Bayesian multiple logistic regression for meta-analyses of GWAS

Saikat Banerjee[1], Lingyao Zeng[2], Heribert Schunkert[2] and Johannes Söding[1,*]

[1]Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany

[2]German Heart Centre, 80636 Munich, Germany

[*]Corresponding author: soeding@mpibpc.mpg.de

Abstract

Genetic variants in genome-wide association studies (GWAS) are tested for disease association mostly using simple regression, one variant at a time. Multiple regression can improve power by aggregating evidence from multiple nearby variants. It can also distinguish disease-coupled variants from variants merely correlated with a coupled variant. However, it requires individual genotype data, limiting its applicability when combining several GWAS. Moreover, multiple *logistic* regression to model binary phenotypes in case-control GWAS requires inefficient sampling schemes to integrate over the variant effect sizes. Our sparse Bayesian multiple LOgistic REgression (B-LORE) method overcomes these two drawbacks. We propose a *quasi-Laplace* approximation to analytically integrate over variant effect sizes. The resulting marginal likelihood functions of individual GWAS are approximated by multivariate normal distributions. Their means and covariance matrices serve as summary statistics for combining several GWAS. Additionally, B-LORE can integrate functional genomics tracks as priors for each variant's causality. To test our method, we simulated synthetic phenotypes for real genotypes. B-LORE improved the prediction of loci harboring causal variants and the variant fine mapping. We also used B-LORE for a metanalysis of five small GWAS for coronary artery disease (CAD). We pre-selected the top 50 loci with SNPTEST / META, which included 11 loci discovered by a 14-fold larger meta-analysis (CARDIoGRAMplusC4D). While simple regression discovered only 3 of them with genome-wide significance, B-LORE discovered all of them with causal probability > 95%. Of the 12 other loci discovered by B-LORE, 3 are known from other CAD GWAS and 6 are associated with well-known CAD risk-related blood metabolic phenotypes. Software availability: https://github.com/soedinglab/b-lore.

Introduction

Common, noninfectious diseases are responsible for over 2/3 of the deaths worldwide. These diseases are usually polygenic, with many variants each contributing only a small fraction of the disease risk. This made them very difficult to investigate using family studies. Genome wide association studies (GWAS) have opened up a fundamentally new approach to explore and understand the causality of disease development. The knowledge about the underlying biological mechanisms is crucial to devise prophylactic and therapeutic treatments.

In the most common GWAS design, patients with diagnosed diseases ("cases") and healthy people ("controls") are genotyped at several hundred thousand positions where single nucleotide polymorphisms (SNPs) are relatively frequent in the population. The data is then statistically analyzed to detect SNPs which have significant associations with the disease. Thousands of GWAS with millions of patients have been conducted in the past decade [1] with which thousands of genetic variants associated with many diseases and complex traits have been identified [2].

The statistical analyses commonly involve hypothesis tests for one SNP at a time, yielding p-values for each SNP independent of all others. Given the enormous volume of genotype data, the computational speed of this simple regression is a major advantage. However, this model only detects association, not statistical coupling. A SNP can show association with the disease simply by being correlated to an actually causal SNP. This situation is the rule rather than the exception, because SNPs are usually stronlgy correlated to dozens of nearby SNPs, up to distances of 100 kbp. This non-random correlation between nearby SNPs is referred to as linkage disequilibrium (LD) and occurs due to the common descent of humankind from a relatively small ancestral population.

The distinction between association (*i.e.* correlation) and coupling is therefore very important for the prediction of causal SNPs. SNPs which are highly correlated with a causal SNP obtain similarly significant p-values, making it difficult to decide which of these SNPs is really causal. Fig. 1 illustrates the importance of this distinction with another example, where two causal SNPs are highly correlated with a noncausal SNP between them. The noncausal SNP may exhibit a more significant p-value for disease association than the two causal SNPs, possibly leading to wrong interpretations of what genes are involved in the disease mechanisms.

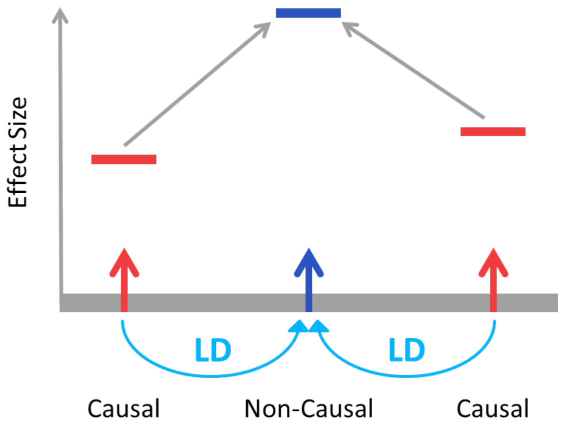

Figure 1. Effect of linkage disequilibrium (LD) with multiple causal variants. A noncausal SNP in LD with multiple causal SNPs may exhibit more significant effect size for disease association than the actually causal SNPs, possibly leading to wrong interpretations of what genes are involved in the disease mechanism.

The low effect sizes of single SNPs in complex polygenic diseases studied in GWAS limits the power of simple regression to detect statistically significant associations. Advances in GWAS have focused on two aspects to improve the power of detecting associations: (1) using multiple regression models for joint analysis of many SNPs (multiple-SNP analyses, or polygenic modelling), and (2) using meta-analysis of multiple independent studies to combine evidence from more and more patients.

Multiple regression models use many SNPs at a genetic region or locus as explanatory variables. This improves the power of GWAS by distinguishing between correlation and coupling as discussed above, as well as by aggregating evidence from many SNPs in the locus with low effect size, each of which would not be detected by single-SNP tests. Bayesian multiple regression models, particularly Bayesian variable selection regression (BVSR) [3,4], have been shown to perform significantly better than simple regression methods, *e.g.* SNPTEST [5–7], when individual-level genotype data are available. However, it is cumbersome to combine studies using multiple regression analyses due to the requirement of individual genotype data and the associated logistical, technical, and ethical restrictions for sharing genetic data from patients.

On the other hand, existing meta-analysis methods efficiently combine the single-SNP summary statistics from many studies and increase the power for detecting associations by collecting evidence from a larger pool of samples. The power gained from increased sample sizes in meta-analyses usually outweighs the power of multiple regression models applied on a more modest number of samples in an individual study.

Approaches that allow to combine multiple regression with meta analysis on many GWAS should lead to more power for detecting associations. Cichonska *et al.* recently made a pioneering contribution in this regard. Their tool metaCCA [8] performs canonical correlation analysis (CCA) of multiple SNPs against multiple traits, based on standard SNPTEST summary statistics and a genotype covariance matrix estimated from individual-level genotype data from the same or a similar population. CCA is used to identify and quantify the linear association between the two sets of variables, but model/variable selection is not included in the model. Therefore, metaCCA requires a pre-selection of SNPs. The authors proposed to select a roughly uncorrelated set of SNPs that together explain the maximum variance in a given genetic locus.

In this work, we present B-LORE, a Bayesian method using multiple logistic regression of the case-control binary variable and a prior distribution for the effect size of SNPs modeled by a two-component Gaussian mixture. The Gaussian mixture prior is motivated by BVSR [3], which has been successfully applied to many GWAS [4].

The weights, means and variances of the two-component Gaussian mixture prior are the model hyperparameters. B-LORE learns the hyperparameters from the data using the empirical Bayes approach of maximizing the *marginal* likelihood, which is obtained from the total likelihood by integrating out the unknown effect sizes of all SNPs. The total likelihood is the product over the regularized likelihoods of each individual study (Eq. (14)). The regularized likelihood of each study is approximated by a multivariate normal distribution, whose mean and covariance matrix is obtained by L_2-regularized multiple logistic regression. Thus the mean and covariance matrices of each study serve as *summary statistics* that can be shared between research groups for downstream meta-analysis.

Through simulations and application to real GWAS data, we show that B-LORE combines the advantages of multiple logistic regression and meta-analysis, successfully incorporates functional information of the SNPs, and outperforms state-of-the-art methods in the prediction of causal loci and in finemapping of causal SNPs.

Materials and Methods

Model likelihood and priors

GWAS data consists of phenotypes $\phi_n \in \{0, 1\}$ (healthy or diseased) and of genotypes $w_{ni} \in \{0, 1, 2\}$, where 0, 1, or 2 signify the number of minor alleles of patient $n \in \{1, \ldots, N\}$ at SNP $i \in \{1, \ldots, I\}$. The genotype is centered and normalized as $x_{ni} = (w_{ni} - 2f_i)/\sqrt{2f_i(1 - f_i)}$, where f_i is the minor allele frequency of the i^{th} SNP. Henceforth, we will denote the vector of normalized genotypes for the n^{th} sample as $\mathbf{x}_n$, and the $N \times I$ matrix of genotypes as $\mathbf{X}$. We model the effect strength, *i.e.* the log-odds ratio of diseased to healthy, by a linear function in which each minor allele contributes independently with additive effect,

$$\log \frac{p\left(\phi_n = 1 \mid \mathbf{x}_n, \mathbf{v}\right)}{p\left(\phi_n = 0 \mid \mathbf{x}_n, \mathbf{v}\right)} = v_0 + \sum_{i=1}^{I} v_i x_{ni} \,. \tag{1}$$

The offset v_0 determines the odds ratio for a patient without any minor allele SNPs, and v_i is the effect size of the i^{th} SNP. For notational convenience we define the $(I + 1)^{\text{th}}$ component of each vector $\mathbf{x}_n$ to be 1 in order to absorb the offset term into the scalar product. We can then write the right-hand side in vector notation, $\mathbf{v}^{\mathsf{T}}\mathbf{x}_n$. Abbreviating $p_n := p\left(\phi_n = 1 \mid \mathbf{x}_n, \mathbf{v}\right)$, we note that the above equation can be transformed to $p_n = 1/(1 + \exp\left(\mathbf{v}^{\mathsf{T}}\mathbf{x}_n\right))$, the standard model of logistic regression. The likelihood for N patients with genotypes $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N))$ is therefore

$$L(\mathbf{v}) = p\left(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{v}\right) = \prod_{n=1}^{N} p_n^{\phi_n} \left(1 - p_n\right)^{(1-\phi_n)} = \prod_{n=1}^{N} \frac{\exp\left(\phi_n \mathbf{v}^{\mathsf{T}}\mathbf{x}_n\right)}{1 + \exp\left(\mathbf{v}^{\mathsf{T}}\mathbf{x}_n\right)} \,. \tag{2}$$

Usually the number of parameters $p = I + 1$ is much larger than the number of samples N ($p \gg N$). Hence, a standard logistic regression approach in which we maximize the likelihood with respect to the effect sizes will lead to gross overtraining on the training data and poor prediction performance on unseen test data.

When learning the model parameters using a maximum likelihood approach in the limit $p \gg N$, one common solution is to add a regularization term to the log likelihood that will push most of the components of $\mathbf{v}$ to zero, or near zero (such as an L_1 regularizer $-\lambda||\mathbf{v}||_1$ or an L_2 regularizer $-\lambda||\mathbf{v}||_2^2$, respectively). From the viewpoint of Bayesian statistics, this approach can be motivated by noting that it is equivalent to maximizing the posterior distribution $p(\mathbf{v} \mid \mathbf{X}, \lambda)$ because according to Bayes' theorem it is proportional to $p(\mathbf{X} \mid \mathbf{v})\, p(\mathbf{v})$. Maximizing the logarithm of the posterior distribution is therefore equivalent to maximizing the log likelihood plus a regularization term $\log p(\mathbf{v})$. Obviously, the L_1 and L_2 regularizers correspond to a Laplace and a normal prior distribution, respectively. Because our prior expectation is that the overwhelming majority of SNPs will have a negligible effect on disease risk, the prior $p(\mathbf{v})$ should have most of its weight narrowly distributed around zero.

We choose to model the prior probability of effect sizes with a more descriptive and realistic two-component Gaussian mixture distribution, in which one component representing the effect sizes of the non-causal SNPs is sharply peaked around $v_i = 0$ and and the second much wider component describes SNPs coupled to the phenotype:

$$p\left(v_i \mid \boldsymbol{\theta}\right) = (1-\pi_i)\,\mathcal{N}\left(v_i \mid 0, \sigma_{\mathrm{bg}}^2\right) + \pi_i\,\mathcal{N}\left(v_i \mid \mu, \sigma^2\right). \tag{3}$$

This prior is similar to various other two-component mixture priors (see [9] for an overview), e.g. the one in BVSR [3], which used a delta function at $v_i = 0$ for the non-causal SNPs and a Students t-distribution for the causal ones. In our prior, $\boldsymbol{\theta} = (\boldsymbol{\beta}_\pi, \mu, \sigma)$ are the model hyperparameters with $\boldsymbol{\beta}_\pi$ defined below. The parameters π_i, as in BVSR, control the sparsity of the model. We set σ_{bg}^2 to 0, corresponding to a delta function at $v_i = 0$. This reduces our prior to a point-normal distribution. The parameter σ^2 describes the variance of the effect size of the causal variants. We assume that all causal SNPs will have the same variance, for which we assume a uniform prior, $p(\sigma^2) = \text{const}$.

We further define $z_i \in \{0, 1\}$ as the hidden indicator variables defining the underlying causality of the SNPs. Here, $z_i = 1$ indicates that SNP i is causal and $z_i = 0$ indicates that it is not. To simplify notations, we define the vectors $\boldsymbol{\mu}_{\mathbf{z}}$ and $\boldsymbol{\sigma}_{\mathbf{z}}^2$, whose i^{th} components are $\boldsymbol{\mu}_{\mathbf{z},i} = z_i\mu$ and $\boldsymbol{\sigma}_{\mathbf{z},i}^2 = \sigma_{\mathrm{bg}}^2 + z_i(\sigma^2 - \sigma_{\mathrm{bg}}^2)$ respectively. This allows us to reformulate Eq. (3) as:

$$p\left(\mathbf{v} \mid \boldsymbol{\theta}\right) = \sum_{\mathbf{z}} \left(\prod_{i=1}^{I} \pi_i^{z_i}\,(1-\pi_i)^{(1-z_i)}\right) \mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \mathrm{diag}(\boldsymbol{\sigma}_{\mathbf{z}}^2)\right)$$

with the sum running over all 2^I possible *causality configurations* $\mathbf{z} \in \{0, 1\}^I$. Using

$$p\left(\mathbf{z} \mid \boldsymbol{\theta}\right) = \prod_{i=1}^{I} \pi_i^{z_i}\,(1-\pi_i)^{(1-z_i)} \tag{4}$$

we can write the prior on the effect sizes as,

$$p\left(\mathbf{v} \mid \boldsymbol{\theta}\right) = \sum_{\mathbf{z}} p\left(\mathbf{z} \mid \boldsymbol{\theta}\right)\,\mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \mathrm{diag}\left(\boldsymbol{\sigma}_{\mathbf{z}}^2\right)\right). \tag{5}$$

The above formulation also helps us to realize that $\pi_i = p\left(z_i = 1 \mid \boldsymbol{\theta}\right)$, which gives the probability of i^{th} SNP being causal before observing the data.

In the simplest case the prior probability of all SNPs are same, *i.e.* $\pi_i = \pi = \text{const}$. Here, we consider a more flexible model in which the π_i can depend on possibly informative local genomic features or annotations of genetic variants by functional consequences. These could potentially help causal inference for the SNPs because they bring independent sources of underlying biological information about each SNP. Data tracks with informative features are becoming available for ever more cell types, *e.g.* ENCODE data on histone modifications, chromatin accessibility enhancer, promoter and coding region annotation [10], summary features from DeepSEA [11], etc. We lump together these additional features into vectors $\boldsymbol{\xi}_i$ $(i \in \{1, \ldots, I\})$, each with K elements or features. We model the dependency of π_i on these functional features as,

$$\pi_i = \frac{1}{1 + \exp\left(-\boldsymbol{\xi}_i^{\mathsf{T}} \boldsymbol{\beta}_\pi\right)} \tag{6}$$

We also add a $(K+1)^{\text{th}}$ “baseline” annotation of $\boldsymbol{\xi}_{i,0} = 1$ for all SNPs whose corresponding coefficient $\beta_{\pi,0}$ can be interpreted as the prior odds for causality of any SNP. Only a few of these $K+1$ features contribute to determine the weights, and we enforce this sparsity by introducing a Laplace hyperprior $p\left(\boldsymbol{\beta}_\pi\right) = \prod_f \exp(-\alpha|\beta_{\pi,f}|)$ with $\alpha = 0.75$, and f runs over all the $K+1$ feature tracks. The dependency of π_i on $\boldsymbol{\beta}_\pi$ (Eq. (6)) can be easily modified, and similar dependencies can be introduced for the mean and variance of the effect size prior.

Inference

Statistical finemapping of causal variants in each locus. The posterior probability for SNP i to be coupled to the disease is obtained by summing the posterior probability over all causality configurations $\mathbf{z}$ for which SNP i is causal (*i.e.* $z_i = 1$):

$$p(z_i = 1 \mid \boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\theta}) = \sum_{\mathbf{z}:z_i=1} p(\mathbf{z} \mid \boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\theta}) \tag{7}$$

Prediction of causality of each locus. Similarly, the probability for a locus to be coupled with the disease phenotype is equal to the probability of the locus harboring at least one causally associated SNP. This is equal to 1 minus the probability of not containing a single causal SNP:

$$\mathrm{Pr}_{\mathrm{causal}} = p(\text{locus is causal} \mid \boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\theta}) = 1 - p(\mathbf{z} = \mathbf{0} \mid \boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\theta}) \tag{8}$$

Both tasks require computing

$$p(\mathbf{z} \mid \boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\theta}) = \frac{p(\boldsymbol{\phi}, \mathbf{z} \mid \mathbf{X}, \boldsymbol{\theta})}{p(\boldsymbol{\phi} \mid \mathbf{X}, \boldsymbol{\theta})} = \frac{p(\boldsymbol{\phi}, \mathbf{z} \mid \mathbf{X}, \boldsymbol{\theta})}{\sum_{\mathbf{z}'} p(\boldsymbol{\phi}, \mathbf{z}' \mid \mathbf{X}, \boldsymbol{\theta})} \tag{9}$$

for any causality configuration $\mathbf{z}$, which in turn requires computing

$$\begin{aligned} p(\boldsymbol{\phi}, \mathbf{z} \mid \mathbf{X}, \boldsymbol{\theta}) &= p(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{z}, \boldsymbol{\theta})\, p(\mathbf{z} \mid \boldsymbol{\theta}) \\ &= p(\mathbf{z} \mid \boldsymbol{\theta}) \int p(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{v})\, p(\mathbf{v} \mid \mathbf{z}, \boldsymbol{\theta})\, d\mathbf{v} \\ &= p(\mathbf{z} \mid \boldsymbol{\theta}) \int p(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{v})\, \mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right) d\mathbf{v}\,. \end{aligned} \tag{10}$$

In contrast to the classical maximum likelihood approach, in which the parameters $\mathbf{v}$ are optimized, the above method integrates out the parameters $\mathbf{v}$. This is a crucial difference in practice, because it eliminates the need to learn a large number $p \gg N$ of parameters and thereby very effectively guards against overtraining. The only parameters that we still have to learn are the hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_\pi, \mu, \sigma)$. Also, by integrating out the parameters we avoid the errors incurred by fixing them to noisy point estimates. We will explain how to solve this integral in the next section.

Optimization of hyperparameters and quasi-Laplace approximation

To learn the hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_\pi, \mu, \sigma)$, we maximize the *marginal likelihood function* $mL(\boldsymbol{\theta}) := p(\boldsymbol{\phi} \mid \mathbf{X}, \boldsymbol{\theta}) = \int p(\boldsymbol{\phi} \mid \mathbf{v}, \mathbf{X}, \boldsymbol{\theta})\, p(\mathbf{v}|\theta) d\mathbf{v}$. This *empirical Bayes* approach [12] is usually robust against overtraining when only few hyperparameters need to be learned from the data. Inserting Eq. (10) into the marginal likelihood yields

$$mL(\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\boldsymbol{\phi}, \mathbf{z} \mid \mathbf{X}, \boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z} \mid \boldsymbol{\theta}) \int p(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{v})\, \mathcal{N}\left(\mathbf{v} | \boldsymbol{\mu}_{\mathbf{z}}, \operatorname{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{z}}\right)\right) d\mathbf{v}\,. \tag{11}$$

The integral on the right hand side does not have an exact solution. A common approach to solve such integrals is to approximate the integrand with a multivariate Gaussian using Laplace's method. The parameters of the Gaussian can be simply determined by finding the integrand's mode using gradient-based optimization and setting the precision matrix to the Hessian at the mode. Unfortunately, the mode depends on the causality configuration $\mathbf{z}$ and the hyperparameters $\boldsymbol{\theta}$. That means we would need to determine mode and precision matrix for every $\mathbf{z}$ in the sum and every time $\boldsymbol{\theta}$ is changed, which is clearly infeasible. One might instead approximate only the likelihood by a Gaussian. This is also problematic because the approximation will become inaccurate as we move away from the mode of the likelihood, and unfortunately the region in $\mathbf{v}$ space which contributes most to the integrand (around the mode of the integrand) can be quite far from the mode of the likelihood.

We propose a novel approximation ("quasi-Laplace approximation") by splitting the integrand into two factors, a *regularized likelihood* that closely approximates the integrand but does not depend on $\boldsymbol{\theta}$ or $\mathbf{z}$ and a correction term that depends on $\boldsymbol{\theta}$ and $\mathbf{z}$:

$$\begin{aligned} &p\left(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \operatorname{diag}\left(\boldsymbol{\sigma}_{\mathbf{z}}^2\right)\right) \\ &= p\left(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}^2\right)\right) \times \frac{\mathcal{N}\left(\mathbf{v} \mid \boldsymbol{\mu}_{\mathbf{z}}, \operatorname{diag}\left(\boldsymbol{\sigma}_{\mathbf{z}}^2\right)\right)}{\mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}^2\right)\right)} \end{aligned} \tag{12}$$

Here, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}$ are constants whose values will be estimated from the data. We approximate the regularized likelihood by a multivariate Gaussian:

$$p\left(\boldsymbol{\phi} \mid \mathbf{X}, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}^2\right)\right) \propto \mathcal{N}\left(\mathbf{v} \mid \tilde{\mathbf{v}}, \tilde{\boldsymbol{\Lambda}}^{-1}\right) . \tag{13}$$

Since the regularized likelihood depends neither on $\mathbf{z}$ nor on $\boldsymbol{\theta}$, we have to perform the gradient-based optimization only once to determine the *summary statistics* $\tilde{\mathbf{v}}$ and $\tilde{\boldsymbol{\Lambda}}$. The regularizer $\mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}^2\right)\right)$ acts as an approximate, simple prior distribution. We learn it from the data in an iterative way (usually two iterations suffice) by maximizing the marginal likelihood given a first estimate of $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}$. The above approximation allows us to analytically solve the integral of Eq. (11) (see the detailed derivation in S1 File).

The regularized likelihood is well approximated by a Gaussian when $N \gg 1$ and the number of cases and controls is similar. To see this, we note that the regularized log likelihood $LL_{\text{reg}}(\mathbf{v}) = \log p(\boldsymbol{\phi}|\mathbf{X}, \mathbf{v}) + \log \mathcal{N}(\mathbf{v}|\boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\sigma}^2))$ is the sum of N concave functions $f_n(\mathbf{v}) = -\log(1 + \exp(\pm\mathbf{v}^{\mathsf{T}}\mathbf{x}_n))$ (see Eq. (2)) plus a quadratic function with respect to $\mathbf{v}$. The Hessians of the concave functions must all have negative or zero diagonal elements and therefore their sum will grow roughly proportionally with the number of patients N. In contrast to the second derivatives, the third and higher partial derivatives will take both positive and negative signs. If the number of diseased and control patients is roughly equal, $p(\phi_n|\mathbf{x}_n, \mathbf{v})$ will mostly lie near $(1/N)\sum_n I(\phi_n = 1) \approx 0.5$, and therefore $\mathbf{v}^{\mathsf{T}}\mathbf{x}_n$ will be roughly as often positive as negative. Therefore the third partial derivatives will tend to be close to zero and have no preferred signs. The same is true of the higher derivatives. The magnitudes of the third and higher derivatives will grow only as $\sqrt{N}$ because their signs fluctuate around 0 for all patients. The second derivatives will increasingly dominate over the higher derivatives as N gets larger, and the log likelihood will be increasingly better approximated by a quadratic function, or in other words, by the logarithm of a multivariate Gaussian.

In summary, B-LORE works in two steps:

1. *Calculate summary statistics for each cohort*. This, in turn, requires two optimizations:
   - Learn $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\sigma}}$ from the data.
   - Learn $\tilde{\mathbf{v}}$ and $\tilde{\boldsymbol{\Lambda}}$ from the data by gradient-based maximization of the logarithm of the regularized likelihood obtained from Eqs. (13) and (2) and setting $\tilde{\mathbf{v}}$ to its mode and $\tilde{\boldsymbol{\Lambda}}$ to the negative of the Hessian matrix at the mode (see S1 File for details).
2. *Meta-analysis using summary statistics*. In this step, B-LORE optimizes the hyperparameters $\boldsymbol{\theta}$ by maximizing the marginal likelihood combining the summary statistics from multiple studies (see below).

In our software, the first step can be run using the command `-summary` and the second step can be run using the command `-meta`.

Factorization over loci

To speed up B-LORE analysis, we recommend to preselect loci with a faster method such as SNPTEST [5–7] and to include SNPs from these preselected loci. Usually these candidate loci will be in linkage equilibrium since LD is highly local. Therefore the covariance matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is approximately block-diagonal, with each block corresponding to a locus. This allows us to factorize the marginal likelihood in Eq. (11) as a product over all the loci (see S1 File). We can therefore calculate the marginal likelihood for each locus independently, which makes the evaluation of the sum over causality configurations and learning the hyperparameters from the summary statistics $\tilde{\mathbf{v}}$ and $\tilde{\boldsymbol{\Lambda}}$ quite efficient.

We note here that several other methods utilize this block-diagonal feature of LD matrix of genotype. For example, BIMBAM [4] uses a factorization over loci to perform multiple regression at each locus independently. However, it does not learn the hyperparameters from the data. Hence it does not need to jointly analyze multiple loci and can compute summary statistics for each locus separately. In contrast, B-LORE analyzes all loci jointly, which requires the summary statistics to be computed for all loci jointly.

Meta-analysis of many studies

The likelihood for a single study is given in Eq. (2). We can combine multiple independent studies simply by computing the total likelihood as the product of the likelihoods of each contributing study s:

$$p\left(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_S \mid \mathbf{X}_1, \ldots \mathbf{X}_S, \mathbf{v}\right) = \prod_{s=1}^{S} p\left(\boldsymbol{\phi}_s \mid \mathbf{X}_s, \mathbf{v}\right) \tag{14}$$

The integrand in Eq. (11) will now have a product over multiple logistic functions. For each study, we estimate the regularizer of the likelihood $\mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}_s, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}_s^2\right)\right)$ and the summary statistics $\tilde{\mathbf{v}}_s$ and $\tilde{\boldsymbol{\Lambda}}_s$. We apply the quasi-Laplace approximation for each study:

$$\prod_{s=1}^{S}\left[p\left(\boldsymbol{\phi}_s \mid \mathbf{X}_s, \mathbf{v}\right) \mathcal{N}\left(\mathbf{v} \mid \tilde{\boldsymbol{\mu}}_s, \operatorname{diag}\left(\tilde{\boldsymbol{\sigma}}_s^2\right)\right)\right] \propto \prod_{s=1}^{S}\left[\mathcal{N}\left(\mathbf{v} \mid \tilde{\mathbf{v}}_s, \tilde{\boldsymbol{\Lambda}}_s^{-1}\right)\right] \tag{15}$$

where,

$$\tilde{\boldsymbol{\Lambda}} = \sum_{s=1}^{S} \tilde{\boldsymbol{\Lambda}}_s \quad \text{and} \quad \tilde{\mathbf{v}} = \tilde{\boldsymbol{\Lambda}}^{-1} \sum_{s=1}^{S} \tilde{\boldsymbol{\Lambda}}_s \tilde{\mathbf{v}}_s\,. \tag{16}$$

We use this approximation to calculate the total marginal likelihood for all studies, and optimize it to obtain estimates of the hyperparameters. We can then perform all the subsequent analyses using the optimized hyperparameters. Unlike conventional meta-analysis methods which pool aggregate allele count data of each individual SNP, the above method allows us to combine information from multiple regression. For details of the derivations see S1 File.

Population stratification and other covariates

Population stratification presents a major challenge in the design and analysis of GWAS. As discussed in the S1 File, current implementation of B-LORE uses a principal component analysis (PCA) for correcting population substructure. The principal components are obtained from the genotype matrix and used as covariates in the model. For all our analysis, we have used 20 principal components. The PCA-based correction assumes that the principal coordinates enter linearly in the argument of the logistic function. Similarly, other external covariates such as age, sex, etc. can be specified in the sample file and will also enter linearly in the argument of the logistic function.

Datasets

To illustrate B-LORE, we used the genotype from five German population cohorts: German Myocardial Infarction Family Study (GerMIFS) I - V [13–18]. Details for quality control and pre-processing of these datasets were described by Nikpay *et al.* [18]. Briefly, there were a total of 6234 cases and 6848 controls with white European ancestry. Each cohort was imputed with phased haplotypes from the 1000 Genomes Project. SNPs were filtered for MAF > 0.05 and HWE p-value > 0.0001. To test the ability for using genome-wide functional genomics tracks to help distinguish causal from merely correlated SNPs, we used DNase-seq data to measure DNA accessibility, published by the ENCODE project [10]. We normalized the DNase-seq data for 112 human samples, as described previously by Sheffield *et al.* [19].

Simulation framework

Simulation studies are popular to evaluate different methods of statistical analyses in GWAS, because they are inexpensive and they give us access to the "ground truth". In our case, we needed to know which SNPs or which loci are causal in the population. The inherent complexity of the genotype data with strong linkage effects are very difficult to simulate realistically from haplotype data. We therefore used real patient genotypes and real DNase-seq tracks in our semisynthetic benchmarking test. We simulated the phenotypes using genomic features as described previously by Kichaev *et al.* [20].

We randomly selected 200 loci, each with 200 SNPs from the whole genome. Random selection of a locus was done by selecting a random SNP from the whole genome and using the chosen SNP plus the nearest 199 SNPs as the locus. While SNPs within a locus can have strong LD, we made sure that all SNP pairs between different loci have LD $r^2 < 0.8$.

Each SNP had 112 functional genomics features, denoted by $\boldsymbol{\xi}_i$. For each simulation, we randomly selected 3 features as significant. We then defined a baseline probability $\pi_0 = 1/(1 + \exp(-\beta_{\pi,0}))$. The enrichment induced by the k^{th} feature can be defined as $\psi_k = \pi_k/\pi_0$, where $\pi_k = 1/(1 + \exp(-\beta_{\pi,0} - \beta_{\pi,k}))$ assuming that the corresponding feature is binary (*i.e.* 1 for enriched SNPs, and 0 for other SNPs). For each selected feature, we randomly chose ψ_k from a uniform distribution between 2 and 8, and calculated the corresponding $\beta_{\pi,k}$. The $\beta_{\pi,k}$ for the remaining 109 cell lines were set to zero. Next, we calculated $\pi_i = 1/\big(1 + \exp(-\boldsymbol{\xi}_i^{\mathsf{T}}\boldsymbol{\beta}_\pi)\big)$ which gives the probability for each SNP to be causal. The prior probability of a locus to be causal is equal to 1 minus the probability of not containing a single causal SNP, which can be obtained as $p_c = 1 - \prod_i(1 - \pi_i)$ where i runs over all SNPs in a given locus. We ranked the 200 loci by this prior probability p_c, and chose the top 100 loci as causal. For each of these causal loci, we sampled causal SNPs with the probability π_i. This gave us a "ground truth" of 100 causal loci and corresponding causal SNPs for each simulation. The number of causal SNPs depended on the choice of π_0 and the number of features. For example, $\pi_0 = 0.01$ and 3 features gave approximately 450 causal SNPs out of 40000 SNPs used for each simulation.

Once we established the causal SNPs, we used a linear model to simulate continuous phenotypes $y_n = \sum_i v_i x_{in} + \varepsilon_j$ such that the causal SNPs aggregated to explain a fixed proportion of the phenotypic variance h_g^2. This phenotypic variance was partitioned equally amongst all the causal SNPs. The environmental contribution given by ε_j, was assumed to be normally distributed $\varepsilon_j \sim \mathcal{N}\left(0, 1 - h_g^2\right)$. In our simulations, we used a heritability of $h_g^2 = 0.25$ as is used typically in GWAS simulations [20].

We obtained the binary phenotype for the 5 cohorts using the classical liability threshold model [21, 22]. The model assumes that the binary disease status results from an underlying continuous disease liability that is normally distributed in the population. If the combined effects of genetic and environmental influences push an individual's liability across a certain threshold level, the individual is affected. In the population, the proportion of individuals with a liability above the threshold is reflected in the disease prevalence. The observations on the risk scale and the liabilities on the unobserved continuous scale can be related by a probit transformation [22]. We used the continuous phenotype y_n as the disease liability. Any individual with disease liability exceeding a certain threshold T was assigned to be a case and a control otherwise. T is the threshold of normal distribution truncating the proportion of K (disease prevalence). We used $K = 0.5$ to obtain roughly equal number of cases and controls.

We repeated the simulations 50 times. For all simulations, we used the same genotype and functional annotations. We resampled which functional features are significant, which loci and which SNPs are causal and simulated new phenotype for every repetition of the simulation.

Results

Prediction of causal loci

We tested two widely used software packages employing conventional methods for association testing in GWAS: (1) SNPTEST [5–7] on each cohort, followed by meta-analysis using META [23] (designated as SNPTEST / META henceforth). (2) Meta-analysis with BIMBAM [4] using the

-psd option to collect summary statistics on each cohort and -ssd to combine them. BIMBAM can perform multiple regression and was earlier shown to outperform SNPTEST when given access to individual genotype data. However, it falls back to simple single-SNP regression when performing meta-analysis.

Furthermore, we tested the recent metaCCA software, which can perform multivariate meta-analysis from summary statistics. The method requires choosing the relevant SNPs beforehand since it has no shrinkage in the model for variable selection [8]. We ranked the SNPs based on p-values obtained from SNPTEST / META, and chose the 10 best SNPs successively such that each chosen SNP had little correlation $\left(r^2 \leq 0.8\right)$ with all previously chosen SNPs. In other words, once a SNP is chosen, all SNPs which had correlation $\left(r^2 > 0.8\right)$ were removed from the ranked list. The method also requires the matrix of LD correlation coefficients between the selected SNPs at each locus for each study. We calculated the LD matrix for each study directly from the individual level genotype using LDstore [24].

It was shown previously that genome-wide annotations and data tracks such as DNA accessibility measurements from DNase-seq are enriched in causal SNPs [25, 26] and can be used for improving the finemapping of causal SNPs [20, 27]. We therefore extended B-LORE, making the mixture weight π_i of the causal part of the effect size distribution for SNP i depend on these functional genomics data tracks at the position of SNP i (see Eq. (6)). We ran B-LORE with and without using the genome-wide functional genomics tracks and denoted the latter as B-LORE FG.

Fig. 2 summarizes the performance of different methods in predicting the causality of loci using synthetic phenotypes. First, we compared the ranking of loci by SNPTEST/META and B-LORE (Fig. 2A). For SNPTEST/META, loci were scored by the $-\log_{10}(p)$ value of their most significant SNP. For B-LORE, loci were scored by the easily interpretable posterior probability of the locus to be causal, i.e., to contain at least one causal SNP (Eq. (8)). The two scores agreed well on those loci that are confidently predicted by SNPTEST/META: All loci with high $-\log_{10}(p)$ values (> 5) also got high posterior probabilities (> 0.95), and all loci with low $-\log_{10}(p)$ values (< 1) also get low posterior probabilities (< 0.05). However, a sizeable fraction of loci with high B-LORE posterior probabilities (> 0.95) had low significance in SNPTEST/META ($-\log_{10}(p)$ values < 3) even though they were all causal. Loci with $-\log_{10}(p)$ values between 1.0 and 4.0 were classified with higher accuracy by B-LORE. Overall, the noncausal and causal loci were much better separated by B-LORE scores along the vertical axis than by SNPTEST/META along the horizontal one.

For a more quantitative analysis of the prediction performance, we performed a precision-recall analysis (Fig. 2B). For each of the prediction tools, the 50×200 scores were ranked, the true positive (TP) and false positive (FP) predictions were counted up to different threshold scores, and the precision (TP/(TP + FP)) was plotted as a function of recall (= sensitivity) (TP/(TP + FN)). BIMBAM and SNPTEST / META perform meta-analyses on single-SNP summary statistics and provide similar accuracy. Surprisingly, metaCCA failed to improve the classification in our benchmarks, probably due to improper preselection of the relevant SNPs. B-LORE significantly improved the ranking of loci over the next best tool. The results suggest that multiple regression can extract more information from the data. B-LORE FG further improved the prediction of causal loci, suggesting that our tool can efficiently incorporate information from functional genomics tracks.

To demonstrate the advantage of distinguishing coupling from correlation, we chose 8 loci from a strongly correlated genomic region of chr6, while the remaining 192 loci were allowed to remain in linkage equilibrium. We also constructed synthetic phenotypes in 50 simulations as before. This ensured that these 8 loci would be correlated among them, and would lead to situations as described in Fig. 1: There would be some noncausal SNPs in a noncausal locus which are correlated with causal SNPs in a nearby locus. B-LORE was much less affected as compared to other methods in presence of such correlations (Fig. 2C). All other methods found many non-causal loci to be highly significant because they were correlated with the coupled loci. leading to many false positive predictions with high significance, seen by the dip at low recall values.

The successful predictions of B-LORE might seem counter-intuitive at first because it neglects the interlocus genetic correlation in order to factorize over loci for the hyperparameter optimization. We explain in Fig. 2D how B-LORE can still trace the correlation / coupling of loci. The schematic figure shows the effect size of 2 correlated SNPs from different loci: a non-causal SNP i in locus

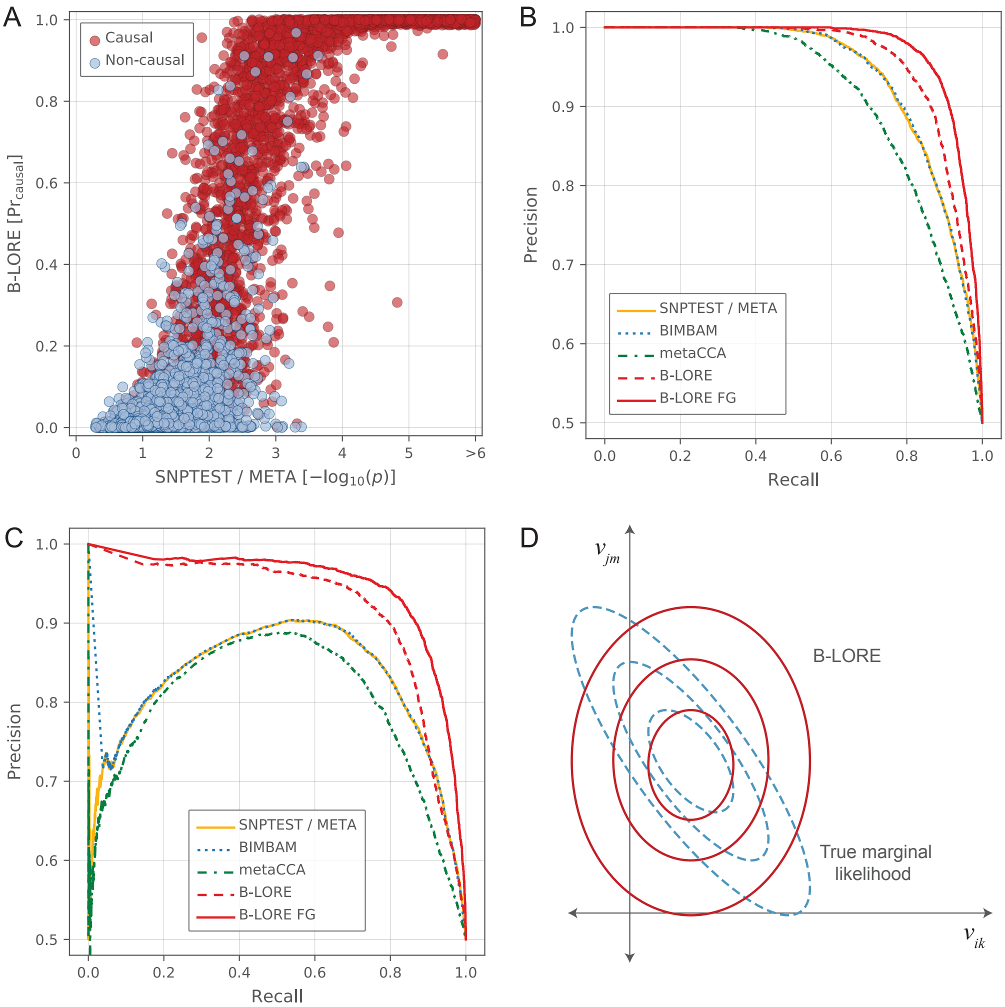

Figure 2. B-LORE improves prediction and ranking of causal loci. We performed 50 simulation runs using 200 loci each with 200 SNPs, obtained from 13082 patient genotypes from the GerMIFS I-V studies. In each simulation run, we chose 100 out of 200 loci to be causal, corresponding causally associated SNPs and their effect sizes, and simulated the case/control phenotypes accordingly (see Simulation framework for details) (A) Scatter plot of significance scores predicted by SNPTEST/META and by B-LORE for the 50×200 causal (red) and non-causal (light blue) loci. (B) Precision-recall curves quantifying the power of 5 methods to predict the 100 causal loci among the 200 loci in each of 50 runs. All loci are sampled such that they are in linkage equilibrium with all others. (C) Same as B, but including 8 loci that are in LD with one another. (D) Schematic diagram to explain how B-LORE can distinguish merely correlated loci from causal loci, in spite of assuming different loci to be in linkage equilibrium. We show the effect sizes of 2 correlated SNPs: a non-causal SNP i in locus k along the $x-$axis and a causal SNP j in locus m along the $y-$axis, their true marginal likelihood (dotted blue contour lines) and the marginal likelihood learned by B-LORE (solid red contours).

k along the $x-$axis and a causal SNP j in locus m along the $y-$axis. The true marginal likelihood of these 2 SNPs shows their correlation (non-zero off-diagonal terms in the covariance matrix). In the calculation of B-LORE the off-diagonal terms are assumed to be zero because they are in different loci. Still, it learns the diagonal terms of the covariance matrix because it maximizes the genome-wide marginal likelihood which includes information from all the loci. The likelihood learned by B-LORE is a fairly good approximation of the true likelihood and is obviously better than getting point estimates for each SNP. It can distinguish the two loci: v_{jm} explains away the effect of v_{ik} and only the true causal locus m is picked. In contrast, methods that deal with each locus separately will find associations for both m and k. Earlier attempts of distinguishing correlation and coupling were confined to each locus independently, for example, in finemapping of causal variants within a locus.

Statistical finemapping of causal SNPs

Prioritization of variants within the associated region or loci (popularly called finemapping) has been an important focus of post-GWAS era to provide insight into disease mechanism. For a recent review of finemapping, see Spain *et al.* [28]. Over the past few years, several methods have been proposed for finemapping of causal variants, *i.e.* to pinpoint individual causal SNPs. We compared B-LORE with 3 finemapping methods, particularly PAINTOR (v3.0) [20, 29], CAVIARBF (v0.1.4.1) [30, 31] and FINEMAP (v1.1) [32]. These methods require only the summary test statistics and a matrix of the pairwise correlation coefficients ($\mathbf{r}^2$) of the variants in each associated region. We used the meta-analysis results from SNPTEST / META as input to these methods. To obtain $\mathbf{r}^2$ of the meta study, we combined the study-specific LD matrices by weighting them by sample size, *i.e.* $\mathbf{r} = \sum_j (\mathbf{r}_j N_j)/\sum_j N_j$, where $\mathbf{r}_j$ and N_j are the correlation coefficient matrix and sample size for the j^{th} study respectively. For all methods, we used the default settings. Owing to computational limitations, we allowed a maximum of 2 causal SNPs per locus in CAVIARBF, and 4 causal SNPs per locus in FINEMAP and B-LORE. We also allowed B-LORE and PAINTOR to use the functional genomics tracks, and denoted them as B-LORE FG and PAINTOR FG respectively.

B-LORE achieved superior accuracy over existing methodologies in identifying causal variants (Fig. 3). For each of the 50 simulations, we ranked the 40000 scores from each of the prediction tools. In Fig. 3A, we plotted the recall (TP/(TP + FN)) against the average number of SNPs which were selected per locus ((TP + FN) /200) at different threshold scores. All characteristic finemapping methods (*i.e.* PAINTOR, CAVIARBF and FINEMAP) performed better than conventional GWAS methods (*i.e.* SNPTEST/META and BIMBAM) when selecting less than 20 SNPs per locus on average. The recall of B-LORE was higher than all other methods at practically relevant low selection thresholds. B-LORE improved the recall by more than 5% from the next best tool when selecting 20 SNPs per locus.

Using ENCODE data for the SNPs significantly improved the performance of B-LORE and PAINTOR. Both methods use the same logistic model (Eq. (6)) to describe the enrichment of causality of a SNP from functional genomic features. Yet B-LORE FG provided better recall than PAINTOR FG. B-LORE FG could identify more than 65% of the causal SNPs when selecting only 20 SNPs per locus, as compared to 55% by PAINTOR FG at the given threshold. It should be noted here that the gain in performance in simulation with functional genomics tracks does not ascertain similar gains in real data, because the true enrichment strengths are yet unclear.

All finemapping methods got worse than conventional GWAS methods at higher selection thresholds (see S3 Figure). Beyond a certain threshold, B-LORE predicts all SNPs with a causal probability of zero when it can no longer distinguish between them. This happens because learning hyperparameters from the data requires summing over causality configurations and we restrict the sum to those configurations with non-negligible probability of being causal to achieve the same in computationally reasonable time frame (see S1 File). Hence configurations without enough evidence are removed from the calculation and the corresponding SNPs get zero causal probability (Eq. (7)). Similarly, all other finemapping methods enforce sparsity in different ways and hence show computational artefacts at high thresholds. However, performance at high selection thresholds is not important for any practical purposes because the aim of finemapping is to pick as few SNPs as possible for downstream analyses and experiments.

In classification of sparse binary data, as the one we are dealing with here, recall can hide a lot of imprecision. However, improving precision presents more of a challenge. Hence, we did a precision-recall analysis of the predictions by the different tools (Fig. 3B). PAINTOR, CAVIARBF and FINEMAP had higher precision than conventional GWAS methods, but all 3 of them performed similarly. B-LORE had higher precision than all other tools at all recall values. Use of functional genomics tracks significantly improved the precision for B-LORE FG and PAINTOR FG. Overall, B-LORE FG provided higher precision than PAINTOR FG.

While finemapping aims to pick truly causal SNPs, it is worthwhile to avoid useless false positives but choose the ones which are correlated to a causal SNP. Hence, we looked at the fraction of false positives (with respect to the total number of false positives) which are in strong

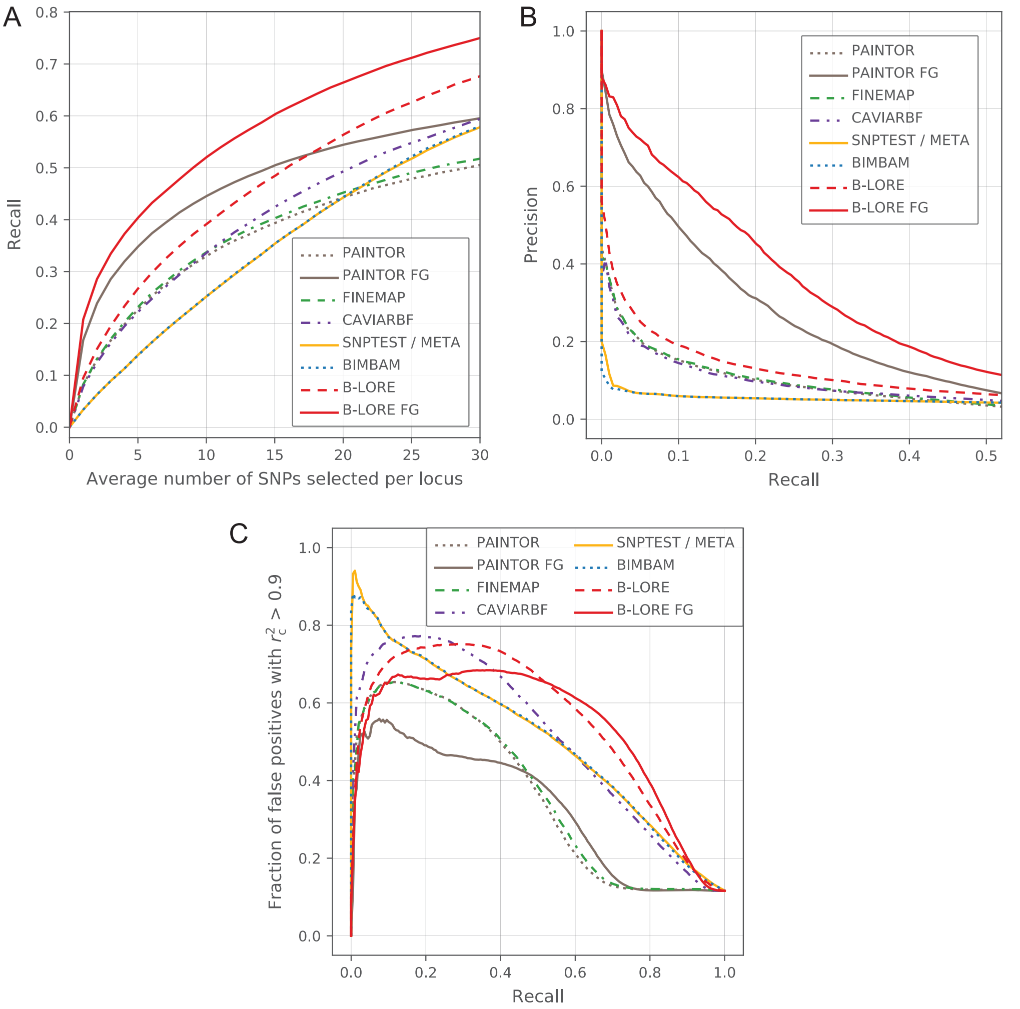

Figure 3. B-LORE with functional genomics tracks improves finemapping of SNPs. Finemapping results from 50 simulations (see Fig. 2 caption for summary and Simulation framework for details) We ranked the significance scores of the SNPs predicted by the different methods mentioned in the legends. A: Proportion of causal SNPs predicted by each method (recall) against the average number of SNPs which are selected per locus at different threshold scores. B: Precision-recall curves quantifying the power of each method to finemap *i.e.* to predict the causal SNPs among the 200 loci, averaged over 50 simulations. C: Fraction of false positives which are in strong LD (correlation coefficient $r_c^2 > 0.9$) with a true causal SNP in the locus as a function of recall, at different selection thresholds. This gives an idea of the "quality" of false negatives.

LD (correlation coefficient $r_c^2 > 0.9$) with a true causal SNP (Fig. 3C). Approximately 10% of the SNPs were correlated to a causal SNP averaged over all simulations. When recall was > 0.3, the false positives from B-LORE had the highest fraction of correlated SNPs as compared to all other methods. B-LORE not only predicted the true positives with high precision, but also the false positives were strongly correlated to actually causal ones.

Application to coronary artery disease

As a proof of concept, we applied B-LORE to identify SNPs affecting risk of coronary artery disease (CAD) from GWAS of five small cohorts (GerMIFS I - V). These cohorts were also used in the largest meta study of CAD GWAS till date, organized by CARDIoGRAMplusC4D Consortium [18], which found 58 significant loci harboring SNPs in genome-wide significant association with CAD. The study leveraged the power from meta-analysis of ~ 185,000 CAD cases and controls, while the GerMIFS I-V cohorts have only ~ 13,000 CAD cases and controls.

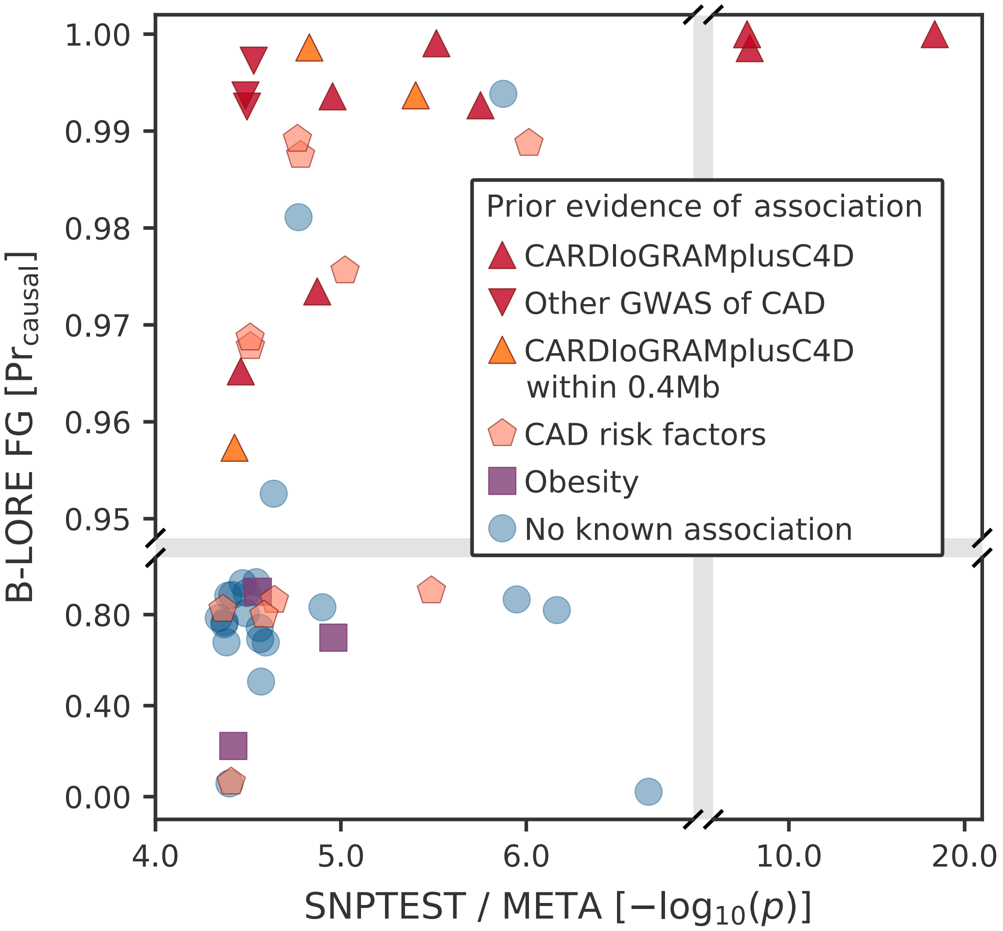

Figure 4. Association of genetic loci with CAD. Comparison of ranking of 50 genetic loci using meta-analysis across 5 cohorts (GerMIFS I-V [13–18]) with a total of 6234 cases and 6848 controls from white European ancestry. We first used meta-analysis of genome-wide SNPTEST summary statistics on these 5 small GWAS to select the top 50 loci and then applied B-LORE on these loci using the genomic features obtained from ENCODE data and assuming a maximum of 4 causal SNPs per locus. On the x-axis of the scatter plot, we show the $-\log_{10}(p)$ values obtained from META, and on the y-axis we show the probability of a locus being causal, obtained from B-LORE FG. The legend shows the classification of all the 50 CAD loci based on prior evidence of association in existing literature (see Application to coronary artery disease and S2 Table). This literature-based classification gives a reasonable "ground truth" of causal and non-causal loci, despite our incomplete knowledge about true underlying association in reality. The noncausal and causal loci are much better separated by B-LORE FG scores along the vertical axis than by SNPTEST/META along the horizontal one.

We performed a meta-analysis using SNPTEST summary statistics of the five cohorts, and found 3 loci to be genome-wide significant, namely the 9p21 locus on chr9, PHACTR1 and SLC22A3-LPAL2-LPA loci on chr6. From this meta-analysis, we ranked the SNPs according to their p-values and selected all the nominally significant SNPs with p-values $< 5 \times 10^{-5}$. We grouped these SNPs together based on their genomic positions. SNPs which were spatially close (within ± 200 kb) were included in the same locus. We then used the top 50 groups, and defined each locus by collecting the top 400 SNPs (ranked according to their p-values) within ± 200 kb regions of each lead group.

We used these 50 loci for meta-analysis with B-LORE. In practice however, one can use as many loci as desired and use more sophisticated definition of a locus, for example, by considering LD blocks. We generated summary statistics for each of the 5 cohorts and combined them using

B-LORE meta-analysis. For the meta-analysis we used the 112 functional genomics tracks from ENCODE data (see Datasets) and allowed a maximum of 4 causal SNPs per locus.

Since the ground truth is unknown for real data, we did an extensive blinded literature search for all these 50 loci. For details about the genomic positions, exons in the region and prior evidence of association with CAD please see S2 Table. We classified these 50 loci into 6 categories based on this prior evidence:

- 8 loci harbor SNPs which were found to be statistically associated with CAD in the CARDIoGRAMplusC4D study [18], the largest GWAS of CAD till date.
- 3 loci have significantly associated CARDIoGRAMplusC4D SNPs within ± 400 kb [18]
- 3 loci harbor SNPs which were found to be statistically associated with CAD in other GWAS.
- 11 loci have evidence of statistical association with risk factors of cardiovascular diseases (CVD), such as myocardial infarction, blood pressure, sudden cardiac arrest, heart failure, high-density lipoprotein cholesterol levels, triglyceride levels, etc.
- 3 loci are associated with obesity and related traits.
- We could not find any statistical association with CAD or its risk pathway for the remaining 22 loci.

We compared the ranking of these loci using B-LORE and SNPTEST/META in Fig. 4, For ranking, we used the same scores as already introduced for Fig. 2A. Unlike simulations, we did not know the "ground truth" in this real data, but we used the literature classification as qualitative indication for accuracy. Despite the modest sample size, B-LORE identified all the 11 CARDIoGRAMplusC4D loci (from the first 2 categories) present in the selected pool of loci with $\mathrm{Pr}_{\mathrm{causal}} > 0.95$. For comparison, SNPTEST / META could identify only 3 genome-wide significant loci. Additionally, B-LORE predicted 12 other novel loci with $\mathrm{Pr}_{\mathrm{causal}} > 0.95$. Three of them are significant hits in other CAD GWAS, and 6 of them are significantly associated with CVD risk-related phenotypes, such as blood pressure, high density lipoprotein cholesterol, triglyceride levels, etc. B-LORE also predicted low probabilities for many loci with no prior evidence of association despite being highly ranked by SNPTEST / META.

Three loci with no prior evidence of association with CAD were ranked with posterior probabilities $\mathrm{Pr}_{\mathrm{causal}} > 0.95$ by B-LORE. One of these loci is located in the AUTS2 gene region in chr7, and another is located in chr10 overlapping with FGFR2 and ATE1 genes. The third one is located in chr4q13.1, with no exons in the region and the nearest gene being LPHN3 ~500kb upstream. Due to lack of validation, it is unclear whether these loci are truely coupled to the disease risk or if they are false positives.

We show the finemapping performance of B-LORE at two example loci – a known risk locus near SMAD3 in chr15 (Fig. 5A) and a novel risk locus at 12q24.31 (Fig. 5B). At the SMAD3 locus, B-LORE picked up a single SNP (rs34941176) for explaining the association, while other SNPs from the region showed negligible probability of being associated. In the novel locus, there are three genes, DNAH10, ZNF664 and CCDC92, which are believed to be associated with multiple CAD risk factors such as high-density lipoprotein cholesterol level, triglycerides levels and waist-to-hip ratio [33–35] Here, B-LORE prioritized three SNPs (rs1187415, rs7961449 and rs6488913) in a region with strong LD. All SNPs in this region showed similarly significant p-values. In both the above cases, B-LORE prioritized SNPs which are different than the ones with lowest p-values in the region.

Indeed in most loci we find that B-LORE prioritizes a few SNPs, while SNPTEST / META predict similar p-values for many SNPs (S4 Figure). Due to lacking validation it is unclear so far whether the SNPs prioritized by B-LORE are actually causal.

In many loci that obtained a high probability to be coupled to disease risk, none of the SNPs have significant probability of being causal (S4 Figure). This observation illustrates the considerable advantage of a Bayesian method that can accumulate evidence for coupling over many, sometimes weak, SNPs. In this way, highly confident predictions can result even though no single SNP is considered significant by single-SNP analyses.

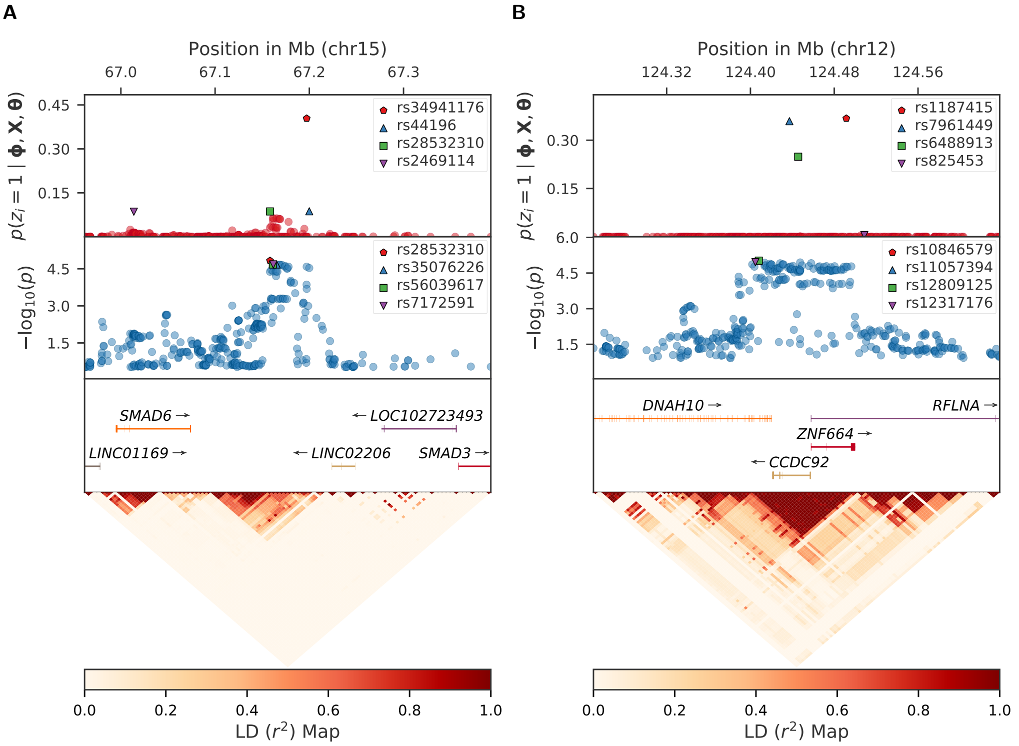

Figure 5. Representative examples of finemapping in CAD-associated loci. The top parts in A and B show the probability for each SNP to be causal as predicted by B-LORE ($p\,(z_i = 1 \mid \boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\theta})$). Below we plot the $-\log_{10}(p)$ values for each SNP obtained from SNPTEST / META. The four best SNPs predicted by B-LORE and SNPTEST / META are marked by special symbols and annotated in the legends. At the bottom, the genes and LD between the SNPs is shown. A: A known locus near SMAD3. (A SNP rs56062135 at 67.45Mb was found associated with CAD by the CARDIoGRAMplusC4D study [18]). The probability for finding at least one causal SNP in the locus is $\text{Pr}_{\text{causal}} = 0.999$. B: A novel locus discovered by B-LORE, with $\text{Pr}_{\text{causal}} = 0.976$.

Discussion

Any method for GWAS meta-analysis that can predict substantially more risk loci at a given precision or that can better distinguish the truly disease-coupled SNPs more accurately from the merely correlated ones has enormous leverage. It can be applied to thousands of GWAS involving millions of patients to help understand the origin of all common diseases in humans [2]. Most GWAS have been analyzed using the simplest type of approach based on testing each single variant in turn, which allows combining datasets using simple aggregate count summary statistics. Previous work has shown that multiple regression and also Bayesian approaches have the potential to yield better predictions [4], but their applicability was limited by the requirement of individual genotype data, which precludes meta analyses.

Our software B-LORE for Bayesian multiple logistic regression can perform meta-analysis using a novel type of summary statistics. B-LORE outputs easily interpretable probabilities for each locus to harbor causal SNPs, as well as probabilities for each SNP to be coupled, not only associated, with the phenotype. These probabilities contain all the information required for further downstream analyses.

This study makes the following technical contributions: (1) It introduces a novel *quasi-Laplace* approximation which makes the Bayesian treatment of the multiple logistic regression case analytically tractable and yields an efficient software implementation obviating the need to use MCMC sampling. (2) B-LORE learns the model hyperparameters from the GWAS data. One set of hyperparameters describes the effect size distribution, another (optional) set describes how the functional genomics tracks modify the prior probability of a SNP to be causal. (3) We show how to calculate the marginal likelihood over many loci by factorizing the likelihood over the loci. (4) The model can integrate genome-wide tracks from functional genomics and other sources.

B-LORE is similar to several successful GWAS analysis methods based on Bayesian multiple regression. It models the effect size distribution of the causal SNPs using a single normal distribution, whereas BVSR models it using a normal distribution whose precision (inverse variance) is sampled by MCMC using a Gamma prior. This is essentially equivalent to assuming a Student's-t prior for the causal component, because the latter is obtained as convolution of a normal distribution with a Gamma distribution for the precision parameter. The Student's-t distribution is more flexible, as its third parameter can be used to tune the heaviness of the distribution's tails. However, this more flexible prior requires computationally expensive Markov Chain Monte Carlo (MCMC) schemes for the integration over the effect sizes. Using a simpler prior allowed us to find an analytical solution.

The analytical integration also allows B-LORE to learn the parameters of the prior distribution from the GWAS data itself, whereas most existing Bayesian frameworks in GWAS, including BVSR, work with a fixed prior distribution of effect sizes. Another method that is able to learn the hyperparameters for the effect size distribution from the data is the Bayesian Sparse Linear Mixed Model (BSLMM) [9], which, similar to B-LORE, uses a mixture of two normal distributions. Any hyperparameter optimization method is limited by computational speed and the requirement of individual level data. Our new quasi-Laplace approximation provides a solution to both these problems. B-LORE thereby extends the scope of Bayesian multiple regression methods to hundreds of loci over hundreds of studies, where the loci can be preselected with a lenient p-value cutoff using a simpler and faster method such as SNPTEST [5–7].

In parallel work, Zhu and Stephens have proposed a clever method with similar aims as B-LORE but adapted to quantitative phenotypes using linear multiple regression, called Regression with Summary Statistics (RSS) [36]. RSS uses summary statistics from simple regression methods like SNPTEST [5–7]. Unlike RSS, our work uses logistic regression and is thus specifically adapted to the case-control GWAS design, by far the most frequent type of GWAS.

B-LORE is robust and should perform well on any sufficiently large dataset. An important caveat is that B-LORE needs a sufficient number of causal variants to be present in the GWAS data it analyzes, because it needs to estimate the hyperparameters from them. If too few causal loci are hidden in the data, the hyperparameter optimization could end up at unrealistic values, *e.g.* by using both components of the Gaussian mixture to describe the non-causal SNPs. In such a situation B-LORE would predict all SNPs as non-causal. We determined from simulations that it is enough to have only 10 coupled SNPs for proper learning of hyperparameters (data not shown). These SNPs could either be present in a single locus or distributed over many loci. This requirement should be easily fulfilled in all practical cases because the number of validated SNPs for most diseases investigated with GWAS are in the range of $20 - 100$.

A central goal of GWAS is to learn what are the genetic factors that determine the risk to acquire a noninfectuous disease or other, quantitative phenotypes. Yet despite ever larger GWAS and advances in analysing the genotype-phenotype relationship, only relatively small fractions of heritable variance can be explained for most diseases investigated [37]. One part of this *missing heritability* might be due to the limitations of logistic regression models used for risk prediction [38]. We hope to investigate this in the future by applying B-LORE to predict the genetic component of disease risk from the genotype. We will also explore more systematically how best to improve predictive performance by adding genome-wide experimental and computational data tracks [11,20,39] that can inform us about the probability of each SNP to be causal.

Supporting Information

S1 File. Additional Methods

S2 Table. Details of the 50 genetic loci used in this study, their literature classification and significance score for association with the diseases.

S3 Figure. Extension of Fig. 2A showing the full range of selection thresholds

S4 Figure. Finemapping of 50 genetic loci using B-LORE meta-analysis from 5 small GWAS (GerMIFS I-V)

Acknowledgments

We thank Christian Roth for pre-processing and providing the DNase-seq data, Clovis Galiez and Gonzallo Parra for helpful discussions and feedback on the manuscript. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grants e:AtheroSysMed 01ZX1313A-2014).

References

1. Visscher, P. M., Brown, M. A., McCarthy, M. I., et al. (2012). Five years of GWAS discovery. *Am J Hum Genet 90*, 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

2. MacArthur, J., Bowler, E., Cerezo, M., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res 45*, 896–901. DOI: 10.1093/nar/gkw1133.

3. Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLOS Genet 3*, 1–13. DOI: 10.1371/journal.pgen.0030114.

4. Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann Appl Stat 5*, 1780–1815. DOI: 10.1214/11-AOAS455.

5. Marchini, J., Howie, B., Myers, S., et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet 39*, 906–913. DOI: 10.1038/ng2088.

6. Consortium, W. T. C. C. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature 447*, 661–678. DOI: 10.1038/nature05911.

7. Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet 11*, 499–511. DOI: 10.1038/nrg2796.

8. Cichonska, A., Rousu, J., Marttinen, P., et al. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics 32*, 1981–1989. DOI: 10.1093/bioinformatics/btw052.

9. Zhou, X., Peter, C., and Matthew, S. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genet 9*, 1–14. DOI: 10.1371/journal.pgen.1003264.

10. Thurman, R. E., Rynes, E., Humbert, R., et al. (2012). The accessible chromatin landscape of the human genome. *Nature 489*, 75–82. DOI: 10.1038/nature11232.

11. Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods 12*, 931–934. DOI: 10.1038/nmeth.3547.

12. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. (Secaucus, NJ, USA: Springer-Verlag New York, Inc.).

13. Samani, N. J., Erdmann, J., Hall, A. S., et al. (2007). Genomewide association analysis of coronary artery disease. *New Engl J Med 357*, 443–453. DOI: 10.1056/NEJMoa072366.

14. Erdmann, J., Groszhennig, A., Braund, P. S., et al. (2009). New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet 41*, 280–282. DOI: 10.1038/ng.307.

15. Schunkert, H., Konig, I. R., Kathiresan, S., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet 43*, 333–338. DOI: 10.1038/ng.784.

16. Erdmann, J., Willenborg, C., Nahrstaedt, J., et al. (2011). Genome-wide association study identifies a new locus for coronary artery disease on chromosome 10p11.23. *Eur Heart J 32*, 158–168. DOI: 10.1093/eurheartj/ehq405.

17. Deloukas, P., Kanoni, S., Willenborg, C., et al. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet 45*, 25–33. DOI: 10.1038/ng.2480.

18. CARDIoGRAMplusC4D. (2015). A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet 47*, 1121–1130. DOI: 10.1038/ng.3396.

19. Sheffield, N. C., Thurman, R. E., Song, L., et al. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res 23*, 777–788. DOI: 10.1101/gr.152140.112.

20. Kichaev, G., Wen-Yun, Y., Sara, L., et al. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLOS Genet 10*, 1–16. DOI: 10.1371/journal.pgen.1004722.

21. Yang, J., Lee, S. H., Goddard, M. E., et al. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet 88*, 76–82. DOI: 10.1016/j.ajhg.2010.11.011.

22. Lee, S. H., Wray, N. R., Goddard, M. E., et al. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet 88*, 294–305. DOI: 10.1016/j.ajhg.2011.02.002.

23. Liu, J. Z., Tozzi, F., Waterworth, D. M., et al. (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet 42*, 436–440. DOI: 10.1038/ng.572.

24. Benner, C. (2017). LDstore, http://www.christianbenner.com. [Online; accessed 9-August-2017].

25. Trynka, G., Sandor, C., Han, B., et al. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet 45*, 124–130. DOI: 10.1038/ng.2504.

26. Pickrell, J. K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet 94*, 559–573. DOI: 10.1016/j.ajhg.2014.03.004.

27. Thompson, J. R., Gögele, M., Weichenberger, C. X., et al. (2013). SNP prioritization using a bayesian probability of association. *Genet Epidemiol 37*, 214–221. DOI: 10.1002/gepi.21704.

28. Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Hum Mol Genet 24*, 111–119. DOI: 10.1093/hmg/ddv260.

29. Kichaev, G., Roytman, M., Johnson, R., et al. (2017). Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics 33*, 248–255. DOI: 10.1093/bioinformatics/btw615.

30. Hormozdiari, F., Kostem, E., Kang, E. Y., et al. (2014). Identifying causal variants at loci with multiple signals of association. *Genetics 198*, 497–508. DOI: 10.1534/genetics.114.167908.

31. Chen, W., Larrabee, B. R., Ovsyannikova, I. G., et al. (2015). Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics 200*, 719–736. DOI: 10.1534/genetics.115.176107.

32. Benner, C., Spencer, C. C. A., Havulinna, A. S., et al. (2016). FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics 32*, 1493–1501. DOI: 10.1093/bioinformatics/btw018.

33. Teslovich, T. M., Musunuru, K., Smith, A. V., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*, 707–713. DOI: 10.1038/nature09270.

34. Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature 518*, 187–196. DOI: 10.1038/nature14132.

35. Dastani, Z., Marie-France, H., Nicholas, T., et al. (2012). Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: A multi-ethnic meta-analysis of 45,891 individuals. *PLOS Genet 8*, 1–23. DOI: 10.1371/journal.pgen.1002607.

36. Zhu, X. and Stephens, M. (2016). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *bioRxiv*. DOI: 10.1101/042457.

37. Manolio, T. A., Collins, F. S., Cox, N. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature 461*, 747–753. DOI: 10.1038/nature08494.

38. Yang, J., Bakshi, A., Zhu, Z., et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet 47*, 1114–1120. DOI: 10.1038/ng.3390.

39. Eraslan, G., Arloth, J., Martins, J., et al. (2016). DeepWAS: Directly integrating regulatory information into GWAS using deep learning supports master regulator MEF2C as risk factor for major depressive disorder. *bioRxiv*. DOI: 10.1101/069096.