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 using a large number of annotations. Moreover, many complex traits are not only genetically related but also potentially share causal mechanisms. We develop an integrative Bayesian fine-mapping model named RiVIERA-MT. The key features of RiVIERA-MT include: 1) the ability to model epigenomic covariance of multiple related traits; 2) efficient posterior inference of causal configuration; 3) efficient 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 accurate estimation of the functional priors. We applied RiVIERA-MT to 27 GWAS summary statistics, in-cis eQTL of risk genes in 44 tissues from GTEx samples, and finally to infer causal GWAS/eQTL SNPs, causal genes, and causal pathways. To leverage potential tissue-specific epigenomic co-enrichments among these traits, we harnessed hundreds of functional annotations compiled from ENCODE/Roadmap consortium. Overall, we observed an improved power in detecting functionally coherent risk SNPs with the Bayesian prior dictated by the weighted tissue-specific annotations compared to solely genetically driven model.

1 Introduction

Genome wide association studies (GWAS) have identified a large number of disease-associated genetic risk loci of many complex human diseases of potentially intricate polygenic structures. Systematic investigation of these loci involving fine-mapping the causal variants can help decipher the underlying epidemiology at the single-nucleotide resolution 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]).

To help interpret non-coding genetic signals and gain insights into potential regulatory functions, several international consortia have recently released a plethora of genome-wide annotations. In particular, ENCODE/Roadmap Epigenome consortium provided diverse epigenomic marks in diverse cell types and tissues, providing a multi-dimensional reference map to elucidate enhancer locations indicative of cis-regulatory functions [6, 7]. The 1000 Genome consortium provided genotype information of hundreds of individuals that share ancestral haplotypes, which can be used to obtain the LD information in terms SNP-by-SNP genetic correlation to disentangle disease causing signals from signals of passenger SNPs in the same haplotypes as the causal SNPs [8]. Moreover, GTEx (Genotype-tissue Expression) provided individual-level genotype and tissue-specific gene expression data for the same individuals, making it possible to link regulatory SNPs to genes in expression quantitive trait loci (eQTL) analysis [9].

Indeed, several studies have implicated enrichment of GWAS variants in putative regulatory elements in enhancer-associated histone modifications, regions of open chromatin, and conserved non-coding elements [3,6,10–14]. Moreover, these enrichments have also been used to predict relevant cell types and non-coding annotations for specific traits [6,11,15,16]. However, it remains a challenge to integrate GWAS signals with functional annotations, LD information, and GTEx eQTL information to generate actionable biological hypotheses to elucidate the underlying disease-causing mechanisms (**Fig.** 1a).

Moreover, many complex traits are genetically related [17,18] and potentially share causal mechanisms such as immune diseases [19,20], metabolic traits [21], and psychiatric disorders [22, 23]. 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. To the best of our knowledge, currently there is a lack of such method.

GWAS data are usually not publicly available as individual-level genotype data due to data transfer and privacy concern but rather become increasingly available in the form of marginal summary statistics for each SNP. Several recently developed methods have leveraged the GWAS summary statistics [24, 25] as well as functional annotations to aid causal inference [26–28]. Recently, we also developed a model called RiVIERA-beta [29], which uses functional annotations to infer causal variants by modeling the GWAS p-values via Beta density. Although RiVIERA-beta works 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.

In this article, we describe a Bayesian framework called RiVIERA-MT (Risk Variant Inference using Epigenomic Reference Annotations to predict Multiple Trait-causing mutations) to fine-map causal variants in a single or multiple related traits using GWAS summary statistics with the aid of LD information from 1000 Genome reference panel and the functional epigenomic annotations. Compared to existing methods, the main features of RiVIERA-MT is the ability to perform efficient 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 fine-map causal variants of 27 well-powered GWAS traits. To leverage potential tissue-specific epigenomic co-enrichments among these traits, we harness a large compendium of epigenomic annotations compiled from ENCODE/Roadmap consortium, including 4 previously implicated epigenomic marks (H3K4me1, H3K4me3, H3K27ac, H3K9ac) across 100 well characterized cell types and tissues [7]. Moreover, we also apply RiVIERA-MT to fine-map variants in eQTL of genes that are within 1 megabase (Mb) distance from the GWAS loci of the 27 traits using the SNP-by-gene summary statistics across 44 tissues from GTEx consortium. This allows us to revisit the 27 complex human traits from a novel perspective by inferring their underlying regulatory variants implicated at the tissue-specific epigenomic contexts and the associated genes based on the joint posterior of the co-localized GWAS and eQTL SNPs.

2 Results

2.1 RiVIERA-MT method overview

We describe a Bayesian model to infer causal variants. **Fig.** 1b illustrates the fine-mapping problems. To fine-map causal variants, we take into account three lines of evidence (from top to bottom) namely (1) linkage disequilibrium (LD as SNP-by-SNP Pearson correlation matrix), (2) genetic signals as z-scores displayed here in terms of GWAS -log10 p-values, (3) epigenomic activities overlapping SNPs and infer the posterior inclusion probabilities (PIP) of each SNP for being causal within the risk locus. The causal variants are the red diamond and highlighted by the vertical line to aid visualization. Our model is able to infer PIP accurately for the cases where the risk locus harbors a single causal variant as well as the cases where multiple causal variants coincide within the same locus. In the first scenario, the risk locus harbors one causal variant (red circle), which drives the genetic signals of other non-causal variants via linkage disequilibrium. 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 [26, 30]. However, these methods become inadequate when there are more than one causal variant within the same locus as shown in the second example 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 [24, 25, 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 [24,25,27,33,34] (**Methods**). However, as illustrated in **Fig.** 2 and detailed in **Methods** (**Supplementary Fig.** S2), RiVIERA-MT has several significant key 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);

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 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 Bayesian treatments (**Fig.** 2c);

4. Simultaneously modeling the underlying heritability parameters as per-SNP additive 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 quadratically 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 [28], 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. It 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 [35].

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.** S4a). At each MCMC iteration, we performed a fixed number of stochastic searches of *the entire local neighborhood* of the current causal configurations [25, 36]. 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 steady 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.** S4b). 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 (**Methods**, **Supplementary Text** 3). RiVIERA-MT demonstrates a consistently improved power in detecting causal variants (**Fig.** 3a). 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 [27], 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 [26] and RiVIERA-beta [29], both of which assume single causal variant per locus. Nonetheless, fgwas and RiVIERA-beta perform better than marginal GWAS -log10 p-value, which assume that all of the SNPs are independent. We also examined the power conferred by each method focusing on loci harboring at most 3 causal variants, which is commonly suggested in several recent studies [30, 37] and the assumption of the PAINTOR model. As expected, PAINTOR performs the second best in this case. Nonetheless, RiVIERA-MT still demonstrates superior performance (**Fig.** 3b).

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.** 1 illustrates a locus harboring 3 causal variants. In this case, 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 reasonably calibrated with 90% credible set identifying over 75% causal variants. More sampling may be required to achieve higher proportion of causal variants to be included in the 90% credible set (**Supplementary Fig.** S6). Nonetheless, our model demonstrates consistently superior performance by identifying additional number of causal variants when selecting more than 100 variants per locus (**Supplementary Fig.** S5). When the number of causal variants to consider go beyond a model's ability to infer, the causal signals are poorly calibrated (**Supplementary Fig.** S6). 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 100 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 detecting causal variants 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.** S7). PAINTOR conferred comparable performance and performed 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 Joint inference of multiple traits

An important 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 related traits. Despite distinct risk loci, we correlated the traits based on the functional co-enrichments of the underlying 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 further gain of power (**Fig.** 4a). As expected, the improvement is primarily attributable to the improved empirical prior inference (**Fig.** 4a) because it is dictated by the annotation weights with covariance reflecting the epigenomic co-enrichments. Indeed, we further confirmed that the improved power conferred by the multi-trait model is precisely due to its capability to model the co-enrichment of the traits by applying the single-trait and multi-trait models to unrelated traits, where we observed no improvement (**Fig.** 4b).

2.6 Applications to 27 GWAS datasets with full summary statistics revealed meaningful cell-type specific enhancer activities

To demonstrate the general utility of RiVIERA-MT, we obtained 27 GWAS summary statistics from the corresponding GWAS consortia (**Supplementary Table** S1). To infer epigenomic functional prior, we used a set of 272 well defined annotations including 52 baseline annotations pre-processed from various sources and 220 cell-type-specific annotations over four transcription-activating histone marks pre-compiled from the original Roadmap Epigenomic Consortium [7, 16] (**Methods**). To gain some intuition for the following analyses, we first applied RiVIERA-MT separately to each of the GWAS data and examined whether the

estimated functional enrichments are consistent with the expected tissue context of some of the well characterized GWAS traits. For instance, we expected that the immune traits are mostly enriched for immune cell types whereas schizophrenia for predominantly enriched for the brain cell group (**Fig.** 5a). To this end, we calculated an aggregated enrichment scores over each of the 10 tissue group (**Methods**). Because our enrichment scores are based on only the SNPs within the GWAS loci and hence not genome-wide, we applied LD score regression (LDSC) [16] to the same GWAS data using the same set of 272 annotations to compare how much our enrichment estimates differ from the genome-wide enrichment approach. Overall, we observed a consistent agreement between RiVIERA-MT and LDSC estimates (**Fig.** 5b; **Supplementary Fig.** S10) with Pearson correlation of 0.29 ($p < 8.41 \times 10^{-7}$; t-test with 2 degrees of freedom). Nonetheless, there are some meaningful tissue-specific enrichments such as 'Year of Education' enriched in 'CNS', 'Fasting glucose' in 'Adrenal/Pancreas' and 'Liver' that are missed by our locus-based approach, underscoring additional informative signals outside of the risk loci.

2.7 Epigenomic prior improved fine-mapping resolution

To fine-map causal variants in the GWAS risk loci of the GWAS traits, we applied RiVIERA-MT without and with the 272 annotations to the 27 GWAS summary statistics (**Methods**). In most of the GWAS traits, using functional annotations significantly reduced the number of SNPs in the 90% credible set compared to the same model without using annotations and/or selecting SNPs based only on common GWAS threshold of 5E-8 or 1E-5, respectively (**Supplementary Table** S2). We attributed the improved confidence to the ability of our RiVIERA-MTto incorporate functional information into the Bayesian prior. Without the annotations, each SNP would have had an equal prior of 1/M, assuming *apriori* one causal variant per locus containing M SNPs. Indeed, this is equivalent to running our model without annotation with only the bias term ($w_0 = \log(1/M) - \log(1 - 1/M)$) in the logistic prior function, which will ultimately penalizes the posterior for selecting multiple variants due to Bayesian Occam's razor. Consequently, this led to individual SNPs with lower posteriors and thus more SNPs are required to form the total 90% posterior mass.

To gain further intuition on how our RiVIERA-MT model dissects functional GWAS signals, we visualized the fine-mapped risk loci especially the ones with strong LD (**Fig.** 6, **Supplementary Fig.** S11). For instance, the multiple sclerosis risk locus harbors 101 (163) SNPs with GWAS p-value < 5E-8 (p-value < 1E-5), making identifying the underlying causal variants challenging without the aid of functional information. On the other hand, we observed a strong functional prior for only a small number of SNPs concentrated around the peak of the H3K27ac weighted by the model enrichment coefficients in immune cells. By taking into account both the GWAS likelihood and the prior, RiVIERA-MT selected only a small set of SNPs as the prominent hits within these loci and by and large refining credible set of SNPs for most the 27 GWAS traits (**Table** S2) facilitating possible functional follow-ups.

2.8 Joint fine-mapping related GWAS traits with distinct risk loci further improved fine-mapping resolution

As observed above, many traits are functionally enriched for enhancer elements active in the same tissue type. We have shown in our simulation above that we gained additional power by modeling these co-enrichments via the trait-trait covariance matrix regardless whether the two related traits share the common risk loci. Here we applied RiVIERA-MT to jointly model three separate sets of related GWAS traits, namely 8 immune traits, 4 psychiatric traits (PSY), and 3 lipid traits (**Supplementary Table** S2). Compared to the single-trait mode, the multi-trait mode RiVIERA-MT model further reduced the 90% credible set in 10 out of 15 GWAS traits, implying its ability to leverage the co-enrichment among related traits in the posterior inference (**Supplementary Table** S2).

We also compared the overlaps of credible sets as the fraction of the selected SNPs in the query set that overlap with the selected SNPs in the subject set (thus, the pairwise comparison table is not symmetrical as the number of selected SNPs by different methods differed from each other). In most traits, substantial number of GWAS SNPs (p<1E-5) are in the credible sets of the SNPs selected by the RiVIERA-MT with single-trait and multi-trait modes, indicating that the SNPs selected by the models are indeed consistent with the genetic signals of GWAS (**Supplementary Fig.** S12). Moreover, a high fraction of the SNPs (median 48.2% across traits) within the 90% credible sets constructed by the joint model overlap with the 90% credible sets of the single-trait model. In contrast, a much lower fraction of the SNPs (median 36.7%) in the single-trait models' 90% credible set were also found in those from the multi-trait inference (**Supplementary Fig.** S13). Thus, the results suggest that we can further improve the fine-mapping resolution for many of the GWAS traits by modeling multiple related traits simultaneously.

Similar as above, we visualized several distinct risk loci in comparison with the inference results between single-trait and multi-trait models. For instance, within a LDL GWAS locus, we observed a strong enhancer activities in the liver tissue from Roadmap Epigenome database. The single-trait and multi-trait models nominate the same top two variants as potential causal variant, neither of them is the most significant GWAS SNP. Nonetheless, the multi-trait model conferred much higher prior and posterior to the top variants compared to those from the single-trait model, implying an increased statistical confidence when the variants coincide within the enhancer peak of the enriched tissue-context (**Fig.** 7). Similarly, we observed a readjustment of posteriors for variants within a schizophrenia risk locus, where the multi-trait model expressed strong preference to the variant right on the summit of the CNS enhancer peak with higher posterior compared to another variants that exhibit slightly higher GWAS signals but is located in a much less prominent enhancer active region.

2.9 Prioritized SNPs by RiVIERA-MT are evolutionarily conserved

To further investigate the functional implications of our prioritized SNPs, we examined their evolution conservation scores. Specifically, we compared the average phastCons46way conservation scores of GWAS SNPs (p<5E-8), SNPs in 90% credible set constructed by single-trait model without annotation (PIP A-) and with annotation (PIP A+ single-trait),

as well as the 90% credible set derived from the joint models (PIP A+ multi-trait; operating on 3 separate groups of related traits) (**Fig.** 8). Indeed, the prioritized SNPs by the fine-mapping model (PIP A-) using both genetic and LD information outperformed GWAS SNPs based on marginal p-value alone in majority of the traits. Moreover, SNPs prioritized by RiVIERA-MT using epigenomic annotations generally exhibit higher conservation scores than the credible set without using the annotations. Furthermore, the prioritized SNPs by our joint RiVIERA-MT model exhibit the highest conservation scores compared to SNPs prioritized by other 3 methods in 7 out of the 15 traits that we examined (**Fig.** 8b). Together, we concluded that SNP prioritization benefits substantially when taking into LD information, epigenomic annotations, and the co-enrichments of related traits.

2.10 Joint fine-mapping GWAS-eQTL variants in shared risk loci

In addition to modeling related GWAS traits, our RiVIERA-MT is sufficiently versatile that enables us to also jointly infer the causal variants co-associated within the same risk locus characterized as both the GWAS locus and expression quantitative trait locus (eQTL) for a specific target gene. In contrast to jointly modeling related GWAS traits with distinct risk loci, in this application we assumed that the risk variants associated with the GWAS trait affect the expression of *in-cis* genes, which serve as an intermediate molecular phenotypes, and that the co-morbidity also manifests in the co-enrichments of shared epigenomic co-enrichments for both the observed and the molecular phenotypes. To this end, we applied RiVIERA-MT to jointly fine-map variants in the GWAS loci and the eQTLs using summary statistics from both GWAS consortia and GTEx consortium (version 6) [9] (**Methods**). In general, we found that the most significant GWAS SNPs and eQTL SNPs based on genetic or our model posterior predictions are not the same SNPs.

For instance, we found that a schizophrenia risk locus harbors rs1894401 as the GWAS variant but rs138682554 as the eQTL variant for gene *FES* (protein-tyrosine kinase) in GTEx whole blood samples. FES is previously shown to accelerate nerve growth factor (NGF)-induced neurite outgrowth [39]. The joint product of the GWAS and eQTL posteriors gave rise to a SNP (rs1894400) that differs from both GWAS and eQTL top SNP. On the other hand, all of the top three variants are centered around the weighted H3K9ac peak in CNS and H3K27ac in immune cell types based on ENCODE/Roadmap data. The general discordance between GWAS SNPs and eQTL SNPs may be attributable to several reasons: (1) due to evolution constraints the disease- or trait-altering SNPs are often under-represented in the natural population that GTEx data were derived from; (2) the relevant tissue context for the disease may be absent; (3) small sample size of GTEx data set (a few hundreds in contrast to tens of thousands in GWAS). Nonetheless, our approach discovered several promising variants where the GWAS and eQTL SNPs coincide.

For instance, based on the RiVIERA-MT inference, we discovered that a potentially regulatory variant rs10903129 is strongly associated with LDL trait and exhibits regulatory potential on affecting the expression of *LDLRAP1* (Low Density Lipoprotein Receptor Adaptor Protein 1) in GTEx liver tissue (**Fig.** 9). Specifically, the SNP is located in the intronic region of another gene (*TMEM57*) and ∼100 kb upstream of the transcription start site of LDLRAP1. We observe strong genetic signals from the GWAS side ($p<5E-8$) but weak signals on the eQTL side ($p<0.05$) (**Fig.** 9a top two tracks). In both cases, the strong LD

makes it difficult to distinguish the causal SNP(s). On the other hand, the epigenomic prior provides additional information on likely functional SNPs based on their locations around the weighted enhancer peaks (**Fig.** 9 bottom four tracks) as well as the implicit epigenomic correlation between the LDL phenotypes and LDLRAP1 expression across epigenomic weights. As a result, the posteriors on both GWAS and GTEx SNPs takes into account the likelihood of the genetic signals and the prior information to nominate rs10903129 as the most prominent variant. Taking the joint product of the two posteriors further improves the fine-mapping resolution.

As another interesting example, our RiVIERA-MT model confers the highest GWAS posterior for a missense mutation rs1042034 for APOB (apolipoprotein B) in HDL trait. This is consistent with the expectation that protein altering mutations usually have the highest impacts on the trait. It is also remarkable because we did not impose higher weights on the coding annotation when training our RiVIERA-MT model. Intriguingly, adjacent to the coding mutation, we also discovered an intergenic SNP rs934197, which exhibits the highest genetic signal in GTEx digestive tract tissue (Esophagus gastroesophageal junction) and modestly significant genetic signals in GWAS ($p<1E-5$). The variant is not only located in the upstream promoter region of the gene but also within the H3K4me3 and H3K9ac peaks specifically in the liver tissue from Roadmap Epigenomic data and therefore it is likely to be a functional SNP despite being in vicinity of the coding mutation. Together, we have made several compelling cases where our proposed data-driven approach led to interesting actionable hypotheses on causal SNPs.

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 [11, 20, 30, 40]. In formulating an efficient fine-mapping strategies, it is natural to incorporate various genome-wide reference annotations in a principled way as a form of Bayesian prior to infer the functional variants that drive the genetic signals of GWAS [11, 27, 41–43]. In this article, we describe a 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 requires 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 (e.g., human height [44]). To detect causal annotations, a general enrichment test should be performed either on *genome-wide* independent loci such as fgwas [26] 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 [16]. In this regards, 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, we may be able to infer the causal genes beyond individual variants by jointly modeling both the GWAS data and eQTL data taking into account not only SNP-level annotations but also gene-level network information such as pathways or protein-protein interactions. Second, we can generalize the model to applying for genome wide SNPs rather than defined risk loci via hierarchical modeling approach to infer risk loci and then the causal variants [26, 45]. 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 and the enrichment sampling processes via a multi-processing computing architecture.

Together, we believe that the success of RiVIERA-MT is an excellent example of integrating functional genomics data from ENCODE and Roadmap epigenomic consortia, population genetic data in terms of LD from 1000 Genome, tissue-specific genotype-expression information from GTEx, and genetic signals from large GWAS consortia in seeking for specific disease mechanisms (**Fig.** 1a), hereby going from an agnostic data-driven approach to a refined set of specific biological hypotheses that are experimentally actionable items in dissecting functional GWAS.

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 [24, 46]:

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

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 [24,25], we can integrate out β by taking the conditional expectation of the mean and variance of $\mathbf{y}$ with respect to β ($\mathbb{E}_\beta$) and leveraging the (log-transformed) linear property of the multivariate Gaussian density function:

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

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 disequilibrium (LD) and estimated either from the corresponding study cohort or a reference population from 1000 Genome Consortium [8].

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:

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

where $\mathbf{z}_c$ and Σ_{cc} denote z-scores and LD for the causal SNPs, respectively. Notably, Eq (14) exploits the property of conditional Gaussian. Consequently, 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 plausible configurations from the neighborhoods of the current configuration 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 [25, 36]. This stochastic search technique was initially developed by [36] as general feature selection algorithm, and was first implemented to fine-map causal variants in the software called FINEMAP [25]. 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 infers 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 [36].

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 [25]. 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 the above BF formulation, $\sigma^2_{a,d}$ is a free hyperparameter. Existing fine-mapping methods such as FINEMAP [25] and CAVIARBF [24] set it to a fixed user-defined value. Here, as first proposed by Guan and Stephen (2011) [31], we associate the additive per-SNP variance $\sigma^2_{a,d}$ 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). Our main goal here is to fine-map causal variants rather than estimating heritability. Rather than estimating the underlying heritability, the main purpose of this step is to approximate the integral of per-SNP variance that is required in the likelihood. Nonetheless, we can assess how well the approximation is by simulating GWAS datasets with heritability values ranging from 0.05 to 0.95. Overall, we observed a consistent increase of the average estimates as we increased the underlying heritabilities (**Supplementary Fig.** S8). This is remarkable compared to the existing fine-mapping methods such as CAVIARBF [24] and FINEMAP [25], 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.

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

We use Gibbs sampling [47] 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 [48]:

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

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) [49, 50], exploiting the fact that the joint posterior of our model is differentiable with respect to the model parameters $\mathbf{w}$ and $\mathbf{b}$ (**Supplementary Text** 1).

4.1.7 Functional enrichment

To assess functional enrichment of a given annotation, we can perform a likelihood ratio tests, which gives Bayes factor as the statistics. 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 ($\mathcal{L}_1^{(t)}$):

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

$$\mathcal{L}_1^{(t)} = \sum_l \log \sum_{\mathbf{c}_{ld}} p(\mathbf{z}_{ld}|\mathbf{c}_{ld}, \Sigma_{ld}, \mathbf{a}_{ldk}, \Theta_1^{(t)}) p(\mathbf{c}_{ld}|\mathbf{a}_{kd}, \Theta_1^{(t)}) d\Theta_0 \tag{46}$$

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

where $\chi^2(1)$ is chi-squared distribution with one degree of freedom that we use to assess the significance of the annotation. We approximate the integral by MCMC outlined above. 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.

In the real application for a large set of annotations, likelihood-ratio tests on each annotation is inefficient. Instead, we aggregated the enrichment parameter weights of each GWAS trait over annotations belonging to each of the 10 tissue groups [16]. Because we allowed negative and positive enrichment weights for each individual annotation, we expected that the aggregated weights centered at zero. Indeed, we observed that our enrichment scores approximately followed a zero-mean Gaussian distribution (**Supplementary Fig.** S9). Accordingly, we tested for significance of enrichment simply using a standard normal distribution as the null distribution.

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 [27]. 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 [51];

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. (multiple related traits), correlate the simulated fold-enrichments between traits with a covariance matrix with non-zero off-diagonal values;

4. 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);

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

4.3 GWAS summary statistics preprocessing

Overall, the summary statistics of the 28 GWAS traits were downloaded from public domains at the GWAS consortia website. For each study, we first removed strand-ambiguous SNP (T/A, C/G) and SNPs with supporting sample sizes lower than a threshold (10,000 individuals). We then imputed summary statistics using ImpG (v1.0.1) (`https://github.com/huwenboshi/ImpG`) [52] 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 GWAS SNPs ($p < 5 \times 10^{-8}$) reported by each study and the (observed/imputed) SNPs within 100 kb genomic distance of the lead SNPs. Overlapping loci were merged. Loci overlapping with the major histocompatibility complex (MHC) region (chr6:28477797-33448354; hg19) were removed from further analyses. The summary statistics of these non-overlapping risk loci were then input to our proposed fine-mapping algorithm. **Supplementary Table**S1 summarizes the data from each individual GWAS study.

4.4 Running existing fine-mapping software on simulated data

The software `fgwas` [26] (version 0.3.4) were downloaded from GitHub. To enable fine-mapping, we issued `-fine` flag and specified the region numbers for each SNP in the input file as required by the software. PAINTOR (version 2.1) was downloaded from GitHub [27]. 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. GPA (0.9-3) [28] 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.

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 [16]. 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 Functional annotations

We obtained the same set of 272 functional annotations as used in the LD score regression (`https://github.com/bulik/ldsc`) as described above.

4.7 Joint inference GWAS-eQTL SNPs

We obtained GTEx summary statistics over 44 tissues from GTEx portal (`gtexportal.org`). To fine-map *in-cis* regulatory SNPs for each gene in each tissue, we selected SNPs 1 Mb up and downstream of each gene. Although our RiVIERA-MT is efficient compared to the state-of-the-art methods, directly fine-mapping the entire 2 Mb locus harboring on average $\sim$8000 common variants is still prohibitive due to the combinatorially large search space of the causal configurations. Here we developed an heuristic sliding window approach. Specifically, we slide a 50 kb non-overlapping window with 25 kb buffer on both sides to account for extended LD spanning across 1 Mb up and downstream of each gene, resulting 40 50-kb windows. We then treat these 40 windows as the risk loci to perform fine-mapping jointly with the overlapping risk loci of a GWAS trait (i.e., including only the GWAS locus or loci defined above that overlap with the eQTL). We then took the posterior for only the SNPs within each 50 kb window (i.e., discarding the overalpped SNP in the 25-kb buffer regions) as well as the GWAS posteriors. The final GWAS-eQTL posterior is simply the joint product of the eQTL posterior and the GWAS posterior for the same SNP.

4.8 Code availability

RiVIERA-MT software (version 0.9.3) implemented as a standalone R package is freely available from Github repository `https://yueli-compbio.github.io/RiVIERA-MT`.

5 Acknowledgements

We thank Alkes Price, Hilary Finucane, and Steven Gazal for providing the full summary statistics of the 27 GWAS traits.

Figure Legends

Figure 1: Integrative approach to infer disease-causing mutations. (**a**) Computational challenge to integrate multiple sources of evidence to study disease mechanisms. These sources may include (but are definitely not limited to): epigenomic reference annotations from Roadmap/ENCODE Epigenomic consortium, LD information from 1000 Genome project, eQTL information from GTEx consortium, and GWAS from various large human disease consortia. (**b**) Fine-mapping causal variants. To fine-map causal variants, we take into account three lines of evidence (from top to bottom) namely (1) linkage disequilibrium (as SNP-by-SNP Pearson correlation matrix), (2) genetic signals as z-scores displayed here in terms of GWAS -log10 p-values, (3) epigenomic activities overlapping SNPs and infer the posterior inclusion probabilities (PIP) of each SNP for being causal within the risk locus. The causal variants are the red diamond and highlighted by the vertical line to aid visualization. Our model is able to infer PIP accurately for the cases where the risk locus harbors a single causal variant as well as the cases where multiple causal variants co-localized within the same locus.

Figure 2: Proposed fine-mapping model RiVIERA-MT. (**a**). We express RiVIERA-MT as probabilistic graphical model (PGM), where shaded nodes are observed data including epigenomic annotations, GWAS summary statistics, and linkage disequilibrium (highlighted in the corresponding color boxes). The unshaded nodes are the latent variables or model parameters of interest. 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 and explained in details in the main text and **Methods** section. The key latent variable is the causal indicator vector c_{ld}, which denote the causal configuration in locus l and disease d. We infer its posterior distribution by taking into account the epigenomic prior and GWAS z-score likelihood. The prior is a logistic function of the linear combination of the epigenomic annotations weighted by the inferred enrichment parameters $\mathbf{w}_k$ for each annotation k, which follows a multivariate Gaussian distribution with fixed mean (default: 0) and the covariance modeling the co-enrichment across D diseases. (**b**). To perform inference on the RiVIERA-MT model, we perform Markov Chain Monte Carlo (MCMC), which can be divided into 3 main phases. We first stochastically sample causal configurations from the posterior. The sampling scheme was adapted from [25, 36]. 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: (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. In the second phase, we marginalize the posteriors over all of the sample configurations to obtain posterior inclusion probabilities (PIP) for each SNP. In the third phase, we sample enrichment parameters (w_k, Σ_w) and variance explained per SNP $(\sigma_{a,d})$. We alternative among the three phases until model converges. 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 6 SNP prioritization methods. The boxplots are based on 500 independent simulations. (**a**) Overall performance of each method based on simulated data where each locus contain 1 or more variants. (**b**) Model performance on cases where each locus contains at most 3 causal variants, which is the same assumption as the PAINTOR model.

Figure 4: Inferring variants across multiple traits. (**a**) We ran RiVIERA-MT single-trait and multi-trait modes on the simulated data containing 2 traits that do not have the same risk loci but related via the correlation of functional enrichments. Same as above, the evaluation is based on the proportion of causal variants identified (y-axis) as a function of increasing number of variants selected (x-axis). We compared multi-trait model with the single-trait model in terms of the posteriors and epigenomic priors. (**b**) As negative control, we applied the two models to data where the two traits are unrelated (i.e., zero epigenomic correlation).

Figure 5: Functional enrichments of GWAS traits. (**a**) List of GWAS traits and their relevant cell types, most of them are implicated in the functional enrichment analyses. (**b**) Comparison of functional enrichments using LD score (LDSC) regression [16] and RiVIERA-MT. For LDSC, the color intensities are the z-score of the transformed co-effiicents output from the software. For RiVIERA-MT, intensities are the aggregated expected enrichment weights estimated from the Bayesian model. "*" and "**" indicate p-values < 0.1 and < 0.01, respectively.

Figure 6: Fine-mapping risk loci of neurologic/psychiatric and immune disorders. A risk locus for multiple sclerosis. The four tracks (top to bottom) GWAS -logP values, empirical prior and posterior from RiVIERA-MT, and weighted H3K27ac peaks by the enrichment parameters that are also learned by RiVIERA-MT. Top SNPs nominated by the posterior are highlighted in red. Due to strong LD, it is difficult to identify the causal SNPs. On the other hand, the empirical priors of SNPs are more prominent around the enhancer peaks. By leveraging the enhancer enrichments as epigenomic prior and GWAS genetic signals as likelihood (taking into account LD), RiVIERA-MT nominates only a handful of SNPs as the putative causal SNPs.

Figure 7: Joint fine-mapping related GWAS traits. (**a**) A risk locus for low density lipoprotein (LDL) trait and (**b**) Schizophrenia (SCZ). In each panel, the track are annotated on the left. The main purpose is to illustrate the difference of inference results between single- and multi-trait models. Top variants selected by GWAS -logP, single-trait model, multi-trait model are in black (square), blue (triangle), red (diamond) colors (shape) respectively. The remaining SNPs are in grey circle. The bottom track display the weighted H3K27ac activities in the loci.

Figure 8: Conservation of prioritized SNPs. For each of the 27 GWAS traits, we obtained SNPs with GWAS $p < 5E\text{-}8$ as well as SNPs included within the 90% credible set by the single-trait model without annotation (PIP A- Single-trait) and with annotation (PIP A+ Single-trait) and the multi-trait models (PIP A+ Multi-trait). The barplots display the averaged conservation score among the selected SNP sets in each GWAS. For clarity, we further separated the GWAS traits by whether we have modeled them together with other related traits in the multi-trait mode.

Figure 9: Joint fine-mapping GWAS-eQTL variants. (**a**) Risk locus for low density lipoprotein (LDL) GWAS trait and eQTL for LDLRAP1 gene in GTEx liver tissue. The 11 tracks (top to bottom; also indicated on the strip at the right of each track) are GWAS -logP, GTEx/eQTL -logP, GWAS prior, GTEx/eQTL prior, GWAS posterior, GTEx/eQTL posterior, joint (product of GWAS and GTEx posteriors for the same SNP), peaks of four epigenomic marks. Color legend indicates the 10 tissue groups. The top GWAS-eQTL co-associated SNP is highlighted in red. In this case, the GWAS and eQTL analyses agree on the most significant SNP namely rs10903129. (**b**) Risk locus for high density lipoprotein (HDL) GWAS trait and eQTL for APOB gene in GTEx digestive tract tissue (esophagus gastroesophageal junction tissue). Two variants are highlighted: the variant in red (rs1042034) is a missense mutation and the most significant SNP in the GWAS and the joint analysis; the variant in blue (rs934197) is an intergenic mutation, the most significant SNP in GTEx samples, and also exhibit strong regulatory potential since it is located in the H3K4me3 and H3K9ac peaks in the liver samples from Roadmap Epigenomic database.

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

9. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015). URL `http://www.sciencemag.org/content/348/6235/648.full`.

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

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

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

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

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

15. 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).

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

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

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

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

20. 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).

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. 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).

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

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

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

26. 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).

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

28. 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).

29. Li, Y. & Kellis, M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research* (2016).

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. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics (Oxford, England)* btw615 (2016).

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

37. Gaulton, K. J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature Genetics* **47**, 1415–1425 (2015).

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

39. Laurent, C. E. & Smithgall, T. E. The c-fes tyrosine kinase cooperates with the breakpoint cluster region protein (bcr) to induce neurite extension in a rac-and cdc42-dependent manner. *Experimental cell research* **299**, 188–198 (2004).

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

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

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

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

44. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569 (2010).

45. 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).

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

47. 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).

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

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

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

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

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

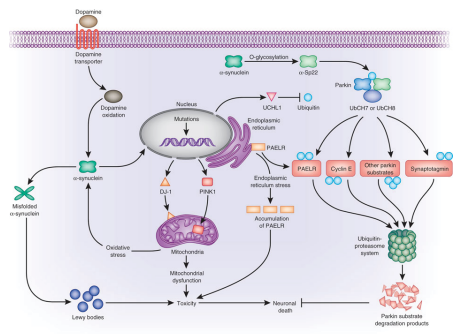
Figure 1

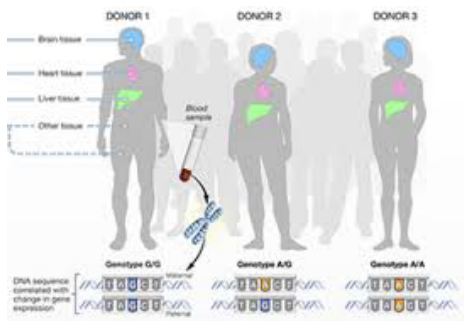
a

Epigenomics Roadmap/ ENCODE

1000 Genome Population genetics/LD

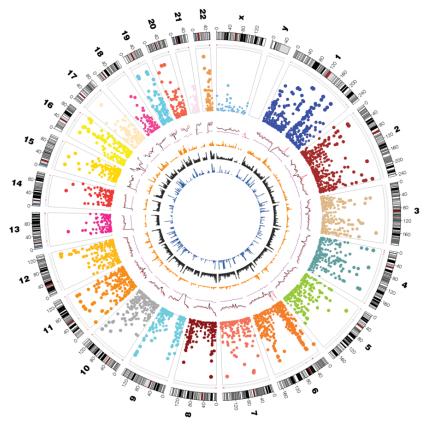
Disease mechanism?





eQTLs/GTEx genotype-expr

GWAS/RVAS Association



b

3 lines of evidence

Linkage Disequilibrium

Genetic evidence

Epigenomic annotations

Our model Inference

Single causal SNP per locus

Multiple causal SNPs per locus

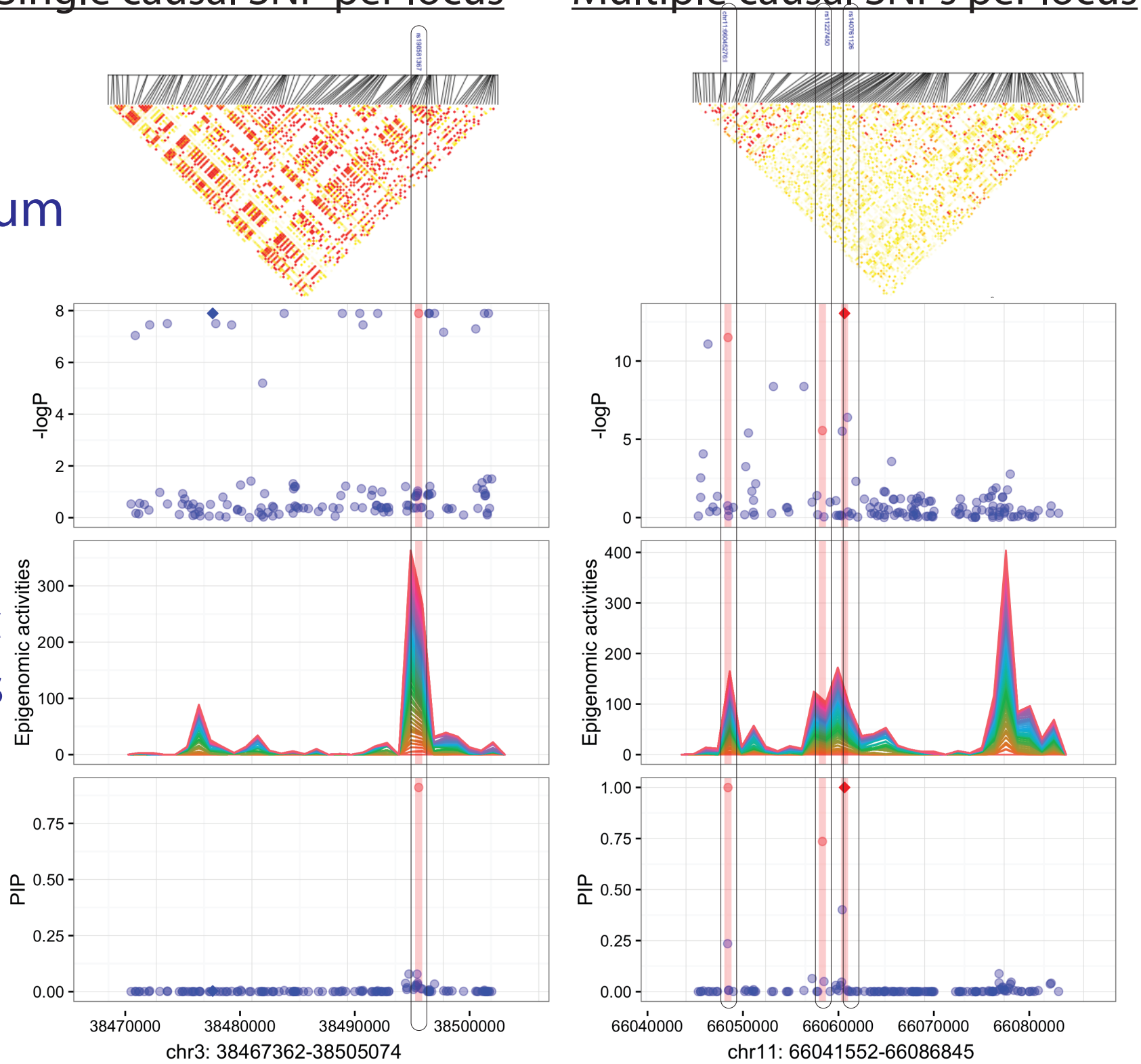
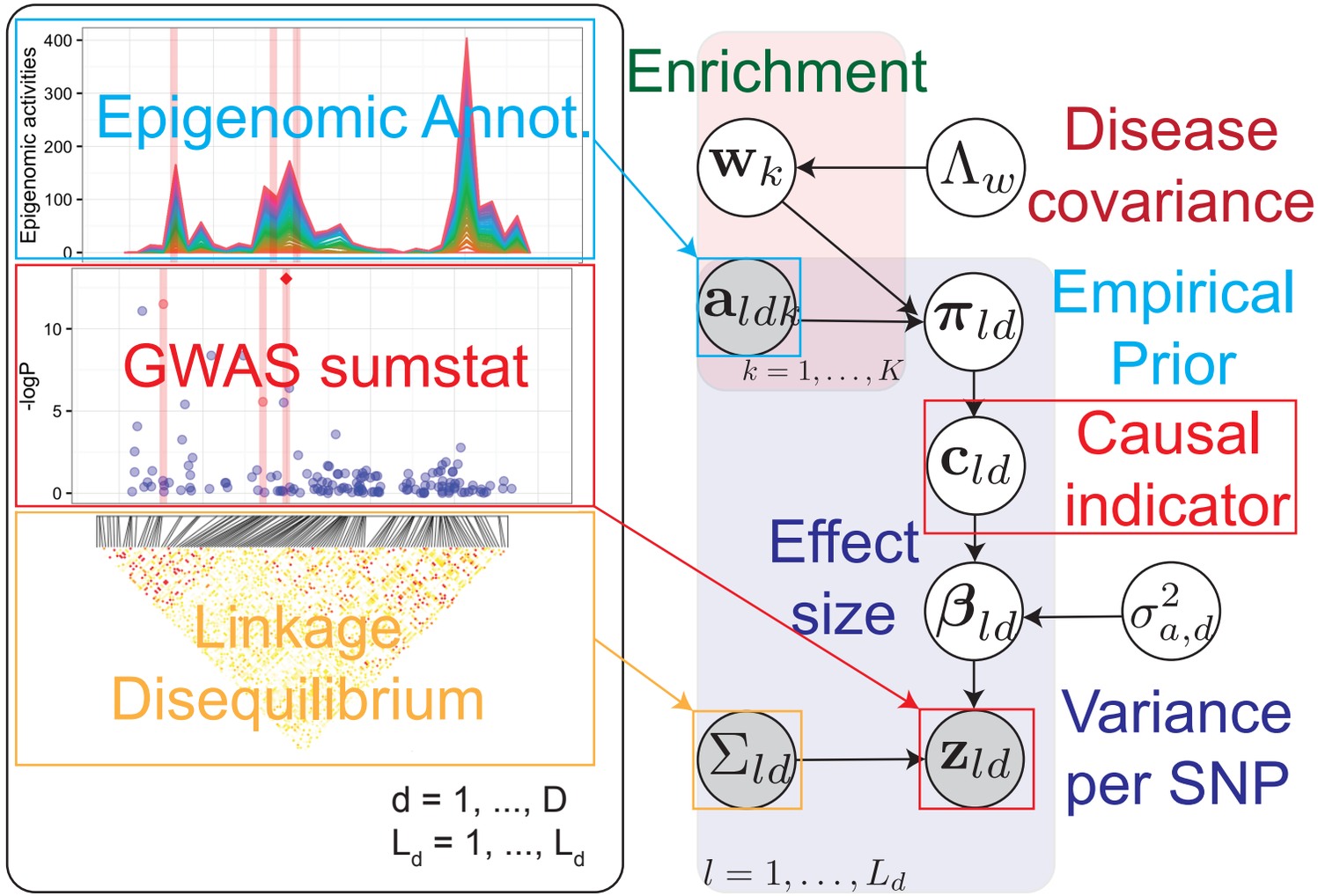


Figure 2

a. RiVIERA-MT model



b. Learning algorithm

Which combination of variants is causal (binary vector = causal config.)?

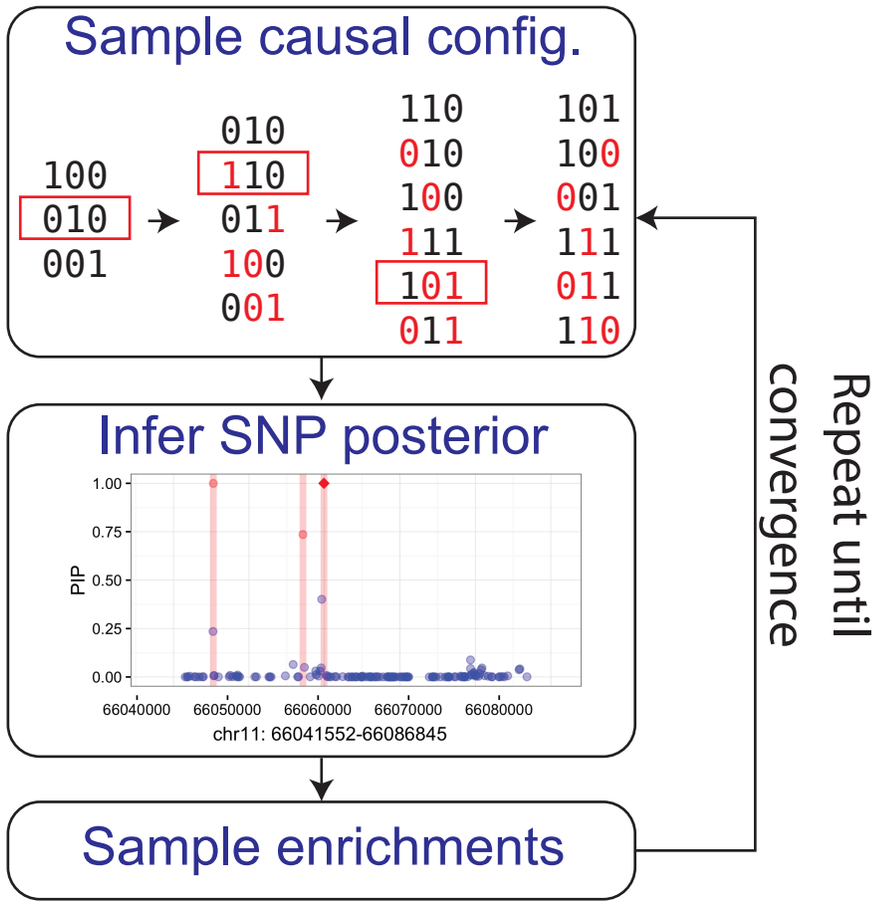
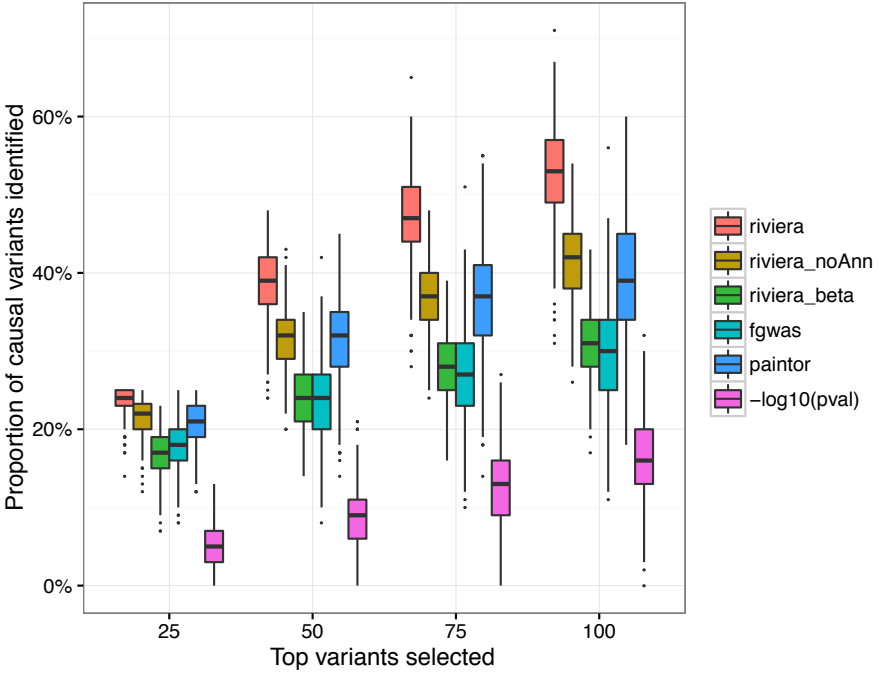


Figure 3

a. overall performance



b. maximum of 3 causal variants

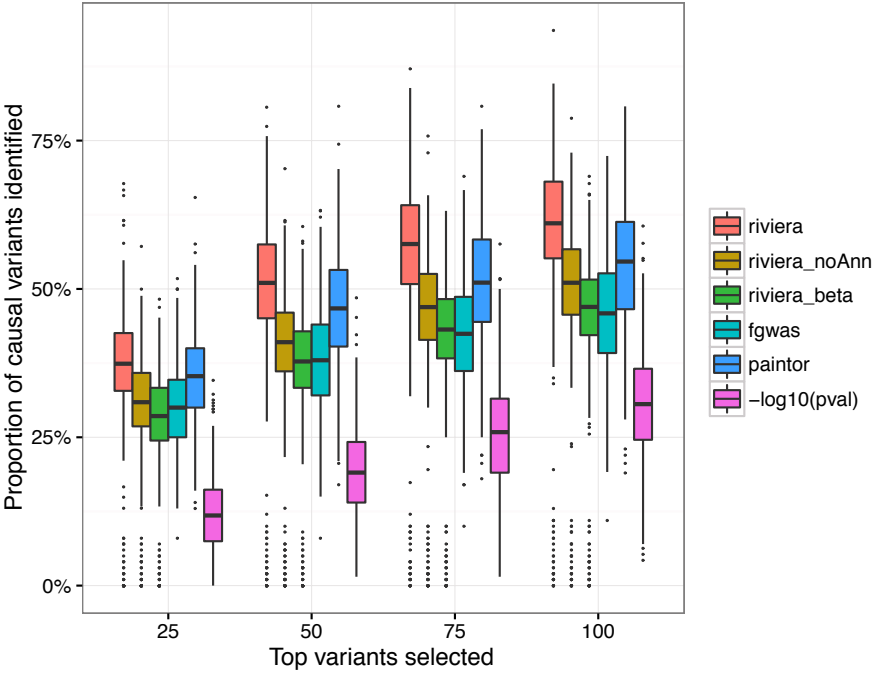


Figure 4

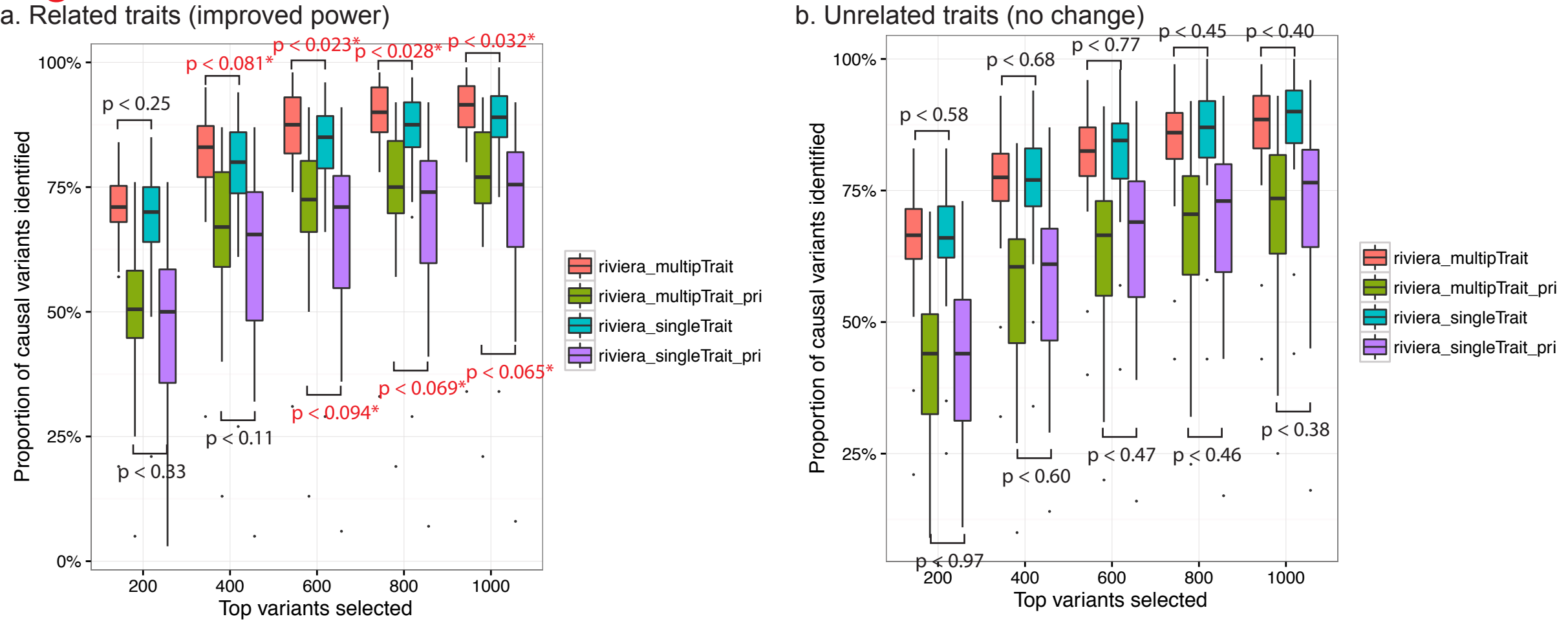

a. Related traits (improved power)
Proportion of causal variants identified
100%
75%
50%
25%
0%
p < 0.25
p < 0.33
p < 0.081*
p < 0.11
p < 0.023*
p < 0.094*
p < 0.028*
p < 0.069*
p < 0.032*
p < 0.065*
200
400
600
800
1000
Top variants selected
riviera_multipTrait
riviera_multipTrait_pri
riviera_singleTrait
riviera_singleTrait_pri
b. Unrelated traits (no change)
Proportion of causal variants identified
100%
75%
50%
25%
p < 0.58
p < 0.97
p < 0.68
p < 0.60
p < 0.77
p < 0.47
p < 0.45
p < 0.46
p < 0.40
p < 0.38
200
400
600
800
1000
Top variants selected
riviera_multipTrait
riviera_multipTrait_pri
riviera_singleTrait
riviera_singleTrait_pri

Figure 5

a

Full sumstat of 26 GWAS traits		Expected enriched tissues	
Anthropometric	Height	ESC	Embryonic stem cell
Immune traits	Celiac disease	Immune	Immune cell types
	Crohn's disease	Immune	
	Ulcerative colitis	Immune	
	Systemic lupus erythematosus	Immune	
	Multiple sclerosis	Immune	
	Primary biliary cirrhosis	Immune	
	Rheumatoid arthritis	Immune	
	Type 1 diabetes	Immune	
Neurologic/ Psychiatric traits	Bipolar disorder	Brain	Brain tissues
	Depression symptom	Brain	
	Education year	Brain	
	Ever smoked	Brain	
	Neuroticism	Brain	
	Mean putamen	Brain	
	Schizophrenia	Brain	
	Subject well being	Brain	
Anthropometric	Body mass index	Brain/Fat	Brain/Adipose
Heart disease	Coronary artery disease	Heart	Heart tissues
Metabolic traits	Type 2 diabetes	P.islets	Digestic tissues
	Fasting glucose-related traits	P.islets	
	HbA1c	P.islets	
	LDL cholesterol	Liver	
	Triglycerides	Liver	
	HDL cholesterol	Liver	
Psychiatric	Autism	Mixed	Mixed/Other

b

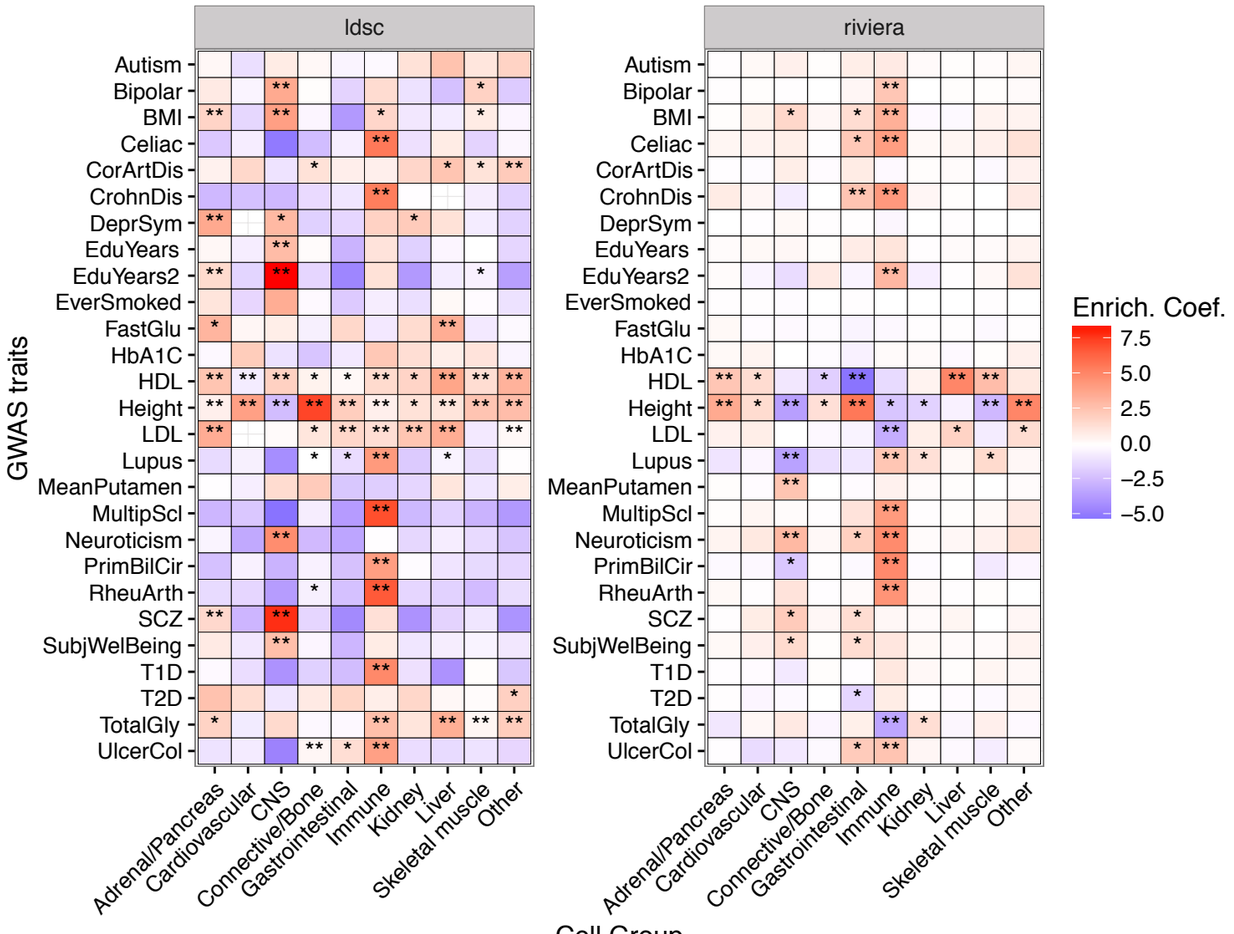


Figure 6

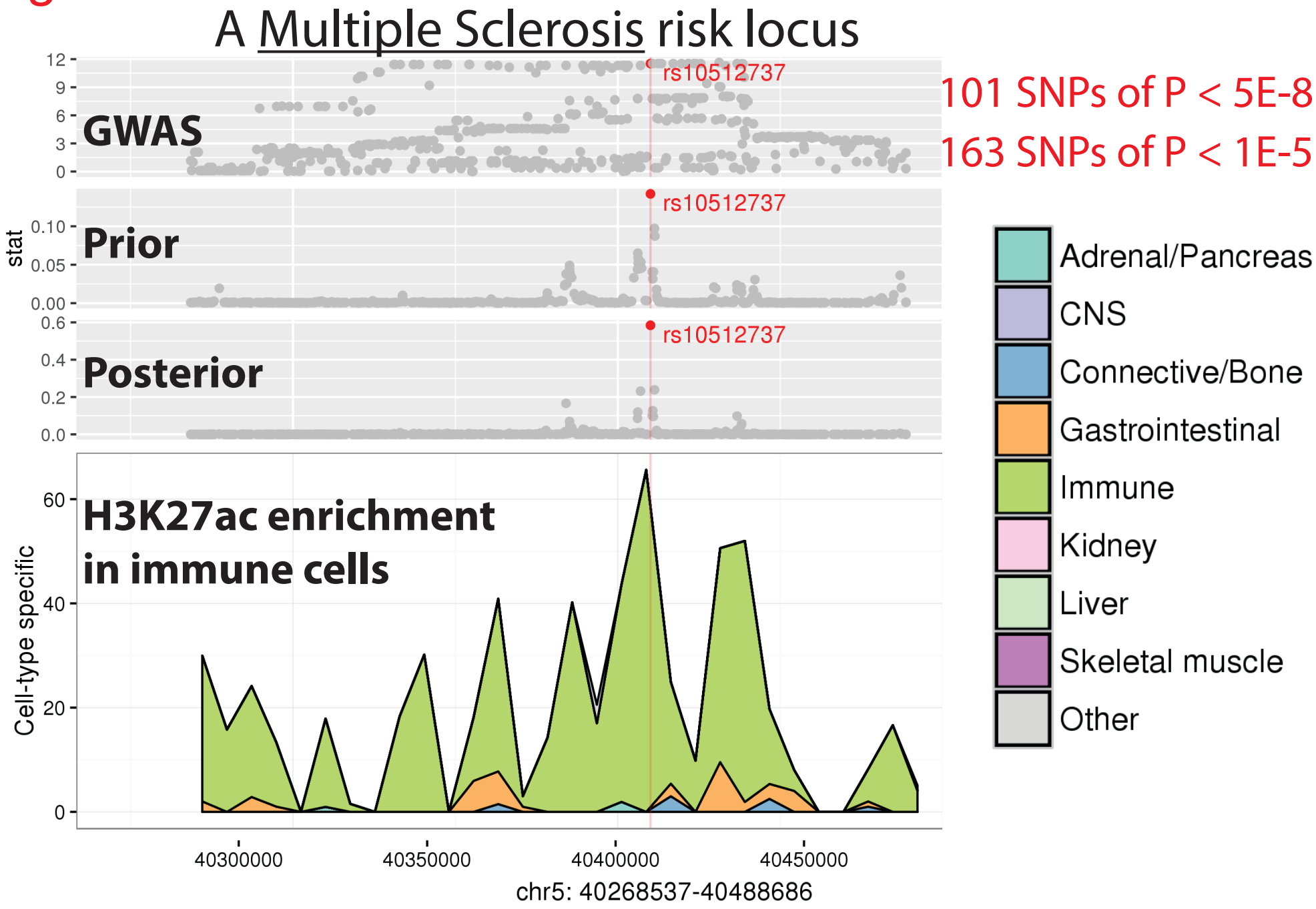

A Multiple Sclerosis risk locus
GWAS
Prior
Posterior
stat
rs10512737
101 SNPs of P < 5E-8
163 SNPs of P < 1E-5
H3K27ac enrichment
in immune cells
Cell-type specific
Adrenal/Pancreas
CNS
Connective/Bone
Gastrointestinal
Immune
Kidney
Liver
Skeletal muscle
Other
40300000
40350000
40400000
40450000
chr5: 40268537-40488686

Figure 7

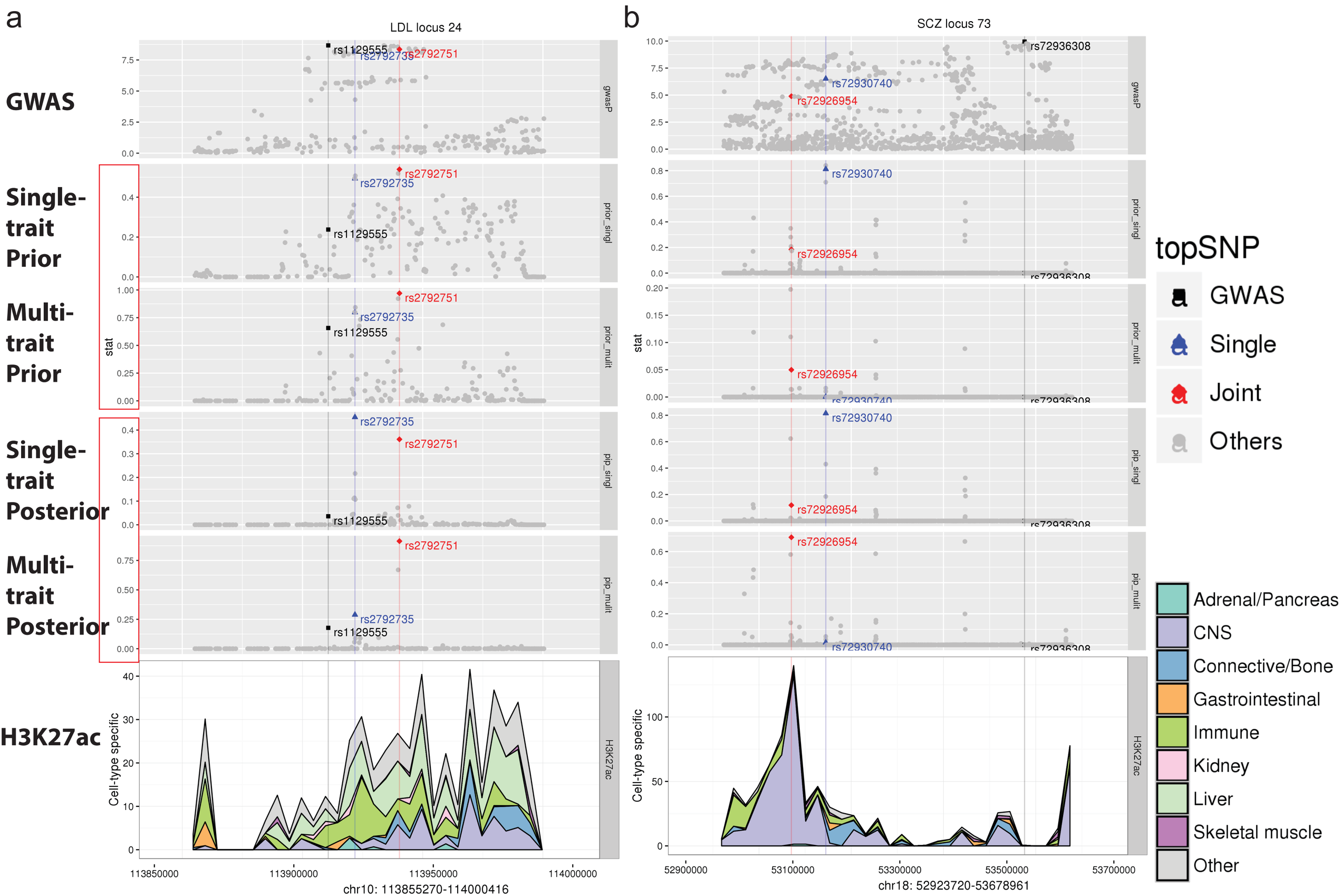

a
LDL locus 24
GWAS
Single-trait Prior
Multi-trait Prior
Single-trait Posterior
Multi-trait Posterior
H3K27ac
rs1129555
rs2792735
rs2792751
stat
gwasP
prior_singl
prior_mulit
pip_singl
pip_mulit
H3K27ac
Cell-type specific
chr10: 113855270-114000416
b
SCZ locus 73
rs72936308
rs72930740
rs72926954
chr18: 52923720-53678961
topSNP
GWAS
Single
Joint
Others
Adrenal/Pancreas
CNS
Connective/Bone
Gastrointestinal
Immune
Kidney
Liver
Skeletal muscle
Other

Figure 8

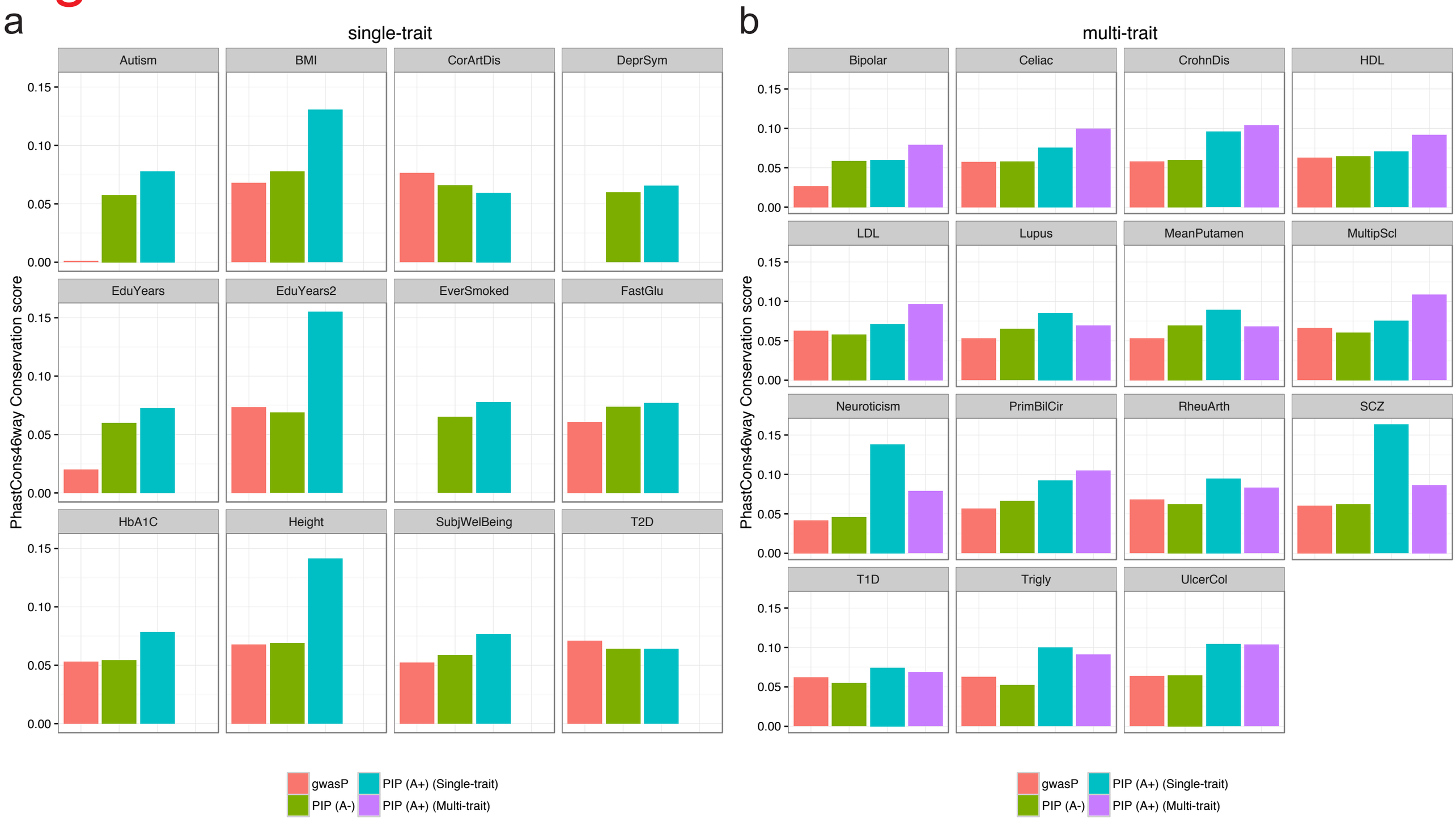
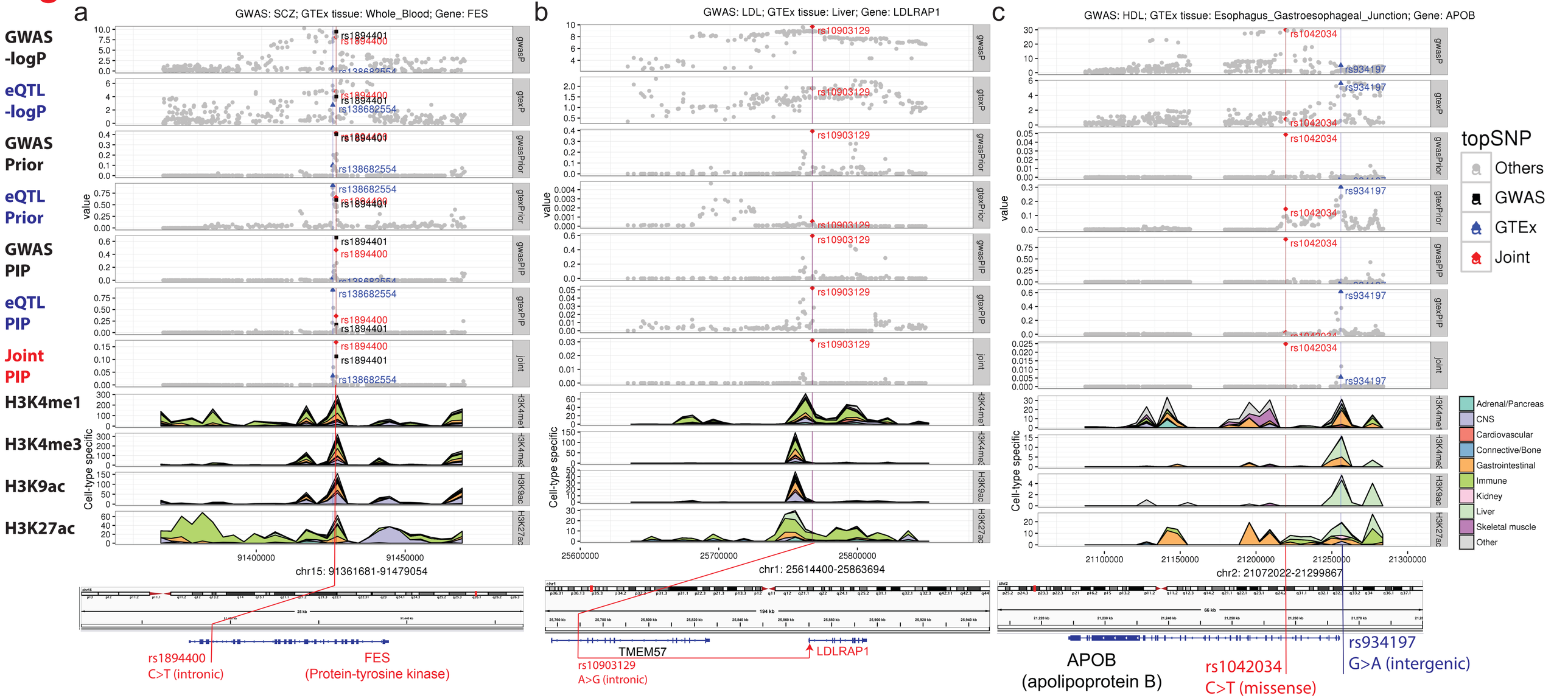

a
single-trait
Autism
BMI
CorArtDis
DeprSym
EduYears
EduYears2
EverSmoked
FastGlu
HbA1C
Height
SubjWelBeing
T2D
PhastCons46way Conservation score
0.15
0.10
0.05
0.00
gwasP
PIP (A-)
PIP (A+) (Single-trait)
PIP (A+) (Multi-trait)
b
multi-trait
Bipolar
Celiac
CrohnDis
HDL
LDL
Lupus
MeanPutamen
MultipScl
Neuroticism
PrimBilCir
RheuArth
SCZ
T1D
Trigly
UlcerCol
PhastCons46way Conservation score
0.15
0.10
0.05
0.00
gwasP
PIP (A-)
PIP (A+) (Single-trait)
PIP (A+) (Multi-trait)

Figure 9


a
GWAS: SCZ; GTEx tissue: Whole_Blood; Gene: FES
b
GWAS: LDL; GTEx tissue: Liver; Gene: LDLRAP1
c
GWAS: HDL; GTEx tissue: Esophagus_Gastroesophageal_Junction; Gene: APOB
GWAS -logP
eQTL -logP
GWAS Prior
eQTL Prior
GWAS PIP
eQTL PIP
Joint PIP
H3K4me1
H3K4me3
H3K9ac
H3K27ac
value
Cell-type specific
rs1894401
rs1894400
rs138682554
rs10903129
rs1042034
rs934197
topSNP
Others
GWAS
GTEx
Joint
Adrenal/Pancreas
CNS
Cardiovascular
Connective/Bone
Gastrointestinal
Immune
Kidney
Liver
Skeletal muscle
Other
chr15: 91361681-91479054
chr1: 25614400-25863694
chr2: 21072022-21299867
rs1894400
C>T (intronic)
FES
(Protein-tyrosine kinase)
TMEM57
LDLRAP1
rs10903129
A>G (intronic)
APOB
(apolipoprotein B)
rs1042034
C>T (missense)
rs934197
G>A (intergenic)

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

Contents

1 HMC method details **2**

2 RiVIERA-MT algorithm outlines **3**

3 GWAS simulation **3**
3.1 Simulating genotypes . . . 3
3.2 Simulating epigenomic enrichments for single trait . . . 4
3.3 Simulating epigenomic enrichments for multiple related traits . . . 5
3.4 Simulating phenotypes and GWAS summary statistics . . . 6

4 Supplementary Fig. **7**

5 Supplementary Table **20**

List of Figures

S2 Full model schematic diagram . . . 8
S3 Simulation pipeline. . . . 9
S4 Model convergence . . . 10
S5 Power comparison . . . 11
S6 Calibration of posterior inclusion probabilities . . . 12
S7 Functional enrichment analysis . . . 13

S8 Estimation of per-SNP variance explained 14
S9 RiVIERA-MT enrichment score of GWAS traits over 10 tissue groups 15
S10 Tissue-specific enrichment scores comparison between RiVIERA-MT and LDSC 15
S11 Fine-mapped risk loci illustrations 17
S12 Overlap percentage of prioritized SNP sets 18
S13 Overlap between multi-trait and single-trait inferred 90% credible sets 19

List of Tables

S1 GWAS data summary 20
S2 Fine-mapping results summary 21

1 HMC method details

To sample epigenomic effects w_{kd} and causal bias b_{ld}, we employ Hamiltonian Monte Carlo (also known as hybrid Monte Carlo) (HMC) [1,2]. HMC is one of the most powerful MCMC methods [1–3]. By extending the state space to include auxiliary momentum variables, and then using Hamiltonian dynamics to traverse long probability contours in the extended state space, HMC is able to move long distances in state space in a single update step and thus largely bypasses the slow exploration of the state space that occurs when Metropolis updates are done using a simple random-walk proposal distribution [3]. However, the essential elements for HMC are partial derivatives or gradients of the potential energy function $U = -\log p(\Theta|\mathcal{D})$ with respect to the variable of interests, which are not always tractable.

Sampling model parameters q by HMC can be intuitively visualized as a particle that slides over a frictionless surface [3]. When climbing on a rising slope, the particles momentum allows it to continue to explore the parameter space with decreasing kinetic energy and increasing potential energy (i.e., decreasing posterior), until the kinetic energy reaches zero, at which point it will slide back down with increasing kinetic energy and decreasing potential energy (i.e., increasing posterior). Thus, the momentum can be built up along the HMC trajectory to tolerate to some extent the deterioration of posterior probability upon hill climbing in order to seek for better parameter space. As a result, the HMC algorithm is able to explore much larger parameter space and much less prone to getting trapped within local suboptimal regions than simple Metropolis random walk.

In our case, solving HMC inference is feasible because $U = -\log p(\Theta|\mathcal{D})$ (i.e., joint posterior) is differentiable with respect to the model parameters w_{kd} and b_{ld}. With some simple linear algebra, we obtain the following gradients:

$$\frac{\partial U}{\partial w_{kd}} = \sum_{d'=1}^{D} w_{kd'}\lambda_{dd'} + \sum_{l=1}^{L_d}\sum_{i=1}^{m_d}(\pi_{ild} - c_{ild})a_{ildk} \tag{1}$$

$$\frac{\partial U}{\partial b_{ld}} = \sum_{i=1}^{m_d}(\pi_{id} - c_{id}) + \lambda_{ld}(b_{ld} - \log\frac{\pi_{ld}}{1-\pi_{ld}})) \tag{2}$$

where $\pi_{ld} = 1/m_{ld}$. The gradient calculations were verified by comparing them individually to the changes in potential energy $\Delta U(q) = \frac{U(q+\xi)-U(q-\xi)}{2\times\epsilon}$, where $\xi = 10^{-4}$ and $\frac{\partial U}{\partial q} \approx \Delta U(q)$ within a very small numerical discrepancy (e.g., 10^{-8}).

Given the gradients, we apply the standard HMC procedure [2,3] known as the "leapfrog" method [2, 3]. In a nutshell, the Hamiltonian dynamics is defined as

$$H(q, p) = U(q) + K(p) \tag{3}$$

where the potential energy $U(q) = -\log p(\Theta|\mathcal{D})$ (i.e., negative log posterior defined as above) and q denotes the position variable in a fictitious physical system. In our case, the position variable $q \in \Theta = \{\mathbf{w}, \mathbf{b}\}$. $K(p)$ is the kinetic energy, where the auxiliary variable p is drawn independently from $\mathcal{N}(0, 1)$ to model the momentum for each position variable q. As described in Algorithm 1, we simulate the Hamiltonian dynamics via the "leapfrog" method. The procedure starts with a half-step update of the momentum, followed by $L - 1$ full-step updates of both position and momentum variables, and ends with another half-step update of the momentum.

At the end of this trajectory, we apply a Metropolis update to accept the proposed Hamiltonian state at the probability $\min(1, \exp[-\Delta H])$, where $\Delta H = H^* - H^{cur}$. Thus, we will always accept the proposed state if the Hamiltonian dynamics is lowered or unchanged in contrast to the current one. Because of the discrete-time procedure, we must use a small step size ϵ to compensate for the numerical imprecision introduced in simulating the Hamiltonian system. Automated procedure adapting the step size ϵ and step number L to the Metropolis acceptance rate is an active research field [3,4]. A rule-of-thumb is to maintain the acceptance rate at around 65%. Here we set the step size ϵ to 0.01 and number of steps L to 100 based on some preliminary runs, where the acceptance rates are within the range of [50%, 90%] over 1000 MCMC samples.

2 RiVIERA-MT algorithm outlines

The pseudocode for RiVIERA is outlined in Algorithm 2.

3 GWAS simulation

3.1 Simulating genotypes

As the first step of the simulation, we generated 10000 genotypes (**Supplementary Fig.** S3). In particular, we use HapGen2 [5] to simulate haplotypes by conditioning on the 1000 Genome CEU population haplotypes (phase 1 version 3) so that the simulated data has the same LD patterns as the reference panel. However, HapGen2 is computationally demanding and can only simulate several thousands of SNPs for the chosen cohort size. To remedy that, we obtained the 1703 approximately independent LD blocks estimated from LDetect [6] and simulated 10,000 genotypes using HapGen2 together for each LD block separately. We only did this once for all of the simulation runs.

Algorithm 1 HMC procedure to simultaneously sample $\mathbf{w}, \mathbf{b}$

Given ϵ, L, and q, where $q \in \{\mathbf{w}, \mathbf{b}\}$
Set $q^{cur} \leftarrow q^0 \leftarrow q$
Sample $p^{cur} \sim \mathcal{N}(0, 1)$
Set $p \leftarrow p^{cur} - (\epsilon/2)\frac{\partial U}{\partial q}$
for $i = 1 \ldots L$ **do**
    Set $q^i \leftarrow q^{i-1} + \epsilon p^{i-1}$
    if $i \neq L$ **then**
        Set $p^i \leftarrow p^{i-1} - \epsilon\frac{\partial U}{\partial q^i}$
    else
        Set $p^i \leftarrow p^{i-1} - (\epsilon/2)\frac{\partial U}{\partial q^i}$
    end if
end for
Set $H^* \leftarrow U(q^L) + \frac{1}{2}\sum_m (p_m^L)^2$
Set $H^{cur} \leftarrow U(q^{cur}) + \frac{1}{2}\sum_m (p_m^{cur})^2$
if $Unif(0, 1) < \exp(-\{H^* - H^{cur}\})$ **then**
    Set $q \leftarrow q^L$
else
    Set $q \leftarrow q^{cur}$
end if

For each simulation run, we randomly sampled 100 LD blocks. For each sampled LD block, we then sampled 500 consecutive SNPs or used the whole block if it contains fewer than or equal to 500 SNPs, which resulted in on average ∼38 kb per locus. Thus, the maximum number of SNPs generated per simulation run is 500,000 SNPs. The sampling step was done mainly for computational reason in order to run many simulation runs in parallel. We then filtered out SNPs with allele frequency lower than 1%.

3.2 Simulating epigenomic enrichments for single trait

We chose 20 representative cell types from the 19 primary groups defined by Epigenome Roadmap Consortium [7]. Whenever available, for each of the 20 chosen tissues or cell types, we obtained epigenomic profiles for DNase and six histone marks namely H3K4me1, H3K4me3, H3K36me1, H3K27me3, H3K9me3, and H3K9ac (**Supplementary Fig.** S3). In total, we obtained 100 epigenomic annotations in BigWig format. For each epigenomic annotation, we then thresholded their values to be 1 for p-values < 0.01 and 0 otherwise.

The fold-enrichment of annotation k is then:

$$f_k = \frac{m_k/m_c}{a_k/m} \tag{4}$$

where

- m_k: number of causal variants harbored in annotation k;
- m_c: the number of causal variants, which we set to 100, giving on average 1 causal variant per locus for the 100 simulated loci per simulated dataset;

Algorithm 2 RiVIERA-MT algorithm

Input:

$\mathbf{z}$: a nested list of GWAS summary statistics matrices as z-scores $\mathbf{z}_{ld}$ for disease $d \in \{1 \ldots, D\}$ in locus $l \in \{1, \ldots, L_d\}$, which harbors m_{ld} SNPs;

$\mathbf{a}$: a nested list of annotation matrices $\mathbf{a}_{ldk}$ annotation k in locus l of disease d;

Σ: a nested list of LD matrices Σ_{ld} of locus l of disease d;

m_{ld}: number of SNPs within each locus l and in disease d;

Output: :

$\mathbf{c}$: estimated posterior inclusion probabilities (PIP) averaged across T' MCMC runs

1: Initialize model parameters $\mathbf{w} \sim \mathcal{N}(0, 1)$
2: **for** epoch = 1 To T **do**
3: Estimate PIP by MH-stochastic configurations search
4: Sampling genetic additive variance $\sigma^2_{a,d}$
5: Sampling model parameters Λ_w, λ (Gibbs sampling), $\mathbf{w}, \mathbf{b}$ (HMC sampling)
6: **end for**
7: Averaging sampled $\mathbf{c}_{ld}$ across T MCMC iterations (after discarding $x\%$ sampled values at the initial iterations (default: $x\% = 20\%$).

- a_k: total number of variants from the 100 loci harbored in annotation k;
- m: total number of variants from the 100 loci;

The maximum fold-enrichment for annotation k is then:

$$f_{\max,k} = \begin{cases} m/a_k, & \text{if } a_k > m_c \\ m/m_c, & \text{if } a_k < m_c \end{cases} \tag{5}$$

We then sampled from uniform distribution the fold-enrichment for each annotation within the corresponding range defined by (5) and subsequently solved for m_k to sample causal variants from that annotation:

$$f_k \sim U(1, f_{max,k}) \tag{6}$$

$$m_k = \frac{f_k a_k m_c}{m} \tag{7}$$

Due to the previously observed enrichments [7, 8], we started sampling causal variants from annotations corresponding to DNase, H3K4me1, and H3K4me3 and then the remaining annotations. The sampling continued until we reached or exceeded the defined number of causal variants (i.e., 100). We then randomly chose 100 causal variants from the candidate causal variants. Because of the overlapping between annotations and the above rounding, we recomputed the fold-enrichment for each annotation using Eq (4).

3.3 Simulating epigenomic enrichments for multiple related traits

To simulate enrichments of multiple related traits, we started by computing the maximum fold-enrichments for all of the D traits based on Eq (5). This gave a $K \times D$ maximum fold-enrichment matrix $\mathbf{f}_{\max}$. We then sampled from uniform distribution independent $\mathbf{f}_d$ and

subsequently correlated the fold-enrichment via an auxiliary pre-defined $D \times D$ covariance matrix Σ_f with off-diagonal equal to a fixed value (i.e., 0.8):

$$\mathbf{f} \sim U(0, \mathbf{f}_{\max})\Sigma_f^{-1} \tag{8}$$

To maintain the desired fold-enrichment correlation and not exceed the maximum fold-enrichment per annotation per trait, we scaled down $\mathbf{f}$ by a constant F equal to $F = \max_{d,k} f_{kd}/f_{\max,kd}$. Based on the scaled fold-enrichments $\mathbf{f}$, we then sampled causal variants the same way as above with no specific order.

3.4 Simulating phenotypes and GWAS summary statistics

We then simulated the effect sizes for causal SNPs as a function of minor allele frequency (p_i):

$$\beta_i = \frac{1}{\sqrt{2p_i(1-p_i)}} \tag{9}$$

We then simulated the quantitative phenotype or liability as:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{10}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_n) \tag{11}$$

Here $\mathbf{X}$ is the standardized genotype of SNPs across individuals, and $\sigma_e^2 = \frac{\sigma_g^2 - h_g^2\sigma_g^2}{h_g^2}$, such that $h_g^2 = \frac{\sigma_g^2}{\sigma_e^2+\sigma_g^2}$ and $\sigma_g^2 = \mathrm{Var}(\mathrm{X}\beta) = \beta'\Sigma_g\beta$ is the phenotypic variance explained by the genetics [8], where Σ_g is the genotypic correlation matrix (i.e., linkage disequilibrium) estimated as the Pearson correlation from the 10,000 simulated individuals. Finally, we simulated p-value and z-score (as t-statistics) by regressing phenotype $\mathbf{y}$ on $\mathbf{X}$ for each SNP i: ($\mathbf{y} \sim \mathbf{x}_i$). We repeatedly generated simulated datasets using the approach described above and evaluated each fine-mapping method on each of the simulated dataset.

4 Supplementary Fig.

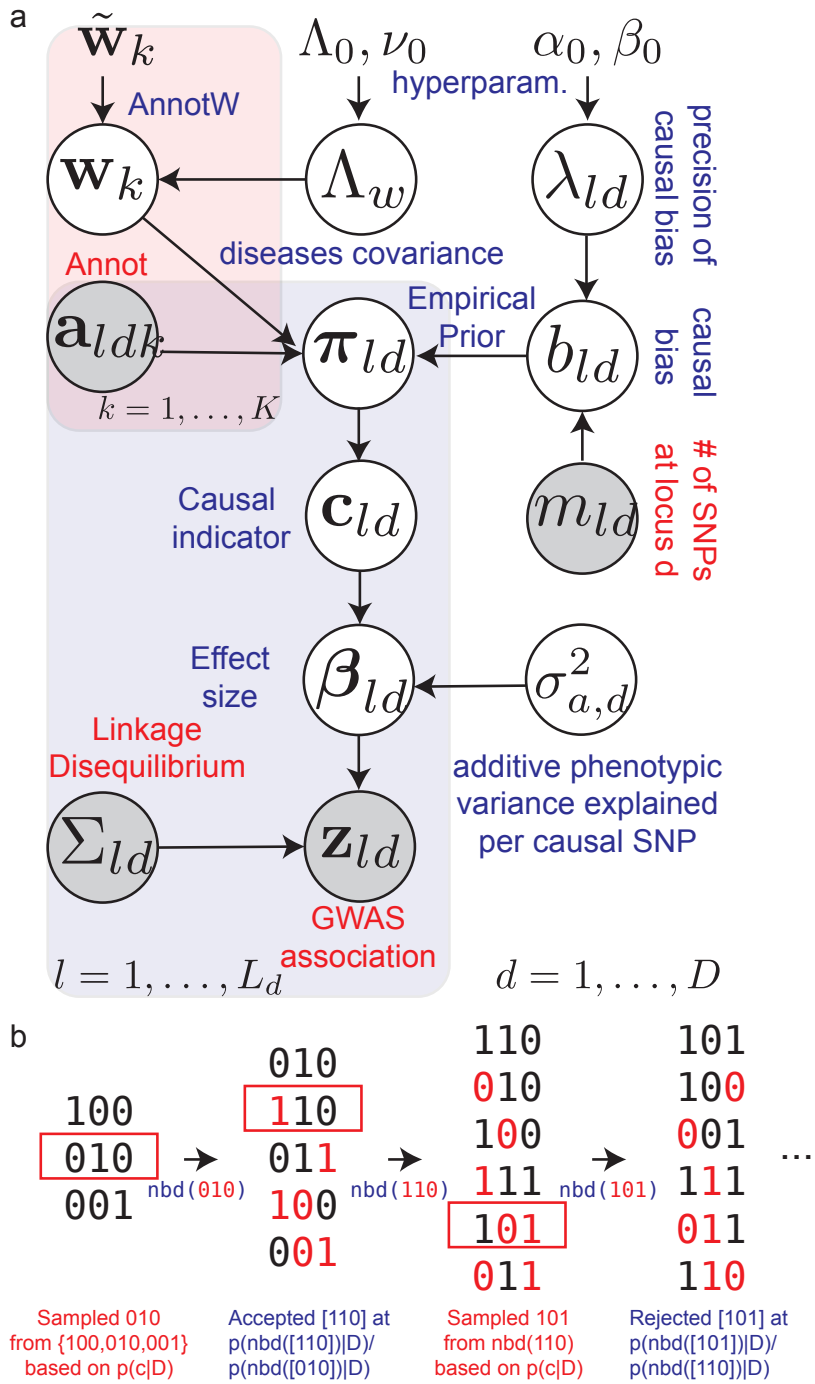


Figure S2: (**a**) RiVIERA-MT expressed in probabilistic graphical model. Shaded nodes are observed data and unshaded are latent variables or model parameters. The plates represent repeated pattern of same entities as indexed by l for loci, k for annotations, and d for diseases. The meaning of each variant is annotated beside each node. (**b**) Stochastic sampling of causal configurations. The sampling scheme was adapted from [9, 10]. 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.

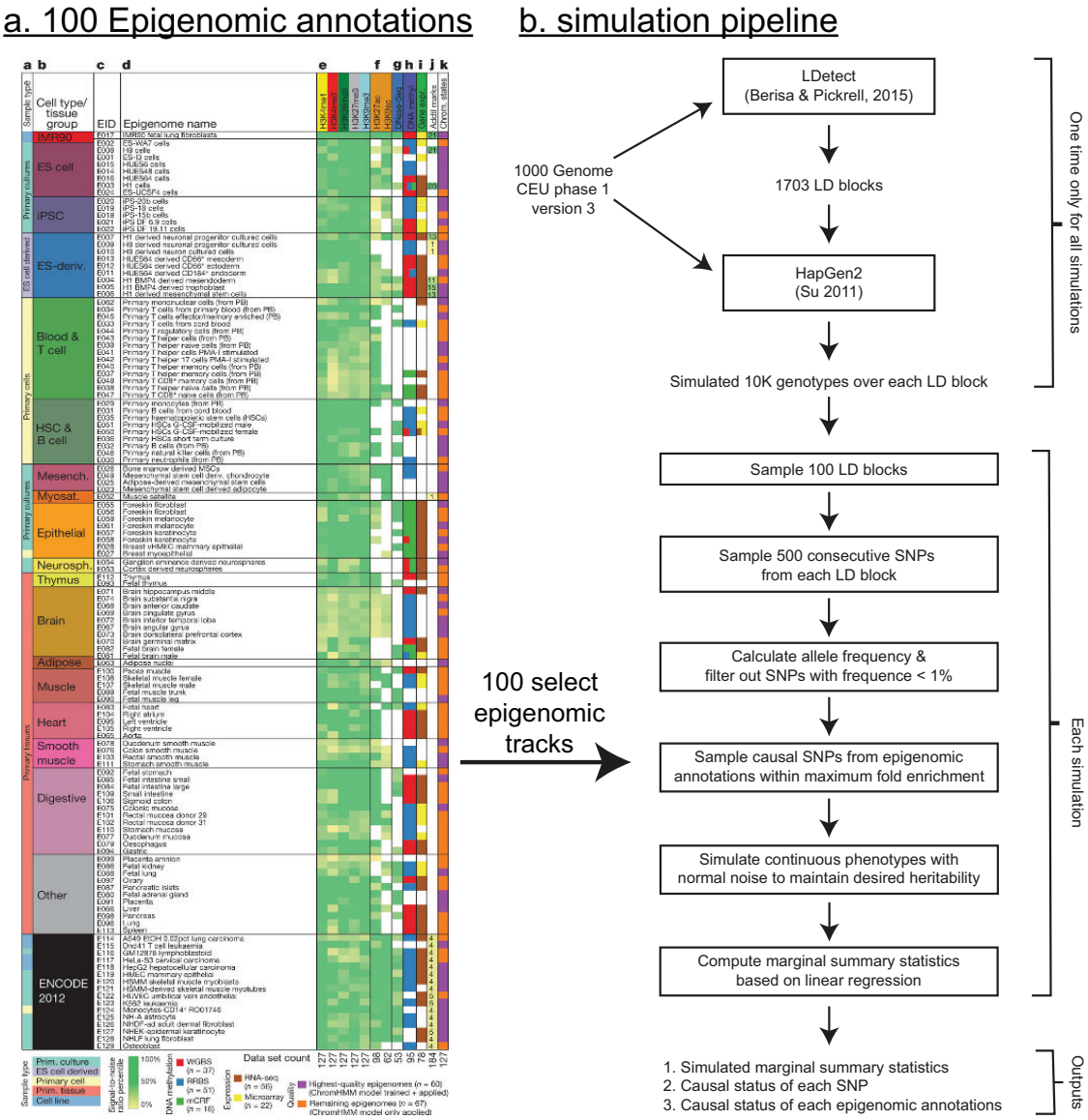


Figure S3: Simulation pipeline. HapGen2 [5] was used to sample haplotypes from 1000 Genome CEU data. For computational efficiency, we applied HapGen2 separately to each LD blocks approximated by LDetect [6]. We simulated 10,000 genotypes for each LD block. At each simulation run, we sampled 100 LD blocks each containing at maximum 500 SNPs to simulate risk loci for the 10,000 individuals. The causal variants were randomly sampled from the epigenomic annotations with predefined fold-enrichment until reaching the requried total number of causal variants. The continuous phenotypes were then simulated as a linear combination of the normalized genotypes weighted by the effect sizes of the causal variants. Finally, least-square linear regression on the 10,000 continuous phenotypes was performed for each variant to generate summary statistics in terms of both p-values and t-statistics that were input to the fine-mapping methods in comparison.

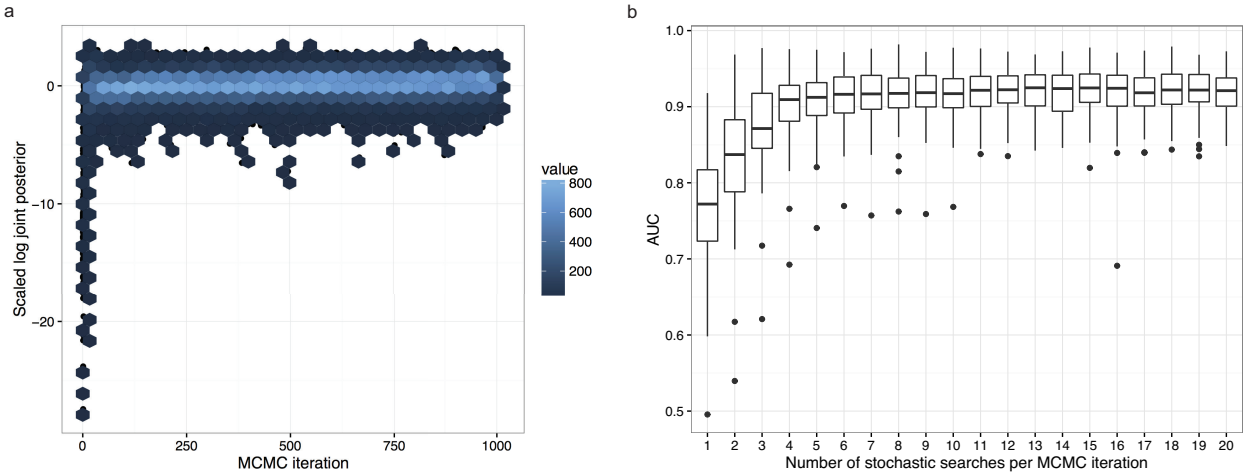


Figure S4: Model convergence. **a**. joint posterior as a function of MCMC iteration. **b**. Model performance as a function of number of stochastic searches per MCMC iteration. Area under of the curve (AUC) is defined as the area under of proportion of causal variants identified as a function of average number of SNPs selected per locus (i.e., **Supplementary Fig.** S5 below).

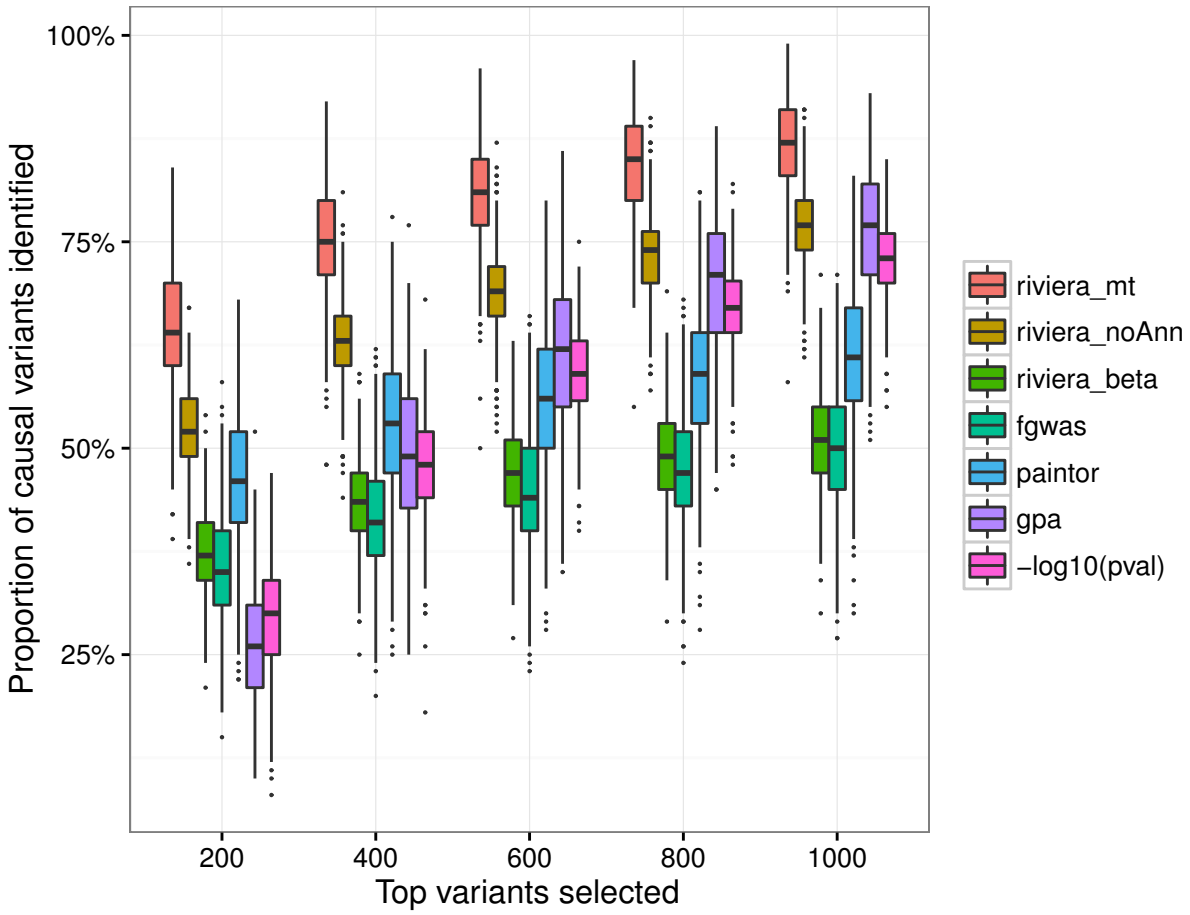


Figure S5: Power comparison as the proportion of causal variants identified as a function of the number of SNPs selected per locus. In this comparison, we also included GPA, which assumes independent SNPs but used annotations to prioritize them in a sensible way. Thus, we observed a reasonable performance when the selecting 1000 variants whereas methods that are limited to small number of variants (i.e., RiVIERA-beta, fgwas, PAINTOR tend to suffer from the loss of power to detect the remaining causal SNPs.

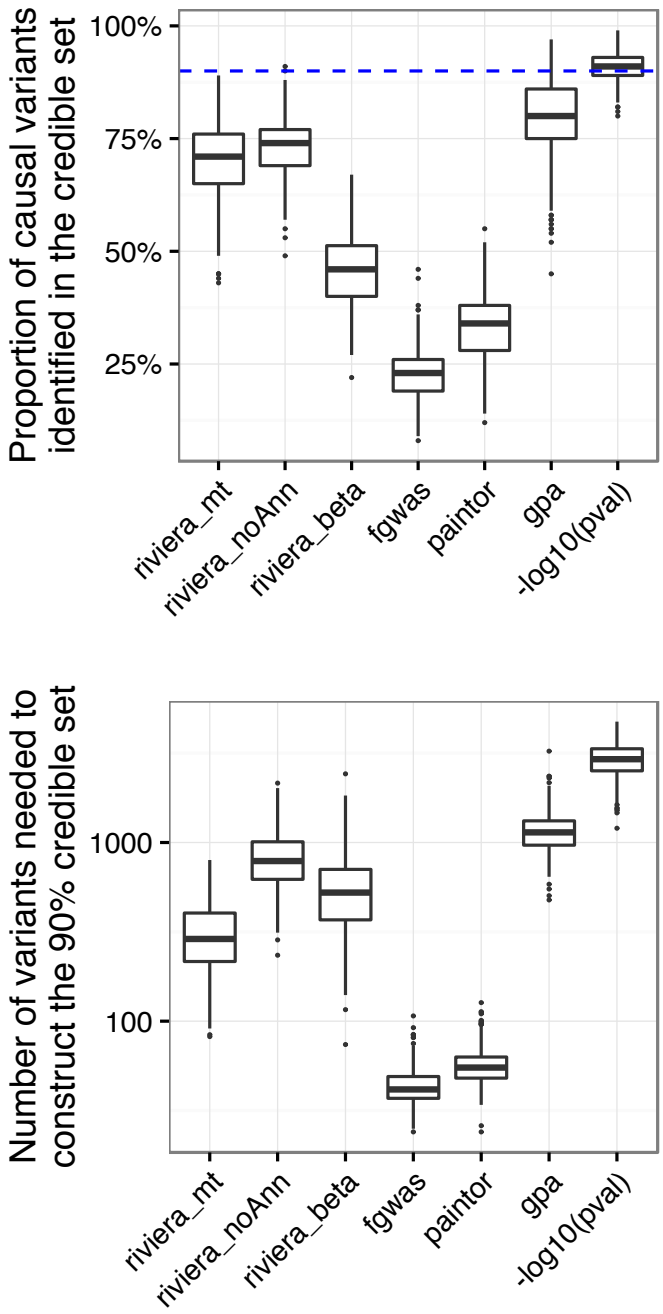


Figure S6: Calibration of posterior probabilities. All of the methods except GWAS -log p-values give posterior probabilities at each SNP, which allows us to construct 90% credible sets (blue dashed line) by adding SNPs in decreasing posterior to the set until reaching the 90% total posterior mass. For GWAS -log p-values, we performed the same by adding SNPs in decreasing order of -logP until reaching the 90% of the total sum of -logP. Upper panel displays the proportion of causal variants detected in the 90% credible set and the lower panel displays the total number of SNPs (with y-axis at log10 scale) required to construct the 90% credible set. As we can see, it is a trade-off between including more variants and recall (i.e., detecting causal variants). As estimated by the median over 500 simulation runs, the posterior of our RiVIERA-MT model is reasonably calibrated with 90% credible set detecting over 75% of the causal variants requiring only fewer than 500 variants.

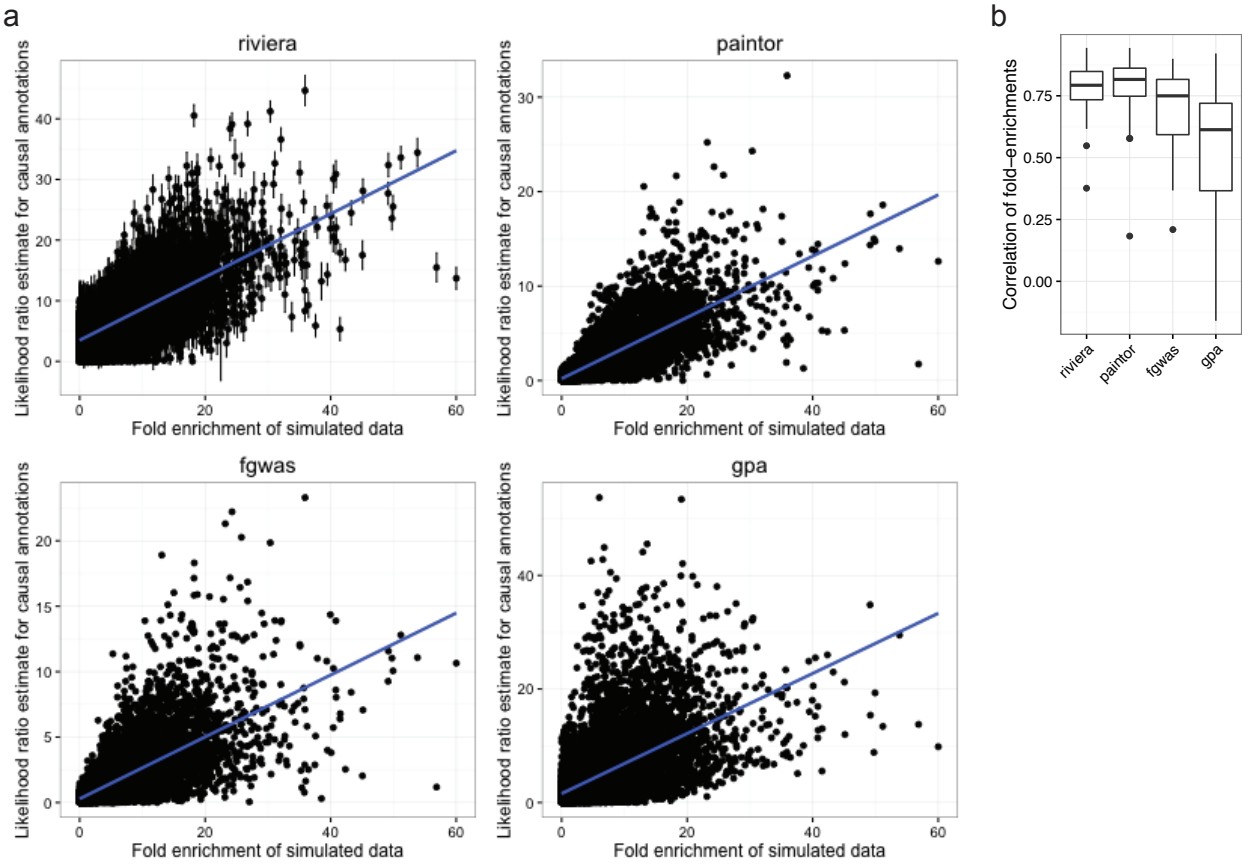


Figure S7: Functional enrichment analysis. (**a**) We estimated the enrichment of each annotations based on likelihood ratio tests. The y-axis is the estimate and the x-axis is the underlying fold-enrichments that were used to sample the causal variants from the corresponding annotation. The error bar indicates the 90% credible interval of the Bayesian LRT estimates by RiVIERA-MT. (**b**) Distribution of correlations for each method across simulated datasets.

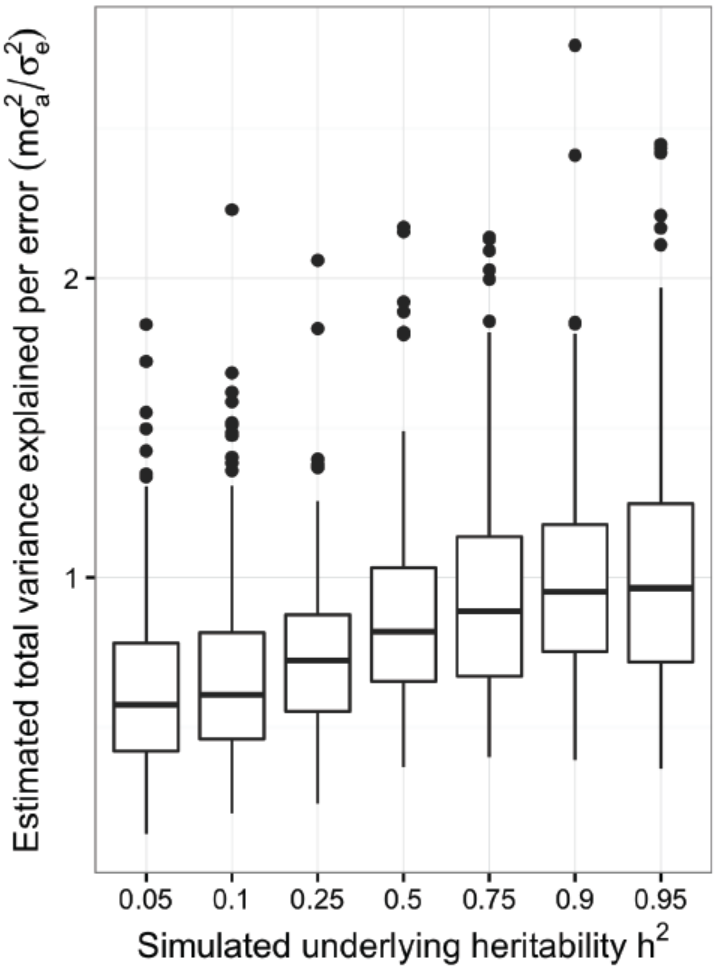


Figure S8: Estimation of per-SNP variance explained. The heritability estimates increased as the underlying heritability increased in the simulated data. The results are based on 100 simulations per heritability value.

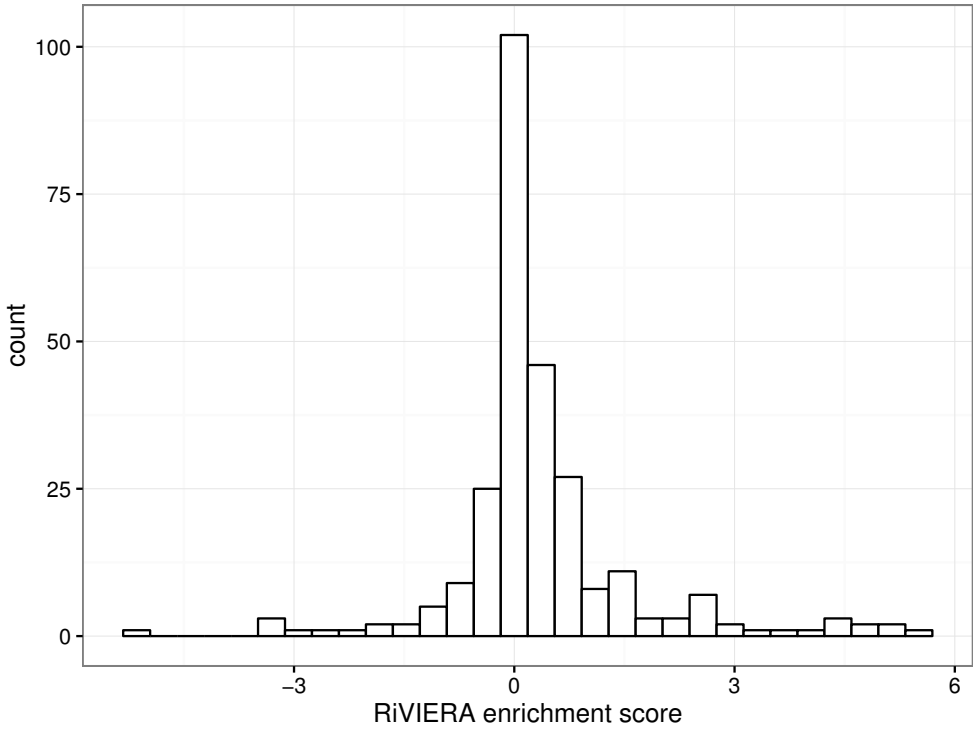


Figure S9: RiVIERA-MT enrichment score of GWAS traits over 10 tissue groups.

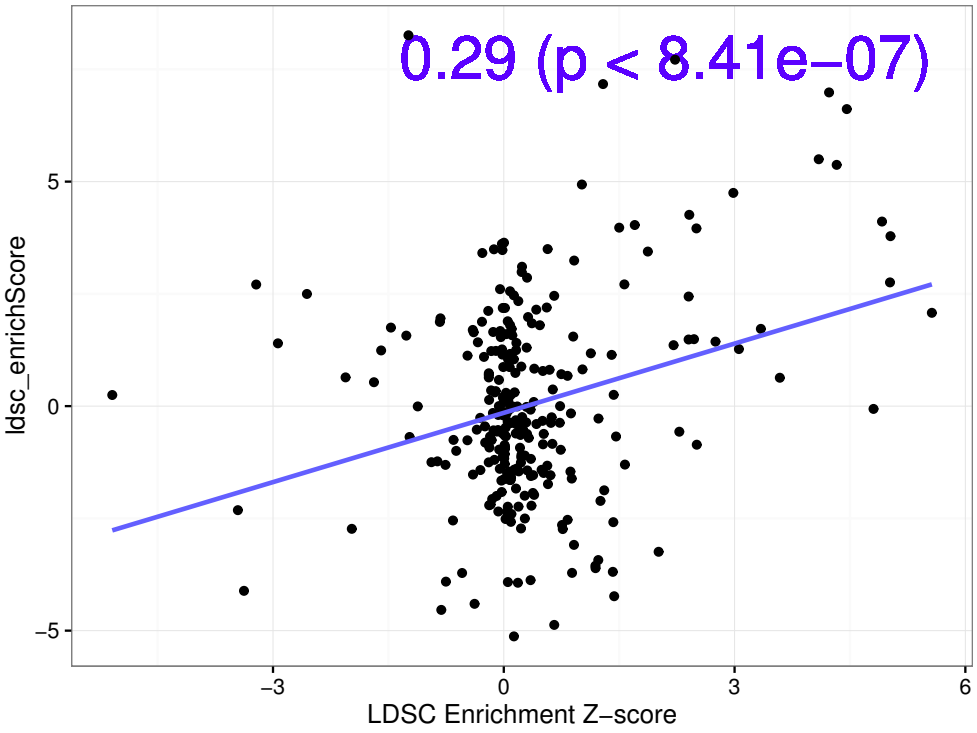


Figure S10: Tissue-specific enrichment scores comparison between RiVIERA-MT and LDSC.

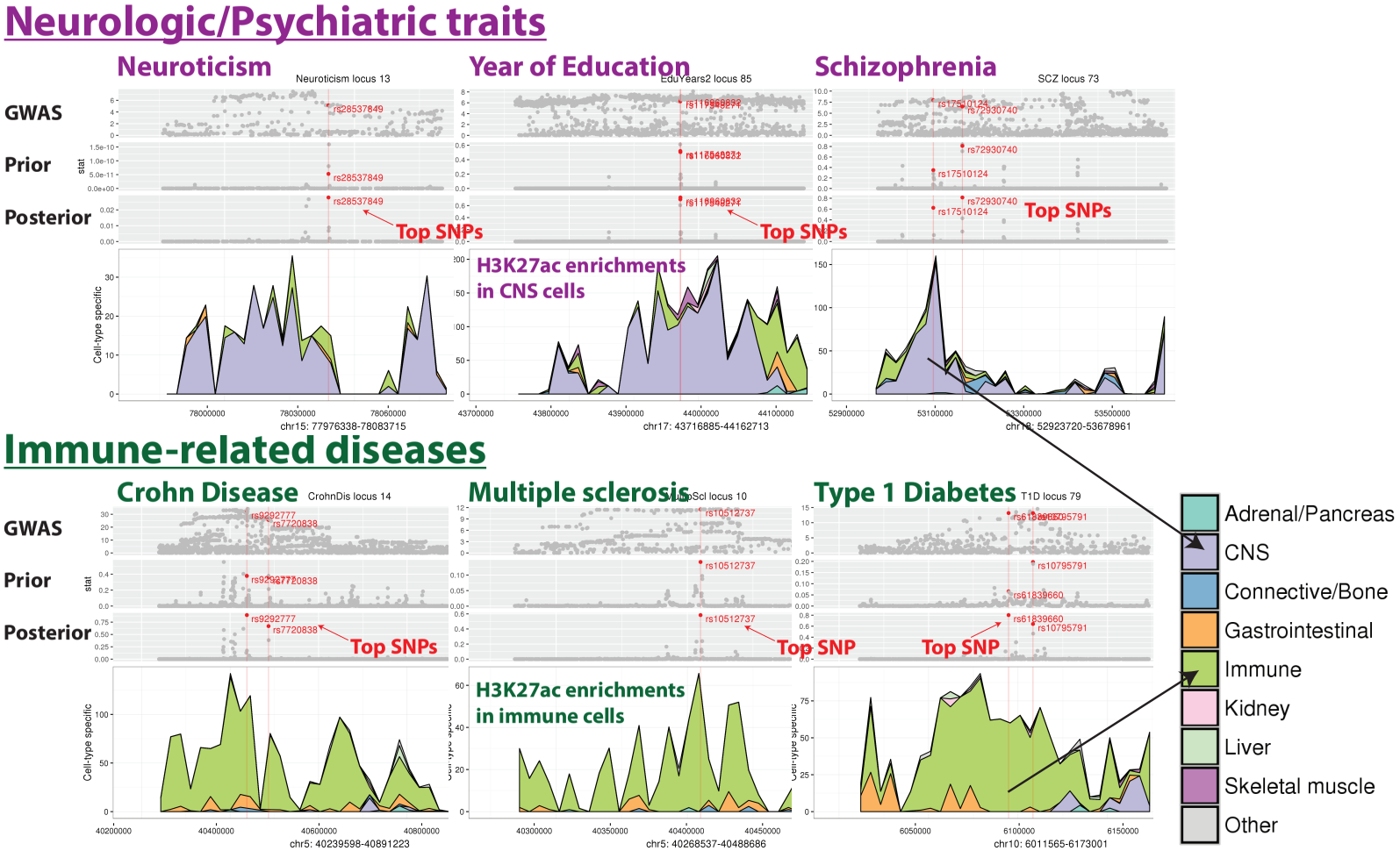

Neurologic/Psychiatric traits
Neuroticism
Neuroticism locus 13
Year of Education
EduYears2 locus 85
Schizophrenia
SCZ locus 73
GWAS
Prior
Posterior
stat
Top SNPs
H3K27ac enrichments in CNS cells
Cell-type specific
chr15: 77976338-78083715
chr17: 43716885-44162713
chr18: 52923720-53678961
Immune-related diseases
Crohn Disease
CrohnDis locus 14
Multiple sclerosis
MultiScl locus 10
Type 1 Diabetes
T1D locus 79
Top SNP
H3K27ac enrichments in immune cells
chr5: 40239598-40891223
chr5: 40268537-40488686
chr10: 6011565-6173001
Adrenal/Pancreas
CNS
Connective/Bone
Gastrointestinal
Immune
Kidney
Liver
Skeletal muscle
Other

Figure S11: Fine-mapping risk loci of neurologic/psychiatric and immune disorders. **a**. Selected distinct risk loci for neuroticism, education year, and schizophrenia. **b**. Selected distinct risk loci for Crohn disease, multiple sclerosis, and type 1 diabetes. In each panel, the four tracks (top to bottom) GWAS -logP values, empirical prior and posterior from RiVIERA-MT, and weighted H3K27ac peaks by the enrichment parameters that are also learned by RiVIERA-MT. Top SNPs nominated by the posterior are highlighted in red. Due to strong LD in these risk loci, it is difficult to identify the causal SNPs. On the other hand, the empirical priors of SNPs are more prominent around the enhancer peaks. By leveraging the enhancer enrichments as epigenomic prior and GWAS genetic signals as likelihood (taking into account LD), RiVIERA-MT nominates only a handful of SNPs as the putative causal SNPs. In the case of the three neurologic/psychiatric traits, the enhancer activities are predominantly acting through central nervous system (CNS) tissue group whereas the enrichment signals primarily come from immune cell types for the 3 immune traits, consistent with the underlying physiological context of the diseases

Celiac

	gwasP	PIP	PIPj
gwasP	100	2	2.2
PIP	55.3	100	54.6
PIPj	64.8	58.5	100

CrohnDis

	gwasP	PIP	PIPj
gwasP	100	1	1
PIP	37.7	100	17.8
PIPj	38.7	18.7	100

UlcerCol

	gwasP	PIP	PIPj
gwasP	100	2.1	1.7
PIP	41.9	100	49.4
PIPj	47	71.1	100

Lupus

	gwasP	PIP	PIPj
gwasP	100	2.9	1.8
PIP	51.1	100	29
PIPj	53.2	48.2	100

MultipScl

	gwasP	PIP	PIPj
gwasP	100	2.9	2.7
PIP	42.2	100	32.6
PIPj	60.2	49.6	100

PrimBilCir

	gwasP	PIP	PIPj
gwasP	100	1.5	1
PIP	42.5	100	31.4
PIPj	49.4	52.6	100

RheuArth

	gwasP	PIP	PIPj
gwasP	100	2.2	1.4
PIP	58.3	100	32.3
PIPj	57.3	48.2	100

T1D

	gwasP	PIP	PIPj
gwasP	100	5.4	3.5
PIP	54.9	100	34.4
PIPj	60.4	58.2	100

SCZ

	gwasP	PIP	PIPj
gwasP	100	0.9	1.4
PIP	57	100	38.9
PIPj	58	24.2	100

MeanPutamen

	gwasP	PIP	PIPj
gwasP	100	5.2	9.7
PIP	61.7	100	51.1
PIPj	64.7	28.8	100

Neuroticism

	gwasP	PIP	PIPj
gwasP	100	0.4	0.8
PIP	50.6	100	42.2
PIPj	54.6	23	100

Bipolar

	gwasP	PIP	PIPj
gwasP	100	13.8	5.5
PIP	67	100	15.1
PIPj	75.5	42.8	100

HDL

	gwasP	PIP	PIPj
gwasP	100	0.8	0.8
PIP	42.7	100	38.2
PIPj	37.2	34	100

LDL

	gwasP	PIP	PIPj
gwasP	100	0.8	0.7
PIP	23.4	100	36.7
PIPj	38.3	61.2	100

Trigly

	gwasP	PIP	PIPj
gwasP	100	1	1
PIP	23.2	100	65.5
PIPj	17.5	54.4	100

Figure S12: Overlap percentage of prioritized SNP sets based on GWAS -logP (gwasP) and RiVIERA-MT single-trait (PIP) and multi-trait mode (PIPj)

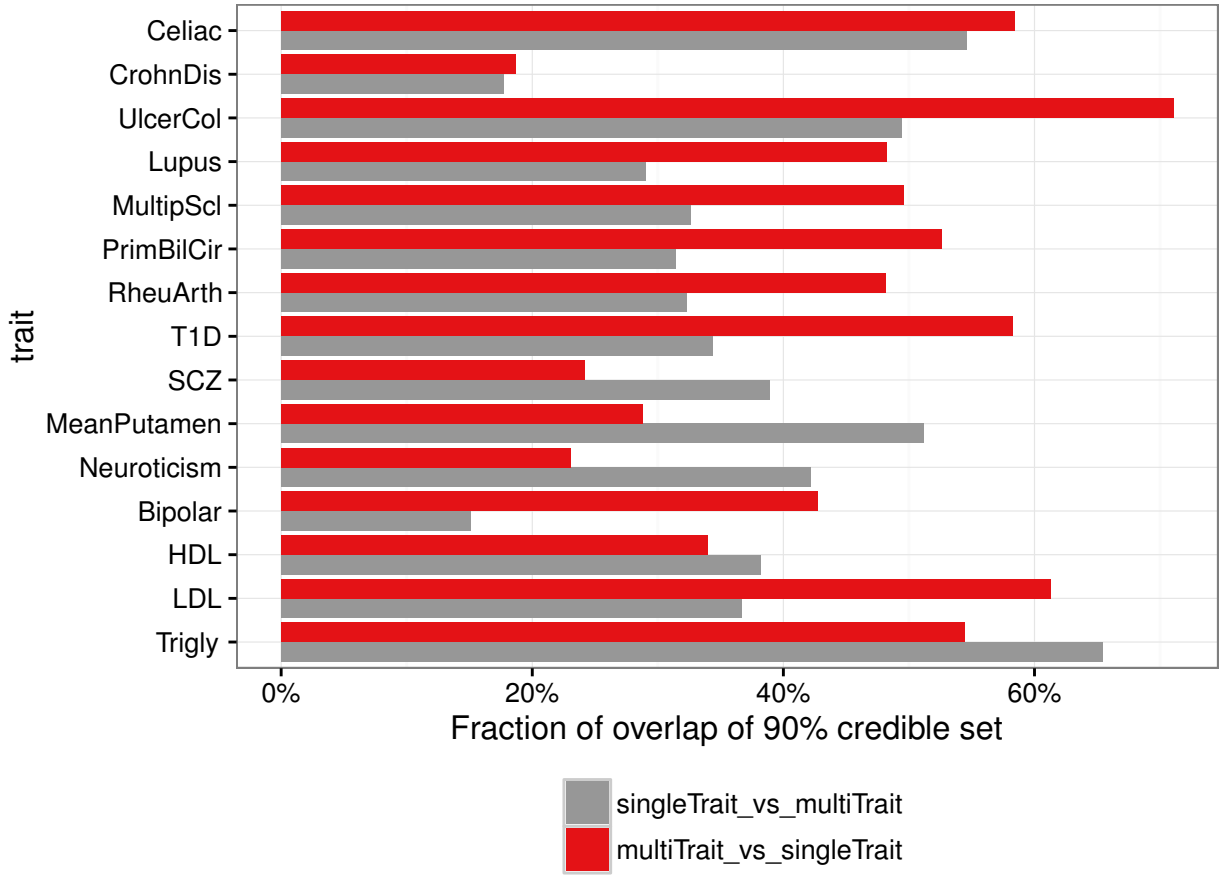


Figure S13: Overlap between multi-trait and single-trait inferred 90% credible sets

5 Supplementary Table

Table S1: GWAS data summary

Trait	Abbrev	Website
Autism	Autism	http://www.med.unc.edu/pgc
Coronary artery disease	CorArtDis	http://www.cardiogramplusc4d.org
Celiac disease	Celiac	https://www.immunobase.org
Depression symptom	DeprSym	http://ssgac.org
Mean putamen	MeanPutamen	http://enigma.ini.usc.edu
Year of education	EduYears	http://ssgac.org
Type 2 diabetes	T2D	http://diagram-consortium.org
Body mass index	BMI	http://portals.broadinstitute.org/collaboration/giant
Height	Height	http://portals.broadinstitute.org/collaboration/giant
High density lipoportein	HDL	http://csg.sph.umich.edu
Hemoglobin A1c	HbA1C	https://www.magicinvestigators.org/downloads/
Low density lipoportein	LDL	http://csg.sph.umich.edu/abecasis/public/lipids2013/
Systemic lupus erythematosus	Lupus	https://www.immunobase.org
Fasting Glucose	FastGlu	https://www.magicinvestigators.org/downloads/
Multiple_sclerosis	MultipScl	https://www.immunobase.org
Neuroticism	Neuroticism	http://ssgac.org
Primary biliary cirrhosis	PrimBilCir	https://www.immunobase.org
Rheumatoid arthritis	RheuArth	https://www.immunobase.org
Subject well being	SubjWelBeing	http://ssgac.org
Triglycerides	Trigly	http://csg.sph.umich.edu/abecasis/public/lipids2013/
Type 1 diabetes	T1D	https://www.immunobase.org
Year of education	EduYears2	http://ssgac.org
Crohn disease	CrohnDis	https://www.ibdgenetics.org/downloads.html
Ulcerative colitis	UlcerCol	https://www.ibdgenetics.org/downloads.html
Schizophrenia	SCZ	www.med.unc.edu/pgc/
Bipolar disorder	Bipolar	www.med.unc.edu/pgc/
Ever smoked	EverSmoked	www.med.unc.edu/pgc/

Table S2: Fine-mapping results summary

Trait	Loci	Total	Joint	GWAS	SubThres	90% CS (A-)	90% CS (A+)	90% CS (joint)
Autism	12	5201		1	99	1744	1160	
CorArtDis	16	5307		573	847	1356	824	
Celiac	11	5405	IMN	680	1170	1084	152	142
CrohnDis	60	35877	IMN	5100	7770	3667	738	700
UlcerCol	35	16774	IMN	1471	2806	2138	716	498
SCZ	81	47484	PSY	4317	10229	8808	565	910
DeprSym	32	13764	PSY	2	520	3839	3400	
EduYears	13	4853		17	771	1188	910	
MeanPutamen	19	10184	PSY	286	1314	2365	747	1327
T2D	12	5283		548	927	764	872	
BMI	21	11052		1764	2895	1588	373	
Height	141	81704		11109	18502	11672	1417	
HbA1C	11	6927		449	773	554	727	
HDL	55	42230	LID	6079	9129	3753	644	724
LDL	47	30482	LID	4769	7239	2383	749	449
Lupus	44	22861	IMN	1502	3080	2097	1137	684
FastGlu	13	5411		587	936	537	388	
MultipScl	19	7567	IMN	993	1889	1275	389	256
Neuroticism	18	25973	PSY	4640	7440	4175	154	282
Bipolar	17	6875	PSY	90	707	1534	1270	449
PrimBilCir	25	14421	IMN	1440	2542	1510	414	247
RheuArth	29	14952	IMN	1013	1933	1957	489	328
SubjWelBeing	21	7485		22	449	1771	824	
EverSmoked	3	972		0	1	456	565	
Trigly	36	24040	LID	4070	6209	1750	802	965
T1D	125	40633	IMN	1983	3286	7796	3688	2176
EduYears2	96	57615		4996	12283	9518	1066	

- Loci: number of risk loci containing GWAS or sub-threshold GWAS SNPs;
- Total: total number of SNPs within the risk loci;
- Joint: the group label of for related traits that were modeled jointly by the multi-trait mode of the RiVIERA-MT model:
  - IMN: immune traits
  - PSY: psychiatric traits
  - LID: lipid traits
- GWAS: p<5e-8;
- SubThres: GWAS sub-threshold p<1E-5;
- CS: the number of SNPs in the credible set (CS);
- A-/+: inferring risk variants without (A-) or with annotations (A+).

References

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

2. Neal, R. M. *Probabilistic inference using Markov chain Monte Carlo methods*. Ph.D. thesis, University of Toronto (1993).

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

4. Homan, M. D. & Gelman, A. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research* **15**, 1593–1623 (2014).

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

6. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics (Oxford, England)* (2015).

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

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

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

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