Base-pair resolution learning improves regional chromatin accessibility prediction in immune cells

Nuria Alina Chandra, Alexander Sasse, Sara Mostafavi

Genome-wide association studies have found thousands of correlations between genetic variants and diseases.[1] However, most of these variants are from regulatory regions of the genome which are difficult to study experimentally due to their cell-type-specific effects.[2] Accurate prediction of the transcriptional consequences of regulatory variants is necessary to decode the genetic basis of disease and facilitate the development of personalized treatment systems.[2] In this study, we focus on unraveling the regulatory mechanisms governing immune cells, a crucial endeavor in hereditary immunological disease research.

Chromatin accessibility provides a window into the complex process of gene regulation. Many regulatory molecules operate by dynamically modifying chromatin accessibility. Transcription factors (TFs) bind to specific DNA sequence motifs in regulatory regions of the genome, opening densely packed chromatin, and enabling transcriptional machinery to access the DNA. Genome-wide chromatin accessibility can be measured by ATAC-seq, which harnesses the Tn5 enzyme's ability to preferentially cut DNA in open chromatin regions (OCRs).[3] Deep learning has been established as a useful tool to identify regulatory motifs that are physically bound by regulatory proteins, such as transcription factors.[4] AI-TAC is a convolutional neural network that was developed to determine the genome-wide sequence patterns that control differential chromatin accessibility in mouse immune cells.[5] AI-TAC is trained to predict regional chromatin accessibility (the total number of ATAC-seq Tn5 cuts per region) across 81 immune cell types from the surrounding genomic sequence alone. While AI-TAC has shown promising results using this training strategy, recent research has established that the base-pair resolution distribution of Tn5 cuts (the ATAC-seq "profile") contains TF "footprints" which provide additional information about the location and strength of TF binding sites.[6] Avsec et al. and Trevino et al. also demonstrated that training deep learning models (BPNet and ChromBPNet) on base-pair resolution chromatin accessibility profiles improves discovery of TF motifs and TF interactions.[7,8] These findings prompted us to ask if learning motifs from base-pair resolution ATAC-seq profiles could also improve prediction of regional chromatin accessibility.

Here, we introduce a multitask model ("bpAI-TAC") that predicts differential regional chromatin accessibility, and base-pair resolution accessibility profiles across 90 different closely-related immune cell types. The body of the model learns a representation of each base-pair with information from surrounding bases using a convolutional layer and a series of dilated convolutional layers with residual skip connections (Figure 1). The representation learned by the body is sent to two output heads: a "scalar head" which predicts the regional accessibility (total number of Tn5 counts in a DNA region), and a "profile head" which predicts the likelihood of a Tn5 cut at each base from the base pair representation. The scalar head applies several additional convolutional layers with max-pooling and a fully connected layer to predict the total Tn5 counts in a region from the learned representation of the entire region.

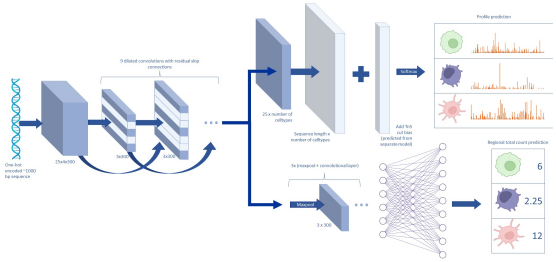

Figure 1: bpAI-TAC architecture

The Tn5 enzyme has a weak but apparent preference for cutting at specific sequence patterns independent of DNA accessibility, which we accounted for by creating a pre-trained bias model of Tn5 cuts on transcription factor free DNA.[9] For each input, the predicted Tn5 bias was added to the profile head likelihood prediction in the bpAI-TAC model. Adding the bias is crucial to enable bpAI-TAC to learn patterns of transcription factor binding and prevent it from learning Tn5 cutting preference patterns. The combined likelihood is normalized with a softmax function and compared to the measured profile with cross entropy. The entire model is trained on a composite loss function that uses the mean squared log error for the scalar head and cross entropy for the profile head.

$$\mathcal{L} = MSE(log(r^{pred}+1), log(r^{obs}+1)) + \lambda * MeanCrossEntropy(\mathbf{p}^{pred}, \mathbf{p}^{obs}))$$

Where r is the regional total Tn5 counts, and $\mathbf{p}$ is a vector of the base-resolution profile. The means are taken across all cell types and chromatin regions. λ controls the relative weight on the profile head prediction error.

bpAI-TAC was trained using ATAC-seq data from 90 different mouse immune cell types collected by the Immunological Genome Project.[10] As input, we used 998 bp long sequences centered around ATAC-seq peaks. We divided chromatin regions into training, validation, and test sets, leaving chromosomes 11, 12, 15, and 16 out for analyzing model performance. We trained models with different weights on the profile head by increasing λ from 0 (only training on regional accessibility) to 10^{10} (heavily focusing on profile). We found that models which included base resolution profiles in their loss showed superior prediction of regional accessibility, with an optimal λ of 0.7 - 0.9. When the model was weighted too strongly towards learning from profiles, the performance of regional chromatin accessibility prediction declined. This is similar to the trade-off pattern found for BPNet.[7]

Our results suggest that base-pair resolution chromatin accessibility profiles contain additional information that can be harvested by our new multi-task deep learning model to improve predictions of regional chromatin accessibility. We hypothesize that the observed performance increase can be attributed to additional information about TF binding interactions derived from base resolution footprints, and therefore that this approach will also improve our ability to extract information about the underlying biological processes from the data. We will seek to test this hypothesis through model interpretation as a next step.

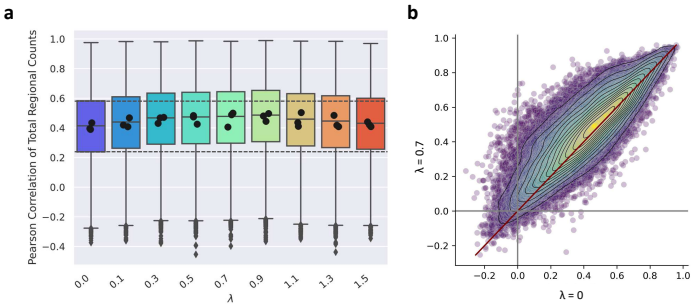

Figure 2: bpAITAC regional accessibility prediction performance. **a)** The Pearson correlation between actual and predicted regional chromatin accessibility (total number of ATAC-seq Tn5 counts in an OCR) across bpAI-TAC trained with different different λ weights on profile learning. Each condition was tested with three different random initializations. Black dots represent the mean correlation for each initialization. Regional chromatin accessibility prediction performance improves when the model learns bp-resolution profiles, with the highest performing initialization found at $\lambda = 0.7$. As weight on profile learning further increases, the regional accessibility prediction decreases demonstrating a trade off. **b)** A comparison of regional accessibility prediction between bpAI-TAC trained without learning from bp-resolution profiles ($\lambda = 0$), and with learning from profiles ($\lambda = 0.7$). Each point is the Pearson correlation between predicted and observed regional chromatin accessibility for an individual chromatin region, averaged over three random initializations, and colored by density.

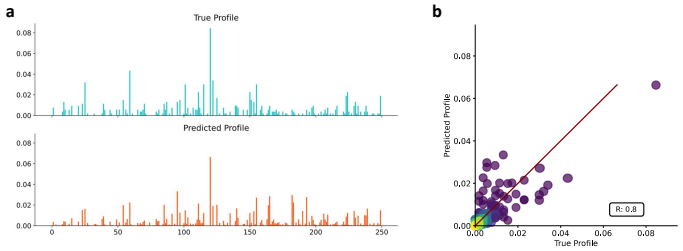

Figure 3: Comparison between a predicted and measured bp-resolution ATAC-seq profile for a single chromatin region in B.FrE.BM immune cells. Profiles were normalized to the center 250 bp of the region. **b** is a scatter plot of the profiles shown in **a** and colored by density.

References

[1] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, August 2021.

[2] Aaron K Wong, Chandra L. Theesfeld Rachel S. G. Sealfon, and Olga G. Troyanskaya. Decoding disease: from genomes to networks to phenotypes. *Nature Reviews Genetics*, 22, December 2021.

[3] Feng Yan, David R. Powell, David J. Curtis, and Nicholas C. Wong. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, 21(1), February 2020.

[4] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.*, 20(7):389–403, July 2019.

[5] Alexandra Maslova, Ricardo N Ramirez, Ke Ma, Hugo Schmutz, Chendi Wang, Curtis Fox, Bernard Ng, Christophe Benoist, Sara Mostafavi, and Immunological Genome Project. Deep learning of immune cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.*, 117(41):25655–25666, October 2020.

[6] Bryan Quach and Terrence S Furey. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics*, 33(7):956–963, April 2017.

[7] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*, 53(3):354–366, March 2021.

[8] Alexandro E. Trevino, Fabian Müller, Jimena Andersen, Laksshman Sundaram, Arwa Kathiria, Anna Shcherbina, Kyle Farh, Howard Y. Chang, Anca M. Pașca, Anshul Kundaje, Sergiu P. Pașca, and William J. Greenleaf. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell*, 184(19):5053–5069.e23, September 2021.

[9] Yan Hu, Sai Ma, Vinay K. Kartha, Fabiana M. Duarte, Max Horlbeck, Ruochi Zhang, Rojesh Shrestha, Ajay Labade, Heidi Kletzien, Alia Meliki, Andrew Castillo, Neva Durand, Eugenio Mattei, Lauren J. Anderson, Tristan Tay, Andrew S. Earl, Noam Shoresh, Charles B. Epstein, Amy Wagers, and Jason D. Buenrostro. Single-cell multi-scale footprinting reveals the modular organization of DNA regulatory elements. March 2023.

[10] Hideyuki Yoshida, Caleb A Lareau, Ricardo N Ramirez, Samuel A Rose, Barbara Maier, Aleksandra Wroblewska, Fiona Desland, Aleksey Chudnovskiy, Arthur Mortha, Claudia Dominguez, Julie Tellier, Edy Kim, Dan Dwyer, Susan Shinton, Tsukasa Nabekura, YiLin Qi, Bingfei Yu, Michelle Robinette, Ki-Wook Kim, Amy Wagers, Andrew Rhoads, Stephen L Nutt, Brian D Brown, Sara Mostafavi, Jason D Buenrostro, Christophe Benoist, and Immunological Genome Project. The cis-regulatory atlas of the mouse immune system. *Cell*, 176(4):897—912.e20, February 2019.