USING WHOLE GENOME SEQUENCING TO IDENTIFY RISK ALLELES FOR SUSCEPTIBILITY TO SCHIZOPHRENIA

By

GILLIAN K. DAVIS

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Microbiology and Molecular Genetics

Written under the direction of

Linda Brzustowicz, M.D.

And Approved By

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2017

ABSTRACT OF THE DISSERTATION

Using Whole Genome Sequencing to Identify Risk Alleles for Susceptibility to Schizophrenia

By GILLIAN K. DAVIS

Dissertation Director:

Linda Brzustowicz, M.D.

Schizophrenia is a complex idiopathic neuropsychiatric illness that affects approximately 1% of the general population. Family, twin, and adoption studies indicate a high heritability and strong genetic element to the disease with first degree relatives demonstrating an increased risk of about 10% and monozygotic concordance rates as high as 50%. These values represent the probability of developing schizophrenia based on the presence of genetic components. The high heritability has led to individual studies and meta-analyses being able to produce significant evidence of linkage to specific locations, but studies that used large number of pedigrees have failed to produce statistically significant linkage results. Genome Wide Association Studies of schizophrenia have also produced similarly mixed results. One interpretation of these mixed linkage and association results is that factors such as small effect size and uncontrolled phenotypic variation require very large samples to overcome. This thesis focuses on a different interpretation: genuine genetic differences between definable

subsets can mask both linkage and association, and that this problem is worsened in studies that use large samples where the entire sample is analyzed as if it were a genetically homogenous group.

The work presented herein begins with linkage studies performed on 22 medium-sized Canadian pedigrees (n=304 individuals) of German or Celtic descent initially recruited if at least three subjects with schizophrenia were available for study. Association studies were conducted on an expanded sample of 30 pedigrees (n=573). Subjects in this sample have been followed for up to 20 years allowing for continued observation of diagnostic stability. We have identified linkage disequilibrium between schizophrenia and single nucleotide polymorphisms (SNPs) from six discrete genomic regions located under linkage peaks within this sample. We hypothesize that SNPs that generated compelling evidence of association (PPLD|L >= 0.2) produce these scores because they either are, or are in, high LD (r^2 >= 0.8) with functional variants that increase susceptibility to schizophrenia. To that end, whole genome sequencing data from ten individuals within this study (n=10) was analyzed to generate a list of variants within 500 kb upstream and downstream of each risk SNP. A pipeline was created to determine whether or not each SNP in this list was a candidate for further analysis by assessing its LD to the risk SNPs identified by the association studies described above. SNPs determined to be candidates were then genotyped in the entire sample (n=378) so that association could be accurately assessed. Finally, association scores were compared between risk SNPs and candidate SNPs, with variants having higher PPLD|L scores than the referring SNP identified as potential functional candidates. Six SNPs from one genomic region produced higher PPLD|L scores than the referring SNP and so will

replace the referring SNP as candidates for further functional analysis. These six SNPs first will be evaluated for additional candidate SNPs 500 kb up- and down-stream in order to determine the best SNP in the region according to the PPLD|L. Additional SNPs have also been identified in some of the other genomic regions that need to be assessed for LD in the full sample. The SNP or SNPs producing the strongest LD signal in each region will need to be further assessed by functional assays to determine their potential role in schizophrenia susceptibility.

DEDICATION

This dissertation is dedicated to the following people:

My mother, Joy Silver, whose financial and emotional support made it possible for me to pursue college seven years after I graduated high school.

My father, Ira Silver, whose unconditional love and guidance made it possible for me to eventually accept his sudden death during high school and continue my pursuit of the dreams he held for me.

My close friend, Ion Hazzikostas, without whose generosity none of what preceded this and certainly all of what follows would have been completely impossible to have achieved.

My husband, Garry Davis, who made it possible for me to achieve balance in my life. This what allowed me to maintain the health and happiness necessary to complete the stressful endeavor of a PhD while balancing a family, all without feeling any pressure at all.

And finally, my son Billy, who is the reason I found the courage to apply to graduate school in the first place, and who made me able to risk the thing I fear most in life, which is failure. Despite your many successes it is the times I have watched you fail and refuse to quit, only to see you later succeed and shine in the success of your hard work, that have motivated me to match the example you set.

Thank you, I love you all.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Linda Brzustowicz, M.D. for allowing me to be a graduate student in her lab, especially because when I asked it was because I was in dire need of a second rotation and had no intention of joining the lab full time. Approximately a month into my 8-week rotation I began to realize that I had found the right lab for me, academically and otherwise. As a non-traditional graduate student, who started her PhD as a 30 year old single mother, finding the right environment was a critical component to the whole endeavor being successful. Linda allowed me the freedom to work on my own schedule, providing me guidance without micro-managing, the swift kick in the rear when necessary, and most importantly, never once put me in the position of feeling like my family should come second to my research. I have had many supervisors over the course of my life, both in academia and out, and never before have I felt so motivated to produce my best work at an accelerated pace simply out of respect and admiration for my boss.

I would also like to thank the members of my committee, Christopher Bartlett, Ph.D., Bonnie Firestein, Ph.D., and Tara Matise Ph.D., who have all been extremely helpful over the years. Dr. Bartlett spent countless hours helping through learning how to use and understand the Kelvin software package. Dr. Firestein spent a lot of time helping me understand the underlying neurobiology for schizophrenia. Dr. Matise took the time to explain the Rutgers Maps and routinely asked questions at my committee meetings that ultimately led to a greater understanding of my project as a whole. I'd also like to thank Jim Millonig, Ph.D. for filling in as a temporary member for my Oral Proposal.

I am also tremendously grateful for the guidance and contributions I received from Veronica Vieland, Ph.D., Jo Valentine-Cooper, John Burian, M.S., and Sarah Bogen of the Vieland Lab at The Research Institute at Nationwide Children's Hospital for their work on this project and the software package Kelvin, without which this project would have been impossible.

I would also like to thank Anne Bassett, M.D. FRCPC, for her dedication and work with the sample used in this project over the past 20 years.

This would work not have been possible without contributions from the Brzustowicz Lab. Special thanks to Jared Hayter and Jaime Messenger for their work genotyping the SNPs of interest. Huge thanks to Marco Azaro, Ph.D., for his software and work on both primer design and genotyping, and for all of his help during my most panicked moments of working on my project. Thanks to William Manley, Ph.D. for being a great mentor and giving a ton of excellent advice over the years, both in the lab and the graduate school in general. Also, thanks for patiently helping me navigate New Brunswick since I still don't really know my way around 5 years later, haha. Thanks also to Judy Flax, Ph.D., Brittany Greene, Vaidhynathan Mahaganapathy, Sherri Wilson, Christine Gwin, Brenda Patel, and Karen Law, for all of the helpful advice over the years on my presentations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Schizophrenia

General Background

Schizophrenia is a complex idiopathic neurological disorder that was first described by Emil Kraepelin, a German psychiatrist, in 1899. Kraepelin referred to the disorder as 'dementia praecox' and formally separated it from manic depression due to the observation that it led to the irreversible loss of cognitive function [1]. Paul Eugen Bleuler renamed the disease at a conference in 1908, arguing that neither dementia nor precociousness was involved and that splitting of the psychic functioning was a more accurate description of the symptomology [2].

Schizophrenia often presents in the mid-teens to mid-twenties, with a slightly earlier onset in males compared to females [3]. There is no laboratory test that can definitively identify schizophrenia. A diagnosis is made by a licensed clinician based on criteria outlined by the Diagnostic and Statistical Manual of Mental Disorders (DSM). Classification by this diagnostic tool created by the American Psychiatric Association (APA) determines treatment recommendations, reimbursement to healthcare providers, and assistance (if any) provided to the patient. The DSM underwent its largest revision in nearly 20 years when it moved from DSM-4, published in 1994, to the current edition DSM-5, published in 2013. One of the most significant changes was the deletion of the subtypes of schizophrenia due to the fact that that they did not adequately explain the heterogeneity of symptoms [4].

The DSM-5 provides a detailed list of criteria that allows for the diagnosis of schizophrenia to be made when symptoms have existed for at least six months.

Symptoms of schizophrenia include positive symptoms such as hallucinations, delusions, disorganized speech, and disorganized behavior. Delusions are firmly held false beliefs, distinct from false or incomplete information. Hallucinations are sensory experiences that do not exist outside of the mind. Hallucinations can affect any of the five senses (sight, hearing, taste, smell, or touch), but the most common type observed in schizophrenia are auditory hallucinations [5]. There are also negative symptoms which include a diminished range of emotions (flattened affect), poverty of speech (alogia), anhedonia (inability to feel pleasure), and lack of motivation to execute daily tasks such as work/school (avolition).

In addition to the clinical symptoms associated with schizophrenia, cognitive impairment is also prevalent, which often has a detrimental impact on both quality of life and functional outcome. Multiple cognitive domains can be affected, such as attention, executive function, memory, social cognition, and language, though deficits will vary by individual [6]. These symptoms are likely caused by structural and functional abnormalities of the brain, which can be caused by: schizophrenia itself, epiphenomena resulting from schizophrenia, or antipsychotics used to treat the disease [7].

Structural changes within the brain are extensively described in schizophrenia throughout all stages of the disease. The most consistent findings are lateral ventricle enlargement, and grey matter volume reduction in prefrontal, temporal, and subcortical regions, as well as decreased white matter fractional anisotropy (a marker for white matter microstructure) providing connectivity between these regions. Though many studies have connected these changes to anti-psychotic interventions, the same abnormalities have been identified in untreated individuals. Substance abuse is often

comorbid in schizophrenia, with 50% of individuals with schizophrenia demonstrating alcohol or illicit drug use, and more than 70% having nicotine dependence [8]. While substance abuse could also be responsible for the observed structural changes to the brain, these changes are still seen in individuals with schizophrenia and little or no substance abuse. This evidence indicates that structural abnormalities frequently seen in individuals diagnosed with schizophrenia are part of the primary disease process (reviewed in [9]).

Cost to Society

Although schizophrenia carries a low lifetime prevalence compared to many disabilities, it is ranked 11th among all disabilities with respect to years lived with disability (YLDs) by the World Health Organization (WHO) [10]. Due to the early onset, the health, social and economic burden related to schizophrenia has been significant, not simply for those affected by the disease, but by their families, caregivers, and society as a whole [11]. There are three types of schizophrenia-related costs: direct cost, indirect cost, and intangible cost. Direct costs refer mainly to the treatment aspect costs of the disease and it includes: hospitalization (both short-term and long-term), outpatient follow-up, residential and day care, medication, laboratory testing, and social security payments. Indirect costs usually affect the income of the individual with the disease, as well as any familial caregivers who provide care at no cost to the patient. Intangible costs are non-financial in nature, and include side-effects of the medications and stress or anxiety caused by the disorder and/or the treatment process [12]. The total economic burden in the United States alone was estimated to be approximately $64 billion in 2002,

split evenly between direct and indirect costs, but this did not take into account the costs of the untreated, uninsured individuals with the disease who undoubtedly add additional economic burdens to society [13].

Though assessing economic burden is a tedious endeavor, it can be quantified because it is based on numbers. The intangible costs associated with the disease burden of schizophrenia should not be overlooked because they cannot be similarly tallied, yet often they are. In 1996, the publication of the results from the Global Burden of Disease Study finally revealed the disabling results of diseases in a comparative framework, demonstrating that mental health disorders should be a major public health concern. The Global Burden of Disease included "disability" in the equation, calculating Disability Adjusted Life Years (DALYs), and this led to mental health disorders ranking near to cardiovascular and respiratory diseases, and to mental health disorders exceeding all forms of cancer and HIV. The Global Burden of Disease Study also determined that the disability caused by major depression was found to be similar to that of blindness or paraplegia, and that disability caused by active psychosis as seen in schizophrenia was found to be somewhere between paraplegia and quadriplegia (as reviewed in [14]). It is therefore not surprising that while schizophrenia is not directly fatal, suicide incidence in those affected with the disease is 10%, which is 12x higher than seen in the general population [15]. Up to 50% of individuals with schizophrenia demonstrate alcohol or illicit drug addiction, and more than 70% are smokers resulting from chronic stress associated with the disease and/or patient attempts to self-medicate aspects of the illness that are not effectively treated by prescribed antipsychotics [8].

Neurobiology

Though the fundamental processes associated with schizophrenia remain uncertain, changes in various neurotransmitter systems have been implicated in the pathophysiological processes that culminate into the presentation of schizophrenia. Two of the most influential hypotheses regarding the neurobiology responsible for the disease involve dopamine and glutamate. Evidence for both hypotheses dates back more than half a century, but new evidence stemming from *in vivo* imaging studies and preclinical data on the role these neurotransmitters has clarified the understanding of dopamine and glutamate dysfunction in schizophrenia (reviewed in[16]).

The dopamine hypothesis originally resulted from several indirect sources of evidence beginning in the 1950s. The primary support came from evidence demonstrating that compounds which elevate extracellular levels of dopamine can cause psychotic symptoms similar to those exhibited in schizophrenia (reviewed in [17]). It was supported by studies demonstrating that drugs which reduce the level of dopamine also reduce those same psychotic symptoms [18]. Observations that the clinical effectiveness of antipsychotic medications was directly related to their ability to bind to dopamine receptors were made towards the end of the 1970s [19-21]. Though compelling, the evidence was nonspecific for two reasons: 1) some of the compounds such as amphetamine and reserpine are demonstrated to affect other brain monoamines besides dopamine and 2) dopamine is a non-specific treatment for any form of psychosis and not a specific treatment for schizophrenia, which in any case it does not fully eradicate (reviewed in [22]). Post-mortem studies beginning in the 1980s continuing to

present day have provided the anatomical and biochemical detail necessary to form more specific links between dopamine and schizophrenia, but these studies are limited by the fact that there is difficulty discerning if presynaptic and postsynaptic changes are due to disease pathology or if these observed changes are iatrogenic in etiology (reviewed in [16]). Imaging techniques such as Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) have provided important details of various elements of dopamine function in the brain through *in vivo* quantification, such as dopamine synthesis, the degree of dopamine release in response to stimuli, and the availability of post-synaptic dopamine receptors and transporters. PET/SPECT over the past two decades has allowed major aspects of the dopamine hypothesis to be examined (reviewed in [23]). Through the above outlined techniques, research has indicated that there is a link between dopamine and schizophrenia, and more specifically it is localized to presynaptic dysfunction, which leads to the symptoms of schizophrenia [16]. There are limitations to the evidence accumulated. First, there is documented treatment resistance in patients who do not respond to medications that address dopaminergic excess indicative of a clinical sub-type of the disease that does not stem from dopamine at all [24-26]. Second, direct causality between dopamine dysfunction and negative and cognitive symptoms has yet to be demonstrated [27]. And lastly, in the case of comorbid substance abuse and psychosis, other pathways may be indicated [28-30].

The glutamate hypothesis dates back as far as 1949 when patients diagnosed with schizophrenia were treated with glutamic acid [31]. In the 1980s a report was published demonstrating reduced cerebrospinal fluid (CSF) glutamate levels in patients with schizophrenia, though later studies were unable to reproduce this finding [32-34].

Glutamate has long been an attractive candidate in the underlying neurobiology of schizophrenia because 60-80% of total brain metabolic activity is utilized by glutamatergic neurons [35]. Neurotransmission handled by these neurons utilizes metabotropic and ionotropic glutamate receptors, with individual receptors classified into one of three groups. Metabotropic glutamate receptors are divided into two distinct groups based on whether they are postsynaptic (Group I) or presynaptic and modulate neurotransmitter release (Group 2). Ionotropic receptors (Group 3) are named after the antagonists originally discovered to selectively activate them (reviewed in [36]). Though the original hypothesis generalized glutamate's role in schizophrenia as being a simple deficit in glutamatergic neurotransmission, it has evolved over time to the prevailing hypothesis of N-methyl-D-aspartate (NMDA) receptor dysfunction (reviewed in [37]). Similar to the historical evolution of the dopamine hypothesis, post-mortem studies provided the first detailed evidence in support of glutamatergic function alteration in schizophrenia, but results have been inconsistent with respect to the causal role of NMDA receptor (NMDAR) density ([38], reviewed in [39]). New evidence suggests that neurobiological abnormalities in schizophrenia may not result from an overall deficit, but instead may be tied to abnormal glutamate receptor localization resulting from changes in glutamate receptor trafficking molecules [40, 41]. Additional support for the NMDAR hypothesis has come from observations that several non-competitive NMDAR antagonists (PCP, MK-801, and ketamine) lead to the acute onset of psychological effects mimicking both the positive and negative symptoms that manifest in schizophrenia ([42], reviewed in [43, 44]). *In vivo* studies began much later compared to studies of the dopaminergic system but still provide significant insight specifically with respect to the

effects of ketamine on brain function in healthy volunteers versus individuals with first episode psychosis with a diagnosis of schizophrenia [16]. SPECT studies showed a deficit in the left hippocampus in unmedicated patients with schizophrenia [45]. Proton Magnetic Resonance Spectroscopy Studies (1H-MRS) have been successfully used to quantify glutamate and glutamine levels in chronic schizophrenia (reviewed in [46]). The areas of the brain demonstrating evidence of involvement include the anterior cingulate cortex and the caudate nucleus, but further studies are needed to conclusively demonstrate that these observations are the result of disease pathology rather than treatment resistance (reviewed in [47, 48]). Limitations persist regarding the role of glutamate in schizophrenia with respect to 1H-MRS being unable to discriminate between intra and extracellular compartments, making it difficult to localize alterations [49]. It also remains unclear what exactly NMDA hypofunction means at a molecular level, and with no glutamatergic medications on the market for schizophrenia and no conclusive evidence from clinical trials of glutamatergic treatments, there is still much research to be done in order to elucidate the full details of glutamate hypothesis (reviewed in [16]).

Current research indicates that dysfunction of dopamine systems in schizophrenia may be the result of a decrease in NMDAR function, known as the NMDAR hypofunction hypothesis. Additionally, it has been demonstrated that dopamine has a regulatory role on glutamate performance, which in turn could affect NMDAR function. Limitations to this combined hypothesis include that specific brain circuits or regions have not been implicated and that dopamine changes have not been demonstrated to account for negative and cognitive symptoms (reviewed in [16]). In order to provide

further refinement to the NMDAR hypofunction hypothesis, GABAergic, opioid, cholinergic, and serotonergic systems have been investigated through treatment studies, animal models, and genetics [50-52].

Diagnosis

Phenotypic expression of the disease can vary significantly between affected individuals. The DSM-5 outlines multiple groups of symptoms that are indicative of schizophrenia: 1) delusions, 2) hallucinations, 3) disorganized speech, 4) disorganized/catatonic behavior, and 5) negative symptoms. Diagnosis of schizophrenia requires that two of these five symptoms be present for a minimum of one month, with at least one of the two symptoms being one of the first three symptoms listed [4]. Additionally, for a significant portion since the onset of symptoms, level of functioning in major areas, such as work, interpersonal relations, or self-care is demonstrably lower than prior to onset. Continuous elements of the disturbance must persist for at least six months (less if successfully treated), but do not need to include all of the elements originally observed. Substance abuse must be ruled out as the catalyst for the change in behavior. Other disorders must have been successfully ruled out, in particular, schizoaffective disorder and depressive or bipolar disorder with psychotic features, and if the patient was diagnosed with autism spectrum disorder or communication disorder, the additional diagnosis of schizophrenia requires that delusions or hallucinations be prominently featured [53]. These diseases share a significant portion of symptomology with schizophrenia, and therefore likely share some underlying neurobiology as well.

Distinguishing between these similar diseases is a necessary component if there is to be any hope of successful treatment [4, 54].

A key component to genetic studies and treatment is a reliable and valid diagnosis. Despite changing definitions, under DSM-IV criteria schizophrenia is reliably diagnosed, with 80-90% of those initially diagnosed with the disease retaining that diagnosis up to ten years later [55]. The DSM-IV construct has fair validity, as evidenced by precursors such as familial aggregation and environmental risk factors, as well as corroborating factors such as diagnostic stability, course of illness, and treatment response [56]. Overall schizophrenia as defined in the DSM-IV conveys useful clinical information resulting in that definition being largely retained in DSM-5. Most individuals diagnosed with schizophrenia under DSM-IV continue to meet the DSM-5 criteria, and those who did not meet the criteria in the earlier version, do not meet it in the new one [4]. Clinical manifestation can vary extensively and the DSM-IV clinical subtypes did not adequately address this variability. Subtypes also have low diagnostic stability, do not demonstrate specific patterns of treatment response, and are not heritable, ultimately leading to their removal under DSM-5 (reviewed in [57]). Psychopathological dimensions were introduced to better account for the phenotypic heterogeneity of schizophrenia with the aim to increase validity and clinical applicability, as well as to improve measurement-based treatment [58]. In the case of complex disease such as schizophrenia, symptomology may initially mimic other neuropsychiatric disorders, such as bipolar disorder, psychotic depression, or substance abuse. To more effectively counter these areas of overlap the DSM-5 specifiers of course of illness were introduced to allow clinicians the ability to document both the current status and the previous course

up to the present observation [54]. To assist efforts in determining etiology of complex disorders such as schizophrenia, linked conditions, including specific symptoms and traits, are often grouped together into spectrum disorders [59-62]. Under DSM-5 several different spectrum disorders share much of their symptomology with a diagnosis of schizophrenia, but important details delineate these diseases from the narrow definition of the disorder.

Schizoaffective disorder is a psychotic illness that shares many features with schizophrenia, but is marked by a predominating mood component. Schizoaffective disorder is usually diagnosed during the period of psychotic illness. Only the first set of criteria for schizophrenia must be met; two or more of the required five symptoms (delusions, hallucinations, disorganized speech, catatonic behavior, and negative symptoms), with one being from the first three. Symptoms that meet criteria for a major mood disorder must be present for most of the total duration of the acute and chronic periods of the overall illness and if these symptoms are depressive in nature, avolition and anhedonia must be ruled out. Delusions or hallucinations must be present for two or more weeks in the absence of a major mood episode in order to rule out depressive or bipolar disorder with psychotic features [53]. It is speculated that greater attention to the longitudinal course and increased specificity regarding mood disorder requirements compared to DSM-IV will lead to a reduction of individuals diagnosed with schizoaffective disorder [4].

Delusional disorder is a psychotic illness marked by delusions, which are also present in schizophrenia. A diagnosis of Delusional disorder therefore requires that the individual has never met the other diagnostic criteria for schizophrenia. If hallucinations

are present, they must not be prominent and must be related directly to the delusional theme. Functioning should not be demonstrably impaired, and the symptoms should not be attributed to substance abuse or medication. Additionally, if manic or depressive episodes have occurred, they have been brief in comparison to the delusional episodes. The criterion for delusional disorder has been further clarified to exclude body dysmorphic disorder and obsessive-compulsive disorder.

Brief psychotic disorder differs from schizophrenia in that diagnostic criteria only contains the first four symptoms outlined above (delusions, hallucinations, disorganized speech, and catatonic behavior), but not negative symptoms. It also features a sudden onset (from a nonpsychotic state to a clearly psychotic state in two weeks), and lasts at least 1 day, but less than one month. Afterwards, the individual eventually returns to normal level of functioning.

Schizophreniform disorder follows the same diagnostic criteria as schizophrenia, with the only difference being duration of at least one month, but less than six months. In order for an individual to receive this diagnosis, he or she must already have recovered from the symptoms and returned to normal functionality. If an individual is still symptomatic, but the duration is less than six months, this diagnosis is given under a provisional status until symptoms either resolve prior to six months elapsing. At that point if the patient is fully recovered the provisional status is removed, or if symptomology remains ongoing the individual is then diagnosed with schizophrenia.

Catatonia remains part of the diagnostic criteria for schizophrenia and other psychotic diagnoses, and it can now be indicated as a specifier for other psychiatric illnesses, or be diagnosed as unspecified if the comorbid disease is not apparent at the

time of evaluation [63]. Catatonia is determined if three of the following criteria are met: stupor, catalepsy, waxy flexibility, mutism, negativism, posturing, mannerism, stereotypy, agitation (not influenced by external stimulus), grimacing, echolalia, and echopraxia [53].

Schizotypal personality disorder, schizoid personality disorder, and paranoid personality disorder are all classified under personality disorders. Schizotypal personality disorder features a pattern of acute discomfort in close relationships, a reduced ability to maintain close relationships, cognitive/perceptual distortions of behavior, and is marked by eccentricity. Schizoid personality disorder features a repeated pattern of detachment from social relationships and a limited range of emotions expressed in interpersonal settings. Paranoid personality disorder features a pattern of distrust and suspiciousness. Individuals suffering from paranoid personality disorder often believe that others' motives are malevolent in nature. For any of the personality disorders to be diagnosed, symptoms cannot occur exclusively during the course of schizophrenia, or any other related psychiatric disorder, such as bipolar or depressive disorder with psychotic features, any other psychotic disorder, or autism spectrum disorder [53].

Treatment Options

Preventative measures are difficult to define when the etiology of schizophrenia remains largely unspecified. Though the disease is believed to be caused when the combination of genetic and environmental risk factors exceeds a certain threshold, without knowing what those risk factors are it is impossible to qualify or quantify what that threshold might be. The focus, at present, is on early identification of severe mental

health symptoms in the effort to allow for intervention in earlier stages of the disorder since it is believed that this may help avoid any lasting cognitive dysfunction [64].

Drug therapies have been the cornerstone of treatment for schizophrenia for half a century and have played an instrumental role in refining the underlying neurobiology. The common target of medications prescribed are dopamine receptors, but no particular type has been shown to be more effective than another, and as mentioned earlier, dopamine targeting does not address the full disease profile of schizophrenia, just the symptom of psychosis [65]. There has been proof of reduction in acute onset of positive symptoms (hallucinations, delusions, etc.). However, reduction of negative symptoms (alogia, avolition, etc.) is less effective [16, 52, 66, 67]. Current prescription treatment of schizophrenia includes first, second, and third generation antipsychotic drugs.

First generation antipsychotics (FGAs), often termed conventional or typical, are high-affinity antagonists of dopamine D2 receptors that have a high association with extrapyramidal symptoms (EPS), such as dystonia, akathisia (motor restlessness), parkinsonism (characteristic symptoms such as rigidity, bradykinesia, and tremor), and tardive dyskinesia (irregular movements), that paired with inefficacy often lead to discontinuation of treatment, either by the patient or by the clinician (reviewed in [68]). FGAs are classified as either low or high potency medications, corresponding to their affinity for D2 receptors [69]. On the basis of chemical structure, FGAs are subdivided into three distinct groups: butyrophenones (ex. haloperidol), phenothiazines (ex. chlorpromazine), and a heterogenous third group (reviewed in [70]).

Second generation antipsychotics (SGAs) and third generation antipsychotics (TGAs) are considered atypical, and are often grouped under the broad heading of new-

generation antipsychotics (NGAs). The first SGA, Clozapine, was found to be effective against psychosis without producing the side effect of EPS. It also demonstrated superiority to chlorpromazine in treatment-resistant schizophrenia. Unfortunately, it was also correlated to an increased risk of hematotoxicity, which can be fatal [71]. As a result, other drugs were developed to pursue the efficacy of clozapine without this side effect, such as risperidone, olanzapine, quetiapine, and ziprasidone, but termination rates remained high due to inefficacy and side effects associated with treatment [68]. Additional SGAs such as paliperidone, asenapine, iloperidone, and lurasidone and the TGA aripiprazole, have been approved by the United States Food and Drug Administration (USFDA), but comparative effectiveness to FGAs has yet to be conclusively determined [72].

Many FGAs carry EPS side effects that, depending on the patient, may rival the symptoms of schizophrenia. Though rare, these antipsychotic drugs can also cause neuroleptic malignant syndrome which carries a mortality rate of 20%. For these reasons, it is often suggested that maintenance doses be tapered down to the lowest therapeutic dose possible, and if multiple drugs are administered, that some be eliminated when the acute episode is over. NGAs, though lacking EPS side effects, are observed to carry an increased risk of weight gain, as well as disturbances in glucose and lipid metabolism, compared to FGAs (reviewed in [70]). Understanding the underlying genetic architecture of schizophrenia may allow for novel drug interventions to be created and may also facilitate identification of off-target medications (medications not intentionally designed for schizophrenia, but target the same pathways) [64, 73, 74].

Complex Genetic Architecture

Schizophrenia is a complex idiopathic neuropsychiatric illness that affects approximately 1% of the general population, but the risk of developing schizophrenia increases in proportion with the amount of DNA shared between an affected individual and his/her relative. The prevalence in third-degree relatives, such as first cousins, who share 12.5% of DNA, increases to 2%. The risk in second-degree relatives, such as half-siblings, who share 25% of their DNA, climbs to 6%. Most first-degree relatives, such as siblings or parents, share about 50% of their DNA, and carry a risk of 9%-13% (depending on the exact relationship between the affected individual and family member). Monozygotic twins share 100% of their DNA, and have a concordance rate of 50%. These increases in risk are present in the specific diagnosis of schizophrenia, but often relatives suffer from schizophrenia spectrum disorders or other neuropsychiatric disorders at higher rates than are seen for those diseases in the general population. Family studies demonstrating increased prevalence among relatives are not enough to conclude a genetic component to schizophrenia risk since related individuals share environments as well as genes. Twin and adoption studies, however, have shown that when biological children of patients with schizophrenia are adopted, and therefore grow up in an entirely different environment, they develop schizophrenia at the elevated rates that are seen in first-degree relatives in family studies. Estimates of the heritability of schizophrenia vary across studies but they have been demonstrated to range anywhere from 70% to as high as 86% (as reviewed in [75]). In some cases, spectrum disorders can

be as severe as the disease itself, but it is not always the case. Some spectrum disorders are mild, and do not involve psychosis. In the case of family studies investigating schizophrenia, adding a broad definition to include unspecified schizophrenia spectrum (previously nonaffective psychotic disorder under DSM-IV), schizotypal personality disorder, and paranoid personality disorder in first-degree relatives can provide additional analytic power, and produce superior results compared to labeling these individuals as unaffected [76]. Individuals diagnosed with unspecified schizophrenia spectrum exhibit the symptoms associated with the narrow definition of schizophrenia, but lack the quantity or severity to qualify for that diagnosis. These individuals also exhibit these symptoms for too long of a period to be diagnosed with Brief Psychotic Disorder. Personality disorders are broadly defined to include impairments in personality (self and interpersonal) and the presence of pathological personality traits. Schizotypal personality disorder and paranoid personality disorder exhibit both positive and negative symptomology that mimic the narrow definition of schizophrenia, but paranoid personality disorder has the primary feature of pervasive distrust and schizotypal personality disorder is marked by eccentricities of behavior [77].

Despite the high heritability, there are complexities underlying the assumption that schizophrenia is strictly a genetic disease, the most obvious being the lack of 100% concordance in monozygotic twins. Additionally, all first-degree relatives of individuals diagnosed with schizophrenia would exhibit consistent increased risk concomitant to their shared DNA, but instead parents carry a 6% risk, siblings 9%, children 13%, and dizygotic twins 17%. The lower risk for parents of affected children may be because affected parents exhibit lower fitness, and are therefore less likely to reproduce. Second

degree relatives would be expected to drop precipitously in terms of risk, but instead range from 2% for uncles/aunts up to 6% in half-siblings (the same as parents, who are first degree relatives and as a result share more DNA with the affected individual) [78]. Therefore, it has been widely accepted that a complex interaction between genetic and environmental factors is responsible for the etiology and overall development of schizophrenia.

Schizophrenia susceptibility is likely linked to multiple genetic factors, as evidenced by the fact that patterns of transmission do not match established Mendelian inheritance patterns of single locus disorders, as well as mounting support for a polygenic component [67], [79]. This complexity, along with phenotypic variation, explains why the search for 'schizophrenia genes' remains ongoing to present day. Conflicting evidence continues to accumulate, with candidate genes being identified in some studies, and later questioned or disputed in others (reviewed in [80, 81]). It remains clear, however, that support is present for many different genetic factors playing a role in the predisposition to schizophrenia, including microdeletions (such as is present in the disease 22q11.2 Deletion Syndrome), microduplications, single nucleotide polymorphisms (SNPs), and copy number variations (CNVs) [82, 83].

CNVs are defined as a gain or loss of a segment of DNA greater than 1 kilobase in size, but in rare cases can be more than 100kb in size. Depending on size and location, CNVs can affect multiple genes and/or regulatory regions (reviewed in [84]). Support for the hypothesis that CNVs are a significant genetic risk factor has accumulated since 2008, with recurrent CNVs identified and replicated at the following locations: 1q21.1, 3q29, 7q11.2, 15q11.2, 15q11.2-13.1, 15q13.3, 16p11.2, 16p13, 17p12, 17q12, and

22q11 (reviewed in [85-87]). Due to shared symptomology between many neurological disorders, these same CNVs produce significant results in other disorders. The CNV located at 16p11.2 was found in bipolar affective disorder [88]. CNVs at 1q21.1, 3q29, 7q11.23, 16p11.2, 15q11.2-13, and 22q11.2 have been found and confirmed in subjects with autism spectrum disorder [89]. A nominal association was found between Alzheimer's Disease and a duplication of 15q11.2 [90]. As discussed in the Diagnosis section of this chapter, many neuropsychiatric diseases share a vast measure of symptomology, and therefore finding shared genetic architecture is not unexpected.

Identification of causal variants in complex disorders, such as schizophrenia and other neurological disorders, can be difficult due to the fact that the underlying biology is not completely understood at present. As such, methods such as linkage analysis and genome-wide association studies (GWAS) can be powerful tools for localizing genetic susceptibility to an inherited disease or symptom to particular regions of chromosomes or genes [91, 92]. Both linkage analysis and GWAS are discussed in greater detail in Chapter 2.

The premise of linkage analysis is based on the observation that regions of the genome that reside physically close remain linked and transmitted together during meiosis. Alfred Sturtevant developed the first linkage map in 1913 while working on *Drosophila* under Thomas Morgan. The first linkage study in psychiatry was published in 1969 [93]. Since that time many individual studies and meta-analyses have produced significant evidence of linkage between schizophrenia, bipolar disorder, or psychosis to regions of every autosome [94-107]. Studies that used very large numbers of pedigrees (>400) for linkage analysis in these same disorders failed to produce statistically

significant results [108-110]. This may indicate that larger sample sizes are not necessarily better. With larger samples it can be more difficult to control for genetic background and phenotypic variation, especially if one is only considering the narrow diagnosis of schizophrenia and ignoring spectrum disorders present in nonpsychotic first-degree relatives.

Another powerful method for causal gene detection in complex disorders is Genome-Wide Association Studies (GWAS). GWAS compares genetic variants among individuals with varying phenotypes of either a trait or disease, but individuals do not have to have any known relationship. For association analysis to work, individuals being assessed must share a common ancestor, so while although a known relationship is not required, genetically homogenous samples are more likely to produce significant results. GWAS can provide insights where linkage analysis fails to for common disorders because the underlying genetic mechanisms differ from rare disorders, which tend to have consistent results across various linkage studies [92, 111]. Despite this, GWAS studies of schizophrenia have produced inconsistent results, similar to linkage analysis.

Some GWAS have reached genome-wide significance for a very limited number of loci, whereas others have failed to produce significant results despite extremely large sample sizes [79, 112-118]. Gene-wide analysis of two European ancestry GWAS datasets, one schizophrenia (479 cases and 2,937 controls) and one bipolar disorder (1,868 cases and 2,938 controls), demonstrated evidence for association across disorders to genes reported in other datasets: *CACNA1C, CSF2RB,* and *DGK1* was observed for both disorders [114]. Meta-analysis of 7,308 schizophrenia cases and 12,834 controls of European ancestry demonstrated strong evidence of association to *ZNF804A*, and was

additionally strengthened by inclusion of bipolar disorder to the affected phenotype [119]. A two-stage GWAS of schizophrenia in the Han Chinese (stage 1: 4,384 cases and 5,770 controls; stage 2: 4,339 cases and 7,043 controls) demonstrated genome-wide significant associations in an exon of *VRK2*, and exon of *GABBR1*, and an intron of *ARL3* [120]. A GWAS of schizophrenia using 871 cases and 863 controls failed to produce significant findings, nor could it reproduce findings from four independent European cohorts comprised of 1,460 cases and 12,995 controls [115]. A GWAS of the Molecular Genetics in Schizophrenia (MGS) European case-control sample (2,681 cases and 2,653 controls) failed to achieve genome-wide significance, but a meta-analysis of European-ancestry subjects (8,008 cases and 19,077 controls) demonstrated a significant association with schizophrenia in a region of linkage disequilibrium on 6p22.1 [116]. GWAS of schizophrenia (738 cases and 733 controls) in a United States population did not provide any evidence of involvement for any genomic region with schizophrenia [118].

Following mixed results in GWAS concerning schizophrenia, the Psychiatric Genetics Consortium (PGC) conducted a large GWAS in 2014 with 36,989 cases and 113,075 controls from 49 studies (46 of European ancestry and three of Asian ancestry). 128 independent association signals across 108 distinct loci achieved genome-wide significance. 83 of those associations have not been previously identified. These findings were supported by the fact that many were within genes expressed in the brain, as well as previously identified genes such as DRD2 and others known to be involved in glutamatergic pathways consistent with current hypotheses for the neurobiology underlying schizophrenia. Additionally, association signals were detected in genes expressed in tissues that play an important role in immunity, adding to growing evidence

for a link between the immune system and schizophrenia [83]. Similar to linkage analysis of large samples, inconsistency with respect to producing significant results could be due to factors such as small effect size and uncontrolled phenotypic variation. The PGC study may have succeeded because it was sufficiently large to overcome these factors [121].

A Hypothesis for Determining Risk Alleles in Schizophrenia

One interpretation of the inconsistent results observed in both linkage and association analyses is that due to factors such as small effect size and uncontrolled phenotypic variation, larger and larger samples will need to be recruited in order for additional susceptibility genes to be discovered [121]. The work performed by the PGC described above demonstrates that this is one potential approach to solving the problem. This thesis focuses on a different interpretation: genuine genetic differences between definable subsets can mask both linkage and association, and that this problem can be worsened in studies that use large samples where the entire sample is analyzed as if it were a genetically homogenous group. We believe that by correcting for this issue and by focusing analysis on regions where linkage and association overlap, which indicates that disease susceptibility in those regions is driving both signals, we can leverage whole genome sequencing to identify risk alleles for susceptibility to schizophrenia using a substantially smaller sample.

The work presented herein begins with 22 medium-sized Canadian families, originally selected for study due to the fact that multiple relatives were clinically diagnosed with schizophrenia. Over time, the sample has grown to include 573

individuals across 30 large Canadian pedigrees of German or Celtic descent initially recruited if at least three subjects with schizophrenia were available for study, the disease appeared to be segregating in a unilineal autosomal dominant pattern with the aim of reducing the number of risk alleles, and minimal severe affective disease to ensure that psychosis results from schizophrenia as opposed to bipolar disorder. Two main advantages to this sample are that these subjects have been followed for up to 20 years allowing for continued observation of diagnostic stability, and the pedigrees are large in size making it possible for the sample to demonstrate statistically significant linkage and association results. Additionally, all individuals were adults at the time of recruitment, past the typical age of onset for schizophrenia, making it unlikely that affected status would change at a later date.

This thesis attempts to identify risk alleles for susceptibility to schizophrenia with the hope of contributing to the elucidation of the molecular pathways involved in the underlying neurobiology of this complex neurological disorder. We believe that greater understanding will lead to objective diagnostic methods and improved treatment options including personalized medicine. Chapter 2 describes the history and methodology of both linkage and association analyses. It places particular emphasis on how different traditional methods work and why they may fail to ascertain statistically significant results. Chapter 2 concludes by describing a novel statistical method for both linkage and association analyses that incorporates the most powerful methodologies and discards those aspects most likely to cause problems in complex disorders such as schizophrenia, with the added benefit of reducing signal-to-noise ratio over traditional methods. Chapter 3 discusses the details of our well-characterized sample, nearly two decades of

preliminary findings tied to that sample, and how those data were leveraged to form a series of analyses to uncover new candidates for susceptibility to schizophrenia. Chapter 3 also covers the full bioinformatics analysis of whole genome sequencing performed on a portion of the sample. Chapter 4 covers harmonization of the data, including the steps necessary to include later individuals to the sample and the subsequent re-analysis of all of the preliminary data described in Chapter 3. Chapter 5 describes the identification of risk SNPs (n=12) from areas where statistically significant linkage and association signals overlap, and how those SNPs were used to search flanking regions for novel candidate variants. It further describes how candidate pools were subjected to a filtering pipeline in order to prune down to 101 SNPs most likely to play a casual role in the etiology of schizophrenia for genotyping. Finally, it describes the results of a comparative analysis between the risk SNPs and the selected candidate SNPs, which shows six SNPs in strong LD with rs7419214 that scored higher than rs7419214 on the PPLD|L and are candidates for further evaluation for a causal role in susceptibility to schizophrenia. Lastly, Chapter 6 describes future directions for the work presented herein.

Chapter 2: Linkage and Association

Linkage

Ever since Gregor Mendel discovered the basic laws of inheritance by studying thousands of pea plants in the mid-1800s, the identification of the underlying genetics responsible for disease, as well as the variation in quantitative traits, such as height, has been the foundation of human and medical genetics. Mendel conducted studies for nearly a decade, growing more than 10,000 pea plants, before publishing his observations in 1865. Mendel's Laws of Heredity remain a staple of biology education more than a century after his death. When Mendel made his observations regarding the independent assortment of traits, he happened to select characteristics that were not located on the same chromosome, otherwise, he may have drawn different conclusions from his experiments. Later studies showed that many genes are linked, and that in those cases, the traits encoded by those genes do not sort or segregate independently, but are instead inherited together [91].

Linkage analysis is a powerful method for localizing genetic susceptibility to a shared disease or trait to regions of the human genome. Linkage is the tendency of two or more loci that are physically close on a chromosome to be transmitted together from parents to offspring during meiosis, in violation of Mendel's Law of Independent Assortment. Linkage analysis was developed to detect this excess co-segregation of alleles underlying a phenotype or trait with the alleles at a marker locus in families. In practice, linkage analysis pertains to a group of statistical methods that when used allow for a gene to be mapped to the chromosome region in which it is located. Since there are

many more genes than there are chromosomes, genes are often transmitted together [122]. During meiosis, a pairing of duplicated homologous chromosomes occurs and a physical exchange of material occurs between them. These exchanges, called chiasmata, lead to a 'crossover' of DNA between the two homologues. Though these exchanges are frequent, the presence of one decreases the chances of another occurring close by. This phenomenon, known as interference, makes it unlikely for double crossovers to occur when two locus are proximate to each other [123]. Therefore, the probability that crossovers will occur between two loci on the same chromosome is dependent on the distance between them [124, 125]. When an odd number of crossovers between two loci occurs, it can be observed by analyzing the genotypes of the parents and the offspring. In this case, the alleles at these loci are transmitted to the offspring in a new combination. When an even number of crossover events occurs, the resulting genotype of the offspring is comprised of the original alleles for each loci. Therefore, two loci that are very far apart on the same chromosome experience observable recombination 50% of the time, giving the appearance of independent assortment (reviewed in [126]).

The recombination fraction, often represented as θ, measures the ratio of recombination events detected between two loci in a group of offspring. It is estimated by counting the number of offspring who show recombination for a given pair of loci, divided by the total number of offspring. Recombination fractions range from 0 to 0.5, with values <0.5 indicating some degree of linkage. Recombination fractions are often converted into map distances in the unit Morgan. The centiMorgan is a unit of recombinant frequency that implies distance along a chromosome taking into account how often recombination occurs within a given region. For small values of θ, map

distance is approximately equal to recombination fraction, but for larger genetic distances, for which all distances have $\theta \approx 0.5$, mapping functions are used. The two most common mapping functions are Kosambi and Haldane, which are based in different formulas, with most genetic maps using Kosambi cM. The Haldane function does not allow for genetic interference (described above), while the Kosambi function models interference.

Once the detection of DNA polymorphisms became possible the study of genetic linkage flourished because analyses were no longer limited to the comparatively rare protein polymorphisms [127]. Restriction fragment length polymorphisms (RFLPs) were first used as a tool for genetic analysis in 1974 when linkage of temperature-sensitive mutations of adenovirus were utilized to locate mutations on a physical map, which measures distances in DNA base pairs [128]. Initially, maps were drawn by hand using RFLPs. This changed following the advent of polymerase chain reaction (PCR) in 1983 and its commercial availability beginning in 1987, which ultimately led to the discovery of a novel class of short tandem repeat (STR) polymorphisms. STRs are dinucleotide, trinucleotide, or tetranucleotide repeats that are multiallelic, typically supplying sufficient heterozygosity permitting maternal and paternal contributions to be distinguished from one another [129]. The discovery of STRs made it possible to create a large number of markers and genetic linkage analysis led to the generation of a genetic map, which measures distance using the centiMorgan (cM).

STRs, also known as microsatellites or MSATs, have limited ability to be scaled up to high-throughput typing because electrophoretic separation must be conducted to properly determine fragment sizes. MSATs occupy 3% of the human genome. Tri- and

hexa-nucleotide repeats have greater abundance in exons, whereas other repeats are more abundant in non-coding regions, making it difficult to construct a map with markers of even genomic distribution while maintaining high heterozygosity (reviewed in [130]). By the late 1990s researchers turned their attention towards SNPs, which are an excellent candidate for comparatively lower cost very-high-throughput genotyping, due to their abundant nature in the genomes of humans and many other organisms. SNPs carry a maximum heterozygosity of 0.5 (due to their biallelic nature), and therefore lack the informativeness of STRs. However, SNPs carry several benefits in addition to abundance including: global genomic distribution, and perhaps most importantly, the adaptability to massively parallel genotyping allowing for denser maps with more markers. The information content (IC) for an individual SNP is based on its minor allele frequency (MAF), the frequency of the less common allele. The higher the value, the more informative the SNP. Denser maps allow SNPs to surpass the IC available from STRs, which are typically spaced ~10 cM apart due to their inability to be scaled up to high throughput [131].

As genotyping technology has evolved so too has the search for an efficient computational algorithm to calculate the evidence for linkage between two loci. Recursive analysis on a simple extended pedigree was pioneered by Elston and Stewart in 1971, and is commonly referred to as the Elston-Stewart algorithm. It allows for decreased penetrance and quantitative traits [132]. Under the Elston-Stewart model the basic assumption is that the phenotype of each individual only depends on its own genotype (reviewed in [133]). Expansion by others has led to its ability to handle more complex data structures [134-136]. Computational time for the Elston-Stewart algorithm

can be prohibitive in multipoint linkage analyses because while it scales linearly with the number of meiosis, it scales exponentially with the number of marker loci, such that large pedigrees cannot be evaluated with a large number of markers (reviewed in [126]). The next major development was made by Lander and Green in 1987. Referred to as the Lander-Green algorithm, it is able to rapidly compute maximum-likelihood multi-locus linkage [137, 138]. For this algorithm, the computational time scales linearly with the number of markers, but exponentially with meiosis, and so it is not suitable for use with large pedigrees. Neither the Elston-Stewart, nor the Lander-Green algorithms are computationally feasible for large extended pedigrees and dense marker maps, such as those that would be needed when using SNPs rather than STRs. Large pedigrees (>25 individuals) are of great value in linkage analysis because these pedigrees are capable of producing strong evidence of linkage on their own with the greatest chance of being genetically homogenous. Trimming pedigrees to circumvent the limitations present in the Lander-Green algorithm can lead to reduction in power, loss of information, and erroneous results. Two commonly employed approaches to bypassing the computational problems presented by large pedigrees are 1) to utilize statistical methods that avoid computation of the full pedigree likelihood (ex. variance-components), and 2) to use Markov chain Monte Carlo (MCMC). MCMC supports utilization of the full likelihood, but complexities arise in the optimization of performance of samplers, which in turn limits the adaptability in handling the trait model (reviewed in [139]).

Linkage analysis in humans carries challenges not present in experimental organisms such as family size, the inability to do test crosses, significantly longer generation times, and inability to discern the parentage of alleles when both parents are

identically heterozygous and have the same genotype as their offspring at the locus being studied. Many approaches have been utilized to query directly or indirectly for lower than expected observed recombination between two loci. These statistical methods fall into two basic categories: parametric and non-parametric [126].

Parametric Linkage Analysis

Parametric (or model-based) linkage analysis, as the name suggests, requires the specification of parameters, which must be known at the time of analysis. For qualitative traits assumed values must be provided for allele frequencies at the trait and marker loci, and penetrance (the relationships between genotypes and phenotypes). For quantitative traits assumed values must be provided for allele frequencies at the trait and marker loci, the means and variances of the phenotype for each genotype. Definitive recombinants can only be defined for qualitative parametric linkage analysis. This is due to the fact that normal probability densities are used to model the genotypic distributions in quantitative linkage analysis. As these densities asymptotically approach zero in both tails, but never reach it, every individual has a non-zero probability for having each genotype. In order to mitigate this problem, which effects the identification of recombination events that assist in the localization of candidate regions, methods have been developed over time to classify individuals based on their most likely genotype [140].

One measure for the likelihood of linkage is the logarithm of the odds score (LOD). LOD score analysis is parametric, and as such requires the assumption of precise genetic models, including penetrance, disease gene frequency, and affection status for the

individuals being tested. The LOD score Z is the logarithm of the odds that the loci are linked divided by the odds that the loci are not linked. Expression of the likelihood as a logarithm allows for summation of the likelihood of linkage observed across different pedigrees [141]. Because the true genetic distance between two loci is often unknown, the LOD score is calculated for several recombination fractions and from there a maximum likelihood estimate (MLE) for the recombination fraction (θ_{max}) at which the greatest LOD score (Z_{max}) is observed can be made. For families in which all of the necessary information is known, a LOD score calculation can be done by hand, but for complex datasets computer programs are, in reality, a necessity to produce timely results [142]. A LOD score of 3.0 is required for evidence of linkage, with a 5% chance of Type I error, which would correspond to a p-value = 0.05. A LOD score of 3.0 corresponds to 1000:1 odds in support of linkage. Because it is improbable that two loci, chosen at random, would be linked, a rigorous standard is applied to demonstrate evidence. Humans have 22 pairs of autosomes, making it unlikely for two randomly chosen loci to be present on the same chromosome, and they would also need to be physically close to one another in order to be linked. The likelihood that two randomly selected loci should be linked (known as the prior probability of linkage) has been debated, but estimates of 1 in 50 are generally accepted. Therefore, using Bayesian calculations, if one multiplies the prior probability (1/50) by the conditional probability (LOD score of 3.0, 1000:1 odds in favor of linkage), the result is a joint probability of 20. Odds of 20:1 correspond to the conventional threshold of statistical significance, p = 0.05. A LOD score below -2 is accepted as evidence against linkage [126].

Locus heterogeneity, where alleles at more than one locus lead to the same phenotype, occurs in complex traits and can have a negative effect on the power of linkage analysis if not considered at the time of analysis [143]. Smith first proposed the mixture model in 1963 [144]. Under this model's framework there are two approaches that may be employed; one may test for homogeneity given linkage or test for linkage allowing for heterogeneity by a likelihood ratio test [145, 146]. The LOD score calculated under the hypothesis of heterogeneity is called a heterogeneity LOD (HLOD), and it is never lower than the LOD score. As such, the threshold for significance for a HLOD score is greater than for a LOD score, with the consensus on that threshold being 3.3 [147]. HLOD scores are particularly useful in a sample comprised of a mixture of families, some linked to a given locus, and others not. 'Unlinked families' is an inclusive term that refers to families linked to a different locus, families misdiagnosed with the disease being investigated, and/or families that contain phenocopies of the disease in question. Using homogeneity LOD scores evidence for linkage may be overlooked because unlinked families will have negative LOD scores, and linked families will have positive LOD score, which may lead to the result failing to meet the threshold for significance when the two are summed. Heterogeneity can also cause a distortion in the estimate of theta calculated under the assumption of homogeneity which can lead to a significant LOD score being reported at an incorrect position. The parameter α is the proportion of families linked to a specific locus. In complex disorders fixed genetic parameters may be unknown and this can lead to inaccurate estimates of α, but overall the HLOD will still be a more useful statistic than the homogeneity or simple LOD score in the analysis of complex disorders [148].

MOD scores were later introduced to address the case of linkage studies for diseases with an unknown mode of inheritance. This analysis method uses variables for both recombination fraction and disease model parameters. Maximizing the MOD score function over all parameters is mathematically equivalent to maximizing the probability of marker data conditional on the affection status (reviewed in [149]). When there is no linkage, the MOD score adheres to a chi-square distribution, with greater degrees of freedom compared to the LOD score [150]. By ignoring the information produced by disease segregation and linkage disequilibrium between marker alleles and functional disease alleles, mod scores are a weaker tool to distinguish between genetic models [151].

Non-Parametric Linkage Analysis

Non-parametric or model-free linkage methods do not require the specification of parameters for the disease inheritance or disease allele frequencies and disease genotypes. Being able to search for evidence of linkage without knowing the mode of inheritance is of particular use when examining complex diseases where it is often unclear. This is especially important because parametric linkage can produce erroneous results when the linkage model is incorrect specified [152]. These methods are based on assessing whether relatives with similar trait phenotypes are also more genotypically similar than expected at a specific marker [126]. Early methods based on Penrose's affected-sib-pair (ASP) test for qualitative and quantitative traits focused on when two individuals share an allele at a specific locus and required pedigrees to be broken up into nuclear families which resulted in the waste of a lot of inheritance information contained in the pedigree

structure [153]. In order to rectify this, Weeks and Lange developed the affected-pedigree-member method (APM) [154, 155]. APM is not truly a linkage method because it does not trace the inheritance pattern within a pedigree, but instead simply focuses on whether two alleles are shared at the same locus. This concept, known as identical-by-state (IBS) does not guarantee that the shared allele is inherited from a common ancestor as defined by Wright's work on the coefficients of relationships termed identical-by-descent (IBD) (reviewed in [156]). The APM approach does not take into account genotype information from additional members in a pedigree in order to distinguish between IBS and IBD, and by failing to extract full inheritance information, it is subject to inaccurate results (reviewed in [138]). Later methods took full advantage of pedigree information in order to discriminate between IBS and IBD [138, 157].

Nonparametric linkage analysis ultimately seeks to determine whether relatives share more marker alleles IBD than expected under the null hypothesis (no linkage). Though several IBD statistics have been suggested, two of the most commonly employed are T_{pairs} and T_{all} [154, 158-161]. T_{pairs} reflects IBD sharing in pairs, which counts pairwise allele sharing among affected relatives. T_{all} represents IBD sharing in larger sets, which in turn increases statistical power by considering larger sets of affected relatives [138]. These two methods have been compared in simulation studies with evidence that T_{all} has greater power than T_{pairs} when assessing linkage in dominant and additive disease in nuclear families, and the reverse being true for recessive diseases [158, 160]. Later studies replicated these findings in three-generation pedigrees [138, 161, 162].

Posterior Probability of Linkage (PPL)

Both parametric and nonparametric methods have their merits. Parametric linkage analysis is decidedly more powerful and easily applied to diseases caused by single and usually rare variants [138]. Complex diseases such as schizophrenia are believed to be caused by multiple variants, at least some of which may be common, and therefore dependence on a specific model can greatly inhibit discovery. Furthermore, attempts to try all models may be too time-consuming and multiple statistical corrections may be needed to interpret results. In these cases, non-parametric linkage analysis may yield better results, but will lack the power of parametric linkage studies since it only uses a portion of available data [163].

The issue that remains with both parametric and nonparametric linkage methods is that neither directly measures the probability that there is linkage, called the posterior probability of linkage (PPL), which is the precise reason why these analyses are performed [164]. Smith first proposed a Bayesian approach to linkage analysis in 1959, but the idea never gained traction despite that probability has a direct meaning, and does not need to be qualified like a significance level [165]. The PPL carries the additional benefit of being model-free, which means that no parameters (allele frequencies, penetrance, admixture) are required to be specified in order to assess for linkage. The parameters that are required for parametric linkage calculations are integrated out of likelihood removing the need to correct for multiple testing as is the case with methods that use maximization. Though seemingly similar to nonparametric linkage analysis, the PPL allows for all available data to be used, making its power to detect similar to parametric linkage analysis [164, 166].

The PPL framework is implemented in the software package Kelvin and was created with the goal of accumulating evidence both for and against linkage. Results are reported on the probability scale and are interpreted as the probability of a trait being linked to the given locus [167]. The PPL assumes a prior probability of linkage in the absence of data of 2%, based on the number and length of human chromosomes [168]. The 2% value becomes the "prior probability of linkage" when no data is available, meaning that there is a 2% chance of linkage between a specific disease and a random marker. Once any amount of data has been evaluated, the "posterior probability of linkage" can be used to leverage the conditional probability based on prior evidence and allows for additional information to be evaluated with no need to correct for multiple testing. When new data becomes available, the "posterior probability of linkage" becomes the new "prior probability of linkage", replacing the 2% value with one more illustrative of the existing data. This allows for evidence either for, or against, linkage to be accumulated across several subsets of pedigree data that may have somewhat different inheritance model parameters (e.g. different values of alpha) that are necessary to perform computations, but are not otherwise important [164, 167].

Kelvin supports two-point (trait-marker or marker-marker) and multipoint linkage analysis based on either sex-averaged or sex-specific genetic maps, and includes the option to allow for imprinting. It has the ability to handle dichotomous trait, quantitative trait, and quantitative trait threshold models, as well as specific types of gene-gene interactions and covariate effects. The original version of Kelvin employs the Elston-Stewart algorithm, which permits analysis of large pedigrees with the ability to handle loops. Multipoint analyses are performed automatically by traversing down each

chromosome following a user-defined number of markers in each calculation performed at user-specified intervals. Kelvin is able to handle mixtures of varying pedigree structures including cases/controls, trios, sib-pairs, nuclear families, and extended pedigrees. It is accompanied by the custom graphing program Kelviz, which allows for results generated by Kelvin to be visualized. Not only does the PPL avoid the issues with maximization, it also offers a better signal-to-noise ratio for observing linkage (**Figure 1**) [169].

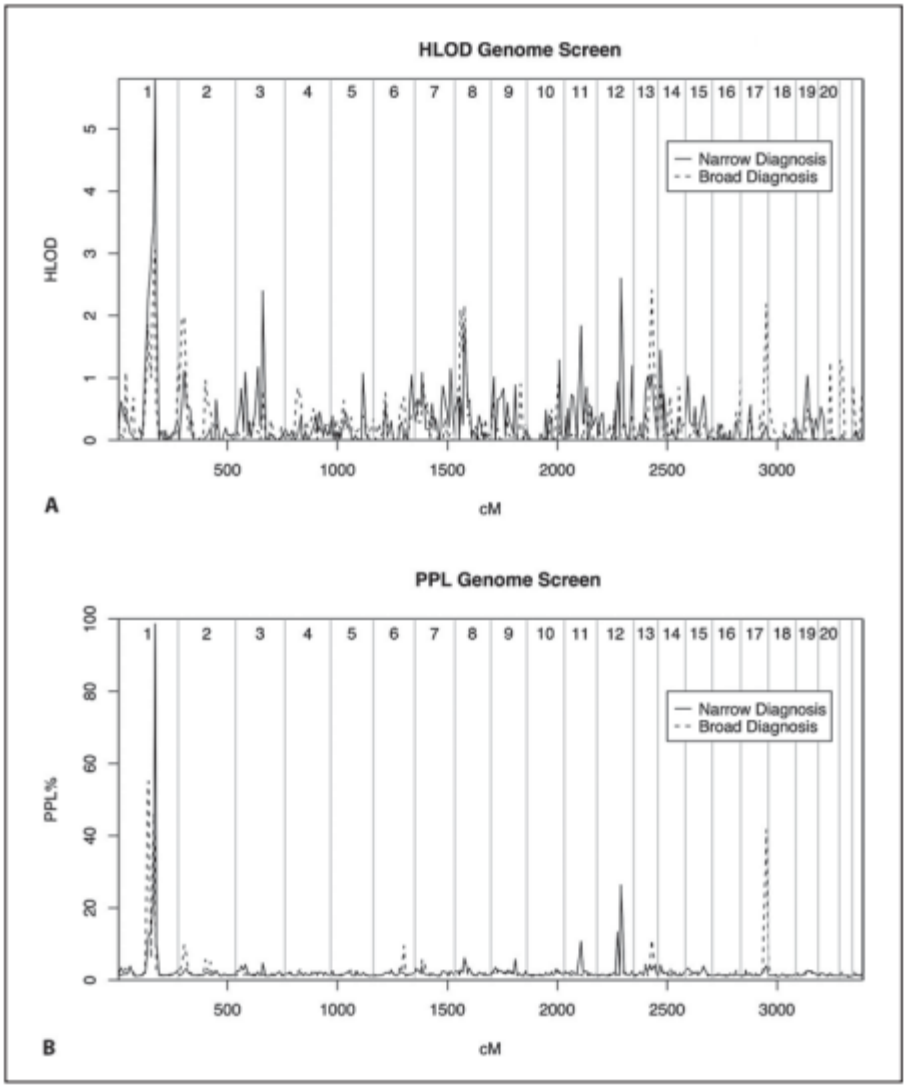

Figure 1: Comparison between HLOD linkage analysis and the PPL (originally published in [169]).

Figure 1 illustrates a comparison between HLOD linkage analysis conducted in 2000 [106] and re-analysis of the same data performed using the PPL in 2005 [169]. There is a significantly improved signal-to-noise ratio using the PPL without loss of statistically significant linkage signals detected by using HLOD analysis.

As mentioned earlier, when high-throughput genotyping techniques became affordable denser marker maps composed of SNPs became ubiquitous. This evolution brought new attention to the issue of how to handle large pedigrees along with the dense marker maps needed for SNPs to match/exceed the information content (IC) of STRs [170]. But, the standard algorithms employed in recursive linkage analysis are unable to simultaneously handle both large pedigrees and dense marker sets. The Elston-Stewart algorithm scales linearly with meioses, but exponentially with the number of markers. The Lander-Green algorithm scales linearly with the number of markers, but exponentially with the number of meioses. As discussed earlier, one approach to circumventing this issue is to use Markov chain Monte Carlo (MCMC) methods, which supports use of the full likelihood. The issue with this approach is that samplers tend to limit the flexibility in handling the trait model. This presents a specific challenge to adaption of MCMC in the context of Kelvin, which handles the trait model by integrating trait parameters out of the likelihood. This flexibility permits new trait models or additional trait parameters to be added to the calculations with ease [139]. The difficulty lies in the order of operations. Kelvin averages the likelihood ratio across pedigrees, calculating at one position at a time as it traverses the chromosome [167]. MCMC performs calculations on a per-pedigree basis for each chromosome, one at a time, in its entirety. The likelihoods are then averaged across iterations. This conflict requires both processes to be redefined with careful tracking at each step and then reconstructed so that first, repeated MCMC marker-sample generation for each pedigree across the chromosome is performed, and then repeated (adaptive) trait-space sampling across

pedigrees at each position on each chromosome takes place, conditioned on the marker data from the MCMC runs and the trait data [139].

Ultimately, the newest version of Kelvin takes the approach of combining marker data generated by MCMC with the trait-model integration implemented in Kelvin. To do this, the graphical-model-based MCMC approach of Thomas et al. for the marker data is combined with the flexible numerical integration algorithm of Seok et al. for the trait data [171, 172]. This allows for the power of MCMC to be employed in the context of the PPL framework [139, 164, 167].

Association

Genome-wide association studies (GWAS) measure and analyze DNA sequence variants across the human genome with the intent of identification of genetic risk factors for common, complex diseases such as schizophrenia. The eventual goal is to use these risk factors in a predictive manner to ascertain who is at risk and elucidate the underlying biology responsible for disease susceptibility. This information will play a crucial role in developing new prevention and treatment strategies [92]. As mentioned earlier, association analyses depend on the presence of common ancestral variants in the individuals being assessed.

One of the early successes of GWAS was the identification of the *Complement Factor H* gene as a major risk factor for age-related macular degeneration. GWAS not only identified the DNA sequence variations in the gene associated with the disease, but also the underlying biological basis responsible for the effect, opening the door for the development of new pharmacological interventions [173-175]. Understanding the

complexities of human disease is not the sole focus of human genetics, and as such techniques are often applied to other disciplines. One of the most successful of these applications is that of GWAS to pharmacology. Pharmacogenetics seeks to identify DNA sequence variations that are associated with drug metabolism and efficacy, as well potential harmful side effects resulting from administration of medications. GWAS uncovered variations in several genes that have a substantial influence on the dosing of warfarin, a blood-thinning medication that prevents the formation of clots [176]. This discovery (and subsequent validations that followed) paved the way for genetic tests that allowed clinicians to tailor warfarin dosage by individual, in order to maximize efficacy and minimum adverse side effects. This type of individualized treatment represents the field of 'personalized medicine', which aims to leverage an individual's genetic background and biological features in the establishment of a treatment plan with a greater likelihood of success compared to more generalized approaches (reviewed in [92]).

Rare genetic disorders can be caused by multiple different genetic variants within a single gene and because the effect is so strong inheritance can often be deduced by studying the inheritance pattern in families affected by them. Linkage analysis fares well in rare diseases such as cystic fibrosis and Huntington's Disease, where genetic markers segregate with the disease across multiple families [177, 178]. Conversely, linkage analysis has not performed as well when applied to common disorders, such as heart disease or cancer. The implication is that the genetic mechanisms responsible for common disorders differ from those that cause rare disorders [111]. Further support for this idea came from the identification of disease susceptibility SNPs for common diseases with high minor allele frequencies [179, 180]. This ultimately led to the common

disease/common variant hypothesis, which states that common disorders/diseases are likely caused by genetic variation that is common in the population [181]. Two major concepts tied to this hypothesis are: 1) if common genetic variants increase susceptibility to a disease, the effect size (penetrance) for an individual variant must be low compared to rare disorders, and 2) given the first condition (if penetrance is low), and heritability is high, then multiple variants must be necessary to increase disease susceptibility (reviewed in [92]).

Linkage Disequilibrium

Association is ascertained by evaluating linkage disequilibrium (LD), which is a property of SNPs within the same genome region that represents how often an allele of one SNP is inherited or correlated with an allele of another SNP. It is similar to linkage, where two markers on a chromosome remain physically joined through generations of a family, except that LD refers to an entire population instead of a single family/pedigree. The opposite of LD, linkage equilibrium, is the result of recombination events over many generations that break apart regions of a chromosome until eventually all combinations of alleles in a given population are inherited together in the ratio expected by chance.

The rate of LD decay depends on several factors, such as population size, the number of founding chromosomes in the population, and the number of generations that the population has existed. Different populations therefore have different patterns of LD, making population matching a critical component when assessing LD. African populations are the oldest, and therefore have smaller regions of LD due to an aggregation of recombination events. European and Asian populations split off from the

African population approximately 100,000 years ago, taking only a small subset of the available genetic variation with them. These populations were created by founder events originating from the African population, which in turn changed all of the factors that affect the rate of LD decay. These populations tend to have larger regions of LD compared to those of African descent [182].

There are many ways to measure LD and all of them are derived from the difference between the observed frequency of co-occurrence for two alleles and the expected frequency if the alleles are independent [183, 184]. The two most commonly employed measurements are r^2 and D' [182-184]. Values for D' range from 0 to 1, with 0 representing linkage equilibrium, and 1 indicating 'complete LD' (which represents no recombination between the two markers within a given population). In genetic analysis LD is usually reported as r^2, which is the statistical measure reflecting the correlation between two SNPs and is also measured on a scale of 0 to 1. The two measures are therefore related in the following manner: r^2 takes into account the allele frequencies for each SNP and can only be high when D' is also high. When r^2 is 1 two SNPs are said to be in 'perfect LD' and genotyping for one SNP will provide full genotype information for the second SNP.

Due to the related nature of linkage and association, the presence of LD between markers can lead to false positive results when assessing linkage. This is of great concern as linkage analysis shifts away from STRs and is performed using SNPs. SNPs are a common source of genetic variation and LD between SNPs, especially those in close proximity on the genome, is not uncommon. Inter-marker LD should therefore be assessed prior to conducting linkage studies using SNPs, with D' being a better indicator

of inflation when allele frequencies are similar, and r^2 being a better predictor when they are disparate [185].

Testing for Association

When LD is detected it still must be determined if the allele producing the signal is causal. In such cases the SNP producing the LD signal is responsible for the trait being assessed, meaning that the phenotype in some manner of statistical measurement is the result of that particular genotype. To assess whether this is the case, additional studies must be conducted to determine if the SNP is functional with respect to the observed phenotype, such as a luciferase assay to assess a sequence for effects on the regulation of gene expression [186]. In a carefully constructed study that properly accounts for population stratification, if the allele is not causal, then the identified association is the result of the LD signal being produced by a SNP that is in LD with the causal SNP, such that signal is being referred. To rule out a referred signal, additional studies may be necessary to further pinpoint the exact location of the causal SNP, such as fine-mapping [187].

The standard assessment of genome-wide association is single-locus statistic tests, which assess each SNP one at a time for association to the provided phenotype. Quantitative traits are usually analyzed using generalized linear model (GLM), the most common of which is the analysis of variance (ANOVA). ANOVA is similar to linear regression and uses genotype classes as the predictor variable. Dichotomous case/control traits are usually analyzed using contingency table methods or logistic regression. The most common form of the contingency table test is the chi-square test and Fischer's exact

test. Logistic regression is an extension of linear regression and is preferred because it allows for adjustments based on clinical covariates and allows for interrogation of the effect size.

The statistical power of a test is affected by how the genotype data is shaped for analysis because degrees of freedom are dictated by how the data are encoded. Allelic association tests assess the association between one allele of a given SNP and the phenotype. Genotypic association tests assess the association between genotypes/genotype classes and the phenotype. Similar to linkage, models can be defined such as dominant, recessive, multiplicative, or additive [188]. Because the general practice of GWAS is to examine additive models only, since the additive model has enough power to test for both additive and dominant effects, it can lead to recessive effects being missed [189]. In addition to model misspecification, association analysis suffers other similar hurdles to those seen in linkage analysis leading to loss of power and a subsequent masking of positive findings, such as population stratification and corrections for multiple testing [190-192].

Posterior Probability of Linkage Disequilibrium (PPLD)

The PPLD is a variation of the LD-PPL (posterior probability of linkage allowing for LD); and both are variations of the PPL, described above. Inclusion of a linkage disequilibrium parameter in the underlying PPL likelihood allows for assessment of trait-marker linkage disequilibrium. This allows for rescaling of the LD-PPL by using linkage as a condition while modeling linkage disequilibrium, which provides a means for separating LD evidence from underlying linkage evidence at any given locus. This is of

critical importance when the multipoint PPL value approaches 1 [169]. If linkage and LD were unable to be disaggregated, a PPLD score might appear high due to linkage when LD may not be present. The PPLD, similar to the PPL, integrates over trait parameters and allows for full use of all pedigree data. The prior probability of LD given linkage (PPLD|L) is also set to 2% because evidence of linkage indicates a strong possibility of LD given the related nature of the two measurements. In regions where there is no linkage the prior probability of LD (PPLD(L)) is set to 0.04%, requiring more evidence to provide a posterior probability in support of a hypothesis of association. This is because the vast majority of the human genome will not exhibit linkage to a given disease, so much greater evidence of co-segregation of a SNP allele and a disease phenotype is required in any unlinked region to overcome the greater initial skepticism that any specific SNP is likely to be in LD with the disease. The PPLD is single locus, and is therefore calculated for one SNP at a time and measures the evidence for or against LD to that SNP. The software package Kelvin also allows for assessment of marker-to-marker LD, making it a great tool for assessing SNPs when creating dense marker maps for use in linkage analysis [167].

Interpreting the PPL and PPLD|L

In order to determine what PPL and PPLD|L values are worth pursuing, simulation studies were conducted using the structure and phenotypes present in the Canadian sample used in this thesis (discussed in detail in Chapter 3). SLINK [193] was used to simulate datasets with no underlying linkage or linkage disequilibrium or with linkage but no linkage disequilibrium. Three marker types were modeled with 2,500

replicates each; a microsatellite with five alleles of equal frequency, a SNP with MAF=0.5, and a SNP with MAF=0.25. No linkage/no LD replicates with each marker type were analyzed for linkage. Across all replicates, 81% produced scores <2%, demonstrating evidence *against* linkage. PPL scores >10% occurred 0.6% of the time and scores >25% appeared only 0.13% of the time. Results were similar for the PPLD|L analysis of the SNP markers generated under linkage, but no LD; 88% of replicates produced PPLD|L scores <2%, indicating evidence *against* association. PPLD|L scores >10% occurred 0.8% of the time, and scores >25% only occurred 0.2% of the time. Note that evidence for LD was low, despite that these replicates were all simulated with a strong linkage signal, highlighting the performance of the PPLD|L in separating evidence for LD from evidence for linkage. As a result, scores over 25%, and even those over 10%, seem worth pursuing, especially if multiple scores of these magnitudes are observed with a discrete genomic region. It seems highly unlikely for all or even most of them to be due to chance.

Chapter 3: Preliminary Data

Sample Details

Recruitment for the sample used for analysis in this thesis began in the early 1990s and was first described in 1993, when 72 members of five families of Celtic origin from a rural Canadian region were assessed using the Positive and Negative Syndrome Sale (PANSS) to demonstrate the validity of positive and negative symptom measures as independent dimensions in familial schizophrenia [194]. More than 20 years later, following ongoing recruitment and reassessment, the sample now consists of 30 pedigrees and is comprised of 573 individuals. Families were selected for enrollment if they had an extended family of two or more generations of adults, with a proband and at least one other relative having a diagnosis of chronic schizophrenia. Schizophrenia and other generally related disorders (as discussed in Chapter 1) were required to have the appearance of segregating in a unilineal autosomal dominant-like inheritance pattern. Families with prevailing bipolar affective disorder, known organic or physical disturbances causing psychiatric illness, or a bilineal segregation of schizophrenia were excluded.

Families were ascertained in their entirety, allowing for proband status to be assigned to all affected subjects, and therefore no subjects were excluded from analysis. Family histories were obtained for each subject from three or more family members and genealogical records were used for confirmation of birth and death dates. Originally, subject families originated in and rarely moved from the same rural island region in Canada where only one psychiatric hospital was available until the 1980s.

Comprehensive records by file-card system documenting admissions allowed for searching of medical records back to 1866 and provided virtually complete ascertainment of psychiatric hospitalization. Records were collected for all subjects with a history of psychiatric evaluation, living subjects were interviewed by a psychiatrist, and diagnostic folders were reviewed independently by two psychiatrists, one of whom was blind to the pedigree structures [195].

The affected individuals were placed into two groups based on a proposed diagnostic hierarchy for genetic studies and sample size considerations. Schizophrenia and schizoaffective disorders were combined into one group, called 'narrow', due to their extensively shared clinical presentation and diagnostic requirements. Psychosis not otherwise specified, schizotypal and paranoid personality disorders were added to the 'narrow' diagnoses into a second group, called 'broad', because these schizophrenia spectrum disorders share similar, but not exact, symptomology with schizophrenia suggesting some extent of shared underlying neurobiology. The remaining participants were assigned to either the 'unaffected' or 'unknown' groups, depending on what phenotypic data were available [195].

As mentioned above, the first study published using this sample was conducted in 1993. Individuals were assessed using the Positive and Negative Syndrome Scale (PANSS) to demonstrate the validity of positive and negative symptom measures as independent dimensions in familial schizophrenia [194]. In 1994 the sample was used to demonstrate that more subjects were diagnosed with psychiatric illness at progressively earlier ages across generations, providing evidence of anticipation in familial schizophrenia [195]. Several confirmatory studies have been performed using this

sample. In those cases, affected status for each individual was determined based on the study replication being attempted and the patient's diagnosis with respect to the current version of the DSM. In 1997 the sample was analyzed in an attempt to replicate previous reports of a schizophrenia susceptibility locus on chromosome 6p. Though parametric linkage analysis using narrow and broad definitions of schizophrenia and sib-pair analysis using categorical disease definitions both failed to provide significant evidence of linkage, sib-pair analysis using positive-symptom (psychotic), negative-symptom (deficit), and general psychopathology-symptom scales as quantitative traits suggested a schizophrenia susceptibility locus on chromosome 6p related to the severity of psychotic symptoms. The results also suggested that assessment of behavioral quantitative traits may provide increased power compared to conventional methods for the detection of linkage in complex psychiatric disorders [196]. In 1999, results generated using this sample provided independent confirmation of significant evidence of linkage of schizophrenia susceptibility locus to microsatellite markers on 13q32 and support for the presence of a second susceptibility locus on 8p21 [107]. In 2005, this sample was used to replicate linkage of schizophrenia spectrum disorders to chromosome 1q44. The results also demonstrated that simulation studies are critical in determining the significance of results obtained with newer statistical methods, when multiple, but not independent, tests are performed and when sample stratification is employed to lessen the impact of heterogeneity or assess the interaction between loci [197].

Beginning in 2000 this sample was used in discovery analysis. At this time, the sample included 22 medium-sized Canadian families of Celtic and German descent recruited using the same criterion described above. On average 13.8 individuals

participated per family, but the five largest families had 20-29 participants. Subjects (n = 304) were again divided into three groups: narrow, broad, and unaffected, as described above. Within each family there was an average of 3.6 individuals in the narrow diagnostic group (according to DSM-IV criteria), with 15 affected individuals in the largest family. On average, two additional participating family members were diagnosed with schizophrenia-related disorders in the broad diagnostic group. Family members diagnosed as affected spanned three generations in 27% of families, and individuals reported by history to be affected spanned three or four generations in 45% of families. DNA on 288 subjects was available for this study. A genome-wide scan for schizophrenia susceptibility loci produced highly significant evidence of linkage to markers on chromosome 1q21-q22, with a maximum HLOD score of 6.5 [106] (under the narrow phenotype). This same sample was used in 2002 to conduct fine mapping of this locus using 15 genetic markers spanning ~15 cM. Parametric linkage analysis provided a maximum multipoint HLOD score of 6.50 with a `Zmax-1` support interval of < 3cM [187]. The data from the 2000 genome scan was later reanalyzed with the PPL (methodology described in detail in Chapter 2) in 2006 and yielded a multipoint PPL of 99.7% in the same location on chromosome 1q. There was also support for two additional loci under the broad diagnostic criteria; a second peak on chromosome 1p13 with a multipoint PPL of 70% and on chromosome 17q25 with a multipoint PPL of 44% [169]. In 2004 an enlarged sample of 24 families was used for an association mapping study under the chromosome 1q peak. As the PPL method was not yet available, LD was assessed with the program PSEUDOMARKER, an Elston-Stewart based parametric analysis method for joint analysis of linkage and LD in family data [198, 199]. Since the

linkage finding in the region was under the narrow phenotype, only that phenotype was analyzed. In this study 330 subjects were phenotypically evaluated with 85 coded as affected (schizophrenia or chronic schizoaffective disorder), 232 as unaffected, and 24 coded as unknown. Fourteen microsatellites and 15 SNPs from the 5.4 Mb region between D1S1653 and D1S1677 were analyzed, and 2 microsatellites and 6 SNPs produced significant evidence of LD ($p < 0.05$) with schizophrenia. All of these markers fall within the genomic extent of *NOS1AP* (formerly *CAPON*), discussed in further detail later in this Chapter [200].

The sample now consists of 30 pedigrees and is comprised of 573 individuals. Recruitment criteria as well as diagnostic groupings remain unchanged from what is described above. In this sample 105 individuals are coded as affected under the narrow diagnostic scheme, an addition 56 are coded as affected under the broad diagnostic scheme, 105 lack phenotypical assessment and are coded as unknown. The remaining 231 individuals are classified as unaffected. Subjects in this sample have been followed for up to 20 years allowing for continued observation of diagnostic stability.

Microsatellite data are available for the original 22 families in the sample comprised of 304 samples, of which 288 had DNA available. These markers consisted of 379 simple tandem repeat markers with an average heterozygosity of 0.76 and an average marker density of 9 cM. Markers were specifically chosen to be informative, equally spaced, and far apart enough to minimize the chance that they would be in LD with one another [106].

SNP data are now available for all 30 families, with genotyping using Affymetrix 6.0 arrays previously completed at The Center for Applied Genomics at the University of

Toronto (TCAG), with an average completion rate of 99.0% across all SNPs, genome-wide. Genotype cleaning was previously performed using PLINK v1.07 [201, 202] and included removal of SNPs with <98% completion rates, monomorphic SNPs, or SNPs with >1% rate of Mendel Errors. All SNPs had Hardy-Weinberg p-values > 0.001. Custom software was then used to identify and remove genotypes causing Mendel errors. After removing these SNPs, 98.8% of attempted genotypes were available. Additionally, custom software using a pattern-based algorithm was used to predict SNPs that alter microRNA binding sites [203]. This generated a panel of 48 candidate SNPs, which were genotyped using bead-based oligonucleotide ligation, and data were cleaned using the same protocols outlined above [204].

DNA is currently available for 376 individuals, though some of the pedigrees have individuals present that serve as linkers for whom no DNA is available either because the patient was unavailable, refused participation, or is deceased. In a small number of cases without cell lines generation of SNP and microsatellite data exhausted the available DNA.

Whole Genome Sequencing (WGS) of 10 affected samples (7 female, 3 male) was performed by Knome Inc. in 2012 using paired end sequencing array by Illumina and Illumina's propriety software the Consensus Assessment of Sequence and Variation (CASAVA) pipeline version 1.9.1 [205]. Individuals were selected for sequencing based on three criteria: 1) coming from the core portion of one of the larger pedigrees, 2) being a patient with a 'typical' case of schizophrenia, featuring diagnostic stability and phenotypic homogeneity with other sample subjects, and 3) having high quality DNA available to send out for sequencing. For the work described in this thesis, BAM files

generated by Knome Inc. were shuffled using SAMtools (v1.3.1) sort, reverted to fastq using bedtools (v2.25.0) bamtofastq, and then aligned to 1000Genomes build 37 of the human genome using BWA-MEM (v0.7.15) [206-209]. Newly aligned BAM files were sorted, indexed, and duplicates removed using SAMtools. Base recalibration and variant discovery were performed using the Genome Analysis Toolkit (GATK) v3.7 [210-212]. Functional annotation and prediction were performed using SnpEff and SnpSift, using dbSNP database v138 for build 37 [213-215].

Linkage Analysis

A genome scan was performed using all available microsatellite data from the original sample (22 families) under both the narrow and broad diagnostic schemes (defined in the Sample Details section above) from 2000, using the PPL (Kelvin version 2.4.0). The newer version allows for easier multipoint analyses and has other computational enhancements compared to published work described earlier [169]. Three-point analysis across the whole genome revealed PPL values >20% in six locations: one on chromosome 1, two on chromosome 2, one on chromosome 8, one on chromosome 11, and one on chromosome 17 (**Figure 2**).

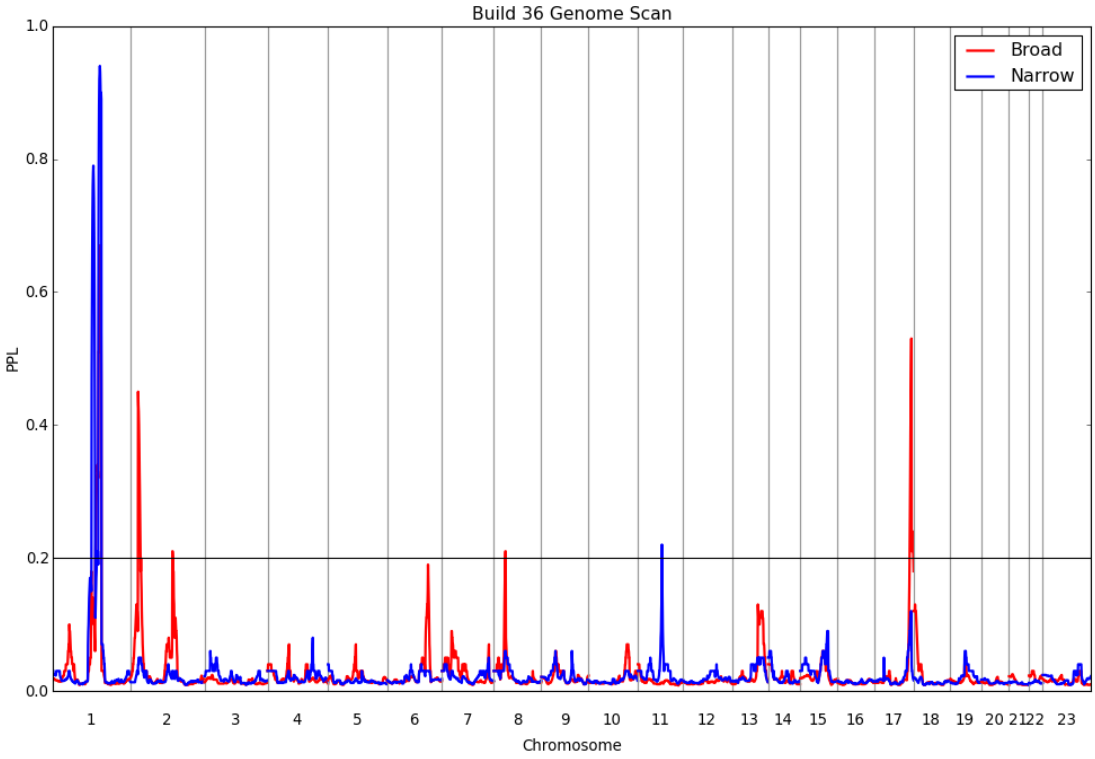

Figure 2: PPL data for original 22 Canadian families, generated in 2012 (*unpublished*).
Red indicates results for individuals diagnosed under the broad definition, and blue indicates results for individuals diagnosed under the narrow definition as described in the Sample Details section of this Chapter.

Linkage peaks of interest were determined in two ways: 1) All PPL scores => 20%, and 2) PPL scores >= 5% that overlapped with regions of interest determined by recalculating results from a large-scale meta-analysis [216], focusing on samples of European ancestry but excluding the results from the Canadian linkage sample. Linkage regions were defined surrounding the linkage peaks using custom software (**Table 1**). Linkage results generated by Kelvin 2.4.0 were parsed for the first location preceding each peak with a PPL score >= 2%, and for the last location with a PPL score >= 2 % following the peak.

Chrom	Phenotype	LP PPL Score	LR Start (cM)	LR End (cM)
1	Broad	0.67	123.0	181.0
1	Narrow	0.94	123.0	182.0
2	Broad	0.45	2.0	59.0
2	Broad	0.21	120.0	167.0
3	Narrow	0.06	0.0	56.0
8	Broad	0.21	24.0	54.0
8	Narrow	0.06	25.0	69.0
11	Narrow	0.22	74.0	115.0
17	Broad	0.53	114.0	137.0

Table 1: Linkage regions for further investigation.
Linkage regions selected for further evaluation based on criteria outlined above. (Chrom = chromosome, LP = Linkage Peak, LR = Linkage Region)

Association Analysis

A GWAS was conducted using the PPLD in 2013 on the expanded sample of 30 pedigrees using the genotype data described above. Results were parsed via Python

script for SNPs with PPLD|L scores => 20% within the linkage regions defined in Table 1 (**Table 2**). SNPs of interest were identified in five genes, further described below.

Chromosome	Phenotype	SNP	PPLD|L	Gene
1	Narrow	rs7419214	0.35	None
1	Narrow	rs17477236	0.28	*VAV3*
1	Narrow	rs641227	0.23	None
1	Narrow	rs465310	0.22	None
1	Narrow	rs12725553	0.27	*NOS1AP*
1	Narrow	rs4411117	0.27	*NOS1AP*
2	Broad	rs12991828	0.21	*DPYSL5*
2	Broad	rs486582	0.2	*DPYSL5*
2	Broad	rs7578749	0.28	None
3	Narrow	rs12494654	0.21	*GRM7*
11	Narrow	rs17631231	0.35	None
17	Broad	rs1060120	0.21	*H3F3B*

Table 2: Risk SNPs identified by GWAS.
GWAS SNPs with a PPLD|L score => 20% and the gene (if applicable) each SNP is located in according to dbSNP.

Genes of Interest

VAV3

VAV3 is located on chromosome 1p13.3 and is a member of the VAV family of proteins. VAV proteins (VAV1, VAV2, and VAV3) are guanine nucleotide exchange factors for Rho family GTPases, which cycle between an inactive GDP-bound state and an active GTP-bound state. They play critical roles in the control of the cytoskeleton, cell motility, gene expression, cell proliferation, cell transformation, and oncogenesis. VAV proteins are differentially expressed, with VAV1 restricted to expression in

hematopoietic cells, whereas VAV2 and VAV3 both demonstrate broader expression profiles. All VAV family members have a zinc finger domain (reviewed in [217]).

VAV3 is closely related to the axon guidance pathways, which have also been identified as playing a role in schizophrenia [218]. During axon guidance when ephrin binds to Ephs the event triggers VAV-dependent endocytosis of the ligand-receptor complex, which changes an attraction interaction into a repulsive one. In the absence of VAV proteins, ephrin-Eph endocytosis cannot occur, leading to defects in growth cone collapse in vitro and also in the ipsilateral retinogeniculate projections in vivo [219]. GABAergic neurons from the hippocampus were used as a model to investigate the specific implication of VAV3 in axonal development. Growth cone collapse was measured in wild-type and VAV3-deficient hippocampal neurons following stimulation with a ligand for EphA receptors. VAV3-deficient cells were less responsive than wild-type cells, indicating that VAV3 is critical for the regulation of axon branching and growth cone morphology, as well as for Ephrin-dependent axon collapsing responses in GABAergic cells [220]. VAV3 has been found to be expressed at high levels in Purkinje and granule cells. Primary neuronal cultures were used to demonstrate that VAV3 is important for dendrite branching in these regions, indicating that VAV3 contributes to the timely developmental progression of the cerebellum [221].

A genome-wide linkage analysis of 236 Japanese families produced significant evidence of linkage (LOD = 3.39) at rs2048839 and 95% CI includes the *VAV3* locus [222]. A GWAS based on meta-analysis for a Japanese sample produced evidence suggestive of association between schizophrenia and rs1410403, which is in the region of *VAV3* [223].

Based on these results, voxel-based morphometry (VBM) was performed (100 cases, 264 healthy controls) and demonstrated rs1410403 might affect volume of the left superior and middle temporal gyri, which were reduced in patients with schizophrenia compared to healthy controls. Additionally, mutation screening of *VAV3* was performed and four missense variants were detected. These mutations were then followed up in a large independent sample. One of those variants was associated with schizophrenia ($P = 0.02$) [224].

NOS1AP

NOS1AP (nitric oxide synthase 1 (neuronal) adaptor protein, formerly known as *CAPON*) is located on chromosome 1q23.3 was first identified in the rat as a neuronal nitric oxide synthase (nNOS) binding protein with the ability to disrupt the association of nNOS with post-synaptic density scaffolding proteins [225]. This association plays a crucial role in targeting nNOS to the post-synaptic N-methyl-D-aspartate receptor (NMDAR), which allows for activation of the NMDAR and nNOS, which in turn generates NMDAR-mediated NO release into synaptic structures [226, 227]. The NMDA receptor channel, a subtype of the glutamate-gated cation channels, was first linked to the neurobiology of schizophrenia in the 1980s and is discussed in detail in Chapter 1. Glycine/D-serine control the ability of L-glutamate to open the NMDAR channel, which is believed to play a dominant regulatory role in neuroplasticity (reviewed in [228]).

Three variants of NOS1AP have been identified in humans. The first isoform is 501 amino acids long, made from all 10 exons of the gene, and contains two functional

domains: one N-terminal phosphotyrosine-binding (PTB) domain and one C-terminal PDZ binding domain. The second isoform is shortened and only contains the last two exons of NOS1AP and produces a truncated protein of 210 amino acids containing only the PDZ domain [229]. The third isoform has a unique 5' exon and transcriptional start site, and it is predicted to be ~18kD. Like the second isoform, it is a truncated version of the full-length variant, but includes a carboxyl-terminal PDZ-binding domain [230]. The first 180 amino acids of the full-length isoform have been previously shown to be required for the binding of the N-terminal targets Dexras1 and Synapsin [231, 232]. Increased expression of the full-length variant has been linked to the reduction of both the number and branching of dendrites in the hippocampal neurons [233]. The truncated isoforms have not been shown to have this functional effect. However, previous work has shown that the terminal 125 amino acids of the full-length protein are enough to bind the PDZ-domain of nNOS and interfere with the binding of nNOS and PSD93/PSD95, which could result in competitive inhibition against the binding of other ligands [225]. Two of the three isoforms, NOS1AP-L and NOS1AP-S were examined in individuals with schizophrenia and healthy controls. A significant change in expression was established for the short isoform. [234]. NOS1AP levels have been demonstrated to modulate cortical neuron migration, resulting in aberrant neuronal connectivity, which could play a role in schizophrenia [235].

A highly significant linkage finding (HLOD score of 6.5; $p < 0.0002$) of schizophrenia to chromosome 1q22 was initially found using a smaller subset of the Canadian sample described above [106]. This study showed evidence of a susceptibility gene within 6 Mb of DNA. It was followed up by a fine-map linkage study with the

same sample and narrowed the target area to 3 Mb, an area containing approximately 50 protein coding genes [187]. Other independent studies have reported linkage to schizophrenia in this region, whereas some have not [108, 110, 236-238]. There are several reasons why linkage analysis may produce inconsistent results, and they are discussed in greater detail in Chapter 2.

An initial association study was conducted with 15 SNPs in the subset of the Canadian sample described above. Three SNPs located within NOS1AP were found to be significantly associated with schizophrenia [200]. Association to schizophrenia has been reported in this area in a Han Chinese sample, but to different SNPs than those identified in the Canadian sample subset [239]. Two additional studies (Spanish and Colombian ancestry) have produced strong evidence of association to schizophrenia for D1S1679, which is located within 25kb of NOS1AP [240, 241]. Further genotyping of 24 SNPs in the Colombian sample from within NOS1AP detected significant association to 8 SNPs, including two that were present in our Canadian sample [242].

Functional studies have been conducted that support a role for NOS1AP in schizophrenia. Increased expression of NOS1AP was observed in post-mortem samples from the dorsolateral prefrontal cortex of patients diagnosed with schizophrenia compared to normal controls in patients from several locations within the United States [234]. In order to confirm that these findings point to causality, the DNA changes were assessed in these samples, regardless of whether a there was an established diagnosis of schizophrenia. All three SNPs previously identified in the Canadian association study described above produced a significant correlation with NOS1AP expression. In all three cases, the sequence variant that led to higher expression in the United States sample was

the sequence variant associated with schizophrenia in the Canadian sample. A schizophrenia-associated noncoding variant, rs12742393, was identified first by the PPLD, with gene expression functionality confirmed by luciferase reporter assay, and further assessed for binding of nuclear proteins by electrophoretic mobility shift assay [186].

DPYSL5

Dihydropyrmidinase-like 5 (*DPYSL5*, formerly known as *CRMP5*) located on chromosome 2p23.3 is a member of the CRMP family, whose members are believed to play a role in growth cone guidance during neural development. A synaptosomal proteomic study on rats treated with MK-801 (a specific NMDAR antagonist that induces NMDAR hypofunction and schizophrenia-like symptoms in rodents) demonstrated altered expression in CRMP5 in cases compared to controls. Differential expression was confirmed by western blot assay [243]. CRMP5 has also been found to regulate neurite outgrowth inhibition and to induce mitophagy, regulating mitochondrion numbers in dendrites [244, 245]. These results provide compelling evidence for *DPYSL5* as a candidate gene for schizophrenia susceptibility.

GRM7

GRM7 (glutamate receptor, metabotropic 7) is located on chromosome 3p26.1 and is a metabotropic glutamate receptor, which are divided into three groups based on sequence homology, putative signal transduction mechanisms, and pharmacologic properties. GRM7 belongs to MGluR group III, which are linked to the inhibition of the

cAMP cascade. GRM7 is activated by L-glutamate, the major excitatory neurotransmitter of the central nervous system and is an important presynaptic regulator of neurotransmission [246, 247].

GRM7 was first associated with schizophrenia in 2008 in a population of Japanese ancestry (cases = 2293, controls = 2382). The sample was screened for mutations in all exons, exon/intron junctions, and promoter regions of the *GRM7* gene. A synonymous mutation in exon 1 showed potential association (allelic p = 0.009) with schizophrenia. Dual-luciferase assay demonstrated suppression of transcription activity by exon 1 and a statistically significant difference in the promoter activity between the T and C alleles [248]. Another study selected 43 common SNPs within *GRM7* and scanned for association with schizophrenia in 100 case-control pairs of Japanese subjects. Two SNPs in *GRM7* demonstrated highly significant haplotype association, and these results were confirmed in an expanded sample (404 cases, 420 controls) [249]. Significant genome-wide linkage to chromosome 3p was followed by a nominally significant association finding in intron 1 of *GRM7* in 124 Indonesian sib-pair families [250]. Evidence of strong association between *GRM7* and schizophrenia was demonstrated in a meta-analysis of an Indo-European and Dravidian population, and replicated in a meta-analysis of that sample and data from the PGC [251]. Investigation of *GRM7* in the Han Chinese population yielded significant association to two SNPs and schizophrenia, as well as to 3 different SNPs and major depressive disorder [252]. These studies collectively demonstrate strong support for a role for *GRM7* as a candidate gene for schizophrenia susceptibility and its associated biological mechanisms fit well within the proposed underlying neurobiology for the disease.

H3F3B

H3F3B (H3 histone, family 3B) is located on chromosome 17q25.1 and belongs to one of the five classes of histone genes that have been reported, all of which are involved in chromosome structure. The SNP originating from this gene is in the 3' UTR region, a microRNA (miRNA) binding site. miRNAs are ~21 nucleotide single-stranded molecules that negatively regulate the expression of 20-30% of human genes. Variations in miRNA binding sites have been found to have profound biological consequences [253, 254]. Reduced expression of miRNAs has been tied to both schizophrenia and bipolar disorder [255]. Changes to the interaction between the miRNA and its binding site would lead to a decrease of H3F3B, which in turn could impact the function of histones and their epigenetic modifications. Both have been demonstrated as playing a role in the development of schizophrenia, as well as other neuropsychiatric disorders [256-258].

Chapter 4: Harmonization of the Data

Updating Previous Analyses

One of the major advantages of the Canadian linkage sample is how long it has been maintained, but that also means that technology has advanced since earlier studies were conducted on this sample. Prior to beginning the bulk of the work described in this thesis, it was necessary to bring all analyses up-to-date with respect to version of the human genome used to ensure consistency. Since the whole genome sequencing was analyzed using build 37 of the human genome, that version was used to update previous analyses.

The first analysis to be updated was the linkage analysis using the PPL, performed on 22 Canadian families of Celtic and German descent. As described in Chapter 3, the original analysis utilized 379 simple tandem repeat (STR) markers with an average heterozygosity of 0.76 and average marker density of 9 cM with positions determined using build 36 of the human genome. A python script was used to pull all available sex-averaged map positions from The Rutgers Maps v.3, which uses dbSNP Build 137 reference SNPs and UniSTS markers from Build 37.3 (GRCh37.p5) [259]. This allowed for direct conversion of 324 of the STR markers to build 37, leaving 55 remaining STR markers to be converted. A python script was used to pull the build 36 base-pair positions for these markers from The Rutgers Maps v.2 [260]. The UCSC LiftOver tool was then used to batch convert the values on this list to build 37 base-pair positions [261]. Finally, a python script was used to pull build 37 sex-averaged map positions in cM,

corresponding to the build 37 base-pair positions (female-averaged map positions were used for chromosome 23).

After the 379 microsatellites had been assigned b37 sex-averaged map positions, a quality control python script was run to assess 1) whether the order of markers had changes as a result of converting the positional values, and 2) how much change had occurred between the two builds. The order of the markers was confirmed to be unchanged, which had been expected given the fact that their average spacing was ~9cM, making it highly unlikely that any two markers so far apart would change relative positions. The minimum change observed between the two builds was 0 cM, the maximum change was determined to be 2.42 cM, and the average change across 379 markers was 0.16 cM (**Table 3**).

Change Between Builds	Number of Markers
0 cM	27
< 1 cM	344
1-2 cM	6
> 2 cM	2
Total	379

Table 3: Breakdown of positional changes between builds.
Break down of position changes when converting microsatellite markers from build 36 to build 37. Number of markers that changed position by the defined range are indicated.

A genome scan was performed using the PPL and build 37 marker positions on the same 22 families. Results were graphed using Kelviz and then compared to build 36 genome scan results (**Figure 3**).

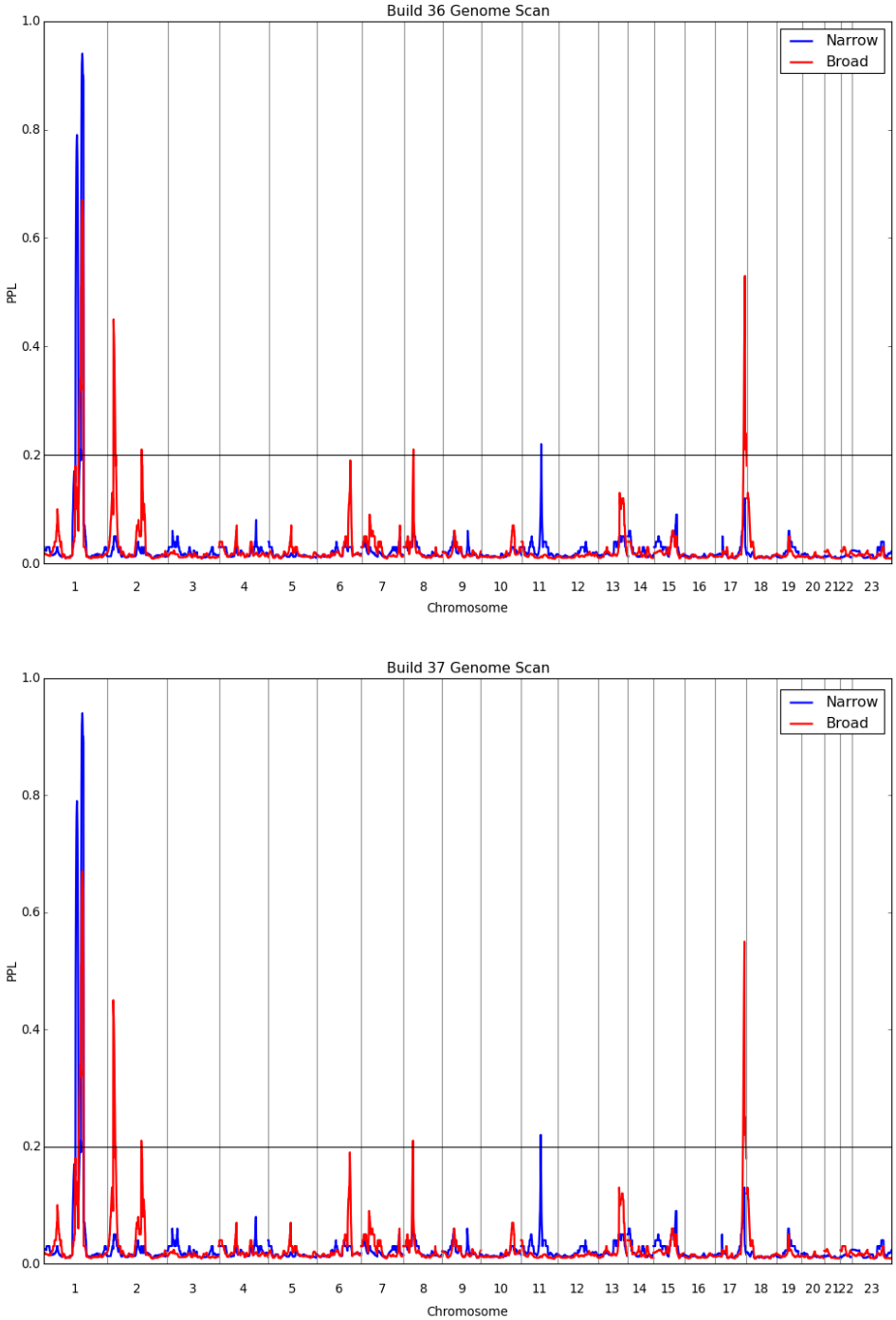

Figure 3: Comparison between build 36 and build 37 genome scans.
Both scans use the same data (family pedigrees, affected status, marker frequencies, genotype data) with the exception of marker positions. Red indicates results for individuals diagnosed under the broad definition, and blue indicates results for individuals diagnosed under the narrow definition as described in the Sample Details section of this Chapter.

The new linkage results were analyzed for linkage regions using the custom script described above and compared to the results generated using build 36 STR positions (**Table 4**).

		Build 36 Genome Scan			Build 37 Genome Scan		
Chr	**Pheno**	**PPL_{max}**	**LR_{start}**	**LR_{end}**	**PPL_{max}**	**LR_{start}**	**LR_{end}**
1	Broad	0.67	123.0	181.0	0.67	123.0	181.0
1	Narrow	0.94	123.0	182.0	0.94	123.0	182.0
2	Broad	0.45	2.0	59.0	0.45	2.0	59.0
2	Broad	0.21	120.0	167.0	0.21	120.0	167.0
3	Narrow	0.06	0.0	56.0	0.06	0.0	56.0
8	Broad	0.21	24.0	54.0	0.21	26.0	54.0
8	Narrow	0.06	25.0	69.0	0.06	26.0	66.0
11	Narrow	0.22	74.0	115.0	0.22	74.0	115.0
17	Broad	0.53	114.0	137.0	0.55	114.0	137.0

Table 4: Comparison of linkage regions between builds.
Comparison of linkage regions between build 36 and build 37 genome scans. Most of the peaks and linkage regions remain the same, with two exceptions (which are highlighted): 1) slight narrowing of the linkage region on chromosome 8 under the narrow and broad phenotypes, and 2) a slight increase to the maximum PPL value for the peak on chromosome 17 under the broad phenotype.

Most of the linkage peaks and regions remained the same following conversion from build 36 to build 37 for the microsatellite marker positions, and only minimal changes were observed anywhere. The linkage regions on chromosome 8, under both the narrow and broad phenotypes, were slightly narrowed. This was not surprising given that marker D8S1130 was located at 22.62 cM under build 36, and at 24.76 cM under build 37, yielding a change of 2.14 cM. The peak on chromosome 17 under the broad phenotype increased from 0.53 to 0.55. This change was also not surprising, since the peak occurred at 129 cM under both builds, and marker D17S784 was located at 129.9

cM under build 36 and 129.43 cM under build 37, with a total change in position of 0.47 cM.

The 2013 GWAS was conducted using Kelvin 2.4.0 and did not need to be reanalyzed because (as described in Chapter 2) the PPLD is currently implemented as a two-point (marker vs. disease) analysis, and is therefore calculated for one SNP at a time and measures the evidence for or against LD between schizophrenia and that SNP. As such, map positions have no bearing on the results generated.

The map positions from the 2013 GWAS were updated to build 37 positions for two reasons: 1) later analysis described in Chapter 5 require positions for these SNPs for work done on the sequence data, which is aligned to build 37, and 2) the development of Kelvin-LKS allows for SNP-based linkage analysis to be performed using many SNPs in large families. Positions for these SNPs were converted in a similar manner to how the microsatellites were converted. A python script was used to match rs identification numbers for all SNPs previously genotyped and retrieve their sex-averaged map positions under build 37 using the Rutgers Maps v.3 (female-averaged map positions were used for chromosome 23). In the case that a SNP was not found on the same chromosome as it was assigned to under build 36, it was removed from the pedigree and map file for that chromosome using PLINK v1.07. Most SNPs were able to be converted (**Table 5**).

Chromosome	SNP Total (build 36)	SNP Total (build 37)	SNPs Removed
1	53771	53542	229
2	56507	56385	122
3	46752	46655	97
4	42938	42868	70
5	43838	43705	133
6	43976	43751	225
7	36646	36447	199
8	37457	37292	165
9	32189	32120	69
10	37021	36888	133
11	34014	33947	67
12	32888	32822	66
13	26295	26049	246
14	21646	21609	37
15	19912	19846	66
16	20937	20839	98
17	15675	15611	64
18	20159	20108	51
19	9098	9012	86
20	17446	17433	13
21	9754	9724	30
22	8711	8639	72
23 (X)	25462	25339	123
Total	693092	690631	2461

Table 5: SNP position conversion between builds.
SNP position conversion between builds 36 and build 37 using the Rutgers Maps v.3. 99.6% of the SNPs genotyped were available for cM position retrieval from The Rutgers Maps.

The newly-generated build 37 pedigree and map files were used along with the frequency files (derived using Mendel v13.2 [262] on the family SNP data) from the 2013 GWAS to create a data set for linkage analysis by Kelvin-LKS (discussed in detail in Chapter 2), subsequently conducted by the Vieland Lab. First a python script was used to find and generate a list of all SNPs with a MAF >= 0.3. PLINK v1.07 was then used to extract these SNPs from the existing pedigree files, to ensure that only the most informative SNPs were considered for further analysis. Next, a custom python script

called PLINK's LD-based pruning method, which prunes based on the variance influence factor ($1/(1-r^2)$), recursively removing SNPs within a sliding window and generates a 'prune.in' file and a 'prune.out' file. A second PLINK call was then made to reduce the pedigree file to include only the SNPs present in the 'prune.in' file. This method was applied recursively by the python script using an r^2 threshold of 0.2, checking three consecutive SNPs at a time against each other, until no SNPs were in LD (the 'prune.out' file was empty). Then a custom python script was used to assess the distance between SNPs on each map to ensure that none were >= 3 cM. The data set was then double-checked for marker-to-marker LD using Kelvin 2.4.0. Kelvin uses offspring genotypes to reconstruct missing genotypes of parents, whereas PLINK does not, so this step ensured that any remaining marker-to-marker LD was identified. A custom python script then removed all SNPs identified by Kelvin as having an $r^2 > 0.4$, and Kelvin marker-to-marker LD was reassessed to ensure no remaining LD existed. The full file sets necessary to run Kelvin-LKS were generated by python script, verified by quality control scripts, and sent out to the Vieland Lab (**Figure 4**). Minor allele frequencies were analyzed in the finalized data set. The minimum MAF was 0.3, the maximum MAF was 0.5, and the average MAF was 0.39. Inter-marker distance was computed for the finalized data set (**Table 6**).

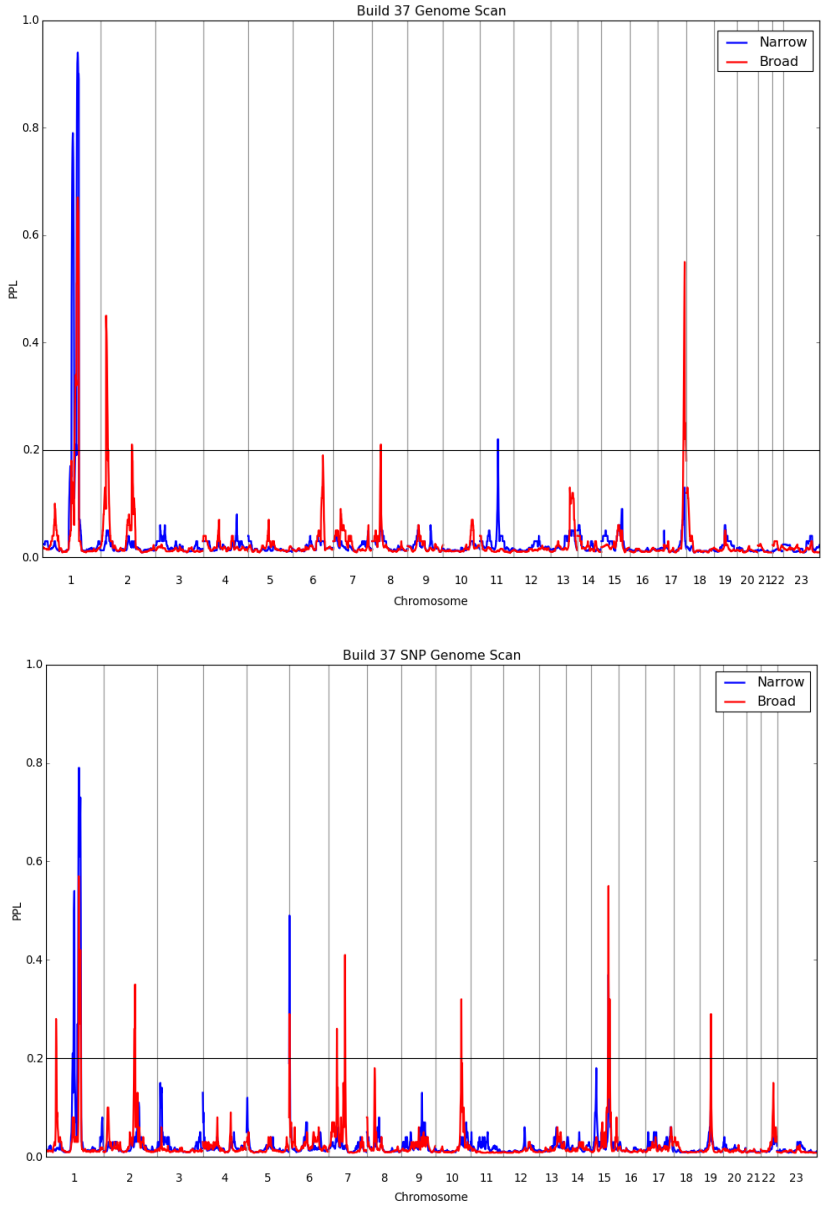

Figure 4: Build 37 SNP-based PPL genome scan.
This analysis used the expanded sample of 30 pedigrees (bottom), compared to the microsatellite scan above which was conducted on 22 pedigrees (top). Red indicates results for individuals diagnosed under the broad definition, and blue indicates results for individuals diagnosed under the narrow definition as described in the Sample Details section of Chapter 3.

Gap Size	Number of Markers
< 1 cM	30,160
1 – 2 cM	90
> 2 cM	5
Total Markers	30,255

Table 6: Breakdown of inter-marker distances in final data set.
The minimum inter-marker distance was 0.00001 cM, the maximum inter-marker distance was 2.5748 cM, and the average inter-marker distance was 0.125 cM.

The new SNP-based linkage results were analyzed for linkage regions using the custom script described above and compared to the results generated using microsatellite markers (**Table 7**).

		MSAT Genome Scan			SNP Genome Scan		
Chr	**Pheno**	**PPL_{max}**	**LR_{start}**	**LR_{end}**	**PPL_{max}**	**LR_{start}**	**LR_{end}**
1	Broad	0.67	123.0	181.0	0.57	122.0	174.0
1	Narrow	0.94	123.0	182.0	0.79	120.0	180.0
2p	Broad	0.45	2.0	59.0	0.10	14.0	38.0
2q	Broad	0.21	120.0	167.0	0.35	126.0	180.0
3p	Narrow	0.06	0.0	56.0	0.15	14.0	54.0
8p	Broad	0.21	26.0	54.0	0.18	36.0	50.0
8p	Narrow	0.06	26.0	66.0	0.08	36.0	72.0
11q	Narrow	0.22	74.0	115.0	0.05	74.0	86.0
17q	Broad	0.55	114.0	137.0	0.06	118.0	136.0

Table 7: Comparing MSAT and SNP genome scans.
Comparison of linkage regions from build 37 microsatellite genome scan to results from build 37 SNP genome scan. Three peaks with a PPL >= 20% on the microsatellite scan are reduced below 20% on the SNP scan. The second peak on chromosome 2 under the broad phenotype and the peak on chromosome 3 under the narrow phenotype increased on the SNP scan compared to the microsatellite scan. In most cases the interval of the linkage region is smaller compared to the microsatellite scan. This was not surprising given the uniformity and ubiquitous nature of SNP markers compared to microsatellite markers.

Three peaks are observed to have a large decrease with respect to their maximum PPL score; chromosomes 2p and 17q under the broad phenotype, and chromosome 11q under the narrow phenotype. Though all three peaks still produce evidence of linkage (PPL > 2%), the marked reduction makes them less compelling than previous results. One reason for these changes may be that one of the microsatellite markers used in the original analyses produced spurious genotypes. In order to investigate this hypothesis further, all three chromosomes were re-analyzed, removing one of the two markers responsible for the original peak at a time. In the re-analysis, a four-point analysis was conducted, to reduce the likelihood that observed changes were due to loss of informativeness (**Table 8**).

Peak	PPL_{max}	Position	Marker 1	Marker 2	Drop M1	Drop M2
2p	0.45	28.0	DS1400	DS1360	0.36	0.11
11	0.22	85.0	D11S2371	D11S2002	0.03	0.19
17	0.55	129.0	D17S1301	D17S784	0.20	0.11

Table 8: Re-analysis of changed linkage peaks.
Re-analysis of three linkage regions, dropping one microsatellite marker from analysis at a time. Dropping either marker on chromosome 2p under the broad phenotype reduces the peak by >50%. Dropping the first marker responsible for the peak on chromosome 11 under the narrow phenotype reduces the PPL score to 0.03. Dropping the second marker responsible for the peak on chromosome 17 under the broad phenotype reduces the PPL score by more than 66%.

The reduced scores for all three peaks are much closer to the scores generated by the SNP genome scan. The differences that remain may be the result of adding more families and more individuals from the families used in the microsatellite genome scan. In addition to the changes in linkage regions, new linkage regions were identified on the SNP scan (**Table 9**).

Chromosome	Phenotype	PPL_{max}	LR_{start}	LR_{end}
1	Broad	0.28	36.0	74.0
6	Broad	0.29	0	14.0
6	Narrow	0.49	0	14.0
7	Broad	0.41	2.0	86.0
10	Broad	0.32	118.0	144.0
15	Broad	0.55	72.0	98.0
15	Narrow	0.37	74.0	100.0
19	Broad	0.29	44.0	62.0

Table 9: New linkage regions identified SNP-based genome scan.
Eight new linkage regions with a PPL_{max} >= 0.2 were identified in the new analysis. The region on chromosome 7 under the broad phenotype is large due to the fact that there are two local maxima (PPL = 0.26 at 40 cM and PPL = 0.41 at 80 cM) and the PPL scores in that region do not go below 2%.

The discovery of new linkage regions with a maximum PPL >= 0.2 was not entirely unexpected. The genome scan using SNPs allowed for finer mapping and also included 8 new families, as well as new individuals from the original 22 families. Of the eight new regions identified in Table 9, six are present on the build 37 microsatellite scan: chromosome 1, chromosome 7, chromosome 10, chromosome 15, chromosome 19 under the broad phenotype, and chromosome 15 under the narrow phenotype had findings suggestive of linkage (PPL > 2%), but now are observed to have demonstrably larger peaks at those locations. The peaks on chromosome 6 under both the broad and narrow phenotype are entirely new, and as such may be driven by the new families and the new individuals in the original families.

Compatibility with Reference Populations

The hypothesis of this thesis includes examining all variants identified by whole genome sequencing that reside 500 kb up-stream and down-stream from the risk SNPs identified in Table 2 in order to identify those in high LD with the risk SNPs. Given that only 10 individuals from the Canadian linkage sample were sequenced, using a larger data set, such as 1000 Genomes, will provide more power to make the LD assessments. Additionally, marker-to-marker LD by population has already been calculated for many of the variants we expect to discover, allowing for a simple filter to be applied as part of our downstream analysis. Before using this resource, the similarity of the genetic background of the Canadian linkage sample and the populations of 1000 Genomes needed to be determined.

In order to assess the ancestry of our sample principal components analysis (PCA) was performed using EIGENSTRAT v6.14 [192]. To prepare for this analysis, HapMap release 23 data was obtained in PLINK format [263]. The pedigree and map files from our 2013 GWAS were merged for chromosomes 1-22. The HapMap map file and our sample map file were compared by Python script. SNPs occurring in both data sets were extracted in each set individually using PLINK v1.07. Strand orientations in our sample were corrected using data derived from Affymetrix releases for the 6.0 array. PLINK was used to merge the HapMap data with our data into a single pedigree/map file set. Because marker-to-marker LD can lead to principal components that are artifacts, PLINK was used to assess LD on the merged file set using the --indep method, with a window of 50 SNPs, a sliding window of 5, and a VIF of 2. SNPs in the prune.in output were extracted. PLINK analysis for LD was run twice more (for a total of 3 runs). PLINK

was then used to assess LD again, this time using the --indep method, with a window of 100 SNPs, a sliding window of 5, and a VIF of 2. SNPs in the prune.in output were extracted. PLINK analysis for LD was run once more (for a total of 2 additional runs) to verify that regions known to contain long-range LD were fully examined [264]. After all pruning, 72,045 SNPs remained on chromosomes 1-22 for analysis. From this LD-pruned file set, two file sets were generated for EIGENSTRAT analysis. In our subset of 10 individuals who were sequenced in two cases, two individuals are from the same pedigree. Because related individuals can influence the results of PCA, these individuals must be separated (**Table 10**).

EIGENSTRAT Run #1	EIGENSTRAT Run #2
001.0026	001.0000
002.0000	002.0000
011.0012	011.0012
029.A038	029.A038
101.0000	101.0000
102.0005	102.0073
105.0000	105.0000
206.0001	206.0001
All of the HapMap Individuals	All of the HapMap Individuals

Table 10: Composition of merged data runs for EIGENSTRAT.
Values listed indicate the pedigree and individual separated by a period for our sample. Two individuals from pedigrees 001 and 102 were sequenced and were subsequently separated for EIGENSTRAT analysis.

The first two eigenvectors from the two PCA returned significant components in the analyses performed using EIGENSTRAT, and were subsequently plotted using R, with separate colors assigned by population (**Figure 5**). Our sequenced individuals are tightly clustered within the CEU population.

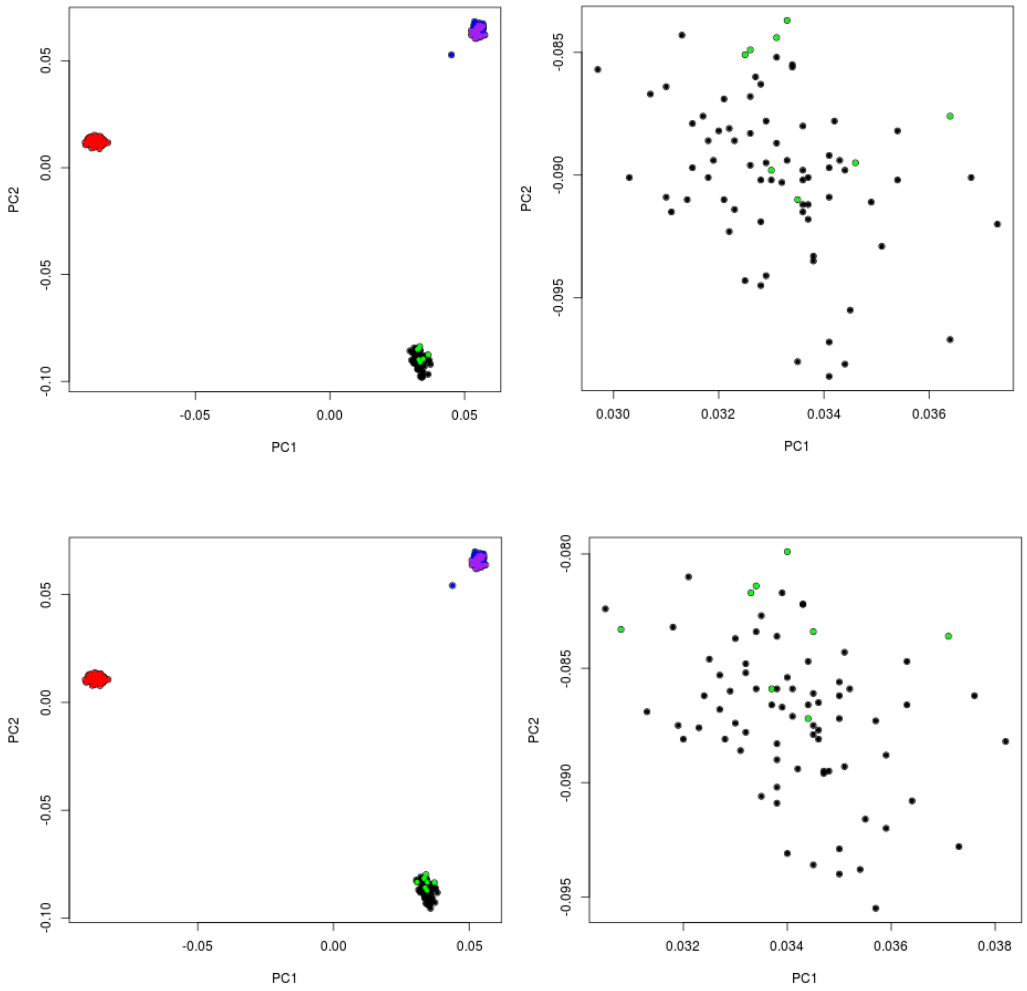

Figure 5: EIGENSTRAT analysis plots.
Run #1 All Populations (top left), Run #1 CEU and Our Sample (top right), Run #2 All Populations (bottom left), Run #2 CEU and Our Sample (bottom right). Red = YRI, Blue = JPT, Purple = CHB, Black = CEU, Green = Sequenced individuals from our sample. Individuals used in each run are described in Table 10.

Chapter 5: Using Risk SNPs from GWAS to Identify Candidate SNVs

Identification of Risk SNPs and Search Parameters

The annotated gvcf file generated by the pipeline described in Chapter 3 was parsed by python script to extract the physical position of each risk SNP identified in Table 2. The script then created a search region of 500 kb upstream and downstream from those physical positions (**Table 11**).

Chrom	SNP	Location (bp)	Search Start	Search End	#Variants
1p21.1	rs7419214	101,859,288	101,359,288	102,359,288	2848
1p13.3	rs17477236	108,124,480	107,624,480	108,624,480	2525
1p13.3	rs641227	111,040,055	110,540,055	111,540,055	2529
1q23.3	rs4656310	161,496,900	160,996,900	161,996,900	3418
1q23.3	rs12725553	162,168,116	161,668,116	162,668,116	3165
1q23.3	rs4411117	162,184,521	161,684,521	162,684,521	3164
2p23.3	rs12991828	27,082,559	26,582,559	27,582,559	2168
2p23.3	rs486582	27,104,131	26,604,131	27,604,131	2153
2q22.3	rs7578749	148,321,813	147,821,813	148,821,813	2018
3p26.1	rs12494654	7,533,393	7,033,393	8,033,393	3798
11q14.1	rs17631231	79,789,629	79,289,629	80,289,629	2647
17q25.1	rs1060120	73,773,000	73,273,000	74,273,000	2707
				TOTAL	**33,140**

Table 11: Risk SNPS to be analyzed.
List of Risk SNPs, their build 37 physical location, the search area for candidate SNPs to be further investigated, and the number of variants within that region to be analyzed. Risk SNP locations were extracted from sequence data, and the search area for candidate SNPs was defined to include 500 kb upstream and downstream from each SNP in order to capture any candidate variants in LD with the risk SNPs.

Some of the search regions overlap, and so the total number of variants (n = 33,140) includes duplicates. Additionally, rs12725553 and rs4411117 are in perfect LD ($r^2 = 1$), so the candidate variants that are generated for both need only be evaluated once.

Nonetheless, even accounting for these duplicates, there are still far too many candidate variants (n = 26,916) to assess them all directly. Variants of interest should be in high LD with the listed risk SNPs. As discussed in Chapter 4, since we only have 10 people genotyped for both the risk SNPs and the candidate variants LD, estimates based purely on our data may be strongly influenced by sampling variation and therefore potentially inaccurate. Since PCA shows high correlation between our sample and the CEU population, we can use that data to draw conclusions about marker to marker LD that should be in applicable to our sample.

The list of 12 risk SNPs was submitted to the rAggr website [265] in order to obtain candidate SNPs 500 kb upstream and downstream from the risk SNPs along with their LD within the CEU population. Though this does not account for all of the variants in our sequenced individuals, for each risk SNP it allowed for evaluation of 82%-93% of the candidates in the designated search region. rAggr is a web-based software for finding SNPs and indels that are in LD with a provided set of markers, using the 1000 Genomes Project and Hapmap genotype databases. It uses an expectation-maximization algorithm adapted from Haploview software [266] to calculate r2 on the fly in real time by the web server. The results were parsed by python script and cross-referenced against the sequence data by matching chromosome, position, reference (REF) allele, and alternative (ALT) allele. If the marker-to-marker LD was below an r^2 of 0.8, the marker was culled from further analysis. If the candidate marker was in high LD with the risk SNP (r^2 >= 0.8), the SNP was cross-referenced against the Affymetrix 6.0 Array SNPs used in the 2013 GWAS. If a SNP had been previously evaluated by the 2013 GWAS, it was also

culled from further analysis. Remaining variants in high LD with risk SNPs (n = 101) were set aside for genotyping (**Table 12**).

Risk SNP	Original Variants	For Genotyping	Culled	Remaining
rs7419214	2848	12	2625	211
rs17477236	2525	0	2166	359
rs641227	2529	0	2297	232
rs4656310	3418	7	2773	638
rs12725553	3165	2	2688	475
rs4411117	3164	2	2676	486
rs12991828	2168	35	1769	364
rs486582	2153	19	1769	365
rs7578749	2018	11	1664	343
rs12494654	3798	8	3512	278
rs17631231	2647	17	2467	163
rs1060120	2707	0	2387	320
Total	**33,140**	**113**	**28,793**	**4234**

Table 12: Summary of candidate SNP LD analysis.
All variants obtained from original parsing of sequence data were assessed for LD within the CEU population against the risk SNPs using rAggr. Candidate variants were culled if they were not in high LD ($r^2 < 0.8$), or were in high LD ($r^2 >= 0.8$), but had already been assessed in the 2013 GWAS described in Chapter 3.

Next, the remaining variants were scanned for monomorphic SNPs. Since our sample correlates to the CEU population, it is expected that population-specific monomorphic sites would result from low mutation rates or positive natural selection. Additionally, given that our sample is comprised of large pedigrees of related individuals from a geographically limited region, it was expected that some of the sequenced variants would be monomorphic in nature. A Python script was used to parse the sequence data and count the occurrences of each allele (REF and ALT) of each SNP. Heterozygotes were counted for both. In the event that a variant had a count of 20 for one allele, and a count of 0 for the other allele, it was added to a list of potentially monomorphic SNPs

(variants with counts below 20 were not subject to this filter, even if all counts were for a single allele). With a sample size of 10, this alone was not enough to definitively call these SNPs monomorphic. These SNPs were submitted to the rAggr website, as described above, and queried against the CEU population. rAggr will not perform analyses if the MAF is 0, and will provide this reason back to the user. Any candidate SNPs with only one allele present in all 10 sequenced individuals that was also found to be monomorphic in the CEU population was culled from further analysis. For each risk SNP, between 11%-55% of the remaining variants were determined to be monomorphic (**Table 13**). Due to the nature of our sample, discussed above, it is likely that some of the SNPs appearing to be monomorphic in our sample, but not in the CEU population, are in fact monomorphic in our entire expanded sample. We therefore expected that some of our genotyped candidate variants may ultimately be uninformative.

Risk SNP	Variants to be Classified	SNPs Culled	Remaining Variants
rs7419214	211	44	167
rs17477236	359	198	161
rs641227	232	83	149
rs4656310	638	62	576
rs12725553	475	164	311
rs4411117	486	168	318
rs12991828	364	175	189
rs486582	365	145	220
rs7578749	343	148	195
rs12494654	278	68	210
rs17631231	163	18	145
rs1060120	320	85	235
Total	**4234**	**1358**	**2876**

Table 13: Summary of analysis for monomorphic SNPs using rAggr.
SNPs with only one allele present in the sequence data were cross-referenced against a list of SNPs known to be monomorphic in the CEU population. These SNPs were removed from further analysis due to being uninformative.

While LD between the remaining variants and the original risk SNPs could be calculated from the set of 10 sequenced individuals, the small sample size would be expected to produce imprecise estimates of LD due to sampling variation. To address this concern, simulations were run in order to determine a 95% CI for the r^2 estimates calculated from a sample size of 10. 100,000 simulation runs were performed for each of six representative minor allele frequencies (0.05, 0.1, 0.2, 0.3, 0.4, and 0.5) using parameters that would generate an r^2 of 0.8 in the population, from which 10 random individuals were drawn to create a sample size of 10. The 95% CI extended to a r^2 of 0.4444. Genotypes were collected for the remaining variants in Table 13 and r^2 computed between those variants and the risk SNPs that referred them. Variants were pruned from further analysis if they produced an r^2 outside of the 95% CI and collected for a later round of genotyping if they produced an r^2 within the 95% CI (**Table 14**).

Risk SNP	Variants To Be Classified	Genotyping	Culled	Remaining
rs7419214	167	7	160	0
rs17477236	161	5	156	0
rs641227	149	8	141	0
rs4656310	576	29	547	0
rs12725553	311	7	304	0
rs4411117	318	7	311	0
rs12991828	189	5	184	0
rs486582	220	3	217	0
rs7578749	195	10	185	0
rs12494654	210	8	202	0
rs17631231	145	4	141	0
rs1060120	235	0	235	0
Total	**2876**	**93**	**2783**	**0**

Table 14: Summary of candidate SNP LD analysis by simulation studies.
All remaining variants were assessed for LD against the risk SNPs using the PPLD. Candidate variants were culled if they were not in moderate LD ($r^2 < 0.4444$). If they were in moderate LD ($r^2 >= 0.4444$), they were set aside for genotyping at a later date.

Genotyping and PPLD|L Analyses

Primers were designed for the 101 SNPs of interest found to be in high LD with the risk SNPs within the CEU population and genotyping was completed on 87 of the 101 SNPs of interest (following the methods described in Chapter 3) at the time of this thesis. Data were cleaned by custom software [267] that computes background and normalizes raw Luminex signals. It then identifies and eliminates ambiguous data points and computes the ratio of corrected median fluorescent intensity (MFI) values between alleles. It then clusters and assigns genotypes to each sample, allowing for real-time viewing and correction. Finally, it combines pedigree and genotype data for error analysis. Resulting genotypes were cleaned as described in Chapter 3. The PPLD|L was reanalyzed, comparing candidate SNPs to risk SNPs. For each candidate SNP/risk SNP comparison, only individuals with a genotype for both SNPs were evaluated, all other genotypes were zeroed out to avoid bias in the results. Six candidate SNPs produced PPLD|L signals of at least 0.1 greater than the risk SNP that referred them (Table 15). All six SNPs exhibited strong marker to marker LD to each other, and to the risk SNP (rs7419214) in our sample (r^2 of 0.923-0.999). LINC01307, a ncRNA, is the closest RefSeq gene to these SNPs and is located ~50,000 bp away using build 37 coordinates.

Risk SNP	PPLD|L	Candidate SNP	PPLD|L	Score Change
rs7419214	0.22	rs4279870	0.74	0.52
rs7419214	0.33	rs12085470	0.80	0.47
rs7419214	0.19	rs59687522	0.43	0.24
rs7419214	0.22	rs12085471	0.38	0.16
rs7419214	0.15	rs10874484	0.29	0.14
rs7419214	0.18	rs12073824	0.28	0.10

Table 15: Six candidate SNPs identified by the PPLD|L.
Six candidate SNPs with stronger PPLD|L scores than the referring risk SNP. All six candidate SNPs and the risk SNP are under the linkage region on chromosome 1p, under the narrow phenotype.

Chapter 6: Conclusions

Concluding Remarks

Schizophrenia is a complex idiopathic neuropsychiatric illness that affects approximately 1% of the general population. Family, twin, and adoption studies indicate a high heritability and strong genetic element to the disease with first degree relatives demonstrating an increased risk of about 10% and monozygotic concordance rates as high as 50%. Schizophrenia susceptibility is likely linked to multiple genetic factors, as evidenced by the fact that patterns of transmission do not match established Mendelian inheritance patterns for single locus disorders, as well as mounting support for a polygenic component [67], [79]. This complexity, along with phenotypic variation, explain why the search for 'schizophrenia genes' remains ongoing to present day. Conflicting evidence continues to accumulate, with candidate genes being identified in some studies, and later questioned or disputed in others (reviewed in [80, 81]). It remains clear, however, that support is present for many different genetic factors playing a role in the predisposition to schizophrenia.

One interpretation of the inconsistent results observed in both linkage and association analyses is that due to factors such as small effect size and uncontrolled phenotypic variation, larger and larger samples will need to be recruited in order for additional susceptibility genes to be discovered [121]. Though studies using extremely large GWAS samples, such as the 2014 study performed by the PGC, have demonstrated that this approach works, we believed that by using a very homogenous sample (in terms of genetic background), and by focusing analysis on regions where linkage and association overlap, which indicates that disease susceptibility in those regions is driving

both signals, we could leverage whole genome sequencing data to identify risk alleles for susceptibility to schizophrenia using a substantially smaller sample.

The work herein identified 12 risk SNPs from areas where strong evidence of linkage and association overlap, and describes how those SNPs were used to search flanking regions for novel candidate variants. Candidate variant pools were subjected to a filtering pipeline which resulted in only 184 variants from 10.7 Mb of sequence analyzed being retained as candidates for further evaluation. Initial genotyping results have identified six SNPs on chromosome 1p under the narrow phenotype, referred by rs7419214, that scored higher on the PPLD|L. Further evaluation is needed to determine the exact role these SNPs play with respect to schizophrenia susceptibility.

Linkage analysis of our sample using SNP markers instead of microsatellites produced six additional regions to be examined, which in turn produced 27 risk SNPs from six discrete genomic regions. These risk SNPs will be evaluated using the procedures described in Chapter 5 to determine if any new candidate SNPs should be pursued.

Future Work

Analyze New Risk SNPs identified by SNP PPL

Linkage regions determined from analysis of the build 37 SNP Genome Scan were examined for PPLD|L scores >= 0.2 in the 2013 GWAS in order to identify additional risk SNPs, in the same manner the microsatellite Genome Scan was used to produce the first group of risk SNPs examined in Chapter 5 (**Table 16**).

Chrom	Phenotype	PPL_{max}	Risk_SNP	PPLD|L	Gene
1	b	0.28	rs12040131	0.45	None
1	b	0.28	rs6600275	0.24	None
1	b	0.28	rs7530233	0.36	None
1	b	0.28	rs7523169	0.39	None
1	b	0.28	rs7534508	0.42	None
1	b	0.28	rs11586139	0.39	None
1	b	0.28	rs11588364	0.39	None
1	b	0.28	rs945338	0.39	None
1	b	0.28	rs2256090	0.39	None
1	b	0.28	rs4660469	0.4	*KCNQ4*
1	n	0.79	rs6428604	0.31	None
1	n	0.79	rs681589	0.45	*HFM1*
1	n	0.79	rs7417055	0.81	*HFM1*
1	n	0.79	rs281979	0.35	*HFM1*
1	n	0.79	rs281935	0.28	*HFM1*
1	n	0.79	rs17131417	0.28	*HFM1*
1	n	0.79	rs4658221	0.33	*HFM1*
7	b	0.41	rs10807764	0.2	None
7	b	0.41	rs10260665	0.27	None
7	b	0.41	rs7784685	0.47	None
7	b	0.41	rs12154666	0.3	None
10	b	0.32	rs12764660	0.53	*RBM20*
10	b	0.32	rs12414939	0.22	*RBM20*
10	b	0.32	rs1341053	0.33	None
15	b	0.55	rs11259948	0.2	*FSD2*
15	b	0.55	rs17273206	0.23	*BLM*
19	b	0.29	rs1030687	0.22	*ZNF536*

Table 16: New risk SNPs identified from the SNP PPL Scan.
New risk SNPs identified by searching under the linkage regions identified by the SNP PPL for GWAS SNPs with a PPLD|L => 20% and the gene (if applicable) each SNP is located in according to dbSNP.

Additional Genes of Interest

KCNQ4

KCNQ4 (potassium channel, voltage-gated, subfamily Q, member 4) is located on chromosome 1p34.2 and is one of five voltage-dependent potassium channels composed of homo- and heterotetrameric complexes of five KCNQ subunits. KCNQ channels (with

the exception of KCNQ1), are broadly expressed in neuronal tissue, including neocortex and hippocampus [268]. Activation of the KCNQ channels is responsible for the initiation of the M current, which inhibits the K^+ current that modulates neuronal excitability [269]. Reduction in KCNQ channel activity as a result of genetic mutation has been implicated in epilepsy, progressive hearing loss, and bipolar disorder [268, 270]. Due to their important role in controlling neuronal excitability, KCNQ channels have become attractive targets for treatment of neurological disorders linked to hyperexcitability and compounds have been proposed for therapeutic potential for the both cognitive and positive symptoms of schizophrenia [271]. Additionally, dopaminergic neurons in the ventral tegmental area express KCNQ4 channels. As discussed in Chapter 1, psychotic symptoms have been shown to be associated with an increased excitability of dopamine cells, and as a result, treatments with KCNQ channel openers may serve as a potential new class of antipsychotics [272]. KCNQ4 is an attractive candidate for susceptibility to schizophrenia.

HFM1

HFM1 (helicase family member 1) is located on chromosome 1p22.2 and is expressed in germline cells, where it is believed to play a role in genome integrity. Formation of most crossover events requires the help of a group of proteins known as ZMM, and HFM1 is in this group. HFM1 is required for normal evolution of homologous recombination and proper synapsis between homologous chromosomes in a number of model organisms (reviewed in [273]). A de novo variant in HFM1 was

identified in Chinese schizophrenic patients during prenatal development, making it a possible candidate for susceptibility to schizophrenia [274].

RMB20

RMB20 (RNA binding motif protein 20) is located on chromosome 10q25. It binds RNA and regulates splicing, and is highly expressed in the heart.

FSD2

FSD2 (fibronectin type III and SPRY containing domain 2) is located on chromosome 15q25.2 and it encodes a protein in the FN3/SPRY family. Alternate splicing leads to multiple transcript variants. A TRIM-related protein minispryn encoded by FSD2 has been demonstrated to exhibit extensive sequence similarity with the C-terminus of myospryn [275]. Myospryn is encoded by CMYA5, which has been identified as a possible risk gene for schizophrenia and major depressive disorder [276, 277]. CMYA5 and FSD2 seem to have originated from chromosome duplication and located within evolutionarily-conserved gene clusters on different chromosomes [275]. Myospryn has also been demonstrated to be a binding partner for Dysbindin in muscle [278]. Dysbindin has been extensively investigated with respect to schizophrenia (reviewed in [279]). Though there is no literature directly linking FSD2 and schizophrenia, the similarities between this gene and CMYA5 warrant further investigation.

BLM

BLM (Bloom symdrome RecQ like helicase) is located on chromosome 15q26.1 and is related to the RecQ subset of DExH box-containing DNA helicases. Mutations causing Bloom syndrome have been shown to delete or alter helicase motifs and may disable the 3'-5' helicase activity. The normal protein is believed to suppress inappropriate recombination. Bloom Syndrome is one of a limited number of rare hereditary diseases marked by genetic defects of DNA repair mechanisms. Though these disorders may present differently, there is overlap in clinical features such as neurological disorders (reviewed in [280]). As such, genes that cause these rare hereditary disorders may play a role in neurological disorders, such as schizophrenia.

ZNF536

ZNF536 (zinc finger protein 536) is located on chromosome 19q12 and encodes a high conserved zinc finger protein. The protein is most abundant in the brain where it negatively regulates neuronal differentiation, which makes ZNF536 an attractive candidate susceptibility gene for schizophrenia.

Identification of Causal SNPs

In each linkage region, we are interested in identifying the variants that produce the strongest evidence for LD with the disease phenotype. Finishing this work includes:

1. *Genotyping the remaining 14 SNPs from the first batch of 101 candidate SNPS.*
    - The 101 SNPs identified in Chapter 5 have been completed preliminary genotyping, with 87 passing data cleaning as described in Chapter 3. The

remaining SNPs are currently being re-processed, and will be assessed using the PPLD|L when completed.

2. *Genotyping the unique candidate (n=83) variants that could not be excluded by r^2 simulations identified in Table 14*.

    - The 83 additional candidate variants identified for genotyping will be analyzed as a second batch. We will follow the same procedures described in Chapter 5 for the first batch, which includes primer design, genotyping, cleaning of genotype data, and re-analysis by the PPLD|L.

3. *Conduct recursive analysis on the 6 new SNPs, and any others of interest determine from steps 1 and 2, above.*

    - As discussed at the end of Chapter 5, six SNPs produced higher PPLD|L scores than the risk SNP that referred them that lie within linkage regions identified using microsatellite markers. In order to further investigate these new SNPs and any others that produce higher PPLD|L scores that are generated by the additional analyses described in steps 1 and 2 above, we will first perform recursive analysis using the methods described in Chapter 5 to search 500 kb up-stream and down-stream from these locations to rule out additional candidates, which will add approximately 4,000 bp to our original search criteria. We will also perform this analysis on the new risk SNPs that lie within linkage regions identified by the genome scan conducted with the SNP markers.

Next, we need to construct a set of candidate variants for further evaluation. This set will include:

1. Original risk SNPs where no variants with stronger LD was found
2. New variants that produce about the same PPLD as the risk SNPs (the 0 to 0.1 increase in PPLD group)
3. New variants that are even better than the original risk SNPs (and replace the original SNP)

All of these may be candidates for functional evaluation. Before we proceed to laboratory evaluation of function, we can perform some bioinformatics work to help prioritize these candidates.

We can use information generated by Kelvin to build a profile of what a causal variant might look like. Because Kelvin retains many aspects of parametric linkage analysis in the course of its computation, examination of the penetrance vectors produced in the calculations for the PPLD|L can be used to construct an expected inheritance model which then can be compared to the pattern of segregation of the candidate variant in our sample. We can also look at the Disease Gene Frequency (DGF) generated by Kelvin and compare it to our candidate SNP minor allele frequency. If they are not similar this may indicate that our candidate SNPs are not causal.

We will also look at evolutionary conservation with the aim of determining if any work on model organisms can help guide prioritization of these candidates. The initial six new candidates are outside of coding regions, in relative ‘gene deserts’. These areas can be particularly prone to long-range LD spanning greater than 500kb, so we may need

to expand our search to include a much larger region. If we discover additional candidate variants the specific laboratory analyses that would need to be done would be tailored to the potential function of the variants, e.g. a variant that causes a missense mutation in a coding region might be tested for nonsense mediated decay or, if a product is made, loss of specific protein function, whereas a variant that is predicted to alter an enhancer sequence might be analyzed for the potential to alter gene expression using a luciferase assay. Testing multiple functional variants using multiple different methods will be costly, but since the cost of sequencing has significantly declined, it will be likely be more economical to first sequence additional individuals to obtain additional data that should help further cull the list of potential candidates.

Sequencing More Individuals

Initial sequencing for this project was performed in 2012. At that time, the cost to sequence a single genome was fairly high, but it has since decreased dramatically, allowing for more individuals from our sample to be sequenced (**Figure 6**).

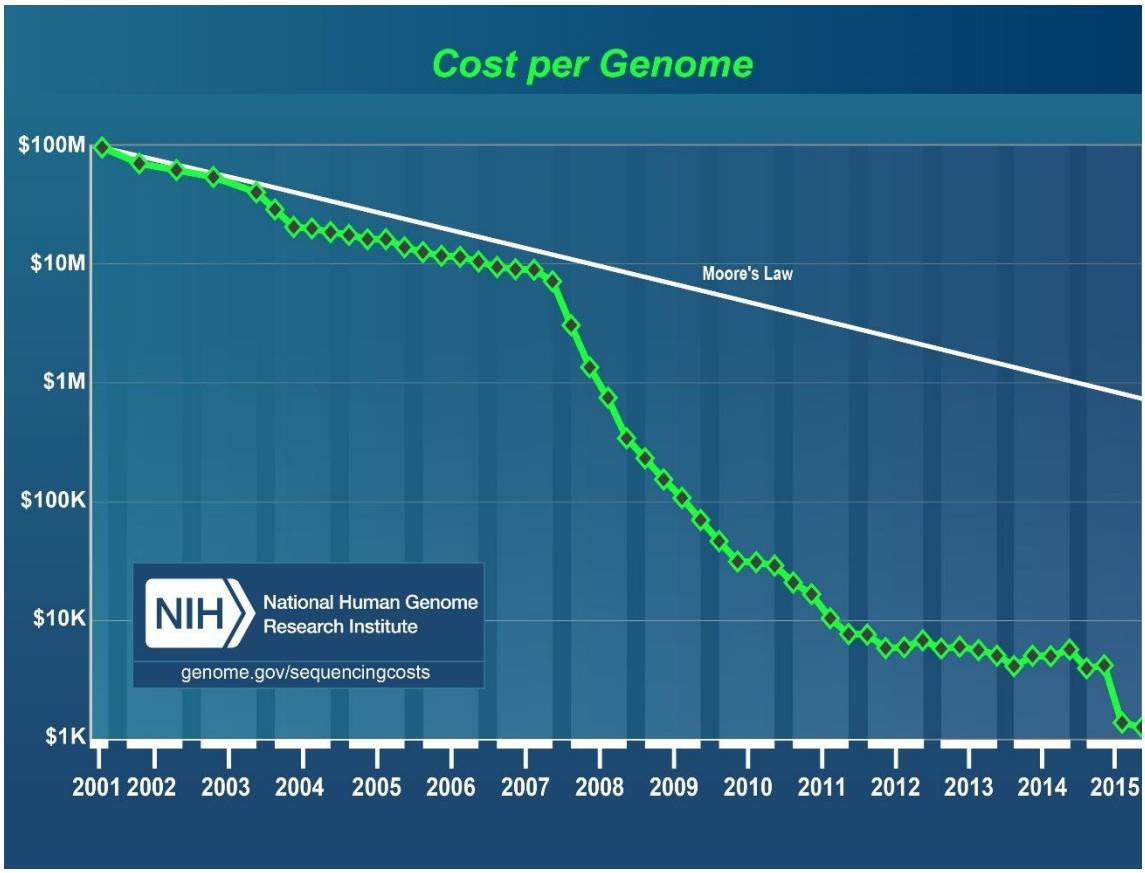

Figure 6: NHGRI cost per enome for individual whole genome sequencing. The precipitous drop in cost that occurred between 2007 and 2008 was the result of major technological advances. These advances not only reduced the cost, but also significantly improved the turn-around time for results.

As discussed in Chapter 3, initially individuals were selected for sequencing based on three criteria: 1) coming from the core portion of one of the larger pedigrees, 2) being a patient with a 'typical' case of schizophrenia, featuring diagnostic stability and phenotypic homogeneity with other sample subjects, and 3) having high quality DNA available to send out for sequencing. New individuals to sequence were selected under the following criterion: 1) At least one sequenced individual in family, 2) Affected under the phenotype tested, 3) Individual contributes to the peak in the same way that the previously sequenced individual does, and 4) DNA is available for analysis.

To identify candidate individuals for sequencing, the original 22 pedigrees for which microsatellite data were available was first divided into the individual pedigrees, and each family was analyzed by itself to assess its contribution to the largest peak on Chromosome 1, using the Build 37 Microsatellite Genome Scan, to determine if enough power exists within that pedigree to reach a PPL score indicative of linkage.

For example, each of the eight families with sequenced individuals were analyzed for the interval 167-171 cM on chromosome 1, under the narrow phenotype. Pedigrees 001 and 102 each produced PPL scores >= 10% when analyzed by themselves (**Table 17**).

Pedigree	**001**	**002**	**011**	**029**	**101**	**102**	**105**	**206**
Position	**PPL**							
167	0.09	0.04	0.016	0.019	0.018	0.13	0.016	0.03
168	0.09	0.04	0.016	0.023	0.017	0.14	0.016	0.03
169	0.16	0.03	0.016	0.03	0.019	0.10	0.016	0.03
170	0.16	0.022	0.016	0.05	0.018	0.09	0.016	0.03
171	0.03	0.017	0.017	0.09	0.018	0.024	0.017	0.03

Table 17: Chromosome 1 results by family under the narrow phenotype. Pedigrees 001 and 102 are the only families that achieved a PPL score >= 10% when analyzed by themselves, demonstrating that they have the power to produce strong linkage signals.

Pedigree 001 has 37 individuals with 14 coded as affected under the narrow phenotype. Pedigree 102 has 21 individuals with 4 coded as affected under the narrow phenotype.

The by-family microsatellite Genome Scan was performed using build 37 positions and the results were analyzed by Python script. Any location where a single family generated a PPL score >= 10% was further evaluated. In the additional analysis,

each individual's affection status was sequentially set to zero and the family score was re-evaluated. If a family PPL score increased with the removal of an individual he or she was determined to be providing evidence against linkage, whereas if the family PPL score decreased with the removal of an individual here or she was determined to be providing evidence for linkage. For each peak by family the affected individual most closely trending with the previously sequenced individual was selected as a possible candidate for sequencing (**Table 18**). Peaks present on the SNP Genome Scan described in Chapter 4 were given priority when determining sequence candidates. It will also be necessary to sequence strategically selected unaffected individuals from these families. This will help us to quickly identify variants that may be monomorphic within our sample. It will also prevent classifying a variant as potentially causal if it is seen in many unaffected individuals.

Peak	Pos (cM)	Ped	Family Score	After Drop	Change	Candidate
ch01n	170	001	0.16	0.09	0.07	001.0115
ch01n	168	102	0.14	0.06	0.08	***102.0073***
ch02b	17	029	0.19	0.11	0.08	029.A039
ch03n	24	102	0.15	0.07	0.08	***102.0073***
ch05b	99	029	0.13	0.08	0.05	029.1029
ch05n	10	102	0.11	0.06	0.05	***102.0073***
ch06b	85	102	0.1	0.06	0.04	102.0079
ch07b	21	029	0.32	0.20	0.12	029.1029
ch07n	50	102	0.14	0.06	0.08	***102.0073***
ch08b	42	029	0.15	0.07	0.08	029.0060
ch10b	137	105	0.17	0.08	0.09	105.1067
ch13b	112	206	0.21	0.10	0.11	206.0004
ch13n	112	206	0.21	0.10	0.11	206.0004
ch14b	85	011	0.17	0.09	0.08	011.0020
ch17b	136	011	0.12	0.07	0.05	011.0018
ch18b	5	105	0.11	0.05	0.06	105.1000
ch21b	11	029	0.13	0.05	0.08	029.0000
ch23b	132	105	0.13	0.05	0.08	105.0001
ch23n	130	105	0.18	0.09	0.09	105.1090

Table 18: Results of analysis of family peaks.
Results of analysis of family peaks with a PPL >= 10% following sequentially zeroing out affected status to ascertain the best sequence targets using criteria described above. Peak includes chromosome and phenotype, where n = narrow and b = broad. Position in cM for PPL max is given, along with the pedigree number generating the Family Score in the following column. After Drop indicates the new score when the Candidate is zeroed out. The Change column indicates the difference between the family score with the individual included and excluded, demonstrating these individuals are contributing to the peaks they have been selected for. Candidate 102.0073 is in bold italics because that individual has already been sequenced.

Another option for additional sequencing targets would be to simply sequence an entire pedigree. In this case, Pedigree 102 is an attractive candidate. Pedigree 102 was able to produce five separate PPL scores >=10%, four of which occur under the narrow phenotype allowing for better homogeneity with respect to diagnosis. Additionally, Pedigree 102 has an uncommon event in it: Two brothers from one nuclear family

married two sisters from another nuclear family. As a result, though their offspring are technically cousins, genetically they are closer to siblings in terms of relatedness.

Sequencing more individuals will benefit many areas of the remaining work to be done, including: recursive analyses where LD is evaluated, functional analyses of proposed candidate variants, and the collection of information on unaffected family members so that direct comparisons can be between individuals with shared genetic backgrounds.

Schizophrenia is a complex mental health disorder that has yielded conflicting results in both linkage and association analysis. This thesis has demonstrated an approach that can be successful using smaller samples. The identification of such an approach holds promise for greater understanding of the complex genetic architecture underlying this devastating disease.

REFERENCES

1. Hippius, H. and N. Muller, *The work of Emil Kraepelin and his research group in Munchen.* Eur Arch Psychiatry Clin Neurosci, 2008. **258 Suppl 2**: p. 3-11.
2. Fusar-Poli, P. and P. Politi, *Paul Eugen Bleuler and the birth of schizophrenia (1908).* Am J Psychiatry, 2008. **165(11)**: p. 1407.
3. Saha, S., et al., *A systematic review of the prevalence of schizophrenia.* PLoS Med, 2005. **2**(5): p. e141.
4. Tandon, R., et al., *Definition and description of schizophrenia in the DSM-5.* Schizophr Res, 2013. **150**(1): p. 3-10.
5. Buchanan, R.W. and W.T. Carpenter, *Domains of Psychopathology: An Approach to the Reduction of the Heterogeneity of in Schizophrenia.* J Nerv Ment Dis, 1994. **182(4)**: p. 193-204.
6. Millan, M.J., et al., *Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy.* Nat Rev Drug Discov, 2012. **11**(2): p. 141-168.
7. Dietsche, B., T. Kircher, and I. Falkenberg, *Structural brain changes in schizophrenia at different stages of the illness: A selective review of longitudinal magnetic resonance imaging studies.* Aust N Z J Psychiatry, 2017. **51**(5): p. 500-508.
8. Brady, K.T. and R. Sinha, *Co-occurring mental and substance use disorders: the neurobiological effects of chronic stress.* Am J Psychiatry, 2005. **162**(8): p. 1483-93.
9. Laskaris, L.E., et al., *Microglial activation and progressive brain changes in schizophrenia.* British Journal of Pharmacology, 2016. **173**(4): p. 666-680.
10. Organization, W.H., *Global Health Estimates 2015: Burden of diseases by Cause, Age, Sex, by Country and Region, 2000-2015.* 2016. **Geneva, World Health Organization; 2016**.
11. Chong, H.Y., et al., *Global economic burden of schizophrenia: a systematic review.* Neuropsychiatric Disease and Treatment, 2016. **12**: p. 357-373.
12. Tajima-Pozo, K., et al., *Understanding the direct and indirect costs of patients with schizophrenia.* F1000Res, 2015. **4**: p. 182.
13. McEvoy, J.P., *The costs of schizophrenia.* J Clin Psychiatry, 2007. **68 Suppl 14**: p. 4-7.
14. Üstün, T.B., *The Global Burden of Mental Disorders.* American Journal of Public Health, 1999. **89(9)**(September): p. 1315-1318.
15. Sham, P.C., C.J. MacLean, and K.S. Kendler, *A typological model of schizophrenia based on age at onset, sex, and familial morbidity.* Acta Psychiatr Scand, 1994. **89(2)**: p. 135-41.
16. Howes, O., R. McCutcheon, and J. Stone, *Glutamate and dopamine in schizophrenia: an update for the 21(st) century.* Journal of psychopharmacology (Oxford, England), 2015. **29**(2): p. 97-115.
17. Lieberman, J.A., J.M. Kane, and J. Alvir, *Provocative tests with psychostimulant drugs in schizophrenia.* Psychopharmacology, 1987. **91**: p. 415-33.
18. Carlsson, A., M. Lindqvist, and T. Magnusson, *3,4-Dihyroxyphenylalanine and 5-hydroxytryptophan as reserpine antagonists.* Nature, 1957. **180**: p. 1200.
19. Creese, I., D. Burt, and S. Snyder, *Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs.* Science, 1976. **80**: p. 481-483.
20. Seeman, P. and T. Lee, *Antipsychotic drugs: direct correlation between clinical potency and presynaptic action on dopamine neurons.* Science, 1975. **188**: p. 1217-1219.
21. Seeman, P., et al., *Antipsychotic drug doses and neuroleptic/dopamine receptors.* Nature, 1976. **261**: p. 717-719.

22. Davis, K.L., et al., *Dopamine in schizophrenia: a review and reconceptualization.* Am J Psychiatry, 1991. **148**(11): p. 1474-86.
23. McGuire, P., et al., *Functional neuroimaging in schizophrenia: diagnosis and drug discovery.* Trends Pharmacol Sci, 2008. **29**(2): p. 91-8.
24. Howes, O.D. and S. Kapur, *A neurobiological hypothesis for the classification of schizophrenia: type A (hyperdopaminergic) and type B (normodopaminergic).* Br J Psychiatry, 2014. **205**(1): p. 1-3.
25. Kapur, S., et al., *Relationship between dopamine D(2) occupancy, clinical response, and side effects: a double-blind PET study of first-episode schizophrenia.* Am J Psychiatry, 2000. **157**(4): p. 514-20.
26. Mortimer, A.M., et al., *Clozapine for treatment-resistant schizophrenia: National Institute of Clinical Excellence (NICE) guidance in the real world.* Clin Schizophr Relat Psychoses, 2010. **4**(1): p. 49-55.
27. Murphy, B.P., et al., *Pharmacological treatment of primary negative symptoms in schizophrenia: a systematic review.* Schizophr Res, 2006. **88**(1-3): p. 5-25.
28. Bloomfield, M.A., et al., *Dopaminergic function in cannabis users and its relationship to cannabis-induced psychotic symptoms.* Biol Psychiatry, 2014. **75**(6): p. 470-8.
29. Mizrahi, R., et al., *Stress-induced dopamine response in subjects at clinical high risk for schizophrenia with and without concurrent cannabis use.* Neuropsychopharmacology, 2014. **39**(6): p. 1479-89.
30. Thompson, J.L., et al., *Striatal dopamine release in schizophrenia comorbid with substance dependence.* Mol Psychiatry, 2013. **18**(8): p. 909-15.
31. Kitzinger, H. and D.G. Arnold, *A preliminary study of the effects of glutamic acid on catatonic schizophrenics.* Rorschach Res Exch J Proj Tech, 1949. **13**: p. 210-218.
32. Kim, J.S., et al., *Low cerebrospinal fluid glutamate in schizophrenia patients and a new hypothesis on schizophrenia.* Neurosci Lett, 1980. **20**: p. 379-382.
33. Korpi, E.R., C.A. Kaufmann, and D.R. Weinberger, *Cerebrospinal fluid amino acid concentration in chronic schizphrenia.* Psychiatry Res., 1987. **20**: p. 337-345.
34. Perry, T.L., *Normal cerebrospinal and brain glutamate levels in schizophrenia do not support the hypothesis of glutamatergic neuronal dysfunction* Neurosci Lett, 1982. **28**: p. 81-85.
35. Rothman, D.L., et al., *In vivo NMR studies of the glutamate neurotransmitter flux and neuroenergetics: implications for brain function.* Annu Rev Physiol, 2003. **65**: p. 401-27.
36. Kew, J.N. and J.A. Kemp, *Ionotropic and metabotropic glutamate receptor structure and pharmacology.* Psychopharmacology (Berl), 2005. **179**(1): p. 4-29.
37. Stone, J.M., P.D. Morrison, and L.S. Pilowsky, *Glutamate and dopamine dysregulation in schizophrenia--a synthesis and selective review.* J Psychopharmacol, 2007. **21**(4): p. 440-52.
38. Sokolov, B.P., *Expression of NMDAR1, GluR1, GluR7, and KA1 glutamate receptor mRNAs is decreased in frontal cortex of "neuroleptic-free" schizophrenics: evidence on reversible up-regulation by typical neuroleptics.* J Neurochem, 1998. **71**(6): p. 2454-64.
39. McCullumsmith, R.E., et al., *Recent advances in targeting the ionotropic glutamate receptors in treating schizophrenia.* Curr Pharm Biotechnol, 2012. **13**(8): p. 1535-42.
40. Funk, A.J., et al., *Decreased expression of NMDA receptor-associated proteins in frontal cortex of elderly patients with schizophrenia.* Neuroreport, 2009. **20**(11): p. 1019-22.
41. Hammond, J.C., et al., *Evidence of Glutamatergic Dysfunction in the Pathophysiology of Schizophrenia*, in *Synaptic Stress and Pathogenesis of Neuropsychiatric Disorders*, M.

Popoli, D. Diamond, and G. Sanacora, Editors. 2014, Springer New York: New York, NY. p. 265-294.
42. Krystal, J.H., et al., *Subanesthetic effects of the noncompetitive NMDA antagonist, ketamine, in humans. Psychotomimetic, perceptual, cognitive, and neuroendocrine responses.* Arch Gen Psychiatry, 1994. **51**(3): p. 199-214.
43. Javitt, D.C., *Glutamate and schizophrenia: phencyclidine, N-methyl-D-aspartate receptors, and dopamine-glutamate interactions.* Int Rev Neurobiol, 2007. **78**: p. 69-108.
44. Morgan, C.J. and H.V. Curran, *Acute and chronic effects of ketamine upon human memory: a review.* Psychopharmacology (Berl), 2006. **188**(4): p. 408-24.
45. Pilowsky, L.S., et al., *First in vivo evidence of an NMDA receptor deficit in medication-free schizophrenic patients.* Mol Psychiatry, 2006. **11**(2): p. 118-9.
46. Poels, E.M., et al., *Glutamatergic abnormalities in schizophrenia: a review of proton MRS findings.* Schizophr Res, 2014. **152**(2-3): p. 325-32.
47. Levitt, J.J., et al., *A selective review of volumetric and morphometric imaging in schizophrenia.* Curr Top Behav Neurosci, 2010. **4**: p. 243-81.
48. Spencer, A.E., et al., *Glutamatergic dysregulation in pediatric psychiatric disorders: a systematic review of the magnetic resonance spectroscopy literature.* J Clin Psychiatry, 2014. **75**(11): p. 1226-41.
49. Miyake, N., et al., *Imaging changes in glutamate transmission in vivo with the metabotropic glutamate receptor 5 tracer [11C] ABP688 and N-acetylcysteine challenge.* Biol Psychiatry, 2011. **69**(9): p. 822-4.
50. Sarter, M., C. Lustig, and S.F. Taylor, *Cholinergic contributions to the cognitive symptoms of schizophrenia and the viability of cholinergic treatments.* Neuropharmacology, 2012. **62**(3): p. 1544-53.
51. Nakazawa, K., et al., *GABAergic interneuron origin of schizophrenia pathophysiology.* Neuropharmacology, 2012. **62**(3): p. 1574-1583.
52. Devor, A., et al., *Genetic evidence for role of integration of fast and slow neurotransmission in schizophrenia.* Mol Psychiatry, 2017. **22**(6): p. 792-801.
53. *Schizophrenia Spectrum and Other Psychotic Disorders*, in *DSM-5® Clinical Cases*.
54. Carpenter, W.T. and R. Tandon, *Psychotic disorders in DSM-5: summary of changes.* Asian J Psychiatr, 2013. **6**(3): p. 266-8.
55. Bromet, E.J., et al., *Diagnostic shifts during the decade following first admission for psychosis.* Am J Psychiatry, 2011. **168**(11): p. 1186-94.
56. Korver-Nieberg, N., et al., *The validity of the DSM-IV diagnostic classification system of non-affective psychoses.* Aust N Z J Psychiatry, 2011. **45**(12): p. 1061-8.
57. Linscott, R.J., J. Allardyce, and J. van Os, *Seeking verisimilitude in a class: a systematic review of evidence that the criterial clinical symptoms of schizophrenia are taxonic.* Schizophr Bull, 2010. **36**(4): p. 811-29.
58. Jager, M., et al., *[Deconstructing schizophrenia. Dimensional models or division into subtypes?].* Nervenarzt, 2012. **83**(3): p. 345-54.
59. Krueger, R.F. and K.E. Markon, *A dimensional-spectrum model of psychopathology: Progress and opportunities.* Archives of General Psychiatry, 2011. **68**(1): p. 10-11.
60. Gurvich, C. and S.L. Rossell, *Editorial: Cognition Across the Psychiatric Disorder Spectrum: From Mental Health to Clinical Diagnosis.* Frontiers in Psychiatry, 2015. **6**: p. 110.
61. Geschwind, D.H. and M.W. State, *Gene hunting in autism spectrum disorder: on the path to precision medicine.* Lancet Neurol, 2015. **14**(11): p. 1109-20.

62. Hafeman, D.M., et al., *Toward the Definition of a Bipolar Prodrome: Dimensional Predictors of Bipolar Spectrum Disorders in At-Risk Youths.* Am J Psychiatry, 2016. **173**(7): p. 695-704.
63. *Schizophrenia Spectrum and Other Psychotic Disorders*, in *Diagnostic and Statistical Manual of Mental Disorders*. 2013, American Psychiatric Association.
64. Rossler, W., et al., *Size of burden of schizophrenia and psychotic disorders.* Eur Neuropsychopharmacol, 2005. **15**(4): p. 399-409.
65. Smyth, A.M. and S.M. Lawrie, *The neuroimmunology of schizophrenia.* Clin Psychopharmacol Neurosci, 2013. **11**(3): p. 107-17.
66. Dickinson, D. and P.D. Harvey, *Systemic hypotheses for generalized cognitive deficits in schizophrenia: a new take on an old problem.* Schizophr Bull, 2009. **35**(2): p. 403-14.
67. Harrison, P.J. and D.R. Weinberger, *Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence.* Mol Psychiatry, 2005. **10**(1): p. 40-68; image 5.
68. Lieberman , J.A., et al., *Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia.* New England Journal of Medicine, 2005. **353**(12): p. 1209-1223.
69. Marder, S.R., et al., *Antipsychotic medications.* The American Psychiatric Press Textbook of Psychopharmacology., 1995. **American Psychiatric Press: Washington, DC**: p. 247-261.
70. Miyamoto, S., et al., *Pharmacological treatment of schizophrenia: a critical review of the pharmacology and clinical effects of current and future therapeutic agents.* Mol Psychiatry, 2012. **17**(12): p. 1206-27.
71. Kane, J., et al., *Clozapine for the treatment-resistant schizophrenic. A double-blind comparison with chlorpromazine.* Arch Gen Psychiatry, 1988. **45**(9): p. 789-96.
72. Miyake, N., S. Miyamoto, and L.F. Jarskog, *New serotonin/dopamine antagonists for the treatment of schizophrenia: are we making real progress?* Clin Schizophr Relat Psychoses, 2012. **6**(3): p. 122-33.
73. Charrier, N., K. Chevreul, and I. Durand-Zaleski, *[The cost of schizophrenia: a literature review].* Encephale, 2013. **39 Suppl 1**: p. S49-56.
74. Rogoz, Z., *Combined treatment with atypical antipsychotics and antidepressants in treatment-resistant depression: preclinical and clinical efficacy.* Pharmacol Rep, 2013. **65**(6): p. 1535-44.
75. Tsuang, M.T., W.S. Stone, and S.V. Faraone, *Genes, environment and schizophrenia.* The British Journal of Psychiatry, 2001. **178**(suppl. 40): p. s18-s24.
76. Tsuang, M.T., W.S. Stone, and S.V. Faraone, *Schizophrenia: family studies and treatment of spectrum disorders.* Dialogues in Clinical Neuroscience, 2000. **2**(4): p. 381-391.
77. Esterberg, M.L., S.M. Goulding, and E.F. Walker, *A Personality Disorders: Schizotypal, Schizoid and Paranoid Personality Disorders in Childhood and Adolescence.* J Psychopathol Behav Assess, 2010. **32**(4): p. 515-28.
78. Gottesman, I., *Schizophrenia Genesis: The Origins of Madness.* New York: W. H. Freeman and Co, 1991.
79. Purcell, S.M., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.
80. Farrell, M.S., et al., *Evaluating historical candidate genes for schizophrenia.* Mol Psychiatry, 2015. **20**(5): p. 555-62.
81. Allen, N.C., et al., *Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database.* Nat Genet, 2008. **40**(7): p. 827-34.

82. Baker, K.D. and D.H. Skuse, *Adolescents and young adults with 22q11 deletion syndrome: psychopathology in an at-risk group.* Br J Psychiatry, 2005. **186**: p. 115-20.
83. Schizophrenia Working Group of the Psychiatric Genomics, C., *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-427.
84. Torres, F., M. Barbosa, and P. Maciel, *Recurrent copy number variations as risk factors for neurodevelopmental disorders: critical overview and analysis of clinical implications.* J Med Genet, 2016. **53**(2): p. 73-90.
85. Rutkowski, T.P., et al., *Unraveling the genetic architecture of copy number variants associated with schizophrenia and other neuropsychiatric disorders.* J Neurosci Res, 2017. **95**(5): p. 1144-1160.
86. Cnv and C. Schizophrenia Working Groups of the Psychiatric Genomics, *Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects.* Nat Genet, 2017. **49**(1): p. 27-35.
87. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.* Science, 2008. **320**(5875): p. 539-43.
88. Georgieva, L., et al., *De novo CNVs in bipolar affective disorder and schizophrenia.* Hum Mol Genet, 2014. **23**(24): p. 6677-83.
89. Sanders, S.J., et al., *Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci.* Neuron, 2015. **87**(6): p. 1215-33.
90. Ghani, M., et al., *Genome-wide survey of large rare copy number variants in Alzheimer's disease among Caribbean hispanics.* G3 (Bethesda), 2012. **2**(1): p. 71-8.
91. Pulst, S.M., *Genetic linkage analysis.* Archives of Neurology, 1999. **56**(6): p. 667-672.
92. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-Wide Association Studies.* PLoS Computational Biology, 2012. **8**(12): p. e1002822.
93. Reich, T., P.J. Clayton, and G. Winokour, *Family history studies: V. The genetics of mania.* Am J Psychiatry, 1969. **125**: p. 1358-1369.
94. Zhao, L.S., et al., *[Genome-wide linkage scan for an ethnic Han Chinese pedigree affected with schizophrenia].* Zhonghua Yi Xue Yi Chuan Xue Za Zhi, 2013. **30**(1): p. 5-8.
95. Alkelai, A., et al., *Identification of new schizophrenia susceptibility loci in an ethnically homogeneous, family-based, Arab-Israeli sample.* Faseb j, 2011. **25**(11): p. 4011-23.
96. Francks, C., et al., *Population-based linkage analysis of schizophrenia and bipolar case-control cohorts identifies a potential susceptibility locus on 19q13.* Mol Psychiatry, 2010. **15**(3): p. 319-25.
97. Hong, K.S., et al., *Genome-widely significant evidence of linkage of schizophrenia to chromosomes 2p24.3 and 6q27 in an SNP-Based analysis of Korean families.* Am J Med Genet B Neuropsychiatr Genet, 2009. **150b**(5): p. 647-52.
98. Hamshere, M.L., et al., *Mood-incongruent psychosis in bipolar disorder: conditional linkage analysis shows genome-wide suggestive linkage at 1q32.3, 7p13 and 20q13.31.* Bipolar Disord, 2009. **11**(6): p. 610-20.
99. Schwab, S.G., et al., *Genome-wide scan in 124 Indonesian sib-pair families with schizophrenia reveals genome-wide significant linkage to a locus on chromosome 3p26-21.* Am J Med Genet B Neuropsychiatr Genet, 2008. **147b**(7): p. 1245-52.
100. Holliday, E.G., B.J. Mowry, and D.R. Nyholt, *A reanalysis of 409 European-Ancestry and African American schizophrenia pedigrees reveals significant linkage to 8p23.3 with evidence of locus heterogeneity.* Am J Med Genet B Neuropsychiatr Genet, 2008. **147b**(7): p. 1080-8.

101. Merette, C., et al., *Replication of linkage with bipolar disorder on chromosome 16p in the Eastern Quebec population.* Am J Med Genet B Neuropsychiatr Genet, 2008. **147b**(6): p. 737-44.
102. Kerner, B., D.L. Brugman, and N.B. Freimer, *Evidence of linkage to psychosis on chromosome 5q33-34 in pedigrees ascertained for bipolar disorder.* Am J Med Genet B Neuropsychiatr Genet, 2007. **144b**(1): p. 74-8.
103. Roche, S., et al., *Candidate gene analysis of 21q22: support for S100B as a susceptibility gene for bipolar affective disorder with psychosis.* Am J Med Genet B Neuropsychiatr Genet, 2007. **144b**(8): p. 1094-6.
104. Freedman, R., et al., *Characterization of allelic variants at chromosome 15q14 in schizophrenia.* Genes Brain Behav, 2006. **5 Suppl 1**: p. 14-22.
105. Mukherjee, O., et al., *Evidence of linkage and association on 18p11.2 for psychosis.* Am J Med Genet B Neuropsychiatr Genet, 2006. **141b**(8): p. 868-73.
106. Brzustowicz, L.M., et al., *Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21-q22.* Science, 2000. **288**(5466): p. 678-82.
107. Brzustowicz, L.M., et al., *Linkage of familial schizophrenia to chromosome 13q32.* Am J Hum Genet, 1999. **65**(4): p. 1096-103.
108. Suarez, B.K., et al., *Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample.* Am J Hum Genet, 2006. **78**(2): p. 315-33.
109. McGuffin, P., et al., *Whole genome linkage scan of recurrent depressive disorder from the depression network study.* Hum Mol Genet, 2005. **14**(22): p. 3337-45.
110. Faraone, S.V., et al., *Genome scan of Han Chinese schizophrenia families from Taiwan: confirmation of linkage to 10q22.3.* Am J Psychiatry, 2006. **163**(10): p. 1760-6.
111. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nat Rev Genet, 2005. **6**(2): p. 95-108.
112. Athanasiu, L., et al., *Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort.* J Psychiatr Res, 2010. **44**(12): p. 748-53.
113. Huang, J., et al., *Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression.* Am J Psychiatry, 2010. **167**(10): p. 1254-63.
114. Moskvina, V., et al., *Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk.* Mol Psychiatry, 2009. **14**(3): p. 252-60.
115. Need, A.C., et al., *A genome-wide investigation of SNPs and CNVs in schizophrenia.* PLoS Genet, 2009. **5**(2): p. e1000373.
116. Shi, J., et al., *Common variants on chromosome 6p22.1 are associated with schizophrenia.* Nature, 2009. **460**(7256): p. 753-7.
117. Stefansson, H., et al., *Common variants conferring risk of schizophrenia.* Nature, 2009. **460**(7256): p. 744-7.
118. Sullivan, P.F., et al., *Genomewide association for schizophrenia in the CATIE study: results of stage 1.* Mol Psychiatry, 2008. **13**(6): p. 570-84.
119. O'Donovan, M.C., et al., *Identification of loci associated with schizophrenia by genome-wide association and follow-up.* Nat Genet, 2008. **40**(9): p. 1053-5.
120. Yu, H., et al., *Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in Han Chinese population.* Mol Psychiatry, 2017. **22**(7): p. 954-960.

121. Craddock, N., M.C. O'Donovan, and M.J. Owen, *Genome-wide association studies in psychiatry: lessons from early studies of non-psychiatric and psychiatric phenotypes.* Mol Psychiatry, 2008. **13**(7): p. 649-53.
122. Bateson, W. and R.C. Punnett, *Experimental studies in the physiology of heredity. Reports of the Evolution Committee.* Roy Soc, 1906. **1906**(3): p. 1-53.
123. Muller, H.J., *The Mechanism of Crossing-Over.* The American Naturalist, 1916. **50**(592): p. 193-221.
124. Morgan, T.H., *RANDOM SEGREGATION VERSUS COUPLING IN MENDELIAN INHERITANCE.* Science, 1911. **34**(873): p. 384.
125. Sturtevant, A.H., *The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association.* Journal of Experimental Zoology, 1913. **14**(1): p. 43-59.
126. Bailey-Wilson, J.E. and A.F. Wilson, *Linkage analysis in the next-generation sequencing era.* Hum Hered, 2011. **72**(4): p. 228-36.
127. Botstein, D., et al., *Construction of a genetic linkage map in man using restriction fragment length polymorphisms.* Am J Hum Genet, 1980. **32**(3): p. 314-31.
128. Grodzicker, T., et al., *Physical mapping of temperature-sensitive mutations of adenoviruses.* Cold Spring Harb Symp Quant Biol, 1975. **39 Pt 1**: p. 439-46.
129. Weber, J.L. and P.E. May, *Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction.* Am J Hum Genet, 1989. **44**(3): p. 388-96.
130. Subramanian, S., R.K. Mishra, and L. Singh, *Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions.* Genome Biol, 2003. **4**(2): p. R13.
131. Matise, T.C., et al., *A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set.* Am J Hum Genet, 2003. **73**(2): p. 271-84.
132. Elston, R.C. and J. Stewart, *A general model for the genetic analysis of pedigree data.* Hum Hered, 1971. **21**(6): p. 523-42.
133. Yuan, A. and G.E. Bonney, *Two new recursive likelihood calculation methods for genetic analysis.* Hum Hered, 2002. **54**(2): p. 82-98.
134. Cannings, C., E.A. Thompson, and M.H. Skolnick, *Probability functions on complex pedigrees.* Adv Appl Prob, 1978. **10**: p. 26-61.
135. Lange, K. and R.C. Elston, *Extensions to pedigree analysis I. Likehood calculations for simple and complex pedigrees.* Hum Hered, 1975. **25**(2): p. 95-105.
136. Ott, J., *Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies.* Am J Hum Genet, 1974. **26**(5): p. 588-97.
137. Lander, E.S. and P. Green, *Construction of multilocus genetic linkage maps in humans.* Proc Natl Acad Sci U S A, 1987. **84**(8): p. 2363-7.
138. Kruglyak, L., et al., *Parametric and nonparametric linkage analysis: a unified multipoint approach.* Am J Hum Genet, 1996. **58**(6): p. 1347-63.
139. Huang, Y., A. Thomas, and V.J. Vieland, *Employing MCMC under the PPL framework to analyze sequence data in large pedigrees.* Front Genet, 2013. **4**: p. 59.
140. Wilson, A.F., et al., *Stepwise oligogenic segregation and linkage analysis illustrated with dopamine-beta-hydroxylase activity.* Am J Med Genet, 1990. **35**(3): p. 425-32.
141. Morton, N.E., *Sequential tests for the detection of linkage.* Am J Hum Genet, 1955. **7**(3): p. 277-318.
142. Ott, J., *Analysis of Human Genetic Linkage, 3rd Edition.* Baltimore, Md: Johns Hopkins University Press, 1999.

143. Xing, C., N. Morris, and G. Xing, *Distribution of model-based multipoint heterogeneity lod scores.* Genet Epidemiol, 2010. **34**(8): p. 912-6.
144. Smith, C.A., *Testing for heterogeneity of recombination fraction values in Human Genetics.* Ann Hum Genet, 1963. **27**: p. 175-82.
145. Hodge, S.E., et al., *The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity.* Am J Hum Genet, 1983. **35**(6): p. 1139-55.
146. Ott, J., *Linkage analysis and family classification under heterogeneity.* Ann Hum Genet, 1983. **47**(Pt 4): p. 311-20.
147. Lander, E. and L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.* Nat Genet, 1995. **11**(3): p. 241-7.
148. Hodge, S.E., V.J. Vieland, and D.A. Greenberg, *HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity.* Am J Hum Genet, 2002. **70**(2): p. 556-9.
149. Clerget-Darpoux, F., *Extension of the lod score: the mod score.* Adv Genet, 2001. **42**: p. 115-24.
150. MacLean, C.J., et al., *Distribution of lod scores under uncertain mode of inheritance.* Am J Hum Genet, 1993. **52**(2): p. 354-61.
151. Clerget-Darpoux, F., et al., *A new method to test genetic models in HLA associated diseases: the MASC method.* Ann Hum Genet, 1988. **52**(Pt 3): p. 247-58.
152. Clerget-Darpoux, F., C. Bonaiti-Pellie, and J. Hochez, *Effects of misspecifying genetic parameters in lod score analysis.* Biometrics, 1986. **42**(2): p. 393-9.
153. Penrose, L.S., *The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage.* Ann Eugen, 1935. **6**: p. 133-138.
154. Weeks, D.E. and K. Lange, *The affected-pedigree-member method of linkage analysis.* Am J Hum Genet, 1988. **42**(2): p. 315-26.
155. Weeks, D.E. and K. Lange, *A multilocus extension of the affected-pedigree-member method of linkage analysis.* Am J Hum Genet, 1992. **50**(4): p. 859-68.
156. Powell, J.E., P.M. Visscher, and M.E. Goddard, *Reconciling the analysis of IBD and IBS in complex trait studies.* Nat Rev Genet, 2010. **11**(11): p. 800-5.
157. Skrivanek, Z., S. Lin, and M. Irwin, *Linkage analysis with sequential imputation.* Genet Epidemiol, 2003. **25**(1): p. 25-35.
158. Whittemore, A.S. and J. Halpern, *A class of tests for linkage using affected pedigree members.* Biometrics, 1994. **50**(1): p. 118-27.
159. Sobel, E. and K. Lange, *Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics.* Am J Hum Genet, 1996. **58**(6): p. 1323-37.
160. McPeek, M.S., *Optimal allele-sharing statistics for genetic mapping using affected relatives.* Genet Epidemiol, 1999. **16**(3): p. 225-49.
161. Feingold, E., K.K. Song, and D.E. Weeks, *Comparison of allele-sharing statistics for general pedigrees.* Genet Epidemiol, 2000. **19 Suppl 1**: p. S92-8.
162. Davis, S. and D.E. Weeks, *Comparison of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluation.* Am J Hum Genet, 1997. **61**(6): p. 1431-44.
163. Lio, P. and N.E. Morton, *Comparison of parametric and nonparametric methods to map oligogenes by linkage.* Proc Natl Acad Sci U S A, 1997. **94**(10): p. 5344-8.
164. Vieland, V.J., *Bayesian linkage analysis, or: how I learned to stop worrying and love the posterior probability of linkage.* Am J Hum Genet, 1998. **63**(4): p. 947-54.
165. Smith, C.A.B., *Some Comments on the Statistical Methods used in Linkage Investigations.* Am J Hum Genet, 1959. **11**(4): p. 289-304.

166. Vieland, V.J., *Thermometers: something for statistical geneticists to think about.* Hum Hered, 2006. **61**(3): p. 144-56.
167. Vieland, V.J., et al., *KELVIN: a software package for rigorous measurement of statistical evidence in human genetics.* Hum Hered, 2011. **72**(4): p. 276-88.
168. Elston, R.C. and K. Lange, *The prior probability of autosomal linkage.* Ann Hum Genet, 1975. **38**(3): p. 341-50.
169. Logue, M.W., et al., *A posterior probability of linkage-based re-analysis of schizophrenia data yields evidence of linkage to chromosomes 1 and 17.* Hum Hered, 2006. **62**(1): p. 47-54.
170. Vieland, V.J., et al., *The value of regenotyping older linkage data sets with denser marker panels.* Hum Hered, 2014. **78**(1): p. 9-16.
171. Thomas, A., et al., *Multilocus linkage analysis by blocked Gibbs sampling.* Statistics and Computing, 2000. **10**(3): p. 259-269.
172. Seok, S.C., M. Evans, and V.J. Vieland, *Fast and accurate calculation of a computationally intensive statistic for mapping disease genes.* J Comput Biol, 2009. **16**(5): p. 659-76.
173. Haines, J.L., et al., *Complement factor H variant increases the risk of age-related macular degeneration.* Science, 2005. **308**(5720): p. 419-21.
174. Edwards, A.O., et al., *Complement factor H polymorphism and age-related macular degeneration.* Science, 2005. **308**(5720): p. 421-4.
175. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.
176. Cooper, G.M., et al., *A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose.* Blood, 2008. **112**(4): p. 1022-7.
177. Kerem, B., et al., *Identification of the cystic fibrosis gene: genetic analysis.* Science, 1989. **245**(4922): p. 1073-80.
178. MacDonald, M.E., et al., *The Huntington's disease candidate region exhibits many different haplotypes.* Nat Genet, 1992. **1**(2): p. 99-103.
179. Corder, E.H., et al., *Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families.* Science, 1993. **261**(5123): p. 921-3.
180. Altshuler, D., et al., *The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes.* Nat Genet, 2000. **26**(1): p. 76-80.
181. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. **17**(9): p. 502-10.
182. The International HapMap Consortium, *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
183. Devlin, B. and N. Risch, *A comparison of linkage disequilibrium measures for fine-scale mapping.* Genomics, 1995. **29**(2): p. 311-22.
184. Guo, S.W., *Linkage disequilibrium measures for fine-scale mapping: a comparison.* Hum Hered, 1997. **47**(6): p. 301-14.
185. Boyles, A.L., et al., *Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing.* Hum Hered, 2005. **59**(4): p. 220-7.
186. Wratten, N.S., et al., *Identification of a schizophrenia-associated functional noncoding variant in NOS1AP.* Am J Psychiatry, 2009. **166**(4): p. 434-41.
187. Brzustowicz, L.M., et al., *Fine mapping of the schizophrenia susceptibility locus on chromosome 1q22.* Hum Hered, 2002. **54**(4): p. 199-209.
188. Lewis, C.M., *Genetic association studies: design, analysis and interpretation.* Brief Bioinform, 2002. **3**(2): p. 146-53.

189. Lettre, G., C. Lange, and J.N. Hirschhorn, *Genetic model testing and statistical power in population-based association studies of quantitative traits.* Genet Epidemiol, 2007. **31**(4): p. 358-62.
190. Hochberg, Y. and Y. Benjamini, *More powerful procedures for multiple significance testing.* Stat Med, 1990. **9**(7): p. 811-8.
191. Falush, D., M. Stephens, and J.K. Pritchard, *Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies.* Genetics, 2003. **164**(4): p. 1567-87.
192. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
193. Weeks D.E., Ott J., and L. G.M., *SLINK: A general simulation program for linkage analysis.* Am J Hum Genet, 1990. **47**: p. A204.
194. Bassett, A.S., et al., *Positive and negative symptoms in families with schizophrenia.* Schizophr Res, 1993. **11**(1): p. 9-19.
195. Bassett, A.S. and W.G. Honer, *Evidence for anticipation in schizophrenia.* Am J Hum Genet, 1994. **54**(5): p. 864-70.
196. Brzustowicz, L.M., et al., *Use of a quantitative trait to map a locus associated with severity of positive symptoms in familial schizophrenia to chromosome 6p.* Am J Hum Genet, 1997. **61**(6): p. 1388-96.
197. Saviouk, V., et al., *Tumor necrosis factor promoter haplotype associated with schizophrenia reveals a linked locus on 1q44.* Mol Psychiatry, 2005. **10**(4): p. 375-83.
198. Göring, H.H.H. and J.D. Terwilliger, *Linkage Analysis in the Presence of Errors IV: Joint Pseudomarker Analysis of Linkage and/or Linkage Disequilibrium on a Mixture of Pedigrees and Singletons When the Mode of Inheritance Cannot Be Accurately Specified.* Am J Hum Genet, 2000. **66**(4): p. 1310-27.
199. Hiekkalinna, T., et al., *PSEUDOMARKER: A Powerful Program for Joint Linkage and/or Linkage Disequilibrium Analysis on Mixtures of Singletons and Related Individuals.* Human Heredity, 2011. **71**(4): p. 256-266.
200. Brzustowicz, L.M., et al., *Linkage disequilibrium mapping of schizophrenia susceptibility to the CAPON region of chromosome 1q22.* Am J Hum Genet, 2004. **74**(5): p. 1057-63.
201. Purcell, S., *PLINK*.
202. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
203. Moreau, M., *Altered microRNA Regulatory Networks in Individuals with Schizophrenia*, in *Graduate School-New Brunswick and The Graduate School of Biomedical Sciences*. 2009, Rutgers University and University of Medicine and Dentistry of New Jersey. p. 223.
204. Bruse, S., et al., *Improvements to bead-based oligonucleotide ligation SNP genotyping assays.* Biotechniques, 2008. **45**(5): p. 559-71.
205. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
206. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
207. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
208. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
209. The Genomes Project, C., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

210. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
211. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
212. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11.10.1-33.
213. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
214. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
215. Cingolani, P., et al., *Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift.* Front Genet, 2012. **3**: p. 35.
216. Ng, M.Y., et al., *Meta-analysis of 32 genome-wide linkage studies of schizophrenia.* Mol Psychiatry, 2009. **14**(8): p. 774-85.
217. Bustelo, X.R., *Regulatory and signaling properties of the Vav family.* Mol Cell Biol, 2000. **20**(5): p. 1461-77.
218. Yaron, A. and B. Zheng, *Navigating their way to the clinic: emerging roles for axon guidance molecules in neurological disorders and injury.* Dev Neurobiol, 2007. **67**(9): p. 1216-31.
219. Cowan, C.W., et al., *Vav family GEFs link activated Ephs to endocytosis and axon guidance.* Neuron, 2005. **46**(2): p. 205-17.
220. Sauzeau, V., et al., *Vav3 is involved in GABAergic axon guidance events important for the proper function of brainstem neurons controlling cardiovascular, respiratory, and renal parameters.* Mol Biol Cell, 2010. **21**(23): p. 4251-63.
221. Quevedo, C., et al., *Vav3-deficient mice exhibit a transient delay in cerebellar development.* Mol Biol Cell, 2010. **21**(6): p. 1125-39.
222. Arinami, T., et al., *Genomewide high-density SNP linkage analysis of 236 Japanese families supports the existence of schizophrenia susceptibility loci on chromosomes 1p, 14q, and 20p.* Am J Hum Genet, 2005. **77**(6): p. 937-44.
223. Ikeda, M., et al., *Genome-wide association study of schizophrenia in a Japanese population.* Biol Psychiatry, 2011. **69**(5): p. 472-8.
224. Aleksic, B., et al., *Analysis of the VAV3 as candidate gene for schizophrenia: evidences from voxel-based morphometry and mutation screening.* Schizophr Bull, 2013. **39**(3): p. 720-8.
225. Jaffrey, S.R., et al., *CAPON: a protein associated with neuronal nitric oxide synthase that regulates its interactions with PSD95.* Neuron, 1998. **20**(1): p. 115-24.
226. Brenman, J.E., et al., *Interaction of nitric oxide synthase with the postsynaptic density protein PSD-95 and alpha1-syntrophin mediated by PDZ domains.* Cell, 1996. **84**(5): p. 757-67.
227. Brenman, J.E., et al., *Cloning and characterization of postsynaptic density 93, a nitric oxide synthase interacting protein.* J Neurosci, 1996. **16**(23): p. 7407-15.
228. Coyle, J.T., *NMDA Receptor and Schizophrenia: A Brief History.* Schizophr Bull, 2012. **38**(5): p. 920-6.
229. Seki, N., et al., *Characterization of cDNA clones in size-fractionated cDNA libraries from human brain.* DNA Res, 1997. **4**(5): p. 345-9.

230. Hadzimichalis, N.M., et al., *NOS1AP protein levels are altered in BA46 and cerebellum of patients with schizophrenia.* Schizophr Res, 2010. **124**(1-3): p. 248-50.
231. Fang, M., et al., *Dexras1: a G protein specifically coupled to neuronal nitric oxide synthase via CAPON.* Neuron, 2000. **28**(1): p. 183-93.
232. Jaffrey, S.R., et al., *Neuronal nitric-oxide synthase localization mediated by a ternary complex with synapsin and CAPON.* Proc Natl Acad Sci U S A, 2002. **99**(5): p. 3199-204.
233. Carrel, D., et al., *NOS1AP Regulates Dendrite Patterning of Hippocampal Neurons through a Carboxypeptidase E-Mediated Pathway.* J Neurosci, 2009. **29**(25): p. 8248.
234. Xu, B., et al., *Increased expression in dorsolateral prefrontal cortex of CAPON in schizophrenia and bipolar disorder.* PLoS Med, 2005. **2**(10): p. e263.
235. Carrel, D., et al., *Nitric oxide synthase 1 adaptor protein, a protein implicated in schizophrenia, controls radial migration of cortical neurons.* Biol Psychiatry, 2015. **77**(11): p. 969-78.
236. Gurling, H.M., et al., *Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23.* Am J Hum Genet, 2001. **68**(3): p. 661-73.
237. Hwu, H.G., et al., *Linkage of schizophrenia with chromosome 1q loci in Taiwanese families.* Mol Psychiatry, 2003. **8**(4): p. 445-52.
238. Shaw, S.H., et al., *A genome-wide search for schizophrenia susceptibility genes.* Am J Med Genet, 1998. **81**(5): p. 364-76.
239. Zheng, Y., et al., *Association of the carboxyl-terminal PDZ ligand of neuronal nitric oxide synthase gene with schizophrenia in the Chinese Han population.* Biochem Biophys Res Commun, 2005. **328**(4): p. 809-15.
240. Rosa, A., et al., *1q21-q22 locus is associated with susceptibility to the reality-distortion syndrome of schizophrenia spectrum disorders.* Am J Med Genet, 2002. **114**(5): p. 516-8.
241. Miranda, A., et al., *Putative association of the carboxy-terminal PDZ ligand of neuronal nitric oxide synthase gene (CAPON) with schizophrenia in a Colombian population.* Schizophr Res, 2006. **82**(2-3): p. 283-5.
242. Kremeyer, B., et al., *Evidence for a role of the NOS1AP (CAPON) gene in schizophrenia and its clinical dimensions: an association study in a South American population isolate.* Hum Hered, 2009. **67**(3): p. 163-73.
243. Zhou, K., et al., *NMDA receptor hypofunction induces dysfunctions of energy metabolism and semaphorin signaling in rats: a synaptic proteome study.* Schizophr Bull, 2012. **38**(3): p. 579-91.
244. Brot, S., et al., *Collapsin response mediator protein 5 (CRMP5) induces mitophagy, thereby regulating mitochondrion numbers in dendrites.* J Biol Chem, 2014. **289**(4): p. 2261-76.
245. Brot, S., et al., *Collapsin response-mediator protein 5 (CRMP5) phosphorylation at threonine 516 regulates neurite outgrowth inhibition.* Eur J Neurosci, 2014. **40**(7): p. 3010-20.
246. Okamoto, N., et al., *Molecular characterization of a new metabotropic glutamate receptor mGluR7 coupled to inhibitory cyclic AMP signal transduction.* J Biol Chem, 1994. **269**(2): p. 1231-6.
247. Gee, C.E., et al., *Blocking metabotropic glutamate receptor subtype 7 (mGlu7) via the Venus flytrap domain (VFTD) inhibits amygdala plasticity, stress, and anxiety-related behavior.* J Biol Chem, 2014. **289**(16): p. 10975-87.

248. Ohtsuki, T., et al., *A polymorphism of the metabotropic glutamate receptor mGluR7 (GRM7) gene is associated with schizophrenia.* Schizophr Res, 2008. **101**(1-3): p. 9-16.
249. Shibata, H., et al., *Association study of polymorphisms in the group III metabotropic glutamate receptor genes, GRM4 and GRM7, with schizophrenia.* Psychiatry Res, 2009. **167**(1-2): p. 88-96.
250. Ganda, C., et al., *A family-based association study of DNA sequence variants in GRM7 with schizophrenia in an Indonesian population.* Int J Neuropsychopharmacol, 2009. **12**(9): p. 1283-9.
251. Jajodia, A., et al., *Evidence for schizophrenia susceptibility alleles in the Indian population: An association of neurodevelopmental genes in case-control and familial samples.* Schizophr Res, 2015. **162**(1-3): p. 112-7.
252. Li, W., et al., *Significant association of GRM7 and GRM8 genes with schizophrenia and major depressive disorder in the Han Chinese population.* Eur Neuropsychopharmacol, 2016. **26**(1): p. 136-46.
253. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
254. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.
255. Moreau, M.P., et al., *Altered microRNA expression profiles in postmortem brain samples from individuals with schizophrenia and bipolar disorder.* Biol Psychiatry, 2011. **69**(2): p. 188-93.
256. Guidotti, A., et al., *Decrease in reelin and glutamic acid decarboxylase67 (GAD67) expression in schizophrenia and bipolar disorder: a postmortem brain study.* Arch Gen Psychiatry, 2000. **57**(11): p. 1061-9.
257. Huang, H.S., et al., *Prefrontal dysfunction in schizophrenia involves mixed-lineage leukemia 1-regulated histone methylation at GABAergic gene promoters.* J Neurosci, 2007. **27**(42): p. 11254-62.
258. Morishita, H., et al., *Interneuron epigenomes during the critical period of cortical plasticity: Implications for schizophrenia.* Neurobiol Learn Mem, 2015. **124**: p. 104-10.
259. Nato, A.Q., S. Buyske, and T.C. Matise, *The Rutgers map: A third-generation combined linkage-physical map of the human genome*, in *Human Genetics Institute of New Jersey, Second Research Day*. 2012: Life Sciences Building, Rutgers University, Piscataway, NJ, USA
260. Matise, T.C., et al., *A second-generation combined linkage physical map of the human genome.* Genome Res, 2007. **17**(12): p. 1783-6.
261. Hickey, G., et al., *HAL: a hierarchical format for storing and analyzing multiple genome alignments.* Bioinformatics, 2013. **29**(10): p. 1341-2.
262. Lange, K., et al., *Mendel: the Swiss army knife of genetic analysis programs.* Bioinformatics, 2013. **29**(12): p. 1568-70.
263. Frazer, K.A., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
264. Price, A.L., et al., *Long-range LD can confound genome scans in admixed populations.* Am J Hum Genet, 2008. **83**(1): p. 132-5; author reply 135-9.
265. Edlund, C.K., D.V. Conti, and D.J. Van Den Berg. *rAggr*. 2017; rAggr is a web-based software program for finding markers (SNPs and indels) that are in linkage-disequilibrium (LD) with a set of queried markers, using the 1000 Genomes Project and HapMap genotype databases. rAggr uses an expectation–maximization algorithm

adapted from the Haploview software (Barrett et al, Bioinformatics. 2005 Jan 15;21(2):263-5) to calculate pairwise r2 and D'. All calculations are done "on the fly" by the web server.]. Available from: http://raggr.usc.edu.

266. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.
267. Azaro, M., *Genotyper*. 2016.
268. Jentsch, T.J., *Neuronal KCNQ potassium channels: physiology and role in disease.* Nat Rev Neurosci, 2000. **1**(1): p. 21-30.
269. Delmas, P. and D.A. Brown, *Pathways modulating neural KCNQ/M (Kv7) potassium channels.* Nat Rev Neurosci, 2005. **6**(11): p. 850-62.
270. Borsotto, M., et al., *PP2A-Bgamma subunit and KCNQ2 K+ channels in bipolar disorder.* Pharmacogenomics J, 2007. **7**(2): p. 123-32.
271. Blom, S.M., et al., *From pan-reactive KV7 channel opener to subtype selective opener/inhibitor by addition of a methyl group.* PLoS One, 2014. **9**(6): p. e100209.
272. Sotty, F., et al., *Antipsychotic-like effect of retigabine [N-(2-Amino-4-(fluorobenzylamino)-phenyl)carbamic acid ester], a KCNQ potassium channel opener, via modulation of mesolimbic dopaminergic neurotransmission.* J Pharmacol Exp Ther, 2009. **328**(3): p. 951-62.
273. Guiraldelli, M.F., et al., *Mouse HFM1/Mer3 is required for crossover formation and complete synapsis of homologous chromosomes during meiosis.* PLoS Genet, 2013. **9**(3): p. e1003383.
274. Wang, Q., et al., *Increased co-expression of genes harboring the damaging de novo mutations in Chinese schizophrenic patients during prenatal development.* 2015. **5**: p. 18209.
275. Benson, M.A., et al., *Ryanodine receptors are part of the myospryn complex in cardiac muscle.* Sci Rep, 2017. **7**.
276. Wang, Q., et al., *The CMYA5 gene confers risk for both schizophrenia and major depressive disorder in the Han Chinese population.* World J Biol Psychiatry, 2014. **15**(7): p. 553-60.
277. Han, S., et al., *Association between CMYA5 gene polymorphisms and risk of schizophrenia in Uygur population and a meta-analysis.* Early Interv Psychiatry, 2015.
278. Benson, M.A., C.L. Tinsley, and D.J. Blake, *Myospryn is a novel binding partner for dysbindin in muscle.* J Biol Chem, 2004. **279**(11): p. 10450-8.
279. Prats, C., et al., *Evidence of an epistatic effect between Dysbindin-1 and Neuritin-1 genes on the risk for schizophrenia spectrum disorders.* Eur Psychiatry, 2017. **40**: p. 60-64.
280. Knoch, J., et al., *Rare hereditary diseases with defects in DNA-repair.* Eur J Dermatol, 2012. **22**(4): p. 443-55.