bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Leveraging supervised learning for functionally-informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs

Qingbo S. Wang[1,2,3], David R. Kelley[4], Jacob Ulirsch[1,2,5], Masahiro Kanai[1,2,3,6], Shuvom Sadhuka[1,7], Ran Cui[1,2], Carlos Albors[1,2], Nathan Cheng[1,2], Yukinori Okada[6,8,9], The Biobank Japan Project[10], Francois Aguet[1], Kristin G. Ardlie[1], Daniel G. MacArthur[11,12], and Hilary K. Finucane[1,2*]

[1]Broad Institute of MIT and Harvard; [2]Analytic and Translational Genetics Unit, Massachusetts General Hospital; [3]PhD program in Bioinformatics and Integrative Genomics, Harvard Medical School; [4]Calico Life Sciences; [5]PhD program in Biological and Biomedical Sciences, Harvard Medical School; [6]Department of Statistical Genetics, Osaka University Graduate School of Medicine; [7]Harvard College, [8]Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University; [9]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University; [10]Institute of Medical Science, The University of Tokyo; [11]Centre for Population Genomics, Garvan Institute of Medical Research; [12]Centre for Population Genomics, Murdoch Children's Research Institute

Abstract

The large majority of variants identified by GWAS are non-coding, motivating detailed characterization of the function of non-coding variants. Experimental methods to assess variants' effect on gene expressions in native chromatin context via direct perturbation are low-throughput. Existing high-throughput computational predictors thus have lacked large gold standard sets of regulatory variants for training and validation. Here, we leverage a set of 14,807 putative causal eQTLs in humans obtained through statistical fine-mapping, and we use 6,121 features to directly train a predictor of whether a variant modifies nearby gene expression. We call the resulting prediction the expression modifier score (EMS). We validate EMS by comparing its ability to prioritize functional variants with other major scores. We then use EMS as a prior for statistical fine-mapping of eQTLs to identify an additional 20,913 putatively causal eQTLs, and we incorporate EMS into co-localization analysis to identify 310 additional candidate genes across UK Biobank phenotypes.

Introduction

Although genome wide association studies (GWAS) have identified large numbers of loci associated with complex traits[1,2], identifying the underlying biological mechanisms is often difficult. Two particular challenges are that (1) the majority of the associated variants are in non-coding regions[1], and (2) the association signals from GWAS studies typically contain a large number of variants in linkage disequilibrium (LD)[3]. Interpreting associations in GWAS to identify the underlying causal mechanisms requires an understanding of the function of non-coding variants at single variant resolution.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Many approaches to characterize non-coding variants exist. Large-scale consortium studies[4,5] have provided a map of functional and regulatory elements across the genome in different cell types that are enriched in various trait heritability[6-10]. Reporter assays have been powerful tools to test variant effects in cellular contexts, but typical high-throughput massive parallel reporter assays (MPRAs)[11,12] do not represent the native chromatin context in the human genome. Direct introduction of single base pair variants in the native genome are still low-throughput[13]. RNA-seq studies combined with genotyping or whole-genome sequencing have highlighted loci that are associated with gene expression in humans (eQTLs)[14-16]. However, as with GWAS, eQTL studies associate loci, rather than individual causal variants, to gene expression.

Statistical fine-mapping[3,17,18] is used to disentangle tightly correlated structures of the nearby genetic variants in LD to elucidate causal variant(s) in a locus identified by a genetic association study such as a GWAS on an eQTL study. For example, Benner *et al*[19] uses stochastic search to enumerate and evaluate possible causal configurations, and Wang *et al*[20] performs iterative Bayesian stepwise selection to prioritize causal variants. Such fine-mapping methods have been applied to identify putative causal eQTLs (i.e., variants that modify gene expression in native chromatin context) that are valuable both for understanding gene regulation and for interpreting GWAS signals at a locus[15,16,21-24]. However, fine-mapped eQTLs fall short of genome-wide characterization of non-coding function, as many variants fail to be identified because of LD or small effect size.

While not providing the same level of confidence as genome editing or fine-mapped eQTLs, computational predictions are informative about variant function in native chromatin in human cells, and can be applied to every variant in the genome. For example, state-of-the-art computational methods predict the effects of non-coding genetic variants on the epigenetic landscape and on gene expression as a function of sequence context using deep neural networks[25-30]. These methods, rather than directly training on gold standard expression-modifying variants, instead predict expression level or other outcomes as a function of sequence, and then score variants based on the difference in predicted expression between the two alleles.

Here, we combine such computational predictions with the large-scale, though not comprehensive, gold standard data provided by statistical fine-mapping of eQTLs, with two goals: to improve on existing computational predictors, and to expand the set of confidently-identified eQTLs. Toward the former goal, we combine an existing sequence-based predictor[28] with epigenetic data and other gene features into a single predictor, leveraging fine-mapped eQTLs[31] as training data. Specifically, we directly train a predictor of whether a variant modifies expression using 14,807 putative expression-modifying variant-gene pairs in humans as training data and utilizing 6,121 features; we call the resulting prediction the expression modifier score (EMS). Toward the second goal, we use EMS as a prior for statistical fine-mapping of eQTLs (analogous to recently-performed functionally-informed fine-mapping of complex traits[32-34]), increasing fine-mapping resolution and identifying an additional 20,913 variants across 49 tissues. Finally, using UK Biobank (UKBB)[35] phenotypes as an example, we show that EMS can

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

be incorporated into co-localization analysis at scale, and we identify 310 additional candidate genes for UK Biobank phenotypes.

Results

Functional enrichment of fine-mapped eQTLs

To define the set of putative expression-modifying variant-gene pairs, we analyzed results of recent fine-mapping of cis-eQTLs (+-1 Mb window) from GTEx v8[16,31], including the 14,807 variant-gene pairs with posterior inclusion probability (PIP) greater than 0.9 according to two methods[19,20] across 49 tissues (**Fig. S1, S2**). The size of our dataset allowed us to quantify the enrichment of putative causal variant-gene pairs for several functional annotations, including deep learning-derived variant effect scores from Basenji[28,29] and distance to canonical transcription starting site (TSS), with high precision (**Fig. 1, S3, S4**). Our results are consistent with previous studies[24,36]: putative causal variant-gene pairs are enriched for a number of functional annotations, such as 5'UTR, H3K4me3 (>10x enrichment compared to random variant-gene pairs) or distance to TSS (>500x enrichment for variant-gene pairs with distance to TSS<100), but are not strongly enriched for introns (0.966x), and are depleted for a histone mark related to heterochromatin state (H3K9me3; 0.510x enrichment).

Building a predictor for putative causal eQTLs [EMS]

Next, we built a random forest classifier of whether a given variant is a putative causal eQTL for a given gene using 807 binary functional annotations including cell-type-specific histone modifications as well as non-cell type specific annotations from the baseline model[4-6], 5,313 Basenji features corresponding to functional activity predictors[28,29], and distance to TSS. We then scaled the output score of the random forest classifier to reflect the probability of observing a positively labeled sample in a random draw from all the variant-gene pairs (**Fig. 2a, Methods**), and named this scaled score the expression modifier score (EMS). We performed the above process for 49 tissues in GTEx v8 individually, to obtain the EMS for variant-gene pairs in each tissue. In other words, EMS is an estimated probability of a variant-gene pair being a putative causal eQTL in a specific tissue, given the >6,000 functional annotations of the variant-gene pair. For whole blood, the Basenji scores together had 55.0% of the feature importance for EMS, and distance to TSS had feature importance of 43.1%. The binary functional annotations together had less than 2% of importance (**Fig. 2b, c**). Analyses of other tissues also showed that (1) distance to TSS is by far the most important single feature, (2) Basenji scores individually explain a small fraction of predictor performance but are collectively equally or more important than the distance to TSS, and (3) compared to the distance to TSS and Basenji scores, the feature importances of both cell-type specific and non-specific binary functional annotations are much smaller (**Supplementary File 1**).

Performance evaluation of EMS:

To evaluate the performance of EMS, we focused on whole blood and compared EMS (calculated by leaving one chromosome out at a time to avoid overfitting) to other genomic scores[26,37-40]. EMS achieved higher prediction accuracy than other genomic scores for putative causal eQTLs (top bin enrichment for held-out putative causal eQTLs 18.3x vs. 15.1x for distance to TSS, the second

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

best, Fisher's exact test p=$3.33 \cdot 10^{-4}$, **Fig. 3a**; AUPRC=0.884 vs. 0.856 when using distance to TSS, the second best, **Fig. S5; Methods**). EMS was among the top-performing methods in prioritizing experimentally suggested regulatory variants from reporter assay experiments[12,41], despite not varying distance to TSS, the most informative feature (**Fig. 3b-c**, **Fig. S6, Methods**). Finally, EMS was also among the top-performing methods in prioritizing putative causal non-coding variants for hematopoietic traits in the UKBB dataset (17.6x for EMS, best, vs 17.1x for DeepSEA, the second best; **Fig. 3d**), although there are known differences between the genetic architectures of cis-gene expression and complex traits[42]. These results were consistent when we performed the same set of analyses in different datasets: hematopoietic traits in BioBank Japan[43] (BBJ) and lymphoblastoid cell line (LCL) eQTL in Geuvadis [14,22] (**Fig. S7**).

Functionally-informed fine-mapping using EMS

Since EMS is in units of estimated probability, one natural way to utilize EMS for better prioritization of putative causal eQTLs is to use it as a prior for statistical fine-mapping. We developed a simple algorithm for approximate functionally-informed fine-mapping and applied it with EMS as a prior to obtain a functionally-informed posterior, denoted PIP_{EMS}, in whole blood (**Methods**). As expected, we found that PIP_{EMS} identified more putative causal eQTLs than the original PIP calculated with a uniform prior, denoted PIP_{unif}. Specifically, 95.4% of variants with $PIP_{unif} > 0.9$ also had $PIP_{EMS} > 0.9$ (2,152 out of 2,255), while only 33.8% of variants with $PIP_{EMS} > 0.9$ had $PIP_{unif} > 0.9$ (1,125 out of 3,277; **Fig. 4a**). Similarly, credible sets mostly decreased in size (**Fig. 4b, Supplementary File 2**). Previous work in functionally-informed fine-mapping[34] adjusted the prior so that the maximum prior value did not exceed 100 times the minimum prior value. We conducted a second round of functionally-informed fine-mapping with a similar adjustment of the prior, identifying fewer additional putative causal eQTLs, as expected (1,125 with EMS as a prior vs 269 with EMS adjusted to a max/min ratio of 100 as a prior; **Fig. S8).**

We evaluated the quality of PIP_{EMS} by comparing it with PIP_{unif} and a publicly available eQTL fine-mapping result that uses distance to TSS as a prior[16,23] (denoted $PIP_{DAP\text{-}G}$) in two ways (Other methods for functionally-informed fine-mapping based on expectation maximization[32,33,36] would be computationally intensive for a dataset this size, while the recently introduced PolyFun[34] is designed for complex traits.). First, PIP_{EMS} had the highest enrichment level of reporter assay QTLs[41] (raQTLs) in the PIP>0.9 bin (16.8x vs 12.9x in PIP_{unif} and 11.4x in $PIP_{DAP\text{-}G}$, Fisher's exact test p=$1.65 \cdot 10^{-2}$ between PIP_{EMS} and $PIP_{DAP\text{-}G}$; **Fig. 4c**). Second, complex trait causal non-coding variants were comparably enriched in PIP>0.9 bins (**Fig. S9**). These results suggest that PIP_{EMS} is a valid measure for identifying putative causal cis-regulatory variants.

Applying functionally-informed PIP (PIP_{EMS}) in gene prioritization across 95 traits

We next compared the utility of PIP_{EMS} to PIP_{unif} for complex trait gene prioritization, as in Weeks *et al*[44]. To do this, we first calculated PIP_{EMS} for 49 GTEx tissues using EMS of matched tissues as priors (**Fig. S10, S11**), resulting in a total of 20,913 additional eQTLs with PIP_{EMS}>0.9 (**Fig. 5, S12; Supplementary File 3**). We then co-localized the eQTL signals with 95 UKBB phenotypes. Using the gold standard gene set described in ref [44], PIP_{EMS} achieved higher precision and higher recall than PIP_{unif} (**Table. 1, Methods**). Overall, PIP_{EMS} elucidated 310 candidate genes for UKBB phenotypes that were not identified with PIP_{unif} (**Supplementary File 4**). On the other hand,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$PIP_{DAP\text{-}G}$ showed lower precision than PIP_{EMS} and PIP_{unif} but higher recall (**Table 1**), suggesting the value of future studies in investigating different priors in eQTL fine-mapping and the trade-off between precision and recall for gene prioritization.

An example of PIP_{EMS} resolving a credible set that is ambiguous with PIP_{unif} is shown in **Fig. 6.** Here, four variants upstream of *CITED4* are in perfect LD in GTEx, giving PIP_{unif} = 0.25 for all four (**Fig. S13)**. In UKBB, the four variants are also in high LD, with PIP for neutrophil count between 0.133 and 0.181 for all four. Thus, standard colocalization analysis does not identify *CITED4* as a neutrophil count-related gene (CLPP less than $4.53 \cdot 10^{-2}$ for all variants; **Methods)**. However, one of the four variants, rs35893233, creates a binding motif of *SPI1*, a transcription factor known to be involved in myeloid differentiation[45,46], and presents epigenetic activity in myeloid-related cell types, such as showing the highest basenji score for cap analysis gene expression (CAGE)[47] activity in acute myeloid leukemia (AML). This variant has >25x greater EMS than the other three variants ($1.73 \cdot 10^{-3}$ vs $6.11 \cdot 10^{-5}$, $1.00 \cdot 10^{-5}$ and $8.62 \cdot 10^{-6}$, respectively), enabling PIP_{EMS} to narrow down the credible set to the single variant (PIP_{EMS} = 0.956 for rs35893233). Integrating EMS into the co-localization analysis thus allows identification of *CITED4* as a neutrophil count-related gene (CLPP=0.173). Additional examples are described in **Fig. S14**.

Discussion

In this study we introduced EMS, a prediction of the probability that a variant has a cis-regulatory effect on gene expression in a tissue. To derive EMS, we trained a random forest model that takes >6,000 features. By analyzing the importance of each feature in the model, we showed that the importance of direct epigenetic measurements such as binary histone mark peak annotation is relatively limited once distance to TSS and deep learning-derived variant effect scores (Basenji) were incorporated. Taking whole blood as an example, we showed that EMS accurately prioritizes putative causal eQTLs, reporter-assay active variants, and putative complex trait causal non-coding variants. We provided a broader set of putative causal variants (n=20,913 across 49 tissues) by using EMS as a prior to perform approximate functionally-informed eQTL fine-mapping, and utilized EMS for co-localization analysis to identify 310 additional candidate genes for complex traits.

Evaluating predictors of non-coding variant function is complicated by the absence of gold standard data. While EMS outperformed other scores for prioritizing putative causal eQTLs, which we believe to be the closest to gold standard of existing large-scale base pair-resolution data sets, it did not outperform existing scores in prioritizing reporter assay active variants or putative complex trait causal non-coding variants. These latter two datasets, while valuable for independent validation, do not fully recapitulate the challenge of prioritizing causal expression-modifying variants in native context[42,48]. On the other hand, we recognize that putative causal eQTLs on a held-out chromosome do not constitute a fully independent validation set. As genome editing technologies continue to improve, we look forward to future large-scale datasets that will enable independent, gold standard evaluation and comparison of scores of non-coding functions at base-pair resolution.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Although our work refines our understanding of cis-gene regulatory mechanisms at single variant resolution, it also presents limitations. First, there are biases in the way the training variants are ascertained: the power to call a putative causal variant is affected by the recombination rate and the allele frequency of the variant[49,50], and the GTEx cohort is highly biased towards adult samples with European ancestry background. Second, although we utilize over 6,000 features in EMS, larger sets of variant and gene annotations such as 3D configuration of genome[51,52], constraint[53-55] or pathway enrichment[44] of genes could allow us to further improve prediction accuracy. Third, we simplified the prediction task by thresholding PIP. We formed a binary classification problem rather than a regression problem to build a predictor due to a highly skewed distribution of PIP, and because of LD-induced biases in variants with intermediate PIPs, but with larger sample size and a more principled hierarchical model, we could potentially take advantage of variants with intermediate PIP as well.

In this work, we focused on the task of predicting putative causal eQTLs. Future work could use a similar framework to predict putative causal splicing QTLs or other molecular QTLs for which statistical fine-mapping has identified a large number of high-PIP variants. In addition, although noisy effect size estimates from eQTL studies present a challenge, future work could explore leveraging features correlated with the sign and magnitude of effect (**Fig. S15**) to estimate these values. As recent studies have suggested, such approaches would also be valuable in understanding the gene expression and complex trait regulation landscape in light of natural selection[56]. Our approach of utilizing statistical fine-mapping of eQTLs to define training data, assembling large number of features to train a predictor, and using the predictor output to expand the set of putative causal eQTLs is highly generalizable. EMS for all variant-gene pairs in GTEx v8 are publicly available for 49 tissues. Our study provides a powerful resource for deciphering the mechanisms of non-coding variation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

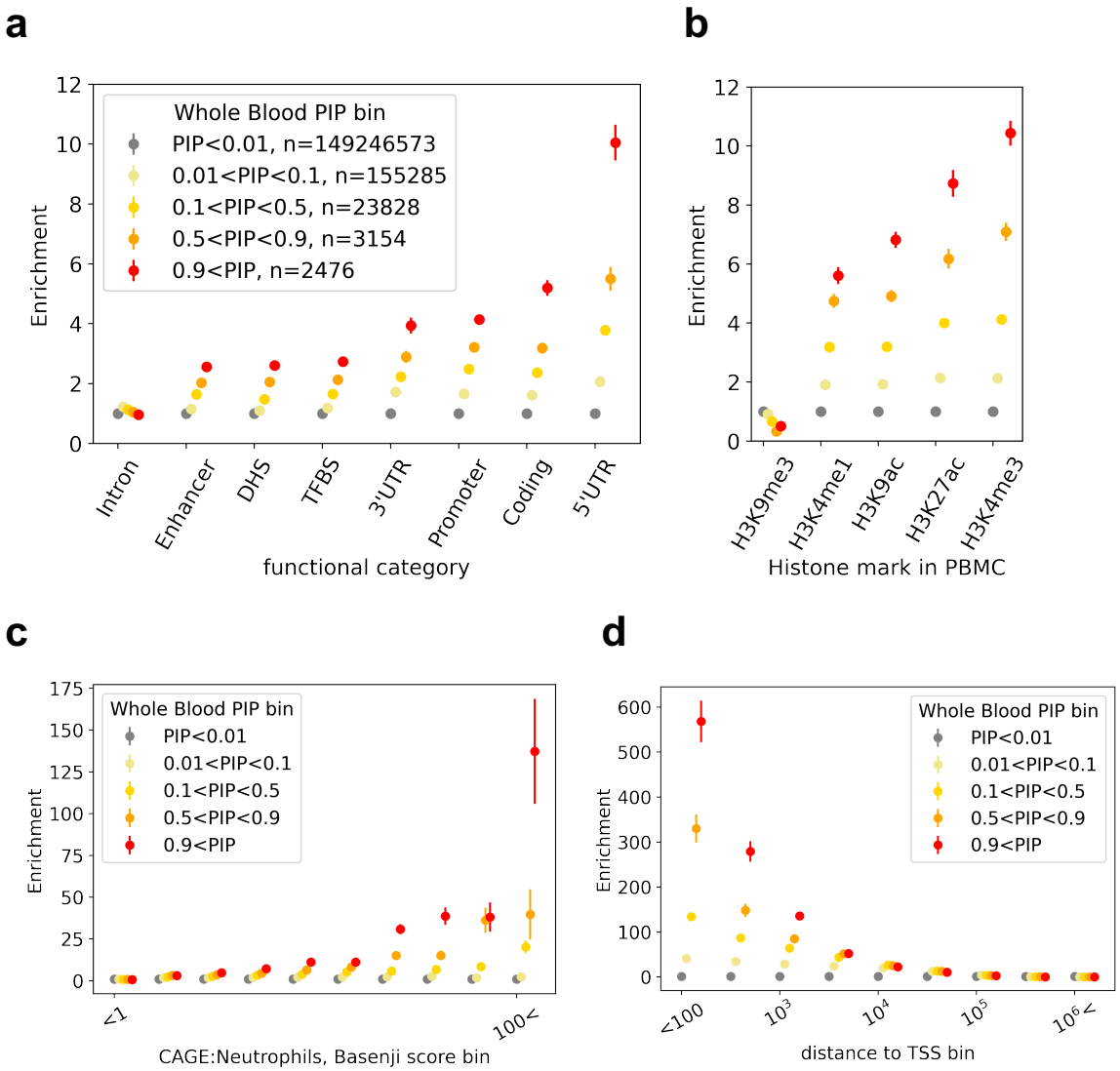

Figure 1. Examples of the enrichment of variant-gene pairs in whole blood eQTL PIP bins for functional genomics features

Enrichments of variant-gene pairs in different PIP bins in binary functional features (non-tissue specific, **a**; tissue-specific in peripheral blood mononuclear cells, **b**), deep learning-derived regulatory activity (CAGE[47]) prediction in Neutrophils (**c**), and distance to TSS (**d**) are shown.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

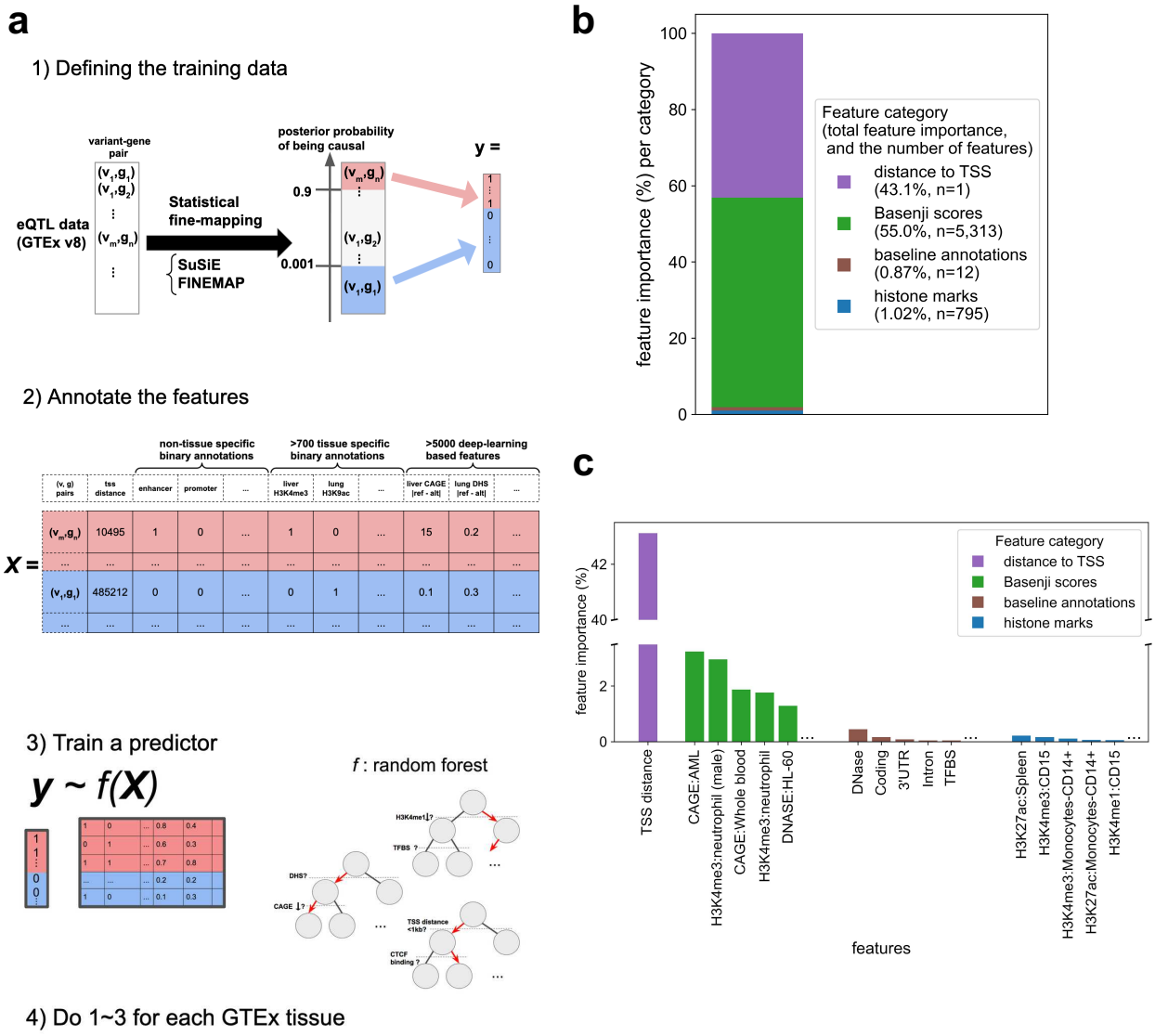

Figure 2. Schematic overview and feature importance of the Expression Modifier Score (EMS)

a. EMS is built by (1) defining the training data based on fine-mapping of GTEx v8 data, (2) annotating the variant-gene pairs with functional features, and (3) training a random forest classifier. We do this for each tissue. **b,c.** Feature importance (Mean Decrease of Impurity = MDI[59]) for four different feature categories (**b**), and top features for each category (**c**). Baseline annotations are non-tissue specific binary annotations from Finucane *et al*[6], and histone marks are tissue-specific binary histone mark annotations from Roadmap[5].

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

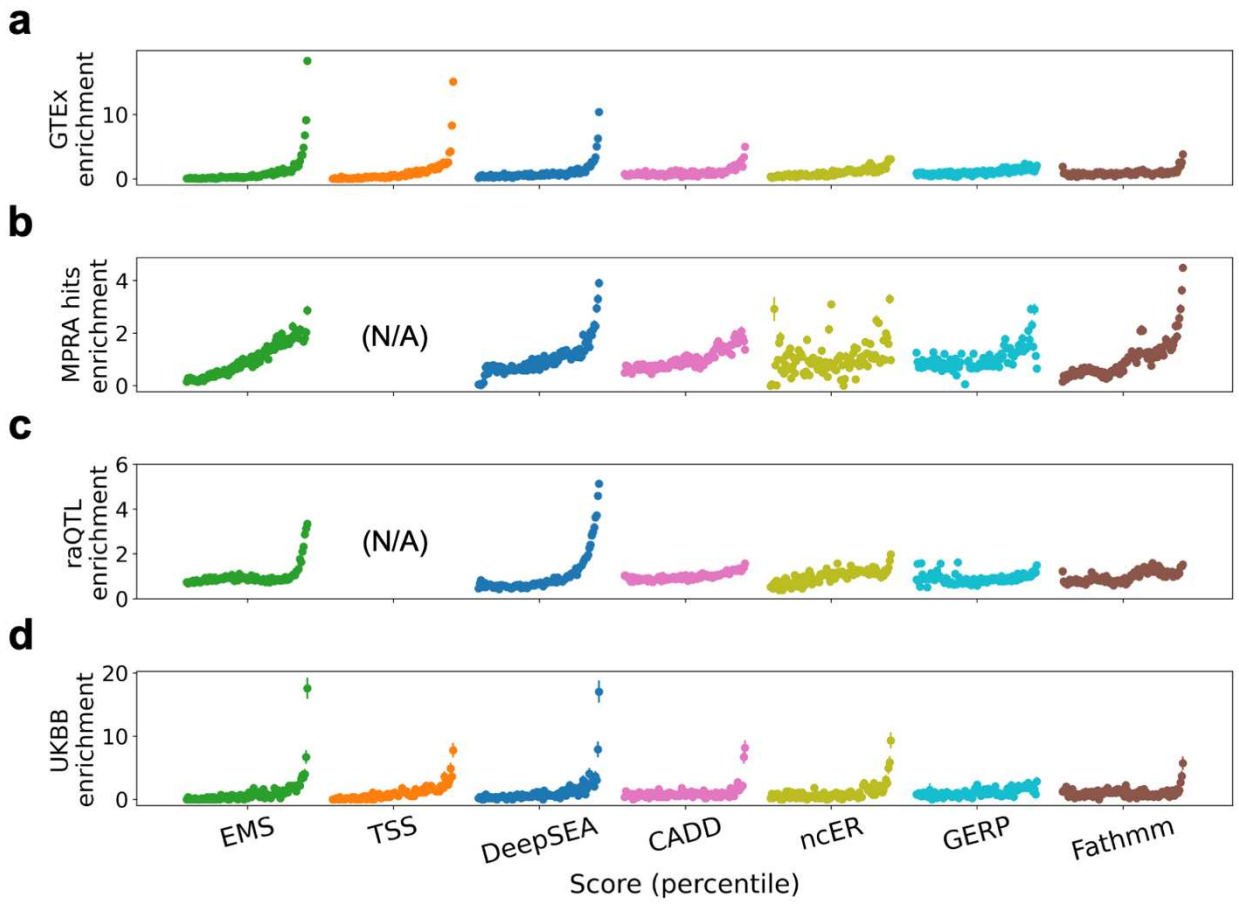

Figure 3: Performance evaluation of EMS

Comparison of the different scoring methods in prioritizing putative causal whole blood eQTLs in GTEx v8 (**a**), massive parallel reporter assay (MPRA) saturation mutagenesis hits[12] (**b**), reporter assay QTLs[41] (raQTLs) (**c**), and putative hematopoietic trait causal variants in UKBB (**d**) in different score percentiles.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

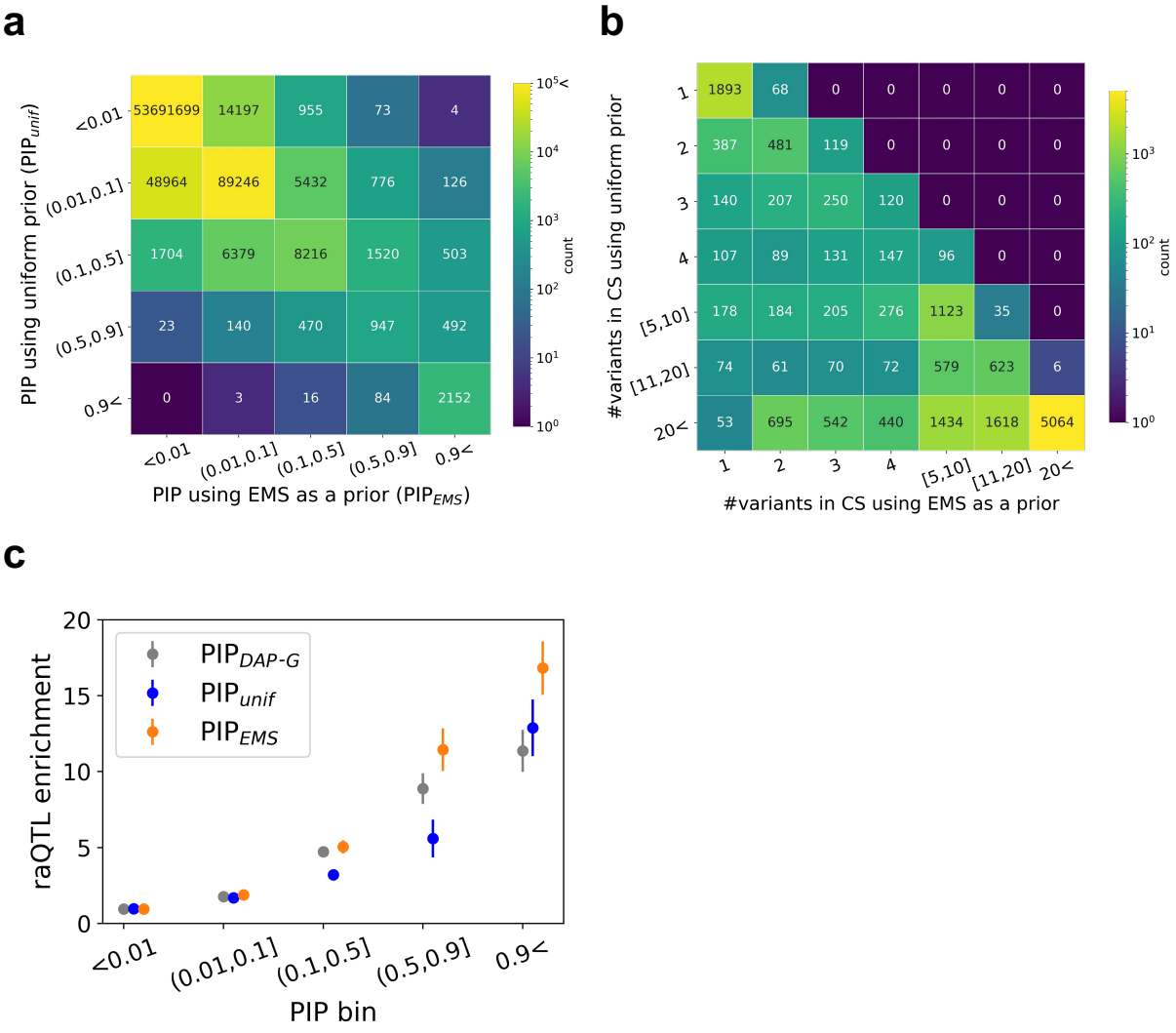

Figure 4. Functionally-informed fine-mapping with EMS as a prior

a. Number of variant-gene pairs in different PIP bins using a uniform prior vs. EMS as a prior. **b**. Number of variants in the 95% credible set (CS) identified by fine-mapping with uniform prior vs. EMS as a prior. **c.** Enrichment of reporter assay QTLs (raQTLs) in different PIP bins (gray: publicly available eQTL PIP using DAP-G[23], blue: uniform prior, orange: EMS as a prior).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

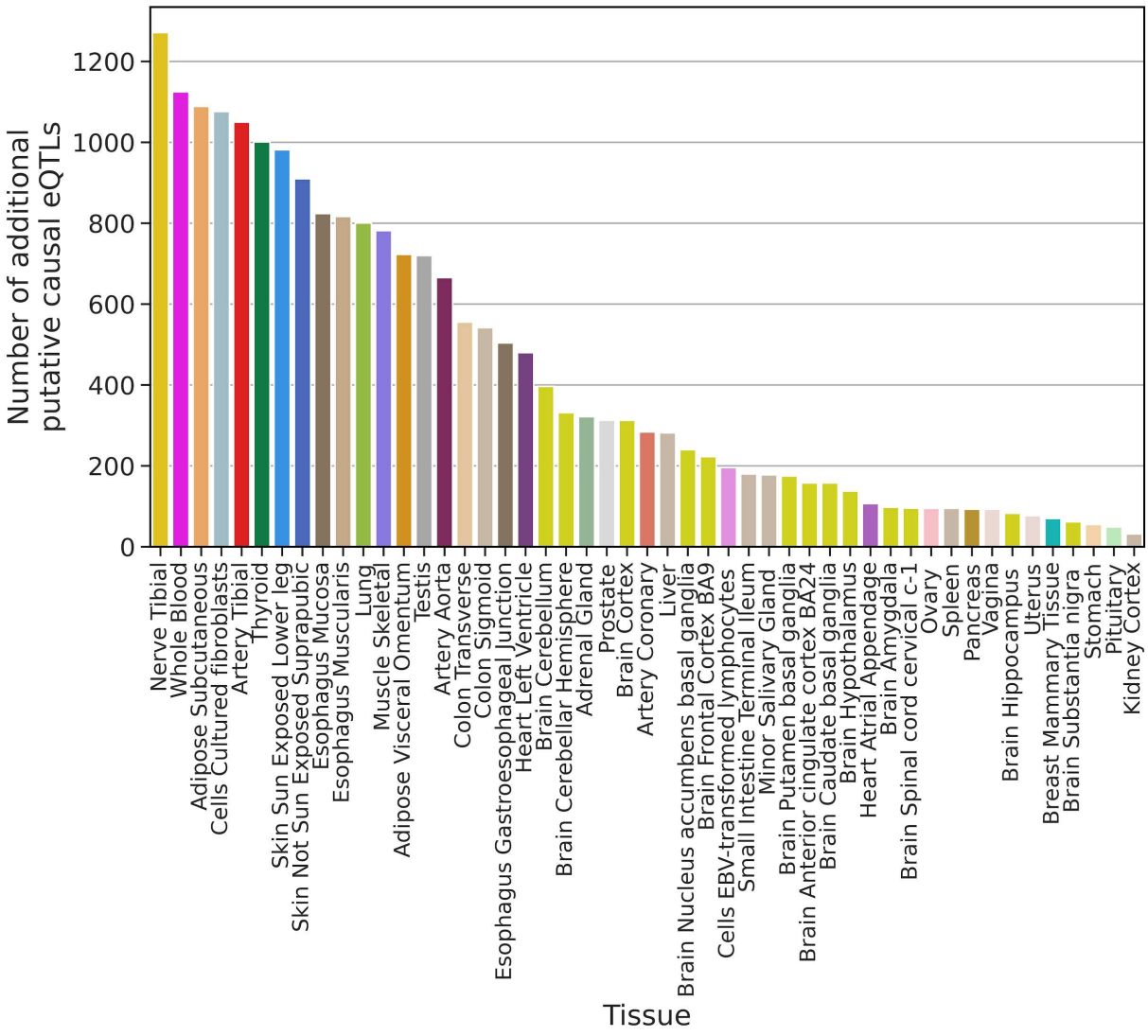

Figure 5. Additional putative causal eQTLs identified with functionally-informed fine-mapping across 49 tissues

The number of additional putative causal eQTLs (defined by PIP_{EMS}>0.9 and PIP_{unif}<0.9) for each tissue is shown in descending order.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

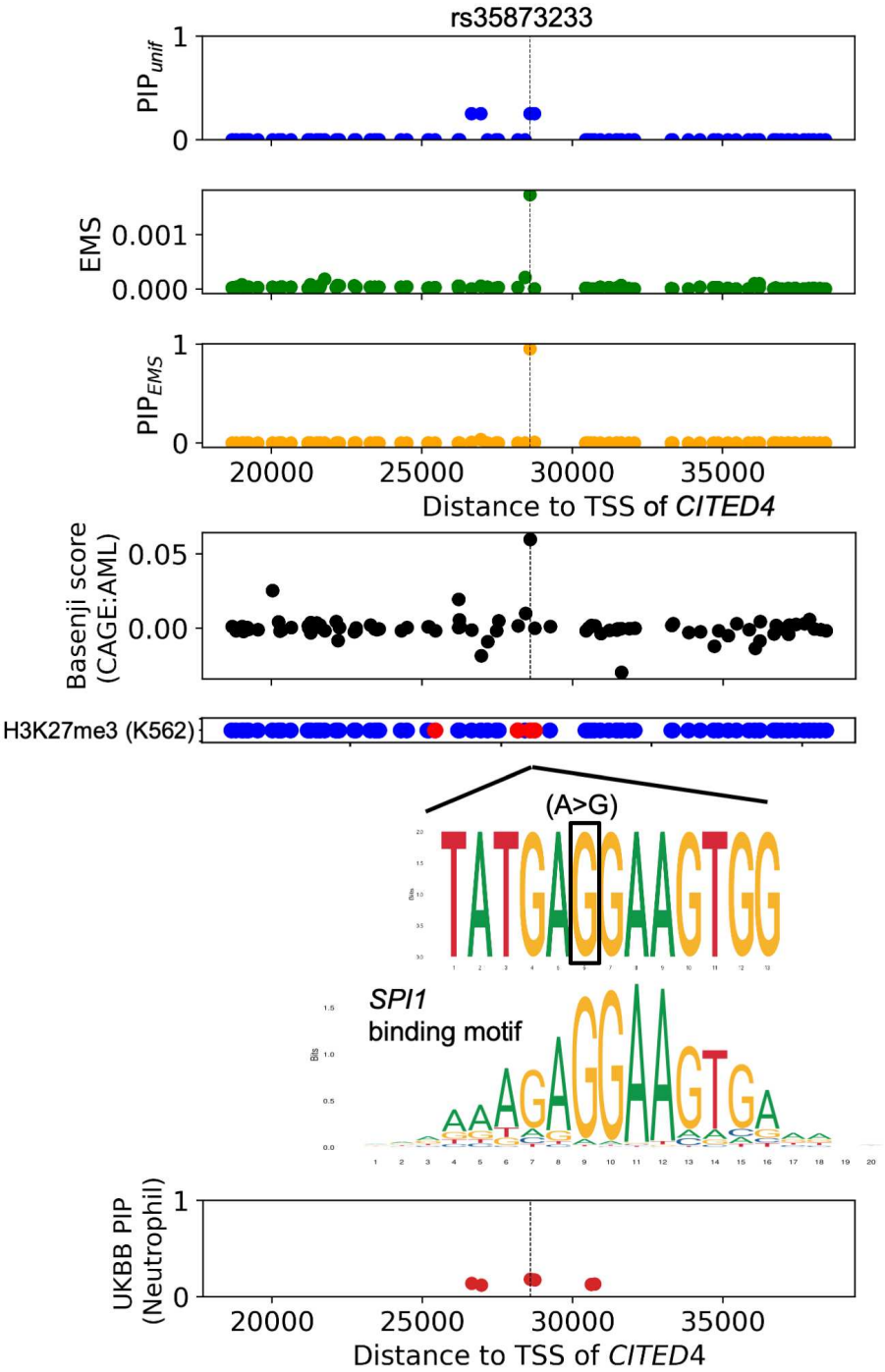

Figure 6. An example of a putative causal eQTL prioritized by EMS

rs35873233, an upstream variant of *CITED4*, was prioritized by functionally-informed fine-mapping using EMS as a prior. From top to the bottom: PIP with uniform prior (PIP_{unif}), EMS, PIP with EMS as a prior (PIP_{EMS}); Basenji score for CAGE[47] activity in acute myeloid leukemia (AML), H3K27me3 narrow peak in K562 cell line (red if the variant is on the peak, blue otherwise), sequence context[60] of the alternative allele aligned with the binding motif[61] of *SPI1*, and PIP for neutrophil count in UKBB[31,35] with uniform prior.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Precision and recall of the gene prioritization task for three different PIPs

method	tool	prior	precision	recall
PIP_{EMS}	SuSiE	EMS	0.556	0.052
PIP_{unif}	SuSiE	uniform	0.525	0.039
$PIP_{DAP\text{-}G}$	DAP-G	Distance to TSS	0.500	0.078

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods:

The Expression Modifier Score (EMS)

Fine-mapping of GTEx v8 data is described in Ulirsch et. al[31] and is summarized in the **Supplementary Methods**. We constructed a binary classification task by labeling the variant-gene pairs with PIP>0.9 for both of the two fine-mapping methods (FINEMAP[19] and SuSiE[20]) as positive, and the ones with PIP<0.0001 for both methods as negative. Each variant-gene pair was annotated with 6,121 features (distance to TSS annotated in the GTEx v8 dataset, 12 non-cell type specific binary features from the LDSC baseline model[6], 795 cell type specific binary features from the Roadmap Epigenomics Consortium[5], where variants falling in narrow peak are annotated as 1, and others are 0, and 5,313 deep-learning derived cell type-specific features generated by the Basenji model[28,29]; **Supplementary File 5**). The 152 most predictive features were selected based on different prediction accuracy metrics such as F1 measure and mean decrease of impurity (MDI) for each feature (**Supplementary Methods**). A combination of random search followed by grid search was performed to tune the hyperparameter for a random forest classifier that maximizes the AUROC of the binary prediction in the held-out dataset (**Supplementary File 6**). Finally, for each prediction score bin, we calculated the fraction of positively labeled samples and scaled the output score, to derive the EMS. Further details are described in the **Supplementary Methods**.

Performance evaluation of EMS

To evaluate the performance of EMS, for each chromosome, we trained EMS using all the other chromosomes to avoid overfitting. CADD[37] v1.4 and GERP[39] scores were annotated using the hail[57] annotation database (https://hail.is), and ncER[40] scores were downloaded from https://github.com/TelentiLab/ncER_datasets. In order to annotate the DeepSEA[26] v1.0 and Fathmm[38] v2.3 non-coding scores, we mapped hg38 coordinates to hg19 using the hail liftover function, removed variants that do not satisfy 1 to 1 matching, and followed their web instructions (https://humanbase.readthedocs.io/en/latest/deepsea.html, and http://fathmm.biocompute.org.uk) to score the variants. Insertion and deletions were not included in the Fathmm scores. For DeepSEA, we calculated the e-values from the individual features, following ref [4]. We computed the area under the receiver operating characteristic curve and the precision recall curve (**Fig. S5**) as well as enrichments of different variant-gene pairs or variants as described in the next sections (**Fig. 3**).

Computation of enrichment

Enrichment of a specific set of variant-gene pairs (e.g. putative causal variants in GTEx whole blood) in a score bin is defined as the probability of drawing a variant-gene pair in the set given that the variant-gene is in the score bin, divided by the overall probability of drawing a variant-gene pair in the set. The error bar denotes the standard error of the numerator, divided by the denominator (we assumed the standard error of the denominator is small enough, since the total number of variant-gene pairs is typically large; >100,000,000 for all the variant-gene pairs in GTEx v8). When testing binary functional features as in **Fig. 1**, the score is the individual functional feature, and the set is defined by the specific PIP bin.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

enrichment analysis of eQTL, complex trait, and reporter assay data

Saturation mutagenesis data[12] was downloaded from the MPRA data access portal (http://mpra.gs.washington.edu). An MPRA hit was defined as having a bonferroni-significant association p-value (lower than 0.05 divided by the total number of variant-cell type pairs) for at least one cell type, regardless of the effect size and direction. The raQTL data[41] was downloaded from https://osf.io/w5bzq/wiki/home/. EMS was re-scaled to have a constant distance to TSS (200 bp, roughly representing the scale of typical distance to TSS in plasmids[12]), which is expected to significantly decrease the performance of EMS compared to in native genome. Similarly, when comparing EMS with other scores for enrichments of MPRA hits or raQTLs, distance to TSS was not used for the comparison.

Fine-mapping of UKBB traits is described in Ulirsch et al[31]. To focus on non-coding regulatory effects, we annotated the variants in VEP[58] v85 and filtered out coding and splice variants for the UKBB dataset. For each (non-coding) variant, we calculated the maximum PIP over all the hematopoietic traits, as well as the maximum Whole-Blood EMS over all the genes in the cis window of the variant, since a variant can have different regulatory effect on different genes, for different phenotypes. A variant was defined as putative hematopoietic trait-causal if it has SuSiE PIP higher than 0.9 in any of the hematopoietic traits. In UKBB, we focused on the variants that exist in the GTEx v8 dataset to reduce the calculation complexity.

For all four datasets, the variants (or variant-gene pairs in GTEx) other than putative causal ones were randomly downsampled to achieve a total number of variants to be exactly 100,000, to reduce the computational burden while keeping enough number of variants to observe statistical significance. GTEx enrichment, MPRA hits enrichment, raQTL enrichment and UKBB enrichment are thus defined as the enrichment of putative causal eQTLs, MPRA hits, raQTLs and putative hematopoietic-trait causal variants in the downsampled dataset respectively.

Approximate functionally-informed fine-mapping using EMS

In the Sum of Single Effects (SuSiE) model, for a given gene, the vector b of true SNP effects on that gene is modeled as a sum of vectors with only one non-zero element each:

$$b = \sum_{l=1}^{L} b_l$$
$$||b_l||_0 = 1$$

where b and b_l are vectors of length m and m is the number of variants in the locus. Intuitively, each b_l corresponds to the contribution of one causal variant. One output of SuSiE is a set of m-vectors $\alpha_1, \ldots, \alpha_L$, with $\alpha_L(v)$ equal to the posterior probability that $b_l(v) \neq 0$; i.e., that the l-th causal variant is the variant v. Credible sets are computed for each l from α_l, and credible sets that are not "pure" -- i.e., that contain a pair of variants with absolute correlation less than 0.5 -- are pruned out. The α_l are also used to compute PIPs.

Our algorithm for approximate functionally-informed fine-mapping takes the approach of re-weighting the posterior probability calculated using the uniform prior, analogous to ref [33], and

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

proceeds as follows. For each gene and each tissue, we start with $\alpha_1, \ldots, \alpha_L$ computed by SuSiE using the uniform prior. For each l, if $\alpha_l$corresponds to a pure credible set, we re-weight each element of α_l by the EMS of the corresponding variant, and we normalize so that the sum is equal to 1, obtaining $\hat{\alpha}_l$. In other words, letting $w_1 \ldots w_m$ denote the EMSs for the m variants, we define $\hat{\alpha}_l(v)$ for the variant v to be

$$\hat{\alpha}_l(v) = \frac{w_v \alpha_l(v)}{\sum_{u=1}^{m} w_u \alpha_l(u)}$$

if α_l corresponds to a pure credible set; otherwise, we set $\hat{\alpha}_l = \alpha_l$. We then use the updated $\hat{\alpha}_1, \ldots, \hat{\alpha}_L$ to compute updated PIPs and credible sets as in the original SuSiE method. See **Supplementary Methods** for further details.

Performance evaluation of PIP_{EMS} and application to gene prioritization

PIP using distance to TSS as a prior ($PIP_{DAP\text{-}G}$) was downloaded from the GTEx portal (https://gtexportal.org/). The raQTL data was downloaded from https://osf.io/w5bzq/wiki/home/, and the negative variants were randomly downsampled to a total of 100,000 variants. For complex trait causal non-coding variant prioritization, a threshold of PIP>0.1 was chosen to account for low sample size. We defined a gene prioritization task using 49 tissues in GTEx v8 and 95 complex traits in UKBB using the following steps (further details are described in Weeks *et al.*[44]):

Across all traits, we identified 1 Mb regions centered at unresolved credible sets (no coding variant with PIP>0.1) that additionally contained at least one "gold standard gene" (protein-coding variant with PIP>0.5) for the same trait. There were 2,897 such regions and 1,161 gold standard genes. Our intuition is that the gene with the fine-mapped protein-coding variant is most likely to be the primary causal signal, and that a nearby non-coding signal is more likely to act through this gene (i.e. via regulation) than through a different gene.

For each gene-region pair, we defined the co-localization posterior probability (CLPP) for the gene to be the maximum of the product of the eQTL PIP and trait PIP, across all tissues and all variants in the unresolved credible set. A gene is prioritized if it has CLPP > 0.1 and it has the maximum CLPP in its region. We compute the precision as the number of correctly prioritized genes (where the prioritized gene is also the gene with the primary, protein-coding signal) divided by the total number of prioritized genes. We compute recall as the number of correctly prioritized genes divided by the total number of gold standard genes. The total number of candidate genes is defined as the number of gene-trait pairs presenting CLPP>0.1 in at least one tissue and variant.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Paul, D. S., Soranzo, N. & Beck, S. Functional interpretation of non-coding sequence variation: Concepts and challenges. *Bioessays* **36**, 191–199 (2014).
3. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294–1301 (2012).
4. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
5. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
6. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).
7. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *Am J Hum Genet* **94**, 559–573 (2014).
8. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
9. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* **45**, 124–130 (2013).
10. Trynka, G. & Raychaudhuri, S. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. *Current Opinion in Genetics & Development* **23**, 635–641 (2013).
11. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529 (2016).
12. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications* **10**, 3583 (2019).
13. Tian, R. *et al.* Pitfalls in Single Clone CRISPR-Cas9 Mutagenesis to Fine-Map Regulatory Intervals. *Genes (Basel)* **11**, (2020).
14. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
15. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
16. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
17. Chen, W. *et al.* Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719–736 (2015).
18. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19**, 491–504 (2018).
19. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
20. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* doi:10.1111/rssb.12388**.**

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

21. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* 198, 497–508 (2014)
22. Brown, A. A. *et al.* Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nature Genetics* 49, 1747–1751 (2017).
23. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *The American Journal of Human Genetics* **98**, 1114–1129 (2016).
24. Wen, X., Luca, F. & Pique-Regi, R. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLOS Genetics* **11**, e1005176 (2015).
25. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Reports* **31**, 107663 (2020).
26. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* **12**, 931–934 (2015).
27. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**, 1171–1179 (2018).
28. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* gr.227819.117 (2018) doi:10.1101/gr.227819.117.
29. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology* **16**, e1008050 (2020).
30. Kopp, W., Monti, R., Tamburrini, A., Ohler, U. & Akalin, A. Deep learning for genomics using Janggu. *Nature Communications* **11**, 3488 (2020).
31. Ulirsch, J. *et al.* in prep
32. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genetics* **10**, e1004722 (2014).
33. Jiang, J. *et al.* Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Communications Biology* **2**, 1–12 (2019).
34. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics* 1–9 (2020) doi:10.1038/s41588-020-00735-5.
35. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203 (2018).
36. Chen, W., McDonnell, S. K., Thibodeau, S. N., Tillmans, L. S. & Schaid, D. J. Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. *Genetics* **204**, 933–958 (2016).
37. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886–D894 (2019).
38. Shihab, H. A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation* **34**, 57–65 (2013).
39. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

40. Wells, A. *et al.* Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat Commun* **10**, (2019).
41. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nature Genetics* **51**, 1160–1169 (2019).
42. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics* **52**, 626–633 (2020).
43. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics* **50**, 390–400 (2018).
44. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv* 2020.09.08.20190561 (2020) doi:10.1101/2020.09.08.20190561.
45. Chen, H. *et al.* PU.1 (Spi-1) autoregulates its expression in myeloid cells. *Oncogene* **11**, 1549–1560 (1995).
46. Burda, P., Laslo, P. & Stopka, T. The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia* **24**, 1249–1257 (2010).
47. Takahashi, H., Kato, S., Murata, M. & Carninci, P. CAGE- Cap Analysis Gene Expression: a protocol for the detection of promoter and transcriptional networks. *Methods Mol Biol* **786**, 181–200 (2012).
48. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res* **27**, 38–52 (2017).
49. LaPierre, N. *et al.* Identifying Causal Variants by Fine Mapping Across Multiple Studies. *bioRxiv* 2020.01.15.908517 (2020) doi:10.1101/2020.01.15.908517.
50. Hutchinson, A., Watson, H. & Wallace, C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLOS Computational Biology* **16**, e1007829 (2020).
51. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics* **21**, 207–226 (2020).
52. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods* **17**, 1111–1117 (2020).
53. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
54. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
55. Iulio, J. di *et al.* The human noncoding genome defined by genetic diversity. *Nature Genetics* **50**, 333 (2018).
56. Schoech, A. P. *et al.* Negative short-range genomic autocorrelation of causal effects on human complex traits. *bioRxiv* 2020.09.23.310748 (2020) doi:10.1101/2020.09.23.310748.
57. Hail Team. Hail 0.2. https://github.com/hail-is/hail
58. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17**, 122 (2016).
59. Louppe, G. Understanding Random Forests: From Theory to Practice. *arXiv:1407.7502 [stat]* (2015).
60. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.20.347294; this version posted December 1, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

61. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87–D92 (2020).

Data availability

EMS for 49 tissues are available at https://www.finucanelab.org/data.

Code availability

Code used in this manuscript is available at https://github.com/FinucaneLab/Expression_Modifier_Score/ .

Acknowledgements

We thank Yakir Reshef, Jesse Engreitz, Elle Weeks, and all the members of Finucane lab for useful conversations. H.K.F. was funded by NIH grant DP5 OD024582 and by Eric and Wendy Schmidt. Q.S.W. and M.K. were supported by the Nakajima Foundation Scholarship.

Contributions

Q.S.W., D.M., and H.K.F. designed the study. Q.S.W., D.R.K., J.U., S.S. analyzed the data. Q.S.W. and H.K.F. wrote the manuscript with input from all authors.

Competing interests

D.G.M. is a founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme.