Expression tissulaire des gènes paralogues : application au cerveau humain et à son état pathologique

Solène Julien

université PARIS-SACLAY

NNT : 2017SACLS545

Thèse de doctorat

université PARIS-SACLAY

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Expression tissulaire des gènes paralogues: application au cerveau humain et à son état pathologique

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°575 Physique et Ingénierie : élec trons, photons, sciences du vivant (EOBE)
Spécialité de doctorat: Imagerie et physique médicale

Thèse présentée et soutenue à Evry, le 19 Décembre 2017, par

Solène Julien

Composition du Jury :

Christophe Ambroise Professeur, Université d'Evry Val d'Essonne (UMR 8071)	Président
Marc Robinson-Rechavi Professeur, Université de Lausanne	Rapporteur
Hervé Isambert Chef d'équipe de recherche, Institut Curie	Rapporteur
Claudine Landés Professeur, Université d'Angers	Examinateur
Jean-François Deleuze Directeur de recherche, CEA/CNRGH, CEPH	Examinateur
Edith Le Floch CEA-E3, CEA/CNRGH	Encadrant
Christophe Battail CEA-E4, CEA/Institut de Biosciences et Biotechnologies de Grenoble	Encadrant
Vincent Frouin CEA-E6, CEA/Neurospin (UNATI)	Directeur de thèse

Remerciements

Mes premiers remerciements vont à mes encadrants de thèse. Je souhaite tout d'abord remercier Christophe Battail qui m'a proposé le sujet de thèse, sans qui le projet n'aurait pas vu le jour, qui a pu gérer pleinement un encadrement à distance et qui m'a prise en stage de M2. Je le remercie de sa patience, de sa franchise et de son optimisme. Un grand merci à Edith Le Floch qui m'a vraiment soutenu sur tous les sujets et qui a tout de suite accepté l'encadrement. Je la remercie pour la gentillesse dont elle fait preuve en permanence ainsi que de son aide pour toutes les épreuves rencontrées au cours de ces années de thèse. Je l'ai embêtée à chaque moment de déprime mais aussi à chaque moment de joie rencontré pendant la thèse et dans le privé. Je tiens particulièrement à remercier mon directeur de thèse, Vincent Frouin, qui a accepté la direction de la thèse malgré un sujet original, qui m'a soutenu et qui m'a beaucoup préservé même dans les moments les plus difficiles. Je les remercie tous les trois pour ces, un peu plus de 3 ans de doctorat. Je suis ravie de mon encadrement de thèse, je n'aurais pas espéré mieux. Je tiens vraiment à vous remercier pour les moments de rédaction et corrections très intenses que ce soit pour l'article ou bien pour la thèse, je suis désolée de vous avoir fait travaillé pendant vos week-ends et même vos vacances mais je vous en remercie, je me suis vraiment sentie soutenu pendant cette thèse et c'est grâce à vous. Je remercie également Jean-François Deleuze qui m'a acceptée dans son laboratoire dès le M2 et qui m'a permis d'être présente quotidiennement au CNRGH afin d'être entourée d'une très bonne équipe de bioinformaticiens. Je le remercie de m'avoir donné l'occasion de présenter mes travaux à des conférences nationales et internationales (JOBIM et ESHG). De plus, je le remercie d'avoir accepté d'être examinateur pour cette soutenance.

Je voudrais ensuite remercier mes rapporteurs, Marc Robinson-Rechavi et Hervé Isambert qui ont accepté de lire et d'évaluer mon manuscrit et aussi d'avoir fait le déplacement jusqu'à Evry pour ma soutenance de thèse. Merci à Christophe Ambroise d'avoir accepté gentiment d'examiner mon manuscrit de thèse mais également de présider mon jury de soutenance. Un remerciement tout particulier va à Claudine Landés qui a accepté d'être examinatrice pour ma soutenance et qui fait partie des personnes sans qui sans leur confiance au départ je n'aurais jamais pu faire de bioinformatique. Merci à tous les membres du jury de vous être libéré afin d'assister à

ma thèse et d'avoir lu mon manuscrit de thèse ainsi que de l'intérêt que vous avez porté à ce sujet de thèse que j'ai ressenti lors de la soutenance.

Pour ces trois années, je tiens à remercier Francois Artiguenave qui m'a accueillie au sein de son équipe de bioinformatique et qui m'a soutenu au début de ma thèse et Vincent Meyer pour son soutien pour la fin de ma thèse et pour son aide pour l'article.
Je remercie Carène Rizzon et Margot Corréa qui m'ont beaucoup apporté pour le lancement du projet et pour la compréhension du vaste domaine de l'évolution. Merci également à Olivier Jaillon pour les discussions autour de mon projet et pour son point de vue extérieur qui nous a aidés notamment pour l'écriture du papier.
Un grand merci à Smahane, ma collègue et amie, pour son aide tout au long de ma thèse, qui a partagé mon bureau, qui m'a réellement soutenu jusqu'au bout et fait part de ses expériences passées. Merci pour toutes les discussions que nous avons eu, de ta gentillesse et du petit texto quand je n'arrivais pas à l'heure de d'habitude de peur qu'il m'arrive quelque chose sur la route… Merci à Florian pour ces critiques très constructives sur les formats de présentation et sur l'esthétique de mes flèches lors des pauses touillettes. Merci à Lilia pour les petits cafés tôt le matin. Je remercie également Morgane pour les petites pauses de la journée et aussi pour m'avoir fait découvrir les bienfaits de l'athlétisme pour le travail. J'en profite pour remercier Christian Rébollo pour ces 4 mois intensifs pour une débutante de course à pied. Merci à Jean-Baptiste, pour son soutien amical au travail pendant une année. Merci à Aurélie pour les quelques discussions dans son bureau nous permettant d'évoquer les vacances. Un grand merci à Isabelle, pour son soutien moral et qui a même pris soin de ma santé en me poussant à faire du sport et en m'apportant de la vitamine C. Je tiens également à la remercier ainsi que Diana et Maryline qui ont tout fait pour que mes missions se passent sans encombre ainsi que pour l'organisation de ma soutenance. Et bien sûr je remercie toute l'équipe du LBI, Eric, Nizar, Claire, Xavier, David, Cédric, Steven, Cham's, Elise, Olivier , Aurélie, Damien, Dmitri, Georges et Sarah et l'équipe statistique Arthur, Mathilde et Claire, qui ont fait de mes pauses et de mes midis des vrais bon moments dans la journée avec des discussions variées sur la nourriture, les jeux vidéo, la nourriture, la musique et la nourriture...
Je remercie également Olivier et Nicolas présents en toutes circonstances ainsi que l'équipe de Delphine. Je remercie Marie-Thérèse Corey, Anne-Marie Brenot et Florence

Petellat pour leur aide concernant les problèmes administratifs rencontrés lors de mon contrat. Je remercie tous les membres du CNRGH qui m'ont aidé et avec qui j'ai eu l'occasion de travailler (Robert, Steven, Elizabeth, Marie-Ange, Céline, Sophie, Masazumi, Anne, Jorg, Florence) et tous les membres du CNRGH qui m'ont soutenu et qui ont été présents ou qui auraient aimé être présents à ma soutenance.

Je remercie aussi certains professeurs que j'ai eu l'occasion de croiser pendant mon cursus, Valérie Chaudru qui s'est battue pour que je puisse venir à Evry en master 1 et qui s'est battue pour me permettre de faire mes enseignements à Evry en thèse et pour tous ses conseils dans les couloirs du CNRGH. Je remercie également Mickael Falconnet, Cyril Dalmasso et Nathalie Boudet pour leur formation en bioinformatique et en biostatistques.

D'un point de vue plus personnel, je tiens à remercier ma famille, ma sœur ainée, Bénédicte et ma maman pour leur « boost » par téléphone dans les moments très difficile et mon père, mon frère, Florian et ma sœur Léa pour leur soutien. Je remercie également mes beaux-parents et beau-frère Jean-Luc, Dominique, Isabelle, Joëlle, Michel et Renaud qui m'ont également soutenu et qui ont compris mes heures à travailler chez eux. Merci à mes petits neveux, Axel, Maxence et Eloi pour leur joie de vivre.
Je remercie aussi mes amis GBI d'origines ou devenus, Mandy, Jocelyn, Jean-Baptiste, Kelly, Anaïs, Christophe, Vimel, Romain et Solène qui me soutiennent depuis le master. Merci à la famille Rameau, qui m'a fait oublier les moments difficiles de la thèse en leur compagnie. Merci à Noémie pour son soutien depuis la maternelle jusqu'à la thèse et qui m'a permis de positiver en toutes circonstances. Merci également à Christine et Caroline qui ne m'ont pas délaissées pendent cette période avec leur petits messages réconfortants. Je remercie également mes amis de la danse, Lydia, Monique, Aurore, Christine, Laurence et Yohan qui ont subi ma fatigue du mardi soir et qui ont vécu toutes les étapes de la thèse par mes humeurs différentes aux cours de danse.
Et bien évidemment je remercie sincèrement mon mari, Mathieu, qui m'a soutenu et supporté tout au long de la thèse, qui m'a encouragé et sans qui je n'aurai j'aimais fait de thèse.

Abréviations

2R : « 2 Rounds » - 2 évènements de WGD

ADN : Acide désoxyribonucléique

ADNc : ADN cyclique

AED : « Asymetrically Expressed Duplicates »

ARI : « Adjusted Rand Index »

ARN : Acide ribonucléique

ARNm : ARN messager

ASD : « Autism spectrum disorder » - Trouble du spectre de l'autisme

BLAST : « Basic Local Alignment Search Tool » - Outil de recherche d'alignement local

CNV : « Copy number variation » - Variation du nombre de copies

DDC : Duplication-Dégénération-Complémentation

DLI : « Duplication/Loss Inference » - Inférence de duplication et de perte

dNTP : désoxy-nucléotide-tri-phosphate

eQTL : « expression Quantitative trait loci »

GO : « Gene ontology » - Ontologie des gènes

GRC: « Genome Reference Consortium »

GTEx: « Genotype-Tissue Expression »

HGNC: « Human Genome Nomenclature Committee »

HSP : « Homologous sequence pairs » - Paires de séquence homologues

IRM : Imagerie à résonnance magnétique

OR : Odds ratio

oSSD : « older SSD »- SSD avant la période de WGD

pb : paire de bases

RIN : « RNA Integrity Number »

RPKM: « Reads Per Kilobase per Million mapped reads »

RT-PCR : « Reverse transcriptase polymerase chain reaction »

SCZ : Schizophrénie

SDI : « Speciation vs Duplication Inference »

SNP : « Single nucleotide polymorphism » - Polymorphisme d'un seul nucléotide

SNV : « Single nucleotide variation » - Variation d'un seul nucléotide

SSD : « Small-scale duplication » - Duplication à petite échelle

TDC : Dispositif à transfert de charge

TOM : « Topological overlap matrix » - Matrice de chevauchement toplogique

TPM: « Transcript Per Million »- Transcrits par millions

UPGMA: « Unweighted Pair Group Method with Arithmetic Mean »- Méthode non pondérée de regroupement de paires par moyenne arithmétique

WGCNA : « Weighted Gene Correlation Network Analysis » - Analyse de réseaux pondéré de corrélation de gènes

WGD : « Whole genome duplication » - Duplication de génome entier

wSSD : « WGD-SSD old » - SSD datant de la période de WGD

ySSD : « younger SSD » - SSD après la période de WGD

Table des matières

Table des figures

Liste des tables

Chapitre 1 : Introduction générale

1.Gènes paralogues

1.1.Définitions sur la paralogie

1.1.1.Paralogie et orthologie

Les arbres phylogénétiques :

L'évolution des espèces peut être représentée sous la forme d'un arbre phylogénétique (Figure 1) (Amemiya et al. 2013). Cet arbre peut être obtenu à partir d'alignements de séquences multiples entre toutes les espèces. Un arbre d'espèces peut être basé sur des données moléculaires correspondant à leur séquence ADN ou protéique. L'arbre phylogénétique peut être construit soit à partir des distances entre les espèces basées sur les alignements des séquences entre espèces, soit avec des méthodes basées sur des modèles d'évolution. Sur la Figure 2 l'arbre S est un schéma d'un arbre d'espèces. Le haut de l'arbre au niveau du « S » correspond à la racine de l'arbre. Les espèces en elles-mêmes correspondent aux feuilles de l'arbre. Les nœuds des arbres des espèces correspondent au dernier ancêtre commun des deux espèces descendant de cet ancêtre. A partir d'un arbre d'espèces, on peut définir un arbre de gènes. Sur la Figure 2 l'arbre G correspond au schéma d'un arbre de gènes. Sur ce schéma est représentée la phylogénie du gène ancestral G donnant des gènes homologues dans différentes espèces. Il existe un arbre de gènes vrai, c'est-à-dire équivalent à l'arbre d'espèces, mais il reste inconnu. La méthode de reconstruction de l'arbre utilisée engendre des erreurs d'estimations. L'arbre de gènes peut différer car sont projetés sur ce dernier les évènements de duplications en plus des évènements de spéciation. Ainsi deux gènes appartenant à la même espèce sont représentés. De plus, dans le cas d'une duplication suivie d'un évènement de spéciation, il peut se produire une perte d'un des deux gènes dupliqués dans une seule des deux espèces. Au niveau des allèles du gène ancestral, soit tous les allèles sont conservés dans les différentes lignées qui vont suivre (« complete lineage sorting ») soit certains allèles ne sont pas conservés entre les espèces (« Incomplete lineage ») (Rogers & Gibbs, 2014). Un arbre phylogénétique ne présente que des relations de descendances. Dès qu'il y a des transferts horizontaux entre les gènes, on parle d'un réseau phylogénétique (Huson & Bryant, 2006).

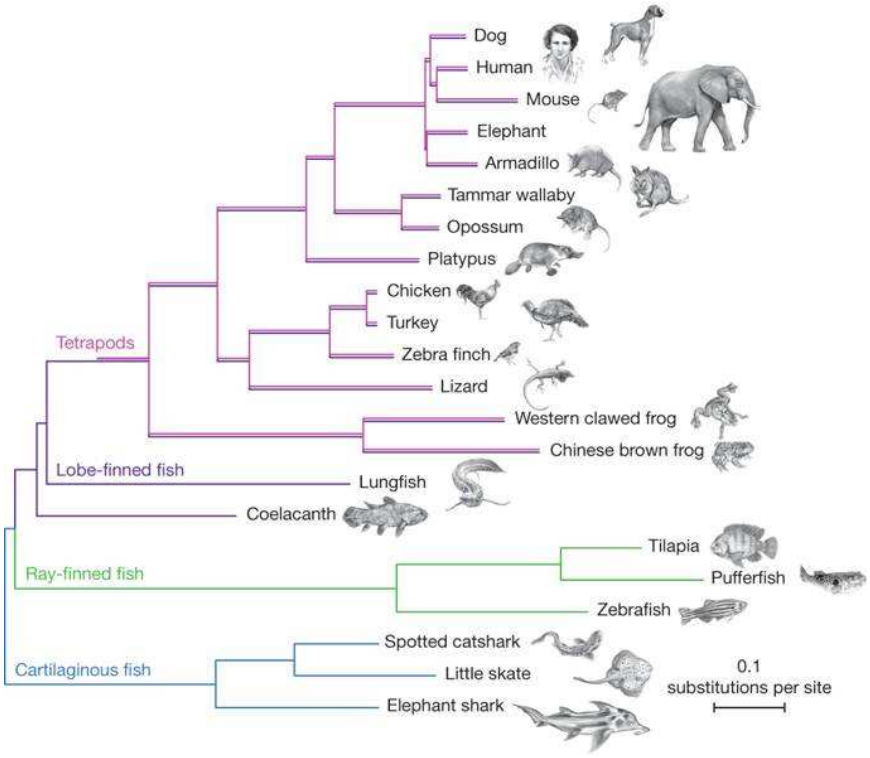


Figure 1: Arbre phylogénétiques des espèces

(Amemiya et al. 2013). Arbre phylogénétique des espèces Tétrapodes.

Les gènes homologues :

L'homologie est la relation de descendance commune entre deux entités (Koonin 2005). En génomique évolutive, ces entités concernent les gènes. Des gènes homologues partagent une origine commune. Les concepts d'orthologie et de paralogie correspondent à deux types d'homologies. Les gènes orthologues sont reliés par un évènement de spéciation. Les gènes paralogues sont reliés par un évènement de duplication (Figure 2). Concernant l'orthologie, les gènes ne sont pas forcément orthologues à un pour un (« 1 :1 »). Un gène qui n'a pas subi d'évènement de duplication dans une espèce peut être orthologue avec plusieurs gènes de l'autre espèce (« 1 : many » ou « many : 1 ») ou des gènes dupliqués dans une espèce peuvent être orthologues avec plusieurs gènes de la seconde espèce (« many : many »). Les gènes paralogues issus d'un évènement de duplication à la suite d'un évènement de spéciation sont des paralogues intra-espèces (« Inparalogs ») et les paralogues issus d'un évènement de duplication avant un évènement de spéciation sont des paralogues inter-espèces (« Outparalogs »).

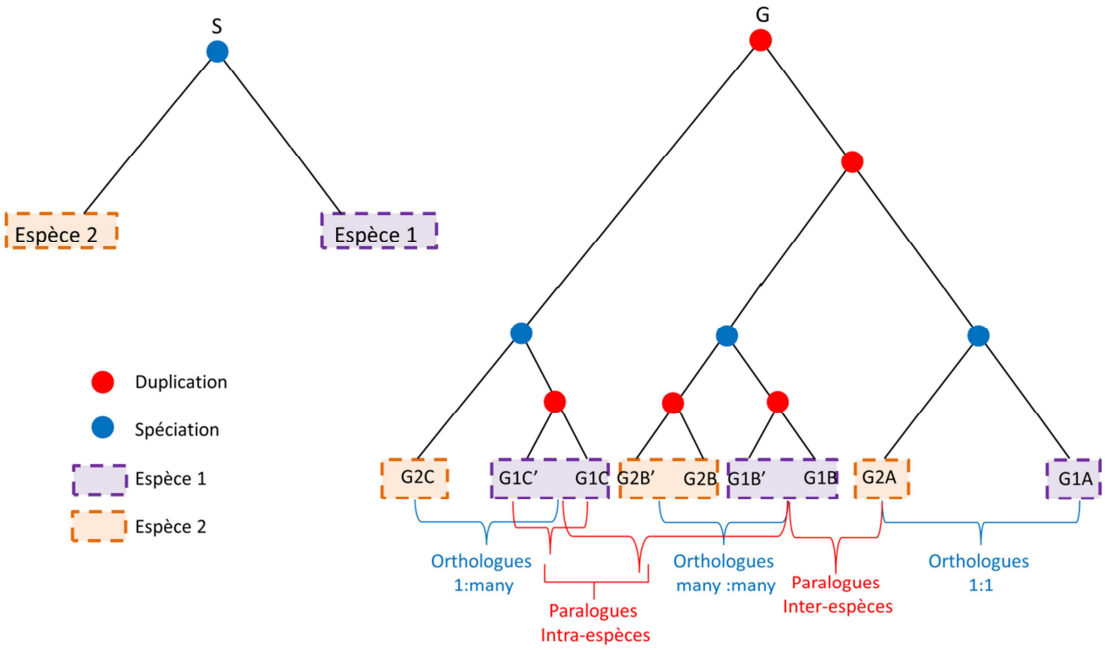


Figure 2: Arbre phylogénétique de gènes homologues

Schémas des arbres phylogénétiques d'espèces (« S ») et de gènes (« G ») d'un gène ancestral G de deux espèces 1 et 2. Définition des différents types d'orthologie et de paralogie.

La pression de sélection :

A la suite d'un évènement de duplication, les gènes paralogues vont avoir une fonction redondante. Certaines copies peuvent accumuler des mutations. Si ces mutations se concentrent sur une même copie de gène, la copie non mutée par sa redondance fonctionnelle va ainsi pouvoir compenser la fonction ou l'expression de la copie de gène mutée (Wagner 2002).

Avec le temps, les copies vont diverger. En effet des évènements de substitutions vont avoir lieu au sein de la séquence ADN ou protéiques des deux copies. Le nombre de substitutions par site obtenu par alignement entre séquences de gènes orthologues correspond à la distance évolutive entre ces gènes.

A partir de cet alignement, nous pouvons mesurer le nombre de substitutions synonymes avec le K_s (pouvant correspondre à des substitutions silencieuses). L'alignement permet aussi de compter le nombre de substitutions non synonymes qui changent l'acide-aminé produit et peuvent donc être délétères (K_a). La métrique de K_a/K_s permet d'estimer la pression de sélection des gènes paralogues. Un ratio obtenu d'une valeur proche de 1 indique qu'autant de substitutions synonymes que non

synonymes ont eu lieues. La pression de sélection est donc neutre. Si le ratio est inférieur à 1, cela signifie qu'il y a eu moins de substitutions non synonymes retenues dans la séquence génique que de substitutions synonymes. Les substitutions de ce gène avaient donc principalement des effets délétères et elles n'ont pas été retenues. Dans ce cas, la pression de sélection est négative. Enfin si K_a/K_s est supérieur à 1, plus de substitutions non synonymes ont été retenues indiquant qu'elles étaient principalement avantageuses pour l'évolution de l'espèce. Il s'agit d'une sélection positive. En comparant les ratios de K_a/K_s entre deux gènes d'une même espèce, nous pouvons identifier si un gène évolue plus rapidement que l'autre.

1.1.2. Types de duplications

Les gènes paralogues sont la résultante des évènements de duplication ayant eu lieu au cours de l'évolution. Il existe deux types d'évènement de duplication : celui à large échelle et celui à petite échelle. La duplication à large échelle correspond à la duplication de génome entier (WGD - « Whole Genome Duplication ») (Hakes et al. 2007). La duplication à petite échelle impacte uniquement un petit fragment du génome impliquant fréquemment un seul gène (SSD – « Small-Scale Duplication »).

La duplication de génome entier (WGD) :

Ce type de duplication correspond à l'hypothèse 2R (« 2 Rounds ») de deux évènements de polyploïdisation (diploïde 2n à tétraploïde 4n ou octoploïdie 8n) survenus il y a environ 450 millions d'années à la base de la lignée des vertébrés (Dehal & Boore 2005) (Figure 3)**.** Les gènes issus de cet évènement de WGD sont des ohnologues (Ohno 1970). A la suite de ces évènements de duplication, des évènements de transposition, de perte de gènes et de rediploidisation ont lieu. Après duplication du génome, la majorité des gènes dupliqués semble être supprimée (Force et al. 1999). Les gènes retenus dans le génome nous laissent penser qu'ils sont porteurs d'un avantage pour l'évolution des vertébrés.

La duplication à petite échelle (SSD) :

Les duplications de type SSD sont des duplications de courts fragments de l'ADN. Ces fragments concernent majoritairement un seul gène mais il est possible d'avoir plusieurs gènes impliqués dans le mécanisme. Ce type de duplication peut avoir lieu à

n'importe quel moment au cours de l'évolution et donc intervenir avant, après ou pendant une période de WGD. Plusieurs mécanismes connus induisent un évènement de SSD (Hurles 2004). La Figure 4 représente les trois mécanismes principaux de ce type de duplication. Une rétrotransposition de l'ARN messager du gène peut avoir lieu. L'ARNm subit alors une transcription inverse puis va s'intégrer aléatoirement dans le génome. Ce mécanisme induit un gène paralogue qui ne possède pas d'intron mais une queue poly-A et qui ne se trouve pas forcément sur le même chromosome que sa paire. Les gènes issus d'une rétrotransposition sont rarement fonctionnels car ils ne possèdent plus de région régulatrice. La duplication segmentale, quant à elle, n'implique pas d'ARNm. La recombinaison segmentale peut consister à une recombinaison homologue à partir d'un élément *Alu* très présent dans le génome humain (~10% du génome). Ce type de recombinaison peut avoir lieu dans le cas où les gènes sur des chromosomes homologues sont décalés l'un par rapport à l'autre. La duplication segmentale peut aussi intervenir à la suite d'une erreur de réplication au niveau du site de terminaison de la réplication. Le gène va donc être répliqué deux fois sur le même brin. Ces duplications segmentales sont la cause des duplications en tandem localisées sur le même chromosome.

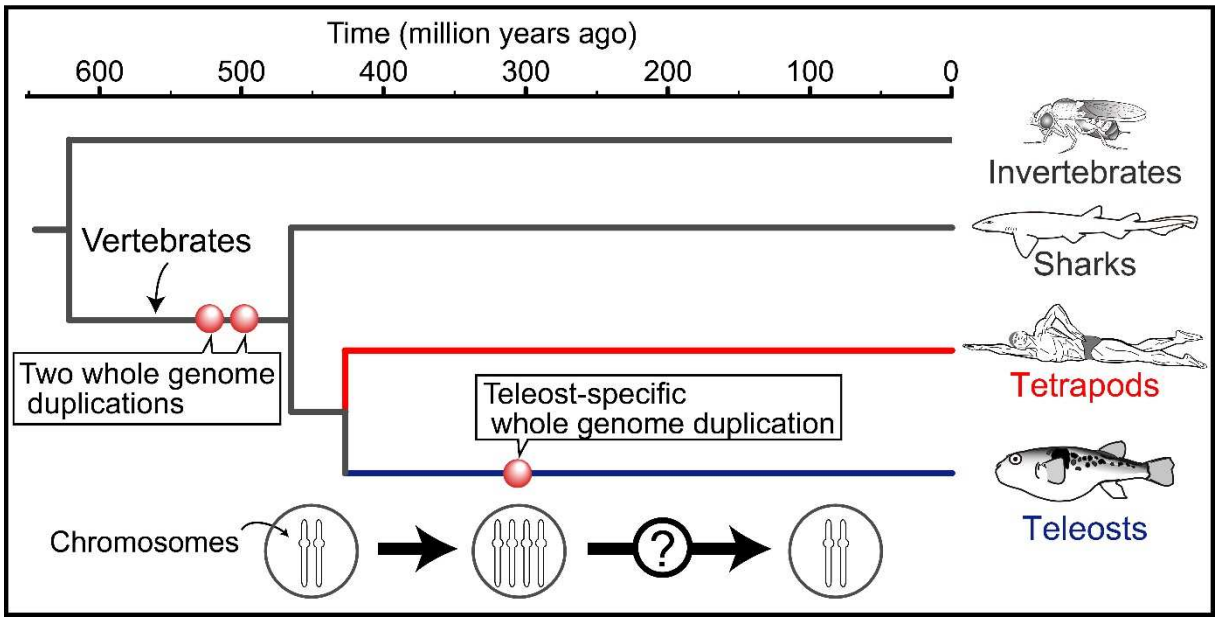


Figure 3: Dates des évènements de duplication de génome entier (WGD)

www.oist.jp. Les points rouges représentent les évènements de WGDs qui ont lieu dans l'histoire évolutive. Les WGDs chez l'homme correspondent aux deux évènements qui ont lieu au début de la lignée des vertébrés.

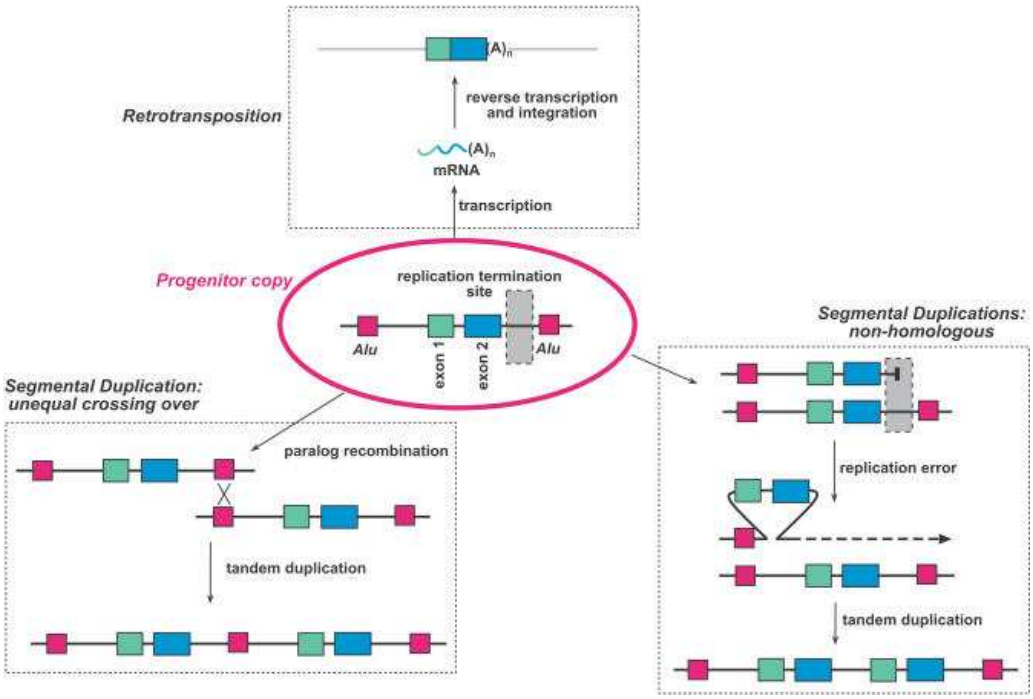


Figure 4: Les mécanismes de la duplication SSD

(Hurles 2004) Gene Duplication: The Genomic trade in Spare Parts. PLOS Biology 2(7). DOI:10.1371/journal.pbio.0020206. Rétrotransposition de l'ARNm : transcription inverse et intégration aléatoire dans le génome. Deux types de duplications segmentales : Recombinaison homologue du gène entre deux éléments *Alu* non en phase ou erreur de réplication au niveau du site de terminaison créant une duplication en tandem.

1.2.Devenir fonctionnel des gènes paralogues

1.2.1.Pseudogénisation

Le phénomène de pseudogénisation correspond à une désactivation d'un gène par l'accumulation de mutations délétères (Ohno 1970). En effet, à la suite d'un évènement de duplication, la majorité des gènes dupliqués ne va pas être retenue dans le génome. Ce mécanisme de dégradation d'un des deux gènes paralogues est la pseudogénisation (Figure 5) (Hurles 2004). Dans le cas où le maintien de la redondance fonctionnelle au sein du génome n'est pas un avantage, le gène ne va pas être gardé en deux copies.

La perte d'une des deux copies peut survenir dans le cas où une mutation dans la région codante du gène va introduire un codon stop prématuré ce qui va détruire la structure du domaine protéique et rendre la protéine non fonctionnelle après traduction. Après une pseudogénisation, le gène dégradé ne va donc plus être fonctionnel. Cependant dans certains cas il est possible que celui-ci soit quand même transcrit. (Harrison et al. 2005).

1.2.2.Equilibre de dosage

L'équilibre de dosage existe entre des gènes qui interagissent ou qui participent à une même voie de signalisation (Makino & McLysaght 2010). Il consiste en un équilibre de production de transcrits entre gènes. Cet équilibre peut favoriser la rétention de la duplication par l'augmentation ou le partage de l'abondance de la production d'ARNm (Zhang 2003; Lan & Pritchard 2016) (Figure 5). L'augmentation de la production de protéines est une contribution possible pour créer les conditions permettant de favoriser la rétention. En effet pour les évènements de WGDs, en cas d'équilibre, la suppression d'un gène n'est pas favorisée car elle peut perturber la production de protéines entre gènes en forte interaction (Papp et al. 2003). Concernant les SSDs, les paires de gènes générées doivent maintenir une interaction avec des singletons et vont donc être sous-régulées pour atteindre un partage d'expression permettant de conserver l'équilibre de dosage (Innan & Kondrashov 2010). Dans le cas du partage de la production d'ARNm, au cours du temps les deux copies peuvent présenter une expression asymétrique (AED – « Asymmetrically Expressed Duplicates ») et le gène le moins exprimé va être dégradé (Lan & Pritchard 2016). Afin que les deux gènes paralogues survivent sur le long terme ils doivent se sous-fonctionnaliser ou se néo-fonctionnaliser.

1.2.3.Néo-fonctionnalisation

La néo-fonctionnalisation est le gain d'une nouvelle fonction pour l'une des deux copies du gène ancestral. Cette nouvelle fonction semble être provoquée par l'accélération du taux de substitution des acide aminés (Figure 5) (Hurles 2004) sur l'un des deux gènes dupliqués. Ces substitutions seraient en grande partie des substitutions en arginine qui auraient lieu peu de temps après l'évènement de duplication (Zhang 2003). Des gènes néo-fonctionnalisés vont donc être retenus au sein du génome.

1.2.4.Sous-fonctionnalisation

La sous-fonctionnalisation correspond au partage de la fonction du gène ancestral entre les deux gènes paralogues. Cette fonctionnalisation peut-être, d'une part, associée au modèle de Duplication-Dégénérescence-Complémentation (DDC) (Force et al. 1999). Ce modèle met en évidence une dégénérescence complémentaire des régions régulatrices

des gènes paralogues. Ainsi les deux gènes vont s'exprimer dans des tissus différents. Les deux copies se partagent donc les régions où s'exprimait le gène ancestral (Figure 5). La sous-fonctionnalisation peut aussi concerner la fonction en elle-même et non pas l'expression. Ainsi les deux paralogues vont se spécialiser dans une des fonctions du gène ancestral. Dans le cas d'une sous-fonctionnalisation, les deux paralogues vont être maintenus dans le génome ce qui participe à la complexité de l'organisme.

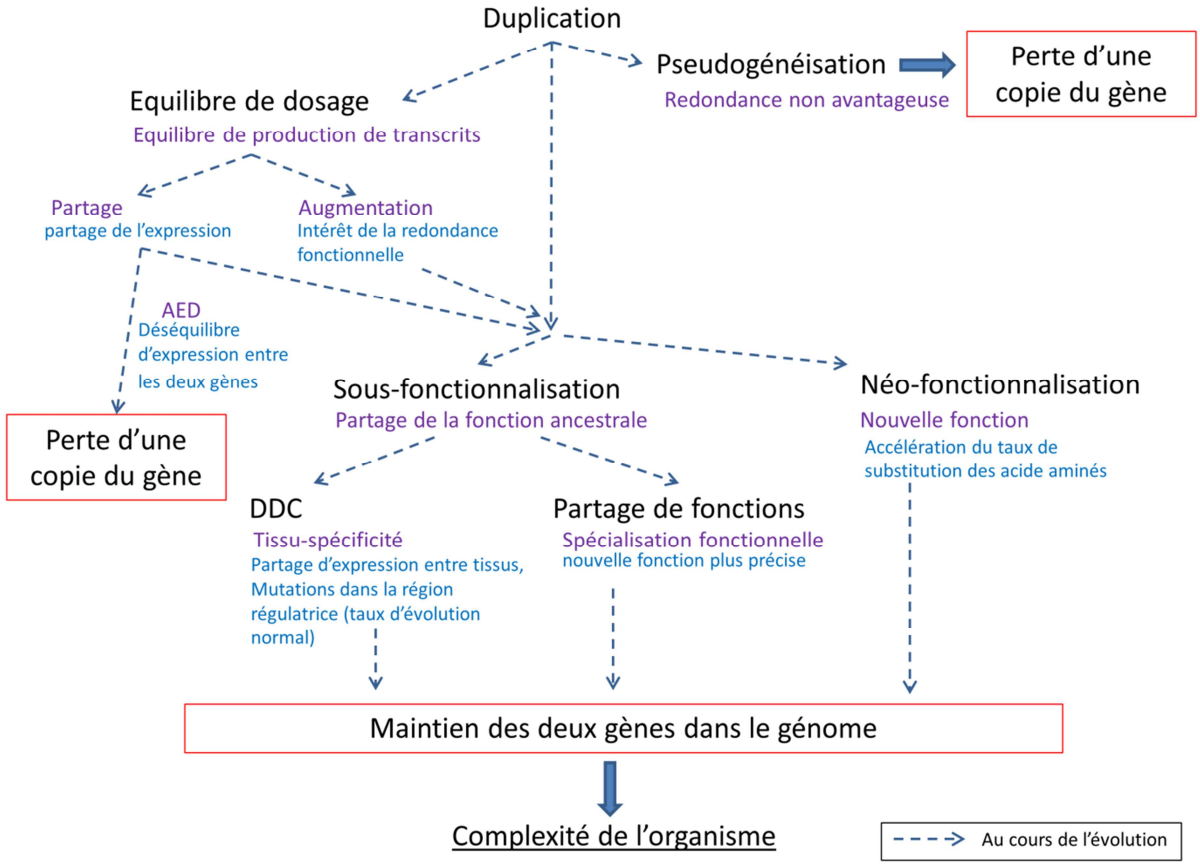


Figure 5: Devenir fonctionnel des gènes paralogues

Après un évènement de duplication, une grande majorité des gènes va être dégradée (pseudogénisation). Les paires de paralogues restantes vont revenir à un équilibre de dosage. Les deux paralogues peuvent ensuite se partager la fonction ancestrale (sous-fonctionnalisation) ou bien l'un des deux gènes va acquérir une nouvelle fonction (néo-fonctionnalisation). Excepté un partage du dosage qui peut dériver vers un déséquilibre d'expression entre les deux paralogues et à la dégradation d'un des deux gènes, les autres devenirs fonctionnels mènent au maintien des deux paralogues dans le génome ce qui va augmenter la complexité de l'organisme.

1.3.Familles de gènes

1.3.1.Définition des familles de gènes

Une famille de gènes est un groupe de gènes descendant d'un unique ancêtre commun (Li et al. 2006). Elle regroupe les gènes orthologues et paralogues. Si une famille contient plusieurs gènes de la même espèce alors ces gènes sont paralogues.

Les gènes d'une même famille peuvent également être fonctionnellement co-régulés permettant de créer de nouvelles interactions ayant un rôle important dans la complexité d'un organisme (Charrier et al. 2012). Par exemple les duplications des gènes de la famille *SRGAP2* (« SLIT-ROBO Rho-GTPase-activating protein 2 ») ont contribué au développement du néocortex humain. En effet le gène *SGAP2C* inhibe *SRGAP2A* permettant le ralentissement de la maturation de la moelle épinière permettant l'élongation de celle-ci ce qui joue un rôle important dans le développement du cerveau humain (Charrier et al. 2012; Dennis et al. 2012).

1.3.2.Tissu-spécificité des familles de gènes

La duplication d'un gène est un facteur de tissu-spécificité des gènes dans l'organisme. Les gènes d'une même famille peuvent diverger suivant le modèle de DDC (« Duplication-Degeneration-Complementation »)(Force et al. 1999) et ainsi se partager l'expression tissulaire du gène ancestral (Freilich et al. 2006). Au cours de l'évolution, ces gènes vont devenir plus tissu-spécifiques que le gène ancestral. Les familles de gènes peuvent présenter différents profils d'expression au sein de différents tissus (Figure 6). Les gènes paralogues peuvent se partager les tissus dans lequel le gène ancestral était actif. La répartition de l'expression des gènes au sein des tissus peut se faire sans (Figure 6– Famille A) ou avec des chevauchements entre ces tissus (Figure 6– Famille C). Tous les gènes d'une famille peuvent également s'exprimer dans les mêmes tissus (Figure 6– Famille B). Dans ce cas les gènes de la famille vont avoir des profils d'expression similaires au travers des différents tissus. Les singletons formant des familles à gène unique au sein d'une espèce peuvent également montrer un profil tissu-spécifique (Figure 6– Singleton B) ou bien un profil plus ubiquitaire (Figure 6– Singleton A).

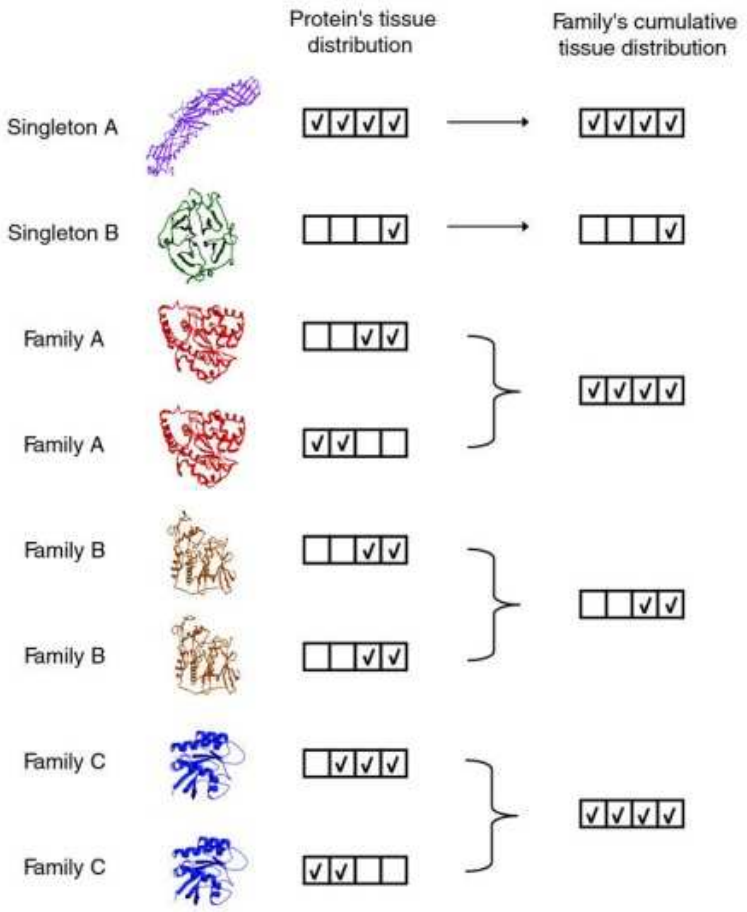


Figure 6: Profils d'expression des familles de gènes au sein des tissus

(Freilich et al. 2006) Profils d'expression cumulés de différentes familles de gènes dans 4 tissus hypothétiques. La famille A est un exemple d'expression complémentaire des gènes de la famille qui une fois cumulé donne l'expression du gène ancestral sans chevauchement. La famille B est un exemple de tissu-spécificité identique entre les gènes de la famille. La famille C est un exemple d'expression complémentaire entre les gènes paralogues mais avec chevauchement.

1.3.3.Méthodes d'identification des familles de gènes

Les familles de gènes sont déterminées de façon bioinformatique. L'identification des familles de gènes est l'une des façons qui permet de retrouver les gènes paralogues. Le principe global de toutes les méthodes d'identification des familles est lié à la comparaison de séquences protéiques de différentes espèces. Les outils et base de données les plus récents utilisent des méthodes basées sur la recherche d'orthologues à partir d'analyses d'arbres phylogénétiques d'espèces (voir Méthode Chapitre 2). Les

bases de données utilisant cette méthodes sont TreeFam (Ruan et al. 2008), Ensembl Compara (Vilella et al. 2009) et HOGENOM (Penel et al. 2009).

2.Méthodologie de séquençage RNA-seq

Le RNA-seq est une méthodologie ayant pour objectif d'obtenir l'identification et la quantification du transcriptome (ARN). Cette méthode de séquençage est classiquement utilisée pour mesurer le niveau d'expression des transcrits afin notamment de réaliser des analyses différentielles d'expression de gènes entre conditions expérimentales (Wang et al. 2009).

Pour séquencer des échantillons d'ARN par RNA-Seq, il faut d'abord réaliser la construction d'une banque de séquençage RNA-Seq. Celle commercialisée par Illumina et adaptée à ses plateformes de séquençage est effectuée au moyen d'un kit nommé TruSeq.

Dans le cas d'un séquençage des ARNm, ceux-ci sont d'abord enrichis, à partir d'une extraction d'ARN total, par hybridation de billes poly-T à la queue poly-A de l'ARNm (Figure 7 A et B). Cette étape de sélection est suivie par une coupure aléatoire afin d'obtenir des fragments d'ARN. Une RT-PCR (« Reverse transcriptase polymerase chain reaction » - Transcription inverse et amplification de l'ARN) est ensuite effectuée avec des amorces aléatoires afin de générer de l'ADNc (ADN complémentaire) à partir des fragments d'ARNm. Une adénine et un phosphate sont ajoutés aux extrémités de chaque brin d'ADNc permettant de les lier aux adaptateurs de séquençage Illumina (Figure 7 C). La dernière étape consiste à l'amplification des fragments (Figure 7 D).

Une fois que les adaptateurs Illumina sont liés aux fragments d'ADNc, ces derniers vont s'hybrider aléatoirement à la surface d'une cellule de séquençage (Figure 8) grâce aux adaptateurs. Des étapes d'amplification effectuées après l'ajout de dNTPs (désoxy-nucléotide-tri-phosphate) vont permettre la génération de clusters de séquences identiques. A la suite de cette étape, le séquençage peut s'effectuer par l'ajout d'amorces de séquençage qui vont s'hybrider aux fragments. Des dNTPs marqués par 4 fluorochromes différents, un pour chaque type de nucléotide, sont également ajoutés et vont s'hybrider grâce à l'incorporation d'une enzyme ADN polymérase dans la solution permettant l'élongation de l'amorce. Comme les clusters sont composés de séquences identiques, l'élongation se fait de manière synchrone pour tout un cluster ce qui permet une amplification du signal de fluorescence. Une image de la cellule est ensuite capturée

par une caméra TDC (dispositif à transfert de charge) puis enregistrée. A chaque cycle de séquençage, une nouvelle base est ajoutée jusqu'à obtenir la longueur de séquence voulue (ex. 100 pb). Pour finir, les images enregistrées à chaque cycle sont superposées et analysées afin de pouvoir lire la séquence du fragment amplifié dans chaque cluster. Le séquençage « pairé » (« paired-end ») qui consiste au séquençage des deux extrémités du fragment, utilise une seconde amorce complémentaire avec le second adaptateur lié aux fragments.

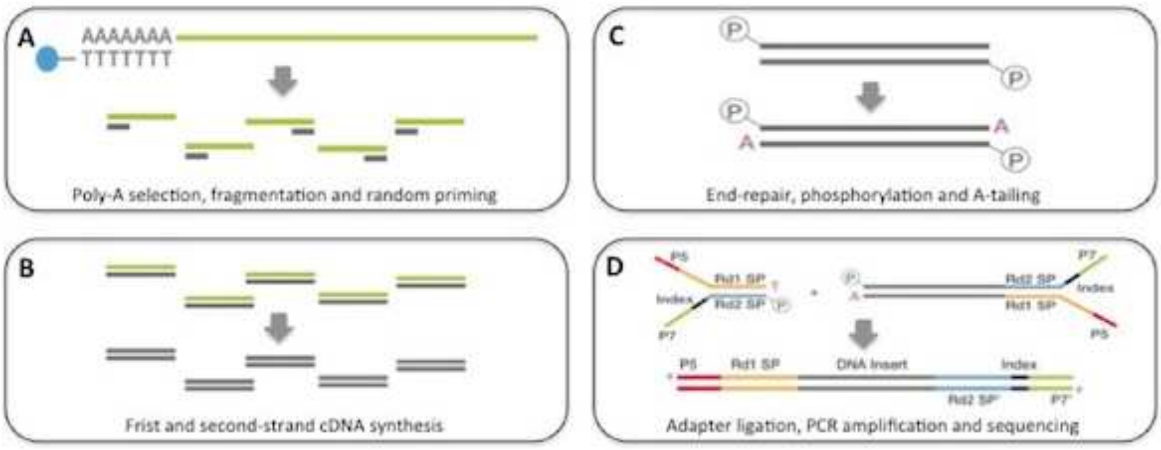


Figure 7:Préparation et séquençage RNA-seq

(Mardis 2008) (A)Sélection Poly-A, fragmentation et amorces aléatoires. (B) Synthèse de l'ADNc. (C) Réparation des extrémités, phosphorylation et ajout d'une adénine à l'extrémité du brin. (D) Liaison aux adaptateurs, amplification et séquençage. www.illumina.com.

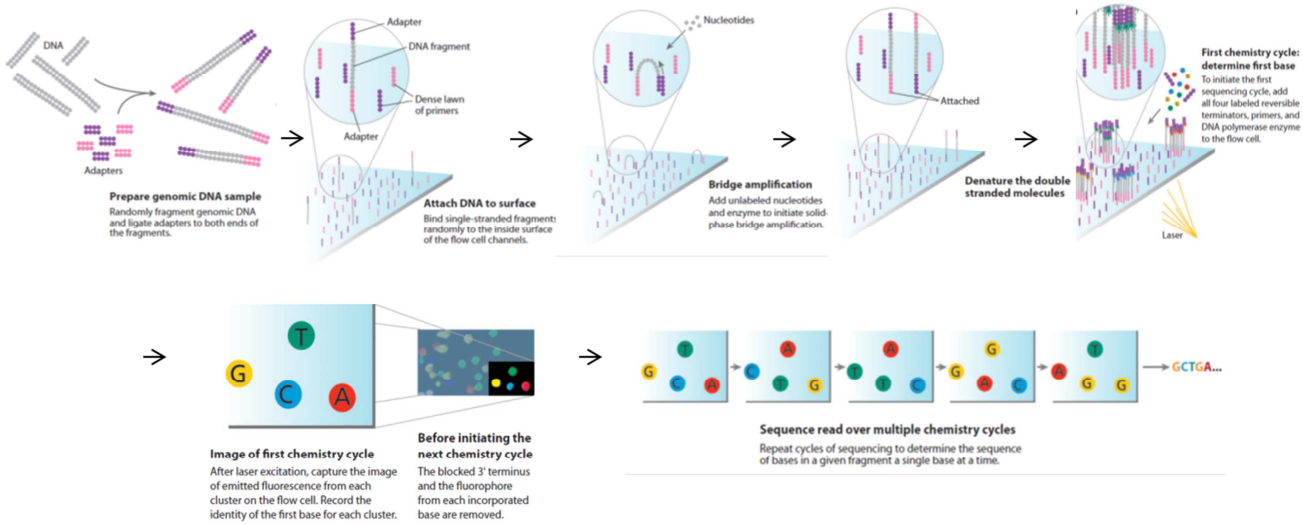


Figure 8: Principe du séquençage Illumina.

Etapes du séquençage RNA-Seq avec liaison des adaptateurs, amplification («bridge amplification »), ajout de dNTPs marqués, capture d'image et reconnaissance des bases (« base calling »).

3. Consortium GTEx

3.1.Description des échantillons

Le consortium GTEx (« Genotype-Tissue Expression ») est un projet qui a pour but de décrire le paysage de l'expression des gènes au travers de différents tissus humains collectés post-mortem sur des individus (Ardlie et al. 2015). De l'ARN et de l'ADN ont été extraits par le consortium à partir de chaque échantillon afin d'effectuer du séquençage RNA-seq et du génotypage. GTEx donne accès aux mesures d'abondance des gènes générées à partir des séquençages RNA-seq (comptages du nombre de lectures par gène et mesure d'expression normalisée) mais également aux données de génotypage et à des eQTL (« expression Quantitative trait Loci »).

Nous utilisons dans le projet de thèse la version V6p des données du consortium. Cette ressource permet de récupérer des données d'expression de gènes pour 53 tissus humains collectés à partir de 544 donneurs, ce qui fait un total de 8555 échantillons. Parmi ces échantillons, un grand nombre d'entre eux a pu être utilisé pour du génotypage. (Figure 9-version V7). Plusieurs tissus différents ont été collectés pour chaque donneur (Figure 10- version V7).

L'extraction de l'ARN de chaque tissu survenant à la suite du décès du donneur (post-mortem), la qualité de l'ARN va baisser à cause de sa dégradation. Une mesure de dégradation de l'ARN, par Agilent (Schroeder et al. 2006) correspond au RIN (« RNA Integrity Number »). Plus la valeur du RIN est faible, plus l'ARN est dégradé (Figure 11). Le consortium GTEx a considéré les extractions d'ARN ayant une valeur RIN supérieure à 6 (max = 10) comme celles pouvant être utilisées. Nous remarquons qu'une majorité des extractions d'ARN obtenues à partir des tissus cérébraux (ex. cortex et cervelet) ont des valeurs de RIN supérieures à 6.

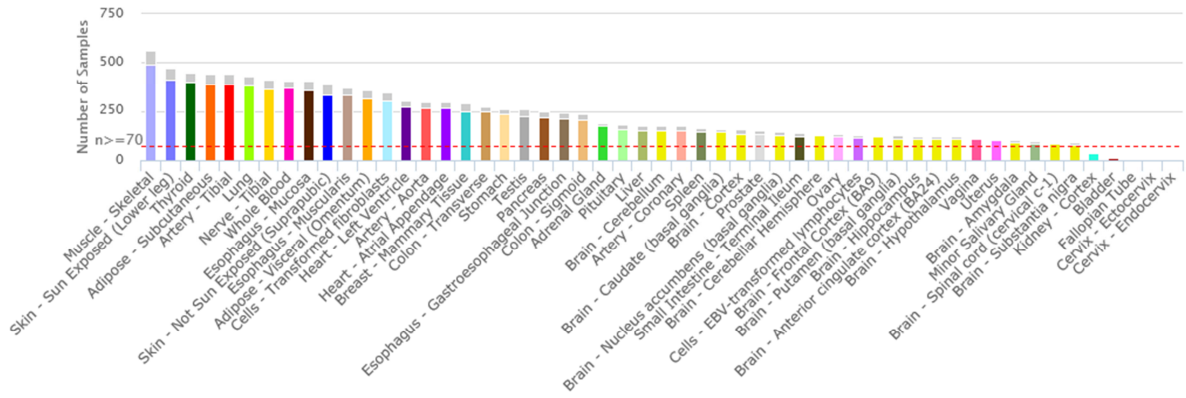


Figure 9: Nombre d'échantillons par tissu.

Portail GTEx. Nombre d'échantillons présents pour les 53 tissus du consortium GTEx . Pour chaque tissu la partie colorée de la barre représente les échantillons qui présentent des données de génotypage.

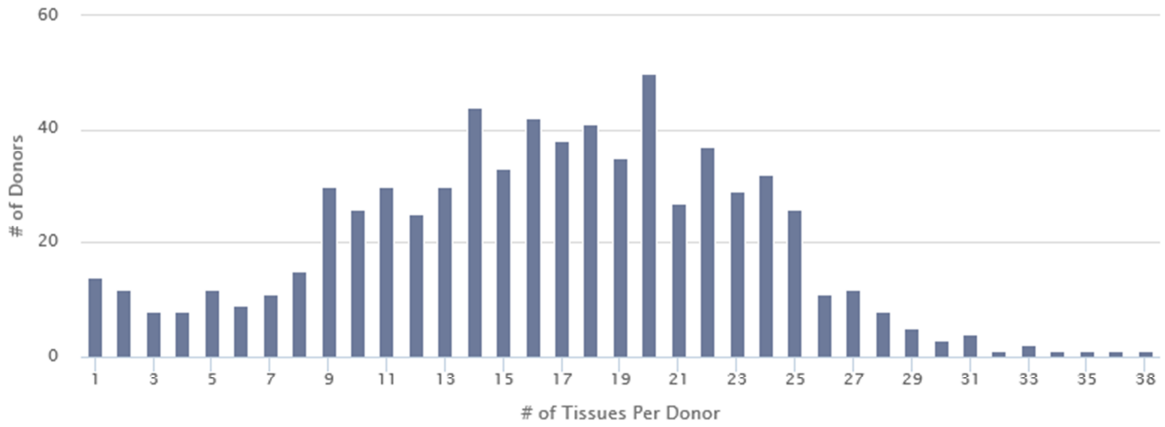


Figure 10: Nombre de donneurs par nombre de tissus donnés.

Portail GTEx. Nombre de donneurs en fonction du nombre de tissus par donneur.

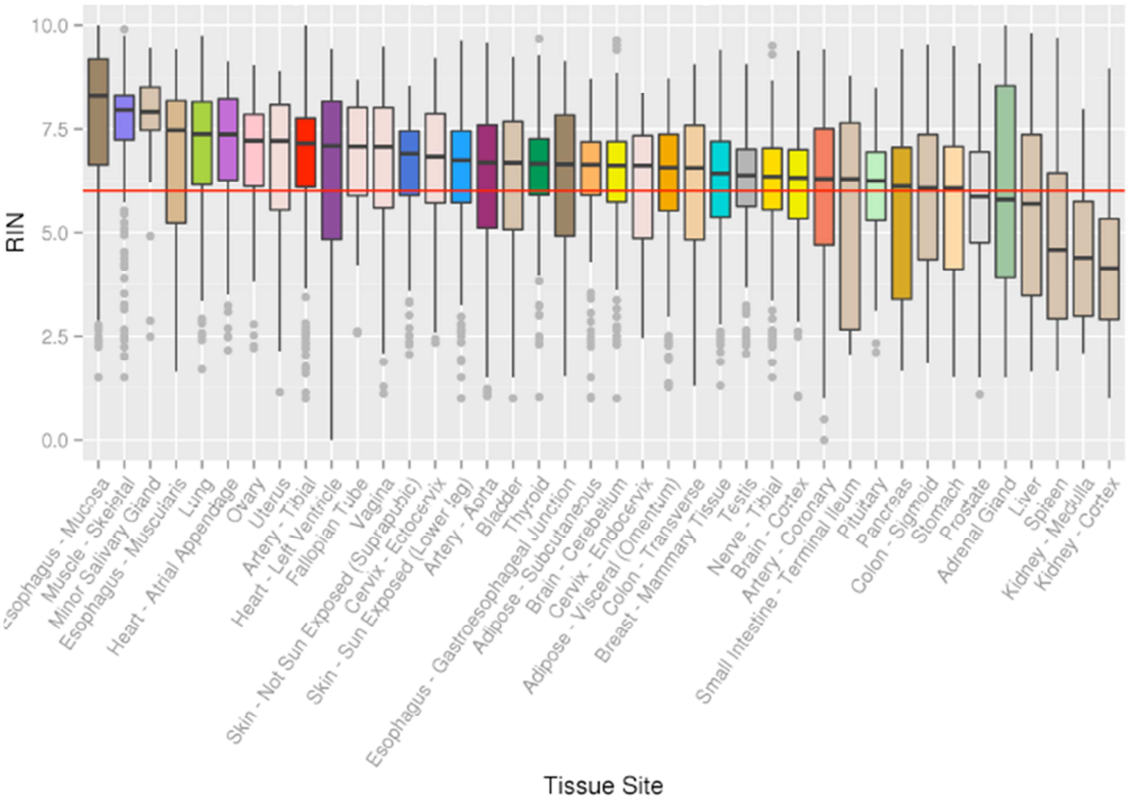


Figure 11: Valeurs moyennes des RIN pour chaque tissu

(Ardlie et al. 2015). Boites à moustache de la qualité de l'ARN (RIN) pour 40 tissus du consortium GTEx (projet pilote). La ligne rouge représente la valeur seuil du RIN.

3.2.Données d'expression de gènes

Les données de séquençage RNA-seq du consortium GTEx ont été produites avec un séquenceur Illumina HiSeq 2000. Elles correspondent à des lectures « pairées » de 2x76pb. La profondeur moyenne de séquençage (nombre total de lectures alignées sur le génome), est de 82,1 millions de lectures alignées par échantillon (Ardlie et al. 2015).
La mesure d'expression de chaque gène repose sur le comptage des lectures de séquençage alignées dans les exons de ce gène. Les données RNA-seq de GTEx (version V6p) ont été alignées avec l'outil TopHat v1.4.1 en n'autorisant que les alignements avec placement unique sur le génome humain de référence (GRCh37 – « Genome Reference Consortium») (http://genome.ucsc.edu/) avec l'annotation de gènes Gencode V19 (Harrow et al. 2012).

Les mesures d'expression ont été normalisées en unité RPKM (« Reads Per Kilobase per Million mapped reads »). Dans la dernière version V7, le consortium produit également des données d'expression normalisées par la méthode TPM (TPM – « Transcript Per

Million »). L'expression d'un gène modèle au travers des tissus humains est présentée en exemple (ENSG00000065613.9) sur la ressource internet du consortium GTEx (Figure 12).

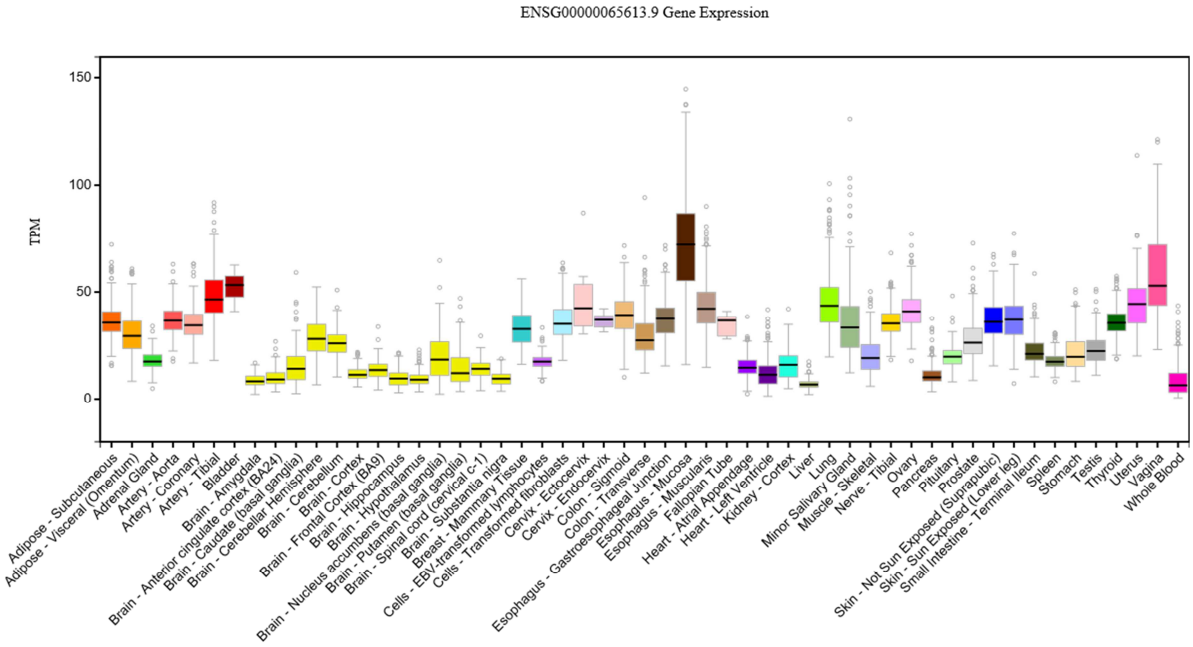


Figure 12: Expression du gène SLK au travers des tissus humains.

Boites à moustache des valeurs TPM (« Transcipt per million ») du gène modèle SLK dans tous les 53 tissus humains (V7).

Le consortium GTEx a publié plusieurs études sur l'expression des gènes aux travers de différents tissus (Ardlie et al. 2015; Saha et al. 2016).

Des analyses de classifications hiérarchiques, réalisées à partir des profils d'expression de gènes par échantillon, permettent de regrouper les échantillons en clusters correspondants aux différents tissus (voir Chapitre 3). L'expression tissulaire des gènes est notamment explorée par GTEx par l'inférence de réseaux de gènes co-exprimés soit au sein d'un tissu au travers de plusieurs individus soit au travers de différents tissus (voir Introduction Chapitre 4).

4.Contexte et objectifs de la thèse

4.1.Contexte du projet de thèse

L'objectif de ce projet de thèse est d'étudier l'expression tissulaire des gènes paralogues au sein des différentes régions du cerveau.

Le travail de thèse a permis d'aborder trois aspects liés à cet objectif :

- Une caractérisation détaillée (propriétés évolutives, de séquence et ontologiques) des gènes paralogues et la mise en place d'une ressource nécessaire pour la suite de la thèse ;
- Une évaluation de la problématique d'identité de séquence entre les gènes paralogues et un travail sur l'estimation de la fiabilité de la mesure d'expression de ces gènes dans des échantillons cérébraux ;
- Une étude de l'expression tissulaire des gènes paralogues dans les différentes régions du cerveau humain appliquée aux pathologies cérébrales.

Dans un premier temps, j'ai travaillé sur la caractérisation des gènes paralogues et sur la découverte de la biologie de l'évolution. J'ai également établi une ressource sur les gènes paralogues que j'ai pu exploiter dans la suite de ma thèse. Cette ressource agrège des informations sur une liste de gènes paralogues comme leur famille de gènes, le type et la datation de leur duplication ainsi que des caractéristiques sur leur identité de séquence.
Ensuite, j'ai commencé à m'intéresser aux identités de séquence des paralogues et aux biais de la mesure d'expression générés par des fortes homologies de séquence. Suite à ces travaux, nous avons considéré que les alignements avec placement unique sur le génome des lectures RNA-seq, réalisés par le consortium GTEx, permettaient d'explorer correctement l'expression d'une grande majorité de gènes paralogues. J'ai donc poursuivi par la normalisation et l'ajustement des mesures de comptage par gène fournies par GTEx. J'ai ensuite travaillé sur l'inférence de réseaux de co-expression de gènes à partir de la méthodologie WGCNA. Cela a consisté à prendre en main cette méthodologie assez complexe et à optimiser ses paramètres afin de produire des réseaux de co-expression compatibles avec nos interrogations sur les familles de gènes paralogues.
Enfin, j'ai entrepris un travail sur la mesure de la spécificité d'expression tissulaire des gènes. Cela m'a permis de m'intéresser à l'analyse intégrée des informations évolutives, de co-expression et de tissu-spécificité des gènes paralogues et de leurs familles. Finalement, j'ai travaillé sur l'exploitation de ces caractéristiques des gènes paralogues afin d'améliorer notre connaissance des gènes impliqués dans les maladies cérébrales. .
En outre, au cours de cette dernière année j'ai également eu l'occasion de travailler sur des projets annexes comme l'analyse bioinformatique des spike-ins en RNA-seq (ERCC en RNA-seq).

4.2.Valorisations pendant la période de thèse

Au cours de ces trois années de thèse, j'ai eu plusieurs occasions pour communiquer sur mes résultats:

- La présentation de mes travaux sur la caractérisation des paralogues et sur les problématiques d'identité de séquences par un poster à la conférence « Biomathematics Conference : Statistical Analysis of Massive Genomic data » à EVRY en 2015 et par un poster aux Journées Ouvertes en Biologie, Informatique et Mathématiques à LYON en 2016;
- La présentation de mes travaux sur l'expression tissulaire et la co-expression des gènes paralogues dans les tissus cérébraux par une communication orale et un poster à la conférence « European Society of Human Genetics» à Copenhague, Danemark en 2017;
- La soumission d'un manuscrit au journal « Molecular Biology and Evolution » en octobre 2017.

4.3. Objectif global de la thèse

Ce projet de thèse présente plusieurs objectifs que nous souhaitons atteindre à la fin du projet.

Un premier objectif préliminaire à toute étude sur les gènes paralogue est de pouvoir obtenir une liste fiable de gènes paralogue. A partir de cette liste, nous souhaitons construire une ressource la plus complète possible sur ces gènes avec notamment des informations essentielles telles que l'information des familles de gènes et la datation des évènements de duplication.

A partir de l'exploitation des caractéristiques évolutives des paralogues, nous souhaitons mieux comprendre la spécificité tissulaire de leur expression dans les différentes régions du cerveau et leur co-régulation.

Enfin, nous souhaitons montrer l'intérêt d'associer les propriétés de tissu-spécificité et de co-expression de ces gènes paralogues pour améliorer notre connaissance des gènes impliqués dans des pathologies cérébrales.

Parte 1 : Caractérisation des gènes paralogues

Chapitre 2 : Base de référence et caractérisation des gènes paralogues

1.Introduction

1.1.Caractéristiques évolutives des gènes issus des duplications WGD et SSD

Les gènes paralogues sont issus des évènements de duplication de type WGD ou de type SSD (voir Chapitre 1). Ces deux catégories de duplications peuvent être comparées au niveau de leur fonction, de leur taux d'évolution ainsi que de leur conservation de séquence dans le génome.

Au cours de l'évolution, les séquences géniques changent plus ou moins rapidement. Le taux d'évolution des gènes peut être estimé par la pression de sélection mesurée par le K_a/K_s avec le K_a, le nombre de substitutions non synonymes et le K_s, le nombre de substitutions synonymes (Wagner 2002) (voir Chapitre 1).). Le taux d'évolution est utilisé pour comparer la divergence entre deux séquences de gènes au cours du temps. Il a été montré chez l'homme que globalement, les WGDs avaient un taux d'évolution significativement plus faible que celui des (Acharya & Ghosh 2016). Cela signifie que les gènes WGDs ont tendance à être plus conservés d'un point de vue évolutif.

Il est également intéressant d'étudier les fonctions moléculaires et biologiques associées à chaque type de duplication. Dans cette même étude (Acharya & Ghosh 2016), la divergence de fonction est abordée par la proportion de termes ontologiques (GO) partagés entre les deux paralogues d'une même paire au sein des WGDs ou des SSDs. Les paires de gènes dupliqués par WGD chez l'homme ont été retrouvées comme ayant des fonctions plus divergentes que les paires de SSDs. Les WGDs ont donc acquis et conservé au cours du temps davantage de nouvelles fonctions biologiques. De plus, cette comparaison entre WGD et SSD a été réalisée pour différentes valeurs de K_a entre les paires de gènes considérées. Cela laisse apparaître que peu de temps après l'évènement de duplication, le taux d'évolution des WGDs serait élevé jusqu'à obtenir des nouvelles fonctions. Une fois ces nouvelles fonctions acquises, le taux d'évolution des WGDs diminuerait pour devenir plus faible que celui des SSDs afin de permettre la conservation de ces fonctions nouvelles.

1.2.Conservation de séquence des gènes ayant subi une duplication

Le risque de forte identité de séquence (faible taux d'évolution) entre certains gènes paralogues peut entraîner des problèmes d'alignement de lectures et de mesure d'expression dans le cadre de l'analyse de données de séquençage RNA-seq.

Le concept de « mappabilité » (Derrien et al. 2012) est utilisé en séquençage à haut débit afin d'estimer, pour une lecture donnée, le nombre d'alignements avec placement unique ou multiple sur le génome. En d'autres termes, cette mesure de « mappabilité » permet d'estimer, pour une longueur de lecture donnée, la proportion du génome qui pose des problèmes d'alignements non-spécifiques. Cette « mappabilité » correspond à l'inverse du nombre de fois qu'un fragment d'ADN de longueur k (ou lecture, ou k-mer) peut s'aligner sur la séquence de référence.

Cette mappabilité peut être mesurée (Derrien et al. 2012) pour un transcriptome humain de référence (GENCODE) (Figure 13) pour différents types de transcrits. La Figure 13A représente le pourcentage de k-mers (taille fixe de 36pb) qui s'alignent de façon unique sur le transcriptome humain de référence pour un nombre de mésappariements (« mismatches ») autorisés de 0 (alignement exact) ou de 2. Plus le nombre de « mismatches » autorisés augmente, plus la proportion de k-mers qui s'alignent de manière unique diminue. Concernant la Figure 13 B, le nombre de « mismatches » autorisés est de 2 et la taille des k-mers varie entre 36 et 100pb. La quantité reportée sur l'ordonnée n'est plus la proportion de k-mers qui s'alignent de manière unique mais celle des nucléotides du transcriptome démarrant un de ces k-mers. Plus la taille des k-mers est grande, plus le pourcentage de nucléotides qui appartiennent à une région d'alignement unique est élevé.

D'après les deux figures, les gènes codants pour des protéines ont presque 95% de leur séquence qui donne lieu à des alignements uniques par des k-mers (proportion de nucléotides) sur le transcriptome. On remarque aussi que les pseudogènes ont un pourcentage de régions uniques beaucoup plus faible (~70%). Les pseudogènes étant issus d'une duplication dont l'une des copies a été dégradée, leur mappabilité peut être représentative de celle des gènes dupliqués, bien que le taux de conservation des pseudogènes soit probablement différent de celui des gènes dupliqués. C'est ainsi que, pour les gènes codants pour des protéines, on peut supposer que cette portion de 5% de séquence qui ne donne pas lieu à des alignements uniques correspond en grande partie à des parties de gènes dupliqués.

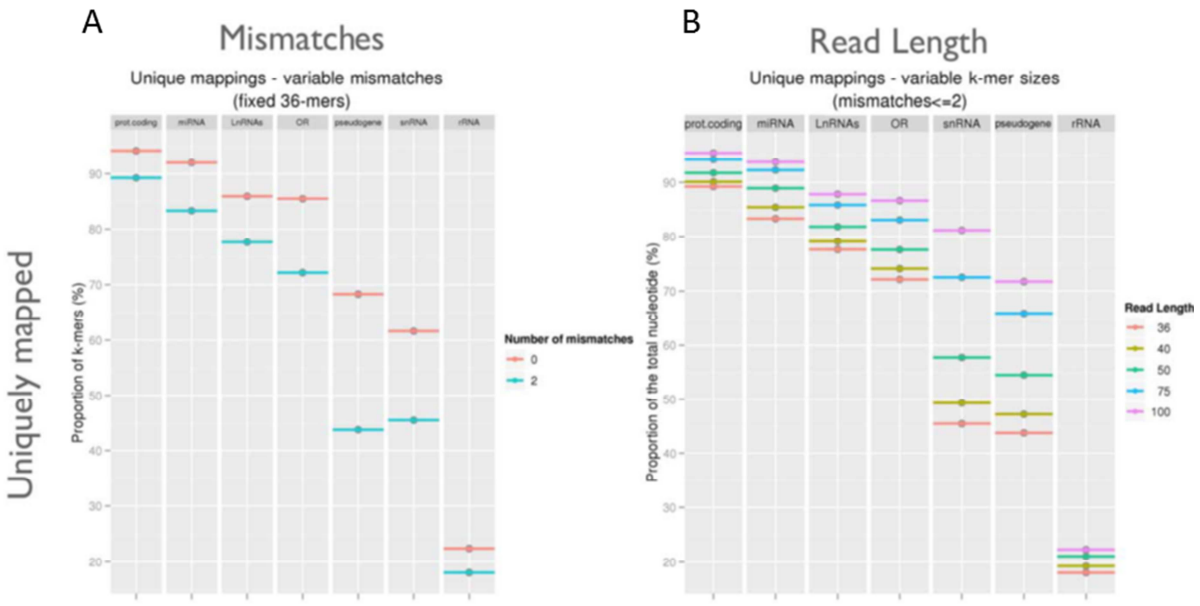


Figure 13:Influence du nombre de mésappariements et de la taille des lectures sur la mappabilité du transcriptome humain

(Derrien et al. 2012). A) Pourcentage de k-mers de 36pb qui s'alignent de façon unique avec un nombre de « mismatches » autorisé de 0 ou de 2 B) Pourcentage de nucléotides commençant une région entre 36 et 100pb autorisant 2 « mismatches » pour différents types de transcrits sur un transcriptome humain.

1.3.Objectifs

Nous allons, au cours de ce chapitre, construire une liste de gènes dupliqués présents chez l'homme et qui sera employée tout le long de ce travail de thèse. Nous nous m'appuyons pour cela sur deux travaux de recherche : l'article de *Chen et al., 2013* et celui *de Singh et al., 2014.*

Ces articles abordent différents aspects des gènes paralogues comme la datation et le type de duplication, mais aussi les groupent en familles de gènes.

Notre objectif de cette première partie du projet est de caractériser de manière exhaustive les paralogues de l'espèce humaine.

Dans un premier temps, nous mettons en place les objets de référence tels que la liste de gènes dupliqués et les familles de gènes paralogues, à partir des travaux de *Chen et al., 2013* et de *Singh et al., 2014* et nous intégrons également des informations complémentaires sur la longueur et la séquence des transcrits (homologie) ainsi que sur la fonction des gènes.

Ces informations nous permettent d'étudier les gènes dupliqués, au niveau évolutif en comparant les types de duplications WGD et SSD par rapport à la taille de leur famille ou

leur fonction. De plus, les résultats sur la mappabilité des pseudogènes nous amènent également à appréhender et à évaluer la problématique de l'identité de séquence entre les gènes dupliqués et les biais potentiels liés à l'utilisation des données d'expression de ces gènes

2.Matériel et méthodes

2.1.Liste de référence des gènes paralogues

2.1.1.Identification des gènes paralogues par Chen et al., 2013

La liste des gènes dupliqués définie par *Chen et al., 2013* correspond à la liste de référence des gènes paralogues qui sera utilisée pour tout le projet de thèse. Cette liste contient plusieurs types d'identifiants de gènes tels que les symboles HGNC (« Human Genome Nomenclature Committee ») (Gray et al. 2015) ou les identifiants Ensembl de la version d'annotation 59 du génome humain GRCh37 (Flicek et al. 2012). Cette liste de gènes paralogues a été construite à partir de la base de données TreeFam (version 8.0) (Ruan et al. 2008) qui permet de retrouver des familles de gènes et l'emplacement dans l'arbre phylogénétique où les évènements de duplication ont eu lieu. Cette version de TreeFam permet d'identifier 12346 gènes dupliqués regroupés en 3692 familles de gènes (identifiants HGNC) (Table 1). La liste donnée associe à chaque gène paralogue, sa paire de gènes dupliqués la plus récente.

TreeFam :

TreeFam (Ruan et al. 2008) est une base de données de EMBL-EBI contenant les familles de gènes de nombreuses espèces animales, de levure et de plantes (riz et arabette). La méthode de recherche de familles implémentée dans TreeFam se base sur une classification basée au départ sur une similarité de séquence entre gènes de différentes espèces ayant un génome de référence de bonne qualité (Ruan et al. 2008). Les clusters obtenus à partir de la classification sont considérés pour construire les familles incomplètes de gènes de référence (« seed families »). Ensuite, afin d'obtenir les familles complètes, les gènes des autres espèces sont ajoutés aux clusters de référence. Pour cela, un BLAST (McGinnis & Madden 2004) (voir Méthode Chapitre 2) est effectué sur chaque gène afin de retrouver une première liste d'orthologues potentiels basée sur l'alignement de séquences des transcrits. Ensuite l'algorithme HMMER (Zmasek & Eddy 2001) est utilisé afin de sélectionner, parmi cette liste, les homologues les plus

probables avec les gènes des clusters initiaux (Li et al. 2006). Les dernières versions de TreeFam s'assurent qu'un gène n'appartienne qu'à une seule famille de gènes. Uniquement le transcrit du gène avec le score d'homologie le plus élevé avec une famille est conservé (obtenu avec HMMER). Une dernière étape permet de définir un arbre phylogénétique de gènes. Afin d'obtenir cet arbre, un alignement multiple des séquences des gènes est effectué avec MUSCLE (Edgar 2004). La dernière version de TreeFam utilise cinq méthodes différentes de construction d'arbre. Ces méthodes sont basées sur l'algorithme du Neighbor-joining (Saitou & Nei 1987) et sur le maximum de vraisemblance (Gu 2001). Finalement, un arbre consensus est construit à partir de ces cinq arbres différents contenant les familles complètes de gènes. La particularité de TreeFam est qu'une fois la partie automatique terminée, une vérification manuelle des familles est effectuée. Les curateurs prennent en compte les annotations déjà publiées notamment sur les familles connues au niveau fonctionnel.

BLAST:

BLAST (« Basic Local Alignment Search Tool") est une méthode d'alignement local de séquences basée sur la recherche de similarités (McGinnis & Madden 2004). Cet outil permet d'établir un pourcentage d'identité entre des paires de séquences nucléotidiques ou protéiques. Afin d'optimiser cette méthode d'alignement de séquences, des heuristiques sont utilisées. Ces dernières recherchent des correspondances entre des courts fragments de la séquence (« k-mers ») puis allongent les alignements pour retrouver les régions qui s'alignent plutôt que de comparer directement les séquences entières. L'algorithme est basé sur des matrices de substitution. BLAST produit un pourcentage d'identité pour chaque région alignée et un score d'alignement (« Bit score »). L'outil donne également une statistique associée (e-valeur – « Expect value ») à chaque région alignée permettant de savoir si l'alignement est significatif (ex. e-valeur < 0,05). Plus le score d'alignement augmente plus la e-valeur diminue. TreeFam sélectionne les séquences qui s'alignent avec une e-valeur inférieure à 0,01. Ce seuillage sur la e-valeur est complétée par le score de HMMER et la vérification manuelle afin de permettre une prédiction robuste d'appartenance de gènes à une famille.

Inférence des paralogues :

La distinction des orthologues et des paralogues s'effectue aussi par TreeFam une fois l'arbre phylogénétique construit. Les gènes homologues sont donc retrouvés par BLAST et HMMER et les familles de gènes sont construites. La distinction entre les orthologues et les paralogues est effectuée par une méthode d'inférence de duplications et de pertes de gènes (DLI – « Duplication/Loss Inference ») (Li et al. 2006). Cet algorithme est basé sur le positionnement des évènements de spéciation et de duplication (SDI – « Speciation vs Duplication Inference ») (Zmasek & Eddy 2001). L'arbre des gènes final est donc comparé à l'arbre des espèces afin de situer les évènements de duplication et d'orthologie sur l'arbre phylogénétique des gènes. Cela explique les différences de topologie entre l'arbre des gènes et celui des espèces. Cette méthodologie minimise le nombre d'évènements de duplication. Cette étape est effectuée automatiquement par TreeFam, mais certaines duplications sont ajoutées par la suite lors de la vérification manuelle.

Datation de l'évènement de duplication :

Afin d'obtenir la datation des évènements de duplication, pour chaque paires de gènes dupliqués, le dernier ancêtre commun est récupéré à partir de l'arbre phylogénétique des gènes. Cet ancêtre identifié peut être commun à cette paire de gènes dupliqués mais également à des gènes orthologues. La totalité des gènes descendant de cet ancêtre sont identifié et sont associés à leur espèce d'appartenance. L'espèce de chaque gène est ensuite repérée sur l'arbre des espèces permettant de situer au niveau de l'arbre des espèces l'évènement de la duplication (Chen et al. 2013). La datation correspond donc à la longueur de branche de l'arbre phylogénétique des espèces entre l'évènement de la duplication et l'espèce humaine (Letunic & Bork 2011).

2.1.2.Annotation des types de duplications via Singh et al., 2014

La liste de gènes dupliqués de *Singh et al., 2014* contient des informations notamment sur le type de duplication ce qui nous permet de comparer les paralogues d'origine différente (Singh et al. 2014). Elle est composée de 15436 gènes dupliqués (HGNC)(Table 1) annotés à partir de la version 70 de la base de données Ensembl pour le génome humain de référence GRCh37.

Dans cette liste on retrouve 7075 gènes SSDs et 9916 WGDs (1422 gènes retenus àprès un évènement de SDD et de WGD). Les gènes restants sont considérés comme des singletons.

Les deux évènements de WGD de la lignée humain ont eu lieu il y environ 450 millions d'années au début de la lignée des vertébrés. Sachant la période des WGDs, il est possible de positionner les évènements de SSDs suivant s'ils ont eu lieu avant (oSSDs), après (ySSDs) ou pendant la période des WGDs (wSSDs). Ces trois périodes de datation des SSDs vont donc permettre de créer 3 sous catégories de SSDs : les ySSDs, les oSSDs et les wSSDs.

Cette seconde liste de référence va donc permettre de compléter la première liste de *Chen et al., 2013* avec ces nouvelles annotations concernant le type et leur catégories en fonction de la datation de leur évènement de duplication.

Identification des gènes issus d'une WGD :

La méthode utilisée pour retrouver les WGDs est celle de *Makino & McLysaght, 2010* (Makino & McLysaght 2010). Un alignement BLASTp « all-against-all » (alignement de toutes les paires de séquences entre elles) est effectué (uniquement les alignements avec une e-valeur < 1e-7 sont gardés) sur les séquences protéiques de plusieurs espèces (homme, poisson zèbre, tétraodon, épinoche, médaka, fugu et ascidie). Les gènes de l'homme et du poisson qui partagent le même meilleur alignement avec un gène de l'espèce ascidie (avant l'évènement de WGD) sont regroupés en clusters. Les clusters retenus contiennent au moins un gène chez le poisson et deux gènes chez l'homme. Au sein de chaque cluster, si le meilleur alignement avec un gène humain est un gène chez le poisson (après l'évènement de WGD), alors les gènes humains sont issus d'une WGD. Si le meilleur alignement concerne deux gènes humains, soit le gène chez le poisson a été perdu au cours de l'évolution et les deux gènes sont issus d'une WGD, soit le gène a subit un évènement de SSD après l'évènement de WGD (Figure 14).

Identification des gènes issus d'une SSD :

A la suite de l'identification des gènes WGDs, un Blastp « all-against-all » est effectué entre toutes les paires de gènes chez l'homme (Singh et al. 2012). Pour un gène donné, si un alignement significatif est retrouvé (e-valeur < 1e-7) avec un autre gène que lui-

même, ces deux gènes sont issus d'une SSD. Si un alignement est retrouvé avec un gène WGD alors il s'agit d'un WGD qui a subi un évènement de SSD (Figure 14).

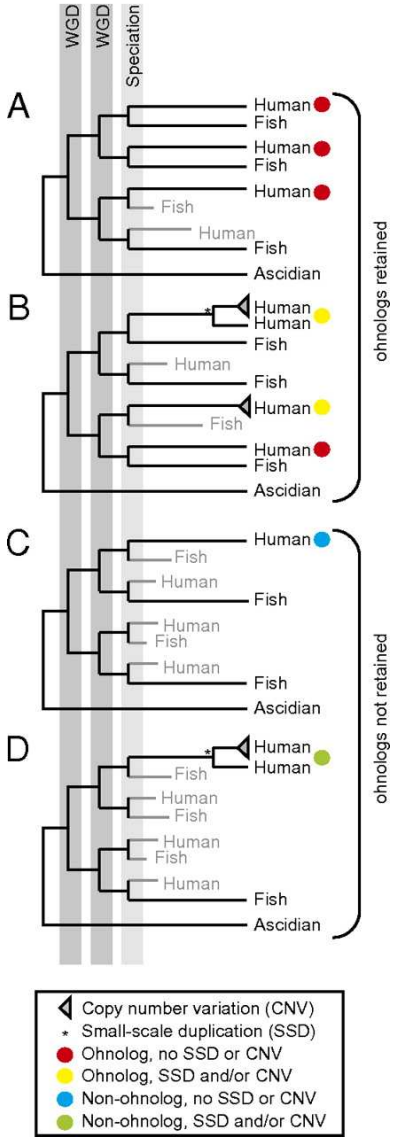


Figure 14: Identification des ohnologues (WGD) et des duplications SSD

Makino & McLysaght, 2010.Familles de gènes tétrapodes hypothétiques (A, B, C, D) où les membres de chaque famille ont été générés par WGD et/ou SSD. Les gènes sont étiquetés avec le nom de l'organisme. De nombreux gènes dupliqués par WGD (ohnologues) sont ensuite perdus, et ceux-ci sont indiqués en gris. (A et B) WGDs conservés. (C et D) WGDs non conservés. Les couleurs représentent les types de duplication associées (rouge : WGDs sans SSDs et CNVs (« Copy number variation » - variation du nombre de copies), jaune : WGDs qui ont connu une SSD ou une CNV, bleu : non WGDs sans SSD et CNV (singleton), vert : non WGDs qui ont connu une SSD ou une CNV.

2.1.3.Construction de notre liste de référence de gènes paralogues

Nous avons décidé de combiner les informations complémentaires sur les gènes paralogues contenues dans les travaux de *Chen et al., 2013* et *Singh et al., 2014* afin d'obtenir une ressource complète sur les gènes paralogues. En résumé, la liste des gènes paralogues, leur regroupement en familles de gènes et la datation de l'évènement de duplication employés pour nos travaux de thèse proviennent de *Chen et al., 2013* . Cette liste a été ensuite annotée avec les types et les datation sous forme de catégories de duplications de *Singh et al., 2014*. Les identifiants de gènes utilisés pour faire l'annotation sont les identifiants HGNC (Gray et al. 2015).

Ce travail nous a permis d'obtenir notre liste de référence définitive de gènes paralogues (Table 1).Nous avons également calculé et agrégé des informations supplémentaires pour ces gènes paralogues.

Table 1: Chiffres sur l'annotation à partir des trois listes de gènes dupliqués
Nombre de gènes (HGNC) ou de familles de gènes dans chacune des listes de référence.

	Chen et al., 2013	Singh et al., 2014	Liste de référence
Nb gènes	17805	20266	-
Nb familles	9616	-	
Nb paralogues	12346	15436	12346
Nb familles paralogues	3692	-	3692

Nous avons annoté notre collection de gènes paralogues constituée précédemment grâce à Ensemble BioMart afin d'obtenir les termes GO associés à chaque transcrit ainsi que la longueur et la séquence de leur ADNc.

Les informations sur l'ontologie des gènes se déclinent en trois domaines : les processus biologiques, les fonctions moléculaires et les composants cellulaires. Ces annotations ont été obtenues à partir de la base de données « Ensembl Gene » sous BioMart (Yates et al. 2016) pour les gènes humains correspondant à la version du génome GRch37 et à la version Ensembl 75.

2.2.Alignement de séquences des gènes paralogues

2.2.1.Alignements globaux et locaux

Les gènes paralogues étant issus d'un ancêtre commun, il est intéressant de se poser la question de la proportion de gènes qui présentent des séquences encore très proches.

L'identité de séquence entre deux gènes peut être obtenue à partir de l'alignement de leur séquence. Deux types d'alignements peuvent être utilisés : l'alignement local ou global. Un alignement global correspond à un alignement sur la totalité de la longueur des séquences, en permettant des substitutions, insertions et délétions, alors qu'un alignement local identifie des portions qui s'alignent au mieux entre les deux séquences.
Dans le projet, les alignements sont effectués à partir des séquences cDNA des transcrits nos gènes dupliqués groupés en familles de gènes.
Concernant les alignements globaux et locaux, des méthodes de « all-against-all » consistant à aligner par paire toutes les séquences entre elles ont été appliquées par famille de gènes. Une fois ces alignements effectués, nous avons supprimé ceux entre transcrits d'un même gène.
Pour les deux types d'alignements, seuls les cas où le pourcentage d'identité est supérieur à 75% sont sélectionnés car nous recherchons la proportion des paires de gènes ayant une forte identité de séquence. De plus pour l'alignement local, seules les régions supérieures à 80pb sont conservées en référence à la taille des lectures de séquençage RNA-seq. En effet, ces résultats vont nous permettre d'estimer les biais potentiels, liés aux gènes à forte similarité de séquence, pour l'alignement des données de séquençage RNA-seq. Nous allons également déterminer la proportion de gènes paralogues pouvant entraîner ces biais d'alignement.

Needle :

L'alignement global est réalisé par paire de séquences entières et son but est d'obtenir un score d'alignement global optimal. L'algorithme utilisé est Needlman & Wunch (Needleman & Wunsch 1970). Pour obtenir ce score et trouver la meilleure configuration pour l'alignement, il faut générer une matrice de scores entre les deux séquences calculée à partir d'une matrice de substitution associant des pénalités différentes pour les mésappariements (« mismatches »), suivant la nature de la substitution. Le calcul des scores prend aussi en compte les pénalités d'ouverture et d'extension de gap. A partir de la matrice des scores, le meilleur alignement gardé est celui qui donne le meilleur score.
L'outil Needle implémenté dans la suite Emboss version 6.6.0 (Rice et al. 2000) a été utilisé. Concernant les scores d'alignement, la matrice de substitution de nucléotide est EDNAFULL et les pénalités des gaps sont de 10 pour l'ouverture du gap et de 0.5 pour

son extension. Le résultat pour chaque alignement est contenu dans un fichier tabulé et seules les informations sur les scores, les longueurs d'alignement et les pourcentages d'identité sont conservées.

Blastall :

L'alignement local entre deux séquences de gènes recherche à optimiser l'alignement de certaines régions des deux séquences de gènes et non pas les séquences dans leur globalité afin de diminuer les pénalités engendrées par les portions de séquences non alignées. La méthode fréquemment utilisée pour l'alignement local est Smith & Waterman (Smith & Waterman 1981). Cet algorithme est basé sur celui de Needlman & Wunch mais au lieu de regarder chaque séquence dans son intégralité, l'algorithme d'alignement local compare des régions des séquences et choisit les longueurs de régions optimisant la mesure de similarité.

Pour l'alignement local au sein de nos familles de gènes, nous avons utilisé la méthode « all-against-all » avec l'outil Blastall, implémenté dans la suite BLAST (McGinnis & Madden 2004). BLAST compare chaque séquence à une base de données de séquences. Pour une famille de gènes donnée, la base de données correspond à tous les transcrits des gènes dupliqués appartenant à la famille. Comme il s'agit d'un alignement « all-against-all », chaque transcrit de la famille est comparé à la base de données pour retrouver des régions ayant une fort homologie avec les transcrits d'autres gènes de la famille. Précisément, pour le calcul des alignements, l'outil utilisé est Blastall version 2.2.26 et comme matrice de substitution correspond à une matrice de BLAST (1 pour les « matches » et -3 pour les « mismatches ») et les pénalités sont de 0 pour l'ouverture ou l'extension d'un gap

2.2.2.Calcul de la mappabilité

La mappabilité permet de prédire les régions uniques et répétées du génome (Derrien et al. 2012), pour une longueur de lecture donnée. Cette méthodologie va nous permettre de retrouver les gènes possédant ou non des régions qui s'alignent à différents endroit dans le génome. Nous estimerons ainsi les problèmes d'alignements non uniques des lectures de séquençage dans le génome ou dans le transcriptome humain.

Pour une lecture de séquençage de longueur k, la fréquence $F_k(x)$ pour la position x de la séquence de référence correspond au nombre de fois que le k-mer commençant à cette position x apparait dans la séquence de référence ou son complément-inverse. La recherche du k-mer dans la séquence de référence peut être exacte ou autoriser des variations (1 ou 2 « mismatches » par exemple).

La mappabilité, $M_k(x)$ à la position x correspond donc à l'inverse de la fréquence du k-mer commençant à cette position:

$$M_k(x) = \frac{1}{F_k(x)}$$

L'avantage de la mappabilité est qu'elle est comprise entre 0 et 1, 1 étant la seule valeur correspondant à un alignement avec placement unique du k-mer sur le génome.

Pour le projet, la mappabilité a été calculée pour le transcriptome des gènes codants pour les protéines et pour le transcriptome réduit aux gènes paralogues. Pour chaque gène, la séquence correspondant au transcrit le plus long a été récupérée à partir d'Ensembl BioMart. Ces séquences ont ensuite été regroupées dans un fichier multi-fasta pour chaque transcriptome.

Les outils employés pour les analyses de mappabilité sont implémentés dans la suite logicielle GEM (« GEnome Multitool ») (Derrien et al. 2012, http://algorithms.cnag.cat/wiki/The_GEM_library). Le programme gem-indexer a permis d'indexer la séquence fasta de chaque transcriptome. Afin de cartographier la mappabilité du transcriptome, le programme gem-mappability a été appliqué en considérant un k-mer de 100pb et une recherche exacte du k-mer, c'est-à-dire en n'autorisant aucun « mismatch ». Le fichier de mappabilité au format gem généré fournit pour chaque position nucléotidique du transcriptome un code qui correspond à la fréquence du k-mer ($F_k(x)$). Enfin, le programme gem-2-wig permet de convertir ce fichier gem au format wig.

Ces résultats de calcul de mappabilité des lectures ont été exploités pour distinguer les gènes ne souffrant d'aucune problématique d'alignement de ceux couverts avec des lectures à placements multiples sur le génome.

2.3.Test statistique de Welch

Le test de Student (t-test) est un test de comparaison de moyennes entre deux échantillons. Il permet de déterminer si ces moyennes sont significativement différentes.

Le test utilisé dans le projet de thèse correspond au test de Welch (Welch 1947), une variante du test de Student qui est adaptée à des échantillons de variances et de tailles différentes. Les tests statistiques sont calculés avec R. Dans ce qui suit nous testerons les moyennes de taille des familles de gènes. Nous considérons que la moyenne entre différents échantillons sera significative si la p-valeur calculée est inférieure à 0,05.
La p-valeur se retrouve à partir de la statistique t et du degré de liberté ν, par rapport à la distribution des t théoriques suivant une loi de Student.
Calcul de la statistique t:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

avec $\overline{X_i}$, la moyenne de l'échantillon i, s_i, l'écart type de l'échantillon i et N_i, la taille de l'échantillon i.
Calcul des degrés de liberté:

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2.(N_1 - 1)} + \frac{s_2^4}{N_2^2.(N_2 - 1)}}$$

3.Résultats

3.1.Gènes paralogues de référence

La première étape de ce projet est d'établir une liste de référence de gènes paralogues correspondant à celle de *Chen et al., 2013*. Chaque gène de cette liste est annoté par son identifiant Ensembl et HGNC. Pour chaque gène, nous avons l'information de sa paire de gène paralogue, correspondant à la duplication la plus récente avec le gène d'intérêt, avec sa datation en longueur de branche. Nous possédons également l'information de la famille à laquelle la paire de gènes appartient et la taille de cette famille. Les gène de cette liste sont ensuite annotés par l'information du type et de la datation leur duplication (Singh et al., 2014). Ces informations sur les gènes dupliqués seront utilisées pour étudier les familles des gènes appartenant à différents types de paralogues et à différentes catégorie de datation.

Plusieurs autres annotations sont également utilisées : l'implication des gènes dupliqués dans les processus biologiques et leur fonction moléculaire (GO) ou bien leur implication dans les maladies (ClinVar ou OMIM).

Enfin, les annotations telles que la longueur des transcrits et les alignements de séquence nous permettront d'interroger les mesures d'expression des gènes obtenus à partir des séquençages RNA-seq.

Ce tableau (Table 2) permet de récapituler la provenance des annotations des gènes dupliqués

Table 2: Sources des annotations de la liste de référence de gènes dupliqués

Récapitulatif des annotations présentes dans la liste de référence de gènes dupliqués. Les sources correspondent aux outils, aux bases de données et aux articles utilisés pour obtenir les annotations.

Sources	Annotations
Chen et al., 2013 Singh et al., 2014	Gènes paralogues
Chen et al., 2013	Familles de gènes
Singh et al., 2014	Types de duplication (WGD et SSD)
Chen et al., 2013 Singh et al., 2014	Datation des évènements de duplication (longueur de branche et catégories)
Ensembl BioMart	Séquence ADNc des transcrits
Ensembl BioMart	Longueur des transcrits
Ensembl BioMart	Nombre de transcrits par gène
Ensembl BioMart	GO
Needle	Identité de séquence – alignement global
BLAST (blastall)	Identité de séquence – alignement local

3.2.Caractérisation des familles de gènes

La première caractéristique d'une famille de gènes est sa taille, soit le nombre de gènes qui la composent. La taille des familles est variable mais elle est au minimum de deux gènes. Une famille de gène de taille 1 correspond à un singleton.

Les familles de gènes ont été collectés à partir des travaux de (Chen et al., 2013) reposant sur la méthode TreeFam (voir Méthode Chapitre 2).

Les familles de gènes de Chen et al., 2013 :

Parmi les 12346 gènes dupliqués, on compte 3692 familles de gènes (Figure 15). La taille de famille la plus faible est donc de 2 gènes et la plus élevée est de 57 gènes. La moyenne de la taille des familles est de 3.8 et la médiane est de 3. On observe un grand nombre de familles de taille 2 (1735 familles) ce qui correspond à 47% des familles de gènes. La majorité des familles sont donc de petite taille (3ème quartile: 4).

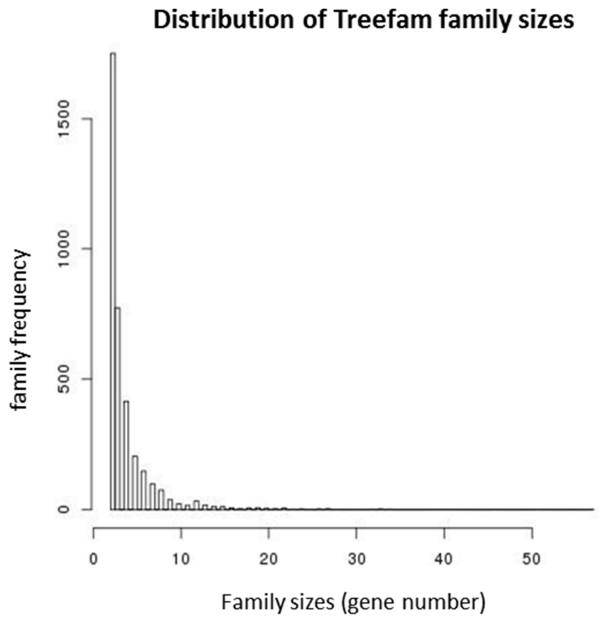


Figure 15: Taille des familles de gènes (TreeFam)

Distribution de la taille des familles de gènes obtenues à partir de *Chen et al., 2013*.

Taille des familles de gènes et type de duplication :

Le type de duplication associé à chaque gène paralogue a été collecté à partir des travaux de *Singh et al., 2014*. Les types de duplication correspondent aux WGDs et aux SSDs. Les SSDs sont aussi caractérisés en fonction de la période de leur évènement de duplication, c'est-à-dire avant (oSSD), après (ySSD) ou pendant (wSSD) la période de WGD. Dans ce chapitre, les gènes qui ont été retenus à la suite d'un évènement WGD et SSD appartiennent aux deux catégories.

Nous cherchons à savoir si la taille des familles est la même en fonction du type de duplication. En effet, nous avons observé pour chaque gène paralogue, la taille de la famille à laquelle il appartient en fonction du type de duplication dont il est issu. La Figure 15 illustre la distribution de la taille des familles en fonction du type de duplication (Figure 16). La taille minimale des familles pour toutes les catégories est de

2, puisque l'on s'intéresse uniquement aux gènes dupliqués, et la taille maximale est de 51 gènes (certains gènes de *Chen et al., 2013* n'ont pas l'annotation sur le type de duplication) retrouvée pour les SSDs.

La distribution des gènes WGDs est très proche de celle des SSDs. La médiane de ces deux distributions est identique (3 gènes). En revanche la moyenne est légèrement différente (4.0 et 4.5 respectivement). Les moyennes des tailles des familles pour ces deux catégories de gènes WGD et SSD sont significativement différentes (t-test de Welch, p-value=0.000228). Les tailles des familles des SSDs sont plus grandes que celles des WGDs.

Concernant les familles des gènes de type ySSD, nous visualisons qu'elles sont plus grandes (médiane : 4 et moyenne : 6.0) que pour les autres catégories de SSDs. Ainsi, la taille des familles de gènes chez les SSDs dépend de la date de l'évènement de duplication. Afin d'évaluer si la différence de taille de familles entre les gènes WGDs et les gènes SSDs peut s'expliquer par l'influence du type de duplication (au-delà d'une possible influence du fait que les duplications WGD ont un âge supérieur ou égal à la majorité des duplications SDD), nous comparons la moyenne des tailles des familles pour les gènes WGDs à celle des gènes wSSD (i.e. l'événement de duplication SSD a eu lieu à la même période que les WGDs) par un test de Welch. Les familles des wSSDs sont significativement plus grandes que les familles des WGDs (p-value = 8.819e-07, moyenne WGD = 4.04, moyenne wSSD = 4.75). La taille des familles dépend non seulement de l'âge de la duplication mais également du type d'évènement de duplication.

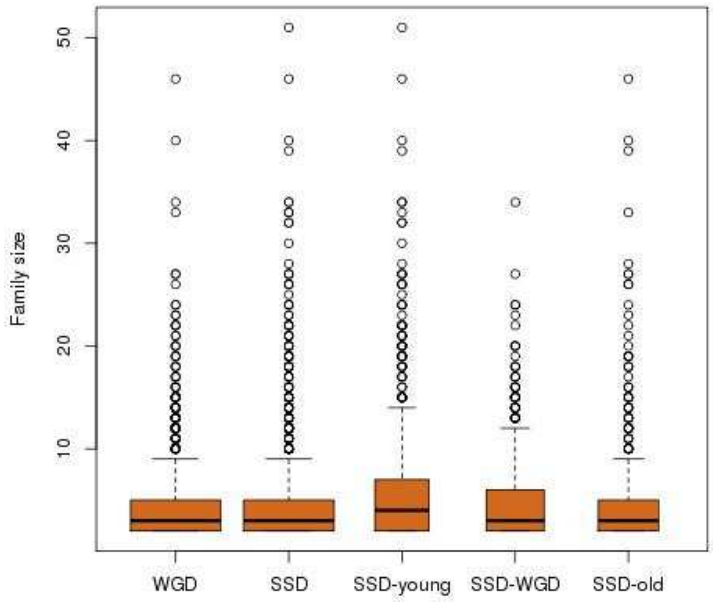


Figure 16: Taille des familles de gènes pour chaque type de duplication

Distribution de la taille des familles de gènes représentée sous forme de boîtes à moustache pour les gènes issus de chaque type de duplication (WGD, SSD, ySSD, wSSD, oSSD). La taille minimum des familles de gènes est de 2 et la taille maximum est de 51 gènes.

Fonction biologique des gènes par rapport au type de duplication :

Nous nous intéressons ensuite à l'ontologie des gènes dupliqués en évaluant la surreprésentation de certains termes reliés aux fonctions moléculaires et aux processus biologiques. Ces analyses ont été effectuées pour les différents types de duplication ainsi que pour les différentes catégories de datation des gènes SSDs avec l'outil GOStat sous R (Table 3).

Des différences importantes d'annotation en termes ontologiques peuvent être observées pour les deux catégories de gènes SSDs et WGDs. Les WGDs ont des fonctions moléculaires fortement liées à la régulation transcriptionnelle tandis que les SSDs sont plutôt associés à l'activité réceptrice des cellules et à la transduction du signal. Concernant les processus biologiques, les WGDs ont tendance à être impliqués dans des processus biologiques reliés aux développements anatomiques et nerveux.

Table 3: Ontologie des groupes de gènes de différents types de duplications

Termes ontologiques significativement (p-valeur < 0,05) associés à chaque type de duplication. Utilisation de l'outil GOStat sous R.

Référence **Gènes paralogues (14472)**	**Termes GO** **Fonctions moléculaires**
WGD (6038)	"DNA binding protein kinase", "transcription activity"
SSD (5170)	"Olfactory reception activity", "transmembrane signaling", "receptor activity"
SSD-younger (2212)	"Olfactory reception activity", "transmembrane signaling", "receptor activity"
SSD WGD-old (1565)	"Cytokine receptor activity", "extracellular matrix structural constituent", "purinergic receptor activity"
SSD-older (1395)	"Catalytic activity", "anion binding", "nucleotide binding", "ion binding"
	Processus biologiques
WGD (6088)	"System development", "anatomical structure development", "nervous system development", "anatomical structure morphogenesis"
SSD (5144)	"Detection of chemical stimulus", "sensory perception of smell"

3.3.Datations des évènements de duplications

Les datations des gènes dupliqués de notre liste de référence ont été obtenues de deux manières.

Dans l'article de *Chen et al., 2013*, les datations de ces évènements sont représentées par des longueurs de branches de l'arbre phylogénétique d'espèces entre l'évènement de la duplication et l'espèce humaine.

L'article de *Singh et al., 2014* propose quant à lui une datation en fonction du type de duplication. Les duplications de type WGD sont considérées comme un repère dans le temps (environ 450 Millions d'années). Concernant les duplications de type SSD, elles sont datées par rapport à la période des WGDs (ySSD, oSSD et wSSD)

Nous remarquons tout d'abord que parmi les SSDs, ce sont les ySSDs qui sont le plus représentés (42% des SSDs) (Figure 17).

Les datations de Singh et al, 2014 correspondent à des valeurs discrètes (catégories de SSD et de WGD) tandis que celles de Chen et al, 2013 à des valeurs continues (longueurs des branches des arbres phylogénétiques). Ainsi, il est intéressant de comparer ces deux sources de datation des duplications afin de s'assurer de leur cohérence.

Sur la Figure 17, sont représentés des histogrammes pour chaque catégorie de SSDs et pour les WGDs. Ces histogrammes représentent la fréquence des différentes longueurs de branche (par intervalles de 0.25). Concernant les WGDs, nous remarquons un pic pour des longueurs de branche autour de 1, par conséquent nous pouvons associer les WGDs à une longueur de branche de 1.

Pour chaque type de SSD, le pic de fréquence est cohérent avec sa catégorie. En effet pour les SSD récents le pic se situe à des longueurs de branches inférieures à 1 et inversement pour les SSD anciens. Concernant les SSDs datant de la période de WGD, le pic le plus élevé est également à 1.

Les deux approches donnent donc des résultats cohérents en ce qui concerne l'estimation de la datation des évènements de duplication.

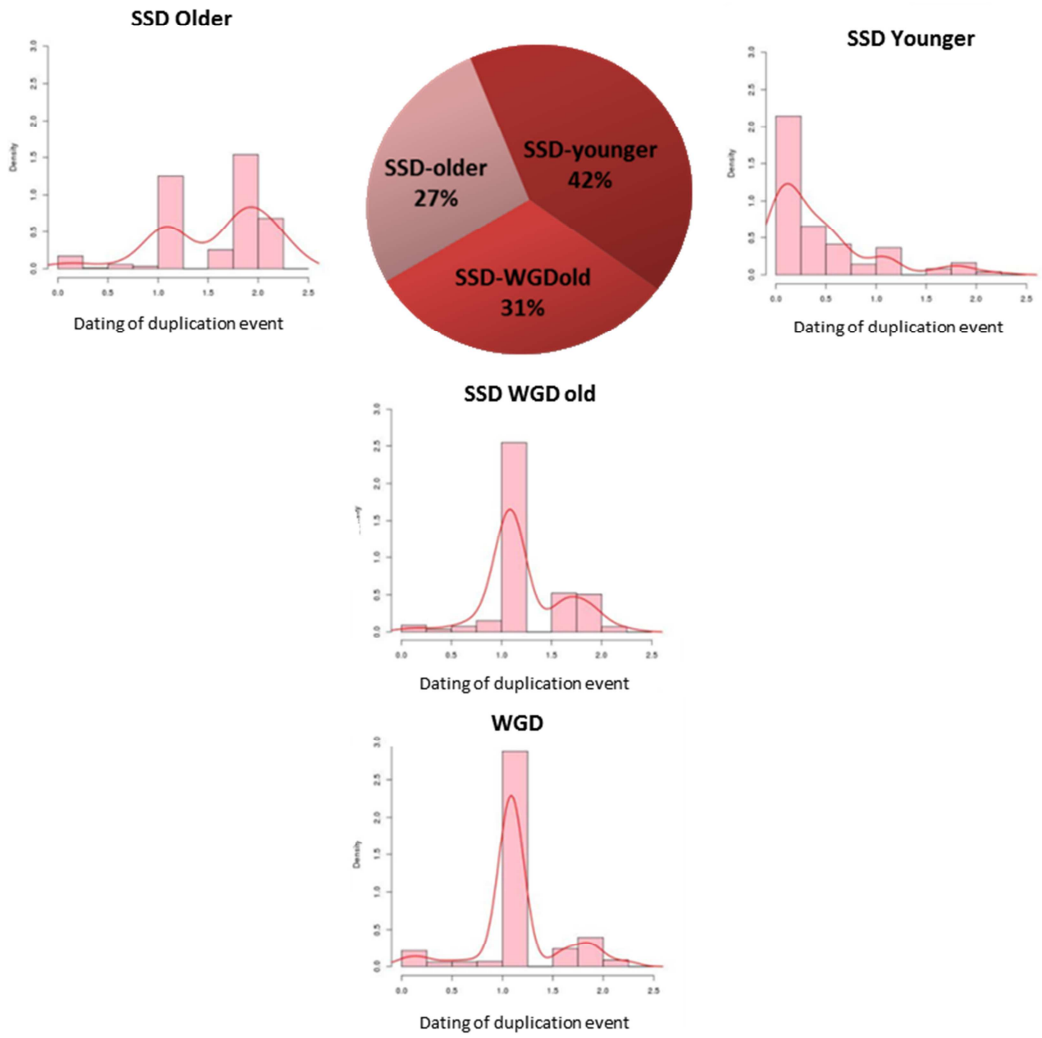


Figure 17: Datation des évènements de duplications

Répartition des SSDs en fonction de la datation de l'évènement de duplication. Histogrammes représentant la fréquence des longueurs de branche par intervalle de 0.25 pour chaque catégorie de SSD et pour les WGDs.

3.4.Caractérisation des séquences des gènes de même famille

3.4.1.Longueur des transcris des familles de gènes

Avant d'étudier les identités de séquences entre les gènes dupliqués, nous caractérisons d'abord cette identité au niveau des familles. La principale caractéristique en lien avec l'identité de séquence est la longueur des transcrits des gènes dupliqués.

Longueur des transcrits des gènes paralogues:

Nous considérons le transcrit le plus long, contenant le plus grand nombre d'exons, comme celui représentatif d'un gène dupliqué donné.

La longueur moyenne des transcrits les plus longs pour chaque gène dupliqué est de 3679pb et la médiane est de 3044pb (Figure 18). Sur l'histogramme nous observons que ces longueurs de transcrits sont très variables de 141 à 27220pb.

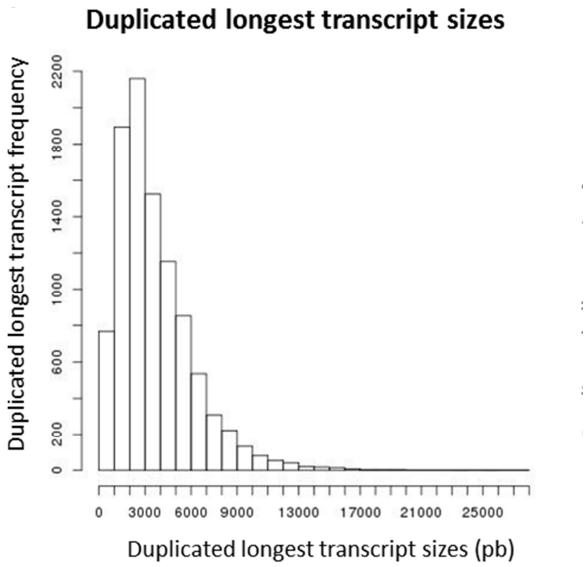


Figure 18: Taille des transcrits des gènes dupliqués

Distribution de la taille des séquences en paires de bases (pb) du transcrit le plus long par gène dupliqué .

Relation entre longueur des gènes des familles et leur identité de séquence:

Nous cherchons maintenant à étudier le lien entre la longueur des gènes contenus dans les familles et leur identité de séquence. Nous cherchons à évaluer si les familles sont composées de gènes de longueur variable et si ces gènes ont une forte identité de séquence entre eux.

La variabilité de la longueur des gènes est calculée par famille. En ce qui concerne l'estimation de l'identité des gènes d'une famille nous procédons à des alignements globaux par paire de gènes d'une même famille, en considérant le plus long transcrit par gène. Ce dernier point est réalisé avec l'outil Needle (voir Méthode Chapitre 2)

La Figure 19 représente la moyenne des pourcentages d'identité calculés entre paires de transcrits par famille (un transcrit par gène) en fonction de l'écart-type de la longueur de ces transcrits. A partir de tous ces points, on calcule une courbe de tendance de régression linéaire locale (« loess ») (Cleveland 1979). Nous remarquons que l'augmentation de la variance de la longueur des gènes par famille s'accompagne d'une baisse de leur identité de séquence. Un test de corrélation entre ces deux variables identifie une anti-corrélation significative de -0.33 (p-value < 2.2e-16) confirmant cette

tendance. Cette anti-corrélation reflète un biais de l'alignement global car plus la différence de longueur des gènes d'une famille est élevée, plus l'identité globale risque de baisser. En revanche, pour les familles avec un écart type de longueur des gènes relativement faible (< 1000 pb), certaines de ces familles ont une identité de séquence très élevée (jusqu'à 100%) ce qui permet de mettre en évidence la très forte identité de séquence entre certains gènes paralogues.

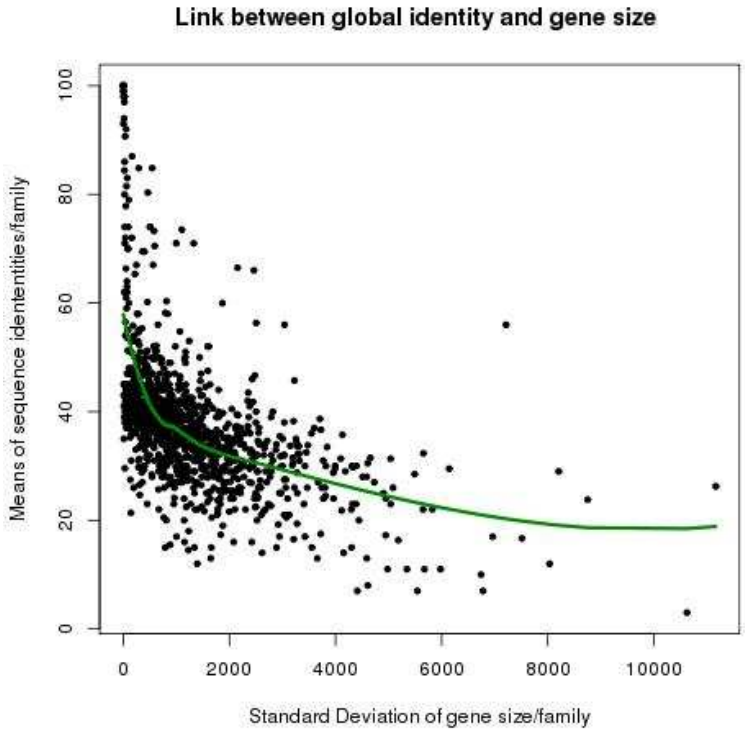


Figure 19: Variabilité de la longueur des gènes des familles et leur identité

Moyenne des pourcentages d'identité entre les transcrits les plus longs de gènes différents d'une même famille en fonction de l'écart-type de la longueur du transcrit le plus long de chaque gène par famille de gènes. Courbe de tendance par loess (Cleveland 1979) (courbe verte).

3.4.2.Identification des gènes à forte homologie de séquence et des régions avec un alignement multiple sur un transcriptome de référence

Les familles de gènes sont composées de gènes de longueur et de séquence plutôt proches. Cette forte identité de séquence laisse à penser que le transcriptome des gènes dupliqués pourrait contenir des régions donc problématiques pour estimer correctement les valeurs d'abondance des gènes associés.

Nous cherchons donc à connaître la proportion de gènes dupliqués dont la forte identité de séquence peut poser des difficultés pour estimer leur mesure d'abondance. Cette question est abordée avec deux approches, l'alignement local entre gènes d'une même famille et la mappabilité.

Alignement local entre transcrits de gènes d'une même famille :

Un bon estimateur permettant d'identifier les gènes qui ont des régions partagées correspond à l'alignement local de toutes les paires de gènes (voir Méthode Chapitre 2). Ces calculs d'identité de séquence locale sont réalisés par paire de transcrits des gènes appartenant à une même famille. Parmi tous les alignements nous avons gardé uniquement ceux générant un pourcentage d'identité supérieur à 75% pour une région alignée supérieure à 80pb. Ces paires de transcrits couvrent 66% des gènes dupliqués soit 8205 gènes. Sur la Figure 20, la ligne verte représente la médiane à 87% d'identité de séquence. Les paires de transcrits en rouge (915 gènes) possèdent des régions avec un pourcentage d'identité supérieur à 98%. Ces paires de transcrits sont donc difficiles à différentier et risquent d'être problématiques pour la mesure d'expression des gènes dupliqués à partir des données RNA-Seq.

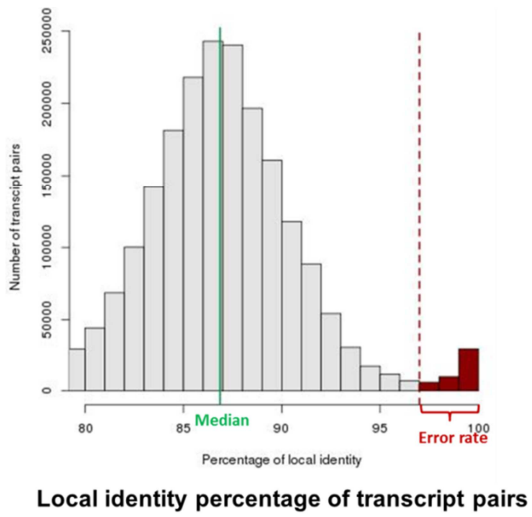


Figure 20: Distribution des identités locales entre paires de transcrits de différents gènes d'une même famille

Alignements locaux entre régions de transcrits de gènes différents d'une même famille A) Distribution du pourcentage des identités locales. Les régions représentées ont un pourcentage d'identité supérieur à 75% et sont de longueur variable (de 80pb à la totalité du transcrit).

Mappabilité des gènes sur un transcriptome total :

La mappabilité (voir Méthode Chapitre 2) est calculée à partir de régions de 100pb afin de simuler un alignement de lectures RNA-Seq sans mésappariement.

Dans un premier temps, la mappabilité est calculée sur un transcriptome total (totalité des gènes transcrits) pour plusieurs biotypes qui sont les gènes codants pour des protéines, les gènes dupliqués et les pseudogènes.

Les nucléotides ayant une mappabilité de 1 appartiennent à une région de 100pb avec placement unique dans le transcriptome de référence. Nous trouvons que 84% des nucléotides des gènes dupliqués ont une capacité d'alignement unique. Ce pourcentage s'élève à 82% pour les gènes codants pour des protéines et 73% pour les pseudogènes.

Mappabilité des gènes dupliqués sur un transcriptome de gènes dupliqués :

Afin d'approfondir les résultats obtenus à partir des alignements locaux entre séquences de transcrits d'une même famille de gènes, il est possible de calculer une nouvelle fois la mappabilité à partir de régions de 100pb mais cette fois sur un transcriptome restreint aux gènes dupliqués.

Nous observons (Figure 21) que la majorité des gènes dupliqués ont des régions avec une mappabilité égale à 1. Toutes les séquences de 100pb de ces gènes sont donc uniques dans le transcriptome de référence et représentent 88% des nucléotides des gènes dupliqués.

Nous trouvons néanmoins 2191 gènes dupliqués ayant au moins une région de 100pb non-unique et donc posant des problèmes d'alignement de séquences RNA-seq.

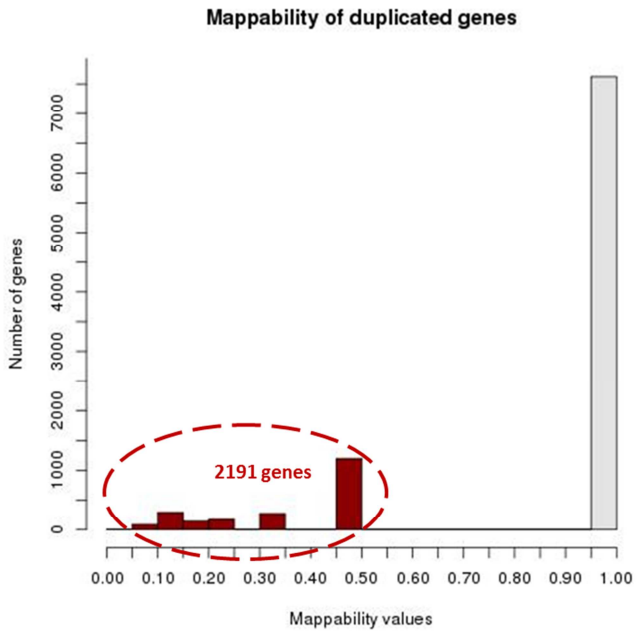


Figure 21: Mappabilité des gènes dupliqués

Distribution des valeurs de mappabilité (comprises entre 0 et 1) parmi les gènes dupliqués (pour des fenêtres de 100bp et seule la région du gène qui possède la mappabilité la plus faible est considérée). En gris sont représentés les gènes dont la région de plus faible mappabilité est de 1 c'est-dire que toutes les séquences de 100pb du gène sont uniques.

4.Discussion

En nous appuyant sur différents travaux de recherche (Chen et al. 2013; Singh et al. 2014), nous avons construit une liste de référence de gènes paralogues regroupant les informations sur les familles de gènes mais aussi le type de duplication associé à chaque gène, ainsi que les datations de type continues (longueur de branches) ou par catégories. Nous avons également ajouté des informations comme la longueur des transcrits (au sens du plus long transcrit) de chaque gène mais aussi celles sur les homologies de séquences entre paires de gènes au sein d'une même famille. Toutes ces annotations et ces caractéristiques donnent une ressource complète sur les gènes paralogues. Cette dernière va donc nous permettre d'interroger les caractéristiques évolutives des gènes paralogues exprimés dans les différentes régions du cerveau.

Nous avons aussi comparé les tailles des familles de gènes pour différents types et différentes datations de gènes paralogues. Ainsi nous avons déterminé que les plus petites familles sont celles des WGDs. L'article *d'Acharya & Ghosh, 2016* montre que les WGD sont plus conservés que les SSD en terme de séquence, ce qui permet aux WGD de conserver les nouvelles fonctions qu'ils tendent à acquérir après la duplication. Grâce

aux comparaisons sur la taille des familles de gènes, nous pouvons ajouter que les familles des WGDs sont plus conservées par rapport aux familles des SSDs. En effet si la taille des familles des WGD est plus faible, cela indique que les duplications de type SSD d'un gène WGD risquent de ne pas être retenues. Nous avons aussi trouvé que la taille la plus élevée des familles de gènes concernait les duplications récentes (ySSD). Par ailleurs nous avons observé que plus la duplication était ancienne, plus la taille de la famille dont elle faisait partie était petite. En effet cela semble pouvoir s'expliquer par la pression de sélection qui n'aurait pas encore eu lieu pour certaines duplications récentes. Au cours du temps, de nombreux gènes qui sont encore dans notre génome risquent de ne pas être retenus et de se pseudogéniser.

Concernant notre étude sur les alignements globaux des transcrits les plus longs des gènes paralogues, nous avons remarqué, une anti-corrélation significative entre l'identité de séquence et la variance de la longueur des gènes d'une même famille. Ainsi, plus la longueur des gènes d'une famille varient, plus l'identité de séquence aura tendance à être faible. Cette anti-corrélation semble s'apparenter un biais lié à la divergence de la taille des transcrit car pour un alignement global, plus la différence de tailles entre les séquences testées sont différentes, plus le score d'alignement va diminuer. A cause de ce biais, nous nous somme intéressé aux alignements locaux des transcrits les plus longs des gènes paralogues. A partir des pourcentages d'identité de séquence obtenus avec les alignements locaux, nous avons pu mettre en évidence qu'au sein des gènes paralogues de mêmes familles, certains possédaient des régions proches en séquence ou mêmes identiques avec un autre gène de la famille. Ce résultat met en évidence un biais qui pourrait se produire en analyse de données de séquençage notamment pour la mesure d'abondance de la production en ARNm des gènes dupliqués à partir d'un séquençage RNA-seq. Afin de comprendre l'impact de cette forte identité de certaines régions des séquences des gènes paralogues, nous avons utilisé la notion de mappabilité qui permet d'identifier les régions uniques ou répétées dans le génome ou bien dans le transcriptome. Nous avons donc retrouvé 2191 gènes, soit 18% des paralogues qui possédaient au moins une région de 100pb qui n'était pas unique dans le génome. Ainsi l'hypothèse du biais se produisant sur la mesure d'abondance à partir d'un séquençage RNA-seq est validée et peut impliquer au moins 18% des paralogues dans un cas d'alignement sans autorisation de mésappariements. Les valeurs

d'expression mesurées pour ces paralogues auraient donc tendance à être sous-estimées par rapport aux gènes qui ne possèdent pas de régions communes.

Une étude qui s'intéresse à la mesure d'expression des gènes paralogues (Lan & Pritchard 2016) met en lumière le problème de la forte identité des séquences des gènes paralogues surtout ceux de duplication récente. Pour les paires de gènes très similaires, les lectures de séquençages RNA-seq vont s'aligner sur les deux gènes. Mais dans le cas d'autorisation d'alignements multiples, il sera difficile d'obtenir une estimation de l'expression des gènes précise ce qui peut être problématique pour étudier les variations d'expression entre gènes (étude des asymétries d'expression et tissu-spécificité des gènes). Ils privilégient donc un alignement unique des lectures de séquençage. Ainsi les ratios des mesures d'abondance obtenus entre paires de paralogues seront plus fiables malgré un risque de la baisse d'expression globale principalement pour les gènes paralogues.

5.Conclusion

Ce chapitre décrit le processus de construction d'une ressource complète sur les gènes paralogues. Cette ressource a été caractérisée du point du vue de la datation des évènements de duplication, de la taille des familles de gènes, de la longueur de leurs transcrits, de l'identité de séquence globale et locale entre gènes de même famille et de leur mappabilité sur un transcriptome de référence.

Nous avons notamment pu mettre en évidence que la forte identité de séquence entre paralogues est un biais à prendre en compte pour la mesure d'expression des gènes paralogues par RNA-seq. Une application potentielle de cette ressource serait de créer une base de données en ligne permettant aux chercheurs de récupérer facilement des annotations évolutives sur les gènes paralogues.

Partie 2 : Expression tissulaire et co-expression des gènes paralogues dans le cerveau humain

Chapitre 3 : La tissu-spécificité d'expression des gènes dans différentes régions du cerveau

1.Introduction

1.1.Expression tissulaire des gènes

Le consortium GTEx :

Le consortium GTEx (Genotype-Tissue Expression) (Ardlie et al. 2015) a étudié entre autre l'expression des gènes dans différents tissus à partir du séquençage RNA-seq. Ces tissus ont été prélevés post-mortem sur une large cohorte d'individus non-porteurs d'une pathologie avérée. Chaque échantillon est caractérisé par des mesures de comptages de lectures par gène et des mesures d'abondance normalisées en RPKM (« Reads Per Kilobase per Million mapped reads »). A partir de ces données d'expression, une classification hiérarchique a été effectuée sur des échantillons (Figure 22) provenant de 24 tissus humains d'organes différents et de 13 tissus cérébraux. Sur cette figure nous pouvons constater que, l'expression des gènes permet de regrouper les échantillons provenant des mêmes tissus. Concernant les tissus cérébraux, bien que la classification produite par GTEx soit plus difficile à interpréter il apparaît néanmoins que les échantillons du cervelet et du cortex semblent se regrouper par tissu. Cette classification des échantillons par tissu suggère une spécificité tissulaire d'expression de certains gènes.

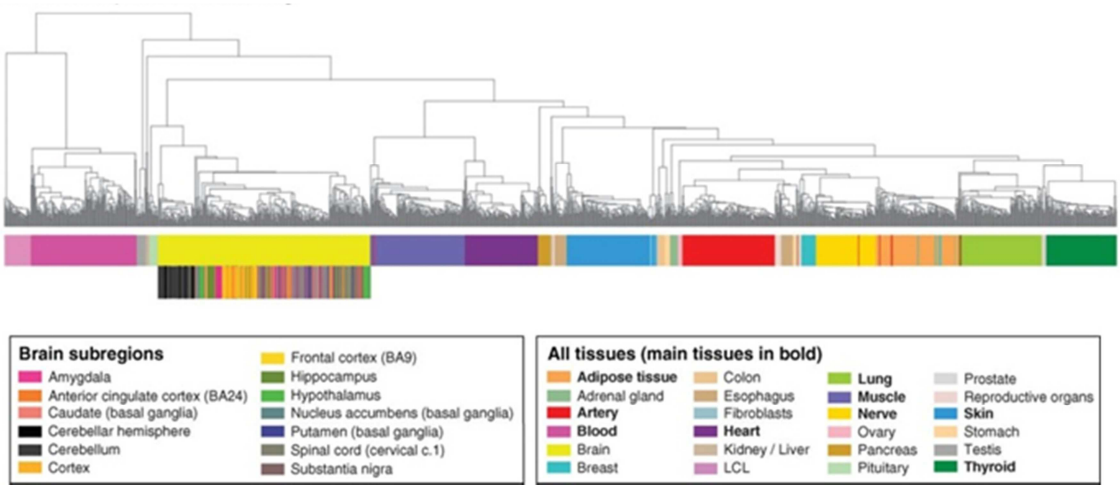


Figure 22: Classification des échantillons GTEX basée sur l'expression des gènes

(Ardlie et al. 2015). Classification hiérarchique des échantillons du consortium GTEx issus de différents tissus humains : 24 tissus d'organes différents et 13 tissus du cerveau. La classification est basée uniquement sur l'expression des gènes dans chacun des échantillons.

La famille *SRGAP2* :

La famille *SRGAP2* (Slit-Robo Rho GTPase activating protein 2) est une famille de gènes paralogues, dupliqués à plusieurs reprises (*SRGAP2A, SRGAP2B, SRGAP2C, SRGAP2D*) à des périodes différentes (Dennis et al. 2012). Les duplications de *SRGAP2*, sont spécifiques à l'homme. L'expression des différents gènes de la famille a pu être observée chez l'homme, chez le chimpanzé et chez le macaque (Figure 23). La visualisation de l'expression de ces gènes en RPKM dans différents tissus illustre une expression plus élevée de certains paralogues de *SRGAP2* dans un tissu en particulier (Figure 23).

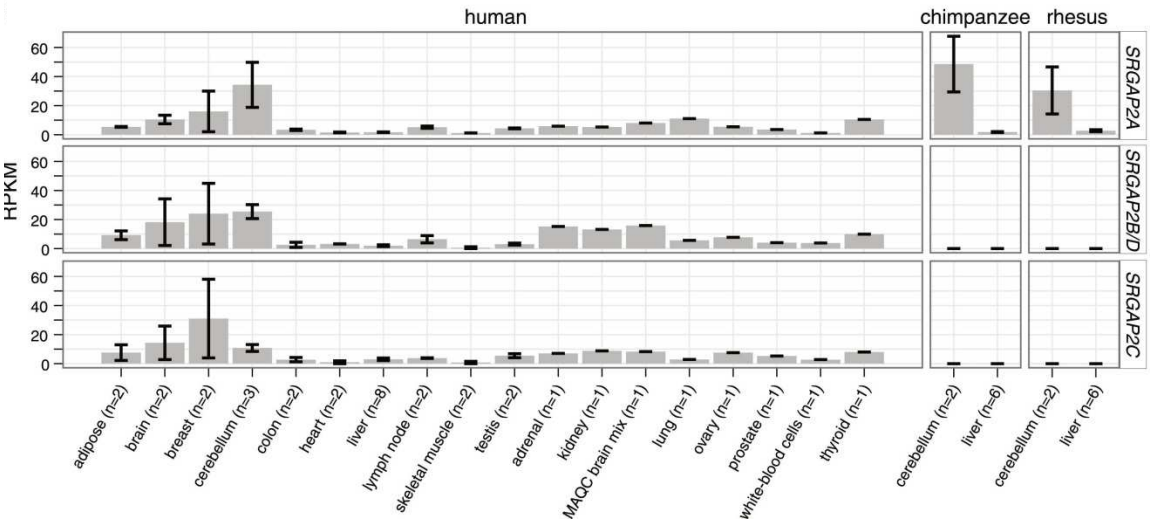


Figure 23: Expression des gènes de la famille SRGAP2 dans différents tissus

(Dennis et al. 2012). Expression des gènes de la famille *SRGAP2* (*SRGAP2A*, *SRGAP2B/D* et *SRGAP2C*) dans différents tissus chez l'homme le chimpanzé et le macaque rhésus.

1.2.Spécificité tissulaire des gènes paralogues

Les gènes peuvent avoir un profil d'expression ubiquitaire (gènes de ménage) au sein de différents tissus ou bien avoir un profil plus tissu-spécifique (exprimés dans un seul ou un nombre limité de tissus) (Kryuchkova-Mostacci & Robinson-Rechavi 2016).

Les gènes paralogues d'une même famille peuvent s'exprimer dans des tissus différents (Figure 24 A). En effet ces gènes semblent conservés dans le génome sur le long terme par le biais d'une divergence de leur expression ou de leur fonction au cours de l'évolution (Lan & Pritchard 2016). Il est alors considéré que ces gènes ont été conservés par leur sous- ou néo-fonctionnalisation. D'autres gènes paralogues vont par exemple

montrer une expression asymétrique au travers des tissus ce qui n'est pas en faveur de la rétention du gène porteur d'une expression globale faible (Figure 24B).

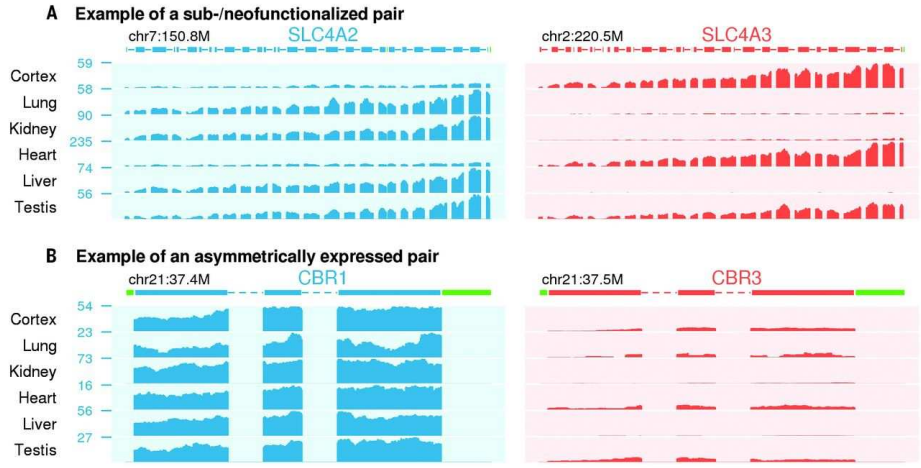


Figure 24: Profils d'expression de gènes paralogues dans 6 tissus humains

(Lan & Pritchard 2016). A) Paires de gènes dupliqués se partageant l'expression tissulaire, chaque gène est spécifique à au moins un tissu. B) Expression asymétrique avec un gène globalement plus exprimé dans tous les tissus que le second.

1.3.Evaluation des de prédiction des gènes tissu-spécifiques

L'observation de profils d'expression tissu-spécifiques laisse présager de la possibilité d'estimer cette spécificité tissulaire d'un gène à partir d'une métrique calculée sur la base de l'expression d'un gène à travers différents tissus. Des travaux récents (Kryuchkova & Robinson-Rechavi 2017) ont permis de comparer plusieurs méthodes de calcul de scores de tissu-spécificité. Tout d'abord, les scores de chaque méthode sont transformés afin que leur gamme de valeurs se situe entre 0 et 1. Une valeur de 0 indique une expression de gène ubiquitaire tandis qu'une valeur de 1 est associée à une expression tissu-spécifique (Figure 25). Cette étude a été réalisée sur 27 tissus humains provenant de différents organes. Une distribution bimodale peut être observée pour la totalité des méthodes. De plus, la proportion de gènes tissu-spécifiques (scores proches de 1) est variable en fonction des méthodes. Cependant, le score τ (voir Méthode Chapitre 3), semble identifier le plus grand nombre de gènes tissu-spécifiques. Des diagrammes de Venn permettent de comparer le nombre de gènes tissu-spécifiques identifiés pour chaque méthode (Figure 26). Le score τ permet d'identifier la quasi-totalité des gènes tissu-spécifiques prédits par les autres méthodes. En effet, seul le z-score prédit des gènes tissu-spécifiques non retrouvés avec τ. Par conséquent, τ semble

être la méthode la plus sensible pour prédire des spécificités tissulaires de l'expression des gènes.

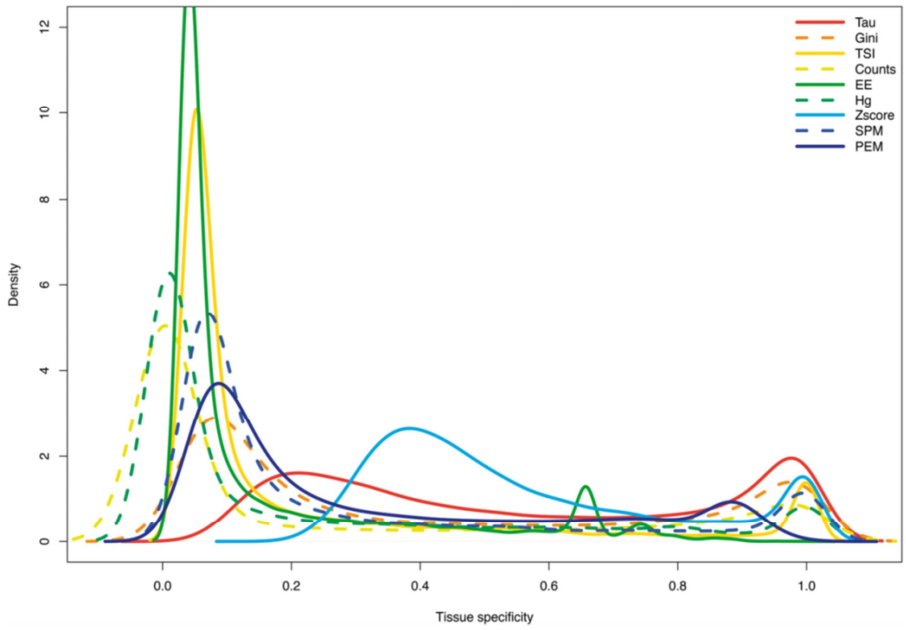


Figure 25: Distribution de différents paramètres de tissue-spécificité

Distribution des scores de tissu-spécificité d'expression des gènes obtenus par différentes méthodes à partir de données d'expression de 27 tissus humains. Kryuchkova-Mostacci & Robinson-Réchavi, 2017.

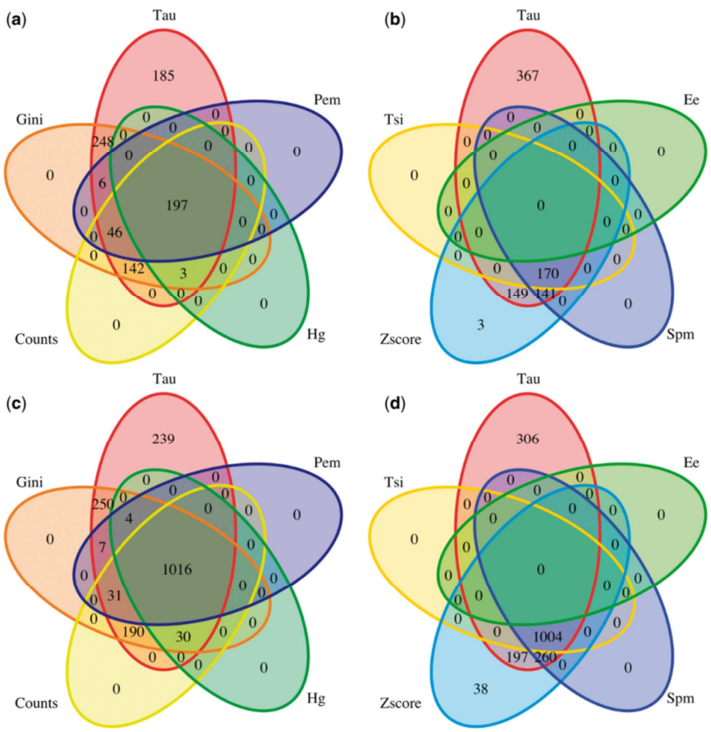


Figure 26: iagrammes de Venn du nombre de gènes prédits pour chaque méthode de calcul de tissu-spécificité

Kryuchkova-Mostacci & Robinson-Réchavi, 2017. A et B) gènes tissu-spécifiques avec la plus forte expression dans le cerveau, C et D) gènes tissu-spécifiques avec la plus forte expression dans les testicules.

1.4.Objectif

Les études d'expression tissulaire des gènes paralogues déjà publiées ayant été effectuées au travers de l'expression de ces gènes dans différents organes, nous nous sommes intéressés au cours de cette thèse à leur expression dans différentes régions du cerveau. Nous cherchons à définir si les gènes paralogues ont une expression plus tissu-spécifiques au sein du cerveau humain que les singletons.

Nous allons tout d'abord comparer la classification des échantillons de tissus cérébraux obtenue à partir de l'expression des gènes dupliqués à celles produites à partir des autres gènes, ce qui permettra de refléter de potentielles différences de spécificité tissulaire. Ensuite, nous effectuerons des analyses permettant de déterminer la proportion de gènes paralogues différentiellement exprimés entre les tissus cérébraux. Enfin, nous utiliserons le score τ pour identifier les gènes spécifiquement exprimés dans une région cérébrale.

2.Matériel et méthodes

2.1.Données d'expression de GTEx pour le cerveau humain

Données GTEx :

Les données d'expression RNA-seq utilisées dans le projet sont celles produites par le Consortium GTEx pour 13 tissus cérébraux humains : le cervelet, l'hémisphère cérébelleux, le cortex, le cortex frontal, le cortex cingulaire antérieur, l'hypothalamus, l'hippocampe, la moelle épinière, l'amygdale, le putamen, le noyau caudé, le noyau accumbens et la substance noire (Figure 27 et Figure 28) (Ardlie et al. 2015). Le consortium a analysé un total de 1259 biopsies post-mortem par séquençage RNA-seq. Sachant que plusieurs biopsies ont été prélevées chez les mêmes individus, il y a entre 63 et 125 individus analysés pour chaque tissu. Les données d'expression collectées auprès des ressources du consortium sont des comptages de lectures par gène et des valeurs d'abondance normalisées en RPKM. Ces deux types de mesure d'abondance ont été générés à partir d'alignements de lectures RNA-seq effectués par GTEx en ne conservant que les lectures qui s'alignent avec placement unique sur le génome.

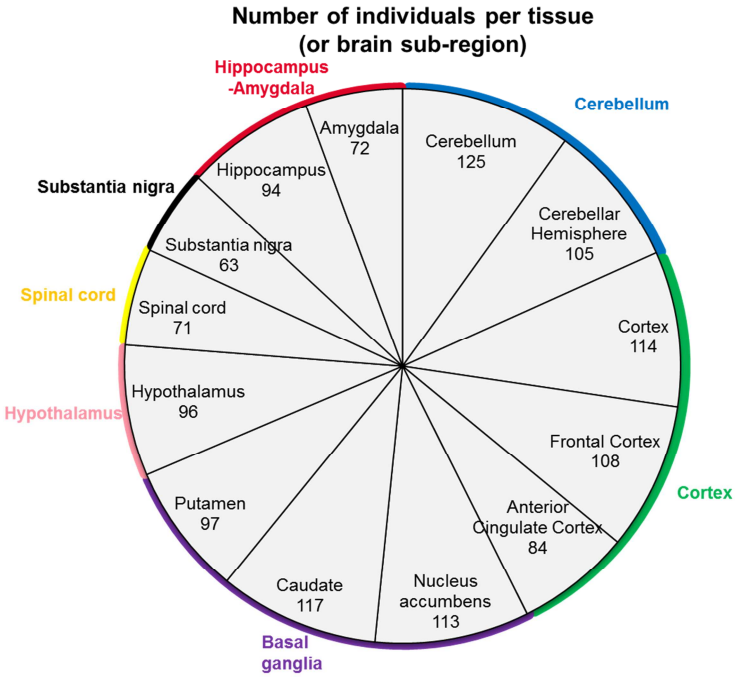
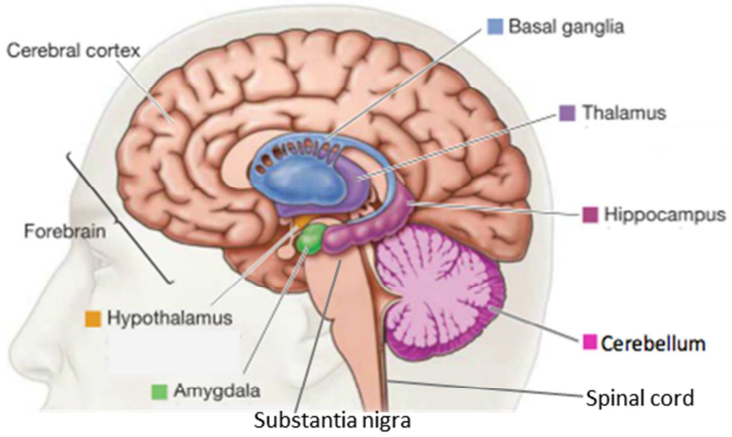


Figure 27: Nombre d'individus par tissu cérébral

Nombre d'individu par tissu cérébral du consortium GTEx. Répartition des tissus par rapport aux régions cérébrales auxquels ils appartiennent.



http://www.proprofs.com/flashcards/story.php?title=intro-mind-and-brain--topic-2-foundation-brains

Figure 28: Visualisation des régions cérébrales utilisées

Prétraitement des données d'expression :

Tout d'abord un filtre est appliqué sur les faibles valeurs d'expression de gènes. Nous éliminons de notre étude les gènes ayant des moyennes d'expression par tissu systématiquement inférieures à 0.1 RPKM. Cela réduit notre collection de données d'expression à 16427 gènes codants pour des protéines répartis en 10335 gènes

paralogues et 6092 gènes singletons. Les gènes paralogues se distribuent quant à eux entre 5114 gènes WGDs, 3719 SSDs dont 1192 ySSDs, 1260 wSSDs et 1267 oSSDs.

Ensuite, les données sont transformées en logarithme (log-transformées) avec la formule suivante :

$$log_2(RPKM + 1)$$

Cette transformation permet notamment d'obtenir des valeurs d'expression positives (voir Résultats Chapitre 3).

L'ajustement des données d'expression de gènes est réalisé afin de prendre en compte plusieurs effets « batch » (effets de lot) pouvant influencer ces valeurs d'abondance. Nous utilisons les mesures des effets « batch » fournis par GTEx, à savoir, des effets techniques, dus à la plateforme de séquençage utilisée, et des effets biologiques, correspondant à l'âge des individus, à leur genre et aux 3 premières composantes principales sur les données de génotypage illustrant la structure de population. Nous réalisons l'ajustement des données d'expression de gènes grâce à une régression linéaire combinant les différents effets:

$$lm(expData \sim Plateforme + Age + Genre + C1 + C2 + C3)$$

Dans ce modèle linéaire, les données résidualisées ne sont pas centrées par l'ordonnée à l'origine afin de conserver les différences de moyennes d'expression entre les gènes.

2.2.Classification hiérarchique des gènes et des échantillons

La classification hiérarchique est une méthode non supervisée permettant de regrouper des observations en fonction de certains critères de similarité (Székely & Rizzo 2005). Les analyses basées sur l'expression de gènes cherchent souvent à regrouper soit les gènes ayant des profils d'expression similaires à travers les échantillons, soit les échantillons présentant des expressions proches à travers les gènes.

Nous effectuons une classification hiérarchique des gènes et des échantillons, à partir des données d'expression de GTEx, log-transformées et ajustées pour les effets techniques et biologiques.

Cette classification est réalisée avec la librairie R « pheatmap » paramétrée pour utiliser la corrélation de Pearson comme mesure de similarité et l'UPGMA (« Unweighted Pair Group Method with Arithmetic Mean ») comme méthode de regroupement des gènes ou des échantillons.

Comparaison des classifications hiérarchiques :

Nous effectuons des classifications hiérarchiques en partant de différents groupes de gènes (gènes codant pour des protéines, singletons et gènes paralogues) afin d'évaluer si l'expression de ces gènes permet de regrouper les échantillons par tissu cérébral. Cette évaluation repose sur la comparaison des classifications d'échantillons produites à partir de ces différents groupes de gènes à celle attendue par la connaissance de l'appartenance des échantillons aux tissus.

Nous utilisons une première approche de comparaison entre les groupes d'échantillons prédits et attendus par le calcul de l'indice de Rand ajusté (Rand 1971). Cet indice est basé sur le nombre de paires d'échantillons au sein de chaque groupe dans la classification attendue qui ne sont pas séparées dans la classification prédite. L'indice est égal à 1 si aucune paire d'échantillons issue de la classification attendue n'est séparée dans la classification prédite. Afin de réaliser ces comparaisons, l'arbre de chaque classification hiérarchique est coupé afin d'obtenir 30 groupes d'échantillons. En effet le nombre de groupes choisi est plus grand que le nombre réel de tissus car sinon les groupes obtenus par classification hiérarchique montrent un grand déséquilibre au niveau de leur taille et il est difficile de faire la correspondance exacte entre ces groupes et les vrais groupes d'échantillons. Ensuite un indice de Rand ajusté est calculé sous R (package « rand.index ») pour chaque classification hiérarchique obtenue à partir de chacune des catégories de gènes testées (gènes codant pour des protéines, singletons et gènes paralogues). La comparaison des valeurs d'indice calculées pour chaque catégorie de gènes permet de déterminer la catégorie produisant le meilleur regroupement des échantillons en tissus.

Nous utilisons également une seconde approche de comparaison entre les groupes d'échantillons prédits et attendus en se basant sur le calcul d'un score F1 (Powers 2011). Le score F1 permet de prendre en compte la sensibilité et la précision de la classification effectuée. L'arbre de la classification hiérarchique est également divisé en 30 groupes d'échantillons puis un score F1 est calculé pour chaque paire groupe/tissu. Pour un tissu donné, le groupe d'échantillons avec le meilleur F1 score est considéré comme celui correspondant au tissu. Cette approche permet donc d'obtenir un score F1 par tissu. La comparaison des scores F1 calculés pour chaque classification obtenue à partir d'une catégorie de gènes permet également de déterminer la catégorie produisant le meilleur regroupement des échantillons en tissus (Figure 31).

Indice de Rand ajusté:

L'indice de Rand ajusté (ARI) permet de comparer deux classifications générant des nombres de groupes différents.

Considérons un premier ensemble R de groupes d'échantillons avec R={R1,…,Ri,…,Rr}, un second ensemble de groupes d'échantillons S={S1,…,Sj,…,Ss} et une collection de n échantillons.

A partir de ces ensembles, une table de contingence est construite indiquant le nombre d'échantillons dans chaque groupe (Table 4).

Table 4: Table de contingence pour le calcul des ARI

Tableau de contingence contenant le nombre d'échantillons dans chaque groupe de C et de R.

	$S_1 \cdots S_j \cdots S_s$	Somme
R_1	$n_{11} \cdots n_{1j} \cdots n_{1s}$	$n_{1.}$
$\vdots$	$\vdots \quad \cdots \quad \vdots \quad \cdots \quad \vdots$	$\vdots$
R_i	$n_{11} \cdots n_{1j} \cdots n_{1s}$	$n_{i.}$
$\vdots$	$\vdots \quad \cdots \quad \vdots \quad \cdots \quad \vdots$	$\vdots$
R_r	$n_{r1} \cdots n_{rj} \cdots n_{rs}$	$n_{r.}$
Somme	$n_{.1} \cdots n_{.j} \cdots n_{.s}$	

A partir de cette table nous pouvons calculer :

- A = $\sum_{ij}^{(r,s)} \binom{nij}{2}$, nombre de paires d'échantillons non séparées ;
- B = $\binom{n}{2}$, nombre total de paires ;
- C = $\sum_i^r \binom{n_{i.}}{2} + \sum_j^s \binom{n_{.j}}{2}$, somme des paires de la classification R et de la classification C ;
- D = $\frac{\sum_i^r \binom{n_{i.}}{2} X \sum_j^s \binom{n_{.j}}{2}}{\binom{n}{2}}$, nombre attendu de paires non séparées si les deux classifications sont indépendantes;

L'indice de Rand ajusté peut donc être calculé :

$$\text{ARI} = \frac{A-D}{C/2-\text{D}}$$

Si ARI = 1, alors toutes les paires sont bien classées, dans ce cas A = C/2. L'indice peut-être négatif si A < D, signifiant un très faible nombre de paires bien classées.

Score F1 :

Calcul de la sensibilité :

TPR = $\frac{VP}{VP+VN}$, avec VP, les vrais positifs et VN, les vrais négatifs

Calcul de la précision :

PPV = $\frac{VP}{VP+FP}$, avec VP, les vrais positifs et FP, les faux positifs.

Calcul du score F1 :

F1 = 2.$\frac{TPR.PPV}{TPR+PPV}$.

2.3.Estimation de la tissu-spécificité

Calcul du score τ:

Pour sélectionner les gènes tissu-spécifiques, nous utilisons le score τ (Yanai et al. 2005; Kryuchkova & Robinson-Rechavi 2017) afin d'estimer le degré de tissu-spécificité de chaque gène au travers de nos différentes régions cérébrales :

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{X}_i)}{n-1};\ \hat{X}_i = \frac{x_i}{\max\limits_{1\leq i\leq n}(x_i)}$$

Avec x_i la moyenne d'expression d'un gène donné dans le tissu i et n le nombre de tissus.

τ varie entre 0 et 1 où 0 signifie que le gène d'étude est uniformément exprimé et 1 que le gène est tissu-spécifique. Pour l'estimation de τ, les gènes doivent avoir une moyenne d'expression positive pour chaque région cérébrale. Bien que les données de GTEx aient été log-transformées avec $\log_2$(RPKM+1) permettant d'obtenir uniquement des valeurs positives, la correction des effets techniques et biologiques peut induire quelques valeurs négatives de l'expression des gènes. Les valeurs négatives sont donc remplacées par zéro afin de pouvoir conserver tous les gènes codants pour des protéines (16427 gènes) pour le calcul du score τ.

Pour estimer la tissu-spécificité des gènes, nous regroupons les données d'expression pour les 13 tissus cérébraux en 7 régions cérébrales afin que le score τ ne diminue pas artificiellement pour les gènes dont l'expression est partagée entre plusieurs tissus anatomiquement proches.

Définition du seuil du score τ par permutations :

Une fois le score τ défini, il doit être utilisé pour retrouver les gènes tissu-spécifiques à une région cérébrale. Comme le score obtenu est une valeur continue, il faut donc définir un seuil général permettant de considérer un gène comme étant tissu-spécifique.

Pour définir ce seuil, nous procédons par une approche empirique basée sur des permutations pour estimer la distribution du score τ sous hypothèse nulle. Nous appliquons 1000 permutations sur les étiquettes des régions assignées à chaque échantillon afin de modifier l'appartenance de l'échantillon à sa région. Pour chaque permutation, les scores τ sont recalculés pour chaque gène. La distribution des 1000 X 16427 scores τ est comparée à la vraie distribution des scores et une p-valeur est calculée pour chaque gène correspondant à la proportion de scores issus des permutations supérieures au score τ original. Une correction de Benjamini-Hochberg (Benjamini & Hochberg 1995) pour le nombre de gènes testés est appliquée sur toutes les p-valeurs. Les gènes avec une p-valeur corrigée inférieure à 0,01 sont considérés comme étant tissu-spécifiques ce qui correspond à un seuil du score τ de 0,525.

Site de tissu-spécificité :

Lorsqu'un gène est considéré comme étant tissu-spécifique à une région cérébrale, nous considérons la région dans laquelle le gène est le plus exprimé comme le site de tissu-spécificité.

2.4.Analyse d'expression différentielle

Les analyses d'expression différentielle (Anders & Huber 2010) permettent de comparer la valeur d'expression des gènes entre deux conditions à partir, par exemple, de données de comptage issues d'un séquençage RNA-seq. Cette analyse est basée sur l'hypothèse d'une distribution binomiale négative (Hilbe 2011) des comptages de chaque échantillon et utilise des méthodes statistiques de comparaison de moyennes et de variances pour retrouver les gènes différentiellement exprimés (calcul d'une p-valeur par gène).

Nous recherchons les gènes différentiellement exprimés entre deux tissus cérébraux. Une analyse différentielle d'expression de gènes est effectuée pour chaque paire de tissus cérébraux (78 paires) avec la méthode DESeq2 (Love et al. 2014) présenté sous forme d'un package R (« DESeq2 ») sous Bioconductor. Nous utilisons les données de comptage (GTEx) des 16427 gènes codants pour des protéines pour lesquels la moyenne

d'expression pour chaque tissu est supérieure à 0.1 RPKM. Comme il s'agit des données de comptage, les effets techniques et biologiques ne sont pas encore corrigés. Pour cela il est possible de les considérer comme covariables dans DESeq2.

Une fois les analyses différentielles effectuées, les p-valeurs sont corrigées pour les tests multiples pour le nombre de gènes testés avec la méthode de Benjamini-Hochberg (Benjamini & Hochberg 1995). Parmi la liste obtenue de gènes significativement différentiellement exprimés (FDR < 0,05), nous considérons uniquement ceux dont la valeur absolue de leur ratio d'abondance $\log_2$-transformé est supérieure à 0,5 ($|\log_2(\text{fold-change})|>0,5$).

3.Résultats

3.1.Expression des gènes dupliqués et singletons

Nous cherchons à étudier l'expression des gènes paralogues et des gènes non paralogues (singletons) dans différents tissus du cerveau. Pour cela nous utilisons les données produites par le consortium GTEx, correspondant à des données de valeur d'expression (RPKM) et de comptage issues de séquençage RNA-seq à partir de 13 tissus cérébraux différents (cervelet, hémisphère cérébelleux, cortex, cortex frontal, cortex cingulaire antérieur, amygdale, hippocampe, putamen, noyau caudé, noyau accumbens, hypothalamus, moelle épinière et substance noire).

Nous effectuons un prétraitement des données d'expression provenant du consortium GTEx (moyennes d'expression inférieur à 0,1 RPKM pour tous les tissus). Nous conservons ainsi des valeurs d'abondance pour 16427 gènes codants pour des protéines, 10335 gènes paralogues et 6092 gènes singletons. Ce seul empirique de 0,1 RPKM a été considéré comme la limite de sensibilité du séquençage RNA-seq au-dessous de laquelle la variabilité d'expression détectée peut être un artefact.

Ensuite, afin de minimiser l'influence des valeurs d'expression extrêmes et de diminuer l'influence dans les analyses de classification hiérarchiques des gènes très fortement exprimés et également d'obtenir une distribution permettant d'appliquer une régression linéaire sur les données d'expression, les données sont transformées en log (log-transformées). Les données ont ensuite été ajustées pour corriger les effets techniques (plateforme) liés au séquençage (effets « batch ») ainsi que certains effets biologiques (genre, C1, C2 et C3) (voir Méthode Chapitre 3).

Nous évaluons deux transformations log des valeurs d'expression : $\log_2$(RPKM + 0.1) et $\log_2$(RPKM + 1). Les distributions de chaque transformation avec ajustement pour les effets techniques et biologiques sont représentées dans la Figure 29**.** Nous sélectionnons la transformation $\log_2$(RPKM + 1) car ainsi presque toutes les valeurs demeurent positives après ajustement.

Ces distributions nous permettent également de comparer l'expression globale à travers les tissus cérébraux entre les gènes paralogues et les singletons. Il semble apparaître que les gènes paralogues soient en moyenne moins exprimés que les singletons.

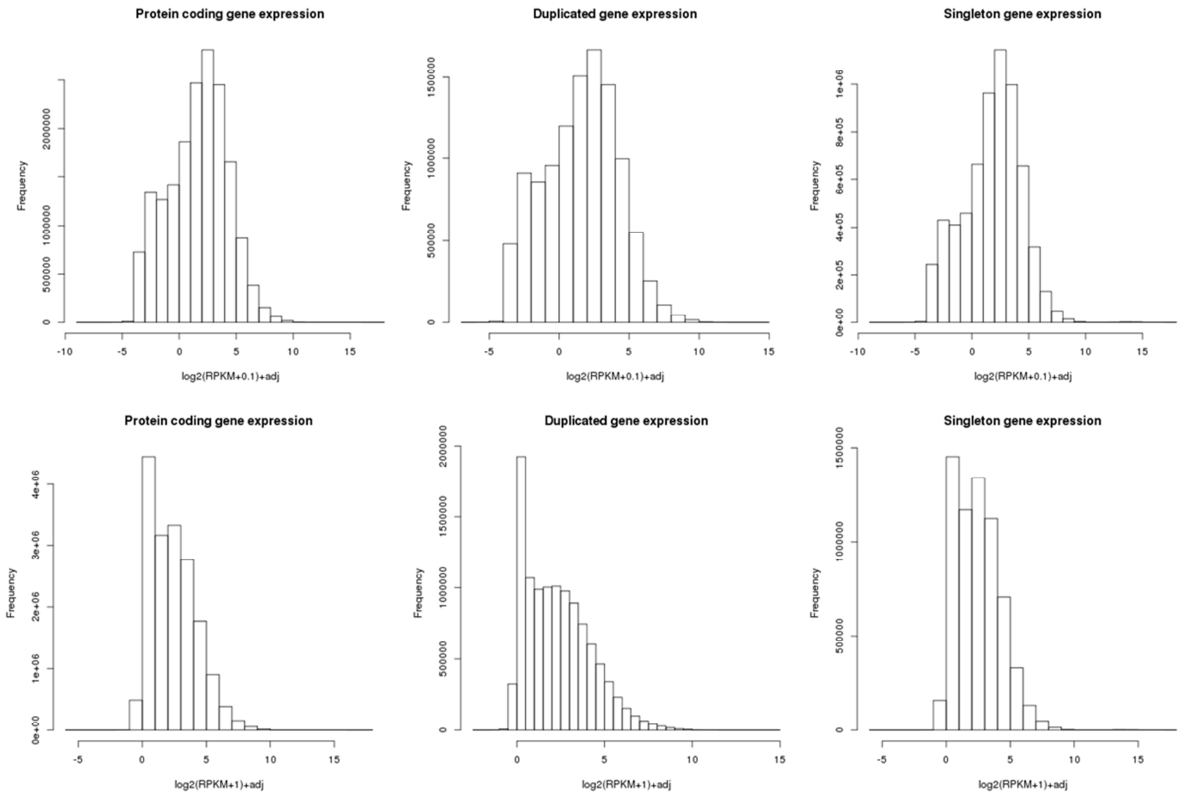


Figure 29: Comparaison de l'expression des gènes singletons et des gènes dupliqués dans les différents tissus cérébraux

Expression en log2(RPKM+0.1) ou log2(RPKM + 1) avec correction des effets batch. Les données ne contiennent pas les gènes qui ont une moyenne inférieure à 0.1 RPKM dans tous les tissus.

3.2.Différentiation des tissus cérébraux par l'expression des gènes dupliqués

3.2.1.Classification des tissus cérébraux à partir de l'expression des gènes

Le consortium GTEx (Ardlie et al. 2015) a montré qu'une classification hiérarchique à partir l'expression de tous les gènes au travers de tous les tissus était capable de regrouper les échantillons par tissus. Nous reproduisons cette classification hiérarchique mais en considérant uniquement les échantillons provenant des tissus cérébraux. Nous remarquons que pour chacune des trois catégories de gènes (gènes codant pour les protéines, singletons et paralogues), les valeurs d'expression permettent de bien regrouper les échantillons en fonction de leurs tissus cérébraux d'origine (Figure 30).

Nous souhaitons aller plus loin que cette inspection visuelle en comparant quantitativement ces trois classifications afin de tester notamment si les gènes paralogues permettent de mieux différentier les tissus cérébraux que les singletons.

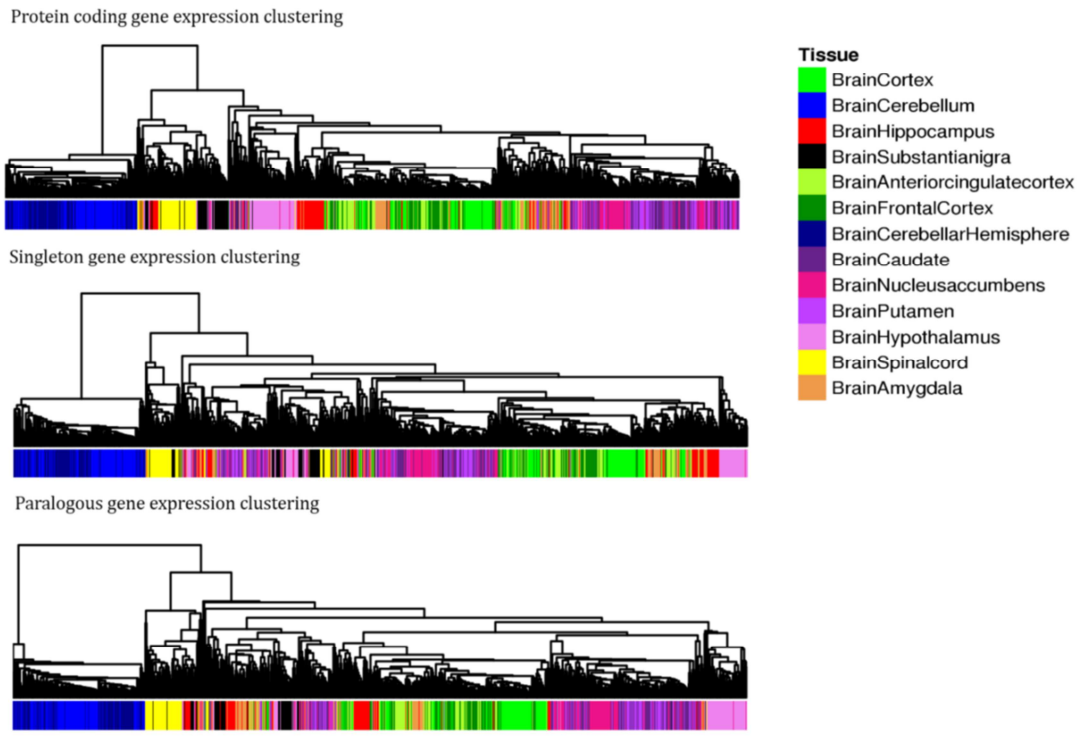


Figure 30: Classification hiérarchique des échantillons par tissu à partir de l'expression des gènes

Comparaison de la classification des tissus suivant le type de gènes (codant pour des protéines, Singletons ou Paralogues). Les couleurs représentent les tissus cérébraux (cortex, cervelet, hippocampe, substance noire, cortex cingulaire antérieur, cortex frontal, hémisphère cérébelleux, noyau caudé, noyau accumbens, putamen, hypothalamus, moelle épinière et amygdale).

Afin de comparer la classification des paralogues et des singletons, nous calculons un score F1 pour chaque tissu. Pour cela nous coupons d'abord l'arbre de la classification hiérarchique en 30 groupes d'échantillons. Nous préférons obtenir plus de groupes d'échantillons que de tissus cérébraux car pour un nombre de groupes égal au nombre de tissus, nous obtenions des tailles de groupes très inhomogènes, avec des groupes d'échantillons qui rassemblaient 2 ou 3 tissus et d'autres groupes constitués par un très faible nombre d'échantillons. Pour chaque groupe d'échantillons, nous calculons un score F1 vis-à-vis de chaque groupe attendu (c'est-à-dire chaque tissu). Nous associons ensuite à chaque tissu le meilleur score F1 obtenu pour ce tissu parmi les groupes d'échantillons observés (Figure 31). Sur cette figure nous remarquons que la majorité des tissus ont des scores F1 meilleurs (sinon égaux) lorsque la classification est réalisée à partir de l'expression des paralogues, par rapport aux scores obtenus avec les singletons. Ces résultats meilleurs sur les paralogues ne semblent pas s'expliquer seulement par le nombre différent de gènes dans chaque catégorie puisque la classification sur l'ensemble des gènes codants est très proche de celle sur les paralogues uniquement.

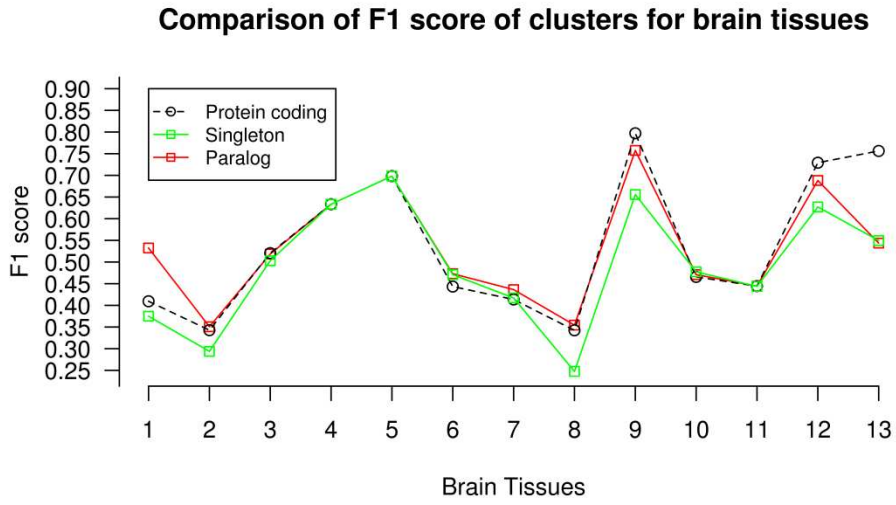


Figure 31: F1 scores des clusters d'échantillons à partir de l'expression des gènes

Meilleur F1 score calculé parmi les différents groupes d'échantillons pour chaque tissu cérébral (1-amygdale, 2-cortex cingulaire antérieur, 3-noyau caudé, 4-hémisphère cérébelleux, 5-cervelet, 6-cortex, 7-cortex frontal, 8-hippocampe, 9-hypothalamus, 10-noyau accumbens, 11-putamen, 12-moelle épinière, 13-substance noire) pour chaque catégorie de gènes testé (gènes codants pour des protéines, Singletons et Paralogues).

Ces mêmes comparaisons sont également effectuées à l'aide de l'indice de Rand ajusté calculé pour chacune des trois classifications obtenues correspondant aux trois catégories de gènes (Rand 1971).Cet indice mesure la similarité entre les groupes d'échantillons attendus et ceux prédits par la classification hiérarchique à partir des données d'expression de gènes. Les indices de Rand ajustés calculés à partir de chaque catégorie de gènes (gènes codants pour des protéines, singletons et paralogues) sont relativement faibles (0.175, 0.182 et 0.197 respectivement) car nous comparons des groupes attendus (13 groupes) et prédits (30 groupes) de tailles différentes. En effet la coupure droite de l'arbre ne permet pas d'identifier certains clusters d'échantillons emboités. Cependant, la valeur obtenue pour les paralogues est plus élevée que celle des singletons et nous pouvons dire que ce résultat est positivement corrélé avec les conclusions obtenues pour les scores F1. Ainsi, les gènes paralogues semblent mieux regrouper les échantillons par tissus cérébraux que les singletons.

3.2.2.Analyse différentielle par paire de tissus

Afin de poursuivre les comparaisons entre la capacité des paralogues et des singletons à différentier les tissus cérébraux, nous décidons de réaliser des études d'analyses différentielles d'expression de gènes entre paires de tissus cérébraux. Pour cela une analyse avec la méthode DESeq2 est effectuée entre toutes les paires de tissus du cerveau (78 paires de tissus). Les gènes différentiellement exprimés (c'est-à-dire avec une p-valeur corrigée (FDR) < 0,05) sont conservés et sont considérer uniquement ceux avec un $\log_2$ (fold-change) supérieur à 0,5 en valeur absolue (ratio minimum entre l'expression du gène dans la condition où il est le plus exprimé et son expression dans la deuxième condition) (Figure 32). Sur la Figure 32 sont représentées deux paires de tissus : une paire de tissus très différents anatomiquement (cortex vs hippocampe) et une paire de tissus proches anatomiquement (cortex vs cortex cingulaire antérieur). Nous retrouvons systématiquement davantage de gènes différentiellement exprimés entre des tissus anatomiquement distants qu'entre des tissus proches (Annexe A).

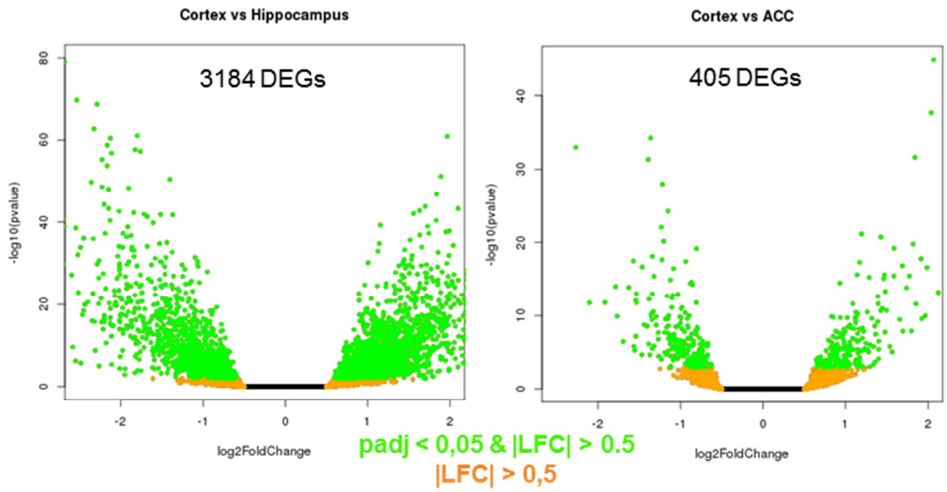


Figure 32: Gènes différentiellement exprimés entre deux paires de tissus cérébraux

Deux exemples d'analyse d'expression différentielle entre paires de tissus cérébraux (cortex vs hippocampe et cortex vs cortex cingulaire antérieur). Les gènes différentiellement exprimés sont représenté en vert (p-value ajustée < 0,05 et $|\log_2$ Fold Change$|$ > 0,5). Les gènes en orange ne sont pas significatifs ($|\log_2$ Fold Change$|$ > 0 ,5 mais p-value ajustée > 0,05). Pour la paire cortex vs hippocampe, 3184 gènes sont différentiellement exprimés alors que seulement 405 gènes le sont pour la paire cortex vs cortex cingulaire antérieur.

Ensuite, nous comparons la proportion de gènes paralogues et de singletons parmi les gènes différentiellement exprimés pour chaque paire de tissus. Nous avons effectué un test d'enrichissement pour chaque paire de tissus afin de déterminer si la proportion de paralogues différentiellement exprimés est significativement plus élevée ou plus faible que la proportion de singletons différentiellement exprimés (test Chi^2, p-valeur < 6,41e-4 après correction de Bonferroni pour le nombre de paires de tissus testées). Sur les 78 paires au total, 76 d'entre elles donnent un test d'enrichissement significatif dans le sens des paralogues, ce qui correspond à 95% des paires de tissus testées. Pour les 2 paires restantes, le test de Chi^2 n'est pas significatif, la proportion de gènes paralogues n'est significativement pas différente de celle des singletons. Ainsi, les gènes différentiellement exprimés sont globalement enrichis en gènes paralogues pour la très grande la majorité des paires de tissus cérébraux.

Ces approches complémentaires de classification et d'analyse différentielle illustrent la contribution majeure des gènes paralogues en terme de capacité à différentier les tissus du cerveau humain par rapport aux singletons.

3.3.Calcul de la tissu-spécificité des gènes

Nous venons de voir que l'expression des gènes, en particulier celle des gènes paralogues, permet de regrouper les échantillons provenant des mêmes tissus cérébraux. Ceci laisse à penser que les gènes vont globalement avoir une expression variable au travers des tissus. Pour poursuivre cette analyse, nous cherchons à identifier individuellement les gènes influençant les classifications précédentes car porteurs d'une propriété de tissu-spécificité.

Pour étudier l'expression tissu-spécifique des gènes, nous regroupons tout d'abord certains tissus proches anatomiquement pour obtenir des régions cérébrales. A partir des 13 tissus initiaux, nous avons obtenu 7 régions cérébrales : le cervelet (cervelet et hémisphère cérébelleux), le cortex (cortex, cortex frontal et cortex cingulaire antérieur), le ganglion de la base (putamen, noyau accumbens et noyau caudé), l'amygdale-hippocampe, l'hypothalamus, la moelle épinière et la substance noire (Figure 27 etFigure 28).

Parmi les différentes méthodes bioinformatiques permettant de mesurer la tissu-spécificité (Kryuchkova & Robinson-Rechavi 2017), notre choix s'est porté sur le calcul du score τ (Yanai et al. 2005). Ce score τ est compris entre 0, pour les gènes uniformément exprimés, et 1 pour les gènes très tissu-spécifiques. Comme pour les études précédentes, nous utilisons les données RPKM de GTEx, log-transformées et ajustées pour les effets techniques et biologiques pour explorer la tissu-spécificité des gènes. Ces valeurs d'expression étant positives, elles peuvent toutes être utilisées pour calculer les scores τ. Comme aucun seuil *a priori* ne peut être appliqué sur le score τpour définir les gènes tissu-spécifiques du fait de l'absence d'une distribution bimodale, nous calculons un seuil empirique statistiquement significatif basé sur des permutations. Nous effectuons 1000 permutations sur les noms des tissus associés aux valeurs d'expression de chaque gène pour chaque échantillon puis nous calculons un score τpour chaque gène à chaque permutation. (Figure 33). La distribution des scores τgénérés à partir des permutations nous permet de définir la valeur de τcorrespondant à un seuil de p-valeur corrigée de 0,01 (Benjamini & Hochberg 1995). Sur la Figure 33, le seuil associé à une p-valeur de 0,01 est visualisé et correspond à un seuil de score τde 0,525. Ainsi tous les gènes avec un score τsupérieur à 0.525 ont été considérés comme tissu-spécifiques. Nous avons également visualisé l'expression des gènes au sein des différentes régions cérébrales (Figure 34) pour différentes fenêtres de score τ. Nous

observons que cette valeur seuil de τde 0,525 permet de révéler des gènes tissu-spécifiques à un tissu.

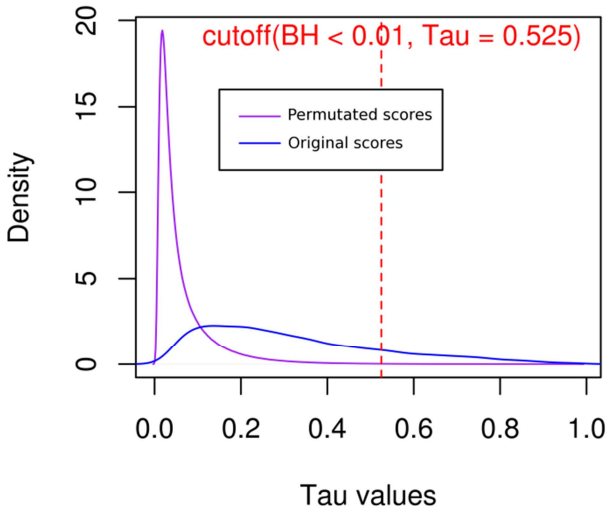


Figure 33: Scores τ réels et obtenus par permutations pour les gènes codant pour des protéines

Distributions des scores τ calculés pour chaque gène codant pour des protéines (courbe bleue) et des scores τcalculés sur ces mêmes données après permutation des étiquettes des échantillons indiquant les régions cérébrales dont ils proviennent (1000 permutations) (courbe violette). Un seuil de tissu-spécificité est défini (0,525) correspondant à un seuil de FDR de 0.01 (ligne rouge pointillée).

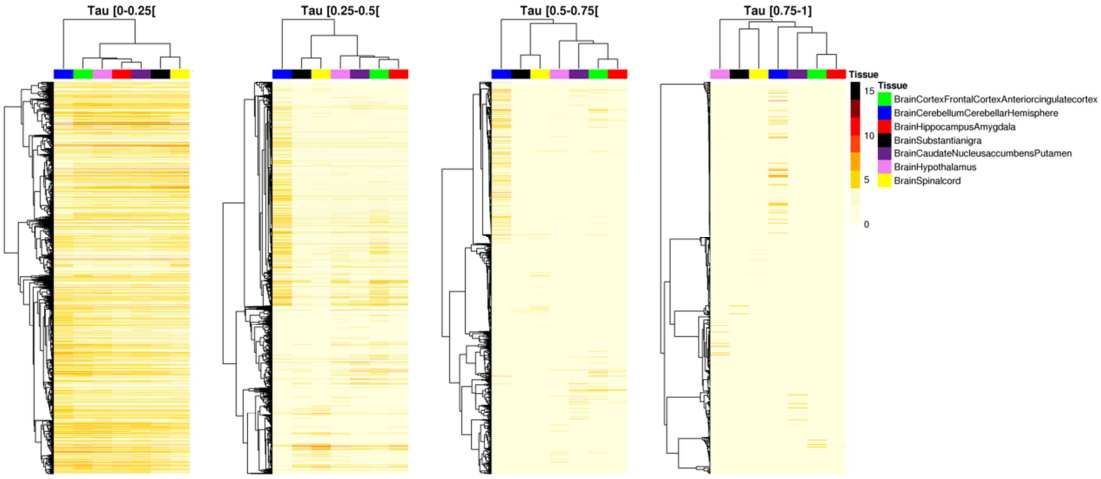


Figure 34: Expression des gènes dans chaque région du cerveau

Représentation des classifications de l'expression des gènes dans chaque région du cerveau (région avec les différents tissus du Cortex, région avec les deux tissus du Cervelet, la région avec l'hippocampe et l'Amygdale, la Substance noire, le ganglion de la base, l'Hypothalamus et la moelle épinière). Chaque représentation considère les gènes ayant un score τ compris dans un intervalle donné ([0-0,25[, [0,25-0,5[, [0,5-0,75[, [0,75, 1]).

Avec cette approche, nous trouvons que 17% des gènes codant pour des protéines sont tissu-spécifiques à une région cérébrale (2829 gènes). Nous interrogeons cette population de gènes tissu-spécifiques et établissons un enrichissement significatif des gènes paralogues en gènes tissu-spécifiques par rapport aux singletons (19.2% des gènes paralogues et 13.9% des singletons sont tissu-spécifiques, test du Chi2, p-valeur = 2.045e-18). Cet enrichissement confirme l'implication majeure des paralogues dans la capacité à différencier les tissus du cerveau. Sur la Figure 35sont représentées les valeurs d'expression des gènes paralogues tissus spécifiques (1985 gènes). Pour chaque gène tissu-spécifique, nous considérons que la région cérébrale associée au maximum d'expression du gène est celle porteuse de la tissu-spécificité.

La Table 5 indique la proportion de gènes spécifiques à chaque région cérébrale au sein de tous les gènes tissu-spécifiques ou uniquement au sein des gènes paralogues tissu-spécifiques, ainsi que la proportion de gènes exprimés (moyenne par gène et par région au travers des échantillons > 0,1 RPKM) dans chaque région. Nous remarquons que la proportion de gènes exprimés dans chaque région cérébrale est très élevée et très stable, que ce soit parmi tous les gènes ou parmi les paralogues. A l'inverse, la répartition des gènes tissu-spécifiques est hautement inhomogène entre régions, que l'on considère les gènes codants pour des protéines ou seulement les paralogues. Cependant, les répartitions des gènes tissu-spécifiques dans ces deux groupes de gènes sont proches. Les gènes tissu-spécifiques se répartissent principalement dans le cervelet (40,2%), la moelle épinière (20,9%) et l'hypothalamus (16,4%). Les 22,5% restants correspondent aux 4 autres régions cérébrales.

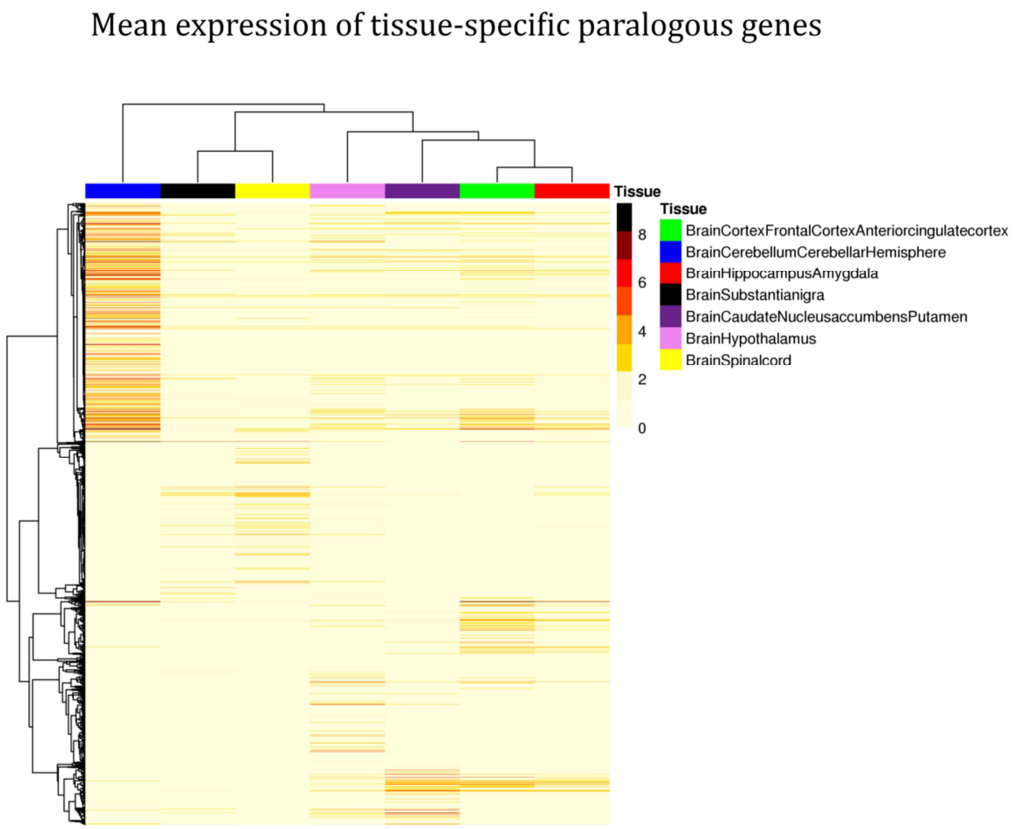


Figure 35: Expression des gènes paralogues tissu-spécifiques dans chaque région du cerveau

Représentation de la classification de l'expression des gènes paralogues tissu-spécifiques dans chaque région du cerveau (région avec les différents tissus du Cortex, région avec les deux tissus du Cervelet, la région avec l'hippocampe et l'Amygdale, la Substance noire, le ganglion de la base, l'Hypothalamus et la moelle épinière).

Table 5: Comparaison du nombre de gènes tissu-spécifiques et exprimés dans chaque région cérébrale

Pour chaque région du cerveau, le nombre de gènes tissu-spécifiques à cette région et le nombre de gènes exprimés sont indiqués. Pour les gènes tissu-spécifiques, les pourcentages attribués à chaque région sont calculés par rapport au nombre total de gènes tissu-spécifiques. Ainsi pour chaque catégorie de gènes (codants et paralogues), la somme des pourcentages de gènes tissu-spécifiques au travers des régions correspond à la totalité des gènes tissu-spécifiques (100%). Pour les gènes exprimés, les pourcentages des différentes régions sont calculés par rapport au nombre total de gènes exprimés dans le cerveau (dans au moins une région cérébrale). La somme des pourcentages de gènes exprimés dépasse 100% puisque de nombreux gènes sont exprimés dans plusieurs régions.

	Nombre de gènes Tissu-spécifiques		**Nombre de gènes exprimés**	
Tissus	Gènes codants 2829 (100%)	Gènes paralogues 1985 (100%)	Gènes codants 16427	Gènes paralogues 10335
Ganglion de la base	261 (9,2%)	182 (9,2%)	15351 (93,4%)	9603 (92,9%)
Cervelet	1137 (40,2%)	761 (38,3%)	15001 (91,3%)	9348 (90,4%)
Cortex	256 (9%)	198 (10%)	15285 (93%)	9568 (92,6%)
Amygdale-Hippocampe	44 (1,6%)	34 (1,7%)	15337 (93,4%)	9614 (93%)
Hypothalamus	463 (16,4%)	304 (15,3%)	15618 (95,1%)	9796 (94,7%)
Moelle épinière	590 (20,9%)	446 (22,5%)	15342 (93,4)	9625 (93,1%)
Substance noire	78 (2,8%)	60 (3%)	15258 (92,9%)	9566 (92,6%)

4.Discussion et conclusion

Les analyses d'alignement de séquences entre paires de gènes paralogues de même famille, nous ont révélé que des fortes homologies de séquences existaient et qu'elles pouvaient donc perturber la fiabilité de la mesure d'expression par gène. Pour rappel, l'alignement des lectures réalisé par le consortium GTEx a été effectué en conservant uniquement celles avec placement unique sur le génome. Par conséquent, les homologies de séquences entre paires de gènes paralogues ont pu entrainer une moindre couverture en lectures de certaines régions génomiques de ces gènes et donc occasionner une diminution de leur mesure d'expression. Ces problématiques d'alignement ont pu notamment mener à une baisse extrême des mesures d'expression, à la limite de sensibilité du RNA-seq, pour les paires de gènes dupliqués dont la séquence est très conservée. Les gènes que nous avons supprimés du fait de leur très faible expression dans tous les tissus incluaient probablement ces paires de gènes problématiques. Cette filtration des données d'expression n'a cependant pas permis de réduire significativement le décalage d'expression moyenne mesurée entre les singletons et les gènes paralogues. Il est possible que les paires de gènes dupliqués partageant seulement une portion de séquence très similaire et problématique pour l'alignement souffrent également d'une moindre couverture. Cependant ce décalage reflète probablement aussi une origine biologique à cette expression globalement plus forte chez les singletons. En effet les gènes dupliqués peuvent avoir une expression plus faible due à l'équilibre de dosage ou bien à une expression asymétrique entre les deux paralogues (Lan & Pritchard 2016).

Les précédentes études sur le calcul d'un score de tissu-spécificité sur des tissus provenant de plusieurs organes (Kryuchkova & Robinson-Rechavi 2017) ont montré une distribution bimodale des valeurs du score Tau. En effet ce type de distribution permet de définir un seuil visuel du score permettant d'obtenir des gènes tissu-spécifiques. Sur notre étude sur les différentes régions cérébrales, nous n'avons pas observé de distribution bimodale et nous avons donc dû estimer un seuil de tissu-spécificité des gènes par une méthode statistique de permutations.

En terme d'expression tissulaire, nous avons également montré que la classification des échantillons en tissus cérébraux réalisée à partir de l'expression des gènes paralogues semblait mieux regrouper les échantillons cérébraux que celle basée sur les singletons.

Une analyse complémentaire, basée sur l'étude d'expression différentielle entre paires de tissus cérébraux, nous a aussi permis de montrer, qu'au sein des gènes différentiellement exprimés, la proportion de gènes paralogues était significativement plus élevée qu'au sein des gènes non différentiellement exprimés. Ces analyses suggèrent des profils d'expression plus tissu-spécifiques des paralogues au travers des tissus cérébraux par rapport aux singletons. Cette hypothèse a pu être confirmée par l'enrichissement des gènes paralogues en gènes tissu-spécifiques, ce qui signifie que les paralogues sont significativement plus tissu-spécifiques que les singletons.

Pour conclure, notre étude de l'expression des gènes dans les tissus cérébraux semble montrer un rôle majeur des paralogues dans la capacité à différencier ces tissus. De plus notre travail sur la tissu-spécificité semble indiquer que les gènes paralogues ont des profils d'expression plus spécifiques à une région cérébrale que les singletons. Ainsi, nos travaux suggèrent la contribution importante des paralogues à la spécialisation des tissus cérébraux.

Chapitre 4 : Etude de la co-expression des gènes paralogues au sein de différents tissus cérébraux

1.Introduction

1.1.Etude des réseaux de co-expression des gènes

Les gènes paralogues appartiennent à des familles de gènes définies par des méthodes phylogénétiques (voir Méthode Chapitre 2). Au sein du génome d'une espèce telle que l'homme, une famille est composée de gènes résultant de duplications se différenciant par les mécanismes moléculaires les ayant générés ainsi que par les dates auxquelles elles sont apparues (Li et al. 2006). Un moyen d'explorer les profils d'expression ou les fonctions biologiques de ces familles repose sur l'étude des réseaux de co-expression des gènes.

De nombreuses études utilisent l'inférence de réseaux de co-expression pour caractériser l'implication conjointe des gènes lorsque les tissus/organismes au sein desquels ils s'expriment, sont exposés à différentes conditions.

Les données à analyser se présentent généralement sous la forme :

$$X = \left[x_g^s\right]$$

où g indexe tous les gènes considérés et s indexe tous les échantillons étudiés.

Ces échantillons peuvent se répartir dans différentes conditions expérimentales suivant la question posée. Ces conditions peuvent correspondre par exemple à un statut malade/sain que l'on échantillonne aux travers des individus.

Nous citons ici deux études emblématiques de l'usage de WGCNA. Une étude a été faite sur l'analyse de co-expression d'individus présentant un syndrome autistique (ASD) (Voineagu et al. 2013) avec la méthode WGCNA (« Weighted Gene Correlation Network Analysis ») (Oldham et al. 2008). Les auteurs ont inféré un réseau de co-expression des gènes s'exprimant dans les tissus du cerveau des seuls sujets contrôle, un second réseau obtenu sur les sujets ASD et enfin un réseau avec tous les échantillons. Ils ont ensuite comparé les groupes de gènes co-exprimés obtenus ou modules produit par WGCNA dans les trois configurations présentées leur permettant d'identifier des modules spécifiques de l'autisme enrichis en fonctions neuronal et du système immunitaire. Une seconde étude a cherché à comparer la co-expression des gènes pour des transcriptomes

des cerveaux de différentes espèces de primates (Konopka et al. 2012). Ils ont utilisé la méthode WGCNA pour générer ces réseaux de co-expression des gènes. Dans ce cas plusieurs jeux de conditions expérimentales ont été étudiés : d'abord un réseau par espèce (humain, chimpanzé et macaque) et un réseau global qui analysait toutes les espèces en même temps. Les différents ensembles de modules ainsi obtenus permettent de procéder à la comparaison des modules de gènes de chaque espèce et par exemple de tenter de trouver les modules spécifiques à une espèce notamment un module spécifique à l'homme enrichis dans les processus de morphologie neuronales ainsi que de l'évolution du langage. Enfin les conditions peuvent correspondre à différents tissus. Le consortium GTEx a effectué des analyses de co-expression par tissu soit avec WGCNA (Ardlie et al. 2015) ou bien avec leur propre outils GNAT (Pierson et al. 2015) dans le but de d'identifier des groupes de gènes pouvant expliquer la spécificité tissulaire.

D'une manière générale, WGCNA est utilisé pour produire des modules qui apportent une information sur la co-expression des gènes pris dans une condition expérimentale donnée sous contrainte de l'existence d'une organisation en réseau sans échelle des gènes. Chaque module peut ensuite être caractérisé ou annoté de manière à apporter de nouvelles hypothèses biologiques sous-jacentes aux conditions expérimentales étudiées.

1.2.Apport des réseaux de co-expression pour la compréhension des transcriptomes de tissus humaines

Au moment de la préparation des données pour la thèse, le consortium GTEx (Ardlie et al. 2015) avait séquencé, par RNA-seq, 53 tissus humains pour 714 donneurs. Parmi les différentes études effectuées à partir de ces données, le consortium GTEx a effectué une étude sur la co-expression des gènes dans différents tissus.

Dans le consortium GTEx, la méthode WGCNA a été utilisée afin d'inférer un réseau de co-expression par tissu pour 9 tissus différents (tissu adipeux, artère tibiale, ventricule gauche du cœur, poumon, muscle squelettique, nerf tibial, peau, thyroïde et sang). Cette approche a permis d'identifier des modules de gènes co-exprimés partagés ou spécifiques à certains tissus. Les modules obtenus pour chaque tissu ont été extraits et une classification hiérarchique pour chaque tissu basée sur les GO ou les facteurs de transcriptions proches a été effectuée. Cette classification a permis d'identifier des certains processus biologiques partagés pour tous les tissus et d'autres spécifiques à certains tissus. La recherche sur les facteurs de transcriptions associés aux modules a

permis de repérer les modules potentiellement co-régulés (même facteurs de transcription). Dans un second temps, le consortium a profité de l'avantage de posséder plusieurs tissus par individu pour définir des modules de gènes co-régulés au sein de chaque individu, en recherchant des corrélations d'expression entre paires de gènes au travers des tissus pour un même individu. Puis une méthode différente a été utilisée, basée sur la projection dans un espace de faible dimension des expressions des gènes moyennées sur les individus au travers de tous les tissus. Le regroupement des gènes sur la base de leur profil d'expression tissulaire permet d'identifier des modules de gènes co-exprimés au travers des tissus et communs à tous les individus. Un total de 117 modules a été retrouvé et les profils moyens d'expression des gènes de chaque module dans chaque tissu (Figure 36) permettent de retrouver certains modules tissu-spécifiques, avec notamment un module contenant des gènes présentant un profil d'expression spécifique pour le cerveau (Mod101).

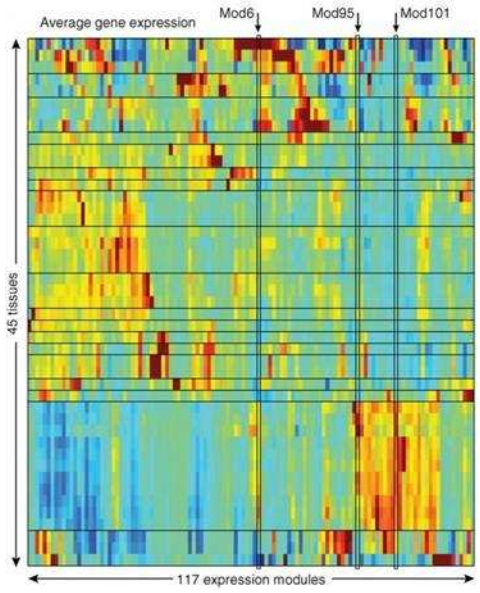


Figure 36: Expression moyenne des modules de co-expression dans chaque tissu

Ardlie et al., 2015.Expression moyenne des gènes des modules de co-expression par tissu.

1.3.Objectif

Nous nous sommes intéressés aux données d'expression de GTEx correspondant à 13 tissus du cerveau. Nous avons étudié dans le Chapitre 3 la tissu-spécificité des gènes au sein de 7 régions cérébrales regroupant ces 13 tissus cérébraux. Nous avons montré notamment que les gènes paralogues étaient enrichis en gènes tissu-spécifiques.

A présent, nous souhaitons identifier des modules de gènes paralogues dont la co-expression pourrait contribuer à la différentiation des tissus cérébraux.

Pour faire cette étude, nous utilisons la méthode WGCNA qui est fréquemment employée pour explorer les réseaux de co-expression (Voineagu et al. 2013; Ardlie et al. 2015). De plus, nous souhaitons interroger ces modules de co-expression, à partir de notre connaissance des familles de gènes, afin d'améliorer notre compréhension de l'expression de ces familles dans le cerveau.

Nous allons réaliser un travail conséquent au niveau du choix des paramètres de l'outil WGCNA afin de pouvoir produire des modules de gènes co-exprimés de taille comparable aux familles de gènes. Une fois ces modules de co-expression obtenus, nous allons effectuer leur caractérisation biologique. Enfin, nous comparerons ces modules avec les familles de gènes pour identifier en particulier celles possédant une homogénéité d'expression de leurs gènes au travers des tissus cérébraux.

2.Matériel et méthodes

2.1.Optimisation des paramètres de WGCNA

Nous avons employé WGCNA pour réaliser l'inférence d'un réseau de co-expression de gènes à partir de tissus cérébraux humains (« Weighted Gene Correlation Network Analysis ») (Langfelder & Horvath 2008). Cette méthodologie permet de générer des réseaux de co-expression et d'identifier des modules (groupes) de gènes co-exprimés.

Elle consiste en l'enchaînement de plusieurs traitements de données :

1) Une mesure de corrélation est calculée pour chaque paire de gènes afin d'estimer la similarité de leur profil d'expression ;
2) Une matrice d'adjacence est générée à partir de ces mesures de corrélation :

$$A = [a_{i,j}]$$

3) A partir de la matrice d'adjacence, une matrice de recouvrement topologique (TOM- « Topological Overlap Matrix ») est construite en convertissant les valeurs de corrélation de la matrice d'adjacence en valeurs représentatives de la similarité de connectivité k entre paires de gènes :

$$T = [t_{i,j}]$$

4) Une classification hiérarchique des gènes est obtenue à partir des valeurs de la matrice TOM.

5) L'identification des modules de co-expression est réalisée par la découpe de l'arbre de classification hiérarchique des gènes.

WGCNA (Langfelder & Horvath 2008) permet de regrouper les gènes en modules à partir de la similarité de leur profil d'expression. La méthode WGCNA définie un module 0 (ou « grey ») contenant les gènes considérés comme étant non co-exprimés du fait d'une très faible variabilité au travers de tous les échantillons. Nous supprimons les gènes avec une variance nulle et ceux ayant une expression très faible sur la totalité des échantillons. Ce filtre permet donc de limiter la taille de ce module « grey » et d'éviter que ces gènes aux profils d'expression peu informatifs soient intégrés dans des modules de co-expression.

Calcul de la matrice d'adjacence :

La matrice TOM permet de construire l'arbre de la classification hiérarchique des gènes à partir duquel sont identifiés les modules de co-expression. Le calcul de cette matrice est basé sur la corrélation de Pearson entre chaque paire de gènes au travers de tous les échantillons. La matrice peut être signée ou non signée. La méthode signée prend en compte uniquement les corrélations positives alors que la méthode non signée prend en compte les corrélations et les anti-corrélations de la même façon. Les valeurs d'adjacence sont donc des valeurs continues. Le calcul de la valeur d'adjacence entre deux gènes *i* et *j* est le suivant pour une matrice non signée :

$$a_{ij} = |corPearson|^{\beta}$$

Avec β, le paramètre de seuillage doux utilisé pour respecter une topologie de réseau invariant d'échelle.

Paramètre de seuillage doux β:

WGCNA fait que la majorité des gènes sont faiblement connectés entre eux du point de vu de la co-expression et que seulement un petit nombre de gènes est très connecté. Cette hypothèse est la traduction au niveau des réseaux de co-expression de l'idée que les gènes interagissent et sont organisés suivant un réseau invariant d'échelle (« scale free »).

La connectivité d'un gène correspond à la somme des valeurs d'adjacence de chaque paire du gène *i* avec tous les autres gènes *u* :

$$k_i = \sum_u a_{iu}$$

Plus précisément, il faut que la fréquence de la connectivité k, au travers des gènes, F(k), soit inversement proportionnelle à k^{α} (avec α un nombre réel positif), ce qui se traduit par une proportionnalité entre log(F(k)) et log(k). Le critère à respecter est donc le suivant:

$$R^2 = cor\left(log_{10}(k), log_{10}\big(F(k)\big)\right)^2 > 0.8$$

Dans WGCNA, la recherche des modules co-exprimés est faite en imposant de respecter une topologie de réseau invariant d'échelle (« scale free »). Le paramètre de seuillage doux β introduit dans le calcul de la matrice d'adjacence pour contraster les faibles et les fortes valeurs de corrélation de la matrice A peut permettre d'obtenir une telle topologie de réseau lorsqu'il est choisi de façon optimale. L'estimation de ce paramètre de seuillage doux est effectuée avec la fonction « pickSoftThreshold » du package « WGCNA » sous R.

Calcul de la matrice TOM :

La mesure (ou matrice) de recouvrement topologique a été introduite (Ravasz 2002) pour rendre compte de la similarité de connectivité (ou inter-connectivité) entre deux gènes. Cette mesure empirique de l'inter-connectivité entre gènes est reprise dans de nombreuses études. Elle se montre robuste pour l'estimation de la co-expression dans des contextes ou le bruit sur les données ne permet pas d'estimer précisément toutes les connections gène à gène. La matrice TOM est essentiellement une covariance sur les valeurs d'adjacence construites précédemment.

Le calcul des valeurs [$t_{i,j}$]de la matrice TOM dépend des valeurs d'adjacence, a_{ij} de la paire de gènes (i,j) et de la connectivité (k_i et k_j) des gènes i et j :

$$t_{ij} = \begin{cases} \dfrac{\sum_u a_{iu}.a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} ; si\ i \neq j \\ 1 \qquad\qquad\qquad ; si\ i = j \end{cases}$$

Paramètres Cuttree:

Le paramètre Cuttree correspond à la hauteur à laquelle l'arbre de la classification hiérarchique des gènes est coupé. Les clusters obtenus correspondent à des modules de co-expression initiaux qui pourront ensuite être sub-divisés. La racine du cluster devra donc être en dessous de la valeur du Cuttree pour que le cluster soit considéré comme un module de co-expression.

Paramètres Deepsplit :

La coupure de l'arbre au niveau du Cuttree a permis de construire les modules de co-expression initiaux. Ces modules peuvent ensuite être sub-divisés en sous-modules emboités. Le paramètre Deepsplit permet de contrôler la sensibilité de la détection des modules. Plus la valeur du Deepsplit est élevée, plus la sensibilité de la détection des modules est élevée et plus la taille des modules sera petite.

Afin d'identifier les modules, tous les clusters possibles de l'arbre de la classification hiérarchique sont testés en partant des plus petits clusters (feuilles de l'arbre) jusqu'aux plus gros (valeur du Cuttree) selon une approche « bottom-up ». A chaque cluster est associée une moyenne $\bar{d}$ (moyenne des dissimilarités des paires de gènes appartenant au cluster) et un gap g (différence entre $\bar{d}$ et la hauteur dans l'arbre de la jonction du cluster avec le reste de l'arbre). Pour identifier les modules il faut définir un d_{max} et un g_{min}. Pour qu'un cluster soit considéré comme un module, son $\bar{d}$ ne doit pas être supérieur à d_{max} et son g doit dépasser g_{min}. Le Deepsplit va donc jouer sur les paramètres d_{max} et g_{min} Plus la valeur du Deepsplit est grande, plus la valeur du d_{max} sera grande et plus la valeur du g_{min} sera petite; ainsi, un cluster sera plus facilement considéré comme un module de co-expression et il sera donc de plus petite taille. Cela signifie également que plus les modules sont petits, plus la co-expression des gènes du module est forte.

Caractérisation d'un module :

Pour chaque module, un « eigengene » est calculé, correspondant à la première composante principale de la matrice de variance-covariance établie à partir des valeurs d'expression des gènes du module. Cette dernière étape peut permettre de faire à nouveau une classification hiérarchique mais à partir des valeurs des « eigengenes » afin de regrouper les modules qui ont des profils d'expression très similaires. L'étape de

regroupement est une option par défaut (« Merge »).

2.2.Données d'expression de gènes

Les données d'expression de gènes utilisées pour l'inférence des réseaux de co-expression sont celles du consortium GTEX filtrées (Les gènes avec une moyenne d'expression par tissu < 0.1 RPKM pour tous les tissus et avec une variance d'expression nulle au travers des tissus sont éliminés), log-transformées ($\log_2$(RPKM + 1)) et ajustées (effets techniques et biologiques) (voir Méthode Chapitre 3). Les gènes pris en compte sont uniquement les gènes paralogues (10335 gènes).

La matrice de données utilisée dans WGCNA est :

$$X = [x_g^s]$$

où g correspond à l'indice des gènes paralogues (10335 gènes) et s correspond à l'indice de tous les échantillons des 13 tissus cérébraux (1259 échantillons).

2.3.Familles de gènes

Les familles de gènes comparées aux modules de co-expression produits par WGCNA sont celles définies dans la section Méthode du Chapitre 2.

2.4.Analyses d'enrichissements de gènes

Enrichissements de gènes en termes ontologiques :

Des enrichissements en groupes de gènes décrits dans (GO) sont effectués sur les modules de co-expression ayant une taille supérieure à 20 gènes. Ces tests d'enrichissements en catégories GO sont effectués sur 81 modules.

L'outil utilisé pour faire ces enrichissements est le package GOStat (version 2.42.0) sous R. Pour chaque module, un enrichissement dans le domaine des fonctions moléculaires et des processus biologiques est effectué. L'impact de la multiplicité des tests d'enrichissement sur leur significativité est pris en compte à l'aide de la correction de Bonferroni (Dunn 1959). Nous considérons le nombre de module pour appliquer un seuil de significativité des p-valeurs obtenues. Le seuil appliqué est de de 6,17 e-04.

Enrichissements de gènes en voies de signalisation :

PANTHER classification system (Version 12.0 release 2017-04-13) est un outil permettant de faire des analyses de surreprésentations en termes d'ontologie de gènes et en voie de signalisation biologiques. Cette méthodologie exploite les ressources Gene Ontology (release 2017-05-25) et Reactome pathway (Version 58 release 2016-12-07) (Mi et al. 2017).

2.5.Comparaison de modules de co-expression aux familles de gènes

Les modules de co-expression de gènes paralogues produits par WGNA ont été aussi extraits pour les comparer aux familles de gènes paralogues.

Afin de pouvoir comparer les gènes appartenant à chaque module avec les gènes appartenant à chaque famille, nous étudions la proportion de gènes d'une famille contenus dans un module et inversement (la proportion de gènes d'un module contenus dans une famille). Nous considérons une liste de modules (ou groupes de gènes) à tester et une liste de familles (ou groupes de gènes) de référence. Nous obtenons une première matrice de la proportion du contenu de chaque groupe testé dans chaque groupe de référence et une seconde matrice pour l'inverse.

Dans ce travail, nous nous intéressons particulièrement aux familles dont les gènes sont contenus dans un seul module de co-expression afin de retrouver les familles de gènes qui possède la majorité de leurs gènes co-exprimés. Une famille de gènes est définie comme homogène, si plus de 60% des gènes qui la composent sont contenus dans un même module de co-expression.

3.Résultats

3.1.Paramétrage de WGCNA

Nous utilisons les données de GTEx (voir Méthode Chapitre 2) comprenant des échantillons correspondants aux 13 tissus du cerveau. Les données considérées ont été précédemment filtrées et ajustées (voir Méthode Chapitre 2). Nous appliquons WGCNA sur les valeurs d'expression des gènes sur tous les échantillons et sur tous les tissus afin de regrouper en module les gènes qui ont des profils d'expression similaires au travers des 13 tissus cérébraux.

Nous faisons le choix de générer les résultats avec WGCNA uniquement sur les gènes paralogues. L'intérêt de ne considérer que les gènes paralogues est de pouvoir comparer

les modules aux familles de gènes paralogues afin de retrouver les familles de gènes co-exprimées.

Nous cherchons également à capturer la co-expression des familles de gènes composites, constituées de profils de gènes corrélés et anti-corrélés, par conséquent nous faisons le choix d'une approche WGCNA non signée considérant de la même façon ces deux types d'associations.

Nous estimons ensuite un paramètre de seuillage doux β qui permet de respecter une topologie de réseau invariant d'échelle (« scale-free ») (voir Méthode). Afin de déterminer ce paramètre β, WGCNA calcule les matrices d'adjacence pour une gamme de valeurs β. Pour choisir le seuillage doux optimal, il faut à la fois que la valeur β permette R^2>0.8 (voir Méthode Chapitre 4) mais que la connectivité reste la plus forte possible sous cette condition. Le paramètre de seuillage doux optimal obtenu est de 6. La visualisation de la matrice d'adjacence (Figure 37), permet de mettre en évidence les réseaux invariants d'échelle qui vont alors être générés.

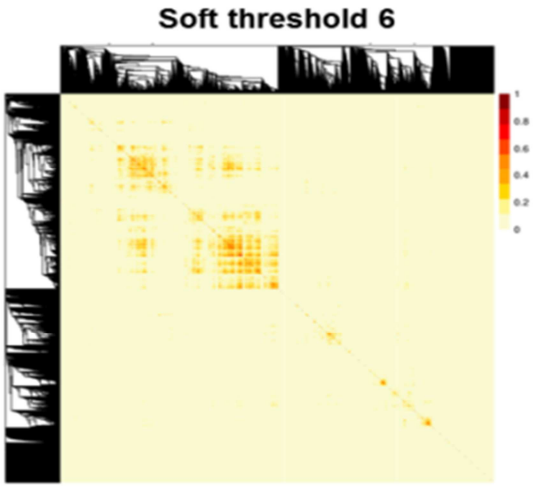


Figure 37: Matrice d'adjacence générée avec un paramètre β de 6

Heatmap (biclustering) de la matrice d'adjacence à partir des corrélations et des anti-corrélations entre l'expression des gènes paralogues au travers de 13 tissus cérébraux.

Optimisation des paramètres :

Afin de pouvoir comparer les modules de co-expression aux familles de gènes, nous cherchons à obtenir des modules de taille comparable aux familles. Nous avons remarqué dans le Chapitre 2 que les familles étaient de petite taille, en effet 47% des 3692 familles sont de taille 2. Par conséquent, afin de comparer de manière optimale les familles aux modules de co-expression, les modules générés par WGCNA doivent être de

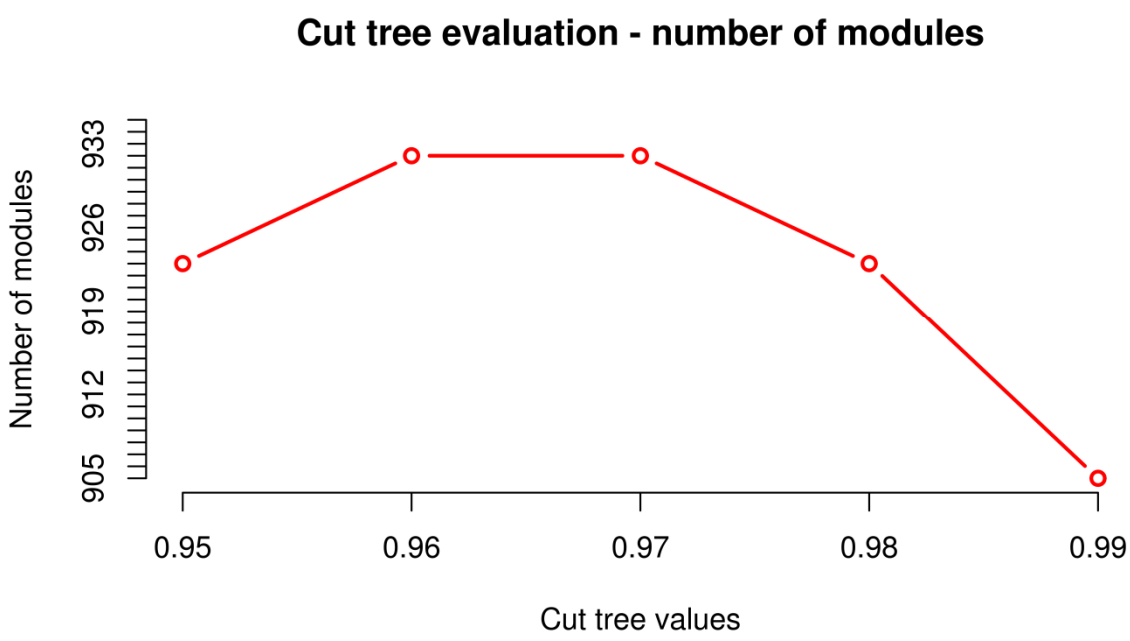
petite taille. Les paramètres de WGCNA agissant sur la taille des modules sont le Cuttree, le Deepsplit et la dernière étape de classification des « eigengenes » (« merge ») (voir Méthode Chapitre 4). Nous évaluons ces différents paramètres afin de réduire la taille des modules de co-expression tout en respectant la propriété « scale-free » des réseaux de gènes générés par WGCNA.

Concernant le paramètre Cuttree, nous évaluons l'impact du changement de valeurs de 0,95 à 0,99 sur la taille des modules de co-expression générés par WGCNA (la valeur par défaut de Cuttre est de 0,99) (Figure 38 A et B). Plus les modules sont de petite taille, plus ils sont nombreux. Le nombre de modules le plus élevé est atteint avec un Cuttree de 0,96 ou de 0,97. Pour les valeurs supérieures à 0,97, le nombre de module diminue signifiant que certains d'entre eux sont regroupés. Concernant les valeurs inférieures à 0,96, le nombre de modules diminue également, car dans ce cas, certains clusters ne sont plus détectés. Ainsi pour départager les deux valeurs (0,96 et 0,97) et pour vérifier que la taille des modules diminue, nous observons la taille des 20 plus grands modules dans les deux cas. Les tailles de modules sont plus petites pour un Cuttree de 0,97 (Figure 38 B**)**, donc nous faisons le choix d'utiliser cette valeur du Cuttree pour générer les modules de co-expression.

Concernant le paramètre Deepsplit, nous visualisons également la taille des 20 premiers plus gros modules obtenus avec un Deepsplit de 0, 2 et 4 (la valeur par défaut de Deepsplit est 1) (Figure 39). Comme attendu par la définition du Deepsplit (voir Méthode Chapitre 4), plus la valeur du Deepsplit est élevée, plus la taille des modules est petite. Ainsi nous choisissons donc une valeur de Deepsplit de 4.

Nous décidons de ne pas appliquer la dernière étape permettant de regrouper certains modules avec des « eigengenes » proches afin que les modules de petite taille ne soient pas regroupés.

A



B

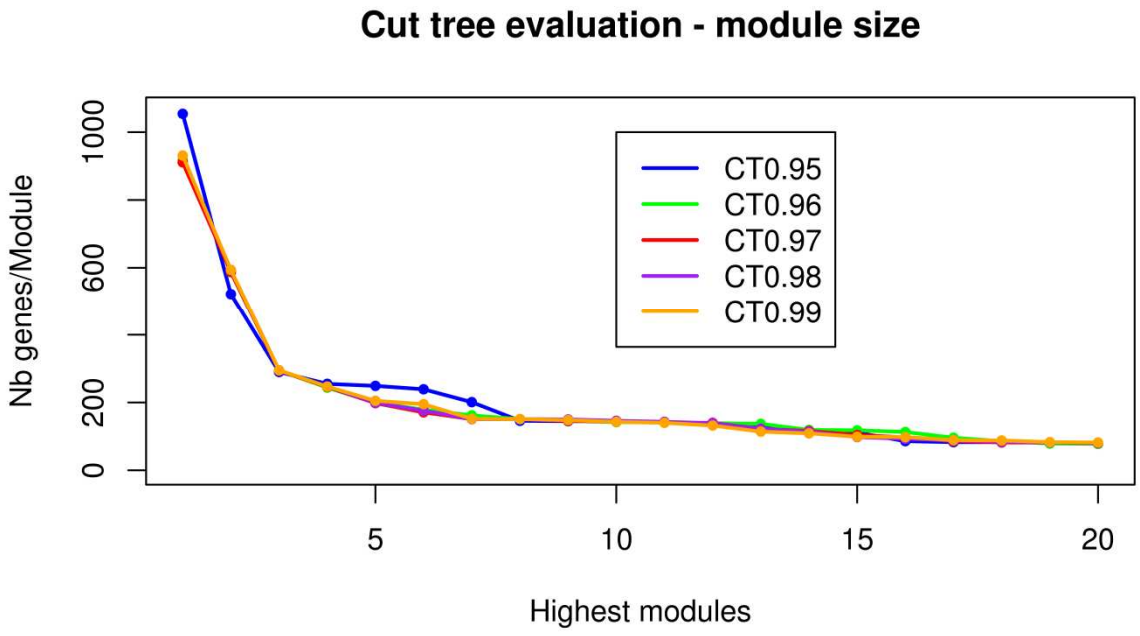


Figure 38: Modules générés pour différentes valeurs de Cuttree

A) Nombre de modules B) Taille des 20 plus gros modules générés à partir d'un Cuttree compris entre 0.95 et 0.99.

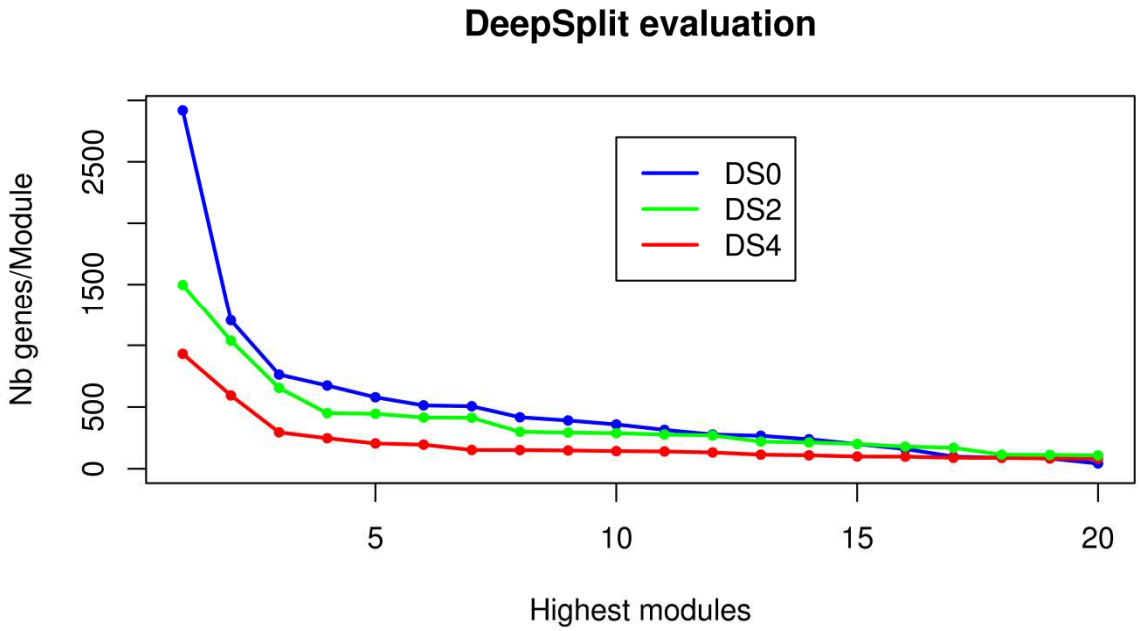


Figure 39: Taille des modules pour différents Deepsplit

Taille des 20 premiers plus gros modules obtenus avec un Deepsplit de 0 (bleu),2 (vert) et 4 (rouge).

3.2.Identification des modules de co-expression

Notre paramétrage de WGCNA nous a permis d'identifier 932 modules de gènes co-exprimés au sein des gènes paralogues. La taille des modules est comprise entre 2 et 911 gènes avec 84% des modules de petite taille (784 modules de moins de 10 gènes). Le filtre sur les gènes très faiblement exprimés au sein de tous les tissus (voir Méthode) permet d'avoir peu de gènes dans le module « grey » de gènes non co-exprimés (104 gènes).

L'efficacité de notre approche pour grouper en modules des gènes avec des profils d'expression proches, au travers des 13 tissus cérébraux, est illustrée par la visualisation des gènes du module antiquewhite1 (Figure 40) représentant pour chaque gène la moyenne d'expression des échantillons de chacun des tissus.

Afin de vérifier la pertinence biologique des modules de gènes co-exprimés, nous effectuons des tests de surreprésentation en termes d'ontologie de gènes (GO). Un test pour les termes des fonctions moléculaires et des processus biologiques est réalisé (avec GOStat) sur chaque module possédant au moins 20 gènes (81 modules). Les p-valeurs fournies par GOStat sont corrigées pour le nombre de termes testés (Benjamini & Hochberg 1995). Une correction de Bonferroni pour le nombre de modules testés est

aussi appliquée (Dunn 1959), donnant un seuil de p-valeur corrigé de 6,17e-04. Ainsi, 67% et 87% des modules testés sont enrichis au moins en une fonction moléculaire ou un processus biologique respectivement. Cette proportion élevée de modules enrichis en termes GO associés aux fonctions moléculaires et aux processus biologiques suggère que notre approche d'inférence de réseaux de gènes est capable de capturer des co-expressions biologiquement significatives au travers des tissus cérébraux.

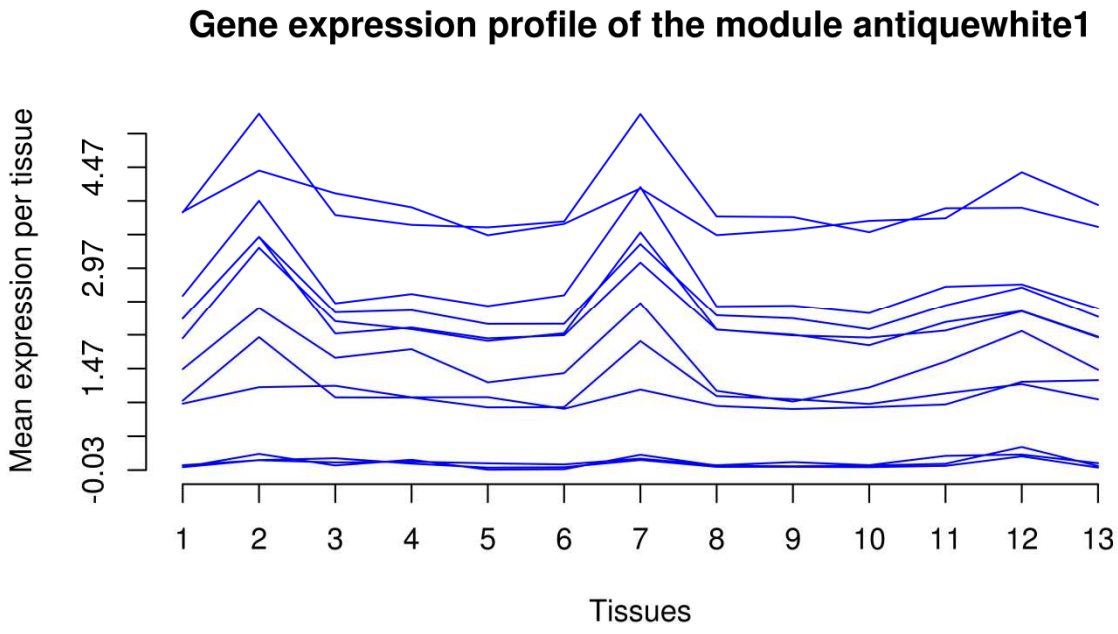


Figure 40: Profils d'expression des gènes d'un module de co-expression

Profils d'expression des gènes du module de co-expression antiquewhite1. 1- Cortex; 2- Cervelet; 3- Hippocampe; 4- Substance noire; 5- Cortex cingulaire antérieur; 6- Cortex frontal; 7- Hémisphère cérébelleux; 8- Noyau caudé; 9- Noyau accumbens; 10- Putamen; 11- Hypothalamus; 12- Moelle épinière; 13- Amygdale

3.3.Analyse de familles de gènes homogènes

Nous exploitons ensuite ces modules de co-expression pour classer les 3692 familles de gènes en deux catégories de familles, les familles homogènes et les familles hétérogènes. Nous définissons les familles homogènes comme celles ayant au moins 60% de leurs gènes contenus dans un seul module de co-expression qu'on appellera module principal. Les autres familles correspondent aux familles hétérogènes.

Concernant les familles homogènes, les gènes de ces familles n'appartenant pas au module principal ne seront pas pris en compte pour toutes les étapes de caractérisation de ces familles homogènes, notamment dans le chapitre 5.

Parmi les 3692 familles, 107 d'entre elles sont considérées comme homogènes dont 51 sont totalement incluses dans un seul module. Au total 15 modules sont des modules principaux pour les familles homogènes, ce qui signifie qu'ils contiennent en moyenne plusieurs familles homogènes.

La caractérisation des familles homogènes débute par des études d'enrichissements en termes GO et en voies de signalisation (Annexe B) avec l'outil PANTHER (Version 12.0) pour un seuil sur les p-valeurs corrigées de 0,05 (correction Bonferroni pour le nombre de termes testés). Les enrichissements en fonctions moléculaires et processus biologiques montrent que les gènes de ces familles homogènes sont particulièrement impliqués dans des mécanismes cellulaires fondamentaux comme la régulation de transcription ou le développement embryonnaire. De plus, nos analyses ont révélé que ces familles homogènes contenaient les familles des facteurs de transcription *AP2* (*TAP2*), *HOX* et des gènes associés à la voie de signalisation *NOTCH*, tous connus pour être impliquées dans le développement neural (Prince & Pickett 2002; Eckert et al. 2005).

4.Discussion et conclusion

De nombreuses études, dont des travaux du consortium GTEx, ont utilisé la méthode WGCNA (Zhang 2003; Konopka et al. 2012; Voineagu et al. 2013; Ardlie et al. 2015; Pierson et al. 2015) pour extraire des modules de gènes co-exprimés.

Nous avons également appliqué cette approche WGCNA, mais pour la première fois sur les gènes paralogues dans leur contexte fonctionnel du cerveau afin de révéler les modules de co-expression de gènes. La corrélation des profils d'expression sous-jacente à WGCNA est une métrique qui permet d'identifier des relations linéaires du niveau d'expression des gènes. Il est reconnu que ce type de modèle permet déjà la production d'interprétations et d'hypothèses pertinentes et riches. Il s'agit donc d'un premier niveau de modèle qui pourrait être complété par d'autres qui modéliseraient des liens non linéaires.

Contrairement aux précédentes études sur les données du consortium GTEx, nous avons fait le choix d'inférer un unique réseau de co-expression en regroupant tous les échantillons sur tous les tissus afin d'identifier les modules de co-expression des gènes paralogues au travers des tissus cérébraux. Les études d'inférence de réseaux de co-expression qui travaillent par tissu ont souvent pour objectif de déterminer les co-

expressions spécifiques à un tissu donné. Nous avons fait le choix d'explorer les réseaux de co-expression générés au travers des tissus cérébraux pour aborder d'autres questions, comme celle de l'homogénéité d'expression au sein des familles de gènes,et d'adresser celle de la tissue-spécificité d'expression des gènes par un score dédié calculé par gène (voir Chapitre 3).

Un de nos objectifs étant de comparer les modules de co-expression aux familles de gènes et sachant que ces familles sont globalement de très petite taille, nous avons décidé d'optimiser le paramétrage de WGCNA pour qu'il infère des petits modules de co-expression composés de gènes fortement corrélés en expression. Parmi les familles de gènes, nous avons identifié des familles homogènes contenant des gènes avec un profil d'expression similaire au travers des tissus cérébraux. Les études d'enrichissement nous ont permis de montrer que ces familles homogènes étaient enrichies pour les familles des facteurs de transcription *AP2*, *HOX* et pour les gènes impliquées dans la voie de signalisation de *NOTCH* (Prince & Pickett 2002; Eckert et al. 2005). De plus, ces familles homogènes ont tendance à être impliquées dans des fonctions fondamentales de la cellule comme la régulation de la transcription et également dans le développement neural.

Des travaux précédents sur l'expression des paralogues ont déterminé que les paires de gènes paralogues ont tendance à être co-exprimées juste après l'événement de duplication, puis évoluent pour finalement être exprimées dans différents tissus par la sous-fonctionnalisation ou la néo-fonctionnalisation (Lan & Pritchard 2016). A partir de nos analyses de co-expression, nous avons identifié un nombre relativement faible de familles de gènes homogènes, selon notre définition des familles homogènes. Les familles identifiées comme hétérogènes sont constituées d'une majorité de gènes qui ne sont pas co-exprimés, comme par exemple des gènes exprimés dans différents tissus en raison vraisemblablement d'une sous-fonctionnalisation ou d'une néo-fonctionnalisation. La fonctionnalisation des paralogues étant associée à leur rétention et donc à une divergence au cours du temps, nous pouvons faire l'hypothèse que les familles homogènes sont composées de gènes ayant peu divergés, donc issus d'une duplication récente, ou encore ayant subi une sous-fonctionnalisation de leur rôle biologique et non de leur expression.

Pour la suite du projet nous allons évaluer si les gènes co-exprimés ont tendance à être issus de duplications récentes. Nous allons également étudier comment l'intégration des

informations de co-expression et de tissu-spécificité permettent de faire progresser notre compréhension de la biologie des gènes paralogues.

Chapitre 5 : Expression tissulaire et co-expression des gènes paralogues dans le cerveau humain

1.Introduction

1.1.Expression et tissu-spécificité des gènes paralogues

La co-expression des gènes paralogues d'une part et leur tissu-spécificité d'autre part ont été étudiées entre différents organes chez l'homme (Acharya & Ghosh 2016; Kryuchkova-Mostacci & Robinson-Rechavi 2016; Lan & Pritchard 2016). Ces approches ont cherché à comparer les différents types de duplications en fonction de leurs propriétés de tissu-spécificité et de co-expression. Une étude récente a également comparé la tissu-spécificité des gènes paralogues à celle de leur gène ancestral (orthologue singleton) dans une autre espèce (Kryuchkova-Mostacci & Robinson-Rechavi 2016). Ce travail de recherche entre espèces a permis de révéler des différences de tissu-spécificité en fonction de la datation des évènements de duplication. Le modèle proposé par cet article est que dans le cas des duplications de type SSD, suite à l'événement de duplication, la copie du gène dupliqué avec l'expression la plus faible tend à devenir rapidement plus tissu-spécifique que le gène ancestral (Figure 41) et la seconde copie semble garder la tissu-spécificité du gène ancestral. Cet effet a pu être observé uniquement pour les duplications récentes (ySSD) (Figure 41) ce qui montre une tendance des ySSD à être plus tissu-spécifiques.

De plus, une étude de corrélation d'expression entre paires de paralogues humains, issus de différents types de duplications, a montré que les profils d'expression entre paires SSDs étaient significativement plus corrélés que ceux des paires WGDs pour différents degrés de divergence de séquence (Acharya & Ghosh 2016). Cette divergence était représentée par les dNs correspondant aux nombres de mutations non synonymes (K_a) et représentant la divergence de séquence des gènes (Figure 42). De plus, cette différence de corrélation est d'autant plus forte que le degré de divergence, qui est lié à l'âge de la duplication pour les SSD, est faible. La corrélation d'expression entre paires de SSDs suggère que dans le cas où les deux gènes de la paire SSD sont tissu-spécifiques, les deux gènes tendent à être spécifiques au même tissu, ce qui a été confirmé sur les ySSDs par l'étude de la tissu-spécificité des gènes paralogues (Kryuchkova-Mostacci & Robinson-Rechavi 2016). Au contraire, la corrélation d'expression plus faible entre paires de WGD leur est cohérente avec le fait que lorsque les deux paralogues d'une

paire datant de la période des WGD sont tissu-spécifiques, ils tendent à être spécifiques dans différents tissus (Acharya & Ghosh 2016; Kryuchkova-Mostacci & Robinson-Rechavi 2016).

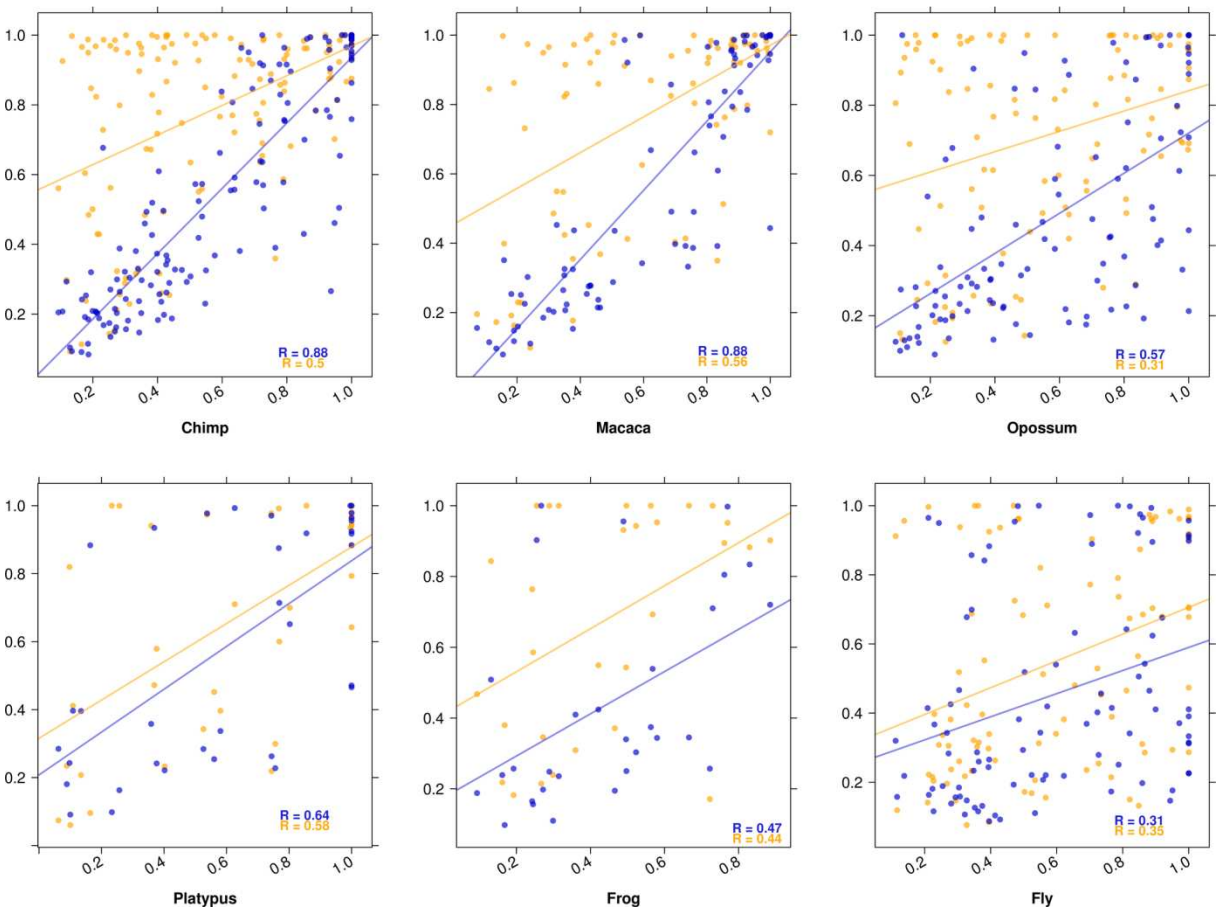


Figure 41: Distribution de la tissu-spécificité des paralogues comparée à leur orthologues

(Kryuchkova-Mostacci & Robinson-Rechavi 2016).Distribution de la tissu-spécificité (scores Tau) des paralogues humains (axes-y) comparé à l'orthologue singleton d'une autre espèce (axe-x). Représentation de la corrélation entre la tissu-spécificité du paralogue avec le gène orthologue. En bleu sont représentés les paralogues les plus exprimés de la paire de dupliqués et en orange les paralogues les moins exprimés.

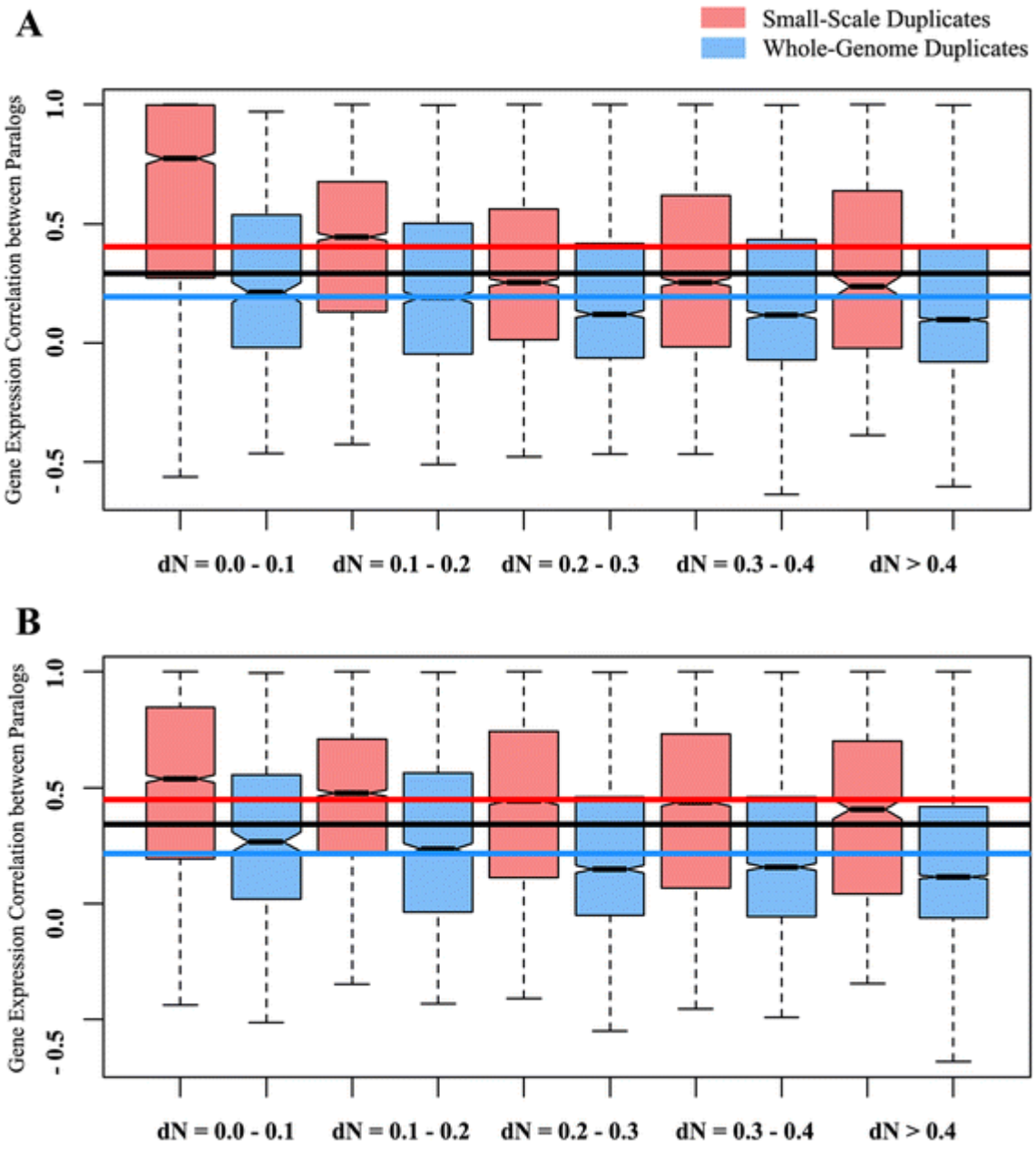


Figure 42: Correlation de l'expression des paires SSD et WGD

(Acharya & Ghosh 2016). Différence de corrélation de l'expression des gènes entre les paires de SSD et de WGD chez l'homme. Les valeurs de corrélation d'expression des gènes des paires SDD ou WGD ont été calculées à partir de données d'expression issues d'un séquençage RNA-seq de A) Human Protein Atlas ety B) Expression Atlas. Les SSDs sont représentés en rouge et les WGDs en bleu. Les lignes rouges et bleues représentent la moyenne des corrélations de l'expression des gènes des paires SSD et des paires WGD respectivement. La ligne noire représente la moyenne des corrélations de l'expression de toutes les paires de gènes paralogues.

1.2.Co-localisation génomique des gènes paralogues

De par les mécanismes de duplication, les paires de gènes paralogues peuvent avoir des positions relatives sur le génome qui sont diverses. Les deux gènes peuvent être co-localisés sur un même chromosome à distance variable ou au contraire sur différents chromosomes. Au sein d'une paire de paralogues, la co-expression peut être influencée par la distance génomique 1D ou la distance spatiale 3D séparant les deux paralogues. Récemment, l'analyse des matrices de contact de la chromatine générées par des expériences Hi-C *in situ* a établi que la co-localisation spatiale des régions génomiques, à l'intérieur et entre les chromosomes, est associée à une co-régulation de l'expression des gènes (Dixon et al. 2012; Schmitt et al. 2016). Les gènes fonctionnellement associés (gènes co-exprimés, gènes dont les produits sont liés dans les interactions protéine-protéine, gènes appartenant aux mêmes voies de signalisations) ont tendance à être surreprésentés au sein des régions génomiques identifiées par les cartes de contact de la chromatine (Xie et al. 2016). De plus, ces expériences de co-localisation de la chromatine ont montré qu'au sein des paires de gènes dupliqués sur différents chromosomes, la co-localisation spatiale 3D est plus fréquente parmi les paralogues provenant d'une WGD que d'une SSD. Ce résultat a été obtenu par la comparaison des fréquences de co-localisation des paires de paralogues dans les régions de contact chromatinien par rapport à la fréquence de co-localisation de paires de gènes sélectionnées aléatoirement (Figure 43).

Concernant les gènes localisés sur le même chromosome, les paires de gènes de duplications récentes (ySSD) sont plus fréquemment organisées en tandem que les duplications plus anciennes (Lan & Pritchard 2016). De plus, dans cette étude, il est également montré que les paires de duplications récentes sont plus co-exprimées que des duplications plus anciennes ce qui confirme des travaux antérieurs sur la corrélation de l'expression des gènes dupliqués par un évènement de SSD récent (Acharya & Ghosh 2016).

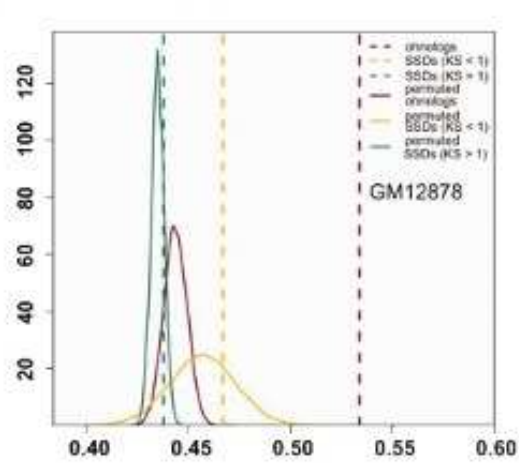


Figure 43: Fréquence de co-localisation entre paires de gènes WGD ou SSD sur différents chromosomes

Co-localisation des paires de gènes WGD ou SSD dans une lignée cellulaire lymphoblastoïde (GM12878) humaine. L'axe-x représente les pourcentages de paires co-localisées (3D) situées sur différents chromosomes. Les lignes verticales pointillées, marron, jaune et verte, indiquent la fréquence de co-localisations observées pour les paires WGDs, ySSDs et oSSD respectivement. Les courbes de même couleur montrent, pour chaque groupe de gènes, la distribution des fréquences de co-localisation pour 10000 permutations aléatoires des paires des données réelles pour le même nombre de paires.

1.3.Objectif

Des travaux publiés ont comparé les gènes SSDs aux gènes WGDs au regard de leur tissu-spécificité (Kryuchkova-Mostacci & Robinson-Rechavi 2016), leur co-expression (Acharya & Ghosh 2016) ainsi que leur distance génomique (1D ou 3D) (Lan & Pritchard 2016; Xie et al. 2016). D'après ces études, il semblerait que les paires de gènes de duplication récente soient plus tissu-spécifiques et soient plus co-exprimées et plus proches en terme de distance génomique en 1D par rapport aux duplications plus anciennes. Cependant, les paires de gènes de type WGD seraient davantage co-régulées du fait de leur co-localisation spatiale (3D), et seraient tissu-spécifiques dans des tissus différents dans le cas où les deux gènes de la paire seraient tissu-spécifiques.

Les questions que nous souhaitons adresser dans ce chapitre est de savoir si les tendances évolutives et d'expression associées aux différentes catégories de paralogues sont conservées dans le cas particulier du cerveau humain. Il serait également intéressant de confronter la co-expression à la localisation relative de ces paires de gènes.

Notre objectif est donc de comparer les deux types de duplications (WGD et SSD) mais aussi les groupes de SSDs à différentes datations des évènements de duplication. Nous focalisons nos études de comparaison sur les propriétés de tissu-spécificité, de co-

expression et de co-localisation génomique des différentes catégories de gènes paralogues au travers des tissus du cerveau.

2.Matériel et méthodes

2.1.Tests d'enrichissements

Les résultats de ce chapitre sont principalement basés sur des tests d'enrichissements. Pour chaque caractéristique analysée, nous examinons si la proportion de gènes possédant cette caractéristique particulière est significativement plus élevée dans un groupe de gènes testé que dans l'ensemble des gènes. L'existence d'une relation entre une caractéristique et un groupe de gènes est testée par l'application d'un test de Chi^2 (chisq.test de R) car nous possédons un nombre d'observations (nombre de gènes dans le test) suffisamment élevé. Nous considérons qu'il y a un enrichissement significatif si la p-valeur est inférieure à 0,05. Dans le cas de tests multiples pour une caractéristique de gènes donnée, nous effectuons une correction de Bonferroni (Dunn 1959) sur la p-valeur.

De plus le calcul de l'odds ratio (OR), à partir de la table de contingence, nous permet d'interpréter le résultat du test en terme d'enrichissement ou bien d'appauvrissement en gènes possédant la caractéristique analysée dans le groupe de gènes considéré.

Tableau de contingence :

Une table de contingence détaille le nombre d'observations contenues dans chaque groupe formé par une combinaison des deux variables testées. En général dans notre étude, nous considérons suivant les lignes de la table les gènes possédant ou non la caractéristique analysée (variable *i*). Suivant les colonnes, nous considérons les groupes des gènes soit le groupe 1 des gènes testés et le groupe 2 rassemblant le reste des gènes (variable *j*). Il ne doit pas y avoir de chevauchement entre les groupes de gènes en ligne (*i*) et en colonne (*j*).

Tableau de contingence des valeurs réelles du nombre d'observations *n* dans chaque groupe:

N_{ij}	Groupe 1	Groupe 2	Total *i*
Présente la caractéristique	n_{11}	n_{12}	$n_{1.}$
Ne présente pas la caractéristique	n_{21}	n_{22}	$n_{2.}$
Total *j*	$n_{.1}$	$n_{.2}$	$n_{..}$

Test du *Chi²*:

Le test du *Chi²* (Stigler 2008) est un test statistique paramétrique qui permet d'évaluer l'association entre deux variables qualitatives. Il se base donc sur une table de contingence et fait l'hypothèse d'une indépendance entre les variables. L'hypothèse d'indépendance des variables induit une symétrie du test et de la table de contingence.

Prédiction du nombre théorique d'individus dans chaque groupe correspondant à la distribution indépendante des individus:

$$n^*_{ij} = \left(n_{i.}.n_{.j}\right)$$

N^*_{ij}	Groupe 1	Groupe 2	Total *j*
Présente la caractéristique	n^*_{11}	n^*_{12}	$n_{1.}$
Ne présente pas la caractéristique	n^*_{21}	n^*_{22}	$n_{2.}$
Total *i*	$n_{.1}$	$n_{.2}$	$n_{..}$

Calcul de l'écart à la valeur d'indépendance :

$$Dev_{ij} = n_{i.} - n^*_{ij}$$

Dev_{ij}	Groupe 1	Groupe 2	Total *j*
Présente la caractéristique	Dev_{11}	Dev_{12}	0
Ne présente pas la caractéristique	Dev_{21}	Dev_{22}	0
Total *i*	0	0	0

Calcul de la valeur du *Chi²* observée :

$$Chi^2_{ij} = Dev^2_{ij} \,/\, n^*_{ij}$$

Chi^2_{ij}	Groupe 1	Groupe 2	Total *j*
Présente la caractéristique	Chi^2_{11}	Chi^2_{12}	-
Ne présente pas la caractéristique	Chi^2_{21}	Chi^2_{22}	-
Total *i*	-	-	$Chi^2_{Obs} = \sum Chi^2_{ij}$

Afin de définir si un test est significatif, on compare la valeur du Chi^2_{Obs} aux valeurs théoriques du *Chi²*. On regarde la proportion de valeurs supérieures au Chi^2_{Obs}parmi les valeurs théoriques. Cette proportion correspond à la p-valeur qui permet de conclure sur la significativité du test statistique. On considère qu'un test est significatif si la p-valeur obtenue est inférieure au seuil de 0,05.

Dans le cas où le test est significatif, les proportions des individus dans chaque groupe ne sont pas indépendantes donc il y a un enrichissement significatif d'un des groupes d'individu de la variable *i* dans un des groupes de la variable *j*.

Calcul de l'odds ratio (OR) :

$$OddsRatio = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}}$$

Ce ratio permet de connaître la direction du test statistique.

Dans le cas où OR > 1 :

- enrichissement en gènes du groupe présentant la caractéristique dans le groupe 1 par rapport au groupe 2 ;
- enrichissement en gènes du groupe 1 dans le groupe présentant la caractéristique par rapport au groupe ne présentant pas la caractéristique.

Dans le cas où OR < 1 :

- appauvrissement en gènes du groupe présentant la caractéristique dans le groupe 1 par rapport au groupe 2 ;
- appauvrissement en individus du groupe 1 dans le groupe présentant la caractéristique par rapport au groupe ne présentant pas la caractéristique
-

2.2.Distances génomiques au sein des paires de gènes dupliqués

La distance génomique (1D) au sein d'une paire de gènes_correspond au nombre de nucléotides (en paires de bases, pb) séparant les deux gènes, lorsque ceux-ci se trouvent sur le même chromosome. Ces distances ont été calculées à partir des coordonnées génomiques des gènes collectées sur Ensembl BioMart (Ensembl Genes 90, GRCh37). Une distance génomique a été obtenue, par paire de gènes dupliqués, à partir de la duplication la plus récente de chaque paire de paralogues (voir Chapitre 2).

2.3.Données d'expression de gènes

Nous avons utilisé les mesures d'expression par gène produites par GTEx sur 13 tissus cérébraux. Nous avons filtré et ajusté ces valeurs d'abondance par gène comme expliqué dans la section Méthode du Chapitre 3. Nous avons considéré différentes catégories de gènes comme définies dans le Chapitre 2. En résumé, sur les 16427 gènes codant pour des protéines, nous avons 5114 gènes WGD (« Whole genome duplication » - duplication

de génome entier), 3719 gènes SSD (« Small-scale duplication » - duplication à petite échelle), 1192 gènes ySSD(« younger SSD » - SSD plus récents que les WGD), 1260 gènes wSSD (« WGD-old SSD »- SSD datant de la période des WGD) et 1267 gènes oSSD (« older SSD » - SSD plus anciens que les WGD, 964 gènes WGD&SSD et 538 gènes sans annotation ainsi que 6092 gènes singletons. Les catégories testées dans ce chapitre correspondront aux gènes qui auront été retenus soit par une WGD, soit par une SSD uniquement. Les gènes WGD&SSD seront contenus dans le groupe de gènes non testés.
Concernant les paires de gènes paralogues, nous dénombrons 6134 paires de gènes paralogues au total dont 3091 paires de gènes WGD, 2291 paires SSD, 703 paires ySSD, 811 paires wSSD et 777 paires oSSD.

3.Résultats

3.1.Associations entre les catégories de gènes paralogues et la tissu-spécificité

Nous avons montré dans le Chapitre 3 que certains gènes paralogues possédaient une expression tissu-spécifique pour une région cérébrale en particulier. Nous cherchons à présent à estimer si cette tissu-spécificité d'expression des gènes dans les régions cérébrales est associée à certaines catégories de gènes paralogues. Nous exploitons les caractéristiques évolutives des paralogues correspondant au type de duplication et la date de l'évènement de duplication. Nous comparons les duplications de type WGD à celles de type SSD concernant leur influence sur la tissu-spécificité. Nous comparons aussi les différentes datations des évènements de duplication SSD : ySSD, oSSD et wSSD (voir Méthode Chapitre 2).
Il est possible qu'un gène ait subi les deux types de duplication (WGD et SSD) et qu'il ait été retenu à la suite de ces deux évènements. Cependant, quand nous nous référons à un type de duplication, nous considérons les gènes qui ont été retenus après ce type de duplication uniquement.
Nous avons précédemment observé dans le Chapitre 3 que les gènes paralogues étaient enrichis en gènes tissu-spécifiques par rapport aux singletons. A présent nous cherchons à savoir si les gènes issus d'une catégorie de duplication sont plus enrichis en gène tissu-spécifiques ce qui suggérerait un lien entre le type ou la date de duplication et la tissu-spécificité.

Premièrement, nous observons que les gènes paralogues tissu-spécifiques sont significativement enrichis en gènes SSDs (42,3% des paralogues tissu-spécifique contre 34,5% des paralogues non tissu-spécifiques, p-valeur = 7,308e-11). Afin d'estimer plus précisément l'influence du type de duplication, nous effectuons cette même analyse mais cette fois en considérant les duplications datant de la période des WGDs (WGDs et wSSDs). Nous observons que parmi ces paralogues les gènes tissu-spécifiques sont significativement appauvris en gènes WGD (72,7% des gènes tissu-spécifiques contre 80,2% des gènes non tissu-spécifiques, p-valeur =2,017e-08). Par conséquent, indépendamment de la date de duplication, les paralogues tissu-spécifiques semblent enrichis en gènes dupliqués à partir d'un évènement SSD (Table 6).

Deuxièmement, nous observons que les gènes paralogues tissu-spécifiques sont également enrichis en duplications récentes (gènes ySSD) (17,1% des paralogues tissu-spécifiques contre 10,2% des paralogues non tissu-spécifiques, p-valeur =1,270e-17). Quand nous effectuons le test parmi les gènes SSDs au lieu des paralogues, les gènes SSDs tissu-spécifiques restent enrichis en gènes ySSD (40,4% des SSDs tissu-spécifiques contre 29,7% des autres SSDs, p-valeur = 6,622e-09) (Table 6). Les gènes ySSD sont donc plus tissu-spécifiques que les autres gènes paralogues probablement dû à l'âge de leur duplication mais également à leur origine SSD.

Table 6: Enrichissements des gènes tissue-spécifiques entre groupes de gènes testés et de référence

Groupes de référence[a]	Groupes testés pour la tissue-spécificité[a]	P-valeurs du test Chi2[b]	Odds ratio[c]
Gènes codants pour des protéines	Gènes paralogues	2.045e-18*	1.48
	Gènes SSDs	8.571e-23*	1.57
	Gènes ySSDs	3.179e-26*	2.03
Gènes paralogues	Gènes SSDs	7.308e-11*	1.39
	Gènes ySSDs	1.270e-17*	1.81
Gènes SSDs	Gènes ySSDs	6.622e-09*	1.60
Gènes WGDs + oSSDs + wSSDs	Gènes WGDs	3.555e-04*	0.80
Gènes WGDs+ wSSDs	Gènes WGDs	2.017e-08*	0.66

[a] Abréviations des catégories des duplications de gènes : WGD (« Whole Génome Duplication » - Duplication globale de génome), SSD (« Small-Scale Duplication »-Duplication à petite échelle), oSSD (« old SSD » - duplication SSD plus anciennes que les WGD), ySSD (« young SSD »- duplication plus récente que les WGD) et wSSD (« WGD-old SSD » - SSD datant de la période des WGD).

[b] Application des tests du Chi2 avec un seuil de p-valeur corrigée = 6,25e-03 (correction de Bonferroni pour 8 tests statistiques).

[c] L'odds ratio (>1 ou <1) indique pour chaque groupe (testé ou non testé respectivement) qu'il y a un enrichissement.

3.2.Exploration des familles homogènes

Les gènes paralogues sont regroupés par famille et comme nous l'avons établi au chapitre précédent, ils sont aussi insérés au sein de modules de co-expression particuliers pour les tissus du cerveau. Nous nous proposons d'étudier la relation potentielle entre les caractéristiques des gènes paralogues d'une part et d'autre part les groupes de gènes se présentant à la fois comme familles de gènes et modules de co-expression dans les tissus du cerveau. Ainsi que défini précédemment de tels groupes de gènes sont les familles des gènes dites homogènes c'est-à-dire les familles contenant une majorité de gènes (> 60%) inclus dans un seul module de co-expression WGCNA. Les familles restantes avec des gènes dispersés sur plusieurs modules de co-expression sont considérées comme hétérogènes. Au total ces familles homogènes représentent 263 gènes.

Tout d'abord, nous nous intéressons à l'enrichissement en un type de duplication spécifique dans les familles de gènes homogènes. L'évaluation de la proportion de gènes WGDs entre les familles homogènes et hétérogènes montre que les celles homogènes sont appauvries en gènes WGD (41,1% des gènes des familles homogènes contre 49,8% des gènes des familles hétérogènes, p-valeur = 6,3e-03).

Concernant la relation entre la date de duplication et les familles homogènes, nous trouvons qu'elles sont enrichies en ySSDs (21,7% des gènes des familles homogènes contre 11,3% des gènes des familles hétérogènes, p-valeur = 3,472 e-07). Ces mêmes résultats sont obtenus sur la sous-population de familles homogènes pour lesquelles la totalité de leurs gènes est inclue dans un seul module de co-expression.

Dans le cas où des gènes sont co-exprimés, si l'un d'entre eux est tissu-spécifique, ils devraient avoir tendance à globalement montrer un profil relativement tissu-spécifiques. Les familles homogènes sont enrichies en gènes ySSDs qui sont plus tissu-spécifiques que les autres gènes. Notre hypothèse est que les gènes de ces familles

devraient être enrichis en gènes tissu-spécifiques. Nous retrouvons donc un enrichissement significatif des familles homogènes en gènes tissu-spécifiques (32,7% des gènes des familles homogènes contre 18,9% des gènes des familles hétérogènes, p-valeur = 3,2e-08) (Table 7).

Table 7: Enrichissements en gènes des familles homogènes entre différents groupes testés et de références

Groupe de référence	Groupe testé de gènes des familles homogènes[a]	P-valeurs du test de Chi2[b]	Odds ratio[c]
Gènes paralogues	Gènes WGD	6.3e-03*	0.70
	Gènes ySSD	3.472e-07*	2.17
	Gènes tissu-spécifiques	3.2e-08*	2.09

Groupe de référence	Groupe testé de gènes des familles homogènes totalement inclues dans un module[a]	P-valeur du test de Chi2[b]	Odds ratio[c]
Gènes paralogues	Gènes WGD	4.565e-03*	0.55
	Gènes ySSD	4.702e-04*	2.30

[a] Abréviations des catégories des duplications de gènes : WGD (« Whole Génome Duplication » - Duplication globale de génome), SSD (« Small-Scale Duplication »-Duplication à petite échelle), oSSD (« old SSD » - duplication SSD plus anciennes que les WGD), ySSD (« young SSD »- duplication plus récente que les WGD) et wSSD (« WGD-old SSD » - SSD datant de la période des WGD).

[b] Application des tests du Chi2 avec un seuil de p-valeur corrigée = 8,3e-03 (correction de Bonferroni pour 5 tests statistiques).

[c] L'odds ratio (>1 ou <1) indique pour chaque groupe (testé ou non testé respectivement) qu'il y a un enrichissement.

3.3.Analyse de la co-localisation génomique des gènes paralogues

Nous évaluons si la co-expression entre deux dupliqués au sein d'une paire de paralogues peut être influencée par leur proximité sur le génome. Nous considérons trois catégories de paires de paralogues: les paires inter-chromosomiques (chaque gène se trouvant sur un chromosome différent), les paires intra-chromosomiques (les deux

gènes sont sur le même chromosome) et les paires dupliquées en tandem avec les deux gènes séparés par moins de 1 Mb (Lan & Pritchard 2016).

Pour les analyses qui vont suivre, nous allons confronter les distances génomiques séparant des paires de paralogues aux caractéristiques évolutives et d'expression de ces paires de gènes dans les tissus du cerveau) (voir Méthode Chapitre 5).

A partir des coordonnées génomiques associées à chaque paire de paralogues, nous allons chercher à savoir si cette paire de gènes dupliqués est sur le même chromosome et, si c'est le cas, calculer la distance génomique en nucléotides qui les sépare (voir Méthode Chapitre 5).

Co-localisation génomique des familles homogènes :

Parmi les gènes appartenant aux familles homogènes, nous sélectionnons uniquement les paires pour lesquelles les deux paralogues sont inclus dans le module principal de co-expression (130 paires) (voir Méthode chapitre 5). Nous observons que les familles homogènes sont appauvries en paires de paralogues inter-chromosomiques (64,6% des paires dans les familles homogènes contre 82,8% des paires dans les familles hétérogènes réparties sur différents chromosomes, p-valeur = 1,266e-07). De plus, nous trouvons que les familles homogènes sont enrichies en paires de gènes dupliqués en tandem (28,5% des paires des familles homogènes contre 11,1% des paires des familles hétérogènes séparées par moins de 1 Mb, p-valeur = 2.187e-09) (Table 8). Ces résultats semblent indiquer que la co-expression des paralogues des familles homogènes est favorisée par leur proximité sur le génome.

Co-localisation génomique des gènes paralogues en fonction des types et datations des duplications :

Nous souhaitons à présent évaluer l'influence du type et de l'âge de la duplication sur la proximité entre les deux gènes dupliqués d'une paire. Pour cela, nous étudions les associations entre la proximité des gènes au sein d'une paire de paralogues et leur catégorie de duplication (SSD, ySSD, wSSD, oSSD et WGD) (Table 8).

Nous observons que les paires de SSDs sont appauvries en paires de gènes présents sur différents chromosomes (69,7% des paires SSDs contre 90% des paires non SSDs, p-valeur = 1,803e-90). Elles sont également enrichies en duplications en tandem (21,8% des paires SSDs contre 5,3% des paires non SSDs, p-valeur = 9,61e-85). Nous obtenons

les mêmes résultats pour les paires ySSDs et nous observons que plus la duplication est récente, plus la proportion de paires dupliquées en tandem augmente (2,6% des paires oSSDs, 10,5% des paires wSSDs et 56% des paires ySSDs). À l'inverse, plus la duplication est récente, plus la proportion de paires inter-chromosomiques diminue (92,7% des paires oSSDs, 82,4% des paires wSSDs et 29,9% des paires ySSDs). En comparant les paires WGDs et SSDs au même âge, les paires de WGDs sont enrichies en paires de paralogues localisés sur différents chromosomes (96% des paires WGDs contre 82,4% des paires wSSDs, p-valeur = 9,065e-42).

En résumé, nos tests d'enrichissement sur des familles homogènes contenant des gènes co-exprimés au travers de tissus cérébraux, indiquent qu'elles ont tendance à être enrichies en SSDs, en particulier en ySSDs. De plus, elles sont associées à un profil d'expression tissu-spécifique. Enfin les paires de gènes dupliqués des familles homogènes ont plus tendance que les autres à être localisées sur le génome en tandem, probablement en raison de la tendance des paires de SSDs à être dupliquées en tandem

Table 8: Enrichissement sur la co-localisation des paires de gènes paralogues
A) Enrichissements des paires de gènes sur différents chromosomes entre groupes de gènes testés et de référence

Groupe de référence[a]	Groupe testé pour les paires de gènes sur différents chromosomes[a]	P-valeur du test du Chi2 [b]	Odds ratio[c]
Paralogues	SSD	1.803e-90*	0.25
	ySSD	0*	0.05
	wSSD	9.812e-1	0.99
	oSSD	1.812e-15*	2.97
	Familles de gènes homogènes[d]	1.266e-07*	0.38
WGD + wSSD genes	WGD	9.065e-42*	5.08

B) Enrichissement en paires de gènes en duplication en tandem entre différents groupes de gènes testés et de référence.

Groupe de référence[a]	Groupe testé pour les paires de gènes en duplication en tandem (Distance génomique < 1Mb)[a]	P-valeurs du test de Chi2[b]	Odds ratio[c]
Paralogues	SSD	9.61e-85*	4.94
	ySSD	0*	21.06
	wSSD	3.701e-1	0.89
	oSSD	1.327e-16*	0.18
	Familles de gènes homogènes[d]	2.187e-09*	3.17
WGD +wSSD	WGD	7.241e-49*	0.07

[a] Abréviations des catégories des duplications de gènes : WGD (« Whole Génome Duplication » - Duplication globale de génome), SSD (« Small-Scale Duplication »-Duplication à petite échelle), oSSD (« old SSD » - duplication SSD plus anciennes que les WGD), ySSD (« young SSD »- duplication plus récente que les WGD) et wSSD (« WGD-old SSD » - SSD datant de la période des WGD).

[b] Application des tests du Chi2 avec un seuil de p-valeur corrigée = 4,1e-03 (correction de Bonferroni pour 12 tests statistiques).

[c] L'odds ratio (>1 ou <1) indique pour chaque groupe (testé ou non testé respectivement) qu'il y a un enrichissement.

[d] Le groupe de référence pour les familles homogènes correspond aux familles homogènes et hétérogènes.

4.Discussion et conclusion

Des travaux publiés sur l'expression des gènes paralogues entre différents organes semblent indiquer qu'ils seraient plus tissu-spécifiques que les autres gènes codants pour des protéines (Freilich et al. 2006; Kryuchkova-Mostacci & Robinson-Rechavi 2016). Ces études montrent également que les ySSDs auraient tendance à être plus tissu-spécifiques que les WGDs (Kryuchkova-Mostacci & Robinson-Rechavi 2016).

Dans nos travaux, nous nous sommes intéressés au cerveau et nous avons montré que les gènes SSDs et plus particulièrement les gènes ySSDs avaient tendance à s'exprimer de manière spécifique dans les régions du cerveau, comparés aux autres catégories de gènes paralogues. De plus, nous avons trouvé que les wSSDs étaient également enrichis en gènes tissu-spécifiques par rapport aux autres paralogues du même âge (WGDs), ce qui suggère que la spécificité tissulaire n'est pas seulement associée à la jeunesse de la duplication mais aussi au type SSD de la duplication.

Dans leurs travaux (Lan & Pritchard 2016) ont également montré que les duplications récentes, c'est-à-dire les ySSDs, auraient tendance à être dupliquées en tandem sur le génome et que ces paires de gènes dupliqués en tandem seraient davantage co-exprimés que les autres paires de ySSDs. Cette co-expression suggère une co-régulation entre les paires de ySSDs potentiellement induite par des régions régulatrices partagées. Nous avons confirmé cette tendance tout d'abord par l'obtention d'un enrichissement significatif en type ySSD pour les paires dupliquées en tandem sur le génome. Nous avons également observé que plus l'événement de SSD est récent, plus la proportion de duplications en tandem est élevée. Cette tendance indique que la date de l'événement de duplication influence sur l'organisation en tandem des paralogues. Afin d'évaluer aussi l'influence du type de duplication, nous avons testé la proximité génomique des SSDs datant de la même période que les événements de WGD (wSSD) et trouvé que les wSSDs sont significativement enrichis en duplications en tandem par rapport aux WGDs.
L'étude précédemment mentionnée (Lan & Pritchard 2016) propose une explication à la localisation préférentielle des ySSDs en tandem sur le génome et fait l'hypothèse que les SSDs auraient été générés, au cours de l'évolution, plus fréquemment par des mécanismes de duplication en tandem que par des mécanismes de rétrotransposition (Hurles 2004). Ils font également l'hypothèse que les remaniements successifs des chromosomes, par des mécanismes de recombinaison du génome, expliqueraient la plus faible proportion des paires de SSDs plus anciens localisés en tandem.
Dans notre étude, nous avons alors abordé le lien entre la co-expression des paralogues dans le cerveau et la proximité génomique en raisonnant notamment au niveau des familles de gènes homogènes. D'après nos tests d'enrichissement, ces familles sont enrichies en paires de gènes situées en tandem sur le génome et sont appauvries en paires de gènes situées sur des chromosomes différents. De plus nous avons montré que les familles homogènes sont enrichies en gènes ySSD. Ces deux derniers résultats et le fait que les ySSDs soient enrichis en duplications en tandem semblent confirmer la co-expression des paires ySSD en tandem.

Nos résultats sont également en accord avec l'étude montrant une tendance à la co-expression des paires de SSDs et en particulier celles issues de duplications récentes, par rapport aux paires de WGDs (Acharya & Ghosh 2016). Ils sont aussi cohérents avec

l'étude montrant que lorsque les deux paralogues d'une paire de ySSDs sont tissu-spécifiques, ils sont spécifiques à un même tissu, tandis que les deux paralogues d'une paire tissu-spécifique datant des WGD sont spécifiques des tissus différents (Kryuchkova-Mostacci & Robinson-Rechavi 2016).

En conclusion, nous avons donc mis en évidence que le type de duplication SSD ainsi que le caractère récent de l'évènement de duplication, contribuaient significativement à la co-expression et à la tissu-spécificité des gènes paralogues dans le cerveau humain.

Chapitre 6 : Application des réseaux de co-expression de paralogues tissu-spécifiques pour l'exploration des gènes associés à une maladie cérébrale

1.Introduction

1.1.Implication des gènes paralogues dans les maladies

Les gènes paralogues contribuent à la complexité du génome du fait des phénomènes de neo et sous fonctionnalisation qui apparaissent à l'occasion de la conservation des gènes dupliqués dans le génome (Force et al. 1999).

Par ailleurs, ils peuvent être associés à des maladies. Par conséquent, il a été démontré que, même si la proportion de gènes associés à des maladies parmi les gènes paralogues est relativement constante au travers des différentes espèces (Dickerson & Robertson 2012) correspondant à différentes complexités, le nombre de gènes associés à des maladies augmente avec le nombre de gènes paralogues et donc avec la complexité de l'organisme.

l parmi les gènes associés à une maladie, jusqu'à 80% d'entre eux auraient été dupliqués au cours de l'évolution (Dickerson & Robertson 2012). Le phénomène de compensation entre les copies des gènes dupliqués, due à leur redondance fonctionnelle primitive, pourrait permettre l'accumulation de mutations sur l'un des deux gènes sans qu'elles puissent s'exprimer phénotypiquement ; certaines de ces mutations pourraient alors constituer un risque de prédisposition à des maladies. En effet, la divergence de fonction ou d'expression d'une des copies de la paire de gènes paralogues pourrait alors rompre la compensation fonctionnelle et contribuer à l'expression de la pathologie (Dickerson & Robertson 2012; Chen et al. 2013). De plus, il a été montré que les gènes associés à une maladie monogénique étaient enrichis en gènes paralogues (Figure 44) (Chen et al. 2013). L'exploration du type et de l'ancienneté des paralogues dans les maladies monogéniques a montré également un enrichissement en gènes paralogues issus d'une duplication ancienne de type WGD ou oSSD (Chen et al. 2013; Singh et al. 2014).

Si l'on considère les gènes essentiels (gène qui induit un phénotype létal ou stérile s'il subit une délétion) (Makino et al. 2009), ceux-ci s'avèrent significativement plus fréquents parmi les WGDs qu'au sein des paralogues SSDs. La proportion des gènes essentiels dans les SSDs est de 4.601 % contre 11.344 % dans les WGDs (p-valeur < 1e-04) (Makino et al. 2009; Acharya & Ghosh 2016). Cette caractéristique d'essentialité des

gènes WGDs, qui présentent une plus faible redondance fonctionnelle entre dupliqués, permettrait d'expliquer que leur mutation ou délétion aient tendance à causer des phénotypes plus létaux que l'altération des gènes SSDs (Makino et al. 2009).

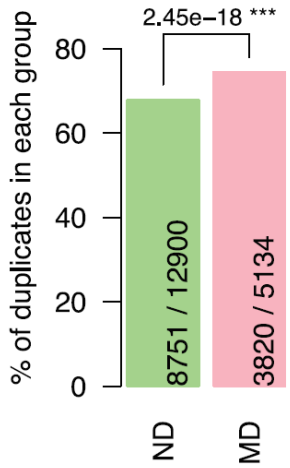


Figure 44 : Les gènes dupliqués sont enrichis en gènes impliqués dans les maladies monogéniques

Chen et al., 2013. Pourcentage de gènes dupliqués parmi les gènes impliqués dans les malaides monogéniques (MD-« monogenic disease genes ») et les gènes non imliqués dans les maldies (ND-« Non disease genes »). Ici les gènes dupliqués sont définis avec TreeFam. La p-valeur a été calculée à partir d'un Test Exact de Fisher ; Niveau de significativité: ***<0.001.

1.2.Régions cérébrales et fonctions biologiques spécifiques à la schizophrénie et à l'autisme

La schizophrénie et l'autisme sont deux syndromes psychiatriques qui provoquent des troubles avérés mais dont le diagnostic, en particulier le diagnostic précoce reste très difficile compte tenu de la non-spécificité des signes cliniques. Pour ces raisons, ces syndromes sont particulièrement étudiés actuellement et l'imagerie médicale est utilisée pour trouver les moyens de poser un diagnostic précis ainsi que mieux comprendre ces maladies. Actuellement, il est considéré que ces deux maladies ont des origines neuro-développementales et que des régions du cerveau peuvent être modifiées chez les sujets atteints (Buckley 2005; Ha et al. 2015).

L'autisme se caractérise par des troubles du comportement et est plus généralement dénommée trouble du spectre autistique (ASD- « Autism spectrum disorder ») (Ha et al. 2015). Les personnes souffrant de cette maladie ont des comportements répétitifs et des troubles de la communication et des émotions. Des études de neuroimagerie suggèrent

que l'organisation ou l'activation des aires cérébrales impliquées dans la communication, les émotions et la reconnaissance faciale seraient défaillantes (Ecker et al. 2015). Chez les patients souffrant d'autisme, la neuro-imagerie, et en particulier l'imagerie par résonnance magnétique (IRM), permet d'étudier des dysfonctionnements éventuels dans les régions cérébrales connues pour être le support de ces activités et comportements. Les séquences IRM fonctionnelles rendent compte de la connectivité fonctionnelle entre les régions cérébrales, tandis que les séquences IRM anatomiques permettent de mener des analyses volumétriques sur diverses structures du cerveau. Les analyses de volumétrie et d'activité des aires cérébrales ont montré que les régions du cortex frontal, temporal et cingulaire pouvaient être touchées chez les patients atteints d'autisme et pouvaient provoquer des déficits de la communication. De même, l'amygdale qui est une structure impliquée dans la gestion des émotions et la reconnaissance d'expression faciale, peut également être touchée. La connectivité fonctionnelle et structurale mesurée par IRM, a montré que les régions perturbées pouvaient concerner le thalamus ainsi que le cortex frontal, temporal et cingulaire.

La schizophrénie est une maladie dont l'origine vraisemblable est neuro-dégénérative et neuro-développementale. Les signes de cette maladie sont des hallucinations ainsi que des troubles cognitifs et de la pensée (pensées négatives) (Buckley 2005). L'IRM permet de mettre en évidence le caractère neuro-développemental de la maladie. Les régions impactées en terme de volume peuvent être le thalamus, le noyau caudé, le lobe temporal et le corps calleux. Le volume de matière grise peut également être réduit dans le cortex cingulaire frontal et temporal.

Des études récentes ont exploré le génome et le transcriptome d'individus atteints par l'une de ces maladies. En ce qui concerne le syndrome autistique, il possède une composante génétique et polygénique (Voineagu et al. 2013). Dans cette dernière étude, les auteurs ont effectué des analyses d'expression différentielle à partir de données de puces Illumina générés sur des échantillons post-mortem de trois tissus du cerveau (lobe temporal supérieur, cortex préfrontal et vermis cérébelleux), entre des patients atteints d'autisme et des sujets contrôles sains. Sur les 444 gènes identifiés comme différentiellement exprimés, des analyses d'enrichissement en catégories ontologiques de gènes ont montré une surreprésentation en fonctions biologiques actives au niveau des synapses ainsi que dans la réponse immunitaire inflammatoire.

La schizophrénie est également une maladie polygénique avec une part héréditaire. Une étude d'expression différentielle a été effectuée à partir de données puces Affymetrix d'échantillons post-mortem provenant du cortex préfrontal dorso-latéral de patients atteints de schizophrénie et de sujets contrôles sains (Hakak et al. 2001). A partir de cette analyse, 89 gènes ont été retrouvés différentiellement exprimés. Ces gènes étaient impliqués dans différents processus biologiques tels que la myélination, le développement neuronal, la neurotransmission (voies de signalisation du neurotransmetteur GABA) (Bouché et al. 2003), la transduction du signal ou la régulation du cytosquelette. Ces analyses ont effectivement permis de confirmer l'implication du cortex préfrontal dorso-latéral dans la schizophrénie (Bunney & Bunney 2000).

Un lien entre ces maladies a été également montré à partir de données RNA-seq de tissus post-mortem de la région corticale du cerveau humain (Ellis et al. 2016). En effet cette étude a montré que le transcriptome de l'autisme serait significativement corrélé au transcriptome de la schizophrénie.

1.3.Objectif

L'étude des profils d'expression et de régulation des gènes au travers de différentes régions cérébrales peut favoriser notre compréhension des maladies cérébrales associées à des mutations dans les paralogues.

L'objectif du travail reporté dans ce chapitre est donc d'intégrer les informations sur la co-expression, la spécificité tissulaire et les caractéristiques évolutives des paralogues pour étudier leur implication potentielle dans les maladies cérébrales que nous avons choisies qui sont la schizophrénie et l'autisme.

Dans un premier temps, nous allons présenter les listes de gènes associés aux maladies cérébrales et plus particulièrement à la schizophrénie et à l'autisme. Ensuite nous détaillerons les réseaux de co-expression de gènes paralogues contenant ceux issus de ces listes, annotés par leur tissu-spécificité ainsi que leur implication dans la schizophrénie ou l'autisme. Nous allons explorer ces réseaux afin d'améliorer la connaissance fonctionnelle et d'expression des gènes connus pour être impliqués dans ces deux maladies et ceux qui leurs sont associés.

2.Matériel et Méthodes

2.1.Gènes associés aux maladies cérébrales dans ClinVar

Gènes muté associés à une maladie cérébrale:

Nous avons utilisé la base de données ClinVar pour collecter des listes de gènes associés à des maladies cérébrales (Landrum et al. 2016). Nous considérons les gènes contenant une mutation pathogène de type « Single Nucleotide Variation » ou variation d'un seul nucléotide (SNV) ainsi que ceux localisés dans une « Copy Number Variation » ou variation du nombre de copies (CNV). Nous sélectionnons les gènes associés, dans ClinVar, aux maladies et phénotypes suivants : « Parkinson », « Alzheimer », « brain », « Autism », « Epilepsy », « Aicardi », « Angelman », « Aphasia », « Apraxia », « Asperger », « Behcet », « spinal », « Canavan », « Charcot », « Chorea », « Dementia », « Dyslexia », « Fabry », « Gaucher », « Gerstmann », « Huntington », « Refsum », « Joubert », « Kennedy », « Klippel », « Krabbe », « learning », « mental », « Leigh », « Leukodystrophy », « migraine », « Niemann », « Rett », « Sandhoff », « syncope », « Tay-sachs », « Tourette », « nervous », « Schizophrenia », « Narcolepsy », « neuro », « cephal », « cortico », « crani », « mening », « psych ». Nous obtenons un total de 9451 gènes impliqués dans des maladies cérébrales.

Gènes associés à la schizophrénie ou à l'autisme.

De la même façon, nous recueillons deux listes de gènes dans ClinVar associés au mot clé "Schizophrenia" et contenant une variation pathogène de type SNV ou CNV. Nous obtenons 384 gènes contenus pour être impliqués dans une CNV (51 variations) et 14 gènes avec au moins une SNV (20 variations). En ce qui concerne l'autisme, nous collectons uniquement la liste de gènes contenant au moins une SNV pathogène ce qui correspond à 53 gènes (85 variations). Nous n'effectuons pas l'analyse sur les CNVs liées à l'autisme car sont trop nombreuses pour nous permettre une analyse détaillée.

2.2.Données de co-expression et d'expression tissulaire

Les travaux de ce chapitre reposent sur les résultats obtenus, au cours des chapitres précédents, sur la co-expression et la tissu-spécificité des gènes dans les différentes régions du cerveau. Nous exploitons les modules de co-expression de gènes paralogues obtenus par la méthode WGCNA à partir de l'expression des gènes paralogues au travers des 13 tissus cérébraux (voir Méthode Chapitre 4).Nous exploitons également les gènes

paralogues identifiés dans le Chapitre 3 comme ayant une expression tissu-spécifique à une région cérébrale parmi les 7 régions du cerveau. Chaque gène tissu-spécifique est associé à la région dans laquelle il est le plus exprimé (voir Méthode Chapitre 3). Les gènes des modules de co-expression sont donc labellisés en termes de tissu-spécificité.
Enfin, nous annotons les modules à l'aide des listes de gènes associés aux maladies cérébrales collectées à partir de la base de données ClinVar (voir Méthode-Chapitre 6).
Les identifiants gènes de ce chapitre sont nommés suivant la nomenclature HGNC (voir Méthode Chapitre 2) également utilisée dans la représentation des réseaux.

2.3.Visualisation des réseaux de co-expression de gènes

La visualisation des modules de co-expression des gènes paralogues est réalisée avec l'outil Cytoscape (version 3.4.0) (Su et al. 2015). Chaque module de co-expression est représenté comme un réseau, où les arêtes représentent des liens de co-expression à l'intérieur du module de gènes et les nœuds correspondent à des gènes. Un gène associé à une maladie cérébrale référencé par ClinVar correspond à un nœud plus grand et est affiché comme un gène central (« hub ») à des fins de visualisation. Pour ces mêmes raisons, les autres nœuds sont donc reliés uniquement au gène central sur le graphe. La couleur de la bordure des nœuds des gènes impliqués dans une maladie est liée au type de maladie (noir : schizophrénie, orange : autisme, rouge : partagé par les deux maladies).
Les couleurs de remplissage des nœuds illustrent la région cérébrale pour laquelle un gène tissu-spécifique est le plus exprimé (bleu : cervelet-hémisphère cérébelleux, vert = Cortex-cortex frontal-cortex cingulaire antérieur, rouge : hippocampe-amygdale, gris : substance noire, violet : ganglions de la base, rose : hypothalamus, jaune : moelle épinière et blanc : pas de tissu-spécificité).

3.Résultats

3.1.Implication des gènes paralogues dans les maladies cérébrales

Chaque gène peut être caractérisé par son expression dans des conditions choisies, son réseau de co-expression, sa tissu-spécificité mais également par son association avec des maladies. Nous nous plaçons dans la lignée des études qui montrent l'importance des paralogues (Dickerson & Robertson 2012) et plus particulièrement ceux de type WGD

(Chen et al. 2013; Singh et al. 2014; Acharya & Ghosh 2016) dans les maladies, pour proposer une nouvelle priorisation des gènes d'intérêt.

Nous décidons d'affiner ces analyses déjà publiées sur l'implication des paralogues dans les pathologies humaines en focalisant sur les maladies cérébrales. Nous collectons depuis la base de données ClinVar (Landrum et al. 2016) une liste de gènes associés à une pathologie cérébrale et contenant une SNV ou présent dans une CNV.

Un test d'enrichissement permet de constater que ces gènes associés à une maladie cérébrale sont significativement enrichis en gènes paralogues (62,8% des gènes impliqués dans les maladies cérébrale contre 52,5% des gènes non identifiés comme impliqués dans une maladie cérébrale, p-valeur = 1,787e-49).

Si nous nous intéressons aux types de duplications, nous constatons que les gènes impliqués dans les maladies cérébrales ne sont ni enrichis ni appauvris en WGDs (p-valeur = 0.848). Cependant, comme nous avons trouvé dans le chapitre précédent que les gènes ySSD étaient tissu-spécifiques et co-exprimés dans les différentes régions et tissus cérébraux, nous souhaitons à présent tester l'enrichissement des gènes associés aux maladies cérébrales en gènes ySSD. En effet, les gènes impliqués dans les maladies cérébrales sont significativement enrichis en ySSD (11,6% des gènes impliqués dans les maladies cérébrales contre 8,9% des gènes non identifiés comme impliqués dans une maladie cérébrale, p-valeur = 1,512e-10).

Par conséquent, les paralogues issus de duplications récentes semblent préférentiellement accumuler des mutations génétiques associées à des troubles cérébraux par rapport à d'autres types de paralogues.

3.2.Réseaux de co-expression des gènes associés à la Schizophrénie ou à l'autisme

Nous cherchons maintenant à améliorer notre connaissance sur les gènes associés aux maladies cérébrales en intégrant les caractéristiques que nous avons obtenus sur leurs réseaux de co-expression avec des paralogues et leur spécificité tissulaire cérébrale. Nous décidons d'illustrer la pertinence de notre approche en l'appliquant aux gènes référencés par la base de données ClinVar (Landrum et al. 2016) comme impliqués dans la schizophrénie ou l'autisme (SCZ/ASD). Deux listes de gènes porteurs d'altérations génétiques sont récupérées de ClinVar : une première liste de gènes reliés à des SNVs et CNVs associés à SCZ et une seconde liste de gènes porteurs de SNVs impliqués dans ASD

(voir Méthode Chapitre 6). Nous extrayons les modules de co-expression connectés à ces gènes et mettons en évidence la spécificité tissulaire de ces réseaux de gènes (Figure 45 et Figure 46).

Nous observons d'abord que lorsque les gènes d'un module sont tissu-spécifiques ils ont tendance à être spécifiques dans le même tissu (Figure 46).

De plus, nous remarquons dans le cas de SCZ et de ASD que les réseaux de co-expression centrés sur les gènes associés à ces pathologies, par des SNVs ou des CNVs, se distribuent préférentiellement dans les mêmes régions cérébrales à savoir le cervelet, la moelle épinière, l'hypothalamus, le cortex et les ganglions de la base .

Pour terminer, nous considérons le réseau de co-expression centré sur le gène *DHH,* dont la forme mutée est associée à la SCZ, et présentant une spécificité d'expression pour la moelle épinière (en jaune) (Figure 46A). Nous constatons que la quasi-totalité des paralogues de ce module sont tissu-spécifiques pour la moelle épinière et correspondent aux gènes de la famille *HOX*. Or, cette famille *HOX* a été retrouvée précédemment, dans le chapitre 4, comme homogène en expression. Ce résultat illustre la tendance, vue dans le chapitre 5, des familles homogènes à avoir une spécificité d'expression par région cérébrale, par rapport aux familles hétérogènes.

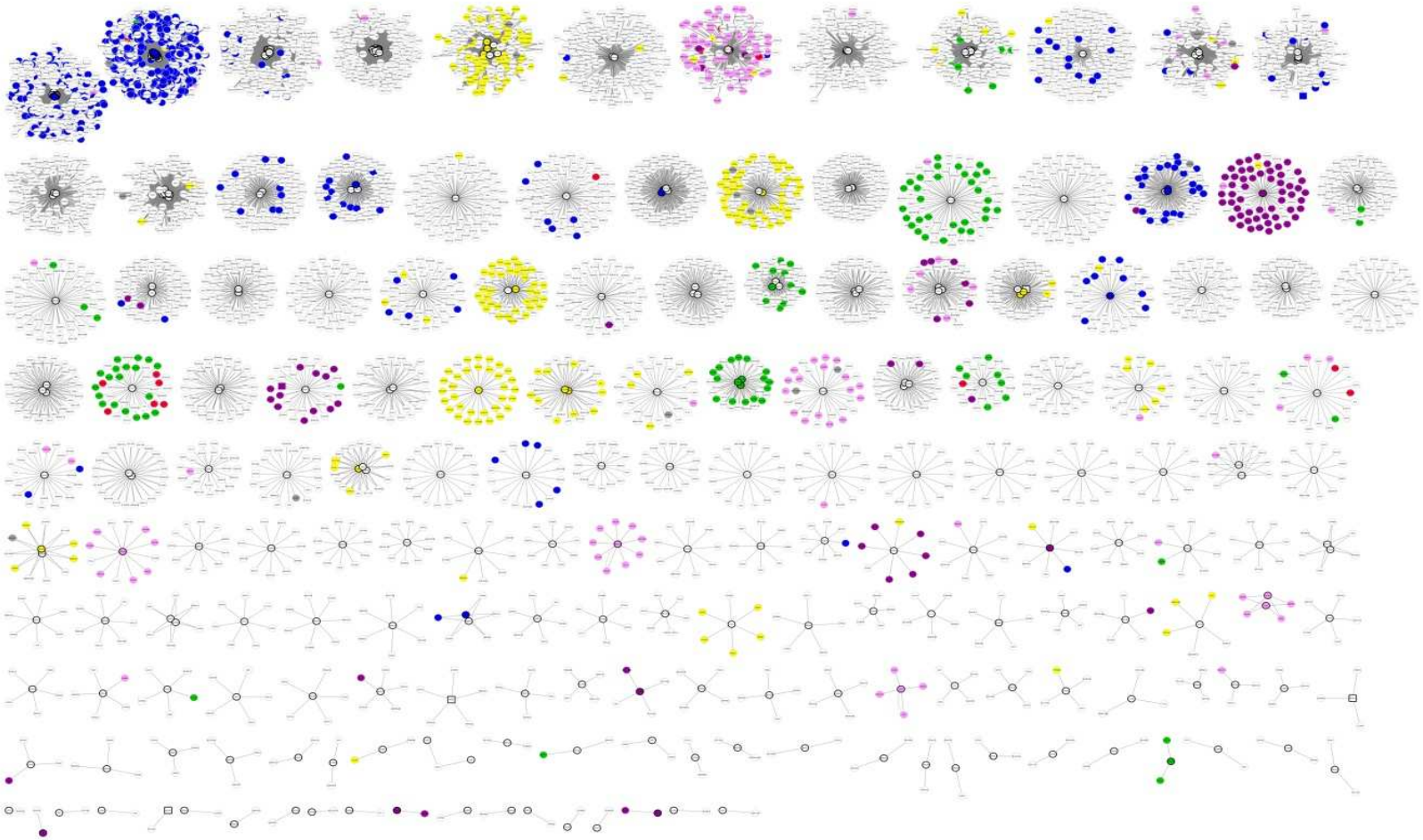
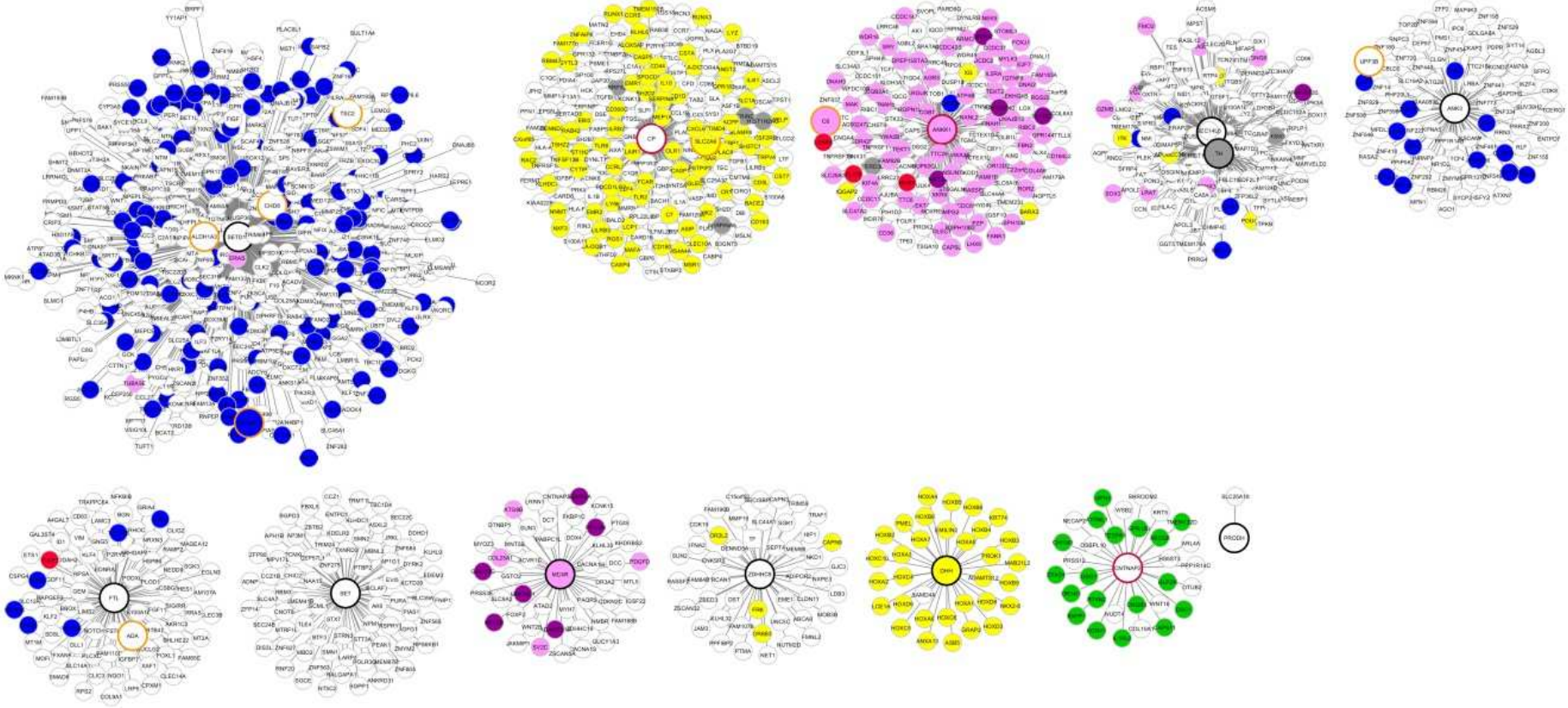


Figure 45: Réseaux de co-expression connectés aux gènes paralogues inclus dans une CNV associée à la SCZ

Réseaux de gènes paralogues co-exprimés, avec leur tissu-spécificité éventuelle, liés à des gènes associés au trouble de la SCZ. Chaque module de réseau de co-expression, où les nœuds représentent les gènes de ce module, comprend au moins un gène dans une CNV associée à la SCZ (base de données ClinVar). L'intérieur d'un nœud est coloré en fonction de la spécificité d'expression à une région cérébrale (rose : région de l'hypothalamus, bleu : région du cervelet, vert : région du cortex, violet : région du ganglion de la base, rouge : région de l'hippocampe-amygdale, gris : substance noire et blanc : pas de tissu-spécificité).

(a)



(b)

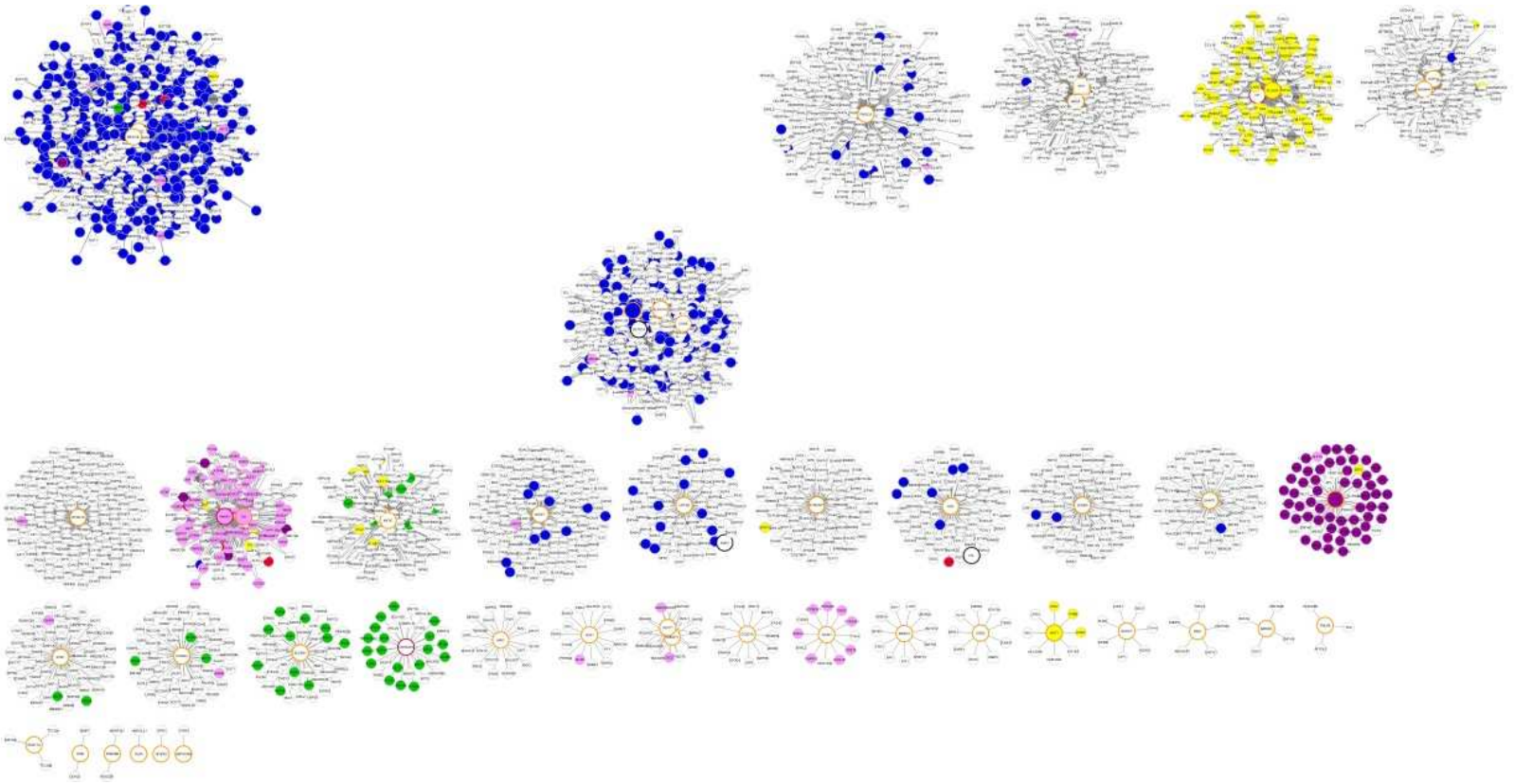


Figure 46: Réseaux de co-expression connectés aux gènes paralogues dont les SNV sont impliquées dans la SCZ ou dans ASD

Réseaux de gènes paralogues co-exprimés, avec leur tissu-spécificité éventuelle, liés à des gènes associés à la SCZ et à ASD. Chaque module de réseau de co-expression, où les nœuds représentent les gènes de ce module, comprend au moins un gène avec une SNV associée (base de données ClinVar) à (a) la SCZ et (b) ASD. L'intérieur d'un nœud est coloré en fonction de la spécificité d'expression à une région cérébrale: rose : région de l'hypothalamus, bleu : région du cervelet, vert : région du cortex, violet : région du ganglion de la base, rouge : région de l'hippocampe-amygdale, gris : substance noire et blanc : pas de tissu-spécificité. La bordure du nœud est colorée en fonction des gènes référencés par ClinVar associés à une maladie du cerveau (noir : gènes associés à la SCZ, orange : gènes associés à ASD et rouge : gènes associés à SCZ et à ASD).

3.3.Mise en évidence du gène *ANKK1*

Parmi les modules de co-expression des gènes associés à la schizophrénie et à l'autisme, nous pouvons mettre en évidence 3 gènes paralogues (*CP*, *ANKK1* et *CNTNAP2*) dont les formes mutées par des SNVs pathologiques, sont associées à ces deux maladies. Dans nos résultats, ces gènes montrent respectivement une expression tissu-spécifique pour la moelle épinière, l'hypothalamus et le cortex (Figure 46).

La connexion dans nos modules entre un gène préalablement associé à une maladie et au moins un gène paralogue co-exprimé et tissu-spécifique est une source d'interprétation du contenu de ces modules. En suivant une logique du type « guilty by association », cela peut nous permettre de mieux comprendre la fonction et l'influence des gènes des modules contenant un gène central associé à une maladie SCZ/ASD.

Pour illustrer ce potentiel, nous considérons le module de co-expression centré sur le gène *ANKK1* (Nymberg et al. 2014) (Figure 47). D'après la littérature, cette protéine appartient à la famille RIP sérine-thréonine kinase impliquée dans le processus de survie, de mort et de différenciation des cellules (Meylan & Tschopp 2005; Declercq et al. 2009; España-Serrano et al. 2017). Nous observons que *ANKK1* et un grand nombre des gènes du module auquel il appartient sont tissu-spécifiques à l'hypothalamus. Une analyse d'enrichissement sur ontologie des gènes de ce module nous permet d'observer une surreprésentation significative des gènes impliqués dans l'organisation du cytosquelette, l'activité motrice des microtubules, le mouvement du cilium et l'assemblage du complexe de la dynéine (Annexe C). Ces résultats sont à rapprocher de ceux de *Fukuda & Yanagi 2017* qui ont très récemment mis en évidence l'implication des anomalies du cytosquelette dans SCZ et ASD. Ainsi, l'exploration des réseaux de co-expression des gènes paralogues tissu-spécifiques peuvent nous renseigner sur les processus biologiques et la région du cerveau dans lesquels sont impliqués des gènes d'intérêt. Dans le cas de *ANKK1*, nous pouvons faire l'hypothèse que sa forme mutée, dans SCZ ou ASD, serait impliquée dans des dérégulations du cytosquelette impactant particulièrement l'hypothalamus du patient (Nymberg et al. 2014).

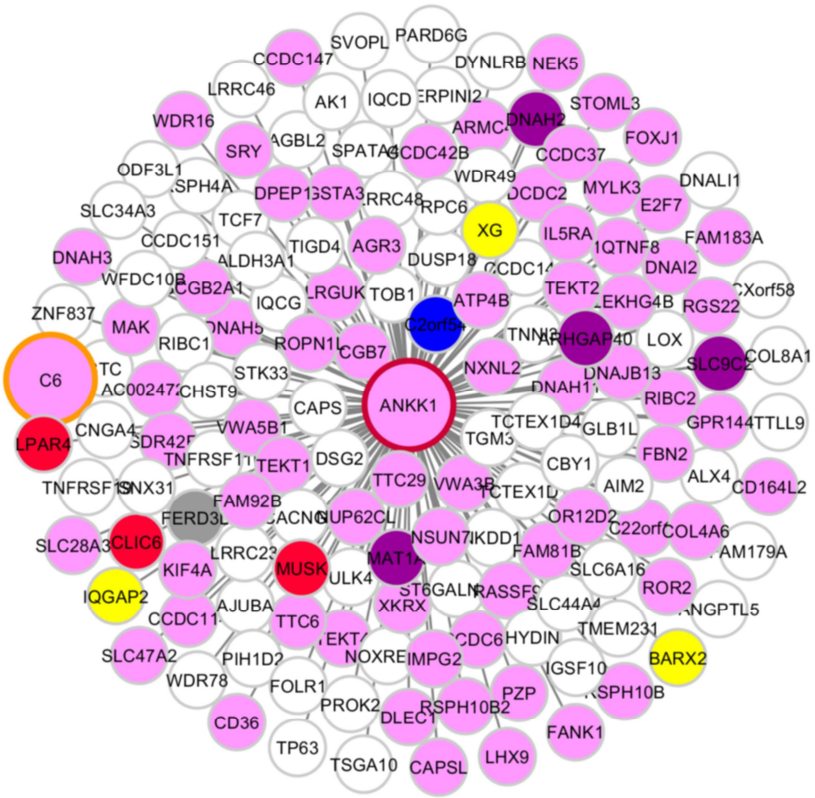


Figure 47: Réseau de co-expression du gène *ANKK1*

Réseau de gènes paralogues co-exprimés, avec leur tissu-spécificité éventuelle, liés à *ANKK1*. Les nœuds représentent des gènes dans le même module de co-expression que *ANKK1*. L'intérieur du nœud est coloré en fonction de la spécificité d'expression à une région cérébrale (rose : région de l'hypothalamus, bleu : région du cervelet, vert : région du cortex, violet : région du ganglion de la base, rouge : région de l'hippocampe-amygdale, gris : substance noire et blanc : pas de tissu-spécificité). La bordure du nœud est colorée selon les gènes référencés par ClinVar comme étant associée à une maladie du cerveau: orange : gène associé à ASD et rouge : gène associé à SCZ et à ASD.

4.Discussion et Conclusion

Des travaux ont montré que si nous considérons l'ensemble des gènes impliqués dans les maladies génétiques monogéniques ou polygéniques, il semblerait que les paralogues soient enrichis en mutations par rapport aux singletons. De plus, les anciens paralogues (WGD et oSSD) ont tendance à être plus associés aux maladies que les autres gènes paralogues (Makino & McLysaght 2010; Chen et al. 2014; Singh et al. 2014; Acharya & Ghosh 2016). Enfin, les gènes impliqués dans une maladie sont enrichis en oSSDs et appauvris en ySSDs par rapport aux gènes non pathogènes (Chen et al. 2014). Nous avons affiné ces résultats en nous concentrant exclusivement sur les gènes associés aux maladies cérébrales et nous avons montré que ces gènes sont enrichis en paralogues. D'autre part, contrairement aux tendances obtenues sur toutes les maladies, nous avons

constaté que les gènes impliqués dans les maladies cérébrales n'étaient pas enrichis en WGDs ou SSDs. Cependant, nous avons constaté que les gènes ySSD tendent préférentiellement à accumuler des mutations associées aux maladies cérébrales par rapport aux autres gènes.

Nous avons ensuite montré que les informations que nous avions agrégées sur la co-expression des gènes paralogues et leur tissu-spécificité dans les régions cérébrales pouvaient améliorer la compréhension des gènes connus comme associés aux maladies cérébrales (Landrum et al. 2016). Nous avons observé que les modules de co-expression des paralogues associés à la schizophrénie ou à l'autisme étaient principalement tissu-spécifiques au cervelet, à la moelle épinière, à l'hypothalamus, au cortex et au ganglion de la base. Des altérations morphologiques ou au niveau fonctionnel de ces régions cérébrales ont déjà été détectées en neuroimagerie chez des patients atteints de schizophrénie ou d'autisme (Buckley 2005; Ha et al. 2015). Cependant nous n'avons pas pu prouver une association formelle car ces mêmes régions sont également surreprésentées lorsque nous examinons la spécificité d'expression tissulaire pour l'ensemble des gènes paralogues.

Nous avons finalement étudié le gène *ANKK1* dont la forme mutée est associée à la fois à la schizophrénie et à l'autisme. Nous avons constaté que ce paralogue était intégré dans un réseau de co-expression dont les gènes étaient fréquemment exprimés spécifiquement dans l'hypothalamus. De plus, nous avons déterminé que ces gènes étaient enrichis en fonctions biologiques liées à l'organisation du cytosquelette, à l'activité motrice des microtubules, au mouvement du cilium et à l'assemblage du complexe de la dynéine. La fonction du gène *ANKK1* a été découverte récemment par des travaux expérimentaux sur le cerveau en développement et adulte (España-Serrano et al. 2017). Ces résultats publiés montrent le rôle de *ANKK1* au cours du cycle cellulaire dans les précurseurs neuronaux qui se produisent pendant la neurogenèse embryonnaire. Une autre étude identifie *ANNK1* comme une kinase ayant un rôle dans le remodelage du cytosquelette dans le processus de reprogrammation des cellules somatiques (Sakurai et al. 2014). Ces résultats déjà publiés résonnent avec nos conclusions sur l'association du gène *ANKK1* avec l'organisation du cytosquelette. Cependant il faut noter que notre approche intégrative sur les réseaux de co-expression des gènes paralogues annotés par leur spécificité d'expression à une région cérébrale, nous a permis d'élargir notre connaissance de *ANKK1* en révélant sa spécificité à

l'hypothalamus et en identifiant, par la co-expression, les gènes de son réseau d'influence.

Ces résultats suggèrent que les gènes appartenant au réseau de co-expression de *ANKK1* pourraient être potentiellement influencés par la mutation pathogène (ClinVar) associée à la schizophrénie et à l'autisme. De plus, nous pouvons également faire l'hypothèse que la mutation de l'un de ces gènes étroitement liés à *ANKK1* aurait un impact sur le processus cytosquelettique et pourrait conduire à un phénotype affectant l'hypothalamus. Les gènes tissu-spécifiques à l'hypothalamus appartenant au même module de co-expression que le gène *ANKK1* pourraient également être des gènes à prioriser pour définir de potentiels candidats pour la schizophrénie et l'autisme.

Conclusion

Les profils de transcription obtenus par le consortium GTEx, sur la plus grande collection à ce jour d'échantillons de différentes régions du cerveau, nous ont permis d'explorer l'expression spatiale des gènes paralogues dans le cerveau humain.
Nous avons constaté que les paralogues, en particulier ceux provenant de duplications récentes (ySSD) contribuent à la spécificité d'expression tissulaire des régions du cerveau. De plus, cette implication s'expliquerait d'une part par la jeunesse de la duplication mais également par le type de duplication SSD.
Par la suite, nous avons mis en évidence que ces paires de paralogues issues d'une ySSD avaient tendance à être plus co-exprimés et à préférentiellement accumuler des mutations associées aux maladies cérébrales, par rapport à d'autres catégories de paralogues.
Finalement, nous avons montré que la co-expression et la spécificité tissulaire des paralogues à travers les régions du cerveau pouvaient être utilisées pour améliorer notre connaissance des gènes associés aux maladies cérébrales.
Ce travail met l'accent sur les progrès pouvant être accomplis sur la compréhension de l'expression tissu-spécifique dans le cerveau humain et des maladies cérébrales, par la prise en compte des caractéristiques évolutives des paralogues dans les analyses transcriptomiques.

En conclusion générale, nous avons montré l'implication des gènes paralogues dans le développement du cerveau humain et donc dans la complexité de cet organe. De plus nous pensons que ces résultats permettent de mettre en évidence l'importance de s'appuyer sur l'histoire de nos gènes pour analyser et interpréter les données de génomique humaine.

En perspective de ce projet de thèse, nous pensons que nos résultats sur l'expression tissulaire des gènes paralogues pourraient permettre d'améliorer notre compréhension de leur régulation.
En effet, nous pourrions utiliser notre cartographie de la co-expression et de la tissu-spécificité des gènes paralogues dans les différentes régions du cerveau comme connaissance *a priori* pour explorer la régulation transcriptionnelle de ces gènes. En

effet, les régulations épigénétiques de la chromatine, comme les marques d'histones, ainsi que la combinatoire liée aux interactions des facteurs de transcription avec les régions régulatrices sont des facteurs contribuant à la tissu-spécificité des gènes (Ong & Corces 2011). Par conséquent, il serait probablement intéressant d'extraire des caractéristiques de régulation d'expression partagées (NIH Roadmap Epigenomics Mapping Consortium) au sein de nos réseaux de co-expression de gènes tissu-spécifiques dans le cerveau. Cela permettrait d'améliorer notre compréhension des marques épigénétiques et des organisations de régions régulatrices en les reliant aux phénotypes d'expression des gènes dans le cerveau.

Nous pourrions également explorer les réseaux de co-expression et la tissu-spécificité des gènes paralogues, en prenant en compte la complexité d'expression de leurs isoformes. Il est en effet maintenant établi que les isoformes transcriptionnelles sont exprimées de manière spécifique entre les différents organes (Saha et al. 2016). Par conséquent, leur exploration permettrait probablement d'affiner encore davantage notre connaissance sur la dynamique d'expression des gènes paralogues dans les régions du cerveau.

Comme dans notre étude, des gènes paralogues co-exprimés et tissu-spécifiques dans le même tissu qu'un gène associé à une pathologie cérébrale pourraient être considérés comme de potentiels candidats pour une maladie en particulier. Une analyse plus en profondeur sur ces gènes (GO, voies de signalisations, analyses d'expression différentielle) pourrait permettre de les prioriser notamment dans les études d'association pangénomiques lorsque l'on recherche des variations génétiques associées à la pathologie.

Finalement, l'intégration de nos réseaux de co-expression et de spécificité tissulaire des gènes paralogues à l'imagerie du cerveau malade ou sain permettrait de potentiellement relier des phénotypes d'imagerie (tels que la taille ou la forme de certaines régions cérébrales, la densité de matière grise au sein de chaque voxel de l'image, le degré d'activation de certaines régions dans une condition donnée, la connectivité entre différentes régions en imagerie fonctionnelle ou anatomique) avec des caractéristiques d'expression de ces gènes dans les différentes régions cérébrales. Cela pourrait permettre également de prioriser un certain nombre de gènes paralogues lors la

recherche des variants génétiques associés à un phénotype d'imagerie spécifique à une région cérébrale donnée.

Références

Acharya D, Ghosh TC. 2016. Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics [Internet]. 17:71. Available from: http://www.biomedcentral.com/1471-2164/17/71

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature [Internet]. 496:311–316. Available from: http://www.nature.com/doifinder/10.1038/nature12027

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol [Internet]. 11:1–12. Available from: http://genomebiology.com/2010/11/10/R106

Ardlie KG, Deluca DS, Segre a. V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge C a., Maller JB, Tukiainen T, Lek M, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-) [Internet]. 348:648–660. Available from: http://www.sciencemag.org/content/348/6235/648.full

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B [Internet]. 57:289–300. Available from: http://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini and Y FDR.pdf\nhttp://engr.case.edu/ray_soumya/mlrg/controlling_fdr_benjamini95.pdf

Bouché N, Lacombe B, Fromm H. 2003. GABA signaling: A conserved and ubiquitous mechanism. Trends Cell Biol. 13:607–610.

Buckley PF. 2005. Neuroimaging of schizophrenia: structural abnormalities and pathophysiological implications. Neuropsychiatr Dis Treat [Internet]. 1:193–204. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18568069\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2416751

Bunney WE, Bunney BG. 2000. Evidence for a compromised dorsolateral prefrontal cortical parallel circuit in schizophrenia. Brain Res Rev. 31:138–146.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, Marchena J De, Jin W, Vanderhaeghen P, Ghosh A. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell [Internet]. 149:923–935. Available from: http://www.cell.com/abstract/S0092-8674(12)00462-X

Chen WH, Zhao XM, van Noort V, Bork P. 2013. Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs. PLoS Comput Biol [Internet]. 9. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003073

Chen W-H, Zhao X-M, Noort V van, Bork P. 2014. Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication. PLoS Comput Biol [Internet]. 10:1–2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117415/

Cleveland WS. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. J Am Stat Assoc [Internet]. 74:829. Available from: http://www.jstor.org/stable/2286407?origin=crossref

Declercq W, Vanden Berghe T, Vandenabeele P. 2009. RIP Kinases at the Crossroads of Cell Death and Survival. Cell. 138:229–232.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Tina A, Nefedov M, Rosenfeld J a, Sajjadian S, Malig M, Curry CJ, et al. 2012. Human-specific evolution of novel SRGAP2 genes by incomplete segmental duplication. cell - unedited Manuscr [Internet]. 149:912–922. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3365555/pdf/nihms378423.pdf

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast Computation and Applications of Genome Mappability. PLoS One [Internet]. 7:e30377. Available from: http://dx.plos.org/10.1371/journal.pone.0030377

Dickerson JE, Robertson DL. 2012. On the origins of mendelian disease genes in man: The impact of gene duplication. Mol Biol Evol. 29:61–69.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Liu JS, Ren B. 2012. Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. Nature. 485:376–380.

Dunn OJ. 1959. Confidence Intervals for the Means of Dependent, Normally Distributed Variables. J Am Stat Assoc [Internet]. 54:613–621. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1959.10501524

Ecker C, Bookheimer SY, Murphy DGM. 2015. Neuroimaging in autism spectrum disorder: brain structure and function across the lifespan. Lancet Neurol [Internet]. 14:1121–1134. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1474442215000502

Eckert D, Buhl S, Weber S, Jäger R, Schorle H. 2005. The AP-2 family of transcription factors. Genome Biol [Internet]. 6:246. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1414101&tool=pmcentrez&rendertype=abstract

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics [Internet]. 5:113. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=517706&tool=pmcentrez&rendertype=abstract

Ellis SE, Panitch R, West a B, Arking DE. 2016. Transcriptome analysis of cortical tissue reveals shared sets of downregulated genes in autism and schizophrenia. Transl Psychiatry [Internet]. 6:e817. Available from: http://www.nature.com/doifinder/10.1038/tp.2016.87

España-Serrano L, Guerra Martín-Palanco N, Montero-Pedrazuela A, Pérez-Santamarina E, Vidal R, García-Consuegra I, Valdizán EM, Pazos A, Palomo T, Jiménez-Arriero MÁ, et al. 2017. The Addiction-Related Protein ANKK1 is Differentially Expressed During the Cell Cycle in Neural Precursors. Cereb Cortex. 27:2809–2819.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. Nucleic Acids Res. 40:84–90.

Force A, Lynch M, Pickett FB, Amores A, Y.-L. Y, Postlethwait J. 1999. Preservation of duplicate genes by subfunctionalization. Genetics [Internet]. 151:1531–1545. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=10101175&retmode=ref&cmd=prlinks

Freilich S, Massingham T, Blanc E, Goldovsky L, Thornton JM. 2006. Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. Genome Biol [Internet]. 7:R89. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1794571&tool=pmcentrez&rendertype=abstract

Fukuda T, Yanagi S. 2017. Psychiatric behaviors associated with cytoskeletal defects in radial neuronal migration. Cell Mol Life Sci [Internet]. 74:1–20. Available from: "http://dx.doi.org/10.1007/s00018-017-2539-4

Gray K a., Yates B, Seal RL, Wright MW, Bruford E a. 2015. Genenames.org: The HGNC resources in 2015. Nucleic Acids Res. 43:D1079–D1085.

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol [Internet]. 18:453–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11264396

Ha S, Sohn I-J, Kim N, Sim HJ, Cheon K-A. 2015. Characteristics of Brains in Autism Spectrum Disorder: Structure, Function and Connectivity across the Lifespan. Exp Neurobiol [Internet]. 24:273. Available from: https://synapse.koreamed.org/DOIx.php?id=10.5607/en.2015.24.4.273

Hakak Y, Walker JR, Li C, Wong WH, Davis KL, Buxbaum JD, Haroutunian V, Fienberg a a. 2001. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. Proc Natl Acad Sci U S A [Internet]. 98:4746–51. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/11296301\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC31905

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol [Internet]. 8:R209. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2007-8-10-r209

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: An intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. 33:2374–2383.

Harrow J, Frankish a, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F. 2012. GENCODE: The Reference Human Genome Annotation for The ENCODE Project. Genome Res [Internet]. 22:1760–1774. Available from: https://doi.org/10.1101/gr.135350.111

Hilbe JM. 2011. Negative binomial regression. [place unknown]: Cambridge University Press.

Hurles M. 2004. Gene duplication: The genomic trade in spare parts. PLoS Biol. 2.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet [Internet]. 11:4. Available from: http://www.nature.com/doifinder/10.1038/nrg2689

Konopka G, Friedrich T, Davis-Turak J, Winden K, Oldham MC, Gao F, Chen L, Wang GZ, Luo R, Preuss TM, Geschwind DH. 2012. Human-Specific Transcriptional Networks in the Brain. Neuron. 75:601–617.

Koonin E V. 2005. Orthologs, Paralogs, and Evolutionary Genomics. Annu Rev Genet [Internet]. 39:309–338. Available from: http://www.annualreviews.org/doi/10.1146/annurev.genet.39.073003.114725

Kryuchkova N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. bioRxiv [Internet]. 18:027755. Available from: http://www.biorxiv.org/content/early/2015/09/28/027755.abstract

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2016. Tissue-specificity of gene expression diverges slowly between orthologs, and rapidly between paralogs. PLoS Comput Biol [Internet].:065086. Available from: http://biorxiv.org/lookup/doi/10.1101/065086

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. Science [Internet]. 352:1009–13. Available from: http://biorxiv.org/content/early/2016/02/02/019166.abstract\nhttp://www.ncbi.nlm.nih.gov/pubmed/27199432

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. 2016. ClinVar: Public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44:D862–D868.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics [Internet]. 9:559. Available from: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559

Letunic I, Bork P. 2011. Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39:475–478.

Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, et al. 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res [Internet]. 34:D572–D580. Available from: https://www.ncbi.nlm.nih.gov/pubmed/16381935

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. 15:550. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. Trends Genet. 25:152–155.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci U S A [Internet]. 107:9270–9274. Available from: http://www.gen.tcd.ie/molevol/pdfs/PNAS-2010-Makino-9270-4.pdf

Mardis ER. 2008. Next-Generation DNA Sequencing Methods. Annu Rev Genomics Hum Genet [Internet]. 9:387–402. Available from: http://www.annualreviews.org/doi/10.1146/annurev.genom.9.081307.164359

McGinnis S, Madden TL. 2004. BLAST: At the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 32:20–25.

Meylan E, Tschopp J. 2005. The RIP kinases: Crucial integrators of cellular stress. Trends Biochem Sci. 30:151–159.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 45:D183–D189.

Needleman, Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol [Internet]. 48:443–453. Available from: http://www.ncbi.nlm.nih.gov/pubmed/5420325

Nymberg C, Banaschewski T, Bokde AL, Büchel C, Conrod P, Flor H, Frouin V, Garavan H, Gowland P, Heinz A, et al. 2014. DRD2/ANKK1 Polymorphism Modulates the Effect of Ventral Striatal Activation on Working Memory Performance. Neuropsychopharmacol

[Internet]. 39:2357–2365. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24713612

Ohno S. 1970. Evolution by Gene Duplication. [place unknown].

Oldham MC, Konopka G, Iwamoto K, Langfelder P, Horvath S, Geschwind DH. 2008. Functional organization of the transcriptome in human brain. Nat Neurosci. 11:1271–1282.

Ong C-T, Corces V. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 12:283–293.

Papp B, Pal C, Hust LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature. 424:194–197.

Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, Duret L, Gouy M, Perrière G. 2009. Databases of homologous gene families for comparative genomics. BMC Bioinformatics [Internet]. 10 Suppl 6:S3. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2697650&tool=pmcentrez&rendertype=abstract

Pierson E, Koller D, Battle A, Mostafavi S. 2015. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLoS Comput Biol. 11:1–19.

Powers MW. 2011. EVALUATION: PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION. 2:37–63.

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet [Internet]. 3:827–837. Available from: http://www.nature.com/doifinder/10.1038/nrg928

Rand WM. 1971. Objective Criteria for the Evaluation of Clustering Methods. J Am Stat Assoc [Internet]. 66:846–850. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356

Ravasz E. 2002. Hierarchical Organization of Modularity in Metabolic Networks. Science (80-) [Internet]. 297:1551–1555. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1073374

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet [Internet]. 16:276–277. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168952500020242

Ruan J, Li H, Chen Z, Coghlan A, Coin LJM, Guo Y, Heacute;riché JK, Hu Y, Kristiansen K, Li R, et al. 2008. TreeFam: 2008 Update. Nucleic Acids Res [Internet]. 36:735–740. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238856/

Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, Consortium Gte, Engelhardt BE, Battle A. 2016. Co-expression networks reveal the tissue-specific regulation of

transcription and splicing. bioRxiv [Internet].:078741. Available from: http://biorxiv.org/content/early/2016/10/02/078741

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–25.

Sakurai K, Talukdar I, Patil VS, Dang J, Li Z, Chang K-Y, Lu C-C, Delorme-Walker V, DerMardirossian C, Anderson K, et al. 2014. Kinome-wide Functional Analysis Highlights the Role of Cytoskeletal Remodeling in Somatic Cell Reprogramming. Cell Stem Cell [Internet]. 14:523–534. Available from: http://www.sciencedirect.com/science/article/pii/S1934590914000952

Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, Li Y, Lin S, Lin Y, Barr CL, Ren B. 2016. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep [Internet]. 17:2042–2059. Available from: http://dx.doi.org/10.1016/j.celrep.2016.10.061

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T. 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol [Internet]. 7:3. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1413964&tool=pmcentrez&rendertype=abstract

Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. 2012. On the Expansion of " Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates. Cell Rep [Internet]. 2:1387–1398. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23168259

Singh PP, Affeldt S, Malaguti G, Isambert H. 2014. Human Dominant Disease Genes Are Enriched in Paralogs Originating from Whole Genome Duplication. PLoS Comput Biol [Internet]. 10. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003754

Smith, Waterman. 1981. Identification of Common Molecular Subsequences. JMolBiol.

Stigler SM. 2008. Karl Pearson's Theoretical Errors and the Advances They Inspired. Stat Sci [Internet]. 23:261–271. Available from: http://projecteuclid.org/euclid.ss/1219339117

Su G, Morris JH, Demchak B, Bader GD. 2015. Biological Network Exploration with Cytoscape 3. [place unknown].

Székely GJ, Rizzo ML. 2005. Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. J Classifcation. 22:151–183.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res [Internet]. 19:327–35. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/19029536%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2652215

Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor R, Blencowe BJ, Geschwind DH. 2013. Transcriptomic Analysis of Autistic Brain Reveals Convergent Molecular Pathology. Nature. 474:380–384.

Wagner A. 2002. Selection and gene duplication: a view from the genome. Genome Biol [Internet]. 3:reviews1012. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC139360/

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet [Internet]. 10:57–63. Available from: http://www.nature.com/doifinder/10.1038/nrg2484

Welch BL. 1947. The generalization of `Student's' problem when several different population variances are involved. Biometrika. 34:28–35.

Xie T, Yang QY, Wang XT, McLysaght A, Zhang HY. 2016. Spatial Colocalization of Human Ohnolog Pairs Acts to Maintain Dosage-Balance. Mol Biol Evol. 33:2368–2375.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 21:650–659.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. Nucleic Acids Res. 44:D710–D716.

Zhang J. 2003. Evolution by gene duplication: An update. Trends Ecol Evol. 18:292–298.

Zmasek CM, Eddy SR. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics. 17:821–828.

Annexes

Annexe A : Gènes différentiellement exprimés parmi les gènes paralogues et les singletons.

Annexe B : Enrichissements GO et voies de signalisation des familles homogènes.

Annexe C : Enrichissement GO pour les gènes du module de co-expression de *ANKK1.*

Annexe A

Tissue_ref	Tissue_test	nbDEG_dup	nbDEG_singl	chi2_Bonferroni	OddRatio
Brain-Cortex	Brain-Cerebellum	5948	2833	9.54503773116808e-41	1,55970014
Brain-Cortex	Brain-Hippocampus	2312	884	1.06270716563776e-32	1,6977344
Brain-Cortex	Brain-Substantianigra	4303	1684	2.3700413519204e-70	1,86728029
Brain-Cortex	Brain-Anteriorcingulatecortex	304	103	6.48437074400183e-05	1,76216447
Brain-Cortex	Brain-FrontalCortex	110	41	1	1,58771543
Brain-Cortex	Brain-CerebellarHemisphere	6357	2963	3.9543923489948e-56	1,68756824
Brain-Cortex	Brain-Caudate	2898	1069	5.54053296193092e-50	1,83099042
Brain-Cortex	Brain-Nucleusaccumbens	2465	934	1.01605891946148e-36	1,72972337
Brain-Cortex	Brain-Putamen	2925	1036	5.181484182592e-58	1,92643772
Brain-Cortex	Brain-Hypothalamus	2992	1117	5.97165832424986e-50	1,81479667
Brain-Cortex	Brain-Spinalcord	5462	2435	2.89803916401032e-55	1,68337657
Brain-Cortex	Brain-Amygdala	2515	963	3.62480841549984e-36	1,71292224
Brain-Cerebellum	Brain-Hippocampus	6533	3190	1.86634730564681e-40	1,56317381
Brain-Cerebellum	Brain-Substantianigra	6898	3381	7.51261851069974e-45	1,60926663
Brain-Cerebellum	Brain-Anteriorcingulatecortex	6323	3136	5.89818598174439e-32	1,48556149
Brain-Cerebellum	Brain-FrontalCortex	6177	3103	2.91560394767047e-26	1,43099216
Brain-Cerebellum	Brain-CerebellarHemisphere	254	51	1.32169381518268e-11	2,98448061
Brain-Cerebellum	Brain-Caudate	6648	3222	2.6490826077346e-45	1,60610611
Brain-Cerebellum	Brain-Nucleusaccumbens	6445	3191	4.53068623033765e-34	1,50624024
Brain-Cerebellum	Brain-Putamen	6652	3254	1.3708555899054e-41	1,57523504
Brain-Cerebellum	Brain-Hypothalamus	6616	3298	8.00995591219816e-34	1,50711041
Brain-Cerebellum	Brain-Spinalcord	7016	3418	9.29399451246646e-50	1,65375691
Brain-Cerebellum	Brain-Amygdala	6745	3335	9.63490159849565e-39	1,55320376
Brain-Hippocampus	Brain-Substantianigra	1477	452	8.21561150022483e-38	2,08058512
Brain-Hippocampus	Brain-Anteriorcingulatecortex	1376	439	1.86796821185223e-31	1,97775897
Brain-Hippocampus	Brain-FrontalCortex	2147	731	2.81996556995009e-44	1,92301486
Brain-Hippocampus	Brain-CerebellarHemisphere	6717	3165	5.4472179495392e-59	1,71694267
Brain-Hippocampus	Brain-Caudate	1661	457	2.27671625531996e-54	2,36117369
Brain-Hippocampus	Brain-Nucleusaccumbens	1943	581	7.19899513224767e-55	2,196148
Brain-Hippocampus	Brain-Putamen	1515	489	4.68952434849131e-34	1,9681392
Brain-Hippocampus	Brain-Hypothalamus	1339	370	3.41394323758481e-42	2,30185127
Brain-Hippocampus	Brain-Spinalcord	3260	1063	2.75608401696836e-85	2,17991484
Brain-Hippocampus	Brain-Amygdala	246	24	5.82875527498185e-20	6,16483299
Brain-Substantianigra	Brain-Anteriorcingulatecortex	3533	1124	1.74765375599857e-101	2,2957378
Brain-Substantianigra	Brain-FrontalCortex	4213	1517	3.09588975665629e-92	2,07540879
Brain-Substantianigra	Brain-CerebellarHemisphere	7082	3394	4.4701459687322e-59	1,73062099
Brain-Substantianigra	Brain-Caudate	2629	934	5.32994275468722e-50	1,8840657
Brain-Substantianigra	Brain-Nucleusaccumbens	3369	1145	1.38316240768476e-79	2,08955576
Brain-Substantianigra	Brain-Putamen	2064	657	9.13128643677721e-51	2,06436198
Brain-Substantianigra	Brain-Hypothalamus	1664	565	6.01940153741852e-33	1,87726314
Brain-Substantianigra	Brain-Spinalcord	1306	313	1.33915516620049e-52	2,67061867
Brain-Substantianigra	Brain-Amygdala	1525	441	1.51158027007941e-44	2,21809761

Brain-Anteriorcingulatecortex	Brain-FrontalCortex	73	19	0.121628782942196	2,27373858
Brain-Anteriorcingulatecortex	Brain-CerebellarHemisphere	6462	3099	2.11626512314732e-46	1,61140459
Brain-Anteriorcingulatecortex	Brain-Caudate	2074	710	8.97782597024796e-42	1,90309941
Brain-Anteriorcingulatecortex	Brain-Nucleusaccumbens	1686	614	1.01851897283265e-26	1,73918319
Brain-Anteriorcingulatecortex	Brain-Putamen	2154	633	1.65633023552673e-64	2,27064212
Brain-Anteriorcingulatecortex	Brain-Hypothalamus	2116	726	1.62222843197108e-42	1,90287706
Brain-Anteriorcingulatecortex	Brain-Spinalcord	5022	2012	2.26152416839093e-82	1,91676592
Brain-Anteriorcingulatecortex	Brain-Amygdala	1096	302	5.96329164107159e-34	2,27434952
Brain-FrontalCortex	Brain-CerebellarHemisphere	6347	3008	4.08939672136267e-49	1,63173597
Brain-FrontalCortex	Brain-Caudate	2802	1016	9.52944688005086e-51	1,85835238
Brain-FrontalCortex	Brain-Nucleusaccumbens	2306	847	7.41965099889295e-38	1,77852481
Brain-FrontalCortex	Brain-Putamen	2848	902	8.14834320290414e-77	2,18873394
Brain-FrontalCortex	Brain-Hypothalamus	2696	897	7.40467261926595e-63	2,04397927
Brain-FrontalCortex	Brain-Spinalcord	5415	2280	6.82993804656485e-75	1,84014228
Brain-FrontalCortex	Brain-Amygdala	2333	752	4.57134268402438e-57	2,07033016
Brain-CerebellarHemisphere	Brain-Caudate	6825	3222	1.58446111948045e-60	1,732016
Brain-CerebellarHemisphere	Brain-Nucleusaccumbens	6667	3153	3.06600254907987e-56	1,69424707
Brain-CerebellarHemisphere	Brain-Putamen	6875	3263	4.1612389902015e-59	1,7227112
Brain-CerebellarHemisphere	Brain-Hypothalamus	6743	3238	5.60654785564524e-51	1,65460357
Brain-CerebellarHemisphere	Brain-Spinalcord	7157	3363	2.85112972465352e-71	1,82748488
Brain-CerebellarHemisphere	Brain-Amygdala	6861	3284	8.08818726325412e-55	1,68869633
Brain-Caudate	Brain-Nucleusaccumbens	266	77	1.54682193194701e-06	2,06366977
Brain-Caudate	Brain-Putamen	139	89	1	0,91952451
Brain-Caudate	Brain-Hypothalamus	2181	678	1.47636597302561e-57	2,13586361
Brain-Caudate	Brain-Spinalcord	4326	1655	1.34782173430392e-77	1,93008192
Brain-Caudate	Brain-Amygdala	1212	377	4.56077275626983e-29	2,01390894
Brain-Nucleusaccumbens	Brain-Putamen	784	274	6.99770409444492e-13	1,74297184
Brain-Nucleusaccumbens	Brain-Hypothalamus	2066	631	3.1120423795182e-56	2,16232088
Brain-Nucleusaccumbens	Brain-Spinalcord	4748	1848	2.24998873228626e-84	1,95166578
Brain-Nucleusaccumbens	Brain-Amygdala	1431	461	4.58794076104521e-32	1,96308491
Brain-Putamen	Brain-Hypothalamus	2329	791	2.72344192783973e-49	1,9495538
Brain-Putamen	Brain-Spinalcord	3877	1446	3.95222360620171e-72	1,92889538
Brain-Putamen	Brain-Amygdala	1250	417	5.53432047024063e-25	1,8724701
Brain-Hypothalamus	Brain-Spinalcord	3994	1470	5.17113867690191e-79	1,98044558
Brain-Hypothalamus	Brain-Amygdala	1366	379	8.32129089255313e-43	2,29578751
Brain-Spinalcord	Brain-Amygdala	3558	1203	3.66138434929903e-87	2,13364847

Annexe B

Reference	Annotation data set	Bonferroni count	Terms	P-value
Homo sapiens (all genes in database) (20972 genes)	Reactome pathway (Reactome version 58 Released 2016-12-07)	1775	Activation of the TFAP2 (AP-2) family of transcription factors	2.33E-05
			Negative regulation of activity of TFAP2 (AP-2) family transcription factors	1.60E-02
			NOTCH2 intracellular domain regulates transcription	3.24E-02
			Notch-HLH transcription pathway	4.43E-02
			Pre-NOTCH Transcription and Translation	1.96E-04
			Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors	9.59E-03
			Pre-NOTCH Expression and Processing	3.58E-03
			Activation of anterior HOX genes in hindbrain development during early embryogenesis	5.50E-04
			Activation of HOX genes during differentiation)	5.50E-04
	GO Cellular Component (GO Ontology database Released 2017-05-25)	1316	nucleoplasm	2.72E-04
			nuclear lumen	1.89E-02
			nucleus	1.30E-03
			membrane	1.22E-02
	GO Molecular Function (GO Ontology database Released 2017-05-25)	3080	lactonohydrolase activity	2.60E-02
			acyl-L-homoserine-lactone lactonohydrolase activity	2.60E-02
			2'-5'-oligoadenylate synthetase activity	2.60E-02
			adenylyltransferase activity	3.83E-02
			transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	4.01E-03
			transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding	6.49E-04
			transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding	1.02E-03
			RNA polymerase II regulatory region sequence-specific DNA binding	3.61E-05
			RNA polymerase II regulatory region DNA binding	4.02E-05
			transcription regulatory region sequence-specific DNA binding	3.57E-05
			sequence-specific double-stranded DNA binding	1.16E-04
			sequence-specific DNA binding	1.77E-07
			double-stranded DNA binding	9.43E-05
			transcription regulatory region DNA binding	9.80E-04
			regulatory region DNA binding	1.03E-03
			regulatory region nucleic acid binding	1.06E-03
			RNA polymerase II transcription factor activity, sequence-specific DNA binding	8.95E-03
			transcription factor activity, sequence-specific DNA binding	2.94E-05
			nucleic acid binding transcription factor activity	3.01E-05
			DNA binding	3.82E-05
			nucleic acid binding	8.85E-06
			heterocyclic compound binding)	5.58E-03

			organic cyclic compound binding	6.22E-03
	GO biological process (GO Ontology database Released 2017-05-25)	8492	embryonic skeletal system morphogenesis	2.99E-05
			embryonic skeletal system development	7.21E-06
			anterior/posterior pattern specification	7.67E-06
			skeletal system morphogenesis	5.52E-03
			embryonic organ morphogenesis	2.52E-04
			regionalization	5.17E-03
			skeletal system development	1.23E-03
			pattern specification process	3.94E-02
			transcription from RNA polymerase II promoter	7.50E-04
			positive regulation of transcription from RNA polymerase II promoter	5.31E-06
			negative regulation of transcription from RNA polymerase II promoter	1.59E-02
			embryo development	6.11E-03
			positive regulation of RNA metabolic process	1.04E-05
			positive regulation of nucleic acid-templated transcription	3.10E-05
			positive regulation of transcription, DNA-templated	3.10E-05
			positive regulation of RNA biosynthetic process	3.17E-05
			positive regulation of gene expression	7.32E-06
			positive regulation of macromolecule biosynthetic process	7.67E-05
			positive regulation of nucleobase-containing compound metabolic process	8.13E-05
			negative regulation of transcription, DNA-templated	4.08E-02
			negative regulation of nucleic acid-templated transcription	3.42E-02
			negative regulation of RNA biosynthetic process	3.55E-02
			positive regulation of cellular biosynthetic process	3.35E-04
			regulation of transcription from RNA polymerase II promoter	9.35E-05
			negative regulation of macromolecule biosynthetic process)	1.73E-02
			negative regulation of cellular macromolecule biosynthetic	3.29E-02
			positive regulation of biosynthetic process	5.65E-04
			negative regulation of biosynthetic process	1.43E-02
			negative regulation of cellular biosynthetic process	2.52E-02
			transcription, DNA-templated	2.09E-06
			nucleic acid-templated transcription	2.12E-06
			RNA biosynthetic process	2.51E-06
			nucleobase-containing compound biosynthetic process	3.61E-05
			heterocycle biosynthetic process	4.15E-05
			aromatic compound biosynthetic process	4.33E-05
			RNA metabolic process	1.53E-05
			organic cyclic compound biosynthetic process	2.77E-04
			positive regulation of macromolecule metabolic process	2.61E-03
			positive regulation of nitrogen compound metabolic process	4.65E-03
			regulation of RNA metabolic process	1.16E-04
			positive regulation of cellular metabolic process	1.62E-02
			nucleic acid metabolic process	9.72E-05
			regulation of transcription, DNA-templated	9.24E-04
			regulation of nucleic acid-templated transcription	1.23E-03

			regulation of cellular macromolecule biosynthetic process	1.62E-04
			regulation of RNA biosynthetic process	1.35E-03
			regulation of macromolecule biosynthetic process	1.20E-04
			gene expression	1.32E-03
			cellular nitrogen compound biosynthetic process	5.24E-03
			cellular aromatic compound metabolic process	1.50E-05
			regulation of nucleobase-containing compound metabolic process	4.02E-04
			positive regulation of metabolic process	4.23E-02
			heterocycle metabolic process	4.69E-05
			regulation of cellular biosynthetic process	2.45E-04
			organic cyclic compound metabolic process	1.92E-05
			regulation of biosynthetic process	2.55E-04
			cellular macromolecule biosynthetic process)	1.18E-02
			regulation of gene expression	4.22E-04
			nucleobase-containing compound metabolic process	6.19E-04
			macromolecule biosynthetic process	2.31E-02
			single-multicellular organism process	1.65E-02
			regulation of macromolecule metabolic process	1.27E-02
			regulation of cellular metabolic process	1.50E-02
			regulation of primary metabolic process	1.76E-02
			regulation of nitrogen compound metabolic process	4.95E-02
			regulation of metabolic process	2.61E-02

Annexe C

Analysis Type:	**PANTHER Overrepresentation Test (release 20170413)**		
Annotation Version and Release Date:	**GO Ontology database Released 2017-08-14**		
GO molecular function complete	**Homo sapiens (21002)**	**Module *ANKK1* (75)**	**P-value**
dynein light chain binding	25	4	7,11E-03
ATP-dependent microtubule motor activity	46	5	2,39E-03
microtubule motor activity	86	5	4,88E-02
GO biological process complete	**Homo sapiens (21002)**	**Module *ANKK1* (75)**	**P-value**
outer dynein arm assembly	17	4	4,29E-03
axonemal dynein complex assembly	31	7	2,39E-07
cilium movement	60	11	3,69E-12
cilium-dependent cell motility	23	4	1,41E-02
cilium or flagellum-dependent cell motility	24	4	1,67E-02
axoneme assembly	58	9	8,56E-09
microtubule bundle formation	85	9	2,47E-07
flagellated sperm motility	67	5	4,07E-02
sperm motility	69	5	4,69E-02
microtubule-based movement	248	13	4,78E-08
cilium assembly	323	14	8,28E-08
cilium organization	334	14	1,28E-07
plasma membrane bounded cell projection assembly	408	14	1,73E-06
cell projection assembly	413	14	2,02E-06
microtubule cytoskeleton organization	383	11	1,07E-03
microtubule-based process	574	16	1,54E-06
organelle assembly	674	17	1,75E-06
cell projection organization	1055	17	1,27E-03
plasma membrane bounded cell projection organization	1025	15	2,55E-02

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : électrons, photons, sciences du vivant (EOBE)

Titre : Expression tissulaire des gènes paralogues : application au cerveau humain et à son état pathologique

Mots clés : gènes paralogues, spécificité tissulaire, RNA-seq, expression cerveau

Résumé : Dans l'histoire évolutive, deux gènes paralogues sont issus d'un évènement de duplication de leur ancêtre commun. Les gènes paralogues sont caractérisés par des duplications globales de génome (WGD) ou à petite échelle (SSD) et par leur datation. Les WGDs ont lieu à deux reprises à la base de la lignée des vertébrés. Les évènements de SSD ont lieu à plusieurs moments pouvant être plus récents, plus anciens ou contemporain de la période des évènements de WGD. La rétention des paralogues dans le génome, associée à une divergence de l'expression spatiale est une contribution importante pour l'augmentation de la complexité de l'organisme au cours de l'évolution. Certaines études ont montré que les duplications anciennes seraient plus associées aux maladies. L'objectif de la première partie de la thèse est de créer une ressource sur les paralogues en collectant et en analysant différentes annotations. Nous avons construit une ressource robuste de paralogues humains à partir de listes publiées mais aussi à partir d'annotations externes. L'exploration de différentes annotations nous a permis d'identifier une identité de séquence élevée entre gènes paralogues pouvant biaiser la mesure d'expression des gènes et diminuer leur expression. L'objectif de la seconde partie, est d'explorer l'expression spatiale et la co- expression des paralogues au sein du cerveau humain, à partir des données RNA-seq du consortium GTEx. Les données d'expression GTEx de 13 tissus cérébraux, nous ont permis de montrer que la datation récente mais aussi que le type SSD contribuaient à une expression plus tissu-spécifique. Nous avons utilisé l'analyse de la co-expression (WGCNA) afin de regrouper les paralogues possédant une expression similaire au travers des tissus et nous avons pu suggérer une co-expression des SSD récents. Nos études sur les maladies ont montré que les SSD récents accumulaient des mutations associées à des maladies cérébrales. Finalement, nous avons trouvé que la co-expression des paralogues et leur tissu-spécificité au travers des régions cérébrales pouvaient enrichir nos connaissances sur les gènes associés à des maladies cérébrales.

Title : Tissue expression of paralogous genes : application on human brain and its pathological state

Keywords : paralogous genes, tissue-specificity, RNA-seq, brain expression

Abstract: In evolution history, two paralogous genes originate from the duplication event of a common ancestor gene. Paralogous genes are characterized by whole genome (WGD) or small-scale (SSD) duplications and their duplication date. The WGDs happened twice in the early vertebrate lineage. SSD events take place at any moment in evolutionary history and can be younger, older or dating to the same period than WGD events. Retention of paralogs in the genome associated with divergence of spatial expression is an important contributor to the increase of organism complexity through evolution. Different studies found that old duplications are more associated with diseases. The objective of the first part of the thesis is to create a resource on paralogs by collecting and analyzing annotations. We built a robust resource of human paralogs from published lists of paralogous genes and also from external annotations. Annotation exploration allowed us to identify a high sequence identity between paralogous genes impacting the gene expression measurement from RNA-seq data and decreasing the gene expression. The objective of the second part is to explore spatial expression and co-expression of paralogs in the human brain, from the GTEx consortium RNA-seq data. The GTEx expression data of 13 brain tissues allowed us to show that duplication youth and SSD type contributed to a more tissue-specific expression. We used co-expression analyses (WGCNA) to group paralogs with similar expression across tissues and we suggested the co-expression of younger SSDs. Our disease studies showed the younger SSD accumulation of mutations associated with brain diseases. We finally found that paralog co-expression and their tissue-specificity across brain regions could enrich information of known brain disease-associated genes.