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:

We present the expression modifier score (EMS), a predicted probability that a variant has a cis-regulatory effect on gene expression, trained on fine-mapped eQTLs and leveraging 6,121 features including epigenetic marks and sequence-based neural network predictions. We validate EMS and use it as a prior for statistical fine-mapping of eQTLs, identifying an additional 20,913 putatively causal eQTLs. Incorporating EMS into colocalization analysis identifies 310 additional candidate genes for UK Biobank phenotypes.

Main text:

Understanding the effects of non-coding variants on gene expression is of fundamental importance[1]. Experimental methods to assess whether an individual variant modifies expression of a nearby gene in native chromatin context via direct perturbation are low-throughput[2]. Existing computational predictors[3-6], which are higher-throughput, thus lack large gold standard sets of regulatory variants for training and validation. Here, we leverage a novel set of 14,807 putative expression-modifying variant-gene pairs in humans, and use 6,121 features to directly train a predictor of whether a variant modifies expression.

To define the set of putative expression-modifying variant-gene pairs, we analyzed results of recent fine-mapping of eQTLs from GTEx[7,8], including the 14,807 variant-gene pairs with posterior inclusion probability (PIP) greater than 0.9 according to two methods[9,10] across 49 tissues (**Fig. S1, S2**). The size of our data set allowed us to quantify the enrichment of putative causal variants for several functional annotations, including deep learning-derived variant

effect scores from Basenji[6] and distance to canonical transcription starting site (TSS), with high precision (**Fig. S3, S4, S5**).

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[11-13], 5,313 Basenji features corresponding to functional activity predictors[6,14], 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. S6, Methods**), and named this scaled score the expression modifier score (EMS). We performed the above process for 49 tissues in GTEx v8 individually. 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. S7**). Results were similar for other tissues (**Supplementary File 1**).

EMS achieved higher prediction accuracy than other genomic scores[4,15-18] for putative causal eQTLs on a held-out chromosome (top bin enrichment for held-out putative causal eQTLs 18.3x vs. 15.1x for distance to TSS, the second best, Fisher's exact test $p=3.33 \cdot 10^{-4}$, **Fig. 1a**; AUPRC=0.884 vs. 0.856 when using distance to TSS, the second best, **Fig. S8; Methods**). EMS was among the top-performing methods in prioritizing experimentally suggested regulatory variants from reporter assay experiments[19,20], despite not varying distance to TSS, the most informative feature (**Fig. 1b-c**, **Fig. S9, Methods**). Finally, EMS prioritized putative causal non-coding variants for hematopoietic traits in the UK Biobank (UKBB) dataset[21] with performance comparable to other scores (17.6x for EMS vs 17.1x for DeepSEA, the second best; **Fig. 1d**), although there are known differences between the genetic architectures of cis-gene expression and complex traits[22]. The results were consistent when we performed the same set of analyses in different datasets: hematopoietic traits in BioBank Japan[23] (BBJ) and lymphoblastoid cell line (LCL) eQTL in Geuvadis [24,25] (**Fig. S10**).

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} (**Methods**). 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. 2a**). Similarly, credible sets mostly decreased in size (**Fig. 2b, Supplementary File 2**).

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[7,25] (denoted $PIP_{DAP\text{-}G}$) in two ways (Other methods for functionally-informed fine-mapping[26,27] would be computationally intensive for a data set this size; the recently introduced PolyFun[28] is designed for complex

traits.). First, PIP_{EMS} had the highest enrichment level of reporter assay QTLs[25] (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. 2c**). Second, complex trait causal non-coding variants were comparably enriched in PIP>0.9 bins (**Fig. S11**). These results suggest that PIP_{EMS} is a valid measure for identifying putative causal cis-regulatory variants.

We next compared the usage of PIP_{EMS} to PIP_{unif} for complex trait gene prioritization, as in Weeks *et al*[29]. We calculated PIP_{EMS} for 49 GTEx tissues (**Fig. S12, S13**), resulting in a total of additional 20,913 eQTLs with PIP_{EMS}>0.9 (**Fig. S14; Supplementary File 3**). We then co-localized the eQTL signals with 95 UKBB phenotypes. Using the gold standard gene set described in ref [29], 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, $PIP_{DAP\text{-}G}$ 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.

An example of PIP_{EMS} resolving a credible set that is ambiguous with PIP_{unif} is shown in **Fig. 2d.** Here, four variants upstream of *CITED4* are in perfect LD in GTEx, giving PIP_{unif} = 0.25 for all four (**Fig. S15)**. 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[30], and presents epigenetic activity in myeloid-related cell types, such as showing the highest basenji score for cap analysis gene expression (CAGE) 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. S16**.

Limitations of our approach include (1) limited power to call putative causal variants in high LD regions or with low minor allele frequency, (2) the simplifications of thresholding and ignoring effect size and direction, and (3) the lack of a comprehensive set of features. EMS for all variants in GTEx v8 are publicly available for 49 tissues. Our study provides a powerful resource for deciphering the mechanisms of non-coding variation.

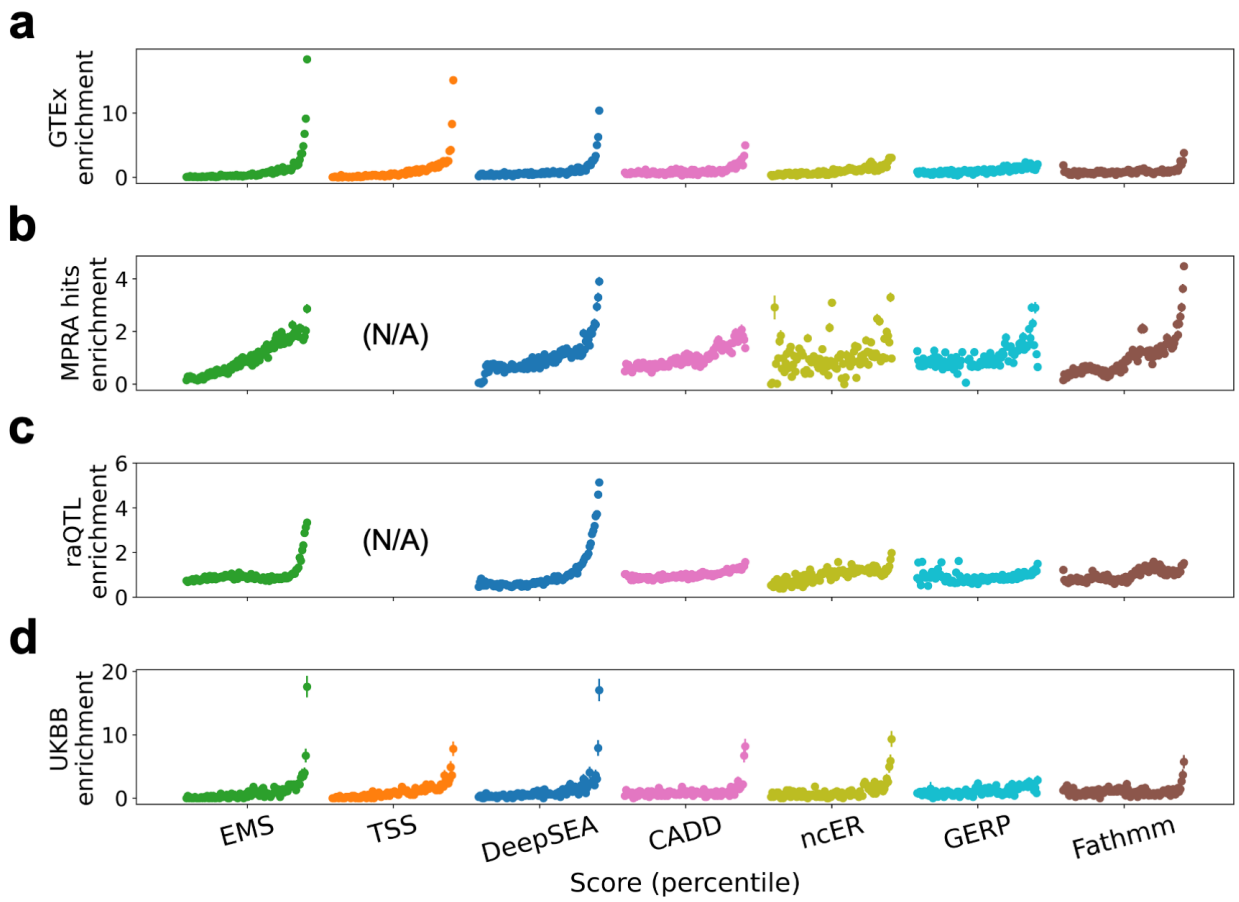

Figure 1. Performance evaluation of EMS. Comparison of the different scoring methods, in prioritizing putative causal whole blood eQTLs in GTEx v8 (**a**), massive parallel reporter assay (MPRA) saturation mutagenesis hits[19] (**b**), reporter assay QTLs[20] (raQTLs) (**c**), and putative hematopoietic trait causal variants in UKBB (**d**) in different score percentiles. Distance to TSS is not defined for reporter assays (**Methods**).

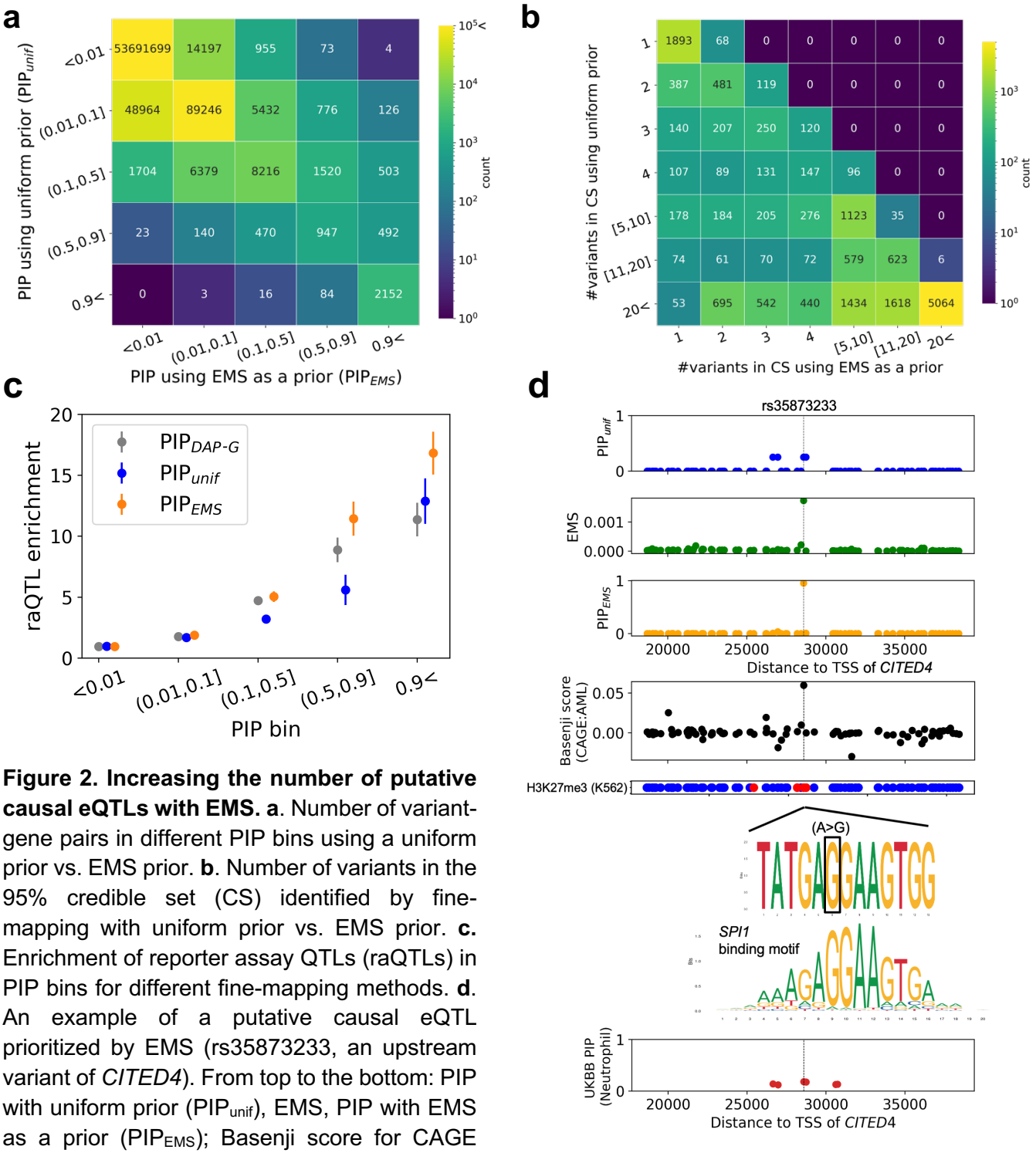

Figure 2. Increasing the number of putative causal eQTLs with EMS. a. Number of variant-gene pairs in different PIP bins using a uniform prior vs. EMS prior. **b**. Number of variants in the 95% credible set (CS) identified by fine-mapping with uniform prior vs. EMS prior. **c.** Enrichment of reporter assay QTLs (raQTLs) in PIP bins for different fine-mapping methods. **d**. An example of a putative causal eQTL prioritized by EMS (rs35873233, an upstream variant of *CITED4*). From top to the bottom: PIP with uniform prior (PIP_{unif}), EMS, PIP with EMS as a prior (PIP_{EMS}); Basenji score for CAGE activity in acute myeloid leukemia, H3K27me3 narrow peak in K562 cell line (red if the variant is in the peak, blue otherwise), sequence context[32] of the alternative allele aligned with the binding motif[33] of *SPI1*, and PIP for neutrophil count in UKBB[8,19] with uniform prior.

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

Methods:

The Expression Modifier Score (EMS)

Fine-mapping of GTEx v8 data is described in Ulirsch et. al[8] 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 (SuSiE[9] and FINEMAP[10]) 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[13], 795 cell type specific binary features from the Roadmap Epigenomics Consortium[12], 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[6,14]; **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 v1.4 and GERP scores were annotated using the hail annotation database (https://hail.is). ncER scores were downloaded from https://github.com/TelentiLab/ncER_datasets. In order to annotate the DeepSEA v1.0 and Fathmm 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. Insertions and deletions were excluded for 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. S8**) as well as enrichments of different variant-gene pairs or variants as described in the next sections (**Fig. 1**).

Computation of enrichment

Enrichment of a specific set of variant-gene pairs (e.g. putative causal eQTLs in the GTEx whole blood dataset) 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). When testing binary functional features as in **Fig. S3, S4**, the score is the individual functional feature, and the set is defined by the specific PIP bin.

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

Saturation mutagenesis data[19] 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[20] 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[5]), 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[8]. To focus on non-coding regulatory effects, we annotated the variants in VEP 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 and raQTL dataset, 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 [31], and 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 α_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 $w_1,...,w_m$ 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. The number of putative causal eQTLs is defined as the number of variant-gene-tissue pairs with PIP_{EMS}>0.9. 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 and 95 complex traits in UKBB using the following steps (further details are described in Weeks *et al.*[26]):

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.

References

1. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. Tian, R. *et al.* Pitfalls in Single Clone CRISPR-Cas9 Mutagenesis to Fine-Map Regulatory Intervals. *Genes (Basel)* **11**, (2020).
3. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Reports* **31**, 107663 (2020).
4. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* **12**, 931–934 (2015).
5. 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).
6. 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.
7. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* 787903 (2019) doi:10.1101/787903.
8. Ulirsch, J. *et al.* in prep
9. 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**.**
10. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
11. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
12. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
13. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).
14. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology* **16**, e1008050 (2020).
15. 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).
16. 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).
17. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
18. Wells, A. *et al.* Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat Commun* **10**, (2019).
19. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nature Communications* **10**, 3583 (2019).
20. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nature Genetics* **51**, 1160–1169 (2019).

21. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203 (2018).
22. 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).
23. 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).
24. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
25. 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).
26. Chen, W. *et al.* Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics* **200**, 719–736 (2015).
27. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genetics* **10**, e1004722 (2014).
28. Weissbrod, O. *et al.* Functionally-informed fine-mapping and polygenic localization of complex trait heritability. *bioRxiv* 807792 (2020) doi:10.1101/807792.
29. 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.
30. Chen, H. *et al.* PU.1 (Spi-1) autoregulates its expression in myeloid cells. *Oncogene* **11**, 1549–1560 (1995).
31. 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).
32. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
33. 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 study is available at https://github.com/FinucaneLab/Expression_Modifier_Score/ .

Acknowledgements

We thank Yakir Reshef and all the members of the 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.G.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.