A GENOME-WIDE ASSOCIATION STUDY OF BIPOLAR DISORDER USING DIMENSIONAL SYMPTOM CLASSIFICATIONS

by

Samuel M. Friedman

Submitted to the Department of Genetics

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Arts

in Genetics

at

Rutgers The State University of New Jersey

April 2021

Written Under the Direction of

Linda Brzustowicz, M.D.

Distinguished Professor

Department of Genetics

Rutgers, The State University of New Jersey

ABSTRACT

Modern psychiatry struggles to reconcile the diagnostic heterogeneity that pervades psychopathologies, including Bipolar Disorder (BD). BD presents on the affective spectrum, oscillating between manic highs and depressive lows. Individuals with varying severity of mania are differentiated by different subtypes of BD (type 1 and 2). However, there is still a substantial amount of diagnostic heterogeneity in any one sub-diagnosis that previous studies have not sufficiently accounted for. For example, sub-diagnoses of BD fail to differentiate the psychotic features that often present alongside affective symptoms. A more nuanced understanding of the distinct biological underpinnings of BD could transform the health outcomes for those battling BD by way of objective susceptibility measures, more concrete diagnoses, and improved treatment planning. Genome-wide association studies (GWAS) are a powerful method for localizing single nucleotide polymorphisms (SNPs) that may be associated with phenotypes of interest. Family-based transmission disequilibrium tests (TDT) are a unique approach to GWAS because they are robust to false-positive findings due to the effects of population stratification. Using rich phenotypic data from previously gathered family studies alongside recently generated SNP data, we set out to investigate associations among individuals with type 1 (severe) BD (BD1), comparing subjects who comorbidly exhibit psychotic features (BD1-P) to subjects whose presentations of BD lack any psychotic features (BD1-NP). We have found a cluster of SNPs on 6q25.3 with p-values of approximately 2.0×10^{-6} among the BD1-NP cohort that are suggestive of findings. Lack of any significant findings at the same locus among BD1-P individuals, and an intermediate level of signal in a mixed group strengthen our findings and demonstrate the striking heterogeneity that pervades the current classification system. Future work will involve aggregating case-control samples and conducting a joint case-control/family association study to further assess the significance of our current findings. Additionally, we hope to continue to distinguish nuanced associations between the human genome and symptoms of psychopathology.

TABLE OF CONTENTS

INTRODUCTION

Challenges Posed by Psychiatric Disorders

Prevalence of Psychiatric Disorders

Psychiatric disorders are a major burden on public health and the global economy[1]. A decade ago, Merikangas et al. found that nearly 1 in 5 U.S. adults will experience living with a mental illness, and that figure still holds today[2,3]. Since then, certain subgroups within the population have been found to have an even higher prevalence of psychiatric disorders[3]. For example, Bruffaerts et al. [2019] has shown that there is a 31.4% likelihood for first year college students to experience mental disorders that can last 12-months, based on a study that tracked nearly 14,000 first-year college students[4]. These striking figures point to an overwhelming need for stronger diagnostic criteria, better preventative measures, and more effective treatments for the whole gamut of existing neuropsychiatric disorders. Several promising novel avenues of research for quantifying variation in the brain and genome have emerged in recent decades; however, few of these studies have actually translated into meaningful health outcome improvements[5]. Due to the extremely complicated architecture and physiology of the human mind, modern psychiatry has not been able to follow the increased pace of advancements that occur within many other medical fields[6].

Clinical Barriers in Psychiatry

Psychiatric help, often in the form of pharmacological intervention, can greatly improve the quality of life for a large proportion of patients struggling with their mental health[7]. However, psychiatrists face a unique set of challenges that impact their ability to care for their patients. For all clinicians, symptomatology is necessary in order to create a framework of terminology through which a patient can be diagnosed and subsequently treated. However, many psychiatric disorders contain overlapping symptoms under the current classification system. This means that psychiatric diagnoses are prone to diagnostic heterogeneity, which creates difficulties for clinicians in pinpointing any given patient's exact illness. For example, two patients with the same diagnosis will often exhibit *different* clinical features while two patients with different

diagnoses may exhibit the *same* clinical features. This not only makes it a challenge to keep diagnoses objective, but it also impacts the efficacy of existing therapeutic treatments[7]. Furthermore, the entire process of moving from diagnosis to selecting a treatment to maximizing the efficacy of a treatment contains much subjectivity and uncertainty in psychiatry[7]. To a large extent, this is because of a lack of understanding about most of the biological underpinnings of psychopathologies. Unlike many other areas of medicine—where a diagnosis implies a specific, identifiable etiology that underlies the pathophysiology—psychiatric nosology focuses primarily on symptoms[7].

Genetic Barriers to Psychiatric Research

The clinical barriers in psychiatric genetics can be diminished with a more complete understanding of the neurophysiology, which—at its core—arises from sets of genes that inform neural development. In other fields of modern medicine with comparatively less complex organ systems, the link between genetics and disease is often very clear. In fact, the advancements made in other fields have been paving the way for a future of “precision medicine”, where care is tailored to a person’s unique genetic makeup. Research in the genetics underlying psychiatric disorders is held back in comparison to other fields due to the scarcity of identifiable biomarkers that have been discovered to date[8,9].

Difficulty identifying biomarkers can be attributed to the notable influence of environmental factors on the mental state[10]. A major model of etiology and risk for psychiatric disorders is founded on the idea that a highly polygenic interplay of multiple genes, on many different chromosomes, all contribute to a person’s overall susceptibility[11]. Hence, there is no one gene or small group of genes that directly causes psychiatric disorders. Thus, most of the relevant molecular pathways for psychiatric disorders are unknown to this day. Furthermore, whether one actually manifests a psychiatric disorder is often dependent on or aggravated by environmental factors such as trauma or substance abuse[12,13]. Thus, since it is possible for individuals to have a sufficient genetic susceptibility for a particular disorder but not manifest it,

psychiatric disorders are fraught with low penetrance (the proportion of individuals who carry genetic risk for a disease and actually express the disease)[14]. The simultaneous interplay of these nuances increases the difficulty of obtaining meaningful findings when conducting genetic analyses on psychiatric disorders.

Locating Genes Associated with Disease

Identifying Sources of Genetic Variation in Humans

Although the human genome contains billions of base pairs, there is shockingly little variation in gene sequence from person to person. In fact, all humans are over 99.8% identical by genetic makeup[15]. However, the uniqueness of each individual's traits with but a 0.2% margin of variation from person to person can be explained by the fact that there are over three billion base pairs in the human genome. These variations come in different forms; one of the most commonly investigated is the single nucleotide polymorphism (SNP).

SNPs are nucleotide positions in the genome at which different members of a population express different alleles; that is, they are sites containing polymorphic alleles[15]. Millions of SNPs exist in the genetic architecture of humans because evolution is driven by a stochastic process, where, occasionally, a nucleotide spontaneously mutates[16]. Pathogenic variants tend to decrease the fitness of affected individuals and thus appear at too low a frequency relative to the population as a whole (less than 1%) to count as polymorphic alleles. In contrast, SNP variants tend to not have perceivable effects on the probands (the individual who serves as the starting point for a genetic study) in which they first appear. Although pathogenic variants are what is truly of interest to investigators, the identification of SNPs that are associated with a disease trait is a much more practical and tangible endeavor.

Assuming an individual has progeny, disease generating alleles may be transmitted to some of the progeny descended from him/her. Furthermore, not only are these changes transmitted, but, due to linkage disequilibrium, adjacent SNPs will tend to segregate along with

the nearby mutation. Thus, the association of specific SNPs with a disease of interest can greatly narrow down the localization of the disease-causing mutation(s).

Although the locations of several million SNPs are known, localizing the genes that contribute to the development of diseases in humans is a sizable, yet critical, task. An increased understanding of a psychopathology's susceptibility variants has the potential to enable clinicians to objectively reach diagnoses and accordingly utilize targeted therapeutics in the treatment process. Furthermore, such findings would allow for better prophylactic treatments and preventative measures to be taken to reduce exposure to potential environmental triggers in individuals who carry a genetic risk burden.

Linkage Studies: Locating Linkage Peaks Associated with Pathologies

Over the past few decades, researchers have conducted several analyses which have identified several loci that confer risk for schizophrenia and affective disorders[17-22]. Linkage studies depend on large blocks of DNA that recombine and are inherited together during meiosis. Linkage peaks identify large regions on any particular chromosome of potential relationship with a phenotype of interest. Association studies, to be discussed in the next section, are aptly suited to narrow in on the numerous specific susceptibility-bearing loci of complex disorders[23]. Additionally, association studies can compare cases and controls from individuals who appear unrelated because SNPs unify genetic ancestry among individuals who are likely not even aware of their distant relationships. However, as whole-genome sequencing is becoming more accessible and is trending towards no longer being prohibitively expensive, linkage studies will again be used widely to identify disease-causing genes[24].

GWAS: Locating SNPs Associated with Pathologies

A genome-wide association study (GWAS or GWA study) is a method used by researchers to pinpoint the specific locations of disease-causing variants[25,26]. Researchers conducting a GWA study can leverage the facts that (1) SNPs occur throughout the genome and (2) are often found adjacent to or within genes that are causative for the manifestation of

diseases in order to locate disease-causing genes[27]. Specifically, in a GWA study, investigators will genotype affected and unaffected (control) samples at hundreds of thousands to several million SNPs and will then statistically determine which SNPs tend to be inherited more frequently by symptomatic individuals than by asymptomatic controls[28]. Through the identification of particular variants that appear overrepresented among an affected population in comparison to unaffected controls, investigators are able to vastly narrow down the probable loci of causative genes through GWA studies[23].

Case-Control Association Studies

When conducting an association case-control studies, it is common to combine datasets of individuals from highly diverse populations. However, the heterogeneity present in the frequencies of alleles between individuals of different ethnic origins should never be disregarded[29]. Certain factors have caused variant frequencies to differ between populations over time. A primary factor includes differences in the *de novo* mutations that have occurred among common ancestors long ago. Other factors include population drift, or a bottleneck or founder effect—all circumstances that alter the gene pool of a population—that cause the frequency of alleles to differ in different populations[30].

If the genetic profile of cases is not matched carefully to controls, the population stratification within the sample space would result in several false-positive associations[23,31,32]. Population stratification is the phenomenon whereby differences in population's underlying allele requires the stratification of multi-ethnic genetic data sets in order to minimize false positive findings. For example, lactose intolerance might be wrongly associated with a disease simply because some populations actively produce lactase enzyme (and thus have varying allele frequencies at the loci that encode lactase), whereas others do not. Thus, in order to not disregard the heterogeneity between the datasets, statistical techniques such as principal components analysis (PCA) can be employed to ensure that cases are matched well with their controls[17,33].

Family-Based Association Studies

Alternatively, family-based studies for association are entirely robust to within-family population stratification[32,34]. Family-based studies work with "trios", two parents and their offspring, as opposed to matching a case with an unrelated control[35]. The two alleles at each genetic locus in the kin is compared to the four alleles of parental DNA (two from each parent); at each locus, the two parental alleles that were not transmitted to the child serve as control alleles. In family-based association studies, the allele frequencies of populations of case and control alleles are perfectly matched because the ethnic background of both parents most closely matches that of their child. However, family trios are more difficult to recruit than unrelated individuals. Additionally, part of the genetic information among trios is used as control information, when it could otherwise serve as valuable case information in a case-control study. Once phenotypes are assigned to individuals, a transmission disequilibrium test (TDT) can be implemented to calculate p-values of association for every SNP within family-based datasets. The TDT is an application of MacNemar's test that compares the proportions of two dichotomous variables (in this case, transmitted and untransmitted alleles). The TDT statistic is calculated as:

$$\frac{(a-b)^2}{(a+b)}$$

where *a* is equal to the number of transmitted alleles and *b* is equal to the number of untransmitted alleles. Under the null hypothesis, the TDT follows a chi-squared distribution with one degree of freedom.

Practical Considerations for GWA Studies

Since GWAS is accomplished by aggregating and comparing information about hundreds of thousands to millions of SNPs in the genome, a multiple-testing correction to the p-value will need to be addressed[36]. Additionally, the ethical implications that arise in the form of healthcare disparities down the pipeline from the initial GWASs will also be addressed[37].

One of the primary difficulties in conducting GWA studies is achieving sufficient statistical power. A large pool of samples is required in order to minimize the effects of false positives and obtain significant results[38,39]. The difficulty exists primarily in amassing and analyzing both phenotypic and genotypic data of a large sample size of study participants; it is incredibly labor intensive, time consuming, and expensive. However, although they may be difficult to orchestrate, GWA studies are particularly important for the investigation of disorders and are a worthwhile economic investment for the future of healthcare.

GWA Studies for Psychiatric Genetics

The introduction of GWAS as a method for investigators to better understand polygenic traits invigorated the scientific community of the late-1990s and 2000s[40]. Over the past thirteen years, more than 1,200 SNPs were found—with genome-wide significance—to be associated with psychopathologies[41]. In more recent years, progress in psychiatric genetics has slowed[42]. In order to keep the momentum of this line of research steady, more nuances—that is, associations between specific symptoms and SNPs—should be discerned, until whole genome sequencing of several study participants is no longer be prohibitively expensive and new tools will yield even finer genetic resolution[24].

Current Efforts in Psychiatric Genetics

Polygenic Risk Scoring

Building off of GWA studies, a relatively recent step forward in genetics has been the application of current knowledge of human susceptibility variants, with varying levels of effect sizes, toward establishing polygenic risk scores (PRS). PRS is a metric that builds off of GWAS data and adds the weighted effects of all of the identified genetic variants to produce a score that is predictive of latent risk for many phenotypes of interest[8,43]. Unlike many of the well-characterized maladies that have genetic influences, psychiatric disorders are known to be highly polygenic, with hundreds or thousands of variant genes that cumulatively contribute to a

disordered phenotype[44]. Thus, PRS is a metric that points toward striking progress in the characterization of the genetic risk liability for individuals with psychiatric disorders[45].

Strengthening Diagnostic Criteria

The emerging realization is that a more dimensional approach to the characterization of phenotypes would profoundly improve research efforts in psychiatric genetics[6,46]. Not only would it improve our understanding of the biology that underlies such complex phenotypes, but by incorporating the biology as a dimension of a larger system, it would also help with the translation of such findings into better health outcomes for the individuals that this research has always sought to help.

NIMH's Research Domain Criteria (RDoC) is one such framework that approaches psychiatric symptomatology in a new light. RDoC breaks the status quo of considering only traditional, subjective symptoms; instead, it focuses on a wider range of data from the whole gamut of academic research in neuroscience, from genetic markers to neurocircuits to self-reports from affected patients[47]. This framework elegantly integrates several levels of analysis into a dimensional model for the modern reclassification of psychiatric disorders. By doing so, rather than solely considering the subjective and non-absolute diagnoses that clinicians garner from their patients, RDoC is superior to traditional methods because it accounts for the wildly varying individual experiences in the realm of psychopathology[47]. Using a dimensional framework for the classification of psychiatric disorders may enable researchers to identify more nuances in the science that drives the manifestation of psychiatric disorders. Significant results from geneticists that are able to discover the genetic component of psychopathologies would vastly improve the utility of a framework such as RDoC in clinical practice.

Data Sharing and Meta-Analyses

Achieving finer detail of the genetics at the root of psychopathologies requires a great deal of collaboration. In order to amass substantial amounts of phenotypic and genotypic data, it is imperative that the psychiatric genetics community work together; for this reason, data-

sharing is becoming standard practice in the field[17]. Given the challenges involved in gathering clinical information and genetic data, the sharing of datasets affords researchers cost-effective ways to conduct their analyses. It also allows for meta-analyses to be applied to a larger dataset than any one study might be able to carry out. Examples of this include the Psychiatric Genomics Consortium (PGC), which combined data from several independent studies to perform highly significant meta-GWA studies[48]. As a direct result of this, in 2018, 44 new risk variants were determined to be associated with major depression[49]. Another group looked at PGC data in the schizophrenia collection and found specific genes (and even point some theories towards potential pathways) that have a strong association with susceptibility for the manifestation of schizophrenia[50]. Over 100 new variants were identified in another meta-analysis on nearly 500,000 individuals in data collections for neuroticism studies, highlighting the strength of data sharing in this field[51]. Finally, PGC found 30 new loci implicated in connection to bipolar disorder in 2019[52].

NIMH Repository for Psychiatric Genetics

In order to confront the logistical obstacles related to the aggregation, curation, and harmonization of shared data, the National Institute of Mental Health (NIMH) has dedicated funding and resources toward the active maintenance of a repository of more than two hundred thousand patient and control samples[53]. The NIMH Repository and Genomics Resource (NRGR) is a centralized location for both the storage and distribution of biosamples and data curation of corresponding phenotypic data related to psychiatric studies. The NRGR's wide array of biosamples includes DNA, RNA, immortalized cell lines, and induced pluripotent stem cells (iPSCs), from subjects who collectively exhibit a wide range of psychiatric disorders, as well as unaffected controls[53]. The NRGR actively makes high-quality clinical phenotypic data as well as biological samples of patients and controls available to investigators globally to enable high-powered research in psychiatric genetics. Datasets such as those that comprise the PGC, and a plethora of others that were constructed for the studies that were conducted over the past

30 years can be requested and studied by qualified investigators. At this point in time, over a thousand publications cited the NRGR as having contributed to their results.

A Dimensional Symptom Classification Approach

Overview

Modern psychiatry struggles to deal with the diagnostic heterogeneity that pervades psychopathologies, including Bipolar Disorder (BD). BD is characterized by oscillations of mood from manic highs to depressive lows and is a spectrum disorder with a clinical presentation that varies greatly in the population. BD is unique in that such variations in clinical presentation tend to bridge together the complexities of mood disorders with canonical symptoms of thought disorders such as psychosis, including delusions and hallucinations.

While BD has been studied extensively, not much is known about how it intersects with psychosis. A closer look at the genetic differences among individuals with BD and psychosis will pave the road to better understanding of both mood and thought disorders.

Sample & Phenotype Availability

Now, several years since phenotypic data were initially collected by various NIMH funded studies, Dr. Brzustowicz and her team at the NRGR are working on the Combined Analysis of Psychiatric Studies (CAPS) project. CAPS is an effort to direct resources toward maximizing already existing phenotypic and genotypic data, as opposed to disregarding previously collected data and launching brand-new studies—a time-consuming and costly task[17,18]. Dr. Brzustowicz's group has aggregated DNA samples from several studies, genotyped the samples with a modern high-fidelity genotyping array, and deposited the resulting data in the NRGR's data collections. At their core, the DNA samples that were originally collected haven't changed; however, researchers can now analyze the genetic data, which were collected several years ago as parts of various independent studies, but only recently genotyped or re-genotyped using denser marker panels[18]. Densely packed SNP arrays with high call rates afford large quantities of genetic material with which to conduct association

studies. Similarly, phenotypic-level data has not changed since the subject interviews took place and can be utilized to reveal detailed insights into the complexities of each unique presentation of BD.

Phenotypic information was gathered by various studies using the Diagnostic Interview for Genetic Studies (DIGS). The DIGS was developed by the NIMH Genetics Initiative to record information about a research subject's psychopathology[54]. The DIGS enables qualified clinicians to reach diagnoses for subjects by way of thorough questioning about past and current episodes of psychiatric illness. Subject responses have been collected by various studies, many of which were retained by the NRGR for future use.

Phenotypes of Interest

BD is diagnosed by assessing for current or past mood episodes (mood symptoms that represent a change in normal functioning and occur over an extended period of time)[55]. Mood episodes include mania, and hypomania. Psychotic features, including delusions and either auditory or visual hallucinations, often present with mood episodes[56].

A major depressive episode is defined by five or more of the following symptoms occurring most of the day, nearly every day, for at least 2 weeks and represent a change in normal functioning: sadness (or irritability), loss of interest (anhedonia), guilt, energy, concentration, appetite, slowed psychomotor behavior, sleep (insomnia/hypersomnia), and suicidality[57]. At least one symptom needs to be depression (irritability) or loss of interest.

Mania is defined as abnormally and persistently elevated or irritable mood, lasting at least one week (or any length of time if hospitalized)[55]. Symptoms include grandiosity or inflated self-esteem, being more talkative than usual or pressure to keep talking, decreased need for sleep, fleeting ideas or racing thoughts, an increase in goal-directed activity, and excessive involvement in pleasurable, and high-risk behavior[55]. The episode causes marked impairment in functioning, social activities or relationships with others, or necessitates hospitalization in order to prevent harm to self or others, or there are psychotic features[55].

Hypomania, a less severe form of mania, is defined as a distinct period of persistently elevated, expansive or irritable mood, lasting throughout at least four days, and is clearly different from the usual non-depressed mood[55]. Symptoms can be same as manic, but the mood disturbance is not severe enough to cause significant impairment in social or occupational functioning and there are no psychotic features[55]. However, the change must be observable by others and uncharacteristic of the person[55].

Bipolar Disorder is diagnosed when one has shown or currently shows evidence of mania, hypomania, or a mixed episode. Bipolar Disorder, Type I (BD1) is defined by one or more manic episodes alongside one or more major depressive episodes[55]. Additionally, psychotic features are commonly present alongside BD1[55]. Bipolar Disorder, Type II (BD2) is defined by one or more major depressive episodes with the presence or history of at least one hypomanic episode[55]. In such a case, there has never been a full-blown manic episode or mixed episode; otherwise, the diagnosis would be BD1. Cyclothymia is a related mood disorder in which the individual fluctuates between subthreshold depressive states and hypomanic states on a recurring basis[55].

Bipolar Disorder has a lifetime prevalence in the range of 1-4% in the American population[11]. It has an evenly split gender distribution, with an average age of onset just under 20[58,59]. Treatments for bipolar depression include pharmacological intervention, family-focused therapy, and education of symptoms, triggers, and habits[60]. Thus, since education and pattern recognition are major parts of treatment for BD and mania, early awareness of susceptibility could help patients increase their preparedness toward any symptoms so that they can have more awareness of, and thus control, over their actions. Additionally, BD is a recurrent disorder in which subsequent episodes are seen in approximately 90% of people who have experienced a first episode[59]. For these reasons, early prediction of genetic susceptibility for such BD could become an essential and global component of medicine in the future.

Dimensional Classification of Phenotypes

As previously mentioned, the PGC recently conducted a high-powered GWAS of BD that identified 30 novel loci associated with BD[52]. This was a high-powered case-control analysis, with over 50,000 cases and control participants. The reported loci are dispersed throughout the genome and are consistent with the polygenic nature of psychiatric disorders. In the publication, it is discussed that “Bipolar I disorder is strongly genetically correlated with schizophrenia, driven by psychosis, whereas bipolar II disorder is more strongly correlated with major depressive disorder.” However, in their analysis, PGC did not ultimately differentiate between BD1 subjects who comorbidly exhibit psychotic features from those who do not. In fact, we would like to draw attention to the possibility that BD1 without psychotic features (BD1-NP) may present a somewhat distinct genetic susceptibility profile from BD1 with psychotic features (BD1-P), which shares symptoms with schizophrenia and schizoaffective disorders. As such, a GWAS of BD1 that is further sub-classified to distinguish the presence or absence of psychotic features may reveal new insights into the ebb and flow from mania to depression that is characteristic of BD. Such a model may unveil findings that were previously hidden by the confounding factor of psychosis in the clinical presentations of just some study participants. We believe this will be the case even though the sample size, and thus the statistical power, of any one sub-classification will be less than the complete dataset. Among the 30 loci discovered by PGC in 2019, eight are also associated with schizophrenia, which may ultimately be accounted for by an analysis of BD1 with comorbid psychotic features.

To the credit of the PGC, one of the reasons that this study did not sub-classify their BD1 individuals could be for a lack of sufficient phenotypic data. A difficulty in psychiatric genetics is that unlike in other fields, a diagnosis rarely conveys the full clinical picture of an individual and may be a limiting factor in many analyses that seek to dive deeper than a broad label will allow. As aforementioned, the CAPS dataset in the NRGR is an aggregation of families who, over the past 30 years, have been interviewed in detail about their psychiatric history.

From these CAPS families, we sought out to conduct a meta-analysis of individuals with Type 1 (severe) BD, comparing subjects who comorbidly exhibit psychotic features (BD1-P) with subjects whose presentations of BD1 lack any psychotic features (BD1-NP). Such a study could aid our understanding in the etiologies of BD, as well as other complex disorders. Ultimately, this study could contribute toward reaching better health outcomes for individuals battling BD, as well as, more broadly, to mood and thought disorders, as well.

Significance

A deeper understanding of the genetic susceptibility factors for nuanced presentations of BD would have several benefits. First and foremost, results could be applied to polygenic risk scoring algorithms and utilized in the assessment of a patient's genetic risk burden for the manifestation of psychiatric disorders. Additionally, nuanced results could help the efforts of frameworks such as the RDoC, which are attempting to harmonize symptomatology with concrete genetic and biochemical data. RDoC aims to make the process of diagnosis and treatment selection in psychiatry less based on the subjective experience of the patient and more based on objective data; such an analysis aligns with this goal. Furthermore, such findings will ultimately help push forward the frontier in our understanding of the clinical heterogeneity with which BD presents and lends understanding to the unique pathophysiology of a specific subset of affected individuals. Finally, given that BD presents on a spectrum, with a variety of intersecting traits that are traditionally recognized as constituents of separate diagnoses, a deeper and more nuanced understanding of the distinct biological underpinnings of BD could inform nosology in modern psychiatry.

Conclusion

Prior GWA studies of psychopathology have aimed to cast a wider net in the hopes that a larger sample size would add sufficient power to association analyses so that strong findings could be uncovered. However, here we show that in the present state of psychiatry, traditional diagnoses blend symptoms to the point that too many variables are being simultaneously

assessed in any one association study. Detailed clinical interviews in our possession enable us to classify subjects with BD1 dimensionally by specific symptoms that are either present or absent from their overall presentation of BD. Specifically, we are particularly interested in psychotic features such as delusions and hallucinations, which are canonically associated with schizophrenia and related thought disorders, but that also pervade the diagnostic picture of patients exhibiting BD. Our study will demonstrate differences in the genetic associations among the different, dimensionally classified, groups of individuals. Even though the dimensional groupings reduce the statistical power of each group in comparison to the overall dataset (via the reduction of sample size), our data shows that signal from groups with more homogenous clinical presentations of BD, and thus a reduction of obfuscating noise, is valuable in addition to signal from a large sample size. Our results support a GWAS model of dimensionally classifying symptoms of BD by the presence or absence of psychotic features in order to discern distinct susceptibility genes for specific clinical presentations of BD. In the future, we will apply this model more broadly to many different psychopathologies and their intersecting symptoms.

MATERIALS & METHODS

Materials

Genotypic Data

Samples were obtained from the Combined Analysis of Psychiatric Studies (CAPS) project, housed at the NRGR. Biospecimens were genotyped with the Affymetrix UK Biobank Array, which used the GRCh37/hg19 reference genome build. This genotyping platform was designed for high-throughput genotyping of large cohorts and yielded genome-wide marker calls for 920,636 SNPs in the CAPS dataset. A total of 7,940 genotyping runs were reported in the Affymetrix quality control (QC) report. Of these 7,940 runs, 242 were failed runs, 84 positive controls, and 84 negative controls (both standards that help report on the quality of samples). At this stage, 7,530 successful runs—potential samples—were captured into CAPS dataset genotyping files.

Phenotypic Data

To identify individuals who exhibit psychotic features or were substance abusers prior to or at the same time as the onset of psychotic symptoms, we employed DIVER, a program developed at the Battelle Center for Mathematical Medicine at Ohio State University. DIVER is a database that stores phenotypic subject data, including DIGS assessment responses for many subjects within the NRGR, including those in the CAPS dataset. DIVER was used to gather all requested variables (i.e., responses) for any question within the DIGS. The output was a batch of comma separated (CSV) files that each contained a subject identifier in one column and a coded response in another column. In typical binary questions, the response ‘no’ was coded ‘0’, whereas ‘yes’ was coded ‘1’.

Methods

“PhenotypeManager” (PM) is a Python class of methods that I constructed to: (1) process identifiers (IDs), (2) link family members, (3) disqualify and prune individuals and/or families that did not meet inclusion criteria, (4) assign formal diagnoses and sub-phenotypes (specific symptoms) to individuals, and, ultimately, (5) filter the larger dataset into smaller

datasets based on specific combinations of phenotypes and sub-phenotypes. This was largely done using “NRGR Identifier Dictionary” (NID), a computational data lookup table that I constructed to aggregate various sources of information and ID conversions for hundreds of thousands of individuals in the NRGR.

“GenotypeProcessor” (GP) is another Python class I constructed that utilizes PM as a dependency and was used to create files for later recoding the identifiers in the genotypic datasets to match the phenotypic IDs, as well as to have proper family IDs, sexes, and parental IDs coded into the files. I used GP to output several files for later recoding by the PLINK association analysis software[61]. The files “recode_ids.txt”, “recode_sex.txt”, and “recode_pars.txt” were outputted to respectively recode identifiers, sexes, and parent identifiers using PLINK commands. The following PLINK flags were utilized on all retained individuals: “--update-ids” utilizing “recode_new_ids.txt”, “--update-sex” utilizing “recode_new_sex.txt”, and “--update-parents” utilizing “recode_new_pars.txt”.

“BuildDiverDicts” is a script I created to assign gathered sub-phenotypes to individuals using DIVER output files, which contained subject IDs and the coded responses. BuildDiverDicts was constructed to convert each file into a data frame and, subsequently, into a dictionary. The dictionaries were all combined into one large dictionary and were exported as a JSON file “diverPhensDicts.json” to be used by PM to assign sub-phenotypes (functionality #4). Upon filtering of the dataset into subsets, counts were collected using pivot tables in Excel.

PLINK is an open-source whole genome association analysis toolset designed at The Broad Institute that performs a range large-scale analyses in a computationally efficient manner[61,62]. PLINK was used to perform quality control steps and remove samples and SNPs with low call rates, remove SNPs that were out of Hardy-Weinberg Equilibrium, and remove samples with a high rate of Mendelian errors[62]. PLINK was then used to run the TDT analysis. Manhattan and Q-Q plots resulting from the analysis were created in the R programming

language. Ggplot2, plotly, and qqman are R libraries that helped to create high-quality vectorized figures[63].

SNPs were reported in the genotypic files as Affymetrix probe IDs, not typical RSIDs as would be found in a database like dbSNP. Some probe IDs were easily converted to RSIDs using an Affymetrix annotation file, but many were not. In order to be able to compare findings with the literature, I used the UCSC Genome Browser (GRCh37/hg19 build) to search for RSIDs of interest manually, using the chromosome number and base pair position of any particular SNP (and verified that the major and minor alleles on the browser were those that I was expecting). I used LDlink to create zoomed in Manhattan plots of the areas of interest and to create heatmap matrices that indicate linkage disequilibrium r^2 and D' values between SNPs in the regions of interest[64,65]. In cases where I wanted to scan a large range of SNPs for LD above a certain threshold, I used the GRCh37 Ensembl Linkage Disequilibrium Calculator. I then searched SNPs of interest (including SNPs in LD with findings from this analysis) on the NHGRI-EBI Catalog of human genome-wide association studies. I used LDlink's summarized queries of the Human Genome Atlas and Roadmap Epigenomics Consortium for RNA expression and histone modification in the brain, for the genes of interest. SpliceAI, a machine learning algorithm that predicts the likelihood that a SNP variant might impact gene splicing was used to analyze the effects of significant SNPs located in an intronic segment of a gene[66].

RESULTS

Phase 1: Harmonizing Genotype IDs & Phenotype IDs

Identifier Mapping

In order to conduct an analysis using the deidentified identifiers (IDs) listed in the CAPS dataset, I needed to first have the ability to (1) pull information on each individual's diagnosis, familial relationships, and demographic background and (2) link genotypic-level data to phenotypic-level data for each subject in the CAPS dataset. NRGR distribution files are spreadsheets that contain pertinent diagnostic, demographic, and family information about each subject that appears in NRGR-distributed datasets, including CAPS. Some variables contained within NRGR distribution files include "ind_id" (individual identifier), "mother_id", "father_id", diagnosis, and race. Infinity BiologiX (IBX) (formerly the Rutgers University Cell and DNA Repository, RUCDR) is responsible for storing biomaterials for the NRGR. IBX catalogs new sample accessions with "RUIDs", IDs that are different from the NIMH identifiers for phenotypic data. RUCDR generated the genotypes for the CAPS project using RUIDs as sample identifiers. Distribution files, along with additional files from IBX, enable the linking of genotypic and phenotypic IDs and harmonization of the two streams of data that exist for each subject in the CAPS dataset.

To be able to quickly access this information programmatically, I created "NRGR Identifier Dictionary" (NID). A dictionary is a highly efficient programming data structure that takes in a "key" input and returns a "value" as the output. To maximize the use of this data structure for NID, I aggregated information from several relevant distribution files (e.g. bipolar, depression, controls) into one data frame, along with the additional mapping files from IBX. When constructing NID, I set each key to the individual's genotypic ID (RUID) and the resultant value to a comprehensive dictionary containing all of that subject's personal deidentified information from distribution and IBX files (including the phenotypic identifier). Additionally, a separate—inverted—version of NID was also created to accept a phenotypic ID as input in order to receive a subject's informational dictionary in return.

Ultimately, NID was converted into a JavaScript Object Notation (JSON) lightweight-interchange formatted file in order to maximize storage and computational efficiency throughout the pipeline that followed. Running the code that built the NID file required approximately 2 minutes of runtime. In comparison, loading NID from a JSON file into the GWAS pipeline consistently required less than 3 seconds of runtime. Additionally, there was no lag when searching for keys in NID because of the high computational efficiency of the dictionary's underlying data structure. Thus, the construction of NID was a worthwhile endeavor because it prevented several occurrences of computationally taxing and time-consuming lookup operations among an already computationally taxing set of programs to be described further in the sections that follow. Additionally, since NID served as an overarching lookup tool, it allowed for quick interchange between phenotypic and genotypic IDs throughout the pipeline described in the following sections.

Pedigree Analysis

The "MapIDs" method from PM used NID to map phenotypic-level data to the genotypic-level data. Using the quality-control report generated by the Affymetrix analysis software, I was able to collect the sample IDs and, using NID, convert them into NRGR phenotypic identifiers. Additionally, at this stage I captured each individual's diagnosis and ethnic background, as well as the IDs of each individual's mother and father, when applicable.

In some cases, individuals who serve as parents in the dataset themselves had parental identifiers map to them through NID. Not all individuals present in the NRGR were part of this CAPS dataset, and, as such, in some cases subjects had too much information for the sake of this analysis. Thus, if these individuals who serve as parents in the CAP dataset do not also serve as children (i.e. in a multigenerational pedigree), then they were intentionally not assigned parents (rather, they were coded '0', representing 'not applicable'). In such a case, the parent of the parent in the CAPS dataset would serve no purpose when running the TDT analysis since they do not themselves contribute to any nuclear family.

Before conducting the family-based association study, it was imperative to determine the familial relationships among the subjects in the CAPS dataset. As previously discussed, mother and father identifiers were added to the original data frame after being passed through the MapIDs method. This information afforded me enough information to design an algorithm, “BuildPedigree”, that (1) links parents to children, and vice versa, and (2) assigns each subject in the study a relative role of mother, father, or child.

Inclusion Criteria

The inclusion criteria for a TDT require that each individual map to a family that ultimately contains a mother, a father, and at least one child. Each parent and affected child must have genotypic level data available for analysis. The TDT uses parent transmitted alleles as case alleles, and untransmitted alleles as control alleles. However, the analysis itself is focused only on the phenotypes of the affected children. Therefore, only the affected children require phenotypic level data for the analysis. Since the individuals being mapped to families were gathered from the genotyping quality-control report, we are confident that each individual was genotyped. Similarly, because each individual successfully maps to the NRGR’s distribution files, we are confident that each individual has accessible phenotypic data within the NRGR. Thus, individuals who do not map to both a mother and a father, or a spouse and at least one child were pruned from the dataset prior to export into a spreadsheet. As a corollary, any families present in the dataset that were not comprised of at least one of each mother, father, and child were disqualified from this study.

Counts Prior to Quality Control & Phenotype Filtering

The resultant families were exported into an Excel file and each placed on a separate row. Each family was assigned an arbitrary number within this dataset, in column A, followed by the ID of each individual in the family, starting with the mother’s ID, followed by the father’s ID, and then each child. The resulting dataset contained 610 family pedigrees (where all extended families were joined into one larger pedigree) comprised of 3229 individuals. Some families

were multigenerational and, as such, a total of 735 nuclear families were available to serve as trios. Among these trios was a minimum family size of 3 (1 child, 2 parents) and a maximum size of 13 (11 children, 2 parents). There was an average of 4.4 members per family in the raw dataset, with a standard deviation of 1.3 members and a mode of 4 member-families.

Recoding Sample Identifiers

The original CAPS genotype files were saved with Registration IDs as individual identifiers, and no family identifiers were saved (all were coded '0' for 'not applicable'). GP uses the Affymetrix QC report and NID to map each unique sample Registration ID to its corresponding NRGR RUID, family identifier, father identifier, and mother identifier. In order for PLINK to recognize family relationships, it was necessary to recode the genotype files so that family identifiers can link different samples (by assigning members of a family the same family ID), and so that individuals can be harmonized with the IDs associated with phenotypic data.

Phase 2: Cleaning & Quality Control

Duplicate Sample Pruning

Some samples were successfully genotyped twice on the Affymetrix Biobank array after an initial poor-quality genotyping. This led to 121 cases where individual genotyping runs mapped to the same RUID. Call rates of duplicate samples were compared; only the sample with the higher call rate was included in the study dataset. Each of the lower-quality samples among the 121 duplicates were saved to a file, 'CELS_REMOVE.txt', for later removal using the PLINK "--remove" flag. At this stage, 7,409 successful runs—potential samples—were captured into CAPS dataset genotyping files.

Cleaning Samples with Pre-Identified Genotyping Issues

Previous analyses of the CAPS dataset by Dr. Veronica Vieland's group at the Battelle Center for Mathematical Medicine have discovered some errors in the genotyping for a subset of individuals. Mendelian errors were detected in some subjects. Among other subjects, family relationships were found to be incorrect. Additionally, some genotyping runs yielded low

resolution along the sex chromosomes and left the true sex of such samples uncertain. Thirty-five samples had high Mendel errors; the genotypes of these samples were removed, or "zeroed-out". Similarly, 14 samples with unclear or borderline ambiguous sexes were zeroed-out at the X chromosome. One sample swap and one family reassignment were also accounted for and was executed with the "--update-pars" command.

'WriteZeroMarkers' is a GP method I wrote to create the files that document the subjects for whom to zero-out all of the chromosomes or just the X chromosome. Zeroing-out works by utilizing the "--zero-cluster" PLINK command after extracting a list of SNPs to remove from the total set of 920,639 SNP markers in the Affymetrix UK Biobank array. Zeroing-out the X-related SNPs of individuals required recoding only SNPs located on chromosome 'X' (coded as 23) or 'XY' (pseudo-autosomal region; coded as 25). Among the X chromosome and pseudo autosomal region alone there were 37,245 SNPs, which were zeroed-out in 14 individuals.

Quality Control in PLINK

Remove Subjects with a High Proportion of Missing Genotype Calls

The "--mind" flag was run with a cutoff of 0.03, which removed any DNA samples with a call rate of less than 97%. In total, 920,636 variants and 7,493 people (3,060 males, 4,443 females) were loaded from the genotype file. Forty people were removed due to missing genotype data; thus, 7,453 individuals passed filters and QC. The number of variants was not changed at this point.

Remove Markers Below Threshold Genotyping Rate

The "--geno" flag was run with a cutoff of 0.02, which removed any SNPs that were missing more than 2% of the genotypes. A total of 920,636 SNP variants were loaded. Of these, 29,977 variants were removed due to missing genotype data; thus, 890,659 variants passed filters and QC. The number of individuals was not changed at this point.

Remove Markers That Are Out of Hardy-Weinberg Equilibrium

The "--hwe" flag was run with a cutoff of 0.00001, which removed any SNPs calculated to be out of Hardy-Weinberg Equilibrium with a p-value less than 0.00001. In total, 890,659 variants and 7,453 people (3,041 males, 4,412 females) were loaded. Of these, 97,409 variants were removed due to having a Hardy-Weinberg exact test statistic below the threshold. Thus, 793,250 variants passed filters and QC. The number of individuals was not changed at this point.

Remove Subjects with a High Frequency of Missing Calls, Again

The "--mind" flag was run again, with a cutoff of 0.05 to remove any DNA samples with a call rate less than 95%. This was just to make sure that the previous two cleaning steps didn't remove enough SNPs to have some samples slip below a good threshold. A total of 793,250 variants and 7453 people (3041 males, 4412 females) were loaded in. No individuals were removed due to missing genotype data and, thus, the original counts were not changed at this step.

Remove Highly Rare Variants

Some SNPs do not show any variation and, as such, can be removed. The "--maf" flag was run with a cutoff of 0.001, to remove SNPs with a minor allele frequency (MAF) below 0.001. Similarly, if only two or three people in the whole sample set have a particular minor allele, that can be indicative of an artifact that should be removed (or at best something that won't be powerful enough to yield any results). In total, 793,250 variants and 7,453 people (3,041 males, 4,412 females) were loaded. Of these, 182,737 variants were below the minor allele threshold and were removed. Thus, 610,513 variants passed filters and QC at this stage. The number of individuals was not changed at this point.

Remove Families or Markers with High Proportion of Mendel Errors

Lastly, the "--me" flag was run with the cutoff values of 0.05 and 0.01. The first cutoff, 0.05, represents the threshold at which PLINK will remove entire families where there is a

problem with greater than 5% of all SNPs. The second cutoff, 0.01, represents SNPs that have greater than 1% of errors across all families. In total, 610,513 variants and 7,453 people (3,041 males, 4,412 females) were loaded in. A total of 3,089,134 Mendel errors were detected. To account for this, 28,164 variants and 123 individuals were excluded. Thus, 582,349 variants and 7,330 people passed filters and QC at this stage. The total genotyping rate was 0.997518.

Reassessing Family Inclusion Status

After completing the quality control steps, I ran a script I wrote that parses individuals in the output files from before the first use of the Mendel command and after the second (and last) use of the Mendel command. This program then collects the discrepant individual IDs—those who were present in the pre-dataset but pruned from the post-dataset. These individuals were saved to an Excel file to be later removed from the phenotypic subsets.

Phase 3: Sub-Phenotype Collection

Traits of Interest

Psychotic features are known to occur in some individuals who are diagnosed with BD1. Typically, they will present as delusions of varying type, intensity, and bizarreness. Sometimes they will be accompanied by auditory hallucinations. At other times, they will be accompanied by visual hallucinations (however, visual hallucinations often point to other neurological conditions). All cases of psychosis will present with some combination of these factors and can be distinguished from individuals who suffer from mood disorders without any psychotic features.

In order to analyze true cases of psychotic features presenting with a mood disorder, it was prudent to remove individuals (and exclude any families that were subsequently disqualified) who had a history of substance abuse prior to occurrences of any particular psychotic event, up to and including the point at which psychotic features began to manifest. Certain substances, such as stimulants like cocaine or hallucinogens like PCP, are known to induce psychosis in individuals, and are sometimes attributed as triggers for the later onset of psychotic disorders. As such, we opted to disqualify any individuals for whom the environmental

pressures leading to the onset of their psychotic features were greater than those of the other individuals in the dataset.

In this relatively small dataset, controlling for other environmental triggers such as trauma from emotional, physical, or sexual abuse, as well as neglect or long-term distress would greatly reduce the power of this analysis through a large reduction of the sample size. At the present moment, this project balances depth of phenotyping and sample size, as they are inversely related since no two individuals share identical life experiences. Moving forward, as we scale up our project, we hope to include more phenotypic filters and, thus, more dimensions to our analysis.

DIGS Variables for Traits of Interest

The variable I17704 was taken from the psychosis section of the DIGS. The item poses the following question: “Since you first began experiencing (psychotic symptoms) have you ever returned to your normal self for at least two months?” Individuals who were prompted to respond to this question have all already been determined to have had previous psychotic episodes, either currently or in the past. Thus, responses to this question give a detailed picture of the subject’s psychotic experiences. We can further deduce that a person who responded ‘no’—that they have not returned to their normal self since the onset of their symptoms—likely has experienced more severe symptoms than someone who responded ‘yes’—that they have had periods of returning to their normal selves since the initial onset of their psychosis. Thus, this single question, unique to the highly descriptive DIGS interview, serves as an effective way to identify psychotic individuals for this study. However, it is possible that this variable should have been asked by the interviewer but was left out or the section was skipped. Thus, it is possible that data from this one question may be missing in some individuals’ responses.

The variable I10940 is a more straightforward question which asks subjects if they have ever had a delusional episode. This question is ancillary to the previous one: individuals who responded ‘yes’ to this question but did not have a response to the psychosis question (i.e. have

never fully expressed a psychotic episode) are informative because they indicate a possible psychotic experience. This variable serves an important purpose in that it may reveal some individuals who have experienced psychotic features alongside their BD, but for some reason were not reported via the previous variable (I17704).

The variable I17748 delineates substance abuse history, when applicable. Individuals who indicated that they were substance abusers (of cocaine, hallucinogens, etc.) prior to or at the same time as the onset of their psychiatric disorder were excluded in order to eliminate any cases of substance induced psychiatric disorder. Individuals who became substance abusers only after their symptoms came about and thereafter were retained, since it is not uncommon for individuals to attempt to self-medicate psychiatric symptoms with substances.

Finally, to further control for other sources of diagnostic variation, we found a DIGS variable “SEV_MANIA_DAYS” that reports on the number of days in which a subject has claimed to have experienced a severe manic episode, in line with BD1. As such, in order to be confident in the severity of the manic episodes exhibited by the subjects of my analysis, we opted to exclude any individuals who did not have a response for this question or had a response equal to “zero days”.

Phase 4: TDT in PLINK

Phenotypic Subset Construction

‘DatasetConstructor’ is an algorithm I constructed that uses the complete dataset with previously added phenotypes and sub-phenotypes, and creates an Excel spreadsheet of individuals for each subset of interest—namely BD1-NP, BD1-P, and BD1-MIX. Note that after this filtering step, all remaining families were reassessed for their fulfillment of the inclusion criteria. Additionally, using the Excel file with QC disqualified individuals created at the end of the QC steps, I removed these disqualified individuals from the phenotypic subsets, as well. Following this pruning step, if exclusion of certain individuals disqualified any family from the

criteria required (two parents and at least one affected child), that entire family was now pruned from the phenotypic subsets.

Sibling Inclusion Status

All siblings who are not documented with the symptoms of interest were removed from any particular subset. In certain cases, unaffected siblings can increase the power of the TDT statistic; however, this is generally not the case when studying psychiatric disorders with relatively low penetrance. In the case of a highly penetrant trait, the genotypes of unaffected siblings would be more likely to display an inverse relationship where unaffected siblings do not have a particular allele that affected children do have. When penetrance is low, however, the segregation of alleles in unaffected siblings will not deviate much from a mendelian ratio of inheritance. In such a case, a mendelian ratio would simply increase the overall randomness of the distribution. This, in turn, would reduce the power of the study at any given locus. Thus, in this study, the retention of unaffected siblings would likely unnecessarily reduce signal and obscure potential findings[34].

Genotypic Subset Construction

Following phenotypic subset construction, the PLINK commands “--keep” and “--pheno” were used on the genotype files to subset individuals of interest and recode the phenotypes of these individuals, respectively. This process was carried out for each of the three groups: BD-MIX, BD1-NP, and BD1-P. Each subset was filtered down from the clean master dataset. In the final datasets, all remaining children are affected and thus assigned a phenotype code of “2”. Furthermore, parent phenotypes are not considered in order to avoid population stratification, and so their phenotypes are coded “0”. As previously mentioned, all unaffected siblings were removed while identifying individuals of interest and their parents.

Counts After Cleaning, QC & Phenotype Filtering

After cleaning, quality control, and filtering for phenotypes, several families were pruned from the analysis. In reporting the counts of the remaining families, the number of “working”

families refers to the number of nuclear families present in the subset, regardless of whether or not certain nuclear families are connected in the same pedigree. However, it is often the case that one individual will appear in one nuclear family as an affected child, but also as a parent in another nuclear family. Additionally, two separate nuclear families may appear in the same family due to sibship among parents. Counts that consider such families as one are referred to by a total count of pedigrees in this study. The following counts were determined using pivot tables in Excel.

BD1-MIX Subset

Among the BD1-MIX subset there remained 236 total pedigrees: 256 working families that consists of 899 individuals (table 1). Of the 236 pedigrees, 103 contain three members (i.e. one affected child), 101 contain four members (i.e. two affected children), and 14 contain five members (i.e. three affected children). There are an additional 18 pedigrees (116 individuals) that contain more than one nuclear family and can be broken down into 38 working families. Five of these are multigenerational pedigrees, where an affected child in one nuclear family is the parent in another. The other 13 pedigrees consist of parents who are siblings and with children who are cousins. In sum, of the 899 individuals, 391 are affected children, 508 are parents.

BD1-NP Subset

Among the BD1-NP subset there remained 129 total pedigrees: 138 working families that consisted of 449 individuals (table 2). Of the 129 pedigrees, 88 contain three members (i.e. one affected child), 31 contain four members (i.e. two affected children), and one contains five members (i.e. three affected children). Additionally, there is one pedigree containing five individuals that can be broken down into two families of three individuals, where one individual is present in both. There are five pedigrees of size six that can be split into separate working families of size three. Finally, there are three pedigrees of size seven that can be split into

separate working families of size three and four. In sum, of the 449 individuals, 173 are affected children, 276 are parents.

BD1-P Subset

Among the BD1-P subset there remained 155 total pedigrees: 161 working families that consists of 537 individuals (table 3). Of the 155 pedigrees, 99 contain three members (i.e. one affected child), 44 contain four members (i.e. two affected children), and six contain five members (i.e. three affected children). There are an additional 6 pedigrees (34 individuals) that contain more than one nuclear family and can be broken down into 12 working families. Two of these are multigenerational pedigrees, where an affected child in one nuclear family is the parent in another. Another has two separate families connected by one mother. The other three pedigrees consist of parents who are siblings, with children who are cousins. In sum, of the 537 individuals, 218 are affected children and 319 are parents (one parent serves in two families).

Transmission Disequilibrium Test

At this point, we set out to run the TDT on each subset using the PLINK "--tdt" flag. The PLINK TDT flag calculates *p*-values for each of the 582,349 SNP variants that remained in the dataset after cleaning and quality control. Each *p*-value represents the probability that the association between a particular SNP and the phenotypes of interest are due to chance. As such, the lower the p-value, the more likely that the association between a SNP and the phenotypes in question are not spurious and may be causally related. Thus, a SNP with a small *p*-value would be indicative of a genetic basis for the phenotypes in question at or nearby the location of the SNP.

Optional Reduction in Computational Intensity of Analysis

Certain methods of analysis are highly computationally intensive. In such cases, using small pedigrees lessens the computational intensity. Therefore, at the risk of inflating association signal, breaking down multigenerational families into separate nuclear families in order to scan the genome for association is often a good approach when seeking to minimize

the intensity of the computations. After resulting areas of interest are discovered, another—standard—association can be run on the areas of interest using the original pedigree structure. In this way, accurate results can be obtained in a less computationally intensive manner.

PLINK generally executes its tasks very promptly; this issue did not arise in the CAPS dataset. However. other methods are much slower, such as calculating the Posterior Probability of Linkage (PPL). Thus, although we did not need to use this two-step approach for this dataset, I have built functionality in my GWAS pipeline to allow for utilization of this approach. This enables strong functionality, even in the face of massive sample sizes.

If initiated, the code will separate nuclear families that are connected by one multifaceted individual in a larger family pedigree. Among the multigenerational families, the family identifiers of the nuclear families in which the multifaceted individual is a child will all be appended by an "R" and saved in a recoding file, "recode_faceted_fams.txt", next to the original identifier. For example, the family identifier "15-71-96065" would be recoded to "15-71-96065R" in whichever family the multifaceted individual was a child. The choice of recoding nuclear families where the multifaceted individual is a child was arbitrary and chosen for the sake of consistency and computational simplicity. Using the "--update-ids" flag, I have set the program to recode all of the subject records saved to "recode_faceted_fams.txt", save for the multifaceted individual in the family.

I had saved the original identifiers of multifaced individuals to a separate file, "multifaceted_inds.txt", and used the "--keep" flag to save each of these individuals' record of genotype data. In a separate file, "recode_multifaceted_inds.txt", I had saved the modified family identifier (identifier plus an appended 'R') and used the "--update-ids" flag on the individuals in this small subset of individuals to create a record for each of these individuals that will connect each subject to the individuals with the modified family identifier. After setting the program to recode the parent IDs in the duplicate multifaceted individual entries, I have set the program to use a flag "--bmerge" to combine the CAPS dataset with the second facet of the

small subset of multifaceted individuals. This procedure allows for the desired result where multigenerational families are separated into unique families of smaller size, and thus, lower computational complexity.

Findings

BD1-MIX Subset

The following loci correspond with the most significant SNP association findings in the BD1-MIX subset (table 4), with *p*-values < 5.0E-06: chr23:19,781,333 (p=2.54E-14), chr8:59,164,297 (p=8.35E-07), chr14:106,321,212 (p=1.5E-06), chr18:29,101,207 (p=1.62E-06), chr9:20,922,554 (p=4.23E-06) (figure 1).

BD1-NP Subset

The following loci correspond with the most significant SNP association findings in the BD1-NP subset (table 5), with *p*-values < 5.0E-06: chr23:19,781,333 (p=4.3E-08), chr6:158,349,893 (p=1.9E-06), chr6:158,329,778 (2.6E-06), chr6:158,317,132 (p=2.6E-06), chr6:158,340,497 (p=4.3E-06). Note that four of these five SNPs are located in the vicinity of one another and are thus suggestive of interesting results (figure 2). The RSIDs for these four adjacent SNPs are, rs2273070, rs3840366, rs6924813, and rs3047738.

BD1-P Subset

The following loci correspond with the most significant SNP association findings in the BD1-P subset (table 6), with p-values < 5.0E-06: chr23:19,781,333 (p=1.137E-07) and chr13:41,995,485 (p=3.46E-06) (figure 3).

DISCUSSION

Significance Thresholds in GWAS

A large multiple testing correction must be made to the typical alpha of 0.05 when dealing with p-values in a study that conducts numerous tests, such as this one with 582,349 polymorphisms analyzed. A Bonferroni correction is likely too conservative for GWA studies because the different variants assessed are not being tested independently due to linkage disequilibrium (LD)[36]. However, a large correction is quite necessary; the current consensus is that genome-wide significance can be asserted at an alpha of $5x10^{-8}$[36,67]. P-values below $5x10^{-8}$ are not conclusively significant in a GWAS because multiple factors could lead to inflated results at any locus. However, studies have also demonstrated several instances where results with p-values below $1x10^{-7}$ are replicable and acceptible[36,67]. Furthermore, depending on the minor allele frequency (MAF) of the SNP in question, there may be more flexibility in terms of setting a significance threshold[36]. In fact, it has been determined that *p*-value thresholds of $1x10^{-6}$, $7x10^{-7}$, $5x10^{-7}$ and $3x10^{-7}$, are reasonable for MAF ≥ 5%, MAF ≥ 1%, MAF ≥ 0.5% and MAF ≥ 0.1%, respectively[36].

Another point to consider depends on the fact that many SNPs are not completely independent of neighboring SNPs, due to LD. If the p-value for association between a SNP and the symptomatology of interest is very small, one would expect that other SNPs in LD with the SNP of interest would also have a smaller p-value than surrounding SNPs that are not in LD with the SNP of interest[68]. As such, we often expect to see clustering of low p-value findings at loci of interest. However, in the event that there is no LD between the significant SNP and its neighboring SNPs, no clustering would be anticipated.

Thus, in this study, results were plotted on Manhattan plots (figures 1-3), with a lower-bound threshold of $5x10^{-6}$ for interesting findings, and an upper-bound of $5x10^{-8}$ for genome-wide significant findings. However, MAF adjustments to the significance threshold are discussed whenever it is prudent to do so.

Summary of Findings

Figure 1a displays results for the BD1-MIX pooled samples in a Manhattan plot. In this sample space, five SNPs surpassed the lower-bound threshold. The lead SNP, rs12010076 (chrX:19781333) returned a p-value of $2.5x10^{-14}$, which surpassed the threshold for genome-wide significance. This SNP is not in LD with any of its neighboring SNPs (figure 1b), which explains the lack of clustering at this locus[64]. The minor allele 'A' has a frequency of 0.016 in this dataset; globally the MAF=0.01, and in European populations it has a MAF=0.00072[69]. This SNP is an intronic variant within the SH3KBP1 gene and has a SpliceAI Δ score of 0.00, indicating that variants at this locus are unlikely to impact splicing of the exons in this gene[66]. The clinical significance of this variant is still unknown today and has yet to be characterized.

The following SNP, rs7003372 (chr8:59164297), returned a p-value of $8.4x10^{-7}$, which surpassed the lower bound threshold for significance. With MAF=0.3968 in European populations, this SNP comfortably surpasses the $1.0x10^{-6}$ MAF-adjusted significance threshold[36,70]. This SNP, rs7003372 is in minimal LD (r^2=0.014, D'=0.515; 323,373 base pair separation) with the 6th most significant SNP finding, rs1992045 (MAF=0.082), of a BD and schizophrenia meta-analysis conducted in 2010 ($p=1.67x10^{-7}$)[71]. An association plot and LD matrix of this specific region and the relationships between these SNPs can be seen in figures 1c and 1d, respectively. Although the r^2 parameter is low to begin with, the lack of a significant finding at rs1992045 itself in this dataset casts doubt on the potential significance of rs7003372. Furthermore, the frequency at which the minor allele was transmitted in the dataset is 0.629, which does not coincide with its MAF in global or European populations.

SNPs rs281865422, rs553299589, and rs117655852, rs76757914, and rs58793557 do not have R' values of LD greater than 0.8 with any neighboring SNPs (determined using Ensembl GRCh37 Linkage Disequilibrium Calculator). Previous findings have demonstrated that rs553299589 is a pathogenic SNP associated with cardiomyopathy with a very small MAF in the European population. Although rs76757914 is an intronic variant, it has a SpliceAI Δ of 0.00 and

therefore does not likely impact splicing patterns. None of these SNPs have been previously associated with BD in the literature and, as such, require further characterization. SNP rs75565666 is in LD with 20 adjacent SNPs; however, none of these other SNPs are present in this dataset to verify the result. SNP rs3840366 on chromosome 6 (European MAF=0.226), with a p-value of $1.2x10^{-5}$ is in strong LD (r^2=0.9472, D'=1) with a neighboring SNP, rs2273070 (European MAF=0.232), which returned a p-value of $9.1x10^{-5}$. These chromosome 6 findings are below the threshold for significance (MAF adjusted or not) but are nonetheless intriguing. Overall, some of these SNPs may be true associations with BD. Future work will include validating these findings by thoroughly analyzing these loci, in a study of greater power.

The BD1-MIX cohort contained 899 individuals, a fairly modest sample size. However, having conducted a family-based GWAS, which is robust to within-family population stratification, and with families of predominantly European descent that minimize between-family population stratification, our dataset does confer some sample-size leniencies in comparison to large scale case-control analyses. As such, the aim of this project was to create smaller subsets from the pooled dataset with the hopes that reducing the clinical heterogeneity seen in cases of BD would allow for further discernment of signal, even in the face of reduced sample sizes.

In taking a look at the BD1-NP cohort, which—with a sample size of 449—has less statistical power than the BD1-MIX cohort, it is interesting to see a breakout cluster of four SNPs with p-values less than $5.0x10^{-6}$ (figure 2a). This cluster of four SNPs are intron variants: rs2273070, rs3840366, rs6924813, and rs3047738 on 6q25.3 and indicate early signs of association specifically with non-psychotic bipolar disorder (figure 2b). The p-values for these variants are $1.85x10^{-6}$, $2.6x10^{-6}$, $2.6x10^{-6}$, and $4.3x10^{-6}$, respectively. Each of these SNPs has a MAF in the range of 0.24-0.26 (figure 2c), which is similar to the observed frequencies in the dataset in the range of 0.29-0.30. These SNPs are all in strong LD with one another (figures 2d-e) and are located along the Sorting Nexin 9 (SNX9) gene, which encodes a protein of the Sorting Nexin family and is involved in intracellular vesicle trafficking[72,73]. Not much is yet

understood about this particular gene, however it has previously been associated with Borderline Personality Disorder[74]. SpliceAI Δ scores at these SNPs are quite low, thus suggesting that the mechanism by which the variants act does not involve altering splicing patterns. Rather, it is possible that these intronic variants could potentially affect regulatory sequences and thus expression patterns, or epigenetic states. Figure 2f shows the RNA expression overview of the SNX9 gene transcript in various tissues from the Human Protein Atlas. SNX9 shows low tissue specificity and is expressed throughout the central nervous system. Figure 2g shows the SNX9 regulatory chromatin states from histone ChIP-Seq from the Roadmap Epigenomics Consortium. This figure demonstrates that certain chromatin states as well as methylation and acetylation patterns are present in brain tissues and are specifically not present in other tissues throughout the body, and thus indicate that this is a potential locus of interest for epigenetic analyses. It is important to note, however, that (due to LD) a nearby gene may truly be causing the association and not SNX9. Nonetheless, this cluster of SNPs is indicative of a potential significant finding that can be further validated through a replication study with greater statistical power.

More interestingly, this association does not hold among the BD1-P group (n=537) (figure 3). In fact, in the BD1-P group, those four SNPs of interest completely lack significance. Whereas the lead SNP of the cluster has a p-value of 1.85×10^{-6} in the BD1-NP cohort, in the BD1-P cohort that same SNP has a p-value of 0.8571 (no significance), a 171,514-fold difference. Furthermore, returning to the BD1-MIX cohort, it is clear that the 4 SNPs of interest on chromosome 6 show an intermediate level of association when the two psychotic/non-psychotic features groups are pooled, further supporting our findings (figure 1a). In fact, two of these SNPs were directly discussed among the top-10 significant findings of the pooled sample space.

Quantile-Quantile (QQ) plots for each of these three groups tend to match the quantiles of a normal distribution overall, except for an unknown effect that has caused the distribution to

stagger along the line of best fit (figures 1e, 2h, and 3b). As such, these plots suggest both that there may be a slight batch effect among the genotypes and that greater power is needed in order to confirm these results.

Especially given that the sample size has decreased in the subsets in comparison to the pooled samples (and thus the statistical power of our analysis has diminished), these highly stratified findings on chromosome 6 are validate our hypothesis. Such a cluster of findings with increased resolution in the BD1-NP cohort relative to the pooled samples, and a clear lack of significance at this locus among individuals with psychotic features, indicate that clinical heterogeneity obstructs potential findings in GWA studies. Such findings demonstrate that more efforts need to be made to homogenize subjects according to the specific traits that together manifest complex diseases with many comorbidities.

Considerations for GWAS

Ethics

As previously discussed, population stratification is a legitimate concern when conducting large-scale genetic analyses. Data analyzed from GWA studies are only meaningful when cases are matched to controls in terms of ancestry. Furthermore, valuable associations gathered from GWA studies can only be applied to the populations on which they were carried out[37,75]. For example, if a GWAS was conducted on a large group of Northern Europeans, results from the GWAS would only benefit individuals of European decent. GWAS and polygenic risk scoring can thus lead to significant healthcare disparities if action is not taken to prevent it. Whereas populations that are selected for the studies will have better diagnostic information to treat disorders that span the globe, those populations that are neglected from studies will be even more so neglected later on[76]. Thus, in order to level the playing field, more studies need to be carried out that represent populations from all over the world[75]. This will allow such work, and other work in the field, to ultimately increase our knowledge of the genetic basis of complex diseases[76,77].

Limitations of the Field

While such an analysis will help contribute to the vast amounts of data that are being amassed—and potentially from a unique perspective that will help researchers to better understand the etiology of specific complex phenotypes related to BD—locating the causative genes in a vacuum will not directly open the door towards a complete and thorough understanding of the disorder[78,79]. Medicine has always had the aim of uncovering the etiologies of diseases in an attempt to better classify and treat them. Furthermore, at this point in time, what with -omics being a nascent field, we are still in the dark about the majority of the pathways that we are studying[78]. As candidate genes are being discovered, how can we move away from discovering plausibly correlated genes to discovering proof for their causative roles in diseases? Chakravarti, et al. have suggested some general rules that take from Koch's postulates of microbiology, including that (1) candidate genes must be enriched in patients, (2) the mutant phenotype needs to be demonstrated in a model system, (3) wild-type human alleles need to be able to rescue the phenotype, and conversely (4) mutant human alleles must fail to rescue the mutant phenotype[78]. Such rules are incredibly difficult to adhere to in the study of psychiatric genetics and pose significant challenges beyond the basic genetic challenges discussed in the introduction. However, as already discussed, polygenic risk scoring is a promising method for combining the effects of numerous genes into one metric of susceptibility[8].

TDT Alternative

Granted that psychiatric disorders are highly polygenic, and therefore no one or handful of genes is directly responsible for the manifestation of psychopathologies, a very large sample size is required in order to achieve sufficient statistical power to obtain meaningful signal. One of the main challenges in a GWAS is the issue of obtaining a large enough sample size to allow the appearance of this signal for association. This is a struggle that commonly pervades studies of complex disorders. Given that family-based models for association have more sample data

than are actually used in the calculation of chi-squared statistics and p-values, a hybrid variation to the traditional TDT has been developed.

ParenTDT is a variation of the TDT that does take parent phenotypes into account and is in many ways a hybrid family/case-control analysis[80]. Overall, it retains the structure of a family-based analysis, but, in addition to the phenotypes of children, the ParenTDT takes parental phenotypes into account in its formula. This has the potential to increase the power of a study, because the effective sample size of subjects being analyzed is increased. However, since the parent genotypes may have arisen from different ancestral lineages, case alleles may not be well matched to control alleles within families. Therefore, the inclusion of parents as subjects in the analysis allows for the possibility of population stratification and, thus, an increased susceptibility to false-positive findings. Thus, although we did run ParenTDT analyses in addition to TDT, we are not reporting any findings with marginally increased significance, as potentially uncontrolled population stratification puts such findings into question. Future work will involve running principal components analysis on our samples in order to effectively match the case alleles to control alleles and minimize the effects of population stratification[33].

Future Direction

Breaking down broad phenotypes into more specific features is a challenging task. This is primarily due to the massive sample size required in order to discern any specific phenotypes, let alone highly polygenic phenotypes. The CAPS dataset, phenotyped with the DIGS assessment, allows for the detailed identification of interesting symptoms at varying levels of severity and paves the way for new associations to be found among specific sub-phenotypes of Bipolar Disorder. This dataset will ultimately serve as a crucial starting point via which loci of interest will be identified for further analysis. After this component of the broader study is complete, the task will shift toward including case-control data from other studies that have been deposited into the NRGR.

To perform a case-control GWAS on additional samples, I will first have to first remove the potential effects of population stratification. This effect has to be minimized prior to conducting an analysis for association using a principal component analysis; otherwise, the analysis will yield false positives. Additionally, case-control data were genotyped using a variety of genotyping chips from different companies. Thus, the next step would be to determine the intersection set of markers that were used on the different genotyping chips used in different studies or to impute missing markers in some platforms[81].

Afterword

By way of creating more objective measures for the determination of susceptibility for psychopathologies, we hope to increase education, prevention and treatment for disorders of mental health among those who are most at risk for manifesting them. We hope to see a future where prevention starts before the first day outside of the womb: where parents will be made aware of their children's susceptibilities early on and will be better equipped to pave the way for better prevention of onset, understanding, and treatment planning for their children. Ultimately and over time, we hope that a deep understanding of the genetic basis for psychopathologies will also decrease the stigma that Bipolar Disorder and other psychiatric disorders carry, so that those stuck in the grasp of such disorders will be treated with more compassion.

APPENDIX

Tables

Individuals Per Pedigree	Num. Individuals	Pedigrees	Working Families
3	309	103	134
4	404	101	108
5	70	14	14
6	53	9	0
7	47	7	0
8	0	0	0
9	16	2	0
TOTAL	899	236	256

Table 1. BD1-MIX subset counts, broken down by the number of individuals per pedigree (first column). Each row contains the number of individuals present for a pedigree of a particular size, followed by the number of pedigrees within the subset with the corresponding number of individuals. Lastly the final column adjusts the number of pedigrees by accounting for all nuclear families, which causes the higher-order pedigrees to be broken down into the multiple smaller pedigrees of which they are comprised.

Individuals Per Pedigree	Num. Individuals	Pedigrees	Working Families
3	264	88	103
4	124	31	34
5	5	1	1
6	35	6	0
7	21	3	0
TOTAL	449	129	138

Table 2. BD1-NP subset counts, broken down by the number of individuals per pedigree (first column). Each row contains the number of individuals present for a pedigree of a particular size, followed by the number of pedigrees within the subset with the corresponding number of individuals. Lastly the final column adjusts the number of pedigrees by accounting for all nuclear families, which causes the higher-order pedigrees to be broken down into the multiple smaller pedigrees of which they are comprised.

Individuals Per Pedigree	Num. Individuals	Pedigrees	Working Families
3	297	99	110
4	176	44	45
5	30	6	6
6	28	5	0
7	6	1	0
TOTAL	537	155	161

Table 3. BD1-P subset counts, broken down by the number of individuals per pedigree (first column). Each row contains the number of individuals present for a pedigree of a particular size, followed by the number of pedigrees within the subset with the corresponding number of individuals. Lastly the final column adjusts the number of pedigrees by accounting for all nuclear families, which causes the higher-order pedigrees to be broken down into the multiple smaller pedigrees of which they are comprised.

CHR	BP	RSID	SNP	OR	CHISQ	P
23	19781333	rs12010076	AX-42738377	0.01639	58.06	2.54E-14
8	59164297	rs7003372	AX-16002599	1.696	24.27	8.35E-07
14	106321212	rs281865422	AX-168647780	0.03846	23.15	1.50E-06
18	29101207	rs553299589	AX-168641298	0	23.00	1.62E-06
9	20922554	rs117655852	AX-36953059	0.04167	21.16	4.23E-06
14	83058722	rs75565666	AX-12838593	0.2708	20.08	7.42E-06
9	78902186	rs76757914	AX-37047267	0	20.00	7.74E-06
14	56928268	rs58793557	AX-12792054	0	20.00	7.74E-06
6	158329778	rs3840366	AX-151210704	0.5957	19.25	1.15E-05
6	158340497	rs3047738	AX-120518563	0.5946	19.07	1.26E-05

Table 4. Top 10 most significant associations found in the BD1-MIX dataset.

CHR	BP	RSID	SNP	OR	CHISQ	P
23	19781333	rs12010076	AX-42738377	0	30.00	4.32E-08
6	158349893	rs2273070	AX-15298641	0.4149	22.74	1.85E-06
6	158329778	rs3840366	AX-151210704	0.4194	22.09	2.60E-06
6	158317132	rs6924813	AX-15298546	0.4194	22.09	2.60E-06
6	158340497	rs3047738	AX-120518563	0.4222	21.12	4.30E-06

Table 5. Top 5 most significant associations found in the BD1-NP subset.

CHR	BP	RSID	SNP	OR	CHISQ	P
23	19781333	rs12010076	AX-42738377	0.03226	28.12	1.14E-07
13	41995485	rs9566764	AX-11695142	0.5241	21.54	3.46E-06
7	28764109	rs2066979	AX-11364907	1.87	20.31	6.58E-06
1	66617553	rs72679119	AX-13088763	2.556	18.38	1.81E-05
9	138058487	rs12379748	AX-42567171	2.083	18.27	1.92E-05

Table 6. Top 10 most significant associations found in the BD1-P subset.

Figures

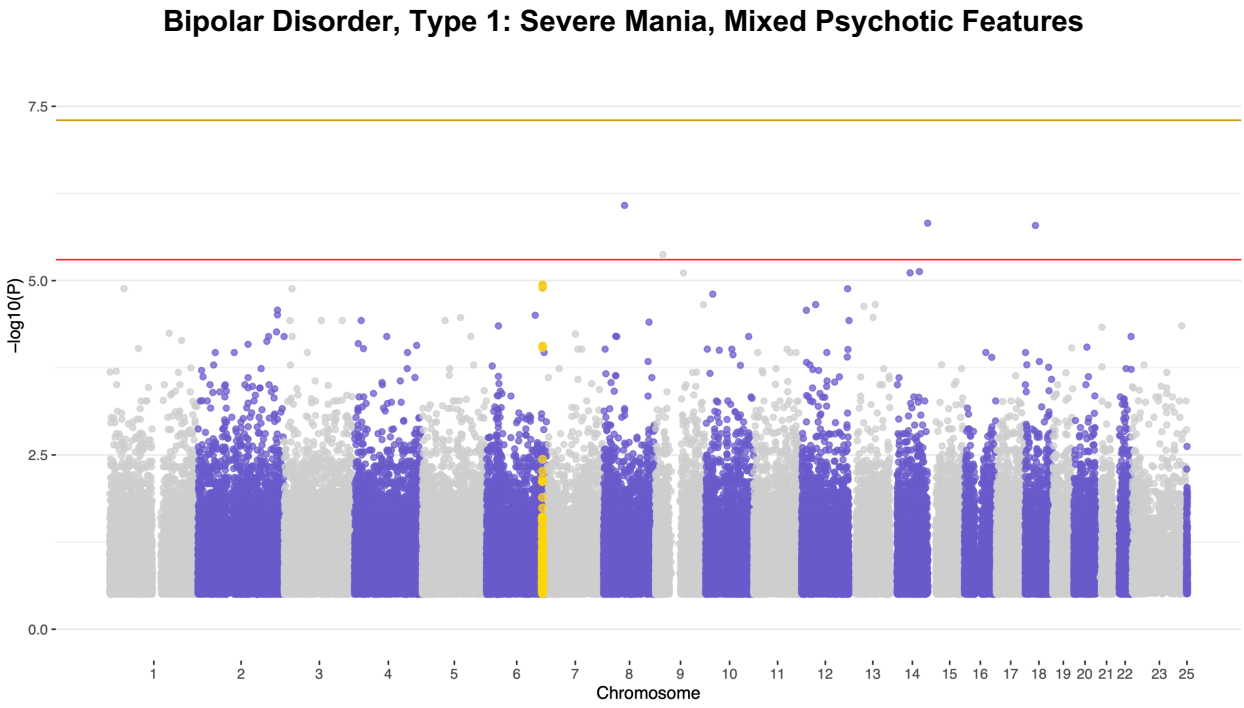

Figure 1a. Manhattan plot of mixed BD1-NP and BD1-P datasets. The total sample size includes 899 individuals (391 affected children, 508 parents). A total of 582,349 variants are plotted as data points. Along the Y-axis, the data points represent the negative logarithm of *p*-values for association between any particular SNP and the phenotype of interest. The x-axis represents chromosomal loci; colors alternate between grey and blue to differentiate the chromosomes. The red and gold horizontal lines represent lenient (5.0E-6) and stringent (5.0E-8) thresholds for significance, although confidence in findings requires clustering of closely positioned SNPs above a threshold line. Overall, no significant findings can be discerned from this heterogeneous group of individuals who experience BD, some with psychotic features and some without. *P*-values highlighted in gold represent SNPs that do indicate significance among the BD1-NP subset. Despite the BD1-NP subset having a smaller sample size, five SNPs do cross the lenient threshold in that subset and thus indicate significance specifically for the BD1-NP subset. Note that the lead SNP on chromosome X (23) is not visible on this figure.

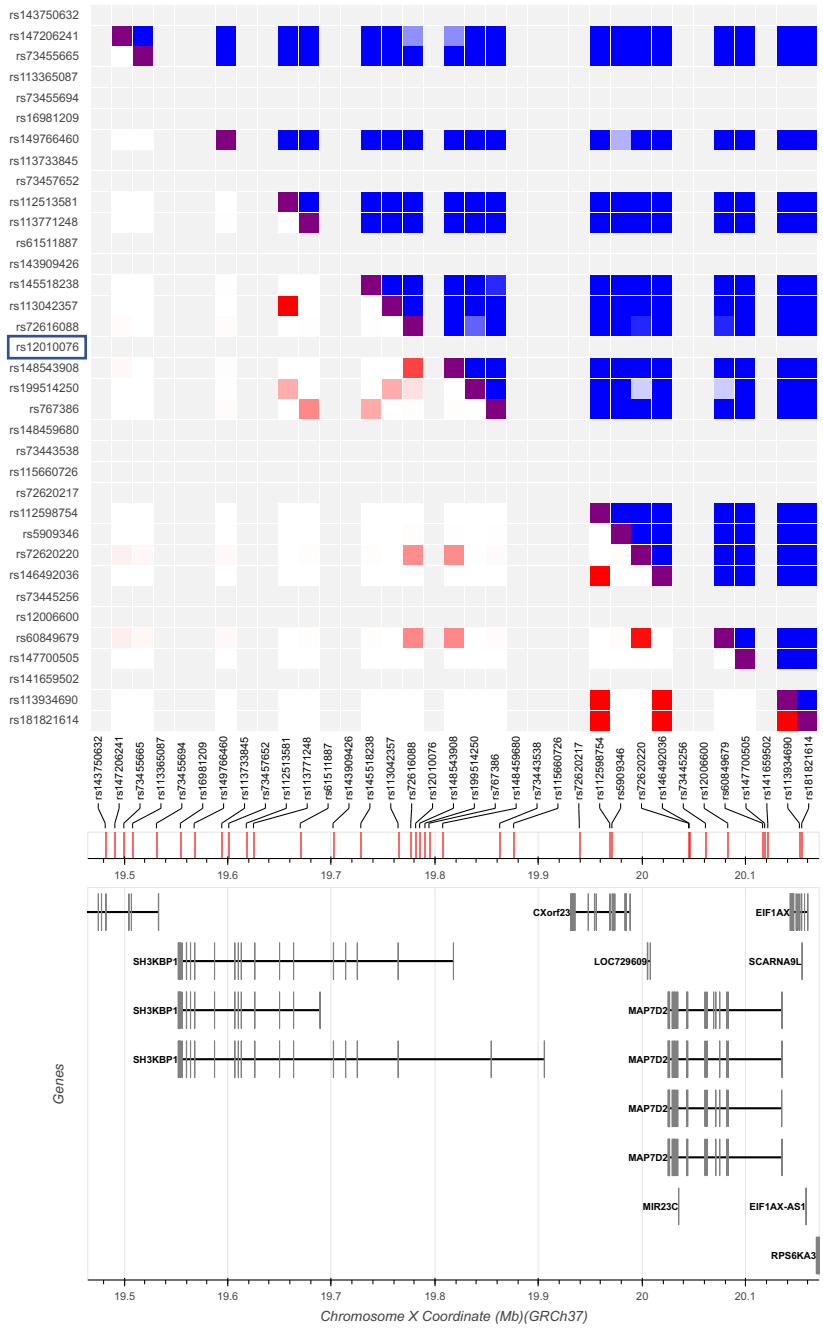

Figure 1b. R^2 and D' coefficients calculated and displayed in a matrix for SNPs including and surrounding BD1-MIX lead SNP, rs12010076. Darker red indicates greater r^2 values, while darker blue indicates greater D' values. This heatmap shows no LD between rs12010076 and the SNPs in its vicinity. Created using LDmatrix[65].

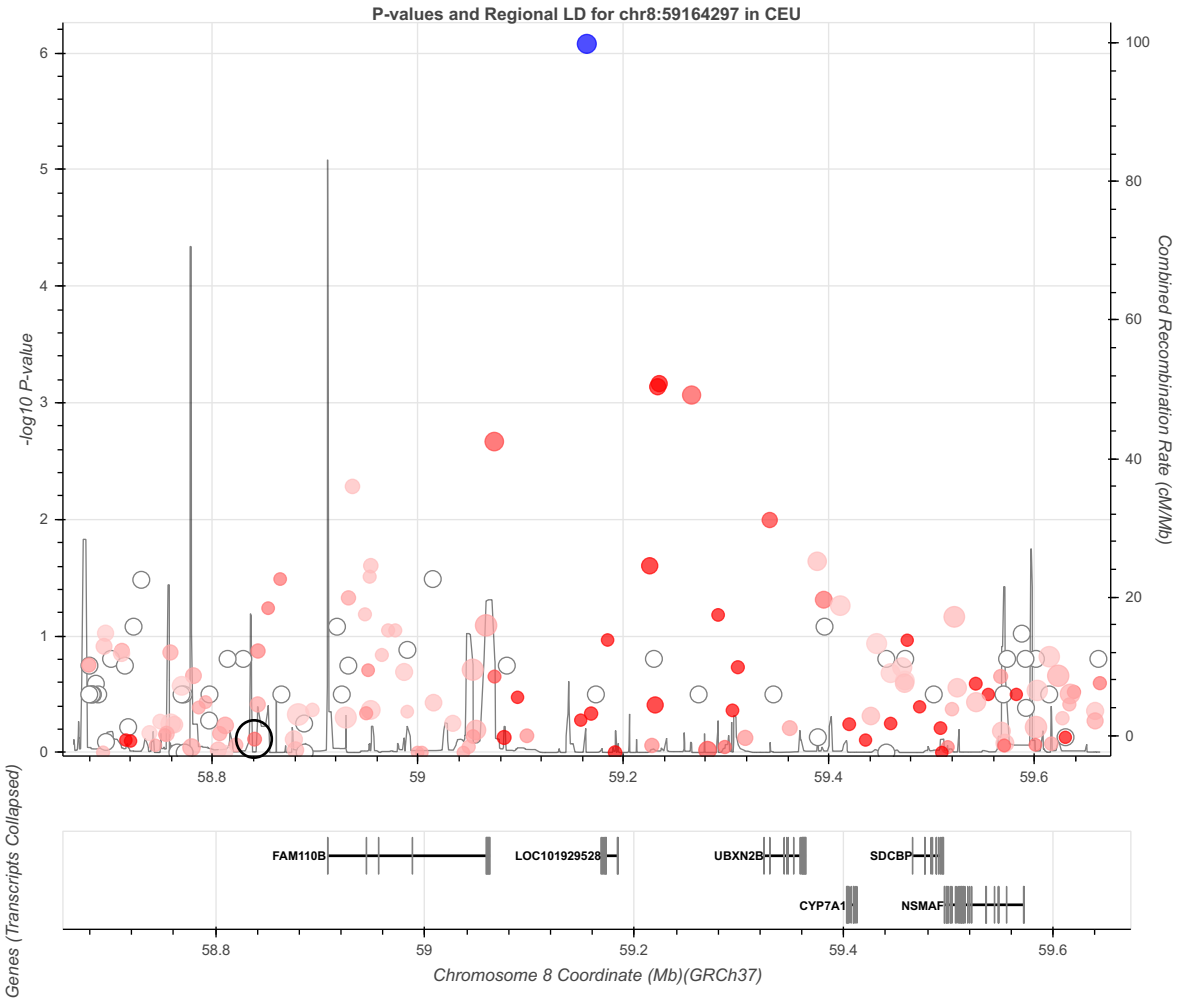

Figure 1c. Association plot of chr8:58664297-59664297. BD1-MIX SNP rs7003372 shown as a blue circle; literature-reported SNP rs1992045 circled in black. Darker red indicates stronger LD (D') between SNP and rs7003372. Smaller circle circumference represents smaller MAF; larger circumference represents larger MAF. Lack of significance at rs1992045 and in between the two SNPs of interest is peculiar and requires further analysis. Created using LDassoc[65].

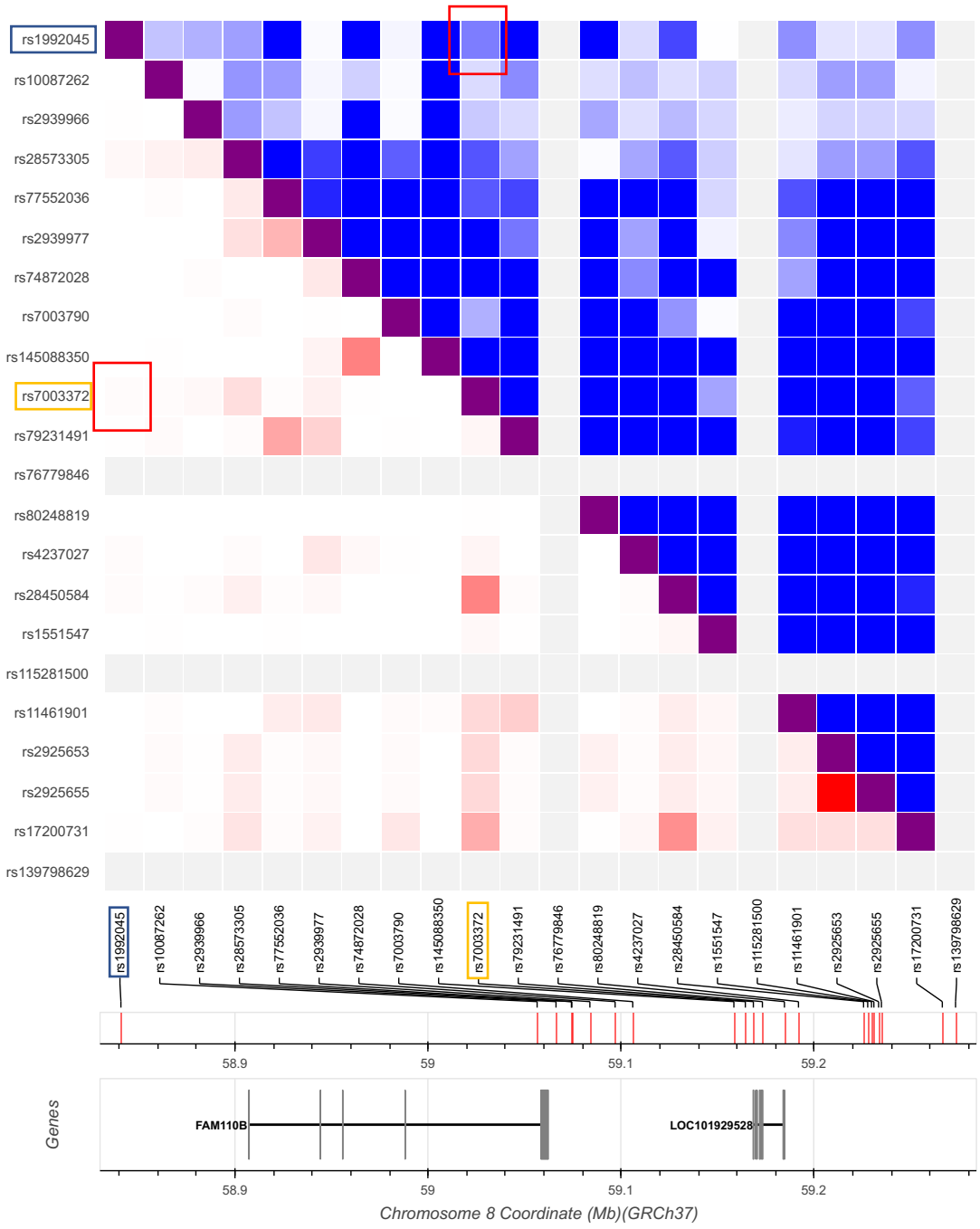

Figure 1d. R^2 and D' coefficients calculated and displayed in a matrix for SNPs including and surrounding BD1-MIX SNP, rs7003372 (boxed in orange along x-axis), and BD associated SNP reported in the literature, rs1992045 (boxed in blue along y-axis). Darker red indicates greater r^2 values and darker blue indicates greater D' values. D'=0.515 (top, middle) and an R^2=0.014 (left, middle). Created using LDmatrix[65].

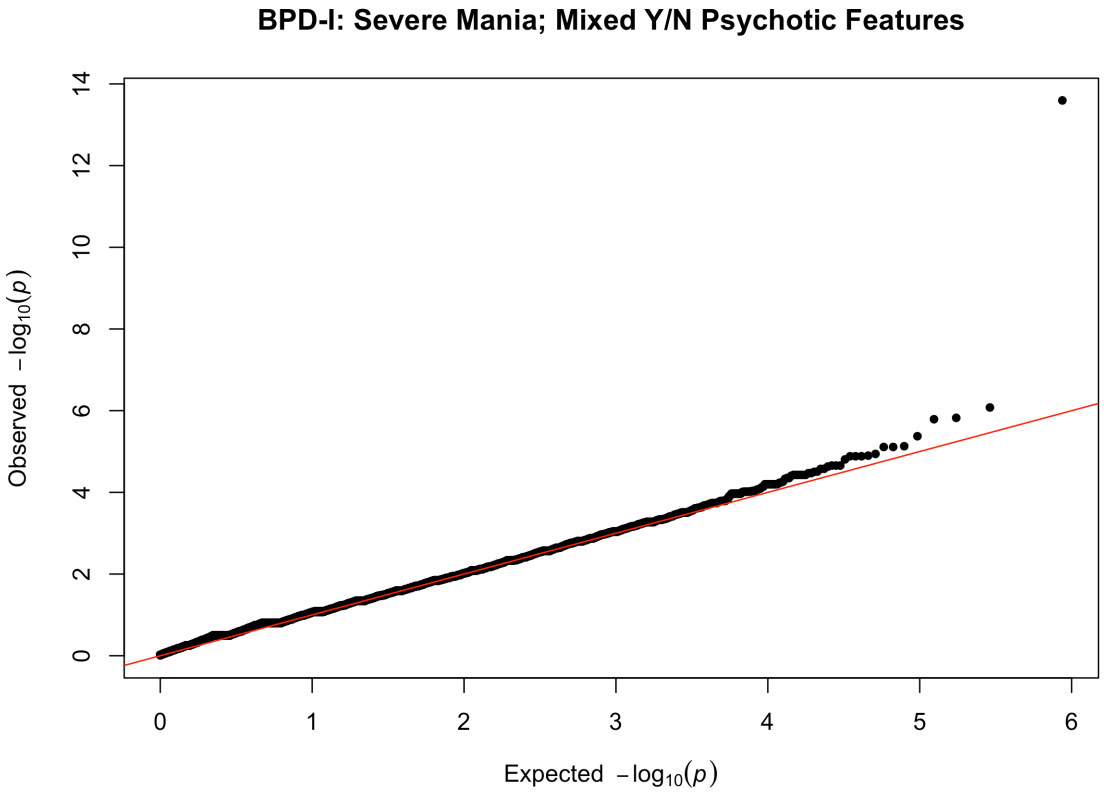

Figure 1e. Q-Q (Quantile-Quantile) plots for the BD1-MIX dataset. Theoretical quantiles following a Gaussian distribution are plotted on the x-axis, whereas quantiles from each of my dataset outputs are plotted on the y-axis. Within $x \in [0,6]$ and $y \in [0,6]$, the overall linearity reflects that association results follow a normal distribution, as expected. However, the slight jaggedness of the nearly linear points suggests either batch effects or truncation of values at some point in the analysis.

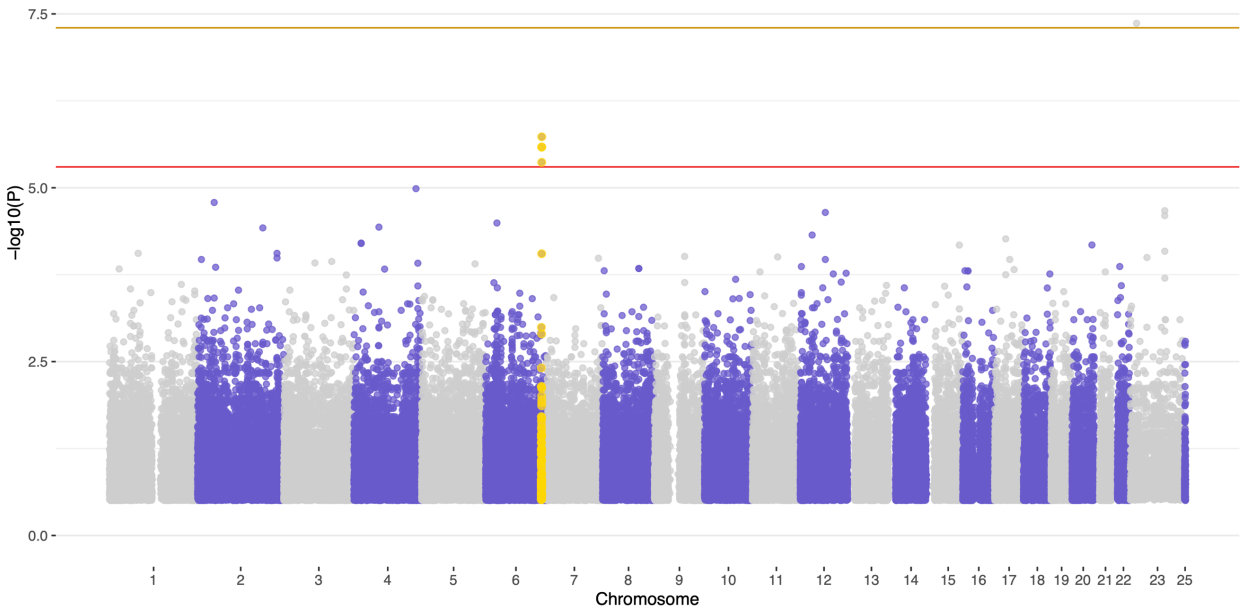

Figure 2a. Manhattan plot of the BD1-NP dataset. The total sample size includes 449 individuals (173 affected children, 276 parents). A total of 582,349 variants are plotted as data points. Along the Y-axis, the data points represent the negative logarithm of *p*-values for association between any particular SNP and the phenotype of interest. The x-axis represents chromosomal loci; colors alternate between grey and blue to differentiate the chromosomes. The red and gold horizontal lines represent lenient (5.0E-6) and stringent (5.0E-8) thresholds for significance, although confidence in findings requires clustering of closely positioned SNPs above a threshold line. Despite the BD1-NP subset having a smaller sample size, a cluster of four intron variants (in gold), rs2273070, rs3840366, rs6924813, and rs3047738 on 6q25.3 indicate early signs of association with non-psychotic BD.

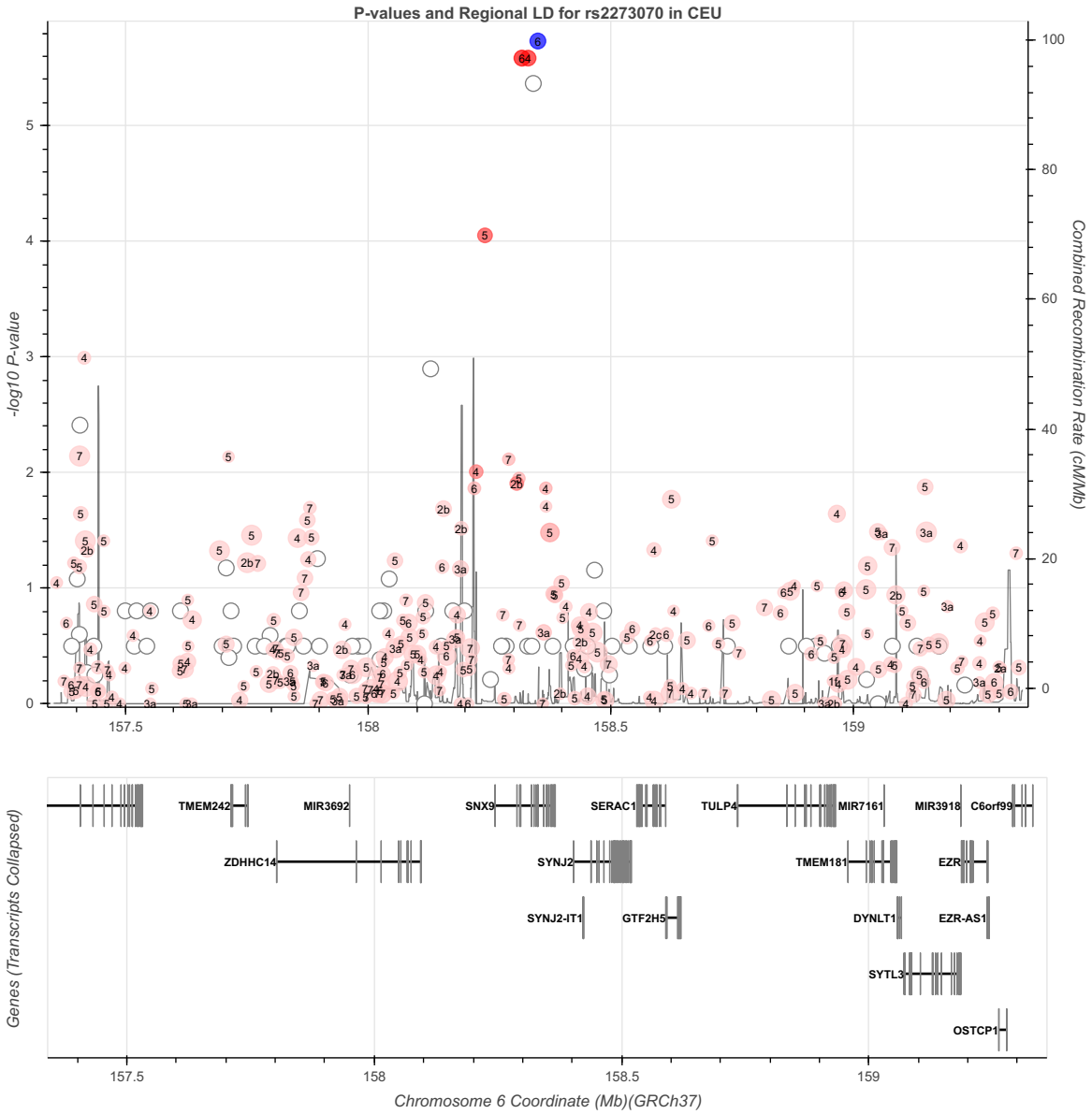

Figure 2b. P-value and Regional LD plot. Close-up view of 6q25.3 SNPs in BD1-NP subset. X-axis displays position of most significant SNP ±500,000 BP. Y-axis displays both the -log of the P-value and the combined recombination rate (cM/Mb). The genes that appear throughout this region are displayed at the bottom of the figure, with SNX9 positioned directly below the lead SNPs. Created using LDassoc.

RS Number	Position (GRCh37)	Allele Frequencies	Haplotypes		
rs6924813	chr6:158317132	T=0.748, C=0.253	T	C	C
rs3840366	chr6:158329778	-=0.758, AC=0.242	-	AC	-
rs3047738	chr6:158340496	CTG=0.758, -=0.242	CTG	-	CTG
rs2273070	chr6:158349893	T=0.748, C=0.253	T	C	C
		Haplotype Count	**147**	**47**	**3**
		Haplotype Frequency	**0.7424**	**0.2374**	**0.0152**

Figure 2c. Haplotypes for the four lead SNPs on chromosome 6 in the BD1-NP subset. These haplotypes were calculated based on the CEU (European) population, which most accurately reflects the ancestry of the dataset used in this analysis, sub-setted from the CAPS-BP dataset. Created using LDlink.

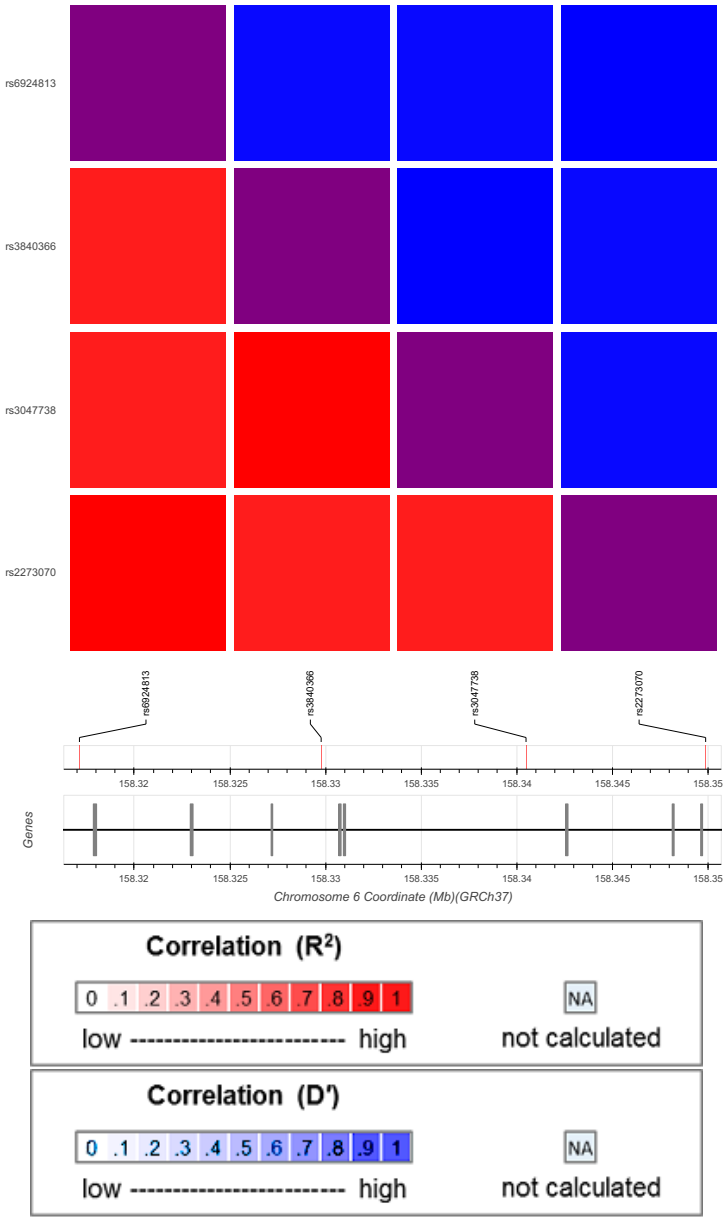

Figure 2d. R^2 and D' coefficients calculated and displayed in a matrix for the four lead SNPs on chromosome 6 in the BD1-NP cohort shows strong linkage disequilibrium among these SNPs. Created using LDmatrix.

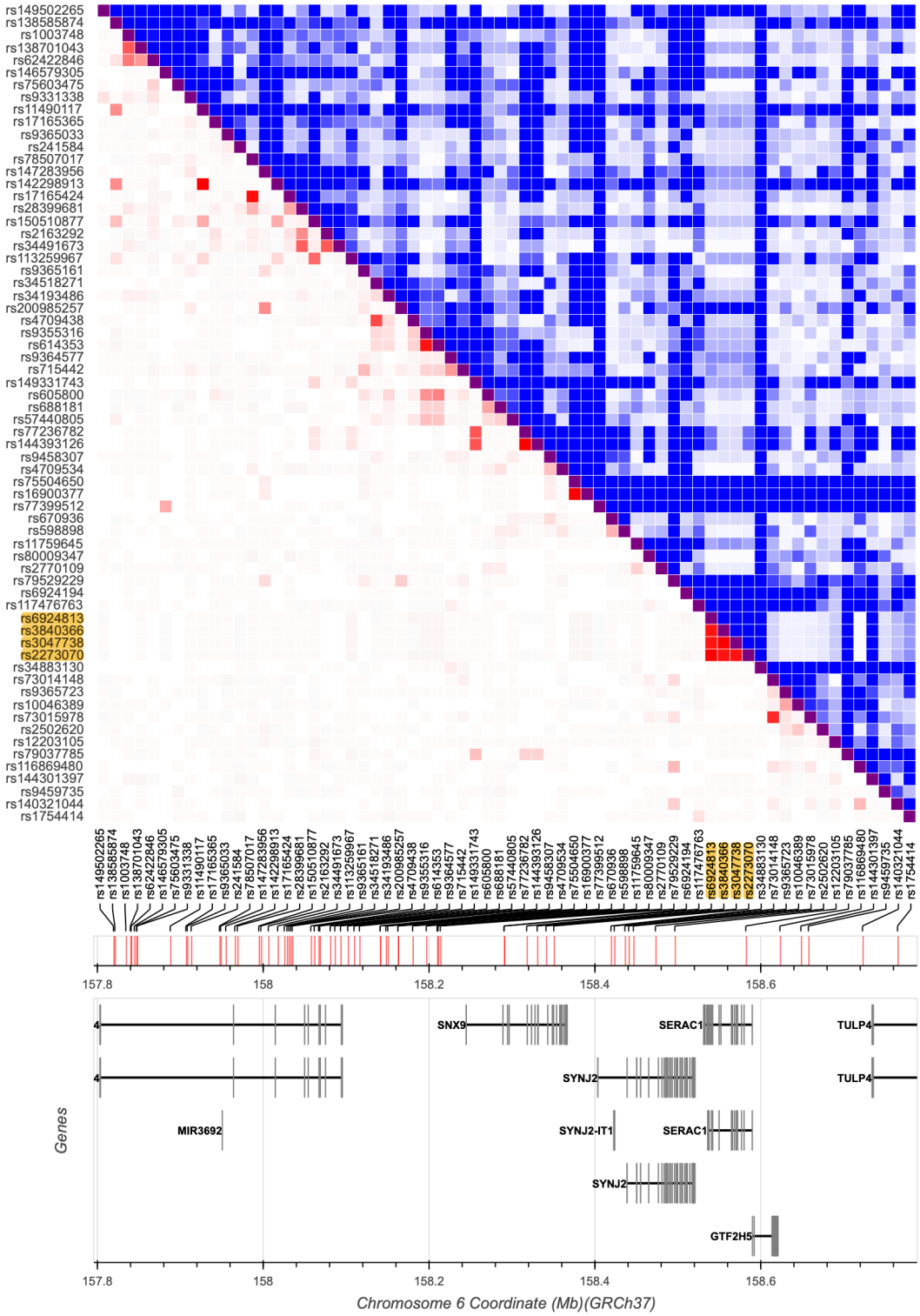

Figure 2e. R^2 and D' coefficients calculated and displayed in a matrix for chr6 SNPs in the BD1-NP cohort. Strong LD among lead SNPs in the context of neighboring SNPs. Created using LDmatrix.

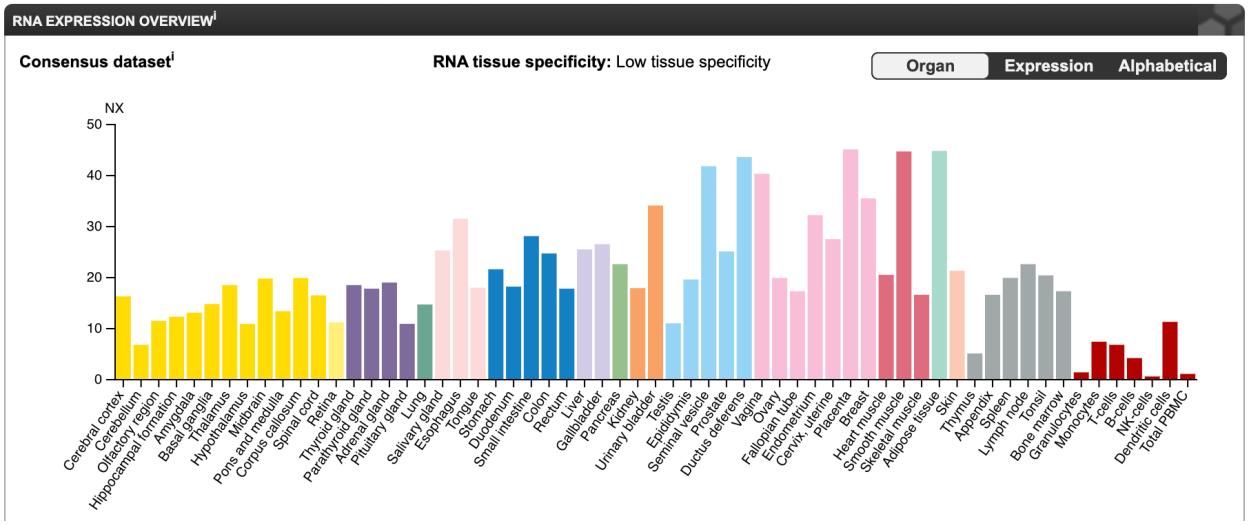

Figure 2f. SNX9 RNA expression overview in various tissues. SNX9 shows low tissue specificity and is expressed throughout the central nervous system (yellow). Data gathered from the Human Protein Atlas.

Epigenome ID (EID)	Group	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase
E093	Thymus	THYM.FET	Fetal Thymus							
E071	Brain	BRN.HIPP.MID	Brain Hippocampus Middle	6_EnhG	11_TxEnh3	H3K4me1_Enh		H3K27ac_Enh		
E074	Brain	BRN.SUB.NIG	Brain Substantia Nigra	6_EnhG	11_TxEnh3	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	
E068	Brain	BRN.ANT.CAUD	Brain Anterior Caudate	6_EnhG	11_TxEnh3	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro	
E069	Brain	BRN.CING.GYR	Brain Cingulate Gyrus	6_EnhG	11_TxEnh3	H3K4me1_Enh	H3K4me3_Pro	H3K27ac_Enh	H3K9ac_Pro	
E072	Brain	BRN.INF.TMP	Brain Inferior Temporal Lobe	6_EnhG	11_TxEnh3	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	
E067	Brain	BRN.ANG.GYR	Brain Angular Gyrus	6_EnhG	11_TxEnh3	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	
E073	Brain	BRN.DL.PRFRNTL.CRTX	Brain_Dorsolateral_Prefrontal_Cortex	6_EnhG	11_TxEnh3	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	
E070	Brain	BRN.GRM.MTRX	Brain Germinal Matrix		11_TxEnh3					
E082	Brain	BRN.FET.F	Fetal Brain Female		11_TxEnh3					
E081	Brain	BRN.FET.M	Fetal Brain Male							
E063	Adipose	FAT.ADIP.NUC	Adipose Nuclei							
E100	Muscle	MUS.PSOAS	Psoas Muscle							
E108	Muscle	MUS.SKLT.F	Skeletal Muscle Female							
E107	Muscle	MUS.SKLT.M	Skeletal Muscle Male							
E089	Muscle	MUS.TRNK.FET	Fetal Muscle Trunk							
E090	Muscle	MUS.LEG.FET	Fetal Muscle Leg					H3K27ac_Enh		
E083	Heart	HRT.FET	Fetal Heart			H3K4me1_Enh				
E104	Heart	HRT.ATR.R	Right Atrium							
E095	Heart	HRT.VENT.L	Left Ventricle							
E105	Heart	HRT.VNT.R	Right Ventricle							
E065	Heart	VAS.AOR	Aorta							
E078	Sm. Muscle	GI.DUO.SM.MUS	Duodenum Smooth Muscle							
E076	Sm. Muscle	GI.CLN.SM.MUS	Colon Smooth Muscle							
E103	Sm. Muscle	GI.RECT.SM.MUS	Rectal Smooth Muscle							
E111	Sm. Muscle	GI.STMC.MUS	Stomach Smooth Muscle							
E092	Digestive	GI.STMC.FET	Fetal Stomach							
E085	Digestive	GI.S.INT.FET	Fetal Intestine Small							
E084	Digestive	GI.L.INT.FET	Fetal Intestine Large							
E109	Digestive	GI.S.INT	Small Intestine							
E106	Digestive	GI.CLN.SIG	Sigmoid Colon							
E075	Digestive	GI.CLN.MUC	Colonic Mucosa							
E101	Digestive	GI.RECT.MUC.29	Rectal Mucosa Donor 29					H3K27ac_Enh		
E102	Digestive	GI.RECT.MUC.31	Rectal Mucosa Donor 31							
E110	Digestive	GI.STMC.MUC	Stomach Mucosa							
E077	Digestive	GI.DUO.MUC	Duodenum Mucosa							
E079	Digestive	GI.ESO	Esophagus							
E094	Digestive	GI.STMC.GAST	Gastric			H3K4me1_Enh				
E099	Other	PLCNT.AMN	Placenta Amnion							
E086	Other	KID.FET	Fetal Kidney							
E088	Other	LNG.FET	Fetal Lung							
E097	Other	OVRY	Ovary							
E087	Other	PANC.ISLT	Pancreatic Islets							
E080	Other	ADRL.GLND.FET	Fetal Adrenal Gland			H3K4me1_Enh				
E091	Other	PLCNT.FET	Placenta							
E066	Other	LIV.ADLT	Liver							
E098	Other	PANC	Pancreas							
E096	Other	LNG	Lung			H3K4me1_Enh				
E113	Other	SPLN	Spleen							
E114	ENCODE2012	LNG.A549.ETOH002.CNCR	A549 EtOH 0.02pct Lung Carcinoma Cell Line							
E115	ENCODE2012	BLD.DND41.CNCR	Dnd41 TCell Leukemia Cell Line							

Figure 2g. SNX9 regulatory chromatin states from histone ChIP-Seq shows epigenetic states in brain tissues that are specifically not present in other tissues throughout the body. Data gathered from the Roadmap Epigenomics Consortium, 2015.

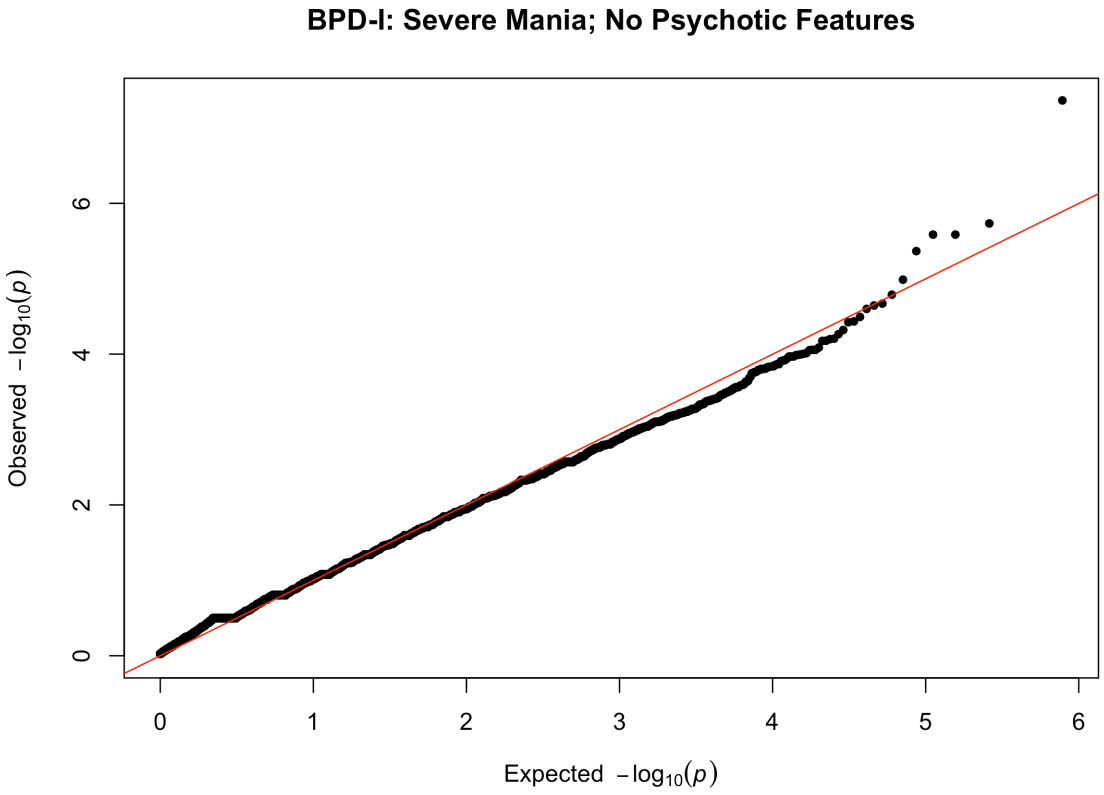

Figure 2h. Q-Q (Quantile-Quantile) plots for the BD1-NP subset. Theoretical quantiles following a Gaussian distribution are plotted on the x-axis, whereas quantiles from each of my dataset outputs are plotted on the y-axis. Within $x \in [0,6]$ and $y \in [0,6]$, the overall linearity reflects that association results follow a normal distribution, as expected. However, the slight jaggedness of the nearly linear points suggests either batch effects or truncation of values at some point in the analysis.

Bipolar Disorder, Type 1: Severe Mania, With Psychotic Features

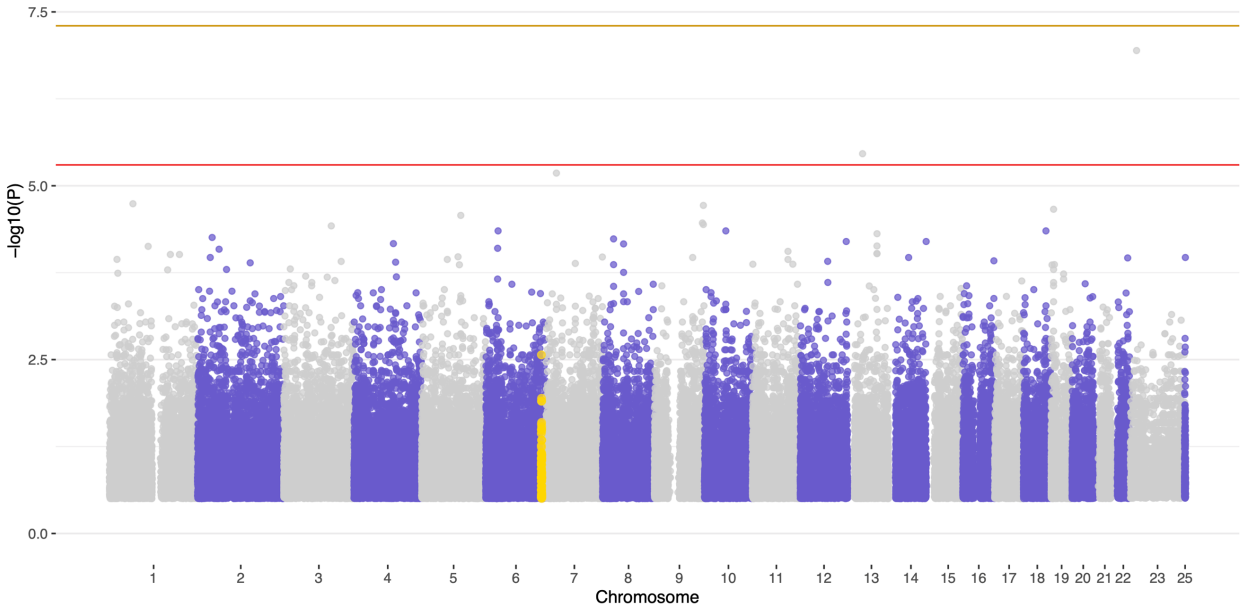

Figure 3a. Manhattan plot of the BD1-P dataset. The total sample size includes 537 individuals (218 affected children, 319 are parents). A total of 582,349 variants are plotted as data points. Along the Y-axis, the data points represent the negative logarithm of *p*-values for association between any particular SNP and the phenotype of interest. The x-axis represents chromosomal loci; colors alternate between grey and blue to differentiate the chromosomes. The red and gold horizontal lines represent lenient (5.0E-6) and stringent (5.0E-8) thresholds for significance, although confidence in findings requires clustering of closely positioned SNPs above a threshold line. Overall, no significant findings can be discerned. However, it is interesting that while SNPs along 6q25.3 (gold) were indicative of association in the BD1-NP dataset, these same SNPs show no association whatsoever with BD1-P. This indicates that a gene within 6q25.3 may be associated with pure manic-depressive BD, but not psychotic features. As such, the introduction of psychotic features in this dataset may have created sufficient noise to eliminate any signal of association.

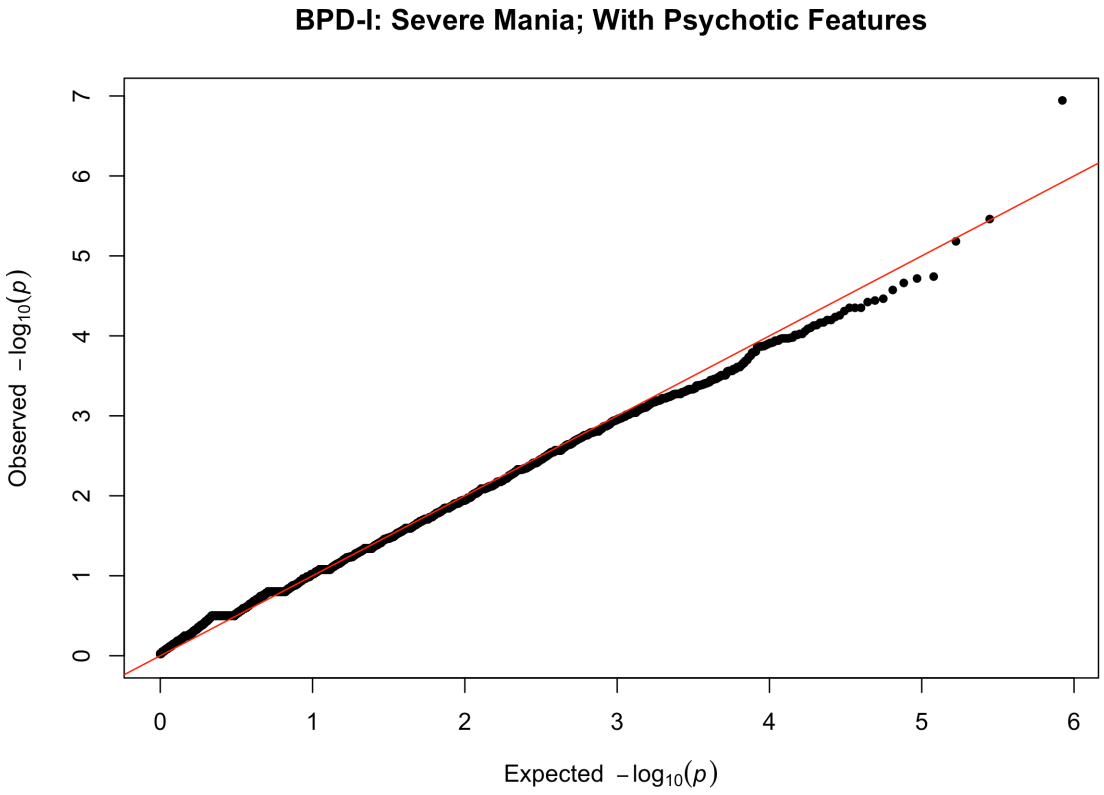

Figure 3b. Q-Q (Quantile-Quantile) plots for the BD1-P subset. Theoretical quantiles following a Gaussian distribution are plotted on the x-axis, whereas quantiles from each of my dataset outputs are plotted on the y-axis. Within $x \in [0,6]$ and $y \in [0,6]$, the overall linearity reflects that association results follow a normal distribution, as expected. However, the slight jaggedness of the nearly linear points suggests either batch effects or truncation of values at some point in the analysis.

Glossary

Comma Separated Values (CSV): A common plain text file-type whereby columns are separated by commas.

Diagnostic and Statistical Manual (DSM): The taxonomic and diagnostic tool for psychiatric disorders, published by the American Psychiatric Association.

Genome-wide association study (GWAS): A technique in genetics used to establish associations between diseases and specific loci or genes.

JavaScript Object Notation (JSON): A lightweight interchange formatted file that serves to maximize storage and computational efficiency (see NID).

Linkage Disequilibrium (LD): The tendency for genes and other genetic markers to be inherited together because of their location near one another on the same chromosome.

Linkage studies: Studies aimed at establishing linkage between genes.

Locus (plural: loci): The position of a gene, marker, or mutation on a chromosome.

Manhattan plot: A type of scatterplot used to display data with a large number of data points. Because the strongest associations have the smallest *p*-values (e.g. 10^{-8}), their negative logarithms will be the greatest (e.g. 8) when plotted on the Y-axis.

Minor Allele Frequency (MAF): the frequency at which the second most common allele appears in a given population. Provides useful information for distinguishing between rare and common variants in a population.

NIMH Repository and Genomics Resource (NRGR): A repository in which decades of phenotypic psychiatric and genetic data are stored for use by grant-funded research studies.

NRGR Identifier Dictionary (NID): A lookup table stored in the form of a highly computationally efficient programming data-structure (see JSON).

Nosology: The branch of medical science dealing with the classification of diseases.

Pleiotropy: The effect whereby one gene has several seemingly unrelated downstream effects and/or products.

Polygenic risk scores (PRS): A metric that builds off of GWAS data and adds the weighted effects of all of the identified genetic variants to produce a score that is predictive of latent risk for many phenotypes of interest.

Population stratification: The phenomenon whereby differences in individuals' genomes strictly due to their ethnic backgrounds appear to stratify multi-ethnic genetic data sets. This effect has to be minimized prior to conducting an analysis for association.

Principal component analysis (PCA): A statistical approach that reduces extremely multidimensional data into a few principal components.

Q-Q (quantile-quantile) plot: A graphical method for comparing two probability distributions, often times one being a theoretical standard normal distribution.

Single nucleotide polymorphism (SNP): A single nucleotide position that is present in populations in many variant forms. SNPs tend to be inherited along with adjacent genes of interest.

Symptomatology: The set of symptoms characteristic of a medical condition or exhibited by a patient.

Type-I error: Also known as a false positive. Occurs when a researcher incorrectly rejects a true null hypothesis. It is important to use stringent P-values in association studies, where false positives are commonplace.

Protocols & Sources

Data Collection

DIVER is a private software developed by William Valentine-Cooper at the University of Ohio.

Bill Valentine-Cooper, Senior Systems Developer

Battelle Center for Mathematical Medicine

Abigail Wexner Research Institute

Nationwide Children's Hospital

Studies Utilized for Analysis

Study BP 0

Data and biomaterials were collected in four projects that participated in the National Institute of Mental Health (NIMH) Bipolar Disorder Genetics Initiative. From 1991-98, the Principal Investigators and Co-Investigators were: Indiana University, Indianapolis, IN, U01 MH46282, John Nurnberger, M.D., Ph.D., Marvin Miller, M.D., and Elizabeth Bowman, M.D.; Washington University, St. Louis, MO, U01 MH46280, Theodore Reich, M.D., Allison Goate, Ph.D., and John Rice, Ph.D.; Johns Hopkins University, Baltimore, MD, U01 MH46274, J. Raymond DePaulo, Jr., M.D., Sylvia Simpson, M.D., MPH, and Colin Stine, Ph.D.; NIMH Intramural Research Program, Clinical Neurogenetics Branch, Bethesda, MD, Elliot Gershon, M.D., Diane Kazuba, B.A., and Elizabeth Maxwell, M.S.W.

Study 1

Data and biomaterials were collected as part of ten projects that participated in the National Institute of Mental Health (NIMH) Bipolar Disorder Genetics Initiative. From 1999-03, the Principal Investigators and Co-Investigators were: Indiana University, Indianapolis, IN, R01 MH59545, John Nurnberger, M.D., Ph.D., Marvin J. Miller, M.D., Elizabeth S. Bowman, M.D., N.

Leela Rau, M.D., P. Ryan Moe, M.D., Nalini Samavedy, M.D., Rif El-Mallakh, M.D. (at University of Louisville), Husseini Manji, M.D. (at Wayne State University), Debra A. Glitz, M.D. (at Wayne State University), Eric T. Meyer, M.S., Carrie Smiley, R.N., Tatiana Foroud, Ph.D., Leah Flury, M.S., Danielle M. Dick, Ph.D., Howard Edenberg, Ph.D.; Washington University, St. Louis, MO, R01 MH059534, John Rice, Ph.D, Theodore Reich, M.D., Allison Goate, Ph.D., Laura Bierut, M.D.; Johns Hopkins University, Baltimore, MD, R01 MH59533, Melvin McInnis, M.D., J. Raymond DePaulo, Jr., M.D., Dean F. MacKinnon, M.D., Francis M. Mondimore, M.D., James B. Potash, M.D., Peter P. Zandi, Ph.D, Dimitrios Avramopoulos, and Jennifer Payne; University of Pennsylvania, PA, R01 MH59553, Wade Berrettini, M.D., Ph.D.; University of California at Irvine, CA, R01 MH60068, William Byerley, M.D., and Mark Vawter, M.D.; University of Iowa, IA, R01 MH059548, William Coryell, M.D., and Raymond Crowe, M.D.; University of Chicago, IL, R01 MH59535, Elliot Gershon, M.D., Judith Badner, Ph.D., Francis McMahon, M.D., Chunyu Liu, Ph.D., Alan Sanders, M.D., Maria Caserta, Steven Dinwiddie, M.D., Tu Nguyen, Donna Harakal; University of California at San Diego, CA, R01 MH59567, John Kelsoe, M.D., Rebecca McKinney, B.A.; Rush University, IL, R01 MH059556, William Scheftner, M.D., Howard M. Kravitz, D.O., M.P.H., Diana Marta, B.S., Annette Vaughn-Brown, M.S.N., R.N., and Laurie Bederow, M.A.; NIMH Intramural Research Program, Bethesda, MD, 1Z01MH002810-01, Francis J. McMahon, M.D., Layla Kassem, PsyD, Sevilla Detera-Wadleigh, Ph.D, Lisa Austin, Ph.D, Dennis L. Murphy, M.D.

Study 2

Data and biomaterials were collected and supported by NIMH grant R01 MH59602 (to Miron Baron, M.D.) and by funds from the Columbia Genome Center and the New York State Office of Mental Health. The main contributors to this work were Miron Baron, M.D. (Principal Investigator), Jean Endicott, Ph.D. (Co-Principal Investigator), Jo Ellen Loth, M.S.W., John Nee, Ph.D, Richard Blumenthal, Ph.D., Lawrence Sharpe, M.D., Barbara Lilliston, M.S.W., Melissa

Smith, M.A., and Kristine Trautman, M.S.W., all from Columbia University Department of Psychiatry, New York, NY, USA. A small subset of the sample was collected in Israel in collaboration with Bernard Lerer, M.D. and Kyra Kanyas, M.S. from the Hadassah - Hebrew University Medical Center, Jerusalem, Israel. We are grateful to the patients and their family members for their cooperation and support, and to the treatment facilities and other organizations that collaborated with us in identifying families.

Study 49

Data and biomaterials used in this research report were collected by the International Neuro-Genetics Association of Spanish America and the United States (INGASU), and funded by NIMH grant MH69856 (Genetics of Bipolar Disorder in Latino Populations) to principal investigator Dr. Michael Escamilla (Paul L. Foster School of Medicine, Texas Tech University Health Science Center, El Paso, Texas). Additional principal investigators who participated in this grant were Dr. Alvaro Jerez (Centro Internacional de Trastornos Afectivos y de la Conducta Adictiva-CITACA, Guatemala), Dr. Ricardo Mendoza (University of California at Los Angeles-Harbor), Dr. Humberto Nicolini (Medical and Family Research Group, Carracci S.C., Mexico City, Mexico), Dr. Henriette Raventos (University of Costa Rica, San Jose, Costa Rica), and Dr. Alfonso Ontiveros (Instituto de Informacion de Investigacion en Salud Mental, Monterrey, Mexico). In addition to Drs. Escamilla and Nicolini, the following contributed to the diagnostic best estimation process: Drs. Salvador Contreras, Albana Dassori and Rolando Medina (University of Texas Health Science Center at San Antonio), Dr. Regina Armas (University of California at San Francisco), Dr. Javier Contreras (University of Costa Rica), and Drs. Mercedes Ramirez and Juan Zavala (Paul L. Foster School of Medicine, Texas Tech University Health Science Center).

WORKS CITED

1. Knapp, M. & Wong, G. Economics and Mental Health: The Current Scenario. *World Psychiatry* **19**, 3-14 (2020).
2. Merikangas, K.R. *et al.* Lifetime Prevalence of Mental Disorders in U.S. Adolescents: Results From the National Comorbidity Survey Replication--Adolescent Supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry* **49**, 980-989 (2010).
3. Mental Illness. *National Institute of Mental Health* (2020).
4. Bruffaerts, R. *et al.* Lifetime and 12-Month Treatment for Mental Disorders and Suicidal Thoughts and Behaviors Among First Year College Students. *Int J Methods Psychiatr Res* **28**, e1764 (2019).
5. Shackman, A.J. & Fox, A.S. Getting Serious About Variation: Lessons for Clinical Neuroscience (A Commentary on 'The Myth of Optimality in Clinical Neuroscience'). *Trends Cogn Sci* **22**, 368-369 (2018).
6. Conway, C.C. *et al.* A Hierarchical Taxonomy of Psychopathology Can Transform Mental Health Research. *Perspect Psychol Sci* **14**, 419-436 (2019).
7. Redish, A.G., J. *Computational Psychiatry: New Perspectives on Mental Illness*, 424 (MIT Press, 2016).
8. Anderson, J.S., Shade, J., DiBlasi, E., Shabalin, A.A. & Docherty, A.R. Polygenic Risk Scoring and Prediction of Mental Health Outcomes. *Curr Opin Psychol* **27**, 77-81 (2019).
9. Lozupone, M. *et al.* The Role of Biomarkers in Psychiatry. *Adv Exp Med Biol* **1118**, 135-162 (2019).
10. Sallis, H.M. *et al.* Genetic Liability to Schizophrenia Is Associated With Exposure to Traumatic Events in Childhood. *Psychol Med*, 1-8 (2020).
11. Gordovez, F.J.A. & McMahon, F.J. The Genetics of Bipolar Disorder. *Mol Psychiatry* **25**, 544-559 (2020).

12. Dvir, Y., Denietolis, B. & Frazier, J.A. Childhood Trauma and Psychosis. *Child Adolesc Psychiatr Clin N Am* **22**, 629-41 (2013).
13. Strålin, P. & Hetta, J. Substance Use Disorders Before, at and After First Episode Psychosis Hospitalizations in a Young National Swedish Cohort. *Drug Alcohol Depend* **209**, 107919 (2020).
14. Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* **179**, 1469-1482.e11 (2019).
15. Shastry, B.S. SNP Alleles in Human Disease and Evolution. *J Hum Genet* **47**, 561-6 (2002).
16. Pugach, I. & Stoneking, M. Genome-Wide Insights Into the Genetic History of Human Populations. *Investigative genetics* **6**, 6-6 (2015).
17. Vieland, V.J. *et al.* Revisiting Schizophrenia Linkage Data in the NIMH Repository: Reanalysis of Regularized Data Across Multiple Studies. *Am J Psychiatry* **171**, 350-9 (2014).
18. Vieland, V.J., Walters, K.A., Azaro, M., Brzustowicz, L.M. & Lehner, T. The Value of Regenotyping Older Linkage Data Sets With Denser Marker Panels. *Hum Hered* **78**, 9-16 (2014).
19. Walters, K.A. *et al.* Meta-Analysis of Repository Data: Impact of Data Regularization on NIMH Schizophrenia Linkage Results. *PloS one* **9**, e84696-e84696 (2014).
20. Berrettini, W.H. Molecular Linkage Studies of Bipolar Disorder. *Dialogues Clin Neurosci* **1**, 12-21 (1999).
21. Chen, W.J. Taiwan Schizophrenia Linkage Study: Lessons Learned from Endophenotype-Based Genome-Wide Linkage Scans and Perspective. *Am J Med Genet B Neuropsychiatr Genet* **162b**, 636-47 (2013).
22. Riley, B. Linkage Studies of Schizophrenia. *Neurotox Res* **6**, 17-34 (2004).
23. Dehghan, A. Genome-Wide Association Studies. *Methods Mol Biol* **1793**, 37-49 (2018).

24. Ott, J., Wang, J. & Leal, S.M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* **16**, 275-84 (2015).

25. Uitterlinden, A.G. An Introduction to Genome-Wide Association Studies: GWAS for Dummies. *Semin Reprod Med* **34**, 196-204 (2016).

26. Hayes, B. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). *Methods Mol Biol* **1019**, 149-69 (2013).

27. Shastry, B.S. SNPs: Impact on Gene Function and Phenotype. *Methods Mol Biol* **578**, 3-22 (2009).

28. Frayling, T.M. Genome-Wide Association Studies: The Good, the Bad and the Ugly. *Clinical medicine (London, England)* **14**, 428-431 (2014).

29. Mattei, J. *et al.* Disparities in Allele Frequencies and Population Differentiation for 101 Disease-Associated Single Nucleotide Polymorphisms Between Puerto Ricans and Non-Hispanic Whites. *BMC Genet* **10**, 45 (2009).

30. Chen, N. *et al.* Allele Frequency Dynamics in a Pedigreed Natural Population. *Proc Natl Acad Sci U S A* **116**, 2158-2164 (2019).

31. Campbell, C.D. *et al.* Demonstrating stratification in a European American population. *Nat Genet* **37**, 868-72 (2005).

32. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New Approaches to Population Stratification in Genome-Wide Association Studies. *Nat Rev Genet* **11**, 459-63 (2010).

33. Price, A.L. *et al.* Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics* **38**, 904-909 (2006).

34. Boehnke, M. & Langefeld, C.D. Genetic Association Mapping Based on Discordant Sib Pairs: The Discordant-Alleles Test. *Am J Hum Genet* **62**, 950-61 (1998).

35. Dudbridge, F. Family-Based Association Studies. *Methods Mol Biol* **713**, 119-27 (2011).

36. Fadista, J., Manning, A.K., Florez, J.C. & Groop, L. The (In)famous GWAS P-Value Threshold Revisited and Updated for Low-Frequency Variants. *European Journal of Human Genetics* **24**, 1202-1205 (2016).

37. Peprah, E., Xu, H., Tekola-Ayele, F. & Royal, C.D. Genome-Wide Association Studies in Africans and African Americans: Expanding the Framework of the Genomics of Human Traits and Disease. *Public Health Genomics* **18**, 40-51 (2015).

38. Hong, E.P. & Park, J.W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics & informatics* **10**, 117-122 (2012).

39. Dauriz, M. & Meigs, J.B. The Power of Numbers. *Diabetologia* **59**, 1400-1402 (2016).

40. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five Years of GWAS Discovery. *Am J Hum Genet* **90**, 7-24 (2012).

41. Horwitz, T., Lam, K., Chen, Y., Xia, Y. & Liu, C. A Decade in Psychiatric GWAS Research. *Mol Psychiatry* **24**, 378-389 (2019).

42. Mehta, D. & Czamara, D. GWAS of Behavioral Traits. *Curr Top Behav Neurosci* **42**, 1-34 (2019).

43. Palk, A.C., Dalvie, S., de Vries, J., Martin, A.R. & Stein, D.J. Potential Use of Clinical Polygenic Risk Scores in Psychiatry - Ethical Implications and Communicating High Polygenic Risk. *Philos Ethics Humanit Med* **14**, 4 (2019).

44. Gratten, J., Wray, N.R., Keller, M.C. & Visscher, P.M. Large-Scale Genomics Unveils the Genetic Architecture of Psychiatric Disorders. *Nat Neurosci* **17**, 782-90 (2014).

45. Fullerton, J.M. & Nurnberger, J.I. Polygenic Risk Scores in Psychiatry: Will They Be Useful for Clinicians? *F1000Research* **8**, F1000 Faculty Rev-1293 (2019).

46. Andreasen, N.C. DSM and the Death of Phenomenology in America: An Example of Unintended Consequences. *Schizophr Bull* **33**, 108-12 (2007).

47. Ford, J.M. *et al.* Studying Hallucinations Within the NIMH RDoC Framework. *Schizophr Bull* **40 Suppl 4**, S295-304 (2014).

48. Sullivan, P.F. The Psychiatric GWAS Consortium: Big Science Comes to Psychiatry. *Neuron* **68**, 182-186 (2010).

49. Wray, N.R. *et al.* Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression. *Nature Genetics* **50**, 668-681 (2018).

50. Goes, F.S. *et al.* Genome-Wide Association Study of Schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet* **168**, 649-59 (2015).

51. Nagel, M. *et al.* Meta-Analysis of Genome-Wide Association Studies for Neuroticism in 449,484 Individuals Identifies Novel Genetic Loci and Pathways. *Nat Genet* **50**, 920-927 (2018).

52. Stahl, E.A. *et al.* Genome-Wide Association Study Identifies 30 Loci Associated With Bipolar Disorder. *Nature genetics* **51**, 793-803 (2019).

53. Senthil, G., Dutka, T., Bingaman, L. & Lehner, T. Genomic Resources for the Study of Neuropsychiatric Disorders. *Mol Psychiatry* **22**, 1659-1663 (2017).

54. Nurnberger, J.I., Jr. *et al.* Diagnostic Interview for Genetic Studies. Rationale, Unique Features, and Training. NIMH Genetics Initiative. *Arch Gen Psychiatry* **51**, 849-59; discussion 863-4 (1994).

55. American Psychiatric Association. Bipolar and Related Disorders. in *Diagnostic and Statistical Manual of Mental Disorders* (Washington, DC, 2013).

56. American Psychiatric Association. Schizophrenia Spectrum and Other Psychotic Disorders. in *Diagnostic and Statistical Manual of Mental Disorders* (Washington, DC, 2013).

57. American Psychiatric Association. Depressive Disorders. in *Diagnostic and Statistical Manual of Mental Disorders* (Washington, DC, 2013).

58. Ferrari, A.J. *et al.* The Prevalence and Burden of Bipolar Disorder: Findings From the Global Burden of Disease Study 2013. *Bipolar Disord* **18**, 440-50 (2016).

59. Sajatovic, M. Bipolar Disorder: Disease Burden. *Am J Manag Care* **11**, S80-4 (2005).

60. Goodwin, G.M. *et al.* Evidence-Based Guidelines for Treating Bipolar Disorder: Revised Third Edition Recommendations From the British Association for Psychopharmacology. *J Psychopharmacol* **30**, 495-553 (2016).

61. Purcell S, N.B., Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC. PLINK: A Toolset for Whole-Genome Association and Population-Based Linkage Analysis. *American Journal of Human Genetics* **81**(2007).

62. Rentería, M.E., Cortes, A. & Medland, S.E. Using PLINK for Genome-Wide Association Studies (GWAS) and data analysis. *Methods Mol Biol* **1019**, 193-213 (2013).

63. Turner, S. qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots. *Journal of Open Source Software* **3**, 731 (2018).

64. Machiela, M.J. & Chanock, S.J. LDlink: A Web-based Application for Exploring Population-specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants. *Bioinformatics* **31**, 3555-7 (2015).

65. Machiela, M.J. & Chanock, S.J. LDassoc: An Online Tool for Interactively Exploring Genome-Wide Association Study Results and Prioritizing Variants for Functional Investigation. *Bioinformatics* **34**, 887-889 (2018).

66. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).

67. Panagiotou, O.A. & Ioannidis, J.P. What Should the Genome-Wide Significance Threshold Be? Empirical Replication of Borderline Genetic Associations. *Int J Epidemiol* **41**, 273-86 (2012).

68. Christoforou, A. *et al.* Linkage-Disequilibrium-Based Binning Affects the Interpretation of GWASs. *Am J Hum Genet* **90**, 727-33 (2012).

69. Bethesda (MD): National Center for Biotechnology Information, N.L.o.M. dbSNP accession: rs12010076. Vol. dbSNP Build ID: 154 (Database of Single Nucleotide Polymorphisms (dbSNP)).

70. Bethesda (MD): National Center for Biotechnology Information, N.L.o.M. dbSNP accession: rs7003372. Vol. dbSNP Build ID: 154 (Database of Single Nucleotide Polymorphisms (dbSNP)).

71. Wang, K.S., Liu, X.F. & Aragam, N. A Genome-Wide Meta-Analysis Identifies Novel Loci Associated with Schizophrenia and Bipolar Disorder. *Schizophr Res* **124**, 192-9 (2010).

72. Ma, M.P., Robinson, P.J. & Chircop, M. Sorting Nexin 9 Recruits Clathrin Heavy Chain to the Mitotic Spindle for Chromosome Alignment and Segregation. *PLoS One* **8**, e68387 (2013).

73. Shin, N. *et al.* Sorting Nexin 9 Interacts With Dynamin 1 and N-WASP and Coordinates Synaptic Vesicle Endocytosis. *J Biol Chem* **282**, 28939-28950 (2007).

74. Witt, S.H. *et al.* Genome-Wide Association Study of Borderline Personality Disorder Reveals Genetic Overlap With Bipolar Disorder, Major Depression and Schizophrenia. *Transl Psychiatry* **7**, e1155 (2017).

75. Clyde, D. Making the Case for More Inclusive GWAS. *Nat Rev Genet* **20**, 500-501 (2019).

76. Peterson, R.E. *et al.* Genome-Wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589-603 (2019).

77. Rosenberg, N.A. *et al.* Genome-Wide Association Studies in Diverse Populations. *Nat Rev Genet* **11**, 356-66 (2010).

78. Chakravarti, A., Clark, A.G. & Mootha, V.K. Distilling Pathophysiology From Complex Disease Genetics. *Cell* **155**, 21-6 (2013).

79. Tam, V. *et al.* Benefits and Limitations of Genome-Wide Association Studies. *Nat Rev Genet* **20**, 467-484 (2019).

80. Shi, M., Umbach, D.M. & Weinberg, C.R. Using Parental Phenotypes in Case-Parent Studies. *Front Genet* **6**, 221 (2015).

81. Marchini, J. & Howie, B. Genotype Imputation for Genome-Wide Association Studies. *Nat Rev Genet* **11**, 499-511 (2010).