2014

Identifying functional variation in schizophrenia GWAS loci by pooled sequencing

Erik Loken
Virginia Commonwealth University

Identifying functional variation in schizophrenia GWAS loci by pooled sequencing

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

Erik Kristen Loken

Director: Brien P. Riley, Ph.D.
Associate Professor, Departments of Psychiatry and Human and Molecular Genetics

Virginia Commonwealth University
Richmond, Virginia
June, 2014

Acknowledgment

I have had no shortage of support during my efforts on this dissertation. I thank my advisor Brien for his guidance in the design and execution of the study and for the insights he has shared with me as my mentor. The work was made possible by the support of Brien and my co-advisor, Ken. In addition to supporting this project, throughout my training Ken has encouraged thinking and broad exploration in the fields of genetics, psychiatry and medicine through reading, discussion, and research. Brandon's support in the laboratory preparing and organizing the samples allowed me to focus on steady completion of the library preparation. I thank my committee, including my co-advisors and Silviu, Mark, and Vlady for their mentoring and help throughout my training. Vlady always emphasized the importance of careful and skilled laboratory technique, and understanding the methods. Silviu guided me on the statistics and was always helpful in answering my questions. Mark advised me on ENCODE data and conservation scores. Finally, I thank my loving wife, Elliott, for all the support she has given me and the sacrifices she has made for me during my studies.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AA	African American
ADHD	attention deficit hyperactivity disorder
ARC	activity-regulated cytoskeleton-associated protein
bp	base pairs
CEPH	Centre d'Etude du Polymorphisme Humain
CEU	CEPH sample of Utah residents with ancestry from northern and western Europe
CNV	copy number variant
CRISP	Comprehensive Read analysis for Identification of SNPs from Pooled sequencing
DISC1	Disrupted-In-Schizophrenia-I
EA	European American
ENCODE	Encyclopedia of DNA Elements
FDR	false discovery rate
FMRP	fragile X mental retardation protein
GATK	Genome Analysis Toolkit
GCTA	Genome-wide Complex Trait Analysis
GWAS	genome-wide association study
GxE	gene-environment interaction
GxG	gene-gene interaction
HLA	human leukocyte antigen
IBD	inflammatory bowel disease
ICCSS	Irish Case-Control Study of Schizophrenia
LD	linkage disequilibrium
LOD	logarithm of the odds
MAF	minor allele frequency
Mb	Megabases
MHC	Major Histocompatibility Complex
miRNA	microRNA
MZ	monozygotic twin
NMDAR	N-methyl-D-aspartate receptor
OR	odds-ratio
PGC	Psychiatric Genomics Consortium
PGC-1	Psychiatric Genomics Consortium Phase 1
PGC-2	Psychiatric Genomics Consortium Phase 2
SNP	single nucleotide polymorphism
SNV	simple nucleotide variant
TFBS	transcription factor binding sites
TSI	Tuscan Italian

UTR............. untranslated region

Abstract

IDENTIFYING FUNCTIONAL VARIATION IN SCHIZOPHRENIA GWAS LOCI BY POOLED SEQUENCING

Erik Kristen Loken, B.S.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2014

Director: Brien P. Riley, Ph.D.
Associate Professor, Departments of Psychiatry and Human and Molecular Genetics

Schizophrenia demonstrates high heritability in part accounted for by common simple nucleotide variants (SNV), rare copy number variants (CNV) and, most recently, rare SNVs Although heritability explained by rare SNVs and CNVs is small compared to that explained by common SNVs, rare SNVs in functional sequences may identify specific disease mechanisms. However, current exome methods do not capture a large proportion of potentially functional bases where rare variation may impact disease risk: as much as two-thirds of conserved sequences lie outside the exome in non-coding regions of cross-species evolutionary constraint. We reasoned that the candidate loci from the Psychiatric Genomics Consortium Phase 1 (PGC-1) schizophrenia study represent good target loci to test for the impact of rare SNVs in non-coding constrained regions. We developed custom reagents to capture mammalian constrained non-coding regions, exons, and 5'- and 3'-untranslated regions (UTRs) in the 12 PGC-1 loci for

pooled sequencing in 912 cases and 936 controls. Compared to our coding targets, our noncoding targets contain substantially more highly conserved bases (46,412 vs. 31,609) and variants (390 vs. 193). Using C-alpha to detect excess variance due to aggregate risk increasing or decreasing rare SNV effects, we identified signals attributable to alleles with MAF < 0.1% in both coding sequences and in functional non-coding sequences, including variants within ENCODE transcription factor binding sites, DNase hypersensitive regions, and histone modification sites in neuronal cell lines. We also observed significant excess risk-altering variation in the CUB domain of CSMD1, a gene expressed in the developing central nervous system. These results support the hypothesis that common and rare variants in the same loci contribute to schizophrenia risk, but highlight the need to expand capture strategies in order to detect trait-relevant sequence variation in a broader set of functional sequences.

Introduction

Relevant Background

Schizophrenia is an idiopathic, complex mental disorder with a lifetime risk of 0.4% (Lichtenstein et al., 2006; Saha, Chant, Welham, & McGrath, 2005). The onset of the disease is typically the early to mid-twenties for males and late twenties for females and can present either acutely with onset of a psychotic episode or with a longer prodromal phase. Schizophrenia as a syndrome was first described as “dementia praecox” by Emil Kraepelin (Kraepelin, 1899). He was the first to recognize the disease as separate from bipolar disorder or, as he called it, manic-depression. To Kraepelin, the negative symptoms, those that reflect a loss of normal functions, including avolition, anhedonia, alogia, and blunted affect were the most relevant in distinguishing schizophrenia. Now the positive symptoms, those that reflect an excess or distortion of normal functions, including delusions, hallucinations, and disorganized speech, are given far more weight. Avolition describes the lack of initiation in goal-directed behavior, anhedonia is the lack of pleasure, alogia is the lack of fluency of thought and speech and blunted affect is a reduction in the range and intensity of emotional expression. Delusions are distortions of inferential thinking, hallucinations are distortions in perceptions and disorganized speech is a distortion in language. According to the DSM-5 (A. P. A. American Psychiatric Association, American Psychiatric Association D. S. M. Task Force, 2013), at least one positive symptom must be present during a one-month period in addition to another positive symptom, catatonia, or negative symptom and the disturbance must be present for six months to diagnose schizophrenia. Previously, the DSM-IV (A. P. A. American Psychiatric Association, American Psychiatric Association Task Force on D. S. M. I. V., 2000) included five subtypes (paranoid, disorganized, catatonic, undifferentiated and residual), but as of the DSM-5 these subtypes are not included.

Schizophrenia has a very high heritability, established as 0.81 in twins (Sullivan, Kendler, & Neale, 2003) and 0.64 in a study of the Swedish population (Lichtenstein et al., 2009). Further evidence of its heritability is a strong sibling recurrence risk of 8.55 (Lichtenstein et al., 2006). Schizophrenia has a high genetic correlation with bipolar disorder (0.68), major depressive disorder (0.43), and a lower genetic correlation with autism spectrum disorder (0.16) (Lee et al., 2013). These correlations have been supported by data from more recent studies detailed in this chapter in the form of joint and independent associations of loci with multiple psychiatric disorders.

Despite the evidence for high heritability, monozygotic twin (MZ) concordance is only 48%, suggesting that genetic risk factors do not entirely explain schizophrenia and that environmental risk factors are also important (Onstad, Skre, Torgersen, & Kringlen, 1991). The number of environmental and non-genetic risk factors studied and found contributing to schizophrenia is numerous, including paternal age (Petersen, Mortensen, & Pedersen, 2011), season of birth (J. J. McGrath & Welham, 1999), famine (St Clair et al., 2005), cannabis use (Hill, 2014), urban birth (J. McGrath & Scott, 2006), migration (Cantor-Graae & Selten, 2005), and prenatal infection (Khandaker, Zimbron, Lewis, & Jones, 2013). A meta-analysis of the prenatal maternal influenza infection literature has found no evidence of contributions to schizophrenia risk from the 1957 pandemic of influenza (Selten, Frissen, Lensvelt-Mulders, &

Morgan, 2010), but the prenatal infections *Toxoplasma gondii* and herpes simplex virus (HSV-2) show some effect (Brown & Derkits, 2010; Khandaker et al., 2013).

Environmental risk factors are divided into common environmental factors to which both twins are exposed, and unique environmental factors to which twins are independently exposed. Common environmental risks are estimated by twin studies to represent 11% of the variance for schizophrenia while unique environmental factors are estimated to represent the remainder of the variance, 8% (Sullivan et al., 2003). Paternal age, season of birth, famine, urban birth and prenatal infection represent common environmental risk factors. Cannabis use and migration, may represent unique environmental risk factors. The examples listed that would contribute to the relatively low twin concordance are cannabis use, migration, measurement error (included in unique environment), and unknown unique environmental risk factors. Future studies investigating the environmental causes of schizophrenia will use prospective birth cohort studies instead of ecological and retrospective designs. By incorporating genotypes with environmental data, it is possible significant gene by environment interaction could be found (Brown, 2011), explaining low MZ concordance.

Linkage and Candidate Genes

These findings have inspired a large molecular genetics effort to identify the source of heritability. Pedigree analysis showed no evidence for one-locus mendelian transmission of schizophrenia (Elston, Namboodiri, Spence, & Rainer, 1978). This lack of evidence for a single causal locus and the swift drop in recurrence risk from monozygotic twins (52.1) to siblings (8.6) and offspring (10), suggested a multilocus model for heritability (Risch, 1990). This led to a large number of linkage studies searching for what investigators thought would be a few loci responsible for schizophrenia. Early studies found no loci in linkage with schizophrenia (Kendler & Diehl, 1993), but later studies point to weak linkage at 22q12-q13, 8p22-p21, 6p24-p22, 13q14.1-q32, 5q21-q31, 10p15-p11, 6q21-q22, 15q13-q14 and 20q11-q22 (McGuffin, Tandon, & Corsico, 2003) (Riley, 2004) with minimal agreement between studies. Positional loci suggested from these findings include *NRG1, G72*, *DAAO*, *DTNBP1,* and *COMT,* and an additional linkage discovery that was named Disrupted-In-Schizophrenia-I (*DISC1*) (Ishizuka, Paek, Kamiya, & Sawa, 2006) (Chubb, Bradshaw, Soares, Porteous, & Millar, 2008). The estimated effect sizes from linkage that were found were relatively small, and the positional candidate loci did not produce any significant variants on follow-up (Kirov, O'Donovan, & Owen, 2005). For the sample sizes collected, the lack of strong linkage in schizophrenia suggests that no locus existed with a recurrence risk > 3 (Owen, Craddock, & O'Donovan, 2005).

Theoretical Candidate Genes

Theoretical candidate genes for schizophrenia such as the dopamine receptors *DRD3* and *DRD2,* the serotonergic receptor *HTR2A*, and the glutamatergic gene *GRM3* have been studied and proposed as candidate genes because of their role in systems thought to be perturbed in schizophrenia, but until recently there has been a complete lack of any robust findings from these candidates (Kirov et al., 2005). More recently, evidence for theoretical candidate loci impacting schizophrenia has been found. The second phase of the Psychiatric Genomics consortium (PGC-2) genome-wide association study (GWAS) (S. W. G. o. t. P. G. Consortium, 2014), a large study of 36,989 schizophrenia cases and 113,075 controls, found associations in *DRD2*, the dopamine receptor target for antipsychotic drugs, and genes involved in glutamatergic neurotransmission (*GRM3*, *GRIN2A*, *SRR, GRIA1*). Genes (*GRM5, PPEF2,* and *LRP1B)* that encode protein products associated with the glutamate receptor N-methyl-D-aspartate receptor (NMDAR) have been found to have rare protein-altering variants in five schizophrenia pedigrees (Timms et al., 2013)

Genome-Wide Association Studies

An early GWAS (Stefansson et al., 2009) of 2,663 schizophrenia cases and 13,498 controls found associations in the Major Histocompatibility Complex (MHC) of chromosome 6, and the genes *TCF4* and *NRGN*. Compared to other GWAS at these sample sizes, the results were not very impressive (W. T. C. C. Consortium, 2007). An additional study (Purcell et al., 2009) of 3,322 schizophrenia cases and 3,587 controls found associations for *MYO18B,* the MHC region, *ZNF804A,* and six imputed human leukocyte antigen (HLA) alleles. The MHC region included over 450 single nucleotide polymorphisms (SNPs) across several Megabases (Mb). To analyze the GWAS signal for polygenicity, the authors summed odds-ratio (OR) weighted allele counts of independent variants per individual, and compared the scores of cases and controls. The aggregation of the signals for a great number of alleles of a small and neutral effect explained 3% of the variance for schizophrenia. This result indicated that although few loci had been observed through GWAS, causal variants were distributed throughout the genome at lower effect sizes, and that increasing sample size could lead to the discovery of more loci. The most successful published efforts so far have been the very large collaborative efforts of the Psychiatric Genomics Consortium (PGC) GWAS to analyze large schizophrenia samples (Ripke S, 2011) (Ripke et al., 2013). The PGC-1 schizophrenia GWAS of 9,394 cases and 12,462 controls identified 8 loci by peak significant SNPs, *MIR137, PCGEM1, TRIM26, CSMD1, MMP16, CNNM2-NT5C2, STT3A* and *CCDC68-TCF4* (hyphens included for loci with multiple genes). In addition to these loci the authors combined the schizophrenia cases with 16,374 bipolar disorder cases for a joint association study, finding *ANK3*, *CACNA1C* and *ITIH3-ITIH4* associated with the combined disorders. SNPs were intragenic to their loci except for *PCGEM1* (343 kb distance to nearest gene), *MMP16* (421 kb), *STT3A* (1 kb) and *CCDC68* (126 kb). A peak significant SNP was intragenic to *TCF4*, a locus previously implicated in schizophrenia (Steinberg et al., 2011). *MIR137* was a novel and interesting result for schizophrenia in that it encodes a microRNA (miRNA) that is predicted to target four of the other significant loci in the study (*TCF4*, *CACNA1C*, *CSMD1* and *C10orf26*). An additional study (Ripke et al., 2013), combining PGC-1 with a Swedish national sample of 5,001 cases and 6,243 controls, expanded the significant loci to 22, of which 13 were new. This study represents the most up-to-date peer-reviewed results for common variation in schizophrenia. Using Genome-wide Complex Trait Analysis (GCTA) (Yang, Lee, Goddard, & Visscher, 2011), a method to estimate the variance in liability explained by all SNPs, researchers estimated SNP heritability to be 0.27, assuming a population risk of 0.004 (Lichtenstein et al., 2006), and 0.33, assuming a population risk of 0.01. The upper bound for schizophrenia heritability is 0.64 to 0.81, based on population and twin evidence, and the lower bound for schizophrenia SNP heritability is 0.27 to 0.33, using a population risk of 0.004 or 0.01. These results suggest that between one-third and one-half of schizophrenia heritability comes from common SNPs and at least half of the heritability of schizophrenia is left to be explained by other sources. A number of sources of this additional heritability have been suggested and studied intensively, including rare SNPs/indels (SNVs) identified by sequencing, CNVs identified directly from array comparative genomic hybridization (aCGH) or indirectly from intensity data on GWAS arrays, gene-environment interactions (GxE) (Iyegbe, Campbell, Butler, Ajnakina, & Sham, 2014; Maric & Svrakic, 2012;

Modinos et al., 2013; Svrakic, Zorumski, Svrakic, Zwir, & Cloninger, 2013), and gene-gene interactions (GxG or epistasis) (Chiesa et al., 2013; Nicodemus et al., 2010; Won et al., 2014).

The PGC-2 schizophrenia study (S. W. G. o. t. P. G. Consortium, 2014) is currently in submission. It expands the samples size to 36,989 schizophrenia cases and 113,075 controls. This is the largest molecular genetics study of schizophrenia or any other neuropsychiatric disorder and it found 108 distinct associated loci, 83 of which were not previously observed in schizophrenia. It is the first to strongly implicate *DRD2*, the target of antipsychotic pharmaceuticals used to treat schizophrenia. It also implicates genes involved in glutamatergic neurotransmission (*GRM3*, *GRIN2A*, *SRR, GRIA1*). More associations (*CACNAB2* and *CACNA1I*) in voltage gated calcium channel subunits were observed including *CACNA1C*. Associations in active enhancers from 56 different tissues and cell lines showed significant enrichment not only in brain, but also tissues with immune functions such as the CD19 and CD20 B-lymphocyte cell lines.

Copy Number Variants

Rare variants implicated in schizophrenia first came in the form of CNVs (Rees et al., 2014) (Walsh et al., 2008). Many of these studies were not measuring heritable contributions from CNVs because they focused on *de novo* mutations (Malhotra et al., 2011) (Stefansson et al., 2008). 22q11.2 deletion syndrome, also known as velocardiofacial syndrome or DiGeorge syndrome was found to be primarily *de novo* in newly diagnosed patients, with 90% of the deletions being *de novo* and 10% being inherited (Bassett, Marshall, Lionel, Chow, & Scherer, 2008). Some standing CNVs affecting schizophrenia, and inherited 2p16.3 deletion affecting *NRXN1*(Kirov et al., 2008), duplications of 16p13.1 (Ingason et al., 2011), and duplications of

16p11.2 (McCarthy et al., 2009), have been discovered. Heritable CNVs show minor allele frequency (MAF) ranges of 0.30% in cases for the 16p11.2 duplication compared to 0.03% in controls (8.4 OR) and 0.12% in cases for the 16p13.1 duplication compared to 0.04% in controls (3.27 OR). Examples of replicated signals in deletions show MAF between 0.23-0.32% for 1q21.1, with an OR ranging 6.6-14.8, and 0.17-0.3% MAF for 15q13.3, with an OR ranging 11.5-17.9 (Sebat, Levy, & McCarthy, 2009). These loci are nonspecific risk factors for other disorders, such as developmental delay and congenital malformations for 1q21.1 and generalized epilepsy and mental retardation for 15q13.3 (Sebat et al., 2009). In addition, 16p11.2 deletions are more common in autism and developmental delay (0.78% MAF, OR 38.7) while not at all more common in schizophrenia. CNVs have low MAFs, high ORs, and pleiotropic effects. The presence of CNVs impacting schizophrenia risk in addition to GWAS data support a complex genetic architecture with rare and common variation.

Exome Studies

The first exome studies of schizophrenia were published in the last few years. One of the first found an elevated *de novo* mutation rate in 14 schizophrenia trios (Girard et al., 2011). Another published concurrently a *de novo* design of 53 case and 22 control trios, identified 40 *de novo* mutations in cases, one of which was in *DGCR2*, a gene located in the 22q11.2 DiGeorge Syndrome locus (B. Xu et al., 2011). Using rare inherited variants for comparison, the authors observed an excess of non-synonymous variants that were *de novo*. This study was later expanded to 231 schizophrenia trios and 34 control trios (B. Xu et al., 2012). The excess signal from *de novo* nonsynonymous SNVs was replicated and these variants were more enriched in genes with greater prenatal expression. A study of 166 cases and 307 controls (Need et al., 2012) with a strategy using a follow-up cohort of 2,756 cases and 1,932 controls for further testing found no significantly associated SNVs. With the relatively small sample size, there was only power to detect variants at 1% MAF with a relative risk of 6 for a nominal association leading to follow-up. One exome study of five large schizophrenia pedigrees (Timms et al., 2013) found rare protein-altering variants implicating glutamatergic neurotransmission. Protein-altering variants from one of three genes, *GRM5, PPEF2,* and *LRP1B*, whose protein products are associated with NMDAR, were discovered in all five pedigrees. A recent exome study (McCarthy et al., 2014) of 57 sporadic and familial schizophrenia trios found a 3.5-fold increase of *de novo* mutations in the sporadic probands compared to the familial probands. These *de novo* mutations were found in excess in genes with a high estimated probability of haploinsufficiency. An overlap of loci with *de novo* mutations was observed with autism (*AUTS2, CHD8* and *MECP2*) and intellectual disability (*HUWE1* and *TRAPPC9*).

A new round of larger exome studies has been published for schizophrenia. The first is a study of *de novo* mutations from trios (Fromer et al., 2014). The study authors acknowledge that they are using the same study design as many *de novo* CNV trio studies, but with exome sequencing they now have the resolution to identify single base *de novo* mutations that impact schizophrenia. Using data from 623 exomes from schizophrenia trios and 731 controls from published data sets, they did not identify any excess rate of *de novo* point mutations in schizophrenia probands. They were able to identify enrichment of nonsynonymous *de novo* mutations in genesets with independent evidence for involvement in schizophrenia. They also found an enrichment of loss-of-function mutations in genes identified in autism and intellectual disability studies of *de novo* variants. Certain genesets also had an enrichment of nonsynonymous mutations, most notably those encoding components of the activity-regulated cytoskeleton-associated protein (ARC) complex, the NMDAR complex, and genes regulated by the fragile X mental retardation protein (FMRP). The authors targeted the first two sets for analysis because of the presence of *de novo* CNVs in the ARC and NMDAR complexes (Kirov et al., 2012). In addition to nonsynonymous mutations, the ARC and NMDAR complexes were also significantly enriched for loss-of-function *de novo* mutations in cases. FMRP and its targets are implicated by *de novo* mutations in autism (Iossifov et al., 2012) that specifically impact these brain-expressed genes related to synaptic function (Darnell et al., 2011). Because of its connection to autism and synaptic function, the authors hypothesized that *de novo* mutations would also be a factor for schizophrenia. They observed that nonsynonymous *de novo* mutations in FMRP target genes were significantly enriched in cases. There was also an enrichment of loss-of-function *de novo* mutations in genes with excess loss-of-function *de novo* mutations found in autism and intellectual disability studies. The genes with loss-of-function *de novo* mutations in autism were also enriched for nonsynonymous *de novo* mutations in schizophrenia.

The other major exome study published in the same issue of Nature was a case-control study of exomes sequences from 2,536 schizophrenia cases and 2,543 controls (Purcell et al., 2014). The authors of this study and the trio study shared data and Purcell et al. were able to confirm the signal in the ARC complex for disruptive (nonsense, essential splice site and frameshift) singletons and < 0.5% MAF variants. The NMDAR complex association from Fromer et al. was not replicated. The genes implicated in the *de novo* SNV studies had an enrichment of < 0.5% MAF distruptive variation. Focusing on a composite geneset of loci previously implicated in schizophrenia by GWAS (Ripke et al., 2013), CNV(Kirov et al., 2012) (Sullivan, Daly, & O'Donovan, 2012), and *de novo* SNV studies (Girard et al., 2011; Purcell et al., 2014; B. Xu et al., 2012), the authors observed a significant enrichment of singleton and <

0.5% MAF disruptive variation in case samples. A significant enrichment of.< 0.1% MAF case variation was observed when including missense variants predicted to be damaging by multiple prediction algorithms in the test set. There was an enrichment of singleton disruptive variants for voltage-gate calcium ion channel genes, especially in *CACNA1C*, which was implicated as a joint bipolar disorder and schizophrenia risk locus in the PGC-1 analysis (Ripke S, 2011). Expanding their analysis to crossover points with autism studies the authors observed enrichment of case disruptive and nonsynonymous alleles in FMRP targets identified from mouse brain (Darnell et al., 2011) for < 0.1% MAF variation. FMRP targets identified from human kidney (Ascano et al., 2012) did not contain any case enrichment of disruptive variants, suggesting that for these FMRP targets the location of the targeting to the brain across species is more meaningful for testing than the targets across tissues but from the same species.

The exome studies build upon the previous GWAS and CNV studies, but are not definitive for the impact of rare variation in schizophrenia. Overall, the results point towards schizophrenia genes having brain functions, specifically synaptic network functions. There is overlap between bipolar disorder and schizophrenia both phenotypically and genotypically. Shared SNP-based coheritability for the two disorders is estimated to be 0.68 (Lee et al., 2013) and certain specific genes, such as the jointly associated voltage-gated calcium ion channel *CACNA1C* (Ripke S, 2011), have now been observed as independent associations for bipolar disorder (Ferreira et al., 2008) and schizophrenia (S. W. G. o. t. P. G. Consortium, 2014). In both the *de novo* SNV and the case-control exome studies, FMRP targets were identified as a source of rare variant enrichment, risk, and overlap with autism. The case-control exome study did not find enrichment in the common SNP loci from the GWAS studies. The test did have a suggestive *p*-value for disruptive MAF < 0.5% variants (0.0037) but this did not meet the significance threshold based on multiple testing. The lack of rare variants impacting the common variant loci could be due to the relatively small samples size from the exome study, 2,536 exome cases compared to a meta analysis of 9,394 PGC-1 cases, and 5,001 Swedish cases.

A final interesting component of the exome studies is their estimation of the variance of schizophrenia risk explained by polygene scores in an overlapping sample of 5,079 individuals due to SNPs (5.7%), rare CNVs (0.2%), and disruptive mutations (0.4%). The authors note that the rare CNVs and disruptive mutations explain an order of magnitude less variance that the SNPs. The authors admit that these estimates represent a conservative lower bound for the true estimates, as the disruptive mutations are only from the composite set (see above) which represents about 10% of the genome (~2,500 genes). Including the rest of the genome, more samples, and potentially a more relaxed set of variants beyond disruptive variation (nonsense, essential splice site and frameshift), should increase the variance explained by rare SNVs. Rare variation is observed less frequently than common variation, requiring larger sample sizes and effect sizes to detect associations. Samples sizes for exome studies so far have been an order of magnitude lower than for GWAS and exome sequencing in schizophrenia has yet to obtain a statistically significant signal for either a single allele or for alleles aggregated across a single gene. Increasing sample size to approximately the level now available in the largest GWAS of schizophrenia seems necessary to provide the power needed for identification of either single loci enriched for variation in cases or specific rare variants associated with schizophrenia.

Non-coding Variation

It is important to consider why the exome and not the genome is the current standard for large sample sequencing studies. The cost for genome sequencing has dropped rapidly since

2001, but the rate of decrease has been less dramatic in the last two years with the estimated cost at about $4,008 as of January 2014 (Figure 1) (Wetterstrand, 2014). Exome sequencing using target capture (Asan et al., 2011) is far cheaper with recent rates quoted as low as $500 per exome (Perkel, 2013). Investigators reason that the most important part of the genome in which to look for variation are the coding regions. Some exome capture reagents also target UTRs and miRNA.

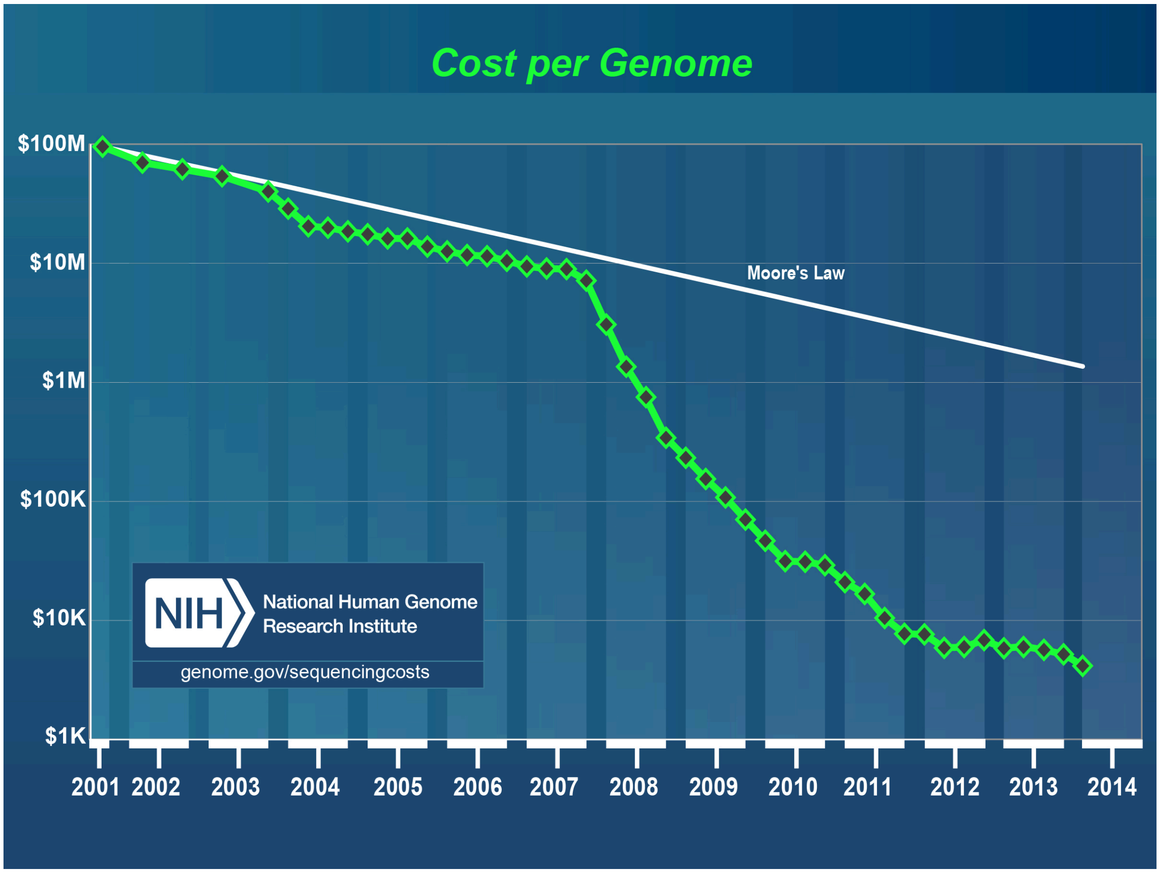

Figure 1: The cost of whole genome sequencing over the years

These approaches are missing potentially important target regions. A study of human evolutionary constraint by comparison of 29 mammalian genomes observed that at least 5% of the human genome is under purifying selection (Lindblad-Toh et al., 2011). Selection implies functionality of the underlying sequence, only 1.5% of which is coding in the human genome. As much as two-thirds of constrained, likely functional sequences are non-coding and are therefore completely missed by exome studies. For a complex genetic disorder like schizophrenia with no observed single rare coding changes associated with disease, this could represent a critical amount of contributing variation. The Encyclopedia of DNA Elements (ENCODE) (Rosenbloom et al., 2013) study is a collaborative study with the goal of creating a complete catalog of functional elements for the human genome. As of 2013 it consists of 2,886 experiments from multiple sites made public by the University of California for download and use on its genome browser (Kent et al., 2002). The ENCODE project has identified regions in the human genome that include sites of modifications to histones which effectively increase or decrease local gene expression, sensitivity to DNase-I, indicating open chromatin and potential for transcription, and transcription factor binding sites. Data have been included in the ENCODE project from many cell lines, including those originating from glia and neurons that are most useful for the study of brain disorders like schizophrenia. Histone modifications sites, DNase-I hypersensitive sites, and transcription factor binding sites occupy largely non-coding regions not included in a traditional exome capture. Current capture reagents (even those including UTRs) do not include these potentially important functional sequences and are likely to be missing an important functional component of the genome.

Common Variant Loci

A great amount of effort in sample collection, planning, and funding went into current exome studies. Two strategies can be used to reduce the scope of work and expense of identifying rare causal variation associated with disease: pooled sequencing (Futschik & Schlotterer, 2010) and targeting of smaller regions than the exome. In the pooled sequencing approach, groups of samples are combined in equimolar amounts and sequenced as one sample. This allows library preparation costs, a significant component of sequencing costs, to be reduced by the factor of the pool size. Common variant loci have been hypothesized to also contain rare associated variation. This has been shown in inflammatory bowel disease (IBD) (Rivas et al., 2011), which includes both Crohn's disease and ulcerative colitis, autoimmune diseases of the whole digestive system and the colon, respectively. Rivas et al. used pooled targeted sequencing of 56 genes from common variant loci identified in GWAS to identify potentially causal rare variation in 350 cases and 350 controls. They identified 70 rare variants that cause a change to the proteins implicated in GWAS of IBD. This study strategy allows low cost sequencing of a reduced target area. Low cost sequencing allows a greater sample size to be sequenced and combined with the reduced target area, increasing the power of the study to detect effects of rare variation. Rivas et al. identified multiple rare variants through follow-up genotyping in 16,054 Crohn's disease cases, 12,153 ulcerative colitis cases, and 17,575 healthy controls, some protective and some damaging. The observation of protective rare variant for a disease is not isolated to IBD, having been observed in coronary heart disease (Cohen, Boerwinkle, Mosley, & Hobbs, 2006) and plasma low-density lipoprotein levels (Cohen, Pertsemlidis, et al., 2006).

Aims

Taking this information in aggregate, it is clear that much of the heritability is from unknown sources (Manolio et al., 2009) (Lee, Wray, Goddard, & Visscher, 2011) and the large majority of heritability that is measurable is polygenic and spread throughout the genome at common and rare allele frequencies. The common variant loci from the PGC-1 schizophrenia study very likely represent the best candidate loci for unbiased follow-up in a targeted rare variant study. Unlike the IBD study, we hypothesized that by including relevant non-coding variation identified in the study of 29 mammals (Lindblad-Toh et al., 2011) we would increase our functional target substantially compared to the coding sequence of these loci alone. We had access to 912 cases in our Irish Case-Control Study of Schizophrenia (ICCSS) and 936 unscreened Irish controls from the Trinity Biobank for the study. We aimed to identify functional rare variation in the coding and non-coding sequences of the top schizophrenia common variant loci. To accomplish this goal, we adapted the pooled, targeted sequencing approach from Rivas et al. to reduce costs by focusing the target, while still maximizing sample size. Instead of using PCR amplification of exons, we used in-solution hybridization capture (Bansal, Tewhey, Leproust, & Schork, 2011) to reduce sample processing load and still be able to sequence many different targets within our loci. For analysis, we not only used Burden tests (Price et al., 2010), which allow for directionally specific aggregate effects, but also C-alpha (Neale et al., 2011), a test which measures bidirectional aggregate effects. This allows us to take advantage of the possibility of a similar observation in schizophrenia of not only damaging variants but protective variants like in IBD, coronary heart disease (Cohen, Boerwinkle, et al., 2006), and low-density lipoprotein levels (Cohen, Pertsemlidis, et al., 2006). Using all these techniques and all this knowledge, we aimed to detect rare variation impacting schizophrenia in common variant loci not only in coding regions, but also the two-thirds of potentially functional variation that is non-coding with goals of improving the understanding of schizophrenia genetics in its top loci and schizophrenia genetics, generally.

Methods

Sample Information

We selected 912 schizophrenia cases from the ICCSS and 936 unscreened Irish controls from the Trinity Biobank for sequencing. The ICCSS was collected by Kenneth Kendler of the Virginia Commonwealth University from 1999 to 2003 (Riley et al., 2010). Affected subjects were selected from inpatient and outpatient psychiatric facilities in the Republic of Ireland and Northern Ireland. Subjects were eligible for inclusion if they had a diagnosis of schizophrenia or poor outcome schizoaffective disorder by DSM-III-R criteria. Diagnoses were confirmed by a blind expert diagnostic review and subjects must have reported all four grandparents as being born in Ireland or the United Kingdom. The use of DSM-III-R maintained consistency with previous DSM-III-R era collections done by the group. Each proband completed a personal interview with a detailed family history. The control subjects are blood donors from the Trinity Biobank in Dublin. Although not given a formal diagnostic interview, all control subjects deny any personal or family history of psychosis. The relatively low prevalence of schizophrenia (~1%) makes these donors suitable controls. The ethnic homogeneity of the sample avoids population stratification in our studies.

Target Capture

Agilent SureSelect Custom solution-based capture allows capture of custom designed regions of the genome. The small size array allowing up to 500kb of sequence was ideal for our application. To define target intervals, we first examined linkage disequilibrium (LD) around associated SNPs from the PGC results using Haploview and HapMap data version 3 release 27 from individuals of European descent, the Centre d'Etude du Polymorphisme Humain (CEPH) sample of Utah residents with ancestry from northern and western Europe (CEU), and Tuscan Italian (TSI) samples. The CEU and TSI samples were chosen to best approximate the PGC-1 sample composition with the available HapMap ethnicities. Pairwise markers > 500 kb apart were ignored and individuals with > 50% missing genotypes were excluded. We included the associated SNPs from the schizophrenia analysis and the joint schizophrenia and bipolar association loci where bipolar and schizophrenia where analyzed as one phenotype (Table 1). The schizophrenia study loci are *MIR137, PCGEM1, TRIM26, CSMD1, MMP16, CNNM2, NT5C2, STT3A, CCDC68*, and *TCF4*. The joint targets are *ITIH3/4*, *ANK3*, and *CACNA1C*. The target interval is defined as the region with $R^2 > 0.8$ with the associated SNP. If the interval overlaps part or all of one or more genes, then all exons and constrained sequence from those genes were included in the target set. For several loci, the LD interval did not overlap any gene (*PCGEM*, *MMP16*, *CCDC68*). The *MMP16* region only contained the original associated SNP from the PGC analysis. These three loci only contributed to 520 bp (base pairs) of constrained regions total (Table 1). Because we were underpowered to detect association for the 3, we were left with 9 out of the 12 loci for locus testing.

Within these regions we selected all coding and UTR sequences in addition to regions from the 29 mammals paper (Lindblad-Toh et al., 2011) including human and primate accelerated regions, regulatory motif instances, peaks indicating constraint structure in promoters and, finally, SiPhy-omega and SiPhy-pi constrained regions with a logarithm of odds (LOD) score requirement of at least 7.325 for inclusion to allow the best regions within the constrained sequence space of 500kb. These regions were then used to design a SureSelect custom capture library consisting of 120 bp baits (Agilent Technologies, Santa Clara, CA) using SureDesign software with moderate masking of repeats. The final Agilent SureSelect target capture design included 84.5% of the non-coding regions that the 29 mammals study (Lindblad-Toh et al., 2011) considered constrained at the 10% false discovery rate (FDR) level.

Table 1: GWAS loci target Intervals

Locus	Chr.	Association	Total Target
MIR137	1p21.3	SCH[1]	10232
PCGEM1	2q32.3	SCH	140
TRIM26	6p21-22	SCH	3888
CSMD1	8p23.2	SCH	29399
MMP16	8q21.3	SCH	1
CNNM2 & NT5C2	10q24	SCH	26609
STT3A	11q24.2	SCH	4846
CCDC68	18q21.2	SCH	379
TCF4	18q21.2	SCH	67448
ITIH3/4	3p21.1	Joint[2]	18640
ANK3	10q21.2	Joint	43303
CACNA1C	12p13.33	Joint	31821
Totals			236706

[1]Schizophrenia association results. [2]Results from joint bipolar and schizophrenia associations

Library Preparation and Sequencing

Samples were run on 1.5% agarose gels to check for the presence of high molecular weight DNA. Only samples with high molecular weight DNA (lanes 1-9 and 11-18, Figure 2) were included in this project; degraded samples (example in lane 10, Figure 2) were excluded. Nanodrop spectrophotometry was used to confirm sample purity using a 260/280 ratio of 1.8 to 2.0. One round of PicoGreen (Life Technologies, Carlsbad, CA) dsDNA quantitation was performed to measure the concentration double stranded DNA only. After adjusting sample concentrations to 50 ng/µL of dsDNA we performed a second round of PicoGreen quantitation and adjusted concentrations to 23 ng/µL, the recommended shearing concentration for library preparation. We then performed a third round of PicoGreen quantitation to determine the precise individual sample concentrations for equimolar pooling.

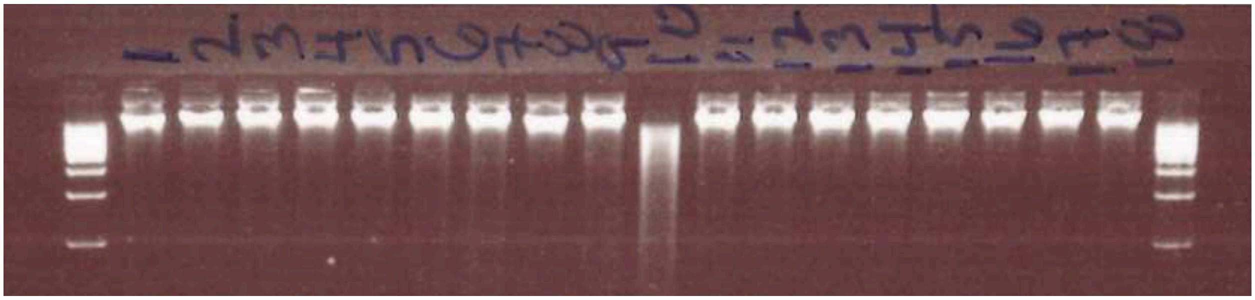

Figure 2: Representative gel results for quality control of cases and controls. Here are 18 controls one of which is degraded (lane 10) and the rest of which contain high molecular weight DNA. The sample from lane 10 was too degraded to include in the study.

We constructed pools of 24 case or 24 control subjects. Twenty-four samples per pool were ideal for several reasons. First, Illumina, our chosen sequencing platform has a nominal 1% error rate. Therefore 50 samples or 100 chromosomes would be the maximum allowable per pool to still allow detection of a singleton allele in the pool at or above the error rate. Reducing pool size to 24 samples (or 48 chromosomes) doubles the signal expected from a singleton allele and substantially improves detection of singleton alleles over errors in the pool. Second, we had reagents for 96 pools based on kit sizes from Agilent, which gave a minimum size of 20 samples per pool to include all selected subjects. Finally, 4 sets of 24 samples fit on a standard 96-well plate, which increases workflow efficiency and allows for spare capture reagents. We pooled each set of 24 case or control samples into equimolar pools basing the exact volume on the final PicoGreen concentration readings using robotic liquid handling to maximize accuracy. The final study sample included 38 case and 39 controls pools.

We sheared 130 µL from each pool stock using the Covaris S2 instrument (Covaris, Woburn, MA) with a duty cycle of 10%, intensity of 5, 200 cycles per burst and the frequency sweeping mode for 6 cycles of 60 seconds each to get a target peak base pair size of 150 to 200 bp. We followed this with purification of the pool using Agencourt AMPure XP beads (Beckman Coulter Inc., Pasadena, CA). Every sheared pools size distribution was assessed using the Agilent 2100 Bioanalyzer with a DNA 1000 chip (see representative trace, Figure 3). Shearing leaves damaged ends to the double stranded DNA, so we followed this step with end repair using T4 DNA polymerase, Klenow DNA Polymerase, T4 Polynucleotide Kinase, and reagents from the SureSelect Library Prep Kit, followed by an additional AMPure XP purification. We added A bases to the 3' end of the fragments for each pool using exo(-) Klenow fragment, dATP, and reagents from the SureSelect Library Prep Kit followed by an AMPure XP purification. We then ligated adapters to the pool fragments using T4 DNA Ligase and SureSelect Library Prep Kit reagents followed by AMPure XP purification. To produce enough library for hybridization we amplified the libraries for 5 cycles of PCR using Herculase II Fusion DNA Polymerase and SureSelect Library Prep Kit reagents. We then purified the pool libraries using AMPure XP and measured concentration and quality using a DNA 1000 chip on the 2100 Bioanalayzer. We dried the pools using a vacuum concentrator and reconstituted them with nuclease-free water at a concentration of 220 ng/µL for hybridization.

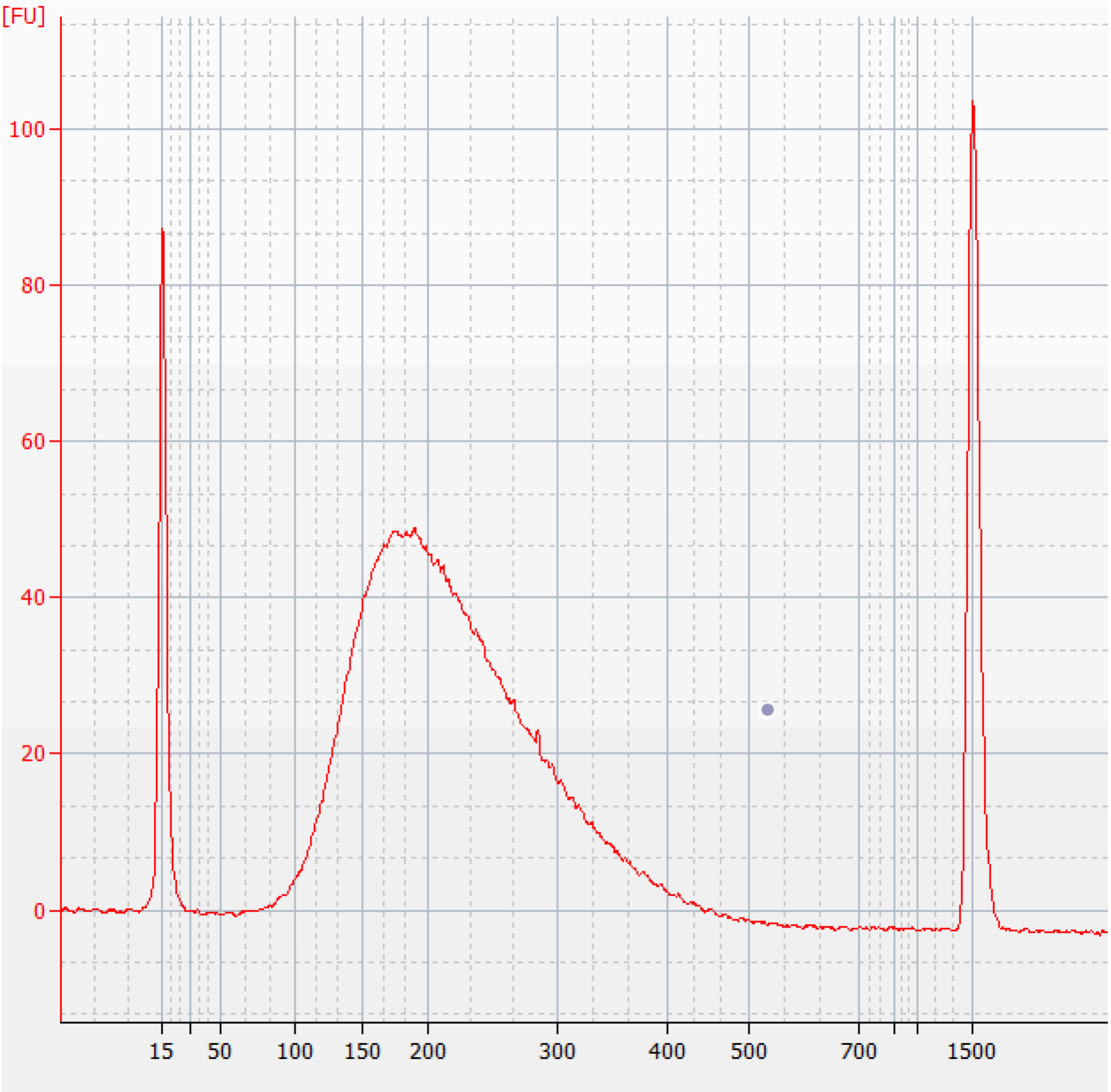

Figure 3: Representative trace of pool 56 after shearing using the Covaris S2. The peak is at 173 bp, between the targets of 150 and 200.

We preformed target capture using solution hybridization of pool library fragments with the 120 bp baits produced for the SureSelect Custom Capture. We denatured the concentrated pool libraries and then combined them with hybridization buffers. We added index blocking oligonucleotides to prevent index and adapter sequences from inhibiting hybridization between the baits and the target sequence. We incubated the final hybridization mixture for 24 hours at 65˚C with a heated lid at 105˚C.

We biotinylated the hybridization baits, allowing selective capture of bait-target hybrid fragments using Dynabeads MyOne Streptavidin-coated T1 magnetic beads (Life Technologies, Carlsbad, CA). We mixed the Streptavidin T1 beads with the hybridization mixture and allowed it cool to room temperature while mixing for 30 minutes on a rotator followed by two washes using SureSelect Kit reagents. In the denaturation step of PCR, the library and bait disassociate and the library fragments move into solution. This allowed us to perform PCR directly from the beads in the PCR solution to amplify the library and add the index tags. We used Herculase II Fusion DNA Polymerase and a random selection of 77 of the 96 Illumina PCR Primer Indexes for the PCR reactions and we ran 16 cycles for each pool. We purified each pool using AMPure XP, which conveniently removes the used Dynabeads MyOne Streptavidin-coated T1 magnetic beads from solution, and analyzed for concentration and library quality using the DNA High Sensitivity Chip on the Bioanalyzer (see representative trace, Figure 4).

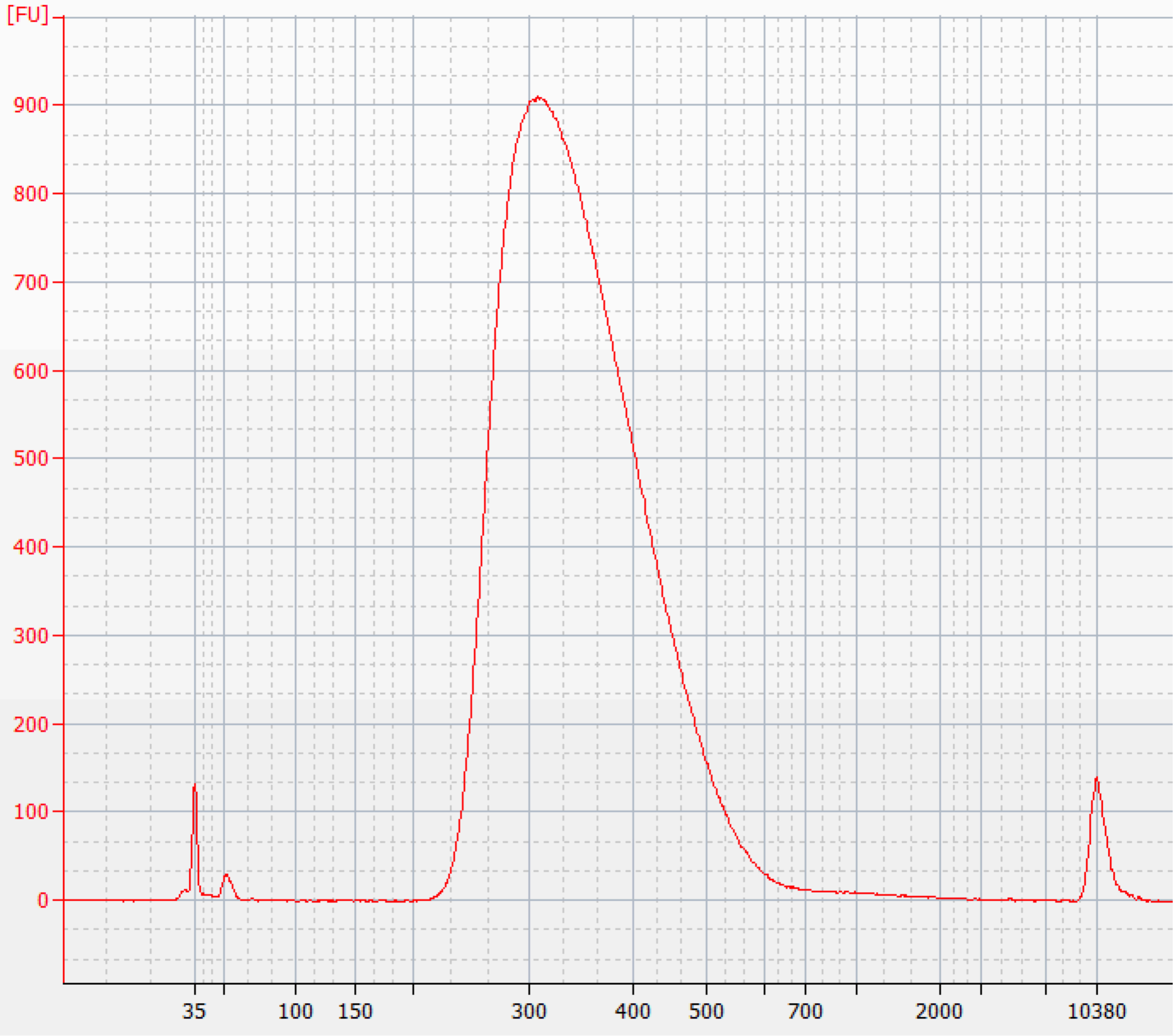

Figure 4: Representative trace of pool 56 after target capture and PCR. The peak is now above 300 bp, due to the addition of the adapters and index tags to the library.

We performed qPCR on each pool using the QPCR NGS Library Quantification Kit (Illumina, San Diego, CA) to measure the concentration of each pool for equimolar megapooling. To test the accuracy of megapool construction, we used these concentrations to make a test megapool of the post-capture library pools using robotic liquid handling and ran this test megapool on an Illumina MiSeq 150 bp paired-end run at the VCU core lab. We aligned the reads from this run using BWA (Li & Durbin, 2009) and we used the coverage results to adjust the 77 individual pool concentrations one more time to construct the final equimolar megapool.

We sent the megapool to the VCU core lab for 103 bp paired-end sequencing on the Illumina HiSeq 2500 where it was cross loaded onto 5 lanes of the flowcell.

Sequence Data Processing

After we received the raw reads from the core lab, we aligned the reads to the genome using BWA version 0.7.0 (Li & Durbin, 2009). BWA uses the Burrows-Wheeler transformation to allow quick alignment of the reads to the genome. We used the Genome Analysis Toolkit (GATK) version 2.5 (DePristo et al., 2011) to do local realignment of reads to reduce mismatches near indels. Each base in a sequencing experiment has a Phred quality score indicating the probability the base call is an error according to the formula $Q = -10log_{10}P$ where P is the error probability and Q is the Phred score. The score is included as output from the sequencing platform's internal software and can be improved using covariates and alignment information to recalibrate the scores. We performed base quality score recalibration using GATK using read group, reported quality score, cycle (base position in read), and context (dinucleotide and trinucleotide) to produce more accurate empirical quality scores for variant calling.

Variant Calling

We called variants from the finalized realigned and recalibrated sequence data using CRISP (Bansal, 2010) (Comprehensive Read analysis for Identification of SNPs from Pooled sequencing), a variant caller developed specifically for read data from pooled sequencing. CRISP is able to call variants in pools by first comparing the allele count distribution across all the pools in the experiment using contingency tables. It then evaluates the probability that multiple non-reference base calls at a locus are due to sequencing errors. To distinguish a sequencing error from a real variant, it uses the distribution of alternate allele counts across the pools as a guide. If the distribution is similar across the pools, this is more likely to reflect sequencing error. Differences in the distribution of apparent allele counts across pools indicate varying allele counts in the pool and support the existence of a true variant at the site. CRISP also takes the sequencing error rate into account and computes a probability that a certain number of alternate alleles would be present at a site given sequencing error alone. The lower this probability, the greater the evidence for a true variant at the site. Finally, the number of chromosomes per pool is considered to ensure that the alternate allele frequency does not deviate too far below $\frac{1}{h}$, where h is the number of haplotypes in the pool. If the frequency of alternate base calls at a site is much lower than expected using a binomial test, it is more likely that the signal represents sequencing error.

Quality Control

We conducted all data processing and analyses in R version 3.0.2 (R Core Team, 2013). We assigned hardcoded allele counts where the alternate allele count of the pool for a particular variant was assigned to the *n* allele count for which its sequence read data alternate allele count was within $\frac{1}{96}$ of $\frac{n}{48}$ (48 chromosomes per pool). This method gives an approximation of the allele count for the pool based very simply on the allele count which most closely matches the read count data. Based on these data, we observed the metric singletons per pool showed serious outliers. We were concerned this may indicate spurious results resulting from potential problems during pooling and library preparation. We identified four pools, two case and two control, greater than the median number of singletons per pool (43) plus the standard deviation (51.25) (Figure 5). We filtered the four pools before final allele count calling and analysis. We filtered variants based on quality score using the R package mclust version 4.2 (Chris Fraley, 2012). Mclust identified 8 clusters of quality scores for which a division between the first 3 and last 5 at a quality score of 875 represented an obvious division between low quality and high quality variants for filtering and maintaining for further analyses (Figure 6). Of the 9928 variants, we filtered 2736 based on quality score. For each pool we calculated the R^2 between allele frequencies for 2,651 variants imputed from Affymetrix array genotypes using the 1000 Genomes data to the pool allele frequencies based on read count for the variants in the pools. The median and standard deviation for the R^2 values across the 77 pools were 0.9919 and 0.0087. We excluded four pools, one case and three controls, that were two standard deviations below the median (Figure 7). We were left with a total 69 pools, 35 case and 34 controls, equivalent to 1,656 samples, 840 cases and 816 controls.

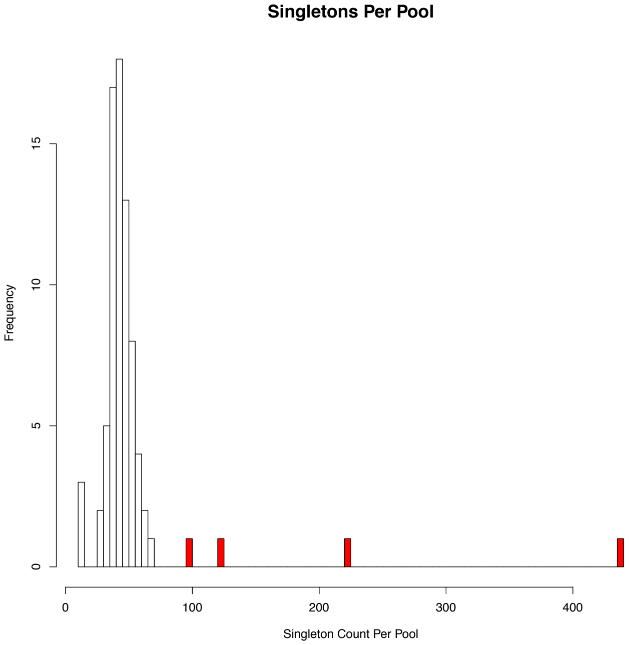

Figure 5: Histogram of sample-wide singletons per pool with pools greater than the median plus standard deviation marked red.

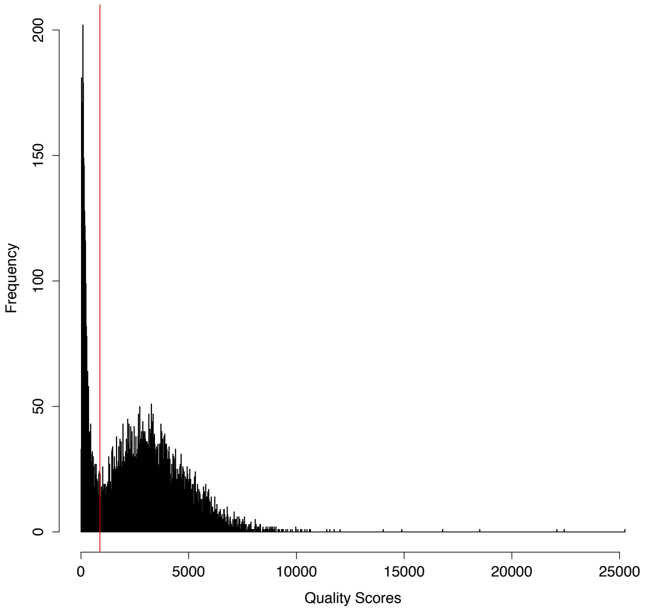

Figure 6: Distribution of variant quality score with vertical red line representing division between filtered low quality variants (left) and maintained high quality variants (right)

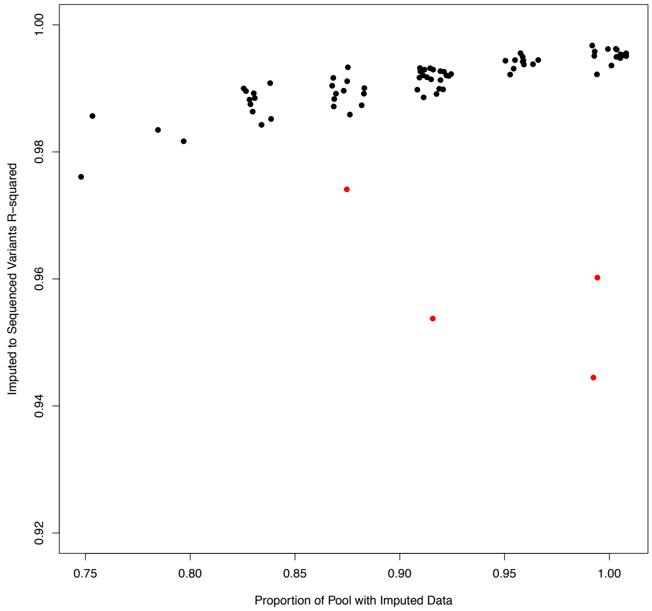

Figure 7: Each pool plotted with its imputed genotypes to pooled sequenced genotypes R^2 and its proportion of pool with imputed data. Some samples included in the project did not have GWAS data for imputation. Filtered pools are in red.

Allele Count Calling

We used a Bayesian method to assign exact allele counts for each variant per pool. We assigned the probability for the allele count for each pool according to the formula $P(allele\ count \mid reads\ in\ pool) = \frac{P(reads\ in\ pool \mid allele\ count)\ P(\ allele\ count)}{\sum P(reads)}$. We calculated the probability that the read count for reference and alternate alleles in the pool occurred using the binomial distribution, taking into account the coverage level and number of alternate bases observed. We assumed there were 0 to 48 alternate alleles in the pool, $p = \frac{count}{48}$, assigning 0 and 48 alleles 0.5% and 99.5% probability, respectively. We used the binomial distribution to determine the probability that a particular pool has allele counts 0 to 48 given the average minor allele frequency across all the pools calculated from the alternate and reference read data. We selected the allele count per variant per pool with the highest probability as the true allele count and we saved the probabilities for each allele count for simulations during permutation significance testing. This allele count fine-tuning lowered the excess rate of intrapool multiple detection for certain rare alleles. For example, for experiment wide doubletons the hardcoded allele counts yields 103 doubleton variants out of 764 where both alleles are observed in a single pool. This is expected only 21 times according to the binomial distribution. For the probabilistic allele counts, intrapool doubletons only occur 44 times representing more than two-fold reduction in the deviation from expectation. The same trend in improvement is observed with tripletons (39 to 26, 18.0 expected) and quadrupletons (34 to 24, 15.3 expected) (Figure 8).

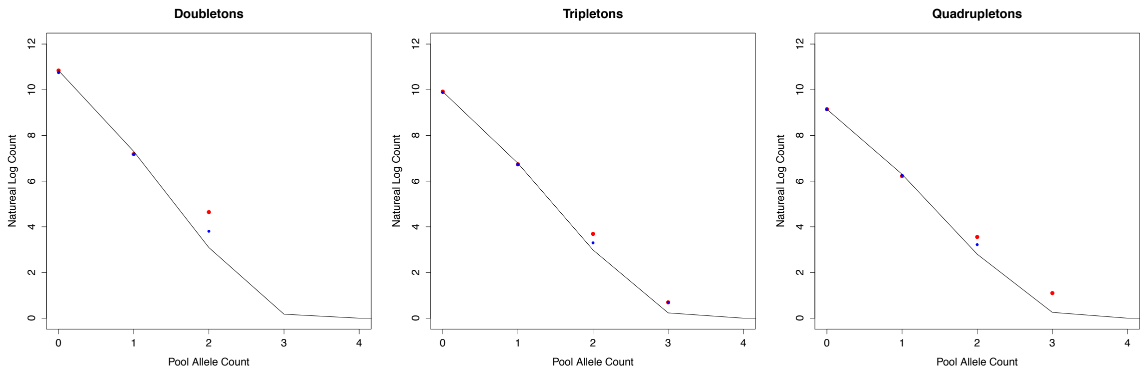

Figure 8: Counts per pool of study-wide doubletons, tripletons and quadrupletons with hardcoded allele counts (red) and Bayesian allele counts (blue). Bayesian counts are closer to the expected number found per pool (black line)

Statistical Tests

We used variable threshold C-alpha (Neale et al., 2011) and Burden (Price et al., 2010) testing. The strategy of the C-alpha test is to measure excess binomial variance in the distribution of allele counts between cases and controls for a set of variants. C-alpha measures general excesses of risk variation bidirectionally, for protective and damaging alleles. Burden testing measures excess enrichment of alleles for a set of variants in one direction, either increasing risk or decreasing risk for the variable in question. C-alpha and Burden tests will both detect excesses unidirectionally but for bidirectional effects in the same set of variants, the signal will be cancelled out for Burden testing and will be increased for C-alpha. Variable threshold tests allow the detection of signals at different MAFs without arbitrarily choosing the threshold for the test. Not only does the variable threshold test allow for detection of signals, but it also provides information regarding the MAF of variants contributing to the signal. We assessed the significance of tests by predicting empirical *P*-values through permutation of case/control pool status and a FDR cutoff of 0.2. We performed a variable threshold version of the tests in which, for every possible allele count in the study under 5% , we calculated a Z-score for all variants at that allele count threshold and below.

For C-alpha we calculated the Z-scores at each threshold using the following formula as outlined in Neale et al., 2011:

$$Z = T/\sqrt{c}$$

where

$$T = \sum_{i=1}^{m} [(y_i - n_i p_0)^2 - n_i p_0 (1 - p_0)]$$

and

$$c = \sum_{n=2}^{max\, n} m(n) \sum_{u=0}^{n} [(u - np_0)^2 - np_0(1 - p_0)]^2 f(u|n, p_0)$$

In *T*, we calculate the variance for each variant where y_i is the alternate allele case count for the i'th variant, n_i is the total alternate allele count, and p_0 is the proportion of the case samples in our case pools out of the total number of samples. The variable *c* standardizes *T*. Here *m(n)* is the number of variants with *n* alternate allele counts. The sum is taken for each possible case alternate allele count *u*.

For Burden tests, we calculated the Z-scores at each threshold using the following formula as outlined in Price et al., 2010:

$$Z = \sum_{i=1}^{m} \sum_{j=1}^{n} \xi_i^T C_{ij} (\pi_j - \bar{\pi}) / \left[\sum_{i=1}^{m} \sum_{j=1}^{n} (\xi_i^T C_{ij})^2 \right]^{1/2}$$

At each threshold, we summed the alternate allele counts in cases *C* where *i* indexes variants and *j* indexes pools. The symbol π is the pools case status and $\bar{\pi}$ *i*s the mean case status.

We performed 10,000 permutations of case status per test. For all Z-scores in the test and permutations at each threshold, we divided by the standard deviation of the permuted Z-scores at each threshold to standardize the Z-scores across the MAF thresholds. We determined *P*-values by comparing the maximum Z-scores from the 10,000 permutations to the maximum Z-score for the test. We assessed significance of tests using a FDR cutoff of 0.2.

ENCODE Regions

For DNase I hypersensitive sites, we used DNase-seq Peaks from SK-N-SH_RA and BE2_C cell lines to represent neuronal positions and from Gliobla, HA-h and NH-A cell lines to represent glial positions. For transcription factor binding sites (TFBS), we used combined TFBS SPP-based peaks and TFBS PeakSeq peaks from SH-SY5Y and SK-N-SH_RA to represent neuronal positions and from Gliobla, U87 and NH-A to represent glial positions. SH-SY5Y data contained peaks for transcription factor GATA2, and SK-N-SH_RA data contained peaks for CTCF, p300, RAD21, USF1, and YY1. Gliobla contained peaks for CTCF and POL2, U87 contained peaks for NRSF, and NH-A contained peaks for CTCF. For histone modifications sites, we used H3K27me3, H3K36me3, and H3K4me3 peaks in SK-N-SH_RA cell lines to represent neuronal positions and used H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3 peaks in NH-A cell lines to represent glial positions.

Results

Sequence Data

We collected 1,612,969,337 reads with 166,135,841,711 base calls across all 77 pools, averaging 20,947,654 reads per pool. An average of 57.1% of the reads were mapped on target with a range of 45.0% to 62.3% and a standard deviation of 3.3%. On average 44.4% of base calls were within the target regions. Average per base coverage per individual was 79.2 with a range of 59.8 to 87.3 and a standard deviation of 5.0. 79.7% of bases within targets had Phred ≥ 20 and the average coverage for Phred ≥ 20 was 59.4 (44.6 to 69.2 range, 3.7 standard deviation). 98.4% of the target bases had at least 20x average coverage per sample.

We called 9,928 variants using CRISP (Bansal, 2010). After filtering pools based on excess singleton detection and poor correlation with known genotypes for the pool, 7,029 variants and 1,656 samples (840 cases and 816 controls) remained. We removed 426 variants for low quality scores (see methods) and an additional 24 variants that fell below our inclusion threshold of 20x coverage for a total of 6,579 passing variants. Finally, we only tested variants with MAFs 5% or lower, excluding an additional 1,129 variants. Of the 5,450 remaining variants 2,944 were singletons that are not tested in C-alpha because they do no provide a deviation from an expected distribution, as random chance will always have a singleton allele in either a case or a control sample. This left 5,450 variants for Burden testing (Price et al., 2010) and 2,506 variants for C-alpha testing (Neale et al., 2011). There were 348 indels called out of the 5,450 variants. For doubletons and greater 182 variants of 2,506 were indels. Of the 6,579 variants that passed filtering 3,881 were novel and 2,698 were previously observed. Figure 9 shows the distribution of variants by MAF.

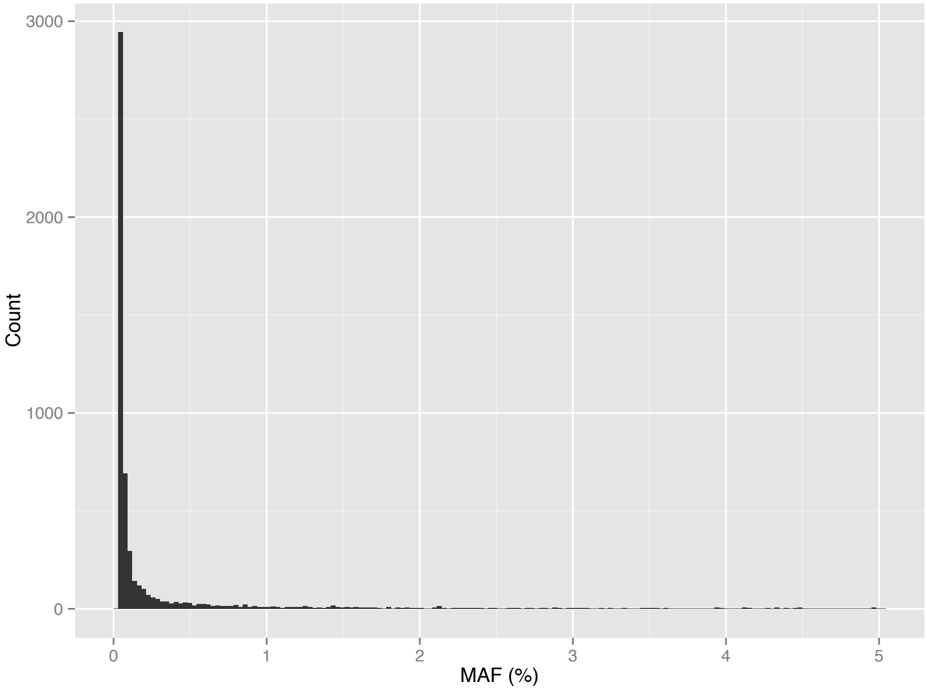

Figure 9: Histogram of MAF in variants that passed all filtering and that are less than 5% MAF.

Summary

Our testing strategy involved applying the Burden and C-alpha pooled association tests to sets of the variants that passed filtering. We selected the variant sets to probe and clarify the role of rare variation in schizophrenia. Starting from the most general set, every variant in the study less than 5% MAF, we then narrowed the focus to sets of study-wide coding and non-coding variants. We used conservation scores to group non-coding variants into low conservation and high conservation positions to better understand the signal in the non-coding regions and we also used conservation scores to compare the sets of coding and non-coding variants with the most highly conserved base positions. We tested variants grouped into sets by ENCODE (Rosenbloom et al., 2013) functional elements. Finally, to detect the influence of rare variation specific to a particular locus, we tested locus variant sets individually. The Burden pooled association test measures an aggregation of excess variants with unidirectional effects, either protective or damaging (see Methods). C-alpha measures an aggregation of excess variants in the test with bidirectional effects, both protective and damaging. These tests are complimentary in that a Burden tests indicates the general direction of effect for the excess risk-altering variation in the set, but will not detect an even mixture of excess protective and damaging variation while C-alpha will detect the mixture, but not indicate a general direction of effect for excess risk-altering variation. No Burden tests were significant at 20% FDR so all results that follow refer to C-alpha results. The results for all tests are summarized in Table 2.

Table 2: Test results. Tests with a (B) indicate the test was a Burden test otherwise it is a C-alpha test.

General Tests	MAF	Allele Count	Variants ≤ Threshold	Z_{max}	*P*-value	*q*-value
Entire Target	0.09%	3	986	4.31	0.011	0.111
Entire Target (B)	0.12%	4	4072	-2.78	0.519	0.601
Non-coding	0.09%	3	828	3.89	0.024	0.111
Non-coding (B)	0.12%	4	3424	-2.99	0.285	0.453
Coding	0.09%	3	158	3.32	0.071	0.182
Coding (B)	2.17%	72	823	-1.48	1.000	0.744
High Impact	0.06%	2	8	1.36	0.489	0.587
High Impact (B)	0.33%	11	89	-0.77	1.000	0.744
Constrained	4.20%	139	1277	3.28	0.089	0.205
Constrained (B)	1.99%	66	2798	-1.32	1.000	0.744
Locus Tests						
MIR137	2.90%	96	116	1.59	0.476	0.580
MIR137 (B)	0.12%	4	176	-1.60	0.995	0.743
ITIH3/4	0.72%	24	94	2.49	0.188	0.353
ITIH3/4 (B)	0.27%	9	226	-1.62	0.977	0.739
TRIM26	1.42%	47	38	2.15	0.128	0.271
TRIM26 (B)	0.06%	2	30	-1.75	0.752	0.686
CSMD1	1.45%	48	669	3.44	0.040	0.142
CSMD1 (B)	0.09%	3	1327	-1.45	1.000	0.744
ANK3	0.09%	3	143	3.22	0.062	0.168
ANK3 (B)	4.26%	141	787	-1.85	0.989	0.742
CNNM2-NT5C2	2.20%	73	143	1.94	0.421	0.550

CNNM2-NT5C2 (B)	1.06%	35	308	1.14	1.000	0.744
STT3A	2.11%	70	20	2.54	0.176	0.338
STT3A (B)	2.11%	70	56	-0.91	0.998	0.743
CACNA1C	0.21%	7	232	1.44	0.608	0.638
CACNA1C (B)	0.09%	3	577	-2.11	0.944	0.733
TCF4	0.30%	10	299	4.06	0.023	0.111
TCF4 (B)	4.23%	140	1017	2.29	0.885	0.720
ENCODE Elements Tests						
DNase glial	0.06%	2	61	2.28	0.322	0.483
DNase glial (B)	3.11%	103	386	-2.70	0.246	0.417
DNase neuronal	4.20%	139	118	3.96	0.031	0.127
DNase neuronal (B)	0.09%	3	197	-1.94	0.880	0.719
Histone glial	2.90%	96	586	4.15	0.022	0.111
Histone glial (B)	0.06%	2	880	-2.36	0.836	0.708
Histone neuronal	0.09%	3	135	3.70	0.048	0.153
Histone neuronal (B)	2.32%	77	657	-1.41	1.000	0.744
TFBS glial	2.39%	79	104	2.26	0.266	0.436
TFBS glial (B)	2.39%	79	235	-2.02	0.808	0.701
TFBS neuronal	2.87%	95	231	4.26	0.015	0.111
TFBS neuronal (B)	0.03%	1	298	1.50	1.000	0.744
Conservation Median Split Tests						
phastCons high conservation	1.12%	37	855	3.17	0.105	0.234
phastCons high conservation (B)	0.21%	7	1849	-1.11	1.000	0.744
phastCons low conservation	3.05%	101	994	4.07	0.013	0.111
phastCons low conservation (B)	0.12%	4	1710	-2.97	0.285	0.453
phyloP high conservation	0.06%	2	297	3.41	0.054	0.160
phyloP high conservation (B)	0.21%	7	1855	-1.58	1.000	0.744
phyloP low conservation	3.14%	104	1009	4.11	0.017	0.111
phyloP low conservation (B)	0.12%	4	1697	-2.86	0.387	0.529
SiPhy-pi high conservation	0.06%	2	286	3.37	0.058	0.164
SiPhy-pi high conservation (B)	0.12%	4	1715	-1.28	1.000	0.744
SiPhy-pi low conservation	0.09%	3	438	4.28	0.010	0.111
SiPhy-pi low conservation (B)	0.12%	4	1709	-3.24	0.110	0.242
Functional Domain Tests						
CUB-CSMD1	1.39%	46	118	4.86	0.005	0.111
CUB-CSMD1 (B)	0.66%	22	304	1.43	0.999	0.744
Sushi-CSMD1	0.09%	3	8	2.34	0.227	0.397
Sushi-CSMD1 (B)	0.12%	4	53	1.26	0.988	0.741
High Conservation Tests						

phyloP > 2	1.12%	37	186	3.44	0.050	0.156
phyloP > 2 (B)	0.06%	2	403	-2.36	0.582	0.628
phyloP > 2 coding	0.06%	2	15	2.93	0.088	0.204
phyloP > 2 coding (B)	0.06%	2	135	-2.19	0.509	0.596
phyloP > 2 non-coding	1.12%	37	129	4.09	0.019	0.111
phyloP > 2 non-coding (B)	0.06%	2	268	-1.19	1.000	0.744

Study-Wide Tests

Our first test was an analysis of all 2506 variants detected in our entire target set. At a threshold of ≤ 0.09% MAF, in our study size representing variants with 2 or 3 alternate alleles detected in the sample as a whole, we observe a significant excess of variance in the distribution of alleles between cases and controls compared to random expectation because these alleles are uniquely or preferentially represented in one phenotypic group or the other (*P* = 0.011, *q* = 0.111, Figure 10). At this low MAF, alleles contributing to signal are uniquely present in cases or controls, for example both doubleton alleles being seen in cases or all three tripleton alleles being seen in controls.

We were interested in comparing independent contributions from coding and non-coding sequences, so we split the variants into a non-coding variants group and an exon and UTR variants group for further testing. Testing each set independently, we again detect peak signals at ≤ 0.09% MAF in both non-coding (*P* = 0.024, *q* = 0.111) and coding (*P* = 0.071, *q* = 0.182) subsets (Figure 10). For both our coding and non-coding variants we observed very rare variation driving the signal of excess unequal case/control allele distribution.

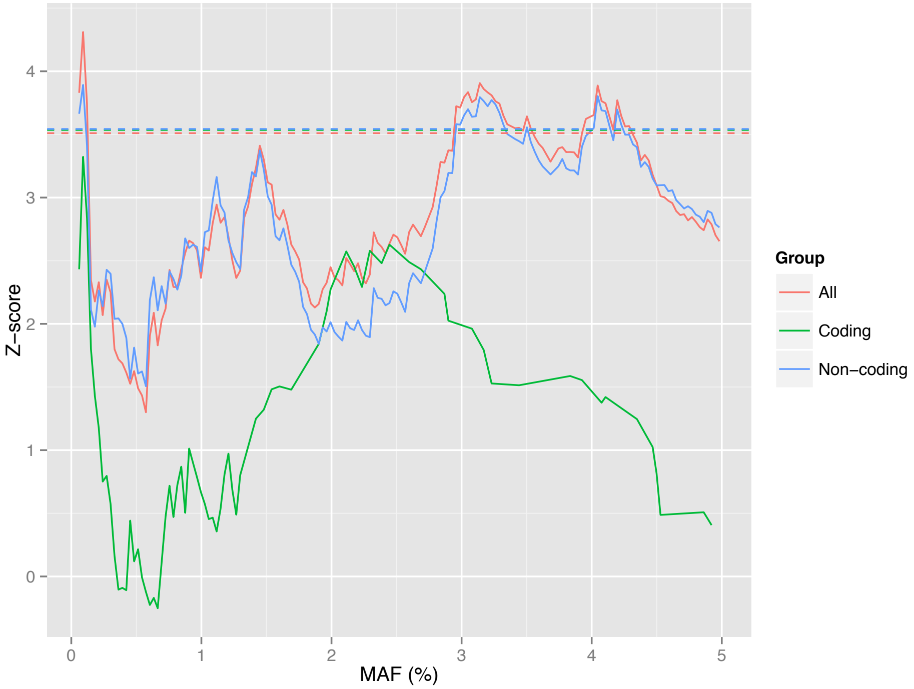

Figure 10: Z-scores at each MAF threshold for all, coding plus UTRs and non-coding variation. For each MAF threshold we calculate a Z-score for all variants at the threshold and below. Each point on the line represents the Z-score for all variants at that MAF and below. The horizontal dotted lines are the levels for which only 5% of the permutation Z-max scores are greater than or equal to for the test.

Conservation

In addition to the peak signal at ≤ 0.09% MAF that we formally tested, the graph of non-coding variant results in Figure 10 also suggests that non-coding alleles with higher MAF from 3.14% to 4.05% are also differentially distributed between cases and controls. We examined these signals further by dividing the non-coding variants into more and less highly conserved sets using a median split for each of three different measures of conservation, phyloP, SiPhy-pi, and phastCons (Garber et al., 2009; Pollard, Hubisz, Rosenbloom, & Siepel, 2010). Scores had been generated using the three methods from Multiz alignments (Blanchette et al., 2004) from 46 species of the placental mammals phylogenetic tree (Murphy et al., 2001). PhyloP and SiPhy-pi measure conservation at a single base while phastCons incorporates flanking bases. Unlike phastCons and phyloP, SiPhy-pi allows for detection of biased substitution patterns. The consideration of the flanking regions in phastCons makes it useful for identifying regions of conservation, instead of positions of conservation. The consideration of biased substitution patterns in SiPhy-pi makes it useful for such special cases. We consider phyloP to be the most general measurement of conservation because it measures conservation at a single base and considers any base changes across species as evidence for lower conservation. We examined the similarity of conservation scores, and since the phastCons measure generally takes values of 0 or 1, our comparison was limited to phyloP and SiPhy-pi scores. PhyloP and SiPhy-pi scores show a reasonable degree of correlation (0.548, Figure 11). Points lying off the diagonal are due to the biased substitution detection in SiPhy-pi for certain sites that increases the conservation score compared to phyloP, which does not detect these patterns.

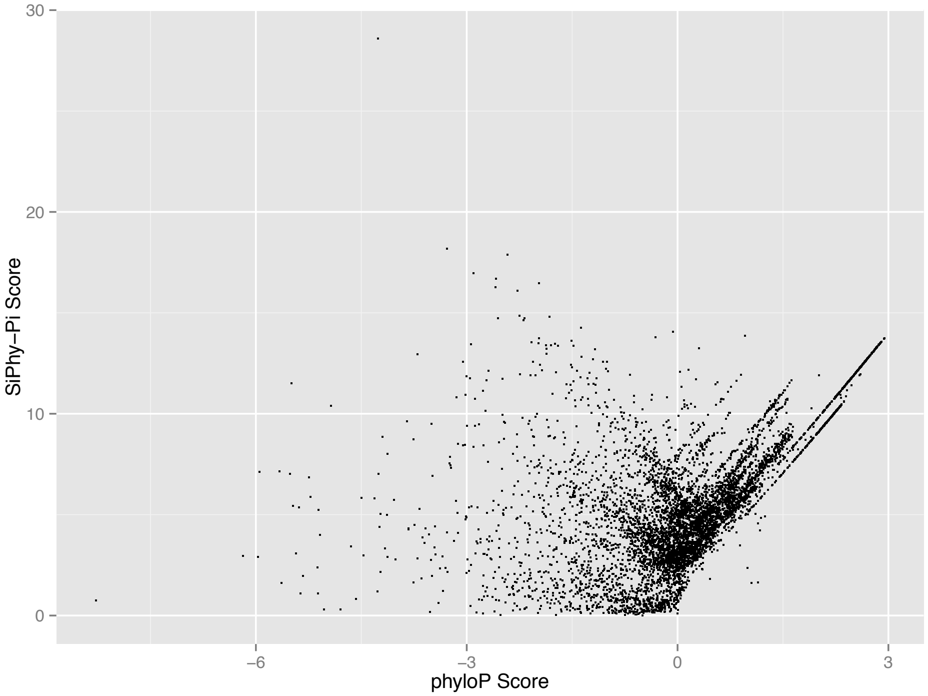

Figure 11: Scatter plot of phyloP scores vs. SiPhy-pi scores for all 6,579 post-filtering variants. The correlation between the scores is 0.548. The fanning off the diagonal trend is due to detection of biased substitution patterns using SiPhy-pi.

For the phyloP (Figure 12) high and low conservation split, we observe a significant excess of variance in the distribution of alleles at 0.06% MAF (*P* = 0.054, q = 0.160) for the highly conserved positions and ≤ 3.14% MAF (*P* = 0.017 *q* = 0.111) for the less conserved positions. The less conserved base position variant set shows a clear excess of signal coming from higher MAF compared to the more conserved base position variant set which has a peak signal from doubleton variants.

Variants at 0.06% MAF (*P* = 0.058, *q* = 0.164) at highly conserved positions and ≤ 0.09% MAF for less conserved positions (*P* = 0.010, *q* = 0.111), defined by SiPhy-pi (Figure 13), were also unequally distributed between cases and controls. While still having a rarer signal in more conserved positions, the SiPhy-pi score MAF difference is not as large as seen in phyloP.

Variants ≤ 3.05% MAF in less conserved positions as defined by phastCons (*P* = 0.013, *q* = 0.111) were unequally distributed between cases and controls (Figure 14). Variants in highly conserved phastCons positions had a similar pattern to phyloP with a lower MAF peak signal (1.12%), but this test was not significant. Although the test was not significant, the lower MAF peak signal in more conserved positions still matched the patterns seen in phyloP and SiPhy-pi.

Across the three measures of conservation, we observe a clear pattern of significant differences in allelic distributions between cases and controls coming from higher frequency variants at less conserved positions. This is strongest and most evident in the most general conservation measure, phyloP. For the variant sets of the more conserved positions, the signal is predominately from very rare doubleton and tripleton variants using phyloP and doubleton variants using SiPhy-pi.

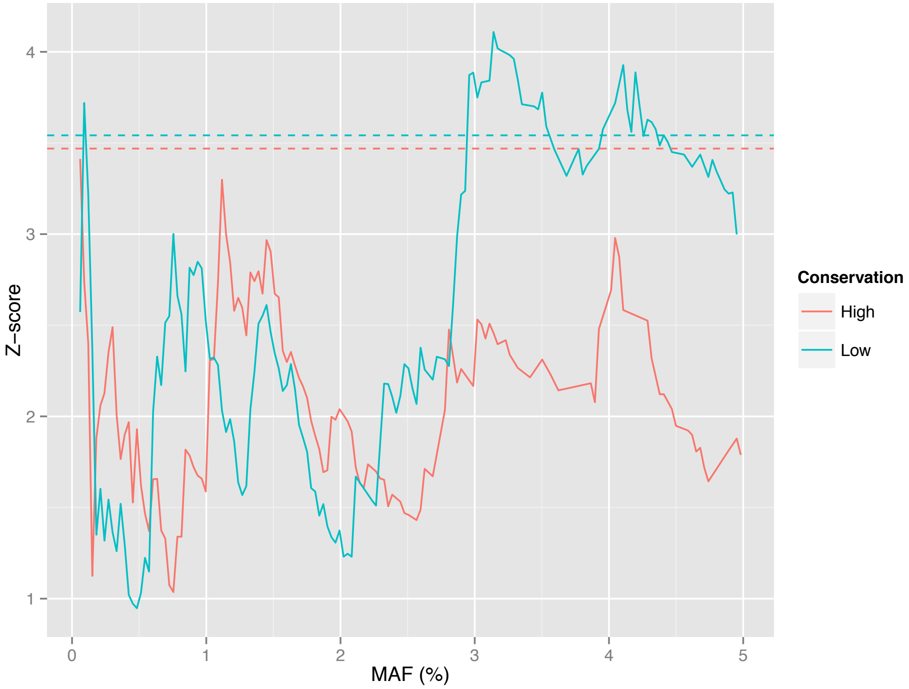

Figure 12: Z-scores for high and low phyloP scores

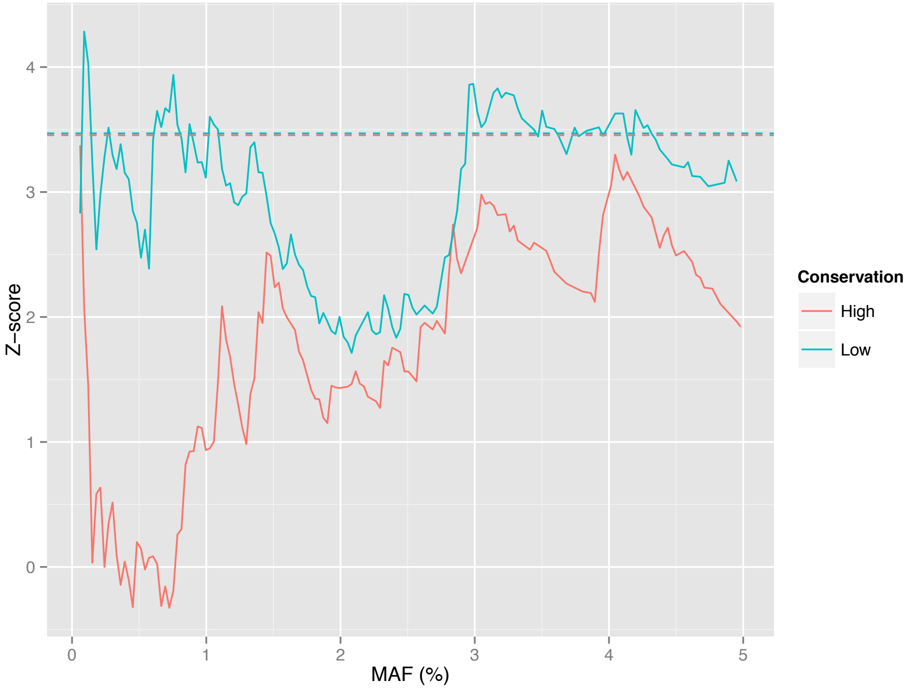

Figure 13: Z-scores for high and low SiPhy-pi scores

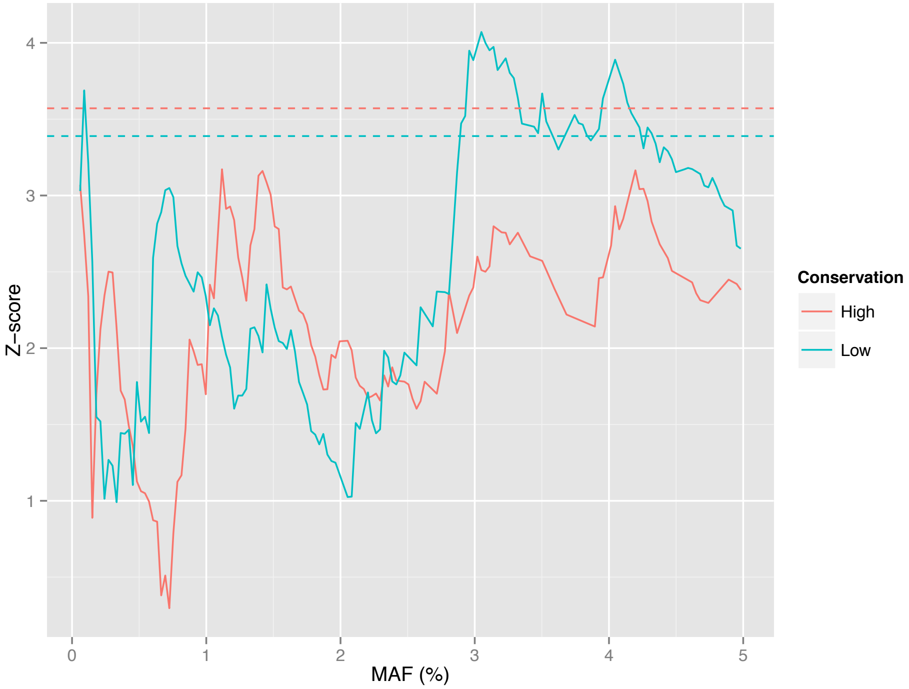

Figure 14: Z-scores for high and low phastCons scores

Narrowing the analysis further to consider only variants in the positions defined as constrained in the 29 mammal comparison (Lindblad-Toh et al., 2011), which includes all coding sequences, we detected a near significant difference ($q = 0.205$) in the distribution of alleles between cases and controls for ≤ 4.20% MAF (Figure 15). This result is weak support for excess unequal case/control allele distribution from much more common variation contributes to the signal from constrained regions.

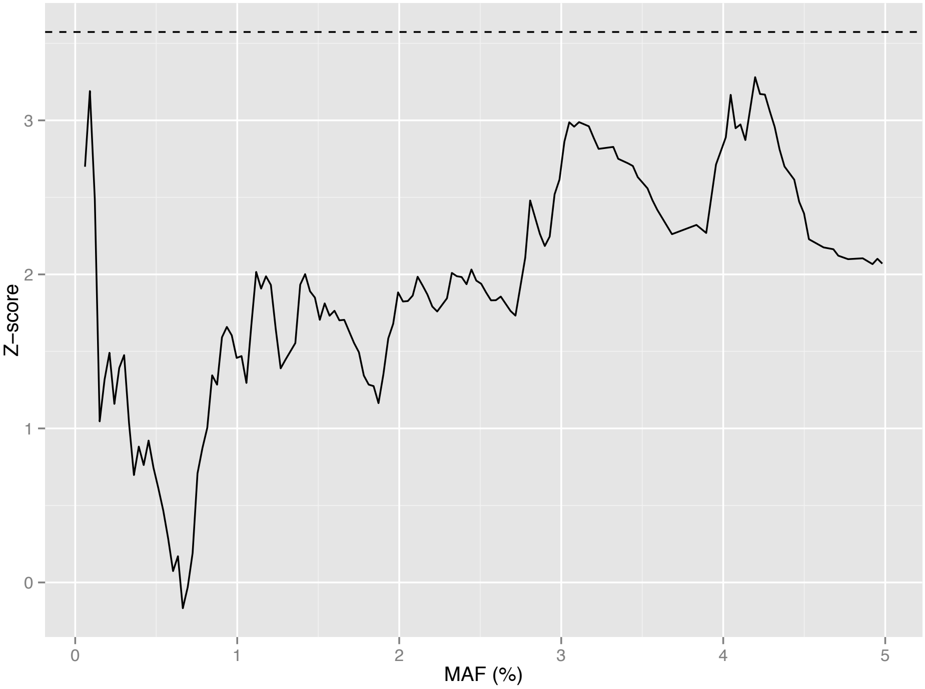

Figure 15: Z-scores for all constrained, coding and UTR variations

High Impact and High Conservation

We also tested frameshift, splice site, nonsense variants in addition to missense variants predicted to be damaging by Condel (Gonzalez-Perez & Lopez-Bigas, 2011) as a high impact variation group. The test was not significant possibly due to the low number of variants available to test with only 8 doubleton variants contributing to its peak Z-score. We pursued a different strategy to select potentially high impact variation by testing coding and non-coding variants with a base position phyloP score above 2, which indicates a less than 1% chance that the base would appear as conserved by chance. This strategy allows comparison of coding and non-coding variants at equally high levels of conservation. The distribution of variant phyloP scores is shown in Figure 16, with line indicating a phyloP score of 2. We observe more variants at these highly conserved positions in non-coding regions (390) than in coding regions or UTRs

(193). For all variants in this category we observed a significant signal ≤ 1.12% MAF (*P* = 0.050, *q* = 0.156). When splitting these variants into coding and non-coding groups we observed a significant signal at ≤ 1.12% MAF (*P* = 0.019, *q* = 0.111) and nonsignificant peak signal (*q* = 0.204) at 0.06% MAF respectively (Figure 17). The high conservation coding test only included 15 doubleton variants contributing to its peak Z-score so the test is relatively underpowered. While we observed that rare variation contributes the predominant signal for the more conserved half of non-coding variation, the non-coding variants in the most highly conserved positions yield a peak signal at a higher MAF. The higher MAF of the C-alpha signal peak for the most highly conserved non-coding positions is not consistent with the very rare peak for the more conserved half of non-coding positions. An explanation for this may be that the highly conserved set only represents a small subset of the more conserved half of non-coding positions (151 variants out of 1039) and therefore is less stable of a signal. We find the non-coding variants at highly conserved positions have more significant differences in allelic case/control distribution at higher MAFs than in coding variants at equivalently conserved positions. It is possible that the conservation of these positions in non-coding sequences does not reflect the same level of importance as in coding sequences, allowing risk-altering variants for schizophrenia to be maintained in the population at a higher MAF and lower effect size at these sites.

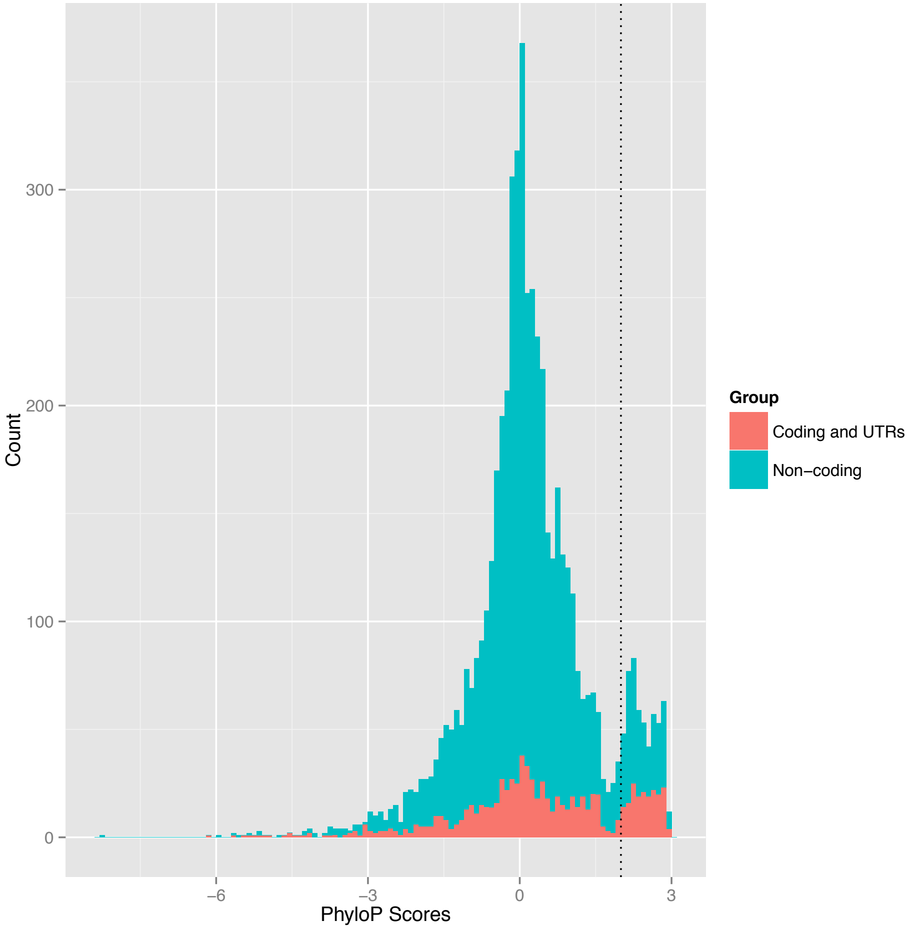

Figure 16: PhyloP base scores for variation in coding regions and UTRs and variation in non-coding regions. The dotted vertical line marks a phyloP score of 2. While the proportion of variation in coding regions and UTRs at bases phyloP > 2 (0.194) is higher than the proportion found in non-coding regions (0.070), the number of non-coding variants at bases with phyloP > 2 (390) is greater than the number found in coding regions and UTRs (193).

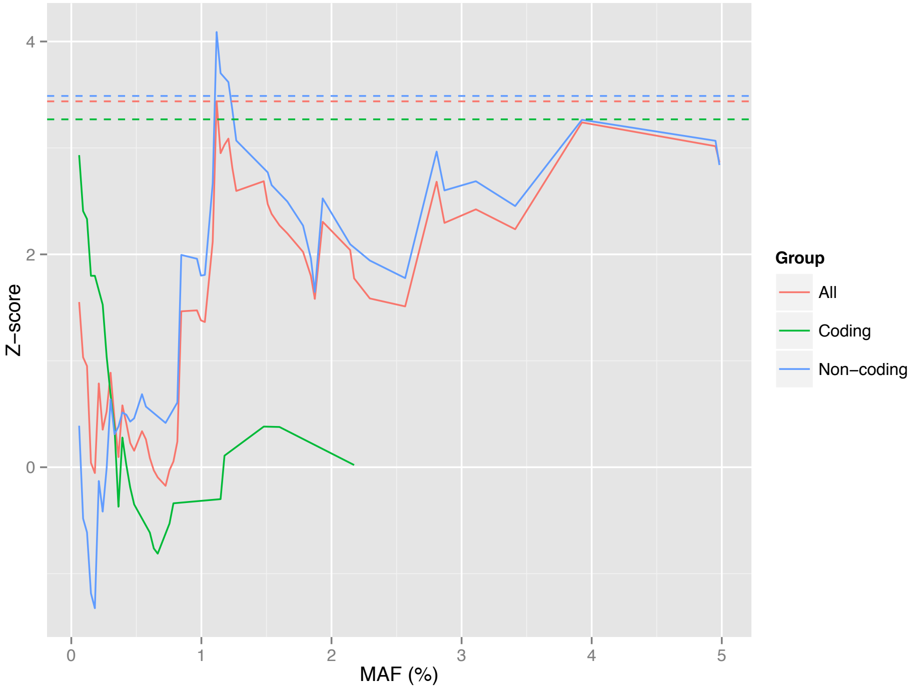

Figure 17: Z-scores for all highly conserved (phyloP > 2) also showing the split into coding and non-coding variation.

ENCODE Regions

We also assigned the variants to DNase I hypersensitive sites, TFBS, and histone modification sites found to be active in neuronal and glial cells in the ENCODE project. Details on how we defined ENCODE test regions are in the methods section.

DNase I hypersensitive sites had a significant excess unequal case/control allele distribution (≤ 4.20% MAF, $P = 0.031$, $q = 0.127$) but DNase I hypersensitive regions in glial cells did not (Figure 18). The excess of unequal distribution of case/control alleles was predominantly in more common variants for the DNase I neuronal variant set.

Transcription factor binding sites had a significant excess unequal case/control allele distribution for neuronal cell regions (≤ 2.87% MAF, $P = 0.015$ $q = 0.111$) (Figure 19) but did not for glial cell regions. For variants in the neuronal cell TFBS the excess in unequal distribution of case/control alleles was at a lower, but still relatively common, MAF than in DNase I hypersensitive sites.

We observed significant excess unequal case/control allele distribution for histone modification sites observed in neuronal cells (≤ 0.09% MAF, $P = 0.048$, $q = 0.153$) and glial cells (≤ 2.90% MAF, $P = 0.022$, $q = 0.111$) (Figure 20). The signal of excess unequal distribution of case/control alleles is very different in neuronal histone modification sites compared to glial histone modification sites, with the former being from rare variants (case and control, only doubletons and tripletons). The results of the tested ENCODE regions require careful interpretation. We include methylated histone markers that activate transcription, H3K4me1, H3K4me3, and H3K36me3, and repress transcription, H3K27me3. For acetylation, we include H3K27ac, which increases openness of chromatin. C-alpha testing is bidirectional, allowing for detection of bidirectional signals as may be seen in a combination of transcription activating and repressing histone modification sites. In histone modification sites from glial cell lines we observe more common variation in the ≤ 2.90% MAF range contributing to excess unequal distribution of case/control alleles, but for neuronal cell line sites this signal comes from very rare variation ≤ 0.09% MAF. For our geneset, variants in neuronal histone modification sites may be under heavier selection than variants in glial histone modification sites.

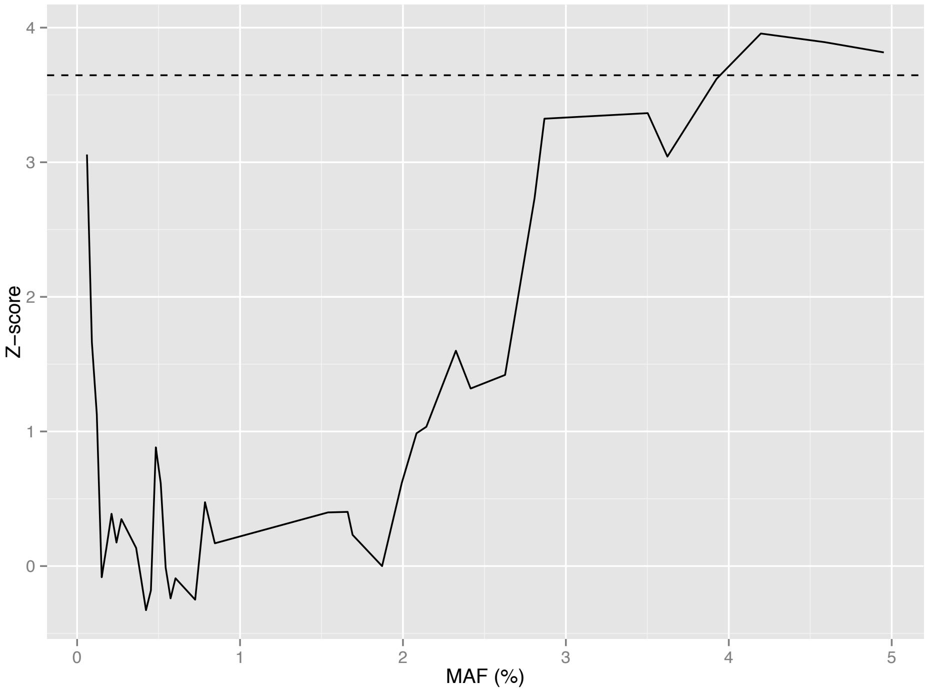

Figure 18: Z-scores for DNase I hypersensitive sites from neuronal cell lines

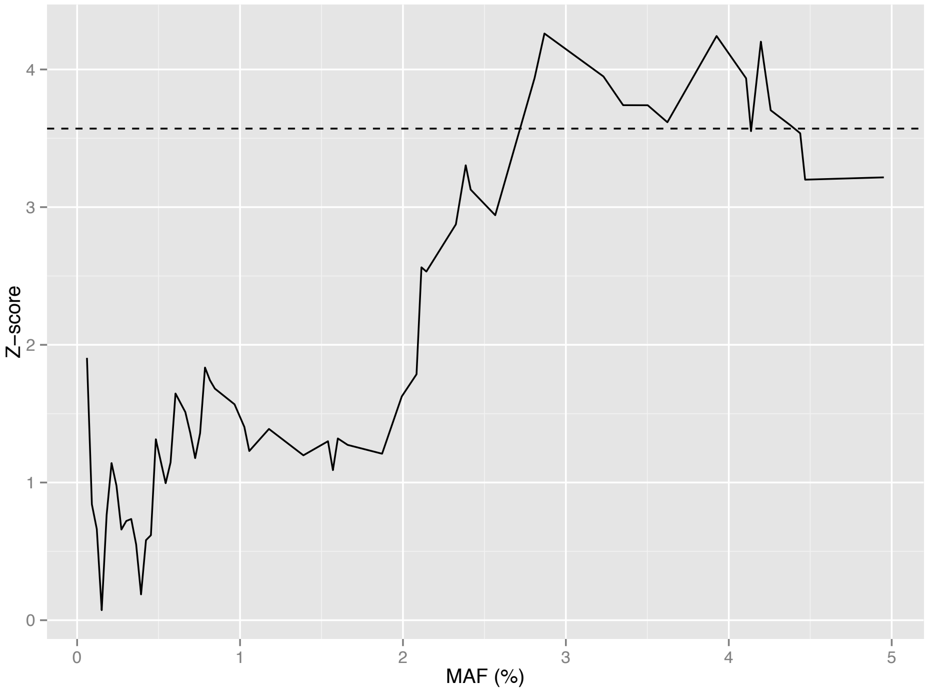

Figure 19: Z-scores for transcription factor binding sites from neuronal cell lines

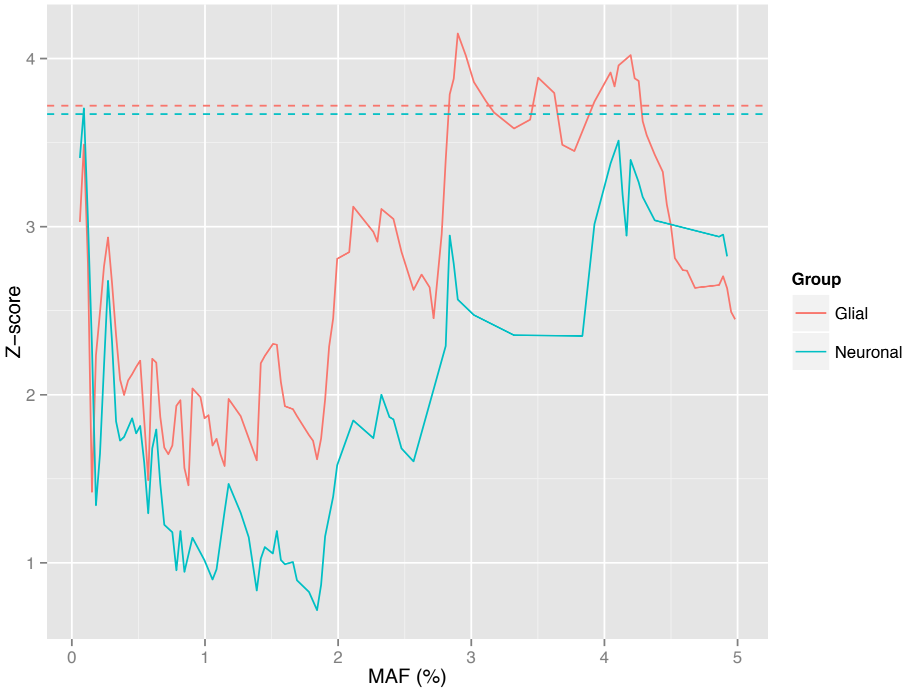

Figure 20: Z-scores for histone modification sites from neuronal and glial cells

Locus Tests

We further examined each of the 9 loci separately. We observed significant results for three individual loci, *ANK3*, a joint schizophrenia and bipolar risk locus that encodes the voltage-gated sodium channel associated Ankyrin G, the miR-137 regulated transcription factor *TCF4*, and complement control related gene *CSMD1*. *ANK3* (Figure 21) had the lowest allele frequency threshold for its C-alpha signal (≤ 0.09% MAF, *P* = 0.062, *q* = 0.168). The peak C-alpha signal for *TCF4* (Figure 22) was significant at ≤ 0.30% MAF (*P* = 0.023, *q* = 0.111) and the peak C-alpha signal for *CSMD1* (Figure 23) was significant at ≤ 1.45% MAF (*P* = 0.040, *q* = 0.142). *CSMD1* has an interesting property in that it consists primarily of two repeating functional domains, the CUB domain and the Sushi domain. We tested the variants in these domains separately and observed a significant excess unequal distribution of case/control alleles in the CUB domains (≤ 1.39% MAF, *P* = 0.005, *q* = 0.111) but not the Sushi domains. The CUB domains of *CSMD1* had the most significant *P*-value from the study. The Sushi domains had relatively few variants (28) in the test set which may have contributed to the low signal. *CSMD1* is an interesting gene because of its role in cancer and inflammation. Expression of CSMD1 mRNA cloned in rats is primarily found in developing CNS and epithelial tissues(Kraus et al., 2006). The CSMD1 protein was detected in the neuronal growth cone in developing fetal rat brains.

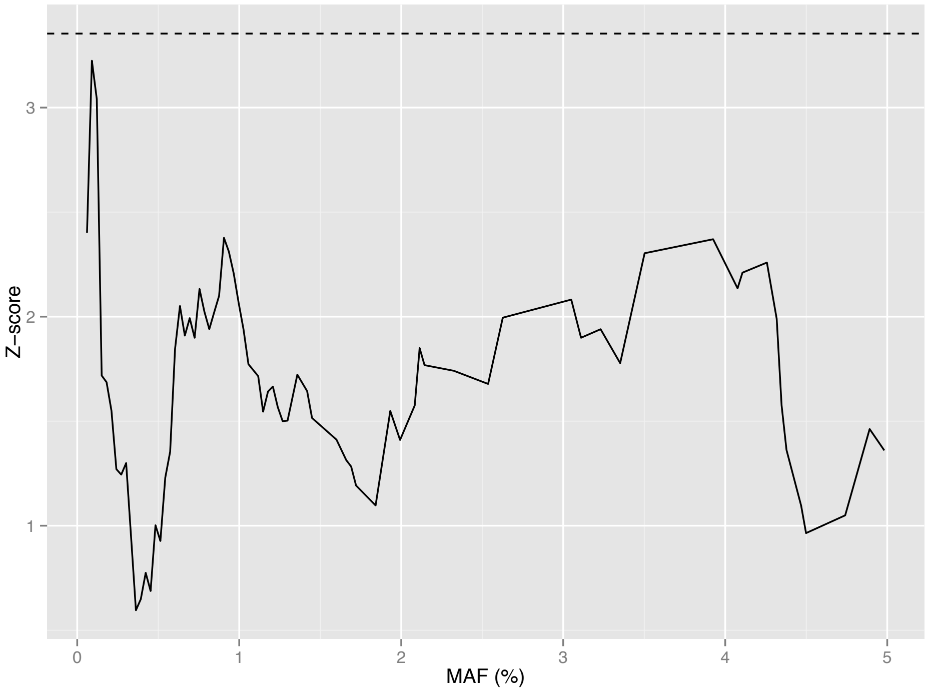

Figure 21: Z-scores for *ANK3*

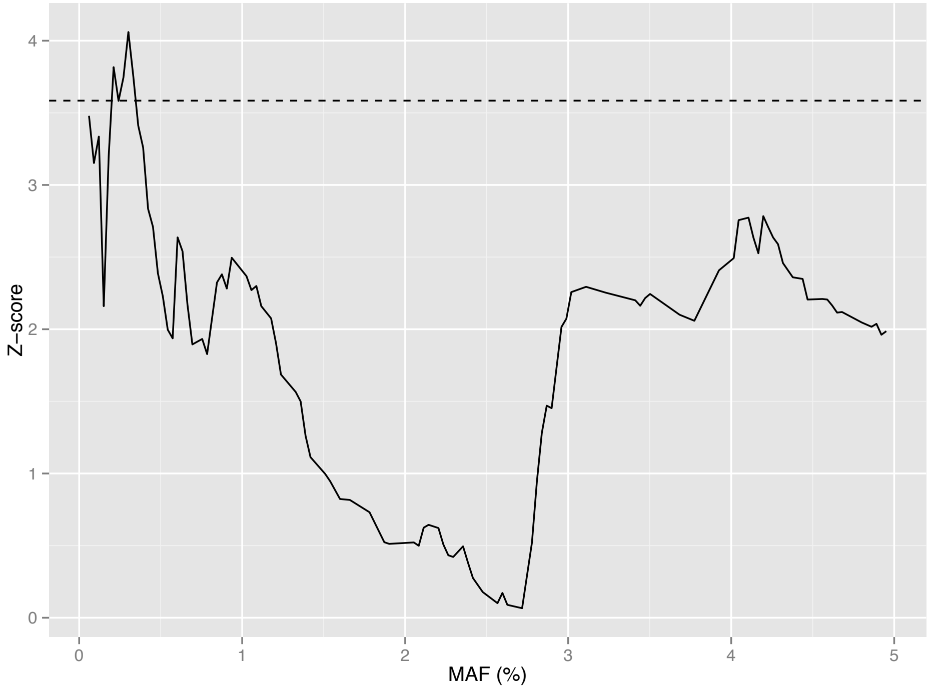

Figure 22: Z-scores for *TCF4*

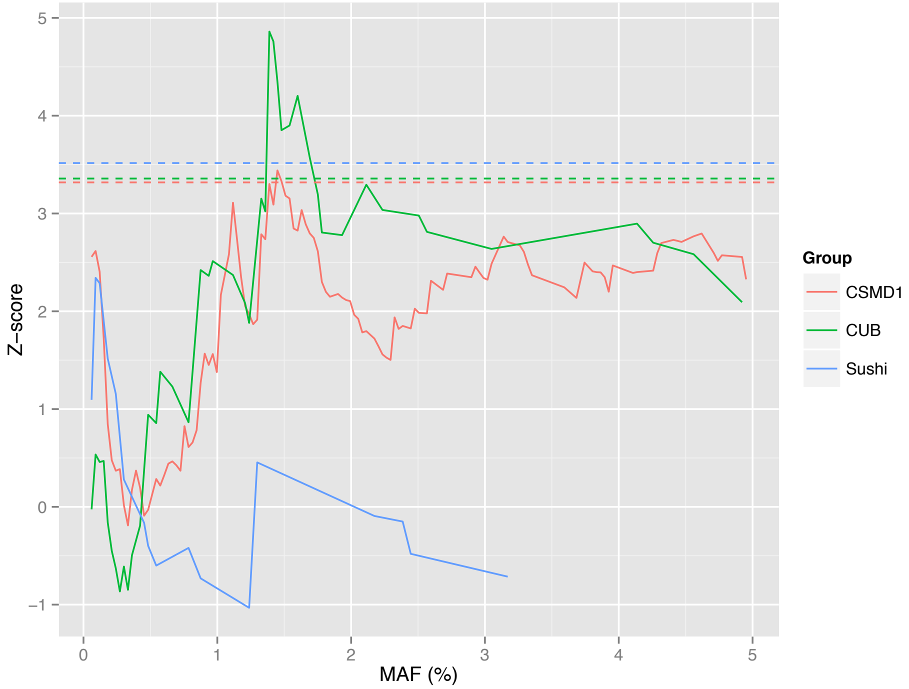

Figure 23: Z-score for *CSMD1* and its functional domains, CUB and Sushi

Discussion

Summary

Our results show that rare variation influencing schizophrenia risk is not limited to the coding exons and UTRs included in current exome sequencing studies. The high levels of conservation across species in these select non-coding positions increase the chance that changes to the sequence affect gene function and disease risk. It is significant that these signals were also observed in regions selected based on common variant associations from GWAS. As in IBD (Rivas et al., 2011) we observe rare variation in the loci selected through common variation association, independently supporting the involvement of these loci in schizophrenia. More loci implicated in recent GWAS (Ripke et al., 2013) (S. W. G. o. t. P. G. Consortium, 2014) should be explored for causal rare variation to help elucidate the mechanism by which these loci affect disease and to assess the additional impact of rare variation.

Rare Non-coding Variation in Disease

In our study 84% of the variation < 0.1% MAF range contributing to the signal was in non-coding regions and the signal was maintained when coding and non-coding variants were tested separately. The addition of non-coding sequence could greatly improve the quality and amount of information obtained from sequencing studies of all diseases. Exome studies could be expanded for targeted deep sequencing in non-coding functional and conserved sequences beyond the exome, until whole genome deep sequencing for large sample sizes is cost-effective.

The observation of rare non-coding variation is not unique to our study. Sequencing of the *IL4* locus in 72 African American (AA) asthma cases and 70 AA controls revealed an excess of private non-coding rare variants in cases (Haller, Torgerson, Ober, & Thompson, 2009). The authors suggest these rare variants cannot be reliably imputed and therefore sequencing of non-coding regions is important for identifying rare genetic variants contributing to disease. A study of 100 genes implicated in asthma using 450 cases and 515 controls found *IL12RB1* to be a susceptibility locus with predominately non-coding variant signals in both AA and European Americans (EA) (Torgerson et al., 2012). Limited work has been done on non-coding rare variation is schizophrenia. A study sequencing 27 kb from six schizophrenia candidate genes, *AKT1, BDNF, DRD3, DTNBP1,* and *NRG1,* found an excess of rare non-coding variants in 37 cases compared to 25 controls (Winantea et al., 2006). The researchers determined the enrichment by calculating Tajima's *D*-value for the cases and the controls. A small Tajima's *D*-value in cases compared to controls indicated an excess of rare variants in the case sample.

Our work is the first large study discovering influence of non-coding rare variants in several loci for schizophrenia. The evidence from asthma, an autoimmune disorder, supports our finding in schizophrenia, a disease for which the immune system is also implicated.

Locus Tests

The loci that we found significantly associated with schizophrenia were *ANK3*, *TCF4* and *CSMD1*. *ANK3* was first identified as a bipolar disorder locus in a GWAS of 4,387 cases and 6,209 controls (Ferreira et al., 2008). The PGC-1 study was the first to observe a joint association for combined bipolar disorder cases and schizophrenia cases compared to controls (Ripke S, 2011). In a study of 516 Han Chinese schizophrenia cases and 400 controls, *ANK3* was implicated as an independent schizophrenia risk locus (Yuan et al., 2012). More recently, coding variation in *ANK3* has been associated with autism (Bi et al., 2012).

Several studies of *ANK3* and its protein product Ankyrin-G and their effect on neurodevelopment and cognitive performance have been published. Ankyrin-G is a scaffolding protein that localizes to the axon initial segment and nodes of Ranvier of neurons (Kordeli, Lambert, & Bennett, 1995). Knockdown of *ANK3* was found to increase β-catenin in the nucleus, causing an increase in neural progenitor proliferation (Durak et al., 2014). The mechanism suggested for the increase in β-catenin is that functional ankyrin-G interacts with E-cadherin and Wnt. One study found damaging mutations in the *ANK3* gene in patients with severe cognitive deficits (Iqbal et al., 2013). The authors found all isoforms of *ANK3* disrupted by a balanced translocation in one patient with autism, attention deficit hyperactivity disorder (ADHD), sleeping problems, and borderline intelligence. They also found a frameshift mutation in the longest isoform segregating in a family with moderate intellectual disability. Reasoning that memory deficits are common to disorders such as intellectual disability, autism and ADHD, the authors disrupted *Ank2*, the closest homolog to human *ANK3* in *Drosophila*. They observed a significant reduction of short-term memory in the flies. The authors observed normal learning and other behaviors, and concluded that *Ank2* was crucial for properly functioning memory in *Drosophila*. A study of cognitive deficits in 173 patients with first episode psychosis found association of a common *ANK3* allele (allele G of rs1938526) with lower cognitive performance, verbal memory, working memory and attention (Cassidy et al., 2014). The authors also observed an association between this allele and cortical thinning. An additional study of 163 patients with first-episode schizophrenia and 42 controls found association of a common *ANK3* allele (allele T of rs10994336) with lower accuracy and longer reaction time in a 2-back test, where the participant must log items flashed on a screen in addition to the 2 items before (Zhang et al., 2014). Our study is the first to implicate rare variation in *ANK3* as a risk factor for schizophrenia.

TCF4 was first identified as a schizophrenia risk factor in an early GWAS (Stefansson et al., 2009). It is a basic Helix-Loop-Helix transcription factor found through knockdown experiments to be involved in cell survival, epithelial to mesenchymal transition, and neurodevelopment (Forrest, Waite, Martin-Rendon, & Blake, 2013). It has been identified as a susceptibility locus for multiple disorders, with common variants identified impacting Fuch's endothelial corneal dystrophy, primary sclerosing cholangitis and, as previously mentioned, schizophrenia (Forrest, Hill, Quantock, Martin-Rendon, & Blake, 2014). Our study is the first to observe rare variation in *TCF4* associated with schizophrenia. Private frameshift, nonsense, splice site and missense variants and deletions (CNVs partially covering the gene or completely covering the gene) of *TCF4* have been found to cause Pitt-Hopkins syndrome (Peippo & Ignatius, 2012), a syndrome characterized by intellectual disability and developmental delay. A study of smokers and never-smokers taken randomly from the German population found that smokers with the rs9960767 risk allele had reduced sensory gating, measured by P50 suppression (Quednow et al., 2012). In patients with first episode psychosis, the rs9960767 risk allele has also been associated with lower performance in the Reasoning/Problem-Solving domain of the WAIS-III and Trail Making Test B (Albanna et al., 2014; Reitan, 1992). There is also evidence for correlation of the risk allele in schizophrenia cases with improved cognitive performance. In one study schizophrenia cases with the rs9960767 risk allele performed better on the Rey Auditory Verbal Learning Test (Helmstaedter, 2001), which measures verbal declarative memory (Lennertz et al., 2011). In a Han Chinese sample, schizophrenia cases homozygous for the rs2958182 risk allele performed better on cognitive tasks compared to cases with the non-risk allele (Zhu et al., 2013). Han Chinese controls homozygous for the rs2958182 risk allele performed worse on cognitive tasks (Zhu et al., 2013). The literature on cognition, schizophrenia, and common variant *TCF4* loci has been mixed, with some positive effects and some negative effects correlated with the risk alleles for schizophrenia patients.

Our significant *TCF4* C-alpha test shows that rare variation has a bidirectional effect. Some variants are protective, decreasing schizophrenia risk, and some variants are damaging, increasing schizophrenia risk. This is supported by the biology of basic Helix-Loop-Helix transcription factors, which may act as transcriptional repressors or activators (Quednow, Brzozka, & Rossner, 2014). In Forrest et al., 2013, *TCF4* knockdown caused increased expression for 494 genes and decreased expression for 710 genes (Forrest et al., 2013). With such a great number of upregulated and downregulated genes, damaging variants in *TCF4* seem likely to perturb regulation of a large number of individual loci with potentially significant impact on neurodevelopment and later function. Such focal changes in a single transcription factor gene could potentially mimic and interact with the effects of numerous variants from other loci on schizophrenia risk, given the strong support for a highly polygenic and distributed genetic structure.

A study of 512 schizophrenia cases and 270 controls searched for causal rare variation in a small, ultraconserved non-coding 227 bp region of *TCF4* with evidence for enhancer activity (UC435) (Gonzalez-Penas et al., 2014). The researchers did not find any variants in UC435. In our 840 case and 816 control sample we found only one singleton in UC435, in a case sample at chr18:53089932 (novel A allele, G reference). Our *TCF4* locus test included many other non-coding regions and our results suggest an aggregation of low frequency *TCF4* variation impacting schizophrenia risk.

CSMD1 is a transmembrane protein, and contains repeating and alternating CUB and Sushi (complement control protein or CCP) domains. Sushi domains are involved in complement inhibition for the classical and lectin pathways of the complement system. The complement system is the immune system's first line defense against foreign antigens. The 15 Sushi tandem repeat inhibits complement deposition on eukaryotic cell surfaces (Escudero-Esparza, Kalchishkova, Kurbasic, Jiang, & Blom, 2013). The CUB domain is a 110-residue domain structure, composed of a β-sandwich fold. Many proteins have CUB domains with a Ca^{2+} binding site, including CSMD1 (Gaboriaud et al., 2011).

CSMD1 is a tumor suppressor gene implicated in multiple cancers, including squamous cell carcinoma, colorectal cancer, melanoma, lung, head and neck, and breast cancer. One study showed copy number losses in *CSMD1* in head and neck squamous cell carcinoma, cutaneous squamous cell carcinoma, cutaneous basal cell carcinoma, lung and breast cancers (Ma et al., 2009). Another study showed that 87% of squamous cell carcinoma cell lines had increased methylation upstream of the *CSMD1* transcription start site, associated with gene silencing, compared to normal upper aerodigestive epithelial cells (Richter, Tong, & Scholnick, 2005). A study has also found low levels of *CSMD1* expression in melanoma (Tang, Wang, Guo, Han, & Wang, 2012). Somatic nonsynonmous mutations have been observed in late stage colorectal cancer (Farrell et al., 2008) and in colorectal cancer diagnosed at an early age (Shull et al., 2013), indicating a role for *CSMD1* the development of aggressive, metastatic disease.

In addition to schizophrenia and cancer, *CSMD1* has been implicated in bipolar disorder through GWAS (Sklar et al., 2008) (Baum et al., 2008) (W. Xu et al., 2014) and autism through CNVs (Glancy et al., 2009) and exome sequencing in families (Cukier et al., 2014). Involvement in neuropsychiatric disorders is supported by the observation that the *CSMD1* gene cloned in rats was expressed primarily in the developing CNS and epithelial tissues, and its protein was enriched in fetal rat nerve growth cones (Kraus et al., 2006).

The *CSMD1* risk variant implicated in schizophrenia (rs10503253) has been studied for impact on several neurocognitive effects. A study of 1,149 healthy Greek Caucasian males found additive effects of the risk allele for poorer general cognitive ability (IQ), strategy formation, spatial and visual working memory, set shifting, target detection, and planning for problem solving (Koiliari et al., 2014). Similar results were found in a study of Irish and German schizophrenia cases and controls with the risk allele being associated with poorer IQ and memory function (Donohoe et al., 2013). With evidence for the role of *CSMD1* in neurodevelopment, and its impact on cognitive abilities, investigators used MRI and fMRI to measure the effect of the rs10503253 genotype on grey and white matter volume and activity during a spatial working memory task (Rose et al., 2013). The authors found that the risk allele was significantly associated with reduced cortical activations in the right middle occipital gyrus, a region involved in spatial working memory. No structural differences in brain volume were found based on genotype.

CSMD1 rare variation has not been strongly implicated in neuropsychiatric disorders. There is no evidence of *CSMD1* rare variation in bipolar disorder and there is only weak evidence of *CSMD1* rare variation in autism. *CSMD1* variants were observed in two separate autism families in a 40 family study (Cukier et al., 2014). Also, unlike *TCF4*'s association with Pitt-Hopkins syndrome, *CSMD1* rare variation is not associated with any syndromes. The only strong evidence for rare variation in *CSMD1* impacting disease has been somatic mutations in cancer.

The *CSMD1* signal we observed differs from both *ANK3* and *TCF4* in that it is a less rare signal, with a signal maximizing when alleles up to 1.45% MAF are included in the test. This result is consistent with the lack of strong association of other neuropsychiatric disorders with rare variation in *CSMD1*. Even though the variation in *CSMD1* driving the signal in our study is not very rare, our results support association of the gene with schizophrenia at a lower MAF than PGC-1 (≤1.45% compared to 19%). Our observation of a signal coming specifically from the CUB domain is a first for schizophrenia. No functional work has been done on the CUB domain or *CSMD1* in schizophrenia, but based on the literature and our results it may be a promising candidate gene and functional domain.

Functional Class Tests

In our study, we found variations within ENCODE functional elements from neuronal and glial cell lines to be associated with schizophrenia. We tested functional elements active in neuronal cell lines because neurons are the primary cells involved in cognition and central nervous system (CNS) function. Our analysis of functional elements active in neuronal cell lines shows that variation in transcription factor binding sites, DNase I hypersensitive sites, and especially rare variation from histone modification sites influences schizophrenia risk.

We tested functional elements active in glial cell lines because glia support neurons and the three major classes of glial cells, oligodendrocytes, microglia and astrocytes, have been implicated in schizophrenia (Goudriaan et al., 2013) (Frick, Williams, & Pittenger, 2013). Dysfunction in oligodendrocytes, glial cells that produce the myelin in the nervous system, has been shown to impact synaptic function and white matter integrity in the brain (Takahashi, Sakurai, Davis, & Buxbaum, 2011) and an analysis of common variant *p*-values from the PGC-1 schizophrenia study (Ripke S, 2011) found the oligodendrocytic gene set pathway associated with schizophrenia (Duncan et al., 2014). Variation in oligodendrocyte specific genes involved in myelin production affects cognitive performance and the integrity of white matter tracts (Voineskos et al., 2013). Microglia are macrophages in the CNS involved in innate immunity. There is evidence that microglia play a role in schizophrenia through neuroinflammation (Monji et al., 2013). Astrocytes support the nervous system in several ways, including by providing structural and nutritional support for neurons and aiding synaptic function (Takahashi & Sakurai, 2013). One study observed a significant decrease in astrocyte density in the cingulate gray matter, cingulate white matter and midline of the corpus callosum in schizophrenia patients compared to controls (Williams et al., 2013). Loss of function in astrocytic receptors and gap junctions of astrocytes may contribute to cognitive impairment in schizophrenia. (Mitterauer, 2011).

The data we used in our glial DNase I hypersensitive sites, transcription factor binding sites, and histone modification sites analyses were all from astrocyte-derived cell lines: Gliobla, HA-h, NH-A, and U87 (Project, 2014). At the time of this study, ENCODE data for microglia and oligodendrocytes were unavailable. We also found damaging and protective variation (C-alpha test) in histone modification sites active in the cell lines. Literature supports the effect of epigenetic modulation of histone deacetylase inhibitors on schizophrenia (Cha, Kudlow, Baskaran, Mansur, & McIntyre, 2014) and we included acetylated histone regions in our glial histone modification test. H3K4me3, a trimethylation of histone H3 at lysine 4 we included in our glial and neuronal histone modification test, is an active mark for transcription and is implicated in increased expression of synapsin genes in bipolar disorder and major depression (Cruceanu et al., 2013). The signal in histone modification sites is very different in neuronal cell line regions compared to glial cell line regions. The signal from neuronal cell line regions is driven by much rarer alleles (≤ 0.09% MAF) than the signal from glial cell line regions (≤ 2.90% MAF), possibly indicating that variation in neuronal cell line regions is under heavier selection pressure.

Limitations

The results of this study must be interpreted with several limitations. We performed sequencing in pools of 24 cases or 24 controls. This method is efficient, but did not allow us to incorporate LD into our analysis. We used permutation testing to compensate for the effects of LD on the test statistic. Pooled sequencing may also introduce some error in allele count calling. Our careful consideration of concentration while pooling, careful quality control, and improved probabilistic allele count calling mitigates this risk. Also, we were not able to include every constrained site within our target region due to the 500kb limit using the specific capture approach we chose, but our target and bait design did include 84.5% of all constrained sites at 10% FDR across 29 mammals (Lindblad-Toh et al., 2011) within our target intervals.

Future Studies

Although pooled association tests can identify associated loci, future studies should use a much larger sample size, probably on the order of the largest GWAS size, to find single rare variant associations. As our data show, future sequencing studies should include non-coding regions because they may contain useful data. Lower capture and library preparation costs will make individual sample sequencing more affordable. Future studies should consider using individual sample sequencing to eliminate the limitations of pooled sequencing studies discussed above. The ultimate future study will use low cost, individual, whole genome sequencing.

There are many influences on the heritability of schizophrenia that remain unexplained and that may be uncovered by increasing sample size and amount of the genome sequenced. The discovery of individual rare variants associated with schizophrenia will promote a new series of functional studies to elucidate the mechanism by which the variants affect schizophrenia risk.

List of References

Albanna, A., Choudhry, Z., Harvey, P. O., Fathalli, F., Cassidy, C., Sengupta, S. M., . . . Joober, R. (2014). TCF4 gene polymorphism and cognitive performance in patients with first episode psychosis. *Schizophr Res, 152*(1), 124-129. doi: 10.1016/j.schres.2013.10.038

American Psychiatric Association, American Psychiatric Association, American Psychiatric Association D. S. M. Task Force. (2013). Diagnostic and statistical manual of mental disorders : DSM-5. from http://dsm.psychiatryonline.org/book.aspx?bookid=556

American Psychiatric Association, American Psychiatric Association, American Psychiatric Association Task Force on D. S. M. I. V. (2000). *Diagnostic and statistical manual of mental disorders : DSM-IV-TR*. Washington, DC: American Psychiatric Association.

Asan, Xu, Y., Jiang, H., Tyler-Smith, C., Xue, Y., Jiang, T., . . . Zhang, X. (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol, 12*(9), R95. doi: 10.1186/gb-2011-12-9-r95

Ascano, M., Jr., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., . . . Tuschl, T. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature, 492*(7429), 382-386. doi: 10.1038/nature11737

Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics, 26*(12), i318-324. doi: 10.1093/bioinformatics/btq214

Bansal, V., Tewhey, R., Leproust, E. M., & Schork, N. J. (2011). Efficient and cost effective population resequencing by pooling and in-solution hybridization. *PLoS One, 6*(3), e18353. doi: 10.1371/journal.pone.0018353

Bassett, A. S., Marshall, C. R., Lionel, A. C., Chow, E. W., & Scherer, S. W. (2008). Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. *Hum Mol Genet, 17*(24), 4045-4053. doi: 10.1093/hmg/ddn307

Baum, A. E., Akula, N., Cabanero, M., Cardona, I., Corona, W., Klemens, B., . . . McMahon, F. J. (2008). A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry, 13*(2), 197-207. doi: 10.1038/sj.mp.4002012

Bi, C., Wu, J., Jiang, T., Liu, Q., Cai, W., Yu, P., . . . Sun, Z. S. (2012). Mutations of ANK3 identified by exome sequencing are associated with autism susceptibility. *Hum Mutat, 33*(12), 1635-1638. doi: 10.1002/humu.22174

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., . . . Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res, 14*(4), 708-715. doi: 10.1101/gr.1933104

Brown, A. S. (2011). The environment and susceptibility to schizophrenia. *Prog Neurobiol, 93*(1), 23-58. doi: 10.1016/j.pneurobio.2010.09.003

Brown, A. S., & Derkits, E. J. (2010). Prenatal infection and schizophrenia: a review of epidemiologic and translational studies. *Am J Psychiatry, 167*(3), 261-280. doi: 10.1176/appi.ajp.2009.09030361

Cantor-Graae, E., & Selten, J. P. (2005). Schizophrenia and migration: a meta-analysis and review. *Am J Psychiatry, 162*(1), 12-24. doi: 10.1176/appi.ajp.162.1.12

Cassidy, C., Buchy, L., Bodnar, M., Dell'elce, J., Choudhry, Z., Fathalli, F., . . . Joober, R. (2014). Association of a risk allele of ANK3 with cognitive performance and cortical thickness in patients with first-episode psychosis. *J Psychiatry Neurosci, 39*(1), 31-39. doi: 10.1503/jpn.120242

Cha, D. S., Kudlow, P. A., Baskaran, A., Mansur, R. B., & McIntyre, R. S. (2014). Implications of epigenetic modulation for novel treatment approaches in patients with schizophrenia. *Neuropharmacology, 77*, 481-486. doi: 10.1016/j.neuropharm.2013.08.038

Chiesa, A., Lia, L., Han, C., Lee, S. J., Pae, C. U., & Serretti, A. (2013). Investigation of epistasis between DAOA and 5HTR1A variants on clinical outcomes in patients with schizophrenia. *Genet Test Mol Biomarkers, 17*(6), 504-507. doi: 10.1089/gtmb.2012.0484

Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca (2012). mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation Technical Report No. 597.

Chubb, J. E., Bradshaw, N. J., Soares, D. C., Porteous, D. J., & Millar, J. K. (2008). The DISC locus in psychiatric illness. *Mol Psychiatry, 13*(1), 36-64. doi: 10.1038/sj.mp.4002106

Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr., & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med, 354*(12), 1264-1272. doi: 10.1056/NEJMoa054013

Cohen, J. C., Pertsemlidis, A., Fahmi, S., Esmail, S., Vega, G. L., Grundy, S. M., & Hobbs, H. H. (2006). Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci U S A, 103*(6), 1810-1815. doi: 10.1073/pnas.0508483103

Consortium, Schizophrenia Working Group of the Psychiatric Genomics. (2014). Common Variant Association Meta-Analysis for Schizophrenia Identifies 108 Genomic Loci and Implicates Postsynaptic and Immune Processes. *in submission*.

Consortium, Wellcome Trust Case Control. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. (1476-4687 (Electronic)). doi: D - NLM: PMC2719288

D - NLM: UKMS4894

Cruceanu, C., Alda, M., Nagy, C., Freemantle, E., Rouleau, G. A., & Turecki, G. (2013). H3K4 tri-methylation in synapsin genes leads to different expression patterns in bipolar disorder and major depression. *Int J Neuropsychopharmacol, 16*(2), 289-299. doi: 10.1017/s1461145712000363

Cukier, H. N., Dueker, N. D., Slifer, S. H., Lee, J. M., Whitehead, P. L., Lalanne, E., . . . Pericak-Vance, M. A. (2014). Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Mol Autism, 5*(1), 1. doi: 10.1186/2040-2392-5-1

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y., Mele, A., Fraser, C. E., . . . Darnell, R. B. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell, 146*(2), 247-261. doi: 10.1016/j.cell.2011.06.013

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498. doi: 10.1038/ng.806

Donohoe, G., Walters, J., Hargreaves, A., Rose, E. J., Morris, D. W., Fahey, C., . . . Corvin, A. (2013). Neuropsychological effects of the CSMD1 genome-wide associated schizophrenia risk variant rs10503253. *Genes Brain Behav, 12*(2), 203-209. doi: 10.1111/gbb.12016

Duncan, L. E., Holmans, P. A., Lee, P. H., O'Dushlaine, C. T., Kirby, A. W., Smoller, J. W., . . . Cohen, B. M. (2014). Pathway analyses implicate glial cells in schizophrenia. *PLoS One, 9*(2), e89441. doi: 10.1371/journal.pone.0089441

Durak, O., de Anda, F. C., Singh, K. K., Leussis, M. P., Petryshen, T. L., Sklar, P., & Tsai, L. H. (2014). Ankyrin-G regulates neurogenesis and Wnt signaling by altering the subcellular localization of beta-catenin. *Mol Psychiatry*. doi: 10.1038/mp.2014.42

Elston, R. C., Namboodiri, K. K., Spence, M. A., & Rainer, J. D. (1978). A genetic study of schizophrenia pedigrees. II. One-locus hypotheses. *Neuropsychobiology, 4*(4), 193-206.

Escudero-Esparza, A., Kalchishkova, N., Kurbasic, E., Jiang, W. G., & Blom, A. M. (2013). The novel complement inhibitor human CUB and Sushi multiple domains 1 (CSMD1) protein promotes factor I-mediated degradation of C4b and C3b and inhibits the membrane attack complex assembly. *FASEB J, 27*(12), 5083-5093. doi: 10.1096/fj.13-230706

Farrell, C., Crimm, H., Meeh, P., Croshaw, R., Barbar, T., Vandersteenhoven, J. J., . . . Buckhaults, P. (2008). Somatic mutations to CSMD1 in colorectal adenocarcinomas. *Cancer Biol Ther, 7*(4), 609-613.

Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L., . . . Craddock, N. (2008). Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet, 40*(9), 1056-1058. doi: 10.1038/ng.209

Forrest, M. P., Hill, M. J., Quantock, A. J., Martin-Rendon, E., & Blake, D. J. (2014). The emerging roles of TCF4 in disease and development. *Trends Mol Med, 20*(6), 322-331. doi: 10.1016/j.molmed.2014.01.010

Forrest, M. P., Waite, A. J., Martin-Rendon, E., & Blake, D. J. (2013). Knockdown of human TCF4 affects multiple signaling pathways involved in cell survival, epithelial to mesenchymal transition and neuronal differentiation. *PLoS One, 8*(8), e73169. doi: 10.1371/journal.pone.0073169

Frick, L. R., Williams, K., & Pittenger, C. (2013). Microglial dysregulation in psychiatric disease. *Clin Dev Immunol, 2013*, 608654. doi: 10.1155/2013/608654

Fromer, M., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., . . . O'Donovan, M. C. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature, 506*(7487), 179-184. doi: 10.1038/nature12929

Futschik, A., & Schlotterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics, 186*(1), 207-218. doi: 10.1534/genetics.110.114397

Gaboriaud, C., Gregory-Pauron, L., Teillet, F., Thielens, N. M., Bally, I., & Arlaud, G. J. (2011). Structure and properties of the Ca(2+)-binding CUB domain, a widespread ligand-recognition unit involved in major biological functions. *Biochem J, 439*(2), 185-193. doi: 10.1042/bj20111027

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics, 25*(12), i54-62. doi: 10.1093/bioinformatics/btp190

Girard, S. L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., . . . Rouleau, G. A. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet, 43*(9), 860-863. doi: 10.1038/ng.886

Glancy, M., Barnicoat, A., Vijeratnam, R., de Souza, S., Gilmore, J., Huang, S., . . . Barber, J. C. (2009). Transmitted duplication of 8p23.1-8p23.2 associated with speech delay, autism and learning difficulties. *Eur J Hum Genet, 17*(1), 37-43. doi: 10.1038/ejhg.2008.133

Gonzalez-Penas, J., Arrojo, M., Paramo, M., Paz, E., Agra, S., Ramos-Rios, R., . . . Costas, J. (2014). Absence of low frequency variants associated with schizophrenia at the ultraconserved non-coding region of TCF4. *Psychiatry Res, 215*(1), 255-257. doi: 10.1016/j.psychres.2013.10.008

Gonzalez-Perez, A., & Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet, 88*(4), 440-449. doi: 10.1016/j.ajhg.2011.03.004

Goudriaan, A., de Leeuw, C., Ripke, S., Hultman, C. M., Sklar, P., Sullivan, P. F., . . . Verheijen, M. H. (2013). Specific Glial Functions Contribute to Schizophrenia Susceptibility. *Schizophr Bull*. doi: 10.1093/schbul/sbt109

Haller, G., Torgerson, D. G., Ober, C., & Thompson, E. E. (2009). Sequencing the IL4 locus in African Americans implicates rare noncoding variants in asthma susceptibility. *J Allergy Clin Immunol, 124*(6), 1204-1209 e1209. doi: 10.1016/j.jaci.2009.09.013

Helmstaedter, C.; Lendt, M.; Lux, S. (2001). *VLMT Verbaler Lern-Und Merkfähigkeitstest*. Göttingen, Germany: Beltz.

Hill, M. N. (2014). Clearing the smoke: what do we know about adolescent cannabis use and schizophrenia? *J Psychiatry Neurosci, 39*(2), 75-77.

Ingason, A., Rujescu, D., Cichon, S., Sigurdsson, E., Sigmundsson, T., Pietilainen, O. P., . . . St Clair, D. M. (2011). Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry, 16*(1), 17-25. doi: 10.1038/mp.2009.101

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., . . . Wigler, M. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron, 74*(2), 285-299. doi: 10.1016/j.neuron.2012.04.009

Iqbal, Z., Vandeweyer, G., van der Voet, M., Waryah, A. M., Zahoor, M. Y., Besseling, J. A., . . . Rooms, L. (2013). Homozygous and heterozygous disruptions of ANK3: at the crossroads of neurodevelopmental and psychiatric disorders. *Hum Mol Genet, 22*(10), 1960-1970. doi: 10.1093/hmg/ddt043

Ishizuka, K., Paek, M., Kamiya, A., & Sawa, A. (2006). A review of Disrupted-In-Schizophrenia-1 (DISC1): neurodevelopment, cognition, and mental conditions. *Biol Psychiatry, 59*(12), 1189-1197. doi: 10.1016/j.biopsych.2006.03.065

Iyegbe, C., Campbell, D., Butler, A., Ajnakina, O., & Sham, P. (2014). The emerging molecular architecture of schizophrenia, polygenic risk scores and the clinical implications for GxE research. *Soc Psychiatry Psychiatr Epidemiol, 49*(2), 169-182. doi: 10.1007/s00127-014-0823-2

Kendler, K. S., & Diehl, S. R. (1993). The genetics of schizophrenia: a current, genetic-epidemiologic perspective. *Schizophr Bull, 19*(2), 261-285.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. doi: 10.1101/gr.229102. Article published online before print in May 2002

Khandaker, G. M., Zimbron, J., Lewis, G., & Jones, P. B. (2013). Prenatal maternal infection, neurodevelopment and adult schizophrenia: a systematic review of population-based studies. *Psychol Med, 43*(2), 239-257. doi: 10.1017/s0033291712000736

Kirov, G., Gumus, D., Chen, W., Norton, N., Georgieva, L., Sari, M., . . . Ullmann, R. (2008). Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet, 17*(3), 458-465. doi: 10.1093/hmg/ddm323

Kirov, G., O'Donovan, M. C., & Owen, M. J. (2005). Finding schizophrenia genes. *J Clin Invest, 115*(6), 1440-1448. doi: 10.1172/jci24759

Kirov, G., Pocklington, A. J., Holmans, P., Ivanov, D., Ikeda, M., Ruderfer, D., . . . Owen, M. J. (2012). De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry, 17*(2), 142-153. doi: 10.1038/mp.2011.154

Koiliari, E., Roussos, P., Pasparakis, E., Lencz, T., Malhotra, A., Siever, L. J., . . . Bitsios, P. (2014). The CSMD1 genome-wide associated schizophrenia risk variant rs10503253 affects general cognitive ability and executive function in healthy males. *Schizophr Res, 154*(1-3), 42-47. doi: 10.1016/j.schres.2014.02.017

Kordeli, E., Lambert, S., & Bennett, V. (1995). AnkyrinG. A new ankyrin gene with neural-specific isoforms localized at the axonal initial segment and node of Ranvier. *J Biol Chem, 270*(5), 2352-2359.

Kraepelin, Emil. (1899). *Psychiatrie : Ein Lehrbuch f‚r Studirende und Aerzte : 6. , vollst. umgearb. Aufl. Z Bd*. Leipzig: Johann Ambrosius Barth.

Kraus, D. M., Elliott, G. S., Chute, H., Horan, T., Pfenninger, K. H., Sanford, S. D., . . . Holers, V. M. (2006). CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. *J Immunol, 176*(7), 4419-4430.

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., . . . Wray, N. R. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet, 45*(9), 984-994. doi: 10.1038/ng.2711

Lee, S. H., Wray, N. R., Goddard, M. E., & Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet, 88*(3), 294-305. doi: 10.1016/j.ajhg.2011.02.002

Lennertz, L., Rujescu, D., Wagner, M., Frommann, I., Schulze-Rauschenbach, S., Schuhmacher, A., . . . Mossner, R. (2011). Novel schizophrenia risk gene TCF4 influences verbal learning and memory functioning in schizophrenia patients. *Neuropsychobiology, 63*(3), 131-136. doi: 10.1159/000317844

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi: 10.1093/bioinformatics/btp324

Lichtenstein, P., Bjork, C., Hultman, C. M., Scolnick, E., Sklar, P., & Sullivan, P. F. (2006). Recurrence risks for schizophrenia in a Swedish national cohort. *Psychol Med, 36*(10), 1417-1425. doi: 10.1017/s0033291706008385

Lichtenstein, P., Yip, B. H., Bjork, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., & Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet, 373*(9659), 234-239. doi: 10.1016/s0140-6736(09)60072-6

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., . . . Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature, 478*(7370), 476-482. doi: 10.1038/nature10530

Ma, C., Quesnelle, K. M., Sparano, A., Rao, S., Park, M. S., Cohen, M. A., . . . Brose, M. S. (2009). Characterization CSMD1 in a large set of primary lung, head and neck, breast and skin cancer tissues. *Cancer Biol Ther, 8*(10), 907-916.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi: 10.1038/nature08494

Maric, N. P., & Svrakic, D. M. (2012). Why schizophrenia genetics needs epigenetics: a review. *Psychiatr Danub, 24*(1), 2-18.

McCarthy, S. E., Gillis, J., Kramer, M., Lihm, J., Yoon, S., Berstein, Y., . . . Corvin, A. (2014). De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Mol Psychiatry, 19*(6), 652-658. doi: 10.1038/mp.2014.29

McCarthy, S. E., Makarov, V., Kirov, G., Addington, A. M., McClellan, J., Yoon, S., . . . Sebat, J. (2009). Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet, 41*(11), 1223-1227. doi: 10.1038/ng.474

McGrath, J. J., & Welham, J. L. (1999). Season of birth and schizophrenia: a systematic review and meta-analysis of data from the Southern Hemisphere. *Schizophr Res, 35*(3), 237-242.

McGrath, J., & Scott, J. (2006). Urban birth and risk of schizophrenia: a worrying example of epidemiology where the data are stronger than the hypotheses. *Epidemiol Psichiatr Soc, 15*(4), 243-246.

McGuffin, P., Tandon, K., & Corsico, A. (2003). Linkage and association studies of schizophrenia. *Curr Psychiatry Rep, 5*(2), 121-127.

Mitterauer, B. J. (2011). Possible role of glia in cognitive impairment in schizophrenia. *CNS Neurosci Ther, 17*(5), 333-344. doi: 10.1111/j.1755-5949.2009.00113.x

Modinos, G., Iyegbe, C., Prata, D., Rivera, M., Kempton, M. J., Valmaggia, L. R., . . . McGuire, P. (2013). Molecular genetic gene-environment studies using candidate genes in schizophrenia: a systematic review. *Schizophr Res, 150*(2-3), 356-365. doi: 10.1016/j.schres.2013.09.010

Monji, A., Kato, T. A., Mizoguchi, Y., Horikawa, H., Seki, Y., Kasai, M., . . . Kanba, S. (2013). Neuroinflammation in schizophrenia especially focused on the role of microglia. *Prog Neuropsychopharmacol Biol Psychiatry, 42*, 115-121. doi: 10.1016/j.pnpbp.2011.12.002

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., . . . Springer, M. S. (2001). Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science, 294*(5550), 2348-2351. doi: 10.1126/science.1067179

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., . . . Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet, 7*(3), e1001322. doi: 10.1371/journal.pgen.1001322

Need, A. C., McEvoy, J. P., Gennarelli, M., Heinzen, E. L., Ge, D., Maia, J. M., . . . Goldstein, D. B. (2012). Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet, 91*(2), 303-312. doi: 10.1016/j.ajhg.2012.06.018

Nicodemus, K. K., Callicott, J. H., Higier, R. G., Luna, A., Nixon, D. C., Lipska, B. K., . . . Weinberger, D. R. (2010). Evidence of statistical epistasis between DISC1, CIT and NDEL1 impacting risk for schizophrenia: biological validation with functional neuroimaging. *Hum Genet, 127*(4), 441-452. doi: 10.1007/s00439-009-0782-y

Onstad, S., Skre, I., Torgersen, S., & Kringlen, E. (1991). Twin concordance for DSM-III-R schizophrenia. *Acta Psychiatr Scand, 83*(5), 395-401.

Owen, M. J., Craddock, N., & O'Donovan, M. C. (2005). Schizophrenia: genes at last? *Trends Genet, 21*(9), 518-525. doi: 10.1016/j.tig.2005.06.011

Peippo, M., & Ignatius, J. (2012). Pitt-Hopkins Syndrome. *Mol Syndromol, 2*(3-5), 171-180. doi: 000335287

Perkel, J. M. (2013). Exome Sequencing: Toward an Interpretable Genome. Retrieved May 18, 2014, from http://www.sciencemag.org/site/products/lst_20131011.xhtml

Petersen, L., Mortensen, P. B., & Pedersen, C. B. (2011). Paternal age at birth of first child and risk of schizophrenia. *Am J Psychiatry, 168*(1), 82-88. doi: 10.1176/appi.ajp.2010.10020252

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res, 20*(1), 110-121. doi: 10.1101/gr.097857.109

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., & Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet, 86*(6), 832-838. doi: 10.1016/j.ajhg.2010.04.005

Project, ENCODE. (2014). ENCODE Common Cell Types. Retrieved June 4, 2014, from http://genome.ucsc.edu/ENCODE/cellTypes.html

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., . . . Sklar, P. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature, 506*(7487), 185-190. doi: 10.1038/nature12975

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

Quednow, B. B., Brinkmeyer, J., Mobascher, A., Nothnagel, M., Musso, F., Grunder, G., . . . Winterer, G. (2012). Schizophrenia risk polymorphisms in the TCF4 gene interact with smoking in the modulation of auditory sensory gating. *Proc Natl Acad Sci U S A, 109*(16), 6271-6276. doi: 10.1073/pnas.1118051109

Quednow, B. B., Brzozka, M. M., & Rossner, M. J. (2014). Transcription factor 4 (TCF4) and schizophrenia: integrating the animal and the human perspective. *Cell Mol Life Sci*. doi: 10.1007/s00018-013-1553-4

R Core Team. (2013). R: A Language and Environment for Statistical Computing.

Rees, E., Kirov, G., Sanders, A., Walters, J. T., Chambert, K. D., Shi, J., . . . Owen, M. J. (2014). Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol Psychiatry, 19*(1), 37-40. doi: 10.1038/mp.2013.156

Reitan, R. (1992). *Trail Making Test: Manual for Administration and Scoring*. Tuscon, AZ: Reitan Neuropsychology Laboratory.

Richter, T. M., Tong, B. D., & Scholnick, S. B. (2005). Epigenetic inactivation and aberrant transcription of CSMD1 in squamous cell carcinoma cell lines. *Cancer Cell Int, 5*, 29. doi: 10.1186/1475-2867-5-29

Riley, B. (2004). Linkage studies of schizophrenia. *Neurotox Res, 6*(1), 17-34.

Riley, B., Thiselton, D., Maher, B. S., Bigdeli, T., Wormley, B., McMichael, G. O., . . . Kendler, K. S. (2010). Replication of association between schizophrenia and ZNF804A in the Irish Case-Control Study of Schizophrenia sample. *Mol Psychiatry, 15*(1), 29-37. doi: 10.1038/mp.2009.109

Ripke S, Sanders AR, Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin DY, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, St Clair D, Corvin A, Gurling H, Werge T, Rujescu D, Blackwood DH, Pato CN, Malhotra AK, Purcell S, Dudbridge F, Neale BM, Rossin L, Visscher PM, Posthuma D, Ruderfer DM, Fanous A, Stefansson H, Steinberg S, Mowry BJ, Golimbet V, De Hert M, Jönsson EG, Bitter I, Pietiläinen OP, Collier DA, Tosato S, Agartz I, Albus M, Alexander M, Amdur RL, Amin F, Bass N, Bergen SE, Black DW, Børglum AD, Brown MA, Bruggeman R, Buccola NG, Byerley WF, Cahn W, Cantor RM, Carr VJ, Catts SV, Choudhury K, Cloninger CR, Cormican P, Craddock N, Danoy PA, Datta S, de Hann L, Demontis D, Dikeos D, Djurovic S, Donnelly P, Donohoe G, Duong L, Dwyer S, Fink-Jensen A, Freedman R, Freimer NB, Friedl M, Georgieva L, Giegling I, Gill M, Glenthøj B, Godard S, Hamshere M, Hansen M, Hansen T, Hartmann AM, Henskens FA, Hougaard DM, Hultman CM, Ingason A, Jablensky AV, Jakobsen KD, Jay M, Jürgens G, Kahn RS, Keller MC, Kenis G, Kenny E, Kim Y, Kirov GK, Konnerth H, Konte B, Krabbendam L, Krausucki R, Lasseter VK, Laurent C, Lawrence J, Lencz T, Lerer FB, Liang KY, Lichtenstein P, Lieberman JA, Linszen DH, Lönnqvist J, Loughland CM, Maclean AW, Maher BS, Maier W, Mallet J, Malloy P, Mattheisen M, Mattinsgsdal M, McGhee KA, McGrath JJ, McIntosh A, McLean DE, McQuillin A, Melle I, Michie PT, Milanova V, Morris DW, Mors O, Mortensen PB, Moskvina V, Muglia P, Myin-Germeys I, Nertney DA, Nestadt G, Nielsen J, Nikolov I, Nordentroft M, Norton N, Nöthen MM, O'Dushlaine CT, Olincy A, Olsen L, O'Neill FA, Ørntoft T, Owen MJ, Pantelis C, Papadimitriou G, Pato MT, Peltonen L, Petursson H, Pickard B, Pimm J, Pulver AE, Puri V, Quested D, Quinn EM, Rasmussen HB, Réthelyi JM, Ribble R, Rietschel M, Riley BP, Ruggeri M, Schall U, Schulze TG, Schwab SG, Scott RJ, Shi J, Sigurdsson E, Silverman JM, Spencer CC, Stefansson K, Strange A, Strengman E, Stroup TS, Suvisaari J, Tereniuis L, Thirumalai S, Thygesen JH, Timm S, Toncheva D, van den Oord E, van Os J, van Winkel R, Veldink J, Walsh D, Wang AG, Wiersma D, Wildenauer DB, Williams HJ, Williams NM, Wormley B, Zammit S, Sullivan PF, O'Donovan MC, Daly MJ, Gejman PV. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat Genet, 43*(10), 969-976. doi: 10.1038/ng.940

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kahler, A. K., Akterin, S., . . . Sullivan, P. F. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet, 45*(10), 1150-1159. doi: 10.1038/ng.2742

Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet, 46*(2), 222-228.

Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., . . . Daly, M. J. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet, 43*(11), 1066-1073. doi: 10.1038/ng.952

Rose, E. J., Morris, D. W., Hargreaves, A., Fahey, C., Greene, C., Garavan, H., . . . Donohoe, G. (2013). Neural effects of the CSMD1 genome-wide associated schizophrenia risk variant rs10503253. *Am J Med Genet B Neuropsychiatr Genet, 162B*(6), 530-537. doi: 10.1002/ajmg.b.32182

Rosenbloom, K. R., Sloan, C. A., Malladi, V. S., Dreszer, T. R., Learned, K., Kirkup, V. M., . . . Kent, W. J. (2013). ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res, 41*(Database issue), D56-63. doi: 10.1093/nar/gks1172

Saha, S., Chant, D., Welham, J., & McGrath, J. (2005). A systematic review of the prevalence of schizophrenia. *PLoS Med, 2*(5), e141. doi: 10.1371/journal.pmed.0020141

Sebat, J., Levy, D. L., & McCarthy, S. E. (2009). Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet, 25*(12), 528-535. doi: 10.1016/j.tig.2009.10.004

Selten, J. P., Frissen, A., Lensvelt-Mulders, G., & Morgan, V. A. (2010). Schizophrenia and 1957 pandemic of influenza: meta-analysis. *Schizophr Bull, 36*(2), 219-228. doi: 10.1093/schbul/sbp147

Shull, A. Y., Clendenning, M. L., Ghoshal-Gupta, S., Farrell, C. L., Vangapandu, H. V., Dudas, L., . . . Buckhaults, P. J. (2013). Somatic mutations, allele loss, and DNA methylation of the Cub and Sushi Multiple Domains 1 (CSMD1) gene reveals association with early age of diagnosis in colorectal cancer patients. *PLoS One, 8*(3), e58731. doi: 10.1371/journal.pone.0058731

Sklar, P., Smoller, J. W., Fan, J., Ferreira, M. A., Perlis, R. H., Chambert, K., . . . Purcell, S. M. (2008). Whole-genome association study of bipolar disorder. *Mol Psychiatry, 13*(6), 558-569. doi: 10.1038/sj.mp.4002151

St Clair, D., Xu, M., Wang, P., Yu, Y., Fang, Y., Zhang, F., . . . He, L. (2005). Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959-1961. *JAMA, 294*(5), 557-562. doi: 10.1001/jama.294.5.557

Stefansson, H., Ophoff, R. A., Steinberg, S., Andreassen, O. A., Cichon, S., Rujescu, D., . . . Collier, D. A. (2009). Common variants conferring risk of schizophrenia. *Nature, 460*(7256), 744-747. doi: 10.1038/nature08186

Steinberg, S., de Jong, S., Andreassen, O. A., Werge, T., Borglum, A. D., Mors, O., . . . Stefansson, K. (2011). Common variants at VRK2 and TCF4 conferring risk of schizophrenia. *Hum Mol Genet, 20*(20), 4076-4081. doi: 10.1093/hmg/ddr325

Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet, 13*(8), 537-551. doi: 10.1038/nrg3240

Sullivan, P. F., Kendler, K. S., & Neale, M. C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry, 60*(12), 1187-1192. doi: 10.1001/archpsyc.60.12.1187

Svrakic, D. M., Zorumski, C. F., Svrakic, N. M., Zwir, I., & Cloninger, C. R. (2013). Risk architecture of schizophrenia: the role of epigenetics. *Curr Opin Psychiatry, 26*(2), 188-195. doi: 10.1097/YCO.0b013e32835d8329

Takahashi, N., & Sakurai, T. (2013). Roles of glial cells in schizophrenia: possible targets for therapeutic approaches. *Neurobiol Dis, 53*, 49-60. doi: 10.1016/j.nbd.2012.11.001

Takahashi, N., Sakurai, T., Davis, K. L., & Buxbaum, J. D. (2011). Linking oligodendrocyte and myelin dysfunction to neurocircuitry abnormalities in schizophrenia. *Prog Neurobiol, 93*(1), 13-24. doi: 10.1016/j.pneurobio.2010.09.004

Tang, M. R., Wang, Y. X., Guo, S., Han, S. Y., & Wang, D. (2012). CSMD1 exhibits antitumor activity in A375 melanoma cells through activation of the Smad pathway. *Apoptosis, 17*(9), 927-937. doi: 10.1007/s10495-012-0727-0

Timms, A. E., Dorschner, M. O., Wechsler, J., Choi, K. Y., Kirkwood, R., Girirajan, S., . . . Tsuang, D. W. (2013). Support for the N-methyl-D-aspartate receptor hypofunction hypothesis of schizophrenia from exome sequencing in multiplex families. *JAMA Psychiatry, 70*(6), 582-590. doi: 10.1001/jamapsychiatry.2013.1195

Torgerson, D. G., Capurso, D., Mathias, R. A., Graves, P. E., Hernandez, R. D., Beaty, T. H., . . . Ober, C. (2012). Resequencing candidate genes implicates rare variants in asthma susceptibility. *Am J Hum Genet, 90*(2), 273-281. doi: 10.1016/j.ajhg.2012.01.008

Voineskos, A. N., Felsky, D., Kovacevic, N., Tiwari, A. K., Zai, C., Chakravarty, M. M., . . . Kennedy, J. L. (2013). Oligodendrocyte genes, white matter tract integrity, and cognition in schizophrenia. *Cereb Cortex, 23*(9), 2044-2057. doi: 10.1093/cercor/bhs188

Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., . . . Sebat, J. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science, 320*(5875), 539-543. doi: 10.1126/science.1155174

Wetterstrand, K. A. (2014). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). from http://www.genome.gov/sequencingcosts/

Williams, M. R., Hampton, T., Pearce, R. K., Hirsch, S. R., Ansorge, O., Thom, M., & Maier, M. (2013). Astrocyte decrease in the subgenual cingulate and callosal genu in schizophrenia. *Eur Arch Psychiatry Clin Neurosci, 263*(1), 41-52. doi: 10.1007/s00406-012-0328-5

Winantea, J., Hoang, M. N., Ohlraun, S., Rietschel, M., Cichon, S., Propping, P., . . . Freudenberg-Hua, Y. (2006). A summary statistic approach to sequence variation in noncoding regions of six schizophrenia-associated gene loci. *Eur J Hum Genet, 14*(9), 1037-1043. doi: 10.1038/sj.ejhg.5201664

Won, S., Kwon, M. S., Mattheisen, M., Park, S., Park, C., Kihara, D., . . . Lange, C. (2014). Efficient strategy for detecting gene x gene joint action and its application in schizophrenia. *Genet Epidemiol, 38*(1), 60-71. doi: 10.1002/gepi.21779

Xu, B., Ionita-Laza, I., Roos, J. L., Boone, B., Woodrick, S., Sun, Y., . . . Karayiorgou, M. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet, 44*(12), 1365-1369. doi: 10.1038/ng.2446

Xu, B., Roos, J. L., Dexheimer, P., Boone, B., Plummer, B., Levy, S., . . . Karayiorgou, M. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet, 43*(9), 864-868. doi: 10.1038/ng.902

Xu, W., Cohen-Woods, S., Chen, Q., Noor, A., Knight, J., Hosang, G., . . . Vincent, J. B. (2014). Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. *BMC Med Genet, 15*, 2. doi: 10.1186/1471-2350-15-2

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet, 88*(1), 76-82. doi: 10.1016/j.ajhg.2010.11.011

Yuan, A., Yi, Z., Wang, Q., Sun, J., Li, Z., Du, Y., . . . Yu, S. (2012). ANK3 as a risk gene for schizophrenia: new data in Han Chinese and meta analysis. *Am J Med Genet B Neuropsychiatr Genet, 159B*(8), 997-1005. doi: 10.1002/ajmg.b.32112

Zhang, C., Cai, J., Zhang, J., Li, Z., Guo, Z., Zhang, X., . . . Fang, Y. (2014). Genetic modulation of working memory deficits by ankyrin 3 gene in schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry, 50*, 110-115. doi: 10.1016/j.pnpbp.2013.12.010

Zhu, X., Gu, H., Liu, Z., Xu, Z., Chen, X., Sun, X., . . . Chen, C. (2013). Associations between TCF4 gene polymorphism and cognitive functions in schizophrenia patients and healthy controls. *Neuropsychopharmacology, 38*(4), 683-689. doi: 10.1038/npp.2012.234

Vita

Erik Kristen Loken was born on May 30, 1985 in Wilmington, Delaware. He graduated from The University of Virginia in Charlottesville, Virginia in 2007 with a Bachelor of Science in Chemistry. While pursuing dual M.D. and Ph.D. degrees at the Virginia Commonwealth University School of Medicine he participated as a student representative on the VCU MD-PhD Program Admissions Committee and as a student representative in a tenure review committee. He began his Ph.D. graduate training in the summer of 2010 at the Virginia Institute for Psychiatric and Behavioral Genetics as the first student in the Clinical and Translational Science with a specialization in the area of Psychiatric, Behavioral, and Statistical Genetics Ph.D. program. In the spring semester of 2012, he was a teaching assistant for Human Genetics 502: Advanced Human Genetics, co-instructed by his advisor Brien P. Riley. He has presented his work at the International Society of Psychiatric Genetics meetings in 2011, 2012, and 2013 and the American Society of Human Genetics in 2012. While working towards his graduate degree, he received funding from the "NIMH Training Program in Psychiatric and Statistical Genetics at the Virginia Institute for Psychiatric and Behavioral Genetics" NRSA T32 Grant from 2012 to 2013. In 2012 he submitted a fellowship application for a NRSA F30 Grant titled "Identifying functional variation in schizophrenia GWAS loci by pooled sequencing" which received funding in 2013. He published "The structure of genetic and environmental risk factors for fears and phobias" with his co-advisor Kenneth S. Kendler in Psychological Medicine in 2014. He received the 2014 Kenneth S. Kendler Award of Excellence in Pre-Doctoral Research.