UCLA

UCLA Previously Published Works

Title

Efficiently Identifying Significant Associations in Genome-wide Association Studies

Permalink

https://escholarship.org/uc/item/25b9w397

Journal

Journal of Computational Biology, 20(10)

ISSN

1066-5277

Authors

Kostem, Emrah
Eskin, Eleazar

Publication Date

2013-10-01

DOI

10.1089/cmb.2013.0087

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number 10, 2013
© Mary Ann Liebert, Inc.
Pp. 817–830
DOI: 10.1089/cmb.2013.0087

Efficiently Identifying Significant Associations in Genome-wide Association Studies

EMRAH KOSTEM[1] and ELEAZAR ESKIN[1,2]

ABSTRACT

Over the past several years, genome-wide association studies (GWAS) have implicated hundreds of genes in common disease. More recently, the GWAS approach has been utilized to identify regions of the genome that harbor variation affecting gene expression or expression quantitative trait loci (eQTLs). Unlike GWAS applied to clinical traits, where only a handful of phenotypes are analyzed per study, in eQTL studies, tens of thousands of gene expression levels are measured, and the GWAS approach is applied to each gene expression level. This leads to computing billions of statistical tests and requires substantial computational resources, particularly when applying novel statistical methods such as mixed models. We introduce a novel two-stage testing procedure that identifies all of the significant associations more efficiently than testing all the single nucleotide polymorphisms (SNPs). In the first stage, a small number of informative SNPs, or proxies, across the genome are tested. Based on their observed associations, our approach locates the regions that may contain significant SNPs and only tests additional SNPs from those regions. We show through simulations and analysis of real GWAS datasets that the proposed two-stage procedure increases the computational speed by a factor of 10. Additionally, efficient implementation of our software increases the computational speed relative to the state-of-the-art testing approaches by a factor of 75.

Key words: genetics, genomics, haplotypes, machine learning, statistical models.

1. INTRODUCTION

Research in complex diseases has progressed rapidly in the last decade with the advent of genomic technologies (Devlin and Risch, 1995; Risch and Merikangas, 1996; International HapMap Consortium, 2005; Hardy and Singleton, 2009). In genome-wide association studies (GWAS), information on millions of single nucleotide polymorphisms (SNPs) across the genome is collected from thousands of case and control individuals. Typically, each SNP is statistically tested for disease association by comparing the minor allele frequency (MAF) between the cases and controls. The significant associations are used to gain insight into the genetic basis of disease, and hundreds of GWASs have been performed on dozens of complex diseases and successfully discovered many novel loci involved in disease susceptibility (Hindorff et al., 2009).

[1]Computer Science Department and [2]Department of Human Genetics, University of California, Los Angeles, California.

More recently, there has been great interest in applying the GWAS approach to genomic data such as gene expression. In these studies, the goal is to identify regions of the genome harboring genetic variation that affect gene expression levels or expression quantitative trait loci (eQTL) (Bochner, 2003; Rockman and Kruglyak, 2006; Cookson et al., 2009). A challenge in applying GWAS to genomic data is that these technologies typically obtain tens of thousands of measurements for each sample resulting in a tremendous computational burden when performing the analysis, including computing billions of tests, and requires substantial computational resources. This challenge is compounded for novel statistical approaches such as linear mixed models, which account for population structure (Kang et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012), yet themselves are computationally intensive.

eQTL studies are already very popular (Brem et al., 2002; Brem and Kruglyak, 2005; Keurentjes et al., 2007) and with rapidly decreasing costs of RNA-seq technologies (Wang et al., 2009; Majewski and Pastinen, 2011) will likely become more popular in the future. These include several major efforts collecting expression from multiple-tissues in humans (Cheung et al., 2005; Stranger et al., 2007; Emilsson et al., 2008; Spielman et al., 2007; Baker, 2012) and mice (Chesler et al., 2005; Bystrykh et al., 2005). More broadly, application of the GWAS approach to phenotypes measured by other genomic technologies such as those reported by the ENCODE consortium (The ENCODE Project Consortium, 2004, 2007, 2011, 2012) will face similar computational challenges.

In this article, we introduce a novel two-stage method that can be applied to reduce the computational burden of a wide range of association studies including those that employ case-control, quantitative trait, and mixed-model statistical testing methodologies. In each trait, typically only a small percentage of the SNPs are significantly associated and the SNPs neighboring a significant association have elevated statistics. Intuitively, one can first test an informative subset of the SNPs, termed proxy SNPs, across the genome to quickly locate these regions and test the SNPs therein. This way, many of the regions with no associations can be discarded from the analysis to reduce the computational burden.

Our novel method for genome-wide rapid association testing (GRAT) guarantees to identify all of the significant associations with high probability while reducing the total number of tests. The proposed method chooses the proxy SNPs and determines which additional SNPs to test based on the observed proxy SNP statistics and the patterns of linkage disequilibrium (LD) in the region. The key insight underlying GRAT is that by taking advantage of how the statistics at SNPs in LD with each other behave, we can estimate the probability that an untested SNP has a significant association and use this probability to only eliminate SNPs from consideration if they are highly unlikely to have significant associations. We have selected a set of proxy SNPs for the 1000 Genomes Project, and any study that imputes to the 1000 Genomes Project SNPs can readily use our approach. We also provide our method for choosing proxy SNPs, which can be applied to any reference dataset. We show through simulations and analysis of real eQTL datasets that the proposed two-stage procedure identifies the significant associations while only testing approximately 10% of the SNPs. GRAT's efficient software implementation reduces the computational time for computing large-scale association studies by a factor of 30 compared to currently used state of the art methods. When our method is applied to association studies that utilize linear mixed models, the speed-up is cumulative with recent efforts that decrease the computational burden of computing the actual association statistic such as EMMAX, FaST-LMM, and GEMMA (Kang et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012).

2. METHODS

2.1. *Genome-wide association studies*

Old version: For the simplicity of description, we consider a balanced case-control genome-wide association study (GWAS) with $N/2$ individuals (N copies of each chromosome) per panel. For our actual experiments, we will use association statistics for quantitative phenotypes, but the approach assuming case-control phenotypes is equivalent. For SNP m_i, p_i denotes its population minor allele frequency (MAF); p_i^+ and p_i^- denote its population case and control MAFs; $\hat{p}_i^+$ and $\hat{p}_i^-$ denote its observed case and control MAFs in the GWAS. Given the relative risk of the SNP, γ_i, in the disease and the prevalence of the disease, F, in the population, it can be shown that the case and control MAFs of the SNP follows,

$$p_i^+ = \frac{\gamma_i p_i}{(1-\gamma_i)p_i + 1}, \quad p_i^- = \frac{p_i - F p_i^+}{1-F}. \tag{1}$$

An SNP is defined as *not* associated if $p_i^+ = p_i^-$.

In case-control GWASs the following statistic is widely used, which is normally distributed for large N with mean $\lambda_i\sqrt{N}$ (the noncentrality parameter), and unit variance,

$$\begin{gathered} S_i = \hat{s}_i = \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{2\hat{p}_i(1-\hat{p}_i)}}\sqrt{N} \sim \mathcal{N}(\lambda_i\sqrt{N}, 1), \\ \text{where } \lambda_i = \frac{p_i^+ - p_i^-}{\sqrt{2p_i(1-p_i)}} \text{ and } \hat{p}_i = \frac{\hat{p}_i^+ + \hat{p}_i^-}{2}. \end{gathered} \tag{2}$$

Given the significance level α and the observed value of the test statistic $\hat{s}_t$, the SNP is deemed as significant, or statistically associated, if $|\hat{s}_i| > \Phi^{-1}(1 - \frac{\alpha}{2})$, where $\Phi^{-1}(.)$ is the quantile function of the standard normal distribution. For simplicity, we use the notation: $t_\alpha \equiv \Phi^{-1}(1 - \frac{\alpha}{2})$. Typically, in a GWAS, the significance level is chosen as $\alpha = 10^{-8}$.

2.2. *A two-stage approach for identifying the significant associations*

We propose the following two-stage testing procedure for dentifying the significant associations within a set of SNPs $\mathcal{M}$. Given a subset of the SNPs $\mathcal{T} \subset \mathcal{M}$, referred to as the proxy SNPs, for each proxy SNP, $m_t \in \mathcal{T}$, its association statistic, $\hat{s}_t$, is computed. In the second stage, a decision rule is exercised for each of the remainder SNP, $m_i \in \mathcal{M}\backslash\mathcal{T}$, in order to determine whether or not to compute the association statistic of the remainder SNP. The decision rule for a remainder SNP m_i is defined using a proxy SNP, $m_t \in \mathcal{T}$, and a threshold, s_t^*, for its observed statistic $\hat{s}_t$. If the observed statistic of the proxy SNP is more extreme than the threshold value, $\hat{s}_t > s_t^*$, the remainder SNP is tested.

2.3. *Performance of the two-stage approach*

In a GWAS, the performance of the two-stage approach can be summarized by the total number of SNPs tested (NT), and the percentage of the significant SNPs identified, or the recall rate (RR). The total number of tests is the sum of the tests performed on the proxy SNPs, plus the remainder SNPs that are tested as a result of the decision rules. We use a standard GWAS simulation model (Kostem et al., 2011) to evaluate a given set of proxy SNPs and decision rules based on their *expected* performance within the simulated data.

The simulation model considers the probability of each SNP being causal, c_i, and the noncentrality parameter (NCP) of the causal SNP, $\lambda_c\sqrt{N}$. For simplicity, we give a brief explanation of the simulation procedure for a single causal SNP using a genomic reference dataset such as the HapMap. Using the given probabilities of each SNP being causal, at most a single causal SNP is randomly selected. Given the disease prevalence F and the NCP of the causal SNP $\lambda_c\sqrt{N}$, the case and control MAFs, p_c^+ and p_c^- are determined. Next, the HapMap haplotypes are divided into two pools according to the minor and major allele of the causal SNP, and case-control panels are sampled using p_c^+ and p_c^-.

For each simulation dataset, each association statistic is computed to identify which SNPs are significant in the dataset. We then apply the two-stage method to observe the NT and RR. The expected recall rate (ERR) and the expected number of SNPs to be tested (ENT) then can be computed by repeatedly simulating datasets, applying the two-stage approach and averaging the observed NT and RR value.

2.4. *Finding the optimal decision rules for given proxy SNPs*

For a given set of proxy SNPs, one can determine the decision rules empirically by evaluating the performance of using different threshold values on the remainder SNPs in the simulated data. The empirical approach can be cumbersome, and instead we derive an analytical framework for estimating the expected performance, which eliminates the need for generating simulated data and saves time. Furthermore, using this analytical framework we show how to determine the optimal decision rules for the remainder SNPs given a set of proxy SNPs.

An SNP that is disease-associated can be either causal in the disease or in LD with the causal SNP. Given that SNP m_i is the causal SNP, the noncentrality parameter (NCP) of a correlated SNP m_t, $\lambda_t\sqrt{N}$, is proportional to the NCP of the causal SNP, $\lambda_c\sqrt{N}$, by their correlation coefficient, r, where $\lambda_t = r\lambda_c$. It can be shown that the joint distribution of the association statistics of the causal SNP m_i and the noncausal SNP m_t follows a bivariate normal distribution (Han et al., 2009). In addition to case-control studies, these principles can also be applied to quantitative traits (Schaid et al., 2002).

We follow a conservative approach in which each remainder SNP m_i is paired with the proxy SNP that is most strongly correlated, referred to as the *best*-proxy and denoted by $m_{b(i)}$. For each remainder SNP m_i, we denote the association statistic of its best-proxy $m_{b(i)}$ with $s_{b(i)}$ and test SNP m_i if its best-proxy SNP association statistic is more extreme than a given threshold, $s_{b(i)} > s^*_{b(i)}$. For simplicity, we assume only the remainder SNP can be causal and express the density function of the joint distribution, $f(s_i, s_{b(i)})$,

$$f(s_i, s_{b(i)}) = c_i\phi\left(\begin{bmatrix} s_i \\ s_{b(i)} \end{bmatrix}; \begin{bmatrix} \lambda_c\sqrt{N} \\ r\lambda_c\sqrt{N} \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right) + (1-c_i)\phi\left(\begin{bmatrix} s_i \\ s_{b(i)} \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right), \tag{3}$$

where $\phi(\boldsymbol{x}; \mu, \Sigma)$ denotes the density of a multivariate normal distribution with mean vector μ and covariance matrix Σ. The first term corresponds to having the remainder SNP as causal, with probability c_i, and the second term to not causal with probability $1 - c_i$.

Assume we are given K proxy SNPs, where $\mathcal{T} = \{m_1, \ldots, m_K\}$. The ENT can be expressed as the fixed cost of testing K proxy SNPs, plus the expected number of decision rules that are triggered,

$$\text{ENT}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) = K + \sum_{i=K+1}^{M} \Pr(|S_{b(i)}| > s^*_{b(i)}). \tag{4}$$

We *approximate* the ERR as the ratio of the expected number of significant SNPs that the two-stage approach discovers, to the expected number of significant SNPs in a GWAS,

$$\text{ERR}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) = \frac{\sum_{t=1}^{K} \Pr(|S_t| > t_\alpha) + \sum_{i=K+1}^{M} \Pr(|S_i| > t_\alpha, |S_{b(i)}| > s^*_{b(i)})}{\sum_{i=1}^{M} \Pr(|S_i| > t_\alpha)}, \tag{5}$$

where the first and the second terms in the numerator correspond to the expected number of significant SNPs obtained from testing the proxy SNPs and the remainder SNPs, respectively. Further, we refer to the second term as the expected recall function, which can be computed using the joint distribution,

$$\begin{aligned} \text{ER}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) &= \sum_{i=K+1}^{M} \Pr(|S_i| > t_\alpha, |S_{b(i)}| > s^*_{b(i)}), \\ \Pr(|S_i| > t_\alpha, |S_{b(i)}| > s^*_{b(i)}) &= \iint_{\Omega_i} f(s_i, s_{b(i)}) ds_i \, ds_{b(i)}, \end{aligned} \tag{6}$$

where $\Omega_i = \{(s_i, s_{b(i)}) \mid |s_i| > t_\alpha \text{ , } |s_{b(i)}| > s^*_{b(i)}\}$.

We are interested in determining the decision rules that lead to the lowest ENT, while the expected recall rate (ERR) satisfies a given target value, ρ, which can be expressed as an optimization problem,

$$\begin{aligned} &\text{minimize} \quad \text{ENT}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}), \\ &\text{such that} \quad \text{ERR}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) = \rho. \end{aligned} \tag{7}$$

We show the problem is convex and outline an efficient iterative solution in the Appendix.

2.5. Choosing the optimal proxy SNPs

The expected number of SNPs to be tested (ENT) in the two-stage approach depends on the number of proxy SNPs and which SNPs are chosen as proxies. It can be shown that the problem of finding the optimal set of proxy SNPs, among all possible sets of proxy SNPs, the set that gives the minimum ENT, is an NP-Hard problem (Bafna et al., 2003). Therefore, we propose a heuristic algorithm for choosing the proxy SNPs using a greedy approach, which incrementally builds the set of proxy SNPs.

Starting with an empty set, let $\mathcal{T}_k$ denote the current set of proxy SNPs with size k, where ENT_k and ERR_k denote the values of its ENT and ERR. ($\mathrm{ENT}_0 = +\infty$ and $\mathrm{ERR}_0 = -\infty$). Each remainder SNP m_i is a candidate to extend the current set of proxy SNPs to become $\{\mathcal{T}_k \cup m_i\}$, which performs $\mathrm{ENT}^{(i)}_{k+1}$. The remainder SNP with the least $\mathrm{ENT}^{(i)}_{k+1}$ is chosen for extending the current set of proxy SNPs:

$$\mathcal{T}_{k+1} = \mathcal{T}_k \cup \underset{m_i \in \mathcal{M} \setminus \mathcal{T}_k}{\operatorname{argmin}} \left(\mathrm{ENT}^{(i)}_{k+1} \right). \tag{8}$$

While the extended set $\mathcal{T}_{k+1}$ improves the ENT, that is, $\mathrm{ENT}_{k+1} < \mathrm{ENT}_k$, the algorithm continues.

For each candidate set of proxy SNPs, the algorithm solves the optimization problem (7) to compute $(\mathrm{ENT}^{(i)}_{k+1}$. This leads to a quadratic computational complexity in the order of the number of the collected SNPs and in practice makes it hard to scale to large numbers. We further introduce a heuristic extension to the above greedy approach to reduce this complexity. While extending the current set of proxy SNPs $\mathcal{T}_k$ to $\mathcal{T}_{k+1}$, the optimization problem (7) is solved $M - k$ times. In particular, solving the optimization problem (7) corresponds to finding the gradient, g^*, at which the ENT function is minimized while satisfying the constraints (see Appendix). We assume that for $\mathcal{T}_k$ and $\mathcal{T}_{k+1}$ the gradient values of their ENT functions are close enough, $g^*_k \approx g^*_{k+1}$. Therefore, while extending the current proxy set, we compute the ENT of each candidate set, $\mathrm{ENT}^{(i)}_{k+1}$, using the gradient value from the previous step, g^*_k. This way, rather than solving the optimization problem $M - k$ times for each possible proxy SNP at each step k, the gradient is updated once after the new set $\mathcal{T}_{k+1}$ is determined. Using this approach, the optimization problem (7) is solved a total of K times, where K is the size of the final set of proxy SNPs.

2.6. *Updating the remainder SNP thresholds in linear mixed models*

We consider the following linear mixed model (LMM) formulation,

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{g} + \boldsymbol{e}, \tag{9}$$

where $\boldsymbol{y}$ is the $(n \times 1)$ vector of phenotypic values, $\boldsymbol{X}$ is the $(n \times p)$ matrix of fixed effects, which includes the mean, covariates, and the SNP to be tested, β is the $(p \times 1)$ vector of fixed effect weights, $\boldsymbol{g}$ is the variance component accounting for the population structure, and $\boldsymbol{e}$ is the iid noise. We assume the random effects, $\boldsymbol{g}$ and $\boldsymbol{e}$, follow multivariate normal distribution, $\boldsymbol{g} \sim \mathcal{N}(0, \sigma^2_g \boldsymbol{K})$, $\mathbf{e} \sim \mathcal{N}(0, \sigma^2_e \boldsymbol{I})$, where $\boldsymbol{K}$ is the known $(n \times n)$ genetic similarity matrix, $\boldsymbol{I}$ is the $(n \times n)$ identity matrix with unknown magnitudes σ^2_g and σ^2_e. We follow the approach taken in EMMAX (Kang et al., 2010) and estimate σ^2_g and σ^2_e in the null model, with no SNP effect, and use these parameters while testing the SNPs. That is, when each SNP is tested, the covariance of $\boldsymbol{y}$ is kept fixed, $\mathrm{Cov}(\boldsymbol{y}) = \Sigma = \hat{\sigma}^2_g \boldsymbol{K} + \hat{\sigma}^2_e \boldsymbol{I}$, where $\hat{\sigma}^2_g$ and $\hat{\sigma}^2_e$ are the restricted log likelihood (REML) estimates (Kang et al., 2010; Lippert et al., 2011).

In GRAT, the threshold value for each remainder SNP is computed after the covariance matrix Σ is estimated, and the alternate model is transformed by the inverse square root of this matrix,

$$\Sigma^{-1/2}\boldsymbol{y} \sim \mathcal{N}(\Sigma^{-1/2}\boldsymbol{X}\beta,\ \sigma^2 \boldsymbol{I}), \tag{10}$$

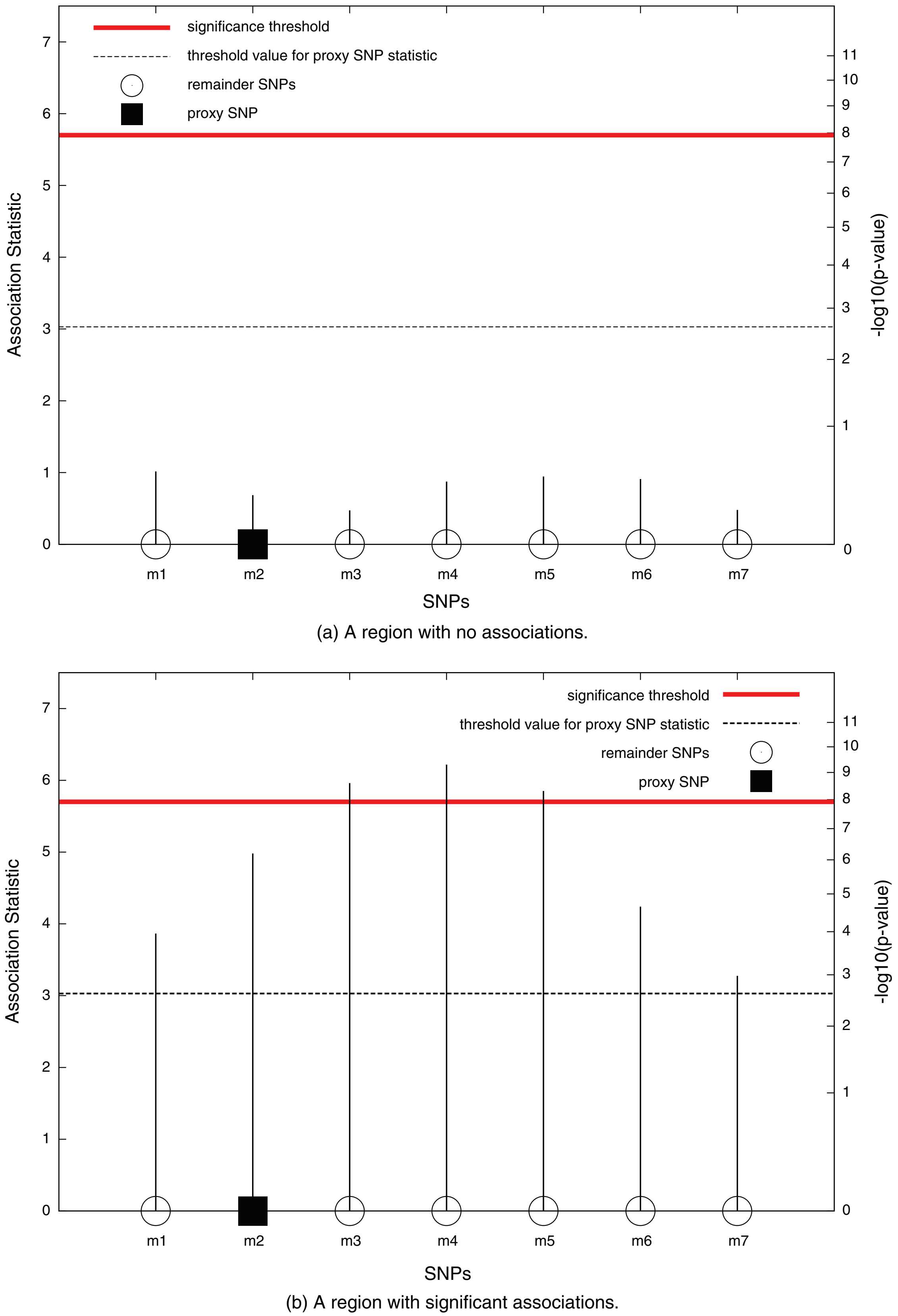
where the residuals are iid. For two SNPs m_i and m_j, let $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ be their $(n \times 1)$ allelic indicator vectors. When the SNPs are tested individually in the above model, the same transformation is applied to the genotype vectors, which may moderately change the pairwise correlation between the SNPs. The transformed genotype vectors are $\tilde{\boldsymbol{x}}_i = \Sigma^{-1/2}\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}_j = \Sigma^{-1/2}\boldsymbol{x}_j$, and their correlation coefficient is,

$$\tilde{r}_{ij} = \frac{\mathrm{Cov}(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)}{\sqrt{\mathrm{Var}(\tilde{\boldsymbol{x}}_i)}\sqrt{\mathrm{Var}(\tilde{\boldsymbol{x}}_j)}}. \tag{11}$$

3. RESULTS

3.1. *Genome-wide rapid association testing (GRAT)*

In Figure 1, we consider two possible scenarios for a genomic region in a GWAS. In (a), the region contains no significant associations, and in (b), the region contains a causal SNP. In (a) and (b), the statistics for each SNP are shown, denoting what could have been observed in each scenario had all the SNPs in the



(a) A region with no associations.

(b) A region with significant associations.

FIG. 1. An example of applying GRAT in two hypothetical regions. First, the proxy SNP (rectangle) is tested and its statistics are compared to the threshold (dashed line). If the statistic is above the threshold, the remaining SNPs in the region are tested. SNP, single nucleotide polymorphism.

region been tested. Let m_2 be the proxy SNP for this region to decide whether or not to test the rest of the SNPs. We refer to the SNPs other than the proxy SNP (m_1, m_3, m_4, m_5, m_6, and m_7) as the ''remainder SNPs.'' If the observed statistic of the proxy SNP is stronger than a threshold value, which in this example is 3.0, the remainder SNPs are tested.

In the first stage, only the proxy SNP is tested, and its association statistic is observed. In (a), where the region contains no associations, the statistic of the proxy SNP is 0.7. The observed statistic of the proxy is less than the threshold value ($0.7 < 3.0$), and hence none of the remainder SNPs within the region are tested. In (b), the region contains associations and the proxy SNP captures this information. The observed statistic of the proxy SNP is stronger than the threshold value ($5.0 > 3.0$), which leads to testing each of the remainder SNPs in the region. This results in identifying all the significant SNPs (m_3, m_4, and m_5).

In the Methods section, we introduce a novel approach for choosing the proxy SNPs and the threshold values, which provide guarantees that all statistically significant associations will be discovered while computing the least amount of association tests. Due to the complexity of linkage disequilibrium (LD) across the genome, we use a separate threshold value for each remainder SNP rather than using a common threshold value for all the remainders SNPs in an LD region. This is performed by pairing each remainder SNP with its most strongly correlated proxy SNP, and a threshold value is used for the pair to decide whether or not to test the remainder SNP. We have precomputed the proxy SNPs for the 1000 Genomes Project, and studies imputing to SNPs in this reference can benefit from our method. Even though the LD structure among the SNPs in the study and the reference dataset may be different, our method guarantees to discover all significant associations with high probability. This is achieved by updating the threshold values using the LD structure observed in the study. We term our novel two-stage testing procedure as genome-wide rapid association testing (GRAT).

GRAT can be applied to a wide range of statistical models, such as case-control studies, quantitative traits, and LMM. In particular, the LMM approach has recently become popular due to its effective control of population structure. Computing the LMM association statistic is computationally expensive, and recently, its efficient computation has attracted great interest (Kang et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012). The speed-up due to GRAT is cumulative with these efforts.

3.2. Application of a large-scale eQTL study

We compared the performance of GRAT to the standard approach of testing all the SNPs using a large-scale eQTL study (Stranger et al., 2012) that contains 47, 292 gene expression traits on 80 HapMap ASN (East Asian ancestry) individuals that are fully sequenced in the 1000 Genomes Project. We obtained the genotype data from the MACH web site (Li et al., 2010) and retained approximately 5.9 million SNPs that are filtered for Hardy-Weinberg equilibrium (HWE) and minor allele frequency (MAF) greater than 5%. We eliminated SNPs with lower MAF frequency since they could not be genome-wide significant due to the sample size.

We performed the standard analysis using PLINK (Purcell et al., 2007), which took approximately 2600 hours. We used a conservative genome-wide significance threshold level, $\alpha = 10^{-8}$, to label the significant SNPs and observed 85,219 significant associations. We repeated the association analysis by applying GRAT using the proxy SNPs precomputed for the 1000 Genomes Project ASN population SNPs. The number of proxies is 276,702, which means GRAT tests approximately 5% of the SNPs in the first stage.

Applying GRAT to the whole eQTL dataset took 35 hours using the same computational resources (single core of an Opteron CPU). In addition to the proxies, GRAT tested 8.5% of the SNPs in the second stage, reducing the computational cost down to analyzing 13.5% of all the SNPs with the rest of the speedup coming from a faster implementation compared to PLINK. GRAT identified all of the significant associations and speeded up the computation by a factor of 75.

3.3. GRAT applied to linear mixed model association

We applied GRAT to a linear mixed model (LMM) association of the eQTL dataset. A challenge in applying GRAT to LMMs is that GRAT utilizes the fact that the joint distribution of traditional association statistics for correlated markers is directly dependent on the correlation between the markers as shown in Pritchard and Przeworski (2001). Unfortunately, when applying LMMs, this relation no longer holds. We derive an analogous relationship between LMM statistics that takes into account both the correlation between the markers and the kinship matrix. Utilizing this relationship, we apply GRAT to LMMs using an efficient implementation (Lippert et al., 2011).

We performed the standard analysis, testing each SNP in each expression trait, which identified 66,818 significant associations ($\alpha = 10^{-8}$). We applied GRAT using the proxy SNPs precomputed for the 1000 Genomes Project ASN population. In two stages, GRAT statistically tested a total of 9.1% of the SNPs, identifying all of the significant associations and demonstrating that GRAT can speed up LMM association by a factor of 10.

3.4. Simulations using the 1000 Genomes Project

To obtain a more robust estimate of the performance, we applied GRAT to thousands of simulated GWAS studies. We simulated the studies using common SNPs (minor allele frequency >5%) available from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) using the phased SNP genotypes obtained from the MACH web site (Li et al., 2010) on four populations: African (AFR), East Asian (ASN), admixed American (AMR), and European (EUR) ancestries.

We divided each chromosome into panels of 1000 SNPs and simulated case-control GWASs by randomly selecting 5% of the panels as alternates, in which we simulated a causal SNP, and the remaining panels as the null panels, without any causal SNPs. In each alternate panel, we randomly selected the causal SNP and set its statistical power to be $\mathcal{P}_c = 50\%$ at the significance level $\alpha = 10^{-8}$. Using this procedure, we simulated 500 GWASs in each population.

We applied GRAT to each simulated GWAS and recorded the recall rate of the significant SNPs and total number of tests performed. In Table 1, we show the performance of GRAT in each population averaged over the simulations. GRAT practically identified all significant associations and reduced the number of tests by 10-fold. Across the simulations, from the total 3,718,126 significant associations, GRAT only missed 1052 significant associations.

3.5. Comparison to traditional tag-SNP–based association testing

Choosing an informative subset of SNPs, termed tag-SNPs, under various criteria has been extensively investigated (Stram, 2004; de Bakker et al., 2005; Stram, 2005; Cousin et al., 2003, 2006; Halperin et al., 2005; Lin and Altman, 2004; Pardi et al., 2005; Qin et al., 2006; Saccone et al., 2006; Carlson et al., 2004; Santana et al., 2010). The main goal of these methods is to reduce the cost of GWASs by genotyping a subset of the SNPs yet collect as much information as possible on the remaining SNPs.

We mimic a two-stage association testing approach using a traditional tag-SNP selection method and compare its performance to GRAT. In the first stage, we test all the tag-SNPs and use a *p*-value threshold, α_{tag}, to choose which of the tag-SNPs to follow. If the *p*-value of a tag-SNP is stronger than the threshold, the remainder SNPs tagged by this tag-SNP are tested.

We simulated association studies using the 10 HapMap ENCODE regions, which are densely genotyped for four HapMap populations (The ENCODE Project Consortium, 2004). In each simulation study, we used the ENCODE regions to generate null regions that harbor no causal SNPs and alternate regions each harboring a causal SNP with 50% statistical power at the genome-wide significance level of $\alpha = 10^{-8}$. Following this approach, we generated 500 association studies in each population.

In each region and in each population, we identified the tag-SNPs with the widely utilized tag-SNP selection method Tagger (de Bakker et al., 2005). Given a set of SNPs and information on their minor allele

Table 1. Performance in Simulations

Population	*Number of SNPs*	*Recall Rate*	*Reduction*
AFR	8.5×10^6	>99.9%	88.2%
AMR	6.7×10^6	>99.9%	92.4%
ASN	6.1×10^6	>99.9%	92.8%
EUR	6.6×10^6	>99.9%	92.6%

The average performance of GRAT in 500 simulated GWASs using 1000 Genomes Project data in four populations. GRAT identified practically all significant associations by only testing 10% of the SNPs.

GRAT, genome-wide rapid association testing; GWAS, genome-wide association studies; SNPs, single nucleotide polymorphisms.

Table 2. The Average Number of Tagging SNPs Chosen by Each Method and the Total Number of SNPs in Each Population

Population	*GRAT*	*Tagger*	*Number of SNPs*
CEU	55.0	234.0	1138.6
CHB	45.7	201.2	1024.3
JPT	42.7	198.7	1058.5
YRI	89.1	486.3	1410.1

In each HapMap population, the average number of tagging SNPs chosen by GRAT and Tagger are shown. The fourth column indicates the average number of SNPs across the ENCODE regions. On average, Tagger chooses 4.7 times more tagging SNPs than GRAT.

frequencies and pairwise correlation coefficients, Tagger selects the minimum number of tag-SNPs such that each of the remaining SNPs correlates to a tag-SNP with a minimum r^2 pairwise correlation value. In our evaluations, we have used the default value of $r^2 = 0.8$. In order to perform a comparison, we also applied GRAT to identify the proxy SNPs and the statistic threshold rules for testing the remainder SNPs to achieve 99% target recall rate on the significant associations. The number of tagging SNPs chosen by Tagger and the number of proxy SNPs chosen by GRAT are summarized in Table 3. On average, Tagger chose more than 4 times the number of proxy SNPs chosen by GRAT.

In Table 3, the performance of GRAT is compared to Tagger in four HapMap populations using various p-value threshold values, $\alpha_{\text{tag}} = \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$. In each population, GRAT achieved more than 99% recall rate while testing approximately 10% of all SNPs. Among all the p-value threshold values used, the traditional tag-SNPs led to testing more than twice the number of SNPs tested by GRAT and only achieved the target recall rate in all populations when the p-value threshold value was $\alpha_{\text{tag}} = 10^{-5}$. Unfortunately, Tagger, unlike GRAT, does not guarantee a recall rate, so it is not clear how to set the threshold and be certain that no associations are missed. Because Tagger selects the tagging SNPs to maintain a particular correlation between the tagging and the nontagging SNPs and using a uniform threshold value, α_{tag}, to choose which SNPs to test; this does not guarantee any sensitivity on the discovery of the significantly associated SNPs.

4. DISCUSSION

In the genome-wide association study (GWAS), information on SNPs across the genome is collected from thousands of case and control individuals. Typically, each SNP is tested individually for disease association, and

Table 3. Performance of GRAT and Tagger in ENCODE Simulations

		Reduction			*Reduction*	
Method	*Recall*	*CEU*	*Speedup*	*Recall*	*CHB*	*Speedup*
GRAT	99.89%	89.7%	9.7×	99.73%	89.6%	9.6×
Tagger$_{\alpha_{\text{tag}} = \text{1e-8}}$	86.25%	78.9%	4.7×	87.78%	79.7%	4.9×
Tagger$_{\alpha_{\text{tag}} = \text{1e-7}}$	95.74%	78.6%	4.7×	97.70%	79.4%	4.8×
Tagger$_{\alpha_{\text{tag}} = \text{1e-6}}$	98.40%	78.3%	4.5×	99.62%	79.0%	4.8×
Tagger$_{\alpha_{\text{tag}} = \text{1e-5}}$	99.30%	77.8%	4.5×	99.97%	78.4%	4.6×
Method		*JPT*			*YRI*	
GRAT	99.63%	90.2%	10.2×	99.72%	88.4%	8.6×
Tagger$_{\alpha_{\text{tag}} = \text{1e-8}}$	88.53%	80.5%	5.1×	87.62%	65.3%	2.9×
Tagger$_{\alpha_{\text{tag}} = \text{1e-7}}$	98.10%	80.1%	5.0×	97.55%	65.3%	2.9×
Tagger$_{\alpha_{\text{tag}} = \text{1e-6}}$	99.52%	79.6%	4.9×	99.39%	65.1%	2.9×
Tagger$_{\alpha_{\text{tag}} = \text{1e-5}}$	99.92%	79.1%	4.8×	99.94%	65.0%	2.9×

In each HapMap population, the average performance of GRAT and Tagger in 500 simulated GWASs are shown. GRAT guarantees to achieve the 99% target recall rate, while reducing the number of tests by 90%. Using Tagger, we test the remainder SNPs that are tagged by the tag-SNPs that exceed a p-value cut-off threshold, α_{tag}. GRAT outperforms the traditional tag-SNPs in all populations.

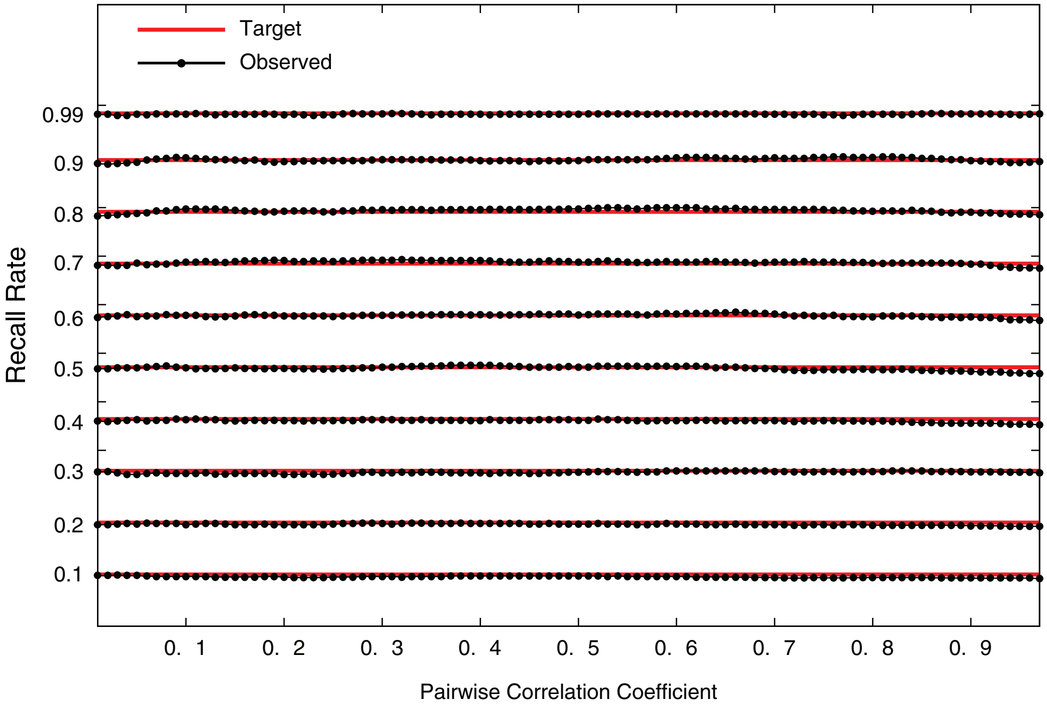


FIG. 2. Performance of the method using a single pair of SNPs. The observed recall rate of the significant causal SNP is shown for different target sensitivity and pairwise correlation values.

the significant SNPs provide insight into the genetics of the disease. Association studies attempt to collect information on as many SNPs as possible to cover the whole genome. However, as the number of collected SNPs increases so does the computational burden to identify the significant associations.

We introduced a novel method, GRAT, for genome-wide rapid association testing to identify all significant associations by testing a small subset of SNPs. Due to the correlation, or LD, testing an SNP provides information about the associations of its neighboring SNPs. Using this intuition, the procedure first tests a subset of SNPs, referred to as the proxy SNPs, across the genome to locate the regions that may contain the significant associations. Once located, additional SNPs are tested from those regions to identify the significant SNPs. Each unobserved, or remainder, SNP is paired with its most strongly correlated proxy SNP, termed best-proxy, and a threshold value is used for the best-proxy's statistic to decide whether or not to test the unobserved SNP. We introduced a novel approach to choose the proxy SNPs and determine the threshold values for each best-proxy SNP. Through simulations and real GWAS data, we showed that the proposed approach can identify more than 99% of the significant SNPs by reducing the number of tests by a factor of 10. Furthermore, GRAT can also be applied to association studies that utilize linear mixed models, where the speed-up is cumulative with recent efforts that decrease the computational burden of computing the actual association statistic. GRAT is implemented in C++ for high performance and is available online.

4. APPENDIX

4.1. Derivatives of the expected number of tests and the expected recall functions

The derivative of the expected number of tests from a single remainder SNP with respect to the decision threshold follows,

$$\begin{aligned}\frac{\partial}{\partial s^*}\mathrm{ENT}(s^*) &= \frac{\partial}{\partial s^*}\left[1-\int_{-s^*}^{s^*} f(s_t)ds_t\right] = -f(s^*)-f(-s^*)\\ &= -c_i\left[\phi\left(s^*-r\lambda_c\sqrt{N}\right)+\phi\left(s^*+r\lambda_c\sqrt{N}\right)\right]-2(1-c_i)\phi(s^*).\end{aligned} \tag{12}$$

Note that the second derivative is negative, hence convex. Therefore, the expected number of SNPs to be tested (ENT) is the sum of convex functions and is also convex. Let us denote the expected recall function by $\text{ER}(s^*) = \Pr(|S_i| > t_\alpha, |S_t| > s^*) = \rho_i$. Its derivative follows,

$$\begin{aligned}\frac{\partial}{\partial s^*}\text{ER}(s^*) &= \frac{\partial}{\partial s^*}\left[\int_{-\infty}^{-s^*}\int_{-\infty}^{-t_\alpha} f(s_i, s_t) ds_i\ ds_t\right] + \frac{\partial}{\partial s^*}\left[\int_{-\infty}^{-s^*}\int_{t_\alpha}^{\infty} + \int_{s^*}^{\infty}\int_{-\infty}^{-t_\alpha} + \int_{s^*}^{\infty}\int_{t_\alpha}^{\infty}\right] \\ &= -\int_{-\infty}^{-t_\alpha}(f(s_i, s_t = -s^*) + f(s_i, s_t = s^*))\ , ds_i - \int_{t_\alpha}^{\infty}(f(s_i, s_t = -s^*) + f(s_i, s_t = s^*))\ , ds_i.\end{aligned} \tag{13}$$

Note that given,

$$f(x, y) = \phi\left(\begin{bmatrix} x \\ y \end{bmatrix}; \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right) \tag{14}$$

it can be shown that a cross-section of the joint distribution at $y = a$ follows,

$$f(x, y = a) = \frac{\sqrt{1 - r^2}}{\sqrt{2\pi}}\exp\left(-\frac{(a - \mu_y)^2}{2}\right)\phi(x;\ \mu_x + r(a - \mu_y), 1 - r^2). \tag{15}$$

Therefore, using the joint distribution of the statistics of a remainder SNP and its best-proxy, Equation (13) can be expressed as,

$$\begin{aligned}&\text{ER}'(s^*) = \\ &\quad \frac{-1}{\sqrt{2\pi}}\Bigg[c_i \exp\left(-\frac{(s^* + r\lambda_c\sqrt{N})^2}{2}\right)\left(\Phi\left(\frac{-t_\alpha - \lambda_c\sqrt{N}(1 - r^2) + rs^*}{\sqrt{1 - r^2}}\right) + 1 - \Phi\left(\frac{t_\alpha - \lambda_c\sqrt{N}(1 - r^2) + rs^*}{\sqrt{1 - r^2}}\right)\right) \\ &\quad + c_i \exp\left(-\frac{(s^* - r\lambda_c\sqrt{N})^2}{2}\right)\left(\Phi\left(\frac{-t_\alpha - \lambda_c\sqrt{N}(1 - r^2) - rs^*}{\sqrt{1 - r^2}}\right) + 1 - \Phi\left(\frac{t_\alpha - \lambda_c\sqrt{N}(1 - r^2) - rs^*}{\sqrt{1 - r^2}}\right)\right) \\ &\quad + 2(1 - c_i)\exp\left(-\frac{(s^*)^2}{2}\right)\Bigg(\Phi\left(\frac{-t_\alpha + rs^*}{\sqrt{1 - r^2}}\right) + \Phi\bigg(\frac{-t_\alpha - rs^*}{\sqrt{1 - r^2}})\bigg)\Bigg)\Bigg].\end{aligned} \tag{16}$$

It can be shown that ER(.) is a monotonic function of the best-proxy statistic threshold, s^*. Therefore, there exists a unique ρ_i such that $\text{ER}^{-1}(\rho_i) = s^*$, where $\text{ER}^{-1}(.)$ is the inverse of the expected recall function. Using this property, the problem can be simplified by linearizing the constraint function, which reads

$$\begin{aligned}&\text{minimize} \sum_{i=K+1}^{M} \Pr(|S_{b(i)}| > \text{ER}^{-1}(\rho_i)), \\ &\text{such that} \sum_{i=K+1}^{M} \rho_i = \rho^*.\end{aligned} \tag{17}$$

Note that $\frac{\partial}{\partial\rho}\text{ER}^{-1}(\rho) = \frac{1}{\text{ER}'(\text{ER}^{-1}(\rho))}$, hence the derivative of the expected number of tests from a single remainder SNP with respect to ρ follows,

$$g = \frac{\partial}{\partial\rho}\Pr(|S_t| > \text{ER}^{-1}(\rho)) = \frac{-f(\text{ER}^{-1}(\rho)) - f(-\text{ER}^{-1}(\rho))}{\text{ER}'(\text{ER}^{-1}(\rho))}. \tag{18}$$

Using the method of Lagrange multipliers, it can be shown that at the optimum solution the expected number of tests from each remainder SNP has the same derivative value, g^*. In GRAT, we determine g^* by using binary-search such that for each remainder SNP m_i, g^* uniquely maps to ρ_i^*, where $\sum \rho_i^* = \rho^*$.

4.2. Performance on a single SNP pair

We apply the proposed method to a pair of SNPs, a causal SNP and noncausal proxy SNP, to verify whether or not the target sensitivity is reached for any value of the pairwise correlation. For each value of the correlation, we sampled thousands of joint statistics for the SNP pair and recorded how many times the causal SNP is significant. The power at the causal SNP is set to $\mathcal{P}_c = 50\%$ using a genome-wide significance level of $\alpha = 10^{-8}$.

We computed the threshold of the proxy SNP statistic for different target sensitivities in each pairwise correlation using a small prior probability for the causal SNP, $c_i = 10^{-5}$. In each correlation value, we applied the decision rules to the samples and recorded the recall rate of significant causal SNPs in each target sensitivity.

In Figure 2, the observed recall rates are shown for different values of target sensitivity and pairwise correlation. The target sensitivities are shown as horizontal lines and are followed closely by the observed recall rates. The variation around a target value is due to the asymptotic distribution of the test statistic and diminishes as the sample size increases.

ACKNOWLEDGMENTS

E.K. is supported by training grant 2T32NS048004-06A1. E.K. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, and 1065276, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, and PO1-HL28481.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Bafna, V., Halldorsson, B.V., Schwartz, R., et al. 2003. Haplotypes and informative SNP selection algorithms: don't block out information. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, RECOMB '03, 19–27.

Baker, M. 2012. Biorepositories: Building better biobanks. *Nature* 486, 141–146.

Bochner, B.R. 2003. Innovations: New technologies to assess genotype-phenotype relationships. *Nature Rev. Genet.* 4, 309–314.

Brem, R.B., and Kruglyak, L. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1572–1577.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755.

Bystrykh, L., Weersing, E., Dontje, B., et al. 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics.' *Nat. Genet.* 37, 225–232.

Carlson, C.S., Eberle, M.A., Rieder, M.J., et al. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *The American Journal of Human Genetics* 74, 106–120.

Chesler, E.J., Lu, L., Shou, S., et al. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* 37, 233–242.

Cheung, V.G., Spielman, R.S., Ewens, K.G., et al. 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437, 1365–1369.

Cookson, W., Liang, L., Abecasis, G., et al. 2009. Mapping complex disease traits with global gene expression. *Nature Rev. Genet.* 10, 184–194.

Cousin, E., Deleuze, J.F., and Genin, E. 2006. Selection of SNP subsets for association studies in candidate genes: comparison of the power of different strategies to detect single disease susceptibility locus effects. *BMC Genetics* 7.

Cousin, E., Genin, E., Mace, S., et al. 2003. Association studies in candidate genes: strategies to select SNPs to be tested. *Human Heredity* 56, 151–159.

de Bakker, P.I.W., Yelensky, R., Pe'er, I., et al. 2005. Efficiency and power in genetic association studies. *Nature Genetics* 37, 1217–1223.

Devlin, B., and Risch, N., 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311–322.

Emilsson, V., Thorleifsson, G., Zhang, B., et al. 2008. Genetics of gene expression and its effect on disease. *Nature* 452, 423–428.

Halperin, E., Kimmel, G., and Shamir, R. 2005. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics* 21 Suppl 1.

Han, B., Kang, H.M., and Eleazar, E. 2009. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet* 5.

Hardy, J, and Singleton, A. 2009. Genomewide association studies and human disease. *N. Engl. J. Med.* 360, 1759–1768.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., et al. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* 106, 9362–9367.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437, 1299–1320.

Kang, H.M., Sul, J.-H., Service, S.K., et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genet.* 42, 348.

Keurentjes, J.J.B., Fu, J., Terpstra, I.R., et al. 2007. Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1708–13.

Kostem, E., Lozano, J.A., and Eskin, E. 2011. Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms. *Genetics* 188, 449–460.

Li, Y., Willer, C.J., Ding, J., et al. 2010. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834.

Lin, Z., and Altman, R.B. 2004. Finding haplotype tagging SNPs by use of principal components analysis. *The American Journal of Human Genetics* 75, 850–861.

Lippert, C., Listgarten, J., Liu, Y., et al. 2011. Fast linear mixed models for genome-wide association studies. *Nature Methods* 8, 833.

Majewski, J., and Pastinen, T. 2011. The study of EQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* 27, 72–79.

Pardi, F., Lewis, C.M., and Whittaker, J.C. 2005. SNP selection for association studies: Maximizing power across SNP choice and study size. *Annals of Human Genetics* 69, 733–746.

Pritchard, J.K., and Przeworski, M. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69, 1–14.

Purcell, S., Neale, B., Todd-Brown, K., et al. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575.

Qin, Z.S., Gopalakrishnan, S., and Abecasis, G.R. 2006. An efficient comprehensive search algorithm for tag SNP selection using linkage disequilibrium criteria. *Bioinformatics* 22, 220–225.

Risch, N., and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.

Rockman, M.V., and Kruglyak, L. 2006. Genetics of global gene expression. *Nature Rev. Genet.* 7, 862–872.

Saccone, S.F., Rice, J.P., and Saccone, N.L. 2006. Power-based, phase-informed selection of single nucleotide polymorphisms for disease association screens. *Genetic Epidemiology* 30, 459–470.

Santana, R., Mendiburu, A., Zaitlen, N., et al. 2010. Multi-marker tagging single nucleotide polymorphism selection using estimation of distribution algorithms. *Artificial Intelligence in Medicine* 50, 193–201.

Schaid, D.J., Rowland, C.M., Tines, D.E., et al. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* 70, 425–434.

Spielman, R.S., Bastone, L.A., Burdick, J.T., et al. 2007. Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* 39, 226–231.

Stram, D.O. 2004. Tag SNP selection for association studies. *Genetic Epidemiology* 27, 365–374.

Stram, D.O. 2005. Software for tag single nucleotide polymorphism selection. *Human Genomics* 2, 144–151.

Stranger, B.E., Montgomery, S.B., Dimas, A.S., et al. 2012. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* 8, e1002639.

Stranger, B.E., Nica, A.C., Forrest, M.S., et al. 2007. Population genomics of human gene expression. *Nat. Genet.* 39, 1217–1224.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467, 1061.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 306, 636–640.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816.

The ENCODE Project Consortium. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 9, e1001046.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* 10, 57–63.

Zhou, X., and Stephens, M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genet.* 44, 821–824.

Address correspondence to:
Dr. Emrah Kostem
Department of Computer Science
University of California at Los Angeles
4732 Boelter Hall
Los Angeles, CA 90095

E-mail: ekostem@cs.ucla.edu