2012

Quantitative Genetic Methods to Dissect Heterogeneity in Complex Traits

T. Bernard Bigdeli
Virginia Commonwealth University

Quantitative Genetic Methods to Dissect Heterogeneity in Complex Traits

by

Tim Bernard Bigdeli

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Ph.D.
in Human and Molecular Genetics
Medical College of Virginia of Virginia Commonwealth University
17 Nov 2011

Doctoral Committee:
Michael C. Neale, Ph.D. & Brion S. Maher, Ph.D. (Co-chairs)
Dr. Danielle M. Dick
Dr. Ayman H. Fanous
Dr. Kenneth S. Kendler
Dr. Brien P. Riley

TABLE OF CONTENTS

ABSTRACT

Quantitative Methods for Dissecting Heterogeneity in Genetic Studies of Complex Traits

by

Tim Bernard Bigdeli

Co-Chairs: Michael C. Neale & Brion S. Maher

Etiological models of complex disease are elusive[99, 68, 10], as are replicable findings of large effect[117, 17, 18, 53]. Commonly-cited explanations have previously invoked low-frequency genomic variation[93], allelic heterogeneity at susceptibility loci[68, 64], variable etiological trajectories[29, 28], and epistatic effects between multiple loci; these have represented among the most methodologically-challenging issues in molecular genetic studies of complex traits. Major sequencing initiatives, such as the 1,000 Genomes Project, are currently identifying human polymorphic sites at frequencies previously unassailable and, not ten years after publication of the first major genome-wide association findings, medical sequencing has already begun to displace GWAS as the standard for genetic analysis of complex traits. However, several recent studies have shown that the cumulative effect of a large number of common SNPs can account for a significant proportion of the variance in liability to complex traits, highlighting a conspicuous discrepancy between the explanatory value of reported GWAS associations and the realized contribution of common genetic variation. Emergent polygenic models posit the influence of thousands of common causal variants, many or most of which will remain obscured by genome-wide significance thresholds. Expectations regarding the number of additional variants “discoverable” by GWAS are sobering, as are implications for risk prediction in complex disease. With studies of complex disease primed for an unprecedented survey of human genetic variation, it is essential that these nascent, impending challenges be addressed.

Of interest herein are methodologies which utilize differential patterns of linkage disequilibrium to resolve the underlying genetic liability to complex traits, the range of allele frequencies for which common association tests are appropriate, and the relevant dimensionality of common genetic variation within ethnically-concordant but differentially ascertained populations. Using high-density SNP genotype data, we consider both hypothesis-driven and agnostic (genome-wide) approaches to association analysis, and address specific issues pertaining to empirical significance and the statistical properties of commonly-applied tests. Lastly, we attempt to place these diverse contributions into a unified framework of human population genetic theory.

CHAPTER I

Introduction & Relevant Background

1.1 Single Nucleotide Polymorphisms

Human genetic polymorphisms represent a diminutive fraction of the total diploid DNA complement, occurring nonetheless within a spectrum of allelic frequencies and varying considerably with respect to molecular information content. Single nucleotide polymorphisms (SNPs) are distinguished by their relative abundance within the genome, occurring with sufficient regularity to permit targeted studies of specific genomic regions by linkage-disequilibrium mapping [18, 5, 94]. Having introduced the class of polymorphism relevant to subsequent analyses, we establish an initial order of complexity for our discussion of genome-wide dimensionality, and introduce a basic metric of genetic diversity for a single locus within a population. Heterozygosity describes the probability of a given individual being a heterozygous at a given polymorphic site,

$$\sum_{i!=j}^{n} p_i p_j$$

[119]

where n is the number of alleles at said locus, and p_i and p_j are the allelic frequencies or, considered together, the genotype frequency for a particular combination of alleles in a diploid organism. The majority of SNPs are biallelic, thereby constraining the effective number of possibilities at a single locus to homozygous for either allelic variant, or heterozygous. Current genome-wide marker panels typically provide coverage of SNPs with minor allele frequencies (MAF) greater than 5%, below which standard associations tests are demonstrably unreliable (Chapter 2). As such, much of variation in complex disease as of yet unaccounted for by genome-wide studies has been attributed to putative effects of rare variants [93, 90, 83, 80, 12]. In addressing the accompanying variability as it pertains to realized dimensionality, it is rather more straightforward to partition our discussion into those components which are directly empirically quantifiable. That is, we consider variable allelic frequencies as manifested in the distributional properties of basic tests of association.

1.2 Association Between a SNP and Disease Outcome

Association between a SNP and disease outcome is commonly evaluated using Pearson's χ^2 test for independence, as applied to a contingency table of allele or genotype counts. A given individual's diploid set of alleles are considered independently, thus representing the relevant unit of analysis for which estimated effects are relevant [2]. Consider the 2×2 contingency table of allele counts by disease state, for a given SNP with alleles $A = \{A, a\}$

	A	a	
cases	A_{cas}	a_{cas}	$2N_{cas}$
controls	A_{con}	a_{con}	$2N_{con}$
	A_{total}	a_{total}	$2N$

As no restrictions are placed on the total number of observations in each cell, the quantities represent binomially-distributed data. The standard χ^2 statistic is calculated as

$$\chi^2 = \frac{N(A_{cas}a_{con} - a_{cas}A_{con})^2}{2N_{cas}N_{con}A_{total}a_{total}}$$

Alternatively, the Cochran-Armitage test for trend represents a modification of a genotypic test with 2 degrees of freedom to incorporate a suspected ordering of effects under a given genetic model [100]. For a given SNP, consider the 2×3 contingency table of genotype count by disease state, with genotype classes $G = \{g_0, g_1, g_2\}$ corresponding to carriers of zero, one, or two copies of the minor allele.

	AA	Aa	aa	Total
cases	AA_{cas}	Aa_{cas}	aa_{con}	$2N_{cas}$
controls	AA_{con}	Aa_{con}	aa_{con}	$2N_{con}$
Total	AA_{total}	Aa_{total}	aa_{total}	$2N$

We calculate the test-statistic as

$$T = \frac{\sum_{i=0}^{2} w_i \cdot (N_{con} \cdot g_{i|con} - N_{cas} \cdot g_{i|cas})}{Var(T)}$$

where $w = \{w_0, w_1, w_2\}$ is the set of weights to be applied to the k genotype categories. Under an additive model, in which each additional copy of a disease-associated allele increases liability to illness, we apply weights $w = \{0, 1, 2\}$. For large N, the approximation of T_{catt} is a normally distributed ($\mathrm{N}_{(0,1)}$), $1d.f.$ random variable and a more powerful test of association than a genotypic χ^2 test.

For a fixed number of minor alleles, there exists a discrete set of possibilities for how these might be arranged within a contingency table of observed counts. Let C represent any even number of alleles, and let us assume equal numbers of cases and controls. From the 2×2 contingency table corresponding to an allelic χ^2 (1.1), it is apparent that our observed data may take on any of $C + 1$ arrangements. For a two-sided test, these $C + 1$ arrangements represent $C/2 + 1$ possible values of the resultant test-statistic. That is, if we restrict all C copies to either cases or controls, we obtain the most-extreme result. For any fixed C, this discrete set of possibilities is illustrated by a normal probability plot of observed quantiles, which will be distributed as a step-function. From the 2 × 3 contingency table of genotypes χ^2 by outcome (1.2), we see that consideration of the diploid state substantially increases the number of possibilities for the observed data. Because the observed count of either genotypic class constrains the value of others, we calculate the total number of possible arrangements as follows:

$$\sum_{i=0}^{C/2} \sum_{j=0}^{C-2i} 1$$

For forty copies of the minor allele, this yields 3,311 possibilities for a 2 × 3 table, compared with 41 possibilities for allelic data. However, for progressively fewer C, the probability of observing a minor allele homozygote becomes vanishingly small, and gains in statistical power from application of the Cochran-Armitage test for trend will be minimal, especially when N is large.

Interpretation of the resultant test-statistic is by approximation to a theoretical distribution, providing a probability estimate of observing so extreme a value by chance. Significance for dense-SNP data is assessed under strong control of the FWER, meaning that we consider only those test-statistic values which are very extreme, or highly improbable under the null hypothesis. The critical range of values corresponds to the asymptotic tail(s) of the reference distribution which, for normally- or χ^2-distributed test-statistics, trends to zero for increasingly extreme values. Of central importance is the tendency of Pearson's χ^2 test to over-estimate significance when observed counts are very small or absent entirely. If inflation of the resultant test-statistic is sufficient to warrant rejection of the null hypothesis, then a Type-I error has occurred. To avoid systematic accruement of false-positive findings, Yates's correction for continuity is applied [126]. The rationale for Yates' correction is, simply, that this inaccuracy arises as a direct consequence of deriving probability estimates for a binomial distribution using the continuous, theoretical χ^2 distribution. In practice, if any cell within a given contingency table contains too few observations, then 0.5 is added to each quantity, thereby lessening the calculated differences between observed and expected values. Common variants with low minor allele frequency (MAF) are subject to this source of Type-I error, especially in underpowered studies. On the other hand, note that the correction is itself prone to returning somewhat deflated probability estimates, thus increasing the Type-II, or false-positive, error rate. Alternatively, Fisher's exact test provides an exact estimate of significance for a given set of values within a contingency table, and is an appropriate method when sample size is limiting [34, 126]. Recall the distinction between sampling procedures that are with or without replacement; Fisher's Test evaluates an exact probability estimate from the hypergeometric distribution of counts within a contingency table with fixed marginal total, whereas Pearson's χ^2 provides an approximation to the theoretical χ^2 distribution for binomial data with the expectation of convergence as N approaches infinity (Central Limit Theorem). As such, Fisher's Exact test represents the canonical test for small sample sizes. Calculation of the exact test statistic is as follows:

$$p = \frac{(2N_{cas})!(2N_{con})!(A_{total})!(a_{total})!}{(A_{cas})!(a_{cas})!(A_{con})!(a_{con})!(2N)!}$$

With respect to significance for low-frequency markers, the sidedness of a particular test-statistic distribution is especially salient. Exploratory genome-wide studies are predominantly hypothesis-free, without any *a priori* expectation that a particular allelic variant is more prevalent among cases or controls. For symmetrical distributions such as the usual χ^2, the two-sided P-value may simply be halved to obtain the one-sided estimate; whether an observed effect is "protective" or conveys some "risk" is gleaned directly from the observed counts. For Fisher's Test, the proper means of obtaining a two-sided P-value has been subject to debate. In some instances, the two-sided test gives a result only slightly larger than the one-sided P-value, reflecting asymmetry in the observed distribution [61].

1.3 Family-Based Approaches to Association

Basic approaches to family-based association test for excess transmission of an allele or alleles at a candidate marker, the most notable being the transmission disequilibrium test (TDT) [108] and the generalized to arbitrary pedigrees and phenotypes, family based association test (FBAT) [98]. An alternative approach to the TDT creates pseudocontrols based on the alleles that were, and were not, transmitted to an affected offspring from genotyped informative parents. Limitations of Spielman's TDT arise upon inclusion of families for which multiple affected offspring or extended pedigrees are available. Whereas Spielman's TDT considers a contribution from any heterozygous parent to be independent, the the pedigree disequilibrium test (PDT) [67] treats each trio as an independent unit of analysis. Because tests of association may not assume independence of trios and discordant sib-pairs (DSPs) derived from the same extended pedigree, PDT defines a summary variable accounting for each possible trio and DSP. For a biallelic marker A with alleles $A1$ and $A2$, let us define for each informative trio and DSP,

$$X_T = (A_1 \text{alleles transmitted})(A_1 \text{alleles not transmitted})$$

$$X_S = (A_1 \text{alleles in affected sib})(A_1 \text{alleles in unaffected sib}).$$

Then, for each extended pedigree, let n_T and n_S represent the numbers of informative trios and DSPs, respectively, as follows:

$$T = \frac{\sum_{i=1}^{N} D_i}{\sqrt{\sum_{i=1}^{N} D_i^2}}$$

Under the null hypothesis of no linkage disequilibrium between the marker and the trait, T is normally distributed, with mean equal to 0 and variance equal to 1.

Tests of association for pedigrees generally feature at least one of two components; the within-family component considers transmissions of alleles from parent to child, or the extent of allele-sharing between siblings, and is typically robust to stratification and, potentially, allelic heterogeneity; the between-family component considers the actual allele transmitted or shared, but is liable to population stratification.

1.4 Multi-locus Diversity and Linkage Disequilibrium Mapping

Consideration of a second, biallelic locus demonstrates the correspondence between the number of independent SNPs and a simple binomial expansion. Examination of Pascal's Triangle illustrates the extent to which progressively greater numbers of variable sites increases the relevant set of possibilities for observed multi-SNP genotypes. That is, if n represents the number of independent, biallelic loci under consideration, $2^{(n+1)}$ gives the corresponding number of possibilities for the allelic complement observed for a given, diploid individual. Given 10 and 100 SNPs, we observe 2048 and 2.53×10^{30} possible combinations, respectively, while the number of "possible" allelic combinations reflecting current marker densities in GWAS corresponds to a binomial expansion which is incalculable by standard analytical packages. While the preceding illustration clearly represents a profound exaggeration of the relevant dimensionality for an individual's genome-wide complement, it nonetheless permits us to establish a discrete set of basic assumptions regarding the biological constraints on individual genetic differences.

In a given population, the extent of genetic sequence diversity is largely attributable to conserved evolutionary mechanisms underlying sexual reproduction and—albeit to a vastly lesser degree—rates of *de novo* molecular changes to DNA sequence itself. Whether of the transmissible or sporadic type, the aforesaid differences arise from a common, founding genetic source (i.e. population), often referred to collectively as the “ancestral” state (e.g. genome, chromosome, haplotype, or allele). At a single, polymorphic locus, the extent of variability is, at bottom, a function of the number and respective frequencies of allelic variants, while the extent of variability for syntenic positions is a function of the rates of recombination within that region. The resulting pattern of intercorrelations between SNPs describes linkage disequilibrium (alternatively, gametic phase disequilibrium), the extent to which co-segregation of alleles at distinct loci deviates from expectation. That is, a hypothetical set of loci with alleles A/a and B/b, is said to be in linkage equilibrium if the possibilities for unique gametes (or haplotypes), given by AB/Ab/aB/ab occur in significantly higher (or lower) proportions than expected based on the frequency of each allele i.e. A*B, A*b, a*B, and a*b. Measures of LD include D, D', and both signed and unsigned r (r^2) [5, 8, 52]. The choice of LD measure generally depends on context and application, as their properties differ somewhat. More generally, LD relationships realized over expansive genomic regions form the basis of indirect approaches to association.

Indirect approaches to association are premised on the presence of unobserved variants, the notion of complete saturation of genomic variation having been, until only recently, seemingly incredible. As exemplified by the additional degree-of-freedom(s) routinely incurred by haplotype-based association tests, the dimensionality of a multi-SNP haplotype is given by the smallest set of constituent alleles which fully differentiate it. Haplotype definitions may therefore include SNP and non-SNP variation, given that variants are both inheritable and linked. Rare haplotypic backgrounds or the presence of nearby causal variation is more readily detected by the combined information of multiple tagging-SNPs. The observation that a disease association is stronger with a haplotype than its composite SNPs is also consistent with *cis*-interactions between these alleles [81]. In contrast, epistatic or *trans*-effects represent interactions between “unlinked” loci.

1.5 Complex, common human disease

1.5.1 Multifactorial Inheritance

The multifactorial model posits that there exists an unobserved continuous liability distribution underlying most complex traits, such that numerous small effects yield the observed phenotypic outcome. Variants of moderate to high effect-size are not likely to be common, and thus would not contribute significantly to liability to common diseases. Rather, a multitude of loci displaying conditional penetrance, possible epistatic, and environmental interactions likely yield the clinically observable disease state. In the context of complex disease etiology, We consider two types of genetic heterogeneity, locus and allelic. Locus heterogeneity arises from the existence of multiple disease genes of convergent effect with respect to pathophysiology, any number of which may or may not be relevant to all instances of disease. Allelic heterogeneity describes the existence of multiple polymorphisms within a particular disease gene, such that carrying a particular allele may only be predictive of disease within select populations.

1.5.2 Schizophrenia

Schizophrenia is a devastating psychiatric condition affecting nearly 1% of the worlds population, and in which a substantial number of patients do not respond to treatment [9]. Additionally, patients vary with respect to particular symptoms, with hallucinations and delusions often accompanied by a combination of cognitive deficits, disorganization symptoms, and severe mood disorder, including manic and depressive episodes. Family studies indicate markedly increased risk of illness among relatives of afflicted individuals, with children of schizophrenics as much as ten times and cousins twice as likely to become ill. Heritability estimates, some upwards of 80% [38], indicate that a substantial proportion of variance in disease risk is attributable to genetic factors.

1.6 GWAS, Multiple-Testing

Available genome-wide platforms offer dense coverage of common (typically, 5% MAF or greater) single nucleotide polymorphisms (SNPs), relying on an indirect approach to association through linkage disequilibrium (LD) to capture common variation in the vicinity of a typed locus, and typically including only those variants with minor allele frequency (MAF) greater than 5%. Although quite efficient, there is little power with these SNPs to detect associations with uncommon, or rare variation. By convention, any SNP with $<$ 5% MAF is considered “rare” in this context. Exceptions include age-related macular degeneration and inflammatory bowel disease, for which large-effect variants were successfully identified using relatively modest sample sizes.

A particularly salient issue in GWAS is what strength of evidence constitutes a genome-wide significant finding. As association is typically assessed on a per-SNP basis, an appropriate correction for multiple-tests is by control of the family-wise error rate (FWER) or Type-I error, the probability of observing a significant finding by chance if no true association exists. Calculation of the corresponding significance threshold, α', is a function of the number of markers analyzed, although methods exist for estimating an effective number of “independent” loci based on LD between markers. Numerous α' thresholds have been proposed to represent significance at the genome-wide level, most being to the order of $\sim 10^{-8}$. Lower-frequency SNPs are more likely to exhibit significant differences between groups, and will yield smaller P-values. It follows that the corresponding distribution of P-values will require more stringent correction to maintain equivalent control of the FWER. This has been demonstrated by simulation of dense-SNP and resequencing data, and comparison to expected distributions for existing genome-wide platforms. Efforts to estimate the "effective" number of tests in the genome illustrate the effect of SNP ascertainment and study design on the expected distribution of P-values [24, 48]. For rare variants, larger samples have the potential to yield smaller P-values, with an accompanying decrease in the α' required to maintain the FWER at a desirable level.

CHAPTER II

Empirical Significance for Single Marker Tests of Low-Frequency Variants

T. Bernard Bigdeli[1,2], Michael C. Neale[1,2,3], and Benjamin M. Neale[4,5] [1]**Department of Human and Molecular Genetics,Virginia Commonwealth University,** Richmond, VA

[2]**Virginia Institute for Psychiatric and Behavioral Genetics Virginia Commonwealth University,** Richmond, VA

[3]**Department of Psychiatry, Virginia Commonwealth University,** Richmond, VA

[4]**Center for Human Genetic Research, Massachusetts General Hospital,** Boston, MA

[5]**Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology,** Cambridge, MA

2.1 Abstract

With the dramatic technological developments of genome-wide association SNP Chips and next generation sequencing, human geneticists now have the ability to assay genetic variation at ever rarer allele frequencies ($\geq .01$). To fully understand the impact of these rare variants on common, complex disease, we must be able to accurately assess the significance of these variants. However, it is well-established that classical association tests are not appropriate for the analysis of low-frequency variation, giving spurious findings when observed counts are too few. To further our understanding of the asymptotic properties of traditional association tests, we conducted a range of null simulations of a typical rare variant and proceeded to test the allelic χ^2, Cochran-Armitage trend, Wald and Fisher's exact tests. We demonstrate that rare variation shows marked deviation from the expected distributional behavior for each test, with fewer minor alleles corresponding to a greater degree of test-statistic deflation. The effect becomes more pronounced at progressively smaller levels. We also show that the Wald Test is particularly deflated at α levels consistent with genome-wide association significance, much more so than the other association tests considered. In general, these classical association tests are inappropriate for the analysis of variants for which the minor allele is observed fewer than 80 times.

Keywords: genome-wide association, next-generation sequencing, significance testing, rare variation

2.2 Introduction

Genome-wide association studies (GWAS) have uncovered hundreds of loci relevant to common, complex disease [65]. These studies assay SNP variation across the allele frequency spectrum, but are limited to studying SNPs with minor allele frequency (MAF) of at least 1-5%. In spite of incomplete coverage of rare alleles in GWAS, a number of rare variants have been implicated in common, complex disease. For example, recent work in Type I Diabetes identified a rare protective mutation in the gene *IFIH1*, with a population allele frequency of approximately 2% [83]. New sequencing endeavors such as the 1,000 Genomes Project are identifying human genetic variation down to frequencies less than one percent. This expanding collection of genetic polymorphisms is, in turn, being made accessible through extending genome-wide association SNP chips to ever decreasing frequencies.

With the increased focus on rare variants, the question of how best to assess their statistical significance arises. For extremely uncommon variation, methods have been developed to test whether a set of variants are implicated in disease [75, 59, 63, 80]. Such methods are better suited to loci for which classical association testing cannot be conducted because of the limited number of observations. Considered another way, a single locus that has only ten copies of the minor allele in a balanced case-control study cannot achieve significance at established genome-wide levels (5×10^{-8}) [99]. One strategy to overcome this problem is to group multiple variants and to conduct tests of association with particular regions rather than with specific variants.

Given that the field has adopted a genome-wide association significance threshold, the accuracy of extreme p-values is also of great importance. For example, in the seminal work of Jonathan Cohen and colleagues, who identified *PCSK9* as a component of LDL cholesterol, the authors used a χ^2 test to assess the role of rare variation in determining risk of coronary heart disease, and reported p-values of 0.008 and 0.003 for African-American and Caucasian samples, respectively [13]. If instead a Fisher's exact test is applied, these p-values shrink to 0.0037 and 0.0024, respectively. Thus, the basic χ^2 test in this circumstance is comparatively conservative in the face of rare observations. To enhance our understanding of the asymptotic properties of these traditional association tests, we have undertaken a range of simulations of the χ^2, Cochran-Armitage trend [78, 120, 106], and Wald tests [101].

2.3 Methods

To assess the asymptotic behavior of rare variant testing, we used a simple null model consisting of a SNP with 1% MAF equally likely to occur in 1,000 cases and 1,000 controls. We initially assigned genotypes for each individual randomly, allowing for sampling variance. That is, each individual replicate may have an observed minor allele frequency of 1%, a little more than 1% or slightly less than 1%. To further constrain the behavior of these tests, we limited the minor allele count to 40 copies among 2,000 individuals. To determine whether the sample size matters, we increased the number of individuals to 10,000, while still fixing the number of minor alleles to 40. We also considered 20 and 80 copies of the minor allele in a sample size of 10,000. We proceeded to analyze each simulated dataset using a suite of common, association tests: the allelic χ^2, the 1 d.f. Cochran-Armitage trend test and the Wald Test for logistic regression. As our goal was to assess the asymptotic behavior of these tests, we chose to conduct a large number of simulations (one billion) for each scenario.

2.3.1 Tests of Association

The allelic χ^2 test compares allele frequencies between cases and controls, and is widely used as a test of association for disease traits [2]. Because the allelic test considers the allele as the relevant unit of analysis, it is assumed that Hardy-Weinberg equilibrium exists. This is equivalent, in the present context, to assuming that the alleles at a locus occur independently within both case and control populations. In other words, non-additive effects of the alleles at a locus are assumed to be absent. The allelic test is known to give spurious results if this condition is not met, although SNPs that show severe departures from Hardy-Weinberg equilibrium are generally unreliable and should be excluded from analysis. Interpretation of odds ratios given by this method is also with respect to alleles, as opposed to individuals, and is discussed elsewhere [100].

The Cochran-Armitage test for trend is a modification of a 2 d.f. genotypic χ^2 to account for an hypothesized ordering of effects across genotype classes, consistent with additive models of disease risk [4, 35]. Applied to common variants, the trend-test is a more powerful test of association than standard allelic and genotypic χ^2, owing to a weighting of genotypic classes which reduces the effective degrees of freedom. Since the individual represents the relevant unit of analysis, the trend test has the additional advantage of not assuming Hardy-Weinberg equilibrium, though the allelic and trend tests are expected to be asymptotically equivalent when this condition is met [100]. Odds ratios from the trend test may be interpreted as the increase in risk to an individual conferred by each additional copy of the non-reference (i.e. minor) allele.

The Wald Test [118, 46] compares the maximum likelihood estimate of a statistical parameter to its expectation under the null, often as an approximation to the theoretical χ^2 distribution. In the present context, we apply the Wald Test to a simple logistic model ($\mathrm{Aff} \sim \beta_0 + \beta_1 \cdot \mathrm{SNP}$) which considers the number of minor alleles carried by an individual. Because it is often desirable to include demographic or clinical covariates in predictive disease models, we extend our regression model to incorporate a covariate predictor of fixed prevalence in the population, but for which carriers of the minor allele are at increased risk of endorsement. As for our basic logistic model, we applied the Wald Test to obtain a χ^2 approximation for the effect of SNP genotype.

2.3.2 Generation of Asymptotic Distributions

Under each scenario, we simulated genotypic data which were identical with respect to the total number of minor alleles, C, the total sample size, N, and the proportions of cases and controls. For each replicate dataset, we sampled N times without replacement from a population of N diploid persons, in which only C chromosomes carry the minor allele, and assigned case-status at random to exactly half of all individuals. It follows that the resultant case-control differences in allele frequency will be identically distributed, as illustrated by the observation that in the most-extreme circumstance, all C copies of the minor allele will occur within cases or controls. By comparison, random simulation of genotypes on a per individual basis, as previously described, might yield instances in which the total number of alleles is slightly greater or slightly fewer than C, thus introducing an additional source of variation in the test-statistic. Stated differently, each replicate dataset represents a standard 2×2 table of allele counts by outcome , but for which the marginal totals of rows and columns are fixed. Similarly, for both the trend test and the logistic model, the data may be arranged as 2×3 contingency tables of genotypic counts by outcome, in which the marginal totals are generally maintained. That is, our focus is on the asymptotic properties of standard association tests as applied to low-frequency variants, for which the occurrence of a minor allele homozygote (MAF^2) is an exceedingly rare event.

Because it is often desirable to include demographic or clinical covariates in predictive disease models [116, 7, 101], we extended the regression models to incorporate a binary covariate predictor of fixed prevalence in the population, which carriers of the minor allele are more likely to endorse. We assume a .10 population endorsement rate across all scenarios, but vary this rate among carriers as .10, .20, .40, .60, and .80. For each replicate dataset, we fitted logistic models which specified case-control status as a function of SNP genotype and a single covariate, and applied the Wald

Test to obtain a χ^2 approximation for the effect of the SNP genotype. Of particular interest is the effect of adding a predictor, unrelated to disease, on the regression of disease outcome on genotype. Note that although the numbers of cases and controls are fixed and equal across permutations, random simulation of a covariate will introduce variance into the observed distribution of test-statistics.

Distributions for Fisher's Exact Test were also derived, but indirectly from the distributions for the allelic χ^2. This is justified by our simulation procedure, as fixing the marginal totals constrains the number of possible configurations of the data within a 2×2 table of counts. That is, each unique value of the allelic χ^2 corresponds to a specific set of observed counts for which the value of Fisher's test is known.

Due to the exceptional number of permutations required to evaluate asymptotic behavior within the critical region, we seeded 100,000 separate instances of our simulation procedure per scenario, making use of several high-performance computing clusters. Rendering of complete null distributions for each test was simplified by tabulating observed test-statistics within each constituent distribution and compounding the resulting counts. We proceeded to quantify departures from expected asymptotic behavior, as defined by the theoretical χ^2 distribution for 10^9 tests.

2.4 Results

2.4.1 Common Association Tests

For each scenario, Table 2.1 gives the number of Cochran-Armitage trend, allelic χ^2, and logistic regression tests (uncorrected for continuity) found to be significant at various α-levels. Corresponding quantile-quantile plots are displayed in Figure 2.1. Expectations regarding asymptotic behavior are based on the theoretical χ^2 distribution (see Central Limit Theorem), to which approximations of binomial SNP data are definitively inexact. At a given threshold, the probability of observing a significant test-statistic under the null is simply the proportion of the total number of permutations. Because our sampling procedure is effectively without replacement, resultant test-statistics occur in discrete quanta. This is illustrated by the step-function-like appearance of the observed quantile plots (Figure 2.1).

Consider the distributions of allelic χ^2 and Fisher's Exact tests, recalling that a 2×2 table of allelic counts will follow a hypergeometric distribution if marginal totals are held constant. For 40 copies of the minor allele in 1,000 cases and 1,000 controls, we observe fewer significant allelic χ^2 tests than expected, with more pronounced discrepancies for progressively smaller α. Comparing the allelic χ^2 and Fisher's Exact methods, significant test counts obtained by each method are indistinguishable for all but the most-extreme α-levels. Given the same number of minor alleles (40) in 5,000 cases and 1,000 controls, we see an overall pattern of deflation similar to that observed for the smaller sample. Inspection of Table 2.1 reveals a slight increase in the number of significant tests observed, less than 2% and 5% for α thresholds of 10^{-2} and 10^{-5}, respectively. However, the larger sample size does not see the allelic χ^2 attain significance at $\alpha < 10^{-8}$. Restricting the number of minor alleles to exactly 20 copies, there is marked decrement in the value of test-statistics by either method, with neither reporting a single p-value less than 10^{-6}. We observe an excess of significant findings by the allelic χ^2 for the $\alpha < 10^{-2}$ critical region, with no such inflation apparent for Fisher's Exact Test. Increasing the number of minor alleles to 80 copies in 5,000 cases and 5,000 controls, asymptotic behavior is visibly restored. Residual deflation is only minimally apparent at genome-wide thresholds, at which both tests report significant findings.

Excepting those additional values indicating one or more observed minor homozygotes, quantiles for the Cochran-Armitage trend test largely parallel those of the allelic χ^2. With 40 minor alleles in 1,000 cases and 1,000 controls, the trend test gives a significant result at $\alpha < 10^{-8}$ which the allelic test failed to identify. Loss of power is evident, with fewer significant permutations observed overall than with either the allelic or exact test. Differences between the allelic χ^2 and trend tests are less marked with 40 copies in 5,000 cases and 5,000 controls, due to the reduced likelihood of observing a minor allele homozygote. With only 20 copies of the minor allele, power for the trend test is diminished further. Under these conditions, the chances of observing a minor homozygote is only one in one million. Like the allelic χ^2 and exact tests, the trend test fails to return a single p-value less than 10^{-6}. An excess of significant findings in the $\alpha < 10^{-2}$ critical region is also apparent, but to a slightly lesser extent than seen for the allelic χ^2. Power for the trend test is restored by increasing the number of minor alleles to 80 copies. Despite the deflation being visibly attenuated, the trend test gives slightly fewer significant differences in the critical region than either the allelic χ^2 or Fisher's Exact test.

Deflation of the Wald Test statistic is considerably more pronounced than those of the allelic χ^2 and trend tests. With 40 minor alleles in either sample size, the Wald Test fails to report a single significant finding at $\alpha < 10^{-5}$, returning p-values 10, 100, and 1,000-fold larger than expected at α thresholds of 10^{-5}, 10^{-7}, and 10^{-8}, respectively. With the total number of minor alleles limited to 20 copies, deviation from expected distributional behavior is particularly extreme. We fail to observe any significant findings for $\alpha < 10^{-3}$, corresponding to a deflation factor of 100,000 at $\alpha < 10^{-8}$. With 80 minor alleles in 5,000 cases and 5,000 controls, the Wald Test is noticeably improved but still gives p-values an order of magnitude larger than expected at $\alpha < 10^{-8}$.

Comparing 80, 40, and 20 copies of the minor allele in 5,000 cases and 5,000 controls, there is an overall increase in the extent of deflation for successively fewer copies of the minor allele, and an increase in the value of α at which this deflation is first apparent. Given the demonstrated non-effect of sample size, it follows that we may take findings for 80 minor alleles in 5,000 cases and 5,000 controls as indicative of expected null behavior for a 2% MAF SNP. Under these conditions, the allelic χ^2 and trend tests exhibit similar asymptotic behavior and return empirical significance estimates which, compared to those obtained by Fisher's Exact Test, are not appreciably misestimated. Equivalently, we take findings for 40 minor alleles in 5,000 cases and 5,000 controls as representative of a 1%, establishing a reasonable lower limit for the allelic χ^2 and trend tests. The Wald Test is particularly sensitive to the number of minor alleles, returning substantially diminished estimates of significance in the genome-wide critical region. At $\alpha < 10^{-6}$, deflation of the Wald Test statistic is at least 4, 40, and 400 times greater than for the allelic χ^2 with 80, 40, and 20 minor alleles, respectively. Whereas the allelic and trend tests both exhibit inflation in the $\alpha < 10^{-2}$ critical region for 20 copies of the minor allele, the counts for Fisher's Test are simply reduced compared to 40 or 80 copies, demonstrating the robustness of Fisher's Test in situations for which our common tests are not suitable.

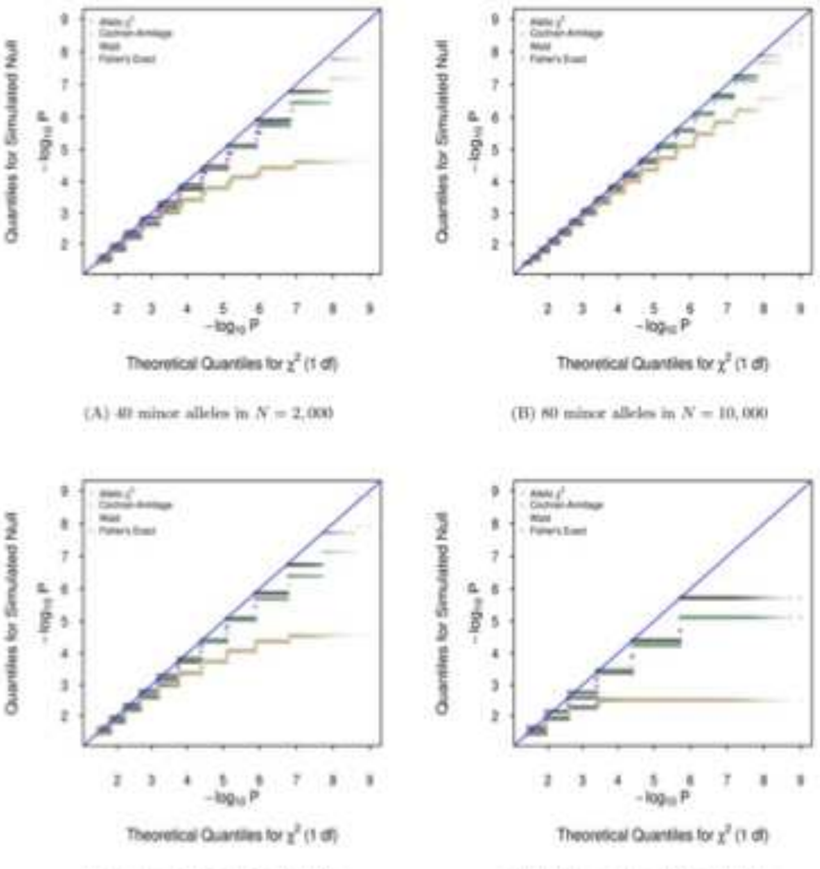
Table 2.1: Expected and observed counts of significant tests at particular α-levels for three 1-$d.f.$ tests, calculated for a fixed number of minor alleles among equal numbers of cases and controls.

Null Distribution			Significance Threshold (α)							
			$\alpha < 10^{-1}$	$\alpha < 10^{-2}$	$\alpha < 10^{-3}$	$\alpha < 10^{-4}$	$\alpha < 10^{-5}$	$\alpha < 10^{-6}$	$\alpha < 10^{-7}$	$\alpha < 10^{-8}$
Theoretical χ^2 (1 $d.f.$)			100,000,000	10,000,000	1,000,000	100,000	10,000	1,000	100	10
	Count$_{MA}$	N								
Cochran-Armitage Trend	40	2,000	79,179,748	6,173,009	639,359	38,879	5,987	143	8	1
Allelic χ^2			79,179,748	6,173,016	640,064	38,933	7,647	148	11	0
Wald			79,179,748	6,167,797	534,239	5,983	0	0	0	0
Fisher's Exact			79,179,748	6,173,016	640,064	38,933	7,647	148	11	1
Cochran-Armitage Trend	80	10,000	92,269,671	9,541,111	1,021,754	68,457	8,128	699	55	3
Allelic χ^2			92,269,671	9,541,208	1,029,744	68,459	8,209	791	65	3
Wald			92,269,671	9,538,127	439,954	24,480	2,215	65	0	0
Fisher's Exact			92,269,671	9,541,208	1,029,744	68,459	8,209	791	65	3
Cochran-Armitage Trend	40	10,000	80,400,565	6,368,800	669,924	41,455	7,745	169	17	0
Allelic χ^2			80,400,565	6,368,800	669,932	41,455	8,147	170	18	0
Wald			80,400,565	6,368,746	178,816	7,745	0	0	0	0
Fisher's Exact			80,400,565	6,368,800	669,932	41,455	8,147	170	18	2
Cochran-Armitage Trend	20	10,000	115,139,237	11,657,125	399,840	39,160	1,884	0	0	0
Allelic χ^2			115,142,423	11,772,128	399,870	39,822	1,927	0	0	0
Wald			115,137,310	2,561,766	0	0	0	0	0	0
Fisher's Exact			41,290,490	2,563,863	399,870	39,822	1,927	0	0	0

2.4.2 Null Covariate Effect

Table 2.2 gives the observed number of Wald Test statistics for logistic models incorporating a null covariate effect of fixed prevalence among controls; corresponding quantile-quantile plots are displayed in Figure 2.2. Regression coefficients, α levels and expectations regarding asymptotic behavior are as described for our previous implementation.

With random assignment of case-status, inclusion of the covariate in our regression analysis should not alter the observed distribution of test-statistics. While generally true, approximations at the extreme tails appear slightly less deflated for higher prevalences of the covariate among carriers of the minor allele (Figure 2.2). Strictly-speaking, this phenomenon may be best described as countervailing inflation, occurring as a result of increased sampling variance. That is, increasing the covariance between minor allele and covariate is accompanied by a gradual degradation of the discrete-valued function seen for our original logistic model. For very small α, at which approximations of binomial data to the continuous χ^2 distribution are exceptionally poor, this additional variance imparts a slight effect on our probability estimates. Comparison of 40 minor alleles in 1,000 and 5,000 cases and 5,000 controls exemplifies our interpretation; the effect is markedly enhanced in the smaller sample, as would be expected for any sampling effect.



(A) 40 minor alleles in $N = 2,000$

(B) 80 minor alleles in $N = 10,000$

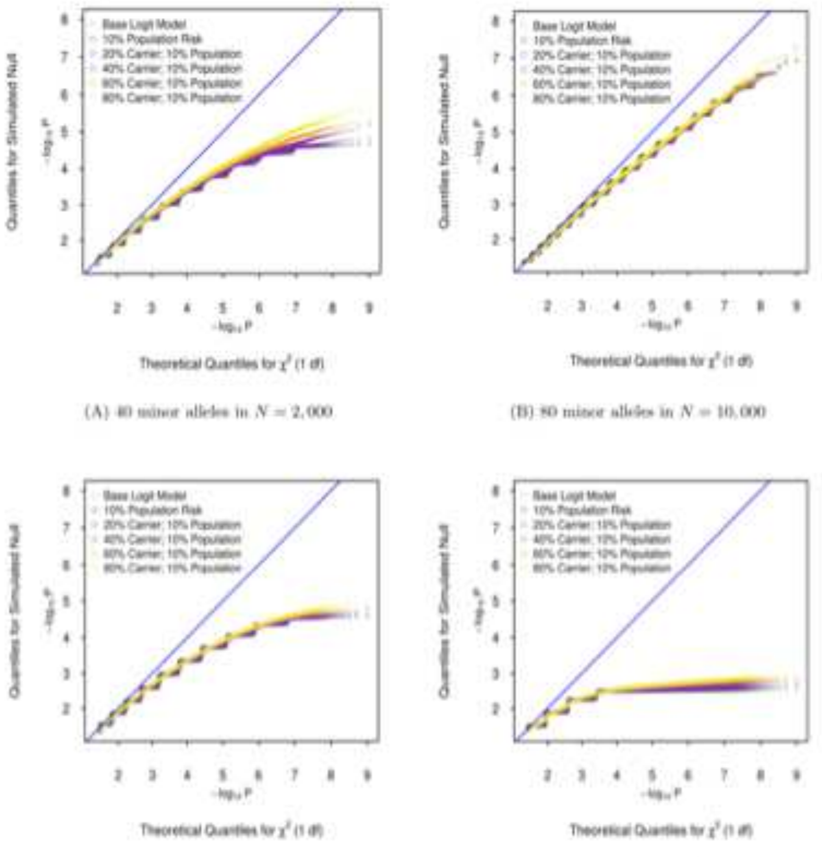
(C) 40 minor alleles in $N = 10,000$

(D) 20 minor alleles in $N = 10,000$

Figure 2.1: Quantile-Quantile plots for simulated null distributions of $1B$ allelic χ^2, Cochran-Armitage Trend, Wald, and Fisher's Exact tests, calculated for a fixed number of minor alleles among equal numbers of cases and controls.

Table 2.2: Expected and observed counts of significant tests at particular α-levels for logistic regression models featuring variable risk of a binary covariate to carriers of the minor allele, given for a fixed number of minor alleles among equal numbers of cases and controls.

Null Distribution					Significance Threshold (α)						
					$\alpha < 10^{-1}$	$\alpha < 10^{-2}$	$\alpha < 10^{-3}$	$\alpha < 10^{-4}$	$\alpha < 10^{-5}$	$\alpha < 10^{-6}$	$\alpha < 10^{-7}$
Theoretical χ^2 (1 $d.f.$)					100,000,000	10,000,000	1,000,000	100,000	10,000	1,000	100
	Trait Risk										
Model ($\sim$ Aff)	Carrier	Pop.	Count$_{MA}$	N							
$\beta_0 + \beta_1 Add$	·	·	40	2,000	79,179,748	6,167,797	534,239	5,983	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	10%	10%			79,782,126	6,180,674	472,267	6,065	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	20%	10%			84,495,240	6,409,079	409,365	6,276	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	40%	10%			95,357,905	7,262,721	371,535	7,344	5	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	60%	10%			96,940,157	7,536,009	397,235	9,753	28	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	80%	10%			97,342,986	7,757,342	438,797	13,473	126	0	0
$\beta_0 + \beta_1 Add$	·	·	80	10,000	92,269,671	9,538,127	439,954	24,480	2,215	65	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	10%	10%			92,271,582	9,529,042	501,357	34,820	2,224	65	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	20%	10%			92,781,985	9,407,137	610,113	40,053	2,156	65	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	40%	10%			97,423,726	8,784,754	659,591	41,297	1,935	68	2
$\beta_0 + \beta_1 Add + \beta_2 Cov$	60%	10%			98,325,325	8,741,198	668,054	42,540	1,993	80	1
$\beta_0 + \beta_1 Add + \beta_2 Cov$	80%	10%			98,425,924	8,793,612	680,545	44,148	2,197	93	1
$\beta_0 + \beta_1 Add$	·	·	40	10,000	80,400,565	6,368,746	178,816	7,745	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	10%	10%			80,405,394	6,368,716	198,507	7,733	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	20%	10%			80,790,880	6,372,733	260,393	7,543	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	40%	10%			86,962,981	6,598,294	343,261	6,396	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	60%	10%			93,681,516	7,082,939	358,649	6,107	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	80%	10%			96,110,140	7,343,465	365,042	6,841	1	0	0
$\beta_0 + \beta_1 Add$	·	·	20	10,000	115,137,310	2,561,766	0	0	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	10%	10%			114,999,456	2,561,820	0	0	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	20%	10%			114,872,450	2,565,090	0	0	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	40%	10%			113,182,826	2,674,582	0	0	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	60%	10%			108,060,024	3,061,018	0	0	0	0	0
$\beta_0 + \beta_1 Add + \beta_2 Cov$	80%	10%			101,803,263	3,564,315	0	0	0	0	0



(A) 40 minor alleles in $N = 2,000$

(B) 80 minor alleles in $N = 10,000$

(C) 40 minor alleles in $N = 10,000$

(D) 20 minor alleles in $N = 10,000$

Figure 2.2: Quantile-quantile plots for simulated null distributions of $1B$ Wald Test statistics for logistic models featuring variable risk of a binary covariate to carriers of the minor allele, and given for a fixed number of minor alleles among equal numbers of cases and controls.

2.5 Discussion

We have demonstrated the tendency of common tests of association to underestimate significance of less-common variants, highlighting the inadequacy of current analytical practices for dense-SNP and re-sequencing data. These results show convincingly that common approaches to multiple-test correction will be subject to inflated Type II error rates, particularly within the genome-wide significance levels. The sampling variance for a 1% allele does slightly improve the continuity of the asymptotic distribution, but does not preclude the deflated estimates of extreme p-values from this distribution.

Table 2.3 gives the number of permutations required to establish significance at various significance thresholds. At the 95% confidence level, our estimates are valid for $\alpha < 10^{-6}$, at which we see a considerable discrepancy between realized and expected test-statistic values for 20, 40, and 80 minor alleles. The required number of simulations to attain equivalent precision at current genome-wide α-levels is prohibitively large. However, this observed trend in distributional behavior is thoroughly convincing at increasingly stringent significance thresholds. Recent estimates of genome-wide α' are commonly of the order of 10^{-8}, and will undoubtedly become even smaller as larger numbers of rare variants are tested. With respect to what constitutes an appropriate correction for genome-wide studies, a reasonable assertion is that α' should reflect the total number of polymorphisms in the genome [48, 24]. Meaningful replication of novel findings demands that p-values be readily interpretable in the context of the entire catalogue of reported associations, and not subject to across-study differences in study design, sample size, or the number of SNPs actually assayed.

The appropriate choice of statistical test for analysis of rare variation is not entirely straightforward. Small samples are typically remedied by Yates correction [126] to the usual χ^2 formula. However, it is well-established that the corrected χ^2 yields a conservative estimate of significance [61], increasing the likelihood of observing a false negative finding. Alternatively, Fisher's exact test provides an exact estimate of significance for a given set of values within a contingency tables, and is an appropriate method when sample size is limited. Intrinsic differences between these approaches demand careful consideration, with non-negligible consequences for both study design and interpretation of findings. We caution readers against casual interpretation of exact tests across studies, and recommend that empirical significance for low-frequency variants be assessed by permutation.

2.6 Acknowledgements

We thank Mark Daly for useful feedback and comments on the manuscript.

Statistical analyses were carried out on the Genetic Cluster Computer (www.geneticcluster.org) which is financially supported by the Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam.

This research was supported by NIH/NIDA grants DA-18673 and DA-26119 and NIH/ARRA grant 1R01MH089208-01.

CHAPTER III

Association of Chromosome 20 Loci with Categorical Diagnoses and Clinical Dimensions of Schizophrenia in 270 Irish High-Density Families

T. Bernard Bigdeli[1,2], Brion S. Maher[1,2,3,4], Zhongming Zhao[1,2,3,5], Edwin J.C.G. van den Oord[2,6], Dawn L. Thiselton[2,3], Jingchun Sun[5], Bradley T. Webb[2,6], Richard L. Amdur[7,11], Brandon Wormley[2,3], Francis A. O'Neill[9], Dermot Walsh[10], Brien P. Riley[1,2,3], Kenneth S. Kendler[1,2,3], and Ayman H. Fanous[2,3,7,8,11]

[1] **Department of Human and Molecular Genetics, Virginia Commonwealth University,** Richmond, VA; [2] **Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University,** Richmond, Virginia; [3] **Department of Psychiatry, Virginia Commonwealth University,** Richmond, Virginia; [4] **Department of Mental Health, Johns Hopkins Bloomberg School of Public Health,** Baltimore, Maryland; [5] **Departments of Psychiatry, Biomedical Informatics, and Cancer Biology, Vanderbilt University Medical Center,** Vanderbilt, TN; [6] **Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University,** Richmond, Virginia; [7] **Mental Health Service Line, Washington VA Medical Center,** Washington, DC; [8] **Department of Psychiatry, Keck School of Medicine of the University of Southern California,** Los Angeles, CA; [9] **Department of Psychiatry, Queens University,** Belfast, UK; [10] **The Health Research Board,** Dublin, Ireland; [11] **Department of Psychiatry, Georgetown University School of Medicine,** Washington, DC

3.1 Abstract

Background: Prior genomewide scans of schizophrenia support evidence of linkage to regions of chromosome 20. However, association analyses have yet to provide support for any etiologically relevant variants.

Methods: We analyzed 2988 LD-tagging single nucleotide polymorphisms (SNPs) in 327 genes on chromosome 20, to test for association with schizophrenia in 270 Irish high-density families (ISHDSF, $N = 270$ families, 1408 subjects). These SNPs were genotyped using an Illumina iSelect genotyping array which employs the Infinium assay. Given a previous report of novel linkage with chromosome 20p using latent classes of psychotic illness in this sample, association analysis was also conducted for each of five factor-derived scores based on the Operational Criteria Checklist for Psychotic Illness (delusions, hallucinations, mania, depression, and negative symptoms). Tests of association were conducted using the PDTPHASE and QPDTPHASE packages of UNPHASED. Empirical estimates of gene-wise significance were obtained by adaptive permutation of a) the smallest observed P-value and b) the threshold-truncated product of P-values for each locus.

Results: While no single variant was significant after LD-corrected Bonferroni-correction, our gene-dropping analyses identified loci which exceeded empirical significance criteria for both gene-based tests. Namely, *R3HDML* and *C20orf39* are significantly associated with depressive symptoms of schizophrenia ($P_{emp} < 2 \times 10^{-5}$) based on the minimum P-value and truncated-product methods, respectively.

Conclusions: Using a gene-based approach to family-based association, *R3HDML* and *C20orf39* were found to be significantly associated with clinical dimensions of schizophrenia. These findings demonstrate the efficacy of gene-based analysis and support previous evidence that chromosome 20 may harbor schizophrenia susceptibility or modifier loci.

3.2 Introduction

With a lifetime prevalence of 1 percent and an estimated annual cost of $62.7 billion in the United States [123], schizophrenia (Scz) is a debilitating neuropsychiatric disorder which poses a significant burden to public health. Whether schizophrenia represents a single or multiple disease processes is a source of persistent controversy, as patients vary considerably in onset, course and outcome of disease, and the particular combination of symptoms endorsed [28, 29]. Models comprising continuous traitsoften extracted in factor analysis of symptom profileshave been adduced, typically distinguishing positive, negative, disorganization, and affective symptoms [92]. One explanation for this variability lies in the existence of more than one putative etiopathogenic mechanism, each imparting susceptibility to a more or less distinct disease subtype or influencing the character of illness dimensionally. Detection and subsequent replication of several putative risk variants, facilitated by genome-wide association studies (GWAS) [87, 113, 82, 96, 104, 111], has seen renewed interest in this question among geneticists and diagnosticians alike [15, 16, 17].

Consistent with the observed variability in clinical presentation is the hypothesis that schizophrenia is likely genetically heterogeneous [113, 96]. Linkage and candidate gene association studies have implicated a number of genes and genomic regions, with varying degrees of subsequent independent replication. Allelic heterogeneity has been demonstrated in meta-analyses of candidate genes such as *DTNBP1* [77, 64]. If the observed clinical heterogeneity of schizophrenia is in fact due to genetic heterogeneity, the use of more clinically homogenous phenotypes may increase the signal-to-noise ratio in gene-finding studies. A previous report by our group described detection of novel linkage to 20p using latent classes of psychotic illness [31]. Linkage analysis of Mania, Schizomania, Deficit Syndrome and Core Schizophrenia latent classes yielded several suggestively significant loci, in regions of chromosome 20 which had previously yielded very little evidence of linkage in our sample. Furthermore, the presence of susceptibility genes in chromosome 20 has been suggested by several previous linkage studies as well [14, 73, 40, 58, 121, 3, 115]. In addition to genes which increase susceptibility to more or less distinct clinical subtypes of illness, other genes may influence clinical features of disease in a dimensional fashion, without altering liability to the illness itself. These have previously been described as modifier loci [28]. Modifier loci may not be resolvable using traditional dichotomous phenotypes (simply affected or unaffected), but rather, by quantitative symptomatic measures. Several examples have been reported [66, 54, 102, 129, 103, 27, 30].

Recent GWAS of schizophrenia support a polygenic model in which potentially thousands of common variants individually impart small effects. Given the unprecedented multiple-comparison burden incurred in a genome-wide approach, hypothesis-based strategies remain viable alternatives for the study of complex disease. A gene-based approach is particularly convenient. In an analysis of bipolar and schizophrenia datasets, Moskvina and colleagues [76] observed significantly more SNPs within genes showing evidence for association than expected, with intergenic SNPs showing no such trend. We describe a comprehensive, gene-based association survey of 327 genes in regions linked to chromosome 20 in our previous studies. In addition to testing for association with traditional diagnostic definitions of schizophrenia, we also sought to assess whether chromosome 20 harbors modifier loci. Association analysis was therefore also performed for five factor-derived scores, representing hallucinations, delusions, depressive symptoms, manic symptoms, and negative symptoms, in schizophrenia cases only. In addition to single-marker tests of allelic association, we employ two gene-based test-statistics, the minimum observed P-value per gene and the truncated product of P-values, to evaluate the efficacy of a gene-based approach as applied to a large, family-based study.

3.3 Methods

Ethics Statement This research was approved by the Institutional Review Boards of Virginia Commonwealth University School of Medicine and the Washington VA Medical Center. All subjects gave verbal assent to participate in research, as this was the norm in Ireland at the time these data were collected.

3.3.1 Sample

Fieldwork for the Irish Study of High Density Schizophrenia Families (ISHDSF) was conducted between April 1987 and November 1992, with probands ascertained from public psychiatric hospitals in Ireland and Northern Ireland [56]. Selection criteria were two or more first-degree relatives meeting DSM-III-R criteria for schizophrenia or poor-outcome schizoaffective disorder (PO-SAD). Diagnoses were based on the Structured Interview for DSM-III-R Diagnosis (SCID) [109]. Independent review of all pertinent diagnostic information was made blind to pedigree assignment and marker genotypes by KSK and DW, with each diagnostician making up to three best-estimate DSM-III-R diagnoses. The Operational Criteria Checklist for Psychotic Illness (OPCRIT) [70] was completed by KSK for all subjects with probable lifetime histories of hallucinations or delusions ($N = 755$; $N = 722$ genotyped). Our diagnostic schema contains 4 concentric definitions of affection: narrow (D2) (schizophrenia, PO-SAD, and simple schizophrenia) ($N = 577$), intermediate (D5) which adds to D2 schizotypal personality disorder, schizophreniform and delusional disorders, atypical psychosis and good-outcome SAD ($N = 700$), broad (D8) (all disorders which significantly aggregated in relatives of probands) ($N = 754$) and very broad (D9), including any psychiatric illness ($N = 961$). Exploratory and confirmatory factor analysis of the OPCRIT was conducted previously by Fanous et al. [32]. This yielded a five-factor solution, comprising depressive, manic, and negative symptoms, delusions and hallucinations. Factor-derived scores were obtained by summing the scores of all items belonging to each factor.

3.3.2 Bioninformatics and SNP-selection

Using WebGestalt [128], a total of 378 genes were initially identified as mapping to the region of chromosome 20 corresponding to the peak NPL and to the positions corresponding to a NPL of at least 1 on either side, based on the Illumina version 4.0 linkage SNP map used for genotyping in a multicenter linkage study funded by R01-MH068881 [49]. While there was very little evidence of linkage in our published microsatellite-based scan [112], we did observe modest evidence using the map in the Holmans et al. study, which included our study sample (results available on request). We included predicted genes and open reading frames (ORFs) from the p-terminal to 45.85 Mb (20q13.13). Physical map positions for 362 genes were obtained from the UCSC Genome Browser (hg17/NCBI Build 35) [55]. Tagging SNPs were selected for each identified genomic region (excluding upstream and downstream regions of genes) using Tagger ($r^2 \geq 0.8$, minor allele frequency (MAF) ≥ 0.1) [18], as applied to the HapMap CEPH dataset [52]. Of these, 31 genes were excluded on the basis of tagging SNPs being unavailable. After removing multiple occurrences of markers resulting from overlap of adjacent genomic regions, 3,386 SNPs in 331 genes were selected for inclusion.

3.3.3 Genotyping

Genotyping was conducted by Illumina, Inc. using a custom iSelect array, which employs the Infinium assay. In total, DNA for 1,128 individuals was submitted for genotyping of 3,386 SNPs. As SNP markers from several ongoing experiments were included on the same array, per-individual summary statistics reflect genotyping across a total of 7,500 SNPs. Average genotyping completion rate across all SNPs was 99.97%. Of 1,128 samples, 21 failed to yield usable genotypes. Genotypes were examined for apparent Mendelian incompatibilities using PEDCHECK v 1.1 [86] and removed for entire families where appropriate.

3.3.4 Association Analysis

We performed association analysis for categorical diagnoses of schizophrenia using PDTPHASE (UNPHASED v. 2.404), an implementation of the pedigree disequilibrium test (PDT) with extensions to deal with uncertain haplotypes and missing data [67, 23]. The PDT is an extension of the transmission disequilibrium test (TDT) to examine general pedigree structures and is similarly a test of association in the presence of linkage. Association with quantitative measures of disease was assessed using QPDTPHASE (UNPHASED v. 2.404), an implementation of the quantitative trait PDT with extensions to deal with uncertain haplotypes and missing data [23, 74]. An LD-corrected significance threshold was obtained using the SNPSpD package for R [85, 97]. For 2,988 SNPs, SNPSpD calculated an estimated 1,569 independent tests, with a corresponding significance threshold of $\alpha_{SNPSpD} \approx 3.18 \times 10^{-5}$, maintaining the type I error rate at 5%.

3.3.5 Gene-wide Tests of Empirical Significance

Estimates of empirical significance for association results were obtained by adaptive permutation of gene-dropping simulations created with MERLIN [1]. Simulated genotypes were of identical frequency, marker spacing, and pattern of missing data as the actual genotypes, with individual phenotypes and pedigree structure also preserved within each simulated dataset. For markers in linkage disequilibrium ($r^2 \geq 0.1$), alleles were simulated using the haplotype frequencies for the marker clusters. To reduce computation time, those pedigrees of complexity greater than 70 bits were omitted from calculation of allele and haplotype frequencies. Each simulated dataset was analyzed as described above in two ways: retaining the minimum

P-value per gene, as well as the calculating the threshold-truncated product of P-values ($\alpha_{trunc} \leq 0.01$) per gene. For the set of single-SNP hypotheses corresponding to a gene, the truncated product method considers the product of only those P-values falling below a specified threshold, evaluating the probability of observing as significant a product by chance. Whereas Fishers Combined Test assesses the overall evidence for departure from the null, the truncated product approach can be used to assess whether suggestive or significant findings are truly significant [127]. Previous reports support the use of a truncated product approach in conjunction with the PDT [45]. Empirical significance was calculated from the proportion of simulated gene-wise test statistics more significant than the actual results ($r_{obs}+1/n_{perm}+1$). We used an adaptive permutation procedure, by which empirical P-values were obtained for 100, 1,000, 10,000, and 100,000 simulations. Only those observed associations for which there were not at least ten more significant simulated results were carried forward to each successive stage of permutation analysis.

3.4 Results

3.4.1 Gene-wide Association Analyses

Following quality-control protocols, 2,988 single nucleotide polymorphisms in 327 genes were tested for association with a diagnosis of schizophrenia (Figure 3.1). Estimates of empiric significance (P_{emp}) were obtained via an adaptive permutation procedure employing the smallest observed P-value, as well as the truncated product of P-values ($\alpha_{trunc} \leq 0.01$) per gene. The number of genes carried forward in successive stages of this procedure, in both approaches, can be found in Table 3.1. Using the minimum gene-wide P-value approach, no genes were observed to be significantly associated with narrow ($N = 1574$), intermediate ($N = 1749$), or broad ($N = 1808$) diagnoses of schizophrenia. Next, we sought to identify those SNPs associated with clinical dimensions of schizophrenia in a subset of cases ($N = 721$) for which the OPCRIT was available. A previous report by Fanous and colleagues [31] supports linkage of latent classes derived from the OPCRIT to chromosome 20 in this sample. No genes were found to be significantly associated with the negative, manic, hallucinations or delusions factors. In the analysis of the clinical dimensions, *R3HDML* demonstrated significant evidence of association ($P_{emp} < 2 \times 10^{-5}$) with the depressive factor using the minimum P-value approach. Using the truncated product of P-values, *C20orf39* was also found to be significantly associated with the depressive factor ($P_{emp} < 2 \times 10^{-5}$). It is important to note that, for both *C20orf39* and *R3HDML*, we observed fewer than ten simulated results more significant than the observed test-statistic after 100,000 permutations. Hence, our estimates of empirical significance may be conservative. However, extending our analyses to $1M$ permutations was not carried out as it was too computationally demanding. Because validation of a truncated product approach in extended pedigrees relies on the permutation procedure faithfully conserving patterns of LD within each replicate dataset, we obtained a quantitative measure of how well haplotype-block structure was maintained for *C20orf39* across actual and simulated datasets. In calculating an LD-corrected significance threshold, SNPSpD estimates the effective number of independent tests present in a set of markers. Using SNPSpD, 1,000 replicate datasets for *C20orf39* were assessed for number of independent tests. When compared to the estimate based on the actual pattern of LD in *C20orf39* (i.e., 26 independent tests), the distribution of these simulation-derived estimates demonstrates that the LD structure within each replicate does not differ significantly from the observed data ($P \approx 0.409$; 95% CI: [26,28]). This increases confidence in the truncated product finding for *C20orf39*. However, this may not hold for every gene and may be sensitive to specific patterns of linkage disequilibrium.

3.4.2 Single Marker Association Analysis

Taking each SNP to represent an independent hypothesis but correcting for LD using SNPsPD, we found that no single marker met experiment-wide criteria for association ($\alpha_{SNPsPD} < 3.18 \times 10^{-5}$) with either the three categorical diagnostic definitions used or our OPCRIT-derived factor scores (Tables 3.2, 3.3). The strongest evidence of association with a diagnosis of schizophrenia was in PLCB1 (20p12.3) (rs6108205, $P \approx 1.00 \times 10^{-3}$, intermediate Scz diagnosis). For the depressive factor, we observed the strongest associations experiment-wide at 20q13.12 (rs3761184, $P \approx 3.31 \times 10^{-5}$) in *R3HDML* This was very close to the LD-corrected significance threshold calculated using SNPSpD ($P = 3.18 \times 10^{-5}$). Furthermore, rs11700002, in *C20orf39* at 20p11.21 attained $P \approx 1.01 \times 10^{-4}$.

Table 3.1: **Number of genes requiring additional simulations at each stage of adaptive permutation.**

	Trait	Min P (100)	Min P (1K)	Min P (10K)	Min P (100K)	Trunc Prod P (100)	Trunc Prod P (1K)	Trunc Prod P (10K)	Trunc Prod P (100K)
Diagnostic Category	Narrow	19	5	0	0	9	5	0	0
Diagnostic Category	Int	12	4	0	0	6	4	2	0
Diagnostic Category	Broad	14	3	0	0	5	3	0	0
Symptom Factor	del	15	4	0	0	8	4	0	0
Symptom Factor	dep	10	6	1	$1^{\dagger}$	9	4	2	$1^{\ddagger}$
Symptom Factor	hal	6	2	0	0	4	1	0	0
Symptom Factor	manic	5	2	0	0	14	6	0	0
Symptom Factor	neg	20	4	0	0	14	7	0	0

For each diagnosis and symptom factor, the number of loci requiring additional permutations after each stage of our adaptive procedure, given for both the Min P and truncated-product methods. For an observed test-statistic to be considered significant at a particular stage of permutation, there may be no greater than 10 simulated null statistics which are more extreme than the observed. For each gene-based test, the number of permutations performed at each stage is displayed parenthetically (100, 1,000, 10,000, and 100,000). *Dim* codes “del”, “dep”, “hal”, “manic”, “neg” are delusions, depressive symptoms, hallucinations, mania, and negative symptoms, respectively, and described elsewhere in full.

†*Min P finding for R3HDML.*

‡*truncated-product finding for C20orf39.*

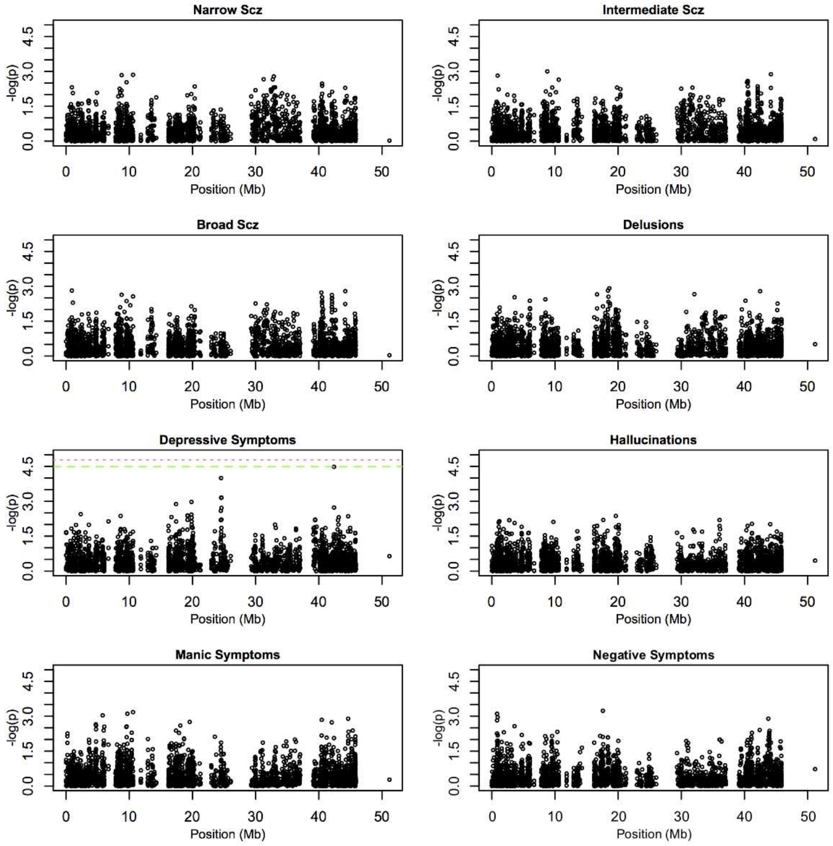


Figure 3.1: **Physical distributions of single-marker associations on chromosome 20, for both categorical diagnoses and clinical dimensions of Scz.** Associations are displayed as log-transformed P-values ($-\log_{10}P$) at genomic positions in megabases (Mb). Where appropriate, a dotted line indicates the Bonferroni-corrected significance threshold, accounting for number of SNPs assayed experiment-wide. Similarly, a dashed line indicates the LD-corrected significance threshold, as estimated by SNPSpD.

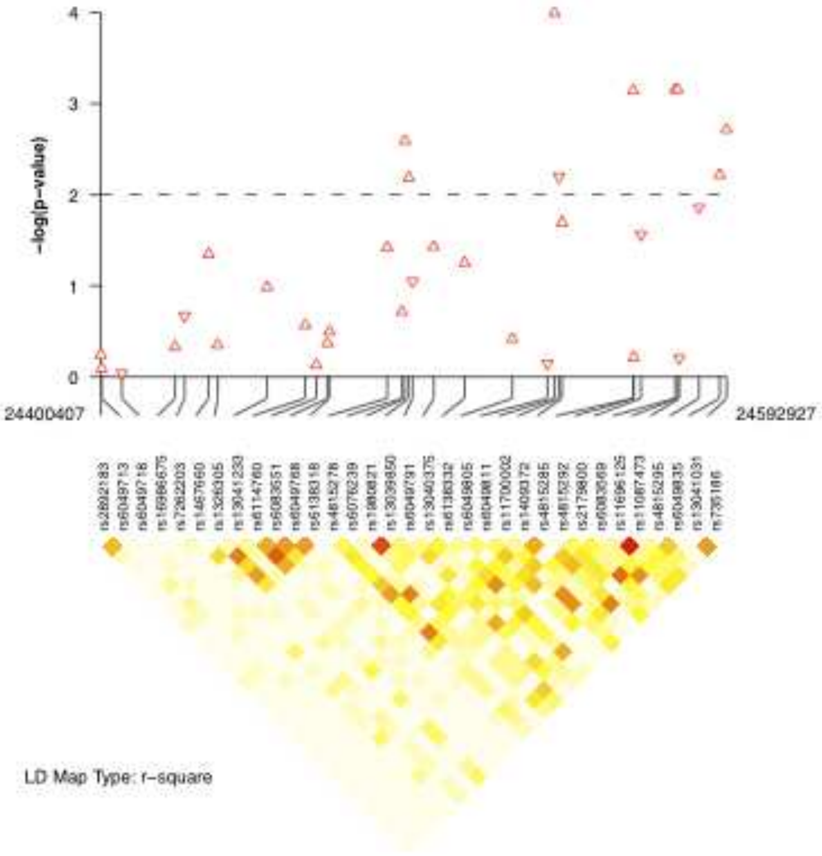


Figure 3.2: **Association of *C20orf39* SNPs with depressive symptoms of Scz.** Magnitudes and directions of associations are displayed in the upper panel, with upwards-oriented triangles indicating a positive correlation with symptom factor score. A dashed line is provided at the inclusion threshold for the truncated product of P-values. Connecting lines relate the physical positions of associations to SNP labels in the corresponding LD-map (r^2). Plot generated using snp.plotter for R [62].

Table 3.2: **Top ten (10) Pedigree Disequilibrium Test results for categorical diagnoses of Schizophrenia in the ISHDSF.**

Chr/Mb	Gene	dbSNP	Nuc.(*Min*)	Assoc.	Frq$_{assoc}$	Trios(Tr/NTr)	DSPs(Aff/Un)	Z	χ^2	P	Dx
20/0.92	*RSPO4*	rs6056462	A/G(G)	A	0.842	94 / 90	868 / 855	3.171	10.05	1.52×10^{-3}	Int
20/0.92	*RSPO4*	rs6056462	A/G(G)	A	0.843	97 / 93	911 / 897	3.170	10.05	1.53×10^{-3}	Broad
20/8.79	*PLCB1*	rs6108205	C/T(C)	C	0.491	67 / 65	536 / 467	3.289	10.82	1.00×10^{-3}	Int
20/8.79	*PLCB1*	rs6108205	C/T(C)	C	0.490	58 / 56	444 / 391	3.182	10.13	1.46×10^{-3}	Narrow
20/10.60	*JAG1*	rs6133987	C/T(T)	C	0.778	85 / 78	678 / 610	3.193	10.19	1.41×10^{-3}	Narrow
20/40.41	*PTPRT*	rs6072690	A/G(A)	A	0.442	67 / 59	586 / 502	3.224	10.40	1.26×10^{-3}	Broad
20/40.51	*PTPRT*	rs6130134	C/T(T)	C	0.792	94 / 93	880 / 826	3.152	9.940	1.62×10^{-3}	Int
20/42.07	*TOX2*	rs6103560	C/T(T)	C	0.705	111 / 95	855 / 822	3.200	10.24	1.37×10^{-3}	Broad
20/44.19	*CD40*	rs3765457	A/G(G)	G	0.175	24 / 24	235 / 170	3.154	9.950	1.33×10^{-3}	Broad
20/44.19	*CD40*	rs3765457	A/G(G)	G	0.176	23 / 23	223 / 160	3.210	10.30	1.61×10^{-3}	Int

For each gene, *Chr/Mb* denotes chromosome and genomic position (Megabases), *dbSNP* is the rs-identifier for the assayed SNP, and *Nuc* is the nucleotide substitution at a SNP. *Frq$_{assoc}$* and *Z*-scores are with respect to the associated allele. Allelic transmissions from parent to affected child is given by *Trios*, where *Tr* and *NTr* represent number of transmissions and non-transmissions. Allele-sharing between phenotypically-discordant sib-pairs is given by *DSPs*, with *Aff/Un* denoting the number of associated alleles in affected and unaffected siblings, respectively. *Dx* codes "Core", "Int", and "Broad" are Core, Intermediate, and Broad diagnoses of schizophrenia, respectively, and are described elsewhere in full.

Table 3.3: **Top ten (10) Quantitative Pedigree Disequilibrium Test results for clinical dimensions of Schizophrenia in the ISHDSF.**

Chr/Mb	Gene	dbSNP	Nuc.(*Min*)	Assoc.	Frq$_{assoc}$	Gametes (maj/min)	Z	χ^2	P	*Dim*
20/0.83	*ANGPT4*	rs976166	C/G(C)	C	0.308	1285/571	3.333	11.11	8.59×10^{-4}	neg
20/9.70	*PAK7*	rs2327225	A/C(C)	A	0.718	1333/523	3.426	11.74	6.13×10^{-4}	manic
20/10.59	*JAG1*	rs6133986	A/G(A)	A	0.089	1691/165	3.523	12.41	4.27×10^{-4}	manic
20/16.66	*SNRPB2*	rs6111262	C/T(T)	C	0.720	1337/519	3.281	10.76	1.04×10^{-3}	del
20/17.58	*RRBP1*	rs3790310	A/T(A)	T	0.809	1502/354	3.451	11.91	5.58×10^{-4}	neg
20/24.54	*C20orf39*	rs11700002	A/G(A)	A	0.187	1509/347	3.930	15.44	8.50×10^{-5}	dep
20/24.56	*C20orf39*	rs4815292	G/T(G)	G	0.353	1201/655	3.362	11.30	7.73×10^{-4}	dep
20/24.58	*C20orf39*	rs11696125	G/T(T)	T	0.230	1430/426	3.392	11.50	6.94×10^{-4}	dep
20/24.58	*C20orf39*	rs11087473	A/G(A)	A	0.230	1430/426	3.392	11.50	6.94×10^{-4}	dep
20/42.40	*R3HDML*	rs3761184	A/G(G)	G	0.169	1543/313	4.120	16.98	3.78×10^{-5}	dep

For each gene, *Chr/Mb* denotes chromosome and genomic position (Megabases), *dbSNP* is the rs-identifier for the assayed SNP, and *Nuc* is the nucleotide substitution at a SNP. *Gametes* represents the number of major and minor alleles (*maj/min*) transmitted from parent to affected child or unshared between phenotypically-discordant sib-pairs. *Frq$_{assoc}$* and Z-scores are with respect to the allele corresponding to a higher trait mean. *Dim* codes "del", "dep", "hal", "manic", "neg" are delusions, depressive symptoms, hallucinations, mania, and negative symptoms, respectively, and described elsewhere in full.

3.5 Discussion

We have conducted a comprehensive gene-based association study of 327 genes on chromosome 20 in an Irish sample of 270 high-density schizophrenia families. This study sought to identify common variants conferring susceptibility to schizophrenia, following up reported linkage in this sample to clinical subtypes of psychotic illness [31], as well as previous studies reporting linkage to chromosome 20. Because those clinical subtypes were derived from quantitative symptom dimensions, we also tested for association with these same dimensions. Although traditional single-marker tests failed to identify any SNPs meeting experiment-wide criteria for significance, application of gene-wide association metrics revealed two previously unimplicated loci, *R3HDML* and *C20orf39*, associated with depressive symptoms. Our findings support the power of gene-based association approaches. They also lend further support to previous evidence suggesting that genetic differences may underlie clinical heterogeneity in schizophrenia [28, 29]. One of the aims of this study was to identify genomic loci predisposing to a particular form of illness or which modifies clinical presentation amongst affected individuals. Such genes have been described previously as modifier or susceptibility-modifier loci and are reviewed elsewhere [28]. Of the two loci showing the strongest associations, namely *R3HDML* and *C20orf39*, neither appears to affect the risk of the illness itself. That is, no single variant in either gene met even nominal significance criteria ($P < 0.05$) for association with narrow, intermediate, or broad diagnoses of schizophrenia. These two genes would therefore fulfill our definition of modifier genes [28]. However, the strength of evidence we observed for *R3HDML* is greater than that observed for *C20orf39*. *R3HDML* was identified by application of the minimum P-value approach. Among affected individuals, those carrying the minor allele (G) of the corresponding SNP, rs3761184, had higher mean depression scores. On the other hand, for *C20orf39*, empirical significance was attained using the truncated product of P-values. This makes it more difficult to identify a specific risk genotype. This is because the truncated product method only considers all variation within a gene jointly.. In Figure 3.2, it is apparent that those markers contributing to the truncated product for *C20orf39* comprise a block of LD distinct from the surrounding region, with the majority showing association of the minor allele with higher depression scores. Whereas individually, none of the single-marker associations were significant after our permutation procedure, the degree of correlation between the SNPs may have been sufficient to produce an empirically significant association for *C20orf39* as a whole. In order to rule out a spurious gene-wise association due to higher LD, we analyzed a set of permutations using SNPSpD, then compared the distribution of estimated number of independent tests (SNPs) to that obtained for the actual data. If our gene-dropping simulations were found to consistently underestimate the extent of LD between adjacent markersindicated by a larger number of independent testswe would expect an inflation of the empiric test-statistic. Alternatively, if the observed LD within simulated datasets tended to overestimate pairwise LD, the corresponding distribution of truncated products would underestimate the empiric test-statistic. For *C20orf39*, the observed SNPSpD estimate of ~ 26 tests was not found to differ significantly from the null distribution of simulated datasets, suggesting that our gene-dropping procedure was faithfully conserving LD-structure across our simulations. As discussed, increased gene-size, especially in the presence of higher LD between markers, might also contribute to over-estimation of the test statistic. To our knowledge, neither *R3HDML* nor *C20orf39* has been functionally characterized to date. Both are predicted genes identified on the basis of domain homology. The *R3HDML* locus encodes a putative serine protease inhibitor belonging to the CRISP family of cysteine-rich secretory proteins, and contains evolutionarily conserved exonic and intronic regions bearing greater than 90% similarity to Rhesus macaque [89]. Interspersed within the conserved intronic sequences are numerous stretches of simple tandem repeats (e.g. CG_n). Our SNP of interest in *R3HDML*, rs3761184, falls just upstream (< 50 bp) of the second exon and 150 bp downstream of one such repeat-rich region. Roles in fertilization, spermatogenesis, and pathogen response have all been proposed for CRISP proteins, but these mechanisms are not immediately supportive of *R3HDML* as a schizophrenia candidate gene. However, recent implication of a number of HLA genes in large-scale GWAS suggest that genes involved in immune-related mechanisms, such as pathogen response, could be reasonable Scz candidates [96]. The presence of specific sequence features in the vicinity of the associated SNP may warrant more thorough bioinformatic inquiry. Additionally, *R3HDML* lies approximately 57 kb downstream of the GDAP1L1 locus, which appears to encode a gluthionine S-transferase (GST). Cell-culture studies have demonstrated a relationship between gluthionine deficiency and oxidative stress, mechanisms frequently purported to contribute to schizophrenia pathophysiology [42, 105]. However, *GDAP1L1* was not significantly associated. Our empirically significant finding for *C20orf39* presents additional challenges for interpretation, given its provisional status as an open reading frame. Provisionally known as *TMEM90B*, this locus encodes a predicted transmembrane protein. Of 33 SNPs assayed within *C20orf39*, the nine included in the truncated product bounded a region of LD corresponding to the coding region of *C20orf39*. The upstream, untranslated region of *C20orf39*, which itself corresponds to a distinct set of ESTs, yielded no SNPs meeting local significance criteria. Whether the markers driving this association simply lie in joint linkage disequilibrium with nearby causal variation, or actually demarcate an etiologically relevant genomic region, is unknown. Depressive symptoms, especially suicidal ideation, comprise a considerable portion of morbidity and mortality in schizophrenia [47]. Therefore, follow up of these two genes could be important in the search for clues to more successful identification and treatment of this clinical dimension. As demonstrated by Moskvina et al., polymorphisms mapping to functional elements are more likely to be associated with complex disease than intergenic variation [76].

Despite ongoing annotation and characterization of functional elements, however, our knowledge of genomic variation, functional or otherwise, remains incomplete. This is exemplified by *C20orf39* and *R3HDML*, which are novel and unannotated. A major benefit of gene-based approaches is that they are robust to allelic and haplotypic heterogeneity across samples. This makes them particularly suited for use in replication and meta-analysis. In traditional replication of single-marker associations, the associated SNP in the discovery sample is usually assayed in all subsequent replication samples. This could inflate Type-II error in the presence of population differences in haplotype structure and allele frequencies [81]. Complex patterns of associations, whether spurious or due to genetic heterogeneity, have been more the rule rather than the exception in candidate gene studies of complex disease, as demonstrated by studies of *DTNBP1* [77, 64]. For discovery-based approaches, adoption of a gene-based strategy may be of even more immediate benefit, specifically by providing a straightforward means of multiple-test correction. Furthermore, traditional methods to correct for multiple-testing, such as Bonferroni correction or the less overtly conservative SNPSpD method, may be less robust in detecting small genetic effects. However, in spite of the advantages of gene-based association studies intergenic causative variants or variants in unrecognized genes might have been missed in this study. Given the poor spatial resolution of linkage and intrinsic differences between these methodologies, we are currently unable to fully relate our association findings with the results of our previously published linkage study of latent classes. However, it is notable that *R3HDML* is located in a region which was linked to the deficit syndrome latent class, for which members were substantially more likely to fall below the median for depressive symptoms. Despite failing to demonstrate any evidence of association with a diagnosis of schizophrenia, *R3HDML* may be associated with a disease subtype characterized by low levels of depression. Because subtyping precludes use of our full sample for association analysis, statistical power is insufficient to test this hypothesis.

Other methods aiming to identify more clinically homogenous subgroups have been applied to linkage analysis of schizophrenia. In a study of 168 affected sibling pairs, Hamshere and colleagues [43] demonstrated that inclusion of major depression as a covariate yielded suggestive evidence of linkage at 20q11.21, while schizophrenia as a whole did not. Taken together, these studies are compelling in their support of 20q11 harboring genes relevant to the affective component of schizophrenia. Emerging evidence supports a role for genetic variants conferring risk of both schizophrenia and bipolar disorder [96, 60]. Furthermore, genome scans of both disorders have consistently implicated regions of chromosome 20 [21, 71, 122, 26, 37, 88]. A recent study of 383 bipolar or schizoaffective relative pairs found suggestive linkage at 20q13.31 when conditioning on the presence of mood-incongruent psychosis, furthering the argument that chromosome 20 loci may have relevance to conditions containing admixtures of mood and psychotic symptoms [44].

The findings presented here provide additional support to published findings suggesting that schizophrenia modifier loci may exist on chromosome 20 and, more generally, that genetic differences underlie clinical heterogeneity in schizophrenia [107]. We await replication of the observed associations between these loci and either categorically defined illness or more or less distinct subtypes or clinical dimensions. There are two main limitations relevant to this study. First, the truncated product of P-values is particularly sensitive to patterns of LD (unpublished results), since markers could be significant only due to their LD with other significant markers. Applied to family-based analysis of extended pedigrees, the validity of gene-based testing relies on the permutation method realistically maintaining LD across simulated datasets. As discussed, for *C20orf39*, the LD structure for a random sample of simulated datasets did not differ significantly from the actual data ($P > 0.05$). Second, our analysis of multiple symptom dimensions may increase the Type-I error rate due to multiple testing. However, as we have previously shown, these dimensions are correlated [32], making Bonferroni correction overly conservative. It remains unclear whether the failure of traditional approaches to detect experiment-wide significant loci reflects the spurious nature of these findings or simply the limited power of this sample. Ultimately, the genotype-phenotype correlations reported herein require confirmation in independent samples for which comparable symptom measures are available. We are unaware of other family-based schizophrenia samples in which OPCRIT data are readily available. However, this is likely to be attempted in case-control samples by the Psychiatric GWAS Consortium Cross-Disorders Group [17].

3.6 Acknowledgments

We are grateful to the patients and their families for their generous participation in these studies. TBB was supported by US National Institutes of Health grant 5R25DA026119-03. BSM, BPR and KSK were supported by US National Institute of Mental Health grant MH083094 and BPR and KSK were supported by grants MH041953 and MH068881. ZZ was supported by US National Institutes of Health grant AA017437 and a NARSAD Maltz Investigator Award. AHF was supported by a grant from the Department of Veterans Affairs Merit Review Program.

CHAPTER IV

Whole-Genome *In Silico* Genotyping and Association Study of the Irish Study of High-Density Schizophrenia Families (ISHDSF)

T. Bernard Bigdeli[1,2] *et al.*

[1]**Department of Human and Molecular Genetics,Virginia Commonwealth University,** Richmond, VA

[2]**Virginia Institute for Psychiatric and Behavioral Genetics Virginia Commonwealth University,** Richmond, VA

4.1 Abstract

Background: Recent WGAS of Schizophrenia have identified major susceptibility loci on 2q32.1, 6p21.3-22.1, 11q24.2, and 18q21.2 but the majority of the genetic variance in disease risk remains unaccounted for.

Methods: The initial sample consisted of $N = 843$ individuals from 234 Irish high-density families (ISHDSF) genotyped on the Illumina 610-Quad platform (557,373 SNPs) and $N = 349$ additional, related subjects genotyped on the Illumina version 4 linkage panel. We conducted *in silico* genotyping to infer WGAS data for sparsely-genotyped subjects using MERLIN (v. 1.1.2). A total of 535,728 SNPs were tested for association with Schizophrenia using the generalized disequilibrium test (GDT). An LD-corrected threshold for experiment-wide significance was obtained using SNPSpD.

Results: We successfully inferred WGAS data for $N = 206$ subjects by *in silico* genotyping. An association at 1q32.1 between narrowly-defined Scz and *PPP1R12B* (rs12734001, $P < 1.2{\times}10^{-7}$) was significanct after LD-corrected Bonferroni-correction.

Conclusions: Using an approach to family-based association which considers all phenotypically-discordant relatives pairs, a SNP in *PPP1R12B* was found to be significantly associated with categorical dimensions of schizophrenia. These findings (1) support the presence of a Schizophrenia susceptibility locus in the vicinity of chromosome 1q32 and (2) demonstrate efficient WGA analysis of multiply-affected pedigrees.

4.2 Introduction

We present herein findings and implications of a whole-genome association study (WGAS) of Schizophrenia (Scz) in 234 Irish high-density Schizophrenia families. That the etiologies of Scz and its spectrum disorders are multifactorial is widely-recognized, as are the high degree of heritability (∼80%) and substantial familiality exhibited by these disorders. Despite an abundance of empirical evidence supporting a fundamental role for genetic factors in Scz pathophysiology, this devastating neuropsychiatric disorder has remained largely recalcitrant towards efforts to identify major susceptibility loci. In recent years, large population-based WGAS have identified several strong associations between genetic loci and Scz, including the major histocompatibility complex (MHC), *ZNF804A*, *TCF4*, and *NRGN*. However, the larger part of the total genetic variance in liability to Scz remains unaccounted for. Possible explanations for the limited progress in delineating the etiology of Scz include its extensive clinical heterogeneity, the inherent complexity of the biological systems and tissues involved, and the possibility of genetic heterogeneity among individual cases.

The results of recent WGAS are appropriately interpreted as supporting a common polygenic model—but of considerably greater multiplicity than previously thought. However, the presence of even moderate genetic heterogeneity is certain to be accompanied by some loss in power to detect individual effects. It follows that some number of "true" effects will go undetected by large, population-based studies of Scz employing traditional, single-marker approaches to association.

The present study, despite employing a relatively-underpowered sample originally intended for linkage analysis, does present two key advantages in this respect. First, recruitment of probands was on the basis of membership in multiply-affected pedigrees; such "high-density" families are conceivably enriched for large genetic effects. Second, the combined sample is relatively un-diverse ethnically, as evidenced by a considerable degree of cryptic relatedness between otherwise unrelated subjects. Using whole-genome SNP data, available for a subset of association-informative subjects, we successfully infer WG data in additional, related subjects typed previously on a low-density SNP panel for linkage. To maximize the amount of information extracted from each pedigree, we utilize a "within-family" approach which considers all phenotypically-discordant relative-pairs to evaluate the evidence of association between common SNPs and categorical diagnoses of Scz.

4.3 Methods

4.3.1 Samples

Fieldwork for the Irish Study of High Density Schizophrenia Families (ISHDSF) was conducted between April 1987 and November 1992, with probands ascertained from public psychiatric hospitals in Ireland and Northern Ireland [56]. Selection criteria were two or more first-degree relatives meeting DSM-III-R criteria for schizophrenia or poor-outcome schizoaffective disorder (PO-SAD). Diagnoses were based on the Structured Interview for DSM-III-R Diagnosis (SCID) [109]. Independent review of all pertinent diagnostic information was made blind to pedigree assignment and marker genotypes by KSK and DW, with each diagnostician making up to three best-estimate DSM-III-R diagnoses. The Operational Criteria Checklist for Psychotic Illness (OPCRIT) [70] was completed by KSK for all subjects with probable lifetime histories of hallucinations or delusions ($N = 755$). Our diagnostic schema contains 4 concentric definitions of affection: narrow (D2) (schizophrenia, PO-SAD, and simple schizophrenia) ($N = 577$), intermediate (D5) which adds to D2 schizotypal personality disorder, schizophreniform and delusional disorders, atypical psychosis and good-outcome SAD ($N = 700$), broad (D8) (all disorders which significantly aggregated in relatives of probands) ($N = 754$) and very broad (D9), including any psychiatric illness ($N = 961$).

4.3.2 Genotyping

In total, 853 individuals representing 237 high-density schizophrenia families were selected for high-throughput genotyping on the Illumina 610-Quad platform, with the selection of particular persons from extended pedigrees based on informativeness of their genetic relationships for association analyses. Moreover, this strategy considered those samplings of family members which optimized potential for application of *in silico* [6] genotyping methods to additional family members for whom less-dense genotyping data was available, namely those selected for a previous genome scan [49]. Following lift-over to the most recent genome assembly (GRCh37.2), 557,373 autosomal SNPs were available for analysis; genotyping completion was greater than 99.9%.

Genotyping for the previous genome scan, described by Holmans *et al.* (2009), was carried out at the Center for Inherited Disease Research (CIDR) using the Illumina GoldenGate assay22 to analyze the Illumina version 4 linkage marker panel. Quality-control filtering of SNPs, described elsewhere, yielded a final, unpruned set of 5,298 autosomal SNPs. Of these, 5,290 SNPs were found to be extant in the current genome build (GRCh37). Across the 234 families retained for whole-genome analysis, genotypes were available for 1,180 individuals, consisting largely of affected sib pairs (ASPs) and parent-offspring trios, of which 349 were untyped for WGA. Following removal of genotypes implicated in Mendelian inconsistencies, genotyping completion in the pre-inference sample ($N = 349$) was 99.3%.

4.3.3 Data pre-processing and quality-control

In order to investigate the possibility of duplicated or erroneously identified DNAs, we compared estimates of genetic relatedness (identity-by-descent) against expectation based on known familial relationships. The majority of observed inconsistencies were instances of a single, duplicated sample labeled falsely as representing an affected sib-pair. For sex-discordant sib-pairs, the true identity of a duplicate sample was resolved by consideration of X-chromosome genotypes, for which males will be haploid. For same-sex pairs of equivalent diagnostic status, duplicated samples were arbitrarily assigned one of the two alleged ids, thus preserving any remaining genetically-informative relationships. Following exclusion of problematic samples, a total of 843 individuals representing 234 pedigrees were retained for analysis.

Strand alignment and merging of WGA and linkage datasets gave a final, pre-inference set of $N = 560,657$ SNPs, of which 2,006 were common to both panels. The concordance rate of these SNPs was greater than 99%. Mendelian inconsistencies represented less than 0.00002% of the total genotypes. Pairwise estimates of genome-wide pi_hat were found to be within expectation for first-degree relationships but were suggestive of excess sharing between more-distantly related persons, including first-

and second-cousins, and avuncular relationships. Although it is conceivable that the observed cryptic relatedness is due to inbreeding or assortative mating, it may also be a consequence of the analytical procedure employed. Estimates of proportions of alleles shared IBD_0, IBD_1, and IBD_2 were generated using PLINK [95], which does not exhaustively delineate phase. Ongoing analyses of the genetic substructure in this population do indicate that the observed over-relatedness does not correlate with nationality (results not shown).

Of 843 individuals for whom WGAS SNPs were available, $N = 93$ represented pedigree founders, though $N = 286$ of the selected samples were demonstrated to be effectively independent, with additional "derived" founders originating from selection of marry-ins or, in pedigrees for which two or more non-founder genotypes were available, by random selection of a single non-founder. Allele-frequency estimates for the 555,367 SNPs unique to the WGA panel were calculated from this "derived" pool of founders. The three-fold increase in the number of contributive samples yielded a visibly improved MAF distribution with fewer monomorphic sites observed overall. Allele-frequency estimates for the 5,290 linkage SNPs were calculated from $N = 766$ founders available in the combined sample.

In silico genotyping of sparsely-genotyped (i.e. linkage) samples at dense-panel (i.e. WGAS) loci was conducted with MERLIN (v. 1.1.2) [1] on a per-pedigree, per-chromosome basis, using sample-wide allele-frequency estimates for SNPs with MAF $> 1\%$. For rarer alleles (i.e. MAF$< 1\%$), pedigree-specific allele-frequency estimates were based on all family members, thus ensuring accurate resolution of uncommon haplotypes in pedigrees segregating a rare allele. That is, SNPs with MAF less than 1% were only presumed to be polymorphic for pedigrees in which at least one copy of the minor allele was observed. In order to reduce computational intensiveness for prohibitively large, multiplex sibships for which dense-panel genotypes were available, all combinations of three siblings, $\binom{n>4}{3}$, were subjected to independent rounds of *in silico* genotyping, then re-merged to attain the final, consensus set of inferred genotypes.

4.3.4 Association Analyses

We performed association analysis for categorical diagnoses of schizophrenia using GDT (v. 0.1.1), an implementation of the generalized disequilibrium test (GDT) [11]. The GDT is a generalization of such "within-family" tests as the transmission disequilibrium test (TDT) and, similarly, the pedigree disequilibrium test (PDT), but is distinguished by its utilization of genotypes for all available phenotypically-discordant relative-pairs. Although the PDT may applied to extended pedigrees, its comparisons are limited to discordant first-degree relatives, namely parent-offspring and discordant-sibling pairs. Because such "within-family" tests consider the number of genotypically-discordant pairs rather than allelic or genotypic frequencies, we did not exclude SNPs on the basis of MAF as these SNPs are not expected to yield increased Type-I error.

In total, 535,728 polymorphic WGAS SNPs were tested for evidence of association with a diagnosis of Schizophrenia, corresponding to a Bonferroni-corrected experiment-wide significance threshold of $(\alpha = .05)/(535,728) \approx 9.33 \times 10^{-8}$. An LD-corrected significance threshold was obtained using the SNPSpD package for R [85, 97]. For each chromosome, an estimate of the "effective" number of independent tests was calculated based on the observed pairwise LD; for metacentric chromosomes[1], the p- and q-arms were taken as effectively independent units, for which estimates of the effective number of tests were calculated separately; for acrocentric chromosomes[2], a singular estimate of the number of effective tests was calculated. For both acrocentric chromosomes and the p- and q-arms of metacentric chromosomes, whole-"chromosome" data was arbitrarily bisected, as necessary, to reduce

[1] Chromosomes 1-12 and 16-20.

[2] Chromosomes 13, 14, 15, 21, and 22.

computational burden. Summation of per-chromosome estimates yielded an approximate LD-corrected experiment-wide significance threshold of $(\alpha = .05)/(265,494) \approx 1.88 \times 10^{-7}$.

4.4 Results

4.4.1 *In silico* Genotyping

Of $N = 349$ sparsely-genotyped individuals in 234 pedigrees, genotypes at $N = 557,373$ untyped WGA loci were successfully inferred for $N = 206$ individuals on the basis of informative relationships with densely-genotyped relatives. Figures 4.1 and 4.2 gives the per-sample and per-locus missingness distributions, displayed by MAF and the probability of the imputed genotype, $\Pr(G)$. We observe the largest gains in genotypic information for SNPs with MAF between 40% and 5% and $\Pr(G) < .95$. In subsequent association analyses, we utilized all inferred genotypes with $\Pr(G) > .90$.

Table 4.1 gives the number of classical linkage- and association-informative relative-pairs for each diagnosis of Schizophrenia considered herein, before and following our imputation procedure. Any realised gains in statistical power to detect disease-related loci are a function of the number of additional, informative relationships contributed to by sparsely-genotyped samples. However, given the variability of the per-sample missingness statistic, gains in sample size are not equivalent for all loci, and notably less at extreme MAFs. For example, at MAF $< .10$ and $\Pr(G) > .90$, we observe in excess of 80 sparsely-genotyped individuals with at least 80% genotyping completion at dense panel loci. At the same threshold of $\Pr(G)$, we observe in excess of 200 individuals with $> 60\%$ genotyping completion for SNPs with MAF $< 5\%$, but no individuals for whom genotyping completion exceeded 70%. From the corresponding distribution of per-SNP missingness at MAF $< 5\%$, it is evident that of 60,000 SNPs, unobserved genotypes for in excess of 40,000 SNPs achieved genotyping completion

$> 90\%$, whereas approximately 22,500 SNPs were uncompletely uninferrable.

Table 4.1: Sample and informative sample-pair counts for each Schizophrenia diagnosis.

	Schizophrenia Diagnosis												
	Narrow (D2)				Intermediate (D5)				Broad (D8)				Unaffected
	N	Trios	DSP	ASP	*N*	Trios	DSP	ASP	*N*	Trios	DSP	ASP	*N*
Observed	432	172	204	185	527	210	234	282	560	222	246	327	217
Imputed	24	15	82	21	28	19	96	29	29	19	98	33	87

For each categorical diagnosis of Scz, *N*, *Trios*, *DSP*, and *ASP* give the total number of affected individuals, parent-offspring pairs, discordant sib-pairs, and affected sib-pairs, respectively. As diagnoses are concentric, observations are unilaterally cumulative across categories.

4.4.2 Association Studies

Table 4.2 gives the thirty most significant GDT results across all categorical diagnoses of Schizophrenia. Corresponding Manhattan and quantile-quantile plots are displayed in Figures 4.3 and 4.4, respectively. While no single variant achieved established significance thresholds for genome-wide significance (i.e. 5×10^{-8}), an association at 1q32.1 between narrowly-defined Scz and *PPP1R12B* (rs12734001, $P < 1.2 \times 10^{-7}$) exceeded our LD-corrected threshold for experiment-wide significance ($\alpha_{SNPSpD} \approx 1.88 \times 10^{-7}$). A nearby SNP in strong LD ($r^2 = 1$) with rs12734001 yielded our second most significant association (rs3817222,2.5×10^{-7}) but did not meet our LD-corrected significance threshold. Both SNPs feature among the most-significant observed associations for intermediate and broad diagnoses of Scz, but are eclipsed somewhat by highly-correlated ($r^2 > .96$) associations at 11p13 (rs4755351, $5.6 \times 10{-}7$; rs12360973 $1.6 \times 10{-}6$) between *TCP11L1* with intermediate Scz, neither of which met experiment criteria for significance. Inspection of the observed quantile distributions (Figure 4.4) reveals a notable correspondence between the most significant associations for the three categorical diagnoses, with the greatest departure from expected behavior observed for narrowly-defined Scz, for which our realized, post-inference sample has least power to detect disease-related loci. Whether this reflects some specificity of effect with respect to core Schizophrenia is unclear.

4.5 Discussion

We have conducted whole-genome *in silico* genotyping and association analysis of categorical diagnoses of Scz in an Irish sample of 234 high-density schizophrenia families for which both linkage and association findings have been reported previously. By optimizing selection of association-informative family members for high-throughput genotyping, we demonstrate gainful inference of dense-panel SNP data ($> 500,000$ markers) among untyped relatives of WGAS samples using a sparse SNP panel of marker density representing less than one percent of the total WGAS SNP content. Combined analysis of observed and inferred genotype datasets by a family-based approach which utilizes all phenotypically-discordant relative pairs yielded substantial improvements to the power to detect disease-related loci using this sample. While no single variant attained established criteria for genome-wide significance, i.e. 5×10^{-8}, an association between narrowly-defined Scz and *PPP1R12B* (1q32.1) met our LD-corrected threshold for experiment-wide significance. Our findings lend support to the use of extended pedigrees in genetic studies of complex disease, with meaningful implications for study-design.

4.5.1 *In silico* Genotyping

One of the aims of this study was to assess the completeness of genotypic data attainable by *in silico* methods as applied to extended pedigrees, and the consequences for power to detect association with disease-related loci. As established by Burdick *et al.* (2006), the contingent of informative familial relationships is crucial to accurate delineation of phase, with three-generation pedigrees representing the ideal circumstance. Exclusion of duplicate and inference-ineligible individuals reduced the total number of applicable samples from $N = 349$ to $N = 206$. For a given DSP typed for our WGAS, the gain of an additional sibling with non-missing affection status corresponds to a minimum increase of one observation for that pedigree. However, the availability of multiple, multiplex sibships guarantees that gains in sample size will be consistently larger than the observed number of individuals for whom dense-panel genotypes were successfully inferred. Furthermore, by utilizing the GDT, we facilitate comparison of second- and third-degree relatives, thus extracting a greater amount of information from large, multi-generational pedigrees than by traditional analytic methods. However, non-uniformity in the per-sample missingness statistic demonstrates some variability in our power to detect allelic effects across the MAF spectrum. At higher allelic frequencies, a corresponding increase in heterozygote frequency is expected to diminish phasing accuracy, especially in genomic regions containing a large number of relatively-common SNPs. In such regions, successful inference of untyped loci relies on the informativeness of parental haplotypes, which is itself a function of local patterns of linkage disequilibrium. Based on the genotypic (allelic) priors alone, inferred genotypes for common SNPs are more likely to be excluded at more stringent thresholds for $\Pr(G)$. Similarly, SNPs with lower MAF are likely to be inferred with either a high degree of certainty or poorly. This "all or nothing" behavior is readily apparent from the distribution of per-locus and per-sample missingness statistics, which are essentially bimodal at MAF $< 5\%$.

By retaining rarer, common alleles (MAF $< 1\%$) we sought to preserve our ability to distinguish rare haplotypes. Given the aforementioned cryptic relatedness observed between allegedly unrelated persons in this sample, we considered our ability to accurately resolve—and, thereby, distinguish in parents—rare haplotypic backgrounds to represent a particularly salient issue. Our analytic treatment of less common SNPs (*i.e.* $<$ MAF 1%), while conservative in this respect, may have yielded slightly biased estimates of allele frequency for these SNPs. Without unanimous genotyping of all pedigree founders, those founders for which genotype data were available cannot be presumed to represent obligate carriers of a rare allele, thus necessitating consideration of all pedigree members. To ensure that this strategy did not indirectly influence the observed distribution of observed case-control differences, we considered the observed GDT test-statistic as a function of the estimated MAF. For narrowly-defined Scz, for which we observed our most significant findings experiment-wide, the correlation between MAF and the magnitude of effect ($|Z_{GDT}|$) was observed to be approximately -0.018, thus effectively validating this expectation. Neale *et al.* (2008) demonstrate a specific bias of Spielman's TDT in the transmission of the major versus the minor allele [79]. Allelic transmission tests represent the "between-family" component of applicable family-based tests and, as such, are not considered by the GDT per se. That is, comparisons of affected/unaffected parent-sibling pairs and vice-versa are without respect to minor/major allele status. However, Neale and colleagues demonstrate that this bias may arise from differential rates of missing genotypes, possibly reflecting the observation that genotyping clustering algorithms typically call the major homozygote genotypic class most accurately. Although differential rates of missing data in cases and controls have been demonstrated by Clayton (insert year) to bias the association test-statistic, the effect of differential rates of imputed genotypes in our sample is not expected to appreciably influence the GDT, as each family contributes at least one "unaffected v. affected" comparison to the calculated test-statistic.

4.5.2 Association Studies

Interpretation of the findings reported herein is not entirely straightforward. Evidence of linkage between chromosome 1q32 and both Schizophrenia and Bipolar Disorder has been reported in diverse samples, including but not limited to Caucasian-American, Finnish, and Korean populations[20, 25, 41, 50]. A recent meta-analysis of 32 genome-scans of Schizophrenia, which included our sample, found that regions of 1q met their threshold for aggregate genomewide evidence of linkage[84]. It is important to note that, as for any unvalidated finding, the observed signal may reflect LD between the typed SNP and nearby causal variation or represent a spurious association. However, associations between Schizophrenia and cytogenic anomalies also support the presence of a susceptibility locus on 1q, most notably *DISC1* (1q42)[110, 72, 19, 114]. More recently, a pharmacogenetic study of neurocognition as a predictor of antipsychotic treatment outcome reported associations between SNPs in two genes, *SLC26A9* (1q32) and *GPR137B* (1q42-43), and response to the drug olanzapine [69].

Of particular interest is the finding that with inclusion of additional, less-severely affected persons, we observe some attenuation in the strength of the reported associations with *PPP1R12B*. Whether this reflects some specificity of effect is a compelling possibility, albeit not immediately conclusive. Similarly, two additional SNPs at 11p13 exhibited an overall increase in significance at more-inclusive diagnostic thresholds. While the latter observation is most easily explained by appreciable gains in sample size with broadly-defined categories, that these SNPs are associated with affective components of disease is not implausible, following reports of linkage evidence at regions of 11p for Bipolar Disorder [22]. Also notable is the observation that our most significantly-associated SNP, rs12734001, exhibits severe departure from Hardy-Weinberg equilibirum (HWE) in our combined pool of founders and “derived” independent subjects ($N = 286$; 4.721×10^{-9}). Given the criteria for inclusion in the original study, we do not presume that the present sample meets the basic assumptions for HWE and therefore we did not exclude SNPs on this basis. However, an excess of heterozygotes and the complete absence of minor-allele homozygotes among said $N = 286$ subjects may be of some consequence. A paucity of minor homozygotes may suggest a particularly deleterious effect of the risk allele. Alternatively, this genotypic distribution might indicate a problematic SNP assay, or reflect a particular bias in genotype-calling.

That we observe our most significant association at a locus with no previous functional evidence of etiological relevance to Scz poses additional challenges to interpretation. Protein phosphatase 1, regulatory subunit 12B (*PPP1R12B*), known alternatively as myosin phosphatase target subunit 2 (*MYPT2*), was identified as a second human isoform of *MYPT*, with which it shares 61% sequence homology. Whereas *MYPT1* is widely distributed in human tissues, western blot analysis detected PPP1R12B protein in only heart and brain [36]. In vitro studies demonstrate that binding of the delta-subunit of protein phosphatase 1 (PP1) by either isoform increases its activity. In human tumor cells, inhibition of MYPT1-PP1-delta was shown to result in deactivation of the tumor supressor merlin, encoded by *NF2*, and downstream activation of Ras. Numerous lines of evidence support a role for a related gene, *PPP1R1B*, in the pathophysiologies of Schizophrenia and Bipolar Disorder. The *PPP1R1B* locus encodes dopamine-and-cAMP-regulated neuronal phosphoprotein (32 kDa), or DARPP-32, is an integral regulatory molecule involved in dopaminergic signaling in the prefrontal cortex. Also of interest are findings from post-mortem studies which indicate lower expression of DARPP-32 in the prefrontal cortex (PFC) of suicide-completed Schizophrenia patients [33]. Given the array of physiological mechanisms in which protein phosphatase 1 participates and its antagonism of protein kinases, it is conceivable that variants in or near *PPP1R12B* contribute to pathogenesis in Scz through dysregulation of protein phosphorylation.

The findings presented here provide additional support to published findings suggesting that Scz loci may exist on chromosome 1q and, more generally, that common SNPs contribute to Scz liability. However, without independent replication, the findings presented herein cannot be taken as confirmatory of a novel susceptibility loci at 1q32.

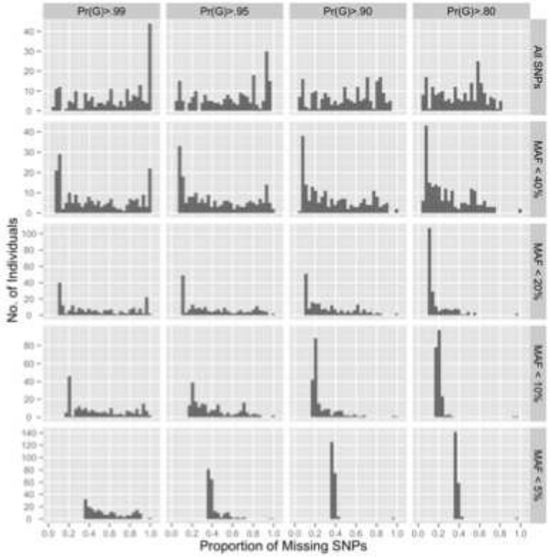


Figure 4.1: **Number of individuals by proportion of missing WGAS SNPs following *in silico* genotyping.** Distributions show the number of sparsely-typed samples as a function of the observed proportion of missing Illumina 610-quad SNP genotypes, and are given for MAFs of ≤ 0.5, .40, .20, .10, and .05 (by row) and genotype probability thresholds of .99, .95, .90, and .80 (by column).

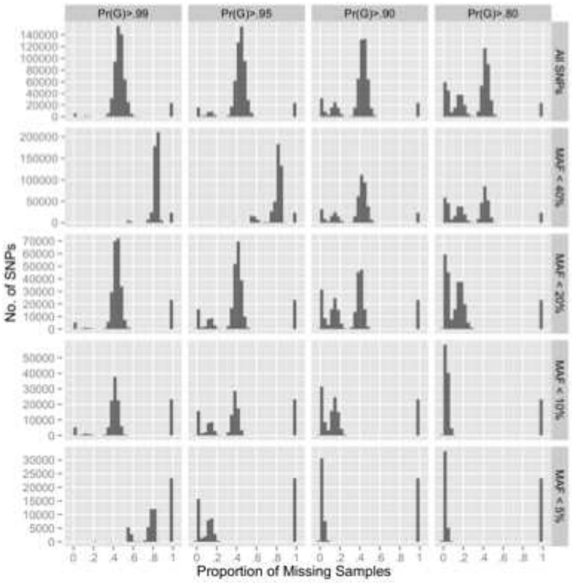


Figure 4.2: **Number of WGAS SNPs by proportion of missing individuals following *in silico* genotyping.** Distributions show the number of Illumina 610-quad SNPs as a function of the observed proportion of missing genotypes among sparsely-typed samples, and are given for MAFs of ≤ 0.5, .40, .20, .10, and .05 (by row) and genotype probability thresholds of .99, .95, .90, and .80 (by column).

Table 4.2: Top thirty (30) GDT results across concentric, inclusive diagnoses of Schizophrenia.

				Schizophrenia Diagnosis											
				Narrow (D2)				Intermediate (D5)				Broad (D8)			
Chr	Mb	SNP	Frq	Pair	Δ_{frq}	Z	P	Pair	Δ_{frq}	Z	P	Pair	Δ_{frq}	Z	P
1	202.39	rs12734001 (T)	0.297	823	0.166	5.298	1.2×10^{-7}	1058	0.142	4.904	9.4×10^{-7}	1161	0.144	5.067	4.0×10^{-7}
1	202.46	rs3817222 (T)	0.296	828	0.160	5.157	2.5×10^{-7}	1070	0.139	4.809	1.5×10^{-6}	1174	0.138	4.907	9.2×10^{-7}
11	33.04	rs4755351 (A)	0.854	903	-0.118	-4.532	5.8×10^{-6}	1158	-0.128	-5.006	5.6×10^{-7}	1269	-0.128	-4.931	8.2×10^{-7}
11	33.09	rs12360973 (A)	0.852	903	-0.115	-4.370	1.2×10^{-5}	1158	-0.125	-4.804	1.6×10^{-6}	1269	-0.125	-4.742	2.1×10^{-6}
5	18.98	rs4398638 (T)	0.606	847	0.151	4.753	2.0×10^{-6}	1078	0.127	4.416	1.0×10^{-5}	1175	0.125	4.448	8.6×10^{-6}
16	12.30	rs1472979 (T)	0.150	867	0.115	4.632	3.6×10^{-6}	1126	0.101	4.459	8.2×10^{-6}	1233	0.091	4.229	2.3×10^{-5}
16	12.30	rs11644280 (T)	0.151	867	0.115	4.632	3.6×10^{-6}	1126	0.101	4.459	8.2×10^{-6}	1233	0.091	4.229	2.3×10^{-5}
8	52.62	rs11989650 (A)	0.032	1157	0.054	4.155	3.3×10^{-5}	1536	0.052	4.552	5.3×10^{-6}	1679	0.052	4.608	4.1×10^{-6}
15	74.02	rs4886520 (T)	0.496	812	-0.142	-4.273	1.9×10^{-5}	1049	-0.141	-4.312	1.6×10^{-5}	1145	-0.147	-4.521	6.2×10^{-6}
4	65.80	rs968827 (A)	0.709	834	-0.144	-4.466	8.0×10^{-6}	1084	-0.113	-4.031	5.5×10^{-5}	1181	-0.113	-4.146	3.4×10^{-5}
15	65.37	rs12904843 (A)	0.531	866	0.145	4.425	9.6×10^{-6}	1120	0.093	3.451	0.00056	1229	0.095	3.444	0.00057
8	52.72	rs12545812 (A)	0.031	1160	0.050	3.969	7.2×10^{-5}	1541	0.049	4.378	1.2×10^{-5}	1684	0.050	4.425	9.7×10^{-6}
6	51.37	rs4355607 (T)	0.075	1076	0.094	4.250	2.1×10^{-5}	1413	0.089	4.403	1.1×10^{-5}	1549	0.082	4.281	1.9×10^{-5}
10	66.38	rs2893959 (T)	0.636	834	-0.150	-4.379	1.2×10^{-5}	1075	-0.120	-3.841	0.00012	1182	-0.122	-3.919	8.9×10^{-5}
7	8.34	rs12666065 (A)	0.918	1053	-0.078	-3.655	0.00026	1372	-0.087	-4.383	1.2×10^{-5}	1492	-0.077	-4.065	4.8×10^{-5}
7	19.35	rs11762830 (G)	0.649	848	0.149	4.199	2.7×10^{-5}	1093	0.134	4.301	1.7×10^{-5}	1196	0.136	4.369	1.2×10^{-5}
4	65.83	rs11131578 (C)	0.708	834	-0.141	-4.347	1.4×10^{-5}	1084	-0.110	-3.869	0.00011	1181	-0.110	-3.987	6.7×10^{-5}
7	19.38	rs10155945 (A)	0.235	865	-0.119	-3.557	0.00038	1113	-0.128	-4.098	4.2×10^{-5}	1219	-0.135	-4.345	1.4×10^{-5}
16	12.29	rs8062913 (A)	0.237	853	0.121	4.300	1.7×10^{-5}	1106	0.093	3.660	0.00025	1212	0.084	3.370	0.00075
10	66.37	rs10995927 (C)	0.598	843	-0.151	-4.287	1.8×10^{-5}	1088	-0.124	-3.898	9.7×10^{-5}	1194	-0.124	-3.899	9.7×10^{-5}
2	219.24	rs11899020 (C)	0.983	1147	-0.040	-4.260	2.0×10^{-5}	1519	-0.030	-3.833	0.00013	1663	-0.027	-3.817	0.00014
5	18.98	rs12109446 (C)	0.527	838	-0.136	-4.261	2.0×10^{-5}	1071	-0.125	-4.113	3.9×10^{-5}	1170	-0.126	-4.247	2.2×10^{-5}
8	38.77	rs10958811 (A)	0.174	901	-0.102	-3.937	8.3×10^{-5}	1180	-0.098	-4.262	2.0×10^{-5}	1290	-0.085	-3.892	9.9×10^{-5}
14	98.48	rs876188 (G)	0.196	877	0.083	3.084	0.002	1116	0.114	4.006	6.2×10^{-5}	1220	0.124	4.267	2.0×10^{-5}
2	174.25	rs7591287 (A)	0.546	897	0.134	4.018	5.9×10^{-5}	1172	0.133	4.212	2.5×10^{-5}	1280	0.133	4.268	2.0×10^{-5}
12	29.53	rs2278094 (A)	0.550	842	0.117	3.404	0.00066	1094	0.132	4.021	5.8×10^{-5}	1202	0.139	4.256	2.1×10^{-5}
19	58.18	rs9749513 (C)	0.340	859	0.110	3.184	0.0015	1110	0.134	4.155	3.3×10^{-5}	1221	0.133	4.254	2.1×10^{-5}
3	45.54	rs3774685 (C)	0.034	1152	0.041	3.965	7.3×10^{-5}	1533	0.046	4.224	2.4×10^{-5}	1676	0.044	4.123	3.7×10^{-5}
3	45.57	rs267219 (G)	0.034	1152	0.041	3.965	7.3×10^{-5}	1533	0.046	4.224	2.4×10^{-5}	1676	0.044	4.123	3.7×10^{-5}
4	159.88	rs17288007 (G)	0.045	1067	-0.050	-3.512	0.00044	1420	-0.055	-4.203	2.6×10^{-5}	1555	-0.050	-4.065	4.8×10^{-5}

Chromosome, genomic position (in megabases), and dbSNP identifier are given by *Chr*, *Mb*, and *SNP*, respectively; reference alleles are denoted parenthetically in *SNP*, with corresponding frequencies given by *Frq*. For each SNP and each diagnosis of Scz, the number of discordant relative pairs, observed allele frequency difference between cases and controls, GDT Z-score and resultant P-value are given by *Pairs*, Δ_{frq}, Z and P, respectively. Diagnostic groups are inclusive, with observations for central diagnoses contributing at more-peripheral strata.

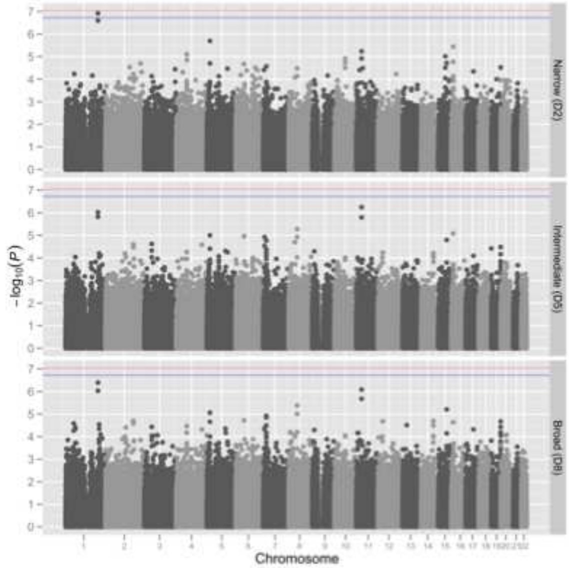


Figure 4.3: **Manhattan plots of observed GDT test-statistics by Scz diagnosis.** In each panel, standard and LD-corrected Bonferroni-correction α-thresholds are displayed in red and blue, respectively.

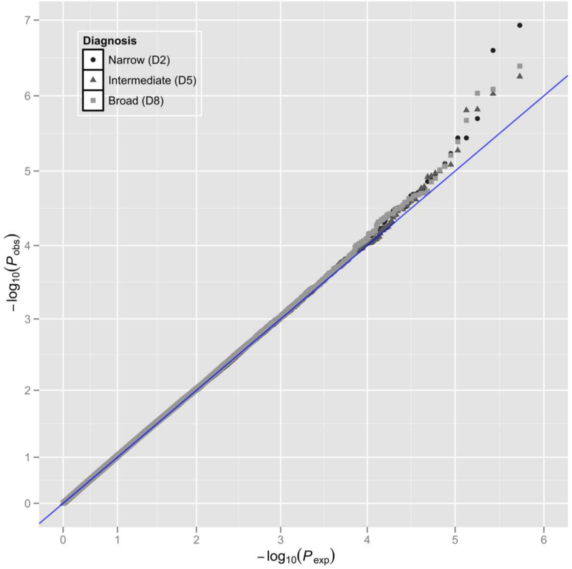


Figure 4.4: **Quantile distributions of observed GDT test-statistics by Scz diagnosis.** For each of 3 concentric diagnoses of Scz, the observed distribution of GDT P-values is given as a function of the expected distribution of quantiles for the theoretical χ^2 with 1 degree of freedom.

CHAPTER V

Global Discussion

5.1 Summary

The preceding chapters may be considered to be a survey of GWAS-era association studies which, while not comprehensive, highlights several salient methodological issues relevant to current genetic studies of complex disease. Briefly, we have thus far discussed issues of study design, including but not limited to, the range of allelic frequencies appropriately analyzed by traditional single-marker approaches, the distributional behavior of several common multi-marker tests with respect to statistical power to detect associations, and various approaches to comparing related, phenotypically discordant individuals. Furthermore, we have considered case-only analysis of clinically heterogeneous traits in the context of modifier and susceptibility-modifier genes, and expectations regarding prevalence and size of genetic effects for differentially ascertained populations. In the subsequent discussion, we attempt to frame these contributions in a unified context, with an emphasis on genetic heterogeneity—of both the allelic and locus types—and population differences in patterns of linkage disequilibrium, as manifested by *post hoc* realizations regarding power to detect significant disease-SNP associations.

5.1.1 Locus Heterogeneity

Locus heterogeneity describes a scenario in which populations, or subsets of affected persons, differ with respect to which pathogenic loci underlie disease presentation. Even moderate levels of locus heterogeneity will severely compromise power to detect significant associations with loci at which the susceptibility-conferring allele/variant is enriched or penetrant in a fraction of affected cases. This loss of statistical power is comparable to an equivalent degree of case misspecification. However, it is reasonable to suggest that for many common diseases, the extent of locus heterogeneity will be considerably less within multiply-afflicted families than in unrelated individuals, especially for populations exhibiting limited evidence of genetic admixture. In chapters 3 and 4, we describe association analysis for a sample of Irish high-density Schizophrenia families (ISHDSF). This study was originally designed for linkage analysis but it also has some advantages for association studies. First, recruitment of probands was on the basis of membership in multiply-affected pedigrees; such "high-density" families are conceivably enriched for large genetic effects. Second, we observe extensive allelic sharing in this sample between allegedly unrelated individuals, which may suggest lower levels of population genetic divergence than would be expected for the general population. Examination of chromosome- and genome-wide patterns of LD revealed extensive collinearity between SNPs, corresponding to estimates of the effective number of independent tests which amounted to roughly half of the total number of assayed SNPs in each study. In addition to substantially lowering the experiment-wide significance threshold, this finding has potential implications for genetic studies of linkage and association. First, to the extent that familial cases overlap in their sources of polygenic liability to disease, the breadth of *etiologically* relevant genetic heterogeneity may also be less in our combined sample than in the general population.

This reduced heterogeneity may also be true of variants that modify the course or presentation of illness, without directly influencing disease risk. For clinically heterogeneous disorders such as schizophrenia, locus heterogeneity may explain the observed variability in disease presentation, trajectory, and course-of-treatment. As reviewed by Fanous and Kendler (2005), distinct susceptibility loci might increase susceptibility to more or less distinct clinical subtypes of illness, whereas other "modifier genes" may influence clinical features of disease in a dimensional fashion without altering liability to the illness itself [28]. In Chapter 3, we sought to identify such modifier loci in the Irish Study of High-Density Schizophrenia Families (ISHDSF), following reported evidence of linkage to regions of chromosome 20 in the same sample using latent classes of psychotic illness. In addition to narrow, intermediate, and broad diagnoses of schizophrenia, we considered five factor-derived scores based on the Operational Criteria Checklist for Psychotic Illness (delusions, hallucinations, mania, depression, and negative symptoms). Using a gene-based approach to association, we demonstrated two empirically significant associations with depressive symptoms of schizophrenia. Most notably, of the two loci showing the strongest associations, namely *R3HDML* and *C20orf39*, neither appeared to affect the risk of schizophrenia. While this may simply reflect the limited power in this sample to detect SNP-disease associations, we take this finding as tentatively supporting roles for *R3HDML* and *C20orf39* as "modifier genes."

In our GWAS of the ISHDSF, we attempted to remedy the unfavorable statistical power of this sample by employing an approach to family-based association which considers all phenotypically-discordant relative pairs, the generalized disequilibrium test (GDT) [11]. This represents a significant departure from previous association studies using this sample, which mainly utilized the pedigree disequilibrium test (PDT). Despite both representing generalizations of Spielman's transmission disequilibrium test (TDT) to extended pedigree structures, the GDT and PDT differ markedly in the sources of associations considered. Specifically, the PDT considers parent-to-affected-child allelic transmissions and allelic sharing between phenotypically discordant sibling pairs (DSP), representing “between-” and “within-family” components of family-based association, respectively. The “within-family” component assumes homogeneity within families with respect to population stratification and, as such, is also a test of linkage. That is, phenotypically discordant relatives are treated as matched pairs without respect to a specific reference allele. Consider the consequences of population stratification or, analogously, the ascertainment of ethnically diverse subjects for Type I error. Sub-sample allele frequency differences will confound tests of association that consider allelic or genotypic counts, including the “between-family” component of the PDT. By comparison, the GDT offers substantially greater power to detect significant disease associations by maximizing the amount of information extracted from extended pedigrees. Nonetheless, our GWAS yielded only one significant association with schizophrenia, further demonstrating the unsuitability of this sample for traditional studies of association. However, that the as yet unsubstantiated association with *PPP1R12B* demonstrates some specificity-of-effect to core schizophrenia is somewhat notable.

5.1.2 Allelic heterogeneity and LD-based methods

A second type of genetic heterogeneity, allelic heterogeneity, describes a phenomenon in which study populations differ with respect to the allele or haplotype shown to confer susceptibility at a disease-associated locus. Excepting the possibility of a spurious finding in either or both populations, this observation is attributable to: (a) a true contrariety of causal variants; (b) etiological differences in the pathogenic mechanism at said locus; or, for indirect approaches to associations; (c) between-group differences in the specific pattern of LD between markers. For rare variation, scenarios (a) and (b) are plausible and perhaps not mutually exclusive. For example, normal gene function may be abrogated by missense or nonsense mutations which may both have arisen in a given population. This is rather less conceivable for common variants which are, in all likelihood, unlikely to represent true causal variation. Instead, allelic heterogeneity in common SNP-disease associations seem most likely to arise as a consequence of (c). Population differences in patterns of LD represent a widely-acknowledged shortcoming of single-marker association studies, contributing to non-replication of several well-established candidate loci. Such inconsistencies are easily resolvable by a gene-based approach, which considers all common variation at a locus jointly. Given some genomic context, functional or otherwise, for clustering common SNPs, combined methods simply represent a comprehensive approach to indirect association. As shown in Chapter 3, while single-marker analyses yielded no significant evidence of association with either diagnoses or dimensions of schizophrenia, our gene-based approach identified two novel associations with depressive symptoms of schizophrenia. Of specific interest is the finding for *C20orf39*, identified by permutation of the truncated product of P-values. From Figure 3.2, it is apparent that the SNP associations contributing to the truncated product for *C20orf39* comprise a distinct LD-block, and that the majority of these indicate an association of the minor allele with higher depression scores. Critically, no single association was sufficient to produce an empirically-significant finding for *C20orf39*.

For extremely rare variation, the very limited number of observed instances precludes any single variant from attaining established genome-wide significance thresholds. In Chapter 2 we found that the approximation that traditional association test-statistics make to the theoretical χ^2 will yield deflated significance estimates when variants are at the rarer end of the common spectrum (1% to 5%). We demonstrated that it is the number of observations, as opposed to the calculated allelic frequencies, that drive the observed deflation. Furthermore, some deflation was apparent at genome-wide significance thresholds given minor allele frequencies as high as 4%. Because accurate representation of p-values is of fundamental importance to genetic association studies, it is crucial that an appropriate statistical test be applied. A number of grouped tests have been developed to assess the evidence of association between a set of rare variants and disease. Conceptually, both grouped rare variant and gene-based approaches presume an equivalence of functional context for multiple variants, in the sense that the aggregate evidence of association is assessed for a given locus or region. However, grouped tests of rare variants differ from gene-based approaches in that the latter invokes the specific pattern of inter-correlation between neighboring SNPs as justification, whereas the grouping of rare alleles reflects the inaccuracy of traditional association tests at low allele frequencies.

Despite the increasing accessibility of re-sequencing technology, LD-based analytical methods are proving ever more important to molecular genetic studies of common, complex disease. The catalogue of human genetic polymorphisms now encompasses variants of unprecedented low frequency, owing to sequencing endeavors such as the 1,000 Genomes Project (1kGP). The current paradigm shift has also been accompanied by an overall reduction in the cost of high-throughput SNP genotyping arrays, an increasing number of which are incorporating this newly identified variation. Emergent reference datasets permit high-resolution imputation of unobserved variants for samples typed for GWAS, greatly extending the coverage of genetic variation proffered by otherwise outmoded platforms. That previous genotyping efforts are thus salvageable is a major boon to genetic studies, facilitating expedient re-analysis, replication, and meta-analysis. In Chapter 4, we describe analogous augmentation of genotypic data in sparsely genotyped (i.e. for linkage) samples using whole-genome SNP data available for related individuals. Whereas imputation approaches rely on Markov sampling of phased, reference haplotypes, family-based *in silico* genotyping utilizes haplotypic phase, where discernible in family members, to construct exact priors for unobserved genotypes. Non-uniformity of the per-SNP missingness statistic was apparent across the MAF spectrum; higher frequencies corresponded to decreased phasing accuracy; low MAF SNPs were “all or nothing.” However, the presence of large sibships in the ISDHSF permitted computationally efficient, high-fidelity inference via an iterative procedure which considered all combinatorial arrangements of GWAS-typed samples. We plan to extend this approach to the ISHDSF at even higher marker densities, following 1kGP-based imputation of GWAS samples.

5.2 Future Directions

Widely-held expectations regarding the relative number and prevalence of risk variants underlying complex, common disease have undergone periodic revision, as evidenced by the forecasted efficacy of GWAS and, more recently, the increased focus on rare and/or structural variants that are not adequately captured by common SNPs. Major medical sequencing efforts are underway but represent an emergent challenge for researchers, both with respect to study design and given the inherent difficulties in identifying and analyzing very low-frequency or structural variants. Therefore, to what extent the “missing heritability” in many complex traits resides in rare and/or structural variation remains to be seen. However, that interrogation of common genetic variation to date has been exhaustive is a dubious assertion. Improved genotype imputation procedures and large meta-analyses have contributed to a growing number of replicated disease-SNP associations. Furthermore, recent demonstrations of polygenic effects involving a large number of common variants suggest that the genetic architectures of many complex traits are of a largely unanticipated degree of complexity, but that a large proportion of the heritability of these traits can be explained by joint consideration of all common SNPs. With much of the focus of modern genetic studies of complex disease gradually shifting towards rarer variation, it is essential that the central lessons of the GWAS era be considered in a manner facilitative of downstream successes.

Although thousands of reproducible associations with human diseases and traits have been detected by GWAS, individual effects are generally quite small and, taken together, may account for only a few percent of the estimated variance. This disparity—between the estimated proportion of variance attributable to genetic factors and the realized contribution of statistically-significant associations—casts significant doubt on the tractability of these diseases by indirect approaches to association. It is conceivable that the perceived "missing heritability" reflects contributions of rare variants, epistasis, G×E interactions, and epigenetics. However, a rather more compelling argument posits the influence of thousands of small effects, many or most of which are undetectable by GWAS. Consider that, given the large number of tests performed in GWAS, a necessarily stringent multiple-testing correction must be applied to buffer against an increased rate of false positives (Type I), but comes at the cost of an increased rate of false negatives (Type II). It follows that, for many causal variants of very small effect, the observed case/control difference in allelic or a frequencies will be insufficient to warrant rejection of the null hypothesis of no association. This highlights a critical distinction between what proportion of variance is explainable by GWAS *significant* findings and the *cumulative* contribution of common variation. For example, a study conducted by the International Schizophrenia Consortium (ISC) demonstrated that ∼3% of the variance in liability to Schizophrenia could be explained by an aggregate risk score composed of a large number of SNPs which did not individually meet criteria for GWAS significance, and that the predictive value of this score across various inclusion thresholds corresponded to a polygenic model in which common variation explained approximately one-third of the variation in Schizophrenia risk[96]. Furthermore, this score was shown to explain ∼1.9% of the variance in risk of bipolar disorder but had no significant predicative value for any of six non-psychiatric disorders, including Crohns disease, type I and type II diabetes. More recently, Lee *et al.* (2011) have demonstrated that simultaneous consideration of all SNPs—in the context of realized genetic relationships between unrelated cases and controls—can explain a large proportion of the heritability in complex disease[57]. Their approach is based on a method described previously by the same group in a study of human height, in which at least 46% of the total variance in height was explained by common variant effects[124]. Extension of this method, which is based on mixed linear model analysis, to disease traits (i.e. “affected” or “unaffected”) entails transformation of the observed scale (0-1) to a continuous liability scale, and thereby must account for an overrepresentation of cases with respect to actual disease prevalence in order to provide an unbiased estimate of the variance explained by common SNPs[57]. As applied to whole-genome data from the Wellcome Trust Case Control Consortium (WTCCC), this method estimated that common SNPs (MAF>5%) accounted for 22, 37, and 28 percent of the variability in liability to Crohn's disease, bipolar diorder, and type I diabetes, respectively. Following a reported finding by the WTCCC between type I diabetes and the MHC[120], the authors also showed that chromosome 6 alone accounted for 18% of the variability in liability to disease.

The majority of demonstrable evidence of polygenic effects has addressed what proportion of genetic variance in disease- or trait-risk can be explained by a model including a large number of SNPs which do not individually attain genome-wide significance. As exemplified by the aforementioned report from the International Schizophrenia Consortium[96], the observed distribution of test-statistics from a primary GWAS can serve as a basis for assembling a genetic risk sum score. Similarly, Peterson et al. (2011) employed a meta-analytic strategy in constructing a genetic risk sum score for obesity. In either scenario, the predictive value of this score is evaluated in an independent sample. Such “evidence-based” strategies are premised on the expectation that the pooled findings from large-scale association studies harbor some number of true susceptibility variants. It follows that, for methods informed by observed case-control differences in allelic or genotypic frequencies, power to substantiate polygenic risk is not unrelated to power to detect individual effects. An alternative approach, implemented in the GCTA software[125], can estimate the variance explained by all GWAS SNPs from the realized genetic relationships between unrelated individuals. We take the latter approach to represent an "agnostic" approach to polygenic association but note that, while presenting a powerful alternative to single-SNP testing, this method is not expected to yield evidence which directly implicates specific loci in complex disease etiology. Taken together, however, the successful application of both "evidence-based" and "agnostic" approaches are of singular implication, as they suggest that a significant number of causal variants—both common and rare—are tagged by common SNPs on commercially available SNP arrays. While compelling, the possibility of an extensively polygenic basis to complex disease presents significant challenges for efforts aimed at improving our understanding of underlying etiopathogenic mechanisms. Consider that tagging SNPs are generally selected for informativeness in the context of indirect association mapping, rather than on the basis of any actual functional relevance. Extricating a true causal variant from the complex networks of LD which tag it represents an extant challenge in the interpretation of GWAS findings, as illustrated by numerous significant associations with intergenic SNPs of no discernible correspondence to nearby coding or regulatory sequence[39, 96, 104, 111]. This issue is further complicated by incomplete functional annotation of the genome and, similarly, incomplete marker saturation. The number of statistically non-significant loci implicated under a polygenic model thus poses a major methodological quandary. Refinement of polygene signals on the basis of demonstrable, pleiotropic effects represents an intriguing strategy, but one that is likely to be of limited applicability. More generally, replication of a reported polygene finding in ethnically diverse populations can lend significant support to its authenticity[96, 51].

Perhaps of more immediate consequence is the outlook for future GWAS, particularly with respect to what sample sizes and marker densities may be required to identify additional variants. That is, reported associations may represent the "low-hanging fruit" of the causal variant spectrum, being of sufficiently large effect-size to have been detected to date. Identification of novel, lower-penetrance variants by GWAS will undoubtedly require larger sample sizes and may also entail refinement of clinical phenotypes. However, particular expectations regarding the number and effect-sizes of *detectable* variants are functions of the unknown, underlying genetic architecture and, therefore, necessarily disease-specific. Park *et al.* (2010) examined recent GWAS of several human diseases and estimate the number of as yet undetected common SNPs of similar effect size, as well as what sample sizes would be required to do so[91]. Consider Crohn's disease, which has a sibling relative risk, λ_{sib}, of 20-35 (compared with ∼10 for Schizophrenia) and prevalence of 0.1%. The authors estimate that 142 loci exist with odds ratios between .07 and 1.96 and that these SNPs are expected to account for approximately 20% of the genetic variance in Crohn's. Projections of the required size of GWAS demonstrate significantly diminishing returns, with "discovery" of 108, 132 and 140 loci entailing sample sizes of 30,000, 40,000, and 50,000. Compare this with the modestly heritable breast, prostate, and colorectal cancers (λ_{sib} of 2-3) for which fewer associations have been reported. Given the same sample sizes cited for Crohn's, the authors estimate that 21, 33, 44 additional susceptibility loci could be discovered, accounting cumulatively for 8.7, 11.4, and 13.5% of the variance in risk in each of these cancers. That a nearly equivalent proportion of the variance in risk could be explained by so fewer loci in the cancers raises important questions regarding allocution of resources to future GWAS.

5.3 Current Directions

Given the robustness of polygenic findings in schizophrenia, the largely unestablished nature of polygenic disease mechanisms, and the projected multiplicity of disease-related effects, we sought to investigate the evidence of aggregate disease-related genetic differences, as manifested by differential patterns of linkage disequilibrium in a sample of unrelated Schizophrenia cases ($N_{cases} = 732$) and controls ($N_{controls} = 933$). Consider that, excepting *de novo* events occurring against otherwise indistinguishable haplotypic backgrounds, population genetic divergence at the level of DNA sequence is predominantly a function of patterns and frequency of recombination between loci. It follows that, if a majority of polygenic risk to schizophrenia were conferred by a large number of *unlinked* common variants, examination of overall patterns of LD in cases and controls should reveal only those differences attributable to sampling variability or stochastic variation.

5.3.1 Multi-SNP estimates of collinearity

Employing a "sliding-window" approach, we conducted long-range linkage disequilibrium mapping on the basis of a fixed window length of 1,000 SNPs with an overlap interval of 100 SNPs between adjacent windows. For cases and controls, we summarized the overall co-linearity of SNP genotypes within a particular window by estimation of the effective number of independent tests using SNPSpD. Figure 5.1 displays the genome-wide mappings for the 22 human autosomes. Case-control differences are displayed as the $\log_e$ ratio of the number of independent tests in cases versus controls, with negative values indicating fewer independent tests in cases compared to controls or, equivalently, that cases exhibit a greater degree of collinearity between SNPs. Alternatively, increased collinearity may be considered to represent an approximation of the observed homozygosity within a given region.

Excepting a minority of chromosomal regions, observed case-control differences in patterns of regional homozygosity are minimal. It is important to note that within a given population, the extent of LD varies between genomic regions. We do not, therefore, derive any singular threshold for what magnitude of case-control difference in the estimated number of independent tests represents a “significant” finding. However, we do interpret gradual inflection in the log-ratio value as reflecting actual, albeit generally small, group differences. Sporadic minima and maxima are most reasonably interpreted as an excess of missing genotypes in in cases and controls, respectively.

We attempted to corroborate observed group differences in regional homozygosity by examining identity-by-descent (IBD) sharing of chromosomal segments within each group for the same set of genomic windows. Figure 5.2(a) displays the observed medians for the proportion of alleles shared identity-by-descent, $\hat{\pi}$, among cases and controls. For comparison, median estimates of identity-by-state (IBS) sharing are displayed in Figure 5.2(b). Although the pattern in IBD sharing is largely concordant between cases and controls, we observe that regions for which either group displays a relative excess of sharing are in agreement with the observed trends in regional homozygosity displayed in Figure 5.1.

5.3.2 *trans*-effects

We sought next to assess whether the overall group differences in regional homozygosity would alter the observed, genome-wide *trans*-SNP interactions between “unlinked” loci. Whereas the spatial distribution of *cis*-interactants is definitively linear and inclusive of intervening loci, *trans*-interactions will be distributed according to the particular definition of “unlinked” applied. At extremes, loci mapping to separate chromosomes represent completey “unlinked” entities though loci mapping to a single chromosome may be considered effectively “independent” if in approximate linkage equilibrium. Under independent assortment (of chromosomes), it is not unreasonable to posit a null hypothesis of no (0) correlation between loci occurring on different chromosomes. Because we are rather less interested in the origin or magnitude of specific SNP×SNP correlations than in gross distributional differences between cases and control, we opted specifically to utilize our full dataset without pruning of highly-correlated SNPs.

Quantiles for the genome-wide distribution of extra-chromosomal SNP×SNP correlations in cases and controls are displayed in Figure 5.3(a). We observe that, excepting some degree of quantization of the correlation test-statistic, the observed distributions for cases and controls are largely concordant. From the CHR×ALL distributions given in Figure 5.3(b), some variability in the distribution of each chromosome is apparent, with respect to both the magnitude of the observed correlations and deviation from expected behavior. For example, the P-values corresponding to the most significant correlations with chromosome 9 SNPs are several orders of magnitude larger than those observed for chromosome 7. This variability is independent of chromosome size and SNP-density, and likely reflects—to some degree—the distribution of low-frequency and monomorphic sites genome-wide. For some chromosomes, most conspicuously chromosome 11, some deviation from expected behavior is apparent, most of which is confined to the upper quantiles. From the CHR×CHR distributions in Figure 5.3(c), the source of this deviation is shown to arise from the correlation with chromosome 7 SNPs. The increased number of significant correlations between chromosomes 7 and 11 might reflect a higher degree of true LD on neighboring SNPs located on either chromosome.

5.3.3 Interpretation

We expect that random variability in patterns of observed *cis*-interactions may give rise to spurious differences, most likely arising as a consequence of sampling variation in the haplotypic diversity. Resolving the conditions under which *cis*-interactions are likely to manifest could benefit from the refinement of boundary definitions. As cited previously in Chapter 3, it has been demonstrated that the vicinity of gene-coding regions are enriched for significant genome-wide associations [76]. Whether common variation in and around genes exhibits greater evidence for *cis*-interactions is not a straightforward question, since many coding sequences will exhibit a higher degree of conservation and, thereby, less divergence than non-coding loci. However, re-sequencing of of associated regions will eventually resolve whether unobserved rare variation captured jointly by multiple SNPs are responsible for an observed signal. The failure of comprehensive, direct association mapping to account for a well-supported WGAS-significant finding may suggest the presence of disease-related *cis*-interactions between genic variation and up/downstream regulatory regions, or across multiple, syntenic loci.

Epistatic or *trans*-interaction between unlinked loci represents a distinctly possible but largely unsubstantiated source of variance in complex traits. Classical definitions of epistasis, much like those of penetrance, are burdened somewhat by a connotation of biological interaction. A "hypothesis-free" approach to detecting epistasis (e.g. GWAS) can entail an overwhelming multiple-testing burden, such that derivation of prior odds of association for interactions is problematic in the absence of functional knowledge. In the present context, we do not presume a specifically multiplicative effect but rather assess the case-control differences in distribution of correlations for all completely unlinked loci.

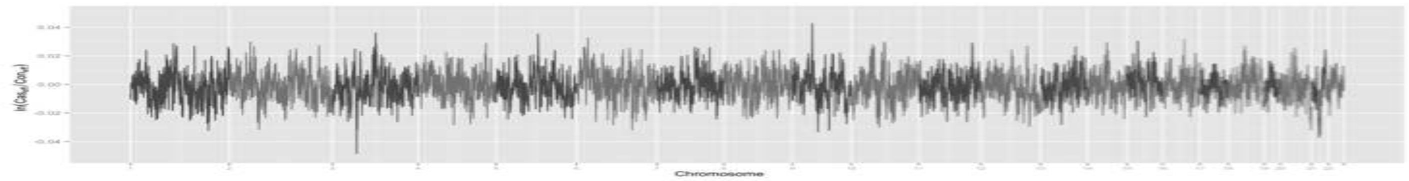
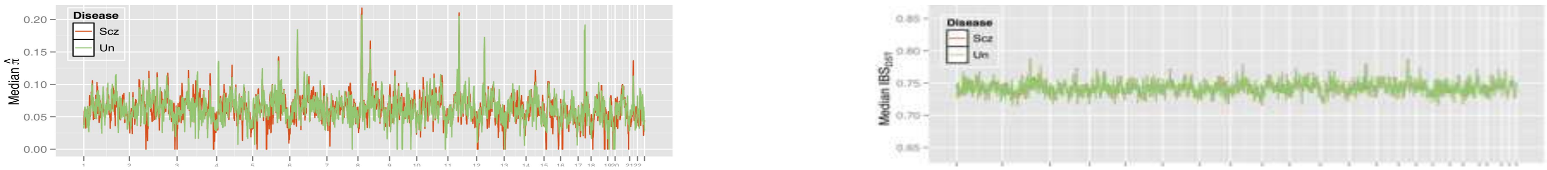


Figure 5.1: **Differential patterns of LD in unrelated cases and controls**. For each genomic window of 1,000 SNPs, an estimate of the effective number of independent tests was obtained using SNPSpD. Case-control differences are displayed as the $\log_e$ ratio of the no. of independent tests in cases versus controls; negative values indicate fewer independent tests in cases compared to controls or, equivalently, that cases exhibit a greater degree of collinearity between SNPs.

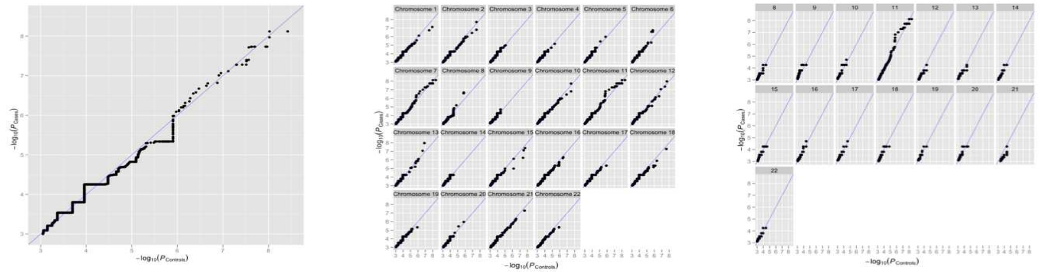


(a) Identity-by-descent ($\hat{\pi}$)

(b) Identity-by-state (IBS_{DST})

Figure 5.2: **Regional IBD/IBS sharing in unrelated cases and controls.** Pairwise estimates of allele-sharing (IBD/IBS) were obtained for a sliding window of 1,000 SNPs (*top*) and for each chromosome (*bot.*). Plotted points represent **(a.)** median $\hat{\pi}^{\dagger}$ and **(b.)** median $\text{IBS}_{DST}{}^{\ddagger}$, displayed separately for each group.

$^{\dagger}\hat{\pi} = P(IBD2) + 0.5 \times P(IBD1)$. $^{\ddagger}\text{IBS}_{DST} = (\text{IBS2} + 0.5 \times \text{IBS1}) / (N\ SNP)$.



(a) ALL×ALL

(b) CHR×ALL

(c) CHR×CHR

Figure 5.3: **Case/control distributions of genome-wide *trans*-associations.** For cases and controls, correlation test P-values representing all SNP×SNP comparisons were assembled to render **(a)** complete quantile distributions from **(c)** quantile distributions for each CHR×ALL comparison. **(c)** Quantile distribution for 7×CHR comparison.

5.4 Closing remarks

Advancement of risk prediction represents a long-held aspiration in complex disease genetics, but one that has been differentially deferred for many common diseases. In particular, the etiologies of common psychiatric diseases have demonstrated a persistent, historical recalcitrance to genetic studies. However, recent developments suggest an imminent revolution in the molecular genetics of several disorders, including Schizophrenia. Refinement of polygenic approaches may aid in disentangling true susceptibility loci from a host of etiologically-irrelevant, statistically-significant associations. More broadly, an extensively polygenic etiology suggests a convergence of individual differences in neurodevelopment and function which, while conceivably unfavorable in terms of risk prediction, may aid in better situating these adverse and divergent diagnostic entities within the purview of natural human phenotypic variation.

APPENDICES

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Goncalo R Abecasis, Stacey S Cherny, William O Cookson, and Lon R Cardon. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101, Jan 2002.

[2] R J Apple, H A Erlich, W Klitz, M M Manos, T M Becker, and C M Wheeler. HLA DR-DQ associations with cervical carcinoma show papillomavirus-type specificity. *Nat Genet*, 6(2):157–162, Feb 1994.

[3] Tadao Arinami, Tsuyuka Ohtsuki, Hiroki Ishiguro, Hiroshi Ujike, Yuji Tanaka, Yukitaka Morita, Mari Mineta, Masashi Takeichi, Shigeto Yamada, Akira Imamura, Koichi Ohara, Haruo Shibuya, Kenshiro Ohara, Yasuo Suzuki, Tatsuyuki Muratake, Naoshi Kaneko, Toshiyuki Someya, Toshiya Inada, Takeo Yoshikawa, Tomoko Toyota, Kazuo Yamada, Takuya Kojima, Sakae Takahashi, Ohmori Osamu, Takahiro Shinkai, Michiko Nakamura, Hiroshi Fukuzako, Tomo Hashiguchi, Shin-ich Niwa, Takuya Ueno, Hirokazu Tachikawa, Takafumi Hori, Takashi Asada, Shinichiro Nanko, Hiroshi Kunugi, Ryota Hashimoto, Norio Ozaki, Nakao Iwata, Mutsuo Harano, Heii Arai, Tohru Ohnuma, Ichiro Kusumi, Tsukasa Koyama, Hiroshi Yoneda, Yasuyuki Fukumaki, Hiroki Shibata, Sunao Kaneko, Hisashi Higuchi, Norio Yasui-Furukori, Yohtaro Numachi, Masanari Itokawa, and Yuji Okazaki. Genomewide high-density SNP linkage analysis of 236 Japanese families supports the existence of schizophrenia susceptibility loci on chromosomes 1p, 14q, and 20p. *Am J Hum Genet*, 77(6):937–944, Dec 2005.

[4] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386, 1955.

[5] Jeffrey C Barrett and Lon R Cardon. Evaluating coverage of genome-wide association studies. *Nat Genet*, 38(6):659–662, Jun 2006.

[6] Joshua T Burdick, Wei-Min Chen, Goncalo R Abecasis, and Vivian G Cheung. In silico method for inferring genotypes in pedigrees. *Nat Genet*, 38(9):1002–1004, Sep 2006.

[7] William S Bush, Stephen J Sawcer, Philip L de Jager, Jorge R Oksenberg, Jacob L McCauley, Margaret A Pericak-Vance, and Jonathan L Haines. Evidence for polygenic susceptibility to multiple sclerosis–the shape of things to come. *Am J Hum Genet*, 86(4):621–625, Apr 2010.

[8] Christopher S Carlson, Michael A Eberle, Mark J Rieder, Qian Yi, Leonid Kruglyak, and Deborah A Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74(1):106–120, Jan 2004.

[9] M Chakos, J Lieberman, E Hoffman, D Bradford, and B Sheitman. Effectiveness of second-generation antipsychotics in patients with treatment-resistant schizophrenia: A review and meta-analysis of randomized trials. *Am J Psychiatry*, 158(4):518–526, Apr 2001.

[10] A Chakravarti. Population genetics–making sense out of sequence. *Nat Genet*, 21(1 Suppl):56–60, Jan 1999.

[11] Wei-Min Chen, Ani Manichaikul, and Stephen S Rich. A generalized family-based association test for dichotomous traits. *Am J Hum Genet*, 85(3):364–76, Sep 2009.

[12] Sven Cichon, Nick Craddock, Mark Daly, Stephen V Faraone, Pablo V Gejman, John Kelsoe, Thomas Lehner, Douglas F Levinson, Audra Moran, Pamela Sklar, and Patrick F Sullivan. Genomewide association studies: History, rationale, and prospects for psychiatric disorders. *Am J Psychiatry*, 166(5):540–556, May 2009.

[13] Jonathan C Cohen, Eric Boerwinkle, Thomas H Jr Mosley, and Helen H Hobbs. Sequence variations in pcsk9, low ldl, and protection against coronary heart disease. *N Engl J Med*, 354(12):1264–1272, Mar 2006.

[14] H Coon, S Jensen, J Holik, M Hoff, M Myles-Worsley, F Reimherr, P Wender, M Waldo, R Freedman, and M Leppert. Genomic scan for genes predisposing to schizophrenia. *Am J Med Genet*, 54(1):59–71, Mar 1994.

[15] N Craddock, M C O'Donovan, and M J Owen. The genetics of schizophrenia and bipolar disorder: Dissecting psychosis. *J Med Genet*, 42(3):193–204, Mar 2005.

[16] Nick Craddock, Michael C O'Donovan, and Michael J Owen. Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. *Schizophr Bull*, 32(1):9–16, Jan 2006.

[17] " " Cross-Disorder Phenotype Group of the Psychiatric GWAS Consortium, Nick Craddock, Kenneth Kendler, Michael Neale, John Nurnberger, Shaun Purcell, Marcella Rietschel, Roy Perlis, Susan L Santangelo, Thomas G Schulze, Jordan W Smoller, and Anita Thapar. Dissecting the phenotype in genome-wide association studies of psychiatric illness. *Br J Psychiatry*, 195(2):97–99, Aug 2009.

[18] Paul I W de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B Gabriel, Mark J Daly, and David Altshuler. Efficiency and power in genetic association studies. *Nat Genet*, 37(11):1217–1223, Nov 2005.

[19] O Demirhan, D Tastemir, and Y Sertdemir. Chromosomal fragile sites in schizophrenic patients. *Genetika*, 42(7):985–992, Jul 2006.

[20] S D Detera-Wadleigh, J A Badner, W H Berrettini, T Yoshikawa, L R Goldin, G Turner, D Y Rollins, T Moses, A R Sanders, J D Karkera, L E Esterling, J Zeng, T N Ferraro, J J Guroff, D Kazuba, M E Maxwell, J I Jr Nurnberger, and E S Gershon. A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2. *Proc Natl Acad Sci U S A*, 96(10):5604–5609, May 1999.

[21] S D Detera-Wadleigh, J A Badner, T Yoshikawa, A R Sanders, L R Goldin, G Turner, D Y Rollins, T Moses, J J Guroff, D Kazuba, M E Maxwell, H J Edenberg, T Foroud, D Lahiri, J I Jr Nurnberger, O C Stine, F McMahon, D A Meyers, D MacKinnon, S Simpson, M McInnis, J R DePaulo, J Rice, A Goate, and E S Gershon. Initial genome scan of the NIMH genetics initiative bipolar pedigrees: chromosomes 4, 7, 9, 18, 19, 20, and 21q. *Am J Med Genet*, 74(3):254–262, May 1997.

[22] S D Detera-Wadleigh, W T Hsieh, W H Berrettini, L R Goldin, D Y Rollins, D Muniec, R Grewal, J J Guroff, G Turner, and D Coffman. Genetic linkage mapping for a susceptibility locus to bipolar illness: chromosomes 2, 3, 4, 7, 9, 10p, 11p, 22, and xpter. *Am J Med Genet*, 54(3):206–218, Sep 1994.

[23] Frank Dudbridge. Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol*, 25(2):115–121, Sep 2003.

[24] Frank Dudbridge and Arief Gusnanto. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32(3):227–234, Apr 2008.

[25] J Ekelund, I Hovatta, A Parker, T Paunio, T Varilo, R Martin, J Suhonen, P Ellonen, G Chan, J S Sinsheimer, E Sobel, H Juvonen, R Arajarvi, T Partonen, J Suvisaari, J Lonnqvist, J Meyer, and L Peltonen. Chromosome 1 loci in finnish schizophrenia families. *Hum Mol Genet*, 10(15):1611–1617, Jul 2001.

[26] B Etain, F Mathieu, M Rietschel, W Maier, M Albus, P McKeon, S Roche, C Kealey, D Blackwood, W Muir, F Bellivier, C Henry, C Dina, S Gallina, H Gurling, A Malafosse, M Preisig, F Ferrero, S Cichon, J Schumacher, S Ohlraun, M Borrmann-Hassenbach, P Propping, R Abou Jamra, T G Schulze, A Marusic, Z M Dernovsek, B Giros, T Bourgeron, A Lemainque, D Bacq, C Betard, C Charon, M M Nothen, M Lathrop, and M Leboyer. Genome-wide scan for genes involved in bipolar affective disorder in 70 European families ascertained through a bipolar type I early-onset proband: Supportive evidence for linkage at 3p14. *Mol Psychiatry*, 11(7):685–694, Jul 2006.

[27] A Fanous, C Gardner, D Walsh, and K S Kendler. Relationship between positive and negative symptoms of schizophrenia and schizotypal symptoms in nonpsychotic relatives. *Arch Gen Psychiatry*, 58(7):669–673, Jul 2001.

[28] A H Fanous and K S Kendler. Genetic heterogeneity, modifier genes, and quantitative phenotypes in psychiatric illness: Searching for a framework. *Mol Psychiatry*, 10(1):6–13, Jan 2005.

[29] Ayman H Fanous and Kenneth S Kendler. Genetics of clinical features and subtypes of schizophrenia: A review of the recent literature. *Curr Psychiatry Rep*, 10(2):164–170, Apr 2008.

[30] Ayman H Fanous, M C Neale, R E Straub, B T Webb, A F O'Neill, D Walsh, and K S Kendler. Clinical features of psychotic disorders and polymorphisms in HT2A, DRD2, DRD4, SLC6A3 (DAT1), and BDNF: A family based association study. *Am J Med Genet B Neuropsychiatr Genet*, 125B(1):69–78, Feb 2004.

[31] Ayman H Fanous, Michael C Neale, Bradley T Webb, Richard E Straub, Francis A O'Neill, Dermot Walsh, Brien P Riley, and Kenneth S Kendler. Novel linkage to chromosome 20p using latent classes of psychotic illness in 270 Irish high-density families. *Biol Psychiatry*, 64(2):121–127, Jul 2008.

[32] Ayman H Fanous, Edwin J van den Oord, Brien P Riley, Steven H Aggen, Michael C Neale, F Anthony O'Neill, Dermot Walsh, and Kenneth S Kendler. Relationship between a high-risk haplotype in the DTNBP1 (dysbindin) gene and clinical features of schizophrenia. *Am J Psychiatry*, 162(10):1824–1832, Oct 2005.

[33] Laura A Feldcamp, Renan P Souza, Marco Romano-Silva, James L Kennedy, and Albert H C Wong. Reduced prefrontal cortex darpp-32 mrna in completed suicide victims with schizophrenia. *Schizophr Res*, 103(1-3):192–200, Aug 2008.

[34] R.A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, Jan 1922.

[35] B Freidlin, G Zheng, Z Li, and J L Gastwirth. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered*, 53(3):146–152, 2002.

[36] M Fujioka, N Takahashi, H Odai, S Araki, K Ichikawa, J Feng, M Nakamura, K Kaibuchi, D J Hartshorne, T Nakano, and M Ito. A new isoform of human myosin phosphatase targeting/regulatory subunit (mypt2): cdna cloning, tissue expression, and chromosomal mapping. *Genomics*, 49(1):59–68, Apr 1998.

[37] Janice M Fullerton, Jennifer A Donald, Philip B Mitchell, and Peter R Schofield. Two-dimensional genome scan identifies multiple genetic interactions in bipolar affective disorder. *Biol Psychiatry*, 67(5):478–486, Mar 2010.

[38] Pablo V Gejman, Alan R Sanders, and Jubao Duan. The role of genetics in the etiology of schizophrenia. *Psychiatr Clin North Am*, 33(1):35–66, Mar 2010.

[39] Pablo V Gejman, Alan R Sanders, and Kenneth S Kendler. Genetics of schizophrenia: new findings and challenges. *Annu Rev Genomics Hum Genet*, 12:121–144, Sep 2011.

[40] H M Gurling, G Kalsi, J Brynjolfson, T Sigmundsson, R Sherrington, B S Mankoo, T Read, P Murphy, E Blaveri, A McQuillin, H Petursson, and D Curtis. Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23. *Am J Hum Genet*, 68(3):661–673, Mar 2001.

[41] H M Gurling, G Kalsi, J Brynjolfson, T Sigmundsson, R Sherrington, B S Mankoo, T Read, P Murphy, E Blaveri, A McQuillin, H Petursson, and D Curtis. Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23. *Am J Hum Genet*, 68(3):661–673, Mar 2001.

[42] Rene Gysin, Rudolf Kraftsik, Julie Sandell, Pierre Bovet, Celine Chappuis, Philippe Conus, Patricia Deppen, Martin Preisig, Viviane Ruiz, Pascal Steullet, Mirjana Tosic, Thomas Werge, Michel Cuenod, and Kim Q Do. Impaired glutathione synthesis in schizophrenia: Convergent genetic and functional evidence. *Proc Natl Acad Sci U S A*, 104(42):16621–16626, Oct 2007.

[43] M L Hamshere, N M Williams, N Norton, H Williams, A G Cardno, S Zammit, L A Jones, K C Murphy, R D Sanders, G McCarthy, M Y Gray, G Jones, P Holmans, M C O'Donovan, M J Owen, and N Craddock. Genome wide significant linkage in schizophrenia conditioning on occurrence of depressive episodes. *J Med Genet*, 43(7):563–567, Jul 2006.

[44] Marian L Hamshere, Thomas G Schulze, Johannes Schumacher, Aiden Corvin, Michael J Owen, Rami Abou Jamra, Peter Propping, Wolfgang Maier, Guillermo Orozco y Diaz, Fermin Mayoral, Fabio Rivas, Ian Jones, Lisa Jones, George Kirov, Michael Gill, Peter A Holmans, Markus M Nothen, Sven Cichon, Marcella Rietschel, and Nick Craddock. Mood-incongruent psychosis in bipolar disorder: Conditional linkage analysis shows genome-wide suggestive linkage at 1q32.3, 7p13 and 20q13.31. *Bipolar Disord*, 11(6):610–620, Sep 2009.

[45] S W Hardy, B S Weir, N L Kaplan, and E R Martin. Analysis of single nucleotide polymorphisms in candidate genes using the pedigree disequilibrium test. *Genet Epidemiol*, 21 Suppl 1:S441–6, 2001.

[46] Walter W. Hauck Jr. and Allan Donner. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72(360):851–853, Dec 1977.

[47] Keith Hawton, Lesley Sutton, Camilla Haw, Julia Sinclair, and Jonathan J Deeks. Schizophrenia and suicide: Systematic review of risk factors. *Br J Psychiatry*, 187:9–20, Jul 2005.

[48] Clive J Hoggart, Taane G Clark, Maria De Iorio, John C Whittaker, and David J Balding. Genome-wide significance for dense snp and resequencing data. *Genet Epidemiol*, 32(2):179–185, Feb 2008.

[49] P A Holmans, B Riley, A E Pulver, M J Owen, D B Wildenauer, P V Gejman, B J Mowry, C Laurent, K S Kendler, G Nestadt, N M Williams, S G Schwab, A R Sanders, D Nertney, J Mallet, B Wormley, V K Lasseter, M C O'Donovan, J Duan, M Albus, M Alexander, S Godard, R Ribble, K Y Liang, N Norton, W Maier, G Papadimitriou, D Walsh, M Jay, A O'Neill, F B Lerer, D Dikeos, R R Crowe, J M Silverman, and D F Levinson. Genomewide linkage scan of schizophrenia in a large multicenter pedigree sample using single nucleotide polymorphisms. *Mol Psychiatry*, 14(8):786–795, Aug 2009.

[50] I Hovatta, T Varilo, J Suvisaari, J D Terwilliger, V Ollikainen, R Arajarvi, H Juvonen, M L Kokko-Sahin, L Vaisanen, H Mannila, J Lonnqvist, and L Peltonen. A genomewide screen for schizophrenia genes in an isolated finnish subpopulation, suggesting multiple susceptibility loci. *Am J Hum Genet*, 65(4):1114–1124, Oct 1999.

[51] Masashi Ikeda, Branko Aleksic, Yoko Kinoshita, Tomo Okochi, Kunihiro Kawashima, Itaru Kushima, Yoshihito Ito, Yukako Nakamura, Taro Kishi, Takenori Okumura, Yasuhisa Fukuo, Hywel J Williams, Marian L Hamshere, Dobril Ivanov, Toshiya Inada, Michio Suzuki, Ryota Hashimoto, Hiroshi Ujike, Masatoshi Takeda, Nick Craddock, Kozo Kaibuchi, Michael J Owen, Norio Ozaki, Michael C O'Donovan, and Nakao Iwata. Genome-wide association study of schizophrenia in a japanese population. *Biol Psychiatry*, 69(5):472–478, Mar 2011.

[52] Consortium International HapMap. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct 2005.

[53] John P A Ioannidis, Gilles Thomas, and Mark J Daly. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet*, 10(5):318–329, May 2009.

[54] R Kaiser, M Konneker, M Henneken, M Dettling, B Muller-Oerlinghausen, I Roots, and J Brockmoller. Dopamine D4 receptor 48-bp repeat polymorphism: No association with response to antipsychotic treatment, but association with catatonic schizophrenia. *Mol Psychiatry*, 5(4):418–424, Jul 2000.

[55] D Karolchik, R Baertsch, M Diekhans, T S Furey, A Hinrichs, Y T Lu, K M Roskin, M Schwartz, C W Sugnet, D J Thomas, R J Weber, D Haussler, and W J Kent. The UCSC genome browser database. *Nucleic Acids Res*, 31(1):51–54, Jan 2003.

[56] K S Kendler, F A O'Neill, J Burke, B Murphy, F Duke, R E Straub, R Shinkwin, M Ni Nuallain, C J MacLean, and D Walsh. Irish study on high-density schizophrenia families: Field methods and power to detect linkage. *Am J Med Genet*, 67(2):179–190, Apr 1996.

[57] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet*, 88(3):294–305, Mar 2011.

[58] Cathryn M Lewis, Douglas F Levinson, Lesley H Wise, Lynn E DeLisi, Richard E Straub, Iiris Hovatta, Nigel M Williams, Sibylle G Schwab, Ann E Pulver, Stephen V Faraone, Linda M Brzustowicz, Charles A Kaufmann, David L Garver, Hugh M D Gurling, Eva Lindholm, Hilary Coon, Hans W Moises, William Byerley, Sarah H Shaw, Andrea Mesen, Robin Sherrington, F Anthony O'Neill, Dermot Walsh, Kenneth S Kendler, Jesper Ekelund, Tiina Paunio, Jouko Lonnqvist, Leena Peltonen, Michael C O'Donovan, Michael J Owen, Dieter B Wildenauer, Wolfgang Maier, Gerald Nestadt, Jean-Louis Blouin, Stylianos E Antonarakis, Bryan J Mowry, Jeremy M Silverman, Raymond R Crowe, C Robert Cloninger, Ming T Tsuang, Dolores Malaspina, Jill M Harkavy-Friedman, Dragan M Svrakic, Anne S Bassett, Jennifer Holcomb, Gursharan Kalsi, Andrew McQuillin, Jon Brynjolfson, Thordur Sigmundsson, Hannes Petursson, Elena Jazin, Tomas Zoega, and Tomas Helgason. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am J Hum Genet*, 73(1):34–48, Jul 2003.

[59] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83(3):311–321, Sep 2008.

[60] Paul Lichtenstein, Benjamin H Yip, Camilla Bjork, Yudi Pawitan, Tyrone D Cannon, Patrick F Sullivan, and Christina M Hultman. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: A population-based study. *Lancet*, 373(9659):234–239, Jan 2009.

[61] Roderick J. Little. Testing the equality of two independent binomial proportions. *The American Statistician*, 43(4):283–288, Nov 1989.

[62] Augustin Luna and Kristin K Nicodemus. Snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. *Bioinformatics*, 23(6):774–776, Mar 2007.

[63] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, Feb 2009.

[64] Brion S Maher, Mark A Reimers, Brien P Riley, and Kenneth S Kendler. Allelic heterogeneity in genetic association meta-analysis: An application to DTNBP1 and schizophrenia. *Hum Hered*, 69(2):71–79, 2010.

[65] Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, Natalia Popova, Stephanie Pretel, Lora Ziyabari, Moira Lee, Yu Shao, Zhen Y Wang, Karl Sirotkin, Minghong Ward, Michael Kholodov, Kerry Zbicz, Jeffrey Beck, Michael Kimelman, Sergey Shevelev, Don Preuss, Eugene Yaschenko, Alan Graeff, James Ostell, and Stephen T Sherry. The ncbi dbgap database of genotypes and phenotypes. *Nat Genet*, 39(10):1181–1186, Oct 2007.

[66] A K Malhotra, D Goldman, C Mazzanti, A Clifton, A Breier, and D Pickar. A functional serotonin transporter (5-HTT) polymorphism is associated with psychosis in neuroleptic-free schizophrenics. *Mol Psychiatry*, 3(4):328–332, Jul 1998.

[67] E R Martin, S A Monks, L L Warren, and N L Kaplan. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet*, 67(1):146–154, Jul 2000.

[68] Mark I McCarthy, Goncalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John P A Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369, May 2008.

[69] Joseph L McClay, Daniel E Adkins, Karolina Aberg, Jozsef Bukszar, Amit N Khachane, Richard S E Keefe, Diana O Perkins, Joseph P McEvoy, T Scott Stroup, Robert E Vann, Patrick M Beardsley, Jeffrey A Lieberman, Patrick F Sullivan, and Edwin J C G van den Oord. Genome-wide pharmacogenomic study of neurocognition as an indicator of antipsychotic treatment response in schizophrenia. *Neuropsychopharmacology*, 36(3):616–626, Feb 2011.

[70] P McGuffin, A Farmer, and I Harvey. A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. *Arch Gen Psychiatry*, 48(8):764–770, Aug 1991.

[71] Melvin G McInnis, Danielle M Dick, Virginia L Willour, Dimitrios Avramopoulos, Dean F MacKinnon, Sylvia G Simpson, James B Potash, Howard J Edenberg, Elizabeth S Bowman, Francis J McMahon, Carrie Smiley, Jennifer L Chellis, Yuqing Huo, Tyra Diggs, Eric T Meyer, Marvin Miller, Amy T Matteini, N Leela Rau, J Raymond DePaulo, Elliot S Gershon, Judith A Badner, John P Rice, Alison M Goate, Sevilla D Detera-Wadleigh, John I Nurnberger, Theodore Reich, Peter P Zandi, and Tatiana M Foroud. Genome-wide scan and conditional analysis in bipolar disorder: Evidence for genomic interaction in the National Institute of Mental Health genetics initiative bipolar pedigrees. *Biol Psychiatry*, 54(11):1265–1273, Dec 2003.

[72] J K Millar, J C Wilson-Annan, S Anderson, S Christie, M S Taylor, C A Semple, R S Devon, D M St Clair, W J Muir, D H Blackwood, and D J Porteous. Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum Mol Genet*, 9(9):1415–1423, May 2000.

[73] H W Moises, L Yang, H Kristbjarnarson, C Wiese, W Byerley, F Macciardi, V Arolt, D Blackwood, X Liu, and B Sjogren. An international two-stage genome-wide search for schizophrenia susceptibility genes. *Nat Genet*, 11(3):321–324, Nov 1995.

[74] S A Monks and N L Kaplan. Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. *Am J Hum Genet*, 66(2):576–592, Feb 2000.

[75] Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutat Res*, 615(1-2):28–56, Feb 2007.

[76] V Moskvina, N Craddock, P Holmans, I Nikolov, J S Pahwa, E Green, M J Owen, and M C O'Donovan. Gene-wide analyses of genome-wide association data sets: Evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol Psychiatry*, 14(3):252–260, Mar 2009.

[77] Mousumi Mutsuddi, Derek W Morris, Skye G Waggoner, Mark J Daly, Edward M Scolnick, and Pamela Sklar. Analysis of high-resolution hapmap of DTNBP1 (dysbindin) suggests no consistency between reported common variant associations and schizophrenia. *Am J Hum Genet*, 79(5):903–909, Nov 2006.

[78] Benjamin M Neale, Jesen Fagerness, Robyn Reynolds, Lucia Sobrin, Margaret Parker, Soumya Raychaudhuri, Perciliz L Tan, Edwin C Oh, Joanna E Merriam, Eric Souied, Paul S Bernstein, Binxing Li, Jeanne M Frederick, Kang Zhang, Milam A Jr Brantley, Aaron Y Lee, Donald J Zack, Betsy Campochiaro, Peter Campochiaro, Stephan Ripke, R Theodore Smith, Gaetano R Barile, Nicholas Katsanis, Rando Allikmets, Mark J Daly, and Johanna M Seddon. Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (lipc). *Proc Natl Acad Sci U S A*, 107(16):7395–7400, Apr 2010.

[79] Benjamin M Neale, Jessica Lasky-Su, Richard Anney, Barbara Franke, Kaixin Zhou, Julian B Maller, Alejandro Arias Vasquez, Philip Asherson, Wai Chen, Tobias Banaschewski, Jan Buitelaar, Richard Ebstein, Michael Gill, Ana Miranda, Robert D Oades, Herbert Roeyers, Aribert Rothenberger, Joseph Sergeant, Hans Christoph Steinhausen, Edmund Sonuga-Barke, Fernando Mulas, Eric Taylor, Nan Laird, Christoph Lange, Mark Daly, and Stephen V Faraone. Genome-wide association scan of attention deficit hyperactivity disorder. *Am J Med Genet B Neuropsychiatr Genet*, 147B(8):1337–1344, Dec 2008.

[80] Benjamin M Neale, Manuel A Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, and Mark J Daly. Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322, Mar 2011.

[81] Benjamin M Neale and Pak C Sham. The future of association studies: Gene-based analysis and replication. *Am J Hum Genet*, 75(3):353–362, Sep 2004.

[82] Anna C Need, Dongliang Ge, Michael E Weale, Jessica Maia, Sheng Feng, Erin L Heinzen, Kevin V Shianna, Woohyun Yoon, Dalia Kasperaviciute, Massimo Gennarelli, Warren J Strittmatter, Cristian Bonvicini, Giuseppe Rossi, Karu Jayathilake, Philip A Cola, Joseph P McEvoy, Richard S E Keefe, Elizabeth M C Fisher, Pamela L St Jean, Ina Giegling, Annette M Hartmann, Hans-Jurgen Moller, Andreas Ruppert, Gillian Fraser, Caroline Crombie, Lefkos T Middleton, David St Clair, Allen D Roses, Pierandrea Muglia, Clyde Francks, Dan Rujescu, Herbert Y Meltzer, and David B Goldstein. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet*, 5(2):e1000373, Feb 2009.

[83] Sergey Nejentsev, Neil Walker, David Riches, Michael Egholm, and John A Todd. Rare variants of ifih1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387–389, Apr 2009.

[84] M Y M Ng, D F Levinson, S V Faraone, B K Suarez, L E DeLisi, T Arinami, B Riley, T Paunio, A E Pulver, P A Holmans, M Escamilla, D B Wildenauer, N M Williams, C Laurent, B J Mowry, L M Brzustowicz, M Maziade, P Sklar, D L Garver, G R Abecasis, B Lerer, M D Fallin, H M D Gurling, P V Gejman, E Lindholm, H W Moises, W Byerley, E M Wijsman, P Forabosco, M T Tsuang, H-G Hwu, Y Okazaki, K S Kendler, B Wormley, A Fanous, D Walsh, F A O'Neill, L Peltonen, G Nestadt, V K Lasseter, K Y Liang, G M Papadimitriou, D G Dikeos, S G Schwab, M J Owen, M C O'Donovan, N Norton, E Hare, H Raventos, H Nicolini, M Albus, W Maier, V L Nimgaonkar, L Terenius, J Mallet, M Jay, S Godard, D Nertney, M Alexander, R R Crowe, J M Silverman, A S Bassett, M-A Roy, C Merette, C N Pato, M T Pato, J Louw Roos, Y Kohn, D Amann-Zalcenstein, G Kalsi, A McQuillin, D Curtis, J Brynjolfson, T Sigmundsson, H Petursson, A R Sanders, J Duan, E Jazin, M Myles-Worsley, M Karayiorgou, and C M Lewis. Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry*, 14(8):774–785, Aug 2009.

[85] Dale R Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*, 74(4):765–769, Apr 2004.

[86] J R O'Connell and D E Weeks. PedCheck: A program for identification of genotype incompatibilities in linkage analysis. *Am J Hum Genet*, 63(1):259–266, Jul 1998.

[87] Michael C O'Donovan, Nicholas Craddock, Nadine Norton, Hywel Williams, Timothy Peirce, Valentina Moskvina, Ivan Nikolov, Marian Hamshere, Liam Carroll, Lyudmila Georgieva, Sarah Dwyer, Peter Holmans, Jonathan L Marchini, Chris C A Spencer, Bryan Howie, Hin-Tak Leung, Annette M Hartmann, Hans-Jurgen Moller, Derek W Morris, Yongyong Shi, GuoYin Feng, Per

Hoffmann, Peter Propping, Catalina Vasilescu, Wolfgang Maier, Marcella Rietschel, Stanley Zammit, Johannes Schumacher, Emma M Quinn, Thomas G Schulze, Nigel M Williams, Ina Giegling, Nakao Iwata, Masashi Ikeda, Ariel Darvasi, Sagiv Shifman, Lin He, Jubao Duan, Alan R Sanders, Douglas F Levinson, Pablo V Gejman, Sven Cichon, Markus M Nothen, Michael Gill, Aiden Corvin, Dan Rujescu, George Kirov, Michael J Owen, Nancy G Buccola, Bryan J Mowry, Robert Freedman, Farooq Amin, Donald W Black, Jeremy M Silverman, William F Byerley, and C Robert Cloninger. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet*, 40(9):1053–1055, Sep 2008.

[88] K J Oedegaard, T A Greenwood, A Lunde, O B Fasmer, H S Akiskal, and J R Kelsoe. A genome-wide linkage study of bipolar disorder and co-morbid migraine: Replication of migraine linkage on chromosome 4q24, and suggestion of an overlapping susceptibility region for both disorders on chromosome 20p11. *J Affect Disord*, 122(1-2):14–26, Apr 2010.

[89] Ivan Ovcharenko, Marcelo A Nobrega, Gabriela G Loots, and Lisa Stubbs. ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res*, 32(Web Server issue):W280–6, Jul 2004.

[90] Orestis A Panagiotou, Evangelos Evangelou, and John P A Ioannidis. Genome-wide significant associations for variants with minor allele frequency of 5 *Am J Epidemiol*, 172(8):869–889, Oct 2010.

[91] Ju-Hyun Park, Sholom Wacholder, Mitchell H Gail, Ulrike Peters, Kevin B Jacobs, Stephen J Chanock, and Nilanjan Chatterjee. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet*, 42(7):570–575, Jul 2010.

[92] V Peralta and M J Cuesta. How many and which are the psychopathological dimensions in schizophrenia? Issues influencing their ascertainment. *Schizophr Res*, 49(3):269–285, Apr 2001.

[93] J K Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69(1):124–137, Jul 2001.

[94] J K Pritchard and M Przeworski. Linkage disequilibrium in humans: Models and data. *Am J Hum Genet*, 69(1):1–14, Jul 2001.

[95] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, Sep 2007.

[96] Shaun M Purcell, Naomi R Wray, Jennifer L Stone, Peter M Visscher, Michael C O'Donovan, Patrick F Sullivan, and Pamela Sklar. Common polygenic variation

contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, Aug 2009.

[97] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.

[98] D Rabinowitz and N Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered*, 50(4):211–223, Jul-Aug 2000.

[99] N Risch and K Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, Sep 1996.

[100] P D Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261, Dec 1997.

[101] Laura J Scott, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, Anne U Jackson, Ludmila Prokunina-Olsson, Chia-Jen Ding, Amy J Swift, Narisu Narisu, Tianle Hu, Randall Pruim, Rui Xiao, Xiao-Yi Li, Karen N Conneely, Nancy L Riebow, Andrew G Sprau, Maurine Tong, Peggy P White, Kurt N Hetrick, Michael W Barnhart, Craig W Bark, Janet L Goldstein, Lee Watkins, Fang Xiang, Jouko Saramies, Thomas A Buchanan, Richard M Watanabe, Timo T Valle, Leena Kinnunen, Goncalo R Abecasis, Elizabeth W Pugh, Kimberly F Doheny, Richard N Bergman, Jaakko Tuomilehto, Francis S Collins, and Michael Boehnke. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, Jun 2007.

[102] A Serretti, E Lattuada, C Lorenzi, R Lilli, and E Smeraldi. Dopamine receptor D2 Ser/Cys 311 variant is associated with delusion and disorganization symptomatology in major psychoses. *Mol Psychiatry*, 5(3):270–274, May 2000.

[103] A Serretti, R Lilli, C Lorenzi, E Lattuada, and E Smeraldi. DRD4 exon 3 variants associated with delusional symptomatology in major psychoses: A study on 2,011 affected subjects. *Am J Med Genet*, 105(3):283–290, Apr 2001.

[104] Jianxin Shi, Douglas F Levinson, Jubao Duan, Alan R Sanders, Yonglan Zheng, Itsik Pe'er, Frank Dudbridge, Peter A Holmans, Alice S Whittemore, Bryan J Mowry, Ann Olincy, Farooq Amin, C Robert Cloninger, Jeremy M Silverman, Nancy G Buccola, William F Byerley, Donald W Black, Raymond R Crowe, Jorge R Oksenberg, Daniel B Mirel, Kenneth S Kendler, Robert Freedman, and Pablo V Gejman. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*, 460(7256):753–757, Aug 2009.

[105] Alison J Shield, Tracy P Murray, and Philip G Board. Functional characterisation of ganglioside-induced differentiation-associated protein 1 as a glutathione transferase. *Biochem Biophys Res Commun*, 347(4):859–866, Sep 2006.

[106] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J Hudson, Alexandre Montpetit, Alexey V Pshezhetsky, Marc Prentki, Barry I Posner, David J Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, Feb 2007.

[107] Renan P Souza, Plabon Ismail, Herbert Y Meltzer, and James L Kennedy. Variants in the oxytocin gene and risk for schizophrenia. *Schizophr Res*, 121(1-3):279–280, Aug 2010.

[108] R S Spielman, R E McGinnis, and W J Ewens. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet*, 52(3):506–516, Mar 1993.

[109] RL Spitzer, JB Williams, and J Gibbon. *Structured Clinical Interview for DSM-III-R Patient Version*, 1987.

[110] D St Clair, D Blackwood, W Muir, A Carothers, M Walker, G Spowart, C Gosden, and H J Evans. Association within a family of a balanced autosomal translocation with major mental illness. *Lancet*, 336(8706):13–16, Jul 1990.

[111] Hreinn Stefansson, Roel A Ophoff, Stacy Steinberg, Ole A Andreassen, Sven Cichon, Dan Rujescu, Thomas Werge, Olli P H Pietilainen, Ole Mors, Preben B Mortensen, Engilbert Sigurdsson, Omar Gustafsson, Mette Nyegaard, Annamari Tuulio-Henriksson, Andres Ingason, Thomas Hansen, Jaana Suvisaari, Jouko Lonnqvist, Tiina Paunio, Anders D Borglum, Annette Hartmann, Anders Fink-Jensen, Merete Nordentoft, David Hougaard, Bent Norgaard-Pedersen, Yvonne Bottcher, Jes Olesen, Rene Breuer, Hans-Jurgen Moller, Ina Giegling, Henrik B Rasmussen, Sally Timm, Manuel Mattheisen, Istvan Bitter, Janos M Rethelyi, Brynja B Magnusdottir, Thordur Sigmundsson, Pall Olason, Gisli Masson, Jeffrey R Gulcher, Magnus Haraldsson, Ragnheidur Fossdal, Thorgeir E Thorgeirsson, Unnur Thorsteinsdottir, Mirella Ruggeri, Sarah Tosato, Barbara Franke, Eric Strengman, Lambertus A Kiemeney, Ingrid Melle, Srdjan Djurovic, Lilia Abramova, Vasily Kaleda, Julio Sanjuan, Rosa de Frutos, Elvira Bramon, Evangelos Vassos, Gillian Fraser, Ulrich Ettinger, Marco Picchioni, Nicholas Walker, Timi Toulopoulou, Anna C Need, Dongliang Ge, Joeng Lim Yoon, Kevin V Shianna, Nelson B Freimer, Rita M Cantor, Robin Murray, Augustine Kong, Vera Golimbet, Angel Carracedo, Celso Arango, Javier Costas, Erik G Jonsson, Lars Terenius, Ingrid Agartz, Hannes Petursson, Markus M Nothen, Marcella Rietschel, Paul M Matthews, Pierandrea Muglia, Leena Peltonen, David St Clair, David B Goldstein, Kari Stefansson, and David A Collier. Common variants conferring risk of schizophrenia. *Nature*, 460(7256):744–747, Aug 2009.

[112] R E Straub, C J MacLean, Y Ma, B T Webb, M V Myakishev, C Harris-Kerr, B Wormley, H Sadek, B Kadambi, F A O'Neill, D Walsh, and K S Kendler. Genome-wide scans of three independent sets of 90 Irish multiplex schizophrenia families and follow-up of selected regions in all families provides evidence for multiple susceptibility genes. *Mol Psychiatry*, 7(6):542–559, 2002.

[113] P F Sullivan, D Lin, J-Y Tzeng, E van den Oord, D Perkins, T S Stroup, M Wagner, S Lee, F A Wright, F Zou, W Liu, A M Downing, J Lieberman, and S L Close. Genomewide association for schizophrenia in the CATIE study: Results of stage 1. *Mol Psychiatry*, 13(6):570–584, Jun 2008.

[114] Deniz Tastemir, Osman Demirhan, and Yasar Sertdemir. Chromosomal fragile site expression in turkish psychiatric patients. *Psychiatry Res*, 144(2-3):197–203, Nov 2006.

[115] Omri Teltsh, Kyra Kanyas, Osnat Karni, Adi Levi, Mira Korner, Edna Ben-Asher, Doron Lancet, Adnan Hamdan, Bernard Lerer, and Yoav Kohn. Genome-wide linkage scan, fine mapping, and haplotype analysis in a large, inbred, arab israeli pedigree suggest a schizophrenia susceptibility locus on chromosome 20p13. *Am J Med Genet B Neuropsychiatr Genet*, 147B(2):209–215, Mar 2008.

[116] Gilles Thomas, Kevin B Jacobs, Peter Kraft, Meredith Yeager, Sholom Wacholder, David G Cox, Susan E Hankinson, Amy Hutchinson, Zhaoming Wang, Kai Yu, Nilanjan Chatterjee, Montserrat Garcia-Closas, Jesus Gonzalez-Bosquet, Ludmila Prokunina-Olsson, Nick Orr, Walter C Willett, Graham A Colditz, Regina G Ziegler, Christine D Berg, Saundra S Buys, Catherine A McCarty, Heather Spencer Feigelson, Eugenia E Calle, Michael J Thun, Ryan Diver, Ross Prentice, Rebecca Jackson, Charles Kooperberg, Rowan Chlebowski, Jolanta Lissowska, Beata Peplonska, Louise A Brinton, Alice Sigurdson, Michele Doody, Parveen Bhatti, Bruce H Alexander, Julie Buring, I-Min Lee, Lars J Vatten, Kristian Hveem, Merethe Kumle, Richard B Hayes, Margaret Tucker, Daniela S Gerhard, Joseph F Jr Fraumeni, Robert N Hoover, Stephen J Chanock, and David J Hunter. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (*RAD51L1*). *Nat Genet*, 41(5):579–584, May 2009.

[117] Sholom Wacholder, Stephen Chanock, Montserrat Garcia-Closas, Laure El Ghormli, and Nathaniel Rothman. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 96(6):434–442, Mar 2004.

[118] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54:426–482, 1943.

[119] K.M. Weiss. *Genetic Variation and Human Disease: Principles and Evolutionary Approaches*. Cambridge University Press, 1993.

[120] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, Jun 2007.

[121] N M Williams, N Norton, H Williams, B Ekholm, M L Hamshere, Y Lindblom, K V Chowdari, A G Cardno, S Zammit, L A Jones, K C Murphy, R D Sanders, G McCarthy, M Y Gray, G Jones, P Holmans, V Nimgaonkar, R Adolfson, U Osby, L Terenius, G Sedvall, M C O'Donovan, and M J Owen. A systematic genomewide linkage study in 353 sib pairs with schizophrenia. *Am J Hum Genet*, 73(6):1355–1367, Dec 2003.

[122] Virginia L Willour, Peter P Zandi, Yuqing Huo, Tyra L Diggs, Jennifer L Chellis, Dean F MacKinnon, Sylvia G Simpson, Francis J McMahon, James B Potash, Elliot S Gershon, Theodore Reich, Tatiana Foroud, John I Jr Nurnberger, J Raymond Jr DePaulo, and Melvin G McInnis. Genome scan of the fifty-six bipolar pedigrees from the NIMH genetics initiative replication sample: chromosomes 4, 7, 9, 18, 19, 20, and 21. *Am J Med Genet B Neuropsychiatr Genet*, 121B(1):21–27, Aug 2003.

[123] Eric Q Wu, Howard G Birnbaum, Lizheng Shi, Daniel E Ball, Ronald C Kessler, Matthew Moulis, and Jyoti Aggarwal. The economic burden of schizophrenia in the United States in 2002. *J Clin Psychiatry*, 66(9):1122–1129, Sep 2005.

[124] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common snps explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–569, Jul 2010.

[125] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, 88(1):76–82, Jan 2011.

[126] F Yates. Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

[127] D V Zaykin, Lev A Zhivotovsky, P H Westfall, and B S Weir. Truncated product method for combining P-values. *Genet Epidemiol*, 22(2):170–185, Feb 2002.

[128] Bing Zhang, Stefan Kirov, and Jay Snoddy. WebGestalt: An integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, 33(Web Server issue):W741–8, Jul 2005.

[129] X Y Zhang, D F Zhou, P Y Zhang, and J Wei. The CCK-A receptor gene possibly associated with positive symptoms of schizophrenia. *Mol Psychiatry*, 5(3):239–240, May 2000.