maxATAC: genome-scale transcription-factor binding prediction from ATAC-seq with deep neural networks

Tareian A. Cazares[1], Faiz W. Rizvi[2], Balaji Iyer[3,4], Xiaoting Chen[5], Michael Kotliar[6], Anthony T. Bejjani[7], Joseph A. Wayman[8], Omer Donmez[5], Benjamin Wronowski[6], Sreeja Parameswaran[5], Leah C. Kottyan[5,9], Artem Barski[6,9,10], Matthew T. Weirauch[3,5,9,11], VB Surya Prasath[3,9,4], Emily R. Miraldi[3,4,8,9]*

[1]Immunology Graduate Program, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA.
[2]Department of Pharmacology and Systems Biology Graduate Program, University of Cincinnati College of Medicine, Cincinnati, OH, 45229, USA.
[3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
[4]Department of Electrical Engineering and Computer Science, University of Cincinnati, OH 45221, USA
[5]The Center for Autoimmune Genetics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
[6]Division of Allergy and Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
[7]Division of Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
[8]Molecular and Developmental Biology Graduate Program, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA.
[9]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA.
[10]Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
[11]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, 45229, USA.
*Correspondence to: emily.miraldi@cchmc.org

Short Title:
Transcription factor binding prediction with maxATAC

Key Words:
Transcription factor; gene regulation; genomics; ATAC-seq; scATAC-seq; deep neural networks, trans-cell type prediction

Abstract:

Transcription factors read the genome, fundamentally connecting DNA sequence to gene expression across diverse cell types. Determining how, where, and when TFs bind chromatin will advance our understanding of gene regulatory networks and cellular behavior. The 2017 ENCODE-DREAM *in vivo* Transcription-Factor Binding Site (**TFBS**) Prediction Challenge highlighted the value of chromatin accessibility data to TFBS prediction, establishing state-of-the-art methods for TFBS prediction from DNase-seq. However, the more recent Assay-for-Transposase-Accessible-Chromatin (ATAC)-seq has surpassed DNase-seq as the most widely-used chromatin accessibility profiling method. Furthermore, ATAC-seq is the only such technique available at single-cell resolution from standard commercial platforms. While ATAC-seq datasets grow exponentially, suboptimal motif scanning is unfortunately the most common method for TFBS prediction from ATAC-seq. To enable community access to state-of-the-art TFBS prediction from ATAC-seq, we (1) curated an extensive benchmark dataset (127 TFs) for ATAC-seq model training and (2) built "**maxATAC**", a suite of user-friendly, deep neural network models for genome-wide TFBS prediction from ATAC-seq in any cell type. With models available for 127 human TFs, maxATAC is the first collection of high-performance TFBS prediction models for ATAC-seq. maxATAC performance extends to primary cells and single-cell ATAC-seq, enabling improved TFBS prediction *in vivo*. We demonstrate maxATAC's capabilities by identifying TFBS associated with allele-dependent chromatin accessibility at atopic dermatitis genetic risk loci.

Author Summary

Proteins called transcription factors interpret the genome, reading both DNA sequence and chromatin state, to orchestrate gene expression across the diversity of human cell types. In any given cell type, most chromatin is “inaccessible”, and only those parts of the genetic code needed or likely to be needed soon are “accessible” for transcription factor binding to affect gene expression and cellular behavior. Hundreds of transcription factors are expressed in a given cell type and context (e.g., age, disease), and knowledge of their context-specific DNA binding sites is key to uncovering how transcription factors regulate cellular behaviors in health or disease. However, experimentally profiling the >1,600 human transcription factors across all cell types and contexts is infeasible. We built a suite of computational models “**maxATAC**” to predict transcription factor binding from a measurement of accessible chromatin, **ATAC-seq**. Importantly, ATAC-seq is feasible even at single-cell resolution. Thus, this data type, in combination with maxATAC, can be used to infer transcription factor binding sites in directly-relevant cell types isolated from physiological and disease settings, enabling insights into disease mechanisms, including how genetic variants and cellular context impact transcription factor binding, gene expression patterns and disease risk.

Introduction:

Most disease-associated genetic polymorphisms fall outside of protein-coding sequences [1]. Instead, they overlap significantly with enhancers, promoters and other locus-control regions [2]. Causal variants are thought to contribute to disease phenotypes by altering gene transcription in specific cell types [3,4]. Gene regulatory networks (**GRNs**) describe the control of gene expression by transcription factors (**TFs**) at genome-scale [5]. GRN reconstruction for human cell types will thus be crucial to identifying how noncoding genetic variants contribute to complex phenotypes through altered TF binding, chromatin looping and other gene regulatory mechanisms.

The Assay for Transposase Accessible Chromatin (**ATAC-seq**) opens new opportunities for GRN inference and genetics. Relative to other assays for chromatin accessibility (e.g., FAIRE-seq[6], DNase-seq[7]), ATAC-seq is an easy-to-use, popular technique that provides high-resolution chromatin accessibility with low sample input requirements[8]. Thanks to advances in single-cell (**sc**)ATAC-seq, it is now possible to resolve the chromatin accessibility profiles of individual cell types from heterogeneous tissues and limited clinical samples [9–11]. ATAC-seq is the only single-cell chromatin accessibility assay available via standard commercial platforms [12]. Whether from single cells or "bulk" populations, integration of TF-binding predictions from ATAC-seq improves GRN inference [13,14]. Although other experimental approaches more directly measure TF occupancy (e.g., ChIP-seq), they require substantial optimization, are costly in time and reagents, and are sometimes impossible due to a lack of quality antibodies. Indeed, hundreds of TFs are expressed in a given cell-type condition, but profiling of >50 TFs has been accomplished for very few human cell types [15]. Furthermore, for some rare cell types and physiological conditions, limited sample material precludes direct measurement of TF occupancies.

Thus, the computational community collectively pioneered methods to predict TF binding sites (**TFBS**) from chromatin accessibility [16]. In 2017, the ENCODE-DREAM *in vivo* TFBS Prediction Challenge established two top-performing TFBS prediction algorithms [17,18] that vastly improved performance over popular motif scanning (median area under precision-recall .4 versus .1). Subsequent efforts (Factornet[19], Leopard[20] and DeepGRN[21]) leveraged deep neural network modeling, yielding further advances in TFBS prediction from chromatin accessibility. Yet these top-performing TFBS methods are not used in popular ATAC-seq analysis pipelines [21–25]. Instead, TF motif scanning in accessible chromatin regions is the norm, and this less accurate method for TFBS prediction ultimately undermines biological inferences (TF enrichment, GRN and genetic analyses), experimental validation rates and overall scientific advances from ATAC-seq and scATAC-seq data.

Several factors limit community access to state-of-the-art TFBS prediction from ATAC-seq:

(1) Current state-of-the-art models for TFBS prediction from chromatin accessibility were trained on DNase-seq data [17–21]. Although both DNase-seq and ATAC-seq provide high-resolution accessibility data, there are notable differences between the technologies [22,23]. Furthermore, the ENCODE-DREAM Challenge DNase-seq training data are single-end sequenced, while paired-end sequencing, which improves mappability, is the standard for modern ATAC-seq datasets [8,12]. Thus, it is risky to assume that the DNase-seq-trained models will perform well on ATAC-seq inputs.

(2) The top-performing DNase-seq TFBS models exist for at most 32 TFs [17–21]. In contrast, TF motifs are available for at least 1200 of the ~1600 human TFs [24]. Thus, exceedingly few biologists are enticed to integrate the small number of state-of-the-art TFBS models into their ATAC-seq analysis pipelines.

(3) While benchmark and training datasets have been developed for TFBS modeling from DNase-seq, such resources have not been curated, quality-controlled and organized for TFBS modeling from the more popular ATAC-seq assay.

Despite the promise of ATAC-seq for GRN inference and human genetics, TFBS prediction from ATAC-seq remains primitive, even as ATAC-seq data generation grows exponentially in both basic and clinical research.

To enable state-of-the-art TFBS prediction from ATAC-seq, we built "**maxATAC**", a collection of user-friendly deep neural networks models for genome-scale TFBS prediction from ATAC-seq (**Fig. 1**). maxATAC currently includes models for 127 human TFs, making it the largest collection of high-performance TFBS models available. This effort required extensive curation of existing ChIP-seq and ATAC-seq datasets as well as select data generation. Benchmarking led to methodological advances for TFBS prediction from ATAC-seq. As a result, our models perform well on both bulk and single-cell ATAC-seq, expanding maxATAC TFBS capabilities to scarce cell types *in vivo*. We use maxATAC to discover TFBS associated with allele-dependent chromatin accessibility at atopic dermatitis genetic risk variants, showcasing the potential for maxATAC to uncover molecular regulatory mechanisms in complex diseases.

Results:

An unprecedented benchmarking resource for cross-cell type TFBS prediction from ATAC-seq

Development and community adoption of state-of-the-art methods for cross-cell type TFBS prediction from ATAC-seq were the objectives motivating our study. Construction of a new benchmark resource was a necessary first step. We needed a benchmark dataset that (1) enabled model training on ATAC-seq and (2) would yield a compendium of TF models large enough to inspire wide-scale community adoption of state-of-the-art methods. Specifically, these needs were not met by the popular DREAM-ENCODE *in vivo* TFBS Prediction Challenge benchmark, which is limited to DNase-seq and only 12 TFs (see **Introduction**).

Cross-cell type TFBS prediction greatly benefits from training in multiple cell types[18]. Thus, for benchmarking, we sought paired TF-binding (e.g., ChIP-seq) and ATAC-seq data in at least three cell types for benchmarking (at least 2 cell types for training and 1 to test generalizability in a new cell type, **Methods**). Limiting to human cell types, we identified >18,000 existing ChIP-seq and ATAC-seq experiments annotated by CistromeDB [25] and ENCODE [15]. Because experimental perturbations can impact TF binding and chromatin accessibility, it was critical that the TF ChIP-seq and ATAC-seq were derived from the same experimental conditions. We excluded data from any perturbed experimental condition (e.g., growth hormone or drug stimulation, genetic modifications, including ectopic TF expression, see **Methods**). To reduce technical variability in our ATAC-seq training data, we restricted to the OMNI-ATAC-seq protocol[26]. (Relative to the original "standard" ATAC protocol[8], OMNI-ATAC-seq has improved signal-to-noise ratio, better recovery of enhancers and reduced mitochondrial contamination.) For improved mappability, we required paired-end sequencing. We manually verified the annotations of each experiment, excluding 2,887 experiments that had experimental perturbation or were incorrectly annotated (**Table S1, Methods**). Finally, we processed data, eliminating experiments that failed our quality-control metrics (**Methods**).

Data curation and quality-control of public data yielded an initial dataset of 361 TF-cell type combinations (110 TFs with at least 2 training examples across 17 cell types). Furthermore, we identified 3 cell lines with abundant TF ChIP-seq but no OMNI-ATAC-seq (HepG2, HEK293, LoVo). To expand the benchmark further, we generated OMNI-ATAC-seq for these cell lines, bringing our final maxATAC benchmark dataset to 438 unique TF-cell type pairs (**Fig. S1A**).

In total, the benchmark spans 463 ChIP-seq and 55 ATAC-seq experiments across 20 cell types, facilitating benchmarking of 74 TF models (≥3 cell types) and model construction for 127 TFs (≥2 cell types, **Fig. 1B**). The 127 maxATAC TFs span 35 TF families, with up to 26 TFs represented per family (**Fig. S1B**). This unique, expansive resource enables rigorous benchmarking of cross-cell type TFBS prediction from ATAC-seq. The underlying quality-controlled, processed datasets are organized and available for community development of ATAC-seq-based TFBS prediction methods (see **Data Availability**).

maxATAC architecture and "peak-centric, pan-cell" training approach

Deep convolutional neural networks (**CNNs**) provide top performance for many sequenced-based prediction tasks, including prediction of TFBS [27–29]. They require no prior knowledge of TF motifs and instead learn complex patterns in input DNA sequence (nonlinear combinations of what often look like TF motifs) de novo. We thus chose deep CNNs to model TFBS from ATAC-seq and DNA sequence (**Fig. 1A, S2**, **Methods**). We utilize dilated convolutional layers to capture spatially distant relationships across the input sequences in a multiscale manner [29,30]. This approach enables both high-resolution TFBS predictions (32bp) and information-sharing between proximal sequence and accessibility signals (+/-512bp).

Given our goal to construct TF models for as many TFs as possible, we developed training approaches to improve performance from minimal data (e.g., only 2-3 training cell types). We built a training strategy to enrich for true positive (**TP**) and challenging true-negative (**TN**) examples of TFBS (**Fig. S3**). Challenging TN, for example, might arise when a TFBS in one cell type is not a TFBS in another cell type, due to potentially subtle differences in the chromatin environment. Because only ~1% of the chromatin is expected to be accessible or bound by TFs in a given cell type [31], randomly sampling genomic intervals for model training is inefficient (i.e., results in too few TFBS examples)[21]. Thus, for each TF model, we defined "regions of interest" as the union of accessible chromatin and TFBS (i.e., ATAC-seq and ChIP-seq peaks) for each training cell type. We found that increasing the representation of accessible chromatin and TFBS in training examples improved performance, relative to randomly selected genomic regions. We refer to this practice as "peak-centric" training. Furthermore, we introduced "pan-cell" training, to increase the number of challenging TN examples. In pan-cell training, regions of interest (e.g., TFBS) in one cell type are equally likely to be selected from the other training cell type(s), which often do not share the TFBS. In this way, the training examples are enriched for challenging TN examples. Peak-centric, pan-cell training outperformed random sampling, with particularly strong gains for smaller training dataset sizes (**Fig. S3A**).

maxATAC models offer high-performance TFBS prediction from ATAC-seq at genome scale

For benchmarking, we compared maxATAC TFBS predictions to "gold-standard" TF ChIP-seq experiments in independent test cell types and chromosomes (**Methods**). By using all possible train-test cell type splits, we report a distribution of precision-recall statistics [32] for each of the

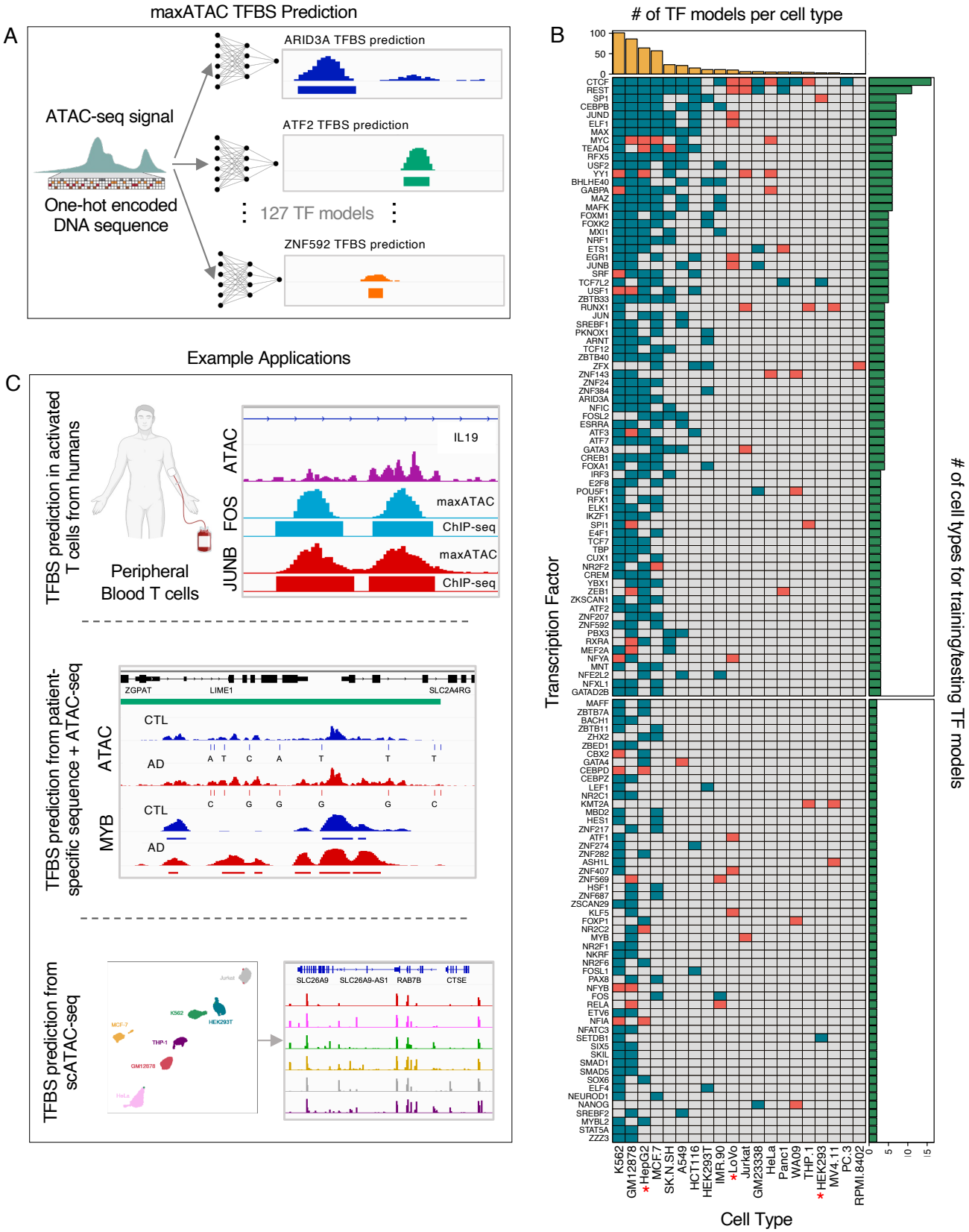


Fig. 1. Overview of maxATAC. (A) maxATAC deep neural network models use DNA sequence and ATAC-seq signal to predict TFBS in new cell types. **(B)** The maxATAC training data per TF and cell type with OMNI-ATAC-seq (top: 74 "benchmarkable" TF models with ≥ 3 cell types available, bottom: 53 TF models with only 2 cell types for training). Teal boxes indicate ChIP-seq from ENCODE, while red boxes indicate data from GEO. Red stars denote cell types for which we generated OMNI-ATAC-seq. **(C)** Example applications of maxATAC TFBS prediction to primary cells, scATAC-seq and clinical studies combining DNA sequencing with ATAC-seq.

74 “benchmarkable” TFs (area under precision recall (**AUPR**) or precision at 5% recall, **Fig. 2, S4A-B, Table S2**).

The maxATAC models advance TFBS prediction from ATAC-seq. We first compared maxATAC model performance to the most popular method of TFBS prediction, TF motif scanning in ATAC-seq peaks[33–36]. maxATAC outperformed standard motif scanning for every TF (**Fig. 2A, S4A**). For the vast majority of TFs (61 of 74), maxATAC also performed favorably to TFBS prediction using the average ChIP-seq signal for that TF across the training cell types (**Fig. 2B, S4B**). Comparison to this simple model[37] ensured that the maxATAC models learned ATAC-seq and sequence features predictive of new TFBS in new cell types, not just TFBS in common with the training cell types.

A direct comparison to the top-performing ENCODE-DREAM Challenge models is problematic due to differences in (1) technologies (ATAC-seq versus DNase-seq), (2) the cell types and number of samples available for model training, and (3) an incompletely overlapping set of TF models (74 for maxATAC and 12 for ENCODE-DREAM with 8 TFs in common). Acknowledging these caveats, we report that maxATAC performance, at $AUPR_{MEDIAN}$ = .43, is roughly on par with the top-performing models at the 2017 ENCODE-DREAM Challenge ($AUPR_{MEDIAN}$ = .44 [18] or .41 [17]) (**Fig. 3D**, **Table S2**).

For the 8 TFs common between the ENCODE-DREAM and maxATAC benchmarks, we compare maxATAC to performance reported by more recent, current state-of-the-art methods: Factornet[19], Leopard[20] and DeepGRN[21]. Overall, the performances are roughly comparable. For example, the maxATAC models on ATAC-seq outperform Factornet on DNase-seq for 5 out of 8 TFs, while maxATAC models outperform Leopard for 3 and DeepGRN for 2 out of 8 TFs (**Fig. 2F**). Given differences in DNase-seq and ATAC-seq technology, we recommend the maxATAC models for TFBS prediction from ATAC-seq, as they were specifically trained and benchmarked on ATAC-seq. (Further analysis of technology-dependent performance follows in subsequent sections.)

Given these positive benchmarking results, we extended maxATAC model construction to all 127 TFs in our training dataset. Because 53 of the TFs had only two cell types available for training (i.e., no opportunity for both cross-cell type training *and* estimation of test performance in a new cell type), we explored the relationship between validation AUPR and test AUPR for the 74 benchmarkable TF models. Here, validation AUPR corresponds to performance on validation chromosome 2 in cell types used for training and validation, while test AUPR corresponds to performance on held-out test chromosome 1 for a cell type independent of training and validation cell types (**Methods**).

We observed a nearly one-to-one correspondence between validation and test performance (**Fig. 2C, S4C-D**). Given this strongly predictive relationship between validation and test performance, the final maxATAC models were constructed using all available cell types for model training. For each TF model, we estimated confidences for maxATAC scores based on interpretable validation performance metrics (precision, recall, F1-score), so that users can choose confidence cutoffs suited to their research goals (**Methods**). For example, in the context of GRN inference with the *Inferelator* [13], initial, noisy TFBS predictions from ATAC-seq are subsequently refined by gene expression modeling, so a GRN modeler might equally prioritize precision and recall (i.e., use F1-score). In contrast, a researcher interested in experimental validation of a TFBS at a particular locus might prioritize high precision. Interpretable confidence cutoffs are a unique aspect of the maxATAC software package, which we further benchmark in primary cells (below).

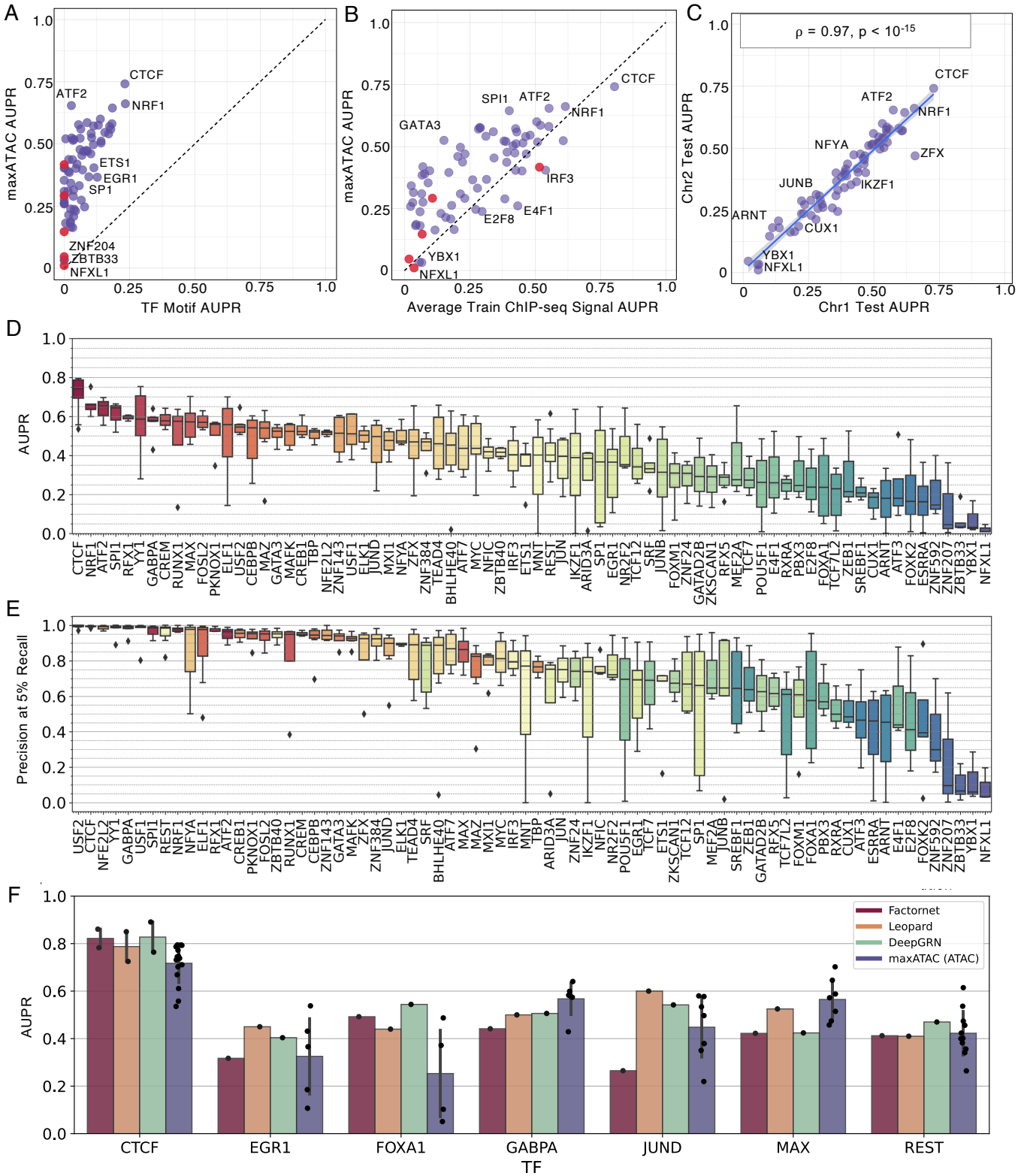


Fig. 2. The maxATAC models offer state-of-the-art TFBS prediction from ATAC-seq. For every TF model, one cell type and two chromosomes (chr1, chr8) were held out during training to assess predictive (test) performance in a new cell type. maxATAC model performance is compared to **(A)** TF motif-scanning in ATAC-seq peaks and **(B)** TFBS prediction using the averaged ChIP-seq signal from the training cell types; each dot represents $AUPR_{MEDIAN}$ across train-test cell type splits. Red dots indicate TFs with no known motifs in CIS-BP. **(C)** Validation performance ($AUPR_{MEDIAN}$) on chr2 (training cell types) as a function of test performance ($AUPR_{MEDIAN}$) on chr1 (held-out test cell type) (n=74; $\rho_{Pearson}$=0.97, $P < 10^{-15}$). Test **(D)** AUPR (median = 0.43) and **(E)** precision at 5% recall (median = 0.85). Boxplots display median (horizontal line), interquartile range (box), 3-quartile range (whiskers) and points outside the 3-quartile range (diamonds). **(F)** Test AUPR of the maxATAC models on ATAC-seq compared to test AUPR reported by state-of-the-art deep learning models (Factornet, Leopard and DeepGRN) on DNase-seq.

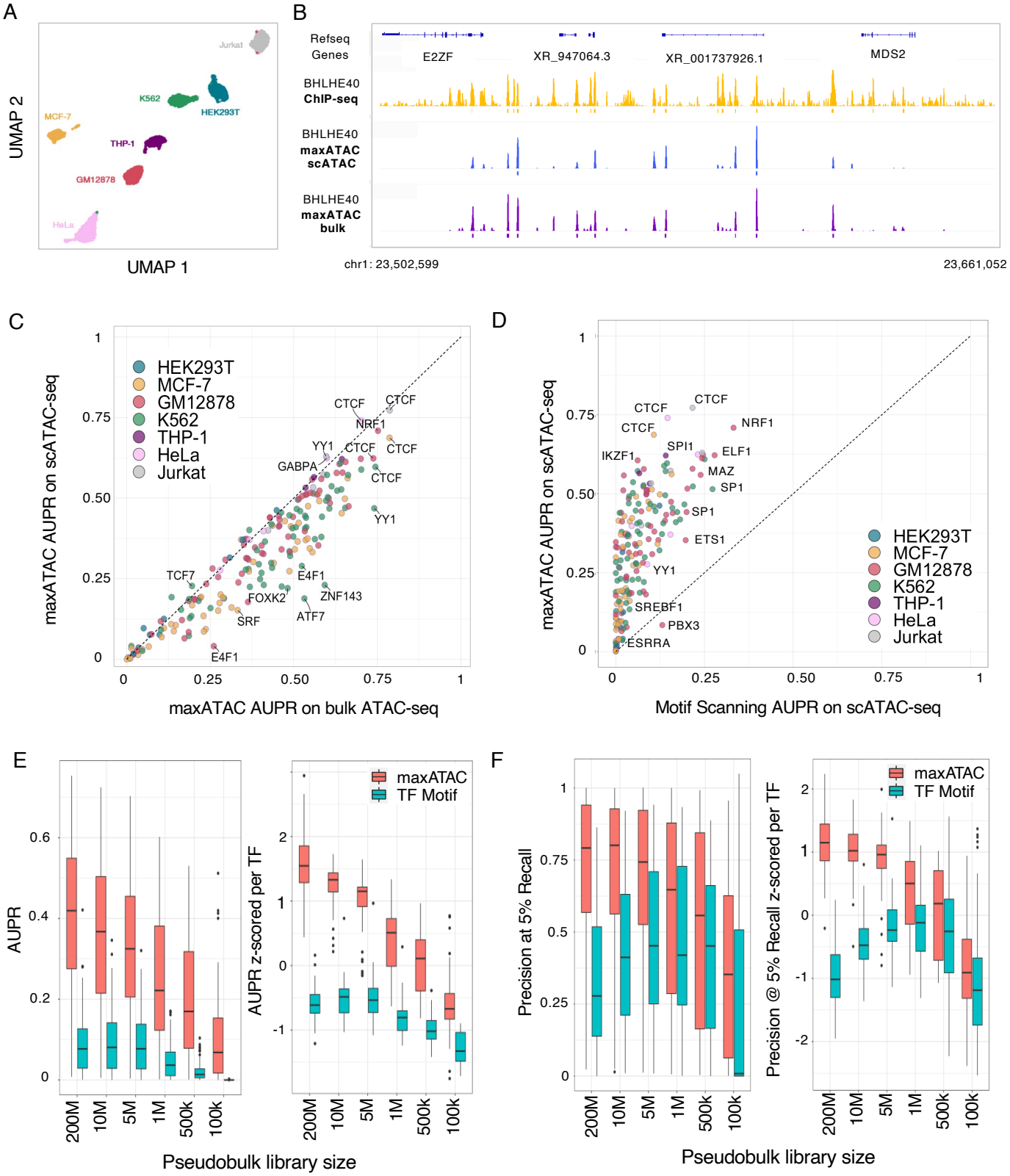


Fig. 3. maxATAC offers state-of-the-art TFBS prediction from scATAC-seq. (A) UMAP of 10x scATAC-seq data from 7 cell types in a cell line-mixing experiment (Granja et al. 2021) that enabled test performance evaluation for 193 maxATAC models. **(B)** IGV tracks comparing BHLE40 TFBS predicted by maxATAC in GM12878 from scATAC-seq (blue) or bulk ATAC-seq (purple), relative to BHLE40 ChIP-seq (–log10(p-value) signal tracks in yellow) located at chr1:23,502,599-23,661,052 (158kb region). **(C)** Test AUPR for maxATAC in scATAC-seq relative to maxATAC performance on bulk ATAC-seq. **(D)** Test AUPR of maxATAC on scATAC-seq versus AUPR of TF motif scanning on scATAC-seq. Test **(E)** AUPR and **(F)** precision at 5% recall performances for maxATAC (red) and TF motif scanning (teal) as a function of down-sampled pseudobulk library sizes from scATAC-seq of GM12878 (n = 55 TFs with maxATAC models and TF motifs available). Given the range of performances per TF model, performances are also z-score normalized per TF, to better visualize model- and library-size-dependent trends.

Our large training dataset afforded new opportunities to explore potential factors driving differential performance across TF models. Test (**Fig. 2D-E**) and validation (**Fig. S4E-F**) performance varied dramatically across TFs, with $AUPR_{median}$ ranging from .75 for CTCF to .01 for NFXL1 (**Fig. 2D**). Given (1) the close correlation between validation and test performance and (2) that validation performance is available for all 127 TF models (versus 74 for test), we analyzed performance variation for both validation and test.

Model performance showed a modest dependence on the number of training cell types available (**Fig. S5A**). While the interquartile range for models trained with 5 cell types was above the overall median AUPR performance (.43), the interquartile range for models trained with 2-4 cell types contained the overall median. We credit robust prediction in the small training data set regime (2-4 cell types) to our peak-centric, pan-cell model training strategy (**Methods, Fig. S3**).

While all maxATAC models outperformed motif-scanning (**Fig. 2A**), maxATAC performance was comparable to averaged train ChIP-seq signal for several TFs (**Fig. 2B**). Relative performance (maxATAC versus averaged train ChIP-seq) was not simply explained by the number of training cell types (ChIP-seq experiments) available, suggesting a role for TF-intrinsic factors (**Fig. S5B**). We hypothesized that training ChIP-seq signal would perform well for TFs whose binding patterns changed little across cell types, while maxATAC integration of context (ATAC-seq) with sequence would be especially critical for TFs whose binding patterns varied across cell types. We used Jaccard overlap [38] of TFBS between pairs of training cell types as a proxy for cell-type specificity (**Fig. S5C**). There was a strong relationship between cell-type specificity and maxATAC performance gains relative to average ρ-seq signal ($\rho_{Pearson}$=.53, $P<10^{-5}$). For cell types with very high specificity (average Jaccard overlap <.05), AUPRs were, on average, >2-fold higher for maxATAC versus average ChIP prediction. For TFs with many shared binding sites across cell types (average Jaccard overlap >.25), performance between maxATAC and average ChIP signal were on-par. Thus, maxATAC modeling is especially important for TFs with context-specific binding sites and a good choice for TFs with many shared TFBS across cell types.

Five of the maxATAC TF models (GATAD2B, NFXL1, ZBTB40, ZNF207, and ZNF592) have no characterized motif in the CIS-BP database [39], suggesting that prediction for these TFs might be especially challenging (motif-less models are highlighted with red dots in **Fig. 2A-B, S4A-B**). Indeed, three of the models (NFXL1, ZNF207, ZNF592) had the 1st, 4th and 5th-lowest test AUPRs. However, two models had good test performance: ZBTB40 at $AUPR_{median}$ = .42 and GATAD2B at $AUPR_{median}$ = .31. Model interpretation, to uncover the predictive sequence and chromatin accessibility features for these – and all 127 – maxATAC models, will be the subject of future investigation.

maxATAC performance extends to scATAC-seq and other chromatin accessibility techniques

Having constructed the largest suite of TFBS deep neural network models for ATAC-seq, we next tested whether these models, trained on bulk data from cell lines, could perform well in new domains: single-cell ATAC-seq and primary cells. The maxATAC models were specifically designed to improve prediction of TFBS from cell types from *in vivo* contexts, where limited sample material or cell sorting strategies would preclude experimental TFBS measurement. Thus, we evaluated the maxATAC models on scATAC-seq.

Clustering of individual cells into cell types and subpopulations is a key first step in scATAC-seq analysis. Next, for each cluster, accessibility per cell is summed to create “pseudobulk” ATAC-seq signal for each inferred population of cells. From pseudobulk profiles, regions of chromatin

accessibility are detected and annotated with TFBS predictions in standard scATAC-seq pipelines [35,36]. Pseudobulk scATAC-seq profiles are natural inputs for TFBS prediction with maxATAC.

For benchmarking, we took advantage of an experiment in which nuclei from 10 cell lines were mixed together for scATAC-seq [35]. Seven of the cell lines overlapped our maxATAC benchmarking cell types (**Fig. 3A**), providing the opportunity to estimate test performance of maxATAC on scATAC-seq for 63 models (**Fig. 3**, **Table S3**). In a side-by-side comparison using the same test cell types, maxATAC performed nearly as well on scATAC-seq as population-level ATAC-seq (**Fig. 3B-C**). As expected, maxATAC outperformed popular TF motif scanning (**Fig. 3D**).

To guide experimental design and application of maxATAC to scATAC-seq, we next examined the relative performance of maxATAC to TF motif scanning as a function of pseudobulk library size, using GM12878 as an independent test cell type (**Fig. 3E-F, S6**, see **Methods**). Depending on the scATAC-seq experiment, the median number of fragments per cell typically ranges from ~5k-10k. Thus, we explored pseudobulk library sizes ranging from 200M to 100K fragments, which roughly correspond to 20-40k to 10-20 cells, respectively. We wanted to know whether there was a pseudobulk library size cutoff at which the simpler method (TF motif scanning) would be preferable to maxATAC. As expected, smaller pseudobulk library sizes (undersampling) diminished maxATAC performance. For example, at library size of 1M (~100-200 cells), median AUPR = .23 is nearly halved relative to the initial library size (200M fragments, ~20-40k cells, median AUPR = .43). Despite this, maxATAC outperformed TF motif scanning at every library size cutoff, even for the sparsest simulated pseudobulk library (corresponding to ~10-20 cells). While this was expected for the AUPR performance metric (because maxATAC is genome-scale and therefore has a recall performance advantage over TF motif scanning), these down sampling experiments importantly demonstrated superior maxATAC performance for the precision at 5% recall metric. At the smallest library size (100k), TF motif scanning provided 0% recall for the majority of TFs (31 of 55, see per-TF performance breakdown **Fig. S6**). This analysis helps set expectations for TFBS prediction from both maxATAC and TF motif scanning as a function of pseudobulk library size. These results nominate maxATAC as the preferable method even for small library sizes.

Given the good performance of the maxATAC models on scATAC-seq, we evaluated maxATAC performance on additional experimental chromatin accessibility protocols. For these analyses, we again focused on the extensively characterized lymphoblastoid cell line, GM12878, which enabled us to benchmark 60 maxATAC TF models on ATAC-seq generated from four different protocols and variable cell-input conditions (**Fig. 4, S7A-E**). In addition, we assessed DNase-seq. The ATAC-seq protocols included the original “standard” ATAC-seq protocol [8], OMNI ATAC-seq [26] (used for maxATAC model training, due to improved signal-to-noise, recovery of enhancers and reduction in mitochondrial contamination), Fast-ATAC [40] (designed for use in primary blood cells as it requires only a single lysis step), and the most recently developed: scATAC-seq (10x Genomics platform) [12]. We also evaluated TFBS prediction by motif scanning, to help distinguish whether protocol-dependent maxATAC performance was inherent to the quality of the assay for TFBS prediction (i.e., generalizable) or simply a limitation of maxATAC training on one assay type (OMNI-ATAC-seq).

For each experimental protocol, maxATAC dramatically outperformed motif-scanning (**Fig. 4A**). To better visualize assay-dependent performance, we normalized AUPR performance per TF, independently for maxATAC and motif scanning predictions (**Fig. 4B-C**). While maxATAC performance was indistinguishable for OMNI-ATAC-seq (50k cells) and scATAC-seq (1k-10k cells), performance was significantly worse for DNase-seq, OMNI-ATAC-seq (500 cells), scATAC-

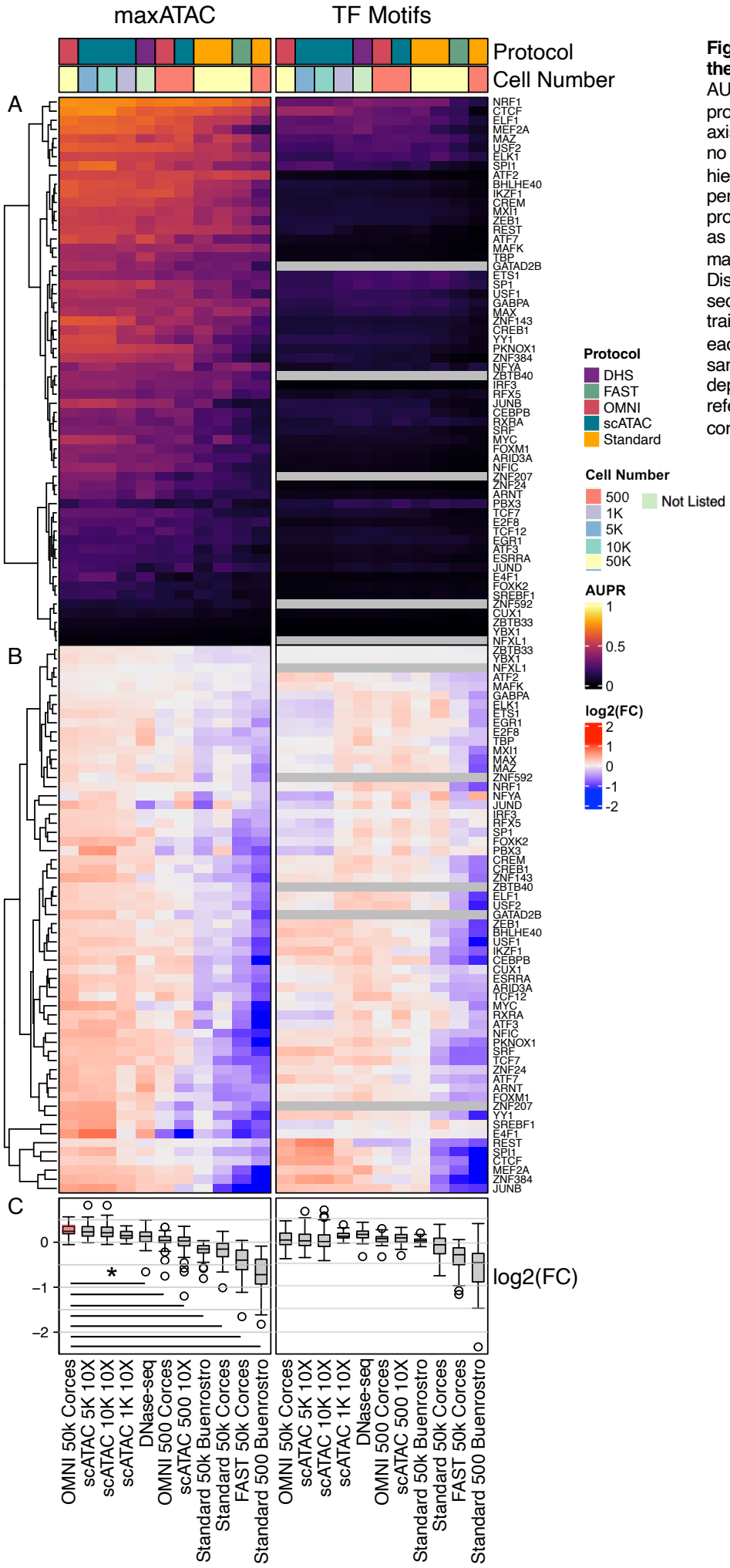


Fig. 4. Protocol and cell input numbers influence the performance of TFBS predictions. (A) Test AUPR for 60 TFs (y-axis) across different ATAC-seq protocols, cell numbers and prediction methods (x-axis) in the GM12878 cell line. Grey squares indicate no predictions, due to lack of TF motif. TFs are hierarchically clustered based on maxATAC performance. **(B)** To visualize trends across protocols, AUPRs were normalized per TF (row-wise) as the $\log_2(\text{AUPR}:\text{AUPR}_{mean})$, independently for maxATAC and TF motif-scanning AUPRs. **(C)** Distribution of the $\log_2(\text{AUPR}:\text{AUPR}_{mean})$ per ATAC-seq sample. Given the maxATAC models were trained on OMNI-ATAC-seq ~50k cells, we compared each experiment to the reference "OMNI 50k Corces" sample (red boxplot). Black lines indicate protocol-dependent performance differences relative to the reference (Student's two-sided *t*-test, Bonferonni-corrected $P < 0.05$).

seq (500 cells), standard ATAC-seq (50k or 500 cells) and Fast-ATAC relative OMNI-ATAC (50k cells), the protocol used for maxATAC training data (two-sided Student's *t*-test, Bonferroni-corrected P_{adj}<.05). Library sizes were comparable (within 2-fold of the OMNI-ATAC-seq) for scATAC-seq (1k, 500 cells), standard ATAC-seq and DNase-seq **(Fig. S7C,D**), suggesting performance differences among these experiments were not simply consequences of library size. Thus, maxATAC performance across all TF models is worse for standard ATAC-seq and DNase-seq relative to OMNI-ATAC-seq or scATAC-seq (1k-10k cells). In contrast, the performance of motif scanning was roughly comparable across standard ATAC-seq (50k cells), OMNI-ATAC-seq, DNase-seq and scATAC-seq. Collectively, these analyses suggest that technical differences between chromatin accessibility protocols undermine cross-protocol model prediction, but they also offer a simple solution: assay-specific modeling training for state-of-the-art prediction from that assay.

maxATAC performance extends to primary cells

We next evaluated the performance of maxATAC on stimulated primary cells. In particular, we examined activation of naïve $CD4^+$ T cells five hours following T-cell receptor (**TCR**) and CD28 stimulation, a timepoint at which standard ATAC-seq and ChIP-seq of TCR-dependent TFs (FOS, JUNB and MYC) had been collected [41] (**Fig. 5A**). maxATAC predictions for all three TFs outperformed TF motif scanning in ATAC-seq peaks (**Fig. 5B-D**). Even at low recall, maxATAC performance was on-par (JUNB) or better (MYC, FOS) than TF motif scanning and unparalleled at recall >10%. The TF motif-scanning precision-recall curves drop off suddenly at ~10% recall because motif scanning is not genome-scale (i.e., it is limited to ATAC-seq peaks). FOS and JUNB maxATAC predictions also outperformed TFBS prediction by average training ChIP-seq signal, while performance for MYC was slightly worse. This performance is especially promising, because the MYC and JUNB models perform worse on standard versus OMNI-ATAC-seq (1.3- and 3.5-fold AUPR decrease, respectively, **Fig. 4A**). Thus, performance evaluation on standard ATAC-seq likely underestimates maxATAC performance on OMNI-ATAC-seq.

We also investigated how the performance metrics associated with the FOS, JUNB and MYC maxATAC models extrapolated to this new test dataset. JUNB and MYC models had 4-5 benchmark cell types, and their associated test and validation performance metrics were good predictors of performance in the $CD4^+$ T cells (**Fig 5E**). In contrast, the FOS model was constructed from only two cell types, and its performance, although superior to motif-scanning, was lower than estimated by validation performance. Aggregation of additional maxATAC training data will be the focus of future work.

We attribute maxATAC's generalizability to several key observations and methodological strategies. Good performance across ATAC-seq protocols required an extended genomic blacklist [42] and a robust normalization strategy [12]. In contrast to DNase-seq, ATAC-seq is variably contaminated with accessibility signal from mitochondrial chromosomes [23,43]. For example, we detected genomic regions of high chromatin accessibility in OMNI-ATAC-seq that were not present in scATAC-seq or DNase-seq (**Fig. S8**). Those regions shared high sequence similarity with mitochondrial DNA, and, in general, we found less mitochondrial contamination in scATAC-seq than bulk. Thus, we expanded our blacklist to include chromatin regions with high mitochondrial sequence similarity. Despite the augmented blacklist, several of the transformed cell lines still had extreme, outlying ATAC-seq signal (**Fig. S9A-C**). Some of these regions overlapped cancer-specific driver super enhancers (e.g., *TRIM37* locus in MCF-7). To be robust to these biological outlier regions, we normalized ATAC-seq data to the 99th-percentile highest signal (in contrast to the absolute max in standard min-max normalization; **Methods**); this strategy proved critical to maxATAC performance in scATAC-seq and primary cells (**Fig. S9D**).

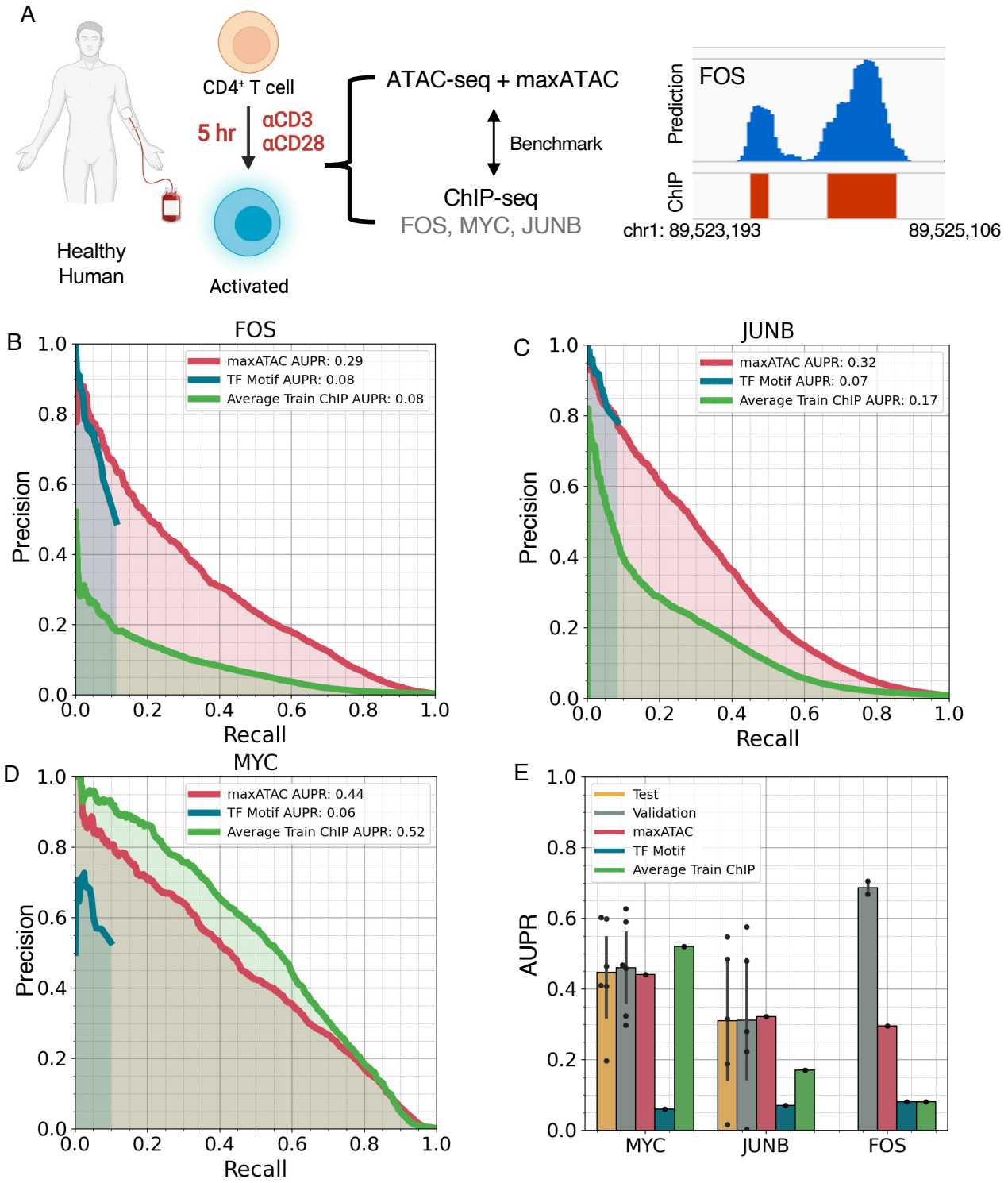


Fig. 5. maxATAC models perform well in primary human cells. (A) Study design for maxATAC benchmarking in primary human CD4+ T cells. Precision-recall curves for TFBS predictions from maxATAC (red line), motif-scanning (blue) and averaged training ChIP-seq signal (green) for **(B)** FOS, **(C)** JUNB and **(D)** MYC, relative to experimentally measured TFBS (ChIP-seq). **(E)** Comparison of test performance in primary cells (red) to the estimates of test (green) and validation (yellow) performance (available with each trained maxATAC model) as well as TFBS prediction by TF motif scanning (blue) or average training ChIP-seq (green). Each point corresponds to a unique train-test cell type split. Error bars indicate standard deviation. Test performance minimally requires 3 cell types and therefore was available for MYC (n=6 cell types) and JUNB (n=5 cell types) but not FOS (n=2 cell types).

435
436 **maxATAC identifies allele-dependent TFBS at atopic dermatitis genetic risk variants**
437
438 Identifying the cellular and molecular drivers of phenotypic diversity is a fundamental goal of basic
439 and translational research. Complex traits are products of genetic and environmental factors.
440 Chromatin accessibility is sensitive to environmental factors, like age [44] and microbiome [45],
441 and therefore a critical complement to genetic profiling of patients across disease spectra, from
442 cancer to autoimmune and obesity-related diseases. While previous work in deep neural network
443 modeling focused on interpretation of genetic variants [28,29,46], maxATAC integrates both
444 genetic (sequence) and environmental (ATAC-seq) signals and is therefore ideally suited to the
445 elucidation of molecular drivers of diseases involving both genetic and environmental
446 components.
447
448 We demonstrate the potential for maxATAC in the context of a complex disease. Atopic dermatitis
449 (**AD**) is one of the most common skin disorders in children. Its etiology involves both genetic and
450 environmental factors, with 29 independent AD risk loci known [47,48]. Here, we take advantage
451 of an important genomics resource in AD: ATAC-seq and RNA-seq of activated $CD4^+$ T cells,
452 along with whole-genome sequencing, of AD patients and age-matched controls (**Fig. 6A**) [49].
453 Previous analysis of this dataset identified several AD risk variants with allele-dependent
454 chromatin accessibility in activated T cells [49].
455
456 Here, we use maxATAC to identify TFBS associated with allele-dependent chromatin accessibility
457 at these AD risk loci (**Table S3**). To avoid error associated with phasing of DNA and ATAC-seq
458 signal (into maternal and paternal strands), we identified a pair of AD and age-matched controls
459 ("AD2", "CTL2") where the AD patient was homozygous for the AD risk haplotypes while the
460 control was homozygous for the AD non-risk haplotype at two independent loci tagged by two
461 variants: rs6062490 and rs1151624 (**Fig. 6B, Table S4**). For both variants, we considered the full
462 haplotype block (linkage disequilibrium R^2>.8), using donor-specific DNA sequence and
463 accessibility to predict TFBS for 105 expressed TFs with maxATAC models available (**Fig. 6C-G,**
464 **Fig. S10A, Methods**).
465
466 The rs6062490 haplotype block is ~34.2kb and contains 30 noncoding SNPs interspersed
467 between exons of two protein-coding genes: *RTEL1* and *TNFRSF6B* (**Fig. 6D**)**.** *TNFRSF6B* is an
468 anti-apoptotic gene, and serum levels of this protein are increased in patients with atopic
469 dermatitis[50,51]. The rs1151624 haplotype block is ~11.3kb and contains 19 noncoding SNPs
470 cis to three genes: *ZGPAT*, *LIME1*, and *SLC2A4RG* (**Fig. 6E**)**.** *LIME1* is a transmembrane protein
471 that controls effector T cell migration to sites of inflammation[52], while *SLC2A4RG* is a known
472 eGene, (associated with an eQTL) in activated $CD4^+$ T cells [53], regulatory T cells [54] and
473 multiple other T cell subsets [55]. Thus, genes at both risk loci have ties to T cell biology, T-cell
474 eQTLs and AD.
475
476 For both loci, we ranked TFs based on the number of predicted differential TFBS regions between
477 AD2 and CTL2 (**Fig. 6F, S11**). Although TFBS were generally increased in the AD patient relative
478 to the control in these loci (53 TFs), three TFs had a predicted decrease in binding sites in AD2
479 (KMT2A, LEF1, STAT5A), while 50 TFs had no TFBS predicted in these loci. MYB and FOXP1
480 showed the greatest differential binding, and their TFBS were predominantly increased in the AD
481 patient. FOXP1 has been previously implicated in the maintenance of T cell quiescence and its
482 expression is typically repressed in activated T cells [56]. MYB is a critical regulator of regulatory
483 T cell differentiation and immune tolerance [57]. Thus, both TFs have known roles in T-cell

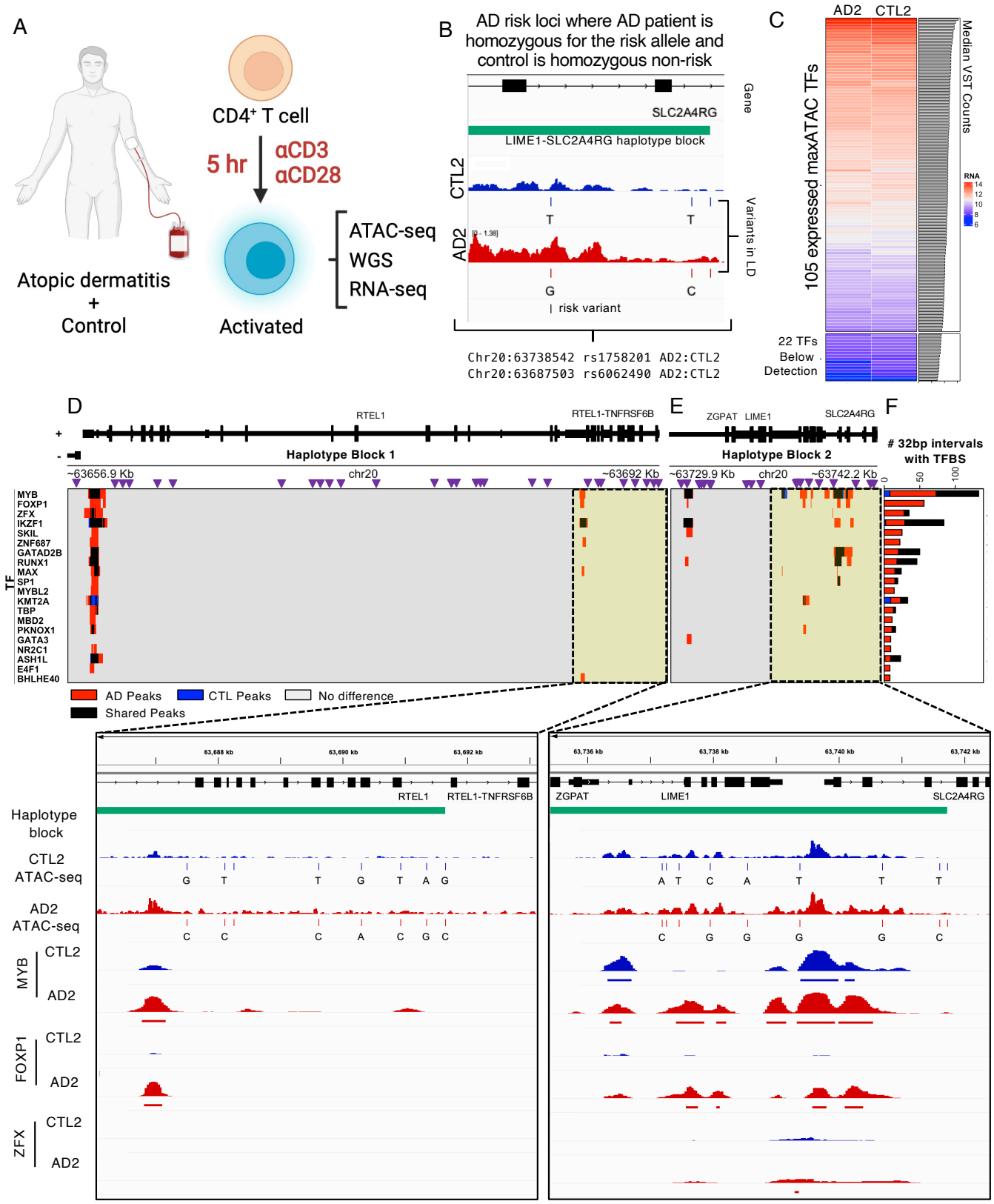


Fig. 6. maxATAC TFBS prediction at atopic dermatitis risk loci in patient-derived CD4+ T cells. (A) In a previous study (Eapen et al., 2021), peripheral CD4+ T cells were isolated from atopic dermatitis (**AD**) patients and age-matched controls (**CTL**) and TCR-stimulated prior to ATAC-seq, RNA-seq and DNA-seq data generation. **(B)** We identified a pair of donors in which the AD patient (AD2) was homozygous for the risk allele and the age-matched control (CTL2) was homozygous for the non-risk allele at two independent loci: rs1758201 and rs6062490. **(C)** 105 of the 127 maxATAC TFs were nominally expressed for the donor pair, and these TFs were selected for TFBS prediction with maxATAC. We identified differential TFBS in the haplotype blocks containing **(D)** rs6062490 and **(E)** rs1758201. Purple triangles represent SNPs in linkage disequilibrium (R^2>.8) with the AD risk alleles. In the heatmap below, red or blue intervals (32bp) indicate respective gain or loss of TFBS in the AD patient relative to control and black denotes intervals of shared TFBS. TFBS are determined using a cutoff that maximizes the predicted F1-score per TF model. **(F)** The 20 TFs with the greatest number of differential binding regions between AD2 and CTL2 are shown. **(G)** IGV screenshots showing regions (highlighted in yellow) of predicted differential TFBS in CTL2 (blue tracks) compared to AD2 (red tracks). The haplotype block is indicated in green. The top 4 signal tracks represent donor-specific ATAC-seq signal and genetic variants. The bottom 6 signal tracks represent predicted TFBS.

484 biology. We visualized regions of each haplotype block, putative cis-regulatory modules, where
485 MYB and FOXP1 were predicted to bind with several other factors (**Fig. 5G, S11**).
486
487 Although limited in statistical power, we examined the correlation between expression of genes
488 at these loci in AD patients and age-matched controls (**Fig. S10B-G**). We observed a trend for
489 increased expression of *SL2A4RG* and decreased expression of *ZGPAT* in AD patients harboring
490 the risk variant rs1151624 (trend held for all 4 homozygous-risk AD patients relative to their
491 homozygous or heterozygous non-risk controls). We observed increased expression of all three
492 genes *(RTEI1, RTEL1-TNFRSF6B*, *TNFRSF6B)* associated with rs6062490, when comparing the
493 single pair of homozygous-risk AD to homozygous non-risk control, but we observed inconsistent
494 trends from the two donor pairs in which the AD patients were homozygous risk and controls were
495 heterozygous non-risk. These trends, in addition to support for *SLC2A4RG* as an eGene in $CD4^+$
496 T cells [53–55], nominate potential molecular mechanisms, specific TFs that might alter the
497 expression of genes important to T cell function (MYB and FOXP1) in AD patients. Interestingly,
498 for the regions shown in **Fig. 6G**, neither FOXP1 nor MYB had motif occurrences in the differential
499 ATAC-seq peaks, highlighting the unique predictive capabilities of maxATAC.
500
501 maxATAC analysis of this clinical genomics study provided “in silico ChIP-seq" for 105 TFs, in a
502 setting where experimental measurement of TF occupancy for 105 factors was infeasible.
503 Application of maxATAC to the growing number of genetic studies with population-level and
504 single-cell ATAC-seq will improve the power of these studies to accurately predict TF mediators
505 of allelic chromatin accessibility and gene expression. Furthermore, many genetic tools use
506 overlap with TFBS to nominate potentially causal risk variants [58–60], where TFBS are predicted
507 based on suboptimal TF motif scanning or ChIP-seq in a potentially suboptimal cell type or
508 condition, due to lack of data in the *in vivo* disease context. Integration of maxATAC TFBS
509 predictions into genetic analysis pipelines will be the focus of future work.
510
511 **Discussion**:
512
513 Genomic measurement and machine learning capabilities are advancing at an unprecedented
514 pace. The possibilities for computational modeling to address fundamental questions in biology
515 and human health have never been greater. Leveraging the fastest-growing chromatin-state
516 measurement ATAC-seq, maxATAC seeks to make this moment’s state-of-the-art in transcription-
517 factor binding prediction readily achievable across the basic and biomedical research spectrum.
518 Rigorously benchmarked across cell lines, primary cells, and single-cell ATAC-seq, maxATAC
519 will improve TFBS predictions and knowledge gain from ATAC-seq studies. With maxATAC and
520 its user-friendly codebase (available from https://github.com/MiraldiLab/maxATAC), genome-
521 scale TFBS prediction can be accomplished for 127 human TFs using a single ATAC-seq or
522 scATAC-seq experiment.
523
524 maxATAC is a unique resource, distinct from previous deep-learning models in genomics. As
525 highlighted in the introduction, current state-of-the-art models for TFBS prediction from chromatin
526 accessibility were trained on DNase-seq (a less popular alternative to ATAC-seq), and their
527 availability is limited to at most 32 TFs (as opposed to 127 for maxATAC)[17–21]. There are also
528 numerous deep learning methods for ATAC-seq data[61–65], but these methods have different
529 modeling objectives, including: prediction of ATAC-seq signal from DNA sequence (e.g., AI-
530 TAC[63], chromBPNet[61], deepMEL[62]), denoising of ATAC-seq and scATAC-seq signal
531 (AtacWorks[65]) and classification of enhancers, promoters and insulators based on ATAC-
532 derived inputs (CoRE-ATAC[64]). None of these models tackle the trans-cell type TFBS prediction

challenge. Finally, there are many deep learning methods that have been used to predict the impacts of genetics on epigenome signals, including TF binding (DeepSea[28], Basset[46], Basenji[29], Enformer[66], ChromBPNet[61]). However, DNA sequence is the *only* input to these models, and none make trans-cell type predictions (i.e., are capable of TFBS predictions outside of the cell types and conditions used for model training). In contrast, maxATAC predicts TFBS based on *both* DNA sequence and ATAC-seq signal, integrating both genetic and environmental effects to predict TFBS in new cell types and contexts (e.g., patient-specific scATAC-seq of a tumor biopsy). This is especially important for elucidation of complex disease mechanisms where environmental context contributes to disease etiology. Thus, maxATAC is a unique resource for TFBS prediction, with no comparable tools available – currently.

That said, we are confident that the training data we curated for maxATAC will spur rapid community advances in trans-cell type TFBS prediction from ATAC-seq. For this purpose, the maxATAC benchmark is organized (quality-controlled, processed, ready-to-go) for community download and development (https://doi.org/10.5281/zenodo.6761768).

We echo a top-performing group in the ENCODE-DREAM *in vivo* TFBS Prediction Challenge [17], acknowledging that state-of-the-art computational prediction of TFBS from ATAC-seq, at median AUPR = ~.4, is not yet accurate enough to replace TFBS measurement, when experimentally feasible. There are many avenues for improvement. For example, maxATAC models TFBS as binary, on-or-off events, despite the quantitative nature of the population-level TFBS measurements used for model training. Recent works [29,67] introduced new loss functions, specifically designed to quantitatively model chromatin state signal from NGS. Furthermore, BPnet's base-pair resolved architecture has already been tested on TF ChIP-seq [67] and appears well-poised to leverage highly-resolved ATAC-seq as model input, too.

TF binding *in vivo* is a multivariate process involving cooperation, competition, and co-binding among TFs. Yet each maxATAC TF model was trained independently, with predictions made for each TF one-at-a-time. In genomics, multi-task modeling of multiple TFs together is a common technique to enhance recovery of predictive sequence features, especially those involved in co-binding, relative to single TF models [68]. The sparsity of our training data (**Fig. 1B**), combined with our goal to predict TFBS in new cell types, limited application of traditional multi-task learning approaches. Pre-training and transfer learning with large genomics resources could bridge the gap, providing richer, multi-task-like features for maxATAC [46], while we await experimental advances in massively parallel TF occupancy measurements. Relatedly, TFs mediate interactions between enhancers and promoters, yet, with maxATAC, each 1kb genomic interval is modeled independently. Chromatin-looping interactions could improve TFBS prediction, and, in the absence of context-specific experimental data, be inferred from existing looping data [69] or estimated from covariance of functional genomics assays [70] (including scATAC-seq [71]). 3D-chromatin interactions could be incorporated using graph neural networks [72,73] or with simple post-processing, like affinity propagation [74] of TFBS labels based on 3D contacts.

For some TFs, experimental-quality TFBS prediction will require more than ATAC-seq signal, especially for signal-activated TFs that bind pre-existing accessible chromatin. Thus, future directions for maxATAC will include addition of new data types, to implicate TFs based on transcriptional activity and methylation status of chromatin. Finally, the coverage of maxATAC models is still small relative to TF motif models (127 versus ~1200 for human TFs [24]). Thus, identification and generation of additional TFBS and ATAC-seq data is a top priority.

In summary, this work represents a significant resource, advancing community access to state-of-the-art TFBS prediction from popular ATAC-seq and scATAC-seq protocols.

584
585
586 **Materials and Methods**:
587
588 **maxATAC Training Data**
589
590 We curated training data for maxATAC from the CistromeDB [75] and ENCODE [15] databases
591 (**Figure 1A-B**), identifying cell types that had (1) a high number of human TF ChIP-seq
592 experiments and (2) ATAC-seq data. We limited ATAC-seq training data to higher quality OMNI
593 [26] and required paired-end sequencing, while, for ChIP-seq, we utilized both paired- and single-
594 end sequencing. We required sequencing depth ≥ 20 million reads per biological replicate (sum
595 of the spot numbers for SRR IDs per SRX ID). We manually verified the cell type, TF, and
596 experimental source of each experiment. To ensure that ATAC-seq and ChIP-seq were derived
597 from the same experimental conditions per cell type, we eliminated any ATAC-seq or ChIP-seq
598 experiment in which the cell type was perturbed (including genetically modified, transfected with
599 exogenous vectors, or treated with vehicle controls, environmental perturbations, metabolic
600 manipulations, or differentiation protocols).
601
602 ***ChIP-seq***
603
604 When available, we utilized processed ENCODE ChIP-seq experiments over other publicly
605 available data. For TF and cell type pairs with multiple ChIP-seq experiments available, we chose
606 the most recent experiment with the greatest number of reproducible TFBS (peaks) detected by
607 IDR analysis [76]. If available, we selected conservative over optimal IDR peak sets. We excluded
608 any experiment with a red flag (i.e., a ”critical issue” was identified by ENCODE) as well as
609 experiments with fewer than 500 TFBS detected. This resulted in 371 TF-cell type combinations
610 from ENCODE.
611 The remainder of our ChIP-seq training data required processing and additional quality
612 control. Using snakemake [77], we followed ENCODE3 [78] standards for ChIP-seq read
613 alignment, read filtering, and peak calling; this workflow is available from the maxATAC codebase.
614 In brief, each biological replicate was summarized according to SRX ID for a total of 316
615 experiments. Fastq files were downloaded (*SRAtools fasterq-dump* v. 2.10.8) with technical
616 replicates concatenated per SRX ID. Fastq files were assessed for adapter contamination and
617 read quality statistics (*FastQC* v. 0.11.9). Samples flagged for high levels of N sequences were
618 removed. *TrimGalore!* (v. 0.6.7) was used to remove adapter contamination and trim the low-
619 quality bases at the 3’ end of the sequencing read with the settings (-q 20). Samples with (1) < 15
620 million reads after filtering or (2) average read length < 20 bp after trimming were excluded from
621 the analysis. Reads were aligned to the hg38 reference genome using *bowtie2* (v. 2.4.4) [79] (--
622 very-sensitive --maxins 2000). The aligned reads were quality filtered with *samtools* (v. 1.9) [80]
623 (-F 1804 -q 30) and PCR duplicates were removed with *samtools markdup* (-r -s). Prior to peak-
624 calling, we also excluded reads mapping to blacklist regions compiled from ENCODE data [42]
625 as well as centromeres, telomeres, and annotated gaps. MACS2 (v 2.2.2.7.1) [81] was used to
626 call peaks on the filtered BAM file with parameters (--nomodel --extsize 147). For cell types and
627 TF combinations with multiple biological replicates, we provided all filtered BAM files during peak
628 calling. Peaks meeting a FDR = 5% cutoff were retained as TFBS for benchmarking. TF-cell type
629 conditions with fewer than 500 TFBS detected were excluded.
630 Further QC of the “non-ENCODE” ChIP-seq involved TF motif analysis and biological
631 replicate Pearson correlation of > .6 (when available). We used HOMER (v. 4.11) [34] with the
632 CIS-BP (v. 2.0) [39] database to test for enrichment of expected motifs in ChIP-seq peaks. CIS-
633 BP provides multiple motifs per TF, so we selected the most highly enriched motif per TF and
634 then ranked motif enrichment scores at the TF level. We excluded ChIP-seq experiments in which

the ChIP'd TF was not ranked among the top-10 enriched TFs. From 316 experiments, we derived 72 TF-cell type pairs. In total, our 127 TF models cover 443 unique cell types and TF combinations across 20 cell types (**Fig 1B, S1A**).

ATAC-seq

We combined public with in-house OMNI-ATAC-seq for our 20 benchmark cell lines. In-house, we ordered HepG2, LoVo and HEK293 cells from ATCC and targeted a median sequencing depth of 20 million reads for each cell type. OMNI-ATAC-seq was later released [78] for one of these cell types (HepG2), so we combined biological replicates for this cell type. To evaluate strategies for alignment, signal normalization and smoothing, we processed ENCODE along with in-house ATAC-seq. ENCODE ATAC-seq were downloaded, converted to FASTQ (*SRAtools fasterq-dump*) and then subsampled to a depth of 30 million reads per biological replicate, to limit compute time for alignment. For both ENCODE and in-house data, we evaluated sequencing quality with *FastQC*. Sequencing adapters and bases with a PHRED score < 30 were trimmed with the package *Trim Galore!* [82] using the parameters (-q 30 -paired). We excluded ATAC-seq experiments with < 20 million reads.

Alignment. We investigated several alignment strategies, using GRCh38 as the reference genome. We tested the performance of STAR (v. 2.7.0a) [83], bowtie2 (v. 2.4.4) [79], and bwa-mem (v. 0.7.17) aligners on TFBS predictions. Two STAR alignment strategies were tested, one with default parameters (--alignIntronMax 1 --alignMatesGapMax 2000 --alignEndsType EndToEnd) and the second with parameters from the TOBIAS[84] TF footprinting package (--alignIntronMax 1 --alignMatesGapMax 2000 --alignEndsType EndToEnd --outMultimapperOrder Random --outFilterMultimapNmax 999 --outSAMmultNmax 1 --outFilterMismatchNoverLmax 0.1 --outFilterMatchNmin 20 --alignSJDBoverhangMin 999 --alignEndsProtrude 10 ConcordantPair). For STAR, a MAPQ score of 255 indicates properly paired reads with a single match and a samflag of 3 indicates properly paired and oriented reads. Thus, we filtered for STAR-aligned reads with a MAPQ score of 255 and samflag of 3 (*samtools view* -f 3 -b -q 255). For bowtie2, we used parameters (-p 8 --very-sensitive --maxins 2000), while we used default parameters for bwa-mem, applying post-alignment filters for reads with MAPQ score of ≥ 30 and samflag 3. For all alignments, we removed duplicates (i.e., potential PCR artifacts) with *samtools rmdup* and *samtools fixmate* with parameter *(-n)*, and further filtered for reads mapping to autosomal chromosomes. In contrast to normalization strategies, the maxATAC models performed robustly across the alignment methods tested (**Fig. S9D**). We chose bowtie2 alignment for subsequent analyses.

Inference of ATAC-seq Tn5 sites and smoothing. The Tn5 transposase dimer inserts sequencing adapters with a strand specific bias that results in a 9bp sequencing extension [8,23,85], therefore, reads are shifted +4 on the (+) strand or -5 on the (-) strand so that the corresponding read ends are centered at the Tn5 cut site. We first converted the filtered BAM files to bed intervals using *BEDtools bamtobed* [86] and *awk (awk 'BEGIN {OFS = "\t"} ; {if ($6 == "+") print $1, $2 + 4, $2 + 5, $4, $5, $6; else print $1, $3 - 5, $3 - 4, $4, $5, $6})'*. In contrast to other pipelines [35,36], we retain both cut sites per fragment, to maximize coverage and ultimately prediction from, e.g., scATAC-seq of rarer cell types *in vivo*. We used a 1bp window around the inferred cut sites to generate a high-resolution cut site signal. Given our goal of applying maxATAC to scATAC-seq in addition to ATAC-seq at typical sequencing depths (~20 million reads), we smoothed the sparse Tn5 cut site signal to overcome noise due to under sampling. We found that extension of Tn5 insertion sites by +/- 20bp (*BEDtools slop*) performed well (on par with +/- 5 or 10bp and better than single-bp resolution, data not shown); the resulting 40bp window also corresponds to the ~38 bp wide Tn5 transposase dimer [85].

Extended blacklist. Initial testing of maxATAC models on scATAC-seq in GM12878 highlighted the need for an extended blacklist [42]. We discovered regions of extreme OMNI-

ATAC-seq signal that were not present in DNase-seq or scATAC-seq data and therefore likely indicative of platform-specific technical artifacts. For example, regardless of alignment method, high signal regions were detected in OMNI-ATAC-seq but not DNase-seq or scATAC-seq of the same cell type (GM12878); these corresponded to mitochondrial chromosome duplication events and regions of low mappability (**Fig. S8-9**). Thus, our final blacklist included (i) blacklisted regions from ENCODE data [42], (ii) centromeres, telomeres, and annotated gaps available from UCSC table browser [87] for hg38, (iii) regions ≥1kb with ≥ 90% sequence identity to chrM [88], and (iv) regions with low mappability on chr21 (**Table S5**). Inferred Tn5 cut sites within blacklisted regions were removed with *bedtools intersect*.

ATAC-seq normalization and signal tracks. For comparison of ATAC-seq signal tracks and combination of biological replicates, we scaled ATAC-seq signal per replicate to 20 million mapped reads (**RP20M**), a process involving signal conversion to BEDgraph interval coverage tracks (*bedtools genomecov*) using the scale factor, derived from **Eq. 1** and **2**, and parameters (-bg -scale *scale factor*). The scale factor is multiplied by the count at each position to yield RP20M normalization:

Eq. 1 $$RP20M\ scaled\ reads = \frac{count}{sequencing\ depth} \times 20{,}000{,}000$$

Eq. 2 $$scale\ factor = \frac{1}{sequencing\ depth} \times 20{,}000{,}000$$

For cell types with multiple replicates, we average the RP20M values across all available samples, using (pyBigWig; v. 0.3.18) to generate bigwig files.

For maxATAC input, we initially applied standard min-max normalization to the RP20M ATAC-seq signal tracks, scaling bp signal by the min and max across RP20M tracks:

Eq. 3 $$minmax_{P\%}(signal) = \frac{signal - min}{max_{P\%} - min},$$

where $\min$ corresponds to the minimum RP20M signal and $\max_{P\%}$ corresponds to the pth-percentile (highest) RP20M signal. Standard min-max normalization uses the absolute maximum or "$\max_{100\%}$". However, despite an ATAC-specific blacklist, regions of extreme ATAC-seq signal persisted in a cell-type-specific manner (**Fig S9**). Although often biological in nature (e.g., *TRIM37* locus in MCF-7 [89], **Fig. S9C**), these outlying signals interfered with min-max normalization of ATAC-seq and cross-platform performance on scATAC-seq (**Fig. S9D**). To improve robustness, we replaced the absolute max (p=100%) in **Eq. 3** with the 95th-percentile and 99th-percentile signals (p=95% or 99%). These strategies enabled high-quality prediction in initial tests of maxATAC on scATAC-seq in GM12878, while a standard max-min strategy did not (**Fig. S9D**). This normalization strategy was then independently tested on another scATAC-seq dataset [35] (**Fig. 3**).

Peak Calling. We called peaks with MACS2[81] to identify "regions of interest" for training (see **Model Training**) and TFBS prediction with PWMs (a key comparator). The Tn5 cut sites (per biological replicate, when available) served as input to MACS2. Our parameter settings (-f BED -shift=0 -ext=40 -keep-dup=all) center the signal over the Tn5 insertion, smooth by extension +/-20bp and ensure that each inferred Tn5 binding site contributes to the peak call. (We keep all duplicate Tn5 cut sites because PCR duplicates were removed in previous steps.) We retained ATAC-seq peaks per cell type if they met an FDR = 5% cutoff.

DNase-seq

DNase-seq for GM12878 was downloaded from ENCODE experiment ENCSR000EMT. Alignment files were downloaded for both biological replicates mapped to the hg38 genome. The BAM files were deduplicated with Picard tools MarkDuplicates then sorted and filtered with samtools (-F 512 -q 10). DNase cut sites were inferred by centering a 40 bp window on the 5' end of the read interval. Coverage tracks were generated as described in *ATAC-seq normalization and signal tracks.* MACS2 peaks were called using the cut sites for both biological replicates as described in *ATAC-seq Peak Calling*.

Model Architecture

The maxATAC models are deep dilated convolutional neural networks that predict transcription factor binding sites as a function of ATAC-seq and DNA sequence (**Fig. S2**). Model inputs are 1,024bp x 5, with four dimensions corresponding to one-hot encoded DNA sequence and the fifth corresponding ATAC-seq signal (processed as described above). The output of each TF model is a 32 x 1 array of TFBS predictions, resolved to 32bp. The CNN is composed of five convolutional blocks, each consisting of two repeating double-layers (ReLU-activated 1D convolutional operations) followed by batch normalization. The max-pooling layer is interspersed between convolutional blocks to reduce the spatial dimensions of the input. The kernel widths are fixed at 7 for all convolutional blocks, while the number of filters grows by a factor of 1.5 per block, from 15 (first block) to 75 (last block). The final output layer uses sigmoid activation for binary prediction of TFBS. The dilation rate of the convolutional filters increases from one, one, two, four, eight, and sixteen across blocks. As a result, the receptive field gradually expands to +/-512bp in the ultimate hidden layer. Thus, information is shared across the length of the 1024bp input, in a distance-dependent manner (i.e., proportional to spatial proximity of the regions), while the resolution of TFBS predictions is preserved at 32bp.

Model Training

Train, validation, and test sets

The goal of maxATAC is TFBS prediction from ATAC-seq in new cell types (i.e., "trans-cell type TFBS prediction). Thus, for a TF with TF ChIP-seq and ATAC-seq available from $N \geq 3$ cell types, N-1 cell types were used for training and validation, while the Nth was reserved for the test set. In addition, to avoid overfitting DNA sequence, the autosomal chromosomes were split into independent training (chromosomes 3, 4, 5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 17,18, and 20), validation (chr2 and chr19) and test (chr1 and 8) sets. Given the 32bp resolution of maxATAC predictions, positive examples of TFBS resulted if the 32bp region had >50% overlap with the set of ChIP-seq TFBS for a given cell type (detailed above).

Training routines

The maxATAC models were trained through optimization of the cross-entropy loss function via stochastic gradient descent using the ADAM optimizer [90], with an initial learning rate of 0.001, for 100 epochs (batch size of 1000 and 100 batches per epoch) and Glorot initialization of the weights [91]. Given the breadth of TFs in our benchmark, the diversity of their binding mechanisms and variable amounts of training data available per model, we used maximum validation dice coefficient to select model parameters (an epoch) per TF model**.**

As described in **Results**, we developed a sampling strategy for model training, to enrich true positive (**TP**) and challenging true-negative (**TN**) examples of TFBS (**Fig. S3**). Because only ~1% of the chromatin is expected to be accessible in a given cell type, for each TF model, we defined "regions of interest" as the union of accessible chromatin and TFBS (for that TF) for each

training cell type. Furthermore, to increase the number of challenging TN examples, we introduced "pan-cell" training, so that regions of interest for one cell type were equally likely to be selected from the other training cell type(s). Using a small subset of the benchmark data (11 TFs, limiting training to A549, HepG2, IMR-90, K562, MCF-7, and SK-N-SH cell lines, and test to GM12878), 100% peak-centric, pan-cell training outperformed commonly used random sampling from the genome [20] and mixed random and peak-centric strategies (e.g., 50-50, **Fig. S3A**). Thus, 100% peak-centric, pan-cell training was used for the final maxATAC models.

Following previous work[92], we doubled the number of training examples by including sequence and signal from the reverse-complement in addition to the forward strand. Although a part of our final maxATAC training routine, this strategy increased training time but did not robustly improve test performance (data not shown). For this reason, it is not the default for model training in the maxATAC codebase.

Performance evaluation

Precision-recall analysis

Genome-wide, the number of true positive (**TP**) TFBS are scarce relative to true negative (**TN**), unbound regions, so we used precision-recall statistics rather than receiver-operator characteristic (**ROC**) [32]. ROC weights TP and TN equally, and therefore is a less informative metric of performance in unbalanced classification problems, like TFBS prediction. We report area under precision-recall (**AUPR**) and precision at 5% recall.

We ranked maxATAC TFBS predictions for precision-recall analysis. maxATAC output is a score, ranging from 0 and 1, indicating the probability of a TFBS in each 32bp genomic interval. Each unique score is a unique rank. We benchmark our predictions by binning the signal from the validation or test chromosome(s) of interest using pyBigWig and report the max value per bin of length 200bp. (200bp was selected for comparison with the DREAM-ENCODE TFBS Prediction Challenge.) The ranking of our predictions is based on the maximum score in the 200bp signal region.

Blacklisted regions are excluded from precision-recall analysis. The ChIP-seq gold standard is binned into 200bp intervals, and a bin is labeled positive if any of the bin overlaps a ChIP-seq peak. For every rank, we calculate precision and recall relative to the ChIP-seq gold standard. We calculate the precision as the number of predictions that were found in the gold standard at each rank (**Eq. 4**). We calculate recall as the percent of the gold standard that was recovered at each rank (**Eq. 5**). Random precision is calculated as the number of bins overlapping the gold standard divided by the total number of bins evaluated (**Eq. 6**). AUPR calculations were implemented using the python package *sklearn* [93].

Eq. 4 $$precision = \frac{\#\ of\ bins\ with\ TF\ binding\ predictions\ overlapping\ ChIP{-}seq\ gold\ standard}{\#\ of\ bins\ with\ TF\ binding\ predictions}$$

Eq. 5 $$recall = \frac{\#\ of\ bins\ with\ TF\ binding\ predictions\ overlapping\ ChIP{-}seq\ gold\ standard}{\#\ of\ bins\ overlapping\ ChIP{-}seq\ gold\ standard}$$

Eq. 6 $$random\ precision = \frac{\#\ of\ bins\ overlapping\ ChIP{-}seq\ gold\ standard}{\#\ of\ bins\ across\ the\ chromosome}$$

For $\log_2$(fold-change) comparisons between AUPR or precision of different methods, we added a pseudocount of .1, to moderate high fold-changes due to small numbers from relatively poorer quality models (e.g., AUPRs < .1).

Comparison to other TFBS methods

TFBS prediction with PWM models. We obtained DNA sequences for ATAC-seq (or scATAC-seq) peaks identified per cell type (*bedtools getfasta*). We used the motif-matching algorithm MOODS[94] together with the TF PWM database CISBP v2[24] to identify motif occurrences with a $P < 10^{-5}$. For TFs with multiple PWM, we used all for our analysis, and, when multiple motif matches occurred within the same genomic region, we removed exact coordinate duplicates, but left overlapping motif matches. To rank TFBS predictions (e.g., at 200bp resolution), we binned the genome into set-width, non-overlapping bins using *bedtools makewindows*. For each bin, we counted the number of TF motif matches that overlapped the bin by at least 1 bp using *bedtools intersect*. We used the number of motif matches per bin to rank our predictions for precision-recall analysis.

TFBS prediction with Average Training ChIP-seq Signal. We used average ChIP-seq signal from the training cell types to predict TFBS in a held-out test cell line. Specifically, for each TF, we averaged the arcsinh of the ChIP-seq signal p-value across training cell types at each genomic position [95], using the averaged signal to rank TFBS for precision-recall analysis.

Comparison to current state-of-the-art deep-learning models for trans-cell type TFBS prediction. For the 8 TFs in common between the ENCODE-DREAM benchmark (DNase-seq train/test data, mapped to hg19) and the maxATAC benchmark (**Fig. 1B,** ATAC-seq train/test data, mapped to hg38), we compared test AUPRs (at 200bp resolution) for maxATAC on the maxATAC benchmark to test AUPRs (also 200bp resolution) of DeepGRN, FactorNet and Leopard on the ENCODE-DREAM benchmark, reported in their respective publications[19–21].

Prediction with the maxATAC Models

The final maxATAC models were constructed using all benchmark cell types available for a given TF (**Fig. 1B**), while maintaining train, validation, and test chromosomes, so that the test performance of these models could eventually be evaluated with new data (e.g., as we did with ATAC-seq and ChIP-seq from primary $CD4^+$ T cells, **Fig. 5**). In addition to publishing the final maxATAC models, we took advantage of the good correlation between validation and test performance (**Fig. 2C, S5C-D**) and mapped maxATAC scores to intuitive validation performance statistics precision, recall, $\log_2(\text{Precision} / \text{Precision}_{random})$, and F1-Score. Per TF model, the validation performance on Chr2 was averaged across each of the validation cell types. Given that precision is not necessarily a monotonic function of maxATAC score, we ensured one-to-one mapping in the following way: For a precision value mapping to multiple maxATAC scores, we selected the maxATAC score that maximized recall.

maxATAC evaluation on scATAC-seq, primary cells and in discovery mode

scATAC-seq

We evaluated maxATAC on scATAC-seq during (i) method development (**Fig. S9D**) and (ii) independent testing (**Fig. 3**). For method development, we downloaded scATAC-seq fragment files for GM12878 (500, 1k, 5k, and 10k cell experiments) from the 10X Genomics website (https://www.10xgenomics.com/resources/datasets) [12]. For testing, we downloaded the high-loading, mixed-cell line scATAC-seq experiment SRX9633387 from GSE162690 [35], which we processed to cell-type specific pseudobulk fragment files.

Fragment files were filtered for reads aligning to the hg38 genome. For each pseudobulk, Tn5 cut sites were identified, signal tracks generated and $\text{minmax}_{99\%}$-normalized for maxATAC prediction, as described for bulk ATAC-seq.

To assess the impact of pseudobulk library size of TFBS prediction from maxATAC and TF motif scanning (**Fig. 3E-F, S6**), we downsampled (*dplyr* (v1.0.8)) the GM12878 scATAC-seq

885 library (10k cells, from the 10x Genomics website). Resulting downsampled libraries were (1)
886 processed for maxATAC prediction or (2) peak calling and motif scanning (see above).
887
888 ***Primary cell types***
889
890 We downloaded ATAC-seq (GSE116696) and ChIP-seq data for 3 TFs (FOS:GSM3258569,
891 MYC:GSM3258570, and JUNB:GSM3258571) in primary $CD4^+$ T cells stimulated with anti-
892 CD3/anti-CD28 beads for 5 hours [41], processing ATAC-seq and ChIP-seq for maxATAC and
893 precision-recall analysis as described above.
894
895 ***Sequence-Specific TFBS prediction in $CD4^+$ T cells from atopic dermatitis patients***
896
897 We derived both DNA sequence and ATAC-seq signal inputs for maxATAC from a genetics atopic
898 dermatitis (**AD**) study [49], combining whole genome sequencing (**WGS**) and OMNI-ATAC-seq
899 (of peripheral blood CD4+ T-cells stimulated with anti-CD3 and anti-CD28 beads for 5 hours) for
900 6 AD patients and 6 age-matched controls. We used patient-specific genetic variant calls,
901 focusing on the nine AD risk variants previously associated with allele-dependent chromatin
902 accessibility identified by Eapen et al. We sought to identify pairs of AD patients and age-matched
903 controls that were homozygous for the risk and non-risk alleles, respectively. Focusing on
904 homozygous-risk and homozygous-nonrisk obviated the need for phasing (i.e., to resolve ATAC-
905 seq signal into maternal and paternal DNA strands). We focused on two independent AD risk loci
906 that met our criteria. For these loci, we applied maxATAC to patient- and control-specific DNA
907 sequence and ATAC-seq signal to make TFBS predictions in the haplotype blocks containing the
908 risk variants (linkage-disequilibrium $R^2 > .8$), for the patient pair, AD2 and CTL2. Haplotype blocks
909 were defined by the homozygous risk variants identified above and all SNPs in linkage
910 disequilibrium with the risk variant ($R^2 \geq .8$) were used for patient-specific sequence prediction
911 with maxATAC. The prediction window was defined by the furthest genetic variant positions in LD
912 with risk variant, extended by an additional + or - 512 bp to contain the most distal variants in LD.
913 We limited TFBS prediction to TFs with nominal mRNA expression (DESeq2 VST counts ≥ 9) in
914 activated $CD4^+$ T cells, measured in parallel RNA-seq by Eapen et al.
915
916 **Data Availability Statement**
917
918 The maxATAC codebase is available from https://github.com/MiraldiLab/maxATAC, including
919 basic usage (maxATAC installation, ATAC-seq data processing and TFBS prediction with the
920 trained maxATAC models) and advanced (model training and benchmarking). OMNI-ATAC-seq
921 data generated by this study is available as accession GSE197009 in the GEO Database. The
922 maxATAC benchmark is organized (QC'd, processed, ready-to-go) for community use from
923 https://doi.org/10.5281/zenodo.6761768.
924
925 **Funding**
926
927 This work was supported by the National Institute of Health [U01AI150748 to ERM, SP, MTW,
928 LCK; R01AI153442 to ERM, AB; R21AI156185 to ERM; R01HG010730, U01AI130830,
929 R01NS099068, R01GM055479, P01AI150585 to MTW; NIH R01AI024717, R01AR073228;
930 R01DK107502, R01AI148276, U19AI070235, U01HG011172, and P30AR070549 to LCK]; and
931 Cincinnati Children's Research Foundation [ARC Award 53632 to MTW, LCK; Center for Pediatric
932 Genomics grants to ERM, AB, MTW, LCK, SP].
933
934 **Acknowledgements**
935

We thank several sources of computational support for this work: the Biomedical Informatics Research IT Group at Cincinnati Children's Hospital Medical Center (P. Velayutham), the Ohio Supercomputer Center and the Advanced Research Computing Center at the University of Cincinnati Research. We acknowledge Christopher Benner, Shaun Mahony, Ivan Marazzi, Brad Rosenberg, Krishna Roskin and Andrea Toth for manuscript feedback and Wen Niu for help cross-referencing ChIP-seq experiments across databases. We used BioRender.com and LucidChart.com to generate some of our images.

Author Contributions

ERM conceived of the study and, with TAC, directed this work. VBSP and ERM supervised the construction of maxATAC models by TAC, FWR, BI and MK, with input from XC, LCK, AB, and MTW. TAC and FWR generated the results, with support and direction from BI, XC, MK, LCK, AB, MTW, VBSP and ERM. JAW, ATB aided TAC in scATAC-seq bioinformatic analyses. OD, BW, LCK and AB contributed experimental data. TAC, FWR, BI, XC, MK and ERM contributed to the codebase. TAC, SP, LCK, MTW and ERM contributed to data curation. TAC, FWR and ERM wrote the manuscript with input from all authors.

Competing Interests Statement

AB is a co-founder of Datirium, LLC.

957 **References**
958

959 1. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential
960 etiologic and functional implications of genome-wide association loci for human diseases
961 and traits. Proceedings of the National Academy of Sciences. 2009;106: 9362–9367.
962 2. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic
963 localization of common disease-associated variation in regulatory DNA. Science (1979).
964 2012; 1222794.
965 3. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and
966 epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518: 337–
967 343. doi:10.1038/nature13835
968 4. Harley JB, Chen X, Pujato M, Miller D, Maddox A, Forney C, et al. Transcription factors
969 operate across disease loci, with EBNA2 implicated in autoimmunity. Nature Genetics.
970 2018;50. doi:10.1038/s41588-018-0102-3
971 5. Davidson EH. Emerging properties of animal gene regulatory networks. Nature.
972 2010;468: 911–20. doi:10.1038/nature09645
973 6. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, et al. A map of open
974 chromatin in human pancreatic islets. Nat Genet. 2010;42: 255–259.
975 7. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible
976 chromatin landscape of the human genome. Nature. 2012;489: 75–82.
977 8. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native
978 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding
979 proteins and nucleosome position. Nat Methods. 2013;10: 1213–8.
980 doi:10.1038/nmeth.2688
981 9. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell
982 chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523: 486–
983 90. doi:10.1038/nature14590
984 10. Cusanovich D, Daza R, Adey A, Pliner H. Multiplex single-cell profiling of chromatin
985 accessibility by combinatorial cellular indexing. 2015 [cited 22 Apr 2017]. Available:
986 http://science.sciencemag.org/content/348/6237/910.short
987 11. Corces M, Buenrostro J, Wu B, Greenside P. Lineage-specific and single-cell chromatin
988 accessibility charts human hematopoiesis and leukemia evolution. Nature. 2016 [cited 22
989 Apr 2017]. Available:
990 http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3646.html
991 12. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, et al. Massively parallel
992 single-cell chromatin landscapes of human immune cell development and intratumoral T
993 cell exhaustion. Nat Biotechnol. 2019;37: 925–936.
994 13. Miraldi ER, Pokrovskii M, Waters A, Castro DM, de Veaux N, Hall J, et al. Leveraging
995 chromatin accessibility for transcriptional regulatory network inference in T Helper 17
996 Cells. Genome Res. 2019;29: 449–463.
997 14. Jackson CA, Castro DM, Saldi G-A, Bonneau R, Gresham D. Gene regulatory network
998 reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse
999 environments. Barkai N, Weigel D, editors. Elife. 2020;9: e51254.
1000 doi:10.7554/eLife.51254

1001 15. Consortium EP. An integrated encyclopedia of DNA elements in the human genome.
1002 Nature. 2012;489: 57.
1003 16. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of
1004 transcription factor binding from DNA sequence and chromatin accessibility data.
1005 Genome Res. 2011;21: 447–455.
1006 17. Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor
1007 binding. Genome Biology. 2019;20: 1–17. doi:10.1186/s13059-018-1614-y
1008 18. Li H, Quang D, Guan Y. Anchor: Trans-cell Type Prediction of Transcription Factor Binding
1009 Sites. 2019; 281–292. doi:10.1101/gr.237156.118.29
1010 19. Quang D, Xie X. FactorNet: a deep learning framework for predicting cell type specific
1011 transcription factor binding from nucleotide-resolution sequential data. Methods.
1012 2019;166: 40–47.
1013 20. Li H, Guan Y. Fast decoding cell type–specific transcription factor binding landscape at
1014 single-nucleotide resolution. Genome Research. 2021;31: 721–731.
1015 doi:10.1101/GR.269613.120
1016 21. Chen C, Hou J, Shi X, Yang H, Birchler JA, Cheng J. DeepGRN: prediction of transcription
1017 factor binding site across cell-types using attention-based deep neural networks. BMC
1018 Bioinformatics. 2021;22: 1–18.
1019 22. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription
1020 factor binding sites using ATAC-seq. Genome Biol. 2019;20: 45. doi:10.1186/s13059-019-
1021 1642-2
1022 23. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to
1023 ATAC-seq data analysis. Genome Biol. 2020;21: 1–16.
1024 24. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription
1025 Factors. Cell. 2018;172: 650–665. doi:10.1016/j.cell.2018.01.029
1026 25. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal
1027 for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res.
1028 2016; gkw983.
1029 26. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et
1030 al. An improved ATAC-seq protocol reduces background and enables interrogation of
1031 frozen tissues. Nat Methods. 2017;14: 959–962. doi:10.1038/nmeth.4396
1032 27. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of
1033 DNA- and RNA-binding proteins by deep learning. Nature Biotechnology. 2015;33: 831–
1034 838. doi:10.1038/nbt.3300
1035 28. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-
1036 based sequence model. Nature Methods. 2015;12: 931–934. doi:10.1038/nmeth.3547
1037 29. Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J. Sequential regulatory
1038 activity prediction across chromosomes with convolutional neural networks. Genome
1039 Res. 2018;28: 739–750.
1040 30. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image
1041 Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected
1042 CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;40: 834–848.
1043 doi:10.1109/TPAMI.2017.2699184

1044 31. 2 Chromatin patterns at transcription factor binding sites. Nature. 2019.
1045 doi:10.1038/nature28171
1046 32. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot
1047 when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10:
1048 e0118432.
1049 33. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. ChromVAR: Inferring transcription-factor-
1050 associated accessibility from single-cell epigenomic data. Nature Methods. 2017;14: 975–
1051 978. doi:10.1038/nmeth.4401
1052 34. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of
1053 lineage-determining transcription factors prime cis-regulatory elements required for
1054 macrophage and B cell identities. Mol Cell. 2010;38: 576–589.
1055 35. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a
1056 scalable software package for integrative single-cell chromatin accessibility analysis. Nat
1057 Genet. 2021;53: 403–411.
1058 36. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis
1059 with Signac. Nature Methods 2021 18:11. 2021;18: 1333–1341. doi:10.1038/s41592-021-
1060 01282-5
1061 37. Schreiber J, Bilmes J, Noble WS. Completing the ENCODE3 compendium yields accurate
1062 imputations across a variety of assays and human biosamples. Genome Biol. 2020;21: 1–
1063 13.
1064 38. Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, et al. Exploring
1065 Massive, Genome Scale Datasets with the GenometriCorr Package. PLOS Computational
1066 Biology. 2012;8: e1002529-. Available: https://doi.org/10.1371/journal.pcbi.1002529
1067 39. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al.
1068 Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.
1069 Cell. 2014;158: 1431–1443. doi:10.1016/j.cell.2014.08.009
1070 40. Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-
1071 specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia
1072 evolution. Nature Genetics. 2016;48: 1193–1203. doi:10.1038/ng.3646
1073 41. Yukawa M, Jagannathan S, Vallabh S, Kartashov A v, Chen X, Weirauch MT, et al. AP-1
1074 activity induced by co-stimulation is required for chromatin opening during T cell
1075 activation. J Exp Med. 2020;217: 647388. doi:10.1084/jem.20182009
1076 42. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic
1077 regions of the genome. Sci Rep. 2019;9: 1–5.
1078 43. Ou J, Liu H, Yu J, Kelliher MA, Castilla LH, Lawson ND, et al. ATACseqQC: a Bioconductor
1079 package for post-alignment quality assessment of ATAC-seq data. BMC Genomics.
1080 2018;19: 169.
1081 44. Hu B, Jadhav RR, Gustafson CE, le Saux S, Ye Z, Li X, et al. Distinct age-related epigenetic
1082 signatures in CD4 and CD8 T cells. Front Immunol. 2020;11.
1083 45. Richards AL, Muehlbauer AL, Alazizi A, Burns MB, Findley A, Messina F, et al. Gut
1084 microbiota has a widespread and modifiable effect on host gene regulation. MSystems.
1085 2019;4: e00323-18.

1086 46. Kelley DR, Snoek J, Rinn JL. Basset: Learning the regulatory code of the accessible
1087 genome with deep convolutional neural networks. Genome Research. 2016;26: 990–999.
1088 doi:10.1101/gr.200535.115
1089 47. Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, Strachan DP, et al. Multi-
1090 ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies
1091 new risk loci for atopic dermatitis. Nature Genetics. 2015;47: 1449–1456.
1092 doi:10.1038/ng.3424
1093 48. Brown SJ. What have we learned from GWAS for atopic dermatitis? Journal of
1094 Investigative Dermatology. 2021;141: 19–22.
1095 49. Eapen AA, Parameswaran S, Forney C, Edsall LE, Miller D, Donmez O, et al. Epigenetic and
1096 Transcriptional Dysregulation in CD4+ T cells of Patients with Atopic Dermatitis. bioRxiv.
1097 2021; 2021.12.03.471059. doi:10.1101/2021.12.03.471059
1098 50. Sun L-D, Xiao F-L, Li Y, Zhou W-M, Tang H-Y, Tang X-F, et al. Genome-wide association
1099 study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han
1100 population. Nat Genet. 2011;43: 690–694.
1101 51. Chen CC, Yang YH, Lin YT, Hsieh SL, Chiang B-L. Soluble decoy receptor 3: increased levels
1102 in atopic patients. Journal of allergy and clinical immunology. 2004;114: 195–197.
1103 52. Park I, Son M, Ahn E, Kim Y-W, Kong Y-Y, Yun Y. The Transmembrane Adaptor Protein
1104 LIME Is Essential for Chemokine-Mediated Migration of Effector T Cells to Inflammatiory
1105 Sites. Mol Cells. 2020;43: 921.
1106 53. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo
1107 J, et al. Impact of genetic polymorphisms on human immune cell gene expression. Cell.
1108 2018;175: 1701–1715.
1109 54. Bossini-Castillo L, Glinos DA, Kunowska N, Golda G, Lamikanra A, Spitzer M, et al.
1110 Immune disease variants modulate gene expression in regulatory CD4+ T cells and inform
1111 drug targets. bioRxiv. 2019; 654632. doi:10.1101/654632
1112 55. Soskic B, Cano-Gamez E, Smyth DJ, Ambridge K, Ke Z, Bossini-Castillo L, et al. Immune
1113 disease risk variants regulate gene expression dynamics during CD4+ T cell activation.
1114 bioRxiv. 2021; 2021.12.06.470953. doi:10.1101/2021.12.06.470953
1115 56. Garaud S, Roufosse F, de Silva P, Gu-Trantien C, Lodewyckx J-N, Duvillier H, et al. FOXP1 is
1116 a regulator of quiescence in healthy human CD4+ T cells and is constitutively repressed in
1117 T cells from patients with lymphoproliferative disorders. European Journal of
1118 Immunology. 2017;47: 168–179. doi:https://doi.org/10.1002/eji.201646373
1119 57. Dias S, D'Amico A, Cretney E, Liao Y, Tellier J, Bruggeman C, et al. Effector Regulatory T
1120 Cell Differentiation and Immune Homeostasis Depend on the Transcription Factor Myb.
1121 Immunity. 2017;46: 78–91. doi:https://doi.org/10.1016/j.immuni.2016.12.017
1122 58. Guo YA, Chang MM, Skanderup AJ. MutSpot: detection of non-coding mutation hotspots
1123 in cancer genomes. npj Genomic Medicine. 2020;5: 26. doi:10.1038/s41525-020-0133-4
1124 59. Liu EM, Martinez-Fundichely A, Diaz BJ, Aronson B, Cuykendall T, MacKay M, et al.
1125 Identification of Cancer Drivers at CTCF Insulators in 1,962 Whole Genomes. Cell
1126 Systems. 2019;8: 446-455.e8. doi:https://doi.org/10.1016/j.cels.2019.04.001
1127 60. Zhu H, Uusküla-Reimand L, Isaev K, Wadi L, Alizada A, Shuai S, et al. Candidate Cancer
1128 Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction

Networks. Molecular Cell. 2020;77: 1307-1321.e10. doi:https://doi.org/10.1016/j.molcel.2019.12.027

61. Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell. 2021;184: 5053–5069.

62. Minnoye L, Taskiran II, Mauduit D, Fazio M, van Aerschot L, Hulselmans G, et al. Cross-species analysis of enhancer logic using deep learning. Genome Res. 2020;30: 1815–1834.

63. Maslova A, Ramirez RN, Ma K, Schmutz H, Wang C, Fox C, et al. Deep learning of immune cell differentiation. Proceedings of the National Academy of Sciences. 2020;117: 25655–25666.

64. Thibodeau A, Khetan S, Eroglu A, Tewhey R, Stitzel ML, Ucar D. CoRE-ATAC: A deep learning model for the functional classification of regulatory elements from single cell and bulk ATAC-seq data. PLoS Computational Biology. 2021;17: e1009670.

65. Lal A, Chiang ZD, Yakovenko N, Duarte FM, Israeli J, Buenrostro JD. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun. 2021;12: 1–11.

66. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18: 1196–1203.

67. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription factor binding reveal soft motif syntax. Nature Genetics. 2020.

68. Setty M, Leslie CS. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. PLoS Comput Biol. 2015;11: e1004271.

69. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell. 2018;174: 744–757.

70. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proceedings of the National Academy of Sciences. 2017;114: E4914–E4923. doi:10.1073/pnas.1704553114

71. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero Predicts cis -Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular Cell. 2018; 858–871. doi:10.1016/j.molcel.2018.06.044

72. Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:171010903. 2017.

73. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017. pp. 1025–1035.

74. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics. 2017;18: 551–562.

75. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Research. 2019;47: D729–D735. doi:10.1093/nar/gky1094

76. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics. 2011;5: 1752–1779. doi:10.1214/11-AOAS466
77. M lder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake [version 2; peer review: 2 approved] . F1000Res. 2021;10. doi:10.12688/f1000research.29032.2
78. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583: 699–710. doi:10.1038/s41586-020-2493-4
79. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923
80. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10: giab008. doi:10.1093/gigascience/giab008
81. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008;9: R137. doi:10.1186/gb-2008-9-9-r137
82. Krueger F, James F, Ewels P, Afyounian E, Schuster-Boeckler B. FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo. 2021 [cited 3 Jan 2022]. doi:10.5281/ZENODO.5127899
83. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012/10/25. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635
84. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nature Communications. 2020;11: 4267. doi:10.1038/s41467-020-18035-1
85. Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biology. 2010;11: R119. doi:10.1186/gb-2010-11-12-r119
86. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033
87. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32: D493–D496. doi:10.1093/nar/gkh103
88. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R v, Schwartz S, et al. Recent segmental duplications in the human genome. (Reports). Science (1979). 2002;297: 1003+. Available: https://link-gale-com.uc.idm.oclc.org/apps/doc/A90752870/BIC?u=ucinc_main&sid=summon&xid=d4c143dc
89. Hampton OA, den Hollander P, Miller CA, Delgado DA, Li J, Coarfa C, et al. A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. Genome Res. 2008/12/03. 2009;19: 167–177. doi:10.1101/gr.080259.108
90. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

91. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Teh YW, Titterington M, editors. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Sardinia, Italy: PMLR; 2010. pp. 249–256. Available: https://proceedings.mlr.press/v9/glorot10a.html
92. Kopp W, Monti R, Tamburrini A, Ohler U, Akalin A. Deep learning for genomics using Janggu. Nat Commun. 2020;11: 1–7.
93. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of machine learning research. 2011;12: 2825–2830.
94. Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics. 2009;25: 3181–3182.
95. Schreiber J, Durham T, Bilmes J, Noble WS. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. Genome Biology. 2020;21: 81. doi:10.1186/s13059-020-01977-6

Supplemental Table Legends

Table S1. **Training data curated for 127 transcription-factor models from GEO and ENCODE**. We tabulate sample IDs of ChIP-seq and ATAC-seq experiments from ENCODE, used for training maxATAC models. This list only includes information for TF and cell type combinations that had at least two cell types for training.

Table S2. Performance metrics for 74 transcription-factor models in bulk ATAC-seq. Performance metrics generated from maxATAC prediction for each cell-type and transcription-factor pair. Results include maxATAC test AUPR for chr1, TF motif scanning AUPR for chr1, and average training ChIP-seq signal AUPR for chr1.

Table S3. Performance metrics for 193 cell type and transcription factor models in scATAC-seq data, including maxATAC AUPR for chr1 in bulk ATAC-seq, maxATAC AUPR for chr1 in scATAC-seq, and TF motif scanning AUPR for chr1 in scATAC-seq for all 193 models tested.

Table S4. Atopic dermatitis risk loci with allele-dependent ATAC-seq signal. Summary of the variants associated with allele-dependent ATAC-seq signal and in linkage disequilibrium with the (tagged) atopic dermatitis risk variant. Each genetic variant is annotated with the patient genotype, Linkage Disequilibrium R^2 value, and gene expression information for genes near these variants.

Table S5. Extended maxATAC blacklist. This BED file (hg38 coordinates) contains the blacklisted regions that were excluded from our analysis. These regions include the low mappability arm of chr21, segmental duplications with high sequence similarity to chrM, telomeres, centromeres, and annotated gaps.

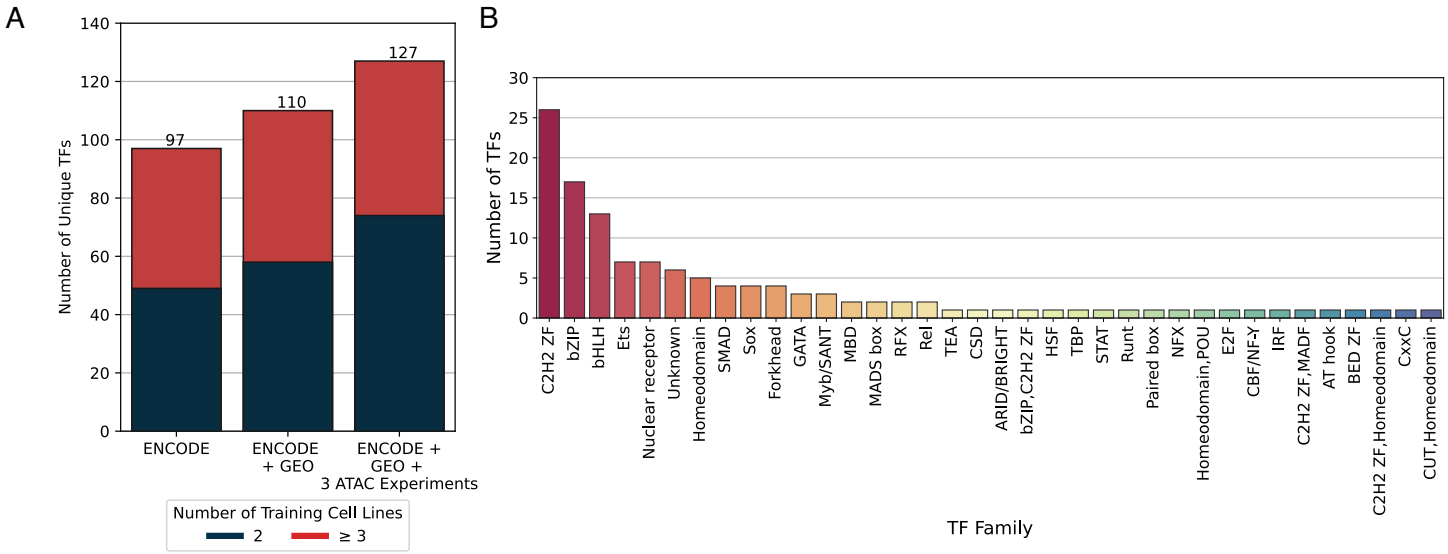


Fig. S1. maxATAC benchmark statistics. (A) The number of unique TF models (y-axis) that can be trained using different combinations of ENCODE, GEO, and in-house generated OMNI ATAC-seq data (x-axis). TFs are broken into two categories: (1) those with only 2 cell lines available (only cross-cell type training is feasible, black bar) and (2) models that have ≥ 3 cell lines (cross-cell type training *and* performance evaluation in a held-out test cell type are feasible, red bar). **(B)** The distribution of maxATAC TF models across TF families. The "unknown" category contains TFs that have not been associated with a TF family.

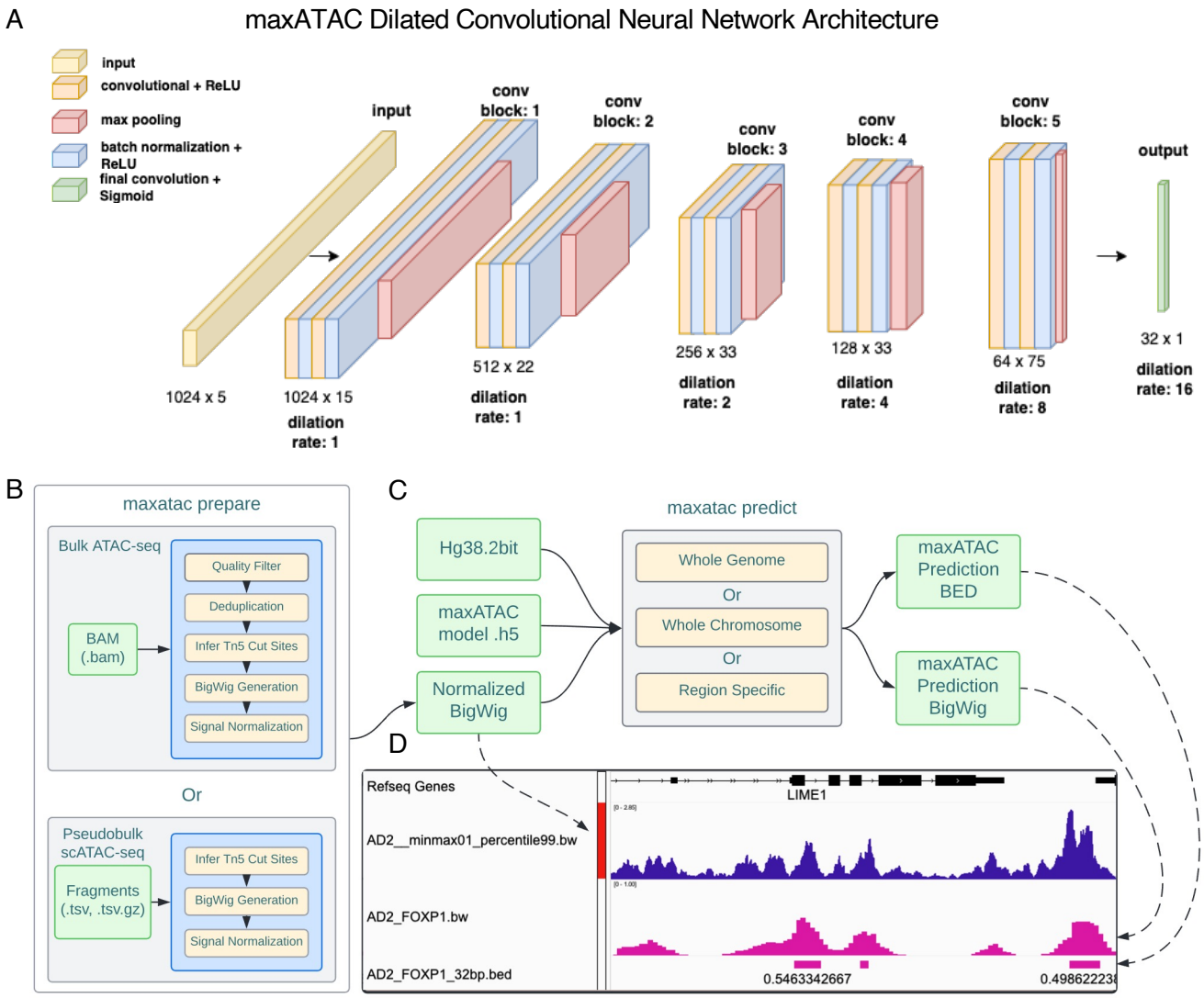


Fig. S2. maxATAC model architecture, inputs, and standard workflow. **(A)** maxATAC inputs are a 1,024 bp one-hot encoded DNA-sequence with ATAC-seq signal for the corresponding region, while maxATAC output is an array of 32 TFBS predictions at 32 bp resolution, spanning the 1024 bp input sequence interval. Inputs go through a total of 5 convolutional blocks. Each convolutional block consists of two layers, each consisting of ReLU activated 1D convolutional operations and batch normalization. A max pooling layer is interspersed between 2 convolutional blocks to reduce the spatial dimensions of the input. The kernel width is fixed at 7 across all convolutional blocks. The model uses 15 filters in the first convolutional block, and the number of filters is increased by a factor of 1.5 for every subsequent block. The final output is produced by a single sigmoid-activated convolutional layer. As a dilated convolutional neural network, the rate of dilation is one, one, two, four, eight, and sixteen from convolutional block 1 to the penultimate convolutional layer. Increasing the dilation rate increases the receptive field, so that spatially distant regions share information. In this network, the receptive field grows to +/-512bp, with information sharing proportional to spatial proximity. **(B)** Schematic overview of a standard maxATAC workflow. maxATAC takes as input a BAM file or scATAC-seq fragments TSV file that is processed to Tn5 cut sites, smoothed and converted to a read-depth-normalized ATAC-seq signal track (robustly min-max normalized between 0-1, see **Methods**). **(C)** The maxATAC predict function takes as input the genome reference DNA 2bit sequence file, a trained maxATAC model h5 file and the normalized ATAC-seq signal track to predict TFBS. **(D)** The outputs of maxATAC are a bigwig file of maxATAC TFBS scores, ranging 0-1, and a BED file that corresponds to regions of TFBS, according to a user-selected confidence threshold (e.g., precision, F1-score, see **Methods**).

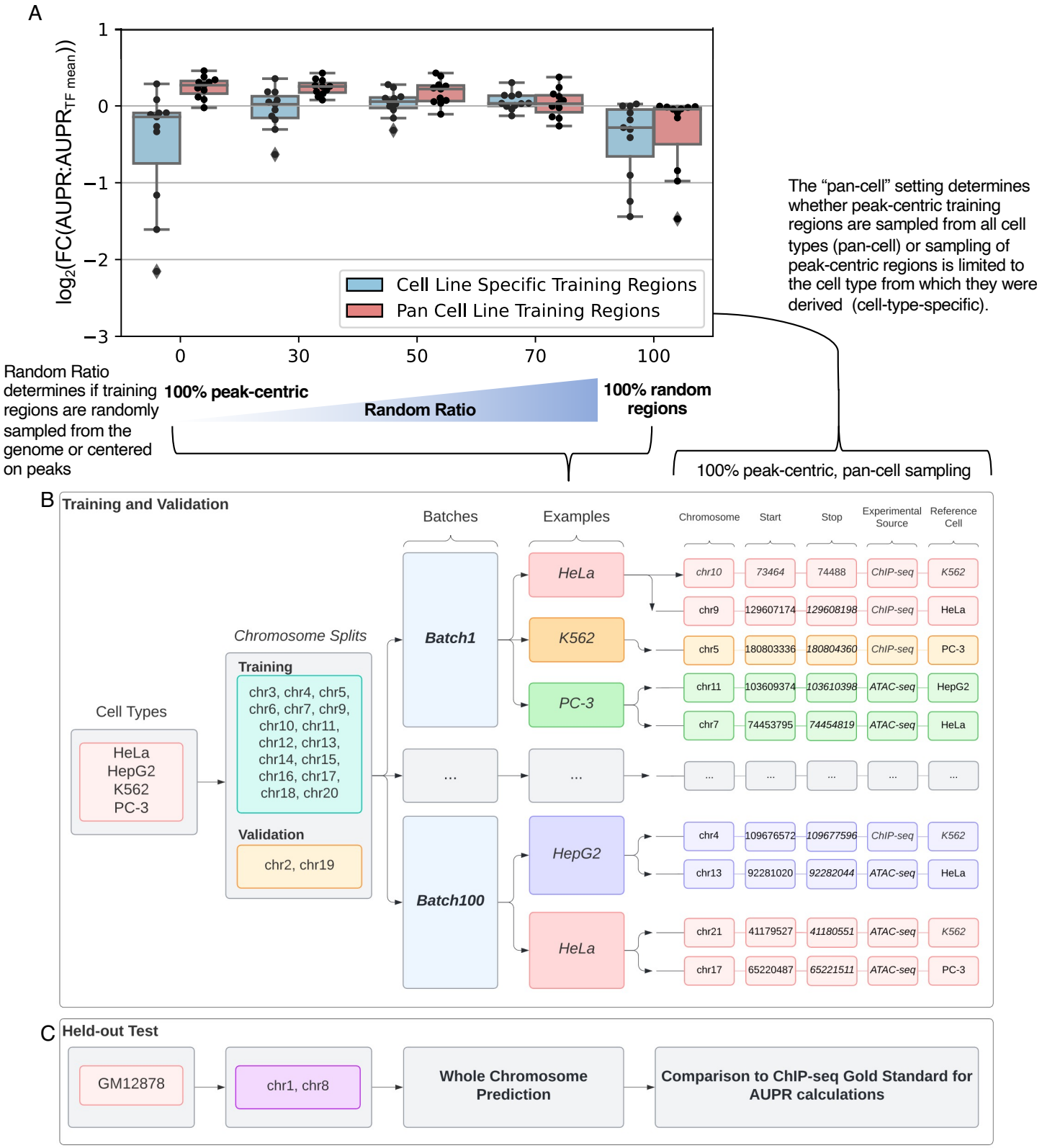


Fig. S3. maxATAC "pan-cell-type, peak-centric" training approach. **(A)** A small subset of our benchmark (11 TFs) was used to test peak-centric, pan-cell training (**Methods**) in a held-out cell type, GM12878, as a function of training strategy. We varied the relative ratio of random and peak-centric regions ("Random Ratio") sampled during model training. Random ratio of 100% indicates that all training examples were randomly sampled from the genome, while 0% indicates that only peak-centric regions (centered on ChIP-seq and ATAC-seq peaks in the training cell types) were used. "Peak-centric" training enriches for TP examples and potentially challenging TN (high ATAC-seq signal but no TFBS). Furthermore, "pan-cell" training (red) additionally enriches for challenging TN examples, by pooling peak-centric regions across the training cell types. During training, those regions are randomly sampled using signal from any of the training cell types (which may or may not have a TFBS in a given peak-centric region). This is in contrast to "cell type-specific" training (blue), in which peak-centric regions are specific to each of the training cell types and sampling of a peak-centric region is limited to signal from the cell types from which the peak-centric region was originally identified. **(B)** shows an example of 100% peak-centric, pan-cell training. (Note how peak-centric examples from HeLa come from both K562 and HeLa.) **(C)** The held-out test cell type, in addition to chr1 and chr8, are independent of model training and thereby enable performance evaluation of maxATAC model predictions in new cell types.

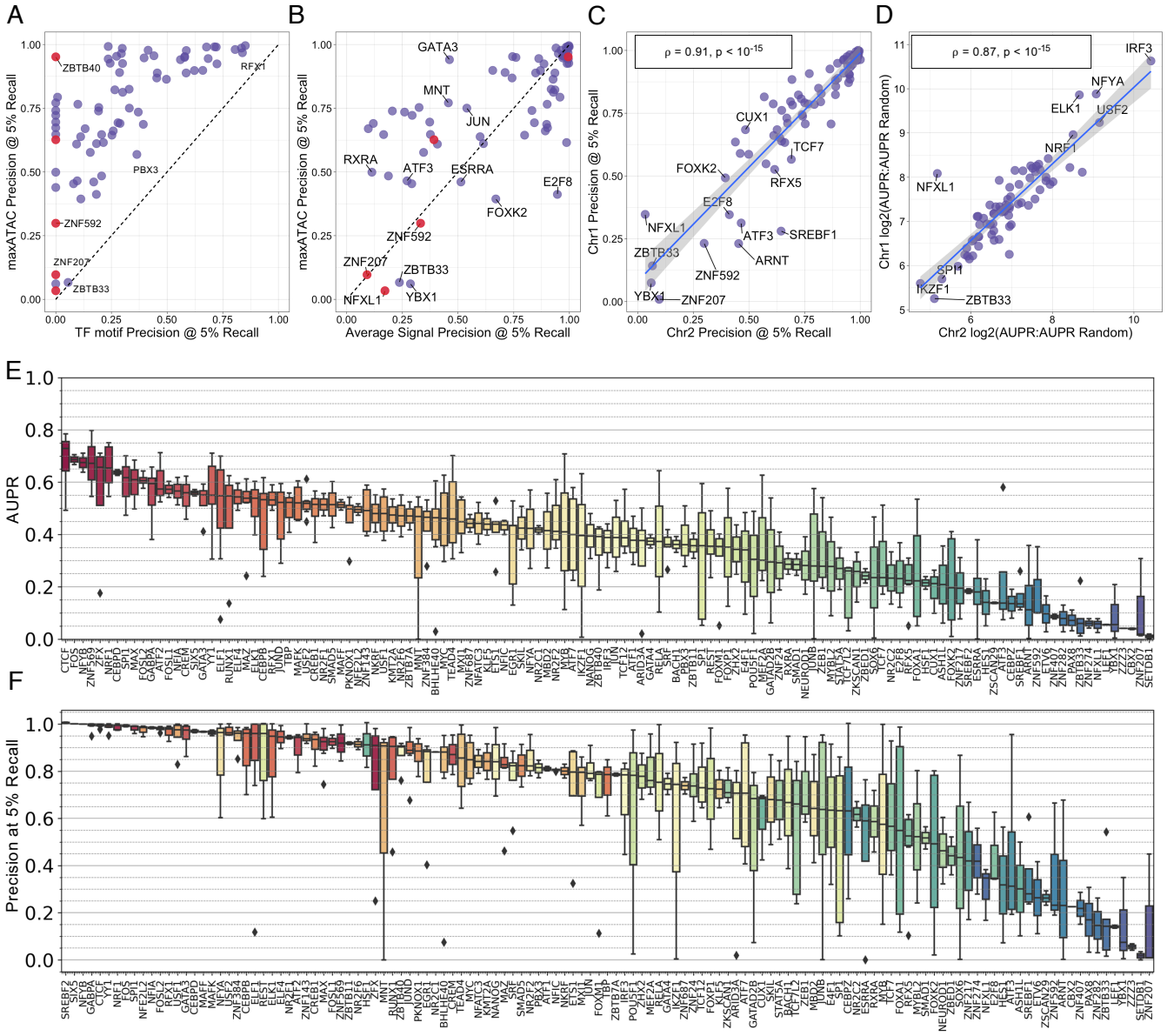


Fig. S4. Extended model performance statistics. Comparison of median test precision at 5% recall between maxATAC TFBS predictions and **(A)** TF motif scanning or **(B)** average training ChIP-seq signal**,** where median is the median performance across each possible train-test cell type split. **(C)** Comparison of median test versus validation precision at 5% recall for maxATAC models constructed using all train-test cell type splits. Test corresponds to the test cell type and chromosome (chr1), while validation corresponds to training cell types and chromosome (chr2); ρ = 0.91 and P < 10^{-15}, n = 74 TFs. **(D)** Test versus validation $\log_2$(AUPR:$AUPR_{RANDOM}$) performance; ρ = 0.87 and P < 10^{-15}. Validation **(E)** AUPR (median = 0.43) and **(F)** precision at 5% recall (median = 0.85) for the final 127 maxATAC TF models.

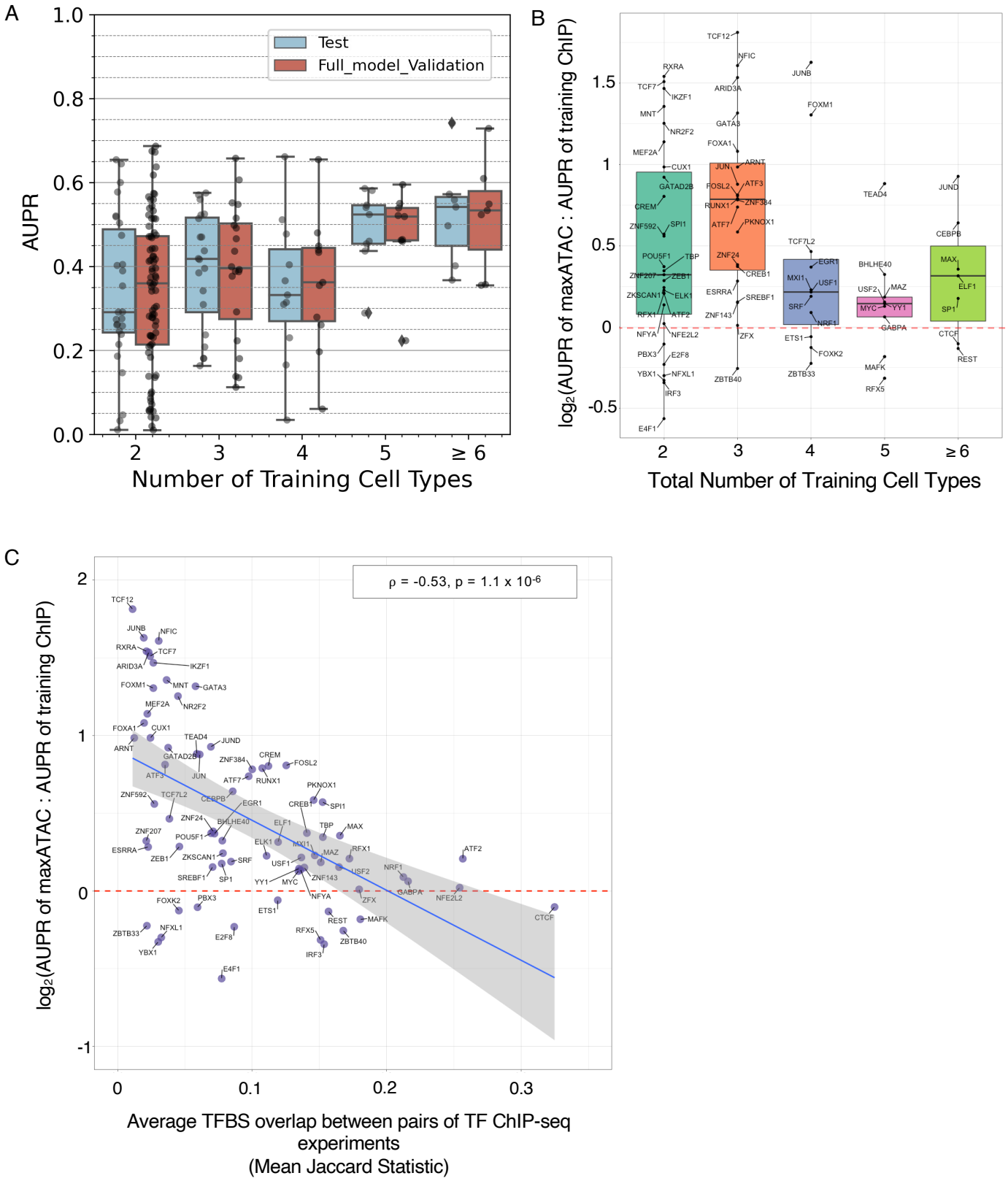


Fig. S5. Factors contributing to maxATAC model performance. (A) $AUPR_{MEDIAN}$ of maxATAC models as a function of number of cell types available for training. Validation performance (red) is estimated for the final maxATAC models (using all cell types available) for model construction, n = 127 TFs, while test performance (blue) is available for 74 TF models. **(B-C)** Factors contributing to relative test performance differences between maxATAC and average training ChIP-seq signal (74 TFs, analysis of median test AUPR)**. (B)** Log2-ratio of maxATAC AUPR relative to the AUPR of TFBS prediction using the averaged training ChIP-seq signal as a function of training cell types. **(C)** Log2-fold-change, of maxATAC AUPR relative to the AUPR of TFBS prediction using training ChIP-seq signal, versus averaged Jaccard overlap between TFBS (from ChIP-seq) in pairs of training cell types.

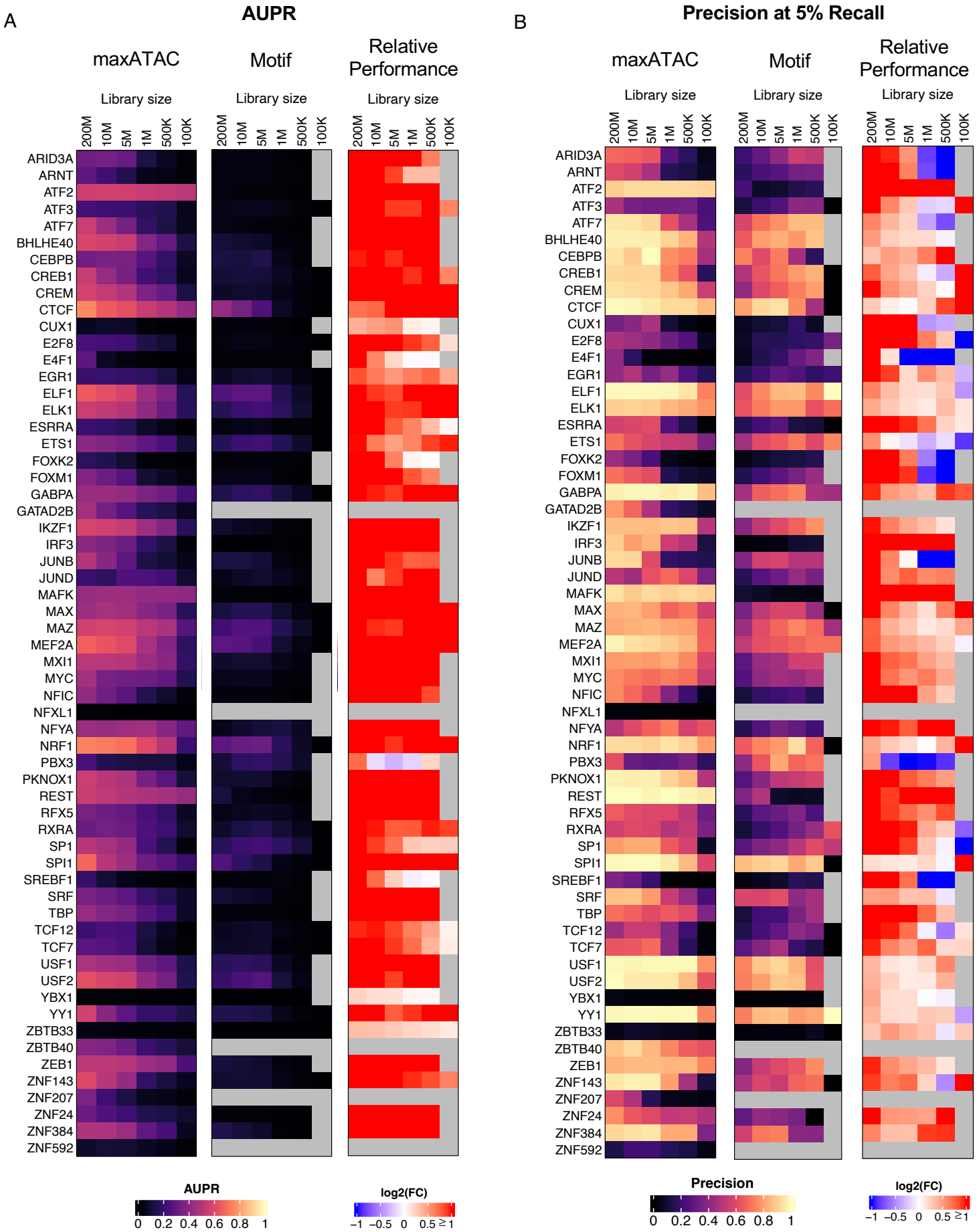
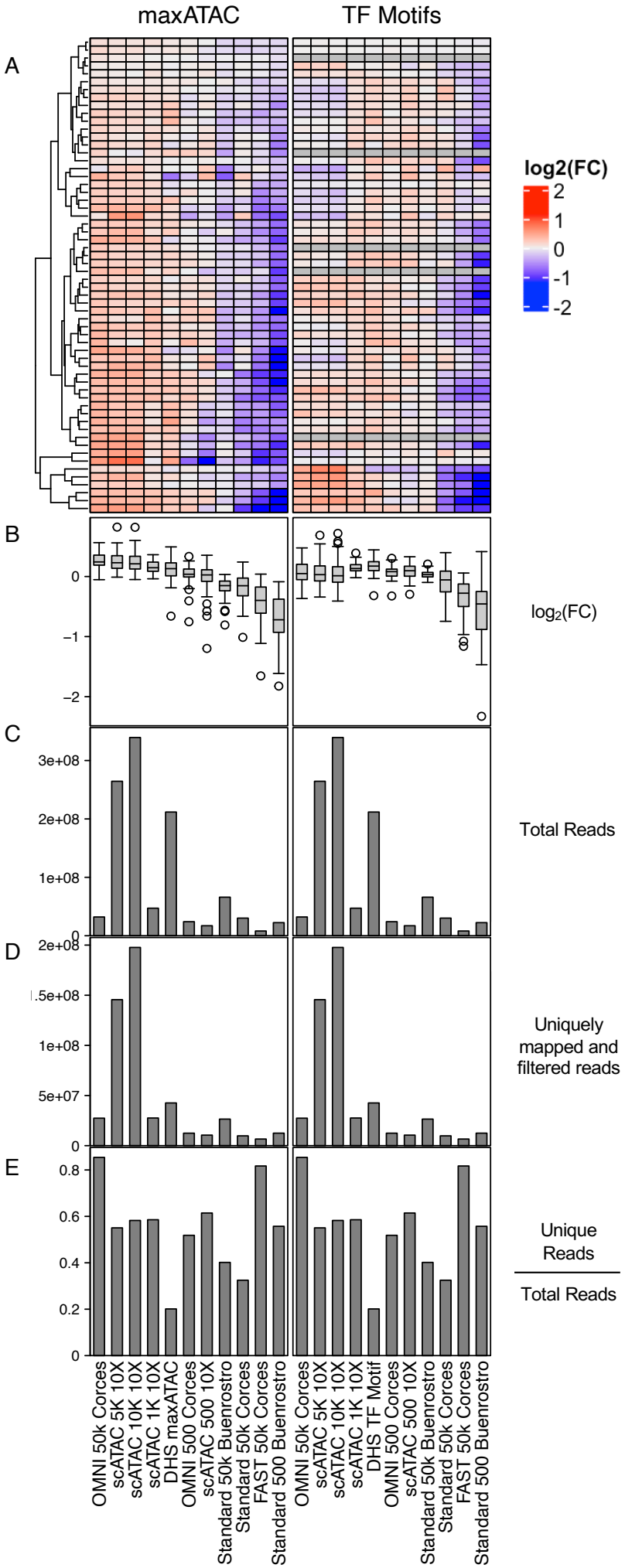


Fig. S6. The impact of scATAC-seq pseudobulk library size on TFBS prediction for maxATAC or TF motif scanning in accessible chromatin regions. scATAC-seq libraries for GM12878 were down-sampled from 200M to 100k fragments, and test **(A)** AUPR and **(B)** precision at 5% recall performances evaluated. Log2(FC) indicates maxATAC performance relative to TF motif scanning for 60 TFs. Gray boxes indicate either (1) no known motif (e.g., GATAD2B, NFXL1, ZBTB40, ZNF207, ZNF592) or (2) no motif predictions for accessible chromatin regions detected at the given library size.

Fig. S7. Experimental factors related to maxATAC performance on scATAC-seq and other ATAC-seq protocols. (A) Log_2(AUPR:$AUPR_{mean\ per\ TF\ and\ TFBS\ method}$) normalized across ATAC-seq protocols per TF, for maxATAC and TF motif scanning separately. **(B)** The distribution of log_2(AUPR:$AUPR_{mean\ per\ TF\ and\ TFBS\ method}$). **(C)** Total reads across biological replicates (when available) for each experimental protocol. **(D)** Total paired-end reads that uniquely map across biological replicates. **(E)** Mean proportion of uniquely mapping reads across biological replicates per protocol.



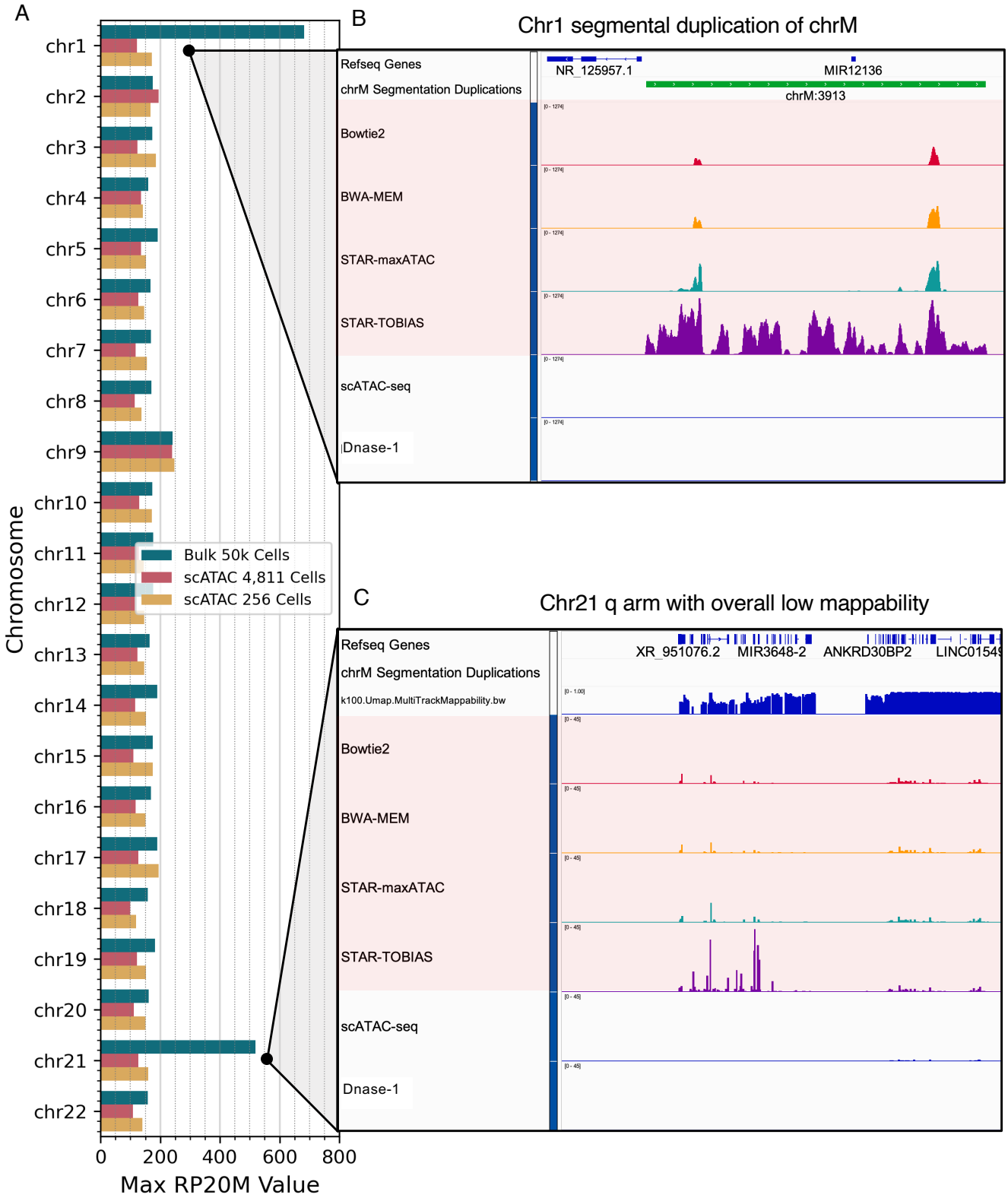


Fig. S8. An extended blacklist is needed for ATAC-seq. (A) The maximum number of Tn5 cut site counts per autosomal chromosome for bulk OMNI-ATAC-seq (blue), 10x scATAC-seq for 500 cells (yellow), and 5,000 cells (red) in GM12878. **(B)** IGV screenshot of a mitochondrial chromosome (**chrM**) segmentation duplication locus (green bar). While chromatin accessibility measurement by DNase-seq and scATAC-seq have no signal in this region, this is a region of extreme signal for OMNI-ATAC-seq, regardless of alignment methods. Pink tracks indicate OMNI-ATAC-seq mapped with multiple alignment methods (**Methods**). **(C)** IGV screenshot of the largely unmappable q-arm of chr21 and the relatively high signal caused by low mappability. Again, DNase-seq and scATAC-seq have no signal in this region, while OMNI-ATAC-seq has extreme signal, regardless of alignment strategy. These observations motivated our extended blacklist (**Methods**).

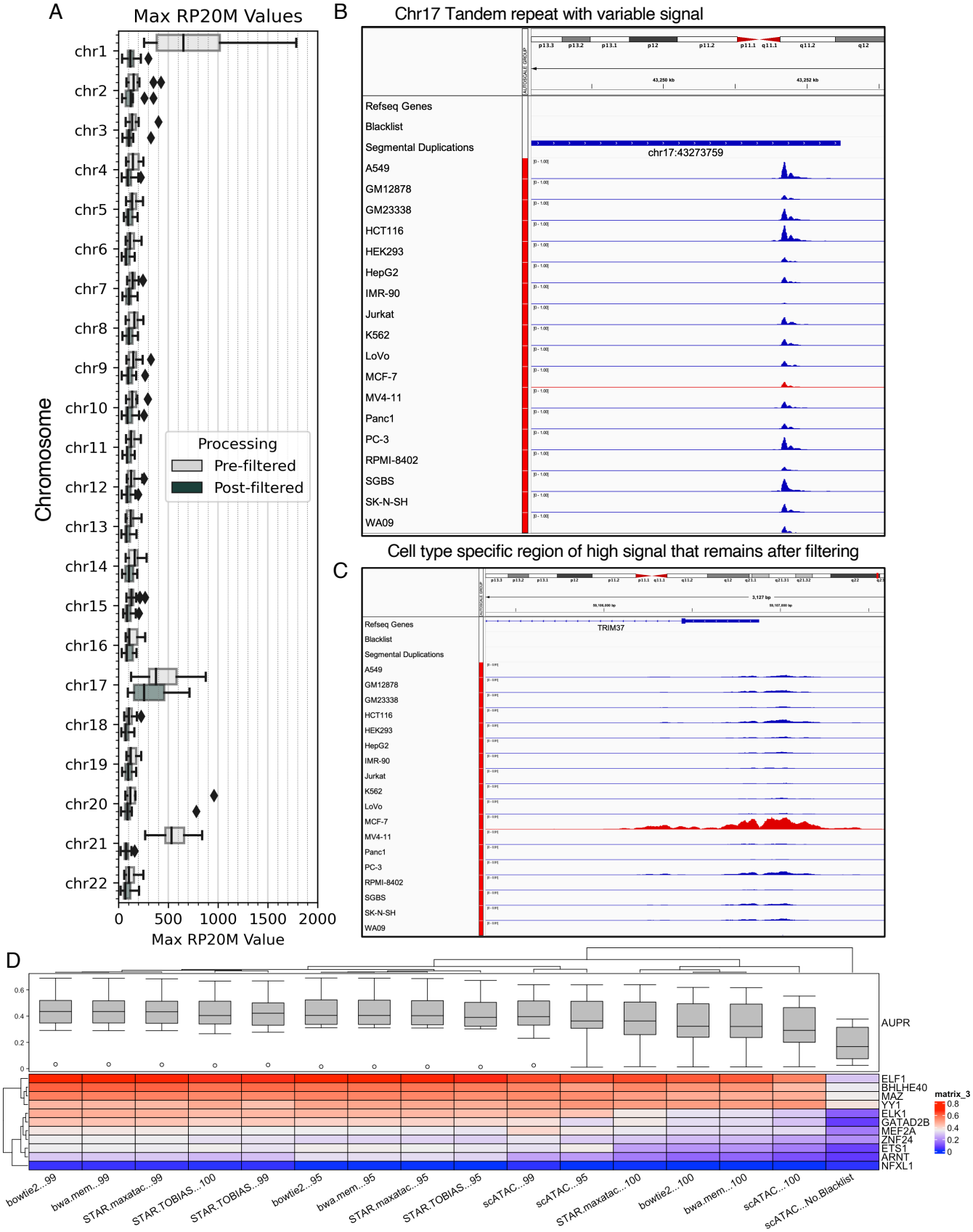


Fig. S9. ATAC-seq-specific normalization is required to for robust prediction across ATAC-seq protocols. (A) Max RP20M value per autosomal chromosome before and after filtering extended blacklist regions. **(B)** An example tandem repeat region with variable signal found on Chr17. **(C)** The *TRIM37* locus on Chr17 exhibits extreme, biologically relevant signal in the breast cancer cell line MCF-7 (red track). **(D)** Test AUPR in GM12878 for different alignment and normalization strategies (**Methods**) for bulk ATAC-seq or scATACseq (pseudobulk of 5k GM12878 cells). PCR duplicates and Tn5 cut sites that mapped to the extended blacklist (**Methods**) were removed prior to normalization. Test data is bulk ATAC-seq unless described as "scATAC". "95", "99", and "100" correspond to minmax normalization to the 95th, 99th or 100th-percentile highest ATAC-seq signal. "100" therefore corresponds to standard minmax normalization to the absolute max; this strategy was not robust to outlying ATAC-seq signal and therefore performs poorly when applied to different ATAC-seq alignment strategies or scATAC-seq. The far-right column represents performance on scATACseq data using (1) standard min-max normalization and (2) without applying the extended blacklist; this strategy has the worst performance generalizability to scATAC-seq.

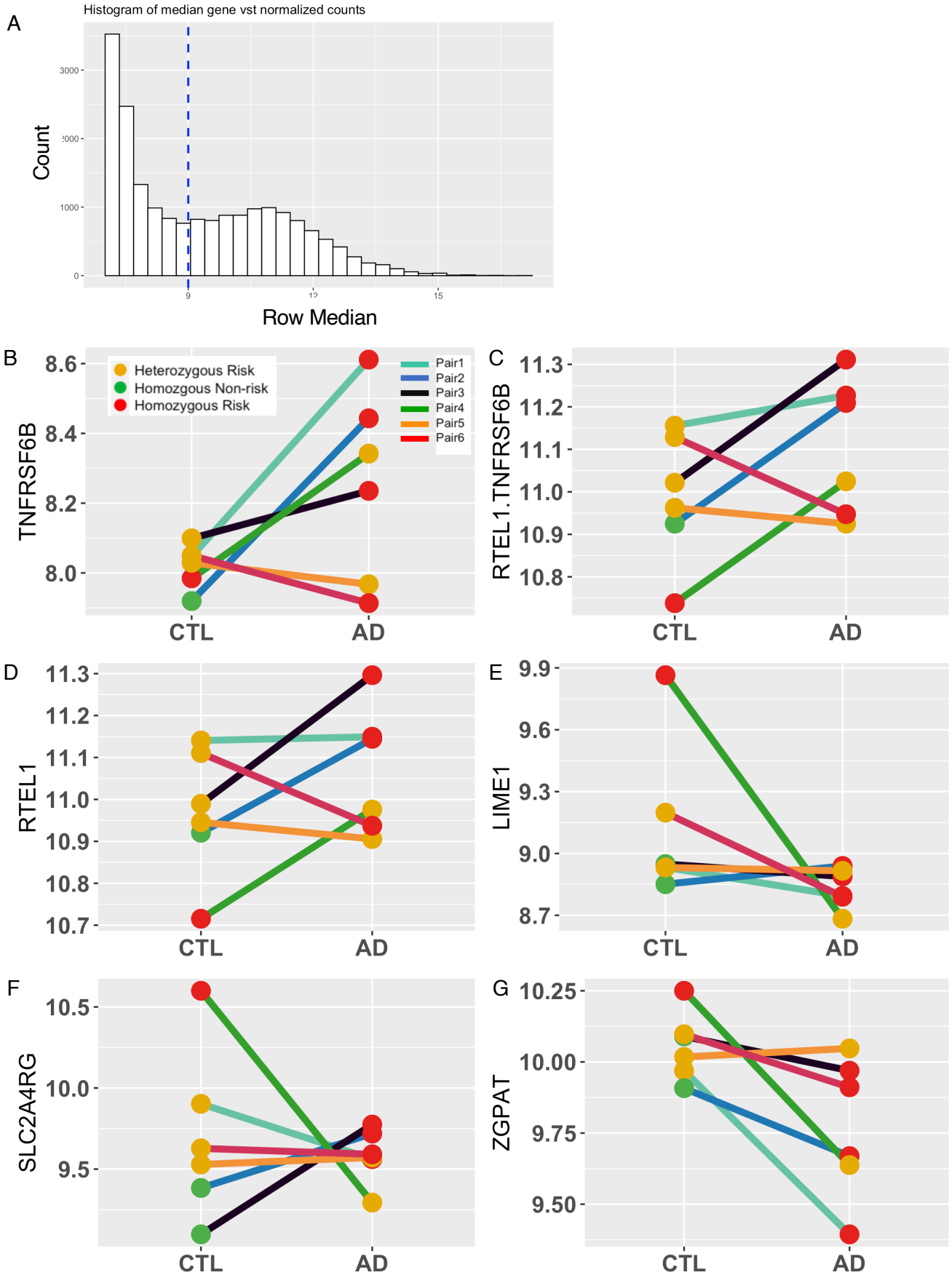


Fig. S10. **Gene expression in AD patients and aged-matched controls (A)** Median expression for each gene (DESeq2 VST-normalized counts) in the activated T cells RNA-seq dataset (6 AD patients and 6 age-matched controls). Blue dotted line indicates the nominal gene expression cutoff applied. Paired line plots showing the difference in gene expression between AD patients and their age-matched controls for **(B)** *TNFRSF6B,* **(C)** *RTEL1-TNFRSF6B,* **(D)** *RTEL1,* **(E)** *LIME1,* **(F)** *SLC2A4RG,* and **(G)** *ZGPAT*. Each point is colored according to whether the donor was homozygous risk (red), heterozygous risk (yellow), and homozygous non-risk (green). Each line is colored according to the donor pair.

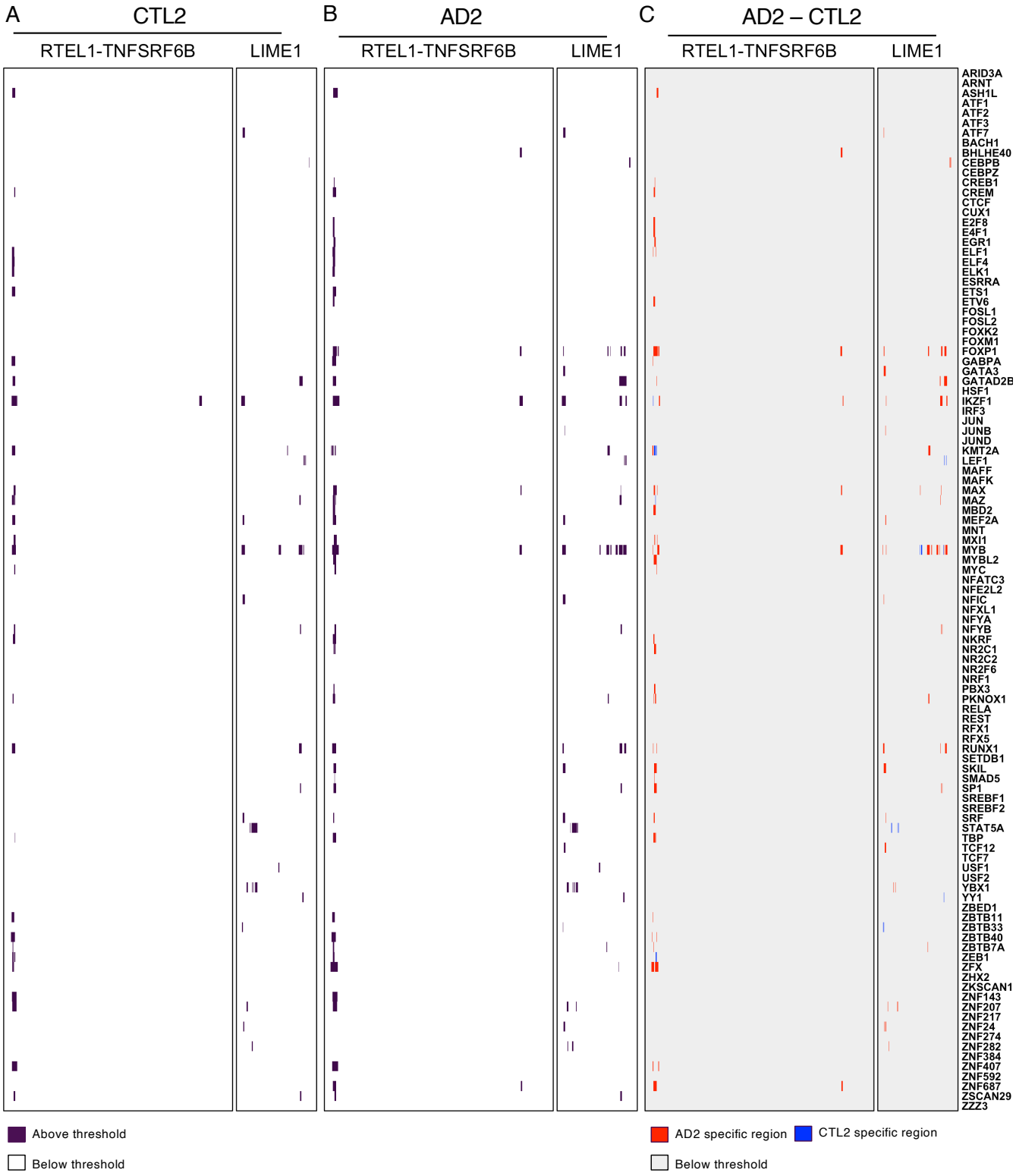


Fig. S11. maxATAC predictions for 105 TFs in AD2 and CTL2. Heatmaps show TFBS predictions (32bp width) for **(A)** CTL2 and **(B)** AD2, using a score cutoff that maximizes the average F1-score across validation cell types. **(C)** Differential TFBS between AD2 and CTL2, where red indicates AD2-specific prediction, blue indicates CTL2-specific prediction and grey denotes no difference.