RiVIERA-beta: Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases

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

Genome wide association studies (GWAS) provide a powerful approach for uncovering disease-associated variants in human, but fine-mapping the causal variants remains a challenge. This is partly remedied by prioritization of disease-associated variants that overlap GWAS-enriched epigenomic annotations. Here, we introduce a new Bayesian model RiVIERA-beta (Risk Variant Inference using Epigenomic Reference Annotations) for inference of driver variants by modelling summary statistics p-values in Beta density function across multiple traits using hundreds of epigenomic annotations. In simulation, RiVIERA-beta promising power in detecting causal variants and causal annotations, the multi-trait joint inference further improved the detection power. We applied RiVIERA-beta to model the existing GWAS summary statistics of 9 autoimmune diseases and Schizophrenia by jointly harnessing the potential causal enrichments among 848 tissue-specific epigenomics annotations from ENCODE/Roadmap consortium covering 127 cell/tissue types and 8 major epigenomic marks. RiVIERA-beta identified meaningful tissue-specific enrichments for enhancer regions defined by H3K4me1 and H3K27ac for Blood T-Cell specifically in the 9 autoimmune diseases and Brain-specific enhancer activities exclusively in Schizophrenia. Moreover, the variants from the 95% credible sets exhibited high conservation and enrichments for GTEx whole-blood eQTLs located within transcription-factor-binding-sites and DNA-hypersensitive-sites. Furthermore, joint modeling the nine immune traits by simultaneously inferring and exploiting the underlying epigenomic correlation between traits further improved the functional enrichments compared to single-trait models.

1 Introduction

Genome wide association studies (GWAS) can help gain numerous insights on the genetic basis of complex diseases, and ultimately contribute to personalized risk prediction and precision medicine [1–4]. However, fine-mapping the exact causal variants is challenging due to linkage disequilibrium (LD) and the lack of ability to interpret the function of noncoding variants, which contribute to about 90% of the current GWAS catalog (40.7% intergenic and 48.6% intronic; [5]). On the other hand, several lines of evidence have been proposed to help interpret non-coding genetic signals, in order to gain insights into potential regulatory functions. In particular, epigenomic annotations can pinpoint locations of biochemical activity indicative of cis-regulatory functions [6, 7]. Indeed, comparison with genome-wide annotations of putative regulatory elements has shown enrichment of GWAS variants in enhancer-associated histone modifications, regions of open chromatin, and conserved non-coding elements [3, 6, 8–12], indicating they may play gene-regulatory roles. These enrichments have been used to predict relevant cell types and non-coding annotations for specific traits [6, 9, 13]. Furthermore, many complex traits potentially share causal mechanisms such as autoimmune diseases [14, 15] and psychiatric disorders [16, 17]. Thus, methods that jointly model the intrinsic comorbidity implicated in the GWAS summary statistics of the related traits may confer higher statistical power of causal variants detection. Recently, several methods were developed to utilize the wealth of genome-wide annotations primarily provided by ENCODE consortium to predict causal variants and novel risk variants that are weakly associated in complex traits. Pickrell (2014) developed a statistical approach called fgwas that models association statistics of a given trait and used regularized logistic function to simultaneously learn the relevant annotations. To account for LD, fgwas assumes at most one causal variants per locus via a softmax function. Kichaev et al. (2014) recently developed a multivariate Gaussian framework called PAINTOR, which allows for more than one causal SNP but at most three to be located within a single locus by considering all of the combinatorial settings [18]. Chung et al. (2014) used a maximum likelihood framework called GPA to infer driver variants shared among multiple traits by modeling the corresponding GWAS p-values as Beta distributions with an option of using one or more sets of annotations to improve the power detecting causal variants [19]. Although useful, these methods are often designed to simultaneously operate on a small number of independent annotations due to some computational constraints. Moreover, most methods only operate on one trait at a time whereas exploiting the correlation between traits at the epigenomic annotation level may prove useful for shared causal mechanisms that go beyond the level of individual variants.

In this article, we describe a novel Bayesian framework called RiVIERA-beta (Risk Variant Inference using Epigenomic Reference Annotations to model Beta likelihood of GWAS summary statistics p-values). The main novelty of RiVIERA-beta is the ability to perform efficient Bayesian inference of the intrinsic causal signals across multiple traits while

simultaneously inferring and exploiting enrichment signals and their correlation between traits over hundreds of tissue-specific epigenomic annotations. We achieve this efficiently via stochastic sampling of loci and powerful Hamiltonian Monte Carlo sampling of model parameters [20]. We first use simulation to demonstrate the utility of RiVIERA-beta in prioritizing driver variants and detecting functional epigenomic annotations. We then apply RiVIERA-beta to some of the most well-powered GWAS datasets, consisting of 9 immunological disorders from ImmunoBase [15] and Schizophrenia 2014 data from Psychiatric Genomic Consortium [21]. To infer tissue-specific epigenomic enrichments, we utilize the largest compendium of epigenomic annotations to date from ENCODE/Roadmap Consortia, consisting of 848 annotations including 8 major epigenomic marks across 127 distinct cell types [7]. This allows us to revisit the GWAS of these 10 common complex disorders by inferring their underlying regulatory variants implicated at the tissue-specific epigenomic contexts.

2 MATERIALS AND METHODS

GWAS summary statistics

The GWAS summary statistics for the nine immune diseases were obtained from ImmunoBase (March 17, 2015) [15]. The nine diseases are: Autoimmune Thyroid Disease (ATD), Celiac Disease (CEL), Juvenile Idiopathic Arthritis (JIA), Multiple Sclerosis (MS), Narcolepsy (NAR), Primary Biliary Cirrhosis (PBC), Psoriasis (PSO), Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D). We imputed the p-values of un-genotyped SNPs using FAPI and 1000 Genome European data (Phase 1 version 3) [22]. We then obtained the p-values of SNPs that fall within the pre-defined risk loci available from ImmunoBase for each of the 9 immune traits. For all analyses, we filtered out risk loci or variants in the MHC regions or sex chromosomes X and Y. The Schizophrenia 2014 (SCZ2) summary data containing 642846 observed and imputed SNPs were obtained from Psychiatric Genomic Consortium (PGC) [21]. Among these, 54132 SNPs fall within the 105 SCZ-associated loci of the autosomes (chr 1-22) defined by PGC (we filtered out the 3 loci on chromosome X). **Table** 1 summarizes the total number of SNPs and risk loci for each individual GWAS that were subject to the proposed fine-mapping analyses.

Roadmap epigenome data

Roadmap epigenome data were obtained from Roadmap epigenomic web portal (March, 2015). Peaks were defined if their p-values were below 0.01 (i.e., following the definition of "Narrow Peaks" [7]). In total, there are 848 epigenome tracks, including 8 epigenomic marks namely H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, H3K9ac and DNase I in in 127 cell or tissue types, which were grouped into 19 categories [7]. To associate each SNP with the annotations, we overlapped their genomic coordinates with each bigWig epigenome track making use of the R packages *rtracklayer* and *GenomicAlignments*. SNPs that fall within a peak of an annotations will have value 1 otherwise 0 for that annotation. The resulting matrix is a $V_d \times K$ input matrix containing the epigenomic values across $K = 848$ marks for each of the V_d SNPs in disease d.

Running existing fine-mapping software on simulated data

fgwas

The software `fgwas` [23] (version 0.3.4) were downloaded from GitHub. We prepared the input for fgwas (1) the Z scores calculated as the t-statistics of the linear coefficients of the genotype of each variant fitted separately by least square regression on the simulated continuous phenotypes (**Materials and methods**) and (2) 100 discretized epigenomic annotations at $p < 0.01$. To enable fine-mapping, we issued `-fine` flag and specify the region numbers for each SNP in the input file as required by the software. As part of the outputs from fgwas, we obtained 'PPA' and 'estimate' for the causal variants and influences of each epigenomic annotations, respectively.

GPA

GPA (0.9-3) [19] was downloaded from GitHub and run with default settings. Same as above, we set the annotations to one at p-value < 0.01 and 0 otherwise. To test for trait-relevant annotations, we followed the package vignette. Briefly, we fit two GPA models with and without the annotation and compared the two models by `aTest` function from GPA, which performs likelihood-ratio (LR) test via χ^2 approximation, and obtained the enrichment scores as the -log10 p-value.

PAINTOR

PAINTOR (version 2.1) was downloaded from GitHub [18]. 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. We then extracted the 'Posterior_Prob' and 'Enrichment.Values' as the model predictions for causal variants and causal annotations, respectively.

Details of RiVIERA-beta Bayesian model

Inference of empirical prior π_{vd}

We first define the empirical prior function of a variant v being associated with disease d as a logistic function:

$$\pi_{vd} = [1 + \exp(-[\sum_k w_{kd} e_{vk} + w_{0d}])]^{-1} \quad (1)$$

where $w_{kd} \in \mathbf{w}_d$ denotes the linear coefficient or the influence of the k^{th} epigenomic mark affecting disease d and w_{0d} is the linear bias.

We assume that epigenomic causal effect w_{kd} follows a multivariate Gaussian distribution with zero mean and unknown covariance:

$$w_{kd} \sim \mathcal{N}(0, \Lambda_w^{-1}) \quad (2)$$

$$\Lambda_w \sim \mathcal{W}(\Lambda_0, \nu_0) \quad (3)$$

where Λ_w is a $D \times D$ inverse covariance matrix $\Lambda_w = \Sigma_w^{-1}$ to model the pairwise epigenomic correlation among D diseases. It follows a Wishart distribution with identity matrix as prior (i.e., by default, we assume *aprior* no correlation between the target traits) and $\nu_0 = 0$ (i.e., by default, we did observe any samples *aprior* that are indicative of the correlation between any two diseases being modeled). The hyperparameters can be easily modified to incorporate prior belief on the correlation between any two diseases of interests.

Additionally, the bias w_{0d} follows a Gaussian distribution with unknown variance and mean determined based on our prior belief of the causal fraction π_0:

$$w_{0d} \sim \mathcal{N}(logit(\pi_0), \lambda_{0d}^{-1}) \tag{4}$$
$$\lambda_{w_{0d}} \sim \Gamma(\alpha_0, \beta_0) \tag{5}$$

where $logit(\pi_0) = \log \frac{\pi_0}{1-\pi_0}$. By default, we set π_0 to 0.01, implying that 1% of the SNPs in the risk loci are expected to be causal when no functional enrichment. We set $\alpha = 0.01$ and $\beta = 0.0001$ to enable a broad hyperprior for w_{0d}.

Notably, w_{kd} can be interpreted as enrichment coefficient for annotation k in disease d, where a positive w_{kd} will increase the causal prior π_{vd} when $e_{vk} = 1$. During the training, however, w_{kd} may become negative, which makes the interpretation difficult. Thus, we constrain w_{kd} to be non-negative values, which involves imposing infinitely high potential energy for negative w_{kd}. More details are described in **Supplementary Text** 1.

Inference of variant causality c_{vd} given prior π_{vd} and model parameters μ_d, ϕ_d

Because the target association variable a_{vd} for variant v in disease d represents p-values, which are continuous and restricted to the interval (0, 1), we assume that it follows a Beta distribution with unknown mean μ_d and unknown precision ϕ_d:

$$a_{vd} \sim \mathcal{B}(\mu_d, \phi_d) \tag{6}$$

Note that we re-parameterize Beta density function from the traditional "rate" p and "shape" q parameters, and instead use mean $\mu = p/(p+q)$ and precision $\phi = p + q$, as per [24, 25]. Specifically, the density function of association variable a_{vd} is defined as follows:

$$f(a_{vd}; \mu_d, \phi_d) = \frac{\Gamma(\phi_d)}{\Gamma(\mu_d\phi_d)\Gamma((1-\mu_d)\phi_d)} a_{vd}^{(\mu_d\phi_d - 1)}(1 - a_{vd})^{(1-\mu_d)\phi_d - 1} \tag{7}$$

Further, we let the mean μ_d and precision ϕ_d follow Beta and uniform prior, respectively:

$$\mu_d \sim \mathcal{B}(\mu_0, \phi_0) \tag{8}$$
$$\phi_d \sim \mathcal{U}(0, \phi_{\max}) \tag{9}$$

where the hyperparameters (μ_0, ϕ_0) reflect *apriori* belief on the p-value signal of a causal variant. By default, we set $\mu_0 = 0.1$ and $\phi_0 = 2$. If $\phi_{\max} = \infty$, ϕ follows an improper prior. Because it is unlikely to have a very large ϕ, by default, we set $\phi_{\max}$ to 1000. Notably, as long as $\phi_{\max}$ is large, the inference results remain the same with different $\phi_{\max}$ values.

With the prior $p(c_{vd}|\mathbf{w}_d, \mathbf{e}_v) \equiv \pi_{vd}$ and likelihood $p(a_{vd}|\mu_d, \phi_d) \equiv f(a_{vd}; \mu_d, \phi_d)$ established, the posterior probability of association (PPA) [26] of variant v being causal for disease d then follows:

$$p(c_{vd}|a_{vd}, \pi_{vd}) = \frac{p(a_{vd}|c_{vd})p(c_{vd}|\mathbf{w}_d, \mathbf{e}_v)}{\sum_{v' \in \mathcal{V}_b} p(a_{v'd}|c_{v'd})p(c_{v'd}|\mathbf{w}_d, \mathbf{e}_{v'})} \tag{10}$$

where $\mathcal{V}_b$ represent all variants within locus b. The 95% credible set $\mathcal{C}_{bd}$ for each locus b is the minimal number of SNPs $v' \in \mathcal{C}_{bd}$ in the locus such that $\sum_{v' \in \mathcal{C}_{bd}} p(c_{v'd}|a_{v'd}, \pi_{v'd}) \geq 0.95$.

Joint posterior distribution

The complete likelihood density function treating c_{vd} as missing values is defined as:

$$\begin{aligned}\mathcal{L} &= \prod_{v,d} f(a_{vd}, \pi_{vd}, c_{vd}, |\mathbf{e}_v, \mu_d, \phi_d) \\ &= \prod_{v,d} [\pi_{vd} p(a_{vd}|\mu_d, \phi_d)]^{c_{vd}} (1 - \pi_{vd})^{(1-c_{vd})}\end{aligned} \tag{11}$$

The logarithmic joint posterior density function is then:

$$\begin{aligned}\log p(\Theta|\mathcal{D}) &= \log f(\mu, \phi, \mathbf{W}, \Lambda_w, \lambda_{0d}|\mathbf{E}, \mathbf{c}_d, \pi_d, \mathbf{a}_d) \\ &\propto \log f(\Lambda_w|\Lambda_0, \nu_0) + \sum_d \log f(\lambda_{0d}|\alpha_0, \beta_0) \\ &+ \log f(\mathbf{W}|\Lambda_w) + \sum_d \log f(w_{0d}|\mu_{w_0}, \lambda_{0d}) \\ &+ \log f(\mu_d|\mu_0, \phi_0) + \log f(\phi_d) \\ &+ \sum_{v,d} \log f(a_{vd}, \pi_{vd}, c_{vd}|\mathbf{e}_v, \mu_d, \phi_d)\end{aligned} \tag{12}$$

In principle, causality is inferred by integrating out all nuisance parameters:

$$\begin{aligned}p(c_{vd}|\mathbf{a}_d, \mathbf{e}_{vd}) = \int & f(c_{vd}|\mathbf{a}_d, \mathbf{e}_v, \mu_d, \phi_d, \mathbf{w}_d, \Lambda_w) \\ & f(\mu_d, \phi_d, \mathbf{w}_d, \Lambda_w|\mathbf{a}_d, \mathbf{e}_v) d\mu_d, \phi_d, \mathbf{w}_d, \Lambda_w\end{aligned} \tag{13}$$

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

Markov Chain Monte Carlo

We use Gibbs sampling [27] 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 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 [28]:

$$\Lambda_w|\mathbf{W} \sim \mathcal{W}((\Lambda_0^{-1} + \mathbf{S})^{-1}, \nu_0 + K) \tag{14}$$

where $\mathbf{S}$ is the sample variance of $\mathbf{W}$, i.e., $S = \mathbf{W}^T\mathbf{W}$.

Similarly, we sample λ_{0d} from Gamma posterior distribution:

$$\lambda_{0d}|w_{0d} \sim \Gamma(\alpha_0 + 0.5, (\beta_0 + \frac{(w_{0d} - \mu_{w_0})^2}{2})^{-1}) \tag{15}$$

To sample epigenomic effects $\mathbf{w}_d$, prior bias w_{0d}, causal mean μ_d, causal precision ϕ_d for disease $d = 1, \ldots, D$, we employ a more powerful gradient-based sampling scheme namely Hamiltonian Monte Carlo (also known as hybrid Monte Carlo) (HMC) [20, 29], exploiting the fact that the joint posterior of our model is differentiable with respect to the model parameters $\mu_d, \phi_d, w_{kd}, w_{0d}$ (**Supplementary Text** S1). Finally, after discarding $t\%$ models accepted before the burn-in period (default: t=20%), we obtain the Bayesian estimates of PPA by averaging the corresponding values computed over the T' individual models accepted throughout the T MCMC runs.

Bayesian fold-enrichment tests for epigenomic annotations

Due to co-linearity among the epigenomic annotations, directly using w_{kd} to assess the epigenomic enrichment for annotation k may be misleading. We propose an heuristic approach to assess the log fold-enrichment of the full prior model over the alternative prior with the effect of annotation k for disease d removed (i.e., $\mathbf{w}_{d\backslash k}, w_{kd} = 0$):

$$f_{kd} = \log \int p(\mathbf{w}_d) \frac{p(\mathbf{c}_d|\mathbf{w}_d, \mathbf{e}_v)}{p(\mathbf{c}_d|\mathbf{w}_{d\backslash k}, w_{kd} = 0, \mathbf{e}_v)} d\mathbf{w}_d \tag{16}$$

$$\approx \frac{1}{T'} \sum_{t=1}^{T'} \log \frac{1}{|\mathcal{C}_d|} \sum_{v \in \mathcal{C}_d} \frac{p(c_{vd}|\mathbf{w}_d^{(t)})}{p(c_{vd}|\mathbf{w}_{d\backslash k}^{(t)}, w_{kd}^{(t)} = 0)} \tag{17}$$

where $p(c_{vd}|\mathbf{w}_d^{(t)}, \mathbf{e}_v)$ is the logistic prior based on Eq 1, $\mathcal{C}_d$ is the union of all the 95% credible sets across loci for disease d: $\mathcal{C}_d = \bigcup_b \mathcal{C}_{bd}$. Notably, under the optional constraint that $w_{kd} \geq 0$, f_{kd} is always positive, which directly translates to fold-enrichment of annotation k conditioned on all the other annotations $k' \neq k$. The 95% Bayesian credible interval for f_{kd} are obtained from the T' MCMC runs. The significance of each annotation k is determined based on the ranking of its lower bound f_{kd} (i.e., the 2.5% quantile of f_{kd}).

Alternatively, we can estimate the fold-enrichment for each annotation simply based on the ratio of estimated fraction of causal variants in an annotation e_{vk} over the fraction of all of the variants in that annotation $\frac{\sum_v c_v e_{vk} / \sum_v c_v}{\sum_v e_{vk}/V}$, where c_v is the PPA for SNP v. This is more efficient and accurate when the underlying causal variants were randomly sampled from the annotations as done in the simulation.

Stochastic gradient updates per locus

Directly updating model parameters based on the gradients of all GWAS loci at each MCMC iteration is inefficient and results in poor HMC acceptance rate. Instead, at each MCMC update, we randomly sample one locus and update the model parameters (which are shared across loci) based on that locus. We find this approach quite efficient in capturing meaningful

causal properties such as causal signals and relevant epigenomes that are shared across all risk loci. Together, we outline the overall algorithm of the proposed Bayesian model in Algorithm S2 (**Supplementary Text** S1).

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 [18]. Briefly, the simulation can be divided into three stages (1) simulate genotypes based on the haplotypes from 1000 Genome European data (phase 1 version 3) using HapGen2 [30] (**Supplementary Fig.** S1); (2) simulate epigenomic enrichments and subsequently sample causal variants accordingly using 100 Roadmap annotations selected from each of the 19 categories of primary tissue/cell types (**Supplementary Fig.** S1); (3) simulate liability phenotype plus the random noise to obtain the desired heritability (fixed at 0.25) and subsequently the GWAS summary statistics in terms p-values and z-scores via ordinary least square regression. Details are described in **Supplementary Text**.

Gene ontology enrichment analysis

We obtained the latest gene annotations from Ensembl database (version 80) programmatically via biomaRt package [31], which resulted in 10,801 gene ontology (GO) terms in biological processes (BP). To assign SNPs to genes, we performed lift-over to map the SNPs from hg19 to hg38 using rtracklayer [32] and assigned each SNP to a gene if it is located within 35 kb up and 10 kb downstream of that gene. The resulting Ensemble gene identifiers were matched with those genes in each GO-BP category. We then performed hypergeometric tests on each GO-BP term for all of potential *in-cis* target genes of the SNPs in each trait and adjusted for multiple testings using Benjamini-Hocherg family-wise Type I error correction method [33]. For the 9 immune traits, the enrichment signals are strong so we set the cutoff at FDR < 0.005; for Schizophrenia, we set FDR < 0.2.

RiVIERA-beta software

RiVIERA-beta is available as an open-source R package with documented functions and walk-through examples described in the vignette. Most functions were implemented in C++ by integrating *Rcpp* and *RcppArmadillo* libraries [34]. These libraries enabled us to apply RiVIERA-beta to large matrices very efficiently with complied code and having much lesser memory overhead than a naïve R implementation. RiVIERA-beta is available at Github (`https://github.mit.edu/liyue/rivieraBeta`).

3 RESULTS

RiVIERA model overview

The fundamental hypothesis of our model is that non-coding disease associations are driven by disruption of regulatory elements of common activity patterns (e.g., motifs of sequence-

specific regulators), thus leading to gene expression changes and ultimately phenotypic changes at the cellular or organism level between case and control individuals. Our RiVIERA-beta Bayesian model aims to infer the probability that a given variant v is a driver for disease d by modeling the corresponding GWAS association statistic for that variant using a vector of genome-wide epigenomic annotations ($\mathbf{e}_v$). Given a set of B risk loci, the inputs to RiVIERA-beta are GWAS summary statistics in terms of p-values and a set of discrete or continuous epigenomic annotations (**Fig.** 1a). In this study, we used binary signals to ease interpretation of the functional enrichments. We train RiVIERA-beta by repeatedly sampling one locus at each iteration to efficiently learn the intrinsic (i.e., locus-independent) causal signals. **Fig.** 1b depicts RiVIERA-beta as probabilistic graphical model [35]. The observed variable of our model is the GWAS association values (in terms of p-values) a_{vd} for each variant v in each disease d. We assume that a_{vd} follows a Beta distribution with unknown mean and dispersion parameters. The effect of each annotation on each trait is learned as global annotation-by-disease weight matrix $\mathbf{w}$, which follows a D-dimensional multivariate normal distribution with zero mean and $D \times D$ disease-disease covariance Λ_w. The prior probability π_{vd} that a variant v is causal in disease d is essentially a linear combination of the weighted genomic annotations $\mathbf{e}_v$, which reflects the disease-associated active histone marks and DNA accessibility in the 127 cell types (**Materials and methods**). The outputs of the model (**Fig.** 1c) are (a) posterior probability of association (PPA) c_{vd} that variant v is causal in disease d; (b) the Bayesian fold-enrichment estimates f_{kd} based on the ratio between the full prior model with all annotations over the null prior model with all annotations except for annotation k.

Method comparison using GWAS simulation

The goal of the simulation is to evaluate the model's power to predict (1) causal variants in each locus; (2) the relevant annotations that determine which variants are causal. To this end, we simulated GWAS summary statistics based on 1000 Genome European data (Phase 1 release 3) (**Supplementary Fig.** S1) and 100 representative epigenomic annotations (**Supplementary Fig.** S1) (**Materials and methods**). We performed a series of power analyses over 500 simulation runs.

First, we examined how well the posterior probabilities were calibrated by taking the credible SNPs that contribute to 95% posterior mass inferred by each method (**Supplementary Fig.** S2). As expected, when our model assumption of single-causal variant per locus holds, we observe that our model is well calibrated (**Fig.** S2), where the 95% credible SNPs indeed correspond to approximately 95% of the causal variants. When there are more than one causal variants per locus, the 95% credible SNPs include on average 50% the true causal SNPs (**Supplementary Fig.** S2)

Because the number of variants within the credible set differs depending on the concentration of the posterior probabilities inferred by each method, we sought to control that bias by evaluating the proportion of identified causal variants as a function of the absolute number of selected variants. When the assumption of one-causal-variant-per-locus holds, we observed comparable or better performance of RiVIERA-beta compared to existing methods (**Fig.** 2). As expected, when the assumption is violated, our current model is second to PAINTOR, which is able to infer multiple causal variants per locus (**Supplementary**

Fig. S3). We also examined the correlation between the functional enrichments estimated by each method and the underlying epigenomic enrichments that were used to simulate the causal variants. The performance of the four methods are comparable with the proposed model achieving a slightly better correlation (**Supplementary Fig.** S4).

Applications to immune and psychiatric disorders

To demonstrate RiVIERA-beta in a real-world application, we used it to investigate 10 complex diseases including 9 immune diseases with summary statistics obtained from ImmunoBase [15] and Schizophrenia from Psychiatric Genomic Consortium (PGC) [21] (**Table** 1). We used 848 epigenomic annotations from ENCODE/Roadmap consortium (**Materials and methods**) to build a functional prior for each trait to aid fine-mapping and conduct cell-type specific epigenomic enrichment analyses [7]. We first applied RiVIERA-beta to the 10 traits separately to examine individual causal signals and then demonstrated RiVIERA-beta's capability to operate on the 9 immune traits and the improved detection power compared to the single-trait model.

RiVIERA-beta detected meaningful tissue-specific enhancers in test GWAS traits

We first sought to confirm the validity of the model through its ability to identify meaningful cell-types or tissues for each trait. To this end, we selected the top 5% (i.e., the top 43) of the 848 annotations for each disease based on the corresponding Bayesian estimates of the lower bounds of the 95% credible interval (**Supplementary Table** S1; **Materials and methods**). We then performed hypergeometric tests on enrichments of each of the 19 categories grouped by Roadmap consortium based on the cell types and tissues [7]. Indeed, we observed a significant enrichment for Blood & T-cell for all 9 immune disorders but not for Schizophrenia, which exhibits exclusive epigenomic enrichments in the Brain category (Hypergeometric adjusted p-values < 0.05) (**Fig.** 3a). Additionally, we also observed modest enrichments for B-cell and Thymus tissue in the 9 immune traits. We then examined the enrichment status for the 8 epigenomic marks. Indeed, enhancer marks namely H3me4me1 and/or H3K27ac are most significantly enriched among all 8 marks (q-values < 0.05). In addition, H3K4me3 associated with promoter is also enriched in most immune traits. Interestingly, we also observed a modest enrichment of H3K9me3 in Schizophrenia but not in the immune traits. We further ascertained the enrichment results by re-running RiVIERA-beta on the permuted data matrix and observed diminishment of the meaningful enrichment observed above (**Supplementary Fig.** S5).

SNPs in the credible set exhibit promising regulatory potentials

The variants in the credible set are more enriched for functional elements. Inspired by the promising tissue-specific enhancer enrichment results obtained above, we refined our RiVIERA-beta model by re-training it on the top 5% (or 43) annotations on each trait using the same GWAS data. For each locus in each trait, we then constructed 95% credible set (**Supplementary Table** S2; **Materials and methods**). On average, we were able

to construct a rather small credible set ranging from 4 to 25 SNPs per locus for the 10 traits (**Table** 1). As a comparison, we extracted the same number of SNPs with the most significant GWAS p-values from each locus. For ease of reference, we named our SNPs in the credible set as "credible SNP" and the GWAS counterpart as "GWAS SNP". Compared to GWAS SNPs, the credible SNPs exhibit substantially higher averaged placental conservation scores (phastCons46way obtained from UCSC database) across most traits (**Fig.** 4 CONS).

Moreover, the credible SNPs were significantly enriched for expression quantitative trait loci (eQTL) that are in the regulatory regions. Specifically, we obtained in total 806,847 GTEx whole-blood eQTL-SNPs (version 6) [36] and retained 122,549 and 23,973 eQTL-SNPs that overlap with transcription factor binding sites derived from 1,772 TF recognition motifs [37] and digital genomic footprints (DGF) at 6-bp resolution derived from DNaseI data in CD cells using method described in [38], respectively as well as 6,743 eQTL-SNPs that overlapped with both the TFBS and DGF regions. We then performed hypergeometric tests to assess the significance of overlap between the credible/GWAS SNPs and the regulatory-eQTL SNPs. Indeed, our credible SNPs exhibit much higher enrichments for those eQTL-SNPs, suggesting their regulatory potentials elucidated based on the enhancer activities by our proposed RiVIERA-beta model (**Fig.** 4; **Supplementary Table** S3).

Gene-centric analysis revealed enrichment for meaningful biological processes

Genes adjacent to the SNPs in credible sets are significantly enriched for disease-specific biological processes. In particular, we observed significant enrichments of many immune-related processes for the *in-cis* genes for which the SNPs in the credible set are within 35 kb upstream or 10 kb downstream (**Fig.** 5; **Supplementary Table** S4; **Materials and methods**). For instance, regulation of T cell homeostatic proliferation, regulation of interferon-gamma-mediated signaling pathway, and regulation of type I interferon-mediated signal pathways are among the most significantly enriched GO terms in 5 or 6 out of the 9 immune traits. In contrast, the enrichments for Schizophrenia are dominated by GO terms involving synaptic processes and neuronal differentiation/development. The enrichment results are mostly consistent between the credible genes and the genes derived from the same number of SNPs chosen based on the GWAS p-values (GWAS-genes).

Intriguingly, we observed a highly significant enrichment for keratinization (GO:0031424) and epidermis (e.g., skin) development (GO:0008544) exclusively for Psoriasis. In particular, 17 genes among the 241 credible genes belong to keratinization and epidermis development, which contain in total 49 and 121 genes, respectively ($q < 9 \times 10^{-18}$, $q < 2 \times 10^{-10}$). Indeed, Psoriasis is mainly characterized as a chronic skin disease with epidermal hyper-proliferation [39, 40]. In contrast, there are only 6 out of 157 GWAS-genes are defined in each of two GO categories ($q < 0.001$).

To further ascertain the RiVIERA-beta fine-mapping results, we created a visualization scheme for each of the 469 risk loci across 10 traits examined (**Supplementary Fig.** S6). **Fig.** 6 displays two example loci for Type 1 diabetes (chr17: 37383069-38239012) and Schizophrenia (chr7: 104598324-105062839). The upper panel displays the RiVIERA-beta model prior, the genetic signals from GWAS -log p-values, and RiVIERA-beta PPA. Red colored

and diamond shape points are GTEx whole-blood eQTL SNP and top SNPs included into 60% credible set (we used 60% to not clutter the plot with the remaining SNPs in the 95% credible set that exhibit low PPA). Intuitively, SNPs with high PPA exhibit both prominent genetic and epigenetic signals. Thus, to infer causal variants, RiVIERA-beta efficiently took into account not only the GWAS signals derived from the genetic data but also the prior signals mainly driven by the weighted epigenomic profiles. The middle panel illustrates the cumulative density for each epigenomic profiles weighted by the tissue-specific enrichment estimates.

Consistent with the overall enrichment results (**Fig.** 3), we observed prominent enrichments for the enhancer regions predominantly in blood T cells for all of the 9 immune traits and brain tissue for Schizophrenia. The bottom tracks display the external functional information (i.e., not in the training data) including conservation score, genes, transcription factor binding sites based on motif matches that may further aid variant selection for downstream experimental validation (please refer to **Supplementary Table** S2 for detailed information). We also visualized the signals within the of Psoriasis-associated risk region ch1:152536784-152785170, which harbors genes involved in keratinization and epidermis development as mentioned above. Interestingly, as an exception of most other immune-susceptible loci, the underlying epigenomic profiles exhibit prominent signals not only in blood T cell but also in epithelia enhancer regions (**Supplementary Fig.** S6). However, the associated SNPs exhibit rather weak genetic signal perhaps due to lower allele frequencies.

Multi-trait causal inference improved functional enrichments in most immune traits

Exploiting epigenomic correlation between highly related immune diseases improved functional enrichments in several traits. We performed multi-trait causal inference over all of the 9 autoimmune traits by jointly applying our RiVIERA-beta to 364 risk loci concatenated together from the 9 immune traits using 174 epigenomic annotations which was a union of unique annotations from the top 43 annotations for each individual trait. We focused only on the 9 immune GWAS (i.e., leaving out Schizophrenia) because they commonly utilized the same genotyping platform namely ImmunoChip. The multi-trait GWAS summary statistics triggered RiVIERA-beta to infer the disease covariance matrix and sample disease-specific epigenomic weights from the joint posterior with modified zero-mean multivariate normal prior that takes into account the sampled disease covariance (**Materials and methods**). As a results, RiVIERA-beta sampled correlated epigenomic weights between traits more frequently compared to the single-trait model.

We constructed the 95% credible sets for each trait using the disease-specific PPA derived from the joint model and assessed the functional enrichments as above (**Supplementary Table** S6). Notably, the cross-trait model exploited 174 annotations as apposed to 43 annotations used by the single-trait model. To examine whether the improved enrichments were due to the increased number of annotations, we re-ran a single-trait model for each of the 9 traits separately each using the 174 annotations. Compared to the 95% credible set constructed based on the single-trait causal inference using the top 43 annotations, we observed smaller 95% credible sets for 8 out of the 9 immune traits (**Table** 1), suggesting

that the mulit-trait joint inference further improved the model confidence in some of the highly related traits.

More importantly, we observed a much more improved enrichments for the GTEx whole-blood eQTL SNPs located within open chromatin regions or digital genomic footprints in most of the immune traits (**Fig.** 7; **Supplementary Table** S5). Thus, the joint inference further improved the regulatory potential through following the Hamiltonian trajectory that is more consistent with the epigenomic correlation pattern between the related immune traits. We also repeated the GO enrichment analysis on the 95% credible set and found that the enriched GO terms were mostly immune-specific biological processes and consistent with the above single-trait analyses (**Supplementary Fig.** S7; **Supplementary Table** S7).

4 DISCUSSION

Understanding the genetic basis of complex traits hinges upon powerful integrative methods to map genotypes to phenotypes [41]. Fine-mapping causal variants has been a highly active and fruitful research in the past several years [9, 18, 42–44]. However, most existing methods typically operate solely on genetic data by estimating each SNP of being causal conditioned on the lead index SNPs in the same LD block, which are typically approximated by the 1000 Genome data [9, 15, 45, 46]. With the recent availability of large-scale functional genomic data provide by ENCODE/Roadmap consortia, there is an urgent need to incorporate these valuable information in a principled way as a form of Bayesian prior. In this article, we describe a novel Bayesian fine-mapping method RiVIERA-beta to re-prioritize GWAS summary statistics based on their epigenomic contexts. The main contribution of RiVIERA-beta is the ability to systematically construct from a targeted set of susceptible loci a Bayesian credible set of SNPs, which exhibit plausible tissue-specific regulatory properties implicated in the large epigenomic data compendium either in a single trait or across multiple traits.

One benefit of using the raw epigenomic annotations rather than using the inferred signals such as ChromHMM [7] or GenoSkyline [47] states derived from the annotations is that it eases the interpretation of the actual relevant epigenomic marks in the relevant tissue types and facilitates downstream experimental efforts to assay the specific marks in the disease-specific cell types. However, the correlation of the epigenomic marks will make difficult estimating the underlying functional enrichments. Moreover, we choose to model the summary statistics rather than genotypes because it is not always possible to obtain individual-level phenotype-genotype data particularly for large-scale meta-analysis. Thus, effective methods based on summary statistics may benefit wider research communities than methods that solely operate on individual-level genotype data [18, 19, 23]. Moreover, our model *requires only p-values* because it uses Beta distribution to model the likelihood. In contrast, fgwas requires both the z-scores and the standard error from the linear regression used in the GWAS to estimate the Wakefield approximate Bayes factors. While some recent GWAS summary statistics include those information, there are many do not have z-scores and/or standard error of the linear model but only p-values (e.g., the ImmunoChip data we used in our studies for the 9 immune traits). When the standard error is not available in a given GWAS summary statistics, fgwas needs to be estimate it from the minor allele

frequency of a reference panel such as 1000 Genome, which may not be accurate depending the study cohorts. Additionally, modeling p-values via Beta density function only has more relaxing model assumption than modeling z-scores via normal density although both methods are highly effective in practice.

Overall, SNPs included into the credible set exhibit both significant GWAS signal and high prior. In some cases, however, SNPs that were added to the credible set in each locus do not exhibit significant GWAS p-values (**Supplementary Table** S2,S6). This generally occurs when the genetic signals in those loci are weak relative to the SNPs in other loci for the same trait, and the model functional prior eventually dominates the SNP prioritization. Thus, we recommend considering these variants cautiously when designing downstream experiments.

One important assumption of our model is that there is one causal variant per locus, which is reflected by the normalization of variants within each locus so that they sum to 1 in order to obtain PPA and construct 95% credible sets [23]. When this assumption holds, the posterior probabilities are well calibrated (**Supplementary Fig.** S2). However, as demonstrated in our simulation, when this assumption is violated, the PPA is not well calibrated (**Supplementary Fig.** S2,S3). Other existing method such as PAINTOR [18] and CAVIAR [48] employ multivariate normal distribution to model all of the variants within a locus using LD reference panel estimated from 1000 Genome data as the covariance matrix, which allows inferring more than one causal variants per locus. While CAVIER used only summary statistics, PAINTOR is able to employ functional annotations to aid fine-mapping. Both methods require computing the likelihood density across a combinatorial set of causal configurations and therefore still needs to assume at most an arbitrarily small number of causal variants, typically below 10 causal SNPs per locus.

As future works, we will explore potential ways to enable efficient inference of more than one causal variants per locus. Furthermore, we will also explore the potential gain of incorporating trans-ethnic data, which was effectively demonstrated by the trans-ethnic version of the PAINTOR model [49]. Moreover, in addition to modeling the epigenomic correlation between traits, variant prioritization may further benefit by jointly inferring the comorbidity at the individual SNP level [19], gene level [50], and/or pathway level [17]. Together, we believe that RiVIERA-beta will serve as a valuable tool complementary to the existing methods in identifying novel risk variants through tissue-specific epigenome-aware fine-mapping as well as aiding the selection of the appropriate cell types and epigenomic marks for more focused investigations of the disruptions of chromatin states by the disease-specific causal variants.

5 ACKNOWLEDGEMENTS

We thank Yongjin Park, Gerald Quon, Abhishek Sarkar, and Zhizhuo Zhang for the helpful discussions.

Supplementary Data

Supplementary Data are available at NAR Online: Supplementary Text S1,S2; Supplementary Tables S1-S5; Supplementary Figures S1-S6.

6 FUNDING

National Institutes of Health (NIH) [R01-HG004037, RC1-HG005334, R01-HG008155]. Funding for open access charge: NIH [R01 HG004037].

6.0.1 Conflict of interest statement.

None declared.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

15. 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).

16. 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).

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

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

19. 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).

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

21. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

22. Kwan, J. S., Li, M.-X., Deng, J.-E. & Sham, P. C. Fapi: Fast and accurate p-value imputation for genome-wide association study. *European Journal of Human Genetics* **NA** (2015).

23. 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).

24. Ferrari, S. & Cribari-Neto, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**, 799–815 (2004).

25. Bayes, C. L., Bazán, J. L. & García, C. A New Robust Regression Model for Proportions. *Bayesian Analysis* **7**, 841–866 (2012). URL `http://projecteuclid.org/euclid.ba/1354024464`.

26. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature reviews Genetics* **10**, 681–690 (2009).

27. 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).

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

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

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

31. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature protocols* **4**, 1184–1191 (2009).

32. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

33. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995).

34. Eddelbuettel, D. *et al.* Rcpp: Seamless r and c++ integration. *Journal of Statistical Software* **40**, 1–18 (2011).

35. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).

36. 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`.

37. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in encode tf binding experiments. *Nucleic acids research* **42**, 2976–2987 (2014).

38. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).

39. of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, G. A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between hla-c and erap1. *Nature genetics* **42**, 985–990 (2010).

40. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics* **44**, 1341–1348 (2012).

41. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics* **16**, 85–97 (2015).

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

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

44. Wallace, C. *et al.* Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. *PLoS Genet* **11**, e1005272 (2015).

45. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

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

47. Lu, Q., Powles, R. L., Wang, Q., He, B. J. & Zhao, H. Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. *PLoS Genetics* **12**, e1005947 (2016).

48. Chen, W. *et al.* Fine mapping causal variants with an approximate bayesian method using marginal test statistics. *Genetics* **NA**, genetics–115 (2015).

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

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

Figure Legends

Figure 1: RiVIERA-beta model overview. **(a)** Inputs to RiVIERA-beta are GWAS summary statistics and epigenomic annotations for B risk loci. At a given iteration, the model samples one locus and tries to learn the intrinsic causal signals implicated in the corresponding GWAS summary data and epigenomic profiles. Highlighted variant is the causal variant based on the simulated data. **(b)** The probabilistic graphical model representation of RiVIERA-beta [35]. Variables for which distribution is defined are in circle. Epigenomic profiles are treated as observed values with no circle. The variable in shaded circle are observed (i.e., GWAS association a_{vd} and variables in unshaded circle are unobserved. The variables in red are observed and variables in blue are the variables of interest (i.e., causal indicator). The two colored plates represent K annotations (red) and V variants (blue). We model the GWAS association a_{vd} of variant v in terms of p-value sampled from Beta distribution with unknown precision ϕ_d and mean μ_d, which respectively follow an uninformative prior and a Beta distribution with hyperparameters μ_0, ϕ_0. The latent variable c_{vd} indicates whether variant v is causal in disease d. On top of it, we dedicate an empirical prior as a linear combination of the epigenomic profile e_{vk} weighted by the epigenomic influence w_{kd}, which follows multivariate normal with zero mean and a $D \times D$ inverse covariance or precision matrix Λ_w^{-1}, where D is the number of traits that are being modeled. The linear bias w_{0d} expresses the prior belief of the causal fraction π_0 (default: 0.01). **(c)** Outputs from the model are posterior probabilities of association (PPA) for each variant in each locus, the 95% credible set containing the minimal number of SNPs whose PPA sum to 0.95 or greater, and the Bayesian estimates of the fold-enrichment of each annotation.

Figure 2: Model performance on simulated datasets. Proportion of causal variants were identified by each method as a function of increasing number of top variants selected.

Figure 3: Predicted tissue-specific epigenomic enrichments in the 10 GWAS traits. **(a)**. Hypergeometric enrichment for each of the 19 primary tissue categories using the top 5% or 43 annotations out of the 848 annotations in total for each trait based on the lower bound of the 95% credible interval of the Bayesian fold-enrichment estimates by our RiVIERA-beta; model; **(b)** enrichments for the 8 epigenomic marks among the top 43 annotations for each trait. Y-axis is the logarithmic q-values, which are the corrected p-values from the hypergeometric tests for multiple testing across traits and tissue groups or marks by Benjamini-Hochberg method [33]. On both plots, horizontal dashed bars indicate standard statistical threshold of FDR < 0.05.

Figure 4: Functional enrichments of credible SNPs. The top left panel displays the averaged phastCons46way conservation scores for variants in the 95% credible set (cred_snp) and the same number of SNPs chosen based on GWAS p-values (gwas_snp). The three other panels illustrate hypergeometric enrichments in terms of the -log10 q-values corrected for multiple testing over the 10 traits of the selected variants for GTEx whole blood eQTL located within transcription factor binding sites based on sequence motif (TFBS) (eQTL+TFBS) and genomic digital footprint (DGF) (eQLT+DGF), and eQTL in both TFBS and DGF (eQTL+TFBS+DGF).

Figure 5: Gene ontology enrichments across the 10 traits. Rows are the GO biological processes and columns are the 10 traits. Color intensities in each cell reflect the significance level in terms of -log10 p-value. Asterisks indicate q-values above significant cutoff after correcting for multiple testings (FDR < 0.2). GO names that match the pattern 'synap|neuro|nerve' are colored blue to highlight their exclusive association with 'Schizophrenia' (also in blue). Notably, GO terms 'keratinization' and 'epidermis development' (highlighted in the red box) are exclusively enriched for Psoriasis. Diseases were ordered based on hierarchical clustering based on the Pearson correlation of their GO enrichment scores.

Figure 6: Visualization of fine-mapping results. Top track: the upper panel display the RiVIERA-beta prior, genetic signals of GWAS -log10 p-values, and RiVIERA-beta PPA; the middle track illustrates the cumulative density of weighted epigenomic profiles colored based on the epigenomic group; the bottom tracks shows the conservation, gene annotations (Gencode 19), transcription factor binding sites (TFBS), and SNP positions. The red colored and bigger diamond plots indicate whole-blood GTEx eQTL SNPs and SNPs included into the 60% credible set, respectively. For illustration purpose, only one risk locus for Type 1 diabetes and one for Schizophrenia are shown above. The full display of 469 risk loci were in **Supplementary Fig.** S6.

Figure 7: Enrichments for eQLT using credible SNPs constructed from multi-trait joint inference. Credible SNPs for each trait were constructed based on PPA inferred by the joint RiVIERA-beta model over the 9 immune traits using 174 annotations, which are the union of the top 43 annotations detected from each trait individually. We then assessed the hypergeometric enrichments of the 95% credible sets for the GTEx whole-blood eQTL that are within DNA hypersensitive sites as defined by the genomic digital footprint data [38]. We compared these enrichment scores derived from the multi-trait model (cred_snp_mt) to the enrichments derived from the single-trait models either running on 43 annotations (cred_snp_st43) or on the 174 annotations (cred_snp_st174). The latter was included to control for the improvements due to the increased number of annotations (from 43 to 174).

Tables

Table 1: GWAS data summary

Abbrev	Trait	Total	Loci	gwSNPs	cSNP_st	cSNPs_mt
ATD	Autoimmune Thyroid Disease	4206	8	630	38	49
CEL	Celiac Disease	29784	39	2592	344	211
JIA	Juvenile Idiopathic Arthritis	13427	22	3	440	223
MS	Multiple Sclerosis	61360	104	2096	884	339
NAR	Narcolepsy	1316	3	62	22	16
PBC	Primary Biliary Cirrhosis	14573	19	2498	172	111
PSO	Psoriasis	24832	34	457	305	171
RA	Rheumatoid Arthritis	38207	78	470	1978	719
SCZ2	Schizophrenia	54132	105	5217	2481	NA
T1D	Type 1 Diabetes	41945	57	2832	826	327

We investigated 10 GWAS traits as listed above. Abbrev: abbreviation of the trait names; Total: total number of SNPs in the risk loci with imputed and observed summary statistics; Loci: total number of risk loci for each trait; gwSNPs: SNPs that pass GWAS cutoff p $<$ 5e-8; cSNP_st: total number of SNPs that are included into the 95% credible set based on single-trait risk inference using RiVIERA-beta; cSNP_mt: SNPs in 95% credible set constructed based on multi-trait joint risk inference using RiVIERA-betaacross the 9 immune traits (without SCZ2).

Figure 1

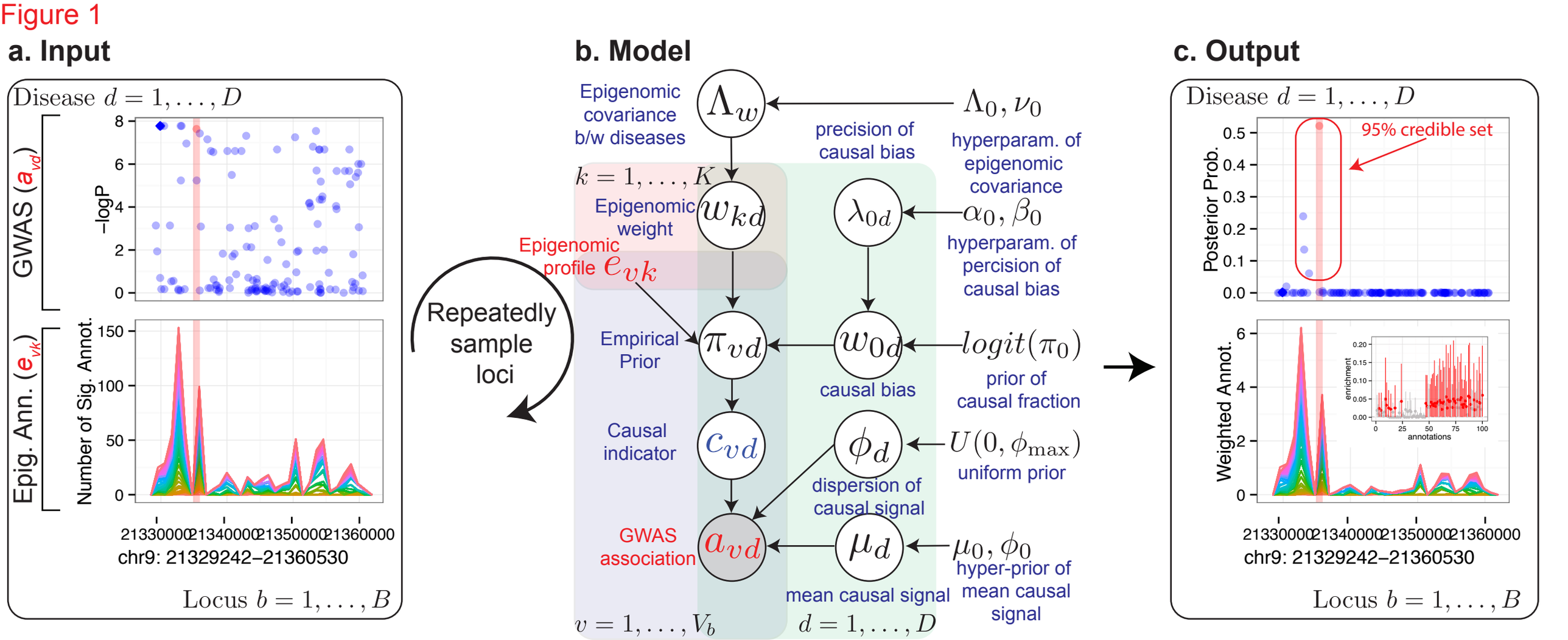

a. Input
Disease $d = 1, \ldots, D$
GWAS (a_{vd})
−logP
Epig. Ann. (e_{vk})
Number of Sig. Annot.
21330000 21340000 21350000 21360000
chr9: 21329242–21360530
Locus $b = 1, \ldots, B$
Repeatedly sample loci
b. Model
Epigenomic covariance b/w diseases
Λ_w
Λ_0, ν_0
precision of causal bias
hyperparam. of epigenomic covariance
$k = 1, \ldots, K$
Epigenomic weight
w_{kd}
λ_{0d}
α_0, β_0
hyperparam. of percision of causal bias
Epigenomic profile e_{vk}
Empirical Prior
π_{vd}
w_{0d}
$logit(\pi_0)$
causal bias
prior of causal fraction
Causal indicator
c_{vd}
ϕ_d
$U(0, \phi_{\max})$
uniform prior
dispersion of causal signal
GWAS association
a_{vd}
μ_d
μ_0, ϕ_0
hyper-prior of mean causal signal
mean causal signal
$v = 1, \ldots, V_b$
$d = 1, \ldots, D$
c. Output
Disease $d = 1, \ldots, D$
Posterior Prob.
95% credible set
Weighted Annot.
enrichment
annotations
21330000 21340000 21350000 21360000
chr9: 21329242–21360530
Locus $b = 1, \ldots, B$

Figure 2

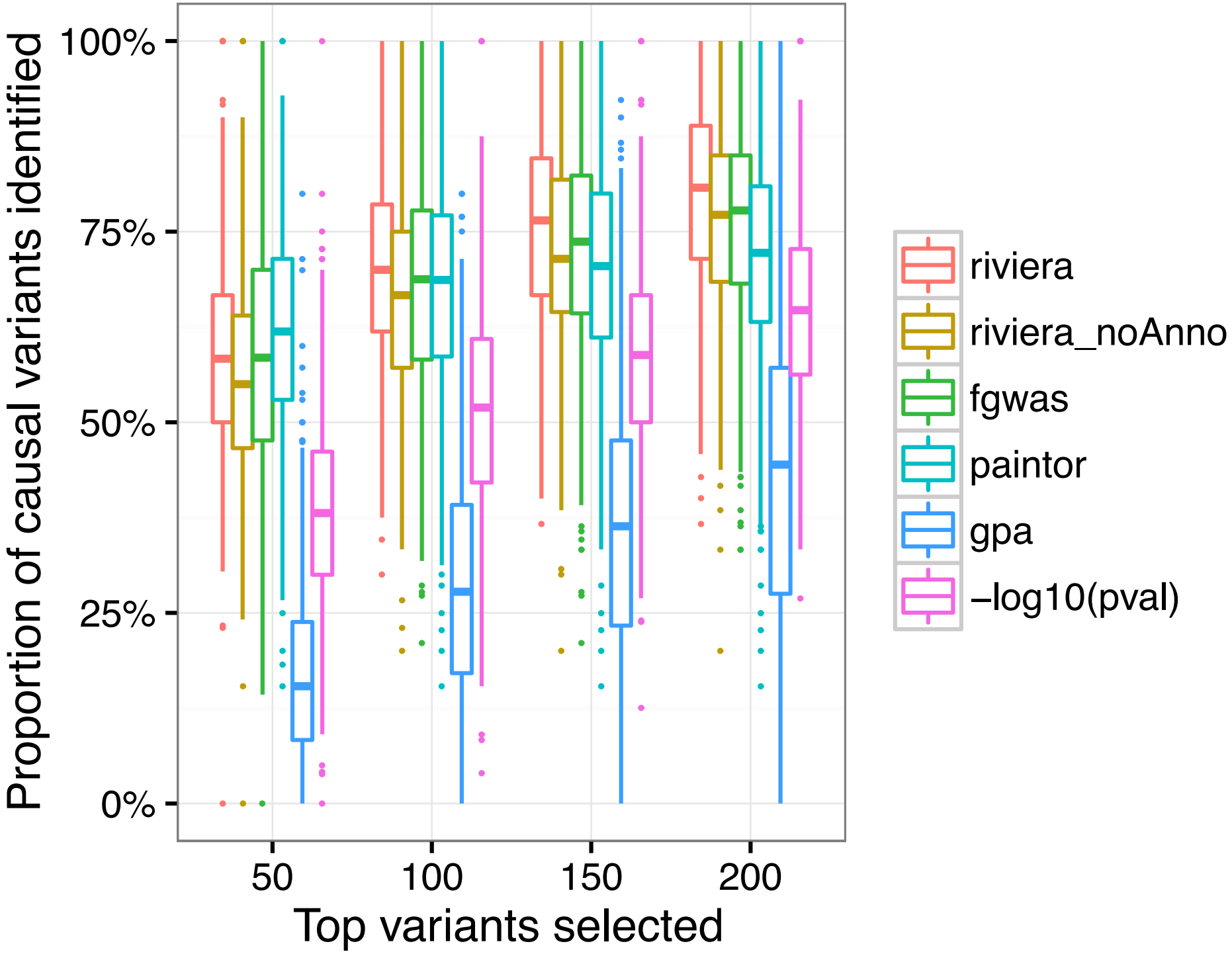

Proportion of causal variants identified
100%
75%
50%
25%
0%
50
100
150
200
Top variants selected
riviera
riviera_noAnno
fgwas
paintor
gpa
−log10(pval)

Figure 3

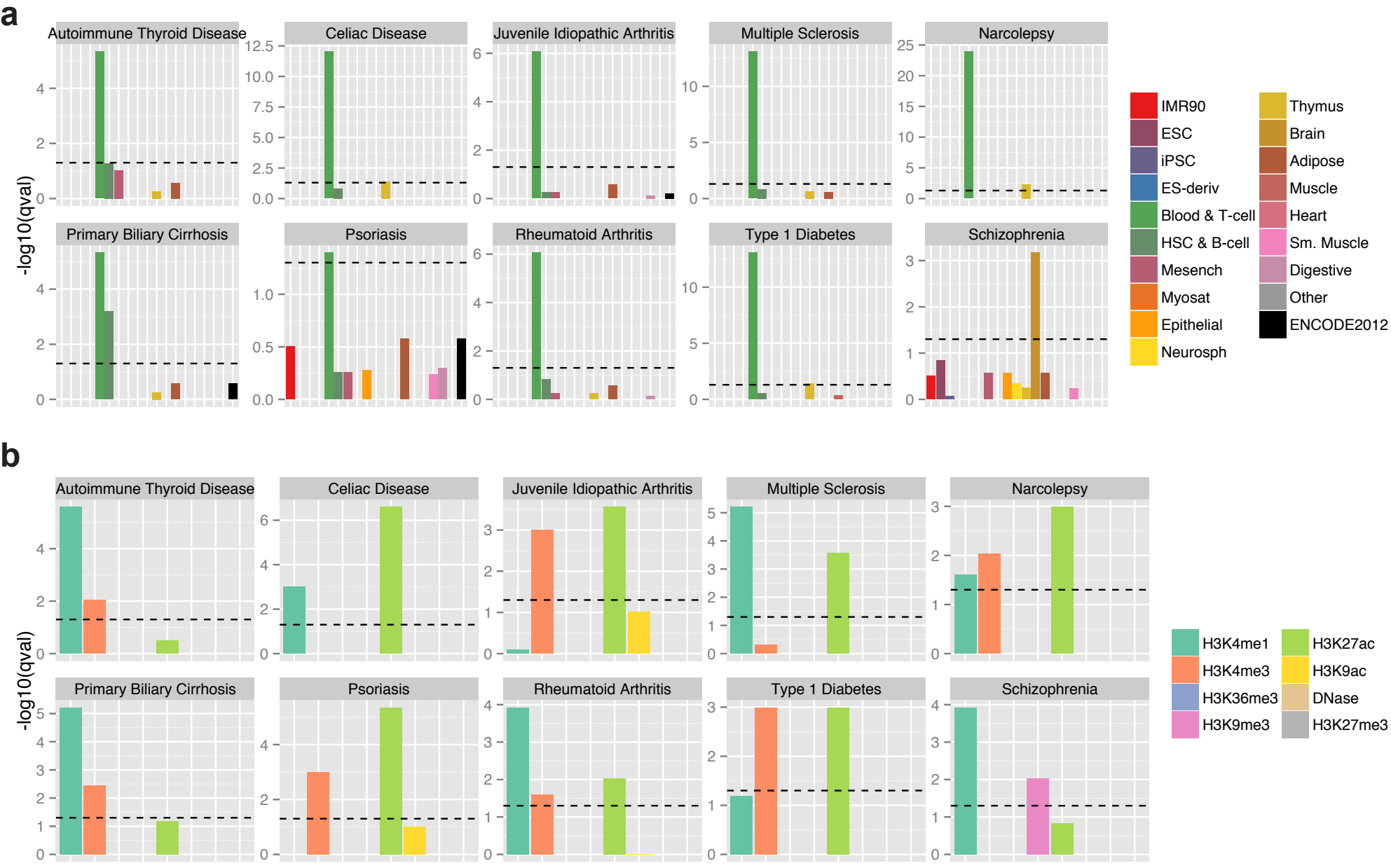

a
Autoimmune Thyroid Disease
Celiac Disease
Juvenile Idiopathic Arthritis
Multiple Sclerosis
Narcolepsy
Primary Biliary Cirrhosis
Psoriasis
Rheumatoid Arthritis
Type 1 Diabetes
Schizophrenia
-log10(qval)
IMR90
ESC
iPSC
ES-deriv
Blood & T-cell
HSC & B-cell
Mesench
Myosat
Epithelial
Neurosph
Thymus
Brain
Adipose
Muscle
Heart
Sm. Muscle
Digestive
Other
ENCODE2012
b
Autoimmune Thyroid Disease
Celiac Disease
Juvenile Idiopathic Arthritis
Multiple Sclerosis
Narcolepsy
Primary Biliary Cirrhosis
Psoriasis
Rheumatoid Arthritis
Type 1 Diabetes
Schizophrenia
-log10(qval)
H3K4me1
H3K4me3
H3K36me3
H3K9me3
H3K27ac
H3K9ac
DNase
H3K27me3

Figure 4

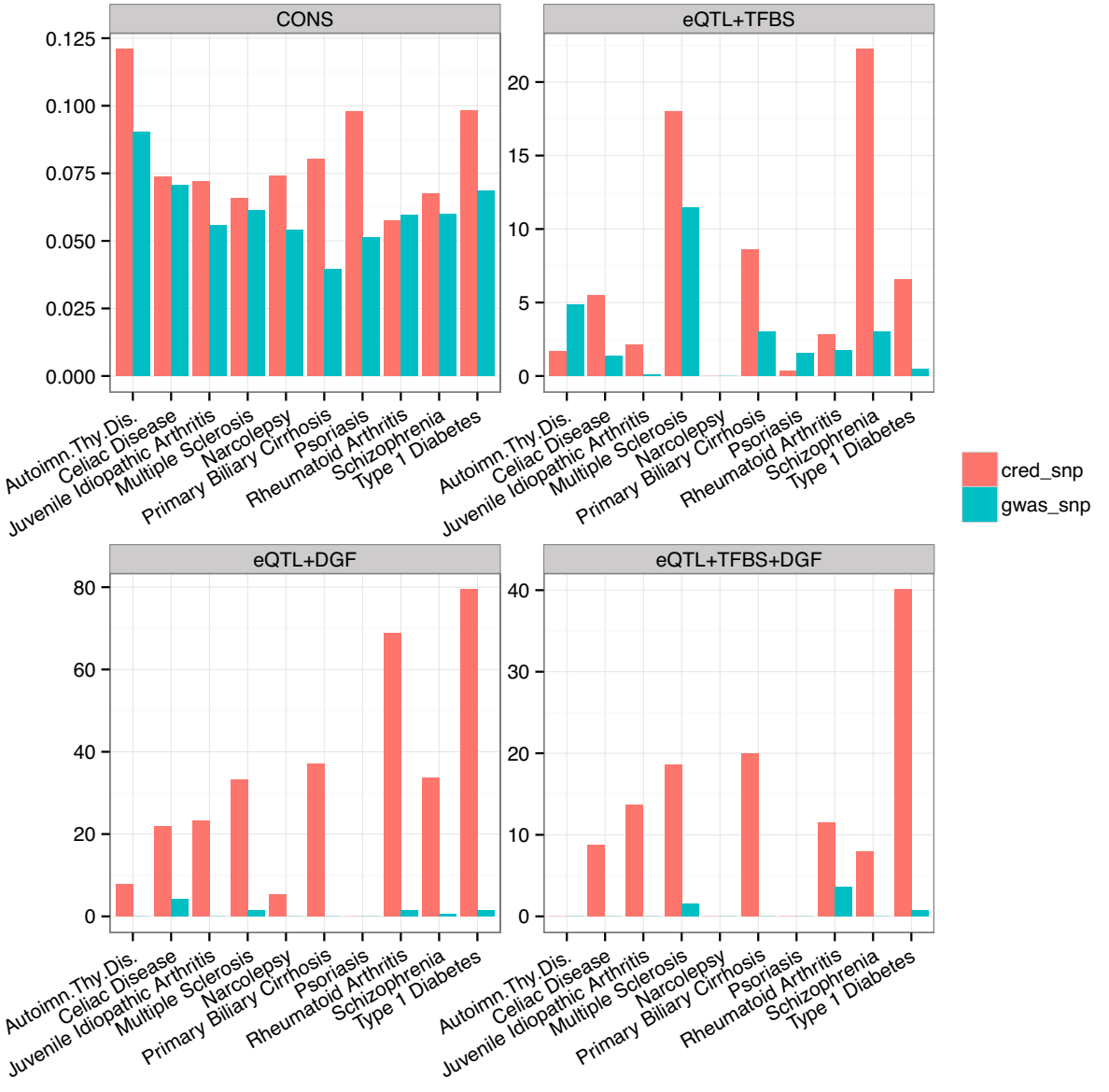

CONS
eQTL+TFBS
eQTL+DGF
eQTL+TFBS+DGF
0.125
0.100
0.075
0.050
0.025
0.000
20
15
10
5
0
80
60
40
20
0
40
30
20
10
0
Autoimn.Thy.Dis.
Celiac Disease
Juvenile Idiopathic Arthritis
Multiple Sclerosis
Narcolepsy
Primary Biliary Cirrhosis
Psoriasis
Rheumatoid Arthritis
Schizophrenia
Type 1 Diabetes
cred_snp
gwas_snp

Figure 5

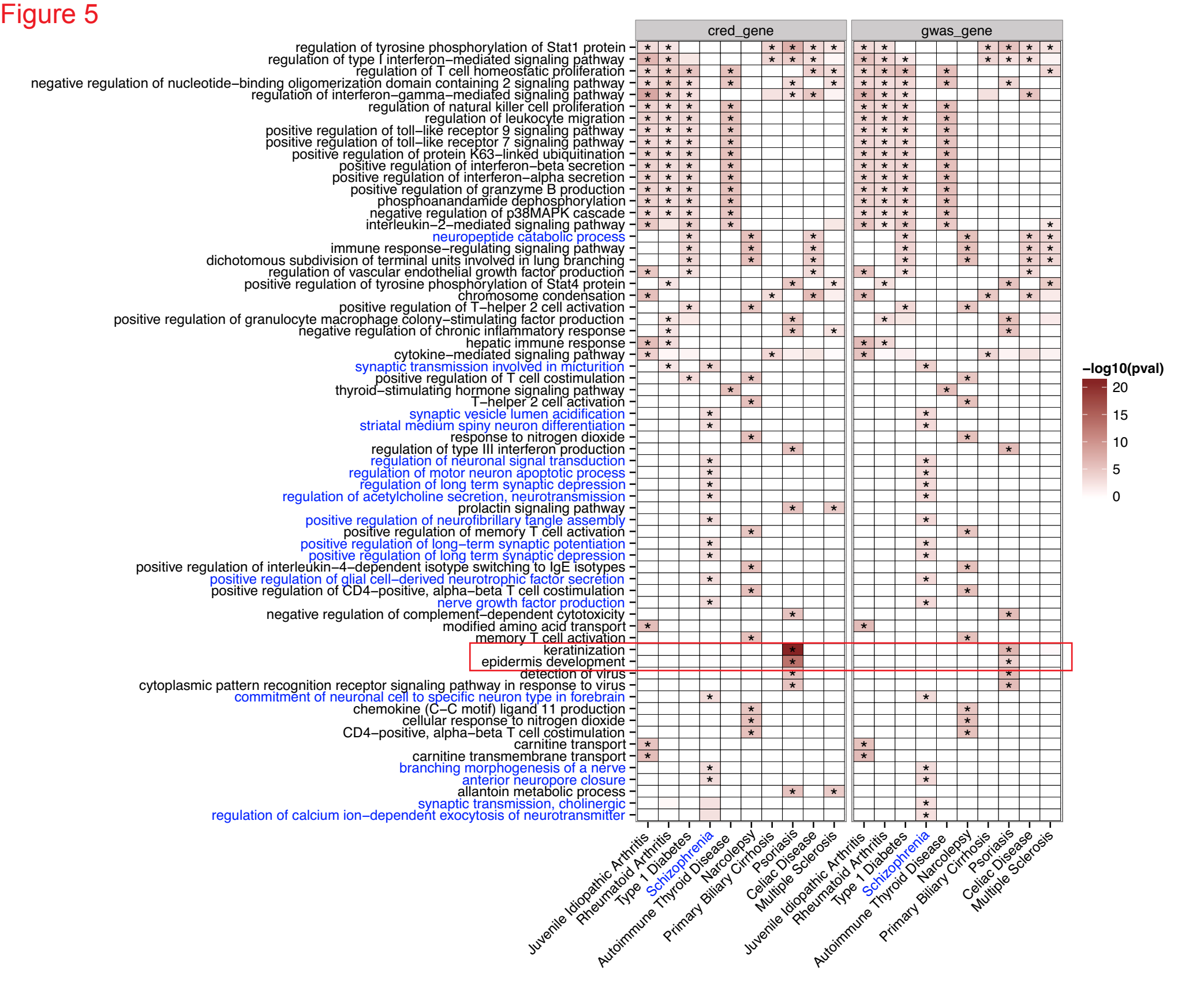

cred_gene
gwas_gene
regulation of tyrosine phosphorylation of Stat1 protein
regulation of type I interferon−mediated signaling pathway
regulation of T cell homeostatic proliferation
negative regulation of nucleotide−binding oligomerization domain containing 2 signaling pathway
regulation of interferon−gamma−mediated signaling pathway
regulation of natural killer cell proliferation
regulation of leukocyte migration
positive regulation of toll−like receptor 9 signaling pathway
positive regulation of toll−like receptor 7 signaling pathway
positive regulation of protein K63−linked ubiquitination
positive regulation of interferon−beta secretion
positive regulation of interferon−alpha secretion
positive regulation of granzyme B production
phosphoanandamide dephosphorylation
negative regulation of p38MAPK cascade
interleukin−2−mediated signaling pathway
neuropeptide catabolic process
immune response−regulating signaling pathway
dichotomous subdivision of terminal units involved in lung branching
regulation of vascular endothelial growth factor production
positive regulation of tyrosine phosphorylation of Stat4 protein
chromosome condensation
positive regulation of T−helper 2 cell activation
positive regulation of granulocyte macrophage colony−stimulating factor production
negative regulation of chronic inflammatory response
hepatic immune response
cytokine−mediated signaling pathway
synaptic transmission involved in micturition
positive regulation of T cell costimulation
thyroid−stimulating hormone signaling pathway
T−helper 2 cell activation
synaptic vesicle lumen acidification
striatal medium spiny neuron differentiation
response to nitrogen dioxide
regulation of type III interferon production
regulation of neuronal signal transduction
regulation of motor neuron apoptotic process
regulation of long term synaptic depression
regulation of acetylcholine secretion, neurotransmission
prolactin signaling pathway
positive regulation of neurofibrillary tangle assembly
positive regulation of memory T cell activation
positive regulation of long−term synaptic potentiation
positive regulation of long term synaptic depression
positive regulation of interleukin−4−dependent isotype switching to IgE isotypes
positive regulation of glial cell−derived neurotrophic factor secretion
positive regulation of CD4−positive, alpha−beta T cell costimulation
nerve growth factor production
negative regulation of complement−dependent cytotoxicity
modified amino acid transport
memory T cell activation
keratinization
epidermis development
detection of virus
cytoplasmic pattern recognition receptor signaling pathway in response to virus
commitment of neuronal cell to specific neuron type in forebrain
chemokine (C−C motif) ligand 11 production
cellular response to nitrogen dioxide
CD4−positive, alpha−beta T cell costimulation
carnitine transport
carnitine transmembrane transport
branching morphogenesis of a nerve
anterior neuropore closure
allantoin metabolic process
synaptic transmission, cholinergic
regulation of calcium ion−dependent exocytosis of neurotransmitter
Juvenile Idiopathic Arthritis
Rheumatoid Arthritis
Type 1 Diabetes
Schizophrenia
Autoimmune Thyroid Disease
Narcolepsy
Primary Biliary Cirrhosis
Psoriasis
Celiac Disease
Multiple Sclerosis
−log10(pval)
20
15
10
5
0

Figure 6

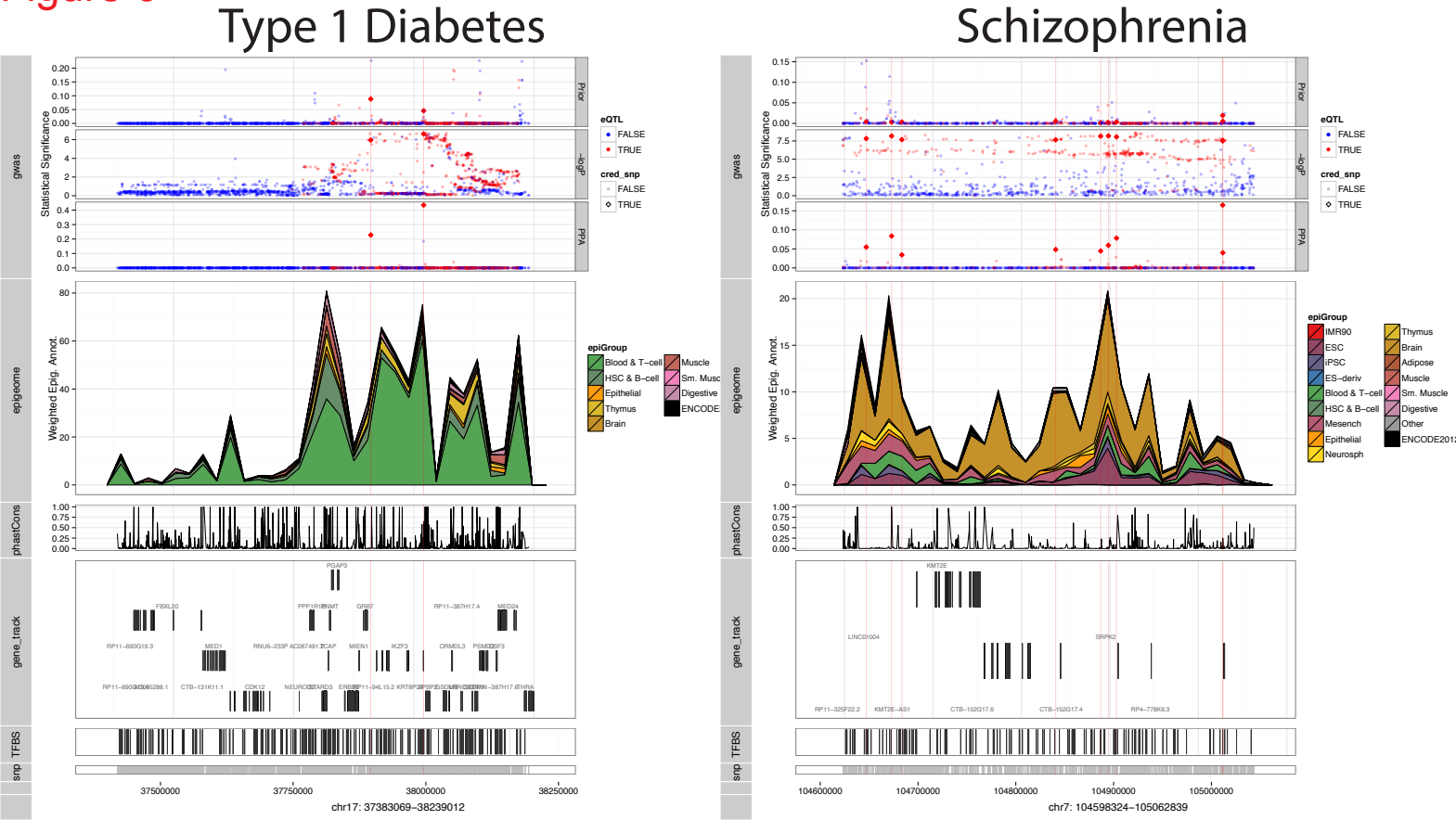
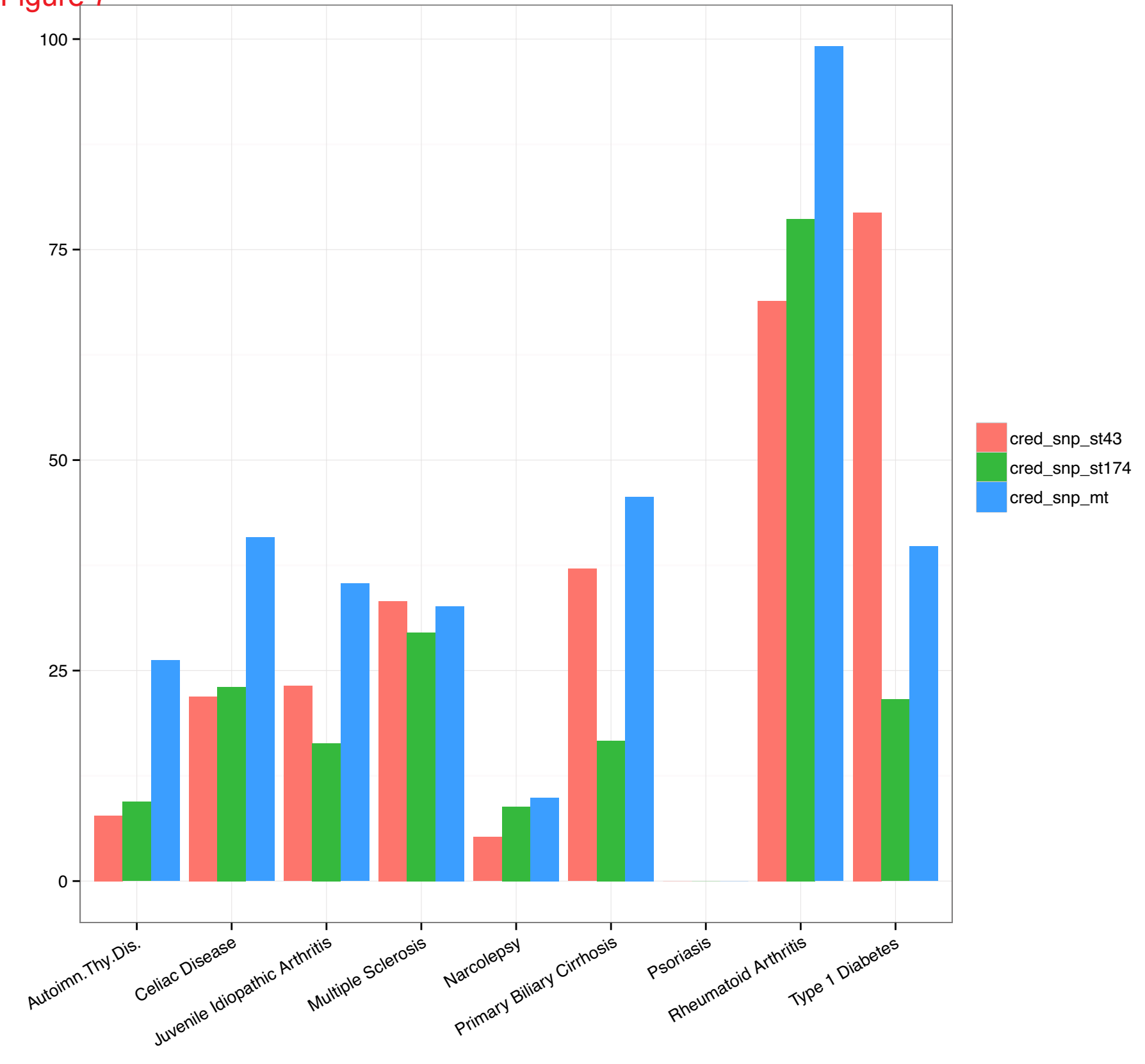

Type 1 Diabetes
Schizophrenia
gwas
epigeome
phastCons
gene_track
TFBS
snp
Statistical Significance
Prior
−logP
PPA
eQTL
FALSE
TRUE
cred_snp
FALSE
TRUE
Weighted Epig. Annot.
epiGroup
Blood & T−cell
HSC & B−cell
Epithelial
Thymus
Brain
Muscle
Sm. Musc
Digestive
ENCODE
IMR90
ESC
iPSC
ES−deriv
Blood & T−cell
HSC & B−cell
Mesench
Epithelial
Neurosph
Thymus
Brain
Adipose
Muscle
Sm. Muscle
Digestive
Other
ENCODE2012
chr17: 37383069−38239012
chr7: 104598324−105062839

Figure 7


100
75
50
25
0
cred_snp_st43
cred_snp_st174
cred_snp_mt
Autoimn.Thy.Dis.
Celiac Disease
Juvenile Idiopathic Arthritis
Multiple Sclerosis
Narcolepsy
Primary Biliary Cirrhosis
Psoriasis
Rheumatoid Arthritis
Type 1 Diabetes