RiVIERA-MT: A Bayesian model to infer risk variants in related traits using summary statistics and functional genomic annotations

Yue Li[1,2,*] and Manolis Kellis[1,2,*]

[1]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, 32 Vassar St, Cambridge, Massachusetts 02139, USA
[2]The Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, Massachusetts 02142, USA
[*]Correspondence to liyue@mit.edu or manoli@mit.edu

Fine-mapping causal variants is challenging due to linkage disequilibrium and the lack of interpretation of noncoding mutations. Existing fine-mapping methods do not scale well on inferring multiple causal variants per locus and causal variants across multiple related diseases. Moreover, many complex traits are not only genetically related but also potentially share causal mechanisms. We develop a novel integrative Bayesian fine-mapping model named RiVIERA-MT. The key features of RiVIERA-MT include 1) ability to model epigenomic covariance of multiple related traits; 2) efficient posterior inference of causal configuration; 3) efficient full Bayesian inference of enrichment parameters, allowing incorporation of large number of functional annotations; 4) simultaneously modeling the underlying heritability parameters. We conducted a comprehensive simulation studies using 1000 Genome and ENCODE/Roadmap epigenomic data to demonstrate that RiVIERA-MT compares quite favorably with existing methods. In particular, the efficient inference of multiple causal variants per locus led to significantly improved estimation of causal posterior and functional enrichments compared to the state-of-the-art fine-mapping methods. Furthermore, joint modeling multiple traits confers further improvement over the single-trait mode of the same model, which is attributable to the more robust estimation of the enrichment parameters especially when the annotation measurements (i.e., ChIP-seq) themselves are noisy. We applied RiVIERA-MT to separately and jointly model 7 well-powered GWAS traits including body mass index, coronary artery disease, four lipid traits, and type 2 diabetes. To leverage potential tissue-specific epigenomic co-enrichments among these traits, we harness 52 baseline functional annotations and 220 tissue-specific epigenomic annotations from well-characterized cell types compiled from ENCODE/Roadmap consortium. Overall, we observed an improved enrichments for GTEx whole blood and

tissue-specific eQTL SNPs based on the prioritized SNPs by RiVIERA-MT compared to existing methods.

1 Introduction

Genome wide association studies (GWAS) can help gain numerous insights on the genetic basis of complex diseases, and ultimately contribute to personalized risk prediction and precision medicine [1–4]. However, fine-mapping the exact causal variants is challenging due to linkage disequilibrium (LD) and the lack of ability to interpret the function of noncoding variants, which contribute to about 90% of the current GWAS catalog (40.7% intergenic and 48.6% intronic; [5]). On the other hand, several lines of evidence have been proposed to help interpret non-coding genetic signals, in order to gain insights into potential regulatory functions. In particular, epigenomic annotations can pinpoint locations of biochemical activity indicative of cis-regulatory functions [6, 7]. Indeed, comparison with genome-wide annotations of putative regulatory elements has shown enrichment of GWAS variants in enhancer-associated histone modifications, regions of open chromatin, and conserved non-coding elements [3, 6, 8–12], indicating they may play gene-regulatory roles. These enrichments have been used to predict relevant cell types and non-coding annotations for specific traits [6, 9, 13].

Recently, several methods proposed to model the summary statistics of GWAS and thus circumvent the difficulties of accessing individual-level genotype data [14–18]. Some of these methods also utilize the wealth of genome-wide annotations primarily provided by ENCODE consortium to predict causal variants. In particular, Pickrell (2014) developed a statistical approach called fgwas that models association statistics of a given trait and used regularized logistic function to simultaneously learn the relevant annotations. To account for LD, fgwas assumes at most one causal variants per locus by normalizing the posterior probabilities of SNPs within the same locis. Kichaev et al. (2014) recently developed a multivariate Gaussian framework called PAINTOR, which allows for more than one causal SNP but at most three to be located within a single locus by considering all of the combinatorial settings [15]. Chung et al. (2014) developed model called GPA to prioritize individual pleiotropic risk variants among multiple related traits by essentially numerating for each SNP all possible configurations across traits with an option of using one or more sets of annotations to improve the power detecting causal variants [16]. GPA does not consider LD and assumes that SNPs are independent. Recently, we also developed a model called RiVIERA-beta , which uses functional annotations to infer causal variants by modeling the GWAS p-values via Beta density. Although RiVIERA-beta works on well on inferring regulatory variants on immune traits using ImmunoChip summary statistics data, it is limited to the assumption of one causal variant per locus (Li and Kellis, bioRxiv 2016).

Moreover, many complex traits are genetically related [19,20] and potentially share causal mechanisms such as lipid traits and coronary artery disease [21], autoimmune diseases [22,23] and psychiatric disorders [24, 25]. Most of these related traits have distinct genome-wide significant loci but it is plausible that they share the causal effects at the pathway level. Thus, we hypothesize that exploiting the correlation between traits at the epigenomic annotation level may prove useful in fine-mapping for shared causal mechanisms that go beyond the level

of individual variants. Currently, there is a lack of fine-mapping method that harnesses the intrinsic comorbidity that manifest as tissue-specific epigenomic correlations among related traits.

In this article, we describe a novel Bayesian framework called RiVIERA-MT (Risk Variant Inference using Epigenomic Reference Annotations to predict Multiple Trait-causing co-localized mutations) to fine-map causal variants across multiple related traits by modeling the distribution of GWAS summary statistics in multivariate normal distribution with the aid of LD information from 1000 Genome reference panel. Compared to existing methods, the main novelty of RiVIERA-MT is the ability to perform efficient full Bayesian inference of multiple causal variants per locus across multiple traits while simultaneously inferring and leveraging the functional co-enrichment signals among traits using related baseline and tissue-specific epigenomic annotations. We achieve this via an efficient Markov Chain Monte Carlo (MCMC) approach by jointly sampling from the posterior distribution causal configurations for each locus and functional effects of each annotation that are shared among loci for the same trait and potentially correlate between traits. To evaluate our proposed model rigorously, we conduct a comprehensive simulation studies using 1000 Genome data and ENCODE/Roadmap epigenomic data.

We then apply RiVIERA-MT to jointly fine-map causal variants of 7 related well-powered GWAS traits, including body mass index (BMI) [26], coronary artery disease (CAD) [27], low density lipoprotein (LDL), high density lipoprotein (HDL), triglycerides (TG), total cholesterol (TC) [28], and type 2 diabetes (T2D) [29]. To leverage potential tissue-specific epigenomic co-enrichments among these traits, we harness the largest compendium of epigenomic annotations to date from ENCODE/Roadmap consortium, including 4 previously implicated epigenomic marks (H3K4me1, H3K4me3, H3K27ac, H3K9ac) across 100 well characterized cell types and tissues [7]. This allows us to revisit the GWAS of these 7 complex human traits by inferring their underlying regulatory variants implicated at the tissue-specific epigenomic contexts.

2 Results

2.1 RiVIERA-MT method overview

We describe a novel full Bayesian model to infer causal variants. **Fig.** 1 illustrates the fine-mapping problems in three representative scenarios using simulated data (**Methods**). In the first scenario, the risk locus harbors one causal variant (red cricle), which drives the genetic signals of other non-causal variants via linkage disequilibrium (LD) (**Fig.** 1a). Notably, the lead SNP (dark diamond) with the most significant p-value is not the causal variant. In this case scenario, the underlying epigenomic activities (middle track) provide a crucial evidence to the inference of functional variants. Methods that assume single causal variant per locus may work well here by normalizing the posterior for each SNP within the locus [14, 30]. However, these methods become inadequate when there are more than one causal variant within the same locus (**Fig.** 1b,c) because they will pull down the true signals of all causal variants in order to maintain a properly normalized posterior probabilities.

Our RiVIERA-MT builds upon some of the existing fine-mapping methods [17,18,31,32]

by utilizing multivariate normal theory to infer the posterior distribution of *causal configurations* and subsequently marginalizes the posterior to infer posterior inclusion probabilities (PIP) for each SNP among all sampled configurations [15, 17, 18, 33, 34] (**Methods**). However, as illustrated in **Fig.** 2 and detailed in **Methods**, RiVIERA-MT has several significant novel features that distinguish it from the existing fine-mapping methods:

1. Ability to model epigenomic covariance Σ_w of multiple related traits, which do not necessarily share the same set of risk loci (**Fig.** 2a and c);

2. Efficient posterior inference of causal configurations $\mathbf{c}_{ld}$ for each locus l and disease d, automatically determining the number of causal variants in each risk locus (**Fig.** 2b);

3. Efficient full Bayesian inference of functional parameters of epigenomic weights ($\mathbf{w}_k$) allowing incorporation of a large number of discrete or continuous annotations $\mathbf{a}_{ldk}$ with lesser concern of overfitting due to the full Bayesian treatments (**Fig.** 2c);

4. Simultaneously modeling the underlying heritability parameters as per-SNP variance explained $\sigma^2_{a,d}$ and leveraging it in the causal inference (**Fig.** 2c);

It is worth noting that the multi-trait feature of RiVIERA-MT allows us to fine-map causal variants simultaneously across a large number of traits because the model complexity grows only linear to the number of traits. Because we only associate traits via the epigenomic covariance, we impose only a weak prior on the underlying relatedness of traits. This contrasts to the direct inference approach of detecting the individual pleiotropic variants that affect zero, one or multiple related traits [16], which is exponential to the number of traits modeled for each SNP in order to consider all of the configurations of the same SNP across traits. Our model also differs from directly inferring causal variants within pleiotropic loci, which requires both the underlying causal variants and the genome-wide significant loci to be same across traits (Kichaev *et al.*, bioRxiv 2016).

2.2 Empirical analysis of model convergence

Convergence is crucial for an MCMC approach to accurately approximate the posterior distributions of causal variants. Although we do not have a theoretical guarantee for the convergence of our model, we examined the joint posteriors at each MCMC iteration across 1000 samplings using 100 simulated datasets (**Methods**). Indeed, we observed that our model converged very fast after a few iterations attributable to the sensible MCMC sampling methods (**Supplementary Fig.** S2a). At each MCMC iteration, we performed a fixed number of stochastic searches of *the entire local neighborhood* of the current causal configurations [18,35]. Thus, the number of searches needed to reach the optimal power is independent from the size of the locus. We assessed the model performance as a function of an increasing number of neighborhood searches. Indeed, we observed a stead improvement as we increased the number of neighborhood searches, and the model reached to the optimal detection power at 10 rounds of stochastic searches (**Supplementary Fig.** S2b). Accordingly, we fixed the number of searches to 10 throughout this study.

2.3 Improved fine-mapping power over existing methods

To assess the power of the proposed fine-mapping model in identifying causal variants and compare it with existing methods, we implemented a simulation pipeline adapted from [15] (**Methods**). RiVIERA-MT demonstrates a consistently improved power in detecting causal variants (**Fig.** 3). For instance, the top 100 selected variants by RiVIERA-MT contain over 50% of the causal variants over most simulation tests in contrast to lower than 40% of the causal variants detected among the top variants chosen by PAINTOR [15], which is the current state-of-art fine-mapping method that integrates annotations with summary statistics. Furthermore, RiVIERA-MT effectively incorporates functional annotations as it performs much better than the same model without annotations. Notably, the methods that explicitly account for LD namely our RiVIERA-MT model and PAINTOR conferred much higher power compared to fgwas [14] and RiVIERA-beta (Li and Kellis, bioRxiv 2016) which assume single causal variant per locus. Nonetheless, fgwas and RiVIERA-betaperform better than GPA and marginal GWAS -log10 p-value, which assume that all of the SNPs are independent.

From our simulation, a single locus can harbor more than one causal variant, some of which may exhibit rather weak genetic signals due to relatively low allele frequency. For instance, **Fig.** 1b and c illustrate two loci containing 3 and as many as 10 causal variants in the same loci, respectively. In these cases, our RiVIERA-MT model is still able to efficiently infer the correct PIP by marginalizing over a large number of sampled causal configurations with high local posteriors, which automatically accounts for the potentially large number of causal variants within the same locus without predefining the number of causal variants per locus. As a result, the posterior probabilities produced by RiVIERA-MT are well calibrated and exhibit consistently superior performance by identifying additional number of causal variants when selecting more than 10 variants per locus (**Supplementary Fig.** S3). On the other hand, when the number of causal variants to consider go beyond a model's ability to infer, the causal signals are poorly calibrated. This is the main reason RiVIERA-beta and fgwas (which assumes 1 causal variant per locus) and PAINTOR (which is only able to infer by default at maximum 2 causal variants per locus due to exhaustive search) perform worse in prioritizing variants beyond top 10 SNPs per locus compared to methods such as GPA and GWAS -log10 p-values that do not impose any normalization constraints on the SNPs.

2.4 Functional enrichments

In addition to the improved power in causal variant detection when using annotations, we sought to further ascertain the ability of RiVIERA-MT to incorporate relevant annotations. To this end, we performed Bayesian log-likelihood ratio tests (LRT) on each annotation by comparing the likelihoods of the null models without the annotation in question with the likelihoods of the alternative models with the annotation incorporated. The Bayesian credible interval of the LRT statistics was obtained naturally from the MCMC samplings (**Methods**). Indeed, we observed a consistent agreement between the predicted LRT and the underlying fold enrichment of each annotation from the simulated data with the median Pearson correlation above 0.78 (**Fig.** 4). PAINTOR and RiVIERA-MT have rather comparable performance and performed much better than fgwas and GPA. Notably, the accuracy

of enrichment tests reflects the accuracy of fine-mapping as all of the four methods jointly infer both the causal variants and the enrichment parameters. Thus, models that generalize to inferring multiple causal variants per locus confer more robust estimate of the enrichment compared to models that do not.

2.5 Inferring variance explained by causal SNPs

As a part of the full Bayesian fine-mapping algorithm, RiVIERA-MT is able to infer the distribution of per-SNP variance via the MCMC sampling scheme, which is related to the narrow-sense heritability [31, 36] (**Methods**). We assessed the variance estimate by simulating GWAS datasets with heritability values ranging from 0.05 to 0.95. Overall, we observed a consistent increase of the estimates as we increased the underlying heritabilities (**Supplementary Fig.** S4). This is remarkable compared to the existing fine-mapping methods such as CAVIARBF [17] and FINEMAP [18], which treat the per-SNP variance as a free parameter defined by the users. For a normalized heritability estimation (ranging between 0 and 1) (which is not the focus of our model), we need to know the standard error from the linear model and *genome-wide* effect sizes of all of the SNPs (Zhu and Stephens, bioRxiv 2016) as apposed to the SNPs in the GWAS loci, which have been integrated out in our model.

2.6 Joint inference of multiple traits

An additional novel feature of RiVIERA-MT is the ability to jointly model the summary statistics of multiple traits, which do not share the same risk loci. To examine whether the multi-trait mode of the model provides any improved performance in detecting causal variants, we simulated 100 datasets for 2 to 10 related traits. Despite distinct risk loci, we correlated the traits based on the functional co-enrichments of the cognate causal variants over the 100 epigenomic annotations (**Methods**). Thus, model that is able to harness this underlying correlation by jointly inferring the causal variants and causal annotations across the related traits should confer better performance than modeling each trait separately. Indeed, compared to the single-trait RiVIERA-MT model, we observed a modest but significant gain of power reflected by the reduced number of SNPs required to identify 90% of the causal variants (**Fig.** 5). The improvements of multi-trait mode over single-trait are consistent over different number of traits.

As expected, the improvement is completely attributable to the improved empirical prior inference (**Fig.** 5 top panels) because it is the only model component that is connected to the estimated covariance of the annotation weights (**Fig.** 2). When the genetic signals are incorporated into the posterior model component, we observed less pronounced but still notable improved resolution (**Fig.** 5 top panels). Additionally, we assessed the robustness of the models in their abilities to infer causal variants when the annotations themselves are noisy estimates of the true underlying annotations. This is a realistic scenario since the epigenomic annotations were based on ChIP-seq experiments, which are often noisy due to imperfect efficacy of antibodies, sequencing errors, read alignment error, peak calling algorithmic errors, etc. To this end, we used standardized continuous annotations instead of binary annotations to fit each model and assessed the number of SNPs required to identify 90% causal variants.

As expected, the general performance decreased when using the noisy annotations (**Fig.** 2 right panels). However, the multi-trait model exhibits better robustness compared to the single-trait model especially when modeling more than 2 traits simultaneously.

2.7 Application to body mass index, lipid traits, type 1 diabetes, and coronary artery disease

We applied RiVIERA-MT to 7 related traits including body mass index (BMI), 4 lipid traits (HDL, LDL, TC, TG), type 1 diabetes (T1D), and coronary artery disease (CAD) using the corresponding publicly available summary statistics imputed to 1000 Genome European SNPs (**Methods**; **Supplementary Table** S1). We first examined the *genome-wide* functional enrichments over 272 well defined annotations including 52 baseline annotations and 220 cell-type-specific annotations over four transcription-activating histone marks via LD score regression on the z-scores of the HapMap3 SNPs from each GWAS trait against European 1000 Genome Phase 1 (version 3) LD reference panel (**Methods**) [37]. Consistent with the published results [37], we observed meaningful cell-type-specific functional enrichments for the GWAS traits conditioned on the 52 baseline functional categories (**Fig.** 2a): BMI is significantly enriched for CNS functional categories; the 4 lipid traits are significantly enriched for liver for one or more histone marks; CAD for heart tissues; T2D for pancreas. Thus, the causal signals are highly implicated in the functional annotations, suggesting an integrative fine-mapping method such as RiVIERA-MT to incorporate them to improve the power of fine-mapping causal variants that potentially disrupt various functional elements and especially the tissue-specific regulatory elements of the genome. However, we also observed a pervasive sharing of functional enrichments between traits, which suggest that jointly modeling these related traits may further improve fine-mapping power over modeling each trait separately.

To fine-map causal variants in independent risk loci for each trait (**Methods**), we applied RiVIERA-MT in multi-trait and single-trait mode as well as the PAINTOR model [15] on each GWAS data using the subset of the baseline and cell-specific annotations with enrichment p-values < 0.05 after Benajmini-Hocherg adjustment for multiple testings over the 7 traits and 272 annotations. We first visualized the fine-mapping results of RiVIERA-MT on single-trait mode with four tracks for each trait including (top-bottom) GWAS p-values, baseline annotations, cell-specific annotations, and the posterior inclusion probabilities (PIP) inferred by RiVIERA-MT (**Supplementary Fig.** S5). Consistent with previously reported results [36–38], the baseline model (second track) suggest that the risk loci are highly enriched for enhancer-related regions/marks such as DNA hypersensitive site (DHS), H3K27ac, and H3K4me1 for all of the 7 traits. More remarkably, we observed a striking distinct tissue-specific epigenomic landscapes (the third track) between different GWAS traits. In particular, the BMI loci are highly enriched for CNS related histone marks whereas CAD exhibit modest enrichment for cardiovascular marks, HDL for liver, and T2D for Adrenal/Pancreas.

The inferred PIP prioritizes SNPs by taking into account 3 sources of information: 1) GWAS signals in terms of z-scores; 2) significant annotations determined by LDSC; 3) linkage disequilibrium from 1000 Genome European reference panel. While many variants with PIP < 0.9 also exhibit significant GWAS signals (p<5e-8), a substantial number of SNPs that are

below the GWAS threshold became prominent after the re-prioritization (**Supplementary Fig.** S5). We then applied RiVIERA-MT to jointly model the data of the 7 GWAS traits. The resulting PIP from the multi-trait mode generally correlate well with the PIP from the single-trait mode (**Supplementary Fig.** S6). Because there is no gold-standard for the causal SNPs of each trait, we compared the inference results from RiVIERA-MT and PAINTOR in terms of the overlap of the 90% credible set from each method, which are determined as the SNPs with PIP that contributed to 90% of the total posterior mass (**Supplementary Fig.** 6). In general, there are substantial overlap between RiVIERA-MT's and PAINTOR's 90% credible sets, implying an overall consistent agreements among these methods. Importantly, the PIP for each method increases as a function of the number of supporting methods (**Fig.** 6).

Moreover, as an empirical evaluation for the functional implication of the prioritized varaints, we ranked the SNPs by the corresponding PIPs inferred by each method and computed the hypergeometric enrichments for the GTEx (version 6) whole blood eQTL SNPs (WB) or tissue-specific eQTL SNPs as a function of the increasing number of top variants selected. For the tissue-specific eQTL SNPs, we chose brain and nerve tissues for BMI, artery and heart tissues for CAD, liver and adipose for the four lipid traits, and pancreas for T2D. Although the results are not monotonically favorable for a single method, SNPs prioritized by RiVIERA-MT single or multi-trait models exhibit higher overall enrichments for the eQTL SNPs compared to PAINTOR and GWAS -logP methods in most traits (**Fig.** 7a).

Since both RiVIERA-MT and PAINTOR provides 90% credible sets, we further examined their functional enrichment for the eQTL SNPs from entire GTEx data over 44 tissues (**Fig.** 7b). Interestingly, the credible SNPs for BMI, HDL, TC and TG exhibit enrichment over majority of the tissues, perhaps implying a multifaceted causal mechanisms for these traits. On the other hand, the credible SNPs for CAD and LDL are highly selective of tissue types with CAD significantly enriched for artery tissues and LDL for liver tissue. T2D exhibits no obviously meaningful enrichment. The enrichment signals are generally consistent among the methods. Nonetheless, RiVIERA-MT achieved more significant eQTL enrichment than PAINTOR in all traits except T2D. However, caution must be taken to interpret these results because the enrichment analysis may be biased for the larger number of SNPs used to construct the 90% credible set and the eQTL SNPs themselves are not independent but rather linked by linkage disequilibrium, which violates the hypergeometric enrichment model assumption.

3 Discussion

Dissecting causal mechanisms of complex traits to ultimately map genotypes to phenotypes becomes plausible with the recent availability of large-scale functional genomic data [9, 23, 30, 39]. In formulating an efficient fine-mapping strategies, it is natural to incorporate the valuable reference annotations in a principled way as a form of Bayesian prior to infer the functional variants that drive the genetic signals of GWAS [9, 15, 40–42]. In this article, we describe a novel Bayesian fine-mapping method RiVIERA-MT to re-prioritize GWAS summary statistics based on their epigenomic contexts and LD information. The main contribution of RiVIERA-MT is the ability to efficiently infer multiple causal variants within

a set of susceptible loci in a single trait or across multiple traits that do not need to share the same risk loci. Through comprehensive simulations and applications to GWAS datasets, we demonstrate the general utilities of RiVIERA-MT. Because our model only require summary statistics, we envision its broad applications in large-scale GWAS meta-analysis on many complex traits.

One caveat in our current model formalism is that the likelihood is based on *the given risk loci* rather than *genome-wide* SNPs. Here we made two implicit assumptions: (1) all of the genetic signals associated with the trait are captured within the risk loci; (2) majority of the SNPs within the risk loci are not causal and serve as background for fine-mapping the causal variants. This is true in our simulation, which was mainly used to assess how sensitive our model is to distinguish causal SNPs and causal annotations with different fold-enrichment of causal variants. In practice, this assumption may not hold especially for highly polygenetic model with small effect sizes. To detect causal annotations, a general enrichment test should be performed either on *genome-wide* independent loci such as fgwas [14] or on genome-wide SNPs such as the recently developed LD-score regression approach, which assesses the proportion of variance explained (PVE) due to the LD-linked SNPs from each functional category over the total estimated heritability [37].

As demonstrated in our applications to GWAS data, users may perform enrichment tests with a software of their choice and input to RiVIERA-MT a select set of annotations for fine-mapping. It is also worth mentioning that annotations that exhibit genome-wide enrichment may not be useful for fine-mapping purpose. Suppose we have an annotation that covers all of the risk loci. The corresponding enrichment for this annotation will be highly significant from genome-wide analysis but no different from background when focused within risk loci. Thus, annotations as such are important for inferring the risk loci but not for inferring individual risk variants. Therefore, it is still necessary to weight each annotations via the fine-mapping model.

As future works, we can extend our current RiVIERA-MT in several ways. First, our current eQTL enrichment analyses may be biased by the LD on the eQTL side and thus may merit more meaningful signals if we can infer the causal genes beyond individual variants by jointly modeling both the GWAS data and eQTL data. Second, we can generalize the model to apply for genome wide SNPs rather than defined risk loci via hierarchical modeling approach to infer risk loci and then the causal variants [14, 43]. Third, for traits that are associated with the same pleiotropic loci, we may provide the users an option to infer the joint posterior of SNP associations with multiple traits. Fourth, the current model can also be easily adapted to model trans-ethnic GWAS using separate LD matrices as effectively demonstrated by the trans-ethnic version of the PAINTOR model [33]. Fifth, instead of using the linear logistic prior model, we will explore other models that take into the spatial information of the genomic sequence and local epigenomic context around each SNP. Finally, the efficiency of our model can be further improved by paralleling the causal configuration searches via a multi-processing computing architecture.

4 Methods

4.1 Model details

4.1.1 Likelihood and Bayes factor

We assume a linear model for quantitative trait of n individuals and p SNPs [17, 44]:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$
$$\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_n) \tag{2}$$
$$\beta \sim \mathcal{N}(0, \sigma_a^2 \sigma_e^2 I_c) \tag{3}$$

where σ_e^2 is the standard error, β is the effect size with equal per-SNP additive variance σ_a^2 for each causal SNP, I_n is identity matrix with ones in the diagonal and zeros elsewhere, and I_c a diagonal matrix such that $I_{i,i} = 1$ if SNP i is causal ($c_i = 1$) otherwise 0 ($c_i = 0$). Notably, we assume the causal indicator $\mathbf{c}$ is given here in order to integrate out the effect size β and then subsequently infer the posterior distribution $\mathbf{c}$ as detailed below.

As previously shown by [17,18], we can integrate out β by taking the conditional expectation of the mean and variance of $\mathbf{y}$ with respect to β (E_β) and leveraging the (log-transformed) linear property of the multivariate Gaussian density function:

$$\mathcal{N}(\mathbf{y}|\mathbf{X}, \sigma_a^2, \sigma_e^2, \mathbf{c}) = \int \mathcal{N}(\mathbf{y}|\mathbf{X}\beta, \sigma_e^2)\mathcal{N}(\beta|0, \sigma_a^2\sigma_e^2 I_c)d\beta \tag{4}$$
$$= \mathcal{N}(y|\mathrm{E}(\mathbf{y}), \mathrm{E}(\mathrm{Var}(\mathbf{y}))) \tag{5}$$
$$= \mathcal{N}(y|\mathrm{E}_\beta(\mathbf{y}|\mathbf{X}\beta), \mathrm{E}_\beta(\mathrm{Var}(\mathbf{y}|\sigma_e^2, \mathbf{X}, \beta)) + \mathrm{Var}(\mathrm{E}_\beta(\mathbf{y}|\sigma_e^2, \mathbf{X}, \beta))) \tag{6}$$
$$= \mathcal{N}(y|0, \sigma_e^2 I_n + \mathbf{X}(\sigma_e^2\sigma_a^2 I_c)\mathbf{X}') \tag{7}$$
$$= \mathcal{N}(y|0, \sigma_e^2(I_n + \mathbf{X}(\sigma_a^2 I_c)\mathbf{X}')) \tag{8}$$

We can then express the likelihood density function of Eq (1) in terms of z-score:

$$\mathbf{y}|\sigma_a^2, \sigma_e^2, \mathbf{c} \sim \mathcal{N}(0, \sigma_e^2(I_n + \mathbf{X}(\sigma_a^2 I_c)\mathbf{X}')) \tag{9}$$
$$\frac{\mathbf{X}'\mathbf{y}}{\sqrt{n}}|\sigma_a^2, \sigma_e^2, \mathbf{c} \sim \mathcal{N}(0, \sigma_e^2(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\mathbf{X}'\mathbf{X}(\sigma_a^2 I_c)\mathbf{X}'\mathbf{X}}{n}) \tag{10}$$
$$\mathbf{z} \equiv \frac{\mathbf{X}'\mathbf{y}}{\sqrt{n\sigma_e^2}}|\sigma_a^2, \mathbf{c} \sim \mathcal{N}(0, \Sigma + \Sigma(n\sigma_a^2 I_c)\Sigma) \tag{11}$$

where $\Sigma = \mathbf{X}'\mathbf{X}/n$ is often referred to as the linkage equilibrium (LD) and estimated either from the corresponding study cohort or a reference population from 1000 Genome Consortium [38].

Thus, given the sample size, z-scores, Σ as the GWAS summary statistics, we can infer $\mathbf{c}$ without the access to the individual-level genotype and phenotype information. Moreover, we do not need to know the effect size β as it has been integrated out or σ_e^2 as it has been cancelled by the z-score calculation in (11).

The Bayes factor is the likelihood ratio of the alternative model over the null model:

$$
\begin{aligned}
&BF(\mathbf{z}|\mathbf{c},\Sigma,\sigma_a^2) \\
&= \frac{\mathcal{N}(\mathbf{z}|0,\Sigma+\Sigma(n\sigma_a^2 I_c)\Sigma)}{\mathcal{N}(\mathbf{z}|0,\Sigma)} && (12)\\
&= \frac{\mathcal{N}(\mathbf{z}_c|0,\Sigma_{cc}+\Sigma_{cc}(n\sigma_a^2 I_c)\Sigma_{cc})\mathcal{N}(\mathbf{z}_n|\Sigma_{nc}\Sigma_{cc}^{-1}\mathbf{z}_c,\Sigma_{nn}-\Sigma_{nc}\Sigma_{cc}^{-1}\Sigma_{cn})}{\mathcal{N}(\mathbf{z}|0,\Sigma)} && (13)\\
&= \frac{\mathcal{N}(\mathbf{z}_c|0,\Sigma_{cc}+\Sigma_{cc}(n\sigma_a^2 I_c)\Sigma_{cc})}{\mathcal{N}(\mathbf{z}|0,\Sigma)}\frac{\mathcal{N}(\mathbf{z}|0,\Sigma)}{\mathcal{N}(\mathbf{z}_c|0,\Sigma_{cc})} && (14)\\
&= \frac{\mathcal{N}(\mathbf{z}_c|0,\Sigma_{cc}+\Sigma_{cc}(n\sigma_a^2 I_c)\Sigma_{cc})}{\mathcal{N}(\mathbf{z}_c|0,\Sigma_{cc})} && (15)\\
&= BF(\mathbf{z}_c|\Sigma_{cc},\sigma_a^2) && (16)
\end{aligned}
$$

where $\mathbf{z}_c$ and Σ_{cc} denote z-scores and LD for the causal SNPs, respectively. Notably, Eq (15) is much more efficient than Eq (12) because it operates only on the causal SNPs instead of all of the SNPs.

Suppose there are D diseases, L_d independent risk loci for disease d, m_{ld} SNPs in locus l. The joint likelihood expressed in terms of Bayes factor is factorized into products of individual likelihoods over loci across traits:

$$\mathcal{L}(\mathbf{c}|\mathbf{z},\Sigma) = \prod_{d=1}^{D}\prod_{l=1}^{L_d} BF(\mathbf{z}_{ld}|\mathbf{c}_{ld},\Sigma_{ld}) \tag{17}$$

4.1.2 Prior

The prior distirbution of being a causal SNP in locus l and disease d follows Bernoulli distribution:

$$p(\mathbf{c}_{ld}|\mathbf{a},\mathbf{w}) = \prod_{i=1}^{m_l} \pi_{ild}^{c_{ild}}(1-\pi_{ild})^{(1-c_{ild})} \tag{18}$$

where π_{ild} is a logistic function of a linear combination of K annotations weighted by model parameters $\mathbf{w}$:

$$\pi_{ild} = [1+\exp(-\sum_{k=1}^{K} w_{kd}a_{ilk} - b_{ld})]^{-1} \tag{19}$$

Here $\mathbf{w} = \{w_{kd}\}_{K\times D}$ follows multivariate normal distribution with $D \times D$ covariance Σ_w modeling the underlying *disease-disease covariance* at the annotation level, and the inverse of the covariance $\Sigma_w^{-1} = \Lambda_w$ follows Wishart distribution:

$$\mathbf{w}|\Lambda_w \sim \mathcal{N}(0,\Lambda_w^{-1}) \tag{20}$$

$$\Lambda_w|\Lambda_0,\nu_0 \sim \mathcal{W}(\Lambda_0,\nu_0) \tag{21}$$

The linear bias b_{ld} in (19) follows a univariate normal with mean equal to the logit function of causal proportion $\pi_{0,ld}$ within locus l of disease d and the inverse variance follows Gamma

distribution:

$$b_{ld}|\pi_{0,ld}, \lambda_{0d}, \sim \mathcal{N}(g(\pi_{0,ld}), \lambda_{0d}^{-1}) \tag{22}$$

$$\lambda_{0d}|\alpha_0, \beta_0 \sim \Gamma(\alpha_0, \beta_0) \tag{23}$$

where $g(\pi_0) = \log(\pi_0)/\log(1-\pi_0)$ and $\pi_0 = 1/m_{ld}$, which implies *apriori* one causal variant per locus, and we set $\alpha = 0.01$ and $\beta = 0.0001$ to enable a broad hyperprior for λ_{0d}.

4.1.3 Approximate posterior inference of causal configurations

Based on the results above, posterior inference of a causal configuration for locus l in disease d can operate on Bayes factors as follows:

$$\begin{aligned}
&p(\mathbf{c}_{ld}|\mathbf{z}_{ld}, \Sigma_{ld}, \mathbf{a}_{ld}, \mathbf{w}_d, \sigma_a^2) \\
&= \frac{p(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld})p(\mathbf{c}_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}{\sum_{\mathbf{c}'_{ld}\in\mathcal{S}_{ld}} p(\mathbf{z}_{ld}|\mathbf{c}'_{ld}, \Sigma_{ld})p(\mathbf{c}'_{ld}|\mathbf{a}, \mathbf{w})} && (24)\\
&= \frac{\frac{\mathcal{N}(\mathbf{z}_{ld}|0, \Sigma_{ld}+\Sigma_{ld}(n_d\sigma_{a,d}^2 I_{c,d})\Sigma_{ld})}{\mathcal{N}(\mathbf{z}_{ld}|0,\Sigma_{ld})}p(\mathbf{c}_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}{\sum_{\mathbf{c}'_{ld}\in\mathcal{S}_{ld}} \frac{\mathcal{N}(\mathbf{z}_{ld}|0, \Sigma_{ld}+\Sigma_{ld}(n_d\sigma_{a,d}^2 I_{c',d})\Sigma_{ld})}{\mathcal{N}(\mathbf{z}_{ld}|0,\Sigma_{ld})}p(\mathbf{c}'_{ld}|\mathbf{a}, \mathbf{w})} && (25)\\
&= \frac{\frac{\mathcal{N}(\mathbf{z}_{c,ld}|0, \Sigma_{cc,ld}+\Sigma_{cc,ld}(n_d\sigma_{a,d}^2 I_c)\Sigma_{cc,ld})}{\mathcal{N}(\mathbf{z}_{c,ld}|0,\Sigma_{cc,ld})}p(\mathbf{c}_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}{\sum_{\mathbf{c}'_{ld}\in\mathcal{S}_{ld}} \frac{\mathcal{N}(\mathbf{z}_{c',ld}|0, \Sigma_{c'c',ld}+\Sigma_{c'c',ld}(n\sigma_{a,d}^2 I_{c',ld})\Sigma_{c'c',ld})}{\mathcal{N}(\mathbf{z}_{c',ld}|0,\Sigma_{c'c',ld})}p(\mathbf{c}'_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)} && (26)\\
&= \frac{BF(\mathbf{z}_{c,ld}|\Sigma_{cc,ld}, \sigma_{a,d}^2)p(\mathbf{c}_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}{\sum_{\mathbf{c}'_{ld}\in\mathcal{S}_{ld}} BF(\mathbf{z}_{c',ld}|\Sigma_{c'c',ld}, \sigma_{a,d}^2)p(\mathbf{c}'_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)} && (27)\\
&\equiv \frac{BF(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \sigma_{a,d}^2)p(\mathbf{c}_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}{\sum_{\mathbf{c}'_{ld}\in\mathcal{S}_{ld}} BF(\mathbf{z}_{ld}|\mathbf{c}'_{ld}, \Sigma_{ld}, \sigma_{a,d}^2)p(\mathbf{c}'_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)} && (28)
\end{aligned}$$

where Eq (25) to Eq (26) utilizes the results from Eq (15). Thus, we can infer Eq (24) by Eq (27) using Bayes factor of only the causal SNPs, which is much more efficient than inferring the likelihood of all of the SNPs in the locus. To simplify notation below, we use Eq (28) instead of Eq (27).

However, the normalization term in the denominator of Eq (28) still requires evaluation of $\sum_{j=1}^{m_{ld}} \binom{m_{ld}}{j}$ causal configurations, which becomes intractable for large m_{ld}. We approximate it by recursively sampling from the neighborhoods of the current configuration plausible configurations based on their posterior normalized only within the neighborhood. By doing so, we ignore the majorities of the highly implausible configurations that likely contribute very little to the normalization [18, 35]. This stochastic search technique was initially developed by [35] as general feature selection algorithm, and was first implemented to fine-map causal variants in the software called FINEMAP [18]. However, FINEMAP infers causal variants on a single locus individually (i.e., no information sharing among loci), works for only a single trait, does not take into account functional annotations, and accepts all proposed configurations. In contrast, our model infer causal variants across multiple loci with model parameters shared among loci, can operate on multiple traits simultaneously, harnessing large-scale functional and epigenomic annotations, and exploits an sampling scheme to ensure the quality of the neighborhood that the model is exploring (detailed as follows).

We apply Metropolis-Hastings (MH) algorithm to accept the proposed configuration $\mathbf{c}^*_{ld}$ at the probability:

$$\min(1, \frac{\sum_{\mathbf{c}''_{ld} \in nbd(\mathbf{c}^*_{ld})} BF(\mathbf{z}_{ld}|\mathbf{c}''_{ld}, \Sigma_{ld}, \sigma^2_{a,d}) p(\mathbf{c}''_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}{\sum_{\mathbf{c}'_{ld} \in nbd(\mathbf{c}^{cur}_{ld})} BF(\mathbf{z}_{ld}|\mathbf{c}'_{ld}, \Sigma_{ld}, \sigma^2_{a,d}) p(\mathbf{c}'_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}) \tag{29}$$

Notably, in contrast to standard MH, our proposed MH step compares the *neighborhood* of the proposed causal configuration (including the proposed configuration itself) with the *neighborhood* of the current causal configuration. Compared to standard MH on single configuration, the neighborhood-based MH approach is more effective in accepting configuration space with larger improvement and thus less prone to random walk behavior [35].

Furthermore, we use an unordered hash table to efficiently keep track of all of the evaluated configurations throughout the MH stochastic samplings and to avoid re-computing the already visited configurations, which is the same as in FINEMAP [18]. However, different from FINEMAP, we need to re-initialize the hash table at each complete iteration because the prior distribution changes (which changes the posterior distribution of the causal configurations) after new annotation weights $\mathbf{w}$ are sampled from the posterior (detailed next).

The posterior inclusion probabilities (PIP) for SNP i in locus l of disease d is then:

$$p(c_{ild}|\mathbf{z}_{ld}, \Sigma_{ld}, \mathbf{a}_{ld}, \mathbf{w}_d, \sigma^2_a) = \sum_{\mathbf{c}_{ld} \in \mathcal{S}^*_{ld}, c_{ild}=1} \frac{BF(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \sigma^2_{a,d}) p(\mathbf{c}_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d)}{\sum_{\mathbf{c}'_{ld} \in \mathcal{S}^*_{ld}} BF(\mathbf{z}_{ld}|\mathbf{c}'_{ld}, \Sigma_{ld}, \sigma^2_{a,d}) p(\mathbf{c}'_{ld}|\mathbf{a}, \mathbf{w})} \tag{30}$$

where $\mathcal{S}^*_{ld}$ is the set of visited configurations.

4.1.4 Joint posterior

Given PIP, the logarithmic joint posterior density function is then:

$$\begin{aligned}
\log p(\Theta|\mathcal{D}) &= \log f(\mathbf{w}, \Lambda_w, \mathbf{b}, \lambda, \sigma^2_a|\mathbf{z}, \Sigma, \mathbf{a}, \mathbf{c}) && (31)\\
&\propto \log f(\mathbf{w}|\Lambda_w) + \log f(\Lambda_w|\Lambda_0, \nu_0) && (32)\\
&+ \sum_{d=1}^{D} \sum_{l=1}^{L_d} \log f(b_{ld}|m_{ld}, \lambda_{ld}) + \sum_{d=1}^{D} \log f(\lambda_d|\alpha_0, \beta_0) && (33)\\
&+ \sum_{d=1}^{D} \log f(\sigma^2_{a,d}) && (34)\\
&+ \sum_{d=1}^{D} \sum_{l=1}^{L_d} \log f(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \sigma_{a,d}) + \log f(\mathbf{c}_{ld}|\mathbf{a}_{ld}, \mathbf{w}_d, \mathbf{b}_{ld}) && (35)\\
& && (36)
\end{aligned}$$

In principle, causal inference requires integrating out all of above parameters:

$$p(c_{ild}|\mathbf{z}_d, \Sigma_d, \mathbf{a}_d) = \int p(c_{ild}|\Theta, \mathcal{D}) p(\Theta|\mathcal{D}) d\Theta \tag{37}$$

which is not tractable. We employ Markov Chain Monte Carlo (MCMC) to sample from the joint posterior in Eq (31).

4.1.5 Sampling genetic additive variance $\sigma^2_{a,d}$

In our BF formulation, $\sigma^2_{a,d}$ is a free hyperparameter. Existing fine-mapping methods such as FINEMAP [18] and CAVIARBF [17] set it to a fixed user-defined value. Here, as first proposed by Guan and Stephen (2011) [31], we associate $\sigma^2_{a,d}$ to the underlying heritability estimate h^2_d of disease d:

$$h^2_d = \frac{m_d\sigma^2_{a,d}}{m_d\sigma^2_{a,d} + \sigma^2_{e,d}} \tag{38}$$

$$= \frac{m_d\sigma^2_{a,d}/\sigma^2_{e,d}}{m_d\sigma^2_{a,d}/\sigma^2_{e,d} + 1} \tag{39}$$

$$\sigma^2_{a,d}/\sigma^2_{e,d} = \frac{h^2_d}{m_d(1-h^2_d)} \tag{40}$$

$$= \frac{h^2_d}{1-h^2_d}(\sum_{l=1}^{L_d}\sum_{i=1}^{m_d} c_{ild})^{-1} \tag{41}$$

In practice, we sample h^2_d from uniform: $h^{2*}_d \leftarrow h_d + U(-0.1, 0.1)$ and re-parameterize it to get $\sigma^2_{a,d}/\sigma^2_{e,d}$ (41). We then apply MH to accept the proposed h^{2*}_d (and hence $\sigma^{2*}_{a,d}/\sigma^2_{e,d}$) at the probability:

$$\min(1, \frac{\prod_l \sum_{\mathbf{c}_{ld}} BF(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \sigma^{2*}_{a,d})p(\mathbf{c}_{ld}|\mathbf{a}, \mathbf{w})}{\prod_l \sum_{\mathbf{c}_{ld}} BF(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \sigma^2_{a,d})p(\mathbf{c}_{ld}|\mathbf{a}, \mathbf{w})}) \tag{42}$$

Note that we do not need to estimate $\sigma^2_{e,d}$ because it is the same for all configurations and thus cancelled out in Eq (24). Also, our main goal here is to fine-map causal variants rather than estimating heritability. For the latter, readers may refer to a recently proposed model on estimating effect size (which we have integrated out) and heritability using summary statistics (Zhu and Stephen, bioRxiv 2016).

4.1.6 Sampling model parameters $\Lambda_w, \lambda, \mathbf{w}, \mathbf{b}$

We use Gibbs sampling [45] to sample the precision matrix Λ_w of epigenomic effects from the posterior distribution. Specifically, Gibbs sampling requires a closed form posterior distribution. Due to the conjugacy of the Wishart prior of epigenomic precision matrix Λ_w to the multivariate normal distribution of epigenomic effect $\mathbf{w}$, the posterior of the epigenomic precision matrix Λ_w also follows Wishart distribution [46]:

$$\Lambda_w|\mathbf{w} \sim \mathcal{W}((\Lambda_0^{-1} + \mathbf{w}'\mathbf{w})^{-1}, \nu_0 + K) \tag{43}$$

Similarly, we sample λ_{ld} from Gamma posterior distribution:

$$\lambda_{ld}|b_{ld} \sim \Gamma(\alpha_0 + 0.5, (\beta_0 + \frac{(b_{ld} - g(\pi_{ld}))^2}{2})^{-1}) \tag{44}$$

To sample epigenomic effects $\mathbf{w}$ and prior bias $\mathbf{b}$ for disease $d = 1, \ldots, D$ and locus $l = 1, \ldots, L_d$, we employ a more powerful gradient-based sampling scheme namely Hamiltonian Monte Carlo (also known as hybrid Monte Carlo) (HMC) [47, 48], exploiting the fact that the joint posterior of our model is differentiable with respect to the model parameters $\mathbf{w}$ and $\mathbf{b}$ (**Supplementary Information**).

4.1.7 Functional enrichment

To assess functional enrichment of a given annotation, we propose a Bayesian likelihood ratio tests. Specifically, at t^{th} MCMC sampling iteration, we compare the likelihood of the null model that does not use annotation k ($\mathcal{L}_0^{(t)}$) with the likelihood of the alternative model that does ($\mathcal{L}_1^{(t)}$):

$$\mathcal{L}_0^{(t)} = \sum_l \log \sum_{\mathbf{c}_{ld}} p(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \Theta_0^{(t)}) p(\mathbf{c}_{ld}|\Theta_0^{(t)}) \tag{45}$$

$$\mathcal{L}_1^{(t)} = \sum_l \log \sum_{\mathbf{c}_{ld}} p(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \mathbf{a}_{ldk}, \Theta_1^{(t)}) p(\mathbf{c}_{ld}|\mathbf{a}_{kd}, \Theta_1^{(t)}) \tag{46}$$

$$\Delta\mathcal{L}_k^{(t)} = -2(\mathcal{L}_0^{(t)} - \mathcal{L}_1^{(t)}) \sim \chi^2(1) \tag{47}$$

where $\chi^2(1)$ is chi-squared distribution with one degree of freedom that we use to assess the significance of the annotation. The Bayesian credible interval of $\Delta\mathcal{L}_k^{(t)}$ forms naturally over the sampled model parameters and causal configurations after discarding the initial 20% of sampled values during burn-in period.

4.2 GWAS simulation

To assess the power of the proposed fine-mapping model in identifying causal variants and compare it with existing methods, we implemented a simulation pipeline adapted from [15]. Details are described in **Supplementary Information**. Briefly, the simulation can be divided into four steps:

1. simulate genotypes based on the haplotypes from 1000 Genome data (phase 1 version 3) using HapGen2 [49];

2. sample epigenomic enrichments from uniform distribution with maximum fold-enrichment defined by the total number of causal variants and the total number of variants harbored in that annotation and then randomly sample causal variants according to the simulated fold-enrichments from each of the 100 epigenomic annotations selected from 19 categories of primary tissue/cell types;

3. simulate phenotypes as a linear combination of the causal effect sizes plus random zero-mean Gaussian noise with pre-determined variance to achieve a given heritability h_g^2 (fixed at 0.25 unless mentioned otherwise);

4. compute p-values and z-scores (as t-statistics) by regressing phenotype on each SNP.

4.3 GWAS summary statistics and imputation

Overall, the summary statistics of each GWAS trait were downloaded from public domains. For each study, we first removed strand-ambiguous SNP (T/A, C/G) as well as SNPs with supporting sample sizes lower than a threshold. For BMI and the four lipid traits (HDL,LDL,TC,TG), we require for SNPs to have the minimum sample size of 80,000.

For CAD (T2D), we obtained only SNPs supported by at least 15,000 (9,000) cases and 50,000 (50,000) controls. We then imputed summary statistics using ImpG (v1.0.1) (`https://github.com/huwenboshi/ImpG`) [50] to 1000 Genome Phase 1 (version 3) data. Only the imputed SNPs with imputation quality measured as $r^2 > 0.6$ were retained. We then obtained the lead SNPs reported by each study and the SNPs within 100 kb genomic distance of the lead SNPs to form the genome-wide significant independent risk loci as the inputs to our fine-mapping algorithm. Table S1 summarizes the data from each individual GWAS study.

4.4 Running existing fine-mapping software on simulated data

The software `fgwas` [14] (version 0.3.4) were downloaded from GitHub. To enable fine-mapping, we issued `-fine` flag and specify the region numbers for each SNP in the input file as required by the software. GPA (0.9-3) [16] was downloaded from GitHub and run with default settings. To test for trait-relevant annotations, we followed the package vignette. Briefly, we fit two GPA models with and without the annotation and compared the two models by `aTest` function from GPA, which performs likelihood-ratio (LR) test via χ^2 approximation, and obtained the enrichment scores as the -log10 p-value. PAINTOR (version 2.1) was downloaded from GitHub [15]. As suggested in the documentation, we prepared a list of input files for every locus including summary statistics as t-statistics, LD matrices, and binary epigenomic annotations. We ran the software with default setting with assumption of at most two causal variants per locus.

4.5 LD score regression for functional enrichment

We obtained the LD score regression software LDSC (v1.0.0) (`https://github.com/bulik/ldsc`) to determine functional enrichments of the GWAS traits by partitioning heritabilities according to the variance explained by the LD-linked SNPs belonging to each functional categories [37]. Following the online LDSC manual (the partitioned heritability page), we first trained a baseline LDSC model using the 52 non-cell-type specific functional categories (plus one category that includes all SNPs) using the observed Z-scores of HapMap3 SNPs for each trait. We then trained 220 models on cell-type-specific annotations including 4 histone marks (H3K4me1, H3K4me3, H3K9ac, H3K27ac) and 100 well-defined cell types. For fine-mapping causal variants, we chose baseline and cell-type-specific epigenomic annotations with p-value < 0.05 adjusted by Benjamini-Hocherg method across 272 annotations over the 7 traits (i.e., 1,904 tests in total).

4.6 Code availability

RiVIERA-MT software implemented as a standalone open-source R package is freely available from Github repository `https://github.mit.edu/pages/liyue/riviera/`.

Figure Legends

Figure 1: Fine-mapping problem illustration. Three case scenarios were simulated to illustrate the fine-mapping problem (left-right): one causal variant, 3 causal variant, 10 causal variants within the locus. In each case, there are four types of data (top-bottom): genetic signals as GWAS -log10 p-values, epigenomic activities as cumulative counts of overlapping SNPs over the 100 epigenomic annotations (where the causal variants were enriched in some annotations), inferred posterior inclusion probabilities (PIP), linkage disequilibrium matrix of the simulated genomic region. The causal variants are the red diamond and highlighted by the vertical line to aid visualization.

Figure 2: Functional enrichments and RiVIERA-MT model. **a**. Functional enrichments of cell-type-specific annotations across 7 traits. Heatmap illustrates the underlying co-enrichment of the 7 related traits (columns) across many cell or tissue types (rows). The intensities reflect the -log10 p-values from LDSC estimate [37]. Red boxes highlighted known relevant tissues for the corresponding traits. **b**. Stochastic sampling of causal configurations. The sampling scheme was adapted from [18,35]. For simplicity, we display a locus of 3 SNPs with 0 and 1 indicating non-causal and causal status. Starting from 1 causal variant per locus on the left, we have 3 choices to place the causal status in each of the variants. Suppose we sample the configuration '010' (highlighted in red box) based on its posterior probabilities relative to the other two configurations. We then apply 3 types of operations that define the "neighborhood" of the current configuration (i.e., `nbd(010)`): (1) adding one causal variant; (2) removing one casual variant (we do not consider this step when there is only one causal variant in the configuration); (3) swapping causal variant with a non-causal variant. We then sample from the posterior normalized within the neighborhood of '010' a new configuration, say '110' and compare the joint posteriors of the proposed configuration namely '110' and that of the current configuration namely '010' to determine whether we should accept the proposal and so on. **c**. RiVIERA-MT expressed in probabilistic graphical model. Shaded nodes are observed data and unshaded are latent variables or model parameters. The plates represent repeated pattern of same entities as indexed by l for loci, k for annotations, and d for diseases. The meaning of each variant is annotated beside each node. Please refer to the main text for details.

Figure 3: Power comparison on inferring causal variant. Proportion of causal variants is plotted as a function of increasing number of variants selected by 7 SNP prioritization methods. The boxplots are based on 500 independent simulations.

Figure 4: Functional enrichment analysis. We estimated the enrichment of each annotations based on likelihood ratio tests. The y-axis is the estimate and the x-axis is the underlying fold-enrichments that were used to sample the causal variants from the corresponding annotation. The error bar indicates the 90% credible interval of the Bayesian LRT estimates by RiVIERA-MT. The inset boxplot display the overall correlations based on 100 simulations.

Figure 5: Inferring variants across multiple traits. We ran RiVIERA-MT single-trait and multi-trait modes on the simulated data containing 2-10 traits that do not have the same risk loci but related via the correlation of functional fold-enrichments. The y-axis indicates the average number of SNPs per locus required to detect 90% causal variants over 100 simulations per number of traits (x-axis). The top and bottom panels are the performance of the empirical prior and posterior, respectively. The left and right columns indicate binary noise-free annotations (i.e., the underlying annotations from which the causal SNPs were sampled from) and noisy annotations (i.e., standardized and scaled continuous annotations ranging from -1 and 1).

Figure 6: Venn diagrams of the number shared variants predicted by each method. For RiVIERA-MT single-trait (riviera_st), RiVIERA-MT multi-trait mode (riviera_mt), and PAINTOR, we constructed 90% credible sets, and for GWAS -logP (gwas_logp) we took the genome-wide significant SNP with p<5E-8. The bottom right plot displaying the model confidence in terms median posterior for each SNP within the 90% credible SNPs as a function of increasing number of supporting methods.

Figure 7: Enrichments of GTEx eQTL SNPs. **a**. The hypergeometric enrichment of SNPs in the GTEx whole blood eQTL SNPs as a function of increasing number of top variants chosen by each method. **b**. Same as **a** except the tissue-specific eQTL SNPs were chosen for each trait. **c**. Heatmap of enrichments of eQTL SNPs across 44 GTEx tissues. We overlapped the 90% credible sets predicted by each method for each trait with the GTEx eQTL SNPs. The color intensities are based on BH-adjusted -log10 p-values of hypergeometric enrichment tests.

References

1. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

2. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk of complex disease. *Current Opinion in Genetics & Development* **18**, 257–263 (2008).

3. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).

4. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of gwas discovery. *The American Journal of Human Genetics* **90**, 7–24 (2012).

5. Welter, D. *et al.* The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research* **42**, D1001–D1006 (2014).

6. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

7. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

8. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* **30**, 1095–1106 (2012).

9. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics* **45**, 124–130 (2013).

10. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory dna. *Science* **337**, 1190–1195 (2012).

11. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).

12. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).

13. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome research* **22**, 1748–1759 (2012).

14. Pickrell, J. K. Joint Analysis of Functional Genomic Dataand Genome-wide Association Studies of 18 Human Traits. *The American Journal of Human Genetics* **94**, 559–573 (2014).

15. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics* **10**, e1004722 (2014).

16. Chung, D., Yang, C., Li, C., Gelernter, J. & Zhao, H. GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. *PLoS Genetics* **10**, e1004787 (2014).

17. Chen, W. *et al.* Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719–736 (2015).

18. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics (Oxford, England)* (2016).

19. Anttila, V. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* 1–9 (2015).

20. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics* 1–10 (2016).

21. Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics* **47**, 1121–1130 (2015).

22. Burton, P. R. *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genetics* **39**, 1329–1337 (2007).

23. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics* **47**, 381–386 (2015).

24. of the Psychiatric Genomics Consortium, C.-D. G. *et al.* Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* **381**, 1371–1379 (2013).

25. O'Dushlaine, C. *et al.* Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature Neuroscience* **18**, 199–209 (2015).

26. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42**, 937–948 (2010).

27. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics* **43**, 333–338 (2011).

28. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

29. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* **44**, 981–990 (2012).

30. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

31. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics* **5**, 1780–1815 (2011).

32. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *bioRxiv* (2016). URL `http://biorxiv.org/content/early/2016/05/21/054684.1`. `http://biorxiv.org/content/early/2016/05/21/054684.1.full.pdf`.

33. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *American journal of human genetics* **97**, 260–271 (2015).

34. Hormozdiari, F., Kichaev, G., Yang, W.-Y., Pasaniuc, B. & Eskin, E. Identification of causal genes for complex traits. *Bioinformatics* **31**, i206–i213 (2015).

35. Hans, C., Dobra, A. & West, M. Shotgun Stochastic Search for "Large p" Regression. *Journal of the American Statistical Association* **102**, 507–516 (2007).

36. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *AJHG* **95**, 535–552 (2014).

37. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

38. Consortium, T. . G. P. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **490**, 56–65 (2013).

39. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

40. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics* **44**, 1294–1301 (2012).

41. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS genetics* **11**, e1005176 (2015).

42. Wallace, C. *et al.* Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. *bioRxiv* 015164 (2015).

43. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *American journal of human genetics* **98**, 1114–1129 (2016).

44. Servin, B. & Stephens, M. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genetics* **3**, e114 (2007).

45. Geman, S. & Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **PAMI-6**, 721–741 (1984).

46. Bernardo, J. M. & Smith, A. F. *Bayesian theory*, vol. 405 (John Wiley & Sons, 2009).

47. Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. Hybrid monte carlo. *Physics letters B* **195**, 216–222 (1987).

48. Neal, R. M. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* **2** (2011).

49. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics (Oxford, England)* **27**, 2304–2305 (2011).

50. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics (Oxford, England)* **30**, 2906–2914 (2014).

Supplementary Information

S1 HMC method details

S2 RiVIERA-MT algorithm

S3 GWAS simulation

S4 Supplementary Fig.

Figure S1: Simulation pipeline

Figure S2: Model convergence

Figure S3: Power comparison

Figure S4: Estimation of per-SNP variance explained

Figure S5: Visualization of fine-mapping results

Figure S6: Correlation of PIP from single-trait and multi-trait models

S5 Supplementary Table

Table S1: GWAS summary statistics used in this study

Figure 1

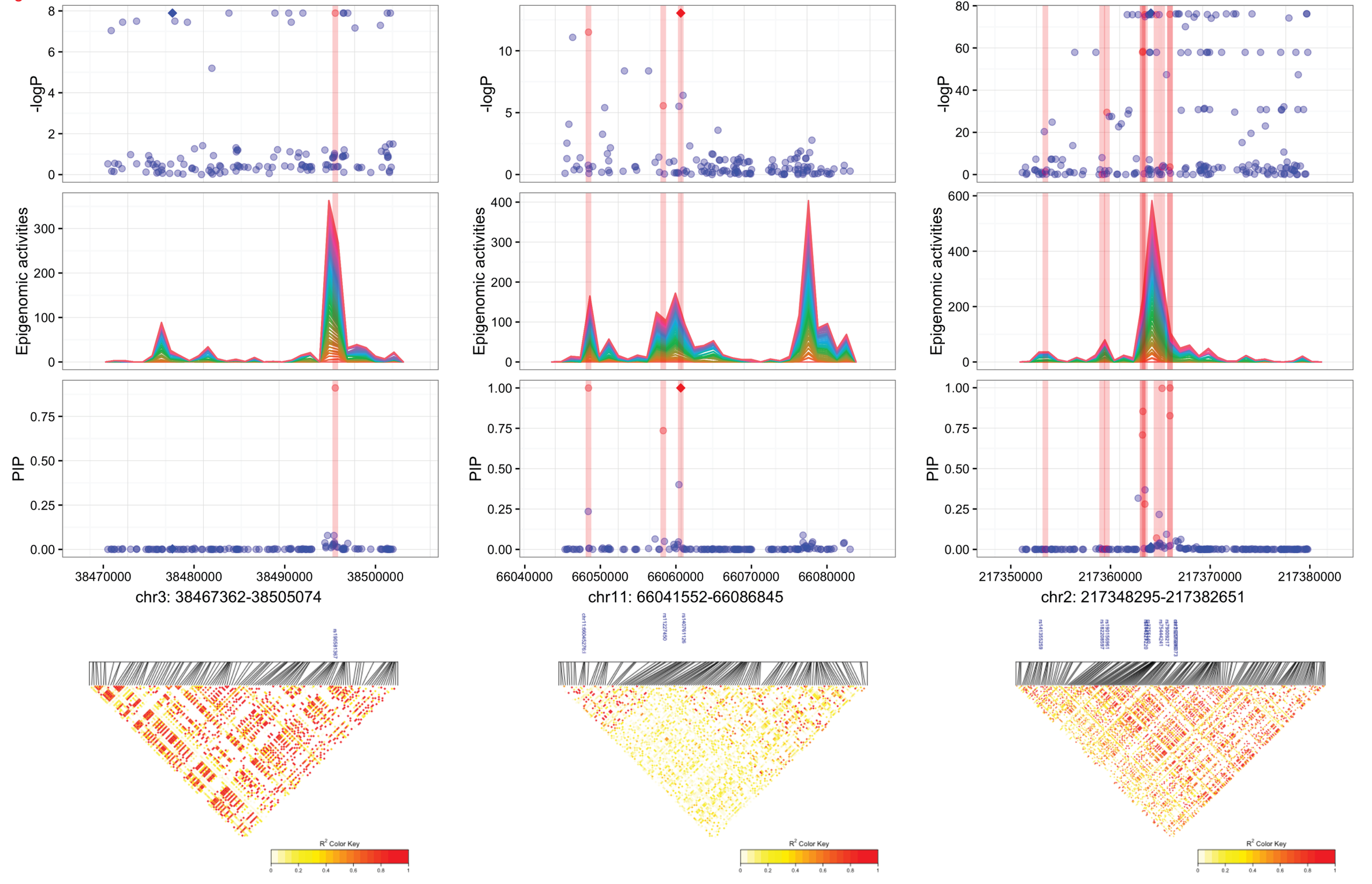

-logP
Epigenomic activities
PIP
chr3: 38467362-38505074
chr11: 66041552-66086845
chr2: 217348295-217382651
R² Color Key

Figure 2

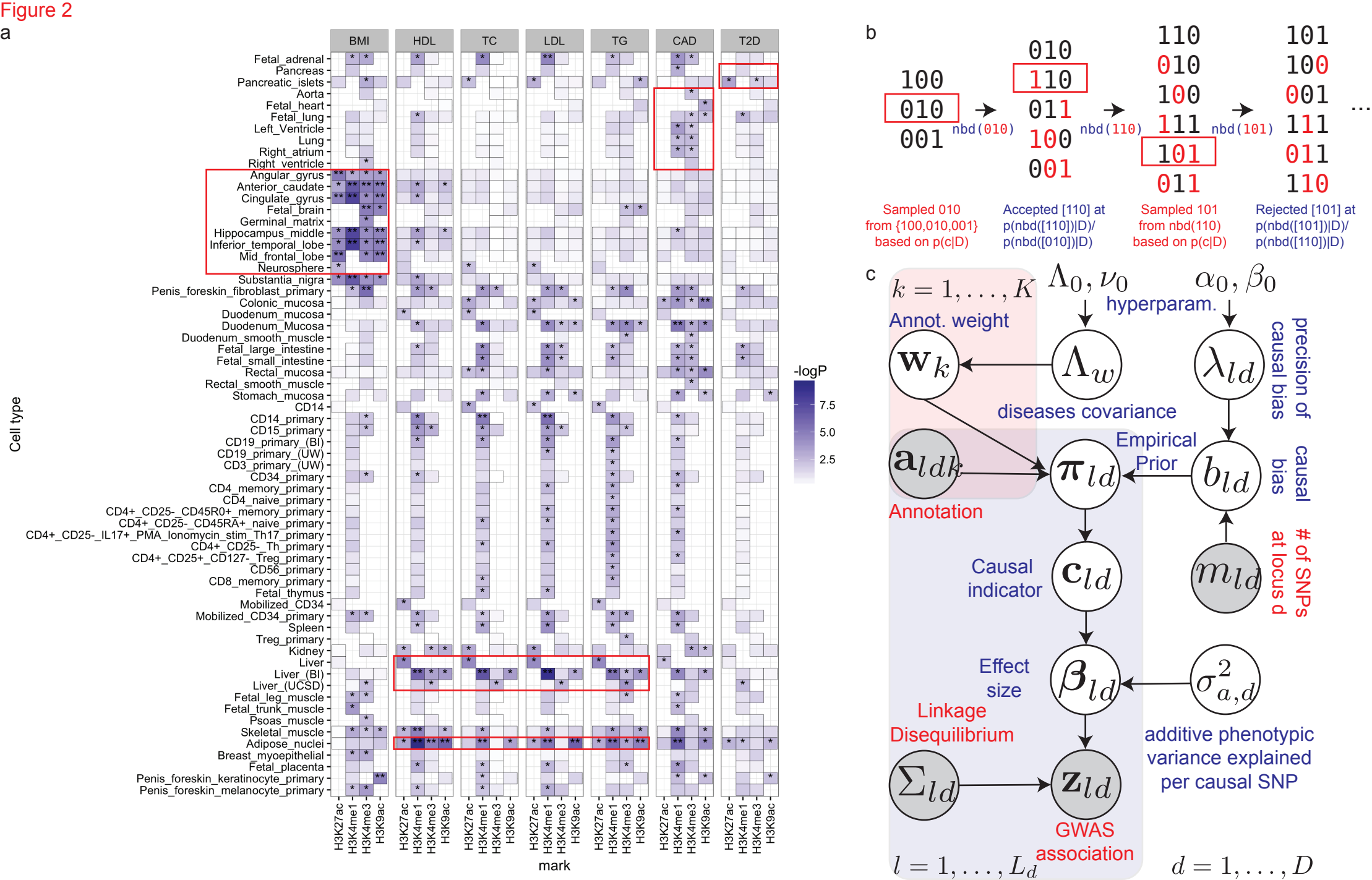

a
BMI
HDL
TC
LDL
TG
CAD
T2D
Fetal_adrenal
Pancreas
Pancreatic_islets
Aorta
Fetal_heart
Fetal_lung
Left_Ventricle
Lung
Right_atrium
Right_ventricle
Angular_gyrus
Anterior_caudate
Cingulate_gyrus
Fetal_brain
Germinal_matrix
Hippocampus_middle
Inferior_temporal_lobe
Mid_frontal_lobe
Neurosphere
Substantia_nigra
Penis_foreskin_fibroblast_primary
Colonic_mucosa
Duodenum_mucosa
Duodenum_Mucosa
Duodenum_smooth_muscle
Fetal_large_intestine
Fetal_small_intestine
Rectal_mucosa
Rectal_smooth_muscle
Stomach_mucosa
CD14
CD14_primary
CD15_primary
CD19_primary_(BI)
CD19_primary_(UW)
CD3_primary_(UW)
CD34_primary
CD4_memory_primary
CD4_naive_primary
CD4+_CD25-_CD45R0+_memory_primary
CD4+_CD25-_CD45RA+_naive_primary
CD4+_CD25-_IL17+_PMA_Ionomycin_stim_Th17_primary
CD4+_CD25-_Th_primary
CD4+_CD25+_CD127-_Treg_primary
CD56_primary
CD8_memory_primary
Fetal_thymus
Mobilized_CD34
Mobilized_CD34_primary
Spleen
Treg_primary
Kidney
Liver
Liver_(BI)
Liver_(UCSD)
Fetal_leg_muscle
Fetal_trunk_muscle
Psoas_muscle
Skeletal_muscle
Adipose_nuclei
Breast_myoepithelial
Fetal_placenta
Penis_foreskin_keratinocyte_primary
Penis_foreskin_melanocyte_primary
Cell type
H3K27ac
H3K4me1
H3K4me3
H3K9ac
mark
-logP
7.5
5.0
2.5
b
100
010
001
nbd(010)
010
110
011
100
001
nbd(110)
110
010
100
111
101
011
nbd(101)
101
100
001
111
011
110
...
Sampled 010 from {100,010,001} based on p(c|D)
Accepted [110] at p(nbd([110])|D)/ p(nbd([010])|D)
Sampled 101 from nbd(110) based on p(c|D)
Rejected [101] at p(nbd([101])|D)/ p(nbd([110])|D)
c
$k = 1, \ldots, K$
Λ_0, ν_0
α_0, β_0
hyperparam.
Annot. weight
$\mathbf{w}_k$
Λ_w
λ_{ld}
precision of causal bias
diseases covariance
$\mathbf{a}_{ldk}$
$\boldsymbol{\pi}_{ld}$
Empirical Prior
b_{ld}
causal bias
Annotation
Causal indicator
$\mathbf{c}_{ld}$
m_{ld}
# of SNPs at locus d
Effect size
$\boldsymbol{\beta}_{ld}$
$\sigma^2_{a,d}$
Linkage Disequilibrium
additive phenotypic variance explained per causal SNP
Σ_{ld}
$\mathbf{z}_{ld}$
GWAS association
$l = 1, \ldots, L_d$
$d = 1, \ldots, D$

Figure 3

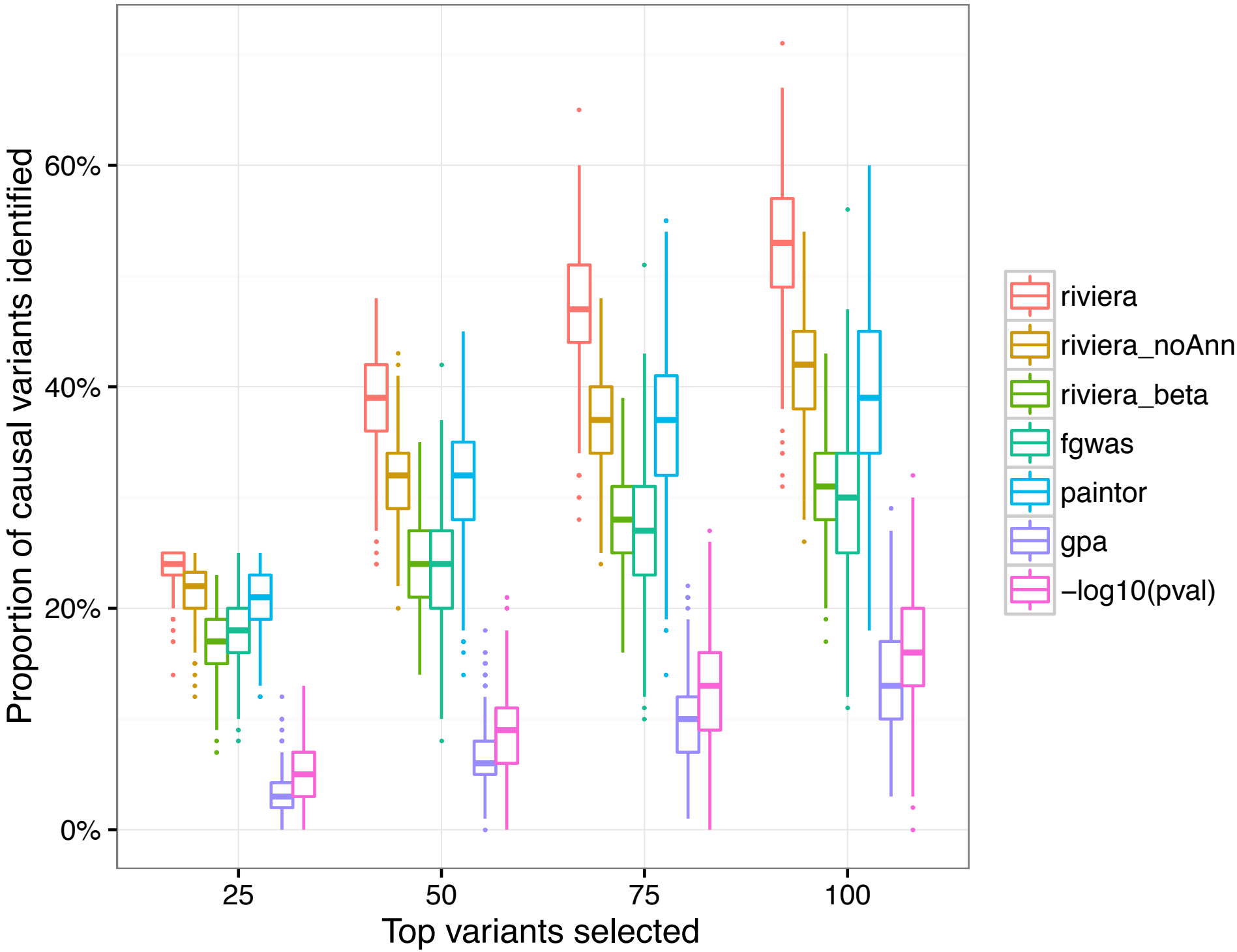

Proportion of causal variants identified
60%
40%
20%
0%
25
50
75
100
Top variants selected
riviera
riviera_noAnn
riviera_beta
fgwas
paintor
gpa
−log10(pval)

Figure 4

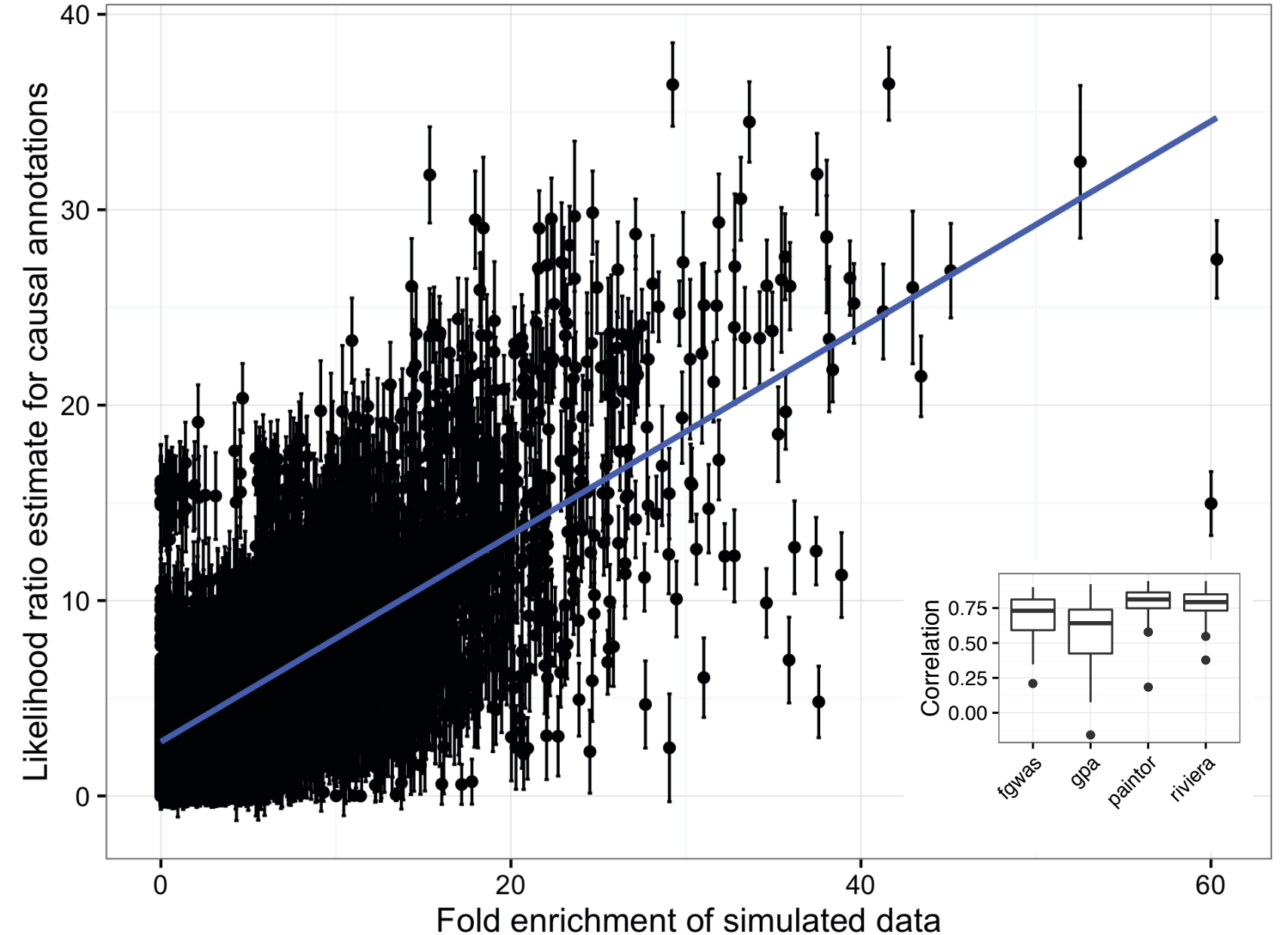

Likelihood ratio estimate for causal annotations
Fold enrichment of simulated data
0
10
20
30
40
0
20
40
60
Correlation
0.00
0.25
0.50
0.75
fgwas
gpa
paintor
riviera

Figure 5

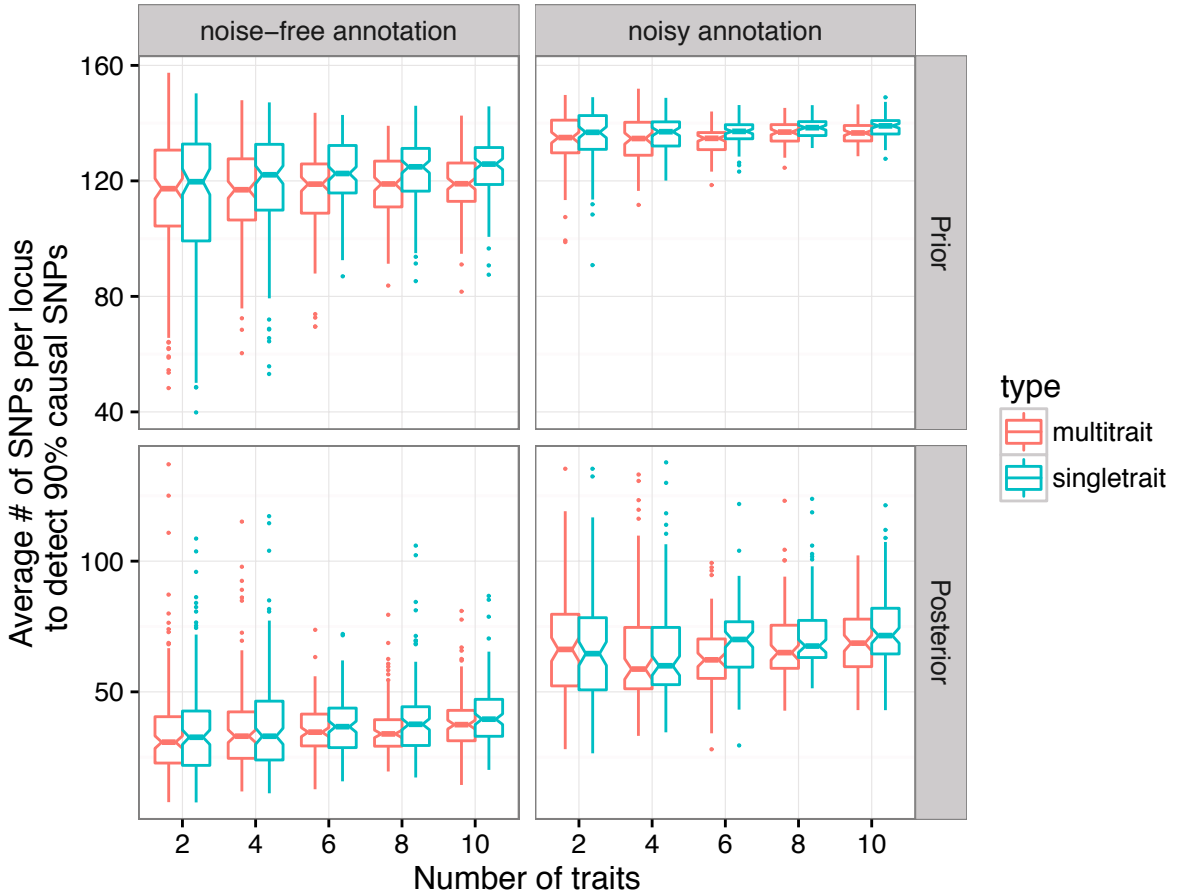
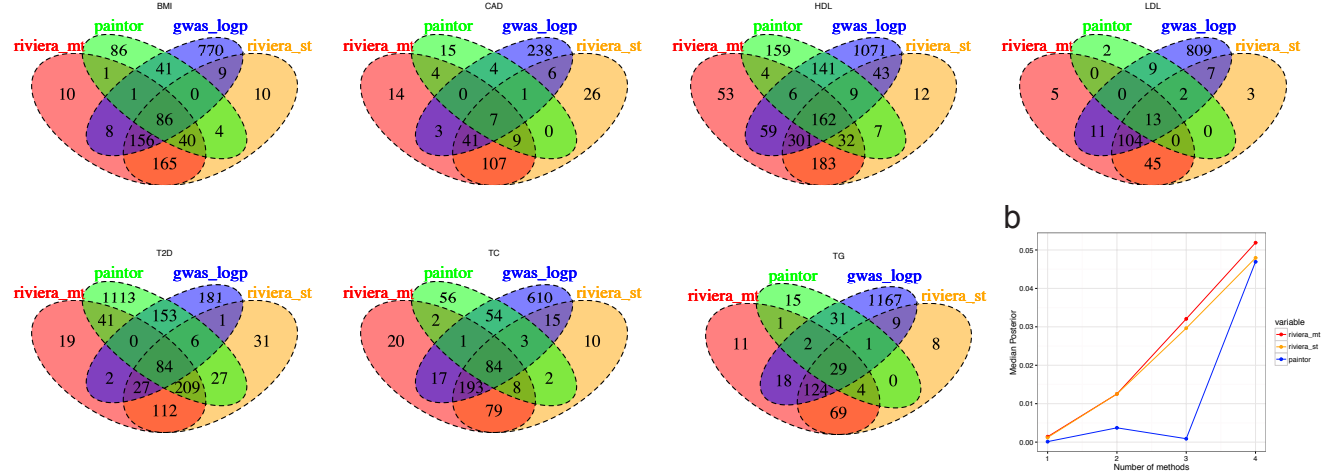

noise−free annotation
noisy annotation
Prior
Posterior
Average # of SNPs per locus
to detect 90% causal SNPs
160
120
80
40
100
50
2
4
6
8
10
Number of traits
type
multitrait
singletrait

Figure 6

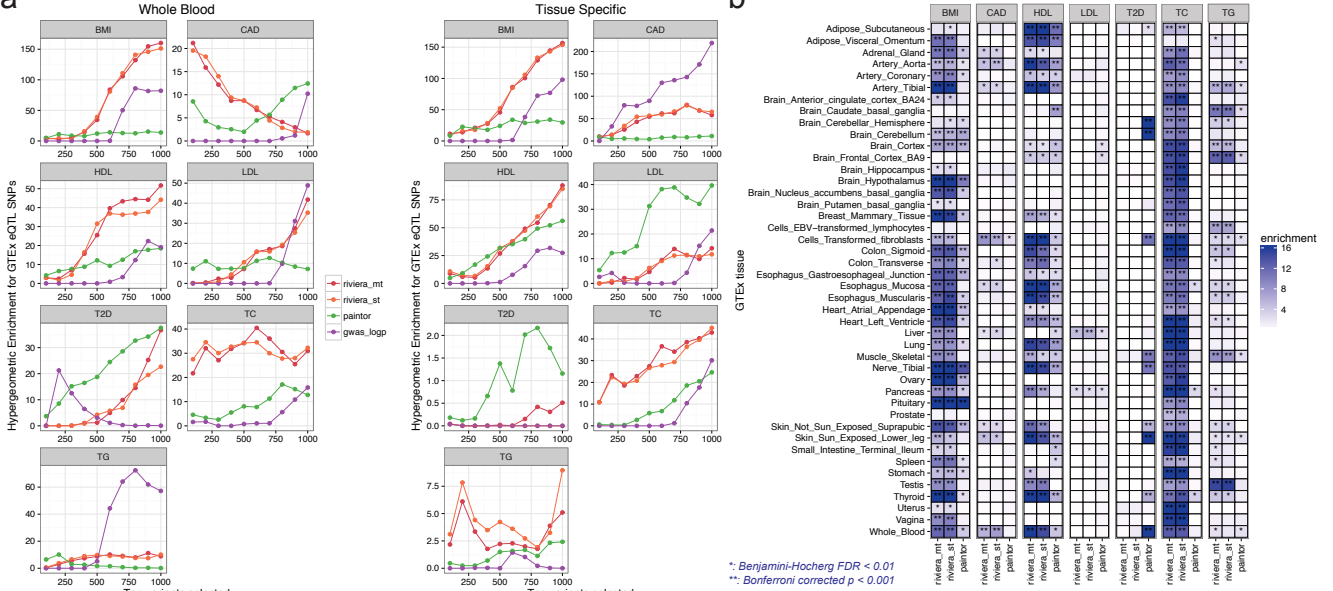
a


BMI
paintor
gwas_logp
riviera_mt
riviera_st
86
770
1
41
9
10
1
0
10
86
8
156
40
4
165
CAD
paintor
gwas_logp
riviera_mt
riviera_st
15
238
4
4
6
14
0
1
26
7
3
41
9
0
107
HDL
paintor
gwas_logp
riviera_mt
riviera_st
159
1071
4
141
43
53
6
9
12
162
59
301
32
7
183
LDL
paintor
gwas_logp
riviera_mt
riviera_st
2
809
0
9
7
5
0
2
3
13
11
104
0
45
T2D
paintor
gwas_logp
riviera_mt
riviera_st
1113
181
41
153
1
19
0
6
31
84
2
27
209
27
112
TC
paintor
gwas_logp
riviera_mt
riviera_st
56
610
2
54
15
20
1
3
10
84
17
193
8
2
79
TG
paintor
gwas_logp
riviera_mt
riviera_st
15
1167
1
31
9
11
2
1
8
29
18
124
4
0
69

b

Median Posterior
0.05
0.04
0.03
0.02
0.01
0.00
1
2
3
4
Number of methods
variable
riviera_mt
riviera_st
paintor

Figure 7


a
Whole Blood
Tissue Specific
BMI
CAD
HDL
LDL
T2D
TC
TG
Hypergeometric Enrichment for GTEx eQTL SNPs
Top variants selected
riviera_mt
riviera_st
paintor
gwas_logp
b
GTEx tissue
Adipose_Subcutaneous
Adipose_Visceral_Omentum
Adrenal_Gland
Artery_Aorta
Artery_Coronary
Artery_Tibial
Brain_Anterior_cingulate_cortex_BA24
Brain_Caudate_basal_ganglia
Brain_Cerebellar_Hemisphere
Brain_Cerebellum
Brain_Cortex
Brain_Frontal_Cortex_BA9
Brain_Hippocampus
Brain_Hypothalamus
Brain_Nucleus_accumbens_basal_ganglia
Brain_Putamen_basal_ganglia
Breast_Mammary_Tissue
Cells_EBV−transformed_lymphocytes
Cells_Transformed_fibroblasts
Colon_Sigmoid
Colon_Transverse
Esophagus_Gastroesophageal_Junction
Esophagus_Mucosa
Esophagus_Muscularis
Heart_Atrial_Appendage
Heart_Left_Ventricle
Liver
Lung
Muscle_Skeletal
Nerve_Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin_Not_Sun_Exposed_Suprapubic
Skin_Sun_Exposed_Lower_leg
Small_Intestine_Terminal_Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole_Blood
enrichment
*: Benjamini-Hocherg FDR < 0.01
**: Bonferroni corrected p < 0.001