Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease

Snehit Prabhu and Itsik Pe'er

Genome Res. published online July 5, 2012
Access the most recent version at doi:10.1101/gr.137885.112

Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease.

Snehit Prabhu[1], Itsik Pe'er[1]

[1]Department of Computer Science, Columbia University in the City of New York, NY 10027.
Correspondence: itsik@cs.columbia.edu
Keywords: Interaction, Bipolar Disorder, Epistasis, Pairwise test, GWAS

ABSTRACT

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. 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. 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.

INTRODUCTION

Genome-wide association studies (GWAS) have successfully identified hundreds of genetic markers associated with a wide range of diseases and quantitative traits (Manolio et al. 2009; Hindorff 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 tend to implicate individual genes in a particular disease or trait, which in turn highlight a single biological entity involved (Hugot et al. 2001; Saunders et al. 1993; 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 two-fold. Firstly, interaction – or statistical epistasis, as defined by Fisher (Fisher 1918) - is hypothesized to explain a part of disease heritability (Marchini et al. 2005; Evans et al. 2006). 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 the sheer size of the combinatorial space (Marchini et al. 2005), both for datasets typed on genotyping arrays ($\sim 10^6$ SNPs) and sequencing technologies ($\sim 10^7$ SNVs). Some methods address this problem by restricting the 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 (Slavin et al. 2011). Others like EPIBLASTER (Kam-Thong et al. 2010) and SHIsisEPI (Hu et al. 2010) make use of specialized hardware like multiple Graphical Processing Units (GPUs) to finish computation on genome-wide datasets in 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.

2006) 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 datasets 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 the Linkage Disequilibrium (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 work, we do not devise a new statistical test – rather, we use a simplified version of the LD-contrast test for interaction (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 either in weak-LD, no LD, or reverse-LD in controls[*].

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 employed) after correcting for the number of tests. The question we ask is this : given a statistical test, is it possible to identify all the significant SNP-pairs with high probability (power), without actually applying the test to all possible combinations genome-wide? In other words, can we design a search algorithm that accepts an arbitrary significance cut-off (as input from the user), and then finds all SNP-pairs which 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*": in this manuscript,

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

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 dataset. Our work focuses on addressing the computational issues which plague an exhaustive search for interaction, leaving issues of statistical power for a separate discussion.

The rest of this paper is structured as follows. First, we briefly review a simple LD-contrast test: which 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 (PAC) testing – which 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 cut-off) and have no element of approximation at all (i.e. 100% complete scan of the interaction space in the case-control dataset). In this paper, 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. approximately 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 results of our analysis on Bipolar disorder (BD) in the Wellcome Trust Case Control Consortium (WTCCC) dataset (Craddock 2007).

METHODS

1. 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.

2. 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 $\vec{v} = (v_1, \ldots, v_s)$, at which an individual can carry any one of 2^s unique allelic combinations. We say an individual carries allelic state $\vec{a} = (a_1, \ldots, a_s) \in \{0,1\}^s$, 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 $\vec{v} = (v_1, \ldots, v_s)$, 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 $\vec{v}$.

3. 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 ≥ 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.

4. 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 $\vec{v} = (v, v')$,

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

eq.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 Supplementary Section 1 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).

5. 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.

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

eq. 2

From eq.S1.1 (see Supplementary Section 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

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

eq. 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 the Supplementary Section 9.

6. Probably Approximately Complete (PAC) Search

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 Supplementary Section 2) – 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.

Group Sampling.

Rationale: From our observation that the LD statistic in cases, is usually more severe than LD-contrast (Supplementary Section 2), we deduce that significant interacting pairs $\vec{v}$ will show a minimum number of excess $\vec{1}$-carriers in cases: $\Delta_{\vec{v}}^{case} \geq Nz_B\sigma_{\vec{v}}^{case}$. 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 illustration 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:

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

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

eq.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

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

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

eq.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 $\vec{v} \in V(w) \times V(w)$.

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.

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 < β) while keeping the probability of a chance observation relatively low (Type I error rate < α). In other words, given fixed specificity and sensitivity constraints α and β (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:

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

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

eq.6

An illustration to visualize this technique is provided in Figure 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 Supplementary Section 4 and the logic provided by Algorithm 2.

This concludes our discussion of a *Probably Approximately Complete* search. PAC-testing offers a powerful computational framework: 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.

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. 2009)) 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 Supplementary Section 5), 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 Supplementary Section 6.

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 (Supplementary Section 7).

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 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 (Figure 2). 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[†]. 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.

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

[†]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.

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 ≤ 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 3×3 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 ($p_{corrected} = 0.0002$).

To get a better picture of the LD-contrast landscape between SNPs in this region, we conducted a wider survey of the area spanning ± 25 SNPs (upstream, downstream and including) both rs2041140 and rs10925490 (i. e. 51×51 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 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 51×51 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 presents a summary of the results along with the single most significant variable pair in the larger test area for each dataset.

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. 2006). 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 Supplementary Section 6). 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, $\ln\left(\frac{P}{1-P}\right) \sim \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$, 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 β_1, β_2 were observed to be not significant, while the epistasis term β_{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.

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.

Acknowledgements:

We are grateful to all patients who contributed data to both the bipolar disorder cohorts used by our work. We thank the Wellcome Trust Case Control Consortium (www.wtccc.co.uk) and the Genetic Information and Analysis Network (GAIN) initiative (http://www.genome.gov/19518664) for making their data publicly available to us, as well as their funding bodies, the Wellcome Trust and the National Institute for Health. This work was funded in part by the Microsoft Research PhD Fellowship awarded to SP, and NSF awards 0829882 and 0845677, NIH grant U54 CA121852-06 awarded to IP.

FIGURE LEGENDS

Figure 1: Group Sampling.

A cohort of N cases is shown on the left, where the cases outlined in red – P, Q, R and S – harbor an interacting pair of recessive variables. In other words, more cases carry the recessive-recessive combination than would be expected by chance, given the marginal frequencies of each recessive allele. By repeatedly drawing random groups of k cases (here $k = 3$), we are guaranteed to have drawn at least one group of individuals that carries both the variables in t attempts with probability $\geq (1 - \beta)$. These variables (and others) are quickly determined by a bitwise-AND operation between the group of cases. Then, all pairs of co-carried variables are enumerated and tested against the stage 1 null-hypothesis (case-only analysis). Rejected combinations are shortlisted and followed-up in stage 2 (case versus control analysis), where an interaction is identified.

Figure 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.

Figure 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).

TABLES

Table 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 (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.

Method	Type of test	Computational approach	Approx. time to process dataset[‡]	Run on specialized hardware
Plink[§]	Odds-ratio Contrast	Brute-force	Weeks	No
FastEpistasis	Linear Regression	Brute-force	Weeks	No
TEAM	Logistic Regression	Check fewer individuals	Weeks[**]	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[††]	No

[‡] All times as self-reported by authors of these tools, or extrapolated from performance metrics provided therein.
[§] Operating in the --fast-epistasis mode
[**] 10K SNPs all-pairs test reported in 1000 seconds, scaling linearly with number of SNP-pairs thereon.
[††] Time taken to find all pairs with LD-contrast p<1e-12 with >90% power, multi-threaded mode.

Table 2: Bipolar Disorder Interaction.

The upper 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 smaller tables below, 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.

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

WTCCC Cases rs2041140, r

rs10925490,	Observed (Expected)	< 2 min allele	= 2 min allele
	< 1 min allele	1617 (1601.6)	17 (32.4)
	≥ 1 min allele	214 (229.4)	20 (4.6)

WTCCC Controls rs2041140, r

rs10925490,	Observed (Expected)	< 2 min allele	= 2 min allele
	< 1 min allele	2533 (2538.3)	52 (46.7)
	≥ 1 min allele	352 (346.7)	1 (6.3)

LIST OF ALGORITHMS

Algorithm 1: Group Sampling Toy Problem.

Algorithm 1.

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.

Algorithm 2. SIXPAC

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

REFERENCES

Achlioptas P, Schölkopf B, and Borgwardt KM. 2011. Two-locus association mapping in subquadratic runtime. *The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

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.

Emily M, Mailund T, Hein J, Schauser L, and Schierup MH. 2009. 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.

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

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.

Greene CS, Sinnott-Armstrong NA, Himmelstein DS, Park PJ, Moore JH, 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.

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.

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.

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.

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.

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.

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.

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.

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.

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).

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 A, 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.

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.

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.

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/.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, and Moore JH. 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.

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.

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 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.

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).

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.

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.

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.

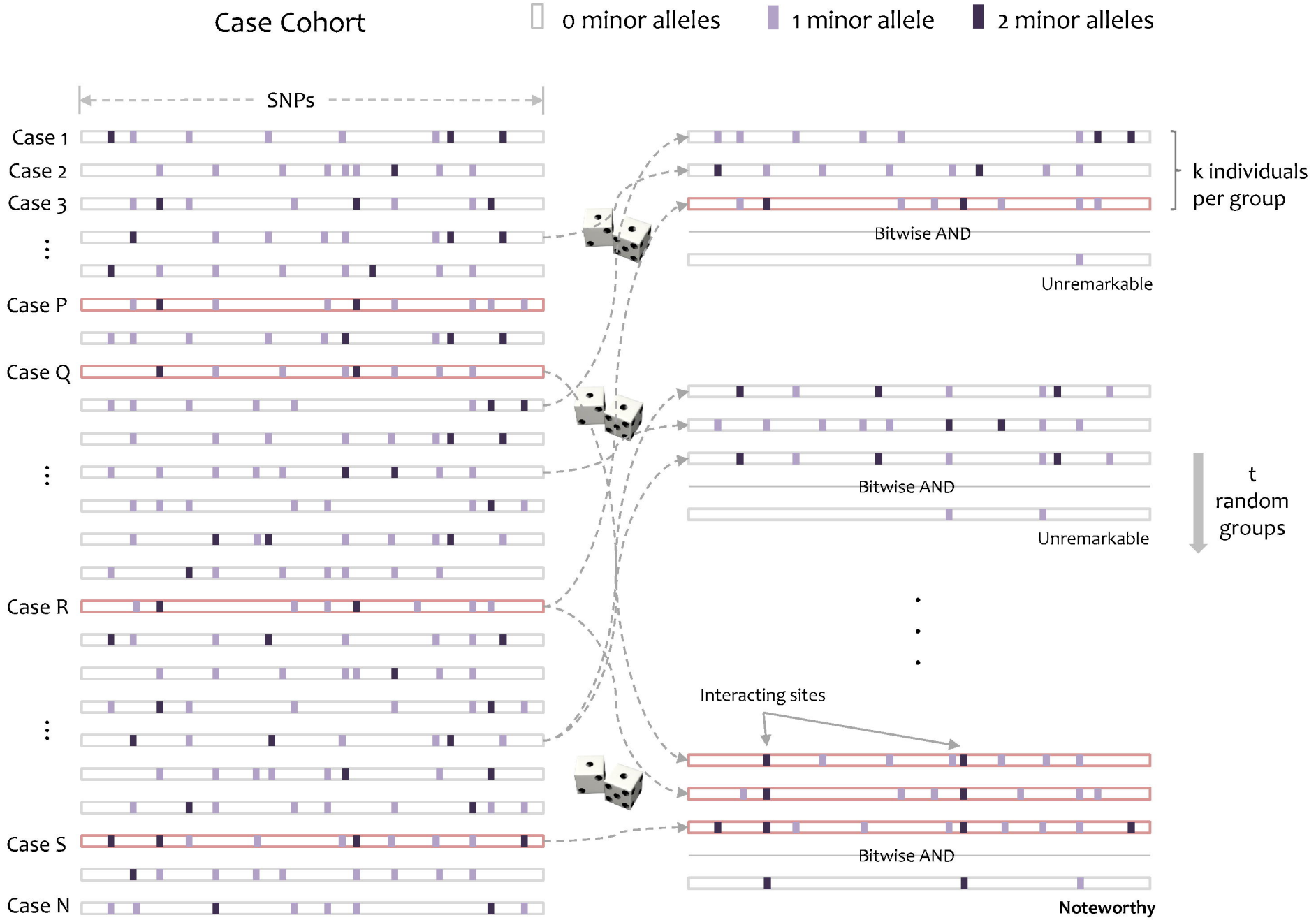
Zhang X, Zou F, and Wang W. 2009. FastChi: an efficient algorithm for analyzing gene-gene interactions. *Pacific Symposium On Biocomputing* 528-539. http://www.ncbi.nlm.nih.gov/pubmed/19209728.

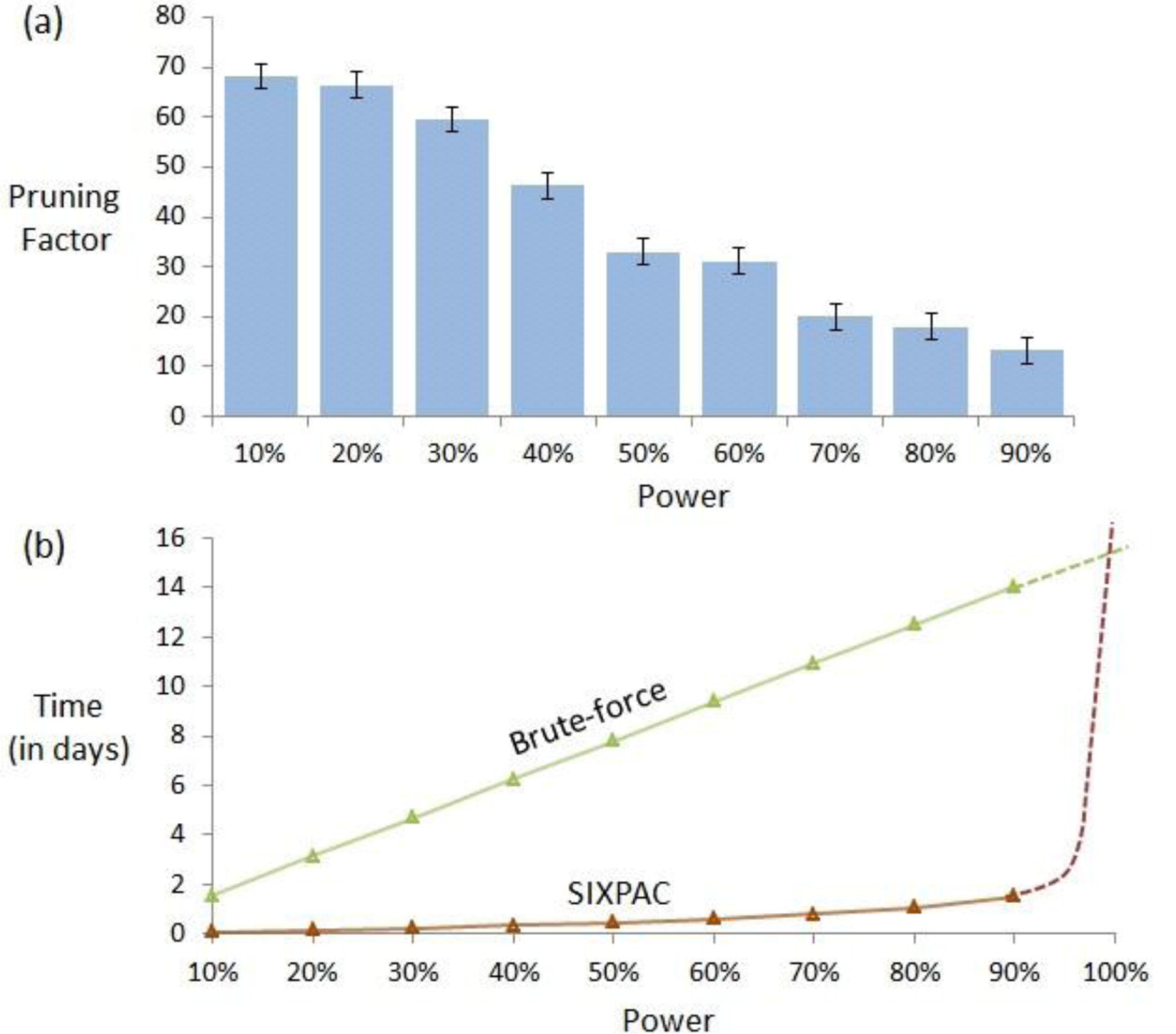
Zhang X, Zou F, and Wang W. 2008. Fastanova: an efficient algorithm for genome-wide association study. *ACM Transactions on Knowledge Discovery from Data* **3**: 821-829. http://portal.acm.org/citation.cfm?id=1401890.1401988.

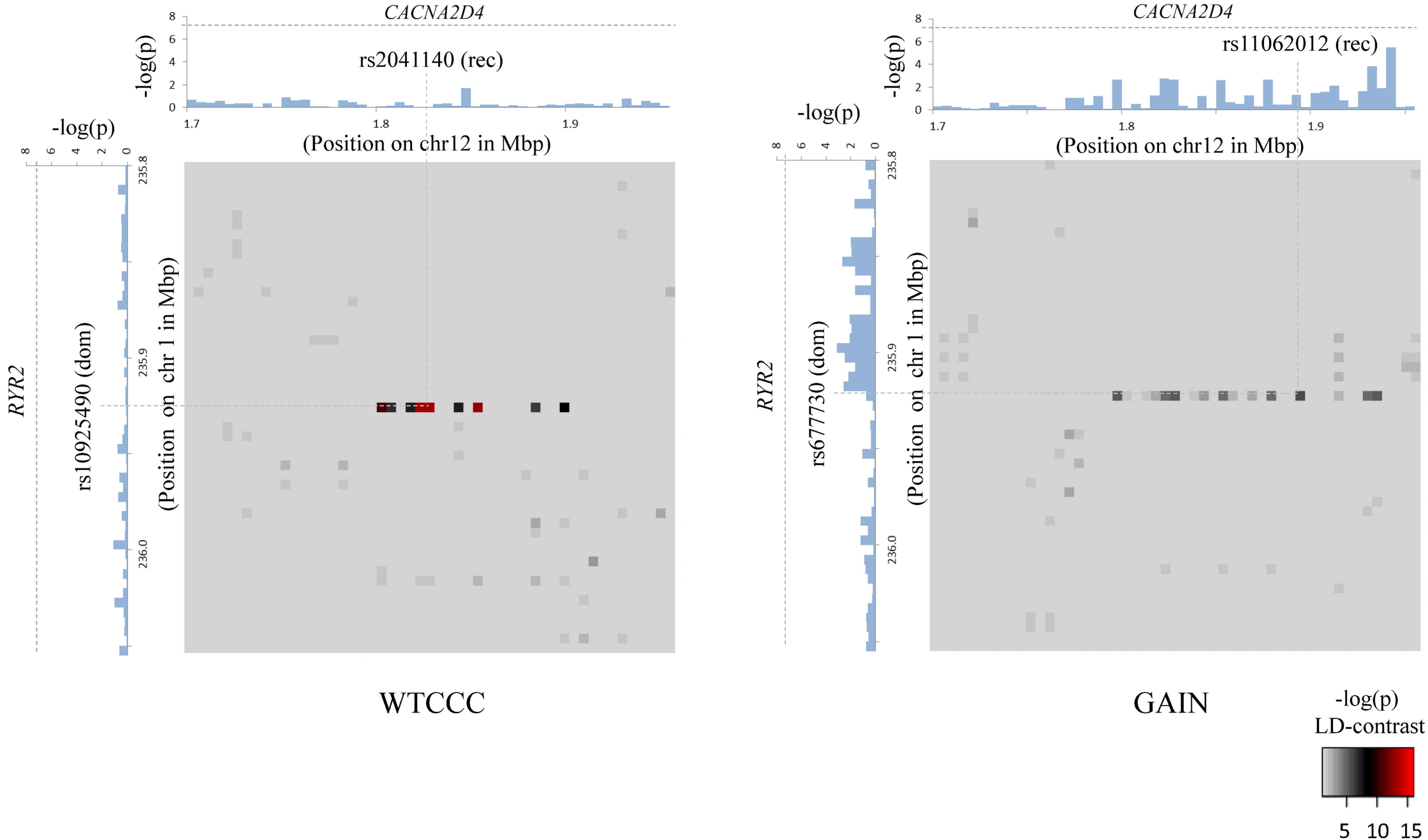
Zhang Y, and Liu JS. 2007. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39**: 1167-1173. http://www.ncbi.nlm.nih.gov/pubmed/17721534.

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*.


Case Cohort
0 minor alleles
1 minor allele
2 minor alleles
SNPs
Case 1
Case 2
Case 3
Case P
Case Q
Case R
Case S
Case N
k individuals per group
Bitwise AND
Unremarkable
Bitwise AND
Unremarkable
t random groups
Interacting sites
Bitwise AND
Noteworthy


(a)
Pruning Factor
80
70
60
50
40
30
20
10
0
10%
20%
30%
40%
50%
60%
70%
80%
90%
Power
(b)
Time (in days)
16
14
12
10
8
6
4
2
0
Brute-force
SIXPAC
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Power


CACNA2D4
rs2041140 (rec)
-log(p)
(Position on chr12 in Mbp)
RYR2
rs10925490 (dom)
(Position on chr 1 in Mbp)
WTCCC
CACNA2D4
rs11062012 (rec)
-log(p)
(Position on chr12 in Mbp)
RYR2
rs677730 (dom)
(Position on chr 1 in Mbp)
GAIN
-log(p)
LD-contrast
5 10 15