bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Organ-specific prioritization and annotation of non-coding regulatory variants in the human genome

Nanxiang Zhao[1,*], Shengcheng Dong[2,*], Alan P Boyle[1,3,†]

1. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA
2. Department of Genetics, Stanford University, Stanford, CA, USA
3. Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

* These authors contributed equally

† Co-Correspondence: apboyle@umich.edu

Abstract

Identifying non-coding regulatory variants in the human genome remains a challenging task in genomics. Recently we advanced our leading regulatory variant database, RegulomeDB, to its second version. Building upon this comprehensive database, we developed a novel machine-learning architecture with stacked generalization, TLand, which utilizes RegulomeDB-derived features to predict regulatory variants at cell or organ-specific levels. In our holdout benchmarking, TLand consistently outperformed state-of-the-art models, demonstrating its ability to generalize to new cell lines or organs. We trained three types of organ-specific TLand models to overcome the common model bias toward high data availability cell lines or organs. These models accurately prioritize relevant organs for 2 million GWAS SNPs associated with GWAS traits. Moreover, our analysis of top-scoring variants in specific organ models showed a high enrichment of relevant GWAS traits. We expect that TLand and RegulomeDB will further advance our ability to understand human regulatory variants genome-wide.

Introduction

Understanding the biological impact of variants located in non-coding regions of the human genome is a significant challenge. Nearly 90% of disease-risk-associated single nucleotide polymorphisms (SNPs) identified from genome-wide association studies (GWAS) are within non-coding regions. Similarly, 75% of patients affected by Mendelian disease have mutations outside of protein-coding regions (1). The abundance of disease-associated non-coding variants makes these regions highly promising for identifying causal SNPs and improving our understanding of their functional consequences.

Prioritizing non-coding variants requires integrating multiple layers of functional information, including regulatory annotations identified from high-throughput sequencing datasets (e.g. DNase-seq (2), ChIP-seq (3), and ATAC-seq (4)). Such annotations provide additional information in pinpointing causal variants, which are often not the lead variants identified in GWAS studies. Despite the benefit of incorporating functional genomics assay-based evidence when examining non-coding variants, the lack of available variant annotation tools

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

limits the use of such data. Most resources developed for clinical purposes have focused on coding regions as an application of exome sequencing-based data (5, 6), which captures less than 5% of human variation (7–9).

Previously we built RegulomeDB, a comprehensive database for prioritizing and annotating variants in non-coding regions, which was highly sought after in the research community (10). RegulomeDB intersects query variants with regulatory regions predicted by functional genomics assays and, by utilizing ranking heuristics, informs users about putative functional consequences to prioritize variants. Recently, RegulomeDB has been upgraded to version 2 (11), improving its annotation power by incorporating thousands of new functional genomics datasets from the ENCODE project (12), Roadmap Epigenomics Consortium (13), and the Genomics of Gene Regulation Consortium. A suite of models, namely SURF and TURF, was developed and integrated in this version to provide accurate probabilistic scores for general and cell-type specific regulatory activities (14, 15).

However, a drawback to this model suite is it was trained using hg19-referenced datasets from ENCODE for only six common cell lines. The resulting models are biased, lacking sufficient statistical power to generalize their predictions when applied to less-studied cell lines. As a result, RegulomeDB scores are less-informative for cell lines, tissues, and organs that lack the abundance of data available for commonly-studied cell lines. This can make it challenging for users to identify variants of interest and formulate hypotheses regarding their regulatory functions. Nonetheless, we anticipate RegulomeDB to further improve for years to come as new datasets are released by the ENCODE (12) and IGVF (16) consortia, spanning additional cell lines, tissues, and assay types: e.g., Hi-C and (17) and Enformer predictions (18). However, the current SURF and TURF models were not designed to incorporate data continuously, and are potentially prone to overfitting as the feature space expands. Given the breadth and volume of functional genomics datasets being released each day, these are serious limitations.

Here we present TLand, a flexible architecture based on stacked generalization (19) to learn RegulomeDB-derived features to predict regulatory variants at a cell-specific level or organ-specific level. TLand's stacked generalization approach groups feature into biologically meaningful subspaces, training individual estimators before assembly to reduce overfitting and enable further integration of features. Cell-specific TLand consistently outperformed state-of-the-art models in benchmarking with a hold-out cell line. By developing a suite of models rather than a single, monolithic model, organ-specific TLand addresses the data availability bias, further improving upon the cell-specific TLand in identifying regulatory variants relevant to the target organ as predicted by GWAS. Furthermore, analysis of top-scored variants in specific organ models showed high enrichment of correlated GWAS traits. Given its superior performance relative to its predecessors and competing methods, we expect TLand to address the ongoing challenge of reliably prioritizing variants, even in less-studied cell lines and organs, thus advancing our ability to identify regulatory variants genome-wide.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Allele-specific binding (ASB) variants

We define a variant as likely-regulatory if it shows evidence for allele-specific binding (ASB), the molecular signature of which is significantly different counts when ChIP-seq reads are phased between the two alleles in a heterozygous individual. We trained our models on ASB variants for all TFs with ChIP-seq data in ENCODE. We included a total of 7,530 ASB variants in 6 cell lines (GM12878, HepG2, A549, K562, MCF7, and H1hESC) called by *AlleleDB* (20). We utilized negative training sets previously produced in our lab (15), encompassing non-ASB variants and randomly-selected background variants. In total, we included 14,773 unique variants in our training set. The complementary ASB data of IMR-90 and H9 used for evaluation were downloaded from Adastra (21) at https://adastra.autosome.org/bill-cipher/search/advanced?fdr=0.05&es=0&cl=IMR90%20(lung%20fibroblasts) and https://adastra.autosome.org/bill-cipher/search/advanced?fdr=0.05&es=0&cl=H9.

Model architecture

TLand is a one-layer stacked architecture that consists of two parts: three base classifiers and one meta-classifier (Fig 1b and Supplementary Fig 1). TLand takes input as 19 generic features, 40 deep learning prediction-derived features, and 5 cell-specific features or 13 organ-specific features (Supplementary Table 1), which are directly derived from RegulomeDB queries. Features were bagged into 3 subspaces: experimental set (generic features and cell/organ-specific features), deep learning set (deep learning features and cell/organ-specific features), and cell/organ-specific set (only cell/organ-specific features). We selected lightGBM (22), random forest (23), and neural network (24) as base classifiers due to their distinct decision boundaries. Our base models were fine-tuned with Optuna (25). We used 300 estimators in random forest models, 250 boost rounds, and a 0.049 learning rate in lightGBM models, 3 layers with 128 neurons per layer, batch size of 128, adaptive learning rate, and with the max iteration of 30 in neuron network models. Probabilities were used as the output of base classifiers and to train our meta-classifier. We calculated interaction terms of probabilities up to the degree of 2 before feeding into our meta-classifier. We customized a ridge classifier to output probabilities with hyperparameter alpha as 1.9 as our final meta-classifier. The ridge classifier was trained with 4-fold group cross-validation. We grouped the training data based on the genomic positions (i.e. variants at the same genomic positions regardless of whether cell lines were in the same group). All models were specified with balanced class weights. LightGBM was implemented with the Python package (22). Random forest, neural network, and relevant pipelines were implemented with scikit-learn (26). The stacked generalization algorithm was implemented with mlxtend (27).

Model training and evaluation

TLand was trained, validated, and tested on the generated ASB datasets. After concatenating across six cell lines, 88,638 (i.e. 14,773 x 6) variants were used for training and validation. The concatenation allowed us to have more data per prediction task. For the task of predicting an unseen cell line, we held out data for one cell line as a test case, then used the rest as training data for the TLand model. The test and train split ratio for negative

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

data is ⅙. Similarly, for organ-specific models, we held out one organ as a test dataset and used organ-specific labels for splitting the data. After evaluation, the final models were trained with all 88,638 variants.

DNase signal quantile normalization

All 1372 DNase bigwig default files processed on GRCh38 assembly (up to May 2022) were obtained from ENCODE (Supplementary Table 5). We designed an efficient pipeline to quantile normalize all signals to achieve the balance between accuracy, runtime, and storage (Supplementary Fig 2). BigWig files were first converted to BedGraph format using bigWigToBedGraph (28). Average signals were extracted for 10bp, non-overlapping windows extracted from each BedGraph file with bedmap (29) and stored in bed files. These bed files were concatenated and converted to parquet format for processing with qnorm (30), a Python package that applies a multithreaded incremental normalization strategy to efficiently normalize extremely large datasets (30). We quantile normalized 10-bp average signals for all 1372 DNase datasets with a batch size of 343 files per normalization iteration. Details of storage, runtime, and memory are shown in Supplementary Fig 2.

Benchmarking other models and organ definition

We benchmarked TLand with state-of-the-art models, including GenoNet, DeepSEA, Sei, and TURF. Pre-calculated GenoNet scores were downloaded from https://zenodo.org/record/3336209/files/. DeepSEA and Sei models were downloaded from the original paper (31, 32). The previous best model, TURF, was re-trained to predict regulatory variants on GRCh38 by querying and inputting the GRCh38 features (15). Final predictions were made by averaging across cell lines instead of the test unseen cell line models. We averaged relevant cell line model predictions to estimate the organ level predictions, using the human organ definitions from ENCODE (https://www.encodeproject.org/summary/?type=Experiment&control_type!=*&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&status=released.). For example, we averaged TURF model predictions of A549 and IMR-90 to predict lung regulatory variants.

GWAS variants and LD data

We first downloaded all known SNV positions significantly associated with traits described in the GWAS Catalog (https://www.ebi.ac.uk/gwas/api/search/downloads/alternative). We then completed LD expansion for each SNV, incorporating all SNVs from the 1000 genome project in strong LD (R^2 threshold of 0.6). R^2 values were downloaded from (gs://genomics-public-data/linkage-disequilibrium). In total, we calculated their TLand model scores for 1,974,549 SNVs.

Target organs annotations for GWAS traits

Gold-standard target-organ associations for 44 GWAS traits were annotated with Open Targets (https://platform.opentargets.org/), EMBL-EBI Ontology Lookup Service (https://www.ebi.ac.uk/ols/index), and ChatGPT (33) for selected GWAS traits (Supplementary Table 3). These annotations were used to evaluate model performance for each GWAS trait.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Prioritization of relevant organs for GWAS traits

We predicted approximately 2 million GWAS variants and strongly-linked SNPs across 51 organs. For each manually annotated GWAS trait, we selected all associated SNPs. For each organ, we calculated the p-value from a one-tailed Student's t-test (34) comparing the sampled organ-specific SNP scores versus the population distribution, defined as the 2 million SNP score distribution for the organ. We ranked organs by the inverse of their p-values, using a significance threshold of 0.05 to select only organs with statistical support for a functional association. We evaluated performance by comparing the ranked (i.e. prioritized) organs with the manually annotated target organs for each trait. Accuracy was defined as the fraction of correct overlaps between model-based and gold-standard organ annotations among the top 4 organs for each model. We combined any two models' ranking lists by taking the union of their list then ranked by p-value and selected the top 5 organs for each combination.

GWAS trait enrichment

Top-scoring GWAS variants (variants with organ-specific scores >0.5) for each organ were selected. We then traced back associations between these variants and the traits with which they are associated by counting the number of appearances of each trait among each variant, weighted by the corresponding organ-specific score. Traits for which total counts (i.e. trait count of all associated GWAS and strongly-linked SNPs) were low were filtered out of the results for each organ. Organ-specific GWAS trait enrichment scores were calculated as the weighted appearance count divided by the total count.

Results

TLand incorporates comprehensive datasets to predict regulatory variants

We developed a new model architecture, TLand, to predict regulatory variants from a comprehensive set of features derived from RegulomeDB (Fig 1b and Supplementary Fig 1). TLand takes input as genomic positions or dbSNP IDs, queries RegulomeDB for features, then outputs the probability that each variant is a cell-type specific (Supplementary Fig 1) or organ-specific (Fig 1b) regulatory variant. Importantly, by combining stacked generalization, features grouping, and interaction terms in meta-classifier training, the TLand model is capable of incorporating growing training datasets and novel data types into its feature space while combating overfitting, whereas its predecessors cannot.

The continuously growing corpus of genomic data captures an increasingly complete view of human regulatory variation. RegulomeDB recently upgraded to version 2, expanding to >650 million and >1.5 billion genomic intervals in hg19 and GRCh38, respectively (11). The large discrepancy in the data availability between the two assemblies, for example in ChIP-seq and open chromatin data (RegulomeDB TF ChIP-seq availability in Fig 1a, open chromatin in Supplementary Fig 3, histone ChIP-seq in Supplementary Fig 4), results from better representation of complex variation and correction of sequencing artifacts in the GRCh38 assembly (12). Features used to predict regulatory variants included experimental and computational features derived in GRCh38, the majority of which were derived from

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

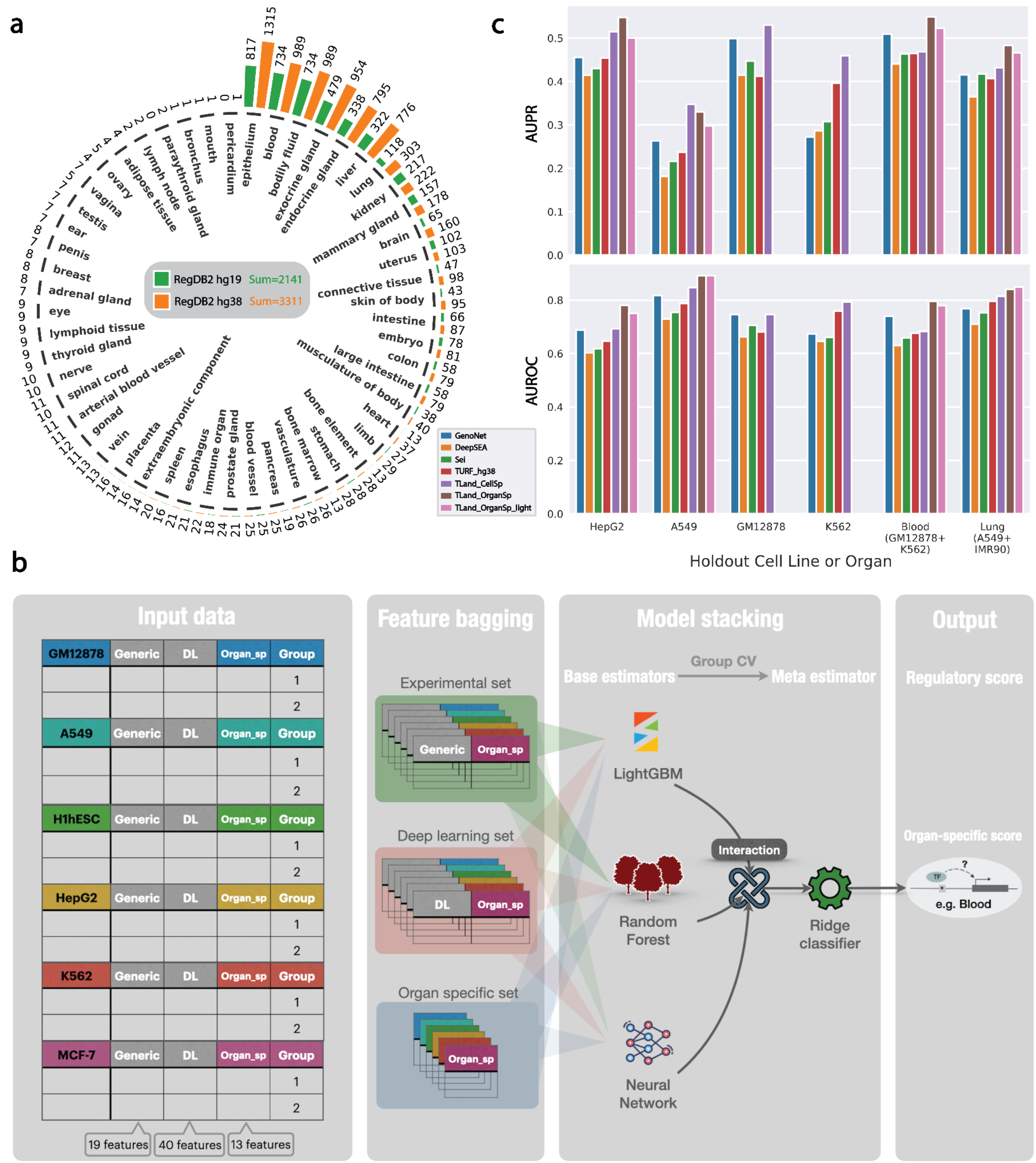

Figure 1. TLand improves regulatory variant predictions. (a) TF ChIP-seq data availability across organs on RegulomeDB v2. Green bar plots represent counts on GRCh38. Orange bar plots represent counts on hg19. The total number of counts for each assembly is in the middle of the gray box. Notice that the summation is not simply adding all numbers together due to some cell lines having multiple corresponding organs. (b) Organ-specific TLand architecture. Organ-specific TLand was trained to predict human regulatory variants in an organ-specific manner by using RegulomeDB-derived features. (c) Benchmarking TLand performance by AUROC and AUPR. X-axis is holdout cell lines or organs. Y-axis is AUPR on the top panel and AUROC on the bottom panel.

RegulomeDB v2 (Supplementary Table 1). 1,372 DNase-seq BigWig files were quantile-normalized and derived as 5 quantile features for each genomic locus. The DeepSEA disease score feature (31) was replaced by its successor Sei (32). The Sei model simultaneously predicted 21,907 binary assay labels which were dimension-reduced to 40

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

features representing sequence classes. Assembling deep learning model predictions and training with comprehensive features reduced the variances of our final models and generalizes well to new cell lines and less-studied organs (35, 36).

Training data across six common cell lines were concatenated to train an agnostic model while the specificities of input features decided the cell or organ specificities for model predictions. Features were bagged into three biologically meaningful subspaces before each set was fed into an individual base classifier. Subspaces included a generic experimental feature set, a computational feature set, and a cell/organ-specific experimental feature set for regulatory variants.

We adapted the stacked generalization algorithm to stack the output of individual classifiers and use a meta-classifier to compute the final prediction (19). Stacking allows for the utilization of the strength of each individual classifier by using their output as input of a final meta-classifier. Cross-validation is required to prevent information leaks during the training of the meta-classifier. We made several modifications to the algorithm. We used group cross-validation to make base classifiers to learn the regulatory function as the conditional probability between generic features and cell/organ-specific features. Interaction terms were calculated before feeding them into the meta-classifier. We used probability rather than binary prediction in the first layer of base classifiers to train the meta-classifier to obtain higher accuracy (37).

TLand improves regulatory variant predictions

Cell-specific TLand models substantially outperformed state-of-the-art models for predicting unseen cell line regulatory variants (Fig 1c and Supplementary Fig 5). On average, across holdout cell lines, cell-specific TLand models outperformed the previous best model (TURF) with the area under the precision-recall curve (AUPR) and the area under the receiver operating characteristic (AUROC) increasing from 0.389 to 0.471, and 0.729 to 0.776, respectively. Although GenoNet outperformed TURF in three cell lines, applying the unseen cell line model of GenoNet to assess the upper limit of the prediction task attributed the performance-gain to information leakage. Cell-specific TLand models still had superior performance compared to GenoNet, improving average AUPR by 0.09 and AUROC by 0.04. The high performance of cell-specific TLand can be explained in two ways. First, we derived comprehensive feature sets for predicting regulatory variants from RegulomeDB, categorizing them into experimental feature sets and a complementary non-correlated computational feature set (i.e. deep learning feature set) (Supplementary Fig 6). The inclusion of experimental features enabled a ~30% performance gain compared to DeepSEA and Sei models, which use only deep learning features, with average TLand average AUPR=0.471 compared to AUPR=0.336 for DeepSea and AUPR=0.364 for Sei. Second, TLand's architecture models different biological domains separately and then calculates their conditional probabilities. The meta-classifier finds the best combination of these probabilities and subtracts redundant information through the inclusion of negative coefficients (Supplementary Fig 7) to cancel out noise, thus improving predictions, yielding improvements of 0.10 in average AUPR and 0.04 in average AUROC when compared to TURF (Supplementary Fig 8).

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Organ-specific TLand models address data availability bias

TLand's flexible architecture allows for the substitution of cell-specific features with organ-specific features, which can be retrained with organ-specific labels to develop organ-specific TLand models. Organ-specific TLand models have the ability to predict regulatory variants specific to organs, even when those organs include only less-studied cell lines or tissues. These models performed better than their cell-specific counterparts in two out of four cell-type specific tasks, specifically in HepG2 (AUPR 0.547 vs. 0.514, AUROC 0.781 vs. 0.692) and MCF-7 (AUPR 0.512 vs. 0.512, AUROC 0.879 vs. 0.840). However, because GM12878 and K562 cell lines have conflicting labels for their same organ blood, the blood organ-specific model was only evaluated in the organ-specific (i.e. heart) task, not in either of the cell-line-specific tasks. The superior performance of organ-specific models, even in hold-out cell line tasks, can be attributed to the fact that cell-type specific regulatory variants in HepG2 and MCF-7 are well-represented by corresponding organ-specific regulatory variants within the RegulomeDB database (see Supplementary Fig 9). In the A549 cell line, which is from the least-represented organ lung, the organ-specific TLand model underperformed the cell-specific TLand model holdout on A549 (AUPR 0.329 vs. 0.347) as the organ model predicted regulatory variants in the lung other than in A549 cell line. To better evaluate model performance in the lung, the ASB dataset from the second-most representative cell line in the lung, IMR-90, was added to complement the holdout dataset. The inclusion of this dataset led to TLand outperforming the cell-specific model when holding out on the lung with IMR-90 datasets (AUPR 0.483 vs. 0.432, AUROC 0.841 vs. 0.814), indicating the organ-specific TLand model can predict a comprehensive set of organ-specific regulatory variants.

However, adding the second most representative cell line in the embryo, H9, did not improve the TLand organ-specific model when compared to the cell-specific model when holding out on the embryo organ (AUPR 0.536 vs. 0.597). Deep learning features, such as DeepSEA disease impact score (31) and Sei sequence classes (32), were more representative of the cell lines or organs with more data availability. This is because the scores were dimensionally reduced from 919 and 21,907 prediction tasks of experimental assays for DeepSEA and Sei, respectively (including TF and histone ChIP-seq and open chromatin assays). Cell lines and organs with more available data dominated the dimension-reduced results. Thus, we removed deep learning features and developed the organ-specific TLand light model (Supplementary Fig 10) to predict regulatory variants in organs with low data availability, such as the embryo. We defined organs with >=100 available ChIP-seq assays as "high data availability" and those with <100 assays as "low data availability". We observed that the TLand (full) model consistently outperformed the TLand light model in organs with high data availability (Fig 1a and c; Supplementary Fig 5), while the light model surpassed the full model in organs with low data availability. For example, the organ-specific TLand light model was the best model when holding out the embryo organ (AUPR 0.639, AUROC 0.774). Those findings indicate that TLand light models are suitable for predicting regulatory variants for organs with low data availability whileTLand (full) models are more suitable for organs with high data availability. We then proceeded with our analysis by training the TLand and TLand light models on all available data. We trained an additional TLand model, TLand lightest, where the organ-specific ChIP-seq features were removed to further reduce bias towards over-represented organs (benchmarking results shown in Supplementary Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TLand prioritizes relevant organs for GWAS traits

To systematically evaluate organ-specific TLand models, we predicted around 2 million GWAS SNPs including SNPs within the same LD blocks across 51 organs defined by ENCODE (12). We found that TLand predictions were more highly correlated with TLand light predictions than TLand lightest (e.g. in the heart organ in Supplementary Fig 11). In addition, we were able to use TLand scores to cluster organs from the same or closely related systems in the human body. For example, TLand (Fig 2a) and TLand light (Supplementary Fig 12) clustered the brain and spinal cord from the central nervous system and the eye and ear from the sensory system together. TLand lightest model, however, was able to cluster the blood vessel, the arterial blood vessel, and the vascular together as part of the circulatory system (Supplementary Fig 13), which was missed by the other two models. This confirmed our earlier observation that TLand light and lightest models are more accurate and appropriate than the full model for organs with low data availability. We manually curated a list of GWAS traits with relevant organs. TLand, TLand light, and TLand lightest models prioritized the relevant organs for 44 GWAS traits (Supplementary Table 3) with an average accuracy of 0.311, 0.340, and 0.466, respectively (Fig 2b. See the definition of accuracy in Methods). By integrating prioritized organs from two distinct models, TLand and TLand lightest, the average accuracy was increased to 0.482. The overall low accuracy was partially due to the uncertainty as to whether low scores reflect low data availability or are true negatives, lacking regulatory function in the target organ.

To better evaluate TLand models in organs with low data availability, we focused on the top-scoring variants in those organs. We found that top-scored variants from organ-specific models were enriched for associated GWAS traits relevant to their target organs (Methods, Fig 2c and results in Supplementary Table 4). For example, two of the three most-enriched GWAS traits identified by the TLand lightest model identified were relevant to the heart organ: atrial fibrillation and resting heart (Fig 2c). Originally, TLand lung models were unable to prioritize the lung organ given the corresponding GWAS SNPs of relevant traits (Supplementary Table 3). However, we found that traits such as physical activity measurement and peak expiratory flow were enriched by top-scored variants in the TLand model. Traits that were associated with multiple organs were able to be properly identified by multiple-organ models. For example, both the bone element and the immune TLand lightest models pinpointed ankylosing spondylitis, a type of inflammatory arthritis affecting the spine and large joints (38), as one of the most enriched GWAS traits for their top-scored variants (Fig 2c). Thus, organ-specific TLand models can accurately prioritize regulatory variants within relevant organs.

Discussion

The identification and characterization of non-coding regulatory variants in the human genome remain a major challenge in the field of genomics. We have developed a novel model architecture, TLand, to accompany the recent advancement of RegulomeDB. TLand utilizes RegulomeDB-derived features to infer regulatory variants at cell-specific and organ-specific levels. By applying stacked generalization to incorporate comprehensive datasets on the GRCh38 reference genome, including experimental and computational features, TLand models consistently outperformed state-of-the-art models in holdout benchmarking. This demonstrates the TLand models' ability to generalize to new cell lines or

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

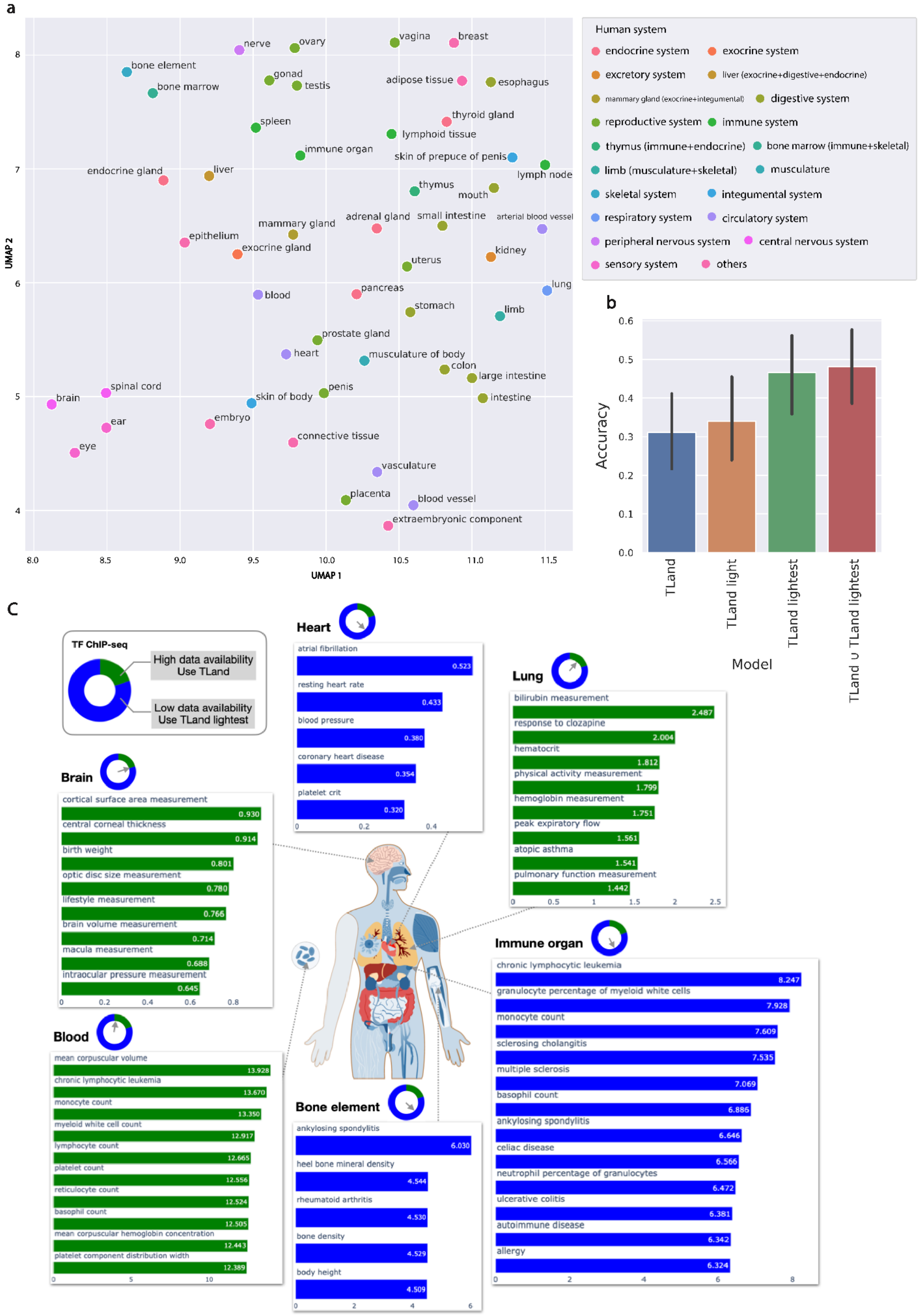

Figure 2. TLand prioritizes relevant organs for GWAS traits. (a) UMAP projections of TLand predictions on GWAS and LD SNPs. They are colored according to human systems. (b) Performance

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of TLand models prioritizing organs. X-axis are TLand models. Y-axis is accuracy (definition see Methods). (c) Prioritization of GWAS traits given top-scored variants by TLand and TLand lightest models. TF ChIP-seq data availability plot is re-plotted in the top left where green represents high data availability and TLand was the model used, and blue represents low data availability and TLand lightest was the model used. Each circle next to the organs indicates the data availability of each organ. The colors of enrichment bars correspond to the model used.

organs, a major advancement over the SURF and TURF models they extend. Much of this improvement can be attributed to better handling of differential data availability issues. In particular, TLand light and lightest models are applicable to organs with limited data, e.g. the embryo, whereas current leading models, such as DeepSEA disease impact scores, perform poorly in such organs.

The TLand models presented here represent a promising foundation for further improvements. These are largely facilitated by TLand's ability to ingest new datasets as they are released. We aim to continuously incorporate new datasets into RegulomeDB as they are released, thus extending its performance and applicability as the number and diversity of training datasets grow across different cells, tissues, and organs. Many of these, including gkm-SVM model predictions (39) and Hi-C contact maps (17) as new features for TLand models. TLand's flexible, modular design allows the grouping of features into biologically-meaningful sets which can be evaluated against others before making the final decision about whether to include novel features in the model.

One limitation of this study is the limited number of allele-specific binding (ASB) sites in our training dataset. Because of data availability limitations, ASB training datasets were derived from the personal genomes of only six common cell lines. In addition, there is an extensive disagreement between existing ASB prediction methods, further reducing the quantity of high-confidence training variants. Recently, Adastra, a new database hosting 652,595 ASBs passing 5% FDR across 647 cell lines and 1,043 TFs, was published (21). However, Adastra's predicts ASB based on a statistical model, not by using personal genomes directly. Therefore, it is unclear whether incorporating these ASB predictions in the high-confidence training datasets will yield meaningful improvements to the TLand models. This remains an area of ongoing interest.

Overall, TLand advances the field of non-coding variant analysis by incorporating comprehensive datasets to predict functional regulatory variants at cell-specific and organ-specific levels. To foster downstream applications, pre-trained TLand models and pre-calculated TLand scores for 2 million GWAS genomic variants are publicly-available at https://doi.org/10.5281/zenodo.7818540. TLand's flexible models and ability to integrate novel datasets in a principled way significantly improve our ability to identify human regulatory variants genome-wide. This ability will only improve as more and larger training datasets, spanning additional organs, tissues, cell types, and assays are released.

Availability

To facilitate the reproducibility of our analyses, we have made the following resources available: Scripts to reproduce the analyses can be found at

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

https://github.com/nsamzhao/tland_supplementary. The data used in the analyses and final trained models can be accessed at Zenodo https://doi.org/10.5281/zenodo.7818540.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Yang,Y., Muzny,D.M., Reid,J.G., Bainbridge,M.N., Willis,A., Ward,P.A., Braxton,A., Beuten,J., Xia,F., Niu,Z., *et al.* (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, **369**, 1502–1511.

2. 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.

3. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A., *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

4. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

5. Worthey,E.A., Mayer,A.N., Syverson,G.D., Helbling,D., Bonacci,B.B., Decker,B., Serpe,J.M., Dasu,T., Tschannen,M.R., Veith,R.L., *et al.* (2011) Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.*, **13**, 255–262.

6. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A., *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.

7. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

8. International HapMap Consortium, Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

9. 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.

10. 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.

11. Dong,S., Zhao,N., Spragins,E., Kagda,M.S., Li,M., Assis,P., Jolanki,O., Luo,Y., Cherry,J.M., Boyle,A.P., *et al.* (2023) Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nat. Genet.*

12. ENCODE Project Consortium, Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.

13. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

14. Dong,S. and Boyle,A.P. (2019) Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum. Mutat.*, **40**, 1292–1298.

15. Dong,S. and Boyle,A.P. (2022) Prioritization of regulatory variants with tissue-specific function in the non-coding regions of human genome. *Nucleic Acids Res.*, **50**, e6.

16. Pazin,M., Gilchrist,D.A. and Morris,S.A. Impact of Genomic Variation on function (IGVF) Consortium. *Genome.gov*.

17. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

18. Avsec,Ž., Agarwal,V., Visentin,D., Ledsam,J.R., Grabska-Barwinska,A., Taylor,K.R., Assael,Y., Jumper,J., Kohli,P. and Kelley,D.R. (2021) Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, **18**, 1196–1203.

19. Wolpert,D.H. (1992) Stacked generalization. *Neural Netw.*, **5**, 241–259.

20. 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.

21. Abramov,S., Boytsov,A., Bykova,D., Penzar,D.D., Yevshin,I., Kolmykov,S.K., Fridman,M.V., Favorov,A.V., Vorontsov,I.E., Baulin,E., *et al.* (2021) Landscape of allele-specific transcription factor binding in the human genome. *Nat. Commun.*, **12**, 2751.

22. Ke,G., Meng,Q., Finley,T., Wang,T., Chen,W., Ma,W., Ye,Q. and Liu,T.-Y. (2017) LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, **30**.

23. Ho,T.K. (1995) Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*.Vol. 1, pp. 278–282 vol.1.

24. McCulloch,W.S. and Pitts,W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, **5**, 115–133.

25. Akiba,T., Sano,S., Yanase,T., Ohta,T. and Koyama,M. (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’19. Association for Computing Machinery, New York, NY, USA, pp. 2623–2631.

26. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Müller,A., Nothman,J., Louppe,G., *et al.* (2012) Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]*.

27. Raschka,S. (2018) MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack. *J. Open Source Softw.*, **3**, 638.

28. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.

29. Neph,S., Scott Kuehn,M., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J., Thomas,S., *et al.* (2012) BEDOPS:

bioRxiv preprint doi: https://doi.org/10.1101/2023.09.07.556700; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

high-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.

30. van der Sande,M. and van Heeringen,S. (2021) qnorm.

31. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, **12**, 931–934.

32. Chen,K.M., Wong,A.K., Troyanskaya,O.G. and Zhou,J. (2022) A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.*, **54**, 940–949.

33. ChatGPT.

34. Student (1908) The probable error of a mean. *Biometrika*, **6**, 1–25.

35. Dey,K.K., van de Geijn,B., Kim,S.S., Hormozdiari,F., Kelley,D.R. and Price,A.L. (2020) Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat. Commun.*, **11**, 4703.

36. Dey,K.K., Kim,S.S., Gazal,S., Nasser,J., Engreitz,J.M. and Price,A.L. (2021) Integrative approaches to improve the informativeness of deep learning models for human complex diseases. *bioRxiv*, 10.1101/2020.09.08.288563.

37. Ting,K.M. and Witten,I.H. (1999) Issues in Stacked Generalization. *jair*, **10**, 271–289.

38. Zhu,W., He,X., Cheng,K., Zhang,L., Chen,D., Wang,X., Qiu,G., Cao,X. and Weng,X. (2019) Ankylosing spondylitis: etiology, pathogenesis, and treatments. *Bone Res*, **7**, 22.

39. Lee,D., Gorkin,D.U., Baker,M., Strober,B.J., Asoni,A.L., McCallion,A.S. and Beer,M.A. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.