

Université de Montréal

Étude sur le rôle des mutations *de novo* dans l'étiologie
génétiq ue de la schizophrénie

Par Simon L. Girard

Programme de Biologie Moléculaire, Faculté de Médecine

Thèse présentée à la Faculté des études supérieures

En vue de l'obtention du grade de Ph. D.

En biologie moléculaire

Août 2013

©Simon Girard 2013

Résumé

La schizophrénie est une maladie psychiatrique grave qui affecte approximativement 1 % de la population. Il est clairement établi que la maladie possède une composante génétique très importante, mais jusqu'à présent, les études ont été limitées au niveau de l'identification de facteurs génétiques spécifiquement liés à la maladie. Avec l'avènement des nouvelles avancées technologiques dans le domaine du séquençage de l'ADN, il est maintenant possible d'effectuer des études sur un type de variation génétique jusqu'à présent laissé pour compte : les mutations *de novo*, c.-à-d. les nouvelles mutations non transmises de manière mendélienne par les parents. Ces mutations peuvent avoir deux origines distinctes : une origine germinale au niveau des gamètes chez les parents ou une origine somatique, donc au niveau embryonnaire directement chez l'individu.

L'objectif général de la présente recherche est de mieux caractériser les mutations *de novo* dans la schizophrénie. Comme le rôle de ces variations est peu connu, il sera également nécessaire de les étudier dans un contexte global au niveau de la population humaine. La première partie du projet consiste en une analyse exhaustive des mutations *de novo* dans la partie codante (exome) de patients atteints de schizophrénie. Nous avons pu constater que non seulement le taux de mutations était plus élevé qu'attendu, mais nous avons également été en mesure de relever un nombre anormalement élevé de mutations non-sens, ce qui suggère un profil pathogénique. Ainsi, nous avons pu fortement suggérer que les mutations *de novo* sont des actrices importantes dans le mécanisme génétique de la schizophrénie.

La deuxième partie du projet porte directement sur les gènes identifiés lors de la première partie. Nous avons séquencé ces gènes dans une plus grande cohorte de cas et de contrôles afin d'établir le profil des variations rares pour ces gènes. Nous avons ainsi conclu que l'ensemble des gènes identifiés par les études de mutations *de novo* possède un profil pathogénique, ce qui permet d'établir que la plupart de ces gènes ont un rôle réel dans la maladie et ne sont pas des artéfacts expérimentaux. De plus, nous avons pu établir une association directe avec quelques gènes qui montrent un profil aberrant de variations rares.

La troisième partie du projet se concentre sur l'effet de l'âge paternel sur le taux de mutations *de novo*. En effet, pour la schizophrénie, il est démontré que l'âge du père est un facteur de risque important. Ainsi, nous avons tenté de caractériser l'effet de l'âge du père chez des patients en santé. Nous avons observé une grande corrélation entre l'âge du père et le taux de mutations germinales et nous avons ainsi pu répertorier certaines zones avec un grand nombre de mutations *de novo*, ce qui suggère l'existence de zone chaude pour les mutations.

Nos résultats ont été parmi les premiers impliquant directement les mutations *de novo* dans le mécanisme génétique de la schizophrénie. Ils permettent de jeter un nouveau regard sur les réseaux biologiques à l'origine de la schizophrénie en mettant sous les projecteurs un type de variations génétiques longtemps laissé pour compte.

Mots-clés : Schizophrénie, génétique, mutation *de novo*, variant rare, âge du père.

Abstract

Schizophrenia is a severe psychiatric disorder that affects roughly 1% of the general population. It has been clearly demonstrated that the disease possesses a strong genetic component, but thus far, studies have had limited success in identifying key schizophrenia genes. With the advent of new DNA sequencing technologies it is now possible to study a type of genetic variation that has been previously looked over: *de novo* mutations (new mutations not transmitted by parents)

The main aim of the present thesis is to better characterize *de novo* mutations in schizophrenia. As the role of these variations is not very well known, it was also necessary to study them in a global context in the human population. The first part of our project was to do a comprehensive study of *de novo* mutations found in the coding section (exome) of patients affected with schizophrenia. We found that the mutation rate was higher than expected. We also observed an aberrant number of nonsense mutations, which suggests a pathogenic profile of mutations. Thus, we strongly suggested that *de novo* mutations are key players in the genetic mechanism of schizophrenia.

The second part of the work builds on the genes bearing mutations identified in the exome sequencing analysis. We sequenced these genes in a larger cohort of cases and controls in order to establish the profile of rare variants for these genes. We were able to conclude that the global mutational profile of the genes identified during *de novo* mutation studies are indeed pathogenic, which confirms that some of

those genes are really involved in the disease and are not sequencing artefacts. Additionally, we were also able to identify some genes that had an aberrant rare variation profile.

The third part of the work aimed to characterize the paternal age effect on the *de novo* mutation rate. Indeed, in schizophrenia, it has been shown numerous times that paternal age is a risk factor for the disease. Thus, we have chosen to characterize this effect in a cohort of healthy subjects. We were able to observe a high correlation between paternal age and an elevated germline mutation rate. We were also able to confirm the existence of genomic regions that present an elevated number of *de novo* mutations, supporting the notion of mutational hotspots.

Our results were amongst the first to be published on the scientific area to directly involve *de novo* mutations in the genetic mechanism of schizophrenia. Those results bring new clues on the biological networks underlying schizophrenia by investigating a genetic variation type long overlooked.

Keywords: Schizophrenia, genetic, *de novo* mutation, rare variant, paternal age.

Table des matières

RÉSUMÉ	II
ABSTRACT	IV
TABLE DES MATIÈRES	VI
LISTE DES TABLEAUX	XI
LISTE DES FIGURES	XIV
LISTE DES SIGLES ET ABRÉVIATIONS	XVI
REMERCIEMENTS	XIX
CHAPITRE 1 : INTRODUCTION	1
1.1 SCHIZOPHRÉNIE : PRÉSENTATION CLINIQUE	2
1.1.1 <i>Symptômes</i>	2
1.1.2 <i>Diagnostic</i>	3
1.1.3 <i>Traitements</i>	5
1.1.4 <i>Épidémiologie</i>	6
1.1.5 <i>Causes</i>	6
1.1.5.1 Facteurs prénataux.....	6
1.1.5.2 Facteurs infectieux.....	7
1.1.5.3 Consommation de drogues	8
1.2 L'ÉTILOGIE GÉNÉTIQUE DE LA SCHIZOPHRÉNIE	9
1.2.1 <i>Introduction</i>	9
1.2.2 <i>L'origine de l'hypothèse génétique</i>	9
1.2.3 <i>Les études de liaisons</i>	10
1.2.3.1 Définition.....	10
1.2.3.2 Les études de liaison et la schizophrénie.....	11
1.2.3.3 Le cas de DISC1	13
1.2.3.4 Perspectives : L'absence de conclusions; une conclusion?.....	13
1.2.4 <i>Les études d'association génomique</i>	15
1.2.4.1 Définition.....	15
1.2.4.2 La première vague d'étude d'association.....	16
1.2.4.3 La deuxième vague : Premières percées	17
1.2.4.4 Troisième vague : La consécration?.....	18
1.2.4.5 Perspectives : Stop ou encore?.....	19
1.2.5 <i>Les variations structurales</i>	21

1.2.5.1 Description	21
1.2.5.2 Les études de prévalence.....	21
1.2.5.3 Les associations spécifiques aux loci de SV	23
1.2.5.4 Perspectives	23
1.2.6 Les études de re-séquençage.....	24
1.2.6.1 Descriptions	24
1.2.6.2 Le séquençage Sanger	25
1.2.6.3 Le séquençage à haut-débit	25
1.2.6.4 Les études de séquençage de gènes candidats.....	27
1.2.6.5 L'implication des mutations de novo	27
1.2.6.6 Perspectives	29
1.3 L'EFFET DE L'ÂGE DU PÈRE SUR LES MUTATIONS	30
1.3.1 Les types de mutations et leurs mécanismes respectifs.....	31
1.3.1.1 Historique.....	31
1.3.1.2 Les types de variants	32
1.3.1.2.1 Les variants à simple nucléotide	32
1.3.1.2.2 Les insertions et délétions	33
1.3.1.2.3 Les variations structurales	33
1.3.2 La spermatogonie	34
1.3.3 L'effet de l'âge paternel.....	36
1.4 PROBLÉMATIQUE	37
CHAPITRE 2 : ÉLÉVATION DU TAUX DE MUTATIONS DE NOVO EXONIQUES DANS DES	
PATIENTS ATTEINTS DE SCHIZOPHRÉNIE	40
2.1 PRÉSENTATION DE L'ARTICLE.....	40
2.1.1 CONTRIBUTION DES AUTEURS.....	41
INCREASED EXONIC DE NOVO MUTATION RATE IN PROBANDS AFFECTED WITH SCHIZOPHRENIA.....	43
2.2 <i>Abstract (Introduction)</i>	44
2.3 <i>Methods (Published online)</i>	44
2.3.1 Recruitment of patients.....	44
2.3.2 Clinical assessment procedures	45
2.3.3 Clinical characteristics	46
2.3.4 Exome Capture and Sequencing.....	46
2.3.5 Read mapping	47
2.3.6 Assessment of enrichment.....	47
2.3.7 Variant calling and annotation.....	48
2.3.8 Variant filtering and segregation analysis	48

2.3.9 Validation of de novo mutations	49
<i>2.4 Results and discussion</i>	49
<i>2.5 Acknowledgements</i>	56
<i>2.6 Tables</i>	57
2.6.1 Table 2.1: Statistics of sequencing and mapping for 42 individuals from 14 trios	57
2.6.2 Table 2.2 : Summary of capture efficiency for the selected trios	59
2.6.3 Table 2.3 : Summary of de novo mutations found during high-throughput sequencing	61
2.6.4 Table 2.4 : Summary of DNM found using deep sequencing and validated using sanger sequencing	66
2.6.5 Table 2.5 : Total of coding sequences (Mbp) with a coverage $\geq 20x$	67
2.6.6 Table 2.6 : DNM rates comparison between SCZ and previously reported studies of normal individuals	67
<i>2.7 Figure</i>	68
2.7.1 Figure 2.1 : Exome Capture Efficiency	68
CHAPITRE 3 : LE FARDEAU DES MUTATIONS RARES DANS LA SCHIZOPHRÉNIE	69
3.1 PRÉSENTATION DE L'ARTICLE.....	69
3.1.1 <i>Contribution des auteurs</i>	70
MUTATION BURDEN OF RARE VARIANTS IN SCHIZOPHRENIA	71
3.2 <i>Abstract</i>	71
3.3 <i>Introduction</i>	72
3.4 <i>Methods</i>	73
3.4.1 Cohorts	73
3.4.2 Samples preparation ,quantification and digestion.....	74
3.4.3 Genes selection	74
3.4.4 Design, Capture and enrichment	75
3.4.5 Sequencing.....	76
3.4.6 Alignment, enrichment assessment and variant calling.....	76
3.4.7 False negative and false positive assessment.....	76
3.4.8 Kernel association testing	77
3.4.9 Sequenom haplotyping.....	78
3.5 <i>Results</i>	78
3.6 <i>Discussion</i>	81
3.7 <i>Tables</i>	84
3.7.1 Table 3.1 List of genes included on the resequencing assay	84
3.7.2 Table 3.2 SKAT results for the dataset grouped by data source.....	87
3.7.3 Table 3.3 Significant association made on a gene and exon basis	87

3.8 Figures	88
3.8.1 Figure 3.1 Network of genes with more than two protein :protein interaction with genes harboring DNM	88
3.8.2 Figure 3.2 Histogram of variant frequencies	89
3.8.3 Figure 3-3 SKAT results for all genes on the assay	90
3.8.4 Figure 3.4 SKAT results for all exons on the assay.....	91
3.8.5 Figure 3.5 Additional genotyping for the SBNO1 haplotype.....	92
CHAPITRE 4 L'ÂGE PATERNEL EXPLIQUE UNE PROPORTION MAJEURE DE LA VARIABILITÉ DES MUTATIONS DE NOVO GERMINALES DES INDIVIDUS EN SANTÉ	93
4.1 PRÉSENTATION DE L'ARTICLE.....	93
4.1.1 <i>Contribution des auteurs</i>	95
PATERNAL AGE EXPLAINS A MAJOR PORTION OF <i>DE NOVO</i> GERMLINE MUTATION RATE VARIABILITY IN HEALTHY INDIVIDUALS.....	96
4.2 ABSTRACT	97
4.3 INTRODUCTION.....	97
4.4 METHODS	98
4.4.1 <i>Sequencing</i>	98
4.4.2 <i>SNP Genotyping</i>	98
4.4.3 <i>CNV Genotyping</i>	98
4.4.4 <i>SNV and indels calling</i>	99
4.4.5 <i>WGS identification of SV</i>	100
4.4.6 <i>Sanger sequencing validation</i>	101
4.5 RESULTS AND DISCUSSION.....	101
4.6 TABLES.....	107
4.6.1 <i>Table 4.1 : Summary of DNM found for each twin quartet</i>	107
4.6.2 <i>Table 4.2 : CNVs identified for each family</i>	108
4.7 FIGURES	109
4.7.1 <i>Figure 4.1 : Association of parental age with germline DNM</i>	109
4.7.2 <i>Figure 4.2 : Association of parental age with CNV called by different algorithms</i> ..	110
4.7.3 <i>Figure 4.3 : Circular plot of the autosomal human genome using combined dataset</i>	111
4.6.4 <i>Figure 4,4 : Circular plot of the autosomal human genome using twins dataset</i>	112
CHAPITRE 5 DISCUSSION ET PERSPECTIVES.....	113

5.1 RETOUR SUR LES TRAVAUX.....	113
5.1.1 <i>Les mutations de novo dans la schizophrénie</i>	113
5.1.2 <i>Le séquençage de gènes candidats</i>	115
5.1.3 <i>L'association avec l'âge parental</i>	119
5.2 LES MUTATIONS <i>DE NOVO</i> DANS LE CONTEXTE GLOBAL DE LA SCHIZOPHRÉNIE.....	123
5.3 LA RÉCONCILIATION DES VARIANTS RARES ET DES VARIANTS COMMUNS.....	126
5.4 LA SÉLECTION ÉGOÏSTE ET LES MALADIES NEUROPSYCHIATRIQUES	128
5.5 PERSPECTIVES	130
CHAPITRE 6 CONCLUSION	134
CHAPITRE 7 RÉFÉRENCES	136

Liste des tableaux

Tableau 2.1 : Statistiques de séquençage et d'alignement pour les 42 individus des 14 trios

(Page 57)

Tableau 2.2 : Résumé de l'efficacité de la capture pour les trios sélectionnés

(Page 59)

Tableau 2.3 : Liste des mutations *de novo* identifiées par séquençage à haut-débit

(Page 61)

Tableau 2.4 : Liste des mutations *de novo* identifiées par séquençage à haut-débit et confirmé par séquençage Sanger

(Page 66)

Tableau 2.5 : Résumé du nombre de séquences codantes (en Mbp) avec une couverture supérieure à 20x.

(Page 67)

Tableau 2.6 Comparaison des taux de mutations *de novo* entre celui trouvé pour la schizophrénie et ceux rapportés précédemment pour des individus normaux

(Page 67)

Tableau 3.1 Liste des gènes sélectionnés pour le reséquençage

(Page 84)

Tableau 3.2 Résultats SKAT pour l'ensemble de données groupé par provenance expérimentale

(Page 87)

Tableau 3.3 Liste des associations significatives faites en utilisant l'ensemble des données groupées par gène et par exon.

(Page 87)

Tableau 4.1 Liste des mutations *de novo* identifiées pour chacun des quartets

(Page 107)

Tableau 4.2 SV identifiées pour chaque famille

(Page 108)

Tableau 5.1 : Âge à la conception des patients utilisés lors de l'étude DNM chez les patients atteints de SCZ (Chapitre 2)

(Page 122)

Liste des figures

Figure 1.1 : Schéma simplifié montrant la division des cellules germinales à la base de la création des spermatozoïdes.

(Page 34)

Figure 2.1 : Résumé de l'efficacité de capture

(Page 68)

Figure 3.1 : Réseau de gènes avec plus de deux interactions protéine-protéine avec les gènes identifiés lors des études de mutations *de novo*

(Page 88)

Figure 3.2 : Histogramme de la fréquence des variants identifiés

(Page 89)

Figure 3.3 : Résultats SKAT pour tous les gènes de l'ensemble de données

(Page 90)

Figure 3.4 : Résultats SKAT pour tous les exons de l'ensemble de données

(Page 91)

Figure 3.5 : Génotypage additionnel pour l'haplotype SBNO1

(Page 92)

Figure 4.1 : Association de l'âge parental avec différents type de mutation *de novo* germinale

(Page 109)

Figure 4.2 : Association de l'âge parentale avec les SV identifiés par différents algorithmes

(Page 110)

Figure 4.3 : Graphique circulaire du génome autosome humain en utilisant les quartets de jumeaux ainsi que d'autres articles de la littérature.

(Page 111)

Figure 5.1 : Modèle simpliste de l'hypothèse des rapports de côtes

(Page 112)

Figure 5.2 : Comparaison de maladies psychiatriques avec d'autres maladies complexes à prévalence et héritabilité comparables

(Page 130)

Liste des sigles et abréviations

À noter : Les noms de gènes et de ressources bio-informatique n'ont pas été inclus à cette liste. Lorsque l'abréviation anglaise est très utilisée, au point d'en devenir un standard (ex : GWAS pour Étude d'association génomique), elle a été préservée. Dans les autres cas, l'abréviation française est utilisée.

1KGP : 1000 genome project

a.a. : Acides aminés

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

ASD : Spectre du désordre autistique

BP : Désordre bipolaire

CCDS : Séquence codante consensus

CGH : Hybridation par génomique comparée

CNV : Variation du nombre de copies

DIGS : Étude diagnostique pour les études génétiques

DL : Déséquilibre de liaison

DNM : Mutation *de novo*

DSM-IV : Manuel diagnostique et statistique des troubles mentaux IV

DSM-V : Manuel diagnostique et statistique des troubles mentaux V

EAP : Effet de l'âge paternel

ELG : Étude de liaison génétique

EQJN – QSNT : Étude Québécoise des Jumeaux Nouveau-nés

FIGS : Entrevue familiale pour les études génétiques

GWAS : Étude d'association génomique

Indel : Insertion et délétion

Mb : Mégabase
Mbp : Mégabase
MR : Retard mental
MS : Mutation faux-sens
MHC : Région d'histocompatibilité majeure
NS : Mutation non-sens
OMS : Organisation Mondiale de la Santé
OR : Rapport de côte
PCR : Réaction en chaine de polymérase
SCZ : Schizophrénie
SCZ-DYS : Schizophrénie dysthymique
S2D : Projet Synapse2Disease
SNP : Polymorphisme à simple nucléotide
SNPchip : Biopuce à ADN
SNV : Variation à simple nucléotide
SV : Variation structurale
TsTv : Ration de transition sur transversion
UTR : Région non-transcrite
UV : Ultra-violet
WGS : Séquençage du génome complet
WTCCC : Wellcome Trust Case Control Consortium

À mes parents

Pour m'avoir conçu à temps

Et bien d'autres raisons

Remerciements

L'écriture d'une thèse est un exercice qui requiert un regard critique et rétrospectif sur plusieurs années de travail. On y réalise les bons coups, mais également les choses que l'ont ferait différemment aujourd'hui. Surtout cependant, on réalise que tout cela n'aurait pas été possible sans plusieurs personnes clés.

En tête de liste, mon directeur de recherche, Guy Rouleau. J'ai la certitude qu'il m'aurait été difficile de trouver une collaboration aussi fructueuse que l'a été celle de Guy et moi. Merci pour les nombreuses heures de discussions et les idées folles. Merci de m'avoir confié des responsabilités dépassant largement le cadre d'un doctorat et d'avoir cru en mes projets.

Également un grand merci à Marie-Pierre Dubé, ma co-directrice de recherche. Grace à toi, j'ai appris que les analyses génétiques requièrent une rigueur et une méthodologie profonde. Mais j'ai également appris qu'un bon scientifique pose un regard critique sur tout ce qu'on lui propose. Une étude est bien ficelée seulement lorsque toutes les failles imaginables ont été comblées.

Au cours des nombreuses années au laboratoire, certains collègues ont eu une place particulière dans mon parcours. Notamment Patrick Dion, que je considère aujourd'hui comme un bon ami. Je tiens également à remercier les gars de bio-info; Alexandre, Dan, Édouard, Ousmane et Philippe pour m'avoir enduré. Également un merci spécial à certaines personnes du labo de Notre-Dame qui m'ont permis de transcender mes connaissances : Jean-Baptiste, Martine, Cynthia, Nancy, Claire, Loubna, Maude, Pascale, Hélène, Annie, Amirtha et Anastasia. Un petit merci spécial

à Lan qui a initié mon premier contact avec la génétique de la schizophrénie, lorsque j'étais encore un étudiant d'été!

Un autre merci spécial à la gang de l'institut de Cardiologie, pour les fous rires, mais aussi pour m'avoir guidé dans plusieurs études : Sylvie, Géraldine, Amina, Yassamine, Steve, Mathieu et Marc-André. Un merci spécial à Louis-Philippe, pour avoir été (presque) aussi fou que moi!

Je tiens également à souligner le travail de certaines personnes au niveau du département de biologie moléculaire, notamment Richard Bertrand, Vivianne Jodoin et Julie Mantovani.

Sur un plan personnel, j'aimerais remercier ma famille : Martine, Bruno et Marc-Olivier. Votre support de chaque instant m'a été précieux. Également, merci à mon beau-papa pour m'avoir inspiré et encouragé.

Finalement, merci à mon épouse et complice, Amélie. Merci d'avoir été là pour moi comme personne d'autre, d'avoir navigué entre les sautes d'humeurs, les excitations et les découragements.

Finalement, merci à TMT.

Sincèrement,

Simon L. Girard

Étude sur le rôle des mutations *de novo* dans l'étiologie génétique de la schizophrénie

Chapitre 1 : Introduction

La schizophrénie (SCZ) est sans doute la maladie psychiatrique la plus connue. Chaque année, elle fait couler beaucoup d'encre et alimente plusieurs films. Malheureusement, le portrait qu'en font les médias à grand déploiement est souvent incomplet ou erroné. Pourtant, la SCZ est l'une des maladies les plus fréquentes, avec une prévalence globale qui varie entre 0.3 % et 0.7 % de la population générale¹. De plus, la SCZ est l'une des rares maladies qui présente une prévalence ainsi qu'une incidence relativement constante dans l'ensemble des populations humaines au niveau mondial. À tort, la SCZ est souvent vue comme une maladie nuisible, mais non mortelle. Or, il a été prouvé que les patients souffrants de SCZ ont une espérance de vie de 12 à 15 ans plus courte que la population générale². Cette espérance de vie écourtée est en partie causée par des décès directement liés à la maladie (suicide), mais surtout par des décès indirects liés à des comportements à risques (mauvaise alimentation, sédentarité, consommation de substances toxiques)

². De ce fait, il est possible de conclure que la SCZ est annuellement directement et indirectement responsable de plus de décès que la majorité des cancers. Cependant, un lourd tabou entouille la SCZ, comme la plupart des maladies mentales. Ainsi, selon l'Organisation mondiale de la santé (OMS), plus de la moitié des patients souffrant de SCZ ne reçoivent pas les soins appropriés³.

Étant donné que la complexité clinique de la schizophrénie a déjà fait l'objet de plusieurs ouvrages complets et qu'elle n'est pas le sujet de la recherche ici présentée, une courte présentation clinique de la maladie sera présentée afin d'ensuite faire le lien avec le sujet principal de l'ouvrage présent : l'étiologie génétique de la schizophrénie.

1.1 Schizophrénie : Présentation clinique

1.1.1 Symptômes

Les symptômes reliés à la SCZ sont vastes et variés, mais affectent tous le processus de la pensée et de la réponse émotionnelle. D'ailleurs, le terme « schizophrénie » vient de la jonction de deux mots grecs : « schizo » qui signifie fractionnement et « phrénie » qui désigne l'esprit; un fractionnement de l'esprit. Cette appellation a été utilisée pour la première fois en 1908 par le psychiatre suisse Eugen Bleuler. Ce terme est aujourd'hui si bien ancré dans le langage populaire qu'il est pratiquement plus souvent utilisé de manière littéraire que pour décrire la maladie qu'il désigne.

On peut classifier les manifestations cliniques de la SCZ en deux grands types de symptômes : positifs et négatifs⁴. Les symptômes positifs sont des manifestations qui sont absentes chez des individus sains, mais que l'on retrouve en spectre chez les patients atteints de SCZ. Il est à noter que les symptômes positifs sont souvent désignés par le terme « psychose ». Cette association, bien qu'elle ne soit pas foncièrement erronée, est incomplète. En effet, il est important de noter que la psychose, souvent décrite comme un état mental de perte de contact avec la réalité,

n'est pas spécifique à la SCZ. Parmi les manifestations cliniques constituant les symptômes positifs, on retrouve évidemment les hallucinations, décrites comme des perceptions sans stimulus. Ces hallucinations peuvent toucher tous les sens, mais les hallucinations auditives sont les plus fréquentes. Parmi les symptômes positifs, on retrouve également les délusions, décrites comme de fausses croyances entretenues par le patient ainsi que les désordres de la pensée qui se manifestent souvent par des discours illogiques ou un cheminement désordonné des idées. Une caractéristique très importante des symptômes positifs est que le patient est incapable de réaliser que les symptômes ne sont pas réels et causés par la maladie⁴. Pour ce qui est des symptômes négatifs, ils consistent en des symptômes ayant un impact nuisible sur les interactions sociales du patient. Ces symptômes incluent notamment un retrait social, une perte de motivation et d'initiative ainsi qu'un discours pauvre et répétitif. Ces symptômes sont rencontrés dans plusieurs autres maladies psychiatriques et ne sont pas des symptômes caractéristiques de la SCZ.

1.1.2 Diagnostic

Le diagnostic de la schizophrénie se fait principalement à l'aide des symptômes positifs. La grande difficulté du diagnostic réside dans le fait que plusieurs autres maladies psychiatriques sont également associées avec des épisodes psychotiques. Une analyse psychopathologique des maladies à épisode psychotique a permis de classer les symptômes en cinq catégories :

- 1) Les symptômes liés à la psychose elle-même
- 2) Altérations de la volonté et de la prise d'initiative

- 3) Altérations neurocognitives
- 4) Dérégulation affective menant à la dépression
- 5) Dérégulation affective menant à la manie⁴.

Cette classification permet de différencier les maladies à profil psychotique en utilisant un facteur de gradation pour chacune des cinq catégories. Ainsi, si la psychose est observée pour la SCZ, le trouble bipolaire (BP) ainsi que pour la schizophrénie dysthymique (SCZ-DYS), le patient SCZ aura une augmentation des symptômes négatifs ainsi que des altérations neurocognitives. Le patient BP sera caractérisé par des épisodes maniaques beaucoup plus importants tandis que le patient SCZ-DYS aura une augmentation marquée des symptômes de la dépression⁵. Il existe deux échelles de classification utilisée pour le diagnostic de la SCZ, le 4^e Manuel diagnostique et statistique des troubles mentaux (DSM-IV) ainsi que la 10^e classification internationale des maladies (ICD-10).

Un débat important existe en ce qui concerne l'avenir du terme « schizophrénie ». En effet, certains pays, comme le Japon, ont complètement cessé d'utiliser le terme « schizophrénie » pour le remplacer par un terme plus descriptif « syndrome de dérégulation de l'intégration ». Tout récemment, une mise à jour des critères diagnostiques de la SCZ a été effectuée avec l'avènement du nouveau DSM-V. La plus grande différence par rapport à son prédécesseur est que le DSM-V ne considère plus aucun sous-type de SCZ.

1.1.3 Traitements

Les médicaments antipsychotiques constituent la première ligne de traitement de la SCZ¹. Typiquement, ces médicaments peuvent réduire les symptômes positifs dans un délai d'environ 7 à 14 jours. Cependant, ces médicaments n'ont pas démontré d'effets majeurs pour le traitement des symptômes négatifs ou bien des dysfonctions cognitives⁶. Il existe deux types de médicaments antipsychotiques; les antipsychotiques typiques (aussi appelés antipsychotiques de première génération) et les antipsychotiques atypiques (aussi appelés antipsychotiques de deuxième génération). Les antipsychotiques typiques ont été identifiés autour des années 1950 et sont pour la grande majorité des bloqueurs des voies dopaminergiques du cerveau. Les antipsychotiques atypiques ont été développés plus tard et bien qu'ils soient également des bloqueurs de la voie dopaminergique, ils seraient beaucoup moins associés au syndrome neurologique extrapyramidal (initiation anormale des mouvements), l'un des effets secondaires majeurs liés aux antipsychotiques typiques. Cependant, il a été montré par une méta-analyse que les antipsychotiques atypiques auraient les mêmes effets indésirables que les typiques⁷, ce qui a ouvert la porte à l'avènement d'un débat à proportion titanesque dans le domaine de la psychiatrie. L'influent journal médical *Lancet* a contribué à amplifier les échanges en déclarant *qu'il était temps d'abandonner les termes première et deuxième génération lorsque l'on fait référence aux antipsychotiques, car la distinction n'est pas méritée*⁸.

1.1.4 Épidémiologie

Comme mentionné précédemment, la prévalence globale de la SCZ est estimée autour de 0.3 % à 0.7 %¹, ce qui nous permet d'estimer qu'à l'échelle mondiale, environ 24 millions d'humains souffrent de SCZ³. Il a été observé que les hommes étaient plus à risque d'être diagnostiqué de la SCZ que les femmes par un facteur de 1,4⁹. Dans la même lignée, il a également été démontré que les hommes avaient tendance à présenter des symptômes plus précocement que les femmes, avec une différence d'environ 4 à 10 ans^{10,11}. Finalement, un fait très intéressant est que la prévalence de la SCZ est plutôt stable au travers du globe, avec une prévalence standardisée pour l'âge variant de 343 à 544 atteints par 100 000 habitants.¹²

1.1.5 Causes

Lorsqu'il s'agit d'identifier les causes de la SCZ, un nombre incalculable de facteurs ont été pointés du doigt. Par contre, dans une majorité des cas, les études n'ont pas été répliquées et il n'a pas été possible de tirer des conclusions. Dans cette section, seuls les facteurs ayant été impliqués dans plus d'une étude et portant sur au moins deux populations indépendantes seront abordés. À noter que les causes génétiques seront discutées exhaustivement dans la section suivante et ne seront pas abordées ici.

1.1.5.1 Facteurs prénataux

Il a été démontré à plusieurs reprises que les complications obstétriques sont associées avec une augmentation des risques de développer la maladie. Cependant,

il est très difficile de quantifier ce risque, car il semble être très petit¹³. Similairement, il a été montré que les individus nés durant l'hiver et l'été seraient plus à risque de développer la SCZ¹⁴. Cependant, encore ici, les risques seraient très petits. L'une des causes environnementales les plus fréquemment liées à la SCZ est l'hypoxie (manque d'oxygène pendant ou après la naissance)¹⁵. Cette hypothèse est également supportée par des preuves biologiques fonctionnelles, notamment par des preuves de modèles cellulaires¹⁶ et animaux¹⁷. Cependant, il est très difficile d'estimer les risques directs de souffrir de SCZ à la suite d'une hypoxie natale et de ce fait, la proportion totale de la SCZ expliquée par l'hypoxie demeure inconnue. Cependant, certaines études ont pu faire le lien entre certains gènes et l'hypoxie, montrant que plusieurs gènes trouvés en association avec la maladie corrélaient positivement dans une approche mécanistique de réseaux d'interactions¹⁸. Cependant, ces résultats doivent être mis en veille, car ils n'ont pas été répliqués avec les trouvailles génétiques des six dernières années.

Finalement, le dernier facteur prénatal, qui sera discuté en détail plus loin, serait l'âge paternel. En effet, il a été démontré qu'un père plus âgé au moment de la conception de son enfant lui confèrera un risque plus élevé de développer la schizophrénie¹⁹.

1.1.5.2 Facteurs infectieux

L'hypothèse d'une source infectieuse pour la SCZ, aussi connue sous le nom de théorie microbienne des maladies, stipule que la SCZ serait le résultat de l'interaction entre le fœtus et des agents infectieux ou des anticorps que la mère

aurait développés en réponse à des agents infectieux²⁰. Ces interactions auraient des effets néfastes sur le développement du cerveau et du système nerveux périphérique. Ainsi, des associations ont été établies entre la SCZ et plusieurs maladies infectieuses, dont l'influenza^{21,22} et le virus de l'herpès²³. Cependant, ces résultats ont été contredits par des études plus récentes qui avaient l'objectif de les répliquer²⁴⁻²⁶.

1.1.5.3 Consommation de drogues

La causalité entre la consommation de drogues et le diagnostic de la schizophrénie est très difficile à établir. Comme les patients souffrants de SCZ sont connus pour être à grands risques pour la consommation de drogues (près de 40 % des patients SCZ présente une consommation pathologique)²⁷, il est difficile d'établir les effets de cause et de conséquence. Malgré la grande médiatisation entre la consommation de cannabis et le risque de développer la SCZ, le lien n'est pas encore clairement établi. Bien que certaines études démontrent que le cannabis augmente le risque d'être diagnostiqué de la SCZ^{28,29}, des facteurs épidémiologiques contredisent ces affirmations. En effet, des pays comme la Suède et le Japon, connus pour avoir une très basse consommation de cannabis, ont une prévalence de la SCZ similaire à des pays comme le Canada ou les États-Unis où la consommation est plus élevée. De ce fait, il faut être très prudent dans les interprétations des facteurs de risques de la SCZ afin de ne pas établir des causalités erronées.

1.2 L'étiologie génétique de la schizophrénie

1.2.1 Introduction

Au début de l'année 2013, une recherche des termes « schizophrenia AND genetics » réalisée dans le cadre de cette thèse rapportait près de 12 000 articles de nouveaux résultats. De ce fait, lorsqu'une démarche de revue de la littérature du domaine est entreprise, il faut être très prudent lorsque l'on établit la liste des « gènes associés ». En effet, la base de données SZGene (<http://www.szgene.org/>) rapporte plus de 1000 gènes associés à la SCZ alors que la base de données SZGR (<http://bioinfo.mc.vanderbilt.edu/SZGR>) en rapporte près de 8000. Dans le cadre de l'exercice courant, ne seront discutées que les trouvailles ayant eu un fort impact sur le domaine de l'étiologie génétique de la SCZ et ayant été répliquées. Il est à noter que ce résumé est fortement inspiré de deux articles de revues préalablement publiés^{30,31} qui sont joints en annexe de cette thèse.

1.2.2 L'origine de l'hypothèse génétique

Encore aujourd'hui, il est très fréquent d'entendre des spécialistes remettre en doute l'hypothèse génétique de la SCZ. Cependant, ces doutes sont injustifiés, car la composante héréditaire de la SCZ a été démontrée à maintes reprises. En effet, afin de comprendre les fondations moléculaires d'une maladie, il est crucial de pouvoir évaluer et distinguer l'apport des composantes génétiques et environnementales. Les études épidémiologiques effectuées sur des paires de jumeaux monozygotiques et dizygotiques constituent la méthode de choix pour y arriver. Ces études

permettent non seulement d'établir si une maladie a une composante héréditaire, mais elles permettent également de la quantifier en établissant une mesure appelée facteur d'héritabilité. Pour la SCZ, plusieurs études de jumeaux ont été réalisées en utilisant plusieurs critères de sélection, que ce soit sur des patients sélectionnés³², sur des sélections systématiques^{33,34} ou bien sûr des populations spécifiques^{35,36}. Toutes les études réalisées ont pu identifier une augmentation majeure de la concordance entre les jumeaux monozygotiques comparés à des jumeaux dizygotiques. Cette observation permet de conclure hors de tout doute que la SCZ possède une importante composante héréditaire. Cette composante fut estimée autour de 81 % lors d'une méta-analyse englobant 12 études indépendantes de jumeaux³⁷. Ceci permet de conclure que la composante génétique de la SCZ est majeure, mais aussi que l'environnement est appelé à jouer un rôle important dans l'architecture de la maladie, puisqu'environ 20 % des cas pourraient être causés par des facteurs environnementaux.

1.2.3 Les études de liaisons

1.2.3.1 Définition

Les études de liaisons génétiques (ELG) se basent sur le principe que les gènes proximaux ont une plus grande probabilité d'être transmis ensemble sur le même fragment chromosomique. Ce processus est rendu possible grâce à la recombinaison chromosomique, qui se produit naturellement lors de chaque transmission génétique. En estimant le taux de recombinaison entre deux marqueurs chromosomique, il est possible d'estimer la distance génétique séparant ces deux

marqueurs. En présence d'une famille où plusieurs membres sont étudiés, il est possible d'étudier chacune des transmissions afin de reconstituer les haplotypes parentaux transmis par les parents. En présence d'une famille où certains des membres sont atteints d'un trait ou d'une maladie, l'ELG identifie les haplotypes spécifiques aux membres de la famille atteints en donnant à chaque région une valeur de confiance statistique. Ainsi, il est possible d'identifier et de raffiner des régions liées aux maladies. Historiquement, les ELG ont souvent été réalisées à partir de marqueurs microsatellites, c'est à dire des motifs de 2 à 10 nucléotides répétés en tandem à de nombreuses reprises. Cependant, plus récemment, les ELG sont plus souvent effectuées à partir des polymorphismes à nucléotide simple (SNP pour *Single Nucleotide Polymorphism*), moins informatifs, mais plus nombreux.

1.2.3.2 Les études de liaison et la schizophrénie

Depuis le début des années 90, les ELG ont connu un immense succès. L'exemple le plus fréquemment cité est le notoire *BRCA1*, identifié grâce à des familles souffrant de cancer du sein précoce³⁸. Plusieurs autres exemples s'appliquent à d'autres maladies, en particulier pour les maladies à profil monogénique, aussi appelé maladies mendéliennes. Cependant, un succès similaire n'a pas été atteint dans le domaine de la SCZ. Malgré le fait qu'on rapporte des études de liaison à échelle génomique (ELEG) dans pas moins de 25 études indépendantes portant sur de grandes familles avec de multiples patients atteints de SCZ³⁹, le succès de la plupart des études a été mitigé, la plupart des loci n'étant pas répliqués par la suite. Le premier locus associé à la schizophrénie a été identifié dans une famille du

Maryland sur le chromosome 22q13.1⁴⁰. Cependant, les efforts de séquençage subséquents n'ont pas mené à l'identification du gène causatif. C'est le gène Neuregulin 1 (*NRG1*) qui fut le premier gène associé à la SCZ lors de l'identification de deux haplotypes « à risque » dans la population islandaise⁴¹. Le gène *NRG1* code pour une protéine qui agit sur les récepteurs de facteur de croissance épidermique, processus impliqué dans le développement du système nerveux, ce qui rendrait plausible son implication dans le mécanisme biologique de la SCZ. Au fil des années, ce locus génétique a été répliqué dans d'autres études portant sur plusieurs populations différentes, faisant de *NRG1* un gène fortement associé avec la SCZ^{42,43}. L'autre locus qui sera présenté dans le présent ouvrage est le locus contenant le gène *DTNBP1*, un gène codant pour la protéine dysbindine, impliquée dans les cellules des muscles squelettiques. Ce gène a été associé à la SCZ par le biais d'une large ELG sur plus de 250 familles irlandaises⁴⁴. Cette même étude a rapporté la présence d'association avec plusieurs marqueurs SNP délimitant deux régions génomiques, l'une d'elles contenant le gène *DTNBP1*. Des études moléculaires fonctionnelles ont par la suite renforcé ces résultats en montrant un impact des haplotypes associés à la SCZ sur l'expression génique⁴⁵.

Les succès mitigés des ELG pour la SCZ furent le premier indicatif que la maladie possède un profil génétique complexe. Ce fut d'ailleurs le cas pour la plupart des maladies à haute fréquence dans la population, qu'on appelle aujourd'hui « maladies complexes ». Bien que les ELG peuvent être réalisées sur une seule famille à la fois, le pouvoir vient dans la combinaison de plusieurs familles afin d'augmenter le pouvoir de détection des haplotypes liés. Ainsi, si plusieurs familles sont liées au

même locus, la détection d'une région liée est facilitée. En revanche, si la majorité des familles possèdent des liaisons distinctes, la détection de ne serait-ce que d'un seul loci s'en trouve complexifiée par le profil génétique hétérogène.

1.2.3.3 Le cas de DISC1

L'association du gène *DISC1* (Disrupted-in-Schizophrenia 1) à la SCZ est sans contredit la plus connue et la plus étudiée de toutes. À ce jour, près de 400 articles ont étudié divers aspects de cette association. L'histoire de *DISC1* est quelque peu particulière. En effet, les premiers indices remontent à l'identification d'une translocation dans une famille écossaise sur le chromosome 1⁴⁶. C'est plusieurs années plus tard que l'on a découvert que cette translocation affectait directement un gène, nommé pour cette occasion *Perturbé-par-la-schizophrénie (Disrupted-in-Schizophrenia-1 DISC1)*⁴⁷. Il a par la suite été démontré que *DISC1* contribuait non seulement à la pathogenèse de la SCZ, mais affectait également d'autres maladies psychiatriques comme BP et la dépression majeure⁴⁸.

1.2.3.4 Perspectives : L'absence de conclusions; une conclusion?

Au plus fort de la vague des ELG, à la fin des années 90 et au début des années 2000, le manque de succès de ces études pour la SCZ pouvait en partie s'expliquer par la complexité phénotypique et une possible non-pénétrance. Bien que peu de liaisons génétiques aient été identifiées et répliquées, certaines, comme *DISC1*, montrent que la SCZ semble posséder une composante héréditaire mendélienne. Cette composante est probablement en grande partie masquée par la portion complexe de

la maladie. Avec l'avènement des nouvelles technologies de séquençage, plus particulièrement avec la technique dite de capture d'exome (définie plus loin dans l'ouvrage), nous aurions été en droit d'espérer des résultats plus encourageants pour les études familiales. Or, il est étonnant de voir que le succès escompté ne s'est pas matérialisé. En effet, pratiquement cinq ans après la globalisation des études de séquençage de deuxième génération, aucun gène à profil de transmission mendélienne n'a été identifié et répliqué pour la SCZ. Serait-il possible que cette absence de résultat soit un indice sur la nature du mécanisme génétique de la SCZ? Peut-on conclure qu'il n'existe pas de famille atteinte d'une forme monogénique de la maladie? Il est encore trop tôt pour tirer une conclusion définitive. En effet, il n'est pas impossible que certains cas de SCZ familiale soient causés par un type de variation génétique encore sous-étudié, comme ce fut le cas pour la sclérose latérale amyotrophique, dont une vaste proportion des cas familiaux est causée par une répétition d'hexanucléotide qui avait jusqu'à tout récemment échappé aux équipes de recherches. Il n'est pas non plus exclu qu'un effet mosaïque soit présent et que les mutations génétiques puissent seulement être détectées dans des tissus spécifiques. Or, pour la SCZ, il est encore impossible de bien situer les mécanismes biologiques. Ainsi, il serait probablement plus judicieux de concentrer les prochaines études génétiques sur des populations plutôt que sur des familles.

1.2.4 Les études d'association génomique

1.2.4.1 Définition

D'une manière simpliste, les études d'association génomique (GWAS, de l'anglais *Genome-Wide Association Study*) peuvent être vues comme le pendant populationnel des ELG. Comme il n'est pas possible d'observer la ségrégation d'un trait parmi les membres affectés, on utilise plutôt la fréquence allélique pour tenter d'établir s'il y a une différence entre un groupe de patients et un groupe de contrôles. Dans le cas présent par contre, il n'est plus possible de travailler avec des haplotypes basés sur les fréquences de recombinaisons. De ce fait, on parle plutôt de déséquilibre de liaison (DL). Si des combinaisons d'allèles se produisent plus souvent que si ces combinaisons étaient dictées par le hasard, il y a présence d'un couplage allélique appelé déséquilibre de liaison. Lorsqu'on a commencé à s'intéresser plus sérieusement aux variations génétiques au niveau populationnel, on a réalisé que le génome était fragmenté en bloc de DL de tailles variables. Un bloc de DL est un segment chromosomique continu qui est déterminé en comparant les haplotypes de plusieurs membres d'une même population, comme le fait le projet HapMap⁴⁹. Avec une connaissance précise des blocs de DL, il est possible de déterminer une liste de SNP « étiquette » qui permet de réduire le nombre de SNP à étudier d'environ quarante millions à moins d'un million.

À l'instar des études de liaisons, une association positive entre un marqueur SNP et un trait phénotypique n'est pas nécessairement concluante de l'identification d'une mutation causative. Il n'est pas exclu que le SNP ainsi identifié soit le variant causal

recherché, mais, dans la plupart des cas, il faut envisager que le variant causal soit en haut DL avec le marqueur SNP, mais qu'il devra être identifié par séquençage ou autre technique.

Le grand essor des études d'association, particulièrement celui des études d'association à l'échelle génomique (GWAS) est très récent. En effet, avant l'achèvement du « projet génome humain » en 2003 et d'un projet dérivé; le projet HapMap (expliqué plus loin), les connaissances des variations génétiques chez l'humain étaient trop limitées pour établir une carte génétique des SNP à l'échelle génomique. Aujourd'hui, plus de 800 GWAS ont été réalisées et plus de 53,000 associations positives ont été réalisées (certains gènes sont associés à plusieurs maladies)⁵⁰.

1.2.4.2 La première vague d'étude d'association

La première vague des GWAS a rapidement confirmé l'une des observations faite lors des ELG; la grande complexité génétique de la SCZ. En effet, la grande majorité des études se sont révélées non concluantes. Il faut dire que ces études, qui eurent lieu à l'aube de l'avènement des GWAS (2008-2009), ont été réalisées sur des échantillons de tailles modestes comparées au standard d'aujourd'hui. Trois de ces études comprenaient des échantillons de taille légèrement plus substantielle. Les deux premières, comprenant respectivement 574 et 733 patients atteints de SCZ furent non-concluantes^{51,52}. La troisième étude, réalisée sur 660 patients, a généré plus d'intérêt suite à la découverte d'une association avec le gène Reelin (*RELN*). Curieusement, l'association rapportée par Shifman et coll. avec RELN était

seulement observable chez les femmes⁵³. Cet intérêt vient en partie du fait qu'une baisse du niveau d'ARN messager du gène *RELN* avait précédemment été identifiée dans des cerveaux de patients atteints de SCZ^{54,55}. Ces résultats avaient par la suite été renforcés par l'observation suivante : dans des cerveaux de patients atteints de SCZ, dans les couches corticales I, il y a surexpression de l'ADN méthyltransferase 1 (DNMT1) qui mène à la baisse d'expression de *RELN*. Cette constatation était d'autant plus intéressante suite à l'observation de l'hyperméthylation du promoteur de *RELN* chez des patients SCZ^{56,57}. Cependant, des études plus récentes ont échoué lorsqu'elles ont voulu répliquer les états d'hyperméthylation^{58,59}. De ce fait, l'association avec le gène *RELN* a été éclipsée par les plus récentes trouvailles génétiques, surtout lorsque l'on tient compte du fait qu'elle n'a jamais été répliquée dans les études génétiques subséquentes.

1.2.4.3 La deuxième vague : Premières percées

La première GWAS sur un échantillon de taille moyenne (479 patients et 2937 contrôles) s'est soldée par l'identification d'une association positive avec un SNP situé dans le gène codant pour la protéine zinc finger 804A (*ZNF804A*)⁶⁰. Il s'agit sans doute de la première association solide identifiée par les GWAS pour la SCZ. En effet, elle a été répliquée à maintes reprises dans plusieurs populations⁶¹⁻⁶³. En modifiant l'expression de *ZNF804A* dans des cellules progénitrices neurales, il a été montré que le gène avait un rôle important dans le processus d'adhésion cellulaire, supportant la théorie que *ZNF804A* serait impliqué dans la migration neuronale et la formation des synapses⁶⁴.

À la suite de cette étude, deux nouvelles GWAS furent publiées conjointement, la première menée par le groupe islandais DECODE sur un échantillon de 2663 cas et de 13498 contrôles⁶⁵ alors que la deuxième a été conduite par le Consortium international sur la Schizophrénie sur 3322 cas et 3587 contrôles⁶⁶. Le point majeur des deux études est que chacune d'elle a rapporté la présence de SNP associés dans le complexe majeur d'histocompatibilité (MHC pour *Major Histocompatibility Complex*). Le MHC est un ensemble de gènes situé sur le chromosome 6 qui code pour la réponse immunitaire d'un individu, et, de par les combinaisons alléliques possibles, est unique pour chaque individu. En plus d'avoir été répliqué par une troisième étude indépendante⁶⁷, le locus MHC avait auparavant été identifié par une métaanalyse d'ELG⁶⁸. De plus, une hypothèse émergente liant la SCZ aux réponses immunitaires supporterait l'implication du MHC dans la pathogenèse de la maladie^{69,70}.

1.2.4.4 Troisième vague : La consécration?

Comme pour la plupart des études génétiques, les premières GWAS ont été réalisées sur des populations à ascendance européenne. Il a fallu attendre la fin de l'année 2011 avant d'obtenir les premiers résultats dans des populations non européennes. Deux études portant sur la population chinoise Han ont été publiées conjointement, la première portant sur 3750 cas et 6468 contrôles alors que la deuxième a été effectuée sur 746 cas et 1599 contrôles^{71,72}. Cependant, plutôt que d'éclaircir le mystère, elles ont contribué à le densifier. En effet, les deux études ne sont pas concordantes et n'identifient aucun signal commun.

Finally, the largest GWAS has been published recently. It covers 9394 cases and 12462 controls. In addition to confirming the association with the MHC, this study has allowed the identification of five new loci associated with the disease⁷³. The most interesting of these new loci is the one where one of the predicted targets of microRNA 137, which was previously implicated in neuronal maturation and the development of the spinal cord⁷⁴. Given the recent publication of this large GWAS, functional analyses reporting a direct link between microRNA-137 and SCZ are still expected. However, it is very likely that several research groups are currently trying to establish a causal link and that several pieces of evidence will reach us soon.

1.2.4.5 Perspectives : Stop ou encore?

When reviewing all the GWAS conducted for SCZ, it is evident that one of the main principles of association studies is at work. The success of GWAS is directly linked to the size of the sample and to the complexity of the disease. It is already known that SCZ is a very complex disease; therefore, large samples will be required to reach positive associations. Thus, if the first studies were not very conclusive, the most recent studies clearly show that several loci are positively linked to SCZ. However, it is important to keep in mind that the results obtained up to today explain only a small portion of the heritability. In addition, each association is linked to a small increase in the absolute risk of the disease. For example, the most recent association of the MHC with microRNA137 was established with a ratio of

cote (OR pour Odds ratio) de 1.14. Converti en risque absolu, cela veut dire qu'un individu porteur de l'allèle à risque voit ses chances de souffrir de la maladie augmentées de 3.2% par rapport à un individu qui n'est pas porteur de l'allèle. Montrée ainsi, la contribution de chacun des variants associés par GWAS est très petite. Même en combinant tous les variants positifs, il n'est pas possible d'expliquer plus de 5 % du facteur héréditaire de la schizophrénie. Cependant, il n'en demeure pas moins que ces locus identifiés sont d'une importance capitale pour les études subséquentes.

En plus des associations individuelles, la globalité des GWAS permet de tirer une autre conclusion sur la complexité de la SCZ. L'une des approches possibles pour pallier à cette complexité est une poursuite en augmentant la taille de l'échantillon. En effet, des études sur d'autres traits complexes supportent la notion relative aux GWAS que « *the bigger the better* ». Ainsi, une méta-analyse portant sur les facteurs génétiques de la taille chez l'humain a été réalisée sur près de 200 000 individus⁷⁵. Malgré cet échantillon de taille sans précédent, il est estimé que la totalité des associations identifiées dans cette étude n'explique pas plus de 10 % de la variabilité observée dans la taille des humains. En ce qui concerne la dernière étude portant sur la SCZ, elle a été réalisée sur près de 10000 patients en phase exploratoire et près du double en phase de confirmation. Cependant, il semble probable que plusieurs associations soient encore cachées, malgré la taille substantielle de l'échantillon. À cet effet, le groupe de Peter Visscher a utilisé les données de la plus récente GWAS comme point de base pour un nouveau calcul. Selon cette étude, en prenant en compte tous les SNP utilisés dans la grande GWAS

de Ripke et coll., on serait en mesure d'expliquer 23 % du facteur héréditaire de la maladie⁷⁶. Ces résultats justifieraient donc l'importance de continuer les études d'associations en SCZ.

1.2.5 Les variations structurales

1.2.5.1 Description

Les variations structurales (SV pour *Structural Variants*) sont des variations génétiques qui adviennent sur de larges fragments chromosomiques (plus de 1000 paires de bases). La complétion du projet « Génome » a révélé que les SV, jusqu'alors très peu connus, pourraient être beaucoup plus fréquents qu'on ne le croyait jusqu'alors⁷⁷. La dernière mise à jour de la base de données *Database of Genomic Variants* enregistre près de 100 000 loci de SV. Plusieurs études ont tenté d'apporter des réponses sur la distribution des SV chez l'humain. Cependant, ces études se révèlent très différentes les unes des autres et brillent par leurs contrastes plutôt que par leurs similitudes. On réalise rapidement que selon les méthodes expérimentales et analytiques employées, les résultats varient énormément. Dans le cas de la SCZ, les études ont été réalisées à partir des données d'intensité de biopuces à ADN (SNPchip). Cette méthode permet de réinterroger les données obtenues lors des GWAS à d'autres fins.

1.2.5.2 Les études de prévalence

Une grande étude réalisée par le *Wellcome Trust Case Control Consortium* (WTCCC) portant sur 8 maladies différentes et réalisée sur plus de 16000 cas et 3000

contrôles partagés a conclu qu'il était très peu probable que les SV communes contribuent significativement aux facteurs héréditaires des maladies⁷⁸. En effet, selon la logique de leur analyse, les signaux de ces SV auraient déjà été détectés lors des GWAS. Cette tendance semble vouloir s'appliquer à la SCZ. En effet, aucune association avec une SV commune n'a été observée. Par contre, l'effet des associations avec des SV rares semble être beaucoup plus important. Avant de parvenir à l'identification spécifique de SV associée à la maladie, plusieurs études de prévalence ont été effectuées. Il a tout d'abord été montré que la prévalence des SV était plus grande dans des cohortes de patients atteints de SCZ comparés à des individus contrôles. Dans une cohorte de 150 patients SCZ, Walsh et coll. ont observé que les longues SV (> 100 kilobases) étaient plus fréquentes chez les patients schizophrènes (15 % comparés à 5 % chez les contrôles). Cette tendance semble être encore plus forte pour les patients atteints de schizophrénie juvénile, une forme rare de la maladie (20 %)⁷⁹. Une autre étude abonde dans le même sens, mais cette fois, ce serait les SV *de novo* (c.-à-d. des mutations spontanées) qui auraient une plus forte prévalence dans un groupe de 152 patients atteints de SCZ. Fait intéressant, cette observation s'applique uniquement aux cas de SCZ sporadiques, la même observation n'ayant pas été répliquée parmi des cas de SCZ familiaux⁸⁰. Une troisième étude a répliqué cette observation, montrant que l'association était inversement proportionnelle avec la fréquence des SV. De plus, cette étude a établi que le nombre de gènes par SV était plus important pour les patients atteints de SCZ que pour les contrôles⁸¹.

1.2.5.3 Les associations spécifiques aux loci de SV

L'étude mentionnée précédemment a également été la première étude à l'échelle génomique qui a rapporté des associations spécifiques pour certaines positions chromosomiques⁸¹. En plus de deux loci situés sur les chromosomes 15q13.3 et 1q21.1, cette étude a permis de répliquer le locus sur le chromosome 22q11.2, le premier locus génétique lié à la SCZ^{82,83}. De manière très contrastée avec les GWAS, une deuxième étude à l'échelle génomique a permis de répliquer deux des loci identifiés par la première étude (1q21.1 et 15q13.3) en plus d'identifier un nouveau locus situé sur le chromosome 15q13.3⁸⁴. Le locus 1q21.1 a été subséquemment lié à la microcéphalie et la macrocéphalie⁸⁵. Par la suite, de nombreuses études ont permis d'effectuer des associations dans des études de SV localisées, notamment sur le chromosome 16p11.2⁸⁶.

1.2.5.4 Perspectives

Il est intéressant de voir le contraste entre les GWAS et les études de distribution de SV. Bien qu'elles soient différentes dans leurs conclusions, elles sont concordantes sur un aspect crucial du mécanisme génétique de la SCZ : les variants communs semblent avoir un rôle très limité dans l'étiologie de la maladie. En effet, bien que certaines associations aient été réalisées dans le cadre des GWAS, les rapports de cotes sont si petits que leurs risques en valeur absolue reste très faibles. En ce qui concerne les SV associés, bien que les fréquences soient très faibles, les rapports de cotes sont beaucoup plus importants, la majorité se trouvant entre 10 et 20 (les rapports de cote du locus 22q11.2 ont été évalués à 21.6!). À titre de comparaison

avec l'exemple donné dans la section précédente, la délétion 1q21.1 a été identifiée avec un rapport de cote de 14.83. Converti en risque absolu, cela veut dire qu'un patient pour lequel on identifie une délétion sur le chromosome 1q21.1 aura 47.85% plus de chances de souffrir de la SCZ qu'un individu non-porteur. Cependant, comme la délétion est très rare (prévalence de moins de 0,04%), ces chiffres confirment les hypothèses génétiques sur l'origine de la maladie : la maladie serait causée par des centaines, voir des milliers de gènes. Ainsi, les SV ont permis par des études de prévalence d'observer un plus haut fardeau des SV chez les patients atteints de SCZ. Elles ont aussi permis d'identifier des loci fortement associés à la maladie. La difficulté des études de SV réside plutôt dans l'identification même des variations. En effet, il a été montré que bien qu'il existe plusieurs méthodes pour identifier les SV à l'échelle génomique, ces études ont un facteur de concordance minimal⁸⁷⁻⁸⁹. Par contre, il a également été montré que le taux de faux positifs de chacune des méthodes est relativement petit. De ce fait, on peut conclure que ces différentes méthodes sont complémentaires plutôt que confirmatoires. Ainsi, il est encore très difficile d'établir avec précision une cartographie fiable de tous les SV d'une population donnée.

1.2.6 Les études de re-séquençage

1.2.6.1 Descriptions

Le séquençage initial du génome humain a été complété pour la première fois en 2001⁷⁷. De ce fait, pour être scrupuleusement exact, toutes les études subséquentes portant sur le séquençage du génome humain sont appelées des études de re-

séquençage (de l'anglais resequencing). Cependant, afin d'alléger le texte, elles seront ici désignées comme études de séquençage.

1.2.6.2 Le séquençage Sanger

Le séquençage Sanger a été développé par Frederick Sanger et son équipe en 1977. Jusqu'à l'avènement des nouvelles technologies de séquençage, c'était la technologie de séquençage la plus utilisée. Le principe consiste à utiliser un petit oligonucléotide, communément appelé amorce, afin d'initier la polymérisation de l'ADN. Par la suite, la réaction est maintenue par des enzymes polymérases, les mêmes que celles utilisées pour les PCR. Par la suite, on ajoute un mélange des quatre désoxyribonucléotides avec une faible quantité de l'un des didésoxyribonucléotides. Ce dernier permet de stopper la réaction d'élongation lors de l'ajout d'une base spécifique. En répétant la réaction à quatre reprises (une pour chacun des quatre didésoxyribonucléotides), on obtient des fragments de tailles différentes que l'on peut faire migrer sur un gel de polyacrylamide afin d'obtenir les positions de chacun des nucléotides dans la séquence. La méthode a par la suite été automatisée, que ce soit par l'ajout de capillaires d'électroporation ou bien en remplaçant les nucléotides standard par des nucléotides phosphorescents.

1.2.6.3 Le séquençage à haut-débit

L'arrivée, dans les quatre dernières années, des nouvelles technologies de séquençage a complètement bouleversé le monde de la génétique. En effet, si pendant de longues années, le séquençage a été effectué par ciblage de petites

régions, il est maintenant possible de paralléliser les réactions pour cibler des millions de régions à la fois. De ce fait, il est désormais réaliste de séquencer un génome complet pour un coût modeste; le premier séquençage du génome humain s'est réalisé pour un total de 4 milliards de \$ US alors que notre laboratoire effectue aujourd'hui une tâche similaire de manière routinière pour moins de 3000 \$ US. Ainsi, il est maintenant possible d'interroger le génome complet sans la nécessité d'avoir une hypothèse de base. Cependant, cette nouvelle accessibilité à l'information génétique ne s'est pas réalisée sans heurts. En effet, il est dorénavant impossible de réaliser une étude génomique sans un support bio-informatique et biostatistique complet. À titre d'exemple, le laboratoire de recherche du Dr Rouleau, est passé d'un laboratoire de génétique traditionnelle à un laboratoire générant approximativement un téraoctet de données par jour. Il est donc nécessaire d'avoir accès à une plateforme d'analyse efficace et rapide afin de pouvoir transformer les données de séquençage en des données utilisables pour des études génétiques.

Le génome humain représente un ensemble complexe d'informations distribuées en séquences codantes et non-codantes. Les séquences codantes représentent les premières cibles de choix des généticiens, car elles sont peu nombreuses (2 % du génome total) et contiennent la quasi-totalité des mutations génétiques associées à des maladies. Ainsi, afin de réduire les coûts globaux et le fardeau des analyses génétiques, une méthode de ciblage appelée «la capture d'exome» a été développée. Cette méthode permet d'isoler et d'amplifier la grande majorité des régions codantes du génome d'un individu de manière à identifier la presque totalité

des variations génétiques ayant un effet direct sur la formation et la structure des protéines.

1.2.6.4 Les études de séquençage de gènes candidats

Dans le domaine de la génétique de la SCZ, plusieurs études de séquençage de gènes candidats ont été réalisées. Il n'est pas possible de toutes les répertorier, mais la plupart ont été réalisées à la suite des études génomiques présentées précédemment, comme c'est le cas pour le gène ZNF804A⁹⁰ ou bien DISC1⁹¹. Cependant, l'interprétation que l'on peut faire de telles études a été drastiquement modifiée par l'arrivée des nouvelles technologies de séquençage. En effet, maintenant qu'il est possible d'interroger l'entièreté des séquences codantes, il a été démontré que les variants à effet nuisible sur les protéines, qu'auparavant on liait directement aux maladies, sont beaucoup plus fréquents qu'on ne le pensait ⁹². Ainsi, il a été démontré qu'un individu sans problème sévère de santé peut avoir jusqu'à 100 variants causant une perte d'effet et on s'attend que 20 de ces variants causent une inactivation complète d'un gène, sans effet phénotypique notable.

1.2.6.5 L'implication des mutations de novo

Il est à noter que les mutations de novo constituent maintenant un facteur important du mécanisme génétique de la schizophrénie. Comme elles sont également le sujet principal de cet ouvrage, une énumération factuelle des études de mutations de novo sera présentée ici et une discussion exhaustive sera effectuée plus loin.

Les mutations de novo (DNM, de l'anglais *de novo* mutations) sont des variations génétiques qui ne sont pas transmises de façon mendélienne; elles sont donc absentes du génome des parents d'un individu. Elles peuvent avoir deux origines différentes: une origine germinale où elles se produisent dans les cellules germinales des parents d'un individu et une origine somatique où elles se produisent à un stade plus ou moins avancé de l'embryogenèse directement chez l'individu. Il est connu depuis longtemps que les DNM sont responsables de plusieurs syndromes génétiques rares comme le syndrome de Rett⁹³ ou la Neurofibromatose de type II⁹⁴. Cependant, il aura fallu attendre beaucoup plus longtemps avant que leur implication soit suggérée dans les maladies plus communes, ce retard étant imputé d'une part à la complexité de leur mécanisme génétique, mais aussi au grand intérêt envers les variations génétiques fréquentes. C'est par le véhicule d'un grand projet de séquençage de gènes synaptiques chez des patients atteints de SCZ ou de troubles du spectre de l'autisme (ASD pour autism spectrum disorder) que notre laboratoire a entrepris la première étude d'envergure des DNM dans les maladies psychiatriques. Ce projet a permis d'identifier des DNM dans certains gènes qui ont été par la même occasion associés à la SCZ, comme *SHANK3*⁹⁵ et *NRXN2*⁹⁶. Mais au-delà des associations locales, c'est l'observation d'une élévation du nombre de DNM à effet nuisible qui a constitué la première preuve que les DNM avaient un rôle important dans la SCZ⁹⁷.

Avec l'avènement des nouvelles technologies, il est dorénavant possible de tester de manière plus efficace l'implication des DNM. Ainsi, des études de capture d'exomes

ont permis d'identifier plusieurs DNM dans des patients atteints de SCZ⁹⁸⁻¹⁰⁰, confortant la thèse de l'implication des DNM dans la SCZ.

1.2.6.6 Perspectives

L'avènement des nouvelles technologies de séquençage milite grandement en faveur de la poursuite des études de séquençage à grande échelle. En effet, le coût du séquençage est de plus en plus abordable et ce qui était impensable il y a à peine quelques années est maintenant presque accessible : le séquençage du génome complet d'un grand nombre d'individus d'une même population.

Il n'y a nul doute que les études de séquençage sont la voie de l'avenir puisqu'elles permettent de tester pour toutes les hypothèses génétiques de par leur nature intégrative et complète. Il est en effet possible de réaliser des ELG, des GWAS ou des études de SV à partir des données de séquençage d'un génome. Et, chaque saison apporte un foisonnement de nouvelles méthodes d'analyses qui permettent d'extrapoler de nouvelles informations à partir des mêmes données.

En ce qui concerne la SCZ, les études de séquençage permettront enfin de réaliser des études intégratives afin de tester l'effet combiné des variants rares et des variants communs. Il y a fort à parier que les cas de SCZ causés de manière mendélienne par une seule mutation sont très rares et que la majorité des cas sont dus à une accumulation de facteurs de risques plus ou moins importants. En effet, s'il y a une leçon que l'on peut tirer des études génétiques réalisées sur la SCZ jusqu'à présent, c'est qu'il est très risqué de mettre tous ses œufs dans le même

panier. Voilà pourquoi les études dites combinatoires commencent à être de plus en plus répandues. Elles permettent de cumuler des signaux d'association dans des ensembles prédéfinis (bien souvent des gènes) afin de tester pour un effet cumulatif de plusieurs variants à fréquences différentes. Il est envisageable que sous peu, il soit possible de tester les effets combinatoires entre des variants venants de différents gènes afin de tester de potentielles combinaisons de multiples variants.

1.3 L'effet de l'âge du père sur les mutations

La réplication fidèle de l'ADN ainsi que la transmission du matériel génétique aux cellules-filles sont des processus critiques. Cependant, malgré l'importance de ces processus, nos génomes sont continuellement exposés à des agents exogènes qui créent des centaines de lésions d'ADN chaque jour, chacune de ces lésions nécessitant une réparation rapide. Les mécanismes cellulaires de réparation et de réplication de l'ADN sont des mécanismes très fiables, mais ils ne sont pas parfaits. Ainsi, dépassant notre compréhension actuelle de ces mécanismes, certaines mutations échappent au contrôle cellulaire et finissent par se retrouver dans le bassin génétique d'une population où elles seront soumises aux lois de la sélection naturelle et de la dérive génétique. D'une part, il a été montré il y a plusieurs décennies déjà que ce mécanisme est crucial pour la survie et l'évolution des espèces¹⁰¹. D'autre part, l'ajout de nouvelles mutations peut aussi avoir des conséquences néfastes qui désavantagent les individus ^{94,102,103}. Les mécanismes par lesquels les mutations sont introduites dans le bassin génétique d'une population ne sont pas très bien compris. Si nous savons que certaines mutations sont introduites

par des erreurs de réplifications ou des cassures double-brins, nous en savons encore très peu sur la fréquence de ces évènements et sur leurs impacts. Cependant, avec les nouvelles technologies de séquençage, il est dorénavant possible d'étudier plus en détail ces mécanismes.

1.3.1 Les types de mutations et leurs mécanismes respectifs

1.3.1.1 Historique

À la suite du séquençage original du génome humain, une grande question s'est imposée : quelle variabilité va-t-on retrouver entre plusieurs génomes d'une même espèce? Pour répondre à cette question, plusieurs des plus grands centres génomiques ont lancé un projet international appelé HapMap⁴⁹ avec pour but avoué d'identifier et de caractériser les variations génétiques entre différentes populations. Suite à l'identification d'un peu plus d'un million de polymorphismes à simple nucléotide (SNP pour Single Nucleotide Polymorphism), leur prévalence a été testée dans quatre populations fondatrices. Par la suite, le projet des mille génomes (1KGP), le successeur spirituel du projet HapMap, a permis de transférer les connaissances acquises sur les populations à un niveau individuel en établissant une carte génétique à haute densité du génome humain^{104,105}. Ces données ayant été disséminées librement dans la communauté scientifique, il est maintenant possible de classifier les différents types de variations génétiques selon leurs effets et leurs fréquences. Ainsi, plusieurs classes générales de variations ont été identifiées.

1.3.1.2 Les types de variants

1.3.1.2.1 Les variants à simple nucléotide

Il existe une ambiguïté relative à la dénomination appropriée pour ce type de variations génétiques. Le nom le plus souvent utilisé est polymorphisme à nucléotide simple (SNP), mais le mot « polymorphisme » sous-entend que le site visé se retrouve chez plusieurs individus dans une population. On ne peut donc pas l'utiliser pour des variations privées (c.-à-d. trouvées uniquement chez un individu). De plus, le seuil au dessous duquel un variant ne peut plus être appelé un SNP n'est pas clairement établi et varie d'une étude à l'autre. Pour ces raisons, le présent ouvrage utilisera le terme de variation à simple nucléotide (SNV pour Single Nucleotide Variant) qui est un terme plus générique que SNP.

Les SNV sont à la fois les variations les plus petites et les plus nombreuses dans le génome humain. Elles résultent d'un mécanisme appelé mutation ponctuelle (aussi appelé substitution de base) qui remplace un nucléotide d'ADN par un autre. Si la substitution remplace une purine par une purine ou bien une pyrimidine par une pyrimidine, alors elle sera appelée transition. Si la substitution remplace une purine par une pyrimidine et vice-versa, on sera plutôt en présence d'une transversion. Comme les effets des transversions sont plus sévères, elles sont généralement placées sous un plus fort contrôle de sélection naturelle et de ce fait, elles sont pratiquement trois fois moins fréquentes que les transitions. La plupart des SNV sont causées par des erreurs lors de la réplication de l'ADN. Or, il arrive également qu'elles soient causées par d'autres facteurs comme le rayonnement ultra-violet

(UV) ou bien les radicaux libres. Comme les SNV affectent seulement un nucléotide, elles sont les plus facilement identifiables. Les SNV ont été impliqués dans presque toutes les maladies affectant l'humain et sont de loin les variations les plus fréquentes.

1.3.1.2.2 Les insertions et délétions

Les insertions et délétions (indels) sont typiquement décrites comme des variations génétiques affectant plus d'un nucléotide et moins de mille. Même si leur taille varie énormément, il a été montré que leur fréquence est inversement proportionnelle à leur longueur¹⁰⁶. Il a également été démontré que la grande majorité des courtes insertions sont des duplications en tandem¹⁰⁶. Bien que les indels de petite taille (un seul nucléotide) puissent être causées par le même mécanisme que les SNV, la plupart des indels sont généralement résultantes d'un glissement de la fourche de réplication ou d'un mécanisme de chevauchement chromosomique inégal¹⁰⁷. La prévalence des indels a longtemps été un mystère, mais récemment, le 1KGP a permis de répondre à cette question en révélant que le séquençage d'environ 1000 individus avait permis d'identifier 1,5 million d'indels (le même nombre d'individus a permis d'identifier 38 millions de SNV)¹⁰⁵.

1.3.1.2.3 Les variations structurales

Aussi appelées variation du nombre de copies (CNV), les variations structurales (SV) sont les plus longues des trois classes. Elles sont plus longues que 1000 nucléotides et peuvent s'étendre sur plusieurs Mégabase (Mb), englobant de longs fragments

chromosomiques. Les SV sont souvent les variations génétiques les plus difficiles à identifier. Bien que le mécanisme à la base des SV ne soit pas encore bien compris, l'implication de certains processus cellulaires a été suggérée, par exemple la jonction des extrémités non homologues¹⁰⁸, la recombinaison allélique non homologues¹⁰⁹ ou la répllication dirigée par la micro homologie¹¹⁰.

1.3.2 La spermatogonie

La spermatogénèse est le processus par lequel les cellules germinales indifférenciées, appelées spermatogonies, se divisent de manière méiotique afin de produire des spermatozoïdes. Ce processus a lieu dans la membrane basale des

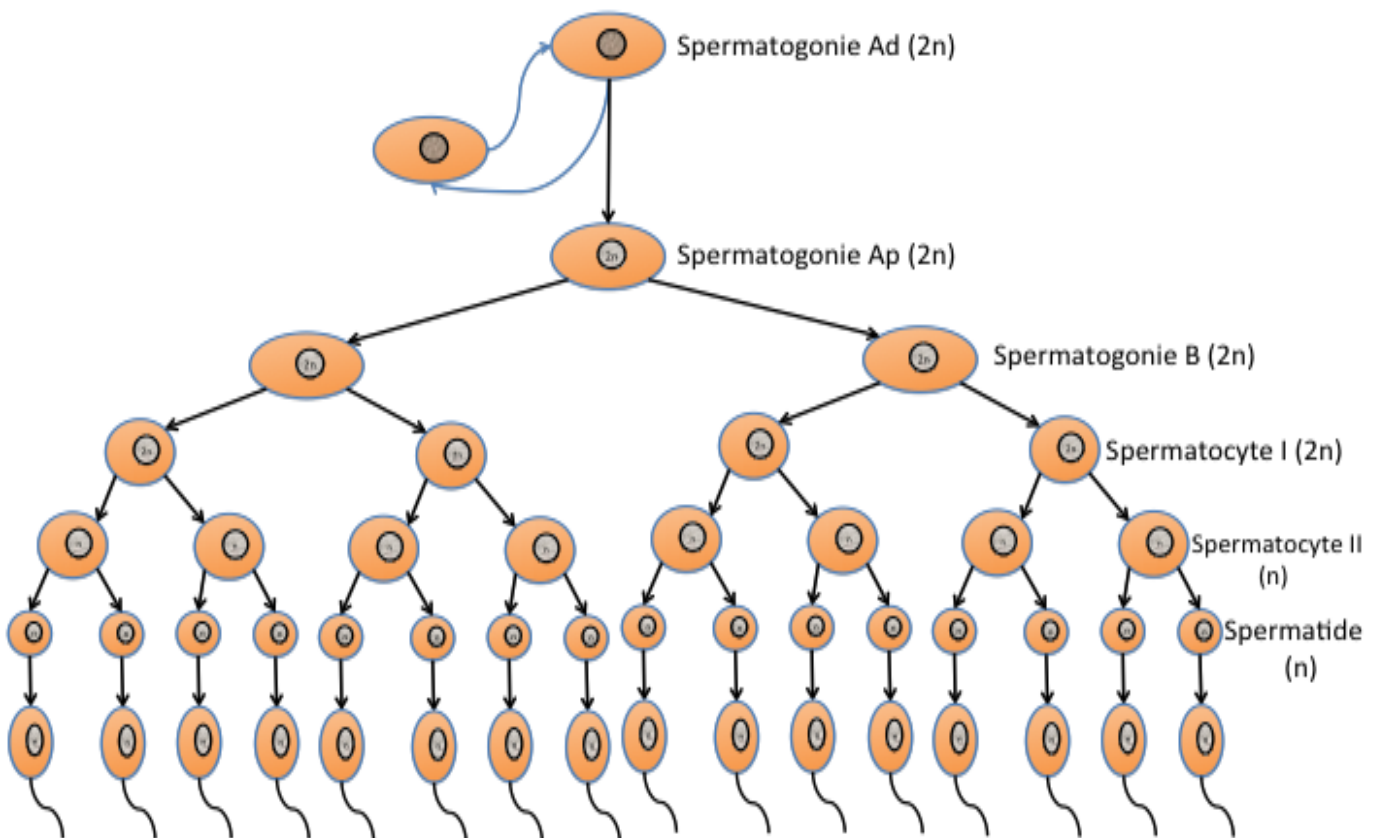


Figure 1 : Schéma simplifié montrant la division des cellules germinales à la base de la création des spermatozoïdes.

tubes séminifères, situés dans les testicules. Chaque testicule est composé de 200 à 300 lobules testiculaires et chacun de ces lobules contient entre 1 et 4 tubes séminifères.

La paroi des tubes séminifères est composée de deux types de cellules : 1) les cellules de Sertoli, qui sont des cellules de soutien nourricières pour la maturation des spermatozoïdes et 2) les cellules germinales. Les cellules germinales sont à la base constituées de deux types de spermatogonies, les types Ad et Ap (voir Figure 1). Bien que les deux types soient des cellules diploïdes, le type Ad est une cellule de maintien qui est responsable de donner un approvisionnement suffisant de cellules souches en ne participant pas directement à la spermatogénèse alors que le type Ap se divise par mitose afin de produire des clones cellulaires connectés par des ponts cytoplasmiques. Ces connexions cytoplasmiques permettent un développement cellulaire synchronisé. Lorsque la division cesse, les cellules ainsi formées se différencient en des spermatogonies de type B. Ce sont ces spermatogonies qui sont responsables de la formation des spermatocytes primaires par division mitotique. Ces spermatocytes primaires, toujours au stade diploïde, seront par la suite déménagés sur le compartiment adluminal où ils se diviseront par méiose en spermatocytes secondaires pour atteindre le stade haploïde. Ces spermatocytes secondaires se diviseront une nouvelle fois afin de former les spermatides haploïdes. Finalement, lors de la spermiogénèse, les spermatides atteindront la maturation afin de former des spermatozoïdes fonctionnels.

1.3.3 L'effet de l'âge paternel

Il est depuis longtemps reconnu que l'âge des parents augmente le risque d'avoir des enfants souffrant de diverses maladies. Parmi les plus connues, on retrouve les anomalies chromosomiques liées à l'âge de la mère, comme le syndrome de Down (Trisomie 21)¹¹¹. Cependant, plus récemment, plusieurs études ont également montré que l'âge du père pouvait indépendamment mener à un risque plus élevé pour certaines maladies, notamment la schizophrénie, mais aussi pour divers cancers¹¹²⁻¹¹⁴ et pour des anomalies congénitales¹¹⁵. Bien que certaines de ces associations soient encore discutables, il existe un petit groupe de maladies, appelées troubles de l'effet de l'âge du père (EAP), pour lesquelles il a été clairement démontré que l'âge du père mène à une augmentation du nombre de DNM dans certains gènes bien précis. Les meilleurs exemples de ces gènes sont les gènes *Fibroblast growth factor receptor 2 (FGFR2)* et *Fibroblast growth factor receptor 3 (FGFR3)* qui peuvent mener à divers syndromes, notamment le syndrome d'Apert¹¹⁶ ou l'achondroplasie¹¹⁷. Ces maladies, bien que rares, représentent un modèle précieux de l'effet de l'âge du père sur les mutations.

Afin de mieux caractériser cet effet, des études de quantification ont été entreprises directement dans le sperme. Cependant, vu la difficulté posée par le grand nombre de cellules différentes, une méthode basée sur la digestion enzymatique suivie par une amplification par PCR a été mise au point pour quantifier la mutation la plus fréquente (c.755C<G) du gène *FGFR2*¹¹⁸. Il a été montré par cette technique que le nombre de cellules porteuses de cette mutation était beaucoup plus élevé dans le

sperme d'hommes en santé que dans le sang de ces mêmes hommes, ce qui suggère une sélection positive des cellules germinales porteuses de cette mutation au niveau de la spermatogenèse. Il a également été démontré que six hommes ayant eu un enfant porteur de la mutation c.755C<G n'avaient pas d'élévation du nombre de cellules germinales porteuses de la mutation comparée à l'ensemble d'hommes en santé. Ces trouvailles ont été confirmées par une étude portant sur la même mutation, mais utilisant une méthode différente¹¹⁹. Suite à ces résultats, le consensus émergent est que l'EAP, plutôt que de causer une hypermutabilité de certaines régions génomiques, serait plutôt responsable de l'expression de protéines mutantes dans les spermatogonies qui seraient associées avec une expansion clonale locale¹²⁰. En effet, il a été démontré que la majorité des mutations associées avec des troubles de l'EAP sont des gains de fonctions, ce qui amènerait les protéines à agir de manière à conférer un avantage sélectif aux spermatogonies. Ce processus est appelé « sélection égoïste ».

1.4 Problématique

Au moment où ce projet a été entrepris (Mai 2010), les ELG avaient été peu concluantes pour la SCZ. À l'instar, les GWAS, que plusieurs présentaient comme la réponse à tous les maux, avaient résulté en des associations marginales. La seule étoile au tableau était les études de SV qui avaient pu identifier plusieurs SV rares fortement associées à la maladie. L'arrivée en 2009 des nouvelles technologies de séquençage a été l'occasion parfaite pour réévaluer la stratégie utilisée pour les

études génétiques portant sur la SCZ. La question était alors de savoir s'il fallait poursuivre les GWAS ou bien s'en détourner et évaluer d'autres hypothèses.

Les données génétiques disponibles à ce moment rendaient plausible la théorie selon laquelle les variants communs (>5 %) jouaient probablement un rôle très faible dans le facteur héréditaire de la maladie. De ce fait, nous avons décidé de nous placer à l'opposé du spectre allélique et d'étudier les variations génétiques privées. Un précédent projet du laboratoire avait rapporté la présence de plusieurs *DNM* dans des gènes synaptiques^{95,96,121}. De plus, avec le séquençage à haut débit, l'équation régissant l'identification des variations génétiques a été complètement débalancée. En effet, les DNM, qui étaient très difficiles à identifier lors de projets de séquençage traditionnel, sont devenues les mutations les plus facilement identifiables, pour peu que l'on ait accès à l'ADN des parents de l'individu atteint.

Au moment où ce projet fut entrepris, très peu d'informations sur les DNM étaient disponibles de manière générale, et encore moins pour la SCZ. De ce fait, notre étude a été divisée en trois objectifs visant à mieux comprendre l'impact des DNM sur la SCZ.

Objectif #1 : Identifier des DNM chez des patients atteints de SCZ et caractériser l'ensemble de ces mutations

Objectif #2 : Appliquer les connaissances acquises lors d'une étude de DNM à une étude populationnelle en vérifiant si des gènes porteurs de DNM ont un plus grand fardeau de mutations rares chez des patients atteints de SCZ.

Objectif #3 : Tester si le taux de DNM est fixe pour l'espèce humaine. Dans le cas contraire, tenter de la corrélérer avec des facteurs externes, comme par exemple avec l'âge du père.

Chapitre 2 : Élévation du taux de mutations *de novo* exoniques dans des patients atteints de schizophrénie

2.1 Présentation de l'article

L'étude présentée ici a été entreprise suite à de nombreux précédents travaux réalisés dans le laboratoire de Dr Rouleau. En effet, en 2005, Dr Rouleau a reçu du financement pour mettre sur pied un grand projet de séquençage (Synapse2Disease, S2D). Le but était de séquencer mille gènes synaptiques chez des patients atteints de SCZ et ASD, avec une emphase particulière sur les mutations *de novo*. Bien que ce projet ait été entrepris il y a moins de dix ans, les nouvelles technologies des dernières années ont fait réaliser un bond de géant à la génétique. Ainsi, lors de l'entreprise du projet S2D, les limitations technologiques ont forcé l'utilisation d'une approche où chaque exon de chaque gène était séquençé séparément et où chaque amorce était dessinée individuellement dans le but de séquençer les fragments par séquençage Sanger.

Malgré ces limites technologiques, le projet S2D a généré de nombreuses retombées, dont l'identification de plusieurs gènes soupçonnés d'être liés à la SCZ^{95,96,122}, notamment le gène SHANK3. Cependant, le plus gros impact du projet S2D a certainement été de démontrer une élévation du taux de mutations néfastes dans les exons de gènes synaptiques, en comparaison aux séquences introniques⁹⁷. Cependant, à la suite de cette découverte, plusieurs questions demeuraient toujours

en suspens. Il n'était, à ce moment, pas possible de savoir si l'effet observé était spécifique aux gènes synaptiques, car très peu de gènes non synaptiques avaient été séquencés. De plus, il était difficile de comparer le taux de mutations observé avec les autres études sur les DNM, car le séquençage était ciblé.

Avec la disponibilité des nouvelles technologies, il a été décidé de stopper le projet S2D peu après avoir passé le cap des 400 gènes. En effet, il était devenu trop coûteux de poursuivre en utilisant les technologies traditionnelles. Nous avons donc décidé d'établir la deuxième phase du projet sur une plateforme de capture d'exomes. Cette décision nous a également forcé à abandonner le concept d'analyse populationnelle puisque le coût encore prohibitif de cette technique ne permet pas de l'appliquer à de grands échantillons.

Cependant, la quantité de séquences obtenue compense largement la diminution de la taille de l'échantillon. En effet, près de 300 Mb pourront être interrogés par la capture d'exomes de 14 individus, comparés à 45 Mb chez 150 individus pour le projet S2D. De ce fait, il sera plus facile d'établir des statistiques globales afin de savoir si le taux de mutations est différent du niveau attendu chez les patients atteints de SCZ.

2.1.1 Contribution des auteurs

Conception de l'étude : JG, LX, SLG, GAR

Recrutement des patients et informations cliniques : MOK, LX, BM, RJ, NJ

Capture de l'exome et séquençage : AR, IB, AHYT

Alignement des séquences et détection des variants : SLG, DS, JYJB, CHL

Validation des variants : AN, SZ, LJ, SLG, PT

Analyses bioinformatiques : SLG, ADL, DS, AHYT, JYJB, CHL, SL

Rédaction du manuscrit : SLG, PAD, MOK, SL, GAR

Increased exonic de novo mutation rate in probands affected with schizophrenia

Article publié dans *Nature Genetics* , Juillet 2011; volume 43(9), pages 860 à 863

Simon L. Girard¹, Julie Gauthier¹, Anne Noreau¹, Lan Xiong¹, Sirui Zhou¹, Loubna Jouan¹, Alexandre Dionne-Laporte¹, Dan Spiegelman¹, Pascale Thibodeau¹, Isabelle Bachand², Jessie Y.J. Bao³, Amy Hin Yan Tong³, Chi-Ho Lin³, Bruno Millet^{4,5}, Nematollah Jaafari^{4,6}, Ridha Joobber⁸, Patrick A. Dion¹, Si Lok³, Marie-Odile Krebs^{4,7}, Guy A. Rouleau^{1,2}.

1 Centre of Excellence in Neuromics of Université de Montréal, Centre Hospitalier de l'Université de Montréal Research Center, and the Department of Medicine, Université de Montréal, Montréal, QC H2L 2W5, Canada

2 Centre of Excellence in Neuromics of Université de Montréal, Centre Hospitalier Universitaire Sainte-Justine Research Center, Montréal, QC H3T 1C5, Canada

3 Genome Research Centre, The Li Ka Shing Faculty of Medicine, University of Hong Kong

4 Université Paris Descartes, Faculté de Médecine Paris Descartes, Paris, France; INSERM U894, Laboratoire de Physiopathologie des Maladies Psychiatriques, Paris, France;

5 Service Hospitalo-Universitaire de Psychiatrie, CH Guillaume Rognier, Rennes, France

6 CIC-P INSERM U 802; Unité de recherche clinique intersectorielle en psychiatrie Centre Hospitalier Henri Laborit ; Université de médecine et de pharmacie ; 86000 Poitiers, France

7 Service Hospitalo Universitaire, Centre Hospitalier Sainte-Anne, 7 rue Cabanis, 75014 Paris, France

8 Douglas Mental Health University Institute, Department of Psychiatry, McGill University, Montréal, QC H4H 1R3, Canada

2.2 Abstract (Introduction)

Schizophrenia (SCZ) is a severe psychiatric disorder that profoundly affects cognitive, behavioral and emotional processes. The wide spectrum of symptoms and clinical variability in SCZ suggest a complex genetic etiology, which is consistent with the numerous loci thus far identified by family linkage, copy number variation (CNV) and association studies^{65,66,84,86}. While SCZ heritability may be as high as ~80%, the genes responsible for much of this heritability remain to be identified³⁶. Using high-throughput sequencing technologies, we sequenced the exome of 14 patients with schizophrenia and their parents. We identified 15 *de novo* mutations (DNMs) in 8 probands, which is significantly more than expected considering the previously reported DNM rate^{97,104,123}. In addition, 4 of the 15 identified DNMs are nonsense mutations, which is more than expected by chance¹²⁴. Our study supports the notion that DNMs may account for a fraction of the heritability reported for SCZ while providing a list of genes possibly involved in disease pathogenesis.

2.3 Methods (Published online)

2.3.1 Recruitment of patients

All patients were recruited through a large clinical genetic SCZ group (MO Krebs). For the current study, we recruited a cohort of 14 schizophrenia patients and their parents (42 individuals in total). DNA was extracted from blood for all samples using the PureGene DNA kit. Paternity/maternity was confirmed using genetic micro satellite markers.

2.3.2 Clinical assessment procedures

All participants, after receiving a full description of the study, gave their informed, written consent. All study procedures were approved by the French ethics committees and were in accordance to the Helsinki declaration. We investigated unrelated French Caucasian adult patients from the Departments of Psychiatry at Sainte-Anne Hospital in Paris, Guillaume Regnier Hospital in Rennes and Henri Laborit Hospital in Poitiers, all taking part to the collaborative network ReFaPsy. All investigators were individually trained in the coordinating center, in order to ensure consensual procedures and ratings.

The patients fulfilled DSM-IV criteria for SCZ or schizoaffective disorders. Consensual diagnosis was reached by two clinicians based on the review of the direct structured interview using the Diagnosis Interview for Genetic Studies (DIGS version 3.0) conducted by trained psychiatrists and psychologists and on information from practitioners and family, and prospectively documented clinical case reports. Family histories of psychiatric disorders were collected using the Family Interview for Genetic Studies (FIGS). Additional specific questionnaires were added to collect information on obstetrical complication and early development. Exclusion criteria for all subjects included neurologic hard signs, a history of head trauma and substance dependence.

The patients included in the present study were selected from the initial cohort using the following criteria: sporadic cases (no family history of psychosis in the

first and second degree relatives, including bipolar disorders with psychotic features) and availability of sufficient amount of native DNA from blood sample.

2.3.3 Clinical characteristics

The total selected sample comprised 14 patients (7M/7F; age: $29\pm 11,6$). The subtype of SCZ were distributed between the undifferentiated subtype (n=7); disorganized subtype (n=6) and schizoaffective (n=1). The number of years of education was 12.9 ± 5.8 . The ages of the father and mother when the probands were born were 29.6 ± 9 and 27.5 ± 8.3 , respectively. Mean birth weight was in the normal range $3,383\pm 1575$ (range 2,750–4,000).

The age at onset of prodromes was 12.9 ± 6.6 , the age at first episode of psychosis was 19.3 ± 8.1 . Only two patients had a lower educational achievement (respectively 6 and 9 years). Six patients had slight language difficulties (including stuttered speech; dyslexia, delayed reading/learning) or delayed developmental milestones (language, walking, wetting). Nine patients had minor physical anomalies as screened using an adapted version of Waldrop.

2.3.4 Exome Capture and Sequencing

Targeted enrichment was performed with Agilent SureSelect All Exome Kit v.1, optimized for Illumina sequencing with 2 ug of genomic DNA. The version of this kit is designed to cover approximately 38 Mb of genomic sequences, mainly protein coding sequences. Exon-enriched DNA libraries from 42 individuals of 14 families were sequenced individually on a single lane of the Illumina Genome Analyzer IIx

platform (Illumina, San Diego, CA), producing 76-base pair end reads in accordance to the manufacturer's direction. Approximately 4 to 5 billion base calls were generated for each sample.

2.3.5 Read mapping

BWA¹²⁵ was used as the main aligner. The mapping was made against the human genome (hg19), indexed using the bwts algorithm included with BWA. The alignment was made using a maximum mismatch penalty of three. All other parameters from BWA were kept at the default value. The alignment was generated using a pair-end mode and SAMTools¹²⁶ was used to store the alignment. All PCR duplicates were removed from the alignments. A second alignment was made based on the default setting of quality filtering, high quality pair-end (76x2 bp) sequencing reads were obtained and mapped to the reference human genome assembly GRCh37 (NCBI Build 37.1) using CLC Genomics Workbench (CLC Bio, Aarhus, Denmark). The nine alternate assembly loci were excluded from the reference. Only uniquely mappable reads with at least 90% similarity with the reference were considered.

2.3.6 Assessment of enrichment

Target specificity was defined as the percentage of reads mapped on the target exon regions out of the number of reads that are uniquely mappable to the whole genome reference. Coverage of the mapped reads to the Exon targets was tabulated by read depth at each exon nucleotide position and by the mean fold coverage within each gene.

2.3.7 Variant calling and annotation

Varscan¹²⁷ and CLC Genomic Workbench were used for variant detection. For a practical reason, the parameters for variant calling were chosen by comparison with NCBI dbSNP Build 131. The minimum coverage threshold was chosen such that increasing the coverage could not improve the proportion of SNVs found in dbSNP 131 significantly. The DNM calls relied on the premise that the parameters for allele calls in parents should be less stringent than that for variant calls in probands due to the non-uniform coverage between individuals. For Varscan, the required parameters for DNM calls to meet the threshold were for the proband: (1) A minimum coverage of 20; (2) Mutation Frequency (defined as the percentage of reads supporting the novel allele) > 20% with at least three reads supporting the mutation; and (3) A minimal PHRED-score quality of 25. For the parents, the criteria were: (1) A minimum coverage of 5; (2) Mutation frequency > 10% with at least two reads supporting the mutation; and (3) A minimal PHRED score quality of 15. For the CLC Genomic Workbench software, the coverage threshold (15,15) respectively for variation calls in probands and allele calls in parents was used to detect DNMs. Annotation was made using ANNOVAR¹²⁸ against database RefSeq v46 and DBSNP v131.

2.3.8 Variant filtering and segregation analysis

Each variant from any proband was compared to his/her own parents, and a common pool of all other parents (26 individuals). Each unique variant was kept. A

further functional filter was applied to remove any non protein-disruptive variant (intronic, intergenic, UTR,).

2.3.9 Validation of de novo mutations

Primers were designed by Primer Select from DNASTar or Exon Primer to allow specific amplification of regions containing variants to validate. Polymerase chain reactions (PCRs) were performed using the AmpliTaq Gold DNA Polymerase (Applied Biosystems, Foster City, California, USA) as per manufacturer's instruction. To visualize DNA fragments, a small fraction of the PCR product will be loaded on a 1.5 % agarose gel containing ethidium bromide for the staining. PCR products were sequenced at the Genome Quebec Innovation Centre (Montréal, Québec, Canada) using a 3730XL DNAnalyzer (Applied Biosystems, Foster City, California, USA) and Mutation surveyor (v.3.10, SoftGenetics, State College, Pennsylvania) was used for mutation detection analysis.

2.4 Results and discussion

We sequenced the captured product of targeted exomes prepared from 14 trios, each trio consisting of a proband affected with SCZ and his/her parents; In order to avoid identifying genetic variations that result from cell culture artifacts, blood DNA was used. Each proband was separately interviewed in person and standardized criteria for diagnosing SCZ (**DSM-IV**) were evaluated during these examinations (See method section). In order to focus on non-familial SCZ cases, none of the selected probands had a first or second-degree family history of psychotic disorders

(including bipolar disorder with psychotic features), and neither probands nor parents had a history of substance dependence. To exclude potentially causative structural genetic variations, genomic DNA from every selected trio was examined using a CNV targeted array (Cytochip 2.0, Affymetrix Inc.); no previously reported *de novo* or other potentially causative CNVs were found.

The exome capture of every individual was performed using SureSelect Human All Exome Kits v.1 (Agilent Technologies Inc.) and the sequencing was done on a Genome Analyzer IIx (Illumina Inc.) (Table 2.1). All sequence reads were mapped to the reference genome using two different mapping algorithms: Burrows-Wheeler Aligner Tool (BWA) (Li, 2009) and CLC Genomics Workbench (CLC Bio, Aarhus, Denmark). An average of approximately 56 million reads was available for each individual, of which approximately 50% aligned to the targeted regions (Table 2.2). The exome capture was relatively effective with an average of ~72% of targeted regions covered with a read depth > 20x (Figure 2.1). For each of the 14 SCZ affected probands sequenced, any variation from the reference genome was examined in his/her unaffected parents, as well as a pool of 26 control individuals (constituted from the parents of the other trios), and every variant found to be unique to the proband was placed on a putative DNM list. Unique variants meeting the quality thresholds were kept for subsequent validation. All mutations within the coding sequences were considered for this study and a total of 73 putative non-synonymous variants were identified (Table 2.3); a direct PCR amplification and Sanger sequencing validation step in concomitant probands and their respective parents, was made for each one of the 73 variants identified. As a result a total of 15

SNVs were validated as genuine DNMs while the remaining 58 variants were found to be either false positives or actually transmitted by one of the two parents. None of those DNMs were reported in dbSNP build 131 or in the latest release of the 1000 genome project. No insertion or deletion mutations were found during this analysis.

This exome captured high-throughput sequencing tandem approach directly identified 15 DNMs in 8 of the 14 SCZ probands (Table 2.4); no DNM were detected in the remaining 6 trios. As we do not cover the totality of the exons, nor are we looking at non-coding sequences or splice junctions, we cannot exclude the possibility that there might be undetected potentially functional DNM in these trios. The DNMs identified can be categorized as either transitions (11) or transversions (4). The genes containing DNMs were searched against the SZgene database¹²⁹ and none of these were previously reported to be associated with SCZ. A predictable observation as DNMs are rare events that cannot be detected by genome wide association (GWA) studies that rely on a methodology better suited for the identification of common variants. Out of the 15 validated DNMs, 4 may warrant particular interest as they are nonsense (NS) that are predicted to lead to a premature stop codon while the rest are missense (MS). Interestingly a 4/15 ratio of NS to MS mutation is significantly higher from the expected ratio of 1/20, as calculated by Kryukov¹²⁴ et al. p-value = 0.005467 using a binomial test, CI 95% = 0.077 – 0.55); furthermore amongst all mutations reported to cause Mendelian diseases (reported in the Human Gene Mutation Database HGMD), the ratio of NS versus MS mutations is roughly 1/4, which is not significantly different from the 4/15 ratio observed in our study (p-value > 0.05).

The first NS mutation, observed in individual SCZ0401, produces a stop codon at R480X that truncates the last 55 amino acids (a.a.) from the ZNF480 protein. A known polymorphism (rs113675780) that creates a stop codon at position R474X has been reported in *ZNF480*, though no population frequency data are available for that SNP. A second NS DNM leading to a premature stop codon, also observed in individual SCZ0401, introduces a stop codon at E448X of the karyopherin alpha 1 (KPNA1) protein and leads to the loss of the last 58 a.a.; no NS mutation has been reported in SNP databases for this gene. Interestingly, KPNA1 regulates and mediates V(D)J recombination through recombination activating gene (RAG) 1 and 2¹³⁰. V(D)J recombination is an important immune system process that generates immunoglobulins and T-cell receptors¹³¹. Given an emerging hypothesis suggests autoimmunity may play a role in SCZ⁶⁵⁻⁶⁷, further study of *KPNA1* in SCZ may be warranted. It is noteworthy that two NS DNMs were found within the same individual, SCZ0401. We may speculate that only one of the two NS mutations may be pathogenic with respect to SCZ, while the other may or may not have a clinical effect. At this time, we don't have enough information to make this distinction. The third NS DNM leading to a premature stop codon, in individual SCZ0201, creates a stop at position Y2200X of the pro low-density lipoprotein receptor-related protein 1 (LRP1). LRP1 is a very large protein of 4,544 a.a. and therefore this DNM truncates the protein to half of its normal size. Interestingly, LRP1 is known to be regulated by the amyloid precursor protein (APP)¹³² and is believed to play a functional role in the pathogenesis of Alzheimer's disease^{132,133}; which is consistent with the notion that it may also play a role in other neuropsychiatric disorders. Three NS mutations

were previously reported in SNP databases (rs75873762, rs79339212 and rs113087094 that create a stop at position E177X, E2105X and Y2428X respectively) for this particular gene but again no population frequencies are yet available for these variants. The last NS DNM leading to a premature stop codon affects the C-terminal portion of the ALS2-like (ALS2CL) protein, at position R733X, and it is predicted to result in the loss of the last 190 a.a. encoded by the longest isoform. ALS2CL is a putative modulator of alsin (ALS2) which is involved in a juvenile form of amyotrophic lateral sclerosis¹³⁴. No NS mutation has been reported for *ALS2CL* in SNP databases. Amongst the 11 remaining missense DNMs identified, 7 are predicted to be deleterious when analyzed using the bioinformatic softwares SIFT¹³⁵, Polyphen¹³⁶, PolyP conservation score¹³⁷ or Grantham Matrix^{138,139}. Interestingly, two missense variants for the genes *CCDC137* and *CHD4* are predicted to be damaging by the four prediction softwares. However until the functional consequences of these 11 missenses variants are investigated, it remains difficult to further predict which actually underlie SCZ pathogenesis.

The data generated by the 1000 genome project¹⁰⁴ has established the normal DNM rate to be $\sim 1.1 \times 10^{-8}$ bp for a human haploid genome, which is consistent with previous estimates^{97,123,140}. With this normal DNM rate in mind, we sought to test whether the number of exonic DNMs identified using 14 SCZ probands was in fact higher than what should be observed. Therefore we calculated what should have been the number of DNMs observed considering both the number of probands and the amount of DNA sequenced from our patients. Using the SureSelect Human All Exome Kit v.1 probes overlapping with regions from the Consensus Coding

Sequences (CCDS), we determined the number of sequence reads that were on target for each of the 14 SCZ probands (Table 2.5). The total number of coding base pairs screened was estimated to be 289.62 Mbp, and with this sequence coverage the number of DNMs identified in our study should have been ~ 6.37 . However, we report here 15 DNMs which actually represents a DNM rate of 2.59×10^{-8} rather than $\sim 1.1 \times 10^{-8}$ bp for a human haploid genome; a binomial test indicates that the number of DNMs observed in our study differs significantly (CI 95% = 1.4493×10^{-8} – 4.2710×10^{-8}) from what should have been observed, considering different studies that have measured the de novo mutation rate in humans (Table 2.6). The DNM rate we observed is specific to an analysis which only takes into consideration exonic coding sequences; on a genome-wide scale, where the vast majority of the DNA is not constrained, the DNM rate would likely be similar in each individual. Our study strongly suggests that the enrichment of DNMs, within the coding sequence of SCZ patients, likely underlies the pathogenesis of a substantial number of SCZ cases. It is important to stress that our observed DNM rate is conservative. Given the huge amount of sequence information, our *de novo* identification algorithms may have missed some DNMs in certain samples. Any missed DNMs would only increase the significance of our results. An analogous study was recently reported for mental retardation (MR) by the group of Veltman which supported the hypothesis that DNMs explain a significant fraction of this diseases heritability¹⁴¹. This group identified 9 DNMs in 7 MR trios, which is similar to our present findings for SCZ (15 DNMs in 8 trios). Interestingly our approach was very similar to theirs and we used the same capture kit. The contribution of DNMs to MR has been recurrently

hypothesized over the years, yet it is only now with the advent of powerful sequencing methods that the hypothesis could be directly tested and confirmed¹⁴². It is somewhat surprising that we find similar DNM mutation rates in SCZ and MR, as this mechanism was well established for MR, but is only recently being explored in SCZ. In a previous smaller scale study screening synaptic genes by direct sequencing, our group reported that the number of functional DNMs appeared to be increased in SCZ patients⁹⁷; a result consistent with the present report where a more powerful approach with no inherent bias for the genes to be screened was used. In fact when re-analyzing data from the previous synaptic gene report, we find that the observed SCZ DNM rate (6 functional DNMs over 91 Mbp) is in agreement with the one reported here (p-value > 0.05). (Awadalla, 2010). In conclusion, we confirmed our initial hypothesis that the deleterious DNM rate is indeed higher in SCZ patients, and we also enlarged this observation to encompass the exonic SCZ DNM rate. Our sequencing of trios where the proband was affected with sporadic SCZ indicates that DNMs likely contribute to the development of SCZ, an observation which may explain part of schizophrenia's missing genetic causes. Those results are also in accordance with the original twin studies used to estimate the heritability attributed to the disease. Furthermore our results provide a list of new candidate genes to be screened in SCZ cohorts of various origins as those mutations are sporadic events independent of Linkage Disequilibrium (LD). Although sporadic cases are not necessarily representative of all patients with SCZ, these new candidates should be further investigated as potentially implicated in SCZ.

2.5 Acknowledgements

The patients were selected from the Psydev study, promoted by Inserm, through the Collaborative Network for Family Study in Psychiatry ("Réseau d'étude familiale en Psychiatrie", REFAPSY), supported by the Fondation Pierre Deniker. We wish to thank Souhail Bannour, Marie-Josée Dos Santos, Mohamed Ali Gorsane, Narjes Benjema, Mélanie Chayet, Sophie Leroy, Faycal Mouaffak, Katia Ossian, Françoise Polides for their participation to the recruitment of the patients and their technical help. This work was supported by Genome Canada and Génome Québec, and received cofunding from Université de Montréal for the Synapse to Disease (S2D) project as well as funding from the Canadian Foundation for Innovation. G.A.R. holds the Canada Research Chair in Genetics of the Nervous System.

2.6 Tables

2.6.1 Table 2.1: Statistics of sequencing and mapping for 42 individuals from 14 trios

Paper ID	Status	Total # of reads	Uniquely mapping reads	On target reads
SCZ0101	Proband	62 528 624	48 092 033 (76,91 %)	30 112 437 (48,16 %)
SCZ0102	Father	64 910 532	48 689 086 (75,01 %)	30 642 110 (47,21 %)
SCZ0103	Mother	61 210 814	45 819 282 (74,85 %)	29 230 302 (47,75 %)
SCZ0201	Proband	66 999 184	52 522 702 (78,39 %)	32 319 350 (48,24 %)
SCZ0202	Father	62 840 990	47 522 849 (75,62 %)	31 660 546 (50,38 %)
SCZ0203	Mother	64 862 226	49 356 003 (76,09 %)	27 934 741 (43,07 %)
SCZ0401	Proband	37 854 748	27 837 257 (73,54 %)	15 094 485 (39,87 %)
SCZ0402	Father	38 183 492	28 692 912 (75,14 %)	15 340 575 (40,18 %)
SCZ0403	Mother	36 353 584	27 224 321 (74,89 %)	15 043 799 (41,38 %)
SCZ0501	Proband	36 893 324	28 755 327 (77,94 %)	16 212 730 (43,94 %)
SCZ0502	Father	38 374 446	29 843 471 (77,77 %)	18 499 793 (48,21 %)
SCZ0503	Mother	37 266 844	29 030 198 (77,90 %)	17 767 387 (47,68 %)
SCZ0601	Proband	38 746 472	30 011 721 (77,46 %)	18 542 761 (47,86 %)
SCZ0602	Father	38 917 488	30 672 191 (78,81 %)	19 379 474 (49,80 %)
SCZ0603	Mother	36 722 572	28 325 706 (77,13 %)	17 685 078 (48,16 %)
SCZ0701	Proband	62 914 980	47 677 500 (75,78 %)	29 115 302 (46,28 %)
SCZ0702	Father	60 172 004	41 837 949 (69,53 %)	25 620 239 (42,58 %)
SCZ0703	Mother	63 591 168	46 705 307 (73,45 %)	29 034 357 (45,66 %)
SCZ0801	Proband	62 158 152	44 390 856 (71,42 %)	28 873 398 (46,45 %)
SCZ0802	Father	62 409 612	45 334 621 (72,64 %)	29 041 455 (46,53 %)
SCZ0803	Mother	60 994 836	43 893 100 (71,96 %)	28 280 642 (46,37 %)
SCZ0901	Proband	58 824 888	44 034 358 (74,86 %)	31 108 100 (52,88 %)
SCZ0902	Father	37 189 437	27 132 103 (72,96 %)	18 971 714 (51,01 %)
SCZ0903	Mother	61 080 136	46 541 238 (76,20 %)	31 622 214 (51,77 %)

Paper ID	Status	Total # of reads	Uniquely mapping reads	On target reads
SCZ1001	Proband	59 247 208	43 752 086 (73,85 %)	29 232 916 (49,34 %)
SCZ1002	Father	55 287 882	41 164 622 (74,46 %)	26 604 485 (48,12 %)
SCZ1003	Mother	52 324 684	38 420 573 (73,43 %)	23 462 464 (44,84 %)
SCZ1101	Proband	64 043 762	45 874 477 (71,63 %)	27 950 730 (43,64 %)
SCZ1102	Father	64 184 762	45 069 516 (70,22 %)	28 146 162 (43,85 %)
SCZ1103	Mother	67 360 678	47 742 015 (70,88 %)	30 423 274 (45,16 %)
SCZ1201	Proband	66 112 476	47 241 908 (71,46 %)	30 467 383 (46,08 %)
SCZ1202	Father	61 280 536	43 541 501 (71,05 %)	27 521 448 (44,91 %)
SCZ1203	Mother	63 493 440	44 861 602 (70,66 %)	28 700 361 (45,20 %)
SCZ1301	Proband	61 292 312	43 794 601 (71,45 %)	27 262 258 (44,48 %)
SCZ1302	Father	71 164 786	49 855 649 (70,06 %)	28 866 286 (40,56 %)
SCZ1303	Mother	71 352 186	51 423 364 (72,07 %)	29 987 500 (42,03 %)
SCZ1401	Proband	72 070 230	52 588 983 (72,97 %)	36 263 159 (50,32 %)
SCZ1402	Father	69 159 648	50 070 427 (72,40 %)	34 105 942 (49,31 %)
SCZ1403	Mother	72 081 330	51 642 587 (71,64 %)	35 075 916 (48,66 %)
SCZ1501	Proband	64 559 638	43 967 239 (68,10 %)	29 693 589 (45,99 %)
SCZ1502	Father	72 840 518	51 674 663 (70,94 %)	35 561 530 (48,82 %)
SCZ1503	Mother	66 428 102	45 432 822 (68,39 %)	30 185 146 (45,44 %)
	Average	56 341 626	41 309 972 (74 %)	25 980 265 (46 %)

*Individuals in bold are proband (affected individuals)

On target reads are defined as the reads overlapped with the target exon regions by at least 1 bp.

Percentage is calculated as bases on of total number of unique reads.

2.6.2 Table 2.2 : Summary of capture efficiency for the selected trios

Paper ID	Status	On Target Reads	Targeted region with coverage \geq 1x	Targeted region with coverage \geq 20x
SCZ0101	Proband	30 112 437	38491652 (98,9 %)	30212865 (77,6 %)
SCZ0102	Father	30 642 110	38474029 (98,8 %)	30324371 (77,9 %)
SCZ0103	Mother	29 230 302	38464746 (98,8 %)	30451835 (78,2 %)
SCZ0201	Proband	32 319 350	38263951 (98,3 %)	30188999 (77,6 %)
SCZ0202	Father	31 660 546	38595142 (99,2 %)	30864491 (79,3 %)
SCZ0203	Mother	27 934 741	38395319 (98,6 %)	29746486 (76,4 %)
SCZ0401	Proband	15 094 485	38312805 (98,4 %)	21925810 (56,3 %)
SCZ0402	Father	15 340 575	38242803 (98,2 %)	22144394 (56,9 %)
SCZ0403	Mother	15 043 799	38254029 (98,3 %)	22095238 (56,8 %)
SCZ0501	Proband	16 212 730	38057328 (97,8 %)	22935536 (58,9 %)
SCZ0502	Father	18 499 793	38264949 (98,3 %)	24946916 (64,1 %)
SCZ0503	Mother	17 767 387	38282783 (98,3 %)	24324596 (62,5 %)
SCZ0601	Proband	18 542 761	38351485 (98,5 %)	24921486 (64,0 %)
SCZ0602	Father	19 379 474	38302319 (98,4 %)	25428424 (65,3 %)
SCZ0603	Mother	17 685 078	38273372 (98,3 %)	24572509 (63,1 %)
SCZ0701	Proband	29 115 302	38428672 (98,7 %)	29653304 (76,2 %)
SCZ0702	Father	25 620 239	38459318 (98,8 %)	29528233 (75,9 %)
SCZ0703	Mother	29 034 357	38316794 (98,4 %)	29829709 (76,6 %)
SCZ0801	Proband	28 873 398	38424868 (98,7 %)	30380808 (78,0 %)
SCZ0802	Father	29 041 455	38479837 (98,9 %)	30131965 (77,4 %)
SCZ0803	Mother	28 280 642	38354914 (98,5 %)	29960107 (77,0 %)
SCZ0901	Proband	31 108 100	38532411 (99,0 %)	30493611 (78,3 %)
SCZ0902	Father	18 971 714	38546894 (99,0 %)	30259300 (77,7 %)
SCZ0903	Mother	31 622 214	38468589 (98,8 %)	30622557 (78,7 %)
SCZ1001	Proband	29 232 916	38545004 (99,0 %)	30062308 (77,2 %)
SCZ1002	Father	26 604 485	38453751 (98,8 %)	28881831 (74,2 %)

Paper ID	Status	On Target Reads	Targeted region with coverage\geq 1x	Targeted region with coverage\geq 20x
SCZ1003	Mother	23 462 464	38393191 (98,6 %)	28070858 (72,1 %)
SCZ1101	Proband	27 950 730	38305188 (98,4 %)	28865994 (74,2 %)
SCZ1102	Father	28 146 162	38445368 (98,8 %)	29476932 (75,7 %)
SCZ1103	Mother	30 423 274	38257343 (98,3 %)	29616705 (76,1 %)
SCZ1201	Proband	30 467 383	38054763 (97,8 %)	26345871 (67,7 %)
SCZ1202	Father	27 521 448	38440149 (98,8 %)	29106896 (74,8 %)
SCZ1203	Mother	28 700 361	38219198 (98,2 %)	29176147 (75,0 %)
SCZ1301	Proband	27 262 258	38394986 (98,6 %)	28929937 (74,3 %)
SCZ1302	Father	28 866 286	38231025 (98,2 %)	28909405 (74,3 %)
SCZ1303	Mother	29 987 500	38267603 (98,3 %)	29695103 (76,3 %)
SCZ1401	Proband	36 263 159	38440272 (98,8 %)	31312670 (80,4 %)
SCZ1402	Father	34 105 942	38540927 (99,0 %)	31500749 (80,9 %)
SCZ1403	Mother	35 075 916	38199432 (98,1 %)	30538365 (78,5 %)
SCZ1501	Proband	29 693 589	36987630 (95,0 %)	24488782 (62,9 %)
SCZ1502	Father	35 561 530	38442057 (98,8 %)	31291967 (80,4 %)
SCZ1503	Mother	30 185 146	38502310 (98,9 %)	31011166 (79,7 %)
Average		25 980 265	38 328 007 (98 %)	25 980 265 (72 %)

*Individuals in bold are proband (affected individuals)

2.6.3 Table 2.3 : Summary of de novo mutations found during high-throughput sequencing

Type	Chr	BP (hg19)	Reference Allele	Mutation Allele	Amino Acid Substitution	Variation Type	Coverage	MutFreq	MutCov	Gene	Proband
SNV	1	183617533	A	C	-	Missense	150	33 %	50	<i>APOBEC4</i>	SCZ010 1
SNV	2	217528675	C	A	L276M	Missense	20	20 %	4	<i>IGFBP2</i>	SCZ010 1
SNV	13	31506870	C	T	P6P	Silent	21	8	2	<i>C13orf26</i>	SCZ010 1
SNV	15	58001403	C	T	A202V	Missense	86	26 %	22	<i>GRINL1A</i>	SCZ010 1
SNV	19	36673714	T	C	H385R	Missense	103	44 %	45	ZNF565	SCZ010 1
SNV	19	40382516	T	C	Q3457R	Missense	20	7	0.35	<i>FCGBP</i>	SCZ010 1
SNV	21	16338349	T	G	K722T	Missense	60	33 %	20	NRIP1	SCZ010 1
SNV	21	46047779	G	A	V231M	Missense	80	50 %	40	<i>KRTAP10-9</i>	SCZ010 1
SNV	1	19612741	T	C	M114V	Missense	20	20 %	4	<i>AKR7A3</i>	SCZ020 1
Indel	10	18087630	T	del-1BP	-	frame-shift	23	22 %	5	<i>FAM23A</i>	SCZ020 1
SNV	12	57579450	C	A	Y2200X	Nonsense	23	52 %	12	LRP1	SCZ020 1
SNV	17	59445778	C	T	T839M	Missense	21	9	1	<i>BCAS3</i>	SCZ020 1
SNV	17	79637360	A	G	Y125C	Missense	21	24 %	5	CCDC137	SCZ020 1
SNV	3	12214647	C	A	E448R	Nonsense	48	42 %	20	KPNA1	SCZ040

		2										1
SNV	14	24621046	G	T	A659S	Missense	44	45 %	20	<i>RNF31</i>	SCZ040	1
SNV	17	57775308	A	T	L11R	Nonsense	20	20 %	4	<i>PTRH2</i>	SCZ040	1
SNV	19	52826001	C	T	R480R	Nonsense	38	37 %	14	<i>ZNF480</i>	SCZ040	1
Indel	20	1559024	C	del-1BP	-	frame-shift	24	25 %	6	<i>SIRPB1</i>	SCZ040	1
SNV	21	45511917	C	T	T995M	Missense	129	37 %	48	<i>TRAPPC10</i>	SCZ040	1
SNV	3	25636012	C	A	D331E	Missense	20	20 %	4	<i>RARB</i>	SCZ050	1
SNV	5	140503451	G	T	R624L	Missense	20	20 %	4	<i>PCDHB4</i>	SCZ050	1
SNV	19	17370460	G	A	Q284X	Nonsense	14	21 %	3	<i>USHBP1</i>	SCZ050	1
SNV	12	57591432	C	T	N3089N	Silent	51	34	0.666666	<i>LRP1</i>	SCZ060	1
SNV	14	10380612	A	G	I351V	Missense	26	46 %	12	<i>EIF5</i>	SCZ060	1
SNV	14	103806120	A	G	I351V	Missense	54	48 %	26	<i>EIF5</i>	SCZ060	1
SNV	15	69003978	T	A	L194Q	Missense	24	21 %	5	<i>CORO2B</i>	SCZ060	1
SNV	19	19654978	C	A	P542T	Missense	20	20 %	4	<i>CILP2</i>	SCZ060	1
SNV	20	40050548	T	C	E1576G	Missense	97	25 %	24	<i>CHD6</i>	SCZ060	1
SNV	19	42795556	C	A	A879D	Missense	120	40 %	48	<i>CIC</i>	SCZ070	1

SNV	19	45397294	T	C	L205P	Missense	20	20 %	4	<i>TOMM40</i>	SCZ070 1
SNV	23	153216416	G	T	L1851M	Missense	20	20 %	4	<i>HCFC1</i>	SCZ070 1
SNV	5	3599732	G	T	D224Y	Missense	20	20 %	4	<i>IRX1</i>	SCZ080 1
SNV	3	46717166	G	A	R733X	Nonsense	77	100 %	77	<i>ALS2CL</i>	SCZ090 1
SNV	3	48603977	C	A	G2775V	Missense	78	100 %	78	<i>COL7A1</i>	SCZ090 1
SNV	7	72436540	T	C	K50R	Missense	29	21 %	6	<i>TRIM74</i>	SCZ090 1
SNV	7	75028366	A	G	K50R	Missense	45	20 %	9	<i>TRIM73</i>	SCZ090 1
SNV	8	38853751	C	T	V70I	Missense	20	20 %	4	<i>TM2D2</i>	SCZ090 1
SNV	10	124753487	C	A	V357F	Missense	24	29 %	7	<i>IKZF5</i>	SCZ090 1
SNV	11	30033935	G	T	Y97X	Nonsense	38	32 %	12	<i>KCNA4</i>	SCZ090 1
SNV	14	24901202	C	A	C245X	Nonsense	20	20 %	4	<i>KIAA0323</i>	SCZ090 1
SNV	15	43585126	T	G	T71P	Missense	20	20 %	4	<i>TGM7</i>	SCZ090 1
SNV	16	27786400	G	T	G1482C	Missense	78	41 %	32	<i>KIAA0556</i>	SCZ090 1
SNV	16	57060554	G	T	G567C	Missense	19	6	9	<i>NLRC5</i>	SCZ090 1
SNV	22	37266475	G	T	V121F	Missense	283	46 %	129	<i>NCF4</i>	SCZ090 1
INDE	1	186276126	C	del-2BP	FS	FrameShift	20	20 %	4	<i>PRG4</i>	SCZ100

L												1
SNV	11	654009	T	G	T446P	Missense	20	20 %	4	<i>DEAF1</i>		SCZ100 1
SNV	12	6707226	G	A	R576W	Missense	45	47 %	21	<i>CHD4</i>		SCZ100 1
SNV	12	12188203	C	T	G748S	Missense	46	52 %	24	<i>KDM2B</i>		SCZ100 1
SNV	18	6974966	T	C	T2187A	Missense	31	35 %	11	<i>LAMA1</i>		SCZ100 1
SNV	23	111195414	C	A	E79X	Nonsense	15	20 %	3	<i>TRPC5</i>		SCZ100 1
SNV	1	42619162	C	A	A14D	Missense	20	20 %	4	<i>GUCA2B</i>		SCZ110 1
Indel	9	104152752	T	del-1BP	-	frame-shift	23	22 %	5	<i>MRPL50</i>		SCZ110 1
SNV	11	10482548	T	C	K39E	Missense	71	55 %	39	<i>CASP4</i>		SCZ110 1
SNV	15	30377047	G	T	S41I	Missense	39	26 %	10	<i>LOC65307</i>		SCZ120 1
SNV	19	15222443	G	C	G473A	Missense	105	49 %	51	<i>SYDE1</i>		SCZ120 1
SNV	20	43926879	C	T	G494S	Missense	20	20 %	4	<i>MATN4</i>		SCZ120 1
SNV	5	54281137	G	C	S70X	Nonsense	20	20 %	4	<i>ESM1</i>		SCZ130 1
SNV	7	882990	C	A	A164D	Missense	20	20 %	4	<i>UNC84A</i>		SCZ130 1
SNV	1	1153867	C	T	D295N	Missense	32	31 %	10	<i>SDF4</i>		SCZ140 1
SNV	1	236703868	C	T	A117V	Missense	26	9	0.34615 4	<i>LGALS8</i>		SCZ140 1

SNV	3	13846154	T	C	I160V	Missense	137	53 %	73	PIK3CB	SCZ140
		3							0.34328		1
SNV	10	94816786	T	A	L759M	Missense	67	23	4	EXOC6	SCZ140
INDEL	12	78540172	G	del-3BP	-	Frameshift	24	14	0.58333	NAV3	SCZ140
		12380498							0.52252		SCZ140
SNV	12	8	G	C	N886K	Missense	111	58	3	SBNO1	SCZ140
											1
SNV	12	133353576	G	T	Q1275K	Missense	20	20 %	4	GOLGA3	SCZ140
									0.41379		1
SNV	15	40328603	T	G	A114A	Silent	29	12	3	SRP14	SCZ140
											1
SNV	19	52570336	C	A	G151C	Missense	20	20 %	4	ZNF841	SCZ140
											1
SNV	8	42202526	G	C	A59P	Missense	20	20 %	4	POLB	SCZ150
											1
SNV	15	93543755	A	G	K1341R	Missense	70	51 %	36	CHD2	SCZ150
											1
SNV	20	35125267	C	A	S603X	Nonsense	20	20 %	4	DLGAP4	SCZ150
											1

* The variants listed in bold are the validated DNM (through Sanger Sequencing). All the rest were either false positive or inherited variants

2.6.4 Table 2.4 : Summary of DNM found using deep sequencing and validated using sanger sequencing

Individual	Chr	BP (hg19)	Gene	Reference Allele	Mutation Allele	Mutation	Nucleotide Substitution	Amino Acid Substitution	PhyloP*	Grantham Score**	Polypeptide	SIFT Prediction
SCZ0101	19	36673714	ZNF565	T	C	Missense	Transition	H385R	4.28987	Conservative	Possibly Damaging	DAMAGING
SCZ0101	21	16338349	NRIP1	T	G	Missense	Transversion	K722T	1.81401	Moderately Conservative	Probably Damaging	DAMAGING
SCZ0201	12	57579450	LRP1	C	A	Nonsense	Transversion	Y2200X	1.52978	-	-	-
SCZ0201	17	79637360	CCDC137	A	G	Missense	Transition	Y125C	3.55235	Radical	Probably Damaging	DAMAGING
SCZ0401	3	122146472	KPNA1	C	A	Nonsense	Transversion	E448X	5.78062	-	-	-
SCZ0401	19	52826001	ZNF480	C	T	Nonsense	Transition	R480X	-2.01835	-	-	-
SCZ0601	14	103806120	EIF5	A	G	Missense	Transition	I351V	2.10831	Conservative	Benign	TOLERATED
SCZ0901	3	46717166	ALS2CL	G	A	Nonsense	Transition	R733X	0.727551	-	-	-
SCZ1001	12	6707226	CHD4	G	A	Missense	Transition	R576W	2.53121	Moderately Radical	Probably Damaging	DAMAGING
SCZ1001	12	121882033	KDM2B	C	T	Missense	Transition	G748S	6.313	Moderately Conservative	Probably Damaging	DAMAGING
SCZ1001	18	6974966	LAMA1	T	C	Missense	Transition	T2187A	-0.76152	Moderately Conservative	Benign	TOLERATED
SCZ1101	11	104825480	CASP4	T	C	Missense	Transition	E39K	-1.1914	Moderately Conservative	Possibly Damaging	TOLERATED
SCZ1401	1	1153867	SDF4	C	T	Missense	Transition	D295N	1.88812	Conservative	Benign	DAMAGING
SCZ1401	3	138461543	PIK3CB	T	C	Missense	Transition	I160V	1.18761	Conservative	Benign	TOLERATED
SCZ1401	12	123804988	SBNO1	G	C	Missense	Transversion	N886K	-0.120827	Moderately Conservative	Probably Damaging	TOLERATED

* The PhyloP all the way 46 for vertebrates was used

**The Grantham matrix was used to evaluate the effect of each mutation and the Li et al. classification was used to classify the effect

2.6.5 Table 2.5 : Total of coding sequences (Mbp) with a coverage $\geq 20x$

Proband	Total coding sequences (Mbp)	Total DNM
SCZ0101	22 031 032	2
SCZ0201	22 116 414	2
SCZ0401	16 768 992	2
SCZ0501	17 244 815	0
SCZ0601	18 757 524	1
SCZ0701	21 755 421	0
SCZ0801	22 175 382	0
SCZ0901	22 355 863	1
SCZ1001	22 111 354	3
SCZ1101	21 085 529	1
SCZ1201	21 474 141	0
SCZ1301	21 334 975	0
SCZ1401	22 870 855	3
SCZ1501	17 542 536	0
Total	289 624 833	15

2.6.6 Table 2.6 : DNM rates comparison between SCZ and previously reported studies of normal individuals

Study	Observed Rate*	Current Study Rate	Binomial P-Value**	Comments
Roach et coll. Science 2010	1.10E-08	2.59E-08	0.002462	Genome Sequencing of a family quartet
Lynch et coll. PNAS 2010	1.28E-08	2.59E-08	0.0144	Survey of genes carrying disease causing mutations
Durbin et coll. Nature 2010 (CEU)	1.20E-08	2.59E-08	0.006336	Deep sequencing of a CEU trio
Durbin et coll. Nature 2010 (YRI)	1.00E-08	2.59E-08	0.0009975	Deep sequencing of a YRI trio
Awadalla et coll. AJHG 2010	1.36E-08	2.59E-08	0.01865	Estimation of neutral mutation rate using candidate sequencing

*The observed de novo mutation rate is expressed in mutation per position in a haploid genome

** The 95% Confidence Intervals (CI) for the current comparison are 1.45E-08 - 4.27E-08

2.7 Figure

2.7.1 Figure 2.1 : Exome Capture Efficiency

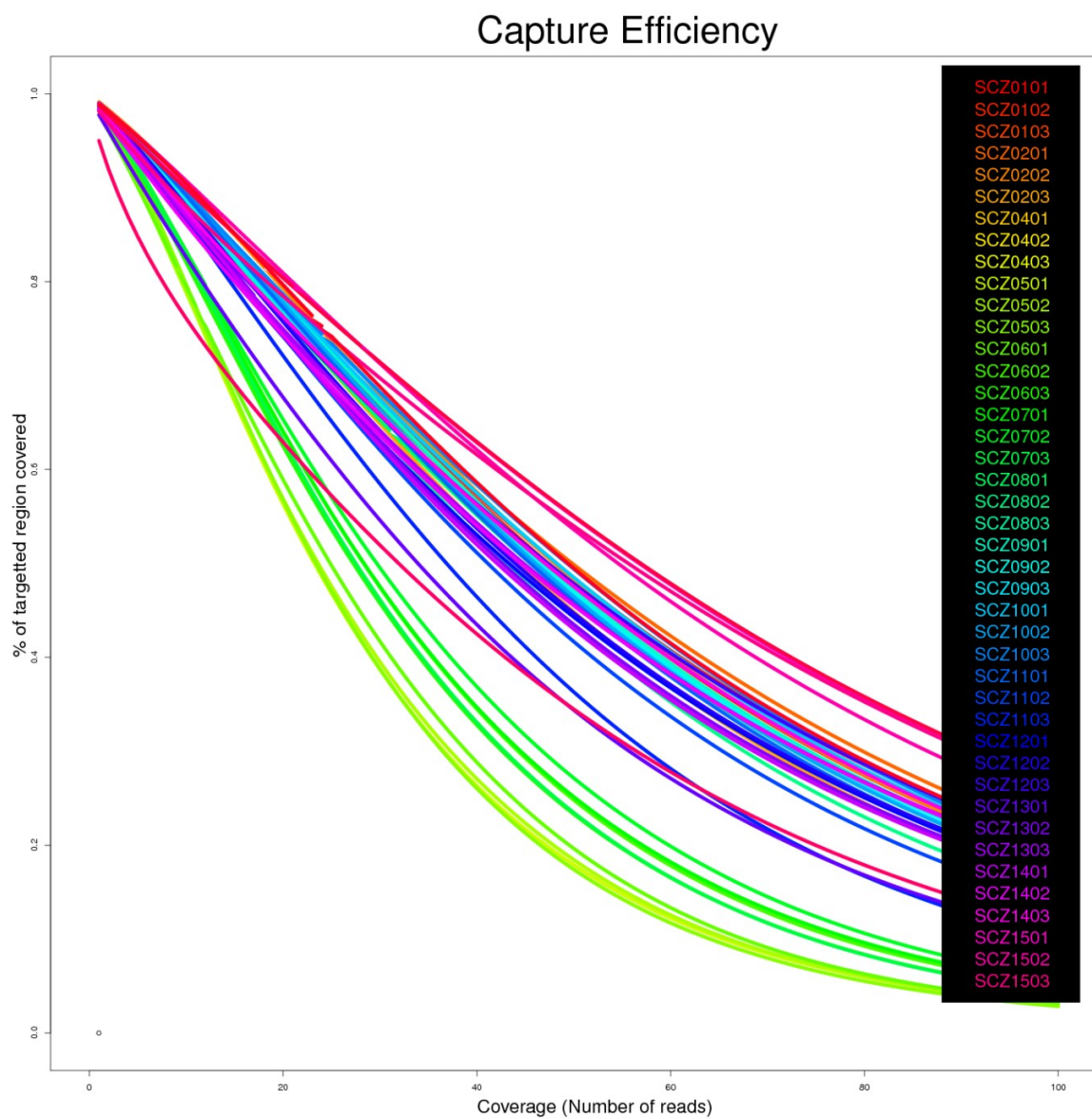


Figure 2.1 : Exome Capture Efficiency : The exome capture efficiency is presented for each individuals sequenced in this study. The x-axis presents the coverage in total number of reads and the y-axis is the percentage of the total targeted region, on a per-bases calculation.

Chapitre 3 : Le fardeau des mutations rares dans la schizophrénie

3.1 Présentation de l'article

À la suite de l'identification des gènes porteurs de DNM dans notre précédente étude, il est difficile de dresser un portrait réaliste des gènes directement liés à la SCZ. En effet, comme les DNM sont des évènements sporadiques qui se produisent naturellement chez tous les individus, il est difficile d'établir avec certitude si les mutations observées chez les patients SCZ sont bel et bien causatives.

Une approche logique serait de prioriser les gènes pour lesquels des mutations à effet tronquant ont été identifiées. C'est ce qui a été entrepris par notre équipe. Un total de 189 patients a été séquencé pour les gènes KPNA1, LRP1, ALS2CL et ZNF480¹⁴³. Cependant, aucun excès de mutations non-sens n'a été détecté pour aucun des 4 gènes. Ce fait en lui-même confirme l'hétérogénéité extrême de la SCZ et suggère qu'une multitude de gènes sont impliqués dans la pathogénèse de la maladie.

Afin de mieux comprendre et de pouvoir quantifier la complexité de ce mécanisme, nous avons décidé de procéder à un reséquençage complet des gènes porteurs de DNM chez des patients SCZ. En utilisant un reséquençage ciblé, nous avons dessiné une étude cas-contrôle (n=480) où il a été possible de tester, pour chacun des gènes, si le fardeau de mutations était plus élevé chez les patients atteints de SCZ.

3.1.1 Contribution des auteurs

Planification de l'étude: SLG, GAR

Échantillons et laboratoire: CB, SG, ML

Analyses bio-informatique: SLG, PLT, MPD

Analyses statistiques: AB, MPD, SLG

Rédaction du manuscrit: SLG, MPD, GAR

Mutation burden of rare variants in schizophrenia

Article soumis à *Molecular Psychiatry*

Simon L. Girard^{1,2,3}, Patrick A. Dion^{1,3,4}, Steve Geoffroy², Pamela Lachance-Touchette³, Cynthia Bourassa¹, Amina Bardhadi², Mathieu Langlois², Ridha Joober⁷, Marie-Odile Krebs^{5,6}, Marie-Pierre Dubé^{2,3}, Guy A. Rouleau¹

1 Montreal Neurological Institute and Hospital, Department of Neurology and Neurosurgery, McGill University

2 Centre de Pharmacogénomique Beaulieu-Saucier, Institut de Cardiologie de Montréal, Montreal, Quebec, Canada

3. Université de Montréal, Faculty of Medicine

4 Département de pathologie et biologie cellulaire, Université de Montréal, Montreal, Canada

5 Université Paris Descartes, Faculté de Médecine Paris Descartes, Paris, France; INSERM U894, Laboratoire de Physiopathologie des Maladies Psychiatriques, Paris, France;

6 Service Hospitalo Universitaire, Centre Hospitalier Sainte-Anne, 7 rue Cabanis, 75014 Paris, France

7 Douglas Mental Health University Institute, Department of Psychiatry, McGill University, Montréal, QC H4H 1R3, Canada

3.2 Abstract

Schizophrenia (SCZ) is a very heterogeneous disease that affects approximately 1% of the general population. Recently, the genetic complexity believed to underlie this condition was further highlighted by two independent studies that identified a total of 55 *de novo* mutations (DNM) in the same number of genes in different SCZ

probands^{98,100}. While these two reports did support the implication of DNM in the pathogenesis of SCZ, the absence of overlap in the genes identified strongly suggests that the number of genes involved in SCZ is likely very large; a notion that has been supported by the moderate success of Genome-Wide Association Studies (GWAS)⁷¹⁻⁷³. To further examine the genetic heterogeneity of this disease, we resequenced 62 genes that were found to have a DNM in SCZ patients, and 40 genes that are encoding proteins known to interact with the products of the genes with DNM in a cohort of 235 SCZ cases and 233 controls. Using a kernel association method, we were able to assess for association for different sets. Although our power of detection was limited, we observed an elevation of the mutation burden in the genes that have DNM.

3.3 Introduction

Schizophrenia (SCZ) is a highly prevalent neurodevelopmental disorder (1.1% of U.S. adult according to NIMH), that severely affects social and vocational development, and that has a strong negative stigmatization. Late adolescence and early adulthood is the peak period for the onset of SCZ (typically ~15-25 yrs)¹⁴⁴. Moreover, according to the World Health Organisation (WHO), nearly half of the patients with SCZ are not receiving appropriate healthcare. The contribution of genetics to SCZ has been widely examined and a recent meta-analysis of 12 twin studies established heritability to be ~81%³⁷. A simplistic view of the genetic architecture of SCZ suggests it involves different common alleles with low penetrance, intermediate frequency alleles with variable penetrance, and/or rare but highly penetrant alleles. More recently, a new hypothesis has emerged for SCZ: the implication of de novo mutations (DNM, i.e. mutations arising sporadically either

in the gamete cells of the parents or at the early stage of embryo development) as a source of rare penetrant variants. The development of high-throughput sequencing technologies facilitated the systematic genome-wide testing of this DNM based hypothesis. An innovative study by our group, using the exome sequencing data of SCZ trios (a proband plus his/her mother and father), found that patients with sporadic SCZ have a higher than expected exonic DNM rate, as well as a high nonsense/missense ratio very similar to pathogenic mendelian mutations¹²⁴. An independent study by Xu *et al.* produced similar results and thus replicating our findings¹⁰⁰. The same group later reported a new DNM study for which they exome-sequenced many trios with a proband affected with SCZ⁹⁹. The importance of DNM in neurodevelopmental diseases has been converging in other psychiatric disorders (e.g. autism¹⁴⁵⁻¹⁴⁷ and intellectual disability¹⁴⁸). However, the individual relevance of genes found to harbour a DNM is still unknown. Thus, we have decided to follow-up on those genes bearing DNM to establish if they have a role to play in the biological mechanism of SCZ.

3.4 Methods

3.4.1 Cohorts

All participants, after receiving a full description of the study, gave their informed, written consent. The CHUM ethics committees approved all study procedures in addition to ethic committees approval from the two recruitment site (French ethic committee and Douglas Hospital ethic committee). A total of 240 cases affected with SCZ were selected to constitute the case cohort. 5 cases were excluded because of their non-Caucasian ethnicity; the remainders of cases were all of European

ancestry. 143 cases were recruited in Canada, 7 in France, 62 in Hungary and 23 in the United States. There were 173 men and 62 women in the group of cases. The cohort was recruited to minimize the number of cases with substance abuse; only 18 cases (~7%) had a history of drug or alcohol abuse. At the time of ascertainment, the average age of the cases was 35.44 +- 10.09 and the average age of onset was 21.40 +- 4.70. The unaffected individuals were from a Caucasian population collected in Canada. In total, 125 controls were men and 115 were women. 7 individuals were excluded because of non-Caucasian ancestry, leaving a total of 233 controls.

3.4.2 Samples preparation ,quantification and digestion

DNA from blood and lymphoblastoid cell lines was extracted following standard protocols. DNA was quantified using the PICO green method and an ABI qPCR sequencer; quantity was adjusted to precisely 900 ng for a concentration of 20 ng/ul. Each DNA was subsequently digested, 8 samples at the time, using a mix of 8 different restriction enzymes, provided by the manufacturer (Agilent Technologies).

3.4.3 Genes selection

The list of genes to be resequenced were those for which de novo mutations were identified in SCZ probands in the Girard et al. and Xu et al. exome sequencing studies^{98,100} and the Awadalla et al. sequencing study⁹⁷, for a total of 62 genes. Furthermore, an additional set of genes encoding close interactors of the proteins encoded by genes from the core set was added. The genes encoding these interactors were selected based on protein:protein interactions listed in the Human Protein Reference Database (HPRD)¹⁴⁹. Node points proteins that showed two or

more interactions with a protein from the core set were all selected (Figure 3.1); in total 40 such genes were selected. In addition, 19 genes from the GABRA synaptic receptors were added. These genes were never genetically associated with SCZ and were used as a control group. (Table 3.1).

3.4.4 Design, Capture and enrichment

Haloplex probes were designed using the Haloplex design wizard tool (now replaced by the SureSelect design tool). In total, 2,041 regions were targeted for a total of 326.6 Kb. After masking of repeated regions and problematic high CG regions, a total of 316.93 Kb (97% of the target) was deemed suitable for Haloplex resequencing. The Haloplex baits covering these 2,041 regions were manufactured according to designs that were made at the Halogenomics headquarter (Uppsala, Sweden) and according to the Haloplex amplification procedures; this included the introduction of a biotin adjunct. For each individual recruited in the study, hybridization of the Haloplex probes was made with a pooled DNA substrate that was assembled following the eight restriction enzymes digestion step. The incubation of probes with the DNA was done overnight. Streptavidin coated magnetic beads and a magnetic stand were used to capture the biotin adjunct of the Haloplex probes to separate beads with the hybridized products from the non-hybridized DNA material. DNA ligase was subsequently added to the hybridized DNA eluted from the beads so that each of the targeted fragments would be enclosed within a circular DNA molecules. Two steps of Haloase treatment were performed in order to carry out the “Halo” PCR. At that stage, each DNA sample became assigned to a unique DNA barcode that is based on a 96-barcode index provided by Halogenomics. Seven DNA pools constituted from unique barcodes were created.

3.4.5 Sequencing

Each PCR pool was ligated to the Illumina standard adapter and loaded on a single lane from an Illumina HiSeq flow cell. Sequencing was done on an Illumina HiSeq 2000 using a paired-end mode that produced 100 bp reads. Once the sequencing was done, data from each lane was demultiplexed using the index of barcode and FASTQ files were generated for every individual.

3.4.6 Alignment, enrichment assessment and variant calling

A first alignment against the whole genome revealed the capture specificity to be >99% for the first ten samples analyzed. Thus, every alignment was made using a custom reference that comprised the targeted regions and an additional 200 flanking base pairs on both 3' and 5' ends. Alignments were performed using BWA v.0.5.9¹²⁵ before they were saved in a BAM file format. The DepthOfCoverage module from the GATK suite was used to assess enrichment efficiency based on the defined targeted regions¹⁵⁰. On average 302.47 Kb \pm 2.76 (95.43% of the effective Haloplex design) had sufficient coverage ($\geq 20\times$) for high quality variant calling. Four of the 468 samples had a significantly lower coverage (with sufficient coverage for 269.16 Kb, 269.41 Kb, 283.29 Kb and 286,71 Kb); these samples were nonetheless kept for the analysis because good variant calling was still possible for $\geq 85\%$ of the targeted regions. Variant calling was performed following the GATK Best Practice V2 and the GATK suite.

3.4.7 False negative and false positive assessment

Using sequencing data from a previous sequencing project from the gene *SHANK3* and other genes from the S2D project, we were able to evaluate the accuracy and

specificity of the sequencing. For a specific genomic region, the S2D project identified 138 variants using Sanger sequencing. 135 of those 138 variants were now correctly confirmed using the Haloplex sequencing dataset, hence the false negative rate was ~2.1%. Conversely, a false positive rate was also assessed based on the variants identified in the Haloplex sequencing dataset. Out of 25 variants that were identified in Haloplex sequencing dataset, Sanger sequencing during the S2D project had not identified 5. However, after revisiting the original Sanger sequencing data of these 5 variants, it was concluded that they were all present in the sequencing traces but the calling processes missed them. Thus, we can conclude that the false positive rate is <4%.

3.4.8 Kernel association testing

In order to test for difference in mutation burden, we used the Sequence Kernel Association Test (SKAT) algorithm¹⁵¹. SKAT is a statistical analysis package using a computationally efficient regression method that tests for associations between genetic variants in a region and a continuous or discontinuous trait. Variants were categorized in different sets. The first set was defined using the experimental origin of each gene (see table 3.2). The second set was defined by the genes encompassing all variants. The third set was defined using all individual exons that included at least one variant. SKAT offers different parameters to give a different weight according to variation frequencies. As our main focus is rare variations, we decided to use the manual recommended settings for rare variants ($B1=0.5$, $B2=0.5$). Those parameters set full weight to rare variants (<1%) while completely ignoring the other variations. Statistical analyses were performed using R statistical software v.2.15.0.

3.4.9 Sequenom haplotyping

Genotyping was performed in accordance with the iPLEX Gold protocol using matrix assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry (Sequenom). Assays were designed using the latest version of AssayDesign 3 with the default parameters for the iPLEX Gold chemistry. Cleaned extension products were conducted on a mass spectrometer and peaks were identified using SpectroTYPER.

3.5 Results

The resequencing effort was conducted on a cohort of 235 SCZ cases and 233 control individuals using the Haloplex-SureSelect method (Agilent Technologies) and a Illumina HiSeq 2000 apparatus. After recalibration and filtering, we identified 6,210 genetic variants ($TsTv = 2.485$) of which 5,892 were Single Nucleotide Variants (SNV) and 319 were indels. Using Sanger Sequencing, we estimated the false positive and false negative rate to be respectively $\sim 4\%$ and $\sim 2.1\%$. Most of the variants were private or shared by only two individuals, but a number of variants were intermediate or common (27%)(Figure 3-2). A higher number of rare variants was expected and is in accordance with recent findings from the 1K-genome project¹⁰⁵. We first sought to test if any individual variant showed a positive association independently of the gene set. For this, we used a Fisher test and a Bonferroni correction adjusted to the total number of variants. No variants reached the significance threshold (data not shown). This is likely explained by the lack of power of such an approach.

We set the SKAT parameters to account only for rare variants. The reason why we decided to focus only on rare genetic variation is that a GWAS recently

conducted for SCZ examined close to 10,000 individuals⁷³. It would be very surprising that our resequencing would pick associations from common or intermediate variants that would not have been detected in this recent GWAS. Also, an increase in rare variant burden is compatible with the elevation of DNM rate we previously demonstrated in SCZ.

The first mutation burden test was performed using study origins (Girard et al., Xu et al, S2D, Protein:Protein Interaction and control genes) as criteria for set definition (n=5). We did this in order to evaluate if genes harbouring DNM have a higher mutation burden than expected. (see table 3.2). The three experimental dataset (Girard et al. Xu et al and S2D) provided borderline associations ($P < 0.05$), but only the S2D dataset and the Girard et al. met the significance threshold once Bonferonni correction was applied ($P < 0.01$). This supports the notion that we can enrich for SCZ predisposing genes by identifying DNM in affected probands, especially when compared to the baseline profile (control genes). Interestingly, the dataset constituted of candidate genes found by looking at protein:protein interactions also shows a low, yet not significant p-value. Even though it is very early to draw any conclusion from this, it could mean that some genes encoding close interactors of the gene products found in DNM studies could also be involved in the disease and that interactome approach to identify candidate genes could be a valid method.

Next, we performed a second mutation burden test, this time treating each gene as a separate set (n=121) and each gene was independently followed, regardless of its experimental origin. The significance threshold was set to 4.1×10^{-4} according to Bonferroni correction. Only one gene reached this significance

threshold, the *strawberry notch homolog 1 (SBNO1)* gene; which was found to carry a DNM in our earlier exome study (See Figure 3.3). Of note, no sample was present in both studies. However, when considering all signals against the group of control genes, there seem to be a distinct profile that some genes might have a differential mutation burden profile but we lack the appropriate power to make a definite conclusion. Thus, we performed a third test, this time using single exons as separate sets. In order to correct for the number of independent exons, we used the simple method, which determine the number of independent tests require for an association study^{152,153}. Using this method, the significance threshold was set to p-value $< 1,08 \times 10^{-4}$. This time, multiple signals met the required threshold (See Figure 3.4). In addition to exons for *SBNO1*, we also found a different mutation burden profile in exons from the genes *EP300*, *MAPK14* and *SHANK3* (See table 3.3). It is not surprising to find the *SHANK3* gene as its implication in SCZ and Autism Spectrum Disorder has been shown many times^{95,121}. Interestingly, *EP300* encodes the p300 protein which plays a role in many tissues¹⁵⁴. It has also been shown that the loss of one copy of EP300 leads to an impairment of the development¹⁵⁵. But even more interesting is the profile difference between the group of control genes and the others. This could indeed suggest that many genes identified during DNM studies in SCZ have an important role in the genetic mechanism of the disease, but further confirmation is required.

An interesting case in the genes found to have a differential mutation burden profile is *SBNO1*. It is the only gene that reaches significance when the statistical test is conducted using sets of genes. When the exons are used as a set, multiple exons from *SBNO1* reach the significance threshold. After an in-depth look at the

individuals carrying variants in *SBNO1*, it was observed that the signal was driven by variants shared by a subset of individuals. To assess if this observation was due to an ancestral haplotype or a technical error, we genotyped all the variants that were driving the signals using a Sequenom platform on a new cohort of 249 SCZ, 256 controls ethnically matched controls. As SKAT is not robust when variants are in strong association as it is the case with a haplotype, we only used a Fisher test to assess if the mutational burden of all tested individuals was different between cases and controls. Out of a total of 22 haplotype variants, a great proportion of the tested individuals had none or only one variant, with the vast majority of individuals showing no more than five variants. Thus, all individuals that carried more than five variants were considered to have the SCZ haplotype. 54 individuals with the haplotype were found in both the SCZ and control cohorts, in agreement with the null hypothesis (p-value=0.91). There was also no difference in the distribution of the number of variants per individuals in the cases and in the controls (Figure 3-5) which led us to suspect a false positive association by the SKAT algorithm due to the strong LD between the variants.

3.6 Discussion

Population studies in SCZ have been widely popular in the last decades. From twin studies to genetic linkage and genome wide association studies, the general conclusion has always been that the genetic mechanism of SCZ is far more complex than previously estimated. Thus, we did not engage in this study expecting that we would identify determinants of a large proportion of the disease's heritability. However, our study is different from previous studies by two key points. The first is that by relying on candidate genes for DNM, we limit the multiple testing burdens to

genes of greater importance to the disease aetiology. There is always the risk that none of the selected genes is really involved in the disease, but our previous studies clearly show that genes carrying DNM are as good candidates as any other genes. Our study is also amongst the first population study to focus only on rare variants. The contribution of rare variants has always been hard to evaluate but many collapsing algorithm were developed to study the datasets generated by the recent advent of high-throughput sequencing methods.

We previously identified many genes harbouring a DNM in schizophrenia patients. However, we are still at the dawn of DNM studies and it remains a challenge to demonstrate the validity of the findings. In this study, we designed a resequencing assay for genes that were reported by us and others to present DNM in SCZ (Table 3.1). In addition, we also looked at the sequence of genes established to encode close interactors of the DNM genes' products. Also on the assay was a set of synaptic genes that were never genetically associated with SCZ. This last set was used as a baseline group to evaluate the pattern of association of other sets. In this study, we were able to identify some genes that present a mutation burden profile that is different between SCZ cases and controls. The genes include one gene that has been previously linked to SCZ but also with ASD (*SHANK3*) as well as two novel genes that were identified through a protein:protein interaction study (*EP300* and *MAPK14*). However, these two genes will need to be replicated in a larger cohort before we can draw a definite conclusion on their implication in SCZ. More importantly, we were able to show that genes found in DNM studies do have a differential mutation profile in SCZ patients, either on a group scale or individual scale. This has also been confirmed by comparing to a baseline group of control

genes consisting of synaptic genes that were not previously associated with SCZ. The use of sets of exons for association tests with rare variants might be a form of analysis worth considering in future studies. Indeed, it has been shown several times that mutations leading to disease can cluster in certain specific exons of a gene¹⁵⁶⁻¹⁵⁸. Thus, using the full gene as set definition could lead to a loss of power from the high proportion of variants from neutral exons.

In this study, we have demonstrated that genes identified in DNM studies likely play a role in the genetic aetiology of SCZ. But more importantly, our work is amongst the first to show the importance of rare variants in the genetic mechanism of the disease. Although the identified signals warrants replication in independent cohorts, those results call for a better integration of rare and private variants in future SCZ studies.

Acknowledgements

We thank the families involved in our study. We are thankful for the efforts of the members of the Genome Quebec Innovation Centre Sequencing and Bioinformatics groups. Guy A. Rouleau is grateful for the support received through his positions as Canada Research Chair in Genetics of the Nervous System and Jeanne-et-J.-Louis-Levesque Chair for the Genetics of Brain Diseases. Simon L. Girard is grateful to the FRQS for financial support.

3.7 Tables

3.7.1 Table 3.1 *List of genes included on the resequencing assay*

Gene ID	Data Source
ALS2CL	Girard et al.
CASP4	Girard et al.
CCDC137	Girard et al.
CHD4	Girard et al.
EIF5	Girard et al.
KDM2B	Girard et al.
KPNA1	Girard et al.
LAMA1	Girard et al.
LRP1	Girard et al.
NRIP1	Girard et al.
PIK3CB	Girard et al.
SBNO1	Girard et al.
SDF4	Girard et al.
ZNF480	Girard et al.
ZNF565	Girard et al.
ACOT6	Xu et al.
ADAMTS3	Xu et al.
ADCY7	Xu et al.
CCDC108	Xu et al.
CELF2	Xu et al.
COL3A1	Xu et al.
DGCR2	Xu et al.
DPYD	Xu et al.
EDEM2	Xu et al.
ESAM	Xu et al.
FAM3D	Xu et al.
FASTKD5	Xu et al.
GIF	Xu et al.
GPR115	Xu et al.
GPR153	Xu et al.
INPP5A	Xu et al.
KLF12	Xu et al.
LAMA2	Xu et al.
MAGEC1	Xu et al.
MTOR	Xu et al.
NPRL2	Xu et al.
OR4C46	Xu et al.
PAG1	Xu et al.
PITPNM1	Xu et al.
PLCL2	Xu et al.
PML	Xu et al.
RB1CC1	Xu et al.

RGS12	Xu et al.
SAP30BP	Xu et al.
SLC26A7	Xu et al.
SLC26A8	Xu et al.
SPATA5	Xu et al.
TEKT5	Xu et al.
THBS1	Xu et al.
TRAK1	Xu et al.
TRRAP	Xu et al.
UGT1A3	Xu et al.
VPS35	Xu et al.
WDR11	Xu et al.
ZNF530	Xu et al.
ATP2B4	S2D project
BSN	S2D project
GRIN2B	S2D project
KIF17	S2D project
MAP2K1	S2D project
NRXN1	S2D project
SHANK3	S2D project
GABRA1	Baseline
GABRA2	Baseline
GABRA3	Baseline
GABRA4	Baseline
GABRA5	Baseline
GABRA6	Baseline
GABRB1	Baseline
GABRB2	Baseline
GABRB3	Baseline
GABRD	Baseline
GABRE	Baseline
GABRG1	Baseline
GABRG2	Baseline
GABRG3	Baseline
GABRP	Baseline
GABRQ	Baseline
GABRR1	Baseline
GABRR2	Baseline
GABRR3	Baseline
APBA1	Protein:Protein Interaction
APBA2	Protein:Protein Interaction
APP	Protein:Protein Interaction
C3	Protein:Protein Interaction
CALR	Protein:Protein Interaction
CAPN1	Protein:Protein Interaction
CASK	Protein:Protein Interaction

CCDC85B	Protein:Protein Interaction
COL7A1	Protein:Protein Interaction
CTBP1	Protein:Protein Interaction
CTSG	Protein:Protein Interaction
DLG1	Protein:Protein Interaction
DLG2	Protein:Protein Interaction
DLG3	Protein:Protein Interaction
DLG4	Protein:Protein Interaction
EP300	Protein:Protein Interaction
ESR1	Protein:Protein Interaction
FYN	Protein:Protein Interaction
HDAC1	Protein:Protein Interaction
HDAC3	Protein:Protein Interaction
JUN	Protein:Protein Interaction
MAPK14	Protein:Protein Interaction
MYC	Protein:Protein Interaction
PDGFB	Protein:Protein Interaction
PLAT	Protein:Protein Interaction
PLAUR	Protein:Protein Interaction
PLCB1	Protein:Protein Interaction
PLG	Protein:Protein Interaction
PRKCD	Protein:Protein Interaction
PTPN11	Protein:Protein Interaction
RARA	Protein:Protein Interaction
SHC1	Protein:Protein Interaction
SIN3A	Protein:Protein Interaction
SKIL	Protein:Protein Interaction
SPARC	Protein:Protein Interaction
SRC	Protein:Protein Interaction
STAT3	Protein:Protein Interaction
TAF9	Protein:Protein Interaction
TRIM27	Protein:Protein Interaction
WNT3A	Protein:Protein Interaction

3.7.2 Table 3.2 SKAT results for the dataset grouped by data source

Set ID	P value	Number of marker per
		set
Girard et al.	0.0015	1073
Xu et al.	0.0260	2322
S2D Project	0.0051	491
Baseline	0.6677	526
Protein:Protein Interaction	0.0413	1686

* The significance threshold is p-value < 0,01

3.7.3 Table 3.3 Significant association made on a gene and exon basis

Gene	Gene as set		Exon	Exon as Set	
	p-value	Number of markers		p-value	Number of markers
SBNO1	6.13x10 ⁻⁵	105			
			30/31	6.52x10 ⁻⁵	5
			29/31	8.80x10 ⁻⁵	7
			28/31	4.53x10 ⁻⁵	2
			24/31	4.62x10 ⁻⁵	2
			13/31	0.00010	5
			7/31	6.52x10 ⁻⁵	2
MAPK14	0,026	30	11/13	6.39x10 ⁻⁶	2
SHANK3	0,0064	68	10/23	5.90x10 ⁻⁵	2
EP300	0,0998	116	9/31	6.26x10 ⁻⁸	3

3.8 Figures

3.8.1 Figure 3.1 Network of genes with more than two protein :protein interaction with genes harboring DNM

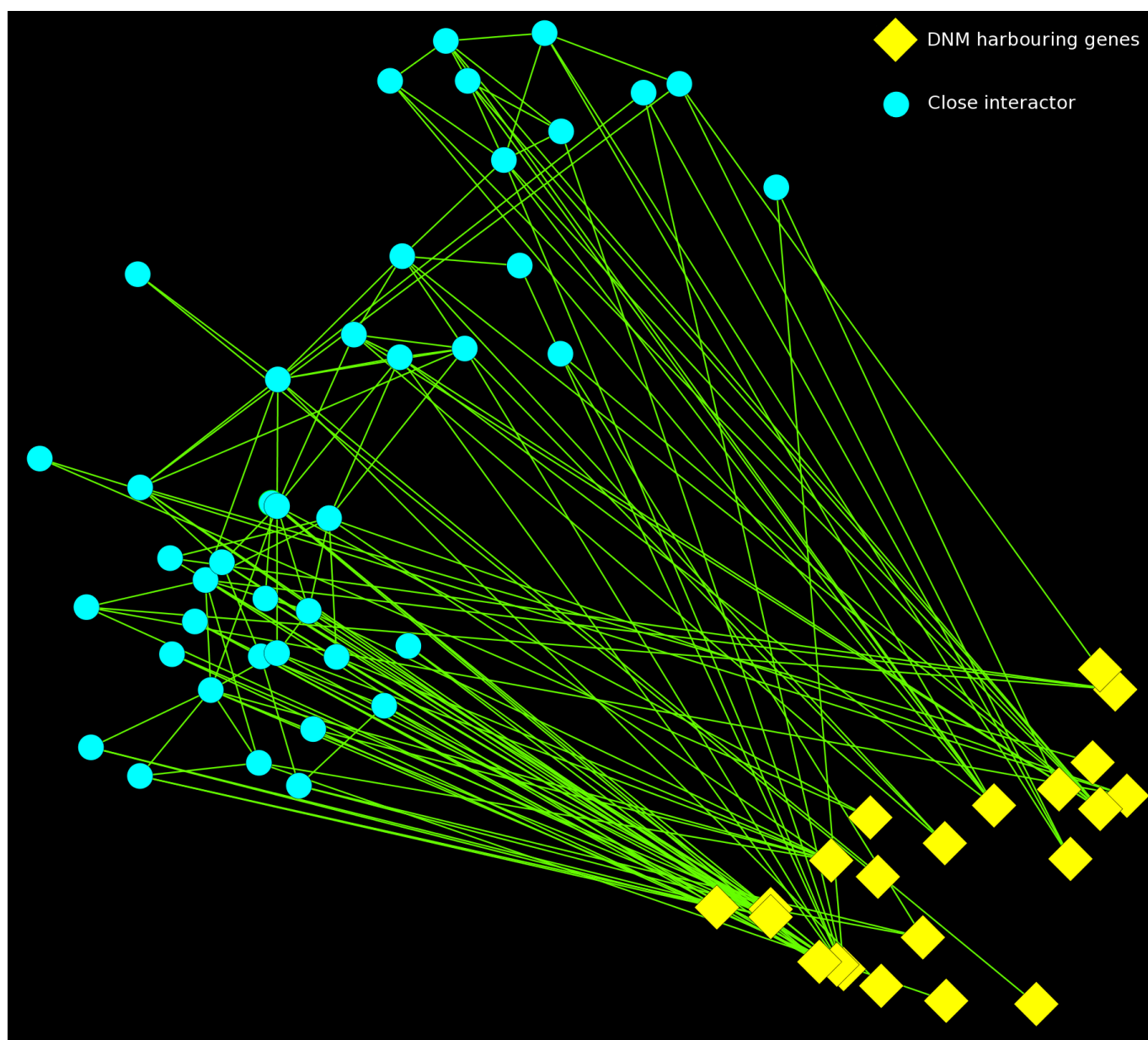


Figure 3.1: Network of genes with more than two protein:protein interaction with genes harbouring DNM : This figure shows the interaction between some DNM harbouring genes (yellow) and close interactor (cyan). The list of gene to be resequenced was primarily constituted of the genes found in the Girard et al. and Xu et al. exome sequencing studies (Girard, 2011, Xu 2011). To this list, 7 genes identified during the synapse2disease projet (S2D) were added to constitute the gene core set The close interactor genes were selected based on protein:protein interactions found in the Human Protein Reference Database (HPRD). All nodes that showed two or more interactions with our gene core set were kept. In total, 121 genes were kept for resequencing in our case control study.

3.8.2 Figure 3.2 Histogram of variant frequencies

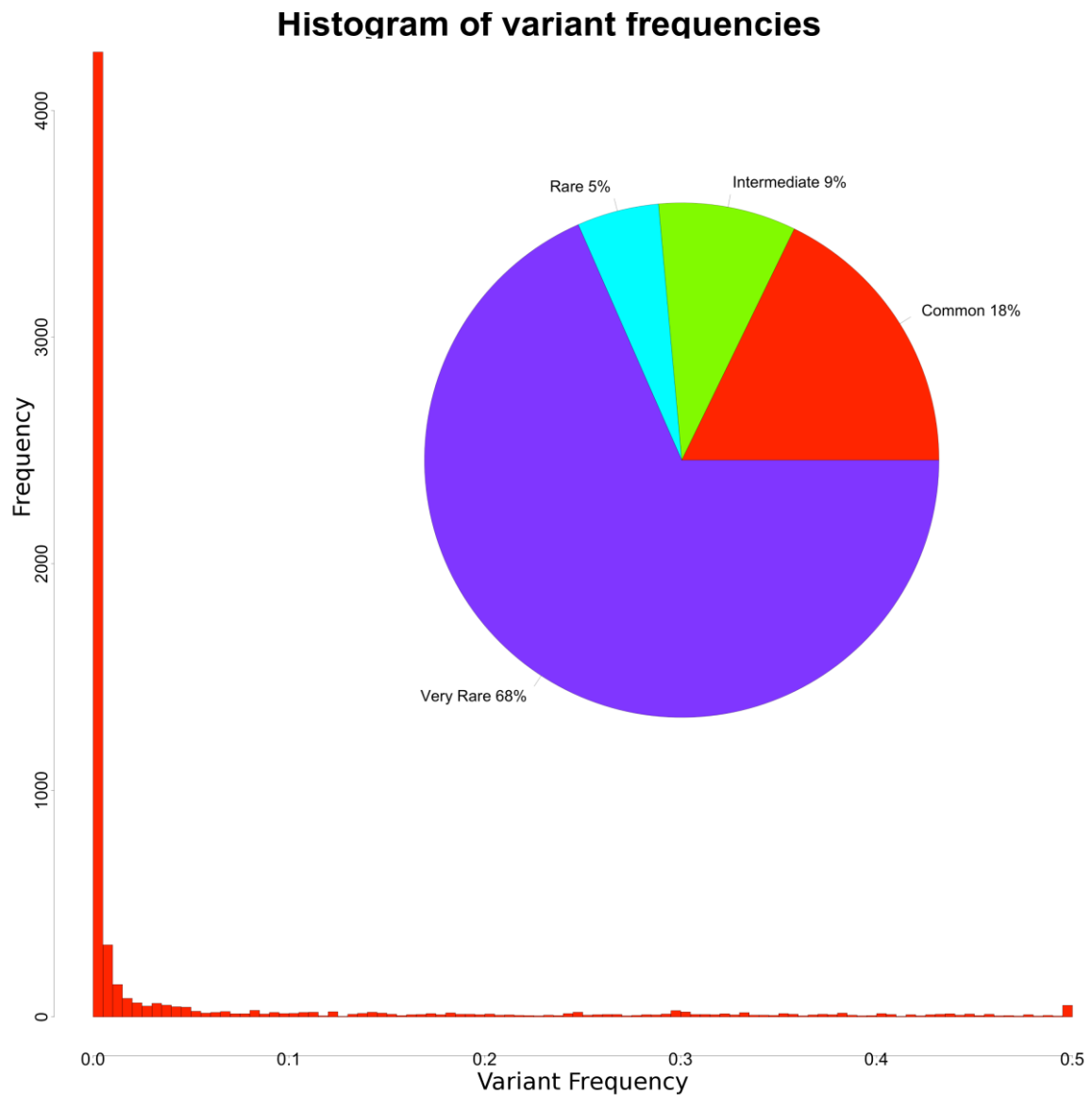


Figure 3.2: Histogram of variant frequencies : For each variant identified in the Haloplex dataset, we established a minor allele frequency. We then established a histogram of those frequencies. In the pie chart, the definition goes as following for variants : common (>5%), intermediate (1-5%), rare (0.5-1%) and very rare (<0.5%)

3.8.3 Figure 3-3 SKAT results for all genes on the assay

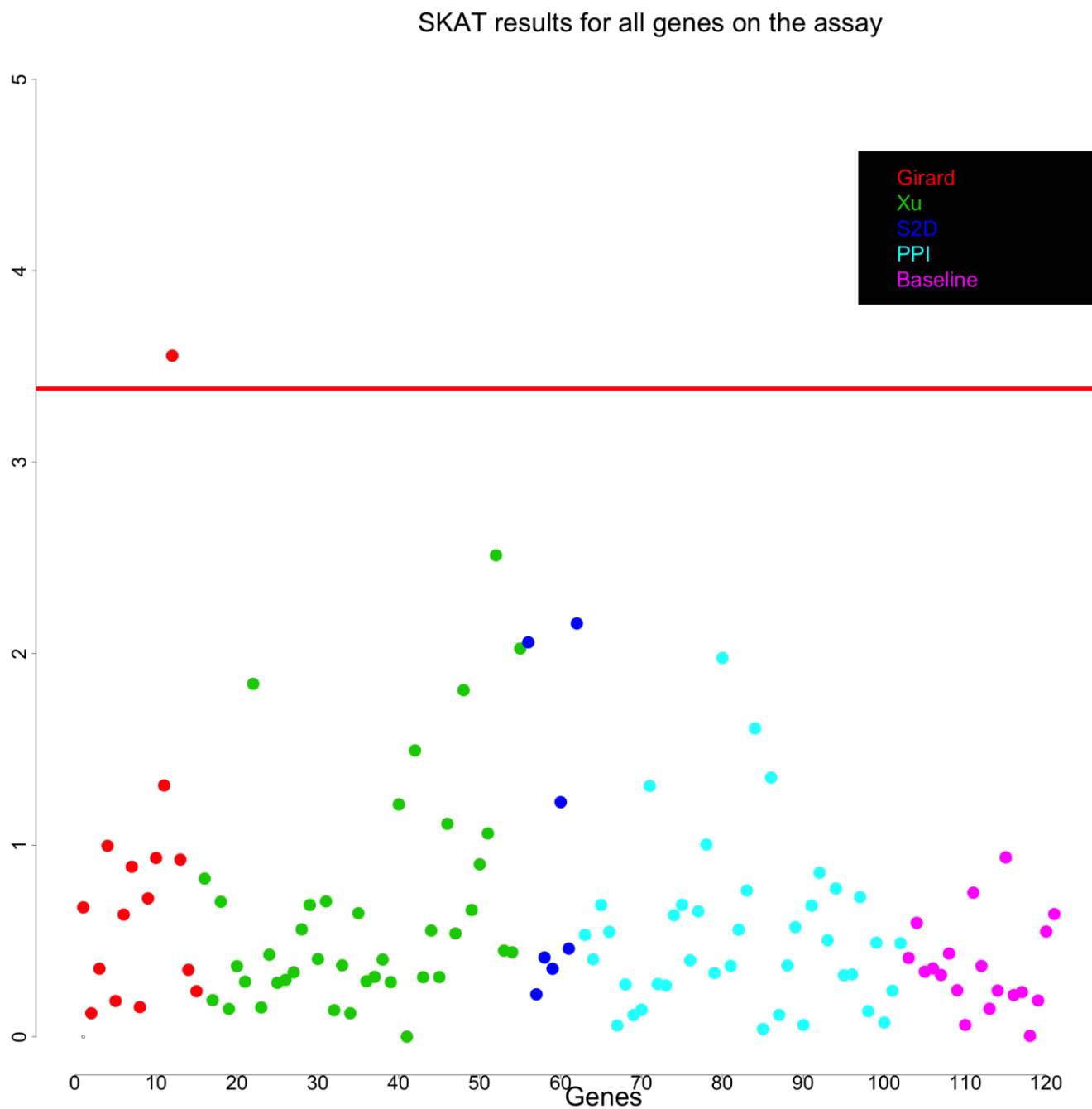


Figure 3.3: SKAT results for all genes on the assay : SKAT analysis were performed using only rare variants ($\leq 1\%$) and using genes as sets. The data sources are shown by the colors (Girard et al. = red, Xu et al. = Green, S2D project = Blue, Protein:Protein Interaction = Cyan, Baseline = Pink). The significance threshold was set using a Bonferonni correction to $4.1 * 10^{-4}$

3.8.4 Figure 3.4 SKAT results for all exons on the assay

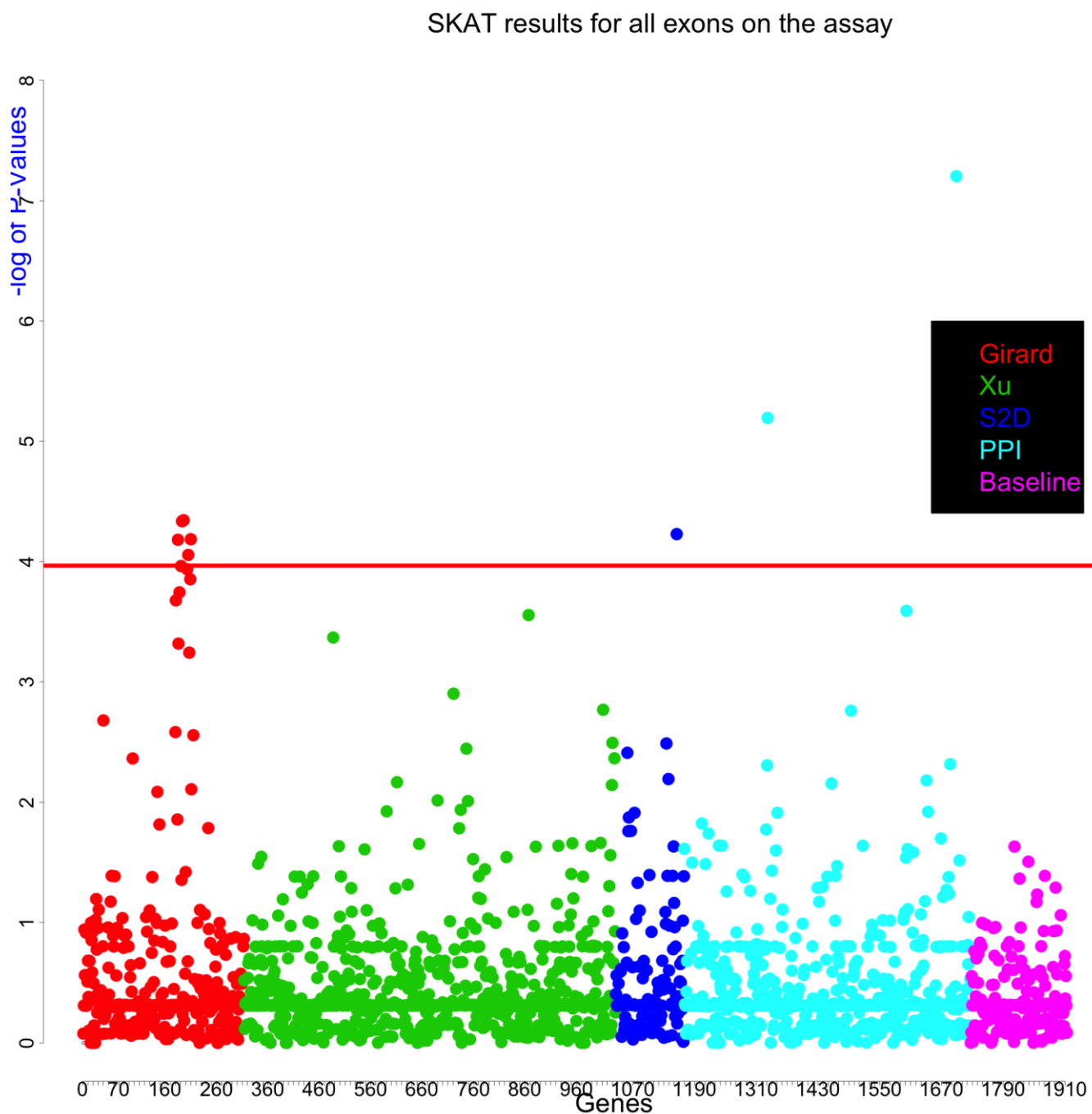


Figure 3.4: SKAT results for all exons on the assay : SKAT analysis were performed using only rare variants ($\leq 1\%$) and using exons as sets. The data sources are shown by the colors (Girard et al. = red, Xu et al. = Green, S2D project = Blue, Protein:Protein Interaction = Cyan, Baseline = Pink). P-values significance threshold was set to $1,0 * 10^{-4}$ using simpleM method.

3.8.5 Figure 3.5 Additional genotyping for the SBNO1 haplotype

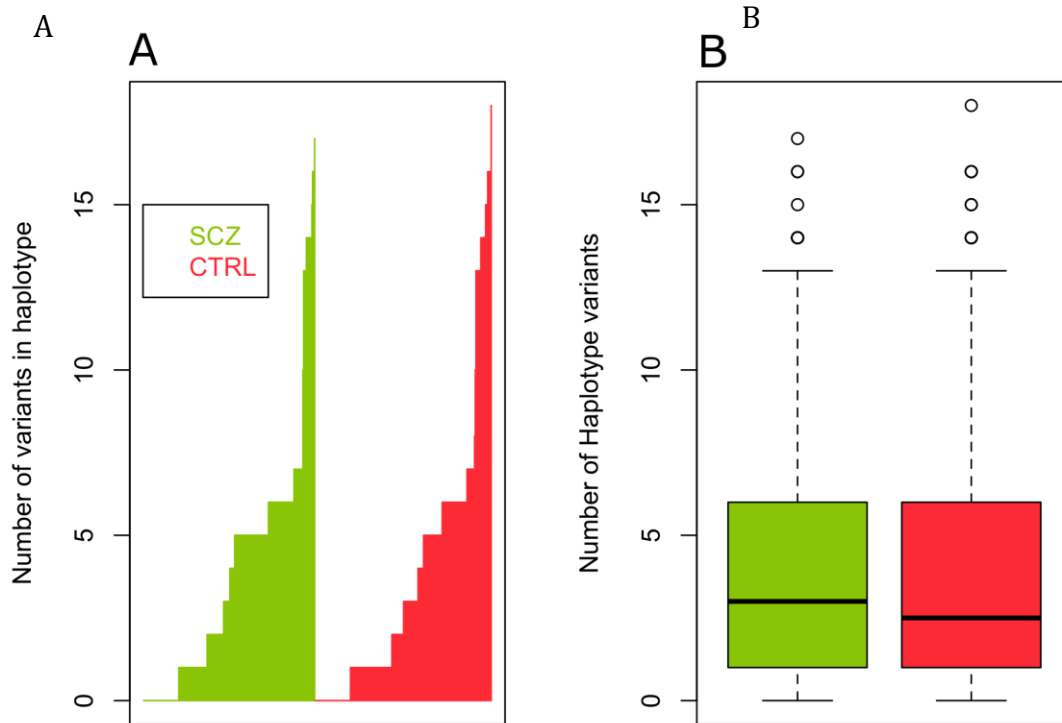


Figure 3.5: Additional genotyping for the SBNO1 haplotype : Sequenom genotyping was used to test for the SBNO1 haplotype in additional cohort of SCZ and controls. Panel A shows the exact distribution of the number of variants per individuals per cohorts and Panel B shows a summary of the same distributions.

Chapitre 4 L'âge paternel explique une proportion majeure de la variabilité des mutations *de novo* germinales des individus en santé

Il est important de mentionner que le chapitre 4 a été réalisé de manière concurrente avec le chapitre 3. Ainsi, il s'agit d'une suite logique de l'article présenté au chapitre 2 et il se consacre à un autre volet des résultats. Il n'y a donc pas de connexion logique avec le chapitre 3

4.1 Présentation de l'article

À la suite des observations faites lors de l'étude de prévalence des DNM chez des patients atteints de SCZ, plusieurs questions sont advenues. En tête de file, l'éternel dilemme, que l'on appelle communément « l'œuf ou la poule »; qui consiste à se demander si le taux de mutations est plus élevé chez certains individus et serait à la racine de la SCZ ou bien est-ce que la SCZ pourrait avoir un impact direct sur le taux de mutation? Ces interrogations nous mènent à poser une question cruciale pour les futures études génétiques : est-ce que le taux de mutation est fixé chez chaque espèce ou au contraire, est-ce qu'il varie en fonction de certains facteurs environnementaux? Jusqu'à tout récemment, on ne possédait que des hypothèses pour répondre à cette question. En ce qui concerne la SCZ, une observation nous permet de penser que le taux de mutations pourrait être directement lié à l'âge parental. En effet, deux études ont observé qu'un âge paternel avancé mènerait à une augmentation des chances

d'avoir un enfant atteint de la SCZ^{19,159}. Puisque nous avons précédemment démontré que le taux de mutations semble plus élevé chez les patients atteints de SCZ (voir chapitre 2), une hypothèse plausible est qu'un père plus âgé transmettrait plus de nouvelles mutations suite à une spermatogénèse moins efficace.

Pour en revenir à l'hypothèse de l'élévation du taux de DNM, il semble logique de penser que les spermatogonies Ad, qui se divisent de nombreuses fois afin de préserver la réserve de cellules germinales, peuvent accumuler de nombreuses mutations au fil des divisions. De plus, comme chaque testicule peut contenir plus de mille tubes séminifères, le nombre de spermatogonies Ad est très grand. De ce fait, l'accumulation de nouvelles mutations est sans doute inégale, mais logiquement, les spermatogonies Ad accumulent de manière générale de nouvelles mutations au fil de la vie d'un homme. De ce fait, l'hypothèse la plus plausible est que l'âge du père est directement lié à la variation du taux de mutations germinales et que, en conséquence, cette élévation du taux de mutations germinales pourrait mener à la SCZ.

Afin de tester cette hypothèse, il est primordial de pouvoir différencier les deux types de DNM, soit les mutations germinales et les mutations somatiques. Les mutations germinales se produisent durant la gamétycogénèse alors que les mutations somatiques se produisent lors de l'embryogenèse. L'une des manières les plus accessibles de différencier les deux types de mutations est de travailler sur des paires de jumeaux monozygotes. En effet, il est attendu que les mutations germinales montrent une ségrégation exclusive aux deux jumeaux alors que les mutations somatiques ne seront observées seulement que chez l'un des deux jumeaux. De plus, l'étude de jumeaux monozygotes permettrait également une identification plus facile

des mutations germinales puisque ces dernières seront observées chez deux individus, ce qui diminue drastiquement le nombre de faux positifs.

De ce fait, nous avons décidé d'effectuer le séquençage du génome complet (WGS pour Whole Genome Sequencing) chez des familles comprenant des jumeaux monozygotes. L'objectif étant de caractériser l'impact de l'âge du père, nous avons décidé de travailler chez des jumeaux ne présentant aucune condition médicale sévère. En se concentrant sur des jumeaux en santé, il est possible de caractériser directement l'impact de l'âge du père sur le taux de DNM sans biais par rapport à une condition médicale. Ainsi, nous avons établi un partenariat avec l'Étude des jumeaux nouveau-nés du Québec (EJNQ) qui est une étude qui a recruté et caractérisé une cohorte de 650 paires de jumeaux nés entre 1995 et 1998, ainsi que leurs parents. Parmi cette cohorte, nous avons identifié 5 familles avec un jeune père à la conception et 5 familles avec un père en âge avancé à la conception et nous avons effectué le WGS de chaque individu afin de caractériser le taux de DNM.

4.1.1 Contribution des auteurs

Planification de l'étude : SLG, MPD et GAR

Étude Nationale Jumeaux Québec : MB, MB, FV et GD

Analyses bio-informatiques : SLG, LPLP, DS, AA, EH, OD et ADL

Études SV: MAL, LPLP et SLG

Rédaction du manuscrit : SLG, PAD, MPD and GAR

Préparation des échantillons : CB et AN

Paternal age explains a major portion of *de novo* germline mutation rate variability in healthy individuals

Article soumis à American Journal of Human Genetics

Simon L. Girard^{1,2,3}, Louis-Philippe Lemieux Perreault^{2,3}, Marc-André Legault^{2,3}, Amina Barhdadi², Cynthia Bourassa¹, Dan Spiegelman¹, Alexandre Dionne-Laporte¹, Édouard Henrion¹, Ousmane Diallo¹, Amirthagowri Ambalavanan¹, Mara Brendgen⁴, Frank Vitaro⁵, Anne Noreau¹, Ginette Dionne⁶, Richard E. Tremblay^{5,7}, Patrick A. Dion^{1,8}, Michel Boivin⁶, Marie-Pierre Dubé^{2,3}, Guy A. Rouleau¹

1 Montreal Neurological Institute, McGill University, Montreal, Canada

2 Centre de Pharmacogénomique Beaulieu-Saucier, Institut de Cardiologie de Montréal, Montreal, Quebec, Canada

3. Université de Montréal, Faculty of Medicine

4 Département de Psychologie, Université du Québec à Montréal, Montreal, Canada

5 Département de Psychologie de l'éducation, Université de Montréal, Montreal, Canada

6 École de Psychologie, Université Laval, Quebec, Canada

7 University of London, London, UK

8 Département de pathologie et biologie cellulaire, Université de Montréal, Montreal, Canada

4.2 Abstract

De novo mutations (DNM) are an important source of rare variants and are increasingly being linked to the development of many diseases. Recently, the paternal age effect has been the focus of a number of studies that are attempting to explain the observation that increasing paternal age increases the risk for a number of diseases^{160,161}. Using disease-free familial quartets we show that there is a strong positive correlation between paternal age and germline DNM in healthy subjects. We also observed that germline CNV does not follow the same trend, suggesting an origin independent of parental age. Finally, we observed that DNM was not evenly distributed across the genome, which adds support to the possibility of DNM hotspots.

4.3 Introduction

Up to recently, little was known about the global prevalence of *de novo* mutations (DNM). Although some studies suggested that different factors could impact the global mutation rate¹⁶²⁻¹⁶⁴, it was not possible to assess it on a genome-wide scale. Recently, the arising of new sequencing technologies has made it possible to interrogate the full genome by sequencing. Using these technologies, it was recently demonstrated that paternal age effect (PAE) modulates the *de novo* mutation rate in patients with psychiatric disorders. Here, we corroborate these findings in ten disease free twins quartets from the same population. We show that germline single nucleotide variants (SNV) and indels but not CNV are associated with parental age.

We also confirm the existence of DNM hotspots, suggesting new mechanism for the genesis of new mutations.

4.4 Methods

4.4.1 Sequencing

DNA from the 40 individuals was extracted directly from blood provided by the Quebec Study of Newborn Twins (QSNT). All families were informed of the aims of the study and signed a consent form. The Ste-Justine hospital ethic committee gave approval for this study. Libraries were constructed using Illumina standard protocols. Sequencing was done on an Illumina HiSeq 2500 at Illumina sequencing facilities. Paired-end mode was used and the median fragment length was found to be ~300 bp for every individual, which was confirm to the library preparations.

4.4.2 SNP Genotyping

Genotyping was done using the Illumina HumanOmni2.5 genotyping array.

4.4.3 CNV Genotyping

Log R Ratio (LRR) and B Allele Frequencies (BAF) were extracted by GenomeStudio (version 2011.1) from Illumina HumanOmni2.5 BeadChips using default parameters. CNVs were identified using QuantiSNP (version 2.3)¹⁶⁵ with local GC correction (hg19) and default parameters. For *de novo* mutations, putative CNV regions were created using the twins' calls of any length but with a logarithm of the Bayes factor higher than 10. Only regions with higher than 25% reciprocal overlap found in both the twins were kept for further analysis. From this set, regions showing any sign of presence in one of the two parents (raw calls) or with an overlap with any regions

found in the Database of Genomic Variants (DGV, version 10, November 2010) were discarded.

4.4.4 SNV and indels calling

We receive sequences from Illumina in the form of alignment (bam) files and variants (vcf) files. In order to have consistent metrics we applied our own bioinformatics pipeline and used the same genome reference (GRCh37), so we could compare with previous runs. We extracted the reads and created pools of 20M reads using Picard SAMtoFastq in order to optimise the use of our computing farm. Each of these pools were aligned using a modified version of the Burrows-Wheeler Aligner (bwa version 0.6.2-r126-tpx)¹²⁵ with threading enabled). The options were 'bwa aln -t 12 -q 5' and 'bwa sampe -t 12'. Next we recombined the pools in a single alignment (bam) file using Picard MergeSamFiles.

We used the realigned bam file in GATK 1.6¹⁵⁰ to process the data with indel realignments, read duplicates removal and quality score recalibration. We made use of the scatter-gather mechanism in GATK to accelerate the process. Once all 40 samples were processed, we generated QC metrics using GATK DepthOfCoverage and custom scripts.

We used GATK UnifiedGenotyper to call SNVs and Indels in a run containing all samples, and splitting by chromosome to accelerate the process on the cluster. Finally, the variants were filtered and ordered using GATK Variant Recalibration, which assigns a validity score to each variant using a training algorithm. For this purpose we strictly followed the procedure from the Broad Institute.

Once the variant calls were produced, these were compared to the SNP-chip calls; all samples had a variant concordance higher than 99.5%.

4.4.5 WGS identification of SV

In order to achieve high-confidence structural variants calls for our samples, two different algorithms were used. The first one, *CNVer* v0.8.1¹⁶⁶, uses information from the paired-end mappings and from the depth of coverage in a given region to call CNVs (insertions and deletions). Prior to the mapping, the *FastqMcf* v1.1.2 (<https://code.google.com/p/ea-utils/wiki/FastqMcf>) utility was used to detect and remove fragments from sequencing adapters and primers and to remove poor quality bases at the end of reads. The *Bowtie* v0.12.9¹⁶⁷ aligner was then used with the “-v -2 -a -m 600 --best --strata” options, as specified by the *CNVer* authors. To reduce the number of false positives, CNVs called by *CNVer* were merged when the distance to their neighbor was fewer than 5 kilobases.

The second algorithm, *BreakDancer*¹⁶⁸, uses paired-end mappings signatures to identify different types of structural variants, including insertions, deletions, inversions and translocations. The reads were processed as described in the *SNV and indels calling* section before using *BreakDancerMax* (cpp implementation, version 1.1.2) to call structural variants. Once again, CNVs were merged with a 5 kilobase proximity threshold.

The germline DNM detection and exclusion from this dataset required a maximum overlap between every twin and his parents of 1% and a minimum overlap to the other twin of 90%.

4.4.6 Sanger sequencing validation

Primers were designed using in-house scripts and the Batch Primer 3 web tool (<http://probes.pw.usda.gov/batchprimer3/>) using enough flanking sequence to have a good coverage of selected variants. PCR was performed using the AmpliTaq Gold DNA Polymerase (Applied Biosystems) according to the manufacturer's instructions. To visualize DNA fragments, a small fraction of the PCR product was loaded on a 1.5% agarose gel containing ethidium bromide. PCR products were sequenced at the Genome Quebec Innovation Centre using a 3730XL DNAalyzer (Applied Biosystems), and SeqMan from the DNASTar suite was used for mutation detection analysis.

4.5 Results and discussion

In the last years, many groups have shown association of *de novo* mutations (DNM) with neurodevelopmental disorders, notably intellectual disability (ID)¹⁶⁹, autism spectrum disorder (ASD)¹⁴⁵⁻¹⁴⁷ and schizophrenia⁹⁸⁻¹⁰⁰. At the time those studies were initiated, it was yet unclear whether the DNM rate was fixed across our species or if it was subject to variation. Interestingly, it was previously shown that paternal age at conception is positively correlated with risk of schizophrenia^{19,159}. In order to evaluate the effect of paternal age on the rate of DNM in healthy subjects, we have selected ten families consisting of two monozygotic twins and their parents. All individuals were recruited by the Quebec Study of Newborn Twins (QSNT) project¹⁷⁰. These families have been prospectively followed since birth and no severe health conditions were reported for the twins, now all around 15 years old. Here, families have been selected according to extremes in paternal age at the time of conception. A group of 5 younger-aged fathers (mean = 23.37±SD 2.06) and 5 older-aged fathers

(mean = 43.33±SD2.75) (Table 1) were selected. Paternal age at conception was significantly different between the two groups (t-test, p-value < 1.00·10⁻⁵). Despite our efforts to select comparably aged mothers, there was also a small but statistically significant difference in maternal age at conception (t-test, p-value < 0.0065). Thus, we will be able to directly assess the effect of parental age effect on DNM variability and we will need to establish parental origin of mutations to directly implicate the paternal or maternal age. Whole Genome Sequencing (WGS) was performed on all 40 family members from the 10 families. Mutation detection was performed using an in-house bioinformatics pipeline to optimize the detection of variants present in the twins against the mutation profiles of the parents. All samples were genotyped using Illumina HumanOmni2.5 BeadChips and the SNP calls were used to calibrate and validate variant detection with the WGS bioinformatics pipeline. Data from the Illumina SNP-chips were also used to identify *de novo* CNVs in twin pairs. We also used the WGS data to identify *de novo* CNVs using two different algorithms. Consistent DNMs in a twin pair were categorized as germline (i.e. inherited from gamete cells of one of the parents), while DNM found in a single twin were categorized as somatic (i.e. mutations happening post-conception). Overall, 889 germline mutations and 106 somatic mutations were detected using these criteria in the 10 twin pairs. In order to test the validity of our *in silico* detection pipeline, we tested a subset of the DNM identified using Sanger Sequencing. A total of 91 germline variants as well as 62 somatic variants were validated in the corresponding quartet. Almost 90% of the germline variants were validated. Based on these results, we calculated the individual false positive rate for each quartet and we use these individual rates to adjust the estimated total number of DNM for each twin pairs. For family quartets for which no false positive were identified, we corrected with the

overall false positive rate of 10.9%. For the somatic variations only 9 of 62 variants were confirmed, resulting in a false positive rate higher than 85%. Thus, given that we would predict that we have detected only 13 somatic DNM, it is impossible to arrive at any conclusions regarding somatic DNM rates, except to say that they are in limited number when looking at DNA extracted from blood.

Assuming a single mutation event for germline mutations by collapsing all twin pairs to a single observation, we calculated the germline *de novo* mutation rate to be $1.37 \cdot 10^{-8}$, which is higher than the previously reported ratio of $1.1 \cdot 10^{-8}$.¹⁰⁴ However, when stratified by paternal age, the estimated DNM rate in the groups of younger and older fathers respectively were $1.01 \cdot 10^{-8}$ and $1.74 \cdot 10^{-8}$ (Rao-Scott chi-square $p=1.38 \cdot 10^{-12}$). We then tried to assess if this elevation was also seen for CNV. For this, we had at our disposal the WGS data as well as SNP genotyping arrays. For the WGS data, we used two algorithms; one focussing on the depth of coverage information (CNVer) and another one that uses read pairs mapped with unexpected distance or orientation (BreakDancer). As getting a consensus for CNV is not trivial and was not needed here, we tested for parental age effect on each separate dataset without building a consensus dataset. The results were concordant across both approaches and no association with parental age could be detected (Table 2).

We next sought to find if there was a specific association between parental sex and the variability in DNM rate. We found that there was a strong correlation between the germline DNM and paternal age (Poisson regression $R^2 = 0.78$, $p=0.00066$) as well as with maternal age (Poisson regression $R^2 = 0.65$, $p=0.004$), though the effect size is deemed to be smaller for the later. Then, we tested for parental effect looking separately at single nucleotide variants (SNV) and small insertions and deletions

(indels) (Figure 4.1). We found a strong correlation between paternal age and single nucleotide germline DNM variants (Spearman correlation $R= 0.7695$, $p=0.0092$). This correlation was also observed in the mother, although not as strong (Spearman correlation $R=0.745$, $p=0.013$) suggesting that parental age accounts for an important portion of the variability in the DNM rate. A positive correlation was also found for indel germline variants for both paternal (Spearman correlation $R= 0.7631$, $p=0.01$) and maternal age (Spearman correlation $R=0.702$, $p=0.0237$). The weaker evidence with indels could in part result from the lower power due to the comparatively less frequent occurrence. This observation is nonetheless in concordance with other studies^{160,161}. Interestingly, germline de novo CNV seems to behave differently than for SNVs or indels, as we did not find an association between parental age and CNVs (Figure 4.2). Although further validation is required, this would suggest that the mechanisms explaining the increasing rate of SNV DNM with increasing parental age do not apply to CNVs, which may have implications in our understanding of this phenomenon.

In order to separate out the paternal and maternal age effects, we used individual WGS reads to assign parental origin of the germline DNM. Working only with germline SNVs, as they were more frequent than indels, we assigned parental origin according to co-transmission of DNM with polymorphisms on single sequence reads. Out of 765 DNM, we were able to trace parental origin for 57 DNM. Of those, 43 DNM were of paternal origin and 14 were of maternal origin; providing an estimated paternal:maternal ratio of $\sim 3:1$ in our dataset. When limiting the comparison to the younger paternal age group, we see 16 paternal germline DNM and 5 maternal germline DNM (ratio of $\sim 3:1$) whereas the older paternal group has 27 paternal

germline DNM and 9 maternal germline DNM (ratio of $\sim 3:1$). These data suggest that DNM rates increase both with maternal and paternal age, though the paternal age explains a greater number of DNMs than maternal age.

To evaluate the distribution of DNM across the genome, we included the publically available dataset from Michaelson et al. (10 ASD twin families) and Kong et al. (78 trios) for a total of 6,892 DNM^{160,161}. Assuming that evenly distributed DNM would be found every ~ 500 Kb, we found the distribution of the DNM to be significantly different (Kolmogorov Smirnov goodness of fit test, p -value $< 8.03 \cdot 10^{-9}$). Indeed, we observed several clusters of more than 6 DNM within a 500 Kb region, the densest regions being on chr7 and chr8 (Figure 4.3). Considering, however, that the Kong and Michaelson datasets consists of psychiatric patients, we repeated the analysis using only the disease free twin families in genomic windows of 4 Mb (adjusted for the total number of DNM). We found a significant deviation from uniformity (KS-test, p -value < 0.0306) with dense clusters containing up to 4 DNM (Figure 4.4).

In this study, we show that parental age, more specifically paternal age, explains a major portion of the variation in the germline DNM rate. Contrary to other studies that have reported similar findings, we have studied disease free families, and we have isolated the germline DNM rate from somatic mutations with respect to parental age. In our study, fathers aged 46 or greater had a rate of germline DNM that was twice as high as that of fathers aged twenty years younger. We also identified a smaller, yet significant increase in the mutation rate associated with maternal age. This could be due to the assortative mating present in our families, but parental origin of mutation suggest that older maternal age also leads to an increase in DNM. In addition, our data suggests that DNMs do not occur randomly throughout the

genome, and that there could be DNM hotspots in the genome, as previously observed in ASD¹⁶¹. The implications of these findings are important, as an increased DNM rate could impact a number of mendelian and complex genetic disorders. A study design that includes the DNA of two parents and their children has proved a valuable approach for the investigation of DNM. The ability to assess DNM rates offers a valuable means to further characterize the impact of rare variants on diseases and traits and should be encouraged in the design of future studies where parents are available.

Acknowledgements

We thank the families involved in our study. We are thankful for the efforts of the members of the Illumina genome sequencing facilities. Guy A. Rouleau is grateful for the support received through his positions as Canada Research Chair in Genetics of the Nervous System and Jeanne-et-J.-Louis-Levesque Chair for the Genetics of Brain Diseases. Simon L. Girard is grateful to the FRQS for financial support.

4.6 Tables

4.6.1 Table 4.1 : Summary of DNM found for each twin quartet

Group	Family ID	Paternal Age	Maternal Age	Twins	Base pair sequenced (in Gbp)	DNM germline	FP Rate	Global germline DNM Rate*
Group 1								
	TQ1	45.27	30.15	Twin1	2,779	142	16,67%	2,14E-08
				Twin2	2,776			
	TQ2	47.15	30.36	Twin1	2,779	96	10,99%	1,54E-08
				Twin2	2,779			
	TQ3	41.42	31.99	Twin1	2,795	91	14,29%	1,41E-08
				Twin2	2,789			
	TQ4	40.8	36.77	Twin1	2,779	117	20,00%	1,69E-08
				Twin2	2,779			
	TQ5	42	34.68	Twin1	2,779	121	10,99%	1,94E-08
				Twin2	2,776			
							old father global DNM rate =	1,74E-08
Group 2								
	TQ6	25.8	26.5	Twin1	2,795	60	10,00%	9,75E-09
				Twin2	2,795			
	TQ7	24.8	19.91	Twin1	2,795	71	15,38%	1,08E-08
				Twin2	2,792			
	TQ8	23.55	21.81	Twin1	2,779	77	10,99%	1,24E-08
				Twin2	2,779			
	TQ9	20.73	21.56	Twin1	2,779	58	10,99%	9,32E-09
				Twin2	2,789			
	TQ10	21.96	19.98	Twin1	2,795	56	23,08%	7,78E-09
				Twin2	2,792			
							young father global DNM rate =	1,00E-08

*The Global germline DNM rate was corrected with the appropriated FP value for each family

4.6.2 Table 4.2 : CNVs identified for each family

Group	Family ID	Paternal Age	Maternal Age	QuantiSNP Loss	QuantiSN P Gain	CNVer Gain	CNVer Loss	BreakDancer Loss*
Group 1								
1652	TQ1	45,27	30.15	8	6	1	10	3
1207	TQ2	47,15	30.36	7	4	4	6	4
1480	TQ3	41,42	31.99	2	3	4	9	5
1389	TQ4	40,8	36.77	1	3	3	6	7
1490	TQ5	42	34.68	0	1	5	6	5
Group 2								
1301	TQ6	25,8	26.5	2	0	5	7	4
1323	TQ7	24,8	19.91	4	3	2	4	7
1443	TQ8	23,55	21.81	2	9	5	5	1
1121	TQ9	20,73	21.56	3	5	2	7	3
1286	TQ10	21,96	19.98	4	4	3	10	6

*Breakdancer was not able to confidently identifies gain in our dataset

4.7 Figures

4.7.1 Figure 4.1 : Association of parental age with germline DNM

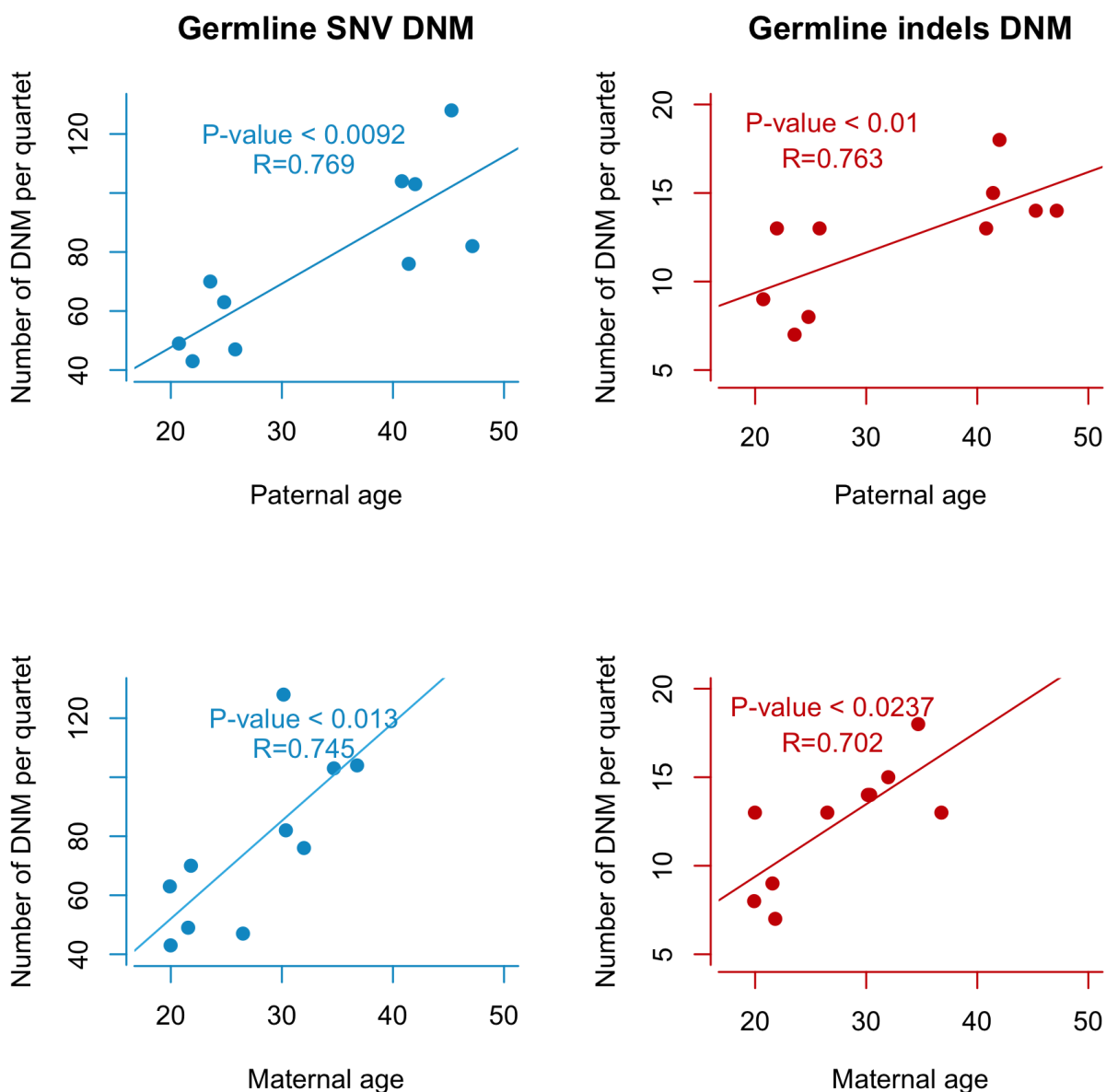


Figure 4.1 : Association of parental age with germline DNM : This figure shows the correlation between different SNV and indel germline DNM (left : germline SNV, right : germline indels) with parental age (Top : paternal age, bottom : maternal age). The X-axis represents the parental age at conception. The Y-axis represents the number of DNM mutations identified through Whole Genome Sequencing.

4.7.2 Figure 4.2 : Association of parental age with CNV called by different algorithms

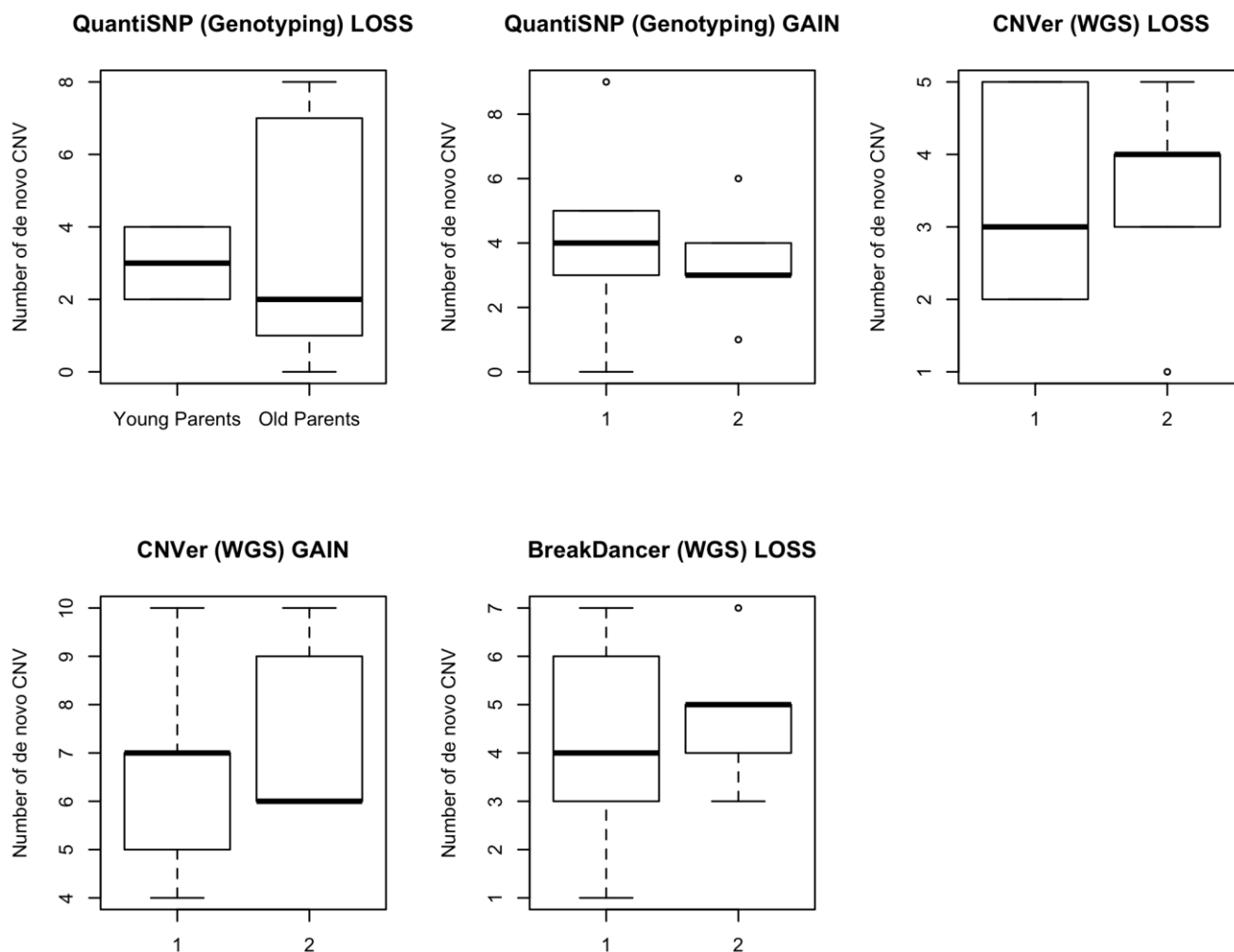


Figure 4.2: Association of parental age with CNV called by different algorithms : We used three different algorithms to detect CNV in our dataset. QuantiSNP was used for genotyping assays while CNVer and BreakDancer were used for WGS. Although the number varies according to which algorithm was used, no difference between young parental age group and old parental group can be detected.

4.7.3 Figure 4.3 : Circular plot of the autosomal human genome using combined dataset

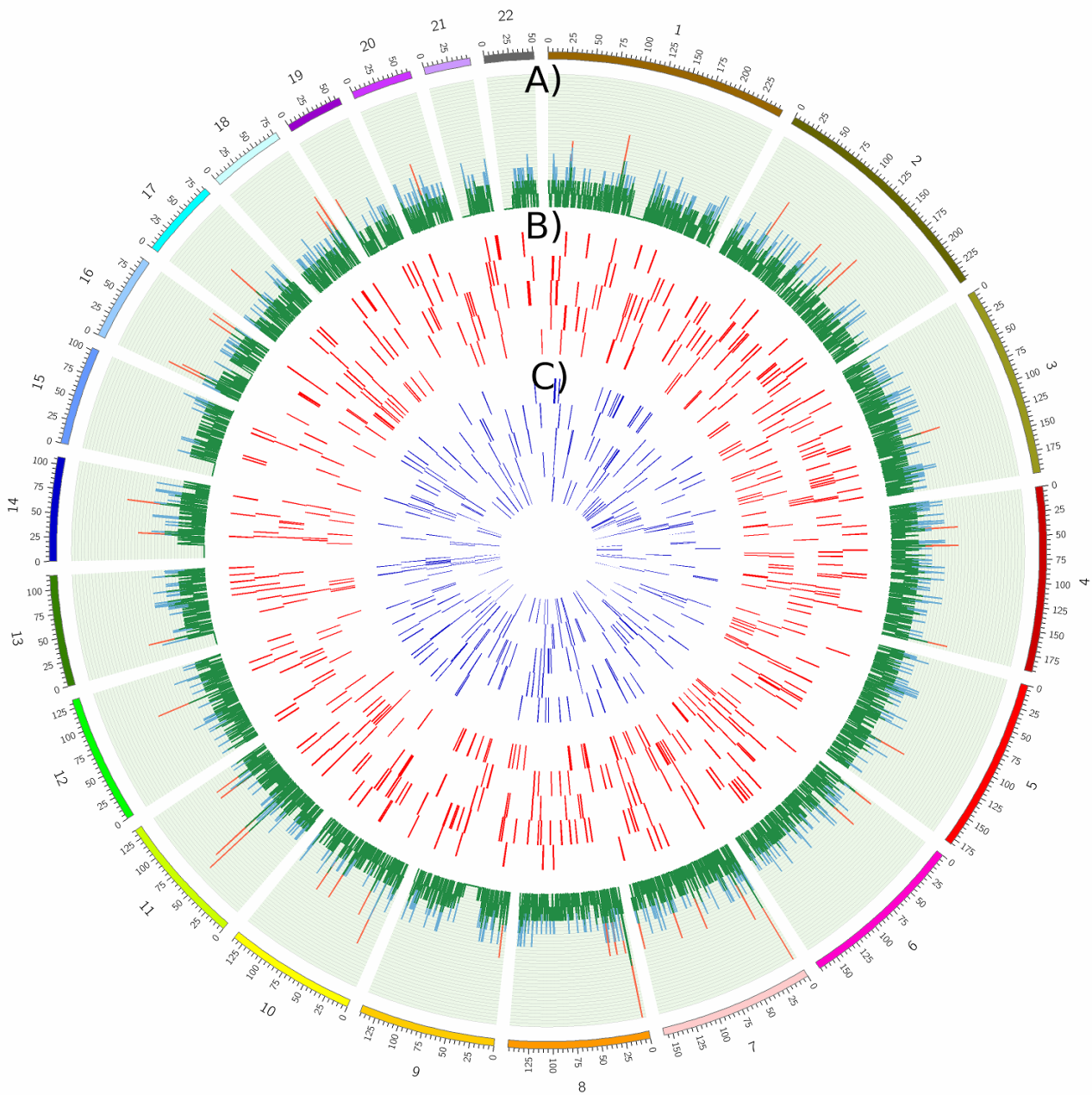


Figure 4.3 : Circular plot of the autosomal human genome. A) Line plot of the number of DNM found in the combined dataset of the twins data with data from Kong et al. and Michaelson et al., with windows size of 500 Kb. B) Density plot of DNM found in the advanced paternal age group C) Density plot of DNM found in the young paternal age group. Figure was made with Circos

4.6.4 Figure 4,4 : Circular plot of the autosomal human genome using twins dataset

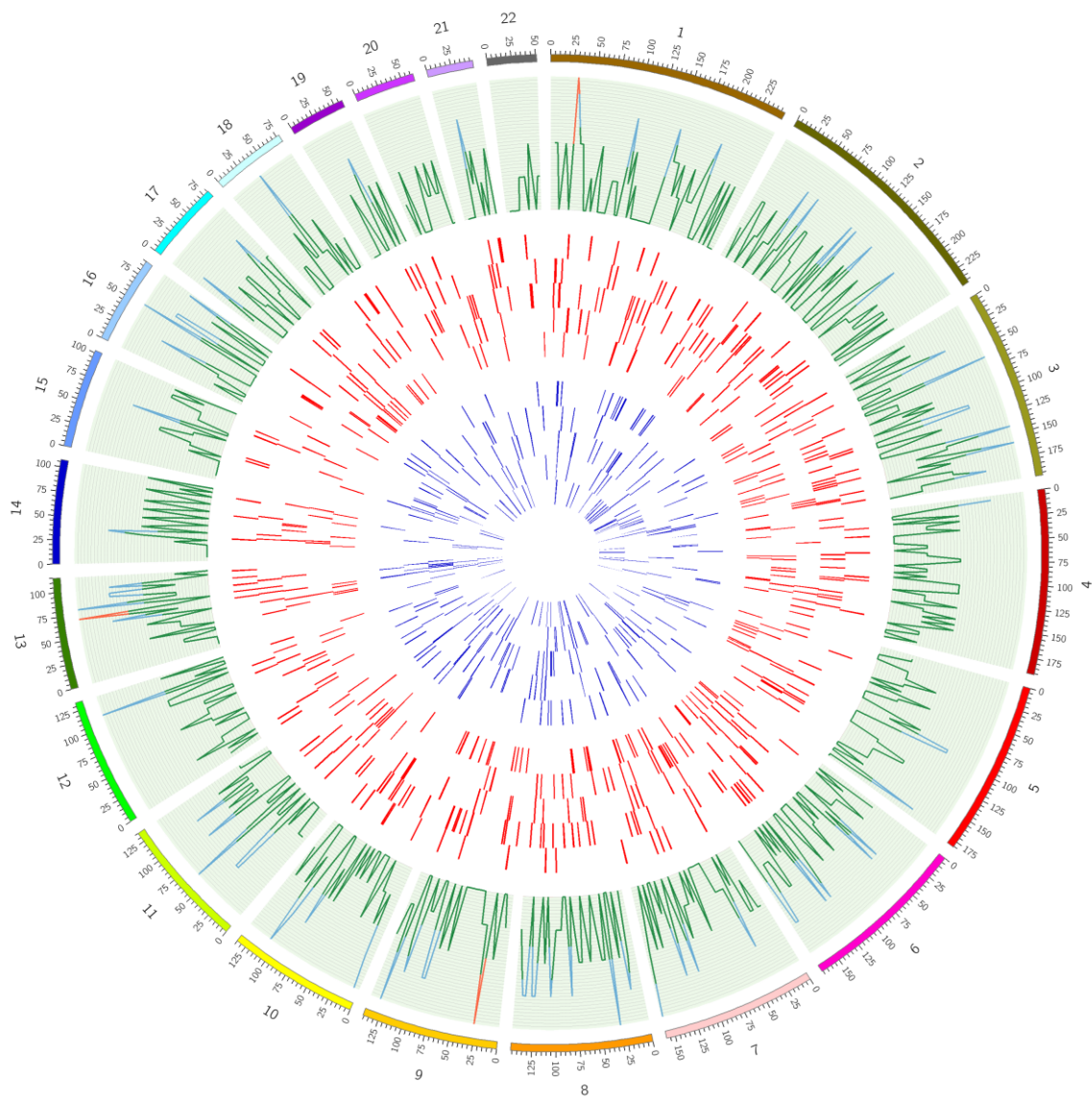


Figure 4-4 : Circular plot of the autosomal human genome. A) Line plot of the number of DNM found in the twins dataset with windows size of 1500 Kb. B) Density plot of DNM found in the advanced paternal age group C) Density plot of DNM found in the young paternal age group

Chapitre 5 Discussion et perspectives

5.1 Retour sur les travaux

5.1.1 Les mutations *de novo* dans la schizophrénie

Lorsque les travaux présentés dans le présent ouvrage ont été entrepris, les connaissances dans la communauté scientifique sur les DNM étaient très limitées. Cependant, les données précédemment générées par le laboratoire (Projet Synape2Disease, visant à séquencer 1000 gènes synaptiques afin de découvrir des DNM chez des patients atteints de maladies psychiatriques) laissaient penser que le rôle des DNM dans les maladies psychiatriques tout particulièrement, pourrait être beaucoup plus important qu'on l'estimait précédemment.

Ainsi, au moment où les nouvelles technologies de séquençage ont commencé à être globalement accessibles, nous avons décidé de lancer la première étude portant sur la découverte de DNM à grande échelle chez des patients atteints de SCZ. Jusque-là, l'approche par gènes candidats était la seule option possible et était lourdement limitée par le biais de sélection de gènes. Ainsi, la technique de capture d'exome offre une approche plus vaste en couvrant la grande majorité des gènes. En contrepartie, nous avons dû laisser tomber la composante populationnelle de l'étude. En effet, le coût du séquençage à haut débit, malgré le fait qu'il soit significativement réduit par l'utilisation de la capture d'exome, demeurait prohibitif pour une étude sur une grande cohorte. Nous avons donc choisi 14 patients avec un phénotype clair. Ces patients ont été diagnostiqués et recrutés par la même équipe

et ont été sélectionnés parmi une grande cohorte. La capture d'exome et le séquençage ont été effectués par notre équipe et une plateforme d'analyse a été conçue afin d'optimiser la détection des DNM. Au total, nous avons identifié 15 DNM chez 8 patients différents. La première conclusion évidente est qu'il y a un nombre important (n=6) de patients pour qui aucune DNM n'a été identifiée. Ce nombre n'est pas pour autant surprenant; le mécanisme génétique derrière la SCZ étant très complexe, il aurait été surprenant que les DNM expliquent la majeure partie du facteur héréditaire de la maladie. Cependant, vu certaines limitations techniques, il n'est pas non plus exclu que certaines DNM n'aient pas pu être identifiées.

Chez les patients pour qui des DNM ont pu être identifiées, nous avons observé certains patients avec plusieurs DNM, deux patients étant porteurs de trois DNM. Ce nombre de mutations nous laisse croire que certaines des DNM identifiées pourraient être causalement liées à la SCZ tandis qu'une importante fraction serait bénigne. L'hypothèse inverse serait que toutes les mutations identifiées soient bénignes et n'aient aucun lien avec la maladie; elles seraient alors dues au fait que les DNM se produisent sporadiquement chez tous les individus.

Afin de déterminer si les DNM identifiées ont un potentiel causal pour la SCZ, nous avons décidé d'établir des statistiques globales pour l'ensemble des DNM. Sans doute, l'observation la plus flagrante était le nombre relativement élevé de mutations non-sens (n=4). Ce ratio de 4/15 paraissait élevé et, en faisant des études de prévalence, nous avons conclu qu'un profil normal (tel que répertorié par la base de données dbSNP) pour les mutations non-sens serait de 1/20 par rapport aux mutations faux-sens. Lorsque l'on compare le ratio observé dans la présente étude avec celui répertorié par une base de données de mutations directement liées à des

maladies (1/4), on observe une bonne concordance, ce qui suggère que le profil des DNM identifiées chez des patients atteints de SCZ est beaucoup plus proche d'un profil pathogénique que d'un profil neutre.

Une autre observation importante est la variation du taux de DNM. Bien que nous n'ayons pas pu séquencer de familles contrôles afin de quantifier un taux de mutations directement comparable avec nos familles de SCZ, plusieurs études contemporaines à la nôtre ont permis d'établir un taux de mutations global chez l'humain d'environ 1.1×10^{-8} mutations par paire de bases par génération. Nous avons établi que le taux de mutations observé dans la totalité de nos patients atteints de SCZ est de plus de deux fois le taux de référence attendu, et ce, même en prenant en compte les patients pour qui aucune DNM n'a pu être identifiée. En effet, selon le taux de référence, nous aurions dû identifier entre 6 et 8 DNM dans notre cohorte, et nous en rapportons le double. Cependant, à la suite de cette observation, une question de relation de cause à effet se pose : est-ce que la SCZ augmente le taux de DNM ou bien est-ce que les DNM sont liées à la SCZ? D'autant que, en choisissant une cohorte de SCZ, nous avons un biais de sélection qui résulte en l'augmentation du taux de DNM. Il n'était pas possible de répondre à cette question à la fin de la première étude, mais une chose était bien claire; il était à ce point justifié de poursuivre l'étude de la piste des DNM dans la SCZ.

5.1.2 Le séquençage de gènes candidats

De ce fait, nous avons entrepris une étude de séquençage de gènes candidats. Le cœur même de l'ensemble de gènes est constitué des gènes identifiés lors de nos études de DNM précédentes^{97,98} ainsi que d'une autre étude publiée après les

nôtres¹⁰⁰. À cela, nous avons ajouté des gènes pour lesquels nous avons identifié un enrichissement d'interaction protéique : protéine avec les gènes porteurs de DNM. Finalement, nous avons également ajouté un groupe « seuil », constitué de gènes synaptiques jamais impliqués dans le mécanisme de la SCZ (gènes de la famille des récepteurs GABRA). Le choix du groupe seuil a évidemment été le sujet de maintes discussions. Nous avons aussi considéré choisir des gènes dits de maintenance mais nous avons plutôt opté pour des gènes exprimés dans le cerveau qui sont plus proche du contexte suggéré pour la schizophrénie.

À la suite des récentes GWAS, il semble clair que les variants communs ont une responsabilité limitée dans le mécanisme génétique de la SCZ. En effet, la dernière étude⁷³ portant sur plus de dix mille patients atteints de SCZ a été en mesure d'identifier seulement quelques variants communs qui semblent avoir une modeste contribution génétique au profil héréditaire. Il serait erroné de croire que notre étude portant sur 240 patients atteints de SCZ puisse identifier des variants communs qui aient échappé à la grande GWAS. De ce fait, nous avons décidé de prioriser l'hypothèse des variants rares. C'est également l'hypothèse la plus compatible avec les DNM. En effet, puisque nous ne disposons pas des parents des patients schizophrènes de la cohorte, nous ne sommes pas en mesure de détecter précisément les DNM. Par contre, parmi les variants privés, il est attendu qu'une certaine partie sera des DNM. De ce fait, une analyse priorisant les variants rares donnera automatiquement un plus grand poids aux DNM. Cependant, vu la nature peu fréquente de ces variants, nous avons dû opter pour un test statistique alternatif, car les tests d'association d'études traditionnelles sont effectués sur les variants individuellement. Ceci ne serait pas une approche suffisamment puissante

pour la détection de variants rares dans notre population de 480 sujets. Ainsi, nous avons opté pour un test dit « d'effondrement » (de l'anglais « collapsing study ») qui permet de regrouper plusieurs variants dans un ensemble défini (tel un gène, un réseau biologique, ou un jeu de données) et d'appliquer un test statistique d'association sur cet ensemble.

Comme les gènes choisis provenaient de différents jeux de données, le premier test a été effectué en utilisant les provenances expérimentales comme ensembles. Le but de ceci était de tester si les mutations provenant des études DNM de SCZ avaient un profil caractéristique plus pathogénique que celui des gènes du groupe contrôle ou bien des gènes ayant été identifiés par le réseau d'interactions. Il était attendu que le profil pathogénique se traduirait par une différence du fardeau de mutations entre les cas et les contrôles. Nous avons effectivement confirmé que les trois ensembles issus des études DNM de la SCZ présentaient un profil de mutations rare significativement différent de celui de la référence ($P < 0.05$). Les deux ensembles de données provenant de notre laboratoire étaient également sous le seuil imposé par une correction multiple de Bonferroni ($P < 0.001$). Une autre observation intéressante est de constater que le groupe contrôle ne montre aucune association. Ainsi, comme prévu, nous avons décidé de poursuivre l'étude en utilisant les seuils requis par une correction de Bonferroni pour tester les associations, mais également, de manière informelle, en comparant au groupe de gènes contrôles.

Lorsque nous avons groupé tous les variants selon leur gène de provenance et que nous les avons utilisés comme ensemble pour l'analyse d'effondrement, nous avons pu constater deux choses. Tout d'abord, un seul gène dépasse le seuil de significativité défini par une correction de Bonferroni, SBNO1. Par contre, en

comparant le profil de chacun des gènes associés avec les DNM et le profil des gènes du groupe seuil, il est possible de voir certaines régions d'association qui se distinguent, sans toutefois atteindre le niveau requis pour avoir des associations significatives. Ceci s'expliquerait facilement par le fait que notre étude, avec un échantillon de 240 individus pour chaque cohorte, n'est sans doute pas assez puissante pour bien distinguer les vraies associations du bruit expérimental. Cependant, les données actuelles suggèrent non seulement que ces gènes sont une intéressante piste à suivre pour les études futures, mais également que les variants rares sont probablement les meilleurs candidats pour expliquer où se trouve l'héritabilité manquante dans la SCZ¹⁷¹.

Lors de l'étude de séquençage de gène candidats, nous avons identifié un gène fortement associé à la maladie. Il s'agit du gène *SBNO1*, qui est défini comme étant un homologue du gène *notch strawberry* chez la drosophile. Malheureusement, très peu est connu sur la fonction de ce gène. Par contre, il a été démontré que SBNO1 est important pour le développement du cerveau chez le poisson zèbre; il semble réguler l'expression d'*otx2*, un acteur important de l'organisation du cerveau^{172,173}. Bien que ce gène semble être un excellent candidat, nous avons poussé plus loin les analyses. En fragmentant les analyses par exon d'appartenance plutôt que par gène, nous avons constaté que plusieurs exons du gène *SBNO1* présentaient une association. En analysant les patients porteurs de ces variants, nous avons pu remarqué que les patients porteurs des variants associés étaient toujours les mêmes. Ainsi, nous avons conclu à l'existence d'un haplotype ancestral, ce qui serait plausible puisque notre échantillon est composé d'une partie de patients et de contrôles canadiens-français. De sorte que le signal d'association statistique

pourrait en partie être dû à un effet de confusion apporté par une structure de population sous jacente. Cependant, il n'est pas non plus possible d'exclure un artéfact expérimental. Afin de statuer sur la question, nous avons conduit une étude de réplication par génotypage sur deux nouvelles cohortes et nous avons ainsi démontré que l'haplotype n'était pas lié à la maladie. Cette constatation nous permet de conclure que les études d'effondrement possèdent de fortes limitations lorsqu'il s'agit de travailler avec des variants liés, car cela ajoute un poids démesuré à l'ensemble testé. En outre, la fragmentation de l'ensemble de données en sets d'exons a également permis l'identification d'association avec les gènes EP300 et MAPK14. Il sera intéressant de voir si ces associations pourront être répliquées dans de plus grandes cohortes.

5.1.3 L'association avec l'âge parental

À la suite de l'étude DNM dans la SCZ, nous nous sommes mis en quête d'une explication plausible aux phénomènes observés pour la SCZ, mais aussi pour l'autisme et le retard mental. Un point en commun entre ces maladies, outre le fait qu'elles soient des maladies psychiatriques, est aussi qu'elles ont comme facteur de risque l'âge paternel^{159,174-176}. Une association a même été effectuée entre les risques de développer l'autisme et l'âge du grand-père paternel¹⁷⁷. L'un des mécanismes pouvant expliquer ceci pourrait être que le taux de DNM augmente graduellement avec l'âge du père. Ainsi, un père de vingt ans qui conçoit un enfant lui transmettrait moins de « nouvelles » mutations (mutations *de novo*) qu'un père concevant un enfant à l'âge de quarante ans. Cette hypothèse, si elle est vraie, serait capitale non seulement pour les maladies psychiatriques, mais également pour la santé publique en général. En effet, dans une population comme la nôtre où les hommes conçoivent

des enfants de plus en plus tardivement, il est raisonnable de se demander si cela pourrait avoir un impact sur le fardeau des maladies dans la population, ou si un dépistage prénatal pourrait être à considérer.

Pour être en mesure de tester cette hypothèse, nous avons décidé de travailler sur un échantillon de la population ne présentant aucune atteinte de santé sévère. Ce choix est facilement justifiable par le fait qu'il est préférable de caractériser un mécanisme dans des individus en santé et par la suite de le transposer dans un modèle de maladie plutôt que de commencer directement par le caractériser dans une population atteinte d'une maladie. Ainsi, nous avons choisi deux ensembles distincts; le premier constitué de jumeaux monozygotes avec de jeunes pères et le second constitué également de jumeaux monozygotes, mais où les pères étaient plus âgés à la conception. Nous avons choisi de travailler sur des familles de jumeaux, car cela permet à la fois de faciliter la détection des DNM et de faire la distinction entre les DNM somatiques et germinales. À l'instar de l'étude des DNM dans la SCZ, nous avons choisi la technologie nous permettant de maximiser le nombre de DNM identifiées par individu, dans ce cas-ci, le séquençage de génome complet (WGS).

À la suite du WGS de chacun des membres des dix familles, nous avons identifié un grand nombre de DNM. La première constatation est qu'il y avait une grande différence entre le nombre de DNM identifiées chez le groupe des jeunes pères et le groupe des vieux pères. Cette observation à elle seule permet de confirmer l'hypothèse que le taux de DNM est fortement modulé par l'âge des parents. Cependant, afin de mieux caractériser cette modulation, nous avons divisé les mutations par type (SNV, indel et CNV) et par origine (germinale et somatique). Nous avons pu montrer une forte corrélation entre les SNV et indels d'origine

germinale et l'âge parentale, mais pas pour les CNV, ce qui laisse croire à un mécanisme différent pour ces dernières. Cependant, en ce qui concerne les SNV et les indels germinales, l'effet est clair et très significatif. Dans nos conditions expérimentales, plus de la moitié de la variance du taux de DNM serait expliquée par l'âge parental.

L'une des limitations de notre étude est que bien que nous ayons pu conclure qu'une majorité de ces DNM provenaient directement d'un haplotype paternel, il n'est pas possible d'exclure que l'âge de la mère soit également associé avec la modulation du taux de DNM. En effet, dû à une homogamie (aussi appelée schéma d'accouplement non aléatoire) qui s'explique d'elle-même, nous n'avons pas pu sélectionner deux groupes pour lesquels l'âge maternel serait fixé. Ainsi, une corrélation significative de l'âge des pères avec l'âge des mères ne permet pas d'exclure la responsabilité de l'âge maternel, bien que l'effet semble beaucoup plus fort avec l'âge paternel.

Une autre observation de l'étude effectuée sur les familles de jumeaux est la présence de points chauds (hotspots) de DNM dans le génome humain. En effet, nous avons observé une distribution inégale et non aléatoire des DNM au travers du génome humain, ce qui nous permet de conclure à l'existence de zones « chaudes » et de zone « froides » de DNM dans le génome. Cette observation pourrait se révéler encore plus importante que l'association du taux de DNM avec l'âge parental, car elle signifierait que certaines régions du génome seraient plus promptes à accumuler des mutations indépendamment de la sélection évolutive.

À la suite des observations présentées précédemment, nous sommes retournés aux familles des patients atteints de SCZ étudiées dans le chapitre 2 de cet ouvrage et avons catalogué l'âge des parents. Le tableau 5.1 présente les résultats obtenus.

Bien qu'il soit difficile de conclure à un effet avec un si petit échantillon, on peut constater une différence entre l'âge moyen des pères entre les individus avec DNM et les individus sans DNM (Avec : 33 ± 2.38 , sans : 27 ± 4.2). Lorsque l'on retire les valeurs aberrantes (une dans chaque groupe), un simple test t de student permet de conclure à une différence significative entre les deux groupes (p -value = 0.0172). Ceci amène en effet un appui à l'hypothèse de DNM accrues chez les pères âgés, en

Familles	Nombre de DNM	Sexe	Age père	Age mère
SCZ0101	2	M	36	32
SCZ0201	2	F	33	24
SCZ0401	2	M	31	27
SCZ0601	1	M	35	34
SCZ0901	1	M	35	30
SCZ1001	3	M	31	27
SCZ1101	1	M	21	23
SCZ1401	3	M	30	27
SCZ0501	0	F	35	37
SCZ1301	0	M	23	24
SCZ0701	0	M	24	23
SCZ1201	0	F	26	22
SCZ1501	0	F	26	27
SCZ0801	0	F	28	28

Tableau 5.1 : Age à la conception des patients utilisés lors de l'étude DNM chez les patients atteints de SCZ (Chapitre 2)
concordance avec nos résultats précédents.

Il est à noter que tout récemment, deux études portant sur la caractérisation de l'effet de l'âge paternel ont démontré des résultats concordants avec notre étude. La première étude a caractérisé l'effet de l'âge du père dans de nombreuses familles, la plupart incluant des patients atteints de diverses maladies psychiatriques¹⁶⁰. Cette étude a été la première à montrer directement que l'âge paternel explique une majeure portion de la variation du taux de DNM. La deuxième étude a été réalisée

sur dix familles avec des patients autistes¹⁶¹. Fait intéressant, cette étude a également été réalisée sur des familles de jumeaux monozygotes. Ils ont non seulement montré l'effet de l'âge du père sur le taux de DNM, mais également l'existence de zones chaudes de DNM. Notre étude, quant à elle, recoupe certains aspects de ces deux études, mais elle est la seule étude ayant été réalisée sur des familles ne présentant aucun problème majeur de santé. Notre étude est également la première à s'étendre sur les CNV.

5.2 Les mutations *de novo* dans le contexte global de la schizophrénie

L'hypothèse de l'implication des DNM dans l'étiologie génétique de la SCZ ne fait toutefois pas l'unanimité. Contrairement à l'autisme ou au retard mental où cette hypothèse est presque globalement acceptée depuis plusieurs années, l'hypothèse des DNM fait lentement son chemin en SCZ, bien que plusieurs aspects de leur rôle soient encore à clarifier¹⁷⁸.

Bien que les GWAS ne puissent pas tenir en compte des DNM à cause de leurs fréquences très rares, il ne va pas sans dire que la poursuite des GWAS est un élément clé pour la SCZ. Il existe cependant des doutes raisonnables à savoir quelle est la taille d'un échantillon qui nous permettrait d'augmenter significativement le nombre de locus associés. En effet, un GWAS sur la taille des individus réalisée sur plus de 200,000 sujets, n'a révélé que 10% de l'hérédité de ce trait. Une approche intéressante serait de combiner les études DNM et les GWAS sur le même échantillon. Il serait intéressant de voir l'évolution des signaux d'association sur un échantillon pour lequel tous les patients porteurs d'une DNM qui semble être associé à la SCZ ont été retirés. Ceci permettrait sans doute de diminuer la

complexité phénotypique de l'échantillon et pourrait permettre une identification de plusieurs nouveaux locus.

Une autre question qui demeure en suspens est le pourcentage de l'héritabilité expliqué par les DNM. Pour cela, aucune étude n'existe pour le moment. Il faut donc se fier aux calculs effectués pour les GWAS. L'année dernière, le groupe à l'origine de la plus grande GWAS en SCZ a revisité les données utilisées pour cette même GWAS pour arriver à la conclusion que 23 % de la variation du risque de la SCZ serait expliqué par les SNP déjà capturés⁷⁶. Pour calculer cette valeur, les auteurs utilisent une combinaison de tous les SNP, qu'ils soient associés ou non. Bien que cette méthode puisse paraître contre-intuitive, les auteurs la justifie en comparant les résultats de SNP trouvés dans des gènes du système nerveux central (SNC) avec les SNP intergéniques ou les SNPs provenant des autres gènes non-SNC. On observe que les SNP du SNC expliquent presque autant de variance observée que les deux autres groupes, bien qu'ils soient moins nombreux et couvrent beaucoup moins de Mbp. Cependant, il serait intéressant de pousser l'observation plus loin : en effet, on remarque que les SNP dans des gènes non SNC expliquent autant d'héritabilité que les SNP intergéniques, ce qui semble très peu probable vu que la majorité des associations effectuées jusqu'à présent sont dans des gènes qui ne sont pas directement liés au SNC. Parallèlement, les auteurs démontrent qu'un modèle qui n'inclut seulement que les variants rares ne pourrait pas expliquer ces résultats. Ces conclusions semblent un peu hâtives, surtout lorsque l'on considère la faible sensibilité des plateformes de génotypage pour les variants rares.

Sur un même ton, un article de perspective paru récemment résume qu'il est difficile de conclure à un rôle clair des DNM pour les maladies psychiatriques et que

les études disponibles jusqu'à présent sont sans doute les précurseurs d'une multitude d'études à venir¹⁷⁸. Les auteurs soulèvent un point fort intéressant lorsqu'ils mentionnent la disparité des méthodes utilisées par les différentes études DNM. Le manque de méthodes communes rend difficile le fait de globaliser les conclusions de chacune des études et ainsi, de pouvoir tirer des conclusions par rapport à l'héritabilité expliquée par les DNM.

Une autre question qui a très peu été avancée mais qui le sera certainement à moyen terme concerne l'importance des variations génétiques non-codantes, plus particulièrement celles dans les régions régulatrices sur le mécanisme génétique de la maladie. Si jusqu'à présent, la majorité des mutations génétiques impliquées dans la maladies ont été identifiées dans les régions codantes, c'est peut-être simplement du au fait que ces régions ont été largement plus investis que les autres. Ou bien se pourrait-il que les variations génétiques régulatrices aient un rôle complémentaire aux variations codantes? Par exemple, les mutations codantes seraient responsables de l'apparition de la maladie et les variations régulatrices influenceraient le spectre des symptômes?

Finalement, une autre question encore non résolue est probablement celle ayant le plus d'impact à court terme; est-ce que les DNM qui sont liées aux maladies psychiatriques sont pleinement pénétrantes? En d'autres termes, est-ce que l'apparition de ces mutations suivra un profil mendélien ou non? Étant donné le manque de succès lors de l'identification de variants ségrégeant parfaitement dans une famille avec plusieurs membres atteints de SCZ, il est peu probable de penser que les DNM causant directement la schizophrénie soient les seules responsables de la maladie. On pourrait penser que ces mutations auraient un plus grand rapport de

côte que les SNP à fréquence commune ou intermédiaire, mais il est difficile de quantifier ces rapports de cote vu la taille encore petite des échantillons en SCZ.

5.3 La réconciliation des variants rares et des variants communs

En résumé, il serait erroné de s'attendre à ce que les variants communs ou rares expliquent à eux seuls la composante majoritaire du facteur héréditaire de la SCZ. La communauté scientifique s'entend pour émettre une hypothèse étiologique combinant à la fois différents types de variants génétiques. La figure 5.1 montre un modèle très simpliste de cette hypothèse.

La figure montre une combinaison des variants communs, rares, et *de novo*. Il est généralement attendu que l'effet de ces variants soit inversement proportionnel à

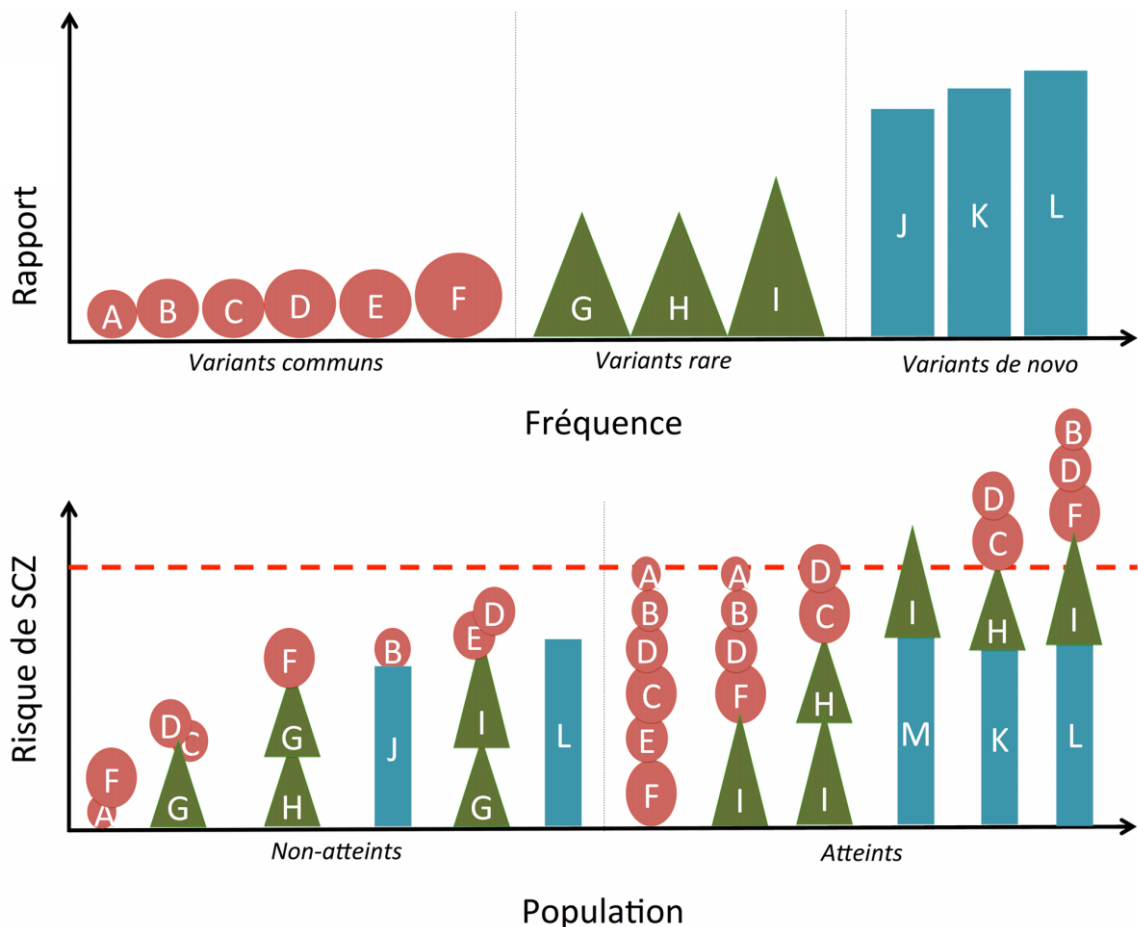


Figure 5-1 : Modèle de l'hypothèse des rapports de cotes

leur fréquence allélique¹⁷⁹. Cependant, des exceptions ont tout de même été répertoriées¹⁸⁰. Pour la schizophrénie, on note plusieurs exemples de variants communs associés à la maladie et quelques exemples de DNM. Par contre, il existe encore une zone grise concernant les variants rares. En effet, bien que leur implication ait largement été suggérée¹⁷¹, elle n'a pas encore été démontrée hors de tout doute. Le modèle présenté ici prendra donc en compte l'implication théorique des variants rares.

De par leur définition, les variants communs et rares se retrouveront plus au moins fréquemment dans la population générale. Les DNM quant à elles auront un profil unique, quoique la nouvelle découverte des zones « chaudes » pour les DNM pourrait laisser penser que certaines DNM pourraient se retrouver à une plus haute fréquence. Ainsi donc, chaque individu possèdera dans son génome un certain nombre de variants lié à la SCZ. C'est l'effet combiné de ces variants qui pourrait statuer sur l'état d'un individu relativement au risque de développer la maladie. L'augmentation du risque pour un individu n'est pas garante d'une certitude de développer la maladie. Il n'existe sans doute pas de « seuil de risque » pour la SCZ. Il est plus probable que ce soit une combinaison des risques génétiques et des risques environnementaux qui fait que certains individus développeront la maladie alors que d'autres individus seront sains. On pourrait également penser que l'accumulation des variants génétiques corrèlerait avec la sévérité des symptômes, mais aucune preuve expérimentale ne peut aujourd'hui supporter cette affirmation. Cependant, il semble clair, comme présenté dans la figure 5.1, qu'aucun variant, qu'il soit *de novo* ou commun, ne permettent à lui seul d'augmenter le risque d'une manière assez importante pour mener à une apparition certaine des symptômes de

la maladie. De ce fait, les études actuelles sont grandement limitées par leur capacité à étudier un modèle de rapport de cote composé.

Il faut également noter que le modèle présenté plus haut ne se concentre que sur les facteurs génétiques. Pour être scrupuleusement exact, il faudrait y inclure également les facteurs environnementaux et sociaux. Cela résulterait sans doute en un modèle d'une complexité désarmante. Voilà pourquoi il demeure de mise de se concentrer indépendamment sur chacun des facteurs et lorsque la base des connaissances sera suffisante, il sera possible de mieux estimer la contribution des facteurs génétiques comparés aux autres facteurs environnementaux.

5.4 La sélection égoïste et les maladies neuropsychiatriques

L'hypothèse de l'effet de l'âge paternel sur le taux de DNM prend tout son sens lorsque l'on revient aux mécanismes de gamétogenèse. En effet, chez la femme, les cellules germinales ne subiront que 22 divisions mitotiques avant de s'arrêter jusqu'à la puberté. Quelques minutes avant l'ovulation, quelques unes des cellules germinales subiront une division méiotique afin de compléter le processus de maturation des oocytes¹⁸¹. Chez l'homme, les cellules germinales subissent 30 divisions mitotiques durant l'embryogenèse et seront en état de latence jusqu'à la puberté. Une fois la spermatogenèse initiée durant la puberté, les spermatogonies entreront en division mitotique environ une fois toutes les 16 jours, ce qui fait qu'un homme dans la quarantaine aura déjà subi plus de 600 divisions cellulaires¹⁸². Ainsi, il semble logique de penser que les spermatogonies accumuleront davantage de mutations que les ovules.

Il a été discuté plus tôt dans cet ouvrage des gènes FGFR2 et FGFR3 qui causent un phénomène appelé la sélection égoïste des spermatozoïdes. Cette sélection semble être contrôlée par un processus qui ressemble au mécanisme de l'oncogenèse, avec entre autres l'implication de la voie de signalisation RAS¹²⁰. Tout récemment, l'un des groupes les plus actifs dans le domaine de la sélection égoïste des spermatozoïdes a publié une perspective intéressante pour les maladies psychiatriques¹⁸³. Ils ont en effet observé que plusieurs des gènes ou des signaux positivement associés à certaines maladies mentales comme l'autisme ou la SCZ semblent être impliqués dans la voie de signalisation RAS ou bien directement avec les gènes FGFR2 et FGFR3. Ainsi, ils suggèrent qu'à l'instar de FGFR2 et FGFR3 plusieurs des gènes associés à l'autisme et à la SCZ pourraient agir comme agent de sélection.

Cette hypothèse, bien qu'intéressante, n'intègre pas toutes les observations réalisées jusqu'à présent. En effet, selon cette hypothèse, les DNM observées dans la SCZ ne seraient que passagères et ne causeraient pas directement la maladie, ce qui ne cadre pas avec les profils pathogéniques que nous avons observés dans le chapitre 2 et le chapitre 4 de cet ouvrage. Ils ignorent également les récents rapports qui montrent que l'effet de l'âge du père augmente directement le taux de DNM. De plus, ce concept remettrait en question la théorie de la transmission des variants et serait contradictoire avec le principe même des GWAS. Néanmoins, l'idée d'une sélection au niveau des spermatozoïdes représente une notion intéressante et pourrait peut-être expliquer une partie des zones «chaudes» de DNM. De plus, il serait intéressant de voir comment une sélection des allèles au niveau des gamètes pourrait modifier notre compréhension de la sélection allélique au niveau des populations.

Bien que la littérature sur la mutagenèse dans les spermatozoïdes soit encore anémique, les observations réalisées sur les gènes *FGFR2* et *FGFR3* laissent présager que le taux de mutation ne serait pas nécessairement aléatoire. Si c'est effectivement le cas, l'apparition des zones chaudes de mutations ne serait alors qu'une formalité. En effet, une distribution non-aléatoire mènera directement à la création de regroupement de mutations. Connaissant la complexité du génome humain, il est finalement bien plus probable que l'apparition des DNM soit contrôlée par divers facteurs biologiques et soient ainsi non-aléatoire. Ainsi, même si les résultats observés au niveau des mutations dans le spermatozoïde s'avéraient non concluants, plusieurs autres processus biologiques pourraient contrôler la distribution des DNM. On peut par exemple penser au repliement de la chromatine. Il est assez logique de penser que selon la disponibilité de l'ADN, les mutations ne se produiront pas aux mêmes endroits. Il serait également intéressant de voir si les zones d'attachements des facteurs de réplifications sont épargnées par les DNM.

5.5 Perspectives

Lorsque l'on considère la génétique de la schizophrénie comme une ligne du temps, les progrès réalisés dans les derniers cinq ans sont phénoménaux. Que ce soit au niveau des GWAS, des études de SV ou bien des DNM, plusieurs associations ont été effectuées. Cependant, nous sommes encore loin de la ligne d'arrivée. Il serait erroné de croire qu'une seule solution pourra venir à bout du mystère du mécanisme génétique de la SCZ. En effet, l'un des points où toutes les études sont concordantes est que la SCZ est probablement l'une des maladies les plus complexes, au même niveau que le diabète, l'hypertension ou d'autres maladies très communes.

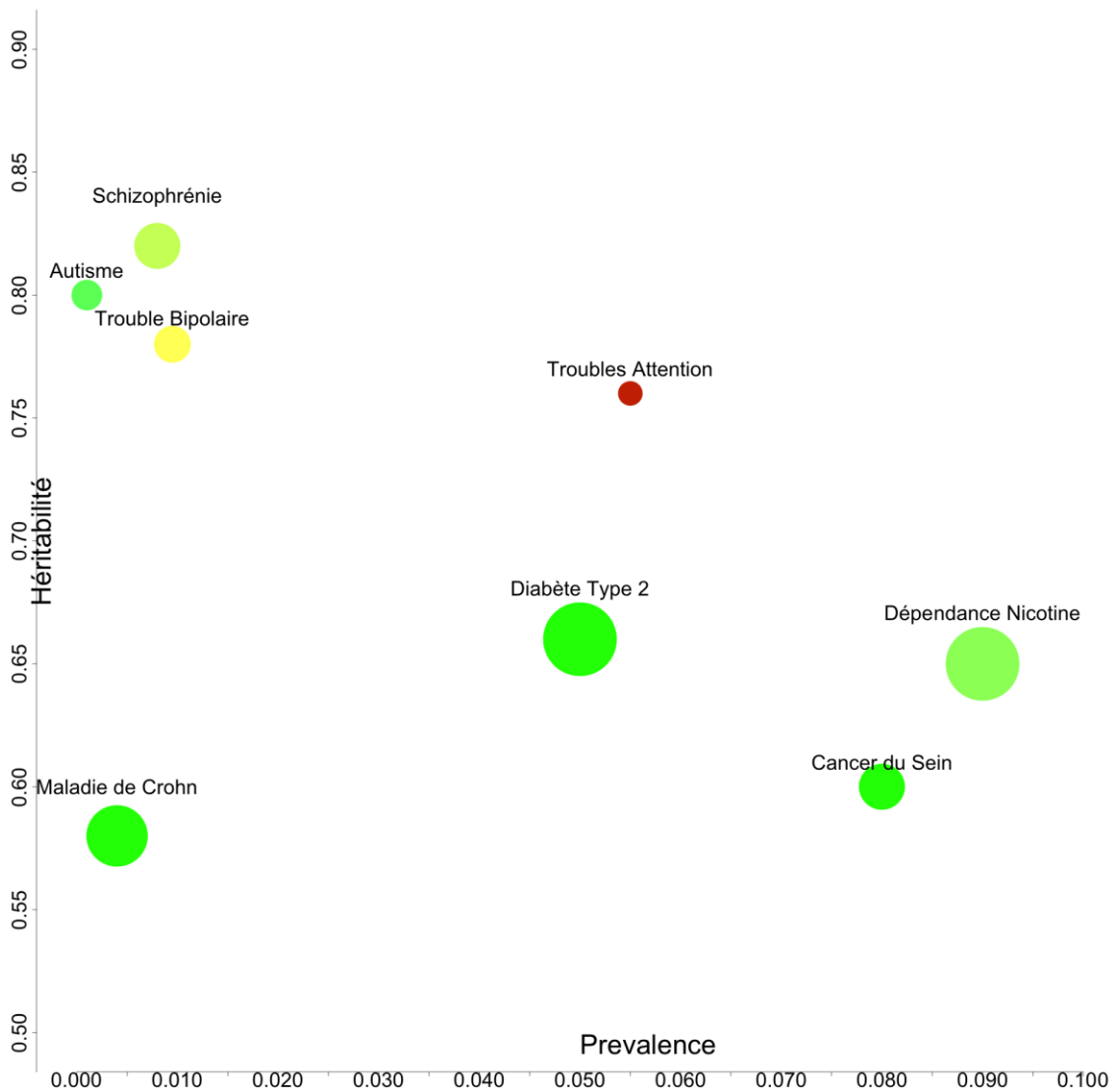


Figure 5-2 : **Comparaison de maladies psychiatriques avec d'autres maladies complexes à prévalence et héritabilité comparable.** La prévalence est présentée sur l'axe des X, alors que l'héritabilité se trouve sur l'axe des Y. La couleur des cercles représente le succès (gradation entre vert (succès) et rouge (échec) et la taille des cercles représente la taille de l'échantillon. Figure adapté de Sullivan et coll.

L'une des pistes pour une meilleure caractérisation génétique de la SCZ pourrait se cacher à même l'aspect clinique. En effet, pour les études génétiques, la SCZ est considérée comme un trait binaire; un individu est atteint ou il est sain. Cependant, cette affirmation fait souvent frémir les psychiatres qui considèrent souvent la SCZ comme une maladie présentant différents symptômes, chacun de ces symptômes pouvant être gradé sur un spectre. De plus, lorsque l'on compare la SCZ avec d'autres maladies complexes, on constate que le succès des études génétiques a été plus modeste que pour des maladies possédant une héritabilité moindre. Ce fait est

également observable pour d'autres maladies psychiatriques comme l'autisme, le trouble bipolaire et le trouble de déficit de l'attention (voir Figure 5-2 adapté de Sullivan et coll.). Il serait ainsi logique de penser que la plus grande hétérogénéité clinique des maladies psychiatriques pourrait en partie expliquer les plus petits succès. Ainsi, au lieu de continuellement augmenter la taille des cohortes, pourquoi ne pas utiliser l'approche inverse: une cohorte de patients atteints de SCZ ultracaractérisée divisée en sous-groupes selon des traits phénotypiques similaires. Cette approche réduirait d'un côté la puissance du nombre, car les cohortes s'en trouveraient ainsi fragmentées, mais l'hypothèse est que cette même fragmentation diminuerait l'hétérogénéité génétique et permettrait d'augmenter la puissance de détection des associations.

Une autre solution que l'on pourra bientôt considérer est l'utilisation d'analyses génétiques intégratives. Pour le moment, les limitations technologiques ne permettent pas d'effectuer des études de WGS sur une grande population. Lorsque ce sera le cas, il sera possible d'obtenir l'identification des variants communs, des variants rares, des DNM et des SV à l'aide d'un seul jeu de données. Ceci nous permettra donc de pouvoir caractériser l'effet de chacun des types de variants sur une même cohorte, de manière à établir avec précision le niveau d'implication des variants génétiques dans l'étiologie de la maladie.

L'émergence des tests d'associations avec des dizaines de gènes amène une nouvelle possibilité : l'utilisation d'une approche intégrative pour mettre en contexte tous les résultats génétiques. Il sera très intéressant de voir comment certains des gènes associés interagissent entre eux et de voir l'implication de certaines voies biologiques. L'établissement de réseaux d'interactions permettra à la fois de

renforcer la confiance envers les trouvailles génétiques, mais également de mieux comprendre les mécanismes biologiques derrière l'apparition de la SCZ. Ces connaissances biologiques seront un atout de taille dans le développement de nouvelles techniques de dépistages et de voies thérapeutiques novatrices.

Chapitre 6 Conclusion

La SCZ est une maladie très complexe et bien qu'elle ait été le sujet de nombreuses études, les mécanismes qui la soutiennent demeurent encore mal connus. Bien que l'héritabilité de la maladie ne soit plus remise en cause, il est encore difficile de déterminer l'apport respectif des mécanismes génétiques impliqués dans l'étiologie de la maladie.

Dans le cadre de cet ouvrage, nous nous sommes concentrés sur un type de variation génétique encore méconnu : les mutations *de novo* (DNM). En séquençant l'exome de 14 familles avec un patient atteint de schizophrénie, nous avons observé que les DNM semblent avoir un rôle significatif dans la maladie. En effet, non seulement le taux de mutations est-il plus élevé que dans d'autres populations contrôles, mais on observe aussi une grande proportion de mutations non-sens qui nous permet de conclure à un profil pathogénique. Ces conclusions nous ont permis d'orienter la suite de nos études sur les DNM.

Tout d'abord, nous nous sommes concentrés sur les gènes identifiés lors de la première étude. En séquençant ces gènes dans une plus grande cohorte, nous avons comparé le fardeau de mutations observé entre les patients et les contrôles. Nous avons non seulement observé un fardeau de mutations beaucoup plus élevé dans les gènes issus d'études DNM, mais nous avons également établi des associations individuelles qui nous permettent de conclure que les variants rares pourraient recéler une importante partie de l'héritabilité manquante de la maladie.

Ensuite, nous avons tenté de caractériser une observation bien connue dans le milieu : l'âge paternel est un facteur de risque de la SCZ. En utilisant une dizaine de familles de jumeaux monozygotiques sans problème de santé sévère, nous avons tenté de confirmer que l'âge du père est lié avec une augmentation du taux de mutations. Nous avons pu observer une forte corrélation entre les petites mutations germinales et l'âge du père qui nous permet de conclure que l'effet de l'âge du père explique plus de 70 % de la variation du taux de mutations *de novo*. En plus de cela, nous avons confirmé la présence de zones « chaudes » de DNM dans le génome, ce qui nous permet d'envisager une refonte complète des mécanismes d'apparition des variations génétiques dans un avenir rapproché.

En conclusion, nos travaux ont été pionniers et déterminants dans la caractérisation du rôle des DNM dans la SCZ, un domaine présentement en fulgurante expansion. Non seulement disposons-nous maintenant de nouvelles pistes pour tenter de résoudre ce mystère qu'est la SCZ, mais les trouvailles obtenues avec les DNM seront applicables dans d'autres maladies complexes et permettront à la génétique de faire un pas de plus l'avant.

Chapitre 7 Références

1. van Os J, Kapur S. Schizophrenia. *Lancet* 2009;374:635-45.
2. Saha S, Chant D, McGrath J. A Systematic Review of Mortality in Schizophrenia: Is the Differential Mortality Gap Worsening over Time? *Archives General Psychiatry* 2007;64:1123-31.
3. Schizophrenia. (Accessed March 05, 2013, at http://www.who.int/mental_health/management/schizophrenia/en/.)
4. Picchioni MM, Murray RM. Schizophrenia. *Bmj* 2007;335:91-5.
5. Eaton WW, Thara R, Federman B, Melton B, Liang KY. Structure and Course of Positive and Negative Symptoms in Schizophrenia. *Archives General Psychiatry* 1995;52:127-34.
6. Smith T, Weston C, Lieberman J. Schizophrenia (Maintenance Treatment). *American family physician* 2010;82:338-9.
7. Leucht S, Corves C, Arbter D, Engel RR, Li C, Davis JM. Second-Generation Versus First-Generation Antipsychotic Drugs for Schizophrenia: A Meta-Analysis. *Lancet* 2009;373:31-41.
8. Tyrer P, Kendall T. The Spurious Advance of Antipsychotic Drug Therapy. *Lancet* 2009;373:4-5.
9. Saha S, Chant D, Welham J, McGrath J. A Systematic Review of the Prevalence of Schizophrenia. *PLoS medicine* 2005;2:e141.
10. Castle D, Wessely S, Der G, Murray RM. The Incidence of Operationally Defined Schizophrenia in Camberwell, 1965-84. *British Journal Psychiatry* 1991;159:790-4.
11. Hafner H, Maurer K, Loffler W, Fatkenheuer B, an der Heiden W, Riecher-Rossler A, Behrens S, Gattaz WF. The Epidemiology of Early Schizophrenia. Influence of Age and Gender on Onset and Early Course. *British Journal Psychiatry Suppl* 1994:29-38.
12. Ayuso-Mateos JL. Global Burden of Schizophrenia in the Year 2000. 2000.

13. Clarke MC, Harley M, Cannon M. The Role of Obstetric Events in Schizophrenia. *Schizophrenia bulletin* 2006;32:3-8.
14. Davies G, Welham J, Chant D, Torrey EF, McGrath J. A Systematic Review and Meta-Analysis of Northern Hemisphere Season of Birth Studies in Schizophrenia. *Schizophrenia bulletin* 2003;29:587-93.
15. Handford HA. Brain Hypoxia, Minimal Brain Dysfunction, and Schizophrenia. *American Journal Psychiatry* 1975;132:192-4.
16. Prabakaran S, Swatton JE, Ryan MM, Huffaker SJ, Huang JT, Griffin JL, Wayland M, Freeman T, Dudbridge F, Lilley KS, Karp NA, Hester S, Tkachev D, Mimmack ML, Yolken RH, Webster MJ, Torrey EF, Bahn S. Mitochondrial Dysfunction in Schizophrenia: Evidence for Compromised Brain Metabolism and Oxidative Stress. *Molecular psychiatry* 2004;9:684-97, 43.
17. Pocock R, Hobert O. Oxygen Levels Affect Axon Guidance and Neuronal Migration in *Caenorhabditis Elegans*. *Nature neuroscience* 2008;11:894-900.
18. Schmidt-Kastner R, van Os J, H WMS, Schmitz C. Gene Regulation by Hypoxia and the Neurodevelopmental Origin of Schizophrenia. *Schizophrenia research* 2006;84:253-71.
19. Malaspina D. Paternal Factors and Schizophrenia Risk: De Novo Mutations and Imprinting. *Schizophrenia bulletin* 2001;27:379-93.
20. Brown AS, Hooton J, Schaefer CA, Zhang H, Petkova E, Babulas V, Perrin M, Gorman JM, Susser ES. Elevated Maternal Interleukin-8 Levels and Risk of Schizophrenia in Adult Offspring. *American Journal Psychiatry* 2004;161:889-95.
21. Cooper SJ. Schizophrenia after Prenatal Exposure to 1957 A2 Influenza Epidemic. *The British journal of psychiatry : the journal of mental science* 1992;161:394-6.
22. Mednick SA, Machon RA, Huttunen MO, Bonett D. Adult Schizophrenia Following Prenatal Exposure to an Influenza Epidemic. *Archives General Psychiatry* 1988;45:189-92.

23. Mortensen PB, Norgaard-Pedersen B, Waltoft BL, Sorensen TL, Hougaard D, Yolken RH. Early Infections of Toxoplasma Gondii and the Later Development of Schizophrenia. *Schizophrenia bulletin* 2007;33:741-4.
24. Conejero-Goldberg C, Torrey EF, Yolken RH. Herpesviruses and Toxoplasma Gondii in Orbital Frontal Cortex of Psychiatric Patients. *Schizophrenia research* 2003;60:65-9.
25. Erlenmeyer-Kimling L, Folnegovic Z, Hrabak-Zerjavic V, Borcic B, Folnegovic-Smalc V, Susser E. Schizophrenia and Prenatal Exposure to the 1957 A2 Influenza Epidemic in Croatia. *American Journal Psychiatry* 1994;151:1496-8.
26. Mino Y, Oshima I, Tsuda T, Okagami K. No Relationship between Schizophrenic Birth and Influenza Epidemics in Japan. *Journal of psychiatric research* 2000;34:133-8.
27. Swartz MS, Wagner HR, Swanson JW, Stroup TS, McEvoy JP, Canive JM, Miller DD, Reimherr F, McGee M, Khan A, Van Dorn R, Rosenheck RA, Lieberman JA. Substance Use in Persons with Schizophrenia: Baseline Prevalence and Correlates from the Nihm Catie Study. *The Journal of nervous and mental disease* 2006;194:164-72.
28. Arseneault L, Cannon M, Witton J, Murray RM. Causal Association between Cannabis and Psychosis: Examination of the Evidence. *The British journal of psychiatry : the journal of mental science* 2004;184:110-7.
29. Bossong MG, Niesink RJ. Adolescent Brain Maturation, the Endogenous Cannabinoid System and the Neurobiology of Cannabis-Induced Schizophrenia. *Progress in neurobiology* 2010;92:370-85.
30. Girard SL, Dion PA, Rouleau GA. Schizophrenia Genetics: Putting All the Pieces Together. *Current neurology and neuroscience reports* 2012;12:261-6.
31. Girard SL, Xiong L, Dion PA, Rouleau GA. Where Are the Missing Pieces of the Schizophrenia Genetics Puzzle? *Current opinion in genetics & development* 2011;21:310-6.
32. FISCHER M, HARVALD B, HAUGE M. A Danish Twin Study of Schizophrenia. *The British Journal of Psychiatry* 1969;115:981-90.

33. Kendler KS, Robinette CD. Schizophrenia in the National Academy of Sciences-National Research Council Twin Registry: A 16-Year Update. *The American journal of psychiatry* 1983;140:1551-63.
34. Onstad S, Skre I, Torgersen S, Kringlen E. Twin Concordance for Dsm-iii-R Schizophrenia. *Acta psychiatrica Scandinavica* 1991;83:395-401.
35. Cannon TD, Kaprio J, Lönnqvist J, Huttunen M, Koskenvuo M. The Genetic Epidemiology of Schizophrenia in a Finnish Twin Cohort. A Population-Based Modeling Study. *Archives of general psychiatry* 1998;55:67-74.
36. Cardno AG, Marshall EJ, Coid B, Macdonald AM, Ribchester TR, Davies NJ, Venturi P, Jones LA, Lewis SW, Sham PC, Gottesman II, Farmer AE, McGuffin P, Reveley AM, Murray RM. Heritability Estimates for Psychotic Disorders: The Maudsley Twin Psychosis Series. *Archives of general psychiatry* 1999;56:162-8.
37. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a Complex Trait: Evidence from a Meta-Analysis of Twin Studies. *Archives General Psychiatry* 2003;60:1187-92.
38. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21. *Science* 1990;250:1684-9.
39. Riley B. Linkage Studies of Schizophrenia. *Neurotoxicity research* 2004;6:17-34.
40. Pulver AE, Karayiorgou M, Wolyniec PS, Lasseter VK, Kasch L, Nestadt G, Antonarakis S, Housman D, Kazazian HH, Meyers D. Sequential Strategy to Identify a Susceptibility Gene for Schizophrenia: Report of Potential Linkage on Chromosome 22q12-Q13.1: Part 1. *American journal of medical genetics* 1994;54:36-43.
41. Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, Gunnarsdottir S, Ivarsson O, Chou TT, Hjaltason O, Birgisdottir B, Jonsson H, Gudnadottir VG, Gudmundsdottir E, Bjornsson A, Ingvarsson B, Ingason A, Sigfusson S, Hardardottir H, Harvey RP, Lai D, Zhou M, Brunner D, Mutel V, Gonzalo A, Lemke G, Sainz J, Johannesson G, Andresson T,

- Gudbjartsson D, Manolescu A, Frigge ML, Gurney ME, Kong A, Gulcher JR, Petursson H, Stefansson K. Neuregulin 1 and Susceptibility to Schizophrenia. *American journal of human genetics* 2002;71:877-92.
42. Petryshen TL, Middleton FA, Kirby A, Aldinger KA, Purcell S, Tahl AR, Morley CP, McGann L, Gentile KL, Rockwell GN, Medeiros HM, Carvalho C, Macedo A, Dourado A, Valente J, Ferreira CP, Patterson NJ, Azevedo MH, Daly MJ, Pato CN, Pato MT, Sklar P. Support for Involvement of Neuregulin 1 in Schizophrenia Pathophysiology. *Molecular psychiatry* 2005;10:366-74, 28.
43. Seshadri S, Kamiya A, Yokota Y, Prikulis I, Kano S-i, Hayashi-Takagi A, Stanco A, Eom T-Y, Rao S, Ishizuka K, Wong P, Korth C, Anton ES, Sawa A. Disrupted-in-Schizophrenia-1 Expression Is Regulated by Beta-Site Amyloid Precursor Protein Cleaving Enzyme-1-Neuregulin Cascade. *PNAS* 2010;107:5622-7.
44. Straub RE, Jiang Y, MacLean CJ, Ma Y, Webb BT, Myakishev MV, Harris-Kerr C, Wormley B, Sadek H, Kadambi B, Cesare AJ, Gibberman A, Wang X, O'Neill FA, Walsh D, Kendler KS. Genetic Variation in the 6p22.3 Gene *Dtnbp1*, the Human Ortholog of the Mouse *Dysbindin* Gene, Is Associated with Schizophrenia. *American Journal Human Genetics* 2002;71:337-48.
45. Bray NJ, Preece A, Williams NM, Moskvina V, Buckland PR, Owen MJ, O'Donovan MC. Haplotypes at the *Dystrobrevin Binding Protein 1 (Dtnbp1)* Gene Locus Mediate Risk for Schizophrenia through Reduced *Dtnbp1* Expression. *Human Molecular Genetics* 2005;14:1947-54.
46. St Clair D, Blackwood D, Muir W, Carothers A, Walker M, Spowart G, Gosden C, Evans HJ. Association within a Family of a Balanced Autosomal Translocation with Major Mental Illness. *Lancet* 1990;336:13-6.
47. Blackwood DH, Fordyce A, Walker MT, St Clair DM, Porteous DJ, Muir WJ. Schizophrenia and Affective Disorders--Cosegregation with a Translocation at Chromosome 1q42 That Directly Disrupts Brain-Expressed Genes: Clinical and P300 Findings in a Family. *American journal of human genetics* 2001;69:428-33.
48. Brandon NJ, Millar JK, Korth C, Sive H, Singh KK, Sawa A. Understanding the Role of *Disc1* in Psychiatric Disease and During Normal Development. *The*

Journal of neuroscience : the official journal of the Society for Neuroscience
2009;29:12768-75.

49. International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M,

- Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. A Second Generation Human Haplotype Map of over 3.1 Million Snps. *Nature* 2007;449:851-61.
50. Zhang Y, De S, Garner JR, Smith K, Wang SA, Becker KG. Systematic Analysis, Comparison, and Integration of Disease Based Human Genetic Association Data and Mouse Genetic Phenotypic Information. *BMC medical genomics* 2010;3:1.
51. Kirov G, Zaharieva I, Georgieva L, Moskvina V, Nikolov I, Cichon S, Hillmer A, Toncheva D, Owen MJ, O'Donovan MC. A Genome-Wide Association Study in 574 Schizophrenia Trios Using DNA Pooling. *Molecular psychiatry* 2009;14:796-803.
52. Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, Liu W, Downing AM, Lieberman J, Close SL. Genomewide Association for Schizophrenia in the Catie Study: Results of Stage 1. *Molecular psychiatry* 2008;13:570-84.
53. Shifman S, Johannesson M, Bronstein M, Chen SX, Collier DA, Craddock NJ, Kendler KS, Li T, O'Donovan M, O'Neill FA, Owen MJ, Walsh D, Weinberger DR, Sun C, Flint J, Darvasi A. Genome-Wide Association Identifies a Common Variant in the Reelin Gene That Increases the Risk of Schizophrenia Only in Women. *PLoS Genetics* 2008;4:e28.
54. Guidotti A, Auta J, Davis JM, Di-Giorgi-Gerevini V, Dwivedi Y, Grayson DR, Impagnatiello F, Pandey G, Pesold C, Sharma R, Uzunov D, Costa E. Decrease in Reelin and Glutamic Acid Decarboxylase67 (Gad67) Expression in Schizophrenia and Bipolar Disorder: A Postmortem Brain Study. *Archives General Psychiatry* 2000;57:1061-9.
55. Impagnatiello F, Guidotti AR, Pesold C, Dwivedi Y, Caruncho H, Pisu MG, Uzunov DP, Smalheiser NR, Davis JM, Pandey GN, Pappas GD, Tueting P, Sharma RP, Costa E. A Decrease of Reelin Expression as a Putative Vulnerability Factor in Schizophrenia. *PNAS* 1998;95:15718-23.

56. Dong E, Agis-Balboa RC, Simonini MV, Grayson DR, Costa E, Guidotti A. Reelin and Glutamic Acid Decarboxylase67 Promoter Remodeling in an Epigenetic Methionine-Induced Mouse Model of Schizophrenia. *PNAS* 2005;102:12578-83.
57. Grayson DR, Jia X, Chen Y, Sharma RP, Mitchell CP, Guidotti A, Costa E. Reelin Promoter Hypermethylation in Schizophrenia. *PNAS* 2005;102:9341-6.
58. Mill J, Tang T, Kaminsky Z, Khare T, Yazdanpanah S, Bouchard L, Jia P, Assadzadeh A, Flanagan J, Schumacher A, Wang SC, Petronis A. Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis. *American journal human genetics* 2008;82:696-711.
59. Tochigi M, Iwamoto K, Bundo M, Komori A, Sasaki T, Kato N, Kato T. Methylation Status of the Reelin Promoter Region in the Brain of Schizophrenic Patients. *Biological psychiatry* 2008;63:530-3.
60. O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I, Hamshere M, Carroll L, Georgieva L, Dwyer S, Holmans P, Marchini JL, Spencer CCA, Howie B, Leung H-T, Hartmann AM, Möller H-J, Morris DW, Shi Y, Feng G, Hoffmann P, Propping P, Vasilescu C, Maier W, Rietschel M, Zammit S, Schumacher J, Quinn EM, Schulze TG, Williams NM, Giegling I, Iwata N, Ikeda M, Darvasi A, Shifman S, He L, Duan J, Sanders AR, Levinson DF, Gejman PV, Cichon S, Nöthen MM, Gill M, Corvin A, Rujescu D, Kirov G, Owen MJ, Buccola NG, Mowry BJ, Freedman R, Amin F, Black DW, Silverman JM, Byerley WF, Cloninger CR. Identification of Loci Associated with Schizophrenia by Genome-Wide Association and Follow-Up. *Nature Genetics* 2008;40:1053-5.
61. Riley B, Thiselton D, Maher BS, Bigdeli T, Wormley B, McMichael GO, Fanous AH, Vladimirov V, O'Neill FA, Walsh D, Kendler KS. Replication of Association between Schizophrenia and Znf804a in the Irish Case-Control Study of Schizophrenia Sample. *Molecular psychiatry* 2010;15:29-37.
62. Steinberg S, Mors O, Borglum AD, Gustafsson O, Werge T, Mortensen PB, Andreassen OA, Sigurdsson E, Thorgeirsson TE, Bottcher Y, Olason P, Ophoff RA, Cichon S, Gudjonsdottir IH, Pietilainen OP, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Athanasiu L, Suvisaari J, Lonnqvist J, Paunio T,

- Hartmann A, Jurgens G, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Breuer R, Moller HJ, Giegling I, Glenthøj B, Rasmussen HB, Mattheisen M, Bitter I, Rethelyi JM, Sigmundsson T, Fossdal R, Thorsteinsdottir U, Ruggeri M, Tosato S, Strengman E, Kiemenev LA, Melle I, Djurovic S, Abramova L, Kaleda V, Walshe M, Bramon E, Vassos E, Li T, Fraser G, Walker N, Toulopoulou T, Yoon J, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Jonsson EG, Terenius L, Agartz I, Petursson H, Nothen MM, Rietschel M, Peltonen L, Rujescu D, Collier DA, Stefansson H, St Clair D, Stefansson K. Expanding the Range of Znf804a Variants Conferring Risk of Psychosis. *Molecular psychiatry* 2010.
63. Williams HJ, Norton N, Dwyer S, Moskvina V, Nikolov I, Carroll L, Georgieva L, Williams NM, Morris DW, Quinn EM, Giegling I, Ikeda M, Wood J, Lencz T, Hultman C, Lichtenstein P, Thiselton D, Maher BS, Malhotra AK, Riley B, Kendler KS, Gill M, Sullivan P, Sklar P, Purcell S, Nimgaonkar VL, Kirov G, Holmans P, Corvin A, Rujescu D, Craddock N, Owen MJ, O'Donovan MC. Fine Mapping of Znf804a and Genome-Wide Significant Evidence for Its Involvement in Schizophrenia and Bipolar Disorder. *Molecular psychiatry* 2010.
64. Hill MJ, Jeffries AR, Dobson RJ, Price J, Bray NJ. Knockdown of the Psychosis Susceptibility Gene Znf804a Alters Expression of Genes Involved in Cell Adhesion. *Human Molecular Genetics* 2012;21:1018-24.
65. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietilainen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Borglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Bottcher Y, Olesen J, Breuer R, Moller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Rethelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemenev LA, Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jonsson EG,

- Terenius L, Agartz I, Petursson H, Nothen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA. Common Variants Conferring Risk of Schizophrenia. *Nature* 2009;460:744-7.
66. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder. *Nature* 2009;460:748-52.
67. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whitemore AS, Mowry BJ, Olincy A, Amin F, Cloninger CR, Silverman JM, Buccola NG, Byerley WF, Black DW, Crowe RR, Oksenberg JR, Mirel DB, Kendler KS, Freedman R, Gejman PV. Common Variants on Chromosome 6p22.1 Are Associated with Schizophrenia. *Nature* 2009;460:753-7.
68. Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, Hovatta I, Williams NM, Schwab SG, Pulver AE, Faraone SV, Brzustowicz LM, Kaufmann CA, Garver DL, Gurling HM, Lindholm E, Coon H, Moises HW, Byerley W, Shaw SH, Mesen A, Sherrington R, O'Neill FA, Walsh D, Kendler KS, Ekelund J, Paunio T, Lonnqvist J, Peltonen L, O'Donovan MC, Owen MJ, Wildenauer DB, Maier W, Nestadt G, Blouin JL, Antonarakis SE, Mowry BJ, Silverman JM, Crowe RR, Cloninger CR, Tsuang MT, Malaspina D, Harkavy-Friedman JM, Svrakic DM, Bassett AS, Holcomb J, Kalsi G, McQuillin A, Brynjolfson J, Sigmundsson T, Petursson H, Jazin E, Zoega T, Helgason T. Genome Scan Meta-Analysis of Schizophrenia and Bipolar Disorder, Part Ii: Schizophrenia. *American journal human genetics* 2003;73:34-48.
69. Brown AS. Prenatal Infection as a Risk Factor for Schizophrenia. *Schizophrenia bulletin* 2006;32:200-2.
70. Brown AS. The Environment and Susceptibility to Schizophrenia. *Progress in neurobiology* 2010.
71. Shi Y, Li Z, Xu Q, Wang T, Li T, Shen J, Zhang F, Chen J, Zhou G, Ji W, Li B, Xu Y, Liu D, Wang P, Yang P, Liu B, Sun W, Wan C, Qin S, He G, Steinberg S, Cichon S, Werge T, Sigurdsson E, Tosato S, Palotie A, Nöthen MM, Rietschel M, Ophoff RA, Collier DA, Rujescu D, Clair DS, Stefansson H, Stefansson K, Ji J, Wang Q, Li W,

- Zheng L, Zhang H, Feng G, He L. Common Variants on 8p12 and 1q24.2 Confer Risk of Schizophrenia. *Nature Genetics* 2011;43:1224-7.
72. Yue W-H, Wang H-F, Sun L-D, Tang F-L, Liu Z-H, Zhang H-X, Li W-Q, Zhang Y-L, Zhang Y, Ma C-C, Du B, Wang L-F, Ren Y-Q, Yang Y-F, Hu X-F, Wang Y, Deng W, Tan L-W, Tan Y-L, Chen Q, Xu G-M, Yang G-G, Zuo X-b, Yan H, Ruan Y-Y, Lu T-L, Han X, Ma X-H, Wang Y, Cai L-W, Jin C, Zhang H-Y, Yan J, Mi W-F, Yin X-Y, Ma W-B, Liu Q, Kang L, Sun W, Pan C-Y, Shuang M, Yang F-D, Wang C-Y, Yang J-L, Li K-Q, Ma X, Li L-J, Yu X, Li Q-Z, Huang X, Lv L-X, Li T, Zhao G-P, Huang W, Zhang X-J, Zhang D. Genome-Wide Association Study Identifies a Susceptibility Locus for Schizophrenia in Han Chinese at 11p11.2. *Nature Genetics* 2011;43:1228-31.
73. Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin D-Y, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, St. Clair D, Corvin A, Gurling H, Werge T, Rujescu D, Blackwood DHR, Pato CN, Malhotra AK, Purcell S, Dudbridge F, Neale BM, Rossin L, Visscher PM, Posthuma D, Ruderfer DM, Fanous A, Stefansson H, Steinberg S, Mowry BJ, Golimbet V, De Hert M, Jönsson EG, Bitter I, Pietiläinen OPH, Collier DA, Tosato S, Agartz I, Albus M, Alexander M, Amdur RL, Amin F, Bass N, Bergen SE, Black DW, Børglum AD, Brown MA, Bruggeman R, Buccola NG, Byerley WF, Cahn W, Cantor RM, Carr VJ, Catts SV, Choudhury K, Cloninger CR, Cormican P, Craddock N, Danoy PA, Datta S, de Haan L, Demontis D, Dikeos D, Djurovic S, Donnelly P, Donohoe G, Duong L, Dwyer S, Fink-Jensen A, Freedman R, Freimer NB, Friedl M, Georgieva L, Giegling I, Gill M, Glenthøj B, Godard S, Hamshere M, Hansen M, Hansen T, Hartmann AM, Henskens FA, Hougaard DM, Hultman CM, Ingason A, Jablensky AV, Jakobsen KD, Jay M, Jürgens G, Kahn RS, Keller MC, Kenis G, Kenny E, Kim Y, Kirov GK, Konnerth H, Konte B, Krabbendam L, Krasucki R, Lasseter VK, Laurent C, Lawrence J, Lencz T, Lerer FB, Liang K-Y, Lichtenstein P, Lieberman JA, Linszen DH, Lönnqvist J, Loughland CM, Maclean AW, Maher BS, Maier W, Mallet J, Malloy P, Mattheisen M, Mattingsdal M, McGhee KA, McGrath JJ, McIntosh A, McLean DE, McQuillin A, Melle I, Michie PT, Milanova V, Morris DW, Mors O, Mortensen PB, Moskvina V, Muglia P, Myin-Germeys I, Nertney DA, Nestadt G, Nielsen J, Nikolov I, Nordentoft M, Norton N, Nöthen MM, O'Dushlaine CT, Olincy A, Olsen L, O'Neill FA, Ørntoft TF, Owen MJ, Pantelis C,

Papadimitriou G, Pato MT, Peltonen L, Petursson H, Pickard B, Pimm J, Pulver AE, Puri V, Quested D, Quinn EM, Rasmussen HB, Réthelyi JM, Ribble R, Rietschel M, Riley BP, Ruggeri M, Schall U, Schulze TG, Schwab SG, Scott RJ, Shi J, Sigurdsson E, Silverman JM, Spencer CCA, Stefansson K, Strange A, Strengman E, Stroup TS, Suvisaari J, Terenius L, Thirumalai S, Thygesen JH, Timm S, Toncheva D, van den Oord E, van Os J, van Winkel R, Veldink J, Walsh D, Wang AG, Wiersma D, Wildenauer DB, Williams HJ, Williams NM, Wormley B, Zammit S, Sullivan PF, O'Donovan MC, Daly MJ, Gejman PV. Genome-Wide Association Study Identifies Five New Schizophrenia Loci. *Nature Genetics* 2011;43:969-76.

74. Smrt RD, Szulwach KE, Pfeiffer RL, Li X, Guo W, Pathania M, Teng ZQ, Luo Y, Peng J, Bordey A, Jin P, Zhao X. MicroRNA Mir-137 Regulates Neuronal Maturation by Targeting Ubiquitin Ligase Mind Bomb-1. *Stem cells* 2010;28:1060-70.
75. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Juntila M, Kaplan LM, Kettunen J, König IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Müller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer

NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiraanen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeny L, Kocher T, Launer LJ, Lehtimäki T, Melander O, Mosley TH, Jr., Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllenstein U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN. Hundreds of Variants Clustered in Genomic Loci and Biological Pathways Affect Human Height. *Nature* 2010;467:832-8.

76. Lee SH, DeCandia TR, Ripke S, Yang J, Schizophrenia Psychiatric Genome-Wide Association Study C, International Schizophrenia C, Molecular Genetics of Schizophrenia C, Sullivan PF, Goddard ME, Keller MC, Visscher PM, Wray NR.

Estimating the Proportion of Variation in Susceptibility to Schizophrenia Captured by Common Snps. *Nature Genetics* 2012;44:247-50.

77. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent

WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing C. Initial Sequencing and Analysis of the Human Genome. *Nature* 2001;409:860-921.

78. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews TD, Arbury H, Attwood A, Auton A, Ball SG, Balmforth AJ, Barrett JC, Barroso I, Barton A, Bennett AJ, Bhaskar S, Blaszczyk K, Bowes J, Brand OJ, Braund PS, Bredin F, Breen G, Brown MJ, Bruce IN, Bull J, Burren OS, Burton J, Byrnes J, Caesar S, Clee CM, Coffey AJ, Connell JM, Cooper JD, Dominiczak AF, Downes K, Drummond HE, Dudakia D, Dunham A, Ebbs B, Eccles D, Edkins S, Edwards C, Elliot A, Emery P, Evans DM, Evans G, Eyre S, Farmer A, Ferrier IN, Feuk L, Fitzgerald T, Flynn E, Forbes A, Forty L, Franklyn JA, Freathy RM, Gibbs P, Gilbert P, Gokumen O, Gordon-Smith K, Gray E, Green E, Groves CJ, Grozeva D, Gwilliam R, Hall A, Hammond N, Hardy M, Harrison P, Hassanali N, Hebaishi H, Hines S, Hinks A, Hitman GA, Hocking L, Howard E, Howard P, Howson JM, Hughes D, Hunt S, Isaacs JD, Jain M, Jewell DP, Johnson T, Jolley JD, Jones IR, Jones LA, Kirov G, Langford CF, Lango-Allen H, Lathrop GM, Lee J, Lee KL, Lees C, Lewis K, Lindgren CM, Maisuria-Armer M, Maller J, Mansfield J, Martin P, Massey DC, McArdle WL, McGuffin P, McLay KE, Mentzer A, Mimmack ML, Morgan AE, Morris AP, Mowat C, Myers S, Newman W, Nimmo ER, O'Donovan MC, Onipinla A, Onyiah I, Ovington NR, Owen MJ, Palin K, Parnell K, Pernet D, Perry JR, Phillips A, Pinto D, Prescott NJ, Prokopenko I, Quail MA, Rafelt S, Rayner NW, Redon R, Reid DM, Renwick, Ring SM, Robertson N, Russell E, St Clair D, Sambrook JG, Sanderson JD, Schuilenburg H, Scott CE, Scott R, Seal S, Shaw-Hawkins S, Shields BM, Simmonds MJ, Smyth DJ, Somaskantharajah E, Spanova K, Steer S, Stephens J, Stevens HE, Stone MA, Su Z, Symmons DP, Thompson JR, Thomson W, Travers ME, Turnbull C, Valsesia A, Walker M, Walker NM,

- Wallace C, Warren-Perry M, Watkins NA, Webster J, Weedon MN, Wilson AG, Woodburn M, Wordsworth BP, Young AH, Zeggini E, Carter NP, Frayling TM, Lee C, McVean G, Munroe PB, Palotie A, Sawcer SJ, Scherer SW, Strachan DP, Tyler-Smith C, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Gough SC, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand W, Parkes M, Rahman N, Todd JA, Samani NJ, Donnelly P. Genome-Wide Association Study of Cnvs in 16,000 Cases of Eight Common Diseases and 3,000 Shared Controls. *Nature* 2010;464:713-20.
79. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* 2008;320:539-43.
80. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong Association of De Novo Copy Number Mutations with Sporadic Schizophrenia. *Nature Genetics* 2008;40:880-5.
81. International Schizophrenia C. Rare Chromosomal Deletions and Duplications Increase Risk of Schizophrenia. *Nature* 2008;455:237-41.
82. Bassett AS, Marshall CR, Lionel AC, Chow EW, Scherer SW. Copy Number Variations and Risk for Schizophrenia in 22q11.2 Deletion Syndrome. *Human Molecular Genetics* 2008;17:4045-53.
83. Karayiorgou M, Morris MA, Morrow B, Shprintzen RJ, Goldberg R, Borrow J, Gos A, Nestadt G, Wolyniec PS, Lasseter VK, al. e. Schizophrenia Susceptibility Associated with Interstitial Deletions of Chromosome 22q11. *PNAS* 1995;92:7612-6.
84. Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV,

Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller H-J, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Touloupoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemenev LA, Franke B, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K. Large Recurrent Microdeletions Associated with Schizophrenia. *Nature* 2008;455:232-6.

85. Brunetti-Pierri N, Berg JS, Scaglia F, Belmont J, Bacino CA, Sahoo T, Lalani SR, Graham B, Lee B, Shinawi M, Shen J, Kang SH, Pursley A, Lotze T, Kennedy G, Lansky-Shafer S, Weaver C, Roeder ER, Grebe TA, Arnold GL, Hutchison T, Reimschisel T, Amato S, Geraghty MT, Innis JW, Obersztyn E, Nowakowska B, Rosengren SS, Bader PI, Grange DK, Naqvi S, Garnica AD, Bernes SM, Fong CT, Summers A, Walters WD, Lupski JR, Stankiewicz P, Cheung SW, Patel A. Recurrent Reciprocal 1q21.1 Deletions and Duplications Associated with Microcephaly or Macrocephaly and Developmental and Behavioral Abnormalities. *Nature Genetics* 2008;40:1466-71.
86. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, Krause V, Kumar RA, Grozeva D, Malhotra D, Walsh T, Zackai EH, Kaplan P, Ganesh J, Krantz ID, Spinner NB, Roccanova P, Bhandari A, Pavon K, Lakshmi B, Leotta A, Kendall J, Lee YH, Vacic V, Gary S, Iakoucheva LM, Crow TJ, Christian SL, Lieberman JA, Stroup TS, Lehtimäki T, Puura K, Haldeman-Englert C, Pearl J, Goodell M, Willour VL, Derosse P, Steele J, Kassem L, Wolff J, Chitkara N, McMahon FJ, Malhotra AK, Potash JB, Schulze TG, Nothen MM, Cichon S, Rietschel M, Leibenluft E, Kustanovich V, Lajonchere CM, Sutcliffe JS, Skuse D, Gill M, Gallagher L, Mendell NR, Craddock N, Owen MJ, O'Donovan MC, Shaikh TH, Susser E, Delisi LE, Sullivan PF, Deutsch CK, Rapoport J, Levy DL, King MC, Sebat J.

- Microduplications of 16p11.2 Are Associated with Schizophrenia. *Nature Genetics* 2009;41:1223-7.
87. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control C, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. Origins and Functional Impact of Copy Number Variation in the Human Genome. *Nature* 2010;464:704-12.
 88. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. Mapping and Sequencing of Structural Variation from Eight Human Genomes. *Nature* 2008;453:56-64.
 89. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. Global Variation in Copy Number in the Human Genome. *Nature* 2006;444:444-54.
 90. Dwyer S, Williams H, Holmans P, Moskvina V, Craddock N, Owen MJ, O'Donovan MC. No Evidence That Rare Coding Variants in Znf804a Confer Risk of Schizophrenia. *American journal medical genetics B* 2010;153B:1411-6.
 91. Green EK, Grozeva D, Sims R, Raybould R, Forty L, Gordon-Smith K, Russell E, St Clair D, Young AH, Ferrier IN, Kirov G, Jones I, Jones L, Owen MJ, O'Donovan MC, Craddock N. Disc1 Exon 11 Rare Variants Found More Commonly in

- Schizoaffective Spectrum Cases Than Controls. *American journal medical genetics B* 2011;156B:490-2.
92. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IH, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Genomes Project C, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* 2012;335:823-8.
 93. Amir RE, Van den Veyver IB, Wan M, Tran CQ, Francke U, Zoghbi HY. Rett Syndrome Is Caused by Mutations in X-Linked *Mecp2*, Encoding Methyl-Cpg-Binding Protein 2. *Nature Genetics* 1999;23:185-8.
 94. Rouleau GA, Merel P, Lutchman M, Sanson M, Zucman J, Marineau C, Hoang-Xuan K, Demczuk S, Desmaze C, Plougastel B, et al. Alteration in a New Gene Encoding a Putative Membrane-Organizing Protein Causes Neuro-Fibromatosis Type 2. *Nature* 1993;363:515-21.
 95. Gauthier J, Champagne N, Lafreniere RG, Xiong L, Spiegelman D, Brustein E, Lapointe M, Peng H, Cote M, Noreau A, Hamdan FF, Addington AM, Rapoport JL, Delisi LE, Krebs MO, Joobar R, Fathalli F, Mouaffak F, Haghghi AP, Neri C, Dube MP, Samuels ME, Marineau C, Stone EA, Awadalla P, Barker PA, Carbonetto S, Drapeau P, Rouleau GA. De Novo Mutations in the Gene Encoding the Synaptic Scaffolding Protein Shank3 in Patients Ascertained for Schizophrenia. *PNAS* 2010;107:7863-8.
 96. Gauthier J, Siddiqui TJ, Huashan P, Yokomaku D, Hamdan FF, Champagne N, Lapointe M, Spiegelman D, Noreau A, Lafreniere RG, Fathalli F, Joobar R, Krebs MO, DeLisi LE, Mottron L, Fombonne E, Michaud JL, Drapeau P, Carbonetto S, Craig AM, Rouleau GA. Truncating Mutations in *Nrxn2* and *Nrxn1* in Autism Spectrum Disorders and Schizophrenia. *Hum Genet* 2011;130:563-73.

97. Awadalla P, Gauthier J, Myers RA, Casals F, Hamdan FF, Griffing AR, Cote M, Henrion E, Spiegelman D, Tarabeux J, Piton A, Yang Y, Boyko A, Bustamante C, Xiong L, Rapoport JL, Addington AM, DeLisi JL, Krebs MO, Joobler R, Millet B, Fombonne E, Mottron L, Zilvermit M, Keebler J, Daoud H, Marineau C, Roy-Gagnon MH, Dube MP, Eyre-Walker A, Drapeau P, Stone EA, Lafreniere RG, Rouleau GA. Direct Measure of the De Novo Mutation Rate in Autism and Schizophrenia Cohorts. *Am J Hum Genet* 2010;87:316-24.
98. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, Dionne-Laporte A, Spiegelman D, Henrion E, Diallo O, Thibodeau P, Bachand I, Bao JY, Tong AH, Lin CH, Millet B, Jaafari N, Joobler R, Dion PA, Lok S, Krebs MO, Rouleau GA. Increased Exonic De Novo Mutation Rate in Individuals with Schizophrenia. *Nature Genetics* 2011;43:860-3.
99. Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y, Levy S, Gogos JA, Karayiorgou M. De Novo Gene Mutations Highlight Patterns of Genetic and Neural Complexity in Schizophrenia. *Nature Genetics* 2012;44:1365-9.
100. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M. Exome Sequencing Supports a De Novo Mutational Paradigm for Schizophrenia. *Nature Genetics* 2011;43:864-8.
101. Haldane JBS. The Part Played by Recurrent Mutation in Evolution. *The American Naturalist* 1933;67:5.
102. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J. Exome Sequencing Identifies Mll2 Mutations as a Cause of Kabuki Syndrome. *Nature Genetics* 2010;42:790-3.
103. Trappe R, Laccone F, Cobilanschi J, Meins M, Huppke P, Hanefeld F, Engel W. Mecp2 Mutations in Sporadic Cases of Rett Syndrome Are Almost Exclusively of Paternal Origin. *Am J Hum Genet* 2001;68:1093-101.
104. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A Map of Human Genome Variation from Population-Scale Sequencing. *Nature* 2010;467:1061-73.

105. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An Integrated Map of Genetic Variation from 1,092 Human Genomes. *Nature* 2012;491:56-65.
106. Messer PW, Arndt PF. The Majority of Recent Short DNA Insertions in the Human Genome Are Tandem Duplications. *Molecular biology and evolution* 2007;24:1190-7.
107. Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. Meta-Analysis of Gross Insertions Causing Human Genetic Disease: Novel Mutational Mechanisms and the Role of Replication Slippage. *Human mutation* 2005;25:207-21.
108. Lee JA, Carvalho CM, Lupski JR. A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell* 2007;131:1235-47.
109. Kim PM, Lam HY, Urban AE, Korbel JO, Affourtit J, Grubert F, Chen X, Weissman S, Snyder M, Gerstein MB. Analysis of Copy Number Variants and Segmental Duplications in the Human Genome: Evidence for a Change in the Process of Formation in Recent Evolutionary History. *Genome research* 2008;18:1865-74.
110. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of Change in Gene Copy Number. *Nature reviews Genetics* 2009;10:551-64.
111. Hassold T, Hunt P. Maternal Age and Chromosomally Abnormal Pregnancies: What We Know and What We Wish We Knew. *Current opinion in pediatrics* 2009;21:703-8.
112. Choi JY, Lee KM, Park SK, Noh DY, Ahn SH, Yoo KY, Kang D. Association of Paternal Age at Birth and the Risk of Breast Cancer in Offspring: A Case Control Study. *BMC cancer* 2005;5:143.
113. Murray L, McCarron P, Bailie K, Middleton R, Davey Smith G, Dempsey S, McCarthy A, Gavin A. Association of Early Life Factors and Acute Lymphoblastic Leukaemia in Childhood: Historical Cohort Study. *Br J Cancer* 2002;86:356-61.

114. Yip BH, Pawitan Y, Czene K. Parental Age and Risk of Childhood Cancers: A Population-Based Cohort Study from Sweden. *International journal of epidemiology* 2006;35:1495-503.
115. Green RF, Devine O, Crider KS, Olney RS, Archer N, Olshan AF, Shapira SK, National Birth Defects Prevention S. Association of Paternal Age and Risk for Major Congenital Anomalies from the National Birth Defects Prevention Study, 1997 to 2004. *Annals Epidemiology* 2010;20:241-9.
116. Wilkie AO, Slaney SF, Oldridge M, Poole MD, Ashworth GJ, Hockley AD, Hayward RD, David DJ, Pulleyn LJ, Rutland P, et al. Apert Syndrome Results from Localized Mutations of Fgfr2 and Is Allelic with Crouzon Syndrome. *Nature Genetics* 1995;9:165-72.
117. Dakouane Giudicelli M, Serazin V, Le Sciellour CR, Albert M, Selva J, Giudicelli Y. Increased Achondroplasia Mutation Frequency with Advanced Age and Evidence for G1138a Mosaicism in Human Testis Biopsies. *Fertility and sterility* 2008;89:1651-6.
118. Goriely A, McVean GA, Rojmyr M, Ingemarsson B, Wilkie AO. Evidence for Selective Advantage of Pathogenic Fgfr2 Mutations in the Male Germ Line. *Science* 2003;301:643-6.
119. Yoon SR, Qin J, Glaser RL, Jabs EW, Wexler NS, Sokol R, Arnheim N, Calabrese P. The Ups and Downs of Mutation Frequencies During Aging Can Account for the Apert Syndrome Paternal Age Effect. *PLoS Genetics* 2009;5:e1000558.
120. Goriely A, Wilkie AO. Paternal Age Effect Mutations and Selfish Spermatogonial Selection: Causes and Consequences for Human Disease. *American Journal Human Genetics* 2012;90:175-200.
121. Gauthier J, Spiegelman D, Piton A, Lafreniere RG, Laurent S, St-Onge J, Lapointe L, Hamdan FF, Cossette P, Mottron L, Fombonne E, Joober R, Marineau C, Drapeau P, Rouleau GA. Novel De Novo Shank3 Mutation in Autistic Patients. *Am J Med Genet B Neuropsychiatr Genet* 2009;150B:421-4.
122. Tarabeux J, Champagne N, Brustein E, Hamdan FF, Gauthier J, Lapointe M, Maios C, Piton A, Spiegelman D, Henrion E, Synapse to Disease T, Millet B, Rapoport JL, Delisi LE, Joober R, Fathalli F, Fombonne E, Mottron L, Forget-

- Dubois N, Boivin M, Michaud JL, Lafreniere RG, Drapeau P, Krebs MO, Rouleau GA. De Novo Truncating Mutation in Kinesin 17 Associated with Schizophrenia. *Biol Psychiatry* 2010;68:649-56.
123. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*;328:636-9.
124. Kryukov GV, Pennacchio LA, Sunyaev SR. Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *American Journal Human Genetics* 2007;80:727-39.
125. Li H, Durbin R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 2009;25:1754-60.
126. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map Format and Samtools. *Bioinformatics* 2009;25:2078-9.
127. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples. *Bioinformatics* 2009;25:2283-5.
128. Wang K, Li M, Hakonarson H. Annovar: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic acids research* 2010;38:e164.
129. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L. Systematic Meta-Analyses and Field Synopsis of Genetic Association Studies in Schizophrenia: The Szgene Database. *Nature Genetics* 2008;40:827-34.
130. Jones JM, Simkus C. The Roles of the Rag1 and Rag2 "Non-Core" Regions in V(D)J Recombination and Lymphocyte Development. *Archivum Immunologiae et Therapia Experimentalis* 2009;57:105-16.

131. Willerford DM, Swat W, Alt FW. Developmental Regulation of V(D)J Recombination and Lymphocyte Differentiation. *Current opinion in genetics & development* 1996;6:603-9.
132. Liu Q, Zerbinatti CV, Zhang J, Hoe HS, Wang B, Cole SL, Herz J, Muglia L, Bu G. Amyloid Precursor Protein Regulates Brain Apolipoprotein E and Cholesterol Metabolism through Lipoprotein Receptor Lrp1. *Neuron* 2007;56:66-78.
133. Hollenbach E, Ackermann S, Hyman BT, Rebeck GW. Confirmation of an Association between a Polymorphism in Exon 3 of the Low-Density Lipoprotein Receptor-Related Protein Gene and Alzheimer's Disease. *Neurology* 1998;50:1905-7.
134. Hadano S, Hand CK, Osuga H, Yanagisawa Y, Otomo A, Devon RS, Miyamoto N, Showguchi-Miyata J, Okada Y, Singaraja R, Figlewicz DA, Kwiatkowski T, Hosler BA, Sagie T, Skaug J, Nasir J, Brown Jr. RH, Scherer SW, Rouleau GA, Hayden MR, Ikeda JE. A Gene Encoding a Putative Gtpase Regulator Is Mutated in Familial Amyotrophic Lateral Sclerosis 2. *Nature Genetics* 2001;29:166-73.
135. Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. *Genome research* 2001;11:863-74.
136. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A Method and Server for Predicting Damaging Missense Mutations. *Nature Methods* 2010;7:248-9.
137. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of Nonneutral Substitution Rates on Mammalian Phylogenies. *Genome research* 2010;20:110-21.
138. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* 1974;185:862-4.
139. Li WH, Wu CI, Luo CC. Nonrandomness of Point Mutation as Reflected in Nucleotide Substitutions in Pseudogenes and Its Evolutionary Implications. *Journal of Molecular Evolution* 1984;21:58-71.
140. Lynch M. Rate, Molecular Spectrum, and Consequences of Human Mutation. *PNAS* 2010;107:961-8.

141. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, Del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA. A De Novo Paradigm for Mental Retardation. *Nature Genetics* 2010;42:1109-12.
142. Hamdan FF, Daoud H, Rochefort D, Piton A, Gauthier J, Langlois M, Foomani G, Dobrzeniecka S, Krebs MO, Joobor R, Lafreniere RG, Lacaille JC, Mottron L, Drapeau P, Beauchamp MH, Phillips MS, Fombonne E, Rouleau GA, Michaud JL. De Novo Mutations in Foxp1 in Cases with Intellectual Disability, Autism, and Language Impairment. *American Journal Human Genetics* 2010;87:671-8.
143. Jouan L, Girard SL, Dobrzeniecka S, Ambalavanan A, Krebs MO, Joobor R, Gauthier J, Dion PA, Rouleau GA. Investigation of Rare Variants in Lrp1, Kpna1, Als2cl and Znf480 Genes in Schizophrenia Patients Reflects Genetic Heterogeneity of the Disease. *Behavioral and brain functions : BBF* 2013;9:9.
144. Messias EL, Chen CY, Eaton WW. Epidemiology of Schizophrenia: Review of Findings and Myths. *The Psychiatric clinics of North America* 2007;30:323-38.
145. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfels R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH, Jr., Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ. Patterns and Rates of Exonic De Novo Mutations in Autism Spectrum Disorders. *Nature* 2012;485:242-5.
146. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE. Sporadic Autism Exomes Reveal a Highly Interconnected Protein Network of De Novo Mutations. *Nature* 2012;485:246-50.

147. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Gunel M, Roeder K, Geschwind DH, Devlin B, State MW. De Novo Mutations Revealed by Whole-Exome Sequencing Are Strongly Associated with Autism. *Nature* 2012;485:237-41.
148. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA. A De Novo Paradigm for Mental Retardation. *Nat Genet* 2010;42:1109-12.
149. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human Protein Reference Database--2009 Update. *Nucleic acids research* 2009;37:D767-72.
150. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A Mapreduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome research* 2010;20:1297-303.
151. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet* 2011;89:82-93.
152. Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the High Bonferroni Penalty in Genome-Wide Association Studies. *Genetic epidemiology* 2010;34:100-5.
153. Gao X, Starmer J, Martin ER. A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms. *Genetic epidemiology* 2008;32:361-9.

154. Eckner R, Ewen ME, Newsome D, Gerdes M, DeCaprio JA, Lawrence JB, Livingston DM. Molecular Cloning and Functional Analysis of the Adenovirus E1a-Associated 300-Kd Protein (P300) Reveals a Protein with Properties of a Transcriptional Adaptor. *Genes & development* 1994;8:869-84.
155. Roelfsema JH, White SJ, Ariyurek Y, Bartholdi D, Niedrist D, Papadia F, Bacino CA, den Dunnen JT, van Ommen GJ, Breuning MH, Hennekam RC, Peters DJ. Genetic Heterogeneity in Rubinstein-Taybi Syndrome: Mutations in Both the Cbp and Ep300 Genes Cause Disease. *Am J Hum Genet* 2005;76:572-80.
156. Hoth CF, Milunsky A, Lipsky N, Sheffer R, Clarren SK, Baldwin CT. Mutations in the Paired Domain of the Human Pax3 Gene Cause Klein-Waardenburg Syndrome (Ws-iii) as Well as Waardenburg Syndrome Type I (Ws-I). *Am J Hum Genet* 1993;52:455-62.
157. Kobayashi A, Miyake T, Kawaichi M, Kokubo T. Mutations in the Histone Fold Domain of the Taf12 Gene Show Synthetic Lethality with the Taf1 Gene Lacking the Taf N-Terminal Domain (Tand) by Different Mechanisms from Those in the Spt15 Gene Encoding the Tata Box-Binding Protein (Tbp). *Nucleic acids research* 2003;31:1261-74.
158. Kwiatkowski TJ, Jr., Bosco DA, Leclerc AL, Tamrazian E, Vanderburg CR, Russ C, Davis A, Gilchrist J, Kasarskis EJ, Munsat T, Valdmanis P, Rouleau GA, Hosler BA, Cortelli P, de Jong PJ, Yoshinaga Y, Haines JL, Pericak-Vance MA, Yan J, Ticozzi N, Siddique T, McKenna-Yasek D, Sapp PC, Horvitz HR, Landers JE, Brown RH, Jr. Mutations in the Fus/Tls Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. *Science* 2009;323:1205-8.
159. Croen LA, Najjar DV, Fireman B, Grether JK. Maternal and Paternal Age and Risk of Autism Spectrum Disorders. *Archives Pediatric and Adolescent Medicine* 2007;161:334-40.
160. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K.

Rate of De Novo Mutations and the Importance of Father's Age to Disease Risk. *Nature* 2012;488:471-5.

161. Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J. Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* 2012;151:1431-42.
162. Antipenko Ye N, Kogut NN. The Experience of Mutation Rate Quantitative Evaluation in Connection with Environmental Pollution (Based on Studies of Congenital Anomalies in Human Populations). *Mutation research* 1993;289:145-55.
163. Ikehata M, Koana T, Suzuki Y, Shimizu H, Nakagawa M. Mutagenicity and Co-Mutagenicity of Static Magnetic Fields Detected by Bacterial Mutation Assay. *Mutation research* 1999;427:147-56.
164. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordonez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, Birney E, Rhodes MD, McKernan KJ, Stratton MR, Futreal PA, Campbell PJ. A Small-Cell Lung Cancer Genome with Complex Signatures of Tobacco Exposure. *Nature* 2010;463:184-90.
165. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. Quantisnp: An Objective Bayes Hidden-Markov Model to Detect and Accurately Map Copy Number Variation Using Snp Genotyping Data. *Nucleic acids research* 2007;35:2013-25.
166. Medvedev P, Fiume M, Dzamba M, Smith T, Brudno M. Detecting Copy Number Variation with Mated Short Reads. *Genome research* 2010;20:1613-22.
167. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome biology* 2009;10:R25.

168. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. Breakdancer: An Algorithm for High-Resolution Mapping of Genomic Structural Variation. *Nature Methods* 2009;6:677-81.
169. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, Scheffer H, de Vries BB, Brunner HG, Veltman JA, Vissers LE. Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *The New England journal of medicine* 2012;367:1921-9.
170. Boivin M, Brendgen M, Dionne G, Dubois L, Perusse D, Robaey P, Tremblay RE, Vitaro F. The Quebec Newborn Twin Study into Adolescence: 15 Years Later. *Twin research and human genetics : the official journal of the International Society for Twin Studies* 2012;1-6.
171. Need AC, McEvoy JP, Gennarelli M, Heinzen EL, Ge D, Maia JM, Shianna KV, He M, Cirulli ET, Gumbs CE, Zhao Q, Campbell CR, Hong L, Rosenquist P, Putkonen A, Hallikainen T, Repo-Tiihonen E, Tiihonen J, Levy DL, Meltzer HY, Goldstein DB. Exome Sequencing Followed by Large-Scale Genotyping Suggests a Limited Role for Moderately Rare Risk Factors of Strong Effect in Schizophrenia. *American Journal Human Genetics* 2012;91:303-12.
172. Takano A, Zochi R, Hibi M, Terashima T, Katsuyama Y. Expression of Strawberry Notch Family Genes During Zebrafish Embryogenesis. *Developmental dynamics : an official publication of the American Association of Anatomists* 2010;239:1789-96.
173. Takano A, Zochi R, Hibi M, Terashima T, Katsuyama Y. Function of Strawberry Notch Family Genes in the Zebrafish Brain Development. *The Kobe journal of medical sciences* 2011;56:E220-30.
174. Malaspina D, Reichenberg A, Weiser M, Fennig S, Davidson M, Harlap S, Wolitzky R, Rabinowitz J, Susser E, Knobler HY. Paternal Age and Intelligence: Implications for Age-Related Genomic Changes in Male Germ Cells. *Psychiatric genetics* 2005;15:117-25.

175. Reichenberg A, Gross R, Weiser M, Bresnahan M, Silverman J, Harlap S, Rabinowitz J, Shulman C, Malaspina D, Lubin G, Knobler HY, Davidson M, Susser E. Advancing Paternal Age and Autism. *Arch Gen Psychiatry* 2006;63:1026-32.
176. Saha S, Barnett AG, Foldi C, Burne TH, Eyles DW, Buka SL, McGrath JJ. Advanced Paternal Age Is Associated with Impaired Neurocognitive Outcomes During Infancy and Childhood. *PLoS medicine* 2009;6:e40.
177. Frans EM, Sandin S, Reichenberg A, Langstrom N, Lichtenstein P, McGrath JJ, Hultman CM. Autism Risk across Generations: A Population-Based Study of Advancing Grandpaternal and Paternal Age. *JAMA psychiatry* 2013;70:516-21.
178. Gratten J, Visscher PM, Mowry BJ, Wray NR. Interpreting the Role of De Novo Protein-Coding Mutations in Neuropsychiatric Disease. *Nature Genetics* 2013;45:234-8.
179. Maher B. Personal Genomes: The Case of the Missing Heritability. *Nature* 2008;456:18-21.
180. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA. Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. *Science* 1993;261:921-3.
181. Manandhar G, Schatten H, Sutovsky P. Centrosome Reduction During Gametogenesis and Its Significance. *Biology of reproduction* 2005;72:2-13.
182. Crow JF. The Origins, Patterns and Implications of Human Spontaneous Mutation. *Nature reviews Genetics* 2000;1:40-7.
183. Goriely A, McGrath JJ, Hultman CM, Wilkie AO, Malaspina D. "Selfish Spermatogonial Selection": A Novel Mechanism for the Association between Advanced Paternal Age and Neurodevelopmental Disorders. *American Journal Psychiatry* 2013;170:599-608.

Annexes

Annexe I : CV de Simon Girard

Annexe I : Curriculum Vitae

Simon Girard

Naissance: 01 Janvier 1986

Curriculum Académique

2010 - 2013	Université de Montréal Doctorat, Biologie Moléculaire (en cours)
2008 - 2010	Université de Montréal Maîtrise, Sciences Neurologiques
2005 - 2008	Université Laval Baccalauréat en bio-informatique

Intérêts de recherche

- Les nouvelles mutations ainsi que leurs transmissions dans la population
- L'implication des mutations de novo dans les traits complexes
- Approche mécanistique pour améliorer la compréhension des traits complexes
- La génétique des maladies psychiatriques
- Caractérisation des génomes en utilisant le séquençage à haut-débit
- Séquençage d'exome dans un cadre clinique

Bourses et Prix

- Bourse d'excellence, CRCHUM, Décembre 2009
- Bourse d'échange international, Ministère des affaires étrangères, May 2009
- Bourse de recrutement, Faculté de Médecine de l'UdM
- Meilleure présentation orale - Rassemblement Médecine Génétique Associé, 2010
- Meilleure présentation orale - Compute Canada symposium, 2011
- ESHG Poster Award Candidates - European Society of Human Genetics, 2011
- Brain star award (Top 3), Canadian Institute of Health Research, 2012
- Bourse de Doctorat (1 an) - Rassemblement Médecine Génétique Associé, 2012
- Bourse de Doctorat (3 ans) - Fonds de Recherche - Santé - Québec - 2013

Présentations

Orale

- American Society of Human Genetics meeting (Invited speaker), San Francisco, 2012
- Annual meeting on RBD disorder, Paris, 2012
- European Society of Human Genetics (Invited speaker), Nuremberg, 2012
- Canadian Association for neuroscience meeting (Invited speaker), Vancouver, 2012
- Human Genome (HUGO) meeting (Invited speaker), Sydney, 2012
- Agilent NGS roadshow (Invited speaker), Saint-Louis, Raleigh and Boston, 2011
- Compute Canada HPC symposium, Montreal, 2011
- Annual Canadian Genetic Epidemiology and Statistics meeting, Toronto, 2011
- RMGA workshop (organizer and main speaker), Montréal, 2011
- Genome Informatics, Hinxton, 2010
- RMGA annual meeting, Montreal, 2010

Affiche

- European Society of Human Genetics, Nuremberg, 2012
- International congress of Human Genetics, Montreal, 2011
- Genome Informatics, Cold Spring Harbor, 2011
- European Society of Human Genetics, Amsterdam, 2011
- European Society of Human Genetics, Vienna, 2009

Articles (sélectionnés)

- **Girard SL**, Gauthier J, Noreau A, Xiong L, [...] Krebs MO, Rouleau GA. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet.* 2011 Jul 10;43(9):860-3, **Facteur d'impact: 35.532**
- Merner ND, **Girard SL**, Catoire H, [...], Rouleau GA. Exome Sequencing identifies FUS Mutations as a Cause of Essential Tremor, *Am J Hum Genet.* 2012 Aug 10;91(2):313-9. **Facteur d'impact 2012 : 10.603**
- **Girard SL**, Dion PA, Rouleau GA. Schizophrenia genetics: putting all the pieces together. *Curr Neurol Neurosci Rep.* 2012 Jun;12(3):261-6 **Facteur d'impact 2012: 2.25**
- **Girard SL**, Rouleau GA. How de novo mutation studies will change our view of the genetics of neurological and psychiatric disorders. *Future Neurology*,
- **Girard SL**, Xiong L, Dion PA, Rouleau GA. Where are the missing pieces of the schizophrenia genetics puzzle? *Current opinion in genetics & development* 2011;21:310-6. **Facteur d'impact 2012 : 8.209**
- **Girard SL**, Xiong L, Dion PA, Rouleau GA. Where are the missing pieces of the schizophrenia genetics puzzle? *Curr Opin*

Genet Dev. 2011 Jun;21(3):310-6. **Facteur d'impact 2012: 8.086**

- Srour M, Rivière JB, Pham JM, Dubé MP, **Girard S**, [...], Rouleau GA. Mutations in DCC cause congenital mirror movements. *Science*. 2010 Apr 30;328(5978):592. **Facteur d'impact 2012 : 31.201**