FILE S1

MATERIALS AND METHODS

Effect of different values of ρ on the causal sets

In our simulations, we used the 100kb region that contains 35 SNPs on chromosome 9, which is centered by the most significantly associated SNP (rs1333049) in the coronary artery disease (CAD) study.

In each simulation, we randomly select one of the SNPs in this region as a causal SNP and generate GWAS statistics for the 35 SNPs using our data-generating model. 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.

We illustrate the performance of our method when we have implanted one causal SNP in Table S1. We range the ρ^* from 0.5 to 0.95. Clearly, we can see as the ρ^* increases the size of the configuration set and the recall rate increase as well. It is worth mentioning the recall rate obtained from the simulation is always higher than the value of ρ^*, as ρ^* is the lower bound for the recall rate guaranteed by our method. Table S2 shows the results when we have implanted two causal SNPs in our simulation data sets.

Comparison between the exact and greedy solution

In this section we perform simulation to indicate the results obtained from the greedy method is close to the solution obtained from solving the exact posterior probability. We compared the size of causal set and the recall rate of both methods. In this simulation we use a region that consist of 15 SNPs, this region is selected from the WTCCC study (Burton, Clayton, Cardon, *et al.* 2007).

We generated the phenotypes similar to previous sections of the paper. As shown in Table S3 for different values of ρ both methods tend to have similar recall rates. Moreover, the size of the causal sets are very close, but the exact solution tends to have smaller causal set (fewer SNPs) compared to the greedy solution.

Conditional method using the marginal z-scores

Here we show how to compute the statistics for the rest of the SNPs given we have selected a SNP as the causal SNP. We use $\hat{z}_i$ and β_i to represent the marginal statistics and the SNP effects of i-th SNP. As both the phenotype and genotype for each SNP are standardized, which has mean zero and variance of one, we have $Var(\boldsymbol{x_i}) = E[\boldsymbol{x_i}^2] - E[\boldsymbol{x_i}]^2 = 1$, thus $\boldsymbol{x_i}^T\boldsymbol{x_i} = n$ where n is the number of individuals in the study. We compute the effect of i-th SNP given we have selected the j-th SNP as follows:

$$
\begin{aligned}
(\hat{\beta}_i|\hat{\beta}_j) &= (\boldsymbol{x_i}^T\boldsymbol{x_i})^{-1}\boldsymbol{x_i}^T[\boldsymbol{y} - \boldsymbol{x_j}(\boldsymbol{x_j}^T\boldsymbol{x_j})^{-1}\boldsymbol{x_j}^T\boldsymbol{y})] && (1)\\
&= (\boldsymbol{x_i}^T\boldsymbol{x_i})^{-1}\boldsymbol{x_i}^T\boldsymbol{y} - (\boldsymbol{x_i}^T\boldsymbol{x_i})^{-1}\boldsymbol{x_i}^T\boldsymbol{x_j}(\boldsymbol{x_j}^T\boldsymbol{x_j})^{-1}\boldsymbol{x_j}^T\boldsymbol{y} && (2)\\
&= \frac{\boldsymbol{x_i}^T\boldsymbol{y}}{n} - \frac{r_{ij}\boldsymbol{x_j}^T\boldsymbol{y}}{n} && (3)\\
&= \mathrm{cor}\,(\boldsymbol{x_i}, \boldsymbol{y}) - r_{ij}\,\mathrm{cor}\,(\boldsymbol{x_j}, \boldsymbol{y}) && (4)\\
&= \frac{\hat{z}_i}{\sqrt{n}} - r_{ij}\frac{\hat{z}_j}{\sqrt{n}} && (5)
\end{aligned}
$$

Where $\hat{z}_i$ is the marginal z-score for the i-th SNP, which is equal to $\mathrm{cor}\,(\boldsymbol{x_i}, \boldsymbol{y})\,\sqrt{n}$. Next, we obtain the variance of the conditional effect size using the equations 5.

$$
\mathrm{Var}(\hat{\beta}_i|\hat{\beta}_j) = \frac{1}{n} - \frac{r_{ij}^2}{n} \qquad (6)
$$

The new z-score is computed using equations 5 and 6. The new z-score is computed as follows:

$$\hat{z}_i^{new} = \frac{(\hat{\beta}_i|\hat{\beta}_j)}{\sqrt{\mathrm{Var}(\hat{\beta}_i|\hat{\beta}_j)}} = \frac{\hat{z}_i - r_{ij}\hat{z}_j}{\sqrt{1-r_{ij}^2}} \tag{7}$$

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 Equation 7. We keep repeating this process until there exist no significant SNP. In our experiment we set the significant threshold value to 0.001. This iterative process is used for the conditional method (CM).

A trade off between the number of individuals collected and the number of SNPs required validation

The number of SNPs selected by CAVIAR decreases with an increase in the number of individuals collected in each study which makes it easier to differentiate the causal SNPs from the other SNPs and this reduces the number of SNPs required to be validated.

We used HapGen (Spencer, Su, Donnelly, and Marchini 2009) to simulate fine-mapping data across European populations in the 1000 Genome project(Abecasis, Altshuler, Auton, *et al.* 2010) across regions consisting of 50 SNPs. We randomly implanted one 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 CAVIAR. We compute the average size of the causal set selected by CAVIAR. The results are shown in Figure S1.

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.

BURTON, P. R., D. G. CLAYTON, L. R. CARDON, et al., 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*(7145)**:** 661–678.

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.

Table S1: Relation between the ρ^*, configuration size and recall rate in regions with low amounts of LD. Recall rate indicates the percentage of times where we picked the true causal SNP in our configuration. The configuration size is the average number of SNPs which is predicated to be causal by our method. For each value of ρ^* we run the experiment for 1000 times.

ρ^*	**Configuration Size**	**Recall Rate(%)**
0.5	1.009862 ± 0.117203	94.67456
0.55	1.04 ± 0.1961554	97.2
0.6	1.066667 ± 0.2572115	96.19048
0.65	1.094412 ± 0.2992068	98.07322
0.7	1.108 ± 0.329474	98.8
0.75	1.136905 ± 0.3498184	99.40476
0.8	1.152642 ± 0.3599944	99.60861
0.85	1.173307 ± 0.3943774	99.8008
0.9	1.177083 ± 0.3981898	99.9
0.95	1.219665 ± 0.4484662	100

Table S2: Relation between the ρ^*, configuration size and recall rate in regions with high amounts of LD. Recall rate indicates the percentage of times where we picked the true causal SNP in our configuration. The configuration size is the average number of SNPs which is predicated to be causal by our method. For each value of ρ^* we run the experiment for 1000 times.

ρ^*	**Configuration Size**	**Recall Rate(%)**
0.5	2.149402 ± 1.047566	62.94821
0.55	2.408348 ± 1.241056	70.96189
0.6	2.663462 ± 1.532	75.96154
0.65	2.921642 ± 1.452493	79.29104
0.7	3.28839 ± 1.716047	81.64794
0.75	3.64497 ± 2.10312	86.39053
0.8	3.978102 ± 2.067303	89.59854
0.85	4.684601 ± 2.73976	93.32096
0.9	5.121377 ± 2.78669	96.37681
0.95	6.598058 ± 3.598475	98.83495

Table S3: Comparison between the solution obtained from solving the posterior probability exactly or using the greedy method.

ρ^*	**Exact Solution**		**Greedy Solution**	
	Configuration Size	**Recall Rate(%)**	**Configuration Size**	**Recall Rate(%)**
0.5	2.025097 ± 0.8759341	67.3	2.015355 ± 0.9007232	67.8
0.55	2.581 ± 0.9276084	70.7	2.132411 ± 1.100085	79.5
0.6	2.420152 ± 1.076278	79.4	2.433962 ± 0.784	78.6
0.65	2.674721 ± 1.225183	81.2	2.750469 ± 1.187191	81.8
0.7	2.82397 ± 1.203071	85.2	2.811429 ± 1.218756	85.4
0.75	3.091085 ± 1.416079	87.4	3.124314 ± 1.37983	87.8
0.8	3.317526 ± 1.598748	91.1	3.274583 ± 1.554082	91.5
0.85	3.514395 ± 1.633862	93.2	3.537402 ± 1.570367	92.7
0.9	3.887064 ± 1.934519	95.6	3.859345 ± 1.922601	96.3
0.95	4.277992 ± 1.968794	99.6	4.165692 ± 1.938969	99.8

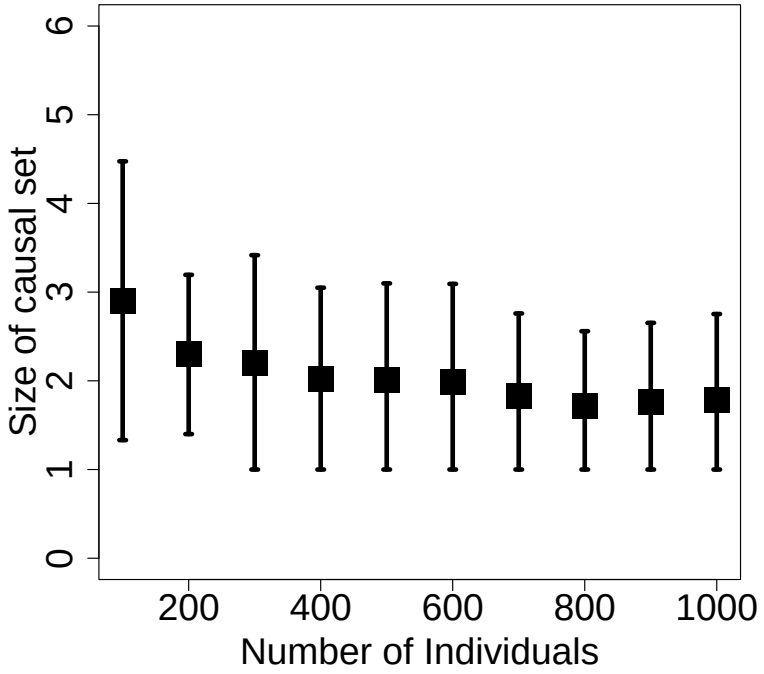

Figure S1: The patterns between the number of individuals collected in each study and the number of causal SNPs selected by CAVIAR. The black squares indicate the mean and the vertical lines indicate the standard deviation of the number of SNPs selected by CAVIAR.