Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE

INDIRIZZO: GENETICA E BIOLOGIA MOLECOLARE DELLO SVILUPPO

CICLO XXVIII

The genetic architecture of schizophrenia and bipolar disorder: identity by descent and exome sequencing in a family-based population sample

Direttore della Scuola: Ch.mo Prof. Paolo Bernardi

Coordinatore di indirizzo: Ch.mo Prof. Rodolfo Costa

Supervisore: Ch.ma Prof.ssa Stefania Bortoluzzi

Co-supervisore: Dott. Giovanni Vazza

Co-supervisore: Ch.ma Prof.ssa Maria L. Mostacciuolo

Dottoranda: Cecilia Salvoro

Table of contents

1. Summary 1

2. Riassunto 4

1. Introduction 7

1.1 Preface 7

1.2 Clinical aspects of schizophrenia and bipolar disorder 9

1.1.1 Schizophrenia 9

1.1.2 Bipolar disorder 10

1.1.3 The clinical overlap of schizophrenia and bipolar disorder 11

1.3 Unravelling the genetics of schizophrenia and bipolar disorder 12

1.3.1 Epidemiological studies 12

1.1.4 Search for common risk variants 12

1.1.5 Search of rare risk variants 16

1.1.6 The genetic overlap between schizophrenia and bipolar disorder 21

1.4 The emerging picture on schizophrenia and bipolar disorder 22

2. Aim of the research 25

3. Materials and Methods 27

3.1 Sample description 27

3.2 Preliminary investigations on the population sample 28

3.3 Copy Number Variants 30

3.4 IBD analysis 31

3.5 Estimation of genetic similarities from IBD data 33

3.6 Cluster Analysis 34

3.7 Haplotype clustering 34

3.8 Whole-exome sequencing and variant filtering 35

3.9 Functional enrichment analyses 37

3.10 Analyses based on inference of variant sharing from IBD map 37

4. Results 39

4.1 Copy Number Variant analysis 39

4.2 The development of IBD mapping methodology 40
4.2.1 Preliminary findings: linkage analysis 40
4.2.2. Identification of Identical-By-Descent (IBD) chromosomal segments 40
4.3 Population studies from IBD data 43
4.3.1 Genome-wide IBD sharing and biological relationships between subjects 43
4.3.2 Cluster analysis 45
4.4. IBD mapping: tracking haplotypes shared by patients 48
4.5 Whole-exome sequencing 51
4.5.1 Variant annotation and filtering 51
4.5.2 Functional enrichment analysis 52
4.5.3 Gene-set enrichment analysis 55
4.5.3 Recurrence of functionally related variants in families 56
5. Discussion and conclusions 61
5.1 The development of a novel approach combining IBD mapping and whole-exome sequencing 61
5.2 Technical considerations about the IBD analysis 62
5.3 Population genetics studies using IBD data 63
5.4 Filtering of whole-exome sequencing data with IBD map 64
5.5 Functional enrichment analyses 65
5.6 Detailed investigation of variants in families 67
5.7 Strengths and limits of the study 69
5.8 Conclusions 72
6. References 73
7. Appendix 83
7.1 The 108 loci significantly associated with SCZ (Ripke et al., 2014) 83
7.2 The 10 loci significantly associated with BPD 88
7.3 List of local origins (Chioggia or Sottomarina) for the 115 ‘family founders’ 88
7.4 Detailed results of the functional enrichment analysis 92
7.5 Detailed information about the investigated variants 96
7.6 Primer list for Sanger sequencing of candidate variants 96

1. Summary

Introduction: Schizophrenia (SCZ) and bipolar disorder (BPD) are complex genetic disorders, each with a prevalence of 1% in the population worldwide. The high frequency and the severity of symptoms place them among the top 20 global causes of disability. Moreover, very few effective drugs are available, as their etiology is largely unknown. Early investigations have revealed a strong genetic contribution, estimated around 80%, and a consistent overlap between SCZ and BPD. Two main models have been proposed for their genetic architecture: the common disease-common variant and the common disease-rare variant. Several genome-wide association studies (GWASs) have demonstrated that common polymorphisms play indeed a role, but they can account only for half of the observed genetic variance in liability. Therefore, attention has turned towards the discovery of rare, high-penetrant risk variants. Recent studies have finally proven the validity of both models, but the works have revealed also how the identification of rare susceptibility factors could be trivial, given the high genetic heterogeneity of SCZ and BPD.

Aim of the study: This works aims at identifying rare risk variants for SCZ and BPD. A unique population sample is investigated, as all the patients come from the same closed community, living in Chioggia, close to the Venetian lagoon. This sample is characterized by an increased prevalence of SCZ and BPD, and about 150 familial cases were available. Considering the high rate of endogamy, a more homogenous genetic background is hypothesized, together with an enrichment in rare susceptibility factors, making the sample particularly valuable for the detection of these type of variants.

Methods and Results: First results excluded a major role for copy number variants (CNVs), the best candidate high-penetrant variants for the disorders to date. Subsequent strategies were then applied with the intent of tracking haplotypes shared between patients. Preliminary data on linkage analyses, however, had provided no significant results. Consequently, a novel approach was set up, to identify haplotypes identical by descent (IBD), considering the common ancestry. This latter analysis is free from segregation constrains, accounting also for the incomplete penetrance and the high genetic heterogeneity that characterize these complex disorders, and that likely had hampered linkage investigations. IBD analysis was carried out with *Relate* software and output data were processes with specifically developed PERL scripts.

Since IBD analysis was completely designed in this work, family information was initially used to verify the robustness of data. Indeed, genetic similarities estimated from IBD regions accurately reflected the biological relationships between pairs of individuals. Additionally, these estimates revealed a substantial relatedness between unrelated subjects, thus the population sample had some features of a population isolate. To get further insights into the connections among families, a cluster analysis was performed. Surprisingly, the outcome highlighted the presence of two main population subgroups, corresponding to the local origins of individuals ($p=2.12\ e\text{-}8$). Samples were in fact collected from two closed areas, named Chioggia and Sottomarina. The studied population could be thus considered as composed by two extended families.

Once the reliability of IBD data had been established, IBD mapping was sought as a strategy to track susceptibility loci for SCZ and BPD. An in-house pipeline was thus developed to obtain a genome-wide map of shared haplotypes among all the subjects. Interestingly, some of the haplotypes were specific for a phenotype (SCZ or BPD) or an origin (Chioggia or Sottomarina); however, a consistent number of loci were common to the two population clusters, indicating a certain level of admixture. IBD map identified haplotypes shared across multiple families and, remarkably, some combinations of loci co-segregating in different families. This last observation was considered particularly relevant as it could indicate possible interplays of genetic factors consistently contributing to risk.

From the total IBD map (IBD_{tot}), a subset of haplotypes were selected (IBD_{sel}), because possibly with a greater impact on the disorders, as they were shared by at least half of the patients of three or more families.

For an in depth investigation of IBD haplotypes, 17 affected subjects were chosen for whole-exome sequencing. Next-generation sequencing was carried out with Ion Torrent™ platform, while annotation and filtering were obtained by integrating *IonReporter*™ software and expressly created algorithms. 6,621 rare and novel variants were then mapped into IBD_{tot} haplotypes; of these, 8.4% were located in IBD_{sel} haplotypes. Interestingly, all these categories of variants were particularly affecting genes involved in axon guidance and synaptic transmission processes. More in details, variants were concerning definite nodes of neurite outgrowth and glutamatergic synapse functions, principally postsynaptic cascades activated by NMDA receptors.

Using IBD map, sharing and segregation of variants could be inferred for the non-sequenced individuals, thus allowing the investigation of the entire population sample. A family-based analysis showed a recurrence of rare or novel alleles affecting axon guidance and synaptic transmission in each pedigree. This scenario is coherent with a polygenic contribution to the disorders. Sharing analysis identified variants segregating across different pedigrees, overall revealing a puzzle of alleles, each with a proper sharing pattern, compatible with a certain genetic heterogeneity. Among the strongest candidate genes with synaptic functions are *GRM7* and *GRM8*, two metabotropic glutamate receptors, presenting two variants co-segregating with SCZ in the same family, and *CACNA1E* and *DLG4*, with variants shared across multiple pedigrees. Similarly, the *NCAM1* and *NCAN* genes, whose products directly interact in neurite formation, were carrying variants in all the patients within the same family.

Under the hypothesis of load of variants affecting the same biological process, it's possible that the risk would be substantially increased by specific combinations of a few functionally related alleles. Coherently, variants on different chromosomes were found co-segregating across multiple pedigrees. For example, a combination of two variants in the *ARHGAP32* and *CDH13* genes was detected in 5 patients of two families. Both these genes are involved in modulating actin cytoskeleton remodelling, the first in NMDA receptor mediated cascades, the latter in response to neuronal adhesion processes. They are thus fundamental in neuronal projection formation and development. Their co-occurrence may pinpoint possible sites of vulnerability, that, when simultaneously affected, confer a high susceptibility for SCZ/BPD.

Discussion and conclusions: In this work a novel approach for the investigation of the genetic architecture of SCZ and BPD was presented, exploiting IBD mapping and WES in a family-based population sample. The approach offered several advantages, such as the possibility to efficiently prioritize variants and to investigate the entire population from a subset of individuals. The IBD map, then, represents a flexible tool to address specific questions relative to candidate risk factors, thus allowing the testing of specific hypothesis, such as the existence of haplotype combinations.

Thanks to the integration of IBD and WES, some insights on SCZ/BPD etiology were obtained. More in details, specific processes related to neurosystem functions and development were implicated and some candidate variants were provided, therefore substantiating the role of rare variation in susceptibility. Considering the unique characteristics of the studied population, it's difficult to predict whether the same alleles could be detected in independent samples. Despite, in the light also of the general convergence of all types of genetic studies, the results could suggest possible genes or nodes of pathways that, when affected, increase vulnerability to these psychiatric disorders. Further replications in other studies would thus support their involvement, finally suggesting new therapeutic targets for these common disorders that cause such a high burden for the society.

2. Riassunto

Introduzione: la schizofrenia (SCZ) e il distrubo bipolare (BPD) sono malattie genetiche complesse, ciascuna con una prevalenza dell'1% nella popolazione mondiale. L'alta frequenza e la severità dei sintomi le rendono due delle prime 20 cause globali di disabilità. Inoltre, i farmaci disponibili sono pochi e solo parzialmente efficaci, conseguenza del fatto che l'eziologia di questi disordini è ancora largamente sconosciuta. Le prime evidenze hanno dimostrato una forte componente genetica, stimata intorno all'80%, e una consistente sovrapposizione tra SCZ e BPD. Due principali modelli sono stati proposti per descriverne l'architettura genetica, il primo definito 'malattia comune-varianti comuni' e il secondo 'malattia comune-varianti rare'. Numerosi studi di associazione sull'intero genoma (GWAS, *genome-wide association study*), hanno dimostrato l'effettivo coinvolgimento di polimorfismi comuni, tuttavia questi possono rendere conto solo di metà della varianza genetica stimata per i disordini. L'attenzione si è dunque spostata verso la ricerca di varianti rare. Recenti lavori hanno infine confermato la validità di entrambi i modelli genetici, ma hanno anche rivelato come l'identificazione di fattori rari di suscettibilità possa essere ardua, data l'estrema eterogeneità che caratterizza la schizofrenia e il disturbo bipolare.

Scopo della ricerca: Questo lavoro ha come obiettivo l'identificazione di varianti rare che possano conferire rischio per SCZ e BPD. A questo proposito, viene investigato un particolare campione di popolazione, che si contraddistingue per il fatto che tutti i pazienti provengono da una comunità chiusa, residente a Chioggia, una città della laguna veneziana. Il campione è caratterizzato da un'elevata prevalenza di SCZ e BPD e circa 150 casi familiari erano disponibili per lo studio. Considerando l'alto tasso di endogamia, l'ipotesi iniziale è quella di una maggiore omogeneità genetica, e in particolare un arricchimento di fattori rari di suscettibilità, che rendono questo campione di estremo valore per l'identificazione di questo tipo di varianti.

Metodi e risultati: i primi risultati ottenuti escludono un ruolo maggiore delle CNV (*copy number variant*), ritenute attualmente tra le migliori candidate per essere varianti rare di rischio per SCZ e BPD. Le strategie successive sono state quindi volte ad identificare aplotipi condivisi dai pazienti. Dati preliminari provenienti da analisi di linkage, tuttavia, non avevano fornito alcun segnale significativo. Di conseguenza, si è proceduto con la messa a punto di un approccio innovativo, per l'individuazione di aplotipi identici per discendenza (IBD, identical by descent), tenendo conto della comune origine dei pazienti. Quest'ultima analisi risulta svincolata da limiti imposti da metodi basati sulla segregazione ed è in grado di contemplare anche la penetranza incompleta e l'alta eterogeneità genetica che caratterizzano queste malattie complesse e che sono la causa dello scarso successo degli studi di linkare. L'analisi IBD è stata realizzata mediante il *software Relate* e l'output del programma è stato processato mediante *script* in linguaggio PERL, appositamente sviluppati.

Essendo l'analisi IBD completamente progettata in questo lavoro, le relazioni familiari sono state inizialmente utilizzate per verificare l'affidabilità dei risultati. La similarità genetica, stimata a partire dalle regioni IBD identificate, ricalcava accuratamente i legami biologici esistenti tra coppie di individui. In aggiunta, queste stime hanno rivelato un elevato grado di similarità tra soggetti non

imparentati, suggerendo che il campione analizzato avesse in effetti alcune caratteristiche di una popolazione isolata. Per ottenere ulteriori informazioni relative alle connessioni esistenti tra le famiglie, è stata eseguita un'analisi di *cluster*. Sorprendentemente, i risultati hanno evidenziato la presenza di due principali sottogruppi di popolazione, corrispondenti all'origine locale degli individui. (p=2.12e-8). I campioni erano stati infatti raccolti in due regioni attigue, Chioggia e Sottomarina. La popolazione studiata può essere dunque considerata come composta di due 'super-familglie'.

Stabilita l'affidabilità dei dati IBD, la mappatura IBD è stata impiegata con l'obiettivo di tracciare loci di suscettibilità per SCZ e BPD. Una specifica procedura bioinformatica è stata quindi sviluppata per ottenere una mappa genomica degli aplotipi condivisi dai soggetti. È interessante notare come alcuni degli aplotipi fossero specifici per un fenotipo (SCZ o BPD) o un'origine (Chioggia o Sottomarina); tuttavia, un consistente numero di loci era comune ai due *cluster* di popolazione, indicando un certo livello di mescolanza. Attraverso la mappa IBD, è stato possibile identificare aplotipi condivisi tra le famiglie e, soprattutto, alcune combinazioni di loci che co-segregavano in famiglie distinte. Quest'ultima osservazione è stata ritenuta particolarmente rilevante perché poteva indicare possibili interazioni di fattoti genetici che contribuiscono in modo sostanziale al rischio.

A partire dalla mappa totale IBD (IBD_{tot}), un sottoinsieme di aplotipi è stato selezionato (IBD_{sel}), in quanto possibilmente con un impatto maggiore nei disordini, perché condivisi da almeno metà dei pazienti di tre o più famiglie.

Per uno studio più approfondito degli aplotipi IBD, gli esomi di 17 soggetti affetti sono stati sequenziati. Il sequenziamento è stato eseguito mediante piattaforma IonTorrent™, mentre l'annotazione e il filtraggio delle varianti sono stati ottenuti tramite l'integrazione del *software IonReporter*™ con algoritmi espressamente creati. 6,621 varianti rare o nuove sono state mappate negli aplotipi IBD_{tot}; di queste, l'8.4% era localizzato in aplotipi IBD_{sel}.Da un punto di vista funzionale, queste varianti interessavano particolarmente geni coinvolti nei processi di *axon guidance* (formazione di proiezioni neuronali) e *synaptic transmission* (trasmissione sinaptica). Più nel dettaglio, le varianti cadevano in precisi nodi dei processi di '*neurite outgrowth*' e funzioni relative alla sinapsi glutamatergica, principalmente in cascate di segnale attivate dai recettori NMDA:

Utilizzando la mappa iBD, è stato possibile inferire la condivisione e la segregazione delle varianti anche in individui non sequenziati, permettendo così l'analisi dell'intero campione. Un'analisi a livello famigliare ha dimostrato una ricorrenza di varianti nei processi di *axon guidance* e *synaptic transmission* in ogni *pedigree*. Questo scenario è coerente con una natura poligenica dei disordini. L'esame della condivisione di tali varianti ha rivelato nel complesso un *puzzle* di alleli, ciascuno con un proprio schema di condivisione, compatibile con un certo livello di eterogeneità genetica. Tra i geni candidati con funzioni sinaptiche si ritrovano *GRM7* e *GRM8,* due recettori metabotropici del glutammato, che presentano due varianti che co-segregano con la SCZ nella stessa famiglia, e *CACNA1E* e *DLG4,* con varianti condivise attraverso multiple famiglie.

Analogamente, i geni *NCAM1* e *NCAN,* i cui prodotti interagiscono direttamente nel processo di formazione di proiezioni neuronali, erano colpiti da varianti in tutti gli affetti di una stessa famiglia.
Sulla base dell'ipotesi di un carico di varianti che interessano lo stesso processo biologico, è possibile che il rischio possa essere sostanzialmente incrementato da specifiche combinazioni di alcuni alleli correlati funzionalmente. Coerentemente, varianti su cromosomi differenti co-segregavano in multipli *pedigree.* Per esempio, una combinazione di due varianti nei geni *ARHGAP32* e *CDH13,* è stata rilevata in 5 pazienti di due famglie. Entrambi questi geni sono coinvolti nel rimodellamento del citoscheletro di actina, il primo in cascate mediate da recettori NMDA, il secondo in risposta a processi di adesione neuronale. Sono quindi fondamentali per la formazione di proiezioni e nello sviluppo neuronale. La loro co-occorrenza potrebbe denotare possibili siti di vulnerabilità, che, quando simultaneamente colpiti, conferiscono un'alta suscettibilità per SCZ/BPD.
Discussione e conclusioni: In questo lavoro, è stato presentato un nuovo approccio per l'investigazione dell'architettura genetica di SCZ e BPD, che sfrutta la combinazione di una mappa genomica IBD con il sequenziamento dell'intero esoma in un campione di popolazione composto da numerose famiglie. L'approccio ha offerto numerosi vantaggi, come la possibilità di prioritizzare in modo efficiente le varianti e di analizzare un intero campione di popolazione a partire da un piccolo gruppo di soggetti sequenziati. La mappa IBD, poi, rappresenta uno strumento flessibile per testare specifiche ipotesi relative a fattori di rischio candidati, come l'esistenza di combinazioni di aplotipi.
Grazie all'integrazione di IBD e sequnenziamento dell'esoma, alcune informazioni sull'eziologia di SCZ e BPD sono state ottenute. Nello specifico, alcuni processi correlati a sviluppo e funzionalità del sistema nervoso sono stati implicati e alcune varianti candidate sono state evidenziate, supportando quindi il ruolo di varianti rare nella suscettibilità alle due malattie. Considerando le caratteristiche uniche della popolazione investigata, è difficile predire se gli stessi alleli possano essere rilevati in altri campioni indipendenti. Tuttavia, alla luce anche della generale convergenza degli studi genetici, questi risultati potrebbero suggerire possibili geni o nodi di processi che, quando mutati, possono incrementare la vulnerabilità a questi disordini psichiatrici. Ulteriori replicazioni sarebbero quindi utili nel confermarne il coinvolgimento, finalmente suggerendo nuovi bersagli terapeutici per queste malattie comuni che rappresentano un tale carico per la società.

1. Introduction

1.1 Preface

Schizophrenia and bipolar disorder: an overview on the study of psychiatric illnesses

Mental health constitutes an actual issue in modern society: according to the World Health Organization (WHO) psychiatric disorders altogether represent the leading cause of disability worldwide (Department of Health Statistics and Information Systems WHO, Geneva, 2013). Schizophrenia (SCZ) and bipolar disorder (BPD) contribute significantly to the estimates, as they are individually placed in the top 20 positions of such a rank of disability causes. This strong impact on global health is ascribable both to the high frequency and the extreme severity of SCZ and BPD. The lifetime prevalence for each disorder is about 1% in the general population (American Psychiatric Association, 2013), with very small differences across countries and ethnic groups. Affected individuals have difficulties in basic personal and social tasks, including self caring and keeping a regular job. In addition, SCZ and BPD are mostly enduring illnesses, so that lifelong assistance is required, creating a consistent burden not only for patients and their families, but also for the society as a whole. SCZ in particular is considered one of the costliest diseases both in terms of human and financial resources (van Os & Kapur, 2009).

Psychosis and mood alterations have been reported as early as ancient Greece age (Möller, 2003), but only with the work of the German psychiatrist Emil Kraepelin, at the beginning of the 20th century, have SCZ and BPD been recognized and classified as proper diseases (Kraepelin, 1899). In a series of books published between 1893 and 1899, Kraepelin investigated mental conditions with the same criteria used for physical ones, identifying specific symptoms and courses. He further distinguished two main groups of insanities: *Dementia Praecox* and manic-depressive illness. The term "*Dementia Praecox*" was used to describe a psychotic disease characterized by rapid cognitive disintegration, that, unlike senile dementia, occurred early in life. Conversely, manic-depressive disorder arose with repeated mood shifts, but without any cognitive involvement. These two categories broadly correspond to the modern SCZ and BPD definitions. *Dementia Praecox* was later renamed as schizophrenia (literally "split mind") by Eugen Bleuer (Bleuer, 1908), who hypothesized at the basis of the disorder a detachment of cognition from emotions, behavior and volition. The term "bipolar disorder" was instead introduced in 1959 by Karl Kleist and Karl Leonhard to highlight the distinction from major depressive disorder, also called unipolar disorder (Angst & Marneros, 2001)

The clear separation between an intellectual functioning disorder (SCZ) and a mood one (BPD) has marked a milestone in clinical practice and is largely known today as the "kraepelinian dichotomy". Although still widely used for diagnostic purposes, this assumption has been recently challenged by growing evidence of an overlap between the two disorders. Whether the

“kraepelinian dichotomy” should be revised is currently a highly debated topic among psychiatrist (Craddock & Owen, 2010; Tesli et al., 2014).

The conceptual definition of SCZ and BPD in early 1900 propelled the examination of their epidemiology in the worldwide population. It became immediately evident that the disorders aggregate in families, suggesting a genetic contribution in liability. This observation has been repeatedly confirmed by a line of studies, initiated by Franz Kallmann around 1930, on families, twins and adopted children. All these works have demonstrated that the genetic influence on SCZ and BPD is in fact substantial (Smoller & Finn, 2003; Patrick F Sullivan, Kendler, & Neale, 2003) and have converged also on a common etiology for the two disorders (Lichtenstein et al., 2009).

In the light of these remarkable data, the genetics of SCZ and BPD has been extensively investigated. However, the genetic architecture has proven to be complex and extremely heterogeneous, with hundreds of genes possibly involved in the pathogenesis of both disorders (Harrison, 2015). Nevertheless, the understanding of the mechanisms underlying these psychiatric conditions is still of extreme importance for diagnosis and, especially, for treatment, given the unfavourable prognosis (American Psychiatric Association, 2013). The number of available drugs is indeed limited, with only partial effectiveness and strong side effects (Werner & Coveñas, 2015). For example, more than 50% of schizophrenic patients show some degree of resistance to antipsychotics (Falkai et al., 2015); in contrast, mood stabilizers can in general counteract manic symptoms, but are less efficient against the depressive state (Frye et al., 2014). Therefore, the identification of involved genes could help uncover new potential and possibly more specific therapeutic targets. In a long term perspective, the determination of the genetic profile would allow to predict the risk for an individual to develop a disease and for affected subjects to get personalized therapies (Purcell et al., 2009).

1.2 Clinical aspects of schizophrenia and bipolar disorder

1.1.1 Schizophrenia

Diagnosis of psychiatric disorders is generally defined following international guidelines. One of the most used reference text on this subject is the Diagnostic and Statistical Manual of mental disorder, published in its fifth edition (DSM-V, American Psychiatric Association, 2013).

SCZ is a devastating mental disorder that affects cognition, behaviour and emotions. According to DSM-V, it is characterized by five main symptom domains. The first four domains are generally known as positive symptoms and include delusions, hallucinations, disorganized thinking and grossly disorganized or abnormal behaviour. Delusions are fixed beliefs that are held with certainty even in light of conflicting evidence; the most common delusions are persecutory, when patients are convinced they're going to be harmed by an individual or an organization. Hallucinations are defined as perception-like experiences that occur without an external stimulus; they can involve any of the five senses, but almost all the schizophrenic subjects undergo auditory hallucinations and declare to hear voices in their heads clearly distinguishable from their thoughts. Disorganized thinking manifests itself with disorganized speech, revealed by the unjustified switch from one topic to another, or even incomprehensible talking. Finally, disorganized or abnormal behaviour may range from childlike "silliness" to extreme agitation; catatonia, a marked decrease in reactivity to the environment, is also classified among this last category.

The fifth symptom domain consists of negative symptoms, that affect specifically the emotional and social sphere. Typical negative symptoms are a diminished emotional expression and avolition, a decrease in the motivation to perform self-directed actions (e.g. going to work or other general routine activities). Social withdrawal and anhedonia, the inability to feel pleasure from positive stimuli, are common as well.

SCZ usually emerges with the first psychotic episode, identified by the persistent presence of at least one positive symptom. The onset of the disorder is in late adolescence to early adulthood, with a peak age in the mid-20s; childhood and late adulthood cases are also described, but are extremely rare. The psychotic episode is frequently preceded by a prodromal phase and followed by a residual one. Both these phases are characterized by mild positive symptoms and, mostly, negative ones. Another important feature that appears evident in non-psychotic periods is cognitive impairment, particularly in attention and memory functions, as originally observed by Kraepelin himself (Falkai et al., 2015). Although not considered as an inclusion criterion in DSM-V, cognitive deficits are present in the majority of patients and are often the first sign of the disorder; for example, children who later develop SCZ have subtle intellectual and motor delays (van Os & Kapur, 2009).

The outcome of SCZ is generally unfavorable. Life expectancy is reduced of about 12-15 years, not only because of suicides, that affect 5-6% of patients, but also due to the poor self-care (e.g. poor diet, little exercise, substance abuse) and the lifelong use of drugs (van Os & Kapur, 2009), that expose individuals with SCZ to metabolic side effects (e.g. diabetes or cardiovascular diseases). Furthermore, only a small fraction of patients report a full recovery. Despite antipsychotic treatments, about a third of affected individuals experience relapses after the first psychotic episode (van Os & Kapur, 2009). Even when drugs are effective against positive symptoms, the residual phase frequently turns into a chronic disorder, because currently available drugs are unable to counteract negative symptoms. Moreover, cognitive alterations become a stable cognitive impairment and a progressive intellectual deterioration occurs. Thus, negative and cognitive symptoms are those mainly contributing to disability, so that patients require constant assistance in everyday life, making SCZ one of the most severe mental disorders worldwide.

1.1.2 Bipolar disorder

BPD is an affective disorder characterized by repeated shifts in mood, that influence thinking and behavior. Two main types of BPD are distinguished, namely type I and II (American Psychiatric Association, 2013).

The typical sign of BPD I is mania. During a manic episode, mood is abnormally and constantly elevated, expansive or irritable. Patients experience an increased energy and a decrease need for sleep, a pressure of keep talking, but also distractibility and agitation, as well as an attraction for high-risk, dangerous activities (reduced risk perception). Besides, a rapid alternation between euphoria, dysphoria and irritability can be seen. All these disturbances strongly affect social and occupational life, and in some cases hospitalization is required. The manic episode by itself is sufficient for the diagnosis of BPD I, but more than 90% of cases have recurrent mood episodes. The typical manifestation is a cyclical change of mood from manic, hypomanic and/or depressive states, with variable intervals of normal mood. Hypomania denotes a less severe mania, where symptoms are similar but not enough to compromise everyday life. Depression is instead a mixture of feelings encompassing sadness, emptiness and hopelessness, combined with a markedly reduced interest in any activity, insomnia or hypersomnia, loss of energy and recurrent thought of death that persist for several days.

In BPD II, proper manic episodes are absent and the predominant feature is major depression, in combination with at least one episode of hypomania. Although hypomanic symptoms are less serious than manic ones, BPD II can't be considered a milder form of the disorder, as patients suffer from more enduring and disabling depressive episodes over time. In addition, the risk of relapses is higher than in BPD I.

The onset of BPD is generally in the late teens, but first manic, hypomanic or depressive symptoms have been described at any age. Patients usually benefit from mood stabilizer drugs,

although at least 15-30% of them do not completely recover from mood episodes and show some degree of dysfunction even in normal mood state (Frye et al., 2014). As previously described, relapses are frequent and BPD becomes a lifelong, chronic disease. Similarly to SCZ, BPD is thus a disabling condition, that compromises personal, social and occupational life. Functional consequences contribute to the exacerbation of depression; as a proof, about one-third of cases attempt suicide (American Psychiatric Association, 2013).

1.1.3 The clinical overlap of schizophrenia and bipolar disorder

In the current diagnostic system, SCZ and BPD are considered distinct nosological entities, following the "kraepelinian dichotomy" (American Psychiatric Association, 2013). But in clinical practice, medical cases are particularly heterogeneous and symptoms are not exclusive of a disorder, transcending diagnostic boundaries. More than 50% of bipolar subjects, in fact, experience a psychotic episode in their lifetime; similarly, mood dysregulation is not rare in schizophrenic individuals and depression is a main feature of the characteristic negative symptoms, that become predominant in later stages (Lin & Mitchell, 2008). Also cognitive deficits, initially described as typical of SCZ, are observed in BPD, although in a less ubiquitous and profound extent (Simonsen et al., 2011). Moreover, both schizophrenic and bipolar cases are responsive to the same medications, such as antipsychotic drugs. As a result of this, it's not uncommon that a patient receives multiple, contrasting diagnoses in the course of his life (Song et al., 2015). Further, a third disorder has been defined, the schizoaffective disorder (SZA), with the intent of grouping the variety of subjects with consistent signs of both SCZ and BPD; whether SZA represent a distinct entity is still controversial (American Psychiatric Association, 2013). In this light, the validity of the "kraepelinian dichotomy" has been strongly questioned, as the strict diagnostic categorization could be trivial and too simplistic to completely describe all the patient manifestations (Cardno & Owen, 2014). In contrast, a psychiatric continuum model has been proposed (Craddock & Owen, 2010; Möller, 2003), where SCZ and BPD symptoms are considered as part of a spectrum, ranging from psychosis to mania and depression. Even if such a hypothesis could more realistically represent the heterogeneity of symptoms found in patients, its application in clinical practice is not yet feasible. Nosological categories are indeed still acknowledged as the best instrument to address course and treatments. Neurobiological understanding of the disorders is actually too limited to reach a more refined diagnosis and at present deeper insights into mechanisms underlying etiology are therefore urgently required (Tesli et al., 2014).

1.3 Unravelling the genetics of schizophrenia and bipolar disorder

1.3.1 Epidemiological studies

Epidemiological studies on SCZ and BPD have provided the first compelling evidence of a strong genetic component in both disorders. The leading observation is that both SCZ and BPD run in families (Cardno & Owen, 2014; Smoller & Finn, 2003). First-degree relatives of schizophrenic probands have almost 10 times higher risk to develop the disorder compared to the general population (Lichtenstein et al., 2009); analogous data have been reported for BPD (Smoller & Finn, 2003). Interestingly, risk decreases with the genetic distance from the affected relative (Song et al., 2015). The confirmation that familiarity is due to genetic factors has come from twin and adoption studies. In multiple cohorts of SCZ and BPD twins it has been observed a significant higher concordance in affection status for monozygotic brothers compared to dizygotic ones (Cardno et al., 2002; Kieseppä et al., 2014). Both type of twin pairs are assumed to share the same environment, including *in utero* and perinatal conditions, thus the higher concordance rate in genetically identical subjects clearly indicate the role of genes in determining the risk. Similarly, adopted away children whose biological parents or siblings have SCZ or BPD have at least a 4-fold risk to be diagnosed with the same disorder (Lichtenstein et al., 2009), again sustaining the importance of genetics over the shared familial environment.

On the basis of these results, the magnitude of genetic influence have been estimated, by statistically modeling liability to disorders. Many studies agree in pointing out that heritability, defined as the fraction of phenotypic variance explained by genes, ranges from 60 to 80% for both SCZ and BPD (Cardno et al., 2002; Song et al., 2015; Sullivan et al., 2003). The remarkable outcome have definitely highlighted that both environmental and genetic factors are involved in these psychiatric disorders, but the latter have the stronger impact in determining SCZ and BPS susceptibility (Lichtenstein et al., 2009).

1.1.4 Search for common risk variants

Since the establishment of the high heritability, considerable efforts have been undertaken to identify the loci involved in SCZ and BPD and their relative contribution, thus the genetic architecture of the disorders. The hypothesis of a major locus with a Mendelian-like inheritance, have been almost immediately rejected, because incoherent the high prevalence of the disorders in the population (Kerner, 2015). Preliminary investigations have suggested that SCZ and BPD are actually polygenic; in addition, the genetic and environmental interplay in defining the phenotype clearly classifies SCZ and BPD as complex disorders (Lichtenstein et al., 2009; Sullivan et al., 2003).

Two genetic models, initially opposing, have been proposed: the 'common disease-common variant' and the 'common disease-rare variant' (Doherty, O'Donovan, & Owen, 2012). According to the 'common disease-common variant' hypothesis, the genetic predisposition to SCZ and BPD is conferred by multiple alleles with a relatively high population frequency (typically >1%), each with a small effect size. Conversely, the 'common disease-rare variant' assumption proposes that much of the disease susceptibility is due to rare, high-penetrant alleles. The debate around the two models has driven the research of genetic factors in the past 20 years.

The 'common disease-common variant' model represented the dominant theory in early phases of SCZ and BPD investigation. The presence of risk factors with low penetrance, likely not subjected to negative selection, was seen as the most plausible scenario considering such a high and constant prevalence of the disorders worldwide (Kerner, 2015).

The conventional methodological approaches for tracking common variants are association studies, that aim at finding alleles significantly more frequent in a cohort of cases versus a cohort of controls. First association studies were conducted by genotyping a small number of SNPs, usually encompassing a single candidate gene in samples of a few hundred cases and controls. These studies have proposed the involvement of a few genes, the most famous ones being *DISC1* (disrupted in schizophrenia 1), *DRD2* (dopamine receptor D2) and *COMT* (Catechol-O-Methyltransferase). The initial enthusiasm was followed by an alternation of positive and negative replications that challenged the consistency of the results (Farrell et al., 2015).

Research received a new impulse with the development of genome-wide association studies (GWASs), that provided the possibility to systematically inspect the 'common disease-common variant' hypothesis (The Wellcome Trust Case Control Consortium, 2007). GWASs rely on the genome-wide genotyping of single nucleotide polymorphism (SNP) markers, feasible with microarray technology. Each marker is tested for association and, since SNP distribution is pervasive, the entire genome can be interrogated in the same experiment, removing biases of candidate-driven analyses. GWASs have thus the advantage of not requiring any *a priori* knowledge on disease biology (Chen et al., 2015; Neale & Sklar, 2015). A fundamental principle in these studies is linkage disequilibrium (LD), defined as the genotype correlation between different SNPs, generally in close proximity, derived from evolutionary conservation or drift (Neale & Sklar, 2015). The degree of LD is measured with the r^2 metric, ranging from 0 (no LD) to 1 (complete LD). LD delineates short haplotype blocks of genotyped and non-genotyped alleles, of which the genotyped SNPs are representative (tag-SNPs). Hence the association with a specific marker must be intended as the association with the genomic locus tagged by the tested SNP (Visscher et al., 2012). The effect size in conferring risk is instead reflected by the odd ratio (OR) of allele frequencies in cases and controls.

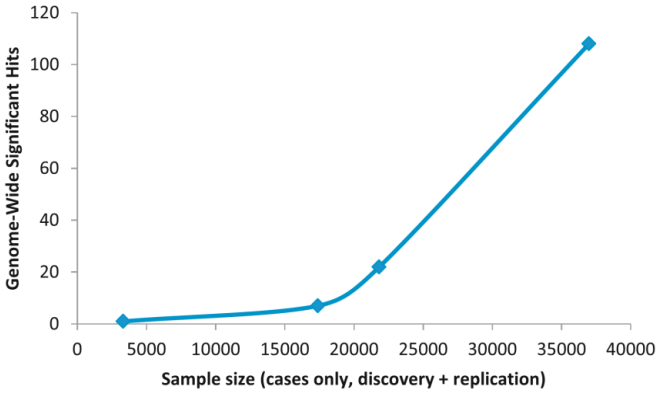

Figure 1.1: Relationship between sample size increase and detection of significant loci in GWASs of SCZ performed by PGC. From Need & Goldstein 2014

The modest success of the first GWASs, where a few or no associations were detected (O'Donovan et al., 2008; Sullivan et al., 2008), have depicted a scenario where common risk alleles individually have very small effects, with typical ORs around 1.1 and always inferior to 1.5 (Doherty et al., 2012; Harrison, 2015). Therefore, a large number of these susceptibility factors should have been implicated in SCZ and BPD etiology. In 2009, Purcell et al. provided indeed the primary demonstration of the extreme polygenic nature of SCZ and BPD (Purcell et al., 2009). Although no single marker could be implicated in this GWAS, cases can be reliably distinguished from controls considering the cumulative effect of hundreds of potentially involved alleles, weighted for their effect size. This cumulative effect, named polygenic risk score, has been also the first proof that common variants contribute to liability. A consequence of the presence of multiple, mostly independent, risk factors is the low probability of sharing an identical susceptibility allele, causing a high heterogeneity of patients. This suggested that the failure of the first attempts was probably due to small sample sizes and highlighted the need for large cohorts to reach the required statistical power. (Chen et al., 2015). In light of this evidence, worldwide collaborations have been established, the largest and most effective being the Psychiatric Genomic Consortium (PGC), founded in 2007 (Sullivan, 2010). From 2011 onward, the schizophrenia division of PGC (PGC-SZ) started to publish meta- and mega-analyses on joint samples, achieving important results (Ripke et al., 2011; Ripke et al., 2013) (Figure1.1). Quality criteria were also established to get trustworthy associations: a genome-wide p-value threshold of $5x10^{-8}$ and a minimum sample size of 10,000 considering cases and controls. The effectiveness of these approaches is summarized in the most recent and comprehensive study (Ripke et al., 2014), that includes almost all the known SCZ samples of Caucasian ancestry worldwide, for a total of 36,989 cases and 113,075 controls. In this mega-analytic GWAS, 108 independent loci were identified, 83 of which had not been previously reported (see appendix 7.1). The results also replicated all but 5 known SCZ loci that meet modern quality criteria; interestingly, neither *DISC1* nor *COMT* genes, initially retained among the strong candidates, found any support. These 108 loci represent today the state of the art on common variants involved in SCZ (figure 1.2).

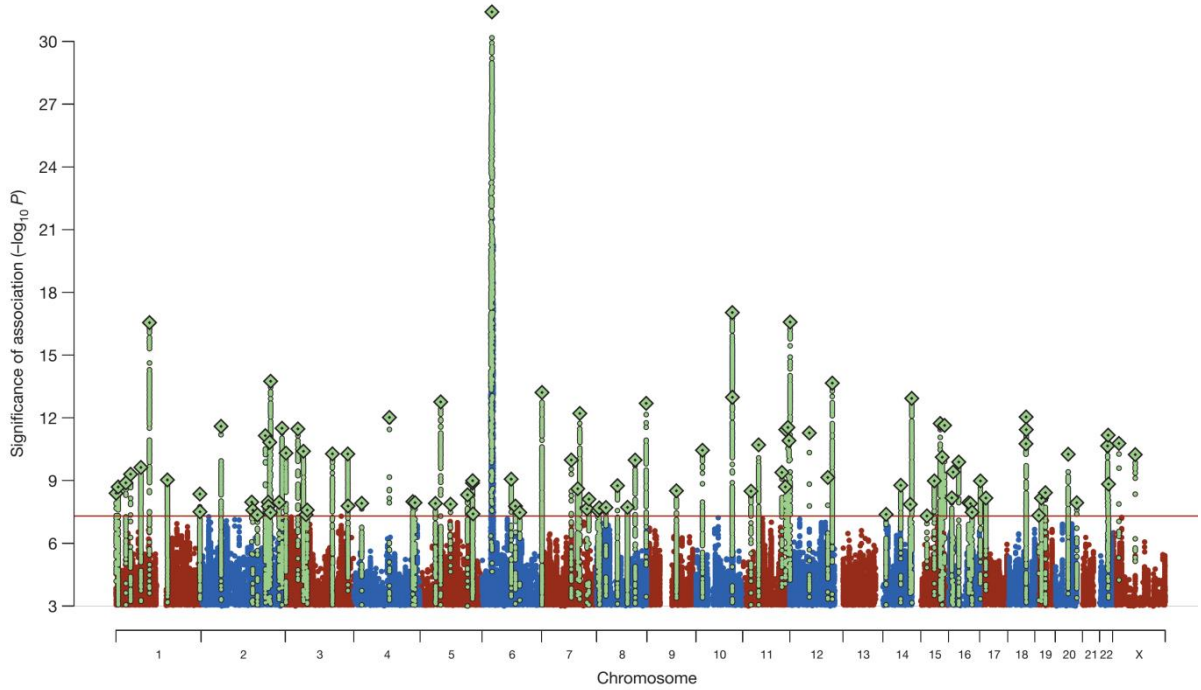

Figure 1.2: Manhattan plot of the largest GWAS on 36,989 SCZ cases and 113,075 controls. Green diamonds indicate the 108 statistically significant loci. From Purcell et al., 2014.

GWASs on bipolar cohorts have been less successful, a difficulty attributable again to the relatively limited number of samples collected so far (Neale & Sklar, 2015). Despite a number of published studies, to date only 10 loci have reached genome-wide significance (Chen et al., 2013; Cichon et al., 2011; Ferreira et al., 2008; Green et al., 2013a; Green et al., 2013b; Sklar et al., 2011; Mühleisen et al., 2014) (appendix 7.2). Indeed, the largest study performed to date analyzed 9,747 cases and 14,278 controls, detecting 4 loci (Mühleisen et al., 2014); these numbers are actually comparable to the ones of the first PGC-SZ publication (Ripke et al., 2011), both in terms of samples and identified loci (Figure 1.1). Thus, the increase of sample sizes will probably reveal new associations.

Although GWASs represent a valuable tool for the identification of risk alleles, their main limitation is that, once an associated locus is detected, they did not provide information about the functional risk variant lying in it. Among the 108 SCZ loci, for instance, 15 have no known gene nearby, and 36 include more than 3 genes; in addition, for each locus, tens to hundreds of alleles are involved (Ripke et al., 2014). This observations suggest that most of the markers showing association are not likely the direct causative variant (Chen et al., 2015). On the other hand, the implication of other alleles inside a specific locus has proven a hard task, considering the difficulty in assigning functional meanings to each sequence change (Need & Goldstein, 2014). Therefore, even if a collection of statistically compelling loci is available for both SCZ and BPD, no common allele, or gene, has been definitively confirmed to date.

Another issue arose with the remark that the identified loci could explain only a small fraction of the heritability estimated from epidemiological studies. For example, the 108 loci together can

account only for 3.4% of the variance in liability due to genetic factors (Ripke et al., 2014). Empirical simulations demonstrated that the whole common variants, including those not yet identified for power issues, explicate about one third of heritability (Lee et al., 2012; Purcell et al., 2009; Ripke et al., 2013). Since the currently available microarrays cover about 70% of the existing common variation, it's possible to affirm that common risk alleles can explain from one third to a half of the genetic susceptibility to SCZ and BPD (Lee et al., 2012). Half of the heritability remains thus theoretically unaccounted. In addition, is has been shown that common markers show very little LD with rare variation (Visscher et al., 2012); for technical reasons, then, rare alleles are invisible to GWASs and can't drive any association signal. This have led to the conclusion that the 'common disease-common variant' model should be integrated with the 'common disease-rare variant' one (Doherty et al., 2012; Sullivan, Daly & O'Donovan, 2012).

1.1.5 Search of rare risk variants

Following the demonstration that common variants alone are not sufficient to explain SCZ/BPD susceptibility , in the last years attention has turned toward the discovery of rare alleles. After re-evaluation, in fact, also the hypothesis of multiple rare risk factors has been considered compatible with common disorders, due to the polygenic inheritance (Kerner, 2015): the high prevalence would be a reflection of a large number of different rare variants. These rare alleles would be likely more penetrant than common ones, thus negatively selected and maintained at low frequencies in the population (George Kirov, 2015). Two types of rare variants have been investigated so far: copy number variants (CNVs) and sequence variants.

1.3.2.1 Copy number variants

CNVs are duplications or deletions of DNA segments with respect to the reference genome, usually in contiguous positions. The size of these structural variants ranges from 1 kb to several Mb, consequently they're submicroscopic and hardly identified with classical karyotyping techniques (Feuk, Carson, & Scherer, 2006). The majority of CNVs are part of the common variation of the human genome, with a frequency of more than 1%, and defined as copy number polymorphisms (CNPs) (Malhotra & Sebat, 2012).

Rare CNVs, instead, have long been known to be involved in neurodevelomental disorders, causing intellectual disability and autism spectrum disorder (Malhotra & Sebat, 2012). The examination of structural variants in SCZ and BPD have been trigged by some precedents from cytogenetic studies: in patients with chromosome deletions of 22q11.2, causing the DiGeorge syndrome, schizophrenic symptoms were recurrently found (Murphy, Jones, & Owen, 1999). Also *DISC1,* historically the first SCZ gene*,* was identified because disrupted by a balanced translocation (Blackwood et al., 2001). Moreover, CNVs had the features of potentially high-penetrant variants because, unlike SNPs, they usually involve several exons or even entire

genes, causing rearrangements affecting coding sequences or dosage imbalances. With the improvement of genotyping platforms, bionformatic tools have been set up for the detection of CNVs exploiting SNP intensity and genotype data. As a result, CNVs have been systematically investigated in SCZ and BPD, in parallel with GWASs, and more recently with the aim of formally testing the 'common disease-rare variant' hypothesis (Malhotra & Sebat, 2012).

An increased burden of rare (<1%) and large (>100 kb) duplications and deletions in SCZ cases compared to controls (Purcell et al., 2008; Walsh et al., 2008). Motivated by this evidence, the search of associations with specific CNVs produced great amount of data, the majority of them hardly replicated; the main reason for inconsistencies is that these variants are extremely rare, so large datasets are required to distinguish truly associated CNVs from neutral ones (Rees et al., 2014). To date, 11 loci harbouring duplications and/or deletions have been confidently recognized, ranging from 120 kb to 4 Mb (Kirov, 2015) (Table 1.1); as hypothesized, their effect size is higher than common variants, with ORs from 2 to more than 50 (Kirov, 2015; Rees et al., 2014). In accordance, the most penetrant SCZ variants known so far are two CNVs, the 3q29 deletion (estimated penetrance 18%) and the previously cited 22q12 deletion (estimated penetrance 12%) (Kirov, 2015).

Locus	Size (kb)	*N* genes	CNV frequency in %			ORs for SZ (95% CI)
			Controls† (47,686–81,821)	SCZ† (12,029–21,450)	BPD† (4,288–9,129)	
1q21.1 del	820	11	0.021	0.17	0.033	8.3 (4.6–15)
1q21.1 dup	820	11	0.037	0.13	0.099	3.45 (1.9–6.2)
NRXN1 exonic del	Variable	1	0.02	0.18	0	9.0 (4.4–18.3)
3q29 del	1610	21	0.0014	0.082	0.025	57.6 (7.6–438.4)
WBS dup	1400	28	0.0058	0.066	0	11.3 (2.6–49.9)
15q11.2 del	290	4	0.28	0.59	0.17	2.2 (1.7–2.7)
PWS/AS dup	3610	13	0.0063	0.083	0	13.2 (3.7–46.8)
15q13.3 del	1350	7	0.019	0.14	0.043	7.5 (4.0–14.2)
16p13.11 dup	790	8	0.13	0.31	0.011	2.3 (1.6–3.4)
16p11.2 dup‡	560	26	0.03	0.35	0.13	11.5 (6.9–19.3)
22q11.2 del	1240	40	0	0.29	0.012	NA (28.2–∞)

Table 1.1: The 11 loci harboring CNVs consistently associated with SCZ. † numbers in brackets report sample sizes. ‡ indicates the CNV associated also with BPD. Adapted from Kirov, 2015.

SCZ patients and, to a lesser extent, BPD subjects show a reduced fitness, thus high penetrant variants such as CNVs undergo strong negative selection (Power et al., 2013). Besides, the recurrence of specific CNVs and the constant prevalence of the disorders implies that these variants are continually re-introduced in populations; consequently, a fraction of rare alleles in general must be *de novo* mutations, rather than inherited from parents. Indeed a high mutational rate (1:4000-1:20000 live births) has been reported for CNVs, due to the presence of flanking low-copy repeats, causing non allelic homologous recombination events (Kirov, 2015). Coherently, *de*

novo CNVs have been found to be more frequent in SCZ cases than controls (Kirov et al., 2012); these rearrangements have substantiated the previously reported CNV loci (table 1.1).
The role of structural variants in BPD is more controversial. In measuring CNV burden, opposite outcomes have been published (Grozeva et al., 2013; D. Zhang et al., 2009); in any case, the load of these variants seems to be lower than SCZ patients, both in terms of quantity and of size (Shinozaki & Potash, 2014). Analogously, *de novo* CNVs are only slightly more frequent in cases (Malhotra et al., 2011). CNV distribution in BPD cohorts is similar to the controls (Table 1.1) and the only CNV consistently associated is a duplication in the 16p11.2 locus (Green et al., 2016). The potential lower pathogenicity in BPD would be also compatible with the lower degree of cognitive defects seen in bipolar subjects compared to schizophrenic ones; large and rare CNVs, in fact, has often been associated with reduced cognitive performance, not only in pathological conditions such as neurodevelopmental disorders, but also in healthy subjects (Kirov, 2015).

1.3.2.2 Sequence variants in case-control samples

With the advent of next generation sequencing (NGS) technologies, the sequencing of whole exome (WES) or even genome (WGS) has become feasible, with affordable costs (Shinozaki & Potash, 2014). Sequencing data have permitted for the first time the investigation of a class of variants impossible to detect with SNP-genotyping: rare sequence variants, that include single nucleotide variants (SNVs) and small insertions/deletion (indels).
The analysis of such rare alleles in case-control cohorts has proved arduous. As previously described for CNVs, in fact, a main issue is the discrimination of low frequency variants from neutral ones. Additionally, unlike structural variants, sequence variants have shown an unexpected heterogeneity (Purcell et al., 2014). Thus huge samples sizes are necessary to achieve sufficient statistical power; for this reason, association studies with sequencing data have been mainly performed so far on schizophrenic cases.
To overcome these problems, first studies focused on the identification of *de novo* coding mutations in family trios of affected probands with unaffected parents. Justified by the hypothesis of a balance of negative selection, the approach offers in fact an important technical advantage: the possibility to focus on a small subset of variants. Several works have supported an increased frequency of potentially damaging *de novo* alleles in SCZ probands compared to controls (Awadalla et al., 2010; McCarthy et al., 2014; Xu et al., 2012). These results have been though challenged by the largest study performed so far, involving 617 SCZ trios and 731 control ones (Fromer et al., 2014). Fromer and colleagues detected no increased rates of *de novo* mutations in cases, rather an enrichment of genes involved in synaptic functions among the ones affected by these class of mutations. Beside this striking result, no single variant or gene could be significantly associated with the disorder.
Despite their proven involvement, anyway, *de novo* mutations don't explain the high heritable risk of SCZ. In this regard, the only real attempt to investigate all rare variants, including the inherited

ones, was published in 2014 (Purcell et al., 2014). In this milestone paper, Purcell et al. demonstrated a polygenic burden of rare disruptive alleles in SCZ patients by comparing WES data from 2,536 cases and 2,543 controls. Interestingly, this burden was driven by genes in synaptic processes, indicating that rare variants could actually have a role in SCZ etiology. Analogously to *de novo* approaches, however, no single allele could be significantly associated, even when restricting the analysis to subsets of genes functionally relevant with the disorder.
Therefore, notwithstanding the great potential, the utility of case-control cohorts in detecting rare risk factors is limited with actual sample sizes and significant results can be detected only when collapsing variants at gene levels.

1.3.2.3 Sequence variants in population- and family-based samples

As outlined in the previous paragraphs, research of genetic risk factors for SCZ and BPD has mainly focused on large case-control cohorts. But considering the emerging limitations in the context of rare susceptibility variants, familial sample gained an increasing importance. Indeed, multigenerational families with several affected individuals offer some advantages to study rare variants as linkage analyses can be exploited to track candidate loci carrying risk alleles. Interesting results have been achieved for example for Alzheimer's disease, again after the discovery of uncommon familial cases with autosomal dominant pattern of inheritance (Sullivan et al., 2012).
Several genome-wide linkage scans on familial samples have been published both for SCZ and BPD each reporting suggestive loci, hardly replicated between studies (Lewis et al., 2003). These results have stressed once more the high genetic heterogeneity of these disorders, hampering the outcome of linkage analyses, unable to detect a single, segregating major gene (Neale & Sklar, 2015).
In this context samples coming from genetic isolates would provide potential advantages. A genetic isolate is a population that, for geographical and/or cultural reasons, has undergone a very low admixture (Escamilla, 2001). Isolates generally originated from a small number of founders who experienced a marked expansion, accompanied by high rates of inbreeding. Since all the alleles descended from a few ancestors, these samples are characterized by a reduced genetic variability and by unique allele sets (genetic drift) (Peltonen et al., 2000). In particular, rare variants carried by founders, if not negatively selected, are pushed to higher frequencies; this is true also for disease variants, as shown by the higher prevalence of some rare disorders in specific isolates, such as multiple sclerosis in Italy in the Sardinia population (Sotgiu et al., 2004).
The diminished genetic heterogeneity and the enrichment in some rare susceptibility factors make population isolates fine opportunities to study simplified models of complex disorders. Moreover, even environmental and societal components are likely more homogeneous than in outbred populations, reducing possible biases introduced by non-genetic effects (Varilo & Peltonen, 2004). Sample sizes are often too low to perform powerful association studies, but the common

ancestry permits the use of distinct approaches for mapping disease loci, based on the identification of haplotypes shared by affected subjects.
Population-wide investigations has been hardly ever undertaken and in practical the success of these approaches has relied on the collection of multiple pedigrees with genealogical ties. In these samples, linkage analyses has been effective in exposing rare causative genes for Hirschsprung disease in Mennonites, a population from North America, and nonsyndromic deafness in Bedouins from Israel. These peculiar samples showed uncommon Mendelian-like segregations of these complex disorders, that facilitated the discovery of high-penetrant risk factors and initiated the unraveling of their etiology (Peltonen et al., 2000).
For SCZ and BPD, the application of such approaches has been more trivial, as no Mendelian-like forms have ever been described (Sullivan et al., 2012). The hypothesis was confirmed by recent works on pedigrees for both SCZ (Timms et al., 2013) and BPD (Georgi et al., 2014), showing different segregating loci. In the work of Georgi and collaborators, a particularly interesting dissection of linkage results in a large Old Order Amish pedigree revealed not only the presence of several linkage signals within and across nuclear families, but also that multiple haplotypes contributed to single linkage peaks (Georgi et al., 2014). In addition, each of the candidate risk haplotype exhibited incomplete penetrance. Thus, even in this well characterize homogeneous population, BPD has a polygenic inheritance typical of a complex disorder.

Despite the proven complexity, the potential of these type of samples has been unfolded with the advent of NGS; the techniques have provided the chance to combine segregation approaches and WES or WGS to obtain a comprehensive view of possible risk factors, necessary to model risk. The prioritization of variants according to the sharing in patients has yielded in general to a ten of candidate alleles per pedigree, selected for further investigations (Cruceanu et al., 2013; Georgi et al., 2014; Timms et al., 2013). Again the variants were not always shared by all the affected individuals and sometimes present also in non-affected subjects, coherently with the complex inheritance (Cruceanu et al., 2013; Georgi et al., 2014).
A further conceptual improvement came with the last works, which finally attempted to explore the polygenic nature of BPD, focusing on the mutational load of a pedigree. Results revealed multiple rare damaging variants in the same patients, affecting genes involved in the same biological processes (Ament et al., 2015; Kerner et al., 2013). Although preliminary, this studies have shown the potential of family- and population-based samples to define sets of rare alleles conferring risk to SCZ and BPD, thus providing realistic models for these complex disorders.

1.1.6 The genetic overlap between schizophrenia and bipolar disorder

Beside the difficulties in defining specific risk factors, genetic studies have consistently converged on an overlap between SCZ and BPD. First evidence came from epidemiological studies, showing that first degree relatives of schizophrenic patients have also a 4-fold risk of developing BPD; the opposite was also true (Lichtenstein et al., 2009).

Subsequently, the findings have been further substantiated by molecular genetics, as the same factors have been repeatedly reported to confer susceptibility for both the disorders (Smoller et al., 2013). Moreover, the polygenic risk score proposed by Purcell et al, when calibrated on SCZ samples, could efficiently predict case-control status non only in SCZ cohorts but also in BPD ones (Purcell et al., 2009). This indicated that some common risk variants have pleiotropic effects for SCZ and BPD. Finally, the 108 loci currently associated with SCZ, include 3 of the 6 genes significant also for BPD (*NCAN*, *TRANK1* and *CACNA1C*) (Ripke et al., 2014).

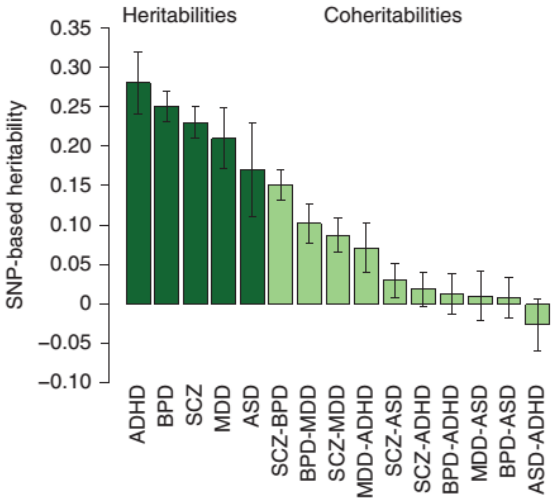

Figure 1.3: SNP-based heritability and co-heritability estimated for 5 major psychiatric disorders. SCZ and BPD show the highest genome-wide pleiotropy, as their co-heritability is about 70% of their individual heritability. From Lee et al., 2013.

In light of these observation, the cross-disorder group of the PGC used common SNPs to estimate the genetic relationships between the five major psychiatric disorders: schizophrenia, bipolar disorder, major depressive disorder (MDD), autism spectrum disorder (ASD) and attention-deficit/hyperactivity disorder (ADHD) (Lee et al., 2013). Results revealed that almost all the disorders share a fraction of genetic etiology, and that the overlap was particularly important between SCZ and BPD, with a genetic correlation of 0.68. Thus about 70% of common risk variation is shared between the two disorders (Figure 1.3). Similar analyses have not been performed yet for rare variants, as these are likely more private and difficult to detect in different samples. More general estimations have however been calculated with family studies, all converging on a co-heritability of at least 60% (Lichtenstein et al., 2009; Song et al., 2015).

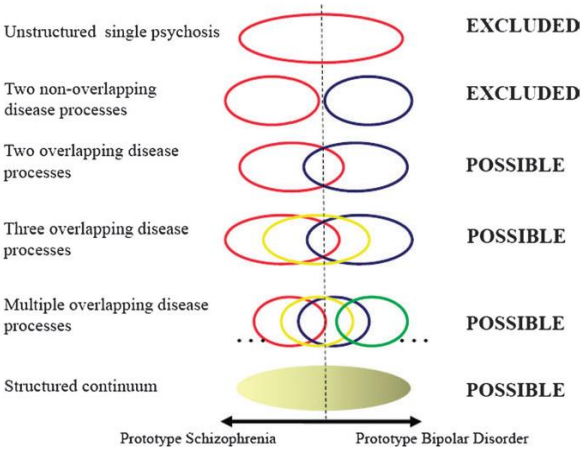

Figure 1.4 .Models of the possible biological relationships underling clinical phenotypes. Each circle represent a group of genes influencing a specific phenotype, defining a biological entity. It is still unclear whether the clinical spectrum is determined by multiple, overlapping entities or by a structured continuum of disease processes. From Craddock et al., 2009.

Despite the high overlap, SCZ and BPD can't be considered identical, for several reasons. Firstly, cross-disorders relative risk in families is always inferior than the within-disorder one. Secondly, some of the variants are specifically associated with a phenotype; CNVs for example, seem to play a minor role in BPD than in SCZ. Consequently, it's realistic to speculate that there must be some characteristic genetic factor driving the outcome towards psychosis or mood symptoms. Nonetheless, the number of distinct biological entities underling the clinical spectrum is still yet to be determined (Figure 1.4) (Craddock, O'Donovan, & Owen, 2009).

1.4 The emerging picture on schizophrenia and bipolar disorder

20 years of genetic studies have led to significant improvements in understanding the genetic architecture of SCZ and BPD. It's clear now that empirical data support both the 'common disease-common variant' and the 'common disease-rare variant' models (Sullivan et al., 2012). The allelic spectrum encompasses so far common polymorphisms, rare sequence variants and even rarer CNVs (figure 1.5). As predictable, the frequency of risk variants has an inverse relationship with penetrance, with common variants only slightly increasing the risk for the disorders and rare variants having instead higher effect sizes(Mowry & Gratten, 2013). The current evidence likely excludes the existence of rare, almost completely penetrant alleles (Mendelian) as well as of common polymorphisms with moderate or large effects (Sullivan et al., 2012). However, the exact total number of loci is still unknown and, assuming an heritability of 80%, more have to be identified. Moreover, while common variants account for about 50% of heritability (Lee et al., 2012), the precise contribution of rare alleles has yet to be determined. Finally, whereas several loci have been confidently reported, no single variant has been definitely implicated (Harrison, 2015; Neale & Sklar, 2015).

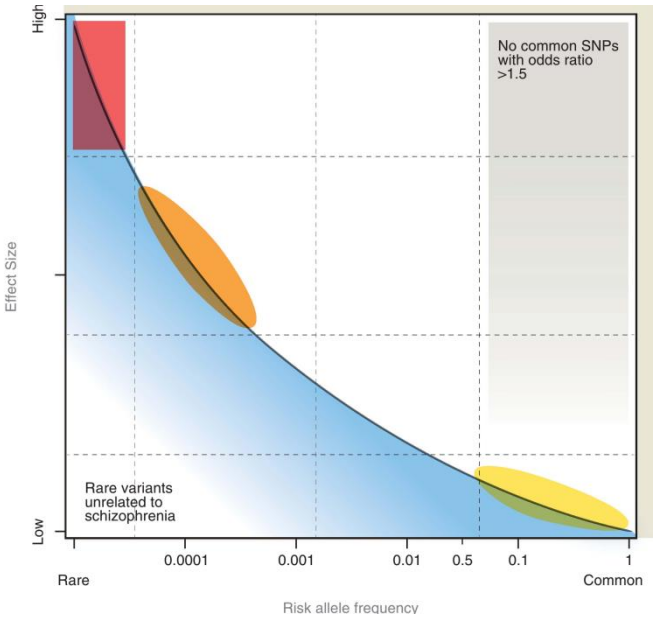

Figure 1.5: Allelic spectrum of risk variants identified for SCZ. From Mowry and Gratten, 2013.

Beside the cited limitations, genetic studies have provided remarkable convergence on some biological networks, clearing the ground for the discovery of possible mechanisms for the etiology of the disorders. First supports have come for the original proposed hypothesis for SCZ, positing a hyperactive dopamine transmission. The theory has been formulated after the discovery of the first antipsychotic drugs, in early 1950s, since these act mainly as antagonist of dopamine

receptors, in particular type 2 (D_2) (Neale & Sklar, 2015). Interestingly, *DRD2* gene, encoding for D_2 receptor, maps in one of the 108 loci associated to SCZ (Ripke et al., 2014), in line with the initial assumptions.

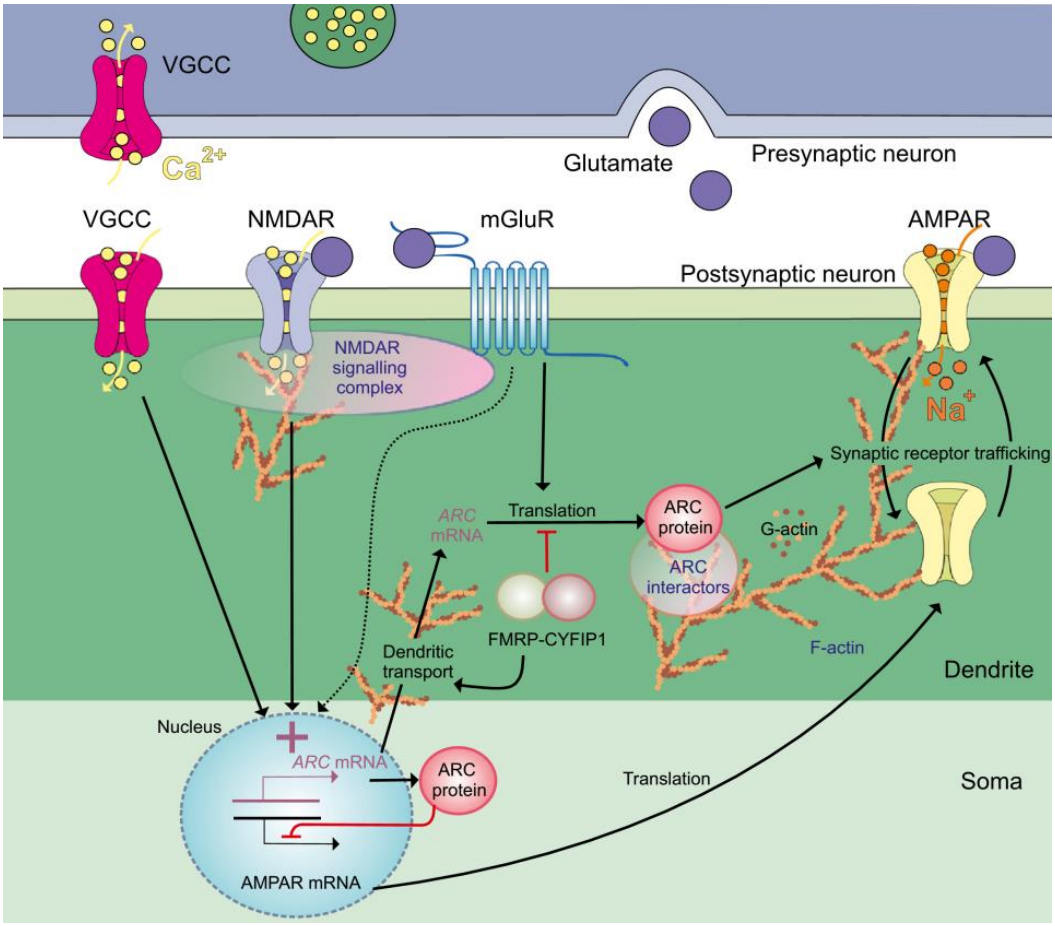

Figure 1.6: Synaptic plasticity modulated by NMDA receptors. From Hall et a., 2015

But another hypothesis has recently received much attention: the glutamatergic synapse dysfunction. Indeed, the known 108 loci include several genes for glutamate receptors, as *GRM3,* (metabotropic), *GRIA1* (AMPA, ionotropic) and *GRIN2A* (NMDA, ionotropic) (Ripke et al., 2014). NMDA receptor signaling, in particular, has become a major candidate pathway for SCZ liability. NMDA receptors are ionotropic Ca^{2+} channels, activated upon membrane depolarization, generally after the opening of AMPA receptors. Glutamate binding on activated NMDA receptors causes an influx of Ca^{2+}, triggering several signaling cascades within the neuron (Figure 1.6). On one side, this results in an enhancement expression of AMPA receptors, leading to a long-term increase of synaptic strength (long-term potentiation). Then again, the transcription of activity-regulated cytoskeleton-associated scaffold protein (ARC) is induced; ARC complex has a major influence on dynamics of dendritic spines through the remodeling of actin cytoskeleton. These glutamate receptors are thus key regulators involved in persistent synaptic changes, determining structural and molecular plasticity, fundamental for memory and cognition, but also for the formation of brain connectivity during development. (Hall et al., 2015). Consistently, it has now been established from postmortem brain studies, that SCZ patients show morphological

alterations in dendritic spines, specifically a reduction in spine density and arborization in the cortex (Moyer et al., 2015). The current model proposes a hypofunction of NMDA receptor in SCZ, causing a reduced plasticity. Indeed, pre-frontal cortex and hippocampus of SCZ patients display a decreased expression of these glutamate receptors, both at the transcriptional and the translational level (Weickert et al., 2013). Moreover, NMDA antagonists can induce also schizophrenia-like psychotic symptoms in humans and rodents (Zhang et al., 2016). Three main works have further provided genetic support for this hypothesis. *de novo* CNVs (G Kirov et al., 2012), *de novo* sequence variants (Fromer et al., 2014) and overall sequence variants (Purcell et al., 2014) associated with SCZ have been found to significantly affect NMDA receptor network. More in detail, the results evidenced an enrichment in proteins belonging to the ARC and the PSD-95/Dlg4 complexes. The latter is composed by several proteins located in the postsynaptic density (PSD) region, which have a major role in modulating NMDA receptor signal transduction and activity.

On the same line, voltage-gated Ca^{2+} channels (VGCCs) have been repeatedly implicated in SCZ. Purcell et al. reported a higher burden of disruptive mutations in VGCC genes in SCZ cases (Purcell et al., 2014); additionally, 3 of such genes overlap the 108 SCZ loci (Ripke et al., 2014). These channels mediate the entrance of Ca^{2+} in excitable cells, above all neurons, after membrane depolarization. In glutamatergic synapses, this regulates axonal growth and guidance, mediated by glutamate signaling (Neale & Sklar, 2015). In a more general view, thus, the emerging hypothesis for SCZ is an impairment in neurodevelopmental processes in the formation of neuron connectivity. Indeed, pre-natal and peri-natal complications are recognized as environmental risk factors, that could act in concomitance with a genetic susceptibility to cause the onset of SCZ (Kotlar et al., 2015).

The majority of the reported theories have been elaborated from SCZ studies, since many more loci have been discovered compared to BPD. Although it's still unclear whether these observation can be extended also to this second disorder, the clinical and genetic overlap suggests that common mechanisms may be involved (Neale & Sklar, 2015). Definitely, some recent evidence sustain this speculation. *CACNA1C* gene, belonging to the VGCC class, is now significantly associated also with BPD (Ferreira et al., 2008) and BPD pedigrees have an increased burden of rare variants in VGCCs and in neuronal ion channels in general (Ament et al., 2015). Glutamatergic synapse seems to be also involved, but NMDA receptors have a specular role for depressive symptoms with respect to SCZ, as antagonists (especially ketamine) can effectively counteract these specific signs (Hashimoto et al., 2013). Brain abnormalities have been less investigated, but diffusion tensor imaging experiments have shown a reduced connectivity between brain areas, particularly with prefrontal cortex (Wessa et al., 2014). Further insights are though required to formulate a more organic hypothesis on BPD pathogenesis

Concluding, genetic investigations have lead to unprecedented advances in understanding the biology of SCZ and BPD. Many questions, however, are still to be solved, starting from the genetic architecture, the pathogenetic mechanisms, that are still hypothetic, to finally unravel the interplay between genetic and environmental factors that determines risk (Harrison, 2015).

2. Aim of the research

This study aims at investigating the genetic component of schizophrenia (SCZ) and bipolar disorder (BPD) in a unique population sample, constituted by several families with a high recurrence of the two psychiatric disorders. Since all the patients have a common ancestry from a closed community, the working hypothesis is an enrichment of rare alleles with large effects on disease risk. On these premises, this sample offers the opportunity to examine the role of rare variation in SCZ/BPD etiology. Several studies have in fact evoked the involvement of rare alleles in these psychiatric disorders, but their precise impact has not yet been determined, nor specific genes have been identified.

These limited results derive from the general difficulties in implicating rare variants in heterogeneous cohorts; as a consequence, more homogenous samples, especially family-based ones, have gained an increasing importance. In these context, however, the majority of available approaches have been developed for Mendelian disorders and are less effective when dealing with complex traits. Indeed, the same approaches had been previously unsuccessful in this sample. For this reason, a novel strategy was envisioned, combining Identity-By-Descent (IBD) mapping and Whole-Exome sequencing (WES). Each of the steps required the development of a customized bioinformatic pipelines and their integration to those already available.

In a preliminary phase of the project, the role of Copy Number Variants (CNVs) wanted to be evaluated, since these structural variants have been implicated in psychiatric disorders, particularly in SCZ.

The following steps were then directed to the set up of IBD analysis, from the selection of the most suitable algorithm to the processing of the output. The final outcome wanted to be a map of the shared haplotypes across the genome, for the identification of loci particularly common in patients. More in details, the focus was put on the tracking of loci encompassing multiple pedigrees, thus accounting for a high number of patients. The last phase envisaged the examination of WES data from a subset of patients, selecting rare or novel variants mapping in IBD haplotypes. Subsequently, the IBD map was intended as an instrument to infer the sharing of each mapped variant in the entire population and the segregation within each pedigree. The final goal was then to establish whether these variants were affecting specific biological functions and to provide a set of candidate genes involved in SCZ and BPD.

3. Materials and Methods

3.1 Sample description

Sample collection was performed in collaboration with the Center of Mental Health of Chioggia. SCZ, BPD and SZA were diagnosed according to DSM-IV criteria (American Psychiatric Association, 2013), after independent evaluations from two expert psychiatrists. All patients had been hospitalized and medical records were available for each of them. Families were collected whenever at least an additional member (first or second-degree relative) met the DSM-IV diagnostic criteria. Informed consent was obtained for all participants and the study was approved by the Medical Ethic Committee of Chioggia, in accordance with current legislation.

Chioggia is a small town built in a group of islands of the southern Venetian lagoon. Total population size is around 60,000 inhabitants; the collected sample, however, belonged to groups of families living in the old parts of the town. More in details, individuals came from two islands, the main Chioggia and its neighborhood Sottomarina (figure 3.1). Pedigrees were tracked by a genealogist back to four/five generations, thanks to church registers and demographical records, to confirm the origin from the historical population.

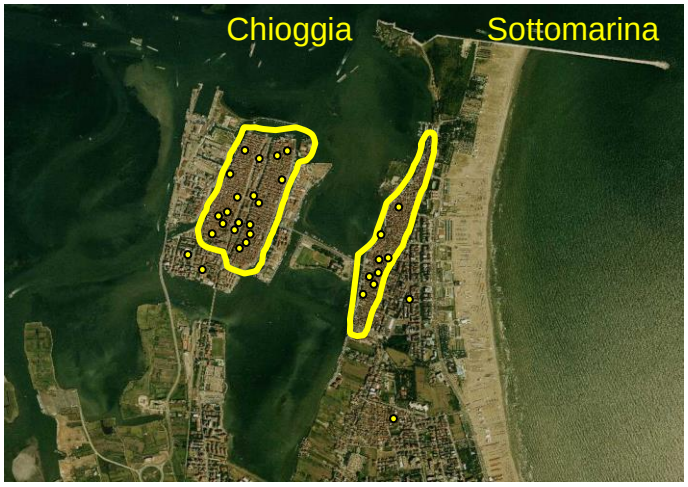

Figure 3.1: Aerial view of Chioggia and Sottomarina, in the Venetian lagoon. Yellow outlines indicate the old parts of the town, while yellow dots represent approximately where the majority collected samples reside.

For socio-cultural reasons, Chioggia had been particularly secluded until 50 years ago; as a proof, several cases of homonymy are observed and 3 family names account for about 50% of the entire population (Gessoni et al., 2010). This evidence suggests a high rate of endogamy, thus a more homogeneous genetic background. Interestingly, the prevalence of the psychiatric disorders of the SCZ/BPD spectrum seems peculiarly elevated, twice as high as the surrounding geographical areas. These features overall indicate a possible enrichment of rare, high-penetrant susceptibility alleles, making this sample ideal for the investigation of genetic etiology of the disorders.

The original sample includes more than 200 patients; the subset presented in this study is composed by a total of 197 subjects, of which 161 affected. 90 individuals were diagnosed with SCZ, 50 with BPD and 21 with SZA. 106 patients belonged to 36 pedigrees, together with 22 healthy relatives (figure 3.2). Additionally, 24 cases had a known family history of the disorders, but no other family member was available for study. On the basis of clinical assessment, 42% of the total pedigrees could be classified as schizophrenic (SCZ patients only), 15 % as bipolar (BPD patients only) and 6% as schizoaffective (SZA patients only); remarkably, the remaining 37% of families (22 pedigrees) comprised multiple disorders (SCZ and BPD or SCZ and SZA), supporting the overlap between SCZ and BPD. Further 31 isolated cases were also analyzed and finally, 14 healthy controls from the same population were collected.DNA was extracted from peripheral blood or buccal swab , with classical phenol-chlorophorm protocols.

3.2 Preliminary investigations on the population sample

Prior to this work, several investigations had been already performed on this sample. Since some of the pedigrees showed an apparent matrilineal transmission of the disorders, a possible role of mitochondrial DNA was analyzed (Bertolin et al., 2011). Haplogroups were determined in 89 patients, but their frequencies didn't significantly differ from the Italian ones, excluding a haplogroup-derived susceptibility. Moreover, the complete mitochondrial genome sequence was obtained for 27 affected subjects, but none of the identified variants had a predicted functional effect. Thus, point variation in mitochondrial genome could not account for the disorders in this population. Attention had thus turned to nuclear genome and a genome –wide linkage analysis based on SNP markers was performed. SNP genotyping data were obtained for the 197 samples using Illumina 370K-quad chip, for the detection of about 370,000 markers. 221,764 autosomal markers were selected after quality filtering, to remove SNPs showing Mendelian inconsistencies or with high rates of genotyping failures. Both parametric, under dominance or recessive models, and non-parametric, model-free settings have been considered. Some LOD signals emerged, but none of them was statistically significant.

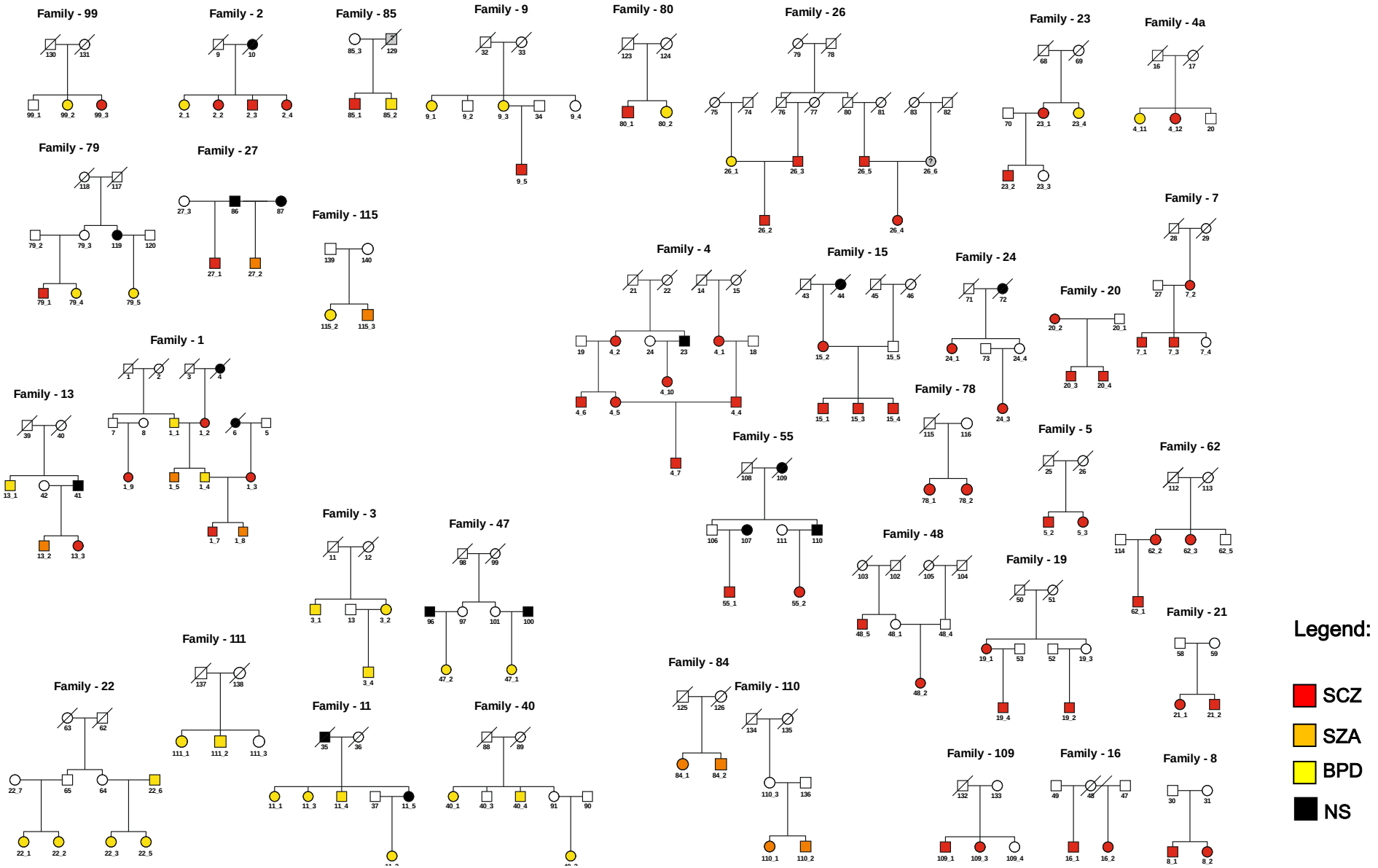

Figure 3.2 : Pedigrees of the 36 families analyzed in the study. Pure schizophrenic families are displayed on the bottom right part, pure bipolar on the bottom left and mixed families on top. SCZ: schizophrenia; BPD: bipolar disorder; SZA: schizoaffective disorder; NS: referred affected with either SCZ or BPD.

3.3 Copy Number Variants

CNVs calling was performed by the CNV-webstore online tool (http://cnv-webstore.ua.ac.be/cnv-webstore/) (Vandeweyer et al., 2011), on the basis SNP-genotyping data. Necessary input files were extracted from Illumina standard output thanks to three PERL scripts provided by the tool itself. The entire SNP set was considered on the 197 available samples. After analysis, 32 samples were excluded for quality reasons.

CNV-webstore integrates three algorithms for the detection of CNVs: QuantiSNP (Colella et al., 2007), PennCNV (Wang et al., 2007) and VanillaICE (Scharpf et al., 2008). A majority vote approach was chosen, thus only CNVs reported by at least two algorithms were retained. All the methods are based on Hidden-Markov Models for the prediction of copy number status exploiting two main parameters: the normalized profile of SNP intensity signals (LogR ratio) and the proportion of signal attributed to the minor allele (BAF) for each SNP. For example, stretches of SNPs with LogR ratio inferior to 0 and BAF of 0 or 1 (homozygous SNPs) are called as deletions (copy number=1); analogously, LogR ratios greater than 0 and BAF of 1/3 or 2/3 indicate a duplication (figure 3.3). Boundaries of differentially called CNVs are automatically adjusted by CNV-webstore.

The overlap between the entire collection of detected CNVs and the International Schizophrenia Consortium (ISC) dataset (Purcell et al., 2008) was explored converting all data in UCSC Genome Browser custom tracks.

For the identification of rare or novel CNVs, a quality filtering was performed in order to focus on reliable variants. On the basis of chip characteristics, the following thresholds were imposed: ≥10 consecutive SNPs, ≥30 kb size, an average density of 1 SNP every 10 kb and an average score of 0.5 per SNP, as calculated by the three algorithms in CNV-webstore. A consensus was then derived across the different samples. The resulting variants were searched in DGV (database of genomic variants, http://dgv.tcag.ca/dgv/app/home) and HapMap (http://hapmap.ncbi.nlm.nih.gov/) resources. Novel or rare variants were initially defined as those with a <70% overlap with a reported DGV polymorphism. Subsequently, retained CNVs were confronted with HapMap data and variants largely reported were excluded. All steps were realized by PERL scripts.

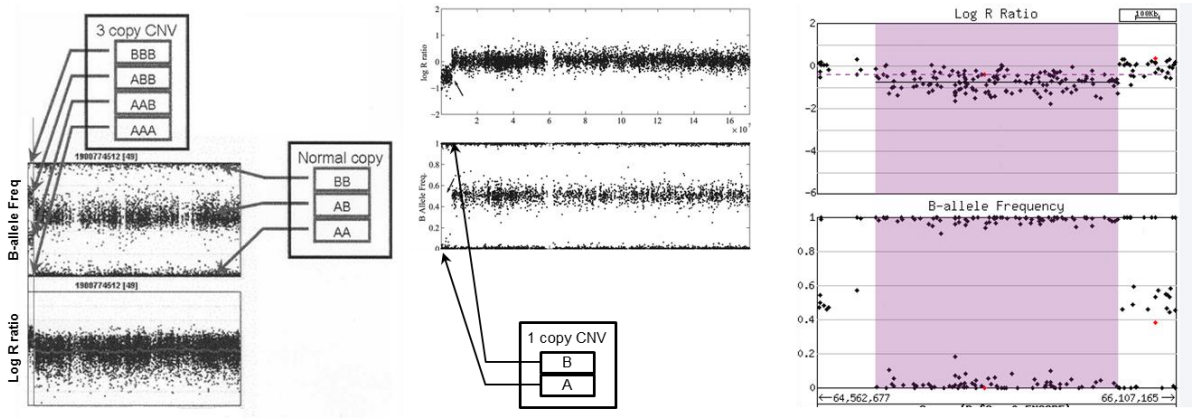

Figure 3.3: Examples of CNV detection from LogR ratio and B-allele frequency estimated from SNP-genotyping data. As illustrated in panel a (adapted from Wang et al., 2007), normal copy numbers (2 copies) are characterized by a LogR of 0 and by B-allele frequencies (BAFs) distributed across three possible values: 0 (AA genotype), 0.5 (AB genotype) and 1 (BB genotype). Panel a shows also a duplication (3 copy CNV), evidenced by increased LogR ratio values and anomalous BAFs (0.33 and 0.66 for AAB and ABB genotypes respectively). In panel b (adapted from Colella et al., 2007)is instead represented a single deletion (1 copy CNV), revealed by a negative LogR ratio and the absence of heterozygous genotypes (BAF ≠ 0.5). In panel c is shown a deletion (1 copy CNV) as displayed in CNV-webstore.

3.4 IBD analysis

Two different free programs have been evaluated for IBD analysis: GERMLINE (Gusev et al., 2009) (http://www.cs.columbia.edu/~gusev/germline/, v1.5.1) and Relate (Albrechtsen et al., 2009) (http://www.popgen.dk/software/index.php/Relate, v0.998). Both methods require SNP-genotyping data as input, in plink file format, and were run on the 221,764 good quality markers previously selected. For technical reasons, again only autosomes could be investigated. The algorithms are based on completely different principles for the detection of IBD regions between all the possible pairs of individuals. GERMLINE relies on an initial phasing step, implemented with BEAGLE software, when SNP genotypes are converted into haplotypes. Since multigenerational pedigrees were available, the phasing results could be checked independently by alternative software (e.g. GeneHunter); GERMLINE procedure was found to be often imprecise. Unfortunately, this critical step compromises the calling of IBD regions, that is obtained by simply aligning the phased haplotypes to detect identical blocks. Consequently, Relate was selected as the best performing software. Relate exploits a first order Markov chain to provide posterior probabilities of IBD status for each marker, on the basis of 3 main parameters:

- a: rate of change between IBD states from a marker to the next.
- k: vector of 3 elements, k0, k1, k2, indicating the overall ratio of IBD0, IBD1 and IBD2
- ε: genotyping error rate

The complete list of options and parameters used for the final analysis are listed in table 3.1. Each of the option was set up by running preliminary tests on sample subsets. Besides ε, that must be indicated, all the other parameters (a, k) were chosen to be optimized by the algorithm itself (doParameter calculation=0, fixA=0); even the estimation of a from the overall allele sharing between a pair was excluded (calculateA=0), since it would be efficient only with a single nuclear family or very distantly related individuals. Also IBD2 status was considered (fixK2=0). As suggested by the manual in case of calculate=0, times_to_run and times_to_converge were set to 10 and 5 respectively, to reach the necessary convergence of the algorithm.

Option/Parameter	Value	Notes
allpairs	1	1= run all pairs; 0= run a single pair
pair[0]	0	first individual of the pair (only if allpairs is set to 0)
pair[1]	1	second individual of the pair (only if allpairs is set to 0)
double recombination	0	rates of double recombination
LD	0	measure to calculate LD (LD=0=rsq2; LD=1=D)
min	0.01	minimum minor allele frequency allowed
alim[0]	0.001	Lower limit of the range of a
alim[1]	1	Upper limit of the range of a
doParameter calculation	0	1= use the specified parameters (par[0,1,2]); 0= optimize parameters
par[0]	0.3	a (only used if doParameter is set to 1)
par[1]	0.25	k2 (only used if doParameter is set to 1)
par[2]	0.5	k1 (only used if doParameter is set to 1)
ld_adj	1	1=use the pairwise emission probabilities to correct for LD; 0=no correction
epsilon	0.01	ε
back	25	number of SNPs to condition on to accomodate LD
doPrune	1	1=prune SNPs to remove LD; 0=no pruning
prune_value	0.5	maximum rsq2 accepted between SNPs (only if doPrune is set to 1)
fixA	0	1=use fixed values for a; 0=optimize a
fixA_value	0	fixed a value (only if fixA is set to 1)
fixK2	0	1=use fixed values for k2; 0=optimize k2
fixK2_value	0	fixed k2 value (only if fixK2 is set to 1)
calculateA	0	1=estimate a from overall allele sharing; 0=optimize a
phi_value	0.013	recombination rate (Morgans/Mb)
convergenge_tolerance	0.1	parameter for convergence of optimization
times_to_converge	5	The number of times the optimization should reach the same optimum
times_to_run	10	The maximum number of times the optimization is run
back2	50	debuggin parameter

Table 3.1: Parameters for Relate analysis

Immediately before proper IBD analysis, Relate executes a SNP pruning step to remove markers in strong LD. LD blocks might be: (I) local associations of alleles that, as a consequence of genetic drift, are particularly common in the population; for these SNPs it's impossible to distinguish identity by state (IBS) from IBD; (II) traces or very ancient ancestors. The aim of the approach was to identify haplotypes inherited by relatively recent ancestors, like population founders, thus LD blocks would have represented a source of false positive results. LD pruning was carried out (doPrune=1) with a threshold of 0.5 (prune_value=0.5): if SNPs with a genotype correlation (measured as r^2, LD=0) greater than 0.5 were present, only one representative SNP would have been retained. LD profile was calibrated on the entire sample using sliding windows of 25 markers (back=25), as recommended with approximately 250,000 input markers. By default, back2 was set to 50, twice the value of back. This configuration yielded 135,702 markers after pruning. Analysis was run on a total of 19,306 pairs of 197 individuals (allpairs=1).

The output is structured as a text file (.post) with 2 rows per pair and a column per marker. These data were used to infer the three possible IBD status:

- IBD0: the two subjects share no IBD.
- IBD1: the two subject share one IBD haplotype on one homologue
- IBD2: the two subjects share two IBD haplotypes both the homogues

The first row of the output file lists the posterior probabilities for each SNP to be IBD0, while the second row report the probabilities of IBD1 status. Probabilities of IBD2 status were instead calculated for each pair, as the complementary to 1 of the IBD0 and IBD1 values. An *ad-hoc* pipeline based on PERL scripts was develop to process this data and obtain a list of IBD regions.

IBD regions were called according to the trend of IBD0, IBD1 and IBD2 probabilities across the markers. IBD status was assigned to every SNP using a threshold of 0.98. The method accounted also for possible fluctuations: in a context of markers with defined status (thus when at least one SNP exceeded the probability of 0.98), the threshold was lowered to 0.70 for that specific status. IBD regions were then defined as segments of consecutive IBD1 or IBD2 markers. Blocks of undefined markers (e.g. markers whose probabilities never reached the imposed thresholds) were by default considered as IBD0, therefore not IBD. These blocks were particularly observed close to centromeres or telomeres on in regions of transition between status.

3.5 Estimation of genetic similarities from IBD data

Genetic similarity was measured through the relatedness coefficient (r), calculated as the sum of the sizes of IBD1 regions and twice the sizes of IBD2 regions over the total genome captured by the 135,702 analyzed markers. The latter was estimated to be 2,784,183,496 bp, approximately the size of euchromatic genome (2.88 Gb) (Platzer, 2006). r was estimated for all the possible pairs and compared to expected values. To evaluate the performance of this method in unrelated individuals, a further sample of 33 subjects of Italian origin was investigated. SNP- genotyping was performed with the same platform and IBD analysis was carried out on the same 135,702 markers. To obtain this configuration, the subset of markers was previously extracted from the total set, then Relate was run avoiding the pruning step (doPrune=0). Output processing and calculation of genetic similarity were accomplished with the identical approaches. Significance of differences in relatedness between the two unrelated cohorts were calculated with Welch two-sample t-test. To evaluate the degree of kinship existing between unrelated pairs, expected genetic similarity was calculated for increasing number of elapsing generations and the number of pairs with corresponding r was counted.

3.6 Cluster Analysis

All the steps of cluster analysis were performed with R software (https://www.r-project.org/). From r coefficients, a matrix of genetic distances was generated for 115 'family founders', where distances were simply calculated as:

$$d=1-r$$

First, multidimensional scaling (or principal coordinates analysis) of the obtained matrix was calculated on 2 dimensions, and resulting values for each pair were plotted.

Second, 53 more samples were added, including the 33 individuals of Italian origin and 20 HapMap individuals of 4 different ancestries: 5 Yoruba (Africa), Yoruba (Africa), Han Chinese in Beijing (China), Japanese in Tokyo (Japan) and Europeans in Utah (Europe). IBD analysis and genetic similarity calculations were carried out with the same strategy adopted for the 33 Italians. After obtaining the total matrix of genetic distances, including 168 subjects, hierarchical clustering was achieved with ward method and results displayed in a dendrogram. Significance of Chioggia and Sottomarina clusters was calculated with a generalized linear model with the known origin as covariate. (didascalia: the covariate origin could significantly predict the cluster).

3.7 Haplotype clustering

Haplotype clustering conceptually consists in identifying groups of individuals sharing identical haplotypes, starting from pairwise-detected IBD regions. Clustering was based on the following logical rule: if the same genomic region was found IBD in all the possible pairs of at least 4 individuals, then a common haplotype had to be assumed (see figure 4.7 for details). The entire procedure was developed using PERL scripts. Briefly, all genome was scanned by a sliding window of 200 SNPs and a pace of 50 SNPs. Haplotype clustering was performed each time considering all the IBD regions overlapping more than 50% of the window; this ensured the effective overlap between the pairwise IBD segments, avoiding wrong outcomes. With this approach, a size exclusion of 100 markers, was also automatically introduced, thus selecting only the larger, most reliable IBD calls. In every window, the clustering algorithm was run three times, each time randomly choosing the order of segments/pairs to be analyzed. A consensus was then calculated within windows, merging clusters with more than 85% identity. These last steps were thought to counteract possible false negative calls, causing cluster interruptions due to the missing detection of an IBD region between a pair; random order in clustering should in fact break the cluster in different ways, then joined in the following stage. Finally, a inter-window consensus was obtained, merging clusters differing in less than 20% of members and in less than 15% in size, eventually estimating the genomic extension of shared haplotypes.

For every cluster, a series of family-based scores were calculated, indicating the number of families and the fraction of the patients per nucleus that were included. In addition, the relative amount of subjects from the two subpopulations (Chioggia or Sottomarina) and the percentage of patients with a specific disorder (SCZ, BPD or SZA) were computed.

3.8 Whole-exome sequencing and variant filtering

Whole-exome sequencing was carried out on 17 selected patients with IonTorrent® technology (Life Technologies) on IonProton™ platform. Template preparation, sequencing reaction, reads alignment and variant calling were performed in collaboration with CRIBI (Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative), from the University of Padova. Target enrichment was achieved by multiplex PCR with Ion Ampiseq™ Exome RDY Kit.
Classical annotation procedure executed by *IonReporter*™ software (v 4.4) was improved thanks to the development of a pipeline, based on PERL scripts. The approach was set up after the surfacing of some issues related to the presence and the frequency of variants in reference databases:

1. Annotation of frequencies with IonReporter™ was sometimes missing or inaccurate.
2. In annotation procedure, only genomic position of variants was considered, without any actual check between the variant and the alleles reported in databases. As a consequence, any variant in the same position of a known SNP would have been always called as a polymorphism, even if the variant itself had never been observed before.
3. This last issue derived from inaccuracies in the reference human genome, where sometimes the reference allele for a SNP is not the most frequent one. Thus, whenever a SNP is not a common polymorphism, the reference becomes a potentially interesting allele, but any caller would never detect it as a variant. This situation has two main drawbacks. First, the reported minor allele frequency (MAF), generally used for filtering, refers to the reference and not to the detected variant. Typically, an individual might look homozygous for a rare allele, but is instead homozygous for the most frequent one. Second, homozygous genotypes for the rarest allele (the reference) are never called.

The pipeline relies on the creation of a variant sharing table, listing unique list of all the identified variants across the 17 exomes. For each variant, a sharing string was computed, composed by 17 digit, corresponding to the 17 patients; for every patient, 0 indicated the absence of the variant, 1 heterozygosity and 2 homozygosity for the variant.
The variants listed in the table were then re-annotated using three different databases: dbSNP (release 142), 1000Genomes and ExAc; frequency annotations from IonReporter™ were ignored.

(http://www.ncbi.nlm.nih.gov/SNP/) is the reference database for known SNPs, collecting both common and rare polymorphisms. Any SNP reported in dbSNP has a relative reference sequence (rs) number. MAF to SNPs in dbSNP is provided by the 1000Genomes project (http://www.1000genomes.org/), currently including data from 2,504 human genomes. ExAc (http://exac.broadinstitute.org/) stands for Exome Aggregation consortium, collecting information relative to 60,706 exomes.
Filtering was then performed on the basis of this information. Variants matching a dbSNP entry (in both position and alleles) were defined as SNPs and classified according to their frequency in 1000Genomes as following:

a. Common SNPs: MAF >=1%
b. Rare SNPs: MAF <1%
c. SNPs with unknown frequency
d. mAiR SNPs: MAF >50%.

The simultaneous analysis of the 17 exome data permitted the identification of mAiR cases. Since the non-reference allele is the most common one, infact, the probability that all the sequenced patients were homozygous for the reference was very low (This event occurring, the variant wouldn't be of interest, as from the IBD map there is no haplotype shared by all the 17 patients). In the majority of the mAiR situations, therefore, the calling of the alternative allele generated an entry in the variant sharing table, recognized as mAiR after annotation and filtering. For these occurrences, real MAF was computed from 1000Genomes, by evaluating all the frequencies of alleles detected for the SNPs. Sharing string was then adjusted by replacing 0 with 2 and vice versa. In this way, homozygous reference calls were revealed. The corrected list of mAiR was then re-filtered according to the actual MAF, and variants with MAF <1% were retained.
Rare SNPs, SNPs with unknown frequency and mAiR rare SNPs were further filtered by ExAc frequencies. Only variants not present or with a frequency <5% were preserved. The higher frequency threshold was determined both by the extremely higher sample size in ExAc and, especially, by the fact that psychiatric patients are included in ExAc collection.

Any variant that was absent in dbSNP was placed in a separate list. From this group, variants detected in more than half (8) patients were excluded, since they were likely technical false positives. The remaining variants were filtered by ExAc frequencies similarly to rare SNPs. The resulting list was joined to the rare SNP one, to get the final subset of rare and novel variants.
Finally, only variants mapping in IBD haplotypes were prioritized. It's important to note that the IBD map was not simply a positional information, but it provided also insights on the sharing of a variant located in a specific haplotype. Both features were taken into account for prioritization.

3.9 Functional enrichment analyses

Functional enrichment analysis was conducted exploiting the Database for Annotation, Visualization and Integrated Discovery (DAVID) tool (https://david.ncifcrf.gov/) (Huang et al., 2009a; Huang et al., 2009b). A list of genes was compiled for each set of variants (all rare or novel, IBD_{tot}, IBD_{sel}, LoF, LoF IBD_{tot}), including genes carrying at least one variant of the specified type. Lists were then tested for enrichment in one of the functional categories annotated by the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) and Reactome (http://www.reactome.org/). A significance threshold of 0.05 was established; multiple correction p-values were calculated with Benjamini method. Only results that were at least nominally significant (nominal p-value <0.05) were considered. Genes from interesting categories were mapped again on individual pathways/processes in the original KEGG and Reactome databases, to assess whether variants were evenly distributed or clustered in specific nodes.

For a further dissection of the synaptic transmission category, enrichment analysis was performed on specific gene sets. PSD-95, ARC and NMDA receptor sets were derived from Fromer et al. (Fromer et al., 2014), who compiled the lists from proteomic studies. VGCC and glutamatergic synapse sets were instead assembled from KEGG annotation. Enrichment test was carried out considering an hypergeometric distribution; multiple test correction was again computed with Benjamini method.

3.10 Analyses based on inference of variant sharing from IBD map

Sharing and segregation of variants mapping in IBD haplotypes were inferred thanks to haplotype clusters. With this approach, the entire population sample could be investigated, starting from the data of the 17 exome-sequenced patients.

Familial load of variants was calculated by simply adding the number of variants per family in genes belonging to the axon guidance or synaptic transmission processes, emerged from functional enrichment analysis. To complete the annotation from KEGG and Reactome, each category was expanded on the basis of Gene Ontology (GO) and DAVID functional classification; for axon guidance, ‘neuron development’ and ‘neuron projection’ terms were considered; terms like ‘synaptic transmission’, ‘synaptic plasticity’ and ‘synapse organization’ were instead acknowledged for synaptic transmission. Burden was then calculated for a total of 87 genes, with at least one novel or rare variant, in the axon guidance cluster, and 66 genes in the synaptic transmission one.

Specific investigation of single variants were conducted on the basis of sharing results. In particular, prediction of effects for variants different from non-synonymous substitutions were obtained from MutationTaster (http://www.mutationtaster.org/) and Human Splicing Finder http://www.umd.be/HSF3/). All the cited variants have been validated by Sanger sequencing, after

amplification using a standard PCR protocol for FastStart Taq (Roche) (30 cycles at melting temperature of 60°) The complete list of primers can be found in appendix 7.6. The presence in both the exome-sequenced individuals and in the inferred cases was confirmed.

4. Results

4.1 Copy Number Variant analysis

In this study, a unique population sample has been deeply investigated for the search of rare risk factors conferring risk to schizophrenia (SCZ) and bipolar disorder (BPD) (see par 3.1). The working hypothesis is the likely enrichment of such rare alleles, on the basis of the high rate of endogamy and the increased prevalence of SCZ/BPD in this closed community.

Copy Number Variants (CNVs) are amongst the first type of rare variants for which an implication with SCZ/BPD has been suggested. Thus the initial step of this work was the examination of CNVs in the population sample. A total of 165 individuals, of which 137 affected, were analyzed with the *CNV-webstore* tool (Vandeweyer et al., 2011), for the calling and visualization of CNVs from SNP-genotyping data. A total of 4827 duplications and deletions were detected, with an average of about 30 variants per sample.

These results were compared with a dataset of CNVs from one of the largest studies conducted on schizophrenic patients (Purcell et al., 2008). No overlap was found, indicating that none of the CNVs previously reported in SCZ patients was present in our sample.

The focus then turned on the investigation of rare or novel CNVs in the 137 patients. An initial quality filtering was performed, selecting only larger (>30 kb) variants called with an average density of 1 SNP every 10 kb. From the obtained 834 CNVs, a consensus was calculated across the samples. The consensus CNV loci were subsequently searched in the reference database of copy number polymorphisms (DGV, Database of Genomic Variants) and in the released data from the international HapMap project. A total of 61 rare (frequency <1%) or novel CNVs in 48 loci emerged. Surprisingly, none of the variants was neither perfectly segregating within a family nor shared by many affected subjects. Indeed, the majority of CNVs were singletons, thus present in a single subject. The most common novel CNV was a deletion in 8p22 (chr8:16,135,246-16,216,372 bp), detected in three subjects from families 85 and 110 (figure 4.1). No known SCZ/BPD locus map in that region (see table 1.1) and no gene resides in the deleted region, making it difficult to infer any pathogenetic hypothesis. These data thus indicated that CNVs likely play a minor role in the pathogenesis of SCZ/BPD in the analyzed sample.

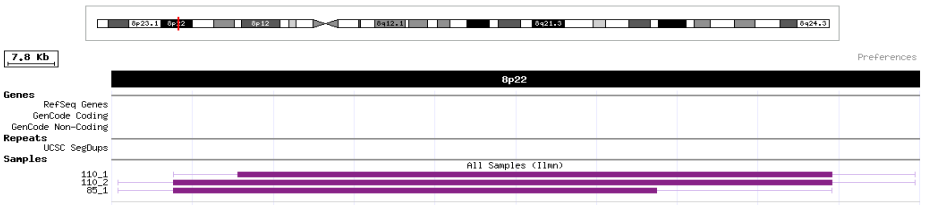

Figure 4.1: An 81 kb deletion (chr8: 16,135,246-16,216,372 bp) detected in two patients of family 110 and one in family 85. As shown, no gene is encompassed by this CNV.

4.2 The development of IBD mapping methodology

4.2.1 Preliminary findings: linkage analysis

The advantage of working with a collection of families, all with a common ancestry, is the possibility to identify risk variants by tracking loci shared by affected subjects. The classical approach for this purpose is linkage analysis, to identify loci that co-segregate with a disorder (see par 1.3.3.3). Genome-wide linkage scans with SNP markers had been previously performed on subsets of families, considering both parametric (dominant or recessive inheritance) and non-parametric models, but no significant results were achieved (see par 3.2 for details). Linkage approaches are likely ineffective when dealing with multiple risk factors rather than a single major gene, thus the preliminary findings suggest that also in this peculiar sample SCZ/BPD have a complex, polygenic inheritance. (Neale & Sklar, 2015). Moreover, even in the context of a reduced genetic complexity, intra-familial heterogeneity may hamper the detection of significant signals. A theoretical scenario to explain this issue is depicted in figure 4.2. In the hypothetic presence in the population of only three susceptibility alleles on different chromosomes, the transmission across generations could result in different combinations of such factors even in patients from the same family. Real risk alleles are probably more than three, but even this simplified example is sufficient to illustrate an expected genetic heterogeneity between patients. Additionally, a specific allele could be present also in non-affected subjects, given the typical incomplete penetrance of non-Mendelian variants. As a consequence, none of the factor would be perfectly co-segregating with the disorder within families and would produce statistically significant LOD score values.

4.2.2. Identification of Identical-By-Descent (IBD) chromosomal segments

In light of the failure of segregation-based studies, an original approach was designed, based on IBD (Identity By Descent) mapping. The rationale of this idea was that all patients belong to the same closed community, therefore risk alleles could have been reasonably located in founder haplotypes, passed through generations and found IBD in the modern population. (figure 4.2). IBD tracking was thus developed as a strategy to map these risk loci. One of the main strengths of the approach is the possibility to overcome segregation constrains, because the detection of IBD regions is conducted between all the existing combinations of analyzed individuals, pair by pair, without any *a priori* knowledge of familial relationships.

Relate software (Albrechtsen et al., 2009) was selected among other available tools to perform IBD analysis from SNP-genotyping data. Parameters were accurately set up on sample subsets to get the most reliable results (see par 3.4). In particular, the algorithm was calibrated in order to remove SNPs in strong Linkage Disequilibrium (LD), a source of false positive results. Only SNPs with $r^2 \leq 0.5$ were

retained; after SNP pruning, analysis was carried out on 135,702 markers on the 22 autosomes, with an average of 1 SNP every 22 kb. Coherently, the average size of LD blocks in non-African populations is estimated around 22 kb (Cardon & Abecasis, 2003). Since *Relate* output files couldn't be immediately interpreted, raw data were processed with a specifically developed pipeline based on PERL scripts. A total of 161,199 IBD regions were finally identified in the 19,306 pairs of 197 individuals, ranging from about 50 kb to entire chromosomes, as between parents and children. The majority of segments were IBD1 (only one IBD homologue), but 2.9% of them were IBD2 (both IBD homologues); the detection of IBD2 sharing in pairs other than full- siblings was consistent with the presence of inbreeding in this population. It's important to underline that IBD2 status doesn't imply homozygosity, since also the sharing of the same two heterozygous haplotypes is possible.

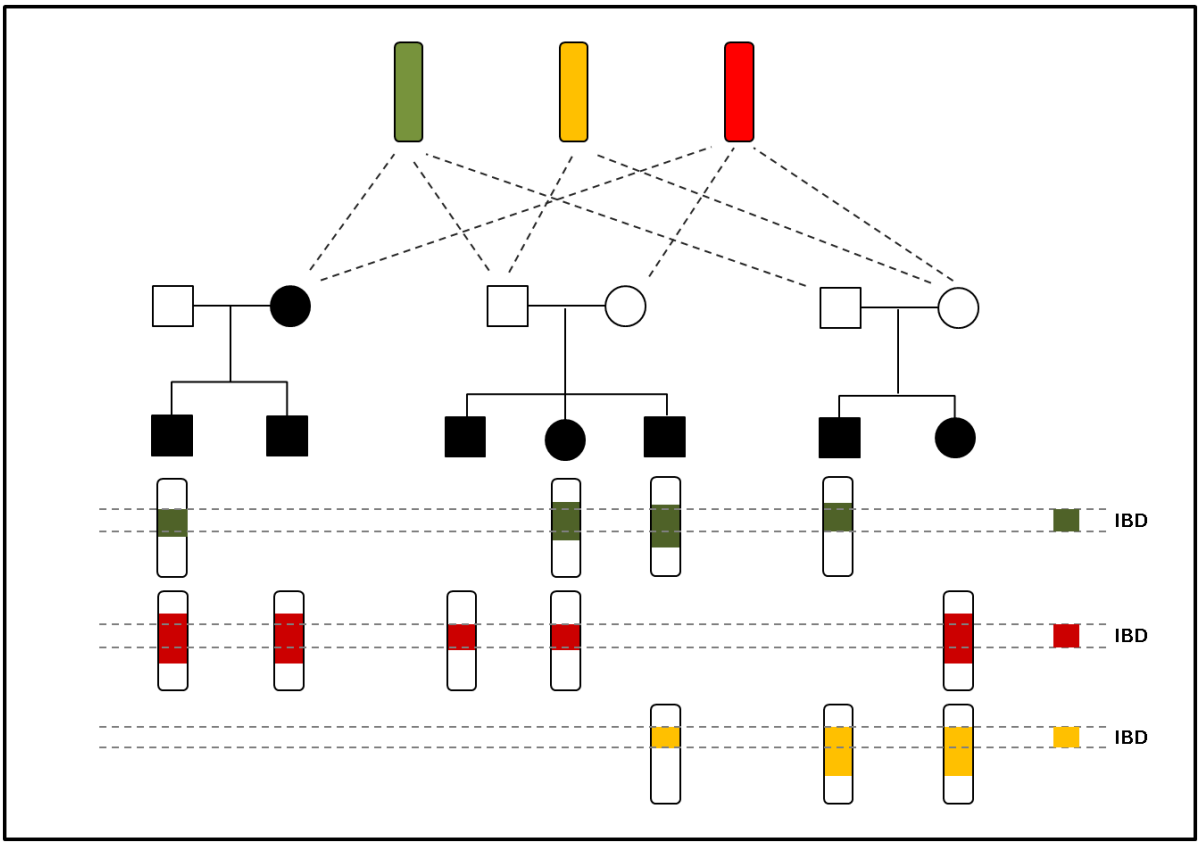

Figure 4.2: Theoretical scenario to explain intra-familial genetic heterogeneity. Here the presence of only 3 risk loci is hypothesized and the sharing across 3 families from the same population is simulated. Since the 3 loci map on different chromosomes, they segregate independently, thus patients from the same family may show different combinations of factors. Consequently, loci don't co-segregate with the disorder and would unlikely produce significant linkage peaks. Considering the features of the investigated sample, however, these loci are probably located in founder haplotypes, that are therefore identical by descent (IBD). A strategy aiming at tracking IBD haplotypes could be thus effective for the identification of these risk factors.

4.3 Population studies from IBD data

4.3.1 Genome-wide IBD sharing and biological relationships between subjects

Since IBD analysis was a novel approach, completely set up in this work, the reliability of IBD data was initially evaluated. For this purpose, IBD regions were used to determine the portion of shared genome between each of the analyzed pair. This measure represents an estimate of the genetic similarity and should reflect the type of biological relationship existing between a specific pair of subjects.

Shared genome was envisioned as the fraction of alleles that were IBD (r coefficient, or relatedness); genetic similarity was thus computed as the sum of the sizes of IBD1 regions and twice the sizes of IBD2 regions over the total genome size captured by the genotyped SNPs.

$$r = \frac{(IBD1 + 2 * IBD2)}{tot\ genome}$$

Calculated r coefficients are plotted in figure 4.3 and summarized in table 4.1; subject pairs were grouped according to their known relationships so that estimated values could be compared with the expected ones for each category. As shown, IBD data provided coherent results for all degrees of kinship. As an explicatory example, for parent-offspring pairs, who should share half of their genome, calculated r was on average 0.49954. Boxplots revealed also some outliers, that were further investigated. The majority of inconsistencies were obtained from an individual from family 4 (4_10); since none of the relationships with the other family members was correctly estimated, this sample had been probably misclassified. Other issues occurred with family 115 and were due to low quality genotyping data. Therefore, in addition to support the robustness of IBD analysis, this approach permitted the identification of some inaccurate data, information then considered in further analyses.

Relationship category	Expected r (relatedness)	Calculated r (mean)	N pairs
Parent-offspring	0.50000	0.49954	54
Full siblings	0.50000	0.49269	63
Half siblings / Grandparent-grandchild / Uncle-nephew	0.25000	0.25303	34
First cousins	0.12500	0.11234	14
First cousins once removed	0.06250	0.06699	5
Second cousins	0.03125	0.04261	1
Unrelated	0.00000	0.00515	19,135
Unrelated (independent population sample)	0.00000	0.00069	528

Table 4.1: Relatedness (r) of pairs of individual according to their biological relationships. Calculated values with IBD data were comparable with expected ones for all the degrees of kinship.

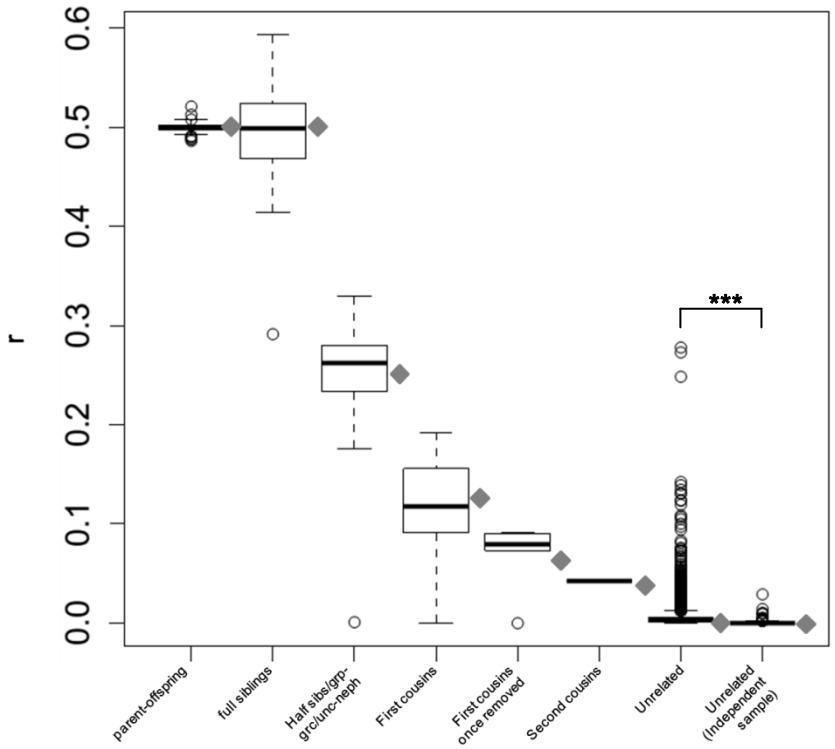

Figure 4.3: Relatedness (r) calculated from IBD data. Pairs of subjects were categorized according to their degree of kinship. Expected values for each category are indicated by grey diamonds. For each degree of kinship, calculated values were coherent with the expected ones. Outliers were ascribable to: the pair of siblings in family 115 in full-siblings category; pairs including the subject 4_10 and her relatives in the other categories. Unrelated pairs from the investigated population show a relatedness significantly greater than the one between unrelated pairs of an independent Italian sample ($p<2.2e-16$).

Since IBD regions could offer a reliable measure of genetic similarity in relatives, unrelated individuals were subsequently inspected. Interestingly, these pairs showed relatedness values considerably higher than expected, since basically no sharing was theoretically predictable. The average r was indeed significantly higher than the one computed in an independent sample of 33 individuals of Italian origin, analyzed with the same metrics ($p<2.2E-16$, Welch two-sample t-test) (figure 4.3). More in details, IBD estimates revealed that a great proportion of unrelated pairs are actually separated by no more than 4-5 generations (figure 4.4). From this evidence it was possible to conclude that this population sample has some characteristics similar to a genetic isolate, where the reduced migration fluxes and the assortative mating have led over time to an average increase of genetic relatedness.

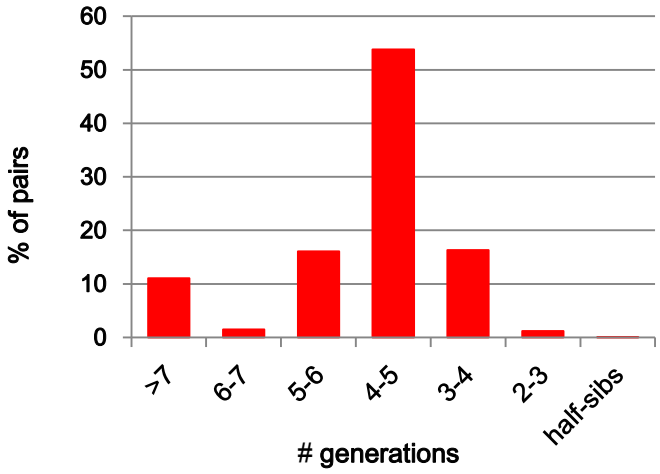

Figure 4.4: Relatedness (r) of unrelated pairs. The number of generations between the members of a pair were estimated according to their r. This computation revealed that more than 50% of pairs were separated by 4-5 generations.

4.3.2 Cluster analysis

After the emerging of genetic links between theoretically unrelated individuals, a connected question was whether the sample was homogeneous or if some of the families were more related than others. To address this issue, a member of each family branch was selected, defining a subset of 115 'family founders'. On the basis of estimated genetic similarity (r), specific analyses were performed to identify eventual population subgroups

Multidimensional scaling revealed the unexpected presence of two main clusters of 'family founders' (figure 4.5). Remarkably, in-depth examination of registry data suggested that the clusters roughly corresponded to the local origin of individuals. In fact, samples were collected from two main close areas (figure 3.1), named Chioggia (the main town) and Sottomarina (its neighborhood), referring to the same Mental Health Center.

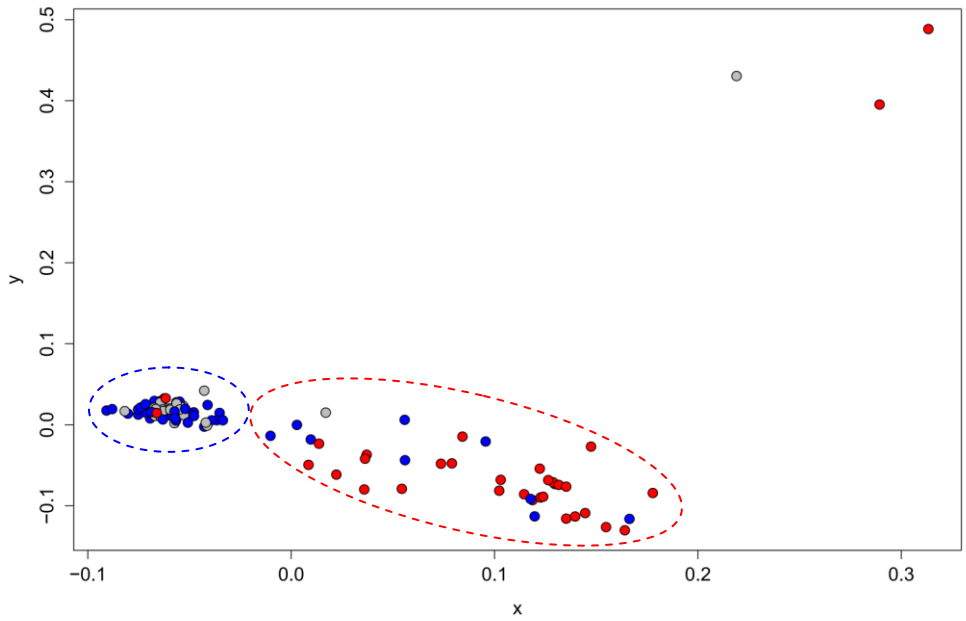

Figure 4.5: Multidimensional scaling of genetic distances (1-r) in 115 'family founders'. Two main clusters of subjects appeared, broadly corresponding to the local origin from Chioggia (blue) or Sottomarina (red). Unknown origins are shown in grey. The three individuals on the top right area of the panel were identified as 79_5, 53 and 71, who are particularly highly related, thus placed in an independent cluster, corresponding to the one indicated with ‡ in the dendrogram below (figure 4.6).

To get more insights, 53 unrelated individuals were further added to the original subset: beyond the 33 samples of Italian origin previously cited, 20 samples were retrieved from the HapMap project, 5 for each main available ancestries, Yoruba (Africa), Han Chinese in Beijing (China, Asia), Japanese in Tokyo (Japan, Asia) and Europeans in Utah (Europe). HapMap samples were processed as already described to obtain genetic similarities. Hierarchical clustering of the total 168 individuals is represented by the dendrogram in figure 4.6. As evident, samples were correctly grouped according to their ancestry. Again, the 115 family founders from the studied population were split in two main

branches, which significantly matched their local origin (p=2.12 e-8). Within the Sottomarina cluster, dendrogram displayed 2 sub-clusters, composed by 2 (¥ in the figure) and 3 (‡ in the figure) individuals respectively. These subjects didn't show an increased genetic distance from the other members, rather a very high relatedness among them. Therefore the separation results from the attempt of the algorithm to accommodate this elevated genetic similarity. About the Chioggia cluster, then, the majority of samples clustered to a branch distinct from the Italian/European subjects; although this might reflects the likely genetic isolation, the difference was not significant.

Concluding, the investigated population can be viewed as composed by two extended pedigrees, each constituted of smaller familiar nuclei connected by substantial genetic links. Considering the high significance of the analysis, clusters were used to infer the belonging to Chioggia or Sottomarina sub-group for the 24 family founders with uncertain origin. Discordances between the predicted origin and the referred information were carefully re-evaluated by checking familial registers and some data were corrected (appendix 7.3).

Figure 4.6 (on the opposite page): Dendrogram of genetic distances calculated from IBD data in 168 subjects (115 'family founders' and additional unrelated 53 individuals from independent populations). Samples were clustered according to their ancestry (African, Asian and European). Moreover, Chioggia and Sottomarina were significantly distinguished (p=2.12 e-8, linear model with origin as co-variate). The two sub-cluster of particularly related individuals are indicated by the ‡ and ¥ symbols (see main text for details). Individuals from Chioggia were grouped separately from the European/Italian ones, indicating a likely genetic isolation, although the difference was not significant.

Sottomarina
‡
¥
Chioggia
Europe/Italy
China/Japan
Africa
0.6
0.8
1.0
1.2
1.4
1.6
Height

4.4. IBD mapping: tracking haplotypes shared by patients

Considering the population features of a genetic isolate, combined with the high prevalence of SCZ/BPD, IBD mapping was sought as a strategy to track susceptibility loci for these complex disorders. The foundation of the approach was the identification of clusters of individuals with identical haplotypes, starting from the list of IBD regions obtained from pairwise IBD analysis. This couldn't be achieved by simply grouping all pairs with IBD segments encompassing the same specific position, since the existence of two homologous chromosomes had to be taken into account. An explicative example is represented in figure 4.7 As shown, if the same region had been found IBD in three pairs of three individuals, this wouldn't necessarily imply that all of them share an identical haplotype; an individual, in fact, could share a different haplotype with each of the other two subjects (figure 4.7a). A simplistic interpretation of IBD data could thus lead to erroneous conclusions. The problem could be overcome thanks to a logical rule: whenever the same region was IBD in all possible pairs of at least 4 individuals, an identical haplotype shared by all the subjects had to be assumed (figure 4.7b) (Thompson, 2013). This was then the minimum requirement for building a cluster (figure 4.7c).

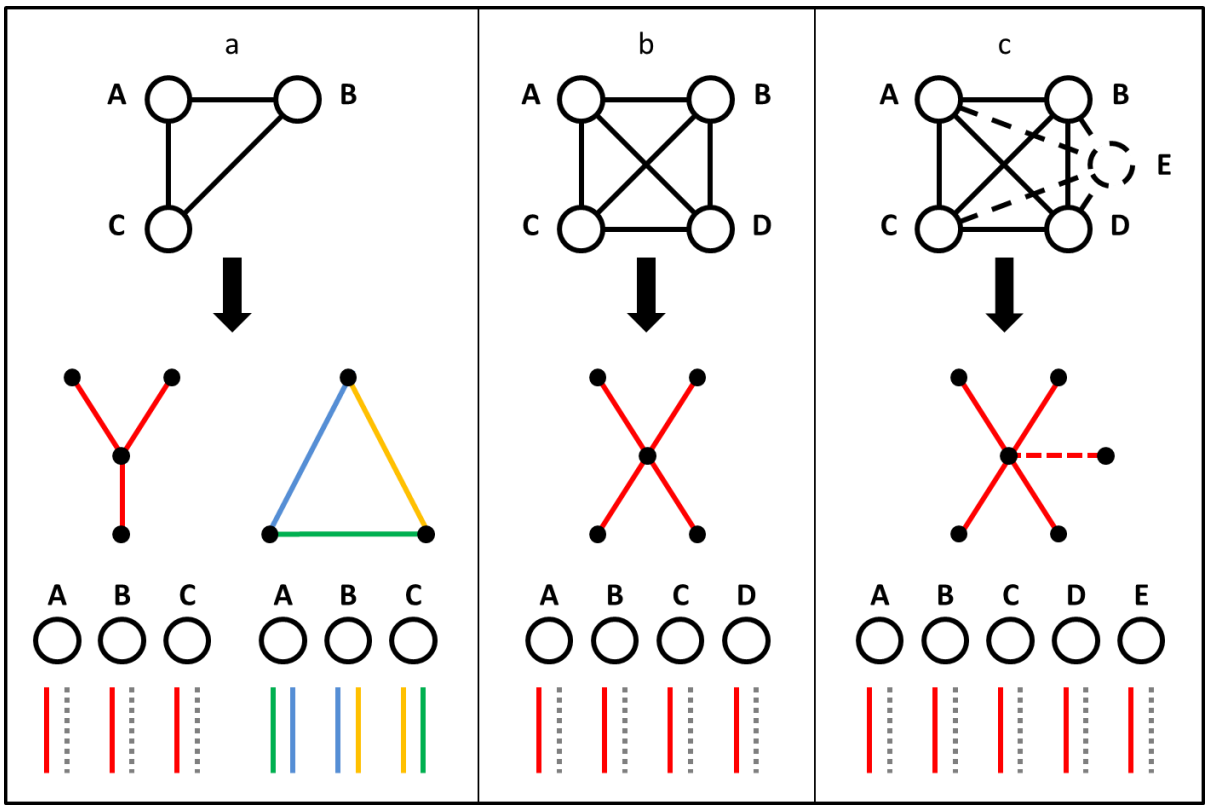

Figure 4.7: Resolution of haplotype clusters from pairwise IBD data. If the same region is IBD in all the three pairs of 3 subjects (A,B and C), resolution of the cluster is not unique (panel a). Thus the assumption of a common haplotype may be incorrect. Conversely, if the same situation occurs for the 6 pairs of 4 individuals (A,B,C, and D), then the resolution have a unique outcome, implying a shared haplotype for all the subjects (panel b). This was therefore the basis for the building of clusters and a further individual (E in the example) was added whenever it displayed IBD in the same region with all the other members of the cluster (panel c).

Haplotype clustering was performed systematically, exploiting a sliding window of 200 markers. The entire genome was scanned and clusters were built position by position, considering pairwise IBD segments overlapping the window. Consensus clusters were then calculated to finally obtain a complete, genome-wide map of IBD shared haplotypes, with relative chromosomal coordinates and a list of subjects carrying the specific alleles.

From the examination of IBD map, important observations emerged. First, a subset of individuals (6 cases from 3 families, 8 single or isolated cases and 1 control) could never be included in any haplotype cluster. Notably, in the previously described dendrogram these subjects were grouped in the sub-branch identified by Italians and Europeans (figure 4.6). This evidence indicated that these individuals were unlikely belonging to the population sample under investigation (appendix 7.3).
Second, identical haplotypes could be tracked across multiple families, revealing a puzzle of shared segments in different chromosomes (figure 4.8). 24% of haplotypes were restricted to a single family, 44% encompassed 2 families, while 32% were shared across 3 or more pedigrees. Interestingly, 20% of them were specifically present in patients with SCZ, while 5% were exclusive of BPD subjects (figure 4.8c,d). The majority of haplotypes, however, crossed diagnostic boundaries. Surprisingly, haplotype sharing didn't reflect the population stratification highlighted by the cluster analysis. A consistent number of regions, in fact, were common to families both from Chioggia and Sottomarina, suggesting some level of admixture between the two subgroups. As a last remark, particular combinations of haplotypes on different chromosomes could be found co-segregating within and between families. This evidence was especially relevant because it could point to possible interplays of genetic factors contributing to risk.

IBD map thus provided a valuable tool to dissect genomic sharing both at familial and at population level. With the aim of focusing on IBD haplotypes mostly relevant to the disorders, a prioritization was performed by selecting from the total IBD map (IBD_{tot}) the haplotypes encompassing more than two pedigrees and shared by at least half of the patients within each nucleus (IBD_{sel}). 648 overlapping haplotypes emerged from this step, with a size between 3 and 67 Mb.
These haplotypes are the most shared within and between families and under the original hypothesis they could carry susceptibility alleles with strong effect on disease liability.

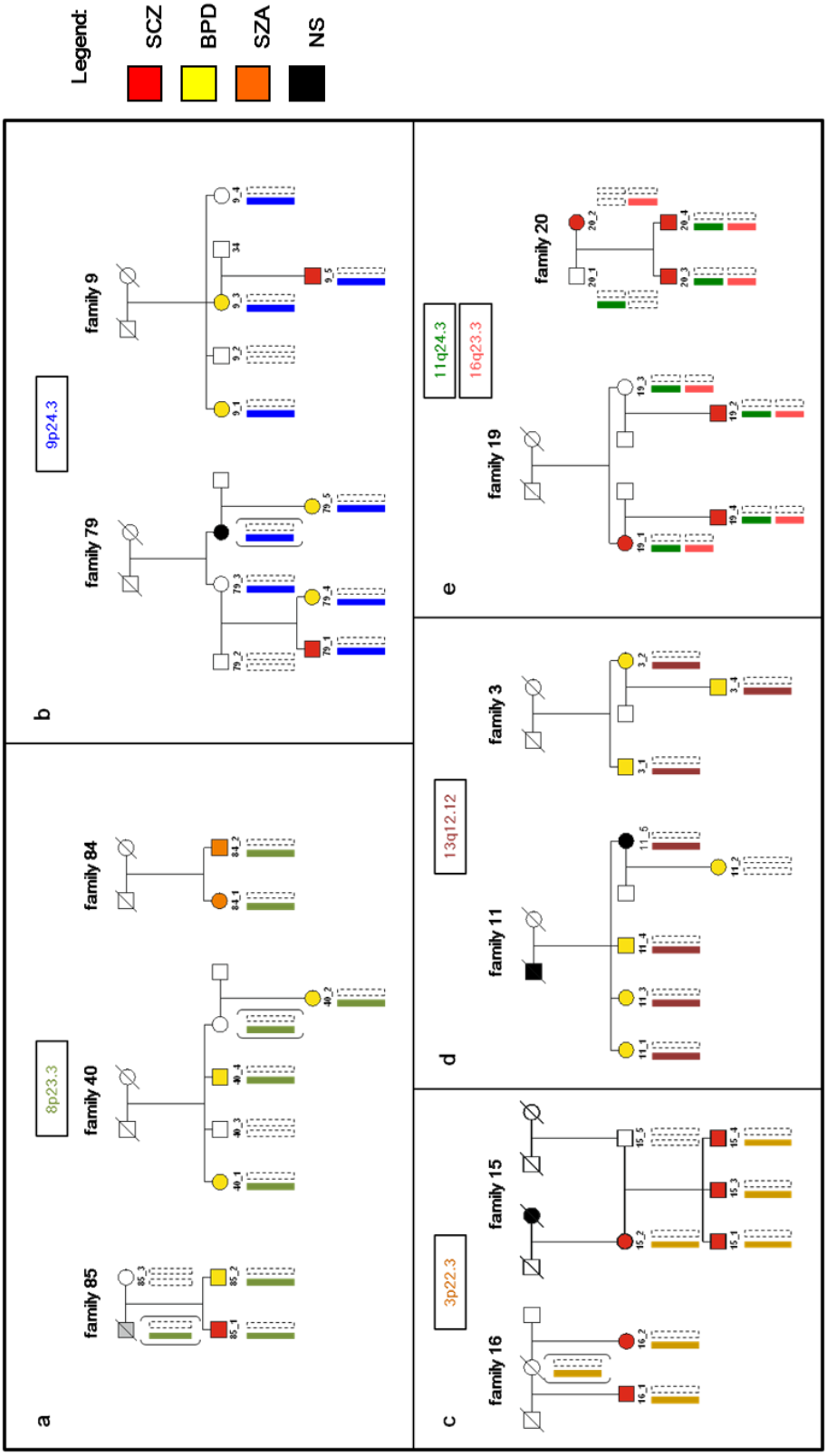

Figure 4.8: sharing and segregation of haplotypes determined with IBD map. Some interesting examples of loci shared across different families are reported. Panel a show a haplotype specific for the Sottomarina sub-group; conversely, the haplotype in panel b was shared by samples from the two population clusters (family 79 from Sottomarina, family 9 from Chioggia). Some phenotype specificities also emerged. Panel c represents a haplotype co-segregating with SCZ, while an panel d illustrates an analogous situation for BPD. Finally, a combination of haplotypes on two different chromosomes is depicted in panel e. SCZ: schizophrenia; BPD: bipolar disorder; SZA: schizoaffective disorder; NS: not specified, referred affected with SCZ or BPD.

4.5 Whole-exome sequencing

4.5.1 Variant annotation and filtering

For the detection of rare variants possibly implicated in SCZ/BPD, whole-exome sequencing (WES) was performed on 17 patients; analyzed subjects were selected on the basis of IBD map, in order to cover the most interesting IBD haplotypes: with this approach, 75% of IBD_{sel} haplotypes were theoretically analyzable. Next-generation sequencing was carried out with IonTorrent™ technology (Life Technologies).

Annotation of variants was initially achieved with *IonReporter*™, the standard software for IonTorrent™ data processing. In a second step, variant frequencies in reference databases were re-annotated exploiting a specifically developed pipeline. This strategy was adopted to overcome some issues encountered with the use of *IonReporter*™, leading to a consistent loss of variants during frequency filtering. Among the most relevant ones is the annotation of variants as known polymorphisms (SNPs) by considering only the genomic position, without any allele check. Another problem derives from inaccuracies in the reference human genome (GRCh37/hg19 assembly); for some SNPs, the reported reference allele doesn't correspond to the most frequent one in the general population (mAiR, minor Allele in Reference). As a consequence, common alleles are typically called as variants, while homozygous genotypes for rare alleles are not detected (see par. 3.8 further details).

The developed pipeline is based on the creation of a variant sharing table, listing all the identified and annotated variants across the 17 exomes, with their relative presence in the sequenced patients. The simultaneous evaluation of all the exome data allowed the detection of mAiR situations, thus the proper classification of rare alleles. Frequency of variants was checked in three different databases: dbSNP, 1000Genomes and ExAc. Rare variants were defined as those reported in dbSNP, with a MAF<1% in 1000Genomes database and absent or with a frequency <5% in ExAc. Novel variants were instead those not present in dbSNP. Another advantage of working with the variant sharing table was the possibility to immediately prioritize variants according to IBD map. In subsequent steps, two narrower classes of rare or novel variants were specified, the ones mapping in any of the IBD haplotypes (IBD_{tot}) and the ones in selected IBD haplotypes (IBD_{sel}).

Of the 138,725 variants identified in the 17 patients, 17,385 were rare and 12,832 novel, for a total of 30,217. All the following investigations were conducted starting from this subset. Remarkably, 4.5% of these selected variants corresponded to mAiR situations. Further, 6,621 variants mapped in IBD_{tot} haplotypes and 558 in IBD_{sel} ones. Interestingly, in all of these categories, the majority of variants were non coding, in line with what reported in other studies, both on common and rare variation (Ament et al., 2015; Ripke et al., 2014) (figure 4.9). Additionally, variant distribution in relation to the effect showed no substantial changes in the different filtering processes; for example, missense

variants represented always about 20% of the total. This indicates that any of the variant types was preferentially more shared in patients. Among the coding variants, a supplementary category was defined, including novel and rare variants with a likely damaging effect (frameshift insertion/deletions, splicing alterations, nonsense, stoploss and missense predicted deleterious by CONDEL software), generally named loss of function (LoF). Also this subset was mapped into IBD_{tot} and IBD_{sel} haplotypes.

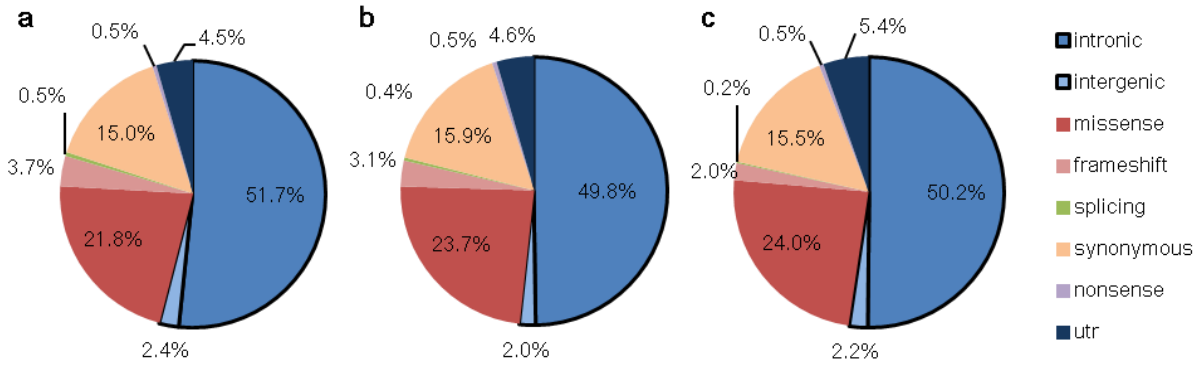

Figure 4.9: Effect of the variants progressively prioritized for frequency (MAF<1% or novel) (a), IBD_{tot} halpotypes (b) and IBD_{sel} haplotypes (c). As shown, the proportion of variants in the different effect categories was almost unchanged across the prioritization process, suggesting that any of the variant type was particularly more shared in patients. Additionally, non coding variants were always predominant, indicating their possible role in etiology of SCZ/BPD.

4.5.2 Functional enrichment analysis

Functional enrichment analysis aimed at determining whether prioritized variants were specifically affecting any biological process. DAVID online tool (Huang et al., 2009a; Huang et al., 2009b) was exploited for this purpose, focusing on KEGG and Reactome functional gene annotations. The previously defined categories of rare or novel variants were tested. As reported in table 4.2, two main processes consistently emerged across all the subsets of IBD variants: axon guidance and synaptic transmission; even more, axon guidance as annotated by Reactome and synaptic transmission showed significance after multiple-test correction in IBD_{tot} (see appendix 7.4 for complete results). Remarkably, both of these processes have been described as possibly relevant for SCZ/BPD (see par 1.4). Four more pathways were repeatedly found and inspected because possibly related to the previous functions: PIP signalling, signalling by rho GTPases, ECM-receptor interactions and cell adhesion molecules. These are in fact basic cellular cascades intervening in a variety of biological processes. Cell-cell and cell-matrix interactions are particularly involved in axon guidance, as fundamental for neuron projection and connectivity during development. Similarly, phosphatidylinositol and rho-GTPase signalling are active in postsynaptic signal transduction. Thus

these outcomes suggest a broad convergence again on axon guidance and synaptic transmission functions, as sustained by the overlap between genes in these latter categories and in the four singalling cascades in the analyzed dataset.

IBD_{sel} set of variants showed no significant enrichment for any pathway reported in DAVID. These results could either simply reflect the considerably lower number of genes tested, or they could also indicate that many of the variants in these functional categories are not largely shared by patients. It's however notable that almost all the nominally significant pathways in this analysis were those already described, ascribable to axon guidance and synaptic transmission. Instead, the lack of enrichment in LoF categories, even after IBD_{tot} mapping, suggests that the significant outcome in IBD_{tot} genes was not driven by coding variants with predicted damaging effects.

Functional annotation class	p-value (adjusted p-value)				
	all	IBD_{tot}	IBD_{sel}	LoF	LoF IBD_{tot}
Synaptic transmission (Reactome)	**2.93E-04 (2.06E-02)**	**6.02E-03 (9.75E-02)**	3.87E-02 (3.61E-01)	NS	NS
Axon guidance (Reactome)	**1.30E-03 (3.03E-02)**	**3.82E-05 (1.30E-03)**	2.60E-02 (5.91E-01)	2.32E-02 (5.5E-01)	3.77E-02 (8.64E-01)
Axon guidance (KEGG)	**6.68E-04 (1.88E-02)**	3.72E-02 (4.14E-01)	1.63E-02 (8.87E-01)	4.97E-02 (5.35E-01)	NS
PIP signalling	**2.31E-04 (9.13E-03)**	5.64E-03 (1.20E-01)	2.45E-02 (8.06E-01)	2.35E-02 (3.71E-01)	NS
Signaling by rho GTPases	**5.48E-04 (1.93E-02)**	**2.79E-05 (1.89E-03)**	NS	2.77E-02 (3.80E-01)	7.80E-03 (3.34E-01)
ECM-rec. interactions (KEGG)	**3.42E-05 (3.40E-03)**	**2.59E-07 (5.11E-05)**	NS	5.68E-07 (1.11E-04)	9.11E-04 (1.44E-01)
Cell Adhesion molecules (KEGG)	NS	NS	4.75E-02 (8.82E-01)	3.22E-02 (4.13E-01)	4.22E-02 (5.59E-01)

Table 4.2 Summary of functional enrichment analyses for different categories of prioritized variants, For each category, a list of genes was compiled, including those genes carrying at least one variant of the specified type. Only annotation classes that were at least nominally significant were considered and the most consistent results are here reported. For multiple testing correction, Benjamini method was selected and significant outcomes are highlighted in bold.

Both axon guidance and synaptic transmission are large and general categories including several genes and functions. For this reason, a more detailed investigation was envisioned. Many signalling pathways are annotated as part of axon guidance both in KEGG and in Reactome. However, rare or novel variants mapping in IBD haplotypes (IBD_{tot} and IBD_{sel}) were affecting almost exclusively *NCAN* signaling for neurite outgrowth. Even more, genes carrying these variants were located in specific nodes of the pathway (figure 4.10). For example, as highlighted in figure 4.10b, neural cell adhesion

molecule 1, encoded by *NCAM1*, interacts with several other proteins for the development of neuron projection, such as neurocan (NCAN), collagens and glial cell derived neurotrophic factor (GDNF); novel or rare variants in IBD haplotypes were detected in all of these cited genes.

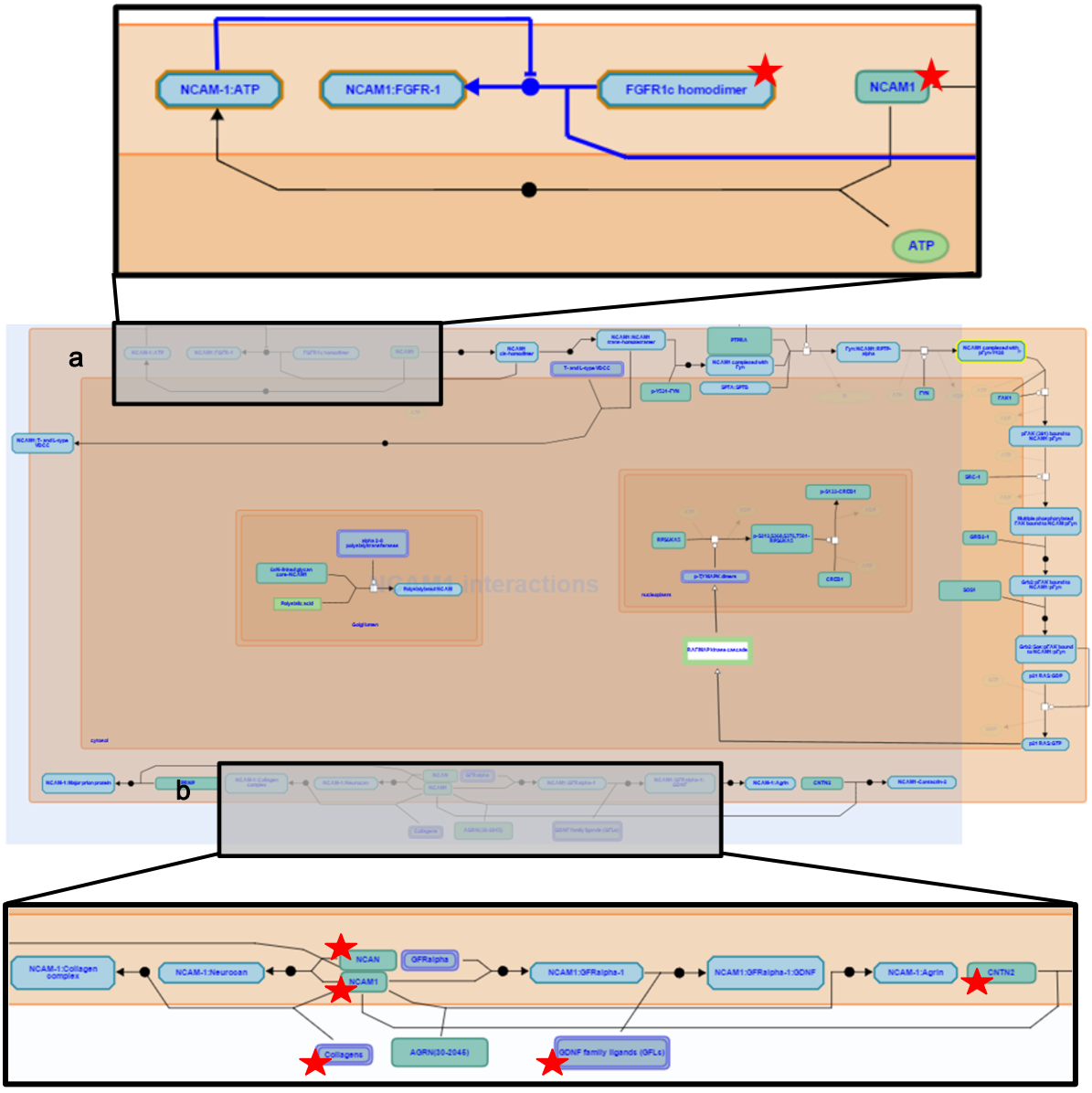

Figure 4.10: NCAM singnalling for neurite outgrowth pathway, a sub-process of axon guidance, according to Reactome annotation. Rare or novel variants mapping in IBD haplotypes were specifically affecting portions of this pathway, as highlighted by panels a and b. Genes carrying at least one variant are marked with a red star. Several of these genes were direct interactors, as *NCAM1* and *FGFR1*, or *NCAM1* and *NCAN* (panel b).

Analogously, synaptic transmission is a broad category that includes all types of presynaptic, synaptic and postsynaptic mechanisms concerning neurotransmitter release and actions. Nevertheless, numerous IBD_{tot} and IBD_{sel} variants were in genes involved in glutamatergic synapse functions, as the glutamate metabotropic receptors *GRM7* and *GRM8* or the NMDA ionotropic receptors *GRIN2A* and *GRIN2B*.

4.5.3 Gene-set enrichment analysis

To further dissect these observations, a collection of gene sets were retrieved from the studies of Fromer et al. and Kirov et al. (Fromer et al., 2014; G Kirov et al., 2012). These gene sets were assembled from proteomic data analyzing the full collection of interactors of specific postsynaptic complexes intervening in NMDA signalling, such as ARC (activity-regulated cytoskeleton-associated scaffold protein), PSD-95 (post-synaptic density protein-95) and NMDA receptors themselves. Additionally, two more sets, glutamatergic synapse and VGCCs (voltage-gated calcium channels), the latter composed by synaptic channels mediating Ca^{2+} influx upon membrane depolarization, were manually assembled from KEGG annotations.

As shown in table 4.3, the previously implicated ARC and PSD-95 complexes were significantly enriched in genes affected by all categories of variants tested. Coherently, NMDA receptor and glutamatergic synapse gene sets evidenced the same outcome. Therefore, the prioritized variants based on the IBD sharing, including IBD_{sel} haplotypes, were selectively clustering in genes involved in glutamatergic synapse transmission, particularly in postsynaptic mechanisms mediated by NMDA receptors. The VGCC gene set, however, still showed a significant enrichment only in IBD_{tot} variants, indicating again a possible more private distribution of these alleles across families.

Taken together, these results indicate a load of rare or novel variants affecting specific portion of axon guidance and glutamatergic synapse processes in SCZ/BPD patients.

Gene set (n genes)	**p-value (adjusted p-value)**		
	all	**IBD_{tot}**	**IBD_{sel}**
Glutamate synapse (92)	3.13E-10 (1.88E-09)	1.70E-06 (5.10E-06)	3.28E-03 (6.68E-03)
NMDA receptors (61)	9.48E-09 (2.84E-08)	7.80E-05 (1.18E-04)	9.12E-03 (1.37E-02)
ARC (28)	4.20E-02 (4.20E-02)	2.41E-05 (4.82E-05)	2.05E-02 (2.46E-02)
PSD-95 (117)	1.15E-03 (1.38E-03)	2.41E-03 (2.41E-03)	3.34E-03 (6.68E-03)
VGCC (26)	9.07E-07 (1.36E-06)	9.29E-04 (1.11E-03)	1.03E-01 (1.03E-01)

Table 4.3: Gene-set enrichment for genes carrying at least one novel or rare variant in IBD_{tot} or IBD_{sel} haplotypes. 5 gene sets were tested considering a hypergeometric distribution. Adjusted p-values were calculated with Benjamini method. Significant outcomes are highlighted in bold.

4.5.3 Recurrence of functionally related variants in families

One of the main advantages of the IBD map was the opportunity to infer the sharing and the segregation in non sequenced individuals of any variant located in IBD haplotypes. On the basis of this feature, the entire sample was investigated at a family and population level.

A global overview of single families revealed that each nucleus carried multiple variants in the axon guidance and synaptic transmission pathways (figure 4.11). Since many of the genes with variants were not annotated in KEGG or Reactome, the original sets were expanded with functionally related categories form DAVID or Gene Ontology, for a total of 87 genes in the axon guidance cluster and 66 genes in the synaptic transmission one. The recurrence of functionally related variants in a family is compatible with a polygenic nature of the disorders, where multiple risk factors contribute to risk. Interestingly, the 13 controls showed in general a lower amount of such variants. This was unlikely attributable merely to the fact that families included several subjects, since the selected variants were mapping in shared IBD haplotypes, reducing the possibility that each family member contributed with different variants independently. Indeed, the higher loads were not observed in the largest families.

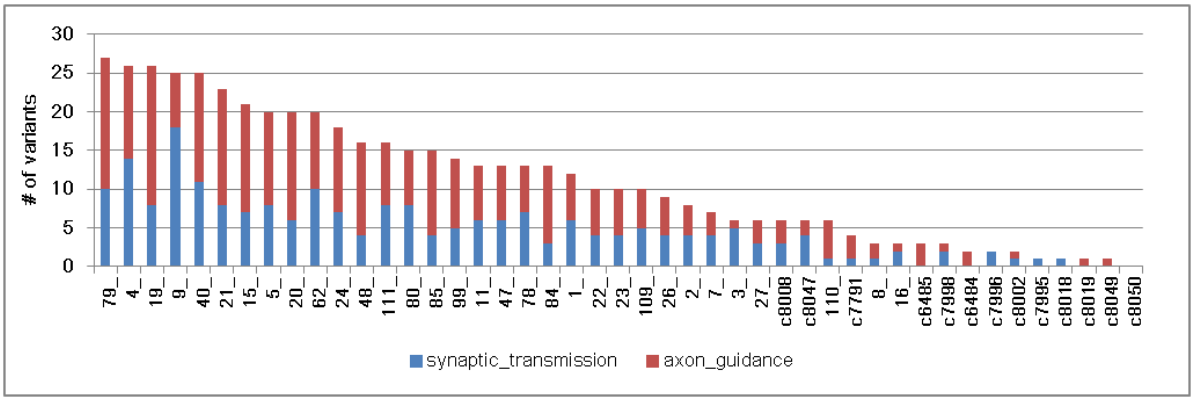

Figure 4.11: Burden of rare or novel variants involved in axon guidance (red) or synaptic transmission (blue) in families. The proportion of variants in the two functional categories is different in the families, indicating a certain heterogeneity. Interestingly, controls seem to have a lower load of these mutations.

At the same time, data evidenced some genetic heterogeneity between different families, with variable loads of variants on the two considered functional processes. For example, family 3 and 9 have a higher proportion of rare or novel alleles affecting synaptic transmission, while the opposite situation can be outlined for families 84 and 85.

To get more insights on variants individually, a further investigation was performed. Detailed information about all the analyzed variants can be found in appendix 7.5 The single variant examination revealed that some of the alleles in functionally related genes were co-segregating in the same pedigree, even if mapping on different chromosomes. Some illustrative results are reported in figure 4.12. As shown, patients from family 85 carry two variants in *NCAN* and *NCAM1* genes, involved neurite outgrowth, as described above (figure 4.10). Analogously, two variants in two metabotropic glutamate receptors (*GRM7* and *GRM8* genes) perfectly co-segregate with the disease in family 62.

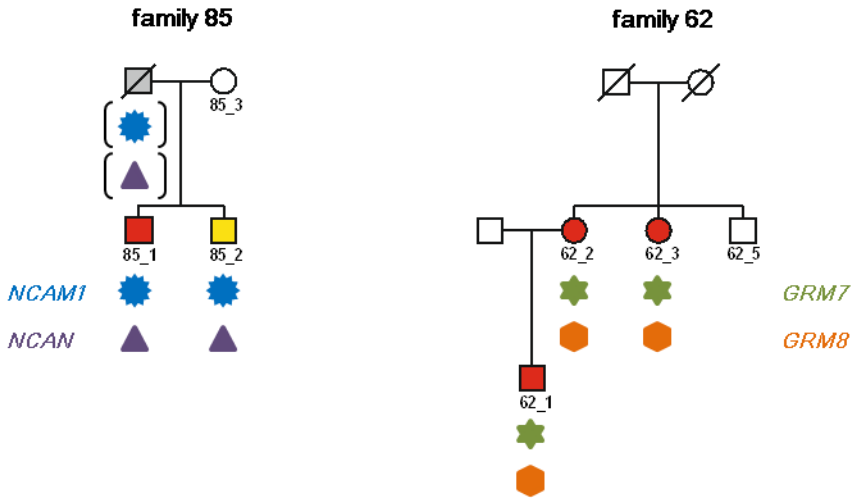

Figure 4.12: Burden of functionally related variants in families. In family 85 an intronic variant in NCAM1 was co-segregating with a missense substitution in NCAN. These genes encode for proteins which directly interact in the neurite outgrowth process. In family 62, two synonymous variants in the glutamate metabotrobic receptors GRM7 and GRM8 were detected in all the patients.

It is noteworthy that some of these variants, especially the ones in IBD_{sel} haplotypes, where shared in different pedigrees. A typical scenario is represented in table 4.4 and figure 4.13 for a subset of families. In the proposed example, a variant in *CACNA1E* gene, a VGCC, is present in family 9 and in one of the two patient of family 111. In family 111, a second variant in the *FNBP1L* gene, involved in Ca^{2+} mediated synaptic plasticity, is shared by both the patients and additionally co-segregates with SCZ in family 15. Interestingly, in this latter family, also a variant in the *DLG4* gene is passed from the affected mother to all the three affected sons; Dlg4, also known as PSD-95, is a key protein in postsynaptic modulation of NMDA receptors and related plasticity. Similar considerations can be outlined for the other shown variants as well.

gene	functional category	families					
		9	111	15	24	78	additional inds
CACNA1E	synaptic transmission						17_1;23_4
FNBP1L	synaptic transmission						47_1;90_2
DLG4	synaptic transmission						4_1;64_1
FGFR1	axon guidance						90_2
NLGN1	axon guidance						fam7;21_2;47_2

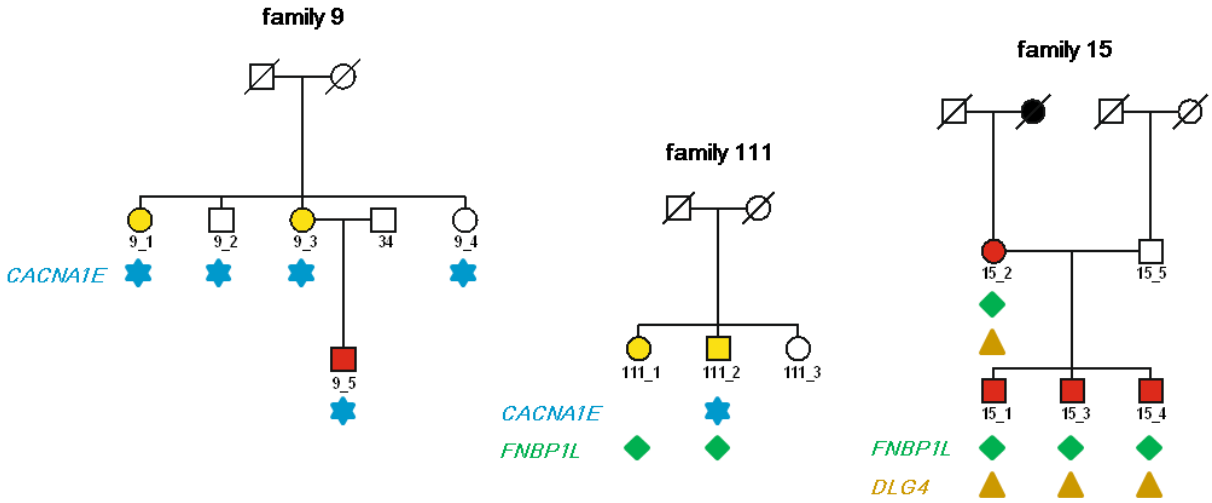

Table 4.4 and Figure 4.13: A schematic representation of the complex puzzle of loci emerging from sharing analysis of variants. The sharing of variants in some of the candidate genes is shown in the table for a subset of families . Each of the variants has a different sharing pattern, encompassing different pedigree. The result is a polygenic load of functionally related alleles in each family. A detailed view of the sharing of three variants is depicted in the figure underneath. Although some of the candidate risk alleles show a perfect co-segregation with the disorders (e.g. FNBP1L, DLG4), some others show incomplete penetrance, or are not shared by all the patients of a family (e.g. CACNA1E); this suggests again the absence of a major gene and instead the presence of an intra-familial heterogeneity, as initially hypothesized, and typical of complex disorders.

Another relevant observation emerging from the sharing analysis was the presence of multiple haplotypes encompassing the same locus, but associated with different variants. For example, in the *DLG4* gene two variants were detected, mapping in just as many haplotypes, shared across distinct groups of families (appendix 7.5). The same situation was identified for the *NLGN1* gene. The evidence that different rare or novel variants on the same genes were shared by patients strongly support the involvement of *DLG4* and *NLGN1* in the etiology of the disorders.

As a final remark, sets of variants could be found co-segregating not only within a family, but in different pedigrees. This led to the hypothesis that risk could be substantially increased by specific combinations of a few functionally related alleles. The most remarkable results were observed for variants in *ARHGAP32* and *CDH13*, two genes acting on the development of neuronal projections through cell adhesion and cytoskeleton remodelling. These two variants, mapping on chromosomes 11 and 16 respectively, were found together in 5 patients from families 19 and 20 (figure 4.14). Other combinations were detected, although with a more limited sharing in patients from the same families. For example, *CACNA1E* and *PRKCB* rare alleles were both presents in families 9 and 78. PRKCB is a protein kinase active in postsynaptic neurons upon Ca2+ signaling, that is mediated by VGCCs. Analogously, the *CNTN2* and *SEMA3C* genes are both playing a role in axon guidance. A rare missense variant with a loss of function prediction was found in each gene; the combination of these two variants was detected in cases of families 40 and 85. Further investigations in this sense will be required to confirm this fascinating theory.

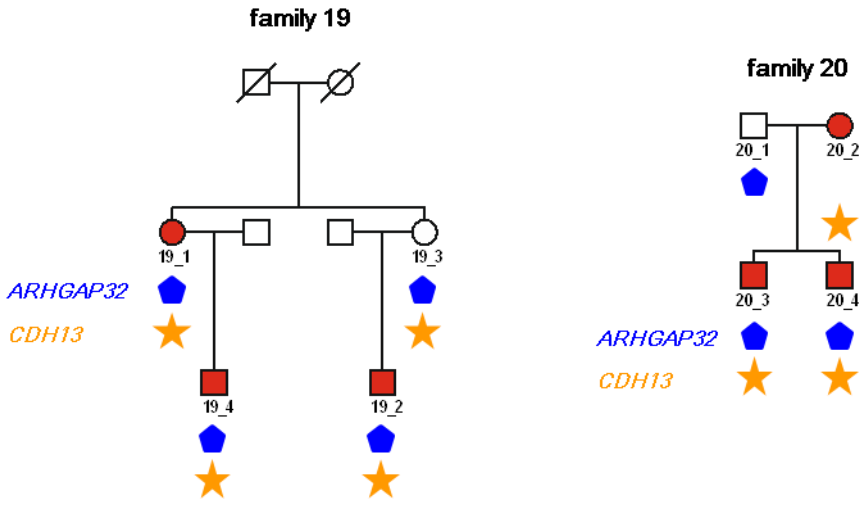

Figure 4.14: Combinations of variants detected in several patients of different pedigrees. Here is shown the example of two missense substitutions in the *ARHGAP32* and *CHD13* genes. The combination is perfectly segregating in family 19 and is further detected in the two bothers from family 20. Both these genes are involved in the formation of neuron projection, thus the simultaneous presence of susceptibility variants could substantially increase the risk of developing SCZ.

Concluding, these analyses overall highlighted a puzzle of different variants, each with a proper sharing pattern across families, resulting in a burden of functionally related rare alleles in a specific pedigree.

5. Discussion and conclusions

5.1 The development of a novel approach combining IBD mapping and whole-exome sequencing

Schizophrenia (SCZ) and bipolar disorder (BPD) are among the top 20 causes of disability worldwide (WHO). Despite this, their etiology is still largely unknown and generally poor outcomes are associated with both the disorders. In fact, available drugs are only partially effective and associated with strong side-effects. The understanding their genetic architecture results therefore extremely important, with the aim of developing more successful treatments. Recently, an involvement of rare variants has been repeatedly proposed (Kerner, 2015; Lee et al., 2012; Neale & Sklar, 2015), but their investigation in large case-control cohorts have produced limited results (Fromer et al., 2014; Purcell et al., 2014). In this work, a different type of sample is analyzed, composed by 36 families with high recurrence of SCZ and BPD and 55 additional cases. The peculiarity of the sample is that all individuals have a common ancestry from a small town in the Venetian lagoon, considered a closed community for cultural reasons. The singular high number of SCZ and BPD cases, and specifically of familial cases, offers a unique opportunity to investigate the genetic architecture of these two disorders, as the homogenous background suggests an enrichment of rare risk alleles. Indeed, a higher prevalence of some rare pathogenic alleles has been reported in this population, such as the Factor V Leiden, a risk factor for venous thrombosis, (Gessoni et al., 2010) and a DSC2 (desmocollin 2) mutation causing arrhythmogenic cardiomyopathy (Lorenzon et al., 2015). The latter was also shown to map in a shared founder haplotype.

Copy number variants (CNVs) have been the first type of rare variants implicated with the disorders, especially with SCZ. For this reason, a CNV analysis was initially performed on the population sample. No previously reported SCZ/BPD structural variant was detected, thus rare and novel CNVs were searched. However, the limited sharing of the identified variants suggests that CNVs are unlikely to play a major role in this sample. Nevertheless, the possibility that some of the variants partially contribute to the disorder can't be completely ruled out. Additionally, the method employed for the calling of CNVs has poor performances with small structural variants, so that a threshold of 30kb size was imposed. Further detailed analyses are thus required to get more insights on these variants.

In the light of these findings , a novel and alternative approach was envisioned, exploiting one of the main advantages of familial-based population samples: the possibility to track loci segregating with

the disorders within a pedigree and shared across patients from different nuclei. The approach combines IBD mapping and whole-exome sequencing (WES) for the identification of rare sequence variants that may contribute to genetic risk for the disorders. Recent studies have actually demonstrated the effectiveness of combining a mapping strategy with next generation sequencing in families. Even though linkage analyses have been largely preferred in pedigrees, IBD has been sometimes more successful in uncovering mutations, particularly with rare recessive Mendelian disorders in consanguineous families. This was the case for retinal dystrophy (Coppieters et al., 2014) and non-syndromic intellectual disability (Mir et al., 2014), where the search of homozygous IBD2 haplotypes, also known as homozygosity mapping, was used to prioritize exome variants. IBD haplotypes have also proven to be even more efficient than linkage in restricting candidate loci when analyzing multiple families with familial late-onset Alzheimer's disease (Kunkle et al., 2016). Linkage analysis can be seen as a particular form of IBD mapping, that tracks haplotypes specifically segregating with a trait within families. For this reason, it has a reduced power when dealing with complex disorders, due to the genetic heterogeneity and the incomplete penetrance that characterize risk alleles, in contrast with Mendelian mutations (Neale & Sklar, 2015). Additionally, the eventuality of phenocopies must be taken into account for common disorders. IBD mapping has been already proposed as an alternative strategy to dissect the contribution of different haplotypes when multiple risk factors are involved (Marchani & Wijsman, 2011). In the present work, IBD mapping is applied for the first time in a systematical fashion, as a tool for the identification of loci shared by SCZ and BPD patients.

5.2 Technical considerations about the IBD analysis

IBD analysis was completely designed in this study, starting from genome-wide IBD predictions obtained with Relate software. Several considerations have been made before the selection of the proper tool for IBD detection. Hidden Markov Model-based algorithms are notoriously heavy from the computational point of view, and numerous days were required to run Relate on the 197 samples. Despite, it was retained the most suitable method for the type of sample analyzed. Alternative algorithms, in fact, rely on the identification of identical segments of markers, then determining if the observed haplotype is IBD rather than IBS on the basis of frequency in the sample (BEAGLE fastIBD) (Browning & Browning, 2011) or length (GERMLINE) (Gusev et al., 2009). The lower the frequency and the greater the size, the lower the probability that the haplotype occurs multiple times independently, thus a common inheritance is assumed (Browning & Browning, 2012). However, in a closed population with a recent ancestry, haplotypes might have been quite common and extended, particularly considering the high number of familial cases and the multiple degrees of kinship existing. Indeed, the majority of methods are developed to track IBD among distant relatives. GERMLINE was

specifically tested on this purpose, but poor outcomes were achieved, especially in the initial step of data phasing, based once again on haplotype frequencies. The main strength of Relate, instead, is provided by the preliminary SNP pruning from LD, that is calibrated on the input sample, avoiding biases from genetically distinct populations. The simultaneous analysis of the two Chioggia and Sottomarina clusters on one hand make it possible to use a sufficient number of sample to get a realistic estimation of LD profiles; on the other hand, the presence of two sub-populations avoided excessive pruning due to the extreme homogeneity. In the phase of haplotype clustering, then, IBD regions shorter than 100 SNPs, more likely to be false positive calls, were automatically excluded from the analysis. Eventual false negative occurrences were then adjusted in the consensus process, when very similar clusters of individual were merged; clusters in the same position, but differing only of one of two individuals, in fact, were likely the result of a missing call for an IBD region between a pair, that was interrupting the cluster building. The random order of extension, repeated multiple times, granted the production of differentially discontinued clusters, highlighting these situations and allowing their resolution.

5.3 Population genetics studies using IBD data

The efficacy of the probabilistic method employed by Relate was revealed by the correct estimation of genetic similarities among pairs of individuals, according to the degree of kinship. The software could thus uncover IBD segments even in the presence of a net of biological relationships. In unrelated pairs, a particularly high relatedness emerged, providing a genetic evidence of the population isolation and inbreeding. Before these findings, these features has been reasonably assumed on the basis of historical data. The town, in fact, was completely destroyed in 1380, during the war counterposing Venice and Genoa. In the following centuries, the population was repeatedly decimated by pestilences and famines and only in the most recent times an exponential growth occurred, accompanied, however, by reduced migration fluxes. Therefore, both genetic and historical information converge in stating that the ancient community in Chioggia has some characteristics of a population isolate. As a further indication, three surnames account for about 50% of inhabitants (Gessoni et al., 2010). Chioggia was also recognized as the town with the highest rate of isonymy in Italy, again stressing the low migration and the endogamy that characterize this area (Barrai et al., 1999).

A further inspection of familial connections within the sample revealed the presence of two main sub-populations, corresponding to two close geographical areas: Chioggia and Sottomarina. Sottomarina is politically a neighborhood of Chioggia, the major town, which is located on a different island 1.5 km apart, connected through a bridge. Despite the proximity, however, the two populations have been maintained historically separated for cultural reasons: Chioggia inhabitants were predominantly

fishermen, while Sottomarina was devoted to agriculture. Although these areas are today popular seaside resorts, ancient residents may still reflect the marked partition. Noteworthy, this sample stratification didn't appear after the construction of the IBD map. The results are not in conflict with each other, as the two approaches highlight different aspects of genome sharing. In the population analysis any type of pairwise IBD region was considered, while haplotype clustering requires that the same region had to be detected in at least four individuals. A relevant implication from these results is therefore that the overall IBD sharing reflects the genetic differences between the two subgroups, while the IBD map highlights the mostly shared haplotypes in patients and these latter are common in part to Chioggia and Sottomarina. This admixture would coherently explain the high prevalence of SCZ/BPD in both the areas, due to a common genetic background. Otherwise, environmental factors should be considered as major determinants, but this would be in contrast both with the known heritability of the disorders and with the familial aggregation observed in the sample.
For the enunciated reasons, beside the outcomes from the cluster analysis, Chioggia and Sottomarina subgroups were examined together in the search of rare risk alleles.

5.4 Filtering of whole-exome sequencing data with IBD map

IBD map was created with the intent to realize a valuable instrument to understand the pattern of sharing and segregation of any genomic locus in the analyzed population. Thanks to the haplotype clustering approach, in fact, far more informative data were obtained compared to the mere IBD output, that provided only pairwise knowledge. As already cited, the map offered the advantage to be free from segregation constrains and offered the possibility to track loci across families. The integration with exome-sequencing data, then, permitted first to strongly prioritize variants and, second, to investigate the entire population from the data of 17 patients, inferring the occurrence of variants in the non-sequenced samples.

Both IBD mapping and exome data processing were realized with in-house pipelines, developed to accomplish specific tasks and overcome some issues emerged with standard variant filtering. In particular, the annotation of variant frequency information was especially curated. The correct identification of novel variants was ensured by the match not only of position but also of the reported alleles. Precise filtering was then obtained by checking which of the reported alleles for a polymorphism was actually the less frequent, resolving inaccuracies of reference human genome. A higher reliability of annotation was achieved by the use of multiple databases (e.g. 1000Genomes, ExAc), collecting large amounts of data from different sources. The immediateness of the approach was based on the creation of a variant sharing table, allowing the simultaneous analysis of all the 17 exomes and the rapid prioritization according to IBD haplotypes. On this purpose, it's important to note that the mapping of variants into IBD haplotypes was not simply a positional localization, but it

accounted also of the sharing of variants. For example, if an haplotype cluster in a defined locus was including two patients among the sequenced ones, a variant in that locus must have been shared by the same two patients, and only them, to be included in the haplotype. Therefore, the filtering according to IBD map was extremely efficient, even considering the relative high number of shared haplotypes. Indeed, from the total 30,217 rare or novel variants, 6,621 (22%) mapped in the total IBD haplotypes (IBD_{tot}), and only 558 (2%) in the IBD_{sel} ones. The latter class of IBD haplotypes was selected because shared by more than two families and by at least half of the patients within each pedigree, thus, according to the original hypothesis, they could carry susceptibility alleles with strong effect on disease liability.

5.5 Functional enrichment analyses

An overall inspection of rare and novel variants mapping in IBD haplotypes revealed that two main biological processes were affected: axon guidance and synaptic transmission. Interestingly, both of these functions are related to neurosystem, thus potentially relevant for psychiatric disorders.

In this and in the following steps, no *a priori* selection based on variant effect was performed, for a series or reasons here described. First, the prioritization according to IBD haplotypes had revealed that no specific classes of variants were particularly more shared in patients. Further, functional enrichment analyses showed that the preferential distribution of variants in axon guidance and synaptic transmission processes was not driven by LoF coding effects, indirectly implicating synonymous or non-coding alleles. The involvement of regulative sequences in SCZ and BPD have been repeatedly suggested (Ament et al., 2015; Ripke et al., 2014). This hypothesis would be also coherent with the complex nature of the disorders, accounting for the environmental contribution and the polygenic inheritance, that likely exclude extremely damaging protein alterations. Unfortunately, current knowledge of regulative sequences is still incomplete and, consequently, prediction tools for changes other than non-synonymous substitutions may be inaccurate. Therefore, for the individual examination of variants, functional annotation of genes was retained more important than single variant effect, together with the fact that each of the variant was extremely rare (MAF<1%) or novel.

The pathway analysis, however, highlighted that functional annotations, described in the mostly used repositories like KEGG and Reactome, include a variety of categories, assembled with different criteria. These comprehend for example large biological processes, like synaptic transmission, or generic signaling cascades, active in several tissues and intervening in multiple cellular functions. Another limitation of this approach is that a consistent fraction of genes is not mapped in any category; as a proof, an annotation was generally available for only a half of the tested genes.

For all these reasons, enrichment was tested for 5 previously established gene sets, assembled after reliable proteomic data, such as the PSD-95 group, or more specific annotation categories, such as

glutamatergic synapse. The sets were selected on the basis of previously reported assumptions of implication, deriving both from GWASs and exome studies (Fromer et al., 2014; G Kirov et al., 2012; Purcell et al., 2014). Four of these were found significantly enriched in genes affected by rare or novel variants, mapping both in any of the IBD haplotypes (IBD_{tot}) or in those most shared by patients (IBD_{sel}). The sets include the broad glutamatergic synapse and three categories involved in glutamatergic postsynaptic functions, particularly the ones ascribable to signalling mediated by NMDA receptors (NMDA receptor pathway, PSD-95 and ARC complexes). These evidence indicated the advantages of the gene-set approach, allowing the test of more narrow functional hypotheses. The lower number of analyzed categories, in addition, reduced the effect of multiple testing correction, increasing the possibility to detect significant outcomes. Finally, the precise definition of gene sets partially overcomes the limitation deriving from the incomplete annotations of functional categories. Thanks to these benefits, it was possible to demonstrate a significant enrichment in the same converging functions also for the IBD_{sel} set of variants, as was initially suggested by the only nominal significance in the general analysis performed with DAVID. The fourth gene set, composed by voltage-gated calcium channels (VGCCs), however, still reached significance only in IBD_{tot} genes.
In the light of these results, also the axon guidance pathway was further inspected, to reveal that again IBD_{tot} and IBD_{sel} variants were not uniformly distributed, but were clustering in genes involved in NCAM signalling for neurite outgrowth, thus in the formation of axon projections.

Noticeably, some of the major pathogenetic hypothesis for SCZ and BPD revolve around the glutamatergic synapse and the formation of neuron projections. SCZ and BPD are considered in general neurodevelopmental disorders, caused by defects in brain pre- and post-natal maturation (Rapoport et al., 2012). More in detail, alterations in dendritic spines have been multiply observed in post-mortem brains from SCZ patients, suggesting an impairment in synaptic plasticity (Moyer et al., 2015). NMDA glutamate receptors are known to be primarily involved in this phenomenon, triggering signalling cascades leading both to synapse potentiation and dendritic remodelling (Hall et al., 2015). In this context, PSD-95 complex is a major regulator of NMDA activity, while ARC proteins are downstream effectors modulating cytoskeleton rearrangements for projection growth. The signalling cascades rely upon a Ca^{2+} influx induced by glutamate, trough both NMDA receptors and VGCCs. The current model for SCZ posits a hypofunction of NMDA receptors, leading to a reduced plasticity, possibly driving the positive symptoms and the cognitive deficits seen in psychiatric cases. For BPD, instead, it has been evidenced that agonists of NMDA receptors improve depressive states. As discussed in paragraph 1.4, the NMDA-based model is also substantiated by several lines of genetic evidence, that have repeatedly implicated postsynaptic processes of glutamatergic transmission (Fromer et al., 2014; G Kirov et al., 2012; Purcell et al., 2014; Ripke et al., 2014). Further studies on structural brain abnormalities have however indicated that, beyond NMDA-mediated mechanisms, neuron projection development in general may be undermined. Rat models have shown for example

that enlarged ventricles and reduced brain volume, some of the most consistently found defects in SCZ, are the consequence of impairments in neurite outgrowth (Q. Zhang, Yu, & Huang, 2016). Additionally, functional Magnetic Resonance Imaging (fMRI) has highlighted a reduced connectivity between hippocampus and pre-frontal cortex in patients (Schmitt et al., 2011); similar observations have been made, although to a more limited extent, in BPD cases (Wessa et al., 2014). The findings presented in this work are consistent with what reported, sustaining a role for neurite formation in SCZ/BPD etiology, involving in particular *NCAM* signalling in axon guidance and NMDA postsynaptic functions for spine plasticity. These hypotheses have been originally advanced by studies on common risk factors (Ripke et al., 2014), then supported by investigations of rare sequencing variants (Fromer et al., 2014; Purcell et al., 2014). Therefore, the here described WES data further substantiate a broad convergence between common and rare variation at a functional level.

5.6 Detailed investigation of variants in families

Functional enrichment analyses are useful to understand which biological processes may be selectively affected by sets of variants. However, these strategies offer only global perspectives and ignore the actual occurrence of variants within the patients. To address this issue, the sharing of functionally relevant variants was inferred in the entire population, starting from the data of the 17 exomes and exploiting the IBD map.

An initial family-based investigation revealed a load of functionally related variants in pedigrees, suggesting a polygenic contribution to the disorders. This polygenic load was further in depth examined by the investigation of sharing and segregation of variants within and across families and some good candidate genes for SCZ/BPD emerged.

CACNA1E encodes for an α_1 subunit of VGCCs, whose role in pathogenesis has been repeatedly proposed (Purcell et al., 2014; Ripke et al., 2014), as already described. α_1 subunits constitute the transmembrane pore, thus determine the conductance properties of these calcium channels; α_1E characterizes $Ca_V2.3$ channels, activated by strong membrane depolarization in neurons and involved in neurotransmission. In particular, they have been shown to mediate presynaptic glutamate release (Heyes et al., 2015). Dysfunctions in $Ca_V2.3$ channels could thus cause alterations of the calcium current in the glutamatergic synapse, affecting both glutamate presence in the synaptic cleft and the Ca^{2+}-triggered synaptic plasticity. In family 111, a synonymous *CACNA1E* variant was found together with an intronic *FNBP1L* (formin binding protein 1-like) one; the latter gene plays a major role in signalling cascades leading to actin cytoskeleton polymerization, necessary also for neuron projection and, again, plasticity. Interestingly, both variants are predicted to potentially cause alterations in splicing process. On the same line, *FNBP1L* allele perfectly co-segregates with SCZ in family 15, together with a variant in the *DLG4* (disc large homolg 4 protein) gene, also known as PSD-95 (post synaptic density protein 95). DLG4 is a scaffolding protein of the posynaptic density, that is part of the

PSD-95 complex. It directly interacts with NMDA receptors, regulating their assembly and modulating downstream signal transduction. More in details, DLG4 binding to NMDA subunits has been shown to suppress dendritic branching, leading to the reduced plasticity typical of mature synapses (Bustos et al., 2014). The *DLG4* gene has been indirectly associated with SCZ in gene-set studies on PSD-95 complex (Fromer et al., 2014; Purcell et al., 2014). Even more, in the present work, two different variants, the first (described for family 111 and 15), synonymous and the latter in 5' UTR region, were detected in this gene, mapping in different IBD haplotypes, shared by distinct group of families. Both of them again were predicted to affect splicing. This evidence further support the involvement of DLG4 in conferring risk for SCZ/BPD, in concomitance with other genes involved in plasticity, such as *FNBP1L*.

Still in the context of glutamatergic hypothesis, *GRM7* and *GRM8* are type III metabotropic glutamate receptors, which regulate glutamate secretion and neuron excitability (Li et al., 2015). Both genes have been significantly associated with SCZ (Jajodia et al., 2015; Li et al., 2015; Ohtsuki et al., 2008) and BPD (Kandaswamy et al, 2014) in candidate-based studies. Additionally, a recent work in Han Chinese cohort not only did confirm the association, but proposed an interactive effect of variants in these two genes in conferring risk to SCZ (Li et al., 2015). Similarly, two synonymous rare alleles with damaging effects in *GRM7* and *GRM8* were found co-segregating in family 62. Interestingly, mouse models with *GRM7* deficiency show behavioural phenotypes, ascribable to increased anxiety and depression, suggesting an implication of type III metabotropic glutamate receptors in psychiatric disorders (Li et al., 2015).

Two more genes are especially of interest as found mutated in the same family and expressly involved in axon guidance: *NCAM1* (neural cell adhesion molecule-1) and *NCAN* (neurocan). *NCAM1* variant is an intronic insertion that likely determines splice site changes; *NCAN* instead carries a LoF missense substitution. As mentioned before, NCAM1 is a key factor intervening in neurite outgrowth; as a transmembrane protein, it acts both as a cell adhesion and signalling molecule. The binding of neurocan, an extracellular matrix glycoprotein, inhibits adhesion and neurite formation (Retzler, Göhring, & Rauch, 1996). Remarkably, *NCAN* maps in one of the 108 genome-wide significant loci for SCZ and has been consistently associated also with BPD (Cichon et al., 2011; Ripke et al., 2014). Conversely, genetic lines of evidence have been so far inconclusive for *NCAM1* association with SCZ/BPD; however, the here presented data suggest that the simultaneous occurrence of variants in *NCAN* and *NCAM1* could increase risk for the disorders.

Finally, the *ARHGAP32* (Rho GTPase activating protein 32) and *CDH13* (cadherin 13) genes emerged because carrying variants whose combination was detected in 5 patients across two different families. Again, these gene have related functions in neuronal development and the neurite formation. *ARHGAP32* directly interacts with NMDA receptors, mediating actin polymerization for dendritic remodelling (Wayman et al., 2008). It has been demonstrated to promote axon growth during development (Kannan, Lee, Schwedhelm-Domeyer, Nakazawa, & Stegmüller, 2012). Similarly,

cadherin 13 is a cell adhesion molecule largely expressed in the brain and a negative regulator of neurite outgrowth, promoting synapse formation. Interestingly, *CHD13* has been consistently associated with attention deficit- hyperactivity disorder (ADHD), another neuropsychiatric disorder that is thought to share part of etiology with SCZ and BPD (Rivero et al., 2013).
As a last remark, other possible combinations analogous to *ARHGAP32*/*CDH13* were identified, although they displayed a more limited sharing. One of these involved the *SEMA3C* (semaphoring 3C) and *CNTN2* (contactin 2), two genes of the axon guidance pathway. Another combination was found between the previously cited VGCC *CACNA1E* and a Ca^{2+} dependent kinase, *PRKCB*. These results could point to a possible interplay of alleles substantially increasing risk for SCZ/BPD.

The overall scenario for the genetic architecture of SCZ and BPD, emerging from this study, is a complex puzzle of loci mapping on several chromosomes, each with a proper sharing pattern across the families. Some of these looked more broadly shared, encompassing multiple pedigrees, some other tended to be more private. This could be the case for variants in VGCCs other than *CACNA1E*, that usually were restricted to a single family; the observation is further supported by the outcome of the gene-set analysis, where the VGCC category was no more significant for IBD_{sel} genes, selected because largely shared by patients.

5.7 Strengths and limits of the study

Exploiting the possibility to check the sharing and the segregation of variants in pedigrees, the reported study represents the first attempt to implicate specific alleles with SCZ and BPD. The work thus offers an example of the potential of family-based samples in the search of rare variants, in contraposition with case-control cohorts, that have so far restricted the investigations at pathway or gene-set level. Further, the homogeneity of the population sample was favourable for the identification of rare variants: alleles with a worldwide frequency <1% were instead multiply detected in the sample, increasing the possibility to track risk factors. Even in this simplified complexity, however, the polygenic nature of the disorders emerged, indicating that classical approaches developed for Mendelian traits are more likely to be unsuccessful. Beside the inter-familial heterogeneity, in fact, also intra-familial genetic diversity was observed, as initially hypothesized. Candidate risk variants were actually sometimes not common to all the patients of a family (e.g. *FNBP1L* in family 111), or showing an incomplete penetrance (e.g. *CACNA1E* in family 9). In this context, IBD mapping proved to be effective in tracking shared loci not only across different pedigrees but also within single pedigrees, where linkage approach has a reduced power given the small family extension.
The sample features could also uncover some combinations of functionally related alleles that may consistently contribute to risk for SCZ/BPD. Interestingly, the analyzed population controls seemed

not only to display a lower load of such alleles, but also to carry none of the identified combinations. These observation can't be however generalized or retained conclusive, considering the small number of control subjects available.

Nevertheless, these outcomes looked particularly interesting, since they could be extremely useful for the identification of specific mechanisms for SCZ/BPD vulnerability. More precise pathogenetic hypotheses could be in fact more easily addressed also in model organisms for SCZ/BPD, such as rodents. If, on one hand, animal models can be extremely valuable for the study of pathogenesis, on the other hand phenotype evaluation in psychiatric disorders has always been a debated issue. Symptoms are indeed mostly behavioural, thus difficult to assess, and only molecular or biophysical alterations could be objectively investigated (Logan & McClung, 2015). In addition, the mutations that are induced in these models are in general extremely damaging, such as a complete knockout (O'Tuathaigh & Waddington, 2015). These type of variants are not coherent with what is currently known about the genetic architecture of SCZ and BPD, specifically with their polygenic nature. On the other hand, the simple presence of a single candidate variant would likely have no effect, for the same reasons. A major improvement could thus be achieved by the simultaneous test of multiple variants, especially those combinations that are more likely to substantially increase risk.

Despite the cited advantages, this work presents some limitations. First, the search of risk factors was focused only on functionally relevant genes, as determined by the initial enrichment analysis. Although the investigated processes were not *a priori* determined, but indicted by the enrichment analysis itself, some biases could have been introduced as a consequence of incomplete annotations, as previously discussed. To improve this preliminary investigations, a more refined approach is currently under evaluation. This method, developed by professor Bortoluzzi's team, implies the use of meta-networks, derived from annotated pathways, to unveil additional connections among genes (unpublished data). An example of a possible outcome is displayed in figure 5.1. The shown meta-network was obtained by the joint analysis of IBD_{sel} and LoF IBD_{tot} variants and highlighted a total of 87 genes, connected by relationships previously hidden. This would thus provide an improved functional prioritization of variants.

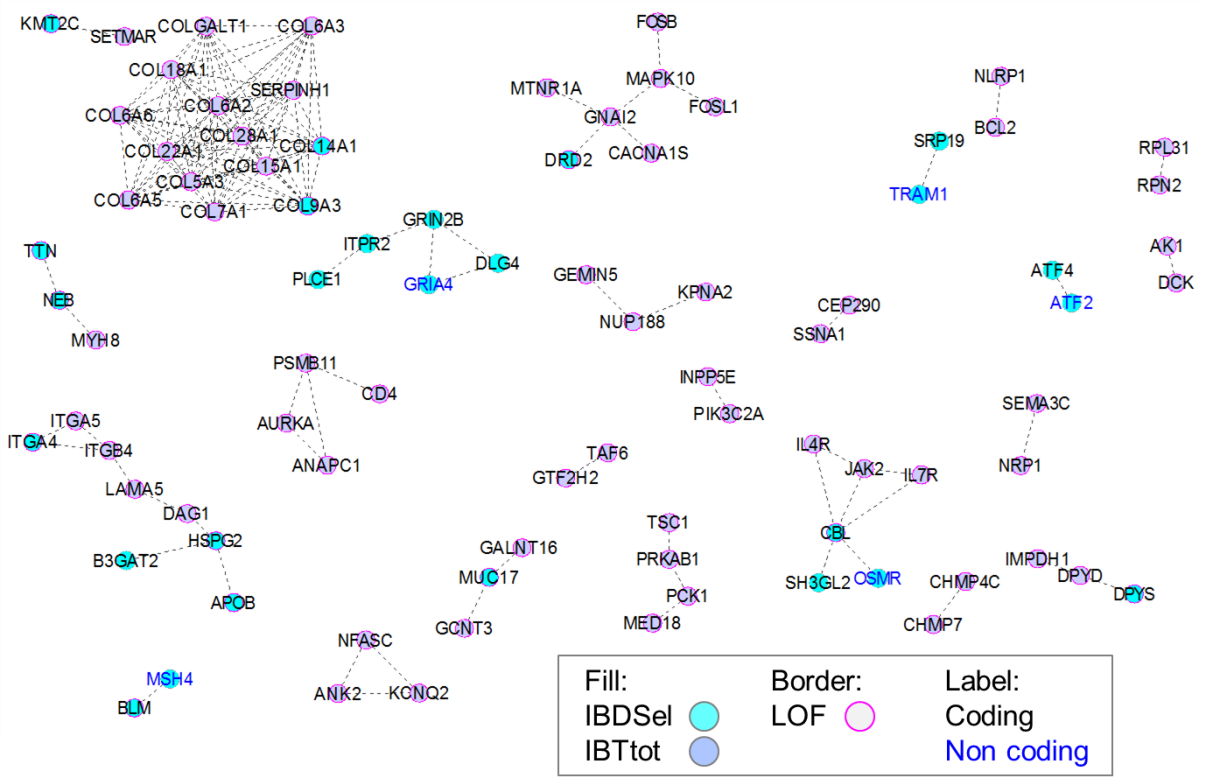

Figure 5.1: Pathway-derived network from genes affected by IBDsel and/or LoF variants. The results unveil connections between genes that were not evident in simply annotated pathways.

A second limitation of the study is related to the main focus on coding sequences, a reflection of WES technique. Several lines of evidence have in fact sustained an important role for con-coding variation in SCZ and BPD. For example, in only 10 of the 108 loci associated with SCZ, the signal could be credibly attributable to an exonic, non-synonymous allele (Ripke et al., 2014). On the opposite, the same 108 loci were enriched in enhancers active in brain tissues. Similarly, in an exome study of bipolar families, 88% of variants within candidate pathways were located in non-coding regions (Ament et al., 2015). Also in the here presented work, about 50% of the prioritized variants, both in IBD_{sel} and IBD_{tot} categories, were located outside coding exons. These type of variants are generally excluded from analyses since their interpretation is complicated. The variants detected in this analysis, however, were all extremely rare or novel and, as described above, in functionally relevant genes. In addition, the high percentage of intronic variants was unlikely due to a simple bias from frequency annotations in databases, since several resources of different type, and collected with different technologies, were considered. Moreover, false positive calls were likely removed by filtering steps, particularly the mapping into IBD haplotypes, that requires a specific sharing of an allele in the sample, improbable to randomly occur. The low prediction abilities on the impact of these variants remain still a main limitation, that needs improvements, especially if a more thorough investigation of non-coding variants is envisioned, through whole-genome sequencing (WGS). The main issues when extending the analysis to the entire genome is, in fact, the interpretation and then the prioritization of

the huge amount of obtained variants. In this regards, IBD map would offer an effective strategy, again allowing the selection of variants according to a desired sharing. An explicative example emerged with the examination of co-segregating haplotypes. Together with *CNTN2* and *SEMA3C* alleles, described above, a third locus on a different chromosome was shared with the same pattern, but no rare or novel variant was detected. Interestingly, the haplotype encompasses the *CACNB4* gene, encoding for a subunit of VGCCs. Both *CNTN2* and VGCCs interact with *NCAM1* for neurite formation; in particular, the binding with VGCCs determines a *NCAM1*-dependent Ca^{2+} influx that triggers several processes promoting the growth of active cones (ref provided by Reactome manual annotation). It would be thus intriguing to assess whether some regulative variant outside exomic targets is actually located in this region.

A final aspect that was not deeply explored is the possibility that some of the risk alleles was more specific for either SCZ or BPD. Beside the consistent overlap, in fact, genetic studies have shown that some characteristic variant must exist, driving the final phenotype towards one of the disorders (Craddock et al., 2009). Indeed, some preliminary observations identified variants shared only by SCZ patients (e.g. *ARHGAP32*). Again IBD map is a valuable tool for this purpose, since for any genomic locus is a priori possible to establish whether there is a preferential sharing in a specific class of phenotypes. Future investigations will thus provide further insights.

5.8 Conclusions

Concluding, this work provided an important contribution in understanding the genetic architecture of two SCZ and BPD. Firstly, it described a novel approach to investigate family-based samples with a common ancestry, relying on the construction of an IBD map. This map represents a flexible tool to address specific questions relative to candidate risk factors, thus allowing the testing of specific hypothesis, such as the existence of haplotype combinations or of phenotype-distinct loci. Moreover, the map can be used to prioritize variants from next-generation sequencing data, such as WES or, in the near future, WGS. Secondly, thanks to the integration of IBD and WES approaches, some insights on SCZ/BPD etiology were obtained. More in details, specific processes related to neurosystem development and functions were implicated and some candidate variants were provided, therefore substantiating the role of rare variation in susceptibility. Further insights will be necessary to establish the actual contribution of these rare variants in etiology. The whole picture is in fact far more complex, likely involving also common variants and environmental effects.

Finally, considering the unique characteristics of the studied population, it's difficult to predict whether the same alleles could be detected in independent samples. Nevertheless, in the light also of the general convergence of all types of genetic studies, the results could suggest possible genes or nodes of pathways that increase vulnerability to these psychiatric disorders. Further replications in other studies would thus support their involvement, ultimately suggesting new therapeutic targets for these common disorders that cause such a high burden for the society.

6. References

Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F. C., & Nielsen, R. (2009). Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology*, *33*(3), 266–74. http://doi.org/10.1002/gepi.20378

Ament, S. A., Szelinger, S., Glusman, G., Ashworth, J., Hou, L., Akula, N., … Roach, J. C. (2015). Rare variants in neuronal excitability genes influence risk for bipolar disorder. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(11), 3576–81. http://doi.org/10.1073/pnas.1424958112

Angst, J., & Marneros, A. (2001). Bipolarity from ancient to modern times: *Journal of Affective Disorders*, *67*(1-3), 3–19. http://doi.org/10.1016/S0165-0327(01)00429-3

Awadalla, P., Gauthier, J., Myers, R. A., Casals, F., Hamdan, F. F., Griffing, A. R., … Rouleau, G. A. (2010). Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *American Journal of Human Genetics*, *87*(3), 316–24. http://doi.org/10.1016/j.ajhg.2010.07.019

Barrai, I., Rodriguez-Larralde, A., Mamolini, E., & Scapoli, C. (1999). Isonymy and isolation by distance in Italy. *Human Biology*, *71*(6), 947–61.

Bertolin, C., Magri, C., Barlati, S., Vettori, A., Perini, G. I., Peruzzi, P., … Vazza, G. (2011). Analysis of complete mitochondrial genomes of patients with schizophrenia and bipolar disorder. *Journal of Human Genetics*, *56*(12), 869–72. http://doi.org/10.1038/jhg.2011.111

Blackwood, D. H. R., Fordyce, A., Walker, M. T., St. Clair, D. M., Porteous, D. J., & Muir, W. J. (2001). Schizophrenia and Affective Disorders—Cosegregation with a Translocation at Chromosome 1q42 That Directly Disrupts Brain-Expressed Genes: Clinical and P300 Findings in a Family. *The American Journal of Human Genetics*, *69*(2), 428–433. http://doi.org/10.1086/321969

Browning, B. L., & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, *88*(2), 173–82. http://doi.org/10.1016/j.ajhg.2011.01.010

Browning, S. R., & Browning, B. L. (2012). Identity by Descent Between Distant Relatives: Detection and Applications.

Bustos, F. J., Varela-Nallar, L., Campos, M., Henriquez, B., Phillips, M., Opazo, C., … van Zundert, B. (2014). PSD95 suppresses dendritic arbor development in mature hippocampal neurons by occluding the clustering of NR2B-NMDA receptors. *PloS One*, *9*(4), e94037. http://doi.org/10.1371/journal.pone.0094037

Cardno, A. G., & Owen, M. J. (2014). Genetic relationships between schizophrenia, bipolar disorder, and schizoaffective disorder. *Schizophrenia Bulletin*, *40*(3), 504–15. http://doi.org/10.1093/schbul/sbu016

Cardno, A. G., Rijsdijk, F. V, Sham, P. C., Murray, R. M., & McGuffin, P. (2002). A twin study of genetic relationships between psychotic symptoms. *The American Journal of Psychiatry*, *159*(4), 539–545. http://doi.org/10.1176/appi.ajp.159.4.539

Cardon, L. R., & Abecasis, G. R. (2003). Using haplotype blocks to map human complex trait loci. *Trends in Genetics : TIG*, *19*(3), 135–40. http://doi.org/10.1016/S0168-9525(03)00022-2

Chen, D. T., Jiang, X., Akula, N., Shugart, Y. Y., Wendland, J. R., Steele, C. J. M., … Strauss, J. (2013). Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Molecular Psychiatry*, *18*(2), 195–205. http://doi.org/10.1038/mp.2011.157

Chen, J., Cao, F., Liu, L., Wang, L., & Chen, X. (2015). Genetic studies of schizophrenia: an update. *Neuroscience Bulletin*, *31*(1), 87–98. http://doi.org/10.1007/s12264-014-1494-4

Cichon, S., Mühleisen, T. W., Degenhardt, F. A., Mattheisen, M., Miró, X., Strohmaier, J., … Nöthen, M. M. (2011). Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *American Journal of Human Genetics*, *88*(3), 372–81. http://doi.org/10.1016/j.ajhg.2011.01.017

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., … Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research*, *35*(6), 2013–2025. http://doi.org/10.1093/nar/gkm076

Coppieters, F., Van Schil, K., Bauwens, M., Verdin, H., De Jaegher, A., Syx, D., … De Baere, E. (2014). Identity-by-descent-guided mutation analysis and exome sequencing in consanguineous families reveals unusual clinical and molecular findings in retinal dystrophy. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *16*(9), 671–80. http://doi.org/10.1038/gim.2014.24

Craddock, N., O'Donovan, M. C., & Owen, M. J. (2009). Psychosis genetics: modeling the relationship between schizophrenia, bipolar disorder, and mixed (or "schizoaffective") psychoses. *Schizophrenia Bulletin*, *35*(3), 482–90. http://doi.org/10.1093/schbul/sbp020

Craddock, N., & Owen, M. J. (2010). The Kraepelinian dichotomy - going, going... but still not gone. *The British Journal of Psychiatry : The Journal of Mental Science*, *196*(2), 92–5. http://doi.org/10.1192/bjp.bp.109.073429

Cruceanu, C., Ambalavanan, A., Spiegelman, D., Gauthier, J., Lafrenière, R. G., Dion, P. A., … Rouleau, G. A. (2013). Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder. *Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada*, *56*(10), 634–40. http://doi.org/10.1139/gen-2013-0081

Doherty, J. L., O'Donovan, M. C., & Owen, M. J. (2012). Recent genomic advances in schizophrenia. *Clinical Genetics*, *81*(2), 103–9. http://doi.org/10.1111/j.1399-0004.2011.01773.x

Escamilla, M. A. (2001). Population isolates: their special value for locating genes for bipolar disorder. *Bipolar Disorders*, *3*(6), 299–317. http://doi.org/10.1034/j.1399-5618.2001.30605.x

Falkai, P., Rossner, M. J., Schulze, T. G., Hasan, A., Brzózka, M. M., Malchow, B., … Schmitt, A. (2015). Kraepelin revisited: schizophrenia from degeneration to failed regeneration. *Molecular Psychiatry*, *20*(6),

671–6. http://doi.org/10.1038/mp.2015.35

Farrell, M. S., Werge, T., Sklar, P., Owen, M. J., Ophoff, R. A., O'Donovan, M. C., … Sullivan, P. F. (2015). Evaluating historical candidate genes for schizophrenia. *Molecular Psychiatry*, *20*(5), 555–62. http://doi.org/10.1038/mp.2015.16

Ferreira, M. A. R., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L., … Craddock, N. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics*, *40*(9), 1056–8. http://doi.org/10.1038/ng.209

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. *Nature Reviews. Genetics*, *7*(2), 85–97. http://doi.org/10.1038/nrg1767

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., … O'Donovan, M. C. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature*, *506*(7487), 179–84. http://doi.org/10.1038/nature12929

Frye, M. A., Prieto, M. L., Bobo, W. V, Kung, S., Veldic, M., Alarcon, R. D., … Tye, S. J. (2014). Current landscape, unmet needs, and future directions for treatment of bipolar depression. *Journal of Affective Disorders*, *169 Suppl* , S17–23. http://doi.org/10.1016/S0165-0327(14)70005-9

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. (2007). *Nature*, *447*(7145), 661–78. http://doi.org/10.1038/nature05911

Georgi, B., Craig, D., Kember, R. L., Liu, W., Lindquist, I., Nasser, S., … Bućan, M. (2014). Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genetics*, *10*(3), e1004229. http://doi.org/10.1371/journal.pgen.1004229

Gessoni, G., Valverde, S., Canistro, R., & Manoni, F. (2010). Factor V Leiden in Chioggia: a prevalence study in patients with venous thrombosis, their blood relatives and the general population. *Blood Transfusion = Trasfusione Del Sangue*, *8*(3), 193–5. http://doi.org/10.2450/2010.0157-09

Green, E. K., Grozeva, D., Forty, L., Gordon-Smith, K., Russell, E., Farmer, A., … Craddock, N. (2013). Association at SYNE1 in both bipolar disorder and recurrent major depression. *Molecular Psychiatry*, *18*(5), 614–7. http://doi.org/10.1038/mp.2012.48

Green, E. K., Hamshere, M., Forty, L., Gordon-Smith, K., Fraser, C., Russell, E., … Craddock, N. (2013). Replication of bipolar disorder susceptibility alleles and identification of two novel genome-wide significant associations in a new bipolar disorder case-control sample. *Molecular Psychiatry*, *18*(12), 1302–7. http://doi.org/10.1038/mp.2012.142

Green, E. K., Rees, E., Walters, J. T. R., Smith, K.-G., Forty, L., Grozeva, D., … Kirov, G. (2016). Copy number variation in bipolar disorder. *Molecular Psychiatry*, *21*(1), 89–93. http://doi.org/10.1038/mp.2014.174

Group, P. G. C. B. D. W. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*, *43*(10), 977–83. http://doi.org/10.1038/ng.943

Grozeva, D., Kirov, G., Conrad, D. F., Barnes, C. P., Hurles, M., Owen, M. J., … Craddock, N. (2013). Reduced burden of very large and rare CNVs in bipolar affective disorder. *Bipolar Disorders*, *15*(8), 893–8. http://doi.org/10.1111/bdi.12125

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., … Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, *19*(2), 318–26. http://doi.org/10.1101/gr.081398.108

Hall, J., Trent, S., Thomas, K. L., O'Donovan, M. C., & Owen, M. J. (2015). Genetic risk for schizophrenia: convergence on synaptic pathways involved in plasticity. *Biological Psychiatry*, *77*(1), 52–8. http://doi.org/10.1016/j.biopsych.2014.07.011

Harrison, P. J. (2015). Recent genetic findings in schizophrenia and their therapeutic relevance. *Journal of Psychopharmacology (Oxford, England)*, *29*(2), 85–96. http://doi.org/10.1177/0269881114553647

Hashimoto, K., Malchow, B., Falkai, P., & Schmitt, A. (2013). Glutamate modulators as potential therapeutic drugs in schizophrenia and affective disorders. *European Archives of Psychiatry and Clinical Neuroscience*, *263*(5), 367–77. http://doi.org/10.1007/s00406-013-0399-y

Heyes, S., Pratt, W. S., Rees, E., Dahimene, S., Ferron, L., Owen, M. J., & Dolphin, A. C. (2015). Genetic disruption of voltage-gated calcium channels in psychiatric and neurological disorders. *Progress in Neurobiology*, *134*, 36–54. http://doi.org/10.1016/j.pneurobio.2015.09.002

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37*(1), 1–13. http://doi.org/10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. http://doi.org/10.1038/nprot.2008.211

Jajodia, A., Kaur, H., Kumari, K., Gupta, M., Baghel, R., Srivastava, A., … Kukreti, R. (2015). Evidence for schizophrenia susceptibility alleles in the Indian population: An association of neurodevelopmental genes in case-control and familial samples. *Schizophrenia Research*, *162*(1-3), 112–7. http://doi.org/10.1016/j.schres.2014.12.031

Kandaswamy, R., McQuillin, A., Curtis, D., & Gurling, H. (2014). Allelic association, DNA resequencing and copy number variation at the metabotropic glutamate receptor GRM7 gene locus in bipolar disorder. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics : The Official Publication of the International Society of Psychiatric Genetics*, *165B*(4), 365–72. http://doi.org/10.1002/ajmg.b.32239

Kannan, M., Lee, S.-J., Schwedhelm-Domeyer, N., Nakazawa, T., & Stegmüller, J. (2012). p250GAP is a novel player in the Cdh1-APC/Smurf1 pathway of axon growth regulation. *PloS One*, *7*(11), e50735. http://doi.org/10.1371/journal.pone.0050735

Kerner, B. (2015). Toward a Deeper Understanding of the Genetics of Bipolar Disorder. *Frontiers in Psychiatry*,

6, 105. http://doi.org/10.3389/fpsyt.2015.00105

Kerner, B., Rao, A. R., Christensen, B., Dandekar, S., Yourshaw, M., & Nelson, S. F. (2013). Rare Genomic Variants Link Bipolar Disorder with Anxiety Disorders to CREB-Regulated Intracellular Signaling Pathways. *Frontiers in Psychiatry*, *4*, 154. http://doi.org/10.3389/fpsyt.2013.00154

Kieseppä, T., Partonen, T., Haukka, J., Kaprio, J., & Lönnqvist, J. (2014). High Concordance of Bipolar I Disorder in a Nationwide Sample of Twins. *American Journal of Psychiatry*.

Kirov, G. (2015). CNVs in neuropsychiatric disorders. *Human Molecular Genetics*, *24*(R1), R45–9. http://doi.org/10.1093/hmg/ddv253

Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., … Owen, M. J. (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular Psychiatry*, *17*(2), 142–53. http://doi.org/10.1038/mp.2011.154

Kotlar, A. V, Mercer, K. B., Zwick, M. E., & Mulle, J. G. (2015). New discoveries in schizophrenia genetics reveal neurobiological pathways: a review of recent findings. *European Journal of Medical Genetics*, *58*(12), 704–714. http://doi.org/10.1016/j.ejmg.2015.10.008

Kunkle, B. W., Jaworski, J., Barral, S., Vardarajan, B., Beecham, G. W., Martin, E. R., … Pericak-Vance, M. A. (2016). Genome-wide linkage analyses of non-Hispanic white families identify novel loci for familial late-onset Alzheimer's disease. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *12*(1), 2–10. http://doi.org/10.1016/j.jalz.2015.05.020

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Sullivan, P. F., Goddard, M. E., … Wray, N. R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, *44*(3), 247–50. http://doi.org/10.1038/ng.1108

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V, Purcell, S. M., Perlis, R. H., … Wray, N. R. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, *45*(9), 984–94. http://doi.org/10.1038/ng.2711

Lewis, C. M., Levinson, D. F., Wise, L. H., DeLisi, L. E., Straub, R. E., Hovatta, I., … Helgason, T. (2003). Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *American Journal of Human Genetics*, *73*(1), 34–48. http://doi.org/10.1086/376549

Li, W., Ju, K., Li, Z., He, K., Chen, J., Wang, Q., … Shi, Y. (2015). Significant association of GRM7 and GRM8 genes with schizophrenia and major depressive disorder in the Han Chinese population. *European Neuropsychopharmacology : The Journal of the European College of Neuropsychopharmacology*. http://doi.org/10.1016/j.euroneuro.2015.05.004

Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., & Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*, *373*(9659), 234–9. http://doi.org/10.1016/S0140-6736(09)60072-6

Lin, P.-I., & Mitchell, B. D. (2008). Approaches for unraveling the joint genetic determinants of schizophrenia and bipolar disorder. *Schizophrenia Bulletin*, *34*(4), 791–7. http://doi.org/10.1093/schbul/sbn050

Logan, R. W., & McClung, C. A. (2015). Animal models of bipolar mania: The past, present and future. *Neuroscience*. http://doi.org/10.1016/j.neuroscience.2015.08.041

Lorenzon, A., Pilichou, K., Rigato, I., Vazza, G., De Bortoli, M., Calore, M., … Rampazzo, A. (2015). Homozygous Desmocollin-2 Mutations and Arrhythmogenic Cardiomyopathy. *The American Journal of Cardiology*, *116*(8), 1245–51. http://doi.org/10.1016/j.amjcard.2015.07.037

Malhotra, D., McCarthy, S., Michaelson, J. J., Vacic, V., Burdick, K. E., Yoon, S., … Sebat, J. (2011). High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron*, *72*(6), 951–63. http://doi.org/10.1016/j.neuron.2011.11.007

Malhotra, D., & Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, *148*(6), 1223–41. http://doi.org/10.1016/j.cell.2012.02.039

Marchani, E. E., & Wijsman, E. M. (2011). Estimation and visualization of identity-by-descent within pedigrees simplifies interpretation of complex trait analysis. *Human Heredity*, *72*(4), 289–97. http://doi.org/10.1159/000334083

McCarthy, S. E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., … Corvin, A. (2014). De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular Psychiatry*, *19*(6), 652–8. http://doi.org/10.1038/mp.2014.29

Mir, A., Sritharan, K., Mittal, K., Vasli, N., Araujo, C., Jamil, T., … Vincent, J. B. (2014). Truncation of the E3 ubiquitin ligase component FBXO31 causes non-syndromic autosomal recessive intellectual disability in a Pakistani family. *Human Genetics*, *133*(8), 975–84. http://doi.org/10.1007/s00439-014-1438-0

Möller, H.-J. (2003). Bipolar Disorder and Schizophrenia: Distinct Illnesses or a Continuum? *The Journal of Clinical Psychiatry*, *64*(suppl 6), 1,478–27.

Mowry, B. J., & Gratten, J. (2013). The emerging spectrum of allelic variation in schizophrenia: current evidence and strategies for the identification and functional characterization of common and rare variants. *Molecular Psychiatry*, *18*(1), 38–52. http://doi.org/10.1038/mp.2012.34

Moyer, C. E., Shelton, M. A., & Sweet, R. A. (2015). Dendritic spine alterations in schizophrenia. *Neuroscience Letters*, *601*, 46–53. http://doi.org/10.1016/j.neulet.2014.11.042

Mühleisen, T. W., Leber, M., Schulze, T. G., Strohmaier, J., Degenhardt, F., Treutlein, J., … Cichon, S. (2014). Genome-wide association study reveals two new risk loci for bipolar disorder. *Nature Communications*, *5*, 3339. http://doi.org/10.1038/ncomms4339

Murphy, K. C., Jones, L. A., & Owen, M. J. (1999). High Rates of Schizophrenia in Adults With Velo-Cardio-Facial Syndrome. *Archives of General Psychiatry*, *56*(10), 940. http://doi.org/10.1001/archpsyc.56.10.940

Neale, B. M., & Sklar, P. (2015). Genetic analysis of schizophrenia and bipolar disorder reveals polygenicity but

also suggests new directions for molecular interrogation. *Current Opinion in Neurobiology*, *30*, 131–8. http://doi.org/10.1016/j.conb.2014.12.001

Need, A. C., & Goldstein, D. B. (2014). Schizophrenia genetics comes of age. *Neuron*, *83*(4), 760–3. http://doi.org/10.1016/j.neuron.2014.08.015

O'Donovan, M. C., Craddock, N., Norton, N., Williams, H., Peirce, T., Moskvina, V., … Cloninger, C. R. (2008). Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nature Genetics*, *40*(9), 1053–5. http://doi.org/10.1038/ng.201

O'Tuathaigh, C. M., & Waddington, J. L. (2015). Closing the translational gap between mutant mouse models and the clinical reality of psychotic illness. *Neuroscience and Biobehavioral Reviews*, *58*, 19–35. http://doi.org/10.1016/j.neubiorev.2015.01.016

Ohtsuki, T., Koga, M., Ishiguro, H., Horiuchi, Y., Arai, M., Niizato, K., … Arinami, T. (2008). A polymorphism of the metabotropic glutamate receptor mGluR7 (GRM7) gene is associated with schizophrenia. *Schizophrenia Research*, *101*(1-3), 9–16. http://doi.org/10.1016/j.schres.2008.01.027

Peltonen, L., Palotie, A., & Lange, K. (2000). Use of population isolates for mapping complex traits. *Nature Reviews. Genetics*, *1*(3), 182–90. http://doi.org/10.1038/35042049

PGC. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, *43*(10), 969–76. http://doi.org/10.1038/ng.940

Platzer, M. (2006). The human genome and its upcoming dynamics. *Genome Dynamics*, *2*, 1–16. http://doi.org/10.1159/000095083

Power, R. A., Kyaga, S., Uher, R., MacCabe, J. H., Långström, N., Landen, M., … Svensson, A. C. (2013). Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry*, *70*(1), 22–30. http://doi.org/10.1001/jamapsychiatry.2013.268

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., … Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, *506*(7487), 185–90. http://doi.org/10.1038/nature12975

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., & Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*(7256), 748–752. http://doi.org/10.1038/nature08185

Rapoport, J. L., Giedd, J. N., & Gogtay, N. (2012). Neurodevelopmental model of schizophrenia: update 2012. *Molecular Psychiatry*, *17*(12), 1228–38. http://doi.org/10.1038/mp.2012.23

Rare chromosomal deletions and duplications increase risk of schizophrenia. (2008). *Nature*, *455*(7210), 237–41. http://doi.org/10.1038/nature07239

Rees, E., Walters, J. T. R., Georgieva, L., Isles, A. R., Chambert, K. D., Richards, A. L., … Kirov, G. (2014).

Analysis of copy number variations at 15 schizophrenia-associated loci. *The British Journal of Psychiatry : The Journal of Mental Science*, *204*(2), 108–14. http://doi.org/10.1192/bjp.bp.113.131052

Retzler, C., Göhring, W., & Rauch, U. (1996). Analysis of neurocan structures interacting with the neural cell adhesion molecule N-CAM. *The Journal of Biological Chemistry*, *271*(44), 27304–10.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K.-H., Holmans, P. a., … O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421–427. http://doi.org/10.1038/nature13595

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., … Sullivan, P. F. (2013a). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*(10), 1150–9. http://doi.org/10.1038/ng.2742

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., … Sullivan, P. F. (2013b). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*, *45*(10), 1150–9. http://doi.org/10.1038/ng.2742

Rivero, O., Sich, S., Popp, S., Schmitt, A., Franke, B., & Lesch, K.-P. (2013). Impact of the ADHD-susceptibility gene CDH13 on development and function of brain networks. *European Neuropsychopharmacology : The Journal of the European College of Neuropsychopharmacology*, *23*(6), 492–507. http://doi.org/10.1016/j.euroneuro.2012.06.009

Scharpf, R. B., Parmigiani, G., Pevsner, J., & Ruczinski, I. (2008). Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *The Annals of Applied Statistics*, *2*(2), 687–713. http://doi.org/10.1214/07-AOAS155

Schmitt, A., Hasan, A., Gruber, O., & Falkai, P. (2011). Schizophrenia as a disorder of disconnectivity. *European Archives of Psychiatry and Clinical Neuroscience*, *261 Suppl* , S150–4. http://doi.org/10.1007/s00406-011-0242-2

Shinozaki, G., & Potash, J. B. (2014). New developments in the genetics of bipolar disorder. *Current Psychiatry Reports*, *16*(11), 493. http://doi.org/10.1007/s11920-014-0493-5

Simonsen, C., Sundet, K., Vaskinn, A., Birkenaes, A. B., Engh, J. A., Faerden, A., … Andreassen, O. A. (2011). Neurocognitive dysfunction in bipolar and schizophrenia spectrum disorders depends on history of psychosis rather than diagnostic group. *Schizophrenia Bulletin*, *37*(1), 73–83. http://doi.org/10.1093/schbul/sbp034

Smoller, J. W., & Finn, C. T. (2003). Family, twin, and adoption studies of bipolar disorder. *American Journal of Medical Genetics. Part C, Seminars in Medical Genetics*, *123C*(1), 48–58. http://doi.org/10.1007/s11920-002-0046-1

Song, J., Bergen, S. E., Kuja-Halkola, R., Larsson, H., Landén, M., & Lichtenstein, P. (2015). Bipolar disorder and its relation to major psychiatric disorders: a family-based study in the Swedish population. *Bipolar Disorders*, *17*(2), 184–93. http://doi.org/10.1111/bdi.12242

Sotgiu, S., Pugliatti, M., Fois, M. L., Arru, G., Sanna, A., Sotgiu, M. A., & Rosati, G. (2004). Genes, environment, and susceptibility to multiple sclerosis. *Neurobiology of Disease*, *17*(2), 131–43. http://doi.org/10.1016/j.nbd.2004.07.015

Sullivan, P. F. (2010). The Psychiatric GWAS Consortium: Big Science Comes to Psychiatry. *Neuron*, *68*(2), 182–186. http://doi.org/10.1016/j.neuron.2010.10.003

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews. Genetics*, *13*(8), 537–51. http://doi.org/10.1038/nrg3240

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*, *60*(12), 1187–92. http://doi.org/10.1001/archpsyc.60.12.1187

Sullivan, P. F., Lin, D., Tzeng, J.-Y., van den Oord, E., Perkins, D., Stroup, T. S., … Close, S. L. (2008). Genomewide association for schizophrenia in the CATIE study: results of stage 1. *Molecular Psychiatry*, *13*(6), 570–84. http://doi.org/10.1038/mp.2008.25

Tesli, M., Espeseth, T., Bettella, F., Mattingsdal, M., Aas, M., Melle, I., … Andreassen, O. A. (2014). Polygenic risk score and the psychosis continuum model. *Acta Psychiatrica Scandinavica*, *130*(4), 311–7. http://doi.org/10.1111/acps.12307

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, *194*(2), 301–26. http://doi.org/10.1534/genetics.112.148825

Timms, A. E., Dorschner, M. O., Wechsler, J., Choi, K. Y., Kirkwood, R., Girirajan, S., … Tsuang, D. W. (2013). Support for the N-methyl-D-aspartate receptor hypofunction hypothesis of schizophrenia from exome sequencing in multiplex families. *JAMA Psychiatry*, *70*(6), 582–90. http://doi.org/10.1001/jamapsychiatry.2013.1195

van Os, J., & Kapur, S. (2009). Schizophrenia. *Lancet*, *374*(9690), 635–45. http://doi.org/10.1016/S0140-6736(09)60995-8

Vandeweyer, G., Reyniers, E., Wuyts, W., Rooms, L., & Kooy, R. F. (2011). CNV-WebStore: online CNV analysis, storage and interpretation. *BMC Bioinformatics*, *12*(1), 4. http://doi.org/10.1186/1471-2105-12-4

Varilo, T., & Peltonen, L. (2004). Isolates and their potential use in complex gene mapping efforts. *Current Opinion in Genetics & Development*, *14*(3), 316–23. http://doi.org/10.1016/j.gde.2004.04.008

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. http://doi.org/10.1016/j.ajhg.2011.11.029

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., … Sebat, J. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (New York, N.Y.)*, *320*(5875), 539–43. http://doi.org/10.1126/science.1155174

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., … Bucan, M. (2007). PennCNV: an integrated

hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17*(11), 1665–74. http://doi.org/10.1101/gr.6861907

Wayman, G. A., Davare, M., Ando, H., Fortin, D., Varlamova, O., Cheng, H.-Y. M., … Impey, S. (2008). An activity-regulated microRNA controls dendritic plasticity by down-regulating p250GAP. *Proceedings of the National Academy of Sciences*, *105*(26), 9093–9098. http://doi.org/10.1073/pnas.0803072105

Werner, F.-M., & Coveñas, R. (2015). New developments in the management of schizophrenia and bipolar disorder: potential use of cariprazine. *Therapeutics and Clinical Risk Management*, *11*, 1657–1661. http://doi.org/10.2147/TCRM.S64915

Wessa, M., Kanske, P., & Linke, J. (2014). Bipolar disorder: a neural network perspective on a disorder of emotion and motivation. *Restorative Neurology and Neuroscience*, *32*(1), 51–62. http://doi.org/10.3233/RNN-139007

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., … Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics*, *44*(12), 1365–9. http://doi.org/10.1038/ng.2446

Zhang, D., Cheng, L., Qian, Y., Alliey-Rodriguez, N., Kelsoe, J. R., Greenwood, T., … Gershon, E. S. (2009). Singleton deletions throughout the genome increase risk of bipolar disorder. *Molecular Psychiatry*, *14*(4), 376–80. http://doi.org/10.1038/mp.2008.144

Zhang, Q., Yu, Y., & Huang, X.-F. (2016). Olanzapine Prevents the PCP-induced Reduction in the Neurite Outgrowth of Prefrontal Cortical Neurons via NRG1. *Scientific Reports*, *6*, 19581. http://doi.org/10.1038/srep19581

7. Appendix

7.1 The 108 loci significantly associated with SCZ (Ripke et al., 2014)

Rank	*P*-value	Position (hg19)	SCZ†	BPD†	Protein coding genes
1	3.48e-31	chr6:28303247-28712247	Y	N	*Locus too broad*
2	3.362e-19	chr1:97792625-98559084	Y	N	*DPYD MIR137 (micro-RNA)*
3	6.198e-19	chr10:104423800-105165583	Y	N	*ARL3 AS3MT C10orf32 CNNM2 CYP17A1 INA NT5C2 PCGF6 PDCD11 SFXN2 TAF5 TRIM8 USMG5 WBP1L*
4	3.217e-18	chr12:2321860-2523731	Y	Y	*CACNA1C*
5	1.737e-15	chr8:143309503-143330533	Y	N	*TSNARE1*
6	7.98e-15	chr4:103146888-103198090	N	N	*SLC39A8*
7	8.2e-15	chr7:1896096-2190096	Y	N	*MAD1L1*
8	1.099e-14	chr5:60499143-60843543	Y	N	*ZSWIM6*
9	1.859e-14	chr12:123448113-123909113	Y	N	*ABCB9 ARL6IP4 C12orf65 CDK2AP1 MPHOSPH9 OGFOD2 PITPNM2 RILPL2 SBNO1 SETD8*
10	5.652e-14	chr2:200715237-200848037	Y	N	*AC073043.2 C2orf47 C2orf69 TYW5*
11	8.296e-14	chr15:91416560-91429040	N	N	*FES FURIN MAN2A2*
12	1.053e-13	chr3:36843183-36945783	N	Y	*TRANK1*
13	1.363e-13	chr14:103996234-104184834	N	N	*AL049840.1 APOPT1 BAG5 CKB KLC1 PPP1R13B TRMT61A XRCC3 ZFYVE21*
14	2.439e-13	chr15:78803032-78926732	N	N	*AC027228.1 AGPHD1 CHRNA3 CHRNA5 CHRNB4 IREB2 PSMA4*
15	3.034e-13	chr7:110843815-111205915	N	N	*IMMP2L*
16	1.088e-12	chr11:130714610-130749330	Y	N	*SNX19*
17	1.53e-12	chr2:185601420-185785420	Y	N	*ZNF804A*
18	1.606e-12	chrX:21193266-21570266	N	N	*CNKSR2*
19	1.971e-12	chr10:18681005-18770105	Y	N	*CACNB2*
20	2.015e-12	chr12:57428314-57682971	N	N	*LRP1 MYO1A NAB2 NDUFA4L2 NXPH4 R3HDM2 SHMT2 STAC3 STAT6 TAC3 TMEM194A*

21	2.025e-12	chr1:73766426-73991366	Y	N	*LRRIQ3 **
22	2.315e-12	chr2:233559301-233753501	Y	N	*C2orf82 EFHD1 GIGYF2 KCNJ13 NGEF*
23	2.804e-12	chr11:124610007-124620147	Y	N	*ESAM MSANTD2 NRGN VSIG2*
24	3.337e-12	chr18:52747686-53200117	Y	N	*TCF4*
25	1.259e-11	chr11:46342943-46751213	Y	N	*AMBRA1 ARHGAP1 ATG13 CHRM4 CKAP5 CREB3L1 DGKZ F2 HARBI1 MDK ZNF408*
26	1.301e-11	chr3:180588843-181205585	N	N	*CCDC39 DNAJC19 FXR1*
27	1.462e-11	chr20:37361494-37485994	N	N	*ACTR5 PPP1R16B SLC32A1*
28	1.473e-11	chr2:57943593-58502192	Y	N	*FANCL VRK2*
29	1.618e-11	chr15:84661161-85153461	N	N	*ADAMTSL3 GOLGA6L4 ZSCAN2*
30	1.971e-11	chr18:53453389-53804154	N	N	*TCF4 **
31	2.064e-11	chr2:198148577-198835577	N	N	*ANKRD44 BOLL COQ10B HSPD1 HSPE1 MARS2 PLCL1 RFTN2 SF3B1*
32	2.069e-11	chr22:41408556-41675156	N	N	*CHADL EP300 L3MBTL2 RANGAP1*
33	2.607e-11	chr8:111460061-111630761	N	N	*KCNV1 **
34	2.692e-11	chr3:2532786-2561686	N	N	*CNTN4*
35	2.749e-11	chr11:113317794-113423994	N	N	*DRD2*
36	3.874e-11	chr11:133808069-133852969	N	N	*IGSF9B*
37	4.264e-11	chr3:52541105-52903405	Y	N	*GLT8D1 GNL3 ITIH1 ITIH3*
38	4.548e-11	chr16:29924377-30144877	N	N	*ALDOA ASPHD1 C16orf92 DOC2A FAM57B GDPD3 HIRIP3 INO80E KCTD13 MAPK3 PPP4C SEZ6L2 TAOK2 TBX6 TMEM219 YPEL3*
39	4.725e-11	chr22:39975317-40016817	N	N	*CACNA1I*
40	7.264e-11	chr3:135807405-136615405	N	N	*MSL2 NCK1 PCCB PPP2R3A SLC35G2 STAG1*
41	1.055e-10	chr5:151941104-152797656	Y	N	*GRIA1 **
42	1.982e-10	chrX:68377126-68379036	N	N	*PJA1*
43	2.862e-10	chr17:2095899-2220799	N	N	*SGSM2 SMG6 SRR TSR1*

44	3.332e-10	chr7:86403226-86459326	N	N	*GRM3*
45	3.384e-10	chr15:61831663-61909663	N	N	*VPS14C **
46	3.394e-10	chr1:44029384-44128084	N	N	*KDM4A PTPRF*
47	3.634e-10	chr19:19374022-19658022	Y	Y	*CILP2 GATAD2A HAPLN4 MAU2 NCAN NDUFA13 PBX4 SUGP1 TM6SF2 TSSK6*
48	4.487e-10	chr1:149998890-150242490	N	N	*ANP32E APH1A C1orf51 C1orf54 CA14 OTUD7B PLEKHO1 VPS45*
49	8.147e-10	chr6:84279922-84407274	N	N	*SNAP91*
50	8.701e-10	chr1:2372401-2402501	N	N	*PLCH2*
51	1.009e-9	chr16:13728459-13761359	N	N	*ERCC4 **
52	1.127e-9	chr7:104598064-105063064	N	N	*MLL5 PUS7 SRPK2*
53	1.166e-9	chr1:8411184-8638984	N	N	*RERE SLC45A1*
54	1.398e-9	chr12:110723245-110723245	N	N	*ATP2A2*
55	1.465e-9	chr4:170357552-170646052	N	N	*C4orf27 CLCN3 NEK1*
56	1.644e-9	chr6:96459651-96459651	N	N	*FUT9*
57	1.709e-9	chr22:42315744-42689414	N	N	*CENPM CYP2D6 FAM109B NAGA NDUFA6 SEPT3 SHISA8 SMDT1 SREBF2 TCF20 TNFRSF13C WBP2NL*
58	1.814e-9	chr2:146416922-146441832	N	N	
59	2.243e-9	chr11:57386294-57682294	N	N	*BTBD18 C11orf31 CLP1 CTNND1 MED19 SERPING1 TMX2 YPEL4 ZDHHC5*
60	2.554e-9	chr11:24367320-24412990	N	N	*LUZP2 **
61	2.86e-9	chr1:30412551-30437271	N	N	
62	3.278e-9	chr7:137039644-137085244	N	N	*DGKI PTN*
63	3.613e-9	chr9:84630941-84813641	N	N	*TLE1*
64	3.73e-9	chr1:243503719-244002945	Y	N	*AKT3 SDCCAG8*
65	4.178e-9	chr15:40566759-40602237	N	N	*ANKRD63 PAK6 PLCB2*
66	4.489e-9	chr19:30981643-31039023	N	N	*ZNF536*
67	4.606e-9	chr5:88581331-88854331	N	N	*MEF2C **

68	4.642e-9	chr3:17221366-17888266	N	N	*TBC1D5*
69	4.666e-9	chr5:137598121-137948092	N	N	*CDC25C CTNNA1 EGR1 ETF1 FAM53C GFRA3 HSPA9 KDM3B REEP2*
70	4.801e-9	chr14:99707919-99719219	N	N	*BCL11B*
71	4.863e-9	chr14:72417326-72450526	N	N	*AC005477.1 RGS6*
72	5.046e-9	chr5:45291475-45393775	N	N	*HCN1*
73	5.97e-9	chr8:60475469-60954469	N	N	*CA8 **
74	7.388e-9	chr2:72357335-72368185	N	N	*CYP26B1*
75	7.539e-9	chr11:123394636-123395986	N	N	*GRAMD1B*
76	8.333e-9	chr2:200161422-200309252	N	N	*SATB2*
77	8.408e-9	chr2:193848340-194028340	Y	N	*PCGEM1 **
78	9.469e-9	chr4:176851001-176875801	N	N	*GPM6A*
79	1.058e-8	chr8:4177794-4192544	Y	N	*CSMD1*
80	1.115e-8	chr2:225334096-225467796	N	N	*CUL3*
81	1.215e-8	chr8:89340626-89753626	Y	N	*MMP16*
82	1.28e-8	chr16:9875519-9970219	N	N	*GRIN2A*
83	1.411e-8	chr14:30189985-30190316	N	N	*PRKD1*
84	1.432e-8	chr3:63792650-64004050	N	N	*ATXN7 C3orf49 PSMD6 THOC7*
85	1.513e-8	chr16:67709340-68311340	N	N	*ACD C16orf86 CENPT CTRL DDX28 DPEP2 DPEP3 DUS2L EDC4 ENKD1 ESRP2 GFOD2 LCAT NFATC3 NRN1L NUTF2 PARD6A PLA2G15 PSKH1 PSMB10 RANBP10 SLC12A4 SLC7A6 SLC7A6OS THAP11 TSNAXIP1*
86	1.585e-8	chr2:149390778-149520178	N	N	*EPC2*
87	1.769e-8	chr17:17722402-18030202	N	N	*ATPAF2 DRG2 GID4 LRRC48 MYO15A RAI1 SREBF1 TOM1L2*
88	1.787e-8	chr15:70573672-70628872	N	N	*TLE3 **
89	1.873e-8	chr16:58669293-58682833	N	N	*CNOT1 SLC38A7*
90	2.096e-8	chr8:27412627-27453627	N	N	*CLU EPHX2*
91	2.205e-8	chrX:5916533-6032733	N	N	*NLGN4X*

92	2.688e-8	chr6:73132701-73171901	N	N	*RIMS1*
93	2.853e-8	chr7:24619494-24832094	N	N	*DFNA5 MPP6 OSBPL3*
94	3.053e-8	chr5:109030036-109209066	N	N	*MAN2A1*
95	3.056e-8	chr4:23366403-23443403	N	N	*MIR548AJ2 **
96	3.145e-8	chr5:153671057-153688217	N	N	*GALNT10*
97	3.695e-8	chr11:109285471-109610071	N	N	*C11orf87*
98	3.713e-8	chr7:110034393-110106693	N	N	*IMMP2L **
99	3.906e-8	chr12:29905265-29940365	N	N	*TMTC1*
100	4.417e-8	chr7:131539263-131567263	N	N	*PODXL **
101	4.448e-8	chr1:177247821-177300821	N	N	*FAM5B*
102	4.468e-8	chr1:207912183-208024083	N	N	*C1orf132 CD46 CR1L*
103	4.559e-8	chr20:48114136-48131649	N	N	*KCNB1 PTGIS*
104	4.591e-8	chr12:92243186-92258286	N	N	*C12orf79 **
105	4.615e-8	chr2:162798555-162910255	N	N	*DPP4 SLC4A10*
106	4.686e-8	chr19:50067499-50135399	N	N	*NOSIP PRR12 PRRG2 RCN3 RRAS SCAF1*
107	4.843e-8	chr12:103559855-103616655	N	N	*C12orf42*
108	4.849e-8	chr5:140023664-140222664	N	N	*AC005609.1 CD14 DND1 HARS HARS2 IK NDUFA2 PCDHA1 PCDHA10 PCDHA2 PCDHA3 PCDHA4 PCDHA5 PCDHA6 PCDHA7 PCDHA8 PCDHA9 TMCO6 WDR55 ZMAT2*

† indicates whether the locus has been previously reported with SCZ (schizophrenia) or BPD (bipolar disorder)

7.2 The 10 loci significantly associated with BPD

Locus	P-value	Genes	Reference
12p13.3	1.5e-8	*CACNA1C*	(Ferreira et al., 2008)
10q21.2	9.1e-9	*ANK3*	(Ferreira et al., 2008)
11q14.1	4.4e-8	*ODZ4*	(Sklar et al., 2011)
19p13.1	2.1e-9	*NCAN*	(Cichon et al., 2011)
3p22.2	2.4e-11	*TRANK1*	(Chen et al., 2013)
6q25.1	2.9e-8	*SYNE1*	(Green et al., 2013a)
12q13.1	9.0e-9	*RHEBL1 DHH*	(Green et al., 2013b)
20q11.2	3.9e-8	*TRPC4AP GSS MYH7B*	(Green et al., 2013b)
5p15.3	9.9e-9	*ADCY2*	(Mühleisen et al., 2014)
6q16.1	1.1e-8	*MIR2113 POU3F2*	(Mühleisen et al., 2014)

7.3 List of local origins (Chioggia or Sottomarina) for the 115 'family founders'

Ind	Family	Reported origin	Revised origin	Outsider	Sample type
1_1	1	C	C	N	familial
1_2	1	C	C	N	familial
1_3	1	C	C	N	familial
2_4	2	C	C	N	familial
3_1	3	C	C	N	familial
4_11	**4a**	**S**	**C**	**Y**	familial
4_2	4	S	S	N	familial
4_4	4	S	S	N	familial
5_3	5	S	S	N	familial
6_1	**6**	**S**	**C**	N	**familial_single**
7_1	7	C	C	N	familial
8_1	8	C	C	N	familial
9_1	9	C	C	N	familial
11_1	11	C	C	N	familial

13_1	13	S	C	Y	familial
14_2	14	S	C	N	familial_single
15_2	15	S	C	N	familial
15_5	15	S	C	N	familial
16_1	16	S	S	N	familial
17_1	17	C	C	N	familial_single
19_1	19	S	S	N	familial
20_1	20	S	S	N	familial
20_2	20	S	S	N	familial
21_1	21	C	S	N	familial
22_1	22	C	C	N	familial
22_6	22	C	C	N	familial
23_2	23	C	C	N	familial
24_1	24	C	C	N	familial
26_1	26	S	C	N	familial
26_3	26	S	C	N	familial
26_6	26	S	C	N	familial
27_1	27	C	C	N	familial
28_1	28	S	C	N	familial_single
34_1	34	S	C	Y	familial_single
34_2	34	S	S	N	familial_single
40_1	40	S	S	N	familial
44_1	44	C	C	Y	familial_single
45_1	45	S	S	N	familial_single
47_1	47	C	C	N	familial
48_1	48	C	C	N	familial
48_4	48	C	C	N	familial
53_1	53	ND	C	N	familial_single
54_1	54	ND	C	N	familial_single
55_1	55	ND	C	Y	familial_single
56_1	56	S	C	N	familial_single
62_1	62	S	S	N	familial
62_4	62	S	S	N	familial_single
63_1	63	ND	C	N	familial_single
64_1	64	ND	C	N	familial_single
69_1	69	S	S	N	familial_single
78_1	78	ND	C	N	familial
79_2	79	S	S	N	familial
79_5	79	S	S	N	familial
80_1	80	S	S	N	familial

84_1	84	S	S	N	familial
85_1	85	S	S	N	familial
86_1	**86**	**ND**	**C**	N	**familial_single**
87_1	**87**	**ND**	**C**	**Y**	**familial_single**
88_1	**88**	**ND**	**C**	N	**familial_single**
90_2	90	C	C	N	familial_single
91_1	**91**	**ND**	**C**	**Y**	**familial_single**
97_1	97	S	S	N	familial_single
99_3	99	C	C	N	familial
100_1	**100**	**ND**	**C**	N	**familial_single**
103_1	103	S	S	N	familial_single
109_1	109	C	C	N	familial
109_2	109	C	C	N	familial_single
110_1	110	C	C	N	familial
111_1	111	C	C	N	familial
115_3	**115**	**ND**	**C**	N	**familial**
29	29	C	C	N	isolated
30	30	S	S	N	isolated
31	31	S	S	N	isolated
32	32	C	C	N	isolated
35	**35**	**S**	**C**	N	**isolated**
36	36	S	S	N	isolated
37	37	C	C	N	isolated
38	**38**	**S**	**C**	N	**isolated**
39	39	C	C	N	isolated
41	41	C	C	N	isolated
42	42	C	C	N	isolated
50	**50**	**ND**	**C**	N	**isolated**
51	51	S	S	N	isolated
52	52	C	C	N	isolated
57	57	S	S	N	isolated
58	**58**	**ND**	**C**	N	**isolated**
59	**59**	**ND**	**C**	N	**isolated**
60	60	C	C	N	isolated
61	**61**	**ND**	**C**	N	**isolated**
67	**67**	**ND**	**C**	N	**isolated**
68	**68**	**ND**	**C**	N	**isolated**
70	**70**	**ND**	**C**	**Y**	**isolated**
71	**71**	**ND**	**C**	**Y**	**isolated**
72	**72**	**ND**	**C**	**Y**	**isolated**

73	**73**	**ND**	**C**	**Y**	**isolated**
74	**74**	**ND**	**C**	**Y**	**isolated**
75	75	C	C	N	isolated
76	**76**	**S**	**C**	N	**isolated**
96	**96**	**ND**	**C**	N	**isolated**
101	101	C	C	N	isolated
104	**104**	**S**	**C**	N	**isolated**
c6484	c6484	S	S	N	control
c6485	c6485	S	S	N	control
c7791	c7791	S	S	N	control
c7863	**c7863**	**S**	**C**	**Y**	**control**
c7995	c7995	S	S	N	control
c7996	c7996	C	C	N	control
c7998	c7998	S	S	N	control
c8002	c8002	S	S	N	control
c8008	c8008	S	S	N	control
c8018	c8018	C	C	N	control
c8019	c8019	S	S	N	control
c8047	c8047	S	S	N	control
c8049	c8049	C	C	N	control
c8050	c8050	C	C	N	control

For each individual is reported the initially reported origin and the revised one in the light of cluster analysis. The outsider column indicate whether the subject was grouped in the Italian cluster outside Chioggia population.

7.4 Detailed results of the functional enrichment analysis

all rare or novel variants							
Category	Term	Matched genes	%	PValue	Tot genes tested	Tot genes in the category	Benjamini
KEGG_PATHWAY	hsa04510:Focal adhesion	153	1.384615	1.24E-05	3133	201	0.00246025
KEGG_PATHWAY	hsa04512:ECM-receptor interaction	70	0.633484	3.42E-05	3133	84	0.00339728
KEGG_PATHWAY	hsa00500:Starch and sucrose metabolism	38	0.343891	1.38E-04	3133	42	0.00909936
KEGG_PATHWAY	hsa00230:Purine metabolism	116	1.049774	2.22E-04	3133	153	0.01100489
KEGG_PATHWAY	hsa04070:Phosphatidylinositol signaling system	61	0.552036	2.31E-04	3133	74	0.00913323
KEGG_PATHWAY	hsa04520:Adherens junction	63	0.570136	2.65E-04	3133	77	0.00874236
REACTOME_PATHWAY	REACT_13685:Synaptic Transmission	65	0.588235	2.93E-04	2061	81	0.02059656
REACTOME_PATHWAY	REACT_11044:Signaling by Rho GTPases	93	0.841629	5.48E-04	2061	123	0.0192842
KEGG_PATHWAY	hsa04360:Axon guidance	98	0.886878	6.68E-04	3133	129	0.01881109
KEGG_PATHWAY	hsa02010:ABC transporters	38	0.343891	0.0010252	3133	44	0.02519145
REACTOME_PATHWAY	REACT_18266:Axon guidance	41	0.371041	0.0012987	2061	49	0.03028768
REACTOME_PATHWAY	REACT_16888:Signaling by PDGF	51	0.461538	0.0021497	2061	64	0.03747722
KEGG_PATHWAY	hsa00562:Inositol phosphate metabolism	44	0.39819	0.0035247	3133	54	0.07510288
KEGG_PATHWAY	hsa04930:Type II diabetes mellitus	39	0.352941	0.0036813	3133	47	0.07076393
KEGG_PATHWAY	hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	59	0.533937	0.0045288	3133	76	0.07883426
REACTOME_PATHWAY	REACT_11061:Signalling by NGF	124	1.122172	0.0055847	2061	176	0.07644444
KEGG_PATHWAY	hsa04144:Endocytosis	131	1.18552	0.0056599	3133	184	0.08983177
REACTOME_PATHWAY	REACT_474:Metabolism of carbohydrates	70	0.633484	0.005945	2061	94	0.06812693
REACTOME_PATHWAY	REACT_13552:Integrin cell surface interactions	61	0.552036	0.0072557	2061	81	0.07120022
KEGG_PATHWAY	hsa04270:Vascular smooth muscle contraction	82	0.742081	0.0106872	3133	112	0.15166258
REACTOME_PATHWAY	REACT_216:DNA Repair	74	0.669683	0.0129107	2061	102	0.1089269
REACTOME_PATHWAY	REACT_17044:Muscle contraction	26	0.235294	0.0146287	2061	31	0.10975318
REACTOME_PATHWAY	REACT_602:Metabolism of lipids and lipoproteins	105	0.950226	0.0149045	2061	150	0.10113142
KEGG_PATHWAY	hsa05414:Dilated	68	0.615385	0.015198	3133	92	0.19562471

	cardiomyopathy			2			
KEGG_PATHWAY	hsa00053:Ascorbate and aldarate metabolism	16	0.144796	0.0169011	3133	17	0.20239126
KEGG_PATHWAY	hsa05410:Hypertrophic cardiomyopathy (HCM)	63	0.570136	0.0181847	3133	85	0.2040781
KEGG_PATHWAY	hsa05200:Pathways in cancer	221	2	0.0194285	3133	328	0.20520063
KEGG_PATHWAY	hsa04730:Long-term depression	52	0.470588	0.0210967	3133	69	0.21000687
KEGG_PATHWAY	hsa04330:Notch signaling pathway	37	0.334842	0.0216158	3133	47	0.20457542
KEGG_PATHWAY	hsa04720:Long-term potentiation	51	0.461538	0.0258583	3133	68	0.22946892
KEGG_PATHWAY	hsa04910:Insulin signaling pathway	95	0.859729	0.0303397	3133	135	0.2531993
KEGG_PATHWAY	hsa00410:beta-Alanine metabolism	19	0.171946	0.0367997	3133	22	0.28762365
KEGG_PATHWAY	hsa00640:Propanoate metabolism	26	0.235294	0.0368289	3133	32	0.27723133
KEGG_PATHWAY	hsa04650:Natural killer cell mediated cytotoxicity	93	0.841629	0.0403246	3133	133	0.28914493
KEGG_PATHWAY	hsa04912:GnRH signaling pathway	70	0.633484	0.0422224	3133	98	0.29064005
KEGG_PATHWAY	hsa00970:Aminoacyl-tRNA biosynthesis	32	0.289593	0.0423995	3133	41	0.2822253
KEGG_PATHWAY	hsa04666:Fc gamma R-mediated phagocytosis	68	0.615385	0.0427692	3133	95	0.27542099
KEGG_PATHWAY	hsa00280:Valine, leucine and isoleucine degradation	34	0.307692	0.0435831	3133	44	0.27145386
KEGG_PATHWAY	hsa04010:MAPK signaling pathway	179	1.61991	0.044951	3133	267	0.27065205
KEGG_PATHWAY	hsa05416:Viral myocarditis	52	0.470588	0.0460626	3133	71	0.26861098
REACTOME_PATHWAY	REACT_604:Hemostasis	156	1.411765	0.0470707	2061	235	0.26743479
rare or novel IBD_{tot} variants							
KEGG_PATHWAY	**hsa04512:ECM-receptor interaction**	**44**	**1.009637**	**2.59E-07**	**1288**	**84**	**5.11E-05**
REACTOME_PATHWAY	**REACT_11044:Signaling by Rho GTPases**	**51**	**1.170262**	**2.79E-05**	**822**	**123**	**0.00189395**
KEGG_PATHWAY	**hsa02010:ABC transporters**	**25**	**0.573658**	**3.06E-05**	**1288**	**44**	**0.00301196**
REACTOME_PATHWAY	**REACT_18266:Axon guidance**	**26**	**0.596604**	**3.82E-05**	**822**	**49**	**0.00129938**
KEGG_PATHWAY	**hsa04510:Focal adhesion**	**74**	**1.698027**	**2.70E-04**	**1288**	**201**	**0.01756438**
REACTOME_PATHWAY	REACT_604:Hemostasis	76	1.743919	0.0031597	822	235	0.06922081
KEGG_PATHWAY	hsa00500:Starch and sucrose metabolism	20	0.458926	0.0038343	1288	42	0.17237915
KEGG_PATHWAY	hsa04070:Phosphatidylinositol signaling system	30	0.688389	0.0056449	1288	74	0.19991724

REACTOME_PATHWAY	REACT_13685:Synaptic Transmission	31	0.711335	0.0060169	822	81	0.0975913
KEGG_PATHWAY	hsa04720:Long-term potentiation	28	0.642497	0.0060544	1288	68	0.1807695
KEGG_PATHWAY	hsa05200:Pathways in cancer	103	2.363469	0.0085123	1288	328	0.2138325
KEGG_PATHWAY	hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	30	0.688389	0.0086864	1288	76	0.1933765
REACTOME_PATHWAY	REACT_13552:Integrin cell surface interactions	30	0.688389	0.0116440	822	81	0.1472244
KEGG_PATHWAY	hsa05410:Hypertrophic cardiomyopathy (HCM)	32	0.734282	0.0140891	1288	85	0.2669347
KEGG_PATHWAY	hsa00052:Galactose metabolism	13	0.298302	0.0168128	1288	26	0.2839257
REACTOME_PATHWAY	REACT_16888:Signaling by PDGF	24	0.550711	0.0221794	822	64	0.2244606
KEGG_PATHWAY	hsa00230:Purine metabolism	51	1.170262	0.0230076	1288	153	0.3408678
KEGG_PATHWAY	hsa00250:Alanine, aspartate and glutamate metabolism	14	0.321248	0.0313358	1288	31	0.4070596
KEGG_PATHWAY	hsa00562:Inositol phosphate metabolism	21	0.481872	0.0368902	1288	54	0.4342650
KEGG_PATHWAY	hsa04360:Axon guidance	43	0.986691	0.0372199	1288	129	0.4135881
KEGG_PATHWAY	hsa00053:Ascorbate and aldarate metabolism	9	0.206517	0.0427946	1288	17	0.4369668
KEGG_PATHWAY	hsa05414:Dilated cardiomyopathy	32	0.734282	0.0433495	1288	92	0.4205380
KEGG_PATHWAY	hsa00982:Drug metabolism	23	0.527765	0.0474149	1288	62	0.4304974
rare or novel IBD_{sel} variants							
KEGG_PATHWAY	hsa04360:Axon guidance	10	1.945525	0.0163544	156	129	0.8865761
KEGG_PATHWAY	hsa04070:Phosphatidylinositol signaling system	7	1.361868	0.0245337	156	74	0.8059506
BIOCARTA	h_blymphocytePathway:B Lymphocyte Cell Surface Molecules	3	0.583658	0.0247183	34	11	0.9095332
REACTOME_PATHWAY	REACT_18266:Axon guidance	5	0.972763	0.0259858	80	49	0.5914742
REACTOME_PATHWAY	REACT_15380:Diabetes pathways	13	2.529183	0.0299183	80	285	0.4033207
REACTOME_PATHWAY	REACT_13685:Synaptic Transmission	6	1.167315	0.0387571	80	81	0.3610864
KEGG_PATHWAY	hsa04514:Cell adhesion molecules (CAMs)	9	1.750973	0.0474614	156	132	0.8822844
all rare or novel LoF variants							
KEGG_PATHWAY	**hsa02010:ABC transporters**	**30**	**0.573504**	**1.65E-07**	**1424**	**44**	**3.22E-05**
KEGG_PATHWAY	**hsa04512:ECM-receptor interaction**	**46**	**0.879373**	**5.68E-07**	**1424**	**84**	**5.54E-05**
KEGG_PATHWAY	**hsa05416:Viral myocarditis**	**34**	**0.649971**	**6.33E-04**	**1424**	**71**	**0.0403444**

REACTOME_PATHWAY	REACT_13552:Integrin cell surface interactions	36	0.688205	9.85E-04	915	81	0.0647 8796
KEGG_PATHWAY	hsa04612:Antigen processing and presentation	36	0.688205	0.0035831	1424	83	0.1605 3626
KEGG_PATHWAY	hsa00500:Starch and sucrose metabolism	21	0.401453	0.0051358	1424	42	0.1819 3541
KEGG_PATHWAY	hsa04510:Focal adhesion	73	1.395527	0.00822	1424	201	0.2352 8651
KEGG_PATHWAY	hsa00563:Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	14	0.267635	0.0093588	1424	25	0.2304 2331
KEGG_PATHWAY	hsa00562:Inositol phosphate metabolism	24	0.458803	0.0144798	1424	54	0.2991 9485
KEGG_PATHWAY	hsa00970:Aminoacyl-tRNA biosynthesis	19	0.363219	0.0206359	1424	41	0.3635 1192
REACTOME_PATHWAY	REACT_18266:Axon guidance	21	0.401453	0.0232278	915	49	0.5502 494
KEGG_PATHWAY	hsa04070:Phosphatidylinositol signaling system	30	0.573504	0.0235308	1424	74	0.3714 4704
KEGG_PATHWAY	hsa05332:Graft-versus-host disease	18	0.344102	0.0260466	1424	39	0.3736 5539
REACTOME_PATHWAY	REACT_649:Phase 1 functionalization	9	0.172051	0.0265009	915	15	0.4559 9121
REACTOME_PATHWAY	REACT_11044:Signaling by Rho GTPases	44	0.841139	0.0277497	915	123	0.3802 3393
KEGG_PATHWAY	hsa04514:Cell adhesion molecules (CAMs)	48	0.917607	0.0322519	1424	132	0.4130 0087
KEGG_PATHWAY	hsa04360:Axon guidance	46	0.879373	0.0497274	1424	129	0.5347 1057
rare or novel LoF IBDtot variants							
KEGG_PATHWAY	hsa04512:ECM-receptor interaction	18	1.123596	9.11E-04	452	84	0.1442 58
KEGG_PATHWAY	hsa00830:Retinol metabolism	12	0.749064	0.0067692	452	54	0.4405 127
KEGG_PATHWAY	hsa00982:Drug metabolism	13	0.811486	0.0073035	452	62	0.3415 2429
REACTOME_PATHWAY	REACT_11044:Signaling by Rho GTPases	19	1.186017	0.0078009	272	123	0.3345 1467
KEGG_PATHWAY	hsa00983:Drug metabolism	10	0.624242	0.0115423	452	43	0.3912 2371
KEGG_PATHWAY	hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	14	0.873908	0.0150983	452	76	0.4056 5755
KEGG_PATHWAY	hsa00860:Porphyrin and chlorophyll metabolism	8	0.499376	0.0231169	452	33	0.4865 3158
REACTOME_PATHWAY	REACT_649:Phase 1 functionalization	5	0.312151	0.0267303	272	15	0.5056 1722
KEGG_PATHWAY	hsa00500:Starch and sucrose metabolism	9	0.561798	0.0288277	452	42	0.5105 9699
REACTOME_PATHWAY	REACT_18266:Axon guidance	9	0.561798	0.0376947	272	49	0.4862 4483
KEGG_PATHWAY	hsa04640:Hematopoietic cell lineage	14	0.873908	0.0383841	452	86	0.5668 2807
KEGG_PATHWAY	hsa04514:Cell adhesion molecules (CAMs)	19	1.186017	0.0421965	452	132	0.5591 8904

7.5 Detailed information about the investigated variants

type	chr	genomic position	ref	variant	gene	effect	prediction
rare	1	93964859	G	A	FNBP1L	intronic	damaging
rare	1	181727109	C	T	CACNA1E	synonymous	damaging
rare	1	205028229	C	T	CNTN2	missense	damaging
rare	3	7494298	G	A	GRM7	synonymous	damaging
new	3	173998947	A	G	NLGN1	missense	damaging
rare	7	80430250	G	T	SEMA3C	intronic	neutral
new	7	126173624	G	A	GRM8	synonymous	damaging
new	8	38271545	C	T	FGFR1	intronic	damaging
new	11	113102000	G	GGCT	NCAM1	intronic	neutral
rare	11	128842742	C	T	ARHGAP32	missense	neutral
rare	16	24231449	C	T	PRKCB	utr_3	damaging
new	16	83520283	T	C	CDH13	intronic	neutral
rare	16	83712083	G	A	CDH13	intronic	neutral
rare	17	7096390	G	A	DLG5	synonymous	damaging
rare	17	7096390	G	A	DLG4	synonymous	damaging
new	17	7120502	T	TCA	DLG4	utr_5	damaging
new	19	19356155	A	C	NCAN	missense	damaging

7.6 Primer list for Sanger sequencing of candidate variants

Primer	Sequence
DLG4a_F	ACTTCTGATGCTGCCACTCC
DLG4a_R	CGAAATCCAGTTCCCCTCCC
DLG4b_F	AGGGGAGAAGGAAGAGGACC
DLG4b_R	GCTGGGCACTATGGGGATTT
CACNA1E_F	GCTAATAACCATGACTTTCT
CACNA1E_R	ATAAGATCACAAAAGGATTC
PRKCB_F	CAGAGAGACAAGAGAGACAC
PRKCB_R	CAGACACAGTAGTTTTGACA
GRM7_F	GTATGTAAGGTGTCACAAAG
GRM7_R	CAGAGATCCTTGTTCATG
GRM8_F	GATGCTGGACTAATGAACT
GRM8_R	GCTGTGAAGGTTACAACT
CNTN2_F	ACTTTCCATCGTTGTGCCCT
CNTN2_R	AGGAGTAGTTGCCCAGGTCT
NCAM1_F	GCAGACAGAAAGGACAGGCT
NCAM1_R	ATTGGTTTGGGGCTAGGTCC

NCAN_F	CACCACTTATCTAACACTGA
NCAN_R	AGACATAGGGTAGGTTGTAG
ARHGAP32_F	TGGAGTAGTTGTCATGTAAG
ARHGAP32_R	TTACTTCAGTTCCCTTAGAC
CDH13_F	TTTCTGCCTTCCACACCCTG
CDH13_R	GTCAGCCACAACCTCCTCTC