

Université de Montréal

**Genetic contribution to the aggregation of schizophrenia
and bipolar disorder in multiplex consanguineous
Pakistani pedigrees**

par Qin He

Programme de Sciences Biomédicales
Faculté de Médecine

Thèse présentée
en vue de l'obtention du grade de Ph. D.
en Sciences Biomédicales

Mars 2019

© Qin HE, 2019

Résumé

La schizophrénie (SCZ) et le trouble bipolaire (TB) sont des troubles mentaux graves qui présentent tous deux des symptômes affectifs et psychotiques. La SCZ est un trouble psychotique primaire caractérisé par des symptômes d'idées délirantes et d'hallucinations. Le TB est principalement un trouble de l'humeur primaire défini des périodes de manie et de dépression. En 2010, ces troubles contribuaient respectivement à 7,4% et 7,0% de la charge mondiale de morbidité. La prévalence élevée (~ 0,4% pour la SCZ et ~ 2,4% pour le TB) et la forte héritabilité estimée (~ 80%) suggèrent toutes deux une forte influence génétique. Les données disponibles démontrent qu'il existe des chevauchements génétiques entre les deux conditions, mais également des composantes génétiques spécifiques à chaque maladie.

Au cours de la dernière décennie, des études d'association pan-génomiques ont identifié des centaines de loci génétiques associés à ces maladies. De plus, d'autres méthodes ont permis de mettre en relief la contribution d'autres types de variations génétiques comme les rares variations du nombre de copies (CNV), de rares polymorphismes de nucléotide simple (SNV) et des mutations de novo (MDN). Bien que notre connaissance de l'architecture génétique de ces conditions est en progression, une grande partie de l'héritabilité demeure toujours non résolue et inexplicée.

Une longue histoire de faible mélange génétique combiné à la pratique répandue de mariages consanguins (50% des unions sont consanguines) rend les familles pakistanaises prometteuses pour des études génétiques médicales basées sur la population. Des études épidémiologiques ont démontré que la consanguinité est associée à un risque accru de nombreux traits. L'étude de familles a largement été appliquée dans la cartographie génétique des caractères mendéliens et complexes. Cependant, peu d'études ont eu recours à de grandes familles consanguines multiplexes pour étudier en profondeur le rôle de la consanguinité dans les troubles neuropsychiatriques tels que la SCZ et le TB.

Les CNVs ont été impliquées dans la SCZ et le TB depuis la découverte des délétions 22q11.2. Malgré que ces derniers soient rares dans la population, ils contribuent de manière significative au risque. Des études d'association de CNV ont révélé un enrichissement de

délétions et de duplications rares et un taux plus élevé de CNV de novo dans les cas relatifs aux témoins. De plus, le séquençage du génome de familles SCZ a révélé une charge accrue de rares CNVs exoniques chez les sujets SCZ ainsi que de l'hétérogénéité génétique. L'utilisation de grandes familles de multiplexes pourrait être statistiquement puissante pour étudier le rôle des CNVs co-ségrégant avec la maladie et éventuellement pathogènes.

Afin de mieux comprendre l'hétérogénéité génétique et résoudre l'héritabilité manquante de ces deux troubles mentaux, nous avons utilisé du génotypage et du séquençage de l'exome afin d'examiner le profil génétique de grandes généalogies consanguines multiplexes d'origine Pakistanaise. Chacune de ces familles comportait plus de dix membres affectés par la SCZ ou le TB. Dans cette thèse, nous caractérisons la population d'origine, ce qui comprend le mélange génétique et la consanguinité récente de notre cohorte. Nous avons testé si le niveau de consanguinité était associé au phénotype binaire et à ses dimensions sous-phénotypiques. Nous avons également inclus un grand ensemble de données de populations contrôles externes et appariées afin de calculer et comparer le coefficient de consanguinité. Notre approche, qui comprenait une analyse de liaison, une cartographie de l'auto-zygosité, la détection de cycles homozygotie et une analyse de ségrégation de variantes homozygotes délétères rares, nous a conduit à rejeter l'hypothèse d'un modèle de transmission récessif sur ces familles (malgré leur forte consanguinité).

Par la suite, nous avons examiné si des CNVs co-ségréguaient avec le phénotype dans certaines familles. Cette étude comportait plusieurs étapes: 1 - une comparaison systématique entre différents algorithmes de détection de CNVs. 2 - une validation croisée de vrais CNVs ou de faux positifs par des approches *in silico* ou expérimentales, 3 - le développement d'un logiciel de ségrégation et d'annotation. Cette étude met de l'avant à la fois les avancées méthodologiques et les limites de l'exploration des CNVs. Au final, aucun des CNVs identifiés ne semblent contribuer à la variance génétique de la SCZ et du TB des familles examinées dans cette étude. Les résultats présentés dans cette thèse étayaient une hypothèse alternative qui impliquerait des interactions polygéniques entre à la fois des variants rares et des variants communs.

Mots-clés: Pakistanais, familles multiplexes, consanguinité, génotypage, séquençage de l'exome, schizophrénie, trouble bipolaire, variation du nombre de copies

Abstract

Schizophrenia (SCZ) and bipolar disorder (BP) are two major psychiatric disorders. SCZ is a primary psychotic disorder that typically involves symptoms of delusions and hallucinations, by comparison BP is a mood disorder engaging mania and depression but it can also involve psychosis. A 2010 estimation of these disorders highlighted that they respectively contributed to ~7.4% and ~7.0% of the global burden of disease. The high prevalence (~0.4% for SCZ and ~2.4% for BP) and estimated heritability (~80%) suggest a strong genetic influence. Evidence shows that there are some genetic overlaps between the two conditions but also disorder-independent genetic components. Over the past decade, genome-wide association studies (GWAS) identified hundreds of SCZ and BP loci, and other approaches identified various forms of potential genetic risk factors, for instance rare copy number variants (CNVs), rare single nucleotide variants (SNVs) and *de novo* mutations (DNMs). While our knowledge of the genetic architecture of these conditions grows, a large portion of the genetic heritability of each disorder still remains unexplained.

The combination of a long history of genetic admixture, and the tradition of consanguineous marriages (50% of unions are consanguineous), makes Pakistani families promising for population based medical genetics studies. Consanguinity has previously been associated with an increased risk of numerous traits in epidemiological studies. Family-based designs have been widely applied in the genetic mapping of Mendelian and complex traits. However, few studies have used large multiplex consanguineous families to thoroughly investigate the role of consanguinity in neuropsychiatric disorders such as SCZ and BP. CNVs have been implicated in SCZ and BP since the discovery of 22q11.2 deletions, however, most of them are rare in the population but contribute significantly to the risk. Association studies of CNVs found enrichment of rare deletions and duplications, and a higher rate of *de novo* CNVs in cases relative to controls. Whole-genome sequencing of multiplex SCZ families reported increased burden of rare, exonic CNV in SCZ probands and genetic heterogeneity. Using large multiplex families could be statistically powerful to investigate the role of segregating, and possibly pathogenic, CNVs.

In order to better understand the genetic heterogeneity and look for missing heritability of these two common disorders in Pakistani families, we used SNP genotyping and whole-exome sequencing to examine the genetic profile of ten large multiplex consanguineous pedigrees; each of these families involved more than ten members affected by SCZ or BP. In this thesis, we characterized the population background which includes admixture and recent inbreeding of our cohort. We tested if the inbreeding level was associated with the binary phenotype and its subphenotype dimensions. We also included large external dataset of matched population control individuals to compute and compare the inbreeding coefficient. Our approach, which included linkage analysis, autozygosity mapping, runs of homozygosity (ROH) and rare deleterious homozygous variants segregation analysis, led us to reject the hypothesis of a recessive inheritance model across these families (despite of their high inbreeding). We subsequently looked if any CNV segregated across some of the families. This examination involved multiple steps: 1 - a systematic comparison of a range of CNV detection algorithms currently available through different platforms, 2 - a cross validation of true and false positive CNV calls through the use of in silico or experimental approaches, 3 - the development of our own segregation and annotation software. This effort both emphasized the methodological advances and limitations of CNV studies. In the end, none of the potentially pathogenic CNV identified appeared to account for the genetic variance of SCZ and BP observed in the families examined here. The results presented in this thesis provide support for an alternate hypothesis that would involve a polygenic pattern where both rare variants and common variants would be at play.

Keywords: Pakistani, multiplex families, consanguinity, SNP chip genotyping, whole-exome sequencing, schizophrenia, bipolar disorder, copy number variants

Table of contents

Résumé.....	i
Abstract.....	iii
Table of contents.....	v
List of tables.....	vii
List of figures.....	viii
List of acronyms.....	ix
Acknowledgements.....	xii
Chapter 1: Introduction.....	14
1.1 The introduction to schizophrenia and bipolar disorder.....	14
1.1.1 The definition and clinical symptoms of schizophrenia and bipolar disorder.....	14
1.1.2 The neurobiological basis of schizophrenia and bipolar disorder.....	15
1.2 The interest of consanguineous families and populations in neuropsychiatric disorders	17
1.2.1 The historical research interest on consanguineous populations and pedigrees...	17
1.2.2 The characteristics of Pakistani populations and their application in genetics	research.....
1.2.3 The genetics of consanguinity and inbreeding in neuropsychiatry.....	24
1.3 The overview of the genetics of schizophrenia and bipolar disorder.....	26
1.3.1 The prevalence and heritability of schizophrenia and bipolar disorder.....	26
1.3.2 The overlap of genetic components of schizophrenia and bipolar disorder.....	27
1.3.3 The history of genetic studies on schizophrenia and bipolar disorder.....	29
1.3.4 The family-based study designs in neuropsychiatric genetics research.....	30
1.3.5 Common disease – common variants (CD/CV) hypothesis on the genetics of	schizophrenia and bipolar disorder.....
1.3.6 Common disease – rare variants (CD/RV) hypothesis on the genetics of	schizophrenia and bipolar disorder.....
Chapter 2: the role of consanguinity in psychotic disorders.....	58
2.1 Preface.....	58

2.2	SNP microarray and whole-exome sequencing of large consanguineous Pakistani families does not support high-penetrance deleterious homozygous variants as a direct cause for the psychiatric phenotypes	60
Chapter 3: The contribution of copy number variants in multiplex Pakistani families		81
3.1	Preface.....	81
3.2	Familial segregation analysis for copy number variations: new software and methodological recommendations	83
Chapter 4: Discussion		105
4.1	The recessive hypothesis and copy number variation calling.....	105
4.2	Alternative hypotheses to explore.....	110
Conclusion		113
Appendix 1: supplementary material for Chapter 2.2.....		115
Appendix 2: Supplementary material for Chapter 3.2		141
BIBLIOGRAPHY.....		i

List of tables

Table I.	Major GWAS studies of schizophrenia and bipolar disorder from 2007 to 2017	51
Table I.	Major GWAS studies of schizophrenia and bipolar disorder from 2007 to 2017, continued.....	52
Table II.	Reported CNV association to schizophrenia and bipolar disorder	55
Table II.	Reported CNV association to schizophrenia and bipolar disorder, continued.....	56
Table III.	Summary of the ten Pakistani families	77
Table IV.	Estimation of the mean inbreeding coefficient and mating types of individuals' parents by pedigree	77
Table V.	Comparison of ROHs in affected and unaffected family members	78
Table VI.	Summary statistics of homozygous variants shared by all the affected individuals in each family.....	78
Table VII.	Number and percentage of likely false positive CNVs and likely true positive CNVs in autosomal chromosomes and estimation of the sensitivity and the specificity for each software	98
Table VIII.	Number and percentage of likely false positive CNVs and likely true positive CNVs in X chromosome and estimation of the sensitivity and the specificity for each software	99
	Likely false positive CNVs are singleton CNVs; likely true positive CNVs are defined as segregating CNVs (in ≥ 2 family members).....	99
Table IX.	Features of likely false positive and likely true positive CNVs called from the genotyping data.....	100
Table X.	Features of likely false positive and likely true positive CNVs called from the WES data.	100
Table XI.	The consequence of using different filtering parameters.....	101

List of figures

Figure 1.	Global prevalence of consanguinity.....	22
Figure 2.	Hypothesized model of the complex relationship between biological variation and some major forms of psychopathology.....	46
Figure 3.	Variance accounted for by genetic, shared environmental, and non-shared environmental effects for schizophrenia and bipolar disorder.....	48
Figure 4.	The relationship between the effective sample size and number of GWAS loci..	49
Figure 5.	F_{ROH_WES} of MNS pedigrees and population controls.....	79
Figure 6.	Example of a putative ROH segment shared by all affected members in MNS0380	
Figure 7.	A- ROC curve of features of CNVs called from the genotyping data. B- ROC curve of features of CNVs called from the WES data.....	102
Figure 8.	A-Venn diagram of true positive CNVs called from the genotyping data. B- Venn diagram of true positive CNVs called from the WES data.....	103
Figure 9.	An example of CNV demonstrated by genotyping intensity (7q31.1 deletion) .	104

List of acronyms

1KGP	1000 Genomes Project
ARC	Activity-regulated cytoskeleton-associated protein
ASD	Autism spectrum disorder
BP	Bipolar disorder
CNP	Consortium of Neuropsychiatric Phenomis
CNV	Copy number variant
D2	Dopamine receptor 2
DGV	Database of Genomic Variants
DSM-5	Diagnostic and Statistical Manual of Mental Disorders fifth edition
ExAC	Exome Aggregation Consortium
FBA	Family-based association
FMRP	Fragile-X mental retardation protein
FoSTeS	Fork stalling and template switching
GWAS	Genome-wide association study
HGDP	Human Genome Diversity Panel
IBD	Identical by decent
ICD-10	International Statistical Classification of Diseases and Related Health Problems 10th Revision
LOF	Loss-of-function
MAF	Minor allele frequency
NAHR	Non-allelic homologous recombination
NHEJ	Non-homologous end joining
NGS	Next-generation sequencing
NMDAR	N-methyl-d-aspartate receptor
PCR	Polymerase chain reaction
PSD	Postsynaptic density
SCZ	Schizophrenia
SNV	Single nucleotide variant
SNP	Single nucleotide polymorphism

TDT	Transmission Disequilibrium Test
URV	Ultra-rare variant
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

To my parents, 何明荣 and 梁素琼,

My sister 何静,

My brother 何磊,

for their unconditional love and support.

Acknowledgements

I would like to thank my supervisor, Dr. Lan Xiong for establishing this project a decade ago, for admitting me in her lab, for sending me to courses and international conferences, for providing me guidance and resources, for discussing and criticizing to move me forward, and for encouraging me to become an independent researcher.

Special thanks to Dr. Guy Rouleau for enlightening me with his wisdom and judgement, for reminding me of basic principles in genetics and in life, for being enthusiastic in my findings in the project, and for accepting me in his big family when my project wasn't going well.

I truly appreciate that Dr. Boris Chaumette lent me a hand during hard times, that you always think out of the box and is also an example of action. I have enjoyed the efficiency and the positivity collaborating with you. You are a teacher, a collaborator, and a good friend.

I want to thank my previous colleague, Amélie, who helped me prepare the samples and process the data. It was a special journey to work with excellent medical students from all over the world, especially Simon, Piotr, and Jom Palada, from whom I had help for my project and harvested enduring friendships.

I have learned a lot from the members in Dr. Rouleau's lab, especially at the last stage of my PhD studies. It was a wonderful experience to work in this balmy atmosphere: Dr. Patrick Dion who inspired me on research and on other pursuits in life such as running a marathon, and patiently helped me on my writing; Cynthia, Jay, Gabby, Fulya, Faezeh and Cal are the best peers I've ever interacted with, who offered me their precious time and constructive feedbacks on my project, my writing, and my thesis; Alex and Dan have always been there if I have any question about bioinformatics; Sandra has been an organized model for my bench training; Sirui, Armitha, and Pingxing for the discussions about PhD projects and future careers. I feel lucky to have had you along my PhD study and in the most challenging time.

I would also like to thank my committee members, Dr. Leila Ben amor, Dr. Marie-Pierre Dubé and Dr. Ridha Joobar, for taking their time to lead me further and evaluate my progress through my study.

The data presented in this thesis is funded by Canadian Institutes of Health Research (CIHR), with a combination of other public datasets, especially the one from European Genome-phenome Archive and kind personal sharing from Dr. Vagheesh Narasimhan. I'm grateful for having the joint scholarship of China Scholarship Council and Université de Montréal to support my PhD studies in Canada.

I would like to sincerely thank Dr. David Saffen for choosing me as his master student in Fudan and giving me an example of a dedicated researcher with eternal curiosity and huge enthusiasm of the unknown. I'm infinitely grateful to my previous lab members in China, especially Yu Tao, who is a reliable colleague and a forever friend.

I wouldn't have come this far without the support and understanding from my family members, especially my father who has faith in me, pushes me to achieve higher goals and is proud of me all the time. I am indebted to my close friends, who have patiently listened to my struggles and frustrations and always stayed positive for my decisions in my study and life. I want to dedicate my special gratitude to Dr. Xu Han (Hopefully we both reach the finishing line by the time my thesis is publically available), with whom I shared a long-term close companionship in graduate studies, in research projects, and in beautiful life experiences, even though there was an ocean between us for most of the time.

Finally, I would thank everyone who appeared in my life, for long or short, to inspire me, to accompany me, and to encourage me.

Chapter 1: Introduction

1.1 The introduction to schizophrenia and bipolar disorder

1.1.1 The definition and clinical symptoms of schizophrenia and bipolar disorder

Schizophrenia is a primary psychotic disorder, and bipolar disorder is a primary mood disorder, but it can also involve psychosis. Schizophrenia and bipolar disorder are characterized as mental and behavioral disorders in both Diagnostic and Statistical Manual of Mental Disorders fifth edition (DSM-5) and International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). In DSM-5, the diagnostic criteria categorized schizophrenia spectrum and other psychotic disorders as including schizophrenia, other psychotic disorders, and schizotypal (personality) disorder. Bipolar disorder and related disorders are separated from the depressive disorders in DSM-5 as a bridge between the diagnostic classes of psychotic disorders and depressive disorders in terms of symptomatology, family history, and genetics¹.

Schizophrenia, by definition, is a disturbance that must last for six months or longer, including at least one month of delusions, hallucinations, disorganized speech, grossly disorganized or catatonic behavior, or negative symptoms¹. Delusions are one type of positive symptom, and these usually involve a misinterpretation of perceptions or experiences. Hallucinations are also a type of positive symptom and may occur in any sensory modality (e.g. auditory, visual, olfactory, gustatory, and tactile). Persecutory delusions and auditory hallucinations are the most common and characteristic types in schizophrenia. Positive symptoms are well known because they are dramatic and the major target of antipsychotic drug treatments. Negative symptoms in schizophrenia, such as apathy and avolition, are commonly considered a reduction in normal functions and are associated with long periods of hospitalization and poor social functioning. Negative symptoms of schizophrenia determine whether a patient ultimately functions well or has a poor outcome. Current drug treatments are

limited in their ability to treat negative symptoms, but psychosocial interventions, along with antipsychotics, can be helpful in reducing negative symptoms.

The symptoms of schizophrenia are widely subcategorized into five dimensions: not only positive and negative symptoms, but also cognitive symptoms, aggressive symptoms and affective symptoms². These symptoms are not necessarily unique to schizophrenia. Positive symptoms can happen in other disorders, like bipolar disorder and schizoaffective disorder. Negative symptoms can occur or overlap with cognitive and affective symptoms but are moderately unique to schizophrenia. Cognitive symptoms of SCZ emphasize “executive dysfunction”, which includes problems representing and maintaining goals, allocating attentional resources, evaluating and monitoring performance, and utilizing these skills to solve problems. Other disorders, like autism, Alzheimer’s disease and other dementias can also be associated with cognitive dysfunctions similar to those seen in schizophrenia. Affective symptoms and aggressive symptoms are also prominent features of other mental disorders.

Bipolar spectrum disorders are characterized by manic-depressive disorder or affective psychosis, where depressive and manic episodes alternate, which causes unusual shifts in mood, energy, activity levels, and the ability to carry out day-to-day tasks. Patients with bipolar I disorder (BP-I) have full-blown manic episodes or mixed episodes of mania plus depression, often followed by a depressive episode. Patients with bipolar I disorder can also have rapid switches from mania to depression and back. This switch occurs at least four times a year. Bipolar II disorder (BP-II) is characterized by at least one hypomanic episode that follows a depressive episode².

1.1.2 The neurobiological basis of schizophrenia and bipolar disorder

The different symptoms of schizophrenia are hypothesized to be regulated by different brain regions². Positive symptoms are hypothetically modulated by malfunctioning mesolimbic circuits, while negative symptoms are hypothetically linked to malfunctioning mesocortical circuits and may also involve mesolimbic regions such as the nucleus accumbens, which is part of the brain’s reward system and thus play a role in motivation and may also be involved in the increased rate of substance use and abuse behavior seen in schizophrenia patients. Affective symptoms are associated with the ventromedial prefrontal cortex, while aggressive symptoms

(related to impulse control) are associated with abnormal information processing in the orbitofrontal cortex and amygdala. Cognitive symptoms are associated with problematic information processing in the dorsolateral prefrontal cortex. The hypothetical model of allocating specific symptom dimensions to brain regions may seem oversimplified, but it assists research and has clinical value.

Two neurotransmitters and their neuronal pathways in the brain – dopamine and glutamate – are the leading hypotheses for explaining the symptoms of schizophrenia, as well as the therapeutic effects and side effects of antipsychotic drugs². One of the five dopamine pathways in the brain is the mesolimbic dopamine pathway. The hyperactivity of this pathway causes the positive symptoms of psychosis, such as delusions and hallucinations. Most antipsychotics work as dopamine antagonists, to block the dopamine receptor 2 (D2), resulting in the decrease of dopamine activity, and therefore stop of positive symptoms. The cognitive, negative and affective symptoms of schizophrenia are believed to be due to a deficit of dopamine activity in mesocortical projections to ventromedial prefrontal cortex. The balance between decreasing dopamine in the mesolimbic pathway and increasing dopamine in the mesocortical pathway generates a dilemma for the therapeutic effects of antipsychotics.

The neurotransmitter glutamate has gained more attention in the pathophysiology of schizophrenia and other psychiatric disorders in recent years². Glutamate, as a ubiquitous excitatory neurotransmitter, seems to be able to excite nearly any neuron in the brain and involves several types of receptors. Molecules targeting the glutamate synapses are serving as either antagonist, to block glutamate release presynaptically, or agonist to facilitate glutamatergic neurotransmission postsynaptically. A major hypothesis for the cause of schizophrenia is that glutamate activity at NMDA (N-methyl-d-aspartate) receptors is decreased, due to abnormalities in the formation of glutamatergic NMDA synapses during the neurodevelopment. This theory is partly based on the use of the NMDA receptor antagonists PCP (phencyclidine) and ketamine in normal human, which could mimic not only positive symptoms but also the cognitive, negative and affective symptoms of schizophrenia. The theory is also partly upheld by the formation of defective synapses at certain GABA interneurons at the cerebral cortex or hippocampus, which causes dysconnectivity of glutamate circuits. This

NMDA hypofunction hypothesis can connect the interaction of glutamate pathways and dopamine pathways, since they display an upstream-downstream relationship.

Similarly, three principal neurotransmitters including norepinephrine, dopamine and serotonin have long been implicated in both the pathophysiology and treatment of mood disorders such bipolar disorder². The neurotransmitter hypothesis suggests that dysfunction, generally due to underactivity of one or more of the three monoamines, may cause depression symptoms, while boosting one or more of the three monoamines in specific brain regions may be linked to symptoms of mania.

1.2 The interest of consanguineous families and populations in neuropsychiatric disorders

1.2.1 The historical research interest on consanguineous populations and pedigrees

Consanguineous marriages (a couple related as second cousins or closer, equivalent to an inbreeding coefficient $F \geq 0.0156$ in their progeny) may have been practiced since the early existence of human society. The potential breeding populations has been estimated to be a minimum of 700 individuals to a maximum of 10,000 persons³⁻⁶ in the out-of-Africa migration of our human ancestors. 60,000-70,000 years ago, extensive inbreeding was basically inevitable, given their hunter-gatherer lifestyle, subdivision into separate small kindred groupings and the suggestion that they exited Africa in two distinct waves^{7,8}.

In Ancient Egypt, when pharaohs ruled, the political and religious leaders performed brother-sister or uncle-niece marriages in order to keep their bloodline pure. The mummy of King Tutankhamun was recently examined, with another ten royal mummies, through the DNA samples taken from their bones. The samples were subjected to microsatellite-based haplotyping and generational segregation of alleles within possible pedigrees, accounting for correlation of identified diseases with individual age, along with archeological and historical evidence. The construction of the five-generation pedigree identified an accumulation of malformations in Tutankhamun's family, and also revealed that King Tut was beset by malaria and a bone disorder – possibly due to his incestuous origins: King Tut's mother and father are siblings⁹. As for King

Tut himself, he married his half-sister and they did not successfully produce an heir (while having two stillborn daughters).

The European royal dynasties of the Early Modern Age provide an example for studying inbreeding in human populations¹⁰. For example, King Charles II of Spain, from the Spanish Habsburg royal family, was physically and mentally disabled, infertile and extremely inbred. Following 16 generations (~200 years) of inbreeding in first cousins and uncles and nieces in the Spanish Habsburg kings, the inbreeding coefficient increased strongly along generations. A statistically significant inbreeding depression for survival to 10 years was detected. Furthermore, King Charles II was believed to suffer from two different genetic, which could explain most of the complex clinical profile of this king. He passed at the age of 38 and this led to the extinction of the dynasty¹¹. Extended study by the same research group suggested the Habsburg royal family might have evolved under natural selection over three centuries to blunt the worst effects of inbreeding, based on their discovery that the childhood mortality decreased while the infant mortality increased over time. They proposed that the genetic basis of inbreeding depression was probably very different for infant and child survival in the Habsburg lineage¹². Of note, this report caused controversial views among senior geneticists in the field¹⁰.

The debate on how deleterious or harmless consanguineous unions could be started in 1858, after the first structured clinical study on the biological effects of inbreeding was published¹³. It was later criticized as having a fallacious study design and conclusion, as most of the other early studies were regarded retrospectively⁷. The debate and early studies on consanguinity in Great Britain and USA led to a radical change of opinion of major public figures such as Charles Darwin on a matter of major personal and also scientific significance⁷. Charles Darwin, who was married to his first cousin Emma Wedgwood, was one of the first experimentalists to demonstrate the adverse effects of inbreeding and to question the consequences of consanguineous mating. Darwin's opinion was somewhat changed by his son, George Darwin, who published a study on consanguinity in the late 19th century^{14,15}. A more recent study (published in 2010) on a sample of 25 Darwin/Wedgwood families of four consecutive generations showed a significant positive association between childhood mortality (including 3 of Darwin's 10 children) and inbreeding coefficient, which might be a result of

increased homozygosity of deleterious recessive alleles produced by the consanguineous marriages¹⁶.

In different geographical regions, the public attitude towards consanguinity vary widely, and is mostly driven by religious and cultural beliefs¹⁷, especially religious ordinances in more traditional rural areas. In general, consanguineous marriage is permitted within Judaism, in some branches of Christianity⁷, Islam, Dravidian Hinduism, Buddhism, the Zoroastrian/Parsi religion, and the Confucian Tradition. However, the prevalence and specific types of marriage permitted vary according to the precepts and traditions of each religion and denomination and, in some cases, these characteristics appear to have altered significantly through time^{7,18}.

A study published in 2008 on the global prevalence of consanguineous unions defined four major global areas¹⁸: 1) Regions in which fewer than 1% of marriages are consanguineous, including North America, most of Europe, and Australasia; 2) Regions in which 1-10% of all marriages are consanguineous, such as the Iberian Peninsula, Japan and South America; 3) Regions where 20% to over 50% of current marriages are consanguineous, represented by North Africa, much of West, Central and South Asia; 4) Some populous countries such as Indonesia, where the status was defined as unknown since the information on consanguinity is partial. The same author updated the distribution map, in which the data was compiled from a comprehensive collection of references, and the majority of national consanguinity levels shown in the map are either the most recent study (up to 2015) or an average of several studies¹⁹. The updated map is included below, as a reference for consanguineous marriages in the Pakistani population, shown in Figure 1.

The topic of consanguinity has its innate complexity and it cannot be easily regarded as a simplistic dichotomy of “good or bad”. Data suggest non-consanguineous progeny have a modest but statistically significant health advantage over their consanguineous counterparts, which is in alignment with the genetic concept of heterozygote advantage⁷. For instance, comparison of pre-reproductive mortality among children of first-cousin marriages, with similar mortality in the children of marriages of unrelated parents, revealed that there is a higher risk for late miscarriage, stillbirth or early death for a child of consanguineous marriages²⁰.

More observations are drawn from the associations of increased morbidity with consanguinity. However, the associated variables are not enough to infer the causality²¹, which means we are not able to suppose that consanguinity of parents would cause certain conditions. In 1902, Garrod observed that the incidence of alkaptonuria, a rare disorder in the general population but frequent in children of first-cousin marriages, conformed to the pattern of recessive inheritance described by Gregor Mendel in his experiments with peas²². Based on discussions with Mendel's advocate Bateson, who suggested that autozygosity increased the risk for the disease, Garrod deduced that alkaptonuria is a recessive disorder. Garrod was also careful to note that it was equally clear that only a minute proportion of the children of consanguineous unions are alkaptonuric.

Consanguinity principally influences the incidence of rare recessive disorders. In fact, a lot of autosomal recessive disorders are found to have an association with consanguinity. In most cases, the affected individuals have the homozygous form of the causative mutation. They inherited identical mutations from each of the biologically related parents. Sometimes, it could also be compound heterozygote mutations in affected siblings. The consanguinity could also be associated with co-expression of different recessive disorders (pleiotropy), or co-existence of multiple mutations for a single disease phenotype (polygenic inheritance)⁷.

With the increasing number of studies which employ homozygosity mapping²³ to identify recessive disease loci, there has been empirical evidence convincingly implicating consanguinity and disorders affecting infancy and childhood, such as non-syndromic hearing loss²⁴, intellectual and developmental disability, and some categories of congenital heart defects⁷. However, no single disorder, or group of disorders, affecting infancy and childhood have been consistently reported in consanguineous offspring since the reports of an association between consanguinity and a specific disorder originating from small endogamous communities (where a founder effect and a genetic drift could be predicted)⁷.

Data on consanguinity and common disease of adulthood are confusing, contradictory and inconclusive⁷. Positive, neutral, and negative association with cardiovascular diseases, diabetes, and various cancers were reported. Despite extensive genome-wide association studies, the fraction of the heritable variance in complex human disorders that can be explained by identified loci remain low. Most of the investigations lacked adequate control for the multiple

non-genetic variables. Consanguinity would be expected to exert a greater influence on the etiology of complex diseases if rare autosomal recessive alleles were causally implicated, whereas if disease alleles that are common in the gene pool are involved, then intra-familial marriage would have a proportionately lesser effect²⁵.

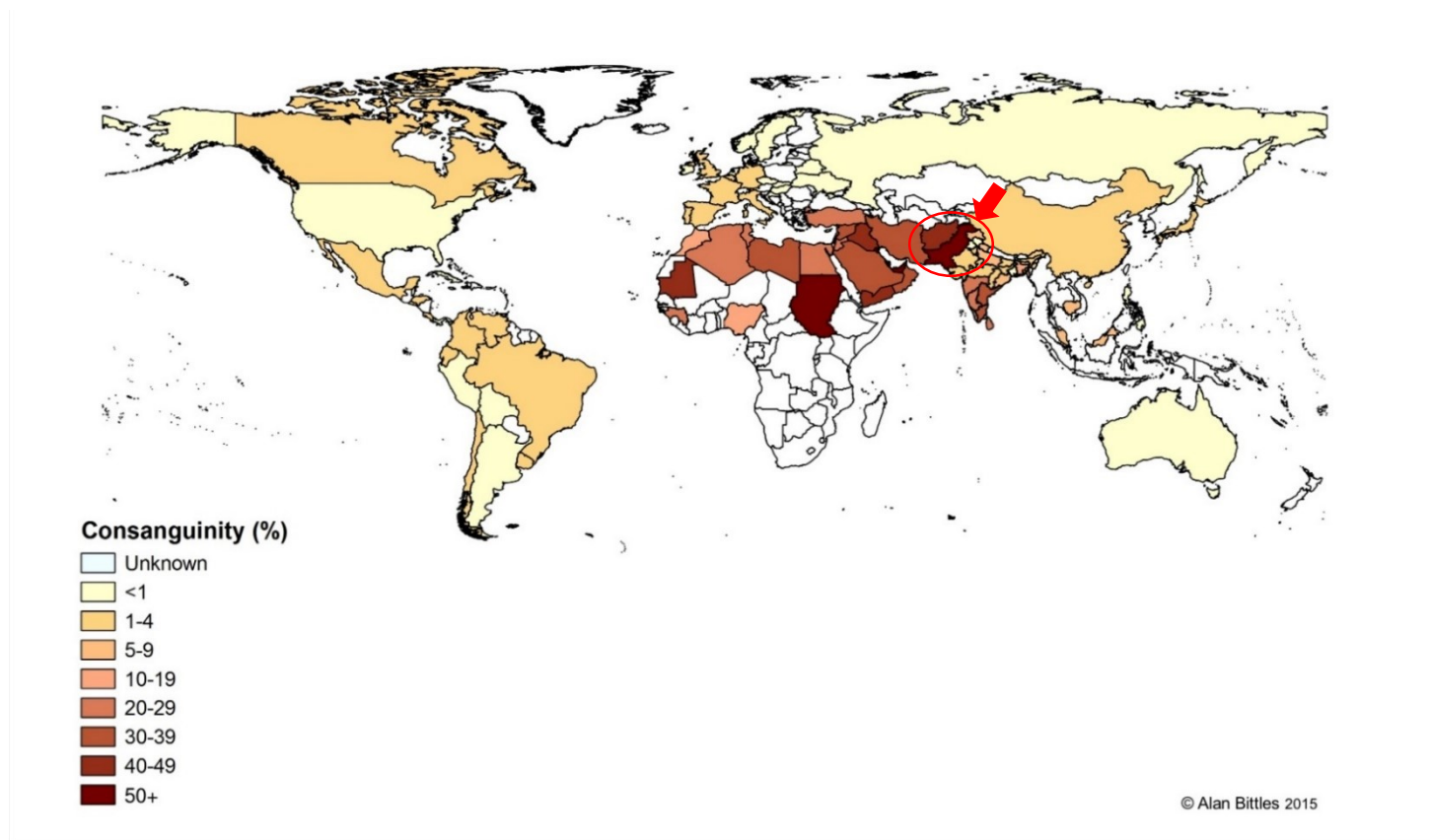


Figure 1. Global prevalence of consanguinity

Reprinted from “Global Patterns & Tables of Consanguinity” by Bittles A.H. and Black M.L., 2015, retrieved from <http://consang.net>. Copyright 2015 by Alan Bittles. The global prevalence of consanguinity map was compiled from a comprehensive collection of references, and the majority of national consanguinity levels shown in the map are either the most recent study (up to 2015) or an average of several studies. The country Pakistan is indicated in the map with red arrow and red circle.

1.2.2 The characteristics of Pakistani populations and their application in genetics research

Pakistan lies on the postulated coastal route from Africa to Australia and the earliest settlers probably came to this area some 60,000 years ago. Evidence from Paleolithic period shows a later occupation dated to around 45,000 years ago²⁶, indicating the Soanian culture of the Lower Paleolithic. Evidence of Neolithic settlements of modern humans have been found at Mehrgarh, dating back to the seventh millennium BCE, in the southern Pakistani province of Balochistan. This site predates the agrarian Harappan culture that flourished in the fertile Indus Valley from about 3300 to 1700 BCE, in what is today the Pakistani provinces of Sindh, Punjab and Balochistan. The Indus Valley Civilization was one of the three early civilizations along with Ancient Egypt and Mesopotamia, and one of the most widespread.

Invaders including the Aryans, Macedonians, Arabs, Mongols, etc. have all contributed to the ethnic variety of Pakistan's population²⁷. In present-day Pakistan, the Balochis (origin from Aleppo, Syria), Brahuīs (Turko-Iranian origin), Makranis (origin from central and southern Africa) and Sindhis (admixed) constitute the major southern populations of Pakistan. The major northern ethnic groups include the Balti (Tibetan origin), Burushos (Greek origin suspected), Hazara (Tartar origin, descents of Genghis Khan's army), Kalash kafirs (descendants of Alexander the Great's army), Kashmiris, Pathans (Greek contribution²⁸) and Punjabis (admixed). The Y-chromosome variations confirmed some of the claimed origin of the populations, but not all of them²⁹⁻³¹.

The long history of admixture and the tradition of consanguineous marriages (50% of unions are consanguineous) made Pakistani populations a good candidate for studying population genetics and medical genetics. However, most studies focused on investigating the role of consanguinity on the epidemiology of mortality/morbidity of certain conditions, especially in the Pakistani communities in European countries such as the UK. Recently, some studies started to look at the homozygous predicted loss-of-function mutations in adult Pakistani with related parents, by linking genetic data with the phenotypic data involving lifelong health records or biochemical and disease traits, to systematically understand the phenotypic consequences of complete disruption of genes^{32,33}. These studies are based on the principal that consanguineous unions are more likely to result in offspring carrying homozygous loss-of-function mutations.

1.2.3 The genetics of consanguinity and inbreeding in neuropsychiatry

Consanguinity has also been reported to be a risk factor for psychiatric conditions³⁴. Associations between consanguinity and common behavioral and psychiatric disorders have been reported in specific communities such as depression in South India³⁵ and Tourette syndrome in Iran³⁶.

Prior attempts to explore the relationship between inbreeding and schizophrenia across various population isolates, in which consanguineous marriages are frequent, have shown both positive and negative results depending on the studied population and sample size³⁷⁻³⁹, e.g. earlier studies in Sudan, Norway, and Saudi Arabia have failed to reveal elevated rates of schizophrenia in the progeny of consanguineous parents^{37,38,40}; however, schizophrenia spectrum psychosis has been associated with consanguinity in genealogy-based studies in Dagestan^{41,42}, the Dalmatian Islands, Croatia⁴³, in Israeli Bedouins⁴⁴, South Indian Tamil communities⁴⁵ and in Egypt³⁹. These studies were primarily based on epidemiological data and did not include in-depth comparisons between the genomic architecture of affected individuals and that of control individuals. Few genetic studies have investigated the molecular genetic factors that could support the link between consanguinity and psychiatric disorders. Nevertheless, some susceptibility loci have been reported previously using genome-wide linkage analysis in schizophrenia and bipolar disorder in extended pedigrees and population isolates⁴⁶⁻⁴⁸.

Genetic data can be used to estimate the degree of consanguinity in an individual (inbreeding coefficient). The inbreeding coefficient (F) measures the proportion of loci at which the offspring of a consanguineous union is expected to inherit identical gene copies from both parents. An individual for whom F is greater than or equal to 0.0156 is deemed to be consanguineous. Knight *et al* investigated five Pakistani children with schizophrenia who were descendants of a first-cousin marriage, and using homozygosity analysis and inbreeding coefficients, they reported two distinct candidate loci⁴⁹.

Genetic data can also be used to identify homozygous chromosomal regions (runs of homozygosity, ROH) resulting from consanguinity. Kurotaki *et al* recruited nine schizophrenia subjects from first-cousin marriages and 92 unaffected control individuals from the Japanese

population. When they compared the distribution of ROHs in offspring of first-cousin marriages and offspring of non-consanguineous marriages, they detected many ROHs in ≥ 3 affected individuals, including one previously reported⁵⁰. In another study using 178 schizophrenia cases and 144 unrelated Caucasian controls from outbred populations, Lencz *et al* used a whole-genome homozygosity association approach and identified nine ROHs that were significantly more common in schizophrenia cases than in controls; thus suggesting that recessive effects of relatively high penetrance might explain a significant proportion of the genetic liability for this disorder⁵¹. Following the extension of their cohort to 9,388 cases and 12,456 controls, the same group estimated that the odds of developing schizophrenia were increased by 17% for each 1% increase in genome-wide autozygosity⁵².

A large study conducted by the Psychiatric Genomics Consortium failed to replicate the significant association between ROH burden and schizophrenia after doubling the sample size⁵³. Another group studied the genome wide ROH burden in a homogeneous Irish cohort of 1,606 cases and 1,794 controls and reported no excess of ROH in schizophrenic cases by comparison to controls⁵⁴. In another study conducted using a cohort of 506 individuals with bipolar disorder and 510 unaffected individuals from the United Kingdom, no excess of ROHs was observed⁵⁵. In summary, results are mixed, and no definitive conclusion has thus far been drawn. The degree to which schizophrenia risk loci identified in genetic isolates are likely to be also found in outbred populations is questionable, but the identification of gene products that may contribute to the phenotypic expression of the disorder could provide useful clues towards successful treatment regimens⁷.

There have been few detailed studies on the possible influence of consanguinity on bipolar disorder, other than a case-control study in the Nile delta region of Egypt based on 64 DNA polymorphisms and self-reported parental relationships, with bipolar I disorder more prevalent among the progeny of consanguineous parents³⁹.

The phenotypic complexity of schizophrenia, bipolar disorder and other psychiatric disorders is a major concern. The accumulating evidence of the substantial polygenic component to the risk of schizophrenia and bipolar disorder involving thousands of common alleles of very small effect, implicates that the effect of consanguinity on these disorders is probably lesser and

an individual's risk of developing them requires a conjunction of rare high-penetrant single nucleotide variants (SNVs), rare copy number variants (CNVs), and common SNVs with epigenetic and environmental factors.

1.3 The overview of the genetics of schizophrenia and bipolar disorder

1.3.1 The prevalence and heritability of schizophrenia and bipolar disorder

Schizophrenia (SCZ) and bipolar disorder (BP) are two non-fatal mental disorders that respectively contribute to ~7.4% and ~7.0% towards the global burden of disease in 2010⁵⁶. Based on previous studies, SCZ has a lifetime prevalence of 0.4%⁵⁷. Bipolar spectrum disorder has an aggregate lifetime prevalence of about 2.4% (0.6% for BP, 0.4% for BP-II and 1.4% for subthreshold BP)⁵⁸.

A family history of SCZ and BP is a major risk factor for the development of these disorders. The genetic influences driving SCZ/BP found through familial aggregation studies were first described by clinicians and researchers before the era of molecular genetics. Familial studies were used to estimate the effects of genetics on phenotypic variance – twin studies were the most commonly used. The concordance between monozygotic twins versus dizygotic twins estimates the genetic variance that explains phenotypic variance – commonly referred to as heritability. A meta-analysis of schizophrenic twin studies estimated heritability in SCZ at 81% (95% CI, 73%-90%), while common or shared environmental influences was estimated to be 11% (95% CI, 3%-19%), suggesting a strong genetic influence⁵⁹. The most recent and largest BP twin study also estimates a strong heritability (75%)⁶⁰. A more comprehensive meta-analysis of the heritability on complex traits in humans can be visualized with the MaTCH (Meta-Analysis of Twin Correlations and Heritability) webtool, where one can view the collection of twin studies for SCZ and BP, albeit the estimates of heritability and shared environment may slightly differ⁶¹. Theoretically, the heritability of a phenotype may differ between populations due to differences in non-additive and additive genetic factors and environmental factors (i.e. differential selection pressures). However, in practice, heritability of some traits can be similar across different ethnic populations⁶².

1.3.2 The overlap of genetic components of schizophrenia and bipolar disorder

SCZ and BP likely share a genetic origin. Nonetheless, the discrete disease entities, with distinct etiology and pathogenesis, are identified by current diagnostic conventions. In the end, the diagnosis of individuals with a mixture of prominent mood and psychotic symptoms somewhat depends on the psychiatrist's subjectivity. The genetic correlation was previously calculated to be 0.68 ± 0.04 SE between SCZ and BP, demonstrating a high level of genetic overlap between the disorders⁶³. Familial coaggregation of SCZ and BP based on a meta-analysis of family studies showed that first-degree relatives of probands with SCZ had significantly increased rate of BP and first-degree relatives of probands with BP had marginally increased rates of SCZ⁶⁴. The comorbidity between the disorders was estimated to be 63% in two million Swedish nuclear families – the largest population-based study to date⁶⁵. Moreover, early linkage studies and meta-analyses have identified some chromosomal regions privy to both SCZ and BP, suggesting further evidence for comorbidity.⁶⁶

Recently, copy number variants in certain genomic loci (e.g. 13q and 22q) were found to be strongly associated with SCZ and BP⁶⁷. Candidate-gene association studies for both SCZ and BP also identified variants in the same genes, suggesting genetic overlap between the two disorders; *DISC1* (*Disrupted in Schizophrenia 1*) and *NRG1* (*Neuregulin 1*) are two examples of genes with variants driving the SCZ and/or BP phenotype(s)⁶⁸. Additionally, genome-wide association studies (GWAS) have identified significant single nucleotide polymorphisms (SNPs) in similar regions of both SCZ and BP. Meta-analyses of GWAS data have shown significant association for *ZNF804A* in both SCZ and BP⁶⁹. Furthermore, variants in *CACNA1C* were found recurrently shared between BP and other psychiatric disorders including SCZ⁷⁰. In fact, pathway analysis implicated a role for calcium channel signaling genes in major psychiatric disorders⁷¹. Besides, polygenic risk analyses have demonstrated that the burden of small-effect SNPs contribute towards the phenotypes of SCZ and BP⁷².

With the increasing evidence of shared genetic components, researchers in the field have proposed a dimensional spectrum, in which five classes of mental disorder are arranged on a single

axis and overlap due to shared risk factors: the changes of genotype influence one or more biological systems, and the relevant biological systems influence specific neural modules that comprise the key relevant functional elements of the brain. The abnormal functioning of the neural modules influences the domains of psychopathological experience and ultimately the clinical syndromes (Figure 2)^{73,74}.

The aforementioned studies suggest that SCZ and BP have an overlap in genetic risk and probably share some genetic components for pathogenesis. This idea does not mean they can fit in a single-disease category regarding clinical symptoms, genetic susceptibility and biological mechanisms. SCZ tends to have a stronger neurodevelopmental component than BP on the gradient of pathology (Figure 2), which is consistent with the evidence showing that structural genomic variations such as CNVs can contribute to neurodevelopmental pathology⁷⁵. Although CNVs do have a role in the risk of BP^{76,77}, they appear to be smaller in size or in effect, compared to the CNVs observed in autism and SCZ^{78,79}. A family study investigating the common genetic determinants of SCZ and BP also presented convincing evidence of unique genetic factors for each disorder (**Figure 3**)⁶⁵.

Despite the shared symptomology and genetics being substantial, a few studies also implicate the genetic architecture differences between these two disorders^{80,81}. As the first evidence for a genetic basis under the differences, Ruderfer *et al.* created a polygenic risk score (PRS) from a case-only SCZ versus BP diagnosis in an independent sample. They showed the PRS was significantly different between BP and SCZ and there was a significant correlation between a BP PRS and the clinical dimension of mania in SCZ patients. They further extended this rationale to a much larger sample size. They then identified genome-wide significant loci shared between disorders, and also genomic regions with disorder-independent causal variants and potassium ion response genes as contributing to differences in biology between disorders. Their PRS analysis identified several significant correlations within case-only phenotypes including SCZ PRS with psychotic features and age of onset in BP. This was the first time to discover specific loci that distinguish between BP and SCZ and identify polygenic components underlying multiple symptom dimensions⁸². Conventionally, large-scale genetic studies on SCZ and BP were carried out separately; a detailed review will be shown further in this thesis.

1.3.3 The history of genetic studies on schizophrenia and bipolar disorder

The success of mapping genes responsible for Mendelian disorders in the linkage era (1980-2005) led researchers to search for co-segregating loci for psychiatric disorders. Many linkage studies have been conducted for both SCZ and BP. However, several linked loci were likely false positives due to lack of replication across multiple independent studies⁸³. The family-based design of individual linkage studies lacked power to detect positive signals, therefore necessitating a collaborative meta-analysis. A meta-analysis of linkage analyses of SCZ suggested many nominally significant chromosomal regions containing SCZ susceptibility loci, but only one genome-wide significant peak was detected in a region never implicated in SCZ⁸⁴. The ‘aggregate’ genome-wide significant loci could not be replicated in a secondary analysis⁸⁵. These loci likely do not confer risk directly to the phenotype or may only contribute a small portion to the heritability in the general population.

The meta-analyses of linkage studies for BP detected no genome-wide significant locus with a rank-based genome scan method⁸⁶. Alternatively, a combined analysis using the original genotype data, comprising the largest scale of BP meta-analysis, established genome-wide significant loci linked to BP on chromosome arms 6q and 8q⁸⁷. These inconsistent results demonstrate that the linkage studies have low power to detect low effect-size genomic loci.

Case-control association studies have been thought to be more powerful than linkage studies at detecting genes with small effect sizes, when performed with an adequately effective sample size. This method tests whether the allele or genotype frequencies differ significantly between cases and controls cohorts. It was initially applied to candidate genes, which were selected based on biological function or positional linkage associated with the disorder. The biological function consists of known and hypothesized functional pathway related to the disorder, or the target proteins of the antipsychotic drugs. For instance, genes involved in dopamine or serotonin neurotransmission tend to be implicated in psychiatric disorders. Approximately 1008 genes have been documented in the SzGene database – an archive of all the candidate gene studies⁸⁸ and meta-analyses of SCZ. Amongst those genes with modest effect sizes and nominal significance, *NRG1*, *DISC1*, *COMT* and *NRX1* were the top candidate genes. However, most of the associations have yet to be confirmed by meta-analysis, independent association studies or functional studies.

1.3.4 The family-based study designs in neuropsychiatric genetics research

Family-based designs are unique in that they use relatives to assess the genetic and molecular epidemiology of disease. The most commonly used studies are of familial aggregation, twins, segregation, linkage, and association. The first three designs evaluate the potential genetic basis of disease using patterns of coaggregation, and the last two directly evaluate genetic markers, usually across the entire human genome, to look for potential risk factors⁸⁹.

The clustering of disease within families usually suggests that a disease may have a genetic component. The familial correlation of a trait could be estimated by comparing the overall population prevalence with the risk of disease to other family members based on their relatedness. Twin studies are more direct evidence of the genetic involvement, assuming they share the same environmental factors. The concordance rate of disease among monozygotic and dizygotic twins is the most commonly used method to calculate the heritability of a disease. Segregation analysis is a type of method performed on family data to establish the genetic inheritance of disease, by testing models of varying degrees of generality. Large pedigrees with many affected individuals are particularly informative both for establishing that genetic component is important and for identifying specific genes. Segregation analysis can be incorporated into further linkage analysis and association analysis, aiding on determining the best-fitting model for model-based linkage analysis and increasing power⁸⁹.

For many years, linkage analysis was the primary tool used for the genetic mapping of Mendelian and complex traits with familial aggregation. In linkage analysis, by investigating the cosegregation of genetic markers and a disease trait within families, one infers that the disease-causing variants are nearby the markers. Linkage analysis has been greatly successful for mapping Mendelian traits but also notably successful in mapping variants that confer susceptibility to common diseases⁹⁰. Parametric (model-based) linkage analysis is used with large pedigrees and non-parametric (model-free) linkage analysis is often used with affected sib-pairs. Linkage can be performed using all or a subset of markers, as single-point linkage analysis takes information from one marker at a time and multi-point combines information from closely spaced markers. The latter provides more power but requires more computational power. Many linkage studies have been conducted for both SCZ and BP. However, several linked loci were likely false positives due to lack of replication across multiple independent studies⁸³.

Typical family designs of linkage include: parent–offspring trios; affected sibling pairs (sib-pairs); unselected sib-pairs or related individuals selected from the extremes of a quantitative trait distribution (for example, concordant or discordant sib-pairs); extended pedigrees with multiple affected individuals; consanguineous families; and families obtained from isolated populations. One of these designs, or a combination of them, may be chosen depending on the questions to be investigated⁹⁰. Discordant sib-pairs have been useful in association analyses of SCZ in the Indonesian population⁹¹.

Linkage analysis lost its predominance to linkage disequilibrium association mapping in recent years. Association studies are routinely carried out on a genome-wide basis on complex traits, examining common variants with a modest effect in large case-control populations. The most common family-based case-control designs for association studies are the use of case-parent trios (Transmission Disequilibrium Test, TDT) and sibling controls. The case-parent analysis looks across numerous trios to assess whether a specific allele or combination of alleles is preferentially transmitted to the cases, indicating an association between the corresponding allele and disease. This case-parent design has been extended to add additional family members, and it is very efficient for rare diseases. A common problem with the TDT is missing parental data, which could lead to bias. Family-based association (FBA) studies are closer to directly identifying disease variants and help address issues of population stratifications, however, recent FBA studies were confirming the significant loci discovered by GWAS⁹².

Common variants detected by genome-wide association studies (GWAS) cannot account for much of the heritability of most common disorders. This observation led to an emerging view that rare variants could be responsible for a substantial proportion of complex diseases risk factors. This hypothesis draws attention back to linkage and other family-based methods to detect rare variants involved in disease etiology, especially with the increased availability of whole-exome and whole-genome sequence data. A recent publication investigated an Icelandic kindred containing ten individuals with psychosis (SCZ, schizoaffective disorder or psychotic bipolar disorder) and found all affected individuals carry a rare nonsense mutation in the gene *RBM12*, and this association was replicated in a Finnish family in which a second *RBM12* truncating mutation segregates with psychosis⁹³. A number of studies combining linkage analysis and WES/WGS reported the genetic contribution to BP^{94,95}.

A combination of linkage and association methodologies should provide the most robust and powerful approach to identify and characterize the full range of disease-susceptibility variants⁹⁰. Family study designs contribute to this combined approach by providing not only the ability to enrich for genetic loci containing rare variants, but also by: providing methods to control for heterogeneity and population stratification; allowing direct estimates of the genetic contribution of different loci; making it possible to follow the transmission of variants with phenotypes; revealing the effects of parental origin of alleles and other applications.

1.3.5 Common disease – common variants (CD/CV) hypothesis on the genetics of schizophrenia and bipolar disorder

“Common disease – common variants hypothesis” implies that a disease is caused by a combination of separate common alleles of modest effect. Since 2007, GWAS have been productive in psychiatric disorders through the development of high-throughput genotyping chips, the documentation by the HapMap Consortium⁹⁶, the 1000 Genomes Project⁹⁷, covering informative SNPs across genomes of different populations, and the collaborative effort of the Psychiatric Genomics Consortium (PGC). GWAS do not rely on any *a priori* selected candidate genes, as they investigate the associations between individual common genomic variations and disorders.

Several large GWAS have been performed both on SCZ and BP. Selected studies are summarized in **Table I**, which includes the sample size of the study, the population ancestry, the number of genome-wide significant loci, and how many new loci were reported. This list is based on the NHGRI-EBI catalog of published GWAS and includes only the studies concentrating on the main SCZ/BP phenotypes, rather than endophenotypes, and the studies that were sufficiently powerful or representative and unique to a new population. As it is shown in the summary table, there is a linear relationship between the discovery sample size and number of reported loci from GWAS, according to their effect sizes for a trait. The statistical framework behind the study design of GWAS is consistent: a stringent significance level ($p\text{-values} < 5 \times 10^{-8}$) is usually set to account for type I error (false-positive) rate. Empirically, SNPs with a p -value less than this threshold are well replicated, which means that type I errors are well controlled. It also indicates that the

associations of SNPs which cannot be replicated had type II errors (false negative) due to their small effect size. Currently, in order to avoid the type II errors, efforts have been made to perform meta-analyses of the GWAS summary statistics or mega-analyses of raw genotype data (not only summary statistics), therefore increasing the sample size and statistical power⁹⁸.

The cumulative number of loci that have been reported for SCZ⁹⁹ and the expected number of BP risk loci¹⁰⁰ that could be found through GWAS were calculated (as shown in Figure 4). In the last decade, the sample sizes of SCZ and BP GWAS have increased from one thousand to one hundred thousand (with approximately equal case to control ratios), and they are still increasing with the aggregation of samples and data across organizations worldwide. This increase makes further discoveries on the pleiotropic nature of psychiatric disorders promising.

The largest schizophrenia GWAS to date using case-control samples (34,241 cases and 45,604 controls from PGC2) of mainly European ancestry have identified 128 significant independent associations spanning 108 conservatively defined loci¹⁰¹, which has provided substantial evidence on previously documented polygenic contribution to SCZ^{72,102}. The polygenic component discovered through GWAS is similar to the results of meta-analyses of other complex traits such as human height¹⁰³, inflammatory bowel disease¹⁰⁴ and breast cancer¹⁰⁵. The most notable associations in this study, which are relevant to major hypotheses of the etiology and treatment of schizophrenia, include *DRD2* – the target of antipsychotic drugs; genes such as *GRM3*, *GRIN2A*, *SRR*, *GRIA* – involved in glutamatergic neurotransmission and synaptic plasticity; and associations with *CACNA1C*, *CACNB2* and *CACNA1I* – encoding voltage-gated calcium channel subunits (this family of proteins have extended previous implications in SCZ and other psychiatric disorders)^{70,71,106–108}. Those discovered loci are also enriched in genes containing *de novo* mutations in schizophrenia, autism, and intellectual disability¹⁰¹. Based on the PGC findings on schizophrenia, people estimated that 8,300 independent, mostly common SNPs, contribute to risk for schizophrenia, and these collectively account for 32% of the variance¹⁰⁶. Further, an overwhelmingly polygenic disease architecture in which $\geq 71\%$ of 1-Mb genomic regions harbor ≥ 1 variant influences schizophrenia risk¹⁰⁹. The highly polygenic nature of the common variants contributing to the risk of SCZ are widely replicated. About 75% of the 108 loci continued to be genome-wide significant in the trans-ancestry analysis with the combination of a Chinese schizophrenia cases/controls and the data from PGC2¹¹⁰. The same study has identified

30 novel genome-wide significant loci, four of which were only significant in the Chinese sample. These findings indicated that most schizophrenia risk loci were shared across two ancestral populations. However, it also suggested common variants explaining the genetic variance are only partially overlapping between European and Chinese populations.

Despite the increasing and unequivocal evidence for common SNPs contributing to schizophrenia risk, some important factors about the GWAS findings should be noted¹¹¹:

1) The associations are to genomic regions (loci), and not to genes. It is not certain which gene is involved for some of the loci since they encompass more than one gene, such as the major histocompatibility complex (MHC) locus. Functional evidence is arising for the involvement of risk alleles: a major study reported structurally diverse alleles of complement component 4 genes (*C4A* and *C4B*) in the MHC locus, which generated widely varying levels of *C4A* and *C4B* expression in the brain, and the allelic association to schizophrenia is related to increased expression of *C4A*¹¹².

2) Almost all the schizophrenia-associated SNPs are in non-coding regions of the genome, either intergenic or intronic, and the scarcity of evidence makes identifying the biological basis of these associations challenging. Further proof for these associations and their potential for therapeutic targeting calls for both caution and collaborative effort.

Similar to the case of GWAS studies on schizophrenia, the GWAS on bipolar disorders started with smaller sample size (as summarized in **Table I**), and therefore most of the susceptibility loci were not replicated. The most often replicated genes are *ANK* and *CACNAIC*^{108,113}. A milestone study was published in 2011 by PGC with a discovery dataset of 7,481 European ancestry cases and 9,250 European ancestry controls⁷¹. They identified a new intronic locus in *ODZ4*, and they confirmed the genome-wide significant evidence of association for *CACNAIC*, though the odds ratio of the susceptibility were both at 1.14 (combined p-values are 1.52×10^{-8} and 4.40×10^{-8} for *CACNAIC* and *ODZ4* respectively), which held a similar magnitude to the risk for schizophrenia¹⁰¹. The small effect size of the associated SNPs makes the signals undetectable under certain sample sizes, hence researchers have attempted to increase the sample size of the discovery GWAS after the PGC study, as shown in **Table I**. However, the number of novel susceptibility loci/genes for bipolar disorder was limited. Subsequent GWAS studies reported novel significant association inside or near genes/regions: an intergenic region on

9p21.3, *EBBB2*, *TRANK1*, *MAD1L1*, *ADCY2*, a region between *MIR2113* and *POU3F2*^{100,114,115}, for which the functional connection to BP is still uncertain. The largest non-European GWAS conducted on the Japanese population (2,964 BP cases and 61,887 controls) found a novel susceptibility locus at 11q12.2, a region known to contain regulatory genes for plasma lipid levels (*FADS1/2/3*). The most recent GWAS study by the PGC reported 30 susceptibility loci including 18 novel ones¹¹⁶; the sample size of the discovery GWAS was tripled compared to their publication in 2011. It was comprised of 20,352 cases and 31,358 controls of European descent and combined an independent sample of 9,412 cases and 137,760 controls. These significant loci contain genes encoding ion channels and neurotransmitter transporters (*CACNA1C*, *GRIN2A*, *SCN2A*, *SLC4A1*), synaptic components (*RIMS1*, *ANK3*), immune and energy metabolism components, and multiple potential therapeutic targets for mood stabilizer drugs. It is also noteworthy that trans-ethnic replication analysis in BP GWAS could be a reasonable way to pinpoint the genuine susceptibility genes, based on the evidence that *FADS* genes were associated with BP in the new PGC data. In sum, there are approximately 40 loci that are significantly associated with the risk of BP from major GWAS studies, with the estimated variance explained by polygenic risk scores (based on the largest GWAS so far) being ~8% – 4% on the liability scale.

One could use GWAS data from human studies to create genetic predictors for disease and other complex traits by estimating the effect size at multiple loci in a discovery sample and using those estimated SNP effects in independent samples to generate a polygenic risk score (PRS) per individual. Additionally, they found that bipolar disorder type I is strongly genetically correlated with schizophrenia, while bipolar disorder type II correlated more with major depression.

1.3.6 Common disease – rare variants (CD/RV) hypothesis on the genetics of schizophrenia and bipolar disorder

In contrast to the “common disease – common variants” model, which implies that a disease is caused by combinations of separate common alleles of modest effect, the alternative model of “common disease – rare variants” hypothesizes that some mutations predisposing to diseases are highly penetrant, individually rare, and of recent origin, even being specific to single cases or families¹¹⁷. A strong effect of the variants is possibly due to the severely reduced fitness of affected patients with schizophrenia and bipolar disorder¹¹⁸ (bipolar disorder appearing to be under weaker

negative selection). Since there is emerging evidence on the involvement of rare variants in schizophrenia and bipolar disorder, this thesis will introduce them by variant type separately.

Rare copy number variants (CNVs)

Copy number variants (CNVs) are chromosomal deletions and duplications that range in size from kilobases to megabases of DNA sequence, and they usually cannot be detected through conventional karyotyping¹¹⁹. The wide usage of microarrays made the discovery of CNVs accessible in large cohorts of patients and controls. Four major mechanisms account for the formation of CNVs: non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS), and L1-mediated retrotransposition¹²⁰. NAHR is responsible for forming the recurrent CNVs at the same chromosomes positions flanked by region-specific, highly repetitive DNA sequences, called low copy repeats (LCRs), which are DNA segments previously duplicated during evolution. Recombination between adjacent and homologous LCRs can occur and leads to deletions or duplications of the DNA stretches between the repeats¹²¹. NHEJ is more error-prone than NAHR because it occurs due to the aberrant repair of DNA double-strand breaks and is guided by the information contained within or near the DNA lesion for repair. It usually forms the breakpoints of CNVs within repetitive elements and it doesn't require extensive sequence homology. Breakpoint analysis revealed that 70.8% of the deletions were attributed to either a nonhomology-based mechanism (i.e. NHEJ) or microhomology-mediated breakpoint-induced replication (generalization of the FoSTeS mechanism), and 89.6% of the insertions/duplications were attributable to retrotransposition activity¹²².

The first and most replicated evidence of structural variants being involved in psychiatric disorders is the case of 22q11.2 microdeletions, which were originally found to be associated with velo-cardio-facial syndrome (VCFS) through karyotyping¹²³. The individuals with VCFS have high rates of psychiatric disorder, especially schizophrenia¹²⁴. There is evidence demonstrating an increased prevalence of chromosome 22q11.2 deletions in schizophrenia patients compared to controls¹²⁵. Additionally, this deletion is associated with multiple neuropsychiatric disorders, such as autism spectrum disorder (ASD) and intellectual disability (ID). Among the genes located in this region, two candidate genes are proposed to contribute to the SCZ phenotype: catechol-o-methyl transferase (*COMT*) and proline dehydrogenase 1 (*PRODH*). *COMT* encodes the

postsynaptic enzyme known to regulate the degradation of dopamine and *PRODH* encodes an enzyme responsible for glutamate production in the mitochondria.

New deletions including 1q21.1 (>1 Mb multi-allelic CNVs), 15q11.2 (470 kb) and 15q13.3 (~1.5Mb) were firstly implicated in schizophrenia^{120,126} in 2008, and these findings were replicated in follow-up studies. Both CNVs in ancestrally matched schizophrenia cases and controls⁶⁰ and large recurrent CNVs proposed to be under negative selection (because of reduced fecundity associated with schizophrenia) were examined⁵⁹. These CNVs have also been observed in other patients with autism, mental retardation and other psychiatric disorders. Deletions of *NRXNI* have been strongly linked to schizophrenia^{127,128}. The successful replication of the *NRXNI* CNVs associated with SCZ might have been a result of high mutation rates and the negative selection acting against them. It has also been shown that rates of these CNVs stay similar in different populations and they are not affected by genetic drift. Meanwhile, the extreme rarity of pathogenic CNVs (frequency is usually less than 1%) even in patient populations requires very large sample size for reaching sufficient statistical power¹²¹. First, scattered studies were used to confirm previously reported CNVs and discover novel ones. Then, large consortia and collaborations were very helpful for increasing the sample size and ruling out false positives. How the most significant CNVs were originally discovered and further replicated in the largest CNV GWAS (SCZ cohort of 21,094 cases and 20,227 controls) to date is summarized in **Table II**. The frequency in cases and controls, the CNV effect sizes and significance levels are also included. The odds ratios of these CNVs range from approximately 2 to 60, with some extremely high effect sizes, such as for 22q11.2 and 3q29. CNVs with higher risk to develop SCZ (higher odds ratios) are rarer (lower frequencies in the population), because higher pathogenic mutations are eliminated from the population faster, due to lower fecundity among their carriers (the SCZ patients). Interestingly, these risks and protective CNVs were recurrent, predominantly mediated by NAHR mechanism¹²⁹.

Most SCZ-associated CNVs are also risk factors for developmental delay (DD) and autism spectrum disorders (ASD)¹³⁰, and the frequencies of CNVs are even higher in DD and ASD compared to SCZ. Although there is no obvious evidence for the genetic link between SCZ and ASD/DD in family or twin studies, they did share the neurodevelopmental component on the hypothesized neuropsychiatric spectrum. In contrast, studies on BP have not produced robust

results. Some studies showed the accumulation of rare *de novo* CNVs in patients with BP, especially in those with early-onset BP^{77,131}. These findings could not be replicated, probably due to small discovery sample size. Most CNVs implicated in SCZ did not play a significant role in BP, except duplications at 1q21.1, 16p11.2 and deletions at 3q29^{120,132}, despite BP being known to share a genetic component with SCZ. A possible explanation could be that SCZ patients suffer from cognitive deficits, while BP patients are more cognitively functional and less impairment-persistent. Structural variants such as CNVs associated with cognitive problems may therefore play a smaller role in BP. It is less clear whether large rare (or *de novo*) CNVs in BP have a smaller magnitude of contribution relative to ASD or SCZ.

Rare single nucleotide variants (SNVs) and indels

Next-generation sequencing (NGS) technology has revolutionized genomic research since its emergence and has empowered researchers studying health sciences. Researchers in neuropsychiatric genetics have applied both whole-exome sequencing (WES) and whole-genome sequencing (WGS) in finding genetic components of schizophrenia, bipolar disorder and other neuropsychiatric disorders. WES has allowed the identification of variants within the 1% protein-coding regions (exons of genes) of the genome (the exome). WES has allowed scanning for variants at a single-base resolution, i.e. SNVs and indels, which are not detected through microarray genotyping. The variations in the exome are likely to have more severe consequences than variations in the remaining 99% genome, since the exome are protein-coding. WGS has not often been applied to large-scale studies due to its high cost per individual. Emerging studies that used WES and WGS to explore SNV and indels in schizophrenia and bipolar disorder, either in family-based design or population-based design, will be discussed.

A family-based design is a powerful way to investigate rare and highly penetrant variants. One study conducted in families with multiple affected members from a Caucasian ancestry, has reported novel private missense variants within *SHANK2* and *SMARCA1* (X-linked)¹³³. Both genes are noteworthy, as the SHANK protein family and the SMARCA protein have multiple plausible connections to schizophrenia and brain function¹³³. In this study, they have examined ninety individuals across nine families with two to six individuals diagnosed as having schizophrenia or schizoaffective disorder. Another outstanding study investigated an Icelandic kindred containing ten individuals with psychosis (schizophrenia, schizoaffective disorder or psychotic bipolar

disorder) and the authors found all affected individuals carried a nonsense mutation in the *RBM12* gene⁹³. They replicated the association in a Finnish family (with five individuals affected by psychosis) in which a second *RBM12* truncating mutation segregates with psychosis. Even though the variants were not fully penetrant for psychosis, they found that carriers unaffected by psychosis resembled patients with schizophrenia in their non-psychotic phase and in their neuropsychological test profile, as well as in their life outcomes. *RBM12* had not been associated with psychosis previously, but it may help to understand the pathogenesis of psychosis or lead to new targets for drug development. This work also provided a template for future familial studies of psychosis: the mutations involved are likely to be recent, incompletely penetrant but to lead to related phenotypes in carriers unaffected by psychosis. They are also likely to act together with other sequence variants.

A combined family-based and case-control approach was used with 36 affected members with BP from 8 multiplex families, and a follow-up meta-analysis in 3 independent case-control samples¹³⁴. The WES revealed an enrichment of rare segregating variants for gene sets previously identified in *de novo* studies of autism and schizophrenia and for targets of the fragile X mental retardation protein (FMRP) pathway. Similar studies carried in BP multiplex families also shed light on the disease pathology. WGS of 41 families, comprising 200 individuals affected with BP, lead to the discovery that these pedigrees had an increased burden of rare variants in genes encoding neuronal ion channels, including subunits of GABA_A receptors and voltage-gated calcium channels. Targeted sequencing of 26 of these candidate genes in an additional 3,014 cases and 1,717 controls confirmed rare variant associations¹³⁵. Cruceanu *et al.* (2017) explored highly penetrant rare variants in 40 well-characterized multiplex families (186 exomes, three to seven affected individuals across one to three generations) and found rare variants segregating with the disease in many genes of clinical interest; an enrichment of deleterious variants in G protein-coupled receptor (GPCR) family genes was observed, which are potentially important drug targets. One variant in particular, a rare and functionally relevant nonsense mutation in the *CRHR2* gene, as a member of GPCR family and the hypothalamic-pituitary-adrenal axis, segregated well in one of these families¹³⁶. Another study combining WES, CNV and linkage analysis in 15 BP families (72 out of 117 subjects are affected) reported that rare predicted pathogenic variants shared among ≥ 3 affected relatives were overrepresented in postsynaptic density (PSD) genes, with no enrichment in unaffected relatives. However, they found no difference in genome-wide burden of

likely gene-disruptive variants in affected versus unaffected relatives. They emphasized the observation of heterogeneity within and between families and a probable genetic model involving variants of modest effect and reduced penetrance⁹⁴.

Most recent, large-scale studies of rare variants in schizophrenia have used case-control approaches. They either showed an enriched burden of gene sets or pinpointed to a specific gene. Purcell et al. (2014) used exome sequences of 2,536 SCZ cases and 2,543 controls to demonstrate a polygenic burden primarily arising from rare (less than 1 in 10,000) disruptive mutations across many genes¹³⁷. These variants were particularly enriched the following gene sets: the voltage-gated calcium ion channel and the signaling complex formed by the activity-regulated cytoskeleton-associated scaffold protein (ARC) of the postsynaptic density, gene sets previously implicated by GWAS and CNV studies in SCZ and targets of the FMRP pathway. Analysis on WES of an extended dataset comprising 4,264 SCZ cases, 9,343 controls and 1,077 trios identified a genome-wide significant association between rare loss-of-function (LOF) variants in *SETD1A* (also known as *KMT2F*, encoding one of the methyltransferases that catalyzes the methylation of lysine residues in histone H3) and risk for schizophrenia¹³⁸. Ten individuals with SCZ carried *SETD1A* LOF variants compared to only two heterozygous LOF variants in 45,376 exomes from individuals in the general population. They also identified carriers of LOF variants in *SETD1A* among samples with severe developmental disorders and notable neuropsychiatric phenotypes. They suggested the epigenetic dysregulation, specifically in the histone H3K4 methylation pathway, is an important mechanism in the pathogenesis of SCZ, based on the evidence that LOF variants in other genes in the same protein family result in dominant Mendelian disorders characterized by severe developmental phenotypes including intellectual disability. By analyzing coding-sequence and splice-site ultra-rare variants (URVs among 4,877 SCZ cases and 6,203 controls) that were present in only 1 of 12,332 unrelated Swedish exomes and never seen in the Exome Aggregation Consortium (ExAC, 45,376 non-psychiatric individuals) cohort, researchers found that gene-disruptive and putatively protein-damaging URVs are more abundant among individuals with SCZ than among controls ($p = 1.3 \times 10^{-10}$)¹³⁹. This elevation rate was several times larger than an analogously elevated rate for *de novo* mutations, suggesting that most rare-variant effects on SCZ risk are inherited. These genes with URVs were concentrated on brain-expressed genes and genes whose RNAs have been found to interact with synaptically localized

proteins, suggesting synaptic dysfunction may mediate a large fraction of strong, individually rare genetic influences on SCZ risk.

The role of rare variants in bipolar disorder has not yet been tested in large-scale population-based studies comprehensively¹⁴⁰. The use of isolated populations might help finding variants with a recent origin, which may have drifted to higher frequency by chance. With the WES of 28 BP cases and 214 controls from the Faroe Islands and follow-up in a British sample of 2025 cases and 1358 controls¹⁴¹, 17 variants in 16 genes in the single-variant analysis and 3 genes in the gene-based statistics were exome-wide significant. The replication confirmed the association with *NOS1* and *NCL* but didn't support the association of two genes (*PITPNM2* and *PIK3C2A*) in significant BP and SCZ GWAS loci. In this sense, large-scale WES and WGS on BP cases and matched controls are necessary to explore the contribution of rare variants in BP pathogenesis.

***De novo* mutations**

De novo mutations (DNM) are genetic alterations that are present for the first time in one family member as a result of a mutation in a germ cell (egg or sperm) of one of the parents, or a variant that arises in the fertilized egg itself during early embryogenesis. Additionally, novel mutations continue to arise through post-natal and adult life in both somatic and germ cells. Only mutations present in germline cells can be transmitted to the next generation¹⁴². A typical human genome varies at 4.1 million to 5.0 million sites from the reference human genome, and only 40,000 to 200,000 of them have a frequency < 0.5% in a population¹⁴³. Those rare genetic variations must have occurred as *de novo* mutations in an individual during human evolution, at the germline *de novo* mutation rate for SNVs in humans of 1.0 to 1.8×10^{-8} per nucleotide per generation. This number translates into 44 to 82 *de novo* single-nucleotide mutations in the genome of an average individual, with 1 to 2 affecting the coding sequence. Moreover, around 2.9 to 9 small *de novo* indels (<50bp), ~0.16 large *de novo* indels and ~0.0154 *de novo* CNVs (larger than 100 kb in length) are also present in an average individual^{144,145}.

De novo mutation frequencies vary between individuals and over time within an individual. The *de novo* rate of SNVs and CNVs shows strong parent-of-origin biases as well as parental age effects, and could be due either to local genomic architecture (segmental duplications) or to variations in specific genes (such as *PRDM9*) mediating homologous recombination¹⁴⁶. The

epidemiological studies provided robust evidence for an association of advanced paternal age to ASD and SCZ¹⁴⁷, as well as a risk factor for BP¹⁴⁸.

High-resolution genomic microarrays allowed the unbiased genome-wide analysis of *de novo* CNVs long before the same could be realized for *de novo* SNVs and indels. Genome-wide CNV data of a case-control study found more *de novo* and rare CNVs in SCZ cases with adult (15%) or young-onset (20%) than in controls (5%), and this association was independently replicated in patients with childhood-onset SCZ as compared with their parents¹⁴⁹. Another study tested directly for association of *de novo* CNVs with SCZ, reporting a frequency of 10% (15 out of 152) in sporadic cases, 1.3% among (2 out of 159) unaffected individuals and none in 48 familial cases¹⁵⁰. Notably, the results of this study confirmed the importance of microdeletions in the 22q11.2 locus as three *de novo* 22q11.2 microdeletions were identified¹⁵¹. Similar to the findings of ASD, around 1% of SCZ cases carries two or more *de novo* CNV events. These two studies highlighted their findings by including notable candidate genes or regions such as *ERBB4*, *SLC1A3*, *RAPGEF4*, *CIT*, *NRXN1* and the 16p11.2 region discussed earlier¹⁵².

Before the popularization of exome sequencing, the search for *de novo* SNVs was mainly in candidate genes that encode proteins known to have physiological roles at the synapse. One study evaluated the contribution of *de novo* SNVs in the synaptic scaffolding protein *SHANK3* in 185 SCZ patients with unaffected parents and 285 unrelated controls and reported 2 *de novo* mutations in this gene¹⁵³. The same group systematically re-sequenced 401 synapse-associated genes in 142 individuals with ASDs and 143 individuals with SCZ, and calculated a direct *de novo* mutation rate which is similar to previous indirect estimates, but a significant excess of potentially deleterious DNMs was observed in ASD and SCZ patients¹⁵⁴.

Girard *et al.* (2011) sequenced the exomes of 14 SCZ probands and their parents and identified 15 DNMs in eight probands. This is significantly more than expected considering the previously reported DNM rate (2.59×10^{-8} in this study versus $1.0-1.28 \times 10^{-8}$ per position in a haploid genome). Additionally, 4 of the DNMs are nonsense mutations, which is more than what is expected by chance. This study suggested that DNMs may account for some of the heritability of SCZ while providing a candidate gene list¹⁵⁵. In the same year, Xu *et al.* also examined rare *de novo* protein-altering mutations in the exomes of 53 sporadic SCZ cases, 22 unaffected controls and their parents. They reported 40 DNMs in 27 cases affecting 40 genes, including a potentially

disruptive mutation in *DGCR2*, a gene located in the SCZ-predisposing 22q11.2 microdeletion region. They assessed the evolutionary conservation of the affected nucleotide by using the phyloP conservation score, and the comparison between DNMs and privately inherited variants in sporadic cases of SCZ showed a statistically significant shift towards higher phyloP scores for DNMs. Based on the thorough comparison to rare inherited variants in SCZ cases, these identified DNMs tended to show a large excess of non-synonymous changes as well as a greater potential to affect protein structure and function. They proposed a major role and a large mutational target for DNMs in the high incidence and persistence of SCZ¹⁵⁶.

One study, which primarily focused on testing the association of *de novo* CNVs with BP in 185 trios, found that the frequencies of *de novo* CNVs were significantly higher in BP as compared with controls (426 trios), and the *de novo* DNMs were enriched among cases with an age at onset younger than 18⁷⁷. A total of 23 *de novo* CNVs were detected and validated in their sample. They have also tested the SCZ samples from Xu *et al.*¹⁵⁰ with the same methodology to confirm the high rate of *de novo* CNVs in SCZ. A similar study design, using 662 Bulgarian SCZ trios and 2,623 Icelandic controls, also reported a higher frequency of rare *de novo* CNVs in cases (5.1%) compared with controls (2.2%)¹⁵⁷. They detected *de novo* CNVs that occurred at known SCZ loci and are known as pathogenic for other genomic disorders. Most significantly, multiple *de novo* CNVs spanned genes encoding members that are components of the postsynaptic density (PSD). The systematic analysis showed that *de novo* variants in cases were enriched for the PSD proteome by merging novel CNVs and proteomics data sets.

With the combination of a USA cohort and the application of WES of previously reported Afrikaner probands, the same group observed an excess of *de novo* nonsynonymous SNVs as well as a higher prevalence of gene-disruptive DNMs in cases relative to controls¹⁵⁸. They found recurrent *de novo* events within or across two distinct populations in four genes (*LAMA2*, *DPYD*, *TRRAP* and *VPS39*), and they examined to what extent the *de novo* events were determined by the developmental pattern of brain expression of the mutated genes. The results showed that DNMs accounted for genes with higher expression in prenatal development, and they affected genes with diverse functions and developmental profiles. These findings may help to understand the genomic and neural architecture of SCZ risk.

Another study tried to identify DNMs by WES of quads and trios comprising of a proband with SCZ (sporadic/singleton case), his/her unaffected parents and an available unaffected sibling¹⁵⁹. This study was conducted on 399 persons, including 105 probands affected with SCZ, 84 unaffected siblings and 210 unaffected parents. The identified genes harboring *de novo* potentially damaging mutations in probands with SCZ. Those genes were then subject to co-expression analysis in different brain regions across development stages and protein interaction profiles. The results showed that 54 genes with damaging DNMs were significantly enriched in the dorsolateral and ventrolateral prefrontal cortex during fetal development, and they are involved in neuronal migration, synaptic transmission, signaling, transcriptional regulation and transport. This study is further evidence that disruptions of fetal development are critical to the pathophysiology of SCZ and is the first to apply genomic and transcriptome analyses to map critical neurodevelopmental processes in time and space in the brain. The enrichment of genes with DNMs was also detected in a Chinese SCZ cohort (45 trios) in transcriptional co-expression profile in prenatal frontal cortex and in prenatal temporal and parietal regions, and four genes (*LRP1*, *MACF1*, *DICER1* and *ABCA2*) harboring DNMs were prioritized¹⁶⁰.

The largest exome sequencing study of DNMs in SCZ to date was published by Fromer *et al.* in 2014, using 623 SCZ trios¹⁶¹. They report: 1) DNMs affecting protein sequences occur in SCZ not at higher than expected rates; 2) genic recurrence of DNMs in SCZ is significant, especially with an increased case/control ratio of rare LOF variants; 3) genes with DNMs are enriched in specific biological processes pathogenic in SCZ; small DNMs (SNVs and indels) in particular are overrepresented among glutamatergic postsynaptic proteins, comprising activity-regulated cytoskeleton-associated protein (ARC), N-methyl-D-aspartate receptor (NMDAR) complexes and fragile X mental retardation protein (FMRP); 4) genes with small DNMs are also enriched in the *de novo* genes implicated in other neurodevelopmental disorders, including autism and intellectual disability. These results were aligned to a parallel case-control study of rare variants¹³⁷ (mentioned earlier in this chapter) and they were consistent to reported the robust and reproducible enrichment signals of LOF variants in ARC complex.

Subsequent studies consistently found an increased proportion of nonsense DNMs and their occurrence in genes less tolerant to rare variation in sporadic probands, and genes with those DNMs overlapped with genes implicated in autism (such as *AUTS2*, *CHD8* and *MECP2*) and intellectual disability (for example, *HUWE1* and *TRAPPC9*)¹⁶². Interesting DNM genes discovered

in one ethnicity through WES and further re-sequenced in ethnically diverse SCZ cases resulted in finding extremely rare and potentially damaging variants in those genes, which could illuminate risk genes that increase the propensity to develop SCZ across ethnicities¹⁶³. Additionally, functional exploration of the gene *TBLIXR1* (previously associated with autism and epilepsy) harboring a DNM in a Japanese cohort (18 trios) concluded that they could alter Wnt/ β -catenin signaling activity, through altering the interaction of *TBLIXR1* with N-CoR and β -catenin.

Alternative view on the contribution of *de novo* synonymous mutations reported that *de novo* near-splice site synonymous mutations changing exonic splicing regulators and those within frontal cortex-derived DNase I hypersensitivity sites are significantly enriched in ASD and SCZ, respectively. *SETDIA* was again found to harbor multiple functional *de novo* synonymous mutations¹⁶⁴. A recent major study, the Deciphering Developmental Disorders (DDD) study, comprising of 7,930 individuals with a severe and undiagnosed developmental disorder and their parents, showed that DNMs in highly evolutionarily conserved fetal brain-active elements are significantly and specifically enriched in neurodevelopmental disorders, which established a robust estimate of the contribution of DNMs in regulatory elements to genetically heterogeneous disorders¹⁶⁵.

The first trio-based WES study on BP (79 probands) to investigate potential roles of DNMs in the disease pathogenesis of BP found significant enrichment of genes highly intolerant to protein-altering variants in the general population, similar to aforementioned reports in autism and SCZ¹⁶⁶. They also observed significantly earlier disease onset among the BP probands with *de novo* protein-altering mutations when compared to non-carriers. However, the gene ontology enrichment analysis did not identify any significant enrichment.

In summary, the historical genetic studies and molecular genetics in the last decade or so demonstrated strong evidence for the genetic contribution to the development of SCZ and BP. Recent studies have shown that the genetic architecture of these two neuropsychiatric disorders is very complex, heterogeneous, and likely follows an omnigenic model¹⁶⁷. The convergence of common and rare variant studies and their consistent overlap at a broad functional level suggests that the common disease – common variants and common disease – rare variants hypotheses are complementary to explain the pathogenesis of the disorders.

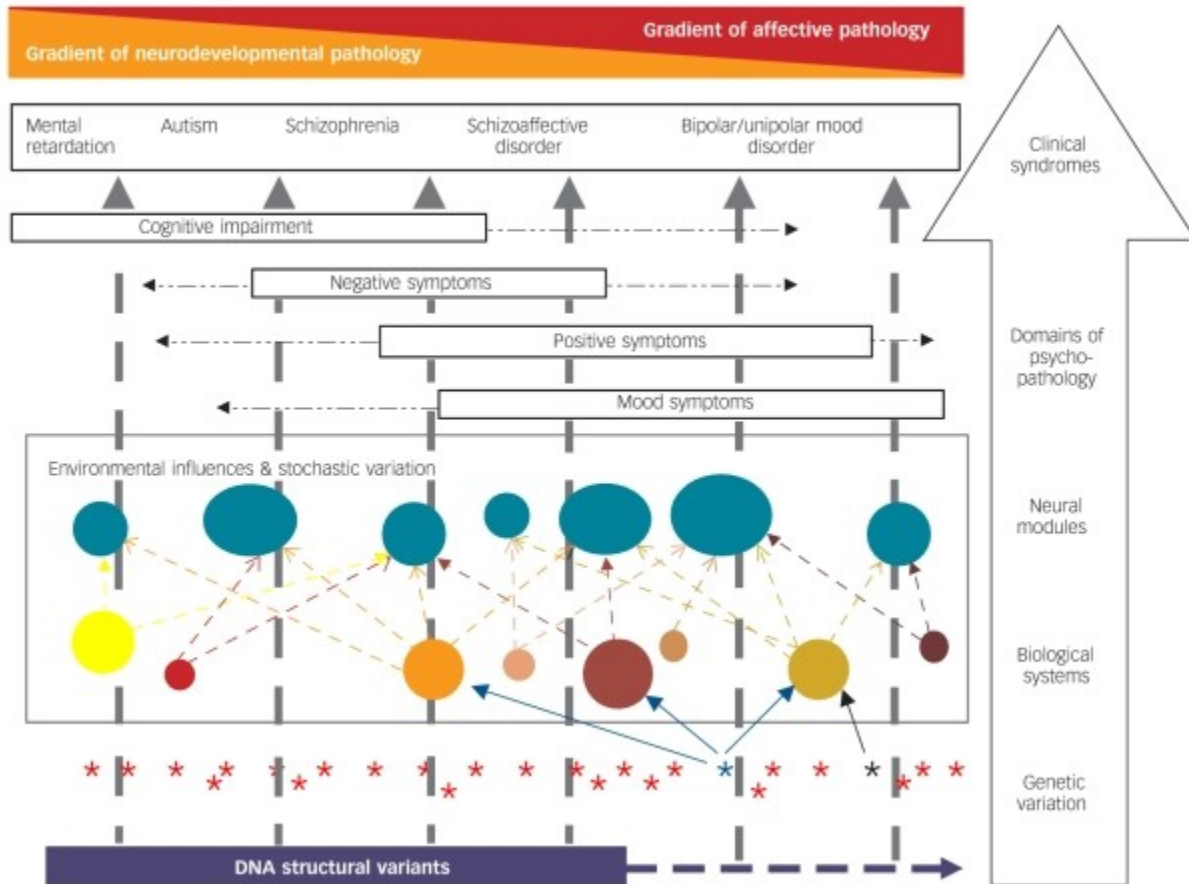


Figure 2. Hypothesized model of the complex relationship between biological variation and some major forms of psychopathology.

Reprinted from “The Kraepelinian dichotomy – going, going ... but still not gone” by Nick Craddock and Michael J. Owen, 2010, *British Journal of Psychiatry*, 196, 92-95. Copy right 2018 by Cambridge University Press. Reprinted with permission.

This figure depicts a simplified model of the relationships between genotype and clinical phenotype. From the lowest tier, the structural variations contribute particularly to neurodevelopmental disorders and are associated particularly with enduring cognitive and functional impairment, while the single-base changes in genes (shown as asterisks) may influence one or multiple biological systems based on the involvement of genes in multiple functions and their interactions with each other and with the environmental exposures/experiences historically and dynamically. Furthermore, the relevant biological systems would influence the neural modules that comprises the relevant functional elements of the brain. The (abnormal) functioning of the neural modules together influences the domains of psychopathology and ultimately the clinical

syndromes. The decreasing proportion of neurodevelopmental contribution and reciprocal increasing proportion of episodic affective disturbance are shown along a simplified axis for some major clinical syndromes.

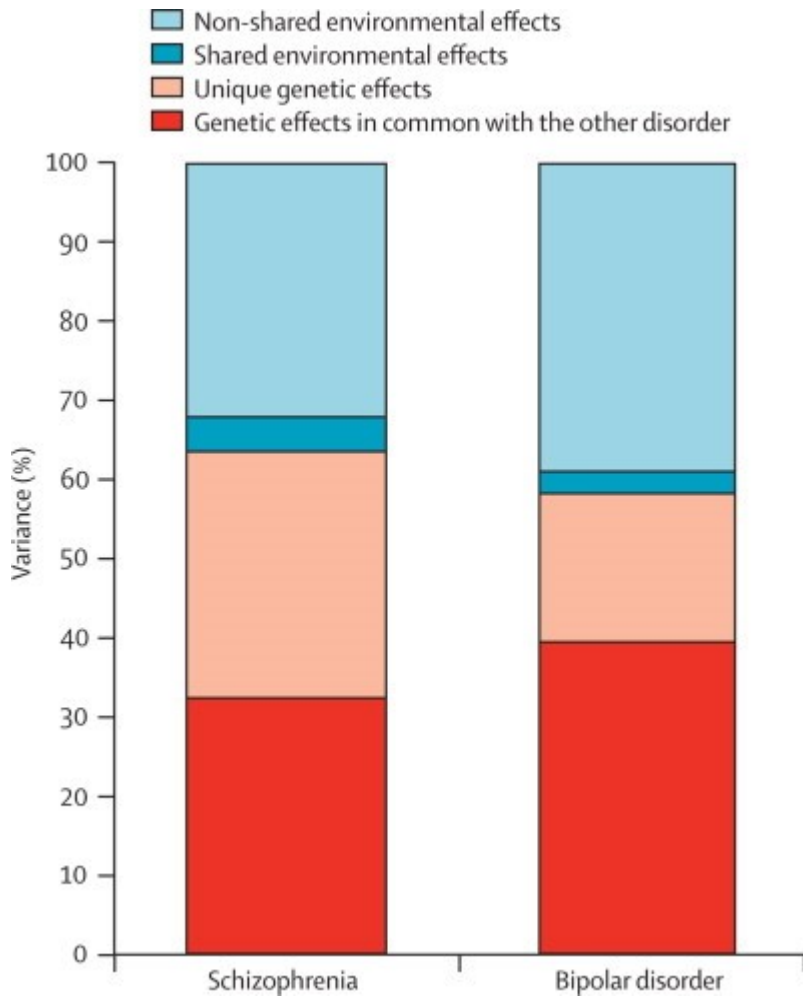


Figure 3. Variance accounted for by genetic, shared environmental, and non-shared environmental effects for schizophrenia and bipolar disorder.

Reprinted from “Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study” by Paul Lichtenstein, Benjamin H Yip, Camilla Björk, Yudi Pawitan, Tyrone D Cannon, Patrick F Sullivan, and Christina M Hultman, 2009, *The Lancet*, 373, 234-239. Copy right 2009 by Elsevier. Reprinted with permission.

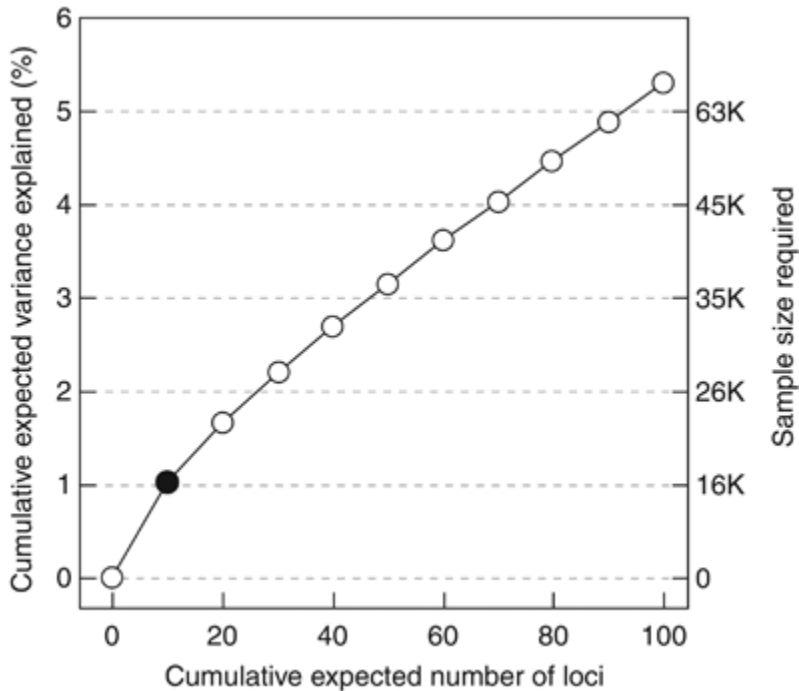
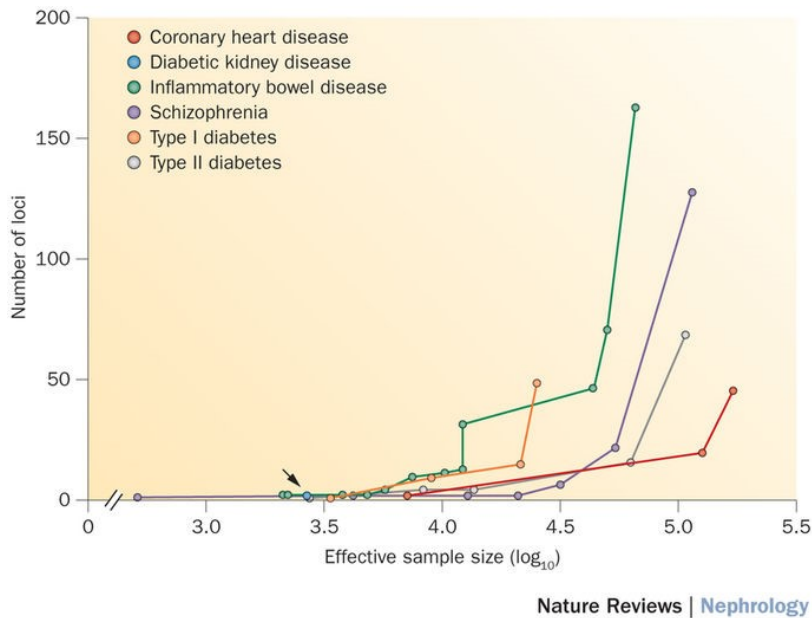


Figure 4. The relationship between the effective sample size and number of GWAS loci.

(A) The maximum number of loci reported from recent genome-wide association studies (GWAS) for selected traits, including schizophrenia, given the effective sample size. Reprinted from “The genetics of diabetic complications” by Emma Ahlqvist, Natalie R. van Zuydam, Leif C. Groop, Mark I. McCarthy,⁹⁹ 2015, Nature Reviews Nephrology, 11, 277-287, copyright 2015 by Springer Nature. Reprinted with permission; and (B) the expected

number of bipolar disorder risk loci through GWAS and its relationship with the effective sample size of the study and genetic variance explained. Reprinted from “Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder” by Chen DT et al.¹⁰⁰, 2013, *Molecular Psychiatry*, 18, 195-205. Copyright 2011 by Springer Nature. Reprinted with permission.

Table I. Major GWAS studies of schizophrenia and bipolar disorder from 2007 to 2017

PMID	Author	Publication Date	Journal	Phenotypes	GW significant loci	# new loci	Discovery cohort ancestry	Replication cohort ancestry	Discovery sample size
18347602	Sullivan PF ¹⁶⁸	3/18/2008	<i>Mol Psychiatry</i>	SCZ	0	0	European	NR	828
18677311	O'Donovan MC ⁶⁹	7/30/2008	<i>Nat Genet</i>	SCZ	1	1	European	European	3,416
19571811	Purcell SM ⁷²	7/1/2009	<i>Nature</i>	SCZ, BP	1	1	European	European	6,909
19571809	Shi J ¹⁶⁹	7/1/2009	<i>Nature</i>	SCZ	1	1	European	European	7,593
19571808	Stefansson H ¹⁷⁰	7/1/2009	<i>Nature</i>	SCZ	3	3	European	European	16,161
20713499	Huang J ¹⁷¹	8/16/2010	<i>Am J Psychiatry</i>	SCZ, BP, MDD	1	1	European	NR	4,186
20832056	Ikeda M ¹⁷²	9/8/2010	<i>Biol Psychiatry</i>	SCZ	0	0	Japanese	Japanese	1,108
21674006	Yamada K ¹⁷³	6/6/2011	<i>PLoS One</i>	SCZ	0	0	Japanese	Japanese	360
21926974	Ripke S ¹⁷⁴	9/18/2011	<i>Nat Genet</i>	SCZ	7	5	European	European	21,856
22037552	Yue WH ¹⁷⁵	10/30/2011	<i>Nat Genet</i>	SCZ	2	0	Han Chinese	Han Chinese	2,345
22037555	Shi Y ¹⁷⁶	10/30/2011	<i>Nat Genet</i>	SCZ	2	2	Han Chinese	Han Chinese	10,218
22688191	Bergen SE ⁷⁹	6/12/2012	<i>Mol Psychiatry</i>	SCZ, BP	1	0	Swedish	Swedish	4,646
22883433	ISGC & WTCCC2* ¹⁷⁷	8/7/2012	<i>Biol Psychiatry</i>	SCZ	1	0	Irish	European	3,400
23894747	Aberg KA ¹⁷⁸	2/1/2013	<i>JAMA Psychiatry</i>	SCZ	7	0	European	European	6,298

Table I. Major GWAS studies of schizophrenia and bipolar disorder from 2007 to 2017, continued

PMID	Author	Publication Date	Journal	Phenotypes	GW significant loci	# new loci	Discovery cohort ancestry	Replication cohort ancestry	Discovery sample size
23453885	Smoller JW ⁷¹	2/27/2013	<i>Lancet</i>	ASD, ADHD, BP, MDD, SCZ	4	0	European	European	61,220
23974872	Ripke S ¹⁰⁶	8/25/2013	<i>Nat Genet</i>	SCZ	22	13	European	European	32,143
24043878	Wong EH ¹⁷⁹	9/16/2013	<i>Schizophr Bull</i>	SCZ	1	1	Han Chinese	Han Chinese	2,506
24166486	Sleiman P ¹⁸⁰	10/29/2013	<i>Sci Rep</i>	SCZ, BP	6	1	European	European	48,070
24253340	Lencz T ¹⁸¹	11/19/2013	<i>Nat Commun</i>	SCZ, BP	1	1	Ashkenazi Jewish	European	2,544
25056061	Ripke S ¹⁰¹	7/22/2014	<i>Nature</i>	SCZ	108	83	European	European	87,534
26198764	Goes FS ¹⁸²	7/21/2015	<i>Am J Med Genet B Neuropsychiatr Genet</i>	SCZ	0	0	Ashkenazi Jewish	NR	4,058
26531332	Kim LH ¹⁸³	11/14/2015	<i>Am J Med Genet B Neuropsychiatr Genet</i>	SCZ	0	1	Korean	Korean	1,050
27922604	Yu H ¹⁸⁴	12/6/2016	<i>Mol Psychiatry</i>	SCZ	3	0	Han Chinese	Han Chinese	10,154
28991256	Li Z ¹¹⁰	10/9/2017	<i>Nat Genet</i>	SCZ	113	30	Han Chinese + European	Han Chinese	108,341

Table I. Major GWAS studies of schizophrenia and bipolar disorder from 2007 to 2017, continued

PMID	Author	Publication Date	Journal	Phenotypes	GW significant loci	# new loci	Discovery cohort ancestry	Replication cohort ancestry	Discovery sample size
17554300	WTCCC ¹⁸⁵	6/7/2007	<i>Nature</i>	BP	1	1	European		4,806
18317468	Sklar P ¹⁸⁶	3/4/2008	<i>Mol Psychiatry</i>	BP	2	2	European	European	3,469
18711365	Ferreira MA ¹⁰⁸	8/17/2008	<i>Nat Genet</i>	BP	2	1	European		10,596
19416921	Scott LJ ¹¹³	5/5/2009	<i>Proc Natl Acad Sci U S A</i>	BP	3	1	European		18,190
19488044	Smith EN ¹⁸⁷	6/2/2009	<i>Mol Psychiatry</i>	BP	4*	4*	European	European	3,049
20351715	Liu Y ¹⁸⁸	3/30/2010	<i>Mol Psychiatry</i>	BP, MDD	1	0	European		14,052
20386566	Lee MT ¹⁸⁹	4/13/2010	<i>Mol Psychiatry</i>	BP	2	2	Han Chinese	Han Chinese	2,000
21926972	Sklar P ⁷¹	9/18/2011	<i>Nat Genet</i>	BP	2	1	European	European	16,731
22182935	Chen DT ¹⁰⁰	12/20/2011	<i>Mol Psychiatry</i>	BP	6	3	European	European	14,755
24280982	Ruderfer DM ⁸¹	11/26/2013	<i>Mol Psychiatry</i>	BP	6	1	European		29,671
24618891	Mühleisen ¹¹⁴	3/11/2014	<i>Nat Commun</i>	BP	5	2	European		24,025
26806518	Hou L ¹⁹⁰	1/22/2016	<i>Lancet</i>	BP	1	1	European		2,563
27329760	Hou L ¹¹⁵	6/21/2016	<i>Hum Mol Genet</i>	BP	6	2	European	European	34,950
27890468	van Hulzen KJ ¹⁹¹	10/18/2016	<i>Biol Psychiatry</i>	BP, ADHD	3	3	European		31,139

Table I. Major GWAS studies of schizophrenia and bipolar disorder from 2007 to 2017, continued

PMID	Author	Publication Date	Journal	Phenotypes	GW significant loci	# new loci	Discovery cohort ancestry	Replication cohort ancestry	Discovery sample size
28115744	Ikeda M ¹⁹²	1/24/2017	<i>Mol Psychiatry</i>	BP	5	2	Japanese + European		81,582
	Stahl EA ¹¹⁶	8/7/2017	<i>BioRxiv</i>	BP	30	18	European	European	51,710

*ISGC & WTCCC2: Irish Schizophrenia Genomics Consortium & the Wellcome Trust Case Control Consortium 2

Table II. Reported CNV association to schizophrenia and bipolar disorder

Locus	CNV type	Gene or region	Direction	Disorder	Refs	Reported p-value	Cases CNV carrier %	Control CNV carrier %	Reported OR	P value in GWAS*	Cases CNV carrier % in GWAS*	Control CNV carrier % in GWAS*	OR in GWAS*
22q11.21	Deletion	Multigenic	Risk	SCZ	¹²³	4.40E-40	0.29	0	Inf	6.2e-13	0.275	0	NA
16p11.2	Duplication	Proximal duplication	Risk	SCZ, BP	^{78, 63}	2.90E-24	0.35	0.03	11.52	2.6e-10	0.299	0.020	13.8
15q13.2-13.3	Deletion	Multigenic	Risk	SCZ	^{119,125}	5.60E-06	0.14	0.019	7.52	6.1e-6	0.142	0.015	10.55
3q29	Deletion	Multigenic	Risk	SCZ, BP	^{193, 63}	1.50E-09	0.082	0.0014	57.65	6.2e-5	0.076	0	NA
2p16.3	Deletion	NRXN1	Risk	SCZ	^{127,194}	1.30E-11	0.18	0.02	9.01	9.4e-5	0.109	0.020	5.87
16p11.2	Deletion	Distal deletion	Risk	SCZ	¹⁹⁵	2.90E-24	0.35	0.03	11.52	1.0e-4	0.052	0.005	12.68
22q11.21	Duplication	Multigenic	Protective	SCZ	¹⁹⁶	8.60E-04	0.014	0.085	0.17	1.6e-4	0.019	0.114	0.18
1q21.1	Deletion	Multigenic	Risk	SCZ	^{119,125}	4.10E-13	0.17	0.021	8.35	2.9e-4	0.156	0.030	5.42
16p13.2	Duplication	C16orf72/USP7	Risk	SCZ	¹⁹⁷	1.00E-04	0.254	0.0197	12.9	3.8e-4	0.114	0.015	9.02
7q11.23	Duplication	Williams-Beuren	Risk	SCZ	¹⁹⁸	6.90E-05	0.066	0.0058	11.35	4.9e-4	0.062	0	NA
15q11.2-13.1	Duplication	AS/PWS	Risk	SCZ	¹⁹⁹	5.60E-06	0.083	0.0063	13.2	6.6e-4	0.071	0	NA
8q11.23	Duplication	FAM150/RB1CC1	Risk	SCZ	²⁰⁰	1.29E-05	0.106	0.014	8.58	9.2e-4	0.066	0	NA
15q11.2	Deletion	Multigenic	Risk	SCZ	¹¹⁹	2.50E-10	0.59	0.28	2.15	1.7e-3	0.450	0.232	1.8
1q21.1	Duplication	Multigenic	Risk	SCZ, BP	^{197, 63}	4.10E-13	0.17	0.021	8.35	2.0e-3	0.090	0.010	6.28
16q22.1	Duplication	WWP2	Risk	SCZ	¹²⁸	NA	NA	NA	NA	3.2e-3	0.024	0	NA
7q36.3	Duplication	WDR60/VIPR2	Risk	SCZ	^{197,201}	0.27	0.11	0.069	1.54	4.1e-3	0.062	0.005	12.12
17q12	Duplication	RCAD duplication	Risk	SCZ	¹²⁸	0.0072	0.036	0.0054	6.64	0.009	0.076	0.020	3.81
9q33.1	Deletion	NA	Risk	SCZ	¹²⁸	NA	NA	NA	NA	0.02	0.043	0.010	4.02
22q11.23	Duplication	Multigenic	Risk	SCZ	¹²⁸	NA	NA	NA	NA	0.02	0.071	0.025	3.28
5q21.2	Deletion	NA	Risk	SCZ	¹²⁸	NA	NA	NA	NA	0.03	0.104	0.049	2.16
8p22	Duplication	SGCZ	Risk	SCZ	¹²⁸	NA	NA	NA	NA	0.03	0.024	0	NA

Table II. Reported CNV association to schizophrenia and bipolar disorder, continued

Locus	CNV type	Gene or region	Direction	Disorder	Refs	Reported p-value	Cases CNV carrier %	Control CNV carrier %	Reported OR	P value in GWAS*	Cases CNV carrier % in GWAS*	Control CNV carrier % in GWAS*	OR in GWAS*
9p24.2	Deletion	SLC1A1	Risk	SCZ	²⁰²	8.40E-03	0.033	0	Inf	0.03	0.038	0	NA
16p12.1	Deletion	Multigenic	Risk	SCZ	²⁰²	1.60E-03	0.15	0.057	2.72	0.03	0.123	0.035	3.22
15q21.3	Duplication	CGNL1	Risk	SCZ	²⁰²	1.90E-03	0.32	0.19	1.71	0.04	0.327	0.168	1.99
17q12	Deletion	RCAD deletion	Risk	SCZ	²⁰³	0.0072	0.036	0.0054	6.64	0.04	0.019	0	NA
16p13.11	Del/Dup	Multigenic	Risk	SCZ	¹⁹⁹	5.70E-05	0.31	0.13	2.3	0.08	0.398	0.272	1.49
7q11.21	Duplication	NA	Protective	SCZ	¹²⁸	NA	NA	NA	NA	0.09	0.123	0.188	0.76
12q23.1	Duplication	ANKS1B/UHRF1BP1 L	Risk	SCZ	¹²⁸	NA	NA	NA	NA	0.1	0.076	0.059	1.23
1p36.33	Duplication	Multigenic	Risk	SCZ	²⁰²	5.00E-04	0.065	0.0075	8.66	0.11	0.057	0.015	3.98
5q33.1	Deletion	NA	Risk	SCZ	²⁰²	NA	NA	NA	NA	0.11	0.043	0.010	4.19
9q21.33	Duplication	AGTPBP1	Risk	SCZ	¹⁹⁷	NA	NA	NA	NA	0.2	0.071	0.035	1.94
9q34.3	Duplication	C9orf62	Risk	SCZ	⁷⁹	1.40E-03	1.47	0.43	3.38	0.23	0.901	1.083	0.8
6q24.2	Duplication	PHACTR2	Risk	SCZ	²⁰²	NA	NA	NA	NA	0.26	0.038	0.010	4.03
3q26.1	Deletion	NA	Risk	SCZ	¹⁹⁷	NA	NA	NA	NA	0.27	0.019	0.005	3.5
4q35.2	Deletion	TRIML1/TRIML2	Risk	SCZ	²⁰²	NA	NA	NA	NA	0.35	0.081	0.044	1.82
18q21.31	Duplication	NEDD4L	Risk	SCZ	¹⁹⁷	NA	NA	NA	NA	0.39	0.009	0	NA
11q25	Deletion	GLB1L3/GLB1L2	Risk	SCZ	¹⁹⁷	3.00E-03	0.38	0.123	3	0.42	0.161	0.119	1.44
9p24.2	Deletion	GLIS3	Risk	SCZ	²⁰²	8.40E-03	0.033	0	Inf	0.43	0.024	0.025	0.99
18q23	Duplication	GALR1	Risk	SCZ	²⁰²	NA	NA	NA	NA	0.57	0.019	0.015	1.22
4q35.2	Duplication	FAM149A/CYP4V2	Protective	SCZ	²⁰²	NA	NA	NA	NA	0.69	0.024	0.035	0.71
2q37.2	Duplication	AQP12A/KIF1A	Risk	SCZ	²⁰²	NA	NA	NA	NA	0.72	0.341	0.262	1.34
17p12	Deletion	HNPP	Risk	SCZ	²⁰⁴	1.20E-03	0.094	0.026	3.62	0.82	0.057	0.049	1.06
4q25	Duplication	ELOVL6	Risk	SCZ	²⁰²	NA	NA	NA	NA	0.9	0.033	0.030	1

Table II. Reported CNV association to schizophrenia and bipolar disorder, continued

Locus	CNV type	Gene or region	Direction	Disorder	Refs	Reported p-value	Cases CNV carrier %	Control CNV carrier %	Reported OR	P value in GWAS*	Cases CNV carrier % in GWAS*	Control CNV carrier % in GWAS*	OR in GWAS*
10q11.21	Duplication	Likely common CNV	NA	SCZ	¹²⁸	NA	NA	NA	NA	NA	NA	NA	NA

This table is adapted from Sullivan *et al.* 2012¹⁰² and Marshall *et al.* 2017¹²⁸. *GWAS is the results on the breakpoint-level CNV association from Marshall *et al.* 2017.

Chapter 2: the role of consanguinity in psychotic disorders

2.1 Preface

Pakistani families and populations are included in recent population genetics and medical genetics studies, due to the combination of a long history of genetic admixture and the higher rate of consanguineous marriages. Epidemiological studies dedicated to find the association of consanguinity with genetic traits and diseases, with the advent of high-throughput genotyping and sequencing technologies in the last decade, geneticists could directly calculate inbreeding coefficient and identify genomic regions that are identical because of ancient or recent common ancestors. However, few studies have examined the relationship between consanguinity and psychiatric disorders in a highly inbred population.

We combined genome-wide SNPchip genotyping and whole-exome sequencing to obtain the genetic profile of large multiplex consanguineous pedigrees. Besides characterizing the admixture and inbreeding population history, we applied different algorithms on both datasets to calculate the inbreeding coefficient and examined the contribution of consanguinity to the psychiatric phenotypes—schizophrenia and bipolar disorder—by the comparison between affected and unaffected family members, and further with a large dataset of population controls. We also tried to correlate the severity of subphenotypes with the inbreeding. The results showed there's no direct association between consanguinity and psychiatric phenotypes.

The recessive mode of inheritance has been successfully implicated in many Mendelian diseases, recent population-based studies suggested an excess of homozygous segments in cases affected with schizophrenia, and the accumulation of rare homozygous variants are expected in a highly inbred population. We analyzed the genetic data of the available family members at different resolutions, in order to test the recessive model in the complex trait. Overall, this study confirmed population-specific and family-specific genetic background, and the phenotype profiling of the affected individuals in the pedigrees. Incapable of identifying causal

homozygous segments or variants, it rejected the hypothesis of recessive model of psychiatric disorders in these families. More complicated genetic inheritance, such as oligogenic or polygenic, is proposed to perform on comprehensive genetic data.

Rational: Homozygosity mapping is a classical method used to map the causal region of Mendelian traits in consanguineous pedigrees. Since the high-density SNP chip data is available, runs of homozygosity (ROH) analysis is used to characterize genomic pattern of world populations, both the number and the length could reveal the history of ancient relatedness and recent inbreeding events. There are also studies reporting association of excess of ROH in psychiatric diseases like schizophrenia.

Hypothesis: The high prevalence of schizophrenia and bipolar disorder in these Pakistani families suggest a recessive mode of inheritance. The possible causal loci could be located through homozygosity mapping even in large consanguineous pedigrees. The inbreeding, the ROH profile or specific homozygous variants which segregate in the pedigrees could contribute to the disease phenotypes.

Methods: SNP chip array and whole-exome sequencing (WES) were generated for the affected and unaffected family members. Different statistical software was applied to calculate the inbreeding coefficient and detect ROHs, and some of them, such as PLINK, can incorporate data from different platforms.

Specific objectives: (1) reconstruct the pedigree trees with genealogical information; (2) confirm the admixture and inbreeding background with genetic data; (3) characterize the ROHs with reference population controls and other world populations; (4) examine the association between inbreeding and psychiatric phenotype; (5) examine the association between ROHs and psychiatric phenotypes; (6) correlate the inbreeding and ROH profile with phenotypic symptoms; (7) use WES data to detect rare deleterious homozygous variants that are shared by the affected family members and segregating in the pedigrees; (8) additional analysis to detect candidate loci, such as linkage analysis and homozygosity mapping.

2.2 SNP microarray and whole-exome sequencing of large consanguineous Pakistani families does not support high-penetrance deleterious homozygous variants as a direct cause for the psychiatric phenotypes

Manuscript: SNP microarray and whole-exome sequencing of large consanguineous Pakistani families does not support high-penetrance deleterious homozygous variants as a direct cause for the psychiatric phenotypes

Authors: Qin He¹, Boris Chaumette², Brohi Qasim³, Vagheesh Narasimhan⁴, Amelie M. Johnson¹, Alexandre Dionne-Laporte², Dan Spiegelman², Lynn E. Delisi⁵, Ridha Joobar⁶, Guy A. Rouleau², Lan Xiong^{1,2}

Affiliations : ¹Université de Montréal, Département de Médecine, Montréal (QC) H3T 1J4, Canada; Centre de Recherche, Institut Universitaire en Santé Mentale de Montréal, Montréal (QC) H1N 3M5, Canada.

²McGill University, Department of Neurology and Neurosurgery and Neurological Institute and Hospital, Montréal (Que), Montréal (QC) H3A 2B4, Canada.

³xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx

⁴xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx xxx

⁵VA Boston Healthcare System, Department of Psychiatry, Harvard Medical School, Brockton (MA) 02301, USA.

⁶McGill University, Department of Psychiatry, Montreal (Que), Canada; Douglas Mental Health University Institute, Montreal (Que), Canada.

Corresponding author: Lan Xiong

The authors declares no conflict of interest.

Word counts: TBC

Number of figures: TBC

Number of tables: TBC

Abstract

Schizophrenia and bipolar disorder are complex mental disorders with significant heritability. Epidemiological studies have suggested that consanguinity could contribute to the development of these two disorders; for instance, an excess of autozygosity was reported to be a risk factor in both cases. We aimed to further explore the contribution of consanguinity in psychotic/affective disorders. In ten large consanguineous Pakistani pedigrees with multiple affected individuals with schizophrenia, schizoaffective disorder, and bipolar disorder, we used genome-wide SNP genotyping ($n = 275$) and whole-exome sequencing ($n = 230$) to systematically investigate the correlation of consanguinity with the psychiatric phenotypes. We first estimated the relationship between the levels of inbreeding and the clinical symptoms. Neither the clinical categories nor the clinical dimensions were associated with the inbreeding coefficient. Secondly, we conducted genome-wide runs of homozygosity (ROHs) analyses, conventional linkage analyses and homozygosity mapping with the SNP genotyping data in each pedigree. Affected family members did not have larger or more numerous ROHs compared to their unaffected relatives. We did not identify homozygous regions shared by all the affected members from the same family, nor did we observe homozygous segments that were significantly more present in affected individuals than in unaffected family members. We did not identify any genome-wide significant loci in linkage analyses under recessive models. Thirdly, using whole-exome sequencing, we did not detect any rare loss-of-function or potentially damaging missense homozygous variants segregating in all the affected individuals of any family. Overall our results do not support either the levels of homozygosity or deleterious homozygous coding variants as the direct cause for the major mental disorders observed in our highly inbred Pakistani pedigrees.

Keywords: inbreeding, runs of homozygosity, linkage analysis, homozygosity mapping, whole-exome sequencing, bipolar disorder, schizophrenia

Introduction

For cultural reasons, consanguineous marriages are frequent in North and Sub-Saharan Africa, the Middle East, and West, Central, and South Asia¹. In particular, Pakistan has one of the highest rate of consanguineous marriages in the world. Consanguinity has been associated with increased risk for infant mortality²⁵, congenital malformations²⁰⁵, neurological diseases²⁰⁶, and intellectual disability²⁰⁷, mainly through the involvement of recessive mutations in the etiology of these disorders²⁰⁸.

Consanguinity has also been reported to be a risk factor for psychiatric conditions³⁴. For instance, consanguinity was reported to be associated with increased risk for type I bipolar disorder in Egypt³⁹. However, previous studies to explore the relationship between inbreeding and schizophrenia across various isolated populations, in which consanguineous marriages are frequent, have shown both positive and negative results depending on the studied population and the sample size³⁷⁻³⁹. These studies were primarily based on the epidemiological data. Fewer genetic studies have investigated the genomic architecture of affected individuals and that of control individuals that could support the link between consanguinity and psychiatric disorders. Nevertheless, some susceptibility loci have been reported historically using genome-wide linkage analysis in schizophrenia and bipolar disorder in extended pedigrees and population isolates⁴⁶⁻⁴⁸. Modern high-resolution genome-wide genetic data can be used to precisely estimate the degree of consanguinity in an individual (inbreeding coefficient) and to reliably identify chromosomal regions homozygous resulting from the consanguinity (runs of homozygosity). The inbreeding coefficient (F) could also be measured by the proportion of loci at which the offspring of a consanguineous union is expected to inherit identical gene copies from both parents. An individual for whom $F \geq 0.0156$ is deemed to be the consequence of a consanguineous union.

Homozygosity mapping is an efficient approach to map causative genes of rare recessive disorders in inbred populations^{23,209}, but is much less applied in genetic studies of complex traits. A number of groups have used the homozygosity mapping method to investigate schizophrenia, both in consanguineous families and in unrelated population samples. Runs of homozygosity (ROHs) were shown to be significantly more common in cases with schizophrenia spectrum disorders than controls⁵¹. Earlier homozygosity analyses on Pakistani and Japanese's offspring

of first-cousin marriages reported either candidate loci or shared ROHs among the schizophrenic probands, without identifying causative variants^{49,50}. More recently, large-scale studies focused on whole-genome homozygosity association on Caucasian schizophrenia cases and unrelated controls. They showed evidence that some ROHs were significantly more common in schizophrenia cases than in controls⁵¹. There was also an increase in the odds of developing schizophrenia with the increase of genome-wide autozygosity⁵². As a result, it has been suggested that recessive mutations of relatively high penetrance effects might explain a significant proportion of the genetic liability for schizophrenia. However, contradicting results were also reported in population-based case-control studies, which failed to replicate the significant association between ROH burden and schizophrenia/bipolar disorder, when doubling the previous sample size⁵³ or testing the same hypothesis in additional independent population cohorts^{54,55}. Therefore, the results were mixed with positive and negative associations, and no definitive conclusion has thus far been drawn. Moreover, none of these studies were performed on large multiplex family cohort, particularly families from isolated population with a high degree of consanguinity.

Over the last decade, rapid advances in genotyping and sequencing technologies have increased the power of genetic discovery and enabled the identification of many common and rare variants as risk factors for psychiatric disorders²¹⁰. Therefore, we explored the role of consanguinity in major psychotic disorders (bipolar disorder and schizophrenia) in ten large multiplex consanguineous pedigrees from Pakistan using combined high-density of SNP chip genotyping and whole-exome sequencing (WES). First, we characterized the families and confirmed their high level of consanguinity, as well as the familial aggregation of multiple affected individuals with psychotic and affective disorders. We then computed the inbreeding coefficient for each individual using SNP genotyping data and investigated its correlation with the clinical presentation. We further investigated whether the lengths or the numbers of the ROHs were associated with the clinical status. We also performed conventional genome-wide linkage analyses, particularly under recessive models. Lastly, we systematically explored the deleterious homozygous coding variants within the ROHs shared more by affected than by unaffected relatives. We present a series of our results here.

Materials and Methods

Samples and phenotyping

Ten large consanguineous pedigrees (further referred to as *MNS* pedigrees) were recruited in the Sindh Province of Pakistan. Each individual was interviewed by a local psychiatrist (B.Q.) using a standardized evaluation that included DIGS, FIGS (v3.0) and detailed medical history. A consensual clinical diagnosis was subsequently reached with two additional expert psychiatrists (R.J., L.D.) who reviewed the DIGS and FIGS separately, and their diagnoses were made blindly to each other. Final diagnosis was made collectively through further discussion among the group in case of discrepancy. The final diagnosis of each individual was based on the DSM-IV criteria. All phenotypic information was digitalized and saved in a secured database. The scale for the assessment of positive and negative symptoms of DIGS were selected to present the severity of phenotype dimensions for patients with schizophrenia.

The pedigree trees were constructed using the Progeny 9 software (Progeny Genetics LLC). Blood was drawn in Pakistan and delivered within 3-4 days to Dr. Rouleau's laboratory (McGill University). DNA was extracted according to a standard salting-out protocol. The full description of the recruitment process and of the clinical assessment is detailed in the **Supplementary text**. Ethical approval of the research project was obtained prior to the study in all involved institutions. All participants have given their written informed consent. All research procedures were carried out according to the Declaration of Helsinki.

Genome-wide SNP genotyping

SNP genotyping data

Internal Pakistani SNP genotyping was performed at the Genome Quebec Innovation Centre (Montréal, Québec, Canada) in two batches using Illumina HumanOmniExpress BeadChip v12 and v24 for 275 samples, including 124 affected and 151 unaffected individuals. The raw data was processed for genotype calling following the recommended GenCall threshold of 0.15 with Illumina GenomeStudio 2.0 software and its PLINK plugin (Illumina, Inc.). 624,015 SNPs were finally retained for further analyses after keeping the overlapping SNPs from both arrays and excluding SNPs: (1) not located on the autosomal chromosomes, (2)

deviating from the Hardy-Weinberg Equilibrium with a threshold p-value $< 1 \times 10^{-4}$, (3) with a call rate less than 90%, (4) with a minor allele frequency < 0.01 . Sex check with PLINK²¹¹ and kinship estimation with KING²¹² were used to check pedigree errors; no sex or Mendelian inheritance error was identified. **Table III** shows a summary of the available genotyping data after quality control. The SNP genotyping data were used to perform the calculation of inbreeding coefficient, analysis of runs of homozygosity, homozygosity mapping and linkage analysis.

Human Genome Diversity Panel (HGDP) SNP genotyping data:

Human Genome Diversity Panel (HGDP) data was downloaded from Stanford HGDP SNP genotyping data (<http://hagsc.org/hgdp/files.html>), corresponding to the genotyping data of 660,918 SNPs in 1,043 individuals from 51 different world populations.²¹³

Whole-exome sequencing (WES)

Internal WES data: We performed WES for all the affected individuals and their unaffected first-degree relatives from all ten pedigrees. A total of 230 samples were used, including 123 affected and 107 unaffected individuals (**Table III**). DNA was captured by the Agilent SureSelect Human All Exon v4 or v5 kit (Agilent Technologies, Inc.). WES was performed using the Illumina HiSeq 2000 platform (paired-end, 100 cycles) at the Genome Quebec Innovation Centre (Montréal, Québec, Canada) and at the Macrogen Korean facility (Macrogen Inc.) in separate batches. The raw fastq files were aligned to the human reference genome (hg19) with Burrows-Wheeler Aligner (BWA)²¹⁴. Duplicates were removed with the MarkDuplicates function in Picard tools. Genome Analysis Toolkit (GATK v3.5)²¹⁵ was used to process the bam files and to call the variants with the HaplotypeCaller algorithm. Quality control by sample included sex check, estimation of the potential DNA contamination level, and depth of coverage. The latter was performed using the GATK DepthOfCoverage tool and VerifyBamID²¹⁶. Eight samples were removed from the subsequent analysis due to possible contamination (VerifyBamID FREEMIX > 0.02). The remaining 222 samples (**Table III**) had an average coverage above 20X in $93.33 \pm 4.43\%$ of the targeted regions. Finally, the VCF files were merged for all the samples and all genotypes were called with the GenotypeGVCF tool implemented in GATK²¹⁵. All the variants were further recalibrated using GATK Variant

Quality Score Recalibration (VQSR) tool and annotated with Variant Effect Predictor (VEP version 88) based on GENCODE basic set version 19²¹⁷. Segregation of the variants in each pedigree was performed by using an in-house script. Final variant segregation files for 10 pedigrees included all the variants with a “PASS” filter of VQSR, corresponding to a minimum read depth (DP) ≥ 10 and a minimum genotype quality (GQ) ≥ 20 .

External WES data: Two external WES data sets were obtained and processed using the same bioinformatic procedures as the internal WES data.

(1) EBI3222 WES data. The UK is home to the largest Pakistani population in Europe. Publicly available WES raw data of the comparable Pakistani samples with related parents from the Born in Bradford Study in UK³² were obtained from the European Genome-phenome Archive (EGAD00001001025, EGAD00001001026, EGAD00001001027, EGAD00001001079, EGAD00001001686) and the processed VCF file with individual genotypes was provided by Dr. Vagheesh Narasimhan from the Wellcome Sanger Institute with the institutional agreement. These EBI Pakistani samples were captured with the Agilent SureSelect V5 and sequenced by 75bp paired-end on HiSeq 2000 with an $\sim 40x$ read-depth. A total number of 3,222 samples (EBI3222) were extracted from the VCF file; and we applied the same above-mentioned filtering criteria as for our WES internal data in terms of VQSR, DP and GQ.

(2) Sequencing data of the 1000 Genomes Project¹⁴¹: CHB (Han Chinese in Beijing, China), CEU (Utah Residents (CEPH) with Northern and Western European Ancestry), YRI (Yoruba in Ibadan, Nigeria) and PJI (Punjabi from Lahore, Pakistan) were also included in our study for population stratification, and the phase 3 release of the VCF files used to extract a subset of individuals.

WES variant annotation and filtration

Loss-of-function variants were defined as stop gain, splice acceptor, splice donor, and frameshift indel. Missense variants were considered to be rare if they were observed at a low frequency in the South Asian ExAC database and EBI3222 dataset (MAF < 0.01), and further considered as deleterious when the CADD⁴⁴ phred score was ≥ 15 . The filtered variants were further examined by referring to the allele frequency in larger database, such as Genome

Aggregation Database (gnomAD). As we aimed to assess the recessive hypothesis, we focused on homozygous variants shared by all affected family members in each pedigree.

Alignment with external datasets

We merged all the overlapping bi-allelic SNPs from our WES dataset, the EBI3222 WES dataset¹⁴¹ and the 1000 Genomes Project phase 3 release set¹⁴¹. Multidimensional scaling (MDS) with PLINK (v1.07)²¹¹ was performed to determine if our study subjects were closely clustered with EBI3222 healthy control samples and Punjabi population from Pakistan in the 1000 Genomes Project.

Estimation of inbreeding level

We applied two methods to estimate the inbreeding coefficient (F) both in our dataset and in the external dataset of the EBI Pakistani samples (EBI3222). First, we used a method implemented in Fsuite²¹⁸ to estimate the inbreeding coefficient. This method infers the full probability distribution of the identity-by-descent (IBD) status of the two alleles of an individual at each marker, along the genome, through a hidden Markov model²¹⁹. This method requires the markers to be in minimal linkage disequilibrium, otherwise it would produce biased estimations of F . To avoid this bias, it is proposed to generate multiple random sparse genome maps (submaps with markers every 0.5 cM) and to take the median of the estimated F from different submaps. We tested different numbers of submaps (5, 100, 1000) to estimate F from our SNP chip genotyping data. All submaps generated similar and highly correlated values. Consequently, the median F was retained for our analysis. This method provided F for correlation with the phenotype. Correlation with dimensional phenotypes was done with Pearson's test in SPSS v24.

A second method was used to measure F , based on the proportion of the autosomal genome located in runs of homozygosity (ROH) divided by the total explored length²²⁰ ($F_{ROH} = \sum L_{ROH} / L_{total}$). This second method is more suitable for WES data and was used for comparing F from our WES data with F from the external dataset. After extraction of the high-quality SNPs (calling rate > 98% and MAF > 0.05), 73,380 WES SNPs remained. Applying the same thresholds, we also extracted 62,041 overlapping WES SNPs from the EBI3222 dataset. The ROHs were identified with PLINK (v1.07)²¹¹ as recommended for WES data²²¹. Parameters

were set by default, except the following: at least 20 SNPs were needed within a 1,000 kb window to call an ROH. This optimal number of SNPs within a 1,000 kb window was chosen to be close to the mean density of the data²²².

ROH analysis

The default parameters were applied to map the ROHs shared by all the affected family members in each pedigree, regardless of the homozygous regions sharing with the unaffected family members of each family. We also used the '--homozyg-group' function in the PLINK program to detect the ROHs and to obtain the pools of overlapping and potentially matching segments from the HomozygosityMapper. The number and the length of ROHs were estimated for each individual in each pedigree. The regions with the ROHs shared by all the affected family members in each pedigree, as well as those more prevalent in affected individuals than unaffected individuals in each pedigree (according to a Fisher's test) were screened for all homozygous missense variants. The ROHs were also called for the overlapped genotyping set for HGDP samples and our samples to confirm the ROH profile of our samples with world populations.

Homozygosity mapping

In order to identify the putative homozygous-by-descent regions corresponding to the ROHs defined by PLINK, the SNP chip genotyping data of each pedigree was converted to AB format with individuals in columns and markers in rows and was uploaded to the HomozygosityMapper²²³ server listed under chip "Illumina: Illumina (any array not listed here)". During the analysis, HomozygosityMapper reads the length of homozygous blocks in all affected samples for every marker and adds them to a homozygosity score for the respective marker. An optimal value of 500kb as the maximum block length was used for our genotyping chip, in order to avoid the inflation of homozygosity score. Genetic homogeneity within single families was required for our dataset, which meant to only detect regions in which all affected individuals are homozygous, and controls were not used to exclude homozygous stretches.

Genome-wide linkage analysis

Linkage analysis was performed on the SNP genotyping data with MLINK (two-point linkage) and SIMWALK (multiple-point linkage). The recessive inheritance model was used in

parametric analysis (two-point and multi-point) and one statistic designed for traits best modeled by recessive inheritance was chosen in non-parametric analyses. The regions with a LOD score of more than 3.0 (parametric analyses) and $NPL-\log [P \text{ value}] > 3.0$ were considered to show significant linkage with the phenotype in each pedigree.

Results

Aggregation of multiple affected individuals with psychotic and affective disorders in each pedigree and phenotype summary

Altogether, 284 individuals were included in the study, 127 affected family members and 157 unaffected family members (**Table III**). Each pedigree included at least ten affected individuals diagnosed with schizophrenia, schizoaffective disorder or bipolar disorder (**Supplementary Figure 1**). The genetic origin and the clinical summary of each family is provided in **Supplementary Table 1a and 1b** and **Supplementary Figure 2a and 2b**. The high prevalence of these phenotypes in these pedigrees implicated they are inherited. Six out of the ten pedigrees mainly aggregated with schizophrenia are presented with the severity of their positive and negative symptoms in **Supplementary Figure 7**. The gradient by severity clearly showed that each patient had a medium to severe phenotype. The age of onset across phenotypes is shown in **Supplementary Figure 3**. The schizophrenia spectrum disorder and bipolar disorder have an age of onset spanning from puberty to 40s, and major depression disorder start to show syndromes in puberty, or 30s. Several pervasive developmental disorder patients from one family had their onset of the disease around 10 years old.

Population characterization

We used the SNP chip genotyping data to both characterize our Pakistani samples and enable its comparison to external samples. First, we determined the population admixture of our samples. By combining the HGDP data with our SNP genotyping data, we inferred the population admixture of our families and compared it with their self-reported ethnic background. Our samples clustered well with EBI3222 samples and Punjabi population from Pakistan in 1000 Genomes Project. As expected, they were also closer to Caucasian samples than to East Asian samples (**Supplementary Figure 4**).

Degree of consanguinity and its correlation with the phenotype

The length of ROH is usually used to infer the population history and the number of ROHs is usually reflective of the degree of inbreeding, i.e., the longer the length of ROH is, the more recent the inbreeding happened. The more ROHs an individual carry, the more closely related are the parents. The recent inbreeding events of our studied pedigrees were confirmed by comparison with the other HGDP populations such as the South Asians, Native Americans, and Middle-East populations known to have high degree of consanguinity (**Supplementary Figure 4**). The level of inbreeding depicted through ROHs seemed higher in our families than in the general Central South Asian population as well as in the Middle-East populations (**Supplementary Figure 4 & 5**).

The mean F of each pedigree, estimated using the SNP genotyping data, across the MNS pedigrees is 0.0758, which is greater than the kinship of a child from a first-cousin marriage (corresponding to an estimated $F = 0.0625$). The level of inbreeding varied from one pedigree to another with pedigree MNS03 showing the highest F (0.170) and pedigree MNS09 showing the lowest F (0.025) (**Table IV**). The distribution of inbreeding coefficient didn't show significant difference between affected and unaffected family members in all samples and in each family (**Supplementary Figure 7**). Due to the smaller number of overlapping SNPs in the WES data, the estimated F is smaller in WES data than in genotyping data. However, the F estimate made using ROH from WES was highly correlated to the F estimate made using the FSuite and the SNPchip genotyping data ($r = 0.987$). Comparing the difference of F estimated from WES data in our pedigrees with the EBI population controls, we found that our pedigrees had a significantly higher F than the EBI Pakistani population controls (0.0667 ± 0.0301 vs 0.0575 ± 0.0232 ; $p\text{-value} = 7.53 \times 10^{-5}$ – **Figure 5**).

We tested if F was associated with both the categorical and quantitative phenotypes presented in the pedigrees. Overall, there was no significant difference between all the affected and all the unaffected individuals from all the pedigrees together ($p\text{-value} = 0.7834$ – **Figure 5**) nor was there any significant correlation between the F value and the positive and negative symptoms among all the affected individuals (**Supplementary Figure 6 & 7**). We further examined several scenarios concerning the association of ROHs with the disease affection status within each pedigree: 1) the total number of ROHs; 2) the total size of ROHs; 3) the average size of ROH; 4) the number of long ROHs ($> 4\text{Mb}$); 5) the number of very long ROHs ($> 8\text{Mb}$).

None of these tests showed any significant difference between the affected and the unaffected individuals in all pedigrees together and in each pedigree respectively (**Table V and Supplementary Table 2**).

Homozygosity mapping

The results of homozygosity mapping were mostly negative. Although the software assigned a higher score to some putative homozygous regions shared more frequently in affected than unaffected family members, none of those regions remained as a true homozygous regions (examples shown in **Supplementary Figure 8**).

Segregation of ROHs

Our ROH analyses identified a large amount of putative ROH regions individual-wise and pedigree-wise. However, no homozygous regions were shared only by all the affected individuals in each pedigree. Using a Fisher's exact test within each family, we identified 42 ROHs that were more present in the affected individuals than in the unaffected ones ($p\text{-value} \leq 0.05$), but none of these remained significant after correction for multiple tests (Bonferroni correction, threshold set at $p\text{-value} \leq 1.89 \times 10^{-5}$ for 2,648 ROH regions tested in total). Considering that long ROHs are due to recent inbreeding, ROHs larger than 1 Mb and containing over 250 SNPs were further examined. The large ROHs that were nominally associated with the disease phenotype were then aligned with the list of homozygous deleterious variants called from the corresponding WES data. **Figure 6** illustrates such a potential candidate region in the MNS03 family on chr4:24676608-25036142 (Fisher's exact $p = 0.0016$). Overall, this analysis failed to detect perfectly segregating ROHs or ROH that were more frequent in affected than in unaffected individuals.

Homozygous variants from WES datasets

In general, the WES genotype data were consistent with the results by the SNP chip genotype data with many homozygous variants in the ROH regions, but no segregating candidate homozygous variants were found in these ROH regions (**Figure 6**). A single variant with incomplete penetrance presented as homozygous variants in more affected family members than unaffected family members of pedigree MNS03 (rs74901868 in *LGI2* gene, missense variant p.Gly61Ser) and this variant was predicted to be deleterious (CADD_phred score: 22.3).

However, this variant is quite common in the South Asian population ($MAF_{\text{ExAC_SouthAsian}} = 0.179$) and in the European populations ($MAF_{\text{ExAC_European}} = 0.270$). Moreover, no rare homozygous variant of interest could be identified in these shared ROHs among the affected individuals in each pedigree.

Finally, we systematically looked for segregating homozygous variants, regardless of autozygosity mapping and ROHs, in each individual pedigree under the assumption of a recessive transmission. No such segregating rare damaging variant was identified in each pedigree. None of the functional variants (loss-of-function/LOF and missense variants) that appeared to be homozygous in all the affected family members remained as potential candidate variants after filtering for allele frequency ($MAF \leq 0.01$ in ExAC South Asian and EBI3222). These potentially damaging homozygous variants were not private to any of our multiplex pedigrees, as they were also found in unrelated population control individuals (**Table VI**), and they were subsequently excluded when referring to the MAF in gnomAD. Additionally, the results of linkage analysis on the recessive inheritance model hardly show significant linkage to the phenotype in each family. We followed the linkage signal on chromosome 2 in one family (MNS05) to search for candidate variants in the region, and we failed to identify any interesting causal homozygous variants. Finally, we also tested if some of these homozygous variants partially segregated, by considering a scenario in which one or two individuals per family would be a phenocopy. However, this approach did not reveal any homozygous variants that partially segregated.

Discussion

To our knowledge, this is the first study to use large multiplex consanguineous pedigrees to investigate the role of inbreeding and autozygosity in psychiatric phenotypes. We used dimensional phenotypes²²⁴ to assess the role of consanguinity on specific symptomatic dimensions. No association was found between the symptom intensity and the inbreeding coefficient. Major studies identified polygenic components underlying multiple symptom dimensions (clinical quantitative measurements) of schizophrenia and bipolar disorder, where they have applied a simplified and adjusted LDPS (the lifetime dimensions of psychosis scale) based on DIGS⁸². Although our trial to explore the association between consanguinity, genomic

profile and phenotyping was primitive, we started from solid phenotype diagnosis of large extended pedigrees.

Furthermore, the large multiplex pedigrees used in our study decrease the likelihood of identifying homozygous variants that would be segregating by chance. As we found no evidence of such association, we concluded that our results were not supporting a recessive mode of inheritance in these major psychiatric disorders. However, we cannot exclude that an unestablished recessive model plays a role in other consanguineous families, as it is well-accepted that psychosis is associated with heterogeneous pathophysiological mechanisms. Despite the technical challenges for aligning different datasets generated from different technologies or platforms (SNP genotyping and WES data), we have confirmed the inbreeding of our families and primitively compared them with larger set of inbred healthy population controls. We would expect to see higher statistical power if our study size gets bigger, and we would also expect to have new findings if we have whole-genome sequencing data available. For all these reasons, we believe that our approach to systematically explore how consanguinity may contribute to the phenotype is of interest. More investigation is thus warranted before the recessive hypothesis can be excluded.

The impact of consanguinity is more striking in autism and intellectual disability where two family-based studies comparing the affected probands with either unaffected parents or unaffected siblings have found positive results, hence the autosomal recessive model explains a large part of the instances of intellectual disability and syndromic autism in consanguineous families^{225,226}. Although several case-controls association studies of genome-wide autozygosity with quantitative and disease phenotypes of schizophrenia and bipolar disorders have been reported in the last decade, the reports looking at the major psychotic and affective phenotypes are often inconsistent²²⁷. Newly published cohort study on Northern Ireland population concluded that children of consanguineous parents are at an increased risk for common disorders (OR, 3.01; 95% CI, 1.24-7.31) and psychoses (OR, 2.13; 95% CI, 1.29-3.51), through assessing the receipt of psychotropic medication in 363,960 individuals (609 of them, around 0.2% were born to consanguineous parents)²²⁸. With the emergence of more genetic and epidemiological evidence, despite the role of consanguinity being a promising hypothesis for schizophrenia and

bipolar disorder, the lack of very consistent positive results suggests that one or several unrevealed mechanism(s) are responsible for the genetic risk.

The successful identification of protein-truncating variants and missense variants in non-consanguineous multiplex families with psychosis has shed light on the rare dominant variant hypothesis^{93,132}. Many publications have also reported the involvement of copy number variations (CNV) in psychotic disorders, a genetic factor that is not directly addressed by our analyses. Whole-genome sequencing might later reveal CNV or regulatory variants. Convergence of rare variants and common variants identified through genome-wide association studies was found upon genes that are implicated in predominant etiological hypotheses of schizophrenia¹⁰¹. More complex genetic models, like gene-environment interactions mediated by epigenetic changes have been proposed to explain the emergence of psychosis⁷¹. The missing heritability of the psychiatric phenotypes in our consanguineous families need to be thoroughly examined. From an overall health perspective, consanguinity is also a much wider and complex topic that involves major social, economic, and demographic influences. Consequently, some environmental adversities might also play a role in the emergence of psychosis in our consanguineous families. Consanguinity would be expected to exert a greater influence on the etiology of complex diseases if rare autosomal recessive alleles were causally implicated. Conversely if the disease alleles are common, then intra-familial marriage would have a proportionately smaller effect²⁵. However, relationships between consanguinity and complex diseases of adulthood are still under-investigated, and more studies are needed before definitive conclusions are drawn.

Acknowledgements:

We thank all the families' participation in this study; this project would not be possible without their cooperation. We are also deeply grateful to Mr. Mehtab Christian's great effort in organizing the sample collection and on-site clinical investigations, and his continuous support for the project. We appreciate EBI's openness of sharing the Pakistani WES data with us, as well as the HGDP genotype data and the 1000 Genomes Project data. The project is supported by a CIHR operating grant to L.X., R.J., and G.A. R. Q.H. is supported by a joint scholarship from the Université de Montréal and the Chinese Scholarship Council. Q.H. and B. C. would

also like to express their gratitude to Dr. Patrick Dion and Cynthia Bourassa, from Dr. Guy Rouleau's lab, for their help with the language editing of the manuscript and helpful discussions.

Author's contributions: Qin He performed data preparation and analyses; Boris Chaumette organized the results and the structure of the manuscript; V. N. provided the EBI dataset VCF file and helped with the alignment of our internal WES data with the external EBI WES data; Amelie M. Johnson was responsible of generating the internal SNP genotyping data; Alexandre Dionne-Laporte and Dan Spiegelman were responsible for the internal WES variant calling and annotation bioinformatic pipelines; Brohi Qasim, Lynn E. Delisi and Ridha Joobar were responsible for the phenotyping and clinical diagnoses; Lan Xiong and Guy Rouleau were responsible for the study design and overall operation of the project, as well as manuscript writing and editing. Lan Xiong was responsible for managing the project and supervising the genotyping, sequencing experiments and all data analyses, and finalizing the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

Figure legends

Figure 5: Froh_wes: Inbreeding coefficient estimated by the runs of homozygosity method using whole-exome sequencing data. aff: affected; unaff: unaffected; EBI3222: cohort of Pakistani control from the EBI consortium; NS.: not significant

Figure 6: A – Run of homozygosity (ROH) region plot for chromosome 4 for all the family members in MNS03. Each row shows one individual (affected in red and unaffected in blue, the shared region highlighted in green). B - Variants from sequencing in the shared region, dark blue, cyan and grey depict heterozygous, homozygous variant and homozygous reference allele respectively. The order of individuals is the same as the top. No predicted pathogenic homozygous variant is segregating in all affected individuals in this ROH. The red star marks the only missense variant that is overrepresented in a homozygous form in affected individuals.

Tables

Table III. Summary of the ten Pakistani families

Pedigree	Main phenotypes	Marriages	Number of samples	DATA		
		(1 st , 2 nd -cousin)	Total (Aff/Unaff)	Genotyping	WES (Aff)	WES (Unaff)
MNS01	SCZ & SAF	12 (4, 7)	39 (13/26)	36	12	9
MNS02	SCZ	7 (3,4)	28 (11/17)	26	4	10
MNS03	SCZ & SAF	25 (12, 13)	31 (13/18)	28	12	11
MNS04	SCZ	2 (2, 0)	23 (11/12)	23	11	8
MNS05	BP	13 (2, 11)	22 (13/9)	22	13	6
MNS06	SCZ	16 (10, 6)	31 (12/19)	31	12	16
MNS07	SCZ, SAF & BP	11 (4, 5)	34 (19/15)	34	17	13
MNS08	PDD, SCZ & BP	7 (1,6)	27 (10/17)	26	10	12
MNS09	SCZ & BP	6 (1, 5)	26 (14/12)	26	14	11
MNS10	BP	4 (1, 3)	23 (12/11)	23	12	9
		110 (45, 62)	284 (127/157)	275	117	105

SCZ: schizophrenia; SAF: schizoaffective disorder; BP: bipolar disorder; PDD: pervasive developmental disorder; Aff: affected; Unaff: unaffected. Marriages indicate the number of 1st and 2nd-cousin mating reported in the genealogical information.

Table IV. Estimation of the mean inbreeding coefficient and mating types of individuals' parents by pedigree

Pedigree	Total	Number of inbred individuals	Mean of <i>F</i>	1 st cousin	2 nd cousin	Double 1 st cousin	Outbred
MNS01	36	17	0.034	4	7	7	18
MNS02	26	25	0.151	0	0	25	1
MNS03	28	28	0.170	0	0	28	0
MNS04	23	16	0.030	3	13	2	5
MNS05	22	18	0.058	5	4	9	4
MNS06	31	31	0.094	4	5	22	0
MNS07	34	30	0.057	15	5	10	4
MNS08	26	23	0.036	9	12	3	2
MNS09	26	22	0.025	6	19	0	1
MNS10	23	23	0.092	13	1	9	0
Total	275	233	0.0758	59	66	115	35

F: inbreeding coefficient. Detailed inbreeding information of the inbred individuals was inferred as from 1st cousin, 2nd cousin and double 1st cousin mating.

Table V. Comparison of ROHs in affected and unaffected family members

Runs of homozygosity	Affected (n = 125)		Unaffected (n = 150)		Wilcoxon Test p-value
	Mean	SD	Mean	SD	
Total number of ROHs	25.51	19.21	25.81	18.44	0.755
Total size of ROHs(Mb)	208.47	186.17	193.60	166.50	0.780
Average size of ROHs(Mb)	6.86	3.12	6.44	2.66	0.244
Total number of ROHs >4Mb_size (Mb)	0.18	0.17	0.17	0.15	0.585
Total number of ROHs >8Mb_size (Mb)	0.15	0.15	0.13	0.13	0.605

SD: standard deviation; Mb: mega-base

Table VI. Summary statistics of homozygous variants shared by all the affected individuals in each family

	MNS01	MNS02	MNS03	MNS04	MNS05	MNS06	MNS07	MNS08	MNS09	MNS10
Number of affected	12	4	12	11	13	12	17	10	14	12
Number of unaffected	9	10	11	9	6	16	13	12	12	9
After filtering by segregation										
Number of SNPs homozygous in all affected individuals	9898	16827	10818	7474	7930	8090	7480	4886	8447	9447
Number of indels homozygous in all affected individuals	1091	1957	1296	1015	994	1028	967	497	1068	1113
Missense SNV	1683	2453	1660	1210	1370	1325	1229	1087	1375	1597
LOF SNV	18	30	20	12	13	12	13	12	11	14
LOF indel	97	120	110	86	95	99	94	58	88	97
After filtering by allele frequency novel or rare (MAF<=0.01) in SAS										
Missense SNV	0	2	0	0	1	1	0	0	1	2
LOF SNV	0	0	0	0	0	0	0	0	0	0
LOF indel	6	7	5	2	7	2	5	0	3	5
Number of family private SNV	0	0	0	0	0	0	0	0	0	0

LOF: loss of function; SNV: Single Nucleotide Variant; SAS: South Asian Population, in both ExAC South Asian population and EBI3222 Pakistani samples.

Figures

Figure 5. F_{ROH_WES} of MNS pedigrees and population controls

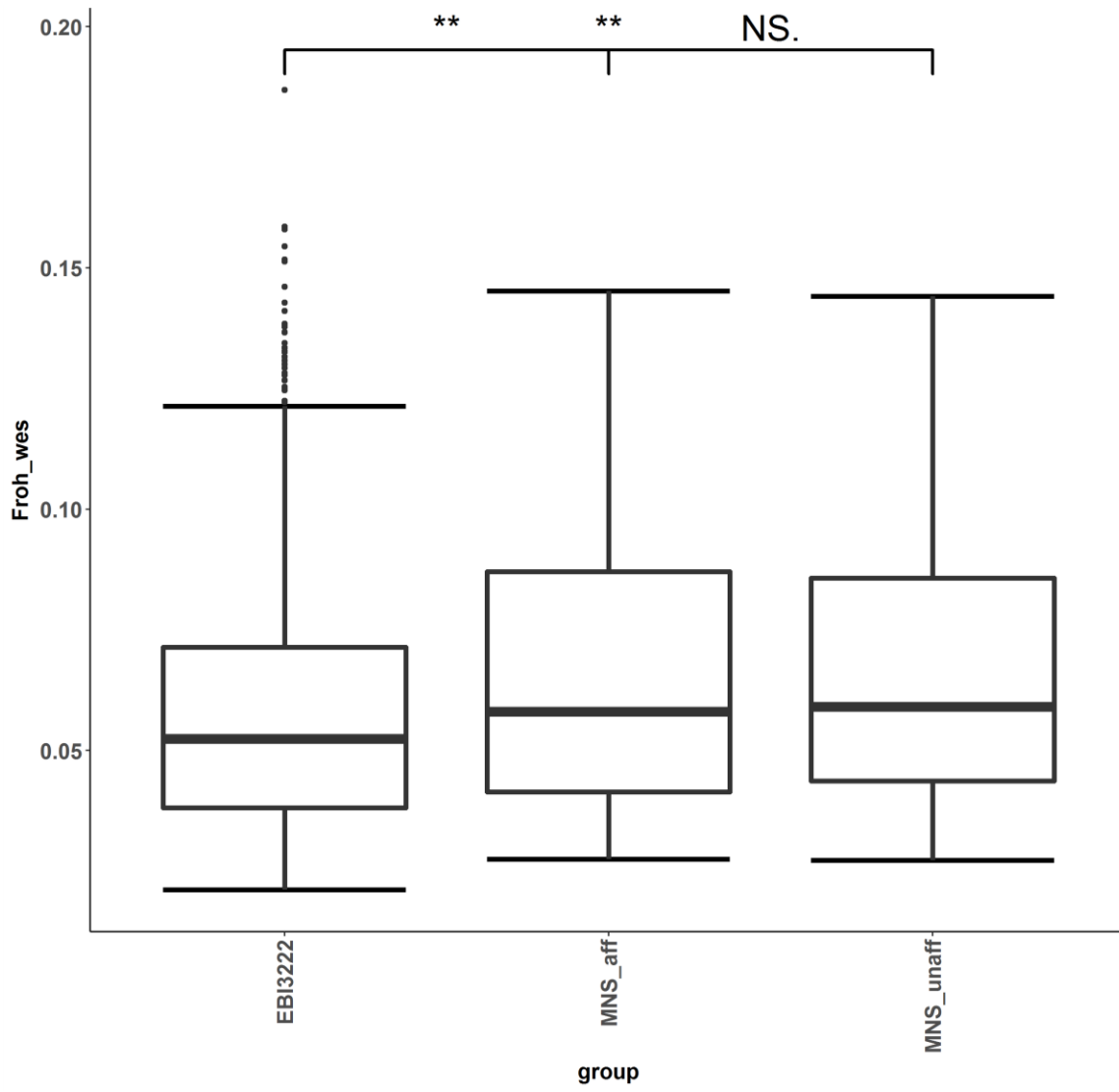
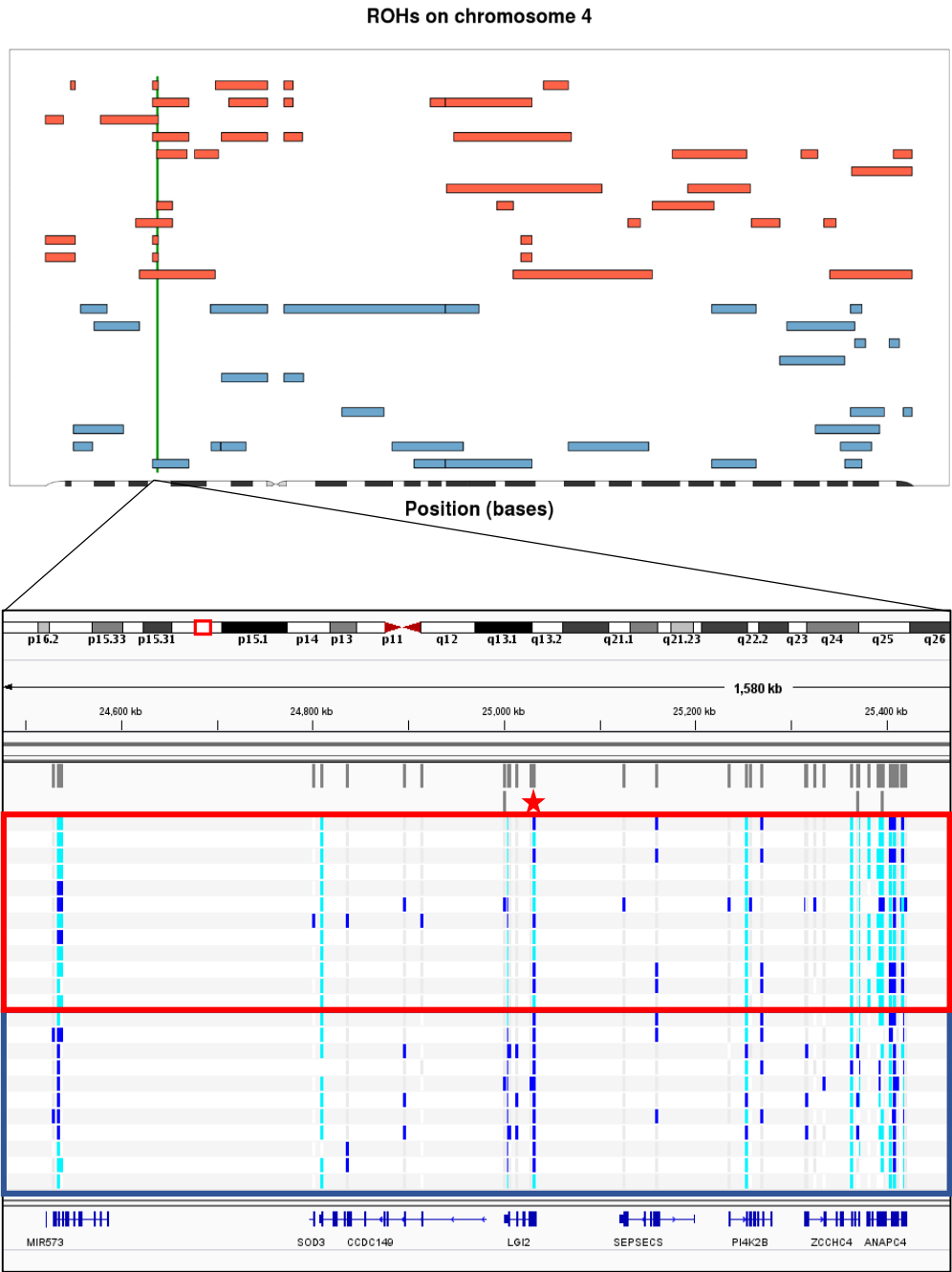


Figure 6. Example of a putative ROH segment shared by all affected members in MNS03



Chapter 3: The contribution of copy number variants in multiplex Pakistani families

3.1 Preface

This chapter is focused on detecting copy number variants (CNVs) from different platforms with different algorithms, since the direct experimental detection methods are still under development. The commonly used data nowadays are genotyping and sequencing, both WES and WGS. In this study, we used genotyping and WES data to call CNVs, usually one is complementary to the other or served as a cross validation.

Here we tried to use the segregating CNVs in these families as a reference to test the sensitivity and specificity of five different detection algorithms. We noticed that the common filters, such as the size of CNVs, number of probes in one CNV, the confidence score generated from the calling algorithms and the intersection calls between algorithms, reduced the number of segregating CNVs as well as the individualized CNVs, which could be technical artifacts or false positives. Those parameters are not good predictors for segregating CNVs. The overlapping rate between different algorithms is limited.

We proposed to combine the segregation analysis and further functional filters for possibly pathogenic CNVs. The CNVs appearing in multiple non-family members could be recurrent in a specific population, although it may not be shown in currently available database of genomic variants. In our results, segregating CNVs demonstrated an incomplete penetrance, where few CNVs are shared among affected family members, let alone CNVs shared only by affected but not in unaffected individuals. We combined the annotation of gene function and previous evidence of association to generate a shortlist of potential pathogenic CNVs. However, these CNVs require further validation.

Rational: CNVs have been largely studied and replicated in major psychiatric disorders, however the detection of CNVs with available technologies and algorithms are still under development. By using our family cohorts which mainly aggregated with psychiatric

phenotypes, we could evaluate the performance of current CNV detection methods from genetic data, especially CNVs segregating in the families.

Hypothesis: The performance of different CNV detection software on data from different platform is variable, and the overlap between them is low. Tuning the parameters of the software separately and together could contribute to a better sensitivity and specificity. The segregating CNVs in families could help filter the likely true positive CNVs.

Methods: Use common software (PennCNV, QuantiSNP, CNVPartition, CoNIFER and XHMM) to detect CNVs from SNP chip genotyping data and whole-exome sequencing data, and compare the parameters of the software and filter the likely positive CNVs that segregate in the families

Specific objectives: (1) to call CNV for the Pakistani families with PennCNV, QuantiSNP, and CNVPartition from SNP chip genotyping data; (2) to call CNVs for the Pakistani families with CoNIFER and XHMM from whole-exome sequencing data; (3) separate the segregating (shared) CNVs in the family members and define likely true positive CNVs; (4) evaluate the sensitivity and specificity of detecting likely true positive CNVs; (5) evaluate the contribution of the software parameters

3.2 Familial segregation analysis for copy number variations: new software and methodological recommendations

Familial segregation analysis for copy number variations: new software and methodological recommendations

Qin He^{1*}, Boris Chaumette^{2*}, Dan Spiegelman², Alexandre Dionne-Laporte², Guillaume Huguet³, Amélie Musa-Johnson¹, Gabrielle Houle^{2,3}, Sébastien Jacquemont⁴, Guy A. Rouleau², Lan Xiong¹

1. Centre de recherche, Institut universitaire en santé mentale de Montréal, Université de Montréal, Montréal, Canada
2. Montreal Neurological Institute and Hospital – McGill University, Montréal, Canada
3. Department of Human Genetics, McGill University, Montréal, Canada
4. CHU Sainte-Justine – Université de Montréal, Montréal, Canada

* These authors contributed equally to this work

Corresponding author: Dr Lan Xiong

Abstract

Background

The detection of Copy Number Variations (CNVs) from large datasets remains challenging due to variable sensitivity and specificity offered by different software and approaches. Moreover, the parameters used for filtering true CNVs are not consistent in literature. Using whole-genome genotyping data (SNP chip) and whole-exome sequencing data (WES) from a large dataset ($n = 243$ individuals from 15 different families), we called CNVs using five common software (PennCNV, QuantiSNP, CNVPartition, CoNIFER and XHMM) and tried to estimate their sensitivity and specificity, as well as the best filtering parameters. Then, we developed and tested a new software called “SV-Segregation” to determine the segregation of CNVs.

Results

Using the genotyping dataset, 14.9% of the CNVs were overlapping with all three algorithms (PennCNV, QuantiSNP, CNVPartition), with a comparable concordance rate for duplication and deletion. Using the WES dataset, XHMM called more CNVs than CoNIFER. The overlap between CNVs called from genotyping and WES data was estimated to 5%. None of the classical parameters (size of the CNV, number of probes, quality score, number of algorithms detecting the CNV) can accurately determine true and false CNVs. Using all the CNV calls without filtering improved our segregation analysis, for which our new software SV-Segregation was utilized.

Conclusions

Our results shed light on the importance of a complementary approach to efficiently detect CNVs. The familial design helped us to determine the sensitivity and the specificity of the

different methodologies, but optimizing the parameters of filtering is not straightforward. For familial segregation analysis, we recommend a thorough look at all segregating CNVs, independently of the quality measurement. Our new software – SV-Segregation – can be helpful for performing familial segregation analysis and is made freely available online.

Keywords

CNV - deletion - duplication - whole-exome Sequencing - microarray genotyping

Background

Genetic variants refer to variations in DNA from the scale of one base-pair to larger structural variations²²⁹. Among the latest, copy number variations (CNVs) are defined as deletion, insertion or duplication of stretches of nucleotides longer than 50 bp²³⁰. In the last decade, CNVs have been mapped to the genomes of control individuals in the general population^[3–7]. They have also been associated with evolutionary adaption^{235,236} and with a wide range of phenotypes, from rare diseases to complex traits including neuropsychiatry, obesity and cancer. Historically, the first primarily used methods to discover CNVs were microarrays with large-insert clones²³⁷ (bacterial artificial chromosomes, BACs) or oligonucleotide arrays²³⁸. Later, comparative genomic hybridization (CGH) was developed to detect fine-scale structural variations in multiple samples^{233,239–241}. The development of high-throughput methods with whole-genome genotyping (SNP array) and next-generation sequencing (whole-exome and whole-genome sequencing) has also opened the way to deeper assessment of CNVs.

For SNP array, the algorithms use two summary measures at each SNP: a measure of normalized total signal intensity, and a measure of normalized allelic intensity ratio. The CNV detection algorithms based on exome sequencing data use the normalized read depth matrix to infer CNV status. Several bioinformatics tools based on different statistical methods have been developed recently. Despite the increasing interest of CNVs in genetics, and the wide use of microarrays and sequencing as CNV-discovery techniques, downstream bioinformatics pipelines can be improved. An accurate CNV detection remains challenging, with a high number of false-positive calls, and an evaluation of the quality of the available tools is needed.

This is particularly relevant for family studies. Missing a CNV in one individual can affect the segregation in the whole family and the conclusion of the study. We aim to estimate the sensitivity and the specificity of well-known software, to determine the threshold of parameters to set for them and to develop reliable segregation analysis.

We employed a large dataset, including genotyping and sequencing data from multiplex families (several affected individuals from each generation) to measure the sensitivity and specificity of five algorithms; three for SNP array data: PennCNV, QuantiSNP, CNVPartition; two for WES data:XHMM and CoNIFER. We then identified segregating CNVs using a new software called SV-Segregation (for Structural Variants-Segregation). This family design, followed by segregation analysis, was used to determine the performance of these algorithms and the influence of filtering parameters used by other studies. Segregating CNVs were considered to be likely positive, and we applied this definition to adjust the filtering parameters of each software. Then, we tested the overlap between the five software based on different parameters.

Results

The approach used the combined results of the five algorithms followed by a segregation analysis. A likely positive CNV was defined as a CNV identified in ≥ 2 related individuals (“segregating CNV”). On the contrary, a likely false-positive CNV was defined as identified in only one individual of the family. We acknowledge that our classification identified the *de novo* CNVs as false positives. Whereas *de novo* CNVs are often clinically significant, their frequency is so rare (~ 0.0154 per generation for CNVs larger than 100kb)¹⁴³ that neglecting them should not impact the performance estimation of the tools. Recurrent CNVs were defined as CNVs

found in ≥ 10 unrelated individuals. They were excluded to avoid potential artifacts that would generate false signals.

We identified a total of 10,944 likely false positive CNVs and 23,414 segregating CNVs (likely to be true positives in family design) and estimated the sensitivity and specificity of each algorithm (**Table VII**). QuantiSNP generated much more calls compared to the other algorithms but the specificity of the calls is lower. A similar trend was observed between deletion and duplication (**Supplementary Table 3**). Specificity and sensitivity for the CNVs located on the X chromosome are reported in **Table VIII**. Since most families are large and with multiple affected family members, our CNV segregation analyses and filtering candidate CNVs are based on CNVs found in affected family members. There was not significant difference between defining a likely positive CNV as identified in ≥ 2 or in ≥ 3 related individuals of one family in terms of the sensitivity and the specificity (**Supplementary Table 4 and 5**, based on analysis of CNVs found in all family members).

We examined if the inbreeding in our studied families would increase homozygous duplications or deletions. A linear regression between the total size of the homozygous deletions/duplications and inbreeding coefficient was carried out, the resulting linear regression p-value is 0.233 for homozygous deletions and 0.374 for homozygous duplication respectively. The total size of CNVs was added from homozygous deletions and homozygous duplications of the autosomal regions of each individual and only limited to the likely positive CNV calls.

Prediction of likely segregating CNVs

Regarding the genotyping data, compared to non segregating CNVs, the segregating CNVs had significantly higher QC scores (as provided by the algorithms), larger sizes, and were identified

more frequently by two or three programs with a greater overlap in terms of size (**Table IX**). However, these features were poor predictors for segregating CNVs because some individualized CNVs could also show a very high-quality score (**Figure 7**).

Regarding the WES data, the segregating CNVs were identified more frequently by two programs, with a greater overlap compared to the likely false positive CNVs; whereas the size of likely true and false calls did not differ (**Table X**). However, these features were, again, poor predictors of true and false-positive CNVs (**Figure 7**).

We tried to filter the CNVs with some commonly used parameters. This strategy missed out a large portion of segregating CNVs. The usage of filtering parameters decreased the number of segregating CNVs (likely positive) and individualized ones (likely negative), consequently reducing the number of potential CNVs for each family. These parameters were widely used by other studies^{71,128}. For example, if we only keep CNVs larger than 100kb, we would eliminate 92% of individualized CNVs and 83% of the segregating CNVs; if we use a combination of parameters (CNV size \geq 20kb, number of probes \geq 5, confidence score \geq 5 and detected by \geq 2 algorithms), we could only obtain 21% of the segregating CNVs and 10% of individualized CNVs (**Table XI**).

Overlap between true-positive CNVs from different sources and software

Using the genotyping data, the three algorithms had a mutual overlap of 14.9%. Each combination of two algorithms achieved a similar overlap (between 15% and 30%) (**Figure 8**). For WES data, the two algorithms had an overlap of 26.3%, with XHMM calling more CNVs than CoNIFER (**Figure 8**). As WES data are inappropriate to detect non-exonic CNVs, we retained only exonic CNVs for comparison between WES and SNP array detection. A total of 10,632 segregating CNVs were tested for overlap. WES alone and SNP array alone identified respectively 1,731 and 8,395 of these CNVs. 506 CNVs were detected by both techniques (5%).

SNP array algorithms identified more CNVs than algorithms using WES data, even in the exonic regions. The overlap between both techniques is limited, suggesting that combining different approaches could be helpful to identify more potential CNVs.

Recommendation for familial segregation analysis and use of the SV-segregation software

Our software SV-segregation was developed in the Python3 programming language. It is based on establishing consensus CNV calls by merging CNVs from multiple samples and calling algorithms, using user-defined thresholds of reciprocal overlap between individual calls. We compared the segregation results of CNVs in our 15 families with and without filters. The advantage of our data is that we have small pedigrees (each includes 3-5 affected individuals) and large multiplex pedigrees (each includes 10-17 affected individuals, the detailed summary of samples is shown in **Supplementary Table 1** and **Supplementary Table 2**). We kept CNVs segregating in at least 5 affected individuals in large pedigrees and at least 2 affected individuals in small pedigrees for further analysis. As a result, we obtained a short list of 135 potential CNVs for 10 large pedigrees and 209 potential CNVs for 5 small pedigrees. The change of segregation pattern according to different filtering parameters in each pedigree is demonstrated in **Supplementary Table 6**. In these families, segregating CNVs have incomplete penetrance. Additionally, we found the stringent filters would rescue some CNVs close to the cut-off (for example, recurrent CNVs that are found across families).

Moreover, potential CNVs require cautious interpretation based on other annotations, such as the frequency in public databases (DGV and 1000 Genomes), the exclusion of segmental duplications, previous evidence of pathogenicity, and thorough examination of the segregation. As a side proof, we visualized one representative CNV with its BAF and logR from raw

genotyping data (**Figure 9**). A short list of possibly pathogenic CNVs is included respectively in **Supplementary Table 7** for large schizophrenia and bipolar disorder pedigrees and **Supplementary Table 8** for small autism pedigrees. The results are interesting but preliminary since the potential candidate CNVs need to be further validated. Consequently, we recommend performing segregation before any filtering.

Discussion

In this study, we applied the most commonly used algorithms for CNV detection on matched samples of a family design, and we did not identify any clear rationale to filter CNVs based on a specific threshold in terms of size, overlap between different programs, number of SNPs, or QC score provided by the algorithms. Adjusting the parameters created a specific choice, which increased the confidence in the detected CNV, but could be missing some relevant variants. True CNVs can be called by one tool only. For familial segregation analysis, we recommend a thorough look at all segregating CNVs, independently of the quality measurement. However, in a case/control design, we might suggest to use stringent filters to improve the confidence in the results.

In the clinical practice, detection of CNVs is done using CGH-array or FISH, which have a strong reliability and reproducibility, but their cost prevents their use in large-scale cohorts. In research, many genotyping or WES data are now available. Using these already-collected data to detect altered copy number of genes, and risk factors for human diseases, is promising. However, CNV calling is subjected to many artifacts and quality control is not standardized across the different cohorts. This lack of standardization decreases the reproducibility of CNV calls and compromises the comparability of the studies. Most of the time, researchers keep only

CNVs called by more than two algorithms, with a certain length or a specific quality score. Some studies argued that CNVs detected by SNP array are more reliable than if they are detected by WES²⁴². The group who invented CoNIFER used CNVs detected by SNPchip as their first validation of CNVs called by CoNIFER and XHMM, and they would further validate a small number of CNVs by arrayCGH or qPCR. They reported higher overlapping rates between SNP array and WES, even though they have more stringent SVD cut-off to call CNVs. In our dataset, the average calls per individual are comparable to their method^{243–245}. WES has been proposed to be better at detecting smaller exonic deletions, compared to SNP array²⁴⁶. Only a small amount of studies have actually investigated systematic comparison of calls from both methods²⁴⁷, while others have suggested a complimentary approach of WES and SNP array to detect intragenic CNVs [21]. Here, we demonstrated that the overlap between these techniques is limited, and our potential pathogenic CNVs preferably need to be validated through experimental procedures.

A statistical framework (iCNV) has recently been released. It combines SNP and sequencing data, by applying platform-specific normalization and utilizing allele-specific reads from integrating matched NGS and SNP array data by Hidden Markov Model²⁴⁸. This software should increase sensitivity and robustness, with the integration of two platforms for CNV detection, comparing to naive intersection or union of platforms. The integration of data from both sequencing and SNP array may result in a bias, due to different spatial coverage of exome-target regions and microarray probes.

With a family-based design to calculate the sensitivity and specificity of CNVs detected by different algorithms and different techniques, we were able to cross-validate our calls and increase the detection of segregating CNVs. We also showed that using some filtering

parameters could not effectively separate true-positive from false-positive calls. Besides, the low percentage of overlap between the different algorithms did not reflect the different coverage of the targeted regions. Using different datasets and different software increased the number of detected true-positive CNVs. The family design is a very powerful tool to filter CNVs. Furthermore, ascertaining the pattern of segregation in families is helpful to determine the pathogenicity of a CNV.

Methods

Population

The samples are members of 15 consanguineous Pakistani families. The description of these multiplex pedigrees is mentioned somewhere else [He Q *et al*, in preparation]. Half of these individuals are affected by various neuropsychiatric diseases (autism spectrum disorder, schizophrenia or bipolar disorder). Whole-genome genotyping was performed on 334 individuals using the Infinium OmniExpress chip (Illumina).

Whole-exome sequences (WES) were available for 241 individuals overlapping with genotyping data. DNA was captured by Agilent SureSelect 50M, Agilent SureSelect V4 and Agilent SureSelect V5. WES was performed using Illumina HiSeq 2000 platform (paired-end, 101 cycles). The raw WES reads were subjected to an in-house pipeline through alignment, quality control and collection of coverage metrics (Burrows-Wheeler Aligner (BWA)²¹⁴ and Genome Analysis Toolkit (GATK v3.5))²¹⁵.

Pipelines for CNV calling

The pipeline for calling CNVs and adjusting parameters for identifying segregating CNVs is depicted in **Supplementary figure 1**. For genotyping data, the final reports were extracted from

GenomeStudio after classical quality control and three CNV calling algorithms were used: QuantiSNP²⁴⁹, PennCNV²⁵⁰, and CNVPartition. QuantiSNP v2.2 was used with MATLAB Compiler Runtime v7.9 and default parameters. For PennCNV, we first generated a population B allele frequency (PFB) file using the whole genotyping dataset. Then the *detect_cnv.pl* script was run using default parameters and the default lib/hh550.hmm model. CNVPartition was run directly from GenomeStudio with default parameters. Detailed information is provided in supplementary text.

CNV calls from WES data were made using two software: XHMM and CoNIFER. Capture kit-specific BED files were used to select the regions to be analyzed. High-complexity and GC-rich regions were excluded from the analysis. We followed the recommended workflow from the tutorial in XHMM using GATK generated DepthOfCoverage files (GATK v3.5)^{215, 251}. CoNIFER²⁴³ calculates RPKM (reads per thousand bases per million reads) for each exome capture targets for each sample from aligned bam files, and the RPKM values were transformed into standardized z-scores based on the mean and standard deviation across all analyzed exomes and organized into an exon-by-sample matrix; the first 7 components were eliminated based on the inflection point of the scree plots (**supplementary text** and **supplementary figure 2**).

Merging and annotation

Before merging, we kept CNV calls larger than 1kb on autosomal chromosomes and X chromosomes (CN=2 is neutral for female and CN=1 is neutral for male). CNV outputs were combined for analysis using the merge function of the CNVision v1.73 software²⁵². The script was slightly modified to return the higher value in terms of length, number of SNPs and confidence score for each merged CNV.

Annotation about mapping the CNV chromosome position to the genes it affects and whether it covers exonic or intronic location was performed by ANNOVAR²⁵³ (region-based annotation) and was based on the GRCh37/hg19 database.

Software for segregation analysis

Segregation analyses were done using an in-house Python script called SV-Segregation (Structural variant-Segregation). Firstly, raw calls from each same sample are filtered by user-defined thresholds and de-fragmented into non-overlapping calls by type (deletion, duplication, inversion and translocation). Filtered calls are then merged by sample between callsets to generate a single unified callset for each sample. Finally, unified calls are overlapped between samples to generate a final consensus set of CNV calls, which are reported once per family, counting the occurrence of each call in all affected and non-affected family and non-family samples (as defined by a standard pedigree input file). In addition, variants are annotated with overlapping CNVs from external datasets (1000 Genomes Project, Database of Genomic Variants), as well as various UCSC tracks (RefSeq genes and exons, micro-exons, repeat regions and segmental duplications) and any other user-defined additional annotations. Each variant merging and annotation step uses independent user-defined thresholds of reciprocal overlap, to fine-tune the analysis as desired. An example of script is given in **Supplementary text**. The software is freely available here https://bitbucket.org/guyrouleaulab/sv_segregation. The parameters were set as follow: length comprises between 1 bp and 100 Mbp; an overlap of 25% between two CNVs is needed to be identified as a segregating CNV.

CNV visualization and validation

Potential CNVs were manually examined in the Illumina Genome Viewer of GenomeStudio. The change of B allele frequency and LogR ratio is compared between copy number neutral and copy number variations.

Declarations

Ethics approval and consent to participate

Consent for publication

Availability of data and material

The SV-Segregation (Structural variant segregation) is freely available here https://bitbucket.org/guyrouleaulab/sv_segregation

Competing interests

The authors declare no conflict of interest.

Funding

This study is funded by Canadian Institutes of Health Research (CIHR).

Authors' contributions

Qin He prepared the genotyping data and called CNVs from whole-exome sequencing data; Boris Chaumette performed the segregation and statistical analysis; Amelie M. Johnson was

responsible of generating the SNP genotyping data and calling CNVs from genotyping data; Dan Spiegelman and Alexandre Dionne-Laporte were responsible for developing the segregation software; Gabrielle Houle provided statistical advice; Guillaume Huguet and Sébastien Jacquemont offered technical perspectives and expertise; Lan Xiong and Guy Rouleau were responsible for the study design and overall operation of the project, as well as manuscript writing and editing. Lan Xiong was responsible for managing the project and supervising the genotyping, sequencing experiments and all data analyses, and finalizing the manuscript.

Acknowledgements

We thank all the families' participation in this study; this project would not be possible without their cooperation. The project is supported by a CIHR operating grant to L.X., and G.A. R. Q.H. is supported by a joint scholarship from the Université de Montréal and the Chinese Scholarship Council. Q.H. and B. C. would also like to express their gratitude for Cynthia Bourassa's, from Dr. Guy Rouleau's lab, help with the language editing of the manuscript.

Algorithm	Likely false positive		Likely true positive		Sensitivity	Specificity
PennCNV	1530	18.85%	6586	81.15%	0.30	0.86
QuantiSNP	7275	30.69%	16426	69.31%	0.74	0.32
CNVpartition	3032	29.80%	7141	70.20%	0.32	0.72
XHMM	999	31.97%	2126	68.03%	0.10	0.91
CoNIFER	460	28.20%	1171	71.80%	0.05	0.96

Table VII. Number and percentage of likely false positive CNVs and likely true positive CNVs in autosomal chromosomes and estimation of the sensitivity and the specificity for each software

Likely false positive CNVs are singleton CNVs; likely true positive CNVs are defined as segregating CNVs (in ≥ 2 family members).

Algorithm	Likely false positive		Likely true positive		Sensitivity	Specificity
PennCNV	44	11.20%	349	88.80%	0.26	0.85
QuantiSNP	195	14.21%	1177	85.79%	0.89	0.32
CNVpartition	77	23.33%	253	76.67%	0.19	0.73
XHMM	23	35.38%	42	64.62%	0.03	0.92
CoNIFER	1	3.70%	26	96.30%	0.02	-

Table VIII. Number and percentage of likely false positive CNVs and likely true positive CNVs in X chromosome and estimation of the sensitivity and the specificity for each software

Likely false positive CNVs are singleton CNVs; likely true positive CNVs are defined as segregating CNVs (in ≥ 2 family members).

Parameter	Likely false positive (m ± sd)	Likely true positive (m ± sd)	p-value
QC score	16.5 ± 68	31.4 ± 87	< 10 ⁻³
Number of SNP	8.6 ± 13.6	10.5 ± 15.5	< 10 ⁻³
Size	44 ± 123 kb	56 ± 132 kb	< 10 ⁻³
Number of algorithms	1.3 ± 0.5	1.5 ± 0.7	< 10 ⁻³
Overlap in 3 algorithms	3 ± 10%	5 ± 17%	< 10 ⁻³
Overlap in 2 algorithms	6 ± 17%	10 ± 22%	< 10 ⁻³

Table IX. Features of likely false positive and likely true positive CNVs called from the genotyping data.

Parameter	Likely false positive (m ± sd)	Likely true positive (m ± sd)	p-value
Size	77 ± 209 kb	87 ± 242 kb	0.184
Number of algorithms	1.13 ± 0.34	1.26 ± 0.44	< 10 ⁻³
Overlap in 2 algorithms	5 ± 17 %	12 ± 25%	< 10 ⁻³

Table X. Features of likely false positive and likely true positive CNVs called from the WES data.

	<i>CNVs from SNPchip</i>		<i>CNVs from WES</i>	
	<i>Likely positives</i>	<i>likely negatives</i>	<i>Likely positives</i>	<i>likely negatives</i>
total	20750	9635	2664	1309
<i>CNV size</i>				
CNV size ≥ 20 kb	10293 (50%)	4090 (42%)	1421 (53%)	643 (49%)
CNV size ≥ 100 kb	2597 (13%)	817 (8%)	525 (20%)	226 (17%)
<i>Number of probes</i>				
number of probes ≥ 5	13000 (63%)	5294 (55%)		
number of probes ≥ 10	6489 (31%)	2320 (24%)		
<i>Confidence score</i>				
confidence score ≥ 5	14762 (71%)	5465 (57%)		
<i>Intersection between algorithms</i>				
detected by ≥ 2 algorithms	8100 (39%)	1967 (20%)	701 (26%)	174 (13%)
detected by ≥ 3 algorithms	3082 (15%)	551 (6%)		
<i>Combination of filtering parameters*</i>				
scenario 1	4818 (23%)	1089 (11%)	567 (10%)	139 (6%)
scenario 2	4376 (21%)	941 (10%)		
scenario 3	4363 (21%)	935 (10%)		
scenario 4	1732 (8%)	274 (3%)		

Table XI. The consequence of using different filtering parameters.

*Combination of filtering parameters:

scenario 1: CNV size ≥ 20 kb, and detected by ≥ 2 algorithms;

scenario 2: CNV size ≥ 20 kb, number of probes ≥ 5 and confidence score ≥ 5 ;

scenario 3: CNV size ≥ 20 kb, number of probes ≥ 5 , confidence score ≥ 5 and detected by ≥ 2 algorithms;

scenario 4: CNV size ≥ 20 kb, number of probes ≥ 5 , confidence score ≥ 5 and detected by ≥ 3 algorithms

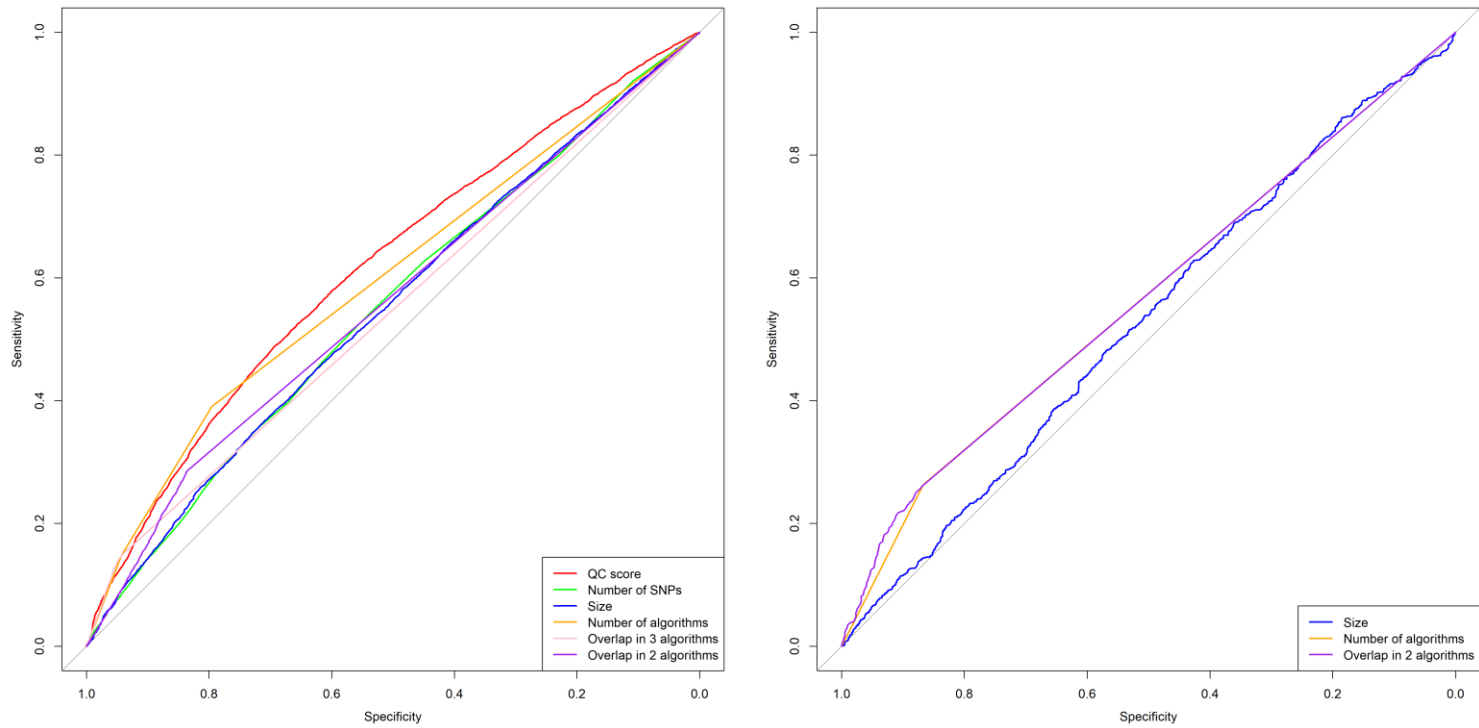


Figure 7. A- ROC curve of features of CNVs called from the genotyping data. B- ROC curve of features of CNVs called from the WES data

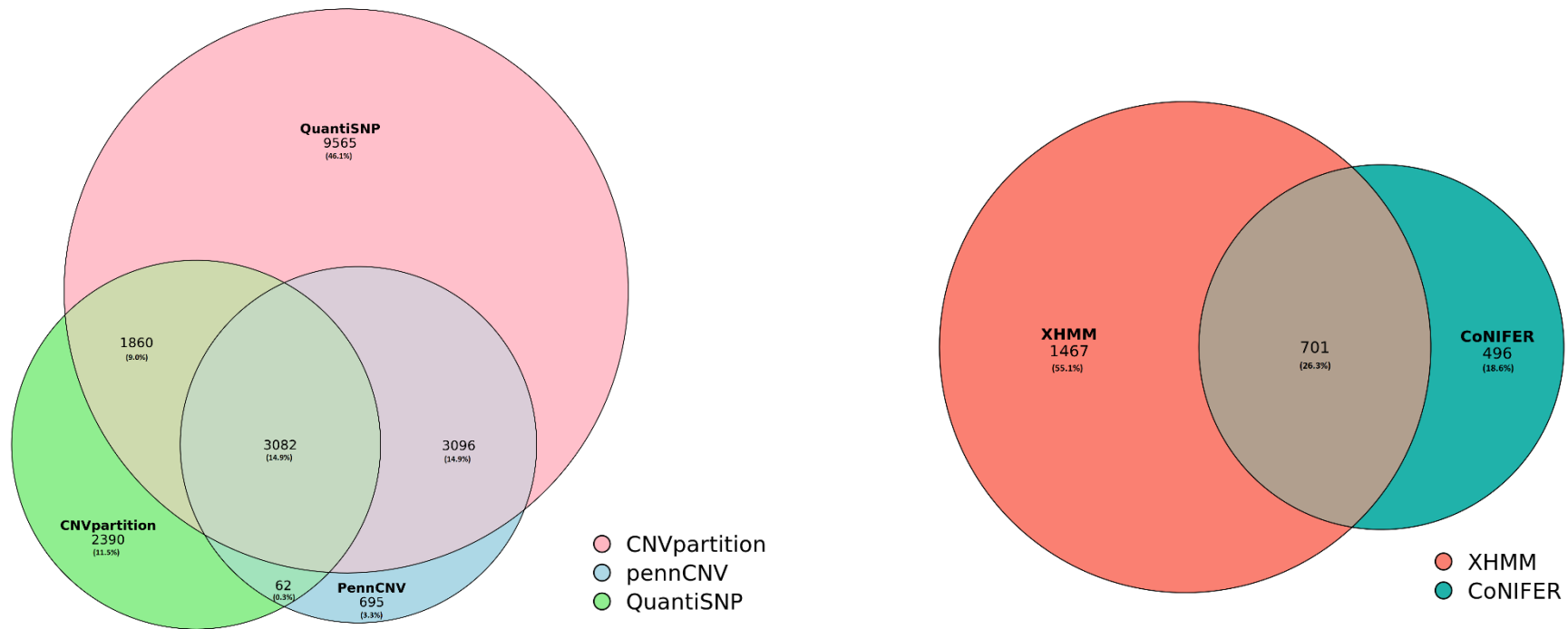


Figure 8. A-Venn diagram of true positive CNVs called from the genotyping data. B- Venn diagram of true positive CNVs called from the WES data

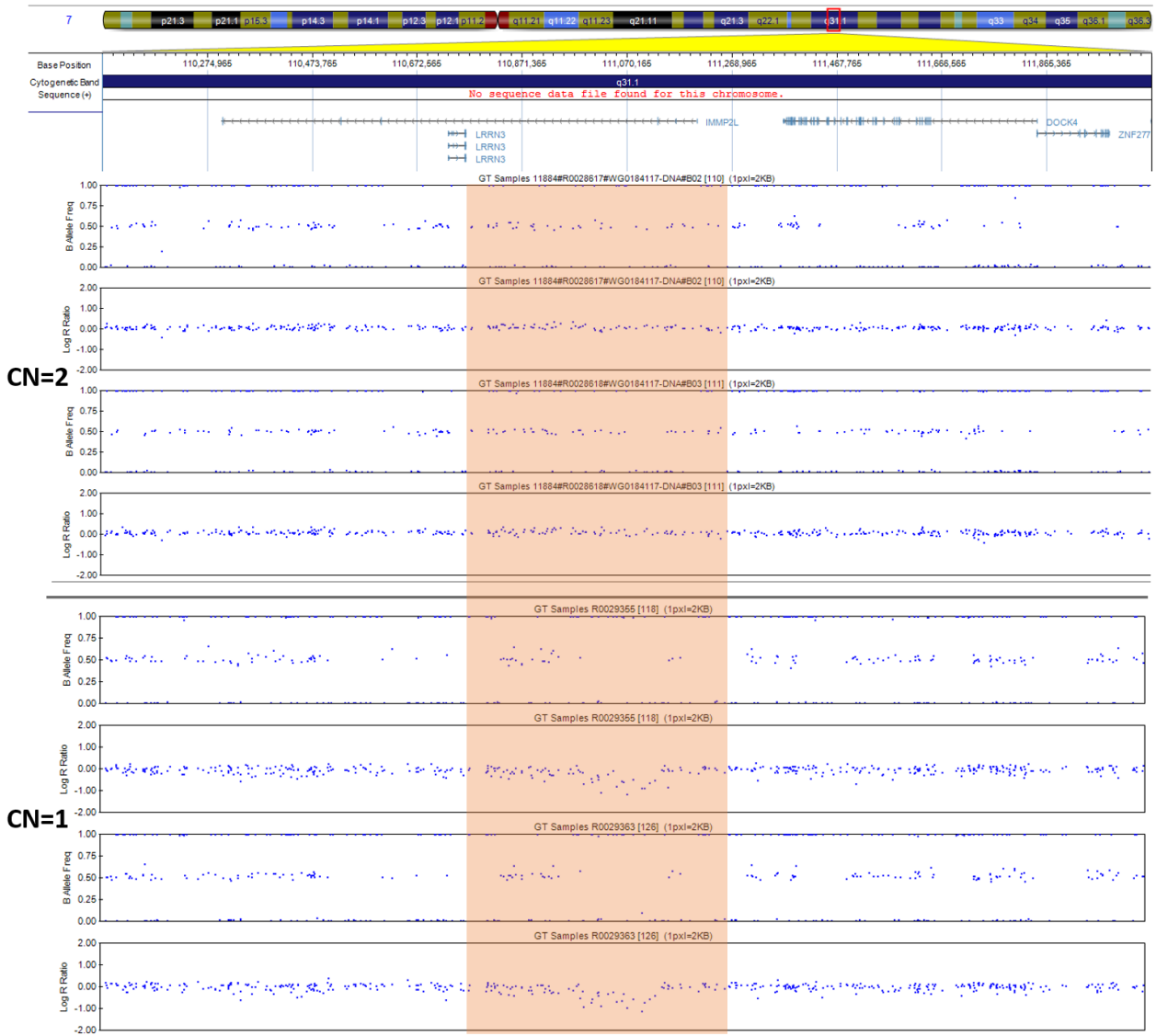


Figure 9. An example of CNV demonstrated by genotyping intensity (7q31.1 deletion)

Chapter 4: Discussion

4.1 The recessive hypothesis and copy number variation calling

The laborious collection of the samples and the phenotypes took time and collaborative efforts from the Pakistani families, local psychiatrists, coordinators, English-speaking psychiatrists, and researchers. The detailed procedure and the construction of the pedigrees are included in the supplementary information of Chapter 2. The update of phenotype in these families was followed up and documented.

In Chapter 2, we validated phenotypic diagnosis based on the profile of the scales on patients' subphenotype dimensions. These families demonstrated an inherited pattern of schizophrenia and bipolar disorder. However, the mode of inheritance was complicated by the consanguineous relationships in these families. The observation on the high rate of consanguineous marriages and the high incidence of the major psychiatric disorders automatically introduced the first question: is consanguinity associated with the phenotypes?

With the available genotyping and sequencing data, we were able to calculate the inbreeding coefficient for each individual with the corresponding methods. We noticed that the exact values of the inbreeding coefficients differed between genotyping and sequencing, but they had a high correlation – the trend stayed the same. We also showed evidence of high inbreeding in these families, but we failed to correlate the inbreeding level with neither the binary phenotype nor the quantitative scale of subphenotypes.

The preliminary comparison with matched population controls showed these extended pedigrees had slightly higher inbreeding levels. However, this comparison was underpowered since our sample size was small. It also brought challenges when we combined external controls to our own dataset. We have included different data sets for different purposes from different consortia, where each of them could be generated by multiple batches and multiple platforms. It is very important to establish systematic quality control procedures to align external resources with internal data sets.

The concerns on combining sequencing data from different resources include the following: the expected sequencing coverage, the sequencing platforms/centers, and batch

effects. When we tried to align our WES data with the WES data of EBI3222, we had difficulties to make them comparable in absolute number of variants per individual. Even though the DNA library were both captured by the same kit and sequenced on the same Illumina sequencers, the difference could be due to the fact that the latter was sequenced with a lower expected coverage (~40X compared to the average of ~100X of our WES samples).

We didn't present our results about the comparison on the genomic burden of homozygous truncating variants and deleterious missense variants between our affected/unaffected family members and the matched population controls, because we had some statistical challenges. We need to apply statistical methods to controls these parameters, while risking losing real interesting candidate variants. However, the large data set of ~3000 Pakistani healthy controls provided a valuable reference for minor allele frequency, besides that of the South Asian population from ExAC dataset, when we need to filter rare variants in Pakistani population.

Major technological issues about our SNP genotyping data worth to mention are: 1) currently available commercialized genotyping microarrays are designed to probe on informative SNPs of European ancestry; therefore, there could be some SNPs neglected by these microarrays that are specifically informative to Pakistani populations; 2) the coverage of our SNP genotyping array was about 700,000 SNPs genome-wide (1 SNP every 50 kb), much smaller compared to the most commonly used ones (~2,5 million SNPs) nowadays. This coverage continued to decrease when we tried to merge our SNP data with the external dataset such as HGDP data (~600,000 SNPs); we only get 300,000 overlapped SNPs and it generated low-resolution results.

Despite the abovementioned technological issues, the long history of admixture and the recent inbreeding of Pakistani populations was well characterized in our data. The co-clustering of the families with other Pakistani populations, through admixture analysis with other world populations and the ROH analysis, showed longer ROHs in these families due to recent common ancestry. However, we noticed that the proportions of ancestry were different from one family to another, which suggests heterogeneity across families.

The conventional linkage analysis with SNP genotyping data was conducted in five out of the ten pedigrees, with different combinations of analyses in two different software, parametric and nonparametric, single-point, two-point and multi-point. The signals detected by these analyses were inconsistent, which is a common problem for most genetic analysis. The family structure of our pedigrees was too complicated for most available linkage analysis to compute, in which the consanguineous loops needed to be broken and the family structure was simplified. For the recessive model, the loci identified through linkage were further examined with autozygosity mapping results and homozygous variants from sequencing data – we were not able to confirm the loci were positive. Nonetheless, in the future, other genetic models could be checked, and a thorough linkage analysis could be done for all ten pedigrees. Otherwise, we cannot conclude that the linkage analysis failed to detect peaks for these families.

We also performed autozygosity mapping and ROH analysis on these pedigrees with different software. Our objective was to identify homozygous regions shared by the affected family members and not by the unaffected members. The methods were complementary, hence the signal detected by one could be validated by another. We were unable to find any homozygous regions segregating with the phenotypes. We further tried to find some homozygous regions presenting more often in affected than in unaffected (considering the possibility of phenocopies in each family), and zoomed in to identify rare homozygous deleterious variants. Most of the homozygous variants shared by majority of the affected members were also common in the Pakistani population, which ruled out their possible role as rare high-penetrant variants contributing to the phenotypes.

Researchers studied the number, the size and distribution of ROH in both inbred and outbred populations and they found either positive or negative association between ROH and schizophrenia. We tested this hypothesis in these families by comparing the ROHs between affected and unaffected family members. We found no excess of ROHs associated with the phenotypes in terms of the size, number and length. The statistical methods used for the test varied from one study to another. The limitations in our results not only lied in choosing the appropriate methods while controlling certain covariates such as relatedness, but also calculating the power based on our sample size and the unknown effect size.

Another type of recessive rare variants – the compound heterozygotes – was also examined, resulting in no rare heterozygotes variants in the same gene observed more often in affected family members than in unaffected family members. In conclusion, the recessive model failed to explain the heritability of phenotypes in these families. The results could be confirmed as negative, if we excluded the following: 1) the possibility of imperfect phenotyping; neuropsychiatric disorders such as BP and SCZ are extremely heterogeneous. The evaluation of the severity and duration of the patient's disorder are based on familial subjective descriptions of the patients to the psychiatrist and several diagnoses were seen in the same family. Since the combination of genomics and phenomics of common complex diseases is still at the early stage, a systematic and detailed phenotyping profile may be needed in the future, with the aid of other objective measurements of the phenotypes such as imaging data. Ideally, the same evaluation should be applied to all the family members including unaffected ones. 2) A later onset of the disease. The individual may not have expressed the phenotype at the time we conducted the study but may develop the disorders in several years, because the ages of onset has a wide range. 3) Incomplete penetrance. To rule it out would require an examination on the full spectrum of the penetrance. However, for a recessive model, we considered variants shared by the majority of the affected family members, regardless of the genotypes of the unaffected relatives. 4) The disease could be polygenic or omnigenic, hence alternative hypotheses and methods are needed.

In chapter 3, we focused on finding a better way to analyze CNVs in our families with popular bioinformatic tools. The CNVs were called from SNP genotyping data with three parallel software and from WES data with another two software. The algorithms take advantage of different data points and make CNV calls. The quality control should be the most important process in combining the calls. Researchers usually set CNVs as true positives if they are detected by more than one software and pass a threshold on size and density of covered SNPs, but interesting CNVs could be missed in the segregation analysis of a pedigree design by doing so. The filtering parameters we applied to get a cleaner set of candidate CNVs would increase the frequency of true positives. The trial solution was to carry out the minimal quality control for calls from each tool and include all segregating CNVs in at least two family members. In this way, we could evaluate the sensitivity and specificity of each tool and the rate of the overlapped calls. Surprisingly, the rate of false positives was very high, and the rate of true

positives was low. Henceforth the sensitivity and specificity were variable and indecisive. However, the method is not a good predictor of true and false positives. The mutual overlap of the software was also very low (~10%), so we suggest combining the calls from different approaches to increase the chance to identify potentially pathogenic CNVs.

The segregation analysis of CNVs in these families failed to identify any CNV segregating perfectly among affected family members. The large number of affected individuals could be one reason that reduces the possibility of locating a shared CNV. We further loosened our criteria to detect CNVs by looking at the ones which are shared by more than half of the affected individuals and have been previously associated with neuropsychiatric disorders. The shortlist of these CNVs would demonstrate an incomplete penetrance. Recent studies reported an increased burden of rare, exonic CNVs in schizophrenia probands and genetic heterogeneity in multiplex families, and included singleton CNVs that have been associated with schizophrenia previously²⁵⁴. Other major studies conducted on case-control cohorts also reported a global enrichment of CNV burden among cases¹²⁸. This means our analysis of CNVs would have to be combined with other types of deleterious variants to gain a thorough examination of potentially pathogenic and biologically relevant variants for affected individuals, even though the individuals were from a more homogeneous family.

In summary, our attempts in finding rare homozygous deleterious SNVs and rare pathogenic CNVs segregating variants are not rewarding thus far. We could partially attribute this to the limitations of genetic data we were using. All the analyses we have done were based on SNP genotyping and WES data, in order to define a genomic landscape of the studied families, while referring to available family tree. The SNP genotyping data could not detect smaller CNVs, while WES dataset is only limited to exonic regions. WES samples are typically sequenced to a higher depth (100X versus 30X WGS), and the reads are focused on only ~2% of the genome. The enrichment step in WES, where DNA or RNA baits are used to hybridize with the coding regions of the genome, lead to non-uniform coverage, generating both regions with too much coverage and too little coverage (resulting in missed variant calls). The PCR-based enrichment steps introduce GC bias and other biases; so, we may have missed rare variants. WGS could be the next strategy for finding causal variants, since our current data is limited by its coverage of full profile of the genome. WGS generates more uniform coverage of

the genome and it can take advantage of longer reads (compared to < 200 bp of the majority of human exons in WES), which allows for better determination of CNVs and other structural variations. Also, the regulatory variants from WGS data could be thoroughly analyzed and annotated. WGS is the trend for identifying SNVs and CNVs in family studies, which is more advantageous and informative than WES, but the latter is more cost-effective for most genetic studies. If the cost of WGS drops and the coverage of sequence increases in the next few years, it would be worthwhile to look through the whole genome for CNVs and even regulatory variants.

4.2 Alternative hypotheses to explore

Causal variants that aggregate in families usually have larger effect sizes than those found in sporadic cases. Additionally, the family-based designs are robust to confounding due to population admixture or substructure. Therefore, family-based designs can be a more powerful approach than population-based designs. In our case, even though we had enough power to detect extremely rare variants in these large families, we were not able to detect any one or any set of variants directly linked to the phenotype. What are the alternate hypotheses? Based on the literature review explored in the introduction, there are some statistical analyses we could apply to these families, on current WES and genotyping data. It includes the burden of rare variants and the polygenic risk of common variants.

Fewer tests have been proposed for family-based studies, compared to population-based studies, of NGS data for rare-variant associations. They are also mostly designed for different family structures. For example, the transmission disequilibrium test (TDT) runs rare-variant association tests for nuclear families with no more than one affected child²⁵⁵; family-based association test (FBAT) analyzes sequence data in the rare-variant burden test on case-parent trio data²⁵⁶; there is another statistical approach applied to affected sibships in nuclear families²⁵⁷; RareIBD analyzes large extended families of arbitrary structure, assuming that only one founder in a family carries a rare variant in a given gene²⁵⁸; and a rare-variant extension of the generalized disequilibrium test (RV-GDT) for both nuclear and extended families also exists²⁵⁹. The last one claims that: it utilizes genotype differences of all discordant relative pairs to assess association within a family, it combines the single-variant GDT statistics over a

genomic region, and it increases power by incorporating the information beyond first-degree relatives. We have tried RV-GDT test with all ten families combined, and the association for each single one did not result in significant results.

Other studies use simple statistical tests, such as non-parametric test or mixed-model analysis, to compare the genomic burden of rare deleterious variants (truncating and missense) between affected and unaffected⁹⁴. This approach gives a direct comparison on the total number of variants. Alternatively, potentially etiologic variants are filtered based on their frequency; rare variants are more likely to have a recent origin and are, therefore, more population-specific than common variants. Using this strategy, it is important to have the correct reference population for the MAF. In addition to their frequency, variants can be filtered by their functional prediction and their cosegregation with the disease. In the latter case, we look for whether they are shared by a reasonable number of affected family members and minimum of unaffected family members. Another list of likely neutral variants can be constructed with the variants predicted to be non-pathogenic and shared by some unaffected relatives and a maximum of one affected relative. We could also explore the role of *de novo* variants in these families, as 34 trios have WES data and 11 of them are from the same pedigree (MNS09).

The polygenic scores can be used to determine whether common alleles associated with SCZ or BP in the general population also confer risk of the phenotypes in our families. It can be calculated based on summary statistics from the most recent mega-analyses²⁶⁰ of SCZ and BP, using the p-value threshold explaining the greatest variance for the relevant disorder. Of note, the largest GWAS thus far were done in European and East Asian populations, and they have reported overlapping GWAS loci and other loci specific to each ancestry. The polygenic scores of affected and unaffected individuals could later be subject to logistic regression, in order to obtain the difference between them.

Using the raw data and summary statistics of large consortia has become the trend in the genetics field. We either use them as a reference, or directly align and compare them with our own data. It is beneficial for increasing the statistical power. However, quality assessment and quality control are necessary and important steps before applying them to any subsequent statistical analyses.

Beyond the genetic contribution, we could also investigate the role of epigenetics (heritable changes in gene expression, active versus inactive genes, a change in phenotype without a change in genotype). For complex diseases like SCZ and BP, other factors including age, environment, lifestyle and disease state could introduce epigenetic changes. Postmortem studies of human SCZ and BP brains show considerable alterations in the transcriptome^{261,262} of a variety of cortical structures, including multiple mRNAs that are downregulated in both inhibitory gamma-aminobutyric acid (GABA)-ergic and excitatory pyramidal neurons, compared with non-psychiatric subjects. Several reports show increased expression of DNA methyltransferases (*DNMT1*) in telencephalic GABAergic neurons^{263,264}. Accumulating evidence suggests a critical role for altered DNA methylation processes in the pathogenesis of SCZ and related psychiatric disorders²⁶⁵. *DNMT1* is selectively overexpressed in GABAergic interneurons of schizophrenic brains, whereas hypermethylation has been shown to repress expression of Reelin (a protein required for normal neurotransmission, memory formation and synaptic plasticity) in brain tissue from patients with schizophrenia and patients with bipolar illness and psychosis.

Phenomics, large-scale phenotyping, was proposed to be the natural complement to genome sequencing as a route to rapid advances in biology²⁶⁶. The Consortium of Neuropsychiatric Phenomics (CNP) is a centrally funded project with a truly phenomic vision and focuses on a set of neural and psychological phenotypes. There is currently a broad chasm between the basic and clinical research strategies used to study these disorders. The ultimate goals of the CNP are to facilitate discovery of the genetic and environmental bases of variation in psychological and neural system phenotypes and to elucidate the mechanisms that link the human genome to complex psychological syndrome. The phenomics of neuropsychiatric disorders is still at the beginning stage, the challenges remain unsolved but there are tools²⁶⁷ and pilot neuroimaging data²⁶⁸ coming out that will help understand the complex dimensions of the neuropsychiatric phenotypes. The combination of genomics and phenomics will be the trend in the near future.

Conclusion

This thesis reviewed the historical interest on consanguineous populations/family and the genetic principles behind. It also gathered genetic studies on schizophrenia and bipolar disorder. With current available genotyping and sequencing data, it examined the role of consanguinity in psychiatric diseases in large multiplex consanguineous pedigrees. In addition to the negative results of the association analysis between consanguinity and the phenotypes, we were unable to identify risk genes with rare homozygous variants, to some extent which disproved the recessive mode of inheritance in these families. The last part of the thesis presented the current limitations of identifying CNVs, both in technology and methodology. We recommended to perform segregation analysis and function filters in a familial design. The resulting CNVs, along with other type of rare damaging variants, putative but inconclusive, that are segregating in these families showed incomplete penetrance and evidence of heterogeneity within and between families. A comprehensive examination of both rare and common variants are needed in these families, in order to estimate the effects of genetics on phenotypic variance.

Appendix 1: supplementary material for Chapter 2.2

Supplementary text: phenotype evaluation, sample collection, and family tree construction

The familial aggregation of major psychiatric disorders was noted by a local senior psychiatrist, Dr. Qasim, who has inquired into the family history and meticulously documented the complicated family trees under the advice and help from an Australian geneticist, Dr. Mike Denton and with a significant contribution from a local biologist, Mr. Mehtab Christian.

1.1 Ascertainment of probands and families: The 10 large multiplex consanguineous pedigrees with major psychiatric disorders were recruited in Sindh (Pakistan). All interviews, clinical examinations, and blood collections were conducted on-site during numerous field trips.

1.2 Clinical assessments: Dr. Qasim was trained to use DIGS and FIGS (v 3.0) at the beginning of the project in 2002. He then performed a DIGS, and a brief psychiatric, neurological and medical examination on each individual with psychiatric symptoms included in this project. All the families have been followed by Dr. Qasim during the ongoing project and newly identified affected individuals were marked. Two independent FIGS were also conducted on each by two reliable informants to recount the observed psychiatric symptoms and to confirm the reliability of the phenotype information. The interviews were conducted using an appropriate translation in Sindhi language.

1.3 Confirmation of clinical diagnosis and evaluation of phenotyping work: Dr. Ridha Joobar (Dept. of Psychiatry, McGill University) and Dr Lynn DeLisi (previously Dept. of Psychiatry, New York University, currently Dept. of Psychiatry, Harvard Medical School) have been responsible for the standardized evaluation of the clinical diagnoses after reviewing a copy of the DIGS, FIGS, along with a clinical narrative. A final diagnosis based on DSM-IV criteria and comments on each individual diagnosis was made by Dr. Joobar and Dr. DeLisi, separately and blinded to each other. Whenever a discrepancy on final diagnosis occurred among the three psychiatrists, a teleconference was arranged. A final consensus diagnosis was usually reached by additional information provided by Dr. Qasim or by further discussion among the

psychiatrists. In case of uncertainty, notes were taken and the individual was marked on the pedigree with either “unknown” or “possibly affected” status.

1.4 Blood sample collection and transformation of cell lines: A batch of blood specimens from one field trip were usually delivered to the DHL office in Hyderabad the same day. Then, blood samples were achemined in Dr. Rouleau’s laboratory (McGill University – Montreal – Canada) within 3–4 working days. Excellent quality (one incidence of sample mixed-up in total collection) and quantity of DNA was obtained for each sample. Dr. Rouleau’s lab. All DNA samples are available upon request.

1.5 Demographic and genealogical information and ethical concerns: Mr. Mehtab Christian has been working with Dr. Qasim and local historians from each village to obtain detailed genealogical and demographic information. Most of the genealogical information was passed on and kept verbally by the senior people in the village. Approval of the research project by the local review committee of ethics, and in particular, consent and approval from each caste/community authority of the families were obtained prior to the study. An informed consent form in the local language was signed by each individual or by their guardian (parent or caste authority) if mentally incapable or illiterate.

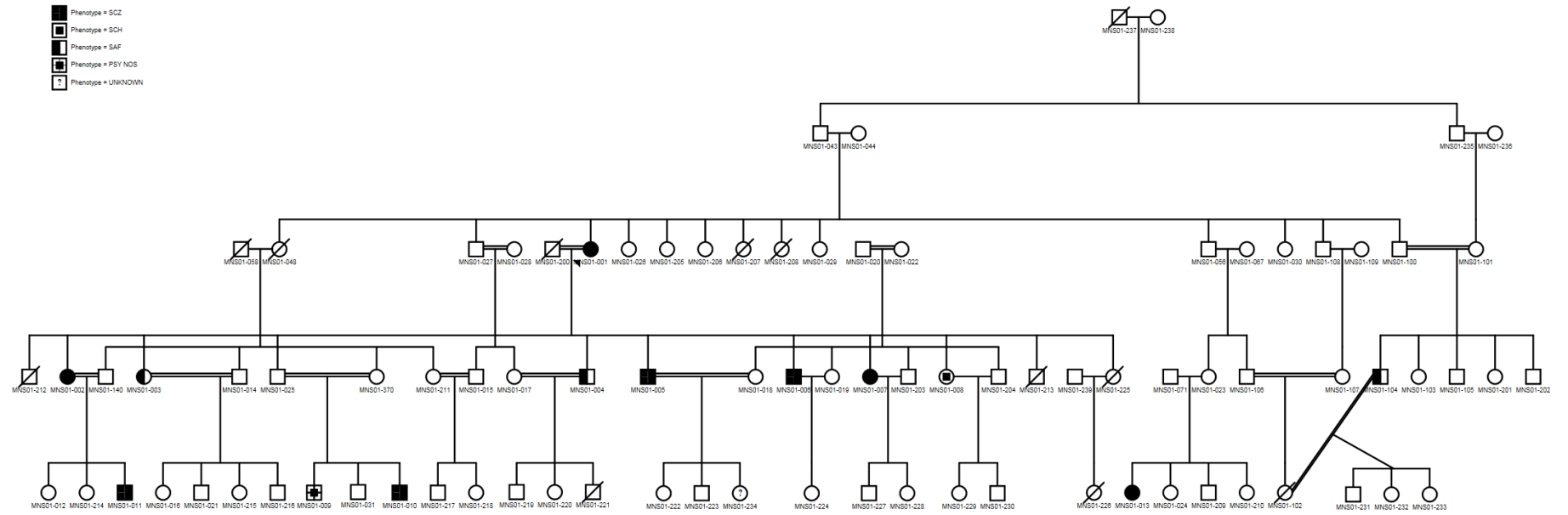
1.6 Reconstruction of pedigrees and phenotype database: Each pedigree was reconstructed using Progeny v9.0 standalone software, including all the collected samples, and integrating the final diagnoses (**Supplementary figure 1**). This work was performed under the supervision of Dr. Xiong, and was verified with Mr. Mehtab Christian. All information from DIGS, FIGS, and clinical summary (over 600 variables for each individual) were digitalized into a database with appropriate security administration.

1.7 Common characteristics of these pedigrees: The relevant common features of these pedigrees and individuals are: (1) all collected individuals/pedigrees are Sindhi-speaking Muslims; (2) each pedigree belong to a different caste or clan and is located in an isolated village or small town; (3) most marriages within these pedigrees are consanguineous (97% of all marriages); inter-marriages of two sets of siblings, and/or first-cousin marriages (42% of

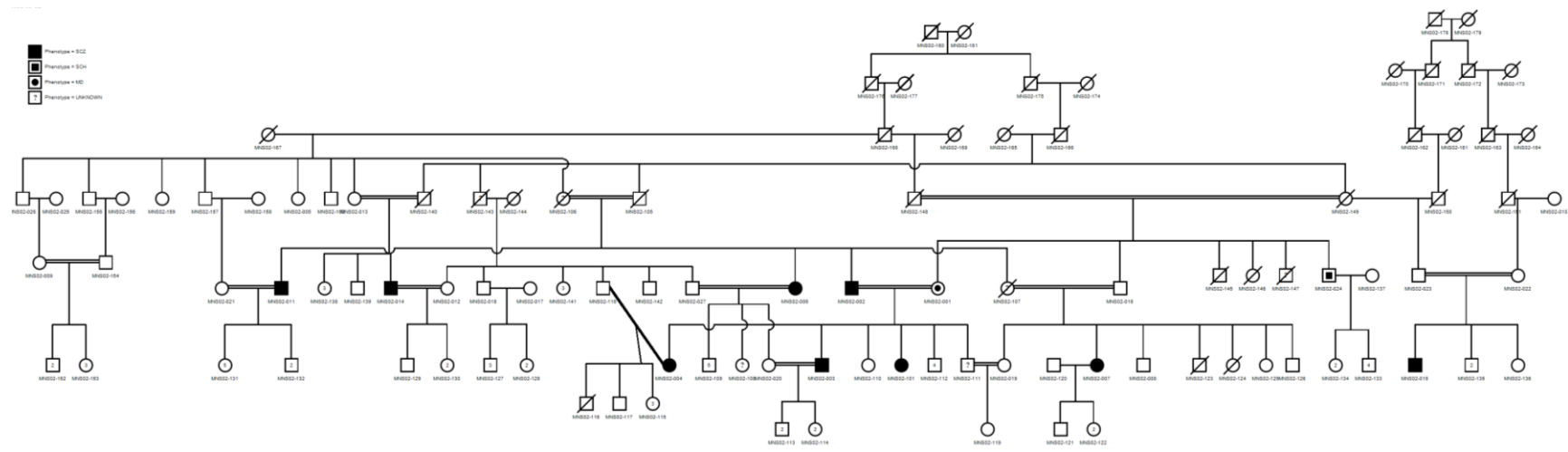
all marriages) are particularly common; and each community is known for long-term strict endogamy; (4) most marriages are arranged; therefore, most of the severely affected individuals are married (69%) and married early in life, so most of them have offspring (62%) prior to the onset or deterioration of disease; (5) some of these pedigrees might have common ancestors, according to the available genealogical information; e.g. MNS06 and MNS08, carry the same caste name; and pedigrees MNS03, 04, 07, and 09, belong to a major Baloch tribe in Sindh; (6) at least 10 affected individuals have been collected for each pedigree, as well as living parents, siblings, and other relatives; (7) though in general the phenotype is highly variable and diverse, each pedigree aggregated with one major phenotype, either schizophrenia, bipolar disorder or schizoaffective disorder.

Supplementary Figure 1. Pedigree tree of the 10 MNS families

MNS01

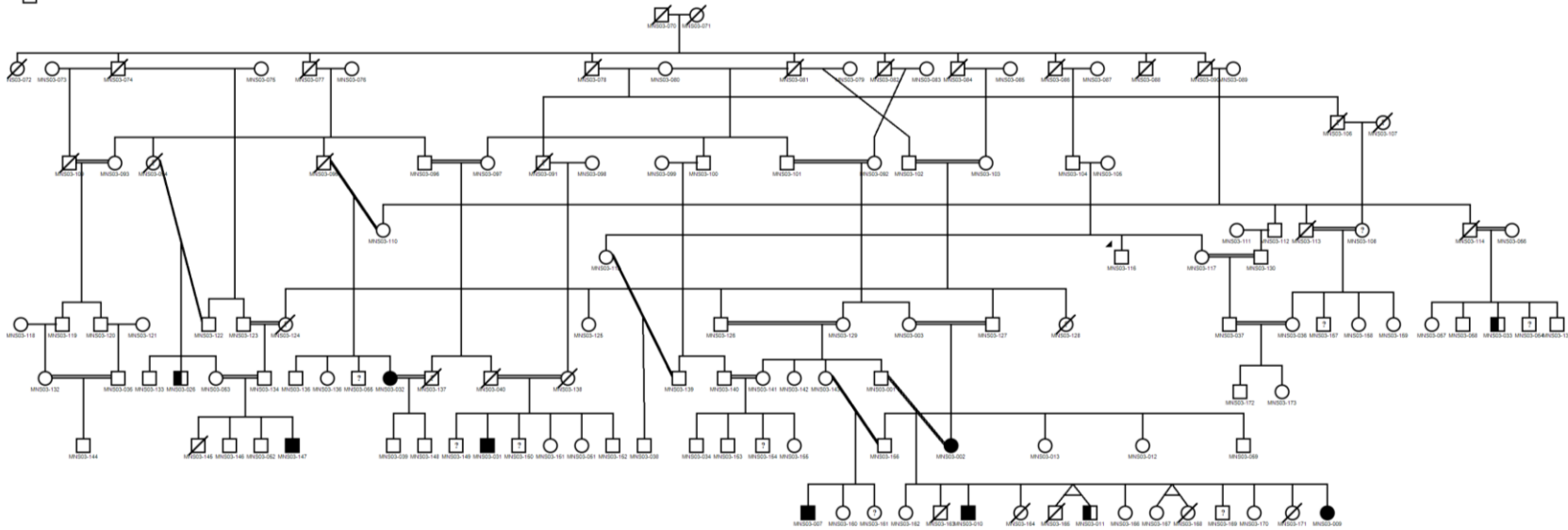


MNS02

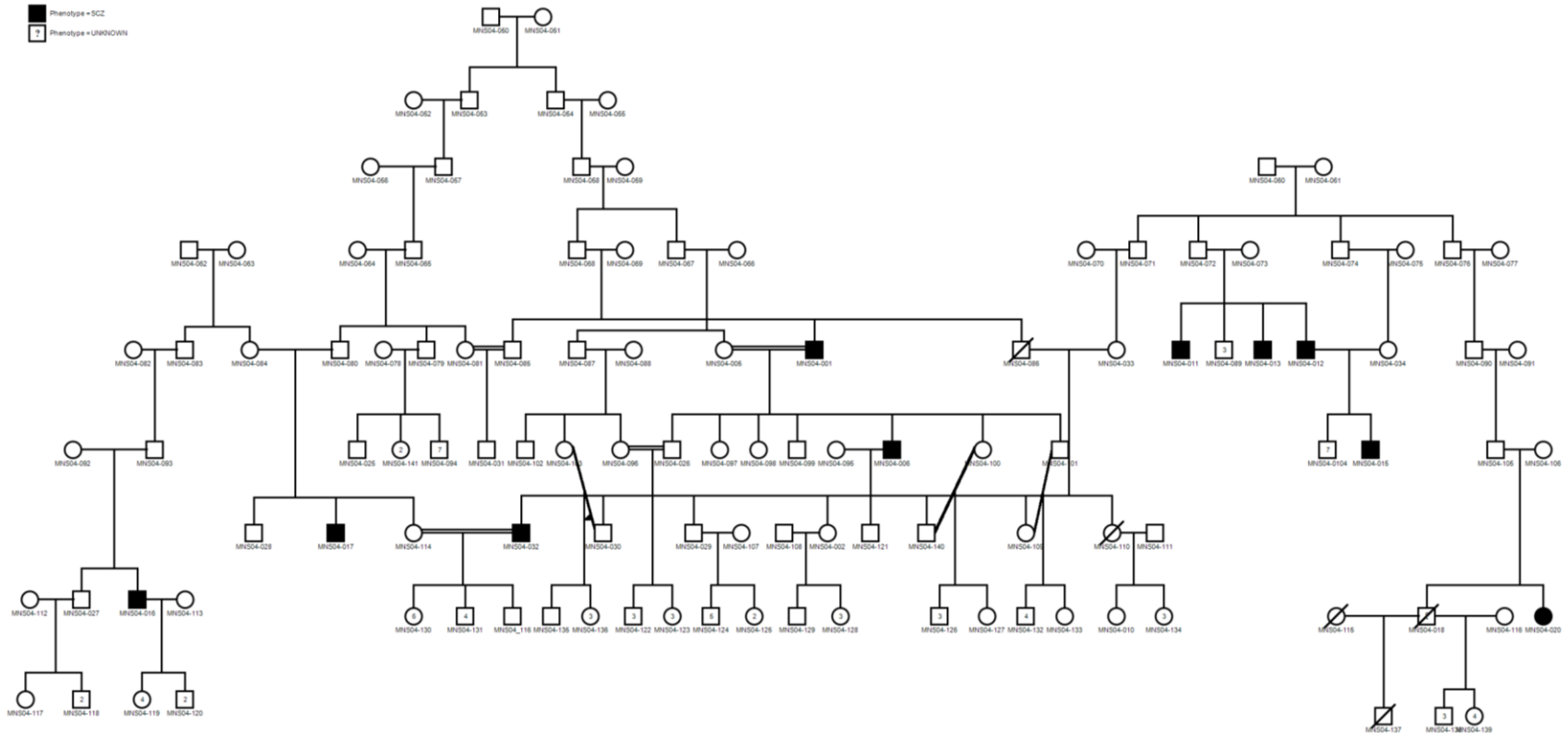


MNS03

- Phenotype = SCZ
- ◼ Phenotype = Sch
- ◻ Phenotype = UNK/NOI

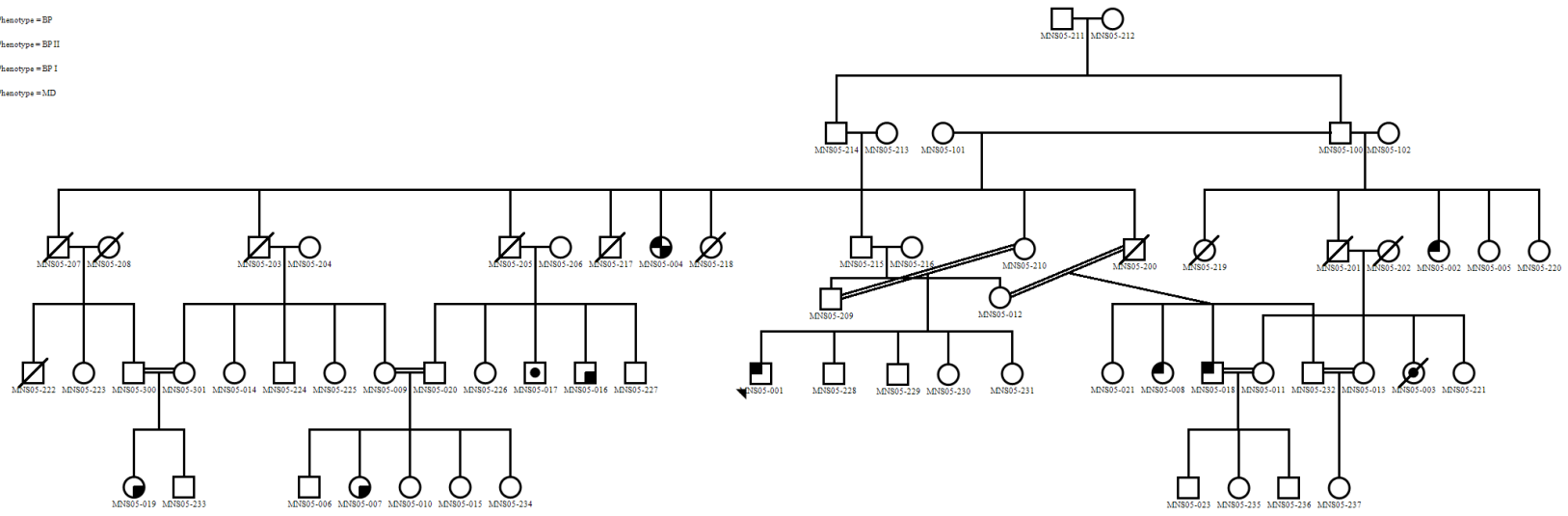


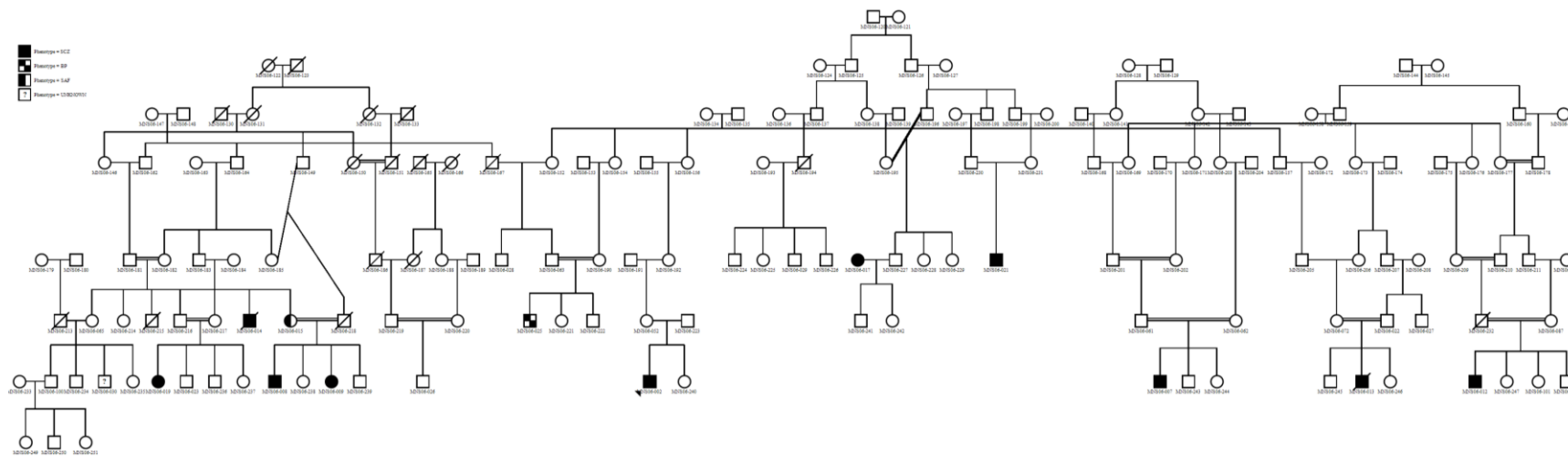
MNS04



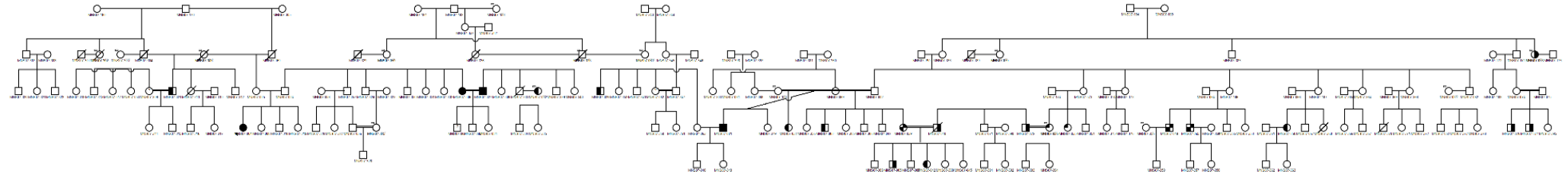
MNS05

- Phenotype = BP
- ◻ Phenotype = BP II
- ◻ Phenotype = BP I
- Phenotype = MD





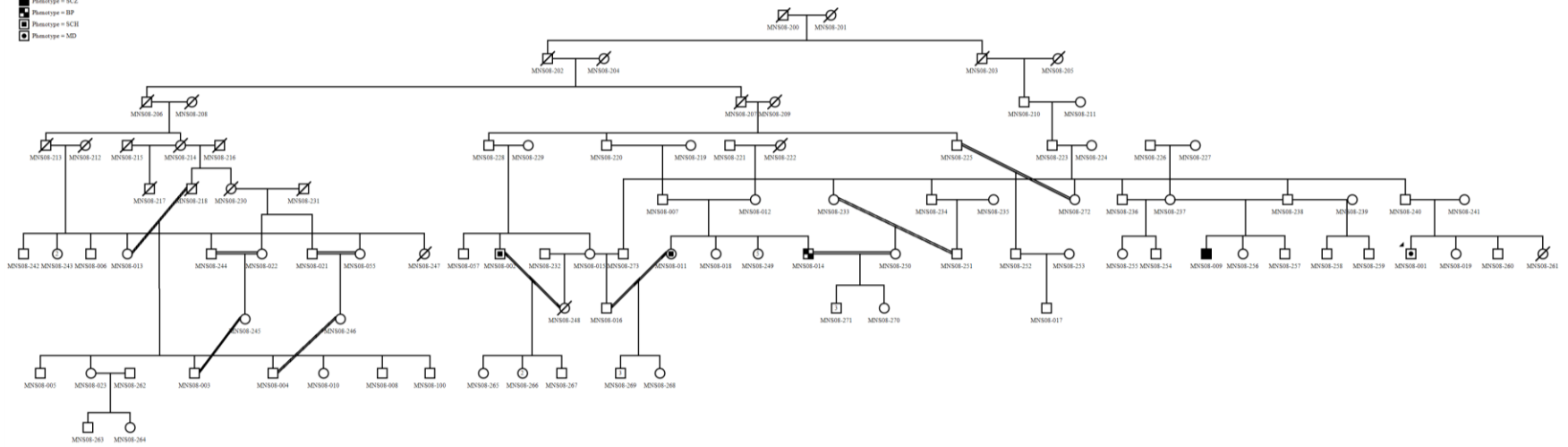
MNS07









- Phenotype: S22
- Phenotype: S22
- Phenotype: S22
- Phenotype: S22
- Phenotype: S22
- Phenotype: S22

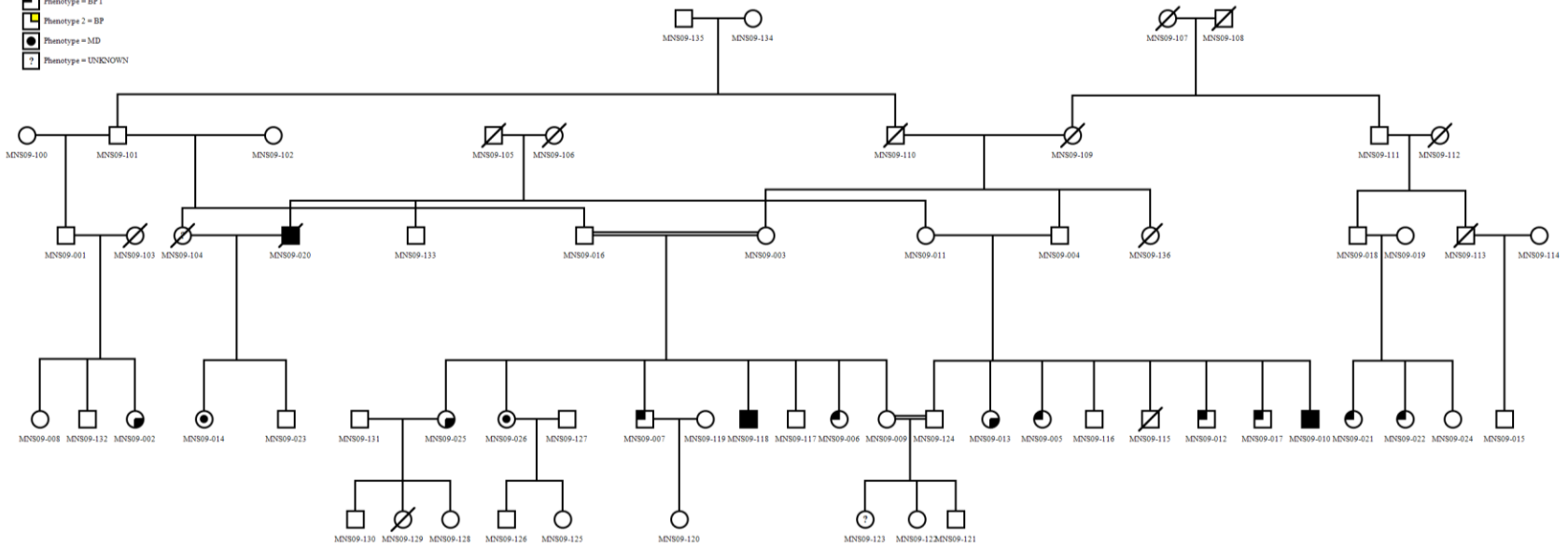
MNS08

- Phenotype = ICZ
- ◼ Phenotype = BP
- ◻ Phenotype = ICE
- ◻ Phenotype = MD

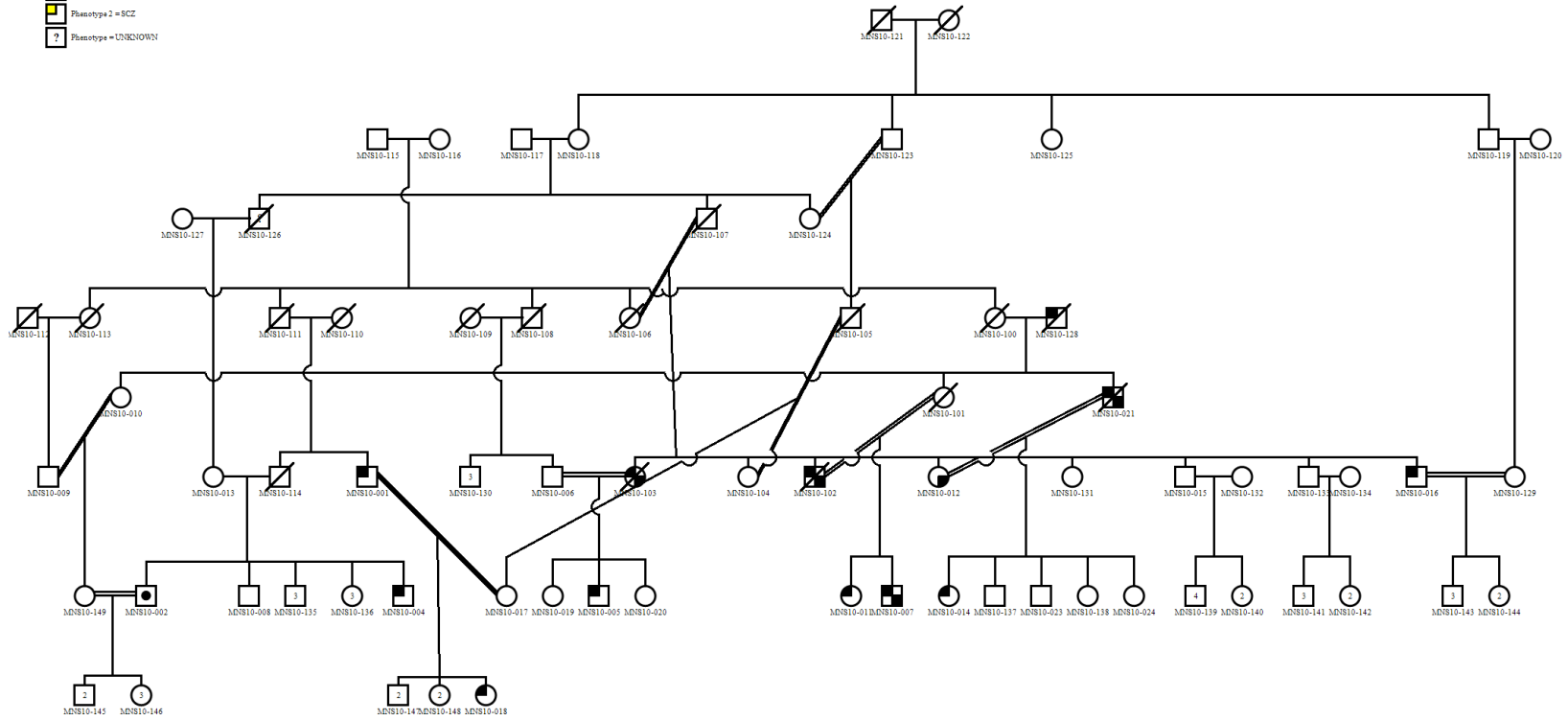


MNS09

-  Phenotype = SCZ
-  Phenotype = BP II
-  Phenotype = BP I
-  Phenotype 2 = BP
-  Phenotype = MD
-  Phenotype = UNKNOWN



MNS10



Supplementary Table 1a: Origin and description of MNS Pedigrees

Pedigree	Location (village/town, District)	Genealogy	Ethnic origin	Main phenotype	Number of individuals collected (affected/unaffected)
MNS01	Nebharo Gaju, Thar	Gaju Caste, Bhanbheer Clan	Rajputs	SCZ&SAF	39 (13/26)
MNS02	Aliabad, Hyderabad	Khosa Bloch Caste, Gohramani Clan	Rind Baloch	SCZ	28 (11/17)
MNS03	Mari, Thatta	Jokhia Caste, Kalaypota Clan	Samma Baloch	SCZ&SAF	31 (13/18)
MNS04	Essa Nohrio, Mirpur	Nohria Caste, Moora Clan	Samma Baloch	SCZ	23 (11/12)
MNS05	Kandiario, Naushahro Feroze	Kalhora Caste	Arabs	BP	22 (11/11)
MNS06	Bachal Soomro, Hyderabad	Soomra Caste, Mulla tribe	Rajputs	SCZ	31 (12/19)
MNS07	Gujo, Thatta	Palija Caste	Samma Baloch	SCZ, SAF&BP	34 (19/15)
MNS08	Essa Soomro, Thatta	Soomra Caste, Mulla tribe	Rajputs	PDD, SCZ&BP	27 (10/17)
MSN09	Tando Ghulam Ali, Badin	Notyar Caste	Samma Baloch	SCZ&BP	26 (14/12)
MNS10	Baradi Panhwer, Dadu	Panhwer Caste	Arabs	BP	23 (12/11)

SCZ: schizophrenia; SAF: schizoaffective disorder; BP: bipolar disorder; PDD: pervasive developmental disorder

Supplementary Table 1b: Summary of individual characteristics of MNS Pedigrees

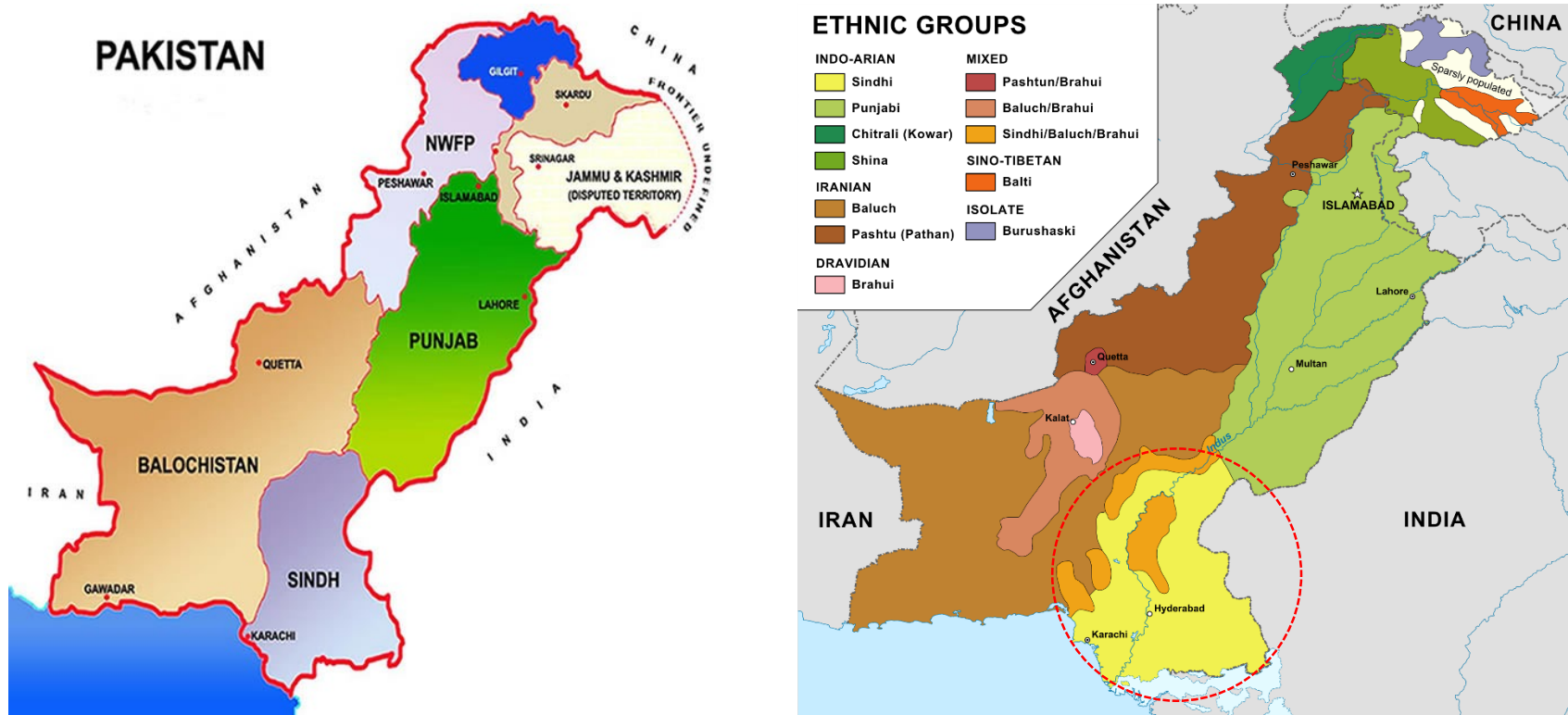
Pedigrees	Number of 1st-cousin marriage	Number of 2nd-cousin & Bradri marriage	Number of total marriages surveyed	Number of individuals surveyed	Number of symptomatic individuals identified	Number of affected individuals married	Number of affected individuals with children
MNS01	4	7	12	118	16	12	12
MNS02	3	4	7	121	19	14	14
MNS03	12	13	25	138	35	21	20
MNS04	2	0	2	193	10	5	5
MNS05	2	11	13	88	14	12	11
MNS06	10	6	16	155	19	5	5
MNS07	4	5	11	231	25	21	20
MNS08	1	6	7	176	12	6	4
MSN09	1	5	6	81	17	11	8
MNS10	3	1	4	51	13	8	8
Total (%)	42 (41%)	58 (56%)	103 (100%)	1352	172	115 (69%)	51 (62%)

Supplementary Table 2: Comparison of ROHs in affected and unaffected family members in each family

	MNS01	MNS02	MNS03	MNS04	MNS05	MNS06	MNS07	MNS08	MNS09	MNS10
number of aff	13	10	12	11	13	13	17	10	14	12
number of unaff	23	16	16	12	9	18	17	16	12	11
total number										
mean±SD (aff)	12 ± 10.9	49.9 ± 13	51.4 ± 10	10.9 ± 6.9	18.8 ± 12.5	33.1 ± 11.6	24.4 ± 10.4	11.6 ± 7.1	12.1 ± 7.4	30.2 ± 17.2
mean±SD (unaff)	10.8 ± 14.4	54.7 ± 7	53.6 ± 9.1	8.4 ± 7.2	19.4 ± 10.4	39.9 ± 9.2	17.9 ± 12.1	16.6 ± 7.9	9.8 ± 5.6	32.4 ± 15.7
Wilcoxon p-value	0.117	0.672	0.593	0.337	0.789	0.144	0.065	0.101	0.679	0.804
total size (Mb)										
mean±SD (aff)	85.8 ± 107.7	377.4 ± 111.5	436.9 ± 133.6	85.4 ± 94	159.7 ± 138.6	217.2 ± 119.9	169.9 ± 94	77.4 ± 67.9	84.8 ± 76.3	229.1 ± 188.1
mean±SD (unaff)	105.8 ± 170.5	442.1 ± 110	501.6 ± 103	79.5 ± 119.9	150 ± 105.9	315.1 ± 134	129.1 ± 112.4	118.4 ± 87.9	49.7 ± 48.6	270.7 ± 151.1
Wilcoxon p-value	0.169	0.310	0.159	0.525	0.896	0.062	0.099	0.121	0.403	0.260
average size (Mb)										
mean±SD (aff)	5 ± 3.3	7.3 ± 1.6	8.4 ± 1.6	6.4 ± 4.1	7.1 ± 2.9	6.1 ± 2.1	6.6 ± 2.3	5.8 ± 2	5.9 ± 2.4	6.9 ± 2.3
mean±SD (unaff)	4.5 ± 4.5	8.1 ± 1.7	9.5 ± 2	6.9 ± 4	6.5 ± 3	7.6 ± 2	6.4 ± 3.1	7.1 ± 2.3	4.6 ± 1.9	8.3 ± 2.2
Wilcoxon p-value	0.328	0.484	0.110	0.651	0.695	0.051	0.734	0.241	0.160	0.118
>4Mb_size										
mean±SD (aff)	72.4 ± 100.9	326.1 ± 101.2	388.3 ± 127.8	73.5 ± 92.3	142.1 ± 131.2	178.5 ± 114	141.1 ± 84.3	64.1 ± 64.3	71.4 ± 70.6	197.8 ± 180.6
mean±SD (unaff)	95.5 ± 159	385.8 ± 115.6	455.2 ± 99.9	71 ± 117.5	133.1 ± 100.5	275.8 ± 134.3	110.8 ± 105.5	101.3 ± 85.1	38.2 ± 44.7	241 ± 141.3
Wilcoxon p-value	0.673	0.310	0.174	0.580	0.893	0.075	0.179	0.133	0.410	0.211
>4Mb_count										
mean±SD (aff)	5.9 ± 8	28.5 ± 8	30.7 ± 7.9	5.7 ± 5.1	10.9 ± 8.7	15.5 ± 8.4	12.8 ± 6.3	5.9 ± 5.4	5.8 ± 4.8	16.9 ± 13.2
mean±SD (unaff)	6.5 ± 10.3	30.8 ± 7.3	33.7 ± 5.9	4.4 ± 5.9	11.4 ± 7.5	22 ± 8.9	9.8 ± 8.4	9.2 ± 5.8	4 ± 3.6	19.6 ± 10.2
Wilcoxon p-value	0.622	0.771	0.295	0.320	0.763	0.089	0.094	0.139	0.483	0.266
>8Mb_size										
mean±SD (aff)	57.1 ± 80.7	251.2 ± 83.4	311.8 ± 126.7	58.9 ± 86.9	114.2 ± 110.2	139.1 ± 103.1	105.2 ± 75.6	52.1 ± 58.7	53.7 ± 60.5	151 ± 150.7
mean±SD (unaff)	82.1 ± 139.2	315.7 ± 111.3	375.9 ± 101.6	60.3 ± 103.7	101.3 ± 86.9	219.9 ± 123.8	84.4 ± 91	78 ± 79.6	27.9 ± 40.4	193.7 ± 126.8
Wilcoxon p-value	0.625	0.220	0.159	0.756	1.000	0.082	0.285	0.215	0.568	0.260
>8Mb_count										
mean±SD (aff)	3.2 ± 4.4	15.6 ± 4.6	17.7 ± 6.4	3.3 ± 3.9	6.1 ± 5.1	8.4 ± 5.7	6.5 ± 4.2	3.9 ± 4.3	2.8 ± 2.7	8.9 ± 8.2
mean±SD (unaff)	4.2 ± 6.9	18.5 ± 6.2	19.4 ± 3.8	2.7 ± 3.7	6 ± 4.8	12.2 ± 6.6	5.2 ± 5.5	5.1 ± 4.4	2 ± 2.6	11.3 ± 7.3
Wilcoxon p-value	0.677	0.559	0.305	0.572	1.000	0.088	0.171	0.264	0.546	0.353

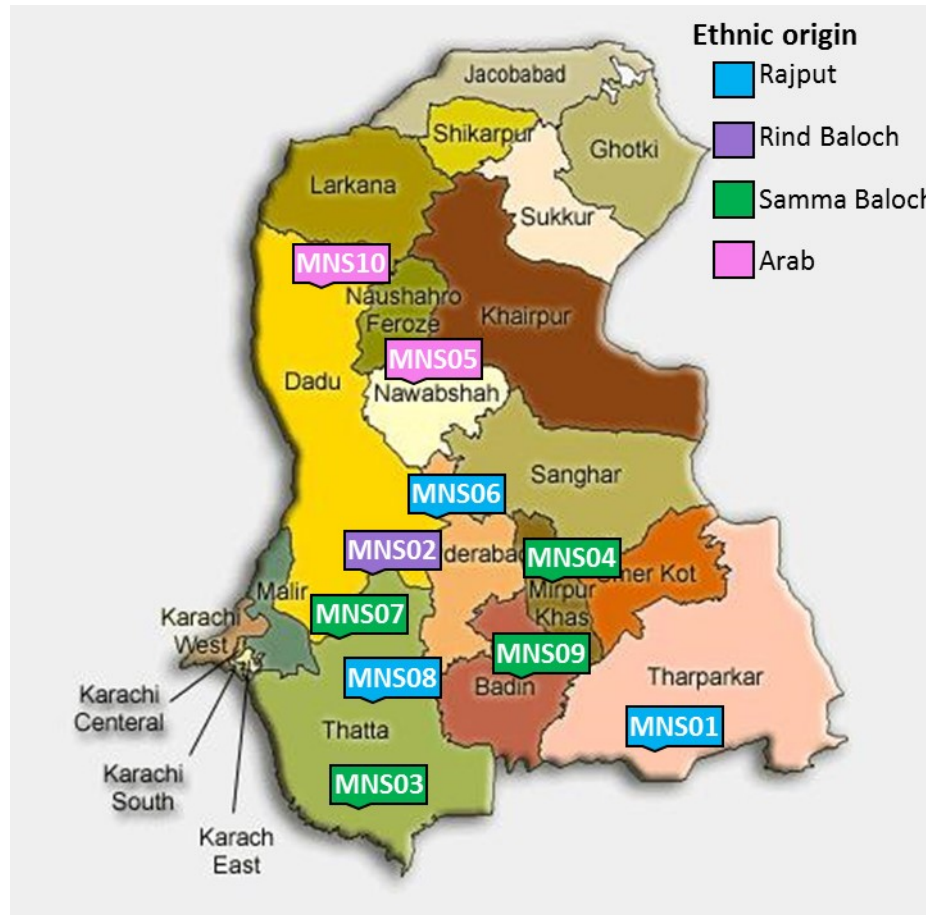
Aff: affected individuals; unaff: unaffected individuals; Wilcoxon p-value, an exact p-value from unpaired Wilcoxon Rank Sum Test; 4Mb_size, total size of ROHs larger than 4 megabases; 4Mb_count, number of ROHs larger than 4 megabases; 8Mb_size, total size of ROHs larger than 8 megabases; 8Mb_count, number of ROHs larger than 8 megabases.

Supplementary Figure 2a: Geographic and ethnic groups map of Pakistan

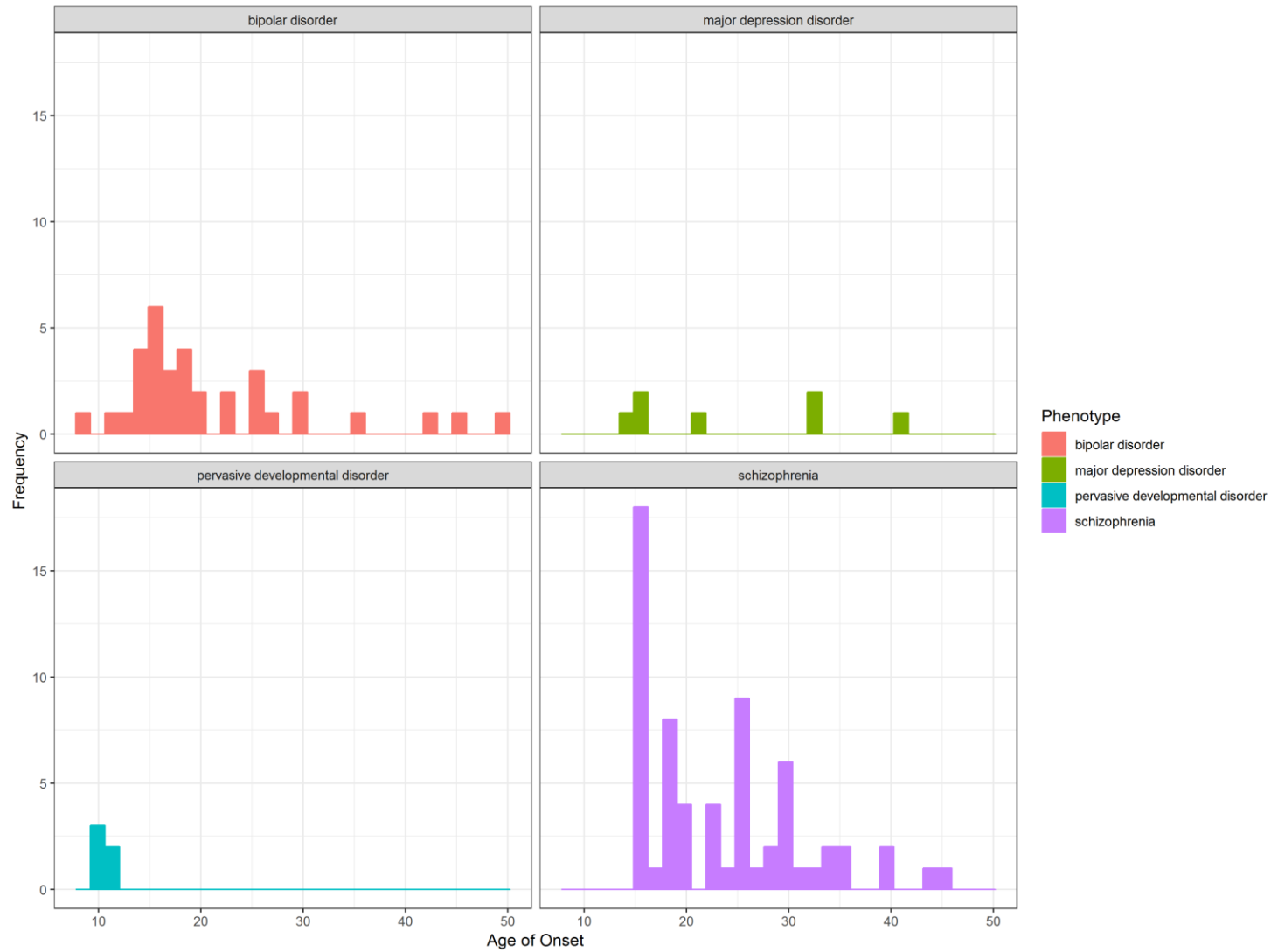


(downloaded from <http://www.lib.utexas.edu/maps/pakistan.html>, January 24, 2009, and https://commons.wikimedia.org/wiki/File:Pakistan_ethnic_map.svg, January 17, 2018)

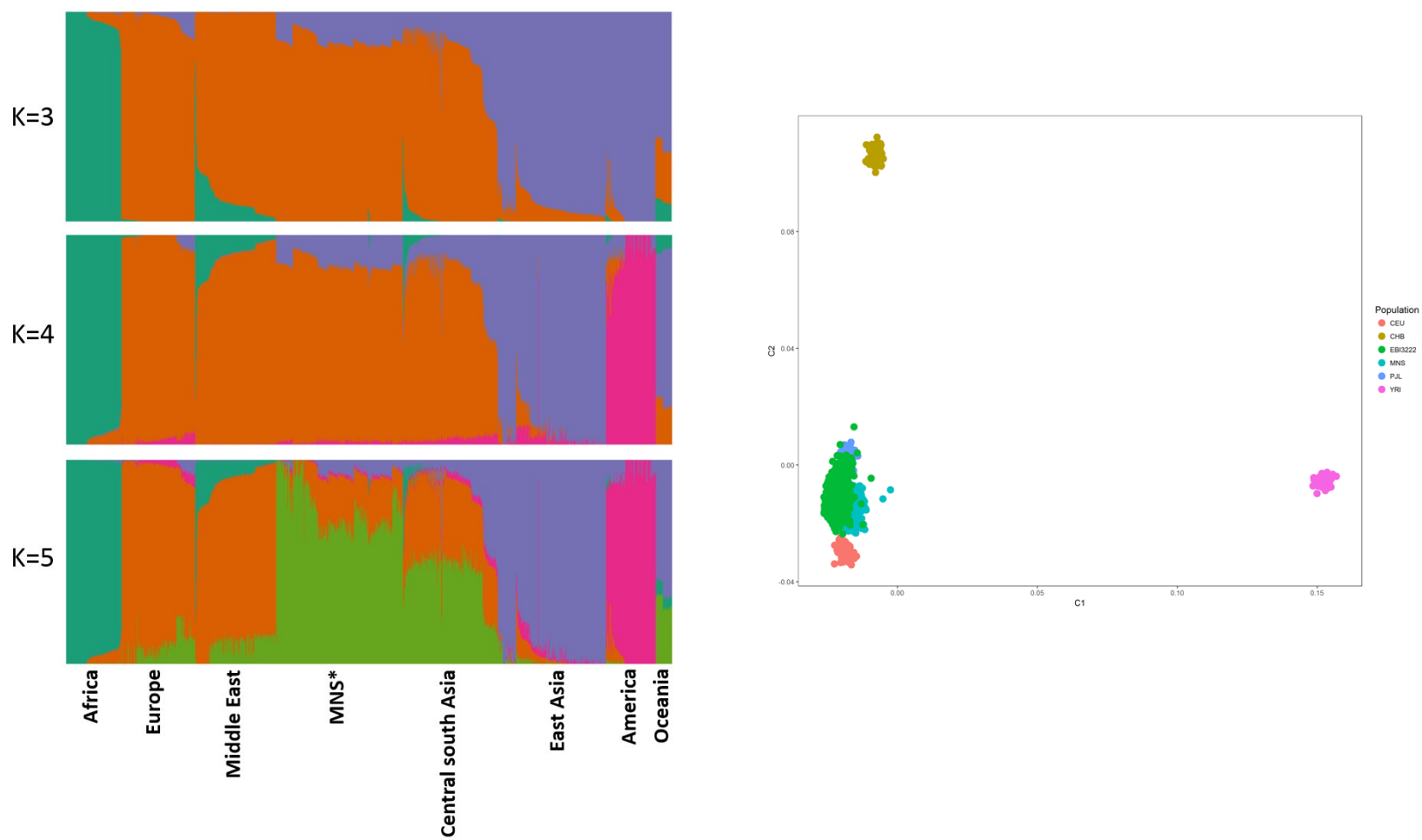
Supplementary figure 2b: Location and ethnic origin of MNS pedigrees in Sindh, Pakistan



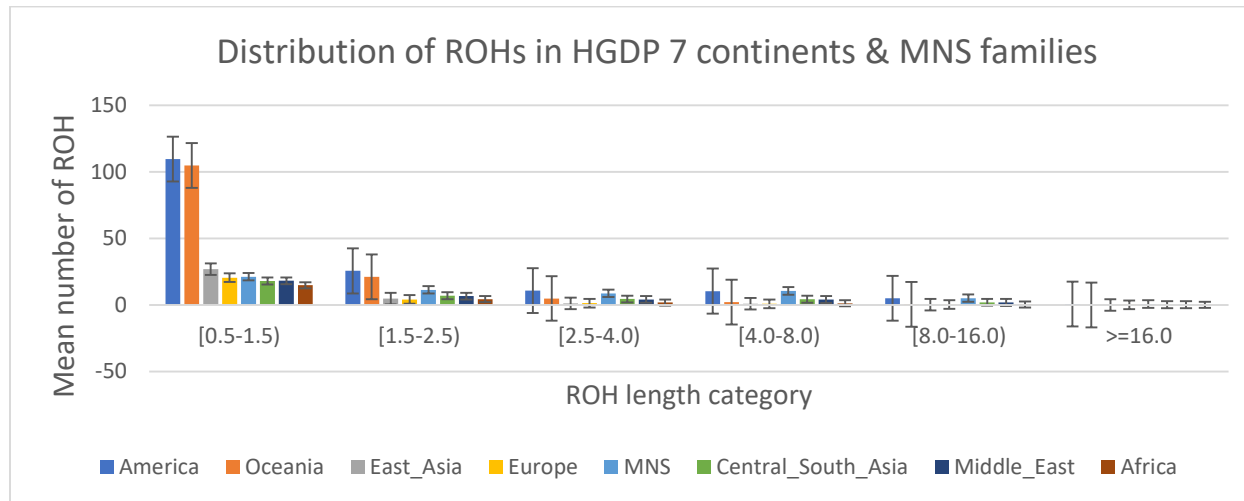
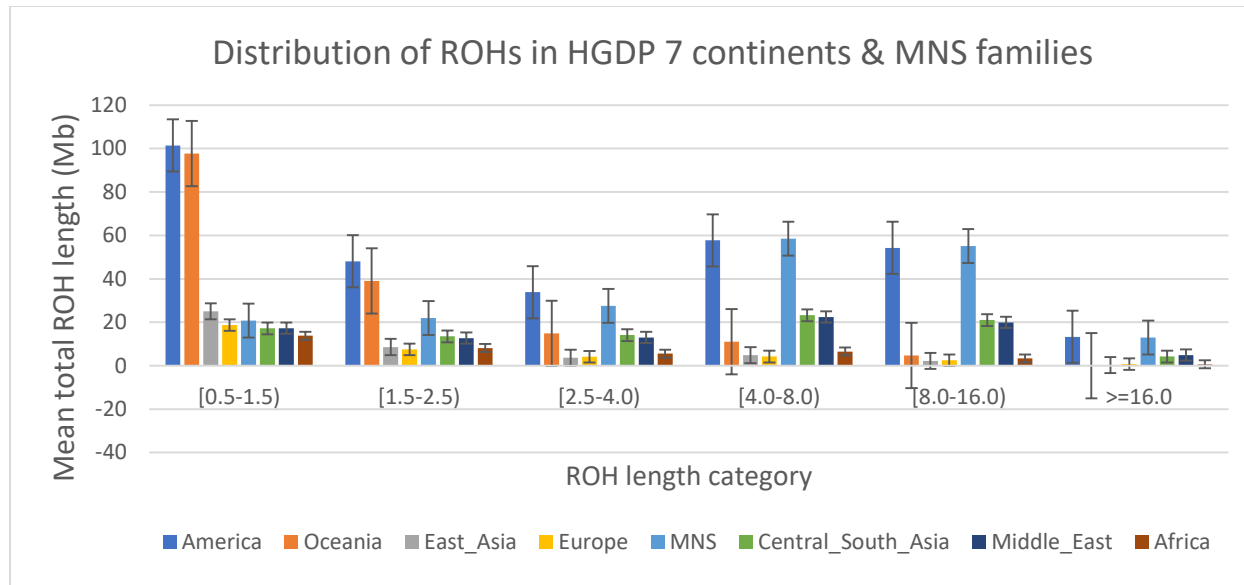
Supplementary Figure 3. Age of onset by phenotypes in MNS pedigrees



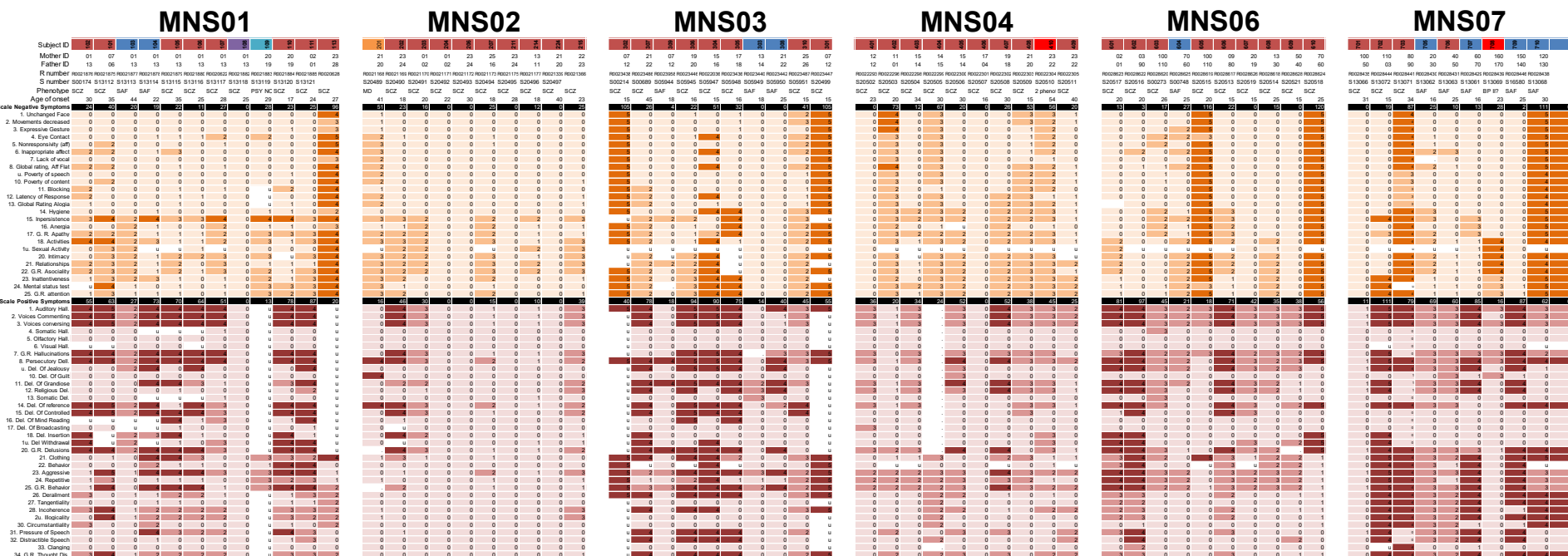
Supplementary Figure 4. Population admixture and population stratification of our pedigrees with other populations and control dataset



Supplementary Figure 5. Distribution of total length and number of ROH compared to other world populations

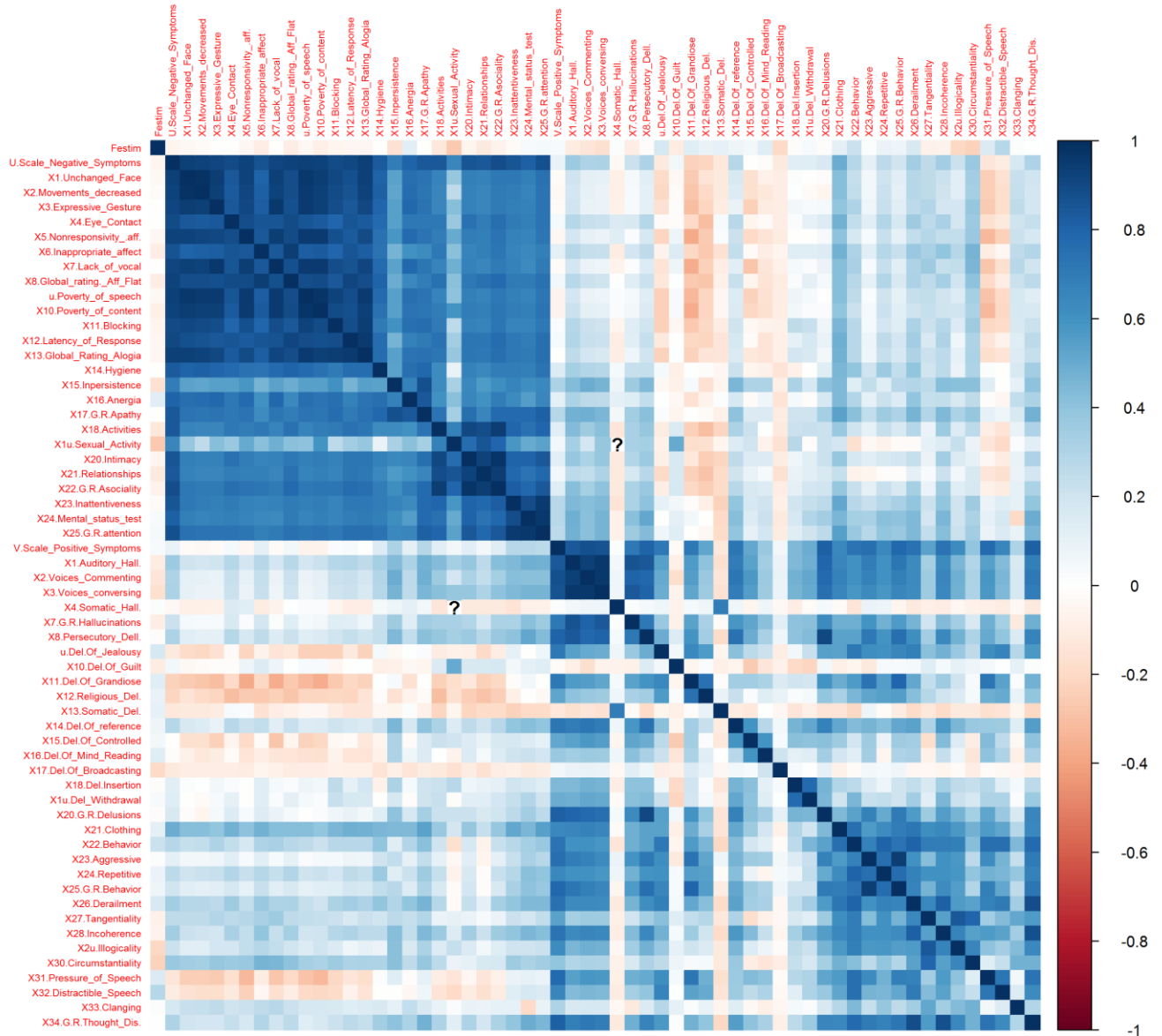


Supplementary Figure 6. Phenotypic profiling of the pedigrees aggregated with schizophrenia



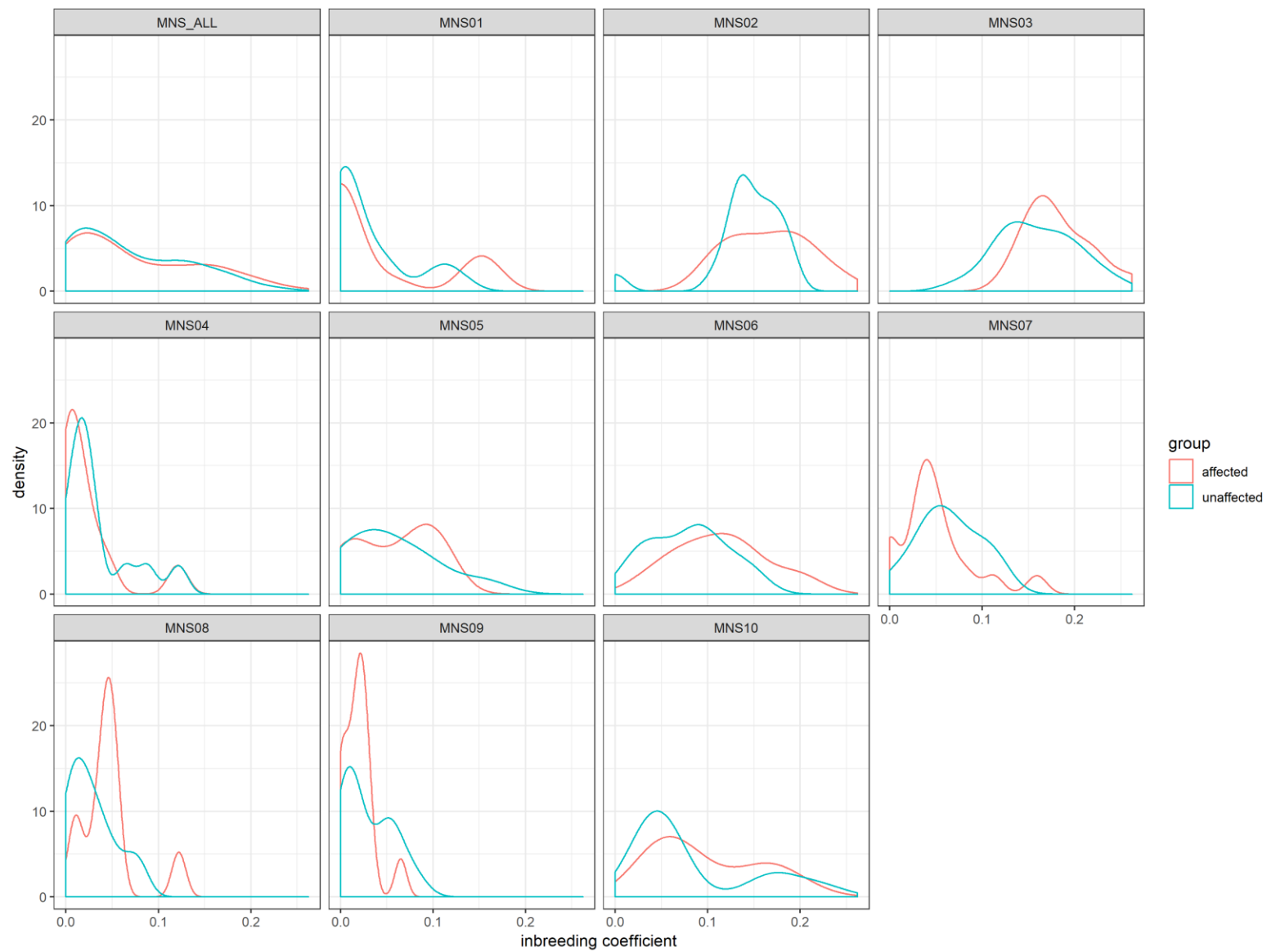
The color gradient depicts the scale of the symptoms, orange color for negative symptoms and red for positive symptoms.

Supplementary Figure 7. Correlation of inbreeding coefficient with positive and negative symptoms



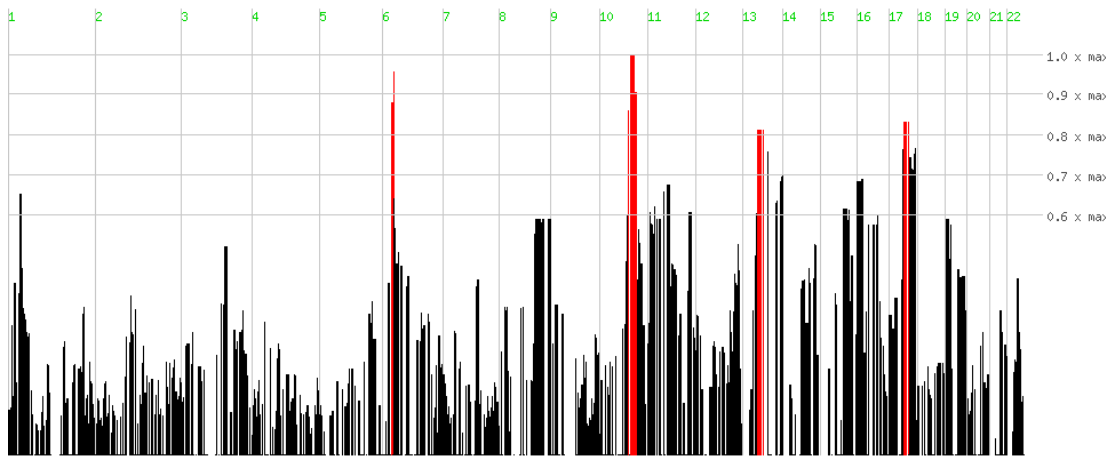
The leftmost side is inbreeding coefficient estimated by FSuite with genotyping data, the left part of the figure depicts negative symptoms, and the right part shows positive symptoms. As expected, there is a strong correlation between symptoms but no correlation between the inbreeding coefficient and the symptoms (first row). Correlation plot done using corrplot R package.

Supplementary figure 7. Density plot of inbreeding in all samples and in samples of each family

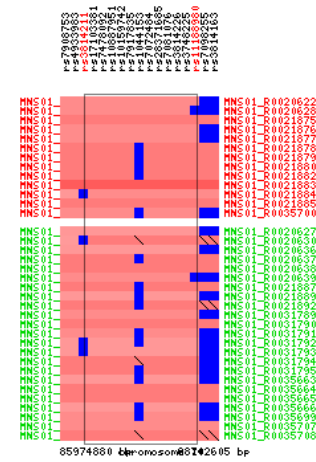


Supplementary figure 8. Example results of homozygosity mapping for the pedigrees

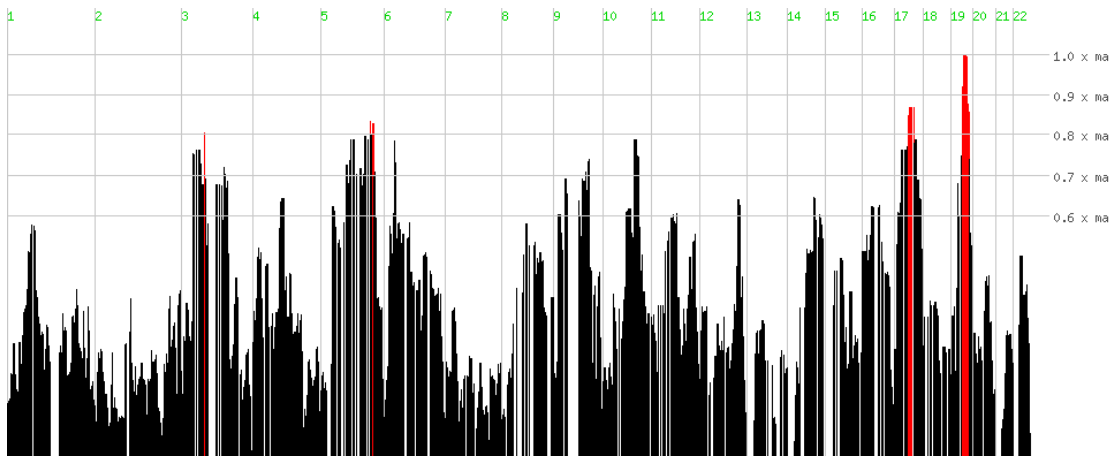
Genome-wide homozygosity in *MNS01*



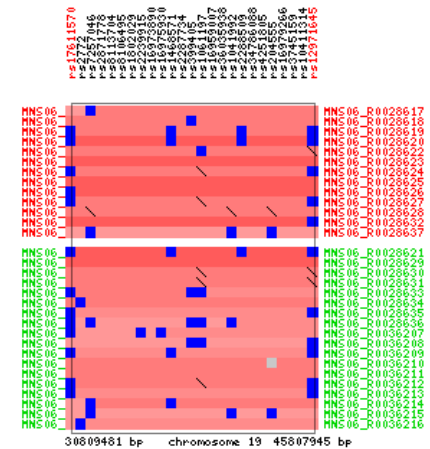
MNS01_chr10



Genome-wide homozygosity in *MNS06*



MNS06_chr19



The plots above demonstrate an example of HomozygosityMapper results in MNS01 and MNS06. Bar charts on the left show the genome-wide homozygosity of one family, and interesting regions are emphasized in red color when the homozygosity score of that region higher than 80% of the maximum score reached in this analysis. MNS01 and MNS06 have a maximum homozygosity score on chromosome 10 and chromosome 19 respectively, and also indicate several other interesting regions. A closer inspection of underlying genotypes is shown on the right. The markers are placed on the x-axis while the samples are on the y-axis, with the affected family members on top in red IDs and with the unaffected family members on the bottom in green IDs. The genotypes are color-coded: unknown genotypes are displayed as grey boxes, heterozygous genotypes as blue boxes and stretches of homozygosity as red bars. The saturation of the red colour reflects length of the homozygous block. A single heterozygous marker (possibly a genotyping error) within the homozygous region is ignored by HomozygosityMapper. As in the plot, the homozygous stretch is shared both by the affected and unaffected family members, which rules out it to be a promising causative region.

Appendix 2: Supplementary material for Chapter 3.2

Supplementary Text

Calling CNVs from SNPchip genotyping data

1. Preparation of intensity files for CNV calling

In total, 367 Pakistani samples were processed, including 179 samples genotyped with OmniExpress v12 SNPchip and 188 samples genotyped with OmniExpress v24. We followed Illumina's technical notes to create a custom cluster file for each SNPchip, in order to gain a more representative reference for the calculations of the normalized intensity values LogR Ratio (LRR) and normalized B allele frequency (BAF) in copy number analysis²⁶⁹. 4 out of the 179 samples were excluded from clustering because of low call rates (<0.98); 3 out of the 188 samples were examined due to very large CNV calls or that they have outlier LogRDev or BAlleleDev in the CN Metrics report. The autosomal SNPs were clustered for all samples, the X chromosome SNPs were clustered on female samples and Y chromosome SNPs on male samples. The SNP statistics was updated and saved after clustering, and used for next step of calling CNVs.

2. Calling CNVs with three algorithms

For CNVpartition (plug-in in GenomeStudio), two sets of parameters were used: confidence threshold 0 or 35. The number of raw calls of using "0" was higher but the raw calls included all the calls made by using stringent "35".

For pennCNV, a population B allele frequency (PFB) file was generated separately using 175 samples out of 179 samples of OmniExpress v12 and 188 samples out of 188 samples of OmniExpress v24. We also tried to a combined PFB file with 363 out of 367 samples, the results of final CNV calls were largely overlapped. Other parameters of CNV calling followed the default setting of the software. X chromosome were treated specifically by providing gender information.

As for QuantiSNP, we followed the default settings and used a configuration files (levels-hd.dat and params.dat), by the command: `run_quantisnp2.sh /home/apps/Logiciels/MATLAB/MATLAB_Compiler_Runtime/v7.9/v79 --outdir`

```
quantisnp/ --levels quantisnp/levels-hd.dat --config quantisnp/params.dat --sampleid  
$variable --input-files quantisnp/quantisnp_$.variable.
```

Calling CNVs from whole-exome sequencing data

1. Preparing coverage files for CNV calling

243 DNA samples was captured by the Agilent SureSelect Human All Exon 38M, 50M, v4 and v5 kit (Agilent Technologies, Inc.). WES was performed using the Illumina HiSeq 2000 platform (paired-end, 100 cycles) at the Genome Quebec Innovation Centre (Montréal, Québec, Canada) and at the Macrogen Korean facility (Macrogen Inc.) in separate batches. The raw fastq files were aligned to human reference genome (hg19) with Burrows-Wheeler Aligner (BWA)²¹⁴. Duplicates were removed with the MarkDuplicates function in Picard tools. Genome Analysis Toolkit (GATK v3.5)²¹⁵ was used to process the bam files. Depth of coverage was performed using the GATK DepthOfCoverage tool.

2. Calling exome CNVs with XHMM and CoNIFER

Since our samples were captured with different capture kit, we separated them by batch and processed with other samples in our database with the same capture kit to obtain a better normalization.

XHMM includes the following process: 1) calculate DepthOfCoverage by sequencing target intervals with GATK; 2) combine GATK DepthOfCoverage outputs for multiple samples captured by the same target intervals; 3) create a list of targets with extreme GC content by GATK or with low complexity by PLINK/seq; 4) filter samples and targets by a range of target size, mean target read depth and mean sample read depth as default; 5) mean-center the targets, run principle component analysis (PCA) on the targets and normalize the mean-centered data using PCA information (using PVE_mean method to remove principle components which individually explain more variance than 0.7 times the average); 6) filter and z-score center the PCA-normalized data by sample; 7) filter the original read-depth data to be the same as filtered, normalized data.

As for CoNIFER, the inflection point for the four capture kits fell on 4 or 5, we chose –SVD 7 for all of them consistently. The data suggested the smaller SVD number is

used, the more CNVs CoNIFER called, and a higher rate of intersecting calls with XHMM software.

Familial segregation of CNVs

We only included CNV calls from 15 families for segregation, and excluded 34 Pakistani population controls since they were all males from different families and they were not sequenced. After merging the CNVs with CNVision, we exclude CNVs shorter than 1kb, and CNVs called “duplication” by one algorithm but “deletion” by other algorithms, and vice versa. The example of using the sv-segregation software is shown as follows:

```
python segregation.py \  
-i merged_PN_CN_QT_1kb.seg.input:popsv:genotyping \  
-r chip \  
-o chip_merged_segregation.output \  
-cfg snpchip.cfg \  
-min 1 \  
-max 100000000 \  
-p PAK366.ped
```

In snpchip.cfg file, one has to assign the minimal and maximal size of the CNVs, the percentage of overlapping region between family members, and percentage of overlap with other public databases, and output CNV found in affected, unaffected or all samples. In the statistics in the main text and shortlist for our families, we focused on variants found in affected samples and search segregation variants. We used CNVs found in all samples when we want to compare the results of defining likely positive CNVs as shared by 2 and 3 family members, shown in **Supplementary Table 4 and 5**.

Supplementary Table 1: summary of samples used for SNPchip genotyping

<i>Ped</i>	<i>total</i>	<i>cases</i>	<i>controls</i>	<i>female</i>	<i>male</i>
ATM01	12	2	10	4	8
ATM02	7	2	5	3	4
ATM03	15	3	12	7	8
ATM04	8	3	5	2	6
ATM05	16	3	13	7	9
MNS01	36	12	24	19	17
MNS02	26	10	16	15	11
MNS03	28	12	16	10	18
MNS04	23	11	12	6	17
MNS05	22	13	9	15	7
MNS06	31	12	19	11	20
MNS07	34	17	17	18	16
MNS08	26	10	16	9	17
MNS09	26	14	12	15	11
MNS10	23	12	11	10	13
PAKcontrols	34	0	34	0	34
total	367	136	231	151	216

ATM01-05 are small pedigrees with probands affected with autism; MNS01-10 are large pedigrees with family members affected with schizophrenia and bipolar disorder; PAKcontrols are population controls collected in different families from the same geographic region.

Supplementary Table 2: summary of samples used for whole-exome sequencing

<i>Ped</i>	<i>total</i>	<i>cases</i>	<i>controls</i>	<i>female</i>	<i>male</i>
ATM01	2	2	0	0	2
ATM02	2	2	0	0	2
ATM03	3	3	0	0	3
ATM04	3	3	0	0	3
ATM05	3	3	0	0	3
MNS01	21	12	9	9	12
MNS02	20	10	10	12	8
MNS03	23	12	11	7	16
MNS04	20	11	9	5	15
MNS05	19	13	6	13	6
MNS06	28	12	16	11	17
MNS07	30	17	13	16	14
MNS08	22	10	12	8	14
MNS09	26	14	12	15	11
MNS10	21	12	9	9	12
PAKcontrols	0	0	0	0	0
total	243	136	107	105	138

ATM01-05 are small pedigrees with probands affected with autism; MNS01-10 are large pedigrees with family members affected with schizophrenia and bipolar disorder; PAKcontrols are population controls collected in different families from the same geographic region.

Supplementary Table 3: Number and percentage of likely false positive CNVs and likely true positive CNVs for deletions and duplications, and the estimation of the sensitivity and the specificity for each software

<i>Algorithm</i>	<i>Likely false positive</i>		<i>Likely true positive</i>		<i>Sensitivity</i>	<i>Specificity</i>
<i>deletion</i>						
PennCNV	1255	18.49%	5533	81.51%	0.32	0.81
QuantiSNP	4567	26.08%	12944	73.92%	0.76	0.32
CNVpartition	2442	29.06%	5961	70.94%	0.35	0.64
XHMM	416	28.30%	1054	71.70%	0.06	0.94
CoNIFER	121	21.96%	430	78.04%	0.03	0.98
<i>duplication</i>						
PennCNV	319	18.54%	1402	81.46%	0.22	0.92
QuantiSNP	2903	38.39%	4659	61.61%	0.73	0.31
CNVpartition	667	31.76%	1433	68.24%	0.22	0.84
XHMM	606	35.23%	1114	64.77%	0.17	0.86
CoNIFER	340	30.71%	767	69.29%	0.12	0.92

Note: Likely false positive CNVs are singleton CNVs; likely true positive CNVs are defined as segregating CNVs (in ≥ 2 family members).

Supplementary Table 4: Number and percentage of likely false positive CNVs and likely true positive CNVs in autosomal chromosomes and estimation of the sensitivity and the specificity for each software

<i>Algorithm</i>	<i>False positive</i>		<i>True positive</i>		<i>Sensitivity</i>	<i>Specificity</i>
<i>Autosomal CNVs</i>						
PennCNV	3029	36.29%	5317	63.71%	0.34	0.86
QuantiSNP	15030	55.36%	12121	44.64%	0.77	0.31
CNVpartition	5023	78.23%	1398	21.77%	0.09	0.77
XHMM	1948	91.67%	177	8.33%	0.01	0.91
CoNIFER	1352	97.69%	32	2.31%	0.00	0.94
<i>CNVs on chromosome X</i>						
PennCNV	45	12.71%	309	87.29%	0.27	0.91
QuantiSNP	289	22.70%	984	77.30%	0.86	0.42
CNVpartition	109	30.62%	247	69.38%	0.21	0.78
XHMM	2	22.22%	7	77.78%	0.01	1.00
CoNIFER	2	12.50%	14	87.50%	0.01	-

Likely false positive CNVs are singleton CNVs; likely true positive CNVs are defined as segregating CNVs (**in ≥ 2 family members**). The number is bigger than Table VII and Table VIII in the main text, since the analysis here was based on the segregation of CNVs found in all samples and the latter is based on CNVs found in affected samples.

Supplementary Table 5: Number and percentage of likely false positive CNVs and likely true positive CNVs in autosomal chromosomes and estimation of the sensitivity and the specificity for each software

<i>Algorithm</i>	<i>False positive</i>		<i>True positive</i>		<i>Sensitivity</i>	<i>Specificity</i>
<i>Autosomal CNVs</i>						
PennCNV	3657	43.82%	4689	56.18%	0.36	0.85
QuantiSNP	17602	64.83%	9549	35.17%	0.72	0.28
CNVpartition	6110	81.38%	1398	18.62%	0.11	0.75
XHMM	1426	88.96%	177	11.04%	0.01	0.94
CoNIFER	1032	96.99%	32	3.01%	0.00	0.96
<i>CNVs on chromosome X</i>						
PennCNV	64	18.08%	290	81.92%	0.26	0.88
QuantiSNP	332	26.08%	941	73.92%	0.86	0.39
CNVpartition	138	38.76%	218	61.24%	0.20	0.75
XHMM	2	22.22%	7	77.78%	0.01	1.00
CoNIFER	2	12.50%	14	87.50%	0.01	-

Likely false positive CNVs are singleton CNVs; likely true positive CNVs are defined as segregating CNVs (**in ≥ 3 family members**). The number is bigger than Table VII and Table VIII in the main text, since the analysis here was based on the segregation of CNVs found in all samples and the latter is based on CNVs found in affected samples.

Supplementary table 6: segregation pattern of CNVs with and without filtering parameters in 15 families

	<i>MNS01</i>	<i>MNS02</i>	<i>MNS03</i>	<i>MNS04</i>	<i>MNS05</i>	<i>MNS06</i>	<i>MNS07</i>	<i>MNS08</i>	<i>MNS09</i>	<i>MNS10</i>	<i>ATM01</i>	<i>ATM02</i>	<i>ATM03</i>	<i>ATM04</i>	<i>ATM05</i>
# affected	13	10	12	11	13	13	17	10	14	12	5	3	3	3	5
unfiltered CNVs from SNPchip, segregating in # affected															
1 affected	554	1829	660	719	430	1778	1348	355	1044	512	99	170	139	622	212
>=2 affected	87	318	55	49	97	94	215	48	216	137	38	14	18	63	36
>=5 affected	16	10	7	1	15	6	9	2	13	18					
all affected	0	0	0	0	0	0	0	0	0	0	2	1	1	7	0
unfiltered CNVs from WES, segregating in # affected															
1 affected	42	141	133	169	59	202	243	120	91	126	48	39	31	15	39
>=2 affected	18	86	42	44	25	64	60	19	51	50	8	4	11	7	10
>=5 affected	0	12	4	1	4	1	4	1	9	2					
all affected	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Filtered CNVs from SNPchip: scenario 1*															
1 affected	57	279	59	162	59	383	116	42	86	44	17	18	22	58	16
>=2 affected	12	42	18	15	31	26	34	15	33	27	14	7	8	14	16
>=5 affected	2	3	3	0	5	0	4	1	4	7					
all affected	0	0	0	0	0	0	0	0	0	0	2	1	2	4	0

* scenario 1: CNV size >= 20kb, and detected by >=2 algorithms;

* scenario 3: CNV size >= 20kb, number of probes >= 5, confidence score >= 5 and detected by >= 2 algorithms;

* scenario 4: CNV size >= 20kb, number of probes >= 5, confidence score >= 5 and detected by >= 3 algorithms.

Supplementary table 6: segregation pattern of CNVs with and without filtering parameters in 15 families, continued

	<i>MNS01</i>	<i>MNS02</i>	<i>MNS03</i>	<i>MNS04</i>	<i>MNS05</i>	<i>MNS06</i>	<i>MNS07</i>	<i>MNS08</i>	<i>MNS09</i>	<i>MNS10</i>	<i>ATM01</i>	<i>ATM02</i>	<i>ATM03</i>	<i>ATM04</i>	<i>ATM05</i>
# affected	13	10	12	11	13	13	17	10	14	12	5	3	3	3	5
Filtered CNVs from WES: scenario 1*															
1 affected	6	16	13	26	15	31	26	20	14	18	8	6	5	4	3
>=2 affected	5	5	15	10	6	16	11	6	9	12	3	1	1	1	1
>=5 affected	0	0	1	0	1	1	2	1	0	0					
all affected	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Filtered CNVs from SNPchip: scenario 3*															
1 affected	43	248	47	149	48	365	90	35	72	27	15	14	20	56	13
>=2 affected	10	39	13	13	29	22	30	13	28	26	14	7	6	14	13
>=5 affected	1	3	3	0	5	0	5	1	4	8					
all affected	0	0	0	0	0	0	0	0	0	0	2	1	1	4	0
Filtered CNVs from SNPchip: scenario 4*															
1 affected	13	18	17	44	32	47	39	25	25	20	12	9	11	13	9
>=2 affected	8	8	9	11	26	13	24	13	16	22	8	2	4	7	10
>=5 affected	2	2	3	0	3	0	4	0	3	6					
all affected	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0

* scenario 1: CNV size >= 20kb, and detected by >=2 algorithms;

* scenario 3: CNV size >= 20kb, number of probes >= 5, confidence score >= 5 and detected by >= 2 algorithms;

* scenario 4: CNV size >= 20kb, number of probes >= 5, confidence score >= 5 and detected by >= 3 algorithms.

Supplementary table 7: A short list of candidate CNVs for small pedigrees aggregated with autism.

family	type ⁺	A	UA	chr	start	end	size	Gene	Func.refGene	source	algos*	linked to Autism*
ATM01	Del	2	0	1	1341167	1444657	103490	<i>ANKRD65</i>	exonic	WES	CO, XH	1p36.33
ATM01	Dup	2	0	7	66648048	69064982	2416934	<i>AUTS2</i>	exonic	WES	CO, XH	7q11.21-q11.22
ATM01	Dup	2	0	11	60899231	61017332	118101	<i>VPS37C</i>	exonic	WES	CO	11q12.2
ATM01	Del	2	0	17	43545526	43596393	50867	<i>PLEKHM1</i>	exonic	WES	CO	17q21.31
ATM02	Del	2	0	8	104825219	105197239	372020	<i>RIMS2</i>	exonic	SNPchip	CN	8q22.3
ATM02	Del	2	0	16	30199185	30234643	35458	<i>BOLA2</i>	exonic	WES	CO	16p11.2
ATM04	Dup	2	0	8	75689462	75749982	60520	<i>PII5</i>	exonic	SNPchip	QT	8q21.11
ATM04	Dup	2	0	10	37467445	37583900	116455	<i>ANKRD30A</i>	exonic	SNPchip	QT	10p11.21
ATM04	Dup	2	1	11	100036291	100072718	36427	<i>CNTN5</i>	exonic	SNPchip	PN, QT	11q22.1
ATM04	Dup	2	0	12	9993452	10024007	30555	<i>CLEC2B, KLRF1</i>	exonic	SNPchip	CN, QT	12p13.31
ATM04	Dup	2	0	12	83354429	83384628	30199	<i>TMTC2</i>	exonic	SNPchip	PN, QT	12q21.31
ATM04	Dup	2	0	13	77803247	77854638	51391	<i>MYCBP2</i>	exonic	SNPchip	QT	13q22.3
ATM04	Del	2	0	17	5337000	5365910	28910	<i>C1QBP, DHX33</i>	exonic	WES	CO, XH	17p13.2
ATM04	Del	2	0	21	44835301	44870150	34849	<i>SIK1</i>	exonic	SNPchip	PN	21q22.3
ATM05	Dup	2	0	2	130951339	130987236	35897	<i>TUBA3E</i>	exonic	WES	XH	2q21.1
ATM05	Del	2	0	2	132200865	132240453	39588	<i>TUBA3D</i>	exonic	WES	XH	2q21.1
ATM05	Del	2	0	6	73975404	74003177	27773	<i>KHDC1</i>	exonic	SNPchip	CN, QT	6q13
ATM05	Dup	2	1	12	80731040	80775056	44016	<i>OTOGL</i>	exonic	SNPchip	QT	12q21.31
ATM05	Dup	2	0	14	20002199	20444740	442541	<i>OR11H2</i>	exonic	WES	CO	14q11.2
ATM05	Dup	2	0	15	20833516	21052456	218940	<i>POTEB</i>	exonic	WES	CO	15q11.2

This table contains a preliminary list of prioritized CNVs in small pedigrees with autism affected probands. We kept CNVs present in at least 2 affected individuals and present more in affected than unaffected family members. We excluded those CNVs are present in 1KG and DGV database. At last, we selected the ones which have been previously linked to autism. Legend: ⁺type: Del, deletion; Dup, duplication. A, number of affected carriers in the family, UA, number of unaffected carriers in the family. *CO, CoNIFER; XH, XHMM;

CN, CNVpartition; PN, PennCNV; QT, QuantiSNP; [€] These Cytoband regions have been previously associated to Autism according to AutDB Home (<http://autism.mindspec.org/autdb/Welcome.do>).

Supplementary table 8: A short list of candidate CNVs for large pedigrees aggregated with schizophrenia and bipolar disorder

family	type	A	UA	chr	start	end	Size	Gene	Func.refGene	source	algos*	to SCZ/BP CNV	to SCZ GWAS	to BD GWAS
MNS02	Del	6	5	15	24385350	24472002	86652	<i>PWRN2</i>	ncRNA_exonic	SNPchip	CN, PN, QT	15q11.2		15q11.2
MNS05	Del	6	2	15	24385350	24472002	86652	<i>PWRN2</i>	ncRNA_exonic	SNPchip	CN, PN, QT	15q11.2		15q11.2
MNS03	Dup	5	4	16	16633361	16682080	48719	<i>NPIPA8</i>	intergenic	SNPchip	QT	16p13.11		
MNS05	Dup	6	5	16	66967835	67070714	102879	<i>CBFB</i>	exonic	WES	CO, XH	16q22.1	16q22.1	
MNS03	Dup	6	0	1	65509	777481	711972	<i>OR4F16</i>	exonic	WES	CO, XH	1p36.33	1p36.33	
MNS03	Del	5	1	1	1634935	1669905	34970	<i>CDK11A</i>	exonic	WES	CO, XH	1p36.33	1p36.33	
MNS10	Del	5	5	22	18626900	18629153	2253	<i>TUBA8,USP18</i>	intergenic	SNPchip	CN	22q11.21	22q11.21	
MNS01	Del	5	1	7	62154874	62159926	5052	<i>NONE,ZNF733P</i>	intergenic	SNPchip	PN, QT	7q11.21		
MNS07	Dup	6	2	2	1.79E+08	1.79E+08	14967	<i>PRKRA</i>	exonic	WES	CO, XH		2q31.2	2q31.2
MNS10	Del	7	3	14	80082435	80115560	33125	<i>NRXN3</i>	intronic	SNPchip	CN		14q31.1	14q31.1
MNS05	Del	8	0	20	25470505	25479064	8559	<i>NINL</i>	exonic	WES	XH		20p11.21	20p11.21
MNS01	Del	7	1	6	77020141	77024665	4524	<i>IMPG1,HTR1B</i>	intergenic	SNPchip	CN, PN, QT		6q14.1	
MNS01	Del	6	0	12	70874726	70877258	2532	<i>KCNMB4,PTPRB</i>	intergenic	SNPchip	QT		12q15	12q15
MNS07	Del	6	0	7	6838829	6864382	25553	<i>CCZ1B</i>	exonic	WES	XH			
MNS09	Del	6	0	16	55844456	55854444	9988	<i>CESI</i>	exonic	WES	XH		16q12.2	16q12.2
MNS01	Del	5	0	3	6651929	6654060	2131	<i>MIR4790,GRM7-AS3</i>	intergenic	SNPchip	QT		3p26.1	3p26.1
MNS02	Del	5	0	16	63574341	63582751	8410	<i>CDH8,CDH11</i>	intergenic	SNPchip	QT		16q21	
MNS05	Del	6	1	2	90010895	90240473	229578	<i>MIR4436A,LOC654342</i>	intergenic	SNPchip	CN, PN, QT		2p11.2	2p11.2
MNS05	Dup	5	0	14	19255726	19328549	72823	<i>NONE,OR11H12</i>	intergenic	SNPchip	QT		14q11.2	14q11.2
MNS09	Del	5	0	9	9796116	9805496	9380	<i>PTPRD</i>	intronic	SNPchip	QT		9p23	9p23
MNS02	Del	5	0	2	67629928	67637212	7284	<i>ETAA1</i>	exonic	WES	XH		2p14	2p14
MNS02	Del	5	0	6	1.17E+08	1.17E+08	6568	<i>KPNA5</i>	exonic	WES	XH		6q22.1	
MNS02	Del	5	0	11	89059898	89073392	13494	<i>NOX4</i>	exonic	WES	XH			
MNS05	Del	5	0	19	49474159	49496516	22357	<i>GYS1</i>	exonic	WES	CO, XH		19q13.33	19q13.33
MNS06	Dup	8	3	X	48054712	48248937	194225	<i>SSX1</i>	exonic	WES	CO, XH			
MNS09	Del	5	0	1	16972036	16974846	2810	<i>MSTIP2</i>	ncRNA_exonic	WES	XH		1p36.13	1p36.13
MNS09	Dup	5	0	8	39311550	39349526	37976	<i>ADAM3A</i>	ncRNA_exonic	WES	XH			

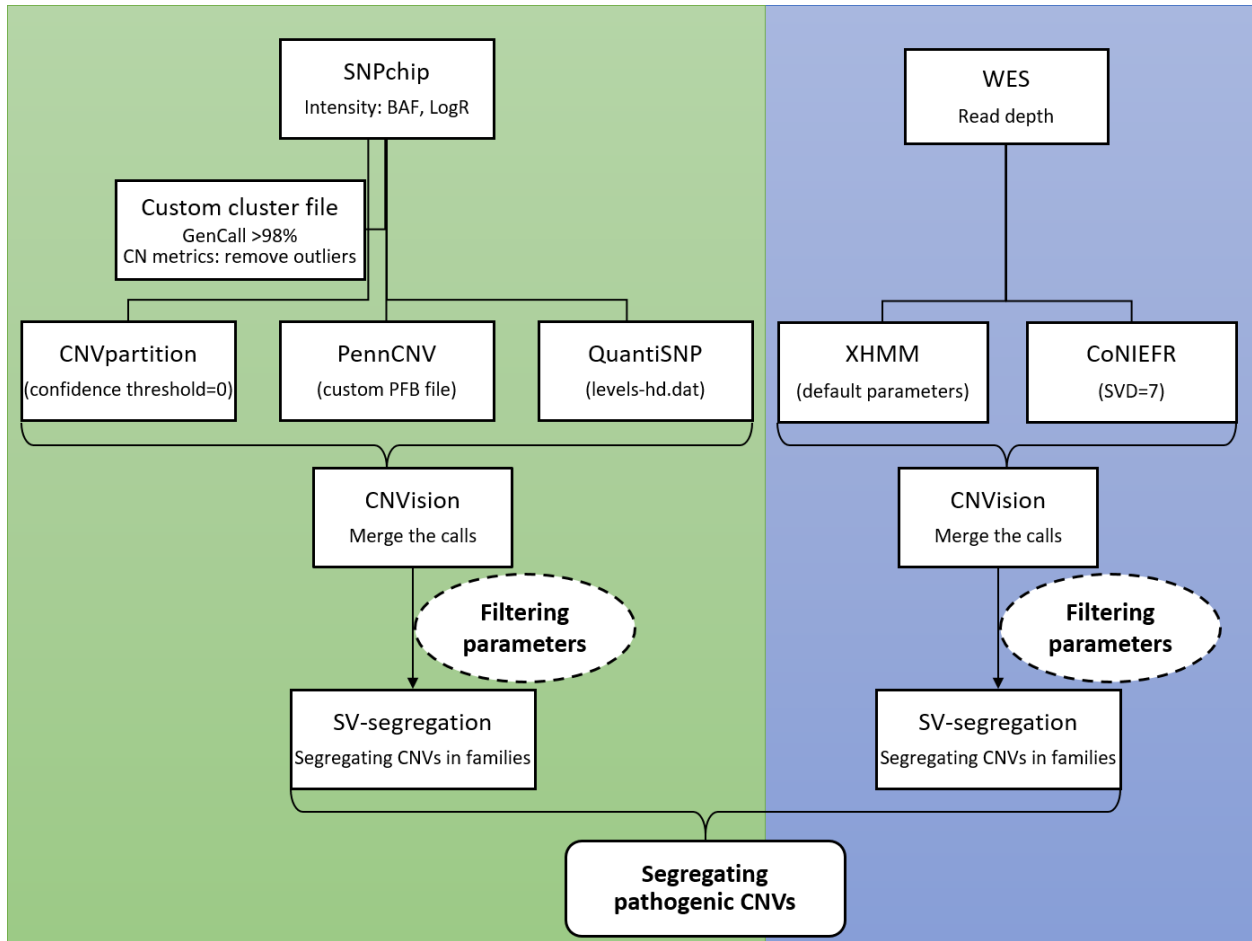
Supplementary Table 8: A short list of candidate CNVs for large pedigrees aggregated with schizophrenia and bipolar disorder, continued

family	type	A	UA	chr	start	end	Size	Gene	Func.refGene	source	algos*	to SCZ/BP CNV	to SCZ GWAS	to BD GWAS
MNS09	Dup	5	0	9	1.36E+08	1.36E+08	7441	<i>CEL</i>	exonic	WES	XH		9q34.2	
MNS09	Dup	6	1	19	54783186	54784774	1588	<i>LILRB2</i>	exonic	WES	XH		19q13.42	19q13.42
MNS10	Dup	5	0	5	1.39E+08	1.39E+08	3752	<i>MATR3</i>	exonic	WES	XH		5q31.2	
MNS01	Del	6	2	3	1.76E+08	1.76E+08	14260	<i>NAALADL2,MIR7977</i>	intergenic	SNPchip	CN, PN, QT		3q26.32	3q26.32
MNS08	Del	5	1	13	54812498	54816326	3828	<i>LINC00458,MIR1297</i>	intergenic	SNPchip	QT		13q14.3	13q14.3
MNS10	Del	8	4	12	70874726	70877258	2532	<i>KCNMB4,PTPRB</i>	intergenic	SNPchip	QT		12q15	12q15
MNS02	Del	5	1	5	64766600	64814448	47848	<i>ADAMTS6,CENPK</i>	exonic	WES	XH		5q12.3	5q12.3
MNS09	Dup	5	1	12	9586597	9590154	3557	<i>DDX12P</i>	ncRNA_exonic	WES	XH		12p13.31	12p13.31
MNS02	Del	6	3	5	1.43E+08	1.43E+08	11840	<i>NR3C1</i>	intronic	SNPchip	CN, PN, QT		5q31.3	5q31.3
MNS05	Dup	5	2	7	33658726	33690732	32006	<i>BBS9,BMPER</i>	intergenic	SNPchip	CN, PN, QT			
MNS06	Dup	8	5	X	48095238	48205223	109985	<i>SSX1</i>	exonic	SNPchip	CN, QT			
MNS07	Dup	5	2	11	51581931	51591253	9322	<i>OR4C46,NONE</i>	intergenic	SNPchip	QT			
MNS09	Del	5	2	3	1.06E+08	1.06E+08	13681	<i>CBLB,LINC00882</i>	intergenic	SNPchip	QT			
MNS09	Dup	5	2	6	31964330	32013891	49561	<i>C4A</i>	exonic	SNPchip	QT		6p21.33	6p21.33
MNS09	Del	6	3	9	1.1E+08	1.1E+08	2629	<i>KLF4,ACTL7B</i>	intergenic	SNPchip	CN, PN, QT		9q31.2	
MNS10	Del	5	2	8	1.16E+08	1.16E+08	1170	<i>CSMD3,TRPS1</i>	intergenic	SNPchip	QT		8q23.3	
MNS02	Del	5	2	2	1.41E+08	1.42E+08	55109	<i>LRP1B</i>	exonic	WES	XH		2q22.1	
MNS02	Del	5	2	11	1.03E+08	1.03E+08	11927	<i>DYNC2H1</i>	exonic	WES	XH		11q22.3	11q22.3
MNS02	Del	5	2	12	82824636	82872861	48225	<i>METTL25</i>	exonic	WES	XH		12q21.31	
MNS02	Del	5	2	17	5047964	5050491	2527	<i>USP6</i>	exonic	WES	XH		17p13.2	17p13.2
MNS07	Del	6	3	14	99182585	99183589	1004	<i>C14orf177</i>	exonic	WES	XH		14q32.2	

This table contains a preliminary list of prioritized CNVs in 10 large pedigrees aggregated with schizophrenia and bipolar disorder. We kept CNVs present in at least 5 affected individuals and present more in affected than unaffected family members. We excluded those

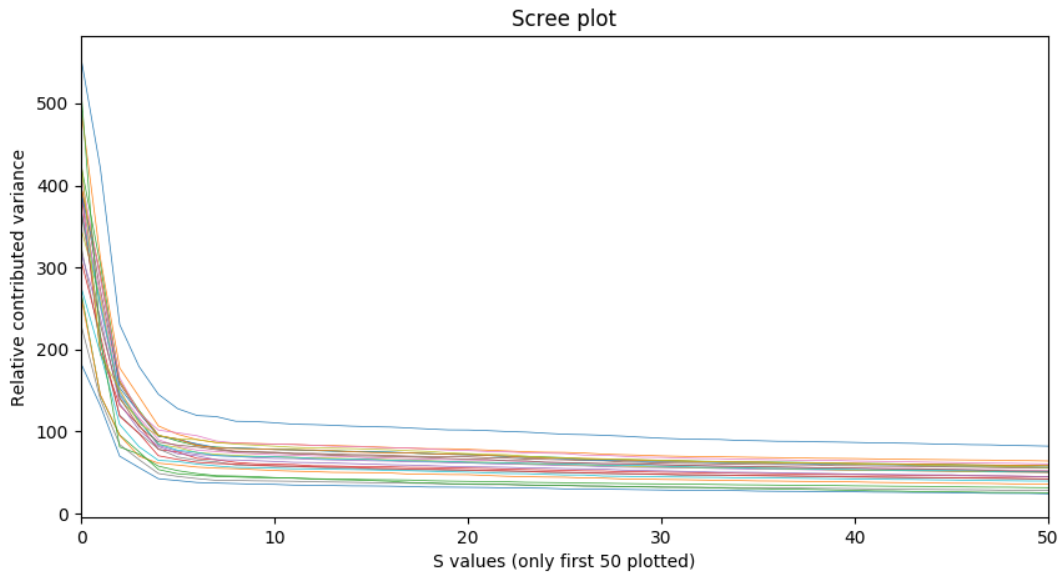
CNVs present in 1KG and DGV database except they have been previously reported to have an association to schizophrenia and bipolar disorder in large scale CNV studies. The left CNVs were checked if the cytoband regions if they are overlapped with significantly associated loci of schizophrenia and bipolar disorder genome-wide association studies. Legend: type: Del, deletion; Dup, duplication. A, number of affected carriers in the family, UA, number of unaffected carriers in the family. *CO, CoNIFER; XH, XHMM; CN, CNVpartition; PN, PennCNV; QT, QuantiSNP. SCZ, schizophrenia; BP, bipolar disorder; GWAS, genome-wide association studies.

Supplementary Figure 1. Pipeline for CNV calling and processing

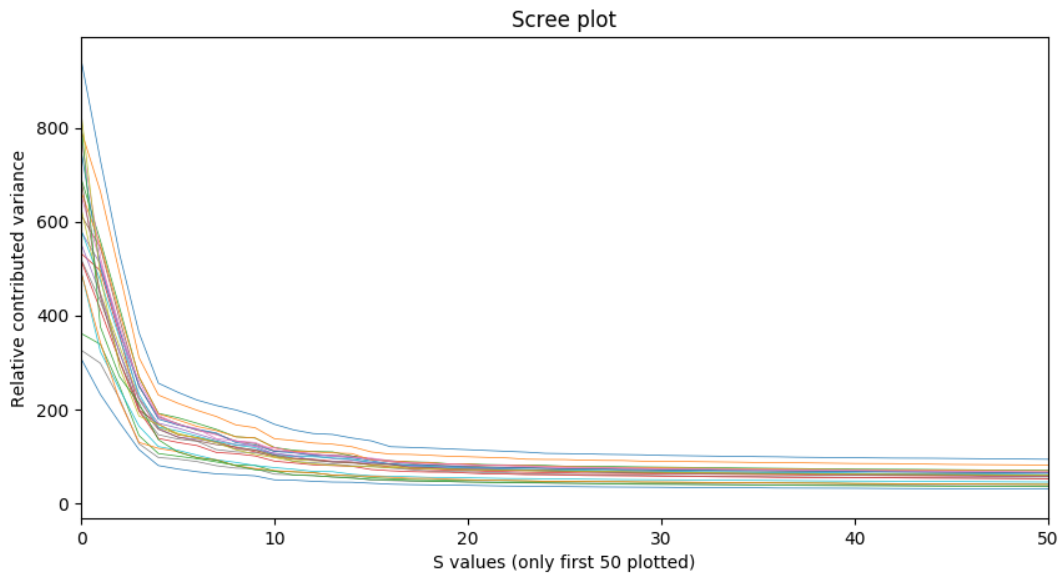


Supplementary Figure 2. Scree plots of SVD components CoNIFER, separated by capture kit.

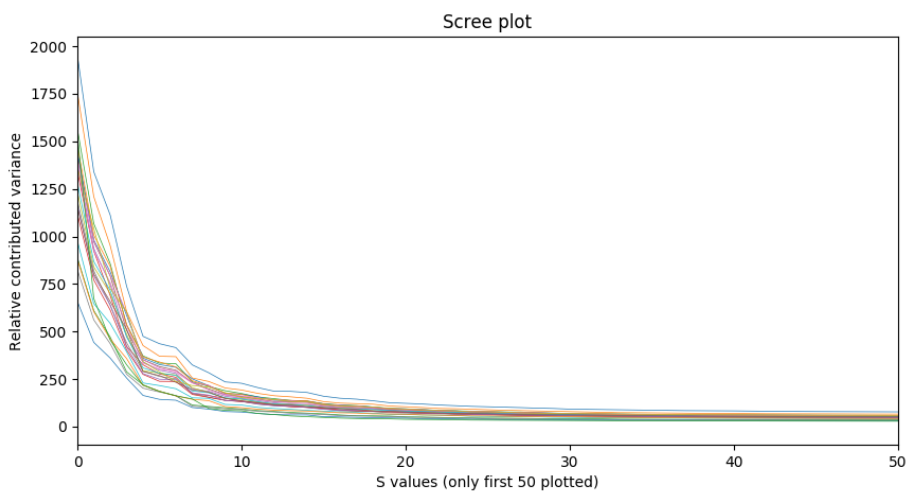
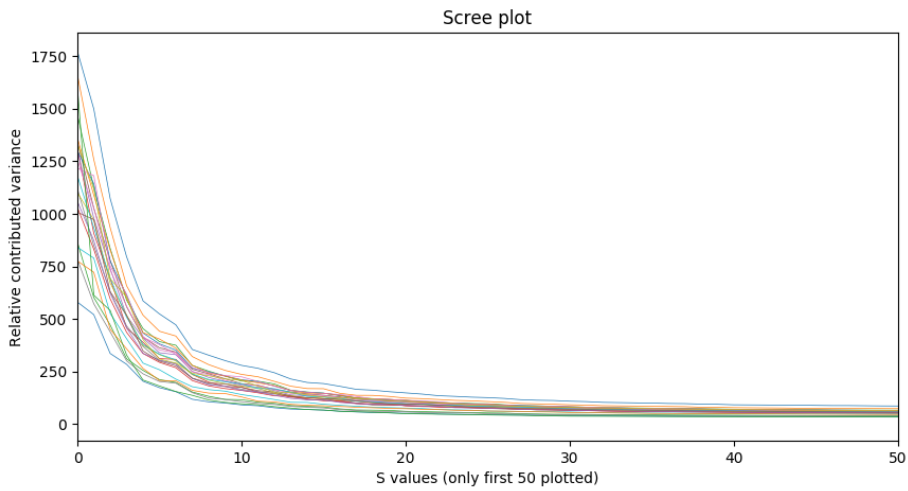
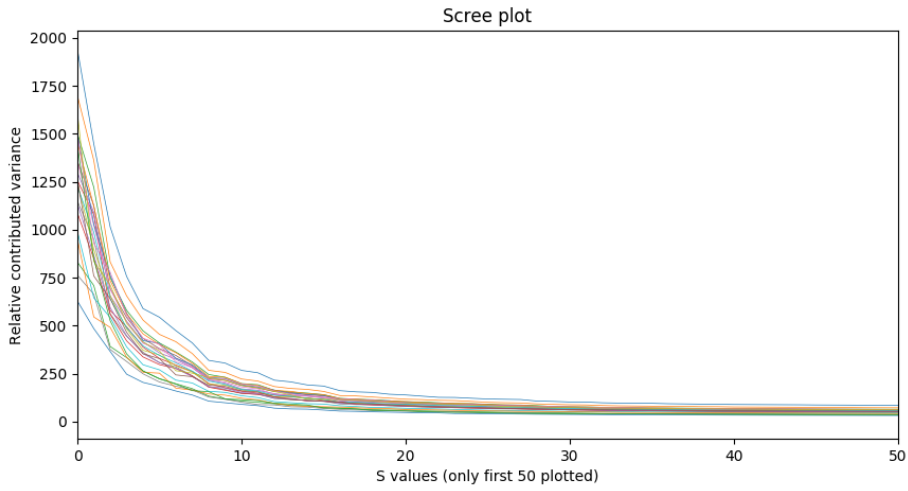
38M Capture Kit (suggested number of SVD components removed, --SVD 4)



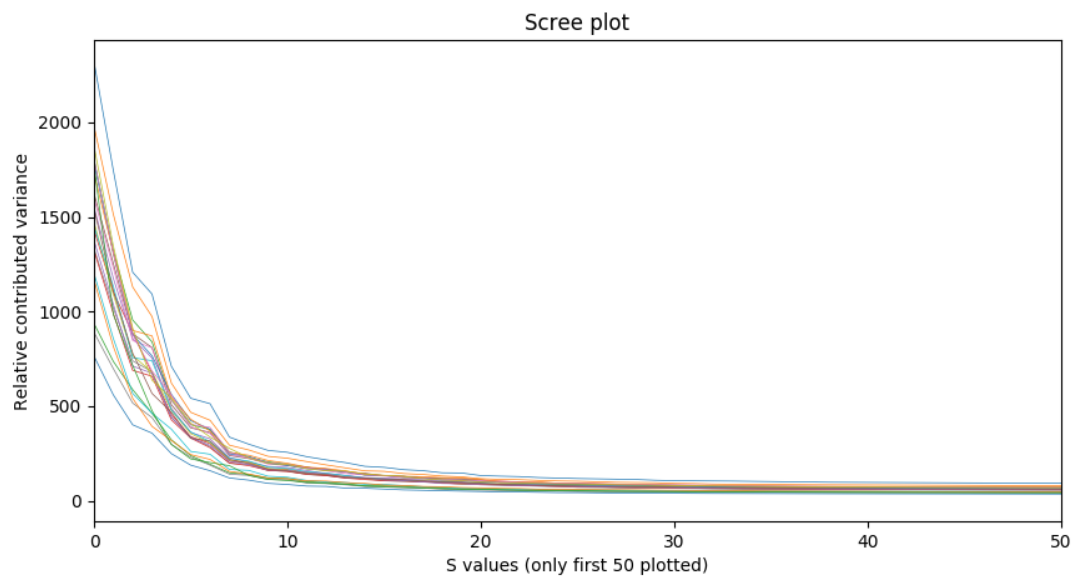
50M Capture Kit (suggested number of SVD components removed, --SVD 4)



V4s1, s2, s3 Capture Kit (suggested number of SVD components removed, --SVD 4)



V5 Capture Kit (suggested number of SVD components removed, --SVD 5)



BIBLIOGRAPHY

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*. (American Psychiatric Association, 2013).
2. Stahl, S. M. *Stahl's Essential Psychopharmacology, 4th Edition*. (Cambridge University Press, 2013).
3. Harpending, H. C. *et al.* Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1961–1967 (1998).
4. Zhivotovsky, L. A., Rosenberg, N. A. & Feldman, M. W. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am. J. Hum. Genet.* **72**, 1171–1186 (2003).
5. Liu, H., Prugnolle, F., Manica, A. & Balloux, F. A geographically explicit genetic model of worldwide human-settlement history. *Am. J. Hum. Genet.* **79**, 230–237 (2006).
6. Tenesa, A. *et al.* Recent human effective population size estimated from linkage disequilibrium. *Genome Res.* **17**, 520–526 (2007).
7. Bittles, A. H. *Consanguinity in Context. Cambridge Core* (2012).
doi:10.1017/CBO9781139015844
8. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98 (2011).
9. Hawass, Z. *et al.* Ancestry and Pathology in King Tutankhamun's Family. *JAMA* **303**, 638–647 (2010).
10. Callaway, E. Inbred royals show traces of natural selection. *Nat. News*
doi:10.1038/nature.2013.12837

11. Alvarez, G., Ceballos, F. C. & Quinteiro, C. The role of inbreeding in the extinction of a European royal dynasty. *PloS One* **4**, e5174 (2009).
12. Ceballos, F. C. & Alvarez, G. Royal dynasties as human inbreeding laboratories: the Habsburgs. *Heredity* **111**, 114–121 (2013).
13. Report on influence of marriages of consanguinity upon offspring. Available at: <https://collections.nlm.nih.gov/catalog/nlm:nlmuid-60531150R-bk>. (Accessed: 21st July 2018)
14. Darwin, G. H. Note on the Marriages of First Cousins. *J. Stat. Soc. Lond.* **38**, 344–348 (1875).
15. Darwin, G. H. Marriages Between First Cousins in England and Their Effects. *J. Stat. Soc. Lond.* **38**, 153–184 (1875).
16. Berra, T. M., Alvarez, G. & Ceballos, F. C. Was the Darwin/Wedgwood Dynasty Adversely Affected by Consanguinity? *BioScience* **60**, 376–383 (2010).
17. Bittles, A. H. A Community Genetics Perspective on Consanguineous Marriage. *Public Health Genomics* **11**, 324–330 (2008).
18. Bittles, A. Consanguinity and its relevance to clinical genetics. *Clin. Genet.* **60**, 89–98 (2001).
19. Global prevalence - ConsangWiki - Consang.net. Available at: http://consang.net/index.php/Global_prevalence. (Accessed: 7th February 2019)
20. Bittles, A. H. & Neel, J. V. The costs of human inbreeding and their implications for variations at the DNA level. *Nat. Genet.* **8**, 117–121 (1994).
21. Sheehan, N. A., Didelez, V., Burton, P. R. & Tobin, M. D. Mendelian Randomisation and Causal Inference in Observational Epidemiology. *PLOS Med.* **5**, e177 (2008).

22. Garrod, A. E. The incidence of alkaptonuria: a study in chemical individuality. *Lancet* **2**, 1616–1620 (1902).
23. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
24. Borck, G. *et al.* Loss-of-function mutations of ILDR1 cause autosomal-recessive hearing impairment DFNB42. *Am. J. Hum. Genet.* **88**, 127–137 (2011).
25. Bittles, A. H. & Black, M. L. Consanguinity, human evolution, and complex diseases. *Proc. Natl. Acad. Sci.* **107**, 1779–1786 (2010).
26. Dennell, R. W., Rendell, H. M., Halim, M. & Moth, E. A 45,000-Year-Old Open-Air Paleolithic Site at Riwat, Northern Pakistan. *J. Field Archaeol.* **19**, 17–33 (1992).
27. Mehdi, S. Q. *et al.* The Origins of Pakistani Populations. in *Genomic Diversity* 83–90 (Springer, Boston, MA, 1999). doi:10.1007/978-1-4615-4263-6_7
28. Firasat, S. *et al.* Y-chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan. *Eur. J. Hum. Genet. EJHG* **15**, 121–126 (2007).
29. Qamar, R. *et al.* Y-chromosomal DNA variation in Pakistan. *Am. J. Hum. Genet.* **70**, 1107–1124 (2002).
30. Mansoor, A. *et al.* Investigation of the Greek ancestry of populations from northern Pakistan. *Hum. Genet.* **114**, 484–490 (2004).
31. Firasat, S. *et al.* Y-chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan. *Eur. J. Hum. Genet. EJHG* **15**, 121–126 (2007).
32. Narasimhan, V. M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* aac8624 (2016). doi:10.1126/science.aac8624

33. Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
34. A., D., Taleb, M., Blecha, L. & Benyamina, A. Genetics and psychotic disorders: A fresh look at consanguinity. *Eur. J. Med. Genet.* **59**, 104–110 (2016).
35. Rao, T. S. S. *et al.* Relationship between consanguinity and depression in a south Indian population. *Indian J. Psychiatry* **51**, 50–52 (2009).
36. Motlagh, M. G., Seddigh, A., Dashti, B., Leckman, J. F. & Alaghband-Rad, J. Consanguineous Iranian kindreds with severe Tourette syndrome. *Mov. Disord. Off. J. Mov. Disord. Soc.* **23**, 2079–2083 (2008).
37. Ahmed, A. H. Consanguinity and schizophrenia in Sudan. *Br. J. Psychiatry J. Ment. Sci.* **134**, 635–636 (1979).
38. Chaleby, K. & Tuma, T. A. Cousin marriages and schizophrenia in Saudi Arabia. *Br. J. Psychiatry J. Ment. Sci.* **150**, 547–549 (1987).
39. Mansour, H. *et al.* Consanguinity and increased risk for schizophrenia in Egypt. *Schizophr. Res.* **120**, 108–112 (2010).
40. Saugstad, L. & ØDegård, Ø. Inbreeding and schizophrenia. *Clin. Genet.* **30**, 261–275 (1986).
41. Bulayeva, K. B. *et al.* The ascertainment of multiplex schizophrenia pedigrees from Daghestan genetic isolates (Northern Caucasus, Russia). *Psychiatr. Genet.* **10**, 67–72 (2000).
42. Bulayeva, K., Bulayev, O. & Glatt, S. *Genomic Architecture of Schizophrenia Across Diverse Genetic Isolates: A Study of Dagestan Populations.* (Springer International Publishing, 2016).

43. Britvić, D., Aleksić-Shihabi, A., Titlić, M. & Dolić, K. Schizophrenia spectrum psychosis in a Croatian genetic isolate: genealogical reconstructions. *Psychiatr. Danub.* **22**, 51–56 (2010).
44. Dobrusin, M. *et al.* The rate of consanguineous marriages among parents of schizophrenic patients in the Arab Bedouin population in Southern Israel. *World J. Biol. Psychiatry* **10**, 334–336 (2009).
45. Holliday, E. G. *et al.* Strong evidence for a novel schizophrenia risk locus on chromosome 1p31.1 in homogeneous pedigrees from Tamil Nadu, India. *Am. J. Psychiatry* **166**, 206–215 (2009).
46. Hovatta, I. *et al.* A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am. J. Hum. Genet.* **65**, 1114–1124 (1999).
47. Garner, C. *et al.* Linkage analysis of a complex pedigree with severe bipolar disorder, using a Markov chain Monte Carlo method. *Am. J. Hum. Genet.* **68**, 1061–1064 (2001).
48. Abecasis, G. R. *et al.* Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *Am. J. Hum. Genet.* **74**, 403–417 (2004).
49. Knight, H. M. *et al.* Homozygosity mapping in a family presenting with schizophrenia, epilepsy and hearing impairment. *Eur. J. Hum. Genet. EJHG* **16**, 750–758 (2008).
50. Kurotaki, N. *et al.* Identification of Novel Schizophrenia Loci by Homozygosity Mapping Using DNA Microarray Analysis. *PLoS ONE* **6**, (2011).
51. Lencz, T. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci.* **104**, 19942–19947 (2007).

52. Keller, M. C. *et al.* Runs of Homozygosity Implicate Autozygosity as a Schizophrenia Risk Factor. *PLoS Genet.* **8**, e1002656 (2012).
53. Johnson, E. C. *et al.* No Reliable Association between Runs of Homozygosity and Schizophrenia in a Well-Powered Replication Study. *PLoS Genet.* **12**, e1006343 (2016).
54. Heron, E. A. *et al.* No evidence that runs of homozygosity are associated with schizophrenia in an Irish genome-wide association dataset. *Schizophr. Res.* **154**, 79–82 (2014).
55. Vine, A. E. *et al.* No evidence for excess runs of homozygosity in bipolar disorder. *Psychiatr. Genet.* **19**, 165–170 (2009).
56. Whiteford, H. A. *et al.* Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet* **382**, 1575–1586 (2013).
57. Saha, S., Chant, D., Welham, J. & McGrath, J. A systematic review of the prevalence of schizophrenia. *PLoS Med.* **2**, e141 (2005).
58. Merikangas, K. R. *et al.* Prevalence and Correlates of Bipolar Spectrum Disorder in the World Mental Health Survey Initiative. *Arch. Gen. Psychiatry* **68**, 241–251 (2011).
59. Sullivan, P. F., Kendler, K. S. & Neale, M. C. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch. Gen. Psychiatry* **60**, 1187–1192 (2003).
60. Smoller, J. W. & Finn, C. T. Family, twin, and adoption studies of bipolar disorder. *Am. J. Med. Genet. C Semin. Med. Genet.* **123C**, 48–58 (2003).
61. Polderman, T. J. C. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).

62. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
63. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
64. Van Snellenberg, J. X. & de Candia, T. Meta-analytic evidence for familial coaggregation of schizophrenia and bipolar disorder. *Arch. Gen. Psychiatry* **66**, 748–755 (2009).
65. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet Lond. Engl.* **373**, 234–239 (2009).
66. Badner, J. A. & Gershon, E. S. Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Mol. Psychiatry* **7**, 405–411 (2002).
67. Lachman, H. M. Copy variations in schizophrenia and bipolar disorder. *Cytogenet. Genome Res.* **123**, 27–35 (2008).
68. Owen, M. J., Craddock, N. & Jablensky, A. The genetic deconstruction of psychosis. *Schizophr. Bull.* **33**, 905–911 (2007).
69. O’Donovan, M. C. *et al.* Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat. Genet.* **40**, 1053–1055 (2008).
70. Green, E. K. *et al.* The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Mol. Psychiatry* **15**, 1016–1022 (2010).
71. Group, P. G. C. B. D. W. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat. Genet.* **43**, 977–983 (2011).
72. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

73. Craddock, N. & Owen, M. J. The Kraepelinian dichotomy – going, going... but still not gone. *Br. J. Psychiatry* **196**, 92–95 (2010).
74. Adam, D. Mental health: On the spectrum. *Nature* **496**, 416–418 (2013).
75. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
76. Zhang, D. *et al.* Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol. Psychiatry* **14**, 376–380 (2009).
77. Malhotra, D. *et al.* High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* **72**, 951–963 (2011).
78. McCarthy, S. E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
79. Bergen, S. E. *et al.* Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol. Psychiatry* **17**, 880–886 (2012).
80. Curtis, D. *et al.* Case-case genome-wide association analysis shows markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. *Psychiatr. Genet.* **21**, 1–4 (2011).
81. Ruderfer, D. M. *et al.* Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. *Mol. Psychiatry* **19**, 1017–1024 (2014).
82. Ruderfer, D. M. *et al.* Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705-1715.e16 (2018).
83. Burmeister, M., McInnis, M. G. & Zöllner, S. Psychiatric genetics: progress amid controversy. *Nat. Rev. Genet.* **9**, 527–540 (2008).

84. Lewis, C. M. *et al.* Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am. J. Hum. Genet.* **73**, 34–48 (2003).
85. Ng, M. Y. M. *et al.* Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol. Psychiatry* **14**, 774–785 (2009).
86. Segurado, R. *et al.* Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. *Am. J. Hum. Genet.* **73**, 49–62 (2003).
87. McQueen, M. B. *et al.* Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. *Am. J. Hum. Genet.* **77**, 582–595 (2005).
88. Allen, N. C. *et al.* Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat. Genet.* **40**, 827–834 (2008).
89. Rebbeck, T. R., Ambrosone, C. B. & Shields, P. G. *Molecular Epidemiology: Applications in Cancer and Other Human Diseases.* (Taylor & Francis, 2008).
90. Ott, J., Kamatani, Y. & Lathrop, M. Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* **12**, 465–474 (2011).
91. Irmansyah *et al.* Genome-wide scan in 124 Indonesian sib-pair families with schizophrenia reveals genome-wide significant linkage to a locus on chromosome 3p26-21. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **147B**, 1245–1252 (2008).
92. Zhang, R. *et al.* Population-based and family-based association studies of ZNF804A locus and schizophrenia. *Mol. Psychiatry* **16**, 360–361 (2011).
93. Steinberg, S. *et al.* Truncating mutations in RBM12 are associated with psychosis. *Nat. Genet.* **49**, 1251–1254 (2017).

94. Toma, C. *et al.* An examination of multiple classes of rare variants in extended families with bipolar disorder. *Transl. Psychiatry* **8**, 65 (2018).
95. Sul, J. H. *et al.* Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. *bioRxiv* 363267 (2018). doi:10.1101/363267
96. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
97. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
98. Ikeda, M., Saito, T., Kondo, K. & Iwata, N. Genome-wide association studies of bipolar disorder: A systematic review of recent findings and their clinical implications. *Psychiatry Clin. Neurosci.* **72**, 52–63 (2018).
99. Ahlqvist, E., van Zuydam, N. R., Groop, L. C. & McCarthy, M. I. The genetics of diabetic complications. *Nat. Rev. Nephrol.* **11**, 277–287 (2015).
100. Chen, D. T. *et al.* Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Mol. Psychiatry* **18**, 195–205 (2013).
101. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
102. Sullivan, P. F., Daly, M. J. & O’Donovan, M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.* **13**, 537–551 (2012).
103. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

104. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
105. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
106. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
107. Hamshere, M. L. *et al.* Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Mol. Psychiatry* **18**, 708–712 (2013).
108. Ferreira, M. A. R. *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.* **40**, 1056–1058 (2008).
109. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
110. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576–1583 (2017).
111. Harrison, P. J. Recent genetic findings in schizophrenia and their therapeutic relevance. *J. Psychopharmacol. Oxf. Engl.* **29**, 85–96 (2015).
112. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
113. Scott, L. J. *et al.* Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7501–7506 (2009).
114. Mühleisen, T. W. *et al.* Genome-wide association study reveals two new risk loci for bipolar disorder. *Nat. Commun.* **5**, 3339 (2014).

115. Hou, L. *et al.* Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum. Mol. Genet.* **25**, 3383–3394 (2016).
116. Stahl, E. *et al.* Genomewide association study identifies 30 loci associated with bipolar disorder. *bioRxiv* 173062 (2018). doi:10.1101/173062
117. McClellan, J. M., Susser, E. & King, M.-C. Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry J. Ment. Sci.* **190**, 194–199 (2007).
118. Power, R. A. *et al.* Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22–30 (2013).
119. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
120. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
121. Kirov, G. CNVs in neuropsychiatric disorders. *Hum. Mol. Genet.* **24**, R45–R49 (2015).
122. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
123. Driscoll, D. A. *et al.* Deletions and microdeletions of 22q11.2 in velo-cardio-facial syndrome. *Am. J. Med. Genet.* **44**, 261–268 (1992).
124. Shprintzen, R. J., Goldberg, R., Golding-Kushner, K. J. & Marion, R. W. Late-onset psychosis in the velo-cardio-facial syndrome. *Am. J. Med. Genet.* **42**, 141–142 (1992).
125. Murphy, K. C., Jones, L. A. & Owen, M. J. High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Arch. Gen. Psychiatry* **56**, 940–945 (1999).

126. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
127. Kirov, G. *et al.* Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum. Mol. Genet.* **17**, 458–465 (2008).
128. Kirov, G. *et al.* Neurexin 1 (NRXN1) deletions in schizophrenia. *Schizophr. Bull.* **35**, 851–854 (2009).
129. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
130. Kirov, G. *et al.* The penetrance of copy number variations for schizophrenia and developmental delay. *Biol. Psychiatry* **75**, 378–385 (2014).
131. Priebe, L. *et al.* Genome-wide survey implicates the influence of copy number variants (CNVs) in the development of early-onset bipolar disorder. *Mol. Psychiatry* **17**, 421–432 (2012).
132. Green, E. K. *et al.* Copy number variation in bipolar disorder. *Mol. Psychiatry* **21**, 89–93 (2016).
133. Homann, O. R. *et al.* Whole-genome sequencing in multiplex families with psychoses reveals mutations in the SHANK2 and SMARCA1 genes segregating with illness. *Mol. Psychiatry* **21**, 1690–1695 (2016).
134. Goes FS, Pirooznia M, Parla JS & *et al.* Exome sequencing of familial bipolar disorder. *JAMA Psychiatry* (2016). doi:10.1001/jamapsychiatry.2016.0251
135. Ament, S. A. *et al.* Rare variants in neuronal excitability genes influence risk for bipolar disorder. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3576–3581 (2015).

136. Cruceanu, C. *et al.* Rare susceptibility variants for bipolar disorder suggest a role for G protein-coupled receptors. *Mol. Psychiatry* (2017). doi:10.1038/mp.2017.223
137. Purcell, S. M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
138. Singh, T. *et al.* Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016).
139. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
140. Craddock, N. & Sklar, P. Genetics of bipolar disorder. *The Lancet* **381**, 1654–1662 (2013).
141. Lescai, F. *et al.* Whole-exome sequencing of individuals from an isolated population implicates rare risk variants in bipolar disorder. *Transl. Psychiatry* **7**, e1034 (2017).
142. Campbell, I. M., Shaw, C. A., Stankiewicz, P. & Lupski, J. R. Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet. TIG* **31**, 382–392 (2015).
143. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
144. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25**, 792–801 (2015).
145. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* **17**, 241 (2016).
146. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).

147. de Kluiver, H., Buizer-Voskamp, J. E., Dolan, C. V. & Boomsma, D. I. Paternal age and psychiatric disorders: A review. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **174**, 202–213 (2017).
148. Frans, E. M. *et al.* Advancing paternal age and bipolar disorder. *Arch. Gen. Psychiatry* **65**, 1034–1040 (2008).
149. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
150. Xu, B. *et al.* Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat. Genet.* **40**, 880–885 (2008).
151. Xu, B. *et al.* Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nat. Genet.* **40**, 880–885 (2008).
152. Jr, E. H. C. & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* (2008). doi:10.1038/nature07458
153. Gauthier, J. *et al.* De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7863–7868 (2010).
154. Awadalla, P. *et al.* Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* **87**, 316–324 (2010).
155. Girard, S. L. *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat. Genet.* **43**, 860–863 (2011).
156. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat. Genet.* **43**, 864–868 (2011).

157. Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol. Psychiatry* **17**, 142–153 (2012).
158. Xu, B. *et al.* De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
159. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
160. Wang, Q. *et al.* Increased co-expression of genes harboring the damaging de novo mutations in Chinese schizophrenic patients during prenatal development. *Sci. Rep.* **5**, 18209 (2015).
161. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
162. McCarthy, S. E. *et al.* De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol. Psychiatry* **19**, 652–658 (2014).
163. Kranz, T. M. *et al.* De novo mutations from sporadic schizophrenia cases highlight important signaling genes in an independent sample. *Schizophr. Res.* **166**, 119–124 (2015).
164. Takata, A., Ionita-Laza, I., Gogos, J. A., Xu, B. & Karayiorgou, M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. *Neuron* **89**, 940–947 (2016).
165. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).

166. Kataoka, M. *et al.* Exome sequencing for bipolar disorder points to roles of de novo loss-of-function and protein-altering mutations. *Mol. Psychiatry* **21**, 885–893 (2016).
167. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
168. Sullivan, P. F. *et al.* Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Mol. Psychiatry* **13**, 570–584 (2008).
169. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753–757 (2009).
170. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460**, 744–747 (2009).
171. Huang, J. *et al.* Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression. *Am. J. Psychiatry* **167**, 1254–1263 (2010).
172. Ikeda, M. *et al.* Genome-wide association study of schizophrenia in a Japanese population. *Biol. Psychiatry* **69**, 472–478 (2011).
173. Yamada, K. *et al.* Genome-wide association study of schizophrenia in Japanese population. *PloS One* **6**, e20468 (2011).
174. Consortium, T. S. P. G.-W. A. S. (GWAS). Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
175. Yue, W.-H. *et al.* Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nat. Genet.* **43**, 1228–1231 (2011).
176. Shi, Y. *et al.* Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat. Genet.* **43**, 1224–1227 (2011).

177. Irish Schizophrenia Genomics Consortium and the Wellcome Trust Case Control Consortium 2. Genome-wide association study implicates HLA-C*01:02 as a risk factor at the major histocompatibility complex locus in schizophrenia. *Biol. Psychiatry* **72**, 620–628 (2012).
178. Aberg, K. A. *et al.* A comprehensive family-based replication study of schizophrenia genes. *JAMA Psychiatry* **70**, 573–581 (2013).
179. Wong, E. H. M. *et al.* Common variants on Xq28 conferring risk of schizophrenia in Han Chinese. *Schizophr. Bull.* **40**, 777–786 (2014).
180. Sleiman, P. *et al.* GWAS meta analysis identifies TSNARE1 as a novel Schizophrenia / Bipolar susceptibility locus. *Sci. Rep.* **3**, 3075 (2013).
181. Lencz, T. *et al.* Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat. Commun.* **4**, 2739 (2013).
182. Goes, F. S. *et al.* Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **168**, 649–659 (2015).
183. Kim, L. H. *et al.* Genome-wide association study with the risk of schizophrenia in a Korean population. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **171B**, 257–265 (2016).
184. Yu, H. *et al.* Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in Han Chinese population. *Mol. Psychiatry* **22**, 954–960 (2017).
185. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

186. Sklar, P. *et al.* Whole-genome association study of bipolar disorder. *Mol. Psychiatry* **13**, 558–569 (2008).
187. Smith, E. N. *et al.* Genome-wide association study of bipolar disorder in European American and African American individuals. *Mol. Psychiatry* **14**, 755–763 (2009).
188. Liu, Y. *et al.* Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. *Mol. Psychiatry* **16**, 2–4 (2011).
189. Lee, M. T. M. *et al.* Genome-wide association study of bipolar I disorder in the Han Chinese population. *Mol. Psychiatry* **16**, 548–556 (2011).
190. Hou, L. *et al.* Genetic variants associated with response to lithium treatment in bipolar disorder: a genome-wide association study. *Lancet Lond. Engl.* **387**, 1085–1093 (2016).
191. van Hulzen, K. J. E. *et al.* Genetic Overlap Between Attention-Deficit/Hyperactivity Disorder and Bipolar Disorder: Evidence From Genome-wide Association Study Meta-analysis. *Biol. Psychiatry* **82**, 634–641 (2017).
192. Ikeda, M. *et al.* A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Mol. Psychiatry* **23**, 639–647 (2018).
193. Mulle, J. G. *et al.* Microdeletions of 3q29 confer high risk for schizophrenia. *Am. J. Hum. Genet.* **87**, 229–236 (2010).
194. Rujescu, D. *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum. Mol. Genet.* **18**, 988–996 (2009).
195. Guha, S. *et al.* Implication of a rare deletion at distal 16p11.2 in schizophrenia. *JAMA Psychiatry* **70**, 253–260 (2013).

196. Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol. Psychiatry* **19**, 37–40 (2014).
197. Levinson, D. F. *et al.* Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am. J. Psychiatry* **168**, 302–316 (2011).
198. Mulle, J. G. *et al.* Reciprocal duplication of the Williams-Beuren syndrome deletion on chromosome 7q11.23 is associated with schizophrenia. *Biol. Psychiatry* **75**, 371–377 (2014).
199. Ingason, A. *et al.* Maternally derived microduplications at 15q11-q13: implication of imprinted genes in psychotic illness. *Am. J. Psychiatry* **168**, 408–417 (2011).
200. Degenhardt, F. *et al.* Duplications in RB1CC1 are associated with schizophrenia; identification in large European sample sets. *Transl. Psychiatry* **3**, e326 (2013).
201. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* **471**, 499–503 (2011).
202. Rees, E. *et al.* CNV analysis in a large schizophrenia sample implicates deletions at 16p12.1 and SLC1A1 and duplications at 1p36.33 and CGNL1. *Hum. Mol. Genet.* **23**, 1669–1676 (2014).
203. Moreno-De-Luca, D. *et al.* Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am. J. Hum. Genet.* **87**, 618–630 (2010).
204. Kirov, G. *et al.* Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum. Mol. Genet.* **18**, 1497–1503 (2009).

205. Toufaily, M. H., Westgate, M.-N., Nasri, H. & Holmes, L. B. Malformations among 289,365 Births Attributed to Mutations with Autosomal Dominant and Recessive and X-Linked Inheritance. *Birth Defects Res.* **110**, 92–97 (2018).
206. Tunca, C. *et al.* ERLIN1 mutations cause teenage-onset slowly progressive ALS in a large Turkish pedigree. *Eur. J. Hum. Genet. EJHG* (2018). doi:10.1038/s41431-018-0107-5
207. Riazuddin, S. *et al.* Exome sequencing of Pakistani consanguineous families identifies 30 novel candidate genes for recessive intellectual disability. *Mol. Psychiatry* **22**, 1604–1614 (2017).
208. Corry, P. C. Consanguinity and Prevalence Patterns of Inherited Disease in the UK Pakistani Community. *Hum. Hered.* **77**, 207–216 (2014).
209. Génin, E. & Todorov, A. A. Homozygosity Mapping. in *Homozygosity Mapping, eLS* (John Wiley & Sons, Ltd, 2001).
210. Raychaudhuri, S. Mapping Rare and Common Causal Alleles for Complex Human Diseases. *Cell* **147**, 57–69 (2011).
211. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
212. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinforma. Oxf. Engl.* **26**, 2867–2873 (2010).
213. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
214. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

215. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
216. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
217. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
218. Gazal, S., Sahbatou, M., Babron, M.-C., Génin, E. & Leutenegger, A.-L. FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics* **30**, 1940–1941 (2014).
219. Leutenegger, A.-L. *et al.* Estimation of the Inbreeding Coefficient through Use of Genomic Data. *Am. J. Hum. Genet.* **73**, 516–523 (2003).
220. McQuillan, R. *et al.* Runs of Homozygosity in European Populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
221. Kancheva, D. *et al.* Novel mutations in genes causing hereditary spastic paraplegia and Charcot-Marie-Tooth neuropathy identified by an optimized protocol for homozygosity mapping based on whole-exome sequencing. *Genet. Med.* **18**, 600–607 (2016).
222. Belkadi, A. *et al.* Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6713–6718 (2016).
223. Seelow, D., Schuelke, M., Hildebrandt, F. & Nürnberg, P. HomozygosityMapper—an interactive approach to homozygosity mapping. *Nucleic Acids Res.* **37**, W593–W599 (2009).
224. Regier, D. A. Dimensional approaches to psychiatric classification: refining the research agenda for DSM-V: an introduction. *Int. J. Methods Psychiatr. Res.* **16 Suppl 1**, S1-5 (2007).

225. Harripaul, R. *et al.* Mapping autosomal recessive intellectual disability: combined microarray and exome sequencing identifies 26 novel candidate genes in 192 consanguineous families. *Mol. Psychiatry* (2017). doi:10.1038/mp.2017.60
226. Hu, H. *et al.* Genetics of intellectual disability in consanguineous families. *Mol. Psychiatry* 1 (2018). doi:10.1038/s41380-017-0012-2
227. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* (2018). doi:10.1038/nrg.2017.109
228. Maguire, A., Tseliou, F. & O'Reilly, D. Consanguineous Marriage and the Psychopathology of Progeny: A Population-wide Data Linkage Study. *JAMA Psychiatry* (2018). doi:10.1001/jamapsychiatry.2018.0133
229. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
230. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986-992 (2014).
231. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
232. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
233. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).

234. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
235. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
236. Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
237. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
238. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
239. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
240. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
241. Park, H. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42**, 400–405 (2010).
242. Hong, C. S., Singh, L. N., Mullikin, J. C. & Biesecker, L. G. Assessing the reproducibility of exome copy number variations predictions. *Genome Med.* **8**, 82 (2016).
243. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res.* gr.138115.112 (2012). doi:10.1101/gr.138115.112
244. Krumm, N. *et al.* Transmission Disequilibrium of Small CNVs in Simplex Autism. *Am. J. Hum. Genet.* **93**, 595–606 (2013).

245. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
246. Fransson, S. *et al.* Estimation of copy number aberrations: Comparison of exome sequencing data with SNP microarrays identifies homozygous deletions of 19q13.2 and CIC in neuroblastoma. *Int. J. Oncol.* **48**, 1103–1116 (2016).
247. de Ligt, J. *et al.* Platform comparison of detecting copy number variants with microarrays and whole-exome sequencing. *Genomics Data* **2**, 144–146 (2014).
248. Zhou, Z., Wang, W., Wang, L.-S., Zhang, N. R. & Birol, I. Integrative DNA copy number detection and genotyping from sequencing and array-based platforms. *Bioinformatics* doi:10.1093/bioinformatics/bty104
249. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007).
250. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
251. Fromer, M. *et al.* Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
252. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
253. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

254. Khan, F. F. *et al.* Whole genome sequencing of 91 multiplex schizophrenia families reveals increased burden of rare, exonic copy number variation in schizophrenia probands and genetic heterogeneity. *Schizophr. Res.* (2018). doi:10.1016/j.schres.2018.02.034
255. He, Z. *et al.* Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am. J. Hum. Genet.* **94**, 33–46 (2014).
256. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. Family-based association tests for sequence data, and comparisons with population-based association tests. *Eur. J. Hum. Genet. EJHG* **21**, 1158–1162 (2013).
257. Epstein, M. P. *et al.* A statistical approach for rare-variant association testing in affected sibships. *Am. J. Hum. Genet.* **96**, 543–554 (2015).
258. Sul, J. H. *et al.* Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees. *Am. J. Hum. Genet.* **99**, 846–859 (2016).
259. He, Z. *et al.* The Rare-Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer Disease WGS Data. *Am. J. Hum. Genet.* **100**, 193–204 (2017).
260. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
261. Wu, J. Q. *et al.* Transcriptome sequencing revealed significant alteration of cortical promoter usage and splicing in schizophrenia. *PLoS One* **7**, e36351 (2012).
262. Ellis, S. E., Panitch, R., West, A. B. & Arking, D. E. Transcriptome analysis of cortical tissue reveals shared sets of downregulated genes in autism and schizophrenia. *Transl. Psychiatry* **6**, e817 (2016).

263. Veldic, M. *et al.* DNA-methyltransferase 1 mRNA is selectively overexpressed in telencephalic GABAergic interneurons of schizophrenia brains. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 348–353 (2004).
264. Zhubi, A. *et al.* An upregulation of DNA-methyltransferase 1 and 3a expressed in telencephalic GABAergic neurons of schizophrenia patients is also detected in peripheral blood lymphocytes. *Schizophr. Res.* **111**, 115–122 (2009).
265. Grayson, D. R. & Guidotti, A. The dynamics of DNA methylation in schizophrenia and related psychiatric disorders. *Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol.* **38**, 138–166 (2013).
266. Houle, D., Govindaraju, D. R. & Omholt, S. Phenomics: the next challenge. *Nat. Rev. Genet.* **11**, 855–866 (2010).
267. Bilder, R. M. *et al.* Cognitive ontologies for neuropsychiatric phenomics research. *Cognit. Neuropsychiatry* **14**, 419–450 (2009).
268. Gorgolewski, K. J., Durnez, J. & Poldrack, R. A. Preprocessed Consortium for Neuropsychiatric Phenomics dataset. *F1000Research* **6**, (2017).
269. Custom Cluster File Creation for Improved Copy Number Analysis. Available at: <https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/custom-cluster-file-cnv-tech-note-770-2017-017.pdf>.

