Assessing the utility of genomic deep learning models for disease-relevant variant effect prediction

Pooja Kathail [1] **Richard Shuai** [2] **Ryan Chung** [1] **Chun Jimmie Ye** [3 4 5 6 7 8] **Gabriel Loeb** [9] **Nilah M. Ioannidis** [1 2 8]

Abstract

Over the past several years, highly accurate deep learning models have been developed to predict epigenetic features such as chromatin accessibility directly from DNA sequence. These models have the potential to assign function to disease-relevant non-coding variation, but their current utility for variant effect prediction remains limited. Here, we identify two features of these models that may limit their performance when applied to disease variant functionalization. First, we find that these models have reduced performance in cell-type specific *cis* regulatory elements, which contain a large fraction of the heritability of complex diseases. Second, we show that typical accuracy metrics used to evaluate the performance of these models on reference sequences are not indicative of performance on variant effect prediction, as measured by allele-specific accessibility. In particular, we show that modeling decisions that improve ”reference accuracy” do not always improve variant effect accuracy. Based on these findings, we propose the use of allele-specific accessibility data in disease-relevant regulatory regions to evaluate future sequence-based epigenetic models.

1. Introduction

The vast majority of disease-associated genetic variants identified by genome-wide association studies (GWAS) are located in non-coding regions of the genome and likely regulate gene expression (Visscher et al., 2017). Pinpointing the causal variants within an associated locus and annotating the mechanisms by which these variants act to modulate disease risk is an open challenge in human genetics.

One promising approach is to use deep learning models that predict epigenetic features such as chromatin accessibility, histone marks, and gene expression directly from DNA sequence (Zhou & Troyanskaya, 2015; Kelley et al., 2016; 2018; Zhou et al., 2018; Agarwal & Shendure, 2020; Avsec et al., 2021). These models are trained on the reference genome sequence, but can make predictions about sequences that differ from reference, e.g. by a single nucleotide. In theory, these variant effect predictions can identify causal variants and annotate the transcription factors and *cis* regulatory elements (CREs) that control expression of nearby genes. However, it was recently noted that variant effect predictions from current deep learning models contain limited unique information about complex disease heritability, when conditioned on a broad set of coding, conserved and regulatory annotations (Dey et al., 2020).

Here, we seek to understand limitations of these deep learning models with respect to disease-relevant variant effect prediction, and why these limitations might not be reflected by typical measures of model accuracy. We base our analysis on two key ideas. First, disease-relevant non-coding variation is not uniformly distributed throughout the genome. Cell-type specific CREs, in particular, are known to harbor a large fraction of the heritability of complex diseases (Finucane et al., 2015). Second, the accuracy metric—which we refer to here as ”reference accuracy”—used to train and evaluate these models does not directly measure their ability to predict variant effects, and may not always correlate with the latter.

Using two ATAC-seq datasets from kidney and immune cells, we train deep learning models, similar to the Basset architecture (Kelley et al., 2016), to predict chromatin accessibility in multiple cell types from DNA sequence. We

[1] Center for Computational Biology, University of California, Berkeley [2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley [3] Division of Rheumatology, Department of Medicine, University of California, San Francisco [4] Institute for Human Genetics, University of California, San Francisco [5] Department of Epidemiology and Biostatistics, University of California, San Francisco [6] Bakar Computational Health Sciences Institute, University of California, San Francisco [7] Parker Institute for Cancer Immunotherapy, San Francisco, CA [8] Chan Zuckerberg Biohub, San Francisco, CA [9] Division of Nephrology, Department of Medicine, University of California, San Francisco. Correspondence to: Pooja Kathail <pooja.kathail@berkeley.edu>, Nilah Ioannidis <nilah@berkeley.edu>.

The 2022 ICML Workshop on Computational Biology.

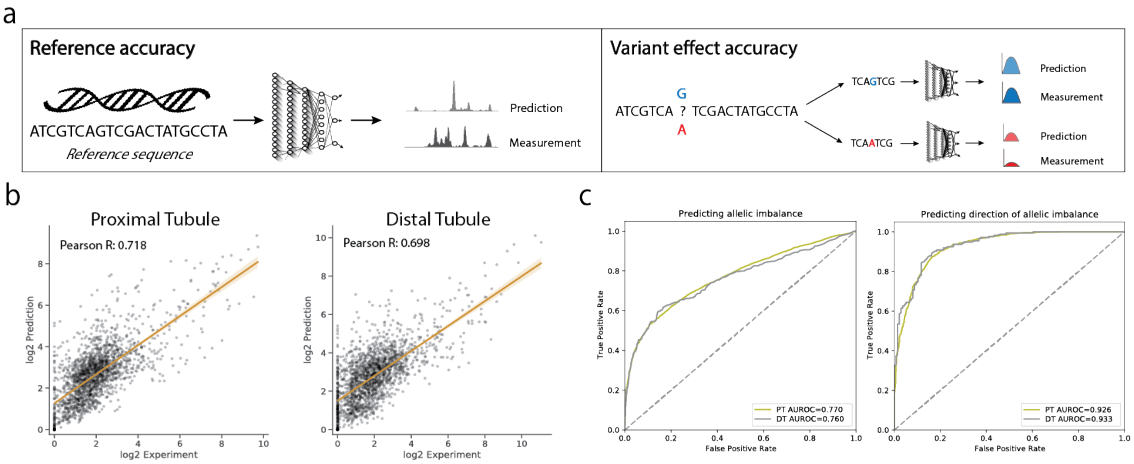

Figure 1. (a) Model evaluation using reference and variant effect accuracy. Reference accuracy reflects the training task, predicting experimentally measured chromatin accessibility from an input reference sequence. Variant effects are predicted by comparing model predictions for sequences containing reference (red) and alternate (blue) alleles as separate inputs. (b) Reference accuracy for two kidney cell types, proximal tubule (PT) and distal tubule (DT): Pearson correlation between experimental and predicted chromatin accessibility for held-out sequences (c) Variant effect accuracy for the same cell types: AUROC for predicting allelic imbalance and imbalance direction at heterozygous variants.

find that our models achieve high genome-wide predictive accuracy, but perform poorly in disease-relevant cell-type specific regulatory elements. We further evaluate the effect of a number of common training decisions on both reference and variant effect accuracy, and find that these decisions can have differing effects on the two metrics. Together, these results reveal factors limiting the utility of current deep learning models for variant effect prediction and highlight targets for future modeling improvements.

2. Data and methods

2.1. Data

We use two ATAC-seq datasets for model training and evaluation throughout our manuscript:

- **Loeb et al. (*manuscript in preparation*)**: Single-cell ATAC-sequencing of primary human kidney tissue from three donors. Data were clustered into 10 cell types, and pseudobulk ATAC data for each cell type was generated. For additional details, see Appendix B.1.
- **Calderon et al. (2019)**: Bulk ATAC-sequencing of 25 primary human immune cell types, sorted by flow cytometry, from four human blood donors.

2.2. Model architecture and training

We train convolutional neural networks (CNNs) to map input DNA sequences (1344 bp) to continuous measures of chromatin accessibility (normalized ATAC-seq read counts). Our architecture is based on an updated version of the Basset model, which consists of 8 convolutional layers followed by 2 fully connected layers (Kelley et al., 2016). We modify the architecture to predict continuous—rather than binary—values, which has recently been shown to improve model generalizability and interpretability (Toneyan et al., 2022), and train models to minimize the Poisson regression loss function. We evaluate 4 types of training procedures, including single task and multitask learning (described in Table 1). In each case, we use chromosomes 7, 14, and 15 for validation, chromosomes 4 and 5 for evaluation, and all other chromosomes for training. We use the Basenji repository (Kelley et al., 2018) for data preprocessing, model training, and evaluation.

Table 1. Description of evaluated models.

MODEL	OUTPUT TASKS	TRAINING REGIONS
1	One cell type (Single task)	All genomic sequences
2	Many cell types (Multitask)	All genomic sequences
3	Many cell types (Multitask)	Sequences overlapping any ATAC peak
4	Many cell types (Multitask)	Sequences overlapping non-ubiquitous ATAC peaks

2.3. Model evaluation

We consider two measures of model performance throughout our manuscript (Fig. 1A):

- **Reference accuracy:** Pearson correlation between experimental ATAC measurements and model predictions for input sequences from the reference genome.

- **Variant effect accuracy:** Classification accuracy (AUROC) of *in silico mutagenesis* predictions of variant effects vs. experimentally assayed effects of variants on chromatin accessibility (allelic imbalance) at heterozygous sites using variant-specific read mapping. For additional details, see Appendix B.2.

As an example, we show the reference accuracy of the **multitask** model (Model 2) for two kidney cell types from the Loeb et al. data (Fig. 1B). The model achieves a genome-wide reference accuracy comparable to previous work (Kelley et al., 2018). We also show the model's variant effect accuracy in the same cell types (Fig. 1C). The model's variant predictions are informative for classifying allelically imbalanced sites and their direction of imbalance.

3. Results

3.1. Cell-type specific CREs contain a large proportion of SNP heritability

It has previously been shown that cell-type specific CREs harbor much of the common genetic variation explaining heritability of human complex traits and diseases. To verify this for both of our datasets, we grouped open chromatin peaks into disjoint clusters based on their accessibility profiles across cell types, giving us one cluster in each dataset corresponding to ubiquitous peaks and additional clusters displaying cell-type specificity. We estimated the fraction of heritability explained by these ubiquitous or cell-type specific open chromatin clusters using partitioned LD score regression (Finucane et al., 2015) for immune-related traits (for the Calderon et al. data) or the kidney function marker creatinine (for the Loeb et al. data) in the UK Biobank (GWAS summary statistics were obtained from the Price lab server for immune-related traits and the Neale lab server for creatinine). We find that cell-type specific peak clusters explain 59% of creatinine heritability, a 7.2-fold increase over the heritability explained by ubiquitous peaks, and an average of 35% of immune-related disease heritability, a 3.4-fold increase over the average heritability explained by ubiquitous peaks (Fig. 2). These results confirm the importance of cell-type specific CREs for understanding the genetic underpinnings of complex diseases.

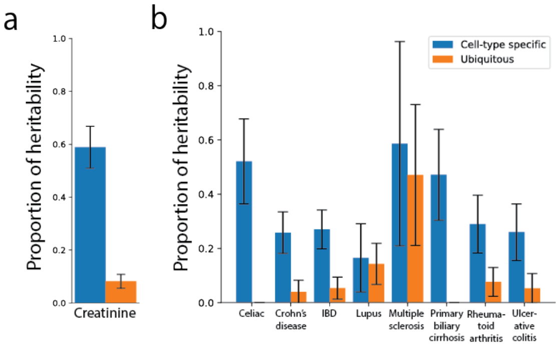

Figure 2. Proportion of (a) creatinine (a kidney function biomarker) heritability and (b) immune-related disease heritability explained by cell-type specific and ubiquitous CREs.

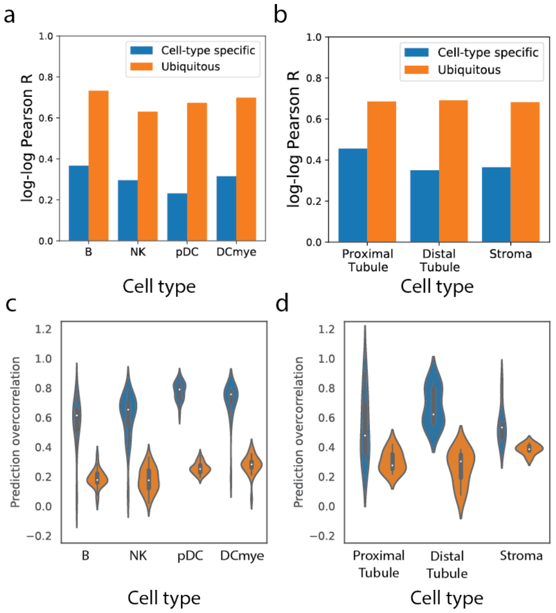

Figure 3. (a,b) Reference accuracy and (c,d) overcorrelation (prediction correlation across cell types - experimental correlation across cell types) for the **multitask** model's predictions in cell-type specific vs. ubiquitous CREs for the Calderon et al. (a,c) and Loeb et al. (b,d) datasets.

3.2. Sequence-to-accessibility models perform poorly in cell-type specific CREs

Next, we quantified the predictive accuracy of the trained models when considering cell-type specific and ubiquitous clusters separately. We used performance of the **multitask** model (Model 2) as a baseline, since prior work has largely focused on training these models using multitask learning across cell types. The **multitask** model's reference accuracy in cell-type specific CREs (0.30 avg. Pearson R for Calderon et al.; 0.39 for Loeb et al.) is markedly lower than in ubiquitous CREs (0.68 for Calderon et al.; 0.69 for Loeb et al.) (Fig 3A,B). While the genome-wide performance of our models is consistent with prior literature, to our knowledge, our results are the first demonstration that predictive accuracy is dramatically lower in cell-type specific CREs.

We tested whether this discrepancy could be caused by higher accessibility on average in ubiquitous peaks. We quantified the distribution of peak heights and found that cell-type specific peaks tended to have lower peak heights than ubiquitous peaks in the single-cell Loeb et al. data (Fig. S1B). However, we did not observe a similar association between peak height and cell-type specificity in the bulk Calderon et al. data (Fig. S1A). After controlling for this bias in the Loeb et al. data, we still observe a drop in predictive accuracy in cell-type specific CREs when compared to height-matched ubiquitous CREs (0.39 vs. 0.50 avg. Pearson R) (Fig. S1C).

3.3. Effect of multitask architecture and training set composition on cell-type specific CRE prediction

We next sought to assess how two common training decisions affect performance in cell-type specific CREs. First, we hypothesized that the multitask architecture might cause the model to learn shared, rather than cell-type specific features. We looked at correlation in the predicted peak heights across cell types, compared to correlation in the experimental peak heights. In ubiquitous regions, which were defined based on the experimental data, distinct cell types have similar levels of accessibility, while accessibility in cell-type specific regions is poorly correlated across cell types. However, we found that the **multitask** model's predictions in cell-type specific regions are highly correlated across cell types (Fig. 3C,D). We observe a slight over-correlation relative to the experimental data in ubiquitous regions as well. To test whether this over-correlation is caused by the multitask architecture, we evaluated the effect of training on a single cell-type, the **single task** model (Model 1). Single task training yielded a drop in overall test set accuracy, but led to a small performance improvement in cell-type specific regions (Fig. S2, Fig. 4A). While predictions across cell types from the **single task** models are less correlated than the **multitask** model, we still observe an over-correlation which is more pronounced in cell-type specific regions (Fig. S3). This suggests that even single task models are primarily learning sequence features that are shared across cell types.

The second decision we explored was what regions of the genome to include in training. For the **single task** and **multitask** models, we include all regions of the genome, apart from assembly gaps and unmappable regions. Cell-type specific CRE sequences make up less than 10% of this training set. We hypothesized that the relative infrequency of these sequences in the training set might contribute to poor performance. We evaluated two simple modifications to our training scheme: (i) training only on sequences overlapping an ATAC-seq peak in at least one cell type, and (ii) training only on sequences overlapping a non-ubiquitous ATAC-seq peak. The **multitask, peaks only** model (Model 3) improved performance in both cell-type specific and ubiquitous regions (Fig. 4A). This is especially striking given that removing non-peaks reduced our training set size by more than 90%. The **multitask, non-ubiquitous peaks only** model (Model 4) further improved performance for cell-type specific peaks (Fig. 4A). However, this model performed poorly on ubiquitous peaks, which is unsurprising as it was given no training examples of this peak type. For reference, we also include the inter-individual correlation across the three donors in cell-type specific and ubiquitous regions as a measure of reproducibility and a reasonable upper bound on performance (Fig. 4A).

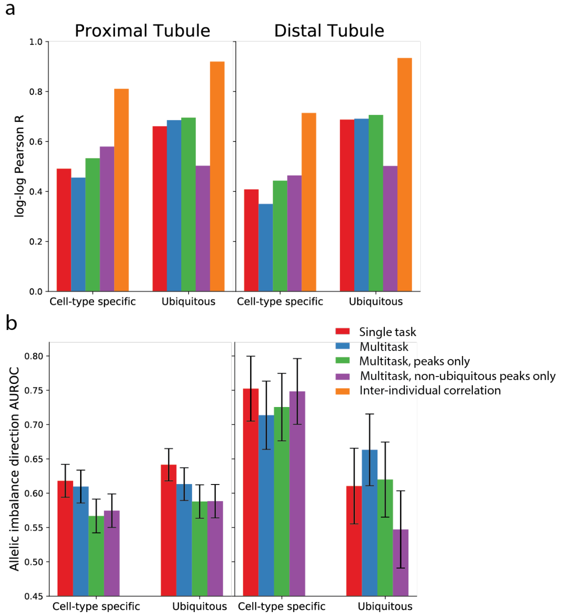

Figure 4. Evaluation of common training decisions on (a) reference and (b) variant effect accuracy in cell-type specific and ubiquitous CREs in the Loeb et al. dataset.

3.4. Training decisions have differing effects on reference accuracy and variant effect accuracy

We also evaluated the effect of the above training decisions on variant effect accuracy, to determine whether reference accuracy is a sufficient proxy for variant effect accuracy in model selection. Using allelic imbalance measurements at heterozygous sites in the experimental ATAC-seq data, we computed an AUROC for the model's ability to predict the direction of imbalance, or the more accessible allele, using *in silico mutagenesis*. In doing this evaluation, we observed that cell-type specific CREs tend to have more imbalanced sites than ubiquitous CREs (Fig. S4). To account for this bias, we matched imbalance distributions between cell-type specific and ubiquitous regions for each cell type. In contrast to reference accuracy, training with the **multitask, peaks only** or **multitask, non-ubiquitous peaks only** models did not improve variant effect prediction (Fig. 4B). Thus, we find that modifications that improve reference accuracy do not necessarily translate into improvements in variant effect accuracy.

4. Discussion

In this work, we identify two factors that may limit the application of deep learning models to disease-relevant non-coding variation. We find that models trained to predict chromatin accessibility from DNA sequence perform poorly in cell-type specific CREs, which contain a large proportion of complex disease heritability. We also find that the commonly reported reference accuracy does not always correlate with variant effect accuracy, and that a number of common training decisions can have different effects on these two accuracy metrics, highlighting the importance of directly evaluating models on their ability to predict variant effects. To facilitate this evaluation, we propose using allele-specific accessibility data, which can be readily obtained from the same experimental data used for model training, with some considerations to facilitate unbiased allele-specific read mapping (van de Geijn et al., 2015). Future work will explore strategies to incorporate allele-specific information into model training, and to extend this work to models of gene expression.

Acknowledgements

We thank members of the Ioannidis and Ye labs for helpful discussions. This work was supported in part by the National Human Genome Research Institute grant R00HG009677 to N.M.I. C.J.Y. and N.M.I. are Chan Zuckerberg Biohub Investigators.

Disclosures

C.J.Y. is a Scientific Advisory Board member for and holds equity in Related Sciences and ImmunAI, a consultant for and holds equity in Maze Therapeutics, and a consultant for TReX Bio. C.J.Y. has received research support from Chan Zuckerberg Initiative, Chan Zuckerberg Biohub, and Genentech.

References

Agarwal, V. and Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.*, 31(7):107663, May 2020.

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. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, 18(10):1196–1203, October 2021.

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Müller, F., Nguyen, V., Lescano, N., Wu, B., Trombetta, J., Ribado, J. V., Knowles, D. A., Gao, Z., Blaeschke, F., Parent, A. V., Burt, T. D., Anderson, M. S., Criswell, L. A., Greenleaf, W. J., Marson, A., and Pritchard, J. K. Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.*, 51(10):1494–1505, October 2019.

Dey, K. K., van de Geijn, B., Kim, S. S., Hormozdiari, F., Kelley, D. R., and Price, A. L. Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat. Commun.*, 11(1):4703, September 2020.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M., and Price, A. L. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, 47(11):1228–1235, November 2015.

Kelley, D. R., Snoek, J., and Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, 26(7):990–999, July 2016.

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.*, 28(5):739–750, May 2018.

Toneyan, S., Tang, Z., and Koo, P. K. Evaluating deep learning for predicting epigenomic profiles. May 2022.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods*, 12(11): 1061–1063, November 2015.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.*, 101(1):5–22, July 2017.

Zhou, J. and Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, 12(10):931–934, October 2015.

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.*, 50(8):1171–1179, August 2018.

A. Supplementary Figures

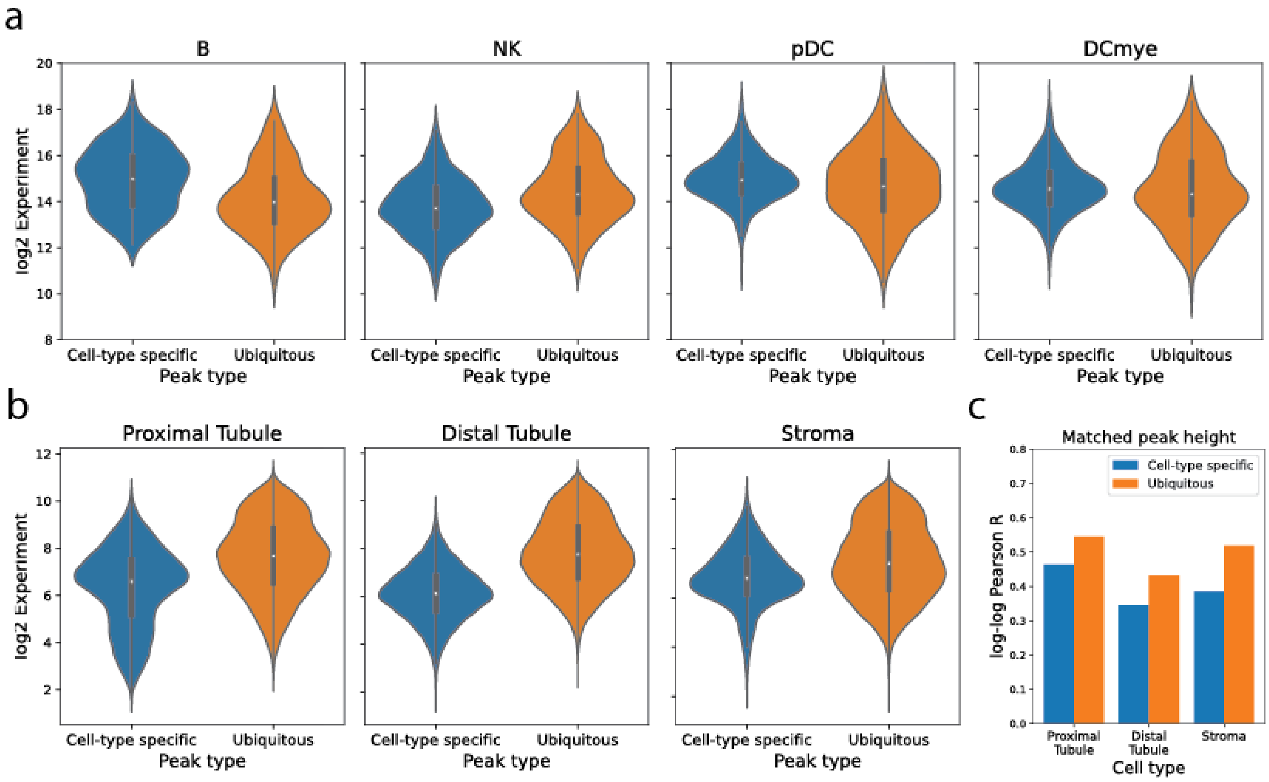

Figure S1. Peak heights in cell-type specific and ubiquitous CREs in the (a) Calderon et al. (2019) and (b) Loeb et al. data. (c) Predictive reference accuracy of the **multitask** model's predictions in cell-type specific and ubiquitous CREs in the Loeb et al. data after matching on peak height.

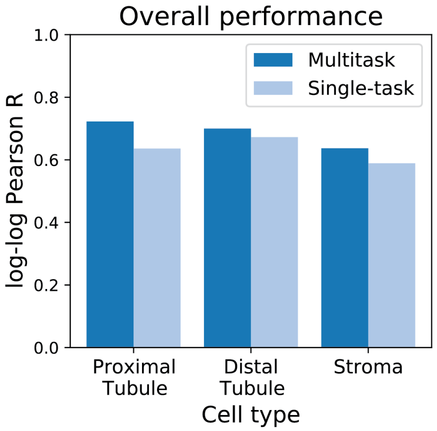

Figure S2. Reference accuracy of the **multitask** and **single-task** models.

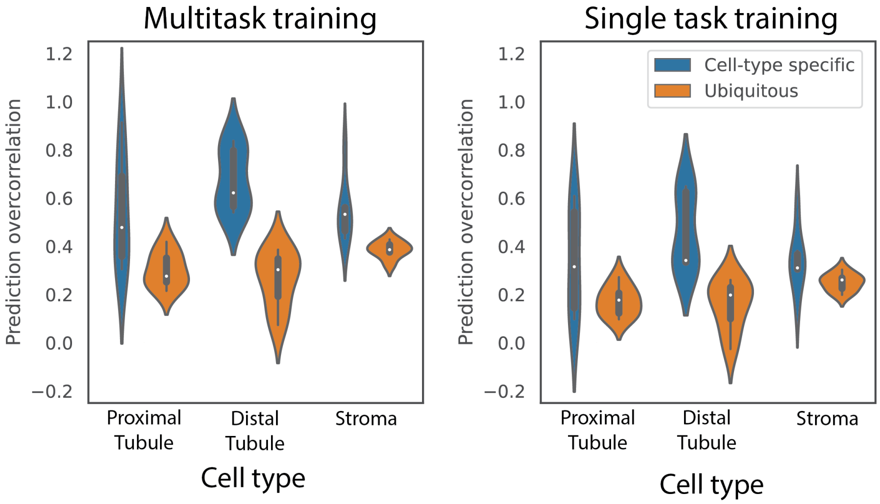

Figure S3. Overcorrelation of the **multitask** and **single-task** models' predictions in cell-type specific and ubiquitous CREs.

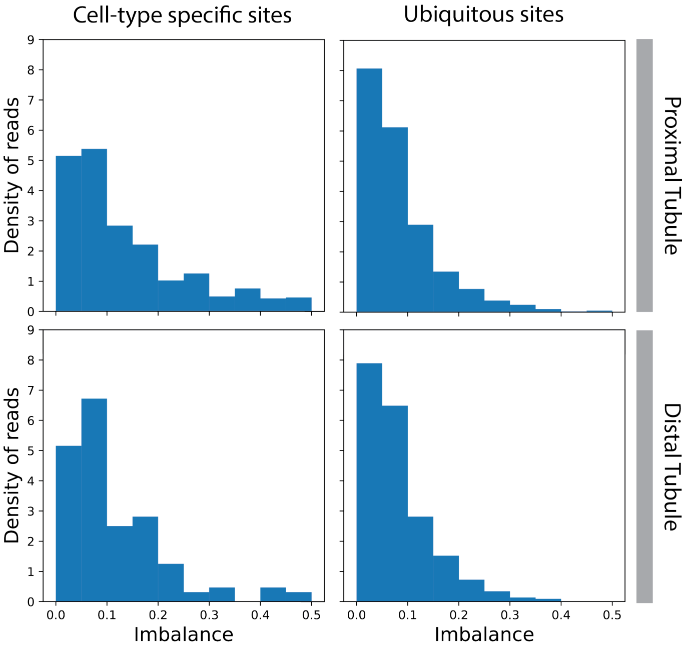

Figure S4. Allelic imbalance distributions at heterozygous sites in cell-type specific and ubiquitous CREs in two cell types of the Loeb et al. data. Allelic imbalance is defined as the absolute value of the proportion of total reads mapping to the reference allele - 0.5 [abs(REF/(REF+ALT) - 0.5)].

B. Additional Methods

B.1. Kidney data collection and processing

Kidney cortex and medullary tissue from deceased donor kidneys rejected for transplant were processed for ATAC-sequencing. Nuclei were extracted, counted, and single cell ATAC-sequencing libaries were generated using the 10x platform (v1.1). Data preprocessing was performed with Cell Ranger ATAC prior to cell quality control, cell clustering, and peak calling using ArchR (Granja et al., 2021). 10 cell types were identified by clustering. Peaks were grouped into disjoint clusters based on their accessibility profiles across cell types, giving the ubiquitous and cell-type specific peak clusters used in our analyses. Transposition sites from each cell cluster were extracted from Cell Ranger generated fragment files to generate bigwigs for model training. Further dataset details in Loeb et al. (in preparation).

B.2. Evaluating variant effect accuracy

To evaluate variant effect accuracy, we constructed a set of variants with allelic imbalance in two cell types of the Loeb et al. data: proximal tubule (PT) and distal tubule (DT). To generate these sets we remapped reads using WASP (van de Geijn et al., 2015), identified heterozygous variants with a read count greater than 20, computed the allelic imbalance ratio REF / (REF + ALT), used a binomial test to obtain p-values for the significance of allelic imbalance, and corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. We defined a set of allelic imbalanced variants in each cell type using a false discovery rate threshold of 0.01. We defined non-allelic imbalanced variant sets as those variants with p-value greater than 0.1 in each cell type. The non-allelic imbalanced sets were adjusted to be as large as possible while matching the read count distribution of the positive set variants on a log scale, resulting in a non-allelic imbalanced set that was 7x the size of the allelic imbalanced set in each of PT and DT.

To evaluate variant effect accuracy genome-wide (Fig. 1), we used two related classification tasks. For both tasks, we defined the model's predictions for the reference and alternate alleles in the 192 bp bin centered at the variant as REF and ALT respectively, and we computed the predicted allelic imbalance as REF / (REF + ALT). In the first task, we tested whether predicted allelic imbalance could classify whether or not a variant showed allelic imbalance in the experimental data (i.e. discriminate between the allelic imbalanced and non-allelic imbalanced sets of variants defined above), as measured by the area under the receiver operating characteristic curve (AUROC). In the second task, we considered only variants in the allelic imbalanced set, and tested whether predicted allelic imbalance could classify the direction of imbalance (i.e. the more accessible allele, REF or ALT), as measured by AUROC.

In order to have a sufficient number of variants to evaluate variant effect accuracy separately within cell-type specific and ubiquitous CREs (Fig. 4), we used all heterozygous variants within these regions with a read count greater than 20. We tested whether predicted allelic imbalance could classify the direction of imbalance (i.e. the more accessible allele, REF or ALT) across all heterozygous variants, as measured by AUROC.