la distribution de HSF2 dans le cortex en développement dépende des conditions environnementales, ce qui peut conduire à modifier le programme transcriptionnel de ce facteur. Ces observations rejoignent celles de Mendillo et collaborateurs portant sur HSF1. Ces auteurs ont montré que la distribution génomique du facteur HSF1 en réponse au stress thermique, était spécifique et différente de celle observée dans des cellules cancéreuses. Ils ont également montré que cette distribution différentielle de ce facteur met en place un programme transcriptionnel distinct dans ces deux conditions (Mendillo et al., 2012).

- En conditions physiologiques, HSF2 se lie à l'ADN sous la forme d'homotrimères, alors que HSF1 est monomérique et ne montre pas de capacité à lier l'ADN (El Fatimy et al., 2014). Lors d'une EPA aiguë, HSF1 est activé par HSF2 et forme un hétérotrimère avec HSF2, capable de fixer l'ADN (El Fatimy et al., 2014). En l'absence d'HSF2, HSF1 n'est pas capable de fixer l'ADN suite à l'EPA, ce qui conforte l'importance de HSF2 dans la réponse au stress alcoolique, par la voie HSFs. Ainsi, contrairement au stress thermique, où HSF1 joue un rôle majeur (McMillan et al., 1998), HSF2 est le facteur essentiel, participant activement à la réponse au stress alcoolique, en activant notamment HSF1 et en régulant l'expression de gènes *Hsps* impliqués dans la réponse au stress par la voie HSF, tels que ceux codant pour HSP70 et HSP90 (El Fatimy et al., 2014). Le facteur HSF1, participe également grandement à la réponse au stress alcoolique, et aux défauts neuro-développementaux pouvant être associés à ce stress, comme en atteste plusieurs études (El Fatimy et al., 2014; Hashimoto-Torii et al., 2014; Ishii et al., 2017; Pignataro et al., 2007).

Dans le cortex cérébral, la formation d'hétérotrimère HSF1-HSF2, au lieu d'homotrimères HSF2 peut engendrer des régulations transcriptionnelles différentes des gènes ciblés en conditions physiologiques et en conditions de stress. Ce type de coopération entre HSF1 et HSF2 a déjà été observée dans d'autres situations, notamment en conditions physiologiques, dans des spermatocytes

murins. Cette coopération HSF1-HSF2, est dans ce système biologique, altérée suite à un stress thermique, ce qui peut moduler l'expression de certains gènes, et rendent ces cellules sensibles à la chaleur (Korfanty et al., 2014).

En conditions physiologiques, mais aussi en réponse au stress, HSF2 fixe des cibles génomiques possédant un motif HSE (*Heat Shock Element*), séquence génomique classiquement reconnue par les facteurs HSFs (El Fatimy et al., 2014). Il est possible que les facteurs HSF1 et HSF2 fixent également d'autres séquences génomiques en réponse au stress alcoolique. Il a été montré dans des neurones corticaux en culture, que HSF1 se liait à une séquence dite « ARE » (*Alcohol Response Element*), au niveau du gène *Gabra4*, induisant alors sa surexpression en réponse à l'alcool (Pignataro et al., 2007). Ce gène, exprimé dans le cortex cérébral et codant une sous-unité d'un récepteur du neurotransmetteur inhibiteur GABA, a été identifié comme étant surexprimé dans divers contextes de stress alcoolique (Sanna et al., 2003). Dans cette étude, l'expression de HSF2 n'a pas été vérifiée. Il est donc impossible de savoir si HSF1 est activé par le stress alcoolique, en absence ou présence de ce facteur, dans ces conditions expérimentales. D'autres gènes surexprimés en réponse à l'alcool portent ce motif ARE, mais la présence des facteurs HSF1 ou HSF2 n'a pas été vérifiée (Pignataro et al., 2007).

Par ailleurs, dans une étude portant sur les défauts de la méthylation de l'ADN causés par le stress alcoolique, le motif HSE a été identifié comme associé aux îlots CpGs de gènes hypométhylés et surexprimés dans un modèle d'EPA de cellules ES humaines et de corps embryoïdes (Khalid et al., 2014). Il est donc possible que les facteurs HSF1 ou HSF2, participent à des mécanismes épigénétiques à l'origine de ces altérations de la méthylation de l'ADN en réponse au stress, pouvant avoir un impact sur l'expression des gènes. D'autant plus que les HSFs participent au remodelage du paysage chromatinien, des interactions entre HSFs et acteurs épigénétiques ayant déjà été mis en évidence (Abane, 2011; Li et al., 2018; de Thonel et al., 2018).

Ainsi, l'activation inappropriée de la voie de réponse au stress des HSFs, par une exposition à l'alcool pourrait jouer un rôle dans les anomalies du développement du cerveau, notamment ceux associés aux TCAF.

Afin de continuer à explorer le rôle des HSFs dans la réponse au stress - et en particulier voir si les HSFs peuvent interagir avec des DNMTs, responsables de la méthylation de l'ADN - l'équipe a poursuivi les recherches à partir de modèles cellulaires *in vitro* (cellules murines neuronales **1C11** et cellules murines des neuroblastome **Neuro2A**) exposés à un stress alcoolique, ou de modèles murins (**cortex cérébral** embryonnaire murin, exposé à une EPA). Ces deux modèles sont complémentaires :

le modèle murin est plus fidèle à la réalité que l'étude de cellules *in vitro* et permet de mieux estimer les effets globaux de l'EPA sur l'organisme, et plus particulièrement sur le cortex embryonnaire. Toutefois, ce tissu est hétérogène, et seule l'étude des lignées cellulaires « *neuron-like* » permettent de cibler le(s) type(s) cellulaire(s) mis en jeu dans l'analyse.

2.1. Résultats observés dans les modèles cellulaires

En utilisant des cellules *in vitro* 1C11, l'équipe a mise en évidence que :

- le taux d'**ARN et de protéines du facteur HSF2 était augmenté** suite à un stress alcoolique dans ces cellules (Miozzo 2014, travaux non publiés). En revanche, les taux de HSF1 restent inchangés suite au traitement alcoolique (Miozzo 2014, travaux non publiés).
- **HSF1 et HSF2 sont capables de lier l'ADN** dans cette lignée cellulaire, à la fois en conditions physiologiques et après un stress alcoolique (Miozzo et al. 2018). Une quantité plus importante de complexe HSF-HSE se forme après l'exposition à l'alcool mais il est difficile de déterminer quel facteur a une activité de liaison à l'ADN augmentée (Miozzo et al. 2018).
- Le **taux d'ARN et de protéines de DNMT3A et DNMT3B est augmenté** suite à un stress alcoolique. Ces augmentations seraient indépendantes des HSFs, mais plutôt corrélées au stress oxydatif provoqué par l'alcool (Miozzo et al. 2018). Même si une diminution transcriptionnelle de DNMT1 est parfois observée suite à un stress alcoolique (selon les conditions du stress), le taux protéique de DNMT1 reste inchangé suite au traitement alcoolique (Miozzo et al. 2018).

En utilisant des cellules *in vitro* Neuro2A, l'équipe a mise en évidence que :

- le taux protéique de **HSF2 est augmenté** suite à un stress alcoolique dans ces cellules (Miozzo 2014, travaux non publiés).
- **La capacité de liaison à l'ADN des facteurs HSF1 et HSF2 est augmentée** suite à un stress alcoolique (Fatimy et al. 2014).

2.2. Résultats observés dans le cortex cérébral embryonnaires murin

Suite à une EPA aiguë, l'équipe a également montré que :

- la quantité de **protéines HSF2 et DNMT3A nucléaires est augmentée au niveau de la zone proliférative** (*i.e.* zone ventriculaire) de cortex cérébral E16.5 (Figure 1.8, résultats non publiés, travaux de AL Schang).
- **HSF2 est capable d'interagir physiquement avec DNMT3A**, en conditions basales (injections PBS) et après EPA. D'après la proportion d'IP par rapport à celle de l'input, il semble y avoir plus de DNMT3A après EPA qui interagit avec HSF2 qu'en conditions PBS. (Figure 1.8, résultats non publiés de F. Miozzo).

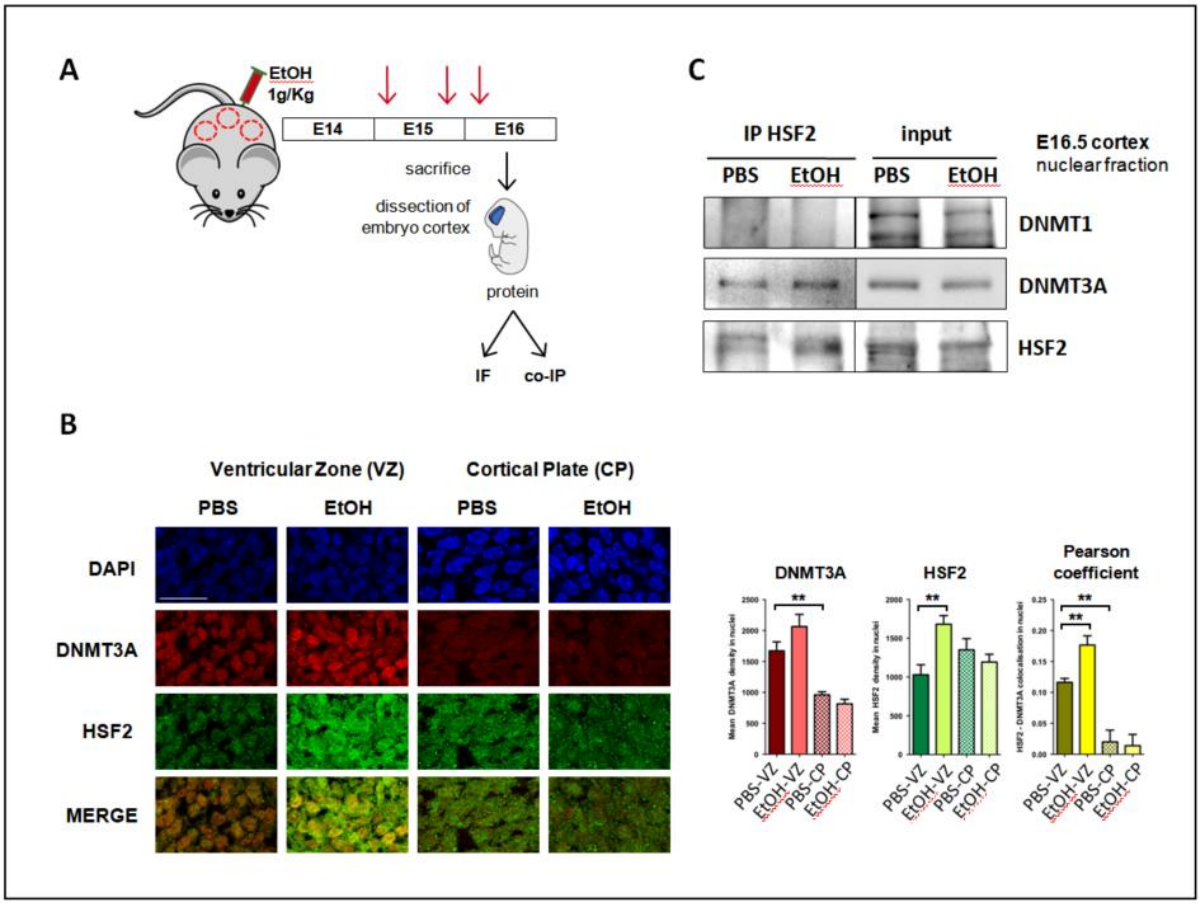


Figure 1.8 : Suite à une EPA aiguë, les protéines HSF2 et DNMT3A sont enrichies dans les noyaux des cellules de la zone ventriculaire, et interagissent physiquement.

(A) modèle d'EPA réalisé : des souris gestantes reçoivent 3 injections intra-péritonéales d'éthanol (1g/kg), aux jours embryonnaires E15, E15.5 et E16. Les embryons sont ensuite sacrifiés à E16.5, et leur cortex récupéré pour des analyses d'immuno-fluorescence (B) et de co-immunoprécipitation (C). Les échantillons contrôles sont traités similairement, avec des injections de PBS.

(B) Expérience d'immunofluorescence réalisée à partir de coupes de cortex embryonnaires murins E16.5, ayant subi un stress EPA ou non (injection PBS). La zone proliférative (zone ventriculaire, ou **VZ**), et la plaque corticale (**CP**) ont été observées. Expérience de l'équipe réalisée par A.L. Schang.

(C) Co-immunoprécipitation de la protéine HSF2 avec DNMT1 et DNMT3A, dans des fractions nucléaires de cellules du cortex cérébrales embryonnaires murins E16.5, exposés ou non à une EPA aiguë. Expérience de l'équipe réalisée par F. Miozzo.

L'ensemble de ces observations suggèrent donc que le HSF2 pourrait être impliqué dans des mécanismes épigénétiques, en conditions physiologiques, susceptibles d'être perturbés après l'EPA. Notamment, l'interaction entre HSF2 et DNMT3A - qui persiste malgré l'EPA - semble indiquer que HSF2 pourrait moduler le dépôt de la méthylation de l'ADN, au niveau de certaines cibles génomiques.

Chapitre 2 : Objectifs de la thèse

Mon projet vise, à caractériser, dans le cerveau en développement, de possibles altérations précoces de la méthylation de l'ADN causées par une exposition prénatale à l'alcool (EPA), et à estimer les conséquences transcriptionnelles et fonctionnelles de ces défauts sur les régions génomiques impliquées dans les fonctions cérébrales. Mon travail de thèse a aussi pour objectif de préciser le rôle du facteur HSF2 dans la réponse à l'EPA, associé ou non aux éventuels défauts du méthylome causés par le stress. Enfin, en estimant également l'implication de HSF2 dans la mise en place de ces défauts, je souhaite proposer un mécanisme épigénétique à l'origine de ces perturbations, dans lequel ce facteur de transcription jouerait un rôle décisif.

L'équipe a mis en évidence le rôle multi-facette de HSF2 dans le cerveau en développement :

- HSF2 régule l'expression de gènes **nécessaires au développement du cerveau et à ses fonctions** (Kallio et al., 2002; Chang et al., 2006). Il a été montré par ailleurs que HSF2 régule, dans des cellules U2OS, des gènes codant des protocadhérines (Pcdh), impliquées dans l'adhésion intercellulaire (Joutsen et al., 2018).
- **Lors d'une exposition à l'alcool, HSF2 est sur-exprimée dans la zone ventriculaire** (Miozzo, 2014, travaux non publiés), et sa **capacité à lier l'ADN est augmentée** dans des cellules N2A et dans le cortex embryonnaire murin de 15.5 jours (El Fatimy et al., 2014).
- Lors d'une EPA, HSF2 est un **médiateur de désordres neurodéveloppementaux typiques du syndrome d'alcoolisation fœtal**. **Suite à l'EPA, HSF2 est redistribué** (perte de fixation au niveau de certaines cibles physiologiques, maintien ou gain de fixation sur des régions génomiques en réponse au stress), ce qui perturbe notamment la migration neuronale, et altère ainsi la formation du cortex (El Fatimy et al., 2014).
- **HSF2 est un partenaire de DNMT3A**, une des protéines responsables de la méthylation *de novo* de l'ADN, qui est surexprimée lors d'un stress alcoolique *in vitro* réalisés dans des cellules 1C11, N2A ou MEFs (Miozzo, 2014; Miozzo et al., 2018). D'après des résultats préliminaires, ce complexe HSF2-DNMT3A, observé en conditions physiologiques, semble renforcée après EPA (Miozzo, 2014).

Ces observations permettent de proposer un **mécanisme précoce de réponse à l'EPA, détectable dès l'exposition au stress, qui expliquerait certains défauts de méthylation de l'ADN observés.** En effet, ces résultats permettent d'émettre l'hypothèse qu'une part des **perturbations du**

méthylome observées à la suite d'une EPA, serait le résultat de l'interaction HSF2-DNMT3A. Le facteur HSF2, mobilisé par l'EPA, interagit avec DNMT3A et le redistribue de façon aberrante suite au stress, selon deux *scenarii* non exclusifs (Figure 2.1) **:**

- HSF2 recrute DNMT3A sur des sites génomiques spécifiques en réponse au stress, induisant leur hyperméthylation, comparée à une situation physiologique.
- HSF2 séquestre DNMT3A hors de ses cibles physiologiques, l'empêchant de les méthyler.

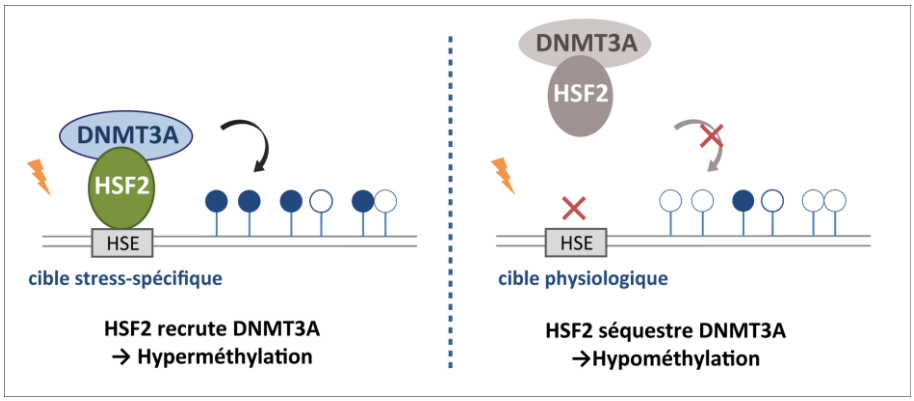


Figure 2.1 : Modélisation du rôle possible du facteur HSF2 dans la mise en place précoce de défauts de méthylation de l'ADN, en réponse à une EPA. Comme HSF2 interagit avec DNMT3A, et que ses sites de fixations sont modifiés lors d'une EPA, il est envisageable que ce facteur recrute DNMT3A sur des sites génomiques spécifiques en réponse au stress, induisant leur hyperméthylation par rapport à une situation physiologique, ou au contraire, qu'il séquestre DNMT3A hors de ses cibles physiologiques, l'empêchant de les méthyler.

Ainsi, à travers l'étude d'un modèle murin d'EPA (stress alcoolique aigu, ou *binge drinking*), et en particulier, en analysant les cortex embryonnaires de 16.5 jours, mutés ou non pour HSF2, collectés quelques heures après le stress, je souhaite répondre à **4 questions complémentaires** (Figure 2.2) :

Q1. Observe-t-on des changements précoces de la méthylation de l'ADN, dans ce modèle d'EPA? Si c'est le cas, dans quelles régions ?

Q2. Les défauts de méthylation de l'ADN, observés après EPA, sont-ils situés préférentiellement dans les régions génomiques modulées dans la période d'exposition au stress, notamment celles, essentielles au neuro-développement ?

Pour le savoir, il est nécessaire de déterminer comment évolue l'accessibilité de la chromatine et l'expression des gènes nécessaires au neurodéveloppement et aux fonctions cérébrales dans la période encadrant l'EPA.

Q3. Comment sont distribués les facteurs HSF2 et DNMT3A peu de temps après ce stress (perte/maintien/gain de fixation ? Existe-t-il des cibles communes aux deux facteurs ou bien sont-elles spécifiques à chaque facteur?).

Q4. Le complexe DNMT3A-HSF2 participe-t-il aux défauts de méthylation observés ?

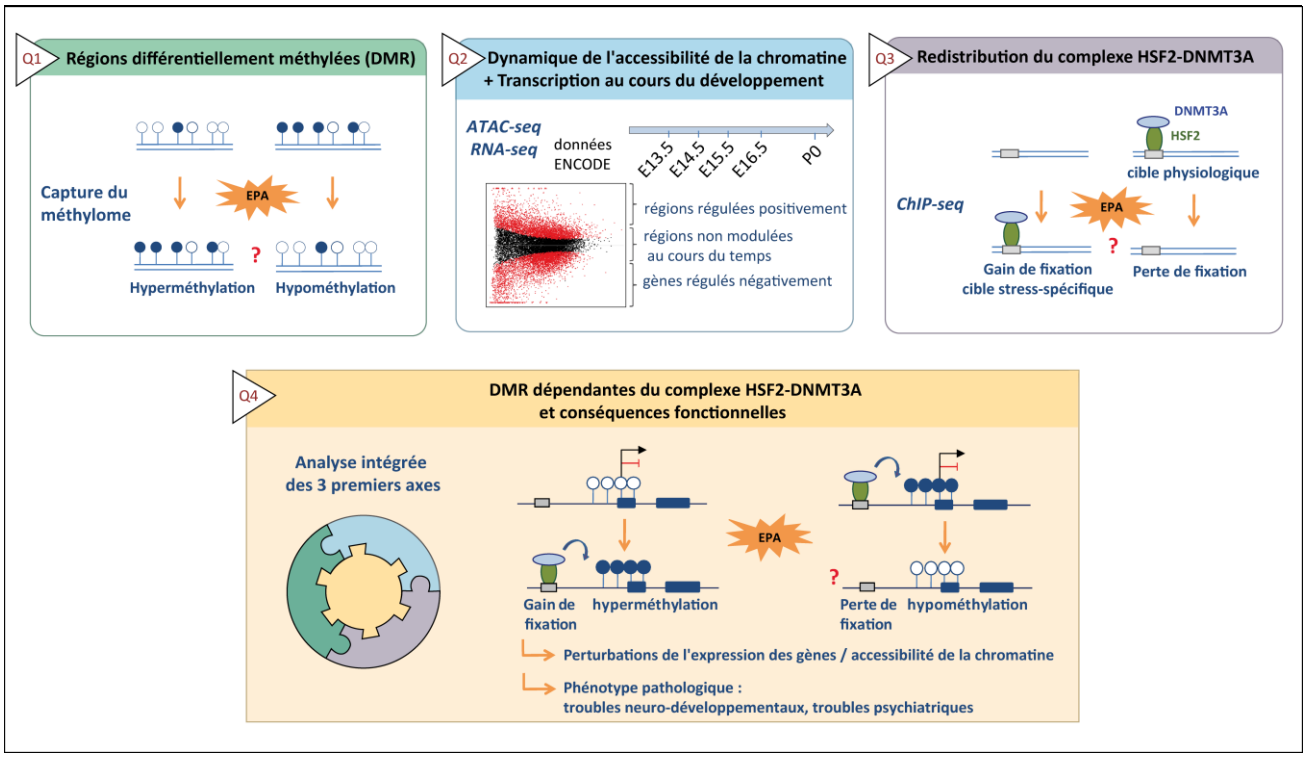


Figure 2.2 : Objectifs du projet de thèse.
Principales questions posées dans le projet de thèse et méthodologie associée.

Afin de répondre à ces différentes questions, j'ai utilisé et intégré des approches de séquençage à haut débit (NGS, Figure 2.2, voir détails dans la partie Matériel et Méthode, Chapitre 3, et dans le *Material & Methods* de l'article, voir Chapitre 4 section 4.1) :

1. Pour identifier les régions différentiellement méthylées, très rapidement après le stress, j'ai analysé une **capture du méthylome** couplé à un séquençage à haut débit (technique *d'EpiCapture, NimbleGen)*.

2. Pour suivre l'évolution physiologique (en absence de stress) de l'accessibilité de la chromatine et l'expression des gènes nécessaires au neurodéveloppement et aux fonctions cérébrales dans la période encadrant l'EPA, j'ai analysé des données d'*ATAC-seq* (*Assay for Transposase-Accessible Chromatin with high- throughput sequencing,* (Buenrostro et al., 2013) et des données transcriptomiques publiques (ENCODE). Afin d'identifier les voies biologiques pouvant être affectées par l'EPA, ces données, obtenues à partir de cortex embryonnaires murins non stressés, de 13 à 16 jours, ainsi que de cortex de nouveau-nés (P0), ont été intégrées aux régions identifiées comme différentiellement méthylées après ce stress.

3. Pour cartographier les cibles de HSF2 et DNMT3A avant et après l'exposition à l'alcool, j'ai réalisé une immunoprécipitation de la chromatine ciblant ces deux facteurs, suivie de NGS (***ChIP-seq***).

4. Pour déterminer si les régions différentiellement méthylées (DMR) sont corrélées à la redistribution de HSF2 et DNMT3A, j'ai **intégré les résultats des analyses précédentes**.

Cette étude apporte une **vision génomique globale de l'effet épigénétique précoce d'une EPA**, alors que les données de la littérature s'intéressent le plus souvent aux effets tardifs de l'alcool, observés chez l'adulte. Or, comme l'activité neuronale peut modifier l'état de méthylation de l'ADN au cours de la vie d'un individu (Su et al., 2017), étudier les perturbations de la méthylation chez l'adulte ayant subi une EPA ne permet pas de comprendre si cette désorganisation du méthylome provient d'une cause directe (une modification de la méthylation se met en place au moment du stress) ou indirecte (les altérations de l'activité neuronale dues au stress perturbent, par ricochets, le méthylome). En identifiant les sites différentiellement méthylés peu de temps après le stress, mon étude permettra - à terme - de revisiter les données de la littérature concernant le méthylome de cerveau d'adultes ayant subi une EPA.

Chapitre 3 : Matériels et Méthode

Les informations concernant le matériel et méthodes relatifs aux articles sont décrites dans le chapitre Résultats, au niveau des sections correspondantes (*cf material & methods* de l'article) et décrites brièvement ci-dessous. Le matériel et les techniques utilisés pour la réalisation et la validation des protocoles de *ChIP-seq* ciblant HSF2 et DNMT3A sont en revanche plus amplement détaillés dans cette section.

1. Modèle murin d'exposition prénatale à l'alcool

1.1. Aspects éthiques

Les expériences ont fait appel à un modèle de souris, suivant un projet déclaré auprès du Ministère de l'Enseignement Supérieur et de la Recherche sous le numéro APAFIS 2016040414515579. Ce projet est autorisé par le Comité d'éthique Buffon (CEB) sous la référence « CEB-06-2016-SABERAN-Etude du développement du cerveau et des effets épigénétiques et comportementaux associés, dans des modèles murins de stress périnatal chez l'enfant ». Ces expériences suivent les recommandations de la structure du bien-être animal de l'université Paris Diderot (SEBA-Buffon). Tous les efforts ont été fournis pour réduire la souffrance et le stress des animaux, et limiter le nombre d'animaux utilisés.

1.2. Alcoolisation prénatale

Le stade embryonnaire 0,5 jour (E0.5) est défini à midi, le jour de détection du bouchon vaginal post-accouplement. Les lignées de souris sont la lignée *C57BL/6N*. (commerciale ; Charles River) ou la lignée *Hsf2*^{tm1Mmr} (Kallio et al., 2002) sur le fond génétique *C57BL/6N*. Afin de mimer un stress alcoolique aigu[2] de type « *binge drinking* », des souris gestantes âgées de 2 à 4 mois reçoivent 3 injections intrapéritonéales d'éthanol, réalisées aux jours embryonnaires E15, E15.5 et E16 (Figure 1A du manuscrit Duchateau *et al.*, en préparation, voir résultats); 3g/kg d'éthanol par injection, dilué dans du PBS pour un volume final de 500 µL injecté). Un tel stress correspond à une dose connue pour induire des défauts neurodéveloppementaux chez les petits rongeurs (Ikonomidou et al., 2000; Olney et al., 2002a; Carloni et al., 2004). Le groupe témoin a été traité de façon similaire, avec 500µL de PBS par injection. Les cortex embryonnaires ont été prélevés 2 heures après la dernière injection,

[2] *Comme seules trois injections d'éthanol sont effectuées en 24h, nous considérons que le modèle d'EPA étudié est un stress alcoolique aigu, en comparaison à un stress alcoolique chronique réalisé sur plusieurs jours.*

à E16.5. Tous les prélèvements d'embryons ont été réalisés sur la glace, dans du milieu L15 froid (Leibovitz Gibco #11415-049). La queue de chaque embryon a également été collectée pour déterminer leur sexe et leur génotype (embryons *Hsf2*$^{+/+}$, issus de croisement entre souris hétérozygotes *Hsf2*$^{+/tm1Mmr}$ ou de croisements entre souris C57Bl/6N d'origine commerciale (appelés « WT » dans la suite du document), embryons *Hsf2*$^{+/-}$, appelés « Hsf2HET » et embryons *Hsf2*$^{-/-}$, appelés « *Hsf2*KO »). Parce que des différences de comportement entre mâles et femelles ont été observées dans la littérature, suite à une alcoolisation fœtale (Hellemans et al., 2010a), nous avons analysé séparément les cortex des embryons mâles et femelles. Nous avons mené l'ensemble de nos expériences à partir d'échantillons d'embryons mâles, afin de pouvoir ultérieurement comparer nos résultats à ceux de la littérature (en ce qui concerne le méthylome, puisque la plupart des études publiées sont basées sur des souris adultes mâles), et croiser de façon optimale nos résultats de *ChIP-seq* et de méthylome. Toutefois, certaines mises au point techniques ont été réalisées à partir de cortex d'embryons femelles.

1.3. Composition des échantillons pour les expériences de méthylome et *ChIP-seq*

Les expériences de capture du méthylome ont été réalisées à partir d'embryons sauvages issus de croisements entre souris *C57BL/6N* d'origine commerciale. Au total, 3 échantillons par condition ont été générés (appelés 1M, 2M, 3M pour les échantillons traités à l'éthanol, et 4M, 5M et 6M pour les échantillons témoins). Afin d'obtenir suffisamment d'ADN et pour réduire la variabilité inter-portée et inter-individuelle, chaque échantillon a été composé de 2 hémi-cortex droits et 2 hémi-cortex gauches, issus de 4 embryons de portées distinctes.

Les expériences de *ChIP-seq* ont été réalisées à partir de cortex d'embryons de souris mâles *Hsf2*WT ou *Hsf2*KO. Les expériences de mises au point techniques ont été réalisées à partir de cortex d'embryons de souris, mâles ou femelles, *Hsf2*WT, *Hsf2*KO ou *Hsf2*HET. En complément de l'a comparaison des cortex d'embryons dont la mère a subi des injections de PBS ou d'éthanol, nous avons aussi analysé des embryons n'ayant subi aucun stress (pas d'injection d'EtOH ou de PBS chez la mère), qui seront dits « naïfs » dans la suite du document.

2. Sexage et génotypage HSF2

Pour déterminer le sexe et le génotype des échantillons, l'ADN génomique a été extrait en incubant les queues des embryons à 95°C pendant 1h, dans un volume de tampon d'extraction (25 mM NaOH, 0.2 mM EDTA). La réaction a ensuite été neutralisée par l'ajout d'un volume équivalent de Tris-HCl à 40mM à pH5, avant d'effectuer une amplification PCR de ces échantillons en utilisant la

Taq polymérase Taq DNA Polymerase with ThermoPol Buffer (M0267S – Biolabs) ou *Taq DNA Polymerase with Standard Taq Buffer* (M0273S – Biolabs) ou encore la *Platinum Taq DNA Polymerase* (Invitrogen). Les amorces PCR utilisées sont détaillées dans le Tableau 3.1.

Tableau 3.1 : Amorces utilisées pour le génotypage de HSF2 et pour le sexage des embryons murins

	Nom des amorces	Séquences des amorces
Génotypage HSF2	HSF2 5'bis	5'-CAGTGAGAATGAATCCCTTTGGAAGG-3'
	HSF2 3'	5'-GCTGGAAGCTTCTTACCTTCCG-3'
	HSF2 lacz2	5'-CAAAGCGCCATTCGCCATTCAGG-3'
Sexage	Ube1R	5'-CACCTGCACGTTGCCCTT-3'
	Ube1F	5'-TGGATGGTGTGGCCAATG-3'

Le couple d'amorces 5'bis/3' permet d'obtenir un amplicon de l'allèle WT pour HSF2 de 95pb, alors que le couple 5'bis /Lacz_2 permet de produire un amplicon de 200pdb de l'allèle muté pour HSF2. Pour la détermination du sexe, l'amplicon obtenu avec le couple d'amorce Ube1R/Ube1F a une taille de 252pb si l'amplification se fait au niveau du chromosome X et de 334pb s'il s'agit du chromosome Y.

3. Immuno-précipitation de la chromatine (*ChIP*) ciblant HSF2 et DNMT3A

3.1. *ChIP* ciblant HSF2

Pour la *ChIP* ciblant HSF2, les embryons prélevés, ayant subi ou non une EPA, sont de génotype WT ou KO. Les cortex femelles et mâles ont été utilisés pour les mises au point techniques alors que seuls les échantillons mâles ayant subi une EPA ont été utilisés pour réaliser un *ChIP-seq* pilote. Directement après le prélèvement, les cortex ont été fixés à 4°C pendant 10 minutes avec 1% de formaldéhyde (Pierce™ *16% Formaldehyde* (w/v), *Methanol-free*, n°28906 Thermo Scientific™). La fixation a ensuite été neutralisée par l'ajout de glycine, à une concentration finale de 125 mM, sous agitation à 4°C pendant au moins 5 minutes, suivi de 2 lavages au PBS et d'un lavage au PBS avec inhibiteurs de protéases (PIC, *SigmaFast™ Protease Inhibitor Cocktail Tablets*, *EDTA-Free*). Les cortex fixés ont été regroupés 2 à 2 et lysés dans 1 mL de tampon de lyse (1% SDS, 10mM EDTA pH8, 50mM Tris-HCl pH8, 1X PIC, 1mM DTT, 1mM PMSF), pendant au moins 30 minutes, à 10°C. La chromatine a été fragmentée à l'aide du sonicateur Bioruptor® Pico (Diagenode), afin d'obtenir des fragments d'ADN d'environ 300 pb (8min, 30 sec ON / 30 sec OFF, vérification de la taille des fragments à l'aide de la TapeStation 2200® - Agilent après *reverse cross link* et purification de l'ADN). Pour éliminer les débris cellulaires, les échantillons fragmentés par sonication ont été centrifugés à 4°C pendant 1h à 14 000 rpm. Un *pre-clearing* a été effectué à 4°C pendant 1h, à partir de 200µL de lysat issu de 6 cortex, dilué dans le tampon servant à l'immuno-précipitation (Tampon d'IP : 0.2% SDS, 1% Triton X-

100, 2mM EDTA pH8, 20mM Tris-HCl pH8, 150mM NaCl, 1X PIC, 1mM DTT, 1mM PMSF), avec 35 µL de billes magnétiques (Dynabeads™, Invitrogen couplées aux protéines G).

L'immuno-précipitation (IP) a été réalisée à 4°C sur la nuit à partir de 2mL de lysat *pré-clearé* et 80 µL de billes magnétiques préalablement incubées une nuit avec l'anticorps. Pour l'IP, 12µL d'anticorps polyclonal *rabbit* anti-HSF2 ont été utilisés (anticorps non commercial SFI58, produit et généreusement procuré par l'équipe de notre collaboratrice, la Pr. Lea Sistonen, Université de Turku, Finlande ; Vihervaara et al., 2013).

Les lavages des immuno-complexes ont été faits à 4°C pendant 5 minutes, 2 fois dans le tampon d'IP, 1 fois dans le tampon de lavage n°2 (0,1% SDS, 1% Triton X-100, 2mM EDTA pH8, 20mM Tris-HCl pH8, 500mM NaCl, 1X PIC, 1mM DTT, 1mM PMSF) et 1 fois dans le tampon de lavage n°3 (20mM Tris-HCl pH8, 2mM EDTA, 10% Glycérol, 50mM NaCl, 1X PIC, 1mM DTT, 1mM PMSF).

L'étape de réversion du pontage chimique (*reverse cross-link*) a été effectuée en incubant les échantillons immuno-précipités à 65°C pendant une nuit, dans un tampon d'élution (50 mM Tris-HCl pH8, 2mM EDTA pH8, 1% SDS, 200mM NaCl, 1X PIC, 1mM DTT, 1mM PMSF). Le *reverse cross-link* a également été effectué pour un aliquot d'input équivalent à 10% d'IP, sans ajout de tampon. Les échantillons ont ensuite été traités à la RNAseA (20 µg/mL final, 1h à 37°C) puis à la protéinase K (200µg/mL, 2h à 55°C), avant de purifier l'ADN sur colonnes MinElute (MinElute® *PCR Purification Kit* n°28004, Qiagen), en suivant les recommandations indiquées par le fournisseur.

3.2. ChIP ciblant DNMT3A

Les embryons prélevés, ayant subi ou non une EPA étaient de génotype WT, HET ou KO pour HSF2. Les cortex femelles et mâles ont été utilisés pour les mises au point techniques alors que seuls les échantillons mâles WT ou KO ayant subi une EPA ont été utilisés pour réaliser un *ChIP-seq* pilote. La même chromatine a été utilisée pour le *ChIP-seq* pilote ciblant HSF2 et DNMT3A. Le protocole de *ChIP* ciblant DNMT3A est identique à celui ciblant HSF2, excepté pour les lavages des immuno-complexes, où seuls deux lavages de 5 minutes à 4°C ont été réalisés dans le tampon servant à faire l'IP.

Deux anticorps différents ont été utilisés selon le type d'expériences réalisé en aval : pour les *ChIP* servant au *western blot*, les anticorps polyclonaux *rabbit* Abcam (8µL / ChIP, ab2850, 1mg/ml) et monoclonaux *mouse* Novus (8µL / ChIP, #64B1446 ; NB120-13888 ; Lot Ab110613-B6 ; 1mg/mL) ont été testés ; pour le *ChIP-seq,* 12µL d'anticorps Novus ont été utilisés.

4. Western blot

4.1. Évaluation de l'intégrité des épitopes après sonication de la chromatine

Pour évaluer l'impact de la sonication sur l'intégrité des épitopes de HSF2 et DNMT3A, des *western blot* ont été réalisés sur des échantillons fragmentés ou non par sonication (8min, 30 sec ON / 30 sec OFF avec l'appareil Bioruptor® Pico, Diagenode, *cf protocole ChIP*). Le dosage protéique des échantillons a été estimé avec une gamme de sérum albumine bovin de concentrations définie selon le protocole du kit *Biorad® Protein Assay*, basé sur la technique de dosage protéique dit de Bradford (mesure de la densité optique à 595nm avec le luminomètre Glomax, Promega).
Les échantillons ont été dénaturés 5 minutes à 95°C dans du tampon Laemmli (2% SDS, 5% Glycérol, 30mM Tris-HCl pH6.8, 50mM DTT, 0.002% bleu de bromophénol, 1.5% β-mercapto-éthanol).

15 à 21µg de protéines par échantillon ont été chargés sur un gel SDS/PAGE 7.5%, puis transférés sur membranes PVDF (GE Healthcare Europe GmbH), dans un tampon borate (50 mM Tris-HCl et 50 mM borate) pendant 1 h30 à 4°C, à voltage constant (47V). Pour limiter les interactions non spécifiques, les membranes ont été incubées 1h dans du tampon de blocage Pierce (Thermo Scientific), avant d'être incubées durant une nuit à 4°C avec les anticorps primaires suivants : anti-HSF2 (G-11 Santa-Cruz sc-74529 X, mouse, 200µg/0.1mL, 1:2500), anti-DNMT3A (Abcam ab2850 rabbit, 1mg/mL, 1:500 ou Novus #64B1446 NB120-13888, mouse, 1mg/mL, 1:500) et anti-actine (Sigma a3853 mouse, 1.5 mg/mL, 1:3000 ou 1:4000).

Les membranes ont ensuite été lavées 3 fois dans une solution PBS-Tween 0.2% pendant 10 minutes, puis incubées 1h avec un anticorps secondaire couplé à une peroxydase de raifort : anti-rabbit (anticorps #111-035-144, Jackson Immunoresearch, 0.8 mg/mL initiale, dilution 1:80 000) et anti-mouse (anticorps #115-035-146, Jackson Immunoresearch, 0.8 mg/mL initiale, dilution 1:50 000). Après 3 lavages de 5 minutes dans du PBS-Tween 0.2%, le signal a été révélé par autoradiographie (développeuse Konica Minolta SRX 101A), en utilisant un réactif chimioluminescent (Pierce® *ECL Plus Western Blotting Substrate* ou Thermo Scientific *Super Signal West Dura Extended Duration Substrate*), sur film photographique (HyperfilmTM ECL, GE Healthcare Amersham).Ces expériences ont été réalisées au moins 3 fois, sur des séries d'échantillons indépendants.

4.2. Vérification de l'enrichissement en protéines après *ChIP*

Les inputs et l'enrichissement[3] en protéines HSF2 ou DNMT3A après *ChIP* ont également été analysés par *western blot*, sur membranes PVDF, selon le même protocole que celui présenté ci-dessus. Pour

[3] *Il nous est impossible de quantifier avec précision cet enrichissement car certains échantillons étaient très dilués et les contraintes de volumes ne permettaient pas de déposer une quantité équivalente de matériel.*

ces expériences, les *ChIP* ont été réalisées jusqu'à l'étape de *reverse-cross link* incluse. Les échantillons immunoprécipités n'ont pas été dosés, mais chargés en quantité maximale[1] sur gel SDS/PAGE 7.5%. Les inputs n'ayant pas subi de *preclearing*, plus concentrés puisqu'ils n'ont pas été dilués lors de l'expérience de *ChIP*, ont été déposés en quantité plus faible. Un input issu de cortex *Hsf2*KO, n'ayant pas subi de *reverse cross-link* a été utilisé comme témoin négatif.

Les *western blots* ont été réalisés avec les anticorps : anti-HSF2 (G-11 Santa-Cruz sc-74529 X, mouse, 200µg/0.1mL, 1:2500), et anti-DNMT3A (Novus #64B1446 NB120-13888, mouse, 1mg/mL, 1:500). Ces expériences ont été faites sur des échantillons indépendants, 2 fois pour la *ChIP* ciblant HSF2, et 1 fois pour les *ChIP* avec chacun des anticorps ciblant DNMT3A.

5. ChIP-seq ciblant HSF2 et DNMT3A

5.1. Préparation des banques et séquençage

Pour vérifier que les protocoles de *ChIP* ciblant respectivement HSF2 et DNMT3A fonctionnent, une expérience de *ChIP-seq* pilote a été effectuée sur un faible nombre d'échantillons (1 échantillon par condition, Figure 3.1A).

Les banques de séquençage (*libraries*) ont été générées par la plateforme Genom'IC (Institut Cochin, France), avec le kit Diagenode MicroPlex à partir de l'ADN extrait sur colonnes *MinElute*. Une analyse au Bioanalyzer a permis de vérifier la qualité des banques et de normaliser les quantités de matériel entre les échantillons.

Le séquençage a été réalisé sur un séquenceur Illumina NextSeq™ 500, en multiplexant les 6 échantillons sur une *flowcell Mid-output*, afin d'obtenir des *reads* de 75pb séquencés en *paired-end*. Le démultiplexage des échantillons a été réalisé par la plateforme de séquençage.

5.2. Workflow bioinformatique

La Figure 3.1B indique les grandes étapes du protocole (*workflow*) bioinformatique mis en place pour analyser le *ChIP-seq.* Les outils utilisés ainsi que les paramètres choisis pour l'analyse du *ChIP-seq* ciblant HSF2 et DNMT3A sont décrits en détail en Annexe 6-3.5 : Workflow-ChIP-seq pilote. Brièvement, les *reads* de chaque échantillon ont été contrôlés (outil FASTQC, http://www.bioinformatics.babraham.ac.uk/projects/fastqc), filtrés pour éliminer les adaptateurs et les *reads* de mauvaise qualité (outil *trim-galore*, https://github.com/FelixKrueger/TrimGalore), puis alignés sur le génome de référence mm9 (outil bowtie2, Langmead and Salzberg, 2012). Comme nous souhaitions faire une analyse en *single-end* pour éviter les biais de comptage liés au chevauchement possible entre les *reads* appariés, seuls les *reads* R1 ont été considérés. Après déduplication des *reads* (outil *samtools markdup*, Li et al., 2009), la recherche de régions enrichies en occupation par

HSF2 ou DNMT3A a été faite, en normalisant les tailles des banques avec celle de l'*input* correspondant (outil *MACS2*, Zhang et al., 2008). Une fois les pics identifiés et filtrés, les coordonnées des pics référencées selon le génome mm9 ont été converties selon le génome mm10 (outil LiftOver, UCSC, https://genome.ucsc.edu/cgi-bin/hgLiftOver) pour permettre le recoupement avec les autres jeux de données (capture du méthylome, *ATAC-seq*, *RNA-seq, cf material & method* de l'article).

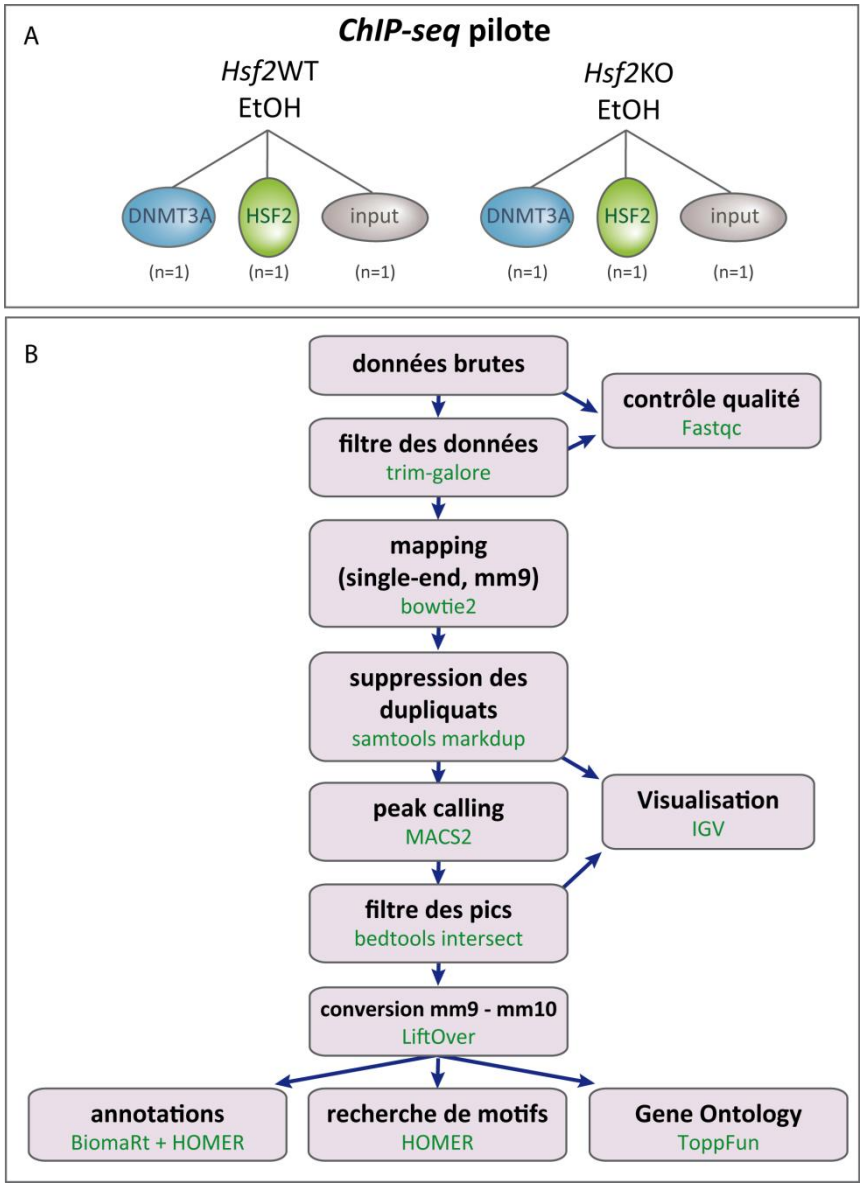


Figure 3.1 : Matériel et *workflow* bioinformatique utilisé pour le *ChIP-seq* pilote ciblant HSF2 et DNMT3A.

(A) Détail des échantillons séquencés lors du *ChIP-seq*. Les *ChIP* ciblant HSF2 et DNMT3A ont été réalisés avec la même chromatine issue de cortex embryonnaires murins E16.5, WT ou KO pour HSF2. Ainsi, seul un input par phénotype (*Hsf2*WT ou *Hsf2*KO) est nécessaire.

(B) Etapes clés du *workflow* bioinformatique suivies pour l'analyse du *ChIP-seq*. Les outils utilisés sont indiqués en vert en-dessous de chaque étape. Pour plus de détails sur les outils et paramètres utilisés, voir Annexe 6-3.5 : Workflow-ChIP-seq-pilote.

Les séquences connues comme étant retrouvées de façon récurrente dans les *ChIP* sont retirées (*blacklist*, voir en Annexe 6-3.5 le workflow détaillé). Les données d'alignement et de *Peak calling* ont été visualisées avec IGV (Robinson et al., 2011). Une recherche d'enrichissement de motifs a été réalisée sur ces données converties (outil *findMotifGenome* de la suite HOMER, Heinz et al., 2010). Les données ont également été annotées selon l'annotation syntaxique fournie par le package R BioMart (Durinck et al., 2009; Huang et al., 2009), avant de réaliser une analyse de *Gene Ontology* selon les sites ToppFun (suite ToppGene, Chen et al., 2009). L'outil *homer_annotatePeaks* (sur le site GalaxEast - www.galaxeast.fr) a permis de définir le type d'éléments génomiques (exons, introns…) ciblés par HSF2.

Chapitre 4 : Résultats

Partie 1 : Résultats de la capture du méthylome

1. Mise en place d'un *workflow* adapté à l'analyse d'une capture du méthylome

1.1. Intérêts de l'étude d'une capture du méthylome

Pour étudier les effets précoces de l'EPA sur la méthylation de l'ADN, dans le cerveau en développement, nous avons choisi de travailler à partir d'une capture du méthylome (technologie d'épicapture *SeqCapEpi Developer Medium Enrichment kit*, NimbleGen, Roche), plutôt que de réaliser une analyse à l'échelle du génome entier. Ce type de méthodologie a déjà été utilisé dans la littérature, aves succès, pour étudier la méthylation de l'ADN (Allum et al., 2015; Ball et al., 2009; Hodges et al., 2009; Lee et al., 2011; Li et al., 2015; Teh et al., 2016). La technique de capture du méthylome que nous avons utilisée est basée sur une hybridation de l'ADN à des sondes couplées à la streptavidine se fixant à des billes biotinylées. Elle permet de sélectionner des régions d'intérêt, spécialement choisies pour l'étude (Figure 4-1.1).

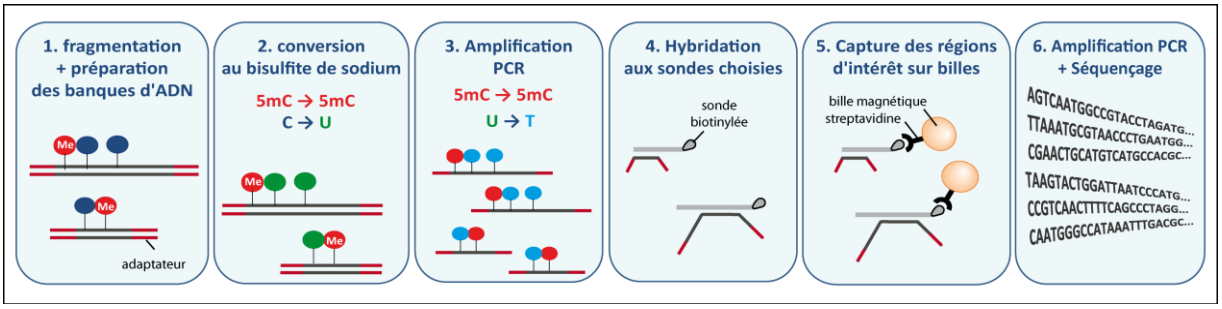


Figure 4-1.1 : Représentation schématique des grandes étapes du protocole de capture du méthylome

Technique de capture à façon permettant l'étude du méthylome de régions d'ADN ciblées, selon la méthode *SeqCap Epi Enrichment* (*NimbleGen, Roche).* L'ADN des échantillons est extrait, puis fragmenté avant l'ajout de séquences adaptatrices permettant le séquençage. Afin de distinguer les cytosines méthylées (5mC et 5hmC) des cytosines non méthylées (C), l'ADN fragmenté subit un traitement au bisulfite de sodium convertissant les cytosines non méthylées en uraciles, puis en thymines après amplification PCR, mais n'affectant pas les cytosines méthylées. Des sondes biotinylées réalisées sur mesure, s'hybrident aux fragments génomiques d'intérêt, avant d'être couplées à des billes couvertes de streptavidine par affinité. Après des lavages permettant d'éliminer les fragments d'ADN non fixés aux billes, seules les séquences d'ADN d'intérêt sont conservées, amplifiées puis séquencées. Une analyse bio-informatique permet ensuite d'identifier les régions différentiellement méthylées (DMR) entre deux conditions.

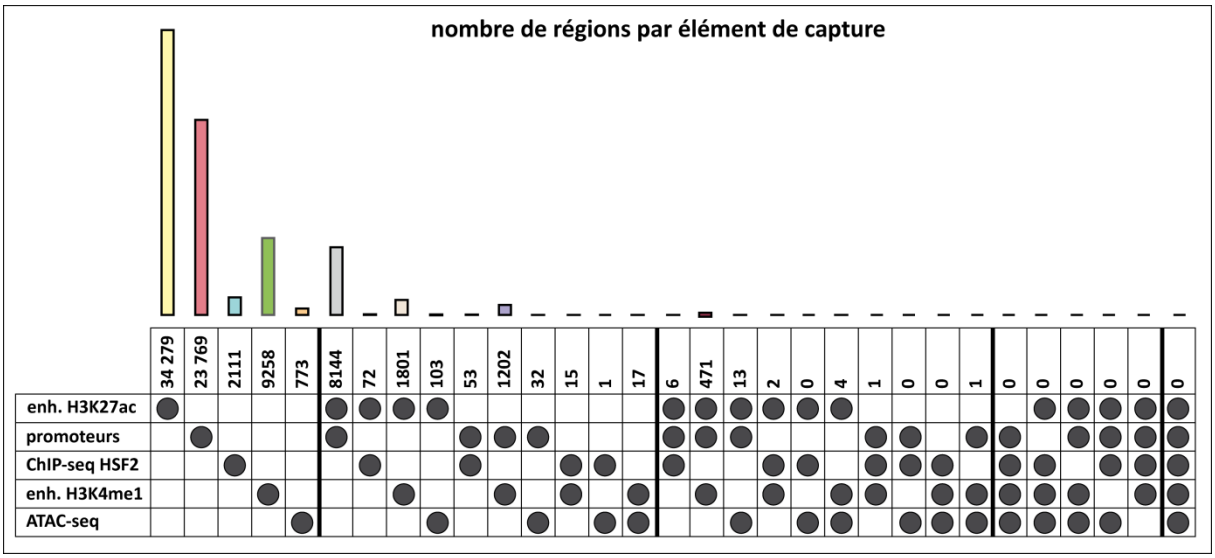


Figure 4-1.2 : Composition de la capture.

Nombre de régions capturées par catégorie (*enh*. H3K27ac, promoteurs…). Ces catégories sont considérées séparément ou au contraire associées les unes aux autres afin de déterminer le nombre de régions appartenant à plusieurs catégories. Dans le tableau, les ronds indiquent les jeux de catégories considérées. ***Abréviations et détails des catégories*** : ***enh. H3K27ac*** : *enhancers* actifs (porteurs de la marque H3K27ac) dans le cortex murin adulte de 8 semaines ; ***ChIP-seq HSF2*** : régions fixées par HSF2 en conditions physiologiques dans le cortex embryonnaires E16.5 d'après un ancien *ChIP-seq,* réalisé en 2010 ; ***enh. H3K4me1***: *enhancers* du cortex murin adulte portant la marque H3K4me1 au sein des gènes identifiés comme différentiellement exprimés dans la microglie suite au stress périnatal IL-1β, à différents stades de développement (Krishnan et al., 2017) ; ***ATAC-seq*** : régions différentiellement ouvertes ou fermées dans des oligodendrocytes, après un stress périnatal IL-1β (Schang et al., 2018).

Cette capture personnalisée, réalisée à partir d'ADN de cortex embryonnaires murins, couvre environ 81 millions de bases, réparties sur 58 611 *loci* différents (Figure 4-1.2) :

- Plus de 75% des promoteurs de souris[4] (+/- 500 bases autour du site d'initiation de la transcription, données issues de la base de données ENCODE)
- l'ensemble des *enhancers* actifs (*i.e.* porteurs de la marque d'histone H3K27ac) dans le cortex de souris adulte de 8 semaines (données issues de la base de données ENCODE)
- des régions d'intérêt selon les résultats d'expériences précédentes de l'équipe ou de nos collaborateurs, *i.e.* :
    - les régions fixées par HSF2 en conditions physiologiques, dans le cortex embryonnaire (E16.5, résultats d'une expérience de *ChIP-seq* réalisé en 2010, travaux de R. El Fatimy, I. Massaoudi, et A. Le Mouël).

[4] *Nous souhaitions capturer l'ensemble des promoteurs murins mais certains d'entre eux n'ont pas pu être étudiés, soit parce qu'ils contenaient des séquences répétées empêchant de les capturer spécifiquement, soit parce qu'aucun nom de gène n'a pu être associé à ces régions promotrices capturées, avec le système d'annotation utilisé (BiomaRt).*

- les régions perturbées par un stress inflammatoire périnatal (injection d'IL-1β à des nouveaux nés de P1 à P5, du premier au 5ème jour après la naissance) :
  - régions différentiellement ouvertes ou fermées après ce stress (résultats d'une expérience d'*ATAC-seq* réalisée sur des précurseurs d'oligodendrocytes isolés (O4+) à partir de cerveaux de souris nouveau-nées de 5 jours (travaux de A.L. Schang et D. Sabéran Djoneidi, (Schang et al., 2018).
  - enhancers du cortex murin adulte portant la marque H3K4me1 (+/- 100 bases autour du centre des pics H3K4me1), au sein des gènes identifiés comme différentiellement exprimés suite au stress IL-1β, à différents stades (uniquement à P45, ou à P1, P5 et P10, ou à P5 et P10, résultats d'une expérience de *micro-array* réalisée à partir de microglie (cellules CD11B+) issues de cerveaux de souris âgées de P5, P10 ou P45 ayant subi ce stress inflammatoire (travaux de A.L. Schang et collaborateurs, équipe de P. Gressens, UMR1141, Hôpital Robert Debré, Paris, Krishnan et al., 2017).

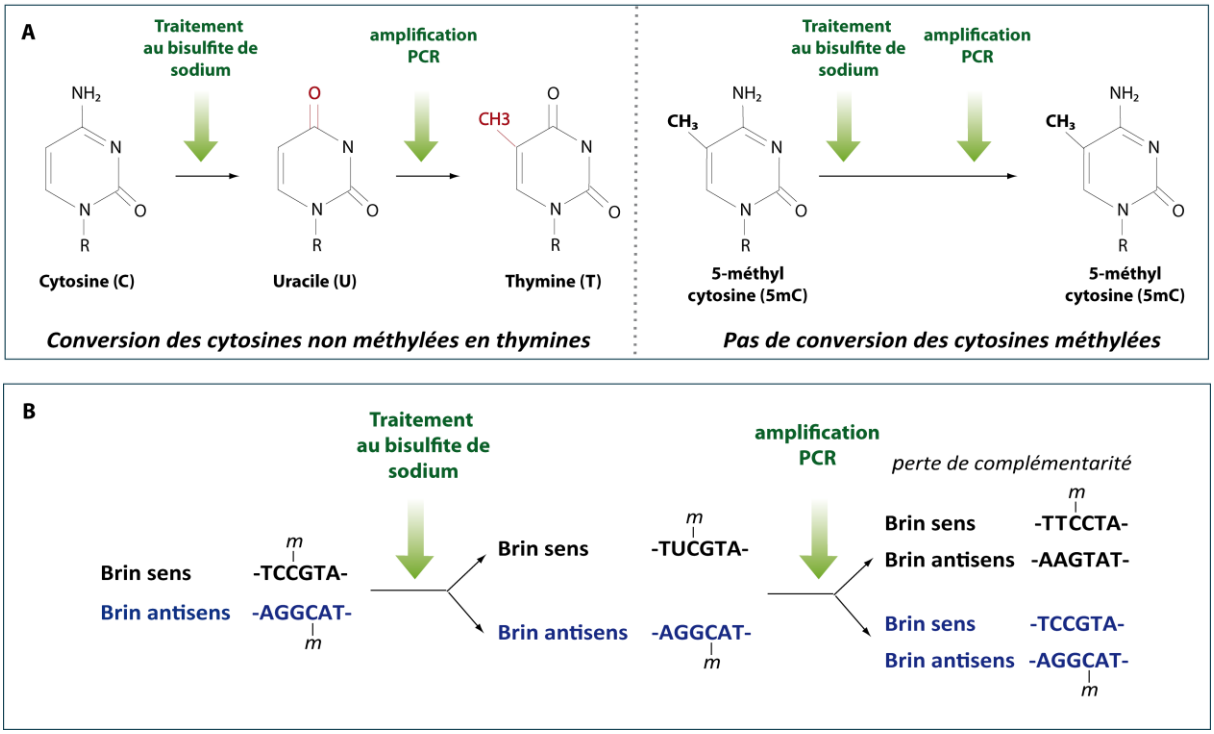


Figure 4-1.3 : Effet du traitement au bisulfite de sodium sur l'ADN.

(A) L'effet du bisulfite de sodium sur les cytosines dépend de leur état de méthylation. Le traitement au bisulfite de sodium convertit les cytosines non méthylées (C) mais n'affecte pas les cytosines méthylées (5mC) ou hydroxy-méthylées (5hmC). Ainsi, seules les cytosines non méthylées sont modifiées en uraciles, eux-mêmes converties en thymines après amplification PCR.

(B) Perte de complémentarité entre les brins d'ADN après traitement au bisulfite de sodium. La conversion des cytosines non méthylées par le bisulfite de sodium engendre une perte de complémentarité entre les brins sens et antisens d'une séquence d'ADN. Après amplification PCR, un fragment d'ADN donné est alors représenté par 4 séquences d'ADN simple brin distinctes. La marque « m » au niveau des cytosines représente un groupement méthyle.

Afin de définir l'état de méthylation des CpG des différents échantillons, l'ADN génomique issus de ces derniers sont traités au bisulfite de sodium, avant d'isoler les fragments d'intérêt sur billes biotinylées. Ce traitement au bisulfite de sodium convertit les cytosines non méthylées en uraciles - puis en thymines après amplification PCR -, sans affecter les cytosines hydroxy-méthylées ou méthylées (Figure 4-1.3A ; Krueger et al., 2012). Deux brins d'ADN initialement complémentaires perdent ainsi leur complémentarité après un tel traitement (Figure 4-1.3B), ce qui nécessite une analyse bioinformatique spécifique, notamment pour tenir compte des dupliquats de PCR. Les sondes permettant la capture du méthylome sont dites "*dégénérées*" (plusieurs sondes distinctes pour un fragment d'ADN donné) afin de prendre en considération les différentes options possibles de conversion au bisulfite de sodium et ainsi capter toute la complexité du méthylome.

Tableau 4-1.1 : avantages et inconvénients de l'analyse du méthylome global capturé

	Whole Genome Bisulfite Sequencing ***(WGBS,* analyse globale*)***	**Capture du méthylome (système *SeqCapEpi*)**
Régions étudiées	Génome entier	Régions d'intérêt, choisies par l'expérimentateur
Profondeur de séquençage	Fort taux de séquençage (plusieurs milliards de *reads*) pour une faible profondeur de couverture au niveau des régions d'intérêt, identification de DMR limitée aux régions dont le taux de méthylation est très différent entre 2 conditions.	Séquençage limité aux régions d'intérêts (dizaines de millions de *reads*) permettant pour un même nombre de *reads*, une plus grande profondeur de couverture de ces régions, donc une recherche de DMR plus résolutive que le *WGBS.*
Coût de séquençage	Elevé	Faible
Analyse bioinformatique	-Fichiers générés volumineux -Analyse lourde (temps de calcul long) mais des outils adaptés sont disponibles	-Fichiers moyennement volumineux -Analyse nécessitant moins de ressources (temps de calcul réduit) mais outils d'analyses pas toujours adaptés à la capture (biais)
Echantillons	Limité	Adapté aux tests diagnostiques à grande échelle

Cette technique de capture du méthylome présente plusieurs avantages par rapport aux études globales du méthylome (*Whole Genome Bisulfite Sequencing - WGBS,* Tableau 4-1.1). En effet, la méthode *WGBS* est onéreuse car elle nécessite le séquençage de l'ensemble du génome, alors que, à titre indicatif, la proportion de CpG dans le génome humain ciblée, et donc informative dans ce genre d'étude, est inférieure à 1% (Robinson et al., 2010a). En utilisant une capture du méthylome, le séquençage étant limité aux régions capturées, le coût de séquençage est réduit lorsque l'on souhaite une grande profondeur de couverture des régions d'intérêt. Ainsi, à partir d'une faible

quantité d'ADN (750ng - 1µg), les possibilités d'identifier des régions significativement différentiellement méthylées entre deux conditions sont maximisées, à la fois par cette grande profondeur de couverture et par son uniformité (Fu et al., 2010). Cet aspect était particulièrement important pour ce projet de thèse en raison de la quantité de matériel limitée et difficile à collecter (analyses de cortex embryonnaires, distinguant les mâles et les femelles, en conditions contrôle et après EPA). Outre le coût important du *WGBS*, la logistique nécessaire au traitement bio-informatique des données (stockage de données, mais aussi mémoire vive) est importante pour de telles analyses, les fichiers générés étant très volumineux (Tableau 4-1.1, Robinson et al., 2010). Ainsi, la capture du méthylome permet de travailler à partir de fichiers moins lourds, ce qui réduit les ressources informatiques nécessaires à l'exploitation des données de séquençage.

1.2. Choix du logiciel d'alignement

Bien que le temps de calcul de l'analyse d'une capture du méthylome soit grandement réduit comparé à celui d'une analyse *WGBS* pour laquelle la quantité de données à analyser est considérable (Tableau 4-1.1), des précautions sont nécessaires pour mener à bien toute étude bio-informatique d'échantillons converties au bisulfite de sodium. Une première difficulté réside dans l'alignement (*mapping*) des séquences sur le génome de référence, puisque la modification des cytosines non méthylées en thymines par le traitement au bisulfite de sodium engendre une perte de complémentarité, entre les *reads* séquencés, et le génome de référence (Figure 4-1.3B). Les outils « classiques » de *mapping* ne peuvent pas être utilisés, mais divers outils bio-informatiques ont été spécialement développés pour pallier ce problème (Bock, 2012; Krueger et al., 2012). La société Roche préconise l'utilisation de l'outil *BSMAP* (Xi and Li, 2009) pour l'analyse de la capture du méthylome. Cet outil est basé sur un algorithme d'alignement de type « lettre joker » (*wild-card bisulphite aligner*), ce qui signifie qu'il modifie le score d'alignement afin de ne pas pénaliser les mésappariements entre les cytosines du génome de référence et les thymines des *reads* séquencés (Bock, 2012; Krueger et al., 2012). Cela revient à remplacer les cytosines (C) du génome de référence par la lettre Y (code nucléotidique IUPAC - www.bioinformatics.org/sms/iupac.html), permettant l'alignement à la fois des cytosines et des thymines des *reads* séquencés (Robinson et al., 2010a; Bock, 2012).

Nous avons décidé de ne pas utiliser cet outil, qui comme les autres logiciels utilisant ce type d'algorithme, permet certes d'obtenir une largeur de couverture importante, mais au détriment de la fiabilité du l'estimation du taux de méthylation (Robinson et al., 2010a; Chatterjee et al., 2012; Bock, 2012). En effet, ces outils présentent un risque accru de définir des niveaux de méthylation plus élevés que ce qu'ils ne sont en réalité, car pour une région donnée, les *reads* méthylés, porteurs de

cytosines, sont constitués de séquences plus complexes que les *reads* non méthylés, ne comportant pas de cytosines mais uniquement des thymines après conversion au bisulfite de sodium. Les *reads* méthylés auront ainsi, de par leur complexité de séquences, plus de chances de s'aligner sur le génome de référence de manière unique et d'être conservés, alors que les *reads* non méthylées seront éliminés de l'analyse car leur alignement risque fortement d'être ambigu car multiple (Robinson et al., 2010a; Krueger et al., 2012; Bock, 2012).

Pour éviter ce biais qui nous semble majeur, nous avons donc décidé de ne pas suivre les recommandations de Roche pour le choix du logiciel d'alignement à utiliser pour la capture du méthylome, ce qui m'a contrainte à générer mon propre *workflow* d'analyse bio-informatique (*cf* Figure 1C et *Materials&Methods* de l'article pages suivantes ainsi que l'Annexe 6-3.1 de ce manuscrit). Nous avons préféré utiliser l'outil Bismark (Krueger and Andrews, 2011), basé sur un algorithme d'alignement de type « 3 lettres » (*three letter alignment*), qui facilite l'alignement de séquences traitées au bisulfite de sodium, en convertissant toutes les cytosines en thymines, à la fois au niveau des *reads* et des deux brins du génome de référence (Bock, 2012; Krueger et al., 2012; Robinson et al., 2010a). De cette façon, l'alignement est réalisé exclusivement sur un génome de référence ne possédant que 3 bases distinctes (A, G et T), à l'aide de l'outil bowtie2 (Langmead and Salzberg, 2012), un outil standard d'alignement de séquences intégré à Bismark (Bock, 2012). En raison de la réduction de la complexité des séquences, (seulement trois bases restantes), un plus grand nombre de *reads* s'aligne à plusieurs endroits du génome de référence et ne sont pas pris en compte dans l'analyse à cause de cette ambigüité de positions d'alignement. La largeur de couverture est donc restreinte, mais le taux de méthylation n'est pas biaisé, puisque l'ambiguïté d'alignement est la même pour tous les échantillons, que l'ADN soit méthylé ou non (Bock, 2012; Krueger et al., 2012; Robinson et al., 2010a).

Une étude de Kunde-Ramamoorthy et collaborateurs (2014) a certes remis en cause l'argument avancé par Chatterjee et collaborateurs (2012) et Bock (2012) concernant le biais d'estimation du taux de méthylation de l'ADN des algorithmes d'alignement de type « lettre joker », mais a conforté notre volonté de ne pas utiliser BSMAP, estimé moins performant que bismark. En effet, cette étude montre qu'une grande proportion de régions d'intérêts (*e.g.* régions présentant des variations de la méthylation de l'ADN tissu-spécifiques) est couverte donc analysable en utilisant l'outil bismark, alors qu'elles ne le sont pas avec BSMAP.

1.3. Création d'un outil adapté à la détection de DMRs au sein de la capture du méthylome

Une fois les *reads* cartographiés sur le génome de référence, l'état de méthylation de chaque cytosine doit être déterminée, dans le but d'identifier de potentielles différences entre le groupe contrôle (injection PBS, n=3) et le groupe ayant subi l'EPA (injection EtOH, n=3 ; *cf* pages suivantes Figure 1A du manuscrit Duchateau *et al.*, en préparation).

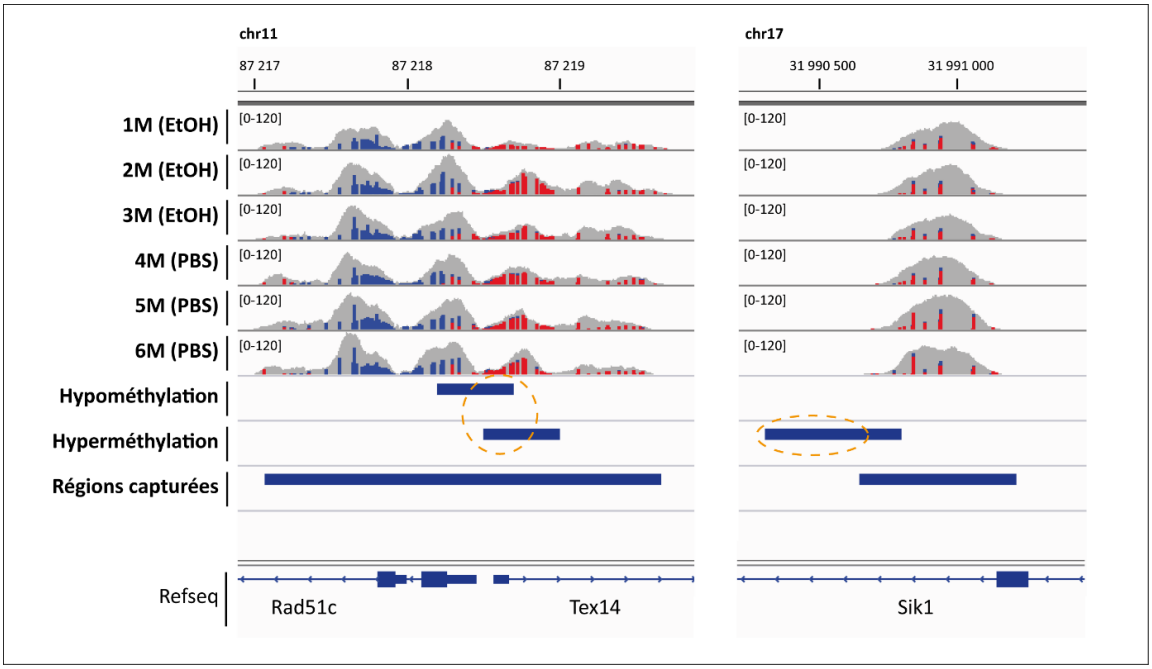


Figure 4-1.4 : Exemples de DMR anormales détectées par *MethylKit* à partir de nos données méthylome, issues d'une capture de régions génomiques spécifiques.

Une même région est parfois identifiée par *Methylkit* comme étant à la fois hypo- et hyperméthylées (à gauche). Comme cet outil effectue un découpage systématique du génome de référence pour détecter les DMR, sans tenir compte des coordonnées des régions capturées, certaines régions sont identifiées comme différentiellement méthylées, sans disposer des données de méthylation pour la partie de la DMR hors capture (à droite).

Aperçus obtenus avec le logiciel IGV. Le code couleur permet de distinguer l'état de méthylation des cytosines en contexte CpG (C méthylée en rouge et C non méthylée en bleu). Analyse *MethylKit* réalisée à partir des cytosines en contexte CpG, limitée aux régions incluses dans la capture. Les données ont été filtrées avant le test statistique (*i.e.* suppression des cytosines dont la couverture est inférieure à 10 et des cytosines anormalement trop couvertes, dont la couverture a une valeur supérieure au 99.9^{ème} centile). Les données restantes ont été normalisées, puis le découpage systématique du génome a été réalisé pour définir des régions de 500bases ayant un chevauchement de 400bases. Sur 769 298 régions totales représentées dans au moins 2 échantillons par groupe (PBS ou EtOH), 5 588 régions hyperméthylées et 4 435 régions hypométhylées ont été détectées par *MethylKit*, après EPA, selon un seuil de *q-value* de 0.05 et un différentiel de méthylation d'au moins 5% (soit respectivement 2 544 et 2 215 régions uniques, une fois les régions chevauchantes associées).

Comme la méthode employée ne repose pas sur une analyse en cellule unique (*single cell*), un « niveau » de méthylation doit être déterminé pour une cytosine donnée, correspondant à la proportion de molécules dans la population cellulaire étudiée qui porte la marque méthylée au niveau de la cytosine correspondance. Ce niveau, ou pourcentage de méthylation, peut être défini pour chaque cytosine grâce à des outils statistiques, comme par exemple *MethylKit* (Akalin et al., 2012) ou RADMeth (suite MethPipe, Dolzhenko and Smith, 2014; Song et al., 2013).

Même s'il arrive que l'état de méthylation d'une cytosine isolée ait un impact sur la régulation de l'expression de certains gènes (Xu et al., 2007), la plupart des publications scientifiques est basée sur des analyse de régions différentiellement méthylées (DMR) entre deux conditions, faisant, entre quelques centaines de bases, et quelques milliers de bases (Bock, 2012). Une seconde difficulté, plus spécifique à l'analyse du méthylome par capture, apparait alors, concernant l'identification des régions différentiellement méthylées (*DMR*). En effet, les outils existants définissent souvent les régions génomiques à étudier en réalisant un découpage systématique du génome de référence entier selon des paramètres définis par l'utilisateur (taille des régions, chevauchement entre les régions…)[5]. Ce type d'outil n'est pas adapté pour identifier des DMR au niveau d'une capture du méthylome. En effet, j'ai testé plusieurs outils (*MethylKit*, RadMeth), qui sont adaptés pour la détection des niveaux de méthylation des CpG isolées, mais insatisfaisants pour l'identification de régions à partir de données issues d'une capture du méthylome réalisée à façon, comme le montre la Figure 4-1.4.

Au vu de la spécificité de la capture que nous étudions, nous avons décidé de créer notre propre outil de détection de *DMR* (Figure 4-1.5, en collaboration avec Pr. S. Ott, Univ. de Warwick, UK). Le principe de notre outil - fonction *get_close_loci()* - est de regrouper les CpGs voisines en une région unique, lorsque qu'elles présentent des différences de méthylation similaires après l'EPA (*i.e.* regrouper les CpGs toutes hypométhylées, ou au contraire les CpGs toutes hyperméthylées, Figure 4-1.6, Annexe 6-3.1). L'association de ces CpGs isolées, parfois peu significatives lorsqu'elles sont considérées séparément, permet alors de former une DMR plus robuste et possédant un potentiel fonctionnel.

[5] Des outils ont également été développés pour étudier des régions pré-sélectionnées, comme celles des puces de type « 450K », mais ils ne sont pas adaptés non plus à une capture faite sur mesure, car les séquences constituants ces puces sont différentes de celles trouvées dans une capture personnalisée.

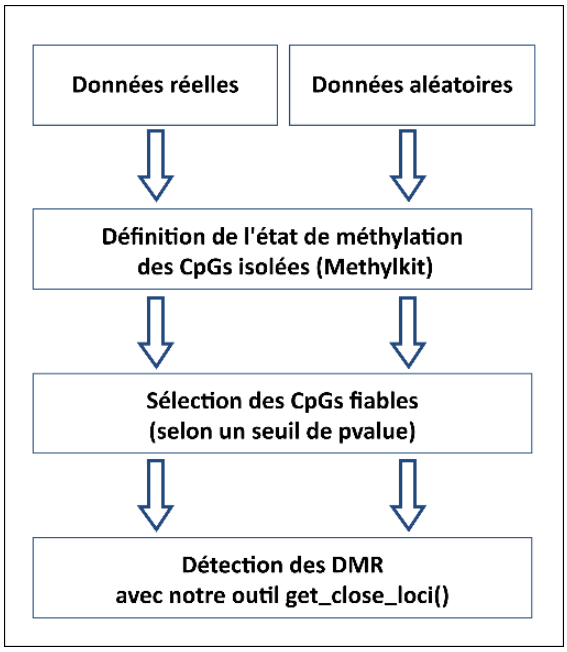


Figure 4-1.5 : Principales étapes de mon approche permettant de détecter les DMR pertinentes dans notre jeu de données provenant d'une capture du méthylome

Notre outil possède des paramètres permettant de moduler la recherche de DMR, selon le jeu de données utilisé et la question biologique posée. Ainsi, il est possible de définir :

- un nombre minimum de CpGs nécessaires pour former la *DMR*
- la taille maximale des *DMR*
- une distance maximale autorisée entre deux CpGs successives, afin de ne regrouper que les CpGs relativement proches.

Nous avons défini le pourcentage de méthylation de la DMR comme correspondant à la valeur médiane de méthylation des CpG qui la compose.

Pour utiliser l'outil que nous avons créé, il est nécessaire de connaître, pour chaque CpG, la différence de méthylation entre le groupe contrôle et le groupe testé. Comme l'outil *MethylKit* est adapté à la détection des niveaux de méthylation à l'échelle des CpGs isolées, et qu'il fournit des informations statistiques fiables (pour chaque CpG : pourcentage du différentiel de méthylation entre les groupes contrôle et EPA, associé à une *p-value* et une *q-value*[6]), basées sur un modèle statistique conçu pour l'étude du méthylome, nous avons choisi d'utiliser ces données comme point d'entrée.

[6] *Comme nous recherchons des différences de méthylation de l'ADN sur un très large nombre de régions, il est nécessaire de tenir compte de ces tests multiples pour éviter d'obtenir un taux trop élevé de « DMR faussement positives ». C'est pourquoi les p-values sont corrigées en q-values.*

Afin de déterminer des paramètres satisfaisants pour identifier les *DMR* exploitables, nous faisons face à un compromis : il faut à la fois utiliser des paramètres stricts, pour ne pas avoir trop de régions « parasites » sans intérêt (faux positifs), mais il faut aussi avoir recours à des paramètres permissifs, pour éviter d'éliminer trop de régions d'intérêt (faux négatifs).

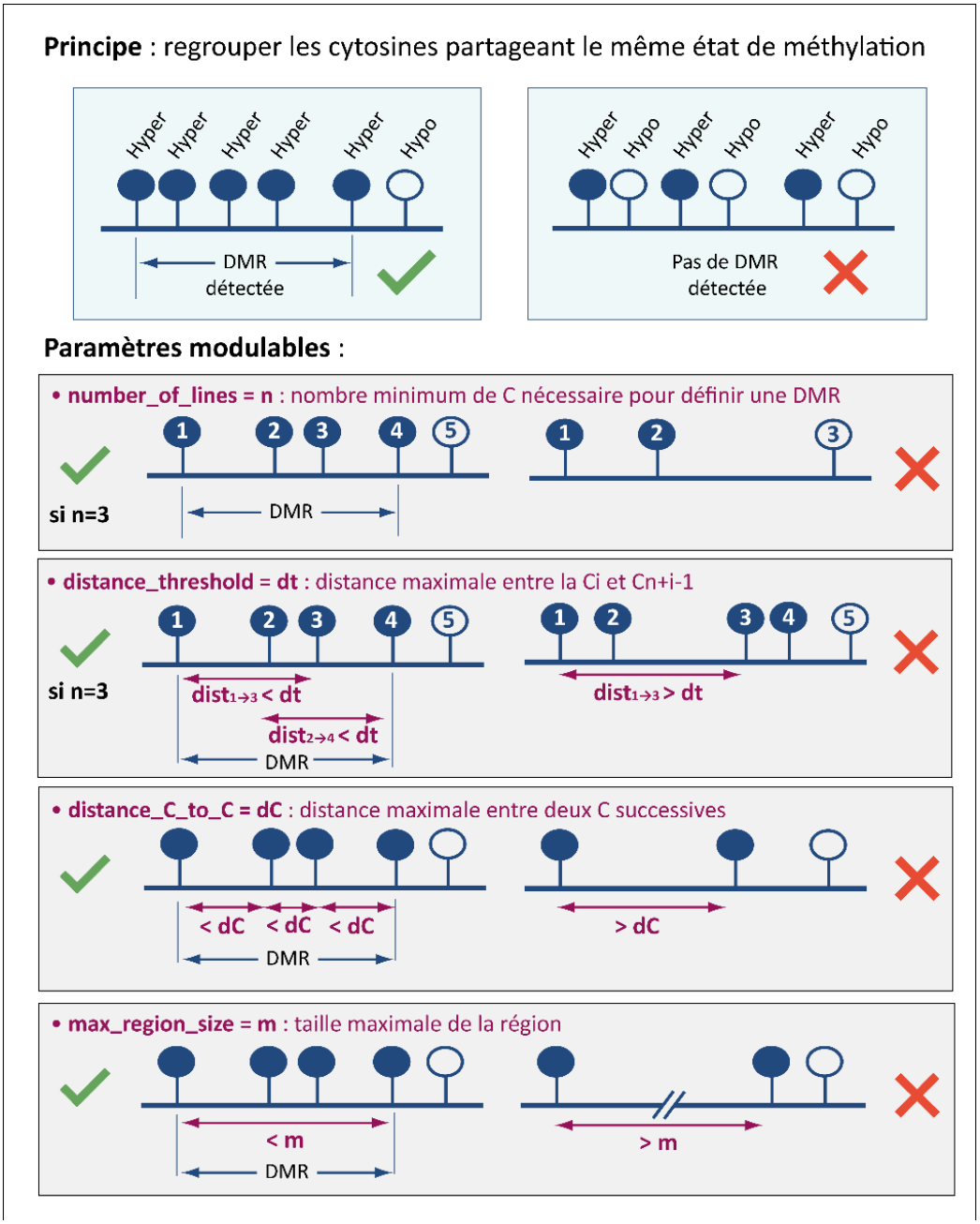


Figure 4-1.6 : Principe et paramètres de la fonction *get_close_loci()* créée spécifiquement pour la détection de DMR au sein de la capture du méthylome.

Plusieurs paramètres peuvent être modulés afin de choisir les critères de détection des DMR, selon le type de données et la question biologique posée (création de l'outil en collab. Pr. S. Ott, Univ. de Warwick, UK). Les ronds pleins représentent des cytosines hyperméthylées entre deux jeux de données (contrôle et test), les ronds vides représentent des cytosines hypométhylées après EPA.

Pour faciliter le choix de paramètres adaptés à l'identification de *DMR* dans notre jeu de données, mais aussi pour valider ces paramètres, j'ai testé différentes options, à la fois sur notre jeu de données (données réelles), mais aussi sur un jeu de données aléatoires que nous avons généré (cf section suivante - Génération de données aléatoires pour définir les paramètres de détection des DMR). La comparaison des résultats obtenus dans les deux situations permet ainsi de définir les paramètres qui permettent de répondre au compromis présenté ci-dessus.

Une pré-sélection des CpGs pertinentes est indispensable pour une détection efficace des *DMR*. En effet, la présence, par exemple, d'une CpG définie faussement comme hypométhylée après l'EPA, parmi un ensemble de CpGs hyperméthylées, suffira à empêcher la détection de l'ensemble de la région hyperméthylée après le stress (Figure 4-1.7). Il est donc nécessaire de ne travailler qu'avec des CpGs pour lesquelles un état de méthylation a été déterminé avec fiabilité.

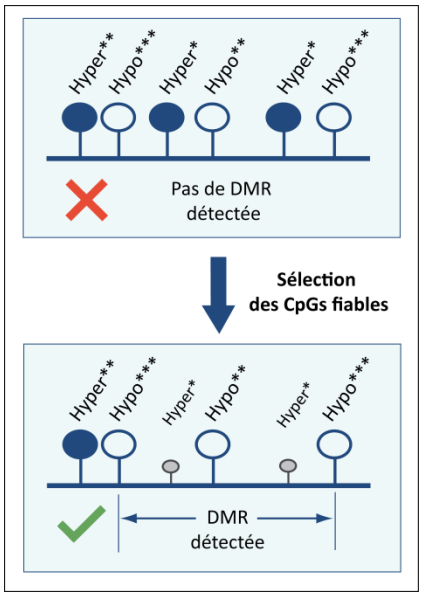


Figure 4-1.7: La sélection des « CpGs fiables » est indispensable pour la détection des DMR.
En conservant, dans le jeu de donnés, des CpGs pour lesquelles l'état de méthylation n'est pas défini de façon fiable, la recherche de DMR peut être compromise. Dans cet exemple, la conservation des deux CpGs considérées, de façon peu fiable, comme hyperméthylées après traitement, empêche la détection de la DMR hypométhylée, alors qu'elle existe de façon fiable. Un seuil de *pvalue* doit donc être défini pour définir les informations de méthylation pertinentes.
Légendes : **ronds pleins** : cytosines hyperméthylées après traitement ; **ronds vides** : cytosines hypométhylées après traitement ; * : *pvalue* peu fiable ; *** *pvalue* fiable.

Nous avons choisi d'effectuer cette pré-sélection selon un seuil de *p-value* associée au différentiel de méthylation déterminé par *MethylKit,* plutôt que d'utiliser la *q-value* pour la raison suivante : cette dernière étant basée sur la méthode SLIM (Wang et al., 2011), qui corrige la *p-value* en tenant compte de la dépendance entres les CpGs successives, elle n'est pas calculée de la même

manière sur le jeu de données réelles et le jeu de données aléatoires. La *q-value* ne peut donc pas être utilisée comme critère de sélection sans introduire un biais dans l'analyse.

Finalement, pour identifier les paramètres satisfaisants pour détecter les DMR, nous avons fixés les paramètres suivants :

- la taille maximale des DMR : **2000 bases**
- la distance maximale autorisée entre deux CpGs fiables[7] successives : **100 bases**

en modulant ces paramètres :

- la valeur seuil de *p-value* en dessous duquel on considère les informations de différentiel de méthylation comme fiables : **0.008, 0.01, 0.03, 0.05, 0.06, 0.07, 0.08, ou 0.1**
- le nombre minimum d*e CpGs n*écessaires pour former la DMR : **3, 4 ou 5**

et en comparant les résultats *o*btenus à partir du vrai jeu de données et du jeu de données aléatoires.

Les résultats de ces tests sont présentés dans le Tableau 4-1.2. Selon le seuil de *p-value* choisi, un nombre plus ou moins important de CpGs est sélectionné, mais quel que soit ce seuil, un plus grand nombre de CpGs est sélectionné dans le vrai jeu de données par rapport aux jeux de données aléatoires. Cette différence indique que les données réelles ne sont pas dues au hasard : le jeu de données généré aléatoirement présente une plus grande variabilité intra-groupe et/ou une plus faible différence inter-groupe que le jeu de données réelles, ce qui rend l'estimation du différentiel de méthylation entre les groupes moins fiables dans ce jeu de données. En sélectionnant le même nombre de CpGs dans les deux jeux de données, j'ai pu identifier les paramètres permettant d'avoir un enrichissement satisfaisant dans le vrai jeu de données, comparé aux données aléatoires. Les critères que nous avons finalement retenus sont un seuil de *p-value* de 0.07 et un nombre minimum de 5 CpGs (« fiables ») par *DMR*. Ils nous conduisent à accepter un taux de faux positif estimé à 22,8%. Avec ces critères, nous obtenons **432 régions différentiellement méthylées** après l'EPA (cf Annexe 6-4.1 – onglet DMR432).

1.4. Génération de données aléatoires pour définir les paramètres de détection des DMR.

Des échantillons aléatoires ont été produits et comparés au vrai jeu de données pour définir des paramètres de détection des *DMR* adaptés à notre jeu de données.

Afin de générer ces données aléatoires, j'ai écrit un script R, qui permet de réattribuer aléatoirement, position par position, les valeurs de méthylation observées dans les échantillons du jeu de données réelles (Figure 4-1.8).

[7] *CpGs fiables* selon le seuil de *p-value* choisi

Tableau 4-1.1 : Résultats des comparaisons entre jeux de données réelles et aléatoires, et paramètres finalement choisis pour la détection de DMR, avec la fonction *get_close_loci()*.

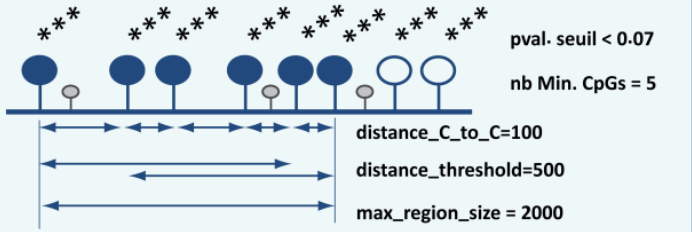
Avec le jeu de données aléatoires n° 1

		Nb C sélectionnées		Nb de DMR détectées			
pvalue seuil	nb min. C	données réelles	données aléatoires	données réelles	données réelles ajustées	données aléatoires	réelles ajust. ÷ aléatoires
0.008	3	12 419	7 225	114	28	9	3.1
	4			16	2	0	-
0.01	3	15 089	8 942	160	40	17	2.4
	4			21	5	0	-
0.03	3	38 843	26 785	1080	453	263	1.7
	4			240	85	20	4.3
	5			64	22	3	7.3
0.05	3	60 936	44 899	2603	1380	924	1.5
	4			672	229	127	2.4
	5			225	94	18	5.2
0.08	3	92 657	73 225	5552	3506	2563	1.3
	4			1684	935	461	2.0
	5			581	287	81	3.5
0.1	3	118 453	95 687	8331	5659	4358	1.2
	4			2733	1679	912	1.8
	5			1016	560	190	2.9

Avec le jeu de données aléatoires n° 2

		Nb C sélectionnées		Nb de DMR détectées			
pvalue seuil	nb min. C	données réelles	données aléatoires	données réelles	données réelles ajustées	données aléatoires	réelles ajustées ÷ aléatoires
0.05	3	60 936	44 584	2603	1280	814	1.6
	4			672	281	98	2.9
	5			225	87	15	5.8
0.06	5	71 157	53 441	329	138	28	4.9
0.07	5	81 494	62 842	432	184	42	4.4
0.08	3	92 657	73 170	5552	3467	2368	1.5
	4			1684	932	429	2.2
	5			581	263	72	3.7

Paramètres finalement sélectionnés pour la détection de DMR



Deux jeux de données aléatoires ont été générés et comparés aux données réelles, en testant différentes combinaisons de paramètres de détection des DMR (*pvalue* seuil, nombre de CpG minimum dans la région), et en fixant des paramètres de distances (*distance_C_to_C* = 100, *distance_threshold*=500, *max_region_size* = 2000). Selon le seuil de *pvalue* choisi, un nombre plus ou moins important de CpGs est sélectionné. A chaque fois, un plus grand nombre de CpGs est sélectionné dans le vrai jeu de données par rapport aux jeux de données aléatoires pour un seuil de *p-value* donné. Or, pour pouvoir comparer le nombre de DMR obtenu dans le jeu de données réelles et dans le jeu de données aléatoires, le même nombre de CpG doit être considéré au départ. Ainsi, le nombre de CpG du jeu de données réelles a été ajusté à celui du jeu de données aléatoires, par sélection aléatoire des CpGs fiables. ***Abréviations*** : **nb min C.** : nombre minimum de CpGs nécessaire pour former une DMR ; **Nb C sélectionnées** : nombre de CpG prise en compte, selon le seuil de *pvalue* choisi ; **réelles ajust. ÷ aléatoires** : ratio du nombre de DMR détectée dans le jeu de données réelles ajustées et du nombre de DMR détectée dans le jeu de données aléatoires.

Ce système de *randomisation* nous parait pertinent car : (i) il permet de redistribuer les valeurs de méthylation observée, et ainsi éliminer la distinction entre groupe traité et non traité en apportant aléatoirement de la variabilité intra-groupe ; (ii) il attribue des valeurs de méthylation biologiquement plausibles, réellement observées pour une cytosine donnée. Si la *randomisation* avait été globale (*i.e.* récupération de toutes les valeurs de méthylation observées dans le vrai jeu de données, puis réattribution des valeurs, sans tenir compte de la position de la cytosine), il y aurait un fort risque d'attribuer des niveaux de méthylation à une cytosine donnée, sans que cela soit le reflet de la réalité biologique (*e.g.* attribution (non physiologique) de fort taux de méthylation de l'ADN, à une cytosine qui est toujours détectée comme non méthylée, quel que soit les conditions, traitement PBS ou EtOH).

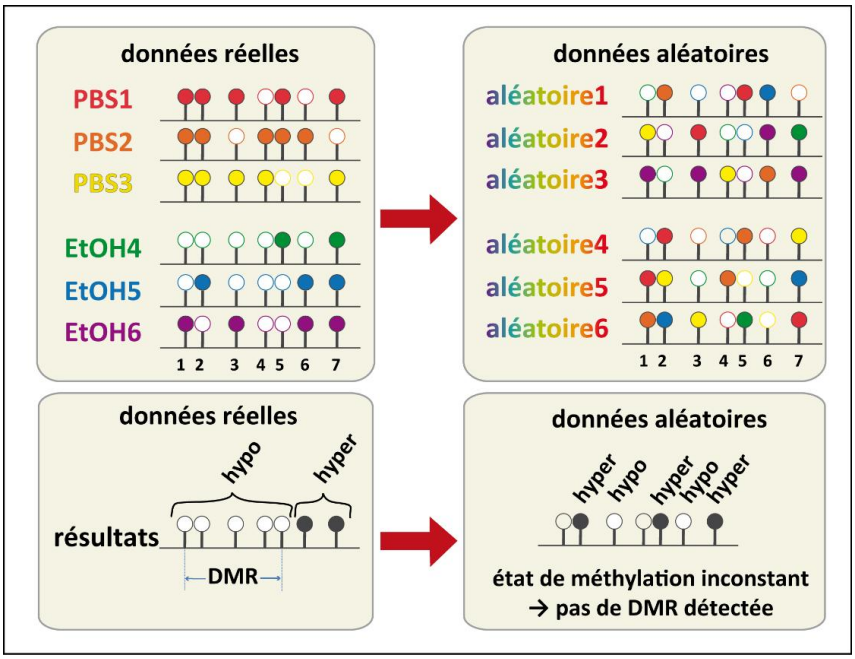


Figure 4-1.8 : Génération d'un jeu de données aléatoires permettant de définir les paramètres pertinents pour la détection des DMR.

Deux jeux de données aléatoires ont été générés à partir du jeu de données réelles (constitués de 3 échantillons contrôles, traités au PBS, et de 3 échantillons tests, traités à l'EtOH). La *randomisation* a été réalisée séparément, pour chaque cytosine donnée (mélange aléatoire du pourcentage de méthylation observé chez chaque échantillon réel, position par position), afin de redistribuer aléatoirement les valeurs sans attribuer des valeurs non observées à une position de CpG donnée.

Légendes du panel supérieur : **ronds pleins** : cytosines méthylées ; **ronds vides** : cytosines non méthylées.

Au total, deux jeux de données aléatoires ont été générés et comparés aux données réelles, pour s'assurer de l'absence de biais d'échantillonnage. Des résultats similaires ont été obtenus pour les comparaisons du jeu de données réelles avec ces deux jeux de données aléatoires.

2. Manuscrit Duchateau *et al.,* en préparation

Prenatal alcohol exposure leads to early disruption of DNA methylation in the developing mouse brain

Agathe DUCHATEAU [1,2,3], Federico MIOZZO [1,2,3*], Anne LE MOUEL [1,2], Myriame MOHAMED [1,2], Sascha OTT [4], Délara SABERAN-DJONEIDI[1,2,$], Valérie MEZGER [1,2,$,&]

[1] Université de Paris, Epigenetics and Cell Fate, CNRS, F-75013 Paris, France

[2] Département Hospitalo-Universitaire DHU PROTECT, Paris, France.

[3] ED 562 BioSPC, Université Paris Diderot Paris 7, F-75205 Paris Cedex 13, France

[4] Department of Computer Science, University of Warwick, Coventry, CV4 7AL, United Kingdom

[$] Co-last authors

[&] corresponding author: valerie.mezger@u-paris.fr

[*] Present address: Department of Genetics and Evolution, Sciences III, University of Geneva, Geneva, Switzerland

Funding information

VM was funded by the Agence Nationale de la Recherche « HSF-EPISAME », SAMENTA ANR-13-SAMA-0008-01) and IREB (Institut de recherches scientifiques sur les boissons (2015-2016). AD was funded by a Doctoral Fellowship from the French Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI). This study contributes to the Université de Paris IdEx #ANR-18-IDEX-0001 funded by the French Government through its "Investments for the Future" program. DSD benefited from a CNRS Délégation de Recherche (2018-2020). OTT benefited travel grants from Université Paris Diderot.

Acknowledgements

We thank Olivier KIRSH (Université de Paris, CNRS Epigenetics and Cell Fate, Paris, France) for sharing his expertise in bioinformatics and biostatistics. We are grateful to Slimane AIT-SI-ALI, (Université de Paris, CNRS Epigenetics and Cell Fate, Paris, France) Maxim GREENBERG and Deborah BOURC'HIS (Institut Curie, Paris) for helpful discussions. We thank Renaud MASSART, former postdoc in the VM Lab, for designing methylome capture. We are grateful to Emeline MUNDWILLER and Yann MARIE (Brain and Spine Institute, Paris, France) for performing the methylome capture libraries and high-throughput sequencing of these samples. We thank Isabelle Le PARCO and the staff from the Buffon animal housing facility at the Jacques Monod Institute (Université de Paris, Paris, France). We are grateful to the VM team members, Aurélie de THONEL and Véronique DUBREUIL (Université de Paris, CNRS Epigenetics and Cell Fate, Paris, France) for helpful discussions and comments on the manuscript.

Abstract

The reshaping of the DNA methylome landscape after prenatal alcohol exposure (PAE) has been well-documented in the adult brain, therefore a long time after the end of the exposure. However, to our knowledge, the question of the early deposition or loss of DNA methylation marks in the prenatal neocortex, just after the end of PAE has not yet been directly addressed. This is of importance for the future identification of biomarkers that, ideally, should be linked to PAE itself (and not to altered neuronal activity, established after PAE). Using a binge-drinking-like protocol of PAE and capture of the DNA methylome, we have identified differentially methylated regions (DMRs) that are established within two hours after the end of PAE. These DMRs are prominently and statistically associated with enhancers that are active in the brain, and remarkably concern genes that, in physiological conditions (no PAE) show dynamic gain in chromatin accessibility or upregulation of their expression in the time-window of exposure. These DMRs, which are associated with GO terms of importance for neurogenesis, neurodevelopment and neuronal differentiation, are thus very susceptible to affect gene expression in a deleterious manner after PAE. Remarkably, among these DMRs, two groups of associated genes are over-represented: imprinted genes and clustered protocadherin genes. DMRs in these two gene families have been previously identified, both in the adult rodent brain that had been prior-exposed to alcohol prenatally and in buccal swabs of children diagnosed with a fetal alcohol syndrome (FASD). Our study therefore strongly suggests that the DNA methylation profile of key regulatory regions of these two gene groups are very quickly disturbed after the end of PAE and that these early altered regions could potentially stay affected long after the stress, which strongly reinforces their potential as future biomarkers of PAE.

Ethical issues

The breeding and treatment of wild type C57BL/6N mice, used for the experimental protocols described in this study have been approved by the Institutional Animal Care and Use Ethical Committee of the Paris University (registration number CEEA-40). The project has been recorded under the following reference by the Ministère de l'Enseignement Supérieur et de la Recherche (#2016040414515579). All efforts were made to reduce stress and pain to animals.

Keywords

Fetal alcohol syndrome; early alcohol-induced modifications in DNA methylation; differentially methylated regions; neurodevelopment; epigenetics; methylome capture

INTRODUCTION

During its prenatal development, the brain is particularly vulnerable to detrimental events that generate neurodevelopmental defects and potentially have long-term consequences in the adulthood (Bale et al., 2010; Schang et al., 2018). Among a diversity of adverse *in utero* stresses, prenatal alcohol exposure (PAE) is a leading cause of non-genetic mental retardation in the Western world (Popova et al., 2012, 2016). Depending on many parameters, such as the timing of exposure, the drinking pattern (chronic or acute) and the amount of alcohol consumed, PAE gives rise to a wide range of neurodevelopmental defects, referred to as fetal alcohol spectrum disorders (FASD), whose prevalence is estimated around 9 for 1000 live births (Burd et al., 2003; Jones and Smith, 1973; Kleiber et al., 2013; Lemoine et al., 1968; Mattson et al., 2011b; Popova et al., 2012, 2016). These defects lead to impairment in cognition, behaviour, executive function, attention (linked or not to hyperactivity), learning, judgment and social adaptation (Gibbard et al., 2003). In addition to these primary defects, FASD individuals are at high risk for neurospychiatric disorders, including anxiety disorders, depression, and addiction in their adulthood (Gibbard et al., 2003; reviewed in Kodituwakku, 2007 and O'Connor and Paley, 2009, in line with the DOHaD concept (Developmental Origins of Health and Disease; Markham and Koenig, 2011; Schlotz et al., 2014; Thompson et al., 2009)). The most severe form of FASD, called fetal alcohol syndrome (FAS), is defined by characteristic facial dysmorphology, growth retardation and severe macroscopic structural abnormalities of the central nervous system (CNS; Jones and Smith, 1973). Early diagnosis of PAE is necessary to enable efficient intervention (Burd et al., 2003). Visible and well-characterized phenotypic defects observed for FAS newborns facilitated the early diagnosis of these individuals. In contrast, diagnosis of FASD children is more challenging, because the history of exposure is often unknown and phenotypic are more cryptic, but still crucial since their central nervous system damages can occur to the same extent than those of FAS children (Gibbard et al., 2003). Indeed, PAE affects neurodevelopment at any stage, therefore impacting all neurodevelopmental processes, such as proliferation of neural progenitors, migration of young post-mitotic neurons and their differentiation, neuronal survival, synaptogenesis, neurotransmission and neuronal plasticity (El Fatimy et al., 2014; Guerri et al., 2009; Hashimoto-Torii et al., 2014; Ishii et al., 2017).

Although defects caused by prenatal alcohol exposure are well identified, the exact molecular mechanisms underlying these alterations and their persistence are still unclear. However, PAE is known to long-lasting impact on gene expression levels. These transcriptional disturbances, observed in the adult brain, occur in a manner dependent on the developmental stage at which the exposure has occurred (Kleiber et al., 2013, 2014). Indeed, transcriptional changes affect genes that fall into distinct GO (Gene Ontology) categories. In particular: 1) cellular organization, cellular development, cell death and proliferation are the most prominent processes affected when the alcohol exposure occurred during the equivalent of the first trimester of pregnancy in humans. PAE at this developmental stage may result in deficits related to quantity of neurons, abnormal axon development, and consequences in synaptic function; 2) Cellular movement, cell death, cell morphology, cell migration and differentiation are mostly affected when PAE occurred during the equivalent of trimester two in humans. 3) When *in utero* alcohol exposure happened during the equivalent of the third trimester in humans, cellular communication and neurotransmission are the most impacted categories in term of gene expression levels (Kleiber et al., 2013, 2014). These long-lasting disturbances therefore affect processes governed by gene networks that are active during the

gestation, at the time of exposure, suggesting a "memory" of the exposure. Epigenetic mechanisms could participate to this "memory".

Epigenetic mechanisms are responsible for heritable changes, in gene function or expression throughout cell divisions, which occur without modifying the DNA sequence, thereby allowing cells or organisms to adapt at diverse environmental cases (Probst et al., 2009). They are essential for gene regulation at multiple levels. On one hand, gene transcription is controlled by small non coding RNAs (ncRNA), DNA methylation and histones modifications (Carthew and Sontheimer, 2009; Probst et al., 2009). On the other hand, ncRNA carry out regulation at post-transcriptional level, in addition to their activity in transcriptional regulation (Carthew and Sontheimer, 2009; Taft et al., 2009). Covalent modification of DNA, including methylation (5mC) and hydroxymethylation (5hmC), involves the recruitment of a methyl group to the 5' position of the cytosine (Penn et al., 1972; Zemach et al., 2010). It is commonly admitted that DNA methyltransferase 1 (DNMT1) is responsible of the maintenance of DNA methylation upon DNA replication (Bestor et al., 1988), whereas DNMT3A and DNMT3B are involved in *de novo* DNA methylation (Okano et al., 1998), but all DNMTs can participate in both processes (reviewed in Jeltsch and Jurkowska, 2014). In mammals, DNA methylation has been mostly studied at cytosines in the CpG-dinucleotide context, but the postnatal brain exhibits also DNA methylation in non-CpG contexts, with differential impacts on gene expression (Lister et al., 2013; Schultz et al., 2015). When located in a promoter (*i.e.* proximal regulatory region), DNA methylation, in a CpG context, is generally associated with gene repression, whereas it is considered as a repressive or active mark in other contexts (*e.g.* gene bodies or intergenic regions), depending on its genomic localization (Portales-Casamar et al., 2016).

Remarkably, brain development is tightly controlled by epigenetic mechanisms, as underlined by the impact of mutations or variants in genes encoding epigenetic actors in the emergence of neurodevelopmental and neuropsychiatric disorders (Rett syndrome, Rubinstein-Taybi syndrome, autism spectrum disorders, etc. ; (Bourgeron, 2015; Gräff et al., 2011; LaSalle et al., 2013). The long-lasting and stage-specific transcriptomic alterations, described above and observed in the PAE-exposed brain, might be underlined by modifications in the epigenetic landscape, including perturbation of DNA methylation profile. Indeed, starting from seminal works on rodent models (*e.g.* Haycock and Ramsay, 2009; Kaminen-Ahola et al., 2010), a number of studies have identified disturbances in DNA methylation on candidate genes or *loci*, or in a genome-wide manner, in the mouse adult brain that was exposed to alcohol prenatally (Kleiber et al., 2013; Laufer et al., 2013; reviewed in Lussier et al., 2017). Alterations in DNA methylation profile observed in mouse brain in response to PAE has been corroborated in peripheral tissues (cheek swabs) in cohorts of FASD children (Laufer et al., 2015; Lussier et al., 2018).

Taken together, these findings highlight the long-lasting consequences of PAE on DNA methylation patterns and gene expression, in the postnatal and adult brain of rodent models and human cohorts. Today, FASD cohorts are still of a relatively modest size, and might present confounding factors in terms of genetic architecture, age, ethnicity, and prenatal alcohol exposures patterns. Interestingly, homologous *loci* affected by PAE at DNA methylation levels are found in rodent models and FASD children (Laufer et al., 2015). Even with these observations, there is a lack of data to fully understand whether these aberrant DNA methylation events are the direct (early) or indirect (late) results of alcohol exposure, since neuronal activity can reshape DNA methylation

throughout lifetime (Su et al., 2017). For this reason, DNA methylation disturbances observed in the adulthood could (**i**) result from early PAE-dependent DNA methylation changes that persist throughout life or (**ii**) be due to PAE-induced brain dysfunction, such as an altered neuronal activity, and appear in the postnatal or adult brain, at temporal distance from PAE.

Strikingly, the short-term impact of PAE on the architecture of DNA methylation in the developing brain has been less studied and there is thus a need to investigate early DNA methylation changes in response to PAE. This is of importance for the field, because, in many cases, the history of exposure is unknown and FASD children are often diagnosed late, which compromised early intervention. The search for accurate and relevant molecular biomarkers, especially for biomarkers of exposure, is thus necessary. Attempts to identify biomarkers of exposure linked to DNA methylation perturbations have been given special attention in the field. Ideally, such DNA methylation changes should occur early after exposure and persist later in life. This remains to be determined.

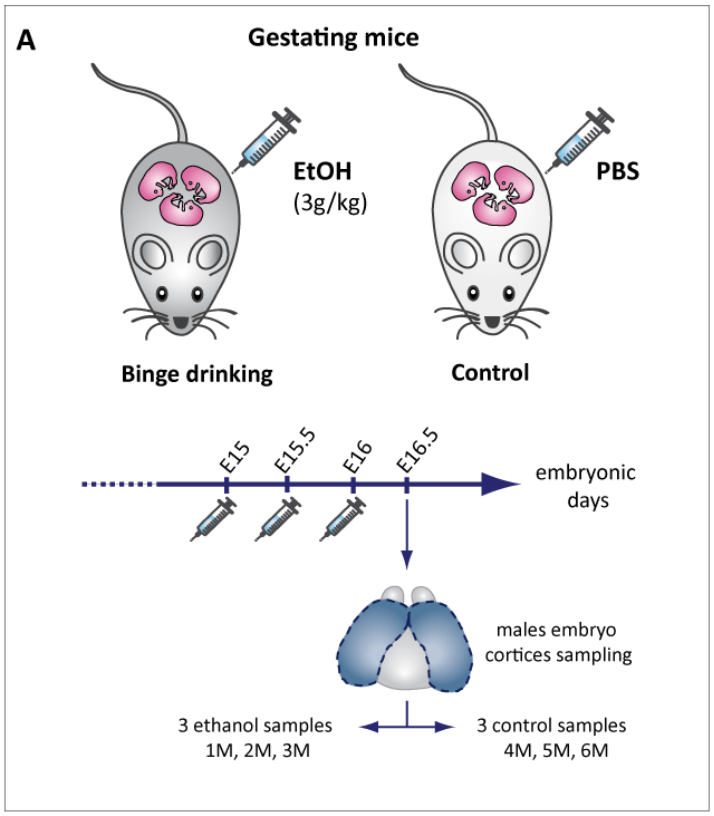
Using a binge-drinking mouse model of PAE, we asked whether early changes in DNA methylation across the genome could be detected in the developing cerebral cortex. Using a tailor-made methylome capture approach, we observed hundreds of differentially methylated regions (DMRs), as early as two hours after the last alcohol injection. Some of these alterations are associated to genes or *loci* which, in physiological conditions, are dynamically regulated at the time of exposure in terms of chromatin accessibility or expression, as determined by our analyses of available time-course ATAC-seq and RNA-seq ENCODE datasets. Early after PAE, we could identify significative DMRs that are linked to biological processes of importance for brain development and functions. Our results therefore show that some genomic regions are the target of differential methylation events, very rapidly after exposure, that seem to persist, according to published results, in the postnatal and adult brain, suggesting that these *loci* might constitute valuable and meaningful biomarkers of PAE in the future.

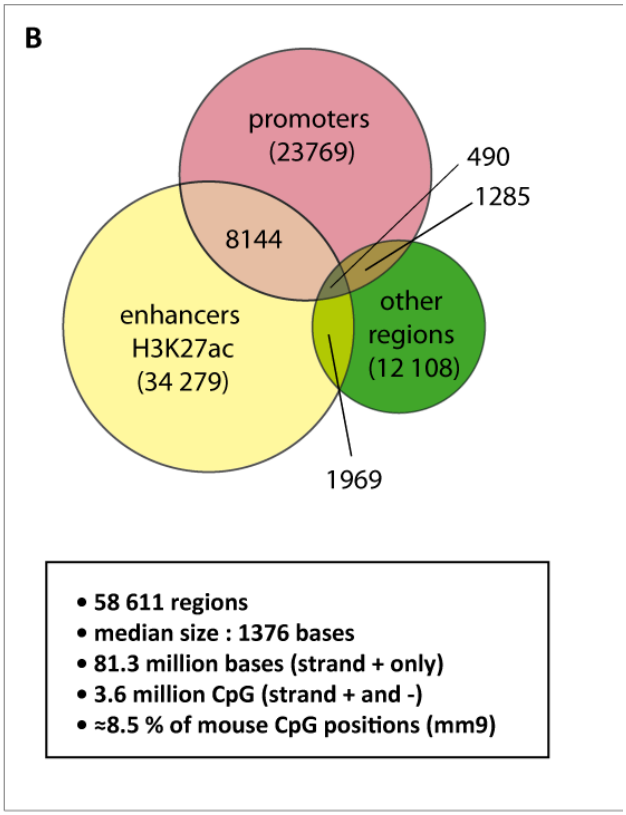
RESULTS

A methylome capture in a murine binge drinking model

In this study, we aim at identifying DNA methylation changes occurring early after an acute prenatal alcohol exposure. For this, we used a murine model of binge drinking at a developmental stage equivalent to the second trimester of pregnancy in human: mouse embryonic cortices were exposed to PAE *in utero* at embryonic days E15 and E16 (Fig. 1A, see Material and Methods).

We chose this binge-drinking model in order to be as close as possible to a human mode of acute consumption (in contrast to a typical mode of consumption for chronic alcoholism). The quantity of injected ethanol corresponded to a dose known to induce brain defects in rodent pups that mimicks FAS (Carloni et al., 2004; Ikonomidou et al., 2000; Olney et al., 2002b). We analysed early methylation events after PAE using a capture of the DNA methylome (EpiCapture technology), which allowed us to study, with a good resolution, 58,611 DNA regions that we have chosen, corresponding to about 81.3 Mb. In particular, the capture design was composed of more than 75% of mouse promoters and all active enhancers in the brain (carrying histone H3 acetylation at the lysine residue K27, Fig. 1B, Supp. Fig. S1A); see **Materials and Methods** for details).


A
Gestating mice
EtOH
(3g/kg)
PBS
Binge drinking
Control
E15
E15.5
E16
E16.5
embryonic
days
males embryo
cortices sampling
3 ethanol samples
1M, 2M, 3M
3 control samples
4M, 5M, 6M


B
promoters
(23769)
490
1285
8144
enhancers
H3K27ac
(34 279)
other
regions
(12 108)
1969
• 58 611 regions
• median size : 1376 bases
• 81.3 million bases (strand + only)
• 3.6 million CpG (strand + and -)
• ≈8.5 % of mouse CpG positions (mm9)

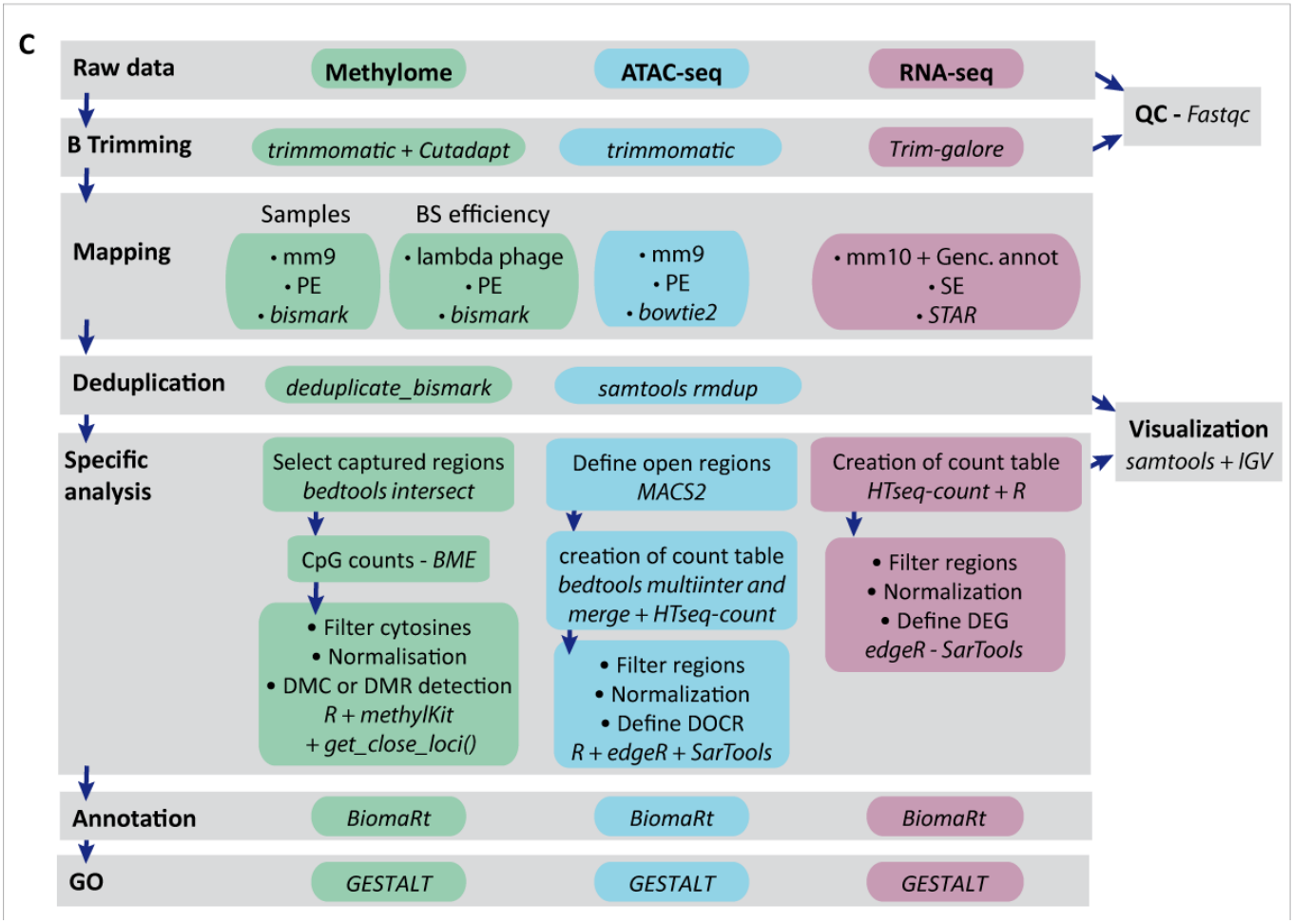

C
Raw data
Methylome
ATAC-seq
RNA-seq
QC - Fastqc
B Trimming
trimmomatic + Cutadapt
trimmomatic
Trim-galore
Mapping
Samples
BS efficiency
• mm9
• PE
• bismark
• lambda phage
• PE
• bismark
• mm9
• PE
• bowtie2
• mm10 + Genc. annot
• SE
• STAR
Deduplication
deduplicate_bismark
samtools rmdup
Visualization
samtools + IGV
Specific
analysis
Select captured regions
bedtools intersect
Define open regions
MACS2
Creation of count table
HTseq-count + R
CpG counts - BME
creation of count table
bedtools multiinter and
merge + HTseq-count
• Filter regions
• Normalization
• Define DEG
edgeR - SarTools
• Filter cytosines
• Normalisation
• DMC or DMR detection
R + methylKit
+ get_close_loci()
• Filter regions
• Normalization
• Define DOCR
R + edgeR + SarTools
Annotation
BiomaRt
BiomaRt
BiomaRt
GO
GESTALT
GESTALT
GESTALT

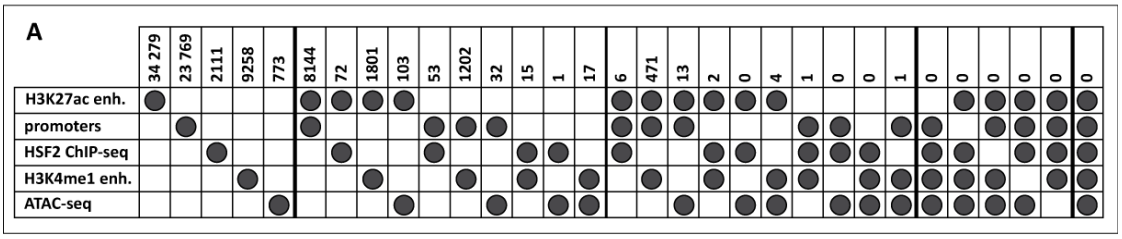
Figure 1 : Binge drinking model and bioinformatic workflows used in this study.

(A) Binge drinking model. Gestating C57BL/6N mice received 3 intra peritoneal injections (IP) of ethanol (3g/kg) between embryonic days E15 and E16. Control is treated similarly with PBS. Embryonic cortices were collected 2 hours after the last IP. Only male samples were used. Three samples per group (called 4M, 5M or 6M for control group and 1M, 2M and 3M for EtOH-treated one) were generated. To reduce variability, each sample is composed of 4 hemi-cortices, from 4 embryos of distinct litters (Supp. Fig. S1F).

(B) Methylome capture composition and key figures. Capture is composed of 58,611 chosen regions, based on ENCODE available data (for enhancers and promoters regions) or based on previous lab results (*other regions*). Most of captured regions correspond to active enhancers in adult (8 weeks-old) mouse cortex (characterized by the H3K27ac histone mark). Capture also includes more than 75% of mouse promoters regions. "Other regions" group corresponds to (**i**) HSF2 binding sites found in unstressed mice cortices of embryos at E16.5 development stage ; and to dynamic regions upon stress : (**ii**) differentially opened or closed regions identified in isolated oligodendrocyte precursors (O4+ cells) from cortices of 5 days-old mice, after a five days inflammatory stress during a period equivalent to the third trimester of pregnancy in human (results of ATAC-seq experiment); (**iii**) enhancers (characterized by the H3K4me1 histone mark) of adult (8 weeks-old) mouse cortex among genes differentially expressed in microglia (CD11B+ cells), at different stages (P1 to P45) upon an inflammatory stress (P1-P5) during a period equivalent to the third trimester of pregnancy in human (results of a microarray experiment). For details see Materials and methods.

(C) Bioinformatic workflow of methylome capture, ATAC-seq and RNA-seq analyses. Methylome capture were performed on control (PBS-treated) and EtOH-treated group (described in Fig. 1A). ATAC-seq and RNA-seq analyses were performed on available ENCODE dataset, obtained from mice forebrains of distinct embryonic developmental stages and newborn samples not exposed to any stress. Key steps and tools used for each bioinformatic analysis are described here. For more details, please see Material & Method and detailed command lines in Supp. Jupyter notebooks n°1 to 4[8].

Abreviations: ***QC:*** quality control ; ***PE :*** paired end analysis ; ***SE:*** single end analysis ; ***Genc. annot.:*** Gencode annotation ***; BME:*** bismark_methylation_extractor ; ***DMC***: differentially methylated isolated CpG observed upon alcohol exposure; ***DMR*** : differentially methylated region observed upon alcohol exposure ; ***DOCR***: differentially opened or closed regions upon physiological brain development stages ; ***DEG***: differentially expressed genes upon physiological brain development stages.

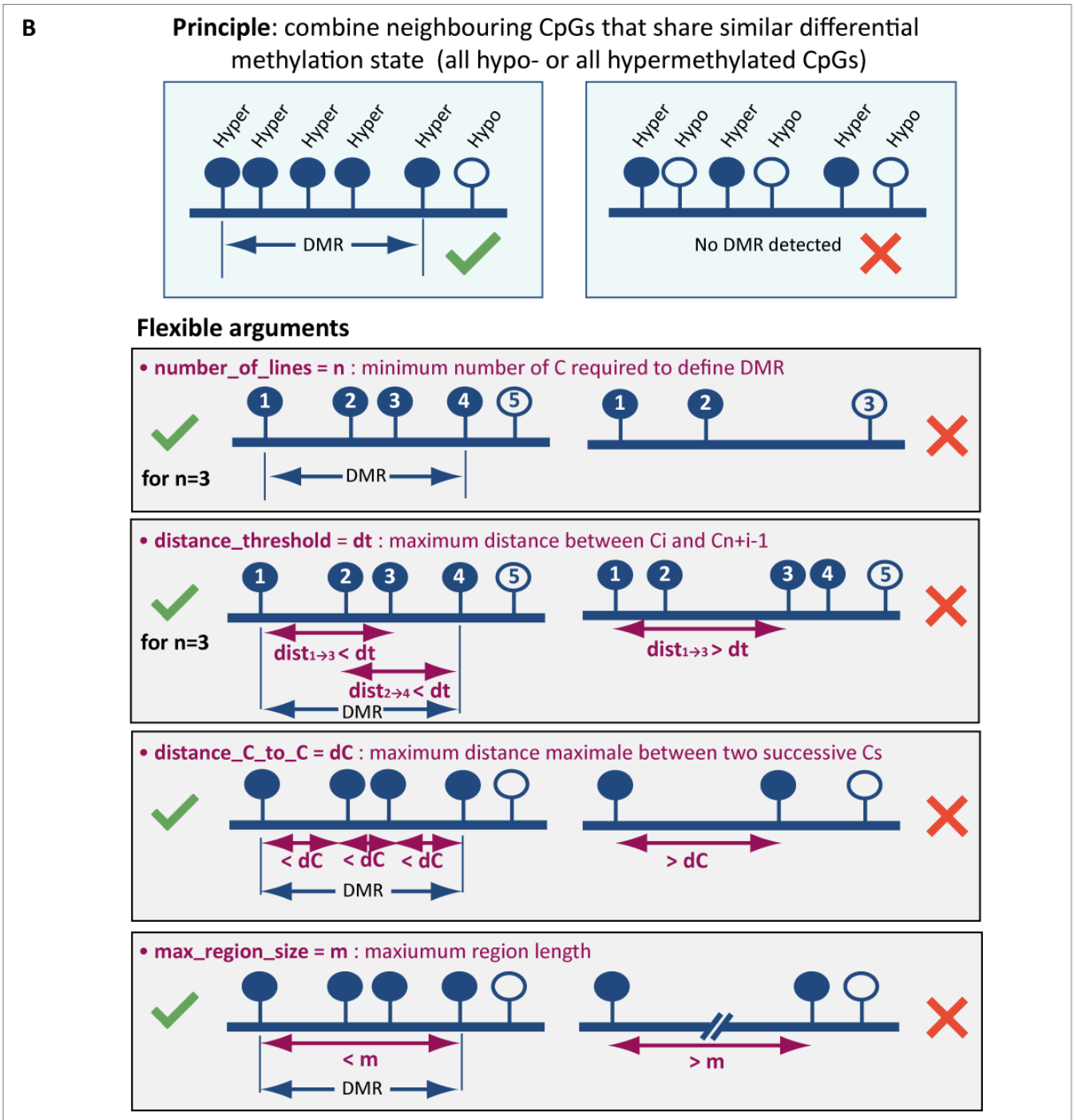


Supplementary Figure S1A: Number of regions in each methylome capture category.

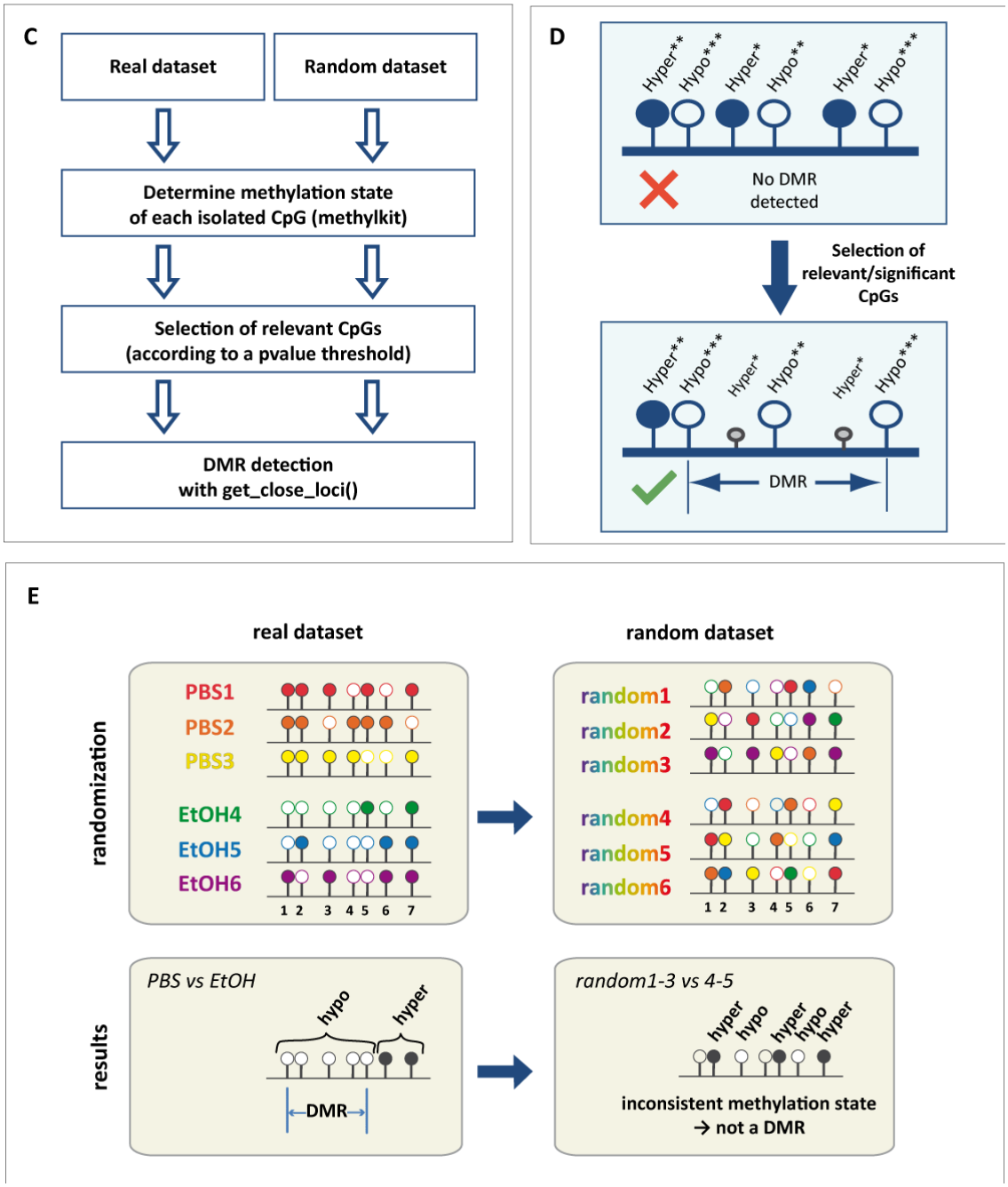
HSF2 ChIP-seq, H3K4me1 enh. and ATAC-seq items correspond to sub-categories of "other" regions shown in Figure 1B**.** Note that categories are not exclusive (*i.e.* some regions are counted several times, *e.g.* 8,144 regions that are included both in H3K27ac enhancers and promoters categories, are also among the 34,279 regions that are included in H3K27ac enhancers alone). Darkgrey dots indicated dataset(s) that are used for the counting.

Abreviations. H3K27ac enh.: active enhancers of adult mouse cortex, characterized by H3K27ac histone mark that; ***HSF2 ChIP-seq*** **:** HSF2 binding sites found in unstressed mice cortices of embryos at E16.5 development stage ; ***H3K4me1 enh***. : enhancers (characterized by the H3K4me1 histone mark) of adult (8 weeks-old) mouse cortex among genes that were differentially expressed in microglia (CD11B+ cells), at different stages (*i.e.* postnatal days 1 (P1) to P45) upon an inflammatory stress (from P1 to P5) during a period equivalent to the third trimester of pregnancy in human (results of a microarray experiment) ; ***ATAC-seq*** : differentially opened or closed regions identified in isolated oligodendrocyte precursor cells (O4+ cells) from cortices of 5 days-old mice, after an inflammatory stress during a period equivalent to the third trimester of pregnancy in human. For details see Materials and methods.

[8] *Annexe 6-3.1 à Annexe 6-3.4 : Jupyter notebooks*



Supplementary Figure S1B: Principle and parameters of get_close_loci() function that we specifically generated for the detection of DMRs into dataset from tailor-made capture. Several parameters can be modulated for DMRs detection, depending on the type of data and the biological question asked. Filled circles represent hypermethylated cytosines observed between two datasets (*e.g.* control group *versus* test group), whereas empty circles represent hypomethylated cytosines.

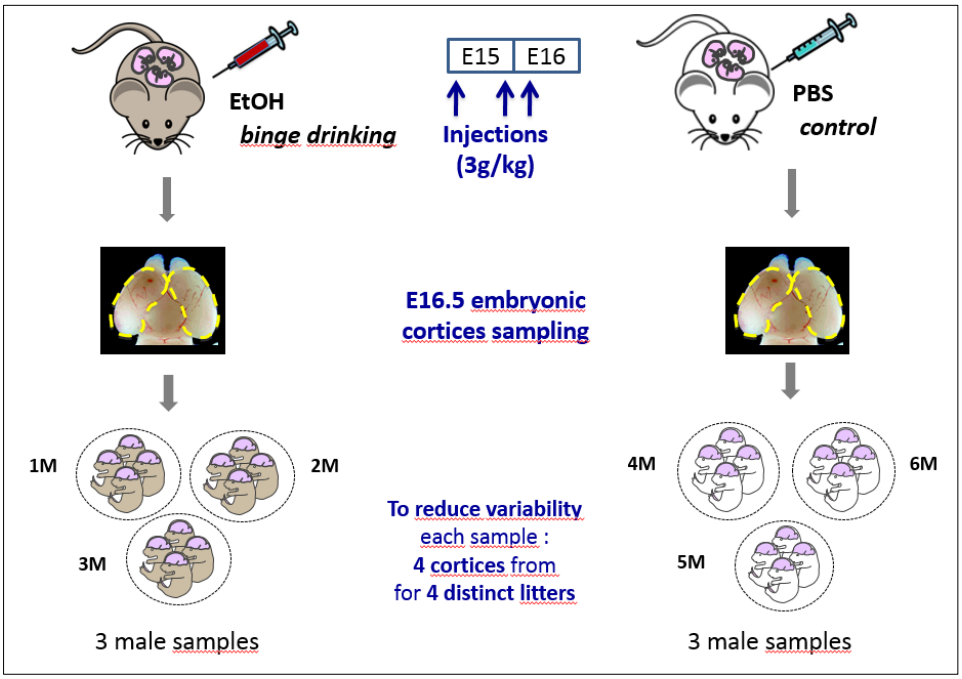


Supplementary Figures S1C-S1E: Approach used to detect DMRs

(C) Main steps of the approach performed to detect relevant DMRs in the real dataset from tailor-made methylome capture. **(D) Selection of relevant CpGs is essential for the detection of DMRs**. DMRs detection could be compromised if CpGs with no reliable methylation state are kept. In this example, the conservation of the two unreliably defined hypermethylated CpGs during DMRs detection process, prevents the detection of hypomethylated DMR, although it exists. Therefore, pvalue threshold must be defined to identify relevant methylation information. Filled circles represent hypermethylated cytosines observed between two datasets (*e.g.* control group *versus* test group), whereas empty circles represent hypomethylated cytosines. *: non significant pvalue; *** significant pvalue. **(E) Creation of a random dataset to define relevant parameters for DMRs detection.** Two random data sets were generated using real dataset (composed of 3 PBS-treated control samples, and 3 EtOH-treated ones). In order to randomly redistribute the values without assigning unobserved values to a given CpG position, randomization was performed separately for each cytosine site (position by position, random shuffling of the methylation percent (percent of the ratio of number of methylated cytosines divided by the total number of at a given position) observed in real samples). For the upper panel, filled circles represent methylated cytosines, whereas empty circles represent unmethylated cytosines.

Definition of DNA methylated regions for capture analysis

This methylome capture approach used sodium bisulfite (BS) conversion, to distinguish between methylated and unmethylated cytosines. Depending on the experimental conditions, the sodium bisulfite conversion may not be complete (Clark et al., 2006), leaving some unmethylated cytosines unchanged, while they should be modified into thymines (Krueger et al., 2012). Thus, we investigated whether the BS conversion correctly occurred for all samples. Abnormal *per-base-sequence-content* plot obtained with FASTQC was observed for each sample, as expected for DNA sequences converted with BS (Supp. Fig. S2A) Indeed, conversion of unmethylated cytosines into thymines leads to an over-representation of thymines and an underrepresentation of cytosines at a given position, Supp. Fig. S2A). Moreover, BS conversion rate of all samples was quantified, using spike-in DNA. Conversion rate is high for all samples, which allows the analysis of the data without conversion biases (Supp. Table 1).



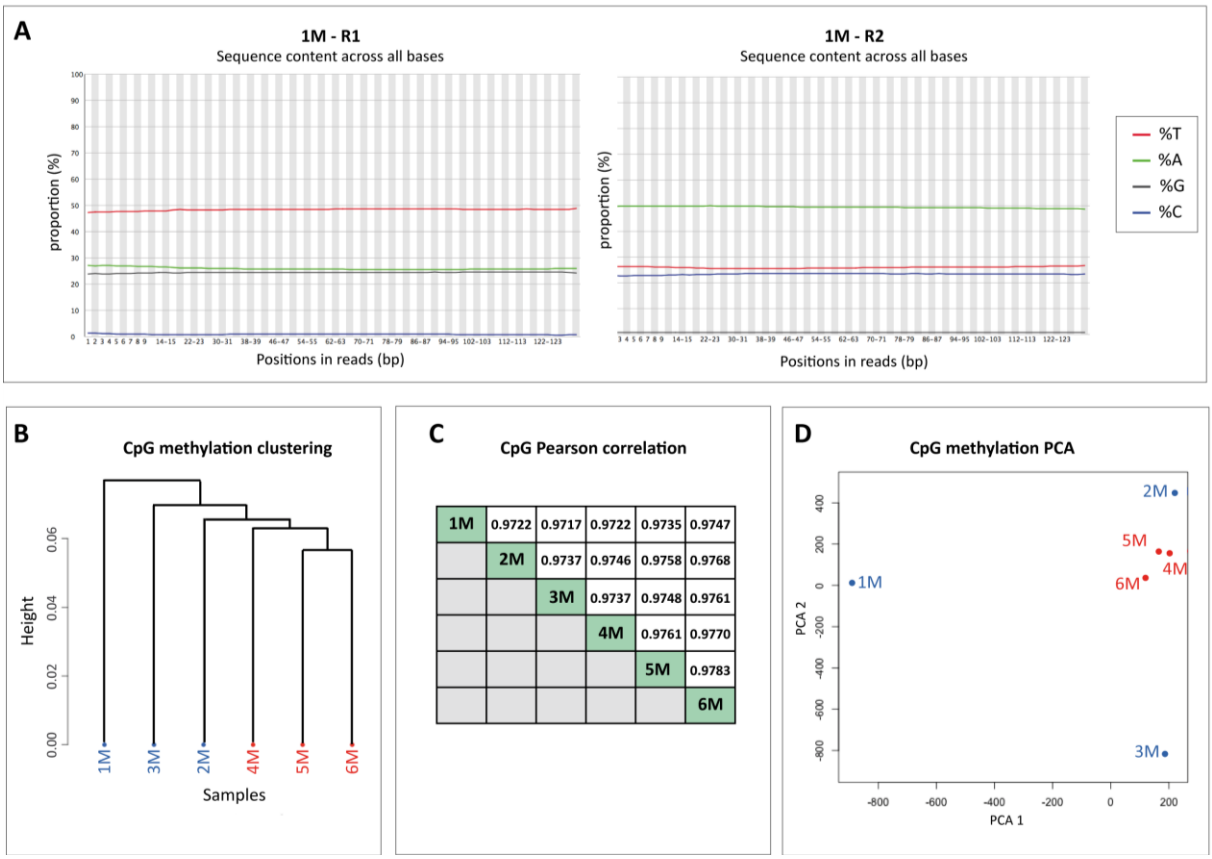
Supplementary Figure S1F: Sampling protocol.
To minimize intra-group variability, each NGS sample was composed of four hemi-cortices of four different male embryos from four different litters.

In line with our sampling choice which aimed at mimimizing inter-individual variability in the same group, by mixing half-cortices from 4 embryos from 4 different litters (Supp. Fig. S1F), we observed that intra-group variability is limited, according to hierarchical clustering analysis and Pearson correlation between samples (Supp. Fig. S2B and S2C). However, triplicates from the EtOH-treated group showed more differences between each other than do triplicates from control group (PCA analysis and hierarchical clustering analysis, Supp. Fig. S2B and S2D), suggesting that, even if samples are globally affected in the same manner by *in utero* alcohol exposure, some subtle differences in term of CpG methylation level could be observed. In addition, CpG base Pearson correlation values indicated that inter-group variability is also limited. This suggests that PAE does not majorly reshape the DNA methylome (Supp Fig. S2C).

Supp. Table S1: Bisulfite conversion efficiency

		1M	2M	3M	4M	5M	6M
deduplication	Total number of mapped reads	22 055	63 017	52 100	62 391	80 753	86 277
	Total count of deduplicated leftover sequences (% of total)	79,16	45,89	54,89	54,25	49,39	59,75
bismark methylation extractor	**Total number of methylation call strings (=R1+ R2 reads)**	**34918**	**57842**	**57196**	**67688**	**79770**	**103108**
	Total number of C's analysed:	**480 679**	**1010776**	**976435**	**1176089**	**1458199**	**2041451**
	Total methylated C's in CpG context:	**1307**	**1873**	**1916**	**2102**	**2369**	**2751**
	Total methylated C's in CHG context:	1671	2560	2625	2804	3147	3872
	Total methylated C's in CHH context:	7617	8635	9137	9188	9601	9803
	Total C to T conversions in CpG context:	136744	297737	285086	346866	429779	606148
	Total C to T conversions in CHG context:	146887	315798	302944	368922	458882	646756
	Total C to T conversions in CHH context:	186453	384173	374727	446207	554421	772121
	C methylated in CpG context:	**0,95**	**0,63**	**0,67**	**0,60**	**0,55**	**0,45**
	C methylated in CHG context:	1,12	0,80	0,86	0,75	0,68	0,60
	C methylated in CHH context:	3,92	2,20	2,38	2,02	1,70	1,25
	Total methylation (i.e. total C non converted by BS) %	2,20	1,29	1,40	1,20	1,04	0,80
	BS conversion efficiency (%)	**97,80**	**98,71**	**98,60**	**98,80**	**98,96**	**99,20**

Few sequences of Lambda phage virus were added to each sample during the sequencing. These sequences are used as DNA-spike in, which allows estimating bisulfite sodium conversion (BS), since these sequences are completely unmethylated before the conversion. Reads that mapped to lambda phage reference genome were deduplicated before the estimation of bisulfite sodium conversion (BS). The high percent observed for the bisulfite (BS) conversion efficiency for all samples allows us to analyse the data.



Supplementary Fig. 2 : Quality controls and correlation between methylome samples

(A) ***Per-base-sequence-content plots*** **obtained with FASTQC, for 1M sample.** An imbalance is observed between bases proportions at each read position. This result is expected, as DNA sequences were converted with sodium bisulfite (BS). The conversion of unmethylated cytosines into thymines by this BS treatment leads to an over-representation of thymines and an underrepresentation of cytosines at a given position for R1 reads (for R2 reads, which are complementary to R1 ones, an over-representation of adenine and an underrepresentation of guanine are logically observed). These plots were obtained after the trimming of the data.

Analyses represented on figures **(B), (C) & (D)** were done using *MethylKit* (default parameters)**.** CpGs were filtered (CpGs covered by less than 10 reads or having coverage > 99.9th percentile were excluded) and coverage was normalized (median method). Only information from CpG positions that are in all samples are kept (1,259,111 CpGs).

(B) CpG methylation Hierarchical Cluster Analysis. This dendrogram was obtained using of ward method with correlation as distance measure.

(C) Correlation matrix. This matrix was obtained using Pearson correlation.

(D) CpG methylation Principal Components Analysis (PCA).

Using the capture approach (Fig. 1B), we only observed three hyper- and one hypomethylated isolated (individual) CpGs when control and alcohol-treated groups were compared (*MethylKit* analysis on normalised and filter datasets, meth. diff < 5%, qval. = 0.05; Table 1), all having a high methylation differential percent (> 35%). Two of these differentially methylated cytosines (DMC) are located in intergenic regions (on chromosomes 5 and 10). One hypermethylated CpG is in the *Kalrn* (*Kalirin RhoGEF Kinase*) gene, encoding a kinase that is involved in various mechanisms, such as neuronal shape regulation, growth and plasticity. The third hypermethylated CpG is associated to *Tiam2* gene, encoding a RAC1-specific guanine nucleotide exchange factor.

Table 1: Only few CpG sites are identified as differentially methylated (DMC) upon the binge drinking stress.

chr	cytosine position	pvalue	qvalue	Meth.diff	Genomic *loci*
chr5	114101310	5.59e-08	0.033	-47.7193	Intergenic region
chr10	117378233	1.64e-07	0.048	37.37742	Intergenic region
chr16	34407296	1.29e-07	0.048	35.69702	*Kalrn*
chr17	3397110	1.45e-08	0.017	57.43663	*Tiam2*

Only cytosine sites, in a CpG context, covered in all samples were investigated. DMC were identified using *MethylKit* (see details in Material & Method). ***Meth.diff***: differential methylation (%) observed between control (PBS-treated) and EtOH-treated groups. If percent is negative, it means that less methylation is observed upon binge drinking stress compared to control. qval. threshold = 0.05, minimum Meth.diff to detect a DMC = 5%

Although the methylation of an individual CpG can affect gene expression (Xu et al., 2007), DMRs are considered as more relevant in terms of impact on gene expression. In line with this consideration, studies of differentially methylated regions (DMRs) have been more frequently performed (Bock, 2012). We therefore pursued our analyses on the identification of DMRs.

Bioinformatic analysis of this methylome capture requires special attention, especially to define potential differentially methylated regions (DMRs) between control and alcohol-treated groups. Indeed, existing tools often considered the whole genome (or well-characterised arrays) as a reference to define genomic regions. This kind of approach is not adapted to a customized capture based on a repertoire of selected regions. For this reason, we developed our own *R* function to define DMRs between control and alcohol-treated groups, which combined neighbouring CpGs that share similar differential methylation states, (all hypo- or all hypermethylated CpGs; see Material & Methods; Suppl. Fig. S1B-D). Bioinformatic generation of random datasets was used to define relevant criteria for DMRs detection (Supp. Fig. S1E, Supp. Table 2).

Early alterations in DNA methylation are detected in brain development upon PAE.

Using this captured-specific bioinformatic workflow (Fig. 1C, Supp. Table S3, details in **Material & Methods** and in Supp. Data Jupyter notebook n°1[9]), we identified 432 regions among the 58,611 regions included in the capture, that were differentially methylated in embryonic cortices, early after PAE (Fig. 2A, Dataset DMRs[10]). Among these DMRs, hypermethylated DMRs (257 “hyper-DMRs”) were predominant, compared to hypomethylated ones (175 “hypo-DMRs”; Fig. 2A). Notably, this result depends on the original capture composition and thus does not necessarily reflect a genome-wide phenomenon. In either hyper- or hypo-DMRs, the median differential levels of DNA methylation in the majority of DMRs reached more than 10% (Fig. 2A).

[9] Annexe 6-3.1 : Jupyter notebook - capture methylome workflow
[10] Annexe 6-4.1 : Tableaux de données - thèse - A. Duchateau - tab DMR432

Suppl. Table 2: Comparison results between real and random datasets, and parameters finally chosen for DMRs detection using *get_close_loci()* function.

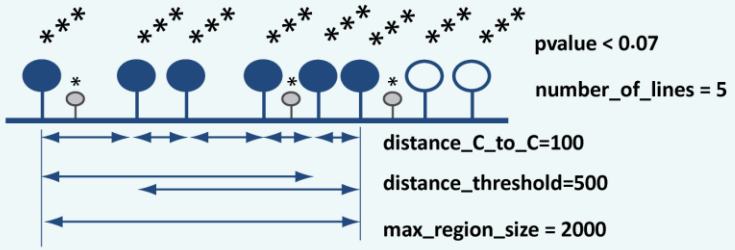
with first random dataset

pvalue thresh.	Min nb of C	Nb of selected C: real data	Nb of selected C: random data	Nb ofdetected DMR: real data	Nb ofdetected DMR: adjusted real data	Nb ofdetected DMR: random data	adjust. real ÷ random
0.008	3	12 419	7 225	114	28	9	3.1
	4			16	2	0	-
0.01	3	15 089	8 942	160	40	17	2.4
	4			21	5	0	-
0.03	3	38 843	26 785	1080	453	263	1.7
	4			240	85	20	4.3
	5			64	22	3	7.3
0.05	3	60 936	44 899	2603	1380	924	1.5
	4			672	229	127	2.4
	5			225	94	18	5.2
0.08	3	92 657	73 225	5552	3506	2563	1.3
	4			1684	935	461	2.0
	5			581	287	81	3.5
0.1	3	118 453	95 687	8331	5659	4358	1.2
	4			2733	1679	912	1.8
	5			1016	560	190	2.9

with second random dataset

pvalue thresh.	Min nb of C	Nb of selected C: real data	Nb of selected C: random data	Nb of detected DMR: real data	Nb of detected DMR: adjusted real data	Nb of detected DMR: random data	adjust. real ÷ random
0.05	3	60 936	44 584	2603	1280	814	1.6
	4			672	281	98	2.9
	5			225	87	15	5.8
0.06	5	71 157	53 441	329	138	28	4.9
0.07	5	81 494	62 842	432	184	42	4.4
0.08	3	92 657	73 170	5552	3467	2368	1.5
	4			1684	932	429	2.2
	5			581	263	72	3.7

Arguments values finally used for DMR detection in the real dataset



Two random datasets (for details on random dataset obtention, see Supp. Fig S1E and **Materials & Methods**) were compared to the actual data, by testing different combinations of DMRs detection parameters (pvalue threshold, minimum number of CpGs in the region), and by fixing some parameters (*distance_C_to_C* = 100, *distance_threshold*=500, *max_region_size* = 2000). Depending on the pvalue threshold chosen, a greater or lesser number of CpGs is selected. In each case, more CpGs are selected in the real dataset than in the random one for a given p-value threshold. To be able to compare the number of DMRs obtained in the real dataset and in the random one, the same number of CpGs must be considered at the outset. Thus, by random selection of relevant CpGs into the real dataset, the number of CpGs in the real dataset was adjusted to the number of CpGs found in the random dataset. ***Abbreviations***: ***Min nb of C***: minimum number of CpGs required to obtain a DMR; ***Nb of selected C***: number of CpG taken into account, depending on the chosen pvalue threshold ; ***adjust. real data ÷ random***: ratio of the number of DMRs detected in the adjusted real dataset and the number of DMRs detected in the random dataset. On the scheme, filled circles represent hypermethylated cytosines observed between two datasets (*e.g.* control group *versus* test group), whereas empty circles represent hypomethylated cytosines. *: non significant pvalue (pval ≥ 0.07); *** significant pvalue (pval < 0.07).

Suppl. Table 3 : Summary of key number of major steps of the bioinformatic analysis of the methylome capture.

		1M-R1	1M-R2	2M-R1	2M-R2	3M-R1	3M-R2	4M-R1	4M-R2	5M-R1	5M-R2	6M-R1	6M-R2
Raw data	Initial number of reads	67 884 931	67 884 931	66 493 707	66 493 707	**66 654 738**	**66 654 738**	**68 515 383**	**68 515 383**	**75 016 568**	**75 016 568**	**73 103 368**	**73 103 368**
Trimmomatic	Nb of reads after trimming	45 188 206	45 188 206	48 803 924	48 803 924	52 093 098	52 093 098	48 406 088	48 406 088	54 295 610	54 295 610	54 238 624	54 238 624
	% of remaining reads	66,57	66,57	73,40	73,40	78,15	78,15	70,65	70,65	72,38	72,38	74,19	74,19
Cutadapt	% of removed bases	19,9	19,2	19,9	19,3	19,9	19,2	19,9	19,2	19,9	19,2	19,9	19,2
Bismark	Nb of mapped reads	27 132 662		33 834 717		32 974 018		33 399 024		37 335 974		37 685 044	
-paired end mapping - mm9 - non directional library	Mapping efficiency (mapped reads ÷ trimmed reads - %)	**60,0**		**69,3**		**63,3**		**69,0**		**68,8**		**69,5**	
	Total number of C's analysed	**1 393 221 218**		**1 723 029 465**		**1 688 063 509**		**1 671 822 446**		**1 915 763 720**		**1 954 602 883**	
	Total 5mC in CpG context	38 343 757		50 120 323		46 162 661		46 221 886		51 128 425		52 101 907	
	% 5mC in CpG context ()	**42,1**		**41,5**		**42,8**		**43,7**		**41,1**		**40,9**	
deduplication	Total number of alignments analysed	27 132 662		33 834 713		32 974 016		**33 399 023**		**37 335 971**		**37 685 043**	
	Total count of deduplicated leftover sequences (% of total)	85,87		78,88		80,66		79,61		78,46		80,85	
bismark_ methylation_ extractor	Total number of methylation call strings (=R1 + R2 reads)	46 595 798		53 377 408		53 191 184		53 179 200		58 586 780		60 938 390	
	Total number of C's analysed:	786 358 255		905 431 146		945 954 930		886 146 203		1 021 808 647		1 076 902 457	
	Total 5mC in CpG context:	21 654 699		26 665 187		26 052 368		24 693 496		27 465 624		28 834 224	
	Total 5mC in CHG context:	609 486		754 954		744 991		712 498		803 420		829 707	
	Total 5mC in CHH context:	2 007 478		2 260 093		2 376 417		2 251 004		2 554 514		2 630 633	
	Tot. C→T conversion (CpG)	29 956 981		36 584 824		34 103 744		31 538 942		38 906 686		41 911 091	
	Tot. C→T conversion (CHG)	199 634 753		235 186 432		237 395 182		223 667 975		256 270 802		271 045 886	
	Tot. C→T conversion (CHH)	532 494 858		603 979 656		645 282 228		603 282 288		695 807 601		731 650 916	
	5mC in CpG context (%)	**42,0**		**42,2**		**43,3**		**43,9**		**41,4**		**40,8**	
	5mC in CHG context (%)	**0,3**		**0,3**		**0,3**		**0,3**		**0,3**		**0,3**	
	5mC in CHH context (%)	**0,4**		**0,4**		**0,4**		**0,4**		**0,4**		**0,4**	
	Total C in CpG (5C + 5mC)	51611680		63250011		60156112		56232438		66372310		70745315	
bedtools intersect	% deduplicated reads on capture regions	57,36		55,38		54,80		59,44		59,34		61,20	
Statistical analysis with methylkit + get-close_loci()	Nb CpG sites covered in all samples (unfiltered)						3 016 144						
	Filtered positions (min. cov. = 10, high. Perc = 99.9 ; coverage normalisation	1 676 294		1 920 575		1 876 952		1 766 642		2 080 535		2 149 843	
	Filtered CpG positions covered by ALL samples						1 259 111						
	DMC (Meth.diff 5%, qval<0,05)	4											
	DMR all	432											

For more details about the bioinformatic workflow, please see Materiel & Methods and command line in Supp. data Jupyter notebook n°1.

Early DNA methylation defects are predominantly located in brain active enhancers

Even if there was a limited number of regions altered in their DNA-methylation levels, at least in the repertoire of the captured *loci*, our results suggest that DNA methylation is rapidly and locally redistributed in the cortices exposed to PAE. To determine whether methylation defects were randomly distributed among capture regions or not, we performed hypergeometric tests either on all DMRs, or by separating hypo-DMRs and hyper-DMRs. We observed that genomic regions were not randomly altered by PAE in their DNA methylation status. Indeed, brain active enhancers, characterized by the H3K27ac histone mark, were significantly over-represented among the DMRs identified in the capture, (345 on 432 DMRs *i.e.*, 79.86% of DMRs; hypergeometric test, pval < 0.05. Fig. 2B, Fig. 3 and Table 2). In contrast, promoters and other regions of interest, which represent respectively 134 and 35 DMRs, were not significantly affected by the binge drinking stress (Fig. 2A and 2B).

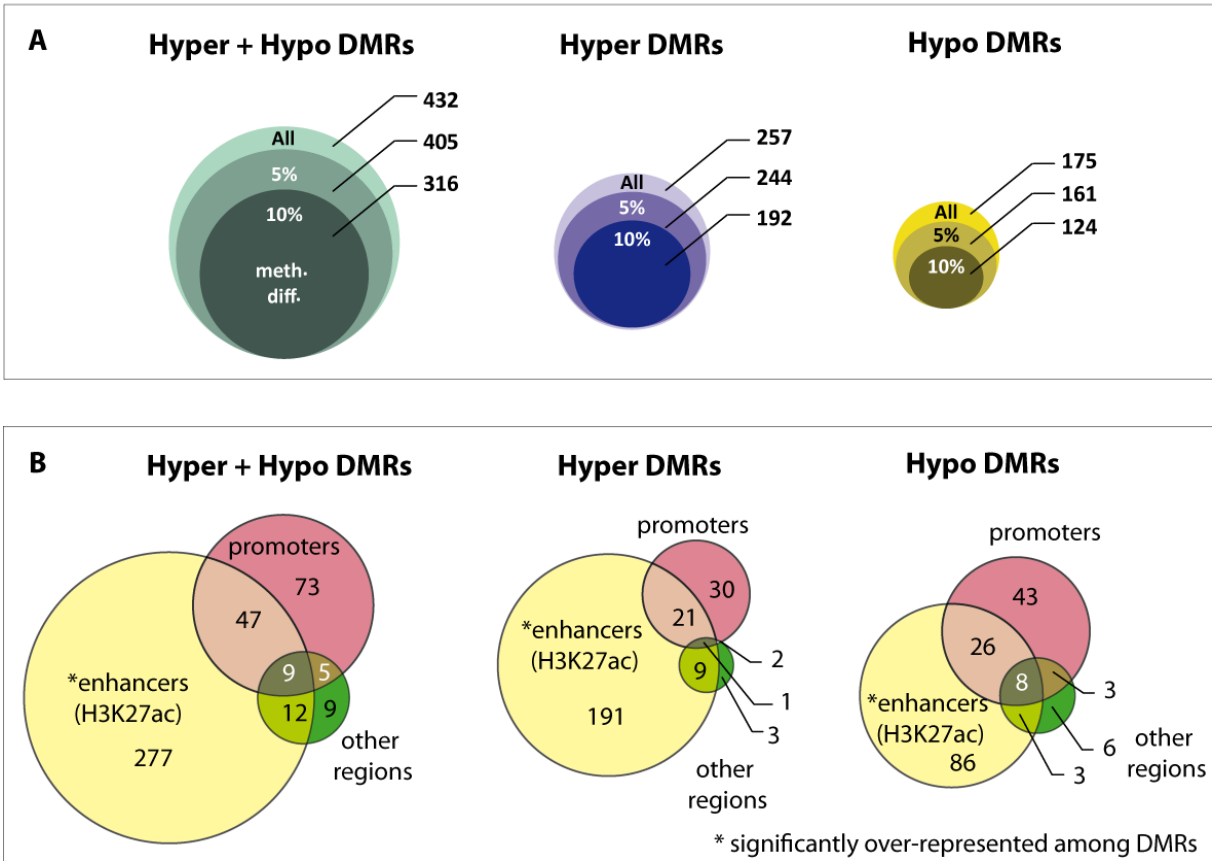


Figure 2 : Early modifications of DNA methylation observed in embryo cortices upon PAE

(A) Number of differentially methylated regions (all, hyper- or hypo- DMRs) promptly observed upon PAE. Numbers of DMRs filtered by methylation differential rate (indicated as percent) observed between alcohol-treated and control groups are also indicated.

(B) Number of DMRs (all, hyper- or hypo- DMRs) promptly observed upon PAE, in each methylome capture category. A DMR is associated to a capture category if it overlaps to or if it is closed to (distance less than 500 bp) a capture region of a given category. Asterisk indicates categories that are significantly over-represented among DMRs, according to hypergeometric tests (see Table 2).

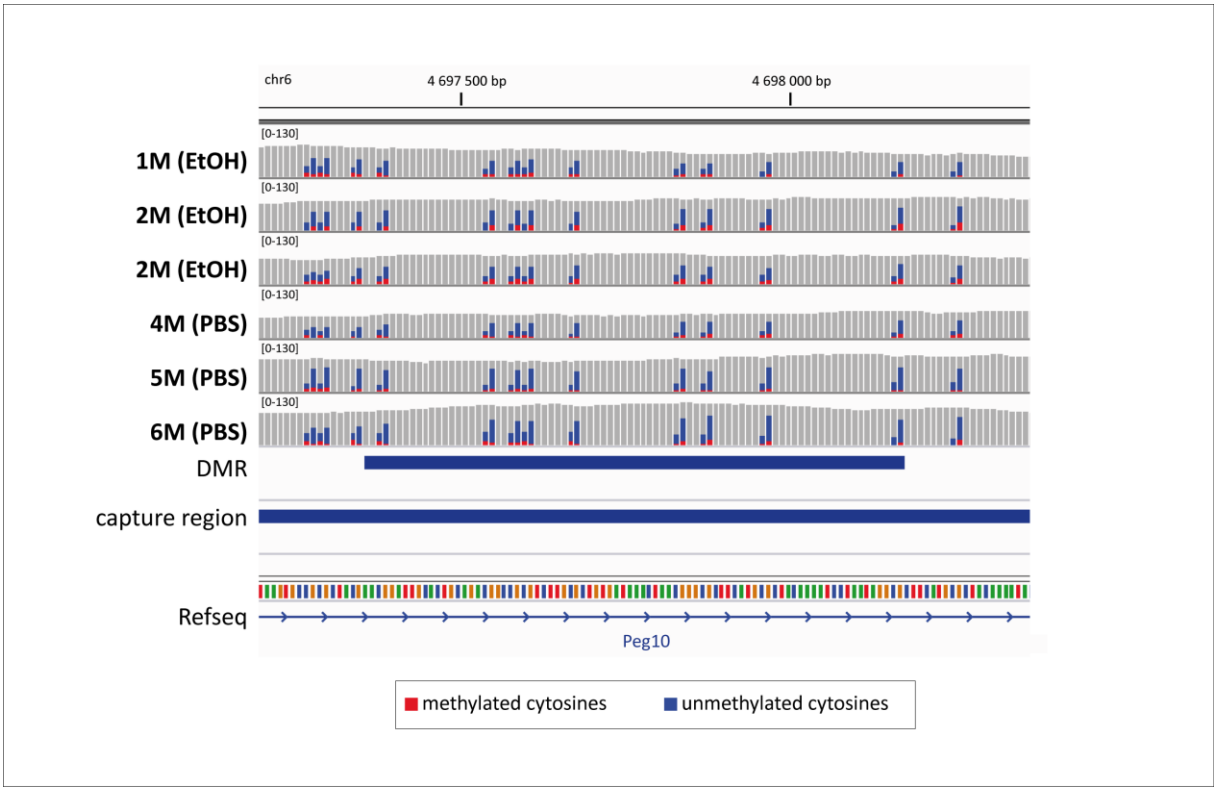


Figure 3 : Example of DMR observed in *Peg10* imprinted gene

Vizualisation of deduplicated read coverage of EtOH- and PBS-treated samples. Hypermethylated region located into *Peg10* imprinted gene is shown (narrow blue bar). Red color represents methylated cytosines, whereas blue color represents unmethylated ones, both in CpG context only. Capture region is also indicated (large blue bar). This screenshot was obtained using IGV. Scale is the same for all samples.

Table 2 : Early DNA methylation defects are predominantly located in brain active enhancers.

	Nb of regions in capture category	Number of DMRs			Hypergeometric test pvalues		
		Nb DMRsAll	Nb DMRsHyper	Nb DMRsHypo	All	Hyper	Hypo
promoters	23 769	135	55	80	1	1	0.09
Enhancers (H3K27ac)	34 279	349	225	124	3.38e-23	1.59e-24	4.68e-4
Other regions	12 108	36	15	21	1	1	1
Global capture	56 811	-	-	-	-	-	-

Results of hypergeometric test designed for determining whether specific capture elements are specifically over-represented among DMRs observed promptly upon binge drinking stress. Values highlighted in red are significant. For this analysis, DMRs are considered to overlap or to be closed to capture regions categories (*i.e.* distance less than 500bp between the regions). Numbers of DMRs that were considered in each analysis are also indicated in the table. They differ very slightly from the numbers indicated in Figure 2B, since some DMRs could overlap several capture regions, and were, in this case, count more than one time.

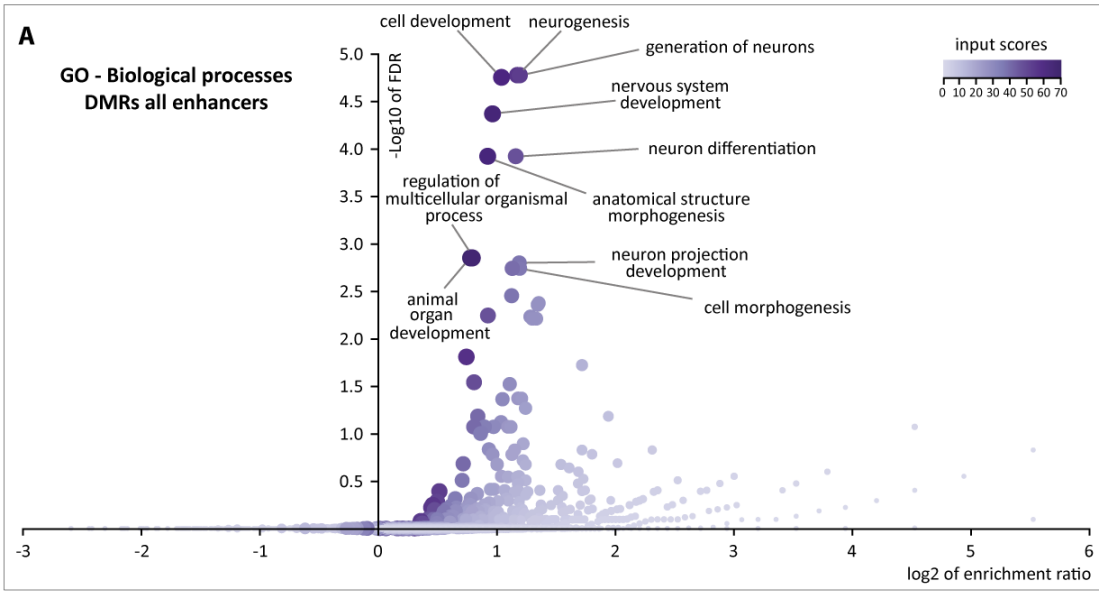
Table 3 : GO results for genes associated to all DMRs (Hypo and Hypermethylation) that are localized in active enhancers (H3K27ac)

GO category	description	size	overlap	expect	Enrichment Ratio	FDR	User ID
GO:0022008	**neurogenesis**	1165	57	25.21	2.260944	0.000017	*Satb2, Klf7, Pbx1, Sdccag8, Prox1, Ank3, Syt1, Dbnl, Bcl11a, Kif3c, Eml1, Atxn1, Itga1, Farp1, Asap1, Trappc9, App, Itsn1, Arid1b, Anks1, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Dab2ip, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Spen, Cit, Cux2, Cux1, Dlx5, Creb3l2, Ctnna2, Mgll, Itpr1, Cd9, Sox5, Mboat7, Zfp536, Shank1, Inpp5f, Fgfr2, Shank2, Efnb2, Lig4, Arhgef10, Unc5d, Gnao1, Snx1, Ephb1, Map4*
GO:0048699	**generation of neurons**	1088	54	23.54	2.293537	0.000017	*Satb2, Klf7, Pbx1, Sdccag8, Prox1, Ank3, Syt1, Dbnl, Bcl11a, Kif3c, Eml1, Atxn1, Itga1, Farp1, Asap1, Trappc9, App, Itsn1, Arid1b, Anks1, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Dab2ip, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Spen, Cit, Cux2, Cux1, Dlx5, Creb3l2, Ctnna2, Mgll, Itpr1, Sox5, Zfp536, Shank1, Inpp5f, Fgfr2, Shank2, Efnb2, Lig4, Unc5d, Gnao1, Snx1, Ephb1, Map4*
GO:0048468	**cell development**	1435	64	31.05	2.060957	0.000018	*Paqr8, Satb2, Klf7, Pbx1, Prox1, Map7, Jmjd1c, Ank3, Chst11, Syt1, Dbnl, Bcl11a, Kif3c, Atxn1, Hrh2, Itga1, Farp1, Asap1, Cldn5, App, Itsn1, Arid1b, Pacrg, Anks1, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Dab2ip, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Spen, Cit, Cux2, Rilpl1, Cux1, Dlx5, Creb3l2, Ctnna2, Antxr1, Mgll, Itpr1, Cd9, Sox5, Sipa1l3, Zfp536, Shank1, Inpp5f, Fgfr2, Shank2, Efnb2, Lig4, Arhgef10, Unc5d, Gnao1, Snx1, Myo1e, Ephb1, Map4*
GO:0007399	**nervous system development**	1583	67	34.26	1.955847	0.000043	*Satb2, Klf7, Pbx1, Sdccag8, Trp53bp2, Prox1, Ank3, Bcr, Apaf1, Syt1, Dbnl, Bcl11a, Abr, Rbfox3, Kif3c, Eml1, Atxn1, Itga1, Farp1, Mal2, Asap1, Trappc9, Cldn5, App, Itsn1, Arid1b, Anks1, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Dab2ip, Zeb2, Slc1a2, Dclk1, Dclk2, Nfib, Nfia, Ncdn, Spen, Mthfr, Cit, Cux2, Cux1, Dlx5, Creb3l2, Ctnna2, Mgll, Itpr1, Cd9, Sox5, Mboat7, Zfp536, Shank1, Inpp5f, Fgfr2, Shank2, Efnb2, Lig4, Arhgef10, Unc5d, Gnao1, Snx1, Ephb1, Map4*
GO:0030182	**neuron differentiation**	990	48	21.42	2.240510	0.000121	*Satb2, Klf7, Pbx1, Prox1, Ank3, Syt1, Dbnl, Bcl11a, Kif3c, Itga1, Farp1, Asap1, Trappc9, App, Itsn1, Arid1b, Anks1, Runx2, Efna5, Ptprm, Tcf4, Ablim1, Dab2ip, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Cit, Cux2, Cux1, Dlx5, Creb3l2, Ctnna2, Mgll, Itpr1, Sox5, Zfp536, Shank1, Inpp5f, Fgfr2, Shank2, Efnb2, Unc5d, Gnao1, Snx1, Ephb1, Map4*
GO:0009653	**anatomical structure morphogenesis**	1604	66	34.71	1.901431	0.000121	*Satb2, Klf7, Cdc73, Pbx1, Sdccag8, Prox1, Map7, Stox1, Jmjd1c, Ank3, Bcr, Chst11, Apaf1, Syt1, Dbnl, Bcl11a, Serpinf2, Abr, Cdc42ep4, Gaa, Pik3cg, Slc24a4, Ryr2, Hrh2, Itga1, Farp1, App, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Dab2ip, Zeb2, Fap, Ocstamp, Dclk1, Nfib, Mthfr, Ajap1, Cit, Cux2, Rilpl1, Cux1, Gna12, Dlx5, Ctnna2, Antxr1, Mgll, Itpr1, Cd9, Mboat7, Sipa1l3, Tshz3, Shank1, Fgfr2, Shank2, Efnb2, Unc5d, Gab1, Gnao1, Sik3, Snx1, Myo1e, Ephb1*
GO:0048513	**animal organ development**	1906	71	41.25	1.721378	0.001418	*Satb2, Klf7, Cdc73, Pbx1, Trp53bp2, Prox1, Map7, Stox1, Jmjd1c, Bcr, Chst11, Apaf1, Syt1, Gas2l1, Bcl11a, Abr, Gaa, Max, Slc24a4, Eml1, Ryr2, Atxn1, Hrh2, Nln, Trappc9, Scn8a, App, Runx2, Plcl2, Efna5, Man2a1, Ptprm, Tmem178, Dab2ip, Zeb2, Mettl8, Slc1a2, Meis2, Ocstamp, Tpd52, Dclk1, Dclk2, Ptpn3, Nfib, Nfia, Ajap1, Cux1, Peg10, Dlx5, Creb3l2, Ctnna2, Itpr1, Cd9, Sox5, Mboat7, Sipa1l3, Zbtb32, Tshz3, Shank1, Tmem143, Fgfr2, Shank2, Efnb2, Lig4, Tcim, Gab1, Gnao1, Cbfa2t3, Sik3, Myo1e, Ephb1*
GO:0051239	**regulation of multicellular organismal process**	1836	69	39.731207	1.736670	0.001418	*Klf7, Cdc73, Pbx1, Prox1, Stox1, Bcr, Syt1, Gas2l1, Grb10, Bcl11a, Sptbn1, Serpinf2, Abr, Gaa, Kif3c, Ryr2, Atxn1, Hrh2, Nln, Farp1, Fam49b, Asap1, Cldn5, Cblb, App, Itsn1, Runx2, Plcl2, Efna5, Man2a1, Ptprm, Tmem178, Prkce, Mapre2, Tcf4, Add3, Dab2ip, Zeb2, Meis2, Ocstamp, Nfib, Spen, Ajap1, Cmklr1, Cit, Cux2, Cux1, Dlx5, Creb3l2, Mgll, Iqsec1, Itpr1, Cd9, Sox5, Zbtb32, Tshz3, Zfp536, Shank1, Sult2b1, Inpp5f, Fgfr2, Shank2, Efnb2, Lig4, Tcim, Unc5d, Gab1, Gnao1, Ephb1*
GO:0031175	**neuron projection development**	768	38	16.619590	2.286458	0.001602	*Klf7, Ank3, Syt1, Dbnl, Bcl11a, Kif3c, Itga1, Farp1, Asap1, App, Itsn1, Arid1b, Efna5, Ptprm, Ablim1, Dab2ip, Zeb2, Dclk1, Nfib, Ncdn, Cit, Cux2, Cux1, Dlx5, Creb3l2, Ctnna2, Mgll, Itpr1, Shank1, Inpp5f, Fgfr2, Shank2, Efnb2, Unc5d, Gnao1, Snx1, Ephb1, Map4*
GO:0000902	**cell morphogenesis**	746	37	16.143508	2.291943	0.001827	*Klf7, Prox1, Map7, Jmjd1c, Ank3, Syt1, Dbnl, Bcl11a, Cdc42ep4, Hrh2, Itga1, Farp1, App, Efna5, Ptprm, Ablim1, Dab2ip, Zeb2, Dclk1, Nfib, Cit, Cux2, Rilpl1, Cux1, Gna12, Dlx5, Ctnna2, Antxr1, Mgll, Sipa1l3, Shank1, Fgfr2, Shank2, Efnb2, Unc5d, Snx1, Ephb1*

Table 4: GO results for genes associated to Hypermethylated regions, observed upon EPA, that are localized in active enhancers (H3K27ac)

geneSet	description	size	overlap	expect	Enrichment Ratio	FDR	User ID
GO:0048699	**generation of neurons**	1088	40	15.55	2.57	0.000126	*Satb2, Pbx1, Sdccag8, Prox1, Ank3, Bcl11a, Kif3c, Eml1, Atxn1, Itga1, Farp1, Trappc9, App, Itsn1, Arid1b, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Spen, Cit, Cux2, Ctnna2, Mgll, Sox5, Zfp536, Inpp5f, Fgfr2, Shank2, Efnb2, Unc5d, Ephb1, Map4*
GO:0022008	**neurogenesis**	1165	41	16.65	2.46	0.000137	*Satb2, Pbx1, Sdccag8, Prox1, Ank3, Bcl11a, Kif3c, Eml1, Atxn1, Itga1, Farp1, Trappc9, App, Itsn1, Arid1b, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Spen, Cit, Cux2, Ctnna2, Mgll, Sox5, Zfp536, Inpp5f, Fgfr2, Shank2, Efnb2, Arhgef10, Unc5d, Ephb1, Map4*
GO:0030182	**neuron differentiation**	990	35	14.15	2.47	0.001253	*Satb2, Pbx1, Prox1, Ank3, Bcl11a, Kif3c, Itga1, Farp1, Trappc9, App, Itsn1, Arid1b, Runx2, Efna5, Ptprm, Tcf4, Ablim1, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Cit, Cux2, Ctnna2, Mgll, Sox5, Zfp536, Inpp5f, Fgfr2, Shank2, Efnb2, Unc5d, Ephb1, Map4*
GO:0007399	**nervous system development**	1583	45	22.62	1.99	0.006082	*Satb2, Pbx1, Sdccag8, Prox1, Ank3, Bcr, Apaf1, Bcl11a, Kif3c, Eml1, Atxn1, Itga1, Farp1, Trappc9, App, Itsn1, Arid1b, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Zeb2, Dclk1, Dclk2, Nfib, Nfia, Ncdn, Spen, Mthfr, Cit, Cux2, Ctnna2, Mgll, Sox5, Zfp536, Inpp5f, Fgfr2, Shank2, Efnb2, Arhgef10, Unc5d, Ephb1, Map4*
GO:0048468	**cell development**	1435	42	20.50	2.05	0.006082	*Satb2, Pbx1, Prox1, Jmjd1c, Ank3, Bcl11a, Kif3c, Atxn1, Hrh2, Itga1, Farp1, App, Itsn1, Arid1b, Pacrg, Runx2, Efna5, Man2a1, Ptprm, Tcf4, Ablim1, Zeb2, Dclk1, Dclk2, Nfib, Ncdn, Spen, Cit, Cux2, Ctnna2, Mgll, Sox5, Sipa1l3, Zfp536, Inpp5f, Fgfr2, Shank2, Efnb2, Arhgef10, Unc5d, Ephb1, Map4*
GO:0000904	**cell morphogenesis involved in differentiation**	549	23	7.84	2.93	0.006492	*Prox1, Jmjd1c, Ank3, Bcl11a, Hrh2, Farp1, App, Efna5, Ptprm, Ablim1, Zeb2, Dclk1, Nfib, Cit, Cux2, Ctnna2, Mgll, Sipa1l3, Fgfr2, Shank2, Efnb2, Unc5d, Ephb1*
GO:0021953	**central nervous system neuron differentiation**	132	10	1.89	5.30	0.029620	*Satb2, Prox1, Zeb2, Dclk1, Dclk2, Nfib, Sox5, Fgfr2, Unc5d, Ephb1*
GO:0051239	**regulation of multicellular organismal process**	1836	47	26.23	1.79	0.029620	*Cdc73, Pbx1, Prox1, Bcr, Grb10, Bcl11a, Sptbn1, Serpinf2, Kif3c, Ryr2, Atxn1, Hrh2, Nln, Farp1, Fam49b, Cblb, App, Itsn1, Runx2, Plcl2, Efna5, Man2a1, Ptprm, Tmem178, Prkce, Mapre2, Tcf4, Zeb2, Meis2, Nfib, Spen, Cmklr1, Cit, Cux2, Mgll, Iqsec1, Sox5, Zbtb32, Tshz3, Zfp536, Inpp5f, Fgfr2, Shank2, Efnb2, Unc5d, Gab1, Ephb1*
GO:0048667	**cell morphogenesis involved in neuron differentiation**	448	19	6.40	2.97	0.029620	*Ank3, Bcl11a, Farp1, App, Efna5, Ptprm, Ablim1, Zeb2, Dclk1, Nfib, Cit, Cux2, Ctnna2, Mgll, Fgfr2, Shank2, Efnb2, Unc5d, Ephb1*
GO:0048812	**neuron projection morphogenesis**	494	20	7.06	2.83	0.030939	*Ank3, Bcl11a, Itga1, Farp1, App, Efna5, Ptprm, Ablim1, Zeb2, Dclk1, Nfib, Cit, Cux2, Ctnna2, Mgll, Fgfr2, Shank2, Efnb2, Unc5d, Ephb1*

To identify potential binding sites of transcription factors within the DNA sequences of enhancers located in DMRs, we explored the presence of DNA sequence motifs using *findMotifsGenome* (HOMER). We could not observe any over-representation of specific motifs. This result suggests that there are no sub-categories of enhancers (*i.e.* characterized by specific transcription factors), that would be more severely affected by PAE than others. Since these enhancers belong to an active repertoire in the mouse brain, we then performed a Gene Ontology (GO) analysis on all the enhancers containing DMRs, in order to unveil specific Biological Processes potentially rapidly altered by PAE. We observed that a significant number of genes associated to these enhancers played a role in neurogenesis, generation of neurons, cell development, nervous system development, and neuron differentiation (Fig. 4A, Table 3). When considered separately, enhancers with a decrease in methylation levels upon PAE were not significantly associated to a specific Biological Process. In contrast, hypermethylated enhancers were predominantly associated with genes that are involved in the above-cited Biological Processes (Fig. 4B, Table 4).



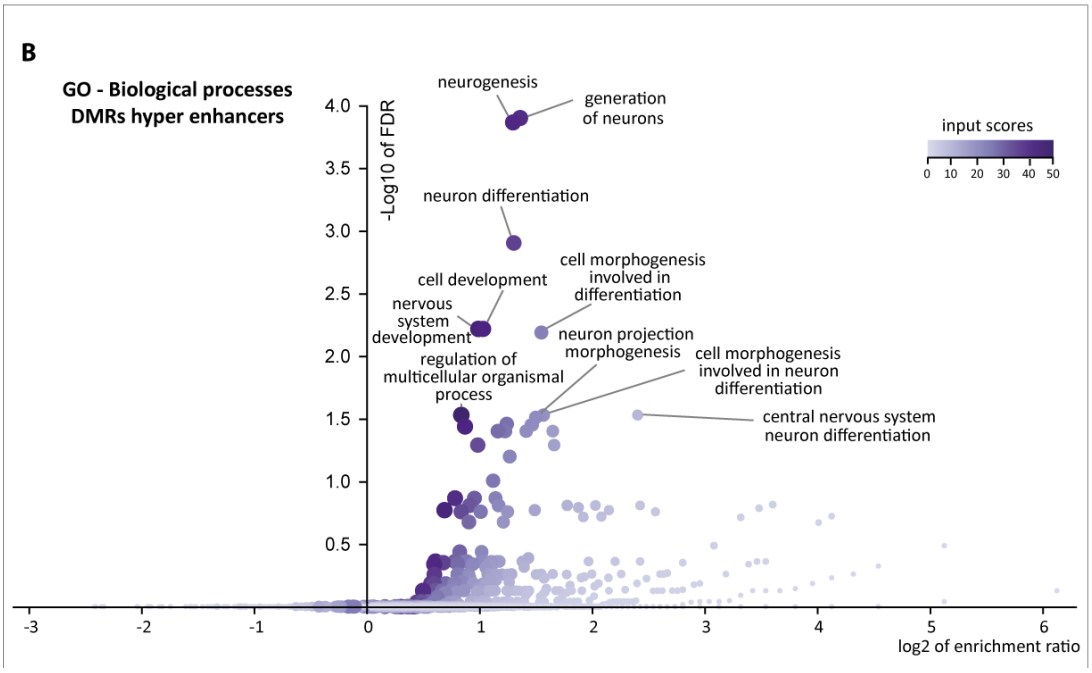


Figure 4: Gene ontology (GO) analyses performed on genes associated to DMRs that are observed into (or close to) active enhancers

Volcano plots showing relevant GO biological processes of genes that are associated to all DMRs (both hypo- and hypermethylated regions) **(A)** or associated to hypermethylated regions only **(B)**, observed into or close to active enhancers (distance between DMRs and active enhancers regions less than 500bp) in both cases. The size and color of the dot is proportional to the number of overlapping of the category (Liao et al., 2019). Genes associated to the ten more relevant GO biological processes of these analyses are described in Table 3 (for GO performed on all DMRs observed in active enhancers, related to Fig. 4A) and Table 4 (for GO performed only on hypermethylated regions observed upon PAE in active enhancers, related to Fig. 4B). These data were obtained using *Web-based GEne SeT AnaLysis Toolkit* (WebGestalt ; Liao et al., 2019).

Early DNA-methylation changes upon PAE affect imprinted genes and genes of protocadherin clusters

As CpG islands (CGis) are potentially relevant targets for DNA methylation variation (Saxonov et al., 2006), we determined whether the global list of DMRs identified after PAE was associated with these kind of *loci*. We found that only 36 out of the 432 DMRs were located into CGis, but, significantly, a quarter of these DMRs corresponded to imprinted genes (9/36, Table 5, Fig. 3). These genes are tightly regulated by DNA méthylation (Perez et al., 2016). In total, 13 DMRs are associated to 10 distinct imprinted genes in our analysis (9 DMRs in CGis, 4 DMRs not associated with a CGi, Table 5, Fig. 3) This over-representation of imprinted genes among DMRs was confirmed by a hypergeometric test (see details in **Material & method**). A majority of DMRs associated to imprinted genes were located within the promoter regions (8/13) and/or active enhancers (11/13, Table 5).

Clustered protocadherin genes (*Pcdhs*), which are also individually regulated by DNA methylation (Phillips et al., 2017), constitute another group of genes that are particularly represented among early DMRs. Indeed, we identified 10 DMRs in 9 distinct protocadherins, from both α- and γ- protocadherins clusters (Table 6). In contrast to imprinted genes, DMRs found on *Pcdhs* were not associated with brain active enhancers (except for Pcdhγ-a12), but rather located in promoter regions (10/10 of DMRs in *Pcdhs*, 9 distinct promoter regions, Table 6). For 4 *Pcdhs*, DMRs observed early upon PAE are also on a CpGi (Table 6).

Imprinted genes (Perez et al., 2016) and protocadherins are crucial for neurodevelopment (Kernohan and Bérubé, 2010) and for normal brain functions (Davies et al., 2008). As an example, clustered *Pcdhs* play important roles in the modulation of dendrites arborisation and synaptogenesis or limitation of autapse formation through cell recognition (Light and Jontes, 2017; Phillips et al., 2017 ; reviewed in (El Hajj et al., 2016; Matsunaga et al., 2017; Molumby et al., 2017 ; Yamagata et al., 2018). The early alteration of their methylation levels by *in utero* alcohol exposure could thus potentially affect their expression levels shortly after the last injection of ethanol and thus leads to neural defects typical of FASD.

Regions showing chromatin remodelling during physiological brain development are significantly associated with PAE-induced DNA methylation events

The striking enrichment of DMRs in enhancer regions that are specifically active in the brain, let us think that these DNA methylation events could occur in regions whose chromatin accessibility is modulated during the brain developmental stages surrounding the PAE period. We thus evaluated global chromatin accessibility changes in parallel to gene expression dynamics throughout mouse

Table 5 : Imprinted genes that are affected in DNA methylation upon *in utero* binge drinking stress.

Imprinted genes	DMRs chr:start-end (mm10 coordinates)	Meth. Diff (%)	Median pval.	Enhancers (H3K27ac) overlap	Promoters Overlap (bp)	CGi start-end	DOCR	DEG
Commd1 / Zrsr1	chr11:22972132-22972265	+16.58	0.056	134	134	22971974-22972993	-	↘ E14.5 vs E15.5
Gab1	chr8:80830430-80830841	+12.99	0.022	412	.	.	-	↘ E14.5 vs E15.5
Gnas	chr2:174297476-174297552	+21.75	0.008	.	-389;77	174297262-174297532	-	↘ E15.5 vs E16.5
Grb10	chr11:12025913-12026141	+17.86	0.063	229	.	12025554-12026332	-	↗ E14.5 vs E15.5
H13	chr2:152686280-152686714	+16.43	0.014	435	218	.	-	-
	chr2:152686874-152686894	+13.09	0.024	21	21	152686809-152687230	-	-
Impact	chr18:12972066-12972404	+24.41	0.028	.	339	.	-	-
Inpp5f	chr7:128687978-128688477	+14.53	0.068	500	.	128687969-128688582	-	
	chr7:128690636-128691043	+9.96	0.042	408	.	.		
Nap1l5	chr6:58906656-58906739	-14.41	0.068	84	.	58906723-58907147	-	
	chr6:58906833-58907186	+15.69	0.066	354	-289	58906723-58907147		
Peg10	chr6:4747935-4748017	+11.42	0.037	-263	83	4747291-4748412	-	↘ E14.5 vs E15.5
Peg13	chr15:72810096-72810197	+14.98	0.009	102	-476	72809537-72810123	-	↗ E14.5 vs E15.5
								↘ E14.5 vs E15.5

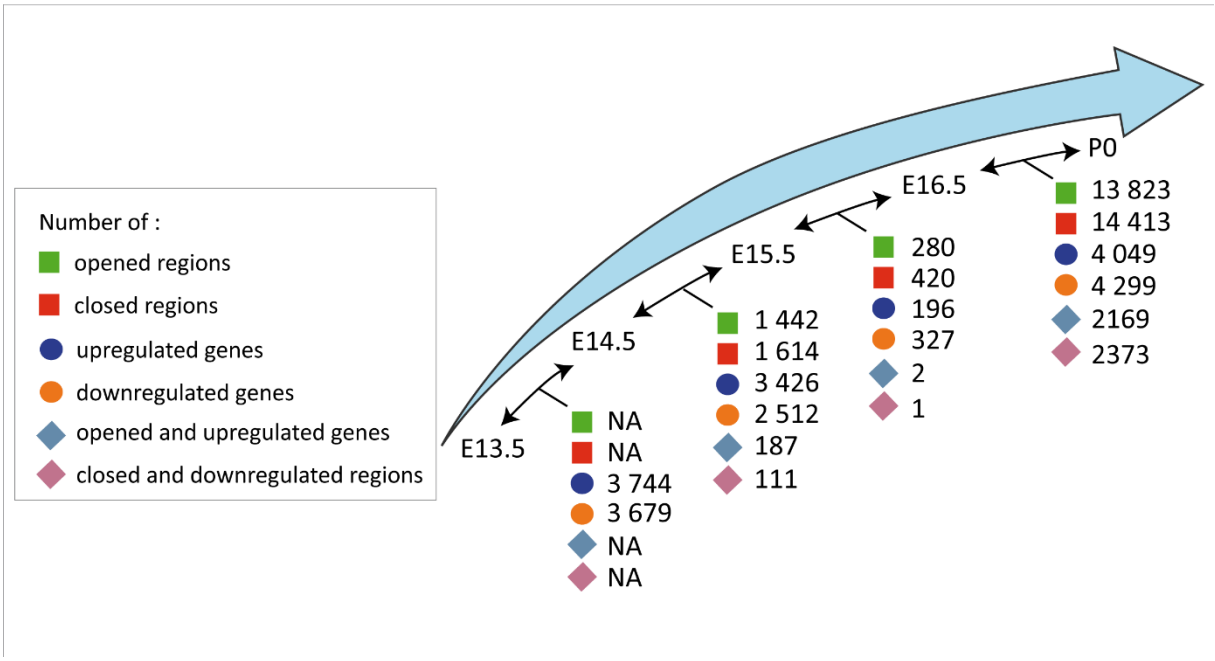
Overlap of PAE affected imprinted genes with capture regions was investigated. Strict overlap and close regions were considered (distance between regions less than 500bp). Thus, negative values in the table for the overlap correspond to close DMRs and capture regions but without any overlap. Median pval. corresponds to the median of pvalues of each relevant CpGs (with pval < 0.07) that composed the DMRs.

Table 6 : Protocadherin genes that were affected in DNA methylation upon *in utero* binge drinking stress.

Pcdh genes	DMR chr:start-end (mm10 coordinates)	Meth. Diff (%)	Median pval.	Enhancers (H3K27ac) overlap	Promoters Overlap (bp)	CGi start-end	DOCR	DEG
Pcdhα3	chr18:36946321-36946496	+15.59	0.0198	.	176	.	-	-
Pcdhα9	chr18:36998102-36998372	-11.72	0.0373	.	271	.	-	-
Pcdhα10/	chr18:37005158-37005610	-13.12	0.0337	.	453	37005415-37005622	-	-
Pcdhα11	chr18:37005677-37005756	+13.20	0.0626		80	.		
Pcdhγa2	chr18:37669155-37669362	-13.06	0.0057	.	208	.	-	-
Pcdhγa5	chr18:37694545-37694935	+10.14	0.0486	.	391	37694513-37694761	-	-
Pcdhγa11	chr18:37755743-37756176	-10.25	0.0426	.	434	37755688-37756215	-	-
Pcdhγa12	chr18:37765890-37766023	-10.79	0.0259	134	134	.	-	↗ E14.5 vs E15.5
Pcdhγb4	chr18:37720141-37720375	-9.71	0.0266	.	235	.	-	↗ E14.5 vs E15.5
Pcdhγb7	chr18:37751794-37751861	+7.29	0.0682	.	68	37751792-37752055	-	-
							-	-

Overlap of PAE affected protocadherin genes with capture regions was investigated. Strict overlap and close regions were considered (distance between regions less than 500 bp). Median pval. corresponds to the median of pvalues of each relevant CpGs (with pval < 0.07) that composed the DMRs.

forebrain development, in physiological conditions. For that, we used our own bioinformatic workflow that allows for pairwise comparisons of successive developmental stages (ATAC-seq and RNA-seq available ENCODE data performed in mouse forebrain at distinct developmental stages in physiological conditions. Results are in Datasets DOCR and Datasets DEG[11]. Numbers of identified regions or genes in each dataset are indicated on Fig. S3. For details on bioinformatic workflow, please see Fig.1C, **Material & Methods**, Supp. Data Jupyter notebook n°2 & 3[12], Supp. Table S4, Supp. Table S5 . Quality control of these data were performed, see Fig. S4 and Fig. S5).

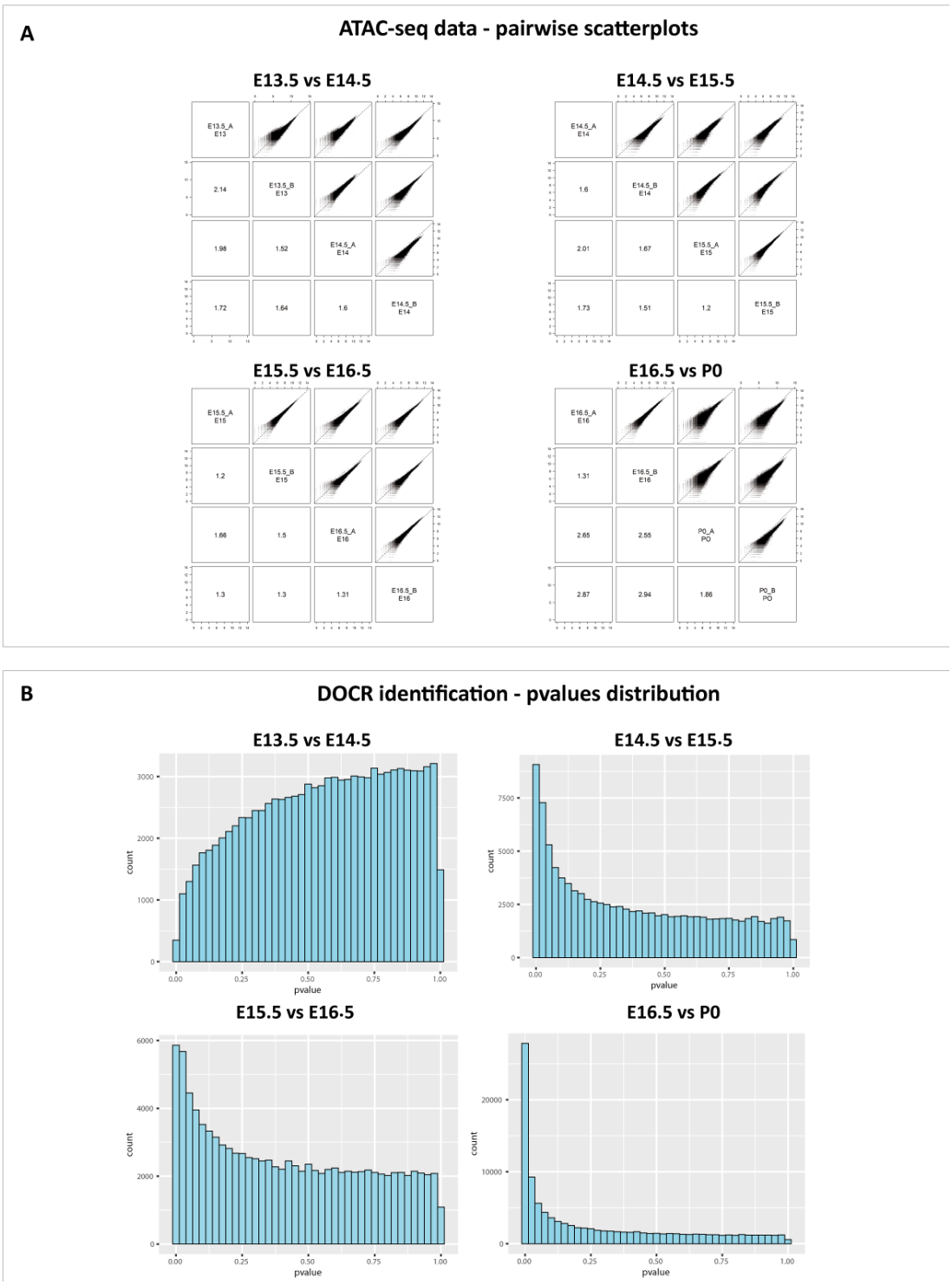


Supp. Figure S3 : Number of differentially opened or closed regions (DOCR) or differentially expressed genes (DEG) identified in the developing brain under physiological conditions.

Available ENCODE ATAC-seq and RNA-seq data were used to identify differentially opened or closed regions (DOCR) and differentially expressed genes (DEG) during brain development, under physiological conditions. DOCR and DEG were obtained in a pairwise comparison manner between two successive developmental stages (embryonic days E13.5, E14.5, E15.5, E16.5 and postnatal day P0). No DOCR could be detected between E13.5 and E14.5 development stages. Indeed, since pvalues distribution was unexpected for this comparison (high number of regions with high pvalues, see in Supp. Fig. S4B DOCR E13.14 adjusted pvalues distribution), it was not recommended to use *edgeR* to detect DOCR between these samples (the data don't follow the right statistical model). Between E15.5 and E16.5 developmental stages, few DOCR and DEG are identified in the developing brain under physiological conditions, compared to E14.5 and E15.5 developmental stages comparisons. These could be explained by the quiet equivalent inter- and intra- variabilities observed for E15.5 and E16.5 samples, whereas inter-group variability was greater between E14.5 and E15.5 samples (Supp. Fig. S4A & S5A). Many DOCR and DEG are identified between E16.5 and P0 stages, compared to other pairwise comparisons. This result is not totally surprising: since more time (few days) separated these two stages (*versus* only one day in the other comparisons), it is plausible that more biological events occurred and modified chromatin accessibility. This is consistent with scatterplots and *Simple Error Ratio Estimate* (*SERE*) values observed between these samples (Supp. Fig. S4A & S5A). Few regions were identified as both DOCR and DEG between each stage, compared to the number of observed DOCR and DEG taken separately. DOCR and DEG were identified using *edgeR*, without logFC threshold but Benjamini Hochberg pvalue adjustment was performed and level of controlled false positive rate was set to 0.05).

[11] *Annexe 6-4.1 : Tableaux de données - thèse - A. Duchateau - tab DOCR ou DEG*

[12] *Annexe 6-3.2 et Annexe 6-3.3: Jupyter notebook : ATAC-seq and RNA-seq workflows*



Supp. Figure S4 : Quality controls and correlation between ATAC-seq samples.

(A) Pairwise scatterplots and matrix of pairwise *Simple Error Ratio Estimate* (*SERE*) values for ATAC-seq samples. Comparisons were performed between two successive developmental stages. *SERE* statistic was used as a similarity index (the more dissimilar the samples are, the higher the SERE value is). Replicates for a given stage, are globally similar, except for E13.5 developmental stages. Indeed, replicates of E13.5 developmental stages are more dissimilar between each other than with E14.5 samples. This observation could explain why we didn't obtain the expected pvalues distribution profile for these samples comparison (see **B**) and why identification of DOCR between these two developmental stages is compromised (Supp. Fig. S3). High inter- and intragroup variabilities are observed for E15.5 and E16.5 samples. These similarities could explain the

lower number of DOCR identified for these sample comparisons, compared to other developmental stages comparisons (Supp. Fig. S3). Scatterplots and SERE values were obtained using *SARTools R* package.
(B) Raw pvalues distribution obtained for ATAC-seq samples pairwise comparisons allowing DOCR identification. For all comparisons except E13.5 *versus* E14.5 stages comparison, pvalues distribution profiles are as expected (*i.e.* a globally uniform distribution with a peak around 0), meaning that DOCR could be detected with *edgeR* statistic tool. For E13.5 *versus* E14.5 stages comparison, distribution didn't follow the expected profile (Supp. Fig. S3), therefore preventing the use of *edgeR* statistical model for DOCR identification. This result could be due to the high dissimilarity observed between E13.5 replicates (see **A**). These pvalue histograms were obtained using *ggplot2 R* package, with data from *SARTools R* package.

With our analysis of ATAC-seq ENCODE data, we observed enrichment of DMRs in regions whose chromatin accessibility changed around the period of PAE for the regions that are more opened between E14.5 and E15.5, in physiological conditions (Fig. 5, Table 7 ; Table 9 ; Supp. Data Jupyter notebook n°4[13] , Datasets Combine_data[14] ; 5 regions out of 313 regions present in the capture repertoire, among 1,442 total regions in the differentially opened regions (DOR) in the original ENCODE dataset; please see the legend of Fig. 5 for normalization details taking into account the constraints of the capture approach). Interestingly, these 5 regions correspond to active enhancers and are associated with important roles for neurodevelopment and brain function (Table 8).

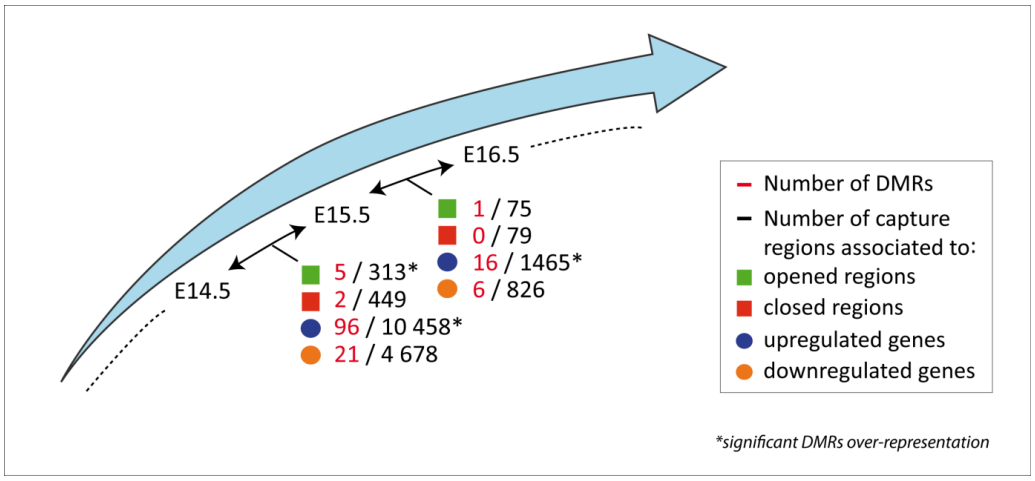


Figure 5 : Some DMRs are significantly located into regions that are modulated during physiological brain development, at the time of ethanol exposure.

Number of capture regions that are associated to differentially opened or closed regions (DOCR) or differentially expressed genes (DEG) observed in the developing brain under physiological conditions. DMRs that are associated to these regions are indicated in red color. DOCR and DEG were obtained in a pairwise comparison manner between two successive developmental studied stages. Asterisks indicated significant results, *i.e.* situations where DMRs are found significantly over-represented (according to hypergeometric tests, see results of these tests in Table 7). Only overlapping regions between DMRs and DOCR/DEG or between capture regions and DOCR/DEG were considered (no distance threshold between distinct regions was allowed). **Note that only DOCR and DEG associated to capture region were taken into account in this analysis**, since no DMR could be detected, even if it exists, in DOCR or DEG that are not represented in methylome capture. Total numbers of DOCR and DEG found during physiological brain development (*i.e.* independently of their presence in capture regions) are indicated in Supp. Fig. S3. Numbers of DOCR / DEG in Supp. Fig. S3 could be lower than the number indicated here, because a given DOCR or DEG could be represent by several capture regions. DOCR and DEG were identified using *edgeR* (using no logFC threshold but Benjamini Hochberg pvalue adjustment was performed and level of controlled false positive rate was set to 0.05).

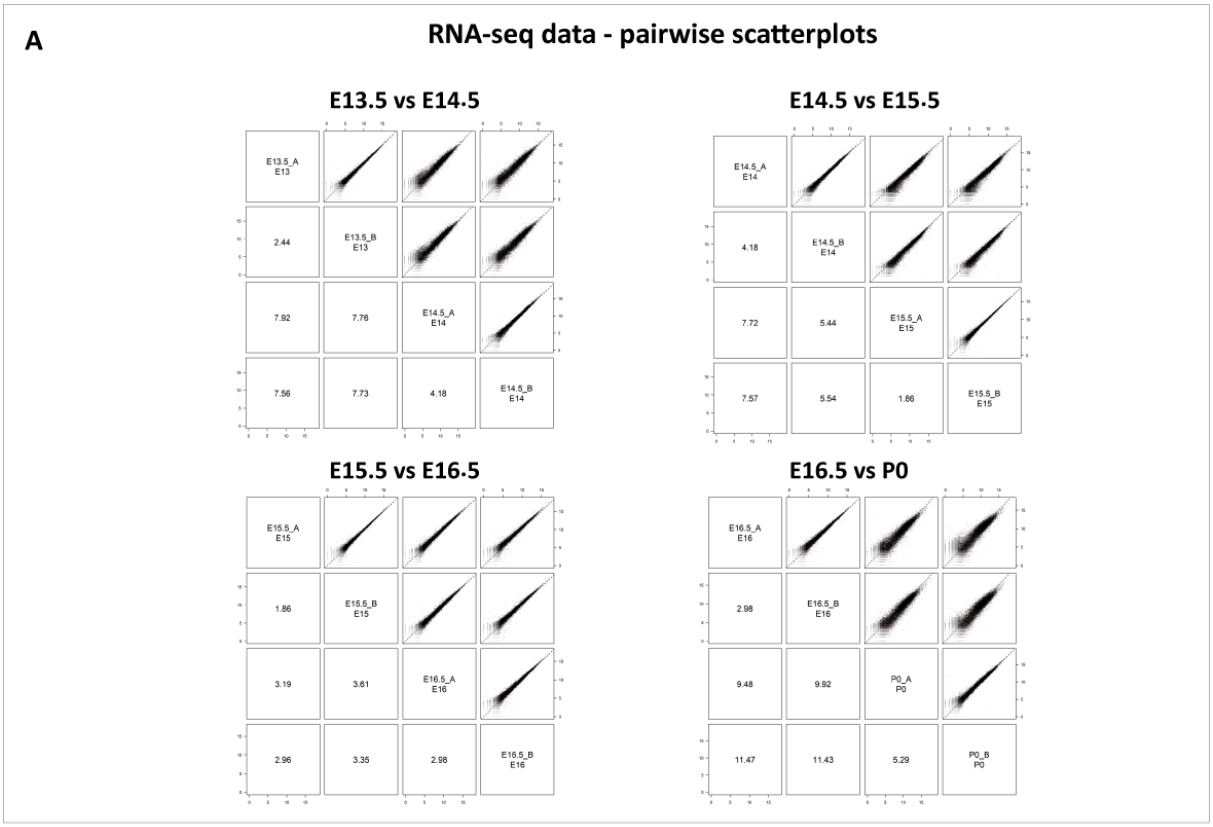
[13] *Annexe 6-3.4: Jupyter notebook : Workflow used to integrate DMRs, DOCR and DEG datasets*
[14] *Annexe 6-4.2 : Tableaux des régions identifiées dans les analyses intégrées (croisement des données)*

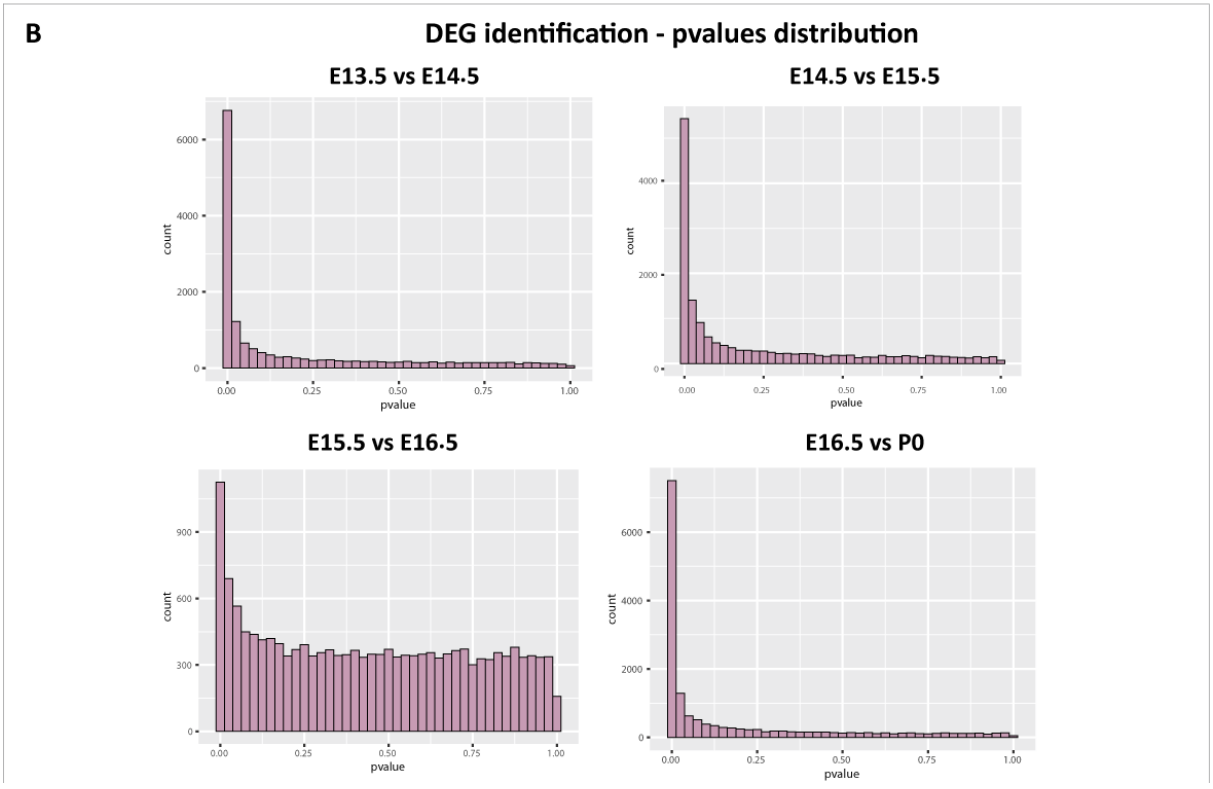
Table 7: Hypergeometric tests results, testing the over-representation of DMRs among differentially opened or closed regions (DOCR) observed during brain development, under physiological conditions.

		Nb of DOCR that overlap to capture regions			**Nb of capture regions associated to DOCR that overlap with DMRs + hypergeometric tests results**			
	Total nb of DOCR (mm9)	nb of regions identified in DOCR dataset (mm9)	nb of regions identified in capture dataset (mm9)	Total nb of association	Nb of regions that are both DMRs and DOCR	Total number of regions in the capture	Nb of capture regions that do not correspond to DOCR	Hypergeometric test results
E14.15_up	1442	303	**313**	315	5	58 611	58 298	0.029
E15.16_up	280	69	**75**	75	1		58 536	0.106
E14.15_down	1614	430	**449**	452	2		58 162	0.644
E15.16_down	420	78	**79**	83	0		58 532	0.443

Values highlighted in red are significant. These tests were done for comparing opened regions between different development stages: E14.15 versus E15.5 and E15.5 versus E16.5 . Hypergeometric tests were done using the number of methylome capture regions that are associated to DOCR (bold values highlight in grey), since DMRs could not be detected even if it exists, in DOCR that are not represented in methylome capture. When comparisons between datasets regions are done, only regions with an overlap are taken into account.

“Up regions” are regions that are significantly more opened for older developmental stage, compared to the younger developmental stage that is considered in the comparison. On the contrary, “Down regions” are regions that are significantly more closed during brain development.





Supp. Figure S5: Quality controls and correlation between RNA-seq samples.

(A) Pairwise scatterplots and matrix of pairwise Simple Error Ratio Estimate (*SERE*) values for RNA-seq samples. Comparisons were performed between two successive developmental stages. *SERE* statistic was used as a similarity index (the more dissimilar the samples are, the higher the SERE value is). For each pairwise comparison, variability between replicates is lower than those of samples from distinct stages. Nevertheless,

variability between replicates is higher than those observed for ATAC-seq data comparisons (Supp. Fig. S4). Inter- and intragroup variabilities are nearly equivalent for E15.5 and E16.5 samples. These similarities could explain the lower number of DEG identified for these sample comparisons, compared to other developmental stages comparisons (Supp. Fig. S3). Scatterplots and SERE values were obtained using *SARTools R* package.
(B) Raw pvalues distribution obtained for ATAC-seq samples pairwise comparisons allowing DEG identification. For all comparisons, pvalues distribution profiles are as expected (*i.e.* a globally uniform distribution with a peak around 0), meaning that DEG could be detected with edgeR statistic tool These pvalue histograms were obtained using *ggplot2 R* package, with data from *SARTools R* package.

Table 8: DMRs significantly associated to genomic *loci* that are more opened between E14.5 and E15.5, under physiological conditions.

Genomic *loci*	DMR chr:start-end (mm10 coordinates)	Meth. Diff (%)	Median pval.	Enhancers (H3K27ac) overlap	Promoters Overlap (bp)
Elmo1	chr13:20600469-20600956	+14.61	0.00310	488	.
Mapre2	chr18:23885054-23885296	+16.23	0.02509	243	.
Antxr1	chr6:87187138-87187332	-14.47	0.03766	195	.
Plcl2	chr17:50641357-50641748	+13.78	0.02776	392	.
Intergenic region	chr1:21726763-21727067	+7.54	0.03234	305	.

Meth.diff : differential methylation (%) observed between control (PBS-treated) and EtOH-treated groups. If percent is negative, it means that less methylation is observed upon binge drinking stress. ***Median pval.*** : median of pvalues observed for each relevant isolated CpG located in the DMRs (pvalue < 0.07, see Material & Method).

Interestingly, these genes were associated with GO terms of importance for brain development: for example, *Elmo1* with cell motility and cell migration; *Plcl2* with synaptic signalling, *Mapre2* with regulation of multicellular organization process and cytoskeleton protein; *Antxr1* with cell development, cell biogenesis, plasma membrane part, abnormal mouse morphology.

Our analysis of RNA-seq ENCODE data (Suppl. Fig. S3) revealed that DMRs were significantly associated with genes that were upregulated between E14.5 and E15.5, and between E15.5 and E16.6 under physiological conditions (Table 9).

DISCUSSION

The impact of PAE on the DNA methylome in the brain has been almost exclusively interrogated in the adult brain and the question whether perturbations of DNA methylation occur quickly after exposure has remained elusive. To the best of our knowledge, our study represents the first analysis of the effect of PAE on the DNA methylome in the developing neocortex, addressing the question of the early deposition of aberrant DNA methylation marks, at very short temporal distance from the exposure.

Table 9 : Hypergeometric tests results, testing the over-representation of DMRs among differentially expressed genes (DEG) observed during brain development, under physiological conditions.

			Nb of DEG that overlap to capture regions			**Nb of capture regions associated to DEG that overlap with DMRs + hypergeometric tests results**			
	Total nb of DEG (mm10)	Total nb of DEG (mm9)	nb of regions identified in DEG dataset (mm9)	nb of regions identified in capture dataset (mm9)	Total nb of association	Nb of regions that are both DMRs and DEG	Total number of regions in the capture	Nb of capture regions that do not correspond to DEG	Hypergeometric test results
E14.15_open	3426	3414	2662	**10 458**	10 708	96	58 611	48 153	0.008
E15.16_open	196	196	188	**1 465**	1 465	16		57 146	0.046
E14.15_close	2512	2500	2216	**4 678**	4 731	21		53 933	0.993
E15.16_close	327	324	304	**826**	831	6		57 785	0.408

Values highlighted in red are significant. These tests were done for comparing gene expression between different developmental stages E14.15 versus E15.5 and E15.5 versus E16.5. Hypergeometric tests were done using the number of methylome capture regions that are associated to DEG (bold values highlight in grey), since DMRs could not be detected even if it exists, in DEG that are not represented in methylome capture. When comparisons between datasets regions are done, only regions with an overlap are taken into account.

The main conclusion of our study is that at least some of these DNA methylation marks are deposited very early after exposure: indeed, we identify functional groups of genes of importance for neurodevelopment that had been associated with DMRs in the adult brain in mice that were prior exposed to alcohol *in utero*. Therefore, this early deposition is interestingly suggestive of the persistence of at least some DNA methylation marks from the prenatal period to the adult life. Since alterations in DNA methylation profile observed in mouse brain in response to PAE has been corroborated in peripheral tissues (cheek swabs) in cohorts of FASD children, these early DNA methylation modifications have putative importance in terms of functional impacts and potentials as relevant biomarkers.

Statistical intra- and inter-group variability

First, we observed that the intra-group variability was limited (Supp. Fig. 2). This is likely due to our sampling mode which aimed at minimizing inter-individual variability in the same group, by mixing half-cortices from four embryos from four different litters in the same NGS sample (Supp. Fig. S1F).

Second, we also observed that triplicates from the EtOH-treated group shown more differences between each other than do triplicates from control group. This might reflect the impact of the ethanol exposure, which as a stress, could disturb the methylome with some stochastic aspects. This is also what we observed, in our lab, in a paradigm of neuroinflammation (at stages equivalent to the third trimester of pregnancy in human) between samples within the exposed group (Schang et al., 2018).

Third, alcohol exposure has multiple potential effects on DNA methylation in terms of availability of precursors of the methyl group or potential redistribution of DNMTs in the genome. Our statistical analyses suggested that inter-group variability (control versus PAE-treated samples) is limited, and that PAE therefore might not majorly reshape the DNA methylome – at least in the captured regions – (Supp. Fig. 2). This observation could mean that, in spite of pleiotropic alcohol impacts, the DNA methylome is globally unchanged after such a binge-drinking-like mode of PAE and could suggest that protective mechanisms might exist, which would be interesting to unravel. Alternatively, this apparent robustness might only reflect the fact that DNA methylomic perturbations occur later after the last ethanol injection, probably in part by abnormal neuronal plasticity that could reshape methylome profile (Su et al., 2017).

Detection of individual differentially methylated cytosines

We detect few individual cytosines showing statistically differential methylation after PAE. Two of them, hypermethylated, reside into genes, *KALRN* and *Tiam2*, that involved in the pathway of guanine nucleotide exchange factors, which is important for neurodevelopment. In particular these genes participate to neuronal shape and polarity, axon growth and/or neuronal plasticity (Cahill et al., 2009; Honda et al., 2017). *KALRN* misexpression has been linked to neuropsychiatric disorders like schizophrenia and addiction, and this gene has been found, in genome-wide association studies, to be related to ADHD, and schizophrenia, all pathologies being relevant for FASD (reviewed in Remmers et al., 2014). Because the methylation of an individual CpG can affect gene expression (Xu et al., 2007), these DMCs could have either at an early stage - just after PAE during neurodevelopment - , or later - in the postnatal or even adult brain -, consequences on *KALRN* and *Tiam2* expression with deleterious effects on neuronal function. The formal demonstration of the

functional importance of these DMCs on gene expression needs to be brought, using epigenome editing approach, based on the CRISPR-dead Cas9 technology, for example.

DMRs early after PAE are mostly found in brain active enhancers

We have mainly focused on the identification and exploration of DMRs, which are, in general, more robustly associated with impacts on gene expression. We identified 432 DMRS, showing that DNA methylation is rapidly redistributed in the genome shortly after PAE. We found that these DMRs appearing just after PAE are not randomly distributed among the capture regions: indeed, in contrast to promoters and other genomic regions, which are unaffected, adult brain active enhancers are significantly over-represented, since more than 75 % of DMRs identified in our capture approach correspond to these genomic features. DMRs in these enhancers, that are particularly important for brain functions, could disturb their activities and explain part of brain defects associated to FASD.

DMRs are linked to genes showing dynamic, physiological changes in chromatin accessibility and gene expression

In line with the predominant location of DMRs in brain active enhancers, we found that DNA methylation changes occur in specific regions that are, under no stress conditions, physiologically modified at the time of exposure, in terms of chromatin accessibility and gene expression. First, DMRs are particularly enriched in regions whose chromatin show a gain in accessibility between E14.5 and E15.5, in physiological conditions, and the corresponding enhancers are associated with genes whose GO terms underline importance for brain development. Any perturbation of their methylation landscape by PAE might thus impair the physiological expression of these genes. Second, DMRs are also significantly associated with genes exhibiting expression upregulation either between E14.5 and E15.5 or E15.5 and E16.5. Notably, *Satb2*, one of the genes that show expression upregulation between E14.5 and E15.5 and between E15.5 and E16.5 in physiological conditions is associated, in our analysis, with a PAE-induced hypermethylated DMR located in an enhancer (DMR in intron 2 on 10). This gene was previously identified by Hashimoto-Torii et al (2011) as the most downregulated gene in a similar paradigm of PAE (ethanol injections between E14 and E16). It was also found dysregulated in human cortices isolated from gestational week E15-E18 fetuses that were *ex vivo* exposed to ethanol. In their mouse model, the perturbation of *Satb2* gene expression persists after birth (postnatal day 14, Hashimoto-Torii et al., 2011), suggesting that the presence of such DMR could potentially alter gene expression at temporal distance from the last ethanol injection. Similarly, the DMR-associated *Mapre2* gene, which encodes a microtubule-associated protein, shows physiological increased chromatin accessibility between E14.5 and E15.5, and upregulated expression between E15.5 and E16.5. DNA methylation changes in *Mapre2* DMR could therefore been followed by alteration in its expression. Interestingly, *Tiam2*, which contains one statistically significant DMC also belongs to the group of genes upregulated between E15.5 and 16.5.

Early DNA-methylation changes upon PAE affect imprinted genes and genes of protocadherin clusters, which have been previously identified as differentially methylated in adults after PAE

DMRs are over-represented in imprinted genes in our PAE paradigm. These DMRs are mainly located within the promoter regions and active enhancers, and in a quarter of the total of CpG

islands containing DMRs. We find that most of imprinted genes are ranked in the top of DMR-containing genes in terms of percentage of differential methylation. The other family of genes prominently altered at methylation level in our PAE paradigm is the clustered protocadherin genes, which are individually regulated by DNA methylation. Interestingly, these two family of genes, which are key for neurodevelopment and neuronal function, have been already identified as differentially regulated in the adult brain in mice (Laufer et al., 2013; Lussier et al., 2017) and in buccal swabs in cohorts of FASD children (Cobben et al., 2019; Laufer et al., 2015), therefore at temporal distance from the cessation of alcohol exposure. Alteration of DNA methylation in imprinted genes have been also historically identified at the *H19/Igf2* control region in mouse preimplantation embryos exposed *in utero* to alcohol and later collected at E10.5 (whole embryos and placentas; (Haycock, 2009).

Early alteration in the DNA methylation status of imprinted genes has also been suggested by Downing et al. (2011), who observed subtle decrease in DNA methylation and gene expression at the mouse *Igf2* locus at E9, four hours after prenatal alcohol exposure in whole embryos and placentas, which were ameliorated by diet supplementation with methyl-group precursors (Downing et al., 2011). Other studies have also identified differential methylation status in imprinted genes after *ex vivo* ethanol exposure of cultured mouse embryos at the early neurulation stage (after 44 hours of exposure, Liu et al., 2009), as well as in human embryonic stem cells exposed to ethanol for 24 or 48 hours (Khalid et al., 2014). These two *ex vivo* studies were suggestive that at least some imprinted genes could show early modifications in their DNA methylation profiles and thus, in line with our *in vivo* findings, in the neocortex.

Among the genes whose expression is physiologically upregulated between E14.5 and E15.5 and that are associated with DMRs in our PAE paradigm, we identify *Grb10,* whose specific expression from the paternally inherited allele is involved in adult behaviour (Dent and Isles, 2014; Perez et al., 2016). Interestingly, *Grb10* has already been identified as differentially methylated in adult mice, after PAE by Laufer et al (Laufer et al., 2013), also by Liu et al. (2009), as well as deregulated expression in response to very brief alcohol exposure in astrocytes (Pignataro et al., 2009). Along these upregulated genes associated to DMRs, we also identified several protocadherin genes, including *Pcdhgb4* and *Pcdhga12*. These observations suggest that expression of these particular genes could be affected both early and later after the PAE.

DMRs and transcriptional regulator-binding motif

We couldn't identify any specific transcriptional regulator-binding motif statistically over-represented within these DMR-associated enhancers. This is in marked contrast to the work by Laufer et al. (2013) who have identified the CTCF binding site (CCCTC) as significantly over-represented in DMRs observed long after the PAE. There might be several reasons for that: 1) our capture design might prevent the identification of such motif by not targeting these sequences ; 2) CTCF motif was observed after alcohol exposure at a different stage (third trimester in human in their case, second in our case) and in promoters; 3) Laufer et al. (2013), investigated DMRs in adult mice (postnatal day 70). Notably, the fact that we observed DMRs in protocadherin genes and imprinted genes, as major targets for associated DMRs (see below), both gene families being regulated by CTCF (Franco et al., 2014; Golan-Mashiach et al., 2012), might nevertheless been in favor of preferential DNA methylation changes occurring and maintained in regions containing CTCF motifs. Interestingly,

Khalid et al. (2014) have also identified specific transcription factor motifs in hESCs exposed to ethanol, including HSF consensus binding sites that have been found in association with CTCF on CTCF-occupied chromatin that gain localized acetylation upon another kind of stress, heat shock (Vihervaara et al., 2017).

CONCLUSION

Importantly, to the best of our knowledge, our data demonstrate for the first time, that DNA methylation alterations are detected in the cortex very close to the cessation of PAE in families of genes that have been previously identified as carrying or associated to DMRs in the adult brain, in human children or mouse models after PAE. These observations therefore strongly suggest that abnormalities in DNA methylation profile can persist at very long temporal distance from the end of exposure. Notably, this persistence might be more qualitative than quantitative in the sense that examination of CpG profile (for example by pyrosequencing) might reveal that a certain DMR globally exhibits differential methylation both at short and long temporal distance from the end of exposure, but with different percentage of individual cytosine methylation within the DMR (*e.g.* although globally the DMR might stay hypo- or hypermethylated, one given cytosine may show more variable methylation levels).

PERSPECTIVES

PAE could potentially affect DNA methylation of cytosines into non CpG-context, rapidly after the stress. DNA methylation in non-CpG context occurs postnatally in the human and mouse brain (He and Ecker, 2015; Lister et al., 2013) and is correlated to the increase in DNMT3A levels after birth. Since we observed an upregulation of DNMT3A after alcohol exposure both in *ex vivo* cell systems (Miozzo et al., 2018) and *in vivo* in the ventricular zone of mouse cortex (*Figure 1.8 in the introduction of thesis*), we suspect that PAE could induced precocious DNA methylation events in non-CpG contexts that could disturbed the interpretation of this mark by the cells of the developing brain in terms of gene expression. This exciting and potentially novel aspect of PAE impact deserves future investigation.

MATERIAL AND METHODS

Mice mating and alcohol exposure

C57Bl/6N murine females of 2- to 4-month of age were time-mated and assessed for mating based on the presence of a vaginal plug. The noon of day of vaginal plug was considered as embryonic day 0.5 (E0.5). For binge drinking stress, pregnant females received intra-peritoneal injections of ethanol (3g/kg, diluted in a final volume of 500 µL of PBS), at embryonic days E15, E15.5 and E16 (Fig. 1A). Control group received similarly PBS alone. Embryonic cortices were collected 2 hours after the last injection.

Tissue collection and genotyping

Embryonic cortices were harvested on ice, in cold L-15 medium (Leibovitz Gibco #11415-049). Only male cortices were used for methylome analysis. Three replicates were generated per group (1M, 2M and 3M samples correspond to ethanol-treated group, while 4M, 5M and 6M design the PBS-treated one, which is considered as the control group). In order to reduce inter-litter variability, each replicate was composed of 2 right hemi-cortices and 2 left-hemi cortices, from embryos from 4 distinct litters, see Supp. Fig. S1F.

Since sexual dimorphism was reported for behaviors of individuals exposed *in utero* to alcohol (Hellemans et al., 2010b), we determined the sex of each embryo and used only male cortices for methylome analysis. Genomic DNA was extracted from animal tails incubation (95°C for 1h) in the extraction buffer (25 mM NaOH, 0.2 mM EDTA) then neutralized in Tris-HCl 40mM, pH5 (vol/vol). Male embryos were identified by PCR genotyping using Ube1R (5'-CACCTGCACGTTGCCCTT-3') and Ube1F (5'-TGGATGGTGTGGCCAATG-3') primers that target *Ube1X* and *Ube1Y* genes (Sugimoto and Abe, 2007). Male DNA amplification gives rise to 2 amplicons (252bp from Ube1X and 334bp from Ube1Y).

ATAC-seq and RNA-seq ENCODE data

To define chromatin accessibility profile of the developing brain, ATAC-seq data time-course analysis was performed, using available public ENCODE data. These ATAC-seq experiments were done using embryonic and newborn forebrains of C57BL/6N mice (experiments from Bing Ren laboratory, UCSD). Developmental stages studied in duplicates were E13.5 (ENCODE accession: ENCFF401VUV (R1) + ENCFF898NRO (R2), ENCFF721LGJ (R1) + ENCFF777UK (R2)), E14.5 (ENCODE accession: ENCFF048MTG (R1) + ENCFF890LGM (R2), ENCFF633MTW (R1) + ENCFF666DRJ (R2)), E15.5 (ENCODE accession: ENCFF248PXW (R1) + ENCFF825UHO (R2), ENCFF906VXU (R1) + ENCFF500SXI (R2)), E16.5 (ENCODE accession: ENCFF058IAE (R1) + ENCFF765HUX (R2), ENCFF776GDQ (R1) + ENCFF588XZG (R2)) and birth (postnatal day 0 - P0, ENCODE accession: ENCFF197GTC (R1) + ENCFF209GGJ (R2), ENCFF296GZG (R1) + ENCFF664RZO (R2)).

To define transcriptomic profiles in the developing brain, RNA-seq data time-course analysis was performed, using available public ENCODE data. These RNA-seq experiments were done using embryonic and newborn forebrains of C57BL/6N mice (experiments of B. Wold, from Caltech laboratory). Developmental stages studied in duplicates were E13.5 (ENCODE accession: ENCFF235DNM (R1), ENCFF959PSX (R2)), E14.5 (ENCODE accession: ENCFF270GKY (R1) + ENCFF460TCF (R2), ENCFF126IRS (R1) + ENCFF748SRJ (R2)), E15.5 (ENCODE accession: ENCFF179JEC (R1), ENCFF891HIX (R2)), E16.5 (ENCODE accession: ENCFF931IVO (R1), ENCFF114DRT (R2)), and birth (postnatal day 0 - P0, ENCODE accession: ENCFF037JQC (R1) + ENCFF358MFI (R2), ENCFF447EXU (R1) + ENCFF458NWF (R2)).

Methylome capture composition.

Capture is composed of 58,611 chosen regions, based on ENCODE available data (for enhancers and promoters regions) or based on previous lab results ("other regions" group). Majority of captured regions corresponds to active enhancers in adult (8 weeks-old) mouse cortex (characterized by the H3K27ac histone mark, regions of +/- 500 bases from the middle of the peaks). Capture also includes mouse promoters regions (+/- 500 bases from the transcription start site - TSS). We estimated that more than 75% of known promoters regions are included in the capture. Due to repetitive sequences that prevent the design of specific probes of some regions, and also because of annotation database information differences, some promoters regions couldn't be studied. "Other regions" group corresponds to (**i**) HSF2 binding sites found in unstressed mice cortices of embryos at E16.5 development stage (results of a ChIP-seq experiment, R. El Fatimy, A.L. Mouël, I. Massaoudi) ; and to dynamic regions upon stress : (**ii**) differentially opened or closed regions identified in isolated oligodendrocyte precursors (O4+ cells) from cortices of 5 days-old mice, after an inflammatory stress from P1 to P5, a period equivalent to the third trimester of pregnancy in human (results of an ATAC-seq experiment, A.L. Schang and D. Sabéran-Djoneidi, (Schang et al., 2018); (**iii**) enhancers (characterized by the H3K4me1 histone mark, +/- 100 bases from the middle of the peaks) of adult (8 weeks-old) mouse cortex among genes (and at +/-20 kb from genes ends) that were differentially expressed in microglia (CD11B+ cells), at different stages (*i.e.* postnatal day 1 (P1) + P5 + P10 or at P5 + P10 or at P45) upon an inflammatory stress from P1 to P5 (results of a microarray experiment, A.L. Schang, Krishnan et al., 2017). For details about the number of regions in each of this sub-category of "other regions" group, please see Supp. Fig. S1A).

Methylome capture and high-throughput sequencing

To analyze the methylation of genomic regions that we selected (see above details about the design of regions selection, Fig. 1B and supp. Fig. S1A), a methylome capture was performed, using *SeqCapEpi Developer Medium Enrichment kit* (*NimbleGen*, *Roche*), following *SeqCap Epi Enrichment System* User's Guide and manufacturer's instructions. Briefly, gDNA samples were fragmented using Covaris technologies and non-directional libraries were prepared using KAPA Biosystems DNA Library Preparation Kits with *NimbleGen SeqCap* Adapter Kits. Then, bisulfite conversion of these DNA samples libraries was performed using the *Zymo EZ DNA Methylation-Lightning* kit, before a DNA amplification using *KAPA HiFi HotStart Uracil* and *ReadyMix*. Capture of these bisulfite-converted sample libraries was done by hybridization, using *SeqCap Developer M Enrichment* Kit, the *Seq*Cap *Epi Accessory* Kit, and the *SeqCap HE-Oligo* Kits. In order to reduce variability between samples, the six libraries of all samples were multiplexed and captured all together on beads. This multiplexing during the 'capture step' has already been successfully tested with similar capture approaches (Allum et al., 2015; van der Werf et al., 2015). DNA was then amplified again using the *KAPA HiFi HotStart ReadyMix*.

Captured and amplified bisulfite-converted gDNA fragments were sequenced using an *Illumina* sequencing instrument (Paired-end, 150bp NextSeq 500 High Throughput). All these steps were performed by *Institut du Cerveau et de la Moelle épinière* (ICM) sequencing platform.

Bio-informatic analysis of methylome capture

Bio-informatic workflow used to analyze methylome capture is described in detail in Supp. data Jupyter notebook n°1[15] and key figures of the analysis are shown in Supp. Table S3. Key steps of the analysis and associated tools are shown in Figure 1C.

Trimming, mapping, methylation count table, and visualization. Briefly, reads quality was verified using *FASTQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). To decrease methylation call errors from poor quality data, a trimming was carried out with *Trimmomatic* (Bolger et al., 2014) and *Cutadapt* (Martin, 2011). Trimmed reads were then paired-end mapped on the whole *Mus musculus* reference genome mm9, using *bismark* (Krueger and Andrews, 2011) which is adapted for alignment of bisulfite converted reads. Libraries were considered as non-directional for the mapping. Deduplication of reads was performed using *deduplicate_bismark* function (included in *bismark* suite), which was adapted for bisulfite converted reads that require special attention due to the loss of complementarity between strands after PCR amplification upon treatment (Krueger et al., 2012).

For each sample, number of methylated and unmethylated cytosines that covered a given cytosine site, limited to cytosines in CpG context, was obtained using *bismark_methylation_extractor* (BME) function (included in *bismark suite*). To prevent double counting of paired-end results, *no_overlap* option was selected for BME function processing. Then, to select only information that corresponds to capture regions, *bedtools intersect* (Quinlan and Hall, 2010) was used to intersect capture regions with mapping files, or with CpG count tables. To avoid removal of potential interesting sequences, at the border of capture regions, each region was extended up and downstream with 150 bases, before the intersections (extended capture regions were named *enlarged capture*). Mapping files restricted to the *enlarged capture* regions were sorted and indexed using *samtools suite* (Li et al., 2009) before their visualization on *IGV* (Robinson et al., 2011).

Detection of differentially methylated isolated CpG (DMCs). Data from BME were formatted on *R* (The R Core Team, 2018) in order to use *MethylKit* R package (Akalin et al., 2012), for detection of significant differentially methylated isolated CpG between control and alcohol-treated samples. For this statistical analysis, formatted data were filtered: cytosine positions displaying abnormal high read coverage (> 99.9th percentile) or low read coverage (number of reads < 10) were excluded from the analysis, to avoid bias

[15] *Annexe 6-3.1 : Jupyter notebook - methylome bioinformatic workflow*

(Krueger et al., 2012). Read coverage was also normalized before statistical analysis, using median to calculate scaling factor. Only methylation state of CpG sites that were covered in all samples was investigated for the analysis (corresponding to 1,259,111 positions). DMCs were considered as significant when difference of methylation state between the two groups was higher than 5%, for a qvalue threshold of 0.05.

Detection of differentially methylated regions (DMRs). We generated an *R* function (called *get_close_loci()*) to define DMRs (Supp Fig. S1B-D). This function combines neighbouring CpGs that share similar differential methylation state, (all hypo- or all hypermethylated CpGs, Supp Fig. S1B and S1D). Because *MethylKit* has been designed for methylome data and provides relevant statistical information about individual CpG, we used *MethylKit* statistical output as an input of *get_close_loci()* function (*i.e.* position of the CpG site, percent of methylation difference between groups for a given site, pvalue and qvalue). Data from BME were thus formatted on *R* (The R Core Team, 2018) before using *MethylKit* (Akalin et al., 2012). Formatted data were filtered and normalized, with the same parameters than those used for DMC detection. Only methylation state of CpG sites that were covered in all samples was investigated for the analysis (corresponding to 1,259,111 positions). To define DMRs with *get_close_loci()*, only CpG sites having a reliable methylation state must be used, since one non relevant CpG is sufficient to affect DMR detection (Supp. Fig. S1D). Thus a pre-selection of *relevant CpGs* was done, according to their pvalue (CpGs with pvalue < 0.07). This pvalue threshold was determined using comparisons of real and random datasets (see below, Supp. Fig. S1E, Supp. Table 2). To define DMRs, parameters were defined as follow: association of at least 5 CpG having a same methylation state, in a region of maximum 2000 bases, with a maximum distance of 100 bases between two successive selected CpG, according to pvalue threshold (pval < 0.07, Supp. Table 2).

Randomization to define appropriate parameters for DMR detection. Random datasets (n=2) were generated and compared to real dataset, to define parameters of *get_close_loci()* function that were appropriated for DMRs detection in the real data. To obtain these random data, methylation states from real samples, were randomly switched, **CpG position by CpG position,** between samples and replicates using our own *R* script (Supp Fig. S1E, see details in Supp. data Jupyter notebook n°1). Each methylation state is thus reassigned to a given sample, but methylation state between distinct CpG positions are not shuffled. This randomization assigns biologically plausible methylation values, actually observed at a given cytosine position (this is not the case if randomization was global, *i.e* redistribution of CpG methylation state, regardless of the CpG position, because there would be a risk of assigning, for example, high DNA methylation rates to a cytosine position that is always detected as non-methylated, regardless the conditions, PBS or EtOH-treatment). Randomization was done on filtered but unormalized real dataset. Read coverage normalization was performed after the randomization, using median to calculate scaling factor. To define appropriate parameters for DMR detection, different parameters were modulated and number of DMRs found in the two datasets (real versus random) was compared (Supp. Table 2). Similar results were obtained by using the two random datasets (Supp. Table 2).

Annotation. To be able to compare DMRs with other bioinformatic analysis (ATAC-seq and RNA-seq results), data obtained with mm9 coordinates were then converted with mm10 coordinates using *LiftOver* software (https://genome.ucsc.edu/cgi-bin/hgLiftOver), available on UCSC website), with default parameters. Annotation file, based on mm10 coordinates was obtained using *BiomaRt R* package (Durinck et al., 2009; Huang et al., 2009). DMRs were annotated in *R* according to this file information (see details in Supp. data Jupyter notebook n°1). A file combining all information (DMRs coordinates, annotation and statistical information) were generated (Supp. Datasets DMRs[16]). To estimate if DMRs are particularly located in a given capture categories (enhancers, promoters…), hypergeometric tests were done on *R* (using *phyper()* function), taking into account the number of each capture category that is represented in the methylome capture. Threshold chosen: pvalue < 0.05.

[16] *Annexe 6-4.1 : Tableaux de données - thèse - A. Duchateau - tab DMR432*

Motifs enrichment. Transcription factors binding site enrichment within the DMRs located in active enhancer regions (H3K27ac mark), were explored using *findMotifsGenome* (*HOMER* suite, see details in Supp. data Jupyter notebook n°1 ; Heinz et al., 2010).

Over-representation of Imprinted genes among DMRs. To estimate whether DMRs are particularly located in imprinted genes (IG), hypergeometric tests were done on *R* (using *phyper()* function). Threshold chosen: pvalue < 0.05. The exact number of capture regions that corresponds to imprinted genes are unknown, but we estimated that about 150 capture regions could be attributed to imprinted gene regions (there are almost 150 mouse IG, according to http://www.geneimprint.com/site/genes-by-species.Mus+musculus website, and we potentially have, at least one promoter for each of this gene in the capture). Using this approximation, results of *phyper* test (phyper(11,150,58611-150,432)) is equal to 1.51e-09, which is significant. Even with a over-estimation of capture regions that correspond to imprinted gene regions, results of *phyper* test is significant (*e.g.* if we considered that 300 capture regions corresponds to imprinted genes regions (*i.e.* two regions per IG), phyher(11,300,58611-300,432) = 2.89e-06).

Estimation of sodium bisulfite conversion efficiency. To ensure that sodium bisulfite treatment correctly converted unmethylated cytosines, we evaluated the conversion efficiency by looking at conversion rate of a spike-in DNA (*i.e.* unmethylated known sequences of non-mammalian DNA), which was added in each sample preparation. This spike-in DNA, corresponding to sequences of lambda phage, was mapped on an appropriate reference genome, using *bismark*. The efficiency of sodium bisulfite conversion is obtained using deduplicated spike-in DNA mapped reads, by calculating the ratio of converted reads (containing thymines) to the total number of reads covering these cytosines (Supp. Table S1).

Pyrosequencing

In progress.

Chromatin accessibility profile of the developing brain: ATAC-seq data time-course analysis

Bioinformatic workflow used to analyze ATAC-seq data is described in details in Supp. data Jupyter notebook n°2[17]. Key steps of the analysis and associated tools are showed in Figure 1C and key figures of the analysis are shown in Supp. Table S4. Quality of sequenced reads was verified for each sample using *FASTQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads having bad sequencing quality were trimmed using *Trimmomatic* (Bolger et al., 2014). Trimmed reads were then paired-end mapped on *Mus musculus* reference genome mm9, using *bowtie2* (Langmead and Salzberg, 2012). Then, reads were deduplicated using *samtools rmdup* (Li et al., 2009). Reads that mapped to mitochondrial chromosome were removed using tools on *Galaxy* platform (Afgan et al., 2016, default parameters), by splitting the mapped reads per chromosome using *bam-splitter tool* (Barnett et al., 2011). Then all the files were merged except the reads mapped on mitochondrial chromosome, using *merge-bam tool*. *MACS2* (Zhang et al., 2008) was then used to identify peaks corresponding to opened genomic regions, in each sample. For that, mapped reads from duplicates were merged, before this Peak calling. Then, table containing all opened regions, detected in at least one sample, was generated, using *bedtools multiinter* and *bedtools merge* (Quinlan and Hall, 2010). To obtain reads count at these regions, in each sample, *htseq-count* (Anders et al., 2015) was runned. This tool prevents double counting of paired-end results since files containing mapped reads were sorted by read name. To identify regions where chromatin accessibility significantly changes during development, pairwise comparisons of successive developmental stages were performed using *edgeR* (Robinson et al., 2010b), an *R* software (The R Core Team, 2018) and *Bioconductor* package (Gentleman et al., 2004). We used (with adaptations), script template called *template_script_edgeR.r* from ***Sartools*** *R* package (Varet et al., 2016), which implements some ***edgeR*** functions (see details in Supp. Data Jupyter notebook n°2). To be able to compare ATAC-seq

[17] *Annexe 6-3.2 : Jupyter notebook - ATACseq bioinformatic workflow*

results with others bioinformatic analysis (methylome capture and RNA-seq results), data obtained with mm9 coordinates were then converted with mm10 coordinates using *LiftOver* software (https://genome.ucsc.edu/cgi-bin/hgLiftOver), with default parameters. Annotation file, based on mm10 coordinates was obtained using *BiomaRt R* package (Durinck et al., 2009; Huang et al., 2009). Differentially open and closed regions (DOCR) observed during brain development were annotated in *R* according to this file information (see details in Supp. Data Jupyter notebook n°2). For each pairwise comparison, a file combining all information (DOCR coordinates, annotation and statistical information) were generated (Supp. Datasets DOCR[18]). To visualize data on *IGV* (Robinson et al., 2011), files containing mapped reads were sorted and indexed using *samtools* suite (Li et al., 2009).

Transcriptome profile of the developing brain: RNA-seq data time-course analysis

Bioinformatic workflow used to analyze RNA-seq data is described in detail in Supp. data Jupyter notebook n°3[19]. Key steps of the analysis and associated tools are showed in Figure 1C and key figures of the analysis are shown in Supp. Table S5. For all samples, sequenced read quality was verified for each sample using *FASTQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Adapter sequences and reads having bad quality were trimmed using *TrimGalore* (https://github.com/FelixKrueger/TrimGalore). Trimmed reads were then singled-end mapped using *STAR* on *Mus musculus* reference genome mm10, which was associated to its corresponding *Gencode* annotation during indexation, to improve accuracy of the mapping (Dobin and Gingeras, 2015; Dobin et al., 2013). For each sample, *htseq-count* (Anders et al., 2015) was used to obtain reads count of all genes contained in *Gencode* annotation file. To identify differentially expressed genes (DEG) during physiological brain development, pairwise comparisons of successive developmental stages were performed using *edgeR* (Robinson et al., 2010b). We used exactly the same script than those executed for ATAC-seq analysis (*i.e.* modified script from *Sartools R* package (Varet et al., 2016, see details in Supp. data Jupyter notebook n°3). Annotation file, based on mm10 coordinates was obtained using *BiomaRt R* package (Durinck et al., 2009; Huang et al., 2009) and was used to annotate DEG observed during brain development with *Unix* commands (see details in Supp. data Jupyter notebook n°3). For each pairwise comparison, a file combining all information (DEG coordinates, annotation, and statistical information) were generated (Supp. Datasets DEG[20]). To visualize data on *IGV* (Robinson et al., 2011), files containing mapped reads were sorted and indexed using *samtools* suite (Li et al., 2009).

Gene ontology analyses

In order to perform Gene ontology analysis, gene lists (based on gene symbol names) were defined for each distinct datasets (*i.e.* genes associated to DMRs that are located in (or close to) captured active enhancers (H3K27ac) and genes associated to DOCR and DEG). For DOCR and DEG, only genes that are modulated between E14.5 and E15.5 and between E15.5 and E16.5 were studied. Gene ontology analyses were performed with WEB-based GEne SeT AnaLysis Toolkit (GESTALT, http://www.webgestalt.org/, using default parameters, except for the minimum number of genes for a category (set to one in our analysis). The reference gene list used for each analysis was either the genome for DOCR and DEG or all the genes of the studied capture category (*i.e.* in case of the "DMRs enhancer" list: all the genes associated to H3K27ac histone mark that are represented in the capture).

[18] *Annexe 6-4.1 : Tableaux de données - thèse - A. Duchateau - tabs DOCR.*
[19] *Annexe 6-3.3 : Jupyter notebook - RNAseq bioinformatic workflow*
[20] *Annexe 6-4.1 : Tableaux de données - thèse - A. Duchateau - tabs DEG*

Integration of bio-informatic (DMRs, DOCR and DEG) datasets

Command lines executed to combine datasets and perform hypergeometric tests are described in details on Supp. data Jupyter notebook n°4[21]. Regions identified in several datasets are reported in Datasets Combine_data [22]).

Integration of DMRs with DOCR or DEG and integration of DOCR with DEG: To determine whether DMR occurs in regions that are differentially opened or closed (DOCR) or in genes that are differentially expressed (DEG) during physiological brain development or not, DMR and DOCR regions, or DMR and DEG regions, were compared using our own *R* function called *find_overlaps_AD_table()* - using mm10 coordinates of regions for the two datasets. Only overlapping regions were investigated by setting *tolerance* argument to 0 (close regions that don't share overlapping regions were not analyzed). The obtained regions were then annotated, using annotation that was already achieved in the two complete datasets.

Similarly, to determine whether DOCR are also DEG during physiological brain development or not, DOCR and DEG were compared.

Estimation of the proportion of DOCR and DEG in methylome capture: To observe overlaps between DOCR and DMR, sequences of DOCR must be in methylome capture, otherwise, it will not be possible to detect a DMR, even if it exists. To determine proportion of DOCR regions and DEG genes that are in methylome capture, DOCR or DEG and methylome capture regions were compared using our own *R* function called *find_overlaps_AD_table()*.Only overlapping regions were investigated by setting *tolerance* argument to 0 (close regions that don't share overlapping regions were not analyzed). Since mm9 coordinates were used for these comparisons, DEG regions with mm10 coordinates were converted to mm9 coordinates using *LiftOver* (https://genome.ucsc.edu/cgi-bin/hgLiftOver), with default settings, before comparisons.

Hypergeometric tests: To estimate whether DMR particularly localize to DEG or DOCR observed during physiological brain development or not, hypergeometric tests were done on *R* (using *phyper()* function), taking into account the number of DEG or DOCR that are represented in the methylome capture. Threshold chosen: pvalue < 0.05.

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. *44*, W3–W10.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., Mason, C.E., and others (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol *13*, R87.

Allum, F., Shao, X., Guénard, F., Simon, M.-M., Busche, S., Caron, M., Lambourne, J., Lessard, J., Tandre, K., Hedman, Å.K., et al. (2015). Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. Nature Communications *6*, 7211.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

[21] *Annexe 6-3.4 : Jupyter notebook – integration of DMRs, DOCR and DEG : bioinformatic workflow*

[22] *Annexe 6-4.2 : Intégration des données - thèse - A. Duchateau*

Bale, T.L., Baram, T.Z., Brown, A.S., Goldstein, J.M., Insel, T.R., McCarthy, M.M., Nemeroff, C.B., Reyes, T.M., Simerly, R.B., Susser, E.S., et al. (2010). Early Life Programming and Neurodevelopmental Disorders. Biological Psychiatry *68*, 314–319.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691–1692.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. Journal of Molecular Biology *203*, 971–983.

Bock, C. (2012). Analysing and interpreting DNA methylation data. Nature Reviews Genetics *13*, 705–719.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bourgeron, T. (2015). From the genetic architecture to synaptic plasticity in autism spectrum disorder. Nat. Rev. Neurosci. *16*, 551–563.

Burd, L., Cotsonas-Hassler, T.M., Martsolf, J.T., and Kerbeshian, J. (2003). Recognition and management of fetal alcohol syndrome. Neurotoxicology and Teratology *25*, 681–688.

Cahill, M.E., Xie, Z., Day, M., Photowala, H., Barbolina, M.V., Miller, C.A., Weiss, C., Radulovic, J., Sweatt, J.D., Disterhoft, J.F., et al. (2009). Kalirin regulates cortical spine morphogenesis and disease-related behavioral phenotypes. Proc. Natl. Acad. Sci. U.S.A. *106*, 13058–13063.

Carloni, S., Mazzoni, E., and Balduini, W. (2004). Caspase-3 and calpain activities after acute and repeated ethanol administration during the rat brain growth spurt. J Neurochem *89*, 197–203.

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642–655.

Clark, S.J., Statham, A., Stirzaker, C., Molloy, P.L., and Frommer, M. (2006). DNA methylation: bisulphite modification and analysis. Nat Protoc *1*, 2353–2364.

Cobben, J.M., Krzyzewska, I.M., Venema, A., Mul, A.N., Polstra, A., Postma, A.V., Smigiel, R., Pesz, K., Niklinski, J., Chomczyk, M.A., et al. (2019). DNA methylation abundantly associates with fetal alcohol spectrum disorder and its subphenotypes. Epigenomics *11*, 767–785.

Davies, W., Isles, A.R., Humby, T., and Wilkinson, L.S. (2008). What are imprinted genes doing in the brain? Adv. Exp. Med. Biol. *626*, 62–70.

Dent, C.L., and Isles, A.R. (2014). Brain-expressed imprinted genes and adult behaviour: the example of Nesp and Grb10. Mamm. Genome *25*, 87–93.

Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR: Mapping RNA-seq Reads with STAR. In Current Protocols in Bioinformatics, A. Bateman, W.R. Pearson, L.D. Stein, G.D. Stormo, and J.R. Yates, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp. 11.14.1-11.14.19.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Downing, C., Johnson, T.E., Larson, C., Leakey, T.I., Siegfried, R.N., Rafferty, T.M., and Cooney, C.A. (2011). Subtle decreases in DNA methylation and gene expression at the mouse Igf2 locus following prenatal alcohol exposure: effects of a methyl-supplemented diet. Alcohol *45*, 65–71.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc *4*, 1184–1191.

El Fatimy, R., Miozzo, F., Le Mouel, A., Abane, R., Schwendimann, L., Saberan-Djoneidi, D., de Thonel, A., Massaoudi, I., Paslaru, L., Hashimoto-Torii, K., et al. (2014). Heat shock factor 2 is a stress-responsive mediator of neuronal migration defects in models of fetal alcohol syndrome. EMBO Mol Med *6*, 1043–1061.

El Hajj, N., Dittrich, M., Böck, J., Kraus, T.F.J., Nanda, I., Müller, T., Seidmann, L., Tralau, T., Galetzka, D., Schneider, E., et al. (2016). Epigenetic dysregulation in the developing Down syndrome cortex. Epigenetics *11*, 563–578.

Franco, M.M., Prickett, A.R., and Oakey, R.J. (2014). The Role of CCCTC-Binding Factor (CTCF) in Genomic Imprinting, Development, and Reproduction1. Biology of Reproduction *91*.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

Gibbard, W.B., Wass, P., and Clarke, M.E. (2003). The neuropsychological implications of prenatal alcohol exposure. The Canadian Child and Adolescent Psychiatry Review.

Golan-Mashiach, M., Grunspan, M., Emmanuel, R., Gibbs-Bar, L., Dikstein, R., and Shapiro, E. (2012). Identification of CTCF as a master regulator of the clustered protocadherin genes. Nucleic Acids Res. *40*, 3378–3391.

Gräff, J., Kim, D., Dobbin, M.M., and Tsai, L.-H. (2011). Epigenetic Regulation of Gene Expression in Physiological and Pathological Brain Processes. Physiological Reviews *91*, 603–649.

Guerri, C., Bazinet, A., and Riley, E.P. (2009). Foetal Alcohol Spectrum Disorders and Alterations in Brain and Behaviour. Alcohol and Alcoholism *44*, 108–114.

Hashimoto-Torii, K., Kawasawa, Y.I., Kuhn, A., and Rakic, P. (2011). Combined transcriptome analysis of fetal human and mouse cerebral cortex exposed to alcohol. Proceedings of the National Academy of Sciences *108*, 4212–4217.

Hashimoto-Torii, K., Torii, M., Fujimoto, M., Nakai, A., El Fatimy, R., Mezger, V., Ju, M.J., Ishii, S., Chao, S., Brennand, K.J., et al. (2014). Roles of Heat Shock Factor 1 in Neuronal Response to Fetal Environmental Risks and Its Relevance to Brain Disorders. Cell Neuron *82*, 560–572.

Haycock, P.C. (2009). Fetal Alcohol Spectrum Disorders: The Epigenetic Perspective. Biol. Reprod. *81*, 607–617.

Haycock, P.C., and Ramsay, M. (2009). Exposure of Mouse Embryos to Ethanol During Preimplantation Development: Effect on DNA Methylation in the H19 Imprinting Control Region. Biology of Reproduction *81*, 618–627.

He, Y., and Ecker, J.R. (2015). Non-CG Methylation in the Human Genome. Annu Rev Genomics Hum Genet *16*, 55–77.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hellemans, K.G.C., Sliwowska, J.H., Verma, P., and Weinberg, J. (2010). Prenatal alcohol exposure: Fetal programming and later life vulnerability to stress, depression and anxiety disorders. Neuroscience & Biobehavioral Reviews *34*, 791–807.

Honda, A., Usui, H., Sakimura, K., and Igarashi, M. (2017). Rufy3 is an adapter protein for small GTPases that activates a Rac guanine nucleotide exchange factor to control neuronal polarity. J. Biol. Chem. *292*, 20936–20946.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44–57.

Ikonomidou, C., Bittigau, P., Ishimaru, M.J., Wozniak, D.F., Koch, C., Genz, K., Price, M.T., Stefovska, V., Hörster, F., Tenkova, T., et al. (2000). Ethanol-induced apoptotic neurodegeneration and fetal alcohol syndrome. Science *287*, 1056–1060.

Ishii, S., Torii, M., Son, A.I., Rajendraprasad, M., Morozov, Y.M., Kawasawa, Y.I., Salzberg, A.C., Fujimoto, M., Brennand, K., Nakai, A., et al. (2017). Variations in brain defects result from cellular mosaicism in the activation of heat shock signalling. Nature Communications *8*, 15157.

Jeltsch, A., and Jurkowska, R.Z. (2014). New concepts in DNA methylation. Trends in Biochemical Sciences *39*, 310–318.

Jones, K.L., and Smith, D.W. (1973). Recognition of the fetal alcohol syndrome in early infancy. Lancet *302*, 999–1001.

Kaminen-Ahola, N., Ahola, A., Maga, M., Mallitt, K.-A., Fahey, P., Cox, T.C., Whitelaw, E., and Chong, S. (2010). Maternal Ethanol Consumption Alters the Epigenotype and the Phenotype of Offspring in a Mouse Model. PLoS Genetics *6*, e1000811.

Kernohan, K.D., and Bérubé, N.G. (2010). Genetic and epigenetic dysregulation of imprinted genes in the brain. Epigenomics *2*, 743–763.

Khalid, O., Kim, J.J., Kim, H.-S., Hoang, M., Tu, T.G., Elie, O., Lee, C., Vu, C., Horvath, S., Spigelman, I., et al. (2014). Gene expression signatures affected by alcohol-induced DNA methylomic deregulation in human embryonic stem cells. Stem Cell Research *12*, 791–806.

Kleiber, M.L., Mantha, K., Stringer, R.L., and Singh, S.M. (2013). Neurodevelopmental alcohol exposure elicits long-term changes to gene expression that alter distinct molecular pathways dependent on timing of exposure. Journal of Neurodevelopmental Disorders *5*, 1.

Kleiber, M.L., Diehl, E.J., Laufer, B.I., Mantha, K., Chokroborty-Hoque, A., Alberry, B., and Singh, S.M. (2014). Long-term genomic and epigenomic dysregulation as a consequence of prenatal alcohol exposure: a model for fetal alcohol spectrum disorders. Frontiers in Genetics *5*.

Kodituwakku, P.W. (2007). Defining the behavioral phenotype in children with fetal alcohol spectrum disorders: a review. Neurosci Biobehav Rev *31*, 192–201.

Krishnan, M.L., Van Steenwinckel, J., Schang, A.-L., Yan, J., Arnadottir, J., Le Charpentier, T., Csaba, Z., Dournaud, P., Cipriani, S., Auvynet, C., et al. (2017). Integrative genomics of microglia implicates DLG4 (PSD95) in the white matter development of preterm infants. Nat Commun *8*, 428.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012). DNA methylome analysis using short bisulfite sequencing data. Nature Methods *9*, 145–151.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

LaSalle, J.M., Powell, W.T., and Yasui, D.H. (2013). Epigenetic layers and players underlying neurodevelopment. Trends in Neurosciences *36*, 460–470.

Laufer, B.I., Mantha, K., Kleiber, M.L., Diehl, E.J., Addison, S.M.F., and Singh, S.M. (2013). Long-lasting alterations to DNA methylation and ncRNAs could underlie the effects of fetal alcohol exposure in mice. Dis Model Mech *6*, 977–992.

Laufer, B.I., Kapalanga, J., Castellani, C.A., Diehl, E.J., Yan, L., and Singh, S.M. (2015). Associative DNA methylation changes in children with prenatal alcohol exposure. Epigenomics *7*, 1259–1274.

Lemoine, P., Harousseau, H., Borteyru, J.P., and Menuet, J.C. (1968). Les enfants de parents alcooliques. Anomalies observées. A propos de 127 cas.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and others (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Research *47*, W199–W205.

Light, S.E.W., and Jontes, J.D. (2017). δ-Protocadherins: Organizers of neural circuit assembly. Seminars in Cell & Developmental Biology *69*, 83–90.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905–1237905.

Liu, Y., Balaraman, Y., Wang, G., Nephew, K.P., and Zhou, F.C. (2009). Alcohol exposure alters DNA methylation profiles in mouse embryos at early neurulation. Epigenetics *4*, 500–511.

Lussier, A.A., Weinberg, J., and Kobor, M.S. (2017). Epigenetics studies of fetal alcohol spectrum disorder: where are we now? Epigenomics *9*, 291–311.

Lussier, A.A., Morin, A.M., MacIsaac, J.L., Salmon, J., Weinberg, J., Reynolds, J.N., Pavlidis, P., Chudley, A.E., and Kobor, M.S. (2018). DNA methylation as a predictor of fetal alcohol spectrum disorder. Clin Epigenet *10*, 5.

Markham, J.A., and Koenig, J.I. (2011). Prenatal stress: Role in psychotic and depressive diseases. Psychopharmacology *214*, 89–106.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. *17*, 10.

Matsunaga, Y., Noda, M., Murakawa, H., Hayashi, K., Nagasaka, A., Inoue, S., Miyata, T., Miura, T., Kubo, K.-I., and Nakajima, K. (2017). Reelin transiently promotes N-cadherin-dependent neuronal adhesion during mouse cortical development. Proc. Natl. Acad. Sci. U.S.A. *114*, 2048–2053.

Mattson, S.N., Crocker, N., and Nguyen, T.T. (2011). Fetal Alcohol Spectrum Disorders: Neuropsychological and Behavioral Features. Neuropsychology Review *21*, 81–101.

Miozzo, F., Arnoux, H., De Thonel, A., Schang, A.L., Saberan-Djoneidi, D., Baudry, A., Schneider, B., and Mezger, V. (2018). Alcohol exposure promotes DNA methyltransferase DNMT3A up-regulation through reactive oxygen species-dependent mechanisms. Cell Stress & Chaperones.

Molumby, M.J., Anderson, R.M., Newbold, D.J., Koblesky, N.K., Garrett, A.M., Schreiner, D., Radley, J.J., and Weiner, J.A. (2017). γ-Protocadherins Interact with Neuroligin-1 and Negatively Regulate Dendritic Spine Morphogenesis. Cell Rep *18*, 2702–2714.

O'Connor, M.J., and Paley, B. (2009). Psychiatric conditions associated with prenatal alcohol exposure. Developmental Disabilities Research Reviews *15*, 225–234.

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat. Genet. *19*, 219–220.

Olney, J.W., Tenkova, T., Dikranian, K., Qin, Y.-Q., Labruyere, J., and Ikonomidou, C. (2002). Ethanol-induced apoptotic neurodegeneration in the developing C57BL/6 mouse brain. Brain Res. Dev. Brain Res. *133*, 115–126.

Penn, N.W., Suwalski, R., O'Riley, C., Bojanowski, K., and Yura, R. (1972). The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. Biochemical Journal *126*, 781–790.

Perez, J.D., Rubinstein, N.D., and Dulac, C. (2016). New Perspectives on Genomic Imprinting, an Essential and Multifaceted Mode of Epigenetic Control in the Developing and Adult Brain. Annu. Rev. Neurosci. *39*, 347–384.

Phillips, G.R., LaMassa, N., and Nie, Y.M. (2017). Clustered protocadherin trafficking. Seminars in Cell & Developmental Biology *69*, 131–139.

Pignataro, L., Varodayan, F.P., Tannenholz, L.E., and Harrison, N.L. (2009). The regulation of neuronal gene expression by alcohol. Pharmacology & Therapeutics *124*, 324–335.

Popova, S., Lange, S., Burd, L., and Rehm, J. (2012). Health Care Burden and Cost Associated with Fetal Alcohol Syndrome: Based on Official Canadian Data. PLoS ONE *7*, e43024.

Popova, S., Lange, S., Shield, K., Mihic, A., Chudley, A.E., Mukherjee, R.A.S., Bekmuradov, D., and Rehm, J. (2016). Comorbidity of fetal alcohol spectrum disorder: a systematic review and meta-analysis. The Lancet *387*, 978–987.

Portales-Casamar, E., Lussier, A.A., Jones, M.J., MacIsaac, J.L., Edgar, R.D., Mah, S.M., Barhdadi, A., Provost, S., Lemieux-Perreault, L.-P., Cynader, M.S., et al. (2016). DNA methylation signature of human fetal alcohol spectrum disorder. Epigenetics & Chromatin *9*, 25.

Probst, A.V., Dunleavy, E., and Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. Nature Reviews Molecular Cell Biology *10*, 192–206.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Remmers, C., Sweet, R.A., and Penzes, P. (2014). Abnormal kalirin signaling in neuropsychiatric disorders. Brain Research Bulletin *103*, 29–38.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24–26.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences *103*, 1412–1417.

Schang, A.-L., Steenwinckel, J. van, Lipecki, J., Rich-Griffin, C., Woolley-Allen, K., Dyer, N., Charpentier, T.L., Schäfer, P., Fleiss, B., Ott, S., et al. (2018). Epigenome and transcriptome landscapes highlight dual roles of proinflammatory players in a perinatal model of white matter injury (Developmental Biology).

Schlotz, W., Godfrey, K.M., and Phillips, D.I. (2014). Prenatal origins of temperament: fetal growth, brain structure, and inhibitory control in adolescence. PLoS ONE *9*, e96715.

Schultz, M.D., He, Y., Whitaker, J.W., Hariharan, M., Mukamel, E.A., Leung, D., Rajagopal, N., Nery, J.R., Urich, M.A., Chen, H., et al. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. Nature *523*, 212–216.

Su, Y., Shin, J., Zhong, C., Wang, S., Roychowdhury, P., Lim, J., Kim, D., Ming, G., and Song, H. (2017). Neuronal activity modifies the chromatin accessibility landscape in the adult brain. Nat Neurosci *20*, 476–483.

Sugimoto, M., and Abe, K. (2007). X Chromosome Reactivation Initiates in Nascent Primordial Germ Cells in Mice. PLoS Genet *3*, e116.

Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M., and Mattick, J.S. (2009). Non-coding RNAs: regulators of disease. J. Pathol. *220*, 126–139.

The R Core Team (2018). R - A Language and Environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Thompson, B.L., Levitt, P., and Stanwood, G.D. (2009). Prenatal exposure to drugs: effects on brain development and implications for policy and education. Nature Reviews Neuroscience *10*, 303–312.

Varet, H., Brillet-Guéguen, L., Coppée, J.-Y., and Dillies, M.-A. (2016). SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. PLoS ONE *11*, e0157022.

Vihervaara, A., Mahat, D.B., Guertin, M.J., Chu, T., Danko, C.G., Lis, J.T., and Sistonen, L. (2017). Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat Commun *8*, 255.

van der Werf, I.M., Kooy, R.F., and Vandeweyer, G. (2015). A robust protocol to increase NimbleGen SeqCap EZ multiplexing capacity to 96 samples. PloS One *10*, e0123872.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J.A., Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proceedings of the National Academy of Sciences *104*, 12377–12382.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Partie 2 : Identification des cibles de HSF2 et de DNMT3A après EPA, dans le cortex embryonnaire murin.

Dans le but de cartographier, par immunoprécipitation de la chromatine suivie d'un séquençage (*ChIP-seq)*, les cibles génomiques de HSF2 et DNMT3A dans le cortex cérébral embryonnaire E16.5 murin, j'ai mis au point des protocoles de *ChIP* ciblant ces deux facteurs, adaptés à cet organe et à ce stade de développement. J'ai choisi de réaliser un *ChIP-seq* pilote, comme validation finale de mon protocole de *ChIP*. Pour cela, je n'ai utilisé que des cortex d'embryons murins ayant subi une EPA (voir les perspectives dans la discussion, chapitre 5). Bien qu'il soit possible de tirer quelques conclusions intéressantes de ce *ChIP-seq* pilote ciblant HSF2, je resterai prudente sur l'interprétation des résultats, notamment au sujet de la redistribution de ce facteur suite à l'EPA, puisque l'expérience n'a été conduite que sur une partie des échantillons.

1. Optimisation des protocoles de ChIP

Un protocole de *ChIP* ciblant HSF2 avait déjà été élaboré au sein de l'équipe (El Fatimy et al., 2014). Toutefois, le matériel de sonication ayant changé et les anticorps utilisés lors des études précédentes n'étant plus commercialisés, de nouvelles mises au point du protocole ont été nécessaires. Des expériences de *ChIP* suivies de *western blots* ont permis d'optimiser et de conforter la validité des protocoles de *ChIP* ciblant DNMT3A et HSF2. En particulier, ces expériences de *western blots* ont permis de :

- Vérifier que les conditions de sonication des échantillons, choisies pour obtenir des fragments d'ADN de tailles optimales pour le séquençage (Figure 4-2.1 A), n'affectent que peu l'intégrité des épitopes de DNMT3A et HSF2 (Figure 4-2.1 B).
- Tester différents anticorps afin de vérifier leur capacité à immuno-précipiter DNMT3A et HSF2 dans les conditions de *ChIP*. Concernant DNMT3A, deux anticorps distincts ont été testés en *ChIP* (*Abcam* ab2850 et *Novus* #64B1446 - NB120-13888). Une faible immunoprécipitation de l'isoforme 1 de DNMT3A est obtenue avec l'anticorps *Abcam*, alors que les deux isoformes de DNMT3A sont fortement immuno-précipités par l'anticorps *Novus* (Figure 4-2.1 C). Ce dernier anticorps a donc été choisi pour le reste des expériences. Un enrichissement en HSF2 est détecté après une *ChIP* réalisée avec l'anticorps SFI58 (Figure 4-2.1 C, anticorps produit et généreusement transmis par notre collaboratrice, Pr. Lea Sistonen, Université de Turku, Finlande ; Vihervaara et al., 2013). La spécificité de cet enrichissement a été confirmée à l'aide de cortex *Hsf2*KO, pour lequel aucune bande n'est visible pour HSF2 après immunoprécipitation (Figure 4-2.1 C).

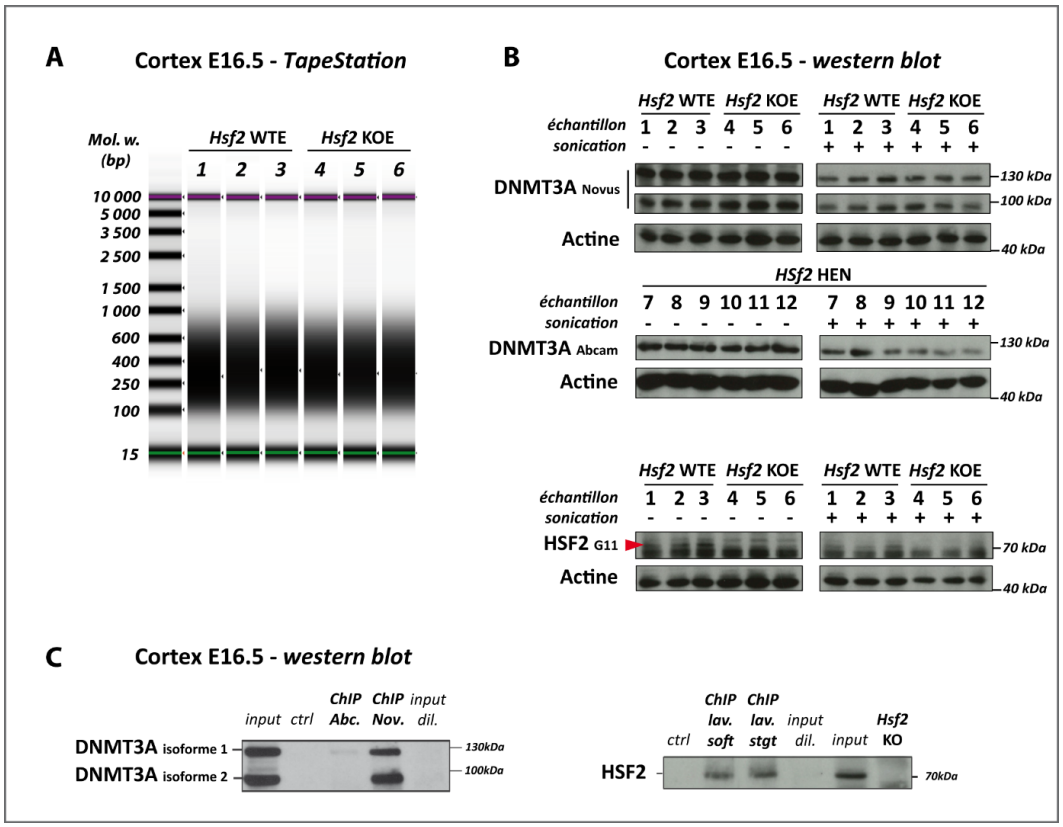


Figure 4-2.1 : Expériences validant le protocole de *ChIP* ciblant DNMT3A et HSF2.

(A) Vérification de la taille des fragments obtenus après sonication. *Electrophorèses TapeStation* réalisées à partir d'ADN purifié, issus de lysats de cortex embryonnaires E16.5 murins, après sonication de 8 minutes - 30sec ON / 30 sec OFF au Bioruptor Pico (Diagenode). La taille souhaitée des fragments d'ADN pour l'expérience de *ChIP-seq* est centrée autour de 300-400 pb. Expérience réalisée avec un kit *Agilent D5000 Screentape* (n > 3 expériences indépendantes). **(B) Vérification de l'intégrité des épitopes HSF2 et DNMT3A après sonication.** *Western blots* ciblant DNMT3A et HSF2, réalisés à partir de lysats de cortex embryonnaires E16.5 murins, avant (-) et après sonication (+) de 8 minutes - 30sec ON / 30 sec OFF au Bioruptor Pico (Diagenode). Même si les conditions de sonication choisies altèrent une fraction des épitopes, les protéines HSF2 (flèche rouge) et DNMT3A sont toujours détectées, donc reconnues par les anticorps en conditions dénaturantes. L'anticorps anti-DNMT3A *Novus* reconnait 2 isoformes de DNMT3A (d'environ 100 et 120 kDa), alors que l'anticorps *Abcam* ne reconnait pas l'isoforme courte (présentant une délétion en partie N-terminale), mais cible l'isoforme longue (d'environ 120 kDa). L'actine a été utilisée comme témoin de charge. (n > 3 expériences indépendantes). **(C) Vérification de l'enrichissement en HSF2 et DNMT3A après *ChIP*.** *Western blots* ciblant DNMT3A et HSF2, réalisés à partir de lysats de cortex embryonnaires E16.5 murins après *ChIP* et *reverse cross link*. Deux anticorps anti-DNMT3A ont été testés en *ChIP* (*Abcam* - Abc. et *Novus* - Nov.). Une faible immuno-précipitation de l'isoforme 1 de DNMT3A est détectée avec l'anticorps *Abcam*, alors que les deux isoformes de DNMT3A sont fortement immuno-précipitées par l'anticorps *Novus*. C'est donc l'anticorps *Novus* qui a été utilisé en *ChIP-seq*. Pour HSF2, après *ChIP* par l'anticorps SFI58, deux types de lavages ont été testés : des lavages *soft* - peu « *stringents* » - correspondants à ceux utilisés pour la *ChIP* ciblant DNMT3A, ou bien lavages « stringents », décrits dans la partie *Matériel et Methodes*. Comme l'enrichissement pour HSF2 est similaire dans les deux cas, les lavages stringents ont été choisis pour le *ChIP-seq* afin de limiter le bruit de fond. (n=1 pour DNMT3A, n=2 pour HSF2 en lavages *soft*, n=3 pour HSF2 en lavages stringents).

Abréviations : ctrl : lysat contrôle traité sans anticorps ; ***Hsf2 HEN*** : cortex de fœtus *Hsf2HET* naïf (non stressé – sans injection) ; ***Hsf2 KOE :*** cortex de fœtus *Hsf2*KO et traité à l'éthanol. ***Hsf2 WTE :*** cortex de fœtus *Hsf2*WT traité à l'éthanol ; ***input dil.*** : *input* dilué ; ***lav. soft*** : lavages soft ; ***lav. stgt*** : lavages stringents ; ***Mol. w. :*** *marqueur de poids moléculaires.*

Tableau 4-2.1 : Chiffres clés résumant les grandes étapes de l'analyse bio-informatique du *ChIP-seq*

		ChIP HSF2 HSF2 +/+	ChIP HSF2 HSF2 -/-	ChIP DNMT3A HSF2 +/+	ChIP DNMT3A HSF2 -/-	Input HSF2 +/+	Input HSF2 -/-
Données brutes	Nb de *reads*	11 529 990	12 163 515	21 349 976	25 838 977	48 706 490	47 853 488
Trimming	Nb de bases	876 279 240	924 437 140	1 622 598 176	1 963 762 252	3 701 693 240	3 636 865 088
	Après *trimming (%)*	93,29	96,52	97,10	97,65	97,21	97,13
Mapping	Nb de *reads* détectés étudié	11 528 388	12 162 451	21 348 997	25 838 338	48 704 582	47 849 756
	Nb de *reads* non alignés (%)	28,54	15,80	10,73	11,29	2,07	1,95
	Nb de *reads* alignés 1 fois (%)	48,22	56,41	60,02	60,84	64,37	64,90
	Nb de *reads* alignés >1 fois (%)	23,24	27,79	29,25	27,87	33,56	33,15
	Nb total de *reads* alignés	8 237 869	10 241 262	19 057 941	22 920 710	47 695 292	46 916 610
	Nb total de *reads* alignés (%)	71.46	84,20	89,27	88,71	97,93	98,05
Suppression des dupliquats	Nb de *reads* après déduplication	4 171 129	6 211 412	10 086 392	19 225 228	40 155 903	39 727 330
	Nb de *reads* éliminés (%)	49,4	39,3	47,1	16,1	15,8	15,3
Peaks calling	Nb pics détectés (**IP norm. à l'input**)	346	120	105	135		
Filtres de pics : suppression des faux positifs de l'IP (norm. à l'input)	Nb pics : IP norm. **sans pic KO**	301	26	22	35		
	Nb pics éliminés	45	94	83	100		
	Nb pics : IP norm. **sans *BL***	304					
	Nb pics éliminés	42					
	Nb pics : IP norm. **sans pic KO + BL**	282					
	Nb pics éliminés	19					
Conversion mm9 vers mm10	Nb de régions converties	280					

Abréviations : norm. : données normalisées par rapport à l'input de l'échantillon. ; **BL** : *blacklist*

Au vu de ces résultats favorables, qui ne permettent cependant pas de vérifier que la protéine d'intérêt est bien fixée à l'ADN fragmenté, un *ChIP-seq* a été réalisé sur un faible nombre d'échantillons, pour une expérience pilote, permettant de tester les conditions de *ChIP* et de séquençage (*cf Matériel et Méthode,* Figure 3.1 A). Cette expérience inclut les ADN immuno-précipités issus de cortex fœtaux murins E16.5 *Hsf2*WT ou *Hsf2*KO ayant subi l'EPA étudiée (échantillons « *ChIP* »), ainsi que l'ADN des chromatines non-immunoprécipitées correspondant (*inputs Hsf2*WT ou *Hsf2*KO).

2. Détection de régions enrichies en HSF2 et suppression des pics artéfactuels

Afin d'identifier les régions génomiques ciblées par DNMT3A et HSF2 après EPA, dans le cortex embryonnaire murin, j'ai effectué une analyse bio-informatique sur les échantillons séquencés (cf *Matériel et Méthode,* Figure 3.1B, Tableau 4-2.1). Les pics identifiés avec l'outil MACS2, ont été d'abord normalisés par rapport à l'*input* correspondant (qval. <0.05), puis les régions incluses dans la *blacklist*[23] ont été retirées.

Ainsi, concernant la *ChIP* ciblant DNMT3A après EPA, **105** pics potentiels ont été détectés dans le cortex *Hsf2*WT, contre **135** dans le cortex *Hsf2*KO (Figure 4-2.2). Néanmoins, une fois filtrée en éliminant les régions incluses dans la *blacklist*, il ne reste respectivement, que **22** et **35** pics, dans le cortex *Hsf2*WT ou *Hsf2*KO (Figure 4-2.2), ce qui semble indiquer que le *ChIP-seq* ciblant DNMT3A n'a pas fonctionné. Ce résultat est confirmé par la visualisation des données sur *IGV*, démontrant que les pics résiduels sont, d'une part, constitués de *reads* ayant un mauvais score d'alignement, et d'autre part, présents pour tous les échantillons, y compris au niveau de la *ChIP* ciblant HSF2, à partir de cortex *Hsf2*KO (Figure 4-2.3). Ces pics ne représentent donc pas un réel enrichissement de l'occupation de la chromatine par DNMT3A, mais vraisemblablement du bruit de fond non significatif, malgré un contexte plutôt favorable. En effet, la chromatine de cortex d'embryons murins est plus ouverte que celles d'autres types cellulaires ou d'autres stades de développement (résultats non publiés du laboratoire). De plus, la présence de DNMT3A est abondante dans le cerveau en développement, à ce stade de développement (Miozzo, 2014). Plusieurs raisons peuvent expliquer ce résultat. Il est notamment possible que les conditions de fixation et/ou de sonication ne conviennent pas. Ainsi, même si la protéine est détectée après sonication (Figure 4-2.1B) et enrichie après *ChIP* (Figure 4-2.1C), elle n'est peut-être plus fixée à l'ADN fragmenté.

[23] La *blacklist* correspond à une liste de régions génomiques retrouvées dans la majorité des *ChIP-seq* quel que soit l'anticorps utilisé. Par conséquent, ces régions sont des cibles non spécifiques du facteur testé.

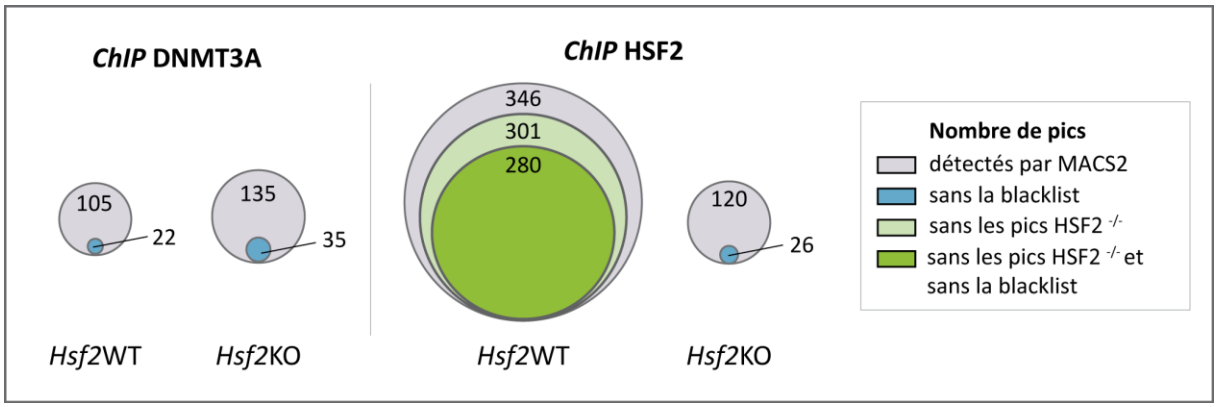


Figure 4-2.2 : Nombre de régions enrichies détectées dans chaque échantillon *ChIP*, avant et après filtrage.

Les régions enrichies ont été identifiées avec l'outil MACS2 (qval. < 0.05, Zhang et al., 2008), puis filtrées en retranchant les régions détectées dans l'échantillon *ChIP* ciblant HSF2 dans le cortex *Hsf2KO* (pour la *ChIP* ciblant HSF2, dans le cortex *Hsf2WT*) et à la *blacklist* (pour l'ensemble des *ChIP*).

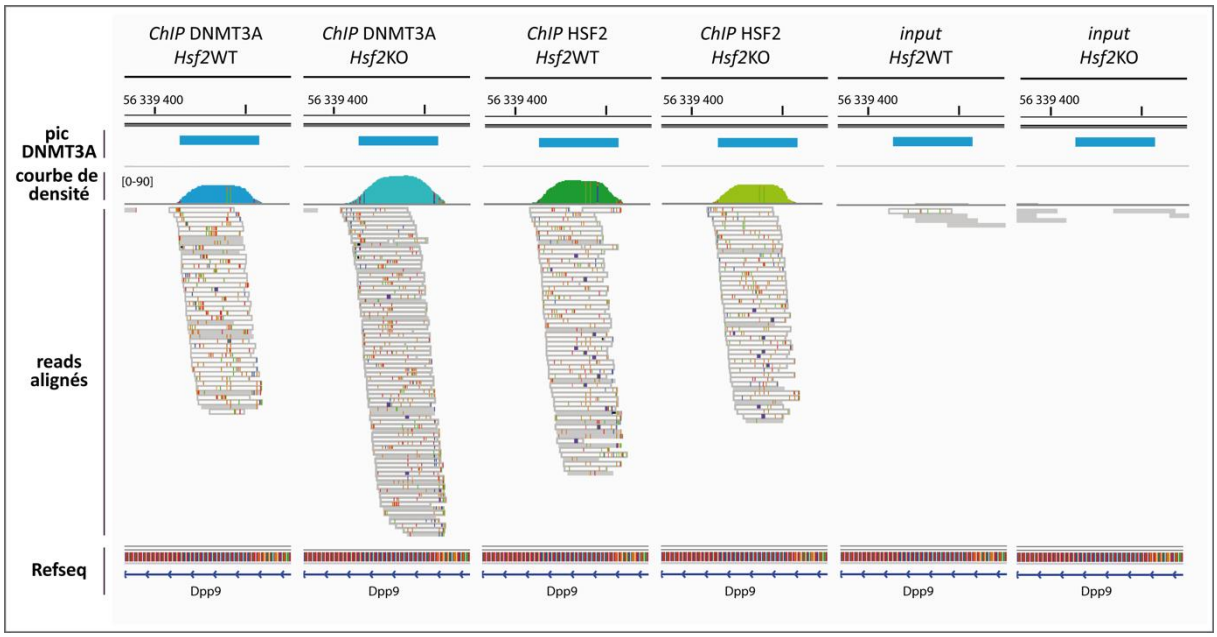


Figure 4-2.3 : Région montrant un enrichissement non significatif d'occupation par DNMT3A

Visualisation des *reads* alignés et du pic d'occupation par DNMT3A après EPA, au niveau du gène *Dpp9* pour l'ensemble des échantillons *ChIP* et *inputs* du *ChIP-seq* pilote. Le pic d'occupation par DNMT3A détecté par l'outil MACS2 (qval. < 0.05, Zhang et al., 2008), ne semble pas correspondre à un enrichissement réel, car ce pic est également présent pour l'échantillon de la *ChIP* ciblant HSF2, à partir de cortex *Hsf2*KO, d'une part, mais aussi parce que la plupart des *reads* alignés à cet endroit a un mauvais score d'alignement (les *reads* représentés en blanc) et/ou présentent de nombreux mésappariements (traits colorés au sein des *reads* alignés). L'échelle de couverture (valeur [0-90]), est identique pour l'ensemble des échantillons visualisés. Visualisation obtenue avec le logiciel IGV (Robinson et al., 2011).

Tableau 4-2.2 : Exemples de gènes connus comme cibles de HSF2 ou connus dans d'autres modèles d'EPA

Nom du gène	Cibles de HSF2 (tissu/cellules)	Conditions	Réf.	Altération après EPA	Contexte	Réf.
Abcc5	-	-	-	Hyperméthylation	Méconium humain - EPA variées	**(Frey et al., 2018)**
Acot7	• Spermatocytes murins • Cellules K562 humaines en mitose	Basales *HS* 30Min à 42°C	**(Korfanty et al., 2014)** **(Vihervaara et al., 2013)**	Répression de l'expression	Cellules NCCIT - carcinome embryonnaire humain - cultivées *in vitro* avec 50 mM (≈0.23%) d'EtOH dans le milieu de culture	**(Halder et al., 2014**)
Acvr1c	Spermatocytes murins	*HS* 5-20Min à 43°C	**(Korfanty et al., 2014)**	-	-	-
Arhgap32	Spermatocytes murins	Basales, perte de liaison après *HS*	**(Korfanty et al., 2014)**	-	-	-
B4galt5	Spermatocytes murins	*HS* 5-20Min à 38°C	**(Korfanty et al., 2014)**	Répression du gène corrélée à l'augmentation de l'expression du *miRNA* régulateur	Cerveaux murins adultes (P60) ayant subi une EPA de type *binge* drinking à une période équivalente au 3ème trimestre de grossesse chez l'humain : 2 injections sous-cutanées (0h et 2h) d'EtOH (2.5 g/kg) dans 0.15M de solution saline entre P4 et P7	**(Kleiber, 2015)**
Basp1	Spermatocytes murins	Basales, perte de liaison après *HS*	**(Korfanty et al., 2014)**	-	-	-
Blcap	-	-	-	Augmentation de l'expression	Placentas humains issus de femmes sud-africaines buvant entre 2 et 3 verres / jour	**(Carter et al., 2018)**
Dcn	-	-	-	Répression de l'expression (mâles)	Foies murins, modèle d'alcoolisation paternel : consommation volontaire chronique d'alcool, (solution contenant 10% (w/v) d'EtOH)	**(Chang et al., 2017)**
Hsp90aa1 ou Hsp90 ab1	• Cortex E16.5 murins • Cellules Jurkat humaines	Basales, fixation augmentée après EPA Basales	**(El Fatimy et al., 2014)** **(Wilkerson et al., 2007)**	Augmentation de l'expression	Cortex embryonnaires murins E16.5, modèle d'EPA par consommation volontaire chronique d'alcool (CAI : *chronic alcohol intoxication*), par nourriture semi-liquide contenant 70mg/ml d'éthanol servie *ab libidum* à des femelles gestantes à partir du 7.5ème jour de gestation.	**(El Fatimy et al., 2014)**

Tableau 4-2.2 (suite) : **Exemples de gènes connus comme cibles de HSF2 ou connus dans d'autres modèles d'EPA**

Nom du gène	Cibles de HSF2 (tissu/cellules)	Conditions	Réf.	Altération après EPA	Contexte	Réf.
Dock2	-	-	-	Répression de l'expression due aux altérations de la méthylation de l'ADN	Cellules souches humaines H1 et H9 traités avec 20 à 50 mM d'EtOH pendant 24 à 48h.	**(Khalid et al., 2014)**
Prdm16	-	-	-	Hyperméthylation	Méconium humain - EPA variées	**(Frey et al., 2018)**
Usp29	-	-	-	Hypométhylation de la région ICR (*imprinting control region*) du cluster Peg3 (incluant les gènes Peg3, Zim2 et Usp29)	Cellules sanguines et cellules buccales de 87 enfants (de 1 à 16 ans) diagnostiqués avec un FAS et 58 individus contrôles (de 17 à 26 ans), vivant et originaire d'Afrique du Sud	**(Masemola et al., 2015)**

Recherches non exhaustives parmi les données de la littérature.

Concernant la *ChIP* ciblant HSF2 après EPA, parmi les **346 régions** identifiées lors de la recherche de régions enrichies, **280 régions** semblent réellement fixées par HSF2 après EPA (Figure 4-2.2, Annexe 6-4.1 - onglet ChIP280). En effet, des pics non spécifiques sont détectés lors de la recherche de régions enrichies, comme en atteste la *ChIP* réalisées dans le cortex d'embryons *Hsf2*KO, pour laquelle **120** pics sont détectés, alors que la protéine HSF2 n'est pas présente dans l'échantillon (Figure 4-2.2). Ces pics aspécifiques ont été éliminés de l'échantillon d'intérêt (*ChIP* HSF2*,* cortex *Hsf2*WT), en recoupant les régions identifiées dans le cortex *Hsf2*WT, à celles détectées dans le cortex *Hsf2*KO.

Au vu de ces résultats, la suite de l'analyse a été réalisée à partir des données ainsi filtrées, issues exclusivement du *ChIP-seq* ciblant HSF2.

3. HSF2 se lie majoritairement au niveau de régions intergéniques et introniques, possédant un HSE.

D'après les données de la littérature, le facteur HSF2 se fixe à l'ADN au niveau de motifs HSE (*Heat Shock Element*), caractéristiques des facteurs HSFs (*cf* Chapitre 1, introduction). La présence de ce type de motif est donc attendue au niveau des régions détectées dans le *ChIP-seq* de cortex *Hsf2*WT. La recherche d'enrichissement de motifs avec le logiciel HOMER[24] que j'ai effectuée, montre que les deux motifs les plus significativement enrichis au sein de ces séquences sont des motifs HSE (Figure 4-2.4). L'enrichissement considérable pour ces deux motifs HSE, présents respectivement dans 88.57% (pval. = 1e-338) et 76.64% (pval. = 1e-289) des régions identifiées, conforte la validité du protocole de *ChIP-seq* ciblant HSF2. De façon notable, nous avons trouvé un motif HSE tripartite et un motif HSE quadripartite, rappelant ceux que l'équipe avait identifiée par une approche Selex dans des cellules embryonnaires (Manuel et al., 2002). Ces motifs HSE comportent la séquence 5'-TTCTAGAA-3' et diffèrent notablement du motif bipartite trouvé à partir de la protéine HSF2 recombinante, synthétisée *in vitro* (Kroeger and Morimoto, 1994). De tels motifs tripartite ou quadripartite sont compatibles avec la liaison d'un trimère HSF. En conditions physiologiques, HSF2 se présente précisément majoritairement sous forme d'homotrimère, dans le cortex en développement, alors que HSF1 est monomérique et ne montre pas de capacité à lier l'ADN dans ce tissu (El Fatimy et al., 2014). Toutefois, lors d'une EPA chronique (*chronic alcoholic intoxication, CAI*[25]), HSF1 est activé et forme un hétérotrimère avec HSF2, capable se fixer à l'ADN (El Fatimy et al., 2014). La liaison d'un hétérotrimère HSF1-HSF2, au niveau des deux motifs HSE que nous observons dans le *ChIP-seq* ciblant HSF2 après EPA, est dont également envisageable.

[24] HOMER *findMotifsGenome*, détails disponibles dans l'annexe 6-3.5 Workflow du *ChIP-seq* pilote

[25] Nourriture semi-liquide contenant 70mg/ml d'éthanol servie *ab libidum* à des femelles gestantes à partir du 7.5ème jour de gestation.

	Nom du facteur + tissu ou lignée cellulaire	p-value	nb séq. d'intérêt avec motif	% séq. d'intérêt avec motif	% séq. de l'univers avec motif
	HSE (HSF) - Striatum	1,00E-338	248	88.57	2.92
	HSE (HSF) - HepG2	1,00E-289	209	74.64	1.96
	ZBTB12 - HEK293	1,00E-43	93	33.21	5.78
	NF1-halfsite - LNCaP	1,00E-11	102	36.43	18.88
	NeuroD1 - Islet	1,00E-10	60	21.43	8.58
	CTCF - CD4+	1,00E-09	22	7.86	1.47
	Atoh1 - Cerebellum	1,00E-07	68	24.29	12.07
	AR-halfsite - LNCaP	1,00E-06	168	60.00	45.10
	NeuroG2 - Fibroblast	1,00E-06	84	30.00	17.96
	Olig2 - Neuron	1,00E-05	103	36.79	24.19
	HIC1 - Treg	1,00E-05	97	34.64	22.39
	Tlx - NPC	1,00E-05	30	10.71	4.10
	PRDM10 - HEK293	1,00E-05	39	13.93	6.37
	NF1 - LNCAP	1,00E-05	25	8.93	3.17
	BORIS - K562	1,00E-05	21	7.50	2.42
	TCF4 - SHSY5Y	1,00E-05	80	28.57	17.84
	PR - T47D	1,00E-04	113	40.36	28.44
	TEAD4 - Tropoblast	1,00E-04	53	18.93	10.39
	Brn1 NPC	1,00E-04	29	10.36	4.54
	SCL - HPC7	1,00E-04	176	62.86	50.99
	Ascl1 - Neural Tubes	1,00E-03	69	24.64	16.61
	Oct2 - Bcell	1,00E-03	27	9.64	4.74
	Oct4 - mES cells	1,00E-03	35	12.50	7.05
	RFX - K562	1,00E-03	9	3.21	0.88

Figure 4-2.4 : Motifs connus de séquences de fixation de facteurs de transcription enrichis dans les séquences ciblées par HSF2 après EPA.

Au total, 24 motifs nucléotidiques reconnus par des facteurs de transcription sont significativement enrichis dans les régions où le facteur HSF2 est lié après EPA. Les deux motifs les plus enrichis sont deux motifs HSE, spécifiques des facteurs HSF, ce qui conforte la validité du protocole de *ChIP-seq*. Analyse effectuée avec le logiciel HOMER *findMotifsGenome* (Heinz et al., 2010), pour lequel un enrichissement est estimé par rapport à un univers (ou *background*) composé de 46 860 séquences choisies aléatoirement dans le génome de référence. Le seuil de significativité de la *p-value* a été défini à 0.02 en réalisant une analyse similaire avec des séquences choisies aléatoirement au sein du génome de référence comparée à un univers composé de séquences choisies elles aussi aléatoirement au sein du génome de référence (cf détails en Annexe 6-3.5).
Abréviations : séq. : *séquences*

Par ailleurs, des cibles de HSF2, déjà connues dans divers contextes physiopathologiques sont présents parmi les régions identifiées dans le *ChIP-seq*, comme par exemple les gènes *Hsp90aa1* et *Hsp90ab1* (*Heat Shock Protein 90,* Figure 4-2.5 , Tableau 4-2.4, (El Fatimy et al., 2014; Korfanty et al., 2014; Wilkerson et al., 2007), *Acot7* (*Acyl-CoA Thioesterase 7*, Figure 4-2.5 , Tableau 4-2.2, Korfanty et al., 2014; Vihervaara et al., 2013), *St13* (*ST13 Hsp70 Interacting Protein,* Vihervaara et al., 2013),

ou *Fkbp4* (*FKBP Prolyl Isomerase 4,* (Vihervaara et al., 2013). De plus, certains gènes identifiés comme cibles de HSF2 dans notre *ChIP-seq*, font partis de catégories de gènes régulées par les HSF. C'est le cas notamment du gène codant la protocadhérine 7 (*Pcdh7*, Figure 4-2.5), appartenant à la famille des cadhérines qui participe à l'adhésion cellulaire et dont certains gènes sont régulés positivement par HSF2, en conditions basales ou après stress protéotoxiques, dans des cellules humaines K562. (Joutsen et al., 2018).

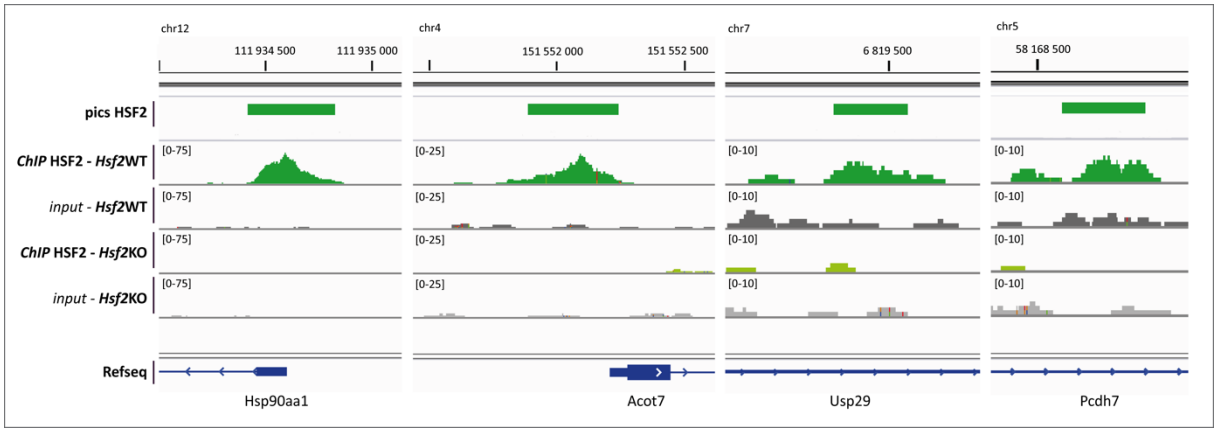


Figure 4-2.5 : Régions enrichies pour l'occupation par HSF2.
Visualisation des *reads* alignés et des pics d'occupation par HSF2, au niveau des gènes *Hsp90aa1*, *Acot7*, *Usp29*, *Pcdh7*, pour les échantillons *ChIP* ciblant HSF2 et *inputs* du *ChIP-seq* pilote. Les pics HSF2 détectés par l'outil MACS2 (qval. < 0.05 , Zhang et al., 2008) dans l'échantillon *ChIP* ciblant HSF2, à partir de cortex *Hsf2*WT, semblent correspondre à des enrichissements réels, car ces pics ne sont présents ni dans la *blacklist*, ni dans l'échantillon de la *ChIP* ciblant HSF2 effectuée à partir de cortex *Hsf2*KO. L'échelle de couverture de chaque échantillon (valeur indiquée entre crochets) est identique pour un gène donné. Visualisations obtenues avec le logiciel IGV (Robinson et al., 2011). ***Abréviations*** *:* chr : chromosome.

Afin de caractériser l'ensemble des régions identifiées, une annotation syntaxique des régions a été réalisée. Les régions fixées par HSF2 après EPA, sont majoritairement intergéniques (45% des régions) et introniques (40,7% des régions, Figure 4-2.6). La plupart des régions génomiques étant intergéniques, la présence majoritaire de HSF2 dans ce type de régions ne semble pas relever d'un enrichissement particulièrement remarquable. Au contraire, la proportion de régions intergéniques ciblée par HSF2, est sous-représentée, en comparaison avec la distribution des éléments génomiques issue du génome global (45% de régions intergéniques ciblées par HSF2, alors que la proportion de régions intergéniques dans le génome est de 66% - Figure 4-2.6). En revanche, la répartition de HSF2, au sein des introns semble plus caractéristique (Figure 4-2.6).

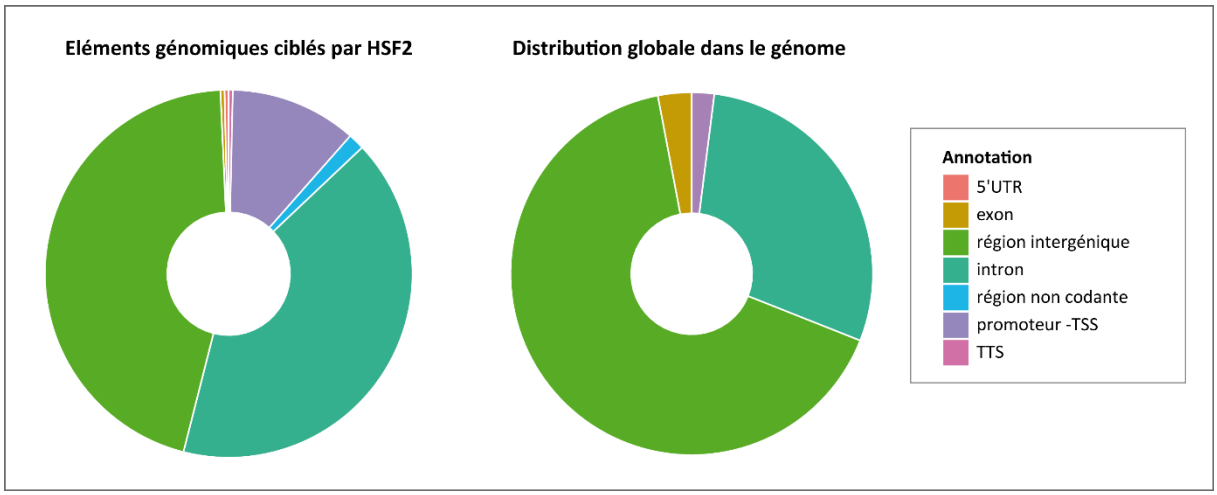


Figure 4-2.6 : Eléments génomiques ciblés par HSF2 après EPA.

(Panneau de gauche) Répartition génomique des 280 régions ciblées par HSF2 après EPA, annotées avec l'outil *homer_annotatePeaks* sur Galaxeast (www.galaxeast.fr). HSF2 se fixe majoritairement au niveau de régions intergéniques (126/280), introniques (114/280) ou promotrices (31/280). (Panneau de droite) Répartition globale au sein du génome indiquée pour comparaison (66% régions intergéniques, 29% introns, 3% exons, 2% promoteurs). Cette distribution globale est extraite de la publication de Achour et collaborateurs (2015), en considérant que les régions régulatrices en amont du gène (de -20 à -1kb du site d'initiation de la transcription - TSS), décrites dans cette publication, correspondent à des régions intergéniques, afin de conserver des catégories similaires (Achour et al., 2015).
Abréviations. ***TSS*** : région entourant le site d'initiation de la transcription, de -1kb à +100pb par rapport à ce site. ***TTS*** : région entourant le site de terminaison de la transcription, de -100pb à +1kb par rapport à ce site. ***5'UTR*** : régions transcrites mais non traduites, situées en 5' du site d'initiation de la traduction.

Plusieurs fonctions biologiques ont pu être attribuées aux introns (épissage alternatif (Brett et al., 2002), formation des nucléosomes (Levitsky et al., 2001), organisation de la chromatine (Chernov et al., 2002; Glazko et al., 2003; Schwartz et al., 2009), régulateur positif de l'expression génique (Callis et al., 1987; Gruss et al., 1979), contrôle du transport de l'ARNm (Palazzo et al., 2007; Valencia et al., 2008), *etc*, revue dans Jo and Choi, 2015). Il est intéressant de noter que HSF2 se fixe principalement au niveau du premier et deuxième intron des gènes cibles (respectivement 48/114 et 24/114 introns, Figure 4-2.7). Certes, les premiers introns de nombreux gènes sont souvent plus longs que les autres introns (Bradnam and Korf, 2008; Hong et al., 2006), ce qui augmente la probabilité d'observer HSF2 dans cette région. Cependant, l'étude des introns des gènes humains, menée par Park et collaborateurs (2014) a montré que, contrairement aux autres introns, les premiers introns possèdent des séquences très conservées, qui sont corrélées positivement à la présence de marques épigénétiques régulatrices (H3K4me1, H3K4me3, Park et al., 2014). De plus, le premier intron étant situé à proximité du promoteur et du site d'initiation de la transcription, les facteurs associés à ce premier intron sont plus susceptibles d'avoir un effet sur la régulation de l'activité promotrice que ceux situés plus en aval (Barrett et al., 2012). Ainsi, la fixation de HSF2 au niveau de ces introns après l'EPA pourrait moduler l'expression des gènes associés.

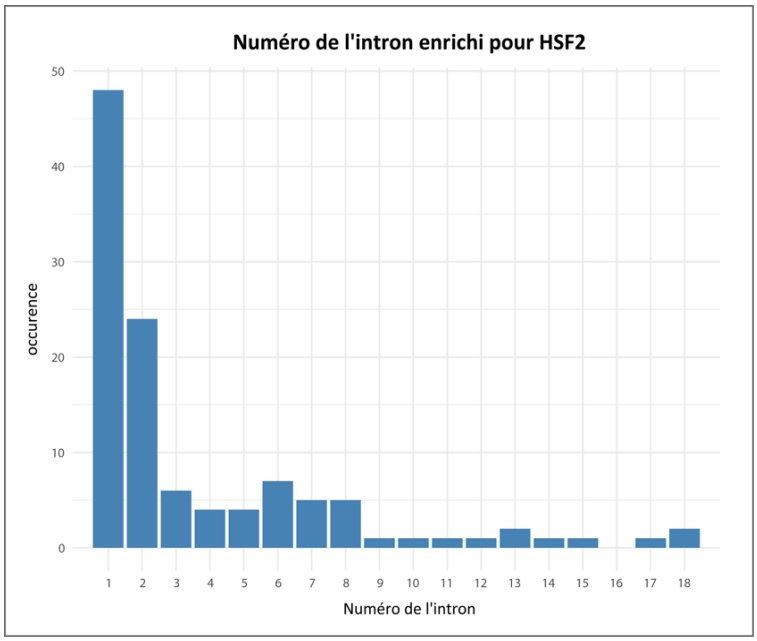


Figure 4-2.7 : Position de l'intron où HSF2 est fixé après EPA.
Analyse réalisée sur les 114 régions introniques fixées par HSF2 après EPA. HSF2 se lie majoritairement au niveau du premier ou du deuxième intron. L'annotation des régions a été réalisée avec l'outil *homer_annotatePeaks* sur Galaxeast (www.galaxeast.fr).

Par ailleurs, une proportion non négligeable de régions ciblées par HSF2 correspond à des séquences promotrices (31/280 régions, soit 11% des cibles de HSF2, alors que le génome contient 2% de promoteurs, Figure 4-2.7). Or, la fixation d'un facteur de transcription comme HSF2 dans une séquence promotrice, peut moduler l'expression du gène. L'équipe a montré par exemple dans des cellules humaines K562, en conditions physiologiques, que HSF2 se fixe sur les séquences promotrices du gène *Cdk5r1* (*cyclin-dependent kinase 5, regulatory subunit 1*, aussi appelé *p35,* Chang et al., 2006) et module son expression, essentiel à la migration neuronale. De même, la fixation de HSF2 au niveau des promoteurs des gènes *Hsp90*, *Hsp27*, et *c-fos* (*FBJ osteosarcoma oncogene*), en conditions physiologiques, module leur expression dans des cellules humaines Jurkat (Wilkerson et al., 2007, Tableau 4-2.2).

4. Les gènes ciblés par HSF2 après EPA codent des protéines impliquées dans la réponse au stress, dans des fonctions cérébrales ou neurodéveloppementales.

Une analyse ontologique a été réalisée pour explorer de potentielles voies fonctionnelles communes aux gènes directement ciblés par HSF2 après EPA. Les compartiments cellulaires dans lesquels sont impliqués les gènes fixés par HSF2 après EPA sont fortement associés aux membranes synaptiques et aux dendrites (Tableau 4-2.3).

Tableau 4-2.3 : Catégories des principaux Compartiments cellulaires (*Gene Ontology*), significativement représentés par les gènes cibles de HSF2 après EPA.

ID	Nom	q-val	nb liste	nb dans génome	Noms des gènes
GO:0099240	*intrinsic component of synaptic membrane*	5.706E-5	12	245	EPHB2, ERBB4, CLSTN2, DCC, CHRNA7, PTPRO, CNTN5, DLG2, GRIA3, GRIK1, NCSTN, CLSTN1
GO:0097060	*synaptic membrane*	5.706E-5	17	549	EPHB2, ARHGAP32, ERBB4, HIP1, CLSTN2, DCC, CHRNA7, OTOF, PTPRO, CNTN5, DLG2, HSPA8, GRIA3, GRIK1, NCSTN, CLSTN1, GRIP1
GO:0099634	*postsynaptic specialization membrane*	5.706E-5	9	136	ERBB4, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, HSPA8, GRIA3, CLSTN1
GO:0098984	*neuron to neuron synapse*	5.706E-5	15	445	ARHGAP32, ERBB4, DAB1, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, HSPA8, DNAJB1, GRIA3, GRIK1, CLSTN1, GRIP1, PRKAR1B
GO:0045211	*postsynaptic membrane*	7.119E-5	14	406	EPHB2, ARHGAP32, ERBB4, HIP1, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, HSPA8, GRIA3, GRIK1, CLSTN1, GRIP1
GO:0098948	*intrinsic component of postsynaptic specialization membrane*	7.119E-5	8	111	ERBB4, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, GRIA3, CLSTN1
GO:0099146	*intrinsic component of postsynaptic density membrane*	7.119E-5	7	78	ERBB4, CLSTN2, DCC, PTPRO, DLG2, GRIA3, CLSTN1
GO:0014069	*postsynaptic density*	7.119E-5	14	420	ARHGAP32, ERBB4, DAB1, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, HSPA8, DNAJB1, GRIA3, GRIK1, CLSTN1, GRIP1
GO:0032279	*asymmetric synapse*	7.119E-5	14	420	ARHGAP32, ERBB4, DAB1, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, HSPA8, DNAJB1, GRIA3, GRIK1, CLSTN1, GRIP1
GO:0030425	*dendrite*	7.947E-5	19	779	EPHB2, ARHGAP32, CHRNA7, FBXO2, PTPRO, NEGR1, DLG2, RBM8A, HSPA8, HSP90AA1, HSP90AB1, DNAJB1, RANGAP1, GRIA3, GRIK1, PEX5L, CLSTN1, GRIP1, PRKAR1B
GO:0097447	*dendritic tree*	7.947E-5	19	779	EPHB2, ARHGAP32, CHRNA7, FBXO2, PTPRO, NEGR1, DLG2, RBM8A, HSPA8, HSP90AA1, HSP90AB1, DNAJB1, RANGAP1, GRIA3, GRIK1, PEX5L, CLSTN1, GRIP1, PRKAR1B
GO:0099572	*postsynaptic specialization*	1.322E-4	14	454	ARHGAP32, ERBB4, DAB1, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, HSPA8, DNAJB1, GRIA3, GRIK1, CLSTN1, GRIP1
GO:0099699	*integral component of synaptic membrane*	1.480E-4	10	227	EPHB2, ERBB4, CLSTN2, DCC, CHRNA7, PTPRO, GRIA3, GRIK1, NCSTN, CLSTN1
GO:0098936	*intrinsic component of postsynaptic membrane*	1.480E-4	9	179	EPHB2, ERBB4, CLSTN2, DCC, CHRNA7, PTPRO, DLG2, GRIA3, CLSTN1

Résultats des 14 intitulés les plus significatifs de l'analyse ontologique réalisé à partir de l'outil ToppFun (Chen et al., 2009). Les *pvalues,* obtenues selon une méthode statistique reposant sur une fonction de densité de probabilité, ont été corrigées en *qvalues* (q-val) par la méthode Benjamini et Hochberg (*FDR - False Discovery Rate*). Au total, 55 catégories sont significativement enrichies parmi les gènes cibles de HSF2 après EPA.

Tableau 4-2.4 : Catégories de principaux Processus Biologiques (*Gene Ontology*), significativement représentés par les gènes cibles de HSF2 après EPA.

ID	Nom	q-val	nb liste	nb dans génome	Noms des gènes
GO:0021955	*central nervous system neuron axonogenesis*	9.206E-4	6	41	EPHB2, DCC, B4GALT5, HSP90AA1, HSP90AB1, NR4A2
GO:0006458	*'de novo' protein folding*	9.206E-4	6	44	HSPH1, ST13, HSPA8, HSPD1, HSPE1, DNAJB1
GO:1900034	*regulation of cellular response to heat*	1.377E-3	7	79	HSPH1, NUP58, FKBP4, HSPA8, HSP90AA1, HSP90AB1, DNAJB1
GO:0021953	*central nervous system neuron differentiation*	1.685E-3	10	220	EPHB2, ERBB4, DCC, RYK, B4GALT5, NFIX, FGFR2, HSP90AA1, HSP90AB1, NR4A2
GO:0051085	*chaperone cofactor-dependent protein refolding*	1.685E-3	5	34	HSPH1, ST13, HSPA8, HSPE1, DNAJB1
GO:0030182	*neuron differentiation*	1.685E-3	28	1580	EPHB2, ERBB4, SNX3, DAB1, DCC, CHRNA7, RYK, B4GALT5, PTPRO, NEGR1, NFIX, CNTN5, DLG2, FGFR2, FKBP4, HSP90AA1, HSP90AB1, NR4A2, ZHX2, UST, YWHAG, GRIP1, TGIF1, TENM3, BHLHB9, AUTS2, LGR6, PBX3
GO:0022008	*neurogenesis*	1.685E-3	31	1870	EPHB2, ERBB4, SNX3, DAB1, DCC, CHRNA7, RYK, B4GALT5, PTPRO, NEGR1, NFIX, CNTN5, DLG2, FGFR2, FKBP4, HSP90AA1, HSP90AB1, NR4A2, NCSTN, ZHX2, UST, YWHAG, GRIP1, TGIF1, TENM3, BHLHB9, AUTS2, LGR6, EML1, PBX3, PRDM16
GO:0061077	*chaperone-mediated protein folding*	1.685E-3	6	62	HSPH1, ST13, FKBP4, HSPA8, HSPE1, DNAJB1
GO:0021954	*central nervous system neuron development*	1.685E-3	7	96	EPHB2, DCC, B4GALT5, FGFR2, HSP90AA1, HSP90AB1, NR4A2
GO:0006986	*response to unfolded protein*	1.685E-3	9	182	HSPH1, DNAJA1, HSPA8, HSP90AA1, HSP90AB1, HSPD1, HSPE1, NAJB1, WIPI1
GO:0006457	*protein folding*	2.172E-3	10	242	HSPH1, ST13, DNAJA1, FKBP4, HSPA8, HSP90AA1, HSP90AB1, HSPD1, HSPE1, DNAJB1
GO:0051084	*'de novo' posttranslational protein folding*	2.172E-3	5	40	HSPH1, ST13, HSPA8, HSPE1, DNAJB1
GO:0048699	*generation of neurons*	2.264E-3	29	1753	EPHB2, ERBB4, SNX3, DAB1, DCC, CHRNA7, RYK, B4GALT5, PTPRO, NEGR1, NFIX, CNTN5, DLG2, FGFR2, FKBP4, HSP90AA1, HSP90AB1, NR4A2, ZHX2, UST, YWHAG, GRIP1, TGIF1, TENM3, BHLHB9, AUTS2, LGR6, EML1, PBX3
GO:0035966	*response to topologically incorrect protein*	3.139E-3	9	205	HSPH1, DNAJA1, HSPA8, HSP90AA1, HSP90AB1, HSPD1, HSPE1, DNAJB1, WIPI1

Résultats des 14 intitulés les plus significatifs de l'analyse ontologique réalisé à partir de l'outil ToppFun (Chen et al., 2009). Les *pvalues,* obtenues selon une méthode statistique reposant sur une fonction de densité de probabilité, ont été corrigées en *qvalues* (q-val) par la méthode Benjamini et Hochberg (*FDR - False Discovery Rate*). Au total, 48 catégories sont significativement enrichies parmi les gènes cibles de HSF2 après EPA.

Tableau 4-2.5 : Catégories de principaux Phénotypes murins (*Gene Ontology*) associés aux gènes cibles de HSF2 après EPA.

ID	Nom	q-val	nb liste	nb dans génome	Noms des gènes
MP:0002067	*abnormal sensory capabilities/reflexes/nociception*	2.233E-2	24	1111	ARHGAP32, TOX, DAB1, CLSTN2, FBXO2, OTOF, NFIX, MITF, DLG2, ELMOD1, HSPD1, STIM2, NR4A2, ABCC5, GRIA3, GRIK1, PEX5L, UST, TGIF1, TENM3, AUTS2, PRKAR1B, EML1, PBX3
MP:0003492	*abnormal involuntary movement*	2.233E-2	24	1113	ARHGAP32, HIP1, TOX, DAB1, CLSTN2, CHRNA7, FBXO2, OTOF, NFIX, MITF, DLG2, FGFR2, ELMOD1, HSPD1, STIM2, NR4A2, ABCC5, GRIA3, PEX5L, UST, TGIF1, AUTS2, EML1, PBX3
MP:0003635	*abnormal synaptic transmission*	2.233E-2	22	966	ARHGAP32, ERBB4, TOX, MAT2A, ACOT7, CHRNA7, OTOF, SLC24A2, SRGAP3, MITF, DLG2, ELMOD1, RAB27B, STIM2, NR4A2, GRIA3, GRIK1, NCSTN, CLSTN1, GRIP1, RBFOX1, PRKAR1B
MP:0002206	*abnormal CNS synaptic transmission*	2.233E-2	19	765	ARHGAP32, ERBB4, TOX, MAT2A, ACOT7, CHRNA7, OTOF, SLC24A2, SRGAP3, DLG2, ELMOD1, STIM2, GRIA3, GRIK1, NCSTN, CLSTN1, GRIP1, RBFOX1, PRKAR1B
MP:0001961	*abnormal reflex*	2.553E-2	21	928	ARHGAP32, TOX, DAB1, CLSTN2, FBXO2, OTOF, NFIX, MITF, DLG2, ELMOD1, HSPD1, STIM2, NR4A2, ABCC5, GRIA3, PEX5L, UST, TGIF1, AUTS2, EML1, PBX3
MP:0003124	*hypospadia*	3.867E-2	3	11	EPHB2, FGFR2, FKBP4
MP:0011796	*abnormal external urethral orifice morphology*	3.867E-2	3	11	EPHB2, FGFR2, FKBP4
MP:0003861	*abnormal nervous system development*	4.534E-2	25	1312	EPHB2, ARHGAP32, ERBB4, SNX3, ZBTB20, DAB1, DCC, CHRNA7, TBC1D32, COMMD1, BASP1, SRGAP3, NFIX, MITF, RBM8A, FGFR2, NR4A2, NCSTN, GRIP1, TGIF1, CELF2, AUTS2, TSHZ1, EML1, PBX3

Résultats significatifs de l'analyse ontologique réalisé à partir de l'outil ToppFun (Chen et al., 2009). Les *pvalues,* obtenues selon une méthode statistique reposant sur fonction de densité de probabilité, ont été corrigées en *qvalues* (q-val) avec la méthode Benjamini et Hochberg (*FDR - False Discovery Rate*).

Une proportion significative des gènes ciblés par HSF2 après EPA code des protéines impliquées dans la réponse au stress (Tableau 4-2.4). Le rôle du facteur HSF2 dans la médiation de la réponse à une EPA, dans un modèle d'alcoolisation chronique, a été montré par l'équipe, au niveau du cerveau en développement (El Fatimy et al., 2014). L'analyse des gènes ciblés par HSF2 suite à l'EPA aigue que nous étudions, conforte notre hypothèse que ce facteur est également impliqué dans la réponse à ce modèle d'EPA. Par ailleurs, les catégories de processus biologiques significativement enrichies par les gènes directement ciblés par HSF2 après EPA concernent également le neurodéveloppement ou des fonctions cérébrales (Tableau 4-2.4). Plus précisément, après EPA, HSF2 se fixe majoritairement au niveau de gènes importants pour la prolifération et la différentiation neuronale, ainsi que l'axonogenèse. Or, en conditions physiologiques, HSF2 est un facteur connu pour être impliqué dans la migration neuronale au stade de développement que nous étudions (Chang et al., 2006; El Fatimy et al., 2014), *via* la modulation de l'expression de gènes qui contrôlent la migration neuronale radiaire (les gènes codant des *microtubule-associated proteins*, MAP ; (Ayala et al., 2007; Francis et al., 2006) qui sont aussi impliqués dans la neuritogenèse, la synaptogenèse et la plasticité neuronale (Bradshaw and Porteous, 2012; Reiner et al., 2009; Su and Tsai, 2011).

Ainsi, au moins deux hypothèses peuvent être émises quant au comportement du facteur HSF2 dans le cerveau en développement, sur la base des résultats du *ChIP-seq* et de ceux de l'équipe sur des gènes candidats (El Fatimy et al., 2014 ; travaux de l'équipe) : (i) soit les cibles de HSF2 après EPA sont identiques aux cibles physiologiques. Dans ce cas HSF2 n'est probablement pas uniquement impliqué dans la régulation de la migration neuronale en conditions physiologiques (Chang et al., 2006 ; El Fatimy et al., 2014), mais possède également un rôle plus vaste, étendu à d'autres fonctions neurodéveloppementales telles que l'axonogenèse, la prolifération et la différentiation neuronale ; (ii) soit l'exposition prénatale à l'alcool détourne HSF2 de ses gènes cibles physiologiques, impliqués dans la migration neuronale, au profit de gènes impliqués dans d'autres fonctions neurodéveloppementales, ce qui peut avoir des conséquences néfastes sur le développement du cerveau et son intégrité à l'âge adulte. La réalisation d'un *ChIP-seq* en condition contrôle (à partir de cortex d'embryons non stressés) me permettra de mettre en évidence les cibles de HSF2 en conditions physiologiques à ce stade, pour mieux caractériser son rôle dans le cerveau en développement en l'absence de stress et ainsi confirmer l'une ou l'autre des hypothèses (voir les perspectives dans la discussion, chapitre 5).

5. Implication de HSF2 dans les défauts causés par l'EPA ?

5.1. EPA de type binge-drinking et alcoolisation chronique : HSF2 fixe-t-il les mêmes cibles ?

Lors d'études précédentes réalisées dans l'équipe, par *ChIP-seq* ou *ChIP-qPCR*, des cibles de HSF2 avaient déjà été identifiées au même stade de développement (*i.e.* au niveau de cortex embryonnaires E16.5) en conditions physiologiques et/ou à l'issu d'un stress prénatal à l'alcool de type « alcoolisation chronique » (*CAI*, El Fatimy et al., 2014 et travaux non publiés de l'équipe). En comparant les résultats du *ChIP-seq* pilote aux anciens résultats de l'équipe, nous avons identifié une cible fixée par HSF2 (*Hsp90aa1*) à la fois sous EPA de type « *binge drinking »* et *« CAI* ». Excepté ce gène, nous ne retrouvons pas, dans notre *ChIP-seq* pilote, les cibles identifiées en conditions *CAI* (notamment des cibles comme *Hspa1a* (codant HSP70) ou *Dclk1*, El Fatimy et al., 2014). De façon surprenante, très peu de cibles identifiées en conditions naïves (résultats d'un ancien *ChIP-seq* ciblant HSF2 réalisé par l'équipe) sont observées parmi les cibles de HSF2 identifiées après EPA dans le *ChIP-seq* pilote actuel (uniquement 5 cibles en communs[26], associés aux gènes *Celf2*, *Cntn5*, *Msra*, *Taok3* et *Prkar1b*). Or, même si nous nous attendions à obtenir un certain nombre de nouvelles cibles, spécifiques du stress EPA, nous pensions également que HSF2 maintiendraient sa liaison avec certaines de ses cibles physiologiques, comme nous avons pu l'observer en réponse à l'EPA de type CAI (El Fatimy et al., 2014). Le nombre de cibles identifiées en commun entre les deux *ChIP*-seq nous semble donc anormalement faible. Plusieurs hypothèses peuvent expliquer ces différences :

- les protocoles de *ChIP* diffèrent : conditions de sonication et appareillage disponible et surtout l'anticorps utilisé (anciennement, anticorps 3E2, monoclonal chez la souris, alors que nous utilisons actuellement l'anticorps SFI58, polyclonal chez le lapin). De plus, la profondeur de séquençage de l'ancien *ChIP-seq* ciblant HSF2 en conditions naïves étant faible, l'identification des cibles physiologiques de HSF2 était plus délicate. Des paramètres d'analyses peu stringents ont dû être utilisés pour identifier les cibles de HSF2 en tenant compte de cette faible profondeur de couverture. Ainsi, il est possible qu'un certain nombre de cibles de HSF2 ne soient pas identifiées, ou que certaines régions génomiques soient faussement considérées comme des cibles de ce facteur. Ces limites techniques peuvent donc expliquer l'absence de concordance entre les résultats obtenus. Cependant, des expériences de *ChIP-qPCR* réalisées par l'équipe, ont toutefois permis de valider certaines des cibles identifiées dans cette analyse.

[26] Il est important de noter, pour la suite de l'analyse - en particulier pour la comparaison de ces données avec les DMRs observées après EPA - que ces régions ne sont pas présentes dans la capture du méthylome, car aucune sonde n'a pu être produite pour cibler spécifiquement ces régions génomiques. De ce fait, aucune DMR potentielle n'a pu être détecté au niveau de ces régions.

- le type d'alcoolisation (*CAI versus binge drinking*) n'est pas le même. Il est donc possible que des cibles différentes soient impliquées selon le type de stress subi par les embryons. En effet, selon le mode d'alcoolisation, les effets sur le développement du fœtus peuvent être variés (Bandoli et al., 2019).

- les fonds génétiques des souris étudiées diffèrent (fond génétique précédent mixte C57BL/6N et C57BL/6J *versus* fond génétique actuel C57BL/6N). Or, nous savons que la pénétrance de la mutation *Hsf2*KO est très différente sur ces deux fonds génétiques (El Fatimy et al., 2014, et travaux non publiés réalisés au sein de l'équipe par M. Mohamed et V. Dubreuil, pour lesquels les phénotypes des animaux *Hsf2*KO diffèrent de ceux précédemment décrits dans l'article de Kallio et collaborateurs en fonction des conditions d'hébergement ; (Kallio et al., 2002). De plus, des études récentes ont identifié un nombre croissant de différences à la fois génomiques et phénotypiques entre les lignées murines C57BL/6N et C57BL/6J, capables de sous-tendre des différences de pénétrance et d'expressivité de mutations et d'avoir un impact sur la sensibilité au stress, dont l'exposition à l'alcool (Hartmann, 2019; Kang et al., 2018; Simon et al., 2013).

L'ensemble de ces éléments peut expliquer la faible concordance entre les cibles HSF2 précédemment décrites et celles identifiées dans notre *ChIP-seq* pilote. Nous souhaitons réaliser un *ChIP-seq* ciblant HSF2 en conditions physiologiques (voir les perspectives : discussion, chapitre 5), ce qui permettra d'obtenir plus d'informations à ce sujet pour pouvoir conclure.

5.2. HSF2 se lie à des régions génomiques associées à des troubles neurodéveloppementaux, et/ou possédant des motifs de liaison à l'ADN de facteurs de transcription associés à l'EPA.

Plusieurs analyses complémentaires ont été menées pour estimer la participation de HSF2 dans les défauts causés par l'EPA. L'analyse ontologique à partir des données de notre *ChIP-seq* pilote, permet de corréler des phénotypes pathologiques murins aux gènes ciblés par HSF2 sous EPA. Parmi les phénotypes enrichis, se trouvent des troubles neuro-développementaux pouvant être assimilés à ceux décrits lors de TCAF (*e.g.* transmission synaptique anormale, développement du système nerveux anormal, Tableau 4-2.5).

De plus, la recherche de motifs enrichis dans les régions identifiées comme liées par HSF2, révèle la présence de motifs de facteurs de transcription impliqués dans le développement ou le fonctionnement du cerveau. C'est le cas, par exemple, des facteurs NeuroD1 (*Neurogenic differentiation 1*, Jahan et al., 2010; Pataskar et al., 2016), Atoh1 (*Atonal BHLH Transcription Factor 1,* Mulvaney and Dabdoub, 2012) et NeuroG2 (Neurogénine 2, Aydin et al., 2019; Mandal et al., 2015) Figure 4-2.4). Il nous est impossible, à ce stade de l'étude, de savoir si ces facteurs de transcription, ressortant de l'analyse d'enrichissement de motifs sont fixés ou non à l'ADN. Néanmoins, la présence

de leurs motifs, à proximité du motif de fixation de HSF2, permet d'envisager une coopération possible entre le facteur HSF2, et ces facteurs de transcription. Cette coopération pourrait être différente selon les conditions (physiologiques ou suite à l'EPA).

Cette recherche de motifs enrichis dans les régions identifiées comme liées par HSF2, révèle aussi la présence de motifs de facteurs de transcription ou régulateurs déjà identifiés dans des études précédentes d'EPA : notamment, les facteurs CTCF (*CCCTC-Binding Factor*) et NeuroG2 (Figure 4-2.4). En effet, le facteur NeuroG2 est décrit comme étant associé au syndrome d'alcoolisation fœtal chez l'Homme (Mandal et al., 2015, base de données Malacards/Genecards), sachant que le motif de liaison de ce facteur est très conservé entre l'espèce humaine et murine (d'après la comparaison du motif NeuroG2 murin obtenu dans notre analyse et du motif NeuroG2 humain issu de la base de données HOCOMOCO[27]). Toutefois, aucune corrélation entre NeuroG2 et HSF2 n'est - pour le moment - établie dans la littérature à notre connaissance.

Concernant CTCF, protéine insulatrice chez les vertébrés (Ong and Corces, 2014), Laufer et collaborateurs ont montré que le motif du facteur CTCF est enrichi dans un grand nombre de régions pour lesquelles la méthylation de l'ADN est altérée, dans le cerveau adulte murin de 70 jours ayant subi une EPA (Laufer et al., 2013), modèle murin d'alcoolisation fœtale à différents stades de développement - injection intrapéritonéale d'alcool de 2.5 g/kg). Par ailleurs, un mécanisme de réponse au stress impliquant HSF1 et CTCF a été récemment décrit dans la littérature (Vihervaara et al., 2017). Dans des cellules humaines K562, il a été montré que HSF1 était recruté, suite à un stress thermique, au niveau de *loci* où CTCF est fixé, favorisant ainsi l'émergence de plateformes de signalisation pour les régulateurs transcriptionnels (Vihervaara et al., 2017). Un mécanisme similaire impliquant HSF2 (en coopération éventuelle avec HSF1) dans le cadre de la réponse à l'EPA est donc envisageable. D'autant plus que le motif de liaison de HSF (HSE) a été identifié comme associé aux îlots CpGs de gènes hypométhylés et surexprimés dans un modèle d'EPA de cellules ES humaines et de corps embryoïdes (Khalid et al., 2014). HSF2 pourrait donc jouer un rôle dans la lecture de la méthylation de l'ADN, en coopérant avec les facteurs *readers* de la méthylation (voir discussion).

Par ailleurs, le motif de liaison de Zbtb12 est le 3ème motif le plus enrichi dans les séquences fixées par HSF2, juste après les motifs HSE (Figure 4-2.4). Or, ce facteur se fixe au niveau des régions d'ADN où les CpGs sont méthylées (Bartke et al., 2010; de Dieuleveult and Miotto, 2018). L'enrichissement en ce motif de fixation de facteurs de transcription dans les séquences ciblées par HSF2 renforce notre hypothèse d'une implication de HSF2 dans la médiation de défauts de méthylation de l'ADN provoqué par l'EPA.

[27] http://hocomoco11.autosome.ru/motif/NGN2_HUMAN.H11MO.0.D

5.3. HSF2 se fixe au niveau de gènes altérés dans divers contextes d'EPA.

De nombreux gènes ciblés par HSF2 dans notre étude basée sur un modèle de *binge drinking,* sont identifiés, dans la littérature, comme perturbés dans diverses modèles d'EPA, sans qu'un lien avec le facteur HSF2 soit systématiquement établi (Tableau 4-2.2). Ces gènes sont altérés par l'EPA, soit au niveau de leur expression (Carter et al., 2018; Chang et al., 2017; Halder et al., 2014; Kleiber, 2015), soit au niveau de leur méthylation (Frey et al., 2018; Masemola et al., 2015, Tableau 4-2.2).

Notamment, quatre gènes liés à l'empreinte sont fixés par HSF2 après EPA d'après notre analyse de *ChIP-seq* pilote (*Blcap*, *Commd1*, *Dcn*, *Usp29,* Figure 4-2.5 ; Annexe 6-4.1 - onglet ChIP280). Or les gènes soumis à l'empreinte, impliqués dans le neurodéveloppement (Kernohan and Bérubé, 2010) et le fonctionnement cérébral (Davies et al., 2008), sont décrits dans la littérature comme perturbés suite à divers contextes d'EPA, soit au niveau de leur méthylation (méthylation du gène ou des régions régulatrices), soit au niveau de leur expression (***Blcap*** dans Carter et al. 2018 ; ***Dcn*** : Chang et al., 2017 ; ***Usp29*** dans Masemola et al., 2015 ; **autres gènes soumis à l'empreinte** : Dietz et al., 2012; Liu et al., 2009; Shukla et al., 2011; Sittig et al., 2011, revue dans Laufer et al., 2017). Cette catégorie de gènes étant régulée par méthylation de l'ADN (Bartolomei and Ferguson-Smith, 2011; Perez et al., 2016), cette observation renforce également notre hypothèse d'une implication de HSF2 dans la médiation de défauts de méthylation de l'ADN provoqué par l'EPA.

6. Conclusions

Comme d'autres tentatives provenant de différents laboratoires, cette expérience pilote a montré que le *ChIP-seq* ciblant la protéine DNMT3A endogène n'a pas été concluant, malgré un contexte plutôt favorable.

Concernant HSF2, même si des validations sont nécessaires pour confirmer les résultats, l'ensemble de nos observations indiquent que le *ChIP-seq* pilote a fonctionné et permet d'identifier des cibles de HSF2, à l'échelle génomique, très peu de temps après l'EPA de type *binge drinking*. Ces cibles peuvent correspondre à une redistribution de HSF2 suite à l'EPA, où à une maintenance de l'occupation de HSF2 au niveau de ces sites. Ces résultats suggèrent que HSF2 (et plus généralement la voie HSF, à travers la formation d'hétérotrimères HSF1-HSF2) pourrait être impliqué dans des mécanismes à l'origine des défauts causés par l'EPA de façon globale, ainsi que le suggéraient déjà nos travaux et ceux de nos collaborateurs[28] (El Fatimy et al., 2014; Hashimoto-Torii et al., 2014; Ishii et al., 2017). Cette expérience pilote de *ChIP-seq* ayant été fructueuse pour identifier des cibles de HSF2, je réaliserai dans un second temps des *ChIP-qPCR* pour valider ces premiers résultats, et un

[28] Indépendamment du mode d'exposition prénatale à l'alcool (chronique ou aigue)

ChIP-seq sur des échantillons issus d'embryons contrôles (naïfs ou injections PBS) ou ayant subi une EPA, afin de définir plus clairement la redistribution du facteur HSF2 après ce type de stress (voir la discussion, chapitre 5).

Partie 3 : le facteur HSF2 est-il impliqué dans les défauts de méthylation observés après l'EPA ?

Pour explorer le rôle du facteur HSF2 - plus précisément du complexe HSF2-DNMT3A - dans la mise en place des défauts de méthylation de l'ADN dans le cortex embryonnaire E16.5 observé après une EPA de type *binge drinking*, j'ai réalisé, d'une part une capture du méthylome, permettant d'identifier les régions différentiellement méthylées sous ce modèle d'EPA, et d'autre part un *ChIP-seq* pilote, ciblant HSF2 après EPA.

L'analyse de la capture du méthylome a montré que **432 régions génomiques présentent des modifications de méthylation de l'ADN, très rapidement après l'EPA**. Celle du *ChIP-seq* ciblant HSF2, dans le cortex *Hsf2*WT exposé *in utero* à l'alcool, a permis d'identifier **280 régions génomiques potentiellement ciblés par HSF2 dans ces conditions de stress**. Ces régions (DMR ou régions fixées par HSF2 après EPA), se trouvent principalement au niveau de gènes impliqués, soit dans le neurodéveloppement, soit dans le fonctionnement cérébral.

Les gènes soumis à l'empreinte sont notamment significativement ciblés (11 gènes soumis à l'empreinte identifiés parmi les DMR, 4 gènes soumis à l'empreinte - *Blcap*, *Commd1*, *Dcn* et *Usp29* - identifiés comme cibles de HSF2 après EPA). Or, l'expression des gènes soumis à l'empreinte est étroitement régulée par méthylation de l'ADN (Bartolomei and Ferguson-Smith, 2011; Perez et al., 2016). L'altération de cette marque épigénétique par l'EPA, au niveau de leur séquence d'ADN peut donc perturber leur expression, et avoir des effets délétères sur le développement du cerveau, ou son fonctionnement.

Ces résultats sont en accord avec les données de la littérature, décrivant les altérations de la méthylation de l'ADN et/ou de l'expression des gènes soumis à l'empreinte, suite à divers contextes d'EPA, sans qu'une implication de HSF2 n'ait été établie (***Blcap*** dans Carter et al., 2018 ; ***Dcn*** : Chang et al., 2017 ; ***Usp29*** dans Masemola et al., 2015 ; **autres gènes soumis à l'empreinte :** Dietz et al., 2012; Liu et al., 2009; Shukla et al., 2011; Sittig et al., 2011, revue dans Laufer et al., 2017).

Par ailleurs, les gènes de *clusters* codant des protocadhérines, eux aussi régulés par méthylation de l'ADN (Phillips et al., 2017), sont également particulièrement altérés, dans notre modèle d'EPA, au niveau de leur méthylation de l'ADN. Or, les protocadhérines sont des cibles connues de HSF2 dans d'autres contextes physio-pathologiques (Joutsen et al., 2018; Korfanty et al., 2014), mais aussi dans notre *ChIP-seq* pilote ciblant HSF2 après EPA (la protocadhérine *Pcdh7*, hors cluster, est ciblée par HSF2 après EPA).

Enfin, outre les gènes soumis à l'empreinte et *Pcdh7*, d'autres gènes ciblés par HSF2 après EPA, dans notre étude, sont déjà connus dans divers modèles d'EPA, comme affectés par ce stress,

notamment au niveau de la méthylation de l'ADN (*e.g. Dock2* dans Khalid et al., 2014 *; Abcc5* ou *Prdm16* dans Frey et al., 2018 ; Tableau 4-2.2).

L'ensemble de ces observations renforce l'idée que le facteur HSF2 puisse être impliqué dans un mécanisme à l'origine des défauts de méthylation de l'ADN observés après EPA. Afin d'identifier une éventuelle corrélation entre l'implication de HSF2 et des changements de méthylation de l'ADN observés après l'EPA, j'ai intégré les régions différentiellement méthylées après EPA, (DMR identifiées par l'analyse de la capture du méthylome), à celle des régions ciblées par HSF2 sous EPA (résultats du *ChIP-seq* ciblant HSF2 ; voir *Jupyter notebook* - Annexe 6-3.6). Nous n'avons pas observé de régions communes aux deux jeux de données : aucune des régions DMRs observées après EPA n'est fixée par HSF2 lors de ce stress. En effet, la plus faible distance observée entre une région fixée par HSF2 et une DMR est de 2 437pb (Tableau 4-3.1). Seulement 8 gènes sont à la fois identifiés comme fixés par HSF2 et altérés au niveau de leur méthylation après EPA. Ces gènes sont tous hyperméthylés après EPA, excepté le gène *Gse1*, présentant à la fois une région hyperméthylée et une région hypométhylée (Tableau 4-3.1).

Tableau 4-3.1 : Gènes possédant à la fois une région différentiellement méthylée et une région fixée par HSF2 sous EPA.

Nom du gène	Distance DMR-liaison HSF2 (pb)	Différentiel de méthylation après EPA (%)	*pvalue* du différentiel de méthylation
Commd1/Zrsr1	2437	+16,58	0.056
Lchn (E330009J07Rik)	3131	+17,94	0.001
Eml1	3320	+15,11	0.0005
Gm4241 (lncRNA)	4015	+6,29	0.011
Cdc88c	>5000	+12,06	0.013
Fgfr2	>5000	+8,24	0.311
Gse1	>5000	+13,16 -11,66	0.006 0.007
Tshz2	>5000	+15,86	0.015

Intégration des 432 régions différentiellement méthylées après EPA, aux 280 régions fixées par HSF2 après ce stress. Résultats obtenus, soit en comparant les régions chromosomiques de ces deux jeux de données, à l'aide de notre fonction R *find_overlap_regions()*, en tolérant une distance seuil de 5000 bases entre les régions, soit en comparant les noms de gènes (selon les identifiants *Entrez gene accession*) au sein des deux jeux de données. Le signe « + » indique une hyperméthylation après EPA, alors que le signe « - » signifie qu'une hypométhylation a été observée après le stress. La valeur de *pvalue* indiquée ici correspond à la médiane des *pvalues* des CpGs appartenant à la DMR et considérés comme fiables (*i.e.* CpGs ayant une *pval.* seuil < 0.07).

Cette observation, allant certes à l'encontre de notre hypothèse, ne nous permet néanmoins pas de conclure formellement sur l'absence d'implication de HSF2 dans la mise en place des défauts de méthylation de l'ADN après EPA. En effet, nous ne disposons pas actuellement du *ChIP-seq* ciblant HSF2 en conditions contrôles (injection PBS, voir les perspectives dans la discussion, chapitre 5). Il nous est donc actuellement impossible de savoir si des régions ciblées par HSF2 en conditions physiologiques, perdent leur liaison avec HSF2 après l'EPA. Or, comme nous avons observé une interaction entre HSF2 et DNMT3A dans les zones du cortex où se trouvent les progéniteurs neuraux en division, il est possible d'envisager que la perte de liaison de HSF2 engendre également la perte de liaison de DNMT3A au niveau de ces cibles physiologiques, s'il conserve son interaction avec HSF2 (Figure 2.1).

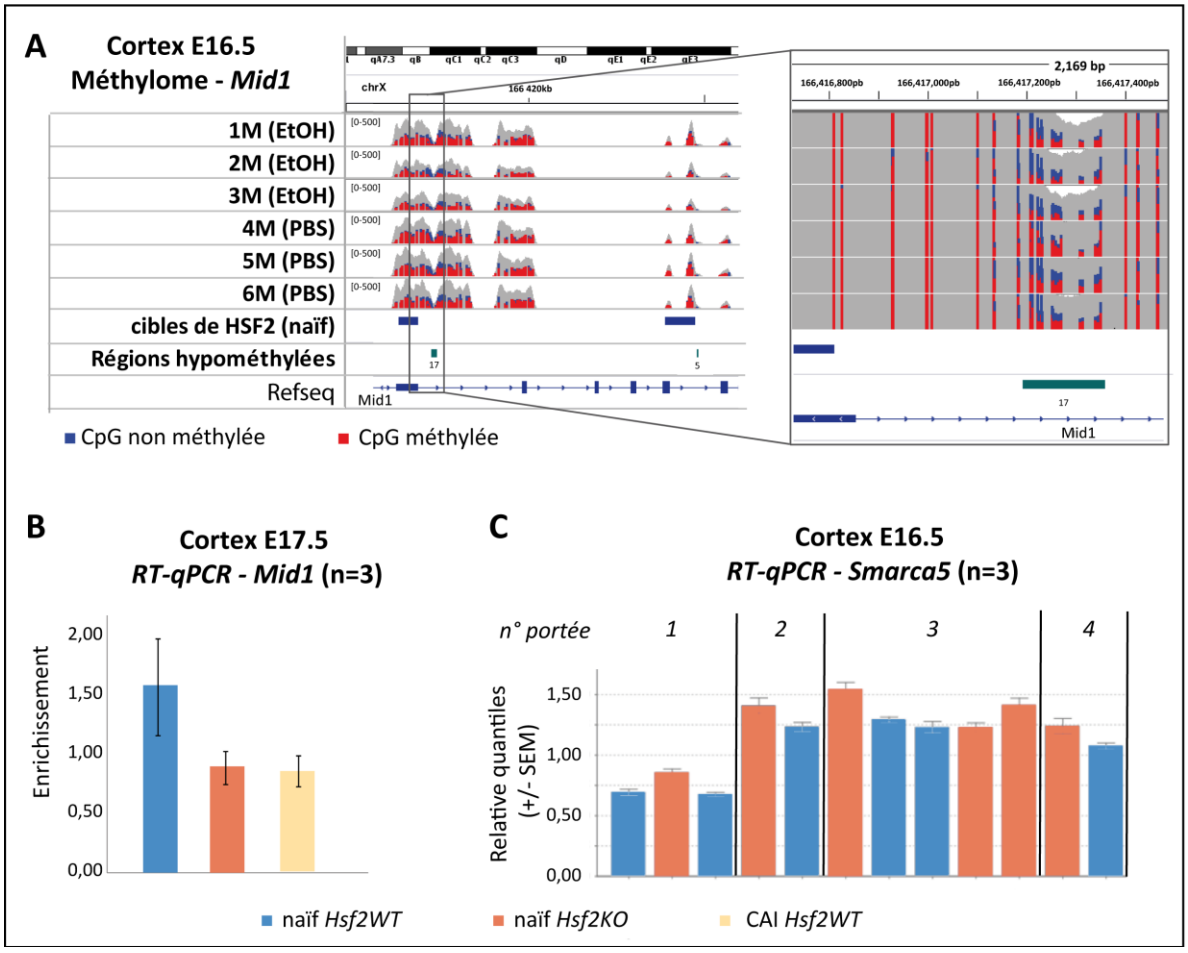


Figure 4-3.1 : *Mid1* et *Smarca5*, cibles de HSF2 en conditions naïves, hypométhylées sous EPA

(A) Aperçu *IGV* (Robinson et al., 2011) des DMRs au niveau du gène *Mid1*. Deux régions hypométhylées après EPA sont identifiées à proximité de sites physiologiques de fixation de HSF2. DMR observées dans le cortex embryonnaires murins E16.5, *Hsf2WT*, identifiées avec l'outil *get_close_loci()* conçu spécifiquement pour l'analyse de la capture du méthylome (collab. S. Ott, Univ. Warwick, UK). (**B**) *RT-qPCR* sur le gène *Mid1*, régulé positivement par HSF2 et réprimé sous stress alcoolique chronique (***CAI*** – *chronic alcohol intoxication*). Expérience réalisée par A. Le Mouël, à partir de cortex embryonnaires E17.5, *Hsf2WT* ou *Hsf2KO*, n = 3 expériences indépendantes. **(C)** *RT-qPCR* sur le gène *Smarca5*, réprimé par HSF2, dans quatre portées distinctes. Expérience réalisée par A. Le Mouël, à partir de cortex embryonnaires E16.5, *Hsf2WT* ou *Hsf2KO*, n = 3 expériences indépendantes.

Dans de tels cas, DNMT3A n'est plus en mesure d'assurer la méthylation *de novo* de l'ADN, ce qui engendrerait l'hypométhylation de ses régions cibles par une déméthylation passive de ces sites. Des résultats allant dans le sens de cette hypothèse ont d'ailleurs été obtenus : parmi les DMR observées dans notre analyse après EPA, j'ai identifié des régions hypométhylées après le stress, dans les gènes *Mid1* (2 DMRs intragéniques, Figure 4-3.1) et *Smarca5* (DMR dans le promoteur), tous deux impliqués dans des fonctions cérébrales (Goodwin and Picketts, 2018; Lu et al., 2013). Or l'équipe a montré par *ChIP-seq*, que :

- HSF2 s'y fixe en conditions physiologiques (Figure 4-3.1A, concernant *Mid1*), mais perd sa fixation après EPA.
- *Mid1* est réprimé sous EPA (Figure 4-3.1B)
- HSF2 régule l'expression de ces deux gènes (Figure 4-3.1C)

La perte de fixation de HSF2, pourrait donc évincer DNMT3A et être ainsi responsable de l'hypométhylation de régions sous EPA, DNMT3A se trouvant « séquestré » par HSF2. Cette corrélation pourrait toutefois être fortuite, étant donné que seulement 3 sites présentent ce profil, parmi les 2111 régions ciblées par HSF2 en conditions naïves (« ancien » *ChIP-seq*), étudiées dans la capture du méthylome. Le complexe HSF2-DNMT3A ne semble donc pas particulièrement intervenir dans les altérations de la méthylation de l'ADN, de type hypométhylation, provoquées par l'EPA. Néanmoins, cette observation doit être confirmée, puisque nous observons une faible concordance des régions ciblées par HSF2 entre les deux expériences de *ChIP-seq* réalisées au laboratoire. L'ancien *ChIP-seq* était réalisé en conditions naïves (sans stress), à partir de cortex d'embryons murins provenant d'un fond génétique mix C57BL/6N et C57BL/6J, alors que le *ChIP-seq* pilote plus récent, est réalisé en conditions EPA, à partir de cortex d'embryons C56BL/6N. La distribution de HSF2 dans le génome peut donc différer, au moins en partie, selon le fond génétique étudiée (discuté précédemment, cf Chapitre 4 section 4.2). Il est, par ailleurs, possible que des contraintes techniques, rendent impossible la comparaison de ces deux expériences, notamment l'utilisation d'anticorps anti-HSF2 différents, le premier utilisé n'étant plus commercialisé, ainsi que la faible profondeur de couverture de l'ancien *ChIP-seq* qui limite l'identification des cibles de HSF2 (discutées précédemment, cf Chapitre 4 section 4.2). De ce fait, aucune conclusion précise ne peut être formulée au sujet de l'implication de HSF2 dans un mécanisme engendrant l'hypométhylation de régions après EPA. La réalisation et l'analyse du *ChIP-seq* ciblant HSF2 en conditions contrôles (souris naïves ou injections PBS, voir les perspectives dans la discussion), comparés aux échantillons ayant subi l'EPA, permettra de définir les régions où HSF2 est fixé en conditions physiologiques, et ainsi comprendre la redistribution du facteur HSF2 après EPA (perte/maintien/gain de fixation). Sous réserve que ces régions soient incluses dans la capture du méthylome (voir plus loin), il sera alors

possible de voir s'il existe une corrélation entre perte de fixation de HSF2 et hypométhylation après l'EPA.

Ainsi, avec les données dont nous disposons actuellement, seules les régions potentiellement hyperméthylées, qui pourraient correspondre à des cibles spécifiques nouvellement fixées par le complexe HSF2-DNMT3A suite au stress, peuvent être analysées.

Par ailleurs, le recours à une approche reposant sur une capture du méthylome ne nous permet d'étudier qu'un nombre limité de régions et non l'ensemble du méthylome. Or, nous avons constaté que seulement 76 des 280 (≈27.1% des régions) régions ciblées par HSF2 après EPA étaient incluses dans le *design* de la capture du méthylome (pourcentage obtenu en recoupant les coordonnées chromosomiques des 280 régions fixées par HSF2 après EPA à celles des régions de la capture, à l'aide de notre fonction R *find_overlap_regions()*, *cf* détails en *Jupyter notebook* - Annexe 6-3.6). Autrement dit, pour **72.9% des sites fixés par HSF2 après EPA**, nous ne connaissons pas l'état de méthylation des CpGs, en conditions contrôle ou après EPA, puisque ces régions hors capture n'ont pas été étudiées. Une capture incluant ces régions[29], ou bien une analyse globale du méthylome (de type *Whole Genome Bisulfite Sequencing* - *WGBS*) permettraient de mieux répondre à notre problématique.

Il est également possible que le complexe HSF2-DNMT3A ne soit pas impliqué dans la mise en place des altérations de la méthylation de l'ADN, directement au niveau du site de fixation de HSF2, mais qu'il agisse à distance, au travers de l'existence de boucles entre régions génomiques distantes. La chromatine compactée dans le noyau est organisée de façon fonctionnelle, en domaines *TADs* (*Topologically Associated Domains*), correspondant à des régions chromatiniennes qui sont plus susceptibles d'interagir entre elles, plutôt qu'avec d'autres régions génomiques hors du domaine. Les *TADs* jouent en un rôle important dans la régulation de l'expression des gènes (Acemel et al., 2017) et dans la régulation temporelle de la réplication (Pope et al., 2014). Il est possible que l'EPA modifie l'organisation des TADs ou des sous-domaines, et/ou que la fixation du complexe HSF2-DNMT3A après EPA sur une séquence génomique donnée puisse engendrer des modifications de la méthylation de l'ADN sur des régions génomiques distantes. Des expériences permettant d'étudier la conformation de la chromatine permettraient de vérifier cette hypothèse (expériences de *Hi-C,* par exemple, Belton et al., 2012).

[29] Nous aurions souhaité réaliser la capture du méthylome après l'analyse du *ChIP-seq* ciblant HSF2 sous EPA, afin d'y inclure les régions ciblées par HSF2 en conditions de stress, mais les contraintes techniques ne nous ont pas permis d'effectuer cela dans un délai raisonnable.

En résumé, aucune corrélation n'a pu être clairement établie dans notre analyse, entre les défauts de méthylation de l'ADN observé rapidement après l'EPA et les sites de fixations de HSF2 après ce stress. Cependant, notre approche par capture du méthylome et le nombre limité de conditions étudiées lors du *ChIP-seq* pilote, ne nous permettent pas d'exclure formellement l'implication de HSF2 dans le mécanisme à l'origine de ces altérations du méthylome. Ainsi, plusieurs hypothèses peuvent encore être envisagées :

- Soit HSF2 (plus spécifiquement, le complexe HSF2-DNMT3A) n'est pas du tout impliqué dans la mise en place des altérations de la méthylation de l'ADN après EPA.
- Soit HSF2, en conservant son interaction avec DNMT3A mais en perdant sa liaison à certaines régions normalement ciblées en conditions physiologiques, engendrera uniquement l'hypométhylation de certains de ces sites, après le stress EPA (perte de fixation), sans intervenir dans la mise en place de régions hyperméthylées (gain de fixation sous stress).
- Soit HSF2 est impliqué dans la mise en place des défauts de méthylation de l'ADN (à la fois hypo et hyperméthylation) après EPA, directement au niveau de son site de fixation, mais notre approche par capture du méthylome ne nous permet pas de l'observer.
- Soit le complexe HSF2-DNMT3A n'est pas impliqué dans la mise en place des altérations de la méthylation de l'ADN, au niveau du site de fixation de HSF2, mais agit à distance au travers la formation de boucles entre régions génomiques.

Chapitre 5 : Perspectives et Discussion

Pendant ma thèse, j'ai contribué à caractériser les modifications précoces de la méthylation de l'ADN observées dans le cortex embryonnaire murin suite à une exposition prénatale à l'alcool aigue (*binge drinking*), subie lors d'un stade de développement équivalent au second trimestre de grossesse chez l'humain. J'ai également estimé les conséquences transcriptionnelles et fonctionnelles de ces défauts sur les régions génomiques impliquées dans les fonctions cérébrales, même si ce travail exploratoire nécessite encore des validations. Par ailleurs, j'ai caractérisé, à l'échelle du génome, les cibles du facteur de réponse au stress HSF2 dans ce modèle d'EPA. Cette **identification globale inédite des cibles de HSF2 dans le cerveau fœtal murin** permet de mieux comprendre l'implication de ce facteur dans la réponse au stress alcoolique subi *in utero*. De plus, cela m'a permis de tester une hypothèse de mécanisme épigénétique, faisant intervenir le complexe HSF2-DNMT3A, qui pourrait être à l'origine d'une partie des perturbations précoces du méthylome observée après EPA. Bien que les résultats obtenus pour le moment ne semblent pas en faveur de cette hypothèse, ils ne nous permettent pas actuellement de confirmer, ni d'infirmer cette hypothèse.

J'ai récemment obtenu un financement EUR G.E.N.E pour un post-doctorat me permettant de poursuivre ce projet pendant une année supplémentaire. Ainsi, je souhaite réaliser des analyses complémentaires, pour valider mes résultats de thèse mais aussi poursuivre les investigations, concernant le rôle de HSF2 dans la mise en place des défauts du méthylome.

1. Des modifications de la méthylation de l'ADN sont observées très rapidement après l'EPA.

1.1. Une étude des effets précoces de l'EPA sur le cerveau en développement

La plupart des études menées sur les défauts de méthylation induites par l'EPA, se sont intéressées aux effets tardifs de ce stress sur le méthylome, en étudiant, notamment la méthylation de l'ADN d'adultes exposés à l'alcool *in utero* (*e.g.* (Laufer et al., 2013; Lussier et al., 2017). Il est possible que l'EPA provoque, au moment de l'exposition, ces changements précoces qui persisteraient à l'âge adulte. Cependant, comme l'activité neuronale peut moduler l'état de méthylation de l'ADN au cours de la vie d'un individu (Su et al., 2017), il est également possible d'envisager que ces modifications du méthylome observées chez l'adulte, ne soient pas mises en place directement lors de l'exposition à l'alcool, mais qu'elles résultent indirectement, d'une activité

neuronale perturbée à la suite de ce stress. Dans ce dernier cas, les changements du méthylome peuvent avoir lieu dans le cerveau postnatal, voire même chez l'adulte, à distance du stress. Il est donc nécessaire de mieux comprendre à quel moment ces perturbations de la méthylation de l'ADN se mettent en place, pour proposer, à terme, des outils diagnostiques fiables des troubles du spectre de l'alcoolisation fœtale, permettant une meilleure prise en charge des enfants à risque, dans des délais brefs (Ehrhart et al., 2019; Lussier et al., 2018).

L'étude que j'ai menée lors de ma thèse apporte une vision génomique globale de l'effet épigénétique précoce d'une EPA de type *binge drinking*. En analysant une capture du méthylome du cortex embryonnaire murin E16.5, j'ai identifié **432 régions génomiques qui présentent des modifications de méthylation des cytosines en contexte CpG, très rapidement après l'EPA**. Suite à ce stress, nous observons aussi bien des régions hypométhylées que des régions hyperméthylées. L'hypothèse selon laquelle des altérations précoces de la méthylation de l'ADN seraient directement provoquées par l'EPA à proximité temporelle de l'exposition, semble donc plausible. Ces résultats sont en accord avec l'étude de Liu et collaborateurs (2009), montrant des altérations précoces du profil de méthylation, observées chez l'embryon murin de 10 jours, ayant subi un stress alcoolique *ex vivo* aigu durant 44h (Liu et al., 2009).

1.2. Les DMR observées rapidement après l'EPA sont associées à des régions génomiques impliquées dans le neuro-développement ou le fonctionnement cérébral

De façon notable, une proportion significative de DMR observées après le stress dans le modèle d'EPA que nous étudions, se situent au niveau de séquences génomiques correspondant à des *enhancers* actifs dans le cerveau murin adulte, caractérisés par la fixation d'histones H3K27ac. Une grande partie des gènes associés à ces *enhancers* altérés précocement par l'EPA, est impliquée, soit dans le neurodéveloppement, soit dans le fonctionnement cérébral (*e.g.* neurogenèse, développement du système nerveux, différentiation neuronale). De plus, grâce à mon analyse intégrée des DMRs avec des données d'*ATAC-seq* et de transcriptomiques, j'ai pu observer qu'un nombre important de DMR, identifiées dans notre modèle d'EPA, est corrélée à des *loci* qui sont normalement - *i.e.* en conditions physiologiques - modulés, dans le cortex préfrontal en développement pendant la période encadrant le stress alcoolique (*binge drinking* réalisé entre les stades E15 et E16, Figure 1A de l'article en préparation dans le chapitre 4, section 4.2). Plus précisément, ces *loci* représentent des régions qui sont **(1)** modulées soit en terme d'accessibilité à la chromatine – notamment les régions chromatiniennes devenant plus accessibles, en conditions physiologiques, entre les stades E14.5 et E15.5 -, **(2)** soit au niveau transcriptionnel - *i.e.* les gènes dont l'expression augmente entre les stades E14.5 et E15.5 ou entre E15.5 et E16.5 – en conditions basales. Sous réserve que certaines de ces régions soient régulées par méthylation de l'ADN, il est

envisageable que la perturbation de la méthylation de l'ADN, précisément au niveau de ces régions génomiques modulées au cours du neuro-développement, puisse avoir des conséquences transcriptionnelles et fonctionnelles directes sur le cerveau en développement. Des altérations de l'expression des gènes, corrélées plus ou moins fortement à des défauts de méthylation causées par l'EPA, ont d'ailleurs été reportées, à distance de l'EPA, avec des conséquences sur les fonctions cérébrales (Kleiber et al., 2013, 2014).

L'étude plus approfondie (*data mining*) des régions identifiées, en conditions physiologiques, comme étant différentiellement accessibles (DOCR) ou différentiellement exprimés (DEG) au cours du développement, et qui sont également associées à une DMR, permettra d'estimer avec plus de précisions les effets potentiels des défauts de méthylation de l'ADN et les conséquences transcriptionnelles associées.

Par ailleurs, pour valider les DMR mais aussi les DEG entre E15.5 et E16.5, identifiées par l'analyse bio-informatique, des expériences de pyroséquençage et de *RT-qPCR* vont aussi être réalisées sur quelques gènes cibles, en conditions contrôles (injections PBS) et après EPA. Des expériences de *RNA-seq* sont également envisagées en conditions contrôles et suite à l'EPA, aux stades embryonnaires E15.5 et E16.5, afin d'obtenir une vision plus globale des altérations transcriptomiques provoquées rapidement par l'exposition prénatale à l'alcool.

1.3. Des modifications précoces de la méthylation de l'ADN par l'EPA, semblent perdurer au cours du temps.

Dans notre modèle d'EPA, les gènes de *clusters* codant des **protocadhérines et les gènes soumis à l'empreinte** constituent des régions génomiques particulièrement ciblées par des changements de leur état de méthylation, rapidement après l'EPA. Comme ces gènes sont régulés par méthylation de l'ADN (Perez et al., 2016), leur expression peut être fortement perturbée par la mise en place de ces marques aberrantes de méthylation suite au stress.

Des altérations au niveau de ces catégories de gènes sont identifiées dans différents modèles d'EPA (Carter et al., 2018; Dietz et al., 2012; Kleiber et al., 2014; Laufer et al., 2017; Liu et al., 2009; Shukla et al., 2011a; Sittig et al., 2011). Il est aussi intéressant de noter que ces mêmes groupes de gènes présentent des altérations de leur méthylation de l'ADN, chez des adultes ayant subi un stress alcoolique *in utero* (Downing et al., 2011; Haycock and Ramsay, 2009; Kleiber et al., 2014; Laufer et al., 2013, 2017). Ces résultats suggèrent donc que des marques de méthylation déposées très tôt après l'exposition prénatale pourraient persister jusque dans la vie adulte. **Ainsi, notre étude suggère que les gènes soumis à l'empreinte et les gènes appartenant aux clusters de protocadhérines, altérés rapidement après le stress, mais aussi chez l'adulte, pourraient être des *loci* génomiques particulièrement vulnérables au stress alcoolique subi *in utero*.**

Ces *loci* représentent donc des **biomarqueurs** potentiellement pertinents pour le **diagnostic des personnes atteintes de TCAF**, d'autant plus que des altérations similaires de méthylation de l'ADN par l'EPA, localisés dans des gènes soumis à l'empreinte ou des gènes codant des protocadhérines ont été observées à la fois chez l'Homme et la souris (Laufer et al., 2015, 2017). Comme évoqué précédemment, l'épigénome peut être modulé par l'activité neuronale (Su et al., 2017). Il est, de ce fait, possible que les marques aberrantes observées chez l'embryon peu de temps après le stress, ne soient pas exactement celles détectées chez l'adulte, mais qu'elles le soient à proximité. Ainsi, le stress alcoolique induirait une perturbation du profil de méthylation sur certains sites, ce qui perturberait localement l'épigénome, et induirait à terme une persistance locale de ces perturbations, sans que la DMR soit nécessairement identique.

2. Méthylation ou Hydroxy-méthylation ?

Les cytosines des mammifères peuvent arborer des modifications chimiques autres qu'un groupement méthyle, susceptibles de jouer un rôle dans la régulation de l'expression des gènes. L'hydroxy-cytosine, forme oxydée de la méthyl-cytosine, portant donc un groupement hydroxyméthyle, a longtemps été uniquement considérée comme un élément intermédiaire du cycle actif de déméthylation de l'ADN par les enzymes *TET.* L'hydroxyméthylation des cytosines est maintenant prise en compte comme une marque épigénétique *per* se, car elle peut moduler l'expression de gènes, en recrutant divers modeleurs de la chromatine (Hahn et al., 2014; Sadakierska-Chudy et al., 2015). Cette marque est particulièrement présente dans le cerveau adulte et en développement, et est activée au cours du développement (Guo et al., 2014; Lister et al., 2013). Chen et collaborateurs (2013), ont par exemple montré, dans l'hippocampe en développement, que le profil de méthylation, et plus particulièrement l'équilibre entre méthylation et hydroxy-méthylation, régule la différenciation neuronale et la maturation spatiotemporelle des neurones. Cette même étude a également mis en évidence qu'une EPA peut altérer le profil de méthylation de l'ADN (*i.e.* à la fois la méthylation et l'hydroxyméthylation), de l'hippocampe en développement, ce qui est corrélé à des défauts de développement, rappelant ceux connus dans le TCAF (Chen et al., 2013).

La conversion au bisulfite de sodium (BS), subie par les échantillons de notre capture du méthylome, ne permet pas de distinguer les méthylations des hydroxyméthylations de l'ADN (Huang et al., 2010). Aussi, comme cette marque est abondante dans le cerveau, il est possible que certaines DMRs identifiées dans notre analyse soient constituées de ces deux formes d'hydroxyméthylaton. L'analyse parallèle de données subissant une conversion BS seule (*BS-seq, bisulfite sequencing)*, ou

précédée d'une étape d'oxydation (*ox-BS seq*, *oxidative bisulfite sequencing*), permettrait de distinguer ces deux marques épigénétiques.

3. Implication du facteur HSF2 dans la mise en place des DMRs précoces ?

3.1. Limites de notre approche pour estimer l'implication de HSF2 dans la mise en place des DMR précoces, dans un contexte CpG.

Dans le but de proposer un mécanisme épigénétique à l'origine des défauts de méthylation précoces observés après l'EPA dans le cortex en développement, je me suis intéressée au complexe HSF2-DNMT3A. Le facteur HSF2 est à la fois essentiel au développement du cerveau (Chang et al., 2006; Kallio et al., 2002), et impliqué dans la réponse au stress, notamment dans la réponse à l'EPA de type alcoolisation chronique (El Fatimy et al., 2014). Dans ces conditions de stress, l'équipe a montré que, dans le cortex murin en développement, HSF2 était capable de se fixer à de nouvelles cibles génomiques, spécifiques du stress, ou de perdre sa liaison avec certaines cibles physiologiques (El Fatimy et al., 2014). De plus, nous avons montré que le facteur HSF2 interagissait avec DNMT3A en conditions basales, et que cette interaction semblait renforcée après EPA (Miozzo, 2014). Ainsi, nous avons posé l'hypothèse qu'en contexte d'EPA, HSF2 pouvait, de par son interaction avec DNMT3A, guider ce facteur, responsable de la méthylation *de novo* de l'ADN, vers de nouvelles cibles, ou au contraire, l'évincer de ses cibles physiologiques, par séquestration. Ces deux *scenarii* peuvent expliquer la mise en place de régions hypo- ou hyperméthylées suite à une EPA.

De façon similaire à l'interaction entre HSF2 et DNMT3A que nous avons observé dans le cerveau murin embryonnaire, une étude récente, réalisée dans une lignée cellulaire de cancer colorectal humain, a montré que le facteur HSF1 pouvait recruter DNMT3A au niveau du promoteur d'un long ARN non codant (MIR137HG), ce qui module son expression et favorise ainsi le développement du cancer colorectal (Li et al., 2018). Les facteurs HSF1 et HSF2 ayant des séquences très similaires, cette étude renforce notre hypothèse mécanistique, d'une redistribution possible de DNMT3A guidée par HSF2, suite à une EPA, potentiellement à l'origine de défauts du méthylome.

Pour tester cette hypothèse, j'ai réalisé et analysé un *ChIP-seq* ciblant HSF2 et DNMT3A, afin de cartographier les cibles de ces facteurs, à l'échelle génomique. Les mises au point techniques du protocole de *ChIP-seq* ayant pris plus de temps que prévu, seuls les échantillons provenant de cortex d'embryons ayant subi l'EPA ont été séquencés, pour une expérience de *ChIP-seq* pilote. Sans le séquençage des échantillons contrôles (traités au PBS), l'interprétation des résultats, concernant une éventuelle redistribution de ces facteurs après l'EPA (perte/maintien/gain de fixation) reste limitée. Néanmoins, même si le *ChIP-seq* ciblant DNMT3A n'a pas fonctionné, possiblement pour les raisons

déjà évoquées précédemment (*cf* Chapitre 4, section 4.2), le ***ChIP-seq* ciblant HSF2**, a permis d'identifier **280 régions génomiques pouvant être fixées** par ce facteur, dans le génome du cortex embryonnaire murin *Hsf2WT,* quelques heures après le stress. Certaines des régions identifiées seront prochainement validées par *ChIP-qPCR* mais plusieurs éléments permettent déjà d'affirmer que l'expérience de *ChIP-seq* pilote a correctement fonctionné car, parmi les cibles de HSF2 identifiées on observe :

- la présence très significative du motif HSE, motif de fixation des HSFs.
- un enrichissement spécifique en *reads* issus de l'échantillon provenant du cortex embryonnaire *Hsf2WT*, par rapport aux résultats obtenus dans le cortex *Hsf2KO.*
- la présence de cibles déjà connues de HSF2 (comme par exemple le gène *Hsp90aa1)*.

Cependant, le croisement des DMR avec les régions identifiées comme cibles de HSF2 sous EPA n'a pas permis d'identifier une corrélation entre ces deux jeux de données. En effet, **aucune des régions DMR observées après EPA n'est fixée par HSF2 lors de ce stress,** ce qui semble invalider notre hypothèse mécanistique. Néanmoins, de nombreux paramètres peuvent expliquer ce manque de corrélation marquée entre les deux jeux de données étudiés, notamment :

- *des contraintes techniques* :

  - La fonction *get_close_loci(),* que nous avons généré spécifiquement pour identifier les DMR au sein de données provenant d'une capture du méthylome, est plutôt stricte puisqu'elle ne regroupe que les cytosines en contexte CpG dont le différentiel de méthylation observé entre le groupe contrôle et le groupe traité, est globalement similaire. Nous avons fait ce choix pour limiter le nombre de DMR faux positifs – malgré tout déjà à ≈22% –, en assumant de tolérer un nombre non négligeable de faux négatifs. En effet, une DMR est définie comme étant constituée uniquement de CpG différentiellement méthylés dans le même sens, (c'est à dire un regroupement des CpGs toutes hyperméthylées ou toutes hypométhylées), sans tolérer la présence d'une, ou quelques cytosines variant en sens opposé. Or, il possible qu'au sein d'une région génomique donnée, quelques cytosines aient un profil de méthylation légèrement différent des cytosines voisines. Certaines DMRs ne sont donc peut-être pas identifiées avec notre démarche.

  - La recherche de DMR est limitée aux régions incluses dans la capture du méthylome, qui comporte peu de régions identifiées comme fixées par HSF2 après EPA (≈27.1% des régions identifiées dans le *ChIP-seq* pilote ciblant HSF2).

  - Nous travaillons à partir d'un tissu cérébral, constitué de populations cellulaires hétérogènes et de sous populations cellulaires elles-mêmes distinctes. Ces différences peuvent masquer des informations. Ainsi, seuls les évènements biologiques majoritaires ou les plus marqués dans

l'échantillon sont mis en évidence. Certaines DMRs et certaines cibles de HSF2 ne sont donc peut-être pas identifiées avec notre démarche. L'isolement de sous populations, en amont de l'analyse, permettraient d'affiner l'étude. Aussi, les avancées technologiques récentes, reposant sur des approches *single cell* permettraient d'obtenir une résolution plus fine de la réponse biologique à l'EPA.

- *des contraintes d'échantillonnage* :

  - Le *ChIP-seq* pilote n'a été réalisé que sur des échantillons ayant subi l'EPA, étant donné que cette expérience servait à la mise au point du protocole et que nous disposions déjà des résultats d'un ancien *ChIP-seq* ciblant HSF2, en conditions naïves.

Par ailleurs, aucune des régions identifiées comme cibles de HSF2 en conditions naïves (ancien *ChIP-seq*) n'est observée après EPA (*ChIP-seq* pilote), alors que nous nous attendions à obtenir quelques régions communes entre les deux conditions, comme cela avait été le cas dans une étude basée sur un modèle de stress alcoolique chronique, menée par l'équipe (El Fatimy et al., 2014). Ici encore, des contraintes techniques peuvent expliquer ces différences de résultats (cf Chapitre 4, section 4.2 et section 4.3), ce qui rend la comparaison des deux *ChIP-seq* délicate. Par conséquent, il est difficile d'identifier les pertes, maintiens ou gains de fixation du facteur HSF2 suite à l'EPA et ainsi d'estimer la redistribution du facteur HSF2 dans le génome, suite à ce stress. Tant qu'un *ChIP-seq* ciblant HSF2 en condition physiologique (naïve) n'est pas réalisé, les seules données du *ChIP-seq* pilote en condition d'EPA ne permet pas l'identification de défauts de méthylation de type hypométhylation, résultant d'une éventuelle perte de fixation de HSF2 après l'EPA.

Au vu de ces éléments, notre approche par capture du méthylome et nos conditions du *ChIP-seq* pilote, ne nous permettent pas de répondre formellement à la question de l'implication ou non de HSF2 dans le mécanisme à l'origine de ces altérations du méthylome.
Plusieurs hypothèses peuvent encore être envisagées :

- Soit le complexe HSF2-DNMT3A n'est pas impliqué dans la mise en place des altérations de la méthylation de l'ADN après EPA.
- Soit HSF2, en conservant son interaction avec DNMT3A mais en perdant sa liaison à certaines régions normalement ciblées en conditions physiologiques, engendre uniquement l'hypométhylation de certains de ces sites, après le stress EPA (perte de fixation), sans intervenir dans la mise en place de régions hyperméthylées (gain de fixation sous stress). L'étude des *loci* des gènes *Mid1* et *Smarca5*, suggère que cela soit le mécanisme en jeu, au moins au niveau de ces deux gènes. Mais, à ce stade de notre analyse, il semble que ce soient les seuls *loci* à présenter ce profil.

- Soit le complexe HSF2-DNMT3A est impliqué dans la mise en place de défauts de méthylation de l'ADN (à la fois hypo et hyperméthylation) après EPA, directement au niveau de site de fixation de HSF2, mais notre approche par capture du méthylome ne nous permet pas de l'observer. Dans ce cas, il serait intéressant de caractériser ces régions sensibles particulières, situées hors des régions capturées (*i.e.* en dehors des régions promotrices et hors des *enhancers* portant la marque H3K27ac dans le cerveau adulte).

3.2. Implication de HSF2 dans la mise en place des DMR précoces, à distance de son site de fixation ?

Une dernière hypothèse – plus probable –, est que le complexe HSF2-DNMT3A soit impliqué dans la mise en place des altérations de la méthylation de l'ADN, non pas directement au niveau de son site de fixation, mais au niveau de séquences génomiques distantes, au travers de l'existence d'interactions privilégiées entre régions chromatiniennes appartenant aux mêmes domaines topologiques associés (TADs). Les régions chromatiniennes associées à un TAD donné interagissent plus facilement entre elles qu'avec les séquences chromatiniennes hors du TAD, en limitant ainsi les interactions avec des séquences régulatrices, tels que des *enhancers* modulant d'autres séquences génomiques (Krefting et al., 2018). Ces domaines TADs contribuent donc à la régulation des gènes (Krefting et al., 2018). La perturbation des TADs ou des sous domaines peut donc engendrer des altérations de l'expression des gènes pouvant être associées à des pathologies (Krefting et al., 2018).

Dans le contexte que nous étudions, au moins deux situations peuvent alors être envisagées quant à **l'implication à distance** du complexe HSF2-DNMT3A dans la mise en place de défauts du méthylome en réponse à l'EPA :

1. Soit, le complexe HSF2-DNMT3A maintient sa fixation au niveau de sa cible suite à ce stress, mais le stress alcoolique modifiant l'organisation des TADs ou des sous-domaines, le complexe n'interagit plus avec ses séquences génomiques habituelles. Dans ce cas, le *loci* génomique où est fixé le complexe HSF2-DNMT3 peut interagir avec d'autres séquences distantes que celles avec lesquelles il s'associe en conditions basales, entrainant l'hyperméthylation de ces nouvelles régions. Et, en miroir, les régions qui, en conditions physiologiques, interagissent normalement à distance avec DNMT3A, se trouvent privées de cette interaction : leur taux de méthylation est alors altéré. Soit ces régions ne sont pas méthylées alors qu'elles devraient l'être, soit elles sont méthylées mais leur méthylation n'est pas maintenue, elles perdent alors, par exemple par dilution passive, leur méthylation de l'ADN.
2. Soit, la fixation du complexe HSF2-DNMT3A est modifiée après EPA (perte ou gain de fixation). Dans ce cas, les séquences génomiques interagissant habituellement avec le *loci*

portant (ou non, selon la situation) le complexe HSF2-DNMT3A, ne seront plus méthylées (ou au contraire méthylées de façon aberrante, dans le cas d'un gain de fixation).

A notre connaissance, aucune étude n'a, pour le moment, montré de perturbations majeures de l'organisation des TADs sous l'effet d'une EPA. En revanche, la redistribution du facteur HSF2 sous EPA, a déjà été mis en évidence suite à une EPA chronique, au moins au niveau de quelques gènes cibles (El Fatimy et al., 2014). De plus, Khalid et collaborateurs (2014) ont montré qu'un motif de liaison de HSF (HSE) a été identifié comme étant associé aux îlots CpGs de gènes hypométhylés et surexprimés dans un modèle d'EPA de cellules ES humaines et de corps embryoïdes.

La deuxième hypothèse formulée semble donc plus probable. Des expériences permettant d'étudier la conformation de la chromatine permettraient de vérifier ces deux hypothèses (expériences de *Hi-C,* par exemple, Belton et al., 2012).

3.3. Implication de HSF2 dans la mise en place des DMR précoces, dans un contexte non CpG ?

La détection des DMR dans notre analyse de la capture du méthylome, a été réalisée en ne s'intéressant qu'aux cytosines en contexte CpG, car c'est à la fois dans ce contexte que la méthylation est la plus répandue chez les mammifères – en particulier dans le cerveau en développement – (Hyun Jang et al., 2017; Lister et al., 2013) mais également la plus analysée dans les études portant sur les défauts de méthylation suite à une EPA. Ainsi, mon analyse pouvait être plus facilement comparée aux données de la littérature. Mais la méthylation de l'ADN a également été observée en contextes non CpG (en contextes CHG ou CHH, avec H = A, T, C), généralement en plus faible proportion. Néanmoins, alors que les méthylations en contextes CHH sont nombreuses dans le cerveau postnatal, notamment dans les neurones, et peuvent avoir des effets plus ou moins marqués sur la régulation de l'expression des gènes (Guo et al., 2014; Lister et al., 2013; Schultz et al., 2015), peu de méthylation en contextes CHH est observée dans le cerveau avant la naissance. Celle-ci se met en place de façon très rapide, peu après la naissance, à la fois chez l'Homme et la souris, alors que la méthylation en contexte CpG, déjà bien présente, évolue peu chez ces deux espèces, en terme de niveau global de méthylation observée (Lister et al., 2013). Cet évènement soudain de méthylation hors contextes CpG, coïncide avec l'augmentation brutale du nombre de synapses (donc avec la synaptogenèse), et l'augmentation de la quantité d'ARN et de protéines DNMT3A (Lister et al., 2013), suggérant l'implication de ce facteur dans la mise en place *de novo* de cette marque.

Comme HSF2 est capable, d'une part d'interagir avec DNMT3A dans le cerveau en développement, à la fois en conditions basales et à la suite d'une EPA, et que d'autre part, son

activité de liaison à l'ADN est renforcée après une exposition à l'alcool (El Fatimy et al., 2014; Miozzo et al., 2018), il est légitime de pense que ce complexe HSF2-DNMT3A puisse perturber la méthylation de l'ADN en contexte non CpG. En effet, il pourrait être à l'origine de la mise en place trop précoce – avant la naissance, au moment de l'exposition à l'alcool – et donc anormale d'une méthylation en contextes CHH, pouvant avoir un impact sur l'expression des gènes dans le cerveau en développement. Cela peut se faire par l'intermédiaire d'une perturbation de l'activité des facteurs *readers* de cette marque épigénétique, comme le suggère les travaux de (Gabel et al., 2015) avec le facteur MeCP2 dans le contexte de gènes longs. Ce facteur est régulé pendant le développement physiologique et présente une affinité particulière pour les méthylations hors contextes CpG. Cette hypothèse est d'autant plus intéressante que de nombreux gènes exprimés dans le cerveau et nécessaires à son fonctionnement sont précisément des gènes longs (Takeuchi et al., 2018).

Pour ces différentes raisons, il serait intéressant et novateur de ré-analyser nos données de capture du méthylome, en contexte CHH, afin de vérifier si une altération de la méthylation de l'ADN est observée suite au *binge drinking*, et dans ce cas, voir si le facteur HSF2 est impliqué dans ces défauts. Nos données de capture et notre démarche bioinformatique actuelle permettraient de le faire.

3.4. Implication de HSF2 dans la régulation de la lecture de la méthylation ?

Parmi les 280 sites génomiques identifiés dans le cortex embryonnaire comme cibles de HSF2 après EPA, un nombre significatif de *séquences* possèdent des motifs de liaison spécifiques de facteurs de transcription capables de reconnaitre et d'interpréter (de lire) la méthylation de l'ADN. Nous avons notamment observé la présence de motifs de fixations des facteurs Zbtb12 (*Zinc finger and BTB domain 12* ; 3ème motif identifié comme le plus enrichi dans les séquences fixées par HSF2, juste après les motifs HSE, parmi 24 motifs enrichis) et CTCF (*CCCTC-Binding Factor* ; 6ème motif).

Peu de données sont disponibles sur Zbtb12, mais il a été montré que ce facteur se fixait au niveau des régions d'ADN où les CpGs sont méthylées, jouant un rôle de *readers* de la chromatine (Bartke et al., 2010; de Dieuleveult and Miotto, 2018). Le facteur CTCF, quant à lui, est impliqué dans de nombreuses fonctions biologiques. Il est notamment impliqué dans la régulation de l'expression des gènes, en tant que facteur répresseur ou activateur de la transcription, mais également de par son activité insulatrice, permettant de délimiter les domaines TADs et de moduler l'organisation tridimensionnelle de la chromatine (Splinter et al., 2006; Williams and Flavell, 2008). Plus de 100 000 motifs de fixation de CTCF sont présents dans le génome murin (Shen et al., 2012), répartit dans différents éléments génomiques (*e.g.* promoteurs, *enhancers*, régions intergéniques ; Laufer et al., 2017). La fixation de CTCF au niveau de ces séquences génomiques spécifiques dépend de leur état

de méthylation (Filippova, 2008; Laufer et al., 2013). Il interagit préférentiellement avec les séquences génomiques non méthylées (Laufer et al., 2013).

Il est intéressant de noter que, dans plusieurs études, des sites de fixation de CTCF sont souvent identifiés au niveau de DMR causées par un stress alcoolique (Haycock and Ramsay, 2009; Knezovich and Ramsay, 2012; Laufer et al., 2013). L'étude de Laufer et collaborateurs (2013) portant sur la méthylation de l'ADN du cerveau adulte murin de 70 jours ayant subi une EPA, a montré que le motif du facteur CTCF était souvent enrichi au niveau des gènes soumis à l'empreinte présentant une DMR après EPA. Or nous retrouvons un grand nombre de gènes soumis à l'empreinte, à la fois parmi les DMRs (11 gènes) précoces observées suite au *binge drinking* mais aussi parmi les régions ciblées par HSF2 après EPA (4 gènes – *Blcap*, *Commd1*, *Dcn* et *Usp29*). L'EPA pourrait donc, en altérant la méthylation de l'ADN, modifier les sites de liaison de CTCF, et ainsi perturber l'expression de gènes, dont ceux des gènes soumis à l'empreinte, parfois impliqués dans le neurodéveloppement (Kernohan and Bérubé, 2010) et le fonctionnement cérébral (Davies et al., 2008).

La présence du motif reconnu par CTCF au niveau des cibles de HSF2 après EPA ne suffit toutefois pas à vérifier la présence ou non de ce facteur. La liaison du facteur CTCF au niveau de ces séquences reste donc à vérifier. Néanmoins, la présence de son motif, à proximité du motif de fixation de HSF2, permet d'envisager une coopération possible entre le facteur HSF2, et CTCF, qui pourrait être différente selon les conditions (physiologiques ou suite à l'EPA), et donc engendrer des conséquences fonctionnelles distinctes dans les deux situations. Un mécanisme de réponse au stress impliquant HSF1 et CTCF a été récemment décrit dans la littérature : suite à un stress thermique, dans des cellules humaines K562, HSF1 est recruté à proximité de sites où est fixé le facteur CTCF, favorisant ainsi l'émergence de plateformes de signalisation pour les régulateurs transcriptionnels (Vihervaara et al., 2017). Il est donc envisageable qu'un mécanisme similaire se mette en place, dans le cadre de la réponse à l'EPA, impliquant HSF2 et CTCF.

L'enrichissement de ces séquences de motifs de fixation de facteurs de transcription dans les séquences ciblées par HSF2 suggère donc que HSF2 puisse être impliqué dans la **médiation de défauts** de méthylation de l'ADN provoqué par l'EPA, en s'associant à des *readers* et/ou des *remodeller*s de la chromatine. Il serait intéressant de voir, en comparant des échantillons contrôles et des échantillons ayant subi l'EPA, si la redistribution de HSF2 suite à l'EPA (perte/maintien/gain de fonction), puisse être corrélée à une redistribution des facteurs Zbtb32 et/ou CTCF, au niveau des cibles de HSF2 possédant un motif pour ce facteur.

4. Rôle de HSF2 dans la réponse au stress alcoolique, indépendamment de la méthylation de l'ADN

4.1. HSF2 est un médiateur de la réponse à l'EPA dans le cerveau, dans divers modèles de stress alcooliques

L'identification des cibles génomiques fixées par HSF2 après l'EPA, permet d'estimer son rôle dans la médiation de la réponse à l'EPA, indépendamment des défauts de méthylation de l'ADN causée par ce stress.

Parmi les gènes ciblés par HSF2 dans notre modèle d'EPA (*binge drinking*) :

- Une proportion significative code des protéines impliquées dans la réponse au stress, et avaient déjà été identifiés comme étant des cibles de HSF2 dans un modèle d'EPA chronique (El Fatimy et al., 2014).
- Certains étaient décrits comme étant altérés par l'EPA, au niveau de leur expression (Carter et al., 2018; Chang et al., 2017; Halder et al., 2014; Kleiber, 2015), ou de leur méthylation (Frey et al., 2018; Masemola et al., 2015) sans qu'un lien avec HSF2 n'ait été établi.

L'analyse des motifs de liaison de facteurs de transcription associés aux sites de fixation de HSF2 dans mon modèle, révèle des motifs de facteurs de transcription déjà associés à l'EPA (*e.g.* CTCF et NeuroG2) (Mandal et al., 2015) ou impliqués dans le développement ou le fonctionnement du cerveau (NeuroD1, Jahan et al., 2010; Pataskar et al., 2016 ; Atoh1 (*Atonal BHLH Transcription Factor 1,* Mulvaney and Dabdoub, 2012 ; NeuroG2 (Neurogénine 2, Aydin et al., 2019; Mandal et al., 2015).

En conditions physiologiques, HSF2 est un facteur connu pour être impliqué dans la migration neuronale au stade de développement que nous étudions (Chang et al., 2006; El Fatimy et al., 2014), *via* la modulation de l'expression de gènes qui contrôlent la migration neuronale radiaire (les gènes codant des *microtubule-associated proteins*, MAP ; Ayala et al., 2007; Francis et al., 2006) qui sont aussi impliqués dans la neuritogenèse, la synaptogenèse et la plasticité neuronale (Bradshaw and Porteous, 2012; Reiner et al., 2009; Su and Tsai, 2011). Après EPA, HSF2 se fixe majoritairement au niveau de gènes importants pour la prolifération et la différentiation neuronale, ainsi que l'axonogenèse.

Enfin, parmi les phénotypes significativement associés aux gènes cibles de HSF2 après EPA, se trouvent des troubles neuro-développementaux pouvant être assimilés à ceux décrits lors de TCAF (*e.g.* transmission synaptique anormale, développement du système nerveux anormal).

Ainsi, l'analyse des gènes cibles de HSF2 et des motifs de fixation de facteurs de transcription associés, dans ce contexte d'alcoolisation aigue, précise l'implication de HSF2 dans la réponse à l'EPA et le neurodéveloppement.

En ce qui concerne le neurodéveloppement, au moins deux hypothèses peuvent être émises : (i) soit les cibles de HSF2 après EPA sont identiques aux cibles physiologiques. Dans ce cas HSF2 n'est probablement pas uniquement impliqué dans la régulation de la migration neuronale en conditions physiologiques (Chang et al., 2006; El Fatimy et al., 2014), mais possède également un rôle plus vaste, étendu à d'autres fonctions neurodéveloppementales telles que l'axonogenèse, la prolifération et la différentiation neuronale ; (ii) soit l'exposition prénatale à l'alcool détourne HSF2 de ses gènes cibles physiologiques, impliqués dans la migration neuronale, au profit de gènes impliqués dans d'autres fonctions neurodéveloppementales, ce qui peut avoir des conséquences néfastes sur le développement du cerveau et son intégrité à l'âge adulte.

Je souhaite réaliser prochainement un *ChIP-seq* ciblant HSF2, en conditions contrôles (à partir de cortex d'embryons non stressés, ou injectés au PBS) et après EPA, qui me permettra de comparer les cibles fixées par HSF2 dans les différentes conditions et ainsi observer la redistribution de ce facteur en réponse au stress (perte/maintien/gain de fixation). L'identification globale des cibles de HSF2 dans le cerveau fœtal murin est inédite, à la fois en conditions physiologiques mais aussi en réponse à l'alcool. Sachant que HSF2 est à la fois un acteur de la réponse au stress et un régulateur de l'expression de gènes impliqués dans des fonctions cérébrales, cela me permettra de mieux caractériser son rôle dans le cerveau en développement dans les différentes conditions et ainsi confirmer l'une ou l'autre des hypothèses.

Par ailleurs, je pourrais intégrer ces données de *ChIP*-seq à celles concernant l'évolution de l'accessibilité de la chromatine et de l'expression des gènes au cours du développement du cortex (résultats de mon d'analyse des données publiques d'ATAC-seq et RNA-seq ENCODE, réalisés à partir de cortex préfrontaux embryonnaires murins). Cela me permettra de mieux estimer l'implication de HSF2 dans le cerveau en développement, en cherchant si les sites de fixation de HSF2, se situent préférentiellement ou non au niveau de régions chromatiniennes différentiellement ouvertes ou fermées, ainsi que les gènes activés ou réprimés entre les stades embryonnaires E13 et E16 dans le cortex en développement.

Des validations par *RT-qPCR* permettront de préciser si HSF2 module l'expression de ses gènes cibles (en comparant des échantillons *Hsf2WT* et *Hsf2KO*), mais aussi de vérifier si la régulation de l'expression de ces gènes par HSF2 est perturbée lors du stress alcoolique (en comparant des échantillons stressés ou non).

L'ensemble de ces analyses me permettront d'avoir une vision beaucoup plus globale de l'implication du facteur HSF2 dans la réponse à l'EPA, et de voir son rôle dans la régulation de la transcription de ses gènes cibles, suite à un stress alcoolique, ou en conditions basales.

4.2. Homotrimères HSF2 ou Hétérotrimères HSF1-HSF2 ? Implication possible de HSF1 dans la réponse à l'EPA précoce.

Le facteur HSF1, qui partage de nombreuses homologies de séquence avec HSF2, est également connu pour son implication dans la réponse au stress alcoolique (El Fatimy et al., 2014; Hashimoto-Torii et al., 2011, 2014; Pignataro et al., 2007).

Nos travaux et ceux de nos collaborateurs montrent qu'en conditions physiologiques, HSF1 n'est pas lié à l'ADN, dans le cortex en développement, contrairement à HSF2, qui se fixe au motif HSE de ses cibles génomiques, sous la forme d'un homotrimère, afin de moduler leur expression (El Fatimy et al., 2014; Hashimoto-Torii et al., 2014). Parmi les cibles physiologiques de HSF2, se trouve les gènes *Maps*, impliqués dans la migration neuronale (Chang et al., 2006; Kallio et al., 2002). Lors d'une EPA chronique, HSF1 est activé par HSF2 et ces deux facteurs coopèrent entre eux pour former un hétérotrimère HSF1-HSF2, capable de fixer l'ADN (El Fatimy et al., 2014). La fixation de HSF1 à l'ADN a également été observée après une EPA aigue (Hashimoto-Torii et al., 2014). HSF1 est ainsi capable de lier des gènes *Hsps* ou *Maps* sous stress alcoolique (Pignataro et al., 2007) mais nécessite la présence de HSF2 pour cela (El Fatimy et al., 2014). Certaines cibles physiologiques de HSF2 peuvent donc être fixées par un hétérotrimères HSF1-HSF2 suite à l'EPA, ce qui peut avoir un impact sur la régulation de ces gènes cibles, particulièrement important pour le bon fonctionnement du cerveau. Le facteur HSF1 peut donc, en coopération avec HSF2 jouer un rôle dans la réponse à l'EPA, et participer ainsi à la mise en place des défauts neuro-développementaux associés à ce stress. Une telle coopération entre HSF1 et HSF2 a également été mise en évidence dans d'autres contextes, notamment dans l'étude du stress thermique (Korfanty et al., 2014).

Il serait donc également informatif de réaliser un *ChIP-seq* pour cartographier la distribution de HSF1, en plus de celle de HSF2, dans le cortex embryonnaire en conditions physiologique et EPA, afin d'avoir une vision intégrée de la redistribution des HSF après EPA, à l'échelle du génome. La présence de HSF1 et/ou HSF2 pourra être mis en lien éventuellement avec la présence de motifs d'autres facteurs de transcription ou facteurs régulateurs identifiés précédemment, comme par exemple NeuroG2 ou CTCF. Des approches d'*ATAC-seq* pourraient permettre aussi d'identifier les « empreintes » (*footprints*) du positionnement de tels facteurs à l'échelle du génome, et leur combinatoire, selon une approche déjà expérimentée par l'équipe dans le contexte de l'exposition périnatale à l'inflammation (Schang et al., 2018).

4.3. HSF2 et Bookmarking : mémoire du stress ?

Lors de la mitose, le génome est globalement fortement compacté (Vagnarelli, 2013), ce qui engendre l'inactivation des gènes. La condensine participe à cette compaction de la chromatine. Le changement de l'organisation de la chromatine avant la mitose s'accompagne de modifications des marques épigénétiques, définissant justement le paysage chromatinien des cellules. Pour autant, le programme transcriptionnel, défini notamment par la présence de ces marques épigénétiques spécifiques, doit être maintenu au cours des divisions cellulaires, pour maintenir l'identité cellulaire au sein d'un tissu. Le *bookmarking* constitue un des mécanismes permettant la réactivation spécifique du programme transcriptionnel après la mitose (Sarge and Park-Sarge, 2005). Il permet donc la ré-expression des gènes dans les cellules filles, à l'identique de l'expression observée dans la cellule mère (Sarge and Park-Sarge, 2005) ou pour préparer la réponse cellulaire à un stimulus (Xing et al., 2005).

Ce *bookmarking* est caractérisé par la présence d'un microenvironnement particulier – absence ou faible condensation de la chromatine locale particulière – maintenu par une marque spécifique, qualifiée de *bookmark (*« marque page »). Il a été montré que HSF2 est impliqué dans un des mécanismes de *bookmarking*, au niveau du promoteur du gène codant pour la protéine HSP70 (Xing et al., 2005). Ce facteur HSF2 permet le recrutement de la phosphatase PP2A, pendant la mitose, qui va localement empêcher la condensine de compacter la région génomique comprenant le promoteur du gène *hsp70* (Xing et al., 2005). Cette faible compaction de la chromatine, permet une régulation transcriptionnelle post-mitotique plus rapide du gène (Sarge and Park-Sarge, 2005; Xing et al., 2005).

Dans notre modèle d'EPA, une proportion importante des régions ciblées par HSF2 après ce stress correspond à des promoteurs. Il est donc possible que HSF2 participe à la mémoire du stress alcoolique subi *in utero* par ce mécanisme de *bookmarking,* soit par son absence – due au stress – de régions qui sont normalement fixées par HSF2, et pour lequel ce rôle de « marquage » n'est plus assuré, soit par un « marquage » inapproprié consécutif au stress subi.

5. Dimorphisme sexuel

Tout comme de nombreux domaines de recherche, la plupart des données de la littérature sur l'EPA ne concernent que l'étude d'échantillons mâles. Pourtant, l'étude du dimorphisme sexuel est d'autant plus pertinente dans le domaine du FASD, que les mâles ayant subi une EPA, présentent un phénotype différent de celui des femelles, au niveau de troubles cognitifs, comportementaux, de la susceptibilité au stress, ou encore de désordres mentaux (Bale, 2015; Hellemans et al., 2008; Oldehinkel and Bouma, 2011). De plus, chez l'Homme comme chez les rongeurs, certaines marques épigénétiques sont très différentes selon le sexe considéré (Gabory et al., 2009; Zhang et al., 2011).

L'étude exclusive des échantillons mâles ne permet pas d'identifier d'éventuelles différences pouvant être dues au dimorphisme sexuel.

Grâce aux échantillons que nous avons déjà collectés au laboratoire, je vais pouvoir en partie adresser cette question du dimorphisme sexuel à partir d'embryons mâles et femelles ayant ou non subi une EPA, en étudiant le profil de leur méthylome et en réalisant un *ChIP-seq* ciblant HSF2 dans ces différents contextes. Ces études devraient permettre de déterminer si le facteur HSF2, possède des sites de fixation qui pourraient varier selon le sexe de l'embryon, en conditions physiologiques, mais également en réponse à l'EPA, ce qui pourraient éventuellement expliquer des différences phénotypiques au niveau du développement physiologique du cerveau, mais aussi au niveau des désordres neuro-développementaux provoqués par l'EPA.

Je dispose déjà de données de méthylome provenant de cortex d'embryons femelles E16.5, qui ont été obtenues de manière similaire à celle des données de méthylome de cortex mâles, étudiés lors de ma thèse (capture SeqCapEpi).

Ainsi, un *ChIP-seq* ciblant HSF2, à partir de cortex embryonnaire E16.5 *Hsf2WT* et *Hsf2KO,* mâles et femelles, en conditions contrôles (naïf ou injectés au PBS) et après *binge drinking,* permettra de répondre à de nombreuses questions, restées jusque-là en suspens :

*Quel est la cartographie globale des sites de fixations du facteur HSF2, dans le cerveau en développement, en conditions physiologiques ?

*Y-a-il une redistribution du facteur HSF2 suite à l'exposition prénatale à l'alcool, dans le modèle d'EPA que nous étudions ? Si c'est le cas :

- sur quels *loci* génomiques s'opèrent les pertes, maintiens, et gains de fixations de HSF2 après EPA ?

-peut-on corréler les pertes de fixations de HSF2 à d'éventuelles hypométhylations observées suite à l'EPA, par notre approche de capture ?

* Existe-t-il un dimorphisme sexuel au niveau des régions génomiques ciblées par HSF2 ?

Conclusions

L'exposition prénatale à l'alcool affecte particulièrement le cerveau du fœtus, sans que les mécanismes moléculaires à l'origine des troubles neuro-développementaux causés par ce stress ne soient complètement élucidés.

Ce travail de thèse a permis de montrer que l'EPA affectait très rapidement la méthylation de l'ADN - en contexte CpG - de certaines régions génomiques dans le cerveau embryonnaire murin. Les données de la littérature s'intéressant plutôt aux effets tardifs de l'EPA, mon étude, basée sur un modèle *in* vivo, constitue l'une des premières analyses globales, portant sur les effets précoces de ce stress sur le méthylome du cerveau. Parmi les *loci* étudiés dans la capture du méthylome, les régions différentiellement méthylées après l'EPA sont significativement associées à des régions génomiques particulières : aux *enhancers* actifs dans le cortex murin adulte mais aussi, aux régions génomiques qui sont normalement modulées dans le cortex au moment du développement, soit en terme d'accessibilité de la chromatine, soit au niveau transcriptionnel. Ces résultats, suggèrent donc que l'exposition *in utero* à l'alcool est susceptible de perturber précocement, l'organisation chromatinienne ou l'expression des gènes dans le cerveau en développement, *via* des altérations ciblées de la méthylation de l'ADN. L'ensemble de ces perturbations peuvent compromettre le bon développement du cerveau et son fonctionnement, et ainsi contribuer à la mise en place des défauts neuro-développementaux typiquement associés à l'alcoolisation fœtale.

De plus, certaines de ces perturbations épigénétiques, notamment celles situées au niveau des *clusters* de protocadhérines ou des gènes liés à l'empreinte, pourraient être des biomarqueurs d'exposition pertinents, puisqu'elles semblent, à la lumière des données de la littérature, persister au cours du temps.

L'analyse de deux éléments additionnels permettrait d'approfondir cette étude des défauts précoces de la méthylation de l'ADN suite à une EPA :

- la méthylation de l'ADN en contexte non CpGs : bien que non présente dans le cerveau embryonnaire en conditions physiologiques, la méthylation de l'ADN en contexte non CpGs pourraient éventuellement être affectée suite à l'EPA.
- Le dimorphisme sexuel : pourtant observé en réponse à l'alcool, il est trop peu étudié dans ce contexte et mériterait d'être exploré.

Ce travail de thèse a également été l'occasion de caractériser, pour la première fois à l'échelle du génome, les cibles du facteur de réponse au stress HSF2 dans ce modèle d'EPA, en vue de proposer un mécanisme épigénétique, impliquant le complexe HSF2-DNMT3A, qui serait à l'origine des défauts précoces de la méthylation de l'ADN causés par l'EPA dans le cerveau en développement. L'analyse intégrée des différents jeux de données NGS, n'a pas permis de mettre en évidence une corrélation entre les sites fixés par HSF2 et les DMRs bien qu'une part significative des gènes ciblés par HSF2 en contexte d'EPA soit régulée par la méthylation de l'ADN. En revanche, les sites de fixation de HSF2 - dans ce contexte d'EPA - étant souvent associés à des sites de liaison de *readers* du méthylome ou des *remodellers* de la chromatine, il est possible que HSF2 soit impliqué dans les conséquences fonctionnelles des perturbations du méthylome dues à l'EPA, plutôt que dans la mise en place de ces défauts.

L'étude du rôle de HSF2, indépendamment des défauts de méthylation causés par l'EPA présente également un intérêt particulier, puisque ce facteur, nécessaire au développement physiologique du cerveau, conduit aussi à des anomalies de développement cérébral, en contexte d'EPA, en changeant de cibles génomiques. L'identification globale de la distribution génomique de HSF2 en contexte physiologique, complèterait celle réalisée au cours de cette thèse, dans le contexte d'EPA. La comparaison de ces données permettrait, (**i**) de caractériser les cibles communes aux deux conditions, qui sont donc peu perturbées par le stress, et, (**ii**) de déterminer les *loci* différentiellement fixés par HSF2 à la suite du stress (perte et gain de fixations). Cette identification globale des cibles de HSF2 dans le cerveau embryonnaire murin sous EPA est inédite dans ces deux conditions et permettrait de mieux comprendre l'implication de ce facteur en conditions physiologiques, mais aussi en réponse au stress alcoolique subi *in utero*. Enfin, comme HSF1, est capable de coopérer avec HSF2, *via* notamment la formation d'hétérotrimères pouvant lier l'ADN dans le cerveau en développement exposé à l'alcool, l'identification globale des cibles de HSF1 en réponse au stress - parallèlement à celle de HSF2 - pourrait contribuer à décrypter l'implication de la voie HSF dans des mécanismes à l'origine des défauts causés par l'EPA.

Bibliographie

Abane, R. (2011). Etude de la régulation des facteurs de choc thermique lors de la réponse au stress des cellules et du cortex cérébral en développement - exemple du syndrome d'alcoolisation foetale. Université Pierre et Marie Curie.

Abane, R., and Mezger, V. (2010). Roles of heat shock factors in gametogenesis and development: Role of the HSF family in development. FEBS J. *277*, 4150–4172.

Acemel, R.D., Maeso, I., and Gómez-Skarmeta, J.L. (2017). Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals: Topologically associated domains. Wiley Interdiscip. Rev. Dev. Biol. *6*, e265.

Achour, M., Le Gras, S., Keime, C., Parmentier, F., Lejeune, F.-X., Boutillier, A.-L., Neri, C., Davidson, I., and Merienne, K. (2015). Neuronal identity genes regulated by super-enhancers are preferentially down-regulated in the striatum of Huntington's disease mice. Hum. Mol. Genet. *24*, 3481–3496.

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., et al. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. *44*, W3–W10.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., Mason, C.E., and others (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol *13*, R87.

Allum, F., Shao, X., Guénard, F., Simon, M.-M., Busche, S., Caron, M., Lambourne, J., Lessard, J., Tandre, K., Hedman, Å.K., et al. (2015). Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. Nat. Commun. *6*, 7211.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Ashwell, K.W.S., and Zhang, L.-L. (1996). Forebrain hypoplasia following acute prenatal ethanol exposure: quantitative analysis of effects on specific forebrain nuclei. Pathology (Phila.) *28*, 161–166.

Ayala, R., Shu, T., and Tsai, L.-H. (2007). Trekking across the brain: the journey of neuronal migration. Cell *128*, 29–43.

Aydin, B., Kakumanu, A., Rossillo, M., Moreno-Estellés, M., Garipler, G., Ringstad, N., Flames, N., Mahony, S., and Mazzoni, E.O. (2019). Proneural factors Ascl1 and Neurog2 contribute to neuronal subtype identities by establishing distinct chromatin landscapes. Nat. Neurosci. *22*, 897–908.

Bale, T.L. (2015). Epigenetic and transgenerational reprogramming of brain development. Nat. Rev. Neurosci. *16*, 332–344.

Bale, T.L., Baram, T.Z., Brown, A.S., Goldstein, J.M., Insel, T.R., McCarthy, M.M., Nemeroff, C.B., Reyes, T.M., Simerly, R.B., Susser, E.S., et al. (2010). Early Life Programming and Neurodevelopmental Disorders. Biol. Psychiatry *68*, 314–319.

Ball, M.P., Li, J.B., Gao, Y., Lee, J.-H., LeProust, E.M., Park, I.-H., Xie, B., Daley, G.Q., and Church, G.M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat. Biotechnol. *27*, 361–368.

Bandoli, G., Coles, C.D., Kable, J.A., Wertelecki, W., Yevtushok, L., Zymak-Zakutnya, N., Wells, A., Granovska, I.V., Pashtepa, A.O., Chambers, C.D., et al. (2019). Patterns of Prenatal Alcohol Use That Predict Infant Growth and Development. Pediatrics *143*, e20182399.

Barlow, D.P. (2011). Genomic Imprinting: A Mammalian Epigenetic Discovery Model. Annu. Rev. Genet. *45*, 379–403.

Barna, J., Csermely, P., and Vellai, T. (2018). Roles of heat shock factor 1 beyond the heat shock response. Cell. Mol. Life Sci. *75*, 2897–2916.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691–1692.

Barrett, L.W., Fletcher, S., and Wilton, S.D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell. Mol. Life Sci. *69*, 3613–3634.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation. Cell *143*, 470–484.

Bartolomei, M.S., and Ferguson-Smith, A.C. (2011). Mammalian genomic imprinting. Cold Spring Harb. Perspect. Biol. *3*.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. Nat. Rev. Cancer *11*, 726–734.

Belton, J.-M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi–C: A comprehensive technique to capture the conformation of genomes. Methods *58*, 268–276.

Bessis, A., Béchade, C., Bernard, D., and Roumier, A. (2007). Microglial control of neuronal death and synaptic properties. Glia *55*, 233–238.

Bestor, T.H. (1992). Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. EMBO J. *11*, 2611–2617.

Bestor, T.H. (1998). The Host Defence Function of Genomic Methylation Patterns. In Novartis Foundation Symposia, D.J. Chadwick, and G. Cardew, eds. (Chichester, UK: John Wiley & Sons, Ltd.), pp. 187–199.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. J. Mol. Biol. *203*, 971–983.

Bird, A. (2007). Perceptions of epigenetics. Nature *447*, 396–398.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bock, C. (2012). Analysing and interpreting DNA methylation data. Nat. Rev. Genet. *13*, 705–719.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bond, J., Roberts, E., Mochida, G.H., Hampshire, D.J., Scott, S., Askham, J.M., Springell, K., Mahadevan, M., Crow, Y.J., Markham, A.F., et al. (2002). ASPM is a major determinant of cerebral cortical size. Nat. Genet. *32*, 316–320.

Bourgeron, T. (2015). From the genetic architecture to synaptic plasticity in autism spectrum disorder. Nat. Rev. Neurosci. *16*, 551–563.

Bradnam, K.R., and Korf, I. (2008). Longer First Introns Are a General Property of Eukaryotic Gene Structure. PLoS ONE *3*, e3093.

Bradshaw, N.J., and Porteous, D.J. (2012). DISC1-binding proteins in neural development, signalling and schizophrenia. Neuropharmacology *62*, 1230–1241.

Brett, D., Pospisil, H., Valcárcel, J., Reich, J., and Bork, P. (2002). Alternative splicing and genome complexity. Nat. Genet. *30*, 29–30.

Briggs, J.A., Wolvetang, E.J., Mattick, J.S., Rinn, J.L., and Barry, G. (2015). Mechanisms of Long Non-coding RNAs in Mammalian Nervous System Development, Plasticity, Disease, and Evolution. Neuron *88*, 861–877.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Burd, L., Cotsonas-Hassler, T.M., Martsolf, J.T., and Kerbeshian, J. (2003). Recognition and management of fetal alcohol syndrome. Neurotoxicol. Teratol. *25*, 681–688.

Burd, L., Deal, E., Rios, R., Adickes, E., Wynne, J., and Klug, M.G. (2007). Congenital heart defects and fetal alcohol spectrum disorders. Congenit. Heart Dis. *2*, 250–255.

Cahill, M.E., Xie, Z., Day, M., Photowala, H., Barbolina, M.V., Miller, C.A., Weiss, C., Radulovic, J., Sweatt, J.D., Disterhoft, J.F., et al. (2009). Kalirin regulates cortical spine morphogenesis and disease-related behavioral phenotypes. Proc. Natl. Acad. Sci. U. S. A. *106*, 13058–13063.

Callis, J., Fromm, M., and Walbot, V. (1987). Introns increase gene expression in cultured maize cells. Genes Dev. *1*, 1183–1200.

Carloni, S., Mazzoni, E., and Balduini, W. (2004). Caspase-3 and calpain activities after acute and repeated ethanol administration during the rat brain growth spurt. J. Neurochem. *89*, 197–203.

Carter, R.C., Chen, J., Li, Q., Deyssenroth, M., Dodge, N.C., Wainwright, H.C., Molteno, C.D., Meintjes, E.M., Jacobson, J.L., and Jacobson, S.W. (2018). Alcohol-Related Alterations in Placental Imprinted Gene Expression in Humans Mediate Effects of Prenatal Alcohol Exposure on Postnatal Growth. Alcohol. Clin. Exp. Res. *42*, 1431–1443.

Carthew, R.W., and Sontheimer, E.J. (2009). Origins and Mechanisms of miRNAs and siRNAs. Cell *136*, 642–655.

Chang, R.C., Skiles, W.M., Chronister, S.S., Wang, H., Sutton, G.I., Bedi, Y.S., Snyder, M., Long, C.R., and Golding, M.C. (2017). DNA methylation-independent growth restriction and altered developmental programming in a mouse model of preconception male alcohol exposure. Epigenetics *12*, 841–853.

Chang, Y., Östling, P., \AAkerfelt, M., Trouillet, D., Rallu, M., Gitton, Y., El Fatimy, R., Fardeau, V., Le Crom, S., Morange, M., et al. (2006). Role of heat-shock factor 2 in cerebral cortex formation and as a regulatorof p35 expression. Genes Dev. *20*, 836–847.

Charness, M.E., Riley, E.P., and Sowell, E.R. (2016). Drinking During Pregnancy and the Developing Brain: Is Any Amount Safe? Trends Cogn. Sci. *20*, 80–82.

Chater-Diehl, E.J., Laufer, B.I., and Singh, S.M. (2017). Changes to histone modifications following prenatal alcohol exposure: An emerging picture. Alcohol *60*, 41–52.

Chatterjee, A., Stockwell, P.A., Rodger, E.J., and Morison, I.M. (2012). Comparison of alignment software for genome-wide bisulphite sequence data. Nucleic Acids Res. *40*, e79–e79.

Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. *37*, W305-311.

Chen, Y., Ozturk, N.C., and Zhou, F.C. (2013). DNA Methylation Program in Developing Hippocampus and Its Alteration by Alcohol. PLoS ONE *8*, e60503.

Chernov, I.P., Akopov, S.B., Nikolaev, L.G., and Sverdlov, E.D. (2002). Identification and mapping of nuclear matrix-attachment regions in a one megabase locus of human chromosome 19q13.12: long-range correlation of S/MARs and gene positions. J. Cell. Biochem. *84*, 590–600.

Clark, S.J., Statham, A., Stirzaker, C., Molloy, P.L., and Frommer, M. (2006). DNA methylation: bisulphite modification and analysis. Nat. Protoc. *1*, 2353–2364.

Clarke, M.E., and Gibbard, W.B. (2003). Overview of Fetal Alcohol Spectrum Disorders for Mental Health Professionals. Can. Child Adolesc. Psychiatry Rev. *12*, 57–63.

Cobben, J.M., Krzyzewska, I.M., Venema, A., Mul, A.N., Polstra, A., Postma, A.V., Smigiel, R., Pesz, K., Niklinski, J., Chomczyk, M.A., et al. (2019). DNA methylation abundantly associates with fetal alcohol spectrum disorder and its subphenotypes. Epigenomics *11*, 767–785.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature *274*, 775–780.

Daneman, R. (2012). The blood-brain barrier in health and disease. Ann. Neurol. *72*, 648–672.

Davies, W., Isles, A.R., Humby, T., and Wilkinson, L.S. (2008). What are imprinted genes doing in the brain? Adv. Exp. Med. Biol. *626*, 62–70.

Dent, C.L., and Isles, A.R. (2014). Brain-expressed imprinted genes and adult behaviour: the example of Nesp and Grb10. Mamm. Genome Off. J. Int. Mamm. Genome Soc. *25*, 87–93.

Dhavan, R., and Tsai, L.-H. (2001). A decade of CDK5. Nat. Rev. Mol. Cell Biol. *2*, 749–759.

Dietz, W.H., Masterson, K., Sittig, L.J., Redei, E.E., and Herzing, L.B.K. (2012). Imprinting and expression of Dio3os mirrors Dio3 in rat. Front. Genet. *3*.

de Dieuleveult, M., and Miotto, B. (2018). DNA Methylation and Chromatin: Role(s) of Methyl-CpG-Binding Protein ZBTB38. Epigenetics Insights *11*, 251686571881111.

Dobin, A., and Gingeras, T.R. (2015). Mapping RNA-seq Reads with STAR: Mapping RNA-seq Reads with STAR. In Current Protocols in Bioinformatics, A. Bateman, W.R. Pearson, L.D. Stein, G.D. Stormo, and J.R. Yates, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.), pp. 11.14.1-11.14.19.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dolzhenko, E., and Smith, A.D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics *15*, 215.

Downing, C., Johnson, T.E., Larson, C., Leakey, T.I., Siegfried, R.N., Rafferty, T.M., and Cooney, C.A. (2011). Subtle decreases in DNA methylation and gene expression at the mouse Igf2 locus following prenatal alcohol exposure: effects of a methyl-supplemented diet. Alcohol *45*, 65–71.

Dulac, C. (2010). Brain function and chromatin plasticity. Nature *465*, 728–735.

Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191.

Ehrhart, F., Roozen, S., Verbeek, J., Koek, G., Kok, G., van Kranen, H., Evelo, C.T., and Curfs, L.M.G. (2019). Review and gap analysis: molecular pathways leading to fetal alcohol spectrum disorders. Mol. Psychiatry *24*, 10–17.

El Fatimy, R., Miozzo, F., Le Mouel, A., Abane, R., Schwendimann, L., Saberan-Djoneidi, D., de Thonel, A., Massaoudi, I., Paslaru, L., Hashimoto-Torii, K., et al. (2014). Heat shock factor 2 is a stress-responsive mediator of neuronal migration defects in models of fetal alcohol syndrome. EMBO Mol. Med. *6*, 1043–1061.

El Hajj, N., Dittrich, M., Böck, J., Kraus, T.F.J., Nanda, I., Müller, T., Seidmann, L., Tralau, T., Galetzka, D., Schneider, E., et al. (2016). Epigenetic dysregulation in the developing Down syndrome cortex. Epigenetics *11*, 563–578.

Faa, G., Manchia, M., Pintus, R., Gerosa, C., Marcialis, M.A., and Fanos, V. (2016). Fetal programming of neuropsychiatric disorders. Birth Defects Res. Part C Embryo Today Rev. *108*, 207–223.

Feng, Y., and Walsh, C.A. (2004). Mitotic spindle regulation by Nde1 controls cerebral cortical size. Neuron *44*, 279–293.

Filippova, G.N. (2008). Genetics and epigenetics of the multifunctional protein CTCF. Curr. Top. Dev. Biol. *80*, 337–360.

Francis, F., Meyer, G., Fallet-Bianco, C., Moreno, S., Kappeler, C., Socorro, A.C., Tuy, F.P.D., Beldjord, C., and Chelly, J. (2006). Human disorders of cortical development: from past to present. Eur. J. Neurosci. *23*, 877–893.

Franco, M.M., Prickett, A.R., and Oakey, R.J. (2014). The Role of CCCTC-Binding Factor (CTCF) in Genomic Imprinting, Development, and Reproduction1. Biol. Reprod. *91*.

Frey, S., Eichler, A., Stonawski, V., Kriebel, J., Wahl, S., Gallati, S., Goecke, T.W., Fasching, P.A., Beckmann, M.W., Kratz, O., et al. (2018). Prenatal Alcohol Exposure Is Associated With Adverse Cognitive Effects and Distinct Whole-Genome DNA Methylation Patterns in Primary School Children. Front. Behav. Neurosci. *12*, 125.

Fu, Y., Springer, N.M., Ying, K., Yeh, C.-T., Iniguez, A.L., Richmond, T., Wu, W., Barbazuk, B., Nettleton, D., Jeddeloh, J., et al. (2010). High-Resolution Genotyping via Whole Genome Hybridizations to Microarrays Containing Long Oligonucleotide Probes. PLoS ONE *5*, e14178.

Fujimori, A., Itoh, K., Goto, S., Hirakawa, H., Wang, B., Kokubo, T., Kito, S., Tsukamoto, S., and Fushiki, S. (2014). Disruption of Aspm causes microcephaly with abnormal neuronal differentiation. Brain Dev. *36*, 661–669.

Fujimoto, M., and Nakai, A. (2010). The heat shock factor family and adaptation to proteotoxic stress: Evolution and function of the HSF family. FEBS J. *277*, 4112–4125.

Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature *522*, 89–93.

Gabory, A., Attig, L., and Junien, C. (2009). Sexual dimorphism in environmental epigenetic programming. Mol. Cell. Endocrinol. *304*, 8–18.

Gardiner-Garden, M., and Frommer, M. (1987). CpG Islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Garro, A.J., McBeth, D.L., Lima, V., and Lieber, C.S. (1991). Ethanol Consumption Inhibits Fetal DNA Methylation in Mice: Implications for the Fetal Alcohol Syndrome. Alcohol. Clin. Exp. Res. *15*, 395–398.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Gibbard, W.B., Wass, P., and Clarke, M.E. (2003). The neuropsychological implications of prenatal alcohol exposure. Can. Child Adolesc. Psychiatry Rev.

Ginhoux, F., and Prinz, M. (2015). Origin of Microglia: Current Concepts and Past Controversies. Cold Spring Harb. Perspect. Biol. *7*, a020537.

Glazko, G.V., Koonin, E.V., Rogozin, I.B., and Shabalina, S.A. (2003). A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. Trends Genet. TIG *19*, 119–124.

Golan-Mashiach, M., Grunspan, M., Emmanuel, R., Gibbs-Bar, L., Dikstein, R., and Shapiro, E. (2012). Identification of CTCF as a master regulator of the clustered protocadherin genes. Nucleic Acids Res. *40*, 3378–3391.

Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.-L., Zhang, X., Golic, K.G., Jacobsen, S.E., and Bestor, T.H. (2006). Methylation of tRNA ^{Asp} by the DNA Methyltransferase Homolog Dnmt2. Science *311*, 395–398.

Gomez-Pastor, R., Burchfiel, E.T., and Thiele, D.J. (2018). Regulation of heat shock transcription factors and their roles in physiology and disease. Nat. Rev. Mol. Cell Biol. *19*, 4–19.

Goodwin, L.R., and Picketts, D.J. (2018). The role of ISWI chromatin remodeling complexes in brain development and neurodevelopmental disorders. Mol. Cell. Neurosci. *87*, 55–64.

Gowher, H., and Jeltsch, A. (2018). Mammalian DNA methyltransferases: new discoveries and open questions. Biochem. Soc. Trans. *46*, 1191–1202.

Gräff, J., Kim, D., Dobbin, M.M., and Tsai, L.-H. (2011). Epigenetic Regulation of Gene Expression in Physiological and Pathological Brain Processes. Physiol. Rev. *91*, 603–649.

Groszer, M., Erickson, R., Scripture-Adams, D.D., Dougherty, J.D., Le Belle, J., Zack, J.A., Geschwind, D.H., Liu, X., Kornblum, H.I., and Wu, H. (2006). PTEN negatively regulates neural stem cell self-renewal by modulating G0-G1 cell cycle entry. Proc. Natl. Acad. Sci. *103*, 111–116.

Gruss, P., Lai, C.J., Dhar, R., and Khoury, G. (1979). Splicing as a requirement for biogenesis of functional 16S mRNA of simian virus 40. Proc. Natl. Acad. Sci. U. S. A. *76*, 4317–4321.

Guerri, C., Bazinet, A., and Riley, E.P. (2009). Foetal Alcohol Spectrum Disorders and Alterations in Brain and Behaviour. Alcohol Alcohol *44*, 108–114.

Guibert, S., and Weber, M. (2013). Functions of DNA methylation and hydroxymethylation in mammalian development. Curr. Top. Dev. Biol. *104*, 47–83.

Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., et al. (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat. Neurosci. *17*, 215–222.

Hahn, M.A., Szabó, P.E., and Pfeifer, G.P. (2014). 5-Hydroxymethylcytosine: A stable or transient DNA modification? Genomics *104*, 314–323.

Haines, T.R., Rodenhiser, D.I., and Ainsworth, P.J. (2001). Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. Dev. Biol. *240*, 585–598.

Halder, D., Park, J.H., Choi, M.R., Chai, J.C., Lee, Y.S., Mandal, C., Jung, K.H., and Chai, Y.G. (2014). Chronic ethanol exposure increases *goosecoid* (GSC) expression in human embryonic carcinoma cell differentiation: Ethanol exposure increases GSC expression during differentiation. J. Appl. Toxicol. *34*, 66–75.

Hard, M.L., Abdolell, M., Robinson, B.H., and Koren, G. (2005). Gene-expression analysis after alcohol exposure in the developing mouse. J. Lab. Clin. Med. *145*, 47–54.

Hartmann, M.C. (2019). Characterization of ethanol-related phenotypic differences between C57BL/6J and C57BL/6NJ substrains: Role of Cyfip2? The University of Maine.

Hashimoto-Torii, K., Kawasawa, Y.I., Kuhn, A., and Rakic, P. (2011). Combined transcriptome analysis of fetal human and mouse cerebral cortex exposed to alcohol. Proc. Natl. Acad. Sci. *108*, 4212–4217.

Hashimoto-Torii, K., Torii, M., Fujimoto, M., Nakai, A., El Fatimy, R., Mezger, V., Ju, M.J., Ishii, S., Chao, S., Brennand, K.J., et al. (2014). Roles of Heat Shock Factor 1 in Neuronal Response to Fetal Environmental Risks and Its Relevance to Brain Disorders. Cell Neuron *82*, 560–572.

Haute Autorité de Santé (2013). Troubles causés par l'alcoolisation foetale : repérage.

Haycock, P.C. (2009). Fetal Alcohol Spectrum Disorders: The Epigenetic Perspective. Biol. Reprod. *81*, 607–617.

Haycock, P.C., and Ramsay, M. (2009). Exposure of Mouse Embryos to Ethanol During Preimplantation Development: Effect on DNA Methylation in the H19 Imprinting Control Region. Biol. Reprod. *81*, 618–627.

He, Y., and Ecker, J.R. (2015). Non-CG Methylation in the Human Genome. Annu. Rev. Genomics Hum. Genet. *16*, 55–77.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hellemans, K.G.C., Verma, P., Yoon, E., Yu, W., and Weinberg, J. (2008). Prenatal alcohol exposure increases vulnerability to stress and anxiety-like disorders in adulthood. Ann. N. Y. Acad. Sci. *1144*, 154–175.

Hellemans, K.G.C., Verma, P., Yoon, E., Yu, W.K., Young, A.H., and Weinberg, J. (2010a). Prenatal Alcohol Exposure and Chronic Mild Stress Differentially Alter Depressive- and Anxiety-Like Behaviors in Male and Female Offspring. Alcohol. Clin. Exp. Res. *34*, 633–645.

Hellemans, K.G.C., Sliwowska, J.H., Verma, P., and Weinberg, J. (2010b). Prenatal alcohol exposure: Fetal programming and later life vulnerability to stress, depression and anxiety disorders. Neurosci. Biobehav. Rev. *34*, 791–807.

Herculano-Houzel, S. (2014). The glia/neuron ratio: How it varies uniformly across brain structures and species and what that means for brain physiology and evolution: The Glia/Neuron Ratio. Glia *62*, 1377–1391.

Hervouet, E., Peixoto, P., Delage-Mourroux, R., Boyer-Guittaut, M., and Cartron, P.-F. (2018). Specific or not specific recruitment of DNMTs for DNA methylation, an epigenetic dilemma. Clin. Epigenetics *10*, 17.

Hodges, E., Smith, A.D., Kendall, J., Xuan, Z., Ravi, K., Rooks, M., Zhang, M.Q., Ye, K., Bhattacharjee, A., Brizuela, L., et al. (2009). High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res. *19*, 1593–1605.

Honda, A., Usui, H., Sakimura, K., and Igarashi, M. (2017). Rufy3 is an adapter protein for small GTPases that activates a Rac guanine nucleotide exchange factor to control neuronal polarity. J. Biol. Chem. *292*, 20936–20946.

Hong, S., Beja-Glasser, V.F., Nfonoyim, B.M., Frouin, A., Li, S., Ramakrishnan, S., Merry, K.M., Shi, Q., Rosenthal, A., Barres, B.A., et al. (2016). Complement and microglia mediate early synapse loss in Alzheimer mouse models. Science *352*, 712–716.

Hong, X., Scofield, D.G., and Lynch, M. (2006). Intron Size, Abundance, and Distribution within Untranslated Regions of Genes. Mol. Biol. Evol. *23*, 2392–2404.

Houchi, H., Pierrefiche, O., Naassila, M., and Daoust, M. (2015). Effets de l'alcoolisation pendant la grossesse. Cah. Nutr. Diététique *50*, 103–108.

Hsieh, J., and Gage, F.H. (2004). Epigenetic control of neural stem cell fate. Curr. Opin. Genet. Dev. *14*, 461–469.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R., and Rao, A. (2010). The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. PLoS ONE *5*, e8888.

Hutson, J.R., Stade, B., Lehotay, D.C., Collier, C.P., and Kapur, B.M. (2012). Folic Acid Transport to the Human Fetus Is Decreased in Pregnancies with Chronic Alcohol Exposure. PLoS ONE *7*, e38057.

Hyun Jang, Woo Shin, Jeong Lee, and Jeong Do (2017). CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. Genes *8*, 148.

Ikonomidou, C., Bittigau, P., Ishimaru, M.J., Wozniak, D.F., Koch, C., Genz, K., Price, M.T., Stefovska, V., Hörster, F., Tenkova, T., et al. (2000). Ethanol-induced apoptotic neurodegeneration and fetal alcohol syndrome. Science *287*, 1056–1060.

Ishii, S., Torii, M., Son, A.I., Rajendraprasad, M., Morozov, Y.M., Kawasawa, Y.I., Salzberg, A.C., Fujimoto, M., Brennand, K., Nakai, A., et al. (2017). Variations in brain defects result from cellular mosaicism in the activation of heat shock signalling. Nat. Commun. *8*, 15157.

Jahan, I., Kersigo, J., Pan, N., and Fritzsch, B. (2010). Neurod1 regulates survival and formation of connections in mouse ear and brain. Cell Tissue Res. *341*, 95–110.

Jäkel, S., and Dimou, L. (2017). Glial Cells and Their Function in the Adult Brain: A Journey through the History of Their Ablation. Front. Cell. Neurosci. *11*.

Jeltsch, A., and Jurkowska, R.Z. (2014). New concepts in DNA methylation. Trends Biochem. Sci. *39*, 310–318.

Jeltsch, A., and Jurkowska, R.Z. (2016). Allosteric control of mammalian DNA methyltransferases – a new regulatory paradigm. Nucleic Acids Res. *44*, 8556–8575.

Jo, B.-S., and Choi, S.S. (2015). Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. *13*, 112.

Johnson, M.B., Sun, X., Kodani, A., Borges-Monroy, R., Girskis, K.M., Ryu, S.C., Wang, P.P., Patel, K., Gonzalez, D.M., Woo, Y.M., et al. (2018). Aspm knockout ferret reveals an evolutionary mechanism governing cerebral cortical size. Nature *556*, 370–375.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. *13*, 484–492.

Jones, K.L., and Smith, D.W. (1973). Recognition of the fetal alcohol syndrome in early infancy. Lancet Lond. Engl. *302*, 999–1001.

Joutsen, J., and Sistonen, L. (2019). Tailoring of Proteostasis Networks with Heat Shock Factors. Cold Spring Harb. Perspect. Biol. *11*, a034066.

Joutsen, J., Silva, A.J.D., Budzynski, M.A., Luoto, J.C., de Thonel, A., Concordet, J.-P., Mezger, V., Saberan-Djoneidi, D., Henriksson, E., and Sistonen, L. (2018). HSF2 protects against proteotoxicity by maintaining cell-cell adhesion. BioRxiv 506881.

Kallio, M., Chang, Y., Manuel, M., Alastalo, T.-P., Rallu, M., Gitton, Y., Pirkkala, L., Loones, M.-T., Paslaru, L., Larney, S., et al. (2002). Brain abnormalities, defective meiotic chromosome synapsis and female subfertility in HSF2 null mice. EMBO J. *21*, 2591–2601.

Kaminen-Ahola, N., Ahola, A., Maga, M., Mallitt, K.-A., Fahey, P., Cox, T.C., Whitelaw, E., and Chong, S. (2010). Maternal Ethanol Consumption Alters the Epigenotype and the Phenotype of Offspring in a Mouse Model. PLoS Genet. *6*, e1000811.

Kang, S.K., Hawkins, N.A., and Kearney, J.A. (2018). C57BL/6J and C57BL/6N substrains differentially influence phenotype severity in the *Scn1a* $^{+/-}$ mouse model of Dravet syndrome. Epilepsia Open epi4.12287.

Kernohan, K.D., and Bérubé, N.G. (2010). Genetic and epigenetic dysregulation of imprinted genes in the brain. Epigenomics *2*, 743–763.

Keshet, I., Lieman-Hurwitz, J., and Cedar, H. (1986). DNA methylation affects the formation of active chromatin. Cell *44*, 535–543.

Kettenmann, H., and Ransom, B.R. (2004). Neuroglia (Oxford University Press).

Kettenmann, H., Hanisch, U.-K., Noda, M., and Verkhratsky, A. (2011). Physiology of Microglia. Physiol. Rev. *91*, 461–553.

Khalid, O., Kim, J.J., Kim, H.-S., Hoang, M., Tu, T.G., Elie, O., Lee, C., Vu, C., Horvath, S., Spigelman, I., et al. (2014). Gene expression signatures affected by alcohol-induced DNA methylomic deregulation in human embryonic stem cells. Stem Cell Res. *12*, 791–806.

Kleiber, M.L. (2015). Ethanol exposure during synaptogenesis in a mouse model of fetal alcohol spectrum disorders: acute and long-term effects on gene expression and behaviour.

Kleiber, M.L., Mantha, K., Stringer, R.L., and Singh, S.M. (2013). Neurodevelopmental alcohol exposure elicits long-term changes to gene expression that alter distinct molecular pathways dependent on timing of exposure. J. Neurodev. Disord. *5*, 1.

Kleiber, M.L., Diehl, E.J., Laufer, B.I., Mantha, K., Chokroborty-Hoque, A., Alberry, B., and Singh, S.M. (2014). Long-term genomic and epigenomic dysregulation as a consequence of prenatal alcohol exposure: a model for fetal alcohol spectrum disorders. Front. Genet. *5*.

Klenova, E.M., Morse, H.C., Ohlsson, R., and Lobanenkov, V.V. (2002). The novel BORIS + CTCF gene family is uniquely involved in the epigenetics of normal biology and cancer. Semin. Cancer Biol. *12*, 399–414.

Knezovich, J.G., and Ramsay, M. (2012). The Effect of Preconception Paternal Alcohol Exposure on Epigenetic Remodeling of the H19 and Rasgrf1 Imprinting Control Regions in Mouse Offspring. Front. Genet. *3*.

Kodituwakku, P.W. (2007). Defining the behavioral phenotype in children with fetal alcohol spectrum disorders: a review. Neurosci. Biobehav. Rev. *31*, 192–201.

Korfanty, J., Stokowy, T., Widlak, P., Gogler-Piglowska, A., Handschuh, L., Podkowiński, J., Vydra, N., Naumowicz, A., Toma-Jonik, A., and Widlak, W. (2014). Crosstalk between HSF1 and HSF2 during the heat shock response in mouse testes. Int. J. Biochem. Cell Biol. *57*, 76–83.

Krefting, J., Andrade-Navarro, M.A., and Ibn-Salem, J. (2018). Evolutionary stability of topologically associating domains is associated with conserved gene regulation. BMC Biol. *16*, 87.

Kriaucionis, S., and Heintz, N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. Science *324*, 929–930.

Kriegstein, A., Noctor, S., and Martínez-Cerdeño, V. (2006). Patterns of neural stem and progenitor cell division may underlie evolutionary cortical expansion. Nat. Rev. Neurosci. *7*, 883–890.

Krishnan, M.L., Van Steenwinckel, J., Schang, A.-L., Yan, J., Arnadottir, J., Le Charpentier, T., Csaba, Z., Dournaud, P., Cipriani, S., Auvynet, C., et al. (2017). Integrative genomics of microglia implicates DLG4 (PSD95) in the white matter development of preterm infants. Nat. Commun. *8*, 428.

Kroeger, P.E., and Morimoto, R.I. (1994). Selection of new HSF1 and HSF2 DNA-binding sites reveals difference in trimer cooperativity. Mol. Cell. Biol. *14*, 7592–7603.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012). DNA methylome analysis using short bisulfite sequencing data. Nat. Methods *9*, 145–151.

Kunde-Ramamoorthy, G., Coarfa, C., Laritsky, E., Kessler, N.J., Harris, R.A., Xu, M., Chen, R., Shen, L., Milosavljevic, A., and Waterland, R.A. (2014). Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res. *42*, e43–e43.

Lacoste, N., and Côté, J. (2003). Le code épigénétique des histones. MS Médecine Sci. *19*, 955–959.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

LaSalle, J.M., Powell, W.T., and Yasui, D.H. (2013). Epigenetic layers and players underlying neurodevelopment. Trends Neurosci. *36*, 460–470.

Laufer, B.I., Mantha, K., Kleiber, M.L., Diehl, E.J., Addison, S.M.F., and Singh, S.M. (2013). Long-lasting alterations to DNA methylation and ncRNAs could underlie the effects of fetal alcohol exposure in mice. Dis. Model. Mech. *6*, 977–992.

Laufer, B.I., Kapalanga, J., Castellani, C.A., Diehl, E.J., Yan, L., and Singh, S.M. (2015). Associative DNA methylation changes in children with prenatal alcohol exposure. Epigenomics *7*, 1259–1274.

Laufer, B.I., Chater-Diehl, E.J., Kapalanga, J., and Singh, S.M. (2017). Long-term alterations to DNA methylation as a biomarker of prenatal alcohol exposure: From mouse models to human children with fetal alcohol spectrum disorders. Alcohol *60*, 67–75.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. *11*, 204–220.

Lee, E.-J., Pei, L., Srivastava, G., Joshi, T., Kushwaha, G., Choi, J.-H., Robertson, K.D., Wang, X., Colbourne, J.K., Zhang, L., et al. (2011). Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing. Nucleic Acids Res. *39*, e127–e127.

Lemoine, P., Harousseau, H., Borteyru, J.P., and Menuet, J.C. (1968). Les enfants de parents alcooliques. Anomalies observées. A propos de 127 cas.

Levitsky, V.G., Podkolodnaya, O.A., Kolchanov, N.A., and Podkolodny, N.L. (2001). Nucleosome formation potential of exons, introns, and Alu repeats. Bioinformatics *17*, 1062–1064.

Li, E., and Zhang, Y. (2014). DNA Methylation in Mammals. Cold Spring Harb. Perspect. Biol. *6*, a019133–a019133.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and others (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J., Song, P., Jiang, T., Dai, D., Wang, H., Sun, J., Zhu, L., Xu, W., Feng, L., Shin, V.Y., et al. (2018). Heat Shock Factor 1 Epigenetically Stimulates Glutaminase-1-Dependent mTOR Activation to Promote Colorectal Carcinogenesis. Mol. Ther.

Li, Q., Suzuki, M., Wendt, J., Patterson, N., Eichten, S.R., Hermanson, P.J., Green, D., Jeddeloh, J., Richmond, T., Rosenbaum, H., et al. (2015). Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. Nucleic Acids Res. *43*, e81–e81.

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. *47*, W199–W205.

Light, S.E.W., and Jontes, J.D. (2017). δ-Protocadherins: Organizers of neural circuit assembly. Semin. Cell Dev. Biol. *69*, 83–90.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global Epigenomic Reconfiguration During Mammalian Brain Development. Science *341*, 1237905–1237905.

Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. Cell *167*, 233-247.e17.

Liu, Y., Balaraman, Y., Wang, G., Nephew, K.P., and Zhou, F.C. (2009). Alcohol exposure alters DNA methylation profiles in mouse embryos at early neurulation. Epigenetics *4*, 500–511.

Liyanage, V.R.B., Zachariah, R.M., Davie, J.R., and Rastegar, M. (2015). Ethanol deregulates Mecp2/MeCP2 in differentiating neural stem cells via interplay between 5-methylcytosine and 5-hydroxymethylcytosine at the Mecp2 regulatory elements. Exp. Neurol. *265*, 102–117.

Lodish, H.F., Berk, A., Baltimore, D., Darnell, J.E., Matsudaira, P., Zipursky, L., Lodish, B., and et al. (1999). Molecular Cell Biology (4th Edition).

Lu, T., Chen, R., Cox, T.C., Moldrich, R.X., Kurniawan, N., Tan, G., Perry, J.K., Ashworth, A., Bartlett, P.F., Xu, L., et al. (2013). X-linked microtubule-associated protein, Mid1, regulates axon development. Proc. Natl. Acad. Sci. *110*, 19131–19136.

Lussier, A.A., Weinberg, J., and Kobor, M.S. (2017). Epigenetics studies of fetal alcohol spectrum disorder: where are we now? Epigenomics *9*, 291–311.

Lussier, A.A., Morin, A.M., MacIsaac, J.L., Salmon, J., Weinberg, J., Reynolds, J.N., Pavlidis, P., Chudley, A.E., and Kobor, M.S. (2018). DNA methylation as a predictor of fetal alcohol spectrum disorder. Clin. Epigenetics *10*, 5.

Maier, S.E., Cramer, J.A., West, J.R., and Sohrabji, F. (1999). Alcohol exposure during the first two trimesters equivalent alters granule cell number and neurotrophin expression in the developing rat olfactory bulb. J. Neurobiol. 10.

Mandal, C., Park, K.S., Jung, K.H., and Chai, Y.G. (2015). Ethanol-related alterations in gene expression patterns in the developing murine hippocampus. Acta Biochim. Biophys. Sin. *47*, 581–587.

Manuel, M., Rallu, M., Loones, M.-T., Zimarino, V., Mezger, V., and Morange, M. (2002). Determination of the consensus binding sequence for the purified embryonic heat shock factor 2. Eur. J. Biochem. *269*, 2527–2537.

Marín, O., and Rubenstein, J.L.R. (2003). Celle migration in the forebrain. Annu. Rev. Neurosci. *26*, 441–483.

Marjonen, H., Sierra, A., Nyman, A., Rogojin, V., Gröhn, O., Linden, A.-M., Hautaniemi, S., and Kaminen-Ahola, N. (2015). Early Maternal Alcohol Consumption Alters Hippocampal DNA Methylation, Gene Expression and Volume in a Mouse Model. PLOS ONE *10*, e0124931.

Markham, J.A., and Koenig, J.I. (2011). Prenatal stress: Role in psychotic and depressive diseases. Psychopharmacology (Berl.) *214*, 89–106.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal *17*, 10.

Masemola, M.L., Merwe, L. van der, Lombard, Z., Viljoen, D., and Ramsay, M. (2015). Reduced DNA methylation at the PEG3 DMR and KvDMR1 loci in children exposed to alcohol in utero: a South African Fetal Alcohol Syndrome cohort study. Front. Genet. *6*.

Matsunaga, Y., Noda, M., Murakawa, H., Hayashi, K., Nagasaka, A., Inoue, S., Miyata, T., Miura, T., Kubo, K.-I., and Nakajima, K. (2017). Reelin transiently promotes N-cadherin-dependent neuronal adhesion during mouse cortical development. Proc. Natl. Acad. Sci. U. S. A. *114*, 2048–2053.

Mattson, S.N., Crocker, N., and Nguyen, T.T. (2011a). Fetal Alcohol Spectrum Disorders: Neuropsychological and Behavioral Features. Neuropsychol. Rev. *21*, 81–101.

Mattson, S.N., Crocker, N., and Nguyen, T.T. (2011b). Fetal Alcohol Spectrum Disorders: Neuropsychological and Behavioral Features. Neuropsychol. Rev. *21*, 81–101.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253–257.

Maunakea, A.K., Chepelev, I., Cui, K., and Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res. *23*, 1256–1269.

McMillan, D.R., Xiao, X., Shao, L., Graves, K., and Benjamin, I.J. (1998). Targeted disruption of heat shock transcription factor 1 abolishes thermotolerance and protection against heat-inducible apoptosis. J. Biol. Chem. *273*, 7523–7528.

Meaney, M.J. (2010). Epigenetics and the biological definition of gene x environment interactions. Child Dev. *81*, 41–79.

Meissner, A. (2010). Epigenetic modifications in pluripotent and differentiated cells. Nat. Biotechnol. *28*, 1079–1088.

Mendillo, M.L., Santagata, S., Koeva, M., Bell, G.W., Hu, R., Tamimi, R.M., Fraenkel, E., Ince, T.A., Whitesell, L., and Lindquist, S. (2012). HSF1 Drives a Transcriptional Program Distinct from Heat Shock to Support Highly Malignant Human Cancers. Cell *150*, 549–562.

Ministère de la santé (2016). Campagne de sensibilisation aux effets de l'alcool pendant la grossesse.

Miozzo, F. (2014). DNA methyltransferases perturbation and cross-talk with the stress response in neural systems exposed to ethanol. Université Paris Diderot.

Miozzo, F., Sabéran-Djoneidi, D., and Mezger, V. (2015). HSFs, Stress Sensors and Sculptors of Transcription Compartments and Epigenetic Landscapes. J. Mol. Biol. *427*, 3793–3816.

Miozzo, F., Arnoux, H., De Thonel, A., Schang, A.L., Saberan-Djoneidi, D., Baudry, A., Schneider, B., and Mezger, V. (2018). Alcohol exposure promotes DNA methyltransferase DNMT3A up-regulation through reactive oxygen species-dependent mechanisms. Cell Stress Chaperones.

Molumby, M.J., Anderson, R.M., Newbold, D.J., Koblesky, N.K., Garrett, A.M., Schreiner, D., Radley, J.J., and Weiner, J.A. (2017). γ-Protocadherins Interact with Neuroligin-1 and Negatively Regulate Dendritic Spine Morphogenesis. Cell Rep. *18*, 2702–2714.

Monk, B.R., Leslie, F.M., and Thomas, J.D. (2012). The effects of perinatal choline supplementation on hippocampal cholinergic development in rats exposed to alcohol during the brain growth spurt. Hippocampus *22*, 1750–1757.

Moore, E.M., and Riley, E.P. (2015). What Happens When Children with Fetal Alcohol Spectrum Disorders Become Adults? Curr. Dev. Disord. Rep. *2*, 219–227.

Mukhopadhyay, P., Rezzoug, F., Kaikaus, J., Greene, R.M., and Pisano, M.M. (2013). Alcohol modulates expression of DNA methyltranferases and methyl CpG-/CpG domain-binding proteins in murine embryonic fibroblasts. Reprod. Toxicol. *37*, 40–48.

Mulvaney, J., and Dabdoub, A. (2012). Atoh1, an Essential Transcription Factor in Neurogenesis and Intestinal and Inner Ear Development: Function, Regulation, and Context Dependency. J. Assoc. Res. Otolaryngol. *13*, 281–293.

Myrie, S.B., and Pinder, M.A. (2018). Skeletal muscle and fetal alcohol spectrum disorder. Biochem. Cell Biol. Biochim. Biol. Cell. *96*, 222–229.

Nagre, N.N., Subbanna, S., Shivakumar, M., Psychoyos, D., and Basavarajappa, B.S. (2015). CB1-receptor knockout neonatal mice are protected against ethanol-induced impairments of DNMT1, DNMT3A, and DNA methylation. J. Neurochem. *132*, 429–442.

Nave, K.-A. (2010). Myelination and support of axonal integrity by glia. Nature *468*, 244–252.

Noctor, S.C., Martínez-Cerdeño, V., Ivic, L., and Kriegstein, A.R. (2004). Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. Nat. Neurosci. *7*, 136–144.

O'Connor, M.J., and Paley, B. (2009). Psychiatric conditions associated with prenatal alcohol exposure. Dev. Disabil. Res. Rev. *15*, 225–234.

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat. Genet. *19*, 219–220.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257.

Oldehinkel, A.J., and Bouma, E.M.C. (2011). Sensitivity to the depressogenic effect of stress and HPA-axis reactivity in adolescence: a review of gender differences. Neurosci. Biobehav. Rev. *35*, 1757–1770.

Olney, J.W., Wozniak, D.F., Jevtovic-Todorovic, V., Farber, N.B., Bittigau, P., and Ikonomidou, C. (2002a). Drug-induced apoptotic neurodegeneration in the developing brain. Brain Pathol. Zurich Switz. *12*, 488–498.

Olney, J.W., Tenkova, T., Dikranian, K., Qin, Y.-Q., Labruyere, J., and Ikonomidou, C. (2002b). Ethanol-induced apoptotic neurodegeneration in the developing C57BL/6 mouse brain. Brain Res. Dev. Brain Res. *133*, 115–126.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Otero, N.K.H., Thomas, J.D., Saski, C.A., Xia, X., and Kelly, S.J. (2012). Choline Supplementation and DNA Methylation in the Hippocampus and Prefrontal Cortex of Rats Exposed to Alcohol During Development. Alcohol. Clin. Exp. Res. *36*, 1701–1709.

Ott, M.O. (1996). L'induction neurale chez la souris. Mini-Synthèse Médecinesciences.

Palazzo, A.F., Springer, M., Shibata, Y., Lee, C.-S., Dias, A.P., and Rapoport, T.A. (2007). The signal sequence coding region promotes nuclear export of mRNA. PLoS Biol. *5*, e322.

Park, S., Hannenhalli, S., and Choi, S. (2014). Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics *15*, 526.

Parker, C.S., and Topol, J. (1984). A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp 70 gene. Cell *37*, 273–283.

Pataskar, A., Jung, J., Smialowski, P., Noack, F., Calegari, F., Straub, T., and Tiwari, V.K. (2016). NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. EMBO J. *35*, 24–45.

Pelizzola, M., and Ecker, J.R. (2011). The DNA methylome. FEBS Lett. *585*, 1994–2000.

Penn, N.W., Suwalski, R., O'Riley, C., Bojanowski, K., and Yura, R. (1972). The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. Biochem. J. *126*, 781–790.

Perera, F., and Herbstman, J. (2011). Prenatal environmental exposures, epigenetics, and disease. Reprod. Toxicol. *31*, 363–373.

Perez, J.D., Rubinstein, N.D., and Dulac, C. (2016). New Perspectives on Genomic Imprinting, an Essential and Multifaceted Mode of Epigenetic Control in the Developing and Adult Brain. Annu. Rev. Neurosci. *39*, 347–384.

Perkins, A., Lehmann, C., Lawrence, R.C., and Kelly, S.J. (2013). Alcohol exposure during development: Impact on the epigenome. Int. J. Dev. Neurosci. *31*, 391–397.

Phillips, G.R., LaMassa, N., and Nie, Y.M. (2017). Clustered protocadherin trafficking. Semin. Cell Dev. Biol. *69*, 131–139.

Pignataro, L., Miller, A.N., Ma, L., Midha, S., Protiva, P., Herrera, D.G., and Harrison, N.L. (2007). Alcohol regulates gene expression in neurons via activation of heat shock factor 1. J. Neurosci. Off. J. Soc. Neurosci. *27*, 12957–12966.

Pignataro, L., Varodayan, F.P., Tannenholz, L.E., and Harrison, N.L. (2009). The regulation of neuronal gene expression by alcohol. Pharmacol. Ther. *124*, 324–335.

Pollard, I. (2007). Neuropharmacology of drugs and alcohol in mother and fetus. Semin. Fetal. Neonatal Med. *12*, 106–113.

Pope, B.D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D.L., Wang, Y., Hansen, R.S., Canfield, T.K., et al. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402–405.

Popova, S., Lange, S., Burd, L., and Rehm, J. (2012). Health Care Burden and Cost Associated with Fetal Alcohol Syndrome: Based on Official Canadian Data. PLoS ONE *7*, e43024.

Popova, S., Lange, S., Shield, K., Mihic, A., Chudley, A.E., Mukherjee, R.A.S., Bekmuradov, D., and Rehm, J. (2016). Comorbidity of fetal alcohol spectrum disorder: a systematic review and meta-analysis. The Lancet *387*, 978–987.

Portales-Casamar, E., Lussier, A.A., Jones, M.J., MacIsaac, J.L., Edgar, R.D., Mah, S.M., Barhdadi, A., Provost, S., Lemieux-Perreault, L.-P., Cynader, M.S., et al. (2016). DNA methylation signature of human fetal alcohol spectrum disorder. Epigenetics Chromatin *9*, 25.

Probst, A.V., Dunleavy, E., and Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. Nat. Rev. Mol. Cell Biol. *10*, 192–206.

Qian, X., Shen, Q., Goderie, S.K., He, W., Capela, A., Davis, A.A., and Temple, S. (2000). Timing of CNS cell generation: a programmed sequence of neuron and glial cell production from isolated murine cortical stem cells. Neuron *28*, 69–80.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Reik, W., and Walter, J. (2001). Genomic imprinting: parental influence on the genome. Nat. Rev. Genet. *2*, 21–32.

Reiner, O., Shmueli, A., and Sapir, T. (2009). Neuronal Migration and Neurodegeneration: 2 Sides of the Same Coin. Cereb. Cortex *19*, i42–i48.

Remmers, C., Sweet, R.A., and Penzes, P. (2014). Abnormal kalirin signaling in neuropsychiatric disorders. Brain Res. Bull. *103*, 29–38.

Rice, D.S., and Curran, T. (2001). Role of the Reelin Signaling Pathway in Central Nervous System Development. Annu. Rev. Neurosci. *24*, 1005–1039.

Robertson, K.D. (2005). DNA methylation and human disease. Nat. Rev. Genet. *6*, 597–610.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Robinson, M.D., Statham, A.L., Speed, T.P., and Clark, S.J. (2010a). Protocol matters: which methylome are you actually studying? Epigenomics *2*, 587–598.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010b). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Roebuck, T.M., Mattson, S.N., and Riley, E.P. (1998). A Review of the Neuroanatomical Findings in Children with Fetal Alcohol Syndrome or Prenatal Exposure to Alcohol. Alcohol. Clin. Exp. Res. *22*, 339–344.

Rubert, G., Miñana, R., Pascual, M., and Guerri, C. (2006). Ethanol exposure during embryogenesis decreases the radial glial progenitorpool and affects the generation of neurons and astrocytes. J. Neurosci. Res. *84*, 483–496.

Sadakierska-Chudy, A., Kostrzewa, R.M., and Filip, M. (2015). A comprehensive view of the epigenetic landscape part I: DNA methylation, passive and active DNA demethylation pathways and histone variants. Neurotox. Res. *27*, 84–97.

Sanna, E., Mostallino, M.C., Busonero, F., Talani, G., Tranquilli, S., Mameli, M., Spiga, S., Follesa, P., and Biggio, G. (2003). Changes in GABA(A) receptor gene expression associated with selective alterations in receptor function and pharmacology after ethanol withdrawal. J. Neurosci. Off. J. Soc. Neurosci. *23*, 11711–11724.

Sarge, K.D., and Park-Sarge, O.-K. (2005). Gene bookmarking: keeping the pages open. Trends Biochem. Sci. *30*, 605–610.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. *103*, 1412–1417.

Schang, A.L., Sabéran-Djoneidi, D., and Mezger, V. (2017). The impact of genomic and epigenomics NGS approaches on our understanding of neuropsychiatric disorders. Clin. Genet.

Schang, A.-L., Steenwinckel, J. van, Lipecki, J., Rich-Griffin, C., Woolley-Allen, K., Dyer, N., Charpentier, T.L., Schäfer, P., Fleiss, B., Ott, S., et al. (2018). Epigenome and transcriptome landscapes highlight dual roles of proinflammatory players in a perinatal model of white matter injury (Developmental Biology).

Schlesinger, F., Smith, A.D., Gingeras, T.R., Hannon, G.J., and Hodges, E. (2013). De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. Genome Res. *23*, 1601–1614.

Schlotz, W., Godfrey, K.M., and Phillips, D.I. (2014). Prenatal origins of temperament: fetal growth, brain structure, and inhibitory control in adolescence. PloS One *9*, e96715.

Schultz, M.D., He, Y., Whitaker, J.W., Hariharan, M., Mukamel, E.A., Leung, D., Rajagopal, N., Nery, J.R., Urich, M.A., Chen, H., et al. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. Nature *523*, 212–216.

Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol. *16*, 990–995.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., et al. (2012). A map of the cis -regulatory sequences in the mouse genome. Nature *488*, 116–120.

Shukla, P.K., Sittig, L.J., Ullmann, T.M., and Redei, E.E. (2011a). Candidate placental biomarkers for intrauterine alcohol exposure. Alcohol. Clin. Exp. Res. *35*, 559–565.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011b). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479*, 74–79.

Simon, M.M., Greenaway, S., White, J.K., Fuchs, H., Gailus-Durner, V., Wells, S., Sorg, T., Wong, K., Bedu, E., Cartwright, E.J., et al. (2013). A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. Genome Biol. *14*, R82.

Sittig, L.J., Shukla, P.K., Herzing, L.B.K., and Redei, E.E. (2011). Strain-specific vulnerability to alcohol exposure in utero via hippocampal parent-of-origin expression of deiodinase-III. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *25*, 2313–2324.

Song, Q., Decato, B., Hong, E.E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J., and Smith, A.D. (2013). A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. PLoS ONE *8*, e81148.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. *20*, 2349–2354.

Streissguth, A.P., and O'Malley, K. (2000). Neuropsychiatric implications and long-term consequences of fetal alcohol spectrum disorders. Semin. Clin. Neuropsychiatry *5*, 177–190.

Su, S.C., and Tsai, L.-H. (2011). Cyclin-dependent kinases in brain development and disease. Annu. Rev. Cell Dev. Biol. *27*, 465–491.

Su, Y., Shin, J., Zhong, C., Wang, S., Roychowdhury, P., Lim, J., Kim, D., Ming, G., and Song, H. (2017). Neuronal activity modifies the chromatin accessibility landscape in the adult brain. Nat. Neurosci. *20*, 476–483.

Sugimoto, M., and Abe, K. (2007). X Chromosome Reactivation Initiates in Nascent Primordial Germ Cells in Mice. PLoS Genet. *3*, e116.

Sulik, K.K. (2005). Genesis of alcohol-induced craniofacial dysmorphism. Exp. Biol. Med. Maywood NJ *230*, 366–375.

Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M., and Mattick, J.S. (2009). Non-coding RNAs: regulators of disease. J. Pathol. *220*, 126–139.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. Science *324*, 930–935.

Takeuchi, A., Iida, K., Tsubota, T., Hosokawa, M., Denawa, M., Brown, J.B., Ninomiya, K., Ito, M., Kimura, H., Abe, T., et al. (2018). Loss of Sfpq Causes Long-Gene Transcriptopathy in the Brain. Cell Rep. *23*, 1326–1341.

Takizawa, T., Nakashima, K., Namihira, M., Ochiai, W., Uemura, A., Yanagisawa, M., Fujita, N., Nakao, M., and Taga, T. (2001). DNA methylation is a critical cell-intrinsic determinant of astrocyte differentiation in the fetal brain. Dev. Cell *1*, 749–758.

Tate, P.H., and Bird, A.P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. Curr. Opin. Genet. Dev. *3*, 226–231.

Teh, A.L., Pan, H., Lin, X., Lim, Y.I., Patro, C.P.K., Cheong, C.Y., Gong, M., MacIsaac, J.L., Kwoh, C.-K., Meaney, M.J., et al. (2016). Comparison of Methyl-capture Sequencing vs. Infinium 450K methylation array for methylome analysis in clinical samples. Epigenetics *11*, 36–48.

The R Core Team (2018). R - A Language and Environment for statistical computing. R Found. Stat. Comput. Vienna Austria.

Thomas, J.D., Biane, J.S., O'Bryan, K.A., O'Neill, T.M., and Dominguez, H.D. (2007). Choline supplementation following third-trimester-equivalent alcohol exposure attenuates behavioral alterations in rats. Behav. Neurosci. *121*, 120–130.

Thompson, B.L., Levitt, P., and Stanwood, G.D. (2009). Prenatal exposure to drugs: effects on brain development and implications for policy and education. Nat. Rev. Neurosci. *10*, 303–312.

de Thonel, A., Ahlskog, J.K., Abane, R., Pires, G., Dubreuil, V., Bertelet, J., Aalto, A.L., Naceri, S., Cordonnier, M., Benasolo, C., et al. (2018). CBP/EP300-dependent acetylation and stabilization of HSF2 are compromised in the rare disorder, Rubinstein-Taybi syndrome. BioRxiv 481457.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tremblay, K.D., Saam, J.R., Ingram, R.S., Tilghman, S.M., and Bartolomei, M.S. (1995). A paternal–specific methylation imprint marks the alleles of the mouse H19 gene. Nat. Genet. *9*, 407–413.

Vagnarelli, P. (2013). Chromatin reorganization through mitosis. Adv. Protein Chem. Struct. Biol. *90*, 179–224.

Valencia, P., Dias, A.P., and Reed, R. (2008). Splicing promotes rapid and efficient mRNA export in mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *105*, 3386–3391.

Valenzuela, C.F., Puglia, M.P., and Zucca, S. (2011). Focus On: Neurotransmitter Systems. Alcohol Res. Health *34*, 106–120.

Varadinova, M., and Boyadjieva, N. (2015). Epigenetic mechanisms: A possible link between autism spectrum disorders and fetal alcohol spectrum disorders. Pharmacol. Res. *102*, 71–80.

Varet, H., Brillet-Guéguen, L., Coppée, J.-Y., and Dillies, M.-A. (2016). SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. PLOS ONE *11*, e0157022.

Vihervaara, A., Sergelius, C., Vasara, J., Blom, M.A., Elsing, A.N., Roos-Mattjus, P., and Sistonen, L. (2013). Transcriptional response to stress in the dynamic chromatin environment of cycling and mitotic cells. Proc. Natl. Acad. Sci. *110*, E3388–E3397.

Vihervaara, A., Mahat, D.B., Guertin, M.J., Chu, T., Danko, C.G., Lis, J.T., and Sistonen, L. (2017). Transcriptional response to stress is pre-wired by promoter and enhancer architecture. Nat. Commun. *8*, 255.

Waddington, C.H. (2012). The Epigenotype. Int. J. Epidemiol. *41*, 10–13.

Wang, H.-Q., Tuominen, L.K., and Tsai, C.-J. (2011). SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. Bioinformatics *27*, 225–231.

Watson, R.R. (2015). Handbook of fertility: nutrition, diet, lifestyle and reproductive health.

van der Werf, I.M., Kooy, R.F., and Vandeweyer, G. (2015). A robust protocol to increase NimbleGen SeqCap EZ multiplexing capacity to 96 samples. PloS One *10*, e0123872.

Wilkerson, D.C., Skaggs, H.S., and Sarge, K.D. (2007). HSF2 binds to the Hsp90, Hsp27, and c-Fos promoters constitutively and modulates their expression. Cell Stress Chaperones *12*, 283–290.

Williams, A., and Flavell, R.A. (2008). The role of CTCF in regulating nuclear organization. J. Exp. Med. *205*, 747–750.

Williamson, J.M., and Lyons, D.A. (2018). Myelin Dynamics Throughout Life: An Ever-Changing Landscape? Front. Cell. Neurosci. *12*, 424.

Woods, C.G., Bond, J., and Enard, W. (2005). Autosomal Recessive Primary Microcephaly (MCPH): A Review of Clinical, Molecular, and Evolutionary Findings. Am. J. Hum. Genet. *76*, 717–728.

Wu, C. (1984). Activating protein factor binds in vitro to upstream control sequences in heat shock gene chromatin. Nature *311*, 81–84.

Wu, H., and Zhang, Y. (2014). Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. Cell *156*, 45–68.

Wu, Y., Dissing-Olesen, L., MacVicar, B.A., and Stevens, B. (2015). Microglia: Dynamic Mediators of Synapse Development and Plasticity. Trends Immunol. *36*, 605–613.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Xing, H., Wilkerson, D.C., Mayhew, C.N., Lubert, E.J., Skaggs, H.S., Goodson, M.L., Hong, Y., Park-Sarge, O.-K., and Sarge, K.D. (2005). Mechanism of hsp70i gene bookmarking. Science *307*, 421–423.

Xu, J., Pope, S.D., Jazirehi, A.R., Attema, J.L., Papathanasiou, P., Watts, J.A., Zaret, K.S., Weissman, I.L., and Smale, S.T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. Proc. Natl. Acad. Sci. *104*, 12377–12382.

Xu, W., Liyanage, V.R.B., MacAulay, A., Levy, R.D., Curtis, K., Olson, C.O., Zachariah, R.M., Amiri, S., Buist, M., Hicks, G.G., et al. (2019). Genome-Wide Transcriptome Landscape of Embryonic Brain-Derived Neural Stem Cells Exposed to Alcohol with Strain-Specific Cross-Examination in BL6 and CD1 Mice. Sci. Rep. *9*, 206.

Yi, R., and Fuchs, E. (2011). MicroRNAs and their roles in mammalian stem cells. J. Cell Sci. *124*, 1775–1783.

Yu, N.-K., Baek, S., and Kaang, B.-K. (2011). DNA methylation-mediated control of learning and memory. Mol. Brain *4*, 5.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919.

Zhang, F.F., Cardarelli, R., Carroll, J., Fulda, K.G., Kaur, M., Gonzalez, K., Vishwanatha, J.K., Santella, R.M., and Morabia, A. (2011). Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. Epigenetics *6*, 623–629.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhou, F.C., Zhao, Q., Liu, Y., Goodlett, C.R., Liang, T., McClintick, J.N., Edenberg, H.J., and Li, L. (2011). Alteration of gene expression by alcohol exposure at early neurulation. BMC Genomics *12*, 124.

Zieff, C.D., and Schwartz-Bloom, R.D. (2008). Understanding Fetal Acohol Spectrum Disorders (FASD) : A Comprehensive Guide for Pre-K - 8 Educators.

Annexes

Pendant ma thèse, j'ai contribué à la réalisation :

- d'un chapitre de livre intitulé « ***Analyse des troubles neurodéveloppementaux à la lumière des modifications de l'épigénome : exemple du syndrome d'alcoolisation fœtale*** **»**. Ce chapitre de livre paraîtra prochainement dans l'ouvrage « La révolution biotechnologique et la médecine de demain ». Co-auteur de ce chapitre : **D. Sabéran Djoneidi** ; coordinateur scientifique : B. Schneider ; directrice éditoriale : V. Parroco - éditions **John Libbey Eurotext**.
  Ce chapitre est présenté en Annexe 6-1.
- d'un article scientifique sur le projet porté par A. de Thonel au sein de l'équipe. Cet article est intitulé « ***CBP/EP300 acetylates and stabilizes the stress-responsive Heat Shock Factor 2, a process compromised in Rubinstein-Taybi syndrome*** **»**.
  A. de Thonel, J. K. Ahlskog, R. Abane, G. Pires, V. Dubreuil, J. Berthelet, A. L. Aalto, S. Naceri, M. Cordonnier, C. Benasolo, M. Sanial, **A. Duchateau**, A. Vihervaara, M. C. Puustinen, F. Miozzo, M. Henry, D. Bouvier, J.-P. Concordet, P. Fergelot, É. Lebigot, A. Verloes, P. Gressens, D. Lacombe, J. Gobbo, C. Garrido, S. D. Westerheide, M. Petitjean, O. Taboureau, F. Rodrigues-Lima, M. Lancaster, S. Passemard, D. Sabéran-Djoneidi, L. Sistonen, V. Mezger
  Article actuellement en pré-print (https://doi.org/10.1101/481457) sur la plateforme bioRχiv (ID bioRχiv : 481457). Cet article est présenté en Annexe 6-2.

Cet article porte sur les mécanismes moléculaires permettant de stabiliser HSF2, qui est une protéine plutôt labile. Cette étude a mis en évidence que :

- l'acétylation de HSF2, au niveau de lysines bien particulières, permet de stabiliser la protéine.
- Cette acétylation est assurée par les lysines acétyl-transférases CBP/EP300, l'étude a notamment montré des interactions entre HSF2 et CBP au niveau de domaines particuliers.
- Dans des cellules mutées pour CBP ou EP300, issues de patients atteints du syndrome Rubinstein-Taybi (RSTS) - maladie rare caractérisée par des défauts congénitaux, dont des défauts neuro-développementaux, présentant une mutation de CBP ou EP300 - le facteur HSF2 est moins présent, et la voie de réponse au stress thermique par les HSF est altérée.

Pour cet article, j'ai réalisé des analyses bioinformatiques afin d'apporter des éléments de discussion quant aux rôles potentiels joués par les complexes HSF2-CBP ou HSF2-EP300. J'ai ainsi analysé des données de ChIP-seq : données ENCODE pour les protéines CBP et EP300, données de ChIP-seq de notre collaboratrice A. Vihervaara (laboratoire de L. Sistonen, Univ. de Turku, Finlande), pour HSF2, qui ont été ré-analysées pour le besoin de l'étude.
L'intégration de ces données a permis d'identifier des cibles génomiques communes à ces différents facteurs, pouvant correspondre à d'éventuels sites de co-occupation de ces facteurs. Cela permet de supposer que l'action conjointe des facteurs HSF2-CBP ou HSF2-EP300 au niveau de cibles génomiques communes particulières pourrait moduler le niveau transcriptionnel de ces cibles et avoir des conséquences fonctionnelles diverses, lors d'un dérèglement, tel que celui observé dans le RSTS.

Annexe 1 : Chapitre de livre

Analyse de troubles neurodéveloppementaux à la lumière de modifications de l'épigénome : exemple du syndrome d'alcoolisation fœtale

Préambule - *Le nom des auteurs* :
Agathe DUCHATEAU[*], Dr Délara SABERAN DJONEIDI[**]
[*]auteur des figures et co-auteur du chapitre, doctorante BioSPC, Université de Paris
[**]auteur correspondant, maître de conférences, Université de Paris

Adresse complète pour les deux auteurs :
Équipe « Interface entre Développement et Environnement »
Université de Paris
Bâtiment Lamarck
35, rue Hélène Brion
75205 Paris Cedex 13
Tél : 01 57 27 89 25
Fax : 01 57 27 89 12
Emails :
delara.saberan@univ-paris-diderot.fr
agathe.duchateau@hotmail.com

Abréviations utilisées :
5hmC : 5-hydroxy-méthyl-cytosine
5mC : 5-méthyl-cytosine
DNMT : *DNA methyl-transferase(s)*
DOHaD : *Developmental Origins of Health and Disease*
EPA : exposition prénatale à l'alcool
MBD : *methyl-binding domain*
MND : maladies neurodéveloppementales
RDM : région différentiellement méthylée
SAF : syndrome d'alcoolisation fœtale
SAM : S-adénosyl-méthionine
TSAF : troubles du spectre de l'alcoolisation fœtale

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

élément sous droit, diffusion non autorisée

Annexe 2 : Article bioRχiv - de Thonel et al. 2018

CBP/EP300 acetylates and stabilizes the stress-responsive Heat Shock Factor 2, a process compromised in Rubinstein-Taybi syndrome

Aurélie de THONEL[1,2,3,§,&], Johanna K. AHLSKOG[4,5,§], Ryma ABANE[1,2,3], Geoffrey PIRES[1,2,3,] Véronique DUBREUIL[1,2,3], Jérémy BERTHELET[2,6], Anna L. AALTO[4,5], Sarah NACERI[1,2,3], Marine CORDONNIER[8,9,10], Carène BENASOLO[1,2,3], Matthieu SANIAL[2,11], Agathe DUCHATEAU[1,2,3], Anniina VIHERVAARA[4,5,ψ], Mikael C. PUUSTINEN[4,5], Federico MIOZZO[1,2,3,θ], Mathilde HENRY[1,2,3,ξ,£], Déborah BOUVIER[1,2,3,] Jean-Paul CONCORDET[12,13,14], Patricia FERGELOT[15,16], Élise LEBIGOT[17], Alain VERLOES[3,18,19,20,21], Pierre GRESSENS[2,3,18,22], Didier LACOMBE[15,16], Jessica GOBBO[8,9,10], Carmen GARRIDO[8,9,10], Sandy D. WESTERHEIDE[23], Michel PETITJEAN[2,7], Olivier TABOUREAU[2,7], Fernando RODRIGUES-LIMA[2,6], Madeline LANCASTER[24], Sandrine PASSEMARD[3,18,19,20,21], Délara SABÉRAN-DJONEIDI[1,2,3], Lea SISTONEN[4,5,&,*], Valérie MEZGER[1,2,3, &,*]

[1]CNRS, UMR7216 Épigénétique et Destin Cellulaire, F-75205 Paris Cedex 13, France,
[2]Univ Paris Diderot, Sorbonne Paris Cité, F-75205 Paris Cedex 13, France,
[3]Département Hospitalo-Universitaire DHU PROTECT, Paris, France.
[4]Faculty of Science and Engineering, Åbo Akademi University, Turku, Finland,
[5]Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Turku, Finland.
[6]Unité BFA, CNRS UMR 8251 Biologie Fonctionnelle et Adaptative (BFA), F-75205 Paris Cedex 13, France.
[7]INSERM UMR-S 973 Molécules Thérapeutiques in silico (MTi), F-75205 Paris Cedex 13, France
[8]INSERM, UMR1231, Laboratoire d'Excellence LipSTIC, Dijon, France.
[9]University of Bourgogne Franche-Comté, Dijon, France.
[10]Département d'Oncologie médicale, Centre Georges-François Leclerc, Dijon, France.
[11]CNRS, UMR 7592 Institut Jacques Monod, Paris F-75205, France.
[12]Muséum National d'Histoire Naturelle, 75231 Paris Cedex 05, France.
[13]CNRS UMR 7196, 75231 Paris Cedex 05, France.
[14]INSERM U1154, 75231 Paris Cedex 05, France.
[15]Laboratoire Maladies Rares: Génétique et Métabolisme (MRGM), Université de Bordeaux, INSERM U1211, Bordeaux, France.
[16]Department of Medical Genetics, CHU de Bordeaux, Bordeaux, France.
[17]Service de biochimie-pharmaco-toxicologie, Hôpital Bicêtre, Hopitaux Universitaires Paris-Sud, 94270 Le Kremlin Bicêtre, France
[18] UMR 1141 PROTECT, INSERM, Université Paris Diderot, Sorbonne Paris Cité, F-75019 Paris, France.
[19]Faculté de Médecine Denis Diderot, Univ Paris Diderot - Sorbonne Paris Cité, Paris, France.
[20]Département de Génétique, Hôpital Robert Debré, AP-HP, Paris, France.
[21]Service de Neuropédiatrie, Hôpital Robert Debré, AP-HP, Paris, France.
[22]Centre for the Developing Brain, Department of Division of Imaging Sciences and Biomedical Engineering, King's College London, King's Health Partners, St. Thomas' Hospital, London, SE1 7EH, United Kingdom.
[23]Department of Cell Biology, Microbiology, and Molecular Biology, College of Arts and Sciences, University of South Florida, Tampa, Florida, United States of America
[24]MRC Laboratory of Molecular Biology, Cambridge Biomedical, Campus, Cambridge, UK.

Present addresses:
ψ Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
θ Department of Genetics and Evolution, Sciences III, University of Geneva, Geneva, Switzerland
ξ INRA, Nutrition et Neurobiologie Intégrée, UMR1286, Bordeaux, France
£ Université de Bordeaux, Nutrition et Neurobiologie Intégrée, UMR1286, Bordeaux, France

[&]Co-corresponding authors: aurelie.dethonel@univ-paris-diderot.fr ; 33 (0)1 57 27 89 25. lea.sistonen@btk.fi; Tel: +358-2-2153311; Fax: +358-2-3338000; valerie.mezger@univ-paris-diderot.fr; Tel: 33 (0)1 57 27 89 14; 33 (0)6 75 77 11 98; Fax: 33 (0)1 57 27 89 11.
[§]Co-first authors: these authors contributed equally to the work.
*Co-last authors : these authors contributed equally to the direction of the work.

Acknowledgments
We warmly thank the patients and their families for their participation in this study. We thank Slimane AIT-SI-ALI for helpful discussions and comments on the manuscript, Anne PLESSIS (Jacques Monod Institute, Paris, France) for helpful discussions on setting the SNAP-TAG technology, Anne VANET (Jacques Monod Institute, Paris, France) for helpful discussions on the HSF2 structural modelling. We thank Lauriane FRITSCH and Slimane AIT-SI-ALI (UMR7216, for HeLa-S3 cells and growth conditions for TAP-TAG analyses). We thank the Imaging Platform IMAGOSEINE and especially Nicole BOGETTO for her help in sorting the GFP-positive HeLa-S3 cells. We are grateful to Heinrich LEONHARDT (Ludwig-Maximilians University, Munich, Germany) for F3H cellular and molecular tools and Pierre-Antoine DEFOSSEZ and Laure FERRY (UMR7216) for helpful guidance in F3H and GFP-Trap experiments, Isabelle LEMASSON (East Carolina University, USA) for the KIX-GST constructs, and Sophie POLO (UMR7216) for SNAP-TAG vectors. We are grateful to Vincent El Ghouzzi for the kind gift of human induced pluripotent stem cells (hIPSCs). We thank Isabelle COUPRY and Benoit ARVEILER (CHU de Bordeaux, France) for primary skin fibroblasts from healthy donors. We thank the Institut Médical Jérôme Lejeune for the gift of Lymphoblastoid cells (patients 4 and 5). We are grateful to Delphine BOHL, and Stéphane BLANCHARD from Pasteur Institute (Rétrovirus et Transfert Génétique, INSERM U622) for their help in producing the retroviruses for TAP-TAG experiments in Hela-S3 cells. We thank Laure FERRY (UMR7216) and the Epigenomics Platform, as well as Sandra PIQUET (UMR7216) and the Microscopy Platform (UMR7216) for access to instruments and technical advice, and Clara GIANFERMI (UMR7216) for microscopy pictures of organoids and nSBs. We thank Isabelle Le PARCO and the staff from the Buffon animal housing facility at the Jacques Monod Institute (Paris Diderot University, Paris, France) and the *Bioprofiler* Platform at the UMR8251 Biologie Fonctionnelle et Adaptative for *in vitro* acetylation assays.

Funding information
VM was funded by the CNRS (Projet International de Coopération Scientifique PICS 2013-2015) for her collaboration with LS and by the Short Researcher Mobility France Embassy/MESRI-Finnish Society of Sciences and Letters; the Agence Nationale de la Recherche («NeuroHSF», Programme Neurosciences, Neurologie and Psychiatrie ANR-06-NEURO-024-01 and « HSF-EPISAME », SAMENTA ANR-13-SAMA-0008-01). LS was funded by the Academy of Finland, Sigrid Jusélius Foundation, Magnus Ehrnrooth Foundation and Cancer Foundation Finland. RA was supported by PhD Fellowships from Neuropôle Ile-de France and Fondation ARC, FM by PhD Fellowships from the CNRS and Fondation pour la Recherche Médicale and from a Postdoc Fellowship from SAMENTA ANR-13-SAMA-0008-01), AD by a Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation (MESRI) Doctoral Fellowship, and ADT, DSD, and DB by the PICS travel grant, and GP and MH by Master 2 Intership Fellowships from SAMENTA ANR-13-SAMA-0008-01. JB was supported by a PhD Fellowship from Région Ile-de-France (Cancéropôle IDF) and University Paris Diderot. JKA was supported by Magnus Ehrnrooth foundation. MCP was supported by the Turku Doctoral Network in Molecular Biosciences and Magnus Ehrnrooth Foundation. The supporting bodies played no role in any aspect of study design, analysis, interpretation or decision to publish this data.

Abstract

Cells respond to protein-damaging insults by activating heat shock factors (HSFs), key transcription factors of proteostasis. Abnormal levels of HSFs occur in cancer and neurodegenerative disorders, highlighting the strict control of their expression. HSF2 is a short-lived protein, which is abundant in the prenatal brain cortex and required for brain development. Here, we report that HSF2 is acetylated and co-localized with the lysine-acetyl transferases CBP and EP300 in human brain organoids. CBP/EP300 mediates the acetylation of HSF2 on specific lysine residues, through critical interaction between the CBP-KIX domain and the HSF2 oligomerisation domain, and promotes HSF2 stabilization. The functional importance of acetylated HSF2 is evidenced in Rubinstein-Taybi syndrome (RSTS), characterized by mutated CBP or EP300. We show that cells derived from RSTS patients exhibit decreased HSF2 levels and impaired heat shock response. The dysregulated HSF pathway in RSTS opens new avenues for understanding the molecular basis of this multifaceted pathology.

INTRODUCTION

Since their discovery three decades ago, our way to envision the regulation and roles of the Heat Shock transcription Factor family (HSFs) has been revolutionized. Originally identified and characterized due to their stress-responsiveness and ability to recognize a consensus DNA-binding site, the heat shock element (HSE), HSFs were more recently shown to perform an unanticipated large spectrum of roles under physiological and pathological conditions (Wu, 1995; Abane and Mezger 2010; Akerfelt et al., 2010; Pastor-Gomez et al., 2018). HSFs are activated by a diversity of stressors that provoke protein damage and govern the highly conserved Heat Shock Response (HSR). The HSR contributes to the restoration of proteostasis, through the regulation of genes encoding molecular chaperones, including the Heat Shock Proteins (HSPs; Hartl et al., 2011). HSFs also control immune/inflammatory pathways, metabolism, and, through dysregulation of their protein levels or activity, shape disease susceptibility to cancer, metabolic and neurodegenerative disorders. These pathophysiological roles are performed through altered expression of a broad repertoire of target genes, beyond the *HSPs* (Xiao et al., 1999; Inouye et al., 2007; Dai et al., 2007; Mendillo et al., 2012; Santagata et al., 2013; Anckar and Sistonen, 2011; Jin et al., 2011; Neef et al., 2011; Nakai, 2016; Pastor-Gomez et al., 2017 and 2018). The multifaceted roles of HSFs are achieved by their fascinating plasticity in terms of multi-modular structure and assembly of homo- or heterodimers or trimers, stress- and context-dependent posttranslational modifications, as well as a diversity of partner networks. As a consequence, HSFs act as fine sculptors of transcriptomic and epigenetic landscapes, through dynamic interactions with other transcriptional activators or repressors and chromatin remodelling complexes (Akerfelt et al., 2010; Miozzo et al., 2015; Pastor-Gomez et al., 2018; Raychaudhuri et al. 2014).

The versatile functions of HSFs have been mostly studied with HSF1 and HSF2, two members of the mammalian HSF family, which in human comprises four additional members, *i.e*. HSF4, HSF5, HSFX and HSFY (Pastor-Gomez et al., 2018). The role of HSF1 in acute and severe proteotoxic stress, including exposures to elevated temperatures (42-45°C), has been extensively documented and has become a paradigm for the *modus operandi* of the HSF family. In contrast, HSF2 appears to be responsive to stresses of relevance for chronic or pathological situations, such as fever-like temperatures at 39-41°C (Shinkawa et al., 2011), alcohol (ethanol) exposure (El Fatimy et al., 2014; Miozzo et al., 2018), and prolonged proteasome inhibition (Lecomte et al., 2010; Rossi et al., 2014). Both factors have been associated with different forms of cancer, in a dual manner; HSF1 acts as a potent facilitator of cancer initiation and progression (Dai et al., 2007; Mendillo et al., 2012; Santagata et al., 2013), whereas HSF2 can counteract tumor progression and invasiveness (Björk et al. 2016). The control of HSF1 protein levels is key for its pathophysiological functions, as elevated expression of HSF1 correlates with poor cancer prognosis (Santagata et al., 2011) and decreased

expression has been found in different neurodegenerative disorders (Kim et al., 2015; Jiang et al., 2013; Pastor-Gomez et al., 2017; reviewed in Pastor-Gomez et al., 2018). HSF1 and HSF2 are also important players in the physiological brain development and adult brain integrity. Furthermore, we and others have shown that deregulated their activities underlie both neurodevelopmental defects (Kallio et al., 2002; Wang et al., 2003; Chang et al., 2006; El Fatimy et al., 2014; Hashimoto-Torii et al., 2014; Ishii et al., 2017; reviewed in Abane et Mezger 2010, Åkerfelt et al., 2010, Pastor-Gomez et al., 2018) and neurodegenerative processes (Shinkawa et al., 2011; Pastor-Gomez et al., 2017).

The amount of HSF2 protein varies in diverse cellular or embryonic contexts and conditions: both transcriptional and post-transcriptional mechanisms have been shown to regulate *HSF2* mRNA levels (Rallu et al., 1997; Björk et al., 2010). In the developing brain, the protein levels of HSF2 seem to correlate with those of *Hsf2* transcripts, emphasizing the importance of the regulatory mechanisms of *Hsf2* gene expression (Rallu et al., 1997; Kallio et al., 2002; Wang et al., 2003). Moreover, HSF2 is a short-lived protein and its stabilization constitutes an important step controlling the DNA-binding activity of HSF2 (Sarge et al., 1993; Mathew et al., 1998; Kawazoe et al., 1998) and mediating its role in physiological processes and stress responses. HSF2 protein levels also fluctuate during the cell cycle, which further shows that stabilization of HSF2 provides with a critical control step in fine-tuning the HSR (Elsing et al., 2014). Indeed, HSF2 modulates the stress-inducible expression of *HSP* genes, which is primarily driven by HSF1 (Östling et al., 2007). This transient modulatory function of HSF2 is due to the rapid poly-ubiquitination and proteasomal degradation in response to acute heat stress (Ahlskog et al., 2010). While diverse posttranslational modifications (PTMs), such as phosphorylation and acetylation, are well known to control HSF1 stability (Kourtis et al., 2015; Pastor-Gomez et al., 2017; Raychaudhuri et al., 2014; reviewed in Pastor-Gomez et al., 2018), the mechanisms regulating the stability of HSF2 are poorly understood, and given its role in chronic stress, cancer, and physiopathological developmental processes, they are crucial to be elucidated.

The histone/lysine-acetyl transferases (HATs/KATs) CBP (CREBBP, CREB-binding protein; KAT3A) and EP300 (E1A-binding protein p300; KAT3B) control the stability of many transcription factors through their acetylation, including HSF1 (Thakur *et al*., 2013; Raychaudhuri et al., 2014). Heterozygous mutations in one of these KATs lead to Rubinstein-Taybi syndrome (RSTS; Lopez-Atalaya et al., 2014; Spena et al., 2015a). RSTS is a rare disease characterized by multiple congenital anomalies, neurodevelopmental defects, childhood cancer susceptibility, and vulnerability to infections (*CREBBP/CBP* mutation, RSTS1, OMIM #180849*; EP300* mutation, RSTS2; OMIM #613684). Here, we show that HSF2 is acetylated during normal brain development in human organoids, where it is expressed in the same territories as CBP and EP300. We demonstrate that CBP/EP300 mediates the acetylation of HSF2 on specific lysine residues, through critical interaction between the CBP-KIX domain and the HSF2 oligomerisation domain, thereby promoting the stabilization of the HSF2 protein. We then interrogate the functional importance of this regulation in the pathological context of RSTS. We observe a proteasomal-dependent reduction in HSF2 protein levels in cells derived from RSTS patients, which results in impairement of their ability to mount a proper heat shock response. The disruption of the HSR pathway in RSTS highlights the importance of the CBP/EP300-dependent regulation of HSF2 by acetylation and provides a new conceptual frame for understanding the molecular basis of this complex pathology.

RESULTS

HSF2 is acetylated and interacts with CBP/EP300 in the developing brain

HSF2 is abundantly expressed in the vertebrate developing brain, where it exhibits spontaneous DNA-binding activity, in unstressed, physiological conditions (Rallu et al., 1997; Kawazoe et al. 1999; Kallio et al., 2002; Wang et al., 2003; Chang et al., 2006). As a first step to determine whether the

acetylation of HSF2 was involved in controlling its stability, similarly to the EP300-mediated acetylation of HSF1 (Raychaudhuri et al., 2014), we compared HSF2 and CBP/EP300 expression profiles, and investigated whether HSF2 acetylation could be detected in the developing mammalian cortex, since CBP and EP300 have key roles in neurodevelopment (reviewed in Chan and La Thangue 2001; Lopez-Atalaya et al., 2014).

To the best of our knowledge, the expression of the HSF2 protein in the human developing cortex has not been reported. By generating brain organoids from human embryonic stem cells (Lancaster et al., 2014), we confirmed that, HSF2 mRNAs are present in human brain organoids, as previously reported (Figure S1A; Camp et al., 2015) and observed that *CBP* and *EP300* mRNA are also present. We found that the HSF2 protein was expressed at different stages, from day 20 (embryoid bodies) to day 60 of differentiation, with a profile similar to that of CBP and EP300 (D20 – D60; Figure 1A). By immunofluorescence on D60 organoids, we found that HSF2 was expressed in neural progenitor cells (NPCs; located in areas of dense DAPI-staining), as verified by SOX2 staining (Figure S1B) and in neurons, expressing beta-III tubulin, in regions displaying a cortical-like morphology (Figure 1B and S1B). In addition, we observed co-labeling of EP300 and HSF2 in NPCs (arrowheads) and neurons (arrows) (Figure 1C). Thus, HSF2 and EP300 (and also CBP; see Figure S1C) exhibited similar expression territories (NPCs and neurons; Figure S1B,C). In addition, HSF2, CBP and EP300 were expressed, in a concomitant manner, in the mouse cortex from E11 to E17 (Figure S1D). The similarity of their expression patterns suggested an interaction between HSF2 and CBP/EP300 in the developing cortex. Accordingly, HSF2 was co-immunoprecipitated with CBP and EP300 (Figure 1D and Figure S1E, left panels). We also detected acetylated HSF2 in the developing mouse cortex (Figure 1E and Figure S1E, middle panels) and in D40 human brain organoids (Figure 1F). Similarly, HSF2 was found acetylated in SHSY-5Y neuroblastoma cells, a human cancer cell line of neural origin (Figure S1F). Altogether, these results show that HSF2 is acetylated and interacts with CBP and EP300 in the developing brain.

Analysis of CBP/EP300-mediated HSF2 acetylation

In order to explore the mechanism of HSF2 acetylation, we first examined whether HSF2 was a substrate for acetylation by CBP/EP300, in human HEK 293 cells co-expressing CBP-HA or EP300-HA and GFP- or Myc-tagged HSF2. We found that the immunoprecipitated exogenous HSF2 protein was acetylated by EP300 or CBP (Figure 2A), but that no acetylation was observed in cells transfected by dominant-negative CBP, unable to catalyze acetylation (Figure 2B).

To identify the acetylated lysine residues in HSF2, we co-expressed Flag-HSF2 with EP300-HA in HEK 293 cells. HSF2 was immunoprecipitated and the acetylation of lysines was analyzed by mass spectrometry (MS). Among the 36 lysine residues of HSF2, we identified eight acetylated lysines: K82 (located in the DNA-binding domain), K128, K135, K197 (all located within the hydrophobic heptad repeat HR-A/B), K209, K210, K395, and K401 (Figure 2C, Figure S2A, and Table S1 and S2). Single point mutations (K82, K128, K135, and K197), or mutation of the doublet K209/K210 to arginine (R, which prevent acetylation), did not abolish global HSF2 acetylation (Figure S2B). This suggests that, in line with our MS data, the acetylation of HSF2 occurs on more than one lysine residue. Indeed, the mutation to either arginine (R) or glutamine (Q) of the three or four lysines K82, K128, K135 and K197, dramatically reduced HSF2 acetylation (Figure 2D and Figure S2C). To dissect the requirement of CBP in the acetylation of HSF2, we used an *in vitro* acetylation assay coupled with HPLC (high-performance liquid chromatography). We found that a synthetic HSF2 peptide containing either K135 or K197 residues was readily acetylated by the purified recombinant full-catalytic domain (Full-HAT) of CBP (Figure 2E; Figure 3A) in an acetyl-CoA-dependent manner, whereas a peptide containing K82 was not (Figure S2D-F). Taken together, our data suggest that HSF2 is acetylated by CBP/EP300 at three main lysine residues, residing in the oligomerization HR-A/B domain: K128, K135 and K197.

HSF2 is a *bona fide* substrate of the core catalytic domain of CBP

Prompted by the finding that a catalytically active CBP is necessary for HSF2 acetylation, we examined whether HSF2 could bind to the core catalytic domain of CBP. The CBP central core catalytic region ("Full-HAT"; Figure 3A) contains the Bromodomain BD, the cysteine/histidine-rich region CH2, and the HAT domain and allows the coupling of substrate recognition and histone/lysine acetyltransferase activity (as in EP300; Delvecchio et al., 2013; Dancy and Cole, 2015; Dyson and Wright 2016). The CH2, in particular, contains a RING domain and a PHD (plant homeodomain; Park et al., 2013). With biolayer interferometry, we observed that the recombinant Full-HAT domain directly interacted with immobilized biotinylated recombinant full-length HSF2 (Figure 3B). Within this region, the recombinant PHD domain, but not the HAT, RING or BD domain, was able to interact with HSF2, in a similar manner as the "Full-HAT" domain (Figure 3B). Interestingly, the interaction of HSF2 with the Full-HAT or the PHD domain was more efficient than with HSP70, which has been reported to interact with HSF2 (Huttlin et al., 2015; Tang et al., 2016). As expected, it is likely that the interaction between HSF2 and the catalytic HAT domain was too transient to be captured in these experiments, because the HAT domain needs other CBP domains to interact with its substrates, including the PHD (Aasland et al., 1995; Bordoli et al., 2001; Kalkhoven et al, 2002). We determined that the K_D of HSF2 interaction with the CBP Full-HAT domain was $1.003E^{-09}$ M (+/-$2.343E^{-11}$; R2= 0.988488; Figure S3A). Our data on HSF2 acetylation and interaction with the Full-HAT domain of CBP, including PHD therein, strongly suggest that HSF2 is a *bona fide* substrate of CBP, and potentially also that of EP300, since their HAT domains display 86% identity.

HSF2 interacts with CBP and EP300 *via* its oligomerisation HR-A/B domain

Our studies in human brain organoids and mouse cortices indicated that endogenous HSF2 and CBP/EP300 could interact (Figure 1). Similar results were observed in the murine neuroblastoma N2A cells (Figure S3B). We started to dissect the mode of anchorage between these proteins. We first tested whether exogenous proteins, tagged-HSF2 and -CBP/-EP300 could interact. Using HEK 293 cells in GFP-Trap assay, we showed that CBP-HA or EP300-HA co-immunoprecipitated with HSF2-YFP (Figure 3C,D). Second, we confirmed that interaction between these tagged proteins occurred *in cellulo* by an independent imaging technique, the fluorescent three-hybrid assay (F3H; Figure 3E; Herce et al., 2013), using GFP-binder and HSF2-YFP, together with CBP-HA or EP300-HA. In negative control experiments, GFP-binder, which was recruited to the *Lac0p* array locus, was unable to recruit endogenous CBP, CBP-HA, or EP300-HA (Figure S3C-E). Similarly, HSF2-YFP was unable to locate at the *Lac0p* array locus in the absence of GFP-binder (Figure S3C-E). Upon co-transfection with HSF2-YFP, CBP-HA or EP300-HA expression resulted in the formation of a red spot in the nucleus, showing co-recruitment to the HSF2-YFP focus (green spot), in 68.4% and 48.6% of the cells, respectively (Figure 3F, upper panels and 3G). The abundance of CBP in BHK cells allowed us to detect the co-recruitment of HSF2-YFP with endogenous CBP in 43.9% of these cells (Figure 3F, lower panels).
Having characterized the interaction between exogenous tagged proteins, we then asked by which specific domains the anchorage between HSF2 and CBP was facilitated. To determine which HSF2 domains were important for its interaction with CBP, we expressed Flag-HSF2 deletion mutants of different domains. We showed that the deletion of HR-A/B domain (but not of the DBD) led to a marked decrease in HSF2 acetylation (Figure S4A-C; see red arrow). These results were in line with our findings that the major acetylated lysine residues reside within the HR-A/B domain. The deletion of the TAD (transcription activation domain; Jaeger et al., 2016) also resulted in decreased acetylation of the Flag-HSF2 (Figure S4A-C) and was associated with decreased interaction with CBP (Figure S4D). This domain may be important for interaction between CBP and HSF2, since the multivalent interactions between CBP/EP300 and many transcription factors generally involve their TADs (Lee et al., 2009; Wang et al., 2012; reviewed in Thakur *et al*., 2013).

The presence of KIX motifs in the HSF2 HR-A/B domain promotes binding to the CBP KIX domain

CBP/EP300 interacts with many transcription factors *via* different binding sites, including the KIX domain (kinase-inducible domain interacting domain; Figure 3A, left). The KIX domain contains two distinct binding sites that are able to recognize the "□XX□□" KIX motif, where "□" is a hydrophobic residue, and "X" is any amino acid residue (Radhakrishnan et al., 1997; Kobayashi et al., 1997; Lee et al., 2009; Zor et al., 2004). Importantly, we identified several conserved, overlapping and juxtaposed KIX motifs in the HR-A/B domain of HSF2 (Figure 4A). We modeled the interaction between the HSF2 HR-A/B KIX motifs and the CBP KIX domain. Based on sequence similarities between the HR-A/B domain, lipoprotein Lpp56, and the transcription factors GCN4, ATF2, and PTRF (Figure S4E; see Experimental procedures), we first developed a structural model of the HSF2 trimeric, triple coiled-coil, HR-A/B domain (Figure S4F; Jaeger et al., 2016). Second, we investigated the possibility of interactions of the KIX recognition motifs in the HR-A/B region with the CBP KIX domain. Best poses suggested that the HR-A/B KIX motif region contacted the so-called "c-Myb surface" within the KIX domain (Figure 4B, Thakur *et al*., 2013), thereby proposing a close interaction of the HSF2 KIX motifs with the tyrosine residue Y650 of CBP (Figure 4C; Figure S4G(b)). We next examined the impact of *in silico* mutations of the K177, K180, F181, V183 residues, which are present within the KIX motifs of HSF2 and involved in the contact with CBP (Figure 4D). Either K177A or Q180A mutation within the HSF2 KIX motifs disrupted HSF2-KIX domain interaction (Figure S4G (e,f); Table S3), in contrast to either F181A or V183A mutation (Figure S4G(c,d)). Finally, we assessed the impact of *in silico* mutation of the Y650 amino acid of the CBP KIX domain, a residue mutated in RSTS patients (Figure 4E). Interestingly, the *in silico* mutation Y650A in CBP profoundly decreased the probability of interaction of the HSF2 KIX motifs with the KIX domain (Figure 4F, upper panel; and Figure S4G (b); Table S3). Using recombinant proteins, we verified that HSF2 directly interacted with the CBP KIX domain in *in vitro* co-immunoprecipitation experiments (Figure 4B) and we confirmed that the Y650A mutation disrupted HSF2 and KIX interaction (Figure 4G). Thereby, we identify Y650 as a residue critical for interaction between the KIX domain of CBP and the KIX motifs within the HSF2 oligomerisation domain.

The acetylation of HSF2 governs its stability under non-stress conditions

To explore the functional impact of the CBP/EP300-mediated acetylation of HSF2, we inhibited CBP/EP300 activity in N2A cells, using the specific inhibitor C646 (Bowers et al., 2010; Dancy and Cole, 2015). The pharmacological inhibition of CBP/EP300 decreased the endogenous HSF2 protein levels, which was abolished by treatment with the proteasome inhibitor, MG132 (Figure 5A and Figure S5A). These results showed that the decrease in the HSF2 protein levels was dependent on the proteasomal activity, and that HSF2 was degraded when CBP/EP300 activity was inhibited (Figure 5A), thereby providing the first evidence for acetylation playing a regulatory role in HSF2 stability. To further investigate the role of acetylation in the regulation of HSF2 protein levels, we generated CRISPR/Cas9 *Hsf2*KO U2OS cell lines (2KO; Figure S5B,C) and measured the protein levels of exogenous wild-type HSF2 or HSF2 acetylation mutants, which mimic either constitutively acetylated (3KQ) or non-acetylated (3KR) HSF2 (Figure 5B-E). We first verified that the HSF2 WT, 3KQ and 3KR were expressed at comparable levels (Figure S5D), and capable of binding DNA *in vitro* (Figure S5E). Our *ex vivo* experiments also verified that they were also able to locate into specific subnuclear structures, called the nuclear stress bodies, nSBs (Jolly et al, 1997, 1999, 2004; Rizzi et al., 2004; Sandqvist et al., 2009), whose formation upon stress is associated to the recruitment of HSF1 and HSF2 on pericentromeric repeats, predominantly at the *SatIII* 9q12 locus (Figure S5F). To monitor the decay of a pre-existing pool of HSF2 molecules, we performed pulse-chase experiments using the SNAP-TAG technology (Bodor et al., 2012). A pool of SNAP-HSF2 molecules was covalently labeled by adding a fluorescent substrate to the cells. At t_0, a blocking non-fluorescent substrate was added, quenching the incorporation of the fluorescent substrate to newly synthesized HSF2 molecules

(Figure 5B), allowing us to measure the decay in the fluorescence intensity of the corresponding labeled HSF2 bands. When 2KO cells were transfected with wild-type SNAP-HSF2 (SNAP-HSF2 WT), a ~50% decay in fluorescence intensity of the corresponding bands was observed within 5 hours (Figure 5C and D). Preventing HSF2 acetylation (SNAP-HSF2 3KR) resulted in a similar decay (Figure 5C and D). In contrast, mimicking acetylation with SNAP-HSF2 3KQ protected HSF2 from decay (Figure 5C and D). Of note, proteasome inhibition with MG132 prevented the decrease in SNAP-HSF2 WT and 3KR fluorescent intensity (Figure 5E). Moreover, we observed that mimicking the acetylation of HSF2 by expressing Myc-HSF2 3KQ limited the poly-ubiquitination of HSF2, when compared to HEK 293 cells expressing either WT or 3KR HSF2 (Figure 5F). (Figure 5F). Altogether these experiments demonstrate that HSF2 acetylation prevents the proteasomal degradation of HSF2.

HDAC1 is involved in the destabilization of the HSF2 protein under non-stress and stress conditions

To identify the enzymes that could function as deacetylases for HSF2, we performed an unbiased screen for HSF2 binding protein partners, using a double-affinity TAP-TAG approach (Bürkstümmer et al., 2006; Figure S6A-C). For this purpose, we generated a HeLa-S3 cell line expressing double-tagged HSF2 (or transfected with the empty vector as a negative control) and analyzed nuclear extracts by MS. We identified HDAC1 as one of the protein partners of HSF2 (Figure 6A). In addition to HSF2, we found nucleoporin Nup62 (Figure 6A), a known HSF2 partner that served as a positive control for the quality of our TAP-TAG/MS analysis (Yoshima et al., 1997). We also performed immunoprecipitation of HSF2 in extracts from the mouse E17 cortices, followed by MS analysis, and found both HDAC1 and HDAC2 as HSF2 partners (Figure S6D). Using the F3H approach (Figure 3E; Figure 6B and Figure S6E), and co-immunoprecipitation in GFP-Trap assays (Figure 6C), we confirmed the interaction between HSF2 and HDAC1 in mammalian cell lines. We then evaluated the impact of HDAC1 and other Class I HDACs on the acetylation of HSF2 by expressing CBP-HA in HEK 293 cells (Figure 6D and Figure S6F). HDAC1 overexpression resulted in marked reduction in HSF2 acetylation levels, whereas HDAC2 and HDAC3 had only limited effects, no effect of HDAC8 was detectable (Figure 6D and Figure S6F).
Because heat shock (HS) provokes the degradation of the HSF2 (Ahlskog et al., 2010), we thus analyzed the impact of acetylation on the heat-shock-induced decay of HSF2, using the SNAP-TAG technology. Mimicking the acetylation of the three major acetylated lysine residues mitigated the decay of fluorescence intensity of SNAP-HSF2 3KQ induced by HS, compared to SNAP-HSF2 WT or 3KR (Figure 6E). The impact of HDAC inhibition on endogenous HSF2 was investigated in N2A cells. We first verified that HS was able to induce HSF2 decay also in N2A cells, although it occurred at a slower rate than in HeLa or HEK 293 cells (Figure S6G; Ahlskog et al., 2010). Treatment with 1 mM of the Class I inhibitor VPA dampened the decline in HSF2 protein levels in N2A cells exposed to HS (Figure 6F and Figure S6H). This indicates that Class I HDAC activity participates to the degradation of HSF2 upon HS, likely through HSF2 deacetylation. We then verified that HS increased HSF2 poly-ubiquitination in HEK293 cells, as previously reported (Ahlskog et al. 2010; Figure S6I, mock transfection). To investigate whether HDAC1 could favor HSF2 poly-ubiquitination, likely through HSF2 deacetylation, we examined the impact of overexpression of a dominant-negative form of HDAC1 on HSF2 ubiquitination. We showed that, indeed, the increase in HSF2 poly-ubiquitination upon HS was mitigated in HEK 293 cells transfected with dominant-negative HDAC1 (Suppl. Figure S6I). As a whole, our results support a role of HDAC1 (and possibly other Class I HDACs) in the destabilization of HSF2 under normal and stress conditions, through HSF2 poly-ubiquitination and proteasomal degradation.

Declined HSF2 protein levels in the Rubinstein-Taybi Syndrome (RSTS)
To determine the functional impact of CBP and EP300 on HSF2 levels in a pathological context, we compared the amounts of HSF2 protein in cells derived from either healthy donors (HD) or RSTS

patients, which are characterized by autosomal-dominant (heterozygous) mutations in the *CBP* or *EP300* genes (see Figure S7A for a description of the mutations). We used human primary skin fibroblasts (hPSFs), at early passages to avoid putative compensation processes during *ex vivo* culture (see Experimental Procedures). We verified the effect of mutated *CBP* or *EP300* in RSTS hPSFs by showing that the amount of acetylated lysine residue K27 in histone H3 (AcH3K27) was reduced in both cases, compared to healthy donors (patient 1 [P1] and patient 2 [P2], respectively, Figure S7B). We observed that HSF2 protein levels were markedly decreased in hPSFs from RSTS patients carrying either *CBP* or *EP300* mutations (Figure 7A-C). HSF2 levels were restored to levels that were comparable to those of healthy donors (HD) when these hPSFs had been treated with the proteasome inhibitor MG132 (Figure 7A and B). However, Class I HDAC inhibition by VPA could not restore the HSF2 levels in $RSTS_{CBP}$ or $RSTS_{EP300}$ hPSFs, although HD and RSTS cells displayed similar levels of HDAC1 (Figure 7B; Figure S7C-E). Notably, the stability of HSF2 was impaired both in the $RSTS_{CBP}$ patient carrying a mutation in the Full-HAT domain of CBP and in the $RSTS_{EP300}$ patient carrying a deletion of the KIX domain of EP300. This finding suggests that both domains are required for the regulation of HSF2 stability, which is in line with our results (Figure 3-5). Altogether these results demonstrate that the proteasomal turnover of HSF2 is increased in RSTS hPSFs, carrying mutated either *EP300* or *CBP*, thereby strongly indicating that EP300 and CBP are key regulators of HSF2 protein stability.

Impaired heat shock response in RSTS cells

HSF1 is the essential driver of the acute heat shock response (HSR) in mammals (McMillan et al., 1998). Although dispensable for the HSR in most cellular contexts (McMillan et al., 1998), HSF2 acts as a fine tuner of the HSR (Östling et al., 2007; Elsing et al., 2014), which determines the magnitude to which the applied heat stress induces *HSP* gene expression. Therefore, we evaluated the ability of RSTS cells to mount a HSR. In the absence of heat stress, we observed that RSTS hPSFs displayed lower amounts of HSP70 and HSP90 than their HD counterparts (Figure 7D). Furthermore, RSTS hPSFs exhibited limited capacity in inducing HSP70 accumulation upon HS and during the recovery phase from heat stress (Figure 7D). Importantly, this limited induction did not result from impairment of HSF1 activation, since HSF1 was activated by HS in RSTS hPSFs, as assessed by its slowed mobility shift in SDS-PAGE (see arrowheads in Figure 7D). This shift is a hallmark of HSF1 hyperphosphorylation, which, although not required for HSF1 activation, accompanies the induction of HSF1 transactivation potential (Sarge et al., 1993; Budzyński et al., 2015; reviewed in Anckar and Sistonen, 2011). As mentioned above, HSF1 and HSF2 do not only control the transcription of the *Hsp* genes in response to acute heat stress, but they also upregulate the transcription of *Sat III* 9q12 heterochromatin regions, where nSBs are formed. We therefore used nSBs as a read-out for assessing the HSR integrity in RSTS cells and observed that the stress-inducible formation of nSBs was reduced by more than 50% in $RSTS_{EP300}$ hPSFs when compared to their HD counterparts (Figure 7E at 43°C and S7F at 42°C). A similar reduction in the formation of nSBs was observed in RSTS lymphoblastoid cells (LBs; Figure S7G).
The functional importance of the regulation of HSF2 stability by CBP and EP300 is therefore highlighted in the RSTS context, under non-stress and stress conditions, which might constitute an interesting novel reading key for this complex disease.

DISCUSSION

During the last decade, HSFs have been associated with a wide spectrum of pathophysiological conditions, and the specific roles of HSFs, either individually or in combination with each other or with other transcription factors, are of great biomedical interest. Especially, the mechanisms by which the expression levels of HSFs are regulated, in a context-dependent manner, have remained poorly understood. However, there is a wealth of documented cases where either excessive or insufficient HSF protein levels favor the development or progression of devastating diseases, including cancer, neurodevelopmental, and neurodegenerative disorders. In this study, we reveal a novel mechanism that regulates the stability of HSF2 protein. Using different cellular systems and pathophysiological conditions, we found that the KATs CBP and EP300 catalyze the acetylation of three highly conserved lysine residues K128, K135, and K197, located in the HR-A/B oligomerization domain, thereby contributing to the stability of the HSF2 protein. Moreover, we demonstrate the importance of this regulation both under normal and stress conditions, as well as in the pathological conditions of the Rubinstein-Taybi Syndrome (RSTS), which is a rare but highly detrimental disease.

HSF2 is acetylated by CBP/EP300 in various contexts including brain development

In addition to ectopically/exogenously expressed proteins, HSF2 is acetylated by the overexpression of CBP or EP300 in cell systems. Importantly, we were able to detect the acetylation of the endogenous HSF2 protein in human and murine neural embryonic tissues and cell lines (Figure 1 and S1), showing that HSF2 acetylation is not restricted to one specific cell context. As does HSF2, CBP and EP300, and more generally HATs play significant roles during neurodevelopment (reviewed by Lopez-Atalaya et al., 2014). HSF2, CBP and EP300 exhibit similar expression patterns along the differentiation of human brain organoids, as they do in the developing mouse cortex. This is, to the best of our knowledge, the first report on the expression profiles of HSF2, CBP, and EP300 proteins in a model of the human developing brain as well as on the interactions between HSF2 and CBP/EP300. Our results point out the physiological importance of HSF2 acetylation and suggest that the acetylation of HSF2 might be a key event involved in the abundant expression and important role of HSF2 in cortical development. Accordingly, it is interesting to note that, our MS analysis revealed HDAC1 as an HSF2 partner in E17 cortices, at which stage the HSF2 protein levels are markedly dowregulated (Figure S7; El Fatimy et al., 2014).

Mode of HSF2 interaction with CBP/EP300

In our attempt to determine the molecular and structural basis for interaction between HSF2 and the acetylating enzymes CBP and EP300, we first show that the full-length HSF2 protein interacts with the CBP core catalytic domain *in vitro*, confirming that HSF2 is a *bona fide* substrate of CBP. In addition, HSF2 strongly interacts with the PHD domain, located within the catalytic core of the CBP/EP300 proteins (Delvecchio et al., 2013). Interestingly, mutations in the PHD domain of CBP/EP300 have been identified in RSTS patients (Kalkhoven et al., 2003). Based on the cell-based analyses combined with the *in vitro* and *in silico* analyses, we found that the HR-A/B oligomerization domain, but not the DNA-binding domain, is necessary for the interaction of HSF2 with CBP and for HSF2 acetylation. Importantly, we show that the HR-A/B domain specifically interacts with the KIX domain of CBP/EP300.

The CBP and EP300 KIX domain serves as a docking site for the binding of many transcription factors and contributes to the properties of CBP/EP300 to act as a molecular bridge, stabilizing the interactions between specific transcription factors and the transcription machinery (Parker et al., 1996; reviewed in Thakur et al., 2014). According to our *in silico* analyses, the KIX-binding motifs that we identified in the HR-A/B oligomerization domain of HSF2 are necessary for its specific interaction with the KIX domain of CBP/EP300, and thus for HSF2 acetylation. In support of the close interaction between the HSF2 KIX recognition motifs and the CBP KIX domain, we show that the mutation of the tyrosine residue Y650 in the CBP KIX domain disrupts HSF2-CBP interaction in *in silico* and *in vitro* experiments. Likely a similar mode of interaction between the HSF2 KIX motif and the homologue of

Y650 in EP300 can be expected (Kauppi et al., 2008). Interestingly, this mutation has been identified in RSTS, and associated with a severe neurodevelopmental phenotype (Spena et al., 2015b). Moreover, our *in silico* analyses indicate that the HR-A/B KIX motifs in HSF2 bind to the c-Myb site of the KIX domain. Indeed, the CBP or EP300 KIX domain can, simultaneously and in a cooperative manner, bind two polypeptide ligands (from two transcription factors), on two distinct surfaces, which have been historically called the "c-Myb" and the "MLL" (*Mixed Lineage Leukemia* protein) sites (Goto et al., 2002; Campbell and Lumb, 2002; reviewed in Thakur et al., 2014). This suggests that the binding of another transcription factor via the "MLL site" might potentially modulate the interaction between HSF2 and CBP, through its the KIX domain.
In addition to the HR-A/B oligomerization domain, we also identify the HSF2 TAD (Jaeger et al., 2016) as a potentially important domain for HSF2 interaction with CBP. This result is reminiscent of the TADs of other transcription factors that bind the KIX domain (reviewed in Thakur et al., 2014). Moreover, two different domains of the same transcription factor can simultaneously bind the KIX domain (Lee et al., 2009; Wang et al., 2012). We therefore hypothesize that HSF2 could simultaneously interact with CBP, through two distinct domains, the HR-A/B and the TAD domains. In addition, the dimeric or trimeric coil-coiled structure of HSF2 might also broadens the possibility of establishing multiple contacts with CBP (and most likely EP300), through KIX or other CBP domains, known to also interact with TADs, which paves the way for future studies.

Dynamics of HSF2 acetylation by CBP/EP300, deacetylation by HDAC1, and degradation
Based on our data, the acetylation of HSF2 by CBP/EP300 limits its proteasomal degradation, which has been observed for other transcription factors, such as p53, STAT3, and HIF1alpha (Grossman, 2001; Jain et al., 2012; reviewed in Yang and Seto, 2008; Geng et al., 2012). However, acetylation does not seem to act by directly preventing the poly-ubiquitination of the three HSF2 lysine residues, K128, K135, and K197. Indeed, only combined mutations of these lysines to glutamines (3KQ), but not to arginines (3KR), prevent HSF2 proteasomal degradation. In addition, 3KQ mutation decreases HSF2 polyubiquitination, whereas 3KR does not (Figure 5). Previous proteome-wide quantitative analyses of the ubiquitin-modified protein have revealed that theubiquitination of HSF2 occurs on multiple residues spanning over the HSF2 protein, including K51, K151, K210 and K420, in addition to K128, K135, and K197. Most of these sites reside in the HR-A/B domain or its vicinity, suggesting a crosstalk between acetylation and ubiquitination (Kim et al., 2011; Wagner et al., 2011; Akimov et al., 2018; www.phosphosite.org). It is important to note that the published studies have not assessed the functional impact of these ubiquitination events on HSF2 turnover (reviewed by Gomez-Pastor et al., 2018). Moreover, we showed that mimicking the acetylation of the lysine residues K128, K135, and K198 limits the degradation of HSF2, also under heat shock conditions. In parallel, we identify HDAC1 as a major lysine deacetylase involved in HSF2 deacetylation and the control of HSF2 proteasomal degradation, both under basal and stress conditions, which may provide an explanation for the rapid degradation of HSF2 by APC/C in response to heat shock (Ahlskog et al., 2010).
Finally, future studies are warranted to determine whether other enzymes (HATs/KATs or HDACs) regulate HSF2 acetylation, deacetylation and thereby stability, and modify other lysine residues that we found acetylated in our MS analysis, as it is the case for HSF1 (Westerheide et al., 2009; Zelin et al., 2012; Raychaudhuri et al., 2014; reviewed by Miozzo et al., 2015).

HSF2 destabilization and impaired heat shock response in RSTS syndrome
The accelerated turnover of the HSF2 protein in primary cells derived from the RSTS patients mutated either for *CBP* or *EP300* is counteracted by proteasome inhibition. This finding confirms the functional importance of HSF2 interaction with the two KAT3 family members, CBP and EP300, and provides further evidence for the role of acetylation on HSF2 stability. RSTS is a rare genetic, autosomal dominant neurodevelopmental disorder, characterized by intellectual disability, heart and skeleton malformations, and elevated susceptibility to infections as well as childhood cancers (Spena et al., 2015). Of the clinically diagnosed RSTS cases, the two identified genes mutated or deleted represent 60% for *CBP* and 8-10% for *EP300*. One mutated allele of *CBP* or *EP300* is sufficient to

provoke this severely disabling disease. This is surprising given that, not only a wild-type copy of the affected gene, but also the two wild-type alleles encoding the other closely related KAT3 are present in the patients. The causal mutations in RSTS patients are extremely diverse and can affect protein expression, protein-protein binding or catalytic activity, which results in loss of specificity for interacting partners and/or substrates. The mutated *CBP (or EP300)* allele *via* a rupture of this subtle equilibrium could thereby exerts a "dominant-negative" effect and compromise compensation by the wild-type copies of the other KAT3 (Merk et al., 2018; Lopez-Atalaya et al., 2014). Accordingly, the presence of either a catalytically inactive *CBP* allele or an *EP300* allele, deleted for the KIX domain, is sufficient to influence the proteasomal turnover of HSF2 in RSTS cell system. These results suggest that the presence of the dominant-negative mutated allele impairs HSF2 acetylation. In addition, VPA was unable to restore the HSF2 levels in RSTS cells, likely because HSF2 is not acetylated in these cells, which reinforces the hypothesis that the integrity of EP300 and CBP function is critical for the control of HSF2 levels in cells derived from RSTS patients.

HSF2 is known to modulate the intensity of the heat shock response (HSR) by fine-tuning the expression of *HSP* genes and the formation of nSBs (Östling et al., 2007; Sandqvist et al., 2009). This occurs through the concomitant binding of HSF2 and HSF1, which involves the formation of heterotrimers, to the regulatory regions of the *HSP* genes and *SatIII* loci (Alastalo et al., 2003; Östling et al., 2007; Sandqvist et al., 2009). In the context of the RSTS model, HSF2 seems to be profoundly deregulated, whereas HSF1 is only slightly affected, and we find that the basal levels of HSP protein are reduced, including HSP70. In response to heat shock, the magnitude of HSP70 induction is clearly impaired, indicating that HSF2 is an important regulator of the HSR in the hPSFs cells, which is in line with our earlier results (Östling et al. 2007). Moreover, we also observe that the reduced HSF2 levels are associated to a decrease in the formation of nSBs in RSTS cells. It is therefore possible that the HSR is regulated at two different levels by the proteasome, providing an exquisite and sophisticated way to control cell proteostasis, *via* the stabilization of: 1) HSF1 in an EP300-dependent manner (at least in some cell systems; Raychaudhuri et al., 2014); and/or 2) HSF2, in a CBP- and EP300-dependent manner, as shown here in hPSFs. The delicate balance between these two arms of regulation, *i.e.* driven by HSF1 or HSF2, could be tipped depending on the cellular context, and potentially by the HSF1/HSF2 ratio (Östling et al., 2007; Sandqvist et al., 2009; Elsing et al., 2014). In conclusion, the repertoire of chaperones is altered in a chronic manner in RSTS cells already under physiological (unstressed) conditions, more extensively upon exposure to stress, and this alteration might confer the diseased cells vulnerability to proteostasis challenges.

The dysregulation of the HSF pathway in RSTS, including the markedly reduced HSF2 levels, might have several implications for this multifaceted disease. Indeed, RSTS is a neurodevelopmental disorder and HSF2, as a transcription factor involved in neurodevelopment both in normal and stress conditions, might contribute to the neurodevelopmental defects characteristic for RSTS. Strikingly, RSTS patients suffer from extreme vulnerability to airway infections, which is mainly due to defects in mounting a response to polysaccharides (Naimi et al., 2006; Herriot et al., 2016). Because the HSF pathway is involved in response to polysaccharides, inflammatory and immune responses, as well as lung protection against stress, its deregulation could also contribute to this aspect of the pathology (Xiao et al., 1999; Inouye et al., 2007; Wirth et al., 2003). Therefore, any imbalance in the delicate composition of the HSPs and other molecular chaperones under normal conditions and the triggering of HSF-driven stress-responses could profoundly influence this vulnerability to multiple severe health problems associated with the complex RSTS syndrome.
Rare diseases are in the center of growing interest based on the recent acceptance that they represent a global public health and economic problem. For example RSTS, despite its rarity (1:100,000 births) represents one on 300 patients institutionalized for intellectual disability (Spena et al., 2015). Moreover, they are conceptually considered as extreme components in the spectrum of a large diversity of diseases, and their study has the strong potential to highlight shared features in related common disorders, and thus of being transposable to other pathologies and deeply

transform our ways to comprehend these pathologies. The functional impact of the regulation of HSF2 stability revealed in the context of RSTS might also represent an important reading key in related diseases, including neurodevelopmental disorders.

EXPERIMENTAL PROCEDURES

CONTACT FOR REAGENT AND RESOURCE SHARING

More detailed information and requests for resources and reagents should be directed to and will be fulfilled by the co-corresponding authors: Aurélie de THONEL (aurelie.dethonel@univ-paris-diderot.fr), Lea SISTONEN (lea.sistonen@btk.fi), and Valérie MEZGER (valerie.mezger@univ-apris-diderot.fr).

Reagents and treatments

Proteasome inhibitor MG132 was used for 6 h (N2A, U20C_2KO) at a final concentration of 20µM. HDACs inhibitor VPA (Interchim, AYJ060) was used at 1 mM for 3 h. The HAT inhibitor C646 (Sigma-Aldrich; SML0002) was used at a final concentration of 20 or 40 µM for 4 h. For all chemicals, DMSO was used as vehicle (control).

Heat shock treatments were performed in water bath at 42 or 43°C for the indicated times.

Table for antibodies

Antibodies	species	Clone	reference	Manufacturer	WB	IP	ImmunoF	IHC
Acetyl-lysine (Pan)	rabbit	pAb	#9441	Cell signalling Technology	1/1000		1/1000	
Actin	mouse	AC40	A3853	Sigma-Aldrich	1/4000			
Alexa Fluor 488	Donkey anti-mouse		715-546-151	J.ImmunoRes			1/800	1/800
Alexa Fluor 488	goat anti-rabbit		A-11008	J.ImmunoRes			1/800	1/800
Alexa Fluor 594	goat anti-rabbit		A-11037	J.ImmunoRes			1/800	1/800
CBP	rabbit IgG	D6C5	#7389	Cell signalling Technology	1/1000		1/100	
CBP	rabbit	A-22	sc-369	Santa-Cruz	1/1000			1/100
Cy3TM-3	Donkey anti-mouse		715-165-150	J.ImmunoRes			1/800	1/800
EP300	rabbit	pAb	sc-584	Santa-Cruz	1/500		1/25	1/25
Flag tag	mouse	M2	F1804	Sigma-Aldrich	1/1000	2µg		
GFP tag	mouse	IgG1	MAB2510	Millipore	1/1000			
GFP-Trap-A	mouse		gta-20	chromotek		25µl		
GST tag	mouse	IgG	AE001	Ab Clonal	1/2000	2,5µl		
H3K18Ac	rabbit	pAb	GTX128943-S	Euromedex				
H3K27Ac	rabbit	pAb	pAb174050	Diagenode	1/1000			
HA tag	mouse	16B12	MMS101R	Covance	1/2000		1/2000	
HDAC1	rabbit	pAb	7028	Abcam	1/4000		1/900	
HRP mouse	Goat anti-mouse		115 035 135	J.ImmunoRes	1/50 000			
HRP mouse Fab	Goat anti-mouse	F(ab')2	115 036 072	J.ImmunoRes	1/50 000			
HRP rabbit	mouse anti-rabbit	IgG1	211 032 171	J.ImmunoRes	1/50 000			
HSC70	rat	mAb	ADI-SPA-815	Stressgen	1/1000			
HSF1	rabbit	pAb	#4356	Cell signalling Technology	1/1000		1/800	
HSF2	mouse	G11	sc-74529	Santa-Cruz	1/250	4µg		
HSF2	mouse	3E2	ab69621	abcam		4µg		
HSF2	rabbit	pAb	H57	Lea Sistonen Lab			1/600	1/600
HSP70	mouse		ADI-SPA-810	Stressgen	1/1000			
HSP90	mouse	H9010	SMC-107	Stressmarq	1/3000			
IgG	mouse		I5381	sigma-Aldrich				
Myc tag	mouse	9B11	#2276	Cell signalling Technology	1/1000			
Myc-Trap-A	mouse		yta-20	chromotek		25µl		
Snap	rabbit	pAb	P9310	NEB	1/1000			
Sox 2	rabbit	pAb	ab97959	abcam				1/500
Trap-A CTL	mouse		bab-20	Chromotek		25µl		
Tuj-1	mouse	2G10	T8578	sigma-Aldrich				1/1000
Ubiquitin	mouse	FK2	BML-PW8810-0100	enzolifes sciences	1/500			

Plasmid Table and constructs

Plasmid name	Reference	vector	Origin	Supplier
Cas9 guide RNA HSF2	Cong et al., 2013	pX300 Cas9	human	
CBP DN-HA				
CBP-HA	kind gift of Pr. Wei Gu	pcDNA3-CMV	mouse	
EP300-HA	kind gift from W Sellers (East Tennessee State University, USA); Duval et al., 2015	1246 pCMVb MycHA	human	Addgene # U10718
GFP binder nanobody (GBP)	Kind gift from P.A. Defossez (University Paris-Diderot, France); Zolghadr et al., 2008			
GFP tag		pEGFP-N1		Clontech (6085-1)
HDAC1 DN$_{(D181A)}$-Flag	Kuzmochka et al., 2014	pcDNA3.1	human	
HDAC1-FLAG	Emiliani et al., 1998	pcDNA3.1	human	
HDAC1-GFP	kind gift from J Steve	eGFP-C3	human	
HDAC2-MYC	kind gift of Tony Kouzarides lab	pcDNA3.1/myc-HisA	mouse	evex 305
HDAC3-MYC	kind gift of Tony Kouzarides lab	pCMV3-Amyc	human	evex 476
His3.3-Snap	kind gift of Dr. S. Polo. (University Paris-Diderot, France)	pSNAPf	human	NEB (N9183)
HSF2alpha WT and mutant-Snap	See Materials and Methods	pSNAPf	human	NEB (N9181)
HSF2alpha-CTAP(GS)-Gw	Bürckstümmer et al., 2006	PCEMM-CTAP	mouse	Euroscarf
HSF2alpha-Myc	Alastalo et al., 2003	pcDNA4™/TO/myc-His-A	human	Life Technologie
HSF2beta-CTAP(GS)-Gw	Bürckstümmer et al., 2006	PCEMM-CTAP	mouse	Euroscarf
HSF2beta-Flag	Pirkkala et al., 2000	pFLAG-CMV-2	mouse	Sigma-Aldrich
HSF2beta-Flag deletion mutants	Alastalo et al., 2003	pFLAG-CMV-2	mouse	Sigma-Aldrich
HSF2beta-YFP	cf materials and methods	pEYFP-C1	mouse	Clontech
Kix WT-GST	kind gift from Dr. Lemassson; Yan et al., 1998	pGEX-2T	mouse	
Kix YAY-GST	kind gift from Dr. Lemassson; Cook et al., 2011	pGEX-2T	mouse	
Myc tag		pcDNA3.1 MycHis		Life technology (V80020)

The human HSF2-Snap (WT/mutants) were constructed from the HSF2-Myc (WT/mutants) plasmid after digestion of the inserts by EcoRI and KpnI and cloning into the EcoRI and EcoRV sites in frame with the C-terminal Flag tag in pSNAPf plasmid using In-Fusion Kit (Clontech). The human HSF2-YFP was constructed by PCR and cloned into the XhoI and SalI sites in frame with the N-terminal YFP tag in EGFp-C1 plasmid using In-Fusion Kit (Clontech). All PCR-amplified products for both plasmids were sequenced to exclude the possibility of second site mutagenesis. The cDNA coding for the acetyltransferase domain of murine CBP (1097-1774) was a kind gift of Pr. Ricardo Dalla-Favera (Columbia University, New York) and was used to generate cDNA coding for key domains of CBP: Full HAT (1096-1700), HAT (1322-1700), RING (1205-1279), PHD (1280-1321), Bromodomain (1096-1205), later sub-cloned in pet28a plasmid (Invitrogen) in order to produce 6 His-tagged proteins.

Patient material

Informed consent for skin biopsy and culture of hPSFs were obtained from the RSTS patients'parents (Patients 1 and 2, [P1] and [P2]; CHU, Robert Debré Hospital, Paris, France; Drs S. Passemard and A. Verloes), in accordance with the Declaration of Helsinki and approved by the local ethics committee, CPP Ile de France, agreement n° P10012 8. Human PSFs from healthy donors have been described in Yehezkel et al., 2008). hPSFs were grown in HAM's F10 supplemented with 12% FBS in humidified atmosphere with 5 % CO2 at 37°C. See Figure S7A for a description of the deletion or mutation carried by the patients.

Cell lines

Cell culture, transfections and treatments: murine Neuro2A (N2A, neuroblastoma, DSMZ # ACC 148), Hamster BHK (kindly provided by Dr.Leonhardt H and cultured as described (Herce et *al.*, 2013), human HEK 293T (ATCC®, CRL-11268™), U2OS (osteosarcoma, ATCC®, HTB-96™), U2OS-Crispr*HSF2*KO (2KO), SH-SY5Y (neuroblastoma, ATCC® CRL-2266™) and HeLa-S3 cells (kindly provided by Dr. Slimane Ait-Si-Ali and cultured as described (Yahi *et al.*, 2008)) were grown in DMEM (Lonza Group Ltd.) supplemented with 4,5 g/L glucose and 10% fetal bovine serum (FBS, Life technology) in humidified atmosphere with 5 % CO2 at 37°C. RSTS lymphoblastoid cells were obtained from Dr. D Lacombe (CHU Bordeaux; patient 3 (P3)) and from CRB-Institut Médical Jérôme Lejeune, CRB-BioJeL (BB-0033-

00016; May 05, 2018, Paris; patients 4 and 5 (P4 and P5)). Lymphoblastoid cells were grown in RPMI (Life technology) supplemented with 4,5 g/L glucose and 10% FBS with L-glutamine (Life technology) in humidified atmosphere with 5 % CO2 at 37°C. See Figure S7A for a description of the deletion or mutation carried by the lymphoblastoid cells. All cell lines were were tested to be mycoplasma free using Venor™ GeM Mycoplasma Detection Kit, PCR-based (Sigma-Aldrich).

Mice model

Specific pathogen-free C57BL/N female mice were purchased from Janvier (Lyon, France) and maintained in sterile housing in accordance with the guidelines of the Ministère de la Recherche et de la Technologie (Paris, France). Rodent laboratory food and water were provided ad libitum. Experiments were performed in accordance with French and European guidelines for the care and use of laboratory animals. The project has been approved by the Animal Experimentation Ethical Committee Buffon (CEEA-40) and recorded under the following reference by the Ministère de l'Enseignement Supérieur et de la Recherche (#2016040414515579).

Patients' primary fibroblasts

For fibroblasts primary cells, The study was conducted in accordance with the Declaration of Helsinki with an approved written consent form for each patient (CPP ESTI: 2014/39 ; N°ID: 2014-A00968-39), and approval was obtained from the local ethics committee ESTI (license: NCT02873832).

Generation of CRISPR/Cas9 *Hsf2*KO U2OS cells

Guide RNAs (gRNA) targeting the exon 1 of *HSF2* were designed using our own software (http://crispor.tefor.net/) and cloned into pMLM3636 guide RNA expression plasmid (Addgene 43860). U2OS cells were transfected with Cas9 and guide RNA expression plasmids using Amaxa electroporation as recommended by the manufacturer (Lonza). Cas9 expression plasmid was from the Church lab (Addgene 41815). One week after transfections, cells were seeded at single cell density. Clones were genotyped by DNA sequencing of PCR products spanning the targeted region of the *HSF2* gene. The selected U2OS clone presented 3 different outframe mutations on *HSF2* exon 1, each corresponding to a different allele (Figure S5B). Guide RNA sequence targeting the 1st AUG on *HSF2* exon 1: 5´-UGCGCCGCGUUAACAAUGAA-3´. Primers used for PCR cloning for validation: forward (hHSF2_Cr_ATG_F): 5´-AGTCGGCTCCTGGGATTG-3´ and reverse (hHSF2_Cr_ATG_R): 5´-AGTGAGGAGGCGGTTATTCAG-3´. The genomic sequences of the mutated h*HSF2* alleles and the resulting lack in HSF2 protein and HSF2 DNA-binding activity are described in Figure S5B.

Production of KIX-GST, His-CBP domain and HSF2 proteins

Escherichia coli 21 (DE3) were transformed with the different 6His-tag CBP constructs for production of the different CBP domains as previously described (Duval et al, 2015). All proteins were stored in 20 mM Tris-HCl, 150 mM NaCl, pH 7.5 and kept at -80°C until use. *E. coli* BL21 bacteria were transformed with the different GST-KIX constructs (Cook et al., 2011) and grown in presence of ampicilline and chloramphenicol at 37°C (4-6h). Bacteria were then grown with 1 mM Isopropyl β-D-1-thiogalactopyranoside over-night. After centrifugation (4,400 rpm at 4°C), bacteria were lysed in PBS pH 8, 300 mM NaCl, Triton X100 1%, 1 mG/mL lysozyme, protease inhibitors under stirring at 4°C for 30 min. Bacteria were sonicated (BRANSON sonicator, power 20%, 10'' ON/20''OFF) and centrifugated at 4°C, 16000g for 30 min. Gluthatione sepharose 4B beads (G4510-10ML Sigma-Aldrich) were added to the cleared supernatants and subjected to rotation for 1 h 30 min at 4°C. The mixture was loaded into a column (Sigma-Aldrich) and washed with PBS/NaCl 300 mM pH8, then Tris 50 mM/NaCl 150 mM pH 8. Proteins were eluted with 5mL of elution buffer (50 mM Tris HCl pH8, 150 mM NaCl, 10 mM GSH). The protein concentration was measured using Bradford method. *In vitro* transcription and translation reactions were performed using a TNT T7 coupled reticulocyte lysate system as recommended by the supplier (Promega, Charbonnieres-les-Bains, France). 1 µG of the plasmid DNA template was transcribed and the protein was translated at 30°C for 90 min.

Immunoprecipitation and Western blotting

Protein extracts from cells were prepared using a modified Laemmli buffer (5% sodium dodecyl sulphate, 10% glycerol, 32.9 mM Tris-HCl pH 6.8) supplemented with protease inhibitors (Sigma-Aldrich). Brain tissues were prepared with a lysis buffer (Hepes 10 mM ph 7.9; NaCl 0.4 M, EGTA 0.1 M; glycerol 5%, dithiothreitol (DTT) 1 mM, PMSF 1 mM, protease inhibitor (Sigma-Aldrich), phosphatase inhibitor (Roche)). Then, 30 µg of proteins from lysates were subjected to migration on 8–12 % acrylamide gels and transferred on to polyvinylidene difluoride membranes (GE Healthcare Europe GmbH) in borate buffer (50 mM Tris-HCl and 50 mM borate) for 1 h 45 at constant voltage (48 V). The membranes were incubated with primary antibodies overnight at 4°C, then washed in Tris-buffered saline–Tween 0.1% and incubated for 1 h with horseradish peroxidase (HRP)-coupled secondary antibody (Jackson Immunoresearch). The signal was revealed using a chemiluminescent reagent (Pierce® ECL Plus Western Blotting Substrate, Thermo Scientific) and was detected using hyperfilm (HyperfilmTM ECL, Amersham Biosciences) and a film processor (Konica Minolta). Poly-ubiquitinated HSF2 was detected as described in Ahlskog et al., 2010.

For immunoprecipitation of exogenous proteins, using GFP/Myc-Trap. GFP-Trap®-A (ChromoTek) contains a small recombinant fragment of alpaca anti-GFP-antibody, covalently coupled to the surface of agarose beads. It enables purification of any protein of interest fused to GFP, eGFP, YFP, CFP or Venus. HEK 293 cells were transfected by a combination YFP- or Myc-tagged hHSF2 and HA-tagged EP300, CBP (WT or DN) or GFP-tagged HDAC1, or mock vector, with XtremGENE HP Reagent (Sigma-Aldrich) following manufacturer's instructions. Cells were lysed in Lysis buffer (50 mM Hepes pH 8, 100 mM NaCl, 5 mM EDTA, Triton X-100 0.5 %, Glycerol 10 %, VPA (1 mM), DTT 1 mM, PMSF 1 mM, proteases inhibitors, phosphatase inhibitors (Roche)) and then, HSF2 was immunoprecipitated using anti-GFP- or anti-Myc-trap antibody, or as a control Trap®-A control (ChromoTek). Immunoprecipitated proteins were run on a 8 % SDS-polyacrylamide gel, followed by an immunodetection of CBP or EP300 protein using anti-HA antibody. The amount of immunoprecipitated HSF2 was determined after reblot of the IP membrane with an anti-GFP or anti-Myc antibody. The amount of HSF2 and CBP or EP300 proteins, in the input samples, were detected with anti-GFP or Myc and anti-HA antibodies, respectively.

For immunoprecipitation of endogenous proteins. Brain cortices or organoids, or cells (SHSY5Y, N2a) were lysed 30min in Lysis buffer A (25 mM Hepes pH 8, 100 mM NaCl, 5 mM EDTA, Triton X-100 0,5%, 1 mM VPA, 1 mM PMSF, proteases inhibitors, phosphatase inhibitors (Roche)). After centrifugation (15 min, 12 000g) and preclearing, cell lysates were subjected to immunoprecipitation overnight using an anti-mouse HSF2 (Santa-Cruz) and a non relevant IgG (Sigma-aldrich) as a negative control that were pre-incubated 1h at RT with Protein G UltraLink Resin beads (53132, Pierce). Protein complexes were then washed 4 times in wash buffer (25 mM Tris HCl pH7.5, 150 mM NaCl, 1 mM EDTA, Triton X-100 0.1 % Glycerol 10 %, 1 mM VPA, 1 mM PMSF, proteases inhibitors, phosphatase inhibitors (Roche)), and suspended in 2x Laemmli buffer. After boiling, the immunoprecipitates were resolved in 8% SDS-PAGE and immunoblots were performed using an anti-rabbit pan acetyl-Lysine, anti-mouse HSF2 (Santa-Cruz), EP300 (Santa-Cruz) and CBP (CST). The amount of HSF2 and CBP or EP300 proteins in the input samples were detected with anti-mouse HSF2 and anti-rabbit CBP (CST) or EP300 (Santa-Cruz) antibodies.

Biolayer Interferometry

For *in vitro* protein-protein interaction experiments, we used biolayer interferometry technology (Octet Red, Forté-Bio, USA). Recombinant HSF2 (TP310751 Origen) was desalted (ZebaTM Spin Desalting Columns, 7K molecular-weight cutoff, 0.5 ml (1034–1164, Fisher Scientific, Germany)) and biotinylated at a molar ratio biotin/protein (3:1) for 30 min at room temperature (EZ-Link NHS-PEG4-Biotin (1189-1195, Fisher Scientific, Germany)). Excess Biotin was removed using ZebaTM Spin Desalting Columns. Biotinylated recombinant HSF2 was used as a ligand and immobilized at 100 nM on streptavidin biosensors after dilution in phosphate-buffered saline (PBS; 600s). Interactions with desalted analytes diluted in PBS at 100 nM (recombinant CBP domains 6His-tag Full HAT, Bromodomain (BD), RING), or HSP70 as a positive control (ADI-SPP-555, Enzo- Life Sciences)) were

analysed after association (600 s). All sensorgrams were corrected for baseline drift by subtracting a control sensor exposed to running buffer only. For Kd determination, each Kd was calculated with a 1:1 stoichiometry model using a global fit with Rmax unlinked by sensor (FortéBio, Data analysis software version 7.1.0.89).

Purification of HSF2 from HEK293 cells and mass spectrometry analysis of acetylated lysines

The protocole used is the same as for HSF1, as described by Westerheide et al., 2009. Briefly, HEK 293 cells were transfected with mouse HSF2-beta Flag with or without CMV-EP300, treated with trichostatin A or nicotinamide as indicated 18 h prior harvesting and lysed in RIPA buffer. HSF2-Flag was immunoprecipitated, using α-Flag M2 affinity gel beads (Sigma F2426), and eluted with Flag peptide. Purified mHSF2-Flag was separated by SDS-PAGE, excised from the gel, digested with trypsin, and subjected to tandem mass spectrometric analysis by a hybrid quadrupole time-of-flight instrument (QSTAR, Applied Biosystems, Foster City, CA) equipped with a nanoelectrospray source. MS/MS spectra were searched against the IPI mouse sequence database (68,222 entries; version 3.15) using Mascot (Matrix Science, Boston, MA; version 1.9.05) and X! Tandem (www.thegpm.org; version 2006.04.01.2) database 4 search algorithms. Mascot and X! Tandem were searched with a fragment and precursor ion mass tolerance of 0.3 Da assuming the digestion enzyme trypsin with the possibility of one missed cleavage. Carbamidomethylation of cysteine was included as a fixed modification whereas methionine oxidation, N-terminal protein and lysine acetylation were included as variable modifications in the database search. Peptide identifications were accepted at greater than 95.0% probability as determined by the Peptide Prophet algorithm (7) and validated by manual inspection of the MS/MS spectra, as shown in Table S2. Related to Figure 2C and S2A and Table S1.

Tandem Affinity Purification (TAP) and M/S identification of HSF2 partners in HeLa cells

We performed retroviral transduction to establish HeLa-S3 cell lines expressing double-tagged HSF2 proteins. *Hsf2*-alpha and *Hsf2*-beta cDNA, cloned from E16 mouse brain were inserted in vector PCEMM-CTAP (Euroscarf P30536; Bürckstümmer et al., 2006) with CMV-driven expression of insert and GFP used as an indirect reporter (IRES; Figure S6A). PCEMM-CTAP allows sequential immunoprecipitation of the tagged protein through two tags: protein G (able to bind IgG) and the streptavidin-binding peptide (GS-TAP; Figure S6A). These two tag-modules are separated by two clivage sites for the Tobacco Etch virus (TEV) protease. Retrovirus production (*Moloney murine leukemia* virus type) and cell transduction with CTAP-empty (no insert), CTAP- Hsf2alpha and CTAP-Hsf2beta were performed as described (Sandrin and Cosset, 2006). GFP-positive Hela-S3 clones were then isolated by clonal dilution, selected by FACS (INFLUX 500, BD BioSciences, IMAGOSEINE Platform, Jacques Monod Institute, Paris) and amplified. By FACS we could isolate four cell sub-populations according to the intensity of the GFP signal to identify a population of cells expressing the recombinant tagged protein at levels similar to that of the endogenous HSF2 protein. GFP-positive Hela-S3 clones stably expressing GS-HSF2alpha, GS-HSF2beta or empty vector, were grown in floating cultures (spinners) and 10 G of cells (3 x 10^9 cells) per cell line was collected as described (Fristch et al. 2010). Nuclear extracts were prepared as described (Fristch et al. 2010; except the DNase 1 treatment was omitted). Total nuclear extracts were incubated with IgG-agarose beads overnight (Sigma; A2909). Then beads were incubated twice with TEV enzyme (Invitrogen, ref. 12575023)) for 45 min transferred to Poly-prep columns (Bio-rad). Eluates were collected in TEV buffer. These eluates were then incubated with Dynabeads My-one streptavidin for 1 h, washed once, and eluted using 5 mM D-biotin. Proteins were concentrated using TCA/acetone precipitation and dissolved in Laemmli sample buffer. Samples from three independent experiments were sequentially sent to TAPLIN Biological Mass Spectrometry Facility (Harvard Medical School, Boston, MA, USA) for MS (LC/MS/MS) analysis as in Fristch et al. (2010).

Immunoprecipitation and MS analysis of the endogenous HSF2 protein from E17 mouse cortices

We also analyzed by MS the partners of the endogenous protein HSF2 from E17 fetal cortices. Brain cortices were mechanically lysed into 4.5 volumes of the following buffer: 10 mM Hepes (pH 7.9),

400 mM NaCl, 5 %, glycerol, 100 mM EGTA) and two cycles of freezing and thawing in liquid nitrogen, and centrifuged at 20,000 g for 30 min. Supernatants were used to immunoprecipitate HSF2 using a monoclonal antibody (Abcam) or PBS as a negative control, and protein G agarose (Roche). Immunoprecipitates were analyzed on SDS-PAGE gel bands, containing the HSF2 protein (staining with colloidal blue) were sent to TAPLIN Biological Mass Spectrometry Facility (Harvard Medical School, Boston, MA, USA) for MS (LC/MS/MS) analysis.

Immunohistochemistry and immunofluorescence

Human organoids sections were prepared as described in Lancaster et al. (2014). Antigen retrieval was performed using citrate buffer (0.1 M Tri-sodium citrate pH 6.0, 10% glycerol, Tween 0,05%) for 1 h at 70°C. Slices were saturated for 30 min with 3% horse serum in PBS -Triton 0,1% and incubated with primary antibody overnight at 4°C. After washing in PBS-Tween 0,1%, slices were incubated with corresponding secondary antibody and DAPI (1 µg/ml) for 1 h at room temperature. Imaging was performed with a confocal microscope Leica TCS SP5 (IMAGOSEINE Imaging Platform in Institut Jacques Monod) and processed with Fiji software.

Cells in basal or heat shock conditions were fixed in 4 % paraformaldehyde on coverslip and stained with HSF1 (CST), HSF2 (Santa-Cruz) or EP300 (Santa-Cruz) antibodies followed by a staining with the corresponding mouse or rabbit fluorescent secondary antibody (Jackson Immunoresearch). Microscopy images were taken on an inverted microscope Leica DMI 6000 and images were analyzed using Fiji software.

SNAP-Tag labelling of HSF2 molecules and analysis of protein decay

CRISPR/Cas9 *Hsf2*KO U2OS cells were transfected with SNAP-HSF2 WT, -HSF2 3KQ or -HSF2 3KR constructs (Xtrem-Gen HP, Sigma-Aldrich), incubated in the presence of the cell-permeable SNAP-Cell® Oregon green fluorescent substrate (1,25 mM) and then with SNAP-Cell® Block (0,5mM) during the pulse chase according to the manufacturer's instructions (New England Biolabs). Cells were lysed in modified Laemmli buffer (5% SDS, 10% glycerol, 32.9 mM Tris-HCl pH 6.8) supplemented with 1 mM DTT (Sigma-Aldrich), and their extracts (15 µG) were run on 10% SDS-PAGE. Gels were then scanned on a Typhoon Trio imager (GE Healthcare; excitation 532 nm, emission 580 nm, PMT 700 V) for determination of signal intensity of the covalently-bound fluorescent products as described in (Sanial et al., 2017).

Fluorescence three-hybrid assay

Fluorescence three-hybrid assay was performed according to Herce *et al.* (2013). BHK cells were transfected with constructs expressing expressing YFP-HSF2, CBP-HA, or EP300-HA, and GBP-LacI, using different combinations (ratio 1:1.5:2) at 70-80 % confluency using reverse transfection by Lipofectamine 2000 (ThermoFisher Scientific), as indicated. Medium was changed after 4 hours for all transfection. After 24 h, the cells were fixed in 4 % paraformaldehyde on coverslip and stained with mouse anti-HA (Covance) or rabbit anti-CBP antibody (Santa-Cruz), followed by a staining with mouse or rabbit fluorescent secondary antibody (Jackson Immunoresearch), respectively. Confocal microscopy images were taken on a confocal microscope Leica TCS SP5 (IMAGOSEINE Imaging Platform in Institut Jacques Monod) and images were analyzed using Fiji software.

Modeling of the HR-A/B domain and KIX domain of CBP

Prediction of secondary structure of the HSF2 HR-A/B domain was performed using Psipred (http://bioinf.cs.ucl.ac.uk/psipred/) and nps@ (https://npsa-prabi.ibcp.fr/). The tertiary structure of the same domain was predicted using http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html. Sequence similarity between human HSF2 HR-A/B (Sandqvist et al., 2009), lipoprotein Lpp56 of *E. coli* (Shu et al., 2000), yeast transcriptional factor GCN4 (mutated on some residues to stabilize heptad repeats; Shu et al., 1999) and murine PTRF (Polymerase I and Transcript-Release Factor) and human ATF2 (Activating Transcription Factor 2; a member of the ATF/CREB family) that is known to interact with CBP/EP300 (Bordoli et al., 2001) was explored using *Uniprot* (Pundir et al., 2017*;*

https://www.uniprot.org/) and *ClustalW* (Thompson et al., 1994; https://www.ebi.ac.uk/Tools/msa/clustalw2/). Step 1: Based on this sequence similarity a sequence alignment of H-RA/B was developed (using *Uniprot and ClustalW*; see Figure S4E), a structural model of the monomer HRA/B was generated using using *Modeller* (v9.19 (Eswar et al., 2006; https://salilab.org/modeller/), verified by *ERRAT* (Colovos et al., 1993; http://servicesn.mbi.ucla.edu/ERRAT/) and *RESprox* (Berjanskii et al., 2012; http://www.resprox.ca/) and Ramachandran plot (see Figure S4F Ramachandran et al., 1963) followed by the development of the trimer using *SymmDock* (Schneidman-Duhovny *et al.*, 2005 ; v Beta 1.0; http://bioinfo3d.cs.tau.ac.il/SymmDock/). Step 2: the interaction between HR-A/B and the KIX domain (pdb: 2LXT; Brüschweiler et al., 2013) was simulated and compared using *Zdock* (v 2.1; Pierce et *al.*, 2014) and *Firedock* (Mashiach et al., 2008; Odoux et al., 2016). More precisely, the ten best results generated by *Zdock* and *Firedock*, and scored according to their Root Mean Square Deviation (RMSD), which were compared thanks to a visualization program ICM (Fernandez-Recio et al, 2005). The mutation of the key residuse involved in the interaction between HR-A/B and the KIX domain have been performed using PyMOL (v2.0) and the docking was done as described above.

RP-UFLC-based separation and quantification of CBP substrate peptides (HSF2) and their acetylated forms

For acetylation assays, we synthetized several 5-fluorescein amidite (5-FAM)-conjugated peptide substrates based on the human HSF2 sequence and containing various lysine residues of interest (Proteogenix):

- 5-FAM-SGIVK82QERD-NH2, referred to as K82 peptide
- 5-FAM-SSAQ135VQIR-NH2, referred to as K135 peptide
- 5-FAM-SLRRK197RPLL-NH2, referred to as K197 peptide

We also synthesized acetylated versions of these HSF2 peptides as standards. Samples containing HSF2 peptides and their acetylated forms were separated by RP-UFLC (Shimadzu) using Shim-pack XR-ODS column 2.0 x 100 mm 12 nm pores at 40°C. The mobile phase used for the separation consisted of 2 solvents: A was water with 0.12 % trifluoacetic acid (TFA) and B was acetonitrile with 0.12 % TFA. Separation was performed by an isocratic flow depending on the peptide:

- 80 % A/20 % B, rate of 1 ml/min for K82 and K135
- 77 % A/23 % B, rate of 1 ml/min for K197

HSF2 peptide (substrate) and their acetylated forms (products) were monitored by fluorescence emission (λ= 530 nm) after excitation at λ= 485 nm and quantified by integration of the peak absorbance area, employing a calibration curve established with various known concentrations of peptides.

***In vitro* acetyltransferase assay**

To determine the activity of recombinant CBP-Full HAT on HSF2 peptides, we used 96-wells ELISA plate (Thermofisher) and assays were performed in a total volume of 50 µL of acetyltransferase buffer (50 mM Tris pH8, 50 mM NaCl) with 500 nM CBP-Full HAT, 50 µM HSF2 peptides and 1 mM DTT. Reaction was then started with the addition of 100 µM Acetyl-CoA (AcCoA) and the mixture was incubated 20 min at room temperature. 50 µL of HClO4 (15 % in water, v/v) was used to stop the reaction and 10 µL of the mixture were injected into the RP-UFLC column for analysis. For time course studies, aliquots of the mother solution were collected at different time points and quenched with 50 µL of HClO4 prior to RP-UFLC analysis.

Statistics

Data are displayed as means ± standard deviation (SD). GraphPad Prism 8 (GraphPad Software, La Jolla, CA, USA) was used for statistical analyses. Statistical significance was assessed using Wilcoxon matched-pairs signed rank test for two groups with paired values (Figure 5-6,S6) or the Mann-

Whitney test for two groups (Figure 7,S7). p–values below 0.05 are considered statistically significant.

LIST OF SUPPLEMENTARY FIGURES

Supplemental Information includes 7 figures. Figure S1 to S7 are relative to Figure 1 to 7.

LIST OF SUPPLEMENTAL TABLES

Table S1. Summary of the HSF2 acetylated peptides identified by MS

Table S2. Original tables and spectra corresponding to the HSF2 acetylated peptides identified by MS

Table S3. Optimal docking area (ODA)/ table of docking scores. Related to Experimental procedures.

FIGURES

de Thonel, Ahlskog et al.

Figure 1

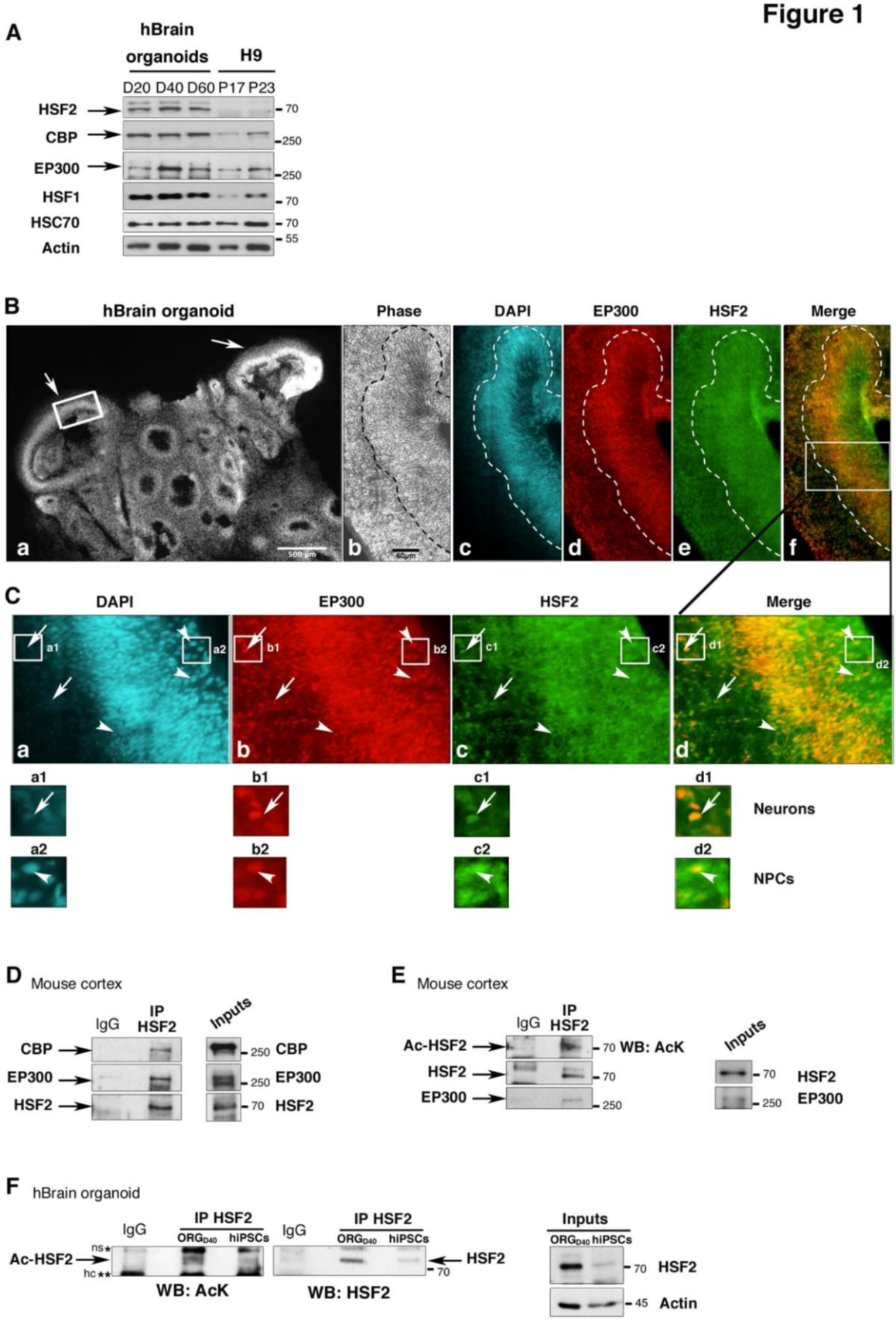


Figure 1. HSF2 expression profiles, acetylation status, and interaction with EP300/CBP in brain development

(A to C) ***HSF2 and CBP/EP300 in human brain organoids exhibit similar expression profiles and territories.***

(A) Representative immunoblot of extracts from human brain organoids at day 20, 40, and 60 of *ex vivo* development (D20, D40, D60) and human embryonic stem (ES) cells (H9) at passage 17 and 23. The position of molecular weight markers is indicated (kDa). HSC70 is heat shock cognate protein, which not induced by stress and is commonly used as a loading control. Actin, which increases during organoid differentiation, was also used for comparison. (See also Figure S1A).

(B) (a) Microscopy epifluorescence images of a D60 human organoid. (a) DAPI-staining of the complete section (image reconstruction), showing structures reminiscent of the developing cerebral cortex (arrows). The thick white rectangle indicates the magnified areas shown in (b-f). (b) Phase contrast; (c) DAPI-staining; (d-f) Immunostaining for EP300 (d, f, red) and HSF2 (e, f, green), and merge (f). The thin rectangle in (f) indicates the area magnified in (**C**). Scale bars: in (a), 500 µm; in (b), 60 µm.

(C) Magnification of the cortical-like area indicated by the thin rectangle in (**B**, f). (a) DAPI-staining; (b-d) Immunostaining for EP300 (b) and HSF2 detection (c), and merge (d). (a1-d1) and (a2-d2) correspond to magnified regions in the zone of neurons (low DAPI density, Tuj1 positive region, see Figure S1B;e,j) and NPCs (high DAPI density, Sox2-positive region, see Figure S1B;h,i), respectively, indicated by white squares in (a, b, c, d). HSF2 and EP300 are co-expressed in some neurons (long arrows;) and NPCs (arrowheads; dense DAPI-stained regions). (See also Figure S1B; d,e).

(D and E) ***HSF2 interacts with EP300 and CBP, and is present in an acetylated form in the developing cortex.***

(D) Endogenous HSF2 and EP300 proteins are co-immunoprecipitated in mouse E16 cortical extracts. (See also Figure S1E (E10 stage)). (Left panels) After immunoprecipitation of HSF2, the co-precipitated CBP or EP300 proteins were detected by Western blot analysis WB. (Right panels) total amounts proteins in the input samples. *: IgG heavy chain. Representative immunoblots (n=3 experiments).

(E) Acetylation of the immunoprecipitated HSF2 protein from E15 mouse cortical extracts was assessed using anti-pan-acetyl-lysine antibody (AcK; see also Figure S1E (E10 stage)). Co-immunoprecipitation of EP300 is shown as a positive control. Representative immunoblots (n= 2 experiments).

(F) HSF2 is acetylated in human brain organoids. (Left panel) Immunoprecipitation of the HSF2 protein in D40 organoids and immunoblotting with an anti-AcK antibody. (Middle panel) reincubation of the blot with anti-HSF2 antibody. (Right panel) total amounts proteins in the input samples. hIPSCs were used as comparison as they contain with low levels of HSF2. ns*, non specific; hc**, IgG heavy chain. Actin was used as a loading control.

de Thonel, Ahlskog et al.

Figure 2

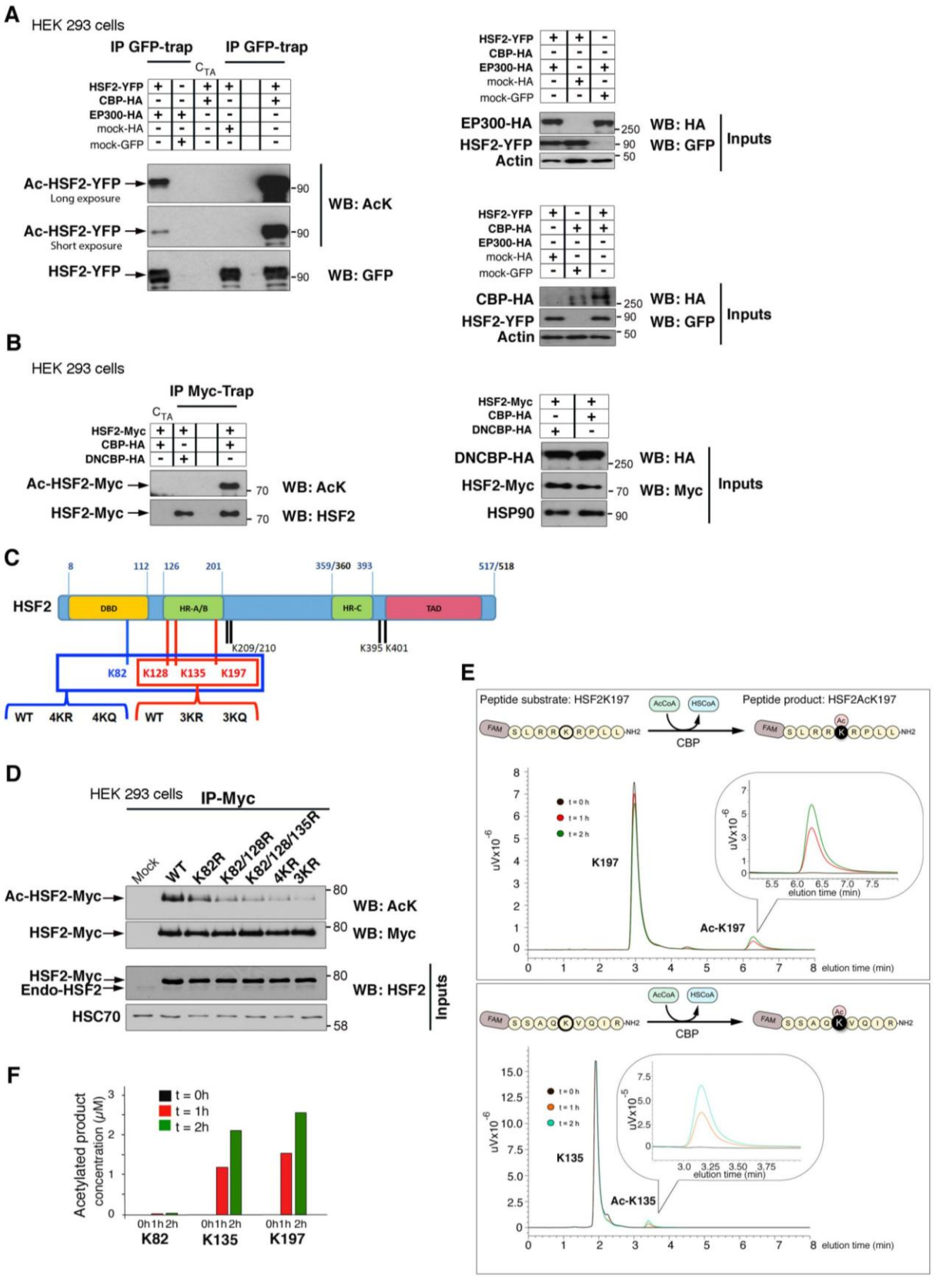


Figure 2. HSF2 is acetylated by CBP and EP300 in normal conditions.
(A) ***The ectopically expressed YFP-HSF2 protein is acetylated by exogenous HA-CBP or EP300.*** Representative immunoblots (n=5 independent experiments). HEK 293 cells were transfected with different combinations of YFP-HSF2, HA-CBP, HA-EP300 constructs, and mock-HA or -GFP constructs. YFP-HSF2 was immunoprecipitated using anti-GFP-trap antibody (IP GFP-Trap) and its acetylation status was determined by WB analyses, using an anti-pan-acetyl-lysine antibody (AcK; left panels). Immunoprecipitated HSF2 was detected using an anti-GFP antibody (WB: GFP). The total amounts of the proteins in the input samples were detected with anti-GFP or anti-HA antibodies (inputs; right panels). Actin, loading control. C_{TA}:Trap®-A beads were used as a negative control (see Experimental Procedures).

(B) The HSF2-Myc protein is not acetylated by a dominant-negative form of CBP (DNCBP-HA). HEK 293 cells were transfected as in (**A**), except that HSF2-Myc was used instead of HSF2-YFP, and immunoprecipitation was performed using anti-Myc-trap antibody (IP Myc-Trap). Representative immunoblots (n=2 experiments). C_{TA}: Trap®-A beads were used as a negative control (see Experimental Procedures). HSP90: loading control.

(C) Schematic representation of the eight main acetylated lysine residues of the HSF2 protein. Purified mouse Flag-HSF2, co-expressed with HA-EP300, immunoprecipitated and subjected to MS analysis for detection of acetylated lysine residues. The three lysine residues K128, K135, K197, located in the oligomerization domain (HR-A/B), are enlightened in red and K82, located in the DBD, in blue; the other four lysine residues (K209/K210, K395/K401) are indicated in regular black. The DNA-binding domain (DBD, orange); the oligomerization domain (HR-A/B; green) and the domain controlling oligomerization (the leucine-zipper-containing HR-C; green); as well as the N-terminal domain (activation domain TAD; red) are illustrated. The boundaries of each domain are indicated in bold and blue. Bold and blue numbers correspond to the number of the amino acids located at boundaries of the domains of the mouse HSF2 protein, numbered from the +1 (ATG); the equivalent in the human HSF2 protein, if different, are indicated in bold and black. These four (K82, K128, K135, K197) or three lysine residues (K128, K135, K197) were mutated into glutamines (4KQ or 3KQ, respectively) or arginines (4KR or 3KR, respectively; see also Figure S2A).

(D) The mutations of three or four lysine (K) residues to arginine (3KR or 4KR) or glutamine residues (3KQ or 4KQ) decrease global HSF2 acetylation levels. HEK 293 cells were co-transfected with EP300-HA and wild-type (WT) or mutated human HSF2-Myc on the indicated lysine residues. After immunoprecipitation of HSF2, using anti-Myc antibody, its acetylation was analysed by WB using an anti-AcK antibody. HSC70, loading control. n=3 independent experiments. (See also Figure S2B,C).

(E) ***In vitro*** **acetylation of HSF2 peptides containing the K135 or K197 residues by recombinant CBP Full-HAT**. Time course of reverse phase-ultra-fast liquid chromatography (RP-UFLC) analysis of the acetylation of HSF2K197 (upper panel) and HSF2K135 peptides (lower panel) by CBP Full-HAT. Aliquots of the reaction were collected at 0 (black), 1 (red) or 2 (green) hours and elution of peptides was monitored by fluorescence emission at 530 nm (excitation: 485 nm, uV: arbitrary unit of fluorescence; see Figure S2D for HSF2K82 peptide).

(F) Quantification of the ***in vitro*** **acetylated HSF2 peptides containing K82, K135, and K197 residues.** The AUC (area under the curve) of the acetylated K82, K135 and K197 peptides was quantified and converted in product concentration using a calibrated curve of various known concentrations of peptides. Note that it was not possible to investigate the acetylation of the HSF2 K128 peptide by CBP, because this peptide was repeatedly insoluble at the synthesis steps (Manufacturer's information; see also Figure S2D-F).

de Thonel, Ahlskog et al.

Figure 3

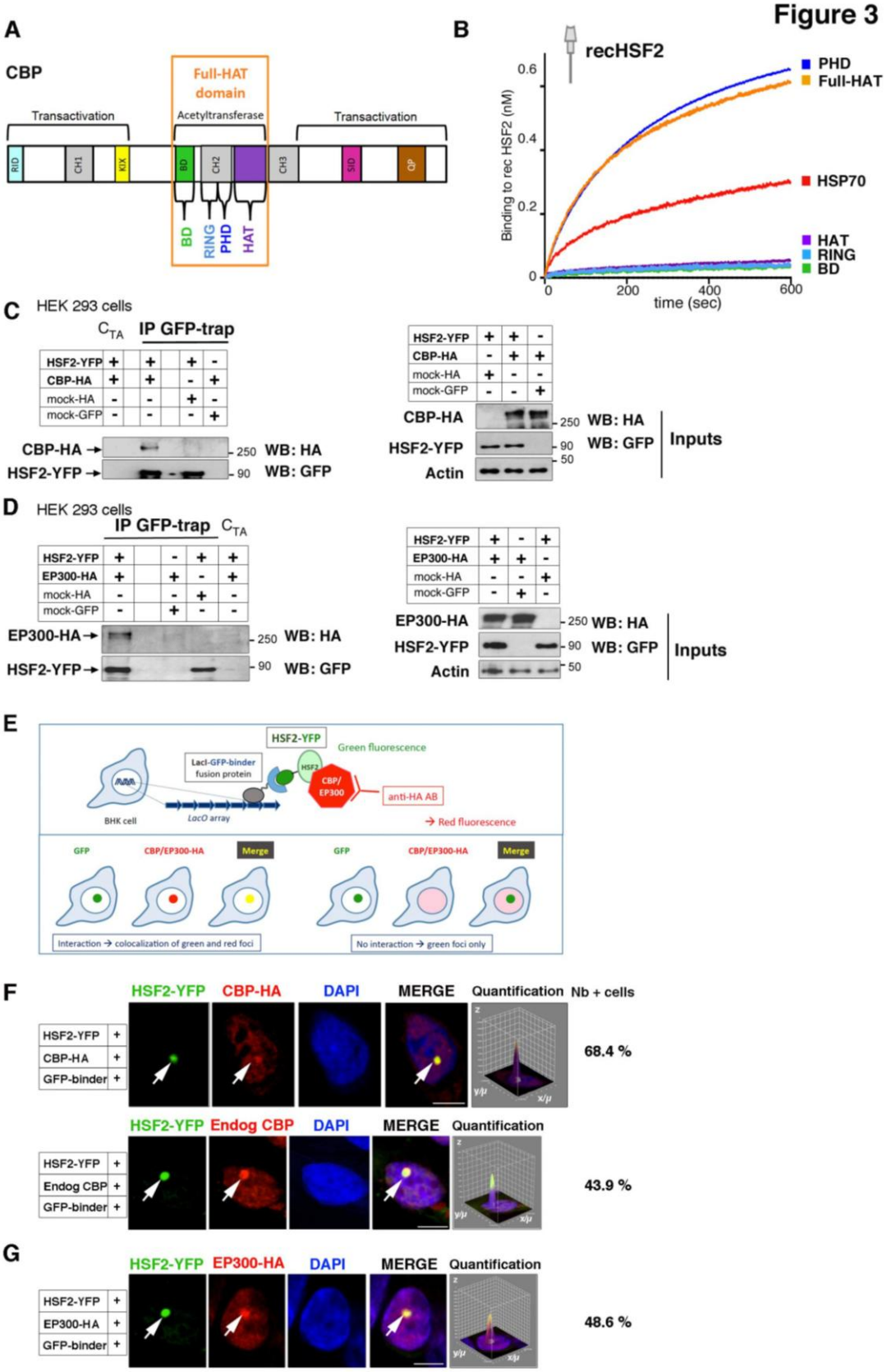


Figure 3. HSF2 interacts with CBP and EP300 in normal conditions.

(A) ***Schematic representation of CBP protein domains***. The ability of CBP to bind a very large number of proteins is mediated by several conserved protein binding domains, including the nuclear receptor interaction *domain* (RID), the cysteine/histidine-rich region 1 (CH1), KIX, Bromodomain (BD), PHD, CH2, HAT, CH3, SID (steroid receptor co-activator-1 interaction domain), and the nuclear coactivator binding domain NCBD (not illustrated here) and QP (Glutamine- and proline-rich domain; Dancy and Cole, 2015; Dyson and Wright 2016).

(B) ***Biolayer interferometry measurement of CBP binding to recombinant HSF2*****.**
Binding of different His-tagged domains of CBP to immobilized biotinylated recombinant HSF2 on streptavidin sensor tips (recHSF2): the catalytic full domain Full-HAT, or one of the following subdomains: PHD, HAT, RING, or Bromodomain (BD). HSP70 was used as a positive control for binding to HSF2 (Tang et al., 2016; see Figure S3A for determination of the Kd).

(C) ***The ectopically expressed YFP-HSF2 protein interacts with exogenous HA-CBP.*** HEK 293 cells were transfected with combinations HSF2-YFP and CBP-HA, or mock-HA (or mock-GFP) constructs. (Left panels) HSF2-YFP was immunoprecipitated using anti-GFP-trap antibody (IP GFP-Trap) and co-immunoprecipitated CBP protein was detected by using anti-HA antibody. The immunoprecipitated HSF2 was detected using an anti-GFP antibody. (Right panels) The total amounts of exogenous HSF2 and CBP proteins in the input samples were detected with anti-GFP and anti-HA antibodies, respectively (inputs). Actin, loading control. Representative immunoblots (n=3 experiments).

(D) ***The ectopically expressed YFP-HSF2 protein interacts with exogenous EP300.*** As in (**C**) except that HEK 293 cells were transfected with an EP300-HA construct. Representative immunoblots (n= 3 experiments).

(E) ***Principle of the fluorescent-3-hybrid (F3H) assay.***

(F) ***F3H assay for the visualization of interaction between HSF2-YFP and exogenous CBP-HA, or HSF2-YFP and endogenous CBP*** (Upper and lower panels, respectively). (Left panels) Confocal sections of BHK cells carrying a stably integrated Lac-operator array that were triple transfected with *LacI* fused to the GFP-binder, HSF2-YFP, and CBP-HA constructs. Exogenous and endogenous CBP was detected using an anti-HA or an anti-CBP antibody, respectively (red signal). Chromatin was counterstained using DAPI. All the experiments involving negative controls are shown in Figure S3C,D. White arrows point out localization of HSF2-YFP and CBP-HA at the *LacO* spot. Scale bar, 10 µm. (Right panels) Graphs represent the quantification of the intensity of the two fluorescence signals, visualizing the co-localization of HSF2-YFP and CBP-HA signals to the *LacO* array (x/µ; y/µ; z, signal intensity in arbitrary units). “Number (Nb) of + cells”: quantification of the percentage of cells showing co-recruitment of YFP-HSF2 and CBP-HA or endogenous CBP, in the *Lac0* array. Representative images, n=3 independent experiments for CBP-HA and endogenous CBP.

(G) ***F3H assay for the visualization of interaction between HSF2-YFP and exogenous EP300-HA*****.** As in Figure 3F, except that exogenous EP300-HA was detected using an anti-HA antibody (red signal). All the negative controls are shown in Figure S3E. Representative images, n=4 independent experiments. Scale Bar, 10 µM.

de Thonel, Ahlskog

Figure 4

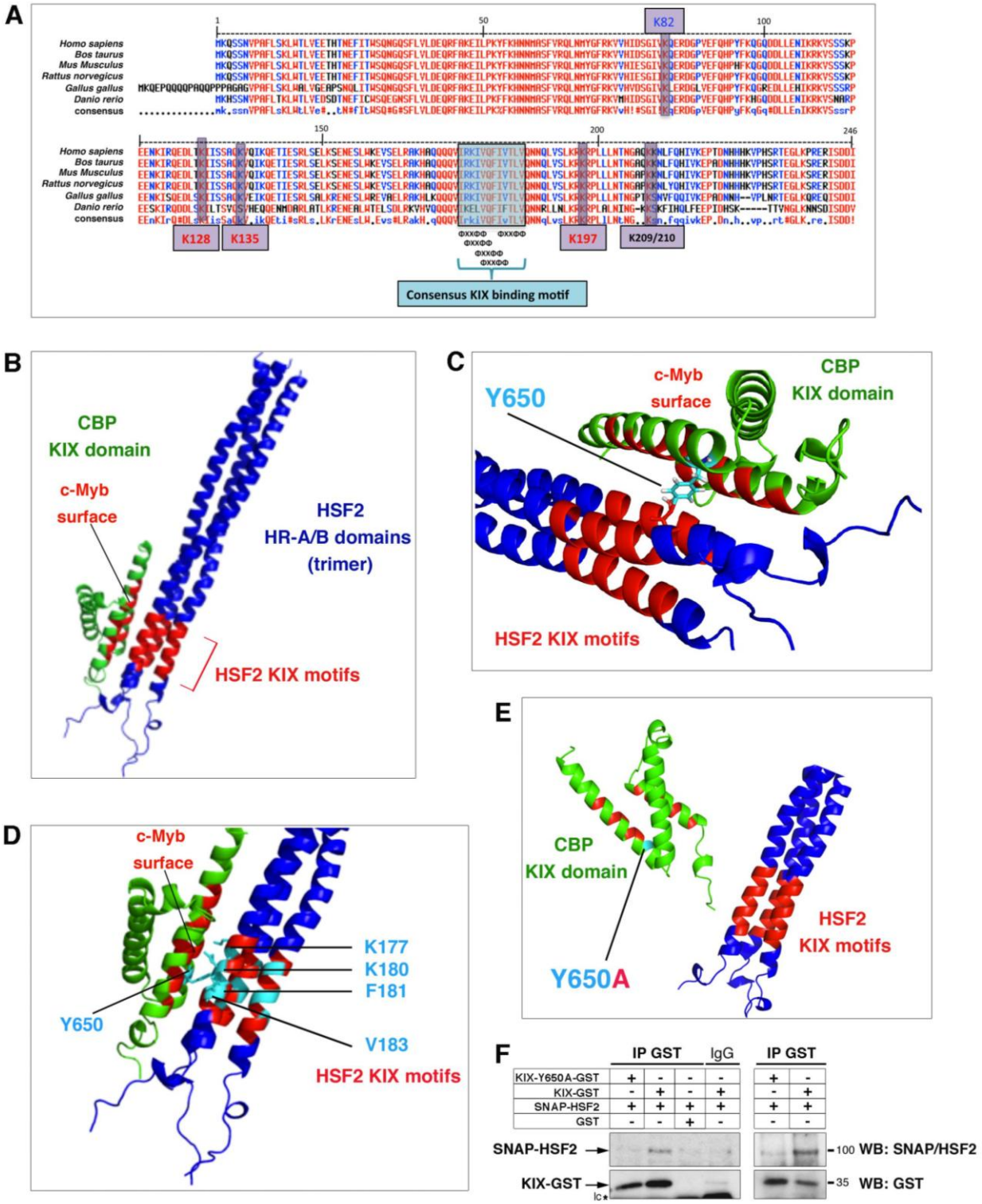


Figure 4. Modelling of CBP and HSF2 interaction.

(A) ***Schematic representation of the KIX-binding motifs located in the HSF2 HR-A/B region*****.**
Conserved KIX-binding motif sequences ("□XX□□") are indicated (blue rectangle). The positions of the very conserved, major acetylated lysine residues are highlighted (purple rectangles): K82 located in the DBD (in blue), K128, K135, and K197 located in the HR-A/B domain (in red), and K209/K210, located downstream the HR-A/B (in black).

(B) In silico ***model structure of CBP KIX domain and HSF2 HR-A/B interaction.***
Representation of the HSF2 HR-A/B domains of the HSF2 trimer (in blue), as a triple-coiled coil. The KIX recognition motifs of HSF2 are indicated in red. Representation of the KIX domain of CBP, a triple helical globular domain (in green). The c-Myb surface of the KIX-domain is indicated in red.

(C) ***Magnification of the*** **in silico** ***representation illustrated in (C)*** showing the contact of the tyrosine residue Y650, located within the c-Myb surface of the CBP KIX domain, to the KIX recognition motifs located within the HSF2 HR-A/B domain.

(D) In silico ***representation of the positioning of four residues located within the HSF2 KIX recognition motifs, and of Y650 within the CBP KIX domain that have been analyzed by*** **in silico** ***mutation*** (see (D-F) and Figure S4G).

(E) In silico ***Y650A mutation disrupts interaction between the HSF2 KIX motifs and the CBP KIX domain*** *Firedock* analysis. (see also Figure S4G(b) for *Zdock* analysis and Table S3).

(F) In vitro ***Y650A mutation disrupts interaction between the HSF2 KIX motifs and the CBP KIX domain*** (see also Figure S4G). The recombinant wild-type or mutated KIX-GST and SNAP-HSF2 proteins were produced in bacteria and reticulocyte lysates, respectively, and then subjected to an *in vitro* co-immunoprecipitation using an anti-GST antibody. The left and right panels correspond to two independent experiments. WB with anti-SNAP antibody (left upper panels) or with anti-HSF2 antibody (right upper panels).

de Thonel, Ahlskog et al.

Figure 5

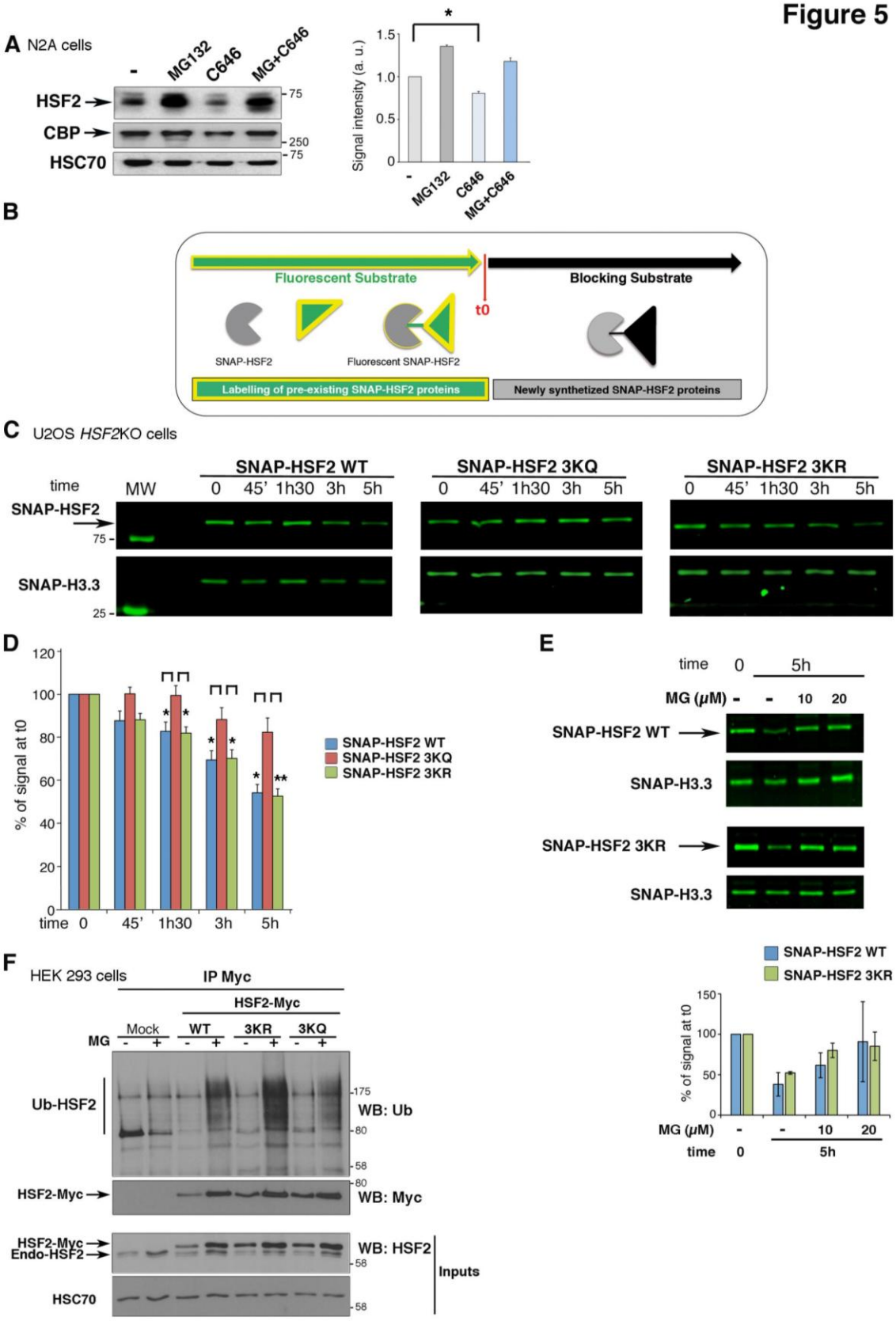


Figure 5. Impact of preventing or mimicking acetylation of lysine residues K128, K135, and K197 on HSF2 protein stability.

(A) ***Inhibiting CBP/EP300 decreases HSF2 protein levels, a process counteracted by proteasome inhibition.*** (Left panels) Representative immunoblot of HSF2 and CBP levels upon treatment of N2A cells with the CBP/EP300 inhibitor C646 (40µM for 4 h) and/or with the proteasome inhibitor MG132 (20 µM for 6h). n = 4 independent experiments. (Right panel) Graph corresponding to quantification of the HSF2 signal intensity relative to the vehicle-treated samples (-) and normalized to the loading control HSC70. Standard deviation is indicated. * $p<0.05$ (relative to non-normalized data).

(B) ***Schematic representation of the principle of SNAP-TAG-based pulse-chase experiments***. Cells expressing SNAP-tagged HSF2 are incubated in the presence of a fluorescent substrate, which, at a given time (t_0), covalently labels the pool of SNAP-HSF2 molecules present in the cell. The addition of a non-fluorescent, blocking substrate prevents further labelling of the newly synthesized HSF2 molecules. It allows the measurement of the decay in the fluorescent signal corresponding to SNAP-HSF2 molecules covalently bound by the fluorescent substrate, thereby allowing an estimation of the decay in HSF2 protein levels.

(C) ***Combined mutations of lysine residues K128, K135, and K197 mimicking HSF2 acetylation (3KQ) slow down the decay of HSF2 protein levels, compared with HSF2 WT***. Representative gel analysis of decay in HSF2 protein levels, carrying 3KR or 3KQ mutations, and labelled by the fluorescent SNAP-substrate, in CRISPR-Cas9 *Hsf2*KO U2OS cells. SNAP-H3.3 is used as a loading control.

(D) ***Quantification of the fluorescent signal*** corresponding to SNAP-HSF2 WT (blue), 3KQ (red), or 3KR (green), as a measure of SNAP-HSF2 protein decay, relative to the control samples (t0) and normalized to the loading control H3.3. n=7 independent experiments. Standard deviation. * $p<0.05$;** $p<0.01$.

(E) ***The decrease in SNAP-HSF2 WT and 3KR protein levels depends on proteasome activity***. (Upper panels) as in (**C**), but cells were pre-treated with 10 or 20 µM MG132 for 6 h. n=2 independent experiments. (Lower panels) quantification as in (D).

(F) ***The HSF2 3KR mutation, favours HSF2 poly-ubiquitination, whereas 3KQ does not*.** HEK 293 cells were co-transfected with Myc-HSF2 WT, 3KR, or 3KQ, and treated or not with the proteasome inhibitor, MG132 (MG; 20 µM for 6 h). HSF2 was immunoprecipitated using anti-Myc antibody and its poly-ubiquitination status was analysed by WB, using an anti-ubiquitin (Ub) antibody. The protein amounts in the input samples were detected with antibodies against HSF2. HSC70, loading control. (n=3 independent experiments).

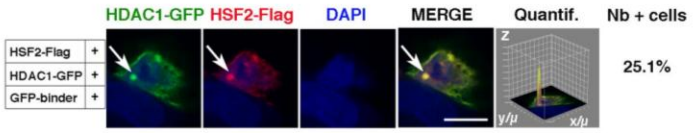
de Thonel, Ahlskog

Figure 6

A

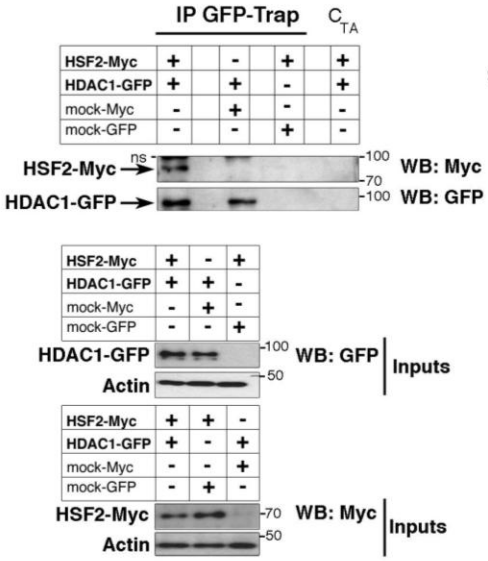
Sample	Number of peptides	Protein match	UniprotKB code
CTAP-HSF2a	1	HDAC1_HUMAN	Q13547
	2	HSF2_HUMAN	Q03933
	1	NUP62_HUMAN	P37198
CTAP-HSF2a	1	HDAC1_HUMAN	Q13547
	1	HSF2_HUMAN	Q03933

B



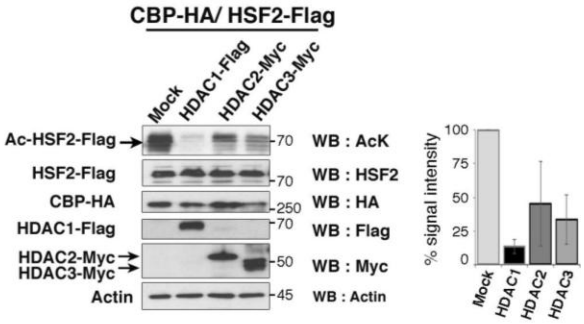
C

HEK 293 cells



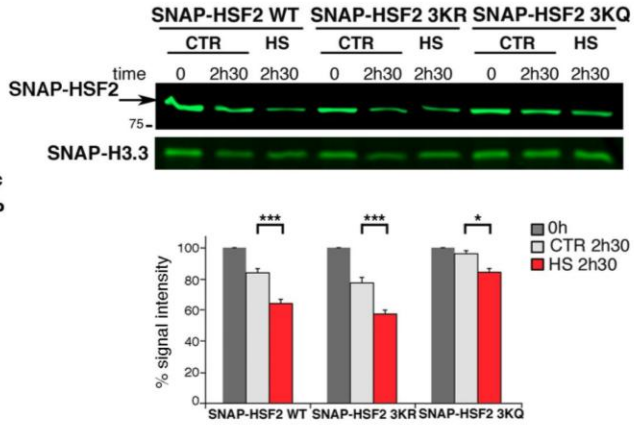
D

HEK 293 cells



E

U2OS 2KO cells



F

N2A

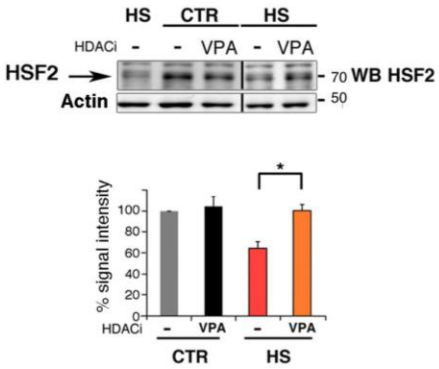


Figure 6. Impact of HDAC1 on HSF2 levels under non-stressed and stress conditions.

(A) ***Identification of HDAC1 as a HSF2 protein partner in TAP-TAG/MS analysis in HeLa-S3 cells***.
After sequential immunoprecipitation of nuclear extracts of HeLa-S3 expressing CTAP-HSF2α and CTAP-HSF2, using two tags (G-protein and Streptavidin-binding peptide; see Figure S6A), eluates were analyzed by MS. The number of unique peptides from each identified protein and their UniProt Knowledgebase (UniProtKB) codes are indicated.

(B) ***Interaction between ectopically expressed HDAC1-GFP and HSF2-Flag in F3H assays***. As in Figure 3F, except that BHK cells were triple transfected with *LacI* fused to the GFP-binder, HDAC1-GFP, and HSF2-Flag constructs. Chromatin was counterstained using DAPI. All the negative controls are shown in Figure S6E. n=3 independent experiments. Scale Bar, 10 µM.

(C) ***The ectopically expressed exogenous HDAC1 interacts with HSF2.*** (Upper panel) GFP-Trap co-immunoprecipitation of HDAC1-GFP and HSF2-Myc in transfected HEK 293 cell extracts. (Middle and lower panels) Immunoblot showing total HDAC1 (WB GFP) or HSF2 levels (WB Myc) in inputs, respectively. n=2 independent experiments. Actin was used as a loading control. ns: non-specific band.

(D) ***Overexpression of HDAC1 markedly reduces the acetylation of HSF2 by CBP*****.** HEK 293 cells were transfected by the following constructs: CBP-HA and HSF2-Flag, and HDAC1-Flag, HDAC2-Myc, or HDAC3-Myc and the acetylation status of the HSF2-Flag protein was checked by using an anti-AcK antibody. n=5 independent experiments. Actin was used as a loading control.

(E) ***The stability of HSF2 3KQ is increased upon HS, compared with HSF2 WT or HSF2 3KR***. (Upper panel) Representative gel analysis of HSF2 protein decay in a SNAP-TAG pulse-chase experiment. As in Figure 5C, except that cells were submitted to HS at 42°C for the indicated times. (Lower panel) Quantification of the fluorescent signal intensity, relative to the control samples (t0) and normalized to the loading control H3.3. n=7 independent experiments. Standard deviation. * $p<0.05$; *** $p<0.001$. SNAP-H3.3 was used as a loading control.

(F) ***Class I HDAC inhibitor VPA prevents the decrease in endogenous HSF2 protein levels induced by HS***. N2A cells were pretreated or not with 1 mM valproic acic (VPA) for 3 h and subjected to HS 42°C for 2h30. (Upper panel) Representative immunoblot. (Lower panel) Quantification on the signal intensity normalized to actin levels (n=4 independent experiments). * $p<0.05$.

de Thonel, Ahlskog et al.

Figure 7

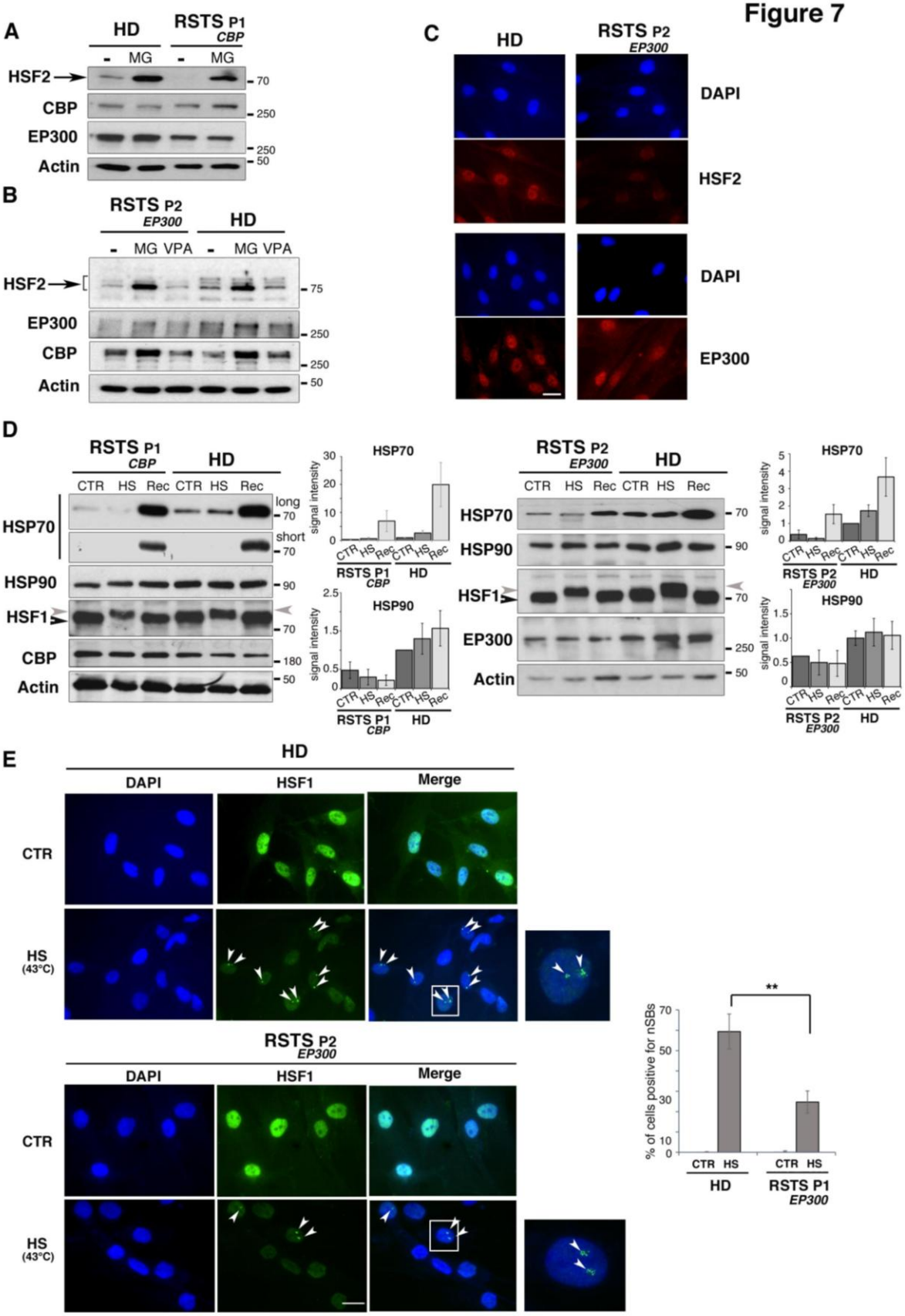


Figure 7. Altered HSF2 protein levels and dysregulated stress response in cells from RSTS patients.

(A) ***HSF2 levels are reduced in RSTS$_{CBP}$ hPSFs (patient P1), but restored in the presence of the proteasome inhibitor MG132***. See Figure S7A for the description of the patient's mutation. Representative immunoblots. Cells were treated with 20 µM MG132 (6 h) and subjected to immunoblot analysis. n=3 experiments. Actin was used as a loading control.

(B) ***HSF2 levels are reduced in RSTS$_{EP300}$ hPSFs (patient P2), but restored in the presence of the proteasome inhibitor MG132.*** See Figure S7A for the description of the patient's mutation. Cells were treated with 20µM MG132 (6 h) or 1mM of the HDAC inhibitor VPA (3 h) and subjected to immunoblot analysis. In contrast to MG132, the HDAC inhibitor VPA does not restores HSF2 levels. n=3 experiments.

(C) ***HSF2 staining is reduced in RSTS$_{EP300}$ hPSFs.*** Representative immunofluorescence experiments. n=3 experiments. Scale Bar, 10 µM.

(D) ***Reduced HSP basal levels and induction by heat shock in RSTS, compared to HD hPSFs.*** (Left panels) Representative immunoblots from ***RSTS$_{CBP}$*** [P1]. CTR, control conditions; HS, heat shock conditions (1 h at 42°C); Rec, recovery at 37°C for 2 h. Two exposure times for HSP70, long and short. Quantification of the HSP70 and HSP90 (HPS90AB1) signal intensity in immunoblots, normalized to actin. (Right panels) as in left panels, but from ***RSTS$_{EP300}$*** [P2]. n=3 experiments. The grey arrow heads point to the hyperphosphorylated and thereby shifted HSF1 band.

(E) ***Altered formation of nSBs by HS in RSTS$_{EP300}$ hPSFs.*** (Left panels) Representative pictures of cells in control (CTR) or heat shock conditions (HS at 43°C for 1 h; arrowheads point to nuclear stress bodies [nSBs]). The white rectangles point out two examples of the magnified cells, positive for nSBs. (Right panel) Quantification of the percentage of fibroblasts positive for nSBs, from 100 – 150 cells in n=3 different experiments. RSTS *hPSFs* were compared to HD *hPSF.* Standard deviations. ** p<0.01. Scale Bar, 10 µM.

References

Ahlskog JK, Björk JK, Elsing AN, Aspelin C, Kallio M, Roos-Mattjus P, Sistonen L. Anaphase-promoting complex/cyclosome participates in the acute response to protein-damaging stress. Mol Cell Biol. 2010 Dec;30(24):5608-20. doi: 10.1128/MCB.01506-09. PMID: 20937767

Alarcón JM, Malleret G, Touzani K, Vronskaya S, Ishii S, Kandel ER, Barco A. Chromatin acetylation, memory, and LTP are impaired in CBP+/- mice: a model for the cognitive deficit in Rubinstein-Taybi syndrome and its amelioration. Neuron. 2004 Jun 24;42(6):947-59. PMID: 15207239

Alastalo TP, Hellesuo M, Sandqvist A, Hietakangas V, Kallio M, Sistonen L. Formation of nuclear stress granules involves HSF2 and coincides with the nucleolar localization of Hsp70. J Cell Sci. 2003 Sep 1;116(Pt 17):3557-70. Epub 2003 Jul 15. PMID: 12865437

Aasland R, Gibson TJ, Stewart AF. The PHD finger: implications for chromatin-mediated transcriptional regulation.Trends Biochem Sci. 1995 Feb;20(2):56-9. PMID: 7701562

Anckar J, Sistonen L. Regulation of HSF1 function in the heat stress response: implications in aging and disease. Annu Rev Biochem. 2011;80:1089-115. doi: 10.1146/annurev-biochem-060809-095203. Review. PMID: 21417720

Akimov V, Barrio-Hernandez I, Hansen SVF, Hallenborg P, Pedersen AK, Bekker-Jensen DB, Puglia M, Christensen SDK, Vanselow JT, Nielsen MM, Kratchmarova I, Kelstrup CD, Olsen JV, Blagoev B. UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. Nat Struct Mol Biol. 2018 Jul;25(7):631-640. doi: 10.1038/s41594-018-0084-y. Epub 2018 Jul 2. PMID: 29967540

Babu A, Kamaraj M, Basu M, Mukherjee D, Kapoor S, Ranjan S, Swamy MM, Kaypee S, Scaria V, Kundu TK, Sachidanandan C. Chemical and genetic rescue of an ep300 knockdown model for Rubinstein Taybi Syndrome in zebrafish. Biochim Biophys Acta. 2018 Apr;1864(4 Pt A):1203-1215. doi: 10.1016/j.bbadis.2018.01.029. Epub 2018 Jan 31. PMID: 29409755

Bhattacherjee V, Horn KH, Singh S, Webb CL, Pisano MM, Greene RM. CBP/p300 and associated transcriptional co-activators exhibit distinct expression patterns during murine craniofacial and neural tube development. Int J Dev Biol. 2009;53(7):1097-104. doi: 10.1387/ijdb.072489vb. PMID: 1959812

Björk JK, Åkerfelt M, Joutsen J, Puustinen MC, Cheng F, Sistonen L, Nees M. Heat-shock factor 2 is a suppressor of prostate cancer invasion. Oncogene. 2016 Apr 7;35(14):1770-84. doi: 10.1038/onc.2015.241. PMID: 26119944

Björk JK, Sandqvist A, Elsing AN, Kotaja N, Sistonen L. miR-18, a member of Oncomir-1, targets heat shock transcription factor 2 in spermatogenesis. Development. 2010 Oct;137(19):3177-84. doi: 10.1242/dev.050955. Epub 2010 Aug 19. PMID: 20724452

Bodor DL, Rodríguez MG, Moreno N, Jansen LE. Analysis of protein turnover by quantitative SNAP-based pulse-chase imaging. Curr Protoc Cell Biol. 2012. Jun; Chapter 8:Unit 8.8. doi: 10.1002/0471143030.cb0808s55. PMID: 23129118

Bordoli L, Hüsser S, Lüthi U, Netsch M, Osmani H, Eckner R. Functional analysis of the p300 acetyltransferase domain: the PHD finger of p300 but not of CBP is dispensable for enzymatic activity. Nucleic Acids Res. 2001 Nov 1;29(21):4462-71. PMID: 11691934

Bowers EM, Yan G, Mukherjee C, Orry A, Wang L, Holbert MA, Crump NT, Hazzalin CA, Liszczak G, Yuan H, Larocca C, Saldanha SA, Abagyan R, Sun Y, Meyers DJ, Marmorstein R, Mahadevan LC, Alani RM, Cole PA. Virtual ligand screening of the p300/CBP histone acetyltransferase: identification of a selective small molecule inhibitor. Chem Biol. 2010 May 28;17(5):471-82. doi: 10.1016/j.chembiol.2010.03.006. PMID: 20534345

Brüschweiler S, Konrat R, Tollinger M. Allosteric communication in the KIX domain proceeds through dynamic repacking of the hydrophobic core. ACS Chem Biol. 2013 Jul 19;8(7):1600-10. doi: 10.1021/cb4002188. Epub 2013 May 20.

Bürckstümmer T, Bennett KL, Preradovic A, Schütze G, Hantschel O, Superti-Furga G, Bauch A. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. Nat Methods. 2006 3:1013-9. PMID: 17060908

Budzyński MA, Puustinen MC, Joutsen J, Sistonen L. Uncoupling Stress-Inducible Phosphorylation of Heat Shock Factor 1 from Its Activation. Mol Cell Biol. 2015 Jul;35(14):2530-40. doi: 10.1128/MCB.00816-14. Epub 2015 May 11. PMID: 25963659

Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Camp JG, Badsha F, Florio M, Kanton S, Gerber T, Wilsch-Bräuninger M, Lewitus E, Sykes A, Hevers W, Lancaster M, Knoblich JA, Lachmann R, Pääbo S, Huttner WB, Treutlein B. Proc Natl Acad Sci U S A. 2015 Dec 22;112(51):15672-7. doi: 10.1073/pnas.1520760112. PMID: 26644564

Campbell KM, Lumb KJ. Structurally distinct modes of recognition of the KIX domain of CBP by Jun and CREB. Biochemistry. 2002 Nov 26;41(47):13956-64. PMID: 12437352

Chan HM, La Thangue NB. p300/CBP proteins: HATs for transcriptional bridges and scaffolds. J Cell Sci. 2001 Jul;114(Pt 13):2363-73. PMID: 11559745

Chang Y, Ostling P, Akerfelt M, Trouillet D, Rallu M, Gitton Y, El Fatimy R, Fardeau V, Le Crom S, Morange M, Sistonen L, Mezger V. Role of heat-shock factor 2 in cerebral cortex formation and as a regulator of p35 expression. Genes Dev. 2006 Apr 1;20(7):836-47. PMID: 16600913

Cook PR, Polakowski N, Lemasson I. HTLV-1 HBZ protein deregulates interactions between cellular factors and the KIX domain of p300/CBP. J Mol Biol. 2011 Jun 10;409(3):384-98. doi: 10.1016/j.jmb.2011.04.003. Epub 2011 Apr 8. PMID: 21497608

Dai C, Whitesell L, Rogers AB, Lindquist S. Heat shock factor 1 is a powerful multifaceted modifier of carcinogenesis. Cell. 2007 Sep 21;130(6):1005-18. PMID: 17889646

Dancy BM, Cole PA. Protein lysine acetylation by p300/CBP. Chem Rev. 2015 Mar 25;115(6):2419-52. doi: 10.1021/cr500452k. Epub 2015 Jan 16. PMID: 25594381 - Erratum 27304234.

Duval R, Fritsch L, Bui LC, Berthelet J, Guidez F, Mathieu C, Dupret JM, Chomienne C, Ait-Si-Ali S, Rodrigues-Lima F. An acetyltransferase assay for CREB-binding protein based on reverse phase-ultra-fast liquid chromatography of fluorescent histone H3 peptides. Anal Biochem. 2015 Oct 1;486:35-7. doi: 10.1016/j.ab.2015.06.024. PMID: 26099937

Delvecchio M, Gaucher J, Aguilar-Gurrieri C, Ortega E, Panne D. Structure of the p300 catalytic core and implications for chromatin targeting and HAT regulation. Nat Struct Mol Biol. 2013 Sep;20(9):1040-6. doi: 10.1038/nsmb.2642. PMID: 23934153

El Fatimy R, Miozzo F, Le Mouël A, Abane R, Schwendimann L, Sabéran-Djoneidi D, de Thonel A, Massaoudi I, Paslaru L, Hashimoto-Torii K, Christians E, Rakic P, Gressens P, Mezger V. Heat shock factor 2 is a stress-responsive mediator of neuronal migration defects in models of fetal alcohol syndrome. EMBO Mol Med. 2014 Aug;6(8):1043-61. doi: 10.15252/emmm.201303311. PMID: 25027850

Elsing AN, Aspelin C, Björk JK, Bergman HA, Himanen SV, Kallio MJ, Roos-Mattjus P, Sistonen L. Expression of HSF2 decreases in mitosis to enable stress-inducible transcription and cell survival. J Cell Biol. 2014 Sep 15;206(6):735-49. doi: 10.1083/jcb.201402002. Epub 2014 Sep 8. PMID: 25202032

Geng H, Liu Q, Xue C, David LL, Beer TM, Thomas GV, Dai MS, Qian DZ. HIF1α protein stability is increased by acetylation at lysine 709. J Biol Chem. 2012 Oct 12;287(42):35496-505. doi: 10.1074/jbc.M112.400697. Epub 2012 Aug 20. PMID: 22908229

Gomez-Pastor R, Burchfiel ET, Neef DW, Jaeger AM, Cabiscol E, McKinstry SU, Doss A, Aballay A, Lo DC, Akimov SS, Ross CA, Eroglu C, Thiele DJ. Abnormal degradation of the neuronal stress-protective transcription factor HSF1 in Huntington's disease. Nat Commun. 2017 Feb 13;8:14405. doi: 10.1038/ncomms14405. PMID: 28194040

Gomez-Pastor R, Burchfiel ET, Thiele DJ. Regulation of heat shock transcription factors and their roles in physiology and disease. Nat Rev Mol Cell Biol. 2018 Jan;19(1):4-19. doi: 10.1038/nrm.2017.73. Epub 2017 Aug 30. Review. PMID: 28852220

Goto NK, Zor T, Martinez-Yamout M, Dyson HJ, Wright PE. Cooperativity in transcription factor binding to the coactivator CREB-binding protein (CBP). The mixed lineage leukemia protein (MLL) activation domain binds to an allosteric site on the KIX domain. J Biol Chem. 2002 Nov 8;277(45):43168-74. Epub 2002 Aug 29. PMID: 12205094

Grossman SR, Deato ME, Brignone C, Chan HM, Kung AL, Tagami H, Nakatani Y, Livingston DM. Polyubiquitination of p53 by a ubiquitin ligase activity of p300. Science. 2003 Apr 11;300(5617):342-4. DOI: 10.1126/science.1080386. PMID: 12690203

Hartl FU, Bracher A, Hayer-Hartl M. Molecular chaperones in protein folding and proteostasis. Nature. 2011 Jul 20;475(7356):324-32. doi: 10.1038/nature10317. PMID: 21776078

Herce HD, Deng W, Helma J, Leonhardt H, Cardoso MC. Visualization and targeted disruption of protein interactions in living cells. Nat Commun. 2013;4:2660. PMID: 24154492

Herriot R, Miedzybrodzka Z2. Antibody deficiency in Rubinstein-Taybi syndrome. Clin Genet. 2016 Mar;89(3):355-8. doi: 10.1111/cge.12671. Epub 2015 Sep 28. PMID: 26307339

Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, Dong R, Guarani V, Vaites LP, Ordureau A, Rad R, Erickson BK, Wühr M, Chick J, Zhai B, Kolippakkam D, Mintseris J, Obar RA, Harris T, Artavanis-Tsakonas S, Sowa ME, De Camilli P, Paulo JA, Harper JW, Gygi SP. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell. 2015 Jul 16;162(2):425-440. doi: 10.1016/j.cell.2015.06.043. PMID: 26186194

Jain S, Wei J, Mitrani LR, Bishopric NH. Auto-acetylation stabilizes p300 in cardiac myocytes during acute oxidative stress, promoting STAT3 accumulation and cell survival. Breast Cancer Res Treat. 2012 Aug;135(1):103-14. doi: 10.1007/s10549-012-2069-6. Epub 2012 May 5. PMID: 22562121

Jiang YQ, Wang XL, Cao XH, Ye ZY, Li L, Cai WQ. Increased heat shock transcription factor 1 in the cerebellum reverses the deficiency of Purkinje cells in Alzheimer's disease. Brain Res. 2013 Jun 26;1519:105-11. doi: 10.1016/j.brainres.2013.04.059. Epub 2013 May 9. PMID: 23665061

Jolly C, Metz A, Govin J, Vigneron M, Turner BM, Khochbin S, Vourc'h C. Stress-induced transcription of satellite III repeats. J Cell Biol. 2004 Jan 5;164(1):25-33. Epub 2003 Dec 29. PMID: 14699086

Jolly C, Morimoto R, Robert-Nicoud M, Vourc'h C. HSF1 transcription factor concentrates in nuclear foci during heat shock: relationship with transcription sites. J Cell Sci. 1997 Dec;110 (Pt 23):2935-41. PMID: 9359877

Jolly C, Usson Y, Morimoto RI. Rapid and reversible relocalization of heat shock factor 1 within seconds to nuclear stress granules. Proc Natl Acad Sci U S A. 1999 Jun 8;96(12):6769-74. PMID: 10359787

Kallio M, Chang Y, Manuel M, Alastalo TP, Rallu M, Gitton Y, Pirkkala L, Loones MT, Paslaru L, Larney S, Hiard S, Morange M, Sistonen L, Mezger V. Brain abnormalities, defective meiotic chromosome synapsis and female subfertility in HSF2 null mice. EMBO J. 2002 Jun 3;21(11):2591-601. DOI: 10.1093/emboj/21.11.2591. PMID: 12032072

Kalkhoven E, Roelfsema JH, Teunissen H, den Boer A, Ariyurek Y, Zantema A, Breuning MH, Hennekam RC, Peters DJ. Loss of CBP acetyltransferase activity by PHD finger mutations in Rubinstein-Taybi syndrome. Hum Mol Genet. 2003 Feb 15;12(4):441-50. PMID: 12566391

Kalkhoven E, Teunissen H, Houweling A, Verrijzer CP, Zantema A.The PHD type zinc finger is an integral part of the CBP acetyltransferase domain. Mol Cell Biol. 2002 Apr;22(7):1961-70. PMID: 11884585

Kauppi M, Murphy JM, de Graaf CA, Hyland CD, Greig KT, Metcalf D, Hilton AA, Nicola NA, Kile BT, Hilton DJ, Alexander WS. Point mutation in the gene encoding p300 suppresses thrombocytopenia in Mpl-/- mice.Blood. 2008 Oct 15;112(8):3148-53. doi: 10.1182/blood-2007-10-119677. Epub 2008 Aug 6. PMID: 18684867

Kawasaki H, Eckner R, Yao TP, Taira K, Chiu R, Livingston DM, Yokoyama KK. Distinct roles of the co-activators p300 and CBP in retinoic-acid-induced F9-cell differentiation. Nature. 1998 May 21;393(6682):284-9. PMID: 9607768

Kawazoe Y, Nakai A, Tanabe M, Nagata K. Proteasome inhibition leads to the activation of all members of the heat-shock-factor family. Eur J Biochem. 1998 Jul 15;255(2):356-62. PMID: 9716376

Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, Sowa ME, Rad R, Rush J, Comb MJ, Harper JW, Gygi SP. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell. 2011 Oct 21;44(2):325-40. doi: 10.1016/j.molcel.2011.08.025. Epub 2011 Sep 8. PMID: 21906983

Kim E, Wang B, Sastry N, Masliah E, Nelson PT, Cai H, Liao FF. NEDD4-mediated HSF1 degradation underlies α-synucleinopathy. Hum Mol Genet. 2016 Jan 15;25(2):211-22. doi: 10.1093/hmg/ddv445. Epub 2015 Oct 26. PMID: 26503960

Kobayashi A, Numayama-Tsuruta K, Sogawa K, Fujii-Kuriyama Y. CBP/p300 functions as a possible transcriptional coactivator of Ah receptor nuclear translocator (Arnt). J Biochem (Tokyo) 1997; 122:703-10. PMID: 9399571

Lamparter CL, Winn LM. Valproic acid exposure decreases Cbp/p300 protein expression and histone acetyltransferase activity in P19 cells. Toxicol Appl Pharmacol. 2016 Sep 1;306:69-78. doi: 10.1016/j.taap.2016.07.001. PMID: 27381264

Lancaster MA, Renner M, Martin CA, Wenzel D, Bicknell LS, Hurles ME, Homfray T, Penninger JM, Jackson AP, Knoblich JA. Cerebral organoids model human brain development and microcephaly.
Nature. 2013 Sep 19;501(7467):373-9. doi: 10.1038/nature12517. Epub 2013 Aug 28. PMID: 23995685

Lee CW, Arai M, Martinez-Yamout MA, Dyson HJ, Wright PE. Mapping the interactions of the p53 transactivation domain with the KIX domain of CBP. Biochemistry. 2009 Mar 17;48(10):2115-24. doi: 10.1021/bi802055v. PMID: 19220000

Marinova Z, Ren M, Wendland JR, Leng Y, Liang MH, Yasuda S, Leeds P, Chuang DM. Valproic acid induces functional heat-shock protein 70 via Class I histone deacetylase inhibition in cortical neurons: a potential role of Sp1 acetylation. J Neurochem. 2009 Nov;111(4):976-87. doi: 10.1111/j.1471-4159.2009.06385.x. PMID: 19765194

Mashiach E, Schneidman-Duhovny D, Andrusier N, Nussinov R, Wolfson HJ. FireDock: a web server for fast interaction refinement in molecular docking. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W229-32. doi: 10.1093/nar/gkn186. Epub 2008 Apr 19. PMID: 18424796

Mathew A, Mathur SK, Morimoto RI. Heat shock response and protein degradation: regulation of HSF2 by the ubiquitin-proteasome pathway. Mol Cell Biol. 1998 Sep;18(9):5091-8. PMID: 9710593

McMillan DR, Xiao X, Shao L, Graves K, Benjamin IJ. Targeted disruption of heat shock transcription factor 1 abolishes thermotolerance and protection against heat-inducible apoptosis. J Biol Chem. 1998 Mar 27;273(13):7523-8. PMID: 9516453

Mendillo ML, Santagata S, Koeva M, Bell GW, Hu R, Tamimi RM, Fraenkel E, Ince TA, Whitesell L, Lindquist S. HSF1 drives a transcriptional program distinct from heat shock to support highly malignant human cancers. Cell. 2012 Aug 3;150(3):549-62. doi: 10.1016/j.cell.2012.06.031. PMID: 22863008

Miozzo F, Sabéran-Djoneidi D, Mezger V. HSFs, Stress Sensors and Sculptors of Transcription Compartments and Epigenetic Landscapes. J Mol Biol. 2015 Dec 4;427(24):3793-816. doi: 10.1016/j.jmb.2015.10.007. Epub 2015 Oct 22. PMID 26482101

Neef DW, Jaeger AM, Thiele DJ. Heat shock transcription factor 1 as a therapeutic target in neurodegenerative diseases. Nat Rev Drug Discov. 2011 Dec 1;10(12):930-44. doi: 10.1038/nrd3453. Review. PMID: 22129991

Naimi DR, Munoz J, Rubinstein J, Hostoffer RW Jr. Rubinstein-Taybi syndrome: an immune deficiency as a cause for recurrent infections. Allergy Asthma Proc. 2006 May-Jun;27(3):281-4. PMID: 16913274

Nakai, A (ed.) Heat Shock Factor. (Springer, 2016).

Östling P, Björk JK, Roos-Mattjus P, Mezger V, Sistonen L. Heat shock factor 2 (HSF2) contributes to inducible expression of hsp genes through interplay with HSF1. J Biol Chem. 2007 Mar 9;282(10):7077-86. Epub 2007 Jan 9. PMID: 17213196

Parker D, Ferreri K, Nakajima T, LaMorte VJ, Evans R, Koerber SC, Hoeger C, Montminy MR.Pierce BG, Phosphorylation of CREB at Ser-133 induces complex formation with CREB-binding protein via a direct mechanism. Mol Cell Biol. 1996 Feb;16(2):694-703. PMID: 8552098

Partanen A, Motoyama J, Hui CC. Developmentally regulated expression of the transcriptional cofactors/histone acetyltransferases CBP and p300 during mouse embryogenesis. Int J Dev Biol. 1999 Sep;43(6):487-94. PMID: 10610021

Rack JG, Lutter T, Kjæreng Bjerga GE, Guder C, Ehrhardt C, Värv S, Ziegler M, Aasland R. The PHD finger of p300 influences its ability to acetylate histone and non-histone targets. J Mol Biol. 2014 Dec 12;426(24):3960-72. doi: 10.1016/j.jmb.2014.08.011. PMID: 25158095

Rallu M, Loones M, Lallemand Y, Morimoto R, Morange M, Mezger V. Function and regulation of heat shock factor 2 during mouse embryogenesis. Proc Natl Acad Sci U S A. 1997 Mar 18;94(6):2392-7. PMID: 9122205

Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963). Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology. 7: 95–9. PMID: 13990617

Raychaudhuri S, Loew C, Körner R, Pinkert S, Theis M, Hayer-Hartl M, Buchholz F, Hartl FU. Interplay of acetyltransferase EP300 and the proteasome system in regulating heat shock transcription factor 1. Cell. 2014 Feb 27;156(5):975-85. doi: 10.1016/j.cell.2014.01.055. PMID: 24581496

Rizzi N, Denegri M, Chiodi I, Corioni M, Valgardsdottir R, Cobianchi F, Riva S, Biamonti G. Transcriptional activation of a constitutive heterochromatic domain of the human genome in response to heat shock. Mol Biol Cell. 2004 Feb;15(2):543-51. Epub 2003 Nov 14. PMID: 14617804

Rothbauer U, Zolghadr K, Muyldermans S, Schepers A, Cardoso MC, Leonhardt H. A versatile nanotrap for biochemical and functional studies with fluorescent fusion proteins. Mol Cell Proteomics. 2008 Feb;7(2):282-9. Epub 2007 Oct 21. PMID: 17951627

Sandqvist A, Björk JK, Akerfelt M, Chitikova Z, Grichine A, Vourc'h C, Jolly C, Salminen TA, Nymalm Y, Sistonen L. Heterotrimerization of heat-shock factors 1 and 2 provides a transcriptional switch in response to distinct stimuli. Mol Biol Cell. 2009 Mar;20(5):1340-7. doi: 10.1091/mbc.E08-08-0864. Epub 2009 Jan 7. PMID: 19129477

Sanial M, Bécam I, Hofmann L, Behague J, Argüelles C, Gourhand V, Bruzzone L, Holmgren RA, Plessis A. Dose-dependent transduction of Hedgehog relies on phosphorylation-based feedback between the G-protein-coupled receptor Smoothened and the kinase Fused. Development. 2017 May 15;144(10):1841-1850. doi: 10.1242/dev.144782. PMID: 28360132

Santagata S, Hu R, Lin NU, Mendillo ML, Collins LC, Hankinson SE, Schnitt SJ, Whitesell L, Tamimi RM, Lindquist S, Ince TA. High levels of nuclear heat-shock factor 1 (HSF1) are associated with poor prognosis in breast cancer. Proc Natl Acad Sci U S A. 2011 Nov 8;108(45):18378-83. doi: 10.1073/pnas.1115031108. Epub 2011 Oct 31. PMID: 22042860

Santagata S, Mendillo ML, Tang YC, Subramanian A, Perley CC, Roche SP, Wong B, Narayan R, Kwon H, Koeva M, Amon A, Golub TR, Porco JA Jr, Whitesell L, Lindquist S. Tight coordination of protein translation and HSF1 activation supports the anabolic malignant state. Science. 2013 Jul 19;341(6143):1238303. doi: 10.1126/science.1238303. PMID: 23869022

Sankar N, Baluchamy S, Kadeppagari RK, Singhal G, Weitzman S, Thimmapaya B. p300 provides a corepressor function by cooperating with YY1 and HDAC3 to repress c-Myc. Oncogene. 2008 Sep 25;27(43):5717-28. doi: 10.1038/onc.2008.181. PMID: 18542060

Sarge KD, Murphy SP, Morimoto RI. Mol Cell Biol. 1993 Mar;13(3):1392-407. Activation of heat shock gene transcription by heat shock factor 1 involves oligomerization, acquisition of DNA-binding activity, and nuclear localization and can occur in the absence of stress. PMID: 8441385

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W363-7. PMID: 15980490

Sierra J, Yoshida T, Joazeiro CA, Jones KA. The APC tumor suppressor counteracts beta-catenin activation and H3K4 methylation at Wnt target genes. Genes Dev. 2006 Mar 1;20(5):586-600. DOI: 10.1101/gad.1385806. PMID: 16510874

Sistonen L, Sarge KD, Phillips B, Abravaya K, Morimoto RI. Activation of heat shock factor 2 during hemin-induced differentiation of human erythroleukemia cells. Mol Cell Biol. 1992 Sep;12(9):4104-11. PMID: 1508207

Spena S, Gervasini C, Milani D. Ultra-Rare Syndromes: The Example of Rubinstein-Taybi Syndrome. J Pediatr Genet. 2015a Sep;4(3):177-86. doi: 10.1055/s-0035-1564571. Epub 2015 Sep 28. Review. PMID: 27617129

Spena S, Milani D, Rusconi D, Negri G, Colapietro P, Elcioglu N, Bedeschi F, Pilotta A, Spaccini L, Ficcadenti A, Magnani C, Scarano G, Selicorni A, Larizza L, Gervasini C. Insights into genotype-phenotype correlations from CREBBP point mutation screening in a cohort of 46 Rubinstein-Taybi syndrome patients. Clin Genet. 2015b Nov;88(5):431-40. doi: 10.1111/cge.12537. PMID: 25388907

Takii R, Fujimoto M, Tan K, Takaki E, Hayashida N, Nakato R, Shirahige K, Nakai A. ATF1 modulates the heat shock response by regulating the stress-inducible heat shock factor 1 transcription complex. Mol Cell Biol. 2015 Jan;35(1):11-25. doi: 10.1128/MCB.00754-14. Epub 2014 Oct 13. PMID: 25312646

Tang S, Chen H, Cheng Y, Nasir MA, Kemper N, Bao E. The interactive association between heat shock factor 1 and heat shock proteins in primary myocardial cells subjected to heat stress. Int J Mol Med. 2016 Jan;37(1):56-62. doi: 10.3892/ijmm.2015.2414. PMID: 26719858

Thakur JK, Yadav A, Yadav G. Molecular recognition by the KIX domain and its role in gene regulation. Nucleic Acids Res. 2014 Feb;42(4):2112-25. doi: 10.1093/nar/gkt1147. Epub 2013 Nov 18. PMID: 24253305

Turnell AS, Stewart GS, Grand RJ, Rookes SM, Martin A, Yamano H, Elledge SJ, Gallimore PH. The APC/C and CBP/p300 cooperate to regulate transcription and cell-cycle progression. Nature. 2005 Dec 1;438(7068):690-5. PMID: 16319895

Vecsey CG, Hawk JD, Lattal KM, Stein JM, Fabian SA, Attner MA, Cabrera SM, McDonough CB,Brindle PK, Abel T, Wood MA. Histone deacetylase inhibitors enhance memory and synaptic plasticity via CREB:CBP-dependent transcriptional activation. J Neurosci 2007;27:6128–6140. PubMed: 17553985

Vihervaara A, Sergelius C, Vasara J, Blom MA, Elsing AN, Roos-Mattjus P, Sistonen L. Transcriptional response to stress in the dynamic chromatin environment of cycling and mitotic cells. Proc Natl Acad Sci U S A. 2013 Sep 3;110(36):E3388-97. doi: 10.1073/pnas.1305275110. PMID: 23959860

Vihervaara A, Sistonen L. HSF1 at a glance. J Cell Sci. 2014 Jan 15;127(Pt 2):261-6. doi: 10.1242/jcs.132605. PMID: 24421309

Wagner SA, Beli P, Weinert BT, Nielsen ML, Cox J, Mann M, Choudhary C. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteomics. 2011 Oct;10(10):M111.013284. doi: 10.1074/mcp.M111.013284. Epub 2011 Sep 1. PMID: 21890473

Wang F, Marshall CB, Yamamoto K, Li GY, Gasmi-Seabrook GM, Okada H, Mak TW, Ikura M. Structures of KIX domain of CBP in complex with two FOXO3a transactivation domains reveal promiscuity and plasticity in coactivator recruitment. Proc Natl Acad Sci U S A. 2012 Apr 17;109(16):6078-83. doi: 10.1073/pnas.1119073109. Epub 2012 Apr 2. PMID: 22474372

Wang G, Zhang J, Moskophidis D, Mivechi NF. Targeted disruption of the heat shock transcription factor (hsf)-2 gene results in increased embryonic lethality, neuronal defects, and reduced spermatogenesis. Genesis. 2003 May;36(1):48-61. PMID: 12748967

Wang X, Asea A, Xie Y, Kabingu E, Stevenson MA, Calderwood SK. RSK2 represses HSF1 activation during heat shock. Cell Stress Chaperones. 2000 Nov; 5(5): 432–437. PMID: 11189448

Westerheide SD, Anckar J, Stevens SM Jr, Sistonen L, Morimoto RI. Stress-inducible regulation of heat shock factor 1 by the deacetylase SIRT1. Science. 2009 Feb 20;323(5917):1063-6. doi: 10.1126/science.1165946. Erratum in: Science. 2013 Nov 22;342(6161):931. PMID: 19229036

Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics. 2014 Jun 15;30(12):1771-3. doi: 10.1093/bioinformatics/btu097. Epub 2014 Feb 14. PMID: 24532726

Wu C. Heat shock transcription factors: structure and regulation. Annu Rev Cell Dev Biol. 1995. PMID: 8689565

Xiao X, Zuo X, Davis AA, McMillan DR, Curry BB, Richardson JA, Benjamin IJ. HSF1 is required for extra-embryonic development, postnatal growth and protection during inflammatory responses in mice. EMBO J. 1999 Nov 1;18(21):5943-52. PMID: 10545106

Xu D, Zalmas LP, and Nicholas B La Thangue NB. A transcription cofactor required for the heat-shock response. EMBO Rep. 2008 Jul; 9(7): 662–669. PMID: 18451878

Yang XJ, Seto E. Lysine acetylation: codified crosstalk with other posttranslational modifications. Mol Cell. 2008 Aug 22;31(4):449-61. doi: 10.1016/j.molcel.2008.07.002. PMID: 18722172

Yao TP, Oh SP, Fuchs M, Zhou ND, Ch'ng LE, Newsome D, Bronson RT, Li E, Livingston DM, Eckner R. Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. Cell. 1998 May 1;93(3):361-72. PMID: 9590171

Yehezkel S, Segev Y, Viegas-Péquignot E, Skorecki K, Selig S. Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. Hum Mol Genet. 2008 Sep 15;17(18):2776-89. doi: 10.1093/hmg/ddn177. Epub 2008 Jun 16. PMID: 18558631

Yoshima T, Yura T, Yanagi H.The trimerization domain of human heat shock factor 2 is able to interact with nucleoporin p62. Biochem Biophys Res Commun. 1997 Nov 7;240(1):228-33. PMID: 9367915

Zhang Y, Wang JS, Chen LL, Zhang Y, Cheng XK, Heng FY, Wu NH, Shen YF. Repression of hsp90beta gene by p53 in UV irradiation-induced apoptosis of Jurkat cells. J Biol Chem. 2004 Oct 8;279(41):42545-51. Epub 2004 Jul 28.

Zelin E, Zhang Y, Toogun OA, Zhong S, Freeman BC. The p23 molecular chaperone and GCN5 acetylase jointly modulate protein-DNA dynamics and open chromatin status. Mol Cell. 2012 Nov 9;48(3):459-70. doi: 10.1016/j.molcel.2012.08.026. Epub 2012 Sep 27. PMID: 23022381

Zor T, De Guzman RN, Dyson HJ, Wright PE. Solution structure of the KIX domain of CBP bound to the transactivation domain of c-Myb. J Mol Biol. 2004 Mar 26;337(3):521-34. PMID: 15019774

For Supplementary data : please see BioRχiv, using doi: https://doi.org/10.1101/481457

Annexes 3 : Jupyter notebooks

3.1. Methylome capture bioinformatic workflow

To be completed

3.2. ATAC-seq bioinformatic workflow

ATAC-seq workflow : analysis of the physiological evolution of chromatin accessibility throughout brain development.

Agathe Duchateau, Délara Sabéran Djoneidi, Yoann Devriendt

First part of the analysis (quality control of the raw data, trimming, mapping and removal of duplicates and mitochondrial chromosome) was done using tools of Galaxy platform (analysis of Délara Sabéran Djoneidi and Yoann Devriendt). Second part of the analysis (from mapping to the end of the analysis) was done on Unix and R environments (analysis of Agathe Duchateau).

Download ENCODE data

To define chromatin accessibility profile of the developing brain, ATAC-seq data time-course analysis was performed, using available public ENCODE data. These ATAC-seq experiments were done using embryonic and newborn forebrains of C57BL/6N mice (experiments from Bing Ren laboratory, UCSD). Developmental stages studied in duplicates were E13.5 (ENCODE accession: ENCFF401VUV + ENCFF898NRO, ENCFF721LGJ + ENCFF777UK), E14.5 (ENCODE accession: ENCFF048MTG + ENCFF890LGM, ENCFF633MTW + ENCFF666DRJ), E15.5 (ENCODE accession: ENCFF248PXW + ENCFF825UHO, ENCFF906VXU + ENCFF500SXI), E16.5 (ENCODE accession: ENCFF058IAE + ENCFF765HUX, ENCFF776GDQ + ENCFF588XZG) and birth (postnatal day 0 - P0, ENCODE accession: ENCFF197GTC + ENCFF209GGJ, ENCFF296GZG + ENCFF664RZO).

Quality control of the raw sata

After simplifying .fastq.gz file names, FASTQC was used to verify the quality of the sequencing raw data.

Trimming

Reads having bad sequencing quality were trimmed using Trimmomatic.

Mapping

Trimmed reads were then paired-end mapped on Mus musculus reference genome mm9, using bowtie2.

Removal of duplicates and reads on mitochondrial chromosome

Then, reads were deduplicated using samtools rmdup. Reads that mapped to mitochondrial chromosome were removed using tools on Galaxy platform, by splitting the mapped reads per chromosome using bam-splitter tool. Then all the files were merged except the reads mapped on mitochondrial chromosome, using merge-bam tool.

Setting up of bioinformatic tools in a conda environment

Following tools were installed in a conda environment to continue ATAC-seq analysis. Version of each tool is indicated. R packages which are cited below were also used to analyse the data.

In [ ]:

```
conda create -n atac #conda version: 4.6.14

conda activate atac

conda install MACS2 # version 2.1.2
conda install bedtools # version 2.28.0
conda install htseq # version 0.11.2

# R version 3.5.2
# R package used for analysis :
library(edgeR) # version 3.24.3
library(ggplot2) # version 3.2.1
library(reshape2) # version1.4.3
library(GGally) # version 1.4.0
library(SARTools) # version 1.6.9
library(BiomaRt) #version 2.38.0
```

Input files download

BAM files from Galaxy, containing deduplicated mapping reads of all chromosomes, except mitochondrial one, were downloaded locally and rename following this nomenclature : stage_rep_ss_mito.bam (e.g. E15_B_ss_mito.bam). MD5sum of files of Galaxy server and downloaded files were compared and are similar.

Peak calling to find potential accessible chromatin regions

MACS2 tool was used to identify enriched regions, corresponding to potential open chromatin regions.

Selected parameters :

- macs2 callpeak : Peakcalling module
- -t treated_file.bam : input file. In case of more than one file, MACS2 merges files before doing the peakcalling.
- --name E11 : prefix name that will be given to output files
- --outdir folder : output folder
- --format BAMPE : it indicates that paired-end bam files are used.
- --gsize mm : size of the reference genome (mm for Mus musculus, mm9 = 1.87e+09)
- --tsize 50 : size of mapped reads
- --qvalue 0.05 : qvalue threshold (= FDR value threshold). Default :0.05. qvalues are based on pvalues and obtained using the Benjamini-Hochberg method
- --keep-dup 1 : if duplicates exists, it keeps only 1 read.
- --bdg : to obtain bedGraph file
- --mfold MFOLD MFOLD : select the regions within MFOLD range of high-confidence enrichment ratio against background to build model. Fold-enrichment in regions must be lower than upper limit, and higher than the lower limit. Default :5 50
- --bw BW : band width for picking regions to compute fragment size. Default :300

In [ ]:

```
%%bash

mkdir -p /mnt/g/projet_DU_DSD/results/06_Peaks_calling/dup_pool/

# Samples name
sample="E11 E12 E13 E14 E15 E16 P0"

cd /mnt/g/projet_DU_DSD/results/05_BAM_wt_mt/

for ech in ${sample}
do
    echo  "---------------------------"
    echo  "Peaks_calling - MACS2 "
    echo  "samples : ${ech}_ss_mito.bam"
    echo  "---------------------------"
    macs2 callpeak -t ${ech}_A_ss_mito.bam ${ech}_B_ss_mito.bam --name ${ech} --outdir /mnt/g/projet_D
U_DSD/results/06_Peaks_calling/dup_pool/${ech} --format BAMPE --gsize mm --tsize 50 --qvalue 0.05 --kee
p-dup 1 --bdg --mfold 5 50 --bw 300
done
```

Obtaining a table with detected peaks from all samples

bedtools multiinter was used to create a table containing all detected peaks, whether they are common between the samples or not. We obtain a table with one region per row and one sample per column. The value "1" is assigned to a sample if region is observed in this sample, otherwise "0" is assigned.

Selected parameters :

- -header : print a header line
- -i : input file
- -names : prefix name to describe each input file

In [ ]:

```
%%bash

#---------------------------
# 1.Files have to be sorted
#---------------------------
```

```
# Samples name
sample="E13 E14 E15 E16 P0"

for ech in ${sample}
do
    echo "sorted des pics de l'ech : ${ech}"
    sort -k1,1 -k2,2n /mnt/g/projet_DU_DSD/results/06_Peaks_calling/dup_pool/${ech}/${ech}_peaks.narro
wPeak > /mnt/g/projet_DU_DSD/results/06_Peaks_calling/dup_pool/${ech}/${ech}_sorted.narrowPeak
    head /mnt/g/projet_DU_DSD/results/06_Peaks_calling/dup_pool/${ech}/${ech}_sorted.narrowPeak
done

#---------------------------
# 2.Obtention of the table with all detected peaks
#---------------------------

mkdir -p /mnt/g/projet_DU_DSD/results/07_all_peaks_MI/dup_pool

cd /mnt/g/projet_DU_DSD/results/06_Peaks_calling/dup_pool/

bedtools multiinter -header -i E13/E13_sorted.narrowPeak E13/E13_sorted.narrowPeak E14/E14_sorted.narro
wPeak E14/E14_sorted.narrowPeak E15/E15_sorted.narrowPeak E15/E15_sorted.narrowPeak E16/E16_sorted.narr
owPeak E16/E16_sorted.narrowPeak P0/P0_sorted.narrowPeak P0/P0_sorted.narrowPeak -names E13_A E13_B E14
_A E14_B E15_A E15_B E16_A E16_B P0_A P0_B > /mnt/g/projet_DU_DSD/results/07_all_peaks_MI/dup_pool/E13_
a_P0_all_peaks_dup_pool.bed
```

Merge overlapping regions

bedtools merge was used to merge overlapping regions of the previous file.

Selected parameters :

- -header : print a header line
- -i : input file
- -d 0 : maximum distance allowed between features for regions to be merged. Default :0, that is, overlapping and book-ended features are merged.

In [ ]:

```
%%bash

#---------------------------
# 1.Files have to be sorted
#---------------------------

cd /mnt/g/projet_DU_DSD/results/07_all_peaks_MI/dup_pool

sed '1d' E13_a_P0_all_peaks_dup_pool.bed > nh_E13_a_P0_all_peaks_dup_pool.bed

sort -k1,1 -k2,2n nh_E13_a_P0_all_peaks_dup_pool.bed > nh_E13_a_P0_all_peaks_dup_pool_sorted.bed

#---------------------------
# 2.Merge overlapping regions
#---------------------------

mkdir -p /mnt/g/projet_DU_DSD/results/08_all_peaks_merge/dup_pool

bedtools merge -header -i nh_E13_a_P0_all_peaks_dup_pool_sorted.bed -d 0 > /mnt/g/projet_DU_DSD/results
/08_all_peaks_merge/dup_pool/E13_a_P0_all_peaks_dup_pool_merge.bed

wc -l  /mnt/g/projet_DU_DSD/results/08_all_peaks_merge/dup_pool/E13_a_P0_all_peaks_dup_pool_merge.bed
# ---> 115752 /mnt/g/projet_DU_DSD/results/08_all_peaks_merge/dup_pool/E13_a_P0_all_peaks_dup_pool_merg
e.bed <----
```

Obtaining a count table of merged regions

htseq-count was used to obtain reads count of regions for each sample. In order to respect the format of input file, I created a 'pseudo' .gff files, containing regions information.

nb : encoding of the line end of the .gff file must be in Unix format, otherwise htseq-count does not work correctly. I changed it using Notepad ++ software

using Notepad ++ software.

To avoid biases from paired-end reads overlaps, files must be sorted by name, to help htseq-count to count paired-end reads features only once.

In [ ]:

```
# Load file containing merged regions
dataset=read.delim("G:/projet_DU_DSD/results/08_all_peaks_merge/dup_pool/nh_E13_a_P0_all_peaks_dup_pool
_merge_sorted.bed", header=FALSE, sep="\t")

colnames(dataset)=c("chr","start","end")

head(dataset)

# Creation of specific columns to obtain a .gff formatted file :
dataset$source = "MACS2" #annotation source
dataset$type = "peaks" #element type
dataset$score ="."
dataset$strand = "."
dataset$phase = "."
dataset$group = paste("peak_id ","\"Peak",1:nrow(dataset),"\"",sep="") #peak_id

head(dataset)

# data reorganization
dataset.gff=dataset[,c("chr","source","type","start","end","score","strand","phase","group")]

head(dataset.gff)

# Save
write.table(dataset.gff, "G:/projet_DU_DSD/results/08_all_peaks_merge/dup_pool/nh_E13_a_P0_all_peaks_du
p_pool_merge_sorted.gff", quote=FALSE,row.names = FALSE,col.names=FALSE, sep="\t")
write.table(dataset.gff, "G:/projet_DU_DSD/results/08_all_peaks_merge/dup_pool/E13_a_P0_all_peaks_dup_p
ool_merge_sorted.gff", quote=FALSE,row.names = FALSE,col.names=TRUE, sep="\t")
```

In [ ]:

```
%%bash

cd /mnt/g/projet_DU_DSD/results/05_BAM_wt_mt/

# Samples name
sample="E13_A E13_B E14_A E14_B E15_A E15_B E16_A E16_B P0_A P0_B"

for ech in ${sample}
do
    echo "-----------------------------------"
    echo "samtools sort - sample : ${ech}"
    echo "-----------------------------------"
    samtools sort -n -o ${ech}_ss_mito_sorted.bam ${ech}_ss_mito.bam

    echo "-----------------------------------"
    echo "htseq-count - sample : ${ech}"
    echo "-----------------------------------"
    htseq-count --stranded=no --format bam --order name -a=10 --type=peaks --idattr=peak_id --mode=unio
n --nonunique=none ${ech}_ss_mito_sorted.bam /mnt/g/projet_DU_DSD/results/08_all_peaks_merge/dup_pool/n
h_E13_a_P0_all_peaks_dup_pool_merge_sorted_Unix.gff > /mnt/g/projet_DU_DSD/results/09_all_peaks_count_t
able/dup_pool/count_${ech}_dup_pool.txt
done
```

Selected parameters :

- --stranded=no : to specify that library is non-strand specific
- --format bam : type of alignment file
- --order name : to specify the sorting order. For paired-end data, the alignment have to be sorted either by read name or by alignment position.
- -a 10 : skip all reads with alignment quality lower than the given minimum value. Default :10
- --type=peaks : element type of the .gff file to be considered, other element types are ignored
- --idattr=peak_id : attribute from .gff file to be used as element ID.
- --mode=union : mode to handle reads overlapping more than one feature. With union mode, a read is count if there is a overlap with a given region, even if the overlap is partial, as long as there is no ambiguous case (such as overlap with two distinct regions)

- --nonunique=none : reads that mapped several genes is not considered

Statistical analysis

To identify regions where chromatin accessibility significantly changes during development, pairwise comparisons of successive developmental stages were performed using **edgeR**, an R software and Bioconductor package.

We used (with adaptations), script template from **Sartools** R package, which implementes some **edgeR** functions, and its script template template_script_edgeR.r. Main steps of the statistical analysis are described below, following by the executing code.

Filtering very-low count regions

Since regions with very low count across all samples could interfere with the statistical approximations of edgeR, these regions were removed using SARTools filter, *i.e.* by selecting regions which contain at least minReplicates (smallest number of replicates = 2) with at least counts per million cutoff (cpmCutoff = 1).

Normalization to avoid regions composition biases

To be able to compare read counts between samples, normalization was carried out according to the edgeR package. Library of each sample was normalised for regions composition using **calcNormFactors()** function, based on a trimmed mean of M-values (TMM) between each pair of samples.
Boxplots of raw- and normalised counts distribution were compared to verify the quality of the normalization process. Normalization is supposed to stabilize distributions across samples.

Statistical test for differential chromatin accessibility of regions between two successive developmental stages

Statistical test

Differential analysis was carried out according to the edgeR model. **edgeR** aims at fitting one linear model per region. The edgeR model assumes that the count data follow a negative binomial distribution which is a suitable way to analysis the data when variance is higher than the mean. First step of the statistical procedure is to estimate the dispersion of the data.
After estimation of dispersions - *i.e.* common (unique value), trended (estimated with splines) and tagwise (estimated from feature counts) dispersions - fitting to a generalized linear model (glm) can be done using **glmFit()** function. Then, statistical test was performed for all pairwise comparisons of successive developmental stages, using a likehood ratio test (lrt method).

Histogram of raw pvalues from the statistical test was plotted to verify the shape of the distribution, that is expected to follow an uniform distribution, with a peak around 0.
Benjamini-Hochberg method was applied to adjust pvalues computed by the statistical test in order to take into account multiple testing and control the false positive rate (FDR). Threshold of statistical significance was set to 0.05.

Differential analysis plots

MA-plot

For each comparison, MA-plot of the data was done. It represents the log ratio of differential expression as a function of the mean intensity for each region.

Volcano plot

Volcano plots for the comparisons were performed. It represents the log of the adjusted pvalue as a function of the log ratio of differential expression.

Description plots to estimate variability within the experiment

Scatterplot

A pairwise scatterplot was produced to verify that intra-group (replicates) similarities are higher than inter-group (samples from distinct developmental stage) ones . This plot is obtained using log2(counts+1), instead of raw count values. This pairwise scatterplot is associated with a SERE statistic, that was used as a similarity index between ATAC-seq samples.

Hierarchical clustering and dendrogram

Hierarchical clustering and dendrogram were performed to estimate variability between samples, *i.e.* see if samples from distinct developmental stages are seperated, while replicates are close to each other. Hierarchical clustering was done after a transformation of the count data as moderated log-counts-per-million.

To plot dendrogram, an euclidean distance is computed between samples. dendrogram is then obtained from CPM data and built upon the Ward criterion.

multidimensional scaling plot

First two dimensions of a multidimensional scaling plot were plotted to visualise experiment variability. If biological variability is the main source of variance in the data, first dimension is expected to separate samples from the different developmental stages.

Adaptations of SARTools functions

- added a filter to remove uncovered reads. counts_nf --> counts. In theory, this step is not necessary because uncovered regions will be removed during the filter proposed by SARTools, which deletes regions with very low-counts. However, it allows to have a double control of this step.
- added a unormalised library size histogram graph plot (ggplot2).
- change the raw pvalues histogram design (ggplot2) : summarizeResults.edgeR() --> summarizeResults.edgeR.AD()
- added supplementary graphs in run.edgeR() function
- added export of supplementary informations (logFC, FDR)
- added a filter using logFC and padj thresholds to obtain a restricted list of differentially open or closed regions, in addition to the list given by SARTools script (based on padj threshold only).

In [ ]:

```
# 1. Creation of a file called target.txt, which contains samples information:
target_cible = matrix(0,ncol=4, nrow=10)
colnames(target_cible)=c("label","files","group","day")
target_cible[,"label"]= c("E13.5_A","E13.5_B","E14.5_A", "E14.5_B","E15.5_A","E15.5_B","E16.5_A", "E16.
5_B","P0_A","P0_B")

target_cible[,"files"]=c("count_E13_A_dup_pool.txt", "count_E13_B_dup_pool.txt", "count_E14_A_dup_pool.
txt","count_E14_B_dup_pool.txt",
                         "count_E15_A_dup_pool.txt", "count_E15_B_dup_pool.txt", "count_E16_A_dup_pool.
txt","count_E16_B_dup_pool.txt",
                         "count_P0_A_dup_pool.txt", "count_P0_B_dup_pool.txt")

target_cible[,"group"]=c(rep("E13",2),rep("E14",2),rep("E15",2), rep("E16",2),rep("PO",2))
target_cible[,"day"]="d1"

head(target_cible)
#     label     files                       group day
# [1,] "E13.5_A" "count_E13_A_dup_pool.txt" "E13" "d1"
# [2,] "E13.5_B" "count_E13_B_dup_pool.txt" "E13" "d1"
# [3,] "E14.5_A" "count_E14_A_dup_pool.txt" "E14" "d1"
# [4,] "E14.5_B" "count_E14_B_dup_pool.txt" "E14" "d1"
# [5,] "E15.5_A" "count_E15_A_dup_pool.txt" "E15" "d1"
# [6,] "E15.5_B" "count_E15_B_dup_pool.txt" "E15" "d1"

write.table(target_cible, "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/target.t
xt", quote=FALSE, col.names=TRUE,row.names=FALSE, sep="\t")



# This file is then splitted to obtain pairwise successive developmental stages.



# 2. Identification of differentially open or closed chromatin regions during brain development.
# edgeR analysis using modified SARTools script

# ---------
# ----------------------
# ------------------------------------------
# E13 vs E14 developmental stages comparison
# ------------------------------------------
# ----------------------
# ---------

dir.create("G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E13.14", recu
rsive=TRUE)

# -------------------
# Parameters setting
# -------------------
```

```
rm(list=ls())

workDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E13.14"
# working directory for the R session

projectName <- "Sartools-edgeR-E13.14"              # name of the project
author <- "Agathe D."                               # author of the statistical analysis/report

targetFile <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/targetE13.14.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",       # names of the features to be removed
                      "ambiguous", "no_feature",    # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                   # factor of interest
condRef <- "E13"                                    # reference biological condition
batch <- NULL                                       # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                       # threshold of statistical significance
pAdjustMethod <- "BH"                               # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                      # counts-per-million cut-off to filter low counts
gene.selection <- "pairwise"                        # selection of the features in MDSPlot
normalizationMethod <- "TMM"                        # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE

nsamples = 4                                       # nb of samples
nstage = 2                                         # nb of developemental stages
nrep = 2                                           # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")


# -------------------
# Load packages
# -------------------

setwd(workDir)
library(SARTools)
library(reshape2)
library(ggplot2)
library(GGally)

#--------------------
# Running script
# checking parameters (unchanged SARTools function)
#--------------------

if (forceCairoGraph) options(bitmapType="cairo")

checkParameters.edgeR(projectName=projectName,author=author,targetFile=targetFile,
                      rawDir=rawDir,featuresToRemove=featuresToRemove,varInt=varInt,
                      condRef=condRef,batch=batch,alpha=alpha,pAdjustMethod=pAdjustMethod,
                      cpmCutoff=cpmCutoff,gene.selection=gene.selection,
                      normalizationMethod=normalizationMethod,colors=colors)

#--------------------
# loading target file (unchanged SARTools function)
#--------------------
target <- loadTargetFile(targetFile=targetFile, varInt=varInt, condRef=condRef, batch=batch)

#-------------------
# loading counts
#-------------------
counts_nf <- loadCountData(target=target, rawDir=rawDir, featuresToRemove=featuresToRemove)

# Delete rows without counts :
counts <- counts_nf[-(which(rowSums(counts_nf) < 1)),]
```

```
#------------------
# description plots
#------------------
# (unchanged SARTools function)
majSequences <- descriptionPlots(counts=counts, group=target[,varInt], col=colors)

# Supplementary graphs
# Print library size (unormalised data) :
counts_reshape = melt(counts, id.vars = colnames(counts), variable.name='samples_name')
head(counts_reshape)
dim(counts_reshape)
colnames(counts_reshape) = c("peak_name","samples_name","filter_count")
counts_reshape$stages = with(counts_reshape, rep(substr(colnames(counts), 1, 3), each = nrow(counts)))

lib_size_unorm = ggplot(counts_reshape, aes(x=samples_name,y=filter_count, fill = stages))
lib_size_unorm + geom_bar(stat="identity") + scale_fill_manual(values=c("skyblue","royalblue", "pink","
tomato2","darkgreen")) +
  ggtitle("library size (unormalised dataset)")+ theme_light() +
  scale_x_discrete(name ="samples_name") + scale_y_continuous("total number of reads") +
  theme(plot.title = element_text(hjust = 0.5))
ggsave(path ="figures/", filename = "01 library size - unormalised dataset.pdf")
ggsave(path ="figures/", filename = "01 library size - unormalised dataset.png")

#---------------
# edgeR analysis (adapted from SARTools function)
#---------------

run.edgeR <- function(counts, target, varInt, condRef, batch=NULL, cpmCutoff=1,
                      normalizationMethod="TMM", pAdjustMethod="BH", ...){

  # filtering very low-count regions : select features which contain at least minReplicates (smallest n
umber of replicates) with at least cpmCutoff counts per million
  minReplicates <- min(table(target[,varInt]))
  fcounts <- counts[rowSums(cpm(counts) >= cpmCutoff) >= minReplicates,]
  cat("Number of features discarded by the filtering:\n")
  cat(nrow(counts)-nrow(fcounts),"\n")

  # building dge object
  design <- formula(paste("~", ifelse(!is.null(batch), paste(batch,"+"), ""), varInt))
  dge <- DGEList(counts=fcounts, remove.zeros=TRUE)
  dge$design <- model.matrix(design, data=target)
  cat("\nDesign of the statistical model:\n")
  cat(paste(as.character(design),collapse=" "),"\n")

  ggsave(path ="figures/", filename ="02 - correlation between samples - filter - unnorm.pdf",ggpairs(a
s.data.frame(dge$counts[,1:nsamples])))
  ggsave(path ="figures/", filename ="02 - correlation between samples - filter - unnorm.png",ggpairs(a
s.data.frame(dge$counts[,1:nsamples])))

  ggsave(path ="figures/", filename ="03 - correlation between log samples - filter - unnorm.pdf",ggpai
rs(as.data.frame(log(dge$counts[,1:nsamples]))))
  ggsave(path ="figures/", filename ="03 - correlation between log samples - filter - unnorm.png",ggpai
rs(as.data.frame(log(dge$counts[,1:nsamples]))))

  # normalization for regions composition bias
  dge <- calcNormFactors(dge, method=normalizationMethod)
  cat("\nNormalization factors:\n")
  print(dge$samples$norm.factors)

  # estimating dispersions
  dge <- estimateGLMCommonDisp(dge, dge$design)
  dge <- estimateGLMTrendedDisp(dge, dge$design)
  dge <- estimateGLMTagwiseDisp(dge, dge$design)

  # graphical representation
  # BCV based on average log CPM (dispersion representation) :
  pdf("figures/04 dispersion plot - glmFit method .pdf")
  plotBCV(dge)
  dev.off()

  png("figures/04 dispersion plot - glmFit method .png")
  plotBCV(dge)
  dev.off()

  # MDS plot :
  pdf("figures/05 MDS plot - glmFit method - bcv.pdf")
  mds = plotMDS(dge,  method="bcv",col=rep(couleur[1:nstage],each=nrep), pch=1, cex=1, xlim=c(-0.55,1.0
```

```
),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],each=nrep), pos=3)
  dev.off()

  png("figures/05 MDS plot - glmFit method - bcv.png")
  mds = plotMDS(dge,  method="bcv",col=rep(couleur[1:nstage],each=nrep), pch=1, cex=1, xlim=c(-0.55,1.0
),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],each=nrep), pos=3)
  dev.off()

  pdf("figures/06 MDS plot - glmFit method - logFC.pdf")
  mds = plotMDS(dge,  method="logFC",col=rep(couleur[1:nstage],each=nrep), pch=1, cex=1, xlim=c(-0.55,1
.0),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],each=nrep), pos=3)
  dev.off()

  png("figures/06 MDS plot - glmFit method - logFC.png")
  mds = plotMDS(dge,  method="logFC",col=rep(couleur[1:nstage],each=nrep), pch=1, cex=1, xlim=c(-0.55,1
.0),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],each=nrep), pos=3)
  dev.off()

  # statistical testing: perform all the comparisons between the levels of varInt
  fit <- glmFit(dge, dge$design, ...)
  cat(paste("Coefficients of the model:",paste(colnames(fit$design),collapse="  ")),"\n")
  colsToTest <- grep(varInt,colnames(fit$design))
  namesToTest <- paste0(gsub(varInt,"",colnames(fit$design)[colsToTest]),"_vs_",condRef)
  results <- list()

  # testing coefficients individually (tests againts the reference level)
  for (i in 1:length(colsToTest)){
    cat(paste0("Comparison ",gsub("_"," ",namesToTest[i]),": testing coefficient ",colnames(fit$design)
[colsToTest[i]]),"\n")
    lrt <- glmLRT(fit, coef=colsToTest[i])
    results[[namesToTest[i]]] <- topTags(lrt,n=nrow(dge$counts),adjust.method=pAdjustMethod,sort.by="no
ne")$table
  }
  # defining contrasts for the other comparisons (if applicable)
  if (length(colsToTest)>=2){
    colnames <- gsub(varInt,"",colnames(fit$design))
    for (comp in combn(length(colsToTest),2,simplify=FALSE)){
      contrast <- numeric(ncol(dge$design))
      contrast[colsToTest[comp[1:2]]] <- c(-1,1)
      namecomp <- paste0(colnames[colsToTest[comp[2]]],"_vs_",colnames[colsToTest[comp[1]]])
      cat(paste0("Comparison ",gsub("_"," ",namecomp),": testing contrast (",paste(contrast,collapse=",
"),")"),"\n")
      lrt <- glmLRT(fit, contrast=contrast)
      results[[namecomp]] <- topTags(lrt,n=nrow(dge$counts),adjust.method=pAdjustMethod,sort.by="none")
$table
    }
  }

  return(list(dge=dge,results=results,lrt=lrt))
}


out.edgeR <- run.edgeR(counts=counts, target=target, varInt=varInt, condRef=condRef,
                       batch=batch, cpmCutoff=cpmCutoff, normalizationMethod=normalizationMethod,
                       pAdjustMethod=pAdjustMethod)

#--------------------
# MDS + clustering (unchanged SARTools function)
#--------------------
exploreCounts(object=out.edgeR$dge, group=target[,varInt], gene.selection=gene.selection, col=colors)

#--------------------
# exporting results of the differential analysis (adapted from SARTools function)
#--------------------

rawpHist_AD <- function(complete_AD, outfile=TRUE){
  ncol <- ifelse(length(complete_AD)<=4, ceiling(sqrt(length(complete_AD))), 3)
  nrow <- ceiling(length(complete_AD)/ncol)

  par(mfrow=c(nrow,ncol))
  for (name in names(complete_AD)){
    ggplot(as.data.frame(complete_AD[[name]]), aes(x=pvalue, fill="tomato2")) + geom_histogram(binwidth
= 0.025, color="black") + scale_fill_discrete(name = "", labels = "pvalues") + labs(title = "Raw pvalue
```

```
s histogram") + theme(plot.margin = margin(2,.8,2,.8, "cm"))
  }
  if (outfile) ggsave(path= "figures/", filename="rawpHist_ggplot.pdf")
  if (outfile) ggsave(path= "figures/", filename="rawpHist.png")
  if (outfile) dev.off()
}


exportResults.edgeR <- function(out.edgeR, group, counts, alpha=0.05, export=TRUE){

  dge <- out.edgeR$dge
  res <- out.edgeR$results

  # raw count, normalised count and baseMean
  tmm <- dge$samples$norm.factors
  N <- colSums(dge$counts)
  f <- tmm * N/mean(tmm * N)
  normCounts <- round(scale(dge$counts, center=FALSE, scale=f))
  base <- data.frame(Id=rownames(counts), counts)
  names(base) <- c("Id", colnames(counts))
  norm.bm <- data.frame(Id=rownames(normCounts),normCounts)
  names(norm.bm) <- c("Id", paste0("norm.",colnames(normCounts)))
  norm.bm$baseMean <- round(apply(scale(dge$counts, center=FALSE, scale=f),1,mean),2)
  for (cond in levels(group)){
    norm.bm[,cond] <- round(apply(as.data.frame(normCounts[,group==cond]),1,mean),0)
  }
  base <- merge(base,norm.bm,by="Id",all=TRUE)

  complete <- list()
  for (name in names(res)){
    complete.name <- base

    # add info from res
    res.name <- data.frame(Id=rownames(res[[name]]),FC=round(2^(res[[name]][,"logFC"]),3),
                           log2FoldChange=round(res[[name]][,"logFC"],3),pvalue=res[[name]][,"PValue"],
                           padj=res[[name]][,ifelse("FDR" %in% names(res[[name]]), "FDR", "FWER")])
    complete.name <- merge(complete.name, res.name, by="Id", all=TRUE)
    # add info from dge
    dge.add <- data.frame(Id=rownames(dge$counts),tagwise.dispersion=round(dge$tagwise.dispersion,4),
                          trended.dispersion=round(dge$trended.dispersion,4))
    complete.name <- merge(complete.name, dge.add, by="Id", all=TRUE)
    complete[[name]] <- complete.name

    if (export){
      # obtain sartools and restricted lists of differentially open and closed regions durong brain dev
elopment
      up.name <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChange>=0),]
      up.name <- up.name[order(up.name$padj),]
      down.name <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChange<=0),
]
      down.name <- down.name[order(down.name$padj),]
      up.name.strict <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChange>
=1),]
      up.name.strict <- up.name.strict[order(up.name.strict$padj),]
      down.name.strict <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChang
e<=-1),]
      down.name.strict <- down.name.strict[order(down.name.strict$padj),]
      for (i in list("up.name", "down.name","up.name.strict","down.name.strict")){
        dimension = nrow(get(i))
        print(paste("number of rows :", i, "=", dimension,sep=" "))
      }

      # exports
      name <- gsub("_","",name)
      write.table(complete.name, file=paste0("tables/",name,".complete.txt"), sep="\t", row.names=FALSE
, dec=".", quote=FALSE)
      write.table(up.name, file=paste0("tables/", name,".up.txt"), row.names=FALSE, sep="\t", dec=".",
quote=FALSE)
      write.table(down.name, file=paste0("tables/", name,".down.txt"), row.names=FALSE, sep="\t", dec="
.", quote=FALSE)
      write.table(up.name.strict, file=paste0("tables/", name,".up.strict.txt"), row.names=FALSE, sep="
\t", dec=".", quote=FALSE)
      write.table(down.name.strict, file=paste0("tables/", name,".down.strict.txt"), row.names=FALSE, s
ep="\t", dec=".", quote=FALSE)
    }
  }
```

```
  return(complete)
}


# summary of the analysis (boxplots, dispersions, export table, nDiffTotal, histograms, MA plot)
summarizeResults.edgeR.AD = function (out.edgeR, group, counts, alpha = 0.05, col = c("lightblue",
                                                                                       "orange", "Medium
VioletRed", "SpringGreen"), log2FClim = NULL,
                                      padjlim = NULL)
{
  if (!I("figures" %in% dir()))
    dir.create("figures", showWarnings = FALSE)
  if (!I("tables" %in% dir()))
    dir.create("tables", showWarnings = FALSE)
  countsBoxplots(out.edgeR$dge, group = group, col = col)
  BCVPlot(dge = out.edgeR$dge)
  complete <- exportResults.edgeR(out.edgeR = out.edgeR, group = group,
                                  counts = counts, alpha = alpha)
  nDiffTotal <- nDiffTotal(complete = complete, alpha = alpha)
  cat("Number of features down/up and total:\n")
  print(nDiffTotal, quote = FALSE)
  rawpHist_AD(complete_AD = complete, outfile = TRUE)
  MAPlot(complete = complete, alpha = alpha, log2FClim = log2FClim)
  volcanoPlot(complete = complete, alpha = alpha, padjlim = padjlim)
  return(list(complete = complete, nDiffTotal = nDiffTotal))
}


summaryResults.AD <- summarizeResults.edgeR.AD(out.edgeR, group=target[,varInt], counts=counts, alpha=a
lpha, col=colors)

#--------------------
# save image of the R session (unchanged SARTools script)
#--------------------
save.image(file=paste0(projectName, ".RData"))

#--------------------
# generating HTML report (adapted from SARTools function)
#--------------------
writeReport.edgeR(target=target, counts=counts, out.edgeR=out.edgeR, summaryResults=summaryResults.AD,
                  majSequences=majSequences, workDir=workDir, projectName=projectName, author=author,
                  targetFile=targetFile, rawDir=rawDir, featuresToRemove=featuresToRemove, varInt=varIn
t,
                  condRef=condRef, batch=batch, alpha=alpha, pAdjustMethod=pAdjustMethod, cpmCutoff=cpm
Cutoff,
                  colors=colors, gene.selection=gene.selection, normalizationMethod=normalizationMethod
)

#--------------------
# save FDR and logCPM information (new in the script)
#--------------------

if (length(names(out.edgeR$results))){
  write.table(out.edgeR$results, file="tables/results_FDR_logCPM_all_filter.txt", sep="\t", row.names=F
ALSE, dec=".", quote=FALSE)
}else{
  print("Certaines données ne seront pas sauvegardées")
}




# ---------
# ----------------------
# ------------------------------------------
# Same code for E14 vs E15 developmental stages comparison with following parameters setting
# ------------------------------------------
# ----------------------
# ---------

dir.create("G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E14.15", recu
rsive=TRUE)

rm(list=ls())

workDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E14.15"
# working directory for the R session
```

```
projectName <- "Sartools-edgeR-E14.15"                # name of the project
author <- "Agathe D."                                 # author of the statistical analysis/report

targetFile <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/targetE14.15.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",         # names of the features to be removed
                      "ambiguous", "no_feature",      # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                     # factor of interest
condRef <- "E14"                                      # reference biological condition
batch <- NULL                                         # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                         # threshold of statistical significance
pAdjustMethod <- "BH"                                 # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                        # counts-per-million cut-off to filter low counts
gene.selection <- "pairwise"                          # selection of the features in MDSPlot
normalizationMethod <- "TMM"                          # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE

nsamples = 4                                        # nb of samples
nstage = 2                                          # nb of developemental stages
nrep = 2                                            # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")



# ---------
# ----------------------
# ------------------------------------------
# Same code for E15 vs E16 developmental stages comparison with following parameters setting
# ------------------------------------------
# ----------------------
# ---------

dir.create("G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E15.16", recu
rsive=TRUE)

rm(list=ls())

workDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E15.16"
# working directory for the R session

projectName <- "Sartools-edgeR-E15.16"                # name of the project
author <- "Agathe D."                                 # author of the statistical analysis/report

targetFile <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/targetE15.16.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",         # names of the features to be removed
                      "ambiguous", "no_feature",      # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                     # factor of interest
condRef <- "E15"                                      # reference biological condition
batch <- NULL                                         # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                         # threshold of statistical significance
pAdjustMethod <- "BH"                                 # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                        # counts-per-million cut-off to filter low counts
```

```
gene.selection <- "pairwise"                          # selection of the features in MDSPlot
normalizationMethod <- "TMM"                          # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE

nsamples = 4                                          # nb of samples
nstage = 2                                            # nb of developemental stages
nrep = 2                                              # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")


# ---------
# ----------------------
# -------------------------------------------
# Same code for E16 vs P0 developmental stages comparison with following parameters setting
# -------------------------------------------
# ----------------------
# ---------

dir.create("G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E16.P0", recu
rsive=TRUE)

rm(list=ls())

workDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E16.P0"
# working directory for the R session

projectName <- "Sartools-edgeR-E16.P0"                # name of the project
author <- "Agathe D."                                 # author of the statistical analysis/report

targetFile <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/targetE16.P0.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/raw_data/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",         # names of the features to be removed
                      "ambiguous", "no_feature",      # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                     # factor of interest
condRef <- "E16"                                      # reference biological condition
batch <- NULL                                         # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                         # threshold of statistical significance
pAdjustMethod <- "BH"                                 # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                        # counts-per-million cut-off to filter low counts
gene.selection <- "pairwise"                          # selection of the features in MDSPlot
normalizationMethod <- "TMM"                          # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE

nsamples = 4                                          # nb of samples
nstage = 2                                            # nb of developemental stages
nrep = 2                                              # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")
```

Obtention of chromosomal coordinates of each region

To associate a chromosomal position at each peak name of region dectected as differentially open or closed regions during brain development, combination of different files is necessary :

In [ ]:

```
%%bash
```

```
# 1. Obtain Region ID
# Keep only ENSEMBL ID of regions dectected as differentially open or closed
cd /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E13.14/tables
sed '1d' E14vsE13.down.txt | cut -d$'\t' -f1 > Peak_ID_down_DOCR_E13.14_mm9.txt
sed '1d' E14vsE13.up.txt | cut -d$'\t' -f1 > Peak_ID_up_DOCR_E13.14_mm9.txt

cd /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E14.15/tables
sed '1d' E15vsE14.down.txt | cut -d$'\t' -f1 > Peak_ID_down_DOCR_E14.15_mm9.txt
sed '1d' E15vsE14.up.txt | cut -d$'\t' -f1 > Peak_ID_up_DOCR_E14.15_mm9.txt

cd /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E15.16/tables
sed '1d' E16vsE15.down.txt | cut -d$'\t' -f1 > Peak_ID_down_DOCR_15.E16_mm9.txt
sed '1d' E16vsE15.up.txt | cut -d$'\t' -f1 > Peak_ID_up_DOCR_E15.16_mm9.txt

cd /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E14.16/tables
sed '1d' E16vsE14.down.txt | cut -d$'\t' -f1 > Peak_ID_down_DOCR_14.E16_mm9.txt
sed '1d' E16vsE14.up.txt | cut -d$'\t' -f1 > Peak_ID_up_DOCR_E14.16_mm9.txt

cd /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E16.P0/tables
sed '1d' POvsE16.down.txt | cut -d$'\t' -f1 > Peak_ID_down_DOCR_E16.P0_mm9.txt
sed '1d' POvsE16.up.txt | cut -d$'\t' -f1 > Peak_ID_up_DOCR_E16.P0_mm9.txt


# 2. Combine ENSEMBL ID wih .gff file containing information of regions/peaks detected by MACS2

# 2.1 dataset 1 (= edgeR output) formatting
# Remove header
cd  /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E13.14/tables
sed '1d' E14vsE13.down.txt > nh_down_DOCR_E13.14_mm9.txt
sed '1d' E14vsE13.up.txt > nh_up_DOCR_E13.14_mm9.txt

cd  /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E14.15/tables
sed '1d' E15vsE14.down.txt > nh_down_DOCR_E14.15_mm9.txt
sed '1d' E15vsE14.up.txt > nh_up_DOCR_E14.15_mm9.txt

cd  /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E15.16/tables
sed '1d' E16vsE15.down.txt > nh_down_DOCR_E15.16_mm9.txt
sed '1d' E16vsE15.up.txt > nh_up_DOCR_E15.16_mm9.txt

cd  /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E14.16/tables
sed '1d' E16vsE14.down.txt > nh_down_DOCR_E14.16_mm9.txt
sed '1d' E16vsE14.up.txt > nh_up_DOCR_E14.16_mm9.txt

cd  /mnt/g/projet_DU_DSD/results/09_all_peaks_count_table/dup_pool/SARTools/Analyse_L/E16.P0/tables
sed '1d' POvsE16.down.txt > nh_down_DOCR_E16.P0_mm9.txt
sed '1d' POvsE16.up.txt > nh_up_DOCR_E16.P0_mm9.txt

#copy and paste files in /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/raw_data folder

mkdir -p /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/raw_data
```

In [ ]:

```
# 2.2 Load dataset 1 in R environment

# List containing "Peaks_ID" files
list_files = list.files("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/raw_data/")
list_files_Peak = list_files[which(grepl("nh",list_files))]

# 2.3 Load dataset 2 (= .gff file containing informations of regions/peaks detected by MACS2) in R envi
ronment

my_gff = read.table("G:/projet_DU_DSD/results/08_all_peaks_merge/dup_pool/nh_E13_a_P0_all_peaks_dup_poo
l_merge_sorted_Unix_tab.gff",
                    sep = "\t", header = FALSE)

# Verifications
dim(my_gff)
head(my_gff)
colnames(my_gff) = c("chr","source","element_type","start","end","V6","strand","V8","V9", "ID")

# 2.4 Combine informations from datasets 1 and 2

for(i in 1:10){
  ID_sd = list_files_Peak[i]
  sartools_file = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/raw
```

```
_data/",ID_sd,sep="/"), header = FALSE, sep = "\t", quote="", dec=".")

  colnames(sartools_file) = c("ID", "cond1_rep1", "cond1_rep2", "cond2_rep1", "cond2_rep2", "norm.cond1
_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2", "baseMean", "cond1", "cond2", "FC", "l
og2FoldChange", "pvalue", "padj", "tagwise.dispersion", "trended.dispersion")
  merge_dataset = merge(sartools_file,my_gff, by="ID")
  print(dim(sartools_file))
  print(dim(merge_dataset))

  write.table(merge_dataset, file=paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/gff_I
D",ID_sd, ".txt",sep="_"), quote=FALSE, sep="\t", row.names = FALSE, col.names = TRUE)
}
```

Conversion of regions coordinates : mm9 to mm10

In order to combine ATAC-seq data with results of others analyses, mm9 coordinates of differentially open or closed regions were converted into mm10 coordinates, using **[LiftOver]**(https://genome.ucsc.edu/cgi-bin/hgLiftOver) tool proposed by **UCSC**. For that, only chromosomal coordinates (chromosome-start-end) are kept to generate a .bed file.

In [ ]:

```
%%bash

mkdir /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10

cd /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/

# Samples name
sample="E13.14 E14.15 E15.16 E16.P0 E14.16"

# keep only bed information (chr - start - end) and sort the .bed file
for file in $sample
do
echo "${file}"
sed '1d' gff_ID_nh_down_DOCR_${file}_mm9.txt | awk '{ print $19"\t"$22"\t"$23 }' | sort -k1,1 -k2,2n >
bed_gff_ID_down_DOCR_${file}_mm9.sorted.txt

sed '1d' gff_ID_nh_up_DOCR_${file}_mm9.txt | awk '{ print $19"\t"$22"\t"$23 }' | sort -k1,1 -k2,2n > b
ed_gff_ID_up_DOCR_${file}_mm9.sorted.txt

echo "bed_gff_ID_down_DOCR_${file}_mm9.sorted.txt"
head bed_gff_ID_down_DOCR_${file}_mm9.sorted.txt
echo "bed_gff_ID_up_DOCR_${file}_mm9.sorted.txt"
head bed_gff_ID_up_DOCR_${file}_mm9.sorted.txt

done
```

Default settings of **LiftOver** tool have been retained (*i.e.* at least 0.95 as the minimum ratio of bases that must remap). Some regions can't be converted for distinct reasons:

In [ ]:

```
# For E14.15 up :
#Partially deleted in new (Sequence insufficiently intersects one chain)
chr9 100999669 101000211

# For E16.P0 up (6 failed) :
#Deleted in new (Sequence intersects no chains)
chr18 11406894 11407504
#Partially deleted in new (Sequence insufficiently intersects one chain)
chr4 130185092 130185630
#Partially deleted in new (Sequence insufficiently intersects one chain)
chrUn_random 126830 127059
#Split in new (Sequence insufficiently intersects multiple chains)
chrUn_random 3318185 3318608
#Split in new (Sequence insufficiently intersects multiple chains)
chrUn_random 3756019 3756421
#Deleted in new (Sequence intersects no chains)
chrUn_random 4062708 4064266
```

Files containing mm10 coordinates have been renamed (e.g. bed_gff_ID_down_DOCR_E15.16_mm10.bed), then sorted and

Files containing mm10 coordinates have been renamed (e.g. bed_gff_ID_down_DOCR_E15.16_mm10.bed), then sorted and associated to mm9 coordinates to keep both annotations.

In [ ]:

```
%%bash

cd /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10

# 1. Sort mm10 .bed files
for file in bed*
do  sort -k1,1 -k2,2n $file > ${file/.bed/.sorted.bed}
echo $file
done

# 2. Manually remove from mm9 regions, those that are missing in mm10 annotation
# For E14.15 up --> bed_gff_ID_up_DOCR_E14.15_mm9_sans_reg_perdu_mm10.sorted.txt
# For E16.P0 up --> bed_gff_ID_up_DOCR_E16.P0_mm9_sans_reg_perdu_mm10.sorted.txt

# 3. Combine mm9 and mm10 coordinates : first coordinate = mm9, second = mm10
cd /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10

sample="down_DOCR_E13.14 down_DOCR_E14.15 down_DOCR_E14.16 down_DOCR_E15.16 down_DOCR_E16.P0 up_DOCR_E
13.14 up_DOCR_E14.16 up_DOCR_E15.16"

for file in $sample
do
    echo "$file"
    paste /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/bed_gff_ID_${file}_mm9.sorted.tx
t bed_gff_ID_${file}_mm10.sorted.bed > mm9_mm10_correspondance_${file}.txt
done

sample_bis="up_DOCR_E14.15 up_DOCR_E16.P0"

for file in $sample_bis
do
    echo "$file"
    paste /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/bed_gff_ID_${file}_mm9_sans_reg_
perdu_mm10.sorted.txt bed_gff_ID_${file}_mm10.sorted.bed > mm9_mm10_correspondance_${file}.txt
done
```

Annotation of differentially open or closed regions during brain development

Syntaxic annotation of regions of interest was performed using **bedtools intersect** with an annotation file (biomart_mm10.txt) obtained with **BiomaRt** R package. This file contains :

- ensembl_gene_id : gene name
- chromosome_name : chromosome/scaffold number
- strand
- start_position : gene start (bp)
- end_position : gene end (bp)
- entrezgene_id :NCBI gene ID
- gene_biotype : gene type
- mgi_symbol : MGI symbol
- entrezgene_accession : NCBI gene accession
- entrezgene_description : NCBI gene description
- uniprot_gn_symbol : UniProtKB gene name symbol

In [ ]:

```
#-------------------------------------
# 1. Obtention of biomart_mm10.txt file
#-------------------------------------

# Load BiomaRt package :
library("biomaRt")

# 1. Select database and reference genome
mm10 = useMart("ensembl", dataset="mmusculus_gene_ensembl")
# mm10 Mus musculus version used:  Ensembl 97 Jul 2019 http://jul2019.archive.ensembl.org
```

```
# 2. Create the dataset with information of interest
# It's necessary to do it in two steps, because there is a limitation in the number of attributes that
can be collected at the same time.
annot_mm10_part1<-getBM(attributes=c("ensembl_gene_id","chromosome_name","strand", "start_position","en
d_position","entrezgene_id","gene_biotype","mgi_symbol","entrezgene_accession"), mart=mm10)
annot_mm10_part2<-getBM(attributes=c("ensembl_gene_id","entrezgene_description","uniprot_gn_symbol"), m
art=mm10)
annot_mm10 = merge(annot_mm10_part1, annot_mm10_part2, by="ensembl_gene_id", all=TRUE)

# Verifications
dim(annot_mm10)
head(annot_mm10)

# Sauvegarde
write.table(annot_mm10,"G:/Chip_pilote/data/mm10_genome/biomart_mm10.txt", quote= FALSE, sep="\t",row.n
ames=FALSE)




#--------------------------------------
# 2. Combine differentially open or closed regions with biomart annotation file
#--------------------------------------
# 2.1 Datasets formatting
setwd("G:/projet_DU_DSD/results/")
biomart_mm10 =read.table("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10/biomart_mm10_so
rted.txt",sep="\t", na.strings = "NA", fill=TRUE, quote="", header=TRUE)

head(biomart_mm10)

# Change "-1" and "1" strand encoding by "-" and "+"
for (i in 1:nrow(biomart_mm10)){
  if(biomart_mm10[i,"strand"]==1){
    biomart_mm10[i,"strand"]="+"
  }else if (biomart_mm10[i,"strand"]==-1){
    biomart_mm10[i,"strand"]="-"
  }else{
  }
}

# Change chromosome format
biomart_mm10$chr = with(biomart_mm10,paste("chr",biomart_mm10[,"chromosome_name"],sep=""))

head(biomart_mm10)

# To combine informations, we will used bedtools intersect. It seems to work with a .bed containing exa
ctly 9 columns.
# I removed uniprot_genename and entrezgene_description informations, because bedtools intersect doesn'
t work if I keep these columns.
# (bedtools intersect is not able to detect the file format if I keep these columns).

biomart_mm10_bis=biomart_mm10[,c("chr","start_position","end_position","strand","entrezgene_accession",
"gene_biotype", "mgi_symbol","ensembl_gene_id", "entrezgene_id")] #ça marche

write.table(biomart_mm10_bis,"G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10/biomart_mm1
0_all_info.bed", sep = "\t", col.names=TRUE, row.names=FALSE, quote=FALSE)
```

In [ ]:

```
%%bash

# 2.2 Sort files
# Sort Biomart annotation file
cd /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10/

sed '1d' biomart_mm10_all_info.bed | sort -k1,1 -k2,2n > biomart_mm10_all_info_sorted.bed

head biomart_mm10_all_info_sorted.bed


# Sort mm10 .bed files containing differentially open or closed regions coordinates
# Already done, I copy and paste files into  /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Bioma
rt/mm10/raw_data
```

```
#---------------------------------
# 3. Combine information using bedtools intersect
#---------------------------------

cd /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10/raw_data

sample="down_DOCR_E13.14 down_DOCR_E14.15 down_DOCR_E14.16 down_DOCR_E15.16 down_DOCR_E16.P0 up_DOCR_E
13.14  up_DOCR_E14.15 up_DOCR_E14.16 up_DOCR_E15.16 up_DOCR_E16.P0"

for file in $sample
do
    echo "$file"
    bedtools intersect -wao -a bed_gff_ID_${file}_mm10.sorted.bed -b /mnt/g/projet_DU_DSD/results/10_an
notations/Sartools/Biomart/mm10/biomart_mm10_all_info_sorted.bed > /mnt/g/projet_DU_DSD/results/10_anno
tations/Sartools/Biomart/mm10/${file}_annot_syntaxique_mm10_biomart.bed

    echo "fichier : $file annoté"
    wc -l /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10/${file}_annot_syntaxique_mm
10_biomart.bed
    head /mnt/g/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10/${file}_annot_syntaxique_mm1
0_biomart.bed
done
```

Combination of annotation and statistical informations

In order to obtain a file where each region of interest is represented by a single line, the annotation elements of a given region that are on separated lines have been grouped together. Then, all data (statistical informations, mm9 and mm10 coordinates regions, annotation informations) are combined in an single file (one file per pairwise comparison).

In [ ]:

```
# Load file : List file containing "AnnotBM" or "ID_sd" patterns
list_files = list.files("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10/")
list_files_Annot = list_files[which(grepl("annot_syntaxique",list_files))]
list_files_corresp = list_files[which(grepl("correspondance",list_files))]

list_files_bis = list.files("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/")
list_files_ID_sd = list_files_bis[which(grepl("gff_ID",list_files_bis))]


for(i in 1:10){
  Annot = list_files_Annot[i]
  DOCR_annot = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10",
Annot,sep="/"), header = FALSE, sep = "\t", quote="", fill=TRUE)

  #---------------------------------------------------------------------
  # A. Concatenate annotation informations : one row = one region of interest
  # 1. Create the ID = chr-start-end
  colnames(DOCR_annot) = c("chr", "start_DOCR","end_DOCR","chr_gene", "start_gene", "end_gene", "stran
d","entrezgene_accession",
                           "gene_biotype","mgi_symbol", "ensembl_gene_id", "entrezgene_id", "overlap_le
ngth")

  DOCR_annot$ID = with(DOCR_annot, paste(DOCR_annot[,"chr"], DOCR_annot[,"start_DOCR"],DOCR_annot[,"en
d_DOCR"],sep="-"))

  # 2. Concatenate informations
  library(dplyr)

  # Liste of unique ID
  liste_DOCR_unique <- unique(DOCR_annot$ID)

  # empty matrice which will be filled during the loop
  # Nb of rows = nb of unique ID | Nb of columns = nb of variables
  treatment_file <- data.frame(matrix(NA,nrow=length(liste_DOCR_unique),ncol=ncol(DOCR_annot)))
  colnames(treatment_file) <- colnames(DOCR_annot)

  treatment_file$ID <- liste_DOCR_unique
  for(k in 1:length(liste_DOCR_unique)){
    # Filter according to ID k
    DOCR_filt <- DOCR_annot %>%
      filter(ID == liste_DOCR_unique[k])
```

```
    # Collapsing of distincts elements of a variables for ID k
    # If variable is composed of only one element ==> we keep this element only once
    # If variable is composed of m distinct elements ==> evene if elements can be repeated, we keep all
the elements (with repetitions)
    for(j in 1:length(colnames(treatment_file))){
      if(length(unique(DOCR_filt[,j])) == 1){
        treatment_file[k,j] <- paste(unique(DOCR_filt[,j]), collapse=';')
      } else {
        treatment_file[k,j] <- paste(DOCR_filt[,j], collapse=';')
      }
    }
  }
  print(paste("nb of uniq ID treatment file : ", nrow(treatment_file), sep=""))



  #-------------------------------------------------------------------------
  # B. Merge treatment_file with mm9 and mm10 regions coordinates

  ID_corresp = list_files_corresp[i]
  corresp_file = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10
",ID_corresp,sep="/"), header = FALSE, sep = "\t", quote="", dec=".")

  colnames(corresp_file) = c("mm9_chr", "mm9_start", "mm9_end", "mm10_chr", "mm10_start", "mm10_end")
  corresp_file$ID = with(corresp_file, paste(corresp_file[,"mm10_chr"], corresp_file[,"mm10_start"], co
rresp_file[,"mm10_end"], sep="-"))
  corresp_file$mm9_ID = with(corresp_file, paste(corresp_file[,"mm9_chr"], corresp_file[,"mm9_start"],
corresp_file[,"mm9_end"], sep="-"))

  ATAC_corresp = merge(corresp_file,treatment_file, by="ID", all = TRUE)

  size_Annot = nrow(treatment_file)
  size_ATAC_corresp = nrow(ATAC_corresp)

  if(size_Annot != size_ATAC_corresp){
    print(paste("Les deux fichiers", Annot, "et", ID_corresp, "n'ont pas la meme taille !", sep =" "))
  }else{
  }

  #-------------------------------------------------------------------------
  # C. Merge ATAC_corresp with statistical informations obtained with SARTools-edgeR for every regions

  ID_sd = list_files_ID_sd[i]
  sartools_file = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart",ID
_sd,sep="/"), header = TRUE, sep = "\t", quote="", dec=".")

  colnames(sartools_file) = c("mm9_ID", "cond1_rep1", "cond1_rep2", "cond2_rep1", "cond2_rep2",
                              "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2
", "baseMean",
                              "cond1", "cond2", "FC", "log2FoldChange", "pvalue", "padj", "tagwise.disp
ersion", "trended.dispersion",
                              "mm9_chr", "source", "element_type", "mm9_start", "mm9_end", "V6", "stran
d", "V8", "V9")
  sartools_file$mm9_ID = with(sartools_file, paste(sartools_file[,"mm9_chr"], sartools_file[,"mm9_start
"], sartools_file[,"mm9_end"], sep="-"))

  Annot_DOCR = merge(ATAC_corresp,sartools_file, by="mm9_ID", all = TRUE)
  Annot_DOCR_clean = Annot_DOCR[,c("mm9_chr.x", "mm9_start.x", "mm9_end.x", "mm9_ID", "mm10_chr", "mm10
_start", "mm10_end", "ID",
                                   "chr_gene", "start_gene", "end_gene", "strand.x", "entrezgene_accession
", "gene_biotype", "mgi_symbol",
                                   "ensembl_gene_id", "entrezgene_id",  "overlap_length", "cond1_rep1", "con
d1_rep2", "cond2_rep1", "cond2_rep2",
                                   "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2
", "baseMean", "cond1", "cond2",
                                   "FC", "log2FoldChange", "pvalue", "padj", "tagwise.dispersion", "trended.
dispersion")]


  colnames(Annot_DOCR_clean) = c("mm9_chr", "mm9_start", "mm9_end", "mm9_ID", "mm10_chr", "mm10_start",
"mm10_end", "mm10_ID",
                                 "chr_gene", "start_gene", "end_gene", "strand", "entrezgene_accession"
, "gene_biotype", "mgi_symbol",
                                 "ensembl_gene_id", "entrezgene_id",  "overlap_length", "cond1_rep1", "
cond1_rep2", "cond2_rep1", "cond2_rep2",
                                 "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_r
```

```
ep2", "baseMean", "cond1", "cond2",
                                "FC", "log2FoldChange", "pvalue", "padj", "tagwise.dispersion", "trend
ed.dispersion")

  write.table(Annot_DOCR_clean, file=paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm
10/stats_mm9_mm10",Annot,sep="_"), quote=FALSE, sep="\t", row.names = FALSE, col.names = TRUE)

}
```

3.3. RNA-seq bioinformatic workflow

RNA-seq workflow - analysis of the physiological modification of gene expression levels throughout brain development

Agathe Duchateau

Download ENCODE data

To study modifications of physiological gene expression levels troughout brain developement, we used available RNA-seq ENCODE data that were performed on embryonic and newborn forebrains of C57BL/6N mice (expriments of B. Wold, Caltech laboratory). Developmental stages that were studied in replicates are :

- embryonic day **E13.5** (Encode accession : ENCFF235DNM,ENCFF959PSX)
- embryonic day **E14.5** (Encode accession : ENCFF270GKY + ENCFF460TCF, ENCFF126IRS + ENCFF748SRJ)
- embryonic day **E15.5** (Encode accession : ENCFF179JEC, ENCFF891HIX)
- embryonic day **E16.5** (Encode accession : ENCFF931IVO, ENCFF114DRT)
- and postnatal day 0 (**P0**, Encode accession : ENCFF037JQC + ENCFF358MFI, ENCFF447EXU + ENCFF458NWF).

These data were obtained with a non strand specificity poly RNA-seq, from oligo-dT primed total RNA. In brief, tissues were lysed and RNA was extracted using Ambion mirVana method. Fragmentation was performed by tagmentation (Nextera) and size selection was done using SPRI beads. For each cDNA sample, fragments of 100 nucleotides were single-ended sequenced on a Illumina HiSeq 2500 platform. For more explanations on experimental procedures, please see *Experiment Summary* on ENCODE database, pages avalaiables with Encode accession of samples cited above.

In [ ]:

```
%%bash

# Download ENCODE .fastq.gz files

cd /data/omics-school/aduchateau/DU_projet/data

# For E16.5 first replicate
wget https://www.encodeproject.org/files/ENCFF931IVO/@@download/ENCFF931IVO.fastq.gz
# For E16.5 second replicate
wget https://www.encodeproject.org/files/ENCFF114DRT/@@download/ENCFF114DRT.fastq.gz
# For E15.5 first replicate
wget https://www.encodeproject.org/files/ENCFF179JEC/@@download/ENCFF179JEC.fastq.gz
# For E15.5 second replicate
wget https://www.encodeproject.org/files/ENCFF891HIX/@@download/ENCFF891HIX.fastq.gz
# For E14.5 first replicate
wget https://www.encodeproject.org/files/ENCFF270GKY/@@download/ENCFF270GKY.fastq.gz
# For E14.5 first replicate
wget https://www.encodeproject.org/files/ENCFF460TCF/@@download/ENCFF460TCF.fastq.gz
# For E14.5 second replicate
wget https://www.encodeproject.org/files/ENCFF126IRS/@@download/ENCFF126IRS.fastq.gz
# For E14.5 second replicate
wget https://www.encodeproject.org/files/ENCFF748SRJ/@@download/ENCFF748SRJ.fastq.gz
# For E13.5 first replicate
wget https://www.encodeproject.org/files/ENCFF235DNM/@@download/ENCFF235DNM.fastq.gz
# For E13.5 second replicate
wget https://www.encodeproject.org/files/ENCFF959PSX/@@download/ENCFF959PSX.fastq.gz
# For P0 first replicate
wget https://www.encodeproject.org/files/ENCFF358MFI/@@download/ENCFF358MFI.fastq.gz
# For P0 first replicate
wget https://www.encodeproject.org/files/ENCFF037JQC/@@download/ENCFF037JQC.fastq.gz
# For P0 first replicate
wget https://www.encodeproject.org/files/ENCFF458NWF/@@download/ENCFF458NWF.fastq.gz
# For P0 second replicate
wget https://www.encodeproject.org/files/ENCFF447EXU/@@download/ENCFF447EXU.fastq.gz

# MD5sum check
# File containing MD5sum of .fastq.gz files
cat md5sum.txt :
# f12d3f32677ecc0208756bf9e05f990b ENCFF931IVO.fastq.gz
# e9b9aae34a242deb0e8ec4b39979fe8c ENCFF114DRT.fastq.gz
# befdcab4ff5b2bbdc9717064c55c02ef ENCFF179JEC.fastq.gz
# f2a836c3ca6b515f3ed60488be900244 ENCFF891HIX.fastq.gz
# 1f6d6da942e693019de92fe5ada06f2f ENCFF270GKY.fastq.gz
```

```
# 97810752f1592e4114ea8c713596a06 ENCFF460TCF.fastq.gz
# ba5e9274ddcf96218d5c152624c73cb3 ENCFF126IRS.fastq.gz
# e8071390598286da2a5c273629a24789 ENCFF748SRJ.fastq.gz
# 66a89656a63f96311f64c7a28cf9102a ENCFF235DNM.fastq.gz
# 1707d6213d80ba3ea18e323dca07eb8d ENCFF959PSX.fastq.gz
# e5f5ef9f88ef582526cf1a54023f5ad0 ENCFF037JQC.fastq.gz
# 1d708f7b64c8b98bb9855453d4e6f6d4 ENCFF358MFI.fastq.gz
# 5a07748fe5a29c6b4fad1ddc850df0a5 ENCFF447EXU.fastq.gz
# 35807543b590dc9a98135aea49a20145 ENCFF458NWF.fastq.gz

md5sum -c md5sum.txt
# --> OK for all samples
```

Setting up of bioinformatic tools in a conda environment

Following tools were installed in a conda environment to perform analysis of RNA-seq from embryonic and newborn mice forebrains, obtained in physiological conditions. Version of each tool is indicated. Since .fasta files are very large, indexation of the reference genome and mapping were done on a distant server (cluster/High Performance Computer, from Institut Français de Bioinformatique - IFB). R packages which are cited below were also used to analyse the data.

In [ ]:

```
# Setting up of conda environment on a distant server (Ubuntu 16.04.5 LTS (GNU/Linux 4.4.0-131-generic
x86_64))
conda create -n rnaseq
conda --version # version 4.5.8

conda activate rnaseq

conda install -y fastqc  #version 0.11.7
conda install -y trim-galore #version 0.5.0
conda install -y cutadapt #version 1.16

# Mapping on a cluster from Institut Français de Bioinformatique (IFB)
STAR  #version 2.6.1d

# Setting up of another conda environment on local computer
conda create -n DU_projet
conda --version # version 4.6.7
conda install -y htseq #version 0.11.2
conda install -y samtools #version 1.9

# R version 3.5.2
# R package used for analysis :
library(edgeR) # version 3.24.3
library(ggplot2) # version 3.2.1
library(reshape2) # version1.4.3
library(GGally) # version 1.4.0
library(SARTools) # version 1.6.9
library(BiomaRt) #version 2.38.0
library(dplyr) #version 0.8.3
library(data.table) # version 1.12.2
library(stringr) # version 1.4.0
```

Quality control of the raw data

After simplifying .fastq.gz file names, **FASTQC** was used to verify the quality of the sequencing raw data.

In [ ]:

```
%%bash

# To rename files
cd /data/omics-school/aduchateau/DU_projet/data
cp ENCFF931IVO.fastq.gz ../results/01_raw_data/E16.5_01.fastq.gz
cp ENCFF114DRT.fastq.gz ../results/01_raw_data/E16.5_02.fastq.gz
cp ENCFF179JEC.fastq.gz ../results/01_raw_data/E15.5_01.fastq.gz
cp ENCFF891HIX.fastq.gz ../results/01_raw_data/E15.5_02.fastq.gz
cp ENCFF270GKY.fastq.gz ../results/01_raw_data/E14.5_01.1.fastq.gz
cp ENCFF460TCF.fastq.gz ../results/01_raw_data/E14.5_01.2.fastq.gz
cp ENCFF126IRS.fastq.gz ../results/01_raw_data/E14.5_02.1.fastq.gz
cp ENCFF748SRJ.fastq.gz ../results/01_raw_data/E14.5_02.2.fastq.gz
```

```
cp ENCFF235DNM.fastq.gz ../results/01_raw_data/E13.5_01.fastq.gz
cp ENCFF959PSX.fastq.gz ../results/01_raw_data/E13.5_02.fastq.gz
cp ENCFF037JQC.fastq.gz ../results/01_raw_data/E00.5_01.1.fastq.gz
cp ENCFF358MFI.fastq.gz ../results/01_raw_data/E00.5_01.2fastq.gz
cp ENCFF447EXU.fastq.gz ../results/01_raw_data/E00.5_02.1.fastq.gz
cp ENCFF458NWF.fastq.gz ../results/01_raw_data/E00.5_02.2fastq.gz


# Quality control of the raw data

mkdir 02_FASTQC
cd ~/DU_projet/results/02_FASTQC

conda activate rnaseq

for file in ../01_raw_data/E*
do
fastqc $file
done

# To move files into appropriate folder
cd ~/DU_projet/results/01_raw_data

for file in *html
do
mv $file ../02_FASTQC/$file
done

for file in *zip
do
mv $file ../02_FASTQC/$file
done
```

Trimming

According to FASTQC reports, we decided to filter the data to remove remaining Nextera Transposase adapters, by using **Trim-galore**. The 15 first bases of all reads were also removed, because of their poor qualities. This trimming was sufficient for the majority of the samples. However some samples (first replicate of E14.5 stage + second replicate of P0 stage), having an over-representation of polyT or primers sequences (Clontech Universal Primer Mix Long), were trimmed again to remove these sequences.

For E14 and P0 replicates, the two files for a given replicate were concatenated before trimming, using **cat unix command**. FASTQC controls were also performed on these concatenate files but reports were similar than those of separated files.

Details of parameters used for the trimming:

- --nextera : to remove Nextera adapters sequences
- -q 20 : to trim low-quality ends from reads in addition to adapter removal. Default value : 20
- --clip_R1 15 : to remove 15 bases in 5' position
- --stringency 5 : overlap with adapter sequence required to trim a sequence. Default value : 1
- --length 20 : to remove reads that are shorter than 20 bases. Default value : 1
- -fastqc : to run a FASTQC report after the trimming

In [ ]:

```
%%bash

# ------------------------------------------------
# To concatenate E14.5 and P0 files of a given replicate
# ------------------------------------------------

cd ~/DU_projet/results/01_raw_data

cat E00.5_01.1.fastq.gz E00.5_01.2fastq.gz > E00.5_01.fastq.gz
cat E00.5_02.1.fastq.gz E00.5_02.2fastq.gz > E00.5_02.fastq.gz
cat E14.5_01.1.fastq.gz E14.5_01.2.fastq.gz > E14.5_01.fastq.gz
cat E14.5_02.1.fastq.gz E14.5_02.2.fastq.gz > E14.5_02.fastq.gz


# To run FASTQC on concatenate files
for file in E00.5_01.fastq.gz E00.5_02.fastq.gz E14.5_01.fastq.gz E14.5_02.fastq.gz
```

```
do
    fastqc $file --outdir ~/DU_projet/results/02_FASTQC/
done

# ------------------------------------------------
# Trimming using trimming.bash script
# First Round of trimming
# ------------------------------------------------

# Script that was runned for trimming:
# Code will stop running if any problem is detected
# (first error, undefined variable, pipe error)
set -euo pipefail

# number of samples to analyse
sample="00.5_01 00.5_02 13.5_01 13.5_02 14.5_01 14.5_02 15.5_01 15.5_02 16.5_01 16.5_02"

for sample in ${sample}
do
    echo "================================================================="
    echo "sample number : ${sample}"
    echo "================================================================="

# Trimming
    echo "================================================================="
    echo "Trimming de l'échantillon ${sample}"
    echo "================================================================="

# To use Trim-galore to remove first bases, reads having bad quality score,
# adapters sequences, too short reads...

    trim_galore --nextera -q 20 --phred33 --clip_R1 15 --stringency 5 --length 20
    -fastqc --output_dir ~/DU_projet/results/03_trimming ~/DU_projet/results/01_raw_data/E${sample}.fas
tq.gz
    > tg_E${sample}.fastq.gz

done

# ------------------------------------------------
# Second round of trimming for E14.5 (first replicate) P0 (second replicate)
# ------------------------------------------------

mkdir -p ~/DU_projet/results/03_trimming/pass1/
for file in E14.5_01*
do
mv $file ~/Du_projet/results/03_trimming/pass1/
done

for file in E00.5_02*
do
mv $file ~/Du_projet/results/03_trimming/pass1/
done

# To remove polyT and Clontech Universal Primer Mix Long sequences
trim_galore -a TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT -q 20 --phred33
--stringency 40 --length 20 -fastqc
--output_dir ~/DU_projet/results/03_trimming ~/DU_projet/results/03_trimming/pass1/E00.5_02_trimmed.fq.
gz

trim_galore -a GTATCAACGCAGAGTACGGGAAGCAGTGGTATCAACGCAGAGTACGGGAA  -q 20 --phred33
--stringency 40 --length 20 -fastqc
--output_dir ~/DU_projet/results/03_trimming ~/DU_projet/results/03_trimming/pass1/E14.5_01_trimmed.fq.
gz

# --------------------
# To save files
# --------------------

for file in *.gz
do
    cp $file ~/DU_projet/results/04_mapping/input/tg_$file
done

cd ~/DU_projet/results/04_mapping/input/
cp ~/DU_projet/results/03_trimming/E00.5_02_trimmed_trimmed.fq.gz ~/DU_projet/results/04_mapping/input/
tg_E00.5_02_trimmed.fq.gz
cp ~/DU_projet/results/03_trimming/E14.5_01_trimmed_trimmed.fq.gz ~/DU_projet/results/04_mapping/input/
```

```
tg_E14.5_01_trimmed.fq.gz
```

Mapping

Since fasta files are very large, indexation of the reference genome and mapping were done on a distant server (cluster/High Performance Computer, from Institut Français de Bioinformatique - IFB), with the help of O. Kirsh (Paris University, CNRS Epigenetics and Cell Fate, Paris, France).

Indexation of mm10 reference genome

Mapping was done using **STAR**, on mm10 Mus musculus reference genome, which was already available on the IFB cluster. When mm10 reference genome was indexed, we specified an annotation file (gencode.vM21.annotation.gtf), to improve accuracy of the mapping, as recommended in STAR user manual. This annotation file was downloaded on Gencode database (in Release M21 (GRCm38.p6) - Comprehensive gene annotation - CHR Regions, containing the comprehensive gene annotation on the reference chromosomes only).

Details of parameters used for the indexation:

- --runMode genomeGenerate : mode to index reference genome
- --runThreadN : number of threads to be used for genome indexation
- --genomeDir : to specify path to the directory where the genome indices are stored. The file system needs to have at least 100GB of disk.
- --genomeFastaFiles : to specify fasta file with the genome reference sequences.
- --sjdbGTFfile : to specify the path to the file with annotation in the standard GTF format. STAR will extract splice junctions from this file and use them to greatly improve accuracy of the mapping. While this is optional, and STAR can be run without annotations, using annotations is highly recommended whenever they are available.
- --sjdbGTFfeatureExon : to specify feature type in .gtf file that have to be used as exons for building transcripts
- --sjdbOverhang : to specify the length of the genomic sequence around the annotated junction to be used in constructing the splice junctions database. Ideally, this length should be equal to the ReadLength-1, where ReadLength is the length of the reads.

In [ ]:

```
%%bash

cd /Volumes/Maxtor/projet_DU_AD/data/mm10_genome/

# -------------------------------------
# Obtention of Gencode annotation file
# -------------------------------------

# Download Gencode Annotation file
wget ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M21/gencode.vM21.annotation.gtf.gz

md5sum gencode.vM21.annotation.gtf.gz
#d08f66b2746d0ae66594fda6ea0c9939   gencode.vM21.annotation.gtf.gz
md5sum gencode.vM21.annotation.gtf
#699f4d30a09157711a3e71df018e52f5   gencode.vM21.annotation.gtf
```

In [ ]:

```
# Copy of this annotation file on IFB cluster

# -------------------------------------
# Genome indexation
# -------------------------------------
# Script that was runned for genome indexation:

#!/bin/bash
#
#SBATCH -o slurm.%N.%j.out
#SBATCH -e slurm.%N.%j.err
#SBATCH --mail-type END
#SBATCH --mail-user olivier.kirsh@univ-paris-diderot.fr
#SBATCH --partition long
#SBATCH --cpus-per-task 12
#SBATCH --mem 40GB

module load conda
```

```
source  activate star-2.6

STAR --runMode genomeGenerate --runThreadN $SLURM_CPUS_PER_TASK
--genomeDir /shared/home/okirsh/mapping/mm10_indexation
--genomeFastaFiles /shared/bank/mus_musculus/mm10/fasta/mm10.fa
--sjdbGTFfile /shared/home/okirsh/mapping/mm10_genome/gencode.vM21.annotation.gtf
--sjdbGTFfeatureExon exon --sjdbOverhang 84

source activate
```

Mapping

Mapping of each sample was performed using **STAR** with the following parameters (see below). Mapping was relatively stringent, since each read can map at only 1 loci, otherwise it is not keep for the analysis (see --outFilterMultimapNmax option).

nb: For the first replicate of embryonic stage E14.5, we lost informations. Indeed, STAR detected not all the reads of this sample (47 515 325 reads were detected during the mapping step, whereas 50 697 328 reads were in the sample, according to FASTQC report). To verify if file was corrupted during transfert, we checked MD5sum of this file, but it was right. Because fasta file of this sample - which contains reads - are unsorted, we supposed that lost of reads was unbiased, because reads that were lost were randomly 'selected'. Thus, we decided to keep this sample for the rest of the analysis.

Detail of parameters used for mapping:

- --runMode alignReads : mode to map sample
- --runThreadN 6 : number of threads to run STAR
- --genomeDir : path containing indexed reference genome
- --genomeLoad NoSharedMemory : do not use shared memory, each job will have its own private copy of the genome
- --readFilesType Fastx : to notify that input file is in FASTA format
- --readFilesIn : path to input file
- --readFilesCommand gunzip -c : command line to execute for each of the input file. Here, we uncompressed input file using gunzip without deleting unzip file.
- --outSAMunmapped Within : to obtain a file containing unmapped reads
- --outReadsUnmapped Fastx : output of unmapped and partially mapped (*i.e.* mapped only one mate of a paired end read) reads in separate fasta/fastq file(s), called Unmapped.out.mate1/2.
- --outSAMtype BAM Unsorted : ouput file will be an unsorted .bam file
- --outSAMattributes Standard : string of desired SAM attributes in the order desired for the output. Standard =Standard NH HI AS nM.

nb: it's impossible to use --outSAMattributes All option (All = NH HI AS nM NM MD jM jI) because file processing fails using htseq-count afterthat.

- --outFilterType BySJout : type of filtering. Using BYSJout, we keep only those reads that contain junctions that passed filtering into SJ.out.tab
- --outFilterMultimapNmax 1 : maximum number of loci the read is allowed to map to. Alignments (all of them) will be output only if the read maps to no more loci than this value. Otherwise no alignments will be output, and the read will be counted as "mapped to too many loci" in the Log.final.out. Default: 10
- --outFilterMismatchNmax 10 : alignment will be output only if it has no more mismatches than this value. Default: 10
- --outFilterMismatchNoverLmax 0.3 : alignment will be output only if its ratio of mismatches to mapped length is less than or equal to this value. Default: 0.3
- --quantMode TranscriptomeSAM GeneCounts : types of quantification requested. Using TranscriptomeSAM : output SAM/BAM alignments to transcriptome into a separate file called Aligned.toTranscriptome.out.bam. Using GeneCounts : count reads per gene. A read is counted if it overlaps (1nt or more) one and only one gene. Both ends of the paired-end read are checked for overlaps. The counts coincide with those produced by htseq-count with default parameters. Both options can be used together.
- --twopassMode None : 1-pass mapping
- --outFileNamePrefix : to specify prefix name of output file.

In [ ]:

```
#-------------------------------------
#B. Mapping (using strict parameters)
#-------------------------------------

#Script that was runned for mapping:
#!/bin/bash

for file in *.gz
```

```
do
    STAR --runMode alignReads --runThreadN 6
    --genomeDir /home/olivier/Bureau/agathe/mm10_indexation --genomeLoad NoSharedMemory
    --readFilesType Fastx --readFilesIn /home/olivier/Bureau/agathe/input/${file}
    --readFilesCommand gunzip -c --outReadsUnmapped Fastx --outSAMtype BAM Unsorted
    --outSAMattributes Standard --outSAMunmapped Within --outFilterType BySJout
    --outFilterMultimapNmax 1 --outFilterMismatchNmax 10
    --outFilterMismatchNoverLmax 0.3 --quantMode TranscriptomeSAM GeneCounts
    --twopassMode None
    --outFileNamePrefix /home/olivier/Bureau/agathe/outputlocal/strict_mm10_Std_${file%%.fq.gz}_
done

#------------------ End of script ---------------------------------


# Mapping output files were copied on the local computer in the following folder:
/Volumes/Maxtor/projet_DU_AD/results/04_mapping/output_strict_Std_mm10/


# To display mapping report (containing statistical informations):
cd /Volumes/Maxtor/projet_DU_AD/results/04_mapping/output_strict_Std_mm10/

for file in *Log.final.out
do
    echo $file
    cat $file
done
```

Creation of count tables

htseq-count was used to generate tables containing read count of each sample, for each gene (see corresponding code after *Mapping visualization section*).

Parameter Details:

- --stranded=no : to specify that RNA-seq data has not been made with a strand-specific protocol.
- --format bam : to specify input file format
- -a 10 : to skip all reads with alignment quality lower than the given minimum value. Default: 10
- --type=exon : to specify feature type (*i.e.* 3rd column of .gtf file) that will be used, all features of other type are ignored. Default: exon
- --idattr=gene_id : to specify the .gff attribute that will be used as feature ID. Several .gff lines with the same feature ID will be considered as parts of the same feature. The feature ID is used to identity the counts in the output table. Default: gene_id.
- --mode=union : to specify mode that will be used to handle reads overlapping more than one feature. With union mode, a read is count if there is a overlap with a given region, even if the overlap is partial, as long as there is no ambiguous case (such as overlap with two distinct regions). Default: union.
- --nonunique=none : reads that mapped several genes are skipped.

Mapping vizualization

To visualize mapping data with **IGV**, mapping output files (*i.e.* .bam files) were sorted and indexed using **samtools sort** and **samtools index**.

In [ ]:

```
%%bash

# On local computer
conda activate DU_projet

cd /mnt/c/Users/Agathe/Desktop/manip_en_cours/projet_DU_remapping

bash script_RNAseq_AD_mm10_htseq_indexIGV.bash

# Script that was runned to generate count tables + .bam indexation (for visualization):

# Code will stop running if any problem is detected
# (first error, undefined variable, pipe error)
set -euo pipefail

# number of samples to analyse
ech="00.5_01 00.5_02 13.5_01 13.5_02 14.5_01 14.5_02 15.5_01 15.5_02 16.5_01 16.5_02"
```

```
#-----------------------------------------------------
# Obtention of count tables using htseq-count
#-----------------------------------------------------

mkdir -p /mnt/g/projet_DU_AD/results/05_count_table/output_strict_mm10_Std/

for sample in ${ech}
do
    echo "========================================================================="
    echo "Obtention de la table de comptage de l'échantillon ${sample}-analyse stricte"
    echo "========================================================================="

    htseq-count --stranded=no --format bam -a=10 --type=exon --idattr=gene_id
    --mode=union --nonunique=none
    /mnt/g/projet_DU_AD/results/04_mapping/output_strict_Std_mm10/strict_mm10_Std_tg_E${sample}_trimmed
_Aligned.out.bam
    /mnt/g/projet_DU_AD/data/mm10_genome/gencode.vM21.annotation.gtf >
    /mnt/g/projet_DU_AD/results/05_count_table/output_strict_mm10_Std/c_strict_Std_tg_E${sample}_trimme
d_mm10.sorted.bam
done

#-----------------------------------------------
# .bam files indexation for visualization
#-----------------------------------------------

cd /mnt/g/projet_DU_AD/results/04_mapping/output_strict_Std_mm10/

echo "============================================================="
echo "Indexation of sorted bam files for visualisation"
echo "============================================================="

for sample in ${ech}
do
    echo "============================================================="
    echo "sample number : ${sample}"
    echo "============================================================="
    samtools sort strict_mm10_Std_tg_E${sample}_trimmed_Aligned.out.bam >
    sorted_strict_mm10_Std_tg_E${sample}_trimmed_Aligned.out.bam

    samtools index sorted_strict_mm10_Std_tg_E${sample}_trimmed_Aligned.out.bam > sorted_strict_mm10_St
d_tg_E${sample}_trimmed_Aligned.out.bai
done
```

Identification of differentially expressed genes (DEG) during physiologigal cortex development

To identify differentially expressed genes (**DEG**) during brain development, pairwise comparisons of successive developmental stages were performed using **edgeR**, an **R** software and **Bioconductor** package.

We used (with adaptations) **Sartools** R package, which implementes some **edgeR** functions, and its script template template_script_edgeR.r. Main steps of the statistical analysis are described below, following by the executing code.

Filtering very-low count regions

Since regions with very low count across all samples could interfere with the statistical approximations of **edgeR**, these regions were removed using SARTools filter, *i.e.* by selecting regions which contain at least minReplicates (smallest number of replicates = 2) with at least counts per million cutoff of 1 (cpmCutoff = 1).

Normalization to avoid regions composition biases

To be able to compare read counts between samples, normalization was carried out according to the **edgeR** package methodology. Library of each sample was normalised for regions composition using **calcNormFactors()** function, based on a trimmed mean of M-values (TMM) between each pair of samples.
Boxplots of raw- and normalised counts distribution were compared to verify the quality of the normalization process. Normalization is supposed to stabilize distributions across samples.

Statistical test to identify DEG between two successive developmental stages

Statistical test

Statistical test

Differential analysis was carried out according to the **edgeR** model. edgeR aims at fitting one linear model per region. The edgeR model assumes that the count data follow a negative binomial distribution which is a suitable way to analysis the data when variance is higher than the mean. First step of the statistical procedure is to estimate the dispersion of the data. After estimation of dispersions - *i.e.* common (unique value), trended (estimated with splines) and tagwise (estimated from feature counts) dispersions - fitting to a generalized linear model (glm) can be done using **glmFit()** function. Then, statistical test was performed for all pairwise comparisons of successive developmental stages, using a likehood ratio test (lrt method).
Histogram of raw pvalues from the statistical test was plotted to verify the shape of the distribution, that is expected to follow an uniform distribution, with a peak around 0. Benjamini-Hochberg method was applied to adjust pvalues computed by the statistical test in order to take into account multiple testing and control the false positive rate (FDR). Threshold of statistical significance was set to 0.05.

Differential analysis plots

MA-plot

For each comparison, MA-plot of the data was done. It represents the log ratio of differential expression as a function of the mean intensity for each region.

Volcano plot

Volcano plots were performed for each comparison. It represents the log of the adjusted pvalue as a function of the log ratio of differential expression.

Description plots to estimate variability within the experiment

Scatterplot

A pairwise scatterplot was produced to verify that intra-group (replicates) similarities are higher than inter-group (samples from distinct developmental stage) ones. This plot is obtained using log2(counts+1), instead of raw count values. This pairwise scatterplot is associated with a SERE statistic, that was used as a similarity index between RNA-seq samples.

Hierarchical clustering and dendrogram

Hierarchical clustering and dendrogram were performed to estimate variability between samples, *i.e.* see if samples from distinct developmental stages are seperated, while replicates are close to each other. Hierarchical clustering was done after a transformation of the count data as moderated log-counts-per-million.
To plot dendrogram, an euclidean distance is computed between samples. Dendrogram is then obtained from CPM data and built upon the Ward criterion.

Multidimensional scaling plot (MDS)

First two dimensions of a multidimensional scaling plot were plotted to visualise experiment variability. If biological variability is the main source of variance in the data, first dimension is expected to separate samples from the different developmental stages.

Adaptations of SARTools functions

Following parameters were modified or added in the original Sartools script, to obtain others graphs or informations:

- adding a filter to remove uncovered reads. counts_nf --> counts. In theory, this step is not necessary because uncovered regions will be removed during the filter initially performed by SARTools (it deletes regions with very low-count). However, it allows to have a double control of this step.
- adding an unormalised library size histogram graph plot (using ggplot2).
- changing the raw pvalues histogram design (using ggplot2): modifications of summarizeResults.edgeR() function --> summarizeResults.edgeR.AD()
- adding supplementary graphs in run.edgeR() function
- export supplementary informations (logFC, FDR)
- adding a filter using logFC and padj thresholds to obtain a restricted list of DEG, in addition to the list given by SARTools script (which is based on padj threshold only).

In [ ]:

```
```

In [ ]:

```
# ------------------------------------------------------------
# 1. Creation of files called targetE13.E14.txt,
# targetE14.E15.txt, targetE15.E16.txt and targetE16.P0.txt,
# which contain samples informations for loading count tables:
# ------------------------------------------------------------

#---------------
#For E13 vs E14
#---------------

target_cible = matrix(0,ncol=4, nrow=4)
colnames(target_cible)=c("label","files","group","day")

# Name of samples
target_cible[,"label"]= c("E13.5_A","E13.5_B","E14.5_A", "E14.5_B")

# Name of files containing count tables
target_cible[,"files"]=c("c_strict_Std_tg_E13.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E13.5_02_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E14.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E14.5_02_trimmed_mm10.sorted.txt")

# Embryonic days informations
target_cible[,"group"]=c(rep("E13",2), rep("E14",2))

# Day of experiment
# to correct eventual batch effect.
# Since we don't know when samples where prepared and if they were prepared at the same day, "d1" is wr
itten by default, for all samples.
target_cible[,"day"]="d1"

write.table(target_cible, "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE13.14.txt",
            quote=FALSE, col.names=TRUE,row.names=FALSE, sep="\t")


#---------------
#For E14 vs E15
#---------------

target_cible = matrix(0,ncol=4, nrow=4)
colnames(target_cible)=c("label","files","group","day")

# Name of samples
target_cible[,"label"]= c("E14.5_A","E14.5_B","E15.5_A", "E15.5_B")


# Name of files containing count tables
target_cible[,"files"]=c("c_strict_Std_tg_E14.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E14.5_02_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E15.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E15.5_02_trimmed_mm10.sorted.txt")


# Embryonic days informations
target_cible[,"group"]=c(rep("E14",2), rep("E15",2))

# Day of experiment
# to correct eventual batch effect.
# Since we don't know when samples where prepared and if they were prepared at the same day, "d1" is wr
itten by default, for all samples.
target_cible[,"day"]="d1"

write.table(target_cible, "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE14.15.txt",
            quote=FALSE, col.names=TRUE,row.names=FALSE, sep="\t")


#---------------
#For E15 vs E16
#---------------

target_cible = matrix(0,ncol=4, nrow=4)
colnames(target_cible)=c("label","files","group","day")

# Name of samples
target_cible[,"label"]= c("E15.5_A","E15.5_B","E16.5_A", "E16.5_B")

# Name of files containing count tables
```

```
target_cible[,"files"]=c("c_strict_Std_tg_E15.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E15.5_02_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E16.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E16.5_02_trimmed_mm10.sorted.txt")

# Embryonic days informations
target_cible[,"group"]=c(rep("E15",2), rep("E16",2))

# Day of experiment
# to correct eventual batch effect.
# Since we don't know when samples where prepared and if they were prepared at the same day, "d1" is wr
itten by default, for all samples.
target_cible[,"day"]="d1"

write.table(target_cible, "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE15.16.txt",
            quote=FALSE, col.names=TRUE,row.names=FALSE, sep="\t")


#---------------
#For E16 vs P0
#---------------

target_cible = matrix(0,ncol=4, nrow=4)
colnames(target_cible)=c("label","files","group","day")

# Name of samples
target_cible[,"label"]= c("E16.5_A","E16.5_B","P0_A", "P0_B")

# Name of files containing count tables
target_cible[,"files"]=c("c_strict_Std_tg_E16.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E16.5_02_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E00.5_01_trimmed_mm10.sorted.txt",
                         "c_strict_Std_tg_E00.5_02_trimmed_mm10.sorted.txt")

# Embryonic days informations
target_cible[,"group"]=c(rep("E16",2), rep("P0",2))

# Day of experiment
# to correct eventual batch effect.
# Since we don't know when samples where prepared and if they were prepared at the same day, "d1" is wr
itten by default, for all samples.
target_cible[,"day"]="d1"

write.table(target_cible, "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE16.P0.txt",
            quote=FALSE, col.names=TRUE,row.names=FALSE, sep="\t")




# ------------------------------------------------------------------------
# 2. Identification of DEG between two developmental stages,
# in physiological development.
# edgeR analysis using modified SARTools script
# ------------------------------------------------------------------------


# ----------------------
# -------------------------------------------
# E13 vs E14 developmental stages comparison
# -------------------------------------------
# ----------------------

dir.create("G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E13.14", recursive=TRUE)

# -------------------
# Parameters setting
# -------------------

rm(list=ls())

workDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E13.14"
# working directory for the R session

projectName <- "Sartools-edgeR-RNAseq-E13.14"        # name of the project
author <- "Agathe D."                                # author of the statistical analysis/report
```

```
targetFile <- "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE13.14.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",        # names of the features to be removed
                      "ambiguous", "no_feature",     # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                    # factor of interest
condRef <- "E13"                                     # reference biological condition
batch <- NULL                                        # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                        # threshold of statistical significance
pAdjustMethod <- "BH"                                # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                       # counts-per-million cut-off to filter low counts
gene.selection <- "pairwise"                         # selection of the features in MDSPlot
normalizationMethod <- "TMM"                         # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE

nsamples = 4                                         # nb of samples
nstage = 2                                           # nb of developemental stages
nrep = 2                                             # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")


# -------------------
# Load packages
# -------------------
setwd(workDir)
library(SARTools)
library(reshape2)
library(ggplot2)
library(GGally)

#--------------------
# Running script
# checking parameters (unchanged SARTools function)
#--------------------

if (forceCairoGraph) options(bitmapType="cairo")

checkParameters.edgeR(projectName=projectName,author=author,targetFile=targetFile,
                      rawDir=rawDir,featuresToRemove=featuresToRemove,varInt=varInt,
                      condRef=condRef,batch=batch,alpha=alpha,
                      pAdjustMethod=pAdjustMethod,
                      cpmCutoff=cpmCutoff,gene.selection=gene.selection,
                      normalizationMethod=normalizationMethod,colors=colors)

#--------------------
# loading target file (unchanged SARTools function)
#--------------------
target <- loadTargetFile(targetFile=targetFile, varInt=varInt, condRef=condRef,
                         batch=batch)

#-------------------
# loading count-table
#-------------------
counts_nf <- loadCountData(target=target, rawDir=rawDir,
                           featuresToRemove=featuresToRemove)

# Delete rows without counts :
counts <- counts_nf[-(which(rowSums(counts_nf) < 1)),]

#------------------
# description plots
#------------------
# (unchanged SARTools function)
majSequences <- descriptionPlots(counts=counts, group=target[,varInt], col=colors)
```

```
# Supplementary graphs
# Print library size (unormalised data) :
counts_reshape = melt(counts, id.vars = colnames(counts), variable.name='samples_name')
head(counts_reshape)
dim(counts_reshape)
colnames(counts_reshape) = c("peak_name","samples_name","filter_count")
counts_reshape$stages = with(counts_reshape, rep(substr(colnames(counts), 1, 3),
                                                 each = nrow(counts)))

lib_size_unorm = ggplot(counts_reshape, aes(x=samples_name,y=filter_count,
                                            fill = stages))

lib_size_unorm + geom_bar(stat="identity") +
    scale_fill_manual(values=c("skyblue","royalblue", "pink","tomato2","darkgreen")) +
  ggtitle("library size (unormalised dataset)")+ theme_light() +
  scale_x_discrete(name ="samples_name") + scale_y_continuous("total number of reads") +
  theme(plot.title = element_text(hjust = 0.5))

ggsave(path ="figures/", filename = "01 library size - unormalised dataset.pdf")
ggsave(path ="figures/", filename = "01 library size - unormalised dataset.png")


#---------------
# edgeR analysis (adapted from SARTools function)
#---------------
run.edgeR <- function(counts, target, varInt, condRef, batch=NULL, cpmCutoff=1,
                      normalizationMethod="TMM", pAdjustMethod="BH", ...){
  # filtering very low-count regions :
  # select features which contain at least minReplicates
  # (smallest number of replicates) with at least cpmCutoff counts per million
  minReplicates <- min(table(target[,varInt]))
  fcounts <- counts[rowSums(cpm(counts) >= cpmCutoff) >= minReplicates,]
  cat("Number of features discarded by the filtering:\n")
  cat(nrow(counts)-nrow(fcounts),"\n")

  # building dge object
  design <- formula(paste("~", ifelse(!is.null(batch), paste(batch,"+"), ""), varInt))
  dge <- DGEList(counts=fcounts, remove.zeros=TRUE)
  dge$design <- model.matrix(design, data=target)
  cat("\nDesign of the statistical model:\n")
  cat(paste(as.character(design),collapse=" "),"\n")

  ggsave(path ="figures/", filename ="02 - correlation between samples - filter - unnorm.pdf",ggpairs(a
s.data.frame(dge$counts[,1:nsamples])))
  ggsave(path ="figures/", filename ="02 - correlation between samples - filter - unnorm.png",ggpairs(a
s.data.frame(dge$counts[,1:nsamples])))

  ggsave(path ="figures/", filename ="03 - correlation between log samples - filter - unnorm.pdf",ggpai
rs(as.data.frame(log(dge$counts[,1:nsamples]))))
  ggsave(path ="figures/", filename ="03 - correlation between log samples - filter - unnorm.png",ggpai
rs(as.data.frame(log(dge$counts[,1:nsamples]))))

  # normalization of regions composition bias
  dge <- calcNormFactors(dge, method=normalizationMethod)
  cat("\nNormalization factors:\n")
  print(dge$samples$norm.factors)

  # estimating dispersions
  dge <- estimateGLMCommonDisp(dge, dge$design)
  dge <- estimateGLMTrendedDisp(dge, dge$design)
  dge <- estimateGLMTagwiseDisp(dge, dge$design)

  # graphical representation
  # BCV based on average log CPM (dispersion representation):
  pdf("figures/04 dispersion plot - glmFit method .pdf")
  plotBCV(dge)
  dev.off()

  png("figures/04 dispersion plot - glmFit method .png")
  plotBCV(dge)
  dev.off()

  # MDS plot:
  pdf("figures/05 MDS plot - glmFit method - bcv.pdf")
  mds = plotMDS(dge,  method="bcv",col=rep(couleur[1:nstage],each=nrep),
                pch=1, cex=1, xlim=c(-0.55,1.0),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],
```

```
                                                    each=nrep), pos=3)
  dev.off()

  png("figures/05 MDS plot - glmFit method - bcv.png")
  mds = plotMDS(dge,  method="bcv",col=rep(couleur[1:nstage],each=nrep),
                pch=1, cex=1, xlim=c(-0.55,1.0),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],
                                                    each=nrep), pos=3)
  dev.off()

  pdf("figures/06 MDS plot - glmFit method - logFC.pdf")
  mds = plotMDS(dge,  method="logFC",col=rep(couleur[1:nstage],each=nrep),
                pch=1, cex=1, xlim=c(-0.55,1.0),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],
                                                    each=nrep), pos=3)
  dev.off()

  png("figures/06 MDS plot - glmFit method - logFC.png")
  mds = plotMDS(dge,  method="logFC",col=rep(couleur[1:nstage],each=nrep),
                pch=1, cex=1, xlim=c(-0.55,1.0),ylim=c(-0.3,0.4))
  text(mds$x, mds$y, labels=rownames(target), col=rep(couleur[1:nstage],
                                                    each=nrep), pos=3)
  dev.off()

 # statistical testing:
 # perform all the comparisons between the levels of varInt
  fit <- glmFit(dge, dge$design, ...)
  cat(paste("Coefficients of the model:",paste(colnames(fit$design),collapse="  ")),"\n")
  colsToTest <- grep(varInt,colnames(fit$design))
  namesToTest <- paste0(gsub(varInt,"",colnames(fit$design)[colsToTest]),"_vs_",condRef)
  results <- list()
  # testing coefficients individually (tests against the reference level)
  for (i in 1:length(colsToTest)){
    cat(paste0("Comparison ",gsub("_"," ",namesToTest[i]),": testing coefficient ",
               colnames(fit$design)[colsToTest[i]]),"\n")

    lrt <- glmLRT(fit, coef=colsToTest[i])
    results[[namesToTest[i]]] <- topTags(lrt,n=nrow(dge$counts),
                         adjust.method=pAdjustMethod,sort.by="none")$table
  }
  # defining contrasts for the other comparisons (if applicable)
  if (length(colsToTest)>=2){
    colnames <- gsub(varInt,"",colnames(fit$design))
    for (comp in combn(length(colsToTest),2,simplify=FALSE)){
      contrast <- numeric(ncol(dge$design))
      contrast[colsToTest[comp[1:2]]] <- c(-1,1)

      namecomp <- paste0(colnames[colsToTest[comp[2]]],"_vs_",
                         colnames[colsToTest[comp[1]]])

      cat(paste0("Comparison ",gsub("_"," ",namecomp),": testing contrast (",
                 paste(contrast,collapse=", "),")"),"\n")

      lrt <- glmLRT(fit, contrast=contrast)
      results[[namecomp]] <- topTags(lrt,n=nrow(dge$counts),
                    adjust.method=pAdjustMethod,sort.by="none")$table
    }
  }

  return(list(dge=dge,results=results,lrt=lrt))
}


out.edgeR <- run.edgeR(counts=counts, target=target, varInt=varInt,
                       condRef=condRef, batch=batch, cpmCutoff=cpmCutoff,
                       normalizationMethod=normalizationMethod,
                       pAdjustMethod=pAdjustMethod)

#--------------------
# MDS + clustering (unchanged SARTools function)
#--------------------
exploreCounts(object=out.edgeR$dge, group=target[,varInt],
              gene.selection=gene.selection, col=colors)

#--------------------
# exporting results of the differential analysis (adapted from SARTools function)
#--------------------
```

```
rawpHist_AD <- function(complete_AD, outfile=TRUE){
  ncol <- ifelse(length(complete_AD)<=4, ceiling(sqrt(length(complete_AD))), 3)
  nrow <- ceiling(length(complete_AD)/ncol)

  par(mfrow=c(nrow,ncol))
  for (name in names(complete_AD)){
    ggplot(as.data.frame(complete_AD[[name]]), aes(x=pvalue, fill="tomato2")) +
      geom_histogram(binwidth = 0.025, color="black") +
      scale_fill_discrete(name = "", labels = "pvalues") +
      labs(title = "Raw pvalues histogram") +
      theme(plot.margin = margin(2,.8,2,.8, "cm"))
  }
  if (outfile) ggsave(path= "figures/", filename="rawpHist_ggplot.pdf")
  if (outfile) ggsave(path= "figures/", filename="rawpHist.png")
  if (outfile) dev.off()
}

exportResults.edgeR <- function(out.edgeR, group, counts, alpha=0.05, export=TRUE){

  dge <- out.edgeR$dge
  res <- out.edgeR$results

  # raw count, normalised count and baseMean
  tmm <- dge$samples$norm.factors
  N <- colSums(dge$counts)
  f <- tmm * N/mean(tmm * N)
  normCounts <- round(scale(dge$counts, center=FALSE, scale=f))
  base <- data.frame(Id=rownames(counts), counts)
  names(base) <- c("Id", colnames(counts))
  norm.bm <- data.frame(Id=rownames(normCounts),normCounts)
  names(norm.bm) <- c("Id", paste0("norm.",colnames(normCounts)))
  norm.bm$baseMean <- round(apply(scale(dge$counts, center=FALSE, scale=f),1,mean),2)
  for (cond in levels(group)){
    norm.bm[,cond] <- round(apply(as.data.frame(normCounts[,group==cond]),1,mean),0)
  }
  base <- merge(base,norm.bm,by="Id",all=TRUE)

  complete <- list()
  for (name in names(res)){
    complete.name <- base

    # To add info from res
    res.name <- data.frame(Id=rownames(res[[name]]),
                           FC=round(2^(res[[name]][,"logFC"]),3),
                           log2FoldChange=round(res[[name]][,"logFC"],3),
                           pvalue=res[[name]][,"PValue"],
                           padj=res[[name]][,ifelse("FDR" %in% names(res[[name]]), "FDR", "FWER")])
    complete.name <- merge(complete.name, res.name, by="Id", all=TRUE)

    # To add info from dge
    dge.add <- data.frame(Id=rownames(dge$counts),
                          tagwise.dispersion=round(dge$tagwise.dispersion,4),
                          trended.dispersion=round(dge$trended.dispersion,4))
    complete.name <- merge(complete.name, dge.add, by="Id", all=TRUE)
    complete[[name]] <- complete.name

    if (export){
      # To obtain 'sartools' and restricted lists of DEG druing brain development
      up.name <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChange>=0),]
      up.name <- up.name[order(up.name$padj),]
      down.name <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChange<=0),
]
      down.name <- down.name[order(down.name$padj),]
      up.name.strict <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChange>
=1),]
      up.name.strict <- up.name.strict[order(up.name.strict$padj),]
      down.name.strict <- complete.name[which(complete.name$padj <= alpha & complete.name$log2FoldChang
e<=-1),]
      down.name.strict <- down.name.strict[order(down.name.strict$padj),]
      for (i in list("up.name", "down.name","up.name.strict","down.name.strict")){
        dimension = nrow(get(i))
        print(paste("number of rows :", i, "=", dimension,sep=" "))
      }

      # To save the data
      name <- gsub(" ","",name)
```

```
      write.table(complete.name, file=paste0("tables/",name,".complete.txt"),
                  sep="\t", row.names=FALSE, dec=".", quote=FALSE)

      write.table(up.name, file=paste0("tables/", name,".up.txt"), row.names=FALSE,
                  sep="\t", dec=".", quote=FALSE)

      write.table(down.name, file=paste0("tables/", name,".down.txt"), row.names=FALSE,
                  sep="\t", dec=".", quote=FALSE)

      write.table(up.name.strict, file=paste0("tables/", name,".up.strict.txt"),
                  row.names=FALSE, sep="\t", dec=".", quote=FALSE)

      write.table(down.name.strict, file=paste0("tables/", name,".down.strict.txt"),
                  row.names=FALSE, sep="\t", dec=".", quote=FALSE)
    }
  }

  return(complete)
}


# Summary of the analysis
# (boxplots, dispersions, export table, nDiffTotal, histograms, MA plot)
summarizeResults.edgeR.AD = function (out.edgeR, group, counts, alpha = 0.05,
                      col = c("lightblue","orange", "MediumVioletRed", "SpringGreen"),
                      log2FClim = NULL, padjlim = NULL)
{
  if (!I("figures" %in% dir()))
    dir.create("figures", showWarnings = FALSE)
  if (!I("tables" %in% dir()))
    dir.create("tables", showWarnings = FALSE)
  countsBoxplots(out.edgeR$dge, group = group, col = col)
  BCVPlot(dge = out.edgeR$dge)
  complete <- exportResults.edgeR(out.edgeR = out.edgeR, group = group,
                                  counts = counts, alpha = alpha)
  nDiffTotal <- nDiffTotal(complete = complete, alpha = alpha)
  cat("Number of features down/up and total:\n")
  print(nDiffTotal, quote = FALSE)
  rawpHist_AD(complete_AD = complete, outfile = TRUE)
  MAPlot(complete = complete, alpha = alpha, log2FClim = log2FClim)
  volcanoPlot(complete = complete, alpha = alpha, padjlim = padjlim)
  return(list(complete = complete, nDiffTotal = nDiffTotal))
}


summaryResults.AD <- summarizeResults.edgeR.AD(out.edgeR, group=target[,varInt],
                                       counts=counts, alpha=alpha, col=colors)


#--------------------
# save image of the R session (unchanged SARTools script)
#--------------------
save.image(file=paste0(projectName, ".RData"))


#--------------------
# generating HTML report (adapted from SARTools function)
#--------------------
writeReport.edgeR(target=target, counts=counts, out.edgeR=out.edgeR,
                  summaryResults=summaryResults.AD, majSequences=majSequences,
                  workDir=workDir, projectName=projectName, author=author,
                  targetFile=targetFile, rawDir=rawDir,
                  featuresToRemove=featuresToRemove, varInt=varInt,
                  condRef=condRef, batch=batch, alpha=alpha,
                  pAdjustMethod=pAdjustMethod, cpmCutoff=cpmCutoff,
                  colors=colors, gene.selection=gene.selection,
                  normalizationMethod=normalizationMethod)

#--------------------
# save FDR and logCPM information (new in the script)
#--------------------

if (length(names(out.edgeR$results))){
  write.table(out.edgeR$results, file="tables/results_FDR_logCPM_all_filter.txt", sep="\t", row.names=F
ALSE, dec=".", quote=FALSE)
}else{
  print("Certaines données ne seront pas sauvegardées")
```

```
}



# ----------------------
# ------------------------------------------
# Same code for  E14 vs E15 developmental stages comparison
# with following parameter settings
# ------------------------------------------
# ----------------------


dir.create("G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E14.15", recursive=TRUE)


rm(list=ls())

workDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E14.15"
# working directory for the R session

projectName <- "Sartools-edgeR-RNAseq-E14.15"       # name of the project
author <- "Agathe D."                               # author of the statistical analysis/report

targetFile <- "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE14.15.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",       # names of the features to be removed
                      "ambiguous", "no_feature",    # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                   # factor of interest
condRef <- "E14"                                    # reference biological condition
batch <- NULL                                       # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                       # threshold of statistical significance
pAdjustMethod <- "BH"                               # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                      # counts-per-million cut-off to filter low counts
gene.selection <- "pairwise"                        # selection of the features in MDSPlot
normalizationMethod <- "TMM"                        # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE

nsamples = 4                                        # nb of samples
nstage = 2                                          # nb of developemental stages
nrep = 2                                            # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")




# ----------------------
# ------------------------------------------
# Same code for E15 vs E16 developmental stages comparison
# with following parameter settings
# ------------------------------------------
# ----------------------

dir.create("G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E15.16", recursive=TRUE)

rm(list=ls())

workDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E15.16"      # working directory
for the R session

projectName <- "Sartools-edgeR-RNAseq-E15.16"       # name of the project
author <- "Agathe D."                               # author of the statistical analysis/report
```

```
author <- "Agathe D."                                  # author of the statistical analysis/report

targetFile <- "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE15.16.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",          # names of the features to be removed
                      "ambiguous", "no_feature",       # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                      # factor of interest
condRef <- "E15"                                       # reference biological condition
batch <- NULL                                          # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                          # threshold of statistical significance
pAdjustMethod <- "BH"                                  # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                         # counts-per-million cut-off to filter low counts
gene.selection <- "pairwise"                           # selection of the features in MDSPlot
normalizationMethod <- "TMM"                           # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE

nsamples = 4                                           # nb of samples
nstage = 2                                             # nb of developemental stages
nrep = 2                                               # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")


# ----------------------
# ----------------------------------------
# Same code for E16 vs P0 developmental stages comparison
# with following parameter settings
# ----------------------------------------
# ----------------------

dir.create("G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E16.P0", recursive=TRUE)

rm(list=ls())

workDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E16.P0"      # working directory
for the R session

projectName <- "Sartools-edgeR-RNAseq-E16.P0"                        # name of the project
author <- "Agathe D."                                  # author of the statistical analysis/report

targetFile <- "G:/projet_DU_AD/results/07_DESEq2_mm10/targetE16.P0.txt"
# path to the design/target file
rawDir <- "G:/projet_DU_AD/results/07_DESEq2_mm10/"
# path to the directory containing raw counts files
featuresToRemove <- c("alignment_not_unique",          # names of the features to be removed
                      "ambiguous", "no_feature",       # (specific HTSeq-count information and rRNA for e
xample)
                      "not_aligned", "too_low_aQual")# NULL if no feature to remove

varInt <- "group"                                      # factor of interest
condRef <- "E16"                                       # reference biological condition
batch <- NULL                                          # blocking factor: NULL (default) or "batch" for e
xample

alpha <- 0.05                                          # threshold of statistical significance
pAdjustMethod <- "BH"                                  # p-value adjustment method: "BH" (default) or "BY
"

cpmCutoff <- 1                                         # counts-per-million cut-off to filter low counts
gene.selection <- "pairwise"                           # selection of the features in MDSPlot
normalizationMethod <- "TMM"                           # normalization method: "TMM" (default), "RLE" (DE
Seq) or "upperquartile"

colors <- c("darkblue","brown3")

forceCairoGraph <- FALSE
```

```
forcedAffinity = TRUE

nsamples = 4                                          # nb of samples
nstage = 2                                            # nb of developemental stages
nrep = 2                                              # nb of replicates per developmental stages
couleur=c("cyan4","brown3","darkgreen","darkblue","plum4","lightgoldenrod3","pink")
```

Annotation of DEG

After edgeR running, we obtained lists of DEG with each gene identified by its ENSEMBL ID.

To annotate DEG, a file (biomart_mm10.txt) containing several annotation data, was generated with **BiomaRt** R package. DEG were annotated using this file and **join unix command**. The file obtained with Biomart contains :

- ensembl_gene_id: gene name
- chromosome_name: chromosome/scaffold number
- strand
- start_position: gene start (bp)
- end_position: gene end (bp)
- entrezgene_id:NCBI gene ID
- gene_biotype: gene type
- mgi_symbol: MGI symbol
- entrezgene_accession: NCBI gene accession
- entrezgene_description: NCBI gene description
- uniprot_gn_symbol: UniProtKB gene name symbol

In [ ]:

```
%%bash

#-------------------------------------
# 1. Obtention of DEG ENSEMBL_ID, without decimal
# to obtain same ID than those contained in annotation file
#-------------------------------------

cd /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E13.14/tables
sed '1d' E14vsE13.down.txt | cut -d . -f1 > ENS_ID_down_DEG_E13.14_mm10.txt
sed '1d' E14vsE13.up.txt | cut -d . -f1 > ENS_ID_up_DEG_E13.14_mm10.txt

cd /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E14.15/tables
sed '1d' E15vsE14.down.txt | cut -d . -f1 > ENS_ID_down_DEG_E14.15_mm10.txt
sed '1d' E15vsE14.up.txt | cut -d . -f1 > ENS_ID_up_DEG_E14.15_mm10.txt

cd /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E15.16/tables
sed '1d' E16vsE15.down.txt | cut -d . -f1 > ENS_ID_down_DEG_15.E16_mm10.txt
sed '1d' E16vsE15.up.txt | cut -d . -f1 > ENS_ID_up_DEG_E15.16_mm10.txt

cd /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E14.16/tables
sed '1d' E16vsE14.down.txt | cut -d . -f1 > ENS_ID_down_DEG_14.E16_mm10.txt
sed '1d' E16vsE14.up.txt | cut -d . -f1 > ENS_ID_up_DEG_E14.16_mm10.txt

cd /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E16.P0/tables
sed '1d' P0vsE16.down.txt | cut -d . -f1 > ENS_ID_down_DEG_16.P0_mm10.txt
sed '1d' P0vsE16.up.txt | cut -d . -f1 > ENS_ID_up_DEG_16.P0_mm10.txt
```

In [ ]:

```
#-------------------------------------
#  2.Obtention of biomart_mm10.txt file
#-------------------------------------

# Loading BiomaRt package :
library("biomaRt")

# Selection of database and reference genome
mm10 = useMart("ensembl", dataset="mmusculus_gene_ensembl")
# mm10 Mus musculus version used:
# Ensembl 97 Jul 2019 http://jul2019.archive.ensembl.org

# Creating the dataset with information of interest
# It's necessary to do it in two steps, because there is a limitation in the number of attributes that
```

```
can be collected at the same time.
annot_mm10_part1<-getBM(attributes=c("ensembl_gene_id","chromosome_name","strand",
                                     "start_position","end_position","entrezgene_id",
                                     "gene_biotype","mgi_symbol",
                                     "entrezgene_accession"), mart=mm10)
annot_mm10_part2<-getBM(attributes=c("ensembl_gene_id","entrezgene_description",
                                     "uniprot_gn_symbol"), mart=mm10)
annot_mm10 = merge(annot_mm10_part1, annot_mm10_part2, by="ensembl_gene_id", all=TRUE)

# Verifications
dim(annot_mm10)
head(annot_mm10)

# Saving
write.table(annot_mm10,
            "E:/projet_DU_AD/results/annotation_mm10/Biomart/biomart_mm10.txt",
            quote= FALSE, sep="\t",row.names=FALSE)
```

In [ ]:

```
%%bash

#-------------------------------------
# 3.Combine DEG files with biomart annotation file
#-------------------------------------

# Copy and paste files containing ENSEMBL ID (of up or down DEG)
# in folder /mnt/g/projet_DU_AD/results/09_annotation/Sartools/Biomart/raw_data

# Datasets formatting:
# To put ENSEMBL ID in the first column:
cd /mnt/g/projet_DU_AD/results/annotation_mm10/Biomart

sed '1d' biomart_mm10.txt | sort > biomart_mm10_sorted.txt

mkdir -p /mnt/g/projet_DU_AD/results/09_annotation/Sartools/Biomart

cp /mnt/g/projet_DU_AD/results/annotation_mm10/Biomart/biomart_mm10_sorted.txt
    /mnt/g/projet_DU_AD/results/09_annotation/Sartools/Biomart

cd /mnt/g/projet_DU_AD/results/09_annotation/Sartools/Biomart/

sample="E13.14 E14.15 E15.16 E14.16 E16.P0"

for ech in ${sample}
do
    echo "============================================="
    echo "file name : ${ech}"
    echo "Annotation"
    echo "============================================="

    sort -k1,1 raw_data/ENS_ID_up_DEG_${ech}_mm10.txt
    > raw_data/ENS_ID_up_DEG_${ech}_mm10_sorted.txt

    sort -k1,1 raw_data/ENS_ID_down_DEG_${ech}_mm10.txt
    > raw_data/ENS_ID_down_DEG_${ech}_mm10_sorted.txt

# To combine informations
    join -1 1 -2 1 raw_data/ENS_ID_up_DEG_${ech}_mm10_sorted.txt
    biomart_mm10_sorted.txt -t $'\t' > AnnotBM_up_${ech}_mm10.txt

    join -1 1 -2 1 raw_data/ENS_ID_down_DEG_${ech}_mm10_sorted.txt
    biomart_mm10_sorted.txt -t $'\t' > AnnotBM_down_${ech}_mm10.txt

    echo "AnnotBM_up_${ech}_mm10.txt"
    wc -l AnnotBM_up_${ech}_mm10.txt
    head AnnotBM_up_${ech}_mm10.txt

    echo "AnnotBM_down_${ech}_mm10.txt"
    wc -l AnnotBM_down_${ech}_mm10.txt
    head AnnotBM_down_${ech}_mm10.txt

done
```

Combination of annotation and statistical informations

In order to obtain a file where each region of interest is represented by a single row, the annotation elements of a given region that are on distinct lines have been grouped together. Then, all data (statistical informations from edgeR, mm10 coordinates regions, annotation informations) are combined in an single file (one file per pairwise comparison).

In [ ]:

```
%%bash

# ------------------------------------------------------------------------
# Dataset n° 1 (files containing statistical informations) formatting
# ------------------------------------------------------------------------

# To remove header
cd  /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E13.14/tables
sed '1d' E14vsE13.down.txt > nh_down_DEG_E13.14_mm10.txt
sed '1d' E14vsE13.up.txt > nh_up_DEG_E13.14_mm10.txt

cd  /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E14.15/tables
sed '1d' E15vsE14.down.txt > nh_down_DEG_E14.15_mm10.txt
sed '1d' E15vsE14.up.txt > nh_up_DEG_E14.15_mm10.txt

cd  /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E15.16/tables
sed '1d' E16vsE15.down.txt > nh_down_DEG_E15.16_mm10.txt
sed '1d' E16vsE15.up.txt > nh_up_DEG_E15.16_mm10.txt

cd  /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E14.16/tables
sed '1d' E16vsE14.down.txt > nh_down_DEG_E14.16_mm10.txt
sed '1d' E16vsE14.up.txt > nh_up_DEG_E14.16_mm10.txt

cd  /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/E16.P0/tables
sed '1d' P0vsE16.down.txt > nh_down_DEG_E16.P0_mm10.txt
sed '1d' P0vsE16.up.txt > nh_up_DEG_E16.P0_mm10.txt

# To combine files, we need a common ID = ENSEMBL_ID (without decimal)
# To replace ENSEMBL_ID (with decimal) to ENSEMBL_ID (without) in identified DEG files, containing stat
istical informations
sample="E13.14 E14.15 E15.16 E14.16 E16.P0"

for ech in ${sample}
do
    cd /mnt/g/projet_DU_AD/results/07_DESEq2_mm10/SARTools/Analyse_L/${ech}/tables

    echo "-----------------------------------"
    echo "${ech}"
    echo "-----------------------------------"

    paste ENS_ID_down_DEG_${ech}_mm10.txt nh_down_DEG_${ech}_mm10.txt >
    ID_double_nh_down_DEG_${ech}_mm10.txt

    paste ENS_ID_up_DEG_${ech}_mm10.txt nh_up_DEG_${ech}_mm10.txt >
    ID_double_nh_up_DEG_${ech}_mm10.txt

    awk '{ print $1"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9"\t"$10"\t"$11"\t"$12"\t"$13"\t"$14"\t"$
15"\t"$16"\t"$17"\t"$18"\t"$19}' ID_double_nh_down_DEG_${ech}_mm10.txt > ID_sd_nh_down_${ech}_mm10.txt
    awk '{ print $1"\t"$3"\t"$4"\t"$5"\t"$6"\t"$7"\t"$8"\t"$9"\t"$10"\t"$11"\t"$12"\t"$13"\t"$14"\t"$
15"\t"$16"\t"$17"\t"$18"\t"$19 }' ID_double_nh_up_DEG_${ech}_mm10.txt > ID_sd_nh_up_${ech}_mm10.txt

    echo "ID_sd_nh_down${ech}_mm10.txt"
    wc -l ID_sd_nh_down_${ech}_mm10.txt
    head ID_sd_nh_down_${ech}_mm10.txt

    echo "ID_sd_nh_up${ech}_mm10.txt"
    wc -l ID_sd_nh_up_${ech}_mm10.txt
    head ID_sd_nh_up_${ech}_mm10.txt
done
```

In [ ]:

```
# ------------------------------------------------------------------------
# To concatenate informations
# ------------------------------------------------------------------------

#------------------------------------------------------------------------
```

```
# 1. To concatenate annotation informations : one row = one region of interest / DEG
# Load files : List files containing "AnnotBM" or "ID_sd" patterns
list_files = list.files("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart")
list_files_Annot = list_files[which(grepl("AnnotBM",list_files))]
list_files_ID_sd = list_files[which(grepl("ID_sd",list_files))]

for(i in 1:10){
  Annot = list_files_Annot[i]
  DEG_annot = read.table(file = paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart",Annot,se
p="/"), header = FALSE, sep = "\t", quote="", fill=TRUE)

  colnames(DEG_annot) = c("ID", "chr", "strand","start_gene","end_gene","entrezgene_id",
                          "gene_biotype","mgi_symbol","entrezgene_accession",
                          "entrezgene_description", "uniprot_gn_symbol")

  # To concatenate informations
  library(dplyr)

  # List of unique ID
  liste_DEG_unique <- unique(DEG_annot$ID)

  # empty matrice which will be filled during the loop
  # Number of rows = nb of unique ID | Nb of columns = nb of variables
  treatment_file <- data.frame(matrix(NA,nrow=length(liste_DEG_unique),ncol=ncol(DEG_annot)))
  colnames(treatment_file) <- colnames(DEG_annot)

  treatment_file$ID <- liste_DEG_unique
  for(k in 1:length(liste_DEG_unique)){
    # Filter according to ID k
    DEG_filt <- DEG_annot %>%
      filter(ID == liste_DEG_unique[k])

    # Collapsing of distincts elements of a variables for ID k
    # If variable is composed of only one element ==> we keep this element only once
    # If variable is composed of m distinct elements ==> even if elements is repeated,
      # we keep all the elements (with repetitions)
    for(j in 2:length(colnames(treatment_file))){
      if(length(unique(DEG_filt[,j])) == 1){
        treatment_file[k,j] <- paste(unique(DEG_filt[,j]), collapse=';')
      } else {
        treatment_file[k,j] <- paste(DEG_filt[,j], collapse=';')
      }
    }
  }
  print(paste("nb of uniq ID treatment file : ", nrow(treatment_file), sep=""))

  #-------------------------------------------------------------------------
  # 2. To merge DEG annotations with statistical informations obtained with SARTools-edgeR for every re
gions

  ID_sd = list_files_ID_sd[i]
  sartools_file = read.table(file = paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart",ID_s
d,sep="/"), header = FALSE, sep = "\t", quote="", dec=".")

  colnames(sartools_file) = c("ID","cond1_rep1","cond1_rep2","cond2_rep1",
                              "cond2_rep2","norm.cond1_rep1","norm.cond1_rep2",
                              "norm.cond2_rep1","norm.cond2_rep2","baseMean",
                              "cond1","cond2","FC","log2FoldChange",
                              "pvalue","padj","tagwise.dispersion","trended.dispersion")

  Annot_DESeq = merge(sartools_file,treatment_file, by="ID", all = TRUE)

  write.table(Annot_DESeq, file=paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/edgeR",An
not,sep="_"), quote=FALSE, sep="\t", row.names = FALSE, col.names = TRUE)

  size_Annot = nrow(treatment_file)
  size_ID_sd = nrow(sartools_file)

  if(size_Annot != size_ID_sd){
    print(paste("Les deux fichiers", Annot, "et", ID_sd,
                "n'ont pas la meme taille !", sep =" "))
  }else{
  }
}
```

Because Gencode and Biomart annotation files contain not always exactly the same informations, due to differences between

available information in the database, few DEG were not associated to a gene name (in 5 of 8 files), using Biomart annotation file.

Number of DEG without annotation:

- 3 downregulated genes between E13 and E14
- 1 downregulated gene between E14 and E15
- 0 downregulated gene between E15 and E16
- 2 downregulated genes between E16 and P0
- 1 upregulated gene between E13 and E14
- 2 upregulated genes between E14 and E15
- 0 upregulated gene between E15 and E16
- 0 upregulated gene between E16 and P0

The DEG lacking annotation information were manually annotated, based on Gencode annotation contained in the file called gencode.vM21.annotation.gtf.

In [ ]:

```
# Automatisation to annotate DEG that lack annotation information, using Gencode annotation database.

library(data.table)
library(stringr)

# Name of samples
nom_sample = list("down_E13.14", "down_E14.15", "down_E14.16", "down_E15.16",
                  "down_E16.P0", "up_E13.14", "up_E14.15", "up_E14.16",
                  "up_E15.16", "up_E16.P0" )
list_sample <- list()

# Load samples files :
for(i in 1:length(nom_sample)){
  list_sample[[i]] <- read.table(paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/edgeR_An
notBM_",
            nom_sample[i] , "_mm10.txt", sep=""), sep = "\t", header= TRUE,
            dec=".", quote="", stringsAsFactors = FALSE)

  print(paste("head_ech",nom_sample[i], sep=""))
  print(head(list_sample[[i]]))
  print(paste("dimension du dataset ech", nom_sample[i], sep = "_"))
  print(dim(list_sample[[i]]))
}
names(list_sample) <- nom_sample
```

In [ ]:

```
%%bash

# Gencode file formatting
# To remove 5 first lines of the file (correponding to the header)

cd /mnt/g/projet_DU_AD/data/mm10_genome
sed -e '1,5d' gencode.vM21.annotation.gtf > nh_gencode.vM21.annotation.gtf
```

In [ ]:

```
gencode <- read.table("G:/projet_DU_AD/data/mm10_genome/nh_gencode.vM21.annotation.gtf", sep = "\t", he
ader= FALSE)

colnames(gencode) <- c("chr","source","element_type","start","end","V6","strand","V8","infos")

head(gencode)
dim(gencode)

for(i in 1:length(nom_sample)){
  # To find rows where annotation informations are missing
  na_values = list_sample[[i]][is.na(list_sample[[i]][,"strand"])==TRUE,]

  if(nrow(na_values)>0){
    print(nom_sample[[i]])
    print(na_values)
    # to replace missing annotation informations by data from Gencode database
    na_values_id <- as.character(na_values$ID)
```

```
    for(j in 1:length(na_values_id)){
      info_to_get <- gencode[(gencode$infos %like% na_values_id[j]) & gencode$element_type =="gene",]
      list_sample[[i]][rownames(na_values)[j],"chr"] <- info_to_get$chr
      list_sample[[i]][rownames(na_values)[j],"strand"] <- info_to_get$strand
      list_sample[[i]][rownames(na_values)[j],"start_gene"] <- info_to_get$start
      list_sample[[i]][rownames(na_values)[j],"end_gene"] <- info_to_get$end

      list_sample[[i]][rownames(na_values)[j],"entrezgene_id"] <- "."

      list_sample[[i]][rownames(na_values)[j],"gene_biotype"] <-
        str_split(str_split(info_to_get$infos,pattern = "; ")[[1]][2],
                  pattern="gene_type ")[[1]][2]

      list_sample[[i]][rownames(na_values)[j],"mgi_symbol"] <-
        str_split(str_split(info_to_get$infos,pattern = "; ")[[1]][3],
                  pattern="gene_name ")[[1]][2]

      list_sample[[i]][rownames(na_values)[j],"entrezgene_accession"] <-
        str_split(str_split(info_to_get$infos,pattern = "; ")[[1]][3],
                  pattern="gene_name ")[[1]][2]

      list_sample[[i]][rownames(na_values)[j],"entrezgene_description"] <- "."
      list_sample[[i]][rownames(na_values)[j],"uniprot_gn_symbol"] <- "."

    }
    print(list_sample[[i]][rownames(na_values),])

  } else {
     write.table(list_sample[[i]],
                 file=paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/edgeR_AnnotBM_et_Man
_",nom_sample[i],"_mm10.txt",sep=""),
                 quote=FALSE,sep="\t",row.names=FALSE,col.names=TRUE)
  }

  # Save data
   write.table(list_sample[[i]],
               file=paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/edgeR_AnnotBM_et_Man_"
,nom_sample[i],"_mm10.txt",sep=""),
               quote=FALSE,sep="\t",row.names=FALSE,col.names=TRUE)

}
```

Generate .bed files for data visualization

To visualize DEG coordinates using **IGV**, .bed files (containing chromosome, start and end informations of each DEG) were generated from files containing DEG informations.

In [ ]:

```
nom_sample = list=c("down_E13.14", "down_E14.15", "down_E14.16", "down_E15.16",
                    "down_E16.P0", "up_E13.14", "up_E14.15", "up_E14.16",
                    "up_E15.16", "up_E16.P0" )

# Load files:
for(name in nom_sample){
  assign(x=paste("ech",name, sep = "_"),
         value = read.table(paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/edgeR_AnnotBM
_et_Man_",
         name , "_mm10.txt", sep=""), sep = "\t", header= TRUE, dec=".", quote=""))
  print(paste("head_ech",name, sep=""))
  print(head(eval(parse(text = paste("ech",name, sep="_")))))
  print(paste("dimension du dataset ech",name, sep = "_"))
  print(dim(eval(parse(text = paste("ech",name, sep = "_")))))
}


setwd("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/")

mes_fichiers = list("ech_up_E13.14", "ech_up_E14.15", "ech_up_E14.16",
                    "ech_up_E15.16", "ech_up_E16.P0", "ech_down_E13.14",
                    "ech_down_E14.15", "ech_down_E14.16", "ech_down_E15.16",
                    "ech_down_E16.P0")

for (i in mes_fichiers){
```

```
  dataset = get(i)
  dataset$chromosome = with(dataset, paste("chr",dataset[,"chr"], sep=""))
  bed_file = dataset[,c("chromosome", "start_gene", "end_gene")]
  print(paste("dimension du fichier bed de ",i, ":", nrow(bed_file), sep =" "))
  print(head(bed_file))
  write.table(bed_file, file=paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/bed_edge_Ann
otBM_et_Man_",i,"_mm10.bed", sep=""), quote=FALSE, row.names=FALSE, col.names=FALSE, sep="\t")
}
```

In [ ]:

```
%%bash

# To sort .bed files
cd /mnt/g/projet_DU_AD/results/09_annotation/Sartools/Biomart

for file in bed*
do  sort -k1,1 -k2,2n $file > ${file/.bed/.sorted.bed}
echo $file
done
```

Gene ontology

To perform a Gene ontology (GO) analysis, Gene names of DEG were isolated from files containing DEG informations. GO analysis was done using WEB-based GEne SeT AnaLysis Toolkit (GESTALT) for DEG that were found between E14.5 and E15.5 and between E15.5 and E16.5.

In [ ]:

```
setwd("E:/projet_DU_AD/results/09_annotation/Sartools/Biomart")

DEG_u14_15 = read.table("edgeR_AnnotBM_et_Man_up_E14.15_mm10.txt",
                        header = TRUE, fill=TRUE, quote="", sep="\t",
                        na.strings = "", stringsAsFactors = FALSE)

DEG_u15_16 = read.table("edgeR_AnnotBM_et_Man_up_E15.16_mm10.txt",
                        header = TRUE, fill=TRUE, quote="", sep="\t",
                        na.strings = "", stringsAsFactors = FALSE)

DEG_d14_15 = read.table("edgeR_AnnotBM_et_Man_down_E14.15_mm10.txt",
                        header = TRUE, fill=TRUE, quote="", sep="\t",
                        na.strings = "", stringsAsFactors = FALSE)

DEG_d15_16 = read.table("edgeR_AnnotBM_et_Man_down_E15.16_mm10.txt",
                        header = TRUE, fill=TRUE, quote="", sep="\t",
                        na.strings = "", stringsAsFactors = FALSE)

head(DEG_u14_15)
head(DEG_u15_16)
head(DEG_d14_15)
head(DEG_d15_16)

dim(DEG_u14_15)
dim(DEG_u15_16)
dim(DEG_d14_15)
dim(DEG_d15_16)

# Function to extract ID
unlist_gene = function(treatment_file, gene_ID, file_name, folder){
  # To extract ID from mm10 Biomart annotation files:
  # Obtentioon of a matrix with only one column, containing ID
  gene_name_ID = as.character(treatment_file[,gene_ID])

  # Obtention of a list with ID which are deconcatenate
  list_gene_splitted <- strsplit(gene_name_ID,split=";")
  vec_gene_splitted <- unique(unlist(list_gene_splitted))

  # Remove missing data = "." :
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  # Obtention of data in a table:
  table_VGS_filter = as.data.frame(vec_gene_splitted_filter)
```

```
  write.table(table_VGS_filter, paste(folder,gene_ID,"_",file_name,".txt",sep=""),
              quote=FALSE, col.names=FALSE, row.names=FALSE)
}


# Extraction of ENSEMBL ID : I then renamed ID files to precise that ID correspond to ENSEMBL ID.
unlist_gene(treatment_file=DEG_u14_15, gene_ID="ID", file_name="DEG_up_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="ID", file_name="DEG_up_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_d14_15, gene_ID="ID", file_name="DEG_down_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="ID", file_name="DEG_down_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

# Extraction of Entrez accession ID
unlist_gene(treatment_file=DEG_u14_15, gene_ID="entrezgene_accession",
            file_name="DEG_up_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="entrezgene_accession",
            file_name="DEG_up_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_d14_15, gene_ID="entrezgene_accession",
            file_name="DEG_down_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="entrezgene_accession",
            file_name="DEG_down_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

#Extraction of mgi ID
unlist_gene(treatment_file=DEG_u14_15, gene_ID="mgi_symbol",
            file_name="DEG_up_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="mgi_symbol",
            file_name="DEG_up_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_d14_15, gene_ID="mgi_symbol",
            file_name="DEG_down_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="mgi_symbol",
            file_name="DEG_down_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")


#Extraction of Entrez_gene_ID
unlist_gene(treatment_file=DEG_u14_15, gene_ID="entrezgene_id",
            file_name="DEG_up_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="entrezgene_id",
            file_name="DEG_up_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_d14_15, gene_ID="entrezgene_id",
            file_name="DEG_down_14_15",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")

unlist_gene(treatment_file=DEG_u15_16, gene_ID="entrezgene_id",
            file_name="DEG_down_15_16",
            folder="E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/ID/")
```

3.4. Workflow used to integrate DMRs, DOCR and DEG datasets

Bio-informatic workflow: combination of data from distinct analyses

Agathe Duchateau

Integration of DMR with DOCR

To determine whether DMRs are located at genes that are differential open or closed (DOCR) during physiological brain development, DMRs and DOCR regions were compared using our own R function called *find_overlaps_AD_table()*. Both overlaping regions and regions less 1000 bases apart were investigated, by setting *tolerance* argument to 0 or 1000. The obtained regions were then annotated, using annotation that was already achieved in the two complete datasets.

In [ ]:

```
%%R

#################################################################
# Integration of DMR and DOCR to find potential overlap
# (mm10 chromosomal coordinates)
#################################################################

# 1.To load DMR432 data - mm10
DMR432 = read.table("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/DMR432
_unique_annot_mm10_BMT_meth_cgi.bed", sep="\t", header = FALSE)
head(DMR432)

# To extract DMR432 chromosomic coordinates
DMR432_bed = DMR432[,c("V1","V2","V3")]
head(DMR432_bed)
dim(DMR432_bed)


# 2. To load DOCR data - mm10 (from edgeR analysis):
# ATAC-seq SORTED files - obtained with SARTools edgeR :
nom_sample = list=c("down_DOCR_E13.14", "down_DOCR_E14.15", "down_DOCR_E14.16",
                    "down_DOCR_E15.16", "down_DOCR_E16.P0", "up_DOCR_E13.14",
                    "up_DOCR_E14.15", "up_DOCR_E14.16", "up_DOCR_E15.16",
                    "up_DOCR_E16.P0" )

# /!\ /!\ Dataset must be in a matrix format, otherwise, function doesn't work /!\/!\
for(name in nom_sample){
  assign(x=name, value = as.matrix(read.table(paste("G:/projet_DU_DSD/results/10_annotations/Sartools/B
iomart/mm10/bed_gff_ID_", name, "_mm10.sorted.bed",
               sep=""), sep = "\t", header= FALSE, dec=".", quote="")))
  print(paste("head_",name, sep=""))
  print(head(eval(parse(text = name))))
  print(paste("dimension du dataset ",name, sep = ""))
  print(dim(eval(parse(text = name))))
}

# Creation of a list containing each DOCR file (up or down for each developmental stage)
ATAC = list(down_DOCR_E13.14 = down_DOCR_E13.14, down_DOCR_E14.15 = down_DOCR_E14.15,
            down_DOCR_E15.16 = down_DOCR_E15.16, down_DOCR_E14.16 = down_DOCR_E14.16,
            down_DOCR_E16.P0 = down_DOCR_E16.P0,up_DOCR_E13.14 = up_DOCR_E13.14,
            up_DOCR_E14.15 = up_DOCR_E14.15 ,up_DOCR_E15.16 = up_DOCR_E15.16,
            up_DOCR_E14.16 = up_DOCR_E14.16, up_DOCR_E16.P0 = up_DOCR_E16.P0)


for (i in 1:length(ATAC)){
  print(names(ATAC)[i])
  print(head(ATAC[[i]]))
  print(dim(ATAC[[i]]))
}


# 3. Function to find overlap or close regions (depending on value given for tolerance argument)
find_overlaps_AD_table <- function (regions_1,regions_2,tolerance=0)
{
  common_region_table = NULL
  number_of_regions_1 <- nrow(regions_1)
```

```
  number_of_regions_2 <- nrow(regions_2)

  overlapped_is <- c()
  overlapped_js <- c()

  for (i in 1:number_of_regions_1) {

    for (j in 1:number_of_regions_2) {

      if (as.character(regions_1[i,1])==as.character(regions_2[j,1])) {
        start_1 <- as.numeric(regions_1[i,2])
        end_1 <- as.numeric(regions_1[i,3])
        expanded_start_1 <- start_1 - tolerance
        expanded_end_1 <- end_1 + tolerance
        start_2 <- as.numeric(regions_2[j,2])
        end_2 <- as.numeric(regions_2[j,3])

        if (expanded_end_1>=start_2) {

          if (end_2>=expanded_start_1) {

            overlapped_is <- c(overlapped_is,i)
            overlapped_js <- c(overlapped_js,j)

            first_region <- paste(regions_1[i,1],start_1,end_1,sep="_")
            second_region <- paste(regions_2[j,1],start_2,end_2,sep="_")
            overlap_size <- min(end_1,end_2)-max(start_1,start_2)+1
            common_region_table = rbind(common_region_table,c(as.character(regions_1[i,1]),start_1,end_
1,
                                                               as.character(regions_2[j,1]),start_2,end_
2, overlap_size))

            if (overlap_size>0) {
              print(paste(first_region,second_region,sep=" "))
            } else {
              print(paste(first_region,second_region, sep=" "))
            }
          }
        }
      }
    }
  }
  print(paste("Number matched in first data set:",length(unique(overlapped_is))))
  print(paste("Number matched in second data set:",length(unique(overlapped_js))))
  if (length(common_region_table) > 0){
    colnames(common_region_table) = c("chr_reg1", "start_reg1", "end_reg1", "chr_reg2", "start_reg2", "
end_reg2","overlap_size")
  }else{
  }
  return(as.data.frame(common_region_table))
}



# 4. To search for common regions between DMR and DOCR
# overlap
for (i in 1:length(ATAC)) {
  print(names(ATAC)[i])
  DMR_ATAC = find_overlaps_AD_table(DMR432_bed, ATAC[[i]], tolerance=0)
  write.table(DMR_ATAC,paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/
mm10_DMR432", names(ATAC)[i], "dist0.txt",sep="_"), sep = "\t", quote=FALSE, row.names = FALSE, col.nam
es=FALSE)
}


# 5. To search for common regions between DMR and DOCR
# maximum distance between 2 regions = 1000bases:
for (i in 1:length(ATAC)) {
  print(names(ATAC)[i])
  DMR_ATAC = find_overlaps_AD_table(DMR432_bed, ATAC[[i]], tolerance=1000)
  write.table(DMR_ATAC,paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/
mm10_DMR432", names(ATAC)[i], "dist1000.txt",sep="_"), sep = "\t", quote=FALSE, row.names = FALSE, col.
names=FALSE)
}
```

```
##################
# Annotation
##################

# To load files or list of files
# To load DMR+DOCR regions
list_DMR_DOCR = list.files("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/DM
R_ATACseq/")
list_DMR_DOCR_dist0 = list_DMR_DOCR[which(grepl("dist0",list_DMR_DOCR))]
list_DMR_DOCR_dist1000 = list_DMR_DOCR[which(grepl("dist1000",list_DMR_DOCR))]

# List of DOCR informations
# Since some files were empty (no overlap), we need to specify list of files we want to annotate.
list_DOCR_Annot=list("stats_mm9_mm10_down_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed",
        "stats_mm9_mm10_down_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed",
        "stats_mm9_mm10_down_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed",
        "stats_mm9_mm10_up_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed",
        "stats_mm9_mm10_up_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed",
        "stats_mm9_mm10_up_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed",
        "stats_mm9_mm10_up_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed")

# To load DMR432 annotation
DMR_Annot = read.table("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/DMR
432_unique_annot_mm10_BMT_meth_cgi.bed", header=FALSE, sep="\t", quote="")
head(DMR_Annot)
dim(DMR_Annot)
colnames(DMR_Annot) = c("mm10_chr","DMR_start","DMR_end","DMR_ID","chr_annot",
                        "start_annot","end_annot","strand","entrezgene_accession",
                        "gene_biotype","mgi_symbol","ENSEMBL_ID","Entrez_ID",
                        "overlap_annot_length","chr_mm9","start_mm9","end_mm9",
                        "chr_mm10","start_mm10","end_mm10", "CpG_nb","pvalue",
                        "qvalue","meth.diff","CGi_chr",  "CGi_start", "CGi_end",
                        "CGi_name","CGi_length", "CGi_cpgNum", "CGi_overlap")

DMR_Annot_clean = DMR_Annot[,c("mm10_chr","DMR_start","DMR_end","DMR_ID",
                               "chr_annot","start_annot","end_annot","strand",
                               "entrezgene_accession","gene_biotype","mgi_symbol",
                               "ENSEMBL_ID","Entrez_ID","overlap_annot_length",
                               "chr_mm9","start_mm9","end_mm9","CpG_nb","pvalue",
                               "qvalue","meth.diff","CGi_chr",  "CGi_start",
                               "CGi_end", "CGi_name","CGi_length", "CGi_cpgNum",
                               "CGi_overlap")]


# for distance=0
for(i in 1:7){

  DMR_DOCR_sample = list_DMR_DOCR_dist0[i]
  Annot_DMR_DOCR = read.table(file = paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croiseme
nt_autre_data/DMR_ATACseq/",DMR_DOCR_sample,sep="/"),
                              header = FALSE, sep = "\t", quote="", fill=TRUE)

  # To rename columns
  colnames(Annot_DMR_DOCR) <- c("DMR_chr","DMR_start", "DMR_end", "DOCR_chr", "DOCR_start", "DOCR_end"
, "DMR_DOCR_overlap")

  # To create 2 ID : DMR_ID et DOCR_ID
  Annot_DMR_DOCR$DMR_ID = paste(Annot_DMR_DOCR$DMR_chr, Annot_DMR_DOCR$DMR_start, Annot_DMR_DOCR$DMR_e
nd, sep=";")
  Annot_DMR_DOCR$DOCR_ID = paste(Annot_DMR_DOCR$DOCR_chr, Annot_DMR_DOCR$DOCR_start, Annot_DMR_DOCR$DO
CR_end, sep=";")

  # To Load Annotation informations of all DOCR
  DMR_sample <- DMR_Annot_clean

  Annot_DOCR_sample = list_DOCR_Annot[i]
  Annot_DOCR = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10",
Annot_DOCR_sample,sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DOCR) = c("mm9_chr", "mm9_start", "mm9_end", "mm9_ID", "mm10_chr",
                           "mm10_start", "mm10_end", "mm10_ID", "chr_gene",
                           "ensembl_gene_id", "entrezgene_id", "overlap_length",
                           "cond1_rep1", "cond1_rep2", "cond2_rep1", "cond2_rep2",
                           "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1",
                           "norm.cond2_rep2", "baseMean", "cond1", "cond2", "FC",
                           "log2FoldChange", "pvalue", "padj", "tagwise.dispersion",
                           "trended.dispersion")
```

```
                                    trended.dispersion" )

  Annot_DOCR$DOCR_ID=paste(Annot_DOCR$mm10_chr, Annot_DOCR$mm10_start, Annot_DOCR$mm10_end,sep=";")

  Annot_DMR_DOCR=merge(Annot_DMR_DOCR, DMR_sample, by="DMR_ID")
  Annot_DMR_DOCR=merge(Annot_DMR_DOCR, Annot_DOCR, by="DOCR_ID")

  # To remove duplicate columns
  Annot_DMR_DOCR_clean = Annot_DMR_DOCR[,c("DMR_chr", "DMR_start.x", "DMR_end.x", "DMR_ID","chr_annot"
, "start_annot", "end_annot", "strand.x",
                                          "entrezgene_accession.x", "gene_biotype.x", "mgi_symbol.x",
"ENSEMBL_ID", "Entrez_ID",
                                          "overlap_annot_length", "chr_mm9" , "start_mm9", "end_mm9" ,
"CpG_nb", "pvalue.x" ,  "qvalue" , "meth.diff",
                                          "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length"
,"CGi_cpgNum", "CGi_overlap",
                                          "DMR_DOCR_overlap", "DOCR_chr", "DOCR_start", "DOCR_end","D
OCR_ID" ,
                                          "mm9_chr", "mm9_start", "mm9_end", "mm9_ID","cond1_rep1", "c
ond1_rep2", "cond2_rep1" , "cond2_rep2",
                                          "norm.cond1_rep1", "norm.cond1_rep2",  "norm.cond2_rep1" , "
norm.cond2_rep2",  "baseMean",  "cond1" , "cond2" ,  "FC",
                                          "log2FoldChange", "pvalue.y",  "padj", "tagwise.dispersion",
"trended.dispersion",
                                          "strand.y", "chr_gene", "start_gene", "end_gene", "entrezgen
e_accession.y", "gene_biotype.y",
                                          "mgi_symbol.y" , "ensembl_gene_id", "entrezgene_id", "overla
p_length")]


  colnames(Annot_DMR_DOCR_clean) = c("DMR_mm10_chr", "DMR_mm10_start", "DMR_mm10_end", "DMR_mm10_ID","m
m10_chr_annot_DMR" , "mm10_start_annot_DMR", "mm10_end_annot_DMR",
                                     "mm_10_strand_annot_DMR",
                                     "entrezgene_accession_annot_DMR", "gene_biotype_annot_DMR", "mgi_s
ymbol_annot_DMR", "ENSEMBL_ID_annot_DMR", "Entrez_ID_annot_DMR",
                                     "overlap_annot_DMR_length", "DMR_mm9_chr" , "DMR_mm9_start", "DMR_
mm9_end" , "DMR_CpG_nb", "DMR_pvalue" ,  "DMR_qvalue" , "DMR_meth.diff" ,
                                     "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length" ,"CGi
_cpgNum", "CGi_overlap",
                                     "DMR_DOCR_overlap", "DOCR_mm10_chr", "DOCR_mm10_start", "DOCR_mm1
0_end","DOCR_mm10_ID",
                                     "DOCR_mm9_chr", "DOCR_mm9_start", "DOCR_mm9_end", "DOCR_mm9_ID","
DOCR_cond1_rep1", "DOCR_cond1_rep2",
                                     "DOCR_cond2_rep1" , "DOCR_cond2_rep2",
                                     "DOCR_norm.cond1_rep1", "DOCR_norm.cond1_rep2",  "DOCR_norm.cond2
_rep1" , "DOCR_norm.cond2_rep2",  "DOCR_baseMean",  "DOCR_cond1" , "DOCR_cond2" ,
                                     "DOCR_FC",
                                     "DOCR_log2FoldChange", "DOCR_pvalue",  "DOCR_padj", "DOCR_tagwise
.dispersion", "DOCR_trended.dispersion",
                                     "strand_annot_DOCR", "mm10_chr_gene_annot_DOCR", "mm10_start_gene_
annot_DOCR","mm10_end_gene_annot_DOCR",
                                     "entrezgene_accession_annot_DOCR", "gene_biotype_annot_DOCR",
                                     "mgi_symbol_annot_DOCR" , "ensembl_gene_id_annot_DOCR", "entrezgen
e_id_annot_DOCR", "overlap_length_annot_DOCR")


  write.table(Annot_DMR_DOCR_clean, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croi
sement_autre_data/DMR_ATACseq/clean_annot",DMR_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names =
FALSE, col.names = TRUE)
  write.table(Annot_DMR_DOCR, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement
_autre_data/DMR_ATACseq/raw_annot",DMR_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names = FALSE,
col.names = TRUE)


  # To keep only gene names (entrezgene_accession_annot_DMR)
  # annotation of DMR - mm10 BIOMART :

  # To uncollapse data
  decondat_Annot_DMR_DOCR_clean = as.character(Annot_DMR_DOCR_clean$entrezgene_accession_annot_DMR)
  print(DMR_DOCR_sample)
  print(head(decondat_Annot_DMR_DOCR_clean))

  # To obtain list of gene names
  list_gene_splitted <- strsplit(as.character(Annot_DMR_DOCR_clean$entrezgene_accession_annot_DMR),spli
t=";")
  vec_gene_splitted <- unlist(list_gene_splitted)

  # To remove " " :
```

```
  # To remove "."
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  # To save data in a table
  table_VGS_filter = as.data.frame(unique(vec_gene_splitted_filter))

  write.table(table_VGS_filter, paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/DMR_ATACseq/list_gene",DMR_DOCR_sample, sep="_"), quote=FALSE, col.names=FALSE, row.names=FALSE)

}



rm(list = ls())

#-------------------
# Idem for distance=1000
# To load files or list of files
# To load DMR+DOCR regions
list_DMR_DOCR = list.files("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/DMR_ATACseq/")
list_DMR_DOCR_dist0 = list_DMR_DOCR[which(grepl("dist0",list_DMR_DOCR))]
list_DMR_DOCR_dist1000 = list_DMR_DOCR[which(grepl("dist1000",list_DMR_DOCR))]

# List of DOCR informations
# Since some files were empty (no overlap), we need to specify, in a list, each file we want to annotate.
list_DOCR_Annot=list("stats_mm9_mm10_down_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed","stats_mm9_mm10_down_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_down_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed", "stats_mm9_mm10_up_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_up_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed", "stats_mm9_mm10_up_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_up_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed")


# To load DMR432 annotation
DMR_Annot = read.table("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/DMR432_unique_annot_mm10_BMT_meth_cgi.bed", header=FALSE, sep="\t", quote="")
head(DMR_Annot)
dim(DMR_Annot)
colnames(DMR_Annot) = c("mm10_chr", "DMR_start", "DMR_end", "DMR_ID", "chr_annot", "start_annot", "end_annot", "strand",
                        "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID", "Entrez_ID", "overlap_annot_length",
                        "chr_mm9", "start_mm9", "end_mm9", "chr_mm10", "start_mm10", "end_mm10", "CpG_nb", "pvalue", "qvalue", "meth.diff",
                        "CGi_chr",  "CGi_start", "CGi_end", "CGi_name","CGi_length", "CGi_cpgNum", "CGi_overlap")

DMR_Annot_clean = DMR_Annot[,c("mm10_chr", "DMR_start", "DMR_end", "DMR_ID","chr_annot", "start_annot", "end_annot", "strand",
                               "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID", "Entrez_ID", "overlap_annot_length",
                               "chr_mm9", "start_mm9", "end_mm9", "CpG_nb", "pvalue", "qvalue", "meth.diff",
                               "CGi_chr",  "CGi_start", "CGi_end", "CGi_name","CGi_length", "CGi_cpgNum", "CGi_overlap")]


# for distance = 1000
for(i in 1:7){

  DMR_DOCR_sample = list_DMR_DOCR_dist1000[i]
  Annot_DMR_DOCR = read.table(file = paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/DMR_ATACseq/",DMR_DOCR_sample,sep="/"),
                              header = FALSE, sep = "\t", quote="", fill=TRUE)

  # To rename columns
  colnames(Annot_DMR_DOCR) <- c("DMR_chr","DMR_start", "DMR_end", "DOCR_chr", "DOCR_start", "DOCR_end", "DMR_DOCR_overlap")

  # To create 2 ID : DMR_ID et DOCR_ID
  Annot_DMR_DOCR$DMR_ID = paste(Annot_DMR_DOCR$DMR_chr, Annot_DMR_DOCR$DMR_start, Annot_DMR_DOCR$DMR_end, sep=";")
  Annot_DMR_DOCR$DOCR_ID = paste(Annot_DMR_DOCR$DOCR_chr, Annot_DMR_DOCR$DOCR_start, Annot_DMR_DOCR$DOCR_end, sep=";")
```

```
CR_end, sep= ; )

  # To load annotation informations of all DOCR
  DMR_sample <- DMR_Annot_clean

  Annot_DOCR_sample = list_DOCR_Annot[i]
  Annot_DOCR = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10",
Annot_DOCR_sample,sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DOCR) = c("mm9_chr", "mm9_start", "mm9_end", "mm9_ID", "mm10_chr", "mm10_start", "mm10
_end", "mm10_ID",
                           "chr_gene", "start_gene", "end_gene", "strand", "entrezgene_accession", "gen
e_biotype", "mgi_symbol",
                           "ensembl_gene_id", "entrezgene_id", "overlap_length", "cond1_rep1", "cond1_r
ep2", "cond2_rep1", "cond2_rep2",
                           "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2",
"baseMean", "cond1", "cond2", "FC",
                           "log2FoldChange", "pvalue", "padj", "tagwise.dispersion", "trended.dispersio
n")


  Annot_DOCR$DOCR_ID=paste(Annot_DOCR$mm10_chr, Annot_DOCR$mm10_start, Annot_DOCR$mm10_end,sep=";")

  Annot_DMR_DOCR=merge(Annot_DMR_DOCR, DMR_sample, by="DMR_ID")
  Annot_DMR_DOCR=merge(Annot_DMR_DOCR, Annot_DOCR, by="DOCR_ID")

  # To remove duplicate columns
  Annot_DMR_DOCR_clean = Annot_DMR_DOCR[,c("DMR_chr", "DMR_start.x", "DMR_end.x", "DMR_ID","chr_annot"
, "start_annot", "end_annot", "strand.x",
                                           "entrezgene_accession.x", "gene_biotype.x", "mgi_symbol.x",
"ENSEMBL_ID", "Entrez_ID",
                                           "overlap_annot_length", "chr_mm9" , "start_mm9", "end_mm9" ,
"CpG_nb", "pvalue.x" ,  "qvalue" , "meth.diff",
                                           "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length"
,"CGi_cpgNum", "CGi_overlap",
                                           "DMR_DOCR_overlap", "DOCR_chr", "DOCR_start", "DOCR_end","D
OCR_ID" ,
                                           "mm9_chr", "mm9_start", "mm9_end", "mm9_ID","cond1_rep1", "c
ond1_rep2", "cond2_rep1" , "cond2_rep2",
                                           "norm.cond1_rep1", "norm.cond1_rep2",  "norm.cond2_rep1" , "
norm.cond2_rep2",  "baseMean",  "cond1" , "cond2" ,  "FC",
                                           "log2FoldChange", "pvalue.y",  "padj", "tagwise.dispersion",
"trended.dispersion",
                                           "strand.y", "chr_gene", "start_gene", "end_gene", "entrezgen
e_accession.y", "gene_biotype.y",
                                           "mgi_symbol.y" , "ensembl_gene_id", "entrezgene_id", "overla
p_length")]



  colnames(Annot_DMR_DOCR_clean) = c("DMR_mm10_chr", "DMR_mm10_start", "DMR_mm10_end", "DMR_mm10_ID","m
m10_chr_annot_DMR" , "mm10_start_annot_DMR", "mm10_end_annot_DMR",
                                   "mm_10_strand_annot_DMR",
                                   "entrezgene_accession_annot_DMR", "gene_biotype_annot_DMR", "mgi_s
ymbol_annot_DMR", "ENSEMBL_ID_annot_DMR", "Entrez_ID_annot_DMR",
                                   "overlap_annot_DMR_length", "DMR_mm9_chr" , "DMR_mm9_start", "DMR_
mm9_end" , "DMR_CpG_nb", "DMR_pvalue" ,  "DMR_qvalue" , "DMR_meth.diff" ,
                                   "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length" ,"CGi
_cpgNum", "CGi_overlap",
                                   "DMR_DOCR_overlap", "DOCR_mm10_chr", "DOCR_mm10_start", "DOCR_mm1
0_end","DOCR_mm10_ID",
                                   "DOCR_mm9_chr", "DOCR_mm9_start", "DOCR_mm9_end", "DOCR_mm9_ID","
DOCR_cond1_rep1", "DOCR_cond1_rep2",
                                   "DOCR_cond2_rep1" , "DOCR_cond2_rep2",
                                   "DOCR_norm.cond1_rep1", "DOCR_norm.cond1_rep2",  "DOCR_norm.cond2
_rep1" , "DOCR_norm.cond2_rep2",  "DOCR_baseMean",  "DOCR_cond1" , "DOCR_cond2" ,
                                   "DOCR_FC",
                                   "DOCR_log2FoldChange", "DOCR_pvalue",  "DOCR_padj", "DOCR_tagwise
.dispersion", "DOCR_trended.dispersion",
                                   "strand_annot_DOCR", "mm10_chr_gene_annot_DOCR", "mm10_start_gene_
annot_DOCR","mm10_end_gene_annot_DOCR",
                                   "entrezgene_accession_annot_DOCR", "gene_biotype_annot_DOCR",
                                   "mgi_symbol_annot_DOCR" , "ensembl_gene_id_annot_DOCR", "entrezgen
e_id_annot_DOCR", "overlap_length_annot_DOCR")
```

```
  write.table(Annot_DMR_DOCR_clean, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croi
sement_autre_data/DMR_ATACseq/clean_annot",DMR_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names =
FALSE, col.names = TRUE)
  write.table(Annot_DMR_DOCR, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement
_autre_data/DMR_ATACseq/raw_annot",DMR_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names = FALSE,
col.names = TRUE)


  # To keep only gene names (entrezgene_accession_annot_DMR) - annotation of DMR mm10 BIOMART :

  # To uncollapse data
  decondat_Annot_DMR_DOCR_clean = as.character(Annot_DMR_DOCR_clean$entrezgene_accession_annot_DMR)
  print(DMR_DOCR_sample)
  print(head(decondat_Annot_DMR_DOCR_clean))

  # To obtain list of genes
  list_gene_splitted <- strsplit(as.character(Annot_DMR_DOCR_clean$entrezgene_accession_annot_DMR),spli
t=";")
  vec_gene_splitted <- unlist(list_gene_splitted)

  # To remove "." :
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  # To save data in a table
  table_VGS_filter = as.data.frame(unique(vec_gene_splitted_filter))

  write.table(table_VGS_filter, paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_au
tre_data/DMR_ATACseq/list_gene",DMR_DOCR_sample, sep="_"), quote=FALSE, col.names=FALSE, row.names=FALS
E)

}
```

Integration of DMR with DEG

To determine whether DMRs are located at genes that are differentially expressed (DEG) during physiological brain development, DMRs and DEG regions were compared using our own R function called *find_overlaps_AD_table()*. Both overlaping regions and regions less 1000 bases apart were investigated, by setting *tolerance* argument to 0 or 1000. The obtained regions were then annotated, using annotation that was already achieved in the two complete datasets.

In [ ]:

```
%%R

#################################################################
# Integration of DMR and DEG to find potential overlap
# (using mm10 chromosomal coordinates)
#################################################################

# 1.Load DMR432 data - mm10
DMR432 = read.table("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/DMR432
_unique_annot_mm10_BMT_meth_cgi.bed", sep="\t", header = FALSE)
head(DMR432)

# To extract DMR432 chromosomic coordinates
DMR432_bed = DMR432[,c("V1","V2","V3")]
head(DMR432_bed)
dim(DMR432_bed)


# 2. Load DEG data - mm10 (from edgeR analysis):
nom_sample = list=c("down_E13.14", "down_E14.15", "down_E14.16", "down_E15.16",
                    "down_E16.P0", "up_E13.14", "up_E14.15", "up_E14.16",
                    "up_E15.16", "up_E16.P0" )

# /!\ /!\ Dataset must be in a matrix format, otherwise, function doesn't work /!\/!\
for(name in nom_sample){
  assign(x=name, value = as.matrix(read.table(paste("G:/projet_DU_AD/results/09_annotation/Sartools/Bio
mart/bed_edge_AnnotBM_et_Man_ech_",
                                                   name , "_mm10.sorted.bed", sep=""), sep = "\t", header= FALSE
, dec=".", quote="")))
  print(paste("head_",name, sep=""))
  print(head(eval(parse(text = name))))
  print(paste("dimension du dataset ",name, sep = ""))
```

```
  print(dim(eval(parse(text = name))))
}

# Creation of a list containing each DEG file (up or down, each developmental stage)
RNA = list(downE13.14 = down_E13.14, downE14.15 = down_E14.15, downE15.16 = down_E15.16, downE14.16 = d
own_E14.16, downE16.P0 = down_E16.P0,
           upE13.14 = up_E13.14, upE14.15 = up_E14.15 ,upE15.16 = up_E15.16, upE14.16 = up_E14.16, upE1
6.P0 = up_E16.P0)

for (i in 1:length(RNA)){
  print(names(RNA)[i])
  print(head(RNA[[i]]))
  print(dim(RNA[[i]]))
}


# 3. Function to find overlap or close regions (depending on value given for tolerance parameter)
find_overlaps_AD_table <- function (regions_1,regions_2,tolerance=0)
{
  common_region_table = NULL
  number_of_regions_1 <- nrow(regions_1)
  number_of_regions_2 <- nrow(regions_2)

  overlapped_is <- c()
  overlapped_js <- c()

  for (i in 1:number_of_regions_1) {

    for (j in 1:number_of_regions_2) {

      if (as.character(regions_1[i,1])==as.character(regions_2[j,1])) {
        start_1 <- as.numeric(regions_1[i,2])
        end_1 <- as.numeric(regions_1[i,3])
        expanded_start_1 <- start_1 - tolerance
        expanded_end_1 <- end_1 + tolerance
        start_2 <- as.numeric(regions_2[j,2])
        end_2 <- as.numeric(regions_2[j,3])

        if (expanded_end_1>=start_2) {

          if (end_2>=expanded_start_1) {

            overlapped_is <- c(overlapped_is,i)
            overlapped_js <- c(overlapped_js,j)

            first_region <- paste(regions_1[i,1],start_1,end_1,sep="_")
            second_region <- paste(regions_2[j,1],start_2,end_2,sep="_")
            overlap_size <- min(end_1,end_2)-max(start_1,start_2)+1
            common_region_table = rbind(common_region_table,c(as.character(regions_1[i,1]),start_1,end_
1,
                                                               as.character(regions_2[j,1]),start_2,end_
2, overlap_size))

            if (overlap_size>0) {
              print(paste(first_region,second_region,sep=" "))
            } else {
              print(paste(first_region,second_region, sep=" "))
            }
          }
        }
      }
    }
  }
  print(paste("Number matched in first data set:",length(unique(overlapped_is))))
  print(paste("Number matched in second data set:",length(unique(overlapped_js))))
  if (length(common_region_table) > 0){
    colnames(common_region_table) = c("chr_reg1", "start_reg1", "end_reg1", "chr_reg2", "start_reg2", "
end_reg2","overlap_size")
  }else{
  }
  return(as.data.frame(common_region_table))
}



# 4. Search for common regions between DMR and DEG - overlap
for (i in 1:length(RNA)) {
```

```
  print(names(RNA)[i])
  DMR_RNA = find_overlaps_AD_table(DMR432_bed, RNA[[i]], tolerance=0)
  write.table(DMR_RNA,paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/m
m10_DMR432", names(RNA)[i], "dist0.txt",sep="_"), sep = "\t", quote=FALSE, row.names = FALSE, col.names
=FALSE)
}

# 5. Search for common regions between DMR and DEG - maximum distance between 2 regions = 1000bases:
for (i in 1:length(RNA)) {
  print(names(RNA)[i])
  DMR_RNA = find_overlaps_AD_table(DMR432_bed, RNA[[i]], tolerance=1000)
  write.table(DMR_RNA,paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/m
m10_DMR432", names(RNA)[i], "dist1000.txt",sep="_"), sep = "\t", quote=FALSE, row.names = FALSE, col.na
mes=FALSE)
}



##################
# Annotation
##################

# Load files or list of files
# Load DMR+DEG regions
list_DMR_DEG = list.files("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/DMR
_RNAseq/")
list_DMR_DEG_dist0 = list_DMR_DEG[which(grepl("dist0",list_DMR_DEG))]
list_DMR_DEG_dist1000 = list_DMR_DEG[which(grepl("dist1000",list_DMR_DEG))]

# Load DEG annotation
list_files_DEG = list.files("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/")
list_DEG_Annot = list_files_DEG[which(grepl("edgeR_AnnotBM_et_Man",list_files_DEG))]

# Load DMR432 annotation
DMR_Annot = read.table("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/DMR
432_unique_annot_mm10_BMT_meth_cgi.bed", header=FALSE, sep="\t", quote="")
head(DMR_Annot)
dim(DMR_Annot)
colnames(DMR_Annot) = c("mm10_chr", "DMR_start", "DMR_end", "DMR_ID", "chr_annot", "start_annot", "end_
annot", "strand",
                        "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID", "Entrez_ID"
, "overlap_annot_length",
                        "chr_mm9", "start_mm9", "end_mm9", "chr_mm10", "start_mm10", "end_mm10", "CpG_n
b", "pvalue", "qvalue", "meth.diff",
                        "CGi_chr",  "CGi_start", "CGi_end", "CGi_name","CGi_length", "CGi_cpgNum", "CGi
_overlap")

DMR_Annot_clean = DMR_Annot[,c("mm10_chr", "DMR_start", "DMR_end", "DMR_ID","chr_annot", "start_annot",
"end_annot", "strand",
                               "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID", "Ent
rez_ID", "overlap_annot_length",
                               "chr_mm9", "start_mm9", "end_mm9", "CpG_nb", "pvalue", "qvalue", "meth.d
iff",
                               "CGi_chr",  "CGi_start", "CGi_end", "CGi_name","CGi_length", "CGi_cpgNum
", "CGi_overlap")]


# for distance=0
for(i in 1:10){

  DMR_DEG_sample = list_DMR_DEG_dist0[i]
  Annot_DMR_DEG = read.table(file = paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisemen
t_autre_data/DMR_RNAseq/",DMR_DEG_sample,sep="/"),
                             header = FALSE, sep = "\t", quote="", fill=TRUE)

  # Change column names
  colnames(Annot_DMR_DEG) <- c("DMR_chr","DMR_start", "DMR_end", "DEG_chr", "DEG_start", "DEG_end", "DM
R_DEG_overlap")

  # Create ID : DMR_ID et DEG_ID
  Annot_DMR_DEG$DMR_ID = paste(Annot_DMR_DEG$DMR_chr, Annot_DMR_DEG$DMR_start, Annot_DMR_DEG$DMR_end, s
ep=";")
  Annot_DMR_DEG$DEG_ID = paste(Annot_DMR_DEG$DEG_chr, Annot_DMR_DEG$DEG_start, Annot_DMR_DEG$DEG_end, s
ep=";")

  # Load Annot of DEG
  DMR_sample <- DMR_Annot_clean
```

```
  DEG_sample <- list_DEG_Annot[i]
  Annot_DEG = read.table(file = paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart",DEG_samp
le,sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DEG)= c("ENS_ID", "cond1_rep1", "cond1_rep2", "cond2_rep1", "cond2_rep2",
                         "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2",
                         "baseMean", "cond1", "cond2", "FC", "log2FoldChange", "pvalue", "padj", "tagwi
se.dispersion",
                         "trended.dispersion", "mm10_chr", "strand", "mm10_start_gene", "mm10_end_gene"
, "entrezgene_id", "gene_biotype",
                         "mgi_symbol", "entrezgene_accession", "entrezgene_description", "uniprot_gn_sy
mbol")

  Annot_DEG$DEG_ID=paste(paste("chr",Annot_DEG$mm10_chr,sep=""), Annot_DEG$mm10_start_gene, Annot_DEG$m
m10_end_gene,sep=";")

  Annot_DMR_DEG=merge(Annot_DMR_DEG, DMR_sample, by="DMR_ID")
  Annot_DMR_DEG=merge(Annot_DMR_DEG, Annot_DEG, by="DEG_ID")

  # Remove duplicate columns
  Annot_DMR_DEG_clean = Annot_DMR_DEG[,c("DMR_chr", "DMR_start.x", "DMR_end.x", "DMR_ID", "chr_annot" ,
"start_annot", "end_annot", "strand.x",
                                       "entrezgene_accession.x", "gene_biotype.x", "mgi_symbol.x", "E
NSEMBL_ID", "Entrez_ID",
                                       "overlap_annot_length", "chr_mm9" , "start_mm9", "end_mm9" , "
CpG_nb", "pvalue.x" ,  "qvalue" , "meth.diff" ,
                                       "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length" ,
"CGi_cpgNum", "CGi_overlap",
                                       "DMR_DEG_overlap", "DEG_chr", "DEG_start", "DEG_end","DEG_ID"
,   "ENS_ID" , "cond1_rep1", "cond1_rep2", "cond2_rep1" , "cond2_rep2",
                                       "norm.cond1_rep1", "norm.cond1_rep2",  "norm.cond2_rep1" , "no
rm.cond2_rep2",  "baseMean",  "cond1" , "cond2" ,  "FC"  ,
                                        "log2FoldChange", "pvalue.y",  "padj", "tagwise.dispersion",
"trended.dispersion", "strand.y" ,
                                        "mm10_start_gene","mm10_end_gene", "entrezgene_id", "gene_bio
type.y" , "mgi_symbol.y" , "entrezgene_accession.y", "entrezgene_description",
                                      "uniprot_gn_symbol")]

  colnames(Annot_DMR_DEG_clean) = c("DMR_mm10_chr", "DMR_mm10_start", "DMR_mm10_end", "DMR_mm10_ID", "m
m10_chr_gene_annot_DMR" , "mm10_start_gene_annot_DMR", "mm10_end_gene_annot_DMR",
                                  "strand_gene_annot_DMR", "entrezgene_accession_annot_DMR", "gene_bio
type_annot_DMR", "mgi_symbol_annot_DMR", "ENSEMBL_ID_annot_DMR", "Entrez_ID_annot_DMR",
                                  "overlap_annot_DMR_length", "DMR_mm9_chr" , "DMR_mm9_start", "
DMR_mm9_end" , "DMR_CpG_nb", "DMR_pvalue" ,  "DMR_qvalue" , "DMR_meth.diff" ,
                                  "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length" ,
"CGi_cpgNum", "CGi_overlap",
                                  "DMR_DEG_overlap", "DEG_mm10_chr", "DEG_mm10_start", "DEG_mm10
_end","DEG_mm10_ID" ,   "DEG_ENS_ID" , "DEG_cond1_rep1", "DEG_cond1_rep2",
                                "DEG_cond2_rep1" , "DEG_cond2_rep2",
                                  "DEG_norm.cond1_rep1", "DEG_norm.cond1_rep2",  "DEG_norm.cond2
_rep1" , "DEG_norm.cond2_rep2",  "DEG_baseMean",  "DEG_cond1" , "DEG_cond2" ,
                                "DEG_FC",
                                  "DEG_log2FoldChange", "DEG_pvalue",  "DEG_padj", "DEG_tagwise.
dispersion", "DEG_trended.dispersion", "DEG_strand" ,
                                  "mm10_start_gene_annot_DEG","mm10_end_gene_annot_DEG", "entrez
gene_id_annot_DEG", "gene_biotype_annot_DEG" , "mgi_symbol_annot_DEG" ,
                                "entrezgene_accession_annot_DEG", "entrezgene_description_annot_DEG"
,
                                  "uniprot_gn_symbol_annot_DEG")



  write.table(Annot_DMR_DEG_clean, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/crois
ement_autre_data/DMR_RNAseq/clean_annot",DMR_DEG_sample,sep="_"), quote=FALSE, sep="\t", row.names = FA
LSE, col.names = TRUE)
  write.table(Annot_DMR_DEG, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_
autre_data/DMR_RNAseq/raw_annot",DMR_DEG_sample,sep="_"), quote=FALSE, sep="\t", row.names = FALSE, col
.names = TRUE)


  # To keep only gene names (entrezgene_accession_annot_DMR) - annotation of DMR - mm10 - BIOMART  :

  # To uncollapse data
  decondat_Annot_DMR_DEG_clean = as.character(Annot_DMR_DEG_clean$entrezgene_accession_annot_DMR)
  print(DMR_DEG_sample)
  print(head(decondat_Annot_DMR_DEG_clean))
```

```
  # To split gene names, when necessary
  list_gene_splitted <- strsplit(as.character(Annot_DMR_DEG_clean$entrezgene_accession_annot_DMR),split
=";")
  vec_gene_splitted <- unlist(list_gene_splitted)

  # To remove "." :
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  #To save data in a table
  table_VGS_filter = as.data.frame(unique(vec_gene_splitted_filter))

  write.table(table_VGS_filter, paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_au
tre_data/DMR_RNAseq/list_gene",DMR_DEG_sample, sep="_"), quote=FALSE, col.names=FALSE, row.names=FALSE)

}



rm(list = ls())


#-------------------------
# Idem for distance=1000
# Load files or list of files
# Load DMR+DEG regions
list_DMR_DEG = list.files("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/DMR
_RNAseq/")
list_DMR_DEG_dist0 = list_DMR_DEG[which(grepl("dist0",list_DMR_DEG))]
list_DMR_DEG_dist1000 = list_DMR_DEG[which(grepl("dist1000",list_DMR_DEG))]

# Load DEG annotation
list_files_DEG = list.files("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart/")
list_DEG_Annot = list_files_DEG[which(grepl("edgeR_AnnotBM_et_Man",list_files_DEG))]

# Load DMR432 annotation
DMR_Annot = read.table("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/DMR
432_unique_annot_mm10_BMT_meth_cgi.bed", header=FALSE, sep="\t", quote="")
head(DMR_Annot)
dim(DMR_Annot)
colnames(DMR_Annot) = c("mm10_chr", "DMR_start", "DMR_end", "DMR_ID", "chr_annot", "start_annot", "end_
annot", "strand",
                        "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID", "Entrez_ID"
, "overlap_annot_length",
                        "chr_mm9", "start_mm9", "end_mm9", "chr_mm10", "start_mm10", "end_mm10", "CpG_n
b", "pvalue", "qvalue", "meth.diff",
                        "CGi_chr",  "CGi_start", "CGi_end", "CGi_name","CGi_length", "CGi_cpgNum", "CGi
_overlap")

DMR_Annot_clean = DMR_Annot[,c("mm10_chr", "DMR_start", "DMR_end", "DMR_ID","chr_annot", "start_annot",
"end_annot", "strand",
                               "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID", "Ent
rez_ID", "overlap_annot_length",
                               "chr_mm9", "start_mm9", "end_mm9", "CpG_nb", "pvalue", "qvalue", "meth.d
iff",
                               "CGi_chr",  "CGi_start", "CGi_end", "CGi_name","CGi_length", "CGi_cpgNum
", "CGi_overlap")]


for(i in 1:10){

  DMR_DEG_sample = list_DMR_DEG_dist1000[i]
  Annot_DMR_DEG = read.table(file = paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisemen
t_autre_data/DMR_RNAseq/",DMR_DEG_sample,sep="/"),
                             header = FALSE, sep = "\t", quote="", fill=TRUE)

  # To rename columns
  colnames(Annot_DMR_DEG) <- c("DMR_chr","DMR_start", "DMR_end", "DEG_chr", "DEG_start", "DEG_end", "DM
R_DEG_overlap")

  # To create 2 ID : DMR_ID et DEG_ID
  Annot_DMR_DEG$DMR_ID = paste(Annot_DMR_DEG$DMR_chr, Annot_DMR_DEG$DMR_start, Annot_DMR_DEG$DMR_end, s
ep=";")
  Annot_DMR_DEG$DEG_ID = paste(Annot_DMR_DEG$DEG_chr, Annot_DMR_DEG$DEG_start, Annot_DMR_DEG$DEG_end, s
ep=";")
```

```
  # Load Annotation for DEG
  DMR_sample <- DMR_Annot_clean
  DEG_sample <- list_DEG_Annot[i]
  Annot_DEG = read.table(file = paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart",DEG_samp
le,sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DEG)= c("ENS_ID", "cond1_rep1", "cond1_rep2", "cond2_rep1", "cond2_rep2",
                         "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2",
                         "baseMean", "cond1", "cond2", "FC", "log2FoldChange", "pvalue", "padj", "tagwi
se.dispersion",
                         "trended.dispersion", "mm10_chr", "strand", "mm10_start_gene", "mm10_end_gene"
, "entrezgene_id", "gene_biotype",
                         "mgi_symbol", "entrezgene_accession", "entrezgene_description", "uniprot_gn_sy
mbol")

  Annot_DEG$DEG_ID=paste(paste("chr",Annot_DEG$mm10_chr,sep=""), Annot_DEG$mm10_start_gene, Annot_DEG$m
m10_end_gene,sep=";")

  Annot_DMR_DEG=merge(Annot_DMR_DEG, DMR_sample, by="DMR_ID")
  Annot_DMR_DEG=merge(Annot_DMR_DEG, Annot_DEG, by="DEG_ID")

  # To remove duplicate columns
  Annot_DMR_DEG_clean = Annot_DMR_DEG[,c("DMR_chr", "DMR_start.x", "DMR_end.x", "DMR_ID", "chr_annot" ,
"start_annot", "end_annot", "strand.x",
                                         "entrezgene_accession.x", "gene_biotype.x", "mgi_symbol.x", "E
NSEMBL_ID", "Entrez_ID",
                                         "overlap_annot_length", "chr_mm9" , "start_mm9", "end_mm9" , "
CpG_nb", "pvalue.x" ,  "qvalue" , "meth.diff" ,
                                         "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length" ,
"CGi_cpgNum", "CGi_overlap",
                                         "DMR_DEG_overlap", "DEG_chr", "DEG_start", "DEG_end","DEG_ID"
,   "ENS_ID" , "cond1_rep1", "cond1_rep2", "cond2_rep1" , "cond2_rep2",
                                         "norm.cond1_rep1", "norm.cond1_rep2",  "norm.cond2_rep1" , "no
rm.cond2_rep2",  "baseMean",  "cond1" , "cond2" ,  "FC"  ,
                                         "log2FoldChange", "pvalue.y",  "padj", "tagwise.dispersion", "
trended.dispersion", "strand.y" ,
                                         "mm10_start_gene","mm10_end_gene", "entrezgene_id", "gene_biot
ype.y" , "mgi_symbol.y" , "entrezgene_accession.y", "entrezgene_description",
                                         "uniprot_gn_symbol")]

  colnames(Annot_DMR_DEG_clean) = c("DMR_mm10_chr", "DMR_mm10_start.x", "DMR_mm10_end.x", "DMR_mm10_ID"
, "mm10_chr_gene_annot_DMR" , "mm10_start_gene_annot_DMR", "mm10_end_gene_annot_DMR",
                                    "strand_gene_annot_DMR", "entrezgene_accession_annot_DMR", "gene_bi
otype_annot_DMR", "mgi_symbol_annot_DMR", "ENSEMBL_ID_annot_DMR", "Entrez_ID_annot_DMR",
                                    "overlap_annot_DMR_length", "DMR_mm9_chr" , "DMR_mm9_start", "DMR_m
m9_end" , "DMR_CpG_nb", "DMR_pvalue" ,  "DMR_qvalue" , "DMR_meth.diff" ,
                                    "CGi_chr" ,"CGi_start" , "CGi_end" ,"CGi_name", "CGi_length" ,"CGi_
cpgNum", "CGi_overlap",
                                    "DMR_DEG_overlap", "DEG_mm10_chr", "DEG_mm10_start", "DEG_mm10_end"
,"DEG_mm10_ID" ,   "DEG_ENS_ID" , "DEG_cond1_rep1", "DEG_cond1_rep2",
                                    "DEG_cond2_rep1" , "DEG_cond2_rep2",
                                    "DEG_norm.cond1_rep1", "DEG_norm.cond1_rep2",  "DEG_norm.cond2_rep1
" , "DEG_norm.cond2_rep2",  "DEG_baseMean",  "DEG_cond1" , "DEG_cond2" ,
                                    "DEG_FC",
                                    "DEG_log2FoldChange", "DEG_pvalue",  "DEG_padj", "DEG_tagwise.dispe
rsion", "DEG_trended.dispersion", "DEG_strand" ,
                                    "mm10_start_gene_annot_DEG","mm10_end_gene_annot_DEG", "entrezgene_
id_annot_DEG", "gene_biotype_annot_DEG" , "mgi_symbol_annot_DEG" ,
                                    "entrezgene_accession_annot_DEG", "entrezgene_description_annot_DEG
",
                                    "uniprot_gn_symbol_annot_DEG")



  write.table(Annot_DMR_DEG_clean, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/crois
ement_autre_data/DMR_RNAseq/clean_annot",DMR_DEG_sample,sep="_"), quote=FALSE, sep="\t", row.names = FA
LSE, col.names = TRUE)
  write.table(Annot_DMR_DEG, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_
autre_data/DMR_RNAseq/raw_annot",DMR_DEG_sample,sep="_"), quote=FALSE, sep="\t", row.names = FALSE, col
.names = TRUE)

  # To keep only gene names (entrezgene_accession_annot_DMR) - annotation of DMR - mm10 BIOMART :

  # To uncollapse data
  decondat_Annot_DMR_DEG_clean = as.character(Annot_DMR_DEG_clean$entrezgene_accession_annot_DMR)
  print(DMR_DEG_sample)
  print(head(decondat_Annot_DMR_DEG_clean))
```

```
  # To split gene names, when necessary
  list_gene_splitted <- strsplit(as.character(Annot_DMR_DEG_clean$entrezgene_accession_annot_DMR),split
=";")
  vec_gene_splitted <- unlist(list_gene_splitted)

  # To remove "." :
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  # To save data in a table
   table_VGS_filter = as.data.frame(unique(vec_gene_splitted_filter))

  write.table(table_VGS_filter, paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_au
tre_data/DMR_RNAseq/list_gene",DMR_DEG_sample, sep="_"), quote=FALSE, col.names=FALSE, row.names=FALSE)

}
```

Integration of DEG with DOCR

To determine whether DEG are also DOCR during physiological brain development or not, DEG and DOCR regions were compared using our own R function called *find_overlaps_AD_table()*. Both overlaping regions and regions less 1000 bases apart were investigated, by setting *tolerance* argument to 0 or 1000. The obtained regions were then annotated, using annotation that was already achieved in the two complete datasets.

In [ ]:

```
%%R

#####################################################################
# Integration of DEG and DOCR to find potential overlap
# (mm10 chromosomal coordinates)
#####################################################################

# 1. Function to find overlap or close regions (depending on value given for tolerance parameter)
find_overlaps_AD_table <- function (regions_1,regions_2,tolerance=0)
{
  common_region_table = NULL
  number_of_regions_1 <- nrow(regions_1)
  number_of_regions_2 <- nrow(regions_2)

  overlapped_is <- c()
  overlapped_js <- c()

  for (i in 1:number_of_regions_1) {

    for (j in 1:number_of_regions_2) {

      if (as.character(regions_1[i,1])==as.character(regions_2[j,1])) {
        start_1 <- as.numeric(regions_1[i,2])
        end_1 <- as.numeric(regions_1[i,3])
        expanded_start_1 <- start_1 - tolerance
        expanded_end_1 <- end_1 + tolerance
        start_2 <- as.numeric(regions_2[j,2])
        end_2 <- as.numeric(regions_2[j,3])

        if (expanded_end_1>=start_2) {

          if (end_2>=expanded_start_1) {

            overlapped_is <- c(overlapped_is,i)
            overlapped_js <- c(overlapped_js,j)

            first_region <- paste(regions_1[i,1],start_1,end_1,sep="_")
            second_region <- paste(regions_2[j,1],start_2,end_2,sep="_")
            overlap_size <- min(end_1,end_2)-max(start_1,start_2)+1
            common_region_table = rbind(common_region_table,c(as.character(regions_1[i,1]),start_1,end_
1,
                                                               as.character(regions_2[j,1]),start_2,end_
2, overlap_size))

            if (overlap_size>0) {
              print(paste(first_region,second_region,sep=" "))
            } else {
              print(paste(first_region,second_region, sep=" "))
```

```
            }
          }
        }
      }
    }
  }
  print(paste("Number matched in first data set:",length(unique(overlapped_is))))
  print(paste("Number matched in second data set:",length(unique(overlapped_js))))
  if (length(common_region_table) > 0){
    colnames(common_region_table) = c("chr_reg1", "start_reg1", "end_reg1", "chr_reg2", "start_reg2", "
end_reg2","overlap_size")
  }else{
  }
  return(as.data.frame(common_region_table))
}



# 2. Load DOCR and DEG data - mm10 (from edgeR analysis):
nom_sample_ATAC = list("down_DOCR_E13.14", "down_DOCR_E14.15", "down_DOCR_E14.16", "down_DOCR_E15.16",
"down_DOCR_E16.P0",
                       "up_DOCR_E13.14", "up_DOCR_E14.15", "up_DOCR_E14.16", "up_DOCR_E15.16", "up_DOCR_E
16.P0" )
nom_sample_RNA = list("down_E13.14", "down_E14.15", "down_E14.16", "down_E15.16", "down_E16.P0", "up_E1
3.14", "up_E14.15",
                      "up_E14.16", "up_E15.16", "up_E16.P0" )

# /!\ /!\ Dataset must be in a matrix format, otherwise function doesn't work /!\/!\
for(i in 1:length(nom_sample_ATAC)){
  assign(x=nom_sample_ATAC[[i]], value = as.matrix(read.table(paste("G:/projet_DU_DSD/results/10_annota
tions/Sartools/Biomart/mm10/bed_gff_ID_", nom_sample_ATAC[[i]], "_mm10.sorted.bed",
                                                             sep=""), sep = "\t", header= FALSE, dec=".", quote=
"")))
  assign(x=nom_sample_RNA[[i]], value = as.matrix(read.table(paste("G:/projet_DU_AD/results/09_annotati
on/Sartools/Biomart/bed_edge_AnnotBM_et_Man_ech_",
                                                                 nom_sample_RNA[[i]] , "_mm10.sorted.
bed", sep=""), sep = "\t", header= FALSE, dec=".", quote="")))


  print(paste("dimension du dataset ",nom_sample_ATAC[[i]], sep = ""))
  print(dim(eval(parse(text = nom_sample_ATAC[[i]]))))

  print(paste("dimension du dataset ",nom_sample_RNA[[i]], sep = ""))
  print(dim(eval(parse(text = nom_sample_RNA[[i]]))))

  # 3. Search for common regions between DEG and DOCR - overlap
  print("distance = 0")
  recoupement = find_overlaps_AD_table(get(nom_sample_RNA[[i]]), get(nom_sample_ATAC[[i]]), tolerance=0
)
  write.table(recoupement,paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_da
ta/mm10_RNA_ATAC",
                                nom_sample_RNA[[i]], "dist0.txt",sep="_"), sep = "\t", quote=FALSE, row
.names = FALSE, col.names=FALSE)

  # 4. Search for common regions between DEG and DOCR - maximum distance between 2 regions = 1000bases:

  print("distance = 1000")
  recoupement_large = find_overlaps_AD_table(get(nom_sample_RNA[[i]]), get(nom_sample_ATAC[[i]]), toler
ance=1000)
  write.table(recoupement_large,paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_au
tre_data/mm10_RNA_ATAC",
                                      nom_sample_RNA[[i]], "dist1000.txt",sep="_"), sep = "\t", quote=F
ALSE, row.names = FALSE, col.names=FALSE)

}



##################
# Annotation
##################

# Load files containing DEG+DOCR regions
list_DEG_DOCR = list.files("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RN
A_ATAC/")
list_DEG_DOCR_dist0 = list_DEG_DOCR[which(grepl("dist0",list_DEG_DOCR))]
list_DEG_DOCR_dist1000 = list_DEG_DOCR[which(grepl("dist1000",list_DEG_DOCR))]
```

```
# List of files containing DOCR informations
# Since some files were empty (no overlap), we need to specify, in a list, each file we want to annotat
e.
list_DOCR_Annot=list("stats_mm9_mm10_down_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed","stats_mm9_mm1
0_down_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_down_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_down_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed", "stats_mm9_mm
10_up_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_up_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed", "stats_mm9_mm10
_up_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_up_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed")


# Load DEG annotation files
# Since some files were empty (no overlap), we need to specify, in a list, each file we want to annotat
e.
list_DEG_Annot = list("edgeR_AnnotBM_et_Man_down_E14.15_mm10.txt", "edgeR_AnnotBM_et_Man_down_E14.16_mm
10.txt" ,
                      "edgeR_AnnotBM_et_Man_down_E15.16_mm10.txt",
                      "edgeR_AnnotBM_et_Man_down_E16.P0_mm10.txt", "edgeR_AnnotBM_et_Man_up_E14.15_mm10
.txt",
                      "edgeR_AnnotBM_et_Man_up_E14.16_mm10.txt", "edgeR_AnnotBM_et_Man_up_E15.16_mm10.t
xt",
                      "edgeR_AnnotBM_et_Man_up_E16.P0_mm10.txt")

# for distance=0
for(i in 1:8){

  DEG_DOCR_sample = list_DEG_DOCR_dist0[i]
  Annot_DEG_DOCR = read.table(file = paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croiseme
nt_autre_data/RNA_ATAC/",DEG_DOCR_sample,sep="/"),
                              header = FALSE, sep = "\t", quote="", fill=TRUE)

  # To rename columns
  colnames(Annot_DEG_DOCR) <- c("DEG_chr","DEG_start", "DEG_end", "DOCR_chr", "DOCR_start", "DOCR_end"
, "DEG_DOCR_overlap")

  # To create two ID : DEG_ID et DOCR_ID
  Annot_DEG_DOCR$DEG_ID = paste(Annot_DEG_DOCR$DEG_chr, Annot_DEG_DOCR$DEG_start, Annot_DEG_DOCR$DEG_e
nd, sep=";")
  Annot_DEG_DOCR$DOCR_ID = paste(Annot_DEG_DOCR$DOCR_chr, Annot_DEG_DOCR$DOCR_start, Annot_DEG_DOCR$DO
CR_end, sep=";")

  # Load annotation of all DEG regions
  DEG_sample <- list_DEG_Annot[i]
  Annot_DEG = read.table(file = paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart",DEG_samp
le,sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DEG)= c("ENS_ID", "cond1_rep1", "cond1_rep2", "cond2_rep1", "cond2_rep2",
                         "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2",
                         "baseMean", "cond1", "cond2", "FC", "log2FoldChange", "pvalue", "padj", "tagwi
se.dispersion",
                         "trended.dispersion", "mm10_chr", "strand", "mm10_start_gene", "mm10_end_gene"
, "entrezgene_id", "gene_biotype",
                         "mgi_symbol", "entrezgene_accession", "entrezgene_description", "uniprot_gn_sy
mbol")

  Annot_DEG$DEG_ID=paste(paste("chr",Annot_DEG$mm10_chr,sep=""), Annot_DEG$mm10_start_gene, Annot_DEG$m
m10_end_gene,sep=";")


  # Load annoation of all DOCR
  Annot_DOCR_sample = list_DOCR_Annot[i]
  Annot_DOCR = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10",
Annot_DOCR_sample,sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DOCR) = c("mm9_chr", "mm9_start", "mm9_end", "mm9_ID", "mm10_chr", "mm10_start", "mm10
_end", "mm10_ID",
                           "chr_gene", "start_gene", "end_gene", "strand", "entrezgene_accession", "gen
e_biotype", "mgi_symbol",
                           "ensembl_gene_id", "entrezgene_id", "overlap_length", "cond1_rep1", "cond1_r
ep2", "cond2_rep1", "cond2_rep2",
                           "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2",
"baseMean", "cond1", "cond2", "FC",
                           "log2FoldChange", "pvalue", "padj", "tagwise.dispersion", "trended.dispersio
n")
```

```
''')


  Annot_DOCR$DOCR_ID=paste(Annot_DOCR$mm10_chr, Annot_DOCR$mm10_start, Annot_DOCR$mm10_end,sep=";")

  Annot_DEG_DOCR=merge(Annot_DEG_DOCR, Annot_DEG, by="DEG_ID")
  Annot_DEG_DOCR=merge(Annot_DEG_DOCR, Annot_DOCR, by="DOCR_ID")

  # To remove duplicate columns
  Annot_DEG_DOCR_clean = Annot_DEG_DOCR[,c("DEG_chr", "DEG_start", "DEG_end","DEG_ID" , "ENS_ID" ,
                                           "cond1_rep1.x", "cond1_rep2.x", "cond2_rep1.x" , "cond2_rep2
.x",
                                           "norm.cond1_rep1.x", "norm.cond1_rep2.x",  "norm.cond2_rep1.
x" , "norm.cond2_rep2.x","baseMean.x", "cond1.x","cond2.x","FC.x",
                                           "log2FoldChange.x", "pvalue.x",  "padj.x", "tagwise.dispersi
on.x", "trended.dispersion.x", "strand.x",
                                           "mm10_chr.x","mm10_start_gene","mm10_end_gene", "entrezgene_
id.x", "gene_biotype.x" , "mgi_symbol.x" , "entrezgene_accession.x", "entrezgene_description",
                                           "uniprot_gn_symbol",
                                           "DEG_DOCR_overlap", "DOCR_chr", "DOCR_start", "DOCR_end","D
OCR_ID" ,
                                           "mm9_chr", "mm9_start", "mm9_end", "mm9_ID","cond1_rep1.y",
"cond1_rep2.y", "cond2_rep1.y" , "cond2_rep2.y",
                                           "norm.cond1_rep1.y", "norm.cond1_rep2.y",  "norm.cond2_rep1.
y" , "norm.cond2_rep2.y",  "baseMean.y",  "cond1.y" , "cond2.y" ,  "FC.y",
                                           "log2FoldChange.y", "pvalue.y",  "padj.y", "tagwise.dispersi
on.y", "trended.dispersion.y",
                                           "strand.y", "chr_gene", "start_gene", "end_gene", "entrezgen
e_accession.y", "gene_biotype.y",
                                           "mgi_symbol.y" , "ensembl_gene_id", "entrezgene_id.y", "over
lap_length")]


  colnames(Annot_DEG_DOCR_clean)=c("DEG_mm10_chr", "DEG_mm10_start", "DEG_mm10_end","DEG_mm10_ID" ,"DEG
_ENS_ID" ,
                                   "DEG_cond1_rep1", "DEG_cond1_rep2", "DEG_cond2_rep1" , "DEG_cond2_re
p2",
                                   "DEG_norm.cond1_rep1", "DEG_norm.cond1_rep2",  "DEG_norm.cond2_rep1"
, "DEG_norm.cond2_rep2",
                                   "DEG_baseMean",  "DEG_cond1" , "DEG_cond2" , "DEG_FC",
                                   "DEG_log2FoldChange", "DEG_pvalue",  "DEG_padj", "DEG_tagwise.disper
sion", "DEG_trended.dispersion", "DEG_strand", "mm10_chr_gene_annot_DEG",
                                   "mm10_start_gene_annot_DEG","mm10_end_gene_annot_DEG", "entrezgene_i
d_annot_DEG", "gene_biotype_annot_DEG" , "mgi_symbol_annot_DEG" ,
                                   "entrezgene_accession_annot_DEG", "entrezgene_description_annot_DEG"
,"uniprot_gn_symbol_annot_DEG",
                                   "DEG_DOCR_overlap",
                                   "DOCR_mm10_chr", "DOCR_mm10_start", "DOCR_mm10_end","DOCR_mm10_ID",
                                   "DOCR_mm9_chr", "DOCR_mm9_start", "DOCR_mm9_end", "DOCR_mm9_ID","DO
CR_cond1_rep1", "DOCR_cond1_rep2",
                                   "DOCR_cond2_rep1" , "DOCR_cond2_rep2",
                                   "DOCR_norm.cond1_rep1", "DOCR_norm.cond1_rep2",  "DOCR_norm.cond2_r
ep1" , "DOCR_norm.cond2_rep2",  "DOCR_baseMean",  "DOCR_cond1" , "DOCR_cond2" ,
                                   "DOCR_FC",
                                   "DOCR_log2FoldChange", "DOCR_pvalue",  "DOCR_padj", "DOCR_tagwise.d
ispersion", "DOCR_trended.dispersion",
                                   "strand_annot_DOCR", "mm10_chr_gene_annot_DOCR", "mm10_start_gene_an
not_DOCR","mm10_end_gene_annot_DOCR",
                                   "entrezgene_accession_annot_DOCR", "gene_biotype_annot_DOCR",
                                   "mgi_symbol_annot_DOCR" , "ensembl_gene_id_annot_DOCR", "entrezgene_
id_annot_DOCR", "overlap_length_annot_DOCR")



  write.table(Annot_DEG_DOCR_clean, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croi
sement_autre_data/RNA_ATAC/clean_annot",DEG_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names = FA
LSE, col.names = TRUE)
  write.table(Annot_DEG_DOCR, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement
_autre_data/RNA_ATAC/raw_annot",DEG_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names = FALSE, col
.names = TRUE)


  # To keep only gene names (entrezgene_accession_annot_DEG) - annotation of DEG - mm10 BIOMART :
  # To uncollapse data
  decondat_Annot_DEG_DOCR_clean = as.character(Annot_DEG_DOCR_clean$entrezgene_accession_annot_DEG)
  print(DEG_DOCR_sample)
  print(head(decondat_Annot_DEG_DOCR_clean))
```

```
  # To obtain list of genes
  list_gene_splitted <- strsplit(as.character(Annot_DEG_DOCR_clean$entrezgene_accession_annot_DEG),spli
t=";")
  vec_gene_splitted <- unlist(list_gene_splitted)

  # To remove "." :
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  # To save data on a table
  table_VGS_filter = as.data.frame(unique(vec_gene_splitted_filter))

  write.table(table_VGS_filter, paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_au
tre_data/RNA_ATAC/list_gene",DEG_DOCR_sample, sep="_"), quote=FALSE, col.names=FALSE, row.names=FALSE)

}



rm(list = ls())

#-----------------------------------
# Idem for distance=1000
# Load files containing DEG+DOCR regions
list_DEG_DOCR = list.files("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RN
A_ATAC/")
list_DEG_DOCR_dist0 = list_DEG_DOCR[which(grepl("dist0",list_DEG_DOCR))]
list_DEG_DOCR_dist1000 = list_DEG_DOCR[which(grepl("dist1000",list_DEG_DOCR))]

# List of files containing DOCR informations
# Since some files were empty (no overlap), we need to specify, in a list, each file we want to annotat
e.
list_DOCR_Annot=list("stats_mm9_mm10_down_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed","stats_mm9_mm1
0_down_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_down_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_down_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed", "stats_mm9_mm
10_up_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_up_DOCR_E14.16_annot_syntaxique_mm10_biomart.bed", "stats_mm9_mm10
_up_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed",
                     "stats_mm9_mm10_up_DOCR_E16.P0_annot_syntaxique_mm10_biomart.bed")


# Load DEG annotation files
# Since some files were empty (no overlap), we need to specify, in a list, each file we want to annotat
e.
list_DEG_Annot = list("edgeR_AnnotBM_et_Man_down_E14.15_mm10.txt", "edgeR_AnnotBM_et_Man_down_E14.16_mm
10.txt" ,
                      "edgeR_AnnotBM_et_Man_down_E15.16_mm10.txt",
                      "edgeR_AnnotBM_et_Man_down_E16.P0_mm10.txt", "edgeR_AnnotBM_et_Man_up_E14.15_mm10
.txt",
                      "edgeR_AnnotBM_et_Man_up_E14.16_mm10.txt", "edgeR_AnnotBM_et_Man_up_E15.16_mm10.t
xt",
                      "edgeR_AnnotBM_et_Man_up_E16.P0_mm10.txt")


# for distance=1000
for(i in 1:8){

  DEG_DOCR_sample = list_DEG_DOCR_dist1000[i]
  Annot_DEG_DOCR = read.table(file = paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croiseme
nt_autre_data/RNA_ATAC/",DEG_DOCR_sample,sep="/"),
                              header = FALSE, sep = "\t", quote="", fill=TRUE)

  # To rename columns
  colnames(Annot_DEG_DOCR) <- c("DEG_chr","DEG_start", "DEG_end", "DOCR_chr", "DOCR_start", "DOCR_end"
, "DEG_DOCR_overlap")

  #To create two ID : DEG_ID et DOCR_ID
  Annot_DEG_DOCR$DEG_ID = paste(Annot_DEG_DOCR$DEG_chr, Annot_DEG_DOCR$DEG_start, Annot_DEG_DOCR$DEG_e
nd, sep=";")
  Annot_DEG_DOCR$DOCR_ID = paste(Annot_DEG_DOCR$DOCR_chr, Annot_DEG_DOCR$DOCR_start, Annot_DEG_DOCR$DO
CR_end, sep=";")

  # Load annotation of all DEG regions
  DEG_sample <- list_DEG_Annot[i]
  Annot_DEG = read.table(file = paste("G:/projet_DU_AD/results/09_annotation/Sartools/Biomart",DEG_samp
le_sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)
```

```
le,sep= "/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DEG)= c("ENS_ID", "cond1_rep1", "cond1_rep2", "cond2_rep1", "cond2_rep2",
                         "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2",
                         "baseMean", "cond1", "cond2", "FC", "log2FoldChange", "pvalue", "padj", "tagwi
se.dispersion",
                         "trended.dispersion", "mm10_chr", "strand", "mm10_start_gene", "mm10_end_gene"
, "entrezgene_id", "gene_biotype",
                         "mgi_symbol", "entrezgene_accession", "entrezgene_description", "uniprot_gn_sy
mbol")

  Annot_DEG$DEG_ID=paste(paste("chr",Annot_DEG$mm10_chr,sep=""), Annot_DEG$mm10_start_gene, Annot_DEG$m
m10_end_gene,sep=";")

    # Load annoation of all DOCR
  Annot_DOCR_sample = list_DOCR_Annot[i]
  Annot_DOCR = read.table(file = paste("G:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10",
Annot_DOCR_sample,sep="/"), header = TRUE, sep = "\t", quote="", fill=TRUE)

  colnames(Annot_DOCR) = c("mm9_chr", "mm9_start", "mm9_end", "mm9_ID", "mm10_chr", "mm10_start", "mm10
_end", "mm10_ID",
                           "chr_gene", "start_gene", "end_gene", "strand", "entrezgene_accession", "gen
e_biotype", "mgi_symbol",
                           "ensembl_gene_id", "entrezgene_id", "overlap_length", "cond1_rep1", "cond1_r
ep2", "cond2_rep1", "cond2_rep2",
                           "norm.cond1_rep1", "norm.cond1_rep2", "norm.cond2_rep1", "norm.cond2_rep2",
"baseMean", "cond1", "cond2", "FC",
                           "log2FoldChange", "pvalue", "padj", "tagwise.dispersion", "trended.dispersio
n")


  Annot_DOCR$DOCR_ID=paste(Annot_DOCR$mm10_chr, Annot_DOCR$mm10_start, Annot_DOCR$mm10_end,sep=";")

  Annot_DEG_DOCR=merge(Annot_DEG_DOCR, Annot_DEG, by="DEG_ID")
  Annot_DEG_DOCR=merge(Annot_DEG_DOCR, Annot_DOCR, by="DOCR_ID")


  # To remove duplicate columns
  Annot_DEG_DOCR_clean = Annot_DEG_DOCR[,c("DEG_chr", "DEG_start", "DEG_end","DEG_ID" , "ENS_ID" ,
                                           "cond1_rep1.x", "cond1_rep2.x", "cond2_rep1.x" , "cond2_rep2
.x",
                                           "norm.cond1_rep1.x", "norm.cond1_rep2.x",  "norm.cond2_rep1.
x" , "norm.cond2_rep2.x","baseMean.x", "cond1.x","cond2.x","FC.x",
                                           "log2FoldChange.x", "pvalue.x",  "padj.x", "tagwise.dispersi
on.x", "trended.dispersion.x", "strand.x",
                                           "mm10_chr.x","mm10_start_gene","mm10_end_gene", "entrezgene_
id.x", "gene_biotype.x" , "mgi_symbol.x" , "entrezgene_accession.x", "entrezgene_description",
                                           "uniprot_gn_symbol",
                                           "DEG_DOCR_overlap", "DOCR_chr", "DOCR_start", "DOCR_end","D
OCR_ID" ,
                                           "mm9_chr", "mm9_start", "mm9_end", "mm9_ID","cond1_rep1.y",
"cond1_rep2.y", "cond2_rep1.y" , "cond2_rep2.y",
                                           "norm.cond1_rep1.y", "norm.cond1_rep2.y",  "norm.cond2_rep1.
y" , "norm.cond2_rep2.y",  "baseMean.y",  "cond1.y" , "cond2.y" ,  "FC.y",
                                           "log2FoldChange.y", "pvalue.y",  "padj.y", "tagwise.dispersi
on.y", "trended.dispersion.y",
                                           "strand.y", "chr_gene", "start_gene", "end_gene", "entrezgen
e_accession.y", "gene_biotype.y",
                                           "mgi_symbol.y" , "ensembl_gene_id", "entrezgene_id.y", "over
lap_length")]


  colnames(Annot_DEG_DOCR_clean)=c("DEG_mm10_chr", "DEG_mm10_start", "DEG_mm10_end","DEG_mm10_ID" ,"DEG
_ENS_ID" ,
                                   "DEG_cond1_rep1", "DEG_cond1_rep2", "DEG_cond2_rep1" , "DEG_cond2_re
p2",
                                   "DEG_norm.cond1_rep1", "DEG_norm.cond1_rep2",  "DEG_norm.cond2_rep1"
, "DEG_norm.cond2_rep2",
                                   "DEG_baseMean",  "DEG_cond1" , "DEG_cond2" , "DEG_FC",
                                   "DEG_log2FoldChange", "DEG_pvalue",  "DEG_padj", "DEG_tagwise.disper
sion", "DEG_trended.dispersion", "DEG_strand", "mm10_chr_gene_annot_DEG",
                                   "mm10_start_gene_annot_DEG","mm10_end_gene_annot_DEG", "entrezgene_i
d_annot_DEG", "gene_biotype_annot_DEG" , "mgi_symbol_annot_DEG" ,
                                   "entrezgene_accession_annot_DEG", "entrezgene_description_annot_DEG"
,"uniprot_gn_symbol_annot_DEG",
                                   "DEG_DOCR_overlap",
                                   "DOCR_mm10_chr", "DOCR_mm10_start", "DOCR_mm10_end","DOCR_mm10_ID",
                                   "DOCR_mm9_chr", "DOCR_mm9_start", "DOCR_mm9_end", "DOCR_mm9_ID", "DO
```

```
                                    DOCR_mm9_chr" , "DOCR_mm9_start" , "DOCR_mm9_end" , "DOCR_mm9_ID" , DO
CR_cond1_rep1", "DOCR_cond1_rep2",
                                    "DOCR_cond2_rep1" , "DOCR_cond2_rep2",
                                    "DOCR_norm.cond1_rep1", "DOCR_norm.cond1_rep2",  "DOCR_norm.cond2_r
ep1" , "DOCR_norm.cond2_rep2",  "DOCR_baseMean",  "DOCR_cond1" , "DOCR_cond2" ,
                                    "DOCR_FC",
                                    "DOCR_log2FoldChange", "DOCR_pvalue",  "DOCR_padj", "DOCR_tagwise.d
ispersion", "DOCR_trended.dispersion",
                                    "strand_annot_DOCR", "mm10_chr_gene_annot_DOCR", "mm10_start_gene_an
not_DOCR","mm10_end_gene_annot_DOCR",
                                    "entrezgene_accession_annot_DOCR", "gene_biotype_annot_DOCR",
                                    "mgi_symbol_annot_DOCR" , "ensembl_gene_id_annot_DOCR", "entrezgene_
id_annot_DOCR", "overlap_length_annot_DOCR")



  write.table(Annot_DEG_DOCR_clean, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croi
sement_autre_data/RNA_ATAC/clean_annot",DEG_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names = FA
LSE, col.names = TRUE)
  write.table(Annot_DEG_DOCR, file=paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement
_autre_data/RNA_ATAC/raw_annot",DEG_DOCR_sample,sep="_"), quote=FALSE, sep="\t", row.names = FALSE, col
.names = TRUE)


  # To keep only gene names (entrezgene_accession_annot_DEG) - annotation of DEG - mm10 BIOMART :
  # To uncollapse data
  decondat_Annot_DEG_DOCR_clean = as.character(Annot_DEG_DOCR_clean$entrezgene_accession_annot_DEG)
  print(DEG_DOCR_sample)
  print(head(decondat_Annot_DEG_DOCR_clean))

  # To obtain list of genes
  list_gene_splitted <- strsplit(as.character(Annot_DEG_DOCR_clean$entrezgene_accession_annot_DEG),spli
t=";")
  vec_gene_splitted <- unlist(list_gene_splitted)

  # To remove "."
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  # To save data on a table
  table_VGS_filter = as.data.frame(unique(vec_gene_splitted_filter))

  write.table(table_VGS_filter, paste("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_au
tre_data/RNA_ATAC/list_gene",DEG_DOCR_sample, sep="_"), quote=FALSE, col.names=FALSE, row.names=FALSE)

}
```

Estimation of the proportion of DOCR included in methylome capture

To observe overlap between DOCR and DMR, sequences of DOCR must be included in methylome capture, otherwise, it will not be possible to detect a DMR, even if it exists. To determine proportion of DOCR regions that are in methylome capture, methylome capture regions and DOCR ones were compared using our own R function called *find_overlaps_AD_table()*. Both overlaping regions and regions les than 1000 bases apart were investigated by setting *tolerance* argument to 0 or 1000. For these comparisons, mm9 coordinates were used.

In [ ]:

```
%%R

# Load DOCR files
setwd("E:/projet_DU_DSD/results/10_annotations/Sartools/Biomart/mm10")

DOCR_u14_15 = read.table("stats_mm9_mm10_up_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed", header = TR
UE, fill=TRUE, quote="", sep="\t", na.strings = "", stringsAsFactors = FALSE)
DOCR_u15_16 = read.table("stats_mm9_mm10_up_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed", header = TR
UE, fill=TRUE, quote="", sep="\t", na.strings = "", stringsAsFactors = FALSE)
DOCR_d14_15 = read.table("stats_mm9_mm10_down_DOCR_E14.15_annot_syntaxique_mm10_biomart.bed", header =
TRUE, fill=TRUE, quote="", sep="\t", na.strings = "", stringsAsFactors = FALSE)
DOCR_d15_16 = read.table("stats_mm9_mm10_down_DOCR_E15.16_annot_syntaxique_mm10_biomart.bed", header =
TRUE, fill=TRUE, quote="", sep="\t", na.strings = "", stringsAsFactors = FALSE)

head(DOCR_u14_15)
head(DOCR_u15_16)
head(DOCR_d14_15)
head(DOCR_d15_16)
```

```
dim(DOCR_u14_15)
dim(DOCR_u15_16)
dim(DOCR_d14_15)
dim(DOCR_d15_16)

# Load capture files containing regions included in the methylome capture
setwd("H:/methylome/fusion/capture/capture_regions/")

capture = read.table("Capture_76800_et150seq.merge.sorted.bed", header = FALSE, fill=TRUE, quote="", se
p="\t", na.strings = "", stringsAsFactors = FALSE)

dim(capture)
head(capture)

# Function to find overlap (or close) regions between two datasets
find_overlaps_AD_table <- function (regions_1,regions_2,tolerance=0)
{
  common_region_table = NULL
  number_of_regions_1 <- nrow(regions_1)
  number_of_regions_2 <- nrow(regions_2)

  overlapped_is <- c()
  overlapped_js <- c()

  for (i in 1:number_of_regions_1) {

    for (j in 1:number_of_regions_2) {

      if (as.character(regions_1[i,1])==as.character(regions_2[j,1])) {
        start_1 <- as.numeric(regions_1[i,2])
        end_1 <- as.numeric(regions_1[i,3])
        expanded_start_1 <- start_1 - tolerance
        expanded_end_1 <- end_1 + tolerance
        start_2 <- as.numeric(regions_2[j,2])
        end_2 <- as.numeric(regions_2[j,3])

        if (expanded_end_1>=start_2) {

          if (end_2>=expanded_start_1) {

            overlapped_is <- c(overlapped_is,i)
            overlapped_js <- c(overlapped_js,j)

            first_region <- paste(regions_1[i,1],start_1,end_1,sep="_")
            second_region <- paste(regions_2[j,1],start_2,end_2,sep="_")
            overlap_size <- min(end_1,end_2)-max(start_1,start_2)+1
            common_region_table = rbind(common_region_table,c(as.character(regions_1[i,1]),start_1,end_
1,
                                                               as.character(regions_2[j,1]),start_2,end_
2, overlap_size))

            if (overlap_size>0) {
              print(paste(first_region,second_region,sep=" "))
            } else {
              print(paste(first_region,second_region, sep=" "))
            }
          }
        }
      }
    }
  }
  print(paste("Number matched in first data set:",length(unique(overlapped_is))))
  print(paste("Number matched in second data set:",length(unique(overlapped_js))))
  if (length(common_region_table) > 0){
    colnames(common_region_table) = c("chr_reg1", "start_reg1", "end_reg1", "chr_reg2", "start_reg2", "
end_reg2","overlap_size")
  }else{
  }
  return(as.data.frame(common_region_table))
}


# Comparison of DOCR files and capture - dist0
capt_DOCR_u14_15 = find_overlaps_AD_table(DOCR_u14_15,capture,tolerance=0)
dim(capt_DOCR_u14_15)
write.table(capt_DOCR_u14_15,
```

```
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/capt_DOCR_up_14_15_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


capt_DOCR_u15_16 = find_overlaps_AD_table(DOCR_u15_16,capture,tolerance=0)
dim(capt_DOCR_u15_16)
write.table(capt_DOCR_u15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/capt_DOCR_up_15_16_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


capt_DOCR_d14_15 = find_overlaps_AD_table(DOCR_d14_15,capture,tolerance=0)
dim(capt_DOCR_d14_15)
write.table(capt_DOCR_d14_15,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/capt_DOCR_down_14_15_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


capt_DOCR_d15_16 = find_overlaps_AD_table(DOCR_d15_16,capture,tolerance=0)
dim(capt_DOCR_d15_16)
write.table(capt_DOCR_d15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/capt_DOCR_down_15_16_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)




# Comparison of DOCR files and capture - dist1000
large_capt_DOCR_u14_15 = find_overlaps_AD_table(DOCR_u14_15,capture,tolerance=1000)
dim(large_capt_DOCR_u14_15)
write.table(large_capt_DOCR_u14_15,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/large_capt_DOCR_up_14_15_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


large_capt_DOCR_u15_16 = find_overlaps_AD_table(DOCR_u15_16,capture,tolerance=1000)
dim(large_capt_DOCR_u15_16)
write.table(large_capt_DOCR_u15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/large_capt_DOCR_up_15_16_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


large_capt_DOCR_d14_15 = find_overlaps_AD_table(DOCR_d14_15,capture,tolerance=1000)
dim(large_capt_DOCR_d14_15)
write.table(large_capt_DOCR_d14_15,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/large_capt_DOCR_down_14_15_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


large_capt_DOCR_d15_16 = find_overlaps_AD_table(DOCR_d15_16,capture,tolerance=1000)
dim(large_capt_DOCR_d15_16)
write.table(large_capt_DOCR_d15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/ATACseq_capture/large_capt_DOCR_down_15_16_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)
```

Estimation of the proportion of DEG included in methylome capture

To observe overlap between DEG and DMR, sequences of DEG must be in methylome capture, otherwise, it will not be possible to detect a DMR, even if it exists. To determine proportion of DEG regions that are included in methylome capture, methylome capture regions and DEG were compared using our own R function called *find_overlaps_AD_table()*. Both overlaping regions and regions less than 1000 bases apart were investigated by setting *tolerance* argument to 0 or 1000. To be able to compare the two datasets, DEG regions with mm10 coordinates were converted to mm9 coordinates using LiftOver, before comparisons (capture regions have mm9 coordinates). Default settings of LiftOver tool have been retained (*i.e.* at least 0.95 as the minimum ratio of bases that must remap). Some regions can't be converted for distinct reasons (see below).

Explanations of conversion problem:

Explanations of conversion problem:

- Deleted in new: Sequence intersects no chains
- Partially deleted in new: Sequence insufficiently intersects one chain
- Split in new: Sequence insufficiently intersects multiple chains
- Duplicated in new: Sequence sufficiently intersects multiple chains
- Boundary problem: Missing start or end base in an exon

In [ ]:

```
%%R

#------------------------------------------------------------------
# Conversion of mm9 coordinates --> mm10 coordinates with LiftOver

# Files that were converted:
# E:\projet_DU_AD\results\09_annotation\Sartools\Biomart\bed\bed_edge_AnnotBM_et_Man_ech_down_E14.15_mm
10.sorted.bed
# + bed_edge_AnnotBM_et_Man_ech_down_E15.16_mm10.sorted.bed
# + bed_edge_AnnotBM_et_Man_ech_up_E14.15_mm10.sorted.bed
# + bed_edge_AnnotBM_et_Man_ech_up_E15.16_mm10.sorted.bed

#---
# For DEG_E14.15_down :
Successfully converted 2500 records
Conversion failed on 12 records.

#Partially deleted in new
chr15 8967949 9067335
#Partially deleted in new
chr4 99829198 99912788
#Partially deleted in new
chr6 30747554 30896794
#Partially deleted in new
chr8 39005845 39165114
#Split in new
chr9 101074101 101104800
#Deleted in new
chrMT 70 1024
#Deleted in new
chrMT 2751 3707
#Deleted in new
chrMT 3845 3913
#Deleted in new
chrMT 10167 11544
#Deleted in new
chrMT 11742 13565
#Deleted in new
chrMT 14145 15288
#Partially deleted in new
chrX 140956907 141164270


#---
# For DEG_down_E15.16
Successfully converted 324 records
Conversion failed on 3 records.

#Split in new
chr5 142724661 142817662
#Deleted in new
chrMT 3772 3842
#Deleted in new
chrMT 3845 3913



#---
#pour DEG_up_E14.15 :
Successfully converted 3414 records
Conversion failed on 12 records.

#Partially deleted in new
chr1 22286251 22805994
#Split in new
chr12 18648214 18649998
```

```
#Partially deleted in new
chr2 58821070 59160683
#Split in new
chr4 156235919 156247616
#Partially deleted in new
chr5 113490333 113589725
#Split in new
chr5 142724661 142817662
#Partially deleted in new
chr7 6343759 6355956
#Partially deleted in new
chr7 138940730 139083977
#Split in new
chr9 124422622 124476898
#Deleted in new
chrMT 3772 3842
#Deleted in new
chrX 170009659 170019281
#Split in new
chrY 90784738 90816465


#---
#Pour DEG_up_E15.16
Successfully converted 196 records


# Data were saved on following files:
#E:\projet_DU_AD\results\09_annotation\Sartools\Biomart\bed\mm9\mm9_bed_DEG_E14.15_down.bed
# + mm9_bed_DEG_E15.16_down.bed
# + mm9_bed_DEG_E14.15_up.bed
# + mm9_bed_DEG_E15.16_up.bed


#-----------------------------------------------------------------
# Comparison of DEG and capture regions
# Load DEG (mm9) files
setwd("E:/projet_DU_AD/results/09_annotation/Sartools/Biomart/bed/mm9/")

DEG_u14_15 = read.table("mm9_bed_DEG_E14.15_up.bed", header = FALSE, fill=TRUE, quote="", sep="\t", na.
strings = "", stringsAsFactors = FALSE)
DEG_u15_16 = read.table("mm9_bed_DEG_E15.16_up.bed", header = FALSE, fill=TRUE, quote="", sep="\t", na.
strings = "", stringsAsFactors = FALSE)
DEG_d14_15 = read.table("mm9_bed_DEG_E14.15_down.bed", header = FALSE, fill=TRUE, quote="", sep="\t", n
a.strings = "", stringsAsFactors = FALSE)
DEG_d15_16 = read.table("mm9_bed_DEG_E15.16_down.bed", header = FALSE, fill=TRUE, quote="", sep="\t", n
a.strings = "", stringsAsFactors = FALSE)

head(DEG_u14_15)
head(DEG_u15_16)
head(DEG_d14_15)
head(DEG_d15_16)

dim(DEG_u14_15)
dim(DEG_u15_16)
dim(DEG_d14_15)
dim(DEG_d15_16)


# Load capture file - mm9
setwd("H:/methylome/fusion/capture/capture_regions/")

capture = read.table("Capture_76800_et150seq.merge.sorted.bed", header = FALSE, fill=TRUE, quote="", se
p="\t", na.strings = "", stringsAsFactors = FALSE)

dim(capture)
head(capture)


# Function to compare the two datasets
find_overlaps_AD_table <- function (regions_1,regions_2,tolerance=0)
{
  common_region_table = NULL
  number_of_regions_1 <- nrow(regions_1)
  number_of_regions_2 <- nrow(regions_2)

  overlapped_is <- c()
```

```
  overlapped_js <- c()

  for (i in 1:number_of_regions_1) {

    for (j in 1:number_of_regions_2) {

      if (as.character(regions_1[i,1])==as.character(regions_2[j,1])) {
        start_1 <- as.numeric(regions_1[i,2])
        end_1 <- as.numeric(regions_1[i,3])
        expanded_start_1 <- start_1 - tolerance
        expanded_end_1 <- end_1 + tolerance
        start_2 <- as.numeric(regions_2[j,2])
        end_2 <- as.numeric(regions_2[j,3])

        if (expanded_end_1>=start_2) {

          if (end_2>=expanded_start_1) {

            overlapped_is <- c(overlapped_is,i)
            overlapped_js <- c(overlapped_js,j)

            first_region <- paste(regions_1[i,1],start_1,end_1,sep="_")
            second_region <- paste(regions_2[j,1],start_2,end_2,sep="_")
            overlap_size <- min(end_1,end_2)-max(start_1,start_2)+1
            common_region_table = rbind(common_region_table,c(as.character(regions_1[i,1]),start_1,end_
1,
                                                               as.character(regions_2[j,1]),start_2,end_
2, overlap_size))

            if (overlap_size>0) {
              print(paste(first_region,second_region,sep=" "))
            } else {
              print(paste(first_region,second_region, sep=" "))
            }
          }
        }
      }
    }
  }
  print(paste("Number matched in first data set:",length(unique(overlapped_is))))
  print(paste("Number matched in second data set:",length(unique(overlapped_js))))
  if (length(common_region_table) > 0){
    colnames(common_region_table) = c("chr_reg1", "start_reg1", "end_reg1", "chr_reg2", "start_reg2", "
end_reg2","overlap_size")
  }else{
  }
  return(as.data.frame(common_region_table))
}



# Comparisons of DEG and capture regions - dist0
capt_DEG_u14_15 = find_overlaps_AD_table(DEG_u14_15,capture,tolerance=0)
dim(capt_DEG_u14_15)
write.table(capt_DEG_u14_15,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/ca
pt_DEG_up_14_15_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


capt_DEG_u15_16 = find_overlaps_AD_table(DEG_u15_16,capture,tolerance=0)
dim(capt_DEG_u15_16)
write.table(capt_DEG_u15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/ca
pt_DEG_up_15_16_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


capt_DEG_d14_15 = find_overlaps_AD_table(DEG_d14_15,capture,tolerance=0)
dim(capt_DEG_d14_15)
write.table(capt_DEG_d14_15,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/ca
pt_DEG_down_14_15_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


capt_DEG_d15_16 = find_overlaps_AD_table(DEG_d15_16,capture,tolerance=0)
```

```
dim(capt_DEG_d15_16)
write.table(capt_DEG_d15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/capt_DEG_down_15_16_dist0.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)



# Comparisons of DEG and capture regions - dist1000
large_capt_DEG_u14_15 = find_overlaps_AD_table(DEG_u14_15,capture,tolerance=1000)
dim(large_capt_DEG_u14_15)
write.table(large_capt_DEG_u14_15,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/large_capt_DEG_up_14_15_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


large_capt_DEG_u15_16 = find_overlaps_AD_table(DEG_u15_16,capture,tolerance=1000)
dim(large_capt_DEG_u15_16)
write.table(large_capt_DEG_u15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/large_capt_DEG_up_15_16_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


large_capt_DEG_d14_15 = find_overlaps_AD_table(DEG_d14_15,capture,tolerance=1000)
dim(large_capt_DEG_d14_15)
write.table(large_capt_DEG_d14_15,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/large_capt_DEG_down_14_15_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)


large_capt_DEG_d15_16 = find_overlaps_AD_table(DEG_d15_16,capture,tolerance=1000)
dim(large_capt_DEG_d15_16)
write.table(large_capt_DEG_d15_16,
            "H:/methylome/fusion/post_methylkit/DMR432_pval0.07/croisement_autre_data/RNAseq_capture/large_capt_DEG_down_15_16_dist1000.txt",
            sep = "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)
```

Are DEG and DOCR over-represented among DMR ?

To estimate whether DMRs are particularly located to DEG or DOCR observed during physiological brain development, hypergeometric tests were done as follow:

In [ ]:

```
%%R

# Hypergeomtric tests on DOCR among DMR :
# We observed:
# 2 DMR that are also DOCR_down between E14.5 and E15.5 (dist0)
# 0 DMR that are also DOCR_down between E15.5 and E16.5 (dist0)
# 5 DMR that are also DOCR_up between E14.5 and E15.5 (dist0)
# 1 DMR that are also DOCR_up between E15.5 and E16.5 (dist0)

# 3 DMR that are also DOCR_down between E14.5 and E15.5 (dist1000)
# 0 DMR that are also DOCR_down between E15.5 and E16.5 (dist1000)
# 5 DMR that are also DOCR_up between E14.5 and E15.5 (dist1000)
# 1 DMR that are also DOCR_up between E15.5 and E16.5 (dist1000)

# We observed 432 DMR in total
# among 58 611 regions (all capture regions)
nb_DMR = 432
nb_capt_tot = 58611


# For DMR-DOCR down E14.15_dist0 :
nb_ov = 2
nb_ATAC_capt = 449
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)
```

```
# For DMR-DOCR down E15.16_dist0 :
nb_ov = 0
nb_ATAC_capt = 79
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)

# For DMR-DOCR up E14.15_dist0 :
nb_ov = 5
nb_ATAC_capt = 313
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)

# For DMR-DOCR up E15.16_dist0 :
nb_ov = 1
nb_ATAC_capt = 75
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)


# For DMR-DOCR down E14.15_dist1000 :
nb_ov = 3
nb_ATAC_capt = 561
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)

# For DMR-DOCR down E15.16_dist1000 :
nb_ov = 0
nb_ATAC_capt = 119
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)

# For DMR-DOCR up E14.15_dist1000 :
nb_ov = 5
nb_ATAC_capt = 377
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)

# For DMR-DOCR up E15.16_dist1000 :
nb_ov = 1
nb_ATAC_capt = 85
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_ATAC_capt, nb_capt_tot - nb_ATAC_capt, nb_DMR, lower.tail = TRUE)




# Hypergeomtric tests on DEG among DMR :
# We observed:
# 21 DMR that are also DEG_down between E14.5 and E15.5 (dist0)
# 6 DMR that are also DEG_down between E15.5 and E16.5 (dist0)
# 96 DMR that are also DEG_up between E14.5 and E15.5 (dist0), but 73 uniq regions from dataset RNA and
88 from dataset DMR
# 16 DMR that are also DEG_up between E15.5 and E16.5 (dist0), but 13 uniq regions from dataset RNA and
16 from dataset DMR

# 27 DMR that are also DEG_down between E14.5 and E15.5 (dist1000), but 27 uniq regions from dataset RN
A and 26 from dataset DMR
# 10 DMR that are also DEG_down between E15.5 and E16.5 (dist1000)
# 106 DMR that are also DEG_up between E14.5 and E15.5 (dist1000), but 83 uniq regions from dataset RNA
and 98 from dataset DMR
# 18 DMR that are also DEG_up between E15.5 and E16.5 (dist1000), but 15 uniq regions from dataset RNA
and 18 from dataset DMR

# We observed 432 DMR in total
# among 58 611 regions (all capture regions)
nb_DMR = 432
nb_capt_tot = 58611

# For DEG down E14.15_dist0 :
nb_ov = 21
nb_RNA_capt = 4678
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)

# For DEG down E15.16_dist0 :
nb_ov = 6
nb_RNA_capt = 826
```

```
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)

# For DEG up E14.15_dist0 :
nb_ov = 96
nb_RNA_capt = 10458
  phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)

# For DEG up E15.16_dist0 :
nb_ov = 16
nb_RNA_capt = 1465
  phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)


# For DEG down E14.15_dist1000 :
nb_ov = 27
nb_RNA_capt = 5250
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)

# For DEG down E15.16_dist1000 :
nb_ov = 10
nb_RNA_capt = 902
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)

# For DEG up E14.15_dist1000 :
nb_ov = 106
nb_RNA_capt = 11141
  phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)

# For DEG up E15.16_dist1000 :
nb_ov = 18
nb_RNA_capt = 1508
  phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = FALSE)
phyper(nb_ov, nb_RNA_capt, nb_capt_tot - nb_RNA_capt, nb_DMR, lower.tail = TRUE)
```

3.5. Workflow de l'analyse bioinformatique des ChIP-seq ciblant HSF2 et DNMT3A

Analyse du ChIP-seq ciblant HSF2 et DNMT3A

Agathe Duchateau

Création d'un environnement conda et installation des outils bio-informatiques

Pour analyser le ChIP-seq ciblant HSF2 et DNMT3A, les outils suivants ont été installés dans un environnement conda. La version de chaque outil utilisé est précisée.

In [ ]:

```
%%bash
#conda version 4.7.11
# Création de l'environnement
conda create -n chip
conda activate chip

# Installation des outils
conda install FASTQC #version 0.11.8
conda install trim-galore #version 0.6.3
conda install bowtie2 #version 2.3.5
conda install samtools #version 1.9
conda install MACS2 #version 2.1.2
conda install homer #version 4.9.1
conda install unzip #version 6.0
conda install bedtools #version 2.25.0

# Package R (version 3.5.2) utilisé :
# BiomaRt #version 2.38.0

# Outil utilisé sur Galaxeast :
# homer_annotatePeaks # Galaxy Version 0.0.5

# autres outils utilisés, disponibles sur internet :
# LiftOver (UCSC)
# DAVID #version 6.8
# ToppFun (suite ToppGene)
```

Contrôle qualité des données brutes

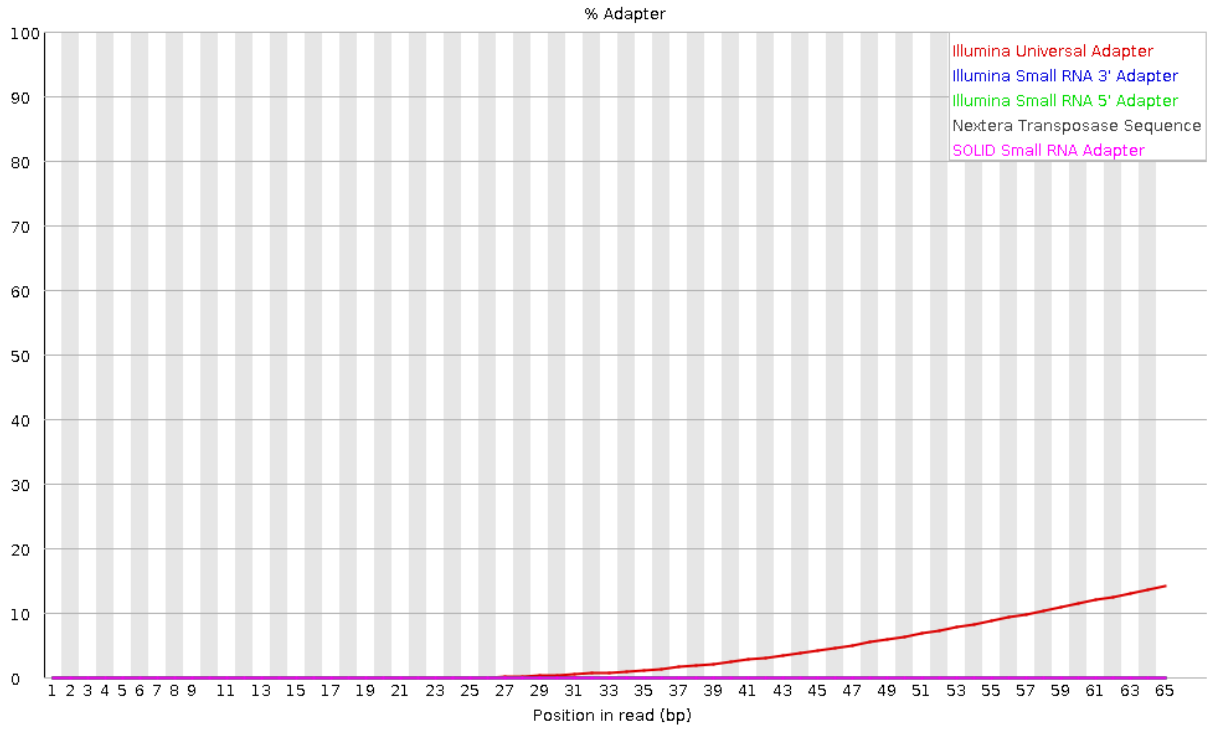
Après le séquençage, nous obtenons 4 fichiers .fastq par échantillon. La commande **cat** permet de les réunir pour ne travailler qu'avec un fichier global par échantillon. La qualité des données issues du séquençage (fichiers séparés) et des données regroupées a été vérifiée avec l'outil **FASTQC**. Pour effectuer ces différentes étapes, le script suivant a été exécuté :

In [ ]:

```
%%bash
#Script 1 - Analyse ChIP-seq HSF2 / DNMT3A pilote

################################################
# FASTQC des données brutes
# + Regroupement des fichiers d'un même échantillon
# + FASTQC des données regroupées
################################################

# le script va s'arrêter
# - à la première erreur
# - si une variable n'est pas définie
# - si une erreur est rencontrée dans un pipe
set -euo pipefail

#----------------------------------------------
# Contrôle qualité des données séparées (FASTQC)
#----------------------------------------------

mkdir -p /mnt/g/Chip_pilote/results/00_raw_data/cat/
```

```
mkdir -p /mnt/g/Chip_pilote/results/01_FASTQC/separated/
mkdir -p /mnt/g/Chip_pilote/results/01_FASTQC/cat/

# nom des échantillons à analyser
sample="IP_WT_HSF2 IP_KO_HSF2 Input_KO Input_WT IP_WT_D3A IP_KO_D3A"

    echo "================================================================="
    echo "FASTQC"
    echo "================================================================="

for file in /mnt/g/Chip_pilote/results/00_raw_data/*R1*
do
    echo "================================================================="
    echo "file name : ${file}"
    echo "================================================================="
    fastqc $file
done

cd /mnt/g/Chip_pilote/results/00_raw_data/
for file in *html *zip
do
    echo "================================================================="
    echo "file name : ${file}"
    echo "================================================================="
    mv $file /mnt/g/Chip_pilote/results/01_FASTQC/separated/
done

#-------------------------------------------------------------
# regroupement des données et FASTQC des données regroupées
#-------------------------------------------------------------

    echo "================================================================="
    echo "cat"
    echo "================================================================="

cd /mnt/g/Chip_pilote/results/00_raw_data/
for ech in $sample
do
    echo "================================================================="
    echo "file name : ${ech}"
    echo "================================================================="
    cat ${ech}_S?_L001_R1_001.fastq.gz ${ech}_S?_L002_R1_001.fastq.gz ${ech}_S?_L003_R1_001.fastq.gz
    ${ech}_S?_L004_R1_001.fastq.gz > /mnt/g/Chip_pilote/results/00_raw_data/cat/cat_${ech}_R1.fastq.gz

    fastqc /mnt/g/Chip_pilote/results/00_raw_data/cat/cat_${ech}_R1.fastq.gz
done

cd /mnt/g/Chip_pilote/results/00_raw_data/cat/
for file in *html *zip
do
    echo "================================================================="
    echo "file name : ${ech}"
    echo "================================================================="
    mv $file /mnt/g/Chip_pilote/results/01_FASTQC/cat/
done
```

Mise à part la présence de quelques séquences adaptatrices du séquençage, les contrôles qualités sont satisfaisants pour l'ensemble des échantillons.



Contenu en séquences adaptatrices (FASTQC) pour l'échantillon ChIP HSF2 - WT HSF2

Trimming des données

Suite au contrôle qualité, un *trimming* des données a été réalisé afin d'éliminer :

- les séquences adaptatrices (*option --illumina*)
- les reads de mauvaise qualité, c'est à dire en dessous d'un *phred score* de 20 (*option -q 20*)
- les reads dont la longueur est inférieure à 20 bases (*option --length 20*)

L'outil **trim-galore** a permis de faire ce filtrage. La qualité des données filtrées a ensuite été vérifiée avec l'outil **FASTQC**.

In [ ]:

```
%%bash
#Script 2 - Analyse ChIP-seq HSF2 / DNMT3A pilote

###########################################
# Trimming
# FASTQC des données filtrées
###########################################

# le script va s'arrêter
# - à la première erreur
# - si une variable n'est pas définie
# - si une erreur est rencontrée dans un pipe
set -euo pipefail

mkdir -p /mnt/g/Chip_pilote/results/02_trimming/
mkdir -p /mnt/g/Chip_pilote/results/01_FASTQC/trim/

# numéro des échantillons à analyser
sample="IP_WT_HSF2 IP_KO_HSF2 Input_KO Input_WT IP_WT_D3A IP_KO_D3A"

    echo "================================================================="
    echo "Trimming"
    echo "================================================================="

cd /mnt/g/Chip_pilote/results/00_raw_data/cat
for ech in $sample
do
    echo "================================================================="
    echo "file name : ${ech}"
    echo "================================================================="
    trim_galore -q 20 --phred33 --fastqc --fastqc_args "--outdir /mnt/g/Chip_pilote/results/01_FASTQC/t
rim/"
    --illumina --stringency 1 --length 20 --output_dir /mnt/g/Chip_pilote/results/02_trimming/
    cat_${ech}_R1.fastq.gz > ct_${ech}_R1.fastq.gz
done
```

Mapping

Une fois filtrées, les séquences ont été alignées sur le génome de référence murin ***mm9***, à l'aide de l'outil **bowtie2**, en *single-end*, en utilisant les options prédéfinies pour une analyse en mode *end-to-end*. Afin d'obtenir un *mapping* fiable., le mode *very sensitive* a également été choisi.

In [ ]:

```
%%bash
#Script 3 - Analyse ChIP-seq HSF2 / DNMT3A pilote

############################
# Mapping - very-sensitive
############################

# le script va s'arrêter
# - à la première erreur
# - si une variable n'est pas définie
# - si une erreur est rencontrée dans un pipe
set -euo pipefail

mkdir -p /mnt/g/Chip_pilote/results/03_mapping/SE_R1_very_sensitive/

# numéro des échantillons à analyser
sample="IP_WT_HSF2 IP_KO_HSF2 Input_KO Input_WT IP_WT_D3A IP_KO_D3A"

cd /mnt/g/Chip_pilote/results/02_trimming/
for ech in $sample
do
    echo "================================================================="
    echo "file name : ${ech}"
    echo "Mapping"
    echo "================================================================="
    bowtie2 -q --trim3 3 --phred33 --very-sensitive --time
    --un-gz /mnt/g/Chip_pilote/results/03_mapping/SE_R1_very_sensitive/unmapped_ct_${ech}_SE_R1_VS.sam.
gz
    -p 4 -x /mnt/g/Chip_pilote/results/03_mapping/index_bowtie2/mm9 -U cat_${ech}_R1_trimmed.fq.gz
    -S /mnt/g/Chip_pilote/results/03_mapping/SE_R1_very_sensitive/ct_${ech}_SE_R1_VS.sam
```

```
done
```

Suppression des duplicats

Pour éviter les biais liés à l'amplification PCR pré-séquençage, l'outil **samtools markdup** a été utilisé pour supprimer les duplicats, c'est à dire les reads qui ont exactement les mêmes coordonnées chromosomiques. Les fichiers .bam résultant ont été ordonnés ét indexés avec les outils **samtools sort** et **samtools index**, afin de visualiser les reads alignés, avec le logiciel **IGV**.

In [ ]:

```
%%bash
#Script 4 - Analyse ChIP-seq HSF2 / DNMT3A pilote

# le script va s'arrêter
# - à la première erreur
# - si une variable n'est pas définie
# - si une erreur est rencontrée dans un pipe
set -euo pipefail

# numéro des échantillons à analyser
sample="IP_WT_HSF2 IP_KO_HSF2 Input_KO Input_WT IP_WT_D3A IP_KO_D3A"

cd /mnt/h/Chip_pilote/results/03_mapping/SE_R1_very_sensitive/

for ech in $sample
do
    echo "============================================================="
    echo "file name : ${ech}"
    echo "Remove Dup"
    echo "============================================================="

# /!\ /!\
# The input file must be coordinate sorted and must have gone
# through fixmates with the mate scoring option on.
# /!\ /!\

    echo "Sorted by name"
    samtools sort -n -o nh_ct_${ech}_SE_R1_VS.sorted.bam  nh_ct_${ech}_SE_R1_VS.bam

    echo "Add ms and MC tags for markdup to use later"
    samtools fixmate -m nh_ct_${ech}_SE_R1_VS.sorted.bam
    nh_ct_${ech}_SE_R1_VS.fixmate.bam

    echo "Markdup needs position order"
    samtools sort -o nh_ct_${ech}_SE_R1_VS.position_sorted.bam
    nh_ct_${ech}_SE_R1_VS.fixmate.bam

    echo "Run markdup"
    samtools markdup -r -s
    nh_ct_${ech}_SE_R1_VS.position_sorted.bam nh_ct_${ech}_SE_R1_VS_markdup.bam

    echo "============================================================="
    echo "Sort and index"
    echo "============================================================="
    samtools sort nh_ct_${ech}_SE_R1_VS_markdup.bam
    -o nh_ct_${ech}_SE_R1_VS_markdup.sorted.bam

    samtools index nh_ct_${ech}_SE_R1_VS_markdup.sorted.bam
    nh_ct_${ech}_SE_R1_VS_markdup.sorted.bai
done


# Détails des options de samtools markdup:
#  -r           Remove duplicate reads
#  -s           Report stats.
```

Recherche de régions enrichies - *Peak calling*

L'outil **MACS2** permet de détecter les régions enrichies en HSF2 ou DNMT3A, en normalisant le nombre de reads dédupliqués des échantillons immunoprécipités à celui de l'input correspondant, pour tenir compte des différences de taille des banques. Pour HSF2, ces pics sont ensuite filtrés :

- Pour supprimer les pics non spécifiques (*i.e.* régions où un enrichissement HSF2 est détecté alors qu'il n'y en a pas dans l'échantillons biologique), les pics de *ChIP* HSF2 détectés à la fois chez un embryon KO HSF2 et WT HSF2 sont éliminés.
- Les pics restants, détectés exclusivement dans l'échantillon WT HSF2 sont recoupés avec la *blacklist* du génome mm9, réunissant des régions apparaissant enrichies dans la plupart des analyses de *ChIP*, quel que soit l'anticorps utilisé pour l'expérience de ChIP.

Détail des options utilisées avec l'outil MACS2 :

- -macs2 callpeak : module permettant de faire le *Peak calling*
- -t treated_file.bam : fichier .bam à traiter
- --control control_file.bam : fichier contrôle à traiter (mock, input ou KO)
- --name HSF2_WT_vs_Input : préfixe donné aux fichiers de sorties
- --outdir folder : nom du dossier où déposer les fichiers de sorties
- --format BAM : pour préciser le format du fichier à traiter
- --gsize mm : pour préciser la taille du génome de référence, réellement séquençable. Indiquer mm pour mm9 *Mus musculus*
- --tsize 70 : taille des reads alignés
- --qvalue 0.05 : qvalue seuil (ou valeur FDR seuil). La q-value est calculée à partir des pvalues selon la méthode de Benjamini-Hochberg. Défaut : 0.05.
- --keep-dup 1 : pour ne conserver qu'un dupliquat lorsque des dupliquats existent. Défaut : 1
- -B : pour obtenir un fichier bedGraph

In [ ]:

```
%%bash

#-------------------------------------------------------------
# 1. Peak calling avec normalisation des données WT par rapport à l'input WT
#-------------------------------------------------------------

#Peak calling - ChIP HSF2 chez un embryon WT HSF2

mkdir -p /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/WT/dedup

cd /mnt/g/Chip_pilote/results/03_mapping/SE_R1_very_sensitive

macs2 callpeak -t nh_ct_IP_WT_HSF2_SE_R1_VS_markdup.sorted.bam
--control nh_ct_Input_WT_SE_R1_VS_markdup.sorted.bam --name HSF2_WT_vs_Input
--outdir /mnt/h/Chip_pilote/results/05_PeakCalling/HSF2/WT/dedup
--format BAM --gsize mm --tsize 70 --qvalue 0.05 --keep-dup 1 -B


#Peak calling - ChIP DNMT3A chez un embryon WT HSF2

mkdir -p /mnt/h/Chip_pilote/results/05_PeakCalling/DNMT3A/WT/dedup

cd /mnt/h/Chip_pilote/results/03_mapping/SE_R1_very_sensitive

macs2 callpeak -t nh_ct_IP_WT_D3A_SE_R1_VS_markdup.sorted.bam
--control nh_ct_Input_WT_SE_R1_VS_markdup.sorted.bam --name D3A_WT_vs_Input
--outdir /mnt/h/Chip_pilote/results/05_PeakCalling/DNMT3A/WT/dedup
--format BAM --gsize mm --tsize 70 --qvalue 0.05 --keep-dup 1 -B

wc -l D3A_WT_vs_Input_peaks.narrowPeak
# -------> 105 D3A_WT_vs_Input_peaks.narrowPeak <--------




#-------------------------------------------------------------
# 2. Peak calling avec normalisation des données KO par rapport à l'input KO
#-------------------------------------------------------------

#Peak calling - ChIP ciblant HSF2 chez un embryon KO HSF2
mkdir -p /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup

cd /mnt/g/Chip_pilote/results/03_mapping/SE_R1_very_sensitive

macs2 callpeak -t nh_ct_IP_KO_HSF2_SE_R1_VS_markdup.sorted.bam
--control nh_ct_Input_KO_SE_R1_VS_markdup.sorted.bam --name HSF2_KO_vs_Input
--outdir /mnt/h/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup
--format BAM --gsize mm --tsize 70 --qvalue 0.05 --keep-dup 1 -B
```

```
#Peak calling - ChIP DNMT3A chez un embryon KO HSF2
mkdir -p /mnt/h/Chip_pilote/results/05_PeakCalling/DNMT3A/KO/dedup

cd /mnt/h/Chip_pilote/results/03_mapping/SE_R1_very_sensitive

macs2 callpeak -t nh_ct_IP_KO_D3A_SE_R1_VS_markdup.sorted.bam
--control nh_ct_Input_KO_SE_R1_VS_markdup.sorted.bam --name D3A_KO_vs_Input
--outdir /mnt/h/Chip_pilote/results/05_PeakCalling/DNMT3A/KO/dedup
--format BAM --gsize mm --tsize 70 --qvalue 0.05 --keep-dup 1 -B

wc -l D3A_KO_vs_Input_peaks.narrowPeak
# -------> 135 D3A_KO_vs_Input_peaks.narrowPeak <--------



#-------------------------------------
# 3. Suppression des "pics HSF2 WT" aussi détectés chez l'embryon HSF2 KO
# pour ne garder que les pics spécifiques
#-------------------------------------

# Pour ça, les données doivent être triées (sorted) : sort -k1,1 -k2,2n in.bed > in.sorted.bed

sort -k1,1 -k2,2 /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/WT/dedup/HSF2_WT_vs_Input_peaks.narrowP
eak
> /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/WT/dedup/HSF2_WT_vs_Input_peaks.narrowPeak_sorted.bed

sort -k1,1 -k2,2 /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup/HSF2_KO_vs_Input_peaks.narrowP
eak
> /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup/HSF2_KO_vs_Input_peaks.narrowPeak_sorted.bed

mkdir -p /mnt/g/Chip_pilote/results/06_PeakFilter/

bedtools intersect -v
-a /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/WT/dedup/HSF2_WT_vs_Input_peaks.narrowPeak_sorted.bed

-b /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup/HSF2_KO_vs_Input_peaks.narrowPeak_sorted.bed

> /mnt/g/Chip_pilote/results/06_PeakFilter/HSF2_WT_vs_Inp_KO_narrowPeak.bed

wc -l /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/WT/dedup/HSF2_WT_vs_Input_peaks.narrowPeak
wc -l /mnt/g/Chip_pilote/results/06_PeakFilter/HSF2_WT_vs_Inp_KO_narrowPeak.bed
# 346 /mnt/g/Chip_pilote/results/05_PeakCalling/HSF2/WT/dedup/HSF2_WT_vs_Input_peaks.narrowPeak
# 301 /mnt/g/Chip_pilote/results/06_PeakFilter/HSF2_WT_vs_Inp_KO_narrowPeak.bed
# -----> On perd 45 régions en recoupant avec l'échantillon KO pour HSF2 <------



#-------------------------------------
# 4. Suppression des "pics" présents dans la blacklist
#-------------------------------------

# Dézipper le fichier /mnt/g/Chip_pilote/results/06_PeakFilter/mm9-blacklist.bed.gz

#----------
# Pour ChIP HSF2 sur cortex WT HSF2
#il faut d'abord trier les fichiers :
cd /mnt/g/Chip_pilote/results/06_PeakFilter/
sort -k1,1 -k2,2n HSF2_WT_vs_Inp_KO_narrowPeak.bed > HSF2_WT_vs_Inp_KO_narrowPeak_sorted.bed

sort -k1,1 -k2,2n mm9-blacklist.bed > mm9-blacklist_sorted.bed

# Recoupement des fichiers
bedtools intersect -v -a HSF2_WT_vs_Inp_KO_narrowPeak_sorted.bed
-b mm9-blacklist_sorted.bed > HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed

wc -l HSF2_WT_vs_Inp_KO_narrowPeak.bed
wc -l HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed
# -----> 301 HSF2_WT_vs_Inp_KO_narrowPeak.bed <-----
# -----> 282 HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed <-----
# -----> On perd 19 pics en recoupant avec la Blacklist <-----
# -----> Après ces filtres, on a donc 282 pics HSF2 chez l'échantillon WT pour HSF2 <----


#------------
# Pour ChIP HSF2 sur cortex KO HSF2
```

```
cd /mnt/h/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup/markdup

sort -k1,1 -k2,2n HSF2_KO_vs_Input_peaks.narrowPeak > HSF2_KO_vs_Input_peaks.narrowPeak_sorted.bed

bedtools intersect -v -a HSF2_KO_vs_Input_peaks.narrowPeak_sorted.bed -b /mnt/h/Chip_pilote/results/06_
PeakFilter/mm9-blacklist_sorted.bed > /mnt/h/Chip_pilote/results/06_PeakFilter/HSF2_KO_vs_Inp_BL_narrow
Peak.bed

wc -l /mnt/h/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup/markdup/HSF2_KO_vs_Input_peaks.narrowPeak
wc -l /mnt/h/Chip_pilote/results/06_PeakFilter/HSF2_KO_vs_Inp_BL_narrowPeak.bed
# -----> 120 /mnt/h/Chip_pilote/results/05_PeakCalling/HSF2/KO/dedup/markdup/HSF2_KO_vs_Input_peaks.nar
rowPeak <-----
# -----> 26 /mnt/h/Chip_pilote/results/06_PeakFilter/HSF2_KO_vs_Inp_BL_narrowPeak.bed <-----
# -----> On perd 94 pics pics en recoupant avec la Blacklist  <-----


#------------
# Pour ChIP DNMT3A sur cortex WT HSF2
cd /mnt/h/Chip_pilote/results/05_PeakCalling/DNMT3A/WT/dedup/markdup

sort -k1,1 -k2,2n D3A_WT_vs_Input_peaks.narrowPeak > D3A_WT_vs_Input.narrowPeak_sorted.bed

mkdir /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9

bedtools intersect -v -a D3A_WT_vs_Input.narrowPeak_sorted.bed -b /mnt/h/Chip_pilote/results/06_PeakFil
ter/mm9-blacklist_sorted.bed > /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9/D3A_WT_vs_Inp_BL_nar
rowPeak.bed

wc -l D3A_WT_vs_Input.narrowPeak_sorted.bed
wc -l /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9/D3A_WT_vs_Inp_BL_narrowPeak.bed
# -----> 105 D3A_WT_vs_Input.narrowPeak_sorted.bed <-----
# -----> 22 /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9/D3A_WT_vs_Inp_BL_narrowPeak.bed <-----
# -----> On perd 83 pics en recoupant avec la Blacklist <-----



#------------
# Pour ChIP DNMT3A sur cortex KO HSF2
cd /mnt/h/Chip_pilote/results/05_PeakCalling/DNMT3A/KO/dedup/markdup

sort -k1,1 -k2,2n D3A_KO_vs_Input_peaks.narrowPeak > D3A_KO_vs_Input.narrowPeak_sorted.bed

mkdir -p /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9

bedtools intersect -v -a D3A_KO_vs_Input.narrowPeak_sorted.bed -b /mnt/h/Chip_pilote/results/06_PeakFil
ter/mm9-blacklist_sorted.bed > /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9/D3A_KO_vs_Inp_BL_nar
rowPeak.bed

wc -l D3A_KO_vs_Input.narrowPeak_sorted.bed
wc -l /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9/D3A_KO_vs_Inp_BL_narrowPeak.bed
# ----->135 D3A_KO_vs_Input.narrowPeak_sorted.bed  <-----
# ----->35 /mnt/h/Chip_pilote/results/06_PeakFilter/DNMT3A_mm9/D3A_KO_vs_Inp_BL_narrowPeak.bed  <-----
# -----> On perd 100 pics en recoupant avec la Blacklist <-----
```

Conversion des coordonnées des régions étudiées : mm9 vers mm10

Afin d'intégrer les résultats des différentes analyses, les coordonnées (*chr,start,end*) des régions enrichies pour HSF2, obtenues avec le génome de référence mm9 ont été converties en coordonnées mm10, grâce à l'outil **LiftOver** proposé par ***UCSC***. Les paramètres par défaut ont été conservés (*i.e.* au moins 95% des bases doivent être présentes après conversion).

J'ai utilisé pour celà, le fichier généré avec la commande suivante :

In [ ]:

```
%%bash

cd /mnt/g/Chip_pilote/results/06_PeakFilter
awk '{ print $1"\t"$2"\t"$3 }' HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed > bed_HSF2_WT_vs_Inp_KO_BL_narrowPe
ak.bed
```

Le fichier .bed obtenu a été nommé **mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed**. Sur les 282 régions enrichies en HSF2, 2 n'ont pas pu être converties :

In [ ]:

```
#Deleted in new (Sequence intersects no chains)
chrUn_random 5881413 5881590
#Split in new (Sequence insufficiently intersects multiple chains)
chrY_random 35398655 35398934
```

Un fichier comprenant à la fois les coordonnées mm9 et mm10 des pics enrichies en HSF2 a ensuite été généré à partir du fichier généré sur LiftOver et du fichier .bed des pics enrichies sur mm9, sans les 2 régions non converties sur le génome mm10 :

In [ ]:

```
%%bash

paste bed_sans_2reg_hors_mm10_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed
mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed > mm9_mm10_ChIP282_correspondance.txt
```

Recherche de motifs

L'outil **findMotifsGenome** (suite d'outil **HOMER**) permet d'identifier l'enrichissement en motifs connus (associés à un facteur de transcription) ou inconnus parmi les séquences des pics identifiés. Afin d'estimer la *pvalue* seuil à partir de laquelle la recherche de motifs s'avère peu fiable, cette recherche de motif a également été réalisée sur un jeu de données *"random"*, regroupant des faux pics générés aléatoirement, mais partageant les mêmes caractéristiques que le vrai jeu de données (*i.e* autant de pics et taille de pics identique). L'outil **bedtools shuffle** a été utilisé dans ce but, en utilisant la taille des chromosomes du génome mm9, obtenue sur le site **UCSC**.
La *pvalue* seuil est fixée à 0.02 pour la recherche de motifs connus, et à 1.10 $^{-8}$ pour la recherche *de novo* de Motifs. Ainsi, au delà de ces valeurs, les résultats sont considérés comme peu fiables

In [ ]:

```
%%bash

#-----------------------------------------------
#1. Recherche de motifs sur le vrai jeu de données
#-----------------------------------------------

#il faut d'abord installer le génome mm10
perl /home/Agathe/miniconda3/envs/chip/share/homer-4.9.1-6/.//configureHomer.pl -install mm10

#Recherche de motifs sur les données réelles
mkdir -p /mnt/h/Chip_pilote/results/07_motifs/mm10/real/

findMotifsGenome.pl /mnt/h/Chip_pilote/results/06_PeakFilter/mm10/mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowP
eak_sorted.bed mm10 /mnt/h/Chip_pilote/results/07_motifs/mm10/real -size given -mset vertebrates


#-----------------------------------------------
#2. Création d'un jeu de données aléatoires
#-----------------------------------------------

mkdir -p /mnt/h/Chip_pilote/results/07_motifs/mm10/random

bedtools shuffle -i /mnt/h/Chip_pilote/results/06_PeakFilter/mm10/mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowP
eak_sorted.bed -g /mnt/h/Chip_pilote/data/mm10_genome/mm10.chrom.sizes -noOverlapping > /mnt/h/Chip_pil
ote/results/07_motifs/mm10/random/random1_mm10_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed


#-----------------------------------------------
#3. Recherche de motifs sur les données générées aléatoirement
#-----------------------------------------------

findMotifsGenome.pl /mnt/h/Chip_pilote/results/07_motifs/mm10/random/random1_mm10_HSF2_WT_vs_Inp_KO_BL_
narrowPeak.bed mm10 /mnt/h/Chip_pilote/results/07_motifs/mm10/random -size given -mset vertebrates
```

Annotation syntaxique des régions enrichies (pics)

L'annotation syntaxique des régions d'intérêt (ayant les coordonnées du génome mm10) a été effectuée à l'aide du fichier d'annotation (biomart_mm10.txt) obtenu avec le package R **BiomaRt**. Ce fichier contient :

- *ensembl_gene_id* : Nom du gène
- *chromosome_name* : Numéro de chromosome
- *strand* : brin
- *start_position* : début du gène (pb)
- *end_position* : fin du gène (pb)
- *entrezgene_id* : ID NCBI du gène
- gene_biotype : type de gène
- mgi_symbol : symbole MGI
- entrezgene_accession : accession NCBI du gène
- entrezgene_description : description NCBI du gène
- uniprot_gn_symbol : symbole UniProtKB du gène

In [ ]:

```
%%R

#---------------------------------
# 1. Obtention du fichier biomart_mm10.txt
#---------------------------------

# Chargement du package BiomaRt :
library("biomaRt")

# 1. Sélection de la bonne base de données et du bon génome de référence
mm10 = useMart("ensembl", dataset="mmusculus_gene_ensembl")
# version Mus musculus mm10 utilisée :  Ensembl 97 Jul 2019 http://jul2019.archive.ensembl.org


# 2. Créer son jeu de données biomaRt
#Je suis obligée de faire en deux fois pour mm10,
#car Biomart n'autorise qu'un nombre restreint d'attributs :
annot_mm10_part1<-getBM(attributes=c("ensembl_gene_id","chromosome_name","strand", "start_position","en
d_position","entrezgene_id","gene_biotype","mgi_symbol","entrezgene_accession"), mart=mm10)

annot_mm10_part2<-getBM(attributes=c("ensembl_gene_id","entrezgene_description","uniprot_gn_symbol"), m
art=mm10)

annot_mm10 = merge(annot_mm10_part1, annot_mm10_part2, by="ensembl_gene_id", all=TRUE)

# Vérifications
dim(annot_mm10)
head(annot_mm10)

# Sauvegarde
write.table(annot_mm10,"G:/Chip_pilote/data/mm10_genome/biomart_mm10.txt", quote= FALSE, sep="\t",row.n
ames=FALSE)




#---------------------------------
# 2. Mise en forme du fichier Biomart
#---------------------------------

#Sur R:
biomart_mm10 =read.table("G:/methylome/fusion/annotation/AD/mm10/Biomart/biomart_mm10.txt",sep="\t", na
.strings = "NA", fill=TRUE, quote="", header=TRUE)

head(biomart_mm10)

#Remplacement des strand "-1" et "1" par "-" et "+"
for (i in 1:nrow(biomart_mm10)){
  if(biomart_mm10[i,"strand"]==1){
    biomart_mm10[i,"strand"]="+"
  }else if (biomart_mm10[i,"strand"]==-1){
    biomart_mm10[i,"strand"]="-"
  }else{
  }
}

#Moficiation du format de chromosomes :
biomart_mm10$chr = with(biomart_mm10,paste("chr",biomart_mm10[,"chromosome_name"],sep=""))

head(biomart_mm10)
```

```
head(biomart_mm10)

#Pour que bedtools intersect fonctionne, il faut un fichier .bed avec 9 colonnes.
# nb : je n'ai pas conservé la colonne uniprot_genename ni entrezgene_description
# qui semblent poser problèmes sur bedtools intersect (format du fichier mal détecté
# si on a cette colonne)
biomart_mm10_bis=biomart_mm10[,c("chr","start_position","end_position","strand","entrezgene_accession",
"gene_biotype", "mgi_symbol","ensembl_gene_id", "entrezgene_id")]

write.table(biomart_mm10_bis,"G:/methylome/fusion/annotation/AD/mm10/Biomart/biomart_mm10_all_info.bed"
, sep = "\t", col.names=TRUE, row.names=FALSE, quote=FALSE)
```

In [ ]:

```
%%bash

#----------------------------------
# 3. Tri (sorted) des fichiers
#----------------------------------

#Pour le fichier d'annotation Biomart :
cd /mnt/g/methylome/fusion/annotation/AD/mm10/Biomart/

sed '1d' biomart_mm10_all_info.bed | sort -k1,1 -k2,2n > biomart_mm10_all_info_sorted.bed

head biomart_mm10_all_info_sorted.bed

#Pour le fichier contenant les régions enrichies en HSF2 - coordonnées mm10
cd /mnt/g/Chip_pilote/results/06_PeakFilter/

sort -k1,1 -k2,2n mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed > mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPe
ak_sorted.bed

head mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak_sorted.bed


#----------------------------------
# 4. Recoupement des fichiers
#----------------------------------

mkdir -p /mnt/g/Chip_pilote/results/09_annotation/Biomart/mm10/

bedtools intersect -wao
-a /mnt/g/Chip_pilote/results/06_PeakFilter/mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak_sorted.bed
-b /mnt/g/Chip_pilote/data/mm10_genome/biomart_mm10_all_info_sorted.bed
> /mnt/g/Chip_pilote/results/09_annotation/Biomart/mm10/ChIP280_annot_syntaxique_mm10_biomart.bed

head /mnt/g/Chip_pilote/results/09_annotation/Biomart/mm10/ChIP280_annot_syntaxique_mm10_biomart.bed

wc -l /mnt/g/Chip_pilote/results/09_annotation/Biomart/mm10/ChIP280_annot_syntaxique_mm10_biomart.bed
#315 /mnt/g/Chip_pilote/results/09_annotation/Biomart/mm10/ChIP280_annot_syntaxique_mm10_biomart.bed
```

On obtient 315 informations d'annotation, pour 280 régions: dans certaines régions, différents gènes (ou autres éléments génomiques annotés) se chevauchent. Afin d'obtenir un fichier où chaque pic est représenté par une ligne unique, les éléments d'annotation d'une même région qui se trouvent sur des lignes séparées ont été regroupés.

In [ ]:

```
%%R

#Charger le jeu de données des "pics ChIP" annotés sur R.
ChIP280_annot = read.table(file =
            "E:/Chip_pilote/results/09_annotation/Biomart/mm10/ChIP280_annot_syntaxique_mm10_biomart.be
d", header = FALSE, sep = "\t", fill=TRUE)
head(ChIP280_annot)
dim(ChIP280_annot)

#Création de l'ID commun
ChIP280_annot_ID = cbind("ID" = paste(ChIP280_annot$V1, ChIP280_annot$V2, ChIP280_annot$V3, sep = ";"),
ChIP280_annot)
head(ChIP280_annot_ID)

# Regrouper pour avoir une région d'intérêt = une ligne.
# Principe : Pour chaque ID, collapser les différentes variables d'intéret pour améliorer le stockage d
e l'information
```

```
e l'information
library(dplyr)

# Liste d'identifiants uniques
liste_chrom_unique <- unique(ChIP280_annot_ID$ID)

# Matrice vide qui sera remplie par la suite avec
# Nombre lignes = nombre d'ID distincts | Nombre colonnes = nombre de variables
treatment_file <- data.frame(matrix(NA,nrow=length(liste_chrom_unique),
                                    ncol=ncol(ChIP280_annot_ID)))
colnames(treatment_file) <- colnames(ChIP280_annot_ID)

treatment_file$ID <- liste_chrom_unique
for(i in 1:length(liste_chrom_unique)){
  # Filtrer sur l'ID i
  ChIP280_filt <- ChIP280_annot_ID %>%
    filter(ID == liste_chrom_unique[i])

  # Pour l'ID i, on collapse les différentes éléments

  # Cas 1 : Si la variable k a une seule élément
  #      ==> on récupère uniquement, une seule fois, cette éléments
  # Cas 2 : Si la variable k a plusieurs éléments (même si plusieurs fois la même élément)
  #      ==> on récupère toute la série
  for(j in 2:length(colnames(treatment_file))){
    if(length(unique(ChIP280_filt[,j])) == 1){
      treatment_file[i,j] <- paste(unique(ChIP280_filt[,j]), collapse=';')
    } else {
      treatment_file[i,j] <- paste(ChIP280_filt[,j], collapse=';')
    }
  }
}


# Nommer les colonnes
colnames(treatment_file) = c("ID","chr_ChIP_peak","start_ChIP_peak","end_ChIP_peak","chr_annot", "start
_annot","end_annot", "strand","entrezgene_accession", "gene_biotype","mgi_symbol","ensembl_gene_id", "e
ntrezgene_id","overlap_length")

#Vérifications
head(treatment_file)
tail(treatment_file)

# Sauvegarde
write.table(treatment_file, "E:/Chip_pilote/results/09_annotation/Biomart/mm10/ChIP280_unique_annot_syn
taxique_mm10_biomart.bed", col.names = TRUE, sep = "\t", quote=FALSE, row.names=FALSE)
```

Annotation des régions enrichies (pics) en élément génomique

L'annotation des pics ayant les coordonnées du génome mm10 en élément génomique (intron, exons, régions intergéniques...) a été effectuée sur le site **Galaxeast** à l'aide du fichier de l'outil **homer_annotatePeaks**, ce qui a permis d'obtenir le fichier Galaxy287-homer_annotatePeaks_on_mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak_sorted_genome_mm10.csv, avec lequel la proportion de chaque élément génomique représenté par les pics détectés a été définie. La proportion dans le génome pris dans sa globalité a aussi été défini, grâce aux informations du papier de Achour et collaborateurs (Achour et al., 2015), en considérant leur catégorie "Upstream regulatory region" comme des régions intergéniques. De plus, lorsque le pic HSF2 est détecté dans un intron, le numéro de l'intron a été extrait.

In [ ]:

```
%%R

# ------------------------------------------------
# Annotation / Répartition des régions identifiées
# ------------------------------------------------

# Chargement du fichier obtenu sur R :
#1. Chargement des donénes
setwd("H:/Chip_pilote/results/09_annotation/mm10_AnnotatePeaks/")

ChIP_annot_HOMER = read.delim("Galaxy287-homer_annotatePeaks_on_mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPea
k_sorted_genome_mm10.csv", header = TRUE)

head(ChIP_annot_HOMER)
dim(ChIP_annot_HOMER)
```

```
#2. Extraction de la colonne contenant l'annotation exon, intron, promoter...
annotation = as.data.frame(ChIP_annot_HOMER[,"Annotation"])
head(annotation)

annotation_light = as.data.frame(sapply(strsplit(as.character(annotation[,1])," "), `[`, 1))
head(annotation_light)

#Comptage
summary(annotation_light)
annotation_count = table(annotation_light)
annot_count_tf = transform(annotation_count)

#Graphe
annot_count_tf$prop = with(annot_count_tf, round((annot_count_tf[,2]/sum(annot_count_tf[,2])), digits=3
))
annot_count_tf[,1]=as.vector(annot_count_tf[,1])
annot_count_tf[1,1]="5'UTR"

donut = ggplot(annot_count_tf, aes(x = 1.5, y = Freq, fill = annotation_light)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  #geom_text(aes(y = Freq, label = prop), color = "white") +
  theme_void()  + labs(fill = "Annotation") + xlim(0.5, 2.5) + ggtitle("Eléments génomiques ciblés par
HSF2") +
  theme(plot.title = element_text(hjust = 0.5))
donut

ggsave("Eléments_ciblés_par_HSF2.pdf",donut)
ggsave("Eléments_ciblés_par_HSF2.png",donut)


# ------------------------------------------------
# Numéro de l'intron
# ------------------------------------------------
# 3. D'autre part : Extraction des lignes contenant "intron"
ind_with_intron <- which(grepl("intron",annotation[,1]))
test_intron <- annotation[ind_with_intron,,drop=FALSE]
colnames(test_intron) = "intron"
# NB: Il faut rajouter "drop=FALSE" car sinon en subsettant on obtient un vecteur au lieu d'un datafram
e

# 4. Récupérer le numéro de l'intron
# Pour ça : splitter les libellés avec le pattern " ", puis récupérer l'indice où il y a le nombre rech
erché
for(i in 1:nrow(test_intron)){
  test_intron$number_intron[i] <- strsplit(as.character(test_intron$intron)," ")[[i]][4]
}
# NB: Cela marche si c'est toujours en 4è position lors du split, sinon il faut adapter

# 5. Comptage des occurences, selon l'intron pec
library(dplyr)
typo_intron <- test_intron %>%
  group_by(number_intron) %>%
  summarise(number_intron_type = n()) %>%
  as.data.frame()
# On groupe par numéro d'intron, puis on vient compter le nombre d'occurences

typo_intron_mef= typo_intron[order(as.numeric(typo_intron$number_intron)),]
typo_intron_mef


#4.Graphes :

#Barplot, numéro d'intron où se trouve HSF2
library(ggplot2)
p <-ggplot(data=typo_intron_mef, aes(x=as.numeric(number_intron), y=number_intron_type)) +
  geom_bar(stat="identity", fill="steelblue") + theme_minimal() + scale_x_discrete(limits=1:18) +
  ggtitle("Numéro de l'intron enrichi pour HSF2") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Numéro de l'intron") + ylab("occurence")
p
ggsave("numéro_intron_enrichi_pour_HSF2.pdf", p)
ggsave("numéro_intron_enrichi_pour_HSF2.png", p)
```

```
# ------------------------------------------------
# Annotation / Répartition du génome global
# ------------------------------------------------

# Chargement des informations sur R :

setwd("H:/Chip_pilote/results/09_annotation/mm10_AnnotatePeaks/")

annot_global = data.frame(matrix(NA, ncol=2, nrow=4))
nom_category = c("exon", "intergenic","intron","promoter")
value_category = c(3,66,29,2)
annot_global[,1] = as.character(nom_category)
annot_global[,2]= value_category
annot_global

#Comptage
table(annot_global)
annot_global_tf = transform(annot_global)

#Graphe
annot_global_tf$prop = with(annot_global_tf, round((annot_global_tf[,2]/sum(annot_global_tf[,2])), digi
ts=3))
annot_global_tf[,1]=as.vector(annot_global_tf[,1])

library(ggplot2)
donut = ggplot(annot_global_tf, aes(x = 1.5, y = prop, fill = X1)) +
  geom_bar(width = 1, stat = "identity", color = "white") +
  coord_polar("y", start = 0)+
  #geom_text(aes(y = Freq, label = prop), color = "white") +
  theme_void()  + labs(fill = "Annotation") + xlim(0.5, 2.5) + ggtitle("Distribution globale dans le gé
nome") +
  theme(plot.title = element_text(hjust = 0.5))
donut

ggsave("distribution_genome_wide- sans Upstr. regul. reg.pdf",donut)
ggsave("distribution_genome_wide- sans Upstr. regul. reg.png",donut)
```

Création d'un tableau regroupant l'ensemble des informations d'intérêt

Un tableau global, regroupant à la fois les informations statistiques du *Peak Calling* effectué par **MACS2**, l'annotation syntaxique Biomart et l'annotation provenant de **homer_annotatePeaks** a été généré :

In [ ]:

```
%%bash

# Pour avoir un ID commun, réattribution des coordonnées mm10 au fichier issu du peak calling :
mkdir -p /mnt/h/Chip_pilote/results/06_PeakFilter/mm10/

cd /mnt/h/Chip_pilote/results/09_annotation/Biomart/mm10/

sed '1d' ChIP280_unique_annot_syntaxique_mm10_biomart.bed
> nh_ChIP280_unique_annot_syntaxique_mm10_biomart.bed

cd /mnt/h/Chip_pilote/results/06_PeakFilter

paste bed_sans_2reg_hors_mm10_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed
/mnt/h/Chip_pilote/results/09_annotation/Biomart/mm10/nh_ChIP280_unique_annot_syntaxique_mm10_biomart.b
ed
> /mnt/h/Chip_pilote/results/09_annotation/Biomart/mm10/
nh_mm9_mm10_ChIP280_unique_annot_syntaxique_biomart.bed

head /mnt/h/Chip_pilote/results/09_annotation/Biomart/mm10/
nh_mm9_mm10_ChIP280_unique_annot_syntaxique_biomart.bed

wc -l /mnt/h/Chip_pilote/results/09_annotation/Biomart/mm10/
nh_mm9_mm10_ChIP280_unique_annot_syntaxique_biomart.bed
#280 /mnt/h/Chip_pilote/results/09_annotation/Biomart/mm10/
#nh_mm9_mm10_ChIP280_unique_annot_syntaxique_biomart.bed
```

In [ ]:

```
%%R

# Charger le jeu de données des "pics CHIP" annotés sur R.
ChIP280_annot = read.table(file =
"H:/Chip_pilote/results/09_annotation/Biomart/mm10/nh_mm9_mm10_ChIP280_unique_annot_syntaxique_biomart.
bed",header = FALSE, sep = "\t", fill=TRUE)
head(ChIP280_annot)
dim(ChIP280_annot)

# Charger le jeu de données issu de MACS2.
ChIP280_stat = read.table(file = "H:/Chip_pilote/results/06_PeakFilter/HSF2_WT_vs_Inp_KO_BL_narrowPeak.
bed",
                          header = FALSE, sep = "\t", fill=TRUE)
head(ChIP280_stat)
dim(ChIP280_stat)

# Création de l'ID
ChIP280_annot$ID = with(ChIP280_annot, paste(ChIP280_annot$V1, ChIP280_annot$V2, ChIP280_annot$V3, sep
= ";"))
head(ChIP280_annot)

ChIP280_stat$ID = with(ChIP280_stat, paste(ChIP280_stat$V1, ChIP280_stat$V2, ChIP280_stat$V3, sep = ";"
))
head(ChIP280_stat)


# Pooler les deux jeux de données
ChIP_complet= merge(ChIP280_stat, ChIP280_annot, by="ID")
head(ChIP_complet)
dim(ChIP_complet)

# Réorganisation du jeu de données - conservation des colonnes d'intérêt
ChIP_complet_save = ChIP_complet[,c("V1.x","V2.x", "V3.x", "ID",
                                    "V5.y", "V6.y", "V7.y", "V4.y",
                                    "V4.y", "V5.x", "V6.x",  "V7.x", "V8.x", "V9.x",
                                    "V10.x","V8.y", "V9.y", "V10.y", "V11", "V12",
                                    "V13", "V14", "V15", "V16", "V17")]
colnames(ChIP_complet_save)=c("mm9_peak_chr","mm9_peak_start","mm9_peak_end","mm9_peak_ID","mm10_peak_c
hr","mm10_peak_start","mm10_peak_end","mm10_peak_ID","peak_name","MACS2_score","MACS2_6th","MACS2_FC",
"MACS2_logpval","MACS2_logqval", "MACS2_summit","chr_annot","start_annot","end_annot", "strand", "entre
zgene_accession", "gene_biotype","mgi_symbol", "ensembl_gene_id", "entrezgene_id","overlap_length")

#MACS2_score : integer score for display. It's calculated as int(-10*log10pvalue) or int(-10*log10qvalu
e) depending on whether -p (pvalue) or -q (qvalue) is used as score cutoff. Please note that currently
this value might be out of the [0-1000] range defined in UCSC Encode narrowPeak format. You can let the
value saturated at 1000 (i.e. p/q-value = 10^-100) by using the following 1-liner awk: awk -v OFS="\t"
'{$5=$5>1000?1000:$5} {print}' NAME_peaks.narrowPeak
#MACS2_FC : fold-change at peak summit
#MACS2_logpval: -log10pvalue at peak summit
#MACS2_logqval: -log10qvalue at peak summit
#MACS2_summit: relative summit position to peak start

head(ChIP_complet_save)

# Sauvegarde
write.table(ChIP_complet_save, "H:/Chip_pilote/results/09_annotation/Biomart/mm10/mm9_mm10_ChIP280_uniq
ue_annot_biomart_et_MACS2.txt",
            col.names = TRUE, sep = "\t", quote=FALSE, row.names=FALSE)




#Recoupement regroupant l'ensemble des informations (peakcalling + annotation Biomart)
# avec le fichier annoté sur Galaxeast (homer_annotatePeaks)

#1. Chargement de l'annotation provenant de HOMER
setwd("H:/Chip_pilote/results/09_annotation/mm10_AnnotatePeaks")

HOMER = read.table("Galaxy287-[homer_annotatePeaks_on_mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak_sorted_g
enome_mm10].csv", sep="\t", header=TRUE, fill=TRUE, quote="")
head(HOMER)
dim(HOMER)
```

```
# homer_annotatePeaks ajoute une base à la valeur du début de la région (start).
# Je dois donc enlever une base de chaqe start, pour retrouver les ID du fichier ChIP280 annoté
HOMER$Start_revu = HOMER$Start-1

#Création d'un ID (chr-start-end) et sélection des colonnes d'intérêts
HOMER$ID = with(HOMER, paste(Chr,Start_revu,End,sep=";"))
HOMER_select = HOMER[, c("ID","Annotation","Detailed.Annotation",
                         "Distance.to.TSS", "Nearest.PromoterID", "Entrez.ID", "Nearest.Unigene",
                         "Nearest.Refseq","Nearest.Ensembl","Gene.Name")]
dim(HOMER_select)
head(HOMER_select)



#2. Chargement du fichier annoté des chIP280
setwd("H:/Chip_pilote/results/09_annotation/Biomart/mm10/")
ChIP280 = read.table("nh_mm9_mm10_ChIP280_unique_annot_syntaxique_biomart.bed", sep="\t", header=FALSE,
fill=TRUE, quote="")
head(ChIP280)
dim(ChIP280)

ChIP280$ID = with(ChIP280,paste(V5,V6,V7,sep=";"))
colnames(ChIP280)[18]="ID"


#3. Combinaisons des données
both = merge(ChIP280,HOMER_select,by="ID")
head(both)
dim(both)



#4. Réorganisation des colonnes
colnames(both)=c("ChIP_mm10_ID", "ChIP_mm9_chr","ChIP_mm9_start", "ChIP_mm9_end", "ChIP_mm9_ID",
                 "ChIP_mm10_chr","ChIP_mm10_start", "ChIP_mm10_end",
                 "ChIP_Annot_chr", "ChIP_Annot_start", "ChIP_Annot_end","ChIP_Annot_strand",
                 "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID",
                 "Entrez_ID", "annot_overlap_length","Annotation","Detailed.Annotation",
                 "Distance.to.TSS", "Nearest.PromoterID", "Entrez.ID", "Nearest.Unigene",
                 "Nearest.Refseq","Nearest.Ensembl","Gene.Name")

both_reorganisation = both[,c("ChIP_mm9_chr","ChIP_mm9_start", "ChIP_mm9_end", "ChIP_mm9_ID",
                              "ChIP_mm10_chr","ChIP_mm10_start", "ChIP_mm10_end","ChIP_mm10_ID",
                              "ChIP_Annot_chr", "ChIP_Annot_start", "ChIP_Annot_end","ChIP_Annot_strand
",
                              "entrezgene_accession", "gene_biotype", "mgi_symbol", "ENSEMBL_ID",
                              "Entrez_ID", "annot_overlap_length","ChIP_mm10_ID","Annotation","Detailed
.Annotation",
                              "Distance.to.TSS", "Nearest.PromoterID", "Entrez.ID", "Nearest.Unigene",
                              "Nearest.Refseq","Nearest.Ensembl","Gene.Name")]

write.table(both_reorganisation, "mm9_mm10_ChIP280_unique_annot_syntaxique_biomart_HOMER.bed",col.names
= TRUE, sep = "\t", quote=FALSE, row.names=FALSE)
```

Gene Ontology (GO)

Afin de réaliser une analyse d'ontologies des gènes d'intérêt avec l'outil ToppFun de la suite d'outil **ToppGene**, les identifiants ID_ENSEMBL correspondants aux régions enrichies pour HSF2 ont été extraits du tableau global d'informations :

In [ ]:

```
%% R

# Récupération des ID ENSEMBL - annotation mm10 BIOMART :
#Pour déconcatener les matrices, selon une colonne donnée
ChIP_decont = as.character(ChIP_complet_save$ensembl_gene_id)
head(ChIP_decont)

#Récupération de la liste des éléments d'intérêts séparés
list_gene_splitted <- strsplit(as.character(ChIP_complet_save$ensembl_gene_id),split=";")

vec_gene_splitted <- unlist(list_gene_splitted)
vec_gene_splitted
```

```
#suppression des "." = information d'annotation non disponible
vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

#Présentation des données sous forme de tableau
table_VGS_filter = as.data.frame(vec_gene_splitted_filter)

write.table(table_VGS_filter, "H:/Chip_pilote/results/09_annotation/Biomart/mm10/ENS_ID_ChIP280_only.tx
t", quote=FALSE, col.names=FALSE, row.names=FALSE)
```

3.6. Workfow de l'intégration des DMRs avec les cibles de HSF2 identifiées par ChIP-seq

Analyse bio-informatique intégrée :

Intégration des données de l'analyse de la capture du méthylome et du *ChIP-seq* ciblant HSF2

Agathe Duchateau

Pour déterminer l'implication de HSF2 dans la mise en place des défauts de méthylation de l'ADN observés après l'EPA, les 432 régions différentiellement méthylées après EPA, (**DMR432**, identifiées avec l'analyse de la capture du méthylome) et les 280 régions ciblées par HSF2 sous EPA (**ChIP280**, résultats du *ChIP-seq* ciblant HSF2) ont été intégré afin d'identifier une éventuelle corrélation.

Croisement des DMR et des régions ciblées par HSF2 selon leurs régions chromosomiques

Pour identifier les gènes qui ont à la fois une région ciblée par HSF2 sous EPA et une DMR après ce stress, les coordonnées des régions identifiées dans les deux jeux de données ont été comparés, en utilisant une fonction que nous avons générée sur R (fonction ***find_overlaps_AD_table()***, en collab. avec Pr. S. Ott, Univ. de Warwick, UK). L'argument *tolerance* a été fixé à 5000, pour regrouper les régions, en tolérant une distance maximale de 5000 bases entre entre les régions.

In [ ]:

```
%%R

#################################################################
# Croisement des régions ChIP280 avec les DMR432 - par coordonnées chromosomiques
# (coordonnées mm10 dans les deux cas)
#################################################################

# Chargement du fichier contenant les régions ciblées par HSF2 après EPA
ChIP280 = read.table("G:/Chip_pilote/results/06_PeakFilter/mm10_bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed
", header = FALSE, sep= "\t")
head(ChIP280)
dim(ChIP280)

# Chargement du fichier contenant les DMR après EPA
DMR432 = read.table("G:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/DMR432
_unique_annot_mm10_BMT_meth_cgi.bed", sep="\t", header = FALSE)
head(DMR432)
DMR432_bed = DMR432[,c("V1","V2","V3")]
head(DMR432_bed)
dim(DMR432_bed)

# Fonction permettant d'extraire les régions présentes dans les deux jeux de données :
find_overlaps_AD_table <- function (regions_1,regions_2,tolerance=0)
{
  common_region_table = NULL
  number_of_regions_1 <- nrow(regions_1)
  number_of_regions_2 <- nrow(regions_2)

  overlapped_is <- c()
  overlapped_js <- c()

  for (i in 1:number_of_regions_1) {

    for (j in 1:number_of_regions_2) {

      if (as.character(regions_1[i,1])==as.character(regions_2[j,1])) {
        start_1 <- as.numeric(regions_1[i,2])
        end_1 <- as.numeric(regions_1[i,3])
        expanded_start_1 <- start_1 - tolerance
        expanded_end_1 <- end_1 + tolerance
        start_2 <- as.numeric(regions_2[j,2])
        end_2 <- as.numeric(regions_2[j,3])

        if (expanded_end_1>=start_2) {

          if (end_2>=expanded_start_1) {
```

```
            overlapped_is <- c(overlapped_is,i)
            overlapped_js <- c(overlapped_js,j)

            first_region <- paste(regions_1[i,1],start_1,end_1,sep="_")
            second_region <- paste(regions_2[j,1],start_2,end_2,sep="_")
            overlap_size <- min(end_1,end_2)-max(start_1,start_2)+1
            common_region_table = rbind(common_region_table,
                        c(as.character(regions_1[i,1]),start_1,end_1,
                       as.character(regions_2[j,1]),start_2,end_2, overlap_size))

            if (overlap_size>0) {
              print(paste(first_region,second_region,sep=" "))
            } else {
              print(paste(first_region,second_region, sep=" "))
            }
          }
        }
      }
    }
  }
  print(paste("Number matched in first data set:",length(unique(overlapped_is))))
  print(paste("Number matched in second data set:",length(unique(overlapped_js))))
  if (length(common_region_table) > 0){
    colnames(common_region_table) = c("chr_reg1", "start_reg1", "end_reg1", "chr_reg2", "start_reg2", "
end_reg2","overlap_size")
  }else{
  }
  return(as.data.frame(common_region_table))
}


# Utilisation de la fonction pour trouver les régions
# communes entre DMR et cibles de HSF2:
ChIP_DMR = find_overlaps_AD_table(DMR432_bed,ChIP280,tolerance=5000)
```

Croisement des DMR et des régions ciblées par HSF2 selon les noms de gènes

Pour identifier les gènes qui ont à la fois une région ciblée par HSF2 sous EPA et une DMR après ce stress, les noms des gènes des régions identifiées dans les deux jeux de données ont été extraits et comparés. Pour extraire le nom des gènes des jeux de données, nous avons utilisé notre propre fonction (fonction ***unlist-gene()***) permettant d'obtenir un seul nom de gène par ligne - les données ayant été concaténées avant pour avoir une région unique par ligne. La comparaison des listes de gènes obtenus a ensuite été réalisée sur Excel avec la fonction RECHERCHEV.

In [ ]:

```
%%R

#################################################################
# Croisement des régions ChIP280 avec les DMR432 - par nom de gènes
# (coordonnées mm10 dans les deux cas)
#################################################################

# 1.Chargement du fichier contenant les régions ciblées par HSF2 après EPA
setwd("H:/Chip_pilote/results/09_annotation/Biomart/mm10")

ChIP280 = read.table("nh_mm9_mm10_ChIP280_unique_annot_syntaxique_biomart.bed", header = FALSE, fill=TR
UE, quote="", sep="\t", na.strings = "", stringsAsFactors = FALSE)
head(ChIP280)
dim(ChIP280)

# Renommer les colonnes le nécessitant
colnames(ChIP280)[12] = "entrezgene_accession"
colnames(ChIP280)[14] = "mgi_symbol"
colnames(ChIP280)[15] = "ENSEMBL_ID"
colnames(ChIP280)[16] = "Entrez_ID"

# 2.Chargement du fichier contenant les DMR après EPA
setwd("H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10")

DMR432 = read.table("DMR432_uniq_annot_mm10_BMT_meth_cgi_capt_dist500_clean_Pcdh.txt", header = TRUE, f
ill=TRUE, quote="", sep="\t", na.strings = "", stringsAsFactors = FALSE)
```

```
head(DMR432)
dim(DMR432)

# 3.Fonction permettant d'extraire les noms des gènes du jeu de données
unlist_gene = function(treatment_file, gene_ID, file_name, folder){
  # Pour extraire les identifiants = nom de gènes
  gene_name_ID = as.character(treatment_file[,gene_ID])

  # Pour obtenir une liste où chaque ligne correspond à un nom de gène
  list_gene_splitted <- strsplit(gene_name_ID,split=";")
  vec_gene_splitted <- unique(unlist(list_gene_splitted))

  # Suppression des données manquantes = "." :
  vec_gene_splitted_filter = vec_gene_splitted[vec_gene_splitted!="."]

  # Obtention de la liste (sous forme de tableau, ne contenant qu'une colonne):
  table_VGS_filter = as.data.frame(vec_gene_splitted_filter)

  write.table(table_VGS_filter, paste(folder,gene_ID,"_",file_name,".txt",sep=""),
              quote=FALSE, col.names=FALSE, row.names=FALSE)
}


# 4.Extraction des différents identifiants du jeu de données DMR432
# (les identifiants ENSEMBL, Entrez, ou mgi)
# Extraction des ID ENSEMBL
unlist_gene(treatment_file=DMR432, gene_ID="ENSEMBL_ID", file_name="DMR432",
            folder="H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/ID_list
/")

# Extraction des ID Entrez accession
unlist_gene(treatment_file=DMR432, gene_ID="entrezgene_accession", file_name="DMR432",
            folder="H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/ID_list
/")

# Extraction des ID mgi
unlist_gene(treatment_file=DMR432, gene_ID="mgi_symbol", file_name="DMR432",
            folder="H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/ID_list
/")

# Extraction des ID Entrez_gene_ID
unlist_gene(treatment_file=DMR432, gene_ID="Entrez_ID", file_name="DMR432",
            folder="H:/methylome/fusion/post_methylkit/DMR432_pval0.07/annotations/Biomart/mm10/ID_list
/")



# 5.Extraction des différents identifiants du jeu de données ChIP280
# (les identifiants ENSEMBL, Entrez, ou mgi)
# Extraction des ID ENSEMBL
unlist_gene(treatment_file=ChIP280, gene_ID="ENSEMBL_ID", file_name="ChIP280",
            folder="H:/Chip_pilote/results/09_annotation/Biomart/mm10/ID_list/")

# Extraction des ID Entrez accession
unlist_gene(treatment_file=ChIP280, gene_ID="entrezgene_accession", file_name="ChIP280",
            folder="H:/Chip_pilote/results/09_annotation/Biomart/mm10/ID_list/")

# Extraction des ID mgi
unlist_gene(treatment_file=ChIP280, gene_ID="mgi_symbol", file_name="ChIP280",
            folder="H:/Chip_pilote/results/09_annotation/Biomart/mm10/ID_list/")

# Extraction des ID Entrez_gene_ID
unlist_gene(treatment_file=ChIP280, gene_ID="Entrez_ID", file_name="ChIP280",
            folder="H:/Chip_pilote/results/09_annotation/Biomart/mm10/ID_list/")


# 6. Sur Excel, Chargement des deux listes de gènes selon ID entrezgene_accession
# Avec la fonction RECHERCHEV:
# recherche des nom de gènes communs entre les deux listes.
```

Proportion de régions ciblées par HSF2 après EPA dans la capture du méthylome

Pour détiminer la proportion de régions ciblées par HSF2 après EPA qui est représentée dans la capture du méthylome, les

régions ciblées par HSF2 ont été croisées avec le fichier .bed contenant l'ensemble des régions capturées. Cette analyse a été réalisée avec la fonction *find_overlaps_AD_table()* présentée ci dessus. Les coordonnées mm9 des régions "ChIP280" ont été utilisés, les régions de la capture du méthylome ayant été obtenue selon ce génome de référence.

In [ ]:

```
%%R

# Chargement du fichier de la capture (fichier de base - region non agrandie) :
Capture = as.matrix(read.table("G:/methylome/fusion/capture/capture_regions/151019_MM9_EDC_RM_EPI_captu
re_targets.bed", sep="\t", header=FALSE))

# Suppression des espaces
Capt_propre = gsub(" ","",Capture)
head(Capt_propre)

# La capture est au format mm9.
# Il faut donc charger la version mm9 des pics ChIP filtrés.
# (i.e. recoupés avec la blacklist et les pics détectés chez Hsf2KO) :
ChIP_mm9 = read.table("G:/Chip_pilote/results/06_PeakFilter/bed_HSF2_WT_vs_Inp_KO_BL_narrowPeak.bed", h
eader = FALSE, sep= "\t")

Capt_chIP= find_overlaps_AD_table(ChIP_mm9,Capt_propre,tolerance=0)

write.table(Capt_chIP,"G:/Chip_pilote/results/08_croisement/mm9_ChIP_all_Capture_non_agrandie.txt", sep
= "\t", quote=FALSE, row.names = FALSE, col.names=FALSE)
```

Annexes 4 : Listes des régions d'intérêts

4.1. Tableaux des régions identifiées dans chaque analyse (DMRs, ChIP-seq, DOCR, DEG)

Voir lien → https://drive.google.com/open?id=1knfPt5j5khC31QDSqWAyuK0ELG0NYKl1

Ce lien renvoie vers un fichier contenant :

- La liste des **DMRs** annotées
- La liste des régions **ciblées par HSF2** sous EPA (résultats du ChIP-seq ciblant HSF2)
- Les listes des **DOCR** annotées : régions différentiellement ouvertes ou fermées au cours du développement cérébral, en conditions physiologiques (entre chaque stade de développement pris deux à deux)
- Les listes des **DEG** annotées : régions différentiellement exprimées au cours du développement cérébral, en conditions physiologiques (entre chaque stade de développement pris deux à deux)

4.2. Tableaux des régions identifiées dans les analyses intégrées (croisement des données)

Voir lien → https://drive.google.com/open?id=1XNGUYyH28Yaw_nRCRJd9VUUlrNJT468f

Ce lien renvoie vers un fichier contenant la liste des régions retrouvées dans plusieurs jeux de données à savoir :

- Les listes des **DMRs étant aussi DOCR** : régions différentiellement méthylées après EPA, qui, en conditions physiologiques, sont différentiellement ouvertes ou fermées dans la période encadrant le stress
- Les listes des **DMRs étant aussi DEG** : régions différentiellement méthylées après EPA associées à au moins un gène qui, en conditions physiologiques, est différentiellement exprimé dans la période encadrant le stress