Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN

Jennifer Sengenès

Thèse de doctorat de l'Université Paris VI

Discipline : Sciences du Vivant

Spécialité : Epigénétique

Réalisée dans le laboratoire d'épigénétique au CEA/CNG

Sous la direction de M. Ivo Gut, encadrée par M. Jörg Tost

Présentée par

Jennifer SENGENÈS

Pour obtenir le grade de Docteur de l'Université Paris VI

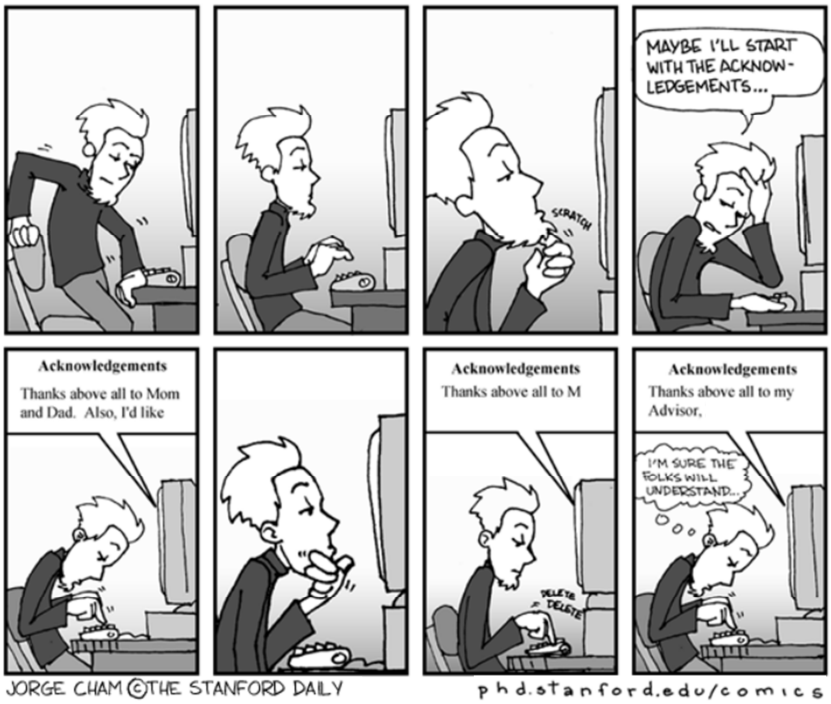
Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN

Soutenue le 30 Mars 2012 devant le jury composé de :

M. Thierry Foulon	Président
M. Patrick Descombes	Rapporteur
M. Zdenko Herceg	Rapporteur
M. Pierre-François Cartron	Examinateur
M. Stéphane Le Crom	Examinateur
M. Michael Weber	Examinateur
M. Jörg Tost	Co-directeur de thèse

Remerciements

A Jörg pour m'avoir confié ce projet, accompagnée et conseillée dans sa réalisation et pour la confiance accordée pendant la durée de ces travaux en m'autorisant à en gérer bon nombre d'aspects de façon autonome. A Ivo pour m'avoir permis d'intégrer son groupe en 2008.



Aux membres du jury qui ont accepté de juger ces travaux : à Messieurs Patrick Descombes et Zdenko Herceg en qualité de rapporteurs, Pierre-François Cartron, Stéphane Le Crom et Michael Weber en tant qu'examinateurs et Thierry Foulon comme Président ainsi que pour sa disponibilité en tant que référent à Paris VI.

A Flo qui a partagé avec moi, pour les avoir vécus en direct, les envies de meurtre sur mon ordinateur, les allergies passagères à ma thèse, les découragements occasionnels mais aussi les « eurêka », les soupirs de soulagement, les bonds de joie et autres cris du cœur ! Merci aussi pour les designs en urgence.

A Christian, avec qui nous sommes passés maintes fois au bord d'un « petit-suiSSide » quand nos ADNs refusaient toute coopération mais avec qui ce fut un plaisir de travailler.

A Marion qui aura connu le CNG sur la même période que moi, et qui a joué un rôle important notamment en me transmettant son savoir-faire des sample-preps.

Aux autres membres du groupe Epigénétique. A Antoine avec qui nous avons mis en place le CQ du MeDIP pour mon premier article. A Sven pour les analyses de MeDIP-chip et Nizar pour celles sur BiQ ainsi que pour les discussions illustrées par graphes et équations sur les paillasses ! A Mikael, Alex, Nicolas, Aurélie et Anne avec qui je n'ai pas eu l'occasion de travailler directement mais qui ont contribué à l'ambiance agréable qui a régné au sein de l'équipe.

A Marie-Thérèse et toute l'équipe Solexa puisque je ne sais entre les mains de quelles personnes sont passés mes échantillons. A Marie-Ange et Yann pour leurs précieuses explications détaillées de la technologie Illumina sur les GAIIx.

A Jinyan et Victor pour les analyses de MeDIP-Seq ainsi que pour les réflexions et leur mise en œuvre qui nous ont menés à la création de la plateforme MeQA et à la publication de l'article correspondant.

A Lotte, Elin et Mats pour m'avoir accueillie dans leur laboratoire à Uppsala dans le cadre du projet Sélectors ; cette expérience suédoise aura été marquante. A Marc, Céline et Christophe pour le séquençage sur le GS Junior.

A nouveau à Marie-Ange, Victor, Christian et Nizar pour les relectures et commentaires judicieux sur certaines parties de ce manuscrit. A Flo pour avoir porté un regard critique sur l'intégralité de cet ouvrage. A Jörg pour m'avoir aiguillée pendant la rédaction de ce manuscrit et pour avoir consacré du temps à la relecture en période d'heureux évènement. A Suzanne et Nath pour leurs « trucs et astuces » concernant la gestion des images, la mise en page et la logistique finale.

A tous ceux qui m'ont permis de concrétiser mes diverses actions de communication scientifique, notamment aux acteurs de « Tu meurs ? » qui ont accepté de se prêter au jeu, parfois en endossant des rôles étranges sortis de mon imagination (« tu veux que je sois une tumeur du colon ???! »).

A toutes les personnes de l'étage et du centre que je ne me risquerai pas à citer de peur d'omettre certains noms, mais qui ont contribué au bon déroulement de ma thèse et à une bonne ambiance générale, que ce soit par un coup de main au niveau administratif, un échange instructif, un mot d'encouragement, un bon moment à la pause déjeuner, la découverte de la salsa ou de l'AEDLP, ou encore la solidarité sur les quais du RER D en temps d'inondation (« mais… il a plu ? »), de feuilles mortes sur les voies (« en août ??? ») ou d'embouteillages de trains…

A ma famille et mes proches, pour m'avoir soutenue dans les périodes les plus éprouvantes et m'avoir donné toujours plus d'énergie, de motivation et d'envie de poursuivre efforts et sacrifices pour mener à bien ce projet. A Boubou tout particulièrement.

Sommaire

Listes des illustrations

Liste des figures

Liste des tableaux

Abréviations et acronymes

A

ADN : Acide Désoxyribonucléique

ADNc : ADN complémentaire

ARN : Acide Ribonucléique

B

B&W: *Binding and washing*

BC-Seq : Bisulphite suivi d'une Capture et du Séquençage

BET : Bromure d'éthidium

BSA: *Bovine Serum Albumin*

BSPP : *Bisulphite Padlock Probe*

BS-Seq : *Shotgun Bisulphite Sequencing*

BWA: *Burrows-Wheeler Aligner*

C

ChIP : *Chromatin ImmunoPrecipitation*

CpG : Dinucléotide contenant une cytosine directement suivi d'une guanine

Ct : *Cycles threshold* (cycle seuil en qPCR)

Cy : Cyanine

D

DIB: *DNA Isolation Buffer*

DMR: *Differentially Methylated Region*

DNMT: ADN méthyltransférase

dNTP: Désoxyribonucléotide triphosphate

ddNTP: Didésoxyribonucléotide triphosphate

E

EtOH: Ethanol

F

FC : *Flow cell*

FDA: *Food and Drug Administration*

G

GA: *Genome Analyzer* (séquenceur Illumina)

GST : *Glutathione-S-Transferase*

H

5-hmC: 5-hydroxyméthylcytosine

hMeDIP: MeDIP ciblant les h-5mCs

HapMap (individu): *Haplotype Map*

HDAC: Histone Déacétylase

HELP: *HpaII tiny fragment Enrichment by Ligation-mediated PCR*

I

IN: *Input*

IP: Immunoprécipitation

IP (ADN): Immunoprécipité

L

LIMS: *Laboratory Information Management System*

LINE: *Long Interspersed Element*

LTR: *Long Terminal Repeat*

M

MBD: *Methyl-CpG-Binding Domain*

MCA: *Methylated CpG island Amplification*

MDA: *Multiple Displacement Amplification*

MeDIP : *Methylated DNA ImmunoPrecipitation*

MeDIP-chip: MeDIP suivi d'une hybridation sur puce à ADN

MeDIP-dep-Seq: *MeDIP-depletion-Sequencing*

MeDIP-Seq : *MeDIP-Sequencing*

MEF: *Mouse Embryonic Fibroblasts*

MeQA : *MeDIP-Seq data Quality assessment and Analysis*

MethylCap : *Methyl-DNA Capture*

MID: *Multiplex IDentifier*

MiGS: *MBD-isolated Genome Sequencing*

MIRA: *Methylated CpG Island Recovery Assay*

MRE-Seq: *Methylation-sensitive Restriction Enzymes sequencing*

MSCC: *Methylation-Sensitive Cut Counting*

MSDK: *Methylation-Specific Digital Karyotyping*

MMSDK: *Modified MSDK*

N

NaOH: Hydroxyde de sodium (soude)

P

pb: Paires de bases

PCR: *Polymerase Chain Reaction*

PEG : Polyéthylène glycol

PGM : *Personal Genome Machine*

PTP : *PicoTiter Plate*

Q

qPCR : *Quantitative PCR* (PCR en temps reel)

qsp :Quantité suffisante pour

R

RCA : *Rolling Circle Amplification*

rpm: Rotation par minute

RRBS : *Reduced Representation Bisulphite Sequencing*

S

SINE: *Short Interspersed Element*

SMRT : *Single-Molecule Real-Time*

SNP: *Single Nucleotide Polymorphism*

SPRI : *Solid-Phase Reversible Immobilization*

T

TA : Température Ambiante

TAE : Tris Acetate EDTA

TE: Tris EDTA

TNM : Classification « Tumeur / Node / Métastases »

U

U : Unités d'enzyme

W

WBA : *Whole Bisulfitome Amplification*

WGA : *Whole Genome Amplification*

WT : *Wild Type* (sauvage)

Introduction

L'analyse des profils de méthylation de l'ADN présente un grand intérêt pour répondre à de nombreuses questions biologiques et cliniques car ceux-ci sont modifiés au cours des processus pathologiques, notamment dans la cancérogenèse. De nombreuses méthodes tirent désormais profit des évolutions spectaculaires des technologies de séquençage haut-débit pour étudier la méthylation de façon toujours plus précise. Ces notions seront introduites dans le chapitre I.

Le MeDIP-Seq est une technique de choix dans ce domaine. Il est basé sur le MeDIP (*Methylated DNA ImmunoPrecipitation*) qui utilise un anticorps dirigé contre les 5-méthylcytidines pour isoler les régions méthylées sur le génome entier. Les fragments immunoprécipités sont ensuite étudiés sur un séquenceur de seconde génération. Nous avons mis en place le MeDIP-Seq dans notre laboratoire afin d'étudier le méthylome : l'établissement de ce protocole est présenté dans le chapitre III. Les outils dont nous disposons pour ces travaux auront été présentés dans le chapitre II.

Le MeDIP-Seq a cependant pour inconvénient d'immunoprécipiter de nombreuses séquences méthylées localisées dans les régions répétées du génome qui restent difficiles à analyser car elles peuvent rarement être alignées sur un génome de référence sans ambigüité. Nous avons donc établi un protocole innovant baptisé MeDIP-dep-Seq pour nous en affranchir, tout en conservant les séquences uniques d'intérêt : celui-ci est décrit dans le chapitre IV. Nous montrerons dans le chapitre V les gains qualitatifs mais aussi quantitatifs qu'il peut apporter.

Des régions d'intérêt identifiées grâce au MeDIP-dep-Seq peuvent ensuite faire l'objet d'une étude plus ciblée. Nous avons développé dans ce but une nouvelle technique basée sur des sondes de capture appelées sélectors dont l'utilisation sera combinée avec le traitement au bisulphite de sodium. L'introduction de séquences identifiantes pour le multiplexage permettra ensuite de séquencer plusieurs échantillons en parallèle sur un séquenceur de paillasse comme nous l'expliquerons dans le chapitre VI.

Nous discuterons enfin dans le chapitre VII de l'intérêt de notre nouvelle plateforme d'analyse de la méthylation à diverses échelles et de ses atouts en comparaison aux technologies existantes.

Chapitre I: Mise en contexte du projet

Dans ce premier chapitre, nous décrirons tout d'abord ce qu'est la méthylation de l'ADN et expliquerons son rôle biologique. Nous détaillerons dans un second temps l'évolution des méthodes de séquençage d'ADN. Ceci permettra de comprendre comment elles se sont imposées, en combinaison à diverses techniques, pour étudier le méthylome, ce que nous verrons dans une troisième partie, ou pour analyser la méthylation sur des régions plus ciblées, ce que nous préciserons dans la quatrième partie de ce chapitre. Enfin, nous replacerons notre étude dans les contextes cités ci-avant pour comprendre l'intérêt des méthodes que nous avons développées.

Certains termes techniques ne trouvent pas de traduction ni d'équivalent corrects dans la langue française. Ils seront donc employés en langue anglaise tout au long de ce manuscrit, et ce en italique.

I.1 La méthylation de l'ADN

I.1.1 Emergence de l'épigénétique

Toutes les cellules d'un organisme vivant renferment le même ADN et portent donc un patrimoine génétique similaire. Cependant, toutes n'ont pas une morphologie identique et ne contribuent pas aux mêmes fonctions biologiques. Cette hétérogénéité est le fruit d'une expression différente du génome en fonction du type cellulaire. La machinerie qui se met en place pour réguler cette expression fait appel à des mécanismes moléculaires complexes appartenant au domaine de l'épigénétique.

En 1942, Conrad Hal Waddington a été le premier à introduire le terme d'épigénétique (1), dont l'une des définitions actuelles les plus généralement admises est l'étude des changements héréditaires dans la fonction des gènes, ayant lieu sans altération de la séquence d'ADN. La génétique, telle qu'elle a été pratiquée pendant de nombreuses années, se focalisant sur les modifications de la structure primaire de notre ADN comme les mutations, n'a effectivement pas suffi à élucider certains des mystères du vivant. Ce n'est qu'il y a une trentaine d'années en s'intéressant aux mécanismes intervenant « au-dessus » de l'ADN (selon l'étymologie du préfixe « épi ») que les chercheurs ont pu comprendre et expliquer certaines pathologies.

Les phénomènes épigénétiques sont de plusieurs natures. Ils impliquent tout d'abord des modifications des histones, protéines s'associant à l'ADN pour former des structures appelées nucléosomes qui constituent l'unité de base de la chromatine. Celle-ci se trouve dans un état ouvert (euchromatine) ou fermé (hétérochromatine) sous le contrôle de modifications affectant les histones, telles que leur méthylation, leur acétylation ou leur phosphorylation (2). D'autres processus épigénétiques se manifestent par l'intervention du complexe protéique polycomb/trithorax (3,4), d'ARNs non-codants (5,6) ou encore de complexes ATP-dépendants provoquant le remodelage de la chromatine (7). Enfin, l'une des modifications épigénétiques étudiées avec un intérêt croissant est la méthylation de l'ADN; elle fera l'objet des travaux présentés tout au long de ce manuscrit. Tous ces mécanismes sont réversibles et les modifications qui en résultent sont transmises entre générations de cellules somatiques après leur division.

I.1.2 Méthylation des cytosines

I.1.2.1 Mécanismes de méthylation

Chez les mammifères, la méthylation affecte la position 5 des cytosines étant directement suivies d'une guanine. Ces dinucléotides CpGs sont statistiquement sous-représentés dans le génome, dû à

leur haut potentiel de mutation en TpG par désamination (8,9). Les 5-méthylcytosines ne représentent ainsi que 1% de toutes les bases composant le génome des mammifères mais occupent la majorité des CpGs (70 à 80%) (10). Certains CpGs sont en revanche une exception car très peu méthylés : ils se situent dans des régions appelées îlots CpG, longues d'environ 0,5 à 4 kb et ayant une teneur en GC très riche puisque de plus de 50% (11). Dans les cellules souches embryonnaires, environ 25% des 5-méthylcytosines peuvent se trouver dans des trinucléotides CpHpG ou CpHpH (où H est A, C ou T) (12,13). Chez les plantes, les cytosines présentes dans le contexte CpNpG ou CpA peuvent également porter cette modification. L'ensemble des bases méthylées du génome constitue le méthylome.

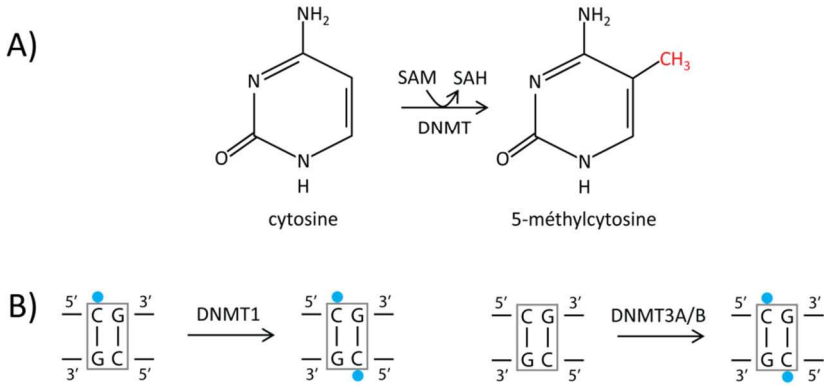


Figure 1: Mécanisme de méthylation de l'ADN
A) Méthylation d'une cytosine par une DNMT (ADN méthyltransférase). SAM : S-adénosyl-L-méthionine. SAH : S-adénosylhomocystéine. B) Mode de fonctionnement simplifié des différentes DNMTs. Points bleus : groupements méthyles.

L'addition d'un groupement méthyle sur une cytosine de façon covalente est catalysée par des ADN méthyltransférases (DNMTs) (14). Ceci est réalisé en présence d'un donneur de groupements méthyles (méthionine et choline) et de co-facteurs (acide folique ou vitamine B12) qui permettent dans un premier temps de synthétiser un donneur universel de groupements méthyles : la S-adénosyl-L-méthionine (SAM). La réaction de méthylation des cytosines est alors réalisée par transfert du groupement méthyle sur la base de l'ADN et libère la S-adénosylhomocystéine (SAH) (voir Figure 1A). A ce jour, 4 ADN méthyltransférases ont pu être identifiées : les DNMT1, DNMT2, DNMT3A et DNMT3B. La DNMT1 est une protéine de maintien de l'information épigénétique à travers les générations cellulaires : elle intervient durant la phase S du cycle cellulaire, en catalysant la méthylation du brin nouvellement synthétisé à partir de l'ADN hémi-méthylé. Les méthyltransférases DNMT3A et B réalisent une méthylation *de novo* en permettant l'ajout de groupements méthyles sur les deux brins de l'ADN non-méthylé (voir Figure 1B). La DNMT2, elle, ne possède qu'une faible activité catalytique *in vitro* (14).

Une autre base a en réalité été découverte comme méthylée dans le génome des mammifères il y a 40 ans : il s'agit de la 5-hydroxyméthylcytosine (15). Cette base est cependant présente en très faible

quantité à travers le génome (16). Elle n'est étudiée avec intérêt que depuis peu (17) et il a pu être montré que des enzymes de la famille TET (Ten-Eleven Translocation) catalysent l'oxydation de 5-méthylcytosines vers cette forme alcool de la base (18). Elle constitue alors un intermédiaire dans le processus de déméthylation de l'ADN (19,20). Jusqu'à présent, la plupart des techniques utilisées pour étudier la méthylation ont échoué à la différencier d'une 5-méthylcytosine (21).

I.1.2.2 Régulation de l'expression des gènes

Plusieurs modifications chimiques interviennent sur la chromatine et contribuent à son organisation dynamique. De façon très simplifiée, dans l'euchromatine, la lysine 9 de l'histone H3 (H3K9) est hyperacétylée et sa lysine 4 (H3K4) est méthylée. Ceci permet de conserver la chromatine en conformation ouverte et l'ARN polymérase II peut accéder au promoteur du gène pour débuter la transcription (voir Figure 2A). Dans l'hétérochromatine, H3K4 est déméthylée tandis que H3K9, H3K27 et H4K20 sont méthylées (2). Ceci provoque la compaction de la chromatine et rend la transcription impossible.

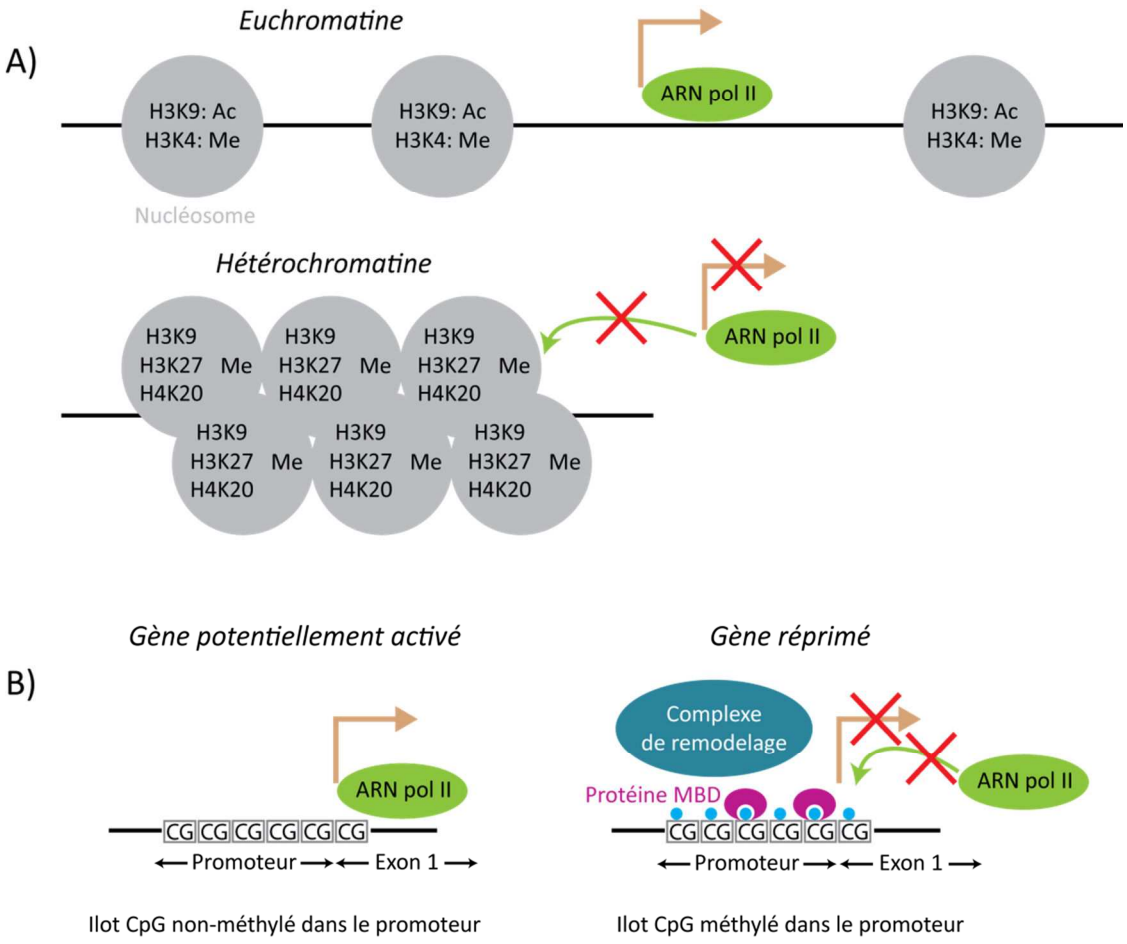


Figure 2: Régulation épigénétique de la transcription
A) Différents états de la chromatine. H : histone, K : lysine, Ac : acétylé, Me : méthylé, ARN pol II : ARN polymérase II. Flèche marron : début de transcription B) Impact de la méthylation de l'ADN sur la transcription. Différents états de méthylation des îlots CpG présents dans le promoteur du gène concerné sont présentés. Point bleu : groupement méthyle.

La méthylation de l'ADN intervient dans la régulation de l'expression des gènes de façon conjointe à ces mécanismes. Chez l'humain, il existe environ 45000 îlots CpG occupant 1 à 2% du génome (11,22), et ceux-ci sont présents dans au moins un des promoteurs de 70% des gènes humains (23). Leur état non-méthylé engendre le maintien d'une structure ouverte de la chromatine permettant ainsi une potentielle transcription (24) (voir Figure 2B). Une minorité de promoteurs peut en revanche être associée à des îlots CpG méthylés et est ainsi reconnue par des protéines MBD (Methyl-Binding Domain) possédant des domaines de liaison à l'ADN méthylé qui empêchent alors la potentielle liaison de facteurs activateurs de transcription et la reconnaissance des séquences consensus. Les protéines MBD recrutent également d'autres protéines possédant des fonctions répressives telles que les HDACs (histones désacétylases) conduisant à la désacétylation des histones et donc à la compaction de la chromatine, ou des complexes de remodelage de la chromatine qui engendrent une configuration inadaptée à la transcription (25).

Par ailleurs, la méthylation de l'ADN contribue à la stabilité du génome en affectant lourdement les éléments répétés qui le composent en grande majorité et qui ont longtemps été considérés seulement comme de l'ADN parasite (26). Il a été observé que 38% des cytosines méthylées étaient localisées dans les éléments répétés et les régions centromériques et sous-télomériques (27). La méthylation des séquences répétées empêche leur transcription ainsi que leur transposition tandis que leur hypométhylation favorise les réarrangements chromosomiques influençant ainsi l'intégrité du génome et pouvant mener à des pathologies diverses, notamment des cancers (28).

Cette régulation est contrôlée de façon spécifique à un tissu au cours du développement et est ensuite maintenue tout au long de la vie d'un individu.

I.1.2.3 Implication au cours du développement

Les profils de méthylation subissent des changements drastiques durant les phases précoces de développement (29). Lors de la fécondation, le génome se déméthyle, probablement pour initier la différenciation cellulaire. Lors de l'implantation de l'embryon, les niveaux de méthylation sont ensuite restaurés par des mécanismes *de novo* ; un dysfonctionnement des DNMTs est alors létal pour l'embryon. Une second phénomène de reprogrammation a lieu durant l'embryogenèse où la méthylation des cellules germinales primordiales disparait sur les gènes (30). Elle sera rétablie sur l'un ou l'autre des loci maternel ou paternel en fonction du sexe, à la naissance chez le mâle ou bien plus tard dans les ovocytes matures chez la femelle.

Ce dernier mécanisme joue également un rôle dans le phénomène d'empreinte parentale (31). Chez les mammifères, les génomes d'origine maternelle et paternelle ne sont pas équivalents d'un point de vue fonctionnel mais tous deux sont nécessaires au développement. Il existe des gènes pour

lesquels seule une copie s'exprime tandis que l'autre reste silencieuse sous régulation de la méthylation. L'absence de la copie d'intérêt conduira alors à de nombreuses maladies, la plus représentative étant le syndrome de Prader-Willi qui se traduit par une hypotonie infantile sévère, une obésité morbide et des troubles du comportement et de l'apprentissage (32). Le nombre de gènes soumis à l'empreinte était estimé à une centaine jusqu'à la découverte en 2010 de 1300 loci concernés par ce phénomène chez la souris (33).

Chez les mammifères femelles, l'un des deux chromosomes X est inactivé dans son intégralité. Ceci permet de contrebalancer le déséquilibre existant avec le génome des mâles au niveau des allosomes, leur chromosome Y ne comportant que très peu de gènes, afin que tous deux ne disposent que d'un seul chromosome X actif. Une autre forme de compensation existe chez d'autres espèces comme *Drosophila Melanogaster* ; elle réside dans l'expression deux fois plus importante de l'X chez le mâle afin de rétablir la balance avec les autosomes (34,35). Le choix du chromosome à inactiver relève du hasard et l'embryon est alors composé de deux types de cellules possédant un chromosome X actif différent (mosaïcisme). Chez le chat calicot, la couleur du pelage est notamment gouvernée par un gène présent sur le chromosome X dont un allèle code pour la couleur noire, l'autre pour la couleur orange. Les mâles, ne possédant qu'un seul X et par là même qu'un seul allèle, seront uniformément colorés tandis que les femelles, à condition d'être hétérozygotes, auront un pelage bicolore dû à l'expression mosaïque. Le mécanisme d'inactivation de l'X débute par la synthèse d'un ARN non-codant Xist (X inhibitory specific transcript) qui va recouvrir le futur chromosome inactif. Ce n'est que bien plus tard que les profils de méthylation sont établis sur ce chromosome pour le conserver silencieux (36).

I.1.3 Epigénétique et santé

Récemment, les études d'association pangénomiques (GWAS : Genome-Wide Association Studies) ont permis d'identifier un grand nombre de SNPs (Single Nucleotide Polymorphisms) associés à des maladies humaines. On assiste désormais à la mise en place d'études à grande échelle permettant d'impliquer les variations épigénétiques dans la compréhension de diverses pathologies (EWAS : Epigenome-Wide Association Studies) (37). Nous présentons ici certaines de ces maladies.

I.1.3.1 Pathologies liées à des altérations épigénétiques

Bien que représentatives du phénomène naturel de vieillissement, l'altération de la méthylation et son évolution vers des profils aberrants peuvent mener à certaines pathologies (38). Peu d'entre elles ont cependant été étudiées. Des modifications de profils épigénétiques ont pu être observées dans des maladies inflammatoires chroniques telles que l'athérosclérose ou l'arthrite rhumatoïde

(39,40). Elles sont également responsables de la dérégulation de la différenciation des lymphocytes T, à l'origine de la pathologie asthmatique et des phénomènes allergiques (41). On peut également citer leur implication dans le diabète de type 2 (42), ou dans des maladies psychiatriques comme l'autisme, voire la schizophrénie (43). Le cancer présente également une forte composante épigénétique et fait l'objet de la majorité des études dans ce domaine.

I.1.3.2 Mécanismes impliqués dans le cancer

I.1.3.2.1 Vers un nouveau biomarqueur

Dans la cancérogenèse, deux évènements épigénétiques majeurs et indépendants dominent : l'hypométhylation globale et l'hyperméthylation ciblée (44-48). Les premières mises en évidence de perte de méthylation et d'hypométhylation du génome dans des cellules cancéreuses humaines remontent à 1983 (49,50). Celles-ci contribuent à une instabilité génomique et ponctuellement à l'activation transcriptionnelle d'oncogènes. Elles peuvent également mener à une perte de l'empreinte parentale. C'est le cas du locus *IGF2*/*H19* pour lequel une perte d'empreinte a été associée à une augmentation de l'incidence de cancer colorectaux. Ce même locus est également impliqué dans des anomalies congénitales : son expression biallélique anormale au cours du développement embryonnaire peut entrainer la mise en place d'une tumeur de Wilms dans le tissu à l'origine du rein. Cependant, un doute subsiste encore quant au fait que l'hypométhylation soit une cause des cancers ou l'une de leurs conséquences en tant qu'effet secondaire du développement tumoral (51).

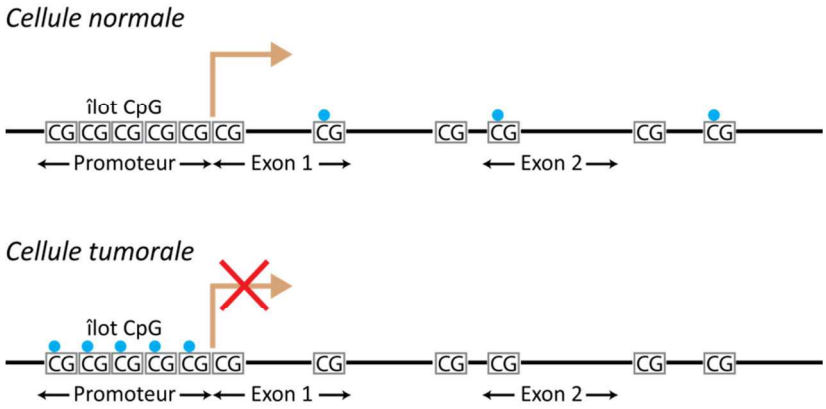


Figure 3: Régulation de l'expression dans des cellules normales et tumorales
Dans les cellules tumorales, l'hyperméthylation de l'îlot CpG situé dans le promoteur entraîne l'inactivation du gène. La méthylation globale est également perdue.

Parallèlement à ce phénomène, une hyperméthylation ciblée se produit au niveau des îlots CpG, notamment dans les régions promotrices de gènes suppresseurs de tumeurs, provoquant ainsi leur inactivation (voir Figure 3). Ceux-ci peuvent être impliqués dans les voies de régulation du cycle

cellulaire (*p16^{INK4a}* par exemple), de réparation de l'ADN (*MGMT*, *MLH1*, *DAPK*) ou de l'apoptose (*RASSF1A*, *CASP8*) (52). Selon le type tumoral, certains de ces gènes se trouveront plus ou moins méthylés : *p14* et *APC* le sont par exemple davantage dans les tumeurs gastro-intestinales tandis que l'hyperméthylation de *GSTP1* est caractéristique des tumeurs du foie ou de la prostate, celle de *BRCA1* des cancers féminins et celle de *MGMT* et *p16* associée aux tumeurs du poumon, de la tête et du cou. La liste de ces gènes pouvant être considérés comme de nouveaux biomarqueurs des cancers n'est pas exhaustive et n'a cessé de s'allonger grâce à des initiatives dans le domaine de l'épigénétique telles que le Human Epigenome Project (53). D'autres exemples figurent dans le Tableau 1 (54-58). Il est donc désormais concevable d'étudier plusieurs types de tumeurs en cherchant à établir des profils de méthylation qui soient propres à chacun (59,60).

Type de cancer	Gènes biomarqueurs	Fluide/tissu utilisés
Ovaire	*BRCA1, RASSF1A*	Sérum
Sein	*RASSF1A*	Liquide mamelonnaire
	APC, DAPK	Sérum
Œsophage	*APC*	Plasma
Poumon	*p16^{INK4a}, MGMT, RASSF1A*	Salive
	GSTP1, DAPK	Sérum
Colorectal	*p16^{INK4a}, SFRP2, MGMT, MLH1*	Selles
	SEPT9, TMEEF2, NGFR	Sérum
Prostate	*RASSF1A, RARβ2, APC*	Urine
	GSTP1	Sperme
	TIG1	Biopsie
Vessie	*RASSF1A, APC, DAPK, BCL2, TERT*	Urine
	p16^{INK4a}	Plasma
Tête et cou	*p16^{INK4a}, DAPK1, MGMT*	Sérum
	FANCF, SOCS1, SOCS3	Biopsie
	DCC, TIMP3, ESR	Salive
Leucémie	*p15^{INK4b}*	Sang

Tableau 1: Potentiels biomarqueurs épigénétiques de divers cancers

L'analyse de la méthylation peut s'avérer être un outil très utile en oncologie clinique (61). En effet, les anomalies présentes dans un tissu tumoral peuvent également être détectées dans les fluides biologiques tels que le sang, la salive, le sperme, les selles ou les urines (voir Tableau 1). Elles vont donc amener à la mise au point de nouveaux tests non-invasifs pour diagnostiquer les cancers à un stade toujours plus précoce, pronostiquer leur évolution mais également suivre leur malignité puisque les taux de méthylation coïncident souvent avec le stade de la maladie. Leur utilisation permettra même de prendre en charge les patients et d'orienter certains traitements dans le cadre de la médecine personnalisée puisque le statut de méthylation de gènes suppresseurs de tumeurs, *MGMT* par exemple dans le cas de gliomes, a pu être associé à une réponse positive à la

chimiothérapie avec, dans le cas cité, le témozolomide, agent alkylant pénétrant la barrière hématoencéphalique (62).

I.1.3.2.2 Thérapie épigénétique

A la différence des altérations génétiques, les évènements épigénétiques sont réversibles, ce qui en fait des outils attractifs pour développer de nouvelles approches thérapeutiques. Des agents capables de restaurer des profils épigénétiques normaux ont peu à peu été mis au point. Il s'agit d'inhibiteurs de HDACs et de DNMTs, dont les plus courants sont présentés dans le Tableau 2 (63,64).

A)

Inhibiteurs de HDAC	Autre dénomination
Acides gras à courte chaîne carbonée	
Butyrate	
Acide valproïque	Depakote®
Acides hydroxamiques	
Acide bishydroxamide *m*-carboxycinnamique	CBHA
Oxamflatin	
Scriptaïd	
Acide suberoïlanilide hydroxamique	SAHA, vorinostat, Solinza®
Trichostatin A	TSA
Tétrapeptides cycliques	
Apicidin	
Romidepsin	Depsipeptide
Trapoxin A	
Benzamides	
N-acétyl-dinaline	CI-994
MS-275	

B)

Inhibiteurs de DNMT	Autre dénomination
Analogues nucléotidiques	
5-azacytidine	Vidaza®
5-aza-2'-déoxycytidine	Decitabin, Dacogen®
5-fluoro-2'-déoxycytidine	Gemcitabin
5,6-dihydro-5-azacytidine	DHAC
Arabinosyl-5-azacytidine	Fazarabine
Zébularine	
Inhibiteurs non-analogues	
Hydralazine	Apresoline
Procaïnamide	Procaïne, novocaïne
Epigallocatichin-3-gallate	ECGC
Psammaplin A	
MG98	DNMT1 anti-sens

Tableau 2: Inhibiteurs pouvant être utilisés en thérapie épigénétique

Les inhibiteurs de HDACs sont sous-divisés en 4 classes en fonction des groupes fonctionnels qu'ils portent. Leur action provoque l'accumulation de groupements acétyles dans les histones et est suivie de modifications des processus cellulaires critiques pour les cellules cancéreuses, notamment grâce à leur pouvoir anti-angiogénique. Les inhibiteurs de DNMTs peuvent être des analogues de la cytosine qui, utilisés à faible dose, ont une action de déméthylation en se liant de manière covalente aux DNMTs. Ils permettent ainsi la réactivation de gènes suppresseurs de tumeurs. Les plus connus sont la 5-azacytidine et la 5-aza-2'-déoxycytidine, approuvés par la FDA (Food and Drug Administration) pour le traitement de la leucémie myéloïde chronique.

Ces thérapies épigénétiques, utilisées en combinaison avec une chimiothérapie conventionnelle, permettront de traiter efficacement certains cas, comme l'utilisation conjointe de vorinostat (inhibiteur de HDAC) et de doxorubicine, paclitaxel ou bevacizumab dans le cas du cancer du sein (56). L'épigénétique ouvre donc désormais de nouveaux horizons à la médecine contemporaine.

I.1.4 Influence environnementale

Des études menées sur des jumeaux monozygotes soumis à des environnements et ayant des modes de vie différents montrent que leurs profils épigénétiques n'évoluent pas de la même manière (65). Ceci offre enfin la possibilité d'expliquer en partie comment l'environnement influence le phénotype.

L'exposition à des polluants chimiques comme certains perturbateurs endocriniens, les changements de températures ou encore des sources de stress diverses sont autant de paramètres environnementaux pouvant provoquer des modifications épigénétiques et induire des effets à long-terme sur le développement, le métabolisme ou la santé d'un individu (66,67). D'autres facteurs, bien qu'insignifiants à première vue, peuvent également influencer les mécanismes épigénétiques, tels que les habitudes de travail (de nuit par exemple), la pratique d'une activité physique ou le mode de vie de façon générale. De plus, il convient de souligner que la consommation d'alcool ou de tabac nuit aussi gravement à notre épigénome (68). Il a ainsi pu être montré que le tabagisme pendant la grossesse modifie le statut de méthylation et dérégule l'expression de plus de 600 gènes dans le placenta, un petit nombre d'entre eux provoquant alors une restriction de la croissance fœtale (69). Les fumeurs passifs seront également concernés : une exposition à la fumée de cigarette a pu être corrélée à une augmentation significative de la méthylation au niveau de gènes impliqués dans la carcinogenèse, notamment dans le cancer de la vessie (70).

Le régime alimentaire est également un facteur de modifications phénotypiques notable. Chez les abeilles, la reine et ses ouvrières issues de larves génétiquement identiques n'ont pas reçu la même alimentation. La reine, nourrie exclusivement à la gelée royale, développera des caractéristiques physiques particulières qui trouvent notamment leur origine dans des différences de méthylation significatives ayant pu être observées sur plus de 500 gènes (71). Chez la souris en gestation, une supplémentation en acide folique, riche en méthionine, mène à une augmentation de la méthylation de l'ADN d'un allèle du locus Agouti. Ceci provoque la répression du gène et une modification du phénotype (ici, la couleur du pelage) de sa progéniture. Le même effet peut être observé lorsque les mères sont exposées à du Bisphénol A (72,73). Par ailleurs, une hypothèse qui émerge d'études épidémiologiques propose que des adaptations métaboliques dues à des conditions nutritionnelles défavorables pendant la vie fœtale puissent affecter la croissance et le développement, notamment par l'altération du statut de méthylation de certains gènes. Pour exemple, pendant la seconde guerre mondiale, les faibles rations alimentaires imposées par l'occupation allemande à la population hollandaise ont amené des femmes, alors à un stade avancé de leur grossesse, à donner naissance à des enfants de très faible poids, et le phénomène s'est reproduit sur leur descendance. Une telle diète pendant la période périconceptionnelle et les étapes précoces du développement de l'embryon semble davantage influencer le risque de troubles chroniques dans la vie future de l'individu puisque

les enfants dont les mères y ont été exposées à un stade initial de leur grossesse avaient un poids normal mais ont développé plus tard des signes d'obésité (74).

Malgré tout, si l'alimentation peut avoir des effets néfastes sur la santé, elle s'avèrerait également être un moyen de prévention efficace des cancers ou de certaines maladies inflammatoires. Les polyphénols (présents dans le thé vert, le curry, le raisin, …) pourraient ainsi agir comme inhibiteurs de DNMTs et permettraient de restaurer l'expression de gènes suppresseurs de tumeurs (75,76). Le bon usage de notre alimentation constituerait donc l'une des rares clés auxquelles nous avons facilement accès pour orienter l'évolution de notre épigénome.

I.2 Séquençage : la course au génome à 1000 $

Le séquençage consiste à déterminer l'enchainement linéaire des nucléotides d'un fragment d'ADN ou, de façon plus générale, d'un génome. Son histoire débute en 1977 lorsque Maxam et Gilbert développent une technique basée sur le marquage radioactif de fragments et leur coupure sélective par dégradation chimique (77). En parallèle, Sanger aborde une tout autre approche.

I.2.1 Le séquençage Sanger et son automatisation

La même année, Sanger base sa technique de séquençage sur une synthèse enzymatique des fragments d'ADN après leur amplification par clonage. Il utilise la propriété qu'ont les ADN polymérases de synthétiser un brin complémentaire d'un brin matrice en présence de dNTPs (désoxyribonucléotides triphosphate). Il ajoute à ce milieu des ddNTPs (didésoxyribonucléotides triphosphate, voir Figure 4) marqués par fluorescence (le marquage étant radioactif dans la méthode d'origine) ; ceux-ci servent de terminateurs d'élongation de façon aléatoire dans la réaction. En faisant migrer sur gel de polyacrylamide les fragments de différentes tailles alors obtenus, il a pu en lire la séquence (78). Le prix Nobel de chimie récompensera Sanger ainsi que Gilbert en 1980 pour leurs découvertes respectives.

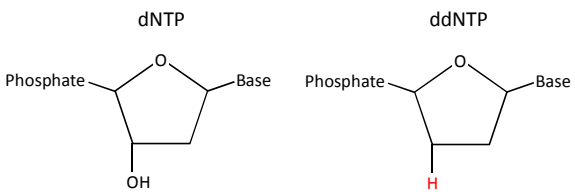


Figure 4: dNTP comparé à un ddNTP
L'absence de groupement hydroxyle sur le ddNTP empêche la poursuite de la synthèse du brin d'ADN.

Depuis, le séquençage Sanger a été automatisé : les premières machines utilisant des gels de polyacrylamide ont rapidement été supplantées par les séquenceurs capillaires. Cette technologie,

principalement représentée par le 3730 DNA Analyzer de Applied Biosystems (maintenant Life Technologies) depuis sa commercialisation en 2002, est maintenant considérée comme la première génération du séquençage haut-débit et a permis de déchiffrer le génome humain. En effet, au début des années 90, la communauté scientifique internationale se fixe pour objectif de séquencer les 3,5 milliards de bases contenues dans le génome humain via le *International Human Genome Sequencing Consortium*. Débute alors une course effrénée vers l'achèvement du séquençage du génome humain. Le désormais célèbre J. Craig Venter crée en 1998 sa société Celera Genomics et a pour ambition d'y parvenir bien avant les laboratoires publics : le privé menace de s'approprier le génome humain. En 2000 et avec trois ans d'avance, les deux acteurs annoncent simultanément avoir atteint leur but, ce qui aboutira à deux publications en 2001, l'une par l'équipe de Craig Venter (79), l'autre par le Consortium (80). On y apprendra notamment et contre toute attente que le génome de notre espèce compte seulement 30000 gènes, et que ceux-ci n'en occupent que 5%. Les génomes complets seront présentés quelques années plus tard, en 2004 par le Consortium (81) pour un coût total de 2,7 milliards de $ puis en 2007 par le J. Craig Venter Institute (82,83). Ce dernier n'est autre que celui de Craig Venter lui-même, séquencé à une profondeur (occurrence moyenne d'un nucléotide) de 7,5x pour un coût total de 70 millions de $.

I.2.2 Le séquençage de 2nde génération : l'explosion de nouvelles technologies

Depuis 2005, l'émergence de nouvelles technologies considérées comme la nouvelle ou seconde génération de séquençage (NGS, next-generation sequencing) a permis de séquencer avec des débits qui évoluent encore aujourd'hui de façon phénoménale. On a pour habitude de comparer cette évolution à l'augmentation exponentielle énoncée par la loi de Moore en 1965 selon laquelle la puissance des processeurs allait doubler tous les ans à prix constant : les technologies de séquençage ont évolué beaucoup plus vite (voir Figure 5).

Après celui de Craig Venter en 2007, plusieurs génomes personnels ont été publiés comme preuve-de-principe de ces nouvelles technologies. En 2008 c'est celui de James D. Watson, biochimiste célèbre pour sa découverte de la double hélice d'ADN (84), qui est généré sur un séquenceur FLX de Roche (85). En seulement deux mois et pour moins d'1 million de $, 24,5 millions de bases sont générées à une profondeur de 7,4x et elles couvrent 95% du génome. En 2009, un des co-fondateurs de la société Helicos Biosciences, Stephen R. Quake, séquence son génome (86) avec une profondeur de 28x et une couverture du génome de 90% pour un coût de 48000 $. En 2009, 4 autres génomes humains ont été décrits : ceux d'un homme yoruba du Nigeria (87) séquencé à une profondeur de 30x, de 2 coréens (88,89) à une profondeur de 28 et 29x et d'un chinois Han (90) à une profondeur

de 36x. Le séquençage du génome a également commencé à servir les intérêts de la médecine et c'est ainsi que la même année ont été publiés les génomes de deux patients atteints de leucémie (91,92) à une profondeur de 33 et 23x. Fin 2010, le nombre de génomes séquencés approche les 3000 et fin 2011 il faut désormais le multiplier par 10 (93).

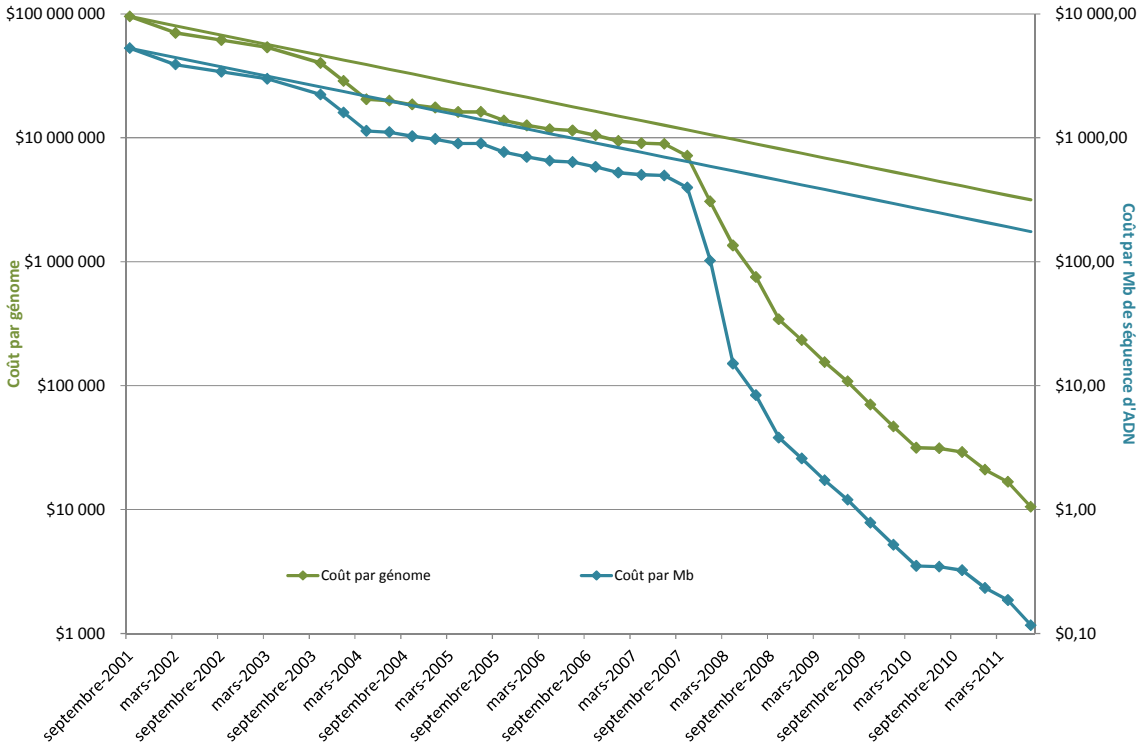


Figure 5: Evolution des coûts du séquençage
En vert : le coût du séquençage d'un génome de la taille du génome humain ; en bleu : le coût du séquençage d'1 million de bases. Les profondeurs utilisées sont les suivantes : 6x pour du séquençage Sanger, 10x pour le 454, 10x pour Illumina et le SOLiD. Les droites représentent l'évolution théorique selon la loi de Moore pour comparaison. Le « décrochage » à partir du 1er trimestre 2008 coïncide avec la période à laquelle les centres de séquençage se sont dotés de séquenceurs de 2nde génération. Ces coûts englobent la main-d'œuvre, les réactifs et consommables, les instruments et leur amortissement sur 3 ans, un service informatique minimum, la construction des librairies et la soumission des données à des bases de données publiques. Ils ne prennent pas en compte le développement technologique, le contrôle-qualité, le matériel et le développement des outils informatiques ni l'analyse complète des données. Chiffres d'après Wetterstrand KA. DNA Sequencing Costs: Données du NHGRI Large-Scale Genome Sequencing Program. Disponible sur: www.genome.gov/sequencingcosts/. Accès le 08/11/11.

Cette augmentation est corrélée avec une diminution drastique des coûts : on approche de l'objectif du génome à 1000 $ (94) fixé quelques années auparavant et les plus optimistes prévoient que ce seuil sera atteint d'ici deux ans même si certains s'accordent à dire que ce coût supportera seulement les réactifs et n'englobera ni l'amortissement des instruments, ni la main-d'œuvre ou encore le stockage et le management des données (95).

Quatre plateformes de séquençage de seconde génération sont actuellement disponibles sur le marché, proposant différentes versions de machines (voir Tableau 3). Elles utilisent des chimies qui leur sont propres ; le choix d'investir dans l'une ou l'autre de ces technologies reposera en partie sur l'application finale et devra prendre en compte les caractéristiques intrinsèques à chacune (96).

Tableau 3: Caractéristiques des technologies de séquençage
Chiffres d'après Glenn 2011 (96). Ces technologies de séquençage ont été, sont ou seront commercialisées sur le marché.

	1ère génération	2e génération													3e génération		
Plateforme	3730	454			Illumina					SOLiD			Ion Torrent			HeliScope	PacBio	Starlight
Entreprise actuelle	Life Technologies	Roche			Illumina					Life Technologies			Life Technologies			Helicos	Pacific Biosciences	Life Technologies
Entreprise d'origine	Applied Biosystems	454			Solexa					Applied Biosystems			Ion Torrent					
Commercialisé depuis	2002	2005			2006					2007			2010			Ne l'est plus	2011	Pas encore
Renommée	Représentatif de la 1ère génération	1er de la nouvelle génération, longs *reads*			1er séquenceur à courts *reads*, leader actuel					2e séquenceur à courts *reads*, faible taux d'erreurs			1er séquenceur sans système de détection optique, 1er à moins de 100.000$			1er séquenceur d'une molécule unique	1er séquenceur en temps réel d'une molécule unique	Séquençage d'une molécule unique avec quantum dots
Méthode de séquençage	Sanger	Synthèse (pyroséquençage)			Synthèse					Ligation			Synthèse (détection de H^+)			Synthèse	Synthèse	Synthèse
Méthode d'amplification	PCR	PCR en émulsion			PCR en ponts					PCR en émulsion			PCR en émulsion			Aucune	Aucune	Aucune
Instrument	3730xl	GS Junior Titanium	FLX Titanium	FLX+	MiSeq	GA IIx	HiSeq 1000	HiSeq 2000	HiSeq 2000 v3	SOLiD 4	SOLiD 5500	SOLiD 5500xl	Ion PGM, puce 314	Ion PGM, puce 316	Ion PGM, puce 318	Helicos[3]	PacBio RS	Starlight
Coût à l'achat ($ US)	376 000	108 000	500 000	500 000	125 000	250 000	560 000	690 000		475 000	349 000	595 000	49 500			nd	695 000	nd
Temps de *run*[1]	2h	10h	10h	18-20h	26h	14 jours	8 jours	8 jours	10 jours	12 jours	8 jours	8 jours	2h			nd	0,5-2h	nd
Millions de *reads/run*	0,000096	0,10	1	1	3,4	320	500	1000	≤ 3000	> 840	> 700	> 1410	0,10	1	4-8	800	0,01	0,01
Nb bases/*read*	650	400	400	700	150+150	150+150	100+100	100+100	100+100	50+35	75+35	75+35	100	>100	>100	35	860-1100	> 1000
Rendement Mb/*run*	0,06	50	500	900	1 020	96 000	100 000	200 000	≤ 600 000	71 400	77 000	155 100	>10	>100	>1000	28 000	5-10	nd
Coût des réactifs/*run*[2]	96	1 100	6 200	6 200	750	11 524	10 220	20 120	23 470	8 128	6 101	10 503	500	750	925	nd	110-900	nd
Coût des réactifs/Mb	1500	22	12,4	7	0,74	0,12	0,1	0,1	<0,04	<0,11	<0,08	<0,07	<50	7,5	0,93	nd	11-180	nd
Erreurs principales	Substitutions	Insertions, délétions			Substitutions					Biais A-T			Insertions, délétions			nd	Délétions CG	nd
Taux d'erreur (%)[4]	0,1-1	1			>0,1					>0,06	>0,01	>0,01	~1			nd	16	nd
Applications principales		a*, b, c*, d, g, h*			a*, b, c*, d, e, f, g, h					c*, e, f, h			a, b, c, d, h			e, h	a, b, c, g, h	a, b, g, h

1 : pour la longueur de lecture maximale
2 : de la préparation de la librairie au séquençage. Pour le séquençage capillaire, uniquement le séquençage
3 : instruments et réactifs ne sont désormais plus commercialisés
4 : pourcentage d'erreur par base sur les reads de longueur maximale (voir ligne nombre de bases/read)

nd: non disponible

Tous les coûts sont donnés en $ US

a : séquençage *de novo* de génomes microbiens
b : caractérisation du transcriptome
c : reséquençage ciblé
d : réquençage de novo de génomes animaux et végétaux
e : reséquençage
f : détection de mutation
g : métagénomique
h : autres applications (ChIP-Seq, MeDIP-Seq, RNA-Seq,...)
* : indexage nécessaire pour une utilisation optimale de cette application

Point fort en vert
Point faible en rouge

I.2.2.1 Les 3 grands du séquençage

Les 3 technologies de séquençage qui dominent actuellement le marché se démarquent les unes des autres par les chimies qui les constituent. Elles se décomposent cependant toutes en 4 grandes étapes principales : la préparation des librairies qui contient une étape d'amplification par PCR, les cycles de réactions de séquençage, la prise d'image après chacun de ces cycles pour déterminer le nucléotide correspondant, puis l'analyse des données. Ces nouvelles générations de machines ont pour avantage leur capacité à analyser de grands génomes à haute résolution grâce à la parallélisation des réactions.

I.2.2.1.1 454/Roche

Le premier des séquenceurs de nouvelle génération a été commercialisé par 454 Life Sciences en 2005, depuis racheté par Roche, mais la plateforme est toujours connue sous le nom générique de 454. Jonathan M. Rothberg en a élaboré la technologie et a démontré sa robustesse avec le séquençage du génome de *Mycoplasma génitalium* (97).

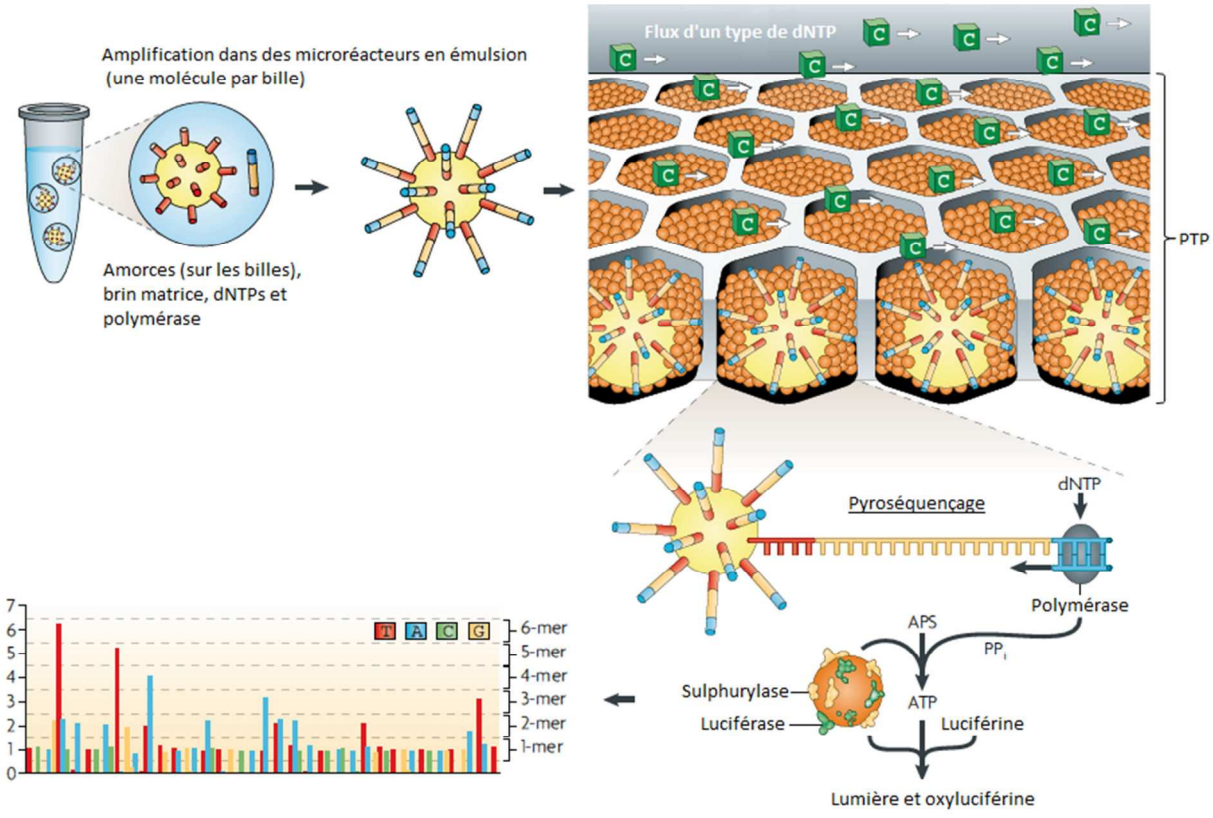


Figure 6: Aperçu de la technologie de séquençage 454
PTP : PicoTiter Plate, PPi : pyrophosphate inorganique, APS : adénosine phosphosulphate, ATP : adénosine triphosphate. Modifié d'après Metzker 2010 (98).

La spécificité de cette technologie repose sur une PCR en émulsion pour l'amplification des fragments à séquencer : la PCR a lieu dans une microgoutte renfermant une microbille d'agarose en

phase aqueuse, séparée des autres billes (plusieurs millions) par de l'huile (voir Figure 6). On obtient ainsi des copies d'un seul fragment d'ADN par bille. Chacune des billes est ensuite déposée dans un des 1,6 millions de puits d'un support solide appelé PTP (PicoTiter Plate).

Des réactions de pyroséquençage ont alors lieu à l'échelle du picolitre dans chaque puits : un flux de nucléotides (chaque nucléotide l'un après l'autre) traverse la PTP et lorsque l'un d'entre eux est incorporé par la polymérase, un pyrophosphate (PPi) est libéré. Deux enzymes, ici contenues dans un autre type de billes beaucoup plus petites, permettent alors une cascade enzymatique : en présence du PPi libéré, une sulphurylase convertit de l'adénosine phosphosulphate (APS) en adénosine triphosphate (ATP). L'ATP permet ensuite la conversion de luciférine en oxyluciférine par une luciférase. Cette dernière réaction produit de la lumière et l'intensité de la réponse lumineuse est directement proportionnelle à la quantité de nucléotides incorporés.

On obtient jusqu'à 900 Mb de données en une dizaine d'heures, soit 15000 fois plus qu'avec la première génération de séquenceurs. Cette technologie présente la plus grande taille de lecture (jusqu'à 700 pb par rapport à 100 pb à ses débuts) et sa grande précision la rend appropriée au séquençage *de novo*. Cependant, l'un de ses défauts, intrinsèque au pyroséquençage, reste la quantification des homopolymères.

Rothberg annonça en 2005 que cette technologie permettrait de séquencer le génome de James D. Watson pour seulement 1 million de $. Le projet s'acheva en 2007 en respectant ce budget et Rothberg et ses collaborateurs publièrent l'un des premiers génomes humains complets (85) après avoir publié celui de Néanderthal en 2006 (99,100) pour lequel on montrera cependant avoir majoritairement séquencé de l'ADN humain moderne contaminant (101).

I.2.2.1.2 Solexa/Illumina

L'entreprise Solexa a commercialisé son premier séquenceur avec succès en 2006 : le Genome Analyzer (GA). La spécificité de cette technologie repose sur une amplification en pont (*bridge PCR)* des fragments à séquencer. Elle a lieu sur une surface de verre appelée *flow cell* (FC), similaire à une lame de microscope, divisée en huit lignes (à l'origine, une ligne par échantillon). Les fragments de la librairie à séquencer possèdent des adaptateurs à leurs extrémités. Ceux-ci vont leur permettre de se fixer de façon aléatoire sur la FC, par hybridation sur les amorces qui en couvrent la surface (voir Figure 7). Un nouveau brin est alors synthétisé par une polymérase (Figure 7a) : il est fixé de façon covalente à la FC. Le brin d'origine est alors éliminé par dénaturation (b) et l'extrémité libre du brin restant s'hybride à une amorce adjacente pour former un pont (c). La polymérase synthétise à nouveau le brin complémentaire pour former un pont d'ADN double brin (d) puis les deux copies sont libérées par dénaturation (e). Le cycle d'amplification en pont (étapes c à e) recommence pour

former à terme un regroupement d'ADN clonal en une zone appelée *cluster* (f). Les brins anti-sens (correspondant aux amorces vertes) sont ensuite clivés (g) : c'est la linéarisation.

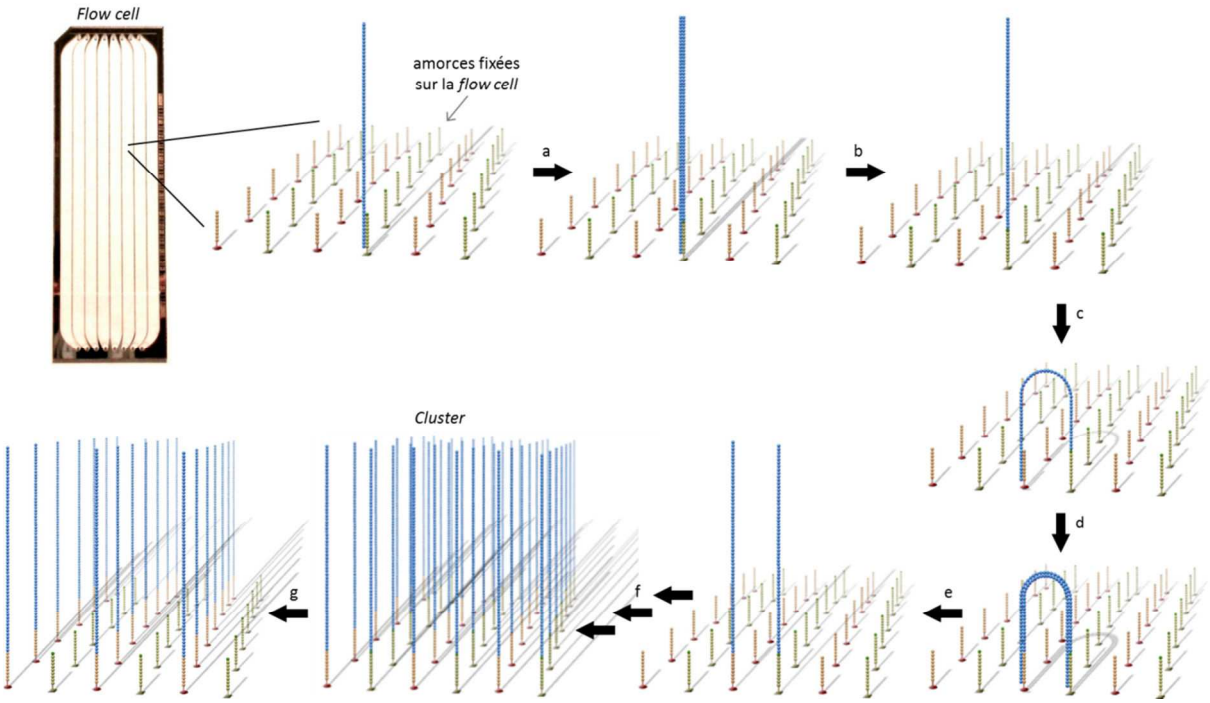


Figure 7: Amplification en pont de la technologie Illumina
a) Synthèse du brin complémentaire par une polymérase. b) Dénaturation. c) Formation d'un pont. d) Synthèse du brin complémentaire par une polymérase. e) Dénaturation. f) Nombreux cycles c à e. g) Linéarisation. Réalisé avec des images issues du site Internet d'Illumina.

L'extrémité 3' libre des fragments d'ADN est bloquée et l'amorce de séquençage s'y hybride (voir Figure 8). Le séquençage s'effectue sur des centaines de millions de *clusters* simultanément, grâce à une chimie de terminateurs réversibles : des nucléotides bloqués marqués par fluorescence sont ajoutés, l'un d'entre eux est incorporé, la fluorescence émise est relevée puis le fluorophore et le bloqueur sont clivés permettant l'ajout d'un nouveau nucléotide. A chaque cycle d'incorporation, une base peut être déterminée.

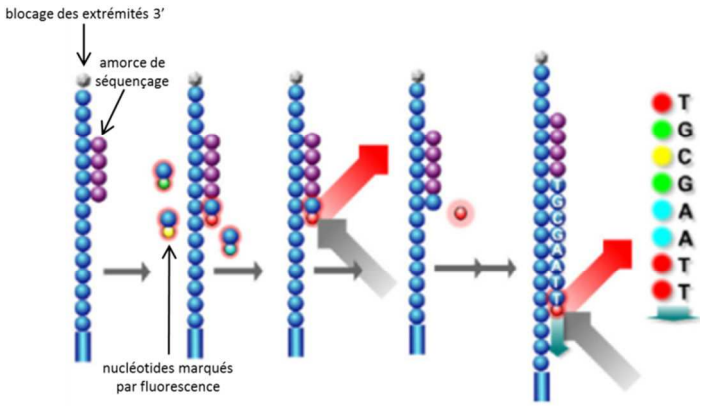


Figure 8: Aperçu de la technologie de séquençage Illumina
Adapté d'une image issue du site Internet d'Illumina.

Cette chimie a pour avantage de séquencer correctement les homopolymères. Le Genome Analyzer a cependant pour inconvénient de lire peu de bases, (36 à ses débuts, jusqu'à 150 bases aujourd'hui), ce qui le rend toutefois approprié à l'analyse de génomes dont on a une bonne annotation. En novembre 2006, Illumina a acquis Solexa et est maintenant le leader sur le marché du séquençage, notamment avec le lancement en février 2010 de son HiSeq2000, à un prix de 690000 $, qui peut produire jusqu'à 600 Gb en 10 jours, soit presque 700 fois plus que le dernier séquenceur de Roche, grâce à une méthode optique qui permet de travailler sur les deux surfaces de la FC et sur deux FCs en parallèle. Elle planifie désormais de l'optimiser pour aboutir mi-2012 au HiSeq2500 qui permettrait le séquençage d'un génome en un seul jour. Illumina possède la plateforme de séquençage la plus largement utilisée à travers le monde, en proposant un coût de séquençage du génome humain de seulement 10000 $.

I.2.2.1.3 SOLiD

Le SOLiD (Sequencing by Oligonucleotide Ligation and Detection) a été la troisième plateforme de séquençage de nouvelle génération, commercialisée par Applied Biosystems (aujourd'hui Life Technologies) depuis 2007. La technologie repose également sur une PCR en émulsion sur billes. Le séquençage ne s'effectue pas par synthèse comme sur les plateformes précédentes mais par ligation. Une amorce de séquençage universelle se fixe sur l'adaptateur puis des oligonucléotides dégénérés de 8 bases, marqués par fluorescence, sont ajoutés (voir Figure 9).

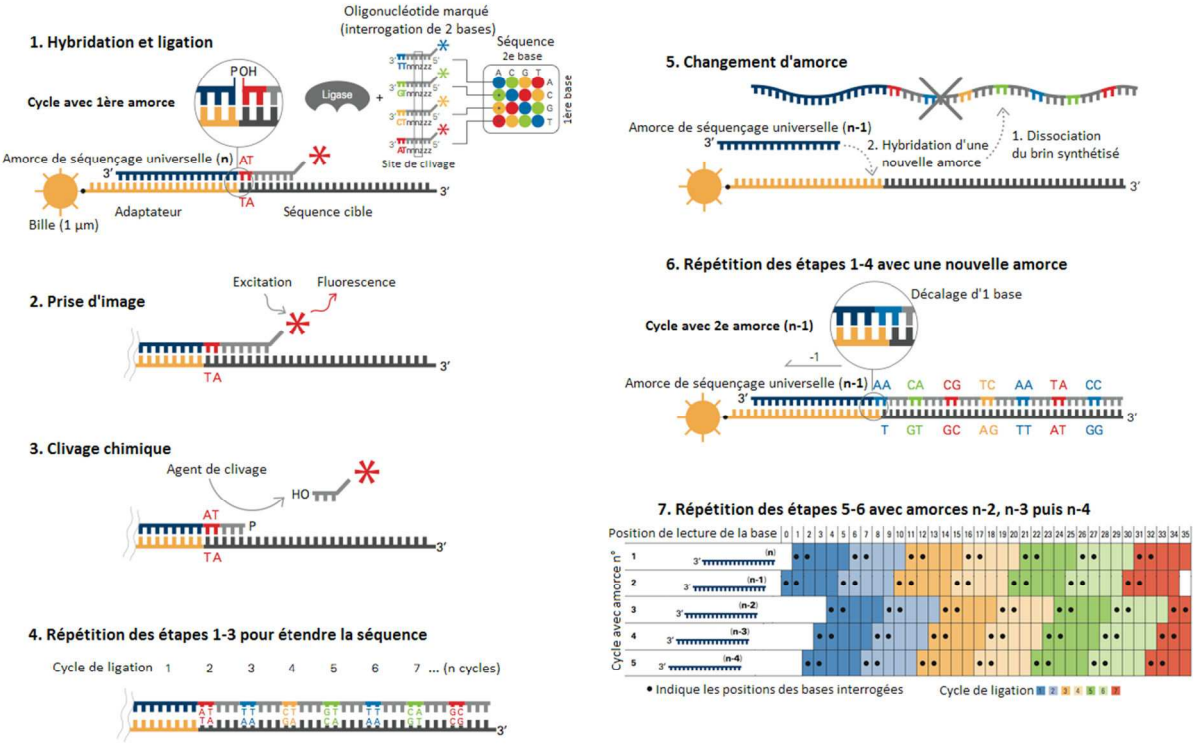


Figure 9: Aperçu de la technologie de séquençage SOLiD.
Adapté d'images issues du site Internet du Next-Generation Sequencing Center, John Hopkins University.

Dès que l'un d'entre eux correspond à la séquence adjacente à l'amorce, la ligase le fixe et de la fluorescence est émise, permettant d'identifier l'oligonucléotide fixé et d'interroger ainsi ses deux premières bases. Un clivage chimique retire les bases 6 à 8 ainsi que le fluorophore puis les oligonucléotides sont alors ajoutés à nouveau : on identifie les bases 6 et 7 de notre fragment à séquencer, puis dans un troisième temps les bases 11 et 12, et ainsi de suite jusqu'aux bases 31 et 32. Un deuxième cycle de ligation est alors entamé avec une amorce universelle se fixant en position n-1 et on identifie les bases en position 0 et 1, 5 et 6, et ainsi de suite jusqu'aux bases 30 et 31. 3 nouveaux cycles de ligation sont effectués grâce à des amorces se fixant en n-2, n-3 puis n-4. Le nombre de cycles de ligation, détection et clivage détermine ainsi la longueur de lecture, de 35 bases dans le cas présenté ici à 75 bases. Chaque base est lue deux fois avec cette technologie, ce qui explique sa grande précision et qui la rend adaptée au reséquençage ou à l'analyse de polymorphismes. Néanmoins, la complexité de fonctionnement de cette technologie en est un inconvénient puisqu'elle implique un lourd travail d'analyse.

I.2.2.2 Les cadets de la 2nde génération

I.2.2.2.1 Ion Torrent

En 2007, Jonathan M. Rothberg fonde la compagnie Ion Torrent, rachetée en 2010 par Life Technologies, qui commercialisera en décembre 2010 son premier séquenceur, le PGM (Personal Genome Machine) pour 50000 $. Sa technologie repose sur la libération naturelle d'un ion H^+ après incorporation d'un nucléotide par une polymérase. Ce phénomène entraine une modification du pH pouvant être détectée par une puce de silicium semi-conductrice composée de plusieurs millions de transistors. Cette technologie est qualifiée de PostLight car aucun intermédiaire de lumière n'est utilisé contrairement aux méthodes citées précédemment : c'est une modification chimique qui entraine la création du signal. L'absence de marquage fluorescent et de système de détection optique ou encore l'utilisation de micropuces standards explique le faible coût de cette machine qui pourrait avoir le même impact sur les technologies que l'ordinateur personnel il y a une trentaine d'années, à condition d'en améliorer encore les taux d'erreurs existants.

La dernière version de ce séquenceur (puce 318) lui permet désormais de lire 1 Gb avec une grande précision en seulement deux heures et la longueur de lecture atteindra les 400 pb en 2012. Le PGM a ainsi servi à décoder le génome de la souche d'*Escherichia coli* qui a fait des ravages en Europe et provoqué un scandale sanitaire en mai 2011 (102). L'équipe de Rothberg a ensuite utilisé cette technologie pour séquencer le génome de Gordon Moore, co-fondateur d'Intel, à une profondeur de 10,6x (103). Life Technologies envisage le lancement d'un nouvel instrument pour mi-2012, le Ion Proton, qui permettrait le séquençage d'un génome en 1 jour pour 1000 $.

I.2.2.2.2 Polonator G007

Le Polonator G007 est distribué par Danaher Motions, une filiale du groupe Dover. C'est une plateforme libre d'accès grâce à laquelle les utilisateurs préparent eux-mêmes leurs réactifs en se basant sur les travaux publiés par le laboratoire de George Church (104). L'accès aux protocoles et aux logiciels est gratuit. Cette technologie est basée sur une amplification en émulsion et sur un séquençage *via* une synthèse par ligation qui fonctionne de la même manière que sur le SOLiD mais seule une base est interrogée à l'aide d'oligonucléotides dégénérés.

I.2.2.2.3 Complete Genomics

Complete Genomics a développé une technologie permettant de séquencer de faibles quantités d'ADN à faible coût en réactifs sur une lame de silicium. La librairie est préparée par RCA (*Rolling Circle Amplification*) qui permet de créer des *nanoballs* d'ADN (105). Le séquençage s'effectue ensuite par ligation. Cette technologie a permis de séquencer 3 génomes humains à une profondeur de 45 à 87x pour 4400 $ par génome. Cependant, Complete Genomics ne propose désormais que des services de séquençage haut-débit.

I.2.2.3 La « miniaturisation » des séquenceurs

En parallèle des différentes versions ou modèles existant pour ces séquenceurs (voir Tableau 3 et Figure 10) les entreprises développent des versions plus petites de leurs machines appelées séquenceurs de paillasse qui seront à la portée de plus petits laboratoires.

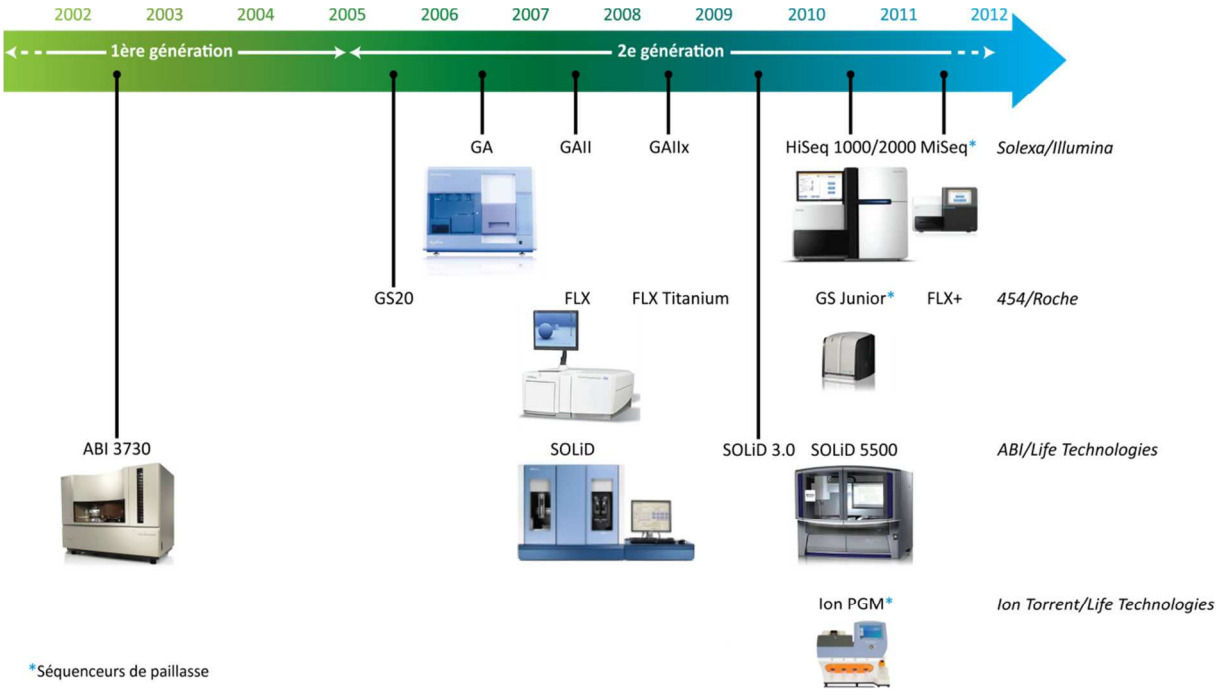


Figure 10: Chronologie de la commercialisation des séquenceurs haut-débit (1ère et 2e générations)
D'après des informations issues des sites Internet respectifs des fournisseurs.

Roche a lancé en 2010 le GS Junior, commercialisé à 100000 $ et capable de générer 35 Mb en seulement 10h. Illumina a présenté son MiSeq en janvier 2011 pour 125000 $, capable de générer 1 Gb en 26h et dont les performances concurrencent directement le séquenceur de paillasse PGM de Life Technologies. Ce dernier est vendu plus de la moitié du prix d'un MiSeq mais les utilisateurs, la plupart ayant une plus grande expérience de la technologie Illumina, préfèreront opter pour un séquenceur Illumina.

I.2.3 Séquençage de 3^{e} génération

Grâce aux avancées en microfluidique et en technologie des nanopores, la troisième génération est bientôt prête à émerger. La différence majeure entre la deuxième et la troisième génération de séquenceurs réside dans leur capacité à séquencer directement des molécules d'ADN de façon individuelle sans aucune amplification préalable. On peut donc également séquencer de l'ARN sans devoir le convertir au préalable en ADNc. Deux plateformes de séquençage de troisième génération ont été ou sont actuellement disponibles sur le marché et une dernière est en stade de développement avancé (voir Tableau 3).

I.2.3.1 HeliScope

Helicos Biosciences a développé le premier séquenceur d'une molécule unique : le HeliScope Single Molecule Sequencer. A la différence de la chimie utilisée sur la plateforme Illumina, ici les nucléotides sont marqués avec le même fluorophore. Helicos ne vend désormais plus d'appareil et propose un service de séquençage.

I.2.3.2 Technologies des nanopores

Le séquençage par nanopores est une technologie prometteuse puisqu'elle permet de déterminer une séquence ADN à la résolution du nucléotide sans aucune amplification, par lecture directe. Les différentes bases et leur statut de méthylation peuvent être déterminés en temps réel grâce au courant qui traverse le pore, avec une grande précision (99,8%) et à faible coût puisqu'aucun système optique n'est nécessaire ainsi que peu de traitement des données (106). La longueur de lecture est aussi plus grande que celle des autres technologies.

Une molécule unique d'ADN traverse un pore formé par une protéine ancrée dans une bicouche lipidique par application d'un potentiel (voir Figure 11). Une exonucléase clive chaque nucléotide à l'entrée du nanopore et celui-ci est alors détecté de façon électronique *via* une cyclodextrine (107). Aucun traitement ni marquage préalables ne sont nécessaires et de faibles quantités d'ADN suffisent.

Le deuxième point fort de cette technologie est sa capacité à identifier les modifications épigénétiques puisqu'une 5-méthylcytosine peut être différenciée d'une cytosine.

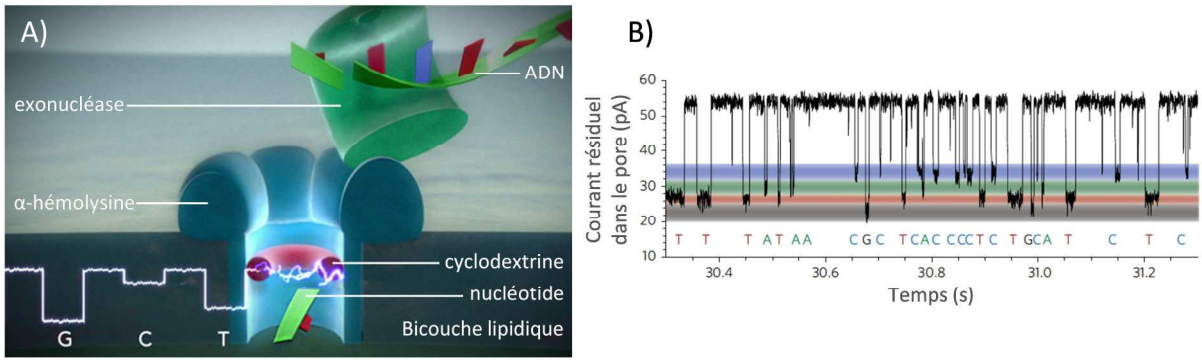


Figure 11: Détection de nucléotides par l'utilisation d'un nanopore
A) Structure utilisée pour le séquençage d'une molécule d'ADN unique ; figure modifiée d'après http://www.technologyreview.com/biomedicine. B) Courant résiduel mesuré à travers le nanopore permettant de discriminer chaque nucléotide ; figure modifiée d'après Clarke 2009 (107).

En 2010 certains groupes ont mis en œuvre 2 principaux types de pores protéiques : le laboratoire de Hagan Bayley a d'abord optimisé un pore à partir d'une protéine mutée d'α-hemolysine A de staphylocoque. Un groupe de l'université de Washington a ensuite développé un nanopore à partir de la porine A de *Mycobacterium smegmatis* ou MspA supposée avoir de meilleures propriétés grâce à sa plus petite taille (108). Oxford Nanopore Technologies, dont les travaux se basent sur ceux de Hagan Bayley, optimise également différents nanopores. Certaines équipes mettent aussi au point des nanopores à base de silicium ou encore de graphène qui est un matériau conducteur et qui a pour avantage de servir de pore et d'électrodes à la fois (109). Enfin on a vu récemment le développement de nanopores hybrides grâce à des protéines intégrées dans une membrane à l'état solide (110). Les défis actuels de ces technologies restent la gestion d'un brin d'ADN qui se meut trop rapidement à l'entrée ou à l'intérieur du pore, et de prouver qu'elles fonctionnent sur de longs fragments de 1 à 5 kb. Aucune machine n'est commercialisée à l'heure actuelle, mais Oxford Nanopore Technologies vient d'annoncer le lancement de son instrument GridION pour 2012 ainsi que d'un système de la taille d'une clé USB, le MinION dont le coût ne devrait dépasser les 900 $.

I.2.3.3 Pacific Biosciences

Pacific Biosciences a développé le premier instrument capable de séquencer une molécule unique en temps réel, là encore par lecture directe et sans amplifier le matériel de départ : il s'agit du PacBio RS. Il utilise une structure composée de cellules SMRT (Single Molecule Real Time) (111,112). Chacune de ces cellules contient 75000 nanostructures appelées détecteurs ZMW (Zero-Mode Waveguide) de 100 nm de diamètre, c'est-à-dire plus petits que les longueurs d'onde utilisées sur la plateforme (532 et 643 nm). La lumière ne peut donc pas s'y propager, d'où le terme de mode-zéro.

Chacun de ces ZMW contient une polymérase qui y est immobilisée et qui incorpore des nucléotides liés à un fluorophore, libérant ainsi celui-ci à l'extérieur du ZMW et mettant fin au signal lumineux. L'intervalle de temps entre chaque pic ainsi que la durée de chaque pic de fluorescence sont propres à chaque nucléotide et permettent ainsi leur identification. Ces paramètres sont également différents pour des nucléotides méthylés et il est donc possible, sans traitement préalable, de différencier une cytosine d'une méthylcytosine ou encore d'une hydroxyméthylcytosine (113,114).

La longueur de lecture moyenne est de plus de 1000 pb, la préparation de l'échantillon dure environ 30 minutes et les données sont produites en quelques minutes seulement, ce qui en fait une des technologies les plus prometteuses des années à venir.

I.2.3.4 Starlight

Life Technologies développe un séquenceur de molécule unique appelé Starlight qui utilise des *quantum dots*, des particules nanométriques à base de semi-conducteur et dont les propriétés de fluorescence peuvent être contrôlées par leur taille. Excitées par un laser, ces particules transfèrent leur énergie aux nucléotides marqués par fluorescence lorsque ceux-ci sont incorporés par la polymérase (115). De plus, il est possible de remplacer la polymérase lorsqu'elle s'essouffle. On peut ainsi théoriquement séquencer la longueur désirée de fragment. En l'absence de communication de la part de Life Technologies, le doute subsiste quant à une éventuelle commercialisation de cet appareil.

I.2.4 La génomique personnelle

Plusieurs compagnies offrent désormais des services de génomique personnelle. Elles s'adressent directement au grand public et proposent de décoder son génome à des prix relativement abordables. Il s'agit en réalité d'étudier des polymorphismes impliqués dans un panel de maladies. C'est le cas de 23andMe, deCODEme, Navigenics et bien d'autres qui proposent pour parfois moins de 100 $ de prédire votre avenir génétique. D'autres grands du séquençage proposent cette fois-ci de séquencer votre génome complet. C'est le cas de Complete Genomics ou encore de Knome qui propose de vous révéler les secrets de votre ADN pour 5000 $.

Tous les séquenceurs déjà commercialisés ont été certifiés pour une utilisation dans des laboratoires de recherche mais les scientifiques tentent de plus en plus de les orienter vers une utilisation clinique grâce à laquelle les médecins pourraient traiter leurs patients de façon personnalisée en leur fournissant des traitements adaptés. Il faut cependant désormais donner un sens biologique à ces quantités pharaoniques de données produites depuis plusieurs années avant d'espérer entrer dans cette nouvelle ère de la biologie. Le *Personal Genome Project* lancé en 2009 recrute des volontaires

prêts à donner leur génome à la science. Il a pour but de séquencer 100000 génomes afin de construire une base de données qui permettrait de mieux les appréhender. De nouveaux challenges informatiques sont alors lancés (116) afin de gérer ces quantités de données, les interpréter de façon correcte mais également les protéger pour assurer une confidentialité aux patients. De nombreux problèmes éthiques sont désormais soulevés (117,118) et la communauté scientifique devra rapidement prendre ce sujet en main pour encadrer au mieux l'utilisation des données de génomique personnelle.

I.3 Etude du méthylome par séquençage

Après la publication du premier méthylome, celui de la plante *Arabidopsis thaliana* (119,120), deux autres publications récentes ont montré que ceux de l'homme et des mammifères en général étaient maintenant à notre portée, en présentant deux méthylomes humains à la résolution du nucléotide (121,122). Cependant, l'analyse de la méthylation sur le génome entier reste très chère et la plupart des stratégies développées enrichissent d'abord une fraction du méthylome. Elles sont principalement basées sur quatre axes (voir Figure 12).

I.3.1 Protéines MBD

Certaines protéines de la famille des MBD (Methyl-CpG-Binding Domain) ont la faculté de détecter les CpGs méthylés de façon symétrique (123). Toutes ces protéines, dont MeCP2 (Methyl CpG Binding Protein 2), MBD1, MBD2, MBD3 et MBD4 (Methyl-CpG Binding Domain protein 1-4) contiennent un domaine de liaison aux CpGs méthylés composé de 70 acides aminés. Il a été montré qu'un environnement spécifique permettait cette liaison ou augmentait son affinité, comme la présence de molécules d'eau environnantes (124). Dans le cas de MeCP2, une séquence AT proche du site de fixation favorise la liaison et induira ainsi une certaine spécificité lors de l'utilisation de cette protéine (125).

Ces différences d'affinité peuvent être utilisées pour sélectionner les fractions méthylées d'un échantillon ou pour séparer ses fractions différentiellement méthylées. C'est ce qu'une équipe a mis au point grâce à une colonne par affinité servant de support solide sur lequel le domaine MBD de MeCP2 a été fixé (126). Diagenode a développé un kit appelé MethylCap (Methyl-DNA Capture) basé sur ce principe : l'ADN est fragmenté de façon aléatoire puis sélectionné par le domaine MBD de MeCP2 fusionné à une protéine GST (Glutathione-S-Transferase) qui permet de récupérer ces fragments après ajout de billes magnétiques couvertes de glutathion. Life Technologies a de son côté développé le kit MethylMiner qui utilise la protéine MBD2 biotinylée et les fragments méthylés sont par la suite isolés grâce à des billes magnétiques recouvertes de streptavidine.

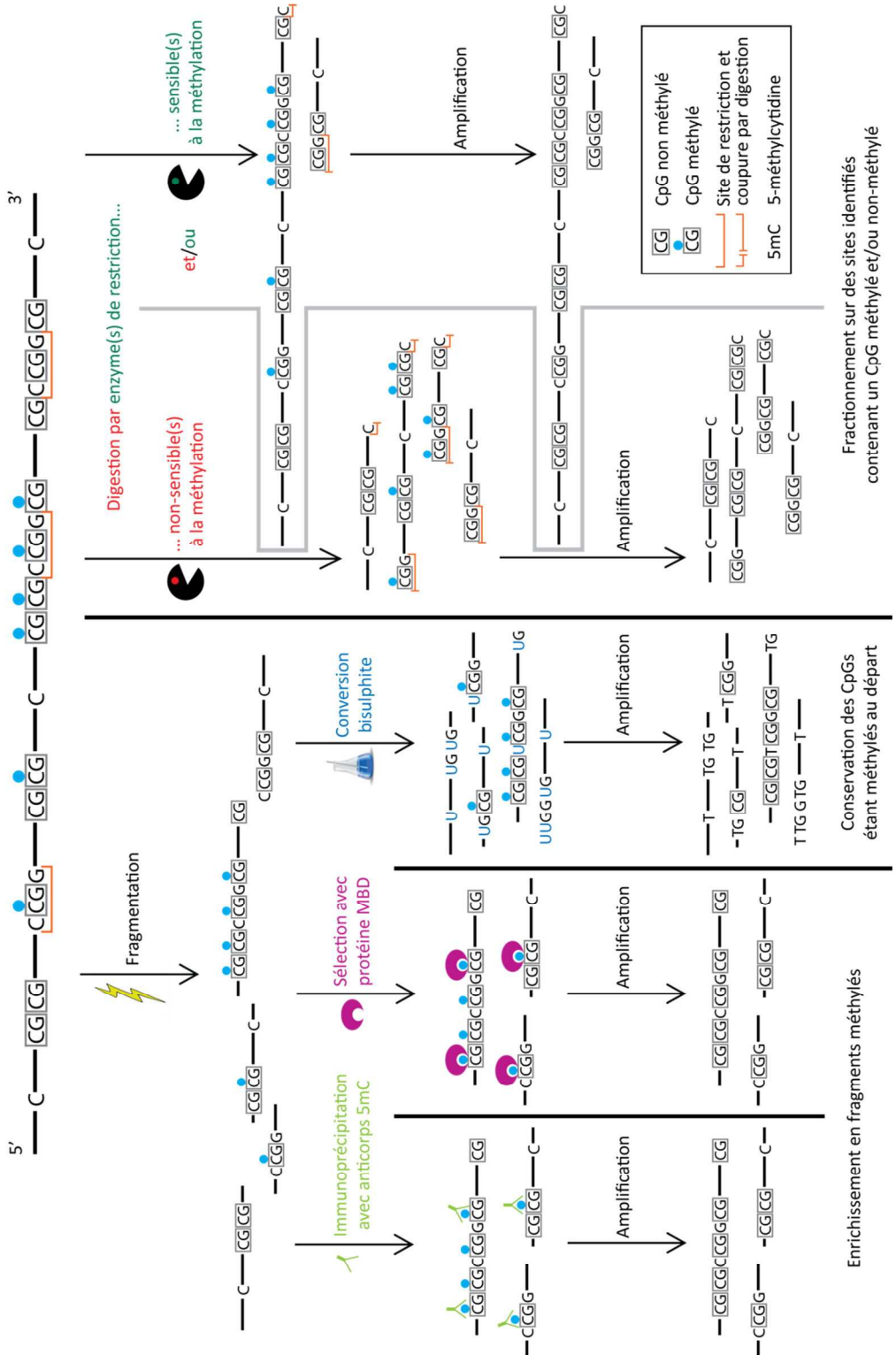


Figure 12: Les 4 techniques utilisées pour l'analyse de la méthylation

Dans les deux cas, après lavage, l'ADN méthylé est élué par ajout d'une solution saline dont la concentration est corrélée à la densité en CpGs des fragments élués.

La précipitation d'ADN méthylé par les protéines de la famille MBD est désormais combinée au séquençage de seconde génération pour constituer le MBD-Seq (MBD2) ou le MethylCap-Seq

(domaine MBD de MeCP2) (127-130). Un protocole similaire mais baptisé MiGS (MBD-isolated Genome Sequencing) a permis de confirmer des données de méthylation existantes et d'identifier de nouvelles régions méthylées du génome humain impliquées dans certains cancers (27).

La présence d'une protéine MBD3L1 (Methyl-CpG-Binding Domain protein 3-Like-1) augmente fortement l'interaction entre la protéine MBD2 et les fragments méthylés en permettant la formation d'un hétérodimère. La technique MIRA (Methylated CpG island recovery Assay) (131,132) en fait la preuve en combinant ces deux protéines pour sélectionner la fraction non-méthylée avant séquençage (MIRA-Seq) (133).

I.3.2 Anticorps dirigé contre les 5-méthylcytidines

Un anticorps dirigé contre les 5-méthylcytidines (le clone 33D3 étant majoritairement utilisé), présente une grande affinité pour cette base et est utilisé pour immunoprécipiter les fragments méthylés du génome. Dans cette technique appelée MeDIP (Methylated DNA ImmunoPrecipitation), l'enrichissement dépend de la densité en CpGs des fragments et la présence minimale de 2 CpGs est nécessaire pour immunoprécipiter un fragment de 100 bases (134).

Le MeDIP a été développé en 2005 (135) puis combiné à une analyse sur puce (MeDIP-chip) sur laquelle un échantillon immunoprécipité est co-hybridé avec l'échantillon n'ayant subi aucun traitement (136,137). Ce type de technique permet d'étudier la méthylation sur certaines régions du génome comme les promoteurs ou les îlots CpG. Récemment, les premiers méthylomes couvrant l'intégralité du génome ont été publiés grâce à l'utilisation du séquençage haut-débit après MeDIP, c'est le MeDIP-Seq. D'un point de vue expérimental, l'ADN est fragmenté puis les adaptateurs (permettant l'amplification puis le séquençage) sont fixés aux extrémités des fragments. Ceux-ci peuvent également être fixés après l'immunoprécipitation (138). Après dénaturation, l'ADN méthylé simple brin est immunoprécipité grâce à l'anticorps spécifique des 5-méthylcytidines. Le complexe ADN-anticorps est alors capturé par un second anticorps anti-IgG (ou la protéine A) fixé sur des billes et l'ADN méthylé en est relâché après lavage et traitement à la protéinase K.

Grâce à cette technique, le méthylome de lignées de cancer du sein a pu être étudié (122). Il a ainsi été montré que les îlots CpG étaient 4 fois plus méthylés dans les échantillons de cancer en comparaison à des échantillons normaux et que cette hyperméthylation affectait également des régions situées à l'extérieur des îlots. Plus tard, le MeDIP-Seq a permis de fournir le profil de méthylation de spermatozoïdes humains en couvrant 60% des CpGs répertoriés (121) ou d'identifier des nouvelles régions impliquées dans la régulation épigénétique des cancers en étudiant différents types de tumeurs du système neveux (139). Dans cette dernière étude, aucune hypométhylation

globale du génome n'a pu être montrée contrairement à ce qui est généralement admis pour d'autres types de tumeurs.

Les méthodes d'enrichissement par affinité (immunoprécipitation et sélection par protéines MBD) ont l'avantage de ne pas être restreintes à l'analyse d'un nombre limité de sites, contrairement aux méthodes basées sur l'utilisation d'enzymes (voir ci-après). Cependant, elles nécessitent une grande quantité d'ADN et ont tendance à enrichir les régions riches en CpGs, notamment les séquences répétées.

I.3.3 Enzymes de restriction sensibles à la méthylation

Les endonucléases de restriction sensibles à la méthylation peuvent être utilisées pour fractionner le génome en portions méthylées et non-méthylées. L'utilisation de ces enzymes a pour caractéristique de simplifier le génome à étudier puisqu'elle en donne une représentation réduite.

Enzyme	Sensibilité si CpG méthylé	Site de restriction
isoschizomères: *Hpa*II		5'… C▼C G G … 3' 3'… G G C▲C … 5'
isoschizomères: *Msp*I		
isoschizomères: *Sma*I		5'… C▼C C▼G G G … 3' 3'… G G G▲C C▲C … 5'
isoschizomères: *Xma*I		
*Aci*I		5'… C▼C G C … 3' 3'… G G C▲C … 5'
*Hin*6I		5'… G▼C G C … 3' 3'… C G C▲G … 5'
*Mlu*I		5'… A▼C G C G T … 3' 3'… T G C G C▲A … 5'
*Dpn*I		5'… G A^{CH_3}▼T C … 3' 3'… C T▲A G … 5' ^{CH_3}
*Mme*I		5'… T C C R A C $(N)_{20}$▼… 3' 3'… A G G Y T G $(N)_{18}$▲… 5'
*Alu*I		5'… A G▼C T … 3' 3'… T C▲G A … 5'
*Nla*III		5'…▼C A T G … 3' 3'… G T A C▲… 5'
*Mcr*BC		5'… R C^{CH_3} ▼ $(N)_{55\text{-}103}$ * R C^{CH_3} … 3' *Séparation optimale entre les 2 sites

Insensible: coupe, que le CpG au niveau du site de coupure soit méthylé ou non

Dépendante: ne coupe que si le CpG au niveau du site de coupure est méthylé

Affaiblie si le CpG au niveau du site de coupure est méthylé

Sensible: ne coupe pas si le CpG au niveau du site de coupure est méthylé

Sensible par chevauchement: ne coupe pas si un CpG dans le palindrome (hors du site de coupure) est méthylé

▼▲ Sites de coupure

R = A/G
Y = C/T

Tableau 4: Endonucléases de restriction majoritairement utilisées et leurs sites de restriction
La sensibilité à la présence d'un CpG méthylé sur le site de coupure ou dans le palindrome est indiquée. Données provenant des sites internet de New England Biolabs et Fermentas.

Une enzyme peut être utilisée en parallèle avec l'un de ses isoschizomères (qui présente une sensibilité différente à la méthylation). Le couple *Sma*I/*Xma*I (voir Tableau 4 pour leurs caractéristiques) constitue par exemple la base de la technologie MCA (Methylated CpG island Amplification) qui a été combinée à des puces à ADN (140). Le couple *Hpa*II/*Msp*I est utilisé dans une autre approche appelée HELP *(Hpa*II tiny fragments Enrichment by Ligation-mediated PCR) qui cible les loci hypométhylés (141). L'ADN est digéré par *Hpa*II sensible à la méthylation et les plus petits fragments (200-2000 pb) sont sélectionnés et amplifiés par LM-PCR (Ligation-Mediated PCR). Les produits de digestion par *Msp*I, isoschizomère de *Hpa*II, subissent le même traitement. Après co-hybridation sur puce, on obtient le statut de méthylation de ces fragments en comparant les 2 librairies : les fragments non présents dans la librairie créée grâce à *Hpa*II sont considérés comme méthylés. Une amélioration de ce protocole appelée nanoHELP a permis une meilleure représentation des régions riches en CpGs ainsi que l'utilisation de plus petites quantités d'ADN (jusqu'à 10 ng) grâce à l'analyse par séquençage haut-débit des librairies générées (142). Methyl-Seq est une méthode similaire où la LM-PCR est remplacée par une sélection des tailles de fragments sur gel (143). Le statut de méthylation de 90000 régions a ainsi pu être analysé par alignement sur des produits de digestion *in silico*. Dans les méthodes précédemment citées, seuls les petits fragments de digestion sont pris en compte de façon à ce que les régions peu riches en CpGs qui résultent en de plus long fragments ne soient pas considérées.

Dans une nouvelle approche appelée MSCC (Methylation-Sensitive Cut Counting) 1,4 millions de sites *Hpa*II ont pu être analysés sur des lymphocytes B (144). Aucune sélection n'a été effectuée en fonction des tailles de fragments et seuls les produits de digestion par *Hpa*II ont été utilisés, sélectionnant des CpGs non-méthylés, sans comparaison avec les produits de digestion par son isoschizomère. Le nombre de fragments séquencés est alors directement corrélé au niveau de méthylation : les sites présentant beaucoup ou pas de *reads* sont estimés comme peu ou très méthylés respectivement (145). Il est également intéressant d'utiliser un cocktail d'enzymes sensibles à la méthylation comme dans le MRE-Seq (Methylation-sensitive Restriction Enzymes Sequencing) où *Hpa*II, *Hin*6I et *Aci*I ont été employées (146). Cette méthode peut être combinée au MeDIP-Seq puisque les 2 techniques se complètent et les scores de méthylation obtenus sont inversement corrélés (147).

Dans un protocole appelé MSDK (Methylation-Specific Digital Karyotyping puis MMSDK pour Modified MSDK) (148,149), l'enzyme *Mlu*I reconnait des sites non-méthylés puis les fragments résultants sont biotinylés à leurs extrémités avant d'être à nouveau fragmentés à l'aide d'une enzyme insensible à la méthylation, *Nla*III. Après liaison à des billes magnétiques de streptavidine, les fragments méthylés sont ainsi séparés des non-méthylés. Ces derniers sont alors fragmentés à

nouveau par digestion par *Mme*I puis amplifiés par PCR pour être séquencés. La combinaison de plusieurs enzymes est désormais courante (150).

La plupart des méthodes présentées ici ciblent les fragments non-méthylés. Il est également possible de cibler les fragments méthylés en utilisant des enzymes dépendantes de la méthylation comme *Mcr*BC (151).

Toutes ces méthodes ne peuvent être considérées comme couvrant tout le génome puisqu'elles limitent l'étude aux CpGs qui se situent dans les séquences spécifiques à ces enzymes. De plus, ces techniques nécessitent la plupart du temps de grandes quantités d'ADN et des biais peuvent facilement y être introduits dans le cas où la digestion par les enzymes serait incomplète. Elles permettent néanmoins de détecter des différences de méthylation et de les quantifier dans des régions de faible densité en CpGs.

I.3.4 Conversion par le bisulphite

Les polymérases utilisées dans les réactions de PCR nécessaires avant séquençage ne peuvent différencier une cytosine méthylée d'une non-méthylée et l'état de méthylation des CpGs est perdu après amplification. Le bisulphite est une méthode chimique qui va permettre de préserver l'information sur leur statut de méthylation. Pour cela, le bisulphite de sodium provoque une désamination des cytosines non méthylées en uraciles, que la polymérase remplacera par des thymines au cours de la PCR, tandis que les cytosines méthylées ne sont pas affectées, la cinétique étant plus lente (152) (voir Figure 13A). La modification épigénétique non analysable dans sa forme brute est donc transformée en un polymorphisme C/T dont l'analyse est plus facilement abordable.

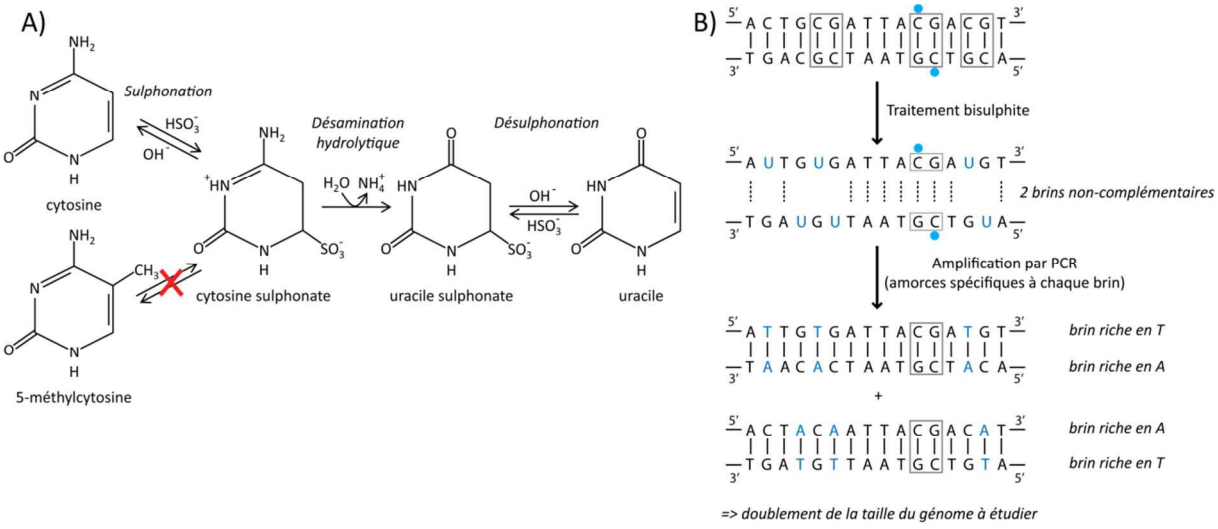


Figure 13: Principe du traitement par le bisulphite
A) Réactions de sulphonation, désamination et désulphonation provoquées par le traitement par le bisulphite de sodium. Le groupement méthyle d'une 5-méthycytosine la protège de la sulphonation et la réaction n'a pas lieu. B) Devenir des différents nucléotides d'un double brin d'ADN face au traitement par le bisulphite suivi d'une amplification par PCR.

Les méthodes basées sur la conversion par le bisulphite sont considérées comme de référence depuis que Frommer a mis au point ce protocole en 1992 (153), amélioré ensuite par Clark en 1994 (154). A la différence des méthodes par affinité, les méthodes basées sur la conversion par le bisulphite fournissent des informations sur la méthylation à la résolution du nucléotide. De plus, l'ADN de départ peut être d'une pureté médiocre puisque le protocole de conversion inclut une purification afin d'éliminer le bisulphite de sodium. Cependant, les conditions drastiques du traitement chimique peuvent dégrader l'ADN. De plus, une dénaturation incomplète de l'ADN peut causer une conversion par le bisulphite incomplète car le traitement n'a pas d'effet sur l'ADN double brin. Après traitement, les deux brins d'ADN ne sont plus complémentaires, la taille du génome à analyser est donc doublée (voir Figure 13B) et la complexité de séquence est diminuée puisque trois bases seulement prédominent, ce qui a pour conséquence d'alourdir les analyses en réduisant la quantité et la qualité de l'alignement après séquençage.

Le BS-Seq (Shotgun Bisulfite Sequencing) a permis de générer le méthylome de deux lignées cellulaires humaines (138) et d'*Arabidopsis thaliana* (120). Cette méthode consiste en une ligation d'adaptateurs contenant d'une part la séquence nécessaire pour le séquençage (avec des cytosines méthylées) et d'autre part un site de restriction enzymatique (avec des cytosines non méthylées). Après conversion par le bisulphite, une première PCR permet de sélectionner les fragments dont la conversion est complète puis une digestion enzymatique révèle les adaptateurs. Une seconde PCR permet de préparer la librairie pour son séquençage. Un des deux groupes a également mis en place un logiciel d'analyse des séquences obtenues appelé BS-Seeker (155).

Une méthode similaire appelée MethylC-Seq dans laquelle la conversion par le bisulphite est réalisée deux fois a également permis d'étudier le génome d'*Arabidopsis* (119) et a été appliquée au génome humain peu de temps après. Les scientifiques de ce groupe ont ainsi pu montrer que les cytosines peuvent être méthylées dans un contexte hors CpGs dans des cellules souches embryonnaires humaines, ce qui en fait une nouvelle caractéristique intrinsèque à ces cellules puisqu'elle est perdue dans les cellules différenciées (12).

Le coût d'une conversion par le bisulphite suivie du séquençage sur le génome entier est conséquent puisqu'une profondeur importante est nécessaire afin de pouvoir quantifier la méthylation avec une grande précision. Une des alternatives est alors de se limiter à des régions présentant un intérêt potentiel dans l'étude de la méthylation, par l'utilisation d'enzymes de restriction notamment. Le RRBS (Reduced Representation Bisulphite Sequencing) génère une représentation réduite du génome par digestion enzymatique. L'utilisation d'enzymes dont le site de restriction contient un CpG, telles que *Msp*I, permet d'obtenir une représentation efficace puisque chaque fragment contient un site d'intérêt. Des adaptateurs contenant des cytosines méthylées sont ensuite ajoutés

aux extrémités et les plus petits fragments sont sélectionnés sur gel d'agarose, soumis au traitement bisulphite puis amplifiés par PCR avant séquençage. Des fragments de taille trop longue seraient sujets à dégradation pendant le traitement au bisulphite. Le laboratoire d'Alexander Meissner qui a développé la technique a ainsi pu analyser la méthylation sur des échantillons murins, notamment des cellules souches pluripotentes (13,156,157) ou encore étudier la déméthylation durant l'érythropoïèse (158).

I.3.5 Etudes comparatives

Il est désormais intéressant de comparer les différentes méthodes présentées ici et certains groupes s'y sont déjà attelé. Les méthodes basées sur la conversion par le bisulphite ont pour avantage d'étudier le génome avec une résolution au nucléotide et elles peuvent quantifier la méthylation. Le MethylC-Seq couvre davantage de CpGs à travers le génome que le RRBS mais ce dernier est plus intéressant d'un point de vue financier si l'on tient compte du nombre de CpGs qu'il couvre (147). Les méthodes d'enrichissement (utilisation d'anticorps ou de protéines) sont encore plus rentables mais ont une résolution réduite et ne permettent pas de donner des valeurs de méthylation précises. Ceci implique de définir des niveaux de méthylation assez grossiers (peu, moyennement ou très méthylé), à moins d'investir dans la mise en place d'outils informatiques adaptés. C'est notamment pour cette raison que ces méthodes trouvent un intérêt à être combinées à des techniques qui permettent de couvrir les sites CpGs non méthylés comme le MRE-Seq.

MeDIP-Seq et MBD-Seq enrichissent sensiblement les mêmes régions, bien que les concentrations en sels utilisées lors de l'élution dans le protocole de MBD-Seq puissent influencer cette conclusion (159,160). Le RRBS, lui, se concentre davantage sur les îlots CpG et les régions promotrices. Un avantage supplémentaire des méthodes utilisant le MeDIP en comparaison aux autres techniques est sa capacité à distinguer une 5-méthylcytosine d'une 5-hydroxyméthylcytosine (161).

Quelle que soit la méthode utilisée, il est théoriquement possible d'augmenter la couverture du méthylome en séquençant à une plus grande profondeur ou en augmentant la longueur de lecture du séquenceur. Enfin la combinaison de plusieurs méthodes, bien que parfois coûteuse, peut permettre de couvrir de plus nombreuses régions de densités différentes en CpGs.

I.4 Etude de la méthylation sur un grand nombre de loci

En parallèle des techniques qui couvrent une grande majorité des CpGs du génome, de nouvelles approches se sont développées et permettent de cibler des régions précises, tout cela simultanément sur plusieurs échantillons. La principale stratégie de multiplexage utilise des amorces

spécifiques portant une séquence d'identification propre à chaque échantillon (162). Les séquences peuvent ainsi être attribuées aux échantillons d'origine après séquençage. Le traitement par le bisulphite est alors introduit dans les protocoles afin de les appliquer à l'étude de la méthylation.

I.4.1 Sélection par PCRs

Pour étudier des fragments du génome bien particuliers, de multiples PCRs peuvent être réalisées puis leurs produits respectifs réunis pour être séquencés (163,164). Afin d'en analyser le profil de méthylation, ceci peut être effectué sur de l'ADN converti par le bisulphite au préalable (165). Néanmoins, dans la plupart des études citées ici, les échantillons sont traités séparément et les cibles amplifiées individuellement, ce qui rend leur préparation fastidieuse. Une plateforme mise au point par Fluidigm permet d'alléger les expérimentations mais ne propose cependant que d'amplifier 96 régions sur 96 échantillons. De plus, une efficacité inégale des amplifications, encore plus prononcée sur de l'ADN bisulphité, pose problème pour le séquençage de ces amplifications multiplexes et aucune application à l'étude de la méthylation n'a à ce jour été rapportée.

I.4.2 Capture des fragments d'intérêt

D'autres approches développées ces dernières années proposent de cibler des régions d'intérêt par l'utilisation d'oligonucléotides avec lesquels elles s'hybrident, en parallèle et dans un même tube.

I.4.2.1 Hybridation sur sondes ou puces

Les sondes de capture s'hybrident sur toute la longueur des fragments ciblés. Un groupe a par exemple utilisé près de 51000 sondes biotinylées qui permettent la capture des fragments cibles grâce à des billes de streptavidine, d'après un protocole récemment publié (166). Le traitement par le bisulphite des fragments capturés a alors permis d'analyser de façon quantitative 1 million de CpGs distribués à travers des îlots CpG et des promoteurs (167). Une autre étude a présenté l'utilisation de près de 20000 sondes appelées *dU probes* couvrant 170000 CpGs pour la capture des fragments d'intérêt. Les sondes sont par la suite éliminées par digestion par une ADN uracyle glycosylase et les cibles traitées au bisulphite (168). Enfin, une technique appelée *bisulphite patch PCR* cible les régions désirées par hybridation d'oligonucléotidess spécifiques qui sont ensuite dégradés par une exonucléase avant le traitement des cibles par le bisulphite (169).

Une équipe a également développé un outil appelé BC-Seq (Bisulphite Capture Sequencing) (170) dans lequel, après sélection des tailles désirées, l'ADN est soumis à une conversion par le bisulphite et amplifié. Après PCR, les nouveaux brins formés et riches en A sont capturés par hybridation sur puce. L'ADN est ensuite élué puis amplifié à nouveau avant séquençage. Grâce à cette méthode, des

hauts niveaux de méthylation ont pu être observés dans certains exons de lignées cellulaires cancéreuses. Cette méthode de capture présente l'avantage d'être efficace quelle que soit la densité en CpGs.

I.4.2.2 Sondes *padlocks* et sélectors

Un nouveau type de sondes a été développé pour étudier des polymorphismes et mutations sur des fragments d'ADN génomique et a également permis de faire du génotypage *in situ* (171). Ces sondes qualifiées de *padlock probes* s'hybrident sur leurs régions cibles respectives d'ADN génomique puis chaque sonde est circularisée par action d'une polymérase et d'une ligase par complémentarité avec la séquence cible (voir Figure 14). Les cercles sont ensuite amplifiés et séquencés (172-174). En concevant des sondes qui s'hybrident à quelques nucléotides d'un SNP d'intérêt, il est ainsi possible d'en étudier l'hétérozygosité. Une équipe a par exemple pu étudier l'hypermutabilité de 24% des CpGs situés sur le chromosome 21 (variation des CpGs en CpAs ou en TpGs) à l'aide de plus de 53000 sondes (175).

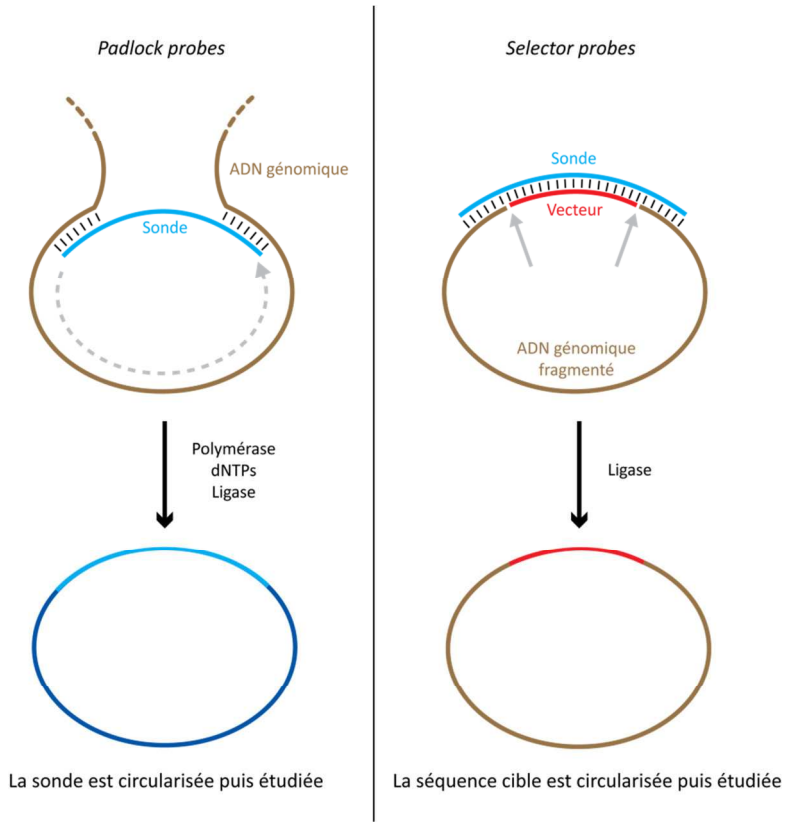


Figure 14: Comparaison entre les technologies de *padlock probes* et de *selector probes*

En parallèle, une équipe suédoise a développé une technique basée sur des sondes appelées *selector probes* ou sélectors qui permettent, après hybridation, de capturer les fragments cibles (voir Figure 14). Ce sont eux, et non pas la sonde elle-même comme dans la technique des *padlock probes*, qui

sont ensuite circularisés par ligation avec un oligonucléotide vecteur. Les fragments non ciblés et donc non circularisés sont éliminés par exonucléolyse (176,177). Il est également possible d'utiliser des sélectors biotinylés afin de sélectionner les cibles circularisées qui y sont hybridées, par capture sur billes de streptavidine (178). Les fragments d'intérêt sont alors amplifiés simultanément *via* le vecteur, puis analysés sur puce ou par séquençage.

La technique des *padlock probes* a été appliquée à l'étude de la méthylation pour la première fois en 2009 (145). Les sondes ont été conçues pour cibler au moins 1 CpG sur de l'ADN converti par le bisulphite et c'est ainsi que plus de 7000 CpGs ont pu être couverts en utilisant plus de 10000 *bisulphite padlock probes* (BSPPs) (144). En adaptant la conception des sondes, des îlots CpGs ont également pu être étudiés (66000 CpGs distribués sur 2020 îlots CpG à l'aide de plus de 30000 sondes puis dans un second temps plus de 480000 CpGs avec plus de 330000 sondes) (179,180). La technique des sélectors, elle, n'a jamais été appliquée à l'étude de la méthylation.

I.5 Projet de thèse

De toutes les techniques citées précédemment, le MeDIP-Seq est une méthode de choix et est au centre du projet présenté ici afin d'étudier le méthylome. Nous avons en complément développé une approche basée sur l'utilisation du bisulphite, permettant l'analyse de la méthylation sur des régions candidates.

I.5.1 Etude de la méthylation sur le génome entier par MeDIP-Seq

I.5.1.1 Eléments répétés du génome

Les séquences codantes pour les protéines représentent seulement 5% du génome humain tandis que plus de 50% sont envahis par des séquences répétées non codantes (80) qui constituent une trace du processus d'évolution des génomes. Celles-ci peuvent être réparties en 2 classes principales (181) elles-mêmes sous-divisées (voir Figure 15).

Les répétitions d'ADN dispersées sont des séquences que l'on retrouve tout le long du génome. Elles englobent :

- Des gènes paralogues dont des pseudogènes obtenus par duplication et qui consistent en des blocs entiers de 10 à 300 pb copiés/collés à travers le génome, et des rétropseudogènes obtenus par rétrotranscription.
- Les éléments transposables qui sont mobiles et comptent pour 45% des séquences répétées dans le génome humain et 39% dans celui de la souris *mus musculus* (182). Les éléments de classe I

sont des rétrotransposons et incluent les LINEs (*Long Interspersed Elements*, longs de 6 à 8 kb, représentant 21% chez l'homme et 19% chez la souris), les SINEs (*Short Interspersed Elements*, 100-400 pb, 13% et 8% respectivement) dont les plus abondants sont les éléments *Alu*, et les LTRs *(Long Terminal Repeats*, 1,5-11 kb, 8% et 10% resp.). Les éléments de classe II sont des transposons d'ADN (80 pb – 3 kb, 3% et 1% resp.).

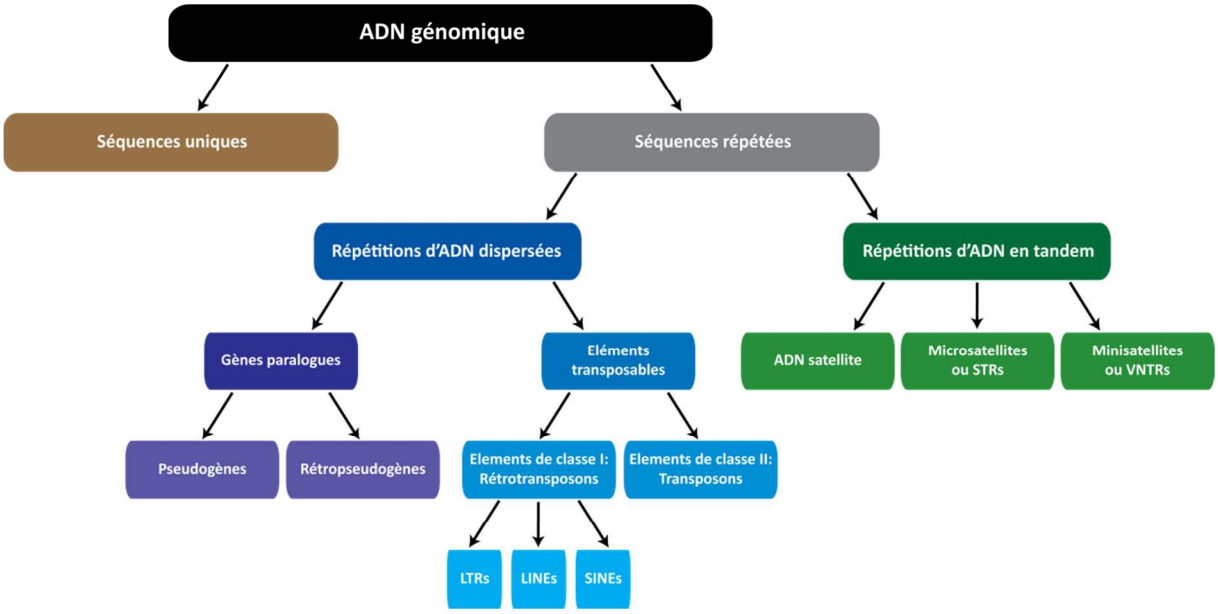


Figure 15: Les différentes familles d'éléments répétés dans le génome

Les répétitions d'ADN en tandem sont des séquences relativement courtes dans lesquelles des oligonucléotides sont répétés. Elles sont de deux natures :

- Des blocs de séquences répétées localisés dans les centromères et les télomères des chromosomes : ce sont les satellites. On peut citer l'α-satellite dont l'unité de répétition est de 171 pb chez l'humain. Dans les chromosomes acrocentriques de la souris en particulier, ils sont sous-divisés en satellites mineurs (l'unité de répétition étant de 120 bases) et majeurs (234 bases) (183).
- Des répétitions localisées dans un autre contexte : ce sont les microsatellites ou STRs (*Short Tandem Repeats*) quand l'unité de répétition est courte (1-13 bases) et les minisatellites ou VNTRs (*Variable Number Tandem Repeats*) quand elle est plus longue (14-500 bases).

L'intérêt croissant pour ces séquences répétées a mené à la création de bases de données comme Repbase au début des années 90 et mise à jour régulièrement (Repbase update) (184) ou encore RepeatMasker (http://www.repeatmasker.org) qui contient les séquences représentatives provenant de plusieurs de ces familles d'éléments répétés.

I.5.1.2 Intérêt de leur déplétion

Le MeDIP immunoprécipite les régions très méthylées dont la plupart sont localisées dans les séquences répétées du génome. Ainsi, 75% des régions répétées peuvent être couvertes par MeDIP-Seq tandis que seulement 3% le sont par MRE-Seq (qui sélectionne les régions non-méthylées) (146). Ceci corrobore le haut degré de méthylation de ces séquences qui représentent une part importante des génomes de mammifères.

De telles séquences posent problème lorsque des échantillons sont analysés par MeDIP-Seq. Bien qu'elles présentent un intérêt potentiel pour répondre à certaines questions biologiques ou cliniques, elles restent très difficiles à aligner de façon non-ambiguë après séquençage et sont, pour la plupart, écartées. Un grand nombre de séquences sont donc inutilisées pour la suite de l'analyse.

Après développement d'un protocole de MeDIP-Seq, nous proposerons une méthode innovante appelée MeDIP-dep-Seq (MeDIP et déplétion des éléments répétés suivi du séquençage) permettant de réduire de façon significative la quantité de ces séquences hautement répétées tandis que les séquences d'intérêt ne sont pas affectées. Ceci permettra alors d'exploiter de façon optimale le potentiel de séquençage de la plateforme choisie en libérant de la surface sur le support où a lieu le séquençage, surface qui sera directement mise à profit pour les séquences d'intérêt.

I.5.2 Etude de la méthylation sur un grand nombre de loci *via* les sélectors

Les régions candidates identifiées dans des études sur le génome entier comme le MeDIP(-dep)-Seq peuvent être validées ou étudiées de façon plus précise grâce à des outils qui permettent de quantifier la méthylation à haute résolution comme le pyroséquençage. Cependant, ces technologies ne peuvent ni supporter l'analyse simultanée de tous les gènes identifiés ni gérer le multiplexage de nombreux échantillons.

En février 2010, une collaboration a été mise en place entre notre laboratoire et celui de Mats Nilsson à l'Université d'Uppsala (Suède) afin de développer un outil permettant d'étudier la méthylation sur un grand nombre de loci d'intérêt et de façon multiplexée. Le protocole de capture spécifique des régions génomiques *via* l'utilisation des sélectors est connu et maitrisé. L'enjeu est donc de combiner cette méthode à l'utilisation du traitement par le bisulphite puis de l'appliquer au séquençage haut-débit afin de quantifier la méthylation sur ces loci.

Chapitre II: Matériel et Méthodes

Nous nous attacherons dans cette partie à décrire les protocoles qui ont été suivis, les appareils qui ont été utilisés et les méthodes que nous avons employées. Il faut donc voir ici une « boite à outils » détaillée dans laquelle nous puiserons les éléments nécessaires et sur laquelle s'appuiera le développement des protocoles dont les enjeux ont été présentés précédemment, démarche que nous expliciterons dans les chapitres III à VI. Les références et fournisseurs respectifs des réactifs, kits et appareils cités se trouvent en Annexe 1.

II.1 Matériel biologique

II.1.1 ADN génomique commercial

Pour la mise au point de tous nos protocoles, nous utiliserons de l'ADN génomique commercialisé par Promega. L'ADN génomique de souris (Mouse Genomic DNA) a été isolé à partir de sang de souris saines. L'ADN génomique humain est issu de plusieurs donneurs anonymes de sexe masculin ou féminin (Human Genomic DNA male, ou female) ou mixtes (Human genomic DNA).

II.1.2 ADN Cot-1

Des expériences de dénaturation d'ADN double brin puis de réhybridation menées en conditions ménagées de refroidissement permettent d'obtenir une « courbe de Cot » : on mesure le pourcentage d'ADN simple brin restant dans le milieu et on trace son évolution en fonction du produit de la concentration initiale (C_0) par le temps (t) en échelle logarithmique ($\log(C_0t)$ ou log(Cot)). Dans le cas d'organismes pluricellulaires, cette courbe se divise en 3 phases dont les 2 premières, ayant le log(Cot) le plus bas, représentent les séquences répétées du génome. Le Cot-1 est un ADN double brin pour lequel cette valeur est de 1, il est enrichi en séquences répétées de diverses familles.

Nous utiliserons des ADNs Cot-1 murin et humain commerciaux (Mouse Cot-1 DNA et Human Cot-1 DNA, Life Technologies ; Mouse Hybloc et Human Hybloc, Applied Genetics Laboratories).

II.1.3 Individus HapMap

Le projet international HapMap (Haplotype Map) a permis l'analyse d'ADNs provenant de différentes populations dans le but d'identifier la plupart des haplotypes existants et de décrire les variations génétiques humaines. Nous disposons de 2 échantillons ayant contribué à cette étude : l'individu NA18506 de la population YRI (Yoruba d'Ibadan, Nigeria) et l'individu NA12802 de la population CEU (résidents de l'Utah originaires de l'Europe du Nord et de l'Ouest, échantillons collectés par le CEPH, Centre d'Etude du Polymorphisme Humain).

II.1.4 Echantillons de placentas humains

Nous disposons d'ADNs issus d'une quarantaine de placentas humains sans aberration pathologique. Nous travaillerons sur une douzaine d'entre eux listés dans le Tableau 5, notamment pour des expériences de contrôle qualité.

Identifiant échantillon placenta	Concentration (ng/µL)				
1	1154,29	5	653,72	9	635,76
2	1143,64	6	538,98	10	606,73
3	642,26	7	551,43	11	494,00
4	431,74	8	552,47	12	471,04

Tableau 5: Liste des échantillons d'ADN issus de placentas à disposition

II.1.5 Lignées cellulaires de MEFs

Des fibroblastes embryonnaires de souris (*Mouse Embryonic Fibroblasts*, MEFs) ont été étudiés par diverses techniques dans le laboratoire et constituent un choix idéal pour des études comparatives. Nous nous intéresserons à 4 lignées fournies par un collaborateur (Dr Chang Zhang, INSERM Lyon) dont 2 lignées sauvages (WT, wild type) et 2 lignées mutées pour le gène *Men*1. La culture des lignées et l'extraction ont été réalisées en externe et nous disposons du matériel décrit dans le Tableau 6.

Individu	Portée	Statut	Sexe	Aliquot	Code-barre identifiant	Quantité disponible (µg)
6.3	A	WT	M	1	B00DR13	31,3
				2	B00DR14	10,0
				3	B00DR15	23,6
				4	D000G15	97,1
8.2	B	WT	F	1	B00DR16	1,8
				2	B00DR17	12,1
				3	D000G16	97,7
8.3	B	MUT	M	1	B00DR18	9,1
				2	B00DR19	48,7
				3	B00DR1A	49,3
				4	B00DR1B	30,4
				5	D000G17	94,0
6.2	A	MUT	F	1	B00DR11	24,5
				2	B00DR12	3,1
				3	D000G18	92,1

WT: Wild Type (sauvage) M: Male
MUT: Mutée F: Femelle

Tableau 6: Liste des échantillons de MEFs à disposition

II.2 Amplification du matériel biologique

II.2.1 Obtention d'un amplicon par PCR

L'amplification est réalisée sur un thermocycleur (Master cycler gradient, Eppendorf) par la HotStar *Taq* polymérase (Qiagen), une forme modifiée d'une *Taq* polymérase classique, dont l'activité est nulle à température ambiante, ce qui réduit considérablement les hybridations non-spécifiques ou l'apparition de dimères d'amorces pouvant être formés à basse température pendant la PCR. Elle

nécessite donc une étape d'activation par chauffage avant toute réaction. Le protocole d'amplification utilisé est présenté dans le Tableau 7.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de PCR
10x PCR Buffer	2,5	1x	
$MgCl_2$ (25 mM)	1,6	1,6 mM	95°C 15 min
dNTPs (8 mM)	1,25	0,4 mM	
Amorce sens (10 µM)	0,5	0,2 µM	95°C 30 sec ┐
Amorce anti-sens (10 µM)	0,5	0,2 µM	$T_{hybridation}$ 30 sec ├ × 50
HotStar Taq polymerase (5 U/µL)	0,4	2 U	72 °C 15 sec ┘
Eau milliQ	17,25		
ADN	1,0		72 °C 5 min
Volume total	25		

Tableau 7: Conditions expérimentales de la PCR par la HotStar Taq polymérase

La température d'hybridation optimale pour un couple d'amorces donné est identifiée par gradient de température sur de l'ADN Promega puis dépôt et migration des produits sur un gel d'agarose à 2%.

On utilisera également la Taq polymérase Platinum High Fidelity (Life Technologies) dans les conditions présentées dans le Tableau 8.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de PCR
High Fidelity buffer (10x)	2,0	1x	
$MgSO_4$ (50 mM)	0,6	1,5 mM	95°C 3 min
dNTPs (25 mM)	0,16	0,2 mM	
Amorce sens (10 µM)	0,2	0,1 µM	95°C 30 sec ┐
Amorce anti-sens (10 µM)	0,2	0,1 µM	50°C 30 sec ├ × 15
Platinum Taq DNA polymerase High Fidelity (5 U/µL)	0,12	0,6 U	72°C 45 sec ┘
ADN	2,0		
Eau milliQ	14,72		72°C 5 min
Volume total	20,0		

Tableau 8: Conditions expérimentales de la PCR par la Platinum Taq polymérase

II.2.2 Amplification de tout le génome

II.2.2.1 Amplification du génome par WGA

10 µL d'ADN (contenant environ 10 ng) sont utilisés pour l'amplification aléatoire du génome avec le kit GenomePlex Complete Whole Genome Amplification (WGA, Sigma-Aldrich) selon les recommandations du fournisseur. L'ADN est d'abord fragmenté puis une librairie est créée et amplifiée par une polymérase grâce à des amorces universelles, avant purification.

II.2.2.2 Amplification du génome par MDA

La MDA (Multiple Displacement Amplification) est une méthode d'amplification qui repose sur l'amplification isotherme par l'ADN polymérase du bactériophage phi29 (Thermo Scientific), en présence d'amorces aléatoires dans les conditions présentées dans le Tableau 9.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de températures
phi29 reaction buffer (10x)	1,0	1x	
dNTPs (25 mM)	0,3	0,37 mM	Amplification
BSA (100 ng/µL)	0,2	1 ng/µL	30°C 90 min
Hexamère aléatoire (100 µM)	1,0	5 µM	
phi29 DNA polymerase (10 U/µL)	2,0	20 U	Inactivation
ADN	10,0		65°C 15 min
Eau milliQ	5,5		
Volume total	20,0		

Tableau 9: Conditions expérimentales de l'amplification par MDA

La réaction est initiée par l'hybridation d'une amorce hexamère aléatoire sur un ADN simple brin puis la polymérase synthétise le brin complémentaire. Lorsqu'elle arrive au niveau d'un autre hexamère qui a lui aussi initié la synthèse d'un brin, elle déplace celui-ci pour poursuivre son amplification. On obtient ainsi un réseau d'ADN composé de multiples branches.

II.2.2.3 Amplification du bisulfitome par WBA

L'ADN génomique converti par traitement au bisulphite de sodium peut également être amplifié (Whole Bisulfitome Amplification, WBA), grâce au kit EpiTect Whole bisulfitome (Qiagen), par la technologie de MDA, selon les recommandations du fournisseur.

II.3 Contrôles qualitatifs et quantitatifs

II.3.1 Quantification d'ADN

II.3.1.1 Par spectrophotométrie : le NanoDrop

Le NanoDrop ND-1000 (Thermo Scientific) est un spectrophotomètre qui ne nécessite pas l'utilisation de cuvette. 1 µL de l'échantillon est déposé directement à l'extrémité d'un câble de fibre optique. Lorsque celui-ci est mis en contact avec un second câble par fermeture de l'appareil, l'échantillon est maintenu par une tension de surface et traversé par un signal lumineux. Il est possible de mesurer des ADNs double comme simple brins. L'un des inconvénients majeurs de ce type de mesure réside dans le fait que toute molécule parasite qui absorbe la lumière (protéines, poussières, …) est prise en compte dans la mesure de la concentration.

II.3.1.2 Par fluorimétrie

Le Qubit (Life Technologies) est un appareil de mesure de concentrations qui utilise des molécules intercalantes fluorescentes. Celles-ci émettent un signal uniquement lorsqu'elles se lient à leur molécule cible d'ADN. Il existe un kit de mesure pour l'ADN double brin (Quant-iT dsDNA HS Assay kit, mesure de 0,2 à 100 ng) et un kit pour l'ADN simple brin (Quant-iT ssDNA HS Assay kit, mesure de 1 à 200 ng).

1 à 20 µL de l'échantillon peuvent être utilisés en complétant (qsp 200 µL) avec une solution préparée avec 1 µL de réactif fluorescent et 199 µL du tampon fourni. Après 2 minutes d'incubation, le tube est placé dans la cavité de l'appareil et la mesure est instantanée.

De la même façon, il est possible d'utiliser le kit Quant-iT dsDNA Assay (Broad Range, pour l'ADN double brin) pour effectuer une mesure de concentration en plaque 96 puits dans le lecteur de fluorescence Spectra Max Gemini XPS (Molecular Devices).

II.3.1.3 Par qPCR

La PCR en temps réel ou PCR quantitative (qPCR) est réalisée sur le Rotor-Gene Q (Qiagen). Des tubes 0,2 mL sont placés dans l'appareil et subissent une rotation à 400 rpm dans une chambre où de l'air circule afin de maintenir la température uniforme.

A) *Pour 1 échantillon*	Volume (µL)	Concentration ou quantité finale	Programme de qPCR
SYBR Green PCR Master Mix	12,5	1x	
Amorce sens (10 µM)	1,0	0,4 µM	95°C 10 min
Amorce anti-sens (10 µM)	1,0	0,4 µM	
Eau milliQ	8,5		95°C 15 sec 60°C 60 sec } × 50
ADN	2,0		
Volume total	25,0		

B) *Pour 1 échantillon*	Volume (µL)	Concentration ou quantité finale	Programme de qPCR
5x HOT FIREPol EvaGreen qPCR Mix	4,0	1x	
Amorce sens (10 µM)	1,0	0,5 µM	95°C 15 min
Amorce anti-sens (10 µM)	1,0	0,5 µM	
Eau milliQ	13,0		95°C 30 sec 60°C 60 sec } × 50
ADN	1,0		
Volume total	20,0		

Tableau 10: Conditions expérimentales de la qPCR

Un agent fluorescent est utilisé et va s'intercaler dans l'ADN double brin nouvellement formé lors de la PCR. La fluorescence alors émise est proportionnelle à la quantité d'ADN présente dans le milieu. Elle franchit une valeur seuil fixée par l'appareil à un cycle de la PCR appelé Ct (cycle threshold ou cycle seuil) qui permet ainsi le calcul de la quantité initiale d'ADN utilisée. On utilisera la SYBR Green (Applied Biosystems) comme agent intercalant fluorescent, dont le protocole de préparation figure

dans le Tableau 10A. Il est aussi possible de suivre le protocole proposé par Qiagen avec le QuantiFast SYBR Green PCR kit. On pourra également utiliser l'EvaGreen (Solis BioDyne, voir Tableau 10B) dont l'avantage est sa non-toxicité pour l'enzyme de PCR, ce qui permet de l'utiliser en quantité saturante pour l'ADN.

Toutes les qPCRs sont effectuées en duplicat. L'acquisition des valeurs de fluorescence est effectuée sur le canal vert du Rotor-Gene et on pourra les exploiter en quantification absolue ou bien en quantification relative lorsqu'il s'agit de comparer 2 échantillons.

II.3.2 Contrôle-qualité : le Bioanalyzer 2100

Le Bioanalyzer 2100 (Agilent) est une plateforme qui permet de séparer des échantillons par électrophorèse en vue de leur contrôle-qualité. 1 µL seulement est analysé dans un puits d'une puce Agilent au fond de laquelle a été déposé un gel contenant une molécule fluorescente qui va s'intercaler entre les brins d'ADN. La préparation de la puce suit les indications données par le fournisseur. 12 échantillons peuvent être analysés sur une puce du DNA 1000 kit (pour des tailles comprises entre 25 et 1000 pb et une gamme de concentrations de 100 pg/µL à 50 ng/µL) et 11 sur une puce du High-Sensitivity kit (tailles de 50 à 7000 pb, concentrations de 5 à 500 pg/µL).

La fluorescence émise est traduite en une image virtuelle de gel sous forme de bandes et en un électrophérogramme sous forme de pics pour identifier les tailles de fragments contenus dans l'échantillon, grâce à une comparaison avec un marqueur de tailles. Il est également possible de quantifier chaque pic de façon indépendante.

II.4 Concentration et purification d'échantillons

II.4.1 Concentration au speedvac

Le speedvac (Savant SC110A Speedvac plus centrifugal vacuum concentrator, Thermo Scientific) est un appareil permettant l'évaporation de tout type de solvant grâce au vide créé pendant la centrifugation des échantillons, avec ou sans chauffage. Ceux-ci sont contenus dans des tubes 1,5 mL ou des plaques 96 puits et peuvent être concentrés ou séchés complètement pour être repris dans le volume désiré.

II.4.2 Purification sur colonnes

Divers types de colonnes sont commercialisés pour la purification, le dessalement et la concentration d'échantillons d'ADN. Nous utiliserons les colonnes du QIAquick PCR purification kit (volume minimum d'élution de 30 µL, fragments élués d'une taille comprise entre 100 pb et 10 kb) et du

MinElute PCR purification kit (volume de 10 µL, tailles de 70 pb à 4 kb) (Qiagen), du GenElute PCR clean-up kit (volume de 50 µL, tailles de 100 pb à 10 kb), les colonnes Microcon Ultracel YM-100, YM-50 et YM-3 (Millipore) et illustra MicroSpin G-50 (GE Healthcare), en suivant les instructions données par les fournisseurs respectifs.

II.4.3 Gel-filtration

La gel-filtration consiste en la purification de molécules grâce à une séparation par leurs tailles. Elles traversent un gel formé de billes dont le diamètre est connu et contenant des pores dans lesquels les petites molécules passent, ralentissant ainsi leur sortie du gel. Les grosses molécules passent entre les billes et quittent le gel les premières.

100 g de poudre (Sephadex G50 Superfine, GE Healhcare ou Bio-gel P100, Bio-Rad) sont hydratés dans 2L d'eau milliQ à +4°C pendant 8 heures puis le volume de liquide en excès est évacué pour obtenir la consistance d'un gel. 400 µL sont déposés dans un puits d'une plaque 96 puits MultiScreen (Millipore) soutenue par une plaque MicroAmp (Millipore). Après centrifugation à 1500 rpm pendant 3 minutes, cette dernière est remplacée par une plaque 96 puits dans laquelle sera récupéré l'échantillon purifié. Le volume d'échantillon est déposé au centre du gel puis il est purifié par centrifugation pendant 1 minute.

La limite d'exclusion de la Sephadex G50 est de 20 bases soit une masse moléculaire de 6,6 kDa (la masse moléculaire d'une base étant de 330 g/mol ou Da). Celle du Bio-gel P100 est de 5 à 100 kDa.

II.4.4 Purification sur billes AMPure

Les billes Agencourt AMPure XP beads (Beckman Coulter) sont utilisées pour purifier des échantillons d'ADN simple ou double brin de plus de 100 pb, notamment des produits de PCR, sans avoir recours à des étapes de centrifugation. Cette technique fait appel à une immobilisation réversible de l'ADN sur phase solide ou SPRI (Solid-Phase Reversible Immobilization) : les fragments sont retenus sur des billes magnétiques puis élués.

100 µL d'échantillon (ou un volume V) sont placés dans un tube de 1,5 mL puis 180 µL de billes (ou un volume de 1,8 × V) remises en suspension par agitation y sont ajoutés et mélangés par aspiration à la pipette. Après 5 minutes d'incubation à TA, les fragments de plus de 100 pb sont liés aux billes. Le tube est placé sur un aimant pendant 2 minutes puis le surnageant est éliminé. Les billes sont lavées 3 fois avec 300 µL d'EtOH 70% fraichement préparé puis les traces d'éthanol sont éliminées par séchage pendant 2 minutes à TA. Les billes sont alors remises en suspension par pipetage dans 30 µL de tampon d'élution (EB, Qiagen). Après 2 minutes d'aimantation, le surnageant contenant le produit purifié est isolé.

II.4.5 Purification automatisée sur billes IPure

L'Auto IPure kit (Diagenode) permet de purifier et concentrer des échantillons directement après leur immunoprécipitation sur le robot du même fournisseur (SX-8G IP Star, voir II.6 Immunoprécipitation de l'ADN méthylé). Il utilise une purification sur billes dans une plaque 96 puits dont chaque demi-ligne est réservée à un échantillon et dans laquelle les réactifs sont distribués comme indiqué dans le Tableau 11 avant d'insérer la plaque dans le robot. A la fin du *run*, les 50 µL purifiés sont récupérés dans les colonnes 6 et/ou 12.

Colonnes n°	Tampons / réactifs	Volume (µL)
1	Buffer C	150
2 et 8	Isopropanol + billes + carrier + échantillon	100 + 15 + 2 + 100
3 et 9	Wash Buffer 1	100
4 et 10	Wash Buffer 2	100
5 et 11	*Elution*	/
6 et 12	*Elution*	/
7	/	/

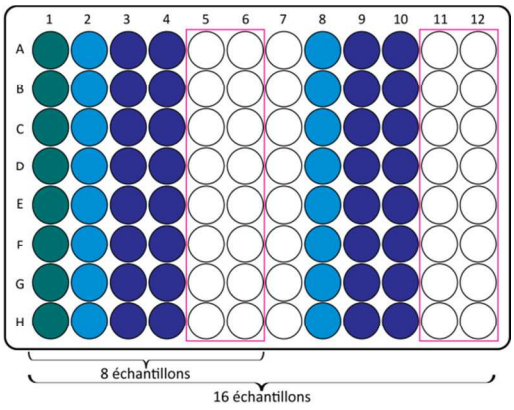


Tableau 11: Distribution des réactifs pour la purification automatisée IPure

II.5 Fragmentation de l'ADN

II.5.1 Fragmentation mécanique

Les sonicateurs utilisent des ondes d'énergie acoustique pour créer une énergie mécanique capable de fragmenter l'ADN. La longueur d'onde des ultrasons utilisés est de l'ordre du millimètre, ce qui permet un bon contrôle de la forme de l'onde qui peut agir de façon précise sur l'échantillon. Les ultrasons provoquent des contractions et expansions successives du liquide dans lequel est dilué l'échantillon, aboutissant à la formation de bulles dont l'implosion suffit à fragmenter l'ADN.

II.5.1.1 Le sonicateur Bioruptor, Diagenode

La technologie ACT (*Adaptative Cavitation Technique*) utilisée dans le Bioruptor UCD-200 permet aux ultrasons de se déplacer dans un bain d'eau et de traverser les échantillons placés dans des tubes en rotation dans ce bain. La sonication est réalisée à 4°C.

Les échantillons sont dilués à une concentration de 25 ng/µL dans du TE (Tris EDTA) et 100 µL sont placés dans un tube 1,5 mL. Les tubes sont positionnés sur le support motorisé (contenant 6 tubes ; compléter avec des tubes contenant de l'eau si moins de 6 échantillons). Les échantillons sont soniqués pendant un temps pouvant varier de quelques secondes à une vingtaine de minutes en

alternant des cycles de marche et d'arrêt, en fonction de la taille désirée. 5 µL sont ensuite chargés sur un gel d'agarose à 1% pour vérifier la taille des fragments obtenus.

II.5.1.2 Le sonicateur E210, Covaris

La technologie AFA (*Adaptative Focused Acoustics*) de l'appareil Covaris permet à l'onde de converger de façon très localisée et reproductible sur le tube contenant l'échantillon.

Les échantillons sont dilués en TE (Illumina) et 100 µL maximum sont placés dans un tube de 100 µL (MicroTube Snap-Cap, Covaris). Les tubes sont positionnés sur le support en plaque (jusqu'à 24 tubes sur le TR246 Snap-Cap, Covaris, en évitant au maximum les bords du support). Les échantillons sont soniqués dans un bain d'eau à 6°C ayant subi un dégazage préalable de 45 minutes, en utilisant des paramètres déterminant la puissance (DC : Duty Cycle, I : Intensity, CPB : Cycles Per Bust) et un temps qui seront définis en fonction de la taille de fragments désirée. 5 µL sont ensuite chargés sur un gel d'agarose à 1% pour vérifier la taille des fragments obtenus.

II.5.2 Digestion enzymatique

Cette fragmentation est utilisée dans le cadre du ciblage de loci par les sélectors. L'échantillon à fragmenter est dilué à une concentration de 50 ng/µL dans un volume total de 15 µL contenant de la BSA 2x (BSA 100x, New England Biolabs (NEB)) puis il est divisé en 3 : 5 µL sont distribués dans 3 puits d'une plaque 96 puits. Chaque aliquot est digéré par ajout de 5 µL d'un mix qui lui est propre contenant 2 enzymes (2,5 U d'enzymes par mix, soit 1,25 U de chacune d'entre elles) dans un tampon adapté aux 2 enzymes, dans notre cas le NEBuffer 4 (NEB). Le choix des enzymes sera régi par les sites que l'on souhaite obtenir aux extrémités de nos fragments.

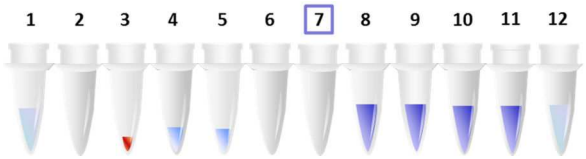
La digestion enzymatique est réalisée par incubation à 37°C pendant 1 heure puis les enzymes sont inactivées par incubation à 80°C pendant 20 minutes. Enfin, les 3 volumes de réaction sont réunis pour obtenir un volume final de 30 µL.

II.6 Immunoprécipitation de l'ADN méthylé

L'immunoprécipitation (IP) de l'ADN méthylé (MeDIP) est automatisée sur un robot (SX-8G IP Star, Diagenode) à l'aide d'un kit contenant tous les réactifs nécessaires (Auto-MeDIP kit, Diagenode).

II.6.1 Préparation des barrettes contenant les tampons

Les barrettes de 12 tubes de 0,2 mL sont préparées sur un support réfrigérant en distribuant les tampons et réactifs tel qu'indiqué dans le Tableau 12.



Tubes n°	Tampons / réactifs	Volume (µL)
1	DIB (DNA Isolation Buffer)	92,5
2	/	/
3	Billes magnétiques	10
4	MagBuffer A 1x	50
5	MagBuffer A 1x	50
6	/	/
7	Mix d'échantillon (voir ci-après)	
8	MagWash buffer 1	100
9	MagWash buffer 1	100
10	MagWash buffer 1	100
11	MagWash buffer 2	100
12	DIB	100

Tableau 12: Distribution des réactifs dans la barrette de tubes pour le MeDIP

II.6.2 Préparation de l'échantillon

L'échantillon est d'abord dilué à 25 ng/µL en TE. 1,0 µg d'ADN est ensuite utilisé pour l'immunoprécipitation et 10% supplémentaires ne subissent pas la sélection : ils seront utilisés en tant que contrôle (*input*, IN). Le mix indiqué dans le Tableau 13 est préparé dans un tube de 0,2 mL puis l'ADN est dénaturé par incubation à 95°C pendant 5 minutes et refroidi immédiatement sur glace.

Tampons / réactifs	Volume pour 1 IP	Volume pour 1 IN	Volume total (µL)
MagBuffer A 5x	20,0	2,0	22,0
MagBuffer B	5,0	0,5	5,5
ADN méthylé contrôle	1,3	0,1	1,4
ADN non-méthylé contrôle	1,3	0,1	1,4
ADN 25 ng/µL	40,0	4,0	44,0
Eau milliQ	7,5	0,8	8,3
Total	75,0	7,5	82,5

Tableau 13: Préparation du mix contenant l'échantillon pour le MeDIP

Les ADNs méthylé et non-méthylé contrôles sont des séquences du génome d'*Arabidopsis thaliana* fournies dans le kit. 75 µL du mix sont déposés dans le puits n°7 de la barrette et le volume restant est conservé à +4°C.

II.6.3 Dilution de l'anticorps anti-5-méthylcytidine

L'anticorps est dilué au demi puis 150 ng sont utilisés dans la préparation du mix présenté dans le Tableau 14.

Tampons / réactifs	Volume pour 1 IP (µL)
MagBuffer A 5x	0,6
Anticorps dilué au ½	0,3
Eau milliQ	2,1
MagBuffer C	2,0
Total	5,0

Tableau 14: Préparation du mix contenant l'anticorps pour le MeDIP

Les 5,0 µL de mix sont ajoutés à l'échantillon dans le puits n°7 de la barrette avant de compléter le volume à 100 µL en ajoutant 20 µL de MagBuffer A 1x.

II.6.4 Immunoprécipitation

Le robot IP-Star est préparé comme dans la Figure 16 : les barrettes de tubes précédemment préparées sont placées sur le bloc Peltier de droite dont la température a été préalablement réglée à 4°C.

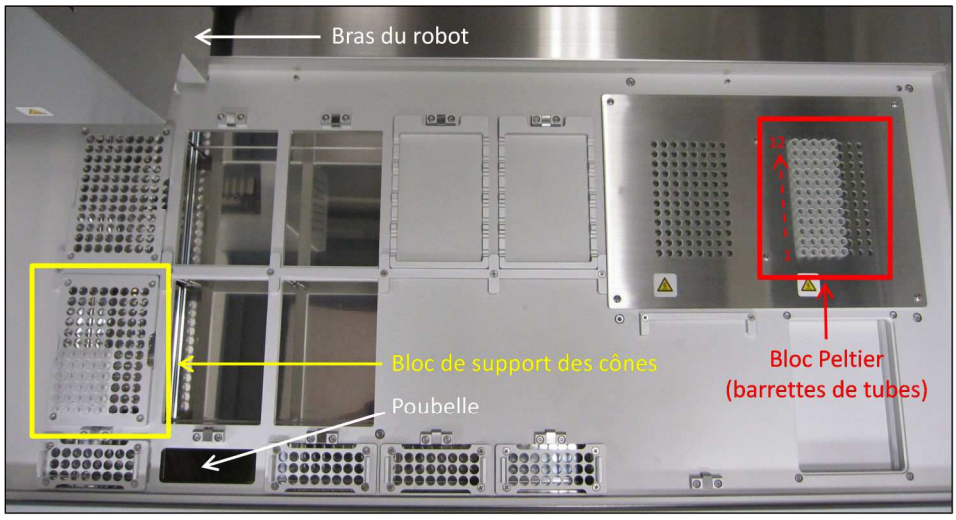


Figure 16: Robot IP-Star pour le MeDIP
Vue de dessus. Les barrettes de tubes contenant tous les réactifs sont placées dans le bloc Peltier de droite et le bloc de support du bas est chargé avec des cônes à filtres. Le MeDIP est réalisé en intégralité dans le robot. Le bras contient un bloc aimanté permettant de retenir les billes magnétiques. Les cônes sont éjectés par la trappe poubelle située en bas.

L'immunoprécipitation est lancée sur la nuit : les billes magnétiques à la surface desquelles se trouve la protéine G sont lavées, puis ajoutées au milieu contenant l'ADN et l'anticorps. L'immunosélection est effectuée pendant 16 heures par agitation du milieu à 4°C : les fragments méthylés (ADN IP) sont sélectionnés par l'anticorps, lui-même sélectionné par la protéine G. Les billes magnétiques sont

retenues sur un aimant permettant ainsi le lavage de l'IP (dans les puits 8 à 11) qui est enfin isolé dans le tampon DIB (*DNA Isolation Buffer*).

Le lendemain, les 7,5 µL de mix contenant l'échantillon et conservés au frais sont ajoutés dans le puits n°1 (IN). 1 µL de protéinase K est ensuite ajouté dans les puits n°1 (IN) et 12 (IP) : elle va digérer l'anticorps pour libérer l'IP. Les tubes des barrettes sont fermés et le robot incube pendant 1 heure à 55°C puis à 95°C pour inactiver l'enzyme.

A la fin de l'incubation, les barrettes de tubes sont centrifugées et placées sur un aimant (DiaMag02, Diagenode) pour récupérer l'IP situé dans le surnageant du puits n°12. L'IN situé dans le puits n°1 est également isolé et tous deux sont stockés à -20°C.

Lorsque le MeDIP est suivi d'une purification avec le Auto IPure kit, la distribution initiale des échantillons diffère légèrement (puits 1, 2 et 3 : laissés vides ; puits 4 et 12 : 50 µL d'IPure elution buffer ; puits 5 : 50 µL de MagBuffer A 1x et 10 µL de billes ; puits 6 : 50 µL de MagBuffer A 1x ; l'IP sera récupéré dans le puits n°4). La digestion par la protéinase K disparait pour faire place à une élution des fragments méthylés grâce à l'IPure elution buffer. Les 7,5 µL d'*input* ne sont plus ajoutés dans le puits n°1 mais dilués dans 92,5 µL d'IPure elution buffer et traités comme un échantillon à part entière dans le processus de purification.

II.6.5 Contrôle-qualité du MeDIP par qPCR

Les oligonucléotides méthylés et non-méthylés, fournis par Diagenode et ajoutés dans le mix de préparation de l'échantillon, servent à contrôler l'efficacité du MeDIP par qPCR (avec le SYBR Green, sur le Rotor-Gene) sur ces séquences, avec des amorces fournies dans le kit.

Le taux de récupération est, pour une séquence étudiée, son pourcentage dans l'IP par rapport à sa quantité dans l'IN. Il est idéalement de 100% si tous les fragments méthylés sont immunoprécipités et se calcule à l'aide de la formule suivante :

$$\text{taux de récupération} = 2^{[(Ct_{IN}-3.32)-Ct_{IP}]} \times 100$$

où Ct_{IN} et Ct_{IP} sont les cycles de seuil (*cycle threshold*) de l'*input* et de l'échantillon immunoprécipité respectivement. Le facteur de correction 3,32 est appliqué puisque seulement 10% de l'échantillon sont utilisés comme *input* ($2^{3,32} = 10$).

En comparant les résultats obtenus entre les oligonucléotides contrôles méthylés et non-méthylés, on déduit l'enrichissement en séquences méthylées avec la formule suivante :

$$\text{facteur d'enrichissement} = \frac{\text{taux de récupération de l'ADN méthylé}}{\text{taux de récupération de l'ADN non-méthylé}}$$

II.7 Traitement par le bisulphite

II.7.1 Protocole de la conversion sur colonne

Le traitement par le bisulphite de sodium permet la conversion des cytosines en uraciles tandis que les 5-méthylcytosines ne sont pas affectées. Ceci implique trois étapes : une sulphonation, une désamination hydrolytique et une désulphonation (185).

Nous utiliserons un kit Qiagen (EpiTect bisulfite kit 48) pour effectuer cette conversion, telle que préconisée par le fournisseur. Brièvement, 1 µg d'ADN (quantité optimale pour la conversion) est dilué ou séché au speedvac et repris dans 20 µL d'eau milliQ. Il est théoriquement possible d'utiliser de 1 ng à 2 µg. La réaction est réalisée dans le bisulphite mix par des cycles de dénaturation à 99°C (5minutes) et de conversion à 60°C (de 25 à 175 minutes) L'échantillon sous forme simple brin est ensuite déposé sur une colonne EpiTect et lavé avec différents tampons puis élué avec 20 µL de tampon d'élution.

Nous utiliserons également le MethylEasy Xceed kit (Human Genetic Signatures) qui offre la possibilité de bisulphiter peu de matériel (de 50 pg à 5 µg d'ADN, selon le fournisseur).

II.7.2 Protocole de la conversion sur billes

Les billes IPure fournies par Diagenode et destinées à la purification d'échantillons ont été récemment intégrées à un kit de traitement au bisulphite de 1 ng à 1 µg de matériel (MagBisulfite kit, voir recommandatins du fournisseur). Elles permettent de réaliser l'étape de désulphonation en remplacement des colonnes proposées dans les autres types de kits, ceci ayant notamment pour avantage de diminuer les pertes d'ADN.

II.7.3 Contrôle de la conversion

La quantité et l'intégrité de l'ADN bisulphité obtenu peuvent être vérifiées par qPCR en utilisant des amorces et une sonde Taqman décrites dans la littérature (186), ciblant une séquence *Alu* dépourvue de CpGs (voir séquences en Annexe 2). Le contrôle est ainsi indépendant du statut de méthylation de la région étudiée. La sonde Taqman est marquée à son extrémité 5' par un fluorophore FAM (6-carboxyfluoresceine) et à son extrémité 3' par un MGBNFQ (Molecular-Groove Binding Non-fluorescence Quencher), un fluorophore qui inhibe la fluorescence du premier lorsqu'ils sont à proximité. Au cours de la PCR, la sonde hybridée est hydrolysée par la polymérase et la séparation des deux fluorophores provoque l'émission d'un signal.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de qPCR
2x Rotor-Gene Probe PCR Master Mix	12,5	1x	
Sonde Taqman (10 pmol/µL)	0,5	0,2 mM	
Amorce sens (10 µM)	1,0	0,4 µM	95°C 3 min
Amorce anti-sens (10 µM)	1,0	0,4 µM	
Eau milliQ	8,0		95°C 3 sec
ADN	2,0		60°C 10 sec × 50
Volume total	25,0		

Tableau 15: Conditions expérimentales de qPCR du contrôle-qualité du traitement au bisulphite

Le protocole suivi sur le Rotor-Gene est présenté dans le Tableau 15. Une droite d'étalonnage est établie avec 5 à 20 ng d'ADN commercial méthylé et bisulphité (Epitect human control DNA, Qiagen) et la quantité d'ADN bisulphité dans notre échantillon en est déduite.

II.8 Biotinylation d'ADN

II.8.1 Biotinylation des extrémités 3'

Une terminal transférase recombinante (Roche ou NEB) peut incorporer aux extrémités 3' de fragments d'ADN un nucléotide biotinylé. La réaction s'effectue par incubation à 37°C (15 minutes pour l'enzyme Roche, 1h pour l'enzyme NEB) en présence de $CoCl_2$, du nucléotide d'intérêt (biotin-16-dUTP, NEB ou biotin-16-ddUTP, Roche), de l'enzyme et de TdT reaction buffer tel qu'indiqué par les fournisseurs respectifs. L'enzyme est ensuite inactivée par chauffage (10 minutes à 95°C pour l'enzyme Roche, 20 minutes à 75°C pour l'enzyme NEB). En parallèle un témoin négatif peut subir la réaction en supprimant l'enzyme dans le mix précédemment cité.

II.8.2 Biotinylation aléatoire par amplification linéaire

Le fragment de Klenow de l'ADN polymérase I peut être utilisé pour biotinyler l'ADN par extension du brin après hybridation d'amorces octamères aléatoires. Nous utilisons le kit BioPrime DNA labelling system (Life Technologies) dans lequel le nucléotide biotinylé incorporé est un biotin-14-dCTP.

Sur glace, 20 µL d'amorces aléatoires (à 750 ng/µL) sont ajoutés à 200 ng d'ADN (20 µL à 10 ng/µL) et le tout est dénaturé en chauffant à 95°C pendant 5 minutes puis immédiatement refroidi sur glace. 5 µL d'un mix de dNTPs contenant le biotin-14-dCTP puis 40 U d'enzyme sont ajoutés. L'amplification se fait par incubation à 37°C de une à plusieurs heures puis elle est arrêtée par ajout de 5 µL d'un tampon d'arrêt. Les échantillons sont ensuite purifiés sur colonnes QIAquick et peuvent être déposés sur gel d'agarose à 2% pour visualiser l'amplification. On estime que le produit obtenu peut se

conserver jusqu'à 3 mois à -20°C. 100 ng d'ADN contrôle (de sperme de saumon) peuvent subir la réaction en parallèle, ainsi qu'un témoin négatif en supprimant l'enzyme dans le mix.

II.9 Méthodes de capture de régions cibles

II.9.1 Hybridation à de l'ADN Cot-1

II.9.1.1 Liaison d'ADN Cot-1 biotinylé à des billes de streptavidine

L'ADN Cot-1 est biotinylé puis fixé sur des billes magnétiques recouvertes de streptavidine. Plusieurs types de billes peuvent être employés ; nous utiliserons les billes M-270 et M-280 streptavidin, MyOne Streptavidin C1 et T1 du Dynabeads streptavidin trial kit (Life Technologies).

1,0 mg de billes est d'abord lavé pour ôter toute trace de conservateur puis les billes sont remises en suspension dans un tampon adapté. Toutes les étapes de lavage se font par aimantation de 2 minutes (DynaMag-96 Side Skirted, Life Technologies), aspiration du surnageant et remise en suspension dans 50 µL de tampon par pipetage, tout ceci en plaque 96 puits. Elles sont ensuite incubées à température ambiante avec 50 µL d'ADN Cot-1 biotinylé sous agitation (700 rpm, sur l'agitateur Thermomixer Comfort avec l'adaptateur 96 microtubes PCR 0.2 mL, Eppendorf). Elles subissent alors à nouveau des étapes de lavage entrecoupées par une étape de dénaturation de l'ADN Cot-1 qui y est fixé. Après le dernier lavage, les billes sont remises en suspension dans un tampon d'hybridation.

Le tampon B&W 4x (binding and washing buffer) est utilisé pour le lavage des billes et est préparé par dissolution de Tris-HCl (10 mM), d'EDTA (10 mM) de NaCl (2 M) et de Tween-20 (0,1%) et ajustement du pH à 7,6. Il peut également être utilisé dilué au demi (B&W 2x). Le tampon d'hybridation (HB, hybridization buffer) est préparé par dilution de SSC (Saline Sodium Citrate) afin de l'obtenir à une concentration de 6x (NaCl 0,9 M et citrate de sodium 0,09 M).

II.9.1.2 Hybridation de l'échantillon à l'ADN Cot-1

L'échantillon est ajouté au milieu HB précédent pour obtenir un volume total de 100 µL. Il est hybridé à l'ADN Cot-1 par incubation à une température fixe (qui devra être déterminée), sous agitation pendant quelques heures (ces paramètres seront optimisés par la suite) dans un tube 0,2 mL. Le tube est ensuite placé sur un support aimanté (DiaMag02, Diagenode) pendant 2 minutes et le surnageant contenant l'échantillon (dont la quantité de séquences répétées s'en trouve diminuée) est isolé et purifié.

II.9.2 Capture de loci par hybridation aux sélectors

Des sondes appelées sélectors, complémentaires des régions à capturer, sont biotinylées en 3'. Elles sont ensuite hybridées à l'échantillon, préalablement fragmenté par digestion enzymatique, grâce au mix et incubations présentés dans leTableau 16.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de températures
Tampon B&W 4x	28,0	0,7x	95°C 10 min
Formamide 100%	36,8	23%	75°C 30 min
Mix de 98 sélectors biotinylés (1 nM)	1,6	0,01 nM	68°C 30 mir
Mix d'ADN fragmenté (25 ng/µL)	30,0	15 ng/µL	62°C 30 mir
Vecteur (1 µM)	0,47	2,94 nM	55°C 30 mir
Eau milliQ	63,13		46°C 10h
Volume total	160,0		10°C stockage

Tableau 16: Conditions expérimentales pour l'hybridation aux sélectors

Le tampon B&W 4x est préparé par dissolution de Tris-HCl (40 mM), d'EDTA (20 mM), de NaCl (4 M) et de Tween-20 (0,4%) et ajustement du pH à 7,5. Le vecteur trouve un intérêt lorsqu'une étape de PCR est réalisée à la fin du protocole (voir explications dans la Figure 58).

Les cibles sont ensuite liées à 10 µL de billes magnétiques de streptavidine (Dynabeads M-280 Streptavidin, Life Technologies) diluées dans 10 µL de tampon B&W 4x et 20 µL d'eau milliQ, grâce aux sélectors biotinylés qui y sont hybridés, par incubation de 10 minutes à TA. Le surnageant est ensuite éliminé par aimantation et pipetage puis un lavage est effectué par ajout de 200 µL du tampon B&W 1x contenant 20% de formamide et incubation pendant 30 minutes à 46°C sous agitation (12 rpm). Le surnageant est alors éliminé par aimantation et pipetage et les billes sont resuspendues dans le mix du Tableau 17 pour la ligation.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale
Ampligase reaction buffer (10x)	5,0	1x
BSA (10 mg/mL)	1,0	10 µg
Ampligase thermostable DNA ligase (5 U/µL)	2,5	12,5 U
Eau milliQ	41,50	
Volume total	50,0	

Tableau 17: Conditions expérimentales de la ligation pour la circularisation des cibles

La ligase (Epicentre Biotechnology) permet la circularisation des cibles et de leur vecteur par incubation de 10 minutes à 55°C. Le surnageant est éliminé par aimantation et pipetage et les billes portant les cibles sont resuspendues dans un tampon adapté à la suite du protocole. Les molécules

circularisées sont enfin isolées dans le surnageant après dénaturation thermique (10 minutes à 95°C) et aimantation.

II.10 Techniques de séquençage

II.10.1 Le pyroséquenceur

Le pyroséquençage est une technologie quantitative en temps réel permettant l'analyse de régions d'intérêt de 300 pb environ. L'échantillon est d'abord traité au bisulphite de sodium puis la séquence d'intérêt est amplifiée par PCR. L'amplicon est ensuite rendu simple brin pour hybrider l'amorce de pyroséquençage.

II.10.1.1 Conceptions d'amorces de PCR et de pyroséquençage

La séquence d'ADN génomique pour laquelle on souhaite concevoir des amorces de PCR peut être obtenue grâce à l'outil bioinformatique de l'UCSC (http://genome.ucsc.edu/) (187) en fournissant le nom du gène d'intérêt ou les coordonnées chromosomiques de la région recherchée. Nous utiliserons les assemblages hg18 et hg19 pour le génome humain et mm8 et mm9 pour celui de la souris. On peut également rechercher un gène ou ses coordonnées si l'on possède sa séquence en réalisant un Blat.

La séquence génomique est ensuite copiée dans le programme MethPrimer (188) (accessible sur http://www.urogene.org/methprimer/index.html). Celui-ci fournit les séquences converties après traitement au bisulphite ainsi que plusieurs couples d'amorces correspondant à la région d'intérêt bisulphitée. Il est possible d'affiner sa recherche en précisant la taille idéale du produit de PCR ainsi que celle des amorces ou encore le nombre de CpGs désirés dans l'amplicon.

On réalise ensuite un Blat de la séquence génomique correspondante sur le site de l'UCSC afin d'identifier les potentiels SNPs présents dans l'amplicon et de vérifier leur absence dans les amorces. Les SNPs peuvent être donnés sur le brin complémentaire de la séquence d'origine. Il faudra alors utiliser la séquence inverse complémentaire (fournie par exemple par Sequence Massager disponible sur http://www.attotron.com/cybertory/analysis/seqMassager.htm) afin de le positionner sur notre séquence.

Enfin, les amorces de pyroséquençage correspondantes aux amplicons sont conçues avec un programme fourni par Qiagen. Il fournit un score à chaque amorce qu'il propose, basé sur son risque d'hybridation non-spécifique, de formation de dimères ou encore de formation de structures en épingles. La température d'hybridation de l'amorce doit également être prise en compte et la plus basse possible car le pyroséquenceur fonctionne à 28°C. On ne sélectionnera pas d'amorce dont le

score est inférieur à 80%, tout comme on évitera de concevoir une amorce de pyroséquençage complémentaire à l'une des amorces de PCR ou encore contenant un SNP. Si l'amorce de pyroséquençage est complémentaire au brin sens (respectivement anti-sens), la PCR sera réalisée avec une amorce anti-sens (resp. sens) biotinylée.

II.10.1.2 Préparation de l'échantillon

Après traitement par le bisulphite, l'échantillon est amplifié par PCR avec la polymérase HotStar Taq. Une des amorces de PCR est biotinylée, ceci permet de retenir le brin correspondant sur un filtre, *via* des billes recouvertes de streptavidine (Streptavidin Sepharose high performance, GE Healthcare) et de purifier le milieu pour éliminer le brin complémentaire. L'amorce de pyroséquençage est enfin hybridée par chauffage à 80°C pendant 2 minutes. Tout ceci est réalisé sur une table de purification (PyroMark Q96 vacuum prep workstation, Qiagen) et a fait l'objet d'une description détaillée dans la littérature (189,190).

II.10.1.3 Pyroséquençage

L'analyse est effectuée sur le système PSQ 96MD avec le kit PyroGold SQA Reagent (Qiagen) et l'exploitation des résultats est réalisée avec le logiciel Q-CpG (V.1.0.9, Qiagen). La séquence à étudier est fournie à l'appareil afin qu'il dispense les nucléotides correspondants. Au niveau de la cytosine d'un dinucléotide CpG, un C et un T sont dispensés : ils correspondront respectivement aux allèles méthylés (non-affectés par le traitement par le bisulphite) et non-méthylés (affectés par ce même traitement). L'incorporation d'une base provoque une cascade enzymatique qui aboutit à l'émission d'un signal lumineux (voir principe du pyroséquençage sur la Figure 6) dont l'intensité est directement proportionnelle à la quantité de nucléotides incorporés. Ainsi, la proportion des allèles méthylés et non-méthylés sur une position CpG peut être directement déduite de la hauteur des pics obtenus et permet ainsi d'en déduire le pourcentage de méthylation. Cette technique est donc quantitative, sa résolution est de 5%.

II.10.2 Séquenceur haut-débit Illumina

II.10.2.1 Ligation des adaptateurs

La ligation des adaptateurs nécessaires au séquençage se divise en 3 parties : la réparation des extrémités, leur adénylation et la fixation des adaptateurs. Plusieurs kits contenant tous les réactifs nécessaires à ces étapes existent, notamment le Paired-End sample prep kit (Illumina, voir recommandations du fournisseur) et le Next DNA sample prep master mix Set 1 (NEB), tous deux

permettant de travailler sur des quantités d'ADN fragmenté de 1 à 5 µg. Nous détaillons ici le protocole impliquant le kit de NEB.

II.10.2.1.1 Réparation des extrémités

75 µL d'ADN préalablement fragmenté sont préparés dans le mix présenté dans le Tableau 18.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale
NEBNext End Repair Reaction Buffer (10x)	10,0	1x
NEBNext End Repair Enzyme Mix	5,0	
[T4 Polynucleotide Kinase (10 U/µL)		0,5 U
T4 DNA Polymerase (3 U/µL)]		0,15 U
Eau milliQ	10,0	
ADN fragmenté (1 - 5 µg)	75,0	
Volume total	100,0	

Tableau 18: Préparation du mix pour la réparation des extrémités

Après 30 minutes d'incubation à 20°C (dans le Thermomixer comfort, Eppendorf) les 100 µL sont purifiés sur colonne QIAquick (élution dans 38 µL d'EB).

II.10.2.1.2 Adénylation des extrémités

Les 37 µL d'ADN résultant de l'étape précédente (perte de volume d'1 µL constatée dans la colonne) sont préparés dans le mix présenté dans le Tableau 19 :

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale
NEBNext dA-Tailing Reaction Buffer (10x)	5,0	1x
Klenow fragment (3'→5' exo$^-$)	3,0	
Eau milliQ	5,0	
ADN	37,0	
Volume total	50,0	

Tableau 19: Préparation du mix pour l'adénylation des extrémités

Après 30 minutes d'incubation à 37°C, les 50 µL sont purifiés sur colonne MinElute (élution dans 27 µL d'EB à 50°C).

II.10.2.1.3 Ligation

Les 25 µL d'ADN résultant de l'étape précédente (perte de volume de 2 µL constatée dans la colonne) sont préparés dans le mix décrit dans le Tableau 20. On utilise des adaptateurs pour le séquençage en *paired-end* du kit PE Oligo only (Illumina).

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale
Quick Ligation Reaction Buffer (5x)	10,0	1x
PE Adapter Oligo Mix (10 µM)	10,0	2,0 µM
Quick T4 DNA Ligase	5,0	
ADN	25,0	
Volume total	50,0	

Tableau 20: Préparation du mix pour la ligation des adaptateurs

Après 15 minutes d'incubation à 20°C, les 50 µL sont purifiés sur colonne QIAquick (élution dans 57 µL d'EB). La concentration finale d'ADN avec adaptateurs est mesurée au Qubit avec le kit Quant-iT dsDNA HS Assay (Life Technologies).

II.10.2.2 *Sizing*

La préparation des librairies se poursuit par la sélection des fragments de tailles différentes, appelée *sizing*. Celui-ci peut être réalisé en préparant un gel d'agarose à 2% en tampon TAE 2x. L'ADN est chargé dans un puits du gel avec un tampon de charge. A la fin de la migration, le gel est plongé dans un bain de SYBR Gold puis placé sur une table produisant une lumière bleue (Dark Reader Transilluminator), et on découpe au scalpel de fines bandes de gel de tailles identifiables par le marqueur de taille (100 pb) qui migre en parallèle. L'ADN est ensuite extrait de chaque bande à l'aide du kit QIAquick gel extraction (Qiagen) selon les recommandations du fournisseur.

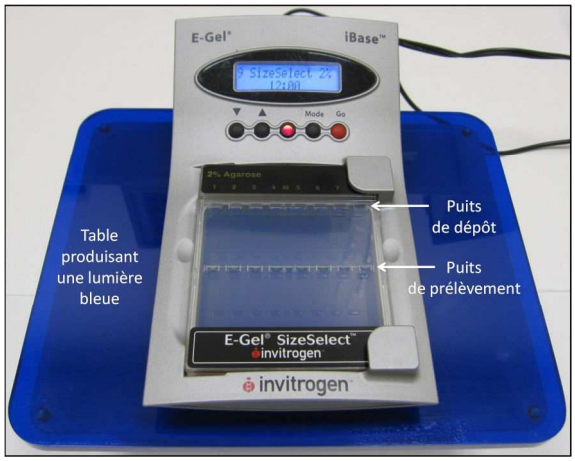


Figure 17: Description de l'E-gel utilisé pour le *sizing*

Il est également possible d'utiliser un gel d'agarose commercial (E-Gel size select 2% agarose, Life Technologies) sur un support adapté (E-Gel iBase Power System, voir Figure 17). Ce système permet de récupérer directement l'échantillon de taille désirée sans avoir à l'extraire du gel. 25 µL d'ADN sont déposés dans un des puits situés en haut du gel ainsi que 10 µL d'un marqueur de taille (100 pb)

dans le puits central. Les puits inutilisés sont complétés avec de l'eau. On dépose ensuite 30 µL d'eau dans les puits situés au milieu du gel (puits de prélèvement) et on lance la migration. Lorsque la taille d'intérêt repérée *via* le marqueur de taille arrive au niveau du puits de prélèvement, la migration est arrêtée et on y prélève le volume correspondant. 12 minutes sont par exemple nécessaires pour obtenir les fragments de 100 pb. On ajoute à nouveau de l'eau dans le puits de prélèvement et la migration est relancée jusqu'à une nouvelle taille d'intérêt.

II.10.2.3 Amplification par PCR et création de la librairie

L'amplification des fragments d'une taille spécifique est réalisée par PCR avec 2 polymérases dont on étudiera les performances : la Platinum Pfx (Life Technologies) et la Phusion High Fidelity (NEB). Les protocoles utilisés à l'origine sont présentés dans le Tableau 21 mais le volume d'ADN utilisé variera selon nos tests.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de PCR
10x Pfx Amplification buffer	5,0	1x	
$MgSO_4$ (50 mM)	2,0	2,0 mM	98°C 30 sec
dNTPs (2,5 mM)	4,0	0,2 mM	
Amorce PE 1.0 (Illumina, 25 µM)	1,0	0,5 µM	98°C 40 sec
Amorce PE 2.0 (Illumina, 25 µM)	1,0	0,5 µM	65°C 30 sec × n
Platinum Pfx DNA polymerase (2,5 U/µL)	0,4	1 U	72°C 30 sec
Eau milliQ	33,6		
ADN	3,0		72°C 5 min
Volume total	50,0		

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de PCR
Phusion HF Reaction Buffer 5x	10,0	1x	98°C 30 sec
dNTPs (10 mM)	1,5	0,3 mM	
Amorce PE 1.0 (Illumina, 25 µM)	1,0	0,5 µM	98°C 10 sec
Amorce PE 2.0 (Illumina, 25 µM)	1,0	0,5 µM	65°C 30 sec × n
Phusion HF DNA polymerase (2 U/µL)	0,5	1 U	72°C 30 sec
Eau milliQ	31,0		
ADN	5,0		72°C 5 min
Volume total	50,0		

Tableau 21: Conditions expérimentales de la PCR par la Platinum Pfx et la Phusion HF

Les produits d'amplification sont ensuite purifiés sur colonne MinElute et élués dans 16 µL d'EB. Leur profil de taille est vérifié sur le Bioanalyzer (avec une puce du DNA1000 kit) et leur concentration y est mesurée : elle doit être supérieure à 10 nM pour permettre à l'échantillon d'être séquencé.

II.10.2.4 Création des *clusters* et séquençage en *paired-end*

Toutes les étapes post-PCR sont effectuées sur la plateforme de séquençage. La première consiste à générer les *clusters* sur la *flow cell* (voir Figure 7). Ceci est effectué de manière automatisée sur un appareil fourni par Illumina, la cBot (voir Figure 18), avec les réactifs correspondants (Paired-end flow

cell v4, GA cBot manifold et TruSeq PE Cluster Kit v2 - cBot – GA). Les fragments sont hybridés sur la *flow cell* puis ils subissent l'amplification en pont, la linéarisation enzymatique, le blocage de leurs extrémités 3' et l'hybridation de la première amorce de séquençage.

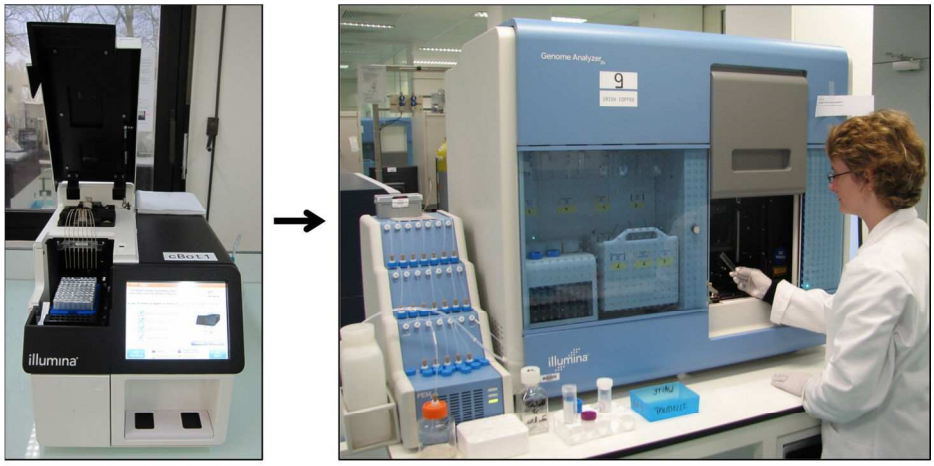


Figure 18: La cBot, le GAIIx et son paired-end module

La *flow cell* est déposée à l'intérieur du GAIIx (voir Figure 18) au-dessus d'un prisme permettant de dévier le laser qui excitera les fluorophores à chaque cycle de séquençage. 76 ou 101 cycles de séquençage sont effectués avec les réactifs Illumina (TruSeq SBS kit v5 – GA). Le paired-end module permet ensuite de retourner la molécule pour la séquencer par sa deuxième extrémité avec la deuxième amorce de séquençage (voir Figure 30) : c'est le *paired-end*. 76 ou 101 cycles de séquençage sont à nouveau réalisés.

II.10.3 Le séquenceur de paillasse GS Junior

II.10.3.1 Préparation de l'échantillon

Nous utilisons ce séquenceur dans le cadre du séquençage multiplexé d'amplicons. Les librairies ont été préparées par PCR grâce à des amorces qui contiennent les séquences d'adaptateurs de séquençage A et B. A la différence du séquenceur FLX du même fournisseur, la PTP du GS Junior ne comporte pas de lignes sur lesquelles déposer les échantillons de façon indépendante ; les amorces portent donc également des séquences MID (Multiplex Identifier) permettant de les identifier (voir Chapitre VI). Leurs concentrations respectives sont mesurées au Bioanalyzer (puce High Sensitivity, Agilent) et leur concentration moléculaire est déduite par la formule suivante :

$$C_{ADN\ double\ brin}(molécules/\mu L) = \frac{C_{ADN\ double\ brin}(ng/\mu L) \times N_A}{656{,}6 \times 10^9} \times \frac{1}{N}$$

où N_A = nombre d'Avogadro = $6{,}02 \cdot 10^{23}$ mol^{-1}, N = nombre de nucléotides (pb) et où 656,6 est la masse molaire (en g/mol) d'une paire de bases.

Les amplicons à séquencer sont ensuite réunis par dilution à une concentration finale de $1 \cdot 10^6$ molécules/µL dans un volume total de 200 µL de façon à ce que chacun soit présent en quantités égales (en terme de nombre de molécules). 12 µL de ce volume sont ensuite purifiés sur billes Agencourt AMPure XP, la dernière étape d'élution étant réalisée dans 40 µL de TE.

II.10.3.2 PCR en émulsion

5 millions de billes de capture A (GS Junior Titanium emPCR kit, Lib-A, Roche), permettant la liaison des fragments par leur adaptateur A, sont lavées, de même pour les billes de capture B qui permettent de les fixer par leur adaptateur B. Un volume d'ADN est ajouté de façon à obtenir deux molécules par bille. Ceci est réalisé en parallèle sur les billes A et sur les billes B qui sont traitées séparément selon les recommandations du fournisseur.

La PCR en émulsion est réalisée par 50 cycles de PCR (durée de 6h) : chaque bille est contenue dans un microréacteur constitué par une migrogoutte de l'émulsion eau/huile qui contient tous les réactifs nécessaires à la PCR. Les deux mix sont réunis, l'émulsion est ensuite cassée et les billes portant de multiples copies d'un fragment sont enrichies. Pour cela, des amorces biotinylées (*enrich primers*), complémentaires aux fragments liés aux billes de capture, sont ajoutées dans le milieu puis liées à des billes magnétiques de streptavidine (*enrichment beads*). Les billes portant l'ADN sont isolées par aimantation, élimination du surnageant puis dénaturation de l'amorce biotinylée. Enfin, les amorces de séquençage A et B (*seq primers*) sont ajoutées au milieu, hybridées aux extrémités respectives des fragments puis leur excès est éliminé par une série de lavages. Le nombre de billes obtenu est estimé grâce au GS Junior Bead Counter : environ 500000 billes doivent être utilisées pour la suite.

II.10.3.3 Préparation de la PTP et séquençage bidirectionnel

Chacune des billes portant les copies d'un seul fragment est déposée dans un puits de la PicoTiterPlate ou PTP (GS Junior Titanium Picotiterplate kit, Roche). Pour cela, la PTP est fixée sur un support (Bead deposition device, Roche) et différentes couches de billes y sont successivement déposées : des billes contenant des enzymes nécessaires à la réaction de pyroséquençage qui aura lieu (*enzyme beads prelayer*), des billes permettant de combler le volume du puits (*packing beads*) pour ne permettre qu'à une seule de nos billes de s'y insérer (*DNA beads*), puis à nouveau des billes contenant des enzymes nécessaires au pyroséquençage (*enzyme beads postlayer* et *PPiase beads*). La PTP est enfin chargée dans le GS Junior (voir Figure 19) pour le séquençage simultané des 2 extrémités de nos librairies grâce aux 2 types de billes utilisées (voir Figure 71).

Figure 19: Le GS Junior

II.10.4 Gestion des échantillons : le LIMS

Au début de l'expérimentation, l'échantillon recoit un code-barre qui lui est propre. A chaque étape du protocole, un nouveau code-barre lui est ensuite attribué et est enregistré dans le système de gestion des échantillons appelé LIMS (Laboratory Information Management System). Ceci permet d'assurer la traçabilité du grand nombre d'échantillons qui transitent sur la plateforme de séquençage et d'en obtenir un historique si désiré.

II.11 Méthodes d'analyse des données de séquençage

II.11.1 ELAND et la plateforme d'analyse d'Illumina

La surface d'une *flow cell* (FC) est divisée en 8 lignes, chacune d'entre elles correspondant dans notre cas à 1 échantillon, et étant divisée en 2 lignes de 60 *tiles* chacune. Un *tile* est la surface correspondant à une prise d'image par la caméra après chaque incorporation d'un nucléotide. Chaque *tile* donne donc lieu à 4 images par cycle de séquençage (une par nucléotide). La première étape de l'analyse par la plateforme d'Illumina, consiste à identifier chaque *cluster* par ses coordonnées sur la FC et à lui attribuer les bases correspondantes à chaque cycle (ce qui est appelé *base calling*). L'enchainement de ces bases fournit un *read*. On transforme donc les fichiers images en fichiers textes contenant les séquences nucléotidiques ainsi qu'un score de qualité de séquençage associé.

La deuxième étape est l'alignement des *reads* obtenus sur un génome de référence grâce à ELAND. Il génère un nouveau fichier contenant les *reads* classés par position d'alignement sur le génome ainsi qu'un score d'alignement associé.

II.11.2 Alignement avec BWA

Il est possible d'utiliser les fichiers textes générés après le *base calling* pour effectuer l'alignement avec BWA (Burrows Wheeler Aligner). Quelques précisions sont apportées sur les critères utilisés pour l'alignement dans la Figure 51A. Cet outil et son utilisation sont décrits dans la littérature (191,192).

II.11.3 Visualisation des données

Nous utilisons l'outil bioinformatique développé par l'UCSC (http://genome.ucsc.edu/) (187) pour visualiser nos données de séquençage. Dans l'onglet « Genomes », nous devons personnaliser notre recherche en cliquant sur « Add custom tracks » puis choisir l'assemblage de référence.

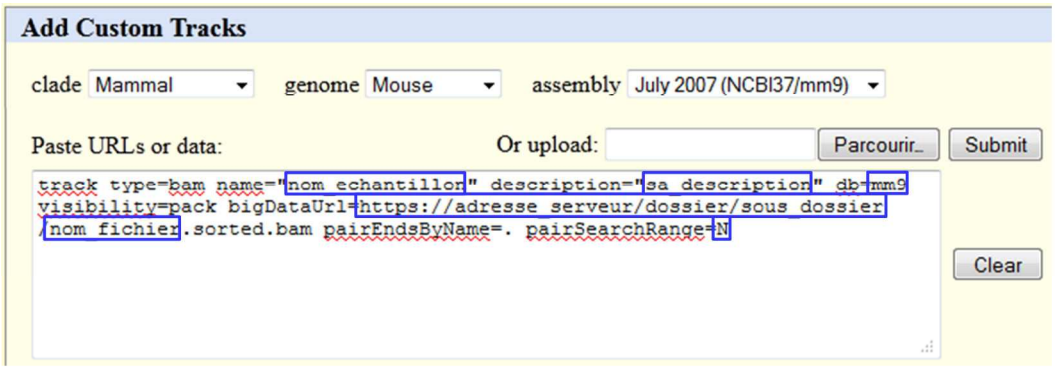


Figure 20: Paramètres utilisés pour la visualisation des données de séquençage en *paired-end*
Les paramètes encadrés en bleu sont à adapter à l'échantillon. mm9 : assemblage utilisé, adresse_serveur : lieu d'hébergement des données. N = distance maximale entre 2 *reads* d'une même paire.

La requête permettant de parcourir les fichiers contenant les séquences issues d'un séquençage en *paired-end*, convertis au préalable au format BAM, est décrite dans la Figure 20.

II.11.4 MEDIPS

MEDIPS est un logiciel d'analyse de données issues de MeDIP-Seq. Il permet d'en déduire des valeurs de méthylation sur des fenêtres d'une taille définie, sur tout le génome. Pour cela, il prend en compte le nombre de *reads* présents dans cette fenêtre par rapport à leur nombre total. Il effectue également une correction des valeurs de méthylation en fonction de la densité en CpGs de la fenêtre étudiée en proposant de compter les CpGs environnants ou de leur conférer un poids lié à leur distance à cette fenêtre. Il est également possible d'identifier des régions différentiellement méthylées (DMRs) entre 2 échantillons. Les méthodes de calcul sont détaillées dans la littérature (193).

II.11.5 Batman

Batman (Bayesian tool for methylation analysis) est un algorithme permettant l'analyse des profils issus de MeDIP-chip ou de MeDIP-Seq et l'estimation de leurs valeurs de méthylation (121). Il prend en compte le contexte des séquences étudiées et leur densité en CpGs pour adapter le taux de méthylation qu'il déduit. Lorsque des données issues de puces (MeDIP-chip) sont étudiées, elles subissent d'abord une normalisation quantile (194).

II.11.6 BiQ Analyzer HT

BiQ est un logiciel permettant l'analyse et la visualisation de données issues du séquençage après traitement par le bisulphite (195). Le script proposé a été développé pour étudier des données obtenues sur la plateforme de séquençage 454 de Roche. Nous l'utiliserons, non pas pour l'analyse du génome entier, mais dans le cadre du séquençage de loci spécifiques.

Les données sont d'abord converties au format FASTA puis chargées dans le logiciel. Il faut ensuite fournir les séquences génomiques correspondantes aux régions que l'on a ciblées et séquencées. Elles serviront de référence pour l'étude puisque l'alignement sera réalisé sur celles-ci après leur conversion au bisulphite *in silico*. L'algorithme utilisé est celui de Needleman-Wunsch qui garantit l'alignement optimal bien que pas nécessairement le plus correct. Nous conserverons les séquences présentant au minimum 90% d'identité avec leur référence. BiQ offre ensuite la possibilité de vérifier le taux de conversion des séquences par le traitement au bisulphite. Il permet enfin de calculer des pourcentages de méthylation de chaque CpG présent dans une séquence en prenant en compte le nombre de *reads* alignés sur cette position. Tous les résultats sont proposés sous forme graphique.

Chapitre III: Mise en place d'un protocole de MeDIP-Seq utilisable en routine

Un protocole de MeDIP-Seq va être développé dans cette partie, afin de l'introduire sur la plateforme de séquençage. Ce protocole devra donc être robuste pour être utilisable en routine. L'échantillon d'intérêt est d'abord fragmenté puis des adaptateurs sont ajoutés à ses extrémités (voir Figure 21). L'ADN est immunoprécipité puis une taille de fragments est sélectionnée (*sizing*). La librairie est enfin créée par PCR avec des amorces dont une partie est complémentaire des adaptateurs, l'autre partie permettant la fixation sur la *flow cell* pour le séquençage sur un GAIIx d'Illumina.

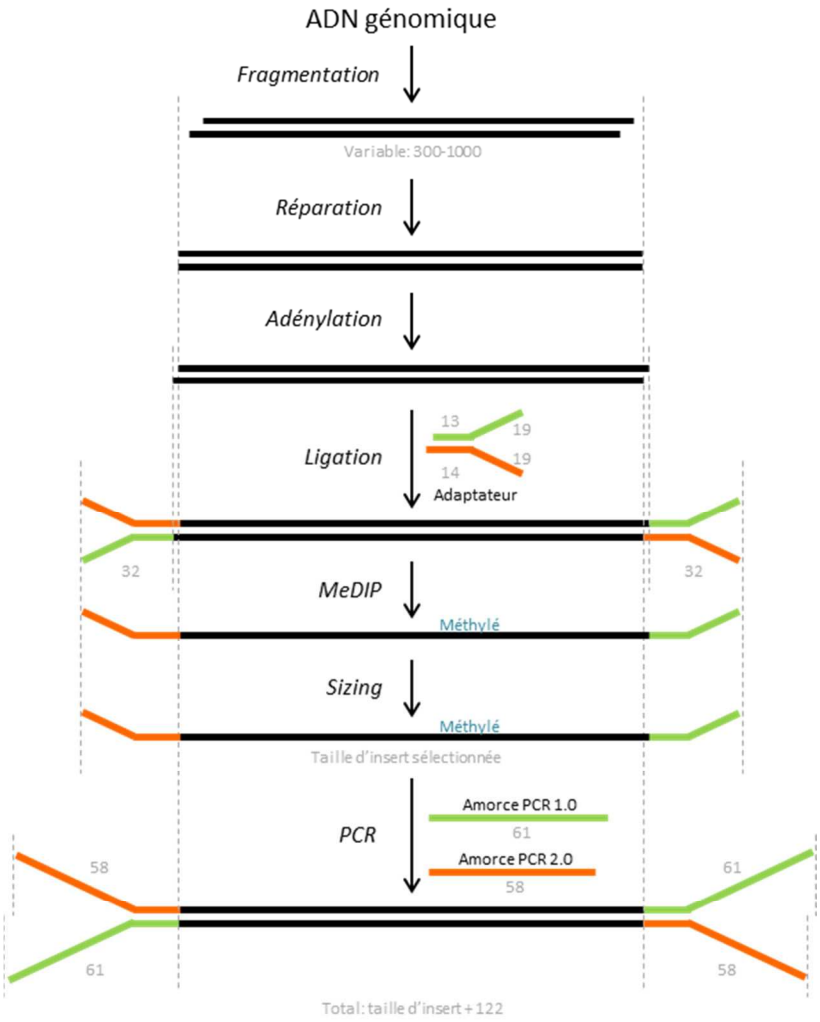


Figure 21: Procédure du MeDIP-Seq
En gris, tailles respectives des fragments (entre 2 lignes en pointillés notamment). La taille finale est égale à la taille de l'insert (en noir) additionnée de celle des amorces les plus longues (2 × 61 pb).

III.1 Optimisation de la fragmentation

La préparation de l'échantillon débute par sa fragmentation. Nous ciblons une distribution de tailles de fragments de 300 à 1000 pb, centrée sur une taille de 600 pb. Les plus longs fragments permettront ainsi de sélectionner des régions pauvres en CpGs et/ou faiblement méthylées lors du MeDIP.

Les premiers tests ont été réalisés sur le sonicateur Bioruptor (Diagenode) avec de l'ADN Promega dilué à 25 ng/µL dans 100 µL de TE (Tris EDTA). Le profil de tailles a été vérifié par dépôt des échantillons sur gel d'agarose à 1% et nous avons établi que la durée de sonication optimale était de 6 minutes. Cependant, nous avons rapidement dû faire face à la non-reproductibilité de cette technique en allongeant cette durée au cas par cas pour chacun de nos échantillons (voir Figure 22).

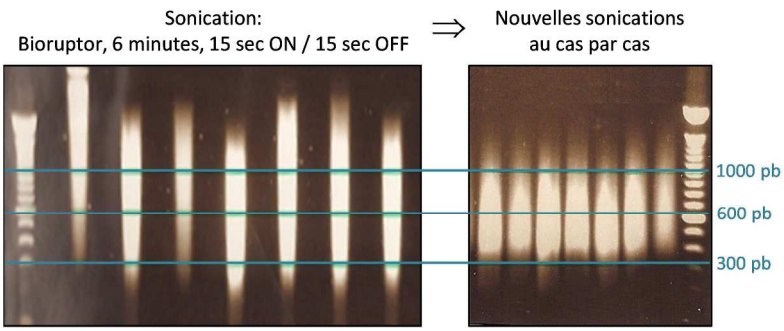


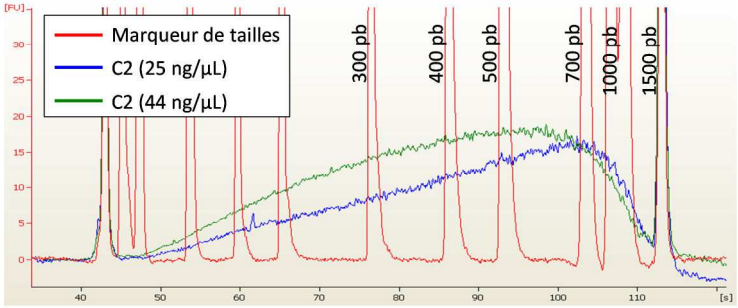
Figure 22: Profils de fragmentation obtenus avec le Bioruptor
Fragmentation de 7 échantillons d'ADN issus de placentas humains, pendant 6 minutes (en mode 15 secondes de sonication / 15 secondes d'arrêt, à puissance maximale). Chaque échantillon a été à nouveau soniqué pendant une durée qui lui est propre pour aboutir au profil de taille désiré. Premier et dernier dépôts : marqueur de taille (100 pb).

Afin de palier à ce problème, nous avons opté pour une fragmentation sur un autre appareil : le Covaris. La combinaison de différents paramètres (*Duty Cycle*, *Intensity*, *Cycle Per Bust*) génère une puissance de fragmentation bien définie. Pour obtenir la gamme de tailles recherchée, le fournisseur conseille d'utiliser les paramètres définis par les conditions A1 et A2 indiquées dans la Figure 23A. Nous les avons donc testées, ainsi que des conditions proches dont les paramètres sont également présentés dans la Figure 23A.

A)

	A1	A2	B1	B2	B3	C1	C2	C3
DC	5			10			5	
I	3			3			2	
CPB	200			200			200	
t (s)	75	85	30	40	50	80	100	120

B)



C)

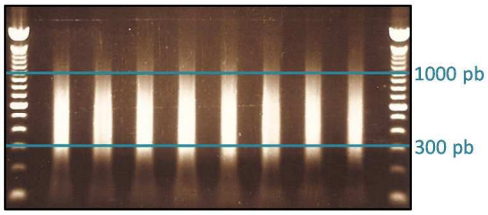


Figure 23: Sonication par le Covaris
A) Paramètres utilisés pour la fragmentation d'échantillons d'ADN Promega par le Covaris. DC = *Duty Cycle*, I = *Intensity*, CPB = *Cycle per Bust*, t = temps en secondes. En rouge, les conditions qui ont conduit aux meilleurs profils. Encadré rouge : meilleure condition. B) Profils obtenus au Bioanalyzer (puce DNA1000) après fragmentation dans les conditions définies par C2 (100 µL à 25 ng/µL et 82 µL à 44 ng/µL. C) Dépôt sur gel d'agarose de réplicats de fragmentation par le Covaris avec les paramètres définis par la condition C2.

Après fragmentation, la distribution des tailles a pu être visualisée sur le Bioanalyzer (puce DNA 1000, Agilent). La condition C2 a abouti au profil le plus proche de celui que nous recherchons (voir Figure 23B) et nous avons confirmé sa reproductibilité sur un échantillon d'ADN Promega (voir Figure 23C). Nous avons également vérifié que la distribution de tailles n'était pas modifiée de façon importante lors d'une augmentation de la quantité d'ADN à fragmenter (voir Figure 23B). Ceci trouvera un intérêt dans le paragraphe suivant.

III.2 Optimisation de la ligation des adaptateurs

La ligation des adaptateurs sur l'ADN fraichement fragmenté se divise en 3 étapes : la réparation des extrémités, leur adénylation puis l'ajout des adaptateurs. Chacune de ces étapes est suivie d'une purification sur colonne (QIAquick ou MinElute, Qiagen), ce qui implique une perte de matériel conséquente à la fin du protocole. Nous avons donc recherché la quantité initiale d'ADN nécessaire pour obtenir les 1,1 µg requis pour le MeDIP.

Nos tests de ligation ont débuté avec 1,9 µg d'ADN Promega de souris fragmenté en utilisant le kit de préparation Illumina (Paired-end sample prep kit). Une mesure au Nanodrop de l'échantillon après ligation nous a indiqué une perte de 40% de l'échantillon (voir Figure 24) et que la quantité finale était à peine suffisante pour poursuivre par une immunoprécipitation.

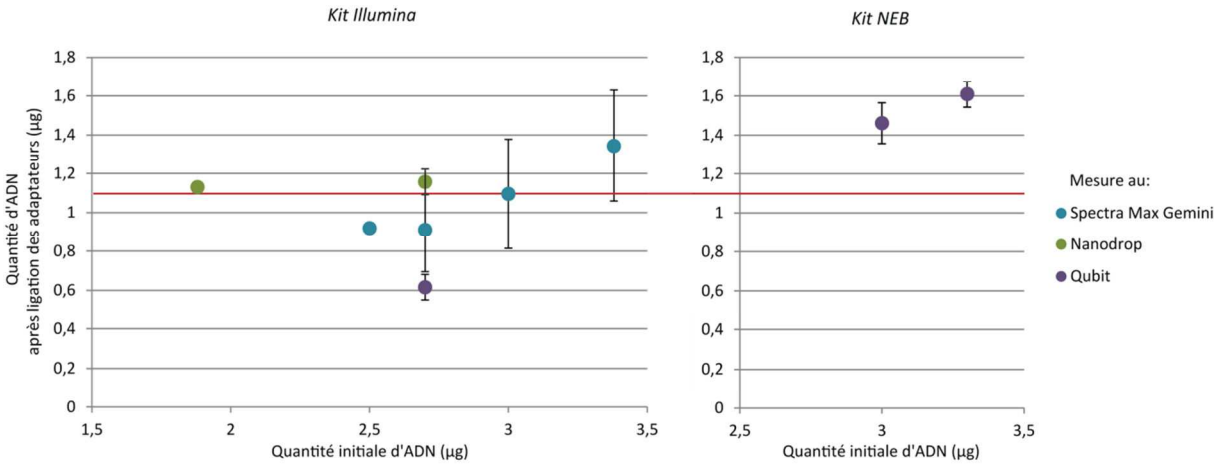


Figure 24: Quantités obtenues après ligation d'adaptateurs sur des quantités d'ADN croissantes
Quantités obtenues par ligation avec 2 kits de ligation (Illumina et NEB). Les quantités obtenues ont été mesurées par plusieurs techniques : avec le Spectra Max Gemini (Quant-iT dsDNA Assay kit, Broad Range), le Nanodrop et le Qubit (kit ds-DNA HS assay). Les écart-types sur ces graphiques représentent la variabilité entre les réplicats de ligation. Les 3 points à 2,7 µg (kit Illumina) représentent les mêmes réplicats mesurés à l'aide des 3 techniques de mesure. La ligne rouge horizontale représente la quantité minimale nécessaire pour le MeDIP (1,1 µg).

Nous avons donc utilisé des quantités croissantes d'ADN Promega jusqu'à obtention d'une quantité raisonnable pour le MeDIP après ligation. En deçà de 3,3 µg d'ADN, le kit Illumina n'a pas satisfait cette condition. Nous l'avons donc comparé au kit NEB (Next DNA sample prep master mix Set 1) en

utilisant cette même quantité. Une mesure au Qubit a montré une perte de 51% d'ADN et l'obtention de 1,6 µg, suffisants pour poursuivre et effectuer un MeDIP. Nous conserverons donc ces conditions.

Sur la Figure 24, nous observons que les différentes techniques utilisées pour mesurer les concentrations ne mènent pas à des résultats comparables. Le Nanodrop a tendance à surestimer les concentrations tandis que le Qubit fournit le chiffre le plus bas. Nous avons donc exclu la mesure au Nanodrop puis, pour une raison pratique d'accès aux appareils, nous avons conservé le Qubit comme appareil de référence.

Nous avons également souhaité tester le protocole de ligation communiqué par un autre laboratoire (Lee M Butcher, laboratoire de Stephan Beck, UCL Cancer Institute, University College London). Les durées d'incubation respectives y sont allongées et la quantité d'adaptateurs dans le milieu revue à la baisse. En effet, une mesure de concentration au Bioanalyzer (puce DNA 1000) à la fin de l'étape d'adénylation montre que nous utilisons un ratio ADN : adaptateurs de 1 : 26 en moyenne, tandis qu'un ratio 1 : 10 permettrait de fournir une quantité suffisante. De plus, dans ce protocole, toutes les étapes de purification sont effectuées sur billes (Agencourt AMPure XP beads, Beckman Coulter), ce qui permettrait de limiter les pertes de matériel. Nous avons donc effectué des tests préalables afin de comparer ce nouveau type de purification à celles que nous réalisons sur colonnes QIAquick et MinElute.

A)

Pertes constatées	Qubit	Puce HS
Billes AMPure	0,4%	20,1%
Colonnes QIAquick	25,4%	24,5%
Colonnes MinElute	22,7%	30,8%

B)

Concentrations obtenues	Qubit	Puce DNA 1000
Protocole standard	35,2 ng/µL	98,3 nM
Protocole adapté	26,0 ng/µL	36,2 nM

Tableau 22: Test d'un nouveau protocole de ligation des adaptateurs
A) Pertes liées à différentes purifications de 100 µL contenant 20 ng d'ADN Promega. Les mesures ont été réalisées au Qubit (kit Quant-iT dsDNA HS assay) et au Bioanalyzer (High Sensitivity DNA kit). B) Concentrations d'ADN obtenues avec 2 protocoles de ligation des adaptateurs. Protocoles réalisés avec 3,3 µg d'ADN. Mesures au Qubit (kit Quant-iT dsDNA HS assay) et au Bioanalyzer (DNA1000 kit). Protocole standard : décrit dans Matériel et méthodes. Protocole adapté : allongement des durées d'incubation (de 30 minutes à 1h pour la réparation et l'adénylation, de 15 minutes à 2h pour l'ajout des adaptateurs), purifications sur billes AMPure et diminution du ratio ADN : adaptateurs à 1 : 10.

Nous avons constaté que la purification sur billes menait effectivement à une perte plus faible de matériel (voir Tableau 22A). Néanmoins, la combinaison des divers paramètres du nouveau protocole n'a pas permis d'augmenter la quantité d'ADN après ligation des adaptateurs (voir Tableau 22B) et nous conserverons donc le protocole initial.

III.3 MeDIP

Le MeDIP permet de sélectionner les fragments méthylés dans 1 µg de notre échantillon portant les adaptateurs, dénaturé au préalable, grâce à un anticorps ciblant les 5-méthylcytidines. La technique du MeDIP est automatisée et nous utilisons la version d'un kit de Diagenode adaptée au robot du même fournisseur (AutoMeDIP kit et SX-8G IP-Star, Diagenode). Nous en vérifions ensuite l'efficacité par PCR et qPCR.

La PCR permet d'amplifier des régions connues pour être méthylées (*IGF2*) et non-méthylées (*PPia*) en tant que contrôles de l'immunoprécipitation.

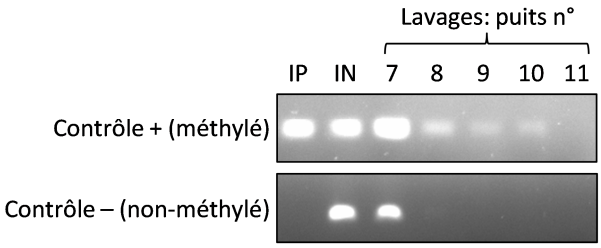


Figure 25: Contrôle du MeDIP par PCR
MeDIP sur de l'ADN Promega. Contrôle négatif : *PPia* (température d'hybridation de la PCR : 64°C). Contrôle positif : *IGF2* (66,7°C). IP : ADN immunoprécipité, IN : *input* (n'ayant pas subi l'immunoprécipitation). Lavages : voir distribution page 49.

Les régions non-méthylées ne sont pas sélectionnées par l'anticorps et sont uniquement présentes dans le premier puits de lavage, ainsi que dans l'IN qui ne subit pas l'immunoprécipitation (voir Figure 25). Les fractions méthylées sont immunoprécipitées et nous les retrouvons dans le puits correspondant à l'IP ainsi que dans celui de l'IN et en quantités décroissantes dans les puits de lavages de l'IP.

Le contrôle par qPCR (avec le SYBR Green, sur le Rotor-Gene) permet d'amplifier les régions contrôles méthylées et non-méthylées introduites dans la préparation de l'échantillon, afin de quantifier l'efficacité du MeDIP. Nous avons obtenu, en moyenne sur toutes nos immunoprécipitations, un taux de récupération des séquences méthylées de 31,7% tandis que celui des séquences non-méthylées était de 0,47%. De façon générale, ces taux sont similaires aux résultats du fournisseur et ont permis de poursuivre le protocole pour des valeurs supérieures à 30% et inférieures à 5% respectivement.

Nous avons envisagé de remplacer la purification manuelle par une purification sur billes des IPs grâce au protocole Auto IPure et au kit homonyme (Diagenode). Celle-ci permet un gain de temps par rapport à une purification manuelle sur colonne puisqu'elle est réalisée après le MeDIP, sur le même robot IP-Star. Cependant, à la fin du protocole, nous avons observé des reliquats de billes de purification au fond des puits de la plaque utilisée, ce qui laisse entendre que l'élution de notre échantillon a été incomplète. De plus, la qPCR standard de contrôle-qualité du MeDIP a révélé des

taux de récupération des fragments méthylés non reproductibles et aberrants (23,8 et 95,3% pour nos duplicats) et une quantification au Qubit (Quant-iT ssDNA HS Assay kit, Life Technologies) a montré une perte de 65,9% du matériel. Cette purification ne sera donc pas introduite en routine dans le protocole.

III.4 Mise en place d'un contrôle-qualité par pyroséquençage

Pour déterminer l'efficacité du MeDIP, des méthodes basées sur des PCRs sont principalement utilisées, que ce soit sur des oligonucléotides contrôles ou sur des régions supposées déméthylées ou très méthylées dans le type cellulaire étudié. Néanmoins, ceci ne renseigne pas sur l'état de méthylation de chaque CpG contenu dans de telles régions et on ne peut vérifier si seuls les fragments totalement méthylés sont immunoprécipités ou non. C'est pourquoi nous avons développé une analyse quantitative de l'ADN immunoprécipité, à la résolution du nucléotide, par pyroséquençage après traitement par le bisulphite. Ceci permettra de vérifier la qualité du MeDIP avant séquençage haut-débit en n'utilisant qu'une partie de l'échantillon immunoprécipité. Le but est de s'assurer qu'une région moyennement méthylée voit son pourcentage de méthylation augmenter nettement après MeDIP.

Le protocole standard de traitement par le bisulphite nécessite 500 ng d'ADN génomique. Nous n'avons pas cette quantité à disposition après MeDIP et souhaitons, de plus, n'utiliser qu'une fraction de l'IP pour ce contrôle qualité. Nous avons donc vérifié dans un premier temps que la conversion par le bisulphite tolérait l'emploi d'une plus faible quantité d'ADN. Nous avons traité des quantités décroissantes d'ADN Promega humain fragmenté en utilisant deux kits commerciaux : le kit Epitect (Qiagen) ainsi que le kit MethylEasy Xceed (Human Genetic Signatures, HGS) pouvant convertir de plus faibles quantités. Nous avons ensuite amplifié les échantillons résultants par PCR (sur *IGF2*). Aucune différence de performance n'a pu être constatée entre les deux kits et la faisabilité du traitement par le bisulphite sur de faibles quantités, jusqu'à 10 ng, a pu être démontrée (voir Figure 26).

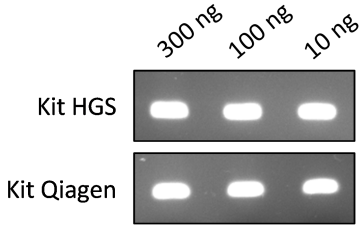


Figure 26: Traitement au bisulphite de quantités décroissantes d'ADN
Une PCR (*IGF2*) a été réalisée après traitement. On compare également 2 kits de bisulphite commerciaux, de Qiagen et de Human Genetic Signatures (HGS).

Nous disposons d'une douzaine d'échantillons de placentas pour lesquels nous possédons des données issues de puces à ADN. Celles-ci nous ont permis d'identifier des régions moyennement méthylées qui seront donc immunoprécipitées, et contenant un nombre de CpGs suffisant pour notre étude de pyroséquençage (5 CpGs minimum). Nous avons ainsi sélectionné des régions dans *OSTM1*, *C6ORF106*, *DLL1* et *FAM50B*. Nous avons également utilisé 2 régions connues pour être partiellement méthylées dans les placentas humains : un îlot CpG du promoteur de *RASSF1A* (196) et *IGF2* (197). Nous avons conçu pour ces régions des amorces de PCR spécifiques à l'ADN bisulphité ainsi que des amorces de pyroséquençage (voir séquences en Annexe 2).

1 µg de chaque échantillon a subi un MeDIP et la quantité d'ADN immunoprécipité a été mesurée au Nanodrop : nous avons obtenu une moyenne de 0,51 µg. Nous avons également contrôlé l'enrichissement par qPCR et avons obtenu un facteur d'enrichissement moyen de 215 sur les différentes immunoprécipitations. Les volumes correspondants à 0,2 µg ont ensuite été évaporés au speedvac pendant 1h et resuspendus dans 20 µL d'eau. Ils ont finalement été traités au bisulphite en parallèle avec les mêmes échantillons avant MeDIP. La quantité d'ADN bisulphité et son intégrité ont été vérifiées par une qPCR (avec le Rotor-Gene Probe PCR kit) utilisant une sonde Taqman et ciblant une séquence *Alu* (186). Nous avons détecté 0,135 µg de matériel amplifiable en moyenne sur nos 12 échantillons immunoprécipités.

1 µL de chaque échantillon bisulphité avant et après MeDIP ont été amplifiés par PCR par la Hot Star Taq polymérase (Qiagen) sur les 6 régions définies ci-dessus. Certaines PCRs ont dû être réalisées à nouveau en utilisant 2 µL d'ADN. En effet, une première amplification n'a pas fourni assez d'amplicon pour une détection sur gel d'agarose, ceci étant probablement dû à la faible quantité de molécules présentes dans l'aliquot utilisé pour le traitement au bisulphite.

10 µL de produit de chaque PCR ont ensuite été analysés par pyroséquençage. Des premiers tests ont montré que les régions sélectionnées de *C6ORF106* et *DLL1* étaient peu méthylées (moins de 10%) et donc moins aptes à être immunoprécipitées. Nous les avons donc écartées. En parallèle de nos échantillons, nous avons analysé un ADN commercial méthylé à 100% (Epitect methylated human control DNA, bisulphite converted) afin de prendre en compte le bruit de fond engendré par la technologie de pyroséquençage.

Pour les 5 régions étudiées, des niveaux de méthylation intermédiaires de 41 à 60% ont été déterminés dans les échantillons de placentas (voir Figure 27). Après MeDIP, des hauts niveaux de méthylation de 77 à 97% comparables à ceux du contrôle 100% méthylé ont été observés. Ceci montre la faisabilité de notre approche et la qualité du MeDIP automatisé sur le robot. Nous avons observé des valeurs de méthylation homogènes à travers les CpGs étudiés sur les différents

amplicons ; ceci indique que le MeDIP immunoprécipite principalement des molécules complétement méthylées et que l'information contenue dans des fragments où la méthylation serait hétérogène est perdue.

Cette méthode permet donc un contrôle qualité rapide de l'ADN méthylé immunoprécipité qui sera alors utilisé pour une analyse haut-débit de la méthylation sur le génome entier. La mise en place de ce protocole a fait l'objet d'une publication en 2010 (198) (voir page XIV).

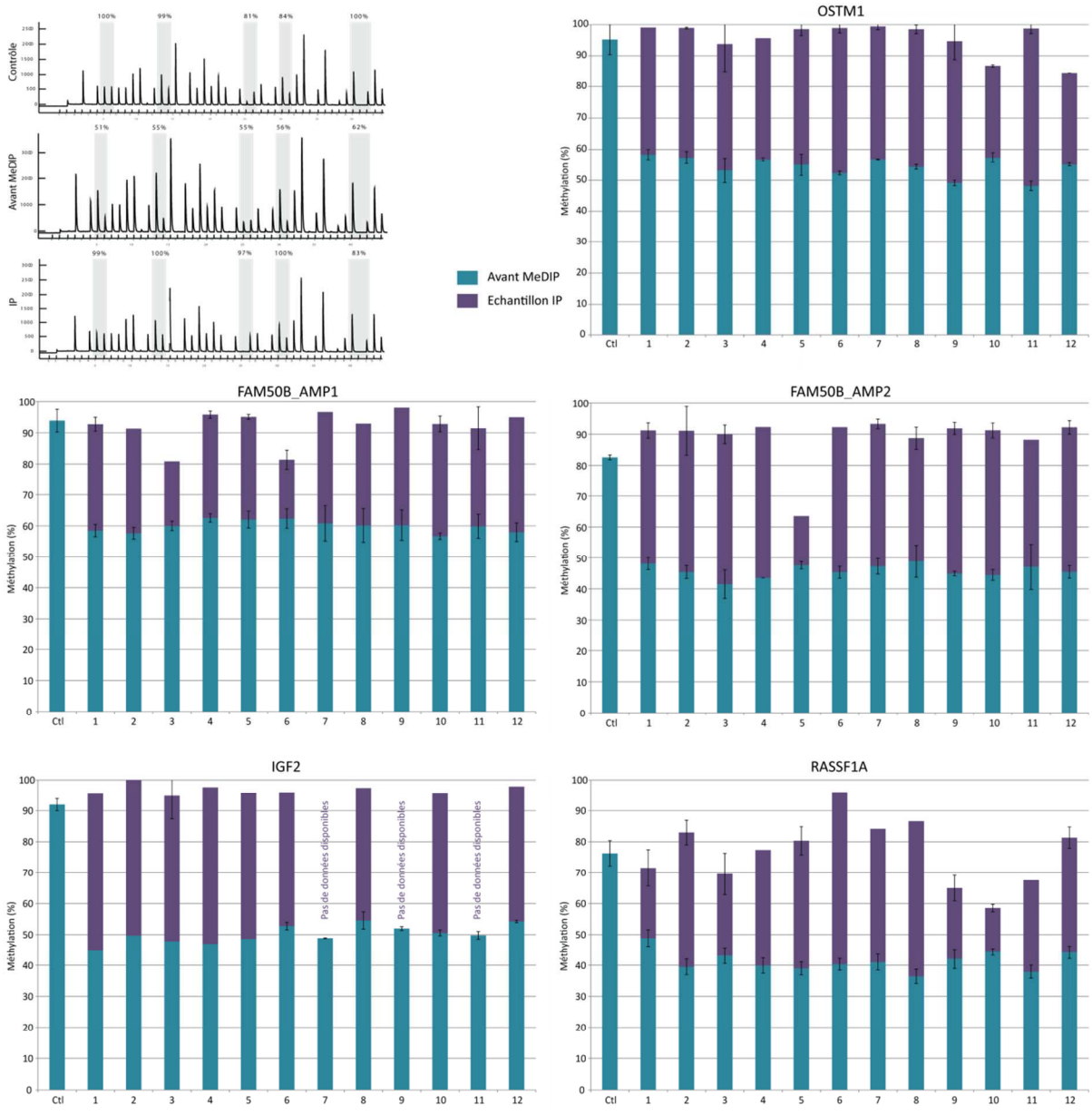


Figure 27: Pyroséquençage d'ADNs issus de placentas humains avant et après MeDIP
Données de méthylation d'un contrôle méthylé (Ctl) et de 12 ADNs issus de placentas (1 à 12) pour 5 régions dans 4 gènes. Une médiane des pourcentages obtenus pour chaque CpG individuel a été calculée pour chaque région. La moyenne de triplicats issus de 3 traitements au bisulphite et immunoprécipitations indépendants est représentée avec leurs écart-types. Pour chaque échantillon de placentas, les données sont fournies avant (en bleu) et après (en violet) MeDIP. En haut à gauche : de haut en bas, pyrogrammes de *OSTM1* de l'ADN contrôle méthylé, d'un échantillon bisulphité avant MeDIP et son équivalent après MeDIP.

III.5 Optimisation du *sizing* sur gel et de l'amplification par PCR

L'échantillon immunoprécipité possède les adaptateurs à ses extrémités qui vont permettre son amplification par PCR. Nous devons au préalable sélectionner les fragments correspondant à une taille précise incluant la taille de l'insert désirée et les 122 pb des amorces complémentaires aux adaptateurs ; ceci est effectué sur gel.

Nous utilisons un E-gel (size select 2%, Life Technologies) pour faire le *sizing*. Un volume maximal de 25 µL est déposé dans un puits du gel avant sa migration. Nous avons donc réduit 85 µL d'ADN immunoprécipité à ce volume de 25 µL par évaporation au speedvac. Au cours de la migration, nous prélevons dans le puits prévu à cet effet le volume d'échantillon correspondant à la taille théorique indiquée par le marqueur de taille qui subit la migration en parallèle.

Nous avons utilisé le matériel prélevé sur E-gel à une taille théorique de 400 pb. 3 µL ont été amplifiés par 10 cycles de PCR avec la polymérase Platinum Pfx (Life Technologies) et le produit de PCR a été purifié. Le protocole ainsi élaboré a permis de préparer un premier échantillon (ADN de MEF (Mouse Embryonic Fibroblast) et de le séquencer sur un Genome Analyzer IIx (Illumina). Les données issues de cet échantillon (non exploitées ici) ont montré un grand nombre de séquences non-spécifiques, absentes des régions riches en CpGs, et nous ont conduit à reconsidérer le protocole d'immunoprécipitation : la quantité d'anticorps utilisée a été revue à la baisse (150 ng au lieu des 300 utilisés jusqu'ici). Cependant, la quantité de matériel disponible après MeDIP s'en est trouvée réduite puisque plus spécifiquement sélectionnée, et la dizaine de cycles de PCR s'est trouvée insuffisante pour permettre la détection du matériel sur le Bioanalyzer. Nous avons donc revu à la hausse le nombre de cycles de PCR : 14 cycles n'ont pas suffi à générer suffisamment de matériel tandis que 18 cycles ont permis d'obtenir les profils de la Figure 28.

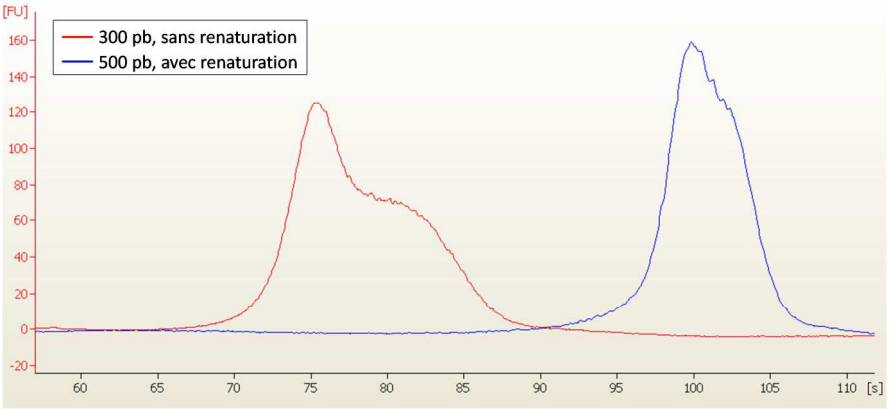


Figure 28: Profils obtenus après 18 cycles de PCR dans le cadre du MeDIP-Seq

Un double pic a pu être observé à la taille des fragments amplifiés (voir Figure 28, profil rouge). Nous l'avons interprété comme un biais lors du *sizing* ; en effet, l'ADN immunoprécipité est à l'état simple

brin et il est concevable que certains réappariemments s'effectuent avant le dépôt sur gel malgré les précautions prises pour conserver l'échantillon à 4°C. Ainsi, un mélange d'ADN simple et double brin serait déposé sur le gel. Ces deux ADNs ne migrant pas à la même vitesse, deux tailles de fragments différentes seraient prélevées et donc amplifiées par PCR. Nous avons envisagé que ce phénomène soit dû au type de gel utilisé. Aussi avons-nous réalisé le *sizing* sur un gel d'agarose fraichement préparé en découpant les bandes de tailles d'intérêt à l'aide d'un scalpel puis en réalisant l'extraction de l'ADN à l'aide du kit QIAquick gel extraction (Qiagen). Cependant, l'inconvénient de ce type de *sizing* est une perte évidente d'ADN due à la purification sur colonne du kit cité. Nous n'avons, de plus, pas observé d'amélioration quant au double pic du produit de PCR. Nous avons donc mis en place et réalisé avant le *sizing* un protocole de renaturation des brins complémentaires par chauffage à 95°C pendant 5 minutes puis diminution de la température de 1°C par minute jusqu'à 37°C et maintien à 4°C. Après PCR dans les conditions précédentes, nous avons observé que ce traitement a permis de réduire la taille du second pic parasite, même s'il est clair qu'un léger épaulement subsiste (voir Figure 28, profil bleu).

Une amplification reproductible a pu être réalisée en ajoutant 2 cycles de PCR, soient 20 au total (voir Figure 29 pour deux échantillons de MEFs différents). Ces résultats ont été obtenus en utilisant directement 25 µL d'ADN immunoprécipité pour le sizing sur E-gel afin d'éviter la dégradation de l'ADN pendant la longue étape d'évaporation au speedvac ainsi que pour se trouver dans les mêmes conditions que dans les tests de déplétion (voir chapitre suivant).

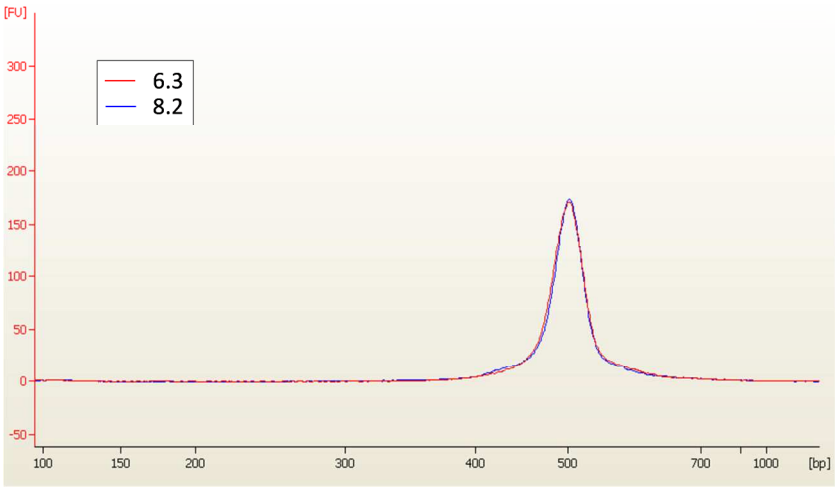


Figure 29: Profils obtenu après 20 cycles de PCR dans le cadre du MeDIP-Seq
Superposition des profils obtenus après PCR pour les échantillons de MEF 6.3 et 8.2.

III.6 Préparation de la *flow cell* et séquençage en *paired-end*

Après la PCR, l'échantillon entre dans le processus de séquençage et est préparé sur la *flow cell*. Les *clusters* sont créés puis la première amorce de séquençage hybridée (voir Figure 30). 76 ou 101

cycles de séquençage sont réalisés sur un GAIIx. Ils donneront lieu au *read* 1 (R1) une fois l'attribution des bases effectuée.

Contrairement au séquençage en *single-read* où le processus s'arrêterait ici, le *paired-end* permet de retourner la molécule pour effectuer à nouveau 76 ou 101 cycles de séquençage à partir de sa deuxième extrémité (voir Figure 30). Ceci donnera lieu au *read* 2 (R2).

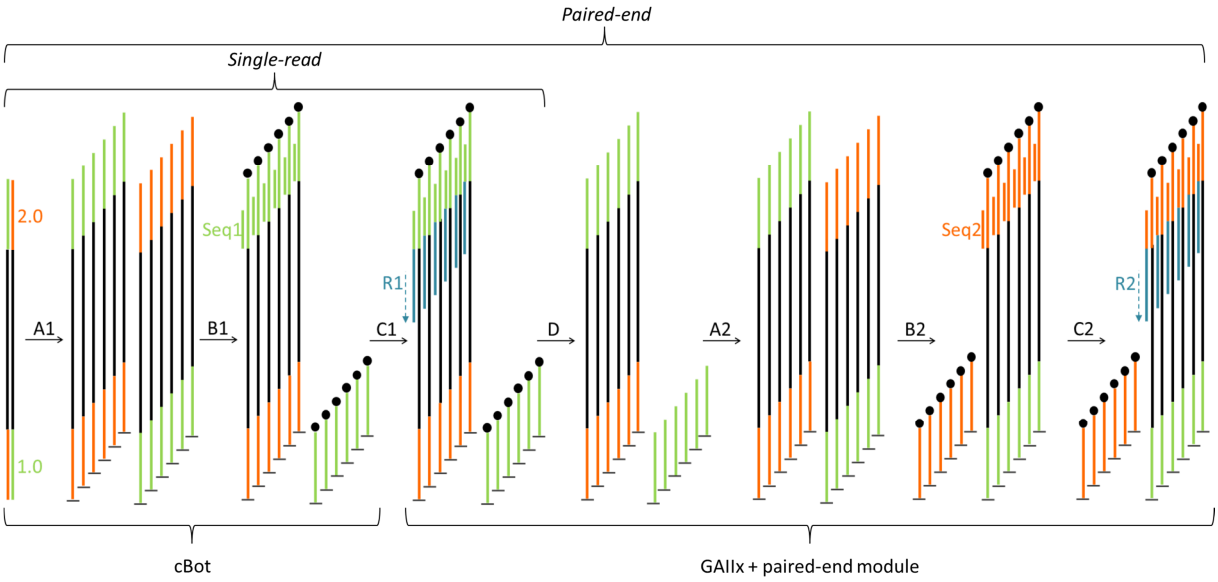


Figure 30: Séquençage en *paired-end* sur le GAIIx
A la surface de la *flow cell* : hybridation des fragments générés par PCR avec les amorces 1.0 et 2.0. A1 (puis A2) : génération des *clusters* par amplification en pont. B1 (puis B2) : préparation au séquençage : linéarisation enzymatique (élimination des brins anti-sens (ou sens pour B2)), blocage des extrémités et des oligonucléotides libres à la surface de la *flow cell* (points noirs) et hybridation de la première amorce de séquençage (Seq1) (resp. la deuxième amorce de séquençage (Seq2)). C1 (puis C2) : cycles de séquençage. Un séquençage en *single-read* termine après C1. D : élimination du brin nouvellement synthétisé et déblocage des extrémités. Les étapes A1 et B1 sont réalisées sur la cBot. C1 à C2 sont réalisées dans le GAIIx. D à B2 font intervenir le paired-end module.

III.7 Développement d'un outil informatique pour l'analyse des données de MeDIP-Seq

De nombreux outils ont été récemment mis en place pour analyser des données de séquençage haut-débit : Pyicos (199), Artemis (200), NARWHAL (201), GeneProf (202) et GenPlay (203) en font partie et la liste est loin d'être exhaustive. Cependant, peu d'entre eux sont adaptés à l'analyse de séquences issues de MeDIP-Seq. Deux méthodes, Batman (121) et MEDIPS (193), ont été développées à cet effet mais elles n'incluent ni le contrôle-qualité des séquences, ni leur alignement, ce qui rend la préparation des données fastidieuse.

Afin de permettre une analyse plus rapide et pratique des données de MeDIP-Seq, nous avons développé en collaboration avec le groupe de bioinformatique du CEPH (Centre d'Etude du Polymorphisme Humain, Paris) une plateforme d'analyse appelée MeQA (MeDIP-Seq data Quality

assessment and Analysis). Celle-ci permet dans un premier temps de contrôler la qualité des séquences en fournissant plusieurs représentations graphiques et d'obtenir un alignement des données brutes grâce à l'utilisation de BWA (191,192). Dans un second temps, MeQA permet de découvrir la distribution des séquences et d'évaluer leur niveau de méthylation en faisant appel à MEDIPS cité ci-avant. La création de MeQA a fait l'objet d'une publication (204) (voir page XVIII).

III.8 Protocole final

Les développements effectués ont conduit à établir un protocole final dont les paramètres sont représentés sur la Figure 31.

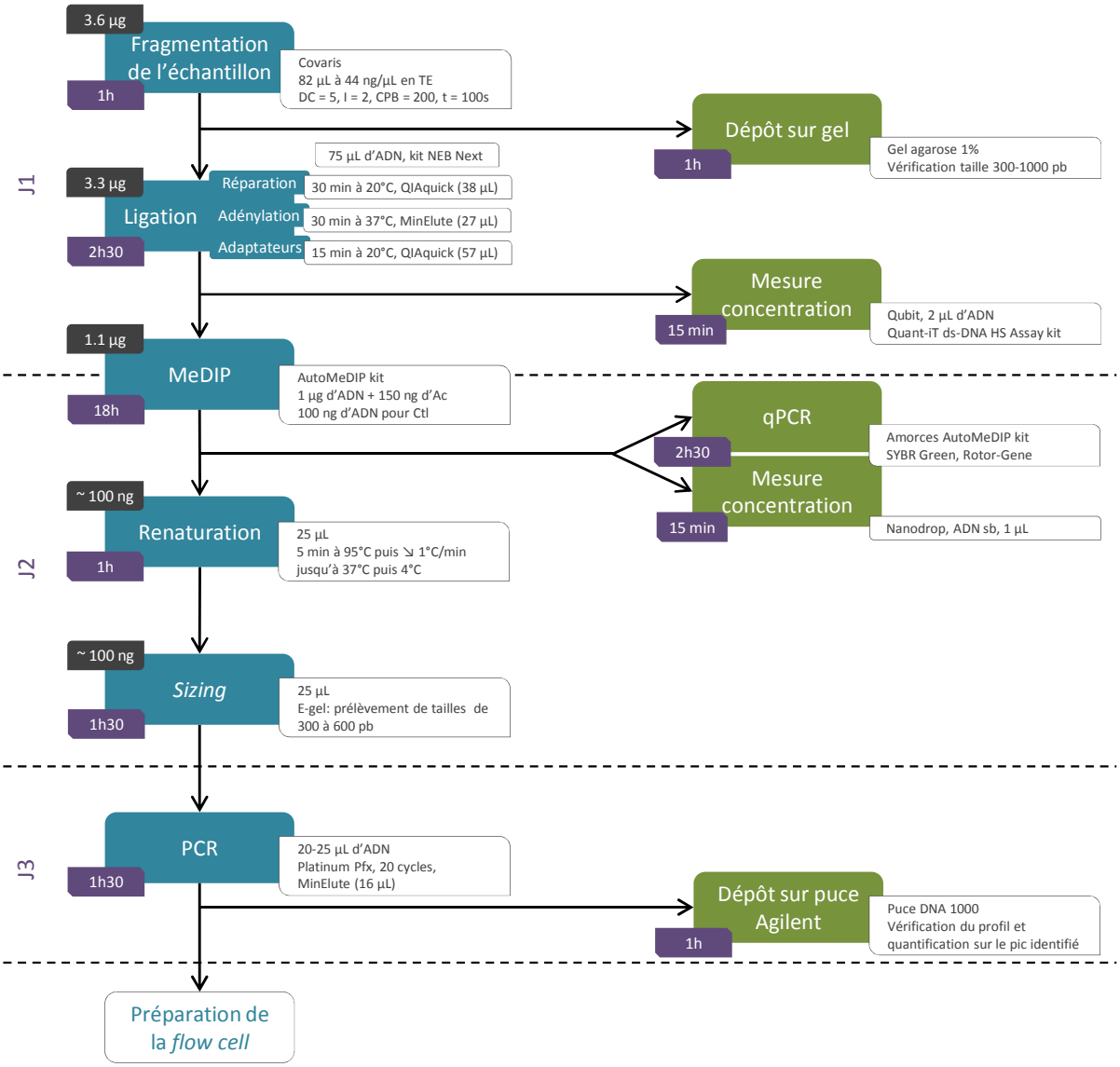


Figure 31: Protocole final de préparation de l'échantillon pour le MeDIP-Seq
Le temps indiqué est la durée nécessaire à la préparation d'un seul échantillon. Le volume indiqué entre parenthèses après le mode de purification est celui de l'élution. J : jour, Ac = anticorps 5-méthylcytidine. Ctl = contrôle, ADN sb : mesure en mode ADN simple brin.

La *flow cell* est ensuite préparée sur la cBot (pendant 5h) afin de créer les *clusters* à sa surface et d'hybrider l'amorce de séquençage. Le séquençage de 101 pb en *paired-end* dure 9,5 jours.

III.9 Discussion

Nous discuterons dans cette partie d'aspects techniques liés à la mise au point du protocole, tout comme ce sera le cas dans les chapitres IV, V et VI. Une discussion plus large aura lieu dans le chapitre VII.

Après son développement en 2005 (135), le MeDIP a fait l'objet d'une introduction dans les processus de séquençage haut-débit à partir de 2008 pour conduire au MeDIP-Seq. Il a depuis permis d'étudier l'implication de la méthylation dans différents types cellulaires (121,147,159,193), dans certaines pathologies humaines (122,139,160) ou très récemment dans différents tissus végétaux (205).

L'intérêt croissant pour cet outil et sa pertinence dans l'étude de la méthylation de l'ADN sur le génome entier nous ont conduits à mettre en place le MeDIP-Seq dans notre laboratoire. Chaque étape du protocole a été évaluée et optimisée. Nous avons d'abord établi des paramètres pour la fragmentation mécanique des échantillons sur le Covaris. Ceci permet de générer des fragments de façon aléatoire afin d'éviter le biais que pourrait introduire un autre type de fragmentation comme une digestion enzymatique. Nous avons ensuite optimisé la ligation d'adaptateurs de séquençage sur nos fragments avant d'effectuer l'immunoprécipitation. Plusieurs kits de MeDIP existent sur le marché et ont été récemment impliqués dans une étude comparative (206) : celui de Diagenode que nous avons utilisé a montré le meilleur enrichissement. Le MeDIP est automatisé sur le robot SX-8G IP-Star (Diagenode) pour une plus grande reproductibilité et a fait l'objet d'une description détaillée dans la littérature (207).

Nous avons également développé un protocole annexe permettant de contrôler la qualité du MeDIP grâce au traitement par le bisulphite et à une analyse par pyroséquençage (198). Il donne la possibilité d'analyser chaque CpG dans un fragment immunoprécipité et renseigne donc sur l'homogénéité du statut de méthylation à travers les régions sélectionnées. Ce protocole pourra s'appliquer de façon plus large au contrôle d'autres types d'immunoprécipitations comme le ChIP.

Nous avons choisi un mode de *sizing* des fragments sur gel. Cette étape n'est pas optimisée pour de l'ADN simple brin, configuration obtenue après MeDIP, aussi avons-nous introduit une renaturation avant cette sélection de tailles. Dans le but de nous affranchir de cette étape, nous envisageons désormais d'amplifier l'ADN immunoprécipité grâce à une dizaine de cycles de PCR afin d'obtenir l'ADN en configuration double brin pour le *sizing*. Ceci est couramment effectué dans le cadre du

MeDIP-Seq (voir Tableau 23, 4e colonne). Il serait également possible de synthétiser le brin complémentaire de nos fragments immunoprécipités comme il a pu être réalisé dans un récent protocole (122) ou par l'utilisation du fragment de Klenow de la polymérase I. Nous pourrons également considérer un autre mode de sélection des tailles en utilisant les billes Agencourt AMPure XP (Beckman Coulter). Ce type de billes a été développé pour la purification d'échantillons grâce à la technique d'immobilisation réversible en phase solide (SPRI) mais récemment, plusieurs équipes les ont appliquées à la sélection de tailles de fragments en adaptant le protocole initial (208-210) et en variant la composition des tampons utilisés à l'origine (211).

La dernière étape de notre protocole consiste à amplifier les fragments immunoprécipités par le biais des adaptateurs de séquençage. 20 cycles de PCR ont été nécessaires pour obtenir une amplification qui soit reproductible car la quantité de matériel obtenue après MeDIP est faible. Il est clair qu'un tel nombre de cycles peut introduire des biais de séquences dans nos librairies mais nous prendrons, lors de l'analyse, les mesures nécessaires pour nous en affranchir (voir Chapitre 5). Différents protocoles de MeDIP-Seq publiés dans la littérature utilisent toutefois davantage de cycles, jusqu'à 24 pour certains (voir Tableau 23, 3e colonne).

	Quantité initiale d'ADN	Quantité d'ADN pour le MeDIP	Cycles de PCR	Enchainement des étapes du protocole de MeDIP-Seq
Down, T.A. *et al.* (2008) (121)	10 µg	1 µg	nd	Ligation / MeDIP / *Sizing* / PCR
Pomraning, K.R. *et al.* (2009) (138)	5 µg	5 µg	18-24 cycles	MeDIP / Ligation / *Sizing* / PCR
Ruike, Y. *et al.* (2010) (122)	4 mg*	4 mg*	18-24 cycles	MeDIP / Ligation / PCR / *Sizing*
Harris, R.A. *et al.* (2010) (147)	2 - 5 µg	nd	15 cycles	Ligation / MeDIP / PCR / *Sizing*
Li, N. *et al.* (2010) (159)	5 µg	nd	16 cycles	Ligation / MeDIP / PCR / *Sizing*
Bock, C. *et al.* (2010) (160)	300 ng	nd	nd	Ligation / *Sizing* / MeDIP / PCR
Chavez, L. *et al.* (2010) (193)	5 µg	4 µg	6 cycles	Ligation / MeDIP / PCR / *Sizing*
Butcher, L.M. *et al.* (2010) (207)	6 µg	1 µg	12 cycles	Ligation / MeDIP / PCR / *Sizing*
Feber, A. *et al.* (2011) (139)	5 µg	1 µg	nd	Ligation / MeDIP / *Sizing* / PCR
Vining, K.J. *et al.* (2012) (205)	10 - 12 µg	nd	15 - 21 cycles	Ligation / MeDIP / PCR (*Sizing* ?)
Sengenès, J. *et al.* (non publié)	3,6 µg	1 µg	20 cycles	Ligation / MeDIP / *Sizing* / PCR

nd : information non disponible

*: Quantité précisée en mg dans la publication

Tableau 23: Comparaison avec des protocoles de MeDIP-Seq de la littérature

Pour diminuer le nombre de cycles de PCR, il aurait été envisageable d'inverser les étapes du protocole de MeDIP-Seq et d'amplifier l'IP pour réaliser la sélection de tailles de fragments sur une quantité plus importante d'ADN. Ceci privilégierait l'amplification des courts fragments par rapport aux plus longs et nous n'avons pas, jusqu'à présent, voulu effectuer ce changement afin de ne pas introduire de biais supplémentaires dans nos librairies. Il aurait en revanche été possible d'augmenter la quantité d'ADN à immunoprécipiter afin de réduire ce nombre de cycles. Certains

groupes effectuent par exemple le MeDIP sur une quantité avoisinant les 5 µg (voir Tableau 23, 2e colonne). Diagenode recommande aussi de dérouler le processus sur plusieurs aliquots du même échantillon. Cependant, tout ceci implique que la quantité initiale de matériel soit revue à la hausse et nous n'avons pas souhaité franchir la limite des 4 µg, au contraire de ce qui est majoritairement réalisé dans les protocoles de MeDIP-Seq (jusqu'à 12 µg, voir Tableau 23, 1ère colonne), car nous ne disposerons en routine que de peu de matériel, surtout dans le cas d'échantillons rares et précieux que pourront nous confier des collaborateurs. Il est donc novateur d'avoir mis au point un protocole qui ne nécessite que 3,6 µg d'ADN initiaux.

Certains obstacles que nous avons rencontrés nécessitent de mettre en place des outils pour y faire face. Nous avons par exemple pu observer à plusieurs reprises que la quantification de nos échantillons n'était pas reproductible selon les appareils de mesure utilisés. Tout d'abord, une mesure au Nanodrop nous a laissé penser que 400 ng étaient obtenus en moyenne après immunoprécipitation. Ce chiffre nous a semblé incorrect car surestimé en comparaison à la trentaine de ng à laquelle on peut s'attendre (121), et nous avons accusé une imprécision due à la mesure par densité optique. Ce type de mesure a en effet pour principal défaut de faire contribuer l'absorbance du milieu de dilution de l'échantillon à la valeur fournie. Nous avons alors vérifié par fluorimétrie (sur le Qubit, avec le kit Quant-iT ssDNA HS Assay) que la quantité réellement obtenue devait être revue à la baisse et avons mesuré en moyenne 22,1 ng d'ADN immunoprécipité. Néanmoins, la fluorimétrie a également parfois fait preuve d'imprécision en fournissant des valeurs aberrantes de quantités de matériel, notamment après l'étape de ligation des adaptateurs (mesure avec le kit Quant-iT dsDNA HS Assay). En effet, nous avons ponctuellement relevé des valeurs de 50 à 60 ng/µL, ce qui laisserait sous-entendre qu'aucune perte n'ait été provoquée par les 3 purifications sur colonnes qui précèdent. Ceci nous a donc amenés, en routine, à considérer l'échantillon à une concentration de 30 ng/µL, moyenne observée sur l'ensemble de nos échantillons, à la fin de la ligation des adaptateurs. Une alternative à ces approches, la qPCR par exemple, devra donc être trouvée pour palier à ces incertitudes de quantification.

Le protocole est désormais robuste et peut-être utilisé en routine. Trois jours seulement seront nécessaires à la préparation de l'échantillon avant son introduction dans le processus de séquençage. Le choix de la plateforme a été guidé par les technologies de séquençage alors disponibles. Le parc de machines dont nous disposions à l'époque (juillet 2010) était essentiellement composé de Genome Analyzer IIx ; nous avons donc opté pour ce séquenceur ainsi que pour les versions de kits de réactifs associés, que ce soit pour la préparation des échantillons ou pour le séquençage en lui-même. Depuis, les HiSeq2000 ont pris le pas sur les GAIIx et ont permis d'augmenter considérablement le potentiel de séquençage des machines d'Illumina (voir Figure 75). Notre

protocole pourra aisément être utilisé avec les nouveaux kits de préparation proposés par le fournisseur (différentes versions de kits TruSeq). Ceux-ci permettront d'analyser davantage d'échantillons en parallèle grâce à une stratégie de multiplexage qui repose sur l'indexage des adaptateurs. Il est important de souligner que l'identification des échantillons par des codes-barres et leur suivi dans le LIMS seront alors d'autant plus nécessaires afin de gérer la quantité croissante d'échantillons impliqués dans la préparation d'une seule *flow cell*.

Enfin, nous avons développé MeQA (204), un outil informatique spécialisé dans l'analyse de données issues de MeDIP-Seq, qui permet de combler le manque existant dans ce domaine. Il s'agit d'une plateforme complète qui, d'une part, analyse et trie les données brutes de séquençage et, d'autre part, en fournit un alignement sur un génome de référence et une quantification de la méthylation sur l'ensemble des séquences obtenues. Nous fournissons ainsi non seulement un protocole de MeDIP-Seq mais aussi les moyens d'en tirer toutes les informations nécessaires à quiconque désirera mettre en place ce type d'étude.

Chapitre IV: S'affranchir des éléments répétés du génome : le MeDIP-dep-Seq

L'ADN immunoprécipité est riche en séquences répétées. Nous cherchons dans ce chapitre à diminuer la quantité de ces séquences tout en conservant les séquences d'intérêt que nous qualifierons de séquences uniques. Cette étape est introduite après le MeDIP dans le processus du MeDIP-Seq pour mener à un protocole que nous baptiserons MeDIP-dep-Seq (MeDIP-Seq et déplétion des séquences répétées). La stratégie utilisée repose sur l'utilisation d'un ADN enrichi en ces mêmes séquences répétées : l'ADN Cot-1. Celui-ci est biotinylé puis fixé sur des billes magnétiques de streptavidine et dénaturé (voir Figure 32). Les séquences répétées contenues dans l'échantillon immunoprécipité (simple brin) peuvent alors s'hybrider au Cot-1 et seront retenues par simple aimantation des billes tandis que l'échantillon enrichi en séquences uniques pourra être isolé dans le surnageant.

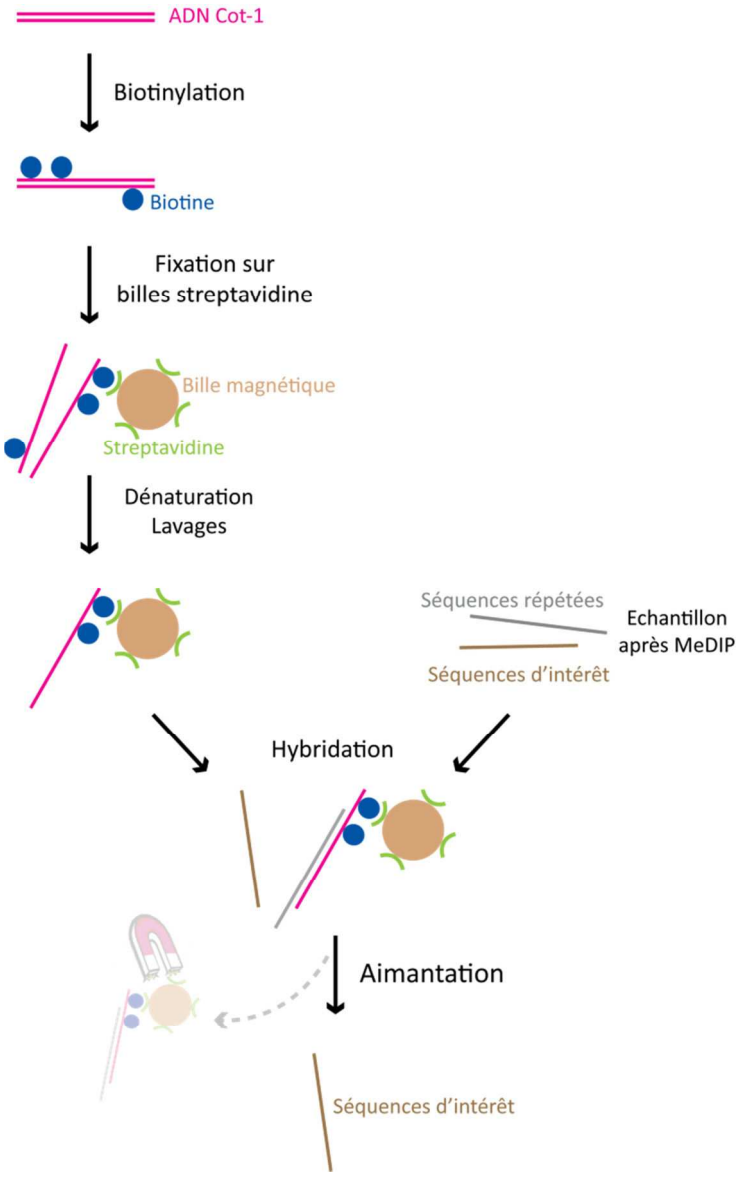


Figure 32: Procédure de la déplétion des séquences répétées

La mise au point du protocole est réalisée sur de l'ADN Promega de souris car nous disposons au laboratoire de données de puces à ADN récoltées sur des échantillons de souris avec lesquelles il sera intéressant de comparer nos résultats.

IV.1 Biotinylation d'ADN Cot-1

L'ADN Cot-1 doit être biotinylé afin d'être fixé aux billes de streptavidine. Nous abordons deux stratégies de biotinylation : l'addition d'une biotine aux extrémités 3' des fragments de Cot-1 et l'incorporation de plusieurs molécules de biotine lors d'une amplification linéaire. Le rendement de la biotinylation pourra être mesuré par marquage fluorescent dans le premier cas, et par mesure de concentration dans le second cas.

IV.1.1 Biotinylation aux extrémités 3'

Nous avons testé cette approche en utilisant 80 ng d'un amplicon de PCR. Celui-ci a été biotinylé par incorporation en 3' de biotin-16-ddUTP par une terminal transférase (Roche). Après incubation à 37°C (de 15 minutes à toute une nuit), désactivation de l'enzyme et purification pour éliminer les nucléotides résiduels, le matériel biotinylé a été fixé à de la streptavidine marquée par fluorescence (Streptavidin-Cy3, GE Healthcare) par incubation à 37°C, ceci dans le but de vérifier l'incorporation de biotine en 3' par fluorimétrie. La streptavidine fluorescente résiduelle a ensuite été éliminée. Nos tests ont révélé que sa purification par gel-filtration ou sur colonne était inefficace. En effet, la masse moléculaire de la streptavidine étant de 60 kDa celle-ci ne peut être retenue sur le gel et la streptavidine-Cy3 résiduelle se retrouve dans les échantillons.

Nous avons donc utilisé une table de purification (PyroMark Q96 vacuum prep workstation, Qiagen) pour éliminer la streptavidine-Cy3 et permettre la vérification de la biotinylation après amplification par PCR. Aucun résultat concluant n'a pu être obtenu quant à ce processus de biotinylation. En effet, il est conseillé d'utiliser 100 pmol d'extrémités 3' quand nous ne disposons que de 100 fois moins de matériel au départ. Nous avons donc orienté notre choix vers un nouveau système de biotinylation qui permet d'amplifier linéairement l'ADN tout en y incorporant des nucléotides biotinylés.

IV.1.2 Biotinylation par amplification linéaire

Des premiers tests ont été réalisés en utilisant 200 ng d'ADN Promega de souris comme matériel de départ. Le but de cette manipulation étant *in fine* l'hybridation à de l'ADN immunoprécipité de taille 300-1000 pb, nous avons cherché à fragmenter l'ADN qui subit la biotinylation dans la même gamme de tailles. 100 µL d'ADN à 25 ng/µL ont donc été soniqués dans le Bioruptor pendant 12 minutes. La biotinylation a été réalisée par incubation à 37°C pendant 2 heures en comparant l'influence du milieu de sonication (en eau et en TE). Les produits résultants ont été dosés au Nanodrop et déposés sur gel d'agarose 1%. Nous avons également cherché à optimiser les conditions de la biotinylation en variant la quantité d'ADN de départ de 200 ng à 400 ng et en allongeant la durée d'incubation de 2h à 4h (voir Figure 33). Nous en avons conclu qu'une plus faible quantité d'ADN était plus efficacement amplifiée (et donc biotinylée) en incubant sur une durée plus longue.

Ces conditions ont ensuite été utilisées sur de l'ADN Cot-1 de souris, dont la sonication a, au préalable, été optimisée à 1 minute seulement pour obtenir la gamme de taille recherchée, le Cot-1 étant déjà fractionné initialement. Nous avons également doublé la quantité d'enzyme pour améliorer l'amplification, sans amélioration du rendement (voir Figure 33). Enfin, un test sur du Cot-1 non-soniqué n'a pas amélioré la qualité de la biotinylation.

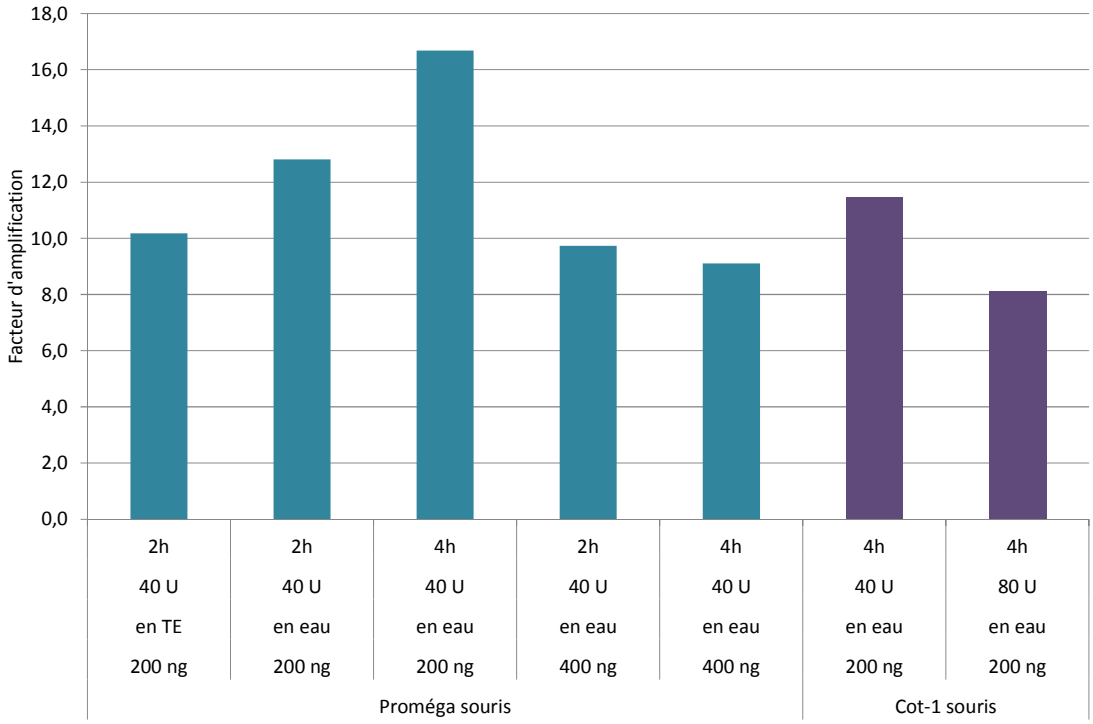


Figure 33: Facteurs d'amplification après biotinylation dans différentes conditions expérimentales
200 et 400 ng sont les quantités initiales d'ADN utilisées, en eau ou en TE désignent le milieu dans lequel elles ont subi la sonication, 40 et 80 U désignent les quantités d'enzyme utilisées pour la biotinylation et 2 ou 4h désignent les temps d'incubation à 37°C.

La biotinylation se fera donc par incubation de 200 ng d'ADN Cot-1 après une courte sonication en eau de 1 minute, avec 40 U de fragment de Klenow à 37°C pendant 4 heures. On obtiendra en moyenne 2,3 µg d'ADN biotinylé, soit un facteur d'amplification de 11,3.

IV.2 Mise au point des PCRs de contrôle

L'optimisation du protocole complet nécessite de pouvoir vérifier à diverses étapes la présence de séquences répétées par PCR ou de les quantifier par qPCR. La conception d'amorces permettant d'amplifier des régions localisées dans les zones hautement répétées du génome est rendue difficile par la nature même de ces régions. Nous avons donc cherché et utilisé des amorces publiées dans la littérature (212) qui ciblent des satellites mineurs et majeurs de l'ADN génomique de souris (*miSat* et *maSat*), ainsi que des éléments transposables (*Line* et *Sine*). Par ailleurs, les séquences d'intérêt que nous désirons conserver seront étudiées par l'analyse d'*Igf2*, gène soumis à l'empreinte qui sera donc immunoprécipité et qui servira de contrôle à la fin de la déplétion. Toutes les séquences d'amorces se trouvent en Annexe 2.

Nous avons adapté le protocole de PCR utilisé de façon classique et élaboré le protocole présenté dans le Tableau 24 pour les différentes familles d'éléments répétés. Nous l'avons validé sur de l'ADN Promega de souris, ainsi que sur ce même ADN immunoprécipité.

Pour 1 échantillon	Volume (µL)	Concentration ou quantité finale	Programme de PCR
10x PCR Buffer	1,0	1x	
$MgCl_2$ (25 mM)	0,7	1,7 mM	95°C 15 min
dNTPs (8 mM)	1,0	0,8 mM	
Amorce sens (10 µM)	0,5	0,5 µM	95°C 30 sec ┐
Amorce anti-sens (10 µM)	0,5	0,5 µM	$T_{hybridation}$ 30 sec ├ × 30
HotStar Taq polymerase (5 U/µL)	0,4	2 U	72 °C 1 min ┘
Eau milliQ	4,9		
ADN	1,0		72 °C 5 min
Volume total	10,0		

Tableau 24: Conditions expérimentales utilisées pour la PCR des éléments répétés

Les températures d'hybridation optimale identifiées sur ces échantillons (68°C pour *Line1Orf2* et 52°C pour *SineB1*, *miSat* et *maSat*) ont été utilisées pour valider ces PCRs sur de l'ADN Cot-1 de souris. Celles-ci ont confirmé la présence de ces séquences bien précises dans l'ADN Cot-1 commercial (résultats confirmés en qPCR). Elles permettront donc de contrôler la fixation de cet ADN biotinylé sur les billes de streptavidine par amplification des surnageants recueillis lors des lavages des billes.

La PCR ne fonctionne pas dans le milieu final du protocole contenant le tampon d'hybridation HB. Une purification sur colonne QIAquick est nécessaire et permet, après PCR, de détecter de faibles quantités d'ADN (jusqu'à 10 ng dilués dans 30 µL d'HB).

IV.3 Couplage de l'ADN Cot-1 biotinylé aux billes de streptavidine

IV.3.1 Optimisation du couplage

Le Cot-1 biotinylé est lié à des billes magnétiques recouvertes de streptavidine. Celles-ci sont d'abord lavées pour être changées de milieu puis le Cot-1 biotinylé y est ajouté. Il est ensuite important d'éliminer le Cot-1 (biotinylé ou non biotinylé résiduel) qui ne se fixe pas aux billes afin qu'il n'interfère pas avec les séquences répétées de notre échantillon dans les PCR de contrôle. Ceci est effectué dans une première série de lavages d'où sont issus les surnageants notés S1a, S1b et ainsi de suite (voir Figure 34). Le Cot-1 est ensuite dénaturé par ajout de NaOH ou par chauffage à 95°C pour l'obtenir en configuration simple brin et permettre l'hybridation dans l'étape suivante, et l'on récupère le surnageant noté S2. Enfin, une dernière série de lavages permet l'élimination de toute trace de Cot-1 résiduel dans des surnageants S3a, S3b et ainsi de suite.

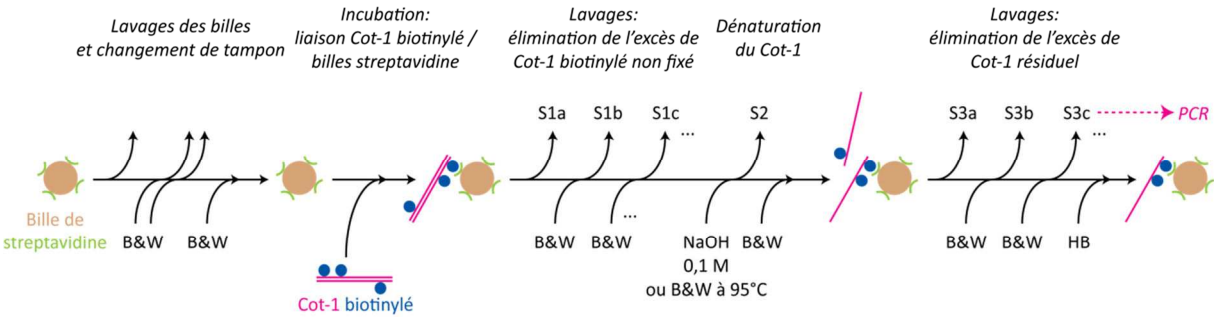


Figure 34: Etapes de lavage pour la liaison du Cot-1 biotinylé aux billes de streptavidine
Les flèches de la partie inférieure indiquent l'ajout de tampon (dont la nature est indiquée ; B&W = binding & washing buffer, dont la concentration sera déterminée plus tard) et la remise en suspension des billes par pipetage. Les flèches de la partie supérieure indiquent le retrait du surnageant par pipetage après aimantation. S : surnageant, HB : tampon d'hybridation (hybridization buffer).

Plusieurs paramètres peuvent influencer la liaison de l'ADN Cot-1 aux billes dont la température des tampons, leur concentration saline, le nombre de lavages ou encore la quantité de Cot-1 introduite. Nous allons faire varier ces paramètres et des PCRs sur des éléments répétés seront réalisées sur les surnageants ainsi isolés (excepté S2 lorsque la dénaturation est effectuée en NaOH) afin de comparer les quantités de Cot-1 dans le milieu après liaison aux billes et de définir les conditions optimales de fixation.

IV.3.1.1 Mode de dénaturation

1,0 mg de billes M-270 est utilisé pour une première approche. Les billes sont lavées 3 fois en tampon B&W 4x puis remises en suspension dans 50 µL de ce même tampon. 50 µL contenant 2,5 µg de Cot-1 biotinylé y sont ajoutés et incubés 15 minutes à TA sous agitation (500 rpm). Le surnageant S1a est alors récupéré puis les billes sont lavées dans le tampon B&W 2x pour donner le surnageant S1b. L'ADN Cot-1 est dénaturé par lavage à la soude 0,1 M et en parallèle par chauffage à 95°C pendant 5 minutes pour donner S2. Un dernier lavage au B&W 2x fournit le S3a avant de resuspendre les billes dans 70 µL de HB.

Une PCR d'un satellite majeur (*maSat*) sur les surnageants obtenus montre, après dépôt sur gel d'agarose, qu'une partie du Cot-1 non fixée sur les billes est libérée lors du premier lavage puisque l'on observe une trainée sur le gel dans les puits correspondant à S1a et S1b. Après dénaturation par la soude, du Cot-1 est à nouveau libéré en plus grande quantité dans le surnageant S3a se traduisant là encore par une trainée sur le gel. Il faut donc ajouter des étapes de lavage avant et après la dénaturation afin d'éliminer tout Cot-1 résiduel qui ne serait pas fixé sur les billes. En revanche, la dénaturation thermique conduit à un décrochage total du Cot-1 biotinylé puisqu'une trainée intense est observée sur le gel dans le puits correspondant à S2 tandis qu'une faible bande est obtenue après le dernier lavage (S3a) ; ce mode de dénaturation est donc écarté.

IV.3.1.2 Quantité d’ADN Cot-1

Nous avons fixé sur les billes de streptavidine une quantité décroissante d’ADN Cot-1 biotinylé (de 2,5 µg à 500 ng) dans les conditions définies ci-dessus tout en incluant deux lavages supplémentaires avant dénaturation (S1c et S1d). L’observation de trainées sur la plupart des surnageants après PCR et dépôt sur gel laisse entendre que du Cot-1, certes en faible quantité, est toujours libéré, y compris dans les surnageants additionnels. Nous avons conclu que la quantité de Cot-1 initiale devait être revue à la baisse et/ou que des lavages supplémentaires étaient à nouveau requis.

IV.3.1.3 Stringence des lavages

1 µg de Cot-1 biotinylé a été fixé sur les billes puis celles-ci ont été lavées 2 fois (S1a à S1c) avant dénaturation (S2) suivie de 3 autres lavages (S3a à S3c). Pour améliorer sensiblement la fixation, nous avons utilisé des conditions plus stringentes en réalisant les lavages à 37°C d’une part et en diminuant la quantité de sel d’autre part (B&W 2x dilué au demi pour donner le B&W 1x contenant 0,05% de Tween-20). Un échantillon contrôle subissant le même protocole avec de l’ADN Cot-1 non biotinylé a également été introduit afin d’apprécier la non-spécificité de la fixation sur les billes.

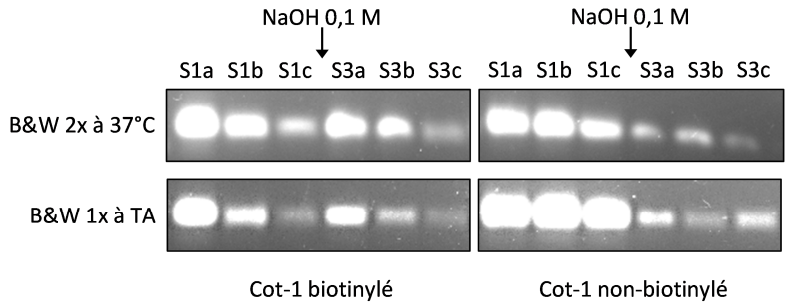


Figure 35: PCR sur les surnageants de lavage lors de la fixation du Cot-1 biotinylé sur les billes
Un élément répété (*Sine*) a été amplifié sur les surnageants pendant le lavage des billes après fixation du Cot-1, dans deux conditions. S2 n’apparait pas car la PCR ne peut être réalisée sur un milieu contenant de la soude (ayant permis de dénaturer l’ADN). TA : Température Ambiante.

Une PCR sur l’élément *Sine* montre bien une diminution de la quantité d’ADN de S1a à S1c (voir Figure 35) puis une ré-augmentation sur S3a. Ceci laisse suggérer un décollement des brins fixés sur les billes de façon non-spécifique et/ou une libération des brins complémentaires dont les biotines n’ont pu être liées à une streptavidine. La quantité d’ADN diminue alors sur les derniers lavages. Notre contrôle, lui, montre que l’ADN non biotinylé peut se fixer sur les billes de façon non-spécifique, mais qu’il est libéré en quasi-totalité au cours des lavages. Les deux conditions ont amélioré de façon identique l’efficacité des lavages. Le tampon B&W 1x sera sélectionné pour les lavages des billes à TA.

Enfin, nous avons quantifié l'ADN Cot-1 biotinylé libéré dans ces conditions en analysant les divers surnageants par qPCR. La majorité du Cot-1 introduit (1,0 µg) se fixe effectivement sur les billes puisque nous n'avons obtenu que 0,28 ng sur l'ensemble des surnageants S1a-c et S3a-c.

IV.3.2 Choix des billes

Les billes utilisées pour déterminer les conditions précédentes (les M-270) n'étant peut-être pas les plus optimales, ce protocole a été mis en œuvre sur 3 autres types de billes magnétiques couvertes de streptavidine se distinguant les unes des autres par les caractéristiques présentées dans le Tableau 25.

Billes magnétiques de streptavidine		M-270 Streptavidin	M-280 Streptavidin	MyOne Streptavidin C1	MyOne Streptavidin T1
Concentration	mg/mL	10			
	billes/mL	$6 - 7 \times 10^8$		$7 - 12 \times 10^9$	
Diamètre des billes (µm)		2,8		1,0	
Capacité de liaison de biotine libre (pmol/mg)		650 - 1350	650 - 900	≥ 2500	1100 - 1700

Tableau 25: Caractéristiques des billes magnétiques de streptavidine

Les billes M-270 semblent augmenter la non-spécificité puisque du Cot-1 est toujours libéré dans les lavages S3a-c (voir Figure 36). De plus, lors des lavages, les billes M-270 semblent être aimantées plus difficilement et sont davantage retenues dans le cône lors de leur remise en suspension. Elles sont donc écartées.

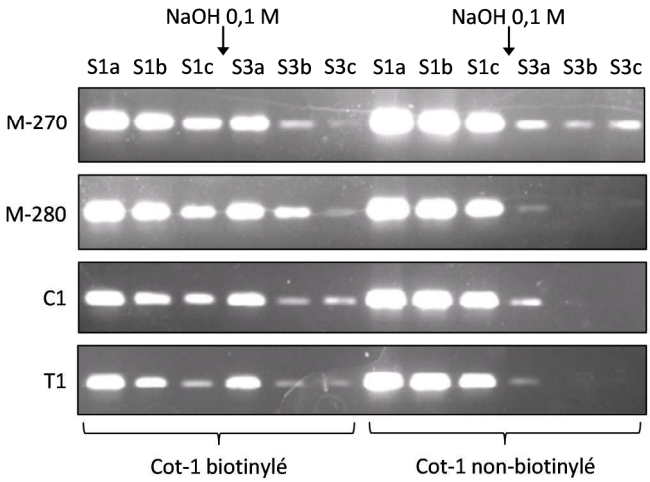


Figure 36: PCR sur les surnageants de lavage lors de la fixation du Cot-1 sur différentes billes

Les billes M-280, C1 et T1 donnent sensiblement les mêmes profils de libération de Cot-1. Des observations expérimentales ont donc guidé notre choix : les billes M-280 présentent l'inconvénient

d'être plus difficiles à laver car retenues dans le cône, tandis que les billes C1 et T1 ont pour avantage de peu sédimenter après incubation avec le Cot-1. Entre ces dernières, les billes possédant la capacité de liaison de biotine libre théorique la plus importante seront conservées: il s'agit des billes C1.

IV.4 Hybridation de l'ADN IP à l'ADN Cot-1

L'échantillon d'ADN d'intérêt immunoprécipité peut maintenant être hybridé à l'ADN Cot-1 fixé sur les billes dans un tampon d'hybridation (voir Figure 37). Ceci permettra, après incubation, de maintenir dans le tube une partie des séquences répétées hybridées au Cot-1 grâce à une aimantation des billes, tandis que le surnageant contenant l'échantillon (dont la quantité de séquences répétées s'en trouve diminuée) sera isolé et purifié sur colonne QIAquick. Des qPCR de divers éléments répétés permettront de quantifier la déplétion dans l'échantillon.

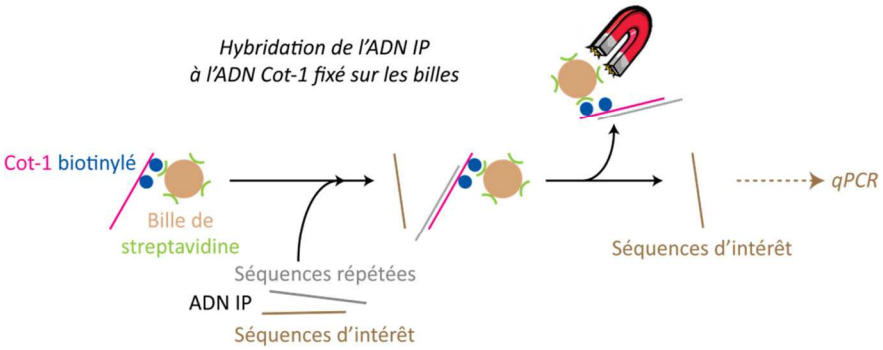


Figure 37: Hybridation de l'échantillon d'ADN IP au Cot-1 fixé sur les billes

Le facteur de déplétion en éléments répétés est mesuré par qPCR d'une séquence répétée telle que l'élément *Line*. Le Ct obtenu est comparé à celui d'un contrôle négatif (Ct_0, voir Figure 38) qui subit tout le protocole sans introduction de Cot-1 biotinylé, selon la formule suivante :

$$\text{facteur de déplétion} = 2^{(Ct-Ct_0)}$$

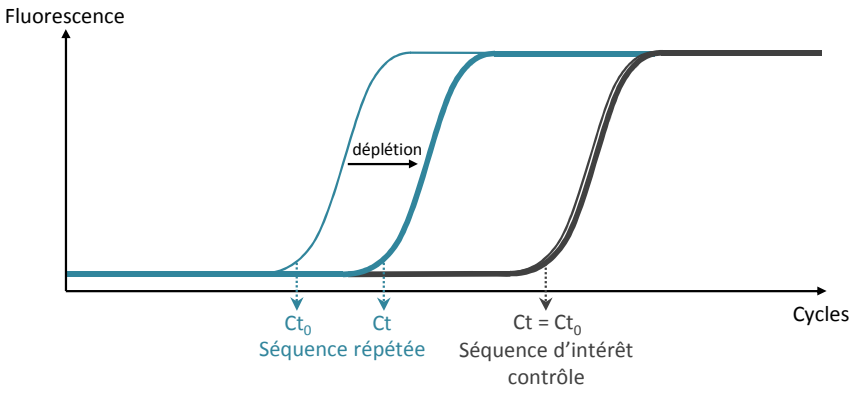


Figure 38: Comparaison des Cts obtenus par qPCR pour la mesure du facteur de déplétion

La même formule appliquée aux Ct recueillis par qPCR sur *Igf2* aura pour but de confirmer que les séquences d'intérêt ne sont pas influencées par ce protocole et leur facteur de déplétion aura idéalement la valeur de 1,0 ($Ct = Ct_0$, voir Figure 38).

Nous chercherons les quantités d'ADN (tant de Cot-1 que de l'échantillon d'ADN IP) qui mènent à une déplétion importante. Nous chercherons également la température et le temps d'incubation au Cot-1 optimaux et nous envisagerons d'améliorer l'hybridation par le blocage préalable des billes avec de la BSA ou un ARN *carrier*. La mise au point du protocole nécessite l'utilisation d'une grande quantité d'ADN immunoprécipité, elle est donc réalisée sur de l'ADN Promega IP de souris ayant subi une WGA (*Whole Genome Amplification*, facteur d'amplification de 59,1 en moyenne).

IV.4.1 Température d'hybridation au Cot-1

Après liaison de l'ADN Cot-1 biotinylé aux billes de streptavidine, l'échantillon d'intérêt est ajouté au billes resuspendues en HB dans un volume total de 100 µL. Les tests d'hybridation sont réalisés sur les billes C1 et nous cherchons d'abord à préciser la température d'incubation optimale. 50 ng d'IP amplifié par WGA sont hybridés à 1 µg d'ADN Cot-1 biotinylé préalablement fixé sur les billes, par incubation de 63 à 67°C sur la nuit, sous agitation à 1000 rpm. Nous avons observé que cette vitesse de rotation minimale était requise pour ne pas provoquer la sédimentation des billes et l'avons ensuite augmenté à 1200 rpm pour une meilleure homogénéisation du milieu. Le lendemain, le surnageant est isolé par pipetage après aimantation puis purifié sur colonne QIAquick et analysé par qPCR sur l'élément *Line* et le contrôle *Igf2*.

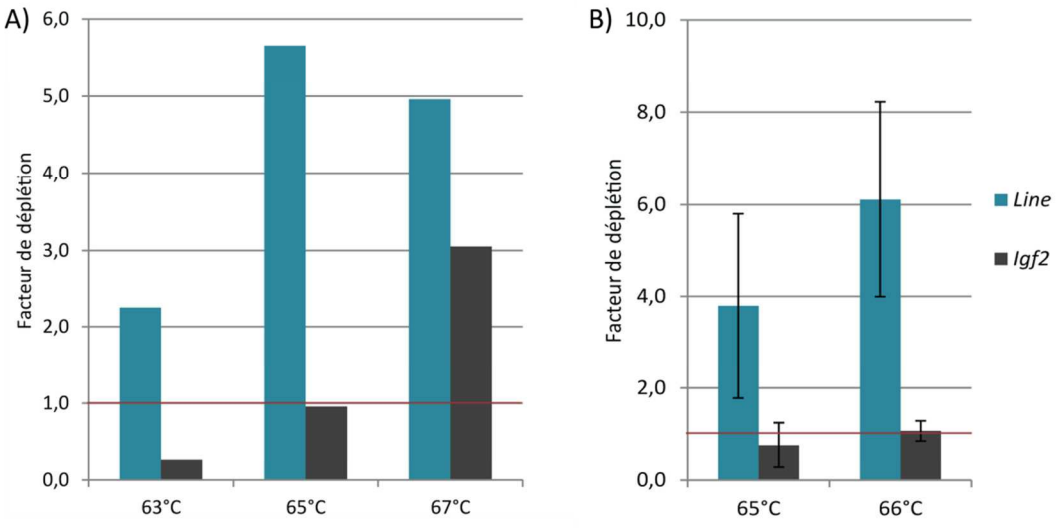


Figure 39: Facteurs de déplétion obtenus par hybridation à différentes températures
La ligne rouge horizontale indique un facteur de déplétion égal à 1,0 : l'élément concerné est conservé. Idéalement, le contrôle *Igf2* aura un facteur de déplétion égal à 1,0.

Une augmentation de la température d'incubation de 63 à 65°C améliore le facteur de déplétion de l'élément répété étudié (voir Figure 39A). De 65°C à 67°C elle contribue cependant à une déplétion des séquences d'intérêt. Nous avons donc affiné la recherche de cette température en réalisant l'hybridation à une température intermédiaire de 66°C. La déplétion est ainsi plus importante tout en permettant de conserver les séquences d'intérêt (voir Figure 39B). Cette température d'hybridation sera donc utilisée pour la suite des tests.

IV.4.2 Quantités d'ADN optimales

Nous avons choisi la quantité d'ADN immunoprécipité qui a été hybridée sur l'ADN Cot-1 de façon arbitraire (50 ng) afin que celui-ci soit en large défaut par rapport à la quantité d'ADN Cot-1 sur les billes (1 µg). Nous avons donc dans un second temps fait varier la quantité d'ADN IP, de 25 à 100 ng (voir Figure 40A). Les séquences d'intérêt représentées par *Igf2* ne sont pas affectées par la variation de cette quantité. En revanche, l'élément répété étudié ici est éliminé d'un facteur d'autant plus grand que ne l'est cette quantité, nous hybriderons donc 100 ng d'ADN IP.

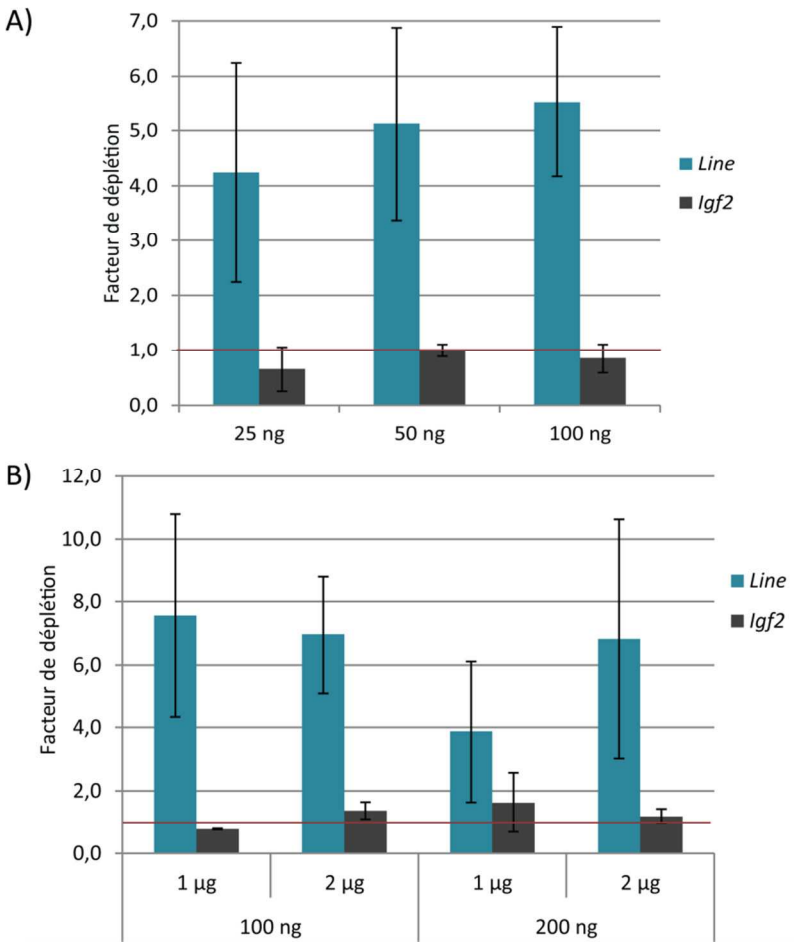


Figure 40: Facteurs de déplétion obtenus par hybridation de quantités croissantes d'ADN IP
A) Hybridation d'une quantité croissante d'ADN IP sur 1 µg d'ADN Cot-1. B) 1 et 2 µg indiquent les quantités d'ADN Cot-1 biotinylé qui ont d'abord été fixées sur les billes de streptavidine. 100 et 200 ng indiquent les quantités d'ADN IP qui ont ensuite été hybridées sur cet ADN Cot-1.

Nous avons ensuite pensé que le ratio entre ces deux quantités d'ADN (échantillon IP et Cot-1 biotinylé) était important et avons donc croisé les conditions suivantes : 100 et 200 ng d'ADN IP ont été hybridés sur 1 et 2 µg d'ADN Cot-1 biotinylé au préalable fixé sur les billes (voir Figure 40B). Un ratio 1 : 10 entre IP et Cot-1 (100 ng pour 1 µg et 200 ng pour 2 µg respectivement) parait optimal puisque l'élément répété étudié est éliminé en quantité importante tandis que la séquence d'intérêt n'est pas affectée. Une augmentation de la quantité de Cot-1 biotinylé (ratio 1 : 20) n'améliore pas le rendement puisque les mêmes résultats sont obtenus tandis que la diminution de celui-ci (ratio 1 : 5) désavantage la déplétion et contribue à l'élimination de séquences d'intérêt. Nous conserverons donc un ratio 1 : 10 en hybridant 100 ng d'ADN IP sur 1 µg de Cot-1 biotinylé.

IV.4.3 Température d'hybridation au Cot-1 (2)

Malgré les bons résultats de déplétion obtenus jusqu'ici, nous nous sommes interrogés sur l'éventuelle fragilité de la liaison biotine/streptavidine et donc de la fixation du Cot-1 sur les billes, lors de l'hybridation. En effet, le système biotine/streptavidine est connu pour être l'une des plus fortes interactions biologiques non-covalentes. Cependant, il a été montré qu'une courte incubation à une température avoisinant les 70°C, donc proche de notre température de travail, pouvait rompre cette interaction (213). Ceci impliquerait la libération de Cot-1 dans le milieu : le Ct mesuré serait donc biaisé puisque plus faible (car davantage de matériel) et donc plus proche de notre contrôle négatif Ct_0, résultant en une déplétion plus faible.

Pour vérifier cette hypothèse, nous avons réalisé l'hybridation à une température bien plus basse que celles testées auparavant, soit 62°C (en comparaison à 66°C). Nous avons également introduit d'autres éléments répétés pour le contrôle de la déplétion par qPCR, à savoir *Sine*, *miSat* et *maSat* (voir Figure 41).

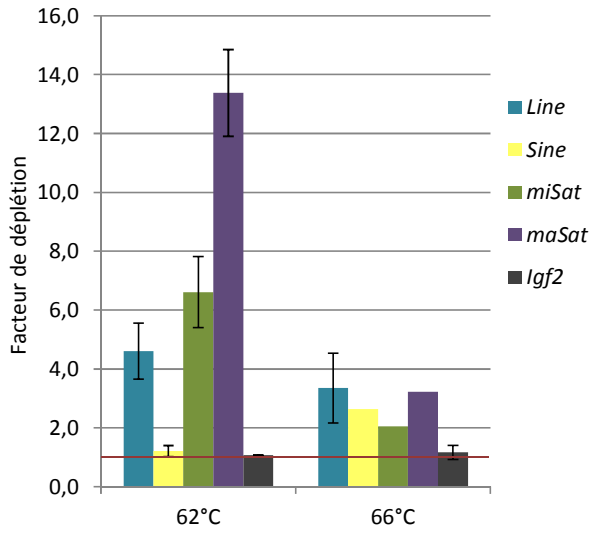


Figure 41: Facteurs de déplétion obtenus par hybridation à 62 et 66°C

Les facteurs de déplétion se sont avérés bien plus intéressants en diminuant la température d'hybridation. De plus, un échantillon contrôle contenant uniquement du Cot-1 (étape d'hybridation à 62°C réalisée avec de l'eau, sans ADN IP) a fourni un Ct bien plus tardif que celui de notre échantillon (8,8 cycles d'écart en moyenne). Ceci montre qu'à 62°C, l'ADN Cot-1, s'il se décroche des billes lors de l'hybridation, l'est en quantité négligeable. La température d'hybridation de 62°C sera donc conservée pour la suite des tests.

IV.4.4 Blocage des billes

L'échantillon d'ADN IP est retenu sur les billes par hybridation au Cot-1 mais on peut envisager qu'il se fixe de façon non-spécifique à la surface des billes. Ceci biaiserait le processus de déplétion puisque cette fraction d'ADN ne s'hybriderait pas sur le Cot-1 et pourrait être libérée par la suite. De plus, ce mécanisme, s'il existe, se trouverait intensifié dans le cas de notre contrôle négatif (qui fournit le Ct_0 avec lequel comparer notre échantillon) puisque celui-ci ne contient pas de Cot-1 biotinylé et que la surface des billes en est d'autant plus accessible.

Afin de vérifier cette hypothèse, nous avons ajouté 500 ng d'ARN *carrier* (Qiagen) dans le tampon B&W 1x du lavage qui suit la dénaturation du Cot-1 sur les billes afin de bloquer, voire saturer, la surface des billes avant l'ajout de l'ADN IP.

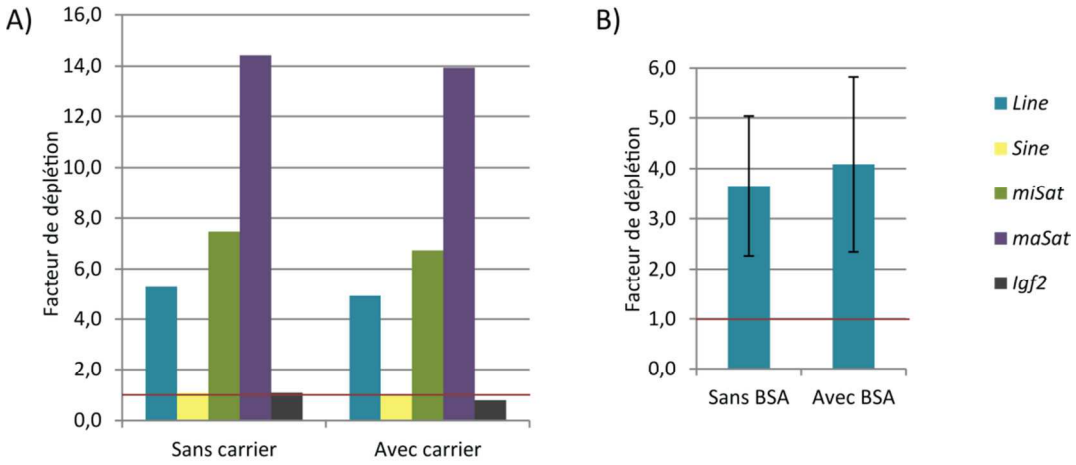


Figure 42: Facteurs de déplétion obtenus avec et sans blocage des billes
A) 500 ng de carrier RNA ont été ajoutés dans la dernière série de lavage (protocole utilisant 1 µg d'ADN Cot-1 biotinylé, 100 ng d'ADN IP, et une hybridation à 62°C). B) 0,1% de BSA a été ajouté dans le tampon B&W 4x de lavage des billes (test réalisé avec le protocole initial utilisant 1 µg d'ADN Cot-1 biotinylé, 50 ng d'ADN IP et une hybridation à 66°C).

Les valeurs de Ct brutes obtenues et utilisées pour le calcul du facteur de déplétion ont été identiques avec et sans ARN *carrier* et un contrôle contenant seulement du *carrier* sans ADN IP a donné un Ct confondu avec celui de l'eau de qPCR. Celui-ci n'impacte donc pas la mesure des Ct de nos échantillons. Aucune amélioration du facteur de déplétion n'a pu être notée par l'ajout de cette

condition de blocage (voir Figure 42A). De plus, on observe que l'élément répété *Sine* ne parait pas impliqué dans la déplétion.

De même, lors des tests préliminaires, 0,1% de BSA a été ajouté dans le tampon B&W 4x de lavage dans le but de saturer les billes, et aucune amélioration du rendement n'a été apportée (voir Figure 42B). Nous ne bloquerons donc pas les billes par l'une ou l'autre des voies envisagées pour la suite de l'optimisation du protocole.

IV.4.5 Durée d'hybridation

Les tests d'hybridation réalisés jusqu'ici ont été effectués sur la nuit (soit environ 15h d'incubation). Nous avons cherché à savoir si ce temps pouvait être diminué tout en conservant les rendements de déplétion obtenus. Nous avons donc comparé ces résultats à ceux obtenus lors d'une hybridation de 3h et de 6h. En parallèle nous avons également réalisé une double hybridation dans laquelle le surnageant résultant de la première hybridation de 3h a subi une seconde hybridation de 3h sur de nouvelles billes portant du Cot-1 (voir Figure 43).

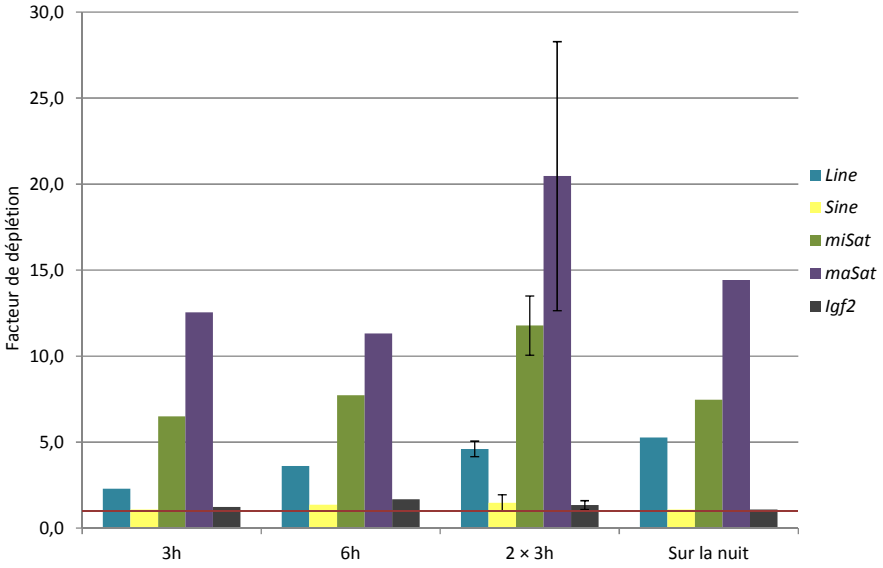


Figure 43: Facteurs de déplétion obtenus avec des durées d'hybridation croissantes

Les 3 hybridations de 3h, 6h et 15h ont provoqué des déplétions sensiblement identiques de divers éléments répétés. En revanche, une nette augmentation du facteur de déplétion a été observée lors de la double hybridation de 2 × 3h, en particulier sur les séquences satellites. Nous noterons néanmoins que l'élément *Sine* n'est, là encore, pas affecté par la déplétion et nous ne l'utiliserons plus comme contrôle par la suite.

Au-delà de 3h d'incubation, l'hybridation semble donc atteindre un équilibre. On peut alors imaginer que toutes les séquences répétées pouvant s'hybrider sur le Cot-1 le sont dans ces premières heures

d’hybridation et que l’allongement de la durée d’incubation est inutile. Or l’introduction d’une seconde hybridation permet d’éliminer une plus grande quantité d’éléments répétés dans l’échantillon. Ceci montre donc qu’il reste du matériel après la première hybridation et laisse supposer que l’ADN Cot-1 disponible n’est plus suffisant ou accessible pour l’hybridation de l’ADN IP, d’où la nécessité de préparer de nouvelles billes portant du Cot-1 et de réaliser à nouveau le protocole. Nous conserverons donc cette double hybridation de 2 × 3h.

IV.4.6 Hybridation à de l’ADN IP non amplifié

La mise au point du protocole a été réalisée sur de l’ADN Promega de souris immunoprécipité amplifié par WGA. A terme, l’échantillon ne sera pas amplifié et nous cherchons donc à valider le protocole réunissant toutes les conditions définies ci-avant (1 µg de Cot-1 biotinylé, 100 ng d’ADN IP, hybridation de 2 × 3h à 62°C) sur l’ADN Promega IP non amplifié.

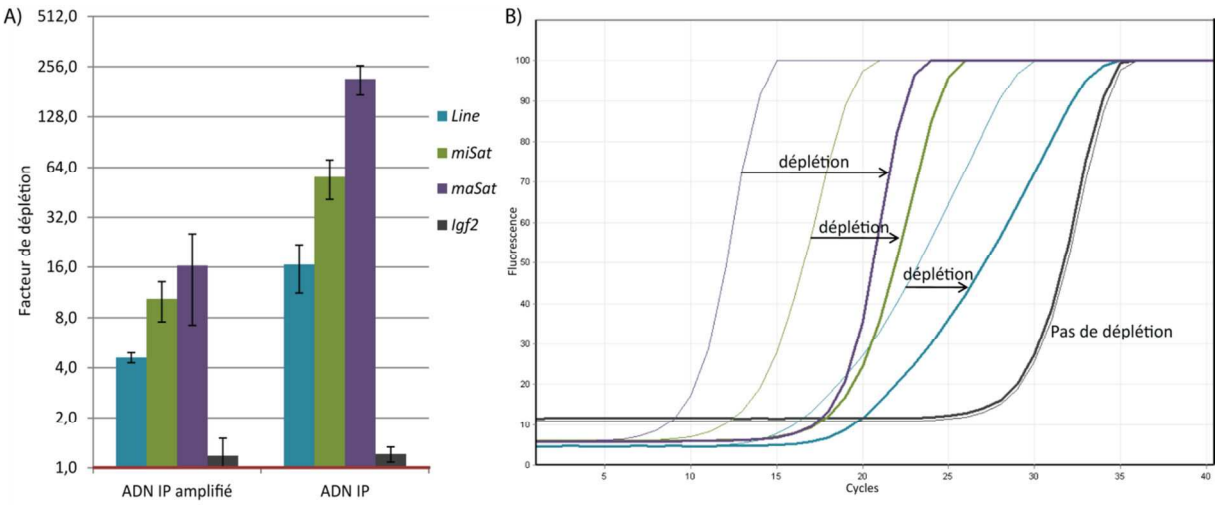


Figure 44: Facteurs de déplétion obtenus sur un échantillon d'ADN IP non amplifié
A) Nous comparons la déplétion réalisée sur un échantillon d’ADN IP amplifié par WGA et sur le même échantillon IP non amplifié. Il faut noter que l’échelle utilisée ici est logarithmique. La ligne rouge horizontale à 1,0 indique là-encore que la séquence concernée est conservée. B) Courbes de qPCR obtenues pour le contrôle de la déplétion sur l’ADN IP non amplifié. Les courbes fines représentent les contrôles négatifs, les courbes plus épaisses leur déplétion correspondante.

Le protocole réalisé sur de l’ADN non amplifié permet d’obtenir une déplétion en éléments répétés beaucoup plus importante, et on atteint même largement le facteur 200 pour le satellite majeur *maSat*, tout en conservant les séquences d’intérêt représentées par *Igf2* (voir Figure 44). Le protocole de déplétion ainsi défini peut donc s’appliquer à un échantillon d’ADN directement après le MeDIP. L’échantillon IP, à l’état simple brin, sera toutefois chauffé à 95°C pendant 3 minutes avant son hybridation au Cot-1 pour s’assurer de la dénaturation de brins éventuellement réappariés.

IV.5 Introduction dans le processus de séquençage

Dans le chapitre précédent, nous avons établi un protocole de préparation d'échantillon pour le MeDIP-Seq. La déplétion des séquences répétées est désormais introduite dans cette procédure : l'échantillon de Promega est fragmenté puis les adaptateurs sont ajoutés aux extrémités des fragments, l'ADN est ensuite immunoprécipité puis il subit la déplétion. 25 µL de l'ADN purifié, soit la majorité de ce que nous obtenons après déplétion, subissent ensuite la renaturation, puis le *sizing* sur E-gel et l'amplification par PCR.

IV.5.1 Enzyme utilisée pour l'amplification par PCR

Contrairement à ce qui fut le cas pour la mise au point du MeDIP-Seq, aucune trace d'ADN n'a pu être observée sur l'E-gel pendant la migration et, par la suite, aucun produit n'a pu être généré par 10 à 14 cycles de PCR (nombre de cycles utilisés dans les premiers essais de préparation d'échantillon pour le MeDIP-Seq). Nous avons alors testé une nouvelle enzyme, la polymérase Phusion High Fidelity (NEB), en comparaison avec l'enzyme utilisée jusqu'alors (Platinum Pfx) en augmentant le nombre de cycles de PCR à 18.

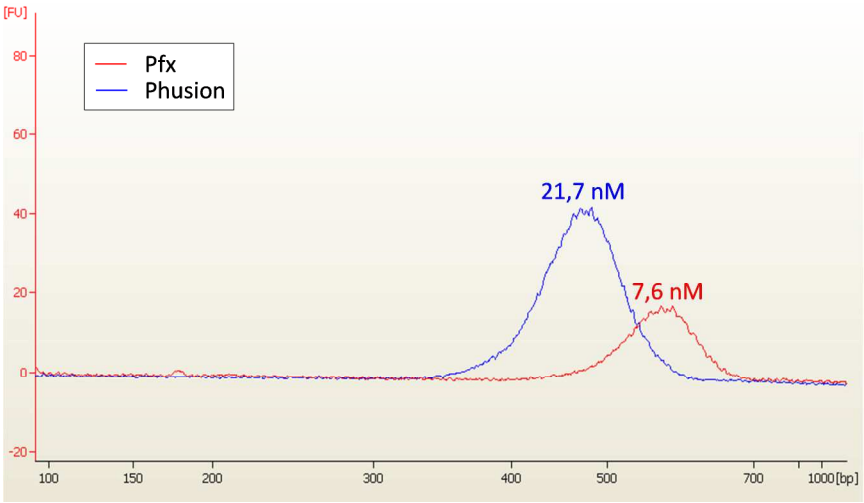


Figure 45: Profils obtenus après 18 cycles de PCR dans le cadre du MeDIP-Seq avec déplétion
En rouge, la PCR a été réalisée avec la Pfx sur l'ADN prélevé sur E-gel à une taille théorique de 450 pb. En bleu, la PCR a été réalisée avec la Phusion sur le même échantillon prélevé sur le même E-gel à une taille théorique de 400 pb. Les molarités obtenues sont indiquées au-dessus du pic correspondant.

La polymérase Phusion a conduit à une meilleure amplification (21,7 nM en comparaison à 7,6 nM, voir Figure 45) et nous l'utiliserons donc préférentiellement dans le protocole de préparation des échantillons qui subissent une déplétion après MeDIP. Il faut toutefois noter que ces tests ont été réalisés avec de l'ADN immunoprécipité par une quantité non-optimisée d'anticorps (voir chapitre III) et donc en plus grande quantité que celle dont nous disposerons dans le protocole final. On envisage

donc d'augmenter d'une part la quantité d'ADN IP utilisée pour la déplétion, et d'autre part le nombre de cycles de la PCR.

IV.5.2 Adaptation des quantités d'ADN nécessaires dans la déplétion

Il nous faut augmenter la quantité d'ADN IP afin d'obtenir suffisamment de matériel pour la suite du protocole. Nous avons déjà établi qu'un ratio 1 : 10 entre les quantités d'ADN IP et d'ADN Cot-1 biotinylé était optimal pour la déplétion. Ces tests ont été réalisés sur de l'ADN IP amplifié par WGA, aussi avons-nous vérifié que ceci était toujours valable lorsque l'on augmente ces quantités et que l'on utilise de l'ADN immunoprécipité non amplifié.

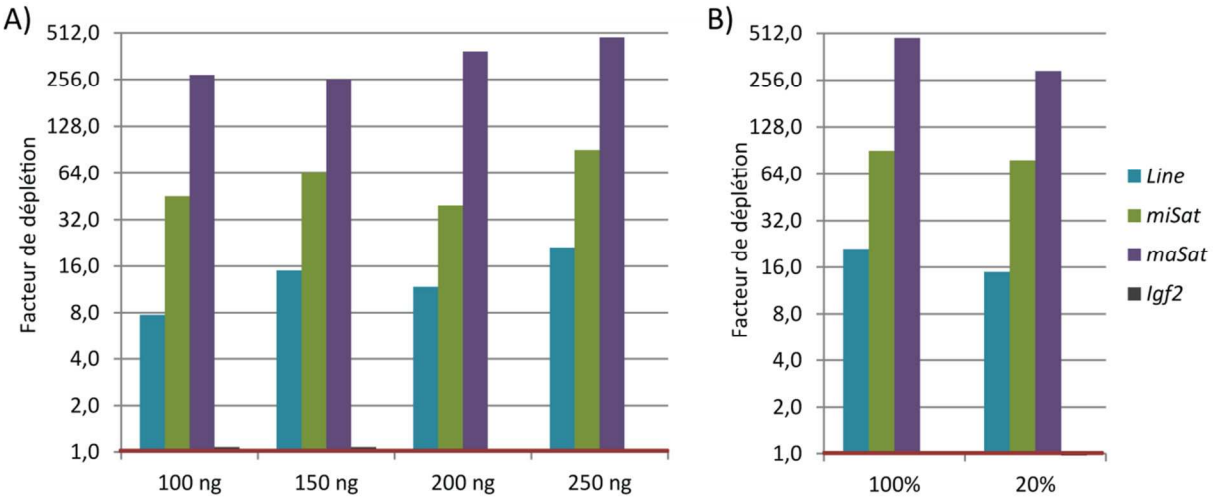


Figure 46: Facteurs de déplétion obtenus avec des quantités variables d'ADN IP
A) Facteurs obtenus avec des quantités croissantes d'ADN IP non-amplifié. B) Facteurs obtenus en réduisant la quantité d'ADN IP pour le contrôle négatif. 100% et 20% désignent la proportion d'ADN IP utilisée dans l'échantillon contrôle négatif qui fournit le Ct_0 pour le calcul de la déplétion. Le ratio ADN IP : Cot-1 biotinylé est de 1 : 10.

Les mêmes ordres de grandeur de facteurs de déplétion ont été obtenus en augmentant la quantité d'ADN IP introduite, tout en conservant le rapport de 1 : 10 avec la quantité de Cot-1 biotinylé (voir Figure 46A). Nous pourrons donc utiliser davantage d'ADN, soit 250 ng, pour la déplétion. Cependant, cette quantité ne sera plus disponible pour réaliser le contrôle négatif en parallèle (fournissant le Ct_0). Nous avons donc vérifié qu'en n'utilisant que 20% de la quantité d'ADN IP pour ce contrôle négatif, il était possible de contrôler le succès de la déplétion à l'aide de la formule suivante :

$$\text{facteur de déplétion} = 2^{[Ct-(Ct_0-2,32)]}$$

Le facteur de correction 2,32 est appliqué puisque seul $1/5^e$ de l'échantillon est utilisé. Les facteurs de déplétion ainsi obtenus apparaissent à peine plus faibles (voir Figure 46B). Ils permettront néanmoins de réaliser un contrôle-qualité correct de la déplétion avant de poursuivre la préparation de l'échantillon dans le but de le séquencer. 25 µL du produit de déplétion ont ensuite été

réintroduits dans le protocole de séquençage et utilisés pour l'amplification par PCR : 22 cycles ont été nécessaires pour obtenir une quantité de matériel suffisante, et ce de façon reproductible. Ces conditions seront conservées pour l'introduction du protocole en routine.

IV.6 Automatisation du protocole

La préparation des billes et la liaison de l'ADN Cot-1 biotinylé à leur surface sont des étapes de préparation longues et répétitives, donc sources d'erreur. Nous les avons donc automatisées sur le robot utilisé pour le MeDIP (SX-8G IP-Star, Diagenode). Les tampons de lavages et les billes sont distribués comme indiqué dans le Tableau 26 dans une barrette de 12 tubes de 0,2 mL qui se positionnera dans le bloc Peltier du robot de la même manière que pour le MeDIP.

Tubes n°	Tampons / réactifs	Volume (µL)
1	Billes MyOne Streptavidin C1	100
2	Tampon B&W 4x	100
3	Tampon B&W 4x	100
4	Tampon B&W 4x	100
5	Tampon B&W 2x + ADN Cot-1 biotinylé 50 ng/µL (ou eau pour le contrôle)	50 + 50
6	Tampon B&W 1x	100
7	Tampon B&W 1x	100
8	NaOH 0,1 M	100
9	Tampon B&W 1x	100
10	Tampon B&W 1x	100
11	Tampon B&W 1x	100
12	Tampon d'hybridation	37,5

Tableau 26 : Distribution des réactifs dans la barrette de tubes pour la déplétion automatisée

Les billes sont d'abord lavées trois fois en tampon B&W 4x puis remises en suspension dans le tube contenant l'ADN Cot-1. Après 15 minutes d'incubation sous agitation par aspiration/refoulement, les billes sont à nouveau lavées, l'ADN Cot-1 qui y est fixé est dénaturé par un lavage à la soude puis trois derniers lavages précèdent la resuspension dans le tampon d'hybridation. Ceci est réalisé en 50 minutes par le robot et en parallèle sur 8 échantillons quand plus d'une heure sera nécessaire à l'expérimentateur pour effectuer le même travail à la main. Des tests préliminaires n'ont pas montré de différence entre les protocoles automatisé et manuel.

IV.7 Protocole final de déplétion

Les développements effectués ont conduit à établir un protocole final dont les paramètres sont représentés sur la Figure 47.

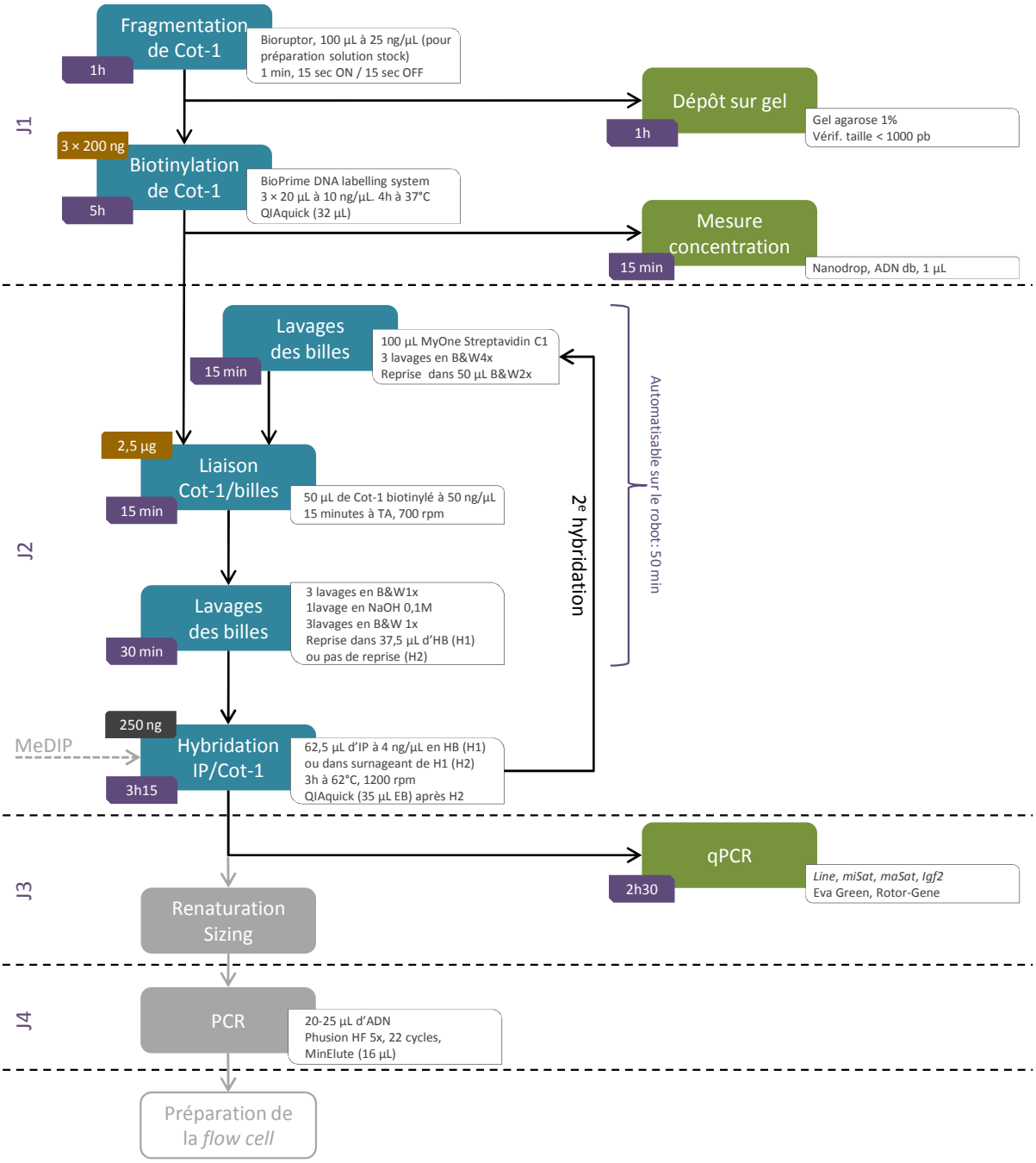


Figure 47: Protocole final de préparation de l'échantillon pour la déplétion
Le temps indiqué est la durée nécessaire à la préparation d'un seul échantillon. Le volume indiqué entre parenthèses après le mode de purification est celui de l'élution. En gris : introduction dans le protocole de MeDIP-Seq. J : jour, ADN db : mesure en mode ADN double brin, H1 : 1ère hybridation, H2 : 2e hybridation.

IV.8 Discussion

Après avoir mis au point un protocole robuste de MeDIP-Seq, nous avons souhaité y introduire une étape permettant de réduire considérablement la quantité de régions répétées obtenue après MeDIP. En effet, ces séquences sont majoritairement méthylées et se trouvent donc

immunoprécipitées. Ceci a été réalisé grâce à un protocole que nous avons baptisé MeDIP-dep-Seq qui inclut une déplétion de ces régions par hybridation à de l'ADN Cot-1.

Chaque étape de ce protocole a fait l'objet d'une étude détaillée pour en optimiser toutes les conditions. La méthode et certains paramètres expérimentaux ont été inspirés par une technique publiée il y a une quizaine d'années, par conséquent dans un contexte tout autre que celui du MeDIP, et dans laquelle l'échantillon était d'abord hybridé à l'ADN Cot-1 biotinylé avant d'être sélectionné par les billes magnétiques de streptavidine (214). Le protocole n'a par la suite été utilisé qu'en 2006 en remplaçant la sélection sur billes par une capture à l'avidine suivie d'une précipitation au phénol/chloroforme (215). Une alternative aurait été d'utiliser une colonne VECTREX (Vector Laboratories), matrice polymérique conjuguée à de l'avidine, pour isoler les fragments non-hybridés au Cot-1 par chromatographie d'affinité. Nous nous sommes basés sur le protocole initial qui utilise des billes magnétiques, moyen de purification le plus rapide et pratique, en modifiant la succession des étapes, ce qui nous a permis d'introduire de nombreux lavages pour nous assurer qu'aucune trace de Cot-1 résiduel ne perturberait la suite de l'expérimentation.

Nous nous sommes d'abord concentrés sur le moyen de biotinyler l'ADN Cot-1. Une biotinylation des extrémités 3' n'a pas donné de résultats concluants. Nous aurions également pu envisager une biotinylation par déplacement de brêche (*nick translation*) (216). Cette technique consiste à ouvrir de façon aléatoire l'un des deux brins du fragment d'ADN et à remplacer grâce à l'ADN polymérase I les nucléotides présents en 3' de cette brêche par des nucléotides marqués, tandis que l'activité exonucléase de l'enzyme hydrolyse ceux qui se trouvent en 5'. Nous avons cependant opté pour une méthode qui a pour avantage, en plus de l'incorporation des nucléotides biotinylés, d'amplifier l'ADN Cot-1 grâce au fragment de Klenow de la polymérase I.

Nous avons ensuite déterminé les meilleures conditions de lavages des billes, intervenant à diverses reprises dans le protocole, et de la fixation du Cot-1 biotinylé à leur surface. Nous avons estimé à trois mois la durée de conservation à 4°C du Cot-1 après biotinylation, après avoir observé que, passé ce délai, les résultats de la déplétion étaient parfois aléatoires. Nous pourrions mettre en place des tests cinétiques pour préciser cette durée de vie ou encore envisager de préparer une grande quantité de billes portant du Cot-1 biotinylé et étudier la stabilité de cette solution stock. Ceci permettrait de réduire considérablement la durée du protocole dont la préparation des billes occupe une place majoritaire. Nous avons d'ailleurs, dans ce but, déjà mis en place l'automatisation de ces étapes très répétitives sur un robot, permettant ainsi à l'expérimentateur de gagner deux heures qui pourront être mises à profit dans des tests ou travaux divers dans le laboratoire.

Nous avons également cherché les quantités de matériel, ADN Cot-1 biotinylé comme ADN IP, qui permettraient d'aboutir à la déplétion la plus importante. Comme évoqué dans le chapitre précédent, nous avons dû faire face à une difficulté de quantification de l'ADN IP. La quantité optimale à utiliser ayant été fixée à 250 ng, mesure donnée par le Nanodrop, nous avons considéré qu'après tout MeDIP la concentration de l'échantillon était de 4,0 ng/µL, moyenne obtenue sur tous nos échantillons, et avons donc utilisé 62,5 µL d'ADN IP de façon systématique dans la déplétion.

Plusieurs durées d'hybridation de l'ADN IP au Cot-1 ont ensuite été envisagées. C'est en enchainant deux hybridations de durées plus courtes (3 heures, comparées à une hybridation sur la nuit) que nous avons obtenu une amélioration conséquente de la quantité de séquences répétées éliminées. Il est à envisager de réduire le temps nécessaire au protocole en diminuant cette durée à deux heures, voire une heure, ou d'améliorer davantage son rendement en ajoutant une troisième hybridation au Cot-1. Ceci sera néanmoins à contrebalancer par le fait que nous diposerons alors d'une quantité de matériel encore plus faible à l'issue du protocole, ce qui nous amènera certainement à adapter à nouveau les étapes de *sizing* et d'amplification par PCR qui précèdent le séquençage. Nous avons effectivement déjà dû rechercher une nouvelle enzyme de PCR, la Phusion, et augmenter le nombre de cycles à 22 pour obtenir une quantité correcte après amplification. D'autres enzymes sont également disponibles et utilisées dans la préparation de librairies de séquençage et pourront être testées comme l'AccuPrime DNA Polymerase High Fidelity (Life Technologies) et son équivalent permettant d'amplifier les régions riches en CpG (AccuPrime GC-rich DNA Polymerase) (217) ou encore la Kapa HiFi (Kapa Biosystems) (218). Notre protocole ainsi développé est cependant adéquat. Il n'allonge le planning d'expérimentation que de deux jours avant la préparation de la *flow cell* pour le séquençage.

Afin d'apprécier le succès du protocole développé, nous avons recherché dans la littérature des séquences d'amorces permettant d'amplifier certaines familles d'éléments répétés puisque celles-ci sont difficiles à concevoir. Nous avons ainsi pu vérifier que notre protocole de déplétion diminue la quantité de l'élément transposable *Line* étudié d'un facteur 20 environ, d'un facteur 4 fois plus élevé pour le satellite mineur concerné et quasiment 15 fois plus important pour le satellite majeur ciblé. Il serait intéressant de quantifier précisément ces différentes familles dans l'ADN Cot-1 utilisé afin de savoir si leurs disparitions respectives après la déplétion sont directement liées à leur présence initiale dans cet ADN, hypothèse qui nous semble la plus probable.

Enfin, un argument de taille en faveur de la déplétion, qui ne saura que convaincre un peu plus de la mettre en place, réside dans son coût. La préparation d'un échantillon et son séquençage en *paired-end* sur 101 bases dans le cadre du MeDIP-Seq voient la facture s'élever à 1506 € (voir calcul en Annexe 1). Il ne faudra débourser que 4,2% de cette somme, soit 64 € supplémentaires, pour

introduire la déplétion des séquences répétées dans la préparation de l'échantillon et mener au MeDIP-dep-Seq. Il est indéniable que ce montant est négligeable, surtout lorsque l'on prend en compte les gains qualitatif et quantitatif que ce protocole apportera et que nous allons découvrir dans le chapitre suivant.

Chapitre V: Du MeDIP-Seq au MeDIP-dep-Seq

Dans le chapitre III, nous avons abouti à la mise en place d'un protocole de MeDIP-Seq qui peut désormais être utilisé en routine sur la plateforme de séquençage. Dans le chapitre IV, nous avons introduit et développé un protocole innovant baptisé MeDIP-dep-Seq, permettant la diminution de la quantité de séquences répétées après le MeDIP et avant le séquençage. Ce dernier protocole fait ici l'objet d'une preuve-de-principe sur des échantillons de MEFs (*Mouse Embryonic Fibroblasts*) et d'une comparaison avec le protocole standard de MeDIP-Seq sur les mêmes échantillons, afin d'en déterminer l'efficacité.

V.1 MeDIP et déplétion

4 échantillons de MEFs (nommés 6.2, 6.3, 8.2 et 8.3, voir Tableau 6) ont suivi le protocole de MeDIP-Seq : ils ont d'abord été fragmentés puis des adaptateurs ont été ajoutés aux extrémités des fragments. Les régions méthylées de chaque échantillon ont ensuite été immunoprécipitées par MeDIP. Tout ceci a été réalisé en doublon sur chaque échantillon et les duplicats ont été réunis à la fin de chaque étape (fragmentation, ligation, MeDIP) puis séparés à nouveau pour l'étape suivante. Ceci permettra une comparaison optimale entre le premier duplicat qui sera utilisé pour le MeDIP-Seq, et le second auquel nous appliquerons la déplétion des séquences répétées et qui sera donc utilisé pour le MeDIP-dep-Seq.

	6.2	6.3	8.2	8.3
Taux de récupération séquences méthylées	44,6%	45,6%	NA	33,7%
Taux de récupération séquences non-méthylées	0,8%	0,3%	NA	0,1%
Quantité après MeDIP	0,43 µg	0,40 µg	0,53 µg	0,44 µg

Tableau 27: Taux de récupération et quantités obtenus après MeDIP de 4 MEFs
Taux de récupération des séquences méthylées et non-méthylées mesurés en qPCR. Quantités après MeDIP mesurées au Nanodrop, en mode simple brin.

Nous avons obtenu en moyenne 0,45 µg d'ADN immunoprécipité avec un taux de récupération des séquences méthylées de 41,3% (voir Tableau 27). Chaque échantillon (la moitié du volume obtenu après MeDIP) a ensuite suivi le protocole de déplétion dans les conditions que nous avons mises au point.

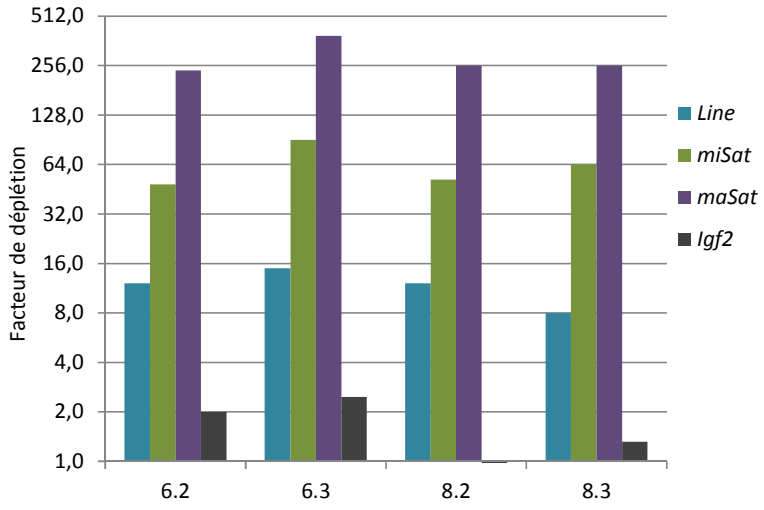


Figure 48: Facteurs de déplétion obtenus sur 4 MEFs pour le MeDIP-dep-Seq

Les différentes séquences utilisées comme contrôles ont donné des facteurs de déplétion très homogènes entre nos 4 échantillons : 11,8 en moyenne pour l'élément *Line*, 63,7 pour le satellite mineur utilisé et 284,7 pour le satellite majeur (voir Figure 48) tandis que les séquences d'intérêt représentées par *Igf2* n'ont pas été affectées (facteur de déplétion de 1,5 en moyenne).

V.2 Préparation des librairies

Les échantillons immunoprécipités, avant (MeDIP-Seq) ou après déplétion (MeDIP-dep-Seq), ont été préparés pour le séquençage : 25 µL de chacun d'entre eux ont été utilisés pour le *sizing* sur E-gel. Un échantillon de MeDIP et l'échantillon correspondant après déplétion ont été sélectionnés en parallèle sur le même E-gel. Des produits de différentes tailles ont ensuite été amplifiés. Les différents paramètres utilisés pour la préparation de chaque échantillon sont résumés dans le Tableau 28 ainsi que les quantités d'ADN obtenues à la fin du protocole. Plusieurs tailles ont été choisies pour le MEF 6.2, conduisant à deux couples MeDIP-Seq/MeDIP-dep-Seq (6.2 a et b).

	6.2 a		6.2 b		6.3		8.2		8.3	
	MeDIP	MeDIP-dep	MeDIP	MeDIP-dep	MeDIP	MeDIP-dep	MeDIP	MeDIP-dep	MeDIP	MeDIP-dep
Taille théorique (prélèvement sur E-gel)	500 pb	300 pb	400 pb	400 pb	450 pb	275 pb	500 pb	500 pb	400 pb	400 pb
Taille obtenue (mesure au Bioanalyzer)	560 pb	400 pb	460 pb	503 pb	501 pb	368 pb	577 pb	476 pb	395 pb	422 pb
Volume d'échantillon dans la PCR	20 µL	23 µL	25 µL	25 µL	25 µL	25 µL	23 µL	25 µL	23 µL	21 µL
Nombre de cycles de PCR	20	22	20	25 *	20	22	20	25 *†	20	25 *†
Quantité après PCR	2,06 pmol	3,43 pmol	2,14 pmol	0,97 pmol	0,86 pmol	0,67 pmol	1,82 pmol	1,47 pmol	1,60 pmol	2,96 pmol
Concentration du produit de PCR sur la FC	10 pM	12 pM	12 pM	12 pM	10 pM	12 pM	12 pM	12 pM	12 pM	12 pM
Longueur de lecture	101 bases		101 bases		76 bases		101 bases		101 bases	

*Utilisation de bétaïne dans la PCR
†Elongation de 45 secondes

Tableau 28: Paramètres utilisés pour la préparation du séquençage de 4 ADNs de MEFs

Pour le MeDIP-Seq, 20 cycles de PCR ont permis d'obtenir en moyenne 1,70 pmol de matériel. Pour le MeDIP-dep-Seq, les 22 cycles déterminés dans le protocole final n'ont, dans certains cas, pas permis d'obtenir une quantité suffisante d'ADN. Nous avons donc augmenté ce nombre de cycles à 25 pour obtenir, en moyenne sur tous les échantillons, 1,90 pmol d'ADN. Nous avons également noté que l'ajout de 1 mmol de bétaïne et l'allongement de l'étape d'élongation de la PCR à 45 secondes amélioraient son rendement.

V.3 Séquençage des librairies et exploitation des données

Chaque MEF IP et l'échantillon correspondant ayant subi la déplétion des éléments répétés ont été séquencés en *paired-end* sur une même *flow cell* (paired-end flow cell v4, Illumina) d'un séquenceur GAIIx d'Illumina (avec les réactifs du kit TruSeq SBS v5, GA, Illumina). 10 à 12 pM de produits de PCR

ont été utilisés et le séquençage a été réalisé avec une longueur de lecture de 76 ou 101 bases (voir Tableau 28).

V.3.1 Qualité du séquençage

Après séquençage, nous avons obtenu des rendements équivalents pour nos 2 techniques : en moyenne 3,6 Gbases après MeDIP-Seq et 3,7 Gbases après MeDIP-dep-Seq.

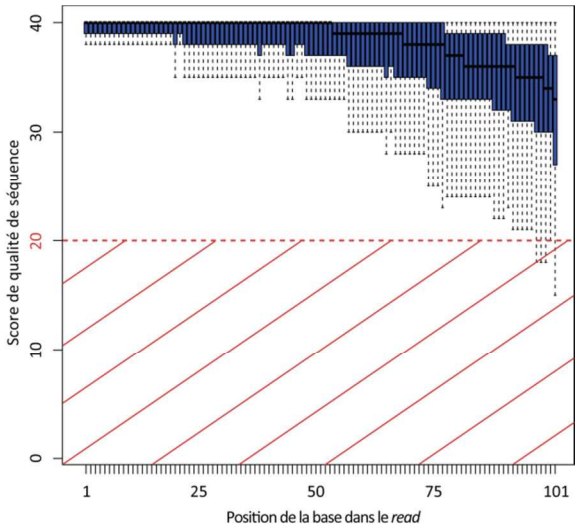


Figure 49: Scores de qualité de séquençage des bases en fonction de leur position de lecture
Scores obtenus pour les *reads* 2 après MeDIP-dep-Seq sur l'échantillon 6.2a. La ligne rouge en pointillés indique le score limite d'une base permettant de la conserver pour la suite des analyses.

Sur l'ensemble de ces bases, celles qui ont été obtenues dans les premiers cycles de séquençage sont de meilleure qualité. A partir d'une cinquantaine de cycles leur qualité commence à baisser (voir Figure 49). Afin de ne pas biaiser les analyses, nous avons effectué un rognage (*trimming*) des bases de basse qualité en fixant une valeur seuil de score de qualité de séquençage à 20. Les bases en fin de séquence dont le score sera inférieur à ce seuil ne seront pas prises en compte pour l'exploitation des données et les *reads* correspondants auront donc une taille inférieure à 101 bases (ou 76 bases pour l'échantillon 6.3).

V.3.2 Alignement

Dans un premier temps, la plateforme d'analyse d'Illumina, ELAND, nous a fourni des résultats préliminaires quant à un premier alignement sur le génome de souris de référence (mm9).

En moyenne, 26,8% des *reads* obtenus après MeDIP-Seq ont pu être alignés (voir distribution sur les échantillons dans la Figure 50) avec un taux d'erreur de 2,04 et un score d'alignement de 81,7. Près de 75% des séquences obtenues sont donc inutilisées car impossibles à aligner. En revanche, l'alignement a été considérablement amélioré par la déplétion des séquences répétées puisqu'après

MeDIP-dep-Seq 81,6% des reads ont pu être alignés avec un taux d'erreur plus faible (1,08 en moyenne) et un score d'alignement nettement plus important (324,2 en moyenne). Le MeDIP-dep-Seq permet ainsi d'approcher les statistiques obtenues en routine pour du reséquençage d'ADN génomique où le taux d'alignement moyen est de 95% (chiffre obtenu en moyenne sur 100 lignes de *flow cells* de GAIIx) avec un taux d'erreur et un score d'alignement considérés comme corrects pour des valeurs inférieures à 2,0 et supérieures à 300 respectivement.

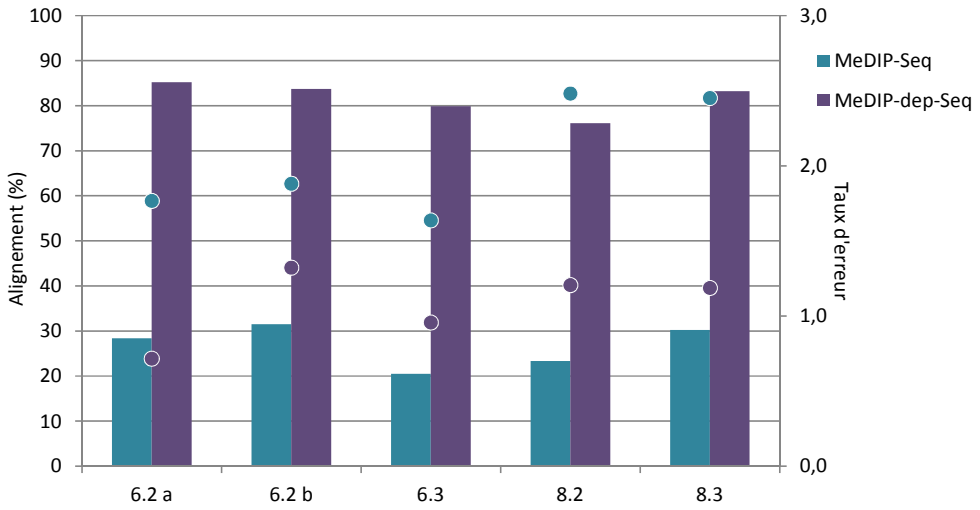


Figure 50: Pourcentages d'alignement et taux d'erreur à l'alignement obtenus avec ELAND
Le diagramme en barres représente le pourcentage d'alignement des *reads* ayant passé les filtres d'ELAND. Les points représentent les taux d'erreur à l'alignement.

Pour des analyses plus complètes, nous avons développé la plateforme MeQA qui fait appel à BWA (191,192) en tant qu'outil d'alignement. Ce dernier a par ailleurs fait preuve de meilleures performances qu'ELAND (219). Une des différences majeures entre ces deux logiciels d'alignement réside dans leur définition de l'alignement propre au séquençage en *paired-end*. Les chiffres cités précédemment et obtenus avec ELAND correspondent aux pourcentages de paires de *reads* ayant trouvé un alignement unique sur le génome de référence : un *read* 1 (R1) aligné de façon unique et dont le *read* 2 (R2) associé trouverait un alignement multiple (sur plusieurs régions du génome) ne se voit pas pris en compte (voir Figure 51A).

BWA applique un alignement moins stringent et permet de conserver des paires dont un *read* seulement trouverait un alignement unique (car ce *read* serait conservé si le séquençage était effectué en mode *single-read*), tout en prenant en compte le score d'alignement du second (Figure 51A, cas 2 à 4). Nous avons fixé une valeur seuil de score d'alignement à 26 (voir Figure 51B) : seuls les *reads* pour lesquels ce score sera supérieur à cette valeur seront pris en compte dans l'analyse. Il faut noter que la grande majorité des séquences obtenues après MeDIP-dep-Seq respectent cette

condition (valeur maximale du score d'alignement) alors que la densité de séquences résultant du MeDIP-Seq et remplissant cette condition est bien plus faible.

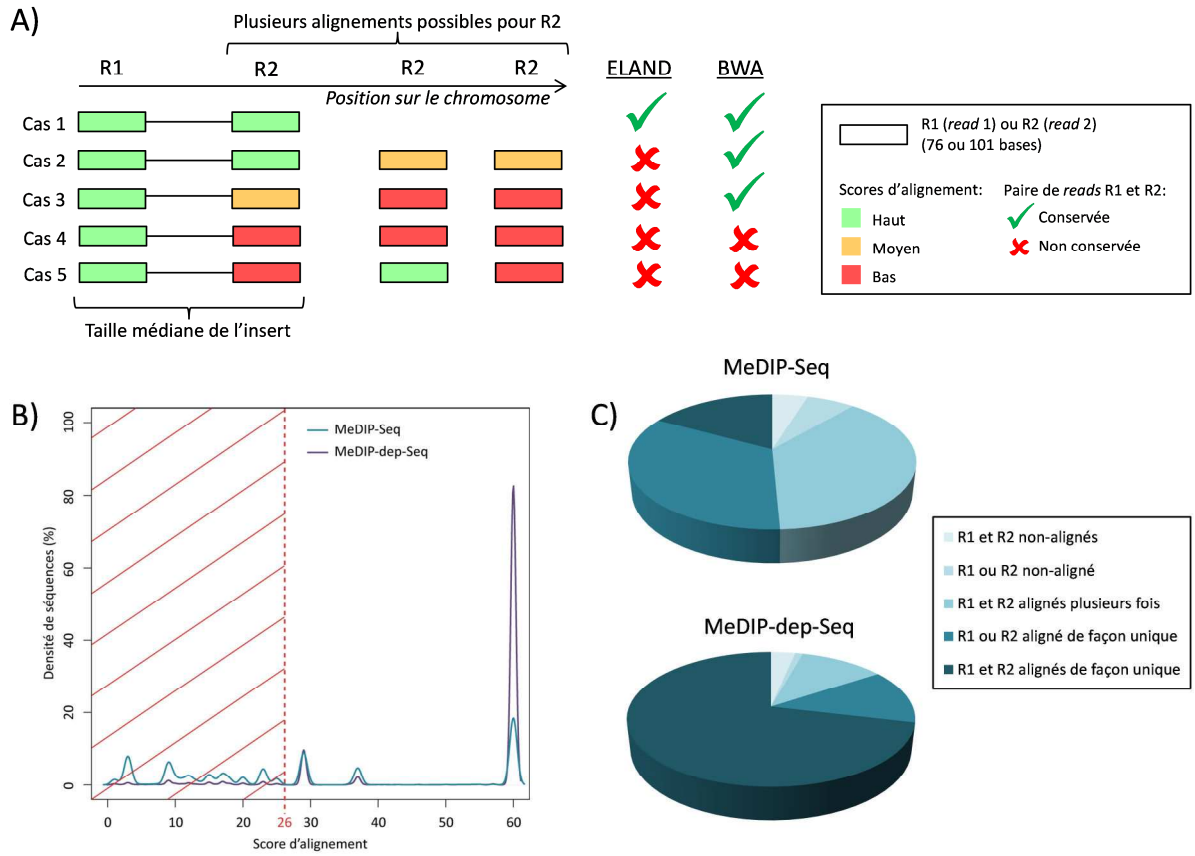


Figure 51: Alignements des paires de *reads* avec ELAND et BWA
A) Traitement des *reads* par ELAND et BWA lorsque plusieurs alignements sont possibles. B) Scores d'alignement obtenus avec BWA sur l'ensemble des reads (ici sur l'échantillon 8.2). La ligne rouge en pointillés indique le score limite d'un *read* permettant de le conserver pour la suite des analyses. C) Classement des séquences en fonction de la nature de R1 (*read 1*) et R2 (*read2*). Moyennes sur les 5 couples MeDIP-Seq/MeDIP-dep-Seq étudiés.

La distance entre R1 et R2 est également prise en compte dans l'alignement (que ce soit avec ELAND ou BWA) et doit être proche de la taille médiane de l'insert calculée sur l'ensemble des séquences, pour permettre de conserver une paire de *reads* pour la suite de l'analyse (voir Figure 51A, cas 5). Enfin, étant donné le nombre de cycles de PCR réalisés dans la préparation des échantillons et pour s'affranchir du biais qu'ils pourraient introduire, nous avons éliminé les *reads* redondants (alignés sur des coordonnées chromosomiques exactement identiques) pour n'en compter qu'un seul.

En utilisant ces paramètres et à l'aide de BWA, nous avons pu aligner sur le génome de référence les paires de *reads* obtenues après MeDIP-Seq et après MeDIP-dep-Seq et nous les avons classées en fonction de la nature de R1 et R2 (alignement et/ou pas d'alignement, unique et/ou multiple, voir Figure 51C). Il est très clair qu'après MeDIP-Seq peu de séquences trouvent un alignement satisfaisant : seules 19,1% des paires R1/R2 sont alignées sans ambigüité et ce chiffre est augmenté à

57,1% si l'on prend également en compte les paires où seul un des deux *reads* est aligné de façon unique. Une amélioration considérable est apportée par notre protocole de déplétion des séquences répétées puisque 74,4% des séquences sont alors alignées de façon unique et 88,1% si l'on considère également les paires où seul un des deux *reads* remplit cette condition.

Le faible alignement obtenu après MeDIP-Seq confirme la présence de séquences répétées en grand nombre puisque la majorité des séquences trouve un alignement multiple ou un alignement trop éloigné du second *read* correspondant, donc non pris en compte, voire ne peut être alignée. Le gain conséquent apporté par le MeDIP-dep-Seq montre ainsi l'efficacité de la déplétion.

V.3.3 Couverture

Nous avons vu que la qualité de l'alignement est nettement améliorée avec le MeDIP-dep-Seq. En revanche, il est à prévoir que l'introduction du protocole de déplétion provoque une suppression d'un grand nombre de séquences et implique donc une diminution de la couverture du génome.

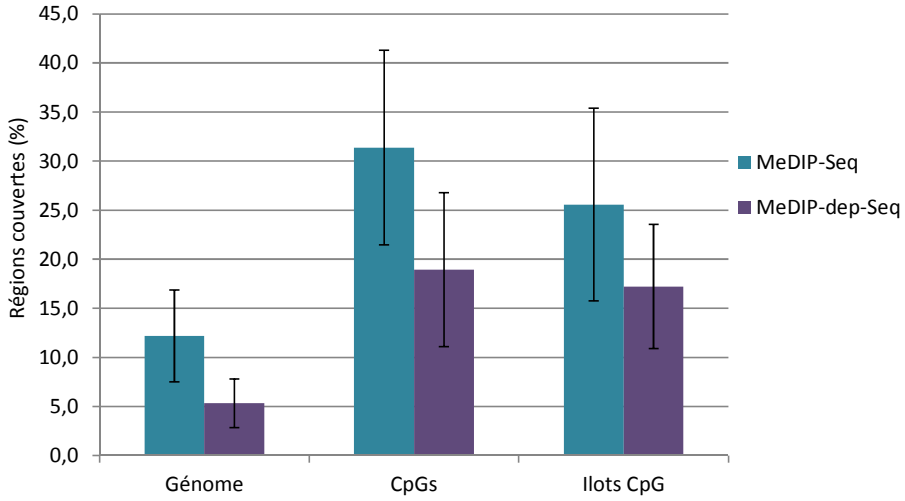


Figure 52: Couverture du génome, des CpGs et des îlots CpG, par MeDIP-Seq et MeDIP-dep-Seq
Moyennes encadrées de leurs écart-types obtenus sur l'ensemble de nos données. Les *reads* alignés de façon unique sont pris en compte. Pourcentage du génome de référence couvert avec les 2 techniques, des CpGs couverts par rapport aux 21342779 CpGs contenus dans le génome de référence et des ilôs CpGs par rapport aux 16026 contenus dans le génome de référence. Un îlot CpG est défini par un contenu en GC de plus de 50%, une longueur de plus de 200 pb et un ratio (nombre observé de CGs/nombre théorique de CGs) de plus de 0,6.

Nous avons effectivement constaté que la moitié de la couverture obtenue par le MeDIP-Seq disparaissait avec le MeDIP-dep-Seq (voir Figure 52) pour lequel 5,3% du génome peut être analysé. 31,4% des CpGs et 25,6% des îlots CpG du génome de référence sont couverts par MeDIP-Seq tandis que leurs couvertures respectives par le MeDIP-dep-Seq approche seulement les 2/3 de ces chiffres. Il est alors intéressant de préciser si la couverture que nous perdons correspond essentiellement à celle des séquences répétées.

V.3.4 Elimination des séquences répétées

Nous avons étudié de plus près l'impact de la déplétion sur les séquences répétées. Une première approche a consisté à visualiser les séquences obtenues après MeDIP-Seq et MeDIP-dep-Seq.

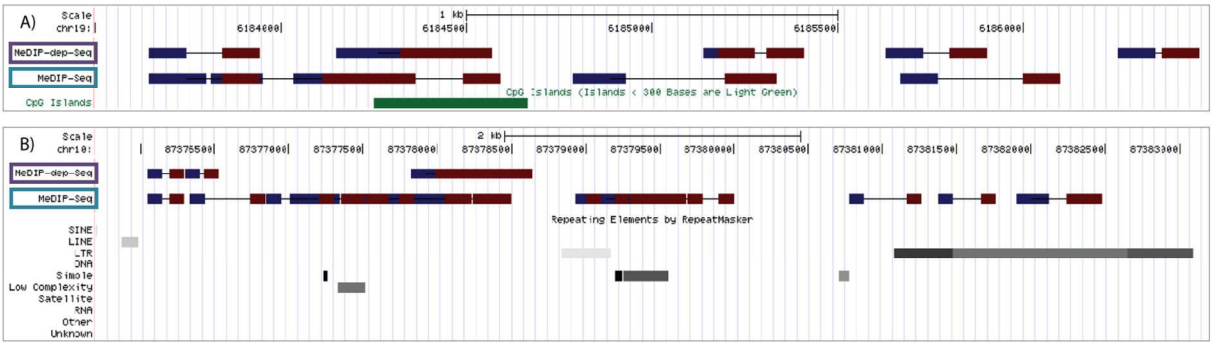


Figure 53: Visualisation des séquences obtenues après MeDIP-Seq et MeDIP-dep-Seq
Données de MeDIP-Seq et de MeDIP-dep-Seq intégrées au navigateur de l'UCSC pour visualisation. Les *reads* sont représentés par des blocs bleus (R1) et rouges (R2) et ceux d'une même paire sont reliés par une ligne noire. Tous les *reads* sont superposés pour une visualisation plus aisée. A) Fenêtre de 3000 pb incluant un îlot CpG (en vert). Nous avons compté 26 paires non-redondantes pour le MeDIP-dep-Seq et 9 pour le MeDIP-Seq. B) Fenêtre de 7500 pb incluant plusieurs familles d'éléments répétés (en gris). Nous avons compté 23 paires non-redondantes pour le MeDIP-dep-Seq et 61 pour le MeDIP-Seq.

Nous avons observé que des régions couvertes par le MeDIP-Seq le sont aussi par MeDIP-dep-Seq ; c'est par exemple le cas de CpGs localisés dans certains îlots CpG. De plus, un nombre plus important de séquences est décompté après MeDIP-dep-Seq (voir Figure 53A). En revanche, des séquences couvrant des régions répétées du génome et représentées dans les données de MeDIP-Seq sont présentes en faible quantité, voire absentes, dans les données issues de MeDIP-dep-Seq (voir Figure 53B), ce qui prouve que l'introduction du protocole de déplétion a permis la suppression de bon nombre d'entre elles dans l'échantillon immunoprécipité.

Nous avons par la suite étudié d'une façon plus large l'impact de la déplétion sur les éléments répétés du génome. Pour cela, nous avons défini de nouvelles références sur lesquelles effectuer l'alignement de nos séquences : nous avons d'une part masqué toutes les régions répétées du génome de souris de référence afin de connaitre la quantité de nos *reads* correspondant à des séquences uniques, et nous avons d'autre part masqué toutes les séquences uniques pour obtenir l'alignement des séquences répétées (voir Figure 54A). En moyenne, seules 11,9% des séquences ont pu être alignées sur des séquences uniques après MeDIP-Seq tandis que 55,5% correspondaient à des séquences répétées. Au contraire, après MeDIP-dep-Seq, 41,4% des *reads* correspondent à des séquences d'intérêt, soit 3,5 fois plus qu'après MeDIP-Seq, tandis que la quantité de séquences répétées (25,4%) est diminuée de plus d'un facteur 2.

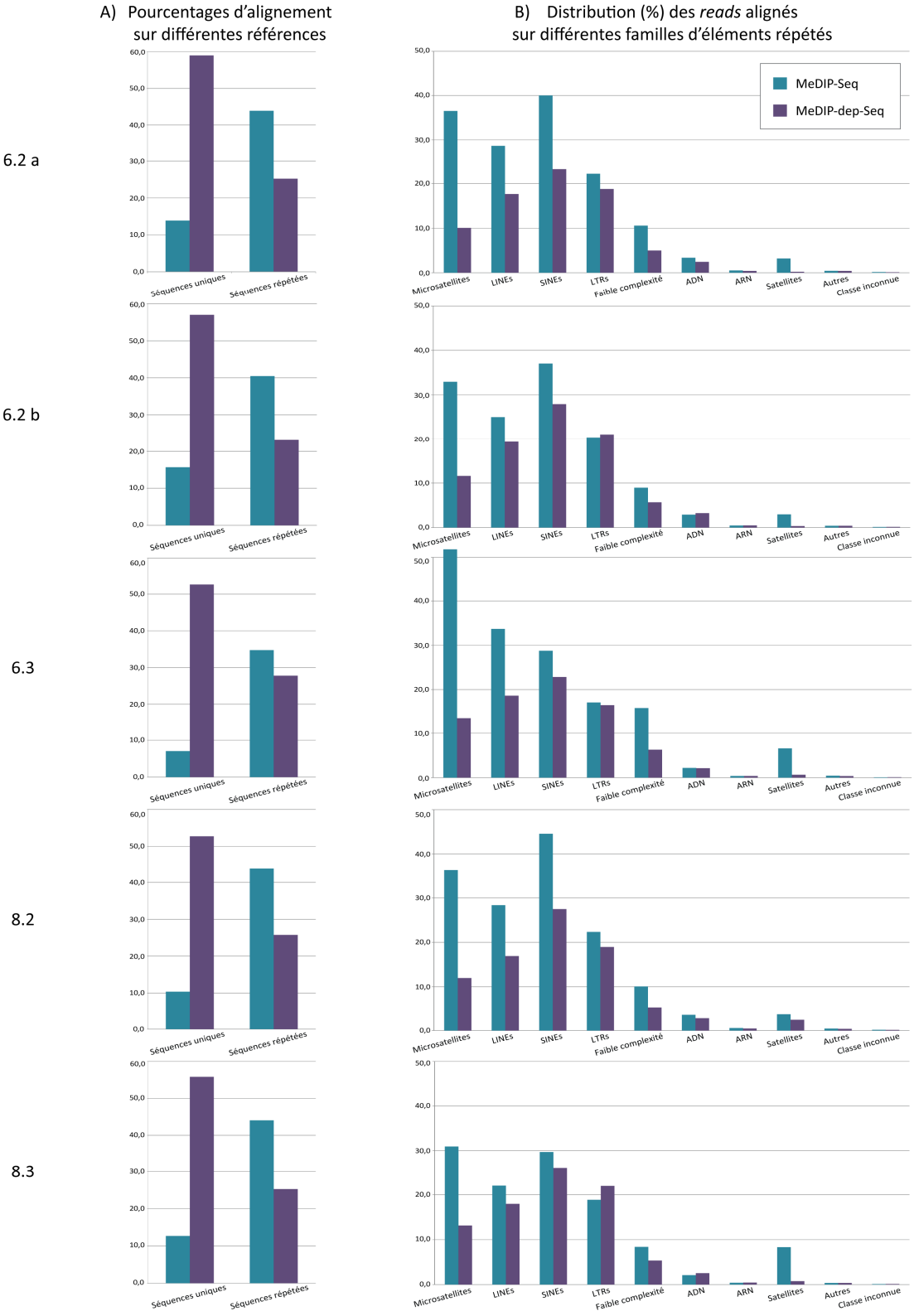


Figure 54: Alignements sur différentes références et distribution sur les éléments répétés
Seules les paires de *reads* alignées de façon unique (R1 et/ou R2) sont prises en compte ici. A) Pourcentages de paires de *reads* trouvant un alignement sur les séquences uniques du génome de souris de référence (nous avons masqué les séquences répétées pour l'alignement) et sur les séquences répétées de ce même génome (nous avons masqué les séquences uniques). B) Distribution des séquences dans différentes familles d'éléments répétés : microsatellites, éléments LINEs et SINEs, LTRs, éléments de faible complexité, éléments répétés ADN et ARN, séquences satellites et autres familles. Pourcentages par rapport au nombre total de *reads* ayant été alignés en masquant les séquences uniques.

La faible couverture du génome par MeDIP-dep-Seq observée dans la Figure 52 peut donc s'expliquer par le fait que les séquences répétées, abondantes parmi les séquences issues de MeDIP-Seq, laissent place à des séquences uniques qui occupent moins de régions dans le génome.

Nous nous sommes par la suite intéressés de plus près aux types d'éléments répétés affectés par la déplétion. Les satellites (dont les satellites mineurs et majeurs), bien que présents en faible proportion dans nos séquences, sont ceux qui subissent la plus grande diminution puisqu'en moyenne 81,7% d'entre eux n'apparaissent plus après MeDIP-dep-Seq en comparaison au MeDIP-Seq (voir leur distribution à travers les différents échantillons sur la Figure 54B). Viennent ensuite les microsatellites avec 68,0% de déplétion. Il faut noter qu'en moyenne 34,3% des éléments LINEs disparaissent avec l'introduction du protocole de déplétion. Ceci corrobore les données de contrôle-qualité de la déplétion par qPCR où les satellites mineurs et majeurs choisis présentaient un facteur de déplétion plus important que celui de l'élément *Line* étudié.

V.3.5 Quantification de la méthylation

V.3.5.1 Comparaison entre MeDIP-Seq et MeDIP-dep-Seq

La plateforme MeQA que nous avons développée permet, après préparation des données et contrôle de leur qualité, de quantifier la méthylation à travers les séquences obtenues grâce à MEDIPS (193). Comme précisé précédemment, afin de s'affranchir du biais que pourrait introduire le nombre de cycles de PCR, nous n'avons compté qu'un seul *read* en lieu et place de tous les *reads* redondants. Ceci trouve toute son importance ici puisque MEDIPS utilise le nombre de séquences présentes dans une fenêtre de taille définie par rapport au nombre total obtenu pour fournir une valeur de méthylation de cette fenêtre.

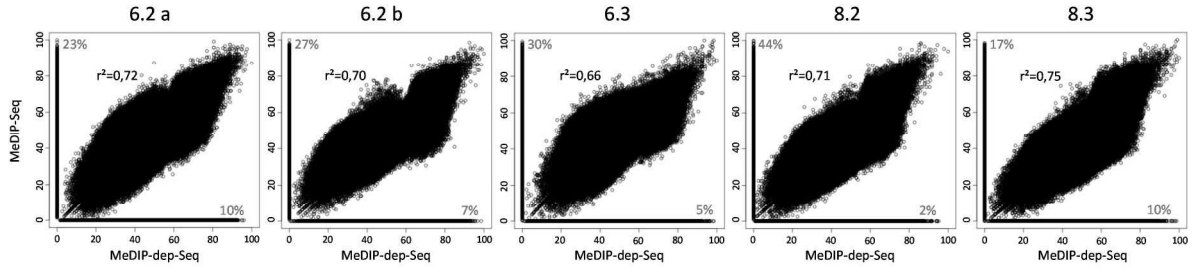


Figure 55: Diagrammes de dispersion des valeurs de méthylation obtenues en MeDIP(-dep)-Seq
Les axes représentent les pourcentages de méthylation obtenus avec MEDIPS dans la plateforme MeQA. Un point représente la valeur obtenue pour une fenêtre de taille définie. Les lignes formées par les points aux abscisses et ordonnées nulles représentent les fenêtres pour lesquelles aucune valeur de méthylation n'a pu être calculée par absence ou insuffisance de *reads*. Le pourcentage de ces fenêtres par rapport à leur totalité est indiqué en gris. Le coefficient de corrélation indiqué ici a été calculé sans prendre en compte les valeurs nulles.

Après MeDIP-dep-Seq, en moyenne 28% des fenêtres définies pour le calcul n'ont pas donné lieu à une valeur de méthylation, par absence ou insuffisance de *reads* dans cette région (voir Figure 55). Ce chiffre n'est que de 7% en ce qui concerne le MeDIP-Seq. En effet, la déplétion a provoqué la suppression de nombreuses séquences (riches en séquences répétées) entrainant donc l'impossibilité de calculer une valeur de méthylation dans ces régions. Nous avons ainsi étudié les pourcentages de méthylation obtenus par MeDIP-Seq et MeDIP-dep-Seq pour chacun de nos échantillons en nous affranchissant de ces fenêtres. Sur l'ensemble du génome, nous avons observé une bonne corrélation entre les deux techniques (coefficient de corrélation de 0,71 en moyenne, voir Figure 55). Ceci montre que l'introduction du protocole de déplétion des séquences répétées après MeDIP permet de conserver l'information quantitative sur la méthylation.

V.3.5.2 Comparaison avec des données de puces à ADN et de pyroséquençage

Nous disposons dans le laboratoire de plusieurs types de données sur les mêmes échantillons de MEFs. Ceux-ci ont été immunoprécipités puis hybridés sur une puce de promoteurs de la plateforme Roche/NimbleGen (Mouse DNA Methylation 385K Promoter Plus CpG Island Arrays) pour fournir les données de MeDIP-chip. Celles-ci ont fait l'objet d'une normalisation quantile et d'une analyse par l'algorithme Batman afin d'en déduire des niveaux de méthylation. Certaines des régions ciblées par cette puce ont également été étudiées par pyroséquençage après MeDIP (voir liste en Annexe 3). Afin de valider notre approche, nous avons comparé ces données aux valeurs de méthylation obtenues par MeDIP-Seq et MeDIP-dep-Seq en utilisant MeQA et en ciblant les mêmes régions. Nous nous sommes intéressés aux données issues du MEF 6.2a.

La comparaison avec une douzaine de valeurs de méthylation provenant du pyroséquençage après MeDIP fournit une corrélation que l'on peut considérer comme satisfaisante, que ce soit avec les valeurs issues du MeDIP-Seq comme avec celles du MeDIP-dep-Seq (voir Figure 56A). En effet, dans les deux cas, un coefficient de corrélation de plus de 0,65 est obtenu, dépassant ainsi celui qui peut être fourni par une comparaison entre pyroséquençage et MeDIP-chip (voir Figure 56C), techniques jusqu'alors les plus usitées. La comparaison avec les données de MeDIP-chip aboutit à une corrélation certes correcte mais moins convaincante (voir Figure 56B). Cependant, les valeurs de méthylation provenant des puces ont été obtenues avec l'algorithme Batman tandis que celles provenant du MeDIP(-dep)-Seq ont été calculées à l'aide de MEDIPS.

Sur les 12 régions étudiées et présentées ici, on observe que le MeDIP-dep-Seq fournit des valeurs de méthylation légèrement supérieures aux valeurs auxquelles le MeDIP-Seq peut conduire (15% en moyenne), en conservant toutefois le même profil global de distribution. Ceci corrobore d'ailleurs les corrélations présentées dans le paragraphe précédent de façon plus générale sur le génome entier.

Une explication résiderait dans le fait que, dans les régions où MeQA parvient à calculer une valeur de méthylation lorsqu'un nombre suffisant de *reads* est disponible, ce nombre est supérieur après MeDIP-dep-Seq puisque la déplétion des séquences répétées a permis l'accès à d'autres séquences. Le nombre de *reads* étant pris en compte pour calculer une valeur de méthylation, il en résulte une valeur légèrement plus forte mais certainement plus proche de la réalité biologique.

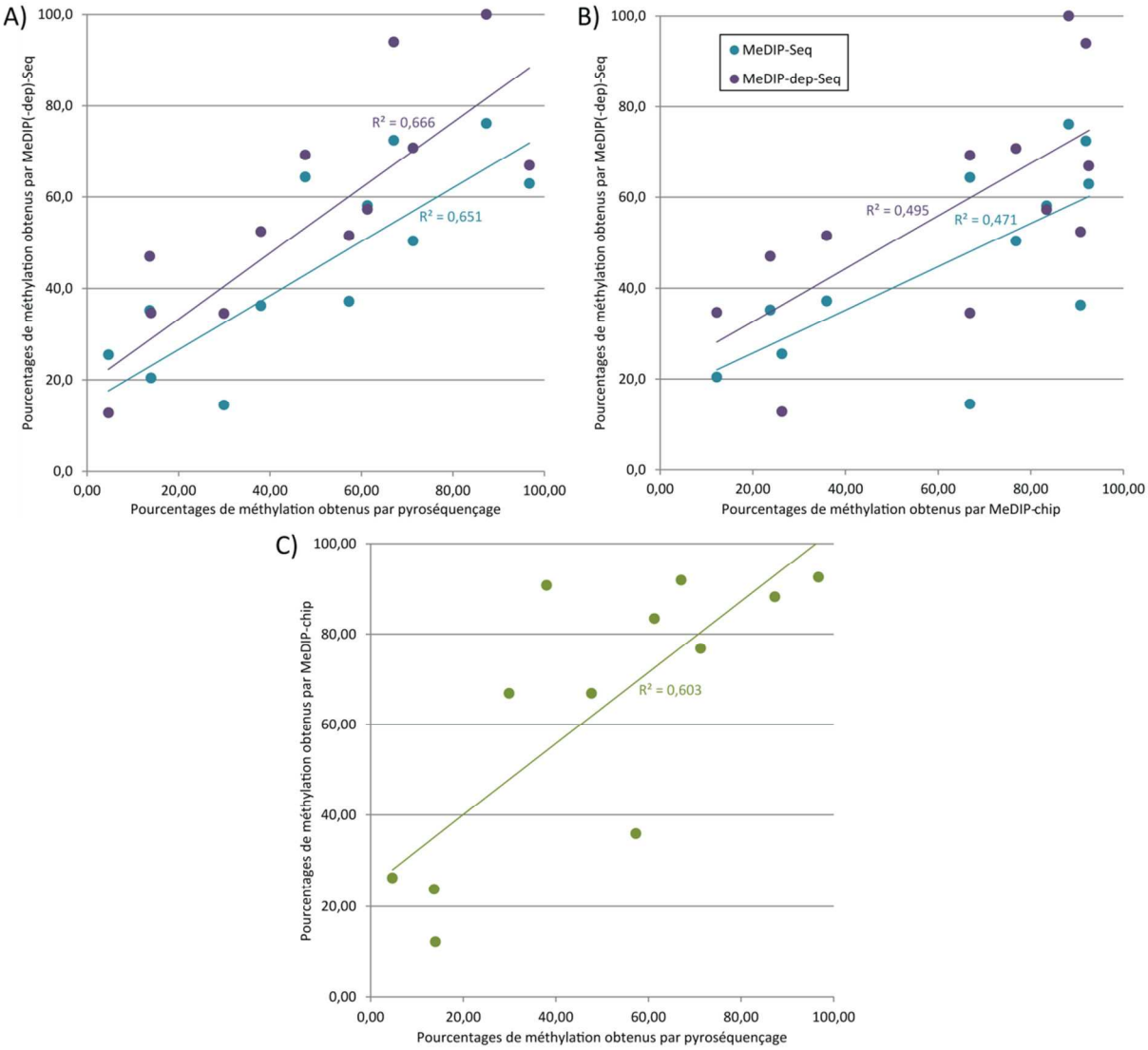


Figure 56: Corrélations entre MeDIP(-dep)-Seq, MeDIP-chip et pyroséquençage
Données du MEF 6.2a. A) Corrélation entre pyroséquençage après MeDIP et MeDIP(-dep)-Seq d'après des valeurs de méthylation obtenues sur 12 régions promotrices. Droites de tendance obtenues sur ces points avec leur coefficient de corrélation associé. B) Corrélation entre MeDIP-chip et MeDIP(-dep)-Seq d'après des valeurs de méthylation obtenues sur les mêmes régions promotrices figurant sur la puce utilisée. C) Corrélation entre pyroséquençage après MeDIP et MeDIP-chip d'après les valeurs citées ci-avant.

V.3.6 Identification de régions différentiellement méthylées (DMRs)

Parmi les échantillons de MEFs étudiés, deux proviennent de lignées sauvages (6.3 et 8.2) et deux de lignées mutées pour le gène *Men1* (6.2 et 8.3). Nous avons cherché à comparer ces individus deux à deux afin d'identifier des régions différentiellement méthylées à travers leur génome, après MeDIP-Seq et après MeDIP-dep-Seq. Pour cela, nous avons utilisé la fonction prévue à cet effet dans le logiciel MEDIPS intégré à notre plateforme MeQA.

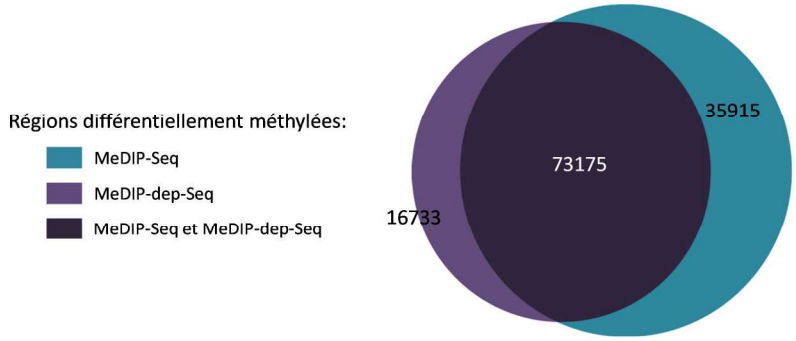


Figure 57: Diagramme de Venn représentant les DMRs
Les individus MEFs provenant des lignées sauvages ont été comparés à ceux provenant des lignées mutées pour identifier les DMRs.

Nous avons identifié plus de 73000 DMRs communes à nos deux protocoles (voir Figure 57). Le MeDIP-dep-Seq permet d'en déterminer quasiment 17000 supplémentaires tandis que près de 36000 sont propres aux données issues de MeDIP-Seq. Ceci appuie les observations que nous avons pu faire jusqu'ici : davantage de séquences sont présentes après MeDIP-Seq, elles permettent de couvrir des régions plus larges, et donc d'accéder à plus de DMRs sur nos échantillons. Néanmoins, 67% des régions identifiées par MeDIP-Seq le sont aussi après introduction du protocole de déplétion. Une grande partie de l'information biologique est donc maintenue après MeDIP-dep-Seq et il serait désormais intéressant de préciser quels types de régions se retrouvent différentiellement méthylés de façon commune au MeDIP-Seq et au MeDIP-dep-Seq.

V.3.7 Ouverture à d'autres espèces

La déplétion ainsi mise au point a fait preuve de son efficacité dans l'élimination de séquences répétées et le gain d'information qu'elle fournit au MeDIP-Seq utilisé de façon standard. Elle peut désormais être adaptée à toute espèce pour laquelle un ADN enrichi en séquences répétées est disponible. Nous souhaitons la mettre en place sur des échantillons humains. Une recherche bibliographique nous a d'abord permis d'obtenir des amorces de qPCR permettant d'amplifier des éléments répétés pour le contrôle de la déplétion. Il s'agit de séquences ciblant des éléments microsatellites (220,221), des éléments Alu (222) ou α-satellites (223) dont les coordonnées figurent

en Annexe 2. Des qPCRs ont confirmé la présence de certaines de ces régions dans l'ADN Cot-1 humain utilisé (Human Cot-1 DNA, Life Technologies), dont les microsatellites présents dans les gènes *CD4*, *PLA2A1* et *CYAR04* ainsi que les trois séquences α-satellites identifiées. Des premiers tests de biotinylation de ce Cot-1 avec le protocole utilisé pour la biotinylation du Cot-1 de souris ont mené à des facteurs d'amplification, et donc des incorporations de biotine, moins importants (4,1 en moyenne en comparaison à 11,3 sur la souris). Une augmentation de la quantité de l'enzyme impliquée dans la biotinylation ou de la durée de l'incubation n'a pas amélioré le rendement. En revanche, la biotinylation d'un ADN enrichi en séquences répétées provenant d'un autre fournisseur (Human Hybloc DNA, Applied Genetics Laboratories) a permis d'augmenter ce facteur à 9,9 en moyenne. Nous pouvons donc désormais débuter les tests de déplétion des séquences répétées sur des échantillons humains.

V.4 Discussion

Le protocole de déplétion des séquences répétées que nous avons établi a été introduit dans le processus du MeDIP-Seq pour mener au MeDIP-dep-Seq. Nous avons cherché à confirmer son efficacité par séquençage en l'appliquant à des échantillons de MEFs. La qPCR de vérification de l'étape de déplétion a fourni des résultats similaires à ceux que nous avons pu obtenir sur de l'ADN commercial durant la mise au point du protocole : en moyenne, l'élément *Line* étudié a vu sa quantité diminuer de plus d'un facteur 10 quand les satellites mineurs ont été supprimés des échantillons avec un facteur proche de 65, facteur 4 fois plus grand en ce qui concerne les satellites majeurs. Le séquençage confirmera que les séquences satellites (mineurs et majeurs) sont les plus affectées par le protocole de déplétion puisque près de 80% d'entre elles disparaissent, tandis que 34% des séquences de la famille *Line* se verront éliminées. Nous avons également pu nous rendre compte de la diminution des éléments répétés en parcourant avec l'outil de visualisation de l'UCSC les séquences obtenues : dans les familles d'éléments répétés, davantage de séquences (non-redondantes) ont pu être comptabilisées après MeDIP-Seq en comparaison au MeDIP-dep-Seq. Ce dernier a, au contraire, permis d'augmenter le nombre de séquences d'intérêt, notamment au niveau des îlots CpG lorsque ceux-ci étaient déjà couverts par le MeDIP-Seq.

Le séquençage a été effectué sur un Genome Analyzer IIx (Illumina) sur 76 ou 101 bases. De telles longueurs de lecture n'étaient pas indispensables dans notre cas, d'ailleurs la grande majorité des études de MeDIP-Seq qui ont pu être décrites dans la littérature utilisent des longueurs de 50 bases au maximum (voir Tableau 29). Cependant, nos échantillons ont été séquencés au fur et à mesure de leurs préparations respectives et ont été ponctuellement insérés dans le flux d'échantillons pris en charge sur la plateforme de séquençage haut-débit du centre. Nous avons donc utilisé les paramètres

imposés par les échantillons voisins des nôtres sur les *flow cells*. Diminuer cette longueur de lecture permettra de réduire les coûts engendrés par les réactifs ainsi que le temps du séquençage (il faut compter une journée pour séquencer 20 bases).

Par ailleurs, le séquençage a été réalisé en *paired-end* quand certaines études se contentent du *single-read*. Ceci a pour inconvénient de multiplier les étapes du travail d'analyse des séquences car les *reads* doivent dans un premier temps être associés par paires. Néanmoins, le *paired-end* permet de conserver davantage de séquences, surtout dans le cadre du MeDIP-Seq où de nombreux éléments répétés immunoprécipités sont séquencés et difficilement alignables. En effet, si un fragment possède sa première extrémité séquencée dans une de ces séquences répétées, il peut être sauvegardé grâce à sa deuxième extrémité qui sera alignée avec un meilleur score. Ceci demeure sous condition d'utiliser un logiciel d'alignement adapté, et nous avons pour cela opté pour BWA en comparaison à ELAND dont le mode de fonctionnement trop stringent élimine un grand nombre de séquences pourtant exploitables. De plus, l'alignement a été réalisé sur le génome de référence de la souris de laboratoire *Mus musculus* (mm9) de souche Black 6 (C57BL/6). Hors, les échantillons d'ADNs que nous avons utilisés proviennent de souris ayant un antécédent génétique différent (souris 129/Sv) (224). Afin de préciser au maximum l'alignement, nous séquençons actuellement l'un de nos échantillons d'ADN génomique dans le but d'en effectuer un assemblage *de novo* local qui servira de nouvelle référence à nos analyses.

La probabilité d'obtenir des *reads* couvrant certaines longues séquences répétées, comme les éléments *Line* ou les LTRs, est forte. En revanche, il est possible d'éviter au maximum les éléments répétés de faible taille (400 pb au maximum) tels que les satellites ou les éléments *Sine* et notamment les séquences Alu, avec une taille d'insert adaptée. Nous l'avons choisie entre 400 et 500 pb de sorte que les deux *reads* à chaque extrémité puissent encadrer ces petites séquences ou, au pire des cas, que seule l'une des deux extrémités séquencées les couvre. Dans le cadre du MeDIP-dep-Seq où un grand nombre d'éléments répétés disparait, nous nous sommes autorisés à utiliser des fragments plus courts.

La mise en place de notre propre plateforme d'analyse de données, MeQA, nous a permis dans un premier temps d'effectuer l'alignement des *reads* obtenus après MeDIP-Seq et MeDIP-dep-Seq. Pour des raisons exposées ci-avant, nous avons par la suite, d'une part pris en compte uniquement les *reads* qui trouvaient un alignement unique sur le génome de référence, et d'autre part supprimé ceux qui étaient redondants. Le taux d'alignement auquel nous prétendions au début de ces travaux était de 40% en ce qui concerne le MeDIP-Seq (Stephan Beck, communication personnelle, chiffre obtenu par séquençage en *single-read* sur 36 bases). Nous avons finalement atteint les 57%, certainement parce que la longueur de nos *reads* permet de conserver davantage de séquences en

les alignant avec un meilleur score. Ce chiffre a pu être augmenté à 88% en introduisant le protocole de déplétion des séquences répétées, ce qui prouve une fois de plus son efficacité. En revanche, la suppression de séquences due à ce protocole a provoqué la diminution de la couverture en CpGs de 31 à 19%. Nous avons alors cherché à comparer ces chiffres avec les données de MeDIP-Seq de la littérature (voir Tableau 29).

	Mode du séquençage	Rendement	Alignement unique (% de *reads*)	*Reads* redondants supprimés	Logiciel d'alignement	Couverture des CpGs	Commentaires
Down, T.A. *et al.* (2008) (121)	36 bases *Paired-end*, GA	0,86 Gb*	nd	nd	MAQ	60%	Mode de calcul de la couverture non précisé
Pomraning, K.R. *et al.* (2009) (138)	36 bases *Single-read*, GA	nd	nd	nd	ELAND	nd	
Ruike, Y. *et al.* (2010) (122)	36 bases *Paired-end*, GAII	nd	nd	nd	MAQ	87%	Mode de calcul de la couverture non précisé
Harris, R.A. *et al.* (2010) (147)	nd *Single-read*, GAII	3,02 et 3,42 Gb	60,9 et 60,5%	nd	MAQ	nd	Nombre de lignes par échantillon non précisé, mode de calcul de l'alignement non plus
Li, N. *et al.* (2010) (159)	45 bases *Paired-end*, GAII	7,9 Gb	62,4%	nd	nd	nd	Nombre de lignes par échantillon non précisé, mode de calcul de l'alignement non plus
Bock, C. *et al.* (2010) (160)	36 bases *Single-read*, GAII	1,08 à 2,16 Gb*	37,7 à 56,4%	Oui	MAQ	nd	2 lignes par échantillon
Chavez, L. *et al.* (2010) (193)	36 bases *Single-read*, GAIIx	nd	nd	nd	MAQ	82 à 90%	Mode de calcul de la couverture non précisé
Butcher, L.M. *et al.* (2010) (207)	nd *Paired-end*, GAII	nd	nd	nd	nd	nd	
Feber, A. *et al.* (2011) (139)	50 bases *Paired-end*, GAIIx	6,5 Gb*	72%	nd	MAQ	nd	Nombre de lignes par échantillon non précisé, mode de calcul de l'alignement non plus
Vining, K.J. *et al.* (2012) (205)	36 bases nd, GAIIx	0,95 à 3,51 Gb*	9,1 à 59,3%	Oui	ELAND/ HashMatch	nd	3 à 5 lignes par échantillon
Sengenès, J. et al. (non publié) MeDIP-Seq	76 ou 101 bases *Paired-end*, GAIIx	3,6 Gb	57,1%	Oui	BWA	31,4%	Chiffres moyennés sur 5 échantillons. 1 échantillon par ligne. Eléménts répétés non-masqués pour le calcul de la couverture
Sengenès, J. *et al.* (non publié) MeDIP-dep-Seq		3,7 Gb	88,1%			18,9%	

nd: information non disponible

*: calculé en multipliant le nombre de (paires) de *reads* figurant dans la publication par la longueur de lecture

Tableau 29: Comparaison avec des données de MeDIP-Seq de la littérature

La comparaison concernant la couverture des CpGs s'est avérée difficile car la plupart des études ne mentionnent pas si les éléments répétés du génome sont masqués ou non pour calculer le pourcentage fourni. Ceci semble pourtant fort probable dans les cas où des chiffres de 60 à 90% sont donnés. De plus, le nombre de lignes requises sur la *flow cell* est rarement précisé et il n'est pas concevable qu'une seule d'entre elles ait été utilisée dans les cas où 3 à 8 Gb de données ont été produites avec des longueurs de lecture si courtes. De tels rendements peuvent d'ailleurs expliquer une plus grande couverture et un alignement plus important. Concernant ce dernier, rares sont les études qui précisent son mode de calcul. Dans les cas où plus de 60% de *reads* peuvent être alignés, il serait pourtant intéressant de savoir si les *reads* redondants ont été comptabilisés, d'autant plus qu'un grand nombre de cycles de PCR a été utilisé dans les protocoles correspondants (15 au minimum, voir Tableau 23), ce qui augmente naturellement la redondance. En revanche, lorsqu'il est mentionné que les *reads* redondants sont supprimés, les pourcentages d'alignement unique sont

certes plus faibles car les longueurs de lecture sont plus courtes, mais aussi beaucoup plus proches de ceux que nous avons obtenus (jusqu'à 56,4 et 59,3% en comparaison à 57,1%).

Par l'intégration du logiciel MEDIPS à notre plateforme d'analyse, nous avons rendu accessible la quantification de la méthylation à travers nos séquences. MEDIPS a été mis au point pour l'analyse de séquences issues d'un séquençage en *single-read* mais a été adapté pour analyser nos séquences en *paired-end* : après l'alignement, les séquences comprises entre les deux *reads* R1 et R2 ont été utilisées pour le calcul du pourcentage de méthylation, en plus des séquences de ces *reads*, à condition que ceux-ci aient été alignés correctement. Ainsi, l'insert dans son intégralité peut être pris en compte, ceci étant un autre avantage du séquençage en *paired-end*. Nous avons pu montrer que l'introduction du protocole de déplétion permet de conserver l'information quantitative sur les régions où un nombre suffisant de *reads* est disponible. Nous avons alors confronté les pourcentages de méthylation obtenus sur une douzaine de régions par nos deux techniques, MeDIP-Seq et MeDIP-dep-Seq, avec ceux fournis par pyroséquençage. Les comparaisons ont mené à des coefficients de corrélation de plus de 0,65. Ce chiffre, s'il n'est pas parfait, reste parmi les meilleurs obtenus dans ce type d'études comparatives. Une autre technique d'analyse aussi précise que le pyroséquençage, la technologie Infinium, a été récemment utilisée : la puce InfiniumMethylation27, qui est une technologie quantitative d'analyse de la méthylation fournie par Illumina, permettant d'interroger 27578 CpGs répartis sur plus de 14000 gènes. L'échantillon est traité par le bisulphite puis hybridé sur des billes portant deux types de sondes, complémentaires au locus méthylé et au locus non-méthylé (225). Dans une étude publiée en 2010, sa comparaison avec le MeDIP-Seq a pu fournir un coefficient de corrélation de 0,56, soit du même ordre de grandeur que celui que nous avons obtenu (160). Nos deux techniques permettent donc de quantifier la méthylation de façon satisfaisante.

Chapitre VI: Analyse multiplexée de loci par la technologie des sélectors

Dans les chapitres précédents, nous avons établi un protocole d'analyse de la méthylation sur le génome entier. Des régions d'intérêt ainsi identifiées peuvent désormais faire l'objet d'une analyse plus ciblée. Dans ce but, nous avons mis en place une collaboration avec le laboratoire suédois dirigé par Mats Nilsson afin d'élaborer une nouvelle technique basée sur l'utilisation de sondes appelées sélectors. Cet outil de capture de séquences génomiques est maîtrisé par nos collaborateurs. En revanche, son utilisation pour l'analyse de la méthylation n'a encore fait l'objet d'aucune étude. Nous cherchons donc à le combiner au traitement par le bisulphite et au séquençage haut-débit avec la technologie 454 sur le séquenceur de paillasse GS Junior de Roche, afin d'étudier la méthylation sur un grand nombre de loci, de façon multiplexée, selon une procédure présentée dans la Figure 58.

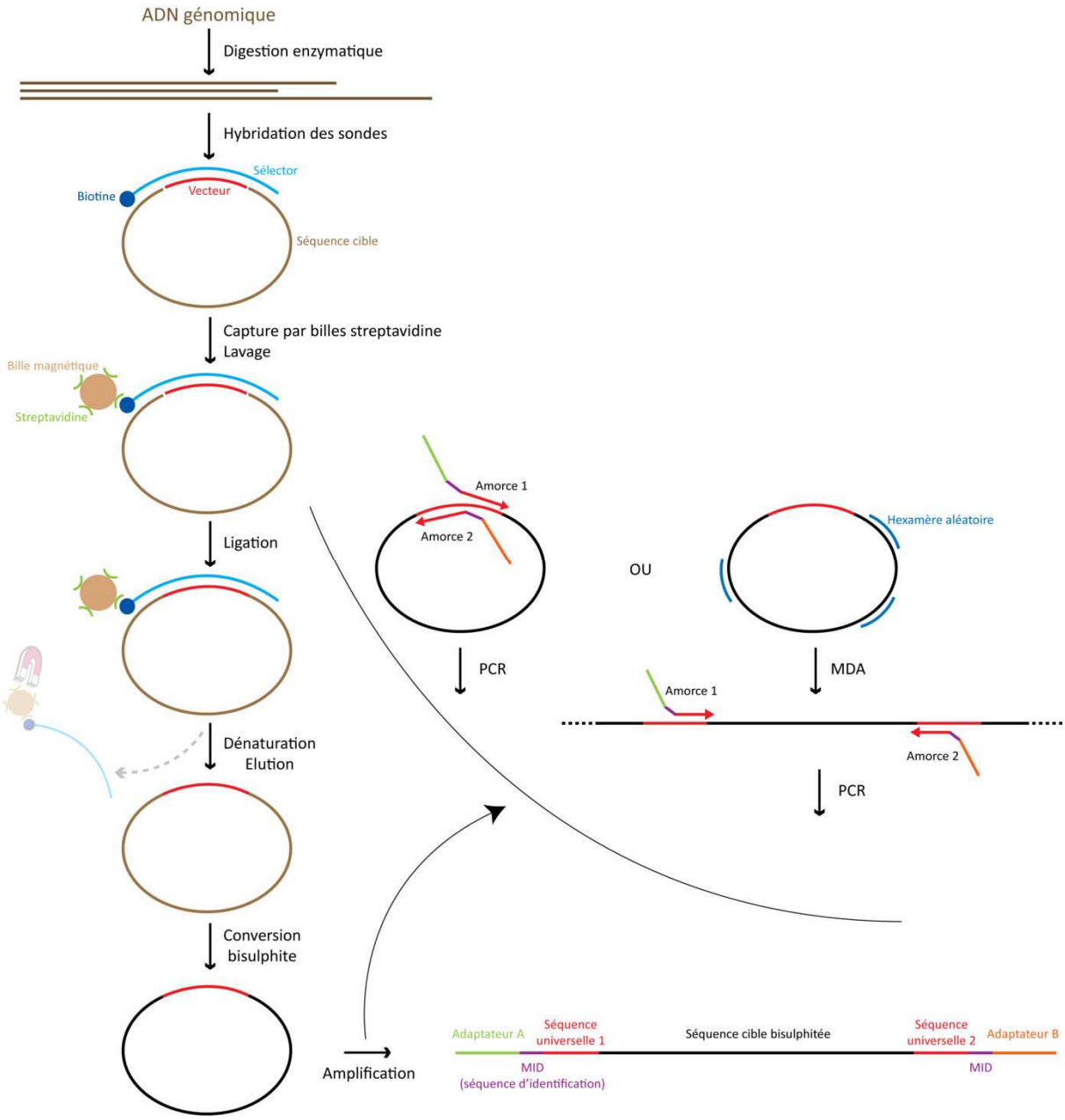


Figure 58: Procédure pour le séquençage bisulphite après sélection par les sélectors

L'échantillon est fragmenté par digestion enzymatique puis les fragments d'intérêt sont sélectionnés par les sélectors. Une séquence appelée vecteur, complémentaire et commune à tous les sélectors, s'intercale entre les extrémités de la séquence ciblée et est circularisée avec la cible par ligation. Des billes magnétiques de streptavidine se lient aux sélectors, biotinylés au préalable, et les cibles sont isolées après dénaturation. Elles sont ensuite traitées par le bisulphite de sodium puis amplifiées par MDA et/ou PCR grâce à des amorces universelles complémentaires au vecteur et portant les séquences d'adaptateurs nécessaires au séquençage. La mise au point du protocole sera effectuée sur de l'ADN Promega humain mâle.

VI.1 Construction des outils utiles à l'étude

Nous cherchons, pour la preuve de principe de ce protocole, à utiliser une centaine de sélectors qui permettront de cibler une trentaine de régions d'intérêt pour le cancer du rein, identifiées dans une étude menée dans notre laboratoire.

La conception des sélectors a été réalisée par nos collaborateurs. Pour cela, le logiciel utilisé développé par leurs soins prend en compte la taille des cibles que nous désirons étudier, soit 200 à 500 pb, et propose l'utilisation de certaines enzymes de restriction, non sensibles à la méthylation, qui permettront, après digestion, d'obtenir ces fragments. 20 à 25 bases sont alors sélectionnées sur les deux extrémités des cibles et les sélectors correspondants sont construits en introduisant la séquence complémentaire au vecteur entre les deux oligonucléotides complémentaires à cette vingtaine de bases (voir Figure 59A). Le vecteur utilisé ne contient que des cytosines méthylées afin que sa séquence ne se trouve pas modifiée par le traitement au bisulphite. Nous avons ainsi pu concevoir 98 sélectors, permettant de cibler des régions d'intérêt de taille moyenne égale à 332 bases (minimum : 204 bases, maximum : 477 bases) et dont les séquences figurent en Annexe 4.

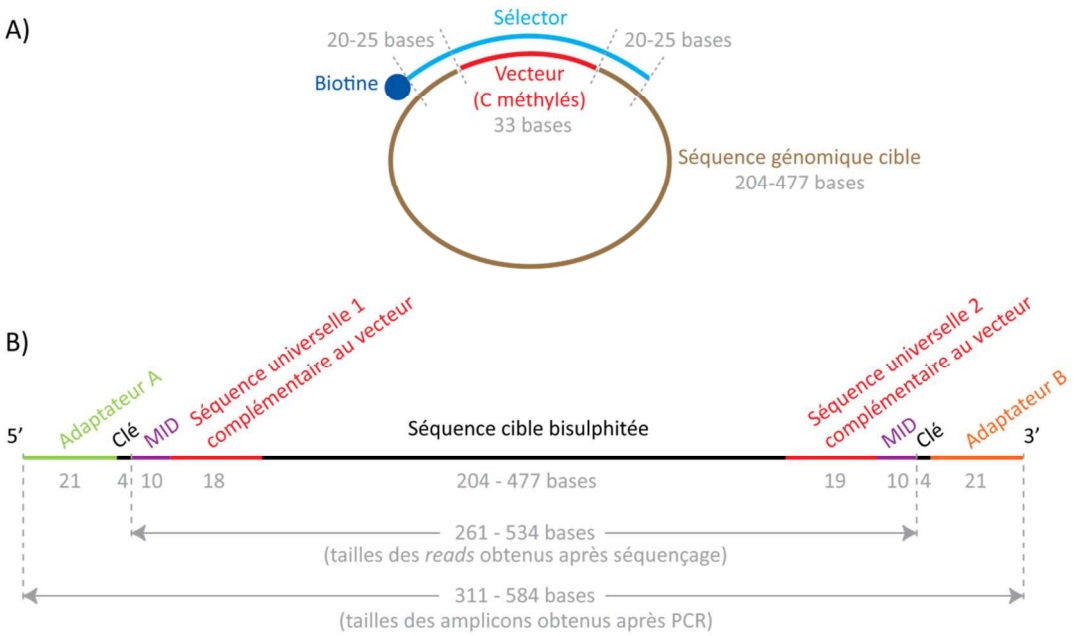


Figure 59: Construction des sélectors et de leurs cibles
En gris, taille des fragments correspondants. A) Construction des sélectors. B) Présentation des cibles après traitement au bisulphite et amplification. La somme des bases des séquences universelles complémentaires au vecteur (19 + 18 bases) est supérieure de 4 bases à la taille du vecteur (33 bases) car les amorces se chevauchent de 4 bases lors de l'amplification. MID : Multiplex IDentifier. La clé intervient dans l'analyse des séquences.

Après la sélection, les cibles sont traitées par le bisulphite et amplifiées grâce à des amorces de PCR conçues par nos soins (voir séquences en Annexe 5), menant à une séquence présentée dans la Figure 59B. Ces amorces possèdent des séquences complémentaires au vecteur qui permettent

d'amplifier toutes les cibles dans le même tube pour créer les librairies. Elles incluent également les séquences des adaptateurs (A et B) qui interviendront dans la préparation de l'échantillon au séquençage, notamment pour la PCR en émulsion et l'hybridation de l'amorce de séquençage. Elles incluent également les séquences identifiantes (MIDs) permettant le multiplexage puisque tous nos échantillons seront séquencés en parallèle sur un même support. Chaque échantillon sera donc amplifié avec une paire d'amorces qui lui est propre. Enfin, la courte séquence de 4 bases accolée aux adaptateurs appelée clé jouera un rôle dans l'analyse des séquences.

VI.2 Protocole de sélection standard

Dans un premier temps, nous avons travaillé sur le protocole de sélection standard qui n'inclut pas de traitement par le bisulphite. Ceci permet de vérifier l'efficacité de nos nouveaux sélectors dans un protocole connu et robuste.

VI.2.1 Test préliminaire de contrôle par qPCR

Nous avons conçu des couples d'amorces de PCR (appelés Gen00 à Gen26) permettant d'amplifier certaines de nos cibles. Nous avons cherché à sélectionner parmi eux par qPCR (avec la Taq Platinum et le SYBR Green) une douzaine de couples qui correspondent à des régions présentes en quantités équivalentes dans l'ADN de départ (ici, de l'ADN Promega mâle) et menant donc à des Cts proches, afin, par la suite, de mieux apprécier le succès de la sélection pour chacune de ces régions.

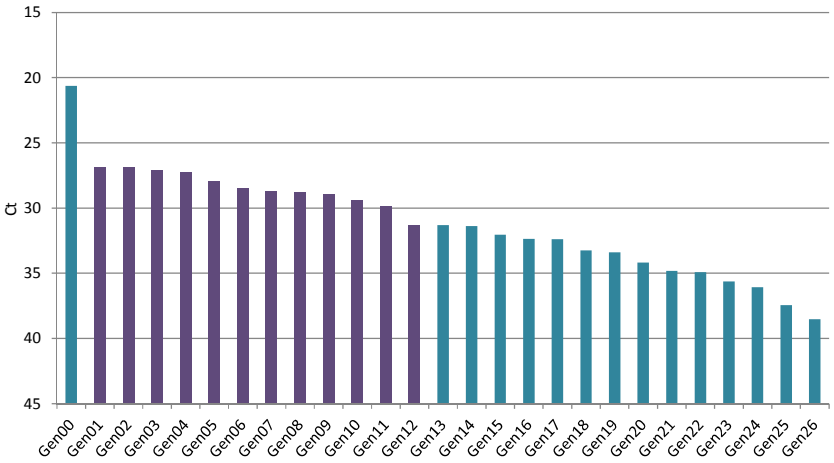


Figure 60: Cycles de seuil obtenus sur de l'ADN génomique avec nos couples d'amorces tests
Ct obtenus par qPCR sur de l'ADN Promega mâle. Gen00 et Gen13 à 26 seront écartés tandis que Gen01 à 12 seront conservés pour le contrôle de la sélection.

Nous avons sélectionné 12 couples d'amorces (Gen01 à Gen12, voir Figure 60 et séquences en Annexe 6) correspondant à des Ct proches mais surtout peu tardifs, que nous conserverons pour le contrôle de la sélection.

VI.2.2 Biotinylation des sélectors

La première étape du protocole consiste à ajouter une molécule de biotine aux extrémités 3' des sélectors (synthétisés par IDT, Integrated DNA Technologies) qui ont été réunis en quantités équimolaires (100 µM au total). Ceci est réalisé sur 1,25 µL du mix de sélectors, par incubation à 37°C avec une terminal transférase (NEB) en présence d'un nucléotide biotinylé (biotin-16-dUTP, Roche) puis désactivation de l'enzyme. Trois purifications sur colonne G-50 suivent la biotinylation puis le mix est dilué à une concentration de 1 nM.

VI.2.3 Digestion enzymatique de l'échantillon

750 ng d'échantillon, ici de l'ADN Promega, sont fragmentés par digestion enzymatique. Ils sont divisés en trois quantités égales de 250 ng afin d'être digérés en parallèle par trois mix contenant chacun deux enzymes de restriction. Les enzymes qui ont été déterminées pour permettre l'obtention de nos fragments cibles figurent dans le Tableau 30.

	Enzyme 1		Enzyme 2	
Mix d'enzymes 1	*Mse*I	5'… T▼T A A … 3' 3'… A A T▲T … 5'	*Eco*O109I	5'… R G▼G N C C Y … 3' 3'… Y C C N G▲G R … 5'
Mix d'enzymes 2	*Nla*III	5'… C A T G▼… 3' 3'…▲G T A C … 5'	*Hpy*CH4V	5'… T G▼C A … 3' 3'… A C▲G T … 5'
Mix d'enzymes 3	*Msc*I	5'… T G G▼C C A … 3' 3'… A C C▲G G T … 5'	*Dde*I	5'… C▼T N A G … 3' 3'… G A N T▲C … 5'

▼▲ Sites de coupure

R = A/G
Y = C/T

Tableau 30: Mix utilisés pour la digestion enzymatique
Les sites de reconnaissance de chaque enzyme sont indiqués ainsi que leur site de coupure.

VI.2.4 Capture par les sélectors

L'intégralité de l'ADN fragmenté est dénaturée puis hybridée aux sondes biotinylées en présence d'un excès de vecteur (3x). Les cibles sont ensuite liées à des billes magnétiques de streptavidine (Dynabeads) *via* leur sélector biotinylé puis lavées dans des conditions stringentes afin que les fragments faiblement liés soient éliminés. Les cibles et le vecteur sont enfin circularisés par ligation. Les cibles sont isolées par dénaturation thermique dans 10 µL d'un milieu qui prépare à l'amplification qui va suivre (phi29 reaction buffer 1x et 10 µM d'hexamère aléatoire) et récupération du surnageant après aimantation.

VI.2.5 Amplification des cibles

Les cibles circularisées sont amplifiées par MDA avec des hexamères aléatoires (produits par IDT). Le produit d'amplification est dilué au 1/100^{e} puis réamplifié par 15 cycles de PCR (avec la Platinum Taq

DNA polymerase High Fidelity, Life Technologies) grâce à des amorces universelles, complémentaires à la séquence du vecteur (voir séquences en Annexe 5). A terme, celles-ci comporteront également les séquences MID d'identification d'échantillon et les séquences d'adaptateurs pour le séquençage.

VI.2.6 Contrôle de la sélection par qPCR

Le produit de MDA est dilué au 1/200^{e} pour être analysé en qPCR. Les 12 régions validées au préalable ont été étudiées, ainsi que l'élément répété Alu et 2 régions non-ciblées (Neg1 et Neg2) que l'on s'attend à voir disparaitre après sélection.

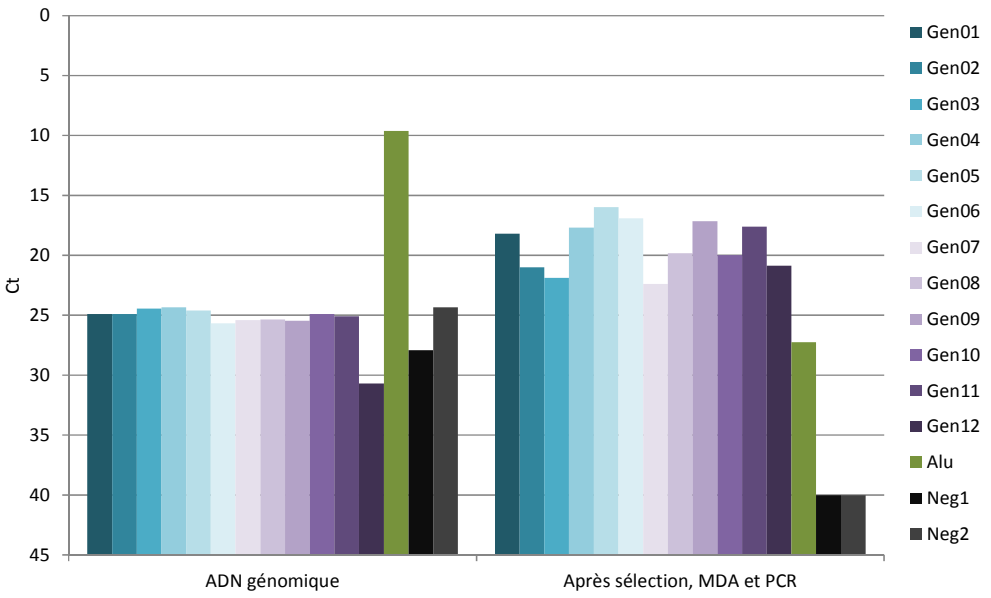


Figure 61: Résultats de la sélection standard par les sélectors
Cts obtenus par qPCR sur l'ADN Promega mâle à 1 ng/µL (ADN génomique) et sur ce même ADN après sélection par le protocole standard. Pour plus de clarté, nous avons attribué la valeur de Ct la plus tardive (40 cycles) aux amplicons pour lesquels aucun Ct n'a été obtenu. Neg1 et Neg2 amplifient des régions non-ciblées par les sélectors.

Après le protocole de sélection, les 12 amplicons que nous avions validés voient leur cycle de seuil diminuer et donc leur quantité augmenter (voir Figure 61), ce qui montre qu'ils ont bien été sélectionnés par les sélectors puis amplifiés par la MDA et la PCR qui ont suivi, tandis que les séquences contrôles non ciblées par les sélectors disparaissent complètement. La quantité d'élément *Alu*, quant à elle, diminue largement dans le milieu après sélection. Nous avons alors introduit le traitement par le bisulphite dans le protocole de sélection pour débuter la mise au point de notre technique.

VI.3 Introduction du traitement par le bisulphite

VI.3.1 Tests préliminaires

VI.3.1.1 Quantités utilisables pour le traitement par le bisulphite

Les quantités obtenues après sélection des cibles grâce aux sélectors sont très faibles, de l'ordre de quelques picogrammes, ce qui est peu courant pour un traitement par le bisulphite de sodium. Nous avons donc vérifié dans un premier temps que de telles quantités pouvaient être traitées efficacement (avec le kit Epitect Bisulfite, Qiagen). A la différence des tests réalisés dans le chapitre III, il ne s'agit pas ici de quantités d'ADN génomique mais d'amplicons. Nous avons donc d'abord amplifié par de nombreuses réactions de PCR une région du gène *IGF*2 sur de l'ADN Promega humain puis quantifié les amplicons résultants avant de les diluer pour en traiter au bisulphite des quantités décroissantes de 500 ng à 10 fg.

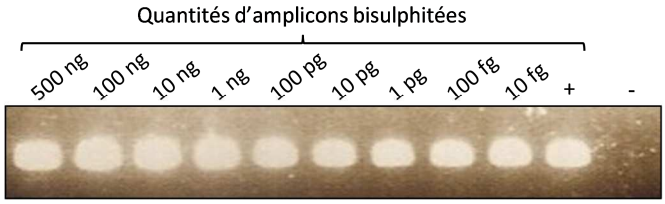


Figure 62: Traitement au bisulphite de quantités décroissantes d'amplicons
+ : Témoin ADN Promega humain traité au bisulphite (500 ng). - : témoin négatif de la PCR (pas d'ADN).

Nous avons ensuite amplifié l'ADN obtenu par PCR avec des amorces spécifiques à l'ADN bisulphité et ciblant l'amplicon d'intérêt, et vérifié sur gel d'agarose si en-deçà d'une certaine quantité initiale d'amplicons il n'était plus possible d'obtenir de matériel (voir Figure 62). La plus faible des quantités introduites dans le protocole de conversion par le bisulphite (10 fg) a permis d'obtenir une bande sur le gel. Ceci a confirmé que le traitement pouvait être réalisé sans biais sur les très faibles quantités dont nous diposerons.

VI.3.1.2 Contrôle par qPCR

De la même façon que pour l'ADN génomique, nous avons conçu des couples d'amorces permettant d'amplifier des régions ciblées par les sélectors et converties par traitement au bisulphite afin de contrôler la sélection. Une qPCR sur notre ADN bisulphité nous a permis de sélectionner 6 de ces couples dont les séquences figurent en Annexe 6 (voir Figure 63), les autres conduisant à des Cts trop tardifs pour être utilisés comme contrôles.

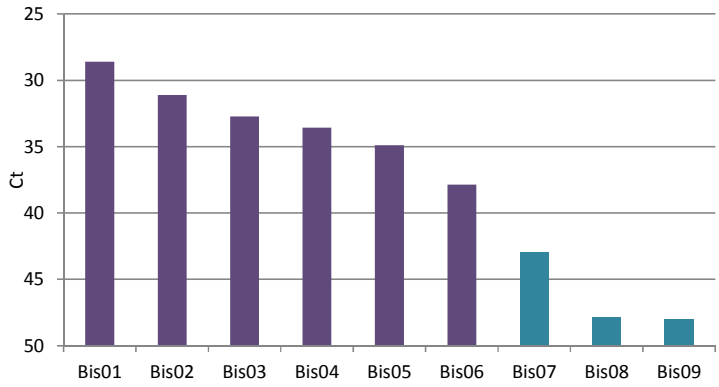


Figure 63: Cycles de seuil obtenus sur de l'ADN bisulphité avec nos couples d'amorces tests
Cts obtenus par qPCR sur de l'ADN Promega mâle à 20 ng/µL traité au bisulphite de sodium. Bis07 à Bis09 seront écartés tandis que Bis01 à 06 seront conservés pour le contrôle de la sélection suivie du traitement au bisulphite.

VI.3.2 Introduction du traitement par le bisulphite par plusieurs biais

Le traitement au bisulphite de sodium est introduit après l'étape de ligation. Nous l'avons testé sur les molécules d'ADN circulaire éluées des billes dans le tampon B&W4x, tampon de dilution des billes magnétiques, ainsi que sur l'ADN circulaire encore lié aux billes (voir Figure 64A). Dans ce dernier cas, la dénaturation a lieu pendant le traitement au bisulphite et les billes sont éliminées avant de déposer l'échantillon sur les colonnes qui interviennent dans le traitement.

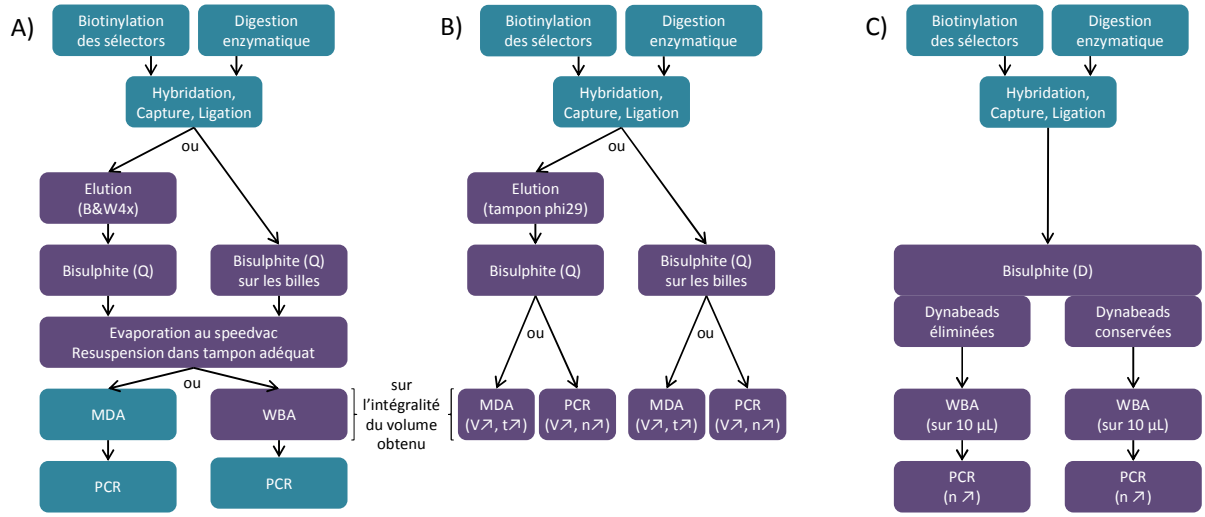


Figure 64: Schémas d'expérimentations testées pour l'introduction du traitement au bisulphite
En bleu, étapes existant dans le protocole standard. En violet, nouvelles étapes. Q : kit Qiagen, D : kit Diagenode. A) Première série de tests réalisée. B) Deuxième série de tests réalisée. Dans les étapes d'amplifications, le volume (V) a été adapté pour s'affranchir de l'étape d'évaporation au speedvac, le temps d'incubation (t) a été augmenté de 15 minutes pour la MDA et le nombre de cycles de PCR (n) a été augmenté de 15 pour le protocole standard à 35. C) Troisième série de tests réalisée. Lors du traitement au bisulphite, les Dynabeads de la sélection ont été éliminées ou conservées pour l'étape de désulphonation sur billes IPure. Lors de la PCR, le nombre de cycles a été augmenté à 20.

Nous avons ensuite évaporé les volumes résultants au speedvac pendant 2h30 afin de resuspendre le premier réplicat dans le mix de MDA pour son amplification avec le protocole standard, et le second

réplicat dans 10 µL de tampon d'élution (EB, Qiagen) dans le but de tester une amplification du bisulfitome (WBA) avec le kit EpiTect Whole bisulfitome (Qiagen). Cette technique fonctionne sur le principe de la MDA. Les produits de ces premières amplifications *via* des amorces aléatoires (dilués au 1/100^{e}) ont alors été amplifiés par PCR avec les amorces complémentaires du vecteur, de la même façon que dans le protocole standard.

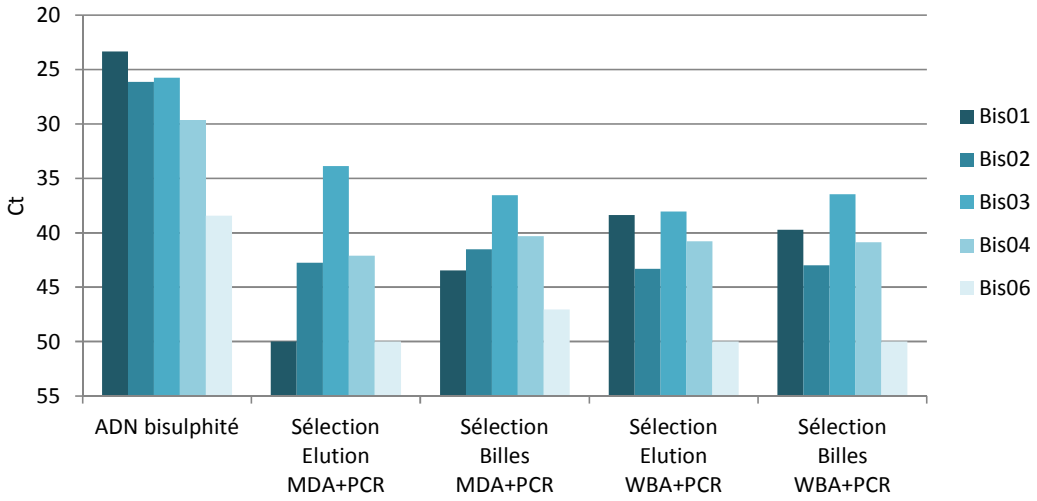


Figure 65: Résultats d'une capture par les sélectors suivie du traitement par le bisulphite
Cts obtenus par qPCR sur l'ADN Promega mâle à 10 ng/µL (ADN bisulphité) et sur ce même ADN génomique après sélection et traitement par le bisulphite dans différentes conditions. Pour plus de clarté, nous avons attribué la valeur de Ct la plus tardive (50 cycles) aux amplicons pour lesquels aucun Ct n'a été obtenu. Elution : l'ADN a été élué des billes dans le tampon B&W4x après la ligation et a subi le traitement par le bisulphite. Billes : le traitement par le bisulphite a été réalisé directement sur les billes après la ligation, sans élution. MDA+PCR ou WBA+PCR : modes d'amplification utilisés après traitement au bisulphite.

Après contrôle par qPCR, tous nos échantillons ont mené à des Cts très tardifs (inférieurs à 40) qu'il est impossible d'exploiter et qui nous ont laissé penser que la sélection et/ou les amplifications n'avaient pas fonctionné (voir Figure 65). Un contrôle sur une puce Agilent (High Sensitivity DNA Assay) de tous les échantillons après chaque étape d'amplification, a confirmé l'absence de matériel. Nous avons alors soupçonné la longue évaporation au speedvac d'être responsable de la dégradation de l'ADN (car l'appareil monte en température après un certain temps) et également le tampon utilisé pour l'élution de ne pas y être adapté.

Nous avons donc débuté une nouvelle série de tests (voir Figure 64B) en effectuant d'une part l'élution dans le tampon utilisé dans le protocole standard (tampon phi29 de la MDA) et en augmentant d'autre part les volumes réactionnels des amplifications afin d'utiliser tout l'ADN obtenu après traitement au bisulphite sans avoir à réduire son volume au speedvac. Nous avons également écarté la WBA pour nous concentrer sur l'amplification par MDA, en allongeant l'incubation à 1h45 (au lieu de 1h30), et sur la PCR utilisée seule et non pas après MDA comme dans le protocole standard, en augmentant le nombre de cycles de 15 à 35. La MDA effectuée seule a pour but de

vérifier s'il est possible d'obtenir un produit à cette étape mais il est clair que ce type d'amplification ne pourra, à terme, être utilisé seul : une étape de PCR est nécessaire pour ajouter les adaptateurs de séquençage *via* les amorces.

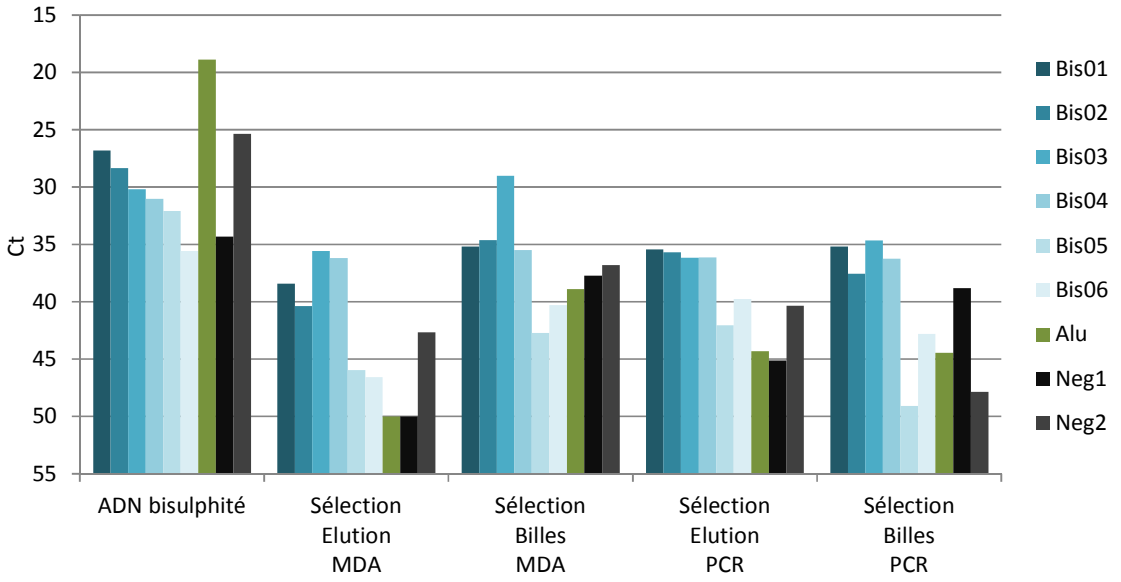


Figure 66: Résultats d'une capture par les sélectors suivie du traitement par le bisulphite (2)
Cts obtenus par qPCR sur l'ADN Promega mâle à 10 ng/µL (ADN bisulphité) et sur ce même ADN génomique après sélection et traitement par le bisulphite dans différentes conditions. Pour plus de clarté, nous avons attribué la valeur de Ct la plus tardive (50 cycles) aux amplicons pour lesquels aucun Ct n'a été obtenu. Elution : l'ADN a été élué des billes dans le tampon phi29 après la ligation et a subi le traitement par le bisulphite. Billes : le traitement par le bisulphite a été réalisé directement sur les billes après la ligation, sans élution.

Les mêmes conclusions ont pu être tirées quant aux résultats de qPCR obtenus (voir Figure 66) : aucune amplification ne semble avoir été efficace, quel que soit le protocole employé, puisque tous les amplicons étudiés après sélection présentent des Cts beaucoup plus tardifs que ceux de l'ADN bisulphité contrôle. Les contrôles négatifs, de plus, fournissent des valeurs de Cts du même ordre que ceux de nos cibles.

Nous avons alors envisagé que l'ADN circulaire reste bloqué sur les colonnes utilisées dans le traitement par le bisulphite. Nous avons également émis l'hypothèse que les conditions d'élution de ces colonnes n'étaient encore pas idéales ou qu'un composé présent dans ces colonnes et se retrouvant dans le milieu final, inhibait les amplifications par MDA ou PCR. Cette dernière hypothèse a été confirmée en éluant sur ces colonnes un produit de sélection par le protocole standard (donc sans traitement au bisulphite) puisqu'aucune amplification, que l'on sait pourtant optimisée, n'a pu être obtenue en utilisant l'intégralité de l'éluat. Il ne faudra donc utiliser que quelques µL du produit bisulphité pour les diverses amplifications, comme le montrent les meilleurs résultats obtenus sur certains amplicons en n'utilisant que 2 µL d'ADN bisulphité pour la PCR (35 cycles, voir Figure 67).

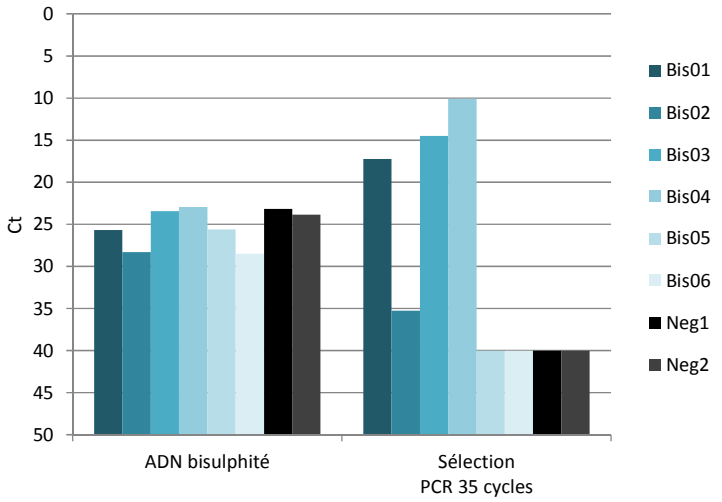


Figure 67: Amélioration des résultats de capture par diminution du volume utilisé dans la PCR
Cts obtenus par qPCR sur l'ADN Promega mâle à 10 ng/µL (ADN bisulphité) et sur ce même ADN génomique après sélection et traitement par le bisulphite puis PCR (35 cycles) sur 2 µL du volume résultant. La qPCR est désormais optimisée avec le QuantiFast kit (Qiagen). Pour plus de clarté, nous avons attribué la valeur de Ct la plus tardive (40 cycles) aux amplicons pour lesquels aucun Ct n'a été obtenu.

Nous avons ensuite souhaité tester un nouveau kit de traitement par le bisulphite fourni par Diagenode (MagBisulfite kit). Celui-ci a pour avantage de proposer une étape de désulphonation (dernière étape du traitement au bisulphite) sur des billes IPure et non plus sur colonne. Nous avons utilisé les réactifs de conversion sur les cibles liées à leurs billes Dynabeads puis nous avons éliminé ces billes avant la suite du traitement sur les billes IPure (voir Figure 64C).

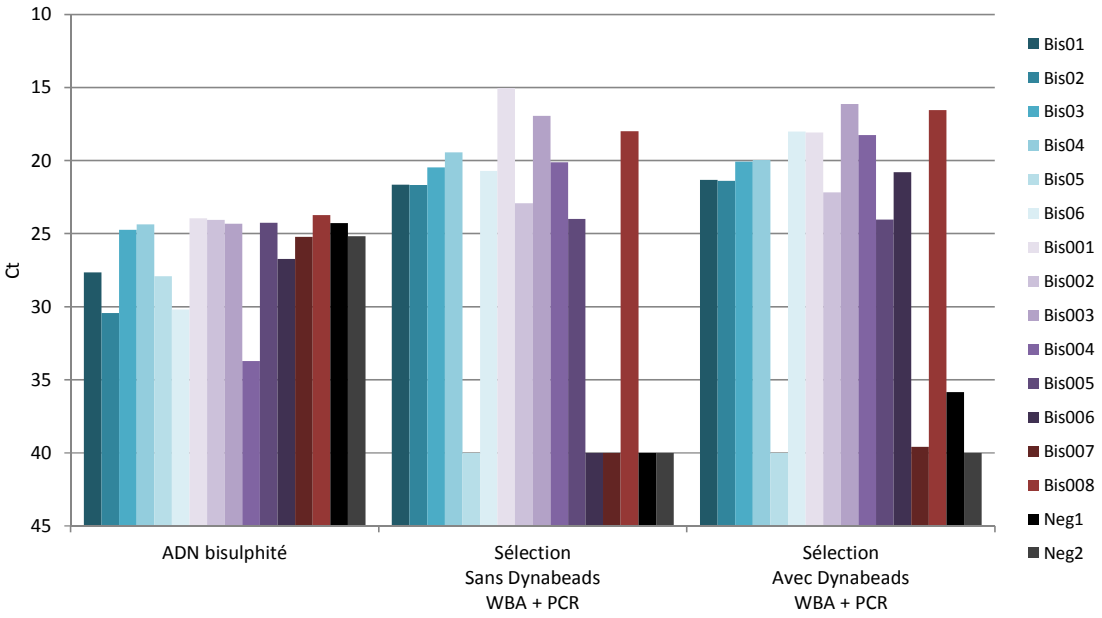


Figure 68: Résultats d'une capture suivie d'un nouveau traitement par le bisulphite
Cts obtenus par qPCR sur l'ADN Promega mâle à 10 ng/µL (ADN bisulphité) et sur ce même ADN génomique après sélection et traitement par le bisulphite avec le kit MagBisulfite (Diagenode) dans 2 conditions : avec et sans les Dynabeads de l'étape de sélection pour l'étape de désulphonation sur billes IPure. Pour plus de clarté, nous avons attribué la valeur de Ct la plus tardive (40 cycles) aux amplicons pour lesquels aucun Ct n'a été obtenu.

Nous avons en parallèle poursuivi le traitement sans élimination des billes Dynabeads. Les interrogations sur les étapes d'amplification ayant trouvé une réponse, nous avons alors réintroduit le kit de WBA en réduisant à 10 µL le volume d'ADN bisulphité à amplifier et nous l'avons fait suivre d'une amplification par PCR sur 2 µL du produit de WBA dilué au 1/100^{e} en n'utilisant que 5 cycles supplémentaires par rapport au protocole standard, soit 20 cycles. Nous avons également conçu de nouvelles amorces pour la qPCR (nommées Bis001 à 012, voir séquences en Annexe 6) afin d'analyser la sélection d'un plus grand nombre de cibles.

La grande majorité des fragments ciblés ont alors montré avoir été sélectionnés et amplifiés avec succès par ce protocole, que les billes Dynabeads soient présentes ou non durant la fin du traitement par le bisulphite (voir Figure 68). Nous avons donc conservé ces protocoles pour la suite des expérimentations.

VI.3.3 Protocoles appliqués à des individus HapMap en vue du séquençage

En vue de séquencer les fragments sélectionnés et amplifiés et de valider notre approche, nous avons appliqué le protocole présenté précédemment à deux individus HapMap. L'avantage de ces échantillons est que les polymorphismes présents dans leurs génomes sont connus et pourront être utilisés pour vérifier la conservation des ratios alléliques après sélection. Ce n'est pas le cas pour l'ADN Promega sur lequel tous les tests précédents ont été effectués puisque cet ADN est obtenu à partir de plusieurs donneurs anonymes.

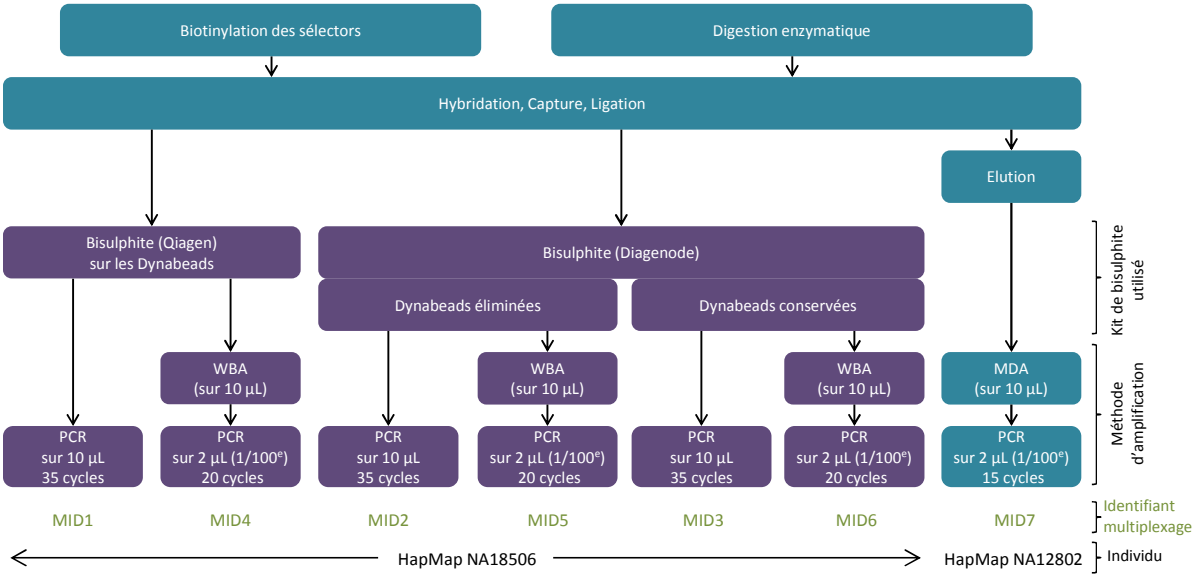


Figure 69: Conditions expérimentales utilisées pour la sélection et le traitement par le bisulphite

Afin de comparer la qualité des conversions par le bisulphite des deux kits utilisés ainsi que les différentes amplifications qui les suivent et établir ainsi un protocole avec les meilleures de toutes les

conditions, nous avons réalisé plusieurs sélections en parallèle sur l'ADN issu de l'individu NA18506 dont les conditions de traitement sont exposées dans la Figure 69. 500 ng d'un ADN lambda (Fermentas) fragmenté à 500 pb au Covaris ont été introduits durant le traitement par le bisulphite afin de réduire la dégradation de nos cibles. Dans les deux conditions de traitements par le bisulphite (sur billes ou colonne), l'élution finale a été réalisée dans du Tris-HCl à pH 8. L'individu NA12802 a subi le protocole standard afin de vérifier l'efficacité de nos sélectors et de confirmer que le vecteur conçu pour cette étude est optimal (l'ADN de l'individu NA18506 n'était plus disponible pour effectuer ce contrôle sur le même échantillon). Les amorces de PCR utilisées ici possèdent, en plus des séquences complémentaires au vecteur, les séquences des adaptateurs qui permettront la PCR en émulsion ainsi que l'hybridation de l'amorce de séquençage. Elles incluent également les séquences identifiantes MID1 à 7 permettant le multiplexage.

Les résultats des qPCR de contrôle obtenus à l'issue des protocoles respectifs sont présentés dans la Figure 70A. De façon générale, les différents protocoles ont permis la sélection et l'amplification des cibles, même si toutes n'ont pas donné lieu à une amplification en qPCR. Seul l'échantillon portant le MID1, traité au bisulphite par le kit Qiagen et amplifié par PCR sort du lot et ne fournit aucun résultat satisfaisant (voir Figure 70A2).

Nous avons ensuite vérifié les profils de taille et de concentration des échantillons au Bioanalyzer (puce High Sensitivity, Agilent). Sans grande surprise, l'échantillon portant le MID1 n'a montré que très peu de matériel tandis que celui correspondant au MID7 qui a subi le protocole standard a confirmé le succès de la sélection (voir Figure 70B). La condition d'amplification par PCR (MIDs 2 et 3) a amplifié préférentiellement les courts fragments tandis que précédée du WBA (MIDs 4 à 6) elle semble avoir favorisé les longs fragments, bien que la distribution de taille soit plus homogène lorsque le traitement par le bisulphite a été réalisé avec le kit Qiagen (MID4). On observe également que la distribution de tailles correspond parfaitement aux tailles théoriques des amplicons obtenus (de 311 à 584 pb, voir Figure 59B). Les échantillons peuvent donc être séquencés pour confirmer ces premières observations sur les différents protocoles utilisés, l'intérêt du séquençage de l'échantillon portant le MID1 restant à confirmer.

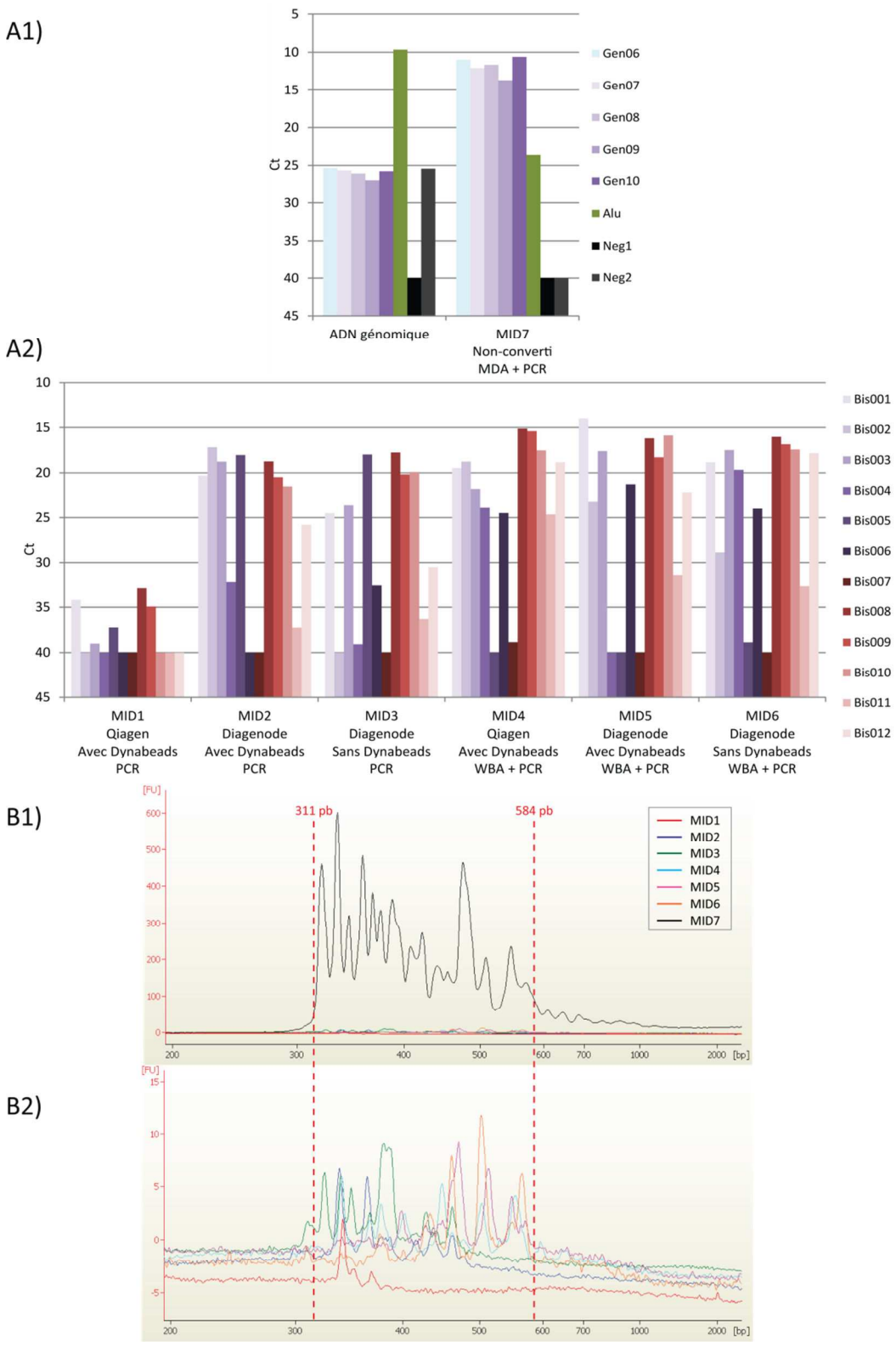


Figure 70: Résultats d'une capture par les sélectors suivie du traitement par le bisulphite
A) Cts obtenus par qPCR après divers protocoles. Pour plus de clarté, nous avons attribué la valeur de Ct la plus tardive (40 cycles) aux amplicons pour lesquels aucun Ct n'a été obtenu. Les échantillons son identifiés par la séquence MID qu'ils portent. A1) Résultats obtenus sur l'échantillon HapMap NA12802 après sélection et amplification. A2) Résultats obtenus sur l'échantillon HapMap NA18506 après sélection, traitement par le bisulphite (avec les kits Qiagen et Diagenode, avec ou sans les billes Dynabeads du protocole de sélection) et amplification (PCR ou WBA + PCR) dans diverses conditions. B) Profils de tailles obtenus au Bioanalyzer. B1) Superposition de tous nos échantillons. B2) Zoom sur les échantillons traités au bisulphite.

VI.4 Séquençage bidirectionnel

La préparation des échantillons pour le séquençage débute par la PCR en émulsion qui va permettre d'amplifier chaque fragment sur une bille. Pour cela, il est nécessaire de quantifier précisément le nombre de molécules dont on dispose dans chaque échantillon afin de les introduire en quantités équivalentes dans le milieu réactionnel. Ceci a été réalisé sur le Bioanalyzer grâce aux profils présentés ci-avant et nous avons obtenu les valeurs figurant dans le Tableau 31.

Identifiant échantillon	Concentration (pg/µL)	Taille moyenne (pb)	Millions de molécules/µL
MID1	7,37	344	19,6
MID2	89,68	386	213,0
MID3	178,77	386	424,6
MID4	113,88	443	235,7
MID5	184,31	459	368,2
MID6	137,51	481	262,1
MID7	7065,91	414	15648,2

Tableau 31: Quantification des échantillons après sélection et amplification
Le nombre de molécules a été calculé en utilisant la taille moyenne fournie par l'appareil et obtenue sur l'ensemble de nos amplicons.

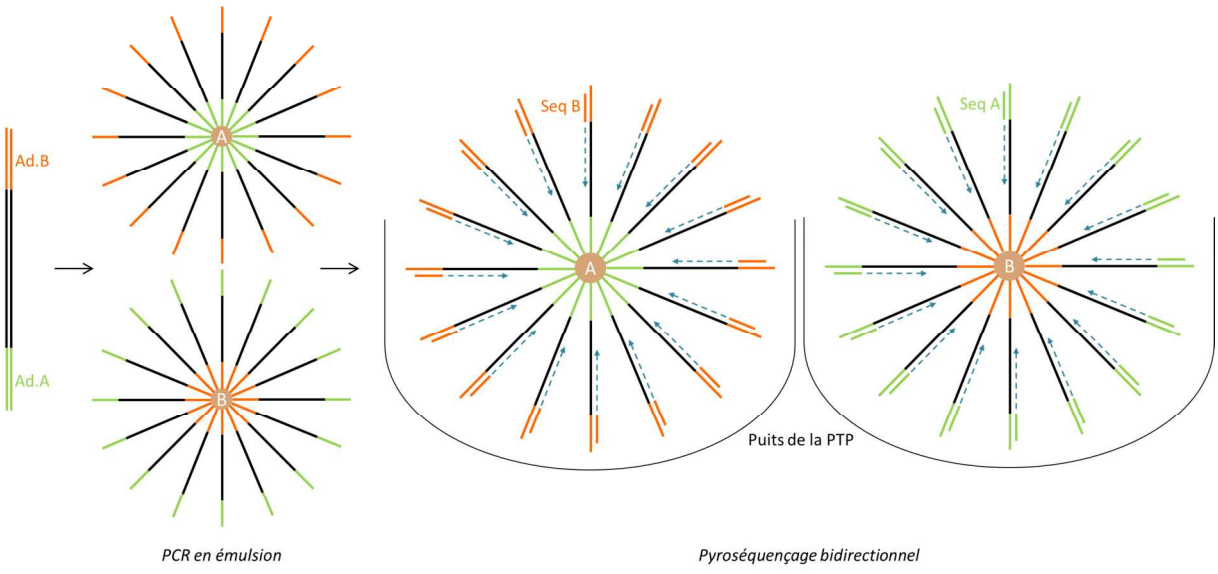


Figure 71: Séquençage bidirectionnel sur le GS Junior avec la technologie 454
Les fragments possédant les adaptateurs de séquençage (Ad.A et B) sont amplifiés par PCR en émulsion sur des billes A et B respectivement (disques marrons, l'échelle n'étant pas respectée). Les amorces de séquençage (Seq B et A respectivement) sont ensuite hybridées aux extrémités libres des fragments puis les billes sont déposées sur la PTP, une bille occupant un puits de ce support. Le pyroséquençage s'effectue alors grâce aux réactifs contenus dans les puits (flèches bleues en pointillés).

L'échantillon portant le MID1 étant très peu concentré par rapport aux autres, nous l'avons finalement exclu pour la suite du protocole. Cela porte donc à 6 le nombre d'échantillons à

séquencer en parallèle. Le nombre moyen de *reads* pouvant être obtenu sur cette plateforme étant estimé à 70000, nous envisageons ainsi d'obtenir une couverture maximale d'environ 120x (sur chacun des 98 amplicons). Nous avons donc dilué les 6 échantillons portant les MIDs 2 à 7 pour préparer la PCR en émulsion. Celle-ci a été réalisée en parallèle sur 2 types de billes (GS Junior Titanium emPCR kit, Lib-A, Roche) afin d'effectuer le séquençage de façon bidirectionnelle : chaque fragment pourra être séquencé par ses 2 extrémités dans des puits différents de la PTP (voir Figure 71). Le séquençage sur le GS Junior dure 10 heures.

VI.5 Exploitation des données de séquençage

VI.5.1 Qualité du séquençage

Les premiers nucléotides séquencés sont ceux correspondant à la clé (voir Figure 59B). Celle-ci permet l'identification du puits dans lequel est située la bille et intervient également dans les calibrations internes de l'appareil au début du séquençage, ce qui constitue un premier filtre des séquences. A l'issue du séquençage, le logiciel fourni par Roche applique un second filtrage des *reads* identifiés et fournit leur distribution de tailles. Dans notre cas, près de 86000 *reads* exploitables ont été obtenus (voir Figure 72a), dépassant ainsi la moyenne généralement admise des 70000. De plus, la gamme de tailles couvertes correspond parfaitement à celle que nous attendions d'après les longueurs de nos cibles (voir Figure 72). En revanche, leur distribution n'est pas homogène : les plus longs fragments sont sous-représentés, ce qui laisse penser que la PCR en émulsion et/ou la PCR lors de la préparation des échantillons ont amplifié de façon préférentielle les fragments de tailles plus modérées (250 à 350 bases environ).

Les séquences des *reads* débutent par la séquence du MID et les séquences universelles (correspondant au vecteur). La première étape de l'analyse consiste donc à rogner les nucléotides de ces séquences et à démultiplexer, c'est-à-dire identifier les *reads* appartenant à chaque échantillon afin de créer un répertoire contenant les séquences des cibles bisulphitées pour chacun d'entre eux. Près de 40% des *reads* ont été attribués à l'échantillon portant le MID7 qui correspond au protocole de sélection standard (voir Figure 72b), ce qui valide les constructions que nous avons créées pour cette étude. Concernant les protocoles où le traitement par le bisulphite a été inclus, ceux pour lesquels la PCR a été l'unique mode d'amplification (MIDs 2 et 3) ont chacun mené à moins de 10% des *reads*. Les échantillons ayant subi une amplification par WBA avant la PCR représentent en revanche près de 18% des *reads* pour ceux où le traitement par le bisulphite a été réalisé en conservant les Dynabeads (MIDs 4 et 6), mais seulement 10% dans le cas où ces dernières ont été supprimées (MID5).

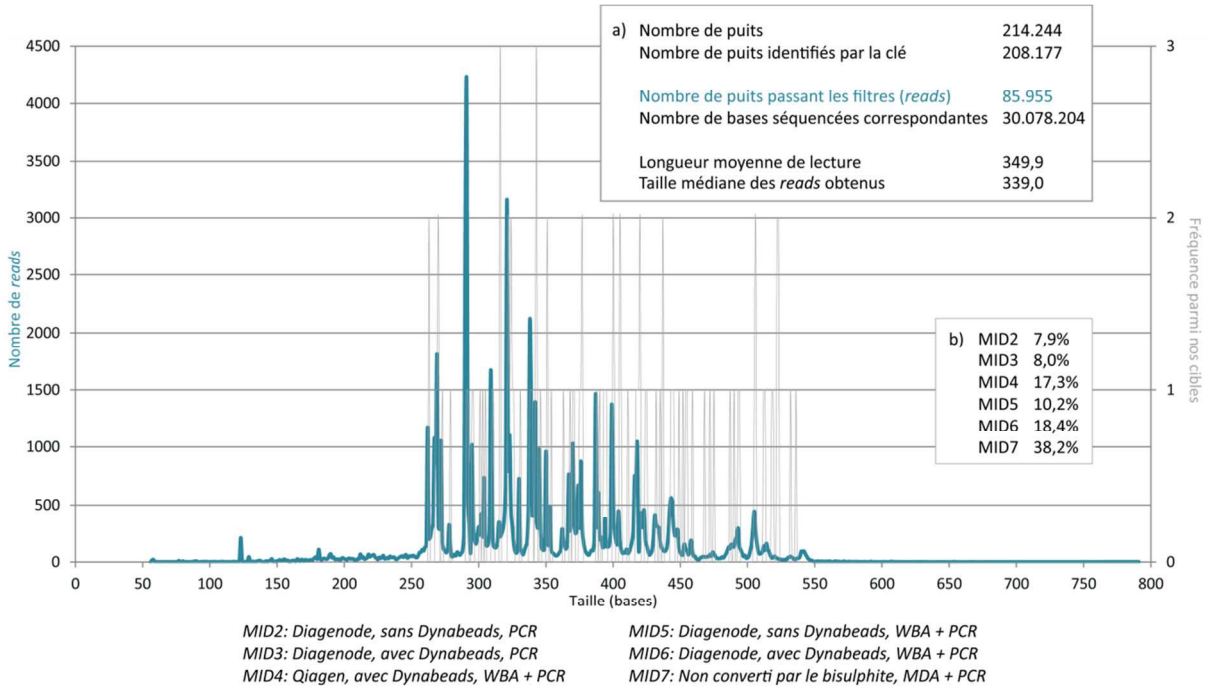


Figure 72: Distribution des *reads* obtenus par séquençage sur le GS Junior
En bleu : distribution de tailles des 85955 *reads* obtenus. Nous l'avons superposée à la distribution théorique des tailles de nos fragments (en gris). a) Statistiques fournies par le logiciel de Roche. b) Pourcentages de *reads* attribués à chaque échantillon, identifié par le MID qu'il porte, après démultiplexage. Pour rappel, les différents protocoles associés aux échantillons portant les MIDs 2 à 7 sont indiqués.

VI.5.2 Profondeur de séquençage

Nous nous sommes intéressés à la profondeur de séquençage obtenue sur l'ensemble de nos cibles. Après les différents protocoles incluant un traitement par le bisulphite (MIDs 2 à 6), près de 80% des bases ciblées ont été couvertes au moins 1 fois pour la meilleure des conditions (MID4) (voir Figure 73).

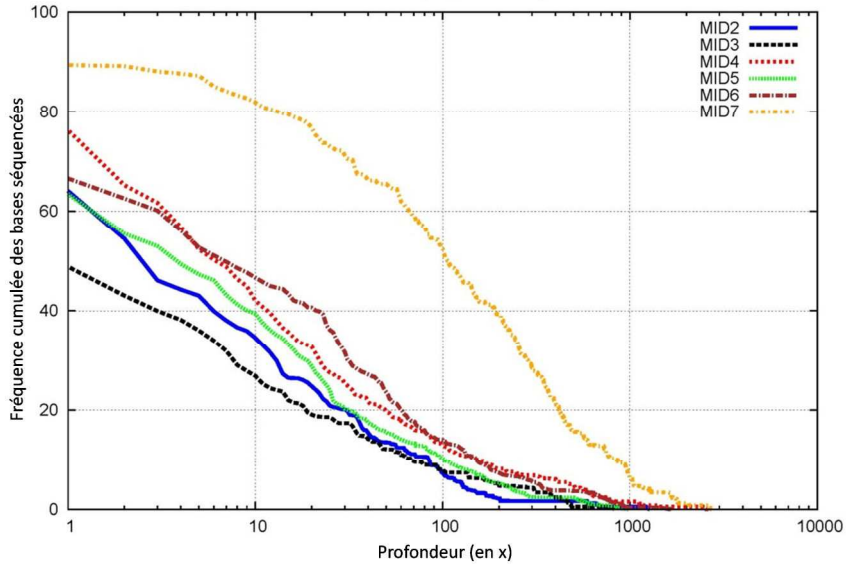


Figure 73: Profondeurs de séquençage atteintes avec les différents protocoles

Cependant, quel que soit le protocole mis en jeu, il existe un biais entre les séquences puisque toutes les bases ne sont pas couvertes de façon équivalente. Il sera donc nécessaire de séquencer davantage pour obtenir une profondeur suffisante pour la suite de l'analyse. On peut par exemple la fixer à 10x au minimum si l'on désire une résolution quantitative de 10%. Cette profondeur est en revanche atteinte pour plus de 80% des bases séquencées après le protocole standard de sélection, prouvant là encore que les sélectors, le vecteur et les amorces de PCR ont été choisis de manière optimale.

VI.5.3 Qualité des traitements au bisulphite et des diverses amplifications

Notre analyse a été approfondie en comparant les différents protocoles qui incluent un traitement par le bisulphite. Nous avons pour cela utilisé BiQ, un outil développé spécifiquement pour l'étude de séquences bisulphitées (195). Pour chacun des échantillons, nous avons cherché à savoir comment nos *reads* s'alignent sur les séquences de référence. Nous avons d'abord pu constater que 90 cibles sur les 98 initiales étaient représentées dans nos séquences. L'amplification par WBA suivie d'une PCR (MIDs 4 à 6) a permis d'obtenir davantage de séquences en comparaison à la PCR utilisée seule (Figure 72b), et les *reads* correspondants ont été plus facilement alignés de façon unique sur nos références (voir Figure 74).

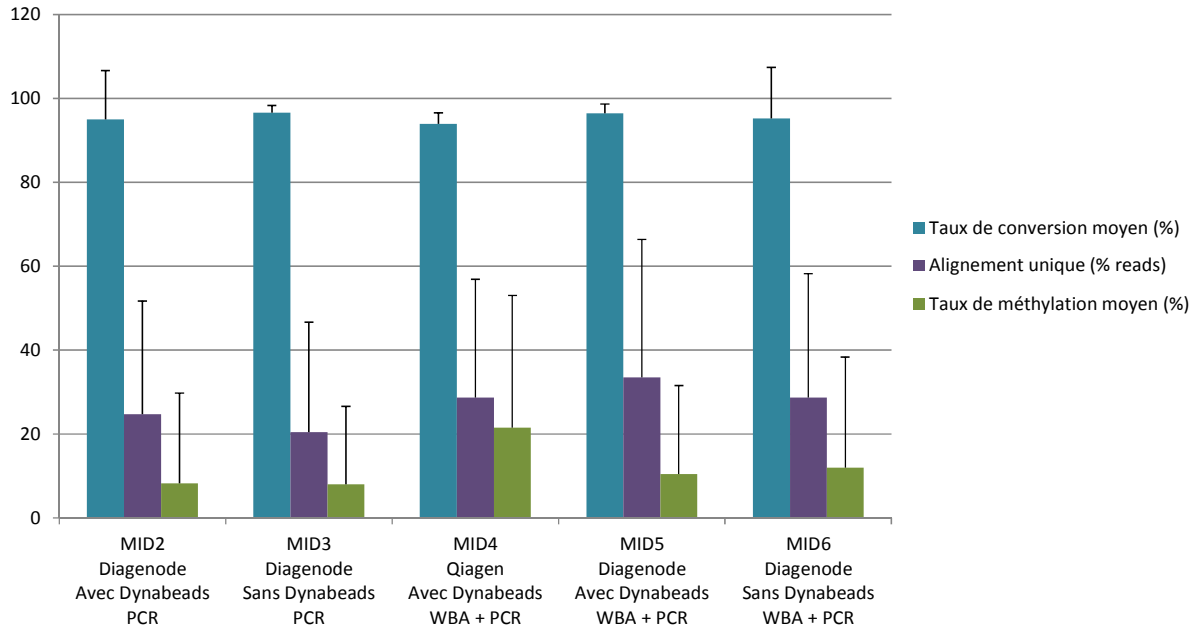


Figure 74: Statistiques fournies par BiQ après séquençage
Moyennes obtenues sur l'ensemble des séquences issues d'un échantillon identifié par son MID. Les barres d'erreur indiquent l'écart-type sur ces séquences. Taux de conversion moyen : pourcentage de cytosines hors CpG converties en thymines, moyenne des chiffres obtenus sur les 98 amplicons après alignement. Alignement unique : nous avons isolé les *reads* qui s'alignent sur chacune des 98 séquences de référence ; parmi eux, nous avons calculé le pourcentage de ceux qui s'y alignent de façon unique (et donc pas également sur une autre séquence de référence). Nous avons ensuite calculé la moyenne de ces chiffres sur les 98 amplicons. Taux de méthylation : pourcentage de cytosines dans le contexte CpG non-converties en thymines, moyenne des chiffres obtenus sur les 98 amplicons.

Concernant la performance des deux traitements au bisulphite, BiQ nous a également permis de vérifier le taux de conversion de nos fragments, c'est-à-dire le pourcentage de cytosines hors CpG converties en thymines. Ce chiffre avoisine les 95% en moyenne sur tous nos échantillons ; seul celui portant le MID4 (traitement avec le kit Qiagen) a montré un léger signe de faiblesse (93,9%, voir Figure 74). Enfin, BiQ offre la possibilité de quantifier la méthylation au niveau de chaque CpG de nos cibles. Pour parvenir à des résultats exploitables, une couverture minimale est requise puisque le logiciel prend en compte le nombre de *reads* couvrant une position CpG pour en calculer son pourcentage de méthylation. Nous avons donc fixé à 10x cette valeur minimale et avons, dans ces conditions, pu obtenir des valeurs de méthylation faibles mais homogènes entre les échantillons. Seul l'échantillon portant le MID4 s'est à nouveau démarqué des 4 autres échantillons par une valeur supérieure, certainement due à son plus faible taux de conversion, mais aussi une variabilité plus importante (voir Figure 74). Ceci nous a donc amenés à exclure le traitement au bisulphite avec le kit Qiagen pour les expérimentations futures. De façon générale, les valeurs de méthylation obtenues restent très faibles, voire nulles, mais ceci était prévisible puisque nous avons ciblé dans cette étude des promoteurs et îlots CpG sur des gènes suppresseurs de tumeurs, sensés être non-méthylés dans de l'ADN issu du sang, comme c'est le cas pour les échantillons HapMap.

VI.6 Discussion

La technologie des sélectors développée par nos collaborateurs a fait ses preuves comme outil de capture de régions d'intérêt (176,178,226). Nous l'avons donc utilisée dans la mise en place d'un protocole permettant pour la première fois l'analyse de la méthylation sur des fragments cibles grâce au traitement par le bisulphite. La réduction de la complexité de séquence qu'engendre ce traitement peut provoquer une hybridation non-spécifique des sélectors et il reste difficile de concevoir des sondes optimales, surtout pour cibler les régions riches en CpGs. Nous avons donc choisi de capturer les loci d'intérêt avant leur conversion par le bisulphite.

Pour réaliser une preuve-de-principe de cette méthode, nous avons conçu 98 sélectors permettant de cibler des fragments de tailles variables, comprises entre 200 et 500 pb. Cette étape de préparation de notre étude a été réalisée avec un logiciel actuellement non disponible pour la communauté scientifique. Il sera cependant possible d'utiliser PieceMaker, un outil proposant différentes enzymes de restriction et les séquences de sélectors correspondantes pour la sélection (227) ou d'opter pour des séquences de sondes figurant dans une base de données qui en contient plus de 21 millions (228). Il faut toutefois noter que la conception des sondes est une étape cruciale à laquelle il sera nécessaire de consacrer du temps afin de n'introduire aucune erreur de séquence dans les sondes choisies car le décalage d'une seule base conduit à l'échec du protocole.

Après capture par les sélectors, les quantités d'ADN obtenues sont infimes et le sont encore davantage après traitement par le bisulphite. Il est donc nécessaire d'introduire des étapes d'amplification. Une PCR est indispensable afin d'ajouter aux extrémités des cibles les séquences d'adaptateurs de séquençage *via* les amorces. Nous l'avons également fait précéder d'une amplification aléatoire par MDA (229). Un équilibre doit être trouvé entre ces deux types d'amplifications car il est connu et admis que la PCR favorise l'amplification des petits fragments dans un milieu où plusieurs tailles sont représentées et ce biais existe davantage sur de l'ADN bisulphité (230). De plus, le traitement par le bisulphite aura déjà fragilisé voire endommagé les plus longs fragments (169).

Le protocole standard de sélection développé par nos collaborateurs a été suivi jusqu'à l'étape de ligation. Nous avons ensuite utilisé différentes techniques de conversion par le bisulphite, sur colonne avec le kit fourni par Qiagen ou sur billes avec celui de Diagenode ainsi que différents modes d'amplification. Nous avons mis en place et optimisé des contrôles par qPCR. Ceux-ci ont permis de confirmer que la majorité des régions d'intérêt ont pu être sélectionnées, traitées par le bisulphite puis amplifiées grâce aux protocoles testés, ce qui a validé le passage à l'étape de séquençage, afin de déterminer quel protocole mène aux meilleurs résultats.

Le séquençage a été réalisé sur le GS Junior de Roche. L'utilisation d'un séquenceur de paillasse a été privilégiée puisque son rendement (50 Mb au maximum) est suffisant pour fournir la profondeur nécessaire à l'analyse de loci. Nous avons introduit des séquences MIDs sur nos amorces de PCR afin d'identifier chacun de nos sept échantillons. Etant donné la longueur de nos fragments, nous avons opté pour un séquençage bidirectionnel (avec le kit GS Junior Titanium emPCR Lib-A du fournisseur) et avons donc introduit les MIDs aux deux extrémités de nos fragments. Un prochain essai avec des cibles plus courtes pourra faire l'objet d'un séquençage unidirectionnel (avec le kit Lib-L du même fournisseur) et ne nécessitera l'introduction du MID que sur l'une des deux amorces de PCR. La première étape de préparation au séquençage est une PCR en émulsion qui peut, elle aussi, accentuer la surreprésentation des petits fragments pour conduire à des biais de séquences et donc de couverture. Dans notre cas, des fragments de tailles diverses sont isolés dans le milieu après la sélection et nous étions donc conscients, dès la conception de l'étude, que les longs fragments pourraient plus difficilement être pris en compte. Nous avons toutefois voulu vérifier quelle distribution de tailles était analysable et nos observations ont confirmé la nécessité de réduire ces tailles.

Les données de séquençage ont ensuite été exploitées avec BiQ (195). Ce logiciel propose une analyse approfondie en fournissant, en plus de l'alignement, une valeur précise de méthylation pour chaque position CpG des *reads* et nous nous sommes donc tournés vers cet outil qui nous a semblé le

plus complet en comparaison à BS Seeker (155), Bismark (231) ou d'autres logiciels qui ont fait l'objet d'une récente comparaison (232). Nous avons ainsi pu comparer les 6 échantillons séquencés, correspondant aux différents protocoles de préparation dont on cherche à valider les meilleures conditions. Le traitement par le bisulphite impliquant des colonnes n'a pas permis de fournir suffisamment de matériel pour le séquençage après amplification par PCR. Lorsque celle-ci a été précédée d'une amplication par WBA, nous avons pu montrer que le taux de conversion provoquée par le bisulphite de sodium était plus faible que celui obtenu après traitement avec un protocole sur billes. Le kit Epitect de Qiagen a donc été exclu pour la mise en place d'un protocole final. Concernant les modes d'amplification, le WBA suivi de la PCR a permis de fournir des séquences de meilleure qualité en alignant davantage de *reads* à leur unique séquence de référence. Cependant, les chiffres obtenus sur l'ensemble des amplicons présentent un écart-type important et il ne semble pas possible, avec ces seules données, de favoriser l'un ou l'autre de ces protocoles.

Afin de maximiser le rendement de séquençage, Roche préconise d'étudier des amplicons du même ordre de taille, à moins de les introduire en quantités inégales, notamment les plus longs fragments en quantité 4 fois supérieure aux plus courts, ce qui n'est pas réalisable dans notre cas puisque tous les fragments sont contenus dans un seul tube après sélection. Il est également envisageable de réaliser la PCR en émulsion dans deux, voire trois mix différents contenant les fragments regroupés par gammes de tailles. Ceci impliquerait, en amont de tout le processus, de partager les sélectors en fonction des tailles de fragments qu'ils ciblent pour effectuer deux ou trois sélections en parallèle, et cela multiplierait donc la quantité initale de matériel nécessaire. Cette option a donc été exclue car l'un des atouts de notre méthode est de n'utiliser que très peu de matériel, 750 ng, et nous envisageons d'ailleurs de diminuer cette quantité à 300 ng.

Afin d'éviter l'introduction de telles étapes qui ne garantissent en rien une amélioration conséquente de la qualité et du rendement de séquençage, nous avons conclu que la sélection devait désormais cibler des fragments de taille plus petite, soit 200 bases au maximum. Dans le but d'augmenter la profondeur de séquençage et la couverture de régions d'intérêt, il sera également intéressant de concevoir des sélectors présentant une certaine redondance. Nous préparons donc actuellement une nouvelle étude qui remplit ces conditions afin de pouvoir confirmer lequel de nos protocoles est le plus performant. De plus, pour concurrencer les technologies existantes permettant l'analyse de loci telle que la plateforme Fluidigm (analyse de 96 amplicons, cependant jamais appliquée à l'étude de la méthylation après traitement par le bisulphite), nous concevrons un millier de sélectors. Il sera alors nécessaire de faire appel à un séquenceur dont le rendement est plus important et le MiSeq, séquenceur de paillasse d'Illumina (rendement de 1 Gb à l'heure actuelle), s'est imposé comme le meilleur choix.

Chapitre VII: Discussion générale

Les travaux présentés dans les chapitres précédents tirent profit des évolutions spectaculaires ayant eu lieu depuis quelques années dans le milieu du séquençage de l'ADN. Nous discuterons dans un premier temps des conséquences de ce phénomène. Nous verrons ensuite comment le MeDIP-Seq a su s'imposer comme une méthode de choix en matière d'étude du méthylome puis nous expliquerons l'intérêt d'y avoir introduit une étape de déplétion des séquences répétées. Enfin, nous montrerons comment la technologie des sélectors a permis d'enrichir notre approche pour aboutir à une plateforme complète d'analyse de la méthylation.

VII.1 Technologies de séquençage

VII.1.1 Une évolution hors de contrôle

Depuis une dizaine d'années, les technologies de séquençage ont connu une évolution impressionnante, d'autant plus depuis 2005 avec l'arrivée des séquenceurs de seconde génération sur le marché (233). Ceci permet de diminuer de plus en plus les coûts du séquençage (voir Figure 5) tout en augmentant le débit fourni par les machines. Pour exemple, nous présentons dans la Figure 75 les changements clés qu'a connu la technologie Illumina depuis la conception du projet décrit dans ce manuscrit.

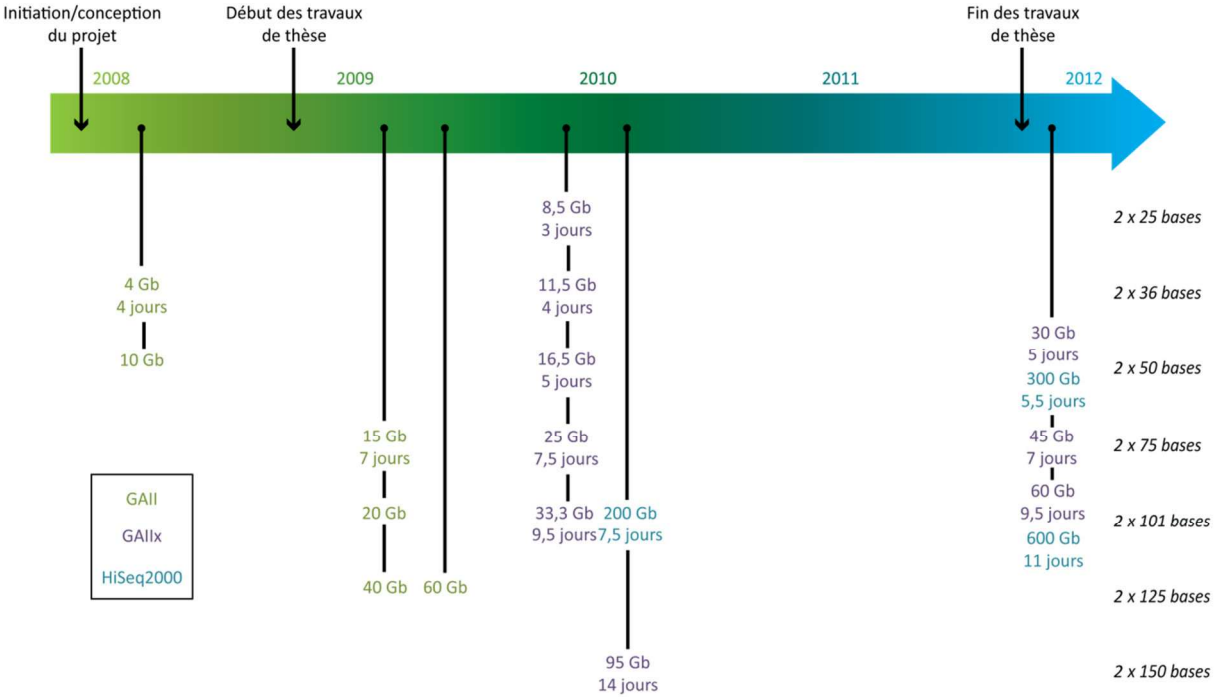


Figure 75: Evolution du séquençage Illumina en *paired-end*
Les évolutions les plus marquantes sont présentées, avec les longueurs de lecture associées. Rendements maximaux en Gb et durée associée du *run*. D'après des données fournies par Illumina.

Au premier trimestre 2008, le séquenceur Genome Analyzer d'Illumina existait sous sa version II et produisait 4 Gb de données en 4 jours (séquençage en *paired-end* sur 36 bases). Un an après le début des travaux de thèse, la version IIx de cette machine produisait plus de 2 fois plus de données dans le même temps. A la fin de ces travaux, le HiSeq2000 produit 75 fois plus de données en seulement 36 heures supplémentaires, et il faut désormais considérer que le HiSeq2500 sera bientôt capable de séquencer un génome entier en un jour seulement.

Ces évolutions engendrent des problèmes de diverses natures. Tout d'abord, les débits des machines sont tels que la préparation manuelle des échantillons devient une étape limitante dans le processus.

Les laboratoires qui se sont dotés d'un grand nombre de séquenceurs optent donc pour des systèmes robotisés afin d'accélérer cette phase en routine. Ceci permet aux expérimentateurs de disposer de davantage de temps pour organiser le lancement des machines, étape pendant laquelle aucun robot ne peut remplacer la main humaine car des gestes précis et minutieux sont requis ainsi que diverses vérifications et interventions propres à chaque nouveau *run*.

Un deuxième point à soulever est celui de la gestion des données. Le doublement de leur capacité de production tous les 9 mois a désormais dépassé les performances de stockage dont nous disposons et la gestion des térabytes de données engendrées à chaque *run* devient critique (234). Il est même parfois plus cher de stocker des données que de séquencer à nouveau un échantillon, particulièrement en prenant en compte le gain de qualité que pourront fournir les évolutions des technologies entre deux analyses. Ces challenges informatiques restent encore à relever.

VII.1.2 Choix d'une plateforme adaptée

Ces évolutions phénoménales mènent à la multiplicité des plateformes disponibles sur le marché (voir Figure 10). Les divers fournisseurs ont également, à partir de 2010, adapté leurs technologies respectives à des séquenceurs de plus petite taille qualifiés comme « de paillasse », certes générant des rendements plus faibles, mais aussi les rendant accessibles à de plus petits laboratoires.

L'application à laquelle est destiné le séquençage orientera le choix de l'une des machines. Dans notre cas, l'étude du méthylome nous a poussés à employer un séquenceur très haut-débit d'Illumina, le GAIIx, machine imposante (dimensions de 92×104×69 cm pour un poids de 181 kg) et dont les temps de *run* sont très longs (9,5 jours avec les paramètres que nous avons utilisés) mais qui permet de fournir la couverture et la profondeur requises pour une analyse sur le génome entier. Dans la seconde partie de ces travaux, nous avons en revanche préféré l'utilisation du séquenceur de paillasse de Roche, le GS Junior, petit appareil (dimensions de 40×60×40 cm pour un poids de 25 kg) qui permet d'obtenir rapidement (temps de *run* de 10 heures) un rendement suffisant pour l'étude d'une centaine d'amplicons. La combinaison de ces deux séquenceurs peut donc s'avérer être un choix judicieux pour étudier la méthylation à diverses échelles.

VII.2 Analyse de la méthylation par MeDIP-Seq

VII.2.1 Substitution aux puces à ADN

Les puces à ADN ont longtemps été considérées comme la méthode standard d'interrogation du génome pour l'analyse de la méthylation. Il faut par exemple citer les puces fournies par Agilent ou celles qui ont été développées pour l'étude de promoteurs par Roche Nimblegen ou Affymetrix, sur

lesquelles peuvent être hybridés des produits de MeDIP ou de sélection par MBD (235). Nous avons également évoqué au chapitre V la technologie Infinium développée par Illumina, basée sur une hybridation spécifique après traitement par le bisulphite, avec l'utilisation de la puce HumanMethylation27. Son optimisation a abouti à la puce HumanMethylation450 qui permet désormais d'étudier le statut de méthylation de plus de 480000 sites CpG (236). Ceci ne représente cependant que 2,2% des CpGs du génome humain et aucune puce n'est encore disponible pour étudier d'autres espèces.

L'utilisation des technologies de séquençage a rapidement supplanté celle des puces à ADN pour l'étude de la méthylation car elle a pour avantages de couvrir beaucoup plus de CpGs, de ne pas nécessiter de conception préalable en fonction du génome étudié (ce qui posait problème pour les puces lorsqu'aucune annotation n'était encore parfaitement connue) et de ne pas être soumise aux biais des hybridations croisées. De plus, les données obtenues sont quantitatives et non pas relatives (dues à une compétition dans la technologie des puces à ADN) et très peu de matériel, quelques nanogrammes, suffisent à une analyse complète (237). Tous ces arguments nous ont donc conduits, lors de la conception de ce projet, à opter pour le séquençage de nos échantillons après MeDIP. Il faut cependant noter que l'analyse sur puce demeure plus rapide et moins coûteuse et que la fabrication de puces multiplexes à façon restera toujours utile pour étudier des régions d'intérêt en aval du séquençage.

VII.2.2 Atouts du MeDIP-Seq

Dans le chapitre I, plusieurs techniques d'analyse du méthylome par séquençage ont été présentées, parmi lesquelles le MeDIP-Seq qui a fait l'objet de l'étude menée dans ce projet. En comparaison aux techniques reposant sur la conversion par le bisulphite, on a pu reprocher au MeDIP-Seq de ne pas être quantitatif et de ne permettre que la définition de niveaux de méthylation grossiers. Ceci n'est le cas qu'en l'absence d'outils informatiques adaptés et nous avons pu faire la preuve de l'intérêt du MeDIP-Seq dans la quantification de la méthylation grâce à la plateforme MeQA que nous avons créée. Il reste cependant vrai que les techniques basées sur la sélection par affinité (MeDIP-Seq et MBD-Seq ou MethylCap-Seq) n'ont qu'une résolution d'une centaine de bases quand le séquençage après traitement par le bisulphite fournit une résolution au nucléotide près. Néanmoins, cette précision a un coût et obtenir une profondeur suffisante pour exploiter ces données demeure extrêmement cher tandis que le MeDIP-Seq est financièrement beaucoup plus abordable (160). Le protocole de MeDIP est également automatisable, ce qui en fait un atout majeur pour des applications en routine.

Le RRBS impliquant un traitement par le bisulphite ne permet de couvrir au maximum que 10% des CpGs tandis que le MeDIP-Seq et le MBD-Seq sont théoriquement capables d'identifier des régions méthylées dispersées à travers tout le génome (160). Ces deux dernières méthodes sont donc adaptées à l'identification de DMRs, y compris dans des régions situées à l'extérieur des îlots CpG.

VII.2.3 Synergie avec MBD-Seq et MethylCap-Seq

La sélection par MBD et MethylCap qui utilisent la protéine MBD2 ou le domaine MBD de MeCP2 respectivement pour isoler les fractions méthylées du génome, ou encore la technique MIRA dans laquelle la protéine MBD3L1 augmente l'affinité de MBD2, sont complémentaires au MeDIP. Il a effectivement été montré que le MeDIP enrichit davantage les régions méthylées de faible densité en CpGs par rapport aux deux autres méthodes ; l'une des explications avancées résiderait dans le fait que l'étape de dénaturation avant immunoprécipitation ne serait pas complète dans les régions riches en CpGs qui se réhybrideraient très rapidement, les rendant alors inaccessibles à l'anticorps (238). En effet, celui-ci est spécifiquement dirigé contre les 5-méthylcytidines portées par l'ADN simple brin tandis que les protéines de la famille MBD reconnaissent la base méthylée sur l'ADN double brin natif.

Cependant cette conclusion peut se trouver inversée (159) en fonction des concentrations en sels utilisées durant l'élution, qui sont directement corrélées à la densité en CpGs des régions qu'elles enrichissent : dans une autre étude, il a été conclu que MeDIP-Seq et MethylCap-Seq enrichissent les mêmes fractions avec une tendance pour les régions riches en CpGs, la couverture du second étant légèrement supérieure (160). Dans tous les cas, la possibilité d'éluer des fractions moins méthylées voire très peu méthylées en diminuant la force ionique du tampon d'élution permet d'étudier des régions que l'anticorps utilisé dans le MeDIP-Seq ne pourrait sélectionner. Néanmoins, les protéines MBD se lient à l'ADN méthylé avec une certaine spécificité de séquence qui engendrera des biais (125).

Nous avons récemment mis en place un protocole de MethylCap-Seq. Une plus faible quantité d'ADN est requise en comparaison à notre protocole de MeDIP-Seq car les adaptateurs de séquençage sont fixés aux extrémités des fragments après sélection par la protéine. De plus, le fait de travailler sur de l'ADN double brin n'a pas soulevé les problèmes que nous avons rencontrés lors de l'étape de *sizing* sur gel. Il serait alors intéressant d'appliquer ce protocole aux mêmes échantillons de MEFs étudiés par MeDIP-Seq et de nous concentrer sur les densités en CpGs des régions sélectionnées afin de tirer nos propres conclusions quant à la spécificité des deux techniques.

VII.2.4 Etude des 5-hydroxyméthylcytosines

Si la 5-méthylcytosine a été considérée comme la cinquième base de l'ADN, il en est désormais une sixième qui suscite toutes les convoitises : il s'agit de la 5-hydroxyméthylcytosine (5-hmC) (239). Les techniques mises en œuvre et présentées jusqu'ici n'ont pas réussi à la distinguer d'une 5-méthylcytosine ou à l'isoler efficacement : le traitement au bisulphite ne la convertit pas en uracile tout comme son homologue (240) et les protéines MBD ne s'y lient pas. L'anticorps utilisé de façon classique dans le MeDIP ne la reconnait pas non plus.

En revanche, Active Motif et Diagenode ont développé dernièrement des anticorps dirigés spécifiquement contre les 5-hmCs et leur utilisation dans le cadre du « hMeDIP » permettra de les étudier avec précision (20,161) et de comprendre enfin leur fonction biologique. Diagenode propose par exemple d'effectuer une double immunocapture appelée « dual MeDIP » en immunoprécipitant par le protocole de MeDIP les fractions de lavages obtenues par hMeDIP. Il est aussi possible de traiter l'échantillon au bisulphite avant l'immunoprécipitation : les 5-hmCs sont alors converties en cytosines 5-méthylène sulfonate et les fragments les contenant sont sélectionnés par un anti-sérum spécifique de cette base (241). On a également vu récemment l'utilisation de l'enzyme de restriction PvuRts1I pour fractionner et isoler les séquences contenant des 5-hmCs (242,243). Une autre stratégie, la glucosylation, consiste à coupler enzymatiquement les 5-hmCs à une molécule de glucose qui est ensuite modifiée chimiquement et couplée à une molécule portant une ou plusieurs biotines, permettant ainsi d'enrichir les fragments d'intérêt grâce à de la streptavidine (114,241).

Cette base est toutefois très rare : chez les mammifères, elle est la plus abondante dans le système nerveux central, mais ceci ne représente que 0,3 à 0,7% des cytosines, tandis que seules 0,15 à 0,17% des cytosines se trouvent sous forme 5-hmC dans les reins, la vessie, le cœur ou les poumons et 0,03 à 0,06% dans le foie et la rate (16). On n'obtiendra donc que très peu de matériel après immunoprécipitation, ce qui le rend peu exploitable.

Les technologies de séquençage vont à nouveau faire preuve de leurs performances car les prémisses de la troisième génération promettent de procurer les outils manquants pour ce type d'étude. Une 5-méthylcytosine peut en effet théoriquement être distinguée d'une 5-hmC grâce à la technologie des nanopores car les différences de potentiel que les deux bases provoquent à leur passage dans le pore sont différentes. Pacific Biosciences a par ailleurs déjà montré sa capacité à différencier ces deux cytosines avec le PacBio RS et ses cellules SMRT dans lesquelles l'incorporation de l'une ou l'autre des deux bases s'effectue avec des cinétiques différentes que l'appareil sait détecter (113). De plus, ces techniques sont applicables sur des quantités très faibles d'ADN sans nécessité d'amplification préalable. Il s'agit là d'un avenir prometteur pour l'épigénétique.

VII.3 Intérêt du MeDIP-dep-Seq

VII.3.1Apport au MeDIP-Seq

Le MeDIP immunoprécipite les régions méthylées du génome et la plupart d'entre elles se trouvent au sein des séquences répétées. Nous avons pu constater que ces séquences sont difficiles à analyser car leur alignement reste problématique : seuls 57% des séquences que nous avons obtenues par MeDIP-Seq ont pu être assignés sans ambigüité à une région de référence, les 43% restants étant alors inexploitables, ce qui engendre des pertes financières conséquentes étant donné le coût actuel du processus de séquençage.

Nous avons mis en place le MeDIP-dep-Seq qui permet d'éliminer un grand nombre de séquences répétées appartenant à diverses familles. Le protocole a été en grande partie automatisé, il n'allonge le processus complet que de deux jours et améliore considérablement la qualité des séquences : nous avons pu aligner 88% d'entre elles grâce à l'introduction de l'étape de déplétion. Parmi elles, la quantité de *reads* alignés sur l'ensemble des séquences répétées du génome de référence (les séquences uniques étant masquées) a été divisée par 2 en comparaison au MeDIP-Seq tandis que celle des *reads* alignés sur les séquences uniques (les séquences répétées étant masquées) a été multipliée par 3,5. Le protocole de déplétion ne permet ainsi pas seulement de diminuer la quantité des séquences répétées, mais aussi de libérer de la surface sur la *flow cell* pour séquencer davantage de régions uniques d'intérêt. Celles-ci occupent un très faible pourcentage du génome et le protocole de déplétion provoque donc une diminution de la couverture des CpGs de 31 à 19%. Ceci est au profit de la profondeur puisque nous avons pu observer en parcourant visuellement nos séquences que davantage de *reads* se trouvaient dans les régions d'intérêt et il serait désormais intéressant de chiffrer ce gain sur le génome entier dans nos deux cas. Cette augmentation de la profondeur permet de quantifier la méthylation de façon plus précise et donc d'identifier davantage de DMRs (160) qui pourront faire l'objet d'études plus ciblées.

Comme il a été souligné à diverses reprises, les technologies de séquençage évoluent à une vitesse toujours plus rapide et on peut imaginer que dans un futur qui n'est pas encore proche, il soit possible de séquencer le génome avec une profondeur suffisante, y compris les séquences uniques qui le composent. Notre protocole de déplétion aura toujours un intérêt dans un tel contexte car il permettra d'augmenter toujours plus la profondeur mais aussi la qualité des séquences obtenues et fournira ainsi la possibilité d'étudier un grand nombre d'échantillons à la fois en faisant usage du multiplexage.

VII.3.2 Utilisation sur de l'ADN tumoral

Il est admis depuis une dizaine d'années que la déméthylation des séquences répétées est l'une des caractéristiques des épigénomes dans les cas de cancers (244,245) et qu'elle serait la base de l'hypométhylation globale que l'on y observe. Cependant, cette caractéristique reste difficile à quantifier (246), y compris récemment en utilisant les technologies de séquençage haut-débit qui permettent de couvrir le génome entier. Une déméthylation dans les satellites a par exemple pu être observée dans un type de cancer (160) mais pas dans les éléments transposables LINEs où, bien au contraire, une hyperméthylation a pu être montrée (139). En revanche, d'autres rétrotransposons tels que les SINEs et notamment les séquences Alu ne semblent pas affectés par l'hypométhylation (51). Notre méthode trouvera donc également un intérêt dans l'étude d'échantillons d'ADNs issus de tumeurs cancéreuses.

L'utilisation des séquences répétées n'est pas envisageable pour établir des associations phénotypiques car leurs profils de méthylation sont communs à de nombreux cancers (247). Néanmoins, leur méthylation mais aussi les mécanismes de leur déméthylation restent encore à approfondir. Notre protocole de déplétion pourra alors trouver une application plus large : il est effectivement tout à fait possible d'imaginer, en complément de ce que nous avons réalisé, l'étude des éléments répétés par cette technique. Il suffirait alors de récupérer l'ADN fixé sur les billes, par chauffage en condition de dénaturation thermique puisque nos tests ont montré qu'elle menait à un décrochage total du Cot-1 biotinylé, et donc des séquences qui y sont hybridées, ou par courte incubation à plus de 70°C dans une solution ionique (213). Le séquençage des éléments répétés ainsi obtenus fournirait un outil adapté à leur étude et permettrait par exemple de compléter l'utilisation de RepArray, l'une des rares technologies actuellement disponibles pour leur analyse sur le génome entier (28). Hormis les cancers, d'autres pathologies pourront bénéficier de la mise en place d'une telle technique, comme la dystrophie facio-scapulo-humérale (FSHD), myopathie à laquelle a été associée une contraction du macrosatellite D4Z4, elle-même liée à une perte de méthylation (248).

VII.4 Une plateforme complète pour l'analyse de la méthylation à diverses échelles

Le protocole de MeDIP-dep-Seq que nous avons mis en place permet l'étude de la méthylation sur le génome entier et l'identification de régions candidates. Celles-ci peuvent alors être validées ou étudiées de façon plus détaillée. A l'heure actuelle, il n'existe pas de technique permettant de réaliser cette étude plus ciblée sur un grand nombre de loci mais aussi d'échantillons : les méthodes de capture, utilisées en combinaison avec le traitement au bisulphite de sodium, rencontrent des

problèmes d'hybridation non-spécifique, elles nécessitent de grandes quantités de matériel et il n'est pas évident de concevoir des sondes adaptées à la sélection d'ADN bisulphité. La technologie Infinium, elle, ne permet de couvrir que très peu de CpGs dont la localisation à travers le génome est, de surcroît, imposée par le fournisseur.

Nous avons donc utilisé une approche basée sur les sélectors pour isoler des fragments d'intérêt de notre choix. Son implication dans l'étude de la méthylation de l'ADN en association à la conversion par le bisulphite est novatrice. Nous avons pu sélectionner avec succès des régions génomiques impliquées dans le cancer du rein et il reste désormais à confirmer l'un des protocoles testés puis à l'appliquer à des lignées cellulaires de ce même cancer afin d'en apprécier la capacité à quantifier la méthylation pour répondre à des questions biologiques.

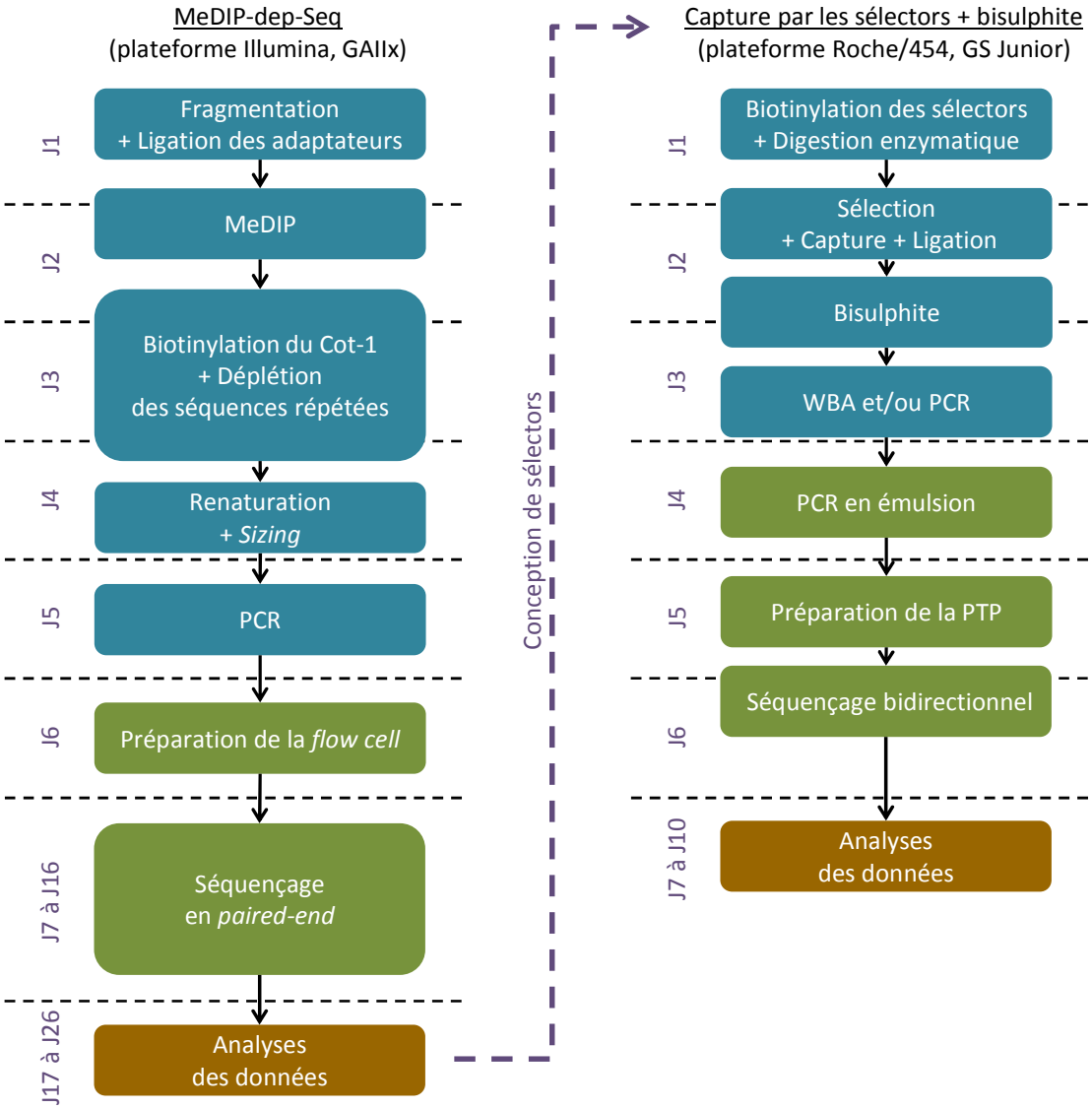


Figure 76: Vue d'ensemble de la plateforme développée pour l'analyse de la méthylation

Dans un premier temps, le MeDIP-dep-Seq et notre outil d'analyse MeQA permettront donc notamment d'identifier des DMRs que l'on pourra ensuite cibler en concevant des sélectors adaptés. En moins de deux mois (voir Figure 76), notre plateforme offre la possibilité d'obtenir une vision d'ensemble du méthylome d'un échantillon *via* le MeDIP puis d'en effectuer une étude plus locale grâce à notre nouvelle technique de capture appliquée à la méthylation, la majorité de ce temps étant occupée par le séquençage et l'analyse des données générées.

Les coûts du séquençage atteindront bientôt leur minimum et il sera un jour sans doute plus intéressant et moins cher de sélectionner *in silico* des régions d'intérêt après un séquençage sur le génome entier. En attendant cette nouvelle ère de la biologie, il faut compter sur la mise en place de grands programmes tels que l'ICGC (International Cancer Genome Consortium) ou l'IHEC (International Human Epigenome Consortium) pour établir des cartes du génome toujours plus détaillées et préciser les CpGs d'intérêt impliqués dans la maladie ou le cancer étudiés afin de concevoir des sélectors appropriés et, de façon plus générale, valider l'identification de nouveaux biomarqueurs.

Conclusion et perspectives

La méthylation de l'ADN suscite une curiosité scientifique toujours plus vive. On assiste désormais au développement de ses applications diverses et variées. La médecine légale se l'est par exemple appropriée (249,250), notamment en mettant en place des nouveaux types de tests de paternité (251). La médecine en général tire profit de ce nouvel outil dans des tests de diagnostic prénatal non-invasifs, pour ne citer que cet exemple, en tant que moyen de discrimination entre l'ADN maternel et celui du fœtus (252). Les applications médicales impliquant la méthylation ont un réel avenir pour comprendre certains des mécanismes engagés dans des pathologies encore obscures telles que les cancers. Ceci nécessite cependant d'avoir accès à des techniques adaptées et potentiellement capables d'identifier ce marqueur biologique de façon exhaustive sur le génome.

Nous avons mis en place au cours de ces travaux un protocole de MeDIP-Seq permettant d'immunoprécipiter les fragments méthylés sur le génome entier ainsi qu'un logiciel d'analyse des données résultantes. L'innovation que nous y avons apportée en comparaison aux études existantes a mené au MeDIP-dep-Seq qui améliore considérablement son rendement. Les régions d'intérêt identifiées par ce biais pourront ensuite faire l'objet d'une analyse plus ciblée grâce à la deuxième technique basée sur les sélectors que nous avons développée. La performance de ces outils réside dans l'utilisation conjointe des technologies de séquençage haut-débit. Il est ainsi possible de générer des cartes du méthylome et d'identifier des nouveaux marqueurs épigénétiques en tant que moyens de diagnostic et de pronostic des cancers et maladies complexes.

Nous entrons désormais dans une nouvelle phase de ce projet et envisageons de valoriser les techniques précédemment citées pour l'étude du cancer du sein. Grâce aux progrès médicaux en matière de diagnostic, la plupart des tumeurs du sein peuvent être prises en charge à des stades peu avancés, nommés T1 et T2 selon la classification TNM. La caractérisation par la taille de ces tumeurs ne diffère que de quelques millimètres mais la distinction entre l'une et l'autre est d'une importance capitale pour orienter le traitement de la patiente, d'où la nécessité d'identifier des signatures moléculaires précises et différentes entre T1 et T2 ; la méthylation peut répondre à ces attentes. Nous désirons dans un premier temps analyser par MeDIP-dep-Seq une soixantaine d'échantillons incluant des ADNs issus de tumeurs de classes T1 et T2 et de leurs tissus adjacents correspondants ainsi que des échantillons témoins provenant de réductions mammaires. L'analyse de ces méthylomes permettra d'identifier des gènes candidats, aptes à discriminer une tumeur T1 d'une tumeur T2, qui pourront alors être étudiés *via* les sélectors. Plusieurs cohortes (plus de 1000 cas)

seront ensuite disponibles pour la validation de ces gènes et de leur pertinence pour une utilisation clinique. Nous chercherons également à savoir si dans des cas de tumeurs identifiées comme T1 par leur taille, une rechute peut s'expliquer par un profil moléculaire T2 qui implique alors la nécessité d'un traitement par chimiothérapie. Des données d'expression pourront éventuellement appuyer les résultats obtenus et affiner le choix des biomarqueurs alors identifiés afin de mieux comprendre certaines voies de régulation impliquées dans la progression et la récurrence tumorales.

Les techniques développées par nos soins, combinées à une sélection adéquate d'échantillons cliniquement caractérisés et moléculairement phénotypés, permettront donc d'identifier de nouveaux biomarqueurs épigénétiques et ainsi de mieux prendre en charge les patientes.

Annexes

Annexe 1 : Matériel et produits utilisés

REACTIFS ET CONSOMMABLES: Du MeDIP-Seq au MeDIP-dep-Seq								
En noir: utilisé dans le protocole final.	Tarifs en € au 16/01/12							
En gris: testé, non-utilisé dans le protocole final	En vert: tarifs en $ US convertis en € au 16/01/12							
	Les tubes, plaques et cônes ne sont pas comptabilisés							
					Pour 1 échantillon			
					MeDIP-Seq		*MeDIP-dep-Seq*	
Nom du produit	*Founisseur*	*Référence*	*Unités*	*Prix public*	*Unités*	*Coût*	*Unités*	*Coût*
Matériel biologique								
Human genomic DNA	Promega	G3041	100 µg					
Mouse genomic DNA	Promega	G 3091	100 µg					
Tubes et plaques								
Plaque Abgene AB800	Dutscher	16214	25	108,40 €				
Microtubes safe-lock 1,5 mL	Eppendorf (Dutscher)	33508	1000	41,60 €				
Barrettes de 8 tubes PCR 0,2 mL	VWR	732-0545	120	158,00 €				
Fragmentation								
MicroTube Snap-Cap 6×16mm	Covaris (Kbioscience)	520045	25	98,63 €	1	3,95 €	1	3,95 €
UltraPure Agarose	Life Technologies	16500500	500 g	450,46 €	1 g	0,90 €	1 g	0,90 €
TBE 10x	Dutscher	091581	5 L	83,40 €	0,01 L	0,17 €	0,01 L	0,17 €
100 bp ladder	Life Technologies	15628-019	50 µL	93,50 €	0,16 µL	0,30 €	0,16 µL	0,30 €
6x DNA Loading Dye Buffer Blue	Solis BioDyne	07-01-00010	10 mL	30,00 €	0,005 mL	0,02 €	0,005 mL	0,02 €
Bromure d'éthidium 1%	Dutscher	091526	1 mL	20,19 €	0,001 mL	0,02 €	0,001 mL	0,02 €
Ligation des adaptateurs								
Paired-end sample prep kit	Illumina	PE-102-1002	40					
PE Oligo Only kit (contient amorces 1.0 et 2.0)	Illumina	PE-102-1003	100	4 349,38 €	1	43,49 €	1	43,49 €
Next DNA sample prep master mix Set 1	NEB (Ozyme)	E6040L	50	1 107,84 €	1	22,16 €	1	22,16 €
MeDIP								
Auto-MeDIP kit	Diagenode	AF-01-0016	16	370,00 €	1	23,13 €	1	23,13 €
200 µL tube strips (12 tubes/strip) + cap strips	Diagenode	WA-001-0080	80	300,00 €	1	3,75 €	1	3,75 €
Cônes à filtres: tips (bulk)	Diagenode	WC-001-1000	1000	300,00 €	6	1,80 €	6	1,80 €
Contrôle-qualité par pyroséquençage								
EpiTect Bisulfite Kit	Qiagen	59104	48					
MethylEasy Xceed	Human Genetic Signatures	ME002	40					
Epitect methylated human control DNA	Qiagen	59655	100					
HotStar Taq DNA Polymerase	Qiagen	203205	1000 U					
Streptavidin Sepharose high performance	GE Healthcare	17-5113-01	5 mL					
PyroMark Q96 HS Plate	Qiagen	979101	100					
PyroMark Gold Q96 SQA Reagents	Qiagen	972812	96					
PyroMark Q96 HS Reagent Tip	Qiagen	979102	4					
PyroMark Q96 HS Nucleotide Tip	Qiagen	979103	8					
Quantification et contrôle qualité								
Quant-iT dsDNA Assay kit, Broad Range	Life Technologies	Q33130	1000					
Quant-iT dsDNA HS Assay kit	Life Technologies	Q32851	100	71,75 €	1	0,72 €	1	0,72 €
Quant-iT ss-DNA HS Assay kit	Life Technologies	Q10212	100					
Qubit assay tubes	Life Technologies	Q32856	500	57,50 €	1	0,12 €	1	0,12 €
Power SYBR Green PCR Master Mix	Applied Biosystems	4367659	200	385,22 €	4	7,70 €	4	7,70 €
5x HOT FIREPol EvaGreen qPCR Mix (no ROX)	Solis BioDyne	08-23-00001	250	37,00 €			8	1,18 €
DNA 1000 kit	Agilent	5067-1504	25	579,00 €	1/12	1,93 €	1/12	1,93 €
High Sensitivity DNA kit	Agilent	5067-4626	10	395,00 €				
Rotor-Gene Probe PCR kit	Qiagen	204374	400					
Purifications								
Auto IPure kit	Diagenode	AL-Auto01-0100	100					
QIAquick PCR purification kit	Qiagen	28106	250	450,00 €	2	3,60 €	4	7,20 €
MinElute PCR purification kit	Qiagen	28006	250	486,00 €	2	3,89 €	2	3,89 €
Biotinylation d'ADN Cot-1								
Mouse Cot-1 DNA	Life Technologies	18440-016	500 µg	329,00 €			0,6 µg	0,39 €
Human Cot-1 DNA	Life Technologies	15279-011	500 µg					
Human Hybloc DNA	Applied Genetics Laboratories	HHB	500 µg					
Terminal transferase recombinant (400 U/µL)	Roche	03 333 574 001	24000 U					
Biotin-16-ddUTP (1 nmol/µL)	Roche	11 427 598 910	25 nmol					
Sephadex G50 Superfine	GE Healthcare	17-0041-01	100 g					
Bio-gel P100 gel	Bio-Rad	#150-4174	100 g					
Plaques MultiScreen-HA	Millipore	MSHAS4510	10					
Plaques MicroAmp	Applied Biosystems	N801-0560	10					
Streptavidin Cy3	GE Healthcare	PA43001	1 mg					
Microcon YM-100	Millipore	42413	100					
BioPrime DNA labelling system	Life Technologies	18094-011	30	308,00 €			3	30,80 €

					Pour 1 échantillon			
					MeDIP-Seq		MeDIP-dep-Seq	
Nom du produit	*Founisseur*	*Référence*	*Unités*	*Prix public*	*Unités*	*Coût*	*Unités*	*Coût*
Fixation du Cot-1 biotinylé sur les billes								
Dynabeads MyOne Streptavidin C1	Life Technologies	650-02	10 mL	1 422,75 €			0,2 mL	28,46 €
Dynabeads Streptavidin trial kit	Life Technologies	658-01D	4×1 mL					
Hybridation IP/Cot-1 biotinylé								
GenomePlex Complete WGA kit	Sigma-Aldrich	WGA2-50RXN	50					
BSA: Albumin from bovine serum	Sigma-Aldrich	A7888	100 g					
Carrier RNA (du Epitect Bisulphite kit)	Qiagen	1017794	1350 µg					
Préparation des librairies								
Low range ultra agarose	Bio-Rad	161-3107	125 g					
TAE buffer 5x	Bio-Rad	161-0773	5 L					
SYBR Gold Nucleic Acid Gel Stain	Life Technologies	S-11494	500 µL					
QIAquick gel extraction kit	Qiagen	28706	250					
E-gel size select 2% agarose	Life Technologies	G6610-02	10	138,00 €	1/8	1,73 €	1/8	1,73 €
Platinum Pfx DNA polymerase	Life Technologies	11708-013	100	176,00 €	1	1,76 €		
Phusion High Fidelity DNA polymerase 5x	NEB (Ozyme)	F530S	100	97,00 €			1	0,97 €
Agencourt AMPure XP beads	Beckman Coulter Genomics	A63881	60 mL					
Séquençage								
TruSeq PE Cluster Kit v2 - cBot - GA								
+ Paired-end flow cell v4 + GA cBot manifold	Illumina	PE-300-2001	8	4 429,89 €	1	553,74 €	1	553,74 €
TruSeq SBS kit v5 - GA (20×36 cycles)	Illumina	FC-104-5020	20	22 153,44 €	6/8	830,75 €	6/8	830,75 €
				TOTAL		**1 505,60 €**		**1 569,25 €**

REACTIFS ET CONSOMMABLES: Analyse multiplexée de loci par la technologie des sélectors			
Nom du produit	*Founisseur*	*Référence*	*Unités*
Matériel biologique			
Human genomic DNA: male	Promega	G1471	100 µg
Biotinylation des sélectors			
Terminal Transferase (20 U/µL)	NEB	M0315S	
Biotin-16-dUTP (1 mM)	Roche	11 093 070 910	50 nmol
illustra MicroSpin G-50 Columns	GE Healthcare	27-5330-02	250
Digestion enzymatique			
BSA (10 mg/mL)	NEB	B9001S	
MseI (10 U/µL)	NEB	R0525S	
EcoO109I (20 U/µL)	NEB	R0503S	
NlaIII (10 U/µL)	NEB	R0125S	
HpyCH4V (5 U/µL)	NEB	R0620L	
MscI (15 U/µL)	NEB	R0534M	
DdeI (10 U/µL)	NEB	R0175S	
NEB Buffer 4	NEB	B7004S	
Hybridation aux sélectors et capture sur les billes			
Dynabeads M-280 Streptavidin	Life Technologies	112-06D	10 mL
Ligation			
Ampligase thermostable DNA ligase	Epicentre Biotechnology	A8101	1000 U
Traitement par le bisulphite			
MagBisulfite kit	Diagenode	AF-106-0024	24
EpiTect Bisulfite kit	Qiagen	59104	48
Lambda DNA	Fermentas	SD0011	500 µg
Amplifications			
phi29 DNA polymerase	Thermo Scientific	EP0094	200 U
Platinum Taq DNA polymerase High Fidelity	Life Technologies	11304-029	
EpiTect Whole Bisulfitome	Qiagen	59205	100
Contrôles par qPCR			
QuantiFast SYBR Green PCR kit	Qiagen	204054	400
Séquençage			
GS Junior Titanium emPCR kit (Lib-A)	Roche	05 996 520 001	
GS Junior Titanium Picotiterplate kit	Roche	05 996 619 001	
GS Junior sequencing kit	Roche	05 996 554 001	

PRODUITS CHIMIQUES			
Nom du produit	*Founisseur*	*Référence*	*Unités*
Ethanol absolu	VWR	20821.321	2,5 L
Tris-HCl	Acros	22-8032500	250 g
EDTA	Sigma-Aldrich	E-5134	250 g
NaCl	Fluka	71376	1 kg
Tween-20	Prolabo	28829.296	1 L
SSC (Saline-sodium citrate) 20x	Sigma-Aldrich	S-6639	1 L
Hydroxyde de sodium	Sigma-Aldrich	S-8045	1 kg

APPAREILS ET SYSTEMES	
Nom du produit	*Founisseur*
Fragmentation	
Bioruptor UCD-200	Diagenode
Sonicateur E210	Covaris
TR-246 Snap-Cap	Covaris
Générateur EPS 1000	Pharmacia Biotech
Ligation des adaptateurs	
Agitateur Thermomixer Comfort	Eppendorf (Dutscher)
Adaptateur 96 microtubes PCR 0.2 mL	Eppendorf (Dutscher)
MeDIP	
Robot: SX-8G IP Star	Diagenode
DiaMag02 - Magnetic rack	Diagenode
Contrôle-qualité par pyroséquençage	
Master cycler gradient	Eppendorf (Dutscher)
Wide Mini-Sub Cell GT Cell	Bio-Rad
PyroMark Q96 vacuum prep workstation	Qiagen
Thermoplate	Grant
PyroMark Q96 HS Sample Prep Thermoplate	Qiagen
PSQ 96MD System	Qiagen
PyroMark Q96 HS Dispensing Tip Holder	Qiagen
Quantification et contrôle-qualité	
Spectra Max Gemini XPS	Molecular Devices
NanoDrop ND-1000	Thermo Scientific
Qubit	Life Technologies
Bioanalyzer 2100	Agilent
Rotor-Gene Q	Qiagen
Purifications	
Heraeus Fresco 21 centrifuge	Thermo Scientific
Centrifuge 5810R	Eppendorf (Dutscher)
Préparation des librairies	
Unité d'électrophorèse H1-SET	Proteigene
Dark Reader - Blue Light Transilluminator	Clare Chemical Research
E-Gel iBase Power System	Life Technologies
DynaMag-2 Magnet	Life Technologies
Master cycler ep gradient S	Eppendorf (Dutscher)
Séquençage	
cBot	Illumina
Genome Analyzer IIx	Illumina
Paired-end module	Illumina
GS Junior emPCR bead counter	Roche
GS Junior bead deposition device	Roche
IKA Turrax	Roche
GS Junior	Roche
Déplétion	
DynaMag-96 Side Skirted	Life Technologies

Annexe 2 : Amorces de PCR utilisées dans le MeDIP-dep-Seq

Les amorces ont été synthétisées par Biotez (Buch, Allemagne)

Contrôle-qualité du MeDIP par pyroséquençage

Amorces spécifiques d'ADN bisulphité humain

Région/gène	Taille amplicon (pb)	Coordonnées chromosomiques (hg18)	Amorce sens Amorce anti-sens	Température d'hybridation	Amorce de pyroséquençage
C6ORF106	218	chr6:34,772,294-34,772,511	Biotine-GTGGTGTTTAGGTAGTTGAATTTTTG ACCCCTATAAAACTATAAAAAACCC	61°C	CRCRAAACTAACTAACAAC
DLL1	233	chr6:170,442,839-170,443,071	Biotine-TGGTTTTTAATTTTTTTTATAGATA CTATCTACATTACCATACAACTAAAC	58°C	AACTACCTTAATAACAACCA
FAM50B	179	chr6:3,794,550-3,794,728	AGGAGTTTGGTATTTTTTTTAGGGTT Biotine-AACCCTTACCTACCCAAACACCCTATC	68°C	GTTTTTATTTTTTGTGTTTA
	160	chr6:3,794,704-3,794,863	Biotine-TAGGGTGTTTGGGTAGGTAAGGGTT AACACCAATTTCCAACATTTTATTTAACCT	67°C	TTTAACCTCAATAAATACAA
IGF2	173	chr11:2,126,043-2,126,215	TTTTTGGGAATGTTTATTTATGTATGA Biotine-ACAAAAACCACTAAACACACAACTCTA	58°C	TTTTTAGGAAGTATAGTTA
OSTM1	218	chr6:108,547,455-108,547,672	AGTATTATTAGTATTGGGGAGATTT Biotine-TCCTACCTATAACACCAAAAAAACC	61°C	TGTTATAGGGGTTGAGTTTA
RASSF1A	191	chr3:50,353,212-50,353,402	Biotine-AGTTTTTGTATTTAGGTTTTTATTG AACTCAATAAACTCAAACTCCCC	60°C	CCCCAAAATCCAAACT

Région/gène	Sonde Taqman	Amorce sens Amorce anti-sens
Alu	FAM-CCTACCTTAACCTCCC-MGBNFQ	GGTTAGGTATAGTGGTTTATATTTGTAATTTTAGTA ATTAACTAAACTAATCTTAAACTCCTAACCTCA

Déplétion des séquences répétées

Amorces spécifiques d'ADN génomique de souris

Région/gène	Taille amplicon (pb)	Coordonnées chromosomiques (mm9)	Amorce sens Amorce anti-sens	Température d'hybridation
Séquences ciblant des régions répétées du génome				
Line L1 (ORF2)	155	/	TTTGGGACACAATGAAAGCA CTGCCGTCTACTCCTCTTGG	68°C
maSat	308, 542, 776	/	GACGACTTGAAAAATGACGAAATC CATATTCCAGGTCCTTCAGTGTGC	52°C
miSat	162, 285, 408	/	CATGGAAAATGATAAAAACC CATCTAATATGTTCTACAGTGTGG	52°C
Sine B1 (Alu)	113	/	GTGGCGCACGCCTTTAATC GACAGGGTTTCTCTGTGTAG	52°C
Séquence contrôle représentant une séquence unique d'intérêt				
Igf2	216	chr7:149,855,219-149,855,434	GGGAGAGTTAGGGTCCTTCAGTT TACTCCTGCTGGGAAGAGAAAG	67°C

Amorces spécifiques d'ADN génomique humain

Région/gène	Taille amplicon (pb)	Coordonnées chromosomiques (hg19)	Amorce sens Amorce anti-sens	Référence
Séquences ciblant des microsatellites (STRs, Short Tandem Repeats)				
TH01	162, 170	chr11:2,192,220-2,192,381	GTGGGCTGAAAAGCTCCCGATTAT GTGATTCCCATTGGCCTGTTCCTC	(220)
CD4	86, 111	chr12:6,897,505-6,897,590	TTGGAGTCGCAAGCTGAACTAGC GCCTGAGTGACAGAGTGAGAACC	(220)
VWA	150, 154	chr12:6,093,104-6,093,253	CCCTAGTGGATGATAAGAATAATCAGTATG GGACAGATGATAAATACATAGGATGGATGG	(221)
PLA2A1	131, 134	chr12:120,764,330-120,764,460	CCCACTAGGTTGTAAGCTCCATGA TACTATGTGCCAGGCTCTGTCCTA	(220,221)
TFIIDA	201	chr6:170,870,947-170,871,147	GCCTATTCAGAACACCAATA TGGGACGTTGACTGCTGAAC	(220,221)
CYAR04	166, 173	chr15:51,519,844-51,520,009	CTCTGGAAAACAACTCGACCCTTC TGGGTGATAGAGTCAGAGCCTGTC	(220,221)
GABARB1	140, 148	chr4:47,427,466-47,427,605	CTAGAAAGCTAGCAAGGTGGAT GCTCATTAAACACTGTGTTCCT	(220,221)
Séquences ciblant des éléments *Alu*				
HBB	158	Non disponible	AGAGATCGCGCCACTGCACA CACAGCCTTTCTTGGTTTTC	(222)
F8	197	Non disponible	GTGATTGTTCCACTGCACTG GTGCCTTGGTGAAAAATAAAGC	(222)
Séquences ciblant des α-satellites				
Kpn I (1)	76	/	ATTCCCACCTATGAGTG GGAAATCATCATTCTCAGT	(223)
Kpn I (2)	55	/	TGCCGCAATAAACATAC GGACTATAAATCATGCTGC	(223)
Kpn I (3)	75	/	TGGCTGCATAAATGTCT AAACAAACAACCCCATC	(223)

Annexe 3 : Régions étudiées par pyroséquençage sur les MEFs

Gène	Coordonnées chromosomiques (mm9)
Myl1	chr1:66,991,474-66,991,755
Ptx3	chr3:66,023,652-66,023,906
Hmgb1	chr5:149,865,178-149,865,439
Cdkn1b	chr6:134,870,204-134,870,419
Ccnd1	chr7:152,126,805-152,126,975
Igf1	chr10:87,320,669-87,320,964
Pttg1	chr11:43,239,469-43,239,762
Myh	chr11:66,983,539-66,983,833
Per1	chr11:68,911,807-68,912,098
Tnfrsf11b	chr15:54,110,559-54,110,864
Has2	chr15:56,526,076-56,526,382
Pvalb	chr15:78,037,315-78,037,671

Tous les oligonucléotides ont été synthétisés par IDT (Integrated DNA Technologies)

Vecteur	Séquence du vecteur
Vecteur (tous les C sont méthylés)	TCGACCGTTAGCAAAGCTTTCTACCGTTATCGT

Nom du sélector	Séquence du sélector (en rouge, séquence complémentaire du vecteur)	Région/ gène ciblé	Coordonnées chromosomiques de la cible (hg18)	Brin ciblé	Longueur de la cible (pb)
NC_000001.9_TARGET_1|MscI/DdeI|-|1	CCCCCTGGGTCGGACGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGTGACCTTGTGCAAGTCACTTTTCC	*HDAC1*	chr1:32529918-32530296	-	378
NC_000001.9_TARGET_4|MscI/DdeI|-|1	GCCAGGTACTCACCTGTATGGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGATCGGTTAAGTAACTGGCTCATAA	*PTGS2*	chr1:184915774-184916006	-	232
NC_000001.9_TARGET_4|MscI/DdeI|+|4	GGAGCACGTCCAGGAACTCCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGATTCCTGGAGAGGAAGCCAAGTGT	*PTGS2*	chr1:184916128-184916591	+	463
NC_000001.9_TARGET_4|MseI/EcoO109I|+|2	CCACCGGGCTTACGCAATTTTTTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAACCTTACTCGCCCCAGTCTGTCC	*PTGS2*	chr1:184916297-184916713	+	416
NC_000003.10_TARGET_2|MscI/DdeI|+|4	CACACGCCTGGCACGGCTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGTGTCTATCAGGCGTCTCAAGGCTG	*MST1R*	chr3:49915197-49915401	+	204
NC_000003.10_TARGET_2|NlaIII/HpyCH4V|+|3	TGTCTCCTACAGTATTGAATACGTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATCTGGCAGCGCCAGCCTG	*MST1R*	chr3:49915297-49915581	+	284
NC_000003.10_TARGET_3|MscI/DdeI|+|2	CTCCTCCCGCCGCTGCCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGACCCGAGGGCCGGGAAGGG	*MST1R*	chr3:49916021-49916232	+	211
NC_000003.10_TARGET_3|MseI/EcoO109I|-|2	CCATCGAGGCGAGCTGGGACACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCGTGGCCAGGCTCTGGACA	*MST1R*	chr3:49915767-49916062	-	295
NC_000003.10_TARGET_3|MseI/EcoO109I|+|5	GCTGTCGGGACCGCTGCTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTGGGCTAGGCCAAGCCTTCC	*MST1R*	chr3:49916148-49916405	+	257
NC_000003.10_TARGET_4|MscI/DdeI|+|3	CCATTGCGCGGCTCTCCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGGGCCGACCACCCATGCCAA	*RASSF1*	chr3:50353364-50353828	+	464
NC_000003.10_TARGET_4|MseI/EcoO109I|-|2	CCACGTGTGCGTGGCGGGCCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAAATCAGCGTATTTTTACATATAAG	*RASSF1*	chr3:50352752-50353068	-	316
NC_000003.10_TARGET_5|MscI/DdeI|-|3	CCCACTGGGTGGCCGCCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGACCCTCGTGGGGCGGAAGA	*FHIT*	chr3:61211857-61212109	-	252
NC_000003.10_TARGET_5|MseI/EcoO109I|-|1	GGTGCTTGGGAATTGGGGGGCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAACACAGACCTGTTGGGACGGATTTT	*FHIT*	chr3:61211630-61211839	-	209
NC_000003.10_TARGET_6|MscI/DdeI|-|2	TCGGAGAGGTTAAGCGACTTCCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGGCTTCAGTCTTTCAGCCTGCAG	*ST6GAL1*	chr3:188130424-188130638	-	214
NC_000003.10_TARGET_6|MseI/EcoO109I|-|7	TGGAAGAGGGCGAGAGAGACATTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCGCGGCGCGCGTAGGGCGC	*ST6GAL1*	chr3:188131414-188131874	-	460
NC_000003.10_TARGET_6|NlaIII/HpyCH4V|-|4	GACCCAGTCTTCGCAACTCTATTTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGTTAGGACCAAAAGCCGGACAC	*ST6GAL1*	chr3:188130685-188130969	-	284
NC_000003.10_TARGET_6|NlaIII/HpyCH4V|-|5	AGCCTCCGCGGCCGAGAGTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATACGGAGGCGACTCGCTTCC	*ST6GAL1*	chr3:188130970-188131312	-	342
NC_000004.10_TARGET_1|MseI/EcoO109I|-|4	TCCCCGTTTCCGGTCAATATATTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTGCGCGCGCGCTCCTGCTG	*IGFBP7*	chr4:57671434-57671644	-	210
NC_000004.10_TARGET_1|NlaIII/HpyCH4V|-|5	CTTTCTGGAACCTCTTAACTGCCTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGGGCGCTCGGTCTCACG	*IGFBP7*	chr4:57671491-57671891	-	400
NC_000004.10_TARGET_1|NlaIII/HpyCH4V|+|1	GATTTTCCGAGACCTCAGACTTCCCACGATAACGGTAGAAAGCTTTGCTAACGGTCGACAAGAGCCGCTACCCGGTGT	*IGFBP7*	chr4:57670508-57670936	+	428
NC_000004.10_TARGET_2|MscI/DdeI|-|1	CGTCGGGACCAGACTCGTCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGTGCTCTTCCTGGATGCTGAATGC	*SPP1*	chr4:89115628-89115939	-	311
NC_000004.10_TARGET_3|MscI/DdeI|-|1	CCCACTCCCTGCTGGGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAGCCCGACTTCTCCTACAAGC	*SFRP2*	chr4:154928899-154929352	-	453
NC_000004.10_TARGET_3|MseI/EcoO109I|-|6	CTCGCCCGCCCTAGGATTTCTTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCGAAAAAGCTGGCAGCCGGC	*SFRP2*	chr4:154929551-154929897	-	346
NC_000005.8_TARGET_1|MseI/EcoO109I|-|2	GAACAACTGTATGCCCAAGGGCTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAACTAGAAACTGACTTGTTGCAACA	*IL12B*	chr5:158690131-158690483	-	352
NC_000005.8_TARGET_2|MscI/DdeI|+|2	GGTGTACACCTTTGCATACCTCTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGACGGCGAGGAAAGTTAGCCCG	*IL12B*	chr5:158690971-158691330	+	359
NC_000006.10_TARGET_2|MscI/DdeI|-|3	CCCCCTGGCCGTGAAACTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGATGCCCCCTGTCAGTTCCCC	*ESR1*	chr6:152169968-152170309	-	341
NC_000006.10_TARGET_2|NlaIII/HpyCH4V|-|11	CCGAGCCCGCTGATGCTACTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATCAGATCCAAGGGAACGAGCTGG	*ESR1*	chr6:152170787-152171074	-	287
NC_000006.10_TARGET_2|NlaIII/HpyCH4V|-|4	CCGTCCTCCAGCACCTTTGTAATGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACAACCGCACACCCCATTCTATCTG	*ESR1*	chr6:152170028-152170459	-	431
NC_000006.10_TARGET_3|MscI/DdeI|-|3	CGCATGATCTTGGTGATCTTGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCACGTACATCCGGCTCTTTTCCG	*THBS2*	chr6:169390470-169390835	-	365
NC_000006.10_TARGET_3|MseI/EcoO109I|+|2	CGCTGTTGGCTCTGGAGGGCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAATTTCCTTACACGCGAGGTTCTGGT	*THBS2*	chr6:169390732-169391092	+	360
NC_000006.10_TARGET_4|MscI/DdeI|-|1	GACTCACTTTTCCTGTCTTCTCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGAGTTCGAGAACTCTAATTTCTTAA	*THBS2*	chr6:169395672-169396149	-	477
NC_000006.10_TARGET_5|MscI/DdeI|+|2	CCCGCCCCTCCCACACCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGTGAGGATCCCCTAGAAGGACGTCT	*THBS2*	chr6:169396506-169396717	+	211
NC_000007.12_TARGET_1|NlaIII/HpyCH4V|+|2	TACCCCTGGATGCGCAAGCTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGGATCTCAGCGTCGGCCG	*HOXA5*	chr7:27149203-27149599	+	396
NC_000007.12_TARGET_10|MscI/DdeI|-|12	CCCAGCGTCATCTTGATATGACTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAGCTCTGCACGGCTGCGG	*MEST*	chr7:129919314-129919709	-	395
NC_000007.12_TARGET_10|MscI/DdeI|+|17	TGATACCTGGGTCTTGAAATCCTAAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGCCTTGATTTCTTCATCTTGTAAAC	*MEST*	chr7:129920131-129920411	+	280
NC_000007.12_TARGET_10|MseI/EcoO109I|+|3	TAAGATAGAATCACAATGCAAGGGCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAACTACCGGTAACCACCAAGGGCA	*MEST*	chr7:129918297-129918628	+	331
NC_000007.12_TARGET_10|MseI/EcoO109I|+|9	CTAGAAATCCTAACCCAAGGAGGTCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAGTCGCACCAGAGCGATGCTG	*MEST*	chr7:129919621-129920006	+	385
NC_000007.12_TARGET_10|NlaIII/HpyCH4V|-|13	TTAACTCATCAGGGGAGGGTTTCTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGAATCTGTTTACTAGCTTGGTG	*MEST*	chr7:129918540-129918901	-	361
NC_000007.12_TARGET_10|NlaIII/HpyCH4V|-|21	GTGCCGAGATCGCCGCGGTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGATTTAGGATTTCTAACCCCCA	*MEST*	chr7:129919814-129920126	-	312
NC_000007.12_TARGET_2|MscI/DdeI|-|2	ACGTGGAAATCTATCCCCATCCTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAGCACGTCGCCAGGCACG	*HOXA11*	chr7:27190967-27191421	-	454
NC_000007.12_TARGET_2|MscI/DdeI|+|1	GTCGTCTTCCGGCCACACTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAAGAGCTCGGCCAACGTCTA	*HOXA11*	chr7:27190600-27190966	+	366
NC_000007.12_TARGET_3|MscI/DdeI|-|3	ACGAGCAGCGCCAGCTCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAGGATGGCGTTCTCCTGCG	*SFRP4*	chr7:37922496-37922880	-	384
NC_000007.12_TARGET_3|MscI/DdeI|+|2	GAACTGGGCGGCCACACCTAAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCATCGAGCAGTACGAGGAGCTG	*SFRP4*	chr7:37922137-37922495	+	358
NC_000007.12_TARGET_3|MscI/DdeI|+|4	GTTCGCGCCCCGAAGGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAGACTAAAAAGGAGGGACTTTAG	*SFRP4*	chr7:37922878-37923157	+	279
NC_000007.12_TARGET_4|MscI/DdeI|-|5	ACGCTGCGCCCGGGGACTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCACCGCTCCACCATCACTTGC	*IGFBP1*	chr7:45894682-45895023	-	341

NC_000007.12_TARGET_4|MseI/EcoO109I|+|6	CGCCACTTGCACCAGGAGGTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTCTGACATCTCCAGGCGCGAG	*IGFBP1*	chr7:45894548-45894785	+	237
NC_000007.12_TARGET_4|NlaIII/HpyCH4V|+|2	AGAAAAAAGCCCTAGAGATCTCCTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACACCCAAGGCCCTACGCAAAC	*IGFBP1*	chr7:45894267-45894487	+	220
NC_000007.12_TARGET_5|MscI/DdeI|+|1	CTGCCTGACGTGCGCACTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGGATCAGCCGCTTCCTGCCT	*IGFBP3*	chr7:45927039-45927368	+	329
NC_000007.12_TARGET_6|MscI/DdeI|-|1	GTTGCGCCGGCTCTGGCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGAGTGTGGGTACGGGGACTTCC	*FZD9*	chr7:72485811-72486055	-	244
NC_000007.12_TARGET_7|MscI/DdeI|-|2	TCATTAGCCAAATGCATGAGCCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGCAGGGATATTGATTCCAAAGGCTA	*MDR1*	chr7:87067221-87067485	-	264
NC_000007.12_TARGET_7|MseI/EcoO109I|-|4	GCGGGCTCAGAGAGCAGGTCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAAAAAGTTCTTCTTCTTTGCTCCTC	*MDR1*	chr7:87067387-87067763	-	376
NC_000007.12_TARGET_8|MseI/EcoO109I|-|1	CCTCAAAAAGAATGGAACCAATTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCGCCCTTTCCCGAAGTCATAAGA	*COL1A2*	chr7:93861492-93861951	-	459
NC_000007.12_TARGET_8|MseI/EcoO109I|+|4	CCTAGGGTGCCTCCAAAAGGGCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAAGTAATTTCCATGTTTCTTATCAA	*COL1A2*	chr7:93862069-93862516	+	447
NC_000007.12_TARGET_8|NlaIII/HpyCH4V|-|7	GTTGGAGGTACTGGCCACGACTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACAGACAACGAGTCAGAGTTTCCCCT	*COL1A2*	chr7:93861878-93862195	-	317
NC_000007.12_TARGET_9|MscI/DdeI|-|1	GGCCGAGAACCTCTGGCCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGATCAGGCCGAATACAATTTTATTC	*MEST*	chr7:129913161-129913407	-	246
NC_000008.9_TARGET_1|MseI/EcoO109I|+|4	CGCCTTTTGTCCCCGGAGGTCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAGGGGTGTTGAGCCGCGTCT	*SFRP1*	chr8:41285917-41286310	+	393
NC_000008.9_TARGET_1|NlaIII/HpyCH4V|+|9	CCAACCTGCTGGAGCACGAGACACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGGGCATCGGGCGCAGCG	*SFRP1*	chr8:41285579-41285836	+	257
NC_000008.9_TARGET_2|MseI/EcoO109I|-|2	CCGCCCTTCGCTCCAAGGCCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAGGAAATACAAGAAAAACCTGACCA	*RAD54B*	chr8:95555946-95556212	-	266
NC_000009.10_TARGET_1|MscI/DdeI|-|2	AGGTGCTATTAACTCCGAGCACTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAGCCAGTCAGCCGAAGGC	*CDKN2A*	chr9:21964778-21965061	-	283
NC_000009.10_TARGET_1|MscI/DdeI|+|1	CGCGTACAGATCTCTCGAATGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCACGGCCGCGGCCCGGGGT	*CDKN2A*	chr9:21964442-21964777	+	335
NC_000009.10_TARGET_1|MscI/DdeI|+|3	CTTCAGGGGTGCCACATTCGCTAAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGTGAACCCCGCGCTCCTGAA	*CDKN2A*	chr9:21965059-21965506	+	447
NC_000009.10_TARGET_2|MscI/DdeI|-|3	CGCGCACCCGCCTTCCCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGTAGCATCAGCACGAGGGCCA	*CDKN2A*	chr9:21984183-21984487	-	304
NC_000009.10_TARGET_2|NlaIII/HpyCH4V|+|1	CCCATTCGGGTTACAACGACTTAGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGGTGCGCAGGTTCTTGGTG	*CDKN2A*	chr9:21983985-21984331	+	346
NC_000011.8_TARGET_1|NlaIII/HpyCH4V|-|2	TGCTCGGCCAGCAGGCGCTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACACGGGCTTGAAAAACTTTGGGGG	*ASCL2*	chr11:2247368-2247832	-	464
NC_000011.8_TARGET_2|MscI/DdeI|-|3	GGGCTTAAGGAGTGGTCGAGACTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAATCGCGCAGGGCGCGGC	*ASCL2*	chr11:2248944-2249209	-	265
NC_000011.8_TARGET_2|MseI/EcoO109I|-|1	CCCAAAGCAAGGTACGCAGGTCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAATGTATAGATAACCCTCCTCCGCG	*ASCL2*	chr11:2248424-2248689	-	265
NC_000011.8_TARGET_2|MseI/EcoO109I|-|7	ATGCGGGGAAATGGTGGGGCACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTGGAGGTCTGCACCCGACC	*ASCL2*	chr11:2249348-2249666	-	318
NC_000011.8_TARGET_2|MseI/EcoO109I|+|2	CGGGAAGGCTCAACCCAGGACACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAGCAGGAACCCAGCTTTGTTAG	*ASCL2*	chr11:2248687-2249048	+	361
NC_000012.10_TARGET_1|MscI/DdeI|+|6	CCTTCTTGGTTCCCAGGGCTCTAAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGGGACTCTGGGAATCAAGGGGTG	*WNT10B*	chr12:47650394-47650784	+	390
NC_000012.10_TARGET_1|NlaIII/HpyCH4V|-|2	CAAGGGGAACTGCTCACCGACTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACAGACCCTGAAGCGCGGACG	*WNT10B*	chr12:47650248-47650511	-	263
NC_000012.10_TARGET_1|NlaIII/HpyCH4V|+|1	CCTCCGCTCAGCTTAATCTGGGTTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACACATCGCGGTCCACGAGTGT	*WNT10B*	chr12:47649861-47650247	+	386
NC_000012.10_TARGET_2|NlaIII/HpyCH4V|-|2	CTGCCGGGCTGACGGAGCTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACACGTCCCTGCCCTCCCTGC	*WNT10B*	chr12:47651467-47651803	-	336
NC_000012.10_TARGET_3|MscI/DdeI|-|3	ACAGGCAAAGGCAAACCGCCTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGAGTTTGGCCCTAGCAGAGGTTCAA	*WNT10B*	chr12:47652367-47652639	-	272
NC_000012.10_TARGET_3|NlaIII/HpyCH4V|+|2	GGCGGTTTGCCTTTGCCTGTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACAACGCCACCACCAAGGAAGAGG	*WNT10B*	chr12:47652616-47652900	+	284
NC_000012.10_TARGET_4|MscI/DdeI|-|2	GTGGGCTCCTGGGAGCTCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGATAGCGCCACTGCGCTCCT	*IRAK3*	chr12:64868822-64869026	-	204
NC_000014.7_TARGET_1|MscI/DdeI|+|5	CGGTTTTTGATCTTTCTTCTGCTTACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGCAGCGGAGATCCCAAAAGCC	*MMP14*	chr14:22375652-22376086	+	434
NC_000014.7_TARGET_1|NlaIII/HpyCH4V|+|5	ACCTCTAAGTTGCCTTTTTGTGGTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGGTCCGAGACCACCGGG	*MMP14*	chr14:22375456-22375869	+	413
NC_000014.7_TARGET_1|NlaIII/HpyCH4V|+|7	AAAAAGAAAGGGAGGGGGAATCCTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGACGAGGAGCCGGGACAG	*MMP14*	chr14:22376096-22376363	+	267
NC_000014.7_TARGET_2|MseI/EcoO109I|-|2	AACTGAAAAGATCACAAGCGACTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAACCACACCTCGAGCGAAGGG	*ESR2*	chr14:63830475-63830848	-	373
NC_000014.7_TARGET_2|NlaIII/HpyCH4V|-|6	GTCTGGAGTAGGGCCTGAGAATATGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACAACTGCCTCCTGGCGGGGG	*ESR2*	chr14:63830773-63831091	-	318
NC_000014.7_TARGET_3|MscI/DdeI|-|3	CACCGCGTCCCCAGAGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGGAGATTCTAAAAGCGCGGTGCTG	*ESR2*	chr14:63874810-63875046	-	236
NC_000014.7_TARGET_3|MseI/EcoO109I|+|3	CCAGGCAGTAATGGGCGGGTCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAGGGAAAAGGAAGTAAGGGCGGG	*ESR2*	chr14:63874982-63875445	+	463
NC_000014.7_TARGET_3|NlaIII/HpyCH4V|-|3	CCAGGTTGCGGTGGAACGTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACAAAGGTGCTTTCTGGCCCGG	*ESR2*	chr14:63875277-63875540	-	263
NC_000016.8_TARGET_1|MseI/EcoO109I|-|5	CAGAATTTGATCCCAACACTGATTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCGGGAGGGGGCCGCAGAGG	*PYCARD*	chr16:31121775-31122210	-	435
NC_000016.8_TARGET_2|MseI/EcoO109I|-|4	CTCTAACCTCAGGACGTCAAGGGCACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTTCCACTGTCTCTGTTTCCATCTC	*MMP2*	chr16:54070083-54070316	-	233
NC_000016.8_TARGET_2|MseI/EcoO109I|-|5	GGCTGCGCATCTGGGGCTTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTAGAGCGACAGATGTTTCCCAGCA	*MMP2*	chr16:54070317-54070664	-	347
NC_000016.8_TARGET_2|NlaIII/HpyCH4V|-|5	GGCAAGCTATTGGAGTGATCTCTTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGAGCCGCTGAGCCGGGC	*MMP2*	chr16:54070736-54071209	-	473
NC_000016.8_TARGET_3|MscI/DdeI|+|4	TTGCAGTTCCGACGCCACTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGGACCCGAACTTTCTTGGAAGAAGG	*CDH1*	chr16:67328694-67328986	+	292
NC_000016.8_TARGET_3|MseI/EcoO109I|-|1	TCCCCGGCCAGCCATGGGCCACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCATAACCCACCTAGACCCTAGCAA	*CDH1*	chr16:67328492-67328826	-	334
NC_000017.9_TARGET_2|MseI/EcoO109I|-|1	CCCCAGAGCATCACTTGGGCCACGATAACGGTAGAAAGCTTTGCTAACGGTCGAAAGACATAGTGTCCCCCTCAAGGCA	*BRCA1*	chr17:38530543-38530785	-	242
NC_000021.7_TARGET_1|MscI/DdeI|-|1	GGGTGCAGCCAGGAGGCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGACGTTTATAAAATGACCGCGGACT	*ERG*	chr21:38955115-38955366	-	251
NC_000022.9_TARGET_1|MscI/DdeI|-|2	CCAGGACTGCAGCGCTGCTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGACGGCTTCTCCTCCTCCTCTTG	*SYN3*	chr22:31527723-31528169	-	446
NC_000023.9_TARGET_1|MscI/DdeI|-|1	GTCGCACGCCATGCTGCTCAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACCAAAGCGCTATTAATCGGCCGC	*ELK1*	chr23:47394984-47395362	-	378
NC_000023.9_TARGET_1|MseI/EcoO109I|-|1	CGCTGATTGGCCAAAGCGCTATTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTCCACCCTCCAGGCCCAGA	*ELK1*	chr23:47394740-47394997	-	257
NC_000023.9_TARGET_1|MseI/EcoO109I|-|4	GGGTTGAACTGCGGCCAGAGTTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGACTGTGTCGGGGGGGCGGGGG	*ELK1*	chr23:47395190-47395599	-	409
NC_000023.9_TARGET_2|MscI/DdeI|+|1	CGACCCGCCCTCTGTTTATTTCTAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGGGCTGGGTTCCCGGCTGCG	*MYCL2*	chr23:106402338-106402619	+	281
NC_000023.9_TARGET_2|MscI/DdeI|+|2	CGGTTCCAGGAGACCAAAGCTGAACGATAACGGTAGAAAGCTTTGCTAACGGTCGAGACCTGGAGCAGGCCTGGATC	*MYCL2*	chr23:106402620-106402929	+	309
NC_000023.9_TARGET_2|NlaIII/HpyCH4V|-|6	CAGTCATCATCGCGGTGCGTGACGATAACGGTAGAAAGCTTTGCTAACGGTCGACATGTGGAGCGGCTTCTCCACC	*MYCL2*	chr23:106402735-106403027	-	292

Taille maximale	477
Taille minimale	204
Taille moyenne	332
Taille médiane	333
Ecart-type des tailles	78

Annexe 5: Séquences des oligonucléotides utilisés dans le séquençage Roche

Oligonucléotide	Séquence
Amorce 1	CGTATCGCCTCCCTCGCGCCATCAG-{MID}-AGCTTTGCTAACGGTCGA (Adaptateur A / Clé / MID / Séquence universelle 1 complémentaire au vecteur)
Amorce 2	CTATGCGCCTTGCCAGCCCGCTCAG-{MID}-AGCTTTCTACCGTTATCGT (Adaptateur B / Clé / MID / Séquence universelle 2 complémentaire au vecteur)
MID1	ACGAGTGCGT
MID2	ACGCTCGACA
MID3	AGACGCACTC
MID4	AGCACTGTAG
MID5	ATCAGACACG
MID6	ATATCGCGAG
MID7	CGTGTCTCTA

Annexe 6: Amorces de PCR utilisées dans les protocoles de sélection

Nom du sélector correspondant à la cible	Gène ciblé	qPCR sur ADN génomique Identifiant qPCR	Amorce sens Amorce anti-sens	qPCR sur ADN bisulphité Identifiant qPCR	Amorce sens Amorce anti-sens
NC_000001.9_TARGET_1|MscI/DdeI|-|1	HDAC1			Bis002	GAGGGTTTTTTGGGTTTGTATTAA AAATTTACAAAACTACTTTCCTCCC
NC_000001.9_TARGET_4|MscI/DdeI|+|4	PTGS2			Bis008	GGGGTAGTTTTTATTTTTTTGTTTGA AAACCAAATATCCTTCTACCCTCC
NC_000003.10_TARGET_3|MscI/DdeI|+|2	*MST1R*	Gen12	AGCCCCAAGATAGCGGAC GGGATTTGGGTTTCACAGG	Bis02	GGTTTGGGTTTGGATTTGGG AAAAAATTTAAATTTCACAAAAACCTAAAA
NC_000003.10_TARGET_3|MseI/EcoO109I|-|2	*MST1R*	Gen05	CTTCAGGTCAGGCCCAAG ACGAGGGCGACAGAAATG		
NC_000003.10_TARGET_4|MscI/DdeI|+|3	RASSF1			Bis001	AAATTTGGGTGTAGGGATTGTG CAATAAAACTAACCTCCAAAAACAC
NC_000003.10_TARGET_4|MseI/EcoO109I|-|2	*RASSF1*	Gen08	AGTGGAGTGCGACAAGGG CCCCATCGCTGAAGAGTG		
NC_000003.10_TARGET_5|MscI/DdeI|-|3	FHIT			Bis005	GTTATTTAGTGGGTATATTTTTAGG ATATTCCACTTAAACTTACCTCCCC
NC_000003.10_TARGET_5|MseI/EcoO109I|-|1	*FHIT*	Gen11	TACACCCCCAGAGCCAAG TTGGGGAAACTGAGGCAC		
NC_000004.10_TARGET_1|NlaIII/HpyCH4V|-|5	*IGFBP7*			Bis03	GGAGGATTTAATAGATGAAATTTTGGAT ATCCCCAATCCCCTACCCC
NC_000004.10_TARGET_3|MscI/DdeI|-|1	SFRP2			Bis007	GTTGTTTAATTTGTTGGGTTA AACACAAAAACTTCTTAATATCC
NC_000006.10_TARGET_5|MscI/DdeI|+|2	*THBS2*			Bis05	GGTTAAGTTTTTTTGGTGTATATTTGTA AAAATTAACTACTACCACTACTTCTC
NC_000007.12_TARGET_1|NlaIII/HpyCH4V|+|2	HOXA5			Bis011	TATTTAGGGGTAGATTTGGGGTT CCACATCAACAACAAAAAAAA
NC_000007.12_TARGET_2|MscI/DdeI|-|2	HOXA11			Bis004	TAAGGTAGTTTAATAATGGATTTTGATGAG CCTAAAACAAATTAAAAAAATAAAAATATA
NC_000007.12_TARGET_3|MscI/DdeI|+|2	*SFRP4*	Gen09	CGGCTTGATAGGGTCGTG GCTGCGCTTCTTCCTCTG		
NC_000007.12_TARGET_3|MscI/DdeI|-|3	SFRP4			Bis012	AGTTTGGGGGAAGAAATTTTTT CCAAATACAACCACAAACACAAC
NC_000007.12_TARGET_4|MscI/DdeI|-|5	*IGFBP1*	Gen07	TGTCAGAGGTCCCCGTTG CCGACCTGGACAGTCAGC		
NC_000007.12_TARGET_4|NlaIII/HpyCH4V|+|2	*IGFBP1*	Gen04	AACTGAGGGCCTGAACCC TGGGAGGAGGGTAAACGG		
NC_000007.12_TARGET_7|MseI/EcoO109I|-|4	*MDR1*			Bis01	GTTGTTAGATTTTTAATTTTGTTTT TTAATACCCCAACTACTCTAACC
NC_000008.9_TARGET_1|NlaIII/HpyCH4V|+|9	*SFRP1*	Gen10	GCTTGGTGTAGAAGCGCC ACGTGAGCTTCCAGTCGG		
NC_000008.9_TARGET_2|MseI/EcoO109I|-|2	RAD54B			Bis003	TTTGGGTAGTGGTTTTTGGG AAAAACCTAACCAATAAAAAAATAAAAAC
NC_000009.10_TARGET_1|MscI/DdeI|+|3	*CDKN2A*	Gen01	AGAGCCCCACCGAGAATC AATCAAGGGTTGAGGGGG		
NC_000011.8_TARGET_2|MscI/DdeI|-|3	*ASCL2*			Bis06	GTAGTTTATTTTTATTTTTAGTAGATTAA AACAAAACTAAATTCCTACTAAACCCC
NC_000012.10_TARGET_1|MscI/DdeI|+|6	*WNT10B*	Gen06	CTCCAACTCTCCCCACCC CGTTGTGCAGTCGGTGAG		
NC_000016.8_TARGET_1|MseI/EcoO109I|-|5	PYCARD			Bis009	TTTAGTATGTGGAATTGAGGGAGTT TTAAAACACCTAAACTTAAAACCTC
NC_000016.8_TARGET_2|MseI/EcoO109I|-|5	*MMP2*	Gen02	GAGGTCGTGCACTGAGGG CTCCACCTCTTTTCCCCG		
NC_000016.8_TARGET_2|NlaIII/HpyCH4V|-|5	*MMP2*	Gen03	CGCCCATCATCAAGTTCC CGCAGCAACTCACCACTG		
NC_000016.8_TARGET_2|NlaIII/HpyCH4V|-|5	MMP2			Bis010	GGAGAGATATTTTTATTTTTTGTTTT ATCTCCTAAACTACCTACTAAACCAC
NC_000017.9_TARGET_2|MseI/EcoO109I|-|1	*BRCA1*			Bis04	GGTTTAAGTGATGTTTTGGGGTAT AACACTCAATACCCCCTTCCTA
NC_000023.9_TARGET_1|MseI/EcoO109I|-|4	ELK1			Bis006	ATTTAGTTTTATGGTTTTGTTTAAT AAAAACTTCTAAACCCCTACCAAAA
Contrôles négatifs (régions non ciblées)					
Alu	*Alu*	Alu	Non disponible		
Contrôle négatif 1	*ACTB*	Neg1	Non disponible		
Contrôle négatif 2	*HLA-B*	Neg2	Non disponible		
Contrôle négatif 1	*DDB2*			Neg1	GGAAAGGTATTAGTTTTTTATAAAGT ATACCAACCAAAACTCTATCTCCC
Contrôle négatif 2	*IGF1R*			Neg2	AAATAAAAGGAATGAAGTTTGGTTT AACACATACTCACTTTCTCCACTC

Communications

Le développement du protocole de contrôle-qualité du MeDIP a fait l'objet de la publication suivante :

Sengenès, J., Daunay, A., Charles, M.A. and Tost, J. (2010) Quality control and single nucleotide resolution analysis of methylated DNA immunoprecipitation products. *Anal Biochem*, **407**, 141-143.

La mise en place du protocole de MeDIP-dep-Seq et les résultats obtenus grâce à cette technique ont été présentés sous forme de poster au congrès *Epigenomics of common diseases*, au Wellcome Trust Science Conference Centre, Hinxton, du 13 au 16 Septembre 2011.

La création de MeQA, une plateforme d'analyse de données issues du MeDIP-Seq, a fait l'objet de la publication suivante :

Huang, J., Renault, V., Sengenès, J., Touleimat, N., Michel, S., Lathrop, M. and Tost, J. (2012) MeQA: a pipeline for MeDIP-seq data quality assessment and analysis. *Bioinformatics*, **28**, 587-588.

Les publications et le poster sont présentés dans les pages suivantes.

Analytical Biochemistry 407 (2010) 141–143

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Notes & Tips

Quality control and single nucleotide resolution analysis of methylated DNA immunoprecipitation products

Jennifer Sengenès [a], Antoine Daunay [a], Marie-Aline Charles [b,c], Jörg Tost [a,d,*]

[a] Laboratory for Epigenetics, Centre National de Génotypage, CEA–Institut de Génomique, 91000 Evry, France
[b] INSERM, CESP Centre for Research in Epidemiology and Population Health, U1018, Team "Epidemiology of Diabetes, Obesity, and Renal Disease: Lifelong Approach," 94807 Villejuif, France
[c] Université Paris Sud 11, UMRS 1018, 94807 Villejuif, France
[d] Laboratory for Functional Genomics, Fondation Jean Dausset–CEPH, 75010 Paris, France

ARTICLE INFO

Article history:
Received 6 May 2010
Received in revised form 8 July 2010
Accepted 17 July 2010
Available online 23 July 2010

ABSTRACT

DNA methylation patterns are altered in many diseases, and their analysis has become of great interest. Methylated DNA immunoprecipitation (MeDIP) is a simple method to enrich the methylated fraction of the genome. However, it has been difficult to assess the quality and the detailed methylation patterns of the immunoprecipitated DNA. Here we present a simple method for the analysis of the immunoprecipitated DNA at single nucleotide resolution by bisulfite treatment and pyrosequencing of genomic regions. The presented method can be used as an initial quality measure prior to genome-wide read-out technologies such as microarrays and second-generation sequencing.

DNA methylation occurring at the 5 position of the pyrimidine ring of cytosines in the context of the dinucleotide sequence CpG forms one of the multiple layers of epigenetic mechanisms controlling and modulating gene expression through chromatin structure. Aberrant DNA methylation changes have been detected in several diseases, particularly cancer where genome-wide hypomethylation coincides with gene-specific hypermethylation [1]. DNA methylation patterns hold the promise to reflect at least a part of the influences of the environment on a phenotype, and analysis of DNA methylation patterns might provide valuable information on the exposure and thereby the susceptibility for a wide range of complex multifactorial diseases of an individual. As a stable nucleic acid-based modification with limited dynamic range that is technically easy to handle, DNA methylation is a promising biomarker for many applications [2] and might also be used for therapeutic interventions [3]. The promise of DNA methylation to be used as a biomarker for early detection carrying prognostic and predictive information has spurred the development of a large number of technologies for DNA methylation analysis [4], whereby the focus has recently shifted to genome-wide studies for the detection of aberrant or differential DNA methylation patterns. A simple method to analyze DNA methylation at a genome-wide scale is based on the enrichment of the methylated fraction of the genome by of methylated DNA immunoprecipitation (MeDIP)[1] with an antibody directed against methylcytidine [5]. MeDIP has been combined with various read-out technologies such as quantitative polymerase chain reaction (qPCR) for locus-specific analyses as well as genome-wide analyses of the methylome using microarrays and, more recently, second-generation sequencing [5–8]. Mainly PCR-based methods have been used to determine the efficiency and recovery rate of MeDIP using either control oligonucleotides or regions that are supposed to be highly methylated or completely unmethylated in the cell type under investigation. The MeDIP protocol requires several methylated cytosines within the fragment to be reproducible. However, there is little information on detailed methylation patterns (i.e., the methylation degree of individual CpGs within the immunoprecipitated fragment), and it is not known whether only completely methylated molecules are immunoprecipitated. Pyrosequencing is a quantitative real-time sequencing technology for small regions of interest such as the promoter region of a gene. It is ideally suited for DNA methylation analysis after bisulfite treatment because it combines the ability of direct quantitative sequencing with single nucleotide resolution, reproducibility, speed, and ease of use [9,10].

Here we introduce the quantitative analysis at single nucleotide resolution of immunoprecipitated methylated DNA using pyrosequencing after bisulfite treatment of a part of the immunoprecipitated DNA that can be used as a measure of quality of the immunoprecipitated DNA prior to analysis on genome-wide read-out technologies such as microarrays and second-generation sequencing.

DNA was extracted from twelve fresh frozen human placentas without any pathological aberrations and adjusted to a concentration of 25 ng/μl. 2.5 μg of the samples were sheared by sonication

* Corresponding author at: Laboratory for Epigenetics, Centre National de Génotypage, CEA–Institut de Génomique, 91000 Evry, France. Fax: +33 160878485.
E-mail address: tost@cng.fr (J. Tost).

[1] *Abbreviations used:* MeDIP, methylated DNA immunoprecipitation; qPCR, quantitative polymerase chain reaction; EB, elution buffer.

0003-2697/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2010.07.013

Table 1
Quality control of MeDIP experiments

	Average enrichment factor	Average IP DNA quantification (μg)	Amplifiable DNA after bisulfite treatment (μg)
1	240	0.62	0.12
2	165	0.53	0.15
3	131	0.59	0.12
4	164	0.62	0.09
5	162	0.42	0.15
6	155	0.52	0.15
7	326	0.56	0.09
8	337	0.50	0.15
9	230	0.47	0.15
10	220	0.41	0.15
11	254	0.41	0.15
12	180	0.48	0.15

Note. The average enrichment factor (second column) is evaluated by qPCR based on unmethylated and fully methylated controls for each of the 12 analyzed placentas (first column). The average quantity of recovered immunoprecipitated DNA (IP DNA) starting with 1 μg of sheared genomic DNA is measured by a Nanodrop (third column). Quantities of amplifiable DNA after bisulfite treatment of 0.2 μg of IP DNA are shown in the fourth column. Deviations from the expected amount are probably due to inaccuracies in the optical density measurement and/or to a loss of material during bisulfite treatment.

(Bioruptor, Diagenode, Sparta, NJ, USA) in 100 μl of 1× TE buffer to the desired size of 300–1000 bp, which was verified on a 2% agarose gel. 1 μg of the sheared DNA was incubated with the antibody and methylated fragments immunoprecipitated using an automated device (SX-8G IP-Star Robot, Diagenode) and the AutoMeDIP kit (Diagenode) as described by the manufacturer. Prior to immunoprecipitation unmethylated and fully methylated control oligonucleotides included in the kit were added to the samples. The efficiency of the MeDIP was subsequently assessed by qPCR on a Mx3005P thermocycler (Stratagene, La Jolla, CA, USA) based on these controls following the manufacturer's instructions. The entire experimental procedure was performed in triplicate using independent immunoprecipitations for each experiment. An average enrichment ratio of 215 (range: 54–524) was obtained for the different immunoprecipitations (Table 1, second column). In order to assess the feasibility of a bisulfite conversion on low amounts of DNA, we tested two commercial kits for bisulfite conversion using a range of 10 to 300 ng of input DNA. Amplification efficiency was subsequently tested using two genomic regions (*IGF2* DMR0 and *IGF2R* DMR, data not shown). We did not observe a major difference in performance between the kits tested: MethylEasy *Xceed* (Human Genetic Signatures, North Ryde, NSW, Australia) and the Epitect Bisulfite kit (Qiagen, Valencia, CA, USA). Amplification products were obtained from all starting amounts using PCR amplification conditions commonly used for pyrosequencing (see below). The available quantity of immunoprecipitated DNA after immunoprecipitation of 1 μg of sheared genomic DNA was roughly assessed through measurement of the optical density using a Nanodrop (Table 1, third column). Then 0.2 μg of the immunoprecipitated DNAs and – as a control – the same amount of the samples that were not immunoprecipitated were bisulfite converted using the Epitect Bisulfite kit according to the manufacturer's instructions (Qiagen). Bisulfite converted DNA was eluted in 30 μl of the provided elution buffer (EB). The quantity of amplifiable material of the 24 bisulfite-treated samples and its integrity were assessed by qPCR as described previously [11]. On average 0.135 μg of amplifiable bisulfite treated DNA was detected (range: 0.09–0.15; Table 1, fourth column). Five regions in four genes known to be partially methylated in human placentas were selected for the validation of our approach: 1) the DMR0 of *IGF2* [12], the promoter CpG island of *RASSF1A*, known to be allele-specifically methylated in human placentas [13] as well as three other regions in *OSTM1* and *FAM50B*, identified to display intermediate methylation levels in human placentas in a genome–wide screen (Tost et al., unpublished). Thus all regions should be present after immunoprecipitation in the placental DNA. Regions of interest were PCR amplified using 1–2 μl after bisulfite treatment and 5 pmol of forward and reverse primers, one of them being biotinylated. Oligonucleotides for PCR amplification and pyrosequencing (Supplementary Table S1) were synthesized by Biotez (Buch, Germany). Standard reaction conditions were 1x HotStar Taq buffer supplemented with 1.6 mM $MgCl_2$, 100 μM dNTPs and 2.0 U HotStar Taq polymerase (Qiagen, Valencia, CA, USA) in a 25 μl volume. The PCR program consisted of a denaturing step of 15 min at 95 °C followed by 50 cycles of 30 s at 95 °C, 30 s at the respective annealing temperature (Supplementary Table S1) and 15 s at 72 °C, with a final extension of 5 min at 72 °C. After verification by standard gel electrophoresis, quantitative DNA methylation analysis of the bisulfite treated DNA was performed by pyrosequencing as previously described using 10 μl of PCR product [10]. Quantitative DNA methylation analysis with the respective sequencing primers (Supplementary Table S1) was carried out on a PSQ 96MD system with the PyroGold SQA Reagent Kit (Qiagen) and results were analyzed using the Q-CpG software (V.1.0.9, Qiagen). In parallel to the samples we analyzed commercial completely methylated DNA (Epitect methylated human control DNA, bisulfite converted, Qiagen) to account for the background/noise level of the pyrosequencing technology. Results for two of the regions are presented in Fig. 1 and their corresponding pyrograms are shown in the Supple-

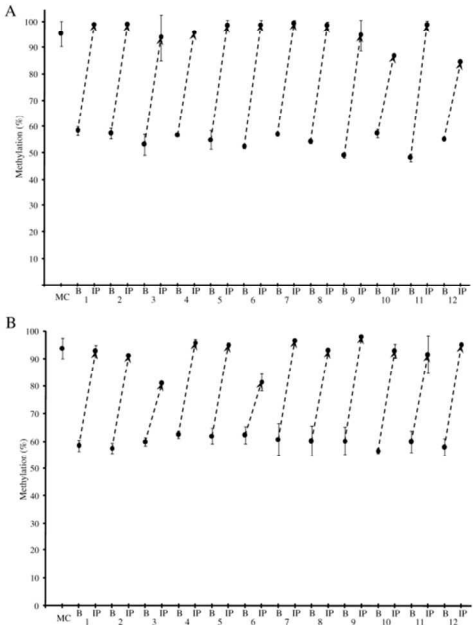


Fig. 1. Pyrosequencing of immunoprecipitated DNAs after bisulfite for A. OSTM1, B. FAM50B AMP1 (see Supplementary Fig. S2 for the corresponding pyrograms and Supplementary Fig. S3 for Pyrosequencing data of FAM50B AMP2, RASSF1A and IGF2). Methylation data of Methylated Control (MC) DNA and twelve placental DNAs. A median of all individual CpG methylation percentages has been calculated over each region. The average of the triplicates derived from three independent bisulfite treatments and immunoprecipitations is represented with their standard deviation. For each sample the data before (B) and after (IP) MeDIP is represented and linked by a dotted arrow for easier visualization.

mentary information (Supplementary Fig. S2) as well as the results for the other three regions analysed (Supplementary Fig. S3). A hallmark of the pyrosequencing technology is the fact that the intensity of the bioluminometric response is directly proportional to the amount of incorporated nucleotides, i.e. a peak corresponding to the incorporation of two consecutive (and identical) nucleotides will have the double height compared to the signal of a single nucleotide incorporation. The peak heights in the resulting output format thus inform on the extent of homopolymeric sequences and proportions of methylated and unmethylated alleles at potentially polymorphic CpG positions can be deduced directly from the relative height of the peaks corresponding to variable nucleotide positions. For all five regions intermediate DNA methylation levels of around 40–60% were found in the bisulfite treated placental samples (middle pyrogram in Supplementary Fig. S2 and in the left panels of Supplementary Fig. S3). After immunoprecipitation and bisulfite treatment a large increase in the DNA methylation levels (Lower pyrogram labelled "IP" in Supplementary Fig. S2 and in the left panels of Supplementary Fig. S3) comparable to the level found in the completely methylated control DNA (Upper pyrogram labelled "Control" in Supplementary Fig. S2 and in the left panels of Supplementary Fig. S3) was observed demonstrating the feasibility of our approach and the high quality of the result of the automated immunoprecipitation. The uniform and high methylation of the immunoprecipitated DNA comparable to the control DNA indicates that mainly completely methylated molecules were immunoprecipitated, which was at least expected for the imprinted *IGF2* DMR. Immunoprecipitation and bisulfite treatment were performed in triplicates to assess the variability of our approach. Percentages of the methylation degree between triplicates differed on average by 4.1% (Interquartile range (IQR): 0.5–2.9%) for *OSTM1*, 6.6% (IQR: 1.9–4.4%) for the first amplification product and 3.9% (IQR: 2.1–4.8%) for the second amplification product in *FAM50B*, 2.6% (IQR: 0.8–2.3%) for *IGF2* DMR0 and 6.3% (IQR: 4.1–8.3%) for *RASSF1A*. Methylation levels were lower for RASSF1A in the immunoprecipitated and the control DNA. This could be due to the high density of CpGs in the analyzed region and methylation of a subset of CpGs might be sufficient for silencing the methylated allele. However no hotspots for DNA methylation were discovered indicating that methylation is uniformly distributed between the CpGs on the large number of molecules analyzed by the pyrosequencing approach.

The developed method permits the rapid quality control of immunoprecipitated methylated DNA independent of the protocol used and should have a wide applicability. As only part of the DNA needs to be bisulfite treated, most of the immunoprecipitated DNA will be available for subsequent genome-wide read-out technologies such as (tiling) microarrays or second generation sequencing. In addition, this approach allows for the detailed analysis of consecutive CpGs and in combination with next-generation sequencing would permit to sequence directly the bisulfite treated DNA at genome-wide scale providing DNA methylation patterns at single nucleotide resolution. This approach would be of great interest especially in regions with heterogeneous methylation levels, but would restrict the methylation analysis to the methylated fraction of the genome avoiding some of the challenges of whole genome bisulfite sequencing which is currently still cost-prohibitive and requires extensive computational resources.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (ANR) under the "Programme National de Recherche en Alimentation et Nutrition Humaine" (ANR-06-PNRA-022-01), the European Union Framework 7 Integrated Project READNA (HEALTH-F4-2008-201418), and the French Ministry of Research. J.S. received an IRTELIS fellowship from the CEA. We thank Sandrine Barbaux (Hopital Cochin, Paris) for extraction of the DNAs from the tissue samples.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2010.07.013.

References

[1] S. Sharma, T.K. Kelly, P.A. Jones, Epigenetics in cancer, Carcinogenesis 31 (2010) 27–36.
[2] J. Tost, DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker, Mol. Biotechnol. 44 (2010) 71–81.
[3] C.B. Yoo, P.A. Jones, Epigenetic therapy of cancer: past, present, and future, Nat. Rev. Drug Discov. 5 (2006) 37–50.
[4] P.W. Laird, Principles and challenges of genome-wide DNA methylation analysis, Nat. Rev. Genet. 11 (2010) 191–203.
[5] M. Weber, J.J. Davies, D. Wittig, E.J. Oakeley, M. Haase, W.L. Lam, D. Schübeler, Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells, Nat. Genet. 37 (2005) 853–862.
[6] M. Weber, I. Hellmann, M.B. Stadler, L. Ramos, S. Pääbo, M. Rebhan, D. Schübeler, Distribution, silencing potential, and evolutionary impact of promoter DNA methylation in the human genome, Nat. Genet. 39 (2007) 457–466.
[7] D. Zilberman, M. Gehring, R. K. Tran, T. Ballinger, S. Henikoff, http://www.nature.com/ng/journal/v39/n1/full/ – a1 Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription, Nat. Genet. 39 (2007) 61–69.
[8] A.L. Brunner, D.S. Johnson, S.W. Kim, A. Valouev, T.E. Reddy, N.F. Neff, E. Anton, C. Medina, L. Nguyen, E. Chiao, C.B. Oyolu, G.P. Schroth, D.M. Absher, J.C. Baker, R.M. Myers, Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver, Genome Res. 19 (2009) 1044–1056.
[9] J.M. Dupont, J. Tost, H. Jammes, I.G. Gut, De novo quantitative bisulfite sequencing using the pyrosequencing technology, Anal. Biochem. 333 (2004) 119–127.
[10] J. Tost, I.G. Gut, DNA methylation analysis by pyrosequencing, Nat. Protoc. 2 (2007) 2265–2275.
[11] M. Campan, D.J. Weisenberger, B. Trinh, P.W. Laird, MethyLight, Methods Mol. Biol. 507 (2009) 325–337.
[12] D. Monk, R. Sanches, P. Arnaud, S. Apostolidou, F.A. Hills, S. Abu-Amero, A. Murrell, H. Friess, W. Reik, P. Stanier, M. Constância, G.E. Moore, Imprinting of IGF2 P0 transcript and novel alternatively spliced INS–IGF2 isoforms show differences between mouse and human, Hum. Mol. Genet. 15 (2006) 1259–1269.
[13] R.W.K. Chiu, S.S.C. Chim, I.H.N. Wong, C.S.C. Wong, W-S. Lee, K.F. To, J.H.M. Tong, R.K.C. Yuen, A.S.W. Shum, J.K.C. Chan, L.Y.S. Chan, J.W.F. Yuen, Y.K. Tong, J.F. Weier, C. Ferlatte, T.N. Leung, T.K. Lau, K.W. Lo, Y.M.D. Lo, Hypermethylation of RASSF1A in human and rhesus placentas, Am. J. Pathol. 170 (2007) 941–950.

MeDIP-dep-Seq: A new protocol for genome-wide analysis of cancer-related DNA methylation signatures

Jennifer Sengenès[1], Jinyan Huang[2], Jörg Tost[1,3]

[1]Laboratory for Epigenetics, CEA, Centre National de Génotypage, Evry, France
[2]Laboratory for Bioinformatics, Fondation Jean Dausset-CEPH, Paris, France
[3]Laboratory for Functional Genomics, Fondation Jean Dausset-CEPH, Paris, France

Introduction

Altered DNA methylation patterns are found in many diseases and especially cancer where genome-wide hypomethylation coincides with gene-specific hypermethylation. MeDIP-Seq (Methylated DNA Immunoprecipitation followed by high-throughput sequencing) is one of the most simple and widely used approaches to enrich the methylated fraction of the genome. MeDIP immunoprecipitates highly methylated sequences many of which are located in the repetitive sequences and represent a substantial part of mammalian genomes: human coding sequences represent only 5% of the genome whereas repeats are spread through at least 50% of it [1]. Although these sequences might be of potential interest for specific biological and clinical questions, they are difficult to align unambiguously after sequencing leading to a large number of sequences that are currently not used for further analysis. We present an innovative method called MeDIP-dep-Seq to deplete a significant part of these highly repetitive sequences while sequences of interest are not affected.

Methods

MeDIP ...

Paired-end sample-prep

3.6 µg DNA were fragmented to a size between 300 and 1000 bp. 3.3 µg were subjected to Illumina's GAIIx paired-end library preparation.

Immunoprecipitation (IP)

MeDIP enrichment was performed on 1.0 µg DNA as previously described [2] using 150 ng of an antibody directed against 5-methylcytidine. We used an automated protocol recently described by Butcher *et al.* [3].

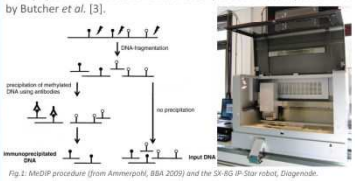


Fig.1: MeDIP procedure (from Ammerpohl, BBA 2009) and the SX-8G IP-Star robot, Diagenode.

Quality control

IP efficiency was controled by a qPCR based on the control oligos provided by the manufacturer. We obtained a mean recovery of 40% and 405 ng of immunoprecipitated DNA. We also developed a quantitative analysis at single nucleotide resolution using pyrosequencing after bisulfite treatment [4].

MeDIP-dep ...

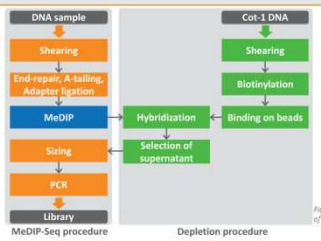


Fig.2: Outline procedure of MeDIP-dep-Seq.

Biotinylation of Cot-1 DNA

Cot-1 DNA was sheared, 200 ng were labeled with biotin and amplified using the BioPrime DNA labeling system.

Depletion of repetitive sequences

2.5 µg of biotinylated Cot-1 DNA were added to 1.0 mg washed streptavidin coated magnetic beads. After incubation, the beads were washed and resuspended in hybridization buffer. 250 ng of immunoprecipitated DNA were added and incubated for 3 hours at 62°C. The supernatant was removed and the whole procedure was repeated. After this double incubation, the supernatant was removed and DNA purified. The depletion was controled by qPCR based on several families of repeat elements including *Line*, *miSat* and *maSat* and *Igf2* as control assays.

Automation of the process

We developed an automated protocol (SX-8G IP-Star) that permits the preparation of the beads and binding to the biotinylated Cot-1 DNA in 50 minutes.

MeDIP-dep-Seq

Preparation of libraries

100 ng were used for gel-based size-selection. We then prepared a library by PCR amplification on the entire volume of the size-selected DNA (20-25 cycles). One lane of paired-end sequencing was performed on an Illumina Genome Analyzer IIx.

Fig.3: A flowcell on which sequencing templates are immobilized and the Genome Analyzer IIx, Illumina.

Results / Discussion

Depletion of repetitive sequences

We used DNA from 4 MEF (Mouse Embryonic Fibroblast) cell lines for the proof-of-principle of the MeDIP-dep-Seq protocol. The qPCR results were compared to a negative control sample for which no Cot-1 DNA was used in the depletion procedure. The current protocol yields a decrease in repetitive elements of more than 100-fold (Fig.4) for several classes of repeats while unique sequences of interest were not affected.

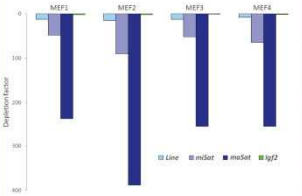


Fig.4: Depletion is measured for several classes of repeats including Line (Long Interspersed Nucleotides Elements), miSat (minor Satellites) and maSat (major Satellites). Unique sequences of interest represented here by Igf2 are saved.

Pipeline for sequencing data analysis

We further developed a pipeline called MeQA [Huang, *J. et al.*, submitted] for the analysis of MeDIP-Seq datasets. It integrates several customized scripting and existing tools for pre-processing and quality assessment of the data, read distribution analysis and methylation estimation by using the MEDIPS package [5].

Fig.5: Screenshot from http://life.tongji.edu.cn/meqa/

MeDIP-Seq *vs* MeDIP-dep-Seq

Distribution of mapped reads

The alignment score substantially increased from 30% only by MeDIP to 80% by introducing the depletion step in the sample-prep (Fig.6). This multiplies thus by 3-fold the amount of usable sequences per sequencing lane. We could for example cover more unique sequences of interest (increase from 5 to 55%) like coding sequences. CpGs located in the CpG islands are still covered (Fig.7) and we can cover more of them with the depletion while sequencing reads are less present in the repeats regions (Fig.6,7).

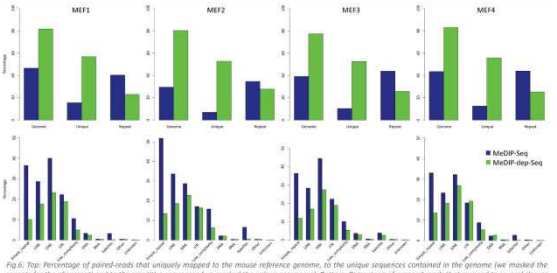


Fig.6: Top: Percentage of paired-reads that uniquely mapped to the mouse reference genome, to the unique sequences contained in the genome (we masked the repeats for the alignment) and to the repetitive sequences (we masked the unique sequences). Bottom: Percentage of mapped reads that mapped to several classes of repeats including simple repeats (micro-satellites), LINEs (Long Interspersed Nucleotide Elements), SINEs (Short Interspersed Nucleotide Elements, which include Alu elements), LTRs (Long Terminal Repeat elements), low complexity repeats, DNA and RNA repeats and satellites.

Families of depleted repeats

Several classes of repeat elements were removed through the depletion step. The depletion rate depends on their proportion in the original Cot-1 DNA (Fig.7).

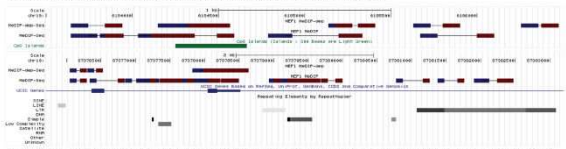


Fig.7: MeDIP-dep-Seq and MeDIP-Seq data integrated into the UCSC genome browser for visualization. Reads are represented with blue and red blocks. Paired reads are linked by a black line. Reads are collapsed for easy visualization. Top: 3000 bp window including a CpG island. 26 non-redundant paired-reads were counted for MeDIP-dep and 9 for MeDIP. Bottom: 7500 bp window including several classes of repeat elements. 23 non-redundant paired-reads were counted for MeDIP-dep and 61 for MeDIP.

Methylation values?

A good correlation was observed between methylation values calculated on MeDIP-Seq and MeDIP-dep-Seq data for the whole genome (Fig.8). We are now developing a robust approach to identify Differentially Methylated Regions in the genome-wide DNA methylation data.

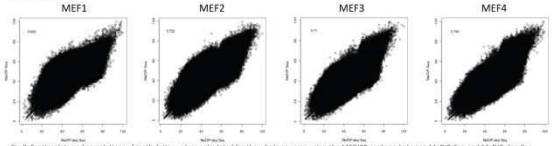


Fig.8: Scatterplots and correlations of methylation values calculated for the whole genome using the MEDIPS package between MeDIP-Seq and MeDIP-dep-Seq.

Conclusion

The proof-of-principle experiments clearly demonstrate the advantages of the developed protocol. The use of the dedicated robotics leads to a significant reduction of the inherent variation due to the multi-step protocol. We also developed quality control assays to validate every step. The innovative depletion we introduced into the MeDIP-Seq procedure leads to a substantial decrease of highly repetitive elements that generally waste a significant part of the data. This would potentially allow multiplexing of several samples in a single lane of second generation sequencers. The combination of this technology with the dedicated analysis pipeline will permit the rapid identification of differentially methylated regions, establish methylome maps and identify novel epigenetic markers for cancer and complex diseases.

References

[1]Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) *Nature*, **409**, 860-921.
[2]Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E.M. *et al.* (2008) *Nat Biotechnol*, **26**, 779-785.
[3]Blutcher, L.M. and Beck, S. (2010) *Methods*, **52**, 223-231.
[4]Sengenès, J., Daunay, A., Charles, M.A. and Tost, J. (2010) *Anal Biochem*, **407**, 141-143.
[5]Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R. and Adjaye, J. (2010) *Genome Res*, **20**, 1441-1450.

Acknowledgments
This work was funded through the EU FP7 Integrated Project READNA (HEALTH-F4-2008-201418).

sengenes@cng.fr

Epigenomics of common diseases, 13-16 September 2011, Wellcome Trust

BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 4 2012, pages 587–588
doi:10.1093/bioinformatics/btr699

Sequence analysis Advance Access publication December 22, 2011

MeQA: a pipeline for MeDIP-seq data quality assessment and analysis

J. Huang[1,2,*], V. Renault[2], J. Sengenès[3], N. Touleimat[2], S. Michel[4], M. Lathrop[2,3] and J. Tost[2,3]

[1]School of life science, Tongji University, 200092 Shanghai, China [2]Fondation Jean Dausset-CEPH, 75010 Paris, [3]CEA, Centre National de Génotypage, 91000 Evry, France and [4]Hannover Medical School, 30625 Hannover, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: We present a pipeline for the pre-processing, quality assessment, read distribution and methylation estimation for methylated DNA immunoprecipitation (MeDIP)-sequence datasets. This is the first MeDIP-seq-specific analytic pipeline that starts at the output of the sequencers. This pipeline will reduce the data analysis load on staff and allows the easy and straightforward analysis of sequencing data for DNA methylation. The pipeline integrates customized scripting and several existing tools, which can deal with both paired and single end data.

Availability: The package and extensive documentation, and comparison to public data is available at http://life.tongji.edu.cn/meqa/

Contact: jhuang@cephb.fr

Received on July 8, 2011; revised on December 4, 2011; accepted on December 18, 2011

1 INTRODUCTION

Methylated DNA immunoprecipitation (MeDIP) enables the rapid identification of genomic regions containing methylated cytosines. MeDIP, in combination with hybridization to high-resolution tiling microarrays or high-throughput sequencing (HTS) techniques, is a useful method for the identification of methylated CpG-rich sequences (Jacinto *et al.*, 2008). Recently, several benchmark publications reported on the use of MeDIP-seq for genome-wide DNA methylation analysis and compared it to several others methods, for example, whole-genome bisulfite-sequencing (BS-seq) and methyl-binding protein-based enrichment of methylated sequences (MBD-seq) (Bock *et al.*, 2010; Harris *et al.*, 2010; Li *et al.*, 2010). Though MeDIP is not substitute for BS-seq to obtain a methylome at single-nucleotide resolution, the generation of genome-wide data derived from MeDIP-seq provides a major tool for epigenetic studies in health and disease (Harris *et al.*, 2010; Li *et al.*, 2010). Further, MeDIP is specific for methylated cytosines and results are not confounded by the presence of hydroxymethylated cytosines unlike bisulfite-based methods. The main challenges resulting from the rapidly advancing technology development in DNA methylation analysis is now the computational analysis of the genome-wide sequencing data (Laird, 2010).

*To whom correspondence should be addressed.

Two methods (Batman and MEDIPS) have been developed for MeDIP-seq data analysis (Chavez *et al.*, 2010; Down *et al.*, 2008). However, both of them do not include quality control of sequencing data or the read mapping. These methods require, therefore, considerable effort to prepare the data and run several other sequencing and quality control packages separately, increasing analysis time and potentially introducing processing errors.

To fill the gap between experimental throughput and processing speed, we developed the MeQA pipeline for pre-processing, data quality assessment and distribution of sequences reads and estimation of DNA methylation levels of MeDIP-seq datasets. The pipeline will also generate files for the UCSC browser. The pipeline presented here runs on the Unix/Linux platform and was written in the popular bioinformatics languages shell script, Python and R. It can be run locally on a single Linux/Unix or Mac server. We have tested this pipeline on our cluster with qsub and bsub commands with excellent performance. On a DELL 16 CPUs (each 2.67 GHz) and 32 GB memory computer server, it takes ~20 h to run the entire pipeline for a mouse genome-wide DNA methylation estimate that contained 20 million 50-bp length single end reads. When several lanes are combined, which would provide a similar sequencing depth as HiSeq2000 data, it will take ~30 h. We recommend running this pipeline on a computer with at least 16 GB memory.

2 METHODS

The execution of MeQA is straightforward and easy. After preparation of the configuration file, a simple command line calls the pipeline. We incorporated several existing computer packages into the MeQA pipeline. As installation of these packages requires some effort, we prepared a script to install these packages conveniently.

The pipeline is described in two parts. Each part can be run independently and be easily exchanged with other software if required. Part A performs the quality control of the DNA sequence information. First, MeQA provides a quick overview of sequence problems and data quality can be quickly assessed since it provides summary graphs. These results are exported to a pdf-based permanent report. The raw sequence is then aligned to a reference sequence genome using BWA (Li and Durbin, 2009). The alignments are saved in the standard SAM format, converted to BAM format and sorted with SAMtools (Li *et al.*, 2009). A shell script provides automatic download of references and index files from UCSC (Fujita *et al.*, 2011), or a local directory can be provided for the alignment to a custom user provided reference. Quality of the sequence read mapping is accessed by SAMStat, and results are presented as a HTML report that includes unmapped, as well as poorly and accurately mapped reads, separately.

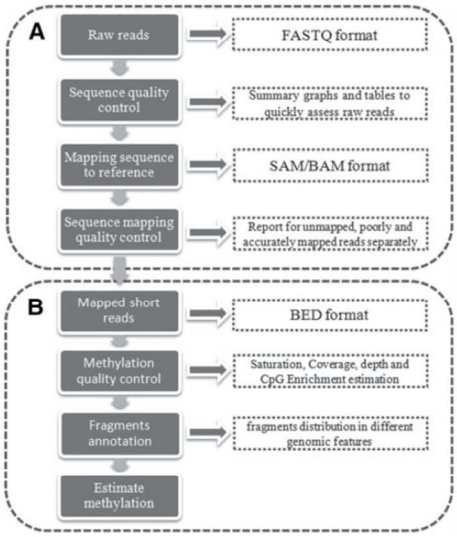


Fig. 1. The MeQA analysis pipeline. (**A**) General sequence quality control. (**B**) Assessment of read distribution and estimation of DNA methylation levels.

Part B deals with the read distribution for different genomic features and estimates of the DNA methylation level. Methylated mapped region can be extracted as BED format by SAMtools and BEDtools (Quinlan and Hall, 2010). These BED files are supplied to MEDIPS that estimates the reproducibility of the genome-wide DNA methylation profile with respect to the total number of given short reads and to the size of the reference genome. MEDIPS also analyzes the coverage of genome-wide DNA sequence patterns (e.g. CpGs) by the given reads, and calculates a CpG enrichment factor as a quality control for the immunoprecipitation. For annotation of the methylated regions, CEAS (Shin *et al.*, 2009) is used to calculate the percentages of regions that correspond to: (i) promoters, (ii) bidirectional promoters, (iii) downstream of a gene and (iv) gene (3′UTRs, 5′UTRs, coding exons and introns). Lastly, MEDIPS summarizes the DNA methylation levels for genome-wide windows of a specified size or for user-defined regions of interest.

A flow chart of the analysis pipeline is shown in Figure 1.

3 DISCUSSION

The MeQA pipeline consists of two parts: the first one provides the general quality control of the sequence reads, and the second assesses the quality of the DNA methylation analysis experiment and provides comprehensive analysis. The first part could in principle also be applied to other sequence data, e.g. for quality assessment of RNA-seq data. The MeQA package allows users to obtain methylation estimates from raw sequence files with a single python function call. The main advantage of MeQA is its ease of use and its availability as open source, as all programs added to the pipeline are open source programs. Widely used file formats are used for each step of the pipeline, e.g. FASTQ, SAM/BAM and BED files. Each step of the pipeline can be replaced without compromising the workflow allowing users to update components, replace some packages and to extend and customize the pipeline for their needs.

This pipeline permits to estimate methylation levels in differentially methylated regions and genes using MEDIPS. Further functional analysis such as GO enrichment and KEGG pathway analysis for differentially methylated genes can be performed using additional Bioconductor packages.

ACKNOWLEDGEMENTS

The authors thank Prof. Howard Cann and Zhongwen CHANG for helpful discussions on this paper.

Funding: National Natural Science Foundation of China (grant no. 30900838); National Basic Research Program of China (973 Program No. 2010CB944904); European Union framework 7 integrated project READNA (contract no. HEALTH-F4-2008-201418).

Conflict of Interest: none declared.

REFERENCES

Bock,C. *et al.* (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, **28**, 1106–1114.

Chavez,L. *et al.* (2010) Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res.*, **20**, 1441–1450.

Down,T.A. *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.*, **26**, 779–785.

Fujita,P.A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.

Harris,R.A. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.

Jacinto,F.V. *et al.* (2008) Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *Biotechniques*, **44**, 35, 37, 39 passim.

Laird,P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,N. *et al.* (2010) Whole genome DNA methylation analysis based on high throughput sequencing technology *Methods*, **52**, 203–212.

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

Shin,H. *et al.* (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics*, **25**, 2605–2606.

Références bibliographiques

1. Waddington, C.H. (1942) The epigenotype. *Endeavour*, **1**, 18-20.
2. Bannister, A.J. and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Res*, **21**, 381-395.
3. Schuettengruber, B., Martinez, A.M., Iovino, N. and Cavalli, G. (2011) Trithorax group proteins: switching genes on and keeping them active. *Nat Rev Mol Cell Biol*, **12**, 799-814.
4. Bantignies, F. and Cavalli, G. (2011) Polycomb group proteins: repression in 3D. *Trends Genet*, **27**, 454-464.
5. Khraiwesh, B., Arif, M.A., Seumel, G.I., Ossowski, S., Weigel, D., Reski, R. and Frank, W. (2010) Transcriptional control of gene expression by microRNAs. *Cell*, **140**, 111-122.
6. Saxena, A. and Carninci, P. (2011) Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *Bioessays*, **33**, 830-839.
7. Prasad, P. and Bartholomew, B. (2010) Control of nucleosome movement: to space or not to space nucleosomes? *Epigenetics*, **5**, 282-286.
8. Pfeifer, G.P. (2006) Mutagenesis at methylated CpG sequences. *Curr Top Microbiol Immunol*, **301**, 259-281.
9. Fryxell, K.J. and Moon, W.J. (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol Biol Evol*, **22**, 650-658.
10. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*, **16**, 6-21.
11. Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J Mol Biol*, **196**, 261-282.
12. 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.
13. Meissner, A., Gnirke, A., Bell, G.W., Ramsahoye, B., Lander, E.S. and Jaenisch, R. (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res*, **33**, 5868-5877.
14. Klose, R.J. and Bird, A.P. (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci*, **31**, 89-97.
15. 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.
16. Globisch, D., Munzel, M., Muller, M., Michalakis, S., Wagner, M., Koch, S., Bruckl, T., Biel, M. and Carell, T. (2010) Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One*, **5**, e15367.
17. Kriaucionis, S. and Heintz, N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, **324**, 929-930.
18. 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.
19. Williams, K., Christensen, J. and Helin, K. (2011) DNA methylation: TET proteins-guardians of CpG islands? *EMBO Rep*, **13**, 28-35.
20. Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S. and Reik, W. (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, **473**, 398-402.
21. Song, C.X. and He, C. (2011) The hunt for 5-hydroxymethylcytosine: the sixth base. *Epigenomics*, **3**, 521-523.
22. Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*, **99**, 3740-3745.

23. Illingworth, R.S. and Bird, A.P. (2009) CpG islands – 'A rough guide'. *FEBS Letters*, **583**, 1713-1720.
24. Antequera, F. (2003) Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci*, **60**, 1647-1658.
25. Joulie, M., Miotto, B. and Defossez, P.A. (2010) Mammalian methyl-binding proteins: what might they do? *Bioessays*, **32**, 1025-1032.
26. Slotkin, R.K. and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, **8**, 272-285.
27. Serre, D., Lee, B.H. and Ting, A.H. (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res*, **38**, 391-399.
28. Horard, B., Eymery, A., Fourel, G., Vassetzky, N., Puechberty, J., Roizes, G., Lebrigand, K., Barbry, P., Laugraud, A., Gautier, C. *et al.* (2009) Global analysis of DNA methylation and transcription of human repetitive sequences. *Epigenetics*, **4**, 339-350.
29. Reik, W., Dean, W. and Walter, J. (2001) Epigenetic reprogramming in mammalian development. *Science*, **293**, 1089-1093.
30. Senner, C.E. (2011) The role of DNA methylation in mammalian development. *Reprod Biomed Online*, **22**, 529-535.
31. Abramowitz, L.K. and Bartolomei, M.S. (2011) Genomic imprinting: recognition and marking of imprinted loci. *Curr Opin Genet Dev*.
32. Cassidy, S.B., Schwartz, S., Miller, J.L. and Driscoll, D.J. (2012) Prader-Willi syndrome. *Genet Med*, **14**, 10-26.
33. Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G.P., Haig, D. and Dulac, C. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643-648.
34. Conrad, T. and Akhtar, A. (2011) Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet*, **13**, 123-134.
35. Kharchenko, P.V., Xi, R. and Park, P.J. (2011) Evidence for dosage compensation between the X chromosome and autosomes in mammals. *Nat Genet*, **43**, 1167-1169; author reply 1171-1162.
36. Heard, E. and Disteche, C.M. (2006) Dosage compensation in mammals: fine-tuning the expression of the X chromosome. *Genes Dev*, **20**, 1848-1867.
37. Rakyan, V.K., Down, T.A., Balding, D.J. and Beck, S. (2011) Epigenome-wide association studies for common human diseases. *Nat Rev Genet*, **12**, 529-541.
38. Fraga, M.F. and Esteller, M. (2007) Epigenetics and aging: the targets and the marks. *Trends Genet*, **23**, 413-418.
39. Wierda, R.J., Geutskens, S.B., Jukema, J.W., Quax, P.H. and van den Elsen, P.J. (2010) Epigenetics in atherosclerosis and inflammation. *J Cell Mol Med*, **14**, 1225-1240.
40. Ospelt, C., Reedquist, K.A., Gay, S. and Tak, P.P. (2011) Inflammatory memories: is epigenetics the missing link to persistent stromal cell activation in rheumatoid arthritis? *Autoimmun Rev*, **10**, 519-524.
41. Kabesch, M., Michel, S. and Tost, J. (2010) Epigenetic mechanisms and the relationship to childhood asthma. *Eur Respir J*, **36**, 950-961.
42. Villeneuve, L.M., Reddy, M.A. and Natarajan, R. (2011) Epigenetics: deciphering its role in diabetes and its chronic complications. *Clin Exp Pharmacol Physiol*, **38**, 401-409.
43. Grayson, D.R. (2010) Schizophrenia and the epigenetic hypothesis. *Epigenomics*, **2**, 341-344.
44. de Fraipont, F. and Richard, M.J. (2009) L'hyperméthylation des gènes suppresseurs de tumeur comme marqueur en cancérologie. *Immuno-analyse & Biologie Spécialisée*, **24**, 9-15.
45. Jones, P.A. and Baylin, S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683-692.
46. Sharma, S., Kelly, T.K. and Jones, P.A. (2010) Epigenetics in cancer. *Carcinogenesis*, **31**, 27-36.
47. Sincic, N. and Herceg, Z. (2011) DNA methylation and cancer: ghosts and angels above the genes. *Curr Opin Oncol*, **23**, 69-76.

48. Jones, P.A. and Baylin, S.B. (2002) The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, **3**, 415-428.
49. Feinberg, A.P. and Vogelstein, B. (1983) Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, **301**, 89-92.
50. Gama-Sosa, M.A., Slagel, V.A., Trewyn, R.W., Oxenhandler, R., Kuo, K.C., Gehrke, C.W. and Ehrlich, M. (1983) The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res*, **11**, 6883-6894.
51. Herceg, Z. and Vaissiere, T. (2011) Epigenetic mechanisms and cancer: an interface between the environment and the genome. *Epigenetics*, **6**, 804-819.
52. Esteller, M. (2007) Epigenetic gene silencing in cancer: the DNA hypermethylome. *Hum Mol Genet*, **16 Spec No 1**, R50-59.
53. Jones, P.A. (2008) Moving AHEAD with an international human epigenome project. *Nature*, **454**, 711-715.
54. Mulero-Navarro, S. and Esteller, M. (2008) Epigenetic biomarkers for human cancer: the time is now. *Crit Rev Oncol Hematol*, **68**, 1-11.
55. Duffy, M.J., Napieralski, R., Martens, J.W., Span, P.N., Spyratos, F., Sweep, F.C., Brunner, N., Foekens, J.A. and Schmitt, M. (2009) Methylated genes as new cancer biomarkers. *Eur J Cancer*, **45**, 335-346.
56. Jovanovic, J., Ronneberg, J.A., Tost, J. and Kristensen, V. (2010) The epigenetics of breast cancer. *Mol Oncol*, **4**, 242-254.
57. Laird, P.W. (2003) The power and the promise of DNA methylation markers. *Nat Rev Cancer*, **3**, 253-266.
58. Lechner, M., Boshoff, C. and Beck, S. (2010) Cancer epigenome. *Adv Genet*, **70**, 247-276.
59. Esteller, M., Corn, P.G., Baylin, S.B. and Herman, J.G. (2001) A gene hypermethylation profile of human cancer. *Cancer Res*, **61**, 3225-3229.
60. Kulis, M. and Esteller, M. (2010) DNA methylation and cancer. *Adv Genet*, **70**, 27-56.
61. Zhu, J. and Yao, X. (2009) Use of DNA methylation for cancer detection: promises and challenges. *Int J Biochem Cell Biol*, **41**, 147-154.
62. Esteller, M., Garcia-Foncillas, J., Andion, E., Goodman, S.N., Hidalgo, O.F., Vanaclocha, V., Baylin, S.B. and Herman, J.G. (2000) Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N Engl J Med*, **343**, 1350-1354.
63. Yoo, C.B. and Jones, P.A. (2006) Epigenetic therapy of cancer: past, present and future. *Nat Rev Drug Discov*, **5**, 37-50.
64. Costa, F.F. (2010) Epigenomics in cancer management. *Cancer Manag Res*, **2**, 255-265.
65. Bell, J.T. and Saffery, R. (2012) The value of twins in epigenetic epidemiology. *Int J Epidemiol*.
66. Feil, R. and Fraga, M.F. (2012) Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet*.
67. Barker, D.J. (1997) Maternal nutrition, fetal nutrition, and disease in later life. *Nutrition*, **13**, 807-813.
68. Alegria-Torres, J.A., Baccarelli, A. and Bollati, V. (2011) Epigenetics and lifestyle. *Epigenomics*, **3**, 267-277.
69. Suter, M., Ma, J., Harris, A.S., Patterson, L., Brown, K.A., Shope, C., Showalter, L., Abramovici, A. and Aagaard-Tillery, K.M. (2011) Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. *Epigenetics*, **6**.
70. Wilhelm-Benartzi, C.S., Christensen, B.C., Koestler, D.C., Andres Houseman, E., Schned, A.R., Karagas, M.R., Kelsey, K.T. and Marsit, C.J. (2011) Association of secondhand smoke exposures with DNA methylation in bladder carcinomas. *Cancer Causes Control*, **22**, 1205-1213.
71. Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C. and Maleszka, R. (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol*, **8**, e1000506.

72. Dolinoy, D.C., Huang, D. and Jirtle, R.L. (2007) Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proc Natl Acad Sci U S A*, **104**, 13056-13061.
73. Jirtle, R.L. and Skinner, M.K. (2007) Environmental epigenomics and disease susceptibility. *Nat Rev Genet*, **8**, 253-262.
74. Schulz, L.C. (2010) The Dutch Hunger Winter and the developmental origins of health and disease. *Proc Natl Acad Sci U S A*, **107**, 16757-16758.
75. Lee, H.S. and Herceg, Z. (2012) The epigenome and cancer prevention: A complex story of dietary supplementation. *Cancer Lett*.
76. vel Szic, K.S., Ndlovu, M.N., Haegeman, G. and Vanden Berghe, W. (2010) Nature or nurture: let food be your epigenetic medicine in chronic inflammatory disorders. *Biochem Pharmacol*, **80**, 1816-1832.
77. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, **74**, 560-564.
78. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**, 5463-5467.
79. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.
80. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
81. Consortium, I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.
82. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol*, **5**, e254.
83. Ledford, H. (2007) All about Craig: the first 'full' genome sequence. *Nature*, **449**, 6-7.
84. Watson, J.D. and Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737-738.
85. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872-876.
86. Pushkarev, D., Neff, N.F. and Quake, S.R. (2009) Single-molecule sequencing of an individual human genome. *Nat Biotechnol*, **27**, 847-852.
87. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53-59.
88. Ahn, S.M., Kim, T.H., Lee, S., Kim, D., Ghang, H., Kim, D.S., Kim, B.C., Kim, S.Y., Kim, W.Y., Kim, C. *et al.* (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res*, **19**, 1622-1629.
89. Kim, J.I., Ju, Y.S., Park, H., Kim, S., Lee, S., Yi, J.H., Mudge, J., Miller, N.A., Hong, D., Bell, C.J. *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011-1015.
90. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60-65.
91. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66-72.
92. Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D. *et al.* (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med*, **361**, 1058-1066.

93. (2010) Human genome: Genomes by the thousand. *Nature*, **467**, 1026-1027.
94. Mardis, E.R. (2006) Anticipating the 1,000 dollar genome. *Genome Biol*, **7**, 112.
95. Sboner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K. and Gerstein, M.B. (2011) The real cost of sequencing: higher than you think! *Genome Biol*, **12**, 125.
96. Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, **11**, 759-769.
97. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.
98. Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.
99. Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M. *et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature*, **444**, 330-336.
100. Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Paabo, S., Pritchard, J.K. *et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science*, **314**, 1113-1118.
101. Wall, J.D. and Kim, S.K. (2007) Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genet*, **3**, 1862-1866.
102. Mellmann, A., Harmsen, D., Cummings, C.A., Zentz, E.B., Leopold, S.R., Rico, A., Prior, K., Szczepanowski, R., Ji, Y., Zhang, W. *et al.* (2011) Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One*, **6**, e22751.
103. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M. *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348-352.
104. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728-1732.
105. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78-81.
106. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X. *et al.* (2008) The potential and challenges of nanopore sequencing. *Nat Biotechnol*, **26**, 1146-1153.
107. Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*, **4**, 265-270.
108. Butler, T.Z., Pavlenok, M., Derrington, I.M., Niederweis, M. and Gundlach, J.H. (2008) Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc Natl Acad Sci U S A*, **105**, 20647-20652.
109. Min, S.K., Kim, W.Y., Cho, Y. and Kim, K.S. (2011) Fast DNA sequencing with a graphene-based nanochannel device. *Nat Nanotechnol*.
110. Hall, A.R., Scott, A., Rotem, D., Mehta, K.K., Bayley, H. and Dekker, C. (2010) Hybrid pore formation by directed insertion of alpha-haemolysin into solid-state nanopores. *Nat Nanotechnol*, **5**, 874-877.
111. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133-138.
112. Korlach, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., Holden, D., Saxena, R., Wegener, J. and Turner, S.W. (2010) In Nils, G. W. (ed.), *Methods in Enzymology*. Academic Press, Vol. Volume 472, pp. 431-455.

113. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J. and Turner, S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*, **7**, 461-465.
114. Song, C.X., Clark, T.A., Lu, X.Y., Kislyuk, A., Dai, Q., Turner, S.W., He, C. and Korlach, J. (2011) Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat Methods*, **9**, 75-77.
115. Pennisi, E. (2010) Genomics. Semiconductors inspire new sequencing technologies. *Science*, **327**, 1190.
116. Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, **11**, 685-696.
117. Lantos, J.D., Artman, M. and Kingsmore, S.F. (2011) Ethical considerations associated with clinical use of next-generation sequencing in children. *J Pediatr*, **159**, 879-880 e871.
118. Greenbaum, D., Sboner, A., Mu, X.J. and Gerstein, M. (2011) Genomics and privacy: implications of the new reality of closed data for the field. *PLoS Comput Biol*, **7**, e1002278.
119. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523-536.
120. Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M. and Jacobsen, S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215-219.
121. Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E.M. *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol*, **26**, 779-785.
122. Ruike, Y., Imanaka, Y., Sato, F., Shimizu, K. and Tsujimoto, G. (2010) Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*, **11**, 137.
123. Ho, K.L., McNae, I.W., Schmiedeberg, L., Klose, R.J., Bird, A.P. and Walkinshaw, M.D. (2008) MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol Cell*, **29**, 525-531.
124. Dhasarathy, A. and Wade, P.A. (2008) The MBD protein family-reading an epigenetic mark? *Mutat Res*, **647**, 39-43.
125. Klose, R.J., Sarraf, S.A., Schmiedeberg, L., McDermott, S.M., Stancheva, I. and Bird, A.P. (2005) DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol Cell*, **19**, 667-678.
126. Cross, S.H., Charlton, J.A., Nan, X. and Bird, A.P. (1994) Purification of CpG islands using a methylated DNA binding column. *Nat Genet*, **6**, 236-244.
127. Martens, J.H., Brinkman, A.B., Simmer, F., Francoijs, K.J., Nebbioso, A., Ferrara, F., Altucci, L. and Stunnenberg, H.G. (2010) PML-RARalpha/RXR Alters the Epigenetic Landscape in Acute Promyelocytic Leukemia. *Cancer Cell*, **17**, 173-185.
128. Lan, X., Adams, C., Landers, M., Dudas, M., Krissinger, D., Marnellos, G., Bonneville, R., Xu, M., Wang, J., Huang, T.H. *et al.* (2011) High Resolution Detection and Analysis of CpG Dinucleotides Methylation Using MBD-Seq Technology. *PLoS One*, **6**, e22226.
129. Brinkman, A.B., Simmer, F., Ma, K., Kaan, A., Zhu, J. and Stunnenberg, H.G. (2010) Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*, **52**, 232-236.
130. Yu, W., Jin, C., Lou, X., Han, X., Li, L., He, Y., Zhang, H., Ma, K., Zhu, J., Cheng, L. *et al.* (2011) Global analysis of DNA methylation by methyl-capture sequencing reveals epigenetic control of Cisplatin resistance in ovarian cancer cell. *PLoS One*, **6**, e29450.
131. Rauch, T. and Pfeifer, G.P. (2005) Methylated-CpG island recovery assay: a new technique for the rapid detection of methylated-CpG islands in cancer. *Lab Invest*, **85**, 1172-1180.
132. Rauch, T.A. and Pfeifer, G.P. (2009) The MIRA method for DNA methylation analysis. *Methods Mol Biol*, **507**, 65-75.
133. Pfeifer, G.P. and Rauch, T.A. (2009) DNA methylation patterns in lung carcinomas. *Semin Cancer Biol*, **19**, 181-187.

134. Thorne, N.P., Marioni, J.C., Rakyan, V., Ibrahim, A.E.K., Massie, C., Curtis, C., Brenton, J.D., Murrell, A. and Tavaré, S. (2009) In Hardiman, E. G. (ed.), *Microarray Innovations - Technology and Experimentation in Drug Discovery and Biomedical Research.*, pp. pp. 175-206.
135. Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L. and Schubeler, D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*, **37**, 853-862.
136. Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M. and Schubeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, **39**, 457-466.
137. Rakyan, V.K., Down, T.A., Thorne, N.P., Flicek, P., Kulesha, E., Graf, S., Tomazou, E.M., Backdahl, L., Johnson, N., Herberth, M. *et al.* (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res*, **18**, 1518-1529.
138. Pomraning, K.R., Smith, K.M. and Freitag, M. (2009) Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods*, **47**, 142-150.
139. Feber, A., Wilson, G.A., Zhang, L., Presneau, N., Idowu, B., Down, T.A., Rakyan, V.K., Noon, L.A., Lloyd, A.C., Stupka, E. *et al.* (2011) Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. *Genome Res*, **21**, 515-524.
140. Estecio, M.R., Yan, P.S., Ibrahim, A.E., Tellez, C.S., Shen, L., Huang, T.H. and Issa, J.P. (2007) High-throughput methylation profiling by MCA coupled to CpG island microarray. *Genome Res*, **17**, 1529-1536.
141. Khulan, B., Thompson, R.F., Ye, K., Fazzari, M.J., Suzuki, M., Stasiek, E., Figueroa, M.E., Glass, J.L., Chen, Q., Montagna, C. *et al.* (2006) Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res*, **16**, 1046-1055.
142. Oda, M., Glass, J.L., Thompson, R.F., Mo, Y., Olivier, E.N., Figueroa, M.E., Selzer, R.R., Richmond, T.A., Zhang, X., Dannenberg, L. *et al.* (2009) High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers. *Nucleic Acids Res*, **37**, 3829-3839.
143. Brunner, A.L., Johnson, D.S., Kim, S.W., Valouev, A., Reddy, T.E., Neff, N.F., Anton, E., Medina, C., Nguyen, L., Chiao, E. *et al.* (2009) Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res*, **19**, 1044-1056.
144. 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.
145. Berman, B.P., Weisenberger, D.J. and Laird, P.W. (2009) Locking in on the human methylome. *Nat Biotechnol*, **27**, 341-342.
146. 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.
147. Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L., Johnson, B.E., Fouse, S.D., Delaney, A., Zhao, Y. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol*, **28**, 1097-1105.
148. Li, J., Gao, F., Li, N., Li, S.T., Yin, G.L., Tian, G., Jia, S.G., Wang, K., Zhang, X.Q., Yang, H.M. *et al.* (2009) An improved method for genome wide DNA methylation profiling correlated to transcription and genomic instability in two breast cancer cell lines. *Bmc Genomics*, **10**.
149. Hu, M., Yao, J. and Polyak, K. (2006) Methylation-specific digital karyotyping. *Nat Protoc*, **1**, 1621-1636.

150. Edwards, J.R., O'Donnell, A.H., Rollins, R.A., Peckham, H.E., Lee, C., Milekic, M.H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H. *et al.* (2010) Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res*, **20**, 972-980.
151. Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B. and Feinberg, A.P. (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res*, **18**, 780-790.
152. Wang, R.Y., Gehrke, C.W. and Ehrlich, M. (1980) Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res*, **8**, 4777-4790.
153. Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A*, **89**, 1827-1831.
154. Clark, S.J., Harrison, J., Paul, C.L. and Frommer, M. (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res*, **22**, 2990-2997.
155. Chen, P.Y., Cokus, S.J. and Pellegrini, M. (2010) BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, **11**, 203.
156. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766-770.
157. Smith, Z.D., Gu, H., Bock, C., Gnirke, A. and Meissner, A. (2009) High-throughput bisulfite sequencing in mammalian genomes. *Methods*, **48**, 226-232.
158. Shearstone, J.R., Pop, R., Bock, C., Boyle, P., Meissner, A. and Socolovsky, M. (2011) Global DNA demethylation during mouse erythropoiesis in vivo. *Science*, **334**, 799-802.
159. Li, N., Ye, M., Li, Y., Yan, Z., Butcher, L.M., Sun, J., Han, X., Chen, Q., Zhang, X. and Wang, J. (2010) Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods*, **52**, 203-212.
160. Bock, C., Tomazou, E.M., Brinkman, A.B., Muller, F., Simmer, F., Gu, H., Jager, N., Gnirke, A., Stunnenberg, H.G. and Meissner, A. (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat Biotechnol*, **28**, 1106-1114.
161. Jin, S.G., Kadam, S. and Pfeifer, G.P. (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res*, **38**, e125.
162. Meyer, M., Stenzel, U. and Hofreiter, M. (2008) Parallel tagged sequencing on the 454 platform. *Nat Protoc*, **3**, 267-278.
163. De Leeneer, K., De Schrijver, J., Clement, L., Baetens, M., Lefever, S., De Keulenaer, S., Van Criekinge, W., Deforce, D., Van Nieuwerburgh, F., Bekaert, S. *et al.* (2011) Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. *PLoS One*, **6**, e25531.
164. Bybee, S.M., Bracken-Grissom, H., Haynes, B.D., Hermansen, R.A., Byers, R.L., Clement, M.J., Udall, J.A., Wilcox, E.R. and Crandall, K.A. (2011) Targeted amplicon sequencing (TAS): A scalable next-gen approach to multi-locus, multi-taxa phylogenetics. *Genome Biol Evol*.
165. Taylor, K.H., Kramer, R.S., Davis, J.W., Guo, J., Duff, D.J., Xu, D., Caldwell, C.W. and Shi, H. (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res*, **67**, 8511-8518.
166. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, **27**, 182-189.
167. 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.
168. Nautiyal, S., Carlton, V.E., Lu, Y., Ireland, J.S., Flaucher, D., Moorhead, M., Gray, J.W., Spellman, P., Mindrinos, M., Berg, P. *et al.* (2010) High-throughput method for analyzing

methylation of CpGs in targeted genomic regions. *Proc Natl Acad Sci U S A*, **107**, 12587-12592.

169. Varley, K.E. and Mitra, R.D. (2010) Bisulfite Patch PCR enables multiplexed sequencing of promoter methylation across cancer samples. *Genome Res*, **20**, 1279-1287.
170. 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.
171. Nilsson, M. (2006) Lock and roll: single-molecule genotyping in situ using padlock probes and rolling-circle amplification. *Histochem Cell Biol*, **126**, 159-164.
172. Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B.P. and Landegren, U. (1994) Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science*, **265**, 2085-2088.
173. Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nat Methods*, **4**, 931-936.
174. Shen, P., Wang, W., Krishnakumar, S., Palm, C., Chi, A.K., Enns, G.M., Davis, R.W., Speed, T.P., Mindrinos, M.N. and Scharfe, C. (2011) High-quality DNA sequence capture of 524 disease candidate genes. *Proc Natl Acad Sci U S A*, **108**, 6549-6554.
175. Li, J.B., Gao, Y., Aach, J., Zhang, K., Kryukov, G.V., Xie, B., Ahlford, A., Yoon, J.K., Rosenbaum, A.M., Zaranek, A.W. *et al.* (2009) Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res*, **19**, 1606-1615.
176. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. and Nilsson, M. (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res*, **33**, e71.
177. Natsoulis, G., Bell, J.M., Xu, H., Buenrostro, J.D., Ordonez, H., Grimes, S., Newburger, D., Jensen, M., Zahn, J.M., Zhang, N. *et al.* (2011) A flexible approach for highly multiplexed candidate gene targeted resequencing. *PLoS One*, **6**, e21088.
178. Johansson, H., Isaksson, M., Sorqvist, E.F., Roos, F., Stenberg, J., Sjoblom, T., Botling, J., Micke, P., Edlund, K., Fredriksson, S. *et al.* (2011) Targeted resequencing of candidate genes using selector probes. *Nucleic Acids Res*, **39**, e8.
179. Deng, J., Shoemaker, R., Xie, B., Gore, A., LeProust, E.M., Antosiewicz-Bourget, J., Egli, D., Maherali, N., Park, I.H., Yu, J. *et al.* (2009) Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol*, **27**, 353-360.
180. Diep, D., Plongthongkum, N., Gore, A., Fung, H.L., Shoemaker, R. and Zhang, K. (2012) Library-free methylation sequencing with bisulfite padlock probes. *Nat Methods*.
181. Richard, G.F., Kerrest, A. and Dujon, B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol Rev*, **72**, 686-727.
182. Jurka, J., Kapitonov, V.V., Kohany, O. and Jurka, M.V. (2007) Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet*, **8**, 241-259.
183. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520-562.
184. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, **16**, 418-420.
185. Patterson, K., Molloy, L., Qu, W. and Clark, S. (2011) DNA methylation: bisulphite modification and analysis. *J Vis Exp*.
186. Campan, M., Weisenberger, D.J., Trinh, B. and Laird, P.W. (2009) MethyLight. *Methods Mol Biol*, **507**, 325-337.
187. Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res*, **40**, D918-923.

188. Li, L.C. and Dahiya, R. (2002) MethPrimer: designing primers for methylation PCRs. *Bioinformatics*, **18**, 1427-1431.
189. Dupont, J.M., Tost, J., Jammes, H. and Gut, I.G. (2004) De novo quantitative bisulfite sequencing using the pyrosequencing technology. *Anal Biochem*, **333**, 119-127.
190. Tost, J. and Gut, I.G. (2007) DNA methylation analysis by pyrosequencing. *Nat Protoc*, **2**, 2265-2275.
191. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589-595.
192. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-1760.
193. Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R. and Adjaye, J. (2010) Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res*, **20**, 1441-1450.
194. Palmke, N., Santacruz, D. and Walter, J. (2011) Comprehensive analysis of DNA-methylation in mammalian tissues using MeDIP-chip. *Methods*, **53**, 175-184.
195. Lutsik, P., Feuerbach, L., Arand, J., Lengauer, T., Walter, J. and Bock, C. (2011) BiQ Analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. *Nucleic Acids Res*.
196. Chiu, R.W., Chim, S.S., Wong, I.H., Wong, C.S., Lee, W.S., To, K.F., Tong, J.H., Yuen, R.K., Shum, A.S., Chan, J.K. *et al.* (2007) Hypermethylation of RASSF1A in human and rhesus placentas. *Am J Pathol*, **170**, 941-950.
197. Monk, D., Sanches, R., Arnaud, P., Apostolidou, S., Hills, F.A., Abu-Amero, S., Murrell, A., Friess, H., Reik, W., Stanier, P. *et al.* (2006) Imprinting of IGF2 P0 transcript and novel alternatively spliced INS-IGF2 isoforms show differences between mouse and human. *Hum Mol Genet*, **15**, 1259-1269.
198. Sengenès, J., Daunay, A., Charles, M.A. and Tost, J. (2010) Quality control and single nucleotide resolution analysis of methylated DNA immunoprecipitation products. *Anal Biochem*, **407**, 141-143.
199. Althammer, S., Gonzalez-Vallinas, J., Ballare, C., Beato, M. and Eyras, E. (2011) Pyicos: A versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics*.
200. Carver, T., Harris, S.R., Berriman, M., Parkhill, J. and McQuillan, J.A. (2011) Artemis: An integrated platform for visualisation and analysis of high-throughput sequence-based experimental data. *Bioinformatics*.
201. Brouwer, R.W., van den Hout, M.C., Grosveld, F.G. and van Ijcken, W.F. (2011) NARWHAL, a primary analysis pipeline for NGS data. *Bioinformatics*.
202. Halbritter, F., Vaidya, H.J. and Tomlinson, S.R. (2011) GeneProf: analysis of high-throughput sequencing experiments. *Nat Methods*, **9**, 7-8.
203. Lajugie, J. and Bouhassira, E. (2011) GenPlay, a multi-purpose genome analyzer and browser. *Bioinformatics*.
204. Huang, J., Renault, V., Sengenes, J., Touleimat, N., Michel, S., Lathrop, M. and Tost, J. (2012) MeQA: a pipeline for MeDIP-seq data quality assessment and analysis. *Bioinformatics*, **28**, 587-588.
205. Vining, K.J., Pomraning, K.R., Wilhelm, L.J., Priest, H.D., Pellegrini, M., Mockler, T.C., Freitag, M. and Strauss, S. (2012) Dynamic DNA cytosine methylation in the Populus trichocarpa genome: tissue-level variation and relationship to gene expression. *BMC Genomics*, **13**, 27.
206. Brebi-Mieville, P., Ili-Gangas, C., Leal-Rojas, P., Noordhuis, M.G., Soudry, E., Perez, J., Roa, J.C., Sidransky, D. and Guerrero-Preston, R. (2012) Clinical and public health research using methylated DNA Immunoprecipitation (MeDIP): A comparison of commercially available kits to examine differential DNA methylation across the genome. *Epigenetics*, **7**.
207. Butcher, L.M. and Beck, S. (2010) AutoMeDIP-seq: a high-throughput, whole genome, DNA methylation assay. *Methods*, **52**, 223-231.

208. Lennon, N.J., Lintner, R.E., Anderson, S., Alvarez, P., Barry, A., Brockman, W., Daza, R., Erlich, R.L., Giannoukos, G., Green, L. *et al.* (2010) A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biol*, **11**, R15.
209. Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J. and Chisholm, S.W. (2010) Unlocking short read sequencing for metagenomics. *PLoS One*, **5**, e11840.
210. Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D. and Lundeberg, J. (2010) Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One*, **5**, e10029.
211. Hawkins, T.L., O'Connor-Morin, T., Roy, A. and Santillan, C. (1994) DNA purification and isolation using a solid-phase. *Nucleic Acids Res*, **22**, 4543-4544.
212. Martens, J.H., O'Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P. and Jenuwein, T. (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *Embo J*, **24**, 800-812.
213. Holmberg, A., Blomstergren, A., Nord, O., Lukacs, M., Lundeberg, J. and Uhlen, M. (2005) The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis*, **26**, 501-510.
214. Craig, J.M., Kraus, J. and Cremer, T. (1997) Removal of repetitive sequences from FISH probes using PCR-assisted affinity chromatography. *Hum Genet*, **100**, 472-476.
215. Lucas, J.N., Wu, X., Guo, E., Chi, L.E. and Chen, Z. (2006) An efficient chemical method to generate repetitive sequences depleted DNA probes. *Am J Med Genet A*, **140**, 2115-2120.
216. Rigby, P.W., Dieckmann, M., Rhodes, C. and Berg, P. (1977) Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *J Mol Biol*, **113**, 237-251.
217. Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, **12**, R18.
218. Quail, M.A., Otto, T.D., Gu, Y., Harris, S.R., Skelly, T.F., McQuillan, J.A., Swerdlow, H.P. and Oyola, S.O. (2011) Optimal enzymes for amplifying sequencing libraries. *Nat Methods*, **9**, 10-11.
219. Kircher, M., Heyn, P. and Kelso, J. (2011) Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics*, **12**, 382.
220. Kimpton, C.P., Gill, P., Walton, A., Urquhart, A., Millican, E.S. and Adams, M. (1993) Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods Appl*, **3**, 13-22.
221. Urquhart, A., Kimpton, C.P., Downes, T.J. and Gill, P. (1994) Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med*, **107**, 13-20.
222. Economou, E.P., Bergen, A.W., Warren, A.C. and Antonarakis, S.E. (1990) The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc Natl Acad Sci U S A*, **87**, 2951-2954.
223. Potter, S.S. (1984) Rearranged sequences of a human Kpn I element. *Proc Natl Acad Sci U S A*, **81**, 1012-1016.
224. Bertolino, P., Radovanovic, I., Casse, H., Aguzzi, A., Wang, Z.Q. and Zhang, C.X. (2003) Genetic ablation of the tumor suppressor menin causes lethality at mid-gestation with defects in multiple organs. *Mech Dev*, **120**, 549-560.
225. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R. and Gunderson, K.L. (2009) Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics*, **1**, 177-200.
226. Goransson, J., Wahlby, C., Isaksson, M., Howell, W.M., Jarvius, J. and Nilsson, M. (2009) A single molecule array for digital targeted molecular analyses. *Nucleic Acids Res*, **37**, e7.

227. Stenberg, J., Dahl, F., Landegren, U. and Nilsson, M. (2005) PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. *Nucleic Acids Res*, **33**, e72.
228. Newburger, D.E., Natsoulis, G., Grimes, S., Bell, J.M., Davis, R.W., Batzoglou, S. and Ji, H.P. (2011) The Human OligoGenome Resource: a database of oligonucleotide capture probes for resequencing target regions across the human genome. *Nucleic Acids Res*.
229. Lovmar, L. and Syvanen, A.C. (2006) Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. *Hum Mutat*, **27**, 603-614.
230. Warnecke, P.M., Stirzaker, C., Song, J., Grunau, C., Melki, J.R. and Clark, S.J. (2002) Identification and resolution of artifacts in bisulfite sequencing. *Methods*, **27**, 101-107.
231. Krueger, F. and Andrews, S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571-1572.
232. Krueger, F., Kreck, B., Franke, A. and Andrews, S.R. (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat Methods*, **9**, 145-151.
233. Mardis, E.R. (2011) A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198-203.
234. Kahn, S.D. (2011) On the future of genomic data. *Science*, **331**, 728-729.
235. Gupta, R., Nagarajan, A. and Wajapeyee, N. (2010) Advances in genome-wide DNA methylation analysis. *Biotechniques*, **49**, iii-xi.
236. Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C. and Fuks, F. (2011) Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, **3**, 771-784.
237. Hurd, P.J. and Nelson, C.J. (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomic Proteomic*.
238. Nair, S.S., Coolen, M.W., Stirzaker, C., Song, J.Z., Statham, A.L., Strbenac, D., Robinson, M.W. and Clark, S.J. (2011) Comparison of methyl-DNA immunoprecipitation (MeDIP) and methyl-CpG binding domain (MBD) protein capture for genome-wide DNA methylation analysis reveal CpG sequence coverage bias. *Epigenetics*, **6**, 34-44.
239. Ndlovu, M.N., Denis, H. and Fuks, F. (2011) Exposing the DNA methylome iceberg. *Trends in Biochemical Sciences*, **36**, 381-387.
240. 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.
241. Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P. *et al.* (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, **473**, 394-397.
242. Szwagierczak, A., Brachmann, A., Schmidt, C.S., Bultmann, S., Leonhardt, H. and Spada, F. (2011) Characterization of PvuRts1I endonuclease as a tool to investigate genomic 5-hydroxymethylcytosine. *Nucleic Acids Res*, **39**, 5149-5156.
243. Wang, H., Guan, S., Quimby, A., Cohen-Karni, D., Pradhan, S., Wilson, G., Roberts, R.J., Zhu, Z. and Zheng, Y. (2011) Comparative characterization of the PvuRts1I family of restriction enzymes and their application in mapping genomic 5-hydroxymethylcytosine. *Nucleic Acids Res*, **39**, 9294-9305.
244. Ehrlich, M. (2009) DNA hypomethylation in cancer cells. *Epigenomics*, **1**, 239-259.
245. Wilson, A.S., Power, B.E. and Molloy, P.L. (2007) DNA hypomethylation and human diseases. *Biochim Biophys Acta*, **1775**, 138-162.
246. Weisenberger, D.J., Campan, M., Long, T.I., Kim, M., Woods, C., Fiala, E., Ehrlich, M. and Laird, P.W. (2005) Analysis of repetitive element DNA methylation by MethyLight. *Nucleic Acids Res*, **33**, 6823-6836.
247. Ting, D.T., Lipson, D., Paul, S., Brannigan, B.W., Akhavanfard, S., Coffman, E.J., Contino, G., Deshpande, V., Iafrate, A.J., Letovsky, S. *et al.* (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. *Science*, **331**, 593-596.
248. Neguembor, M.V. and Gabellini, D. (2010) In junk we trust: repetitive DNA, epigenetics and facioscapulohumeral muscular dystrophy. *Epigenomics*, **2**, 271-287.

249. Frumkin, D., Wasserstrom, A., Davidson, A. and Grafit, A. (2010) Authentication of forensic DNA samples. *Forensic Sci Int Genet*, **4**, 95-103.
250. Lee, H.Y., Park, M.J., Choi, A., An, J.H., Yang, W.I. and Shin, K.J. (2012) Potential forensic application of DNA methylation profiling to body fluid identification. *Int J Legal Med*, **126**, 55-62.
251. Zhao, G., Yang, Q., Huang, D., Yu, C., Yang, R., Chen, H. and Mei, K. (2005) Study on the application of parent-of-origin specific DNA methylation markers to forensic genetics. *Forensic Sci Int*, **154**, 122-127.
252. Nygren, A.O., Dean, J., Jensen, T.J., Kruse, S., Kwong, W., van den Boom, D. and Ehrich, M. (2010) Quantification of fetal DNA by use of methylation-based DNA discrimination. *Clin Chem*, **56**, 1627-1635.

Développement de méthodes de séquençage de seconde génération pour l'analyse des profils de méthylation de l'ADN

Résumé

L'analyse des profils de méthylation présente un grand intérêt car des altérations du méthylome sont impliquées dans de nombreuses pathologies. Le MeDIP (Methylated DNA ImmunoPrecipitation) immunoprécipite les séquences méthylées sur le génome entier, la plupart étant localisées dans les séquences répétées. De telles séquences sont difficiles à aligner après séquençage (MeDIP-Seq) et bon nombre d'entre elles ne peuvent donc être utilisées pour la suite des analyses. Nous présentons une méthode innovante appelée MeDIP-dep-Seq permettant de supprimer une quantité significative de plusieurs familles de ces éléments répétés (diminution d'un facteur de 300 au maximum) tandis que les séquences uniques d'intérêt ne sont pas affectées. Après séquençage sur un séquenceur de seconde génération (GAIIx, Illumina), le taux d'alignement est amélioré de façon conséquente permettant ainsi d'augmenter la quantité de séquences analysables. Nous avons également développé une plateforme d'analyse des données issues du MeDIP-Seq.
De potentielles régions candidates identifiées par cette technique sur le génome entier peuvent ensuite être validées en utilisant des sélectors, sondes permettant la capture de régions génomiques d'intérêt. Nous avons introduit un traitement au bisulphite dans le protocole de sélection afin de développer un nouvel outil pour une analyse multiplexe. 98 loci ont été enrichis dans 6 échantillons puis séquencés en parallèle sur un séquenceur de paillasse (GS Junior, Roche).
La combinaison de ces technologies permettra d'établir des cartes du méthylome et d'identifier des nouveaux biomarqueurs épigénétiques pour diagnostiquer et pronostiquer les cancers et maladies complexes.

Mots-clés : méthylation de l'ADN, séquençage, MeDIP-Seq, éléments répétés, déplétion, sélectors

Development of second generation sequencing technologies for the analysis of DNA methylation signatures

Summary

The analysis of DNA methylation patterns has become of great interest as methylome alterations have been found in many diseases. MeDIP (Methylated DNA ImmunoPrecipitation) immunoprecipitates genome-wide methylated sequences many of which are located in the repetitive sequences. Such sequences are difficult to align unambiguously after sequencing (MeDIP-Seq) leading to a large number of sequences that are currently not used for further analysis. We present an innovative method called MeDIP-dep-Seq which depletes a significant part of several classes of these highly repetitive sequences (up to 300-fold decrease), while unique sequences of interest are not affected. After sequencing on a second generation sequencer (GAIIx, Illumina) the alignment rate substantially enhanced increasing thus the amount of usable sequences. We have further developed a pipeline for the analysi of MeDIP-Seq datasets.
Potential candidate regions identified in this genome-wide assay can then be validated by the use of selector probes that specifically capture genomic regions of interest. We introduced a bisulfite treatment in the selection protocol and developed a novel multiplex assay. 98 gene loci were enriched in 6 samples and were then sequenced in parallel on a bench sequencer (GS Junior, Roche).
The combination of these technologies will permit the establishment of methylome maps and the identification of novel epigenetic biomarkers for cancer and complex diseases diagnostics and prognosis.

Keywords : DNA methylation, sequencing, MeDIP-Seq, repetitive elements, depletion, selectors