# Computational Contributions Towards Scalable and Efficient Genome-wide Association Methodology

Snehit Prabhu

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

Columbia University
2013

ABSTRACT

Computational Contributions Towards Scalable and Efficient Genome-wide Association Studies

Snehit Prabhu

Genome-wide association studies are experiments designed to find the genetic bases of physical traits: for example, markers correlated with disease status by comparing the DNA of healthy individuals to the DNA of affecteds. Over the past two decades, an exponential increase in the resolution of DNA-testing technology coupled with a substantial drop in their cost have allowed us to amass huge and potentially invaluable datasets to conduct such comparative studies. For many common diseases, datasets as large as a hundred thousand individuals exist, each tested at million(s) of markers (called SNPs) across the genome.

Despite this treasure trove, so far only a small fraction of the genetic markers underlying most common diseases have been identified. Simply stated - our ability to predict phenotype (disease status) from a person's genetic constitution is still very limited today, even for traits that we know to be heritable from one's parents (e.g. height, diabetes, cardiac health). As a result, genetics today often lags far behind conventional indicators like family history of disease in terms of its predictive power. To borrow a popular metaphor from astronomy, this veritable "dark matter" of perceivable but un-locatable genetic signal has come to be known as missing heritability.

This thesis will present my research contributions in two hotly pursued scientific hypotheses that aim to close this gap: (1) gene-gene interactions, and (2) ultra-rare genetic variants - both of which are not yet widely tested. First, I will discuss the challenges that have made interaction

testing difficult, and present a novel approximate statistic to measure interaction. This statistic can be exploited in a Monte-Carlo like randomization scheme, making an exhaustive search through trillions of potential interactions tractable using ordinary desktop computers. A software implementation of our algorithm found a reproducible interaction between SNPs in two calcium channel genes in Bipolar Disorder. Next, I will discuss the functional enrichment pipeline we subsequently developed to identify sets of interacting genes underlying this disease. Lastly, I will talk about the application of coding theory to cost-efficient measurement of ultra-rare genetic variation (sometimes, as rare as just one individual carrying the mutation in the entire population).

# Table of Contents

**List of Main Figures**

# Acknowledgements

The study and advancement of science is rarely a solitary adventure. I am the beneficiary of many fortuitous relationships – personal, professional and often both – relationships that have contributed in immeasurable ways towards helping me realize this modicum of work and my professional identity as a life scientist.

To my graduate mentor, Itsik: thank you for believing in my unrealized potential and accepting me as your protégé at a crucial phase in your own academic career. To my thesis committee – Ken Ross, Shaun Purcell, Iuliana Ionita-Laza and Rocco Servedio: thank you for your investment and your thoughtful advice. To Microsoft Research: thank you for having the courage and vision to fund promising new scientific areas that are only tangentially related to your own business. To all my wonderful lab-mates (past and present) at Columbia, particularly Eimear, Sasha, Pier, Vlada, Ben, Nate and Yufeng: thank you for your professional collegiality, your mentorship and your friendship. To Jagir, Deepak, Umasu, Albee, Dr. Bera and other colleagues at IBM Bangalore: if it wasn't for the seed of scientific curiosity you planted in me during my term at the TIC, none of this may have come to pass. And to Columbia University: you are the right mixture of frustrating and brilliant. Among all institutions whose walls I have had the honor of learning within, you have taught me the most.

Finally, Mum, Dad, Utsav and Manasi: thank you for your patience, your support, your infinite love and your very real sacrifices. Such an unlikely sequence of events, don't you think?

# Chapter 1. Introduction

In June of 2000 at a crowded press conference in the white house, President Clinton announced that the first effort to map the entire DNA sequence of a human being had been successfully completed (White House Press Release). From its inception, the effort had been eulogized as a landmark scientific endeavor: one that would see the creation of the first roadmap of the human genome, the bedrock upon which many a future medical breakthrough would rest, and the forbearer of a genetic information based public healthcare revolution. High stakes – some commercial, some controversial, and most pertaining to a scientific legacy (a.k.a. "props") – saw the emergence of a fierce race to the finish line between two groups, with the large government-funded Human Genome Project (HGP) initiative led by Francis Collins pitted against a private biotechnology startup called Celera Genomics founded by Craig Venter. As it turned out, the teams were relying on two very different experimental strategies, revealing an important difference in ideology. The HGP would use the more painstaking and deliberate approach of sequencing DNA in fragments of staggered length in the wet-lab, thereby making the genome-jigsaw easier to assemble computationally. Meanwhile Celera would rely on the (at the time quick-and-dirty) shotgun sequencing approach that only provided short, less informative DNA reads, but it did so more economically and at a faster pace. These short DNA strings would be used in conjunction with sophisticated algorithms running on large supercomputers to manage a more difficult assembly process. Eventually, the race was called a tie and credit was officially awarded to both teams, with both Venter and Collins standing beside the president during the historic announcement.

In a way, the effort to sequence the first genome marked the unofficial birth of the now burgeoning field of computational genetics. Along with a few surprises (e.g. the human genome only had ~20K genes, not 100K as predicted by most models at that time), the project had underscored the importance of advanced computational and statistical methods in studying all things *genome*. Another outcome was that dropping costs and established protocols would allow concurrent, cheaper and more informative sequencing efforts. Two such projects, the Human Genome Diversity Panel (HGDP) and the Haplotype Map (Hapmap) (Hapmap Consortium 2003; Cann 2002) were outlined to study the genetics different human populations and to develop a better understanding of the entire spectrum of genetic variation – both between and across populations – on a genome-wide scale. A feature of particular interest was SNPs (Single Nucleotide Polymorphisms): sites at which at least an estimated 5% of humans were polymorphic. These "common" genetic variations were initially defined as positions in the DNA where >5% of chromosomes in the tested populations harbored one nucleotide (say A), while the <95% remaining chromosomes in the population carried a different nucleotide (say G). The subsequently revised, broader definition used today considers any polymorphisms straddling a 1%-99% division to be a SNP. Although the existence of these polymorphisms was well known, their incidence across the whole genome was systematically characterized for the first time.

Importantly, identifying all SNPs would bring us closer to understanding the connections between genetic variation and differences in physical characteristics. Since the genomes of two randomly chosen individuals differ in very few locations (a current estimate is 0.1%, although the exact sites difference may vary from individual-pair to individual-pair), in some sense the most commonly occurring sites of variation in a population could be the most important drivers

of trait variation. The first genetic studies on a genome-wide scale were to focus on these "*eigen*"-sites[1]. Once a identified, much cheaper technology could be used to test an individual at a particular site of interest (for example, whether an individual was AA, AG or GG). This technology, called genotyping by hybridization, had so far only been applied to test a few hundred known polymorphisms on a few genes of interest. With the complete catalog of SNPs, new assays containing hundreds of thousands of sites allowed genotyping human DNA at an unprecedented scale.

Although the choice of a 5% (later 1%) frequency threshold was arbitrary and stemmed largely from technological limitations, the availability of a low-cost high-quality picture of genetic variation across the whole genome (and not just a few genes or loci) gave rise to a decade of genome-wide association studies (GWAS). These were experimental designs that compared the genomes of related or unrelated individuals in the hope of finding polymorphisms that were correlated with phenotypic (i.e. trait) differences. The following decade saw an onslaught of successful studies on a wide variety of diseases and traits, each uncovering a few causal variants by examining millions of markers (Manolio et al. 2009). However, it quickly became apparent that runaway successes – finding a single genetic variant that explained all or most of disease's incidence – were mostly limited to rare genetic aberrations affecting a small minority of the population. These were diseases whose genetic "simplicity" could be inferred by observing their

[1] It is important to note the difference between correlation and causality. Although it was unlikely that SNPs themselves would be responsible for differences in traits, they were expected to reveal causal variants in their close proximity that were. This ability of genetic polymorphisms to "tag" one another is due to the existence of linkage blocks – stretches of DNA that are likely to be inherited together from one's parents. Intuitively, this results in polymorphisms close to each other becoming correlated. Asymptotically, the $r^2$ (linkage disequilibrium) between two sites on the DNA decreases quadratically with physical distance.

classical inheritance pattern in families, in accordance with Gregor Mendel's canonical principles.

While the utility of genetic analysis on so-called *Mendelian* traits was unquestionable, for more common diseases of greater public health interest like diabetes, heart disease, immune disorders and most cancers, the genetic bases remained difficult to elucidate. Given their non-mendelian (and hence, "complex") segregation patterns within families, an early guesstimation was that a set of variants, each contributing a partial effect towards the phenotype status were more likely to be responsible. That is, instead of a single fully penetrant mutation like in the case of Mendelian traits, there would be multiple variants, each of incomplete penetrance (see Figure 1-1). This prevailing opinion was formally stated in 1996 as the Common Disease Common Variant (CDCV) hypothesis (Risch and Merikangas 1996). CDCV held that a few common variants like SNPs – perhaps half a dozen or so – would be responsible for the entire genetic component of each common complex disease. This cautiously optimistic statement was made testable by the rapidly improving technology of the day ($\sim 10^5$ SNP sites ascertained across the whole genome of each individual). Since then, genome-wide association studies under the CDCV design have had quite a few successes (GWAS catalog).

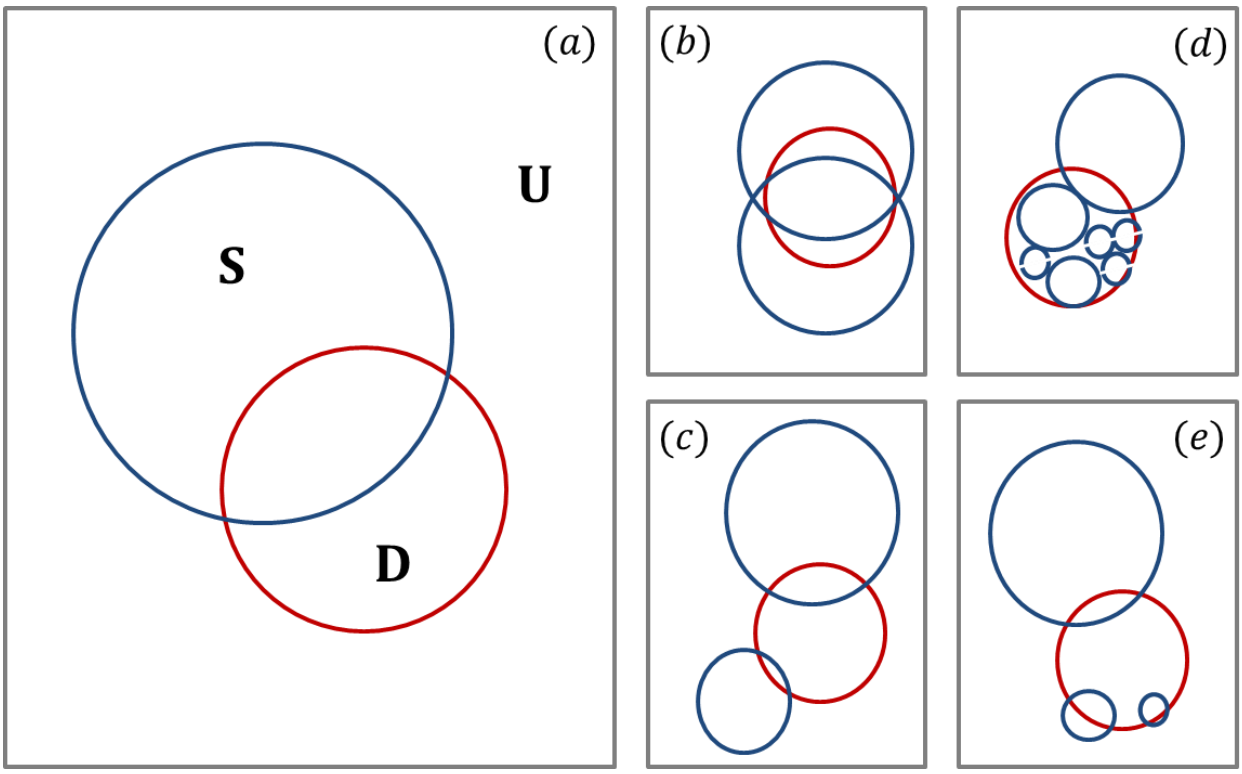



Figure 1-1. Venn diagram panels to visualize key terms.

(a) Consider a population $U$ (grey box) containing $D$ individuals affected by some common disease (red circle). The disease has an estimated $prevalence = {}^{D}/_{U}$. Consider a SNP (polymorphic site of DNA) in this population (blue circle). The individuals carrying the mutant (non-reference) allele lie within $S$, while the rest lie outside. The estimated effect size of this SNP is usually represented either by its $penetrance = {}^{S \cap D}/_{S}$, its $odds = {}^{S \cap D}/_{S \cap \bar{D}}$, or its $odds\ ratio = \left({}^{S \cap D}/_{S \cap \bar{D}}\right) \div \left({}^{\bar{S} \cap D}/_{\bar{S} \cap \bar{D}}\right)$.

(b) The "Common Disease Common Variant" hypothesis predicted that a few commonly occurring SNPs (large blue circles) would account for most or all of the incidence of common diseases.

(c) After a decade of CDCV designed GWAS, statistically significant SNP associations were made for most diseases, but accounted for very little of the disease incidence –a few notable exceptions include Age-related Macular Degeneration.

(d) The present-day zeitgeist is to seek rare, high-penetrance mutations (small blue circles) each of which account for the high-disease incidence in genetic relatives (this may include socially unrelated individuals who are *identical by descent*), and cumulatively account for most of the disease in the population.

(e) Again, notwithstanding a few exceptions, recent results seem to indicate that this is not the case for a majority of common diseases, at least concerning variants within the 1-5% frequency range. However, the hypothesis itself is far from spent, as whole-genome sequence (as opposed to genotype tagging) based GWAS designs that capture every single variation become economical.

An unequivocal way to measure the contribution of genetic associations is to test our ability to predict an individual's phenotype (like disease status) from genotype. Stated in terms of a linear regression for rhetorical purposes, if we code the trait as a dependent variable ($Y$), and genetic

variation like SNPs as predictive features ($X$), we can ask questions about the predictive power of genetics: $Y \sim f(X)$. We recall here that one arrives at the best-fit in a linear regression by minimizing squared residuals – the proportion of variation in the dependent variable $Y$ that simply cannot be accounted for using a linear combination of the predictors $X$. As it turned out, although GWAS had statistically associated thousands of SNPs with nearly a hundred common diseases, the trait variance explained by these genetic features remained miniscule (Manolio et al. 2009). The fact that the phenotypes chosen were known to have a clear and perceivable genetic basis was established through orthogonal methods – for example, monozygotic twins bearing strong resemblance in their trait value, or, the ability to predict offspring trait values from parental trait values with high accuracy. Despite this, finding the precise genetic loci responsible for this trait (predictive power) and therefore by extension, explaining variation of the trait in the larger population (by minimizing residual variance) has turned out to be quite difficult. This has come to be known as the "missing heritability" problem (Maher 2008).

Over the past few years (roughly 2008 onwards, which was the period of my PhD), several hypotheses have been put forward to bridge this gap between heritable phenotypic variance and explained genotypic variance. One ambitious viewpoint held that the functional forms used to model genotype-phenotype relationship – i.e. $f$ in $Y \sim f(X)$ – were inadequate. So far, GWAS methods have used kernels that model SNPs acting in isolation, and in a manner that lets their cumulative effect on disease predisposition behave additively. The alternate viewpoint predicted that higher-order models incorporating combinations of variants (e.g. interaction kernels on pairs or triples of SNPs) might expose causality (Marchini et al. 2005). There was certainly sufficient biological precedent for this phenomenon in model organisms like plants and mice. Moreover,

several theoretical results and studies on simulated human datasets suggested the widespread existence of interactions that leave a negligible-to-nil association signature on their constituent SNPs (Cordell 2009; Evans et al. 2006a). If this were true, it would explain to some extent why traditional GWAS have found it so difficult to associate these loci. More pertinently to me, the reason this hypothesis had not made much headway as yet was due in no small part to the formidable computational challenge of testing pairwise (or indeed three-way) combinations of millions of SNPs routinely assayed on contemporary genotyping chips – a perfect recipe for stimulating CS PhD research.

The first research section of my doctoral thesis (Chapter 2) describes contributions made towards genome-wide SNP-SNP interaction testing. Specifically, we focused on alleviating the $\mathbb{O}(n^2)$ burden of testing all pairwise combinations of SNPs adopted by most state-of-the-art methods. While a few methods had attempted this by making lossy, simplifying assumptions (pre-selection of a few candidate SNPs based on biology, or in a greedy approach based upon their marginal association level), these had failed to get at the heart of the problem. Our objective was to address these limitations from both a theoretical and a software implementation standpoint.

First, we noted that in accordance with established methodology, SNP-SNP interactions would need to exceed a threshold of statistical significance (in this case, a *p-value*) before they might be robustly claimed as interactions. Since the number of SNP-pairs we test is in the order of trillions, statistical procedure dictates that this threshold is set fairly high – or in our case, a very low p-value cutoff. This is done in order to avoid false positive discoveries: red herrings where the data might look indicative of interaction, but in fact can occur purely by chance. Conversely,

given the limited size of human GWAS datasets (empirical evidence), any SNP-pair that exceeded the significance cutoff would also leave behind telltale statistical signatures. To our advantage, one of these signatures was exploitable by a classic Monte-Carlo randomization scheme. Our software implementation of this algorithm, SIXPAC, provides a solution to this computationally hard problem with high probability of success, within a fraction of the compute time used by comparable deterministic algorithm. As it turned out, the "limitation" of low statistical power (i.e. dataset size) could be exploited for a computational benefit.

Chapter 3 of this thesis addresses the other side of the coin – statistical limitations of genome-wide interaction mapping due to insufficient data points (i.e. samples). Besides building more powerful statistical indicators, the most widely used computational approach to extract biological insight from vanilla GWAS is called functional enrichment. Briefly, the procedure involves grouping SNPs into functional categories and to test the significance of each group. We used one of the most widely used and well established functional definition database, the Gene Ontology (GO). As the name indicates, the GO database assigns function to genes rather than SNPs, and therefore necessitates a staggered approach: combining SNP p-values into gene p-values, before subsequent testing for enrichment of gene-groups is possible. Although techniques to do ontology enrichment from a list of SNP associations were adequately described in the literature, ontology-ontology enrichment from a list of SNP-SNP interactions had never before been attempted in an exhaustive, genome-wide manner. The availability of high-speed software like SIXPAC now made this possible: motivating us to adapt the entire single-locus enrichment pipeline to interaction analysis for the first time. In this chapter, we describe the algorithm implemented by the software module – FUNGI – and some of its early results on bipolar

disorder. Interestingly, these results suggest the presence of interacting/dependent biological function groups underpinning this trait, although additional validation will be required to make a strong claim.

Rather than interaction, a different hypothesis seeking to explain missing heritability has garnered the most scientific attention of late. This theory posits that rare variations – perhaps even mutations limited to one individual or one affected family in the population – might cumulatively be responsible for a large portion of common disease incidence (McClellan and King, 2010). In other words, although nosological labeling of patients may suggest the same disease, many and disparate genetic aberrations might underlie each ailment, and perhaps the perceived pathophysiological effect of each variation is what appears to be the same. Essentially, the hypothesis furthers the idea that a main (linear additive) effects architecture – the simplistic $f$ in our metaphorical $Y \sim f(X)$ – is still the most likely to find new loci. The difference is that instead of genotyping a few common markers as we did earlier, we can apply cheap sequencing technology to directly ascertain these rare and ultra-rare variants today. Moreover, if this increase in our set of predictor variables is accompanied by a loss of statistical power (known as a propensity for *overfitting* in machine learning literature), then we might best address this by increasing the size of our datasets. Lastly, to ameliorate any concern that a single ultra-rare variant can only account for so much disease incidence (see very small blue circles in Figure 1-1) is the fact that rare genetic variation is turning out to be more ubiquitous than we expected (McClellan and King, 2010). Some estimates suggest that almost every site in the DNA has existed in a polymorphic state in some population at some point in history. Overall, this

hypothesis has recently gathered considerable momentum today, concurrently with the onset of cheap yet high-quality sequencing technology (although the direction of causality is debatable).

If it is true, then validating the rare-variant hypothesis with current technology will require a tremendous economic investment – whole-genome sequencing of tens of thousands of individual does not come cheap. However, it is far more likely that advances in *in silico* methods will help us make experimentation feasible, just like it has in the past. As an indication, the first human genome sequence cost the HGP an estimated USD $3 \times 10^9$ (or approximately $1 per site, given the 3.3 billion nucleotide sequence of human DNA), while it cost Celera approximately USD $3 \times 10^8$. Since then, although sequencing throughput has increased rapidly, costs have spiraled downward to an approximate USD $3 \times 10^3$ per genome today, representing a million fold reduction *in vitro*. An ironic outcome of this pace of development (faster than Moore's law (NHGRI whitepaper)) has been a shift in the cost bottleneck from sequencing chemistries to information storage and processing. Whereas early projects would handle a few million precious DNA reads, today's machines routinely generate billions of reads that present a high-quality but several-fold redundant picture of the sequenced genome. The erstwhile intangibles of storing, computational mapping and assembling of the genome from its pieces now dictate a sizeable fraction of most projects' budgetary allocations. Regardless of the source of cost, it is safe to say that an efficient strategy is to exercise restraint with sequencing throughput, but to retain enough redundancy that allows us to confidently discover ultra-rare variation.

One promising idea is pooled DNA sequencing. As the name implies, a single culture is prepared from DNA belonging to several individuals, and this mixed sample is processed on a single

sequencing lane – as if we were sequencing DNA from a single individual as usual. Although this technique allowed us to detect the presence of a rare variant and to estimate its frequency in the dataset, it is obvious that the identity of the mutation carrier is lost. The utility of intelligently designed pools – sequencing each individual in a unique set of lanes so that it was possible to reconstruct carrier identity – was apparent, and it presented us with an opportunity to apply off-the-shelf coding schemes that are so widely studied in computer science. Besides introducing relatively straightforward concepts of logarithmic codes and error correcting codes to this application, we also studied more applied issues of random/probabilistic errors that creep in at various stages of the sequencing pipeline. Details and results are discussed in Chapter 4.

# Chapter 2. Ultrafast SNP-SNP interaction mapping in large genetic datasets.

*Summary:* Long range gene-gene interactions are biologically compelling models for disease genetics and can provide insights on relevant mechanisms and pathways. Despite considerable effort, rigorous interaction mapping in humans has remained prohibitively difficult due to computational and statistical limitations. In this section, we introduce a novel algorithmic approach to find long-range interactions in common diseases using a standard two-locus test which contrasts the linkage disequilibrium between SNPs in cases and controls. Our ultrafast method overcomes the computational burden of a genome × genome scan by employing a novel randomization technique that requires 10X to 100X fewer tests than a brute-force approach. By sampling small groups of cases and highlighting combinations of alleles carried by all individuals in the group, this algorithm drastically trims the universe of combinations while simultaneously guaranteeing that all statistically significant pairs are reported. We show that our implementation can comprehensively scan large datasets (2K cases, 3K controls, 500K SNPs) to find all candidate pairwise interactions (LD-contrast $p < 10^{-12}$) in a few hours – a task that typically took days or weeks to complete by methods running on equivalent desktop computers. We applied our method to the Wellcome Trust bipolar disorder data and found a significant interaction between SNPs located within genes encoding two calcium channel subunits: *RYR2* on *chr1q43* and *CACNA2D4* on *chr12p13* (LD-contrast test $p = 4.6 \times 10^{-14}$). We replicated this pattern of inter-chromosomal LD between the genes in a separate bipolar dataset from the GAIN project, demonstrating an example of gene-gene interaction that plays a role in the largely uncharted genetic landscape of bipolar disorder.

## 2A. Background.

Genome-wide association studies (GWAS) have successfully identified hundreds of genetic markers associated with a wide range of diseases and quantitative traits (Hindorff et al. 2009; Manolio et al. 2009). Unfortunately, for most common diseases, nearly all associated variants have small effect sizes and taken together explain very little of the genetically heritable variation of the phenotype (Craddock 2007) - a phenomenon often posed as the conundrum of "missing heritability" (Maher 2008). Furthermore, single-locus association methods (e.g. considering one SNP at a time and measuring its association levels to the phenotype) tend to implicate individual genes in a particular disease or trait, which in turn highlight a single biological entity involved (Saunders et al. 1993; Hugot et al. 2001; Benjamin M Neale et al. 2010). They do not, by definition, seek to implicate links between the functional elements of a system or elucidate pathway connections that may be broken. Investigation of joint gene–gene effects can therefore improve the explanatory ability of genetics twofold. Firstly, interaction - or statistical epistasis, as defined by (Fisher 1918) - is hypothesized to explain a part of disease heritability (Evans et. al., 2006; Marchini et. al., 2005) Secondly, finding significant statistical links (epistatic or otherwise) between genes could provide strong indications of molecular-level interactions that differ between cases and controls.

However, an all-pairs (or all-triples) scan of SNPs genome-wide still poses widely discussed computational challenges due to the sheer size of the combinatorial space (J Marchini et al. 2005), both for data sets typed on genotyping arrays ($\sim 10^6$ SNPs) and sequencing technologies ($\sim 10^7$ single nucleotide variants - or SNVs). Some methods address this problem by restricting their combinatorial analysis to a small subset of "candidate" markers - those identified through

single-locus analysis or those of biological interest (Emily et. al., 2009) or by only checking for interactions between SNPs that are physically close to one another on the genome (Slavin et al. 2011). Others like EPIBLASTER (Kam-Thong et al. 2010) and SHIsisEPI (Xiaohan Hu et al. 2010) make use of specialized hardware like multiple Graphical Processing Units (GPUs) to finish computation on genome-wide data sets on the order of days, rather than weeks or months. While it is known that reductionist, candidate SNP-based approaches can miss many real interactions (Culverhouse et al. 2002; Evans et al. 2006b) and fail to provide novel biological insights in an unbiased manner, brute-force approaches that rely on hardware for speedup may also scale poorly as data sets increase in size and interaction tests increase in complexity.

For genome-wide interaction analysis to become pervasive, there is a pressing need for algorithmic insights that make interaction testing on large datasets a scalable proposition, without placing undue computing or hardware demands on the investigator. The contribution of our work is such a method. Recently, others had exploited the fact that contrasting/comparing the linkage disequilibrium (Jinying Zhao et al. 2006), Pearson correlation (Kam-Thong et al. 2010) and log-odds ratio (Plink *"--fast-epistasis"* option) between a pair of SNPs in cases and controls could be computed more efficiently than maximum likelihood estimates in a logistic regression. Usefully, these computationally efficient contrast tests showed high congruence with statistical epistasis under a variety of genetic models. In this study, we do not devise a new statistical test; rather, we use a simplified version of the LD-contrast test for interaction (Jinying Zhao et al. 2006) to demonstrate our computational principles. Our version seeks pairs of physically

unlinked (often inter-chromosomal) SNPs that are in strong LD in cases, but in weak LD, no LD, or reverse LD in controls.[2]

Our computational approach is driven by the intuition that most genome-wide interaction methodologies only report SNP pairs that are statistically significant (as per the test used) after correcting for the number of tests. The question we ask is this: given a statistical test, is it possible to identify approximately all the significant SNP pairs with high probability (power), without actually applying the test to all possible combinations genome-wide? In practice, can we design a search algorithm that accepts an arbitrary significance cutoff (as input from the user), and then finds all SNP pairs that will pass this cutoff without a brute-force search? We show here that for some contrast tests, this is indeed possible. At this juncture, it is imperative that we point out the two distinct meanings of "*power*": Here, unless otherwise specified, we mean the power of an algorithm to identify SNP pairs for which a test statistic is large (i.e., significant), whereas in the broader context of genome-wide interaction mapping literature, power is the ability of a statistical test to detect a real interaction in the data set. Our work focuses on addressing the computational issues that plague an exhaustive search for interaction, leaving issues of statistical power for a separate discussion.

The rest of this chapter is structured as follows. First, we briefly review a simple LD-contrast test that compares LD between binary allelic states (rather than 0/1/2 genotypes) in cases and controls. Next, we present a novel computational framework—probably approximately complete

[2] Disequilibrium between physically unlinked loci is also often called Gametic Phase Disequilibrium (Wang et al. 2010), but for purposes of this study, we consider both terms equivalent—in particular, we do not imply physical linkage/proximity on the genome with the term LD.

(PAC) testing—that quantifies the power of a search done by an algorithm. PAC is an intuitive concept: For example, a brute-force method that tests all-pairs of SNPs genome-wide is considered fully powered at finding all significant pairs in our framework (i.e., 100% probability of finding all pairs whose test statistic clears the significance cutoff) and have no element of approximation at all (i.e., 100% complete scan of the interaction space in the case-control data set). In this study, we design a two-stage PAC test for common complex diseases that is guaranteed to find all significant pairwise interactions with high power (e.g., probability >95% of finding all pairs with a significant statistic) by looking at almost the entire space of possibilities (e.g., ~99% complete scan of interaction space). In return for accepting a small loss of certainty and power, we show that algorithms that offer tremendous computational gains can be designed. We evaluate the performance of our implementation of this framework (SIXPAC) on genome-scale data and then present the results of our analysis on bipolar disorder (BD) in the Wellcome Trust Case Control Consortium (WTCCC) data set (Craddock 2007).

## 2B. Methods.

***Outline:*** The goal of our method is to efficiently identify the set of SNP-pairs which have vastly different LD in cases and controls from the universe of pairs genome-wide - if any such pairs exist at all. First, we define the LD-contrast statistic and establish a minimum cutoff value that determines whether a pair of SNPs has a statistically significant contrast in a genome-wide study or not. Next, we devise a stage-1 filtering step that identifies potential case-control differences in LD by looking for LD in cases alone. We quantify the losses that stage-1 incurs (false negatives) by applying this "approximate" version of the full LD-contrast test.

In stage 2, the candidates shortlisted based on their LD in cases are tested using the full cases-versus-controls LD-contrast test, and either validated or discarded based on the difference. Stage 2 is needed to distinguish stage-1 shortlisted candidates that are true interactions from false positives. False positives may include SNP-pairs drawn by pure chance, and also pairs which show large LD in cases, but also show large LD in controls in the same direction. Such a systemic inflation of disequilibrium between alleles in cases and controls might be due to other factors like population stratification, technical artifacts or ascertainment bias and is, by definition, not associated with phenotype.

The motivation for dividing the search into two stages is because the stage-1, case-only, "*approximate*" filtering step can be processed extremely rapidly by exploiting computer bit-wise operations, making it much faster than a brute-force approach. We present the novel randomization technique called *group-sampling* with which we can efficiently find SNP-pairs that are in strong LD in cases. However, like every randomization algorithm, we need to stop

sampling when we are reasonably certain that all significant (high LD) candidates have already been encountered and shortlisted. Consequently, at the end of stage-1, we are left with a "*probably complete*" list of pairs that demonstrate severe LD in cases. Taken in conjunction, this design outputs a "*Probably Approximately Complete*" (PAC) catalog of interacting SNP-pairs at the end of the filtering stage, which are subsequently screened by the full test. We demonstrate that our software implementation of this PAC-testing framework can find approximately all significant SNP-pairs in current GWAS datasets with arbitrarily high power (e.g. >99% probability) at a fraction of the computational cost of an exhaustive search.

**Definitions and Notation:** For purposes of illustration, consider two binary matrices $X_{N\times M}$ and $x_{n\times M}$, representing the cohorts of $N$ haploid cases and $n$ haploid controls typed at $M$ polymorphic sites respectively (we will extend this to the diploid human case later). $X_{i,v}$ denotes the allele carried by case $i$ at variant site $v$ (0 for major, 1 for minor), while $x_{j,v}$ similarly denotes the allele carried of control $j$ at that site. Further, we respectively denote $X_{.v}(a) = \left|\{i|X_{i,v} = a\}\right|$ and $x_{.v}(a) = \left|\{j|x_{j,v} = a\}\right|$ as the number of cases and controls that carry allele $a = \{0,1\}$ at $v$. Therefore, $P_v(a) = X_{.v}(a)/N$ and $p_v(a) = x_{.v}(a)/n$ are the corresponding allele $a$-frequencies of $v$ in cases and controls. Since we are only discussing binary carrier states $(0/1)$, for ease of notation we henceforth use $P_v$ instead of $P_v(1)$, and $(1 - P_v)$ instead of $P_v(0)$ (and analogously, $p_v$ and $1 - p_v$ for controls).

We are interested in examining whether a haploid individual carries a certain combination of alleles at two (or more) sites. Consider $s$ different binary sites $\vec{v} = (v_1, \dots, v_s)$, at which an

individual can carry any one of $2^s$ unique allelic combinations. We say an individual carries allelic state $\vec{a} = (a_1, \dots, a_s) \in \{0,1\}^s$ , at these sites if she carries allele $a_i$ at each one of the respective sites $v_i$. Analogous to individual sites, we can also denote the $2^s$ different $\vec{a}$-frequencies of $\vec{v}$ by $P_{\vec{v}}(\vec{a}) = X_{\vec{v}}(\vec{a})/N$ in cases and $p_{\vec{v}}(\vec{a}) = x_{\vec{v}}(\vec{a})/n$ in controls, where $X_{\vec{v}}(\vec{a}) = |\{i|X_{i,\vec{v}} = \vec{a}\}|$ and $x_{\vec{v}}(\vec{a}) = |\{j|x_{j,\vec{v}} = \vec{a}\}|$ are the number of $\vec{a}$ carriers at $\vec{v}$ in cases and controls respectively. For example, if an individual carries 1-alleles (i.e. minor alleles) at each of the sites $\vec{v} = (v_1, \dots, v_s)$, then we say she is a $\vec{1}$-carrier of $\vec{v}$. The $\vec{1}$-frequency of $\vec{v}$ in cases (controls) is the fraction of cases (controls) that are $\vec{1}$-carriers of $\vec{v}$.

**Binary representation of diploid genomes:** For diploid genomes like humans, equivalent matrices of cohorts would be $G_{N\times M}$ for cases and $g_{n\times M}$, for controls, where each entry $\{0,1,2\}$ in these matrices represents the number of minor alleles at the site, rather than presence or absence of a minor allele. Depending on the model of interaction the investigator is interested in, these may be transformed into an appropriate binary representation in several ways. For our purpose, we represent each ternary genotype as two binary variables. The first variable asks whether the individual carries $\geq 1$ copies of the minor allele (i.e. is dominant) at this SNP, while the second asks whether the individual carries exactly 2 copies of the minor allele (i.e. is recessive) at this SNP. In this format, cases and controls are represented by the binary matrices $X_{N\times 2M}$ and $x_{n\times 2M}$ respectively, where each genotype $G_{i,v}$ is recoded as two binary values $\{X_{i,2v-1}, X_{i,2v}\}$ for cases,

$$X_{i,2v-1} = \begin{cases} 0 & if\ G_{i,v} < 1 \\ 1 & if\ G_{i,v} \geq 1 \end{cases} \quad \text{and} \quad X_{i,2v} = \begin{cases} 0 & if\ G_{i,v} < 2 \\ 1 & if\ G_{i,v} = 2 \end{cases}$$

and $g_{j,v}$ is recoded equivalently as $\{x_{j,2v-1}, x_{j,2v}\}$ for controls. For example, case #6 is represented as a recessive carrier of SNP #12 (variable coordinates: row 6, column $2 \times 12 = 24$) by setting $X_{6,24} = 1$. If case #6 is a dominant carrier of SNP #12 then we set both $X_{6,23} = 1$ and $X_{6,24} = 1$. The notations for number of carriers and frequency of variables (and combination of variables) all follow analogously.

**Statistical Test for Two-Locus Effect:** We adapt the LD-contrast test for interaction between a pair of unlinked genotypes (Zhao et al., 2006) into a similar two-tailed test between a pair of unlinked binary variables $\vec{v} = (v, v')$ ,

$$LD_{\vec{v}}^{diff} = \frac{D_{\vec{v}}^{case} - D_{\vec{v}}^{control}}{\sqrt{\left(\sigma_{\vec{v}}^{case}\right)^2 + \left(\sigma_{\vec{v}}^{control}\right)^2}} \sim \mathcal{N}(0,1)$$

Equation 2-1

where $D_{\vec{v}}^{case}$ and $D_{\vec{v}}^{control}$ represent the estimated LD between these variables in cases and controls respectively, while $\sigma_{\vec{v}}^{case}$ and $\sigma_{\vec{v}}^{control}$ represent the standard error of these estimators (see 7A. for derivation and details) and $LD_{\vec{v}}^{diff}$ is their LD-contrast. This normalized statistic behaves as a Z-score, and for variable-pairs that pass the significance cutoff in a genome-wide pairwise analysis (typically $p < 10^{-10}$ or less on present day datasets), this statistic will assume large values (typically 6 or more).

Variable-pairs with large differences in LD are of interest to several genetic models, and their signal can be dissected to either reveal statistical (epistatic) or biological interaction. Based on what is known about the genetic architecture of a specific disease, the relevant community of geneticists can bring different model assumptions to bear on a test for interaction. Here, we do not attempt to dictate a specific model that might cause such a difference in LD between the cases and controls. Rather, we focus on presenting a general method that can report all SNP-pairs with a significant contrast and provide expert users with the flexibility to filter the results from such an analysis according to relevant assumptions. This can be done either *apriori* (e.g. removing SNPs with marginal signals before running a search for interaction), or *aposteriori* (e.g. discarding reported SNP-pairs that do not provide evidence for statistical epistasis).

**Two-stage testing design:** A widely used simplification (Cordell 2009; Piegorsch et al., 1994; Yang et al., 1999) in genome-wide interaction scans is to divide the search effort into two stages - first filter candidates, and then verify interaction. The crucial insight that permits this step is that we can expect physically unlinked markers to be in (or almost in) linkage equilibrium in large outbred populations. Even for common diseases, the general population is mostly comprised of healthy controls (disease prevalence < 50%). We show that in the absence of confounding factors like population stratification a pair of physically unlinked variables showing large LD-contrast will be a pair which has large LD in cases rather than large LD in controls. Without loss of generality, we focus our discussion on identifying pairs with strong positive LD in cases $\left(LD_{\vec{v}}^{cases} > 0\right)$. Pairs with strong negative LD between variables are easily modeled (with a trivial change in binary encoding) as strong positive LD between the major allele at one and a minor allele at the other. Alternative variable-pairings of this kind would only require a

different binary encoding scheme, but introduce more confusing notation. A separate (but limiting) issue is that of the statistical testing burden incurred by encoding alternate models, which we address in the discussion. A sequential two-stage testing strategy is designed as follows.

*Stage 1 (Shortlisting):* The stage 1 null-hypothesis states that any pair of distal variables $\vec{v} = (v, v')$ should be in linkage equilibrium in cases.

$$\mathbb{H}_0' : LD_{\vec{v}}^{case} = \frac{D_{\vec{v}}^{case}}{\sigma_{\vec{v}}^{case}} = 0$$

Equation 2-2

From Equation 7-1 we know the distribution of $LD_{\vec{v}}^{case}$ is $\mathcal{N}(0,1)$. We shortlist only those variable-pairs that reject the stage 1 null hypothesis at a significance level of $\mathcal{B}'$. In other words, for a pair to be shortlisted as a candidate for follow-up, we require that the LD in cases between its variables should exceed some threshold - i.e. $LD_{\vec{v}}^{case} \geq z_{\mathcal{B}}'$. We will determine this threshold to satisfy sensitivity/specificity constraints later.

*Stage 2 (Validating):* Next, we apply the LD-contrast test on candidates shortlisted by stage 1. This helps us to determine, for each candidate, whether the observed LD is indeed case-specific (and therefore a putative indicator of interaction) or pervasive in the population (and hence unrelated to disease). The stage 2 null-hypothesis posits that there is no LD difference between cases and controls

$$\mathbb{H}_0 : LD_{\vec{v}}^{diff} = 0$$

Equation 2-3

Putative significant pairs will reject this null hypothesis at a significance level of $\mathcal{B}$ *( i.e.* $LD_{\vec{v}}^{diff} \geq z_{\mathcal{B}}$ ).

In order to appreciate how such a two-stage design can capture almost all significant pairs in the dataset, and what the appropriate significance cutoff $z'_{\mathcal{B}}$ in the stage 1 analysis must be, we now introduce the concept of a Probably Approximately Complete Search. A numerical example depicting the concepts that follow is provided in 0

**Probably Approximately Complete (PAC) Search:** We now apply this two-stage hypothesis test into our probabilistic framework.

*A. Complete Search:*

To find all significant variable-pairs in the dataset, current algorithms would sequentially visit each pair of SNPs, genome-wide, and check whether each LD-contrast exceeds the user-prescribed significance threshold $\left(LD_{\vec{v}}^{diff} \geq z_{\mathcal{B}}\right)$ by comparing cases and controls.

*B. Approximately Complete Search:*

Here we ask, what threshold $LD_{\vec{v}}^{case} \geq z'_{B}$ can we apply in the filtering step, so as to capture almost all significant pairs by means of their disequilibrium in cases alone. In other words, can most significant pairs (pairs for which $LD_{\vec{v}}^{diff} \geq z_{\mathcal{B}}$) be captured without explicitly

determining $D_{\vec{v}}^{control}$ at all? Furthermore, we wish to determine the proportion of significant pairs that such an approximation might miss. We show that for most common diseases, an adequate cutoff for LD in cases is usually $z'_B > z_{\mathcal{B}}$ (see 7B. ) – i.e. SNP-pairs with a severe LD-contrast (difference in LD between cases and controls) are usually observable from their severe LD in cases alone.

*C. Probably Approximately Complete (PAC) Search:*

So far, our two-stage design has reduced the cumbersome task of counting the number of carriers for all variable-pairs (genome-wide) in cases and then again in controls, to the simpler task of shortlisting the small set of pairs which demonstrate $LD_{\vec{v}}^{case} \geq z'_B \geq z_B$. From a complexity standpoint however, such a simplification (restricting the stage 1 analysis to cases only) does not change the order or magnitude of the number of tests: this is still quadratic in the number of SNPs genome-wide. To address this computational problem we now introduce the novel randomization technique called group sampling, which can rapidly perform the case-only shortlisting with arbitrarily high power, without explicitly checking all pairs of variables.

Next, we describe a novel Monte-Carlo sampling technique called "Group Sampling" that allows us to identify all interacting variable-pairs in a dataset, within the PAC paradigm.

***Group Sampling***: From our observation that the LD statistic in cases, is usually more severe than LD-contrast (7B. ), we deduce that significant interacting pairs $\vec{v}$ will show a minimum number of excess $\vec{1}$-carriers in cases: $\Delta_{\vec{v}}^{case} \geq N z_B \sigma_{\vec{v}}^{case}$. In a genome-wide analysis, as the

universe of variable-pairs tested grows, so does the burden of multiple test correction that is applied to characterize statistical significance. Consequently, the number of excess of $\vec{1}$-carriers required in order for $\vec{v}$ to achieve statistical significance in cases - $\Delta_{\vec{v}}^{case}$ - grows commensurately. Group sampling overcomes the computational burden of a genome-wide analysis by using this "side-effect" of multiple-test correction to its advantage: the larger the number of variants typed, the larger is the universe of pairs to be tested, and the larger the excess $\vec{1}$-carriers needed to make statistically significant pairs stand apart from the crowd - this observation allows us to quickly prune the universe of pairs into a much smaller candidate set that is "guaranteed" to contain all significant pairs with arbitrarily high probability.

For illustrative purposes, let us consider a simplified version of the problem at hand. In this version, we are only interested in searching through pairs of distal variables $\vec{v} = (v, v')$, where both variables have 1-frequencies ($P_v$ and $P_{v'}$) that lie within the narrow frequency window $w = [\tilde{P}, \tilde{P} + \epsilon)$. Let the set of all variables that lie within this frequency window be labeled $V(w)$. We wish to determine whether there exists a pair $\vec{v} \in V(w) \times V(w)$, such that $\vec{v}$ rejects $\mathbb{H}_0'$. We can compute a lower bound on $\Delta_{\vec{v}}^{case}$ for all such $\vec{v}$ as:

$$\min_{\mathrm{w}\times\mathrm{w}}(\Delta_{\vec{v}}^{case}) \geq N \min_{\mathrm{w}\times\mathrm{w}}(\hat{\sigma}_{\vec{v}}^{case})\, \mathrm{z}_{\mathcal{B}}$$

$$= \sqrt{N}.\tilde{P}(1-\tilde{P})\, \mathrm{z}_{\mathcal{B}}$$

Equation 2-4

This is because the excess $\vec{1}$-carriers required for any $\vec{v} \in V(w) \times V(w)$ to reject $\mathbb{H}_0'$ is at least as many as the excess $\vec{1}$-carriers required by the least frequent $\vec{v}$ in that set: when $P_v = P_{v'} = \tilde{P}$. Therefore the $\vec{1}$-frequency of all pairs that reject $\mathbb{H}_0'$ is at least

$$P_{\vec{v}} \geq \tilde{P}^2 + \frac{\min_{\text{w×w}}(\Delta_{\vec{v}}^{case})}{N}$$

$$= \tilde{P}^2 + \delta_{w \times w}$$

Equation 2-5

where $\delta_{w \times w} = \frac{\tilde{P}(1-\tilde{P})}{\sqrt{N}}\, z_{\mathcal{B}}$ is the minimum LD in cases for all significant pairs $\vec{v} \in V(w) \times V(w)$.

*Randomly sampling a single group:* Consider a group of $k$ cases drawn randomly (with replacement). If $\vec{v}$ rejects $\mathbb{H}_0'$, then the probability that all $k$ cases in the group will be $\vec{1}$-carriers of $\vec{v}$ has a lower bound $(P_{\vec{v}})^k \geq \left(\tilde{P}^2 + \delta_{w \times w}\right)^k$. On the contrary, if $\vec{v}$ does not reject $\mathbb{H}_0'$, then the probability that such a group will contain all $\vec{1}$-carriers of $\vec{v}$ purely by chance has an upper bound $(P_{\vec{v}})^k \leq (\tilde{P} + \epsilon)^{2k}$ – corresponding to the most frequent variable-pair in $V(w) \times V(w)$. It is easy to see that if $\delta_{w \times w} > \epsilon$, we are much more likely to observe a random group of cases that are all $\vec{1}$-carriers of $\vec{v}$ when it rejects $\mathbb{H}_0'$.

The reason for drawing cases in groups (as opposed to one by one) is that it allows us to rapidly find the subset of variables for which all $k$ cases are $\vec{1}$-carriers. This is done with a native bitwise AND operation using computers, which is very fast in practice. In fact, the larger the group size, the exponentially smaller the subset of variables carried by all cases in the group becomes.

Furthermore, long stretches of binary genotype data can be processed per CPU clock cycle, making this step even more attractive. Subsequent to finding this small subset of variables, it is computationally efficient to enumerate all pairs (or indeed, triplets) among them, and pass them on to stage 2.

*Randomly sampling multiple groups*: If the group of cases we draw is sufficiently large (i.e. $k$ is high), then it is extremely unlikely to contain only $\vec{1}$-carriers, not only when $\vec{v}$ accepts $\mathbb{H}_0'$, but also when this null is rejected : because both $(\tilde{P} + \epsilon)^{2k}$, $\left(\tilde{P}^2 + \delta_{w\times w}\right)^k \ll 1$. We can counter this by drawing up to $t$ independent groups (each containing $k$ random cases), so that the probabilities of not witnessing even a single group containing only $\vec{1}$-carriers decreases at diverging rates for the two realities:

$$\left(1 - (\tilde{P} + \epsilon)^{2k}\right)^t \ll \left(1 - \left(\tilde{P}^2 + \delta_{w\times w}\right)^k\right)^t$$

In fact, if $\vec{v}$ does reject $\mathbb{H}_0'$, then by varying the two parameters $k$ and $t$ the probability of observing at least one group of all $\vec{1}$-carriers can be driven arbitrarily high (Type II error rate < $\beta$ ) while keeping the probability of a chance observation relatively low (Type I error rate < $\alpha$). In other words, given fixed specificity and sensitivity constraints $\alpha$ and $\beta$ (provided as input by the user), when $\delta_{w\times w} > \epsilon$ we can always find group-sampling parameter values $k$ and $t$ for which:

$$Sensitivity:\ 1-\left(1-\ \left(\tilde{P}^2+\ \delta_{w\times w}\right)^k\right)^t \geq 1-\beta$$

$$Specificity:\ 1-\ \left(1-\ \ (\tilde{P}+\epsilon)^{2k}\right)^t \leq \alpha$$

Equation 2-6

An illustration to visualize this technique is provided in Figure 2-1, while the simple algorithm implied by our toy problem logic is provided by Algorithm 1. The general formulation for PAC-testing across all frequency windows (genome-wide) is described in 7D. and the logic provided by Algorithm 2.

**Algorithm 1: Group Sampling Toy Problem.**

Given all variables within frequency range $V(w)=\{v \mid P_v \in w=[\tilde{P},\ \tilde{P}+\epsilon)\}$
Calculate significance threshold $\delta_{w\times w}$
Calculate sampling parameters $k$ and $t$
Repeat $t$ times :
    Randomly choose a group $C$ of $k$ cases ($k$ rows from $X_{N\times 2M}$)
    Co-carried variables $CV \leftarrow Bitwise\ AND\ (C)$
    For all unique combinations $\vec{v}=(v,v') \in CV \times CV$ :
        If $LD_{\vec{v}}^{case} \geq z'_B$ do $Shortlist \leftarrow Shortlist \cup \{\vec{v}\}$

**Algorithm 2: Group Sampling Genome-wide.**

Assign all variables genome-wide to frequency windows $W = \{w_0, \ldots, w_{r-1}\}$

For every pair of windows $\{w_A, w_B\} \in W \times W$ :

  Calculate significance threshold $\delta_{A\times B}$

  Calculate sampling parameters $k_{A\times B}$ and $t_{A\times B}$

  Repeat $t_{A\times B}$ times:

    Randomly choose group $C$ of $k_{A\times B}$ cases

    Co-carried variables $CV \leftarrow Bitwise\ AND\ (C)$

    Identify variables $CV_A \leftarrow V(w_A) \cap CV$

    Identify variables $CV_B \leftarrow V(w_B) \cap CV$

    For all unique combinations $\vec{v} = (v, v') \in CV_A \times CV_B$ :

      If $LD_{\vec{v}}^{case} \geq z'_B$ do $Shortlist \leftarrow Shortlist \cup \{\vec{v}\}$

    For all shortlisted variables $\vec{v} \in Shortlist$:

      If $LD_{\vec{v}}^{diff} \geq z_B$ output $\vec{v}$ as an interaction

This concludes our discussion of a PAC search, and the group-sampling algorithm by which one might construct a *Probably Approximately Complete* list of interacting SNPs. These theoretical results can be realized in a powerful software implementation: as we shall demonstrate next, we can find approximately all significant SNP-pairs genome-wide with high power in a fraction of the time that an exhaustive search would require.


Figure 2-1 Group sampling illustration.

Sample $k$ individuals per group, perform a simple and rapid binary AND operation that shortlist alleles carried by all individuals in the group, hash these pairs of co-occurring alleles and proceed to next sampling step. Stop sampling after $t$ iterations. We show that this sampling technique provides a probably approximately complete picture of the SNP-SNP interaction landscape at a fraction of the computational cost, when compared against enumerating and testing all pairs of SNPs genome-wide.

## 2C. Results.

The major methodological contribution of this work is a novel randomization algorithm (group sampling), which can focus the computational effort towards finding significant pairwise interaction candidates, without testing all pairs genome-wide. To determine whether a candidate SNP-pair is significant or not, and to minimize risk of false positives, in all our analyses we subject the results to the most conservative threshold for significance in a genome-wide analysis - the Bonferroni corrected p-value of 0.05 – unless otherwise stated. More sophisticated treatment of the multiple testing issues in interaction testing (e.g. (Emily et al. 2009a)) are equally applicable and can be plugged into our method without violating any of the principles or assumptions. We also restrict our analysis to pairs of genetic markers (SNPs) only, and choose to ignore gene-environment interactions for the moment. These simplifications serve to highlight the fundamental concepts of our approach, without loss of interpretable results. Our software implementation of this algorithm (SIXPAC) is available for download at http://www.cs.columbia.edu/~snehitp/sixpac.

*Dataset:* SIXPAC was used to analyze 1868 cases of the Bipolar disorder (BD) cohort in the WTCCC against 2938 combined controls from the 1958 British birth cohort (58C) and UK national blood service (NBS), all typed on the Affymetrix 5.0 platform, after cleaning all data as per requirement (Craddock 2007). Each of the remaining 455,566 SNPs remaining in the dataset was encoded into two binary variables (dominant and recessive), giving 911,132 binary variables genome-wide and a universe of $\binom{455566}{2} \times 4 = 4.15 \times 10^{11}$ potential variable-pairs to be tested. Although we only report pairwise interactions that are significant at the Bonferroni level in this

dataset ($p < 1.2 \times 10^{-13}$), investigators who employ less stringent multiple test correction can use SIXPAC to discover interactions at a different cutoff as well.

To verify that the LD-contrast statistic follows a standard normal distribution, we drew random variable-pairs genome-wide and constructed a QQ plot. Like others before (Liu et al. 2011), we observed that WTCCC data cleaning was inadequate for interaction analysis and systematically applied more stringent filters to preemptively screen out false positives which can be a result of bad genotype-calls on a few individuals. Specifically, 81085 additional SNPs which had <95% confidence calls (CHIAMO) in >1% of the individuals (cases and controls combined) were removed. For the cleaned dataset of 374,481 SNPs that remain, we verified that the LD-contrast statistic $LD_{\vec{v}}^{diff}$ for randomly drawn pairs of unlinked variables >5cM apart was indeed a Z-score (QQ plots and additional cleaning details in 7E. ), in agreement with our null hypothesis.

*Power analysis on spiked data:* Next, we tested SIXPAC's computational sensitivity by searching for synthetic interactions inserted into the bipolar cases while keeping the joint controls unchanged. 11 recessive-recessive interaction pairs between 22 SNPs on successive autosomal chromosomes (chr1 and chr2, chr3 and chr4, etc.) were simulated over a range of different parameters. Interactions between each pair of SNPs were simulated in a manner not to introduce and main effect, but effectively introduce only interaction effects. Details of this procedure are outlined in 7F.

Algorithm 2 configures the search parameters according to two user inputs: (i) a significance cutoff (LD-contrast test p-value), and (ii) the minimum search power (defined as the power to

discover all variable pairs that exceed the given significance cutoff, assuming such interactions exist). We tested SIXPAC on the synthetic datasets over a range of different input value combinations, to check whether we could discover the spiked interactions in accordance with theoretical estimates, and confirmed finding all of them at (or above) the power guaranteed to the user (0).

*Computational savings from group-sampling:* To put the computational savings of our novel approach in context, we reviewed the literature for published, high-performance, genome-wide pairwise search methodologies that either (i) contrast a statistic for a pair of SNPs between cases and controls or (ii) directly test for statistical epistasis between a pair of SNPs using a regression model. Plink (Purcell et al. 2007) offers a *--fast-epistasis* option that tests pairs of SNPs using a statistic similar to ours: specifically, it collapses each pair of SNPs completely into a 2x2 table of major vs. minor allele counts, and subsequently contrasts the odds ratios of each combination between cases and controls. On the other hand, EPIBLASTER (Kam-Thong et al. 2010) operates on the entire 3x3 table of genotypes to contrast the exact Pearson's correlation of each SNP-pair between cases and controls. Like Plink, SHEsisEPI (Hu et al. 2010) also contrasts odds-ratios of all SNP pairs reduced to a 2×2 table. Both EPIBLASTER and SHEsisEPI achieve speedup through the use of a GPU stack.

Among the methods that directly test for statistical epistasis, we report TEAM (Zhang et al. 2010) and FastEpistasis (Schüpbach et al. 2010). The authors of FastCHI (Zhang et al. 2009), FastANOVA (Zhang et al. 2008), COE (Zhang, et al. 2010) and TEAM presented a review (Zhang et al. 2011) in which TEAM was reported as the most appropriate for handling human

datasets, and was therefore chosen to represent the family of methods. TEAM achieves computational speedup by a novel approach that allows it to accurately identify interacting SNP-pairs (for most statistical tests) by checking only a small subset of individuals in the cohort. Unlike EPIBLASTER, Plink --*fast-epistasis* and SIXPAC, TEAM works directly on the logistic regression framework – giving it the ability to test a broader range of interaction models. The other method, FastEpistasis, reports epistasis in the analysis of quantitative traits (and is particularly built for gene-expression analysis) by implementing a rapid linear regression that takes advantage of multi-core processor architectures. Notable among methods omitted in this comparison are Multifactor Dimensionality Reduction (Ritchie et al. 2001) and Restricted Partition Method (Culverhouse et al. 2004), both of which partition the data according to genotypic effect in a relatively model agnostic manner. Consequently both methods test a variety of interaction models (alternate parameterizations) that are not currently captured by high-performance computational techniques like ours and others previously discussed. Another widely cited method, BEAM (Zhang and Liu 2007) does not scale to present day datasets (Cordell 2009) and was left out of this analysis. There are numerous other methods which perform whole-genome interaction scans (Liu et al. 2011; Emily et al. 2009; Achlioptas et al. 2011; Greene et al. 2010; Zhang et al. 2009), and an older review of a few of these is provided elsewhere (Cordell 2009).

Except for SIXPAC, all the time-scales presented in

Table 2-1 are performance figures as self-reported by the authors of each method (or in the case of TEAM, extrapolated from performance figures reported therein) on a dataset of this size. Our

synopsis does not constitute a comprehensive methods comparison, and is presented solely to highlight the computational savings achieved by group-sampling. The reason SIXPAC is able to achieve its speedup without GPUs is because it does not need to exhaustively test all pairs of SNPs to identify the significant combinations[3]. On the other hand all other methods are burdened by a brute-force test of all pairs to identify such combinations. In confirmation of our estimates, they also report that genome-wide testing on ordinary CPUs requires several weeks of compute time (some report weeks even on a small cluster of computers). The application of group-sampling was able to reduce this computational investment to around 8 hours.

| Method | Type of test | Computational approach | Approx. time to process dataset[4] | Run on specialized hardware |
|---|---|---|---|---|
| Plink[5] | Odds-ratio Contrast | Brute-force | Weeks | No |
| FastEpistasis | Linear Regression | Brute-force | Weeks | No |
| TEAM | Logistic Regression | Check fewer individuals | Weeks[6] | No |
| EPIBLASTER | Correlation Contrast | Brute-force | ~1 day | Yes (4 GPUs) |
| SHEsisEPI | Odds-ratio Contrast | Brute-force | ~1 day | Yes (2 GPUs) |
| SIXPAC | LD Contrast | Group Sampling | 8 hours[7] | No |

Table 2-1 Methods comparison.

We list the approximate times reported by five other recent pairwise interaction methods (all perform an exhaustive, genome-wide search) to process a dataset the size of WTCCC bipolar disorder (approximately 2K cases, 3K controls, 450K SNPs, 1 genetic model tested per distal SNP-pair, ≈100 billion pairwise tests). For methods that do not use a GPU cluster, reported times were measured on a comparable desktop computer configuration to the one that SIXPAC was benchmarked on (Intel i7 quad core processor, 2.67Ghz with 8GB RAM). For TEAM, we extrapolated runtime based on performance figures reported on a smaller dataset. Graphical Processing Units

[3] However, we report that the SIXPAC implementation currently takes advantage of multi-core CPU architectures with large reserves of RAM to speed up computation, as well as cluster computing infrastructures to distribute computational burden across multiple nodes - all with little or no effort on the part of the end user. Details are provided on the software webpage.
[4] All times as self-reported by authors of these tools, or extrapolated from performance metrics provided therein.
[5] Operating in the --fast-epistasis mode
[6] 10K SNPs all-pairs test reported in 1000 seconds, scaling linearly with number of SNP-pairs thereon.
[7] Time taken to find all pairs with LD-contrast p<1e-12 with >90% power, multi-threaded mode.

(GPUs) are computing chips which provide around 100X speedup over regular CPUs, and were therefore used by two recent high-performance implementations. Despite not using such specialized hardware, SIXPAC is the only method which can scan a GWAS dataset of this size in a few hours. This is because while most methods effectively need to test each pair to find the few significant combinations, group-sampling allows SIXPAC to drastically prune the search space while simultaneously guaranteeing that all the statistically significant pairs will make it through such a pruning.

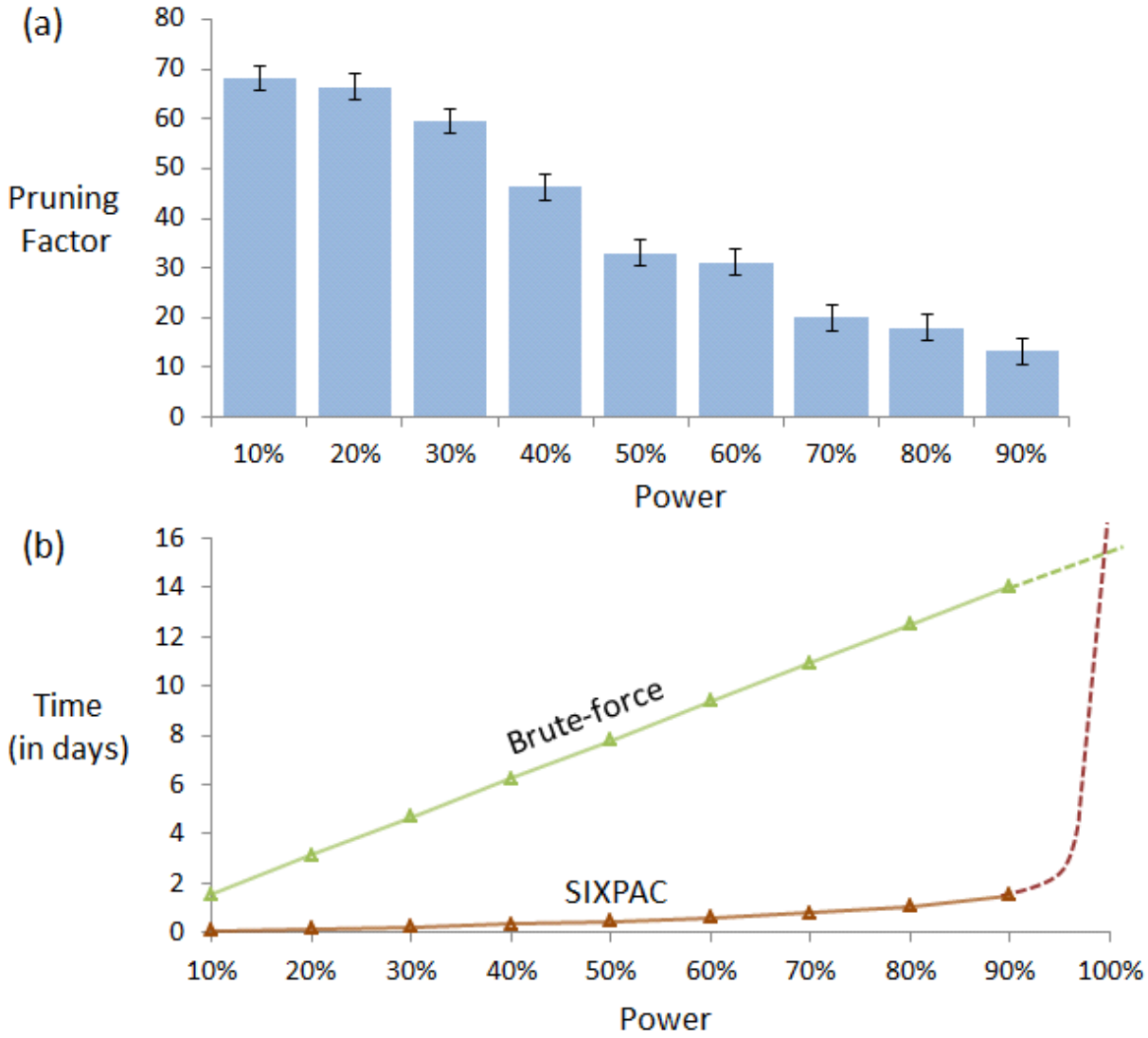



Figure 2-2. Computational Efficiency.

Our implementation of the two-stage PAC-testing framework (SIXPAC, orange line) was benchmarked on the cleaned WTCCC bipolar disorder dataset (approximately 2K cases, 3K controls, 450K SNPs, 4 genetic models tested per distal SNP-pair, 400 billion pairwise tests genome-wide). Part (a) shows the factor reduction in the

universe of SNP-pairs achieved by stage 1, for each power setting. Note that unlike brute-force, this does not mean down-sampling the universe of SNP-pairs, but rather involves reducing the probability of identifying any one of them. For example, a brute-force method would presumably test 40 billion pairs (and ignore the remaining 360 billion) to achieve 10% power on this dataset. However, PAC-testing scans all 400 billion pairs, but simply reduces the probability of finding the significant interactions among them to 10%. This results in shortlisting approximately 68X fewer combinations through stage 1. Part (b) shows the efficiency of our software implementation of this method. We compare the performance of SIXPAC against the time taken by a brute-force approach of applying the LD-contrast test directly to all pairs (green line). All tests were benchmarked on a common desktop computer configuration (Intel i7 quad-core processor, 2.67 GHz with 8GB RAM). The last data-point shows the 90% power benchmarks, followed by dotted lines which illustrate how these estimates may continue as we approach 100% power. SIXPAC, like any randomization algorithm, will require infinite compute time to achieve 100% power, but can approach very close at a small fraction of the brute-force cost. Lastly, we note that these measurements only reflect the performance of our java program rather than what might be feasible with a different implementation of the algorithm.

*Novel Significant Interaction in Bipolar Disorder:* We ran SIXPAC on the BD dataset with >95% power to check whether there exist any significant LD-contrasts between pairs of physically unlinked variables (SNPs >5cM apart). We report the presence of only one statistically significant two-locus contrast (BD cases vs. NBS+58C controls LD-contrast $p < 1.2 \times 10^{-13}$) between SNPs lying within two calcium channel genes : rs10925490 within *RYR2* on *chr1q43*, and rs2041140 and rs2041141 within *CACNA2D4* on *chr12p13.33*. We successfully replicated the signal from this region at Bonferroni significance levels in a different bipolar dataset of Europeans (653 BARD cases, 1034 GRU controls) from the GAIN initiative (Manolio et al. 2007; Smith et al. 2009; also see www.genome.gov/19518664) which were typed on a different platform (Affymetrix 6.0). Deeper investigation revealed that the SNP in *CACNA2D4* is 200Kbp away from *CACNA1C* – a known calcium channel gene whose association to BD was only recently confirmed by combining large GWAS datasets for meta-analyses (Ferreira et al. 2008; Sklar et al. 2008). Functional experiments have also confirmed the role played by genes at this locus in bipolar disorder (Perrier et al. 2011). Although channel ideopathies (and more specifically faults in calcium channels and signaling) have long been

known to play a major role in bipolar disorder, single-locus association methods were underpowered to implicate genes in these pathways without considerably boosting their sample sizes (Craddock 2007; Sklar et al. 2008; Ferreira et al. 2008). Neither gene that we report – either at the known locus or novel locus - was identified as a candidate by the original WTCCC analysis (Craddock 2007) which focused on effects visible to single-locus association.

Specifically, we found that the dominance variable of rs10925490 (one or more minor alleles) was in severe positive linkage disequilibrium with the recessive variables of adjacent SNPs rs2041140 and rs2041141 (two minor alleles each) in BD cases, and slight negative disequilibrium with them in controls, giving an LD-contrast $p = 4.6 \times 10^{-14}$. To verify that this signal was not due to any unaccounted biases, we first confirmed that high LD between the two variables was specific to BD cases only, even when contrasted against samples from all other WTCCC disease phenotypes (6 tests of BD vs. other-disease-cases all show LD-contrast $p < 10^{-9}$). Next, we performed a permutation analysis to characterize the empirical distribution of the LD-contrasts statistic at the theoretical significance level of $p = 4.6 \times 10^{-14}$ (i.e. to check if $p_{corrected} \leq 0.05$). We ran SIXPAC on 100 phenotype permuted versions of the same dataset (i.e. 100 whole-genome, all-pairs scans for interaction) and observed $p \leq 4.6 \times 10^{-14}$ between a pair of SNPs in only 1 such permutation ($p_{corrected} \approx 0.01$).

Finally, we sought to replicate the observed difference in LD at these loci. In the GAIN dataset, we considered all LD-contrasts in an area of 1 SNP immediately upstream and downstream of rs10925490 in the dominant allelic mode, against 1 SNP immediately upstream and downstream of rs2041140 in the recessive allelic mode. In other words, we tested $3 \times 3 = 9$ pairs (around

and including the original interaction), to test if any pair in this area bore an LD-contrast that passed the conservative Bonferroni significance cutoff $\alpha = \frac{0.05}{9} \approx 0.005$. This roughly translates to a region $\leq$ 5Kbp upstream and downstream of each SNP in the original pair. Although there was no appreciable difference in LD between the same SNPs (rs2041140/rs10925490 shows LD-contrast $p > 0.01$), we observed a significant LD-contrast ($p = 4 \times 10^{-5}$) between rs2041140 and rs677730 (the SNP immediately upstream of rs10925490 on the Affymetrix 6.0 platform). To confirm that this observation was not likely by chance, we randomly picked 5000 pairs of physically unlinked (>5cM apart) SNPs genome-wide and tested an equal neighborhood of $3 \times 3$ LD-contrasts around each pair in the GAIN dataset. Only 1 out of 5000 random areas contained a SNP-pair with a more significant LD-contrast ($p_{corrected} = 0.0002$).

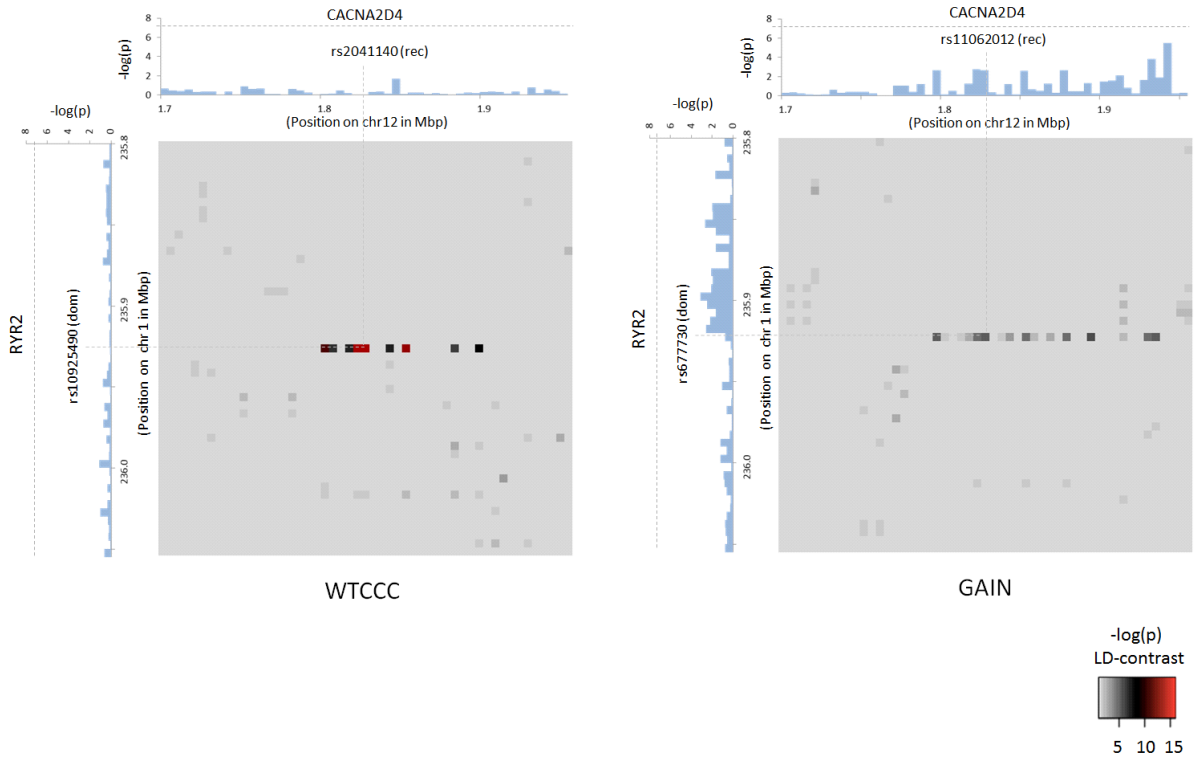



Figure 2-3 Bipolar Disorder Interaction.
In a genome-wide scan of all 400 billion variable-pairs (4 genetic models tested per SNP-pair) in the WTCCC bipolar disorder dataset (Affymetrix 500K), SIXPAC found one significant interaction ($p < 1.2 \times 10^{-13}$) between SNPs >5cM apart that satisfied all our filtering criteria. The SNPs rs10925490 and rs2041140 lie within the *RYR2* gene on *chr1q43* and the *CACNA2D4* gene on *chr12p13.33* respectively. Each figure shows the -log(pvalue) from a standard single-locus association test (allelic model) of the two SNPs as well as 25 SNPs immediately upstream and downstream from each of them, along the X and Y axis. Also shown in the grayscale area is the –log(p) from the pairwise LD-contrast test of all $51 \times 51 = 2601$ variable-pairs. As suggested by the original finding, SNPs around rs10925490 were considered in dominant allelic mode, while SNPs around rs2041140 were in recessive mode. We replicated this signal by similarly testing 2601 dominant-recessive pairs of variables around the very same SNPs in a much smaller bipolar disorder dataset from the GAIN consortium (Affymetrix 6.0). In the replication dataset, we observe several pairs that cross the significance threshold and a strikingly similar visual pattern in the LD-contrast landscape (see main text for a permutation analysis). The top pair (rs677730- rs11062012) in this area is pinpointed with dashed lines (see main text for permutation analysis). Standard single-locus association analysis does not yield any significant result in either dataset, as seen in the marginal Manhattan plots (gray dashed line represents genome-wide significance level).

To get a better picture of the LD-contrast landscape between SNPs in this region, we conducted a wider survey of the area spanning $\pm 25$ SNPs (upstream, downstream and including) both rs2041140 and rs10925490 (i. e. $51 \times 51$ tests). The scan reveals several additional pairs of

SNPs that show differences in LD going in the same direction (strong LD in cases, weak negative LD in controls) – arranged in a strikingly similar pattern in both datasets, presenting strong evidence of an inter-locus effect. The 2 dimensional LD-contrast spectrum for this larger area is presented in Figure 2-3, alongside the Manhattan plots for marginal association at each locus. The top SNP-pair in the area (rs677730,dom × rs11062012,rec) had LD-contrast $p = 1.19 \times 10^{-6}$ in GAIN: a similar phenotype permutation analysis as earlier reveals that only 19 out of the 5000 randomly chosen $51 \times 51$ areas genome-wide contained a more significant pair ($p_{corrected} = 0.0038$). It can also be seen that there is no marginally significant association at these loci in either dataset.

Table 2-2 presents a summary of the results along with the single most significant variable pair in the larger test area for each dataset.

| Dataset | 1q43 (RYR2) | | 12p13 (CACNA2D4) | | LD-cases (Z-score) | LD-controls (Z-score) | Interaction p-value | |
|---|---|---|---|---|---|---|---|---|
| | SNP, mode | p-value (Marginal) | SNP, mode | p-value (Marginal) | | | LD-contrast test | Logistic Regression |
| WTCCC | rs10925490, d | 0.5974 | rs2041140, r | 0.6594 | +7.7 | −2.3 | 4.61e-14 | 1.28e-09 |
| GAIN | rs677730, d | 0.17 | rs11062012, r | 0.05 | +5.1 | −1.2 | 1.19e-06 | 0.0001 |

Table 2-2. Bipolar Disorder Interaction.

The lower table lists the most significant LD-contrast SNP-pair spanning two calcium channel genes RYR2 and CACNA2D4, in both the original (WTCCC) as well as the replication datasets (GAIN). Columns 2 and 3 present the apparent mode of action for this SNP-pair (represented as SNP rsid, allelic mode – dominant d, recessive r), and the p-value for each SNP using single-locus association analysis. Columns 4 and 5 show the LD between these SNPs in

cases and controls (each normalized into a Z-score), which are derived by comparing the expected to the observed co-carriers in cases and controls (see boxes below). These counts are outlined in the tables above, and show a clear enrichment of observed minor allele co-carriers in cases and depletion in controls (against their corresponding null expectations, assuming linkage equilibrium). Column 5 reports the LD-contrast significance (note that the LD-contrast statistic is not a simple difference in Z-scores). Although LD-contrast does not seek or imply statistical epistasis, we can see that the pair is also a nominally significant candidate as per a logistic regression based 1 d.f. test for interaction term, as shown in column 6.

## 2D. Discussion

In this work we introduced a novel method that defuses the computational challenge of a genome × genome interaction scan by using the statistical constraint towards, rather than against our goal. Focusing only on interactions that have a chance of achieving statistically significant association, we developed a rapid filter that does not require the naïve arduous scan of all pairs of variants. To demonstrate its utility, we implemented an established test for interaction which contrasts LD between cases and controls, to demonstrate how an exhaustive genome-wide multi-locus association search is possible while saving an order of magnitude or more in computational resources. Usefully, we are also able to provide performance guarantees and quantify the approximate nature of our output, and our algorithm brings genome-wide three-locus scans into the realm of feasibility.

While the focus of this contribution is computational methodology, we prove applicability in practice to a classical GWAS dataset. Among widely investigated common diseases, bipolar disorder remains one of the most recalcitrant phenotypes to GWAS methodology (Craddock and Sklar 2009), perhaps in part because of the limitations of single locus association analysis. We highlight the power and utility of multi-locus effects in terms of uncovering molecular processes by exposing two calcium channel coding genes as affecting bipolar disorder, supporting recent

discoveries that were only made possible through a significant increase in dataset size. We have replicated this observation in an independent dataset, strongly suggesting a *bona-fide* underlying interaction between members of a gene-family known to be functionally associated with bipolar disorder, making it suitable for further investigation.

Compared to the number of single-locus associations, GWAS of common phenotypes in humans have uncovered very few reproducible gene-gene effects so far. This is partly because interaction analyses for human populations are difficult to design and interpret (Cordell 2002; Phillips 2008). A conventional test for statistical epistasis is expected to only identify loci whose combined effect on phenotype is not explained by the addition of their individual effects, for an appropriately chosen scale. In case-control studies, this typically involve applying a logistic regression to check for significance of the interaction term(s) after accounting for main effects (Wang et al. 2010): which is equivalent to a test for deviation from multiplicative odds (or additive log-odds). However, there are several limitations to this approach – scale of choice (Mani et al. 2008), assumption of a genetic model by which two-loci combine their effects (Hallander and Waldmann 2007), limited models of interaction that can be tested (Li and Reich 1999; Hallgrímsdóttir and Yuster 2008) and limited sensitivity of logistic regression to non-normal residuals, among others. How these factors might cumulatively affect a test for other models of genetic interaction has not yet been decisively established.

Further, true biological interaction between two or more loci may or may not manifest itself as a departure from additivity. Two loci whose main effects appear to combine in an additive manner might also indicate their biological co-involvement (and hence "interaction") underlying the

disease (Wang et al. 2011). In general, two-locus association tests are known to contribute signal independent from what is seen by conventional single locus association tests (Kim et al. 2010; Marchini et al. 2005) and comprehensive multi-locus association strategies may be worth undertaking despite the increased multiple testing burden (Evans et al. 2006a). Indeed, recent work (Zuk et al. 2012) showing that alternate models of biological interaction could confound estimates of heritability have redirected the attention of the genetics community on the potential of interaction studies.

A previous genome-wide scan for statistical epistasis on the same bipolar disorder dataset had reported Bonferroni significant epistasis between rs10124883 and four other SNPs (Hu et al. 2010). As expected, all four pairs approached (but did not clear) Bonferroni significance levels as per the LD-contrast test as well ($p \approx 10^{-12}$) – and could therefore be captured simply by lowering the significance cutoff. This congruence between tests for statistical epistasis and contrast tests has been exploited by others (Plink, EPIBLASTER) and indeed, also holds for the binary LD-contrast test (see tables in 7F. ). But whereas other methods would employ a brute-force testing strategy to identify candidate SNP-pairs, PAC testing will accomplish the same result much quicker by looking at a small fraction of the pairs.

Our findings do suggest that unlike stepwise regression approaches that sequentially attribute residual variance/deviance to each of their components, tests that make fewer assumptions regarding scale may indeed be more powerful at capturing a wider range of interactions. Conversely, a distinct advantage of regression over our LD-contrast test remains its clear interpretation and measurement of effect size: though the difference in LD between cases and

controls is consistent and reproducible across datasets, it does not immediately suggest a clear causal genetic model underlying this signal. We dissected this interaction using the standard logistic regression, $\ln(\frac{P}{1-P}) \sim \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$, where $X_1 = \{0,1\}$ codes for dominance carrier status at rs10925490 while $X_2$ codes for recessive carrier status at rs2041140. The main effects $\beta_1$, $\beta_2$ were observed to be not significant, while the epistasis term $\beta_{12}$ was considerable ($p \approx 10^{-9}$), suggesting deviation from multiplicative odds is one option. We also considered the standard full genotype model (0/1/2 parameterization of predictor variables) with 8 degrees of freedom (Cordell and Clayton 2002) as implemented by INTERSNP (Herold et al. 2009), where the most significant test (Test 6, $p \approx 10^{-9}$) was the one comparing the full model against a model that accounts for just within-SNP additive and dominance effects. In a genome-wide search for interaction using logistic regression, these levels are likely to fall short of significance cutoffs after correcting for hundreds of billions of tests performed: which explains why other methods seeking statistical epistasis on the same BD dataset did not report LD between the *RYR2-CACNA2D4* as a significant finding. A true etiological understanding of this persistent difference in LD may require sequencing at each locus to identify the interacting variants.

*Results in other WTCCC datasets:* Subsequent to the publication of the original paper (Prabhu and Pe'er 2012b), SIXPAC was extended to implement the LOD-contrast statistic (discussed in section 7K. ). Unlike LD-contrast, contrasting log of odds exposes interaction as measured on its most widely adopted scale: logistic regression. When we ran SIXPAC on the 6 other phenotypes collected by the WTCCC (Crohn's disease, Coronary Artery disease, Type I and Type 2 diabetes, Rheumatoid arthritis and Hypertension), not a single epistatic interaction was observed

at $p < 10^{-12}$. In the absence of replication datasets, it is unadvisable to apply the more powerful LD-contrast test which tends to suffer from an inflated type I error rate.

*Limitations and Extensions:* The major contribution of this work is a computational technique to rapidly identify SNP-pairs with large values of a test statistic without performing a brute-force search. While we assessed the issue of power with regards to our randomization algorithm, we left the separate (but equally important) concept of statistical power unaddressed – i.e. the ability of an interaction test to spot a true biological interaction in the dataset. Although contrasting LD, correlation and odds-ratios between cases and controls have all separately been characterized as powerful tests for interaction, each test makes specific model assumptions and is powerful only under its own regime. Consequently, the absence of interaction reported by SIXPAC (or indeed, by any other software) does not imply the absence of interaction itself, but could simply mean lack of statistical power of the test, inadequate number of samples, or simply, incorrect model assumptions. During the course of publishing this method, minor corrections were suggested for a range of contrast statistics to improve their power and decrease type I error rate (Ueki and Cordell 2012). Again, we note that modifications to these tests can be easily adopted into our computational methods – which are agnostic of statistics.

In contrast to the performance gains offered by group-sampling are its two notable weaknesses. First - like any other randomization algorithm - group-sampling can never achieve 100% power (probability of completion), whereas brute-force approaches will. Second, by virtue of limiting itself to binary features, testing for genetic models that incorporate allelic dosage and trend effects using group-sampling does not appear straightforward. Although extending our

computational principles to implement rapid correlation and odds-ratio contrast tests (among others) may be appealing, the loss of statistical power from increasing the number of tests is less easily addressed. Where we currently encode recessive and dominance binary status, each additional test may require a different encoding of features (genotypes, or combinations thereof), thereby adding to the multiple testing burden. Overcoming these limitations appears non-trivial, and increases in sample size will almost certainly play a crucial role in discovering these hidden genetic connections.

Extrapolating from the hardware speedups reported by others (Kam-Thong et al. 2010; Hu et al. 2010) may suggest that a high-performance GPU-enabled implementation of our method might offer a scan of all-pairwise interactions in a few minutes, and all 3-way interactions on the order of a day(s) in large GWAS datasets. But a more immediate concern related to testing 3-way interactions would be the statistical power and semantic interpretation of such a test (conceivably devised on a 2×2×2 binary table). In conclusion, we note that while the transition of association studies from SNP arrays to full ascertainment of variants may have led to analytical emphasis on rarer alleles, it has only increased the impetus to examine the spectrum of multi-locus effects. With so many more variants to consider, the computational limitations will only become more severe, but the solutions reported will be ever more essential.

# Chapter 3. Functional enrichment of SNP-SNP interactions

## 3A. Background

Over the past two decades, genome-wide linkage and association mapping methods have investigated millions of genetic markers (Hindorff et al. 2009; GWAS catalog), predominantly SNPs, and implicated a few thousand of these underlying a wide range of disease phenotypes. Importantly, these associations can help us better understand the genetic architecture of these heritable traits, bringing us closer to treatment and prevention. Although these genome-wide significant findings have helped us make some crucial insights into the biology of these phenotypes, they cumulatively capture only a fraction of the total genetic signal underlying their respective traits. For most common diseases, a large part of the genetic variance remains unexplained, suggesting that most of the loci and mechanisms underlying disease are yet to be discovered (Maher 2008).

In this regard, pathway-based association methods that utilize functional information ascribed to various loci on the genome have been tremendously useful. Instead of looking at SNPs in isolation, these methods typically group SNPs by functional annotation, devise statistics to measure each group's significance from its constituent SNPs, and then test which groups contain a significantly greater number of highly associated SNPs than others. This general approach has developed well-established statistical principles and has been implemented by a wide variety of methods (Holmans et al. 2009; Mootha et al. 2003; Miao-Xin Li et al. 2011; Poirel et al. 2011). By incorporating functional information into the analysis, signals emanating

from SNPs well below genome-wide significance levels can be utilized – and this strategy that has borne high returns.

In this work we present a pipeline that takes this idea a step further. We consider a framework that allows us to test for enrichment in interacting SNPs across a pair of functional groups, as opposed to marginally associated SNPs in a single group. Specifically, we determine all epistatic (non-additive) effects between pairs of SNPs straddling two functional groups (where each SNP in the pair belongs to one group) and ask whether there is any overall inflation in the level of epistasis between these two groups. Mechanistically, we extend single-group enrichment methods in a relatively straightforward manner but use recent, high-speed genome-wide interaction methods that make this experiment computationally feasible.

In this paper, we consider functional groups for human genes defined by the Gene Ontology database (Ashburner et al. 2000). The goal of this work is to find whether there is any evidence of ontology-ontology epistasis underlying common disease. By definition, statistical epistasis between two SNPs suggests that their combined effect deviates from what one would expect if each SNP acted independently of the other (there are several excellent references that discuss types, causes, scales and interpretation of epistasis (Cordell 2002; Cordell 2009; Wang et al. 2010; Moore and Williams 2009)). Under the most widely accepted definition of independence for dichotomous phenotypes, the effect of two genetic features on phenotype status is expected to be additive. In case versus control genome-wide association studies, this typically entails applying a logistic regression to find pairs of SNPs whose combined effect deviates from the sum of their log of odds. Such a departure suggests a level of biological inter-dependence or

connectivity between the two loci and can potentially help us make useful insights into the trait's genetic architecture. By extension then, evidence of increased epistasis between two ontologies (if it indeed exists) can offer a broader view into whether certain functional groups are more closely intertwined than others in the context of a disease's etiology.

A major concern for such an experiment is the massive computational investment required by a genome-wide epistasis scan. Until recently, the burden of even a single genome-wide scan of $\sim 10^{10-12}$ SNP-pairs (assuming a contemporary assay $\sim 10^{5-6}$ SNPs) required in a brute-force approach was justly considered to be quite computationally expensive (Marchini et al. 2005). Furthermore, to establish the significance of an ontology-pair in a robust empirical manner, it is conceivable that we will have to undertake such a scan not just once but several thousand times under a permutation testing scheme. The computational impracticality of such an effort has necessitated several simplifying assumptions by earlier analyses (Emily et al. 2009b; Hannum et al. 2009). Principal among these assumptions is that certain SNPs are more important than others – where "importance" was measured by a high marginal association score, or by whether the SNP lies in the exome, or causes an amino acid change, or some other inferred biological motivation – and such SNPs may therefore be preselected. Subsequently, pairwise interaction scans were restricted test the chosen candidates. Even so, these early semi-exhaustive experiments for functional enrichment of pathway pairs using SNP interaction signatures have been tremendously successful at identifying pathway and protein complex interaction signatures. Although these assumptions made functional enrichment experiments practical, their findings only provide a partial picture of the epistatic landscape. It is well known that epistasis between SNPs may not necessarily imply their marginal effects, and furthermore, selections that are based

on our (presently) incomplete understanding of genome biology can cause statistical biases (Evans et al. 2006; Marchini et al. 2005). In light of this, an unbiased and exhaustive genome-wide analysis merits consideration.

The publication of several of rapid genome-wide SNP-SNP interaction mapping tools (Prabhu and Pe'er 2012; Zhang et al. 2010; Kam-Thong et al. 2010; Hu et al. 2010; Schüpbach et al. 2010) have made functional interaction enrichment an attractive and viable proposition. In this work, we use SIXPAC (Prabhu and Pe'er 2012a), a tool developed by our group, for its speed and robustness. However, the principles of the subsequent analysis pipeline and associated software can be applied in conjunction with any of the other methods. We applied our method to Wellcome Trust Bipolar Disorder dataset (Craddock 2007) to find compelling evidence for pervasive interaction between biological relevant gene groups.

The rest of this paper is structured as follows. First, we describe in detail the functional enrichment pipeline and the methodological modifications required to parse through an input list of SNP-SNP associations (in place of SNP associations). Subsequently, we describe the results of our analysis on Bipolar Disorder dataset. In conclusion, we discuss our findings and suggest avenues of further research.

## 3B. Methods

Functional enrichment methodology is considerably mature, and here we list a few best-practices that have become established in the literature. First, SNPs within a physical window (usually 20Kbp to 50kbp) of a gene's boundaries to that gene (Holmans et al. 2009; Li et al. 2011). Note that this definition allows for a single SNP to be assigned to multiple genes. To derive a gene's association score from its SNPs, various schemes such as choosing the minimum, mean, and average top percentile of p-values of each gene's constituent SNPs have been explored (Lehne et al. 2011). Subsequently, to derive an ontology enrichment score, the most widely used approach is to apply a running sum statistic like Kolmogorov-Smirnov (KS) on a list of gene scores in each ontology (Subramanian et al. 2005; Mootha et al. 2003) . Intuitively, the KS statistic is a measure of the overall inflation in gene scores within the ontology against a background of gene-scores that are not. To account for potential statistical biases like variation in number of tests (SNPs per gene, as well as genes per ontology), normalization procedures are used to adjust the statistics. Further biases that can skew statistical scores are due to LD between SNPs in a gene (resulting in non-independence of tests) or genes in an ontology (due to clustering of genes belonging to certain functional groups) are the most difficult to correct analytically (Holmans et al. 2009). Consequently, the most reliable and statistically robust way to account for such biases is through permutation testing. Deriving an empirical null distribution of the enrichment score for each ontology provides us with the most unbiased estimate of its significance.

We systematically extend the principles gleaned from functional enrichment on single-locus association studies to our multi-locus scheme. For a gene-pair, we assign all those SNP-pairs

where each SNP in the pair is assigned to one gene of the pair. For example, to derive an association score for a pair of genes $gg'$ (with $n$ and $m$ distinct SNPs respectively), we consider the sorted list of p-values $p_{11} < p_{12} \ldots < p_{nm}$ for all SNP-pairs[8]. We then apply the Simes multiple-test correction procedure on this list:

$$\mathrm{p}_{\mathrm{gg}'}^{\mathrm{simes}} = \min_{\mathrm{i,j}} \left( p_{ij} \times \frac{nm}{ij} \right)$$

Equation 3-1

This is less conservative and more powerful multiple test correction than the family wise error rate (FWER) control procedures like the Bonferroni, particularly when the multiple hypotheses tested are not independent of each other (i.e. correlated tests), as in our case [\cite]. Unlike Bonferroni, the Simes procedure gives us a p-value for all the $nm$ null hypotheses tested being true simultaneously, and does not try to control for a false positive rate.

Analogously, to devise an association score for an ontology-pair, we consider all those gene-pairs where one gene belongs to one ontology and the other to the other. For example, to derive an enrichment score $ES_{OO'}$ for a pair of ontologies $OO'$ (with $N$ and $M$ distinct genes respectively), we consider the sorted list of Simes corrected p-values for all gene-pairs $p_{11} < p_{12} \ldots < p_{NM}$ and apply the standard Kolmogorov-Smirnov statistic.

[8] If, on the other hand, genes $g$ and $g'$ contain $c$ SNPs in common as a result of their proximity, then we only consider the $(nm - c^2)$ unique SNP-pairs straddling these genes.

$$ES_{OO'} = \max_{t} \left\{ \sum_{\substack{gg' \in OO' \\ p(gg') < t}} \frac{1}{NM} - \sum_{\substack{gg' \notin OO' \\ p(gg') < t}} \frac{1}{(U - NM)} \right\}$$

Equation 3-2

where $U = \binom{G}{2}$ is the size of the gene-pair universe if we consider $G$ genes, and $t$ is the cutoff at which the KS statistic is maximized. The statistic captures the difference in proportion of gene-pairs with a score below $t$ that belong to the ontology-pair and gene-pairs that do not. We extend our earlier caveat to subsume ontologies which have genes in common by ignoring any gene-pairs where both genes are the same (i.e. we do not consider the case of a gene interacting with itself just because it belongs to both ontologies).

As input, a genome-wide scan for epistasis is used to provide a list of p-values for each SNP-pair. In this work, we employed SIXPAC for its speed and completeness. SIXPAC now provides an implementation of the LOD-contrast statistic (Ueki and Cordell 2012; Wu et al. 2010), which is equivalent to testing for epistasis using the traditional logistic regression method but is computationally much faster. In other words, the p-value for each SNP-pair is the same as the p-value one would obtain by testing the interaction term in a logistic regression. Details are provided in the section 0

From the list of SNP-pair p-values, gene-pair p-values and subsequently ontology-pair KS scores can be calculated. The significance of an ontology-pair KS score is calculated through a permutation scheme – where we shuffle the case-control labels, redo a genome-wide scan for

SNP-pair epistasis, and repeat the procedure to compute a KS score under permutation. This is a computationally intensive task that is made feasible by the speed of SIXPAC. We perform five thousand permutations (genome-wide, exhaustive SNP-pair interaction scans) to derive an empirical assessment of the null distribution of enrichment scores.

At the ontology-pair level as well, FWER corrections like Bonferroni can be too conservative and impractical. For example, assuming just 1000 ontologies of interest, and consequently $\binom{1000}{2}$ ontology-pairs, achieving Bonferroni significance at $\alpha = 0.05$ would require around 10 million permutations. In light of this computational impasse, we describe a different technique to control for experiment wide false discovery rate.

Firstly, just as we accounted for variation in SNPs per gene in the gene-pair p-value, we need to incorporate variation in ontology size into the enrichment score. An elegant technique has been described in (Wang et al. 2007). Given an observed enrichment score $ES_{OO'}$ for a pathway-pair $OO'$, its normalized enrichment score is calculated as

$$\mathrm{NES}_{\mathrm{OO'}} = \frac{\mathrm{ES}_{\mathrm{OO'}} - \mathrm{mean}[\pi(\mathrm{ES}_{\mathrm{OO'}})]}{\mathrm{SD}[\pi(\mathrm{ES}_{\mathrm{OO'}})]}$$

Equation 3-3

where $\pi(ES_{OO'})$ represents the enrichment score for $OO'$ under phenotype permutation $\pi$. Subsequently, we say the proportion of false discoveries $(\alpha)$ to true positives $(1-\beta)$ at any arbitrary enrichment threshold $NES^*$ is controlled at the q-value

$$q_{\mathrm{FDR}} = \frac{\alpha}{(1-\beta)} = \frac{\%\ \text{of permuted } OO'_{\pi} \text{ with } \pi(\mathrm{NES}_{OO'}) \geq \mathrm{NES}^{*}}{\%\ \text{of observed } OO' \text{ with } \mathrm{NES}_{OO'} \geq \mathrm{NES}^{*}}$$

Equation 3-4

Since we are interested in results that are highly likely to be real, we apply a stringent threshold of $q_{\mathrm{FDR}} < 0.01$ on our findings. We have implemented this method in a *java* software package called FUNGI (Functional enrichment of genetic interactions).

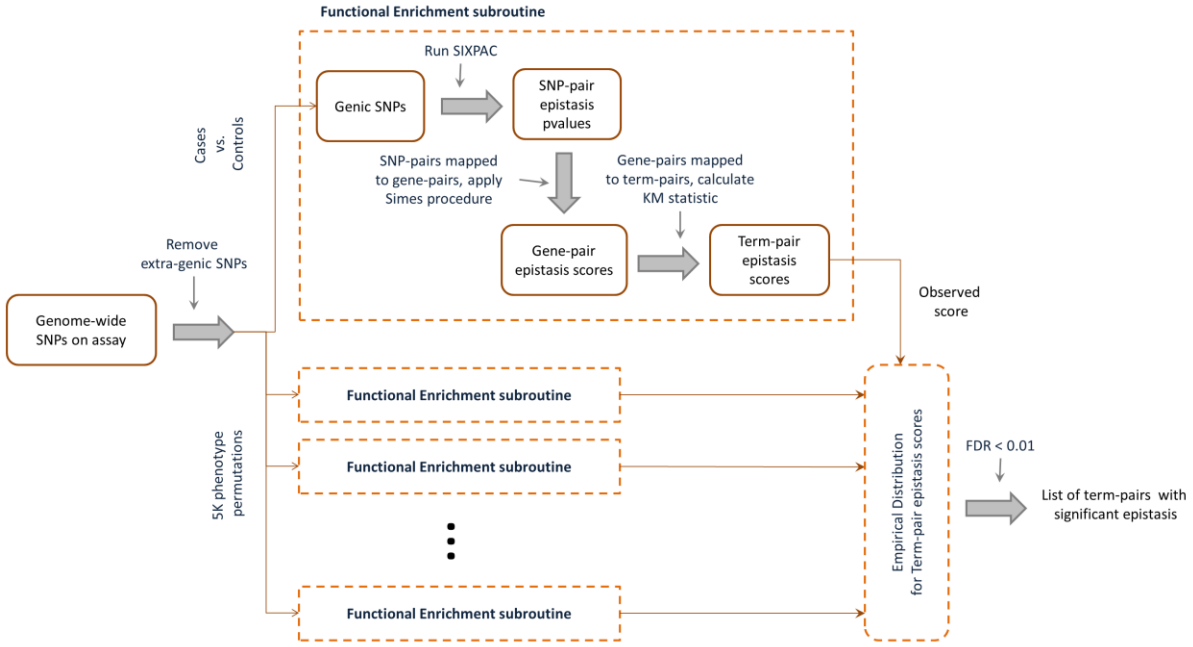



Figure 3-1 Functional enrichment pipeline.

This figure illustrates the flow of logic implemented in FUNGI.

## 3C. Results

We applied our method (pipeline illustrated in Figure 3-1) to the Bipolar Disorder (BD) dataset provided by the Wellcome Trust Case Control Consortium (WTCCC). The data consists of around 2000 cases and 3000 controls (combined from the 1958 birth cohort and National Blood Service), all typed on the Affymetrix 5.0 assay. To alleviate the risks of false positives due to batch effect, population structure and other experimental bias, in addition to prescribed cleaning (Craddock 2007) we further subjected the data to extremely stringent filters (see section 7E. ), leaving us with a cleaned dataset of 370234 SNPs genome-wide in 2881 controls and 1818 cases.

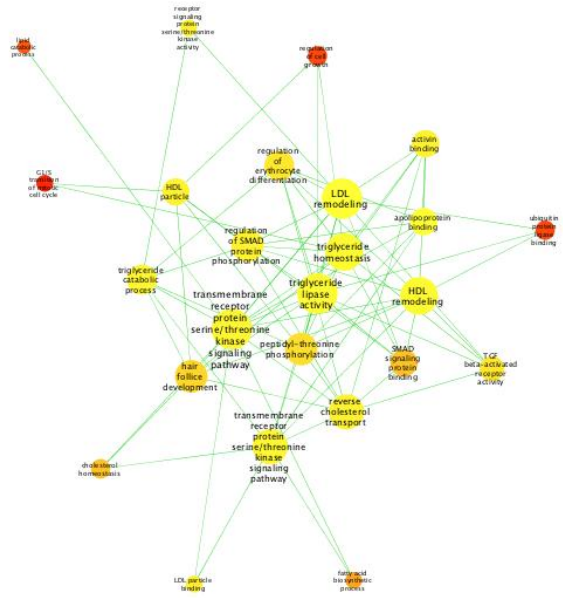
Next, we discarded autosomal SNPs that do not reside within the genes, leaving us with 113626 genic SNPs. These are defined as lying within a 50Kbp window around gene boundaries (Kbp window used to cover proximal promoter and enhancer sites). Gene start and end coordinates were obtained from UCSC genome browser the March 2006/hg18 reference build and are those used by the WTCCC dataset. Functional annotations for *homo sapiens* genes were downloaded from the curated GO database (Ashburner et al. 2000), release 01/21/2013. Ontologies that were too general (>1000 genes) or too specific (<5 genes) in their definition were discarded for their limited utility and also to alleviate the multiple testing burden. Since our experiment hinges on a permutation testing scheme, the effect of any other source of confounding (such as overlapping ontologies, redundant definitions, etc.) on type 1 error is minimized.

As described in pipeline (see methods, figure 1), we first performed an exhaustive SIXPAC search on BD cases vs. controls to find 311 SNP-pairs that demonstrate epistasis at $p < 10^{-7}$. The other genic SNP-pairs that were not reported in this list were all set to $p = 1$ (see 7J. for

details and explanation). We then applied SIXPAC with the same search parameters on 5000 phenotype permuted versions of this dataset to make an equivalent list of epistatic SNP-pairs for each one, giving us a background of $324\ (\pm 32)$ SNP-pairs discovered on average. This suggests that there may be nothing remarkable about the overall number of epistatic interactions between genic SNPs, assuming our data had sufficient statistical power to find all *bona-fide* interactions at $p < 10^{-7}$ or that this estimate holds as sample sizes increase. This is also in accord with the lack of reported and robustly replicated interactions at Bonferroni significance levels on GWAS datasets so far.

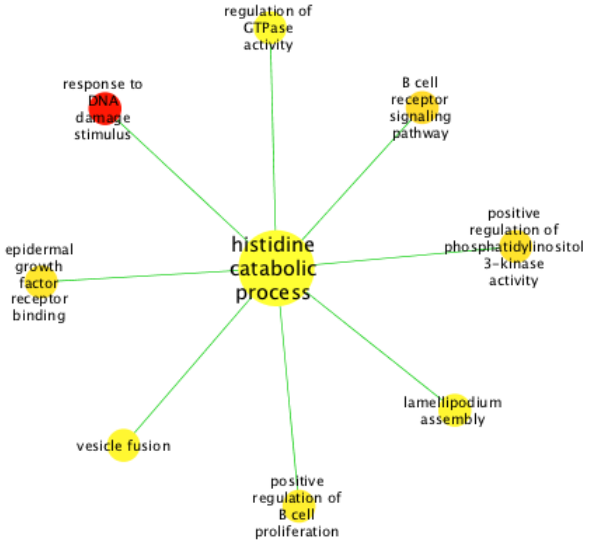
The objective of this work is to group epistatic SNP-pairs according to their annotations and check whether there is anything notable about the case vs. control hits – if not in their number, then perhaps in their location and functional characterization. We ran the FUNGI pipeline on the observed and permuted datasets to find just 157 interacting ontology-pairs (edge) between 110 ontologies (nodes) at a false discovery rate of $q_{FDR} < 0.01$, as illustrated by the graph in Figure 3-2. The complete graph (not shown) contains 17 connected components, the ranging from a large dense cluster (25 nodes, 78 edges) to several small single edge cliques. The largest cluster is also the most intriguing, as it contains the most highly-connected ontologies (gene groups comprising the trans-membrane receptor protein serine/threonine kinase signaling pathway, LDL and HDL remodeling, triglyceride lipase activities and cholesterol transport).

Cluster 1




Cluster 2


Cluster 3




Cluster 4

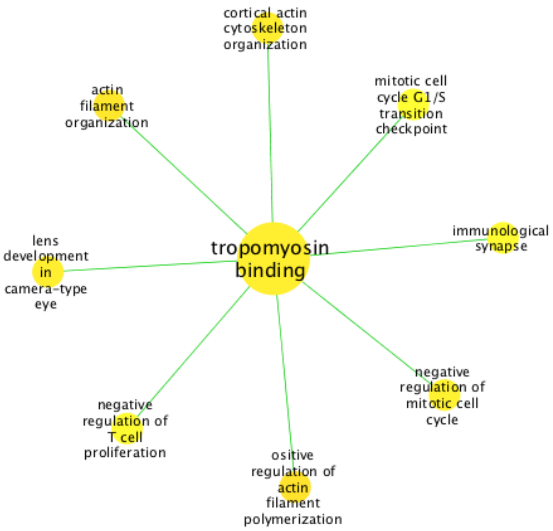


Figure 3-2 Graph of ontology interactions.

Clusters 1 to 4. Full graph not shown.

Interestingly, the serine/threonine pathway is also the target of the most widely prescribed compound for Bipolar Disorder – Lithium (Machado-Vieira et al. 2009; Chalecka-Franaszek and Chuang 1999; Hall et al. 2002; Tsuji et al. 2003). The role of lipid levels in pathophysiology has also been well documented, both in Bipolar Disorder and Schizophrenia (Ghaemi et al. 2000; Chung et al. 2007; Vila-Rodriguez et al. 2011).

The second largest cluster highlights the role of cerebellar Purkinje cells which have been implicated in a variety of psychiatric disorders, including mood disorders like schizophrenia and depression (Tsai et al. 2012; Maloku et al. 2010; Lingärde et al. 2000; Tran et al. 1998). The third and fourth clusters are isomorphic and show a hub-spoke pattern. Cluster 3 outlines the role of histidine catabolism: histidine is an amino acid precursor to histamine, a major neurotransmitter widely studied in psychiatric phenotypes (Ruitenbeek et al. 2009; Nuutinen and Panula 2010; Jin et al. 2009). Cluster 4 elucidates a known inter-connections between genes involved with tropomysin and actin filaments, groups have a known role in muscle disease, but whose role is less understood in this context. Interestingly, new drugs are being developed to antagonize Tropomysin-related kinase B for the treatment of mood disorders (Harrison 2011).

## 3D. Discussion

Over the past few years, advances in computational and statistical methodology in conjunction with increases in sample sizes ($\sim 10^{3-5}$ individuals) and assay resolution ($\sim 10^{6-7}$SNVs) have made it feasible to map genetic interactions in human populations. Despite this, few robust genome-wide significant interactions have been reported in common disease datasets. More pertinently, the findings present insufficient evidence to resolve the hotly debated question of whether interactions are pervasive (Zuk et al. 2012) or anecdotal (Hill et al. 2008) in human genetics, and how much genetic variance they might cumulatively explain.

In this work, we attempted to characterize the functional interaction landscape of a widely studied common disease with a large genetic component and one of the highest heritability estimates among psycho-affective disorders in europeans ($H^2 \approx 0.8$, McGuffin et al. 2003; Edvardsen et al. 2008). To do so, we implemented a well-established but hitherto computationally impractical functional enrichment pipeline. We limited our models to a list epistatic interactions: defined as deviation from additive log of odds in a logistic regression. As a caveat, we note that though the logistic scale is the most widely used (interaction on this scale is commonly referred to as *statistical epistasis*, as per the original Fisherian definition (Fisher 1918)), it is by no means the only one (Cordell 2009). It remains to be seen whether other definitions of interaction will present an equally compelling case for their ubiquity and potential utility.

Our minor result is that the number of such interactions between genic-SNPs when comparing bipolar cases to healthy controls does not appear to be any different than what one would expect by chance. Our major result is that by integrating functional information associated with these interacting loci, a clear picture of interaction between elements of biologically and clinically relevant pathways arises.

To the best of our knowledge, this work describes the first tractable approach to a rigorous, unbiased and genome-wide assessment of ontology-ontology epistasis in human genetic datasets. While previous methods severely restricted their search space in the interest of computational practicality, our work attempts a more exhaustive scan, made possible by the speed of recent interaction methods. In particular, we were able to apply rigorous permutation testing to establish the empirical significance of our results in a computationally feasible manner, rather than relying on analytical methods that often rely on scale-limit approximations and sample homogeneity. By deriving an empirical null distribution for the enrichment statistic, our permutation design implicitly corrects for any subtle biases that might be introduced by confounders like linkage disequilibrium and population stratification. This is all the more important on large GWAS datasets like WTCCC, where data can be collected from multiple centers and represent several sub-populations that are often processed separately.

In conclusion, we note that though our method offers a broader picture of the epistatic landscape underlying a common disease than those before us, our study is currently limited to the genic region of the genome. Recent advances in genome annotation, particularly those made by the ENCODE project (Dunham et al. 2012), offer the intriguing possibility of searching for a wider

and more expressive range of interacting functions. However, the endeavor is not as straightforward as it first seems : a drastic increase in the number of annotations means a concomitant increase in the multiple-testing burden. Given that our experiment was designed to overcome the very issue of low statistical power in genome-wide interaction scans in the first place, this can appear self-defeating in outcome. Careful thought needs to be given to tradeoffs between statistical power and the number and specificity of annotations. Lastly, increasing the coverage of function from genes to the whole genome, would mean $\sim 3X$ increase in number of SNPs, and $\sim 9X$ corresponding increase in the number of pairwise SNP interactions to be tested by each permutation. Considering that we will also have to test a larger number of function pairs, the overall experiment will be even more computationally expensive than our current pipeline.

In the past, several theoretical models had predicted that under the expected realm of small effect sizes, finding epistatic interactions would require tremendous increases in statistical power – presumably by collecting larger and larger cohorts (Evans et al. 2006a; Zuk et al. 2012). Today, our results offer additional evidence that epistasis is both pervasive and is a relevant mechanism of pathogenesis. However, rather than trying to map isolated events of SNP-SNP interaction – an approach that is almost always pathologically underpowered – pursuing alternate strategies to characterize broad patterns of interaction between biologically meaningful genomic elements might be worth considering.

# Chapter 4. Cost-effective DNA sequencing for large cohorts.

*Summary:* Resequencing genomic DNA from pools of individuals is an effective strategy to detect new variants in targeted regions and compare them between cases and controls. There are numerous ways to assign individuals to the pools on which they are to be sequenced. The naïve, disjoint pooling scheme (many individuals to one pool) in predominant use today offers insight into allele frequencies, but does not offer the identity of an allele carrier. We present a framework for overlapping pool design, where each individual sample is resequenced in several pools (many individuals to many pools). Upon discovering a variant, the set of pools where this variant is observed reveals the identity of its carrier. We formalize the mathematical framework for such pool designs and list the requirements from such designs. We specifically address three practical concerns for pooled resequencing designs: (1) false-positives due to errors introduced during amplification and sequencing; (2) false-negatives due to under-sampling particular alleles aggravated by non-uniform coverage; and consequently, (3) ambiguous identification of individual carriers in the presence of errors. We build on theory of error-correcting codes to design pools that overcome these pitfalls. We show that in practical parameters of resequencing studies, our designs guarantee high probability of unambiguous singleton carrier identification while maintaining the features of naïve pools in terms of sensitivity, specificity, and the ability to estimate allele frequencies. We demonstrate the ability of our designs in extracting rare variations using short read data from the 1000 Genomes Pilot 3 project.

## 4A. Background

DNA sequencing is being revolutionized by new technologies, replacing the methods of the past decade. “Second generation” sequencing currently offers several orders of magnitude better throughput at the same cost by massively parallel reading of short ends of genomic fragments (Mardis 2008). This enables addressing new questions in genomics, but poses novel technical challenges. Specifically, it is now feasible to obtain reliable genomic sequence along a considerable fraction of the human genome, from multiple individual samples. Such high-throughput resequencing experiments hold the promise of shifting the paradigm of human variation analysis and are the focus of this study.

Connections between genetic and phenotypic variation have traditionally been studied by determining the genotype of prescribed markers. This cost-effective strategy for large-scale analysis has recently led to multiple successes in detecting trait-associated alleles in humans (Wang 1998; Risch 2000). However, genotyping technologies have two fundamental drawbacks: First, they are limited to a subset of segregating variants that are predetermined and prioritized for typing; second, this subset requires the variant to have been previously discovered in the small number of individuals sequenced to date. Both of these limitations are biased toward typing of common alleles, present in at least 5% of the population. Such alleles have been well characterized by the Human Haplotype Map (The International Hapmap Project, 2003) have been associated with multiple phenotypes. On the other hand, rare alleles are both under-prioritized for association studies, and a large fraction of them remain undiscovered (Reich et al. 2003; Brenner 2007).

Resequencing can fill in the last pieces of the puzzle by allowing us to discover these rare variants and type them. Particularly, regions around loci that have previously been established or suspected for involvement in disease can be resequenced across a large population to seek variation. However, finding rare variation requires the resequencing of hundreds of individuals: something considered infeasible until now. With the arrival of low-cost, high-fidelity, and high-throughput resequencing technology, however, this search is feasible, albeit expensive. At the time of this work, Illumina's Genome Analyzer (Gunderson et al. 2004), ABI's SOLiD sequencer (Fu et. al., 2007), 454 Life Sciences' (Roche) Genome Sequencer FLX (Margulies et al. 2005), to name a few, were the primary technology providers offering throughputs on the order of giga-base pairs in a single run (Elaine R Mardis 2008).

Resequencing is typically done on targeted regions rather than the whole genome, making throughput requirements to sequence an individual much less than what is provided by a single run. A costly option is to utilize one run per individual, but in a study population of hundreds or thousands, such an approach is prohibitively expensive. In such cases, “pooled” sequence runs may be used.

The central idea of pooling is to assay DNA from several individuals together on a single sequence run. Pooled Genotyping has been used to quantify previously identified variations and study allele frequency distributions (Shaw et al. 1998; Ito et al. 2003; Zeng and D Y Lin 2005) in populations. Given an observed number of alleles and an estimate of the number of times an allelic region was sampled in the pool, it is possible to infer the frequency of the allele in the pooled individuals being studied. Pooled resequencing can be used to reach similar ends, with

the added advantage of being able to identify new alleles. At least one recent work has analyzed the efficacy of pooled resequencing for complete sequence reconstruction (Hajirasouliha et al. 2008). The investigators of that work studied the problem of reconstructing multiple disjoint regions of a single genome while minimizing overlap between regions. Our work addresses the problem of identifying rare variations contained within a single region across multiple individuals.

Historically, the primary trade-off of a pooled approach has been the inability to pinpoint the variant carrier from among the individuals sequenced in a pool. Retracing an observed variant back to its carrier required additional sequencing (or genotyping) of all of these individuals, one at a time. Barcoding is an upcoming experimental method that involves ligating a "signature" nucleotide string (~5 bp) to the start of all reads belonging to an individual. These nucleotides serve as the barcode that identifies which individual a given sequenced read came from. If/when established, barcoding technology may essentially offer a more complex assay for a wet-lab solution to the same problem we address through computational means. Other purely computational approaches to pooled sequencing have arisen from the fields of group testing and compress sensing (Ding-Zhu Du and Hwang 1993; Margraf et al. 2011; Erlich et al. 2009). In our work, we focus purely on the application of computer coding theory to this problem.

The rest of this chapter is organized as follows: in the Methods section, we first introduce a generic mathematical model that can be used to represent the pooled resequencing process. We develop figures of merit to evaluate a pool design's robustness to error, and coverage under given budgetary constraints. We then propose two algorithms for pool design: logarithmic signature

designs and error-correcting designs. In the Results section we compare the efficacies of our designs against each other and against current practices using synthetic data as well as real short-read data from the 1000 Genomes Pilot 3 project (www.1000genomes.org), where we quantify the abilities and trade-offs of the designs in the context of various sources of noise.

## 4B. Methods

A resequencing experiment is characterized by the target region and a cohort. We consider a cohort $I = \{i_1, \ldots, i_N\}$ of $N$ diploid individuals. These individuals are to be sequenced for a target region of $L$ base pairs using $R$ pools (or sequence runs) labeled $P = \{P_1, \ldots, P_R\}$. Each pool offers a sequencing throughput of $T$ base pairs mapped to the reference sequence. A key factor in such an experiment is the mean expected coverage of diploid individuals in the cohort. This is the number of reads $\hat{C}$ in which each haploid nucleotide of that individual is expected to be observed, summed over all pools, and averaged over all individuals and sites. Mean expected coverage is given by:

$$\hat{C} = \frac{\text{total sequencing capacity}}{\text{total region to be sequenced}} = \frac{RT}{2NL}$$

Equation 4-1

We introduce notation for a pool design as an $R \times N$ binary matrix, $\mathbf{D}$

$$\mathbf{D}_{p,i} = \begin{cases} 1, & \text{if individual } i \text{ is sequenced on pool } p \\ 0, & \text{otherwise} \end{cases}$$

Equation 4-2

We further define notation for column and row sums of the design matrix: For each pool $p$ we denote the number $n(p) = \sum_i \mathbf{D}_{p,i}$ of individuals in that pool; for each individual $i$ we denote the number $k(i) = \sum_p \mathbf{D}_{p,i}$ of pools with that individual. Whenever $n(p)$ and $k(i)$ are constant, as will be evident from context, we shall omit the parameters $p$ and $i$, respectively.

This setup facilitates a discussion of expected coverage of sites across several parameters. The actual coverage, or number of reads that observe a particular nucleotide $x$ on a single haplotype of individual $i$ in a pool $p$, is a random variable $C_{p,i}^{x}$ with mean $\hat{C}_{p,i}$ across all sites $x \in L$. The distribution of this random variable around its mean may be technology specific. We demonstrated elsewhere (Sarin et al. 2008) that $C_{p,i}^{x}$ for Illumina's short read alignments mirror the Gamma distribution (see Figure 4-1 below).

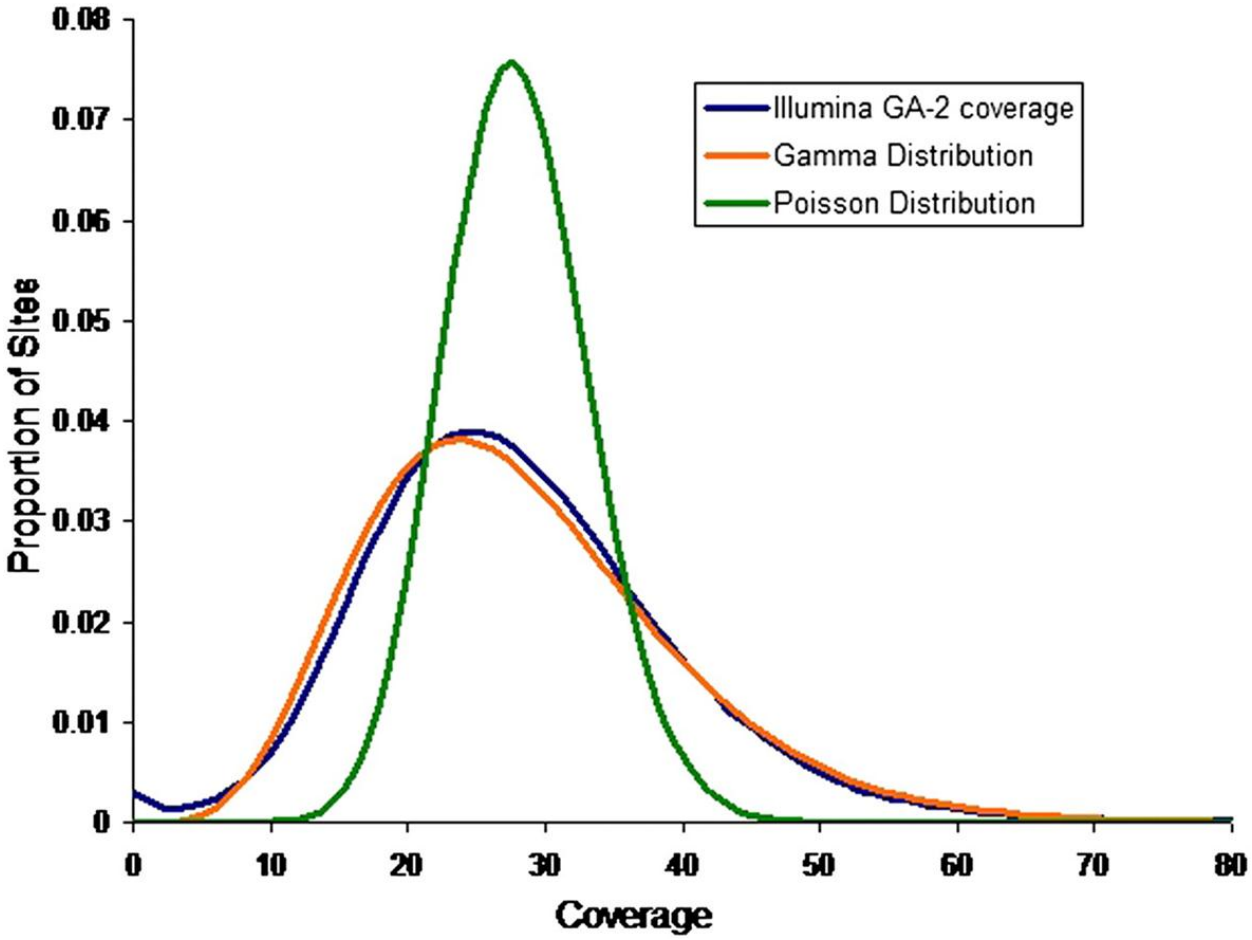



Figure 4-1 Coverage Distribution.

Observed distribution of coverage of Illumina's Genome Analyzer-2 with a mean coverage $\hat{C}_{p,i} = 28X$ over a 4-Mbp region of C. elegans. The distribution best fits a heavy-tailed Gamma distribution Γ(α,β) with shape parameters α = 6.3 and $\beta = \frac{28}{6.3}$. A Poisson distribution is also shown in the figure to compare fits. These results have also been reported by the authors in .(Sarin et al. 2008)

The mean expected coverage of each haplotype of the diploid individual $i$ in a particular pool $p$ is

$$\hat{C}_{p,i} = \frac{\mathbf{D}_{p,i} \times T}{2L \times n(p)}$$

Equation 4-3

Using this, we normalize the binary entries of **D** (presence or absence of a site on pool $p$) to formulate $\mathbf{D}'$ (expected coverage of a site on pool $p$):

$$\mathbf{D}'_{p,i} = \begin{cases} \hat{C}_{p,i} & \text{if individual is sequenced on pool p} \\ 0 & \text{otherwise} \end{cases}$$

Equation 4-4

Summing over a column of $\mathbf{D}'$, we get the expected coverage of a site from an individual accumulated over all pools as $\hat{C}_i = \sum_p \hat{C}_{p,i}$. Likewise, summing across a row gives the expected coverage of a site across all individuals on the pool $\hat{C}_p = \sum_i \hat{C}_{p,i}$. Finally, the expected cumulative coverage of a site across the whole populations in the pooled arrangement, $\hat{C}$ satisfies

$$\hat{C} = \sum\nolimits_i \hat{C}_i = \sum\nolimits_p \hat{C}_p$$

Equation 4-5

Next, we model the sequence of alleles carried by individuals in $I$ as an $N \times L$ matrix $\mathbf{M}$. Each element $\mathbf{M}_{i,x}$ can take on the values $\{0, 1, 2\}$ to register how many copies of the minor allele are present in the diploid genome of $i$ at site $x$. $\mathbf{M}$ is the ground reality: It is not known to us a priori, but rather is what we wish to ascertain. Reconstructing as much of $\mathbf{M}$ as possible is the objective of this work.

Lastly, our expected sequencing results are captured by an $R \times L$ matrix $\mathbf{E}$ of nonnegative integers. The pool design, ground truth, and expected results are linked by the equation

$$\mathbf{D} = \mathbf{M} \times \mathbf{E}$$

Equation 4-6

Each entry $\mathbf{E}_{p.x}$ is a tally of the expected number of minor alleles at site $x$ across all individuals in pool $p$.

**Design properties**

As a first step toward successfully designing overlapping pools, we focus on engineering $\mathbf{D}$, such that it satisfies the following properties of a good design as best as possible.

**Property 1: D retains carrier identity**

This property states that $\mathbf{D}$ must have unique column vectors. Since unique columns serve as unique pool signatures of each individual in the cohort, matching the occurrence pattern of a variant in $\mathbf{E}$ to a column vector in $\mathbf{D}$ suggests that the variant is carried by the individual

associated with that column. Therefore, the design matrix **D** needs to have at least $N$ unique columns.

Pool signatures for all individuals in the cohort may be defined through a function $\mathcal{D}: I \longrightarrow \mathcal{P}(P)$ mapping individuals to sets of pools. Here, $\mathcal{P}(P)$ denotes the power set (set of all subsets) of $P$, while the pool signature of individual $i$ is denoted $\mathcal{D}(i)$. Formally, the pool signature is defined as the set: $\mathcal{D}(i) = \{p \mid \mathbf{D}_{p,i} = 1\}$.

**Property 2: *D* achieves an equitable allocation of sequencing throughput**

All else being equal, there is an equal probability of observing a rare variant carried by any individual in the cohort. It can therefore be shown that any unequal allocation of throughput (coverage) to certain individuals increases the overall probability of missing a variant in the population. While there may often be biological motivation to focus on certain sites (for example, where variation is known or expected to be functional), current technological limitations restrict selective allocation of coverage within the region of interest.

Additionally, the goal of resequencing includes discovery of rare variants, rather than investigating sites that are already known to be polymorphic. We therefore assume no such deliberate preferential coverage, and our aim is to cover all $2L$ sites in each of $N$ diploid individuals as equally as possible using the $P$ pools at our disposal.

One direct way to achieve equitability would be for the throughput of each pool to be divided equally by the individuals sequenced on it, and the number of pools assigned per individual is

constant. In other words, $\forall i, k(i) = k$ and $\forall p, n(p) = n$. Summing coverage assigned to an individual over all of the pools it is sequenced in, we then get $\forall i,\ \hat{C}_i = k \times {}^{T}/_{2n}$.

**Property 3: D is error tolerant**

Modeling the empirical errors introduced into pooled resequencing requires review of the different experimental stages and their associated sources of error. The first step in targeted resequencing experiments is typically pull-down of the target genomic region by standard direct PCR with primers for each amplicon, or by tiling oligonucleotide probes and universal amplification. For pooled resequencing, we assume that the entire pool is amplified in a single reaction. The region of interest is then randomly sheared into short library fragments, which are then single-molecule amplified and end-sequenced. Such sequencing protocols provide millions of single or paired-end reads, which are computationally mapped against the reference genomic sequence. Errors occur during several stages, depending on the sequencing technology. A good pool design should account for errors introduced at each stage and use the redundancy of information in high-throughput sequencing for robustness against such errors.

**Modeling error**

We now quantify errors that occur in the sequencing process within the framework of our model. Equation 4-6 represents an ideal pooling arrangement, where each value $\mathbf{E}_{p,x}$ is an expectation of the number of rare alleles we should observe. In reality, the observed number of alleles at $\mathbf{E}_{p,x}$ is a random variable whose mean is the corresponding expectation. The reason for this randomness is a variety of errors that can cause differences between the expectation and observation. We

address three primary sources of error: read error, error due to under-sampling, and error during amplification (PCR).

**Read error**

Sequence read errors that cause consensus mismatches occur anywhere in the range of one per 50–2000 bases (Smith et al. 2008). These are more likely to occur at non-variant sites, and therefore show up as false-positives, than occur at variant sites, and be observed as false-negatives. Traditionally, sequence assembly methods (Li et al. 2008) have used base-call quality of reads to assess veracity of base calls across multiple reads. However, in the absence of long-established support of the base-call quality used by current technologies, likelihood may still be evaluated by requiring a minimum threshold $t$ of reads that report a variant in order to call it a variant. We assume that this read error occurs at the technology dependent rate of $\mathbf{err}_{read}$ per base pair and is uniform across all pools.

**Under-sampling error**

An individual $i$ is said to be under-sampled at base $x$ if $C_i^x$ is too small to confidently call $x$. Undersampling is intrinsic to all shotgun sequencing, whether pooled or single sample (E S Lander and Waterman 1988). However, pooled experiments are generally carried out due to cost/throughput constraints with coverage distributions more prone to under-sampling than traditional sequencing (Smith et al. 2008). We define a site to be under-sampled if it is read less than $t$ times. The distribution of $C_i^x$ is therefore key for quantifying under-sampling. We propose the density of the Gamma distribution (Sarin et al. 2008) at integer values of coverage as an

approximation of the distribution of practical coverage (see Figure 4-1). The shape parameters for this distribution that we use in our analysis are elaborated in the Appendix.

$$C_i^x \sim \Gamma(\alpha, \beta)$$

Equation 4-7

$$\mathbb{P}(C_i^x = r) = \int_{c=r}^{r+1} c^{\alpha-1} . \frac{\exp(-\hat{C}_i/\beta)}{\beta^c \Gamma(\alpha)} . dc$$

Equation 4-8

The number of under-sampled sites is therefore:

$$\mathbf{err}_{us} = \int_{c=0}^{t} c^{\alpha-1} . \frac{\exp(-C_i/\beta)}{\beta^c \Gamma(\alpha)} . dc$$

Equation 4-9

Equation 4-9 applies to pools just as well as to single-sample sequencing: The parameters that determine under-sampling remain unchanged. Furthermore, it applies to overlapping pools, with the mean coverage across pools $\hat{C}_i$ still determining under-sampling probability, even if the individual coverage per pool is smaller than naïve pooling.

These principles are best demonstrated by an example. Consider a naïve pool design that offers $\hat{C} = 12X$ coverage to each pooled individual and recommends an under-sampling threshold of $t = 3X$. In other words, if we observe a variant less than three times, we do not make a confident

call. The probability of under-sampling a site in this case is relatively small: $\mathbb{P}(C_i < 3) = 0.3\%$. However, to accommodate an overlapping pool design within the same resources as a naïve pool design, we would have to distribute total available throughput $C_i$ over the $k$ pools that an individual $i$ occurs in. If our design leads us to sequence each individual of our experiment in $k = 3$ pools, then our per-pool coverage would be $\hat{C}_{p,i} = \frac{12}{3} = 4$. The chance of under-sampling at the given threshold in a specific pool is now high: $\mathbb{P}\left(C_{p,i} < 3\right) \approx 29\%$, and yet, base calling that is aware of the pool design can distribute the $t$ observations required for calling a new variant across all $k$ pools to formulate a new threshold $t' = \frac{t}{k}$, justifying the same probability of under-sampling.

Note that if a variant fails to be observed at all in a particular pool, the signature of its carrier will not be observed accurately, although the presence of the variant will be detected.

**Amplification error**

These errors occur when PCR chemistries erroneously introduce variants like base substitutions in the replicated DNA (Freeman et al. 1999; Raeymaekers 2000; Huggett et al. 2005). Depending on the enzymes and protocols used, these errors range in frequency from traditional specifications of $\mathbf{err}_{pcr} = 10^{-4}$ errors per base pair to negligible magnitude for high-fidelity chemistries (of the order of $\mathbf{err}_{pcr} = 10^{-6}$ errors per base pair). In pooled resequencing, PCR is most economically pooled as well, and PCR errors may affect multiple reads in that pool. In principle, overlapping pools, each involving separate amplification, are more robust to PCR error, as the error would be introduced to only a small fraction of the independent pools in which

a particular individual participates. Empirically, we observe that practical PCR error rates are negligible enough to be ignored compared with other sources of errors.

**Logarithmic signatures**

The binary representation of numbers $\{i, \dots, N\}$ uses bit-words of size $\log_2 N$. One potential design is to use an encoding function $\mathcal{D}_1 : I \longrightarrow \mathcal{P}(P)$, where each signature $\bar{\sigma}$ is one of $N$ unique bitwords. For example, a small study cohort of 16 individuals would require $\log_2 16 = 4$ pools to generate unique signatures as shown by the first four rows of Figure 4-2. The encoding clearly maintains carrier identity: A variant noticed only in pools {1, 4} points to individual 11, whereas a variant observed on {1, 3, 4} is carried by individual 12, and so on. However, not every individual is sequenced on the same number of pools through this scheme (individual 1 is not on any of the first three pools for that matter).

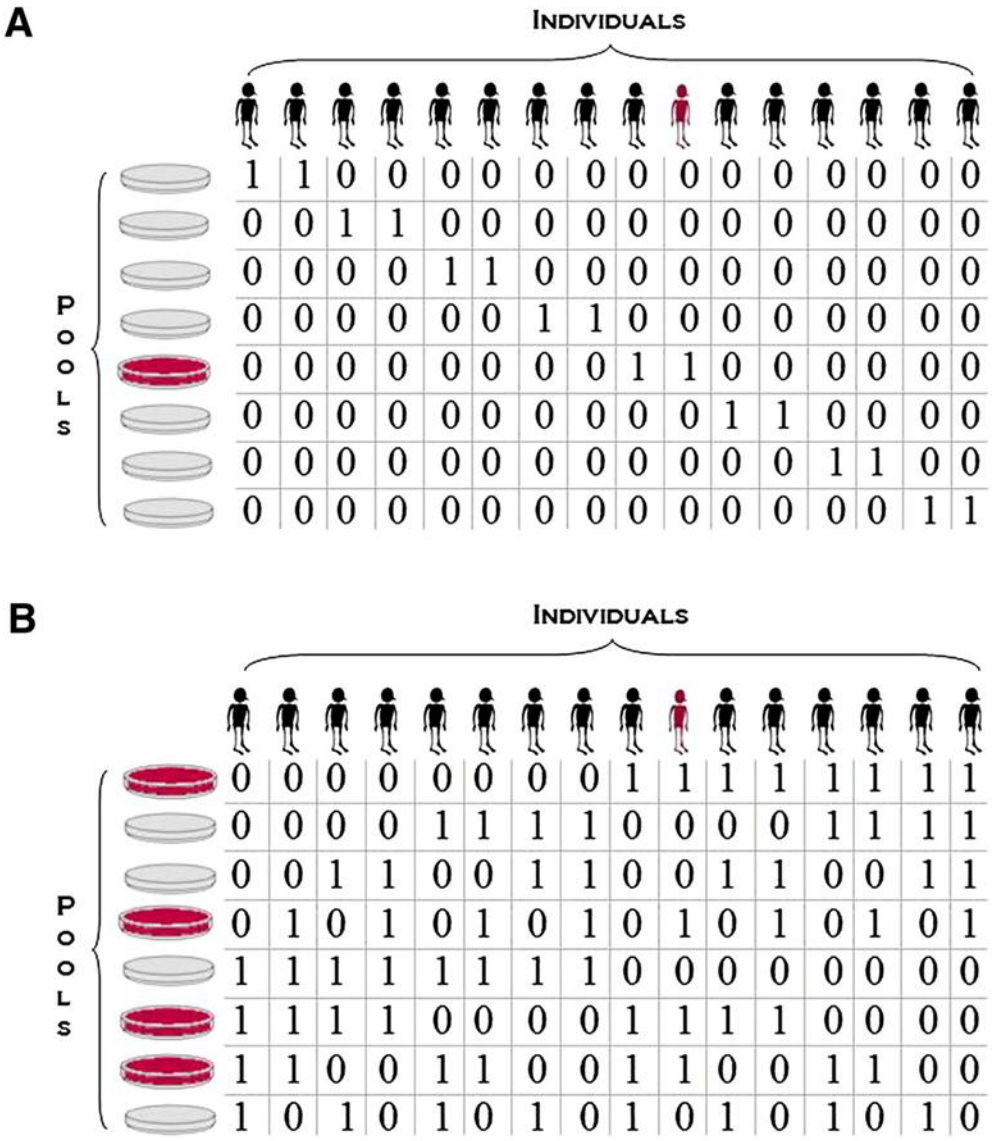



Figure 4-2 Resequencing with naïve and log pool designs.

(A) A total of 16 individuals are divided into groups of two and pooled; (B) 16 distinct pool signatures are created using just eight pools. In both cases, the pools on which the variant appears and the variant carrier are marked in crimson.

We revised the design to satisfy the equitability property by appending the 1's complement of each word to itself, represented by the last four rows of Figure 4-2. The resulting signatures require $2 \times \log_2 N$ pools. Observe that each individual in this example is now pooled on exactly

four out of eight pools. The ratio of the number of individuals sequenced to the number of pools utilized is given by the code efficiency:

$$\text{Code Efficiency} = \frac{N}{2\log_2 N}$$

Equation 4-10

which grows with $N$. This schema is extendible to $\theta$-ary encodings as follows. Each of the $N$ individuals is uniquely indexed base- $\theta$ by $\log_\theta N$ numerals in the range $\{0, \dots, \theta - 1\}$. A base $\theta$ numeral is then mapped to its unique binary signature given by its corresponding vector from the standard basis of order $\theta$. In other words, first numeral 0 to first $\theta$-bit basis vector 0…0001, numeral 1 to vector 0…010, numeral 2 to vector 0…100, and so on. A total of $\theta\log_\theta N$ pools are required to construct the design. For the general case:

$$\text{Code Efficiency} = \frac{N}{\theta\log_\theta N}$$

Equation 4-11

The continuous version of this expression maximizes at $\theta = e$, and for natural $\theta = 3$ (ternary encoding). We call this family of encodings "logarithmic signature designs," because the design uses the order of a logarithmic number of pools in the size of the study cohort.

Regardless of under-sampling, determining allele frequency is no more difficult than with a naïve pool design. The total number of copies of a site $x$ sequenced across all pools is $\hat{C}$

(Equation 4-5). If $m$ minor alleles are observed cumulatively, then assuming equitable coverage (Property 2), we deduce the maximum likelihood estimate of the allele frequency $f$ as $\hat{f} = \frac{m}{2\hat{C}}$.

More often than not, *N* may not be a perfect power of any integral value $\theta$. In such a case, we may not use the entire spectrum of $\theta$-ary signatures (e.g., in a study of only 14 out of 16 individuals in Figure 4-2). The result is that some pools may sequence fewer individuals than others (i.e., $n(p)$ is not constant), violating the equitable coverage dictum. However, by the nature of the design, this variation of $n(p)$ across pools is restricted to 1. In such a case, allele frequency calculations are normalized as $f = \frac{\sum_{p=0}^{R} m_p}{\sum_{p=0}^{R} n(p)\hat{C}_{p,i}}$, where $\sum_p m_p = m$.

**Error-correcting signatures**

While economical, logarithmic signatures fail to satisfy Property 3. They are prone to ambiguous carrier identity in the presence of false-negative variant calls. For example, in the design illustrated in Figure 4-2, individual 1 is sequenced in pools {5, 6, 7, 8}. Suppose the variant call is a false-negative in pool 8 due to under-sampling, but is observed on all others. The resulting "incomplete" signature {5, 6, 7} is not sufficient to unambiguously identify the carrier as individual 1. In fact, it is equally likely that individual 2 under-sampled in pool 4 elicited such an observation. In the general case, a false-negative call in a $\theta$-ary signature ambiguates precisely $\theta$ individuals as potential carriers.

We now develop error-correcting designs that are able to unambiguously identify the variant carrier, even in the presence of false-negatives. Borrowing from results in coding theory (MacWilliams and Sloane 2006), we formulate a one-to-one mapping $\mathcal{D}_2: I \longrightarrow \mathcal{P}(P)$, that retain

identity in the face of false-positives. Consider an individual $I$ with a pool signature $\bar{\sigma}_i$. Intuitively, under-sampling error can be identified and corrected if the incomplete signatures resulting from the loss of "1" bits in $\bar{\sigma}_i$ all continue to point to the same individual $i$. Rather than associating an individual with a single signature, such a design reserves an entire set of signatures within an "error-space" of $\bar{\sigma}_i$ to individual $i$. This error space is the set of all the signatures $\{\bar{\sigma}'\}$ generatable by converting up to some $\epsilon < k(i)$ number of "1"s to "0"s in $\bar{\sigma}_i$. The larger the $\epsilon$, the more signatures reserved per individual, and consequently, the fewer individuals we can multiplex into the pool design. This is the trade-off between efficiency and error correction. With a fixed number of pools at its disposal, an error-correcting design has to maximize the number individuals it can identify while maintaining a disjoint error space.

An estimate of the expected coverage of each site $C_{p,i}^{x}$ also allows us to calculate the expected number of pools on which a site $x$ of $i$ may be under-sampled: $e = k \times \mathbb{P}[C_{p,i}^{c} < t']$. We may then choose an error-correcting scheme that can handle up to some $e$ false-negatives by setting the parameter $\epsilon > e$. By definition then, a variant observed in as few as $q = k - \epsilon$ out of some $k$ pools is sufficient to identify the carrier individual.

A fixed-length block code assigns each individual in a set some $I = \{i_1, \ldots, i_N\}$ to code words such that each code word is of the same length (but not necessarily the same Hamming weight). Logarithmic signatures are a type of fixed-length block code without error-correction ability. There is an extensive theory regarding such codes that do offer error correction. Extended binary Golay codes (EBGC) are such a type of error-correcting block code. Formally, the EBGC consists of a 12-dimensional subspace of the space $\Re = F_2^{24}$ over the binary field $F_2 = \{0,1\}$,

such that any two elements in $\mathfrak{R}$ differ in at least eight co-ordinates. The code words of $\mathfrak{R}$ have Hamming weight 0, 8, 12, 16, or 24. To satisfy Property 2, we only use those code words of Hamming weight 8 (i.e., every pool signature assigns its individual to exactly eight pools). These code words of weight 8 are elements of the $S(5, 8, 24)$ Steiner system. The error space of these Hamming weight 8 code words is a hyper-sphere of radius $\epsilon = 3$. In other words, all signatures generated by up to three false-negatives of $\bar{\sigma}_i$ are reserved for the individual $i$.

EBGC has 759 code words of Hamming weight 8. We therefore repeat this coding separately for $[N/759]$ subsets of the individuals. We note that similar to logarithmic signatures, equitable coverage (Property 2) holds for specific values of $N$, which in this case its values are divisible by 759. For other values, coverage is only approximately equitable, as different pools may accommodate different numbers of individuals.

## 4C. Results

We assessed the performance of our designs by simulating pools of short read data. We downloaded short read sequences from the 1000 Genomes Pilot 3 project (www.1000genomes.org) that were available on the Short Read Archive (SRA) in January, 2009. The 1000 Genomes Pilot 3 project states that it is a targeted sequencing of the coding region of ~1000 genes, while the SRA annotates it as sequence from 1000 to 2000 gene regions and conserved elements (5 KB average length), giving an expected total of 5 Mbp sequence. Illumina runs from 12 individuals, sequenced using single-end, 51-bp read-length libraries, were selected. We created a 123.4-Mbp region of interest from the Human Genome, as outlined in the Supplemental material. The individuals show between 4.2 and 5.3 Mb of mapped sequence with ≥3× coverage, with the notable exception of one individual. From the coverage profile, we verified that the exception was due to poor fidelity/low scoring reads for that run, possibly due to experimental error. Merging the coverage of all individuals, we identified 6.41 million unique sites of high significant coverage.

We constructed two simulated pool designs of 12 individuals on eight sequencing lanes by mixing reads from multiple individuals as detailed in the Supplemental material. Reads for individuals and pools were then independently aligned against the same 123.4-Mbp reference using MAQ (Li et al. 2008), and SNPs were called on the alignment. Since available algorithms call alleles under the assumption that they are looking at reads from a single individual (allele frequency 0, 1, or 2), we built our own SNP-calling algorithm for pooled data (refer Supplemental Methods and Analysis). A combined total of 13,022 single nucleotide variants

were detected across the 123.4-Mbp region in the identity design (i.e., combining independent calls made on each of 12 data sets), of which 10,668 were detected by log pools and 10,868 by ECC Pools. Both designs demonstrate a high-fidelity allele frequency prediction, as evidenced by data outlined in the Supplemental material.

Based on the pool signature of each detected variant, we associated a distribution over possible carrier individuals. Out of a total of 8618 singletons and doubletons, log pools detected 6270 of these variants, while ECC pools detected 6478 of these variants (refer to the table in Supplement on Allele Detection). In truth, we ascertained (using the 12 data sets) that 5332 of the variants detected by log pooling had a single carrier (either homozygous causing singleton or heterozygous causing doubleton), while 5539 of the variants detected by ECC pools had a single carrier individual.

At each of these sites, our algorithm uses the variants pool signature to output a set of equally likely candidate individuals (uniform distribution) to be the variant carriers. Log pools associated 4798 variants with a candidate carrier distribution, while being unable to assign the rest. Likewise, ECC pools assigned 5060 variants with a distribution. In some cases, the call is ambiguous (multiple individuals are given a uniform probability of being potential carriers), while in other cases, the design identifies a single variant carrier.

Of these calls, 3130 distributions in log design captured the correct individual as one of the prospective carriers, while 2907 distributions in ECC design captured the same. Some variants strongly identified single individuals as their carriers instead of offering a distribution over

multiple prospective individuals. The degree of correctness of these calls show a strong correlation to what coverage the site enjoyed on the carrier individual's data set (and, consequently, on the pools in which the individual was sequenced). The results confirm our hypothesis that error-correcting designs enjoy a considerable advantage in terms of numbers of correct calls. In the absence of a suitable paired-end data set, we were unable to assess the ability of our designs to characterize structural variants like indels, copy-number changes, and transposons.

We also assessed the performance of our pool designs on synthetic data. In particular, the logarithmic and error-correcting pool designs were compared against a no-pooling strategy (one individual per pool), and the barcoding strategy. Naïve pooling does not claim to establish carrier identity in the first place, and therefore is irrelevant as a benchmark for these results.

We ran our simulations to identify rare mutations on 500 human individuals, each harboring a targeted region of interest whose size we varied from 300 Kbp to 3 Mbp. We pooled these individuals over 24 sequence runs, mirroring eight lanes on three Illumina GA-2 machines. Each sequence run was given a throughput of 0.5-Gbp mapped sequence, resulting in a total throughput offering of 12 Gbp. This translated to a realistic expected per-haplotype coverage range of 40× to 4× per individual for the corresponding pool sizes.

Recent literature (Levy et al. 2007) suggests that ≈518 K high-confidence variations were found in a newly sequenced genome, which were undocumented in dbSNP, giving a genome-wide approximate new variation incidence rate of 1 in 6.5 Kbp. Assuming most of these variants occur

at a 1% allele frequency in the general populace, we approximate the likelihood of a singleton in a 500 individual (1000 chromosome) cohort to be one in 65 Kb. We randomly inserted mutations at this rate to the data set, and further subjected it to PCR and read error. From the resulting noisy observations $E_{obs}$, we predicted a reality matrix $M_{predict}$, which we then compared against the ground truth.

The no-pooling scheme was used to sequence 24 individuals chosen at random from the 500 individual cohorts. The EBGC scheme uses 24 pools to generate up to 759 code words as discussed earlier. We used the first 500 of these in lexicographic order. We used a value of $\theta = 8$ for logarithmic designs; consequently, also giving us a total of $8 \log_2 8 = 24$ pools. Barcoding also used 24 pools, albeit, effectively simulating 500 distinct pools from their cumulative throughput. Each individual was offered a pool of $T/N = 1$ Mbp.

## 4D. Discussion

In this study we tackle the design of resequencing pools, a very current challenge for large-scale analysis of genetic variation. To the best of our knowledge, this is the first attempt to develop a framework for the design of such pools. We were able to represent real experimental error (as observed on Illumina Genome Analyzer-2 runs) within this framework. We introduced a few properties that represent quantitative figures of merit by which any pooling scheme may be judged. Finally, we presented two original design schemes: logarithmic design and error-correcting design. Each scheme demonstrated a unique set of advantages and disadvantages, but both held much promise compared with a naïve pooling strategy. A comparison between our two

designs themselves reveals that they are both valid approaches, each suited for a varying set of requirements. In fact, logarithmic signatures are appropriate when under-sampling is a negligible consideration. If there is a relatively high-per-pool throughput available vis-à-vis the amount of DNA to be sequenced, a scenario that marginalizes considerations of false-negatives, logarithmic designs offer the most promise to find rare variation. Error-correcting designs are best suited for more trying experimental conditions, when large population studies must be done within minimal resources. These designs can effectively identify variant carriers in spite of noisy signals, but concomitantly run the risk of assigning lesser sequencing throughput to other carrier individuals in the cohort.

Our results indicated that both error-correcting designs and logarithmic designs detect most of the variation in the cohort, with fidelity and ability both dropping as a function of coverage. Our algorithms currently do not attempt to determine carrier identity of more common variations that might have higher incidence (doubletons, tripletons, and polytons), and will be the subject of future work.

In conclusion, our proposed framework motivates both analytical and experimental downstream studies. Analytically, this study focused at identification of rare mutation carriers. Information content in the pooled sequences may facilitate such recovery, particularly if samples are known to be related and carry variants identical by descent at the resequenced locus, as demonstrated for genotype pools (Beckman et al. 2006). Our computational contribution is particularly useful for characterizing human variation by enabling pooled resequencing studies to be conducted with overlapping pools.

# Chapter 5. Advice to graduate students in computational genetics

Despite the giant strides made over the last two decades, much of the science to extract "meaning" from genetics remains to be done. An important genetic feature outside the scope of this thesis has been the impact of non-single-nucleotide variation: particularly structural variants like inversions, copy number alterations, deletions and duplications. Unlike SNPs, the technology to accurately measure these features remained relatively underdeveloped during the tenure of my PhD, making them harder to study (and with a high risk of false positive results from noisy data). This is not the case anymore. Genetics aside, in my opinion the most interesting heritable features will be both non-environmental and non-genetic. This realm of *epigenetic* features will undoubtedly play a huge role in explaining phenotypic variation – either independently, or in conjunction with genetics. Another important point that is never sufficiently repeated is that heritability, by definition, is a measure of a population and not of an individual. Consequently, an imperfect understanding of population membership can result in inflated estimates of population variance, and hence unattainable heritability components. Methods to accurately attribute population membership and estimate phenotype variation will play an important role in improving predictive power. If you are going to make your PhD research about explaining heritability (i.e. building better predictive models), then my advice to you would be to internalize the subject material well (Re: textbooks $\geq$ review papers) before plunging head-first into method development with an imperfect understanding of the domain – a mistake that I made and would not care to revisit. Lastly, methods to resolve naïve phenotype classifications due to unavoidable gaps in our understanding of biology also hold great potential. Besides these research areas, consider the following while looking for a thesis topic: (a) build methods to

characterize genome-wide trends, as opposed to methods looking for anecdotal results, and (b) new datasets that use the latest technologies can often provide the first insight into an untested feature (i.e. low hanging fruit) – but be wary of the price paid in terms of the inaccuracy and unreliability of cutting-edge assays. The race to sequence the genome was the zeitgeist of my generation of graduate students: the race to interpret it is yours. Carpe diem.

# Chapter 6. References


Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**: 25–9. http://dx.doi.org/10.1038/75556 (Accessed January 29, 2013).

Beckman KB, Abel KJ, Braun A, and Halperin E. 2006. Using DNA pools for genotyping trios. *Nucleic acids research* **34**: e129. http://nar.oxfordjournals.org/content/34/19/e129.abstract?ijkey=3ab774386fcb739605ab2653f7df97c8d11f009f&keytype2=tf_ipsecsha (Accessed February 6, 2013).

Brenner SE. 2007. Common sense for our genomes. *Nature* **449**: 783–4. http://dx.doi.org/10.1038/449783a (Accessed November 30, 2012).

Cann HM. 2002. A Human Genome Diversity Cell Line Panel. *Science* **296**: 261b–262. http://www.sciencemag.org/content/296/5566/261.2.short (Accessed February 2, 2013).

Chalecka-Franaszek E, and Chuang DM. 1999. Lithium activates the serine/threonine kinase Akt-1 and suppresses glutamate-induced inhibition of Akt-1 activity in neurons. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 8745–50. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=17587&tool=pmcentrez&rendertype=abstract (Accessed February 7, 2013).

Chung K-H, Tsai S-Y, and Lee H-C. 2007. Mood symptoms and serum lipids in acute phase of bipolar disorder in Taiwan. *Psychiatry and clinical neurosciences* **61**: 428–33. http://www.ncbi.nlm.nih.gov/pubmed/17610669 (Accessed February 7, 2013).

Cordell H J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11**: 2463–8. http://www.ncbi.nlm.nih.gov/pubmed/12351582.

Cordell Heather J. 2009. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**: 392–404. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2872761&tool=pmcentrez&rendertype=abstract.

Cordell Heather J, and Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *The American Journal of Human Genetics* **70**: 124–141.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=384883&tool=pmcentrez&rendertype=abstract.

Craddock N. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661–678. http://www.ncbi.nlm.nih.gov/pubmed/20360734.

Craddock N, and Sklar Pamela. 2009. Genetics of bipolar disorder: successful start to a long journey. *Trends in Genetics* **25**: 99–105. http://www.ncbi.nlm.nih.gov/pubmed/19144440.

Culverhouse R, Klein T, and Shannon W. 2004. Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiology* **27**: 141–152. http://www.ncbi.nlm.nih.gov/pubmed/15305330.

Culverhouse R, Suarez BK, Lin J, and Reich T. 2002. A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics* **70**: 461–471. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=384920&tool=pmcentrez&rendertype=abstract.

Du D-Z, and Hwang FK. 1993. *Combinatorial Group Testing and Applications (Series on Applied Mathematics)*. World Scientific Pub Co Inc http://www.amazon.com/Combinatorial-Testing-Applications-Applied-Mathematics/dp/9810212933 (Accessed February 25, 2013).

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74. http://dx.doi.org/10.1038/nature11247 (Accessed January 28, 2013).

Edvardsen J, Torgersen S, Røysamb E, Lygren S, Skre I, Onstad S, and Oien PA. 2008. Heritability of bipolar spectrum disorders. Unity or heterogeneity? *Journal of affective disorders* **106**: 229–40. http://dx.doi.org/10.1016/j.jad.2007.07.001 (Accessed February 6, 2013).

Emily M, Mailund T, Hein J, Schauser L, and Schierup MH. 2009a. Using biological networks to search for interacting loci in genome-wide association studies. *European journal of human genetics EJHG* **17**: 1231–1240. http://www.ncbi.nlm.nih.gov/pubmed/19277065.

Emily M, Mailund T, Hein J, Schauser L, and Schierup MH. 2009b. Using biological networks to search for interacting loci in genome-wide association studies. *European journal of human genetics : EJHG* **17**: 1231–40. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2986645&tool=pmcentrez&rendertype=abstract (Accessed March 19, 2012).

Erlich Y, Chang K, Gordon A, Ronen R, Navon O, Rooks M, and Hannon GJ. 2009. DNA Sudoku--harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome research* **19**: 1243–53. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2704425&tool=pmcentrez&rendertype=abstract (Accessed February 10, 2013).

Evans DM, Marchini Jonathan, Morris AP, and Cardon Lon R. 2006a. Two-Stage Two-Locus Models in Genome-Wide Association ed. T. MacKay. *PLoS Genetics* **2**: 9. http://www.ncbi.nlm.nih.gov/pubmed/17002500.

Evans DM, Marchini Jonathan, Morris AP, and Cardon Lon R. 2006b. Two-stage two-locus models in genome-wide association. ed. T. MacKay. *PLoS genetics* **2**: e157. http://dx.plos.org/10.1371/journal.pgen.0020157 (Accessed April 20, 2012).

Ferreira Manuel A R, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, Fan Jinbo, Kirov G, Perlis Roy H, Green EK, et al. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nature Genetics* **40**: 1056–1058. http://eprints.bournemouth.ac.uk/14295/.

Fisher RA. 1918. On the correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399–433.

Freeman WM, Walker SJ, and Vrana KE. 1999. Quantitative RT-PCR: pitfalls and potential. *BioTechniques* **26**: 112–22, 124–5. http://www.ncbi.nlm.nih.gov/pubmed/9894600 (Accessed February 5, 2013).

Ghaemi SN, Shields GS, Hegarty JD, and Goodwin FK. 2000. Cholesterol levels in mood disorders: high or low? *Bipolar disorders* **2**: 60–4. http://www.ncbi.nlm.nih.gov/pubmed/11254022 (Accessed February 7, 2013).

Greene CS, Sinnott-Armstrong NA, Himmelstein DS, Park PJ, Moore Jason H, and Harris BT. 2010. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics* **26**: 694–695. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2828117&tool=pmcentrez&rendertype=abstract.

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, et al. 2004. Decoding randomly ordered DNA arrays. *Genome research* **14**: 870–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=479114&tool=pmcentrez&rendertype=abstract (Accessed November 16, 2012).

GWAS catalog. Catalog of Published Genome-Wide Association Studies. http://www.genome.gov/gwastudies/.

Hajirasouliha I, Hormozdiari F, Sahinalp SC, and Birol I. 2008. Optimal pooling for genome re-sequencing with ultra-high-throughput short-read technologies. *Bioinformatics (Oxford, England)* **24**: i32–40. http://bioinformatics.oxfordjournals.org/content/24/13/i32.full (Accessed November 15, 2012).

Hall AC, Brennan A, Goold RG, Cleverley K, Lucas FR, Gordon-Weeks PR, and Salinas PC. 2002. Valproate regulates GSK-3-mediated axonal remodeling and synapsin I clustering in developing neurons. *Molecular and cellular neurosciences* **20**: 257–70. http://www.ncbi.nlm.nih.gov/pubmed/12093158 (Accessed February 7, 2013).

Hallander J, and Waldmann P. 2007. The effect of non-additive genetic interactions on selection in multi-locus genetic models. *Heredity* **98**: 349–359. http://www.ncbi.nlm.nih.gov/pubmed/17327874.

Hallgrímsdóttir IB, and Yuster DS. 2008. A complete classification of epistatic two-locus models. *BMC Genetics* **9**: 17. http://arxiv.org/abs/q-bio/0612044.

Hannum G, Srivas R, Guénolé A, Van Attikum H, Krogan NJ, Karp RM, and Ideker T. 2009. Genome-wide association data reveal a global map of genetic interactions among protein complexes. ed. K. Kerr. *PLoS genetics* **5**: e1000782. http://dx.plos.org/10.1371/journal.pgen.1000782 (Accessed February 6, 2013).

Hapmap Consortium. 2003. The International HapMap Project. *Nature* **426**: 789–96. http://www.ncbi.nlm.nih.gov/pubmed/14685227 (Accessed November 14, 2012).

Harrison C. 2011. Mood disorders: Small-molecule neurotrophin antagonist reduces anxiety. *Nature reviews. Drug discovery* **10**: 415. http://www.ncbi.nlm.nih.gov/pubmed/21629289 (Accessed February 7, 2013).

Herold C, Steffens M, Brockschmidt FF, Baur MP, and Becker T. 2009. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics* **25**: 3275–3281. http://www.ncbi.nlm.nih.gov/pubmed/19837719.

Hill WG, Goddard ME, and Visscher PM. 2008. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits ed. T.F.C. Mackay. *PLoS Genetics* **4**: 10. http://www.ncbi.nlm.nih.gov/pubmed/18454194.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 9362–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2687147&tool=pmcentrez&rendertype=abstract.

Holmans P, Green EK, Pahwa JS, Ferreira Manuel A R, Purcell Shaun M, Sklar Pamela, Owen MJ, O'Donovan MC, and Craddock N. 2009. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American journal of human genetics* **85**: 13–24. http://www.cell.com/AJHG/fulltext/S0002-9297(09)00209-2 (Accessed February 2, 2013).

Hu X, Liu Q, Zhang Zhao, Li Z, Wang S, He L, and Shi Y. 2010. SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell Research* **20**: 854–857. http://www.ncbi.nlm.nih.gov/pubmed/20502444.

Huggett J, Dheda K, Bustin S, and Zumla A. 2005. Real-time RT-PCR normalisation; strategies and considerations. *Genes and immunity* **6**: 279–84. http://www.ncbi.nlm.nih.gov/pubmed/15815687 (Accessed January 31, 2013).

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, et al. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**: 599–603. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11385576.

Ito T, Chiku S, Inoue E, Tomita M, Morisaki T, Morisaki H, and Kamatani N. 2003. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *American journal of human genetics* **72**: 384–98. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=379231&tool=pmcentrez&rendertype=abstract (Accessed January 3, 2013).

Jin CY, Anichtchik O, and Panula P. 2009. Altered histamine H3 receptor radioligand binding in post-mortem brain samples from subjects with psychiatric diseases. *British journal of pharmacology* **157**: 118–29. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2697790&tool=pmcentrez&rendertype=abstract (Accessed February 7, 2013).

Jon McClellan, and Mary-Claire King. Genetic Heterogeneity in Human Disease. *Cell*. http://www.cell.com/abstract/S0092-8674(10)00320-X.

Kam-Thong T, Czamara D, Tsuda K, Borgwardt K, Lewis CM, Erhardt-Lehmann A, Hemmer B, Rieckmann P, Daake M, Weber F, et al. 2010. EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European journal of human genetics EJHG* **19**: 465–471. http://eprints.pascal-network.org/archive/00007993/.

Kim S, Morris NJ, Won S, and Elston RC. 2010. Single-marker and two-marker association tests for unphased case-control genotype data, with a power comparison. *Genetic Epidemiology* **34**: 67–77. http://www.ncbi.nlm.nih.gov/pubmed/19557751.

Lander E S, and Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**: 231–9. http://www.ncbi.nlm.nih.gov/pubmed/3294162 (Accessed February 6, 2013).

Lehne B, Lewis CM, and Schlitt T. 2011. From SNPs to genes: disease association at the gene level. *PloS one* **6**: e20133. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3128073&tool=pmcentrez&rendertype=abstract (Accessed February 6, 2013).

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, et al. 2007. The diploid genome sequence of an individual human. *PLoS biology* **5**: e254. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1964779&tool=pmcentrez&rendertype=abstract (Accessed January 28, 2013).

Li H, Ruan J, and Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**: 1851–8. http://genome.cshlp.org/content/18/11/1851.abstract?ijkey=f1ac7a63ecf5fd205dee265c988dbc6f7814fc4b&keytype2=tf_ipsecsha (Accessed January 28, 2013).

Li M-X, Gui H-S, Kwan JSH, and Sham Pak C. 2011. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *American journal of human genetics* **88**: 283–93. http://www.cell.com/AJHG/fulltext/S0002-9297(11)00049-8 (Accessed February 4, 2013).

Li W, and Reich J. 1999. A complete enumeration and classification of two-locus disease models. *Human Heredity* **50**: 334–349. http://arxiv.org/abs/adap-org/9908001.

Lingärde B, Jönsson SAT, Luts A, and Brun A. 2000. Cerebellar abnormalities in mental illness. A study on Purkinje cell density in schizophrenic men. *European Child & Adolescent Psychiatry* **9**: 0021–0025. http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s007870050112 (Accessed February 7, 2013).

Liu Y, Xu H, Chen S, Chen X, Zhang Zhenguo, Zhu Z, Qin X, Hu L, Zhu J, Zhao G-P, et al. 2011. Genome-Wide Interaction-Based Association Analysis Identified Multiple New Susceptibility Loci for Common Diseases ed. D.B. Allison. *PLoS Genetics* **7**: 16. http://dx.plos.org/10.1371/journal.pgen.1001338.

Machado-Vieira R, Manji HK, and Zarate CA. 2009. The role of lithium in the treatment of bipolar disorder: convergent evidence for neurotrophic effects as a unifying hypothesis. *Bipolar disorders* **11 Suppl 2**: 92–109. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2800957&tool=pmcentrez&rendertype=abstract (Accessed February 7, 2013).

MacWilliams F, and Sloane N. 2006. The theory of error-correcting codes. http://www.citeulike.org/group/972/article/541435 (Accessed February 6, 2013).

Maher B. 2008. Personal genomes: The case of the missing heritability. *Nature* **456**: 18–21. http://www.nature.com/news/2008/081105/full/456018a.html (Accessed November 3, 2010).

Maloku E, Covelo IR, Hanbauer I, Guidotti A, Kadriu B, Hu Q, Davis JM, and Costa E. 2010. Lower number of cerebellar Purkinje neurons in psychosis is associated with reduced reelin expression. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 4407–11. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2840121&tool=pmcentrez&rendertype=abstract (Accessed February 7, 2013).

Mani R, St Onge RP, Hartman JL, Giaever G, and Roth FP. 2008. Defining genetic interaction. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 3461–3466. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265146&tool=pmcentrez&rendertype=abstract.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon Lon R, Chakravarti Aravinda, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–53. http://www.nature.com/nature/journal/v461/n7265/abs/nature08494.html (Accessed July 14, 2010).

Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly Peter, Faraone Stephen V, Frazer K, Gabriel S, et al. 2007. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genetics* **39**: 1045–1051. http://www.ncbi.nlm.nih.gov/pubmed/17728769.

Marchini J, Donnelly P, and Cardon L R. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics* **37**: 413–417. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15793588.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**: 133–141. http://www.ncbi.nlm.nih.gov/pubmed/18262675.

Margraf RL, Durtschi JD, Dames S, Pattison DC, Stephens JE, and Voelkerding K V. 2011. Variant identification in multi-sample pools by illumina genome analyzer sequencing. *Journal of biomolecular techniques : JBT* **22**: 74–84. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3121147&tool=pmcentrez&rendertype=abstract (Accessed February 25, 2013).

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–80. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1464427&tool=pmcentrez&rendertype=abstract (Accessed October 26, 2012).

McGuffin P, Rijsdijk F, Andrew M, Sham P, Katz R, and Cardno A. 2003. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Archives of general psychiatry* **60**: 497–502. http://www.ncbi.nlm.nih.gov/pubmed/12742871 (Accessed February 6, 2013).

Moore J H, and Williams SM. 2009. Epistasis and Its Implications for Personal Genetics. *The American Journal of Human Genetics* **85**: 309–320. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2771593&tool=pmcentrez&rendertype=abstract.

Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, et al. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34**: 267–73. http://dx.doi.org/10.1038/ng1180 (Accessed February 6, 2013).

Neale BM, Fagerness J, Reynolds R, Sobrin L, Parker M, Raychaudhuri S, Tan PL, Oh EC, Merriam JE, Souied E, et al. 2010. Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC). *Proceedings of the National Academy of Sciences of the United States of America* **107**: 7395–7400. http://www.ncbi.nlm.nih.gov/pubmed/20385826.

NHGRI. DNA Sequencing Costs. http://www.genome.gov/sequencingcosts/ (Accessed February 6, 2013).

Nuutinen S, and Panula Pertti. 2010. Histamine in Neurotransmission and Brain Diseases. *Advances in experimental medicine and biology* **709**: 95–107. http://www.ncbi.nlm.nih.gov/pubmed/21713693 (Accessed January 30, 2013).

Panagiotis Achlioptas BS and KB. 2011. Two-locus association mapping in subquadratic runtime. *The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Perrier E, Pompei F, Ruberto G, Vassos E, Collier D, and Frangou S. 2011. Initial evidence for the role of CACNA1C on subcortical brain morphology in patients with bipolar disorder. *European psychiatry the journal of the Association of European Psychiatrists* **26**: 135–137. http://www.ncbi.nlm.nih.gov/pubmed/21292451.

Phillips PC. 2008. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* **9**: 855–867. http://dx.doi.org/10.1038/nrg2452.

Piegorsch WW, Weinberg CR, and Taylor JA. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13**: 153–162. http://www.ncbi.nlm.nih.gov/pubmed/8122051.

Poirel CL, Owens CC, and Murali TM. 2011. Network-based functional enrichment. *BMC Bioinformatics* **12**: S14. http://www.biomedcentral.com/1471-2105/12/S13/S14 (Accessed February 6, 2013).

Prabhu S, and Pe'er I. 2012a. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res* gr.137885.112–. http://genome.cshlp.org/cgi/content/abstract/gr.137885.112v1.

Prabhu S, and Pe'er I. 2012b. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome research* **22**: 2230–40. http://genome.cshlp.org/content/22/11/2230.long (Accessed February 11, 2013).

Purcell S. Plink epistasis models documentation. http://pngu.mgh.harvard.edu/~purcell/plink/epidetails.shtml.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M A R, Bender D, Maller J, Sklar P, De Bakker PIW, Daly M J, et al. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**: 559–575. http://pngu.mgh.harvard.edu/purcell/plink/.

Raeymaekers L. 2000. Basic principles of quantitative PCR. *Molecular biotechnology* **15**: 115–22. http://www.ncbi.nlm.nih.gov/pubmed/10949824 (Accessed February 6, 2013).

Reich DE, Gabriel Stacey B, and Altshuler D. 2003. Quality and completeness of SNP databases. *Nature genetics* **33**: 457–8. http://www.ncbi.nlm.nih.gov/pubmed/12652301 (Accessed December 6, 2012).

Risch N, and Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)* **273**: 1516–7. http://www.ncbi.nlm.nih.gov/pubmed/8801636 (Accessed February 6, 2013).

Risch NJ. 2000. Searching for genetic determinants in the new millennium. *Nature* **405**: 847–56. http://dx.doi.org/10.1038/35015718 (Accessed November 3, 2010).

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, and Moore Jason H. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics* **69**: 138–147. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1226028&tool=pmcentrez&rendertype=abstract.

Van Ruitenbeek P, Sambeth A, Vermeeren A, Young SN, and Riedel WJ. 2009. Effects of L-histidine depletion and L-tyrosine/L-phenylalanine depletion on sensory and motor processes in healthy volunteers. *British journal of pharmacology* **157**: 92–103. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2697785&tool=pmcentrez&rendertype=abstract (Accessed February 7, 2013).

Sarin S, Snehit P, O'Meara MM, Pe'er I, and Hobert O. 2008. Caenorhabditis elegans mutant allele identification by whole- genome sequencing. *Nature Methods* **5**: 865–867.

Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PHS, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Alberts MJ, et al. 1993. Association of apolipoprotein E allele E4 with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43**: 1467–1472. http://www.neurology.org/cgi/content/abstract/43/8/1467.

Schüpbach T, Xenarios I, Bergmann S, and Kapur K. 2010. FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* **26**: 1468–1469. http://www.ncbi.nlm.nih.gov/pubmed/20375113.

Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, and Chakravarti A. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome research* **8**: 111–23. http://www.ncbi.nlm.nih.gov/pubmed/9477339 (Accessed November 21, 2012).

Sklar P, Smoller J W, Fan J, Ferreira M A R, Perlis R H, Chambert K, Nimgaonkar VL, McQueen MB, Faraone S V, Kirby A, et al. 2008. Whole-genome association study of bipolar disorder. *Molecular Psychiatry* **13**: 558–569. http://eprints.bournemouth.ac.uk/14297/.

Slavin TP, Feng T, Schnell A, Zhu X, and Elston RC. 2011. Two-marker association tests yield new disease associations for coronary artery disease and hypertension. *Human Genetics*. http://www.ncbi.nlm.nih.gov/pubmed/21626137.

Smith DR, Quinlan AR, Peckham HE, Makowsky K, Tao W, Woolf B, Shen L, Donahue WF, Tusneem N, Stromberg MP, et al. 2008. Rapid whole-genome mutational profiling using

next-generation sequencing technologies. *Genome research* **18**: 1638–42. http://genome.cshlp.org/content/18/10/1638.abstract?ijkey=2c2242b2b46c3c031f760c563a64dc24bf5ebde5&keytype2=tf_ipsecsha (Accessed January 28, 2013).

Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W, Byerley W, Coryell W, Craig D, Edenberg HJ, et al. 2009. Genome-wide association study of bipolar disorder in European American and African American individuals. *Molecular Psychiatry* **14**: 755–763. http://www.ncbi.nlm.nih.gov/pubmed/19488044.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander Eric S, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 15545–50. http://www.pnas.org/content/102/43/15545.long (Accessed January 28, 2013).

Tran KD, Smutzer GS, Doty RL, and Arnold SE. 1998. Reduced Purkinje cell size in the cerebellar vermis of elderly patients with schizophrenia. *The American journal of psychiatry* **155**: 1288–90. http://www.ncbi.nlm.nih.gov/pubmed/9734558 (Accessed February 7, 2013).

Tsai PT, Hull C, Chu Y, Greene-Colozzi E, Sadowski AR, Leech JM, Steinberg J, Crawley JN, Regehr WG, and Sahin M. 2012. Autistic-like behaviour and cerebellar dysfunction in Purkinje cell Tsc1 mutant mice. *Nature* **488**: 647–51. http://www.ncbi.nlm.nih.gov/pubmed/22763451 (Accessed February 4, 2013).

Tsuji S, Morinobu S, Tanaka K, Kawano K, and Yamawaki S. 2003. Lithium, but not valproate, induces the serine/threonine phosphatase activity of protein phosphatase 2A in the rat brain, without affecting its expression. *Journal of neural transmission (Vienna, Austria : 1996)* **110**: 413–25. http://www.ncbi.nlm.nih.gov/pubmed/12658368 (Accessed February 7, 2013).

Ueki M, and Cordell Heather J. 2012. Improved Statistics for Genome-Wide Interaction Analysis ed. Nicholas J. Schork. *PLoS Genetics* **8**: e1002625. http://dx.plos.org/10.1371/journal.pgen.1002625 (Accessed April 5, 2012).

Vila-Rodriguez F, Honer WG, Innis SM, Wellington CL, and Beasley CL. 2011. ApoE and cholesterol in schizophrenia and bipolar disorder: comparison of grey and white matter and relation with APOE genotype. *Journal of psychiatry & neuroscience : JPN* **36**: 47–55. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3004975&tool=pmcentrez&rendertype=abstract (Accessed February 7, 2013).

Wang DG. 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* **280**: 1077–1082. http://www.sciencemag.org/content/280/5366/1077.abstract (Accessed October 26, 2012).

Wang K, Li M, and Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics* **81**: 1278–83. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2276352&tool=pmcentrez&rendertype=abstract (Accessed January 30, 2013).

Wang X, Elston RC, and Zhu X. 2011. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nature Reviews Genetics* **12**: 74. http://www.ncbi.nlm.nih.gov/pubmed/21102529.

Wang X, Elston RC, and Zhu X. 2010. The meaning of interaction. *Human Heredity* **70**: 269–277. http://www.ncbi.nlm.nih.gov/pubmed/21150212.

White House Press Release. Announcement on the Human Genome Project. http://www.ornl.gov/sci/techresources/Human_Genome/project/clinton1.shtml (Accessed February 6, 2013).

Wu X, Dong H, Luo L, Zhu Y, Peng G, Reveille JD, and Xiong M. 2010. A Novel Statistic for Genome-Wide Interaction Analysis ed. Nicholas J Schork. *PLoS Genetics* **6**: 15. http://dx.plos.org/10.1371/journal.pgen.1001131.

Yang Q, Khoury MJ, Sun F, and Flanders WD. 1999. Case-only design to measure gene-gene interaction. *Epidemiology Cambridge Mass* **10**: 167–170. http://www.ncbi.nlm.nih.gov/pubmed/10069253.

Yutao Fu, Heather Peckham, Stephen McLaughlin, Jingwei Ni, Michael Rhodes, Joel Malek KM and AB. 2007. Solid system sequencing and 2-base encoding. *Cold Spring Harbor Laboratory Press*. http://marketing.appliedbiosystems.com/images/Product_Microsites/Solid_Knowledge_MS/pdf/CSHL_Fu.pdf (Accessed January 3, 2013).

Zeng D, and Lin DY. 2005. Estimating haplotype-disease associations with pooled genotype data. *Genetic epidemiology* **28**: 70–82. http://www.ncbi.nlm.nih.gov/pubmed/15558554 (Accessed January 2, 2013).

Zhang X, Huang S, Zou F, and Wang W. 2010. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics* **26**: i217–i227. http://www.bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq186.

Zhang X, Huang S, Zou F, and Wang W. 2011. Tools for efficient epistasis detection in genome-wide association study. *Source code for biology and medicine* **6**: 1. http://www.ncbi.nlm.nih.gov/pubmed/21205316.

Zhang X, Pan F, Xie Y, Zou F, and Wang W. 2010. COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study. *Journal of computational*

*biology a journal of computational molecular cell biology* **17**: 401–415. http://www.ncbi.nlm.nih.gov/pubmed/20377453.

Zhang X, Zou F, and Wang W. 2008. Fastanova: an efficient algorithm for genome-wide association study. *ACM Transactions on Knowledge Discovery from Data* **3**: 821–829. http://portal.acm.org/citation.cfm?id=1401890.1401988.

Zhang X, Zou F, and Wang W. 2009. FastChi: an efficient algorithm for analyzing gene-gene interactions. *Pacific Symposium On Biocomputing* 528–539. http://www.ncbi.nlm.nih.gov/pubmed/19209728.

Zhang Y, and Liu JS. 2007. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39**: 1167–1173. http://www.ncbi.nlm.nih.gov/pubmed/17721534.

Zhao J, Jin L, and Xiong M. 2006. Test for interaction between two unlinked loci. *The American Journal of Human Genetics* **79**: 831–845. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1698572&tool=pmcentrez&rendertype=abstract.

Zuk O, Hechter E, Sunyaev S, and Lander E. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*.

# Chapter 7. Appendix

Please note that notation is carried over from the appropriate sections of the main text.

## 7A. Statistical Test for Interaction

Under the null assumption that distal SNPs segregate independently in the population (and are therefore in linkage equilibrium), the expected $\vec{1}$-frequency for a pair of distal variables in cases can be estimated by $\mathrm{E}\left[P_{\vec{v}}(\vec{1})\right] = P_v P_{v\prime}$: where $P_v$ and $P_{v\prime}$ are the empirical 1-frequencies of $v$ and $v'$ respectively. Positive LD in cases results in an increase in the frequency of $\vec{1}$-carriers ($P_{\vec{v}}$), where LD is measured as a difference between the observed and expected frequency ($D_{\vec{v}}^{case} = P_{\vec{v}} - P_v P_{v'}$). If we denote $\Delta_{\vec{v}}^{case} = N D_{\vec{v}}^{case}$ as the number of excess $\vec{1}$-carriers in cases, then we derive the statistic

$$\begin{aligned}\left(LD_{\vec{v}}^{\mathrm{case}}\right)^2 &= \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{NP_vP_{v'}} + \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{NP_v(1-P_{v'})} + \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{N(1-P_v)P_{v'}} + \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{N(1-P_v)(1-P_{v'})} \\ &= \frac{\left(\Delta_{\vec{v}}^{case}\right)^2}{NP_v(1-P_v)P_{v'}(1-P_{v'})} \sim \chi^2 \; with \; 1 \; d.o.f.\end{aligned}$$

From which, we get

$$LD_{\vec{v}}^{case} = \frac{D_{\vec{v}}^{case}}{\sigma_{\vec{v}}^{case}} \sim \mathcal{N}(0,1) \quad \text{where } \sigma_{\vec{v}}^{case} = \sqrt{\frac{P_v(1-P_v)P_{v'}(1-P_{v'})}{N}}$$

Equation 7-1

A similar analysis for controls gives us

$$LD_{\vec{v}}^{control} = \frac{D_{\vec{v}}^{control}}{\sigma_{\vec{v}}^{control}} \sim \mathcal{N}(0,1) \quad \text{where } \sigma_{\vec{v}}^{control} = \sqrt{\frac{p_v(1-p_v)p_{v'}(1-p_{v'})}{n}}$$

Equation 7-2

Consequently, the LD in cases and controls can be contrasted to derive an LD-contrast statistic

$$LD_{\vec{v}}^{diff} = \frac{\left(D_{\vec{v}}^{case} - D_{\vec{v}}^{control}\right)}{\sigma_{\vec{v}}^{diff}} \sim \mathcal{N}(0,1) \quad \text{where } \sigma_{\vec{v}}^{diff} = \sqrt{\left(\sigma_{\vec{v}}^{case}\right)^2 + \left(\sigma_{\vec{v}}^{control}\right)^2}$$

Equation 7-3

Under our null-hypothesis, we would expect to see no difference in LD between cases and controls,

$$\mathbb{H}_0 : LD_{\vec{v}}^{diff} = 0$$

Significant variable-pairs $\vec{v}$ are those for which $LD_{\vec{v}}^{diff} \geq z_{\mathcal{B}}$, where the $z_{\mathcal{B}} = \Phi^{-1}(1-\mathcal{B})$ represents the number of standard deviations of the standard normal distribution $\mathcal{N}(0,1)$ required to achieve a *significance level of* $\mathcal{B}$. In the interest of clarity, we use the Bonferroni significance level without loss of generality (any other multiple test correction approach can be plugged in as easily). For example, in a dataset of $M = 450{,}000$ SNPs, if we perform $\left(\binom{450{,}000}{2} \times 4\right)$ pairwise tests (4 models tested per SNP-pair as per our binary encoding) genome-wide, giving us a significance threshold of $p = 1.2 \times 10^{-13}$. The LD-contrast cutoff required to achieve this significance level is $z_{\mathcal{B}} \approx 7$.

## 7B. Stage-1 filtering step

Consider a common disease with prevalence $\tau$ in the population. The LD between any two variables $\vec{v} = (v, v')$ in the entire population can be considered a mixture of two distributions

$$D_{\vec{v}}^{pop} \sim \tau\, D_{\vec{v}}^{case} \;+\; (1-\tau) D_{\vec{v}}^{control}$$

Equation 7-4

Assuming that physically unlinked alleles are in population-wide linkage equilibrium – i.e. $D_{\vec{v}}^{pop} \approx 0$ - we estimate that $\mathbb{E}\left[\, D_{\vec{v}}^{control} | D_{\vec{v}}^{pop} \approx 0 \,\right] = \frac{-\tau}{(1-\tau)} \mathbb{E}\left[\, D_{\vec{v}}^{case} | D_{\vec{v}}^{pop} \approx 0 \,\right]$. If variable-pairs exceed a disequilibrium cutoff $LD_{\vec{v}}^{case} \geq z'_B$ in cases (i.e. if $D_{\vec{v}}^{case} \geq z'_B \sigma_{\vec{v}}^{case}$), then for the variables to remain in population-wide equilibrium, the expected reverse disequilibrium in control required to counter the imbalance created by these cases is $\mathbb{E}\left[\, D_{\vec{v}}^{control} \mid D_{\vec{v}}^{pop} \approx 0,\; D_{\vec{v}}^{case} \geq z'_B \sigma_{\vec{v}}^{case} \right] \leq \frac{-\tau}{(1-\tau)} z'_B \sigma_{\vec{v}}^{case}$ . Substituting in Equation 2-1 (main text)**,** we get

$$\mathbb{E}\left[ LD_{\vec{v}}^{diff} \mid D_{\vec{v}}^{pop} \approx 0 \right] \;\geq\; \frac{1}{(1-\tau)} \times \frac{\sigma_{\vec{v}}^{case}}{\sigma_{\vec{v}}^{diff}} \times z'_B$$

Equation 7-5

For significant pairs $\left(LD_{\vec{v}}^{diff} \geq z_{\mathcal{B}}\right)$, by assuming the marginal frequencies of both variables are approximately equal in cases and controls (i. e. $P_v \approx p_v$ and $P_{v'} \approx p_{v'}$), we get

$$z'_B \;\geq\; (1-\tau) \times \sqrt{\frac{N+n}{n}} \times z_B$$

Equation 7-6

This result expresses the disequilibrium cutoff $z'_B$ in cases as a function of the disequilibrium-contrast cutoff $z_B$ in cases versus controls. To extend the example in 7A. , if $z_B \approx 7$ is the LD-

contrast Bonferroni cutoff for a common disease with population prevalence 5% in a dataset with an equal number of cases and controls, from Equation 2-3 we can estimate that in cases $z'_B \geq 9.4 \geq z_B$.

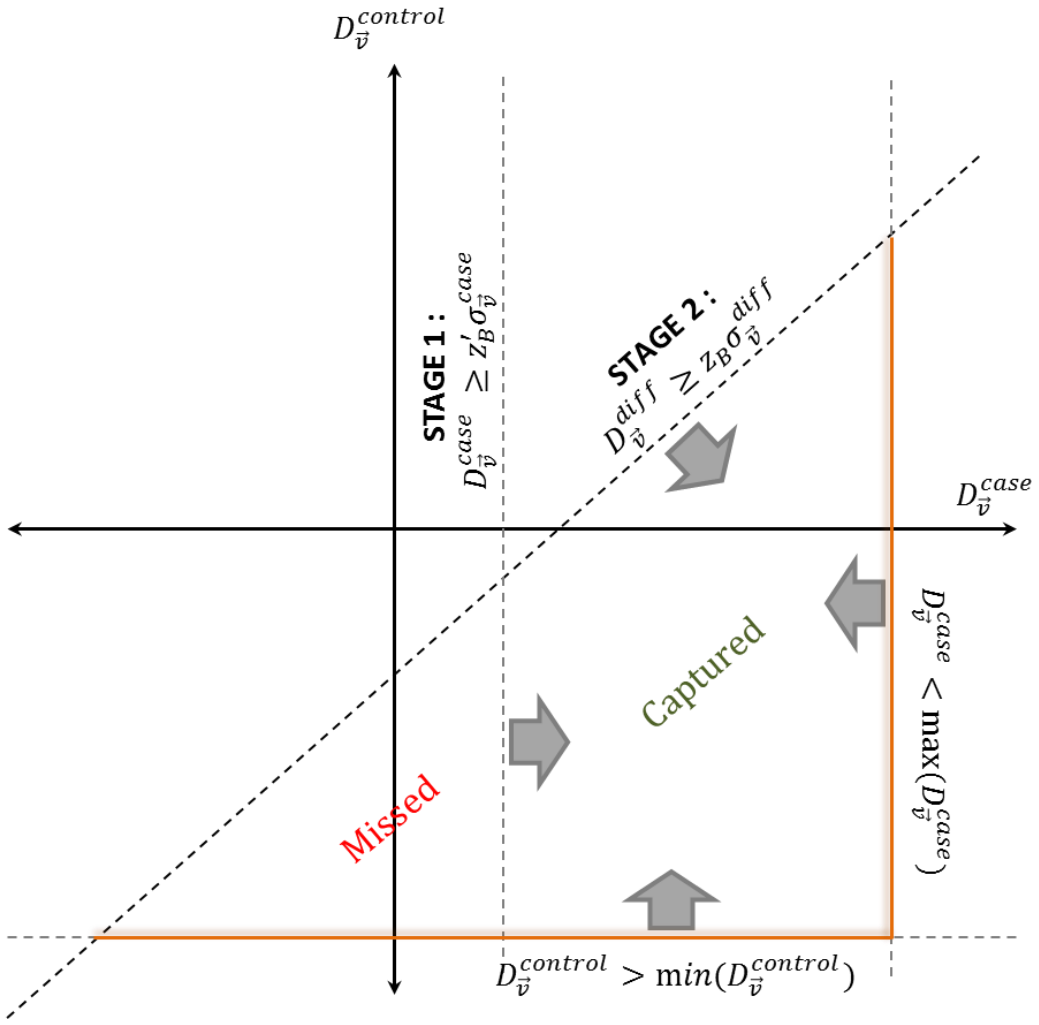
We say a stage 1 case-only analysis is *approximately complete* under this prescription, because while we have determined the expected value of the statistic for significant pairs, we have not characterized the full distribution.

This approximation depends on standard assumptions, particularly that most alleles have similar frequencies in cases and controls. For interactions between variables with large causal or protective marginal signals, we make the following observations: (i) For variables whose minor allele is enriched in cases (i.e. causal association) the approximation is actually violated in our favor: it underestimates our power to find interactions. (ii) For the converse case, where the minor allele is depleted in cases (i.e. protective association), the approximation does indeed falter. However, using an ultra-conservative approach in which we lower the stage 1 cutoff for candidates to be the same as the stage 2 cutoff (i.e. $z'_B = z_B$), is observed to be sufficient to accommodate such violations in practice. In other words, we can largely capture interactions between SNPs whose frequency is greater in cases as well as greater in controls (SNPs with main effects). Lastly and importantly, these loci have typically already been identified by single-locus association and are therefore accessible to a candidate gene based analysis.

## 7C. The approximate nature of a stage-1 case-only analysis

The area enclosed within the orange lines represents the region occupied by significant variable-pairs in linkage disequilibrium space (LD in cases on the X-axis, controls LD on the Y-axis). The three sides of the triangle are the maximum (i.e. positive) LD in cases, the minimum (i.e. negative) LD in controls and the LD-contrast threshold used to demarcate significant pairs. The dotted vertical line represents the stage 1 cutoff used for shortlisting tuples through a case only analysis. The number of significant pairs in the area labeled “Missed” and “Captured” depends upon the density distribution of pairs in this space. We show that for common diseases almost all

statistically significant pairs lie in the captured region (see section on Approximately Complete Search).




## 7D. Applying group sampling to a genome-wide scan.

In the toy example described in the main text, we restricted our discussion to finding pairs of variables that occupy a narrow frequency window $w$. To generalize the approach to a genome-wide search for significant LD-contrasts, we first partition the entire spectrum of frequencies [0,1] into $R$ windows $W = \{w_0, \dots, w_{r-1}\}$ of ranges $E = \{\epsilon_0, \dots, \epsilon_{r-1}\}$ respectively, where each $w_r = [\eta_r, \eta_{r+1})$; $\eta_{r+1} = \eta_r + \epsilon_r$ and $\eta_0 = 0, \eta_r = 1$. We then allocate each of

the $2M$ variables genome-wide to their appropriate frequency windows (see 7H. ). As before, the number of variables in a window $w_r$ is denoted $V(w_r)$, and every variable is assigned to exactly one window: $\sum_r |V(w_r)| = 2M$. In practice, we find that using around 50 windows is adequate to cover the frequency spectrum even in large datasets of $\geq 10^{5-6}$ SNPs.

Consider any pair of windows $\{w_A, w_B\}$. There are $\binom{W}{2} + W$ such window pairs (including the possibility that A=B). For all $\vec{v} = (v_A, v_B)$ comprising of one variable $v_A \in V(w_A)$ and the other $v_B \in V(w_B)$, the minimum and maximum expected $\vec{1}$-frequencies can be derived as per Equation 2-5 (main text). If $\mathbb{H}_0'$ holds for $\vec{v}$, then $(P_{\vec{v}})^k \leq (\eta_{A+1}\eta_{B+1})^{2k}$. If $\mathbb{H}_0'$ is rejected by $\vec{v}$ then $(P_{\vec{v}})^k \geq (\eta_A\eta_B + \delta_{A\times B})^{2k}$, where $\delta_{A\times B}$ is derived as

$$\delta_{A\times B} = \sqrt{\frac{\eta_A(1-\eta_A)\eta_B(1-\eta_B)}{N}}\, z_{\mathcal{B}}$$

Equation 7-7

We are required to find the group-sampling parameter values $(k_{A\times B}, t_{A\times B})$ for this window pair, at which we can guarantee that all significant-pairs which reject $\mathbb{H}_0'$ are observed in at least one group with probability greater than the user-specified threshold $(1-\beta)$. Furthermore, our solution $(k_{A\times B}, t_{A\times B})$ has to be "optimal" in two ways : (i) the false positive rate should be low (which requires number of individuals per draw - $k_{A\times B}$ - to be large) because these candidates will have to be kept in memory, only to be screened out later by stage 2, and (ii) the solution should not consume too many compute cycles (large $k_{A\times B}$ requires a large number of random draws $t_{A\times B}$ to achieve the desired power, which in turn drives up the number of compute cycles). Details of the optimization procedure we employed to find the best $k_{A\times B}$ and $t_{A\times B}$ are provided below.

To summarize, group-sampling lets us restrict our test to an exponentially (in $k_{A\times B}$) small fraction $f_{A\times B} \leq t_{A\times B}.(\eta_{A+1}\eta_{B+1})^{2k_{A\times B}}$ of the pairs of variables in $V(w_A) \times V(w_B)$ at minimum computational "cost" (as discussed in the optimization section below), and simultaneously

guarantees that all significant variable pairs will be captured in this fraction of the universe with power $\geq (1-\beta)$. This makes stage 1 of our search experiment extremely rapid.

Although the sheer size of the universe of combinations $U_{AB} = |V(w_A) \times V(w_B)|$ can suggest a large number of false-positives $\alpha U_{AB}$ in stage 1 overall. We make three observations to alleviate this concern: (i) This constitutes an upper bound which (by definition) is rarely encountered in empirical data, (ii) Most false-positive pairs are observed in more than one sampled group. However, these are stored in memory using a hash-table the very first time they are encountered, and have to be tested only once by the stage 2 analysis. The upper bound appears large because it does not account for such "over counting", and finally (iii) A poor stage 1 false-positive rate comes at a computational cost, but does not affect the accuracy of the algorithm. False-positive candidates like these are screened out in stage 2.

*Details of the optimization procedure:*

Translating Equation 2-5 to our current setting, and expressing $k$ as a function of $t$ gives,

$$k_{A\times B}(t) \leq \frac{\log(1-\beta^{1/t})}{\log(\eta_A \eta_B + \delta_{A\times B})}$$

Equation 7-8

There are a total of $\binom{|V(w_A)|+|V(w_B)|}{2}$ potential variable-pairs between these windows. The number of pairs that emerge purely by chance over $t_{A,B}$ random draws is estimated as

$$Chance_{A\times B}(t) \approx \binom{|V(w_A)| + |V(w_B)|}{2} . t_{A\times B} . (P_{\vec{v}})^{k_{A\times B}}$$

And since $(P_{\vec{v}})^{k_{A\times B}} \leq (\eta_{A+1}\eta_{B+1})^{2k}$ for these chance pairs, we get

$$Chance_{A\times B}(t) \leq \binom{|V(w_A)| + |V(w_B)|}{2} . t_{A\times B} . (\eta_{A+1}\eta_{B+1})^{2k_{A\times B}}$$

Equation 7-9

This gives us an upper bound on the number of pairs that turn up purely by chance. We confirmed this bound in practice: since most co-occurring variables are encountered in several random draws, they need not be investigated more than once if we record them in a hash-table in memory. We can now find the optimal parameter values $k_{A\times B}$ and $t_{A\times B}$ that satisfy Equation 7-9 while minimizing the overall cost function,

$$k_{A\times B}, t_{A\times B} \longleftarrow \text{argmin}_t(\lambda_1 Chance_{A\times B}(t) + \lambda_2 t + \lambda_3 \mathbb{I}(k < 2))$$

Equation 7-10

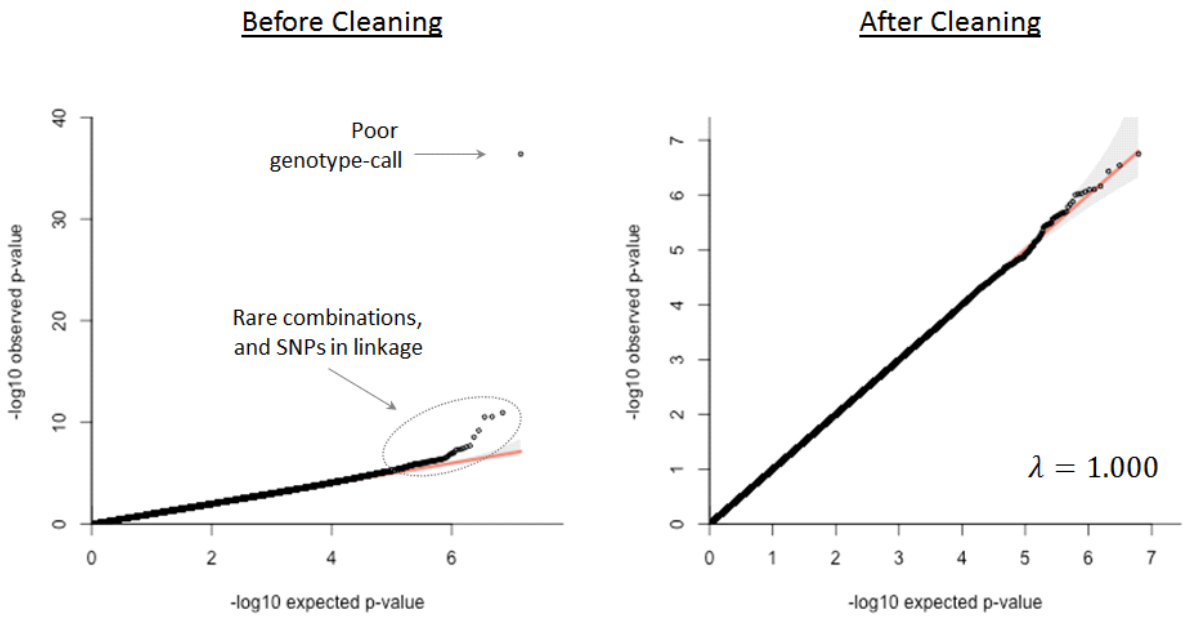
where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are the cost of shortlisting and validating a chance pair, the cost of a random draw, and a Lagrange multiplier to avoid degenerate values of $k$ respectively, while $\mathbb{I}(.)$ is the indicator function. These costs depend on the particular software implementation and data-structures used.

## 7E. QQ plots for LD-contrast test (sub genome-wide)

We drew 15 million random pairs of binary variables $\vec{v} = (v, v')$ from the cleaned WTCCC dataset, and contrasted their LD between bipolar cases and joint controls. We consider this a representative sample of the full space of 400 billion pairs which is computationally difficult to test. We observed over-dispersion that suggested a deviation from the null-hypothesis. We filtered out pairs with an extremely low expected number of minor allele co-carriers (i.e. remove pairs with $NP_vP_{v'} < 4$ or $np_vp_{v'} < 4$ in cases or controls respectively) because these might inflate the statistic due to unstable variance estimates. Further, we filtered out pairs comprising of SNPs in genetic linkage (<5cM apart) which cannot be treated as independent random variables when calculating co-carrier expectations of the 2×2 table : we observed an over-

dispersion of LD-contrast p-values on random pairs of such physically linked SNPs. However, since interactions between nearby markers (e.g. neighboring genes or markers within a gene) quite possibly comprise a significant portion of the interaction space, modifications to the test that can adjust for this variance inflation due to multi-collinearity are the subject of future work. Lastly, in addition to WTCCC prescription, we removed SNPs whose CHIAMO genotype-calling confidence was <95% in >1% of the individuals of the dataset. Conservative filters help us avoid false positives when we report pairs in the genome-wide significance range ($p < 10^{-12 \text{ to } 13}$). The resulting QQ plots show that for pairs that pass our filters, the LD-contrast test operates on a robust null-hypothesis and does not suffer from any residual over dispersion in BD data ($\lambda = 1.000$).

We note here that each QQ plot only presents a subset of randomly chosen variable-pairs out of the potential $4.15 \times 10^{11}$ pairs that exist genome-wide in this dataset. It is computationally prohibitive to test (on the order of) a trillion pairs of variables, sort their p-values and plot as many points without specialized software and computational infrastructure (indeed, avoiding this is the primary motivation of our work) but genomic over-dispersion (inflation of the median value) can be estimated robustly with a representative sample of the universe of pairs. Additionally, for SNP-pairs that do pass the Bonferroni cut-off genome-wide, we also perform a permutation analysis to verify their significance.




## 7F. Synthetic dataset construction.

We tested the accuracy of the randomization algorithm under various simulated scenarios. In particular, we were interested in whether SIXPAC always finds SNP-pairs with a significant LD-contrast level at (or above) the computational power requested by user.

Using the original WTCCC BD case-control cohort, we simulated 3 datasets to contain SNP-pairs with significant LD-contrast. These datasets capture a range of different scenarios concerning disease prevalence levels in the population (1% to 25%), minor allele frequencies of interacting SNPs (5% to 40%), as well as mode of interaction (recessiveness and dominance). The datasets were synthesized through a technique called *chromosomal shuffling*, as follows:

i. First choose one SNP on each chromosome (All MAF 40% for dataset 1, MAF 30% and MAF 10% on alternate chromosomes for dataset2, and all MAF 5% for dataset3).

ii. Ascertain that these SNPs presented no discernible marginal significance (p >0.1) according to five standard single-locus association tests (allelic, genotypic, trend, dominance and recessive) offered by Plink (Purcell et al. 2007).

iii. Next, we introduce LD-contrasts into the dataset without changing the marginal frequencies of the SNPs either in cases on controls. To do this, we swap entire chromosomes among cases (controls) so as to increase (decrease) the number of co-carriers of minor alleles in these cohorts. The ratio of cases/controls to shuffle during each iteration is determined by the prevalence estimate.

For example, we create one additional recessive-recessive co-carrier in the case dataset (without affecting the marginal signal) we follow these steps.

Let case 2 be recessive at SNP A (chr21) and case 3 be recessive at SNP B (chr22), then:

(i) swap chromosome 21 of case 1 and case 2

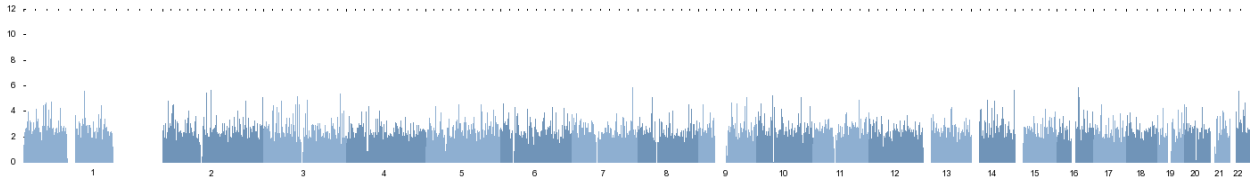
(ii) swap chromosome 22 of case 1 and case 3.

Case 1 is now a carrier of the recessive-recessive pair at SNPs A and B. The controls can have the number of co-carriers depleted by analogous shuffling.

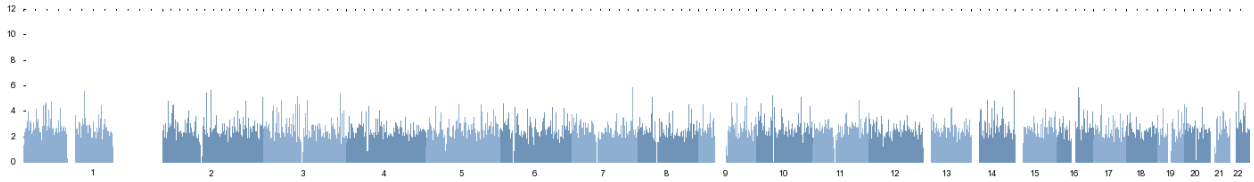
iv. By shuffling case and control chromosomes in this manner, we simulated 11 interactions (between SNPs on 22 autosomes) in each of the 3 datasets – each interaction at different levels of LD-contrast significance from $p = 10^{-2}, 10^{-4}, \dots, 10^{-22}$ (decrements of $\approx 10^{-2}$). Note that because of the discrete nature of the shuffle, it is not always possible to achieve accurate LD-contrast p-values in the synthetic data (e.g. $p = 1.0 \times 10^{-2}, 1.0 \times 10^{-4}, \dots, 1.0 \times 10^{-22}$). Instead, after each swap we perform an LD-contrast test between this SNP-pair to check if the co-carrier imbalance introduced between cases and controls is sufficient to provide the required level of significance. We stop when we cross this level.

Chromosomal shuffling allows us to effectively manipulate LD between SNPs in cases (and controls) without changing the marginal association signal at all. This can be verified by checking that the Manhattan plot (single-locus, allelic model) *before* chromosomal shuffling in

cases



And the corresponding Manhattan plot *after* simulating recessive-recessive co-carriers in the cases are exactly identical.



Although differentiating LD in the case and control datasets is not intended to directly confer statistical epistasis, we can also analyze these simulated SNP-pairs using the traditional model for interaction in a case-control study. This involves applying logistic regression, which tests whether the two loci - when considered in conjunction - result in a deviation from multiplicative odds.

First we test whether the SNPs in the synthetic datasets have any main effects – by testing the term $\beta_1$ or $\beta_2$ term in a logistic regression $\ln(\frac{P}{1-P}) = \beta_1 \mathrm{G}_1 + \beta_2 G_2$, where $G_1$and $G_2$ are binary predictor variables that encode :

i. recessive carrier status for interacting SNPs in dataset1
i. recessive and dominant carrier status respectively for interacting SNPs in dataset2, or
ii. dominance carrier status for interacting SNPs in dataset3.

As the tables below confirm, in all 3 scenarios, LD-contrast does not inflate main-effect estimates. Next we test for the multiplicative interaction effects – by testing for significance of

the $\beta_{12}$ term in a logistic regression using the full model: $\ln\left(\frac{P}{1-P}\right) = \beta_1 \mathrm{G}_1 + \beta_2 G_2 + \beta_{12} G_1 G_2$. We note that increasing LD-contrast is strongly indicative of increasing statistical epistasis on this scale (see 3 dataset tables below), although the correlation between the 2 tests is not perfect (see discussion elsewhere (Shaun Purcell; Kam-Thong et al. 2010)). Fully elucidating the wide range of models and alternate parameterizations that may be visible through such LD-contrasts is the subject of future work.

Table 7-1 Dataset 1: Common × Common Interaction. LD-contrast simulated between a 40% MAF SNP (in recessive mode) with a 40% MAF SNP (in recessive mode), disease prevalence 25%.

| Simulated interactions (approximate significance) | Main effects models | | Full model | Empirical significance (p-value) | |
|---|---|---|---|---|---|
| | odds ratio ($e^{\beta_1}$) | odds ratio ($e^{\beta_2}$) | odds ratio ($e^{\beta_{12}}$) | $\beta_{12}$ term (epistasis) | LD-contrast |
| chr 1 – chr 2 ($10^{-2}$) | 1.0 | 1.01 | 1.6 | 0.01 | 8.2E-03 |
| chr 3 – chr 4 ($10^{-4}$) | 0.99 | 1.0 | 2.0 | 2.0E-04 | 6.8E-05 |
| chr 5 – chr 6 ($10^{-6}$) | 1.0 | 1.0 | 2.6 | 3.7E-06 | 6.0E-07 |
| chr 7 – chr 8 ($10^{-8}$) | 1.0 | 1.0 | 2.9 | 1.5E-07 | 8.2E-09 |
| chr 9 – chr 10 ($10^{-10}$) | 1.0 | 1.0 | 3.5 | 1.7E-09 | 4.2E-11 |
| chr 11 – chr 12 ($10^{-12}$) | 1.0 | 0.99 | 4.0 | 1.2E-11 | 8.9E-13 |
| chr 13 – chr 14 ($10^{-14}$) | 1.0 | 0.99 | 4.2 | 2.6E-12 | 7.2E-15 |
| chr 15 – chr 16 ($10^{-16}$) | 0.99 | 1.0 | 4.5 | 3.6E-14 | 5.5E-17 |
| chr 17 – chr 18 ($10^{-18}$) | 1.0 | 1.0 | 5.3 | 3.2E-16 | 8.2E-19 |
| chr 19 – chr 20 ($10^{-20}$) | 1.0 | 0.99 | 6.3 | <2E-16 | 5.7E-21 |
| chr 21 – chr 22 ($10^{-22}$) | 1.0 | 1.0 | 6.5 | <2E-16 | 3.3E-23 |

Table 7-2 Dataset 2: Rare × Common Interaction**.** LD-contrast simulated between a 10% MAF SNP (in dominant mode) with a 30% MAF SNP (in recessive mode), disease prevalence 10%.

| Simulated interactions (approximate significance) | Main effects models | | Full model | Interaction significance (p-value) | |
|---|---|---|---|---|---|
| | odds ratio ($e^{\beta_1}$) | odds ratio ($e^{\beta_2}$) | odds ratio ($e^{\beta_{12}}$) | $\beta_{12}$ term (epistasis) | LD-contrast |
| chr 1 – chr 2 ($10^{-2}$) | 1.01 | 0.99 | 1.7 | 0.01 | 8.8E-03 |
| chr 3 – chr 4 ($10^{-4}$) | 1.01 | 1.0 | 2.3 | 4.0E-04 | 7.9E-05 |
| chr 5 – chr 6 ($10^{-6}$) | 0.99 | 0.99 | 2.8 | 4.7E-06 | 5.9E-07 |
| chr 7 – chr 8 ($10^{-8}$) | 1.01 | 1.0 | 3.4 | 2.4E-07 | 6.9E-09 |
| chr 9 – chr 10 ($10^{-10}$) | 1.0 | 0.99 | 3.8 | 2.7E-09 | 3.5E-11 |
| chr 11 – chr 12 ($10^{-12}$) | 1.0 | 1.0 | 4.2 | 2.6E-10 | 6.9E-13 |
| chr 13 – chr 14 ($10^{-14}$) | 1.01 | 0.99 | 5.4 | 1.6E-12 | 6.7E-15 |
| chr 15 – chr 16 ($10^{-16}$) | 0.99 | 0.99 | 5.7 | 5.7E-14 | 5.8E-17 |
| chr 17 – chr 18 ($10^{-18}$) | 1.0 | 1.0 | 5.6 | 1.3E-14 | 5.3E-19 |
| chr 19 – chr 20 ($10^{-20}$) | 1.0 | 1.0 | 6.0 | 7.6E-16 | 9.9E-21 |
| chr 21 – chr 22 ($10^{-22}$) | 0.99 | 1.0 | 7.1 | <2E-16 | 3.0E-23 |

Table 7-3 Dataset 3: Rare × Rare Interaction. LD-contrast simulated between a 5% MAF SNP (in dominant mode) with a 5% MAF SNP (in dominant mode), disease prevalence 1%.

| Simulated interactions (approximate significance) | Main effects models | | Full model | Interaction significance (p-value) | |
|---|---|---|---|---|---|
| | odds ratio ($e^{\beta_1}$) | odds ratio ($e^{\beta_2}$) | odds ratio ($e^{\beta_{12}}$) | $\beta_{12}$ term (epistasis) | LD-contrast |
| chr 1 – chr 2 ($10^{-2}$) | 1.0 | 1.0 | 1.9 | 0.017 | 9.4E-03 |
| chr 3 – chr 4 ($10^{-4}$) | 1.01 | 1.04 | 2.8 | 4.4E-04 | 5.2E-05 |
| chr 5 – chr 6 ($10^{-6}$) | 1.0 | 0.99 | 3.1 | 1.8E-05 | 6.4E-07 |
| chr 7 – chr 8 ($10^{-8}$) | 1.0 | 1.0 | 3.7 | 7.9E-07 | 4.2E-09 |
| chr 9 – chr 10 ($10^{-10}$) | 0.99 | 0.99 | 4.0 | 9.7E-08 | 8.5E-11 |
| chr 11 – chr 12 ($10^{-12}$) | 1.0 | 1.01 | 5.3 | 4.6E-10 | 2.9E-13 |
| chr 13 – chr 14 ($10^{-14}$) | 0.99 | 1.01 | 5.1 | 8.2E-11 | 7.5E-15 |
| chr 15 – chr 16 ($10^{-16}$) | 1.0 | 1.0 | 6.2 | 3.0E-12 | 8.1E-17 |
| chr 17 – chr 18 ($10^{-18}$) | 1.0 | 1.0 | 6.3 | 1.7E-13 | 1.9E-19 |
| chr 19 – chr 20 ($10^{-20}$) | 1.0 | 1.0 | 6.5 | 7.9E-14 | 3.4E-21 |
| chr 21 – chr 22 ($10^{-22}$) | 1.0 | 0.99 | 7.9 | 5.1E-16 | 4.0E-23 |

## 7G. Power of Algorithm

To confirm that theoretical estimates of algorithm power were matched or exceeded by our implementation, we tested SIXPAC on the three simulated datasets, each containing 11 pairwise SNP-SNP interactions (LD-contrast) at different levels of significance as described in Supplementary Section 3.

SIXPAC accepts two critical inputs from the user, based on which it calculates search parameters

1. Significance cutoff as a p-value – all LD-contrasts above this cutoff must be reported.
2. Power (probability) to find these significant pairs, demanded by the user.

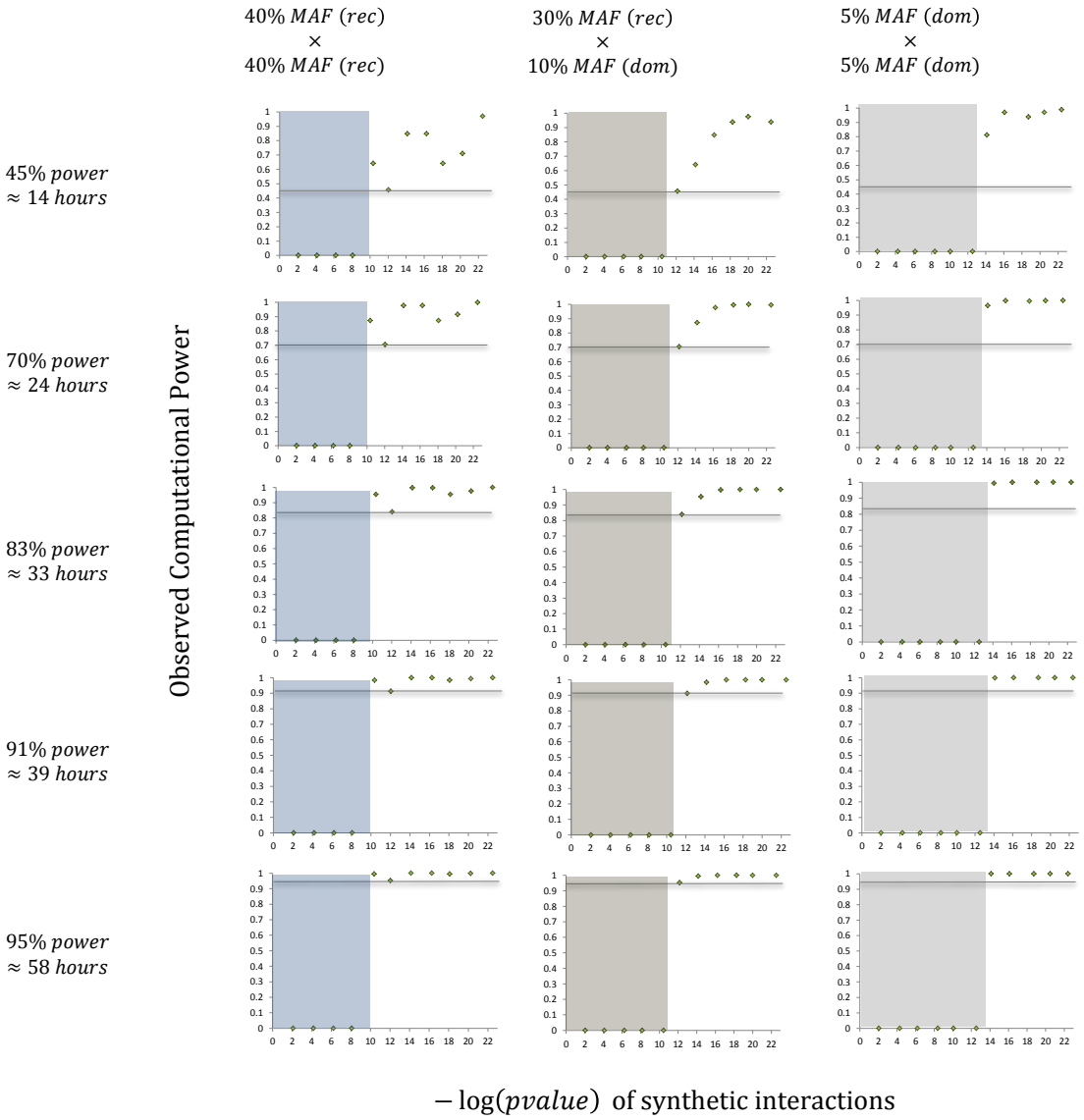
For the purposes of this simulation experiment, we arbitrarily defined the LD-contrast significance cutoff at 3 different realistic values of $p < 10^{-10}, 10^{-11}$ and Bonferroni ($1.2 \times 10^{-13}$) for datasets 1, 2 and 3 respectively. We note that any arbitrary cutoff value, lower or higher than these values, can be provided by the user. For the computational power parameter,

we measured results over 5 different realistic values – 45%, 70%, 83%, 91% and 95% probability respectively. Here, power of the algorithm is defined as the probability of finding all SNP-pairs in the dataset with a significant LD-contrast. As we discussed in the main text, this is different from statistical power.

Each panel in the figure below represents the result of a SIXPAC run with a particular combination of power and significance cut-off. The shaded rectangle in each panel represents the significance cut-off : interactions below this threshold are not reported. The solid line represents the theoretical power required by the user – and guaranteed as per theoretical estimates. We wish to determine whether the interactions to the *right* of the shaded area are *above* the cutoff threshold line, as promised.

In the panels below, each interaction is represented by a green dot: the X-axis co-ordinate gives the –log(p) value of its LD-contrast, while its Y-axis co-ordinate gives the average observed probability of spotting the interaction by SIXPAC (100 runs) - under each particular power, cut-off setting. We can see that as per guarantees, pairs with an LD-contrast above the significant cut-off are always reported with probability greater than the user-prescribed baseline.

For each dataset, SIXPAC scanned approximately 400 billion pairwise tests (4 tests per SNP-pair). We report the times taken by each SIXPAC run on a Single Intel i7 processor (quad-core) with 8GB RAM alongside. We note that like any randomization algorithm, SIXPAC will require an infinite amount of compute time to reach 100% certainty of finding everything in a dataset, but can approach close to 100% with large compute savings.




## 7H. Frequency Binning

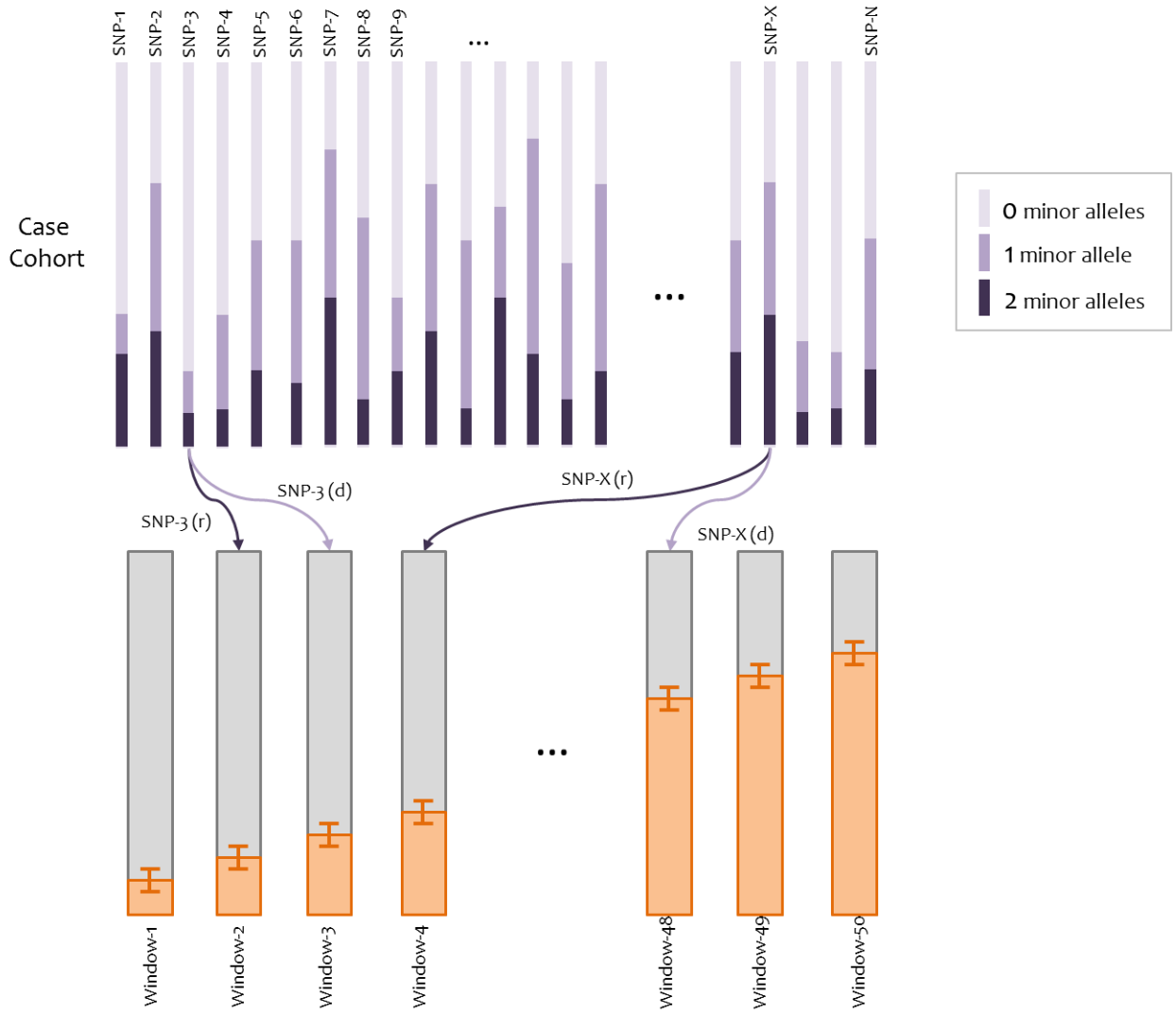
For each SNP, we consider the empirical frequency of the 2 encoded binary variables in cases (recessive and dominance carrier status). Each variable is then assigned to a narrow frequency

bin, as shown. Our algorithm operates by considering pairs of windows: since LD is a function of the frequency of two variables, we can conversely estimate how frequent a combination would need to be in order for the LD to be statistically significant. Group sampling exploits this difference in frequency between pairs with significant LD and pairs in equilibrium to rapidly shortlist interaction candidates.

The optimal width of a frequency window is difficult to characterize analytically : it depends on the significance cut-off, statistical test being implemented, number of SNPs typed in the dataset as well as the number of samples.

A. On the one hand, having many windows with a narrow frequency range makes it easy to distinguish between statistically significant LD - $\left(\tilde{P}^2 + \delta_{w\times w}\right)$ - and a SNP-pair that is at the upper end of the frequency spectrum - $\left(\tilde{P} + \epsilon\right)^2$. On the positive side, this reduces the number of shortlisted candidates per window-pair. However, many narrow windows means a quadratic increase in the number of window-pairs that have to be considered – and each pair must go through millions of group-sampling iterations, which can be computationally expensive.
B. On the other hand, having fewer but wider frequency windows does not allow group sampling to distinguish between a pair with a statistically significant increase in frequency, and a pair that is at the upper end of the permitted frequency spectrum - $\left(\tilde{P}^2 + \delta_{w\times w}\right)$ and $\left(\tilde{P} + \epsilon\right)^2$ respectively, from the main text. This can result in a large false positive rate at the stage 1 shortlisting step.

For WTCCC size case-control datasets, using the LD-contrast test, at a significance level of $p < 10^{-12}$, we found that using 50 to 60 windows provided the best performance.




## 7I. Numerical Example.

We describe a particular example of a joint-effect we pursue, in order to provide sense of the actual numbers involved. Consider a realistic GWAS dataset of 10,000 case and 40,000 control

samples from a population. If the disease prevalence in this population is 4%, then cases are oversampled 5-fold by the ascertainment of this study.

Consider two unlinked SNPs of 5% MAF each (in HWE, same MAF in both cases and controls and thus no marginal signal at either). The dominant-variable of each SNP (which encodes whether an individual carries $\geq 1$ minor alleles at the SNP) has a frequency of 9.75% in both datasets, and hence under the null hypothesis we expect 975 and 3,900 dominant carriers among the cases and controls respectively. Consequently, around ~95 cases and ~380 controls are expected to be "co-carriers" of these alleles when they are in perfect linkage equilibrium in both datasets ($LD_{\vec{v}}^{control} = LD_{\vec{v}}^{case} = 0$).

Now let us assume the specific alternative hypothesis under a certain interaction model (note: this may not be the only interaction that an LD-contrast captures, but is used here simply for illustrative purposes). Suppose that the disease penetrance for individuals carrying 1 or more minor alleles at both SNPs is 5%. If so, we expect to observe just ~19 fewer controls as co-carriers, leaving ~361 control co-carriers ($LD_{\vec{v}}^{control} = -1.08$). However, this small deflation in control co-carriers will be counterbalanced by an overabundance of ~95 co-carriers among cases due to the ascertainment bias (5-fold oversampling of cases), resulting in ~190 observed case co-carriers. This addition of 95 carriers to the background marginal count of 975 dominant carriers for each SNP, results in an observed marginal frequency of 10.7% in cases (up from 9.75%). Given these marginal frequencies, we would expect ~114.5 dominant co-carriers, which our observations exceed by ~75.5 ($LD_{\vec{v}}^{case} = 7.9,\ p \approx 1.4 \times 10^{-15}$).

Note that a signal of -19 (out of 40,000) vs. +75.5 (out of 10,000) co-carriers is highly significant ($D_{\vec{v}}^{control} = -0.000475$, $D_{\vec{v}}^{case} = +0.0075$, $LD_{\vec{v}}^{diff} = 7.6$, $p < 1.5 \times 10^{-14}$), and will pass the multiple testing burden of all pairs of variables in most experiments. In particular, we note here that the LD-case statistic was even more extreme and indicative of a significant LD-contrast, just as we had concluded in the section on Approximately Complete Search.

Group sampling utilizes this difference in co-carrier frequencies as follows. If the dominant × dominant allelic combination was in perfect linkage equilibrium in both datasets, then by randomly sampling ($k = 4$) cases, the probability of all 4 cases being co-carriers by chance is $0.0095^4 = 8.1 \times 10^{-9}$. In the alternative situation when there is penetrance of co-carriers, this probability is $0.0190^4 = 1.3 \times 10^{-7}$. If we draw $10^6$ such groups of cases at random, then the probability that we will sample all co-carriers *at least once* is >12% if they are synergistic, while it is <0.7% if they are not. In this manner, group sampling makes it highly plausible that the joint-effects pair of variants will be observed under the alternative, not so under the null. Because group sampling utilizes Binary computer operations, even a million random draws can be accomplished in relatively insignificant amount of time.

## 7J. Application of SIXPAC in functional enrichment designs.

SIXPAC requires its user to provide two search parameters which are pertinent to this experiment: (i) a significance threshold, and (ii) a power setting. The significance threshold, which is provided as a p-value, tells SIXPAC to only report SNP-pairs whose epistasis term significance crosses this cutoff. Lowering the significance threshold (i.e. raising the p-value

cutoff) results in an increase in computational costs that can quickly make our study design intractable. In addition to the computational savings, there is also a statistical rationale to this: high p-values represent diminished evidence of epistasis, thereby increasing the proportion of false-positives (truly non-epistatic combinations) in our list. In other words, the proportion of statistical signal gleaned to statistical noise added (our metaphorical bang-for-buck) by lowering the threshold keeps decreasing. As a practical balance, we use a threshold of $p_{\text{cutoff}} = 10^{-7}$, which is roughly a million-fold more relaxed that the Bonferroni significance levels ($\sim 10^{-13}$). The pvalues of SNP-pairs that are not reported by this search are set to 1.

The power parameter of SIXPAC denotes the probability with which the list of SNP-pairs reported at the end of the search is complete. One of the key properties of SIXPAC is its use of randomization to achieve computational speedup at the cost of certainty. Once again, increasing certainty comes at the cost of disproportionate compute time. We apply SIXPAC to 95% certainty – in other words, there remains a 5% chance that the list is incomplete (alternately, we expect to miss 5% of SNP-pairs whose epistasis stands at $p < 10^{-7}$). However, since we apply the same search parameters on every permuted dataset as well, we are reasonably protected against any systematic biases.

## 7K. LOD contrast statistic vs. Logistic regression

The usual scale applied to test for statistical epistasis between a pair of variants $S_A = \{0,1\}$ and $S_B = \{0,1\}$ – where 0 and 1 represent reference and mutant allele respectively – is to check whether the combined effect size of carrying mutant alleles at both loci (measured as the log of odds ratio) deviates from the addition of their individual effect sizes. This is typically accomplished by comparing the fit of two separate logistic regression models on the dataset of cases (labelled as $Y = 1$) and controls (labelled as $Y = 0$).

Under the null model $\mathbb{H}_0$ of additive effect (i.e. no epistasis)

$$\log\left(\frac{P[Y=1]}{P[Y=0]}\right) = \beta_0 + \beta_A S_A + \beta_B S_B$$

provides a sufficient fit, whereas under the alternate model $\mathbb{H}_1$ of deviation from additive log odds (i.e. epistasis)

$$\log\left(\frac{P[Y=1]}{P[Y=0]}\right) = \beta_0 + \beta_A S_A + \beta_B S_B + \beta_{AB} S_A S_B$$

provides a significantly better fit. When the residual deviance captured by the interaction term $\beta_{AB}$ in the alternate model is large, the likelihood ratio of $\mathbb{H}_1/\mathbb{H}_0$ suggests a significant p-value.

| $N$ Cases $(Y = 1)$ | $S_A = 0$ | $S_B = 1$ |
|---|---|---|
| $S_A = 0$ | $N_{00}$ | $N_{01}$ |
| $S_B = 1$ | $N_{10}$ | $N_{11}$ |

| $n$ Controls $(Y = 0)$ | $S_A = 0$ | $S_B = 1$ |
|---|---|---|
| $S_A = 0$ | $p_{00}$ | $p_{01}$ |
| $S_B = 1$ | $p_{10}$ | $p_{11}$ |

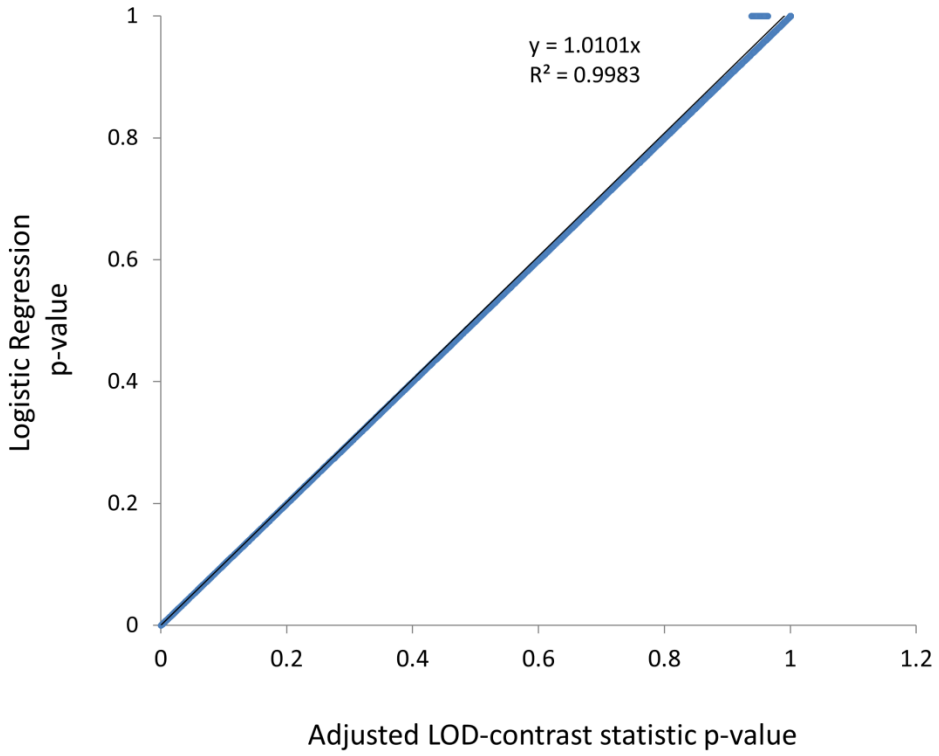
The disadvantage of this approach is that each logistic regression requires a computationally expensive numerical iteration procedure to converge to the MLE of its coefficients. Instead of performing a logistic regression, SIXPAC applies an analytical derived closed form that is faster to compute, and yet provides an equivalent result.

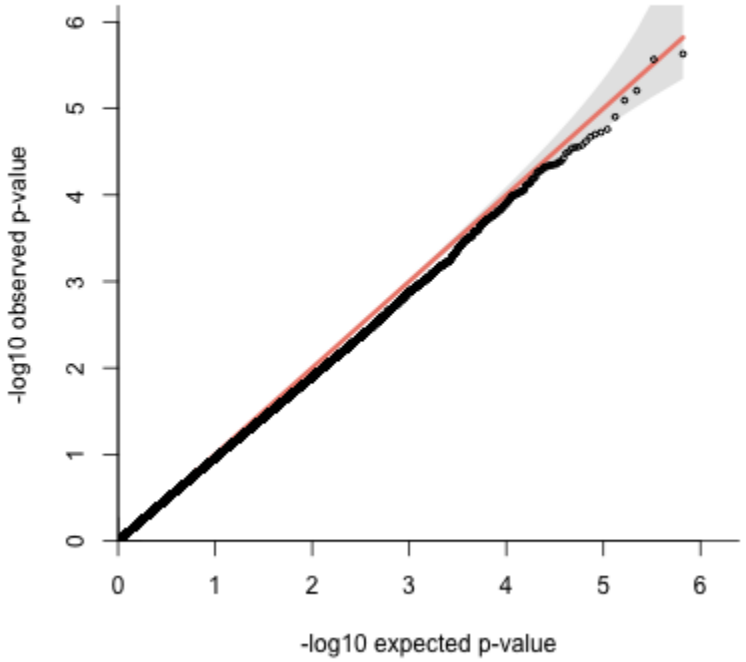
The LOD-contrast statistic (described here (Wu et al. 2010), and corrected here (Ueki and Cordell 2012)) is easy to compute and mirrors the logistic regression based likelihood-ratio statistic for statistical epistasis. Therefore, by applying the LOD-contrast statistic SIXPAC manages the same search in a few hours on a single computer – allowing us to perform the thousands of genome-wide scans required by our permutation scheme. The LOD contrast statistic is given by

$$LOD_{contrast} = \frac{\log\left(\frac{N_{11}N_{00}}{N_{10}N_{01}}\right) - \log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right)}{\left(\frac{1}{N_{11}} + \frac{1}{N_{10}} + \frac{1}{N_{01}} + \frac{1}{N_{00}}\right) + \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}\right)} \sim \chi^2_{1dof}$$

Equation 7-11

The perfect concordance of this test with logistic regression is seen below in figure, while the validity of the null hypothesis ( that the statistic is distributed as $\chi^2_{1dof}$ ) is seen from the QQ plot in the next figure.






## 7L. Functional enrichment statistics

The statistics we employ have been well studied and characterized in marginal enrichment. In particular, both Simes correction and Fisher's combination test have been proven to be the most powerful. We directly extend the insights made by others to our work. We note that though the Simes correction is a fairly conservative procedure for positively correlated tests, it tightly controls the false positive rate (Li et al. 2011). Other tests like Fisher's combination test combine information from multiple epistatic SNP-pairs which gives an intuitively more appealing measure of gene-gene epistasis, but can be too liberal when tests are correlated – as is our case (all pairwise SNPs combinations contained in a pair of genes). In other words, we can be very

confident of the results, but future extensions may need to devise more statistically powerful alternatives to find even more pathway interactions.

## 7M. Equitable distribution of sequence coverage

Equitable distribution mandates an equal coverage to all pooled individuals in order to maximize the probability of observing a rare variant and identifying its carrier. This may be seen as follows: Consider a pooling of two individuals *a* and *b*, given unequal overall coverages $\hat{C}_a > \hat{C}_b$. By Equation 4-9, we get a total number of false negatives for a site as

$$\mathbf{err}_{us}(C_a^x + C_b^x) \sim \mathbf{err}_{us}.\int_0^t c^{(\alpha_a-1)} \frac{\exp(-\hat{C}_a/\beta_a)}{\beta_a^c\Gamma(\alpha_a)}\, dc + \mathbf{err}_{us}.\int_0^t c^{(\alpha_b-1)} \frac{\exp(-\hat{C}_b/\beta_b)}{\beta_b^c\Gamma(\alpha_b)}$$

Equation 7-12

where $\alpha_a = \alpha_b = 6.3$, while $\beta_a < \beta_b$ are shape parameters by Equation 4-8.

Under equitable allocation $\hat{C} = \frac{(\hat{C}_a+\hat{C}_b)}{2}$, it may be shown that $2C^x.\mathbf{err}_{us} < \mathbf{err}_{us}(C_a^x + C_b^x)$. That is,

$$2\int_0^t c^{(\alpha_a-1)} \frac{\exp(-C_a/\beta_a)}{\beta_a^c\Gamma(\alpha_a)}\, dc < \int_0^t c^{(\alpha_a-1)} \frac{\exp(-\hat{C}_a/\beta_a)}{\beta_a^c\Gamma(\alpha_a)} + \int_0^t c^{(\alpha_b-1)} \frac{\exp(-\hat{C}_b/\beta_b)}{\beta_b^c\Gamma(\alpha_b)}$$

Equation 7-13

For example, if individual *a* has an overall coverage 8×, while individual *b* has overall coverage 4×, their independent under-sampling rates at threshold 2 are 0.3% and 7.7%, respectively. However, at a mean coverage of 6× across both, the probability of a **FN** is 1.4%.