scooby: Modeling multi-modal genomic profiles from DNA sequence at single-cell resolution

Johannes C. Hingerl[1,2†], Laura D. Martens[1,4†], Alexander Karollus[1,2], Trevor Manz[5], Jason D. Buenrostro[6,7], Fabian J. Theis[1,4], Julien Gagneur[1,2,3,4*]

[1] School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[2] Munich Center for Machine Learning, Munich, Germany
[3] Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany
[4] Computational Health Center, Helmholtz Center Munich, Neuherberg, Germany
[5] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[6] Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA 02142 USA
[7] Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA

[†] Equal contribution
[*] Corresponding author

Abstract

Understanding how regulatory DNA elements shape gene expression across individual cells is a fundamental challenge in genomics. Joint RNA-seq and epigenomic profiling provides opportunities to build unifying models of gene regulation capturing sequence determinants across steps of gene expression. However, current models, developed primarily for bulk omics data, fail to capture the cellular heterogeneity and dynamic processes revealed by single-cell multi-modal technologies. Here, we introduce scooby, the first model to predict scRNA-seq coverage and scATAC-seq insertion profiles along the genome from sequence at single-cell resolution. For this, we leverage the pre-trained multi-omics profile predictor Borzoi as a foundation model, equip it with a cell-specific decoder, and fine-tune its sequence embeddings. Specifically, we condition the decoder on the cell position in a precomputed single-cell embedding resulting in strong generalization capability. Applied to a hematopoiesis dataset, scooby recapitulates cell-specific expression levels of held-out genes and cells, and identifies regulators and their putative target genes through in silico motif deletion. Moreover, accurate variant effect prediction with scooby allows for breaking down bulk eQTL effects into single-cell effects and delineating their impact on chromatin accessibility and gene expression. We anticipate scooby to aid unraveling the complexities of gene regulation at the resolution of individual cells.

Introduction

Predicting molecular traits from genomic sequence, also called sequence-to-function modeling[1], is an effective strategy for uncovering the genetic basis of gene regulation and complex traits. Deep learning architectures can predict and interpret not only genomic readouts but also the effects of genetic variants directly from DNA sequence[2–8]. With the advent of single-cell technologies, sequence-based models of single-cell omics assay were developed aiming at unraveling sequence determinants of cell-state specific regulation and cell fate decisions[9–13]. Early work modeled local chromatin accessibility by predicting scATAC-seq data using convolutional neural networks operating on small input sequences (1 kb)[9,11]. While scATAC-seq provides valuable insights into chromatin state and transcription factor binding, gene expression remains the ultimate functional readout of regulatory activity. Therefore, local predictors of binarized or pseudobulked gene expression have been proposed[10,13], yet their sequence context is limited and fails to capture cell-type specific gene regulation[12].

To address this, the recently presented seq2cells model adapts Enformer[7], a state-of-the-art model for gene expression prediction trained on thousands of bulk omics assays to infer single-cell gene expression counts from 200 kb of sequence. However, seq2cells only predicts gene expression, relies on potentially ambiguous transcription start site annotations[8], and models each cell with a separate output track, leading to computational intractability for large datasets.

To overcome these limitations, we present scooby, which jointly models scATAC-seq and scRNA-seq genomic profiles without the need for annotations and easily scales to large datasets. scooby builds upon Borzoi[8], a recently released bulk omics model that uses RNA-seq coverage as an annotation-free representation of gene regulation. Using Borzoi as a foundation model, we equip it with a cell-specific decoder, and fine-tune its sequence embeddings to adapt it to the single-cell setting. We call this adapted model "scooby", a nod to both its single-cell (sc) focus and its canine-named model ancestry. We demonstrate scooby's capabilities to accurately model genomic profiles at single-cell resolution, identify lineage-specific regulators and their putative target genes, and delineate cell-type-specific variant effects on a multiome hematopoiesis dataset, benchmarking it against state-of-the-art models.

Results

scooby enables the modeling of genomic profiles at single-cell resolution

Here, we present scooby, a model predicting single-cell accessibility and expression profiles from DNA sequence (Fig. 1a). Our model builds upon Borzoi, a state-of-the-art sequence-based model for RNA-seq coverage prediction, and leverages its trained convolutional and transformer-based architecture to extract informative sequence embeddings at 32 bp resolution. To tailor these embeddings to individual single-cell datasets and achieve single-cell resolution profile predictions, scooby introduces two key innovations (Fig. 1a, Extended Data Fig. 1).

First, to enable scooby to efficiently adapt to dataset-specific features, we finetuned its sequence embeddings using low-rank adaptation (LoRA)[14], a parameter-efficient fine-tuning strategy. Following the LoRA approach, we kept pre-trained weights frozen and added trainable low-rank matrices into the transformer and convolutional layers (Methods). Advantageously, these matrices can be merged into the existing weights after training, resulting in no overhead during model inference[14]. We reasoned that this would allow scooby to capture effects of regulatory sequences relevant to cell states that are absent from or weakened in the bulk data Borzoi was trained on, but also to adjust to characteristics inherent to single-cell assays, such as the 3' coverage bias common in scRNA-seq[15].

Second, we implemented a lightweight decoder for gene expression and accessibility prediction at single-cell resolution. To this end, scooby leverages low dimensional, multiomic representations of cell states, in this case derived from Poisson-MultiVI[16,17], to decode the fine-tuned sequence embedding generated by Borzoi in a cell-specific manner. This design differs from approaches that rely on separate output heads for each cell[12], whose number of parameters scales with the number of cells and cannot, by design, effectively leverage similarities between cells. For efficient analysis of large single-cell datasets with scooby, we developed an accessible workflow by adapting SnapATAC2.0[18] to store single-cell profiles in the widely-used AnnData[19,20] format which facilitates memory-efficient model training (Extended Data Fig. 2 and 3, Methods).

We ensured robust evaluation by following the same sequence-level train and test splits as our underlying foundation model Borzoi. Moreover, genes and scATAC-seq peaks overlapping with validation or test regions were excluded from the input data used to generate the single-cell embeddings to avoid data leakage.

We trained scooby on a 10x Single Cell Multiome dataset (joint single-nuclei RNA-seq and ATAC-seq) comprising 63,683 human bone marrow mononuclear cells (NeurIPS dataset, Fig. 1b,

Methods)[21] across 8 NVIDIA A40 GPUs for 2 days until convergence. A typical example of the model prediction is shown in Figure 1c for the *SLC25A37* locus. Despite the inherent sparsity of single-cell data, the profile predicted for a single cell captures the expected profile as indicated by the empirical averages over 100 neighboring cells. Moreover, the model precisely predicted the localization of scRNA-seq signal at the 3'-end of transcripts. This shows that scooby successfully adapted to the scRNA-seq assay, despite Borzoi being originally trained to predict full-length RNA-seq coverage. Importantly, the model effectively captured the differential regulation of the *SLC25A37* locus, accurately inferring lower expression in a megakaryocyte erythroid progenitor cell compared to an erythroblast cell and highlighting distinct accessibility patterns between these cells (Fig. 1c).

To assess the performance of scooby globally, we first computed the Pearson correlation between predicted profiles and the observed profiles averaged across 100 cell neighbors, on a logarithmic scale for a random subset of cells of each cell type across all test sequences (Methods). We found that the profile predictions were highly correlated with the observations both for scRNA-seq (Mean Pearson correlation = 0.63) and for scATAC-seq (Mean Pearson correlation = 0.70, Fig. 1d). This compares to 0.74 and 0.81 for RNA-seq and DNase-seq for Borzoi on high-quality bulk data. Since averaging observed profiles across 100 cells can be a noisy estimate, we also compared our performance averaged across all cells of a cell type (scRNA-seq: Mean Pearson R = 0.67, scATAC-seq: Mean Pearson R = 0.72). These results demonstrate promising concordance between predicted and observed profiles, approaching the performance of bulk-pretrained models, even with the inherent challenges of sparse single-cell readouts.

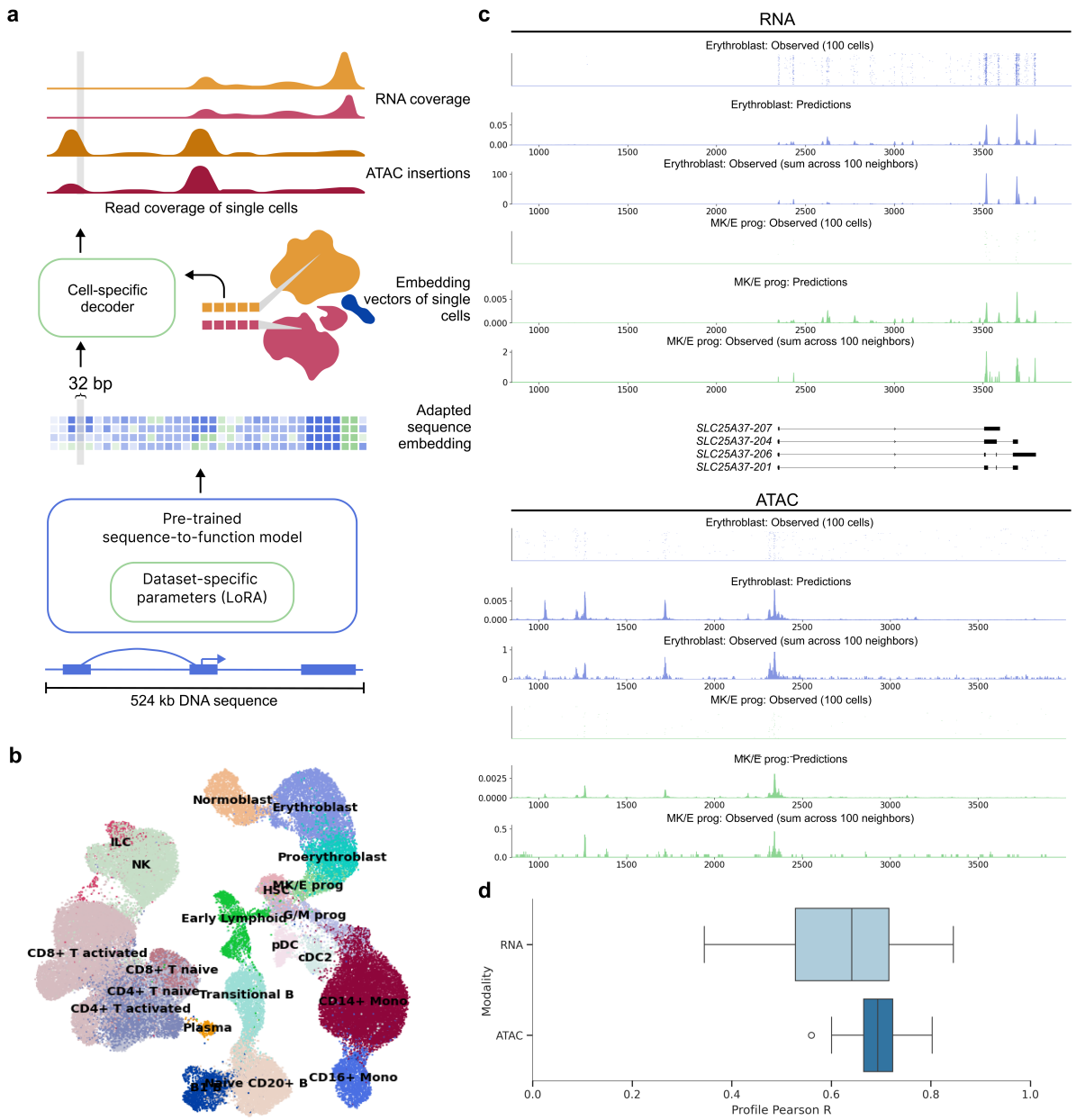

Figure 1: scooby accurately predicts cell-state-specific expression and accessibility profiles from single-cell data. a, scooby integrates a pre-trained sequence-to-profile model with a cell-state-specific decoder to predict genomic profiles at single-cell resolution. The pre-trained model is fine-tuned on the target dataset using a parameter-efficient strategy, generating an adapted sequence embedding at 32 bp resolution. The cell-state-specific decoder takes this sequence embedding together with embedding vectors of single cells as input to predict scATAC-seq insertion and scRNA-seq coverage profiles at single-cell level. **b,** Uniform Manifold Approximation and Projection (UMAP) visualization of the 10x multiome NeurIPS bone marrow dataset[21] integrated with Poisson-MultiVI, colored by cell type. **c,** Representative example of predicted and observed gene expression (top) and accessibility (bottom) profiles of an erythroblast and a megakaryocyte-erythroid progenitor cell and its 100 nearest neighbors at the *SLC25A37* locus (part of the

test set). **d,** Distribution of the correlation between predicted and observed profiles on test sequences (Methods, n=210 cells). In all boxplots, the central line denotes the median, boxes represent the interquartile range (IQR) and whiskers show the distribution except for outliers. Outliers are all points outside 1.5 × IQR. B, B cell; T, T cell; Mono, Monocyte; prog, progenitor; HSC, Hematopoietic stem cell; ILC, Innate lymphoid cell; Lymph, Lymphoid; MK/E, Megakaryocyte and Erythrocyte; G/M, Granulocyte and Myeloid; NK, Natural Killer cell; cDC2, Classical dendritic cell type 2; pDCs, Plasmacytoid dendritic cells.

scooby accurately predicts cell-type-specific gene expression counts

Given that marker genes exhibit distinct expression patterns across cell types, we reasoned that accurate prediction of these genes would provide initial evidence for scooby's capacity to capture cell-state-specific gene expression. To derive gene expression counts, we summed the predicted scRNA-seq coverage across exons for each cell and compared these predicted counts to the observed single-cell counts (Fig. 2a, Methods). The model accurately predicted cell-state-specific expression levels for marker genes unseen during training, even for small cell populations (Fig. 2b). In particular, it accurately predicted expression profiles of the markers *ANK1, DIAPH3, SLC25A37* and *AUTS2* that distinguish different cell types of erythroid differentiation (MK/E progenitors, (Pro-) Erythroblasts, Normoblasts).

For a quantitative analysis, we next evaluated scooby's performance at predicting pseudobulked gene expression profiles for each cell type (Fig. 2c). We calculated the log–fold change between predictions and the observed pseudobulk expression levels for each gene within each annotated cell type. Across all cell types, we observed a mean Pearson correlation ranging from 0.82 to 0.88 (mean Pearson R = 0.86, Fig. 2d and Extended Data Fig. 4), demonstrating accuracy comparable to the original Borzoi model trained on bulk RNA-seq data (0.86 mean Pearson R). To assess the extent to which scooby captures differential expression patterns, we calculated the correlation after subtracting both the gene-wise and cell-type-wise means on a logarithmic scale (Fig. 2c). This analysis, which focused on deviations from the global mean expression, yielded a Pearson correlation of 0.54, indicating that scooby successfully recapitulated a significant portion of the biological variation in gene expression across cell types (Fig. 2e). Notably, scooby substantially outperformed the count-based seq2cells model retrained on the NeurIPS dataset on shared Enformer and Borzoi test genes, with mean correlation across genes increasing from 0.77 to 0.87 and mean correlation across cell types increasing from 0.43 to 0.55 (Methods, Extended Data. Fig. 5).

We next performed some comparative analyses to understand the basis of scooby’s performance (Extended Data Fig. 6). A scooby model trained on the scRNA-seq data and using embeddings

purely derived from this modality performed worse than the multiomic model (Across genes Pearson R = 0.848, across cell types Pearson R = 0.496). Nevertheless, these results remain better than those of the seq2cells model and indicate that the increased performance of scooby is not limited to the fact it can leverage more data modalities.

Moreover, removing LoRA matrices to finetune Borzoi decreased prediction accuracy, particularly for relative expression between cell types (across cell types Pearson R = 0.501). We also designed a comparable model for the direct prediction of cell-type level, pseudobulked profiles, which are predicted in a multitask fashion while keeping the rest of the architecture the same. Still, scooby performed on par or slightly better on both cell-type level metrics, showing that the scooby architecture does not trade cellular resolution for performance.

scooby extrapolates gene expression dynamics to unseen cell states

A salient feature of scooby over seq2cells is that cell-specific predictions are achieved using a single-cell embedding representation instead of considering every single cell as a distinct task. In principle, this model design allows scooby to be applied to cells not seen during training. To assess this generalization capacity in a non-trivial setting, we evaluated a model where all normoblast cells, constituting the terminal cell type of the erythroid lineage, were withheld during training. Remarkably, using the normoblast embeddings yielded predictions with an accuracy close to the model trained on the full dataset (0.78 Pearson R compared to 0.81 for the model trained with normoblasts). Moreover, the best predictions were obtained using the normoblast embeddings as input to the decoder, followed by imputing with closely related cell types, consistent with scooby making cell-type specific predictions based on information from the single-cell embedding space (Fig. 2f).

Building on this observation, we investigated whether scooby's capacity to generalize to unseen cell states extends to capturing the continuous gene expression changes that occur during differentiation. As a case study, we considered Hemogen (*HEMGN*), a gene known to be upregulated during erythroid differentiation that was part of the sequences held out during model training. Using diffusion pseudotime[22] to order cells along the erythroid trajectory, we compared *HEMGN* expression dynamics predicted by both the full scooby model and the model trained without normoblasts (Fig. 2g). Both models, including the one trained without normoblasts, accurately recapitulated the regulation of *HEMGN* along the erythroid lineage (0.939 Pearson R for the full model and 0.937 for the ablated model).

Altogether, these results indicate that scooby can be applied to investigate unseen cell types and continuous regulatory programs similar to the ones observed during model training, a step

towards its potential application in reference atlas integration workflows which involves integrating a query dataset into large reference atlases to better resolve cellular heterogeneity and sub-cell type level gene regulation.

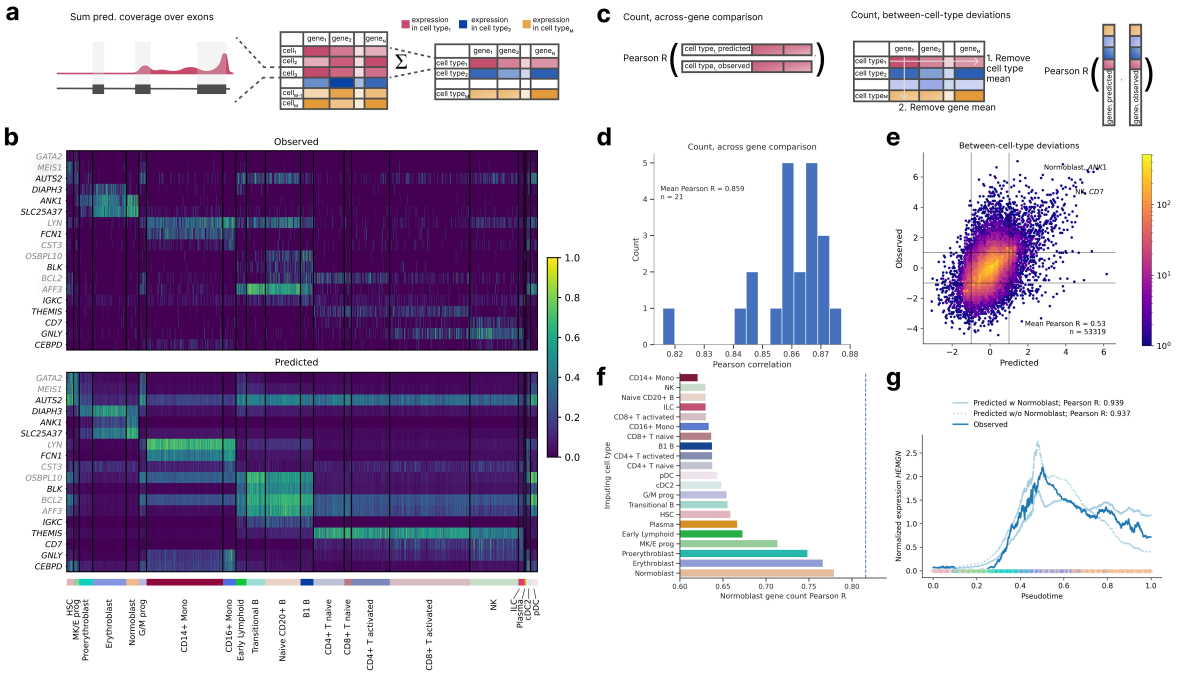

Figure 2: scooby accurately predicts cell-type-specific gene expression counts and generalizes to unseen cell states. a, Predicted and observed profiles are aggregated into a gene expression count matrix by summing coverage over exons. We obtain pseudobulk counts by summing over all predictions of every cell for each cell type. **b,** Normalized gene expression matrix (Methods) for cell-type-specific genes, observed (top) and predicted (bottom). Each row is a marker gene from test (black) or validation (gray), each column is a randomly selected cell. Cells are grouped by cell type (bottom track) **c,** We evaluate scooby's performance using two metrics: the correlation between predicted and observed gene expression counts within each cell type (left) and the model's ability to capture cell-type-specific deviations of gene expression to gene mean (right). **d,** Distribution of gene-level Pearson correlation between log-transformed predicted and observed counts of scRNA-seq reads overlapping exons across cell types. **e,** Predicted against measured between-cell-type deviations of gene expression. Exemplarily highlighted combinations of marker gene and cell show strong deviations from the mean expression level. **f,** Across-gene Pearson correlation between log-transformed predicted and observed normoblast gene expression counts using an ablated model which was not trained on normoblast cells. Each bar corresponds to predictions done using the single-cell embeddings of cells of a different cell type. **g,** Mean-normalized observed and predicted gene expression of *HEMGN* along the diffusion pseudotime axis[22] representing erythropoietic differentiation. Both the full and the no-normoblast model accurately recapitulate the expression dynamics. Dots are colored by cell type, lines are smoothed with a rolling mean (window size: 200 cells).

TF motif effect scoring allows the investigation of TFs and target gene regulation

Given the central role of transcription factors (TFs) in regulating gene expression, we sought to identify TFs that drive lineage-specific gene expression predictions. To quantify the importance of TFs on gene expression, we introduced a TF motif effect score which measures the impact of mutating TF binding sites on predicted gene expression in single cells. Specifically, we focussed our analysis on 83 TFs that are significantly differentially expressed between cell types of the bone marrow dataset. For each TF, we used established position weight matrices[23] to map candidate TF binding sites located within 524 kb centered on the gene body of the 3,681 genes differentially expressed across the dataset (Fig. 3a, Methods). We mutated all matching sites in silico at the same time by substituting them with random sequences. Cell-level TF motif effect scores were defined as the log-fold-change of scooby's predicted expression between reference sequence and in silico mutated sequences, averaged across genes. TF motif effect scores are directional, being positive for activators and negative for repressors.

To assess the reliability of scooby's TF motif effect scores, we compared their agreement with observed cognate TF expression, as a proxy for true TF activity (Methods). We benchmarked against chromVAR[24] and scBasset[9], two widely used sequence-based methods for inferring TF activity from scATAC-seq data by comparing the correlation of their scores with TF expression. We found that scooby's TF motif effect scores correlate significantly better with gene expression than those of chromVAR (P = 9x10^{-8}, Wilcoxon two-sided, Fig. 3b) and scBasset (P = 0.04, Wilcoxon two-sided, Extended Data Fig. 7a). Remarkably, training scooby only with scRNA-seq data led to TF motif effect scores on par or better than the two alternative methods chromVAR and scBasset, which use scATAC-seq data (Fig. 3c, Extended Data Fig. 7b, Methods). This result indicates that scooby could alleviate the need for scATAC-seq data for the purpose of TF activity inference.

Having established TF motif effect scores, we next leveraged them to investigate the regulatory role of activating or repressing TF sequence elements on gene expression in a cell type-specific manner. We observed scooby to recapitulate the importance of known motifs for cell types of the main hematopoietic lineages (Figure 3d, Extended Data Fig. 8). For example, the GATA1 motif family exhibited the highest score in erythroblasts[25], the EBF1 motif in B1 B cells[26], the C/EBP motif family in monocytes[27], and the RUNX motif family in T cells[28]. The SOX motif family, containing a TF known to drive multipotent hematopoietic stem cells toward the B cell lineage (SOX4[29]), displayed the strongest effect in these cell types, and showed stronger activity in the early stages of each differentiation lineage. Furthermore, scooby captured early lineage

commitment within G/M progenitor cells, as their TF motif effect scores closely resembled those of both differentiated myeloid cells and the progenitor populations. Additionally, the model identified repressors of gene expression such as BACH2, a TF known to be repressive of the myeloid program in B cells[30]. However, we also observed that TFs with similar motifs are scored similarly (i.e., GATA1, TRPS1, and GATA3, Extended Data Fig. 8), which is a limitation of transcription factor binding site matching by motifs only. Despite this caveat, scooby's ability to distinguish lineage-specific patterns of TF activity suggests its potential for further exploration of the regulatory mechanisms underlying cell-fate decisions.

TFs can exert distinct effects on accessibility and expression due to temporal lags and repression mechanisms, some of which operate independently of chromatin accessibility changes. To support the investigation of motif effects on those two regulatory layers, we defined a TF motif effect score on the overall chromatin accessibility across a gene locus, analogously to the TF motif effect score on expression (Fig. 3a, Methods). Applied to GATA1, an established master regulator during erythropoiesis, in silico alteration of its binding sites indicated an early impact on chromatin accessibility across loci, whereas the effect on gene expression was delayed (Fig. 3e). This is consistent with the role of GATA1 as a pioneer factor and shows that scooby can be used to delineate the effect of motifs on accessibility from the effect on expression[31].

scooby allows exploration of cell-state-specific TF target genes

Identifying the genes regulated by a TF in a specific cell state is important to pinpoint the regulatory cascades driving cell fate decision and differentiation. With scooby, we can obtain TF target genes by investigating the TF motif effect on a per-gene basis. We applied scooby to explore putative target gene regulation by three key erythroid regulators - GATA1, TAL1, and KLF1. Target genes were defined as genes predicted to show differential expression in erythroid cell types upon cognate motif mutation (Fig. 3f, Methods). While direct validation of these predicted targets using experimental data like ChIP-seq is beyond the scope of this study, these putative target genes were overall enriched for Gene Ontology terms related to erythropoiesis (Heme Biosynthetic Process P=2.5x10^{-8}, Regulation of Erythrocyte differentiation P=5.7x10^{-6}, Methods), consistent with the known roles of GATA1, TAL1, and KLF1. As expected, the erythroid master regulator GATA1 was predicted to affect the largest number of genes. For TAL1, known to bind cooperatively with GATA1[32], the model recapitulated the connection to known target genes such as *HBB*, *SLC4A1* and *TRIM10*, as well as other erythroid regulators (*GATA1, KLF1*)[33]. Finally, we observed distinct clusters of genes controlled by combinations of GATA1, TAL1, and KLF1. For instance, cluster 6 (*RHAG*, *RHD*, *ALAS2*, *TFRC*, *TSPO2*, i.a.), enriched for iron ion

homeostasis and ammonium transmembrane transport GO terms, was shown to be impacted by both GATA1 and KLF1 deletion. These predictions are corroborated by previous reports in human and mice (as reviewed in Perkins et al.[34]). Altogether, our analysis demonstrates that scooby can be used to investigate the complex regulatory roles of lineage-specific transcription factors on target genes.

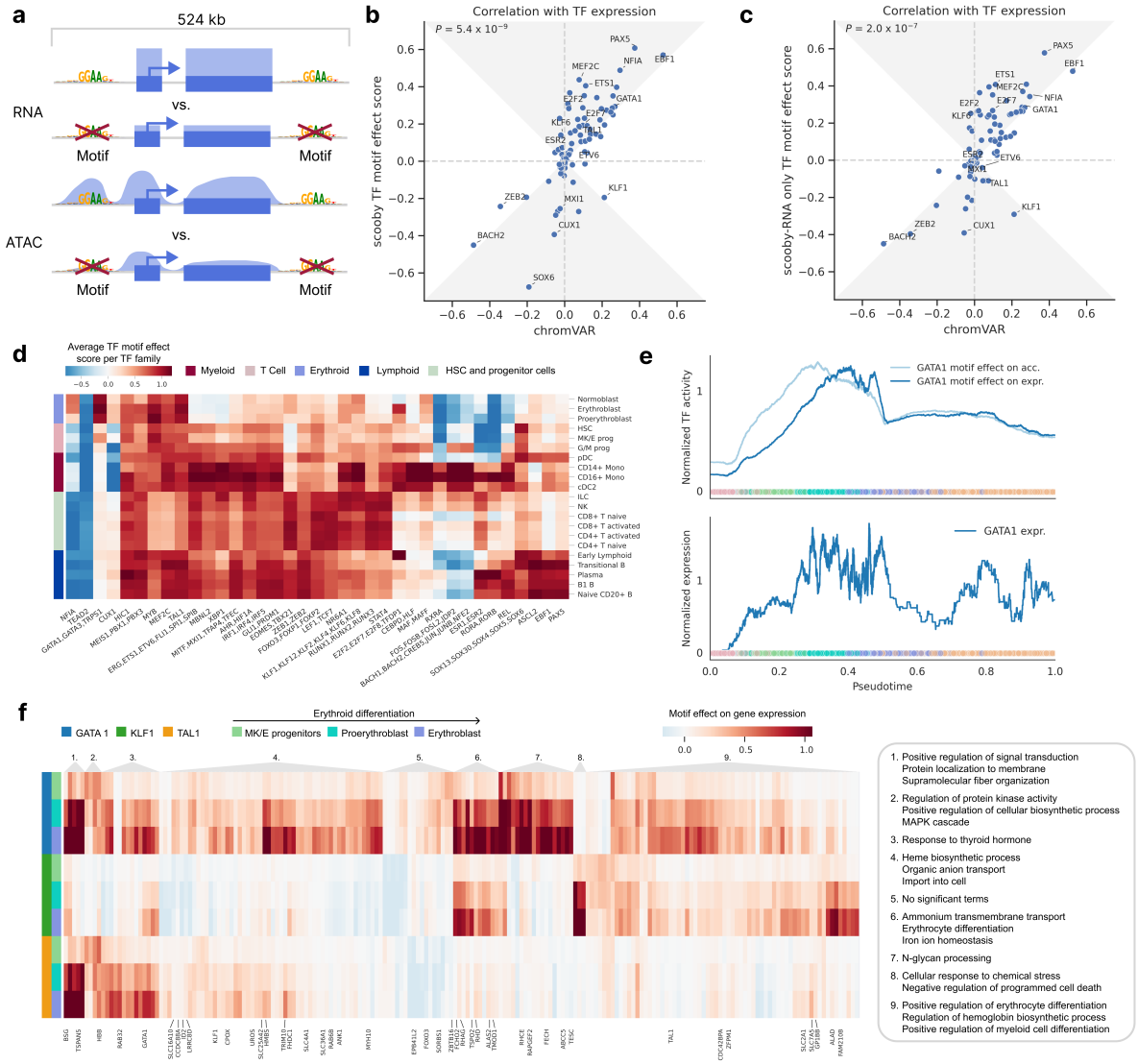

Figure 3: In silico motif mutation enables TF motif effect scoring and reveals lineage and gene-set-specific regulators. a, Schematic of TF motif effect scoring via in silico motif mutation. **b,** Pearson correlation of TF motif effect score with TF expression for scooby against chromVAR. The gray area marks the zone of improvement. **c,** Same as **b** for a scooby trained on scRNA-seq only. **d,** Heatmap of average TF motif effect score per TF family (columns) across cell types (rows). **e,** Median-normalized effect of GATA1 in silico motif mutation on accessibility and expression (top) and median-normalized expression of *GATA1* along the diffusion pseudotime axis representing erythropoietic differentiation (bottom). Dots are

colored by cell type, lines are smoothed with a rolling mean (window size: 200 cells). **f,** Heatmap of genes with high motif mutation effects of GATA1, KLF1 and TAL1 in the erythrocyte lineage. Genes are clustered according to their motif effect score. The three most significantly enriched GO terms for each cluster are shown. Gene names are only shown for genes ascribed to these terms.

scooby improves variant effect prediction of bulk eQTLs

Accurately predicting the regulatory impact of genetic variants on cell-state-specific gene expression remains a major challenge in genomics[35–37]. While sequence-based models including Borzoi and its predecessor Enformer have shown promise in distinguishing causal eQTLs from those in linkage disequilibrium[7,8], these analyses were limited to bulk data such as the tissue-specific eQTLs collected by the GTEx project[38]. Seq2cells recently demonstrated single-cell level variant effect prediction on gene expression and compared their results to GTEx v8 whole blood bulk eQTLs. However, concordance with statistically mapped eQTLs remains modest, with the seq2cells study reporting no effect for the vast majority of eQTLs and otherwise often disagreeing on the effect direction[12].

To assess scooby's capability for variant effect prediction and to compare to seq2cells, we considered fine-mapped GTEx whole blood eQTLs[39] both because many mature blood cell types originate from lineages present within the bone marrow dataset, and because this eQTL resource has also been used in the seq2cells and Borzoi studies.

We predicted bulk-level eQTL target gene expression for the reference and for the alternative eQTL allele by summing scooby's single-cell level gene expression on the natural scale. We defined the $\log_2$-ratio of the alternative compared to the reference prediction as the predicted variant effect (Methods). To directly compare with seq2cells, we used the model trained on the NeurIPS dataset and predicted variant effects for the subset of eQTLs located within both receptive fields of seq2cells and scooby (full set: n = 1,548; common set: n = 1,219). Overall, we observed a Spearman correlation of 0.495 between scooby's predicted fold-changes and the reported eQTL effects from GTEx on the full set (Fig. 4a), demonstrating that scooby generalizes to unseen variants. On the shared subset, scooby significantly improved upon seq2cells with a Spearman correlation of 0.517 ([0.474;0.556] 95% confidence interval), against 0.380 ([0.331;0.427] 95% confidence interval; Extended Data Fig. 9). In their original study, Schwessinger et al.[12] observed that seq2cells predicts a high proportion of variants to have negligible effects (with $\log_2$ effect less than 0.05, i.e. 3.5% fold-change). We recapitulated their observation both when training seq2cells on the NeurIPS dataset and with scooby (71.6%, absolute fold-change less than 3.5%, within red boundaries in Fig. 4a). However, on the remaining eQTLs, scooby achieved significantly higher sign agreement than seq2cells (91.7% vs. 83.9% on

the common subset, Fischer exact test *P*= 0.003, Extended Data Fig. 9). Notably, the fraction of concordant predictions with a predicted effect size greater than 3.5% exceeded 50% when the fine-mapped variant was located within 1 kb of the transcription start site and then declined with distance (Fig. 4b). This suggests that capturing the effects of distal regulatory elements, such as enhancers, remains a challenge for sequence-based models, consistent with previous findings[7,8,35].

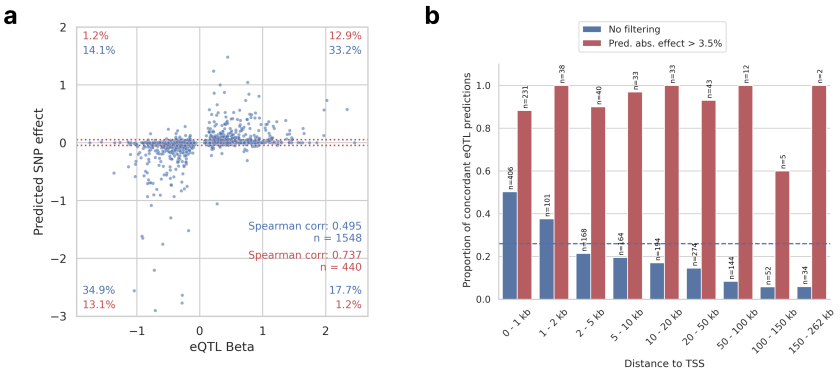

Figure 4: scooby predicted variant effects are concordant with reported eQTL effects. a, Predicted aggregated effects (log-fold change) vs observed whole-blood normalized eQTL effect sizes. Red dotted lines mark thresholds below which predicted fold-changes are deemed negligible (absolute fold change 3.5%; matching the threshold by Schwessinger et al.[12] for comparability). Percentages quantify variants within each quadrant: black - all variants; red - variants passing the 3.5% predicted effect threshold. **b,** Proportion of concordant eQTL predictions (non-negligible predicted aggregated effect and with the same direction as observed), as a function of distance to the transcription start site when filtering for non-negligible predicted effect (red) or without filtering (blue). Dashed blue line indicates the mean proportion of concordant eQTL predictions across all distances (0.26).

scooby allows cell-type-specific delineation of bulk eQTLs

Having demonstrated better variant effect prediction than seq2cells on bulk-tissue eQTLs, we then focussed on the subset of 370 concordant eQTL predictions, i.e. with a predicted aggregated absolute fold change greater than 3.5% and the same direction as observed. Leveraging scooby's single-cell resolution, we then investigated the regulatory effects of these variants at the finer granularity of individual cells and cell types. While predicted eQTL effects generally agreed across cell types, substantial variations were observed, reflecting relationships between cell types (Fig 5a). Notably, the erythroid, the monocyte, and the early progenitor cell types each showed distinct predicted eQTL effects. To explore the potential functional relevance of this cell-type specificity, we focussed on the 15% most-variable eQTLs previously associated with a human trait in genome-wide association studies (GWAS Catalog[11]; Figure 5b, Methods). We found several

cases where predicted cell-type-specific effects were consistent with the biology of the associated GWAS traits. For example, the *SLC14A1* eQTL, linked to "Immature fraction of reticulocytes" exhibited strong effects specifically within the erythroid lineage. Similarly, the *MIR34AHG* and *NDST1* eQTLs, both associated with "Monocyte count" showed pronounced effects in the monocytes. These findings suggest that scooby can provide insights into the cellular context of GWAS associations.

As an illustrative example of how scooby can reveal cell-type-specific regulatory mechanisms, we examined an eQTL (variant rs143664050) with a negative effect on Testin (*TES)* expression in CD14+ monocytes yet a negligible effect in erythroblasts (Fig. 5c). The alternative allele was associated with a predicted loss of an accessible region in monocytes, which could explain the observed reduction in *TES* expression. In contrast, no change in accessibility was predicted in erythroblasts, consistent with the negligible predicted effect on expression. Applying a gradient-based model interpretation method indicated that the eQTL disrupts a predicted binding site for the transcription factor SPI1 (Fig. 5c, bottom, Methods). Notably, *SPI1* is only expressed in myeloid cells, including monocytes, but not in erythroblasts (Fig. 5d), providing an explanation for the observed cell-type specificity of this eQTL's effect (Fig. 5e). In contrast, the variant rs62032983 provided an example of an eQTL predicted to reduce expression of the gene *DCTN5* across all cell types. Model interpretation attributed this effect to the disruption of a predicted binding site for the ubiquitously expressed transcription factor *ELF1* (Extended Data Fig. 10). Altogether, these analyses demonstrate scooby's ability to link cell-type specific variant effects, that would be missed in bulk-level analyses, to the underlying regulatory mechanism.

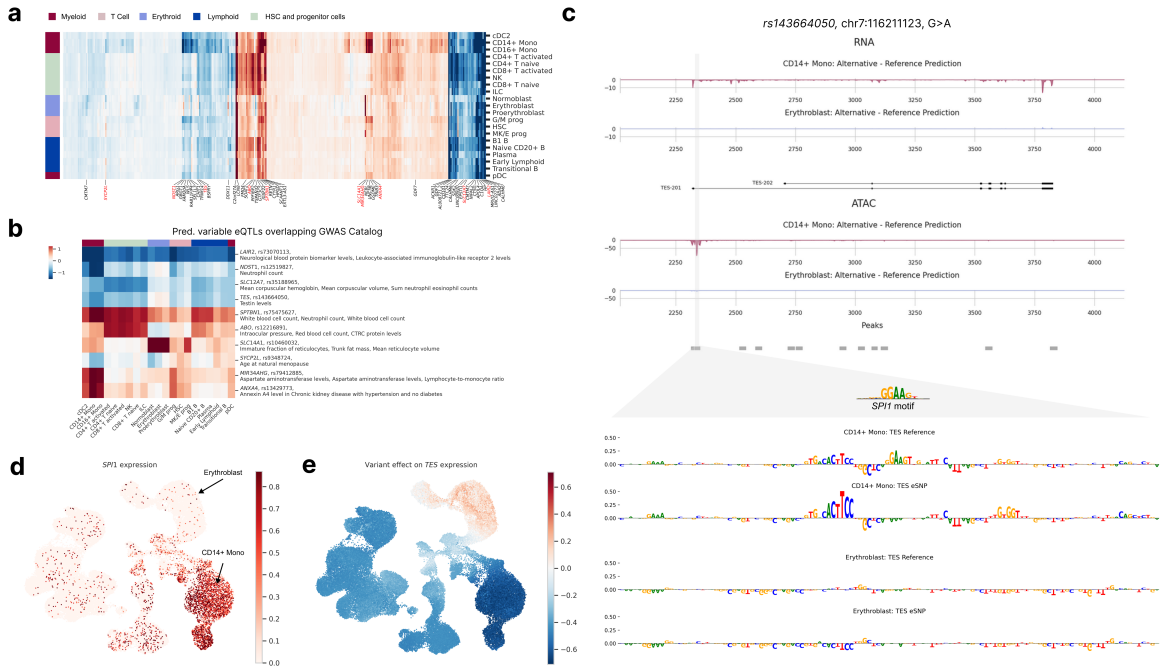

Figure 5: scooby allows cell-type-specific delineation of bulk eQTLs. a, Clustermap of eQTL effect size predictions across cell types. Left color bar indicates lineage membership. Genes were clustered according to their predicted effect size per cell type. Highlighted genes (and their fine-mapped eQTLs) have predicted variable effects (black) and were considered for an overlap with the GWAS Catalog. eQTL-GWAS term matches are colored in red. **b,** Heatmap of gene-variant pairs with strong cell-type specific effects and matching GWAS terms. **c,** Predicted fold change in gene expression (top) and accessibility (bottom) between the alternative and reference alleles of variant rs143664050 in CD14+ Monocytes and Erythroblasts. Sequence attributions revealed the destruction of a *SPI1* motif to only affect model outputs in CD14+ Monocytes (Methods). **d,** UMAP of the NeurIPS dataset colored by observed normalized *SPI1* expression levels. **e,** UMAP of variant rs143664050 effect on *TES* expression levels.

Discussion

This work introduced scooby, a model that enables the prediction of single-cell gene expression and chromatin accessibility profiles directly from half a megabase sequence context while scaling efficiently with the number of cells. This is achieved by equipping the pretrained multi-omics profile predictor Borzoi with a cell-specific decoder and fine-tuning its sequence embeddings. The model shows generalizability across cells and cell types and improved the state-of-the-art single-cell gene expression prediction from 0.77 to 0.87 Pearson correlation on unseen sequences. In silico motif mutations led to TF motif effect scores showing strong concordance with TF expression levels, improving upon state-of-the-art TF activity inference methods. Strong concordance also held for a scooby model when training exclusively on RNA-seq data, suggesting that it can

effectively leverage sequence information to infer TF motif effect scores without requiring matched accessibility data. Furthermore, scooby variant effect prediction showed high concordance with observed eQTL effect size and direction for established whole-blood eQTLs. Lastly, leveraging scooby's single-cell resolution together with interpretation methods allows for finer-grained analysis of variant effects, uncovering cell-type-specific eQTLs that are masked in bulk studies and the underlying transcription factors.

The architecture of scooby allows incorporating further modalities in two ways. Firstly, profile-based modeling is a generic approach that flexibly permits the prediction and interpretation of a wide range of additional single-cell modalities, such as methylation or ChIP-seq in contrast to genome annotation-based methods. Secondly, scooby can in principle work with any cell state representation. For instance, incorporating other data modalities such as CITE-seq into the embedding could allow for a finer resolution of cellular state.

We chose to evaluate scooby on a 10x multiome hematopoiesis dataset, as it provides both paired scATAC-seq and scRNA-seq profiles for joint modeling and well-characterized differentiation lineages for validating TF motif effects and target gene predictions, making it an ideal testbed for our study. As scooby predicts RNA-seq coverage, it could in principle be employed to predict differential isoform usage, however, the 10x scRNA-seq 3' coverage bias limits the signal for splice sites and TSS choice[15]. Furthermore, we observed little evidence for differential isoform usage in the dataset, concordant with studies finding that alternative transcript usage is most pronounced in brain and muscle tissues[40]. Thus, future work applying scooby to more diverse cell types or applications to alternative single-cell protocols such as SMART-seq[41] or long-read sequencing[42] is needed to assess the potential of scooby to model isoform-specific expression.

We obtained promising results when comparing scooby-based variant effect prediction with eQTLs. In particular, our model did not show frequent disagreement in effect directions, in contrast to reports on the bulk model Enformer[36,37] and our underlying model Borzoi[8]. Perhaps, fine-tuning models with single-cell data is one reason for this improvement, as the seq2cells study also reported moderate disagreement in directions. However, more extensive analyses using larger datasets and eQTL from more tissues are needed to generalize those results and understand the potential reasons for the improvements.

In the future, we envision scooby to aid interpretation of large single-cell atlases[43] as the framework efficiently scales to large numbers of cells. Moreover, we foresee its application to learn about conserved cell-type specific regulation by integrating diverse multi-species datasets. To facilitate adoption, we provide a streamlined workflow for applying scooby to new datasets.

In summary, scooby establishes a modeling paradigm for connecting single-cell genomics and sequence-to-function predictors. Its modular nature and ability to integrate multi-modal data and to capture cell-state-specific gene expression dynamics positions it as a valuable tool for uncovering the genetic basis of gene regulation and complex traits at single-cell resolution.

Methods

Data acquisition and processing

We obtained scRNA-seq and scATAC-seq data for the multiome hematopoiesis dataset used in the NeurIPS 2021 challenge[21]. Specifically, we downloaded scRNA-seq BAM files from the Sequence Read Archive (SRA) under accession SRP356158 and scATAC-seq fragment files from the AWS bucket s3://openproblems-bio/public/post_competition/multiome/. Pre-processed gene count and peak count matrices were retrieved from the Gene Expression Omnibus (GEO) under accession code GSE194122.

We performed all scRNA-seq data analyses using scanpy (v1.10)[44]. We utilized the predefined filtered cell and gene sets, as well as the highest-resolution cell type annotations (l2_cell_type key) provided in the original publication. We identified and removed doublet cell populations using Scrublet with default parameters[45]. Doublet calls were based on a threshold that primarily captured cells clustering in discrete locations on the Uniform Manifold Approximation and Projection (UMAP) embedding.

We normalized raw gene expression counts using the normalize_total function in scanpy and applied a log(expression + 1) transformation for downstream analyses. We inferred pseudotime trajectories on the integrated dataset using diffusion pseudotime[22] with default parameters.

Generation of cell embeddings

To prevent information leakage from the cell embeddings to the gene expression and accessibility prediction models, we identified and excluded genes and peaks present in the test and validation sets using pyRanges (v0.0.129)[46] prior to computing the embedding.

We further filtered out genes and peaks present in fewer than 1% of all cells to reduce dimensionality. We then used the MultiVI model from the scvi package (v1.1.2, https://github.com/lauradmartens/scvi-tools/tree/poissonmultivi)[17,47] to generate a unified embedding of both scRNA-seq and scATAC-seq data. Following previous work demonstrating improved performance[16], we adapted the model to utilize raw fragment counts for scATAC-seq data, modeling these using a Poisson distribution instead of binarized counts (Poisson-MultiVI).

Otherwise, we trained the model with default parameters, incorporating sample information as the batch key during data integration. This process generated an embedding vector of dimension 14 for each cell. The 100 nearest neighbors for the profile evaluations for each cell were computed using the scanpy function sc.pp.neighbors.

For the RNA-only model, we reran the embedding generation using only the RNA modality by running the scVI model with n_latent set to 14.

Efficient read coverage extraction for RNA-seq and ATAC-seq data

To generate the scRNA-seq and the scATAC-seq profiles used for training, we used an adapted version of SnapATAC2 (v2.6.1, https://github.com/lauradmartens/SnapATAC2)[18], an efficient Rust software package initially designed for ATAC-seq data processing. SnapATAC2 achieves efficient storage of scATAC-seq data within the AnnData[19] format by recording only the start position and length of each fragment and supports out-of-memory reading.

In contrast to scATAC-seq data, scRNA-seq aligned reads contain split reads due to RNA splicing. Therefore, we modified the code of sp.pp.make_fragment_file function of SnapATAC2 to parse and store split reads as multiple entries in the AnnData, with each entry representing a contiguous fragment and its corresponding length (see Extended Data Fig. 3).

Processing of RNA bam files and ATAC fragment files

Since the 10x BAM files included all reads (also low quality ones), we filtered for reads that were marked as valid by the cellRanger pipeline (xf:i:25 flag) using the subset-bam package (v1.1.0, https://github.com/lauradmartens/subset-bam). We then utilized our modified sp.pp.make_fragment_file function, specifying the appropriate barcode and UMI tags ("CB" and "UB") for our data to convert the reads in the BAM file into a fragment file. We removed duplicate reads using SnapATAC2's automatic read deduplication. For processing of scRNA-seq reads, we set the is_paired argument to False and both shift_left and shift_right arguments to 0.

We imported the processed fragment file into an AnnData object using sp.pp.import_data and the hg38 genome assembly. To ensure consistency with our scRNA-seq count data, we set min_num_fragments=0 and used the whitelist argument to retain only cells present in the pre-filtered scRNA-seq AnnData object.

We imported ATAC-seq fragment files using the sp.pp.import_data function with the parameters described above. To model Tn5 insertion sites, we converted fragment locations into insertion sites by recording the fragment ends.

Data preparation for training

We used the snapatac-processed coverage AnnData directly during training and created coverage tracks per cell on the fly. Following the procedure described for Borzoi[8], we first aggregated the coverage and insertions in a 32 bp window. For RNA profiles, we followed the same squashed-scale approach as Borzoi, but set clip_soft to 5, such that the fraction of soft-clipped values at single cell level is similar to the one of Borzoi's tracks. For ATAC profiles, we scaled the output by 0.05 to ensure they were on the same scale as the RNA coverage tracks. For faster evaluation on pseudobulks, we exported read aggregates per cell type to the bigWig format using sp.ex.export_coverage.

Model

scooby builds upon Borzoi[8], the first deep learning model for predicting RNA-seq profiles, which operates at 32 bp resolution on 524,288 bp of DNA sequence, outputting profiles for the center 6,144 bins (corresponding to 196,608 bp). We adapted a publicly available PyTorch implementation of Borzoi (https://github.com/johahi/borzoi-pytorch), removing the original human and mouse-specific output heads and retaining the convolutional and transformer layers responsible for sequence encoding. We integrated LoRA modules[14], each with a rank of 8, into all convolutional layers and the query, value, and MLP projection matrices within the transformer layers using an adapted version of peft (v0.10.1, https://github.com/lauradmartens/peft)[48]. While used separately during training, these LoRA modules were merged back into the original model weights after training, resulting in no additional overhead during inference.

Furthermore, we introduced a trainable layer with GELU nonlinearity on top of Borzoi's penultimate layer. The weights of this layer were randomly initialized and trained from scratch, allowing for potential refinement of Borzoi's embeddings for the single-cell context. The output of this layer was then passed to the cell-state-specific decoder. This decoder operates on the sequence embeddings of the center 6,144 bins. Specifically, the decoder consists of a 1x1 convolution along the sequence dimension (effectively a position-wise linear transformation). The weights of this convolutional filter were not fixed but were dynamically generated for each cell based on its corresponding cell embedding. To this end, a small multilayer perceptron (MLP) was added that receives the cell embedding as input and outputs a vector to parameterize the convolutional filter used to produce the final predicted profiles for that cell from the sequence embedding. We visualize the exact model architecture in Extended Data Fig. 1. For stranded RNA predictions, the MLP outputs a weight matrix of shape (1921, 2, 1), encompassing the filter

weights (1920 dimensions) and biases (1 dimension) for each strand. For ATAC predictions (unstranded), it outputs a weight matrix of shape (1921, 1, 1).

To efficiently scale scooby to a large number of cells, we implemented two optimizations. First, we introduced a caching mechanism for the sequence embeddings, reducing redundant computations when predicting profiles for multiple cells from the same genomic region. Second, we performed cell-state-specific decoding for expression only on the embedding slices that overlap with exons (or gene body) of interest.

Training Procedure

We initialized scooby's Borzoi backbone with pre-trained weights from Borzoi's replicate 0 (test fold 3, validation fold 4), converted from the original TensorFlow implementation to PyTorch. These pre-trained weights correspond to a model trained on the human and mouse reference genome (hg38, mm10 assembly). We maintained the same train-val-test split as Borzoi for scooby's training. During training, we only updated the parameters of the LoRA modules, the cell-state-specific convolutional filter weights, and the weights of the additional layer with GELU nonlinearity introduced after Borzoi's penultimate layer. With PyTorch (v2.1.0), we used the AdamW optimizer with a learning rate of 4×10^{-4} for the cell-state convolutional layer and 2×10^{-4} for the LoRA modules and the additional layer to stabilize training. Both learning rates were warmed up over the first 1,000 steps and decayed linearly afterward over 40 epochs. To stabilize training during the first step, we froze the batch normalization layers from the pre-trained Borzoi model and disabled dropout within Borzoi. We then unfroze the batch normalization layers and enabled dropout to prevent overfitting.

We monitored validation performance using the Pearson correlation between the predicted and observed total counts (c.f. Gene count evaluation, $\log_2$-transformed pseudobulk counts with an added pseudocount of 1) and retained the model with the largest correlation across cell types.

All models were trained with a batch size of 8 sequences across 8 A40 GPUs, using mixed precision to accelerate training. Per training sequence, the model predicted RNA and ATAC profiles for 64 randomly sampled cells in a multi-task learning fashion. We randomly augmented training sequences by shifting them by up to 3 base pairs in either direction and reverse-complementing them. To ensure consistent strand orientation, we reverse-flipped the target profiles when training on reverse-complemented sequences. We used the same weighting scheme for the Poisson and multinomial loss terms as in the original Borzoi implementation. The gradient clipping threshold was set to 1.0, and weight decay was set to 10^{-6} for all trainable parameters. Due to computational constraints, extensive hyperparameter optimization was not performed.

Ablations and other models

To train the model without normoblasts, we used the same hyperparameters, cell embeddings, and training procedure as described above. However, normoblast cells were excluded from the random sampling of cells during training. To train the RNA-only ablation, we followed the above steps, but instead removed scATAC-seq targets and output heads and used an embedding based on RNA-seq only. For the model without LoRA, we removed the trainable LoRA weights while maintaining the model architecture elsewhere.

We downloaded seq2cells from the official github repository (https://github.com/GSK-AI/seq2cells)[12] and processed files to fit the required format. For comparability, we used a cell x gene matrix with counts generated by summing over the observed profile. We matched gene IDs to gene names, retaining 15,892 genes, and followed the same data split as Enformer[7], upon which seq2cells is based. We trained the model using the training configuration provided in the repository, but longer for up to 40 epochs to stay comparable to scooby, and evaluated the best checkpoint.

Inference

To obtain model predictions, we performed inference on both the input sequence and its reverse complement. The output tracks for the reverse complement were reverse-flipped, and both predictions were averaged to produce the final profiles. We then reversed the squashed scale transformation to obtain raw expression values, and scaled accessibility profiles by 20 to reverse the transformation applied during training. We used mixed-precision inference to accelerate computations.

Profile level evaluation

To quantitatively assess profile prediction accuracy, we calculated the Pearson correlation on a logarithmic scale between the predicted profiles and the observed profiles averaged across 100 cell neighbors over all test sequences for a random subset of cells of each cell type. Genomic annotations were plotted using trackplot[49].

Gene count evaluation

To generate cell-type-specific counts, we first centered the input sequence on the gene body as annotated with GENCODE release v32[50]. We summed the expression profile for each bin overlapping an exon using the output track that matched the strand of the gene. This was repeated for each cell and summed across all cells of the same cell type to obtain pseudobulk counts. Finally, we $\log_2$-transformed both the predicted and target pseudobulk counts, and added a pseudocount of 1 to calculate the Pearson correlation across genes for each cell type. To obtain

a metric quantifying how well the model captures cell-type-specific expression, we subtracted the mean across genes of the cell-type x gene $\log_2$-transformed expression matrix, and then subtracted the mean across cells for both predictions and observations and correlated the results. For the no-normoblast ablation, we predicted all test gene counts pseudobulked for each cell type as described above and correlated their predicted expression with the observation of the normoblast cell type. For the comparison with seq2cells, we only retained genes overlapping the Enformer test set (used in seq2cells) and the Borzoi test set (used in this study). Predicted and observed gene counts along the diffusion pseudotime axis were smoothed using a rolling window of 200 cells with mean aggregation.

Motif deletion experiments

To investigate the impact of transcription factor (TF) binding sites on scooby's predictions, we performed in silico motif deletion experiments. We first obtained a list of TF position weight matrices (PWMs) from the HOCOMOCO v12 core database (https://hocomoco12.autosome.org/downloads_v12)[23]. To focus on TFs with potential regulatory roles in the relevant cell types, we filtered the list for TFs overlapping the set of differentially expressed genes per lineage. Differential gene expression analysis was conducted using the Wilcoxon rank-sum test, with a significance threshold of $P \leq 0.05$ after correcting for multiple testing with the Benjamini-Hochberg procedure using the scanpy function sc.tl.rank_gene_groups with groupby="l1_cell_type". We selected the motif with the most evidence and lowest motif subtype for each TF, drawing randomly if multiple candidates exist.

We used tangermeme FIMO (v0.2.3)[51] to scan input sequences centered on differentially expressed genes, identifying putative TF binding sites based on their PWMs using default significance cutoffs of 0.0001 after converting pwm-matching log-odds scores into p-values. We generated alternative sequences by substituting each predicted binding site with a random nucleotide sequence of the same length, repeating this procedure 10 times per sequence to mitigate spurious motif introduction.

We used scooby to predict gene expression and accessibility profiles for both the original and motif-deleted sequences in each cell. For gene expression, we summed the predicted RNA-seq coverage over all exons within a gene. For accessibility, we summed the predicted scores across the entire 6,144 x 32bp output bins. We averaged over the 10 distinct random replacements. The prediction for each cell was divided by the size factor of its corresponding reference prediction and scaled by its median reference size factor. For each cell, we calculated the mean $\log_2$ fold-

change between the reference and alternative sequence predictions across differentially expressed genes, yielding a single TF score per cell.

For the chromVAR comparison, we used pychromvar (v0.0.4)[24] with default configuration, except that we used the PWMs of the HOCOMOVO v12 core database for consistency. For scBasset[52], we followed the scVI tutorial (https://docs.scvi-tools.org/en/stable/tutorials/notebooks/atac/scbasset.html) to obtain TF activity scores.

To compare TF scores and target TF expression, we then computed the Pearson correlation of each TF score (from scooby, chromVAR and scBasset) with log-normalized TF expression.

eQTL analysis

We evaluated the ability of scooby to pinpoint likely causal nucleotide variants driving gene expression changes within fine-mapped expression quantitative trait loci (eQTLs) from GTEx v8[38]. We used the publicly available summary statistics and fine-mapping results generated with SuSiE (https://www.finucanelab.org/data)[53]. Following Linder et al.[8], we used single nucleotide variants with a posterior causal probability (PP) ≥0.9 and filtered out all non-single-nucleotide variants such as indels and deletions.

For each variant, we centered the input sequence on the identified eQTL target gene and recorded the effect on gene expression (sum over exons) and accessibility (sum over the whole region) for the reference and the alternative nucleotide of the true eSNP. We then computed cell-type level variant effects by summing the predicted gene expression levels on the natural scale over all cells of that cell type and computing the $\log_2$-fold change of alternative versus reference predictions, adding a pseudocount of 1. For a general variant effect, we summed the predicted expression levels over all cells.

To identify eQTLs with cell-type-specific effects, we focused on variants demonstrating both a strong overall effect and substantial variability across cell types. We first filtered for variants with a substantial average effect size ($\log_2$(mean effect) > 0.05) across all cell types. We labeled eQTLs as cell-type-specific if their effect sizes showed high variability across cell types, specifically within the top 15% of variance. We clustered the variants on their effect size using seaborn clustermap with metric = "seuclidean" and method = “ward”. We downloaded v1.0 of the GWAS Catalog from https://www.ebi.ac.uk/gwas/docs/file-downloads and matched GTEx variants using rs_id_dbSNP151_GRCh38p7, manually adding matches where possible using

dbsnp[54]. We report up to three randomly selected terms matching variant and target gene combination.

To link eSNPs to TF motifs, we generated gradient-weighted position weight matrices (PWMs) encompassing a 10-bp window centered on each variant. Specifically, we derived the PWM of the 10-bp window by exponentiating 2 with the gradient of each nucleotide per position, and subsequently sum-normalized each position. We then used Tomtom to scan these PWMs against the HOCOMOCO v12 core database using the default MEME parameters[55]. To account for potential redundancy among TFs with similar motifs, we filtered the results to include only TFs expressed in at least 1% of all cells. Since TF nomenclature can vary between databases, we used a python interface to Uniprot (Unipressed v1.3.0) to map protein names from HOCOMOCO to their corresponding gene symbols.

Writing

A large language model (LLM) was used to assist with refining the phrasing and clarity of the manuscript. All suggestions generated by the LLM were carefully reviewed and edited by the authors.

Data availability

The scRNA-seq, scATAC-seq, and pre-processed count matrices for the multiome hematopoiesis dataset are available from the NeurIPS 2021 challenge[21], SRA (accession SRP356158), AWS (s3://openproblems-bio/public/post_competition/multiome/), and GEO (accession GSE194122). TF position weight matrices were obtained from HOCOMOCO v12 (https://hocomoco12.autosome.org/downloads_v12). GTEx eQTL summary statistics and fine-mapping results are available from https://www.finucanelab.org/data. The GWAS Catalog (v1.0) was downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads.

Code availability

The scooby model including training scripts and data loaders are available at https://github.com/gagneurlab/scooby. Jupyter notebooks and scripts to reproduce our analysis and figures are available at https://github.com/gagneurlab/scooby_reproducibility. The adapted version of SnapATAC2 is available at https://github.com/lauradmartens/SnapATAC2.

References

1. Sasse, A., Chikina, M. & Mostafavi, S. Unlocking gene regulation with sequence-to-function models. *Nat. Methods* **21**, 1374–1377 (2024).
2. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
3. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **31**, (2020).
4. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931–934 (2015).
5. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
6. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
7. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
8. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. 2023.08.30.555582 Preprint at https://doi.org/10.1101/2023.08.30.555582 (2023).
9. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088–1096 (2022).
10. Michielsen, L., Reinders, M. J. T. & Mahfouz, A. Predicting cell population-specific gene expression from genomic sequence. *Front. Bioinforma.* **4**, (2024).
11. Trevino, A. E. *et al.* Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053-5069.e23 (2021).
12. Schwessinger, R., Deasy, J., Woodruff, R. T., Young, S. & Branson, K. M. Single-cell gene

expression prediction from DNA sequence at large contexts. 2023.07.26.550634 Preprint at https://doi.org/10.1101/2023.07.26.550634 (2023).

13. Li, J. *et al.* Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. *Nat. Genet.* **54**, 1711–1720 (2022).
14. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).
15. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58**, 610–620 (2015).
16. Martens, L. D., Fischer, D. S., Yépez, V. A., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat. Methods* **21**, 28–31 (2024).
17. Ashuach, T. *et al.* MultiVI: deep generative model for the integration of multimodal data. *Nat. Methods* **20**, 1222–1231 (2023).
18. Zhang, K., Zemke, N. R., Armand, E. J. & Ren, B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat. Methods* **21**, 217–227 (2024).
19. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Annotated data. 2021.12.16.473007 Preprint at https://doi.org/10.1101/2021.12.16.473007 (2021).
20. Virshup, I. *et al.* The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.* **41**, 604–606 (2023).
21. Luecken, M. D. *et al.* A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. in (2021).
22. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).
23. Vorontsov, I. E. *et al.* HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* **52**, D154–D163 (2024).
24. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring

transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

25. Cheng, Y. *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Res.* **19**, 2172–2184 (2009).
26. Györy, I. *et al.* Transcription factor Ebf1 regulates differentiation stage-specific signaling, proliferation, and survival of B cells. *Genes Dev.* **26**, 668–682 (2012).
27. Friedman, A. D. Transcriptional regulation of granulocyte and monocyte development. *Oncogene* **21**, 3377–3390 (2002).
28. Collins, A., Littman, D. R. & Taniuchi, I. RUNX proteins in transcription factor networks that regulate T-cell lineage choice. *Nat. Rev. Immunol.* **9**, 106–115 (2009).
29. Laurenti, E. *et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nat. Immunol.* **14**, 756–763 (2013).
30. Itoh-Nakadai, A. *et al.* The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program. *Nat. Immunol.* **15**, 1171–1180 (2014).
31. Ludwig, L. S. *et al.* Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis. *Cell Rep.* **27**, 3228-3240.e7 (2019).
32. Han, G. C. *et al.* Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. *Mol. Cell. Biol.* **36**, 157–172 (2016).
33. Kassouf, M. T. *et al.* Genome-wide identification of TAL1's functional targets: Insights into its mechanisms of action in primary erythroid cells. *Genome Res.* **20**, 1064–1083 (2010).
34. Perkins, A. *et al.* Krüppeling erythropoiesis: an unexpected broad spectrum of human red blood cell disorders due to KLF1 variants. *Blood* **127**, 1856–1862 (2016).
35. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**, 56 (2023).

36. Sasse, A. *et al.* Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).
37. Huang, C. *et al.* Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat. Genet.* **55**, 2056–2059 (2023).
38. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
39. Wang, Q. S. *et al.* Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
40. Ling, J. P. *et al.* ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nat. Commun.* **11**, 137 (2020).
41. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
42. Joglekar, A. *et al.* Single-cell long-read sequencing-based mapping reveals specialized splicing patterns in developing and adult mouse and human brain. *Nat. Neurosci.* **27**, 1051–1063 (2024).
43. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).
44. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
45. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).
46. Stovner, E. B. & Sætrom, P. PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics* **36**, 918–919 (2020).
47. Gayoso, A. *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).

48. Mangrulkar, S. *et al.* PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. (2022).
49. Zhang, Y., Zhou, R., Liu, L., Chen, L. & Wang, Y. Trackplot: A flexible toolkit for combinatorial analysis of genomic data. *PLOS Comput. Biol.* **19**, e1011477 (2023).
50. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
51. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
52. Yuan, H. & Kelley, D. R. *scBasset: Sequence-Based Modeling of Single Cell ATAC-Seq Using Convolutional Neural Networks*. 2021.09.08.459495 https://www.biorxiv.org/content/10.1101/2021.09.08.459495v1 (2021) doi:10.1101/2021.09.08.459495.
53. Wang, Q. S. *et al.* Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
54. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
55. Bailey, T. L. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).

Acknowledgements

We thank Jan Engelmann for insightful discussions regarding variant effect prediction, Ruochi Zhang for computational advice as well as Thomas Mauermeier for his talented cluster support. We thank Pedro Tomaz da Silva and Nils Wagner for comments on the manuscript. L.D.M. acknowledges support by the Helmholtz Association under the joint research school Munich School for Data Science. L.D.M and J.G. acknowledge support from the Deutsche

Forschungsgemeinschaft (SFB/TransRegio TRR267, Project-ID 403584255) and J.C.H and J.G from the European Research Council (ERC) (EPIC, Grant number: 101118521). F.J.T. acknowledges support by the European Research Council (ERC) (DeepCell 101054957). The views and opinions expressed are those of the authors and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. J.D.B. acknowledges support from the Gene Regulation Observatory at the Broad Institute of MIT & Harvard, NIH 5P01HL131477, the NHGRI IGVF consortium (UM1 HG011986) and the NIH New Innovator Award (DP2 HL151353).

Author contributions

L.D.M. and J.C.H. conceived the project with J.G. J.C.H. designed the model. L.D.M. and J.C.H. implemented the framework. L.D.M. and T.M. developed the data pipeline. L.D.M. and J.C.H. designed the evaluations and analyzed the data with help from A.K. and J.D.B.. J.G., J.D.B. and F.J.T. supervised the project. J.C.H, L.D.M, and J.G. wrote the manuscript with input from all authors.

Competing interests

J.D.B. holds patents related to ATAC-seq and is an SAB member of Camp4 and seqWell. F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd and Omniscope Ltd, and has ownership interest in Dermagnostix GmbH and Cellarity.