Nucleic Acids Research, 2021 **1**
https://doi.org/10.1093/nar/gkab924

Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome

Shengcheng Dong[1] and Alan P. Boyle[1,2,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA and [2]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

Received March 18, 2021; Revised September 21, 2021; Editorial Decision September 24, 2021; Accepted September 27, 2021

ABSTRACT

Understanding the functional consequences of genetic variation in the non-coding regions of the human genome remains a challenge. We introduce h ere a computational tool, TURF, to prioritize regulatory variants with tissue-specific function by leveraging evidence from functional genomics experiments, including over 3000 functional genomics datasets from the ENCODE project provided in the RegulomeDB database. TURF is able to generate prediction scores at both organism and tissue/organ-specific levels for any non-coding variant on the genome. We present that TURF has an overall top performance in prediction by using validated variants from MPRA experiments. We also demonstrate how TURF can pick out the regulatory variants with tissue-specific function over a candidate list from associate studies. Furthermore, we found that various GWAS traits showed the enrichment of regulatory variants predicted by TURF scores in the trait-relevant organs, which indicates that these variants can be a valuable source for future studies.

INTRODUCTION

Characterizing the biological impact of variation in the non-coding regions of the human genome remains a challenge in the interpretation of human diversity. Genome-wide association studies (GWAS) have identified millions of genetic variants that are associated with diverse disease traits (1). Most of these variants (∼90%) map to the non-coding regions of the human genome (2). Due to the lack of understanding of these regulatory elements within non-coding regions, it is important to assess the functional consequences of these disease-related variants from GWAS.

To facilitate studies of non-coding genomic regions, large consortia, including ENCODE (3,4) and the Roadmap Epigenomics projects (5), have defined the human regulatory landscape using high-throughput functional genomics assays. For example, DNase-seq locates open chromatin regions in the genome (6,7), while ChIP-seq identifies chromatin modification patterns and transcription factor (TF) binding sites within regulatory elements (8–10). With further incorporation of variant genotypes into these methods, variants associated with differential TF binding and chromatin states have been described (11–13). In addition, massively parallel reporter assays (MPRA) identify regulatory variants that affect gene expression levels directly (14–16). More recently, SNP-SELEX assays assess the binding affinity between TFs and SNP-containing genomic sequences in the non-coding regions (17). These studies demonstrate that a significant number of variants drive regulatory state variation across the population, and potentially explain the diversity in disease risk and phenotype observed from GWAS studies.

Computational tools have helped prioritize regulatory variants in non-coding regions by leveraging knowledge from functional genomics assays. Prediction scores of functional probability for variants are available from tools including RegulomeDB (18), GWAS3D (19), HaploReg (20), gkm-SVM (21), DeepSEA (22), DeepBind (23), DanQ (24) and Basenji (25). The process of narrowing down a candidate list of variants using these prediction scores can reduce time-consuming validation experiments. However, most current computational tools overlook the uniqueness of gene regulatory networks found within different tissues by only providing a prediction score at an organism level. This can be misleading for research groups focused on tissue-specific functional variants. New tools have recently become available that provide tissue-specific prediction scores or prioritize relevant tissues for candidate regulatory variants, such as FUN-LAD (26), GenoNet (27), cepip (28), GenoSkyline (29) and Motif-Raptor (30). However, they mainly utilize epigenetic data from the Roadmap Epigenomics project (5) making it hard to leverage their results against other tissues not included in the Roadmap project. The ENCODE project currently houses thousands of ChIP-seq and DNase-seq datasets in over 200 tissues and

*To whom correspondence should be addressed. Tel: +1 734 763 7382; Fax: +1 734 763 7382; Email: apboyle@umich.edu

cell types, including those from the Roadmap project, that can further increase the scale and accuracy of tissue-specific function prediction.

Here, we introduce a computational tool, TURF (Tissue-specific Unified Regulatory Features), that prioritizes regulatory variants in the non-coding regions of the human genome. TURF is built on our RegulomeDB framework to allow for easy delivery of our predictions as well as constant updates in the functional annotations across the human genome. We extend our previous algorithm SURF (31) to predict tissue-specific functional variants in addition to the tool's original generic context at an organism level. To construct a high-quality training set, we called 7,530 allele-specific TF binding (ASB) single nucleotide variants (SNVs) in six cell lines from over 600 ChIP-seq datasets. We then trained a random forest model using features from functional genomic annotations across all available tissues from ENCODE. This classifier greatly improves the robustness of RegulomeDB v1.1 ranking scores. We then incorporated annotations of histone marks and open chromatin regions in a particular tissue to train a separate random forest model and obtain a final tissue-specific score. The tissue-specific score leverages information from other tissues, as well as retaining the uniqueness of individual tissues and surpasses other top-performing tools. Moreover, we extended the tissue-specific scores to organ-specific scores in the 51 organs with available genomics data from the ENCODE project. The pre-calculated organ-specific scores for all GWAS SNVs from the GWAS Catalog are available at https://github.com/Boyle-Lab/RegulomeDB-TURF and TURF is currently being integrated into RegulomeDB v2.0.

MATERIALS AND METHODS

Training dataset generation

We identified 7530 allele specific transcription factor (TF) binding (ASB) SNVs in six cell lines (GM12878, HepG2, A549, K562, MCF7 and H1hESC), which are defined as variants that result in stronger binding of a TF to one allele at heterozygous sites in an individual. The *AlleleDB* protocol was used to call ASB SNVs (32).

The SNVs in GM12878 and H1hESC were obtained from the 1000 Genome Project (33) and NCBI GEO database (accession number: GSE52457) separately. For the other four cell lines, variants were called from their whole genome sequencing data (data accessible at NCBI SRA database with accession numbers: DRX015191, SRX2598759, SRX285595 and SRX1705314) by *HaplotypeCaller* from the Genome Analysis Toolkit (GATK) v3.6 (34) following GATK's Best Practices (https://gatk.broadinstitute.org/). Their diploid personal genomes were constructed using *vcf2diploid* v0.2.6 (35) to avoid alignment biases favoring reads containing reference alleles by mapping to maternal and paternal genomes separately. Copy number variation regions with a read depth of <0.5 or >1.5 called from *CNVnator* v0.3.3 (36) were filtered out.

The *AlleleDB* pipeline was run on 864 ChIP-seq datasets in the six cell lines from the ENCODE project. In addition to the standard steps in *AlleleDB*, our ASB set was refined by performing beta-binomial tests only within reads overlapping their corresponding TF binding peaks called from the same ChIP-seq dataset. In total, 7530 ASB SNVs were identified from 638 ChIP-seq datasets.

The ASB SNVs were treated as positive examples in our random forest model. To generate a comparable negative set, we included SNVs from three sources: (i) The 55 611 non-allelic TF binding SNVs, defined by having equal ChIP-seq read counts on two alleles at heterozygous site. (ii) The closest variants from each of the SNVs in positive set and outside ChIP-seq peaks (6373 unique variants in total). (iii) A randomly selected set of 1000 genome variants scored no hits on functional annotations from RegulomeDB v1.1. Those three negative sets were combined and weighted equally in our model. The number of training SNVs in each cell line is shown in Supplemental Table S1.

Building random forest models

For TURF generic scores, seven binary and eight numeric features were created for each variant in the training set (Supplemental Table S2). The seven binary features represent if the variant position overlaps corresponding functional genomic regions by querying RegulomeDB 2.0. Custom scripts were written to retrieve annotations from the RegulomeDB web server. The maximum information content change from PWM was calculated based on the query. Quantiles and variations in ChIP-seq signals pre-calculated from all available bigwig files in ENCODE and functional significance scores from DeepSEA were also incorporated. A random forest model was trained to make predictions on the probability of a query variant being functional. The *scikit-learn 0.20.3* python package was used to train the random forest model, setting the number of trees to 500. The feature importance was calculated based on the mean decrease of impurity from random forest model (Supplemental Table S3).

For TURF tissue-specific scores, a separate random forest model was built with seven binary tissue-specific features (see feature list in Supplemental Table S2). When training with each ASB cell line, the ASB SNVs in the corresponding cell line were labeled as positive variants, while the other variants were labeled as controls. The *scikit-learn 0.20.3* python package was used, setting the class_weight option as 'balanced'.

Generic scores performance assessment

We evaluated our generic model performance on an independent dataset from an MPRA assay in GM12878 (14). The labels of the MPRA variants (435 positive variants, 2670 control variants) and prediction scores from DeepSEA (22) and regBase were downloaded from regBase database (37). The performance of different tools was assessed on the Area Under ROC Curve (AUROC) and the Area Under Precision-Recall Curve (AUPR).

Tissue-specific scores performance assessment

The tissue-specific model's performance was evaluated first on three MPRA datasets in GM12878 (E116), HepG2 (E118) and K562 (E123). The labels for the MPRA variants were obtained from GenoNet (27). The authors labeled

the MPRA variants in GM12878 from (14) with a slightly different criteria than regBase (37), resulting in 293 positive variants and 2772 control variants. The MPRA variants in HepG2 and K562 were from (15), where 524 positive variants and 1439 control variants were in HepG2, and 339 positive variants and 1361 control variants were in K562. The same evaluation process as described in GenoNet (27) was used to compare TURF to other available tools, including DeepSEA (22), CADD (38) and GenoSkyline (29). In detail, we calculated AUROC, AUPR and the correlation coefficient using 1000 replicates of 4:1 random partition of each MPRA dataset. For the divided five parts, four parts were used for training while the remaining part was used for testing.

When evaluating performance with allele specific TF binding SNVs, pre-calculated scores from GenoNet (27) and GenoSkyline (29) were downloaded from https://zenodo.org/record/3336209/files/ and http://zhaocenter.org/GenoSkyline.

Extension to organ-specific scores

The mapping from tissues and cell types (i.e. biosamples) to organ names was downloaded from the ENCODE website (https://www.encodeproject.org/report/?type=BiosampleType). When generating organ-specific prediction scores, we combined the annotations from functional genomics data in all biosamples belonging to the corresponding organ. Fifty-one of fifty-five organs had available ChIP-seq data of histone marks and DNase-seq data to generate organ-specific scores.

Organ-specific significance scores

We calculated organ-specific significance scores relative to a background set from GWAS variants. The GWAS variants were downloaded and assigned to their mapped traits from the GWAS Catalog (1). SNVs on chromosomes 1–22 and chromosome X were the only ones considered for the organ-specific scoring. Linkage disequilibrium (LD) expansion was performed by including SNVs from the 1000 genome project that are in strong LD (R^2 threshold of 0.6, precalculated R^2 values downloaded from gs://genomics-public-data/linkage-disequilibrium) with any GWAS SNV (39). To convert each organ-specific score to a significance score, we calculated the portion of GWAS variants with a greater score in the corresponding organ and did a negative $\log_{10}$ transformation on to the portion (Figure 5).

Organ-specific scores enrichment of GWAS traits

In the enrichment analysis, we focused on the GWAS traits with the enrichment of regulatory variants, which have at least 20 GWAS SNVs and at least 5% of the LD-expanded GWAS SNVs in the trait that have TURF generic scores no less than 0.8 (400 traits in total).

To test the enrichment of organ-specific regulatory variants, each GWAS trait set was first sampled with an equal sized background set from all GWAS SNVs from any trait. Subsequent LD expansion was performed on both the trait set and background set (with a more strict R^2 threshold of 0.8). To reduce the dependencies across SNVs within each set, the SNVs were pruned on each organ individually so that no two SNPs were within 1MB of each other in the same set. Each SNV in decreasing order on organ-specific score was considered, and only retained a SNV if there was no other SNV within 1 Mb. After the pruning process, a *P*-value was computed from the Mann–Whitney U test for each organ-trait combination, with the alternative hypothesis as SNVs in the trait set have greater organ-specific scores than the background set. This test was repeated by sampling 100 versions of the background set and a total of 100 *P*-values were obtained for each organ-trait pair. 159 traits had at least one organ passing multiple test correction with an FDR of 5%, applied with the Holm-Sidak test from the python package *statsmodels* v0.12.1.

To determine the top organ for each trait, overall high scores of the trait were compared to other organs. The negative log-transformed p-values from the U tests were used to compute the z-score of each organ over all 51 organs. The mean z-scores over 100 iterations for each organ-trait pair were calculated and hierarchical clustering on the 51 organs was performed using the ward linkage method. The final heatmap (Figure 6 and Supplemental Figure S5) only shows organ-trait pairs with a z-scores mean higher than 0 and passing multiple test correction (FDR threshold of 5%).

RESULTS

Overview of the TURF algorithm

TURF prioritizes non-coding variants with both generic scores and tissue-specific scores (Figure 1). It first uses a random forest model built by training on features from functional genomics annotations in all available tissues and cell types from the ENCODE project (3). It uses a similar feature set to our previously successful algorithm SURF (31), including binary features retrieved from the original RegulomeDB ranking scheme and functional significance scores from DeepSEA (22). Furthermore, it includes continuous signals from ChIP-seq assays to increase the resolution of the algorithm (see features list in Supplemental Table S2). Generic scores from the first random forest model predict whether the query variant is functional in any human tissue. Tissue-specificity is further predicted by using a separate random forest model trained on functional genomic annotation features only from a particular tissue. To avoid data availability bias for different tissues, TURF takes advantage of DNase-seq and well-studied histone mark ChIP-seq data that cover most tissues (see features list in Supplemental Table S2). By combining the probability score from the second random forest model with the generic score from the first model, the resulting tissue-specific score predicts the probability of the query variant being functional in a specific tissue.

TURF generic score improves the performance of RegulomeDB v1.1 ranking score

TURF improves on the original heuristic ranking score in RegulomeDB v1.1 by providing a probabilistic score generated from a random forest model. By replacing the sin-

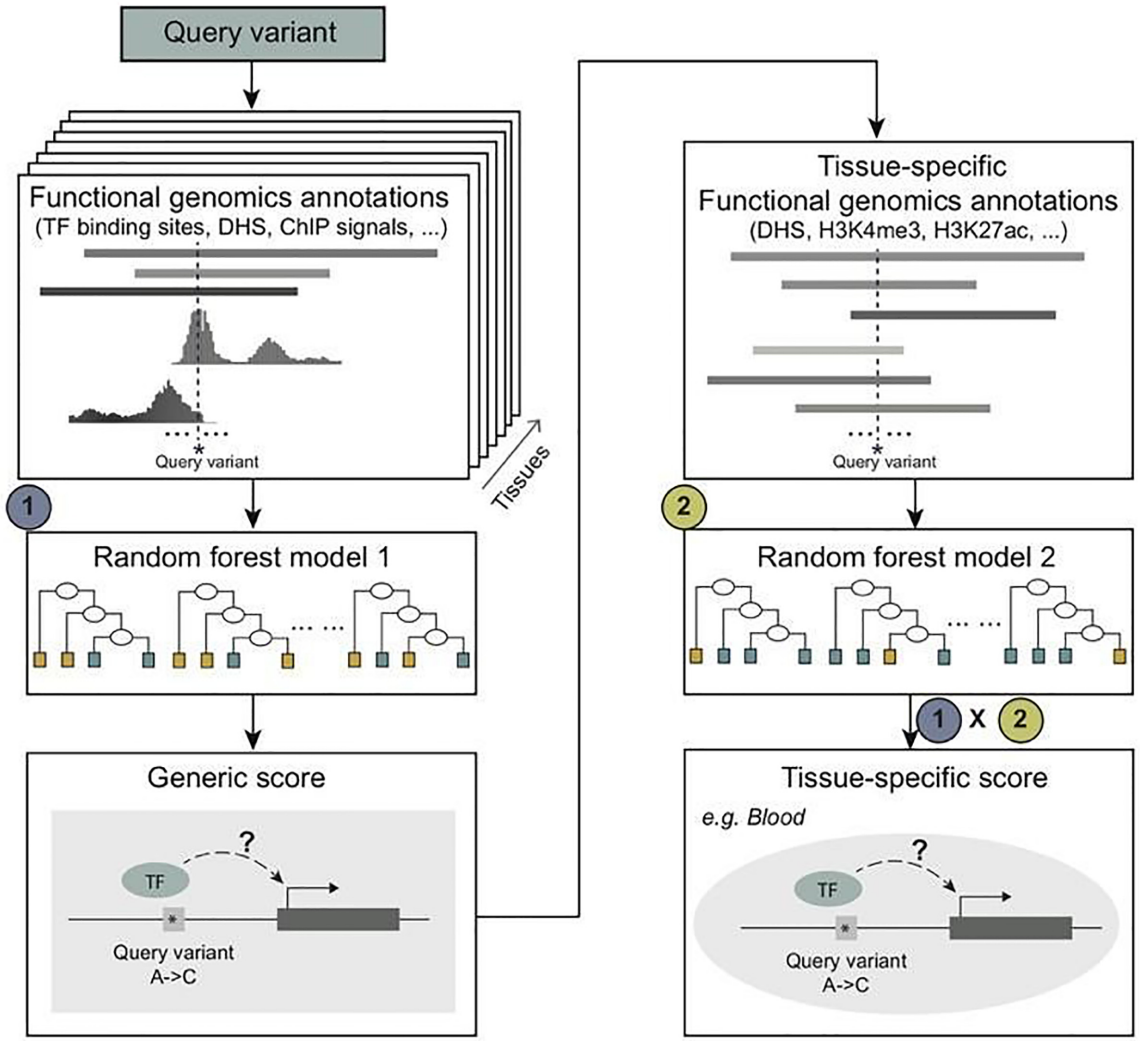

Figure 1. Overview of TURF algorithm. TURF generic score predicts the probability of a query variant being functional in any tissue from the first random forest, which used features of functional genomics annotations from all available tissues. By further incorporating annotations from a given tissue, a tissue-specific prediction score is computed by multiplying generic score with the prediction score from a second random forest model.

gle empirical decision with sets of decision trees, the model avoids issues caused by excessive reliance on only a few functional genomic annotations. To develop a training set for the model, we generated a set of variants with high confidence functional confidence through identification of 7,530 allele specific transcription factor binding (ASB) single nucleotide variants (SNVs) in six cell lines (GM12878, HepG2, A549, K562, MCF7 and H1hESC) by reprocessing 864 ChIP-seq datasets from the ENCODE project using *AlleleDB* v2.0 (32). ASB SNVs were called if different TF binding affinity with a single nucleotide change at heterozygous sites was observed. We defined a background set using non-allele specific TF binding SNVs as well as a set of variants outside TF binding regions (see methods).

We evaluated the TURF generic score performance on an independent and orthogonal dataset from a massively parallel reporter assay (MPRA) (14). This dataset was also utilized as a test set in a previous paper (37), where the authors found DeepSEA scores provided the best prediction model for calling variants functional in tissues. TURF performed on-par with DeepSEA scores on this MPRA test set with a larger AUROC and the same AUPR (Figure 2). To compare with the original ranking score from RegulomeDB v1.1, we calculated TURF generic scores for all common SNPs from dbSNP153 (40). The SNPs that originally scored in the highest category, which was largely dominated by eQTL evidence, now show a wider range of scores that better predicted their functionalities, while the overall trend was unchanged (Supplemental Figure S1).

TURF tissue-specific scores performance on MPRA data in three cell lines

We further evaluated TURF tissue-specific predictions with MPRA datasets from three cell lines (GM12878, HepG2 and K562) using the same strategy as He et al. (27). Tissue-specific predictions by TURF had the best performance in GM12878 versus other top performing computational tools (Figure 3A and Supplemental Table S4). TURF also has the top AUROC in HepG2 with the second largest

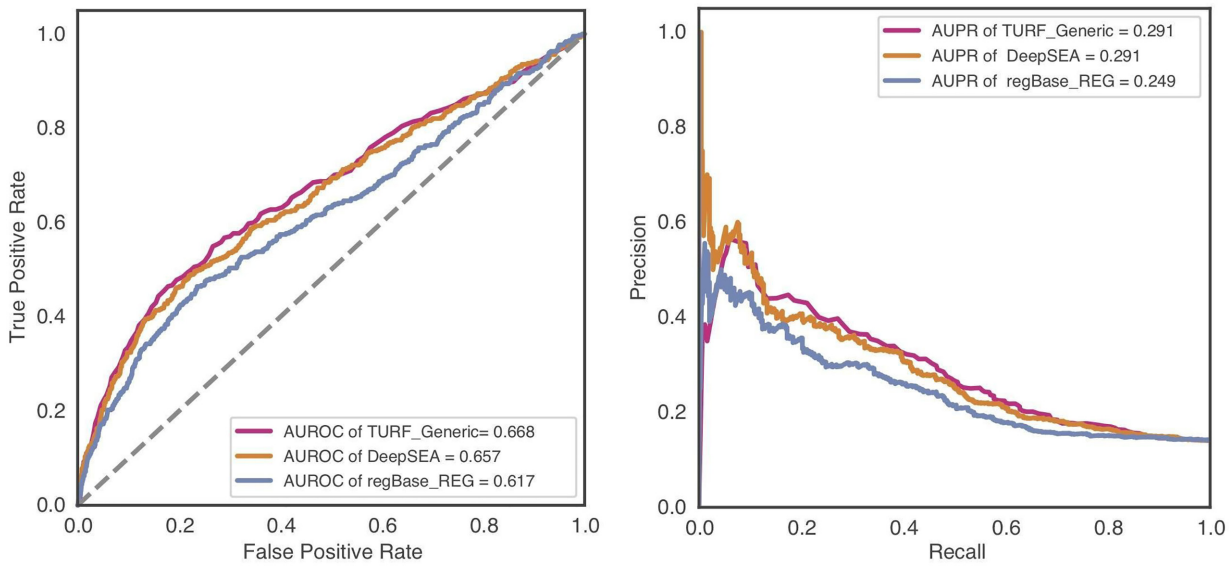

Figure 2. TURF generic scores performance on test data from massively parallel reporter assay (MPRA) in GM12878. Performance was evaluated by Area Under ROC Curve (AUROC) and Area Under Precision-Recall Curve (AUPR). 435 positive variants versus 2670 control variants were called in this MPRA validated dataset.

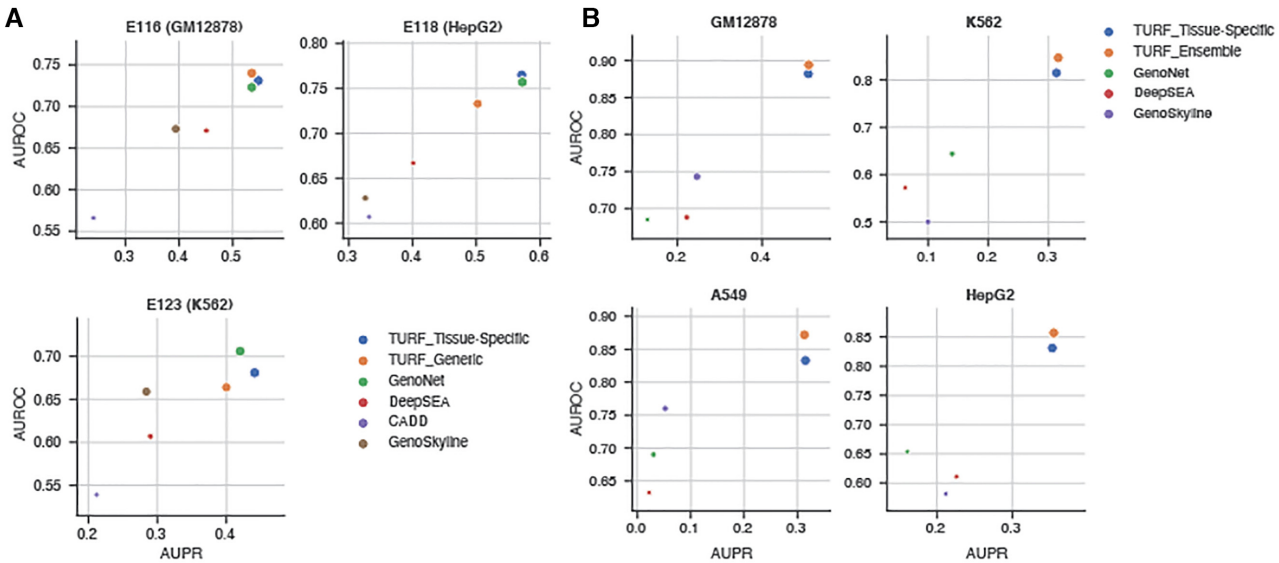

Figure 3. Tissue-specific predictions performance comparisons. Each plot shows the AUPR (area under the precision recall curve) on x axis and the AUROC (area under the receiver operating characteristics curve) on y axis. The size of each point represents the pearson correlation. (**A**) Performance on MPRA data in three cell lines (GM12878: 693 positive variants, 2772 control variants; HepG2: 524 positive variants, 1439 control variants; K562: 339 positive variants, 1361 control variants). (**B**) Performance on allele specific transcription factor binding SNVs (see the number of variants in Supplemental Table S1).

AUPR (0.571 compared to 0.572 from GenoNet) and the largest AUPR in K562. Noticeably, the tissue-specific features in the second random forest model have significantly improved the performance of the TURF generic scores. Among all tissue-specific features, open chromatin regions from DNase-seq in the corresponding cell lines are the most important predictors in all three MPRA datasets. Tissue-specific DNase footprints and active histone marks, including H3K4me2, H3K4me3 and H3K27ac, also play essential roles in variant prediction (Supplemental Figure S2).

TURF tissue-specific predictions on allele specific TF binding (ASB) SNVs

Despite the power of using MPRA datasets as training sets, they are currently limited in terms of the number of tested variants and the variety of tissues. To obtain a more robust tissue-specific model, we called allele-specific TF binding (ASB) SNVs from 6 cell lines. When trained on ASB SNVS, our tissue-specific models greatly outperformed other methods (Figure 3B and Supplemental Table S4). Among the

tissue-specific features, DNase-seq peaks and several active histone marks, such as H3K4me2 and H3K27ac, were important predictors of tissue-specific functional variants, similar to what was observed in the MPRA datasets (Supplemental Figure S3). However, DNase footprints show more variation in feature importance ranking within the 6 cell lines. This indicates the diversity of DNase-seq data quality in different cell lines, and suggests that utilization of a more robust model to compensate for this variation is needed when extending to other tissues not used in the training data.

We then trained an ensemble tissue-specific model using the average predictions from 6 models with feature weights individually learnt from six ASB cell lines. The histone mark features were restricted to five histone marks that ranked high in feature importance, and had available datasets covering most tissues (i.e. H3K27ac, H3K36me3, H3K4me1, H3K4me3 and H3K27me3). The ensemble model outperformed the individual tissue-specific models when predicting ASB SNVs (Figure 3B and Supplemental Table S4). Moreover, this ensemble model trained on ASB SNVs performed better than most of the other tools when tested on the previous independent MPRA datasets in all three cell lines. The exception was GenoNet, which used labels from the MPRA datasets in their training step (Supplemental Table S4). Predictions were computed from this ensemble tissue-specific model on the ASB SNVs in six cell types and most exhibited the highest prediction scores in their corresponding functional cell line (Figure 4). However, HepG2 ASB SNVs had the least enrichment of high HepG2-specific scores, perhaps due to DNase-seq noise in the dataset as only 25% were in DNase peaks. Some H1hESC ASB SNVs had high scores in K562 and MCF7, implying that a many stem cell regulatory variants are involved in regulation of pathways in differentiated cell lines.

Extension of TURF tissue-specific scores to organ-specific scores

To expand the scale of prediction for TURF, we leveraged tissue-specific functional genomic annotations of tissues belonging to the same organ and generated combined organ-specific scores across 51 organs. We were able to recover the organ-specific function of some well-studied regulatory variants in specific genomic loci with TURF scores. For example, TURF's organ-specific scoring was able to pick out the regulatory SNP rs12740374 that affects liver-specific *SORT1* gene expression levels in the 1p13 cholesterol locus (41) (Figure 5). The liver-specific function of rs12740374 was also validated in HepG2 MPRA assays (42). The position of rs12740374 overlaps several active histone mark peaks from ChIP-seq (H3K27ac, H3K4me3 and H3K4me1) and DNase peaks in liver tissues. These multiple lines of genomics evidence prioritized rs12740374 as the top SNP for liver-specific scores within a list of candidates from previous association studies. In addition to liver, rs12740374 has a high significance score in other organs relevant with cholesterol metabolism, such as adipose tissue and gonad. As another example, TURF also detected a regulatory SNP at the *GATA4* locus in the heart (Supplemental Figure S4) that was initially discovered in a genome-wide association scan on 466 bicuspid aortic valve cases (43).

TURF organ-specific scores prioritize genetic variants associated with traits in relevant organs

We examined TURF organ-specific scores on variants identified from genome-wide association studies (GWAS) using the GWAS Catalog portal (1). GWAS variants were found to be enriched in regulatory elements of non-coding regions (44,45). We tested the enrichment of putative regulatory variants prioritized by TURF scores for a variety of traits. For each trait, the top organ with the highest z-score showed the most significant enrichment of organ-specific regulatory variants relative to the background set from all traits within the GWAS catalog, as well as 50 other organs with the same trait (Figure 6 and Supplemental Figure S5).

The top enriched organs from diverse traits were consistent with current trait-relevant organ knowledge. For example, many immune system related diseases, such as autoimmune disease, celiac disease and chronic lymphocytic leukemia, showed a high enrichment for regulatory variants functional in immune-related organs, including immune organ, spleen, and lymph node. Traits of immune cells, such as leukocyte, eosinophil and platelet, were also enriched in immune organs. Cardiac traits, including PR interval, which is a measurement in electrocardiography, and coronary artery disease, were enriched in heart and arterial blood vessel. Enrichment in the colon and immune-related organs was demonstrated for Crohn's disease and ulcerative colitis, both inflammatory bowel diseases. Furthermore, several traits of measurement were enriched for organs involved in relevant metabolic pathways, such as cholesterol measurement in liver, apolipoprotein A1 measurement in small intestine (46), renin-angiotensin system (RAS) use measurement in adrenal gland, and alcohol consumption measurement in exocrine gland (i.e. salivary gland). Of note, the enrichment of variants in some traits could be affected by cofactors, such as gender for body height enrichment within the vagina and ovary. Also, some organs seem to share similarities in gene regulatory networks, partly due to overlapping of tissues, or tissues with similar functions across different organs. This explains a mixture of brain and optic traits enriched in either brain or eye, as the optic nerve gene expression pattern was found to be similar to brain tissue (47).

The most enriched organ for potential regulatory variants provides new directions for understudied diseases or traits. For instance, drugs of calcium channel blockers were found to increase the risk of pancreatic cancer in postmenopausal women (48), while the underlying mechanisms remain unclear. Interestingly, pancreas was the top organ for the calcium channel blocker use measurement trait, which indicates an enrichment of putative regulatory variants functional in pancreas. Thus, additional studies on top variants prioritized by TURF pancreas-specific scores may help further explain the association between pancreatic cancer risk and the use of calcium channel blocker drugs. Similar workflow can be applied to other diseases, such as Alzheimer's disease in immune organs, to determine the causal variants in non-coding regions.

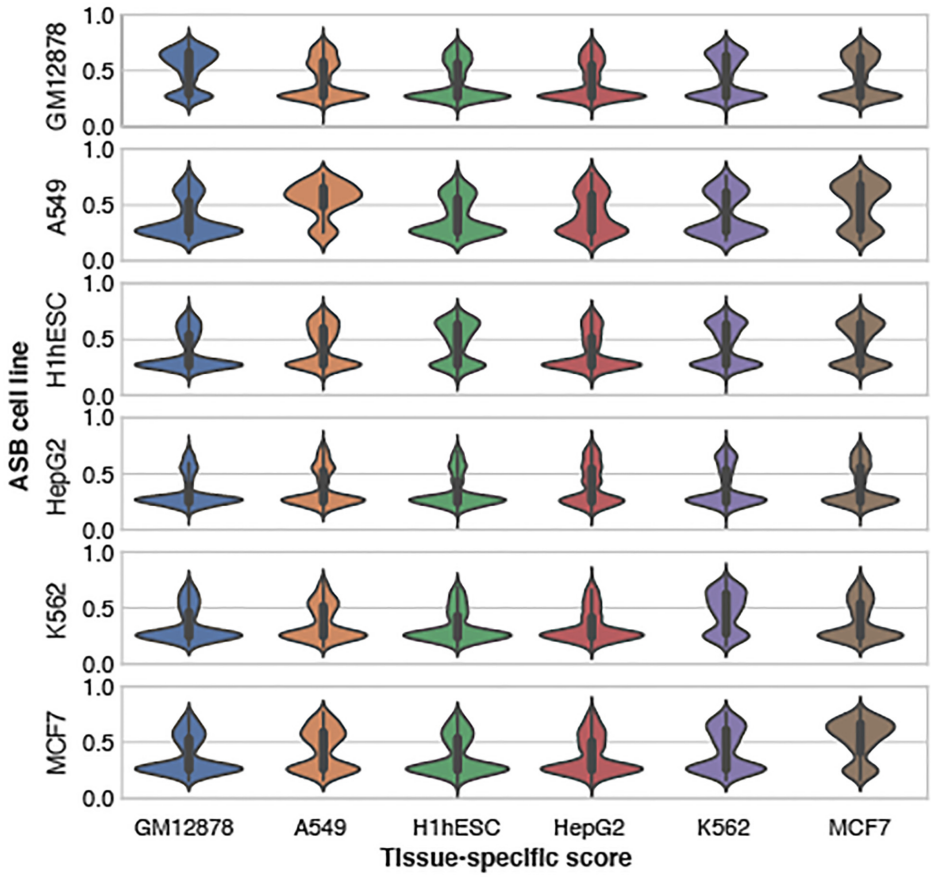

Figure 4. TURF tissue-specific scores on allele-specific transcription factor binding (ASB) SNVs called from six cell lines. The ASB cell line represents the functional tissue for ASB SNVs in each row. The tissue-specific scores are shown in violin plots with a given cell line in each column. ASB SNVs have overall the highest tissue-specific scores in their functional cell line.

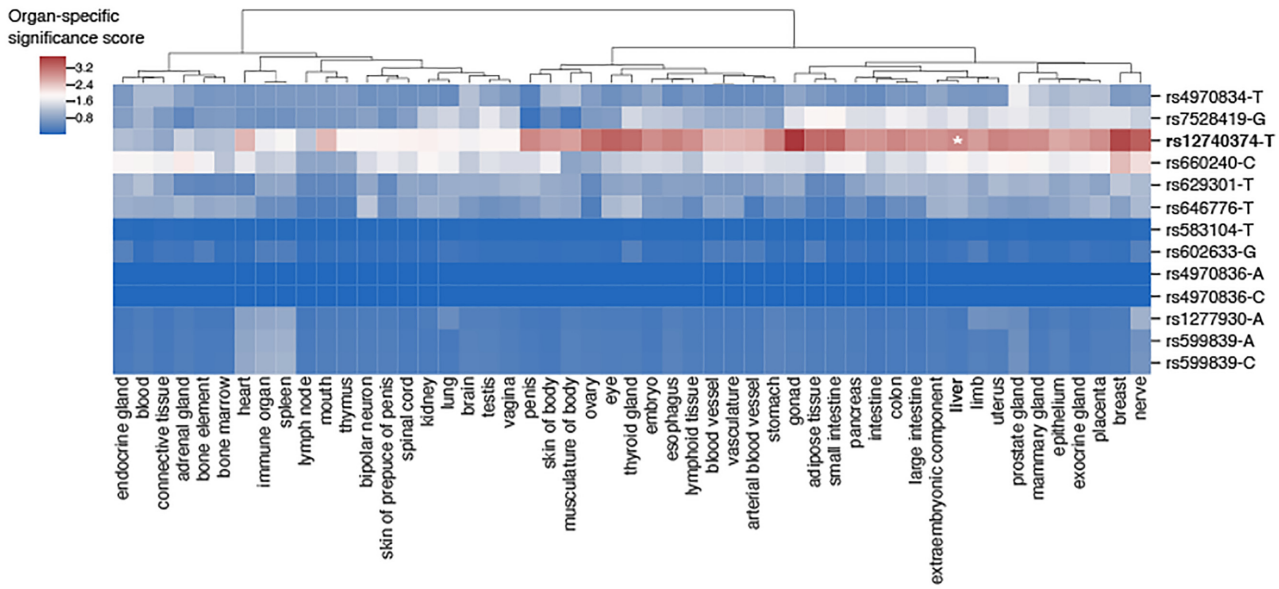

Figure 5. Organ-specific significance scores of variants in the 1p13 cholesterol locus. rs12740374 has the top liver-specific significance score compared to other nearby candidate SNPs from association studies, which was validated to affect *SORT1* gene expression level in liver tissue. The organ-specific significance scores were calculated relative to a background set from GWAS variants (see methods).

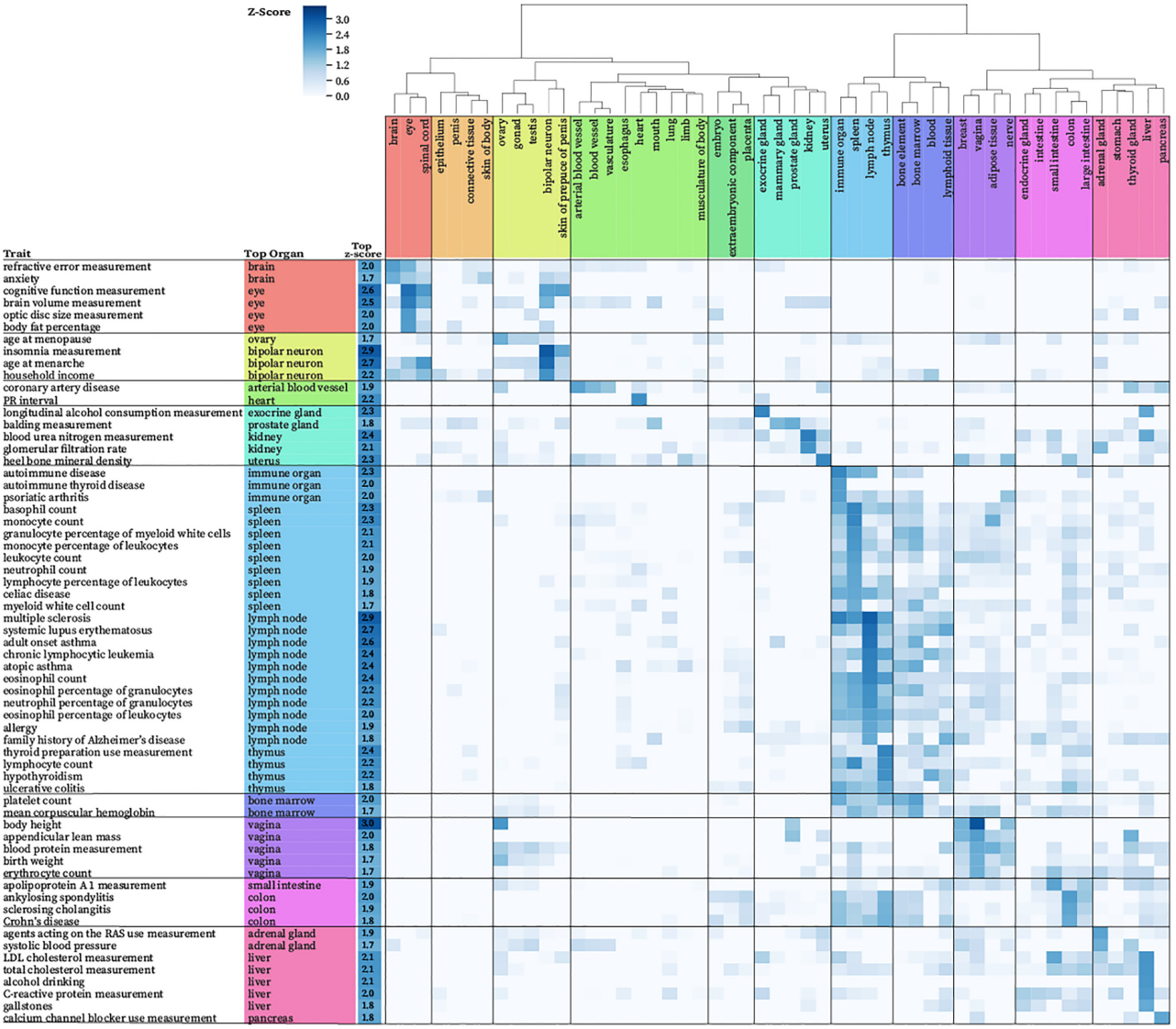

Figure 6. Enrichment of regulatory variants with high organ-specific scores over variants associated with diverse traits. The z-scores of organs (column) for a given trait (row) are shown. The organ with the highest z-score for each trait is shown in additional columns on left. Only organ-trait pairs with z-scores higher than 0 and passing multiple test correction (FDR threshold of 5%) are shown. Traits with top z-scores <1.7 were ignored in this plot, see complete plot Supplemental Figure S5.

DISCUSSION

In this study, we developed TURF, a computational tool that prioritizes variants in non-coding regions. Evidence was incorporated from various functional genomic assays to produce robust predictions that were verified via MPRA assays in both generic and tissue-specific contexts. The workflow was designed to identify regulatory variants from association studies with tissue/organ-specific regulatory function. Moreover, we found GWAS variants were enriched with regulatory variants predicted by TURF organ-specific scores in trait-related organs.

To balance between prediction accuracy and data availability, we trained TURF on ASB SNVs identified from ChIP-seq to determine the weight of features in a tissue-specific context, then extended the scale of annotation to an organ-specific level. The TURF tissue-specific scores leverage information gained from other tissues while retaining the uniqueness of the gene regulatory network in individual tissues. We were able to prioritize putative organ-specific regulatory variants across 51 organs in diverse pathways. A number of computational tools have been developed recently for similar purposes however, most focus on genomic assays and tissues from the Roadmap project (27,29). This makes it difficult to utilize their results for tissues not included in the Roadmap project. As an alternative, we took advantage of over 3000 genomic assays in >200 tissues and cell types available from the ENCODE project, expanding the annotation scope and enhancing the robustness of our predictions. Most relevant organs of various GWAS traits were recovered from the organ-specific scores, including some well-studied traits, such as LDL cholesterol measurement and immune diseases. These results were mirrored in active histone marks using epigenomics data

from the Roadmap project (45). In addition, we observed novel organ-trait pairs, including pancreas in calcium channel blocker use measurement, which can help elucidate underlying disease mechanisms. As more functional genomics datasets are generated, our algorithm is flexible allowing for addition of new tissues by querying histone mark and DNase features within the new tissue and then computing new tissue/organ-specific scores. In addition to the variants from GWAS studies, TURF scores can also be implemented on any personal genome as a way to prioritize potential regulatory variants for individuals (Supplemental Figure S6).

Despite the large scale of annotation utilizing the 51 ENCODE organs, further refinement of the organ terms and the tissues assigned to each organ is possible. Some traits in Figure 5 showed enrichment in non-relevant organs, such as household income in the bipolar neuron. This could be partly due to cofactors within individual GWAS samples, but can also imply an imbalance in the number of genomic datasets across diverse organs as the bipolar neuron (i.e. ear) only contains one ENCODE biosample. Due to the limitation of data availability, we only used seven tissue-specific binary features when building the second random forest model. With more functional genomics data being generated, especially those targeting more histone marks, we can expand our feature set and generate a wider spectrum of prediction scores. The organ-specific scores can then be normalized across different organs to eliminate bias from data availability. The organ-specific scores for a variant will be more comparable over a list of interested organs. In addition, as more single cell data is being generated, we can explore more complex models other than simply combining all annotations at organ-levels, for example to separate cancer cells and normal cells.

We used MPRA data to validate our method as these assays provide more direct evidence of variants affecting gene expression than other association analyses, such as eQTLs, which can be affected by variants that are in strong linkage equilibrium. However, we could only test our model in three MPRA cell lines when comparing performance to other tools. We found one tool used MPRA data labels causing overfitting when tested on ASB SNVs. We built an ensemble model trained on SNVs from six cell lines to avoid the overfitting. With more MPRA data becoming available in the future, we can provide a more thorough comparison of performance and further refine our model by including training variants from more cell types or more types of assays. In addition, it will be possible to integrate features from 3D conformation assays, such as Hi-C and ChIA-PET, in a tissue-specific manner to further improve TURF performance as more datasets in high resolution become available. Moreover, although the ENCODE project already includes datasets from human tissues in addition to cell lines, incorporating more datasets from other large consortia, such as the GTEx project, will further enhance TURF performance on annotating non-coding variants. While initially focusing on single nucleotide variation, TURF has the potential to be leveraged to other classes of variants, such as small INDELs, but might not be appropriate for other more complex structural variation.

Overall, TURF is able to prioritize regulatory variants with either generic or tissue-specific functions. We expect our tool to enhance future studies on functional consequences of regulatory variants associated with diseases from GWAS. The organ-specific scores generated here will be incorporated into the RegulomeDB database. We also include GWAS variants as a part of a Docker pipeline and as a public Amazon Machine Image to allow annotation of the most updated GWAS datasets, making it a useful tool for broad communities.

DATA AVAILABILITY

The RegulomeDB TURF pipeline including a docker instance and public Amazon Machine Image as well as pre-calculated TURF generic and organ-specific scores on GWAS variants and data used in this study are available at: https://github.com/Boyle-Lab/RegulomeDB-TURF.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank all members in the Boyle lab for helpful discussions and constructive feedback. We thank Yunhai Luo and Ben Hitz for help in retrieving annotations from the RegulomeDB web server and Pedro Assis for help in deployment of the Amazon Machine Image.

FUNDING

NIH [U24 HG009293 to A.P.B.]. Funding for open access charge: NIH [U24 HG009293].
Conflict of interest statement. None declared.

REFERENCES

1. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
2. Tam,V., Patel,N., Turcotte,M., Bossé,Y., Paré,G. and Meyre,D. (2019) Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.*, **20**, 467–484.
3. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
4. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
5. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
6. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
7. Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, **2010**, db.prot5384.
8. Barski,A., Cuddapah,S., Cui,K., Roh,T.-Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
9. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

10. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.-K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
11. Kasowski,M., Kyriazopoulou-Panagiotopoulou,S., Grubert,F., Zaugg,J.B., Kundaje,A., Liu,Y., Boyle,A.P., Zhang,Q.C., Zakharia,F., Spacek,D.V. *et al.* (2013) Extensive variation in chromatin states across humans. *Science*, **342**, 750–752.
12. Kasowski,M., Grubert,F., Heffelfinger,C., Hariharan,M., Asabere,A., Waszak,S.M., Habegger,L., Rozowsky,J., Shi,M., Urban,A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.
13. Degner,J.F., Pai,A.A., Pique-Regi,R., Veyrieras,J.-B., Gaffney,D.J., Pickrell,J.K., De Leon,S., Michelini,K., Lewellen,N., Crawford,G.E. *et al.* (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–394.
14. Tewhey,R., Kotliar,D., Park,D.S., Liu,B., Winnicki,S., Reilly,S.K., Andersen,K.G., Mikkelsen,T.S., Lander,E.S., Schaffner,S.F. *et al.* (2018) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, **172**, 1132–1134.
15. Kheradpour,P., Ernst,J., Melnikov,A., Rogov,P., Wang,L., Zhang,X., Alston,J., Mikkelsen,T.S. and Kellis,M. (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.*, **23**, 800–811.
16. Patwardhan,R.P., Hiatt,J.B., Witten,D.M., Kim,M.J., Smith,R.P., May,D., Lee,C., Andrie,J.M., Lee,S.-I., Cooper,G.M. *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.*, **30**, 265–270.
17. Yan,J., Qiu,Y., Ribeiro Dos Santos,A.M., Yin,Y., Li,Y.E., Vinckier,N., Nariai,N., Benaglio,P., Raman,A., Li,X. *et al.* (2021) Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, **591**, 147–151.
18. Boyle,A.P., Hong,E.L., Hariharan,M., Cheng,Y., Schaub,M.A., Kasowski,M., Karczewski,K.J., Park,J., Hitz,B.C., Weng,S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.
19. Li,M.J., Wang,L.Y., Xia,Z., Sham,P.C. and Wang,J. (2013) GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res.*, **41**, W150–W158.
20. Ward,L.D. and Kellis,M. (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.*, **44**, D877–D881.
21. Beer,M.A. (2017) Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.*, **38**, 1251–1258.
22. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
23. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
24. Quang,D. and Xie,X. (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, **44**, e107.
25. Kelley,D.R., Reshef,Y.A., Bileschi,M., Belanger,D., McLean,C.Y. and Snoek,J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, **28**, 739–750.
26. Backenroth,D., He,Z., Kiryluk,K., Boeva,V., Pethukova,L., Khurana,E., Christiano,A., Buxbaum,J.D. and Ionita-Laza,I. (2018) FUN-LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. *Am. J. Hum. Genet.*, **102**, 920–942.
27. He,Z., Liu,L., Wang,K. and Ionita-Laza,I. (2018) A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. *Nat. Commun.*, **9**, 5199.
28. Li,M.J., Li,M., Liu,Z., Yan,B., Pan,Z., Huang,D., Liang,Q., Ying,D., Xu,F., Yao,H. *et al.* (2017) cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. *Genome Biol.*, **18**, 52.
29. Lu,Q., Powles,R.L., Wang,Q., He,B.J. and Zhao,H. (2016) Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLos Genet.*, **12**, e1005947.
30. Yao,Q., Ferragina,P., Reshef,Y., Lettre,G., Bauer,D.E. and Pinello,L. (2021) Motif-Raptor: a cell type-specific and transcription factor centric approach for post-GWAS prioritization of causal regulators. *Bioinformatics*, **37**, 2103–2111.
31. Dong,S. and Boyle,A.P. (2019) Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum. Mutat.*, **40**, 1292–1298.
32. Chen,J., Rozowsky,J., Galeev,T.R., Harmanci,A., Kitchen,R., Bedford,J., Abyzov,A., Kong,Y., Regan,L. and Gerstein,M. (2016) A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.*, **7**, 11101.
33. 1000 Genomes Project Consortium, Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
34. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
35. Rozowsky,J., Abyzov,A., Wang,J., Alves,P., Raha,D., Harmanci,A., Leng,J., Bjornson,R., Kong,Y., Kitabayashi,N. *et al.* (2011) AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.*, **7**, 522.
36. Abyzov,A., Urban,A.E., Snyder,M. and Gerstein,M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974–984.
37. Zhang,S., He,Y., Liu,H., Zhai,H., Huang,D., Yi,X., Dong,X., Wang,Z., Zhao,K., Zhou,Y. *et al.* (2019) regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants. *Nucleic Acids Res.*, **47**, e134.
38. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
39. Slatkin,M. (2008) Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, **9**, 477–485.
40. Sherry,S.T. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
41. Musunuru,K., Strong,A., Frank-Kamenetsky,M., Lee,N.E., Ahfeldt,T., Sachs,K.V., Li,X., Li,H., Kuperwasser,N., Ruda,V.M. *et al.* (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**, 714–719.
42. Shigaki,D., Adato,O., Adhikari,A.N., Dong,S., Hawkins-Hooker,A., Inoue,F., Juven-Gershon,T., Kenlay,H., Martin,B., Patra,A. *et al.* (2019) Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.*, **40**, 1280–1291.
43. Yang,B., Zhou,W., Jiao,J., Nielsen,J.B., Mathis,M.R., Heydarpour,M., Lettre,G., Folkersen,L., Prakash,S., Schurmann,C. *et al.* (2017) Protein-altering and regulatory genetic variants near GATA4 implicated in bicuspid aortic valve. *Nat. Commun.*, **8**, 15481.
44. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
45. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
46. Glickman,R.M. and Green,P.H. (1977) The intestine as a source of apolipoprotein A1. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 2569–2573.
47. Diehn,J.J., Diehn,M., Marmor,M.F. and Brown,P.O. (2005) Differential gene expression in anatomical compartments of the human eye. *Genome Biol.*, **6**, R74.
48. Wang,Z., White,D.L., Hoogeveen,R., Chen,L., Whitsel,E.A., Richardson,P.A., Virani,S.S., Garcia,J.M., El-Serag,H.B. and Jiao,L. (2018) Anti-hypertensive medication use, soluble receptor for glycation end products and risk of pancreatic cancer in the women's health initiative study. *J. Clin. Med. Res.*, **7**, 197.