bioRxiv preprint doi: https://doi.org/10.1101/2022.01.28.478235; this version posted January 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Tareian Cazares[1], Faiz W. Rizvi[2], Balaji Iyer[3,4], Xiaoting Chen[5], Michael Kotliar[6], Joseph A. Wayman[7], Anthony Bejjani[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[3,7,9,4]*

[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

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.28.478235; this version posted January 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. Yet, while Assay-for-Transposase-Accessible-Chromatin (ATAC)-seq datasets grow exponentially, suboptimal motif scanning is commonly used for TFBS prediction from ATAC-seq. Here, we present “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 largest collection of state-of-the-art TFBS models to date. maxATAC performance extends to primary cells and single-cell ATAC-seq, enabling state-of-the-art TFBS prediction *in vivo*. We demonstrate maxATAC’s capabilities by identifying TFBS associated with allele-dependent chromatin accessibility at atopic dermatitis genetic risk loci.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.28.478235; this version posted January 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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**)[5] at genome-scale. 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. This easy-to-use, popular technique provides high-resolution chromatin accessibility with low sample input requirements[6]. Thanks to advances in single-cell (**sc**)ATAC-seq, it is now possible to computationally resolve the chromatin accessibility profiles of individual cell types from heterogeneous tissues and limited clinical samples[7–9]. Whether from single cells or "bulk" populations, integration of TF-binding predictions from ATAC-seq improves GRN inference[10,11]. 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[12]. 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[13]. In 2017, the "ENCODE-DREAM *in vivo* TFBS Prediction Challenge" established two top-performing TFBS prediction algorithms[14,15] that vastly improved performance over popular motif scanning (median area under precision-recall .4 versus .1).

Yet these top-performing TFBS methods are rarely used in popular ATAC-seq analysis pipelines[16–20]. Several factors are to blame: (1) The coverage of top-performing TFBS models is poor relative to motif-scanning. State-of-the-art models exist for fewer than 30 TFs, while there are motifs available for at least 1200 of the ~1600 human TFs[21]. (2) Most models for TFBS prediction from chromatin accessibility were trained on DNase-seq rather than ATAC-seq. Although both data types provide high-resolution accessibility data, there are notable differences between the technologies[22], and the DNase-seq-trained models[14,15,23,24] have yet to be tested on ATAC-seq inputs. Thus, despite the promise of ATAC-seq for GRN inference and human genetics, the quality of TFBS prediction from ATAC-seq remains primitive, even as ATAC-seq data generation grows exponentially in both basic and biomedical research.

To enable state-of-the-art TFBS prediction from ATAC-seq, we built "**maxATAC**", a collection of top-performing, 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 state-of-the-art 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 state-of-the-art TFBS capabilities to rare 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.28.478235; this version posted January 29, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results:

maxATAC models offer state-of-the-art TFBS prediction from ATAC-seq at genome scale

State-of-the-art TFBS prediction methods use supervised learning and thus generally require paired TF-binding (e.g., ChIP-seq) and chromatin accessibility data in at least three cell types for benchmarking (2 for training, 1 to test generalizability in a new cell type, **Methods**). We identified existing ChIP-seq and ATAC-seq data from cistromeDB[25] and ENCODE[12] (**Fig. 1A**, **Methods**). We generated new OMNI-ATAC-seq data for three cell lines with abundant available TF ChIP-seq, enabling benchmarking for 74 TF models (≥3 cell types) and model construction for 127 TFs ≥2 cell types, **Fig. 1B**).

Deep convolutional neural networks (**CNNs**) provide state-of-the-art performance for many sequenced-based prediction tasks, including prediction of TFBS[26–28]. 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. 1C, S1A**, **Methods**). We utilize dilated convolutional layers to capture spatially distant relationships across the input sequences in a multiscale manner[28]. Our maxATAC models are thus capable of high resolution TFBS prediction (at 32bp) using information-sharing between proximal sequence and accessibility signals (+/-512bp).

Given our objective to construct state-of-the-art 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. Challenging TN, for example, might arise when a TFBS in one cell type is a TN (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[29], for each TF model, we defined "regions of interest" as the union of accessible chromatin and TFBS (e.g., ATAC-seq and ChIP-seq peaks) for each training cell type. We found that increasing the representation of ATAC-seq and TFBS in training examples (relative to randomly selected genomic regions) improved performance, and 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 and, in that way, the training examples are enriched for challenging TN examples (see **Methods**). Peak-centric, pan-cell training outperformed random sampling, with particularly strong gains for smaller training dataset sizes (**Fig. S1B**).

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[30] for each of the 74 "benchmarkable" TFs (area under precision recall (**AUPR**) or precision at 5% recall, **Fig. 2, S2, Table S1**).

The maxATAC models offer state-of-the-art 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. maxATAC outperformed standard motif scanning for every TF (**Fig. 2A, S2A**). For the vast majority of TFs, maxATAC also performed favorably to TFBS prediction using the average ChIP-seq signal for that TF across the training cell types (**Fig. 2B, S2B**). Comparison to this important null model[31] ensured that the maxATAC models learned ATAC-seq and

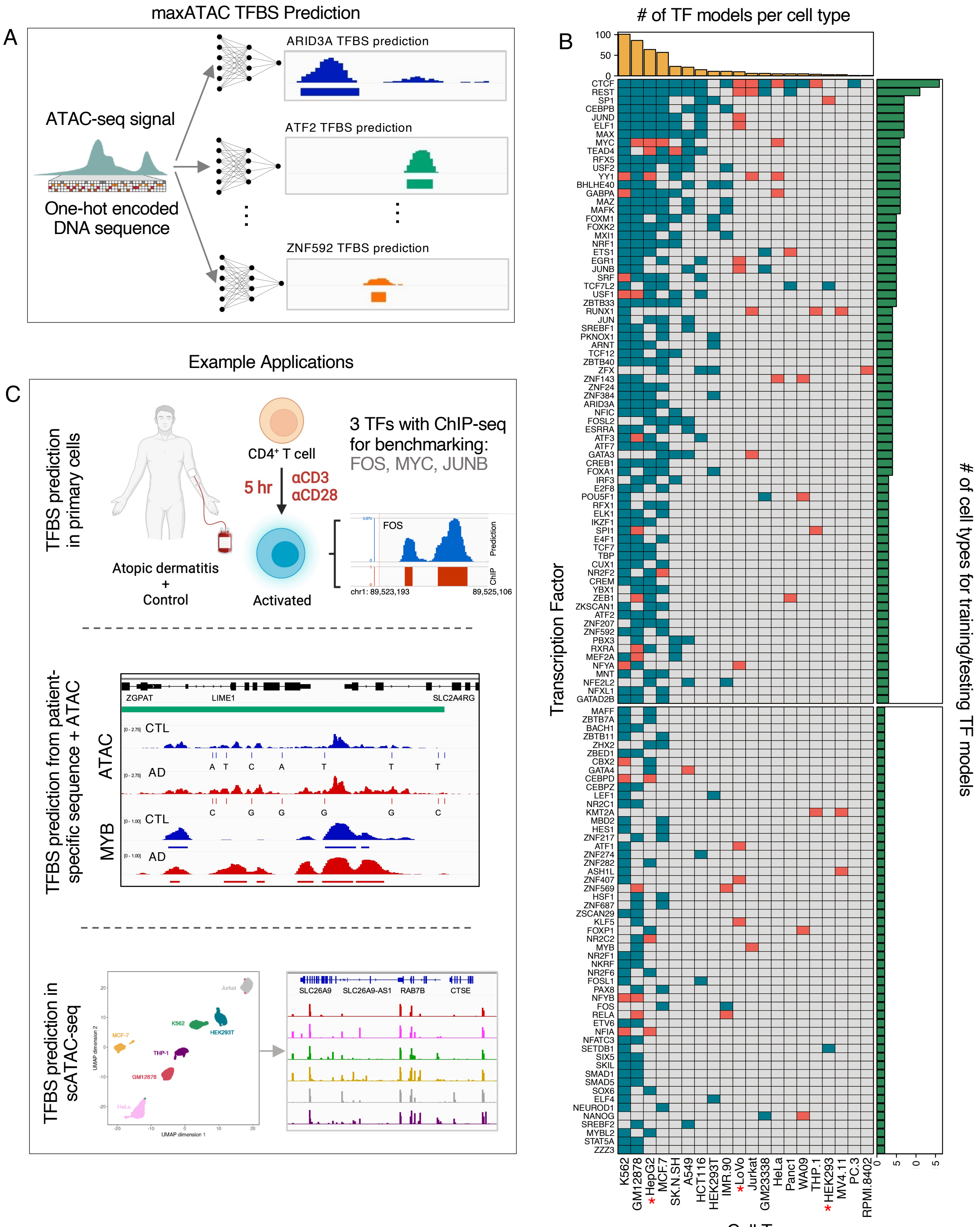

Fig. 1. Overview of maxATAC. (A) Trained maxATAC models use DNA sequence and ATAC-seq signal to predict TFBS in new cell types. **(B)** The maxATAC training data per TF and cell type with ATAC-seq (top: 74 "benchmarkable" TF models with ≥ 3 cell types available, bottom: 53 TF models with only 2 cell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

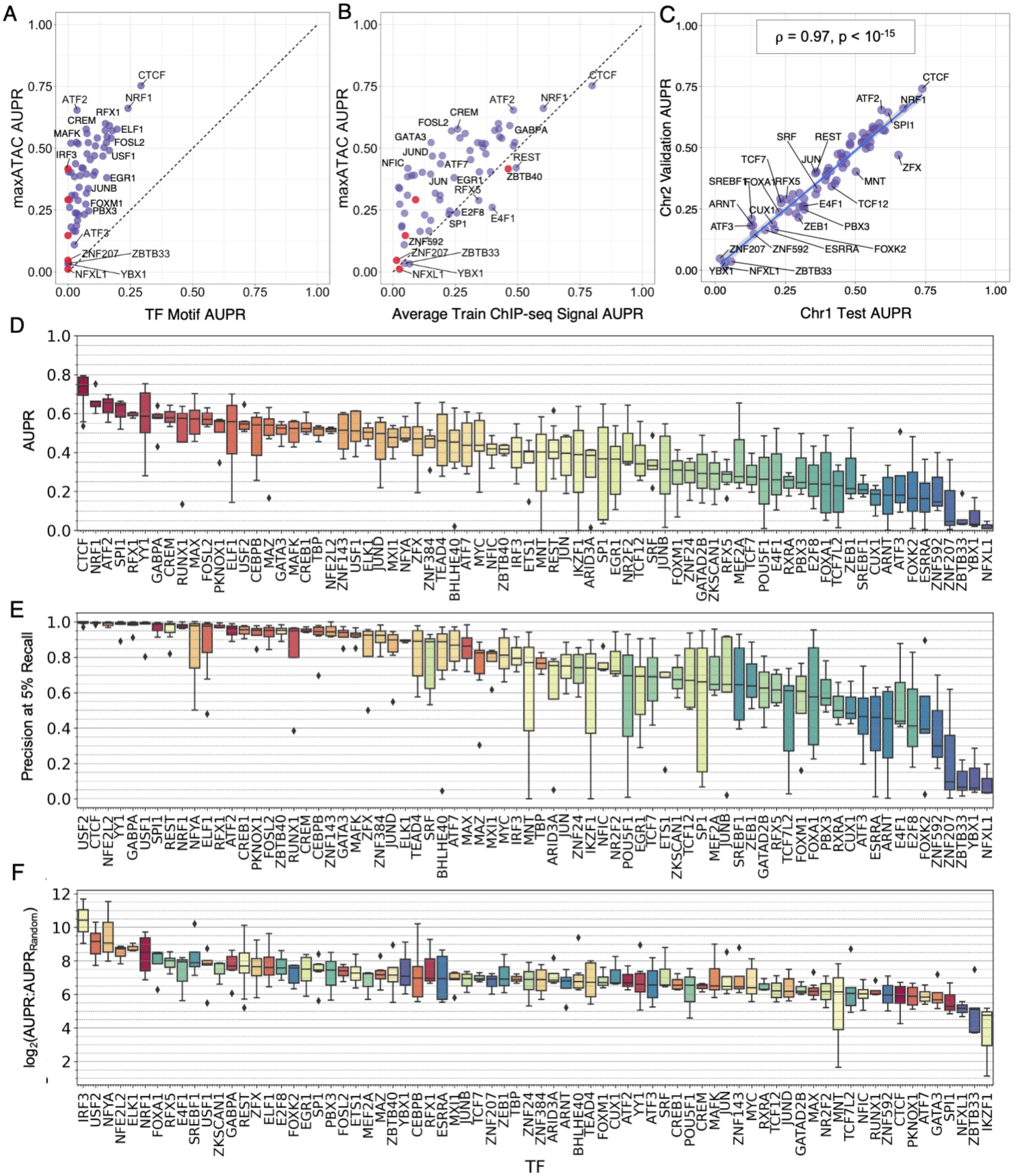

Fig. 2. The maxATAC models offer state-of-the-art TFBS prediction from ATAC-seq. For every TF model, one cell type and two chromosomes (chr1, chr8) were held out during training to assess predictive (test) performance in a new cell type. maxATAC model performance is compared to **(A)** TF motif-scanning in ATAC-seq peaks and **(B)** TFBS prediction using the averaged ChIP-seq signal from the training cell types; each dot represents $\mathrm{AUPR_{MEDIAN}}$ across train-test cell type splits. Red dots indicate TFs with no known motifs in CIS-BP. **(C)** Test performance ($\mathrm{AUPR_{MEDIAN}}$) on chr1 (held-out test cell type) as a function of validation performance ($\mathrm{AUPR_{MEDIAN}}$) on chr2 (training cell types) (n=74; ρ_{Pearson}=0.97, $P < 10^{-15}$). Boxplots of test **(D)** AUPR (median = 0.43), **(E)** precision at 5% recall (median = 0.85), and **(F)** $\log_2(\mathrm{AUPR:AUPR_{RANDOM}})$ (median = 6.87) per model, where each train-test cell type split yields a model and performance estimate, for a total of 74 TFs.

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 9 TFs in common). Acknowledging these caveats, we report that maxATAC performance, at $\text{AUPR}_{\text{MEDIAN}}$ = .43, is roughly on par with the top-performing models at the ENCODE-DREAM Challenge ($\text{AUPR}_{\text{MEDIAN}}$ = .44[15] or .41[14]) and a state-of-the-art method for TFBS prediction from ATAC-seq **(Fig. 2D**, **Table S1).**

Given this good performance, 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 to estimate test performance), 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, S2C-D**). Given the strongly predictive relationship between validation and test performance, the final maxATAC models were constructed using all data. 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*[32], initial, noisy TFBS predictions from ATAC-seq are subsequently refined by gene expression modeling, so a GRN modeler might prioritize high recall over high precision. 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).

Uncovering determinants of performance for the maxATAC TF models

Test (**Fig. 2D-E**) and validation (**Fig. S2E-F**) performance varied dramatically across TFs, with $\text{AUPR}_{\text{median}}$ ranging from .75 for CTCF to .01 for NFXL1 (**Fig. 2D**). We, thus, explored potential factors that could explain the performance disparities across TF models. 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. S3A-B**). While the interquartile range for models trained with 5 cell types was above the overall median AUPR performance (.40), 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**).

We also visualized model performances per TF family (**Fig. S3C-D**). The 127 maxATAC TF models span 25 TF families, with 1-13 TF models per family. Ets and bZIP family models had above-average performances across metrics. Extension of maxATAC to construct models for new TFs will improve representation per family and help resolve these emerging trends.

There was a strong relationship between model test (or validation) performance and the number of TFBS in the test (or validation) cell types (**Fig. S3D-E**, Pearson correlation = .61 or .60, $P<10^{-15}$ for both test and validation comparisons). Test cell types with more TFBS will have a higher AUPR, because, just by random chance, it is more likely that a genomic region is a positive example (TFBS). Thus, the number of positive examples (TFBS) in a test cell type directly influences AUPR, and it is helpful to account for this background by calculating random precision or AUPR (**$\mathbf{AUPR_{RANDOM}}$.**, **Eqn. 6**, **Methods**). As a complementary metric to compare performances across TFs, we therefore also report $\log_2$-foldchange of AUPR relative to $AUPR_{RANDOM}$ (**Fig. 2F, S2G**). By this metric, the top-ranked models by test AUPR (CTCF, NRF1, ATF2) fell to 67th, 6th, and 45th, respectively (**Fig. 2F**). By $\log_2(AUPR{:}AUPR_{RANDOM})$, performance ranged from a median ~1,300-fold over background (IRF3) to ~26-fold (IKZF1).

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 ChIP-seq experiments available, suggesting a role for TF-intrinsic factors (**Fig. S4B**). 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[33] of TFBS between pairs of training cell types as a proxy for cell-type specificity (**Fig. S4C**). High cell-type specificity explained some of the biggest maxATAC performance gains (e.g., for TCF12, JUNB, TCF7). On the other hand, for TFs with the least cell-type specificity, maxATAC had small performance gains over averaged ChIP-seq signal (~35% higher AUPR for NFE2L2, GABPA, ATF2) and on-par performance for CTCF and NRF1. Thus, maxATAC modeling is especially important for TFs with context-specific binding sites but also beneficial 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[34], suggesting that prediction for these TFs might be especially challenging (motif-less models are highlighted with red dots in **Fig. 2A-B, S2A-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 models extend state-of-the-art prediction to scATAC-seq and primary cells

Having constructed the largest suite of top-performing TFBS models for ATAC-seq, we next tested whether these models, trained on population-level 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 rare cell types and *in vivo* settings, 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 computationally inferred population of cells. From pseudobulk profiles, regions of chromatin accessibility are detected and annotated with TFBS predictions in standard scATAC-seq pipelines[18,19]. 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[19]. 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 S2**). As expected, maxATAC outperformed popular TF motif scanning (**Fig. 3B**). In side-by-side comparison in the same test cell types, maxATAC performed nearly as well on scATAC-seq as population-level ATAC-seq (**Fig. 3C**). The comparative performance was cell-line dependent. Predictions from scATAC-seq of HeLa, THP-1, Jurkat and HEK293T performed similarly to bulk (**Fig. 3D**). In contrast, maxATAC predictions on scATAC-seq were, on average, slightly worse than bulk ATAC-seq for MCF-7, K562 and GM12878 (**Fig. 3C**). Undersampling is a potential concern with any genomic assay, and single-cell technologies in particular[35]. Thus, we examined whether undersampling might explain the poorer performance for MCF-7, K562, and GM12878 relative to the other "on-par" cell lines. However, there was no strong correspondence between relative performance and the number of cells or total number of fragments per pseudobulk population (**Fig. S5A-C**). Given the near-decade spanning ChIP-seq, ATAC-seq and scATAC-seq data generation, changes due to tissue culture or passaging of cell lines are potential confounders driving relative performance differences. Nevertheless, performance of maxATAC on scATAC-seq is state-of-the-art.

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 ATAC-seq and ChIP-seq of TCR-dependent TFs (FOS, JUNB and MYC) had been collected[36] (**Fig. 4A**). In addition to being a test of performance on primary cells, the ATAC-seq data were generated using the standard ATAC-seq protocol (in contrast to the OMNI-ATAC-seq benchmark data). maxATAC predictions for all three TFs outperformed TF motif scanning in ATAC-seq peaks (**Fig. 4B-D**). Even at low recall, maxATAC performance was on-par (JUNB) or better (MYC, FOS) than TF motif scanning, with 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) (**Methods**).

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 4E**). In contrast, the FOS model was constructed from only two cell types, and its performance, although far 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[37] and a robust normalization strategy. We highlight that these strategies were developed using an independent scATAC-seq dataset in GM12878[38], so the performance reported above truly represents test performance **(Fig. 4).** In contrast to DNase-seq, ATAC-seq is variably contaminated with accessibility signal from mitochondrial chromosomes[39,40]. For example, in GM12878, we detected genomic regions of high chromatin accessibility in OMNI-ATAC-seq that were not present in scATAC-seq (or DNase-seq) (**Fig. S6**). 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

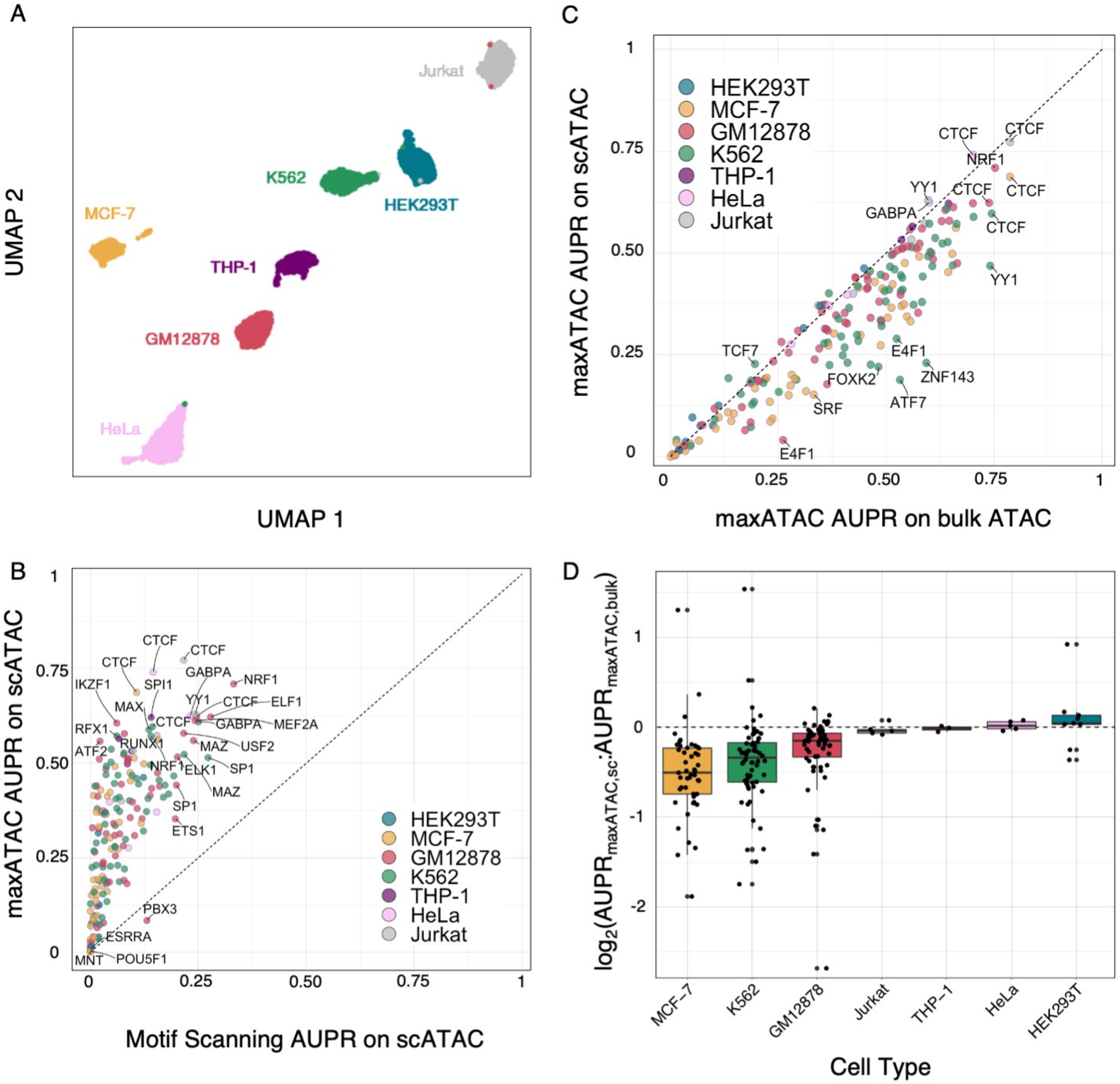

Fig. 3. maxATAC offers state-of-the-art TFBS prediction from scATAC-seq. (**A**) UMAP of 10X scATAC-seq data from 7 cell types in a cell line-mixing experiment (Granja et al. 2021), enabling test performance evaluation for 193 maxATAC models. (**B**) AUPR of maxATAC on scATAC-seq versus AUPR of TF motif scanning on scATAC-seq. **(C)** AUPR for maxATAC in scATAC-seq relative to maxATAC performance on bulk ATAC-seq. **(D)** Boxplot of the $\log_2(\text{AUPR}_{\text{maxATACsc}}:\text{AUPR}_{\text{maxATACbulk}})$ per cell type. Each dot indicates test performance in a TF model.

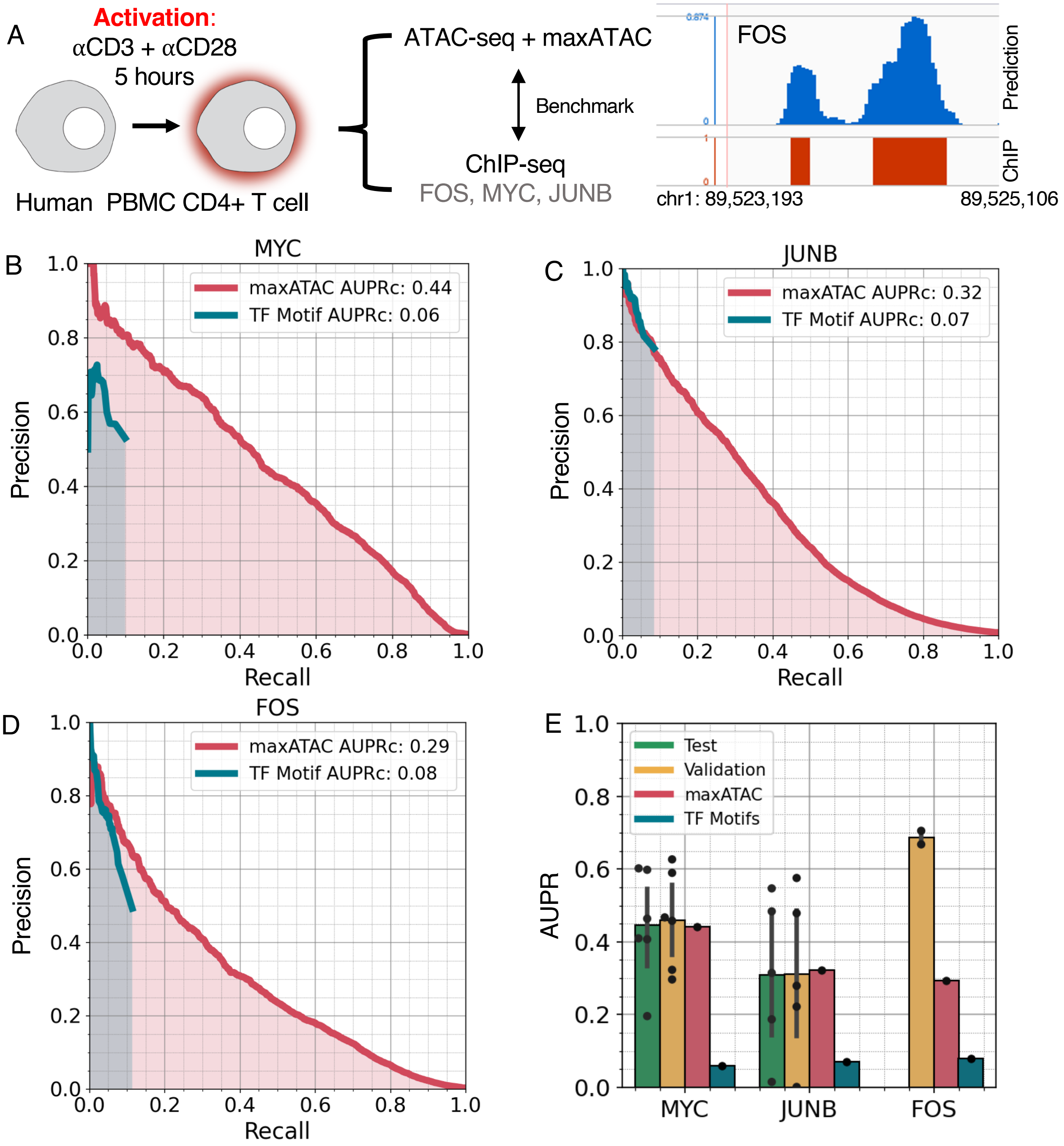

Fig. 4. maxATAC models perform well in primary human cells. (A) Study design for maxATAC benchmarking in primary human CD4^{+} T cells. Precision-recall curves for maxATAC predictions (red line) and TF motif predictions (blue line) for **(B)** FOS, **(C)** JUNB, and **(D)** MYC compared to TFBS from 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 (teal). Each point corresponds to a unique train-test cell type split, and 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).

cell lines still had extreme, outlying ATAC-seq signal (**Fig. S7**). Some of these regions overlapped cancer-specific driver super enhancers (e.g., TRIM37 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. S8**).

maxATAC identifies allele-dependent TFBS at atopic dermatitis genetic risk variants

Identifying the cellular and molecular drivers of phenotypic diversity is a fundamental goal of basic and translational research. Complex traits are products of genetic and environmental factors. Chromatin accessibility is sensitive to environmental factors, like age[41] and microbiome[42], and therefore a critical complement to genetic profiling of patients across disease spectra, from cancer to autoimmune and obesity-related diseases. While previous work in deep neural network modeling focused on interpretation of genetic variants[27,28,43], maxATAC integrates both genetic (sequence) and environmental (ATAC-seq) signals and is therefore ideally suited to the elucidation of molecular drivers of diseases involving both genetic and environmental components.

We demonstrate the potential for maxATAC in the context of a complex disease. Atopic dermatitis (**AD**) is one of the most common skin disorders in children. Its etiology involves both genetic and environmental factors, with 29 independent AD risk loci known[44,45]. Here, we take advantage of an important genomics resource in AD: ATAC-seq and RNA-seq of activated CD4+ T cells, along with whole-genome sequencing, of AD patients and age-matched controls (**Fig. 5A**)[46]. Previous analysis of this dataset identified several AD risk variants with allele-dependent chromatin accessibility in activated T cells[46].

Here, we use maxATAC to identify TFBS associated with allele-dependent chromatin accessibility at these AD risk loci (**Table S3**). To avoid error associated with phasing of DNA and ATAC-seq signal (into maternal and paternal strands), we identified a pair of AD and age-matched controls ("AD2", "CTL2") where the AD patient was homozygous for the AD risk haplotypes while the control was homozygous for the AD non-risk haplotype at two independent loci tagged by two variants: rs6062490 and rs1151624 (**Fig. 5B**). For both variants, we considered the full haplotype block (linkage disequilibrium R^2>.8), using donor-specific DNA sequence and accessibility to predict TFBS for 105 expressed TFs with maxATAC models available (**Fig. 5C-G, Fig. S9A Methods**).

The rs6062490 haplotype block is ~34.2kb and contains 30 noncoding SNPs interspersed between exons of two protein-coding genes: *RTEL1* and *TNFRSF6B* **(Fig. 5D).** *TNFRSF6B* is an anti-apoptotic gene, and serum levels of this protein are increased in patients with atopic dermatitis[46,48]. The rs1151624 haplotype block is ~11.3kb and contains 19 noncoding SNPs cis to three genes: *ZGPAT*, *LIME1*, and *SLC2A4RG* **(Fig. 5E).** *SLC2A4RG* is a known eGene, (associated with an eQTL) in activated CD4+ T cells[47], while *LIME1* is a transmembrane protein that controls effector T cell migration to sites of inflammation[47]. Thus, genes at both risk loci have ties to T cell biology, a T-cell eQTL, and AD.

For both loci, we ranked TFs based on the number of predicted differential TFBS regions between AD2 and CTL2 (**Fig. 5F, S9B**). Although TFBS were generally increased in the AD patient relative to the control in these loci (53 TFs), three TFs had a predicted decrease in binding sites in AD2 (KMT2A, LEF1, STAT5A), while 50 TFs had no TFBS predicted in these loci. MYB and FOXP1 showed the greatest differential binding, and their TFBS were predominantly increased in the AD

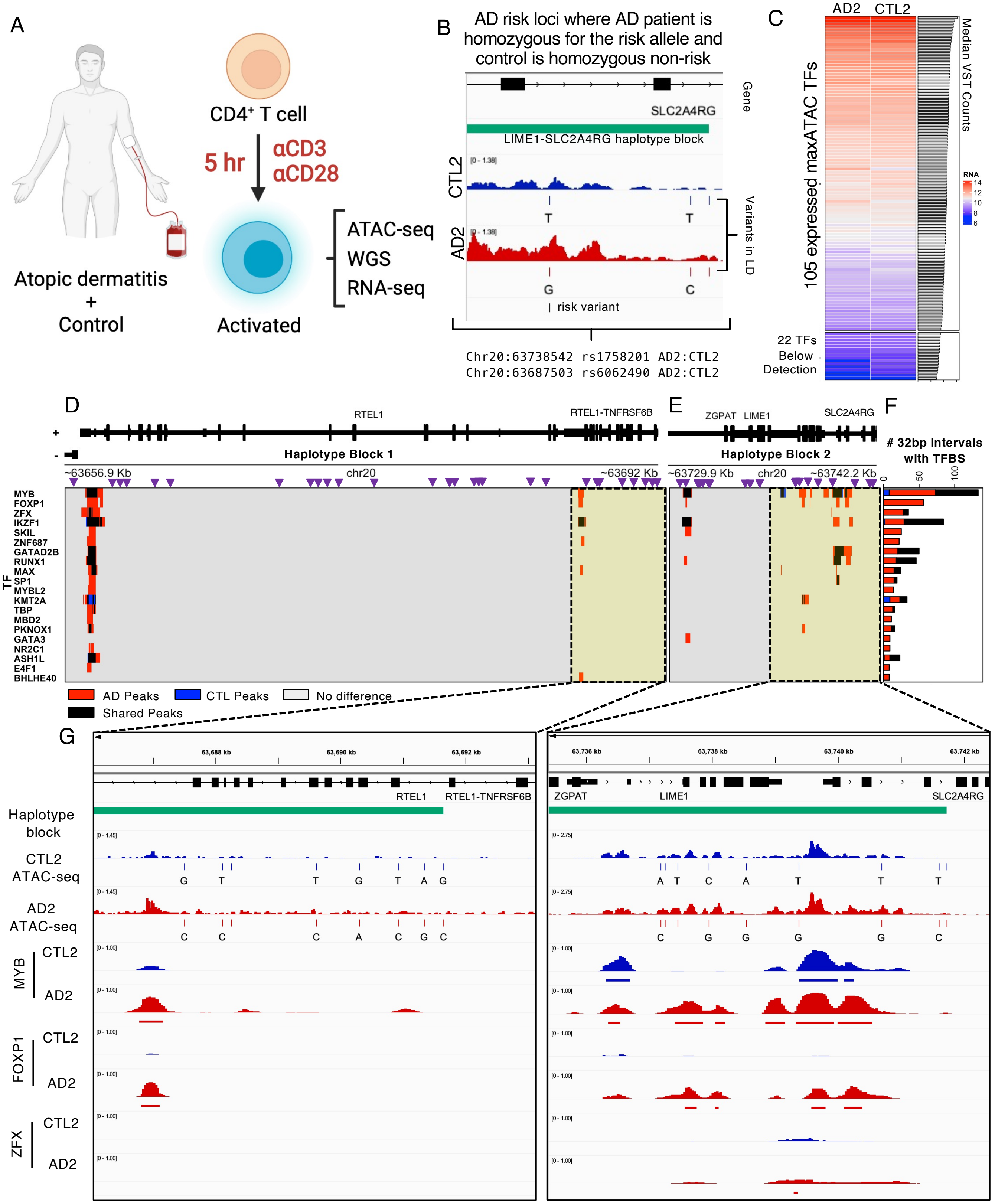

Fig. 5. maxATAC TFBS prediction at atopic dermatitis risk loci in patient-derived CD4⁺ T cells. (**A**) In a previous study (Eapen et al., 2021), peripheral CD4⁺ T cells were isolated from atopic dermatitis (**AD**) patients and age-matched controls (**CTL**) and TCR-stimulated prior to ATAC-seq, RNA-seq and DNA-seq data generation. (**B**) Of nine AD risk loci shown to exhibit allele-dependent chromatin accessibility in T cells by Eapen et al., 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 expressed in 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 represents intervals with 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 TF binding sites.

patient. FOXP1 has been previously implicated in the maintenance of T cell quiescence and its expression is typically repressed in activated T cells[48]. MYB is a critical regulator of regulatory T cell differentiation and immune tolerance[49]. Thus, both TFs have known roles in T-cell biology. We visualized regions of each haplotype block, putative cis-regulatory modules, where MYB and FOXP1 were predicted to bind with several other factors **(Fig. 5G, S10).**

Although limited in statistical power, we examined the correlation between expression of genes at these loci in AD patients and age-matched controls **(Fig. S9B-G).** We observed a trend for increased expression of *SL2A4RG* and decreased expression of *ZGPAT* in AD patients harboring the risk variant rs1151624 (trend held for all 4 homozygous-risk AD patients relative to their homozygous or heterozygous non-risk controls). We observed increased expression of all three genes *(RTEl1, RTEL1-TNFRSF6B*, *TNFRSF6B)* associated with rs6062490, when comparing the single pair of homozygous-risk AD to homozygous non-risk control, but we observed inconsistent trends from the two donor pairs in which the AD patients were homozygous risk and controls were heterozygous non-risk. These trends, in addition to support for *SLC2A4RG* as an eGene in CD4$^+$ T cells[47], nominate potential molecular mechanisms, specific TFs that might alter the expression of genes important to T cell function (MYB and FOXP1) in AD patients. Interestingly, for the regions shown in **Fig. 5G**, neither FOXP1 nor MYB had motif occurrences in the differential ATAC-seq peaks, highlighting the unique predictive capabilities of maxATAC.

maxATAC analysis of this clinical genomics study provided “in silico ChIP-seq" for 105 TFs, in a setting where experimental measurement of TF occupancy for 105 factors would have been infeasible. Application of maxATAC to the growing number of genetic studies with population-level and single-cell ATAC-seq will improve the power of these studies to accurately predict TF mediators of allelic chromatin accessibility and gene expression. Furthermore, many genetic tools use overlap with TFBS to nominate potentially causal risk variants, where TFBS are predicted based on suboptimal TF motif scanning or ChIP-seq in a potentially suboptimal cell type or condition, due to lack of data in the *in vivo* disease context. Integration of maxATAC TFBS predictions into genetic analysis pipelines will be the focus of future work.

Discussion:

Genomic measurement and machine learning capabilities are advancing at an unprecedented pace. The possibilities for computational modeling to address fundamental questions in biology and human health have never been greater. Leveraging the fastest-growing chromatin-state measurement ATAC-seq, maxATAC seeks to make this moment’s state-of-the-art in transcription-factor binding prediction readily achievable across the basic and biomedical research spectrum. Rigorously benchmarked across cell lines, primary cells, and single-cell ATAC-seq, maxATAC will improve TFBS predictions and knowledge gain from ATAC-seq studies. With maxATAC and its user-friendly codebase (soon to be available from https://github.com/MiraldiLab/maxATAC), state-of-the-art, genome-scale TFBS prediction can be accomplished for 127 human TFs using a single ATAC-seq or scATAC-seq experiment.

We echo a top-performing group in the “ENCODE-DREAM *in vivo* TFBS Prediction Challenge”[14], 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[28,50] 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[50] 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[51]. 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[43], 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[52] or estimated from covariance of functional genomics assays[53] (including scATAC-seq[54]). 3D-chromatin interactions could be incorporated using graph neural networks[55,56] or with simple post-processing, like affinity propagation[57] 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[21]). 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.

Acknowledgements

We thank the 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 and Brad Rosenberg for manuscript feedback and Wen Niu for help cross-referencing ChIP-seq experiments across databases. This work was supported by the National Institute of Health (**NIH**) U01AI150748 to ERM, SP, MTW, LCK; NIH R01AI153442 to ERM, AB; NIH R21AI156185 to ERM; NIH R01HG010730, U01AI130830, R01NS099068, R01GM055479, P01AI150585 to MTW; NIH R01AI024717, R01AR073228, and Cincinnati Children's Research Foundation (**CCRF**) ARC Award 53632 to MTW, LCK; NIH R01DK107502, R01AI148276, U19AI070235, U01HG011172, and P30AR070549 to LCK; CCRF Center for Pediatric Genomics (**CPG**) grants to Cincinnati Children's Research Foundation to ERM, AB, MTW, LCK, SP. We used BioRender.com and plotgardener to create some of the figure images.

Author Contributions

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

Competing Interests Statement

AB is a co-founder of Datirium, LLC.

Methods:

Data and Code Availability

The maxATAC codebase will soon be available from https://github.com/MiraldiLab/maxATAC, including basic usage (maxATAC installation, ATAC-seq data processing and TFBS prediction with the trained maxATAC models) and advanced (model training and benchmarking). OMNI-ATAC-seq data generated by this study will be deposited in the GEO database.

maxATAC Training Data

We curated training data for maxATAC from the CistromeDB[58] and ENCODE[12] databases (**Figure 1A-B**), identifying cell types that had (1) a high number of human TF ChIP-seq experiments and (2) ATAC-seq data. We limited ATAC-seq training data to higher quality OMNI[59] and required paired-end sequencing, while, for ChIP-seq, we utilized both paired- and single-end sequencing. We required sequencing depth ≥ 20 million reads per biological replicate (sum of the spot numbers for SRR IDs per SRX ID). We manually verified the cell type, TF, and experimental source of each experiment. To ensure that ATAC-seq and ChIP-seq were derived from the same experimental conditions per cell type, we eliminated any ATAC-seq or ChIP-seq experiment in which the cell type was perturbed (including genetically modified, transfected with exogenous vectors, or treated with vehicle controls, environmental perturbations, metabolic manipulations, or differentiation protocols).

ChIP-seq

When available, we utilized processed ENCODE ChIP-seq experiments over other publicly available data. For TF and cell type pairs with multiple ChIP-seq experiments available, we chose the most recent experiment with the greatest number of reproducible TFBS (peaks) detected by IDR analysis[60]. If available, we selected conservative over optimal IDR peak sets. We excluded any experiment with a red flag (i.e., a ”critical issue” was identified by ENCODE) as well as experiments with fewer than 500 TFBS detected. This resulted in 371 TF-cell type conditions from ENCODE.

The remainder of our ChIP-seq training data required processing and additional quality control. Using snakemake[61], we followed ENCODE3[62] standards for ChIP-seq read alignment, read filtering, and peak calling; this workflow is available from the maxATAC codebase. In brief, each biological replicate was summarized according to SRX ID for a total of 316 experiments. Fastq files were downloaded (*SRAtools fasterq-dump* v. 2.10.8) with technical replicates concatenated per SRX ID. Fastq files were assessed for adapter contamination and read quality statistics (*FastQC* v. 0.11.9). Samples flagged for high levels of N sequences were removed. *TrimGalore!* (v. 0.6.7) was used to remove adapter contamination and trim the low-quality bases at the 3’ end of the sequencing read with the settings (-q 20). Samples with (1) < 15 million reads after filtering or (2) average read length < 20 bp after trimming were excluded from the analysis. Reads were aligned to the hg38 reference genome using *bowtie2* (v. 2.4.4)[63] (--very-sensitive --maxins 2000). The aligned reads were quality filtered with *samtools* (v. 1.9)[64] (-F 1804 -q 30) and PCR duplicates were removed with *samtools markdup* (-r -s). Prior to peak-calling, we also excluded reads mapping to blacklist regions compiled from ENCODE data[23] as well as centromeres, telomeres, and annotated gaps[24]. MACS2 (v 2.2.2.7.1)[65] was used to call peaks on the filtered BAM file with parameters (--nomodel --extsize 147). For cell types and TF combinations with multiple biological replicates, we provided all filtered BAM files during peak calling. Peaks meeting an FDR = 5% cutoff were retained as TFBS for benchmarking. TF-cell type conditions with fewer than 500 TFBS detected were excluded.

Further QC of the "non-ENCODE" ChIP-seq involved TF motif analysis and biological replicate Pearson correlation of > .6 (when available). We used HOMER (v. 4.11)[16] with the CIS-BP (v. 2.0)[34] database to test for enrichment of expected motifs in ChIP-seq peaks. CIS-BP provides multiple motifs per TF, so we selected the most highly enriched motif per TF and 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 conditions.–In total, our 127 TF models cover 443 unique cell types and TF combinations across 20 cell types (**Fig 1B**).

ATAC-seq

We combined ENCODE with in-house OMNI-ATAC-seq for our 20 benchmark cell lines. In-house, we ordered HepG2, LoVo and HEK293 cells from ATCC[66] tand targeted a median sequencing depth of 20 million reads for each cell type. OMNI-ATAC-seq was later released[62] 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!*[67] 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)[68], bowtie2 (v. 2.4.4)[63], 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[69] 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 (e.g., 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. S8**). 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[70–72], 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*[73] *bamtobed* 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[74,75], 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[72].

Extended blacklist. Initial testing of maxATAC models on scATAC-seq in GM12878 highlighted the need for an extended blacklist[37]. 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. S6-7**). Thus, our final blacklist included (1) blacklisted regions from ENCODE data[76], (2) centromeres, telomeres, and annotated gaps available from UCSC table browser[77] for hg38, (3) regions ≥1kb with ≥ 90% sequence identity to chrM[78], and (4) regions with low mappability on chr21 (**Table S4**). 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 S7**). Although often biological in nature (e.g., TRIM37 locus in MCF-7[79], **Fig. S7C**), these outlying signals interfered with min-max normalization of ATAC-seq and cross-platform performance on scATAC-seq (**Fig. Supp. S8**). To improve robustness[80], 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[81], while a standard max-min strategy did not (**Fig. S8**). This normalization strategy was then independently tested on another scATAC-seq dataset[19] (**Fig. 3**).

Peak Calling. We called peaks with MACS2[65] 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.

Model Architecture

The maxATAC models are deep dilated convolutional neural networks[28,82] that predict transcription factor binding sites as a function of ATAC-seq and DNA sequence (**Fig. Supp. 9**). 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 7bps 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. 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 derived per TF-cell type conditions (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[83], 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[84]. 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 minimum validation cross entropy loss 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. Because only ~1% of the chromatin is expected be accessible in a given context, 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 train 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 (1) commonly used random sampling from the genome[85] and mixed random and peak-centric strategies (e.g., 50-50, **Fig. S2A**). Thus, 100% peak-centric, pan-cell training was used for the final maxATAC models.

Following previous work[25], we doubled the number of training examples by including sequence and signal from the reverse-complement in addition to the typically used 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**)[30]. ROC weights TP and TN equally, and therefore a less informative metric of performance in unbalanced classification problems, like TFBS prediction. We report area under precision-recall (**AUPR**), precision at 5% recall and $\log_2(\text{AUPR}_{\text{model}}/\text{AUPR}_{\text{random}})$, as described in **Results**.

We ranked maxATAC TFBS prediction 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 for 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*[86].

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

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[18] together with the TF PWM database CISBP v2[2] to identify motif occurrences with a P< $1x10^{-5}$. 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[87], using the averaged signal to rank TFBS for precision-recall analysis.

Prediction with the maxATAC Models

The final maxATAC models were constructed using all benchmark cell types available for a given TF (**Fig. 1C**), 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. 4**). In addition to publishing the final maxATAC models, we took advantage of the good correlation between validation and test performance (**Fig. 2C, S2C-D**) and mapped maxATAC scores to intuitive validation performance statistics precision, recall, $\log_2$(Precision / $\text{Precision}_{\text{random}}$), 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 cutoff that maximized recall.

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

scATAC-seq

We evaluated maxATAC on scATAC-seq during (1) method development (**Fig. S8**) and (2) independent testing (**Fig. 3**). For method development, we downloaded scATAC-seq fragment files for GM12878 (500 and 5000 cell experiments)[88] from the 10X Genomics website (https://www.10xgenomics.com/resources/datasets). For testing, we downloaded the high-loading, mixed-cell line scATAC-seq experiment SRX9633387 from GSE162690[75], which we processed to cell-type specific pseudobulk fragment files. We annotated cells and removed doublets using demuxlet.

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.

Primary cell types

We downloaded ATAC-seq (GSE116696) and ChIP-seq data for 3 TFs (FOS:GSM3258569, MYC:GSM3258570, and JUNB:GSM3258571) in primary CD4+ T cells stimulated with anti-CD3/anti-CD28 beads for 5 hours[89], processing ATAC-seq and ChIP-seq for maxATAC and precision-recall analysis as described above.

Sequence-Specific TFBS prediction in CD4$^+$ T cells from atopic dermatitis patients

We derived both DNA sequence and ATAC-seq signal inputs for maxATAC from a genetics atopic dermatitis (**AD**) study[46], combining whole genome sequencing (**WGS**) and OMNI-ATAC-seq (of peripheral blood CD4+ T-cells stimulated with anti-CD3 and anti-CD28 beads for 5 hours) for 6 AD patients and 6 age-matched controls. We used patient-specific genetic variant calls and the nine AD risk variants previously associated with allele-dependent chromatin accessibility identified by Eapen et al. We sought to identify pairs of AD patients and age-matched controls that were homozygous for the risk and non-risk alleles, respectively. Focusing on homozygous-risk and homozygous-nonrisk obviated the need for phasing (i.e., to resolve ATAC-seq signal into maternal and paternal DNA strands). We focused on two independent AD risk loci that met our criteria. For these loci, we applied maxATAC to patient- and control-specific DNA sequence and ATAC-seq signal to make TFBS predictions in the haplotype blocks containing the risk variants (linkage-disequilibrium R^2>.8), for the patient pair, AD2 and CTL2. Haplotype blocks were defined by the homozygous risk variants identified above and all SNPs in linkage disequilibrium with the risk variant ($R^2 \geq .8$) were used for patient-specific sequence prediction with maxATAC. The furthest genetic variant positions in LD with risk variant defined the prediction window, which was extended + or - 512 bp to contain the most distal variants in LD. We limited TFBS prediction to TFs with nominal mRNA expression (DESeq2 VST counts ≥ 9) in activated CD4$^+$ T cells, measured in parallel RNA-seq by Eapen et al.

References

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).
2. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 1222794 (2012).
3. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
4. Harley, J. B. *et al.* Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. *Nature Genetics* **50**, (2018).
5. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468**, 911–20 (2010).
6. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213–8 (2013).
7. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–90 (2015).
8. Cusanovich, D., Daza, R., Adey, A. & Pliner, H. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. (2015).
9. Corces, M., Buenrostro, J., Wu, B. & Greenside, P. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature* (2016).
10. Miraldi, E. R. *et al.* Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. *Genome research* **29**, 449–463 (2019).
11. Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R. & Gresham, D. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife* **9**, e51254 (2020).
12. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
13. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research* **21**, 447–455 (2011).
14. Keilwagen, J., Posch, S. & Grau, J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biology* **20**, 1–17 (2019).
15. Li, H., Quang, D. & Guan, Y. Anchor: Trans-cell Type Prediction of Transcription Factor Binding Sites. 281–292 (2019) doi:10.1101/gr.237156.118.29.
16. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576–589 (2010).
17. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods* **14**, 975–978 (2017).
18. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nature Methods 2021 18:11* **18**, 1333–1341 (2021).
19. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics* **53**, 403–411 (2021).

20. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nature Communications* **12**, 1337 (2021).
21. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
22. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome biology* **20**, 45 (2019).
23. Li, H. & Guan, Y. Fast decoding cell type–specific transcription factor binding landscape at single-nucleotide resolution. *Genome Research* **31**, 721–731 (2021).
24. Quang, D. & Xie, X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Doi.Org* 151274 (2017) doi:10.1101/151274.
25. Mei, S. *et al.* Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic acids research* gkw983 (2016).
26. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**, 831–838 (2015).
27. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* **12**, 931–934 (2015).
28. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research* **28**, 739–750 (2018).
29. 2 Chromatin patterns at transcription factor binding sites. *Nature* (2019) doi:10.1038/nature28171.
30. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**, e0118432 (2015).
31. Schreiber, J., Bilmes, J. & Noble, W. S. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome biology* **21**, 1–13 (2020).
32. Miraldi, E. R. *et al.* Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. *Genome research* **29**, 449–463 (2019).
33. Favorov, A. *et al.* Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. *PLOS Computational Biology* **8**, e1002529- (2012).
34. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).
35. Kharchenko, P. v. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods* 1–10 (2021).
36. Yukawa, M. *et al.* AP-1 activity induced by co-stimulation is required for chromatin opening during T cell activation. *The Journal of experimental medicine* **217**, 647388 (2020).
37. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Scientific reports* **9**, 1–5 (2019).
38. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature biotechnology* **37**, 925–936 (2019).
39. Ou, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC genomics* **19**, 169 (2018).

40. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome biology* **21**, 1–16 (2020).
41. Hu, B. *et al.* Distinct age-related epigenetic signatures in CD4 and CD8 T cells. *Frontiers in immunology* **11**, (2020).
42. Richards, A. L. *et al.* Gut microbiota has a widespread and modifiable effect on host gene regulation. *MSystems* **4**, e00323-18 (2019).
43. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* **26**, 990–999 (2016).
44. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics* **47**, 1449–1456 (2015).
45. Brown, S. J. What have we learned from GWAS for atopic dermatitis? *Journal of Investigative Dermatology* **141**, 19–22 (2021).
46. Eapen, A. A. *et al.* Epigenetic and Transcriptional Dysregulation in CD4+ T cells of Patients with Atopic Dermatitis. *bioRxiv* 2021.12.03.471059 (2021) doi:10.1101/2021.12.03.471059.
47. Schmiedel, B. J. *et al.* Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715 (2018).
48. Garaud, S. *et al.* FOXP1 is a regulator of quiescence in healthy human CD4+ T cells and is constitutively repressed in T cells from patients with lymphoproliferative disorders. *European Journal of Immunology* **47**, 168–179 (2017).
49. Dias, S. *et al.* Effector Regulatory T Cell Differentiation and Immune Homeostasis Depend on the Transcription Factor Myb. *Immunity* **46**, 78–91 (2017).
50. Avsec, Ž. *et al.* Base-resolution models of transcription factor binding reveal soft motif syntax. *Nature Genetics* (2020).
51. Setty, M. & Leslie, C. S. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. *PLoS computational biology* **11**, e1004271 (2015).
52. Quinodoz, S. A. *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744–757 (2018).
53. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data. *Proceedings of the National Academy of Sciences* **114**, E4914–E4923 (2017).
54. Pliner, H. A. *et al.* Cicero Predicts cis -Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell* 858–871 (2018) doi:10.1016/j.molcel.2018.06.044.
55. Veličković, P. *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
56. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 1025–1035 (2017).
57. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics* **18**, 551–562 (2017).
58. Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Research* **47**, D729–D735 (2019).
59. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* (2017) doi:10.1038/nmeth.4396.

60. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5**, 1752–1779 (2011).
61. M□lder, F. *et al.* Sustainable data analysis with Snakemake [version 2; peer review: 2 approved] . *F1000Research* **10**, (2021).
62. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
64. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
65. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
66. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature methods* **14**, 959–962 (2017).
67. Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo. (2021) doi:10.5281/ZENODO.5127899.
68. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)* **29**, 15–21 (2013).
69. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications* **11**, 4267 (2020).
70. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213–1218 (2013).
71. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21**, 22 (2020).
72. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology* **11**, R119 (2010).
73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
74. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nature Methods* **18**, 1333–1341 (2021).
75. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics* **53**, 403–411 (2021).
76. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* **9**, 9354 (2019).
77. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic acids research* **32**, D493–D496 (2004).
78. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. (Reports). *Science* **297**, 1003+ (2002).
79. Hampton, O. A. *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 research* **19**, 167–177 (2009).
80. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, 1–12 (2010).

81. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37**, 925–936 (2019).
82. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**, 834–848 (2018).
83. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
84. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. in *Proceedings of the thirteenth international conference on artificial intelligence and statistics* 249–256 (JMLR Workshop and Conference Proceedings, 2010).
85. Li, H. & Guan, Y. Fast decoding cell type–specific transcription factor binding landscape at single-nucleotide resolution. *Genome Research* **31**, 721–731 (2021).
86. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
87. Schreiber, J., Durham, T., Bilmes, J. & Noble, W. S. Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biology* **21**, 81 (2020).
88. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology* **37**, 925–936 (2019).
89. Yukawa, M. *et al.* AP-1 activity induced by co-stimulation is required for chromatin opening during T cell activation. *Journal of Experimental Medicine* **217**, e20182009 (2019).

Supplementary Table Legends

Table S1. Performance metrics for 127 transcription factor models in bulk ATAC-seq. This table consists of results 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, Average ChIP-seq signal AUPR for chr1, and analgous metrics for validation performance on chr2. Additional annotations such as the number of ChIP-seq peaks and TF family annotations are also included.

Table S2. Performance metrics for 193 cell type and transcription factor models in scATAC-seq data. This table consists of 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 S3. Atopic dermatitis risk loci with allele-dependent ATAC-seq signal. This table contains the variants that are associated with allele-dependent ATAC-seq signal and are in linkage disequilibrium with an 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 S4. 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 annotation gaps.