Genetics: Early Online, published on August 7, 2014 as 10.1534/genetics.114.167908

Identifying causal variants at loci with multiple signals of association

Farhad Hormozdiari[1*], Emrah Kostem[1*], Eun Yong Kang[1], Bogdan Pasaniuc[2,3,†], Eleazar Eskin[1,2,†,‡‡]

1 Department of Computer Science, University of California, Los Angeles, California 90095, USA

2Department of Human Genetics, University of California, Los Angeles, California 90095, USA

3 Department of Pathology & Laboratory Medicine, University of California, Los Angeles, California 90095, USA

∗ **These authors contributed equally to this work**

† **These authors contributed equally to this work**

‡‡ **corresponding author : eeskin@cs.ucla.edu**

ABSTRACT

Although genome-wide association studies have successfully identified thousands of risk loci for complex traits, only a handful of the biologically causal variants, responsible for association at these loci, have been successfully identified. Current statistical methods for identifying causal variants at risk loci either use the strength of association signal in an iterative conditioning framework, or estimate probabilities for variants to be causal. A main drawback of existing methods is that they rely on the simplifying assumption of a single causal variant at each risk locus which is typically invalid at many risk loci. In this work, we propose a new statistical frameworks that allows for the possibility of an arbitrary number of causal variants when estimating the posterior probability of a variant being causal. A direct benefit of our approach is that we predict a set of variants for each locus that under reasonable assumptions will contain all of the true causal variants with a high confidence level (e.g. 95%) even when the locus contains multiple causal variants. We use simulations to show that our approach provides 20-50% improvement in our ability to identify the causal variants compared to the existing methods at loci harboring multiple causal variants. We validate our approach using empirical data from a eQTL study of *CHI3L2* to identify new causal variants that affect gene expression at this locus.

INTRODUCTION

Although genome-wide association studies (GWAS) reproducibly identified thousands of risk loci (Zeggini *et al.* 2007; Sladek *et al.* 2007; Hakonarson *et al.* 2007; Yang *et al.* 2011; Kottgen *et al.* 2013; Lu *et al.* 2013; Ripke *et al.* 2013) only a handful of causal genetic variants (i.e. variants that biologically alter

disease risk) have been found (McCarthy *et al.* 2008; Manolio *et al.* 2008; Altshuler *et al.* 2008), thus prohibiting the mechanistic understanding of genetic basis of common diseases. The linkage disequilibrium (LD) (Reich *et al.* 2001; Pritchard and Przeworski 2001) structure of the human genome has greatly benefited GWAS in interrogating only a subset of all variants to assay common variation across the genome. Unfortunately, LD hinders the identification of causal variants at risk loci in fine-mapping studies as at each locus, there are often tens to hundreds of variants tightly linked to the reported associated single-nucleotide polymorphism (SNP) (Malo *et al.* 2008; Yang *et al.* 2012; Maller *et al.* 2012). In a continued effort to identify causal variants, many fine-mapping studies that assess genetic variation at known GWAS risk loci are currently underway (Bauer *et al.* 2013; Coram *et al.* 2013; Diogo *et al.* 2013; Gong *et al.* 2013; Marigorta and Navarro 2013; Peters *et al.* 2013; Wu *et al.* 2013).

Fine-mapping studies typically follow a two-step procedure. First, a statistical analysis of the association signal is performed to identify a minimum set of SNPs that can explain the signal. Second, the SNPs that are putatively casual are functionally tested using laborious and expensive functional assays. Therefore, the objective of the statistical component of fine-mapping is to minimize the number of SNPs that need to be selected for follow-up studies while identifying the true causal SNPs. In this work, we focus on developing approaches for statistical refinement of the association signal with the goal of identifying the minimum set of variants to be tested to identify all the causal variants. Although in this work we primarily focus on common variants, our work can be extended to rare variants through careful regularization of normalized association scores (z-scores) (Navon *et al.* 2013).

The basic statistical fine-mapping approach is to select SNPs for functional validation based on the strength of the association signal. A standard statistical association test is performed followed by the selection of the top k SNPs with the highest evidence of association for functional assays. The value of k depends on the budget and resources assigned for the follow-up study. This procedure is suboptimal as it does not properly account for the LD at a particular locus (Lawrence *et al.* 2005; Udler *et al.* 2009; Faye *et al.* 2013). For example, two SNPs in perfect LD will always show the same association statistic and it is unclear how to prioritize these SNPs for functional assays. In addition, the finite sampling of individuals in the fine-mapping study induces statistical noise in the association statistics that can result in higher association statistics at neighboring SNPs as opposed to the true causal SNP. Furthermore, even when the sample sizes are large enough such that the statistical noise can be ignored, the local LD structure can induce higher association statistics for neighboring SNPs rather than causal variants at loci with multiple causal variants (Udler *et al.* 2009). More fundamentally, this approach provides no guarantees that the actual causal SNPs are contained in the top k SNPs selected for functional assays.

In this paper, as opposed to the basic top k approach, recent works (Maller *et al.* 2012; Beecham *et al.* 2013) have proposed to estimate the probability of each SNP to be causal at a given locus under the simplifying assumption that each GWAS associated locus harbors exactly one causal variant. Under this assumption the approximation of the posterior can be computed using only the marginal per-SNP association statistics. This induces a one-to-one relationship between marginal association statistics and the estimated posterior probabilities that yields the same ranking of SNPs within each locus. A major advantage of this approach is that confidence intervals (i.e. sets of SNPs that account for the 95% of all the posterior probability of causal variants in the locus) can be estimated and used to determine the number of SNPs for each locus to follow up in functional assays. A major drawback of this approach is that the confidence intervals rely on the assumption of a single causal variant per locus. As we show below, when applied to loci where there are more than one causal variants (Trynka *et al.* 2011; Galarneau *et al.* 2010; Haiman *et al.* 2007; Flister *et al.* 2013; Chung *et al.* 2011; Allen *et al.* 2010; Stahl *et al.* 2012), the confidence intervals may not contain any causal variants with a much higher than expected likelihood.

As opposed to approaches above that yield same ranking of SNPs, conditioning approaches to dissect association signal that may change ranking of variants have also been proposed (Trynka *et al.* 2011; Galarneau *et al.* 2010; Flister *et al.* 2013; Chung *et al.* 2011; Allen *et al.* 2010; Stahl *et al.* 2012). The conditional approach relies on an iterative selection of most associated SNPs followed by re-computation of the statistical score for the remaining SNPs conditional on the already selected SNPs. The iterations continue until no significant signal remains in the locus at a nominal or Bonferroni corrected significance (Yang *et al.* 2012; Allen *et al.* 2010; Sklar *et al.* 2011; Yang *et al.* 2011; Udler *et al.* 2009). Although conditioning is amenable for identifying the presence of multiple signals within the locus, it can also lead to the unfavorable situation of selection of no causal SNPs for follow-up assays. For example, in the case of two SNPs in perfect LD, where only one of the SNPs is the causal variant, the conditioning approach will drop one of the SNPs from the analysis depending on the order in which the SNPs are selected in the iterative procedure. Since the statistics at these two SNPs are mathematically equal, the order can only be random (in the absence of other sources of information) leading to conditioning not finding any casual variants in 50% of the cases. This underlines a major drawback of the conditioning approach that can lead to highly suboptimal scenarios when searching for variants to test in functional assays.

As compared to previous work, we propose CAVIAR (CAusal Variants Identification in Associated Regions), a statistical framework that quantifies the probability of each variant to be causal while allowing with arbitrary number of causal variants. We accomplish this by jointly modeling the observed association statistics at all variants in the risk locus; posterior probabilities for sets of variants to be causal are then

estimated using the conditional distribution of all association statistics in the locus conditional on the set of causal variants. The output of our approach is a set of variants that with a certain probability (e.g. 95%) contain all of the casual variants at that locus. Intuitively, the 95% causal confidence set is akin to a 95% confidence interval around an estimated parameter. Through extensive simulations we show that our method attains superior performance over all existing methods with comparable results at loci where there is a single causal variant. We validate our approach using empirical data from an eQTL study of the *CHI3L2* gene(Cheung *et al.* 2005) where the true causal variants are known. CAVIAR correctly identifies the true causal variant.

RESULTS

Overview of statistical fine-mapping. Our approach, CAVIAR (CAusal Variants Identification in Associated Regions), takes as input the association statistics for all of the SNPs (variants) at the locus together with the correlation structure between the variants obtained from a reference dataset such as the HapMap(Gibbs *et al.* 2003; Frazer *et al.* 2007) or 1000 Genomes project (Abecasis *et al.* 2010) data. Using this information, our method predicts a subset of the variants that has the property that all the casual SNPs are contained in this set with the probability ρ (we term this set as the "ρ causal set"). In practice we set ρ to values close to 100%, typically greater than or equal to 95% and let CAVIAR find the set with the fewest number of SNPs which contains the causal SNPs with probability at least ρ. The casual set can be viewed as a confidence interval. We use the casual set in the follow up studies by validating only the SNPs that are present in the set. While in the paper we discuss SNPs for simplicity, our approach can be applied to any type of genetic variants including structural variants.

We used simulations to show the effect of LD on the resolution of fine-mapping. We selected two risk loci (with large and small LD) to showcase the effect of LD on fine-mapping (see Figure1(a) and (b)). The first region is obtained by considering 100kbp upstream and downstream of the rs10962894 SNP from the coronary artery disease (CAD) case control study. As shown in the Figure 1(a), the correlation between the significant SNP and the neighboring SNPs is high. We simulated GWAS statistics for this region by taking advantage that the statistics follow a multivariate normal distribution, as shown in Han et al. and Zaitlen et al. (2009,2010) (see Methods). CAVIAR selects the true causal SNP which is SNP8, together with 6 additional variants (Figure 1(a)). Thus, when following up this locus, we only have to consider these SNPs to identify the true causal SNPs. The second region showcases loci with lower LD (see Figure 1(b)). In this region only the true causal SNP is selected by CAVIAR (SNP18). As expected,the size of the ρ causal set is a function of the LD pattern in the locus and the value of ρ with higher values of ρ resulting in larger sets (see Table S1 and S2).

We also showcase the scenario of multiple casual variants (see Figure 2). We simulated data as before and considered SNP25 and SNP29 as the causal SNPs. Interestingly, the most significant SNP (SNP27, see Figure 2) tags the true causal variants but it is not itself causal making the selection based on strength of association alone under the assumption of a single causal or iterative conditioning highly suboptimal. To capture both causal SNPs at least 11 SNPs must be selected in ranking based on p-values or probabilities estimated under a single causal variant assumption. As opposed to existing approaches, CAVIAR selects both SNPs in the 95% causal set together with 5 additional variants. The gain in accuracy of our approach comes from accurately disregarding SNP30-SNP35 from consideration since their effects can be captured by other SNPs.

Iterative conditioning is suboptimal in statistical fine-mapping. We performed simulations to assess the performance of various approaches for identification of the causal variants in fine-mapping studies. In each simulation, we randomly selected one of the SNPs in this region as a causal SNP and generated association statistics for the 35 SNPs using our data-generating model (see Methods). We set the statistical power at the causal SNP to be 50% at the genome-wide significance level of $\alpha = 10^{-8}$. This way, on average, the causal SNP statistic is significant in half of the simulation panels, and the causal SNP does not always attain the peak statistic in the region. Using this procedure, we generated 1000 simulation panels. Figure 1(c) and 1(d) indicate the ranking of the causal SNP for both regions, where the x-axis is the ranking of the true causal SNP and the y-axis is the number of simulations where the true causal SNP have that specific ranking. We observe the top k SNP where k is set to one, fails to find the true causal SNP in 5%-40% of the time depending how complex is the LD pattern in the region. Furthermore, this result illustrates that the first step of the conditional method, which selects the most significant SNP, will fail to select the right SNP in 5%-40% of the time.

CAVIAR outperforms existing approaches in fine-mapping. We used HapGen (Spencer *et al.* 2009) to simulate fine-mapping data across European populations in the 1000 Genome project(Abecasis *et al.* 2010) across regions consisting of 50 SNPs. We randomly implanted one, two, or three causal SNPs in each region and then simulated case-control studies. We perform a t-test for each SNP to obtain the marginal statistical scores for each SNP. After obtaining the statistical scores and the LD correlation between each SNP, we apply our method. Figure 3 illustrates the recall rate and the size of causal set for our method and the two competing methods (conditional and posterior methods). We define recall rate as the fraction of simulations where all the true causal SNPs are identified. The x-axis indicates the number of true causal SNPs implanted in each region. First we compared the recall rate of a probabilistic method that assumes a

single causal variant (1-Post, Maller et al. (Maller *et al.* 2012)) and CAVIAR. In simulations of a single causal variant both methods are well-calibrated while in scenarios with multiple causals CAVIAR is the only approach that maintains a well-calibrated recall rate. Our simulations suggest that the approach that assumes a single causal variant will attain miss-calibrated recall rates at loci with multiple casual variants.

In the above experiments, CAVIAR shows the best recall rate compared to the competing methods. However, the number of SNPs selected by CAVIAR in the causal set is slightly higher than those methods. In order to make the comparison among these methods fair, we extended the CM and 1-Post methods such that the number of SNPs selected by each method is equal to the number of SNPs selected by CAVIAR. The extension of CM and 1-Post methods are referred to as the ECM and E1-Post methods. As shown in Figure 4, our method has the highest recall rate among the competing methods for all the scenarios. Furthermore, we compared the ranking of the causal SNPs for each method. We vary the number of SNPs selected by each method from one SNP to ten SNPs and compare the recall rate and the results are shown in Figure 5. The x-axis is the number of SNPs selected by each method and the y-axis is the recall rate for each method.

We also assessed the impact of the number of individuals in the fine-mapping study. As expected, we find that CAVIAR's confidence set decreases with increased sample size (see Figure S1).

Fine-mapping of the CHI3L2 locus. To validate simulation results, we applied CAVIAR to the *CHI3L2* region using the gene expression as a phenotype. This locus was extensively fine-mapped with the true causal variant already identified(Cheung *et al.* 2005; Malo *et al.* 2008; Chen and Witte 2007). We obtained marginal statistical scores for each SNP from the Malo et al.(Malo *et al.* 2008) study and inferred LD patterns from the HapMap data for 57 unrelated individuals of European ancestry (CEU), the same set of individuals used by previous studies. The result of our method and the LD pattern is shown in Figure 6. CAVIAR selects rs755467, rs961364, rs2764543, rs2477578, rs3934922 and rs8535 for the causal set. Cheung et al.(Cheung *et al.* 2005) illustrate the rs755467 SNP is the causal SNP through luciferase reporter and haplotype-specific chromatin immunoprecipitation assays. Furthermore, using the CM method and conditioning on the known true causal SNP (rs755467), we obtain the secondary signal in the region which is rs2764543. The E1-Post 95% causal set selected the same 6 SNPs as CAVIAR. The ECM selects rs755467, rs2274232, rs2182115, rs2764543, rs2820087, and rs11583210 for the causal set.

DISCUSSION

Over the past few years, genome-wide association studies (GWAS) have identified hundreds of genetic loci harboring genetic variation affecting disease risk for hundreds of common diseases (Bauer *et al.* 2013; Coram *et al.* 2013; Diogo *et al.* 2013; Gong *et al.* 2013; Marigorta and Navarro 2013; Peters *et al.*

2013; Wu *et al.* 2013). Identifying the causal genetic variants affecting disease risk at these loci has the potential of providing clues to the mechanism of the disease which can lead to identification of better targets for drug terrapins. Unfortunately, the pervasive linkage disequilibrium (LD) and the uncertainty of data makes the task of deconvoluting causal variants from tagging ones very challenging.

In this paper, we present a novel framework for identifying the causal variants underlying GWAS risk loci. The key idea behind our framework is that instead of considering each variant one at a time, we instead analyze all of the variants in the entire locus simultaneously. The result of our method is a set of variants which with high probability contains (or captures) all the causal variants. Through extensive simulation results, we show that our approach is superior to existing methods in reducing the overall number of variants to be examined in functional follow-up to identify the causal variants.

In our method we make a series of assumptions to ease the computational burden and to simplify the model. We make the assumption the number of causal SNPs in a region, in which we are interested to preform fine-mapping, is at most 6. Our method also makes the standard assumption of the Fisher's polygenic model that effects size follow a normal distribution with mean zero. This assumption is the basis of many recent approaches to estimate heritability (Kostem and Eskin 2013; Yang *et al.* 2011; Speed *et al.* 2012) and to correct for population structure in GWAS (Kang *et al.* 2008; Lippert *et al.* 2011; Listgarten *et al.* 2012; Zhou and Stephens 2012; Segura *et al.* 2012).

Our method also assumes that we have genotyped each variant in the locus. With the increasing cost efficiency of high throughput sequencing, this assumption is becoming more and more realistic. One future direction of research is to extend this approach to handle imputed association statistics. In this case, only a relatively small number of individuals in a GWAS must be fully sequenced at the locus while the rest of the individuals can use the sequenced individuals as an imputation reference panel.

Our method takes as input the association statistics and linkage disequilibrium patterns in the locus to identify the set of variants which are likely to contain the causal variants. The minor allele frequencies of the variants will both affect the magnitude of the observed statistics as well as the linkage disequilibrium patterns. However, our approach is only applied to loci which harbor significant association signals at individual's variants. These types of signals are most likely driven by common variants. Most likely, additional rare variants in the locus which also have effects on the phenotype will not be selected because their association statistics are low. Extending our approach to discover additional rare variants in a locus is an interesting direction for future work.

CAVIAR can easily take into account data on putative function of variants either from functional genomic data (Bernstein *et al.* 2012) or eQTL data which has been recently shown to help facilitate fine mapping

studies (Hoffman *et al.* 2012; Edwards *et al.* 2013). The way that this information can be incorporated is by assigning each variant a prior probability of affecting the trait (Darnell *et al.* 2012; Eskin 2008; Jul *et al.* 2011). In this framework, the functional genomic data is converted to a probability between 0 and 1 of that variant having affect on the trait. These priors then affect the likelihood of each causal status and then ultimately are incorporated into the final causal set.

The method presented in this paper has some conceptual similarities to methods for identifying associations in regions where there is more than one associated variant. These methods have become very popular in the context of rare variant association studies (Navon *et al.* 2013; Jul *et al.* 2011; Long *et al.* 2013; Li and Leal 2008; Madsen and Browning 2009). However, there are other methods which also consider common variants as well (Yi *et al.* 2011; Wu *et al.* 2011). Our method differs from these approaches in that our goal is to narrow down the possible set of variants in a locus which we suspect is associated while the previous approaches utilize multiple variants to attempt to identify an associated locus.

Compared to methods for association testing, methods for fine mapping, including the proposed method, are more complicated and make many implicit or explicit assumptions. For example, our method makes explicit assumptions about the effect size of causal variants while association methods make no such assumptions. In our view, this is inherent to the fact that fine mapping methods attempt to control false negatives compared to association methods which attempt to control false positives. In order to control false negatives, fine mapping methods must make explicit assumptions about the "alternate" distribution in order to understand how well the data fits the assumptions. Association method on the other hand in order to control false positives, only need to make assumptions about the null distribution which in the case of association studies is the assumption that all of the variants at a locus have no effects. This asymmetry characterizes the fine mapping problem and complicates attempts to merge fine mapping and association into a single framework.

MATERIAL AND METHODS

The traditional fine-mapping study approach A fine-mapping study is a procedure to identify, or predict, the disease causing single-nucleotide polymorphisms (SNPs) from a given genome-wide association study (GWAS) dataset. It is assumed that the genotype data is dense enough, such that all the causal SNPs are genotyped, including the SNPs that are perfectly correlated to the causal variants other than SNPs. With the development of sequencing technologies, this assumption is becoming more realistic. Therefore, we assume that there exists a true label for each genotyped SNP on whether or not the SNP is causal in disease.

The traditional fine-mapping study approach performs the following iterative procedure to predict the causal SNPs within a genomic region. First, the association statistic of each SNP is computed and the

most strongly associated SNP is chosen as a causal SNP. Intuitively, if the region contains a single causal SNP, then the most significantly associated SNP is likely to be the causal SNP itself (the assumption in the traditional fine-mapping approach). However, the region may contain multiple causal SNPs, and furthermore these SNPs may be correlated, or in linkage disequilibrium (LD). In this scenario, the association statistic at a causal SNP may be contaminated by the presence of the causal SNPs that are in LD. In order to control for this contamination, at each iteration, the traditional approach re-computes the association statistic of the SNPs while conditioning on the presence of the causal SNPs which are identified in each iteration of the method. Given a statistic threshold, if the statistic of the most strongly associated SNP exceeds the threshold, the SNP is chosen as a causal SNP, or otherwise the procedure terminates.

We show through empirical and theoretical results that the traditional approach is under-powered to identify the causal SNP compared to our method. In the next section we present a data-generating model for fine-mapping studies.

Data-generating model for fine-mapping studies We consider a genome-wide association study (GWAS) on a quantitative trait where n individuals are genotyped on m SNPs. For individual k, we are given the phenotypic value y_k and the genotype values at m SNPs, where for SNP i, $g_{ik} \in \{0, 1, 2\}$ is the minor allele count. Let $\boldsymbol{y}$ denote the $(n \times 1)$ vector of the phenotypic values and $\boldsymbol{x}_i$ denote the $(n \times 1)$ vector of normalized genotype values at SNP i such that $\mathbf{1}^T \boldsymbol{x}_i = 0$ and $\boldsymbol{x}_i^T \boldsymbol{x}_i = n$.

Let's assume that a SNP c is the only SNP involved in the disease. We assume the data generating model follows a linear model,

$$\boldsymbol{y} = \mu \mathbf{1} + \beta_c \boldsymbol{x}_c + \boldsymbol{e}$$

where $\mathbf{1}$ denotes the $(n \times 1)$ vector of ones, μ is the intercept, β_c is the effect-size of SNP c and $\boldsymbol{e}$ is the $(n \times 1)$ vector of i.i.d. and normally distributed residual noise, where $\boldsymbol{e} \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}\right)$ with covariance scalar σ and $(n \times n)$ identity matrix $\boldsymbol{I}$.

The estimates for μ and β_c are obtained by maximizing the likelihood function,

$$\boldsymbol{y} \sim \mathcal{N}\left(\mu \mathbf{1} + \beta_c \boldsymbol{x}_c, \sigma^2 \boldsymbol{I}\right),$$

$$\mathcal{L}\left(\boldsymbol{y} | \mu, \beta_c, \sigma^2\right) = |2\pi\sigma^2 \boldsymbol{I}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \mu\mathbf{1} - \beta_c \boldsymbol{x}_c\right)^T \left(\boldsymbol{y} - \mu\mathbf{1} - \beta_c \boldsymbol{x}_c\right)\right),$$

$$\frac{\partial \mathcal{L}\left(\boldsymbol{y} | \mu, \beta_c, \sigma^2\right)}{\partial \mu} = 0 \; \hat{\mu} = \frac{1}{n} \mathbf{1}^T \boldsymbol{y}, \; \hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

$$\frac{\partial \mathcal{L}\left(\boldsymbol{y} | \mu, \beta_c, \sigma^2\right)}{\partial \beta_c} = 0 \; \hat{\beta}_c = \frac{\boldsymbol{x}_c^T \boldsymbol{y}}{n}, \; \sqrt{n}\frac{\hat{\beta}_c}{\sigma} \sim \mathcal{N}\left(\frac{\beta_c}{\sigma}\sqrt{n}, 1\right).$$

The association statistic for SNP c, denoted by $S_c = \hat{s}_c$, follows a non-central t-distribution, which is the ratio of a normally distributed random variable to the square root of an independent chi-squared distributed random random variable,

$$\hat{s}_c = \frac{\frac{\sqrt{n}\hat{\beta}_c}{\sigma}}{\sqrt{\frac{1}{n}\frac{\sqrt{\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}}}{\sigma}}} = \frac{n\hat{\beta}_c}{\sqrt{\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}}} \sim t_{(\lambda_c, n)},$$

with non-centrality parameter (NCP) $\lambda_c = \frac{\beta_c}{\sigma}\sqrt{n}$, and n degrees of freedom. Note that,

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\mu}\mathbf{1} - \hat{\beta}_c\boldsymbol{x}_c,\ \frac{\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}}{\sigma^2} \sim \chi^2_n,$$

where χ^2_n denotes the chi-squared distribution with n degrees of freedom and it can be shown that $\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}$ is independent of $\hat{\beta}_c$.

For simplicity, we assume the sample size n is large-enough, such that the association statistic S_c is well approximated by a normal distribution with NCP λ_c and unit variance,

$$S_c \sim t_{\lambda_c, n} \approx \mathcal{N}(\lambda_c, 1).$$

Furthermore, if SNP i is correlated with a disease involved SNP c with coefficient r, i.e. $\frac{1}{n}\boldsymbol{x}_i^T\boldsymbol{x}_c$, the estimate of its effect-size follows,

$$\hat{\beta}_i = \frac{\boldsymbol{x}_i^T\boldsymbol{y}}{n},\ \sqrt{n}\frac{\hat{\beta}_i}{\sigma} \sim \mathcal{N}\left(r\frac{\beta_c}{\sigma}\sqrt{n}, 1\right).$$

The covariance between the two normal random variables reads,

$$\mathrm{Cov}\left(\sqrt{n}\frac{\hat{\beta}_i}{\sigma}, \sqrt{n}\frac{\hat{\beta}_c}{\sigma}\right) = \frac{1}{n\sigma^2}\boldsymbol{x}_i^T\,\mathrm{Var}(\boldsymbol{y})\,\boldsymbol{x}_c = r.$$

Therefore, the joint distribution of the association statistics of two SNPs in a region follows a multivariate normal distribution,

$$\begin{bmatrix} S_i \\ S_j \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \lambda_i \\ \lambda_j \end{bmatrix}, \begin{bmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{bmatrix}\right),$$

If we assume the i-th SNP is causal we have $\lambda_j = r_{ij}\lambda_i$ and if we assume the j-th SNP is causal we have $\lambda_i = r_{ij}\lambda_j$. Given the significance level α and the observed value of the test statistic $\hat{s}_i$, the SNP is deemed as significant, or statistically associated, if $|\hat{s}_i| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$, where $\Phi^{-1}(.)$ is the quantile function of the standard normal distribution.

The equivalent derivation showing that the joint distribution of the association statistics in case/control studies follows the MVN has been shown in Han et al., (2009).

A new framework for computing the posterior probability of causal SNP statuses from GWAS data Consider we are given a set of m SNPs $\mathcal{M}$, with their pairwise correlation coefficients $\mathbf{\Sigma}$. We introduce a new parameter, $\boldsymbol{c}$, an $(m \times 1)$ causal status indicator vector, c_i denoting an element for that vector. There are three possible causal status for each SNP: positive effect ($c_i = +1$), negative effect ($c_i = -1$), and no effect $c_i = 0$. The indicator vector $\boldsymbol{c}$ can take 3^m possible causal statuses, denoted by the set $\mathcal{C}$, with $3^m - 1$ of them having at least one causal SNP.

We denote the association statistics of the SNPs by the $(m \times 1)$ vector $\boldsymbol{S} = \begin{bmatrix} S_1 & \dots & S_m \end{bmatrix}^T$, which follows a multivariate normal distribution,

$$\boldsymbol{S} \sim \mathcal{N}\left(\lambda_c \mathbf{\Sigma} \boldsymbol{c}, \mathbf{\Sigma}\right), \tag{1}$$

where, for simplicity in presenting the model, we assume all causal SNPs have the same NCP, λ_c. Later, we will relax this assumption by utilizing the standard Fisher's polygenic model that effects size follow a normal distribution with mean zero. Although the above equation holds for common variants, we can extended it to rare variants through careful regularization of normalized association scores (z-scores) (Navon *et al.* 2013).

Let $\boldsymbol{c}^* \in \mathcal{C}$ denote a particular causal status. We define a prior probability over the possible causal statuses, P(c) which assumes that each variant has a probability of being causal in either direction, γ,

$$\mathrm{P}\left(\boldsymbol{c}\right) = \prod \gamma^{|c_i|}(1-2\gamma)^{(1-|c_i|)}$$

Below, we extend the prior to allow for incorporating functional information into our approach.

Given the observed association statistics of the m SNPs, $\hat{\boldsymbol{s}} = \begin{bmatrix} \hat{s}_1 & \dots & \hat{s}_m \end{bmatrix}^T$, the posterior probability of the causal status $\mathrm{P}\left(\boldsymbol{c}^* | \hat{\boldsymbol{s}}\right)$ can be expressed as,

$$\mathrm{P}\left(\boldsymbol{c}^* \mid \hat{\boldsymbol{s}}\right) = \frac{\mathrm{P}\left(\hat{\boldsymbol{s}} \mid \boldsymbol{c}^*\right)\mathrm{P}\left(\boldsymbol{c}^*\right)}{\sum_{\boldsymbol{c}\in\mathcal{C}} \mathrm{P}\left(\hat{\boldsymbol{s}} \mid \boldsymbol{c}\right)\mathrm{P}\left(\boldsymbol{c}\right)}. \tag{2}$$

Given a set of SNPs $\mathcal{K} \subset \mathcal{M}$, we denote the set of causal SNP configurations rendered by $\mathcal{K}$ with $\mathcal{C}_\mathcal{K}$, which excludes all causal SNP configurations having a SNP outside of $\mathcal{K}$ as causal. Note that, our definition for $\mathcal{C}_\mathcal{K}$ includes the null configuration of having no causal SNPs as well. Using $\mathcal{C}_\mathcal{K}$, we can compute the posterior probability of $\mathcal{K}$ to include, or capture, all the causal SNPs,

$$\mathrm{P}\left(\mathcal{C}_\mathcal{K} | \hat{\boldsymbol{s}}\right) = \sum_{\boldsymbol{c}\in\mathcal{C}_\mathcal{K}} \mathrm{P}\left(\boldsymbol{c} \mid \hat{\boldsymbol{s}}\right).$$

We denote the value of this posterior probability with ρ, where $\rho = \mathrm{P}\left(\mathcal{C}_\mathcal{K} | \hat{\boldsymbol{s}}\right)$, and refer to it as the confidence level of $\mathcal{K}$ in capturing the causal SNPs. Similarly, we refer $\mathcal{K}$ as a "*ρ confidence-set of causal SNPs*" or a "*ρ confidence-set*".

Given a minimum confidence threshold ρ^*, there can be many confidence-sets, each having a confidence level that is greater than the threshold. Among all these sets, the ones with smaller number of SNPs are more

informative, or have higher resolution, in locating the causal SNPs. Then, the problem we are interested is to find the ρ^* confidence-set with the minimum size,

$$\mathrm{P}\left(\mathcal{C}_{\mathcal{K}^*}|\hat{\boldsymbol{s}}\right) \geq \rho^*,$$

where $\mathcal{K}^*$ has the minimum size.

Generalized framework for locus with multiple causal SNPs with different NCP values In the previous section we consider the case where all the causal SNPs in a locus have the same NCP. Thus, $\lambda_c \boldsymbol{c}$ indicates a point in a R^m space and the coordinates corresponding to the causal SNPs have value of $\pm\lambda_c$ and the coordinates corresponding to the non-causal SNPs have a value of zero. We relax this assumption to instead have the NCP for each causal SNP drawn from a distribution with mean 0 and variance σ^2. This is the standard assumption of the Fisher's polygenic model.

We define the prior probability on the vector of NCP $\boldsymbol{\lambda_c}$ for a given causal status $\boldsymbol{c}$ using the multivariate normal probability,

$$\left(\boldsymbol{\lambda_c}|\boldsymbol{c}\right) \sim \mathcal{N}\left(0, \boldsymbol{\Sigma_c}\right),$$

where $\boldsymbol{\Sigma_c}$ is constructed as follows:

$$\boldsymbol{\Sigma_c}\{i,j\} = \begin{cases} 0 & i \neq j \\ \sigma & \text{if i is causal} \\ \epsilon & \text{if i is not causal} \end{cases}$$

ϵ is a small constant which ensures that the matrix $\boldsymbol{\Sigma_c}$ is of full rank. The final prior is then

$$\begin{aligned} \mathrm{P}\left(\boldsymbol{c}, \boldsymbol{\lambda_c}\right) &= \mathrm{P}\left(\boldsymbol{c}\right) \mathrm{P}\left(\boldsymbol{\lambda_c}|\boldsymbol{c}\right) \\ &= \textstyle\prod_{i=1} \gamma^{|c_i|}(1-\gamma)^{1-|c_i|} f(\boldsymbol{\lambda_c}, 0, \boldsymbol{\Sigma_c}), \end{aligned} \tag{3}$$

where $f(\boldsymbol{\lambda_c}, 0, \boldsymbol{\Sigma_c})$ is the probability density function of the causal status $(\boldsymbol{\lambda_c}|\boldsymbol{c}) \sim \mathcal{N}\left(0, \boldsymbol{\Sigma_c}\right)$. We use the above generalization as a prior on the mean of the distribution indicated in equation (1). We know the LD between two SNPs is symmetric ($\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$) and the NCP $\boldsymbol{\lambda} = \boldsymbol{\Sigma\lambda_c}$,

$$\boldsymbol{\lambda} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma\Sigma_c\Sigma}\right),$$

Thus, the association statistics of the SNPs follows a multivariate normal distribution,

$$\boldsymbol{S} \sim \mathcal{N}\left(0, \boldsymbol{\Sigma} + \boldsymbol{\Sigma\Sigma_c\Sigma}\right),$$

Optimization In order to compute the posterior probability for each set, which is shown in equation (2), we calculate the summation over the likelihood of all the possible causal status. Unfortunately, computing this summation that is the denominator of the equation (2) is computationally intractable in the general case (multiple causal SNPs with different NCP values). Thus, in order to simplify the calculation we assume the total number of causal SNPs in a region is bounded by at most 6 causal SNPs. Although this assumption simplifies the denominator in the equation (2), to detect the minimum causal set still we have to consider all the possible causal status. We utilize the following greedy algorithm to make the detection of minimum causal set tractable. In each iteration of the greedy algorithm we select a SNP to be causal that increases the posterior probability the most. The process of selecting SNPs to be causal continues as long as the posterior probability of the causal set is at least ρ fraction of the total posterior probability of the data.

Using simulated data, we show in Table S3 the proposed greedy method results are similar to the results obtained by solving the equation (2) exactly. In addition, for each causal status we define a prior. In order to compute the prior, we assume each SNP is independent and the probability of a SNP to be causal is equal to 10^{-2} (Eskin 2008).

In order to identify the causal SNP sets, we need to consider all possible subsets of the SNPs which numbers 2^m (in the case of multiple causal SNPs with different NCP values, we consider two causal status for each SNP: have effect or have no effect) when m is the number of SNPs in the region. In the process of computing the posterior probability for each of these possible subsets, we need to enumerate over each possible causal status for each SNP. There are two possible causal status for each SNP. SNP has an effect or SNP has no effect. Thus for each possible subset of SNPs, we need to consider 2^m possible causal statuses for the SNPs. For each of these statuses, the multivariate normal distribution is utilized to compute the likelihood of the data given the causal statuses. Thus in order to identify the best causal SNP set, we must perform a significant amount of computation.

The computational burden is high because we need to consider every possible subsets of SNPs to be in causal set and for each subset, we need to enumerate all of the possible causal SNP status. We propose two ideas to reduce the computational burden. The first idea only reduces the possible causal status which we need to consider for each subset. The second idea utilizes a greedy algorithm to identify the subset of SNPs in the causal set by eliminating our need to consider all possible subsets.

In order to reduce the computational burden, we assume in each region we have at most 6 causal SNPs. If we only consider causal status that have a total of i causal SNPs, there are $2^i\binom{m}{i}$ possible different causal status. Thus, for the case where we only consider at most 6 causal SNPs we have $\sum_{i=1}^{6} 2^i\binom{m}{i}$ possible causal status which reduces the number of possible causal statuses. The intuition behind this assumption lies in the

fact that causal variants are relatively rare. Using the simulated data we show (Table S3) the set obtained by considering only 6 causal SNPs in a region is highly similar to the set obtained by considering all the 2^m causal status.

The assumption of at most 6 causal SNPs reduces the computational burden to compute the posterior probability for each subset of SNPs. However, to identify the causal SNP sets, we need to select the smallest subset of SNPs that has the desired posterior probability. This process can be extremely slow in some cases as we need to consider all the possible subsets of SNPs. We proposed an efficient greedy method where in each iteration of the method we select a SNP which increases the posterior probability the most. We continue the process of adding SNPs to causal set until we have the desired posterior probability for the causal set.

Incorporating functional data as prior into CAVIAR Although we consider a simple prior in our model, CAVIAR can easily be extended to incorporate external information such as functional data or knowledge from previous studies. These external information can be incorporated to CAVIAR as a prior. We allow the probability that a variant is part of causal set to vary from variant to variant depending on prior information. This variant specific probability is denoted γ_i. We extend equation (3) and instead of $\mathrm{P}(\boldsymbol{c})$ as the prior for each causal status, we compute $\mathrm{P}(\boldsymbol{c}|\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_m])$ as follow:

$$\mathrm{P}(\boldsymbol{c}|\boldsymbol{\gamma}) = \prod_{i=1}^{m} \gamma_i^{|c_i|}(1-\gamma_i)^{1-|c_i|}$$

Conditional method for fine-mapping Here we show how to compute the statistics for the rest of the SNPs given we have selected a SNP as the causal SNP. For simplicity we only use two SNPs to compute the conditional statistics. Thus, we have :

$$(S_i|S_j = \hat{s}_j) \sim \mathcal{N}\left(\beta_i + r_{ij}\left(\hat{s}_j - \beta_j\right), 1 - r_{ij}^2\right)$$

Conditioning on one SNP is equivalent to make the statistics for that SNP equal to zero. Moreover, the variance of the remaining SNP is one. As a result,

$$(S_i^{new}|\hat{s}_j) \sim \mathcal{N}\left(\frac{\hat{s}_i - r_{ij}\hat{s}_j}{\sqrt{1-r_{ij}^2}}, 1\right)$$

We use the iterative method to obtain all the causal SNPs. In each iteration of the method we pick the SNP with the lowest p-value (the highest statistics) and re-compute the statistics of the renaming SNP using the formula mentioned above. We keep repeating this process until there exist no significant SNP. In our experiment we set the significant threshold value to 0.001.

ACKNOWLEDGMENT

F.H., E.K., E.Y.K., and E.E. are supported by National Science Foundation (NSF) grants 0513612, 0731455, 0729049, 0916676, 1065276,1302448, and 1320589 and National Institutes of Health (NIH) grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-MH101782 and R01-ES022282. We acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691). B.P is supported in part by the National Institutes of Health (R03 CA162200 and R01 GM053275). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

LITERATURE CITED

Abecasis, G., D. Altshuler, A. Auton, et al., 2010 A map of human genome variation from population-scale sequencing. Nature *467*(7319): 1061–1073.

Allen, H. L., K. Estrada, G. Lettre, et al., 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*(7317): 832–838.

Altshuler, D., M. J. Daly, and E. S. Lander, 2008 Genetic mapping in human disease. Science *322*(5903): 881–888.

Bauer, D. E., S. C. Kamran, S. Lessard, et al., 2013 An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science *342*(6155): 253–257.

Beecham, A. H., N. A. Patsopoulos, D. K. Xifara, et al., 2013 Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nature Genetics *45*(11): 1353–1360.

Bernstein, B. E., E. Birney, I. Dunham, et al., 2012 An integrated encyclopedia of DNA elements in the human genome. Nature *489*(2): 57–74.

Chen, G. K. and J. S. Witte, 2007 Enriching the analysis of genomewide association studies with hierarchical modeling. The American Journal of Human Genetics *81*(2): 397–404.

Cheung, V. G., R. S. Spielman, K. G. Ewens, et al., 2005 Mapping determinants of human gene expression by regional and genome-wide association. Nature *437*(7063): 1365–1369.

Chung, C. C., J. Ciampa, M. Yeager, et al., 2011 Fine mapping of a region of chromosome 11q13 reveals multiple independent loci associated with risk of prostate cancer. Human Molecular Genetics *20*(14): 2869–2878.

Coram, M. A., Q. Duan, T. J. Hoffmann, et al., 2013 Genome-wide Characterization of Shared and Distinct Genetic Components that Influence Blood Lipid Levels in Ethnically Diverse Human Populations. The American Journal of Human Genetics *92*(6): 904–916.

Darnell, G., D. Duong, B. Han, and E. Eskin, 2012 Incorporating Prior Information Into Association Studies. Bioinformatics *28*(12): i147–i153.

Diogo, D., F. Kurreeman, E. A. Stahl, et al., 2013 Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. The American Journal of Human Genetics *92*(1): 15–27.

Edwards, S. L., J. Beesley, J. D. French, and A. M. Dunning, 2013 Beyond GWASs: Illuminating the Dark Road from Association to Function. The American Journal of Human Genetics *93*(2): 779–797.

Eskin, E., 2008 Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. Genome Research *18*(7319): 653–60.

Faye, L. L., M. J. Machiela, P. Kraft, S. B. Bull, and L. Sun, 2013 Re-Ranking Sequencing Variants in the Post-GWAS Era for Accurate Causal Variant Identification. PLoS Genetics *9*(8): e1003609.

Flister, J., S. Tsaih, C. O'Meara, and others, 2013 Identifying multiple causative genes at a single GWAS locus. Genome Research *467*(7319): 1061–1073.

Frazer, K. A., D. G. Ballinger, D. R. Cox, et al., 2007 A second generation human haplotype map of over 3.1 million SNPs. Nature *449*(7164): 851–861.

Galarneau, G., C. D. Palmer, V. G. Sankaran, et al., 2010 Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nature Genetics *42*(12): 1049–1051.

Gibbs, R. A., J. W. Belmont, P. Hardenbol, et al., 2003 The international HapMap project. Nature *426*(6968): 789–796.

Gong, J., F. Schumacher, U. Lim, et al., 2013 Fine Mapping and Identification of BMI Loci in African Americans. The American Journal of Human Genetics *93*(4): 661–671.

Haiman, C. A., N. Patterson, M. L. Freedman, et al., 2007 Multiple regions within 8q24 independently affect risk for prostate cancer. Nature Genetics *39*(5): 638–44.

Hakonarson, H., S. F. A. Grant, J. P. Bradfield, et al., 2007 A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature *448*(7153): 591–594.

Hoffman, M. M., J. Ernst, S. P. Wilder, et al., 2012 Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Research *93*(2): 779–797.

Jul, J. H., B. Han, and E. Eskin, 2011 Increasing power of groupwise association test with likelihood ratio test. Journal of Computational Biology *18*(11): 1611–1624.

Kang, H. M., N. A. Zaitlen, C. M. Wade, et al., 2008 Efficient Control of Population Structure in Model Organism Association Mapping. Genetics *5*(4): e1000456.

Kostem, E. and E. Eskin, 2013 Improving the Accuracy and Efficiency of Partitioning Heritability into the Contributions of Genomic Regions. The American Journal of Human Genetics *92*(2): 558–564.

Kottgen, A., E. Albrecht, A. Teumer, et al., 2013 Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. Nature Genetics *45*(2): 145–154.

Lawrence, R., D. M. Evans, A. P. Morris, et al., 2005 Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants. Genome Research *15*(11): 1503–1510.

Li, B. and S. M. Leal, 2008 Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics *83*(3): 311–321.

Lippert, C., J. Listgarte, Y. Liu, et al., 2011 FaST linear mixed models for genome-wide association studies. Nature Methods *8*(7): 833835.

Listgarten, J., C. Lippert, C. M. Kadie, et al., 2012 Improved linear mixed models for genome-wide association studies. Nature Methods *9*(7): 525526.

Long, N., S. P. Dickson, J. M. Maia, et al., 2013 Leveraging Prior Information to Detect Causal Variants via Multi-Variant Regression. PLoS computational biology *9*(6): e1003093.

Lu, Y., V. Vitart, K. P. Burdon, et al., 2013 Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. Nature Genetics *45*(2): 155–163.

Madsen, B. E. and S. R. Browning, 2009 A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics *5*(2): e1000384.

Maller, J. B., G. McVean, J. Byrnes, et al., 2012 Bayesian refinement of association signals for 14 loci in 3 common diseases. Nature Genetics *44*(12): 1294–1301.

Malo, N., O. Libiger, and N. J. Schork, 2008 Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. The American Journal of Human Genetics *82*(2): 375–385.

Manolio, T. A., L. D. Brooks, and F. S. Collins, 2008 A HapMap harvest of insights into the genetics of common disease. The Journal of Clinical Investigation *118*(5): 1590–1605.

Marigorta, U. M. and A. Navarro, 2013 High trans-ethnic replicability of GWAS results implies common causal variants. PLoS Genetics *9*(6): e1003566.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, et al., 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics *9*(5): 356–369.

Navon, O., J. H. Sul, B. Han, et al., 2013 Investigations-Genetics of Complex Traits: Rare Variant Association Testing Under Low-Coverage Sequencing. Genetics *194*(3): 769–79.

Peters, U., K. E. North, P. Sethupathy, et al., 2013 A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study. PLoS Genetics *9*(1): e1003171.

Pritchard, J. K. and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics *69*(1): 1–14.

Reich, D. E., M. Cargill, S. Bolk, et al., 2001 Linkage disequilibrium in the human genome. Nature *411*(6834): 199–204.

Ripke, S., C. O'Dushlaine, K. Chambert, et al., 2013 Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nature Genetics *45*(10): 1150–1159.

Segura, V., B. J. Vilhjlmsson, A. Platt, et al., 2012 An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature Genetics *44*(7): 825 – 830.

Sklar, P., S. Ripke, L. J. Scott, et al., 2011 Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nature Genetics *43*(10): 977.

Sladek, R., G. Rocheleau, J. Rung, et al., 2007 A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature *445*(7130): 881–885.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding, 2012 Improved Heritability Estimation from Genome-wide SNPs. The American Journal of Human Genetics *91*(2): 1011–1021.

Spencer, C. C., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genetics *5*(5): e1000477.

Stahl, E. A., D. Wegmann, G. Trynka, et al., 2012 Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nature Genetics *44*(5): 483–489.

Trynka, G., K. A. Hunt, N. A. Bockett, et al., 2011 Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics *43*(12): 1193–1201.

Udler, M. S., K. B. Meyer, K. A. Pooley, et al., 2009 FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. Human molecular genetics *18*(9): 1692–1703.

Wu, M., S. Lee, T. Cai, et al., 2011 Rare variant association testing for sequencing data with the sequence kernel association test (SKAT). The American Journal of Human Genetics *89*(2): 8293.

Wu, Y., L. L. Waite, A. U. Jackson, et al., 2013 Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. PLoS Genetics *9*(3): e1003379.

Yang, J., T. Ferreira, A. P. Morris, et al., 2012 Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics *44*(4): 369–375.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics *88*(1)**:** 76–82.

Yang, J., T. A. Manolio, L. R. Pasquale, et al., 2011 Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics *43*(6)**:** 519–525.

Yi, N., N. Liu, and J. Li, 2011 Hierarchical Generalized Linear Models for Multiple Groups of Rare and Common Variants: Jointly Estimating Group and Individual-Variant Effects. PLoS Genetics *12*(7)**:** e1002382.

Zeggini, E., M. N. Weedon, C. M. Lindgren, et al., 2007 Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science *316*(5829)**:** 1336–1341.

Zhou, X. and M. Stephens, 2012 Genome-wide efficient mixed model analysis for association studies. Nature Genetics *44*(7)**:** 821–824.

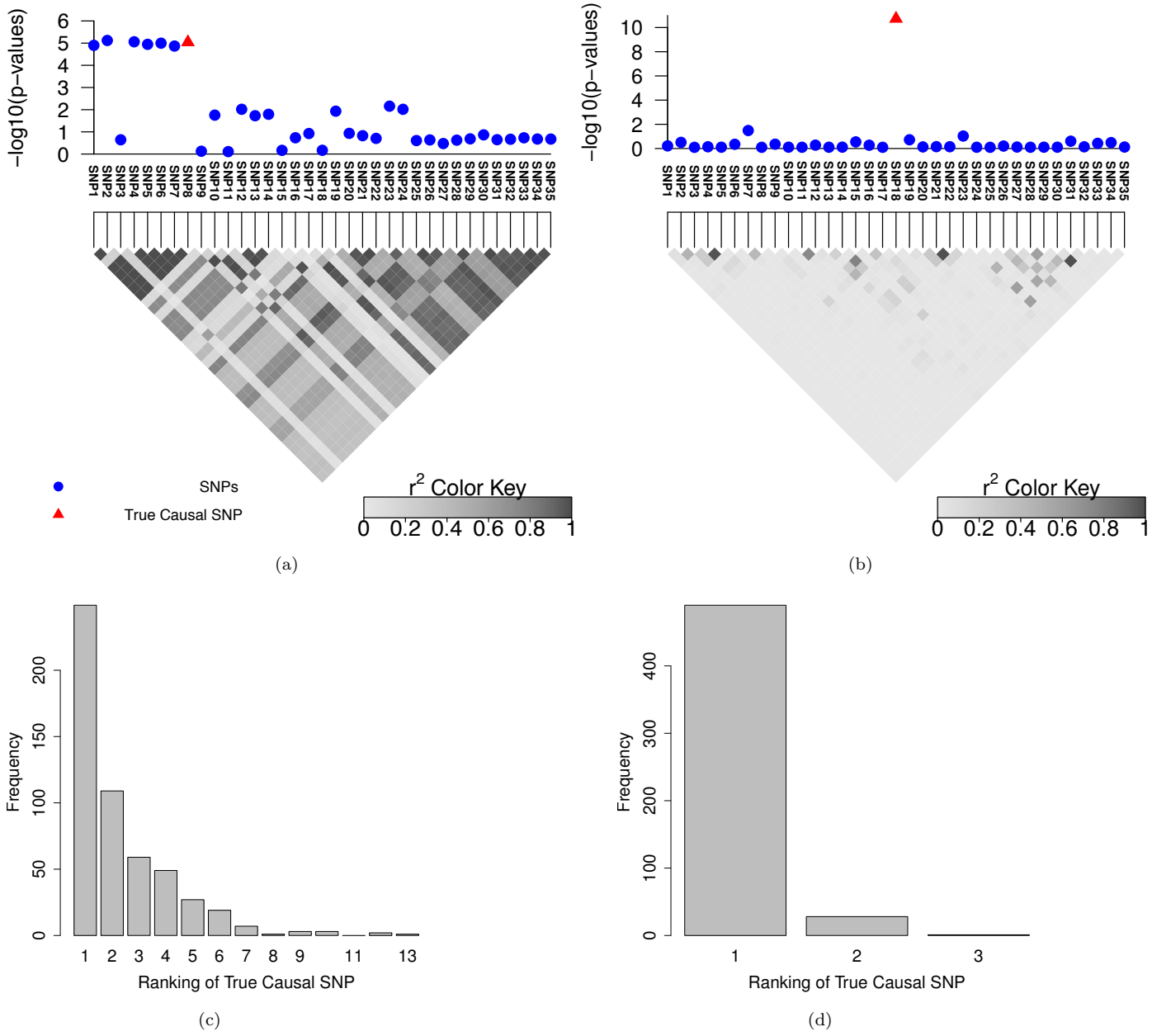
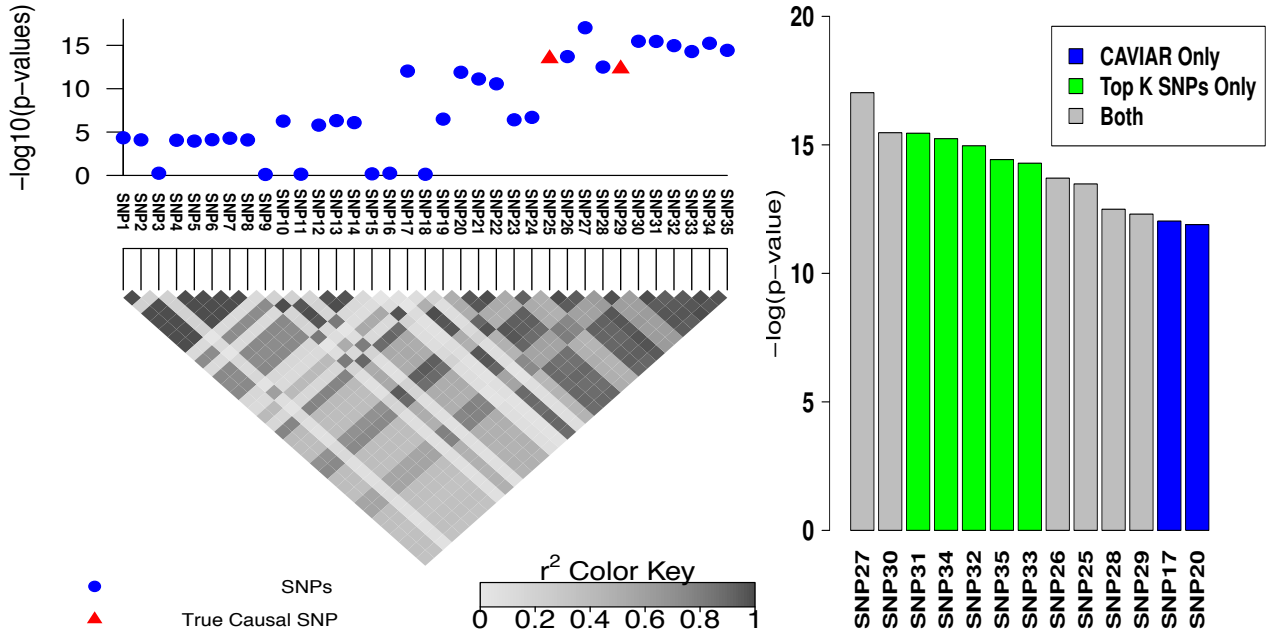
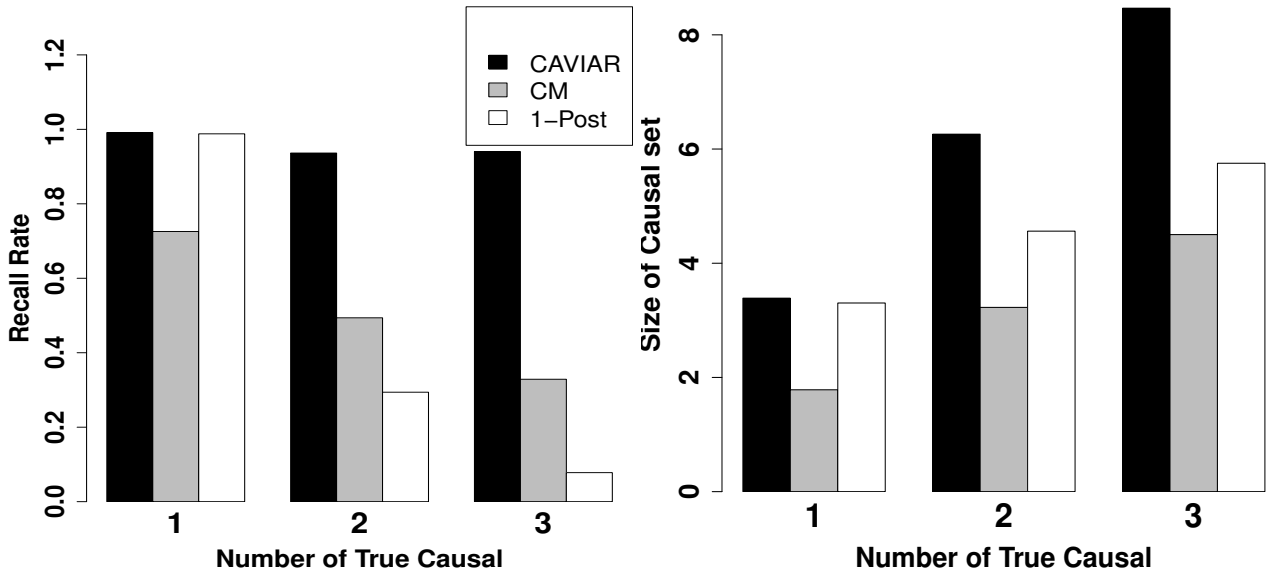


Figure 1: Panel (a) and (b) show simulated data for two regions with different LD patterns that contain 35 SNPs. Panel (a) and (b) are obtained by considering the 100Kbp from upstream and downstream of the rs10962894 and rs4740698 respectively from the WTCCC study for the coronary artery disease (CAD). Panel (c) and (d) indicate the rank of the causal SNP in additional simulations for the regions in panel (a) and (b) respectively. We obtain these histograms from simulation data by randomly generating GWAS statistics using multivariate normal distribution. We apply the simulation for 1000 times.



(a) (b)

Figure 2: Simulated association with two causal SNPs. Panel (a) shows the 100Kbp region around the rs10962894 SNP and simulated statistics at each SNP generated assuming two SNPs are causal. In this example SNP25 and SNP29 are considered as the causal SNPs. However, the most significant SNP is the SNP27. Panel (b) indicates the causal set selected by CAVIAR (our method) and top k SNPs method. We ranked the selected SNPs based on the association statistics. The grey bars indicate the selected SNPs by both methods, the green bars indicate the selected SNPs by the top k SNPs method only, and the blue bars indicate the selected SNPs by CAVIAR only. The CAVIAR set consists of SNP17, SNP20, SNP21, SNP25, SNP26, SNP28 and SNP29. In order for the top k SNPs method to capture the two causal SNPs we have to set k to 11, as one of the causal SNPs is ranked 11th based on its significant score. Unfortunately, knowing the value of k before hand is not possible. Even if the value of k is known the causal set selected by our method excludes SNP30 up to SNP35 from the follow-up studies and reduces the cost of follow-up studies by 30% compared to the top k method.



(a) (b)

Figure 3: Comparison of each method performance on the simulated GWAS data. Panel (a) illustrates the recall rate for each method and panel (b) illustrates the number of causal SNPs selected by each method. CM is the conditional method and 1-Post is the method proposed by Maller et al. (Maller *et al.* 2012). In both panels the x-axis is the true number of causal SNPs that we have implanted in each region. In the scenario of one causal SNP both our method and 1-Post have similar results as both methods use the 95% confidence interval to select a SNP as causal. However, for scenarios in which we have more than one causal SNPs, our method outperforms the 1-Post.

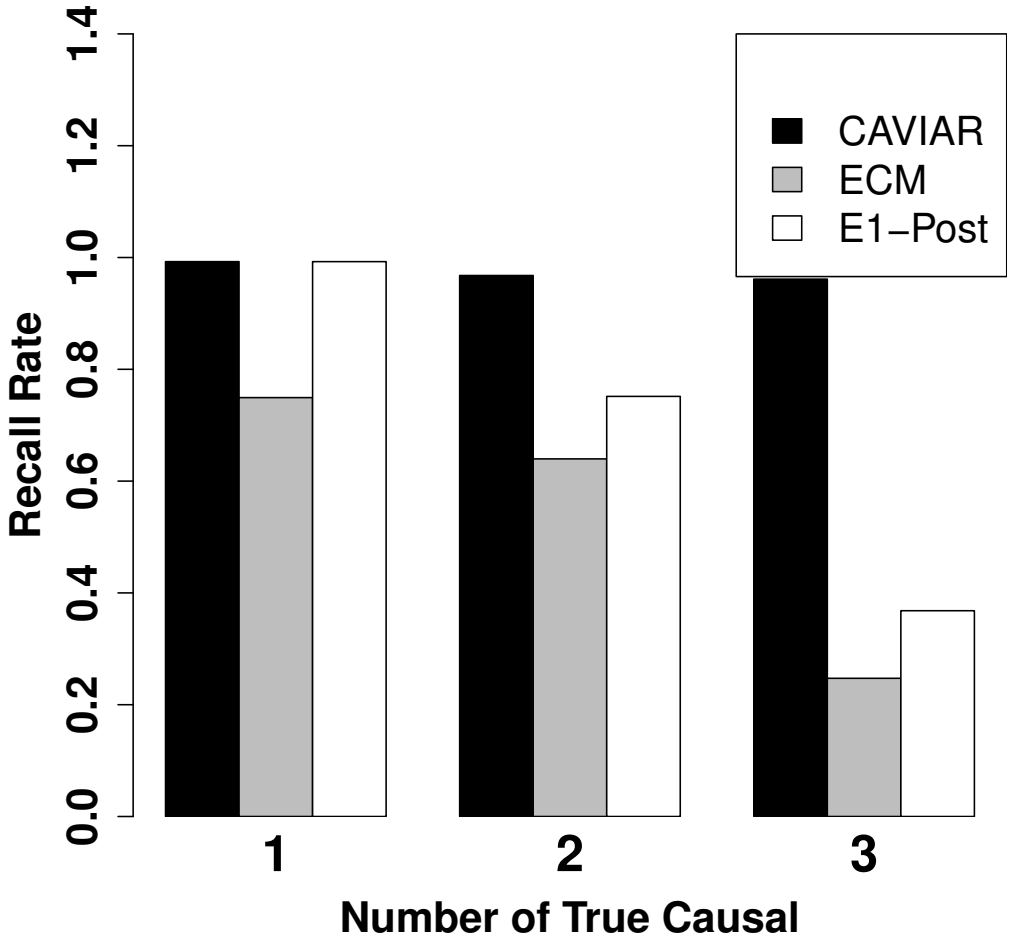
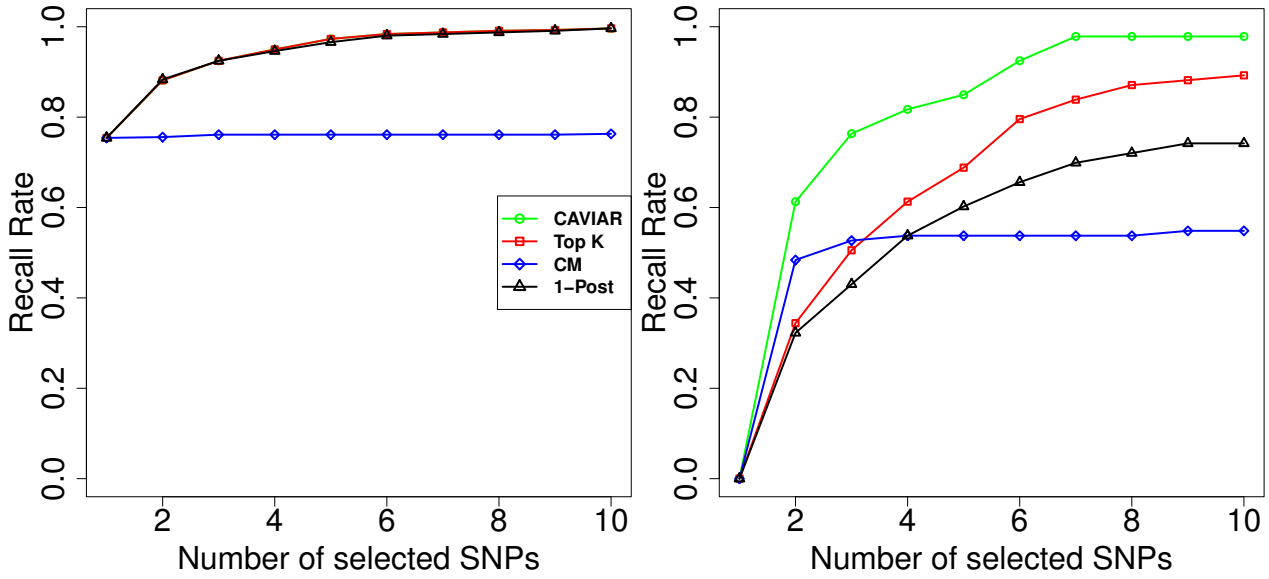
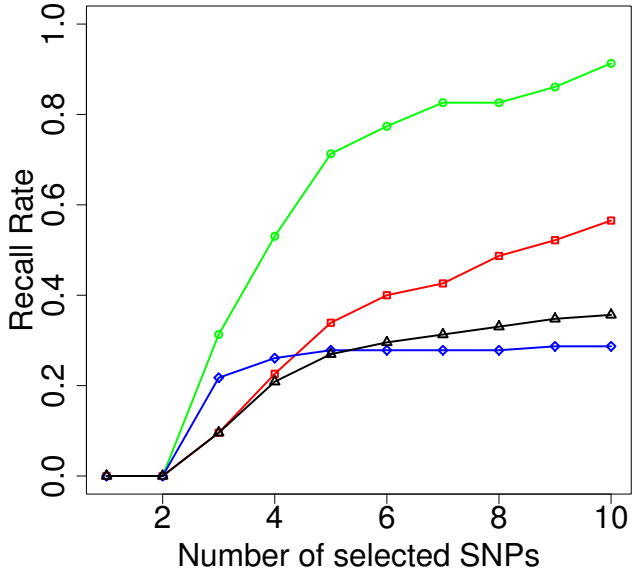


Figure 4: Comparison of recall rates. ECM and E1-Post are our extension of CM and 1-Post methods respectively, where we allow them to select the same number of causal SNPs as CAVIAR.



(a)

(b)



(c)

Figure 5: The recall rate compression for different methods while selecting the same number of causal SNPs. The x-axis is the number of SNPs selected by each method and the y-axis is the recall rate for each method. Panel (a), (b) and (c) represent the scenarios where we have implanted one, two and three causal SNPs respectively. In the scenario of only one causal SNP CAVIAR, top k SNPs, and the 1-Post method obtain similar ranking for SNPs.

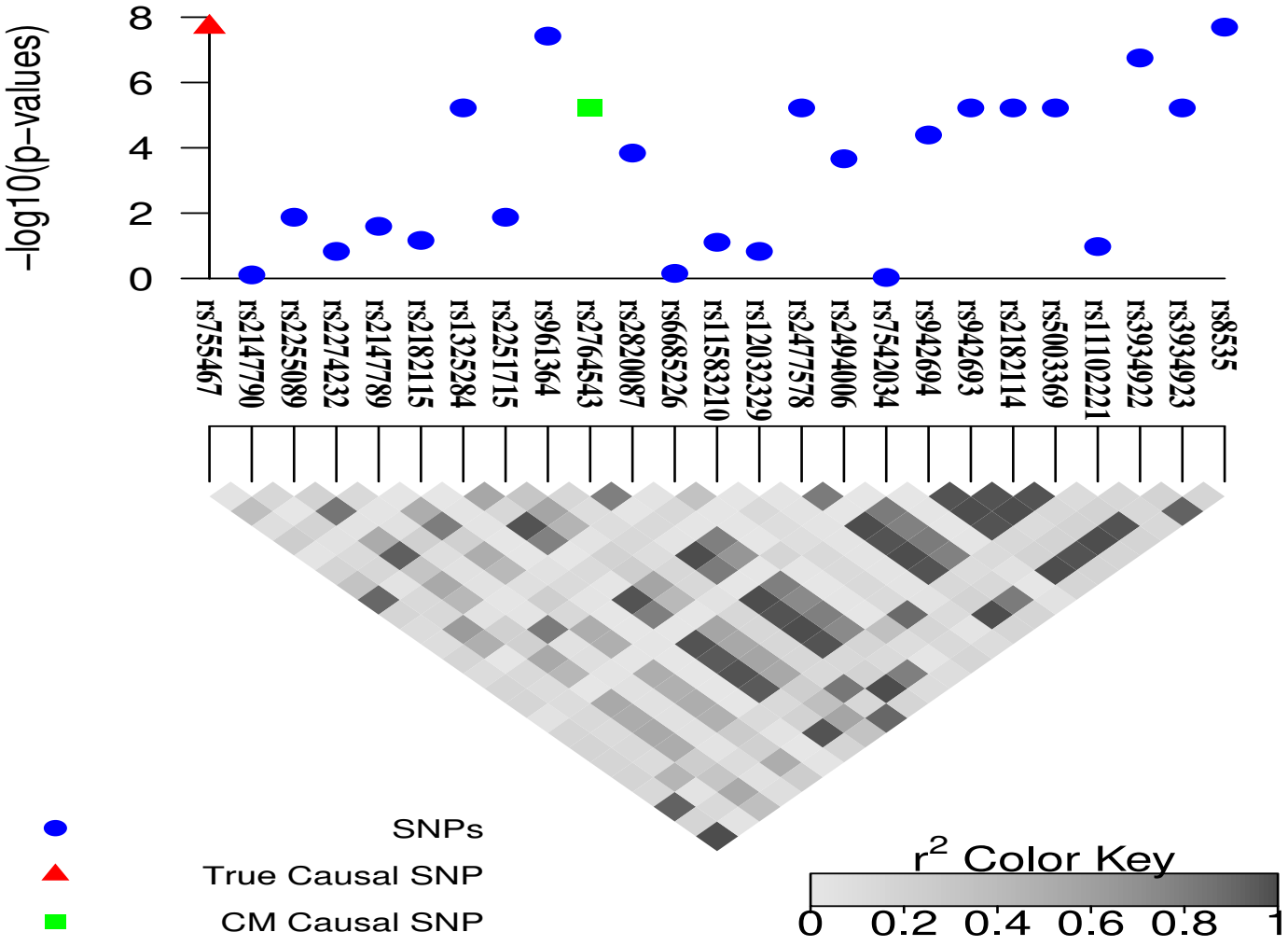


Figure 6: The 95% causal set selected by CAVIAR for the *CHI3L2* region. The red triangle represents the true causal SNP that is known using experimental methods (Cheung *et al.* 2005) and the green square represents the causal SNP detected using the CM method conditional on the true causal SNP (rs755467).