









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

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

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 6121 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.

<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>2</sup>Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. <sup>3</sup>PhD program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. <sup>4</sup>Calico Life Sciences, South San Francisco, CA, USA. <sup>5</sup>PhD program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. <sup>6</sup>Department of Statistical Genetics, Osaka University Graduate School of Medicine, Osaka, Japan. <sup>7</sup>Harvard College, Cambridge, MA, USA. <sup>8</sup>Laboratory of Statistical Immunology, Immunology Frontier Research Center (WPI-IFReC), Osaka University, Osaka, Japan. <sup>9</sup>Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Osaka, Japan. <sup>10</sup>Centre for Population Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia. <sup>11</sup>Centre for Population Genomics, Murdoch Children's Research Institute, Parkville, VIC, Australia. \*A list of authors and their affiliations appears at the end of the paper. ✉email: [qingbow@broadinstitute.org](mailto:qingbow@broadinstitute.org); [finucane@broadinstitute.org](mailto:finucane@broadinstitute.org)

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

Many approaches to characterize noncoding variants exist. Large-scale consortium studies<sup>4,5</sup> have provided a map of functional and regulatory elements across the genome in different cell types that are enriched in various trait heritability<sup>6–10</sup>. Reporter assays have been powerful tools to test variant effects in cellular contexts, but typical high-throughput massive parallel reporter assays (MPRAs)<sup>11,12</sup> 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<sup>13</sup>. RNA-seq studies combined with genotyping or whole-genome sequencing have highlighted loci that are associated with gene expression in humans (eQTLs)<sup>14–16</sup>. However, as with GWAS, eQTL studies associate loci, rather than individual causal variants, to gene expression.

Statistical fine-mapping<sup>3,17,18</sup> 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.<sup>19</sup> uses stochastic search to enumerate and evaluate possible causal configurations, and Wang et al.<sup>20</sup> 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<sup>15,16,21–24</sup>. However, fine-mapped eQTLs fall short of genome-wide characterization of noncoding 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 noncoding genetic variants on the epigenetic landscape and on gene expression as a function of sequence context, using deep neural networks<sup>25–30</sup>. 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<sup>28</sup> with epigenetic data and other gene features into a single predictor, leveraging fine-mapped eQTLs (<https://www.finucanelab.org/data>) 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 6121 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<sup>31–33</sup>), increasing fine-mapping resolution and

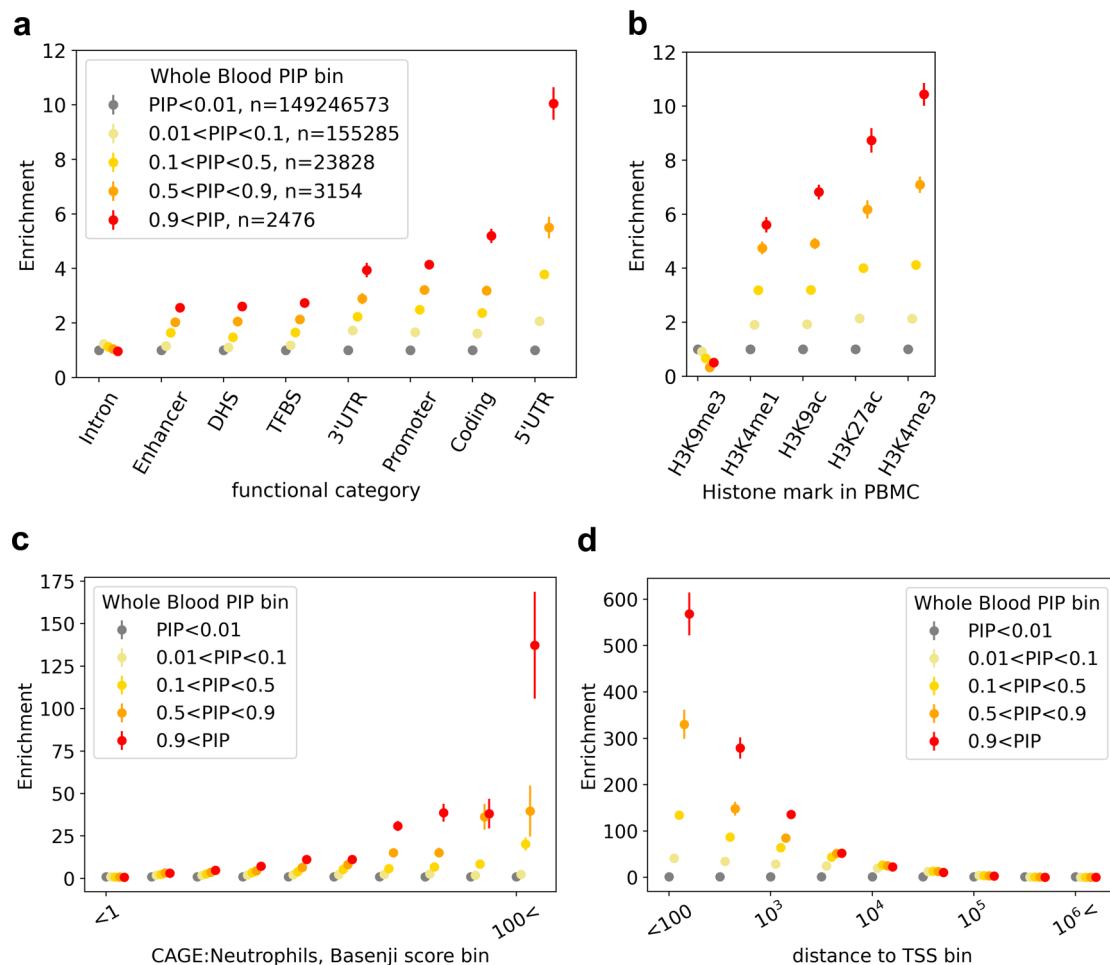
identifying an additional 20,913 variants across 49 tissues. Finally, using UK Biobank (UKBB)<sup>34</sup> phenotypes as an example, we show that EMS can be incorporated into colocalization analysis at scale, and we identify 310 additional candidate genes for UKBB phenotypes.

## 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 ( $\pm 1$  Mb window) from GTEx v8 (ref. 16; <https://www.finucanelab.org/data>), including the 14,807 variant–gene pairs with posterior inclusion probability (PIP) > 0.9 according to two methods<sup>19,20</sup> across 49 tissues (Supplementary Figs. 1 and 2). 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<sup>28,29</sup> and distance to canonical transcription starting site (TSS), with high precision (Fig. 1, and Supplementary Figs. 3 and 4). Our results are consistent with previous studies<sup>24,35</sup>: putative causal variant–gene pairs are enriched for a number of functional annotations, such as 5'UTR, H3K4me3 (>10 $\times$  enrichment compared to random variant–gene pairs) or distance to TSS (>500 $\times$  enrichment for variant–gene pairs with distance to TSS < 100), but are not strongly enriched for introns (0.966 $\times$ ), and are depleted for a histone mark related to heterochromatin state (H3K9me3; 0.510 $\times$  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<sup>4–6</sup>, 5313 Basenji features corresponding to functional activity predictors<sup>28,29</sup>, 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 and “Methods”), and named this scaled score the 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 >6000 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 <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 nonspecific binary functional annotations are much smaller (Supplementary Data 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<sup>26,36–39</sup>. EMS achieved higher prediction accuracy than other genomic scores for putative causal eQTLs (top bin enrichment for held-out putative causal eQTLs 18.3 $\times$  vs 15.1 $\times$  for distance to TSS, the second best, Fisher's exact test  $p = 3.33 \times 10^{-4}$ , Fig. 3a; AUPRC = 0.884 vs 0.856 when using distance to TSS, the second best, Supplementary Fig. 5 and “Methods”). EMS



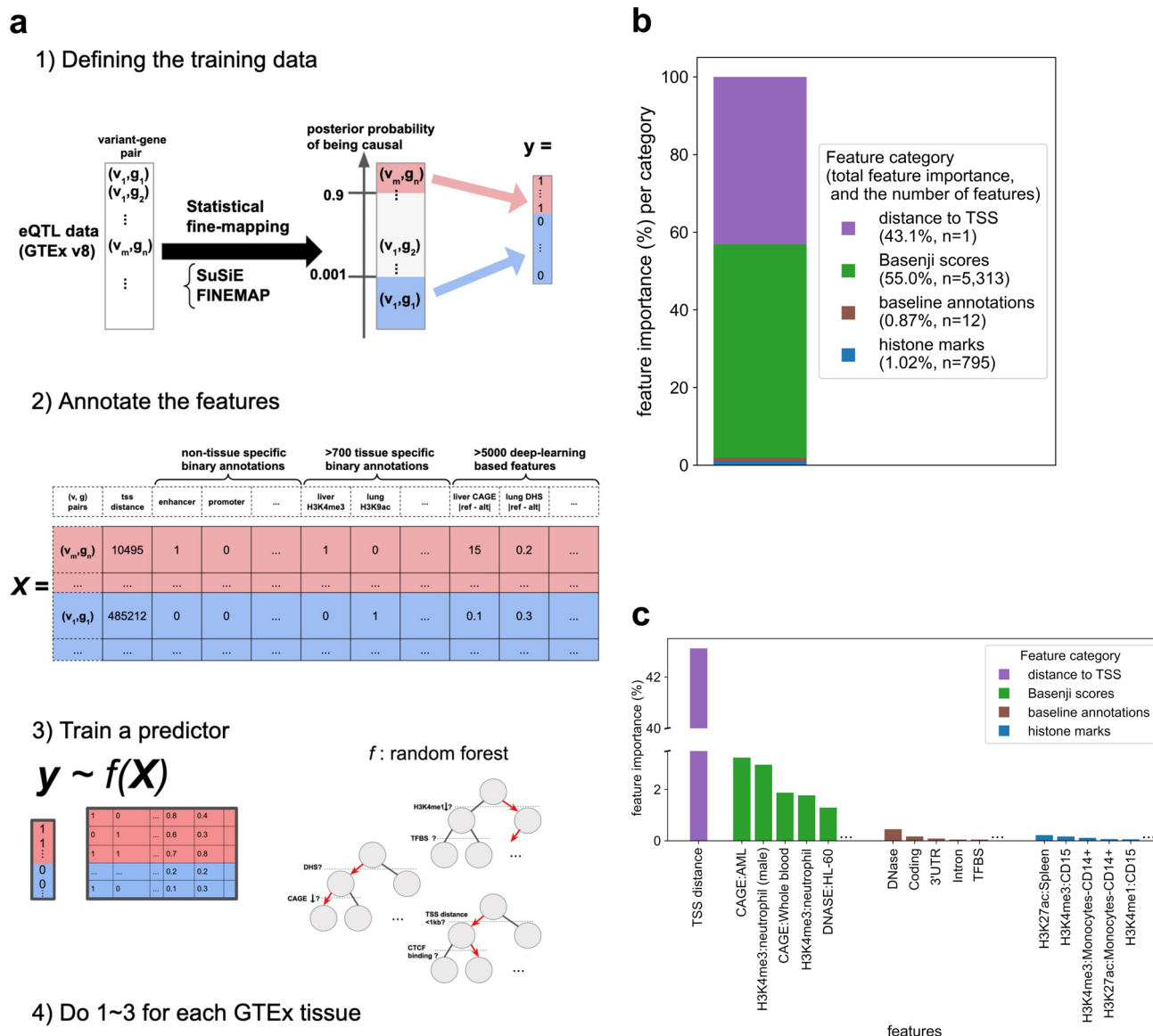
**Fig. 1** Examples of the enrichment of variant-gene pairs in whole-blood eQTL PIP bins for functional genomics features. Enrichments of variant-gene pairs in different posterior inclusion probability (PIP) bins in binary functional features (non-tissue specific (a), tissue-specific in peripheral blood mononuclear cells (b), deep learning-derived regulatory activity (CAGE<sup>46</sup>) prediction in neutrophils (c), and distance to TSS (d) are shown (*n* is the number of variant-gene pairs).

was among the top-performing methods in prioritizing experimentally suggested regulatory variants from reporter assay experiments<sup>12,40</sup>, despite not varying distance to TSS, the most informative feature (Fig. 3b, c, Supplementary Fig. 6, and “Methods”). Finally, EMS was also among the top-performing methods in prioritizing putative causal noncoding variants for hematopoietic traits in the UKBB dataset (17.6× for EMS, best, vs 17.1× for DeepSEA, the second best; Fig. 3d), although there are known differences between the genetic architectures of *cis*-gene expression and complex traits<sup>41</sup>. These results were consistent when we performed the same set of analyses in different datasets: hematopoietic traits in BioBank Japan<sup>42</sup> and lymphoblastoid cell line (LCL) eQTL in Geuvadis<sup>14,22</sup> (Supplementary Fig. 7).

**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<sub>EMS</sub>, in whole blood (“Methods”). As expected, we found that PIP<sub>EMS</sub> identified more putative causal eQTLs than the original PIP calculated with a uniform prior, denoted PIP<sub>unif</sub>. Specifically, 95.4% of variants with PIP<sub>unif</sub> > 0.9

also had PIP<sub>EMS</sub> > 0.9 (2152 out of 2255), while only 33.8% of variants with PIP<sub>EMS</sub> > 0.9 had PIP<sub>unif</sub> > 0.9 (1125 out of 3277; Fig. 4a). Similarly, credible sets mostly decreased in size (Fig. 4b and Supplementary Data 2). Previous work in functionally informed fine-mapping<sup>33</sup> 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 (1125 with EMS as a prior vs 269 with EMS adjusted to a max/min ratio of 100 as a prior; Supplementary Fig. 8).

We evaluated the quality of PIP<sub>EMS</sub> by comparing it with PIP<sub>unif</sub> and a publicly available eQTL fine-mapping result that uses distance to TSS as a prior<sup>16,23</sup> (denoted PIP<sub>DAP-G</sub>) in two ways (other methods for functionally informed fine-mapping based on expectation maximization<sup>31,32,35</sup> would be computationally intensive for a dataset this size, while the recently introduced PolyFun<sup>33</sup> is designed for complex traits). First, PIP<sub>EMS</sub> had the highest enrichment level of reporter assay QTLs<sup>40</sup> (raQTLs) in the PIP > 0.9 bin (16.8× vs 12.9× in PIP<sub>unif</sub> and 11.4× in PIP<sub>DAP-G</sub>, Fisher’s exact test  $p = 1.65 \times 10^{-2}$  between PIP<sub>EMS</sub> and PIP<sub>DAP-G</sub>; Fig. 4c). Second, complex trait causal noncoding variants were comparably enriched in PIP > 0.9 bins (Supplementary Fig. 9). These results suggest that PIP<sub>EMS</sub> is a valid measure for identifying putative causal *cis*-regulatory variants.

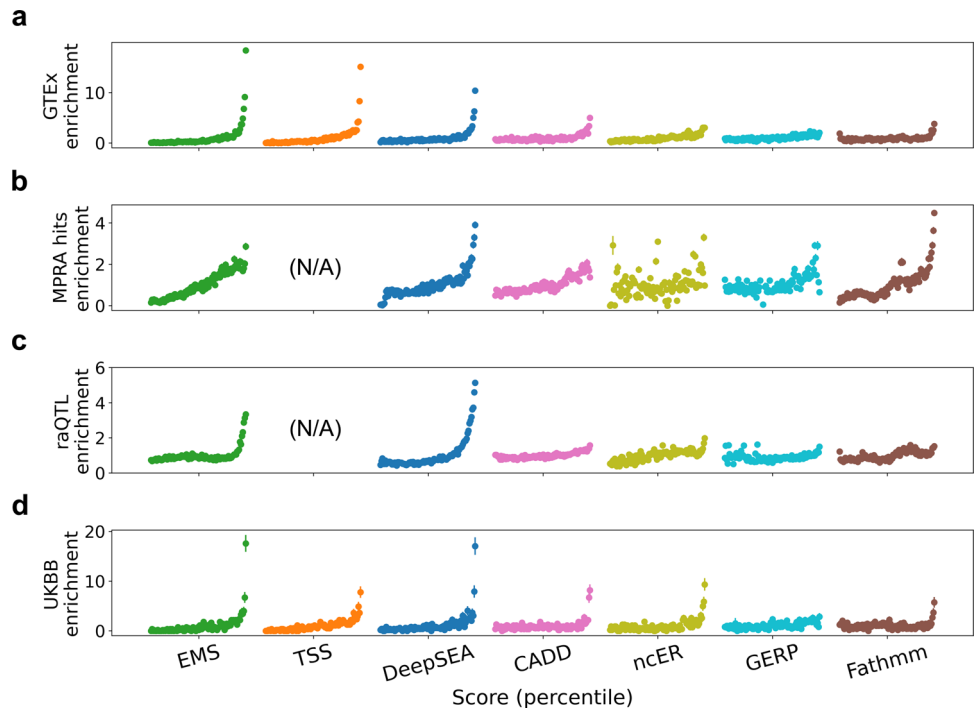


**Fig. 2 Schematic overview and feature importance of the expression modifier score (EMS).** **a** EMS is built by (1) defining the training data based on fine-mapping of GTEx v8 data, (2) annotating the variant–gene pairs with functional features, and (3) training a random forest classifier. We do this for each tissue. **b, c** Feature importance (mean decrease of impurity MDI<sup>59</sup>) 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.<sup>6</sup>, and histone marks are tissue-specific binary histone mark annotations from Roadmap<sup>5</sup>. In **b**,  $n$  is the number of features in the category.

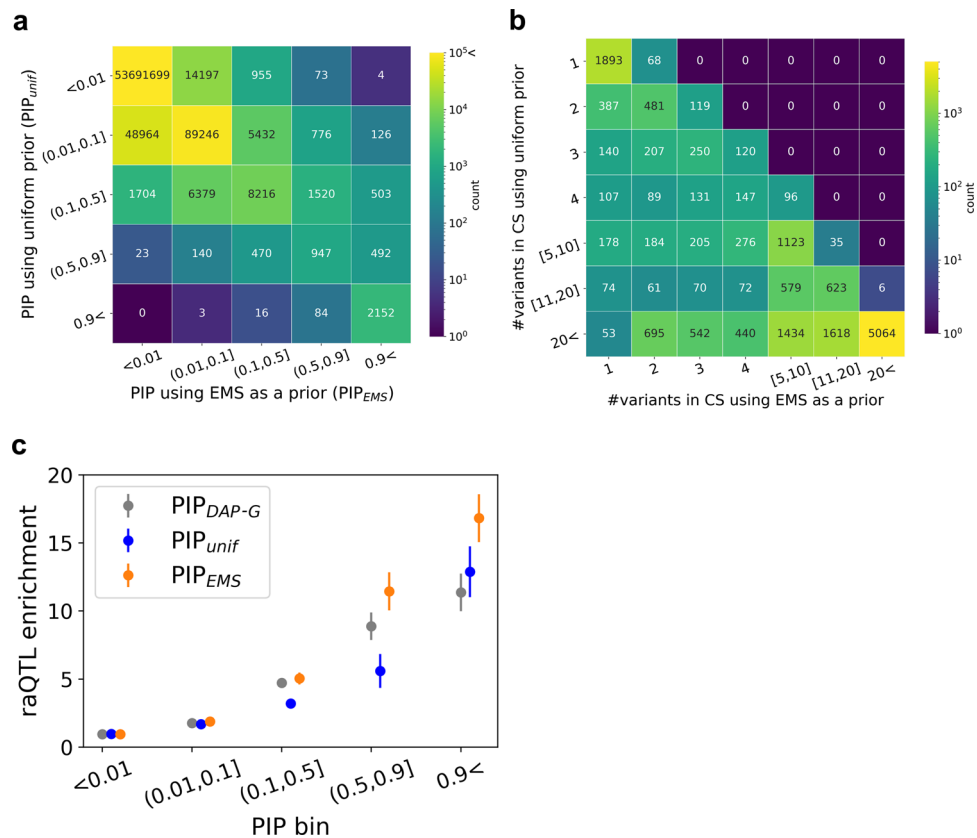
**Applying functionally informed PIP (PIP<sub>EMS</sub>) in gene prioritization across 95 traits.** We next compared the utility of PIP<sub>EMS</sub> to PIP<sub>unif</sub> for complex trait gene prioritization, as in Weeks et al.<sup>43</sup>. To do this, we first calculated PIP<sub>EMS</sub> for 49 GTEx tissues using EMS of matched tissues as priors (Supplementary Figs. 10 and 11), resulting in a total of 20,913 additional eQTLs with PIP<sub>EMS</sub> > 0.9 (Fig. 5a, Supplementary Fig. 12, and Supplementary Data 3). Tissue-specificity of putative causal eQTLs were characterized by enrichments of corresponding tissue-specific transcription factor (TF) activity scores in the Basenji model (Fig. 5b–d, Supplementary Figs. 13 and 14, and “Methods”). We then colocalized the eQTL signals with 95 UKBB phenotypes. Using the evaluation gene set described in ref. 43, PIP<sub>EMS</sub> achieved higher precision and higher recall than PIP<sub>unif</sub> (Table 1 and “Methods”). Overall, PIP<sub>EMS</sub> elucidated 310 candidate genes for UKBB phenotypes that were not identified with PIP<sub>unif</sub>

(Supplementary Data 4). On the other hand, PIP<sub>DAP-G</sub> showed lower precision than PIP<sub>EMS</sub> and PIP<sub>unif</sub> 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.

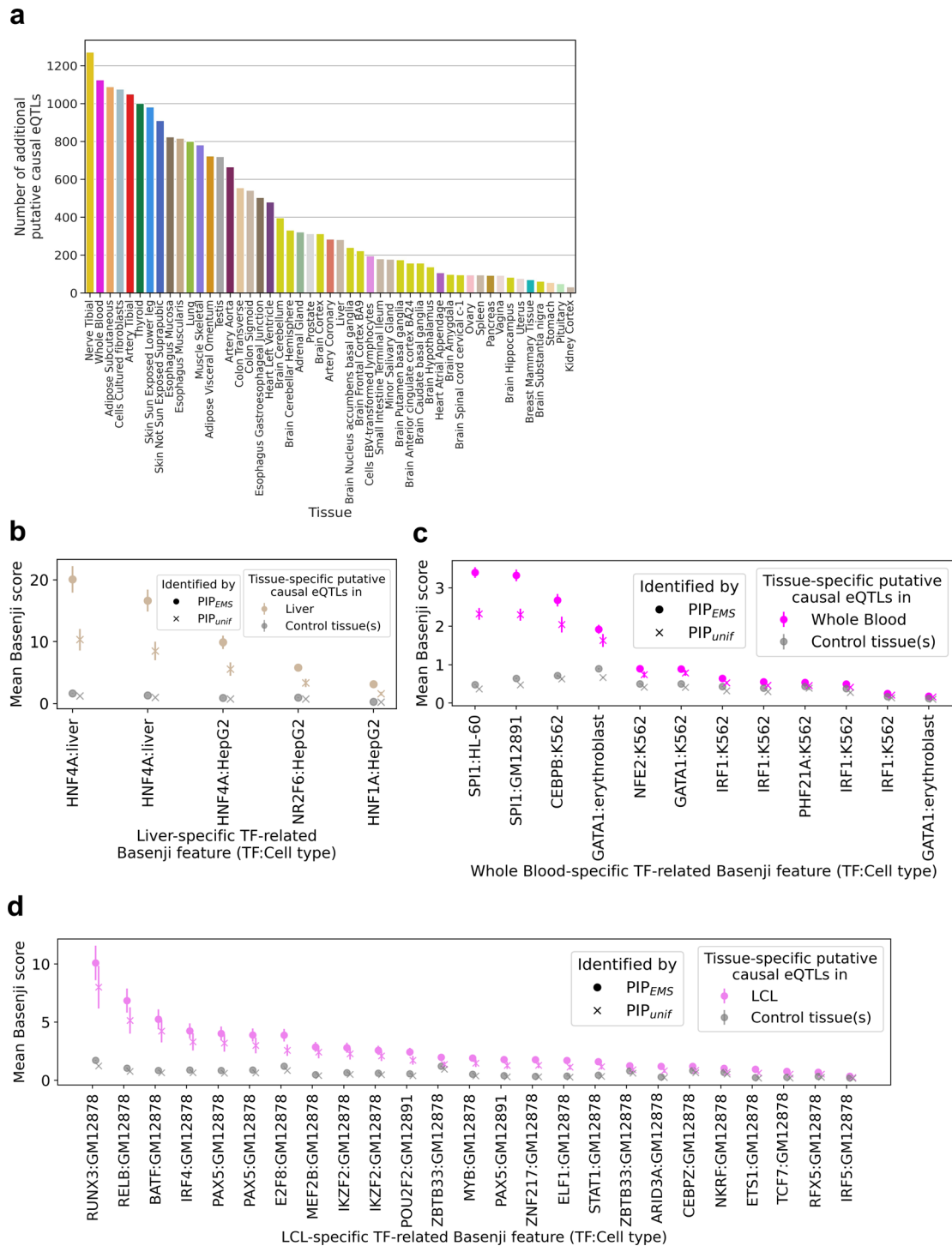
An example of PIP<sub>EMS</sub> resolving a credible set that is ambiguous with PIP<sub>unif</sub> is shown in Fig. 6. Here, four variants upstream of *CITED4* are in perfect LD in GTEx, giving PIP<sub>unif</sub> = 0.25 for all four (Supplementary Fig. 15). In UKBB, the four variants are also in high 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 < 4.53 × 10<sup>-2</sup> for all variants; “Methods”). However, one of the four variants, rs35893233, creates a binding motif of *SPI1*, a TF known to be involved in myeloid differentiation<sup>44,45</sup>, and presents epigenetic activity in myeloid-



**Fig. 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<sup>12</sup> (b), reporter assay QTLs<sup>40</sup> (raQTLs) (c), and putative hematopoietic-trait causal variants in UKBB (d) in different score percentiles.



**Fig. 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<sup>23</sup>, blue: uniform prior, orange: EMS as a prior).



**Fig. 5 Functionally informed fine-mapping across 49 tissues.** **a** 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. **b-d** Mean Basenji score in different classes of tissue-specific putative causal eQTLs for tissue-specific TF-related Basenji features for liver (**b**), whole blood (**c**), and LCLs (**d**). In 39 out of all 42 features across all three tissues, the mean Basenji score in tissue-specific putative causal eQTLs identified by  $PIP_{EMS}$  is significantly higher in the corresponding tissue than in control tissues ( $t$  test  $p < 0.05/42$ ). This changes to 36 in 42 when using  $PIP_{unif}$  instead of  $PIP_{EMS}$ . The enrichment of mean Basenji score in putative causal eQTLs in the corresponding tissue compared to control tissues is higher for  $PIP_{EMS}$  than  $PIP_{unif}$  for all 42 tissues ( $p < 10^{-100}$  in aggregate), consistent with our understanding that functionally informed fine-mapping using EMS utilizes cell-type-specific functional enrichments, identified from putative causal eQTLs identified with a uniform prior, to identify additional putative causal eQTLs. Duplicated names are distinct features corresponding to biological replicates in the TF activity measurements. Out of 17,960 tissue-specific putative causal eQTLs,  $n = 222$  were for liver (**b**),  $n = 1758$  were for whole blood (**c**), and  $n = 140$  were for LCL (**d**).

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

Method	Tool	Prior	Precision	Recall
PIP <sub>EMS</sub>	SuSiE	EMS	0.556	0.052
PIP <sub>unif</sub>	SuSiE	Uniform	0.525	0.039
PIP <sub>DAP-G</sub>	DAP-G	Distance to TSS	0.500	0.078

related cell types, such as showing the highest basenji score for cap analysis gene expression (CAGE)<sup>46</sup> activity in acute myeloid leukemia. This variant has >25× greater EMS than the other three variants ( $1.73 \times 10^{-3}$  vs  $6.11 \times 10^{-5}$ ,  $1.00 \times 10^{-5}$  and  $8.62 \times 10^{-6}$ , respectively), enabling PIP<sub>EMS</sub> to narrow down the credible set to the single variant (PIP<sub>EMS</sub> = 0.956 for rs35893233). Integrating EMS into the colocalization analysis thus allows identification of *CITED4* as a neutrophil count-related gene (CLPP = 0.173). Additional examples are described in Supplementary Fig. 16.

## 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 >6000 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 noncoding 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 colocalization analysis to identify 310 additional candidate genes for complex traits.

Evaluating predictors of noncoding 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 datasets, it did not outperform existing scores in prioritizing reporter assay active variants or putative complex trait causal noncoding 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<sup>41,47</sup>. 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 noncoding functions at base-pair resolution.

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<sup>48,49</sup>, and the GTEx cohort is highly biased towards adult samples with European ancestry background. Second, although we utilize over 6000 features in EMS, larger sets of variant and gene annotations, such as 3D configuration of genome<sup>50,51</sup>, constraint<sup>52–54</sup>, or pathway enrichment<sup>43</sup> 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.

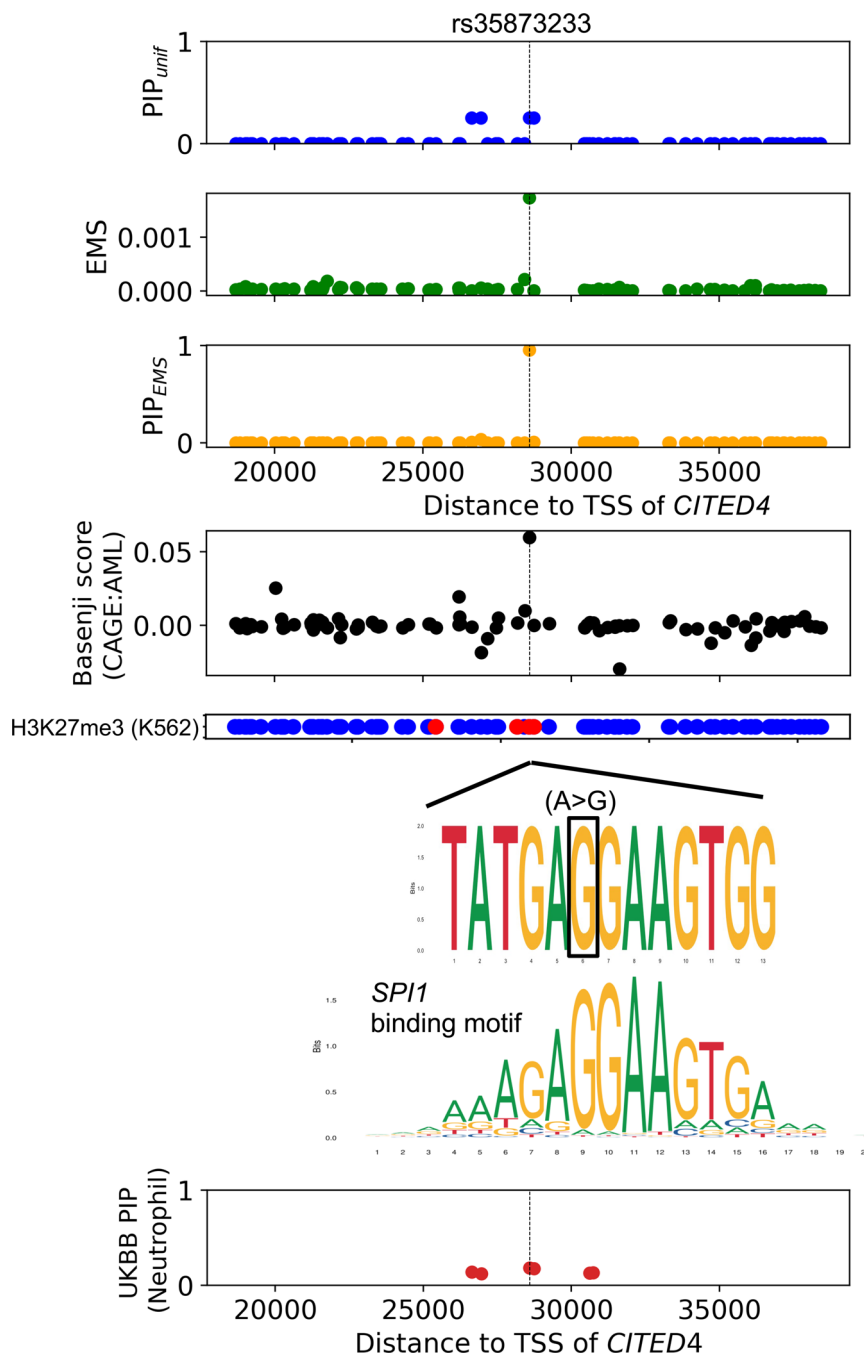
In this work, we focused on the task of predicting putative causal eQTLs. Future work could use a similar framework to predict putative causal splicing QTLs or other molecular QTLs for which statistical fine-mapping has identified a large number of high-PIP variants. In addition, although noisy effect size estimates from eQTL studies present a challenge, future work could explore leveraging features correlated with the sign and magnitude of effect (Supplementary Fig. 17) 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<sup>55</sup>. 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.

## Methods

**The expression modifier score.** Fine-mapping of GTEx v8 data is described in <https://www.finucanelab.org/data> 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<sup>19</sup> and Sum of Single Effects, SuSiE<sup>20</sup>) as positive, and the ones with PIP < 0.0001 for both methods as negative. Each variant–gene pair was annotated with 6121 features (distance to TSS annotated in the GTEx v8 dataset, 12 non-cell-type-specific binary features from the LDSC baseline model<sup>6</sup>, 795 cell-type-specific binary features from the Roadmap Epigenomics Consortium<sup>5</sup>, where variants falling in narrow peak are annotated as 1, and others are 0, and 5313 deep learning-derived cell-type-specific features generated by the Basenji model<sup>28,29</sup>; Supplementary Data 5). The 152 most predictive features were selected based on different prediction accuracy metrics, such as F1 measure and mean decrease of impurity 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 Data 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<sup>36</sup> v1.4 and GERP<sup>38</sup> scores were annotated using the hail<sup>56</sup> annotation database (<https://hail.is>), and nCR<sup>39</sup> scores were downloaded from [https://github.com/TelentiLab/ncER\\_datasets](https://github.com/TelentiLab/ncER_datasets). In order to annotate the DeepSEA<sup>26</sup> v1.0 and Fathmm<sup>37</sup> v2.3 noncoding 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 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 (Supplementary Fig. 5), 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 of enrichment 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.



**Fig. 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<sup>46</sup> 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<sup>60</sup> of the alternative allele aligned with the binding motif<sup>61</sup> of *SPI1*, and PIP for neutrophil count in UKBB (<https://www.finucanelab.org/data>, ref. <sup>34</sup>) with uniform prior.

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

Saturation mutagenesis data<sup>12</sup> was downloaded from the MPRA data access portal (<http://mpras.washington.edu>). An MPRA hit was defined as having a bonferroni-significant association  $p$  value ( $<0.05$  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<sup>40</sup> was downloaded from <https://osf.io/w5bzq/wiki/home/>. EMS was rescaled to have a constant distance to TSS (200 bp, roughly representing the scale of typical distance to TSS in plasmids<sup>12</sup>), 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 <https://www.finucanelab.org/data>. To focus on noncoding regulatory effects, we annotated the variants in VEP<sup>57</sup> v85 and filtered out coding and splice variants for the UKBB dataset. For each (noncoding) 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  $> 0.9$  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 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  $\alpha_1, \dots, \alpha_L$ , 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  $\alpha_l$ , and credible sets that are not pure—i.e., that contain a pair of variants with absolute correlation  $< 0.5$ —are pruned out. The  $\alpha_l$  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. <sup>32</sup>, and proceeds as follows. For each gene and each tissue, we start with  $\alpha_1, \dots, \alpha_L$  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  $\alpha_l$  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  $w_1 \dots w_m$  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  $\alpha_l$  corresponds to a pure credible set; otherwise, we set  $\hat{\alpha}_l = \alpha_l$ . We then use the updated  $\hat{\alpha}_1, \dots, \hat{\alpha}_L$  to compute updated PIPs and credible sets, as in the original SuSiE method. See Supplementary Methods for further details.

**Performance evaluation of PIP<sub>EMS</sub> and application to gene prioritization.** PIP using distance to TSS as a prior (PIP<sub>DAP-G</sub>) 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 noncoding 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.<sup>43</sup>):

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 “evaluation gene” (protein-coding variant with PIP  $> 0.5$ ) for the same trait. There were 2897 such regions and 1161 evaluation 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 noncoding 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 colocalization 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 evaluation 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.

**Tissue-specific putative causal eQTL analysis.** Tissue-specific putative causal eQTL in a tissue was defined as a variant–gene pair with PIP<sub>EMS</sub>  $> 0.9$  in the tissue and PIP<sub>EMS</sub>  $< 0.1$  in all the other tissues (including cases where a variant is missing in a tissue; Supplementary Data 7). A tissue-specific putative causal eQTL pair was defined as a pair of tissue-specific putative causal eQTL on a same gene in two different tissues, existing within 10 kb distance (Supplementary Fig. 14 and Supplementary Data 8). Basenji features were classified as TF related if the feature name contains the gene symbol classified as a human TF in an external database<sup>58</sup> (<http://humantfs.ccrb.utoronto.ca/download.php>).

Then for each TF, we defined it as specific for tissue  $T$  if the expression level (TPM) of the TF was higher in  $T$  than in all other tissues and was  $> 2$  standard deviations away from the mean expression level across tissues. All the tissues for which the TF had expression level ten times lower than that of tissue  $T$  were defined as control tissues. TF-related Basenji features with no specific tissue, or lacking control tissues were filtered out. We also filtered out the features where the TF specificity and the assay cell type did not clearly match (Supplementary Data 9).

This resulted in 42 TF-related Basenji features corresponding to 30 unique TFs. Enrichment of each TF-related Basenji feature was examined by comparing the average score in the tissue-specific putative causal eQTLs for the corresponding tissue with the average in the control tissues, using a  $t$  test (Supplementary Data 9).

**Statistical analysis.** All the statistical tests were two-sided. No adjustment was made in the  $p$  value we report.

Error bar in Fig. 5b–d and Supplementary Fig. 13 is defined as the standard error of the mean.

Error bar in the enrichment analyses (all the other figures, where error bars are present) are explained in the “Computation of enrichment” section in the “Methods”. The set of software used for data generation, statistical analysis, and plotting in the study are listed below:

SuSiE v0.8.1.0521 (<https://github.com/stephenslab/susie-paper>)

FINEMAP v1.3.1 (<http://www.christianbenner.com>)

ggseqlogo (<https://cran.r-project.org/web/packages/ggseqlogo/index.html>)

basenji v0.0.1 (<https://github.com/calico/basenji>)

brokenaxis v0.3.1 (<https://pypi.org/project/brokenaxes/>)

joblib v0.11 (<https://joblib.readthedocs.io>)

hail v0.2.26 (<https://hail.is>)

matplotlib v3.2.0 (<https://matplotlib.org>)

numpy v1.18.1 (<https://numpy.org>)

pandas v1.0.1 (<https://pandas.pydata.org>)

scikit-learn v0.21.3 and v0.23.2 (<https://scikit-learn.github.io/stable>)

scipy v1.2.1 (<http://scikit-learn.github.io/stable>)

seaborn v0.9.0 (<https://seaborn.pydata.org>).

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

EMS for 49 tissues are available at <https://www.finucanelab.org/data>. 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](https://github.com/TelentiLab/ncER_datasets). DeepSEA v1.0 scores were downloaded from <https://humanbase.readthedocs.io/en/latest/deepsea.html>. Fathmm v2.3 noncoding scores were downloaded from <http://fathmm.biocompute.org.uk>. Saturation mutagenesis data was downloaded from the MPRA data access portal (<http://mprags.washington.edu>). The raQTL data was downloaded from <https://osf.io/w5bzq/wiki/home/>. Human transcription factor (TF) data was downloaded from <http://humantfs.ccrb.utoronto.ca/download.php>. The UKBB fine-mapping results are deposited at <https://www.finucanelab.org/data>.

## Code availability

Code used in this manuscript is available at [https://github.com/FinucaneLab/Expression\\_Modifier\\_Score/](https://github.com/FinucaneLab/Expression_Modifier_Score/).

Received: 23 October 2020; Accepted: 15 April 2021;

Published online: 07 June 2021

## References

- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Paul, D. S., Soranzo, N. & Beck, S. Functional interpretation of non-coding sequence variation: concepts and challenges. *Bioessays* **36**, 191–199 (2014).
- Maller, J. B. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- 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).
- Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
- Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
- Trynka, G. & Raychaudhuri, S. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. *Curr. Opin. Genet. Dev.* **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. *Nat. Commun.* **10**, 3583 (2019).
13. Tian, R. et al. Pitfalls in single clone CRISPR-Cas9 mutagenesis to fine-map regulatory intervals. *Genes* **11**, 504 (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. *Nat. Rev. Genet.* **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. *J. R. Stat. Soc. B* **82**, 1273–1300 (2020).
21. Hormozdiani, F., Kostem, E., Kang, E. Y., Pasiński, 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. *Nat. Genet.* **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. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).
24. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* **11**, e1005176 (2015).
25. Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
26. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. 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.* **28**, 739–750 (2018).
29. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).
30. Kopp, W., Monti, R., Tamburrini, A., Ohler, U. & Akalin, A. Deep learning for genomics using Janggu. *Nat. Commun.* **11**, 3488 (2020).
31. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
32. Jiang, J. et al. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun. Biol.* **2**, 1–12 (2019).
33. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
34. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203 (2018).
35. 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).
36. Rentsch, 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).
37. Shihab, H. A. et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
38. Cooper, G. M. et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
39. Wells, A. et al. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* **10**, 5241 (2019).
40. van Arensbergen, J. et al. High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160–1169 (2019).
41. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).
42. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
43. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Preprint at *medRxiv* <https://doi.org/10.1101/2020.09.08.20190561> (2020).
44. Chen, H. et al. PU.1 (Spi-1) autoregulates its expression in myeloid cells. *Oncogene* **11**, 1549–1560 (1995).
45. 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).
46. 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).
47. Inoue, F. et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).
48. LaPierre, N. et al. Identifying causal variants by fine mapping across multiple studies. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.01.15.908517> (2020).
49. Hutchinson, A., Watson, H. & Wallace, C. Improving the coverage of credible sets in Bayesian genetic fine-mapping. *PLOS Computational Biol.* **16**, e1007829 (2020).
50. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* **21**, 207–226 (2020).
51. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* **17**, 1111–1117 (2020).
52. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
53. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
54. Iulio, J. et al. The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333 (2018).
55. Schoech, A. P. et al. Negative short-range genomic autocorrelation of causal effects on human complex traits. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.09.23.310748> (2020).
56. Hail Team. Hail 0.2. <https://github.com/hail-is/hail>(2020).
57. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
58. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
59. Louppe, G. Understanding random forests: from theory to practice. Preprint at <https://arxiv.org/abs/1407.7502> (2015).
60. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
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).

## Acknowledgements

We thank Yakir Reshef, Jesse Engritz, 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.

## Author contributions

Q.S.W., D.G.M., and H.K.F. designed the study. Q.S.W., D.R.K., J.U., and S.S. analyzed the data. Q.S.W. and H.K.F. wrote the manuscript with input from all authors (D.R.K., J.U., M.K., S.S., R.C., C.A., N.C., Y.O., B.B.J., F.A., K.G.A., and D.G.M.).

## 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.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-23134-8>.

**Correspondence** and requests for materials should be addressed to Q.S.W. or H.K.F.

**Peer review information** *Nature Communications* thanks Anshul Kundaje and the other, anonymous, reviewer for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

## The Biobank Japan Project

Koichi Matsuda<sup>12,13</sup>, Yuji Yamanashi<sup>14</sup>, Yoichi Furukawa<sup>15</sup>, Takayuki Morisaki<sup>16</sup>, Yoshinori Murakami<sup>17</sup>, Yoichiro Kamatani<sup>13,18</sup>, Kaori Muto<sup>19</sup>, Akiko Nagai<sup>19</sup>, Wataru Obara<sup>20</sup>, Ken Yamaji<sup>21</sup>, Kazuhisa Takahashi<sup>22</sup>, Satoshi Asai<sup>23,24</sup>, Yasuo Takahashi<sup>25</sup>, Takao Suzuki<sup>26</sup>, Nobuaki Sinozaki<sup>26</sup>, Hiroki Yamaguchi<sup>27</sup>, Shiro Minami<sup>28</sup>, Shigeo Murayama<sup>29</sup>, Kozo Yoshimori<sup>30</sup>, Satoshi Nagayama<sup>31</sup>, Daisuke Obata<sup>32</sup>, Masahiko Higashiyama<sup>33</sup>, Akihide Masumoto<sup>34</sup> & Yukihiro Koretsune<sup>35</sup>

<sup>12</sup>Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>13</sup>Laboratory of Clinical Genome Sequencing, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. <sup>14</sup>Division of Genetics, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>15</sup>Division of Clinical Genome Research, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>16</sup>Division of Molecular Pathology, IMSUT Hospital Department of Internal Medicine, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>17</sup>Department of Cancer Biology, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>18</sup>Laboratory of Complex Trait Genomics, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. <sup>19</sup>Department of Public Policy, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. <sup>20</sup>Department of Urology, Iwate Medical University, Iwate, Japan. <sup>21</sup>Department of Internal Medicine and Rheumatology, Juntendo University Graduate School of Medicine, Tokyo, Japan. <sup>22</sup>Department of Respiratory Medicine, Juntendo University Graduate School of Medicine, Tokyo, Japan. <sup>23</sup>Division of Pharmacology, Department of Biomedical Science, Nihon University School of Medicine, Tokyo, Japan. <sup>24</sup>Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan. <sup>25</sup>Division of Genomic Epidemiology and Clinical Trials, Clinical Trials Research Center, Nihon University School of Medicine, Tokyo, Japan. <sup>26</sup>Tokushukai Group, Tokyo, Japan. <sup>27</sup>Department of Hematology, Nippon Medical School, Tokyo, Japan. <sup>28</sup>Department of Bioregulation, Nippon Medical School, Kawasaki, Japan. <sup>29</sup>Tokyo Metropolitan Geriatric Hospital and Institute of Gerontology, Tokyo, Japan. <sup>30</sup>Fukujuji Hospital, Japan Anti-Tuberculosis Association, Tokyo, Japan. <sup>31</sup>The Cancer Institute Hospital of the Japanese Foundation for Cancer Research, Tokyo, Japan. <sup>32</sup>Center for Clinical Research and Advanced Medicine, Shiga University of Medical Science, Shiga, Japan. <sup>33</sup>Department of General Thoracic Surgery, Osaka International Cancer Institute, Osaka, Japan. <sup>34</sup>IIZUKA-HOSPITAL, Fukuoka, Japan. <sup>35</sup>National Hospital Organization Osaka National Hospital, Osaka, Japan.