Effective Genetic-Risk Prediction Using Mixed Models

David Golan[1,2,*] and Saharon Rosset[1,*]

For predicting genetic risk, we propose a statistical approach that is specifically adapted to dealing with the challenges imposed by disease phenotypes and case-control sampling. Our approach (termed Genetic Risk Scores Inference [GeRSI]), combines the power of fixed-effects models (which estimate and aggregate the effects of single SNPs) and random-effects models (which rely primarily on whole-genome similarities between individuals) within the framework of the widely used liability-threshold model. We demonstrate in extensive simulation that GeRSI produces predictions that are consistently superior to current state-of-the-art approaches. When applying GeRSI to seven phenotypes from the Wellcome Trust Case Control Consortium (WTCCC) study, we confirm that the use of random effects is most beneficial for diseases that are known to be highly polygenic: hypertension (HT) and bipolar disorder (BD). For HT, there are no significant associations in the WTCCC data. The fixed-effects model yields an area under the ROC curve (AUC) of 54%, whereas GeRSI improves it to 59%. For BD, using GeRSI improves the AUC from 55% to 62%. For individuals ranked at the top 10% of BD risk predictions, using GeRSI substantially increases the BD relative risk from 1.4 to 2.5.

Introduction

Despite the huge investment and considerable progress in the study of the genetic causes of human diseases, the underlying genetic mechanisms of many common diseases, including type 1 diabetes (T1D), bipolar disorder (BD), schizophrenia, multiple sclerosis, and Alzheimer disease, are still largely unknown. The leading methodology for finding the genetic causes of disease is the genome-wide association study (GWAS). In a typical GWAS, one collects thousands of sick and healthy individuals, genotypes them, and searches for SNPs that are more abundant in one group or the other. To date, GWASs have flagged thousands of SNPs as associated with hundreds of diseases. However, our ability to accurately predict an individual's disease status on the basis of these SNPs still falls considerably short of what is expected given the high heritability of these diseases.

This remarkable gap between the predictive power of significantly associated SNPs and the expected predictive capacity based on the high heritability of the phenotypes has been termed the "mystery" or "problem" of the missing heritability. One leading theory attempting to explain this mystery is that many phenotypes are driven by a plethora of common SNPs with small effects and that present-day GWASs are underpowered to detect these SNPs because of their small effects. Goldstein[1] estimated the overall number of SNPs affecting height at 93,000.

In light of this theory, the traditional naive approach of using only the SNPs found to be significantly associated with the disease in calculating genetic-risk scores (GRSs) is expected to perform poorly because it overlooks the lion's share of causal SNPs, whose effect is not large enough to be declared significant. Instead, recent efforts in computing GRSs have attempted to include a larger number of SNPs, primarily by adopting much more lenient inclusion criteria for SNPs.

Using a more permissive threshold has two effects: (1) capturing more "true" causal signal through the inclusion of causative SNPs with small effects and (2) increasing the noise in the genetic prediction, given that every additional parameter estimated (for either a true association or a false-positive signal) adds uncertainty to the predictions. A good choice of p value threshold would be such that the trade-off between signal and noise is beneficial. Recent studies have demonstrated that GRSs computed in this manner have a significant association with disease—one above and beyond that of GRSs computed with only significantly associated SNPs—when up to half of the genotyped SNPs are included (see, e.g., Purcell et al.[2]). These results suggest that for at least some diseases, there is considerable information in the long tail of insignificantly associated SNPs and that the benefit from including more true positives trumps the cost of estimating more parameters. Dudbridge[3] and Chatterjee et al.[4] provide an in-depth mathematical analysis of this approach and variants thereof, and it was recently applied to predicting risk of Celiac disease with remarkable success.[5]

These approaches for computing GRSs fall under the category known in the statistics literature as "fixed-effects" modeling. In such models, the effects of SNPs are assumed to be parameters (i.e., fixed but unknown quantities). These parameters are estimated and used in subsequent analysis. For example, one would estimate the odds ratio of a given SNP from a GWAS and use this estimate to predict the risk of new individuals. The main difference between the methods lies in the way these parameters are estimated, ranging from simple SNP-by-SNP regression to shrinkage-based estimates, such as Lasso and more, as reviewed by Dudbridge[3] and Abraham et al.[5] They also differ

[1]Department of Statistics, Tel Aviv University, Tel Aviv 69978, Israel
[2]Present address: Department of Genetics, Stanford University, Stanford, CA 94305, USA
*Correspondence: golandavid@gmail.com (D.G.), saharon@post.tau.ac.il (S.R.)
http://dx.doi.org/10.1016/j.ajhg.2014.09.007.

in the way SNPs are chosen to be included in computation of the risk scores.[3,5] However, they all share the fundamental treatment of the effects as parameters that require estimation.

An alternative approach to computing GRSs is "random-effects" modeling. The basic premise of this approach is that our goal is not to estimate the individual effect of every SNP but rather to estimate their cumulative effect. Hence, it is expected to be beneficial to circumvent estimating each and every effect and instead target this cumulative effect directly. To accomplish this goal, effect sizes are treated not as parameters but rather as random variables with some common distribution. They can then be "integrated out," thus mitigating the need to estimate them separately. Instead, a correlation (or kinship) matrix G is estimated with the genotypes and models the correlations between the GRSs. The correlation between GRSs of individuals who are more genetically similar would be higher, and vice versa.

This random-effects approach has been adopted in the context of GWASs for association tests[6–10] and heritability estimation[11–14] with much success. All of these approaches rely on treating the phenotype as a normally distributed variable, which is the sum of a genetic component and an environmental component, and utilizing well-established linear-mixed-model (LMM) methodologies to draw inferences about their quantities of interest.

In the context of risk prediction, the animal-breeding literature has long used similar approaches to model and estimate the "breeding value," which is closely related to the genetic risk. In the scenario of an observational study of a quantitative phenotype, a well-established methodology for estimating the breeding value is known as the best linear unbiased predictor, or BLUP.[15,16] When breeding values are estimated, pedigree data are usually available. In GWASs, the kinship can be estimated from genotype data (referred to as genetic BLUP, or gBLUP[17]), and this method is implemented in the widely used GCTA software.[18] It was recently extended in various ways,[19,20] resulting in a considerable improvement in prediction accuracy.

However, case-control studies present a much more challenging statistical setup. First, the phenotype is binary rather than quantitative and so cannot be accurately modeled by a multivariate normal (MVN) distribution. Additionally, affected individuals (cases) are highly overrepresented in the sample, in comparison to the population, and so many of the typical statistical assumptions (namely normality and independence of the genetic and environmental effects) are no longer legitimate.

A common approach to random-effects modeling in case-control GWASs is to treat the phenotype as quantitative and apply LMM methodologies, possibly followed by post-hoc corrections to account for violation of the underlying assumptions.[12,13,19] Although this approach has proven successful in practice, its reliance on probabilistic models that are known to be inaccurate is expected to result in suboptimal performance. In the context of GRS estimation, the natural extension of the LMM approach to case-control data is to use gBLUP and its extensions, but this is subject to similar inaccuracy concerns, as our simulations below demonstrate.

We describe a statistical approach for Genetic Risk Scores Inference (GeRSI). GeRSI is based on a Markov-chain Monte-Carlo (MCMC) method utilizing Gibbs sampling to estimate the GRSs of individuals given the genotypes of a case-control study under a random-effects model. We use the well-known normal liability-threshold model to account for the dichotomous nature of the phenotype. Additionally, our Gibbs-sampling approach conditions explicitly on the selection of individuals to the study and thus accounts directly for the overrepresentation of the case group in the study. By properly conditioning on the selection, we can sample from the true posterior distribution of the GRS. This is in contrast to using LMM-based approaches, which treat a case-control disease phenotype as if it were a randomly sampled quantitative one.

In addition to accounting for disease phenotypes and nonrandom selection in prediction, GeRSI naturally accommodates fixed effects within the probabilistic framework. Hence, our approach also allows "mixed-effects" modeling, where SNPs with considerable effects can be included as fixed effects and the long tail of insignificantly associated SNPs is accounted for with random effects. We distinguish between random-effects GeRSI (which treats all SNPs as random effects) and mixed-effects GeRSI (which, in our basic implementation, includes SNPs below a certain p value threshold as fixed effects and treats the rest of the SNPs as random effects). Additionally, introducing fixed effects to the model allows accounting for additional covariates such as sex, ethnicity, and known environmental risk factors (e.g., smoking habits). Mixed-effects GeRSI can also utilize other schemes for selecting fixed-effects SNPs and estimating their effects (such as Lasso[21]) or including covariate effects estimated from published data.[22] Hence, it can combine state-of-the-art approaches for fixed-effects estimation with proper inference on random effects.

Material and Methods

Generative Model of a Polygenic Disease

A polygenic quantitative trait y is typically modeled with the following additive model:

$$y_i = \mu + \sum_{j \in C} z_{ij} u_j + e_i,$$

where C is the set of causal SNPs, u_i is the effect of the i^{th} causal SNP, e_i is the environmental effect associated with individual i, and z_{ij} is the genotype of the j^{th} SNP of the i^{th} individual. The term $\sum_{j \in C} z_{ij} u_j$ is often referred to as the genetic effect and denoted g_i. Under mild independence assumptions, we have $\sigma_g^2 = \mathrm{Var}(g) = |C| \sigma_u^2$.

We note that the choice to use standardized SNPs rather than just centering the SNPs hides an implicit assumption that SNPs with lower frequencies have larger effect sizes (as noted in Zhou et al.[13]). However, because we focus on common SNPs, the effects of this assumption are minimal.

Polygenic disease phenotypes are modeled with the liability-threshold model.[12] We assume the existence of a latent quantitative "liability" phenotype. If an individual's liability exceeds a certain threshold, she is part of the case group. The liability is modeled as a quantitative trait with mean 0 and variance 1. Under these assumptions, the threshold corresponding to disease prevalence K is $\Phi^{-1}(1-K) = Z_{1-K}$, where Φ is the cumulative distribution function of the standard normal distribution and Φ^{-1} is its inverse (percentiles of the distribution), often denoted Z.

Random-Effects Modeling

We are interested in predicting risk, and the actual values of u_i are not of direct interest. We therefore model them as random variables drawn from a distribution with mean 0 and variance $\sigma_g^2/|C|$. Typically, one assumes a normal distribution (e.g., Yang et al.[11] and Zhou et al.[13]), but other distributions were suggested as well as mixture distributions.[4,13] We note that as long as the number of causal SNPs is large enough and the effect sizes are independent of each other, the genetic effect approximately follows a normal distribution, regardless of the underlying distribution of the effect sizes, by virtue of the central-limit theorem.

Assuming normality of the random effects and random sampling, this implies the following MVN distribution of the liabilities: $l \sim \mathrm{MVN}(0; G\sigma_g^2 + I\sigma_e^2)$, where G is the correlation matrix of the genetic effects and is given by

$$G_{ij} = \frac{1}{|C|}\sum_{k\in C} z_{ik}z_{jk},$$

and it is assumed without loss of generality (because l is unobserved) that $\sigma_g^2 + \sigma_e^2 = 1$. Given that the identity of the causal SNPs is unknown and additionally they are often not genotyped, we follow previous works[11,12] and estimate G by using all genotyped SNPs and use this estimate throughout. This matrix is often referred to as the observed kinship matrix. The estimation of G is the subject of much recent research (see, e.g., Golan and Rosset,[14] Speed et al.,[23] and Crossett et al.[24]) but is out of the scope of the current paper.

GeRSI Sampling Scheme

Assume that we have a group of n individuals with known genotypes but that the phenotypes are known only for the first $n-1$ individuals. We are interested in predicting the genetic risk of the n^{th} individual.

We denote g and e the vectors of latent genetic and environmental effects, respectively. The heritability (and hence σ_g^2) is assumed to be known and in practice can be estimated directly from the data[11,12] or obtained from family studies.

Our goal is to predict $P(l_n > t)$, conditional on our entire data (namely the genotypes of all n individuals and the phenotypes of the first $n-1$ individuals). Had we known g_n, the (optimal) risk prediction, r_n, under the model would have been

$$r_n = P(l_n > t \,|\, g_n) = 1 - \Phi\left(\frac{t-g_n}{\sigma_e}\right).$$

However, because g_n is unknown, we generate k samples, $g_{n,1},\ldots,g_{n,k}$, from the posterior distribution of g_n, conditional on all the observed data, and estimate the risk as

$$\hat{r}_n = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \Phi\left(\frac{t-g_{n,i}}{\sigma_e}\right)\right).$$

To generate samples from the posterior distribution of g_n, we note that

$$\begin{aligned} P\left(g_n \,|\, y_{-n}; G, \sigma_g^2\right) &= P\left(g_n \,|\, g_{-n}, y_{-n}; G, \sigma_g^2\right) P\left(g_{-n} \,|\, y_{-n}; G, \sigma_g^2\right) \\ &= P\left(g_n \,|\, g_{-n}; G, \sigma_g^2\right) P\left(g_{-n} \,|\, y_{-n}; G, \sigma_g^2\right). \end{aligned}$$

In other words, sampling the posterior can be decomposed into two separate problems. The first problem is the problem of sampling g_n given the values of the other genetic effects, g_{-n}. Because we are conditioning on the genetic effects, g_n is independent of the phenotypes of the other individuals.

As we show below, even when the sampling is not random (e.g., in case-control studies), the conditional distribution of g_n is given by

$$g_n \,|\, g_{-n}; G, \sigma_g^2 \sim \mathrm{MVN}\left(G_{n,-n}G_{-n,-n}^{-1}g_{-n}, \sigma_g^2\left(G_{n,n} - G_{n,-n}G_{-n,-n}^{-1}G_{-n,n}\right)\right),$$

where positive or negative indices indicate the extraction or removal, respectively, of rows or columns.

The second problem is the problem of sampling from $g_{-n}|y; G, \sigma_g^2$, which is more involved. We introduce another set of variables, namely the environmental effects e_{-n}. It is then possible to write down the conditional distribution of each variable in the set (g_{-n}, e_{-n}), conditional on the rest of the variables in the set.

The knowledge of the phenotype induces dependence between g_i and e_i, given that knowing the phenotype implies that we have an upper or lower bound on their sum (the sum is either above or below the threshold, depending on the phenotype). Additionally, e_i is independent of the other environmental effects and is also independent of the other genetic effects conditional on g_i. Hence,

$$e_i \,|\, g, e_{-i}, y_i; \sigma_g^2 = e_i \,|\, g_i, y_i; \sigma_g^2 = \begin{cases} \sigma_e Z \,|\, \sigma_e Z + g_i > t & y_i = \text{case} \\ \sigma_e Z \,|\, \sigma_e Z + g_i < t & y_i = \text{control} \end{cases},$$

where $Z \sim N(0,1)$ (hence, the distribution is simply a truncated normal distribution). Intuitively, when g_i is known and the phenotype is known, the posterior distribution of the environmental effect is a truncated normal distribution with mean 0, variance σ_e^2, and a truncation point above or below $t - g_i$, depending on whether i is in the case or control group.

The conditional distribution of g_i is slightly more complicated because it depends on the other genetic effects via the correlation between genetic effects. Thus, we need to explicitly describe its conditional distribution, conditional on the other g_{-i} genetic effects and all environmental effects. By the independence assumption between genetic and environmental effects, we only need to consider dependence on e_i and the other genetic effects (g_{-i}) via the correlation between genetic effects.

Denote μ_i and σ_i^2 the mean and variance of g_i, conditional on g_{-i}, respectively. Then, $\mu_i = G_{i,-i}G_{-i,-i}^{-1}g_{-i}$ and $\sigma_i^2 = \sigma_g^2(G_{ii} - G_{i,-i}G_{-i,-i}^{-1}G_{-i,i})$.

Similarly to the conditional distribution of the environmental effects, conditioning on the phenotype results in a truncation of the aforementioned normal distribution:

$$g_i \,|\, e_i, g_{-i}, y_i; \sigma_g^2 = \begin{cases} \mu_i + \sigma_i Z \,|\, \mu_i + \sigma_i Z + e_i > t & y_i = \text{case} \\ \mu_i + \sigma_i Z \,|\, \mu_i + \sigma_i Z + e_i < t & y_i = \text{control} \end{cases}.$$

Again, the conditional distribution can be seen as a truncated normal distribution, but with a mean term and variance term that capture the influence of the other genetic effects on the genetic effect in question.

Once all of the conditional distributions are specified, Gibbs sampling[25] can be used to draw samples from the posterior distribution of g_{-n}, and the risk is estimated as described above. This is done is a similar fashion to Campbell et al.[26]

Conditional Sampling in Case-Control Studies

The fact that the observed samples are obtained via a case-control sampling scheme and are therefore not a random sample from the population renders the usual mixed-effects model incompatible. In particular, under assumptions of (1) normality of genetic and environmental effects in the population and (2) independence of the genetic and environmental effects in the population, the actual distribution of genetic and environmental effects in the study is nonnormal, and they are not independent as a result of selection, as noted in Lee et al.[12] Because the distribution of the genetic effects is no longer normal, their joint distribution is no longer MVN, and so naive application of Gibbs sampling might be inaccurate. However, we show here that the same sampling scheme can be used to sample the posterior genetic effects in a case-control study.

To model and account for the effects of selection, we define an event S, which signifies that individuals 1,…,$n-1$ were selected via a case-control scheme and not by random sampling. Hence, the conditional distributions above now require additional conditioning on S. However, we note that

$$f\left(g_n \mid g_{-n}, \mathcal{S}; G, \sigma_g^2\right) = \frac{f\left(S \mid g_{-n}, g_n; G, \sigma_g^2\right) f\left(g_n \mid g_{-n}; G, \sigma_g^2\right)}{f\left(S \mid g_{-n}; G, \sigma_g^2\right)},$$

but because S signifies only the selection of individuals 1,…,$n-1$, it is independent of g_n. Hence, $f(S|g_{-n}, g_n; G, \sigma_g^2) = f(S|g_{-n}; G, \sigma_g^2)$ and $f(g_n|g_{-n}, \mathcal{S}; G, \sigma_g^2) = f(g_n|g_{-n}; G, \sigma_g^2)$, i.e., Gibbs sampling of the genetic effect of the individual in question can be carried out as if there were no selection, given that the samples of the genetic effects $g_1,\ldots,g_{n-1}$ are drawn by correct conditioning on S. Moreover, for an individual i in the reference group, we have

$$f\left(g_i \mid e_i, g_{-i}, y_i, \mathcal{S}; \sigma_g^2\right) = \frac{f\left(\mathcal{S} \mid e_i, g_{-i}, y_i, g_i; \sigma_g^2\right) f\left(g_i \mid e_i, g_{-i}, y_i; \sigma_g^2\right)}{f\left(\mathcal{S} \mid e_i, g_{-i}, y_i; \sigma_g^2\right)},$$

but because we assumed that the selection is driven only by the phenotypes, we have $f(\mathcal{S}|e_i, g_{-i}, y_i; \sigma_g^2) = f(\mathcal{S}|y_i; \sigma_g^2)$ and $f(\mathcal{S}|e_i, g_{-i}, y_i, g_i; \sigma_g^2) = f(\mathcal{S}|y_i; \sigma_g^2)$, so again the sampling boils down to the same Gibbs scheme. Lastly, we need to take care of the sampling of the environmental effects, but because this is done per individual, the selection has no effect. To conclude, the same Gibbs sampling scheme can be applied to case-control studies and yield correct posterior risk estimates.

Simulation Setup

Our simulations adopt the "spike-and-slab" model of genetic risk, recently explored by Zhou et al.[13] and Chatterjee et al.[4] and found to provide a good fit for the observed effect sizes for a wide variety of GWASs. In this model, all SNPs have effects on the phenotype, but the SNPs are divided into a small fraction of "slab" SNPs with considerable effect sizes and a bulk of "spike" SNPs with very small but nonzero effects.

Given the prevalence of a disease in the population (K), the desired proportion of cases in the study (P), the desired study size (n), the total number of SNPs (m), the proportion of slab SNPs (π_1), the overall variance of the genetic effects (σ_g^2), and the fraction of the heritability explained by the slab SNPs (f_{slab}), we simulated data sets by using the following procedure:

1. We randomly sampled the minor allele frequencies (MAFs) of m SNPs from $U[0.05,0.5]$.
2. We randomly sampled SNP effect sizes for $\pi_1 m$ slab SNPs and $(1-\pi_1)m$ spike SNPs from $N(0, f_{\text{slab}}\sigma_g^2/(\pi_1 m))$ and $N(0, (1-f_{\text{slab}})\sigma_g^2/((1-\pi_1)m))$, respectively.
3. For each individual, we (a) randomly generated a genotype by using the MAFs, (b) computed the genetic effect as described above, (c) sampled an environmental effect from $N(0, 1-\sigma_g^2)$, (d) computed liability and phenotype, and (e) if the phenotype was in the case group, automatically included the individual in the study. Otherwise, we included the individual in the study with probability $K(1-P)/(P(1-K))$ to maintain the expected proportion of cases in the study at P.
4. We repeated steps 2 and 3 until n individuals were accumulated.

Setting $f_{\text{slab}} = 1$ results in a model where only π_1 of the SNPs are causal and the rest of the SNPs have no effects on the phenotype.

We note that our choice of working with SNPs in linkage equilibrium was motivated by a result of Patterson et al.[27] They showed that for the purpose of generating correlation matrices, using SNPs in linkage disequilibrium (LD) is equivalent to using a smaller number of SNPs in linkage equilibrium. They also suggested a method for estimating the effective number of SNPs (i.e., the number of SNPs in linkage equilibrium that lead to the same distribution of correlation matrices as a given set of SNPs in LD). We thus find that our simulations using $m = 50{,}000$ SNPs in equilibrium are of realistic size.

Computing GRSs with "Standard" Fixed-Effects Models

To compute GRSs with a fixed-effects approach, we follow the spirit of Dudbridge[3] and Chatterjee et al.[4] For each SNP, we estimate the effect size $\hat{u}$ by using univariate linear regression. Denote by ν_i the p value of the null hypothesis of $u_i = 0$. We then define the estimated risk score by using a p value threshold c as

$$\text{risk score } (c)_j = \sum \hat{u}_i Z_{ij} \mathbb{I}\{\nu_i < c\}.$$

When dealing with real data, where LD structure is present, we select a subset of significant SNPs—such that the distance between included SNPs is at least 1 Mb—by choosing the SNP with the lower p value within any such window. We note that other alternative definitions exist, e.g., using shrinkage estimates, but generally there is very little difference between the methods, as noted by Dudbridge.[3] For the real data, we try several p value thresholds, namely 5×10^{-c} for $c \in \{1, \ldots, 8\}$. We then choose the threshold that maximizes the area under the ROC curve (AUC). Hence, AUC estimates of the fixed-effects model are expected to be slightly elevated. Our bootstrap scheme for computing confidence intervals accounts for this selection scheme, as detailed below.

Random- and Mixed-Effects GeRSI

We use the genotyped SNPs to estimate the genetic correlation matrix G. When applying random-effects GeRSI, we use this matrix as the correlation matrix in the sampling scheme described above. When applying mixed-effects GeRSI with a p value threshold c, we keep only SNPs with a p value below that threshold in the univariate association test. We then use logistic regression to estimate the personal in-study risk due to the fixed effects. We then convert this risk to the liability scale by using the following transformation:

$$\widehat{t}_i = \Phi\left(1 - \frac{C\widehat{P}_i}{1 + C\widehat{P}_i - \widehat{P}_i}\right),$$

where $\widehat{P}_i$ is the estimated in-study risk, $\widehat{t}_i$ is the individual-specific liability threshold, and $C = K(1-P)/(P(1-K))$. We note that this method is reminiscent of the method of Zaitlen et al.[22] but differs in the data utilized for estimating the covariate effects, given that Zaitlen et al.[22] take advantage of external data for this purpose.

gBLUP

For comparison, we compute the gBLUP by coding the discrete phenotype as 0/1 and treating it as a quantitative and randomly sampled phenotype. In other words, the phenotype is modeled as

$$y \sim \mathrm{MVN}(0, \Sigma),$$

where $\Sigma = G\sigma_g^2 + I\sigma_e^2$. Hence, if the phenotype of individual i is unknown, then the conditional mean of her phenotype is easily obtained with the formula for the conditional mean of a MVN distribution:

$$\mathbb{E}\left[y_i \,|\, y_{-i}; \Sigma, \sigma_g^2\right] = \Sigma_{i,-i}\Sigma_{-i,-i}^{-1} y_{-i}.$$

The gBLUP method could be similarly extended to account for fixed effects (see, e.g., Yang et al.[18]).

Extending to Multiple Correlation Matrices

Recently, Speed and Balding[19] extended gBLUP to account for multiple correlation matrices ("multiBLUP"). The SNPs are divided into k subsets according to some criteria (e.g., functional annotation), and correlation matrices $G_1,\ldots,G_k$ are estimated for each subset separately. The corresponding variances of the effect sizes, $\sigma_{g_1}^2, \ldots, \sigma_{g_k}^2$, are either estimated or taken from published sources, as before. This formulation allows SNPs from different sets to have a typically larger or smaller effect on disease risk. The correlation of the genetic effect is then given by

$$G^{\mathrm{multi}} = \sum_{i=1}^{k} G_i \sigma_{g_i}^2.$$

MultiBLUP is defined as running gBLUP with G^{multi} instead of the previously defined G. We similarly extend GeRSI to "multi-GeRSI," i.e., running random-effects GeRSI with G^{multi}.

Controlling for Population Structure

When attempting to control for population structure in association studies or heritability estimation, it is customary to include several top principal components as covariates. However, in the context of risk prediction, this would result in inflated estimates of the predictive accuracy. Instead, we remove the top k principal components directly from the correlation matrix. In other words, denote $\lambda_1,\ldots\lambda_n$ the sorted eigenvalues of the correlation matrix G and denote $v_1,\ldots,v_n$ their corresponding eigenvectors. We define a cleaned correlation matrix as

$$G^*(k) = G - \sum_{i=1}^{k} \lambda_i v_i v_i^{\top}.$$

Describing Real Data

We obtained genotypes and phenotypes from the Wellcome Trust Case Control Consortium (WTCCC). Following Lee et al.[12] we applied a stringent quality-control (QC) process to the WTCCC data to avoid overestimation of the predictive capacity due to genotyping differences between case and control groups or between the different control groups. We removed SNPs with a MAF $<$ 5%, SNPs with a missing rate $>$ 1%, and SNPs that displayed a significantly different missing rate between case and control groups (p value $<$ 0.05). We also removed SNPs that deviated from Hardy-Weinberg equilibrium in the control groups (p value $<$ 0.05) or were noted for "bad clustering" in the genotype-calling step. Additionally, we removed SNPs that displayed a significant difference in frequency between the two control groups. Only autosomal chromosomes were included in the analysis. We removed all the individuals appearing in the WTCCC exclusion lists. These included duplicate samples, first- or second-degree relatives, individuals not of European descent, and other reasons. In addition, we removed individuals with a missing rate $>$ 1% and all individual pairs with an estimated genetic correlation $<$ 0.05 according to the correlation matrix. We performed the last step to ensure that individuals in the study were not closely related. In addition, when computing the correlation matrix G, we used estimated MAFs from HapMap's CEU panel (Utah residents with ancestry from northern and western Europe from the CEPH collection)[28] to mitigate any possibility of leakage between the train and test sets. Our approach requires specification of population parameters of each disease (prevalence and heritability), and we detail the parameters we used and their sources in Table S1, available online.

Inference

We use the bootstrap[29] to estimate the SD of our AUC estimates by resampling the GRSs or genetic-risk predictions. For GeRSI, this is a straightforward bootstrap scheme. Denote $\mathrm{AUC}(r_1,\ldots,r_n,y_1,\ldots,y_n)$ the estimated AUC given risk predictions $r_1,\ldots,r_n$ (obtained via the method described above) and phenotypes $y_1,\ldots,y_n$. We sample with return n indices $i_1,\ldots,i_n \in \{1,\ldots,n\}$. The j^{th} bootstrap AUC sample is then

$$\mathrm{AUC}_j = \mathrm{AUC}\left(r_{i_1}, \ldots, r_{i_n}, y_{i_1}, \ldots, y_{i_n}\right),$$

and we report the empirical SD of 100 such bootstrap samples.

When dealing with fixed-effects models, we must account for the fact that the p value threshold is selected on the basis of the AUC. Obtained with c as a threshold, the risk score of the i^{th} individual, is denoted r_i^c. A bootstrap sample accounting for threshold selection is thus given by

$$\mathrm{AUC}_j = \max_c \left\{\mathrm{AUC}\left(r_{i_1}^c, \ldots, r_{i_n}^c, y_{i_1}, \ldots, y_{i_n}\right)\right\}.$$

Lastly, when testing whether one method performs better than another, we note that comparing AUCs by using estimated SDs is considerably conservative, given that AUCs obtained with the same set of observations are expected to be highly correlated.

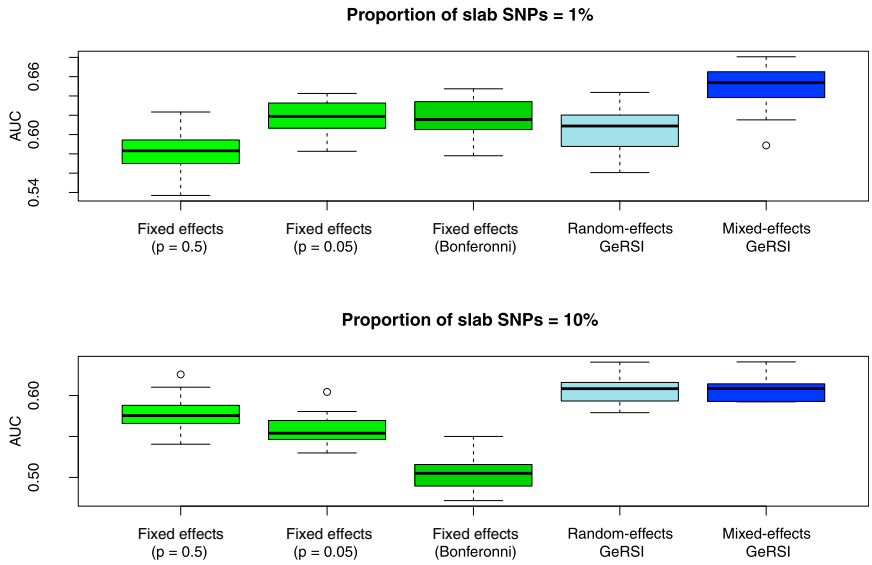

Figure 1. Comparison of the Performance of Fixed-, Random-, and Mixed-Effects Models in Predicting Disease Risk in a Spike-and-Slab Model

We simulated balanced case-control studies of a disease with 5% prevalence and 50% heritability and for which the fraction of slab SNPs with large effects was either 1% or 10% out of a total of 50,000 simulated SNPs and for which these slab SNPs accounted for 90% of the heritability, in line with values from Chatterjee et al.[4] We show the performance of the fixed-effects approach with (A) a Bonferroni-adjusted p value threshold, (B) a p value threshold of 0.05, and (C) a p value threshold of 0.5. In addition, we computed the correlation matrix G and used it to predict risk with the random-effects GeRSI approach, as well as with mixed-effects GeRSI treating the SNPs from (A) as fixed effects. In each simulation, we used a train set of 3,000 individuals and estimated the AUC for each method by using a test set of 1,000 individuals. We used the results from 20 independent simulations to draw the box plots.

Instead, we follow a similar scheme to estimate the SD of the difference in AUC directly.

Estimating Relative Risk

We are interested in estimating the relative risk (RR) of individuals at the top X of the risk predictions. A subtle aspect is that we wish to do so with case-control data because we don't have a random sample from the population. For a given risk threshold v, we estimate the fraction of the population with risk predictions higher than this threshold as

$$p_{\text{pop}} = \frac{K}{n_{\text{cases}}} \sum_{y_i=\text{case}} \mathbb{I}\{r_i > v\} + \frac{1-K}{n_{\text{controls}}} \sum_{y_i=\text{control}} \mathbb{I}\{r_i > v\}.$$

We then search for a value v such that p_{pop} is the desired value. The fraction of cases with risk predictions higher than the threshold is $p_{\text{cases}} = (1/n_{\text{cases}}) \sum_{y_i=1} \mathbb{I}\{r_i > v\}$, and the RR is estimated as $p_{\text{cases}}/(1 - p_{\text{cases}}) \times (1 - p_{\text{pop}})/p_{\text{pop}}$.

Results

We tested GeRSI in extensive simulations, as described in the Material and Methods. Prediction quality is measured by the AUC, which is the probability that a randomly sampled affected individual (case) attains a higher GRS than a randomly sampled unaffected individual (control).

Our results demonstrate, as expected, that fixed-effects modeling is generally effective when the phenotype is driven by a small number of SNPs with sizable effects (i.e., most causal SNPs are easily identified) and that the random-effects approach is most effective when the phenotype is driven by a large number of SNPs with small effects. Mixed-effects GeRSI performs well in both scenarios, as well as in intermediate scenarios, and was never inferior to fixed-effects modeling or gBLUP in any of our simulations.

In Figure 1, we present the results for the spike-and-slab model of genetic risk, recently explored by Zhou et al.[4] and Chatterjee et al.[13] and described in the Material and Methods. The results illustrate the power and flexibility of GeRSI and its superior performance in comparison to fixed-effects modeling (superior in all 20 simulation runs; for both setups, p value $< 10^{-6}$ with sign tests). When slab effects were relatively small (as in the bottom panel), random-effects GeRSI and mixed-effects GeRSI performed similarly, but in the presence of large slab effects (as in the top panel), the mixed-effects version allowed us to capture these as fixed effects and was far superior to the random-effects version. The results for other simulation settings are presented in Figures S1–S21.

In Figure 2, we compare GeRSI to gBLUP, which similarly utilizes random-effects modeling to reduce the number of parameters. Here, too, GeRSI's performance was uniformly superior, as expected from the fact that it utilizes the correct probabilistic model rather than an approximated model. We also investigated the recently described multiBLUP method of Speed and Balding,[19] which extends the BLUP model to include several variance components. We did this by constructing two correlation matrices in our spike-and-slab model—one for the spike SNPs and one for the slab SNPs. Using this refined correlation structure yielded considerably more accurate results. Importantly, GeRSI can be similarly extended to accommodate several variance components (multi-GeRSI), thus taking advantage not only of the refined correlation structure but also of GeRSI's improved statistical-modeling approach. As expected, multi-GeRSI outperformed multiBLUP in our simulations (Figure 2). Lastly, multi-GeRSI can be extended to incorporate fixed effects (mixed multi-GeRSI) so that SNPs with remarkably significant effects are treated as fixed effects while the rest are modeled

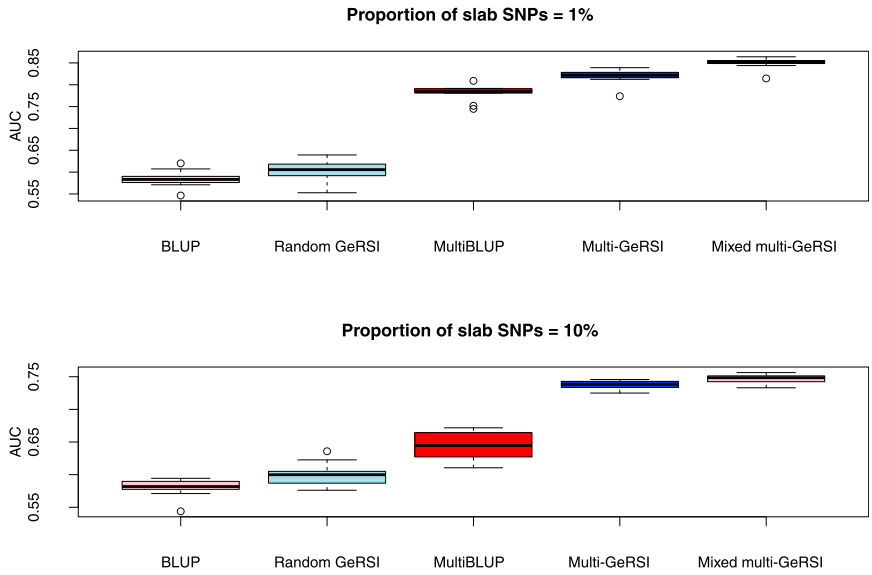

Figure 2. Comparison of the Performance of BLUP and GeRSI Methods in Predicting Disease Risk in a Spike-and-Slab Model

We compared the performance of BLUP, multiBLUP, GeRSI, multi-GeRSI, and mixed multi-GeRSI by using the same simulation setup as in Figure 1. We observed that GeRSI outperformed BLUP by utilizing the correct probabilistic setup. MultiBLUP takes into account the different effect-size distributions of spike and slab SNPs and therefore outperformed both. Multi-GeRSI enjoys the best of both worlds—correct sampling scheme and improved correlation structure—and so trumped all previous methods. Lastly, mixed multi-GeRSI improves over multi-GeRSI by including the most significant SNPs as fixed effects in addition to the other advantages of the multi-GeRSI approach.

as random effects. This approach improved the performance even further (Figure 2).

We then proceeded to apply GeRSI to seven WTCCC[30] case-control studies on BD, coronary artery disease (CAD), Crohn disease (CD), hypertension (HT), T1D, type 2 diabetes (T2D), and rheumatoid arthritis (RA). For each phenotype, we first performed stringent QC as suggested by Lee et al.[12] and detailed in the Material and Methods to mitigate batch effects. We then estimated the AUC with 4-fold cross-validation by using both the fixed-effects method of Dudbridge[3] and the random-effects and mixed-effects GeRSI approaches. To demonstrate the potential clinical utility of the risk-prediction approaches examined, we also estimated the RR of an individual found to be in the top 1% and 10% of risk predictions. For the fixed-effects approaches, we considered a range of possible p value thresholds (5×10^{-c} for $c \in \{1, \ldots, 8\}$) and display here the best result for each phenotype.

Comparing the fixed-effects approach to random-effects GeRSI, we observed that random-effects GeRSI obtained significantly higher AUC than the optimized fixed-effects approach for four of the seven phenotypes: BD, T2D, CAD, and HT (p value $< 10^{-9}$, 10^{-3}, 10^{-3}, and 10^{-5}, respectively; Table 1; see the Material and Methods for details). Specifically for HT, no SNPs were found to be associated at the 5×10^{-8} genome-wide-significance level in the WTCCC data,[30] and very few associations have been found in any studies to date (only seven associations are listed in the NHGRI GWAS Catalog [see Web Resources]). The optimal p value threshold for the fixed-effects model was 0.5, indicating a true polygenic architecture and thus an ideal phenotype for random-effects modeling. As expected, random-effects GeRSI yielded substantially better predictive power with an AUC around 0.6 and a considerable increase in RR for individuals at the top 10% of GeRSI risk predictions (1.61 versus 1.31). Figure 3 contrasts the ROC curves and the risk-prediction behavior of the two approaches on HT data.

Contrary to HT, BD has numerous replicated associations, some of which were identified in the original WTCCC study. However, in agreement with other studies,[2] we found that using a permissive p value threshold for risk predictions is appropriate (the optimal p value threshold was 0.5). Here, too, using random-effects modeling was beneficial, as expected. This is reflected in the AUC and 10% RR numbers in Table 1. Additionally,

Table 1. Comparison of the Fixed-Effects Approach and Random-Effects GeRSI on Four Phenotypes from the WTCCC

	Fixed			Random	
Phenotype	**Best Threshold**	**AUC (SE)**	**RR at Top 10% [CI]**	**AUC (SE)**	**RR at Top 10% [CI]**
BD	0.5	0.55 (0.01)	1.4 [1.03–1.65]	0.62 (0.01)	2.5 [2.16–2.96]
T2D	0.005	0.55 (0.01)	1.48 [1.2–1.83]	0.59 (0.01)	1.67 [1.34–1.99]
CAD	5×10^{-5}	0.65 (0.01)	1.85 [1.17–2.3]	0.67 (0.01)	2.16 [1.79–2.72]
HT	0.5	0.54 (0.01)	1.31 [1.06–1.55]	0.60 (0.01)	1.61 [1.42–1.97]

The AUC and RR of the top 10% and top 1% of individuals were estimated by 4-fold cross-validation. We compared the predictions obtained from our random-effects GeRSI approach to the predictions obtained by a fixed-effects approach. We computed the AUC of the fixed-effects approach for a wide range of p value thresholds (5×10^{-c} for $c \in \{1,\ldots,8\}$) and display here results for the value with the highest AUC (note that GeRSI has no such parameter). All analyses included sex as a covariate. CI stands for confidence interval.

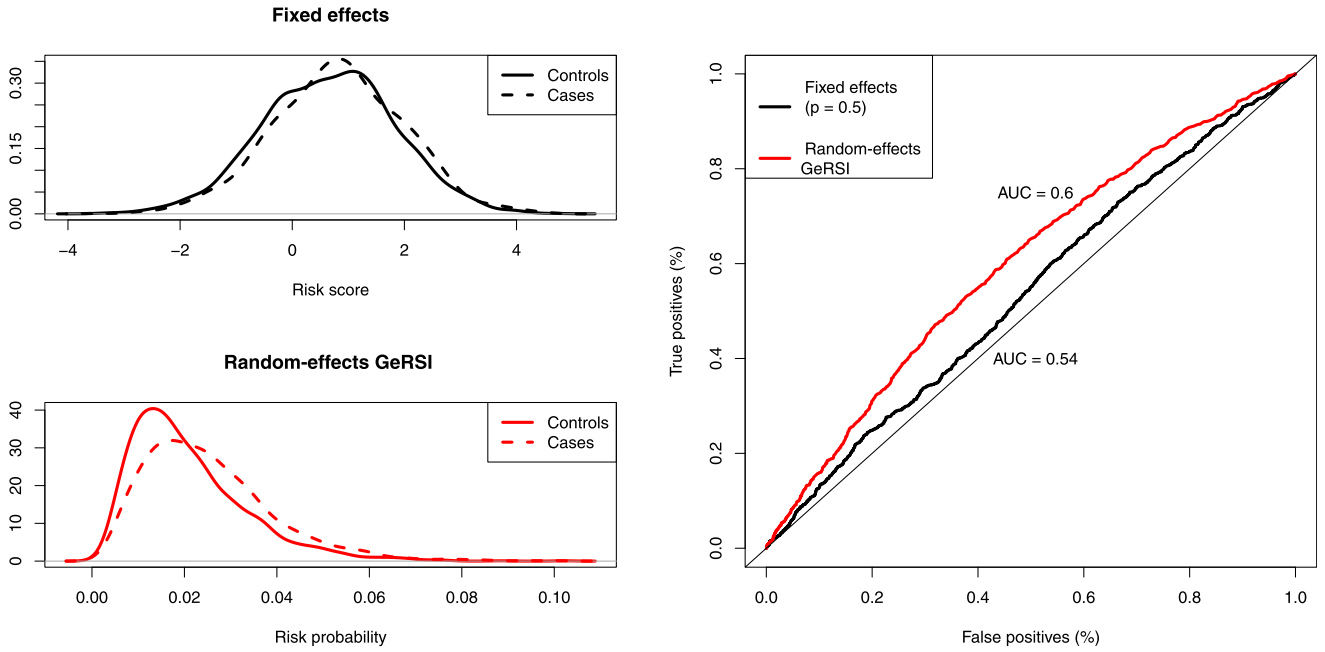

Figure 3. Comparison of HT Risk Predictions with Fixed-Effects Models and Random-Effects GeRSI
We used the fixed-effects approach with a p value threshold of 0.5. With fixed effects, there is very little difference between the distribution of risk scores of cases and control (top-left panel), but with random-effects GeRSI, out-of-sample risk predictions for cases is clearly skewed to the right (bottom-left panel). This is also evident in the comparison of the ROC curves of both methods (right panel).

the top 1% of risk scores attained a RR of almost 4 in random-effects GeRSI but only 1.26 with the fixed-effects approach.

Random-effects GeRSI did not improve over fixed-effects modeling for CD and performed significantly worse for RA and T1D (Table 2). This is consistent with our knowledge regarding the genetic architecture of these diseases: all three are autoimmune diseases with strongly associated SNPs with considerable effect sizes, primarily in the MHC region on chromosome 6. Gusev et al.[31] recently showed that a significant portion of the heritability of these diseases is due to variants in the vicinity of previously identified causal SNPs and is not uniformly distributed along the genome. To demonstrate the flexibility of mixed-effects GeRSI to combine the power of fixed- and random-effects modeling, we also show its performance in Table 2, where it is generally comparable to that of the fixed-effects approach (slightly superior for CD, p value $< 10^{-3}$).

Discussion

An important aspect in applying prediction methodology to case-control data is that of population structure. It is well established that population structure must be accounted for in GWASs, and this is often done with mixed models[8] or by inclusion of several top principal components as covariates.[32] In the context of heritability estimation, population structure can inflate the estimated heritability if unaccounted for, and typically a number of principal components of G are included as fixed effects to control for that structure. During risk prediction, the role of population structure is more complicated. We distinguish between two types of population structure: actual and induced. Actual population structure is structure that is truly present in the population and is properly reflected in the study group. Importantly, taking advantage of this type of structure for the purpose of risk prediction is legitimate, even if the effect of the structure on the phenotype is not via genetics. For example, if the diet of individuals of a certain ethnicity affects their disease risk, but this is not accounted for with fixed effects (e.g., if dietary information is not collected), this can still be captured in GeRSI via the genetic differences between these individuals and others in the population.

On the other hand, induced structure is an artifact of the sampling procedure. For example, a certain subpopulation might be considerably more likely to be sampled as cases rather than controls. In this case, GeRSI predictive-power estimates based on the study sample might be illegitimately inflated if this structure is not accounted for.

The WTCCC studies are considered to have relatively little structure,[30] and we are not able to separate the structure that does exist into its legitimate and induced components. To examine the robustness of our results to removal of structure, we reran our analyses while removing the top ten principal components from the correlation matrix (Table 3). This had a small negative effect on the performance of GeRSI for some of the phenotypes, such as BD and RA, but the general spirit of the results remained unchanged.

Table 2. Comparison of the Random- and Mixed-Effects GeRSI Approach and the Fixed-Effects Approach for Three Autoimmune Diseases in the WTCCC Data

	AUC (SE)			RR at Top 10% [CI]		
Phenotype	**Fixed**	**Random**	**Mixed**	**Fixed**	**Random**	**Mixed**
CD	0.59 (0.01)	0.59 (0.01)	0.62 (0.01)	2.02 [1.67–2.34]	2.18 [2.12–2.53]	2.63 [2.32–2.81]
T1D	0.72 (0.01)	0.55 (0.01)	0.71 (0.01)	3.4 [3.19–3.63]	1.46 [1.31–1.83]	3.45 [2.76–3.57]
RA	0.67 (0.01)	0.63 (0.01)	0.68 (0.01)	2.85 [2.62–3.61]	1.83 [1.44–2.07]	2.93 [2.55–3.54]

The AUC and RR of the top 10% of individuals were estimated by 4-fold cross-validation. We compared the predictions obtained from our random-effects GeRSI approach, our mixed-effects GeRSI approach, and a fixed-effects approach. We computed the AUC of the fixed-effects approach and the mixed-effects GeRSI approach for a wide range of p value thresholds (5×10^{-c} for $c \in \{1,\ldots,8\}$) and display here results for the value with the highest AUC. All analyses included sex as a covariate. CI stands for confidence interval.

Another key point is that of QC. Following Lee et al.,[12] we applied very stringent QC as described in the Material and Methods. Such stringent QC is particularly important during the evaluation of predictions, given that different cohorts are genotyped at different times and different centers, and so systematic genotyping errors might manifest as inflated estimates of the predictive capacity. On the other hand, stringent QC results in fewer SNPs and fewer individuals in the inspected data sets, which could result in conservative estimates of the predictive power.

The major computational bottleneck in GeRSI is the computation of the genetic correlation matrix (time complexity of $\mathcal{O}(n^2m)$, where n is the number of individuals in the reference sample and m is the number of SNPs). However, it is very easy to parallelize this task, and the correlations for the reference panel could be computed offline. Additionally, faster implementations of correlation-matrix computations could be utilized for improved performance.[33] The second bottleneck is the inversion of the correlation matrix for the purpose of computing the conditional means and variances (time complexity of $\mathcal{O}(n^3)$). This task could also be completed for the reference set offline, and rank-one updates could be used once the correlations between the reference set and the new individual are computed. Lastly, the time complexity of the Gibbs sampling itself is linear in the number of individuals. In conclusion, if proper prepossessing is carried offline, the time complexity of predicting the risk of a newly observed genotype is $\mathcal{O}(nm)$. Our main simulations used a reference set of 3,000 individuals to predict the phenotype of 1,000 individuals, in which case the three parts—computing the correlation matrix, inverting the matrix, and predicting risk—were completed in approximately 2 hr, 20 min, and 2 min, respectively, with a single core of a standard laptop computer. Our implementation is available in the Web Resources.

Table 3. Investigating the Possible Effect of Sampling-Induced Population Structure on Estimating Accuracy of Risk Prediction

	AUC (SE)		RR at Top 10% [CI]	
Phenotype	**0 PCs**	**10 PCs**	**0 PCs**	**10 PCs**
BD	0.62 (0.01)	0.59 (0.01)	2.5 [2.3–3.05]	1.81 [1.66–2.23]
CD	0.59 (0.01)	0.57 (0.01)	2.18 [1.83–2.55]	1.74 [1.49–2.01]
T1D	0.55 (0.01)	0.52 (0.01)	1.46 [1.09–1.7]	1.03 [0.94–1.16]
T2D	0.59 (0.01)	0.57 (0.01)	1.67 [1.58–1.8]	1.66 [1.46–1.76]
CAD	0.67 (0.01)	0.68 (0.01)	2.16 [1.75–2.61]	2.5 [2.24–2.93]
RA	0.63 (0.01)	0.6 (0.01)	1.83 [1.47–1.91]	1.77 [1.45–1.97]
HT	0.59 (0.01)	0.6 (0.01)	1.61 [1.48–1.91]	1.68 [1.46–1.85]

We compared the AUCs and RRs for the top 10% of individuals when using the correlation matrix G directly or after removing the top ten principal components (PCs) as described in the Material and Methods. As expected, we observed a minor decrement in performance in some phenotypes, but others (such as CAD and HT) displayed no real change in performance. All analyses included sex as a covariate. CI stands for confidence interval.

To study the scalability and performance of GeRSI when larger reference panels are used, we repeated the simulations of Figure 1 with reference panels of 6,000 or 9,000 individuals (Figures S22 and S23). For the largest studies simulated, the entire simulation and prediction process took about 2 days with a single CPU with 16 GB RAM. As expected, the performance of all methods improved as the size of the reference panel increased, but qualitatively, the results remained the same. Importantly, mixed-effects GeRSI still outperformed all other methods.

In conclusion, GeRSI is a method that takes advantage of the full power of random-effects modeling to accumulate evidence from the entire genome for the purpose of obtaining accurate risk predictions from GWASs. This is accomplished through an appropriate probabilistic-inference approach, also allowing for inclusion of relevant fixed effects, including associated SNPs and other covariates. Thus, any method for selecting a subset of SNPs or estimating covariate effects for the purpose of fixed-effects modeling can be used to model the fixed part of GeRSI and gain power from treating the rest of the SNPs as random effects.[3,4,21,22,34,35] Our results demonstrate the significant benefits of using this approach on both simulated and real data. Additionally, approaches designed for predicting quantitative phenotypes and that already utilize random effects could be easily adapted to take advantage of the GeRSI sampling scheme when applied to case-control data and benefit from a more accurate probabilistic model. We have demonstrated the improvement in performance for BLUP and multiBLUP, but similar improvements can be expected for other methods designed

for predicting quantitative phenotypes by using both fixed and random effects (e.g., LMM-Lasso[20]).

Specifically for BD, random-effects GeRSI allows us to identify 1% of the population at 4-fold risk of disease and 10% of the population at 2.5-fold risk. These numbers represent a major improvement over current state-of-the-art approaches and bring us closer to the ultimate goal of obtaining clinically useful risk predictions from GWAS data. We believe that random-effects modeling is a key component in this quest.

Appendix A

Matrix Identities for Fast Computation of the Conditional Mean and Variance

This section contains some matrix identities that are used in the software for faster computation.

Let $\overline{x} \sim \mathrm{MVN}(\overline{\mu}, \Sigma)$. We are interested in the conditional distribution $x_i|x_{-i}$. It is well known that for the MVN case, it is given by

$$x_i \,|\, x_{-i} \sim N\left(\Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}(x_{-i}-\mu_{-i}), \Sigma_{ii}-\Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}\Sigma_{-i,i}\right).$$

We wish to quickly compute the mean and variance for every i. Naively, this implies inverting an $(n-1)\times(n-1)$ matrix for n individuals in the train set, resulting in running-time complexity of $o(n^4)$, which is infeasible for realistic values of n. We are therefore interested in computing the conditional mean and variance for every individual in a more effective fashion.

To do so, focusing on $i = n$ WLOG, we write

$$\Sigma = \begin{pmatrix} A & b \\ c^\top & d \end{pmatrix} \text{ and } \Sigma^{-1} = \begin{pmatrix} E & f \\ g^\top & h \end{pmatrix},$$

where A and E are $(n-1)\times(n-1)$ matrices, so b, c, f, and g are column vectors, and d and h are scalars.

The first result is

$$A^{-1} = E - \frac{fg^\top}{h}$$

(for a derivation, see Mathematics Stack Exchange in the Web Resources). In other words, the inverses of all principal submatrices can be computed from the inverse of the overall matrix.

However, we are not interested directly in A^{-1} but rather in the conditional mean and variance.

The conditional mean is given by $\Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}(x_{-i}-\mu_{-i})$, so all we need to compute is $\Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}$, in other words,

$$c^\top A^{-1} = c^\top\left(E - \frac{fg^\top}{h}\right) = c^\top E - \frac{c^\top fg^\top}{h} = -dg^\top - \frac{(1-dh)g^\top}{h} = -\frac{g^\top}{h},$$

where we use the identities $c^\top E + dg^\top = 0$ and $c^\top f + dh = 1$, which stem from the identity $\Sigma\Sigma^{-1} = I$.

With our notation, the variance is

$$\Sigma_{ii} - \Sigma_{i,-i}(\Sigma_{-i,-i})^{-1}\Sigma_{-i,i} = d - c^\top A^{-1} b = d - c^\top\left(E - \frac{fg^\top}{h}\right)b.$$

Again, using the fact that $c^\top f + dh = 1$ and $c^\top E + dg^\top = 0$, we get

$$\ldots = d - c^\top\left(E - \frac{fg^\top}{h}\right)b = d - \left(c^\top E - \frac{c^\top fg^\top}{h}\right)b = d - \left(-dg^\top - \frac{(1-dh)g^\top}{h}\right)b = d + \frac{g^\top}{h}b = d + \frac{1-dh}{h} = \frac{1}{h}.$$

Supplemental Data

Supplemental Data include 23 figures and 1 table and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2014.09.007.

Acknowledgments

This study made use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available at www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113. D.G. was partially supported by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and by a fellowship from the Colton Family Foundation. This research was partially supported by Israeli Science Foundation grant 1487/12.

Received: July 13, 2014
Accepted: September 12, 2014
Published: October 2, 2014

Web Resources

The URL for data provided herein is as follows:

GeRSI, https://sites.google.com/site/davidgolanshomepage/software/gersi

Mathematics Stack Exchange, http://math.stackexchange.com/questions/208001/are-there-any-decompositions-of-a-symmetric-matrix-that-allow-for-the-inversio

NHGRI GWAS Catalog, http://www.ebi.ac.uk/fgpt/gwas/

References

1. Goldstein, D.B. (2009). Common genetic variation and human traits. N. Engl. J. Med. *360*, 1696–1698.
2. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P., Ruderfer, D.M., McQuillin, A., Morris, D.W., et al.; International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

3. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
4. Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. *45*, 400–405, e1–e3.
5. Abraham, G., Tye-Din, J.A., Bhalala, O.G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. PLoS Genet. *10*, e1004137.
6. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat. Methods *8*, 833–835.
7. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354.
8. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. *11*, 459–463.
9. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824.
10. Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods *11*, 407–409.
11. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
12. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
13. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
14. Golan, D., and Rosset, S. (2011). Accurate estimation of heritability in genome wide studies using random effects models. Bioinformatics *27*, i317–i323.
15. VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. *91*, 4414–4423.
16. Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics *157*, 1819–1829.
17. Clark, S.A., and van der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. Methods Mol. Biol. *1019*, 321–330.
18. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
19. Speed, D., and Balding, D.J. (2014). MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. *24*, 1550–1557.
20. Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics *29*, 206–214.
21. Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). The Bayesian lasso for genome-wide association studies. Bioinformatics *27*, 516–523.
22. Zaitlen, N., Lindström, S., Pasaniuc, B., Cornelis, M., Genovese, G., Pollack, S., Barton, A., Bickeböller, H., Bowden, D.W., Eyre, S., et al. (2012). Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet. *8*, e1003032.
23. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.
24. Crossett, A., Lee, A.B., Klei, L., Devlin, B., and Roeder, K. (2013). Refining genetically inferred relationships using treelet covariance smoothing. Ann. Appl. Stat. *7*, 669–690.
25. Casella, G., and George, E.I. (1992). Explaining the gibbs sampler. Am. Stat. *46*, 167–174.
26. Campbell, D.D., Sham, P.C., Knight, J., Wickham, H., and Landau, S. (2010). Software for generating liability distributions for pedigrees conditional on their observed disease states and covariates. Genet. Epidemiol. *34*, 159–170.
27. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.
28. Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al.; International HapMap Consortium (2003). The international hapmap project. Nature *426*, 789–796.
29. Efron, B., and Tibshirani, R.J. (1994). An introduction to the bootstrap (Boca Raton: CRC Press).
30. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., et al.; Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
31. Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B.J., Diogo, D., Stahl, E.A., Gregersen, P.K., Worthington, J., Klareskog, L., Raychaudhuri, S., et al. (2013). Quantifying missing heritability at known GWAS loci. PLoS Genet. *9*, e1003993.
32. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
33. Gray, A., Stewart, I., and Tenesa, A. (2012). Advanced complex trait analysis. Bioinformatics *28*, 3134–3136.
34. Zaitlen, N., Paşaniuc, B., Patterson, N., Pollack, S., Voight, B., Groop, L., Altshuler, D., Henderson, B.E., Kolonel, L.N., Le Marchand, L., et al. (2012). Analysis of case-control association studies with known risk variants. Bioinformatics *28*, 1729–1737.
35. Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010). Risk prediction using genome-wide association studies. Genet. Epidemiol. *34*, 643–652.

The American Journal of Human Genetics, Volume *95*

Supplemental Data

Effective Genetic-Risk Prediction Using Mixed Models

David Golan and Saharon Rosset

Supplemental data

Additional simulation results

Large number of SNPs with small percentage of them causative

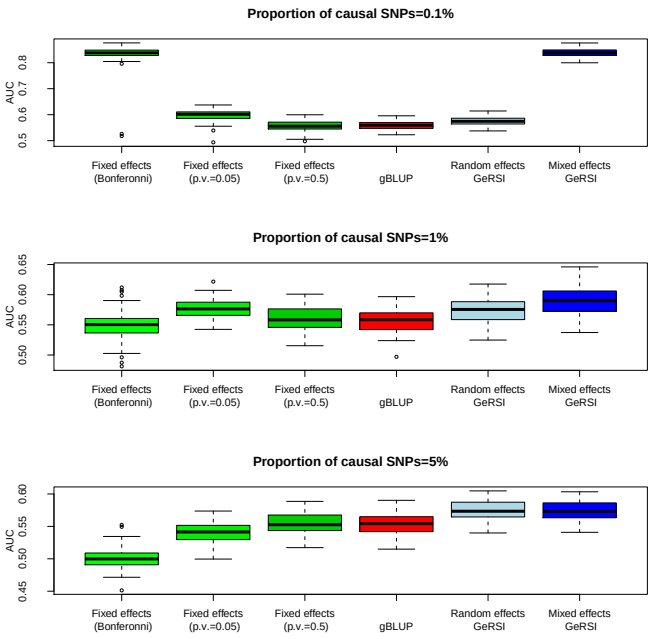

Figure S1: We explored the performance of the various methods when a fraction π_1 of the SNPs are causal and have a normally distributed effect size, while the rest of the SNPs have no effects. This allows for simulations with a larger number of SNPs as only the causative SNPs need to be simulated for the entire population, and the non-causal SNPs can be simulated only for the individuals selected for the study. In these simulations we simulated $100,000$ SNPs. The results are given in figures S1-S3. In this figure the prevalence is K=0.05.

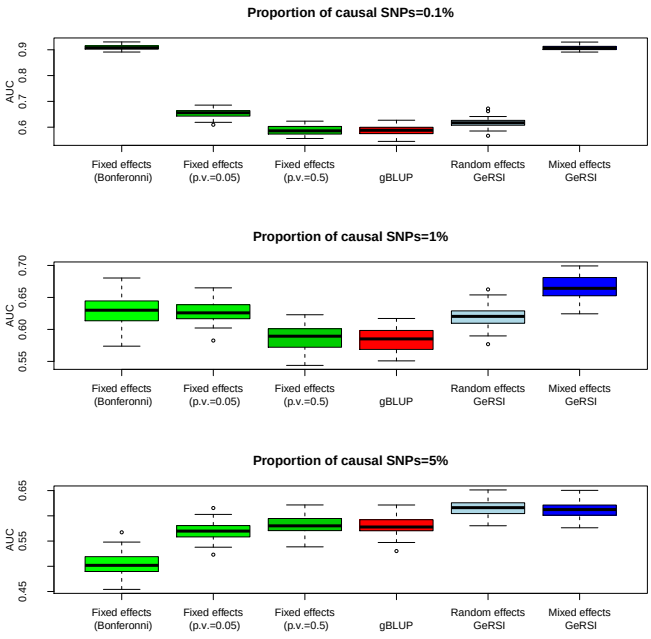

Figure S2: Prevalence K=0.01.

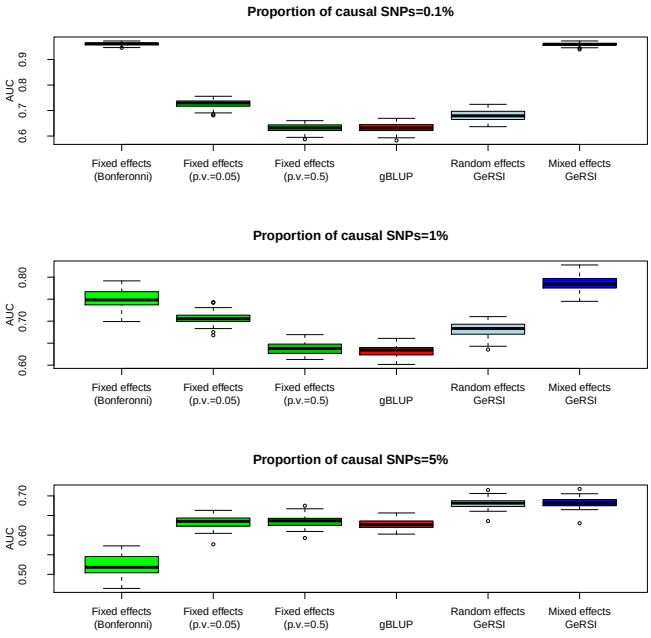

Figure S3: Prevalence K=0.001.

Using the double-exponential distribution to model effect sizes

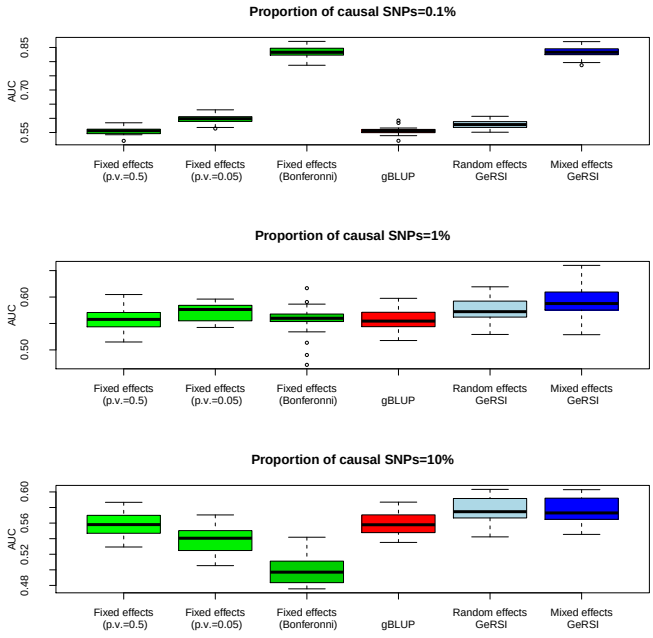

Figure S4: To study robustness of GeRSI to assumptions about the distribution of effect sizes, we reran the same simulation scheme, but with the effect sizes drawn from a double-exponential distribution with the parameter calibrated to achieve the appropriate variance. The results are given in figures S4-S6. Qualitatively they are similar to the normal distribution case and spike and slab case. Here K=0.05.

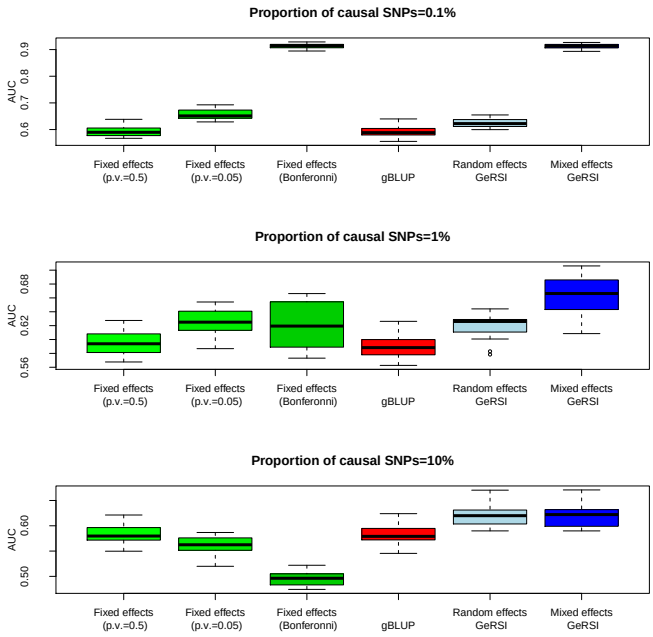

Figure S5: Simulating effects using the double exponential distribution. K=0.01.

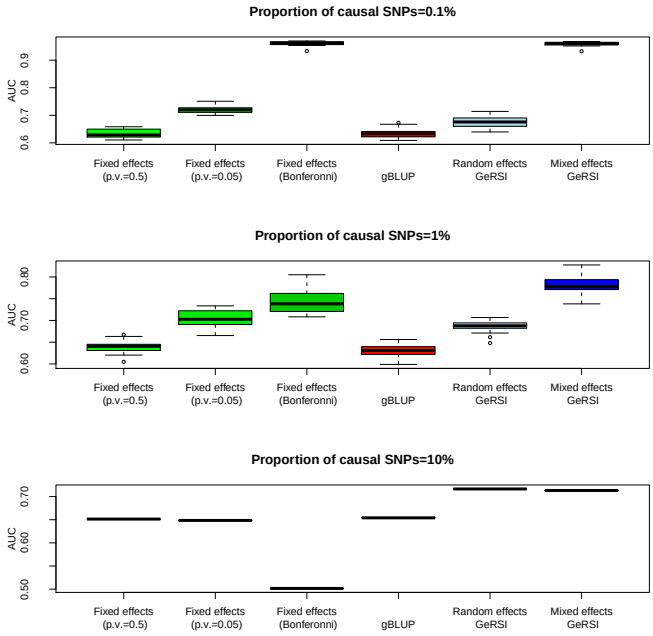

Figure S6: Simulating effects using the double exponential distribution. K=0.001.

Additional simulations under "spike and slab" mixture model

Adopting the "spike and slab" model proposed by Chatterjee et al. (2013) [1] and Zhou et al. (2013) [2] as realistic for modeling the distribution of effect sizes, we propose here additional simulations under a wide range of parameters. We simulate genotypes and phenotypes from this model for all combinations of $\pi_1 \in \{0.1\%, 1\%, 10\%\}$, and $\frac{\sigma^2_{spike}}{h^2} \in \{10\%, 30\%, 50\%, 70\%, 90\%\}$, while fixing the heritability at 50% and simulating $50,000$ SNPs. The results of these simulations are shown in figures S7-S21.

We note that Chatterjee et al. (2013) estimate the components of such mixtures for a wide range of phenotypes. resulting in a wide range of values. We chose the values of our simulations to cover a considerable part of the parameters estimated by Chatterjee et al. (2013).

As might be expected intuitively, under a mixture model, mixed-effect GeRSI considerably outperforms all other approaches.

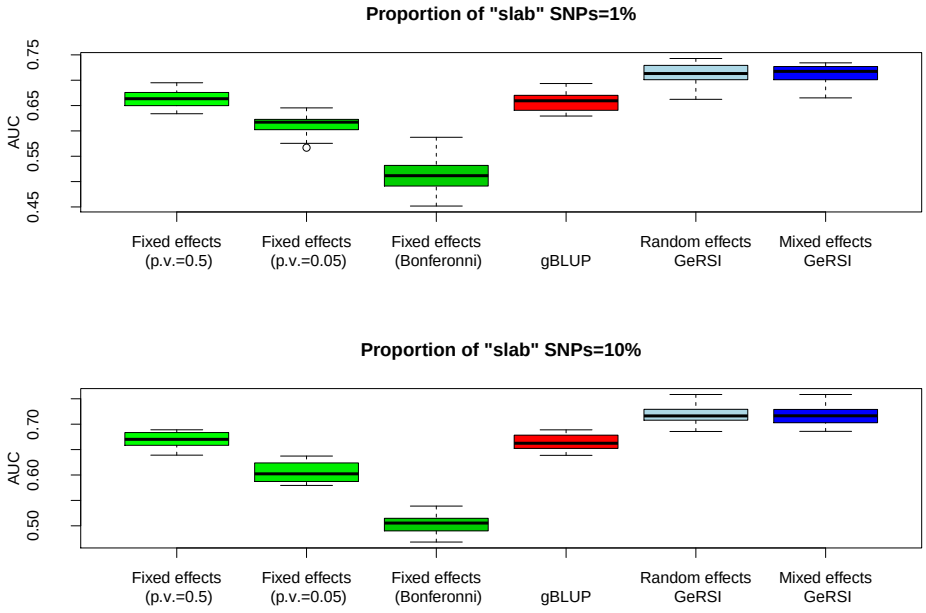

Figure S7: K=0.05, proportion of heritability due to slab=0.1.

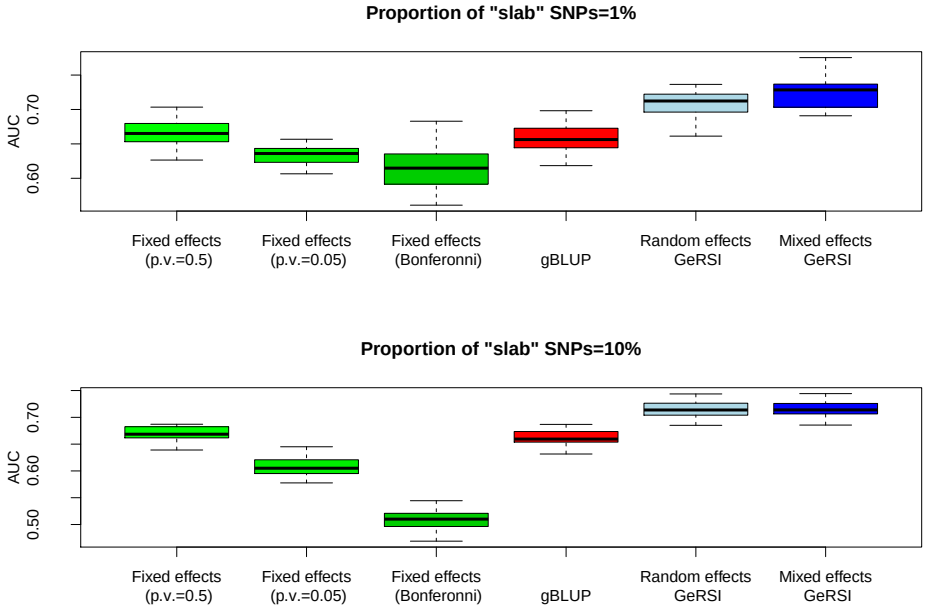

Figure S8: K=0.05, proportion of heritability due to slab=0.3.

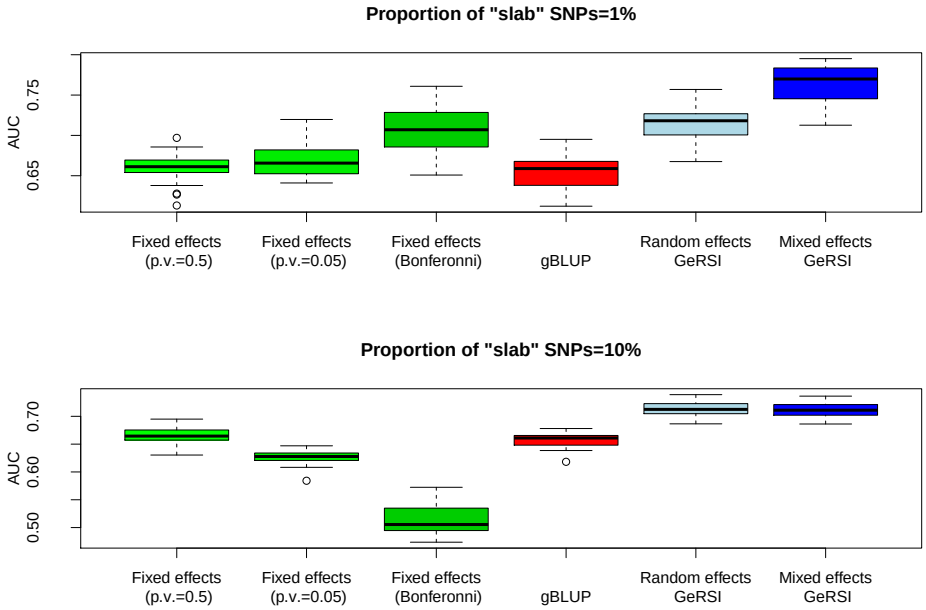

Figure S9: K=0.05, proportion of heritability due to slab=0.5.

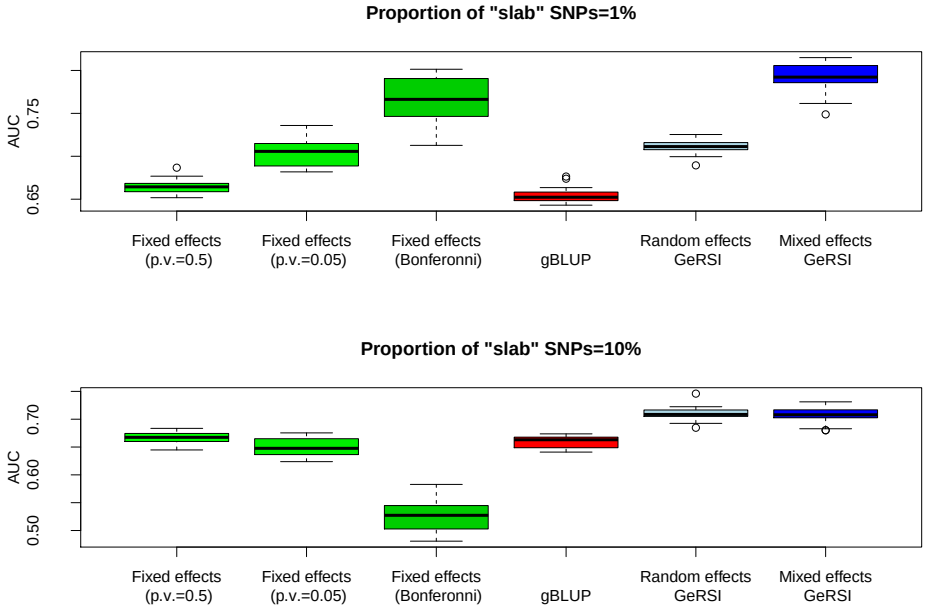

Figure S10: K=0.05, proportion of heritability due to slab=0.7.

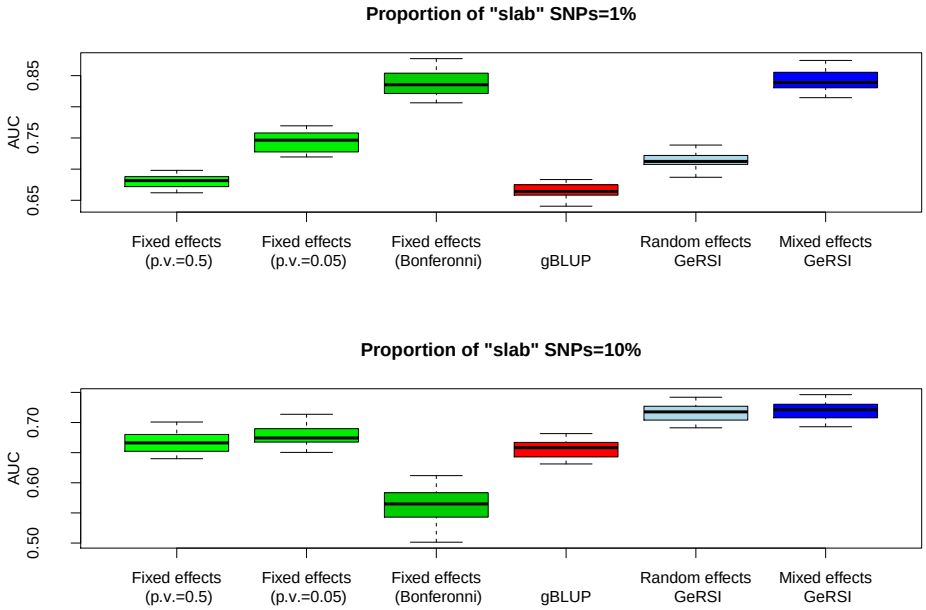

Figure S11: K=0.05, proportion of heritability due to slab=0.9.

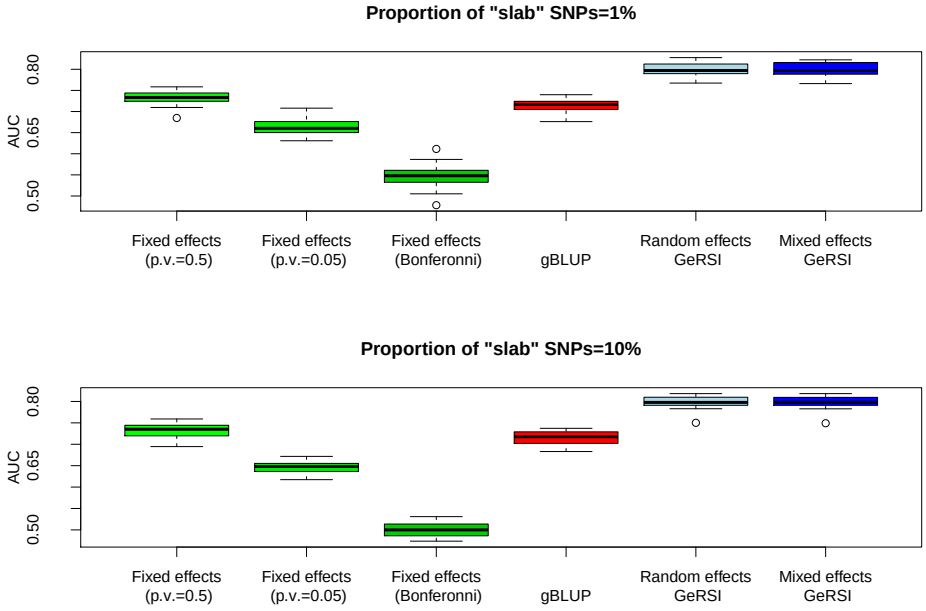

Figure S12: K=0.01, proportion of heritability due to slab=0.1.

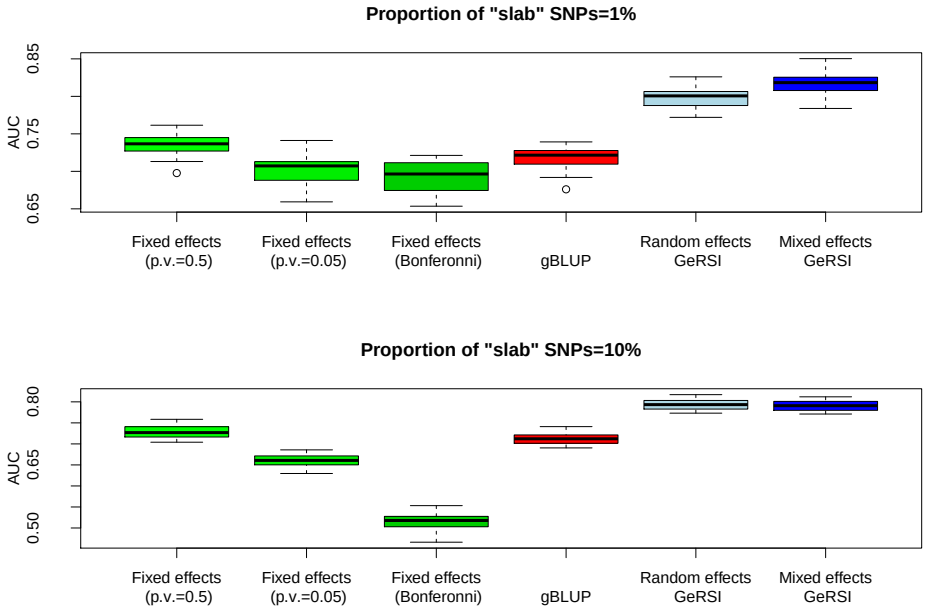

Figure S13: K=0.01, proportion of heritability due to slab=0.3.

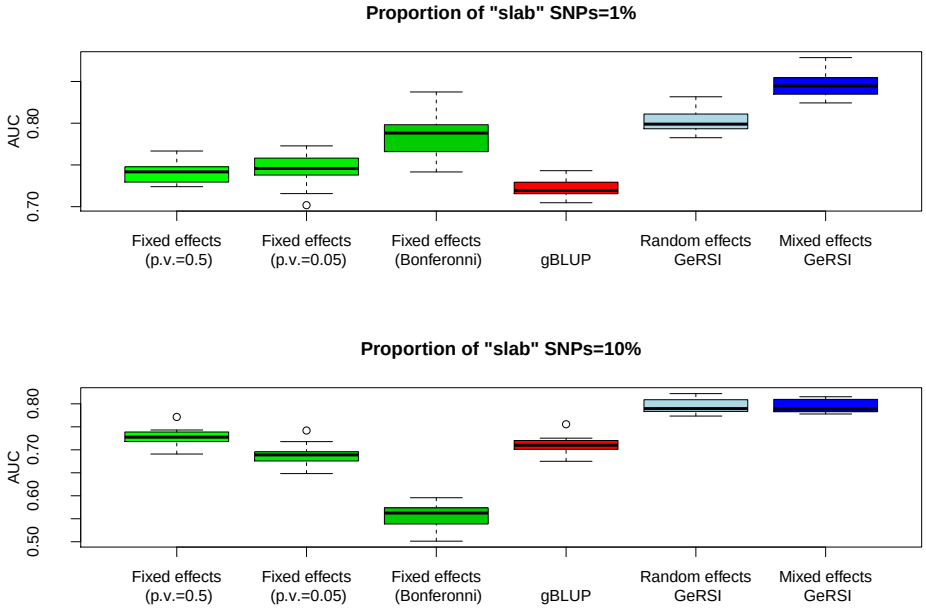

Figure S14: K=0.01, proportion of heritability due to slab=0.5.

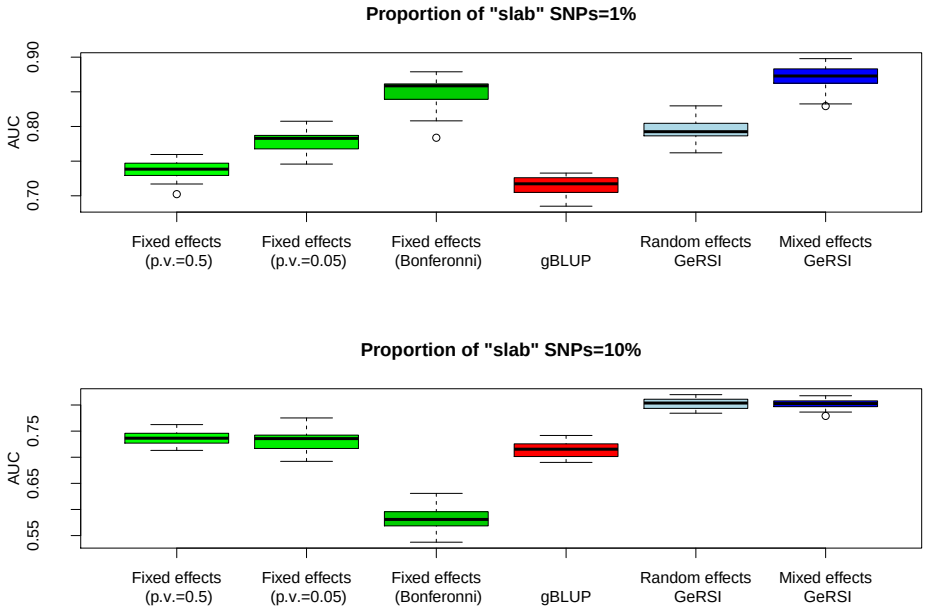

Figure S15: K=0.01, proportion of heritability due to slab=0.7.

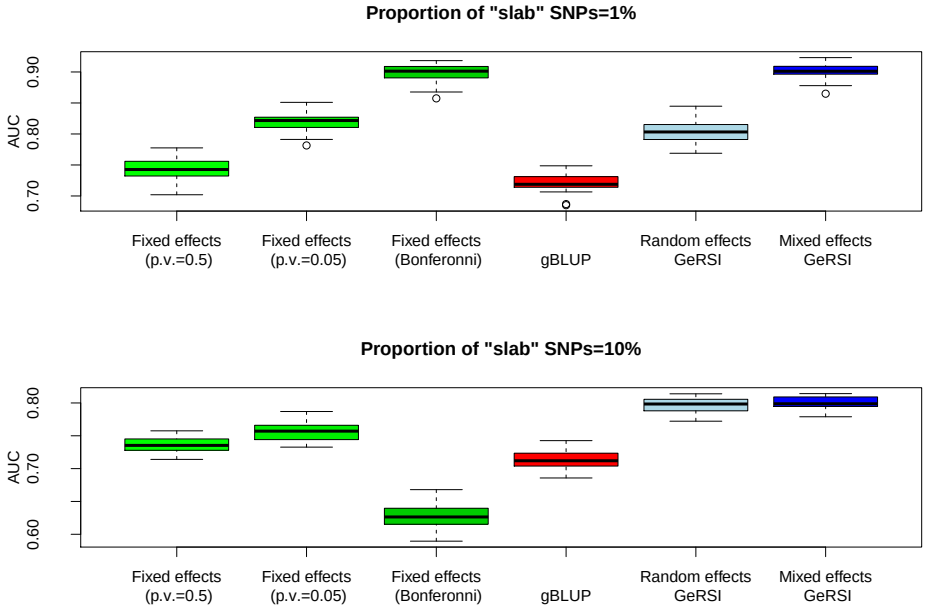

Figure S16: K=0.01, proportion of heritability due to slab=0.9.

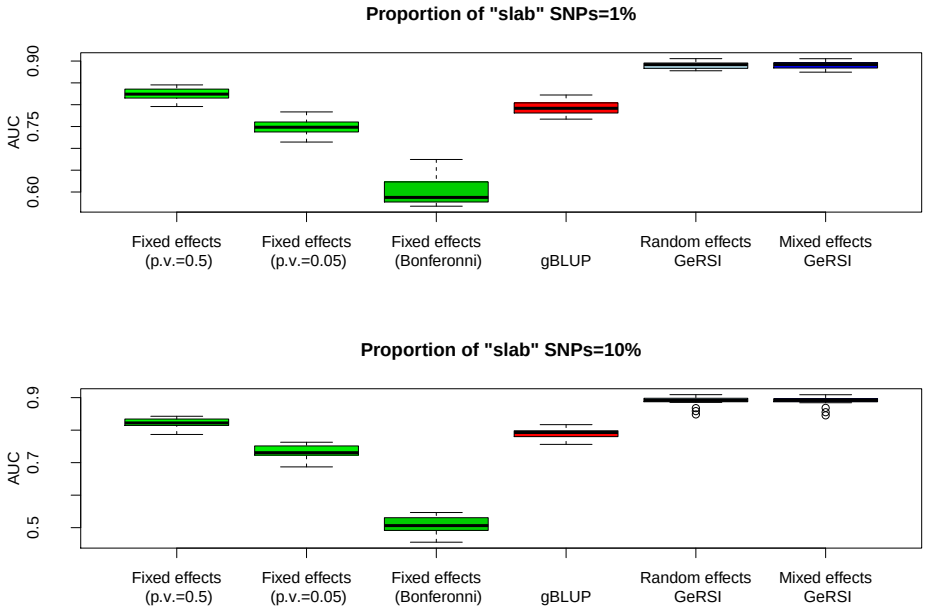

Figure S17: K=0.001, proportion of heritability due to slab=0.1.

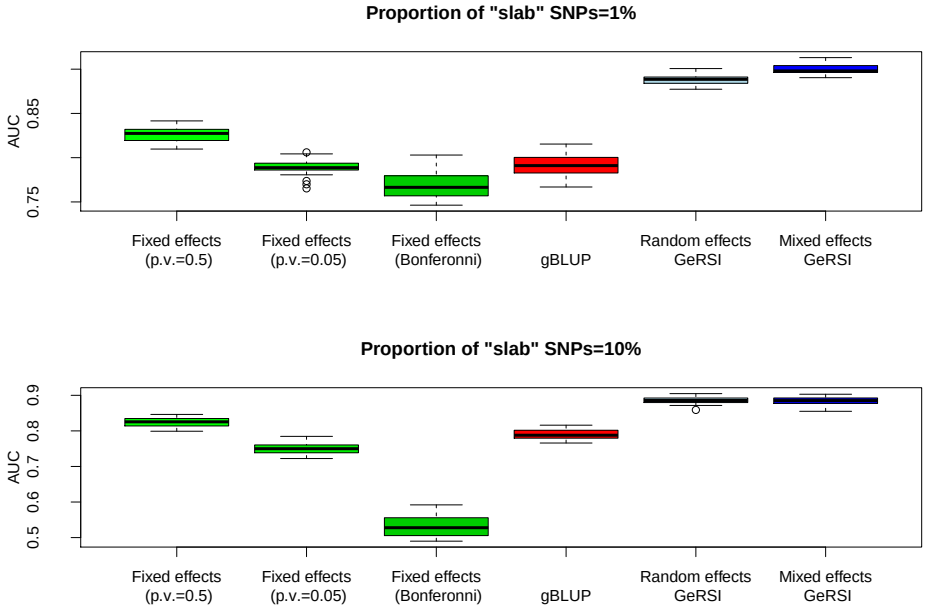

Figure S18: K=0.001, proportion of heritability due to slab=0.3.

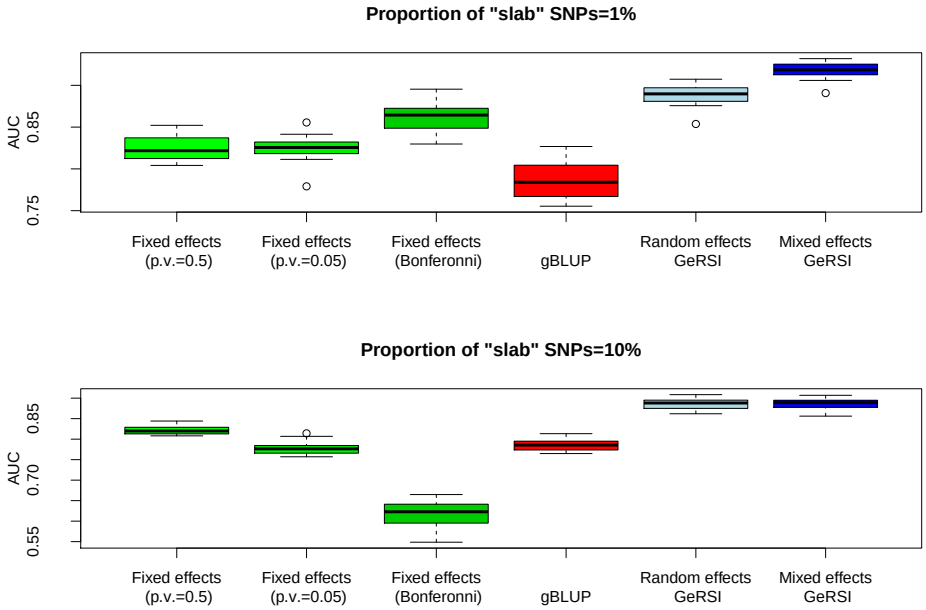

Figure S19: K=0.001, proportion of heritability due to slab=0.5.

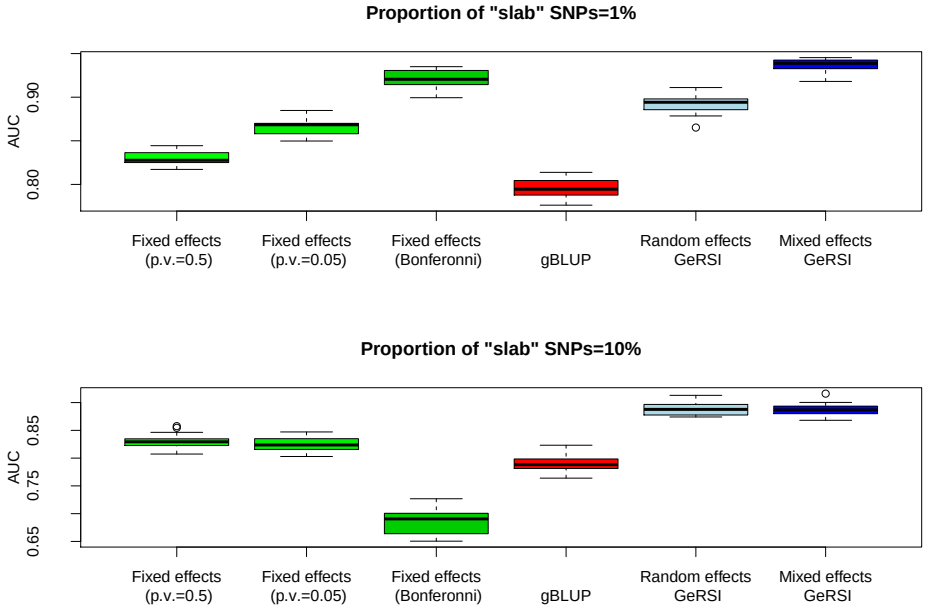

Figure S20: K=0.001, proportion of heritability due to slab=0.7.

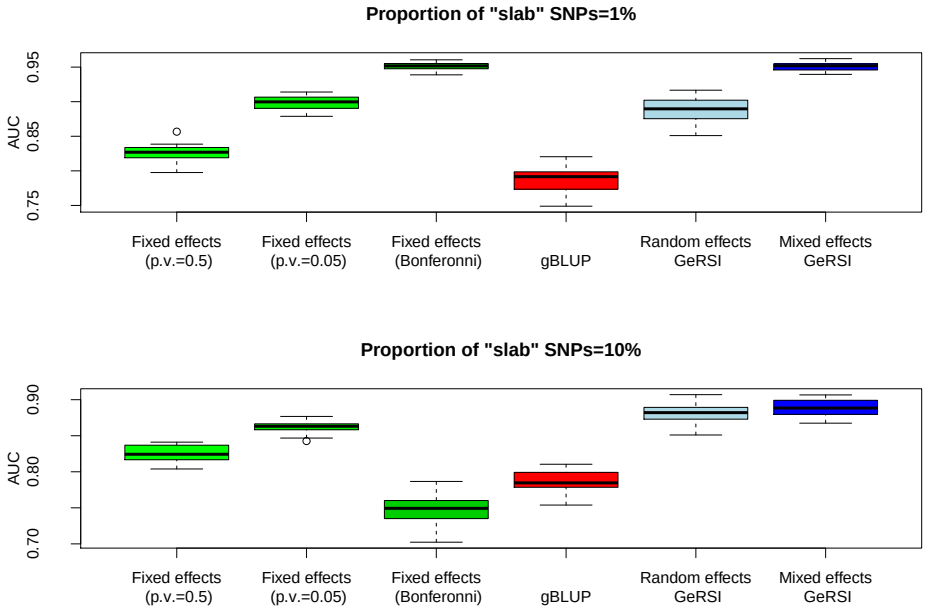

Figure S21: K=0.001, proportion of heritability due to slab=0.9.

Simulations using larger studies

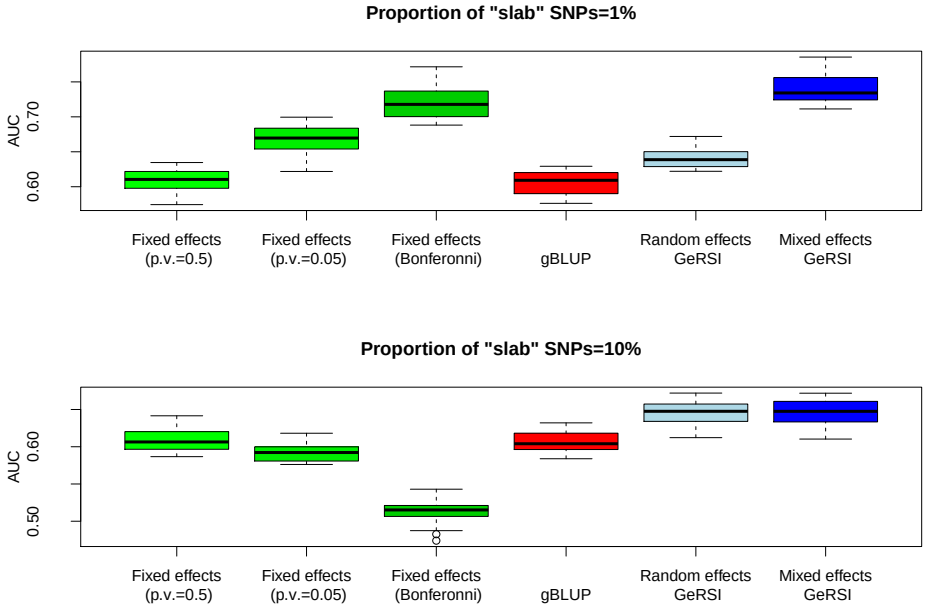

Figure S22: K=0.05, proportion of heritability due to slab=0.9. The size of the reference set was $6,000$ individuals.

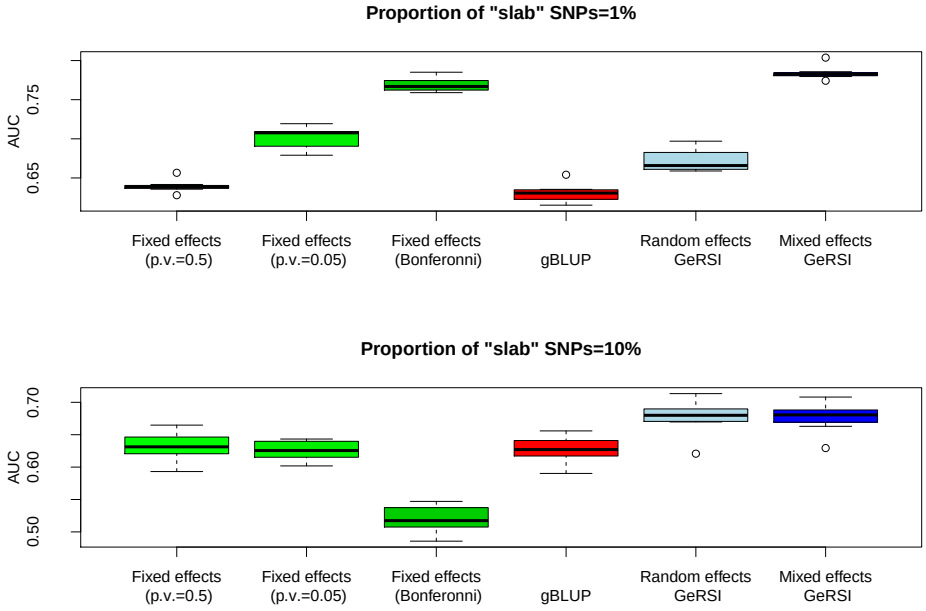

Figure S23: K=0.05, proportion of heritability due to slab=0.9. The size of the reference set was $9,000$ individuals.

Population parameters used in real data analysis

Phenotype	K (%)	h^2 (%)
BD	0.5	60
CD	0.1	70
T1D	0.5	80
T2D	3	70
CAD	3.5	50
RA	0.75	60
HT	5	60

Table S1: Population parameters (prevalence and heritability) used for estimating the performance of risk-prediction methods, taken from [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]

References

[1] Nilanjan Chatterjee, Bill Wheeler, Joshua Sampson, Patricia Hartge, Stephen J Chanock, and Ju-Hyun Park. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics*, 45(4):400–405, 2013.

[2] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics*, 9(2):e1003264, 2013.

[3] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[4] Naomi R Wray, Jian Yang, Michael E Goddard, and Peter M Visscher. The genetic interpretation of area under the roc curve in genomic profiling. *PLoS genetics*, 6(2):e1000864, 2010.

[5] Frank Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9(3):e1003348, 2013.

[6] P Almgren, M Lehtovirta, B Isomaa, L Sarelin, MR Taskinen, V Lyssenko, T Tuomi, and L Groop. Heritability and familiality of type 2 diabetes and related quantitative traits in the botnia study. *Diabetologia*, 54(11):2811–2819, 2011.

[7] P Poulsen, K Ohm Kyvik, A Vaag, and H Beck-Nielsen. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*, 42(2):139–145, 1999.

[8] J Sofaer. Crohn's disease: the genetic contribution. *Gut*, 34(7):869–871, 1993.

[9] Jack Edvardsen, Svenn Torgersen, Espen Røysamb, Sissel Lygren, Ingunn Skre, Sidsel Onstad, and Per Anders Øien. Heritability of bipolar spectrum disorders. unity or heterogeneity? *Journal of affective disorders*, 2008.

[10] Kirsten O Kyvik, Anders Green, and Henning Beck-Nielsen. Concordance rates of insulin dependent diabetes mellitus: a population based study of young danish twins. *BMJ*, 311(7010):913–917, 1995.

[11] Valma Hyttinen, Jaakko Kaprio, Leena Kinnunen, Markku Koskenvuo, and Jaakko Tuomilehto. Genetic liability of type 1 diabetes and the onset age among 22,650 young finnish twin pairs a nationwide follow-up study. *Diabetes*, 52(4):1052–1055, 2003.

[12] Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[13] Nina Kupper, Gonneke Willemsen, Harriëtte Riese, Daniëlle Posthuma, Dorret I Boomsma, and Eco JC de Geus. Heritability of daytime ambulatory blood pressure in an extended twin design. *Hypertension*, 45(1):80–85, 2005.