Methodology and Applications for studying the Heterogeneity and Sequence Determinants of cis-Regulatory Elements

by

Tal Ashuach

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nir Yosef, Chair
Professor Sandrine Dudoit
Professor Nadav Ahituv
Professor Craig Miller

Spring 2022

Methodology and Applications for studying the Heterogeneity and Sequence Determinants of cis-Regulatory Elements

Abstract

Methodology and Applications for studying the Heterogeneity and Sequence Determinants of cis-Regulatory Elements

by

Tal Ashuach

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Nir Yosef, Chair

cis-regulatory elements (CREs) are non-coding segments of the genome which regulate the transcription of nearby genes. They can be broadly divided to two categories: 1) promoters, positioned directly upstream of their target gene, and 2) enhancers, positioned distally to their target gene. Enhancers are thought to be the main drivers of cell type-specific and state-specific transcription, and regulate gene expression by fine-tuning the rate of transcription, as opposed to the more binary (on or off) regulatory function that promoters typically have. Understanding how enhancers function is therefore crucially important to understanding how cells obtain and maintain certain fates and determine response to stimuli. Despite their importance, much is still unknown about the roles enhancers play in many biological processes, and how their sequence determines their regulatory function.

The first part of this dissertation deals with single-cell chromatin accessibility data (e.g as produced by single-cell ATAC-seq) as a means for systemically studying heterogeneity of CREs, and specifically enhancers. In chapter 2 this is demonstrated in the innate immune system's response to vaccination: in a subset of cells, a distinct state of chromatin accessibility maintains long-term epigenetic changes that prime these cells to a different response to stimuli, and provides non-specific viral protection. However promising, the unique properties of this data modality poses significant challenges. These are addressed in chapter 3, which introduced PeakVI, a deep generative model that provides a comprehensive statistical framework for analyzing data generated by scATAC-seq assays. Recent advances in sequencing technologies now enable obtaining these measurements alongside gene expression measurements (i.e single cell RNA-seq), providing the ability to directly measure the relationship between the heterogeneity of the chromatin landscape and that of the transcriptional profile. Chapter 4 introduces MultiVI, a general framework for the joint analysis of multi-modal single-cell data, using single-cell ATAC-seq and single-cell RNA-seq as the main example. These models enable exploration of cis-regulatory programs, identification of putative key

enhancers, and generating hypotheses about their regulatory functions.

The second part of this dissertation focuses on analyzing high-throughput functional data produces by massively parallel reporter assays (MPRAs). These assays enable direct functional characterization of thousands of synthetically generated candidate regulatory sequences. However, these assays include both DNA-seq and RNA-seq observations, and require controlling for various technical confounders within both assays, posing substantial computational challenges. Chapter 5 describes MPRAnalyze, a nested generalized linear model that provides a comprehensive statistical framework for analyzing MPRA data. Chapter 6 then uses MPRAnalyze extensively to identify key enhancers and novel trancription factors involved in early neural differentiation. In chapter 7, systemic perturbation of binding sites in the identified enhancers reveal the specific sequence features that determine enhancer function, and elucidates how multiple functional sites interact in a single enhancer sequence to reach the desired functional output.

To my family

Contents

Acknowledgments

My deepest gratitude to my advisor, Nir Yosef. For teaching me how to think like a scientist, guiding me through the frustrations and triumphs of research, and always encouraging me to disagree. And to all members of the Yosef lab, past and current, for fostering a collaborative environment where ideas can be shared and materialized. A special thanks to Nick Everetts, Anat Kreimer, Adam Gayoso, Shaked Afik, Carlos Buen Abad Najar, Michael Cole, David DeTomaso, Matt Jones, Alyssa Morrow, Galen Xing, and Michal Rozenwald.

I want to thank all the brilliant scientists I had the privilege of collaborating with throughout my doctorate. A special shout-out to Anat Kreimer and Florian Wimmers, for numerous enlightening discussions and brilliant ideas.

I also wish to thank those who started my journey in computational biology at the Hebrew University. Eran Meshorer, my previous advisor, for letting me do things that made me want to do more. Tommy Kaplan, Hanah Margalit, and Nir Friedman, for introducing me to the field and for the priceless mentorship throughout my years there. And the friends made through hours of studying and laughing and eating and drinking: Ehud Karavani, Tamar Amitai, Itay Dalmedigos, Chaim Hoch, and Adi Watsman. And most of all, for making Jerusalem feel like home, I am eternally grateful to Shiran Woland.

My academic journey at Berkeley would not have been the same without wonderful supporting faculty. Sandrine Dudoit, a rotation advisor, instructor, qualifying exam chair, and thesis committee member, and at all times was a brilliant scientist and awesome person. Haiyan Huang, who was my instructor, rotation advisor, and program director, and always a genuine joy to speak to and work with. Nadav Ahituv, who provided invaluable feedback and assistance both as a collaborator and a member of my thesis committee. And Craig Miller, Axel Visel, and Lexin Li, who made my qualifying exam and committee meetings useful and productive experiences, and far less painful than they could have been.

I want to thank the computational biology PhD program, and specifically Xuan Quach and Kate Chase, for doing everything they do for the program and the students in it, and for doing it with so much kindness, patience, and a genuine desire to make it the best program it can be. And the fellow students in the program, past and present. I could not have asked for a better student community. Special thanks to Shaked Afik, who went out of his way to welcome me into the program before I was even in the program, and Prakruthi Burra, who I had the privilege of mentoring and the pleasure of befriending.

I want to thank my cohort for making my time at Berkeley so rewarding. Nick Everetts: beyond the well-appreciated humor and nonsense, you were a great friend and a wonderful person throughout. Your friendship means a lot. Amanda Mok and Sandra Hui: for making me feel slightly less crazy when I grew impatient with the world. I am a better person because of you, and I am forever grateful. Jared Bennett: thank you for being an awesome person and a great drinking buddy.

I owe a debt of gratitude to all the people who supported me through these years. Peleg Dvir, for believing in me far more than I believed in myself. Tomer Fridel, for never backing down from digging deeper, for never judging, and for never telling me what I wanted to hear.

I don't know what I would have done without you. Ehud Karavani, for being the smartest person I know and still letting me bounce ideas off you whenever I needed to. Nir Moneta, for knowing me for as long and as well as you do, and still putting up with me, somehow. You are a remarkable person and a constant inspiration. And Constance Thorsnes, for going above and beyond to teach me how to hold my head above water.

This dissertation is dedicated to my family. My parents, for encouraging me to live life on my own terms even when it meant moving away, and for always reminding me that I may be far away but I am never alone. My siblings, Stav, Omer, and Yarden: for being the incredibly inspiring people you are, for never taking me too seriously, and for not letting the physical distance get between us. I am beyond lucky to have you in my life. And Ruby and Nelly: missing out on spending time with you was the hardest part of this entire experience, by far. I can't wait to make up for lost time.

Finally, I am endlessly grateful to Lennon Zheng. For being the absolute best person I know, and challenging and inspiring me to be the best I can be. Every moment has been made better by having you there with me. Thank you.

Chapter 1

Introduction

One of the primary ways by which cells control cellular processes and responses to stimuli is by accurately adjusting the expression levels of select genes. This process is orchestrated by a complex network of interdependent factors, including chromatin conformation, regulatory regions of DNA, and a class of peptides termed transcription factors (TFs). Transcription factors typically regulate gene expression by binding specific sequence patterns in the genome (DNA binding motifs) and interacting with the transcriptional machinery at the transcription initiation sites of nearby gene. These binding sites, termed Transcription Factor Binding Sites (TFBSs), reside within regulatory areas of the genome known collectively as cis-regulatory elements (CREs). CREs that reside directly upstream of the transcription start site of their target gene are termed promoters, and distal sites are termed enhancers. CREs are crucial elements in regulatory networks, providing the mechanism by which TFs target specific genes. Promoters are considered essential for transcription initiation, whereas enhancers allow for the expression levels to be adjusted in a context-specific manner, controlling the timing, location, and precise levels of expression [1]. Their importance is evidenced by the effects of variations in enhancers regions, including morphological effects and diseases [2, 3, 4]. Understanding the roles and mechanisms of cis-regulation, and specifically enhancers, has been an outstanding challenge and significant focus of the scientific community [5]. The work presented in this dissertation relates to two of the main open questions regarding cis-regulation: (1) cis-regulatory heterogeneity and single-cell chromatin accessibility; and (2) high-throughput functional characterization of putative enhancers.

1.1 cis-Regulatory Heterogeneity and Single-Cell Chromatin Accessibility

The DNA of eukaryotes is looped around protein complexes (termed histones) that together with the DNA for a structure called chromatin. The basic unit of chromatin is called a nucleosome, which is a complex of two copies each of four types of histones (H2A, H2B, H3, and H4; total of eight proteins) and the segment of DNA wrapped around them. This structure can have varying degrees of compactness: tightly-packed ("closed") chromatin limits interactions between the DNA and the nuclear environment and therefore inhibits transcription; accordingly, loosely-packed ("open") chromatin more easily allows for interactions

and enables transcription. Chromatin compactness can be regulated by chemical modifications of the underlying histones. Enhancers typically reside in regions of open chromatin which have characteristic histone modifications, such as acetylation of lysin 27 of the H3 histone (H3K27ac), and methylation of lysic 4 of the H3 histone (H3K4me1). This allows enhancers to be identified in a genome-wide manner using assays that identify where specific histone marks are bound to the DNA, like chromatin immunopercipitation sequencing (ChIP-seq [6]). Alternatively, one can identify functional regions of the genome by assaying the chromatin accessibility landscape, which determines which areas of the genome are accessible (open chromatin) or inaccessible (closed chromatin). One popular chromatin accessibility technique is Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq [7]). Briefly, ATAC-seq utilizes an enzyme (a mutated hyperactive Tn5 Transposase) that cleaves the DNA in accessible regions of the genome and tags them with sequencing adapters. These adapters enable these tagged fragments of DNA to be purified, amplified using PCR, sequenced, and computationally mapped back to the reference genome. Regions of open chromatin are more likely to be cleaved by the transposase, and are therefore characterized by a localized accumulation of fragments. In this manner ATAC-seq can map the chromatin accessibility landscape of the entire genome, which can be highly informative: both by identifying putative enhancers, and by comparing the landscapes between biological conditions or stimuli, thereby exposing the regulatory underpinnings of the cellular response.

Single-cell sequencing technologies make it possible to perform these assays in single-cell resolution. Specifically, single-cell ATAC-seq (scATAC-seq), which identifies accessible regions in the genome, has become increasingly popular. This technology enable studying the heterogeneity in chromatin landscapes within seemingly homogeneous cell populations, shedding light on potentially distinct subpopulations that are transcriptionally identical but are epigenetically primed for different responses or cell fates. This is especially promising in processes and biological systems in which heterogeneity plays an important role, such as differentiation and within the immune system.

The promise of this technology is exemplified in chapter 2 of this dissertation. Briefly, single cell ATAC-seq was used, alongside various other assays, to study the effects of the flu vaccine on the innate immune system. The results showed that the vaccine for avian influenza (H5N1) with the added AS03 adjuvant caused a subset of classical monocytes to activate distinct epigenetic programs. Specifically, CREs that harbor binding motifs the correspond to TFs of the pro-inflammatory AP-1 family had decrease accessibility 30 days after vaccination, whereas CREs harboring motifs of the anti-viral interferon response factors (IRF) family had increased accessibility. This epigenetic priming, encoded in the accessibility landscape of these cells, translated to long-term non-specific protection to viral infections that are unrelated to the H5N1 flu (Zika, Dengue).

This previously unknown effect, rooted in changes to the regulatory network, demonstrates the potential of single-cell accessibility data. However, the computational methodology used for analyzing this data remains limited. The statistical properties of mostly-binary scATAC-seq data differ substantially from those of bulk assays or count-based single cell assays. However, methods developed for mostly-binary data, largely designed for natural

language processing, do not account for issues intrinsic to biological and single-cell data: high noise levels, high sparsity, and batch effects. For this reason, chapter 3 of this dissertation describes PeakVI, a deep generative model that uses computational neural networks to model scATAC-seq data while accounting for all unique properties of the data. PeakVI outperforms current methods in batch correction and the ability to robustly and accurately identify differentially accessible regions.

Accessibility-based assays can be highly informative but, since not all accessible regions having a regulatory function [8, 9], are inherently limited. This means that accessibility alone is not sufficient to infer the regulatory effect on transcription of a given CRE. Recent advances in sequencing technologies now enable assaying multiple different types of biological data (data modalities) from the same single cells [10, 11]. This allows the chromatin accessibility landscape to be measured alongside the transcriptional profile of each single cell, which in turn enables directly correlating changes in accessibility with changes in transcription. This technology comes with significant computational challenges. Namely, obtaining a representation of cell state that reflects both gene expression and chromatin accessibility, which can then be used to identify different populations of cells and response gradients. Additionally, in cases where some cells only have one data type available, the ability to estimate the missing modality from the observed modality (i.e estimating the transcriptional profile of a cell for which only the accessibility landscape was observed, or vice versa). To address these needs, chapter 4 of this dissertation describes MultiVI, a deep generative model which build upon existing models developed in the lab (including PeakVI [12], scVI [13], totalVI [14]) and enables multi-modality analysis. MultiVI fits a low-dimensional probabilistic space that uses information from all observed modalities, and allows for integration of multi-modal datasets with single-modality datasets. This additional ability can be used to reanalyze existing single-modality datasets and reach novel discoveries.

Accessibility-based assays can be highly informative, but are inherently limited by not all accessible regions having a regulatory function [8, 9]. This means that accessibility alone is not sufficient to infer the regulatory effect on transcription of a given CRE. However, recent advances in sequencing technologies now enable assaying multiple different data modalities from the same single cells [10, 11]. This allows the chromatin accessibility landscape to be measured alongside the transcriptional profile of each single cell, which in turn enables directly correlating changes in accessibility landscapes with changes in transcriptional profiles. This molecular technology comes with significant computational challenges, namely the ability to jointly model different data modalities that reflect a single underlying biological state. To address this, chapter 4 of this dissertation describes MultiVI, a deep generative model which build upon existing models developed in the lab (including PeakVI [12], scVI [13], totalVI [14]) and enables multi-modality analysis. MultiVI fits a latent space that uses information from all modalities, and allows for integration of multi-modal datasets with single-modality datasets. This additional ability can be used to reanalyze existing single-modality datasets and reach novel discoveries.

1.2 High-throughput Functional Characterization of Enhancer Sequences

Single-cell assays for chromatin accessibility are promising techniques for identifying systemic regulatory changes and generating hypotheses about the location and function of CREs. However, they do not provide direct evidence of regulatory function, and do not isolate the function of specific CREs within a given biological context. On the other hand, classical reporter assays are typically limited in scope to a handful of sequences that can be tested simultaneously, whereas genome-wide assays can generate thousands of testable hypotheses. Recent advances in reporter assays address these limitations with a set of procedures collectively termed Massively Parallel Reporter Assyas (MPRAs).

In MPRAs, synthetic DNA constructs that contain a minimal transcriptional unit are introduced to cells, each with a distinct candidate regulatory sequence of interest, along with a minimal promoter and a unique "barcode" sequence. The candidate regulatory sequence is assumed to regulate this transcriptional unit similarly to how it would regulate the native target gene. The transfected cells then undergo both DNA and RNA sequencing, where the DNA sequencing captures the baseline abundance of each construct, and the RNA sequencing captures the transcriptional output of each construct. The normalized transcriptional output is then a direct quantitative measure of the regulatory function of the candidate sequence. Relying on the vast combinatorial space of sequence barcodes (as opposed to fluorescent markers used in classical reporter assays [15]) enables measuring the activity of many thousands of sequences in a single experiment. However, To mitigate potential regulatory and post-transcriptional effects from the barcodes themselves, each candidate sequence is typically associated with several barcodes, ranging from < 10 to > 100. The technology therefore introduces several non-trivial normalization and quantification steps that require specialized methods.

Chapter 5 of this dissertation introduces MPRAnalyze, a model for analyzing MPRA data using two nested generalized linear models (GLMs). MPRAnalyze models technical effects in both DNA and RNA counts and shares dispersion information between them, allowing for robust quantification of induced transcriptional activity, as well as comparative analyses of both biological conditions and sequence variants.

MPRAnalyze is then used extensively in chapter 6, in which high-throughput assays were used to identify candidate enhancers that may play a role during early neural differentiation from human embryonic stem cells (hESCs) to neuro-progenitor cells (NPCs). Of these select candidates, several thousands were then included in an MPRA experiment that included seven time points along the course of 72 hours after induced differentiation (using dual SMAD inhibition). MPRAnalyze was then used to quantify the regulatory effect of each candidate, classify validated functional enhancers, and identify enhancers with a significant temporal pattern along the different time points. The analysis results showed that sequences with binding motifs corresponding to TFs known to have a regulatory role in early differentiation indeed corresponded to consistent temporal activity. For instance, pluripotent TFs (e.g

NANOG, POU3F1) were enriched in sequences that were preferentially active in early time points. The analysis also identified novel TFs, the regulatory role of which in early neural differentiation was then validated using additional experiments.

A single enhancer region can harbor multiple functional TFBSs, and several models have been presented as to how these interact to produce a given regulatory effect. The two leading models are the "billboard" model, in which each TFBS has an independent and additive contribution to the overall, and the "enhanceosome" model of all-or-nothing fully dependnt interactioneffect [16, 17]. The length of each sequence measured in the above MPRA experiement was 171 base pairs, whereas TF binding sites are typically around 10 base pairs long [18], so each tested sequence included in the MPRA has the potential of harboring multiple functional binding sites. While the MPRA allowed us to identify enhancers that are involved with neural differentiation, it did not identify the exact elements within each sequence that have a regulatory function. A follow-up study, presented in chapter 7 of this dissertation, was then designed to identify and characterize the specific functional elements within each enhancer sequence. For this purpose, validated sequences were selected and known binding motifs were identified in those sequences and systematically perturbed (by replacing the known motif with some supposedly inert sequence). This approach, termed Perturbation MPRA, demonstrates another advantages of MPRAs - in the ability to test designed sequences that are not native to the genome. These perturbed sequences were included in another MPRA with a similar experimental design. The data was analyzed with MPRAnalyze to identify functions regulatory sites (FRSs) within each sequence, as well as different classes of regulatory functions and cooperation patterns.

The results indicate that enhancer sequences with an overall activating effect (the sequence increases the transcription rate of the target gene) often harbor binding sites with an inhibiting effect (perturbing the site increases the induced transcription rate), indicating that many enhancers are composed of mixed activating and inhibiting sites that fine tune the regulatory effect. The sequences included in the experiment also included combinatorial perturbation of pairs of sites (i.e perturbing either site separately as well as both at the same time), which showed that neither model for TFBS cooperation explains all results, and that different enhancers are governed by different cooperation patterns. Overall our studies demostrate the power of MPRAs in precisely characterizing both the functional activity of a given sequence, and the elements within the sequence that encode that functional activity.

1.3 References

[1] Fumitaka Inoue and Nadav Ahituv. "Decoding enhancers using massively parallel reporter assays". en. In: *Genomics* 106.3 (Sept. 2015), pp. 159–164. ISSN: 0888-7543, 1089-8646. DOI: 10.1016/j.ygeno.2015.06.005. URL: http://dx.doi.org/10.1016/j.ygeno.2015.06.005.

[2] Lucia A Hindorff et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". en. In: *Proc. Natl. Acad. Sci. U.*

S. A. 106.23 (June 2009), pp. 9362–9367. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0903103106. URL: http://dx.doi.org/10.1073/pnas.0903103106.

[3] Sean B Carroll. "Evolution at two levels: on genes and form". en. In: *PLoS Biol.* 3.7 (July 2005), e245. ISSN: 1544-9173, 1545-7885. DOI: 10.1371/journal.pbio.0030245. URL: http://dx.doi.org/10.1371/journal.pbio.0030245.

[4] Danielle Welter et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations". en. In: *Nucleic Acids Res.* 42.Database issue (Jan. 2014), pp. D1001–6. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt1229. URL: http://dx.doi.org/10.1093/nar/gkt1229.

[5] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11247. URL: http://dx.doi.org/10.1038/nature11247.

[6] Mark J Solomon, Pamela L Larsen, and Alexander Varshavsky. "Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene". In: *Cell* 53.6 (1988), pp. 937–947.

[7] Jason D Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". en. In: *Curr. Protoc. Mol. Biol.* 109 (Jan. 2015), pp. 21.29.1–21.29.9. ISSN: 1934-3639, 1934-3647. DOI: 10.1002/0471142727.mb2129s109. URL: http://dx.doi.org/10.1002/0471142727.mb2129s109.

[8] Nathaniel D Heintzman et al. "Histone modifications at human enhancers reflect global cell-type-specific gene expression". en. In: *Nature* 459.7243 (May 2009), pp. 108–112. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07829. URL: http://dx.doi.org/10.1038/nature07829.

[9] Axel Visel et al. "ChIP-seq accurately predicts tissue-specific activity of enhancers". en. In: *Nature* 457.7231 (Feb. 2009), pp. 854–858. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07730. URL: http://dx.doi.org/10.1038/nature07730.

[10] Junyue Cao et al. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". en. In: *Science* 361.6409 (Sept. 2018), pp. 1380–1385. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0730. URL: http://dx.doi.org/10.1126/science.aau0730.

[11] Eleni P Mimitou et al. "Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells". In: *Nature Biotechnology* (2021), pp. 1–13.

[12] Tal Ashuach et al. "PeakVI: A Deep Generative Model for Single Cell Chromatin Accessibility Analysis". en. Apr. 2021. DOI: 10.1101/2021.04.29.442020. URL: https://www.biorxiv.org/content/10.1101/2021.04.29.442020v1.

[13] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". en. In: *Nat. Methods* 15.12 (Dec. 2018), pp. 1053–1058. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: http://dx.doi.org/10.1038/s41592-018-0229-2.

[14] Adam Gayoso et al. "Joint probabilistic modeling of single-cell multi-omic data with totalVI". In: *Nature Methods* 18.3 (2021), pp. 272–282.

[15] Axel Visel et al. "VISTA Enhancer Browser–a database of tissue-specific human enhancers". en. In: *Nucleic Acids Res.* 35.Database issue (Jan. 2007), pp. D88–92. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkl822. URL: http://dx.doi.org/10.1093/nar/gkl822.

[16] François Spitz and Eileen E M Furlong. "Transcription factors: from enhancer binding to developmental control". en. In: *Nat. Rev. Genet.* 13.9 (Sept. 2012), pp. 613–626. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3207. URL: http://dx.doi.org/10.1038/nrg3207.

[17] Hannah K Long, Sara L Prescott, and Joanna Wysocka. "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution". en. In: *Cell* 167.5 (Nov. 2016), pp. 1170–1187. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.09.018. URL: http://dx.doi.org/10.1016/j.cell.2016.09.018.

[18] Alexander J Stewart, Sridhar Hannenhalli, and Joshua B Plotkin. "Why transcription factor binding sites are ten nucleotides long". en. In: *Genetics* 192.3 (Nov. 2012), pp. 973–985. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.112.143370. URL: http://dx.doi.org/10.1534/genetics.112.143370.

Part I

cis-Regulatory Heterogeneity and Single-Cell Chromatin Accessibility

Chapter 2

The single-cell epigenomic and transcriptional landscape of immunity to influenza vaccination

This chapter was published in *Cell* (2021), and is included here as published. The authors on the paper are:

Florian Wimmers[1], Michele Donato[1,2,*], Alex Kuo[1,3,*], Tal Ashuach[4,*], Shakti Gupta[5], Chunfeng Li[1], Mai Dvorak[1,3], Mariko Hinton Foecke[1,3], Sarah E. Chang[1,3], Thomas Hagan[1,13], Sanne E. De Jong[1], Holden T. Maecker[1], Robbert van der Most[6], Peggie Cheung[3], Mario Cortese[1], Steven E. Bosinger[7], Mark Davis[1,8,9], Nadine Rouphael[10], Shankar Subramaniam[5], Nir Yosef[4,11], Paul J. Utz[1,3], Purvesh Khatri[1,2], Bali Pulendran[1,8,12,†]

1. Institute for Immunity, Transplantation and Infection, School of Medicine, Stanford University, Stanford, CA 94305, USA
2. Department of Medicine, Division of Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA 94305, USA
3. Department of Medicine, Division of Immunology and Rheumatology, Stanford University School of Medicine, Stanford, CA 94305, USA
4. Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, Berkeley, CA 94720, USA
5. Department of Bioengineering, University of California, 9500 Gilman Drive MC 0412, San Diego, La Jolla, CA 92093, USA
6. GSK, 1330 Rixensart, Belgium
7. Emory Vaccine Center, Emory University School of Medicine, Atlanta, Georgia 30322, USA
8. Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA
9. Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA
10. Hope Clinic of the Emory Vaccine Center, Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Decatur, GA 30030, USA
11. Chan-Zuckerberg Biohub, San Francisco, CA, USA
12. Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
13. Present Address: Division of Infectious Diseases, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

∗ these authors contributed equally to the work

† Corresponding author

2.1 Abstract

Emerging evidence indicates a fundamental role for the epigenome in immunity. Here, we mapped the epigenomic and transcriptional landscape of immunity to influenza vaccination in humans at the single-cell level. Vaccination against seasonal influenza induced persistently diminished H3K27ac in monocytes and myeloid dendritic cells (mDCs), which was associated with impaired cytokine responses to TLR stimulation. Single-cell ATAC-seq analysis revealed an epigenomically distinct subcluster of monocytes with reduced chromatin accessibility at AP-1-targeted loci after vaccination. Similar effects were observed in response to vaccination with the AS03-adjuvanted H5N1 pandemic influenza vaccine. However, this vaccine also stimulated persistently increased chromatin accessibility at interferon response factor (IRF) loci in monocytes and mDCs. This was associated with elevated expression of antiviral genes and heightened resistance to the unrelated Zika and Dengue viruses. These results demonstrate that vaccination stimulates persistent epigenomic remodeling of the innate immune system and reveal AS03's potential as an epigenetic adjuvant.

2.2 Introduction

Recent research has highlighted a central role for the epigenome in the regulation of fundamental biological processes. The epigenome can maintain particular chromatin states over prolonged periods of time that span generations of cells, thus enabling the durable storage of gene expression information [1]. In the context of the immune system, epigenomic events have been described during hematopoiesis [2, 3, 4], generation of immunological memory and exhaustion in T lymphocytes [5, 6, 7], and the development of B and plasma cells [8, 9]. Recent studies have also revealed that epigenomic changes in monocytes [10, 11, 12] and NK cells [13] imprint a form of immunological memory in the innate immune system [14].

The concept of epigenetic imprinting on the innate immune system has acquired a particular significance in the context of vaccination [15]. Vaccination with live-attenuated BCG has been shown to induce epigenomic changes in monocytes [10, 11], and it has been suggested that such changes result in a durable state of innate activation. However, the extent to which such epigenomic imprinting, observed with BCG vaccination, reflects a more general phenomenon with other vaccines is an open question. Furthermore, the critical parameters that determine vaccination-induced epigenomic imprinting, such as the type of vaccine or adjuvant used, or the impact of the microbiome, are not known. Notably, previous studies identified transcriptional and protein level heterogeneity within monocyte and dendritic cell populations [16, 17, 18, 19, 20, 21, 22, 23]. How this cellular heterogeneity affects epigenomic imprinting during and immune response to a vaccine or to any stimulus is entirely unknown. Recently, researchers have used systems biology approaches to comprehensively analyze the transcriptional, metabolic, proteomic and cellular landscape in response to vaccination in humans, and identified correlates and mechanisms of vaccine immunity [24, 25, 26, 27, 28, 29, 30, 31, 15]. Despite these advances, a comprehensive systems biology assessment of the

human epigenomic landscape during an immune response, particularly at the single-cell level, is missing.

In the current study, we used single-cell techniques, including EpiTOF (Epigenetic landscape profiling using cytometry by Time-Of-Flight) [32], single-cell ATAC-seq, and single-cell RNA-seq, to study the epigenomic and transcriptional landscape of immunity to seasonal and pandemic influenza vaccination in humans. We found that vaccination with the trivalent inactivated seasonal influenza vaccine (TIV) induced global changes to the chromatin state in multiple immune cell subsets, which persisted for up to six months after vaccination. These changes were most pronounced in myeloid cells, which demonstrated a transition to inaccessible chromatin in loci targeted by AP-1 transcription factors, and reduced cytokine production. Single-cell analysis revealed distinct subclusters within the monocyte population that were characterized by differences in AP-1 accessibility. Vaccination with the AS03-adjuvanted H5N1 pre-pandemic influenza vaccine additionally induced increased chromatin accessibility at IRF and STAT loci, and heightened resistance against heterologous viral infection during in-vitro culture.

2.3 Results

Global epigenomic reprogramming of immune cell subsets after vaccination with TIV

To determine how immunization with TIV affects the epigenomic landscape of the immune system at the single-cell level, we employed EpiTOF to analyze a cohort of 21 healthy individuals aged 18-45 before and after TIV administration (DataS1, related to STAR Methods). All subjects received TIV on day 0. To determine the impact of the gut microbiota on the epigenomic immune cell landscape, a subgroup of ten subjects received an additional oral antibiotic regimen, consisting of neomycin, vancomycin, and metronidazole, between days -3 and 1 (Figure 1A). Our previous work with this cohort had demonstrated that antibiotics administration induced significant changes in the transcriptional and metabolic profiles of peripheral blood mononuclear cells (PBMCs) [33]. Therefore, we hypothesized antibiotics administration would induce epigenomic reprogramming of PBMCs. To test these hypotheses, we developed two EpiTOF panels and probed the global levels of 38 distinct histone marks, including acetylation, methylation, phosphorylation, ubiquitination, citrullination, and crotonylation, in 21 distinct immune cell subsets (DataS2, related to STAR Methods). Using EpiTOF, we analyzed PBMCs (Figure S1A, related to Figure 1) isolated at day -21 and 0 prior to vaccination, and days 1, 7, 30, and 180 after vaccination. While the frequency of immune cell populations did not change significantly between these time points, we observed a trend towards reduced fractions of myeloid cells in some subjects at later time points, and a transient increase in the proportion of pDCs in response to antibiotics treatment (days 0, 1) (Figure S1B), in line with previous observations [33].

Next, we extracted the histone modification information for each subset and generated a

UMAP representation of the epigenomic immune cell landscape (Figure 1B). In the UMAP space, lymphoid cells separated from myeloid cells while hematopoietic progenitors (CD34+) showed a unique epigenetic pattern distinct from fully differentiated immune cells. When generating a sample-level UMAP representation, there was a segregation of samples isolated at various times after vaccination, especially at day 30 relative to baseline samples, indicating TIV-induced epigenetic changes (Figure 1C, left). In contrast to previous studies using blood transcriptomics and metabolomics [33], antibiotics status had no measurable impact on histone modification levels and samples from antibiotics-treated and control subjects were intermixed (Figure 1C, right). Rather, we observed changes in the acetylation, methylation, and phosphorylation states of several histone marks in response to vaccination, regardless of exposure to antibiotics (Figure S1C, D, related to Figure 1). To enhance statistical power, we combined both groups for downstream analyses. We detected an increase in several histone methylation marks in CD34+ cells and a decrease in multiple acetylation marks in myeloid cells in day 30 samples over baseline (Figure 1D). Also, we observed increased H2BS14ph in multiple immune cell subsets at day 30 after vaccination (Figure S1C). Elevated H2BS14ph has been shown to occur during apoptosis [34, 35, 36]. However, we did not observe reduced cell viability at any time point (Figure S1A, related to Figure 1), suggesting H2BS14ph functions independent of apoptosis in post-vaccination immune cells. Notably, H2BS14ph is catalyzed by Mst1/STK4, whose immune modulatory role has been reported [37, 38, 39, 40].

Persistent epigenomic reprogramming in myeloid cells

Classical monocytes and myeloid dendritic cells (mDCs) were characterized by repressed H2BK5ac, H3K9ac, H3K27ac, and H4K5ac at day 30 after vaccination (Figure 1E). PADI4, an arginine deiminase catalyzing histone citrullination, was also repressed in these cells. Notably, pairwise correlation analysis identified high correlation coefficients between acetylation marks and PADI4 (Figure S1E, related to Figure 1). PADI4 has been implicated in monocyte development, and inflammation [32, 41, 42, 43]. Longitudinal analysis demonstrated a time-dependent decrease of the four histone acetylation marks and PADI4, which showed the greatest repression at day 30 and largely returned to baseline levels at day 180 (Figure 1F). Blood transcriptomics data obtained from PBMCs of the same subjects at early time points after vaccination [33] revealed downregulation of histone acetyltransferases CREBBP/CBP [44] and KAT6A [45] at days 1, 3, and 7 after vaccination (Figure S2A, related to Figure 1). In contrast, various histone deacetylases showed increased expression (Figure S2A, related to Figure 1). Moreover, the expression of lysine methyltransferase EZH2 was elevated (Figure S2A, related to Figure 1), consistent with increased H3K27me3, an antagonist of H3K27ac, in classical monocytes and myeloid dendritic cells (mDCs) (Figure 1E). Epigenomic and transcriptional analysis thus both point towards a, potentially repressive, state of hypoacetylation in myeloid cells after immunization with TIV.

Next, we investigated the TIV-induced epigenomic alterations in myeloid cells at the single-cell level (Figure 1G). By performing sub-clustering and UMAP-based dimensionality reduction analysis of mDCs and classical monocytes using the H3K27ac, H2BK5ac, H4K5ac,

H3K9ac, and PADI4 marks, we constructed the single-cell histone modification landscape. Importantly, in both cell types, single cells segregated according to vaccination time point with cells at day 0 and 1 clustering together on one side of the 2D space, and cells at day 30 occupying the opposite side (Figure 1G). Interestingly, and undetected by the bulk analysis (Figure 1F), cells at day 180 did not return to the baseline position occupied by day 0 cells but assumed an intermediate state (Figure 1G), indicating persistent epigenetic alterations that can be detected up to 6 months after immunization with TIV.

These observations raise the question of how persistent epigenetic changes lasting up to 6 months, can be maintained in monocytes and mDCs, given that these cell types have a rapid turnover of less than 7 days. Recent studies indicate that such persistent changes in circulating myeloid cells are associated with changes in the hematopoietic progenitor cell compartment in the bone marrow [46, 47, 48]. To determine if this was also evident here, we calculated the epigenomic distance of CD34+ cells to a consensus profile of differentiated lymphoid or myeloid cells (Figure S3A, related to Figure 1). We detected multiple populations of CD34+ cells based on their epigenomic distances with minor populations showing relatively short distances to differentiated immune cells, possibly resembling pre-committed clones (Figure S3B, related to Figure 1). After vaccination, the overall distance between CD34+ cells and differentiated cells was greatly reduced (Figure S3B-D, related to Figure 1) indicating a potential shift of the stem cell pool towards an immature phenotype after vaccination. At day 180, the distances returned to their pre-vaccination state.

TIV induces persistent functional changes in innate immune cells

To determine the functional consequence of the observed epigenetic changes in myeloid cells, we stimulated PBMCs from vaccinated individuals prior to vaccination or at various time points after vaccination, with cocktails of synthetic TLR ligands mimicking bacterial (LPS, Flagellin, Pam-3-Cys) or viral (pI:C, R848) pathogen-associated molecular patterns (Figure 2A). After 24h of stimulation, we measured the levels of 62 secreted cytokines in culture supernatants using a multiplexed bead-based assay. Our previous work using intracellular staining (data not shown), as well as work by others [49] has demonstrated that monocytes are the most abundant and dominant contributors to cytokine production upon in vitro stimulation of PBMCs with the aforementioned stimuli. To determine whether PBMCs from time points after vaccination showed any alterations in cytokine production, we calculated the relative change in cytokine levels compared to day 0 (Figure 2B). Indeed, using hierarchical clustering, we identified a subset of cytokines that displayed a significant reduction at day 30 after vaccination (Figure 2B red box, C). These cytokines include TNF-α, IL-1b, IL-1RA, IL-12, and IL-10, the monocytic chemokines MCP1, MCP3, ENA78 (CXCL5), and IP-10 (CXCL10), as well as the monocyte growth factor GCSF. Similar to the kinetics of the epigenomic changes, reduced cytokine responses were evident as early as day 1 to 7 after vaccination, reaching a nadir at day 30, and returning to near-baseline levels at day 180 (Figure 2D). All these cytokines were strongly induced by both TLR cocktails (Figure S4A,

related to Figure 2) and a reduction relative to day 0 was observed in both antibiotics-treated and control subjects (Figure S4B, related to Figure 2).

Next, we investigated whether there is a direct relationship between global histone modification levels and TLR-induced cytokine production. We used pairwise correlation analysis to correlate the cytokine levels in a sample with the EpiTOF histone modification levels in classical monocytes and with monocyte frequency in the PBMCs of the same sample and cell viability (Figure 2E). Strikingly, the histone acetylation marks previously identified in Figure 1C, especially H3K27ac, and PADI4 showed positive correlation with cytokine production (Figure 2E, F). In contrast, H2BS14ph and several repressive methylation marks, including H3K27me3 and H4K20me3 [50, 51], were negatively correlated with cytokine production (Figure 2E).

Next, we determined if perturbations of global histone acetylation or PADI4 activity affect TLR-induced cytokine secretion. We conducted an ex-vivo stimulation experiment using specific inhibitors for the histone acetyl transferases CBP/p300 (A-485, inhibits acetylation at H3K27, H2BK5, and H4K5, [52, 44] and PADI4 (Cl-Amidine) followed by stimulation with synthetic TLR ligands. Using flow cytometry, we assessed expression of H3K27ac and the intracellular accumulation of IL-1b and TNFa. As expected, treatment with the histone acetyl transferase inhibitor A-485 led to a concentration-dependent decrease in global histone H3K27ac levels in classical monocytes while treatment with the HDAC inhibitor TSA generated a concentration-dependent increase (Data not shown). Furthermore, treatment with the PADI4 inhibitor Cl-Amidine led to similar reductions in H3K27ac (Data not shown) in line with the strong correlation of PADI4 and H3K27ac levels in EpiTOF (Figure S1E) and the previously observed ability of PADI4 to regulate CBP/P300 [53]. Notably, none of these inhibitors influenced cell viability (Data not shown). Next, we asked whether inhibition of CBP/P300 and PADI4 has an impact on cytokine production. Indeed, treatment with A-485 led to a major diminution in the frequency of IL-1b and TNFa positive monocytes after stimulation with LPS or R848 (Figure 2G, H). Cl-Amidine treatment, strikingly, led to a complete abrogation of cytokine production in these cells (Figure 2H).

TIV induces reduced chromatin accessibility of AP-1 targeted loci in myeloid cells

To gain greater insight into the epigenomic changes induced by vaccination, we conducted ATAC-seq analysis of FACS purified innate immune cell subsets before and after vaccination (Figure 3A). After preprocessing, we retained a high-quality dataset of 51 unique samples (DataS3, related to Figure 3). To identify the molecular targets of the TIV-induced epigenomic changes, we determined genomic regions with significantly changed chromatin accessibility at day 30 after vaccination compared to day 0. Overall, we detected more than 10,000 differentially accessible regions (DARs) in CD14+ monocytes and 4,500 DARs in mDCs, while pDCs showed only minor changes (Figure 3B). In line with reduced histone acetylation levels detected by EpiTOF, the majority of DARs in monocytes and mDCs

showed a reduction in chromatin accessibility indicating reduced gene activity (Figure 3B). In contrast, comparing samples from day -21 before antibiotics treatment and day 0 during antibiotics treatment showed a modest increase in chromatin accessibility (Figure S5A, related to Figure 3). Overrepresentation analysis of antibiotics-induced DARs demonstrated enrichment of pathways associated with PAX3 targets and inflammation (Figure S5B,C, related to Figure 3) in line with upregulated expression of these genes in a previous analysis of the cohort [33]. Notably, we did not observe an antibiotics-induced change in the accessibility of any of the cytokines altered in Figure 2 and D0vD30 DARs correlated well between antibiotics-treated and control subjects (Figure S5D, related to Figure 3). Among the top 200 vaccine-induced DARs in CD14+ monocytes, we identified many immune-related genes with reduced accessibility, including several cytokines and chemokines and their associated receptors (IL18, CCL20, CXCL8, CXCL3, IL4R, IL6R-AS1), pathogen recognition receptors (CLEC5A, CLEC17A), and adhesion molecules (CD44, CD38) (Figure 3C). We also observed reduced accessibility in regions coding for molecules associated with Ras-MAPK-AP-1 signaling (RAP2B, ETS1, MAP3K8, DUSP5) (Figure 3C). Importantly, genomic loci associated with seven of the ten cytokines with reduced post-vaccination levels during ex-vivo stimulation showed reduced accessibility (Figure 2, Figure 3C, right panel). Interestingly, these reduced DARs were predominantly located in non-promoter regions (Figure 3C) suggesting the involvement of distal regulatory elements such as enhancers. Pathway analysis followed by network analysis of all DARs in CD14+ monocytes revealed two major biological themes: TLR and cytokine signaling, and genome rearrangement (Figure 3D). The TLR and cytokine cluster was dominated by pathways with mostly reduced chromatin accessibility while terms in the genome rearrangement cluster were mixed. Notably, DARs associated with signaling pathways around Ras and MAPK signaling were enriched as well.

Next, to identify regulatory patterns, we determined whether the identified DARs in each cell type were enriched for transcription factor (TFs) binding motifs. Indeed, we observed an enrichment for bZIP TFs of the AP-1 family including c-Jun and c-Fos in DARs of monocytes and mDCs, and on average DARs carrying such a motif showed a reduction in chromatin accessibility after vaccination, especially in non-promoter regions (Figure 3E). Gene set analysis of the DARs in classical monocytes further confirmed this finding and showed strong enrichment for target genes of c-Jun in DARs with reduced accessibility at day 30 (Figure 3F). In addition, we observed reduced expression of several AP-1 family members, including c-Jun, at day 30 after vaccination (Figure 3F). Using bulk transcriptomics data from previous systems vaccinology studies [54, 55, 56, 57, 58, 31], we confirmed the reduced c-Jun expression in up to 9 independent TIV studies and additionally identified a reduction in expression of the AP-1 members JUND, ATF3, FOS, FOSL2, and FOSB (Figure 3G). Similar to the histone acetylation changes, the reduction in AP-1 TF expression was first detected at day 7 after vaccination and was most pronounced at day 28 (Figure 3G).

To determine whether the observed reduction in AP-1 accessibility is related to the reduced levels of histone acetylation, we correlated the normalized accessibility levels of all genomic regions in every sample to histone mark levels in classical monocytes or mDCs isolated from the same sample. Using enrichment analysis on the highly correlated peaks (cor

coef > 0.5), we identified a significant enrichment of target genes for multiple AP-1 family members, including c-Fos and c-Jun (Figure 3H). These findings suggest the possibility of a causal link between reduced histone acetylation/PADI4 and reduced AP-1 accessibility. Indeed, previous studies described a direct physical interaction and functional co-dependence between AP-1 and the histone acetyl transferases CBP/P300 [59, 60, 61]. To investigate whether AP-1 activity and histone acetylation are also functionally linked in classical monocytes, we conducted an ex-vivo stimulation experiment using the same specific inhibitors of histone acetylation and PADI activity as in Figure 2. To gauge AP-1 activity, we used a monoclonal antibody specific for the activated form of c-Jun, phosphorylated at serine 73. While treatment with LPS or R848 alone induced a robust upregulation of p-c-Jun that can be readily detected by flow cytometry (Figure 3I, J), pre-treatment with A-485 or Cl-Amidine, which lead to reduced histone acetylation, abolished c-Jun activation completely (Figure 3I, J).

Single-cell epigenomic and transcriptional landscape of innate immunity to TIV

Previous studies using transcriptomics and proteomics approaches detected heterogeneity within monocyte and dendritic cell populations at steady state [16, 62, 20, 22]. However, it is unclear how this heterogeneity affects the epigenomic landscape in such cells and their response to vaccination. To address this, we used scATAC-seq and scRNA-seq and constructed the single-cell landscape of the innate immune response to TIV at the epigenomic and transcriptional level. PBMCs from vaccinated individuals were isolated, enriched for DC subsets, and analyzed using droplet-based single-cell gene expression and chromatin accessibility profiling (Figure 4A). After initial pre-processing, we obtained chromatin accessibility data from 62,101 cells with an average of 4,126 uniquely accessible fragments. These cells displayed the canonical fragment size distribution and showed high signal-to-noise ratio at transcription start sites (data not shown). Using UMAP representation and chromVAR TF deviation patterns, we generated an epigenomic map of the innate immune system and identified clusters for all major innate immune cell subsets, including classical and non-classical monocytes, mDCs, and pDCs (Figure 4B, DataS4, related to Figure 4). In parallel, we used the scRNA-seq data to construct a gene expression map. After pre-processing, we retained 34,368 high-quality transcriptomes with an average of 2,477 genes and 8,951 unique transcripts detected per cell. UMAP representation in combination with clustering allowed us to identify all major innate immune cell subsets (DataS4, related to Figure 4). These subsets were found at all vaccine time points and in sample of all subsets (DataS4, related to Figure 4).

Next, we used chromVAR to determine the TIV-induced changes in TF chromatin accessibility. AP-1 accessibility was strongly reduced at day 30 after vaccination in classical monocytes and mDCs (both cDC1 and cDC2) (Figure 4C) confirming our findings with bulk ATAC-seq (Figure 3). In addition, using the single-cell dataset, we observed that the re-

duction in AP-1 accessibility starts early, at day 1 after vaccination (Figure 4C) suggesting that the TIV-induced epigenomic reprogramming is imprinted during the acute phase of the vaccine response. At the gene level, we observed a reduction in the expression of multiple AP-1 members including ATF3, JUND, JUNB, FOS, and FOSL2 (Data not shown).

Next, we determined the impact of cellular heterogeneity on TIV-induced epigenomic changes. Sub-clustering analysis of classical monocytes revealed the presence of four distinct populations based on chromatin accessibility (Figure 4D, DataS4, related to Figure 4) with different temporal patterns (Figure 4E): while clusters 6 and 8 dominated the classical monocyte pool at day 0, most cells at day 30 belonged to cluster 5 (Figure 4D, E). Notably, the observed heterogeneity between the classical monocyte populations was driven by differences in AP-1 and, to a lesser extent, CEBP accessibility (Figure 4F). While the dominating clusters at day 0 (cluster 6 and 8) were high in AP-1 accessibility, cells in cluster 5, which was predominantly found at day 30, were low in AP-1 (Figure 4G, H). Cells in cluster 3 exhibited intermediate AP-1 and CEBP accessibility (Figure 4G) and their relative abundance was stable throughout vaccination (Figure 4E). Using Hotspot [63], we determined a set of genomic regions underlying the observed heterogeneity (Figure 4I). This set was enriched for regions associated with the production of proinflammatory cytokines and TLR signaling (Figure 4J), and included regions associated with the AP-1 members FOS and JUN, multiple MAP kinases, and NFKB. Importantly, cells with high AP-1 accessibility using the motif-based chromVAR analysis also displayed high accessibility at the regions coding for inflammatory genes (Figure 4H, I). Finally, using the scRNA-seq dataset, we determined the cellular heterogeneity at the transcriptional level. Although genes in the Hotspot module varied in their expression between single classical monocytes, this heterogeneity was less distinct compared to the epigenomic landscape (Figure 4K).

AS03 adjuvanted H5N1 influenza vaccine induces reduced chromatin accessibility of AP-1 loci in myeloid cells

The effects of the inactivated seasonal influenza vaccination in inducing reduced chromatin accessibility of AP-1 loci, and reduced H3K27ac and refractoriness to TLR stimulation by myeloid cells, was unexpected and seemingly at odds with prior work on the live attenuated BCG vaccine showing enhanced and persistent innate responses to vaccination, termed "trained immunity" [10]. This raised the possibility that the live BCG vaccine delivered potent adjuvant signals that stimulated persistent epigenomic changes in myeloid cells, whereas the seasonal influenza vaccine, devoid of an adjuvant, was unable to stimulate trained immunity and instead induced a form of trained tolerance. We hypothesized that the addition of an adjuvant to an inactivated influenza vaccine would induce enhanced and persistent innate responses. We used AS03, a squalene-based adjuvant containing alpha-tocopherol that induces strong innate and adaptive immune responses [64] and is included in the licensed H5N1 avian influenza vaccine [65, 66], and is being developed for COVID-19 vaccines [67, 68, 69]. We investigated the effect of AS03 on the epigenomic immune cell landscape in a

cohort of healthy individuals that were vaccinated with an inactivated split-virion vaccine against H5N1 influenza administered with or without AS03 (Figure 5A). The vaccine was administered in a prime-boost regimen and individuals received injections at day 0 and day 21.

First, we determined if AS03 affected the vaccine-induced chromatin mark changes observed after vaccination with TIV. Using EpiTOF, we analyzed PBMC samples from 18 vaccinated subjects (9 H5N1, 9 H5N1+AS03) at day 0, 7, 21, 28, and 42 and constructed the histone modification profile landscape (Figure 5B, Figure S6A, related to Figure 5). Comparing histone modification profiles at day 0 with day 42, we, unexpectedly, observed that vaccination with H5N1+AS03 induced a significant reduction in H3K27ac, H4K5ac, H3K9ac, and PADI4 in classical monocytes, four of the five highly correlated marks associated with myeloid reprogramming after TIV (Figure 5C). In contrast, vaccination with H5N1 alone did not induce significant changes in these chromatin marks. In line with these findings, we also observed in the H5N1+AS03 group but not the H5N1 group significantly reduced production of most of the innate cytokines and chemokines that were diminished after vaccination with TIV (Figure 5D). Notably, we did not detect a change in the frequency or viability of classical monocytes and all detected cytokines were strongly induced by TLR stimulation (Figure S6A, B, related to Figure 5).

Next, we analyzed subjects using scATAC-seq and scRNA-seq. PBMC samples from 4 vaccinated individuals (2 H5N1, 2 H5N1+AS03) at days 0, 21, and 42 were enriched for DC subsets and analyzed using droplet-based single-cell gene expression and chromatin accessibility profiling. After initial pre-processing, we obtained high quality chromatin accessibility data from 58,204 cells with an average of 2,745 uniquely accessible fragments which we used to generate an epigenomic map of the single immune cell landscape during H5N1 vaccination (Figure 5E, DataS5, related to Figure 5). In parallel, we used the scRNA-seq data to construct a gene expression map of the single immune cell landscape. We retained 11,213 high-quality transcriptomes with an average of 2,462 genes and 9,569 unique transcripts detected per cell and identified all major innate immune cell subsets (Figure 5E, DataS5, related to Figure 5). The different immune cell subsets were evenly distributed over all vaccine conditions and time points (DataS5, related to Figure 5).

Notably, using the scATAC-seq data, we observed a significant reduction in AP-1 accessibility in H5N1+AS03 but not H5N1 alone (Figure 5F). To further investigate the nature of the epigenomic changes after H5N1+AS03, we determined the differentially accessible regions at day 42 after vaccination compared to day 0 using a logistic regression model that corrects for library-size differences. Using overrepresentation analysis, we found that similar to TIV, the predominantly negative DARs were enriched for TLR-, and cytokine-signaling pathways as well as innate immune activity (Figure 5G). Additionally, we observed a reduction in the expression of multiple AP-1 family members, including c-Fos and c-Jun as observed after vaccination with TIV (Figure 5H).

AS03 adjuvanted H5N1 influenza vaccine induces enhanced chromatin accessibility of the antiviral response loci

Despite the reduction in AP-1 accessibility, our scATAC-seq analysis revealed an increase in chromatin accessibility at day 42 compared to day 0 for several TFs of the interferon-response factor (IRF) and STAT families (Figure 6A). These changes were observed in innate immune cell populations of subjects vaccinated with H5N1+AS03, but not with H5N1 alone. Further analysis of the kinetics revealed that these IRF- and STAT-related changes were already present after administration of the first vaccination at day 21 (pre-boost) (Figure 6B). Using the scRNA-seq dataset, we compared the expression of IRF and STAT family TFs before (day 0) and after prime (day 21) or boost (day 42) vaccination. We observed significant increases in the expression of IRF1 and STAT1 in multiple innate immune cell subsets after vaccination with H5N1+AS03, but not with H5N1 alone (Figure 6C). Notably, at a single-cell level, IRF accessibility was generally negatively correlated with AP-1 accessibility (Figure 6D), especially in dendritic cells. Next, we determined the log fold change in chromatin accessibility for peaks containing the IRF1 binding motif (Figure 6E). Indeed, we observed a significant change in accessibility in many peaks, many of which showed increased accessibility (Figure 6E). Importantly, amongst the genes with increased accessibility, we identified many interferon- and antiviral-related genes, including DDX58 (encoding the viral detector RIG-I), several interferon response genes (IFIT1, IFIT3, IFI30, ISG20, OASL), as well as the transcription factors IRF1 and IRF8. Enrichment analysis further demonstrated an enrichment of genes related to antiviral immunity (Figure 6F). In contrast to this H5N1+AS03-induced effect, vaccination with TIV induced only a transient type I IFN response (Figure S7A, related to Figure 6) and only a transient increase in IRF and STAT accessibility at day 1 after vaccination (Figure S7B, related to Figure 6). At day 30 after vaccination with TIV, the majority of significantly changed peaks with an IRF motif in fact showed reduced chromatin accessibility (Figure S7C, related to Figure 6) and scRNA-seq gene expression analysis showed no change in IRF and STAT TF gene expression at day 30 (Figure S7D, related to Figure 6). This is in line with the findings from bulk ATAC-seq where we observed an enrichment of IFN response pathway genes in DARs with reduced accessibility but not in those with enhanced accessibility (Figure S7E, related to Figure 6). Together, this indicates that, only vaccination with H5N1+AS03, but not H5N1 alone or TIV, leads to an overall enhanced accessibility of the antiviral response loci. IRF1, together with STAT1 and IRF8, orchestrates monocyte polarization in response to interferon gamma exposure [70]; IFN signaling, via JAK/TYK, leads to phosphorylation of IRF and STAT TFs [71]. Indeed, we observed an increase in IFN gamma levels in plasma of vaccinated subjects immediately after prime and boost vaccination with H5N1+AS03, but not with H5N1 alone (Figure 6G). The levels of IP10, a cytokine that is produced by monocytes in response to IFN signaling, were also elevated (Figure 6G). This raises the possibility that IFN signaling could have induced the increased IRF accessibility.

Next, we determined whether the observed epigenomic changes were translated to changes in gene expression in resting monocytes. We assessed the relationship between the change

in accessibility for significantly changed peaks carrying the IRF1 motifs and the change in gene expression for the same gene (Figure 6H). Notably, we detected a weak association between changes in accessibility and gene expression (Pearson correlation R = .082, p = .017, Chi-square p-value = 0.62) indicating that the increased accessibility has limited impact on homeostatic gene expression in these cells. Instead, we hypothesized that the changes in chromatin accessibility enhance the induced response to viral stimuli in activated cells. To test this hypothesis, we analyzed bulk RNA-seq data from 50 (16 H5N1, 34 H5N1+AS03) vaccinated subjects at time points before and after the prime (days 0, 1, 3, 7) and booster (days 21, 22, 24, 28) vaccination. As expected, antiviral- and interferon-related genes were upregulated at day 1 after each vaccination, especially in the group that received H5N1+AS03 (Figure 6I). Importantly, subjects receiving a H5N1+AS03 booster vaccination (day 22 vs day 21) displayed even higher levels of antiviral gene expression compared with the response to the prime vaccine (day 1 vs day 0) (Figure 6I). The booster vaccine was given at a time when the chromatin accessibility landscape of the innate immune system was altered suggesting that the increased accessibility in IRF loci might enable the enhanced response to the booster vaccine. To further test this hypothesis, we compared the increase in gene expression of antiviral- and interferon-related genes during booster compared to prime with the change in chromatin accessibility at day 21 compared to day 0 (Figure 6J). Indeed, we observed a significant association between both variables (Chi-square p-value = 0.01) and most genes with increased expression after booster vaccination also showed increased chromatin accessibility at the time the booster vaccine was administered. Genes with increased accessibility and enhanced expression were enriched for IRF1 transcription factor target genes (Figure 6K). This is in line with the elevated levels of IP-10 and IFN gamma observed in plasma of individuals after booster compared to prime vaccination (Figure 6G).

To determine if the observed epigenomic changes resulted in enhanced resistance to viral infections, we infected PBMCs at days 0, 21 and 42 with Dengue or Zika virus (Figure 7A). Previous studies have shown that the primary targets of these viruses in PBMCs are monocytes and dendritic cells [72, 73]. After infection, we cultured cells for 0, 24, and 48h and determined the viral copy number using qPCR (Figure 7B). We observed increased numbers of Zika and Dengue virus copies at 24h and reduction at 48h following the expected cycle of infection, replication, and eventual death of the host cells (Figure 7C). Next, we compared the viral titers at day 21 and 42 after vaccination with the pre-vaccination titers at day 0 for each subject. Strikingly, we observed a significant reduction in viral titers for both Dengue and Zika virus at day 21 after vaccination (Figure 7D). Importantly, in many subjects, we observed reduced viral titers as late as 42 days after initial vaccination (Figure 7D). Next, we determined the cytokine concentration in infected PBMC cultures at 24h after infection (Figure 7E). While Dengue and Zika virus induced the production of both IFNa and IFNg, we observed that Dengue virus suppressed the production of IP10 (Figure 7E). Finally, we correlated the change in viral titers at d0 compared to d21 with the change in vaccine-induced expression of antiviral genes that were associated with open chromatin (Figure 6J red quadrant). The majority of these genes correlated negatively with viral titers (Figure 7F). Strikingly, IRF1 was amongst the top genes negatively correlating ($r < -0.8$)

with both Dengue and Zika titers (Figure 7F). Subjects with enhanced IRF1 expression at day 21 showed reduced viral titers at the same time point (Figure 7G). In addition, the antiviral gene ANKRD22, which is involved in immunity to both Dengue and Chikungunya infection [74], was also highly negatively correlated with Zika and Dengue titers.

2.4 Discussion

Our results demonstrate that the seasonal inactivated influenza vaccine TIV and the adjuvanted pre-pandemic influenza vaccine (H5N1+AS03), induce profound and persistent global epigenomic changes blood myeloid cells, and that these epigenomic changes are linked to alterations in their function. The observed changes were most pronounced at three to four weeks after vaccination but traces of an altered epigenomic landscape were still detectable as late as 180 days after vaccination. In contrast to vaccination, antibiotics-treatment only had a transient and subtle impact on the epigenomic immune cell landscape.

Based on their molecular and functional characteristics, the observed epigenomic changes can be broadly classified into two distinct types: 1) a state of innate immune refractoriness that is characterized by reduced histone acetylation, reduced PADI4 levels, reduced AP-1 accessibility and diminished production of innate cytokines; 2) a state of heightened antiviral vigilance defined by increased IRF accessibility, elevated antiviral gene expression, increased interferon production, and, most importantly, enhanced control of heterologous viral infections. Importantly, both states can occur simultaneously and in the same single cell. While seemingly paradoxical, this superimposition might represent an evolutionary adaptation to avoid excess inflammatory host damage during late stages of infections, while maintaining a state of immunological vigilance against viral infections.

Our findings were unexpected as researchers previously observed that the live-attenuated BCG vaccine induces elevated H3K27ac levels in CD14+ monocytes which coincided with enhanced cytokine production in these cells [10]. In contrast, our results suggest that vaccine-induced epigenomic reprogramming of immune cells is more complex. Given the observed reduction in H3K27ac levels in association with immune refractoriness in this study, the possibility arises that histone acetylation could represent a bi-directional regulator, powered by epigenomically distinct states at the single-cell level, that can be raised or lowered to manipulate monocyte cytokine production accordingly, akin to a thermostat dial (Figure 7H). In addition, our data demonstrate that multiple distinct epigenomic states, such as antiviral vigilance and immune refractoriness, can be superimposed within the same cell. Importantly, this superimposition is encoded at the single-cell level as single monocytes and dendritic cells displayed elevated IRF and diminished AP-1 accessibility at the same time.

Single-cell analysis further revealed multiple clusters within the classical monocyte population based on differences in chromatin accessibility. Notably, all these epigenomic subclusters existed before vaccination and their abundance within the pool of circulating cells shifted post vaccination driving the observed bulk level changes. The transcription factor families underlying the observed heterogeneity, AP-1 and CEBP, were previously described

as key players in monocyte-to-macrophage differentiation [75] and classical-to-non classical monocyte differentiation [62], respectively. AP-1 signaling is also a central regulator of inflammation [76, 77, 78, 79, 80, 81] and our Hotspot analysis revealed differences in accessibility at inflammatory loci between epigenomic subclusters. This might suggest that distinct functional and ontogenetic fates could be imprinted within the epigenome of single monocytes. Indeed, it was recently hypothesized that classical monocytes could represent a heterologous population of cells, some pre-committed to tissue infiltration and macrophage differentiation and others primed for differentiation into non-classical monocytes [62]. The functional relevance of these epigenomically distinct subsets of myeloid cells, and their developmental relationships deserve further exploration.

With respect to the molecular mechanisms driving the epigenomic changes, we observed that the state of antiviral vigilance was associated with enhanced IRF1 and STAT1 activity and enhanced accessibility of many, but not all, loci targeted by IRF. It is established that IRF and STAT signaling promotes antiviral immunity [71] and KO models lacking IRF1 or STAT1 are more susceptible to viral infection [82, 83]. In contrast, the state of immune refractoriness was associated with a global reduction in histone acetylation and chromatin accessibility. The magnitude of the observed changes suggests a comprehensive switch towards a broadly restrictive chromatin state [1]. Our TF motif-based analysis revealed that AP-1 loci are affected by this process. AP-1 is a dimeric TF composed of different members of the FOS, JUN, ATF, and JDP families and our gene expression analysis suggests that multiple members including FOS, JUN, JUNB, and ATF3 are involved. While the role of AP-1 as a key regulator of differentiation, inflammation and polarization in myeloid cells is well described [84, 76, 77, 78, 79, 80, 85, 75, 86, 81], recent research also positions it as a central epigenomic regulator [59, 87, 88, 75, 61].

A fundamental question concerns the mechanisms by which vaccination induces such long-lasting epigenetic changes in myeloid cells. The half-lives of most DC and monocyte subsets are known to be only a few days [89, 90]. Therefore, it is unclear how epigenetic changes acquired by a DC or monocyte responding to a vaccine might be maintained for several weeks or months. Multiple explanations are conceivable: for instance, the phenomenon of innate memory could simply be caused by the effects of an ongoing adaptive immune response on innate immune cells (via paracrine signaling of cytokines such as interferon-gamma), rather than being an intrinsic property of innate immune cells. Furthermore, innate memory could be maintained by some long-lived population of innate immune cells, like memory T and B cells, and such cells could respond with enhanced vigor to a secondary vaccination or infection. Finally, it is possible that epigenetically reprogrammed myeloid cells in the periphery are continually replenished by altered myeloid cell precursors in the bone marrow [46, 47, 48]. In the context of our results, it might be possible that soluble mediators related to the vaccine response, such as interferons, could act on progenitor cells in the bone marrow [91].

Our results from the H5N1+AS03 vaccine revealed that PBMCs from vaccinated individuals control infection with the heterologous Dengue and Zika virus more efficiently than pre-vaccination PBMCs. These results, in combination with the enhanced expression of

antiviral genes and increased levels of IP-10 and IFN gamma production in-vivo, suggest that the epigenomic state of antiviral vigilance might provide broad protection against viral infections unrelated to the vaccine virus. Elevated levels of IFN gamma production at day 1 after booster vaccination were also detected with AS03-adjuvanted vaccines in the context of Hepatitis [92] suggesting that antiviral vigilance might also be induced by other vaccines containing AS03. In contrast, TIV induced a profound state of immune refractoriness at four weeks after vaccination. Nevertheless, it is important to highlight that there is ample evidence that TIV does prevent influenza (Centers for Disease Control and Prevention, 2020) and our own study found induction of robust anti-influenza antibody titers [33]. Additionally, the severity of many viral infections, including influenza and COVID-19, is closely linked to the level of inflammation-related immunopathology. Therefore, it is tempting to speculate that in addition to stimulating robust antigen-specific antibody responses, TIV vaccination could also promote disease tolerance that could conceivably help ameliorate the immunopathology caused by excessive inflammation in severe disease caused by influenza. Taken together, these results suggest that it could be beneficial to administer TIV together with an adjuvant, such as AS03. This adjuvanted TIV would exploit the beneficial effects of both epigenomics-driven states observed here: IRF-driven antiviral vigilance, and AP-1 driven immune refractoriness and disease tolerance. Indeed, a phase 3 clinical trial comparing the response to TIV vs TIV+AS03 in more than 43,000 Elderly individuals demonstrated that TIV+AS03 led to a profound reduction in all-cause death and pneumonia compared to TIV alone while influenza-specific immunity was only somewhat increased [93].

In conclusion, our results demonstrate that vaccination with AS03-adjuvanted pandemic influenza vaccine induces persistent epigenomic changes in myeloid cells, leading to an antiviral state and protection against heterologous viruses. These findings have implications for the design of future vaccines consisting of epigenetic adjuvants that provide broad, non-specific protection by manipulating the epigenomic landscape.

2.5 Methods

Experimental Subject Details

TIV

The study design was as described in phase 1 of the original publication [33] and the study was conducted in Atlanta, GA. In brief, during the 2014-2015 seasons, we enrolled a total of 21 healthy adults who were randomized into antibiotics-treated (n = 10) and control (n = 11) groups. Subjects were males and non-pregnant females between the ages of 18-40 who met the eligibility criteria as listed on clinicaltrials.gov (NCT02154061). Subject demographics are listed in (DataS1, related to STAR Methods). The antibiotics treatment consisted of a cocktail of neomycin, vancomycin, and metronidazole, all given orally, for five days. Antibiotic treatment started 3 days before the day of vaccination and continued until one day after for the antibiotics-treated group. All the study participants were vaccinated

with Sanofi Pasteur's TIV vaccine, Fluzone, for the 2014-2015 season (DataS1, related to STAR Methods). Written informed consent was obtained from each subject and protocols were approved by Institutional Review Boards of Emory University.

H5N1/H5N1+AS03

This study was conducted in Atlanta, GA. Subjects were males and non-pregnant females who met the eligibility criteria as listed on clinicaltrials.gov (NCT01910519). We enrolled a total of 50 healthy adults who were randomized into two groups receiving either the adjuvanted (H5N1+AS03, n=34) or unadjuvanted (H5N1, n=16) GSK avian influenza vaccine. While both vaccines contained split-virion (A/Indonesia/5/2005) inactivated hemagglutinin antigen, the adjuvanted vaccine additionally contained the AS03 adjuvant system (containing DL-alpha-tocopherol and squalene in an oil-in-water emulsion). Subject demographics are listed in (DataS1, related to STAR Methods). Written informed consent was obtained from each subject and protocols were approved by Institutional Review Boards of Emory University.

In-vitro stimulation and intracellular flow cytometry experiments

Samples from healthy subjects were collected at Stanford Blood Center or derived from the before-vaccination time point of a previous vaccination trial [57]. All subjects provided a confidential medical history card and completed informed consent to donate blood for clinical or research uses. We exclude subjects with known diseases, including but not limited to HIV, and hepatitis infections. Purification of buffy coat or LRS chamber from whole blood was performed at Stanford Blood Center to enrich for leukocytes prior to PBMC isolation. From the vaccination trial [57], only samples from subjects aged 26 – 41 were selected for this paper. Samples were only selected from the before vaccination time point at day 0. Written informed consent was obtained from each subject with institutional review and approval from the Emory University Institutional Review Board.

Method Details

Cells, plasma, and RNA isolation

Peripheral blood mononuclear cells (PBMCs) and plasma were isolated from fresh blood (CPTs; Vacutainer with Sodium Citrate; BD), following the manufacturer's protocol. For samples from Stanford Blood Center, PBMCs isolated from whole blood, buffy coat or LRS chamber by Ficoll density gradient centrifugation using Ficoll-Paque PLUS (GE Healthcare, #17-1440-02). PBMCs were frozen in DMSO with 10% FBS and stored at –80C and then transferred on the next day to liquid nitrogen freezers (–196C). Plasma samples from CPTs were stored at –80C. Trizol (Invitrogen) was used to lyse fresh PBMCs (1 mL of Trizol to $\sim 1.5x10^6$ cells) and to protect RNA from degradation. Trizol samples were stored at –80C.

Mass cytometry sample processing, staining, barcoding and data acquisition

Cryopreserved PBMCs were thawed and incubated in RPMI 1640 media (ThermoFisher) containing 10% FBS (ATCC) at 37°C for 1 hour prior to processing. Cisplatin (ENZO Life Sciences) was added to 10 μM final concentration for viability staining for 5 minutes before quenching with CyTOF Buffer (PBS (ThermoFisher) with 1% BSA (Sigma), 2mM EDTA (Fisher), 0.05% sodium azide). Cells were centrifuged at 400g for 8 minutes and stained with lanthanide-labeled antibodies (DataS2, related to STAR Methods) against immunophenotypic markers in CyTOF buffer containing Fc receptor blocker (BioLegend) for 30 minutes at room temperature (RT). Following extracellular marker staining, cells were washed 3 times with CyTOF buffer and fixed in 1.6% PFA (Electron Microscopy Sciences) at 1x106 cells/ml for 15 minutes at RT. Cells were centrifuged at 600g for 5 minutes post-fixation and permeabilized with 1 ml ice-cold methanol (Fisher Scientific) for 20 minutes at 4°C. 4 ml of CyTOF buffer was added to stop permeabilization followed by 2 PBS washes. Mass-tag sample barcoding was performed following the manufacturer's protocol (Fluidigm). Individual samples were then combined and stained with intracellular antibodies in CyTOF buffer containing Fc receptor blocker (BioLegend) overnight at 4°C. The following day, cells were washed twice in CyTOF buffer and stained with 250 nM 191/193Ir DNA intercalator (Fluidigm) in PBS with 1.6% PFA for 30 minutes at RT. Cells were washed twice with CyTOF buffer and once with double-deionized water (ddH2O) (ThermoFisher) followed by filtering through 35μm strainer to remove aggregates. Cells were resuspended in ddH2O containing four element calibration beads (Fluidigm) and analyzed on CyTOF2 (Fluidigm).

Bulk stimulation experiment

Aliquots of thawed PBMCs from the EpiTOF experiment described above were washed and resuspended in RPMI 1640 (Corning, 10-040-CV) containing 10% FBS (Corning, 35-011-CV) and 1x Antibiotics/Antimycotics (Lonza, 17-602E) [complete media abx] at $4x10^6$ cells/mL. 100 μL of cell solution were added to each well of a 96-well round-bottomed tissue culture plate and mixed with 100 μL of either complete media abx (unstim), a cocktail of synthetic TLR ligands mimicking bacterial pathogens (bac: 0.025 μg/mL LPS, 0.3 μg/mL Flagellin, 10 μg/mL Pam3CSK4), or a cocktail of synthetic TLR ligands mimicking viral pathogens (vir: 4 μg/mL R848, 25 μg/mL pI:C). Depending on cell numbers, PBMCs from each sample were stimulated with all 3 conditions in duplicate. After 24h of incubation at 37C and 5% CO2, cells were spun down, supernatant was carefully transferred into new plates, and immediately frozen at -80C until further analysis using Luminex.

Luminex TIV

The Luminex assay was performed by the Human Immune Monitoring Center, Stanford University School of Medicine. Human 62-plex custom Procarta Plex Kits (Thermo Fisher Scientific) were used according to the manufacturer's recommendations with modifications as follows: Briefly, Antibody-linked magnetic microbeads were added to a 96-well plate along

with custom Assay Control microbeads (Assay Chex) by Radix Biosolutions. The plates were washed in a BioTek ELx405 magnetic washer (BioTek Instruments). Neat Cell culture supernatants (25ul) and assay buffer (25ul) were added to the 96 well plate containing the Antibody-coupled magnetic microbeads, and incubated at room temperature for 1 h, followed by overnight incubation at 4°C. Room temperature and 4°C incubation steps were performed on an orbital shaker at 500–600 rpm. Following the overnight incubation, plates were washed in a BioTek ELx405 washer (BioTek Instruments) and then kit-supplied biotinylated detection Ab mix was added and incubated for 60 min at room temperature. Each plate was washed as above, and kit-supplied streptavidin–PE was added. After incubation for 30 min at room temperature, wash was performed as described, and kit Reading Buffer was added to the wells. Each sample was measured in two technical replicates where cell numbers allowed. Plates were read using a FlexMap 3D Instrument (Luminex Corporation). Wells with a bead count < 50 were flagged, and data with a bead count < 20 were excluded.

Luminex H5N1/H5N1+AS03

This assay was performed by the Human Immune Monitoring Center at Stanford University. A custom 41 plex from EMD Millipore kits was assembled and included: 1. A Premixed 38 plex Milliplex Human Cytokine/Chemokine kit (CAT# HCYTMAG-60K-PX38) 2. ENA78/CXCL5 (CAT# HCYP2MAG-62K-01) 3. IL-22 (CAT# HTH17MAG-14K-01). 4. IL-18 (HIL18MAG-66K). Manufacturer's recommendations were followed with modifications described. Briefly: neat supernatant samples (25ul) were mixed with antibody-linked magnetic beads in a 96-well plate containing assay buffer, for an overnight incubation at 4°C. Cold and Room temperature incubation steps were performed on an orbital shaker at 500-600 rpm. Plates were washed twice with wash buffer in a BioTek ELx405 washer (BioTek Instruments). Following one-hour incubation at room temperature with biotinylated detection antibody, streptavidin-PE was added for 30 minutes. Plates were washed as above, and PBS was added to wells for reading in the Luminex FlexMap3D Instrument with a lower bound of 50 beads per sample per cytokine. Each sample was measured in duplicate wells where cell numbers allowed. Custom Assay Chex control beads were added to all well (Radix Biosolutions). Wells with a bead count < 50 were flagged, and data with a bead count < 20 were excluded.

H3K27ac antibody conjugation

α-H3K27ac antibody was labeled using the Lightning-Link Rapid DyLight 488 Antibody Labeling Kit according to manufacturer's instructions (Novus Biologicals, 322-0010). In brief, 100 μg of antibody was mixed with 10 μL of LL-Rapid modifier reagent and added onto the lyophilized dye. After mixing, solution was incubated at room temperature overnight in the dark. The next morning, 10 μL of LL-Rapid quencher reagent was added.

In-vitro stimulation and intracellular flow cytometry experiments

Cryopreserved PBMCs were thawed, counted, and resuspended in RPMI 1640 (Corning, 10-040-CV) supplemented with 10% FBS (Corning, 35-011-CV) [complete media] at a concentration of $4x10^6$ cells/mL. Next, 150μL of cell suspension ($6x10^5$ cells) was added to each well of a 96-well round-bottomed tissue culture plate and mixed with 50 μL of inhibitor solution containing either Trichostatin A (TSA; CST, 9950S), A-485 (Tocris, 6387), or Cl-Amidine (EMD Millipore, 506282) in complete media. After 2h of incubation at 37C and 5% CO2, the cells were stimulated by adding either LPS (0.025 μg/mL; Invivogen, tlrl-pb5lps) or R848 (4 μg/mL; Enzo Life Sciences, ALX-420-038-M005) to the cultures. After another 2h of incubation, Brefeldin A (10 μg/mL; Sigma Aldrich, B7651-5MG) was added to all cultures and cells were incubated for a final 4h. After a total of 8h of incubation, cells were washed twice with 150 μL PBS (GE Life Sciences, SH30256.LS) and stained for viability using 100 μL of Zombie UV Fixable Viability Dye in PBS (1:1000; Biolegend, 423108). After incubating for 30 minutes at 4C in the dark, cells were washed twice with 150 μL PBS and blocked with 100 μL of PBS supplemented with 5% FBS, EDTA (2 mM; Corning, 46-034-cl), and human IgG (5 mg/mL; Sigma Aldrich, G4386-5G) [blocking buffer] for 15 minutes at 4C in the dark. After incubation, cells were stained for surface markers with 100 μL of antibody cocktail containing α-CD14 BUV805, α-CD3, CD19, CD20 BUV737, α-CD123 BUV395, α-HLA-DR BV785, α-CD16 BV605, α-CD56 PE-CY7, α-CD11c APC-eFluor780 in blocking buffer for 20 minutes at 4C in the dark. Next, cells were washed twice with 150 μL PBS, and fixed in 200 μL eBioscience Foxp3 Fixation/Permeabilization solution (ThermoFisher Scientific, 00-5523-00) for 30 minutes at 4C in the dark. Afterwards, cells were washed twice with 100 μL eBioscience Foxp3 permeabilization buffer and blocked with 100 μL permeabilization buffer containing human IgG (5 mg/mL) overnight at 4C in the dark. Cells were washed and stained for intracellular markers with 25 μL of antibody cocktail containing α-IL-1b Pacific Blue, α-H3K27ac DyLight 488, α-TNFa PE-Dazzle, α-p-c-Jun PE, and α-H3 AF647 in permeabilization buffer containing human IgG (5 mg/mL) for 60 minutes at 4C in the dark. Finally, cells were washed twice with 150 μL of permeabilization buffer, resuspended in 100 μL PBS containing 0.5% FBS and 2 mM EDTA [FACS buffer], and acquired using a BD FACSymphony flow cytometer. Data was analyzed using Flowjo X software (BD). Briefly, cells were identified via FSC/SSC, doublets were discarded via SSC-A/SSC-H and FSC-A/FSC-H gates, and dead cells were discarded as Zombie UV Fixable Viability Dye high. Monocytes were then identified as CD3-CD19-CD20- and CD14+.

FACS sorting – bulk ATAC-seq/RNA-seq

Cryopreserved PBMCs were thawed, washed, counted, and resuspended in PBS (GE Life Sciences, SH30256.LS). $5-10x10^6$ cells were washed once more with 2 mL of PBS and stained for viability using 500 μL of Zombie UV Fixable Viability Dye in PBS (1:1000; Biolegend, 423108). After incubating for 30 minutes at 4C in the dark, cells were washed with 2 mL of PBS and resuspended in 500 μL blocking buffer. After spinning cells down, supernatant

was discarded, and cells were resuspended in 50 μL antibody cocktail containing α-CD3, CD19, CD20 BUV737, α-CD123 BUV395, α-HLA-DR BV785, α-CD14 BV605, α-CD56 BV510, α-CD1c BV421, α-CD327 AF488, α-CD370 PE, α-CD11c APC-eFluor780, α-CD15 AF700, and α-Axl APC in blocking buffer. Cells were stained for 15 minutes at 4C in the dark. Finally, cells were washed with 2 mL of FACS buffer, resuspended in PBS containing 5% FBS at $10 - 20x10^6$ cells/mL, and stored at 4C before sorting on a FACS Aria Fusion (BD). During sort, live innate cells were identified by gating on Viability Dye- CD3-CD19-CD20- cells. Within this population, CD14+ monocytes were identified as CD14+, mDCs were identified as CD14-CD56-HLA-DR+CD16-CD11c+CD123-, and pDCs were identified as CD14-CD56-HLA-DR+CD16-CD11c-CD123+.

Omni ATAC-seq of purified immune cells

Atac was performed on purified innate immune cell subsets immediately after sorting based on the low-input Omni-Atac protocol described before [94]. In brief, 1,500 – 5,500 cells were washed with ATAC resuspension buffer (10 mM Tris-HCl pH 7.5 [Invitrogen, 15567027], 10 mM NaCl [Invitrogen, AM9760G], 3 mM MgCl2 [Invitrogen, AM9530G], in water [Invitrogen, 10977015]) and supernatant was carefully aspirated, first using a P1000, then a P200 pipette. Next, 10 μL transposition mix (0.5 μL Tn5, 0.1 μL 10% Tween-20, 0.1 μL 1% Digitonin, 3.3 μL PBS, 1 μL water, and 5 μL tagmentation buffer) was added to the pellet and cells were resuspended by pipetting up and down 6 times. Tagmentation buffer was prepared locally by resuspending 20 mM Tris-HCl pH 7.5, 10 mM MgCl2, and 20% Dimethyl Formamide (Sigma Aldrich, D4551-250ML) in water. Cells were incubated at 37C for 30 minutes under constant mixing. After tagmentation, the reaction was cleaned up using the MinElute PCR Purification Kit (Qiagen, 28006) according to manufacturer's instructions. Cleaned DNA was eluted in 21 μL of elution buffer, stored at -20C, and shipped to Active Motif for sequencing library preparation. At Active Motif, tagmented DNA was amplified with 10 cycles of PCR using customized Nextera PCR Primers 1 and 2 (see Key Resource table), and purified using Agencourt AMPure SPRI beads (Beckman Coulter, A63882). Resulting material was quantified using the KAPA Library Quantification Kit for Illumina platforms (Roche, 07960255001), and sequenced with PE42 sequencing on the NextSeq 500 sequencer (Illumina).

Bulk RNA-seq of purified immune cells

Bulk RNA-seq was performed on purified CD14+ monocytes after sorting. In brief, after sorting, 5,500 cells were washed, resuspended in 350 μL chilled Buffer RLT (Qiagen, 79216) supplemented with 1% beta-Mercaptoethanol (Sigma, M3148-25ML), vortexed for 1 minute, and immediately frozen at -80C. RNA was isolated using the RNeasy Micro kit (Qiagen, 74004) with on-column DNase digestion. RNA quality was assessed using an Agilent Bioanalyzer and total RNA was used as input for cDNA synthesis using the Clontech SMART-Seq v4 Ultra Low Input RNA kit (Takara Bio, 634894) according to the manufacturer's instruc-

tions. Amplified cDNA was fragmented and appended with dual-indexed bar codes using the NexteraXT DNA Library Preparation kit (Illumina, FC-131-1096). Libraries were validated by capillary electrophoresis on an Agilent 4200 TapeStation, pooled at equimolar concentrations, and sequenced on an Illumina NovaSeq6000 at 100SR, yielding 20-25 million reads per sample.

FACS sorting – scATAC-seq/RNA-seq

Cryopreserved PBMCs were thawed, and innate immune cell subsets were isolated using FACS as described above (FACS sorting – bulk ATAC-seq/RNA-seq). Within the live gated cells, CD14+ monocytes were identified as CD14+ (fraction A) while a mixture of the remaining monocyte and dendritic cell subsets was identified as CD14-CD56-HLA-DR+ (fraction B). After sorting and depending on the number of isolated cells, fraction A and B were mixed at a 2:1 ratio to yield a solution of monocytes and dendritic cells enriched for CD14- cells.

scRNA-seq

FACS-purified cells were resuspended in PBS supplemented with 1% BSA (Miltenyi), and 0.5 U/μL RNase Inhibitor (Sigma Aldrich). About 9,000 cells were targeted for each experiment. Cells were mixed with the reverse transcription mix and subjected to partitioning along with the Chromium gel-beads using the 10X Chromium system to generate the Gel-Bead in Emulsions (GEMs) using the 3' V3 chemistry (10X Genomics). The RT reaction was conducted in the C1000 touch PCR instrument (BioRad). Barcoded cDNA was extracted from the GEMs by Post-GEM RT-cleanup and amplified for 12 cycles. Amplified cDNA was subjected to 0.6x SPRI beads cleanup (Beckman, B23318). 25% of the amplified cDNA was subjected to enzymatic fragmentation, end-repair, A tailing, adapter ligation and 10X specific sample indexing as per manufacturer's protocol. Libraries were quantified using Bioanalyzer (Agilent) analysis. Libraries were pooled and sequenced on an NovaSeq 6000 instrument (Illumina) using the recommended sequencing read lengths of 28 bp (Read 1), 8 bp (i7 Index Read), and 91 bp (Read 2).

scATAC-seq

FACS-purified cells were processed for single nuclei ATAC-seq according to the manufacturer's instructions (10x Genomics, CG000168 Rev D). Briefly, nuclei were obtained by incubating PBMCs for 3.20 minutes in freshly prepared Lysis buffer following manufacturer's instructions for Low Cell Input Nuclei Isolation (10x Genomics, CG000169 Rev C). Nuclei were washed and resuspended in chilled diluted nuclei buffer (10x Genomics, 2000153). Next, nuclei were subjected to transposition for 1h at 37C on the C1000 touch PCR instrument (BioRad) prior to single nucleus capture on the 10x Chromium instrument. Samples were subjected to post GEM cleanup, sample index PCR, cleanup, and library QC prior to sequencing according to the protocol. Samples were pooled, quantified, and sequenced

on NovaSeq 6000 instrument (Illumina) with at least minimum recommended read depth (25000 read pairs/nucleus).

Detection of IFNa and IFNg in plasma and cell culture supernatants

Frozen plasma or supernatant was thawed at room temperature and analyzed using the IFNa and IFNg Human ProQuantum Immunoassay Kits according to manufacturer's instructions. In brief, samples were mixed with assay dilution buffer at a 1:5 or 1:2 ratio and protein standard was serially diluted in assay dilution buffer. Next, Antibody-conjugates A and B were mixed with Antibody-conjugate dilution buffer and added to each well of a 96-well qPCR plate (Bio-Rad, #HSP9601). Next, diluted sample or standard were added to each well and mixtures were incubated for 1h at room temperature in the dark. Finally, Master max and Ligase were mixed and added to each well. QPCR was conducted on a CFX96 Touch Real-Time Detection System (Biorad) using the recommended instrument settings. After measurements were completed, CT values were calculated using a regression model and exported to the ProQuantum Cloud app that accompanied the kit (apps.thermofisher.com/apps/proquantum). ProQuantum Cloud app was then used to construct a standard curve and calculate protein concentrations from CT values.

IP-10 plasma Luminex

Plasma biomarker levels were assayed using a 10-analyte multiplex bead array (fractalkine, IL-12P40, IL-13, IL-1RA, IL-1b, IL-6, IP-10, MCP-1, MIP-1α, TNFβ; Millipore) prepared according to the manufacturer's recommended protocol and read using a Bio-Plex 200 suspension array reader (Bio-Rad). Data were analyzed using Bio-Plex manager software (Bio-Rad).

Viral infection assay

Dengue virus (DENV- 2, Strain Thailand/16681/84) and Zika virus (PRVABC59) were propagated and titrated on Vero cells and stored at -80C until infection. Cryopreserved human PBMCs were thawed, washed, counted, and resuspended in RPMI 1640 (Thermo Fisher, 72400-047) supplemented with 10% FBS (Corning, 35-011-CV), 1mM Sodium pyruvate (Lonza, 13-115E), and 1x Penicillin/Streptomycin (Lonza, 17-602E) at $1.5x10^6$ cells/mL. 200 μL of cell solution ($3x10^5$ cells) was added to each well of a 96-well round-bottomed tissue culture plate and cells were rested in plates for 4h at 37C and 5% CO2. After resting, PBMCs were infected with DENV-2 or ZIKV at MOI 1. At 0h, 24h, 48h post infection, PBMCs and supernatant were collected for RNA purification and cytokine analysis, respectively. Supernatants were immediately frozen at -20C and stored until analysis. Cells were suspended in RNA lysis buffer and kept at -20C until analysis. RNA was purified using the Purelink RNA kit according to manufacturer recommendations (Thermo Fisher Scientific, #12183052). For viral load detection, quantitative reverse transcription PCR (qRT-PCR) was conducted using Luna universal probe one-step RT-PCR kit (NEB, #E3006) on a CFX96

C1000 Touch Real-Time Detection System with 96-well plates (Bio-Rad, #HSP9601). RNA standards (ATCC, # VR-3229SD, VR-1843DQ) were used to generate standard curves. Viral RNA copies were normalized by cell number. Utilized primers and probes are listed in the Key Resources table.

Detection of IP-10 in culture supernatant

Culture supernatants were thawed at room temperature and analyzed using the IP-10 enzyme-linked immunosorbent assay (R&D Systems, DIP100) according to the manufacturer's instructions. In brief, samples were thawed at room temperature and mixed with assay dilution buffer at 1:2 ratio. Protein standard was serially diluted in assay dilution buffer. Samples and standards were incubated in plate for 2h at room temperature. Plates were washed and then incubated with human IP-10 conjugate for 2h at room temperature. After wash, substrate solution was added for 30min. Finally, stop solution was added, A450 and A595 were read on a plate reader (Bio-Rad, iMARK). The concentration of IP-10 was determined by the number of A450-A595 based on the standard curve.

Quantification and Statistical Analysis

Immune Cell Population Definitions and EpiTOF Data Pre-Processing

Raw data were pre-processed using FlowJo (FlowJo, LLC) to identify cell events from individual samples by palladium-based mass tags, and to segregate specific immune cell populations by immunophenotypic markers. A detailed gating hierarchy is described in DataS2, related to STAR Methods (TIV & H5N1/H5N1+AS03). Single-cell data for various immune cell subtypes from individual subjects were exported from FlowJo for downstream computational analyses.

EpiTOF analysis

The exported Flowjo data were then normalized following the approach described in [32]. In brief, the value of each histone mark was regressed against the total amount of histones, represented by measured values of H3 and H4. For sample level analyses, the values of each histone mark were averaged for each cell type in each sample. Distances of HSC from lymphoid and myeloid epigenetic profiles were obtained by first computing centers of the epigenetic profiles for the two lineages, and then computing Euclidean distances from the centers for each individual HSC. Distances of HSC from epigenetic profiles of specific cell types were similarly obtained by computing Euclidean distances from the centers of the epigenetic profiles for each cell type. Statistical significance of the differences between groups at the sample level was assessed by computing an effect size with Hedges' g formula [95]. All p-values were corrected for multiple comparisons with the Benjamini-Hochberg method [96]. Dimensionality reduction was performed with applying UMAP [97]. For single cell analyses, the normalized values were used as input. Correlation between variables was computed using

Pearson's correlation coefficient. All the analyses were performed using the R framework for statistical computing (Version 3.6.3) (R Core Team, 2020).

TIV bulk gene expression analysis

Processed data was normalized in Bioconductor by RMA [98], which includes global background adjustment and quantile normalization. Samples from phase1 subjects in the antibiotics and control arm of the study were selected and statistical tests and correlation analyses were performed using MATLAB. Test details and significance cutoffs are reported in figure legends.

Luminex analysis

Statistical analysis was conducted in R (v 4.0.2) (R Core Team, 2020). First, MFI data was log2 transformed and average MFI and CV was calculated from duplicate cultures where available. For samples with CV > 0.25, the duplicate that was closer to the average of all samples of that subject was kept and the other discarded. In case no other sample was available and CV > 0.25, the sample was discarded. Wells without indication of cytokine production were excluded. Statistical tests, correlation analysis, and hierarchical clustering were performed using the R packages stats (v 4.0.2), ggpubr (v 0.4.0) and pheatmap (v 1.0.12). Test details and statistical cutoffs are reported in the figure legends.

Bulk ATAC-seq pre-processing

Reads were aligned using the BWA algorithm (mem mode; default settings; v 0.7.12) [99]. Duplicate reads were removed, only reads mapping as matched pairs and only uniquely mapped reads (mapping quality >= 1) were used for further analysis. Alignments were extended in silico at their 3'-ends to a length of 200 bp and assigned to 32-nt bins along the genome. The resulting histograms (genomic "signal maps") were stored in bigWig files. Peaks were identified using the MACS algorithm (v 2.1.0) [100] at a cutoff of p-value 1e-7, without control file, and with the –nomodel option. Peaks that were on the ENCODE blacklist of known false ChIP-Seq peaks were removed. Signal maps and peak locations were used as input data to Active Motif's proprietary analysis program. For differential analysis, reads were counted in all merged peak regions (using Subread), and the replicates for each condition were compared using DESeq2 (v 1.24.0) [101].

Bulk ATAC-seq analysis

Quality control analysis of ATAC-seq data was performed using Rockefeller University workshop on analysis of ATAC-seq data in R and Bioconductor (`https://rockefelleruniversity.github.io/RU_ATAC_Workshop.html`). Of 57 unique samples processed, 51 passed QC criteria and, on average, we detected more than 42,000 genomic regions and more than 15x106 unique ATAC tags per sample while the average fraction of reads in peaks was larger than

35% (DataS3, related to Figure 3). Passed samples showed the characteristic fragment length and TSS enrichment distribution (DataS3, related to Figure 3). DARs were annotated as promoter, distal and trans regulatory peak for a particular gene based on the distance from the middle of the peak to the nearest transcription start site (TSS) using the ChIPpeakAnno package in R (v.3.24.1). Promoter, distal and trans regulatory peaks were defined as -2000 bp to +500 bp, -10kbp to +10kbp – promoter, and $< -10kbp$ or $> +10kbp$ from TSS, respectively. The hypergeometric distribution-based enrichment analysis was performed to identify the significance of the DARs. Reactome pathways and TF-target relationship using Chip-seq data from ENCODE (both downloaded from `https://maayanlab.cloud/chea3/`) were used to identify overrepresented pathways and TFs. EnrichmentMap Pipeline Collection (v1.1.0) [102] for CytoScape (v3.8.2) [103] was used to create the pathway network. Significantly enriched Reactome pathways (p <= 0.05) for each genomic region were used as input. Pathways were clustered and annotated using the AutoAnnotate function within the pipeline. To test for enrichment of TF motifs in DARs, the chromVAR (v1.8.0) [104] and motifmatchr (v1.8.0) packages were used in R (v3.6.0) (R Core Team, 2020). In brief, TF motifs were downloaded from the JASPAR2016 core homo sapiens database [105] and merged regions were annotated for the presence of all TF binding motifs using the matchMotifs (motifmatchr) function with standard settings. Hypergeometric distribution-based enrichment analysis was then performed to identify enrichment of TF motifs in DARs. To determine the relationship between EpiTOF and ATAC-seq data, the Pearson correlation was computed between EpiTOF H3K27ac levels and normalized read counts in each merged peak region. Positively correlated merged peak regions with p-value <= 0.05 were selected for functional annotation. Enrichment analysis was performed as described above.

Bulk RNA-seq of purified immune cells

Alignment was performed using STAR version 2.7.3a [106] and transcripts were annotated using GRCh38 Ensembl release 100. Transcript abundance estimates were calculated internal to the STAR aligner using the algorithm of htseq-count [107]. DESeq2 version 1.26.0 [101] was used for differential expression analysis using the Wald test with a paired design formula and using its standard library size normalization.

Analysis of bulk transcriptomics data from previous TIV studies

Processed bulk transcriptomics data from nine independent TIV studies conducted between 2007 and 2012 were obtained from GEO (accessions: GSE47353, GSE59635, GSE29619, GSE74813, GSE59654, GSE59743, GSE74811, GSE29617, GSE74816) [54, 55, 56, 57, 58, 31]. After removing samples and genes with missing values as well as extraordinary vaccine time points, we selected only samples from subjects matching the same age range as the current study: 18 – 45 years of age. The remaining samples were batch corrected using ComBat from the sva package in R (v 3.36.0) with study as batch, no covariates, and otherwise standard

settings. Statistical tests were performed using the R base and ComplexHeatmap (v 2.4.3) packages. Test details and statistical cutoffs are reported in the figure legends.

scATAC analysis

The CellRanger-atac pipeline (v1.1.0) by 10X Genomics was used for alignment (GRCh38 reference genome), de-duplication, and identification of cut sites for each sample. The samples were then combined using the CellRanger-atac aggregation procedure without depth normalization (–normalize=none). The resulting fragment file was read into SnapATAC [108]. SnapATAC was used to bin the genome (bin size of 5K) and create a cell-by-bin count matrix. Cells were identified as barcodes with at latest 1000 UMIs, and a promotor ratio (defined as: (fragments in promoter regions + 1) / (total fragments + 1)) of at least 0.1. Bins that mapped to chrY, mitochondrial DNA, or bins that overlap with ENCODE blacklist regions [109], were removed. The remaining bins were used for dimensionality reduction using Truncated SVD with the irlba R package [110], and the first 50 dimensions were then used for clustering. Macs2 [100] was then used to call peaks within each cluster using recommended parameters for ATACseq data (–nomodel –shift 100 –ext 200 –qval 5e-2 -B –SPMR). The cluster-specific peaks were merged to a single combined set. SnapATAC was then used to map the fragments to the combined peaks set and create a peak-by-cell binary matrix. In the H5N1/H5N1+AS03 dataset, deeply-sequenced libraries were downsampled to an average of 1500 fragments per barcode by randomly removing counts from these samples at a probability p=1500/(mean fragments per cell in the sample). The dimensionality reduction and clustering procedure described above was then repeated on the peak-by-cell matrix. ChromVAR [104] was used with default parameters and the JASPAR2016 [105] motif database to calculate motif accessibility scores and compute differentially accessible motifs in the data. Hotspot was used to identify informative gene modules that explain heterogeneity within the monocyte population [63], using the Bernoulli model and the top 2500 regions (ranked by highest autocorrelation z-score) for module calculation. Modules were then identified using the create_modules function, with $min_gene_threshold = 200$. Similar modules were manually identified and merged by taking the average score across modules. Differentially accessible regions were identified using logistic regression with the glm function in R with the design: y $\sim$ timepoint + donor + log_fragments to control for donor and library size effects. The coefficient corresponding to the time point was then used as the logFC value, and a Wald test was used to get p-values. For numerical stability, we only included peaks that were detected in at least 5% of the cells included in each comparison. All custom scripts for preprocessing, correlation analysis, and differential accessibility analysis are posted in zenodo [doi: 10.5281/zenodo.4446316]. The hypergeometric distribution-based enrichment analysis was performed to identify the significance of the DARs (p <= 0.05 and detected in at least 5% of cells). Reactome pathways database (both downloaded from `https://maayanlab.cloud/chea3/`) were used to identify overrepresented pathways. Enrichr [111] was used to conduct enrichment analysis of genomic regions within Hotspot modules 2, 3. Enrichr was also used to conduct enrichment analysis of DARs containing an IRF1 motif. Briefly, significant DARs

($p <= 0.05$ and detected in at least 5% of cells) carrying an IRF1 motif, as determined by chromVAR, were selected. Next, gene names with multiple associated DARs were collapsed in case all DARs changed in the same direction or otherwise discared. Subsequently, gene list was submitted to Enrichr for enrichment using the Reactome_2016 database. Similarly, we used Enrichr together with the ChEA_2016 databases to identify TF target genes enriched in genes that were enhanced after booster vaccination with H5N1+AS03 and that overlapped with changes in accessibility at promoter regions.

scRNA analysis

The CellRanger pipeline (v3.1.0) by 10X Genomics was used for alignment (GRCh38 reference genome), demultiplexing, cell-calling, and filtering. The filtered count matrices from each sample were then aggregated using the CellRanger aggregation procedure without depth normalization (–normalize=none). The resulting count matrix was analyzed with scVI (scvi-tools v0.7.1) [112] with default hyperparameters to fit a low-dimensional latent space, using the experiment annotation for each sample as a batch label for batch correction. Visualization, clustering, and exploratory analyses were performed with VISION (v2.1.0) [63]. Differential expression analysis between time points was performed with edgeR [113] as described in the package documentation, using the exactTest hypothesis testing for each pairwise analysis.

Bulk transcriptomics vax010

Initial data quality was assessed by background level, 3' labeling bias, and pairwise correlation among samples via the arrayQualityMetrics package in Bioconductor [114]. CEL files were normalized via RMA [98], which includes global background adjustment and quantile normalization. Probes mapping to multiple genes were discarded, and the remaining probes were collapsed to gene level by selecting the probe for each gene with the highest mean expression across all subjects. Statistical tests were performed in MATLAB and R.

2.6 Figures

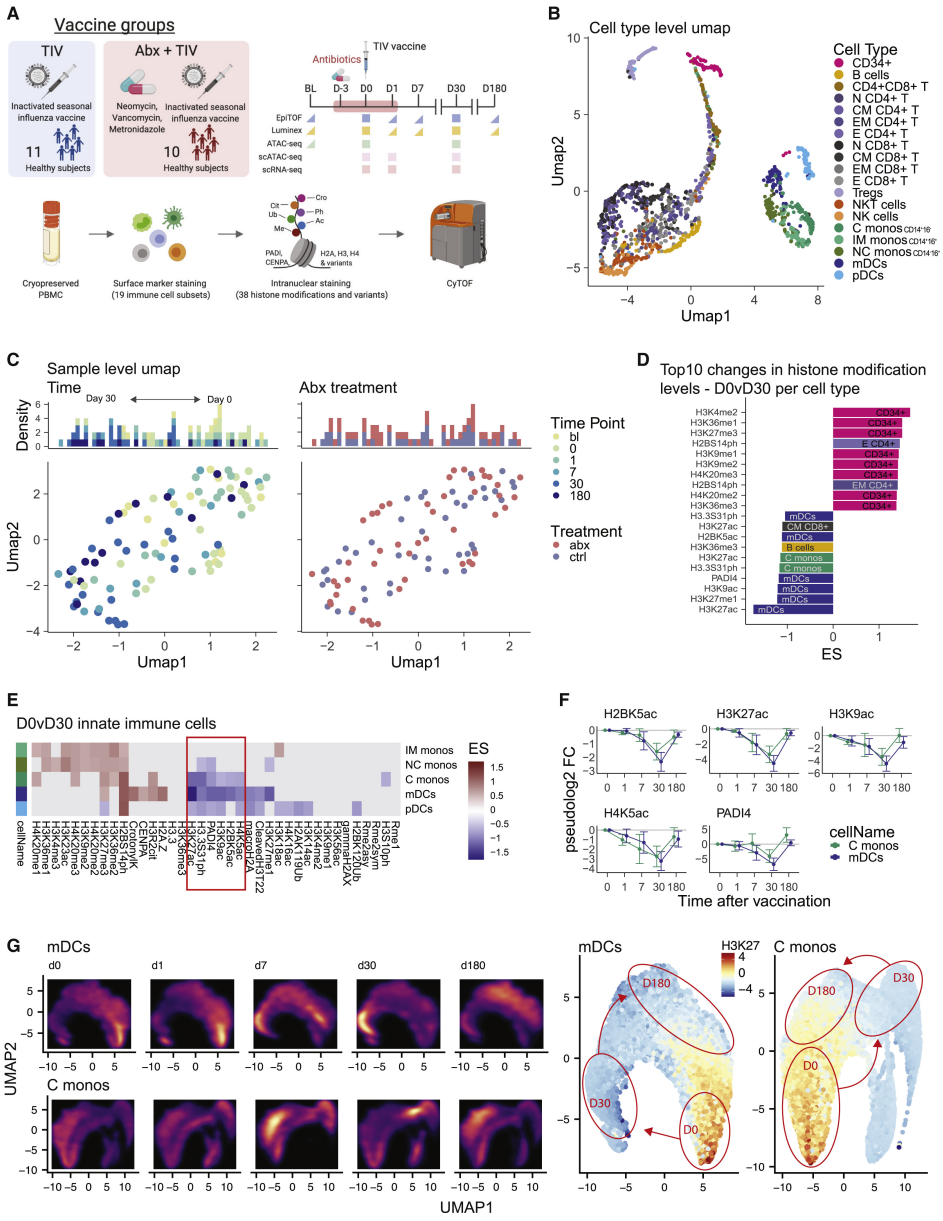


Figure 1

Figure 1: **TIV alters the global histone modification profile of immune cells. (A)** Study overview. **(B)** UMAP was used to create a dimensionality-reduced representation of the global histone mark profiles of all immune cell subset. **(C)** UMAP was used to visualize epigenomic profiles at the sample level. **(D, E)** The effect size of vaccine-induced changes to the global histone modification profiles at day 30 vs day 0 were calculated. D) Top-10 most significantly increased and reduced histone modifications. E) Heatmap showing histone modification changes in innate immune cells. Only changes with an FDR <= 20% are shown. **(F)** Change in histone modification levels relative to day 0 before vaccination for a set of highly reduced histone modifications in C monos and mDCs. Dots and lines indicate average modification levels, error bars indicate the standard error of mean. **(G)** UMAP representation of single monocytes and mDCs using H2BK5ac, H3K37ac, H3K9ac, H4K5ac, and PADI4. Left panel: cell density at each time point, right panel: H3K27ac levels in each single cell. Red ellipses indicate high-density areas corresponding to bright areas in left panel. See also Figure S1, S2, S3

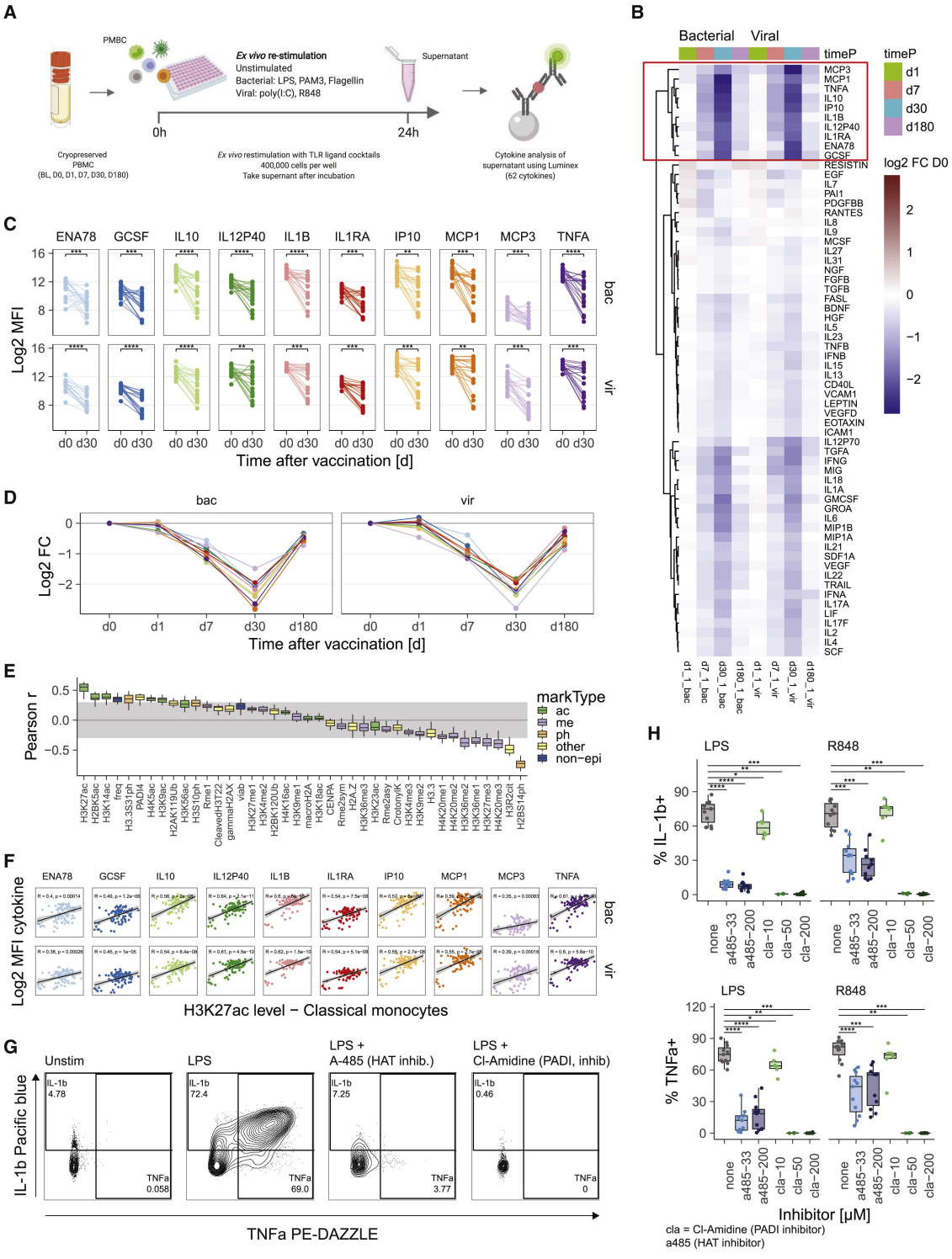

A
PMBC
Ex vivo re-stimulation
Unstimulated
Bacterial: LPS, PAM3, Flagellin
Viral: poly(I:C), R848
Supernatant
0h
24h
Cryopreserved PBMC (BL, D0, D1, D7, D30, D180)
Ex vivo restimulation with TLR ligand cocktails 400,000 cells per well Take supernant after incubation
Cytokine analysis of supernatant using Luminex (62 cytokines)
B
Bacterial
Viral
timeP
d1
d7
d30
d180
log2 FC D0
MCP3
MCP1
TNFA
IL10
IP10
IL1B
IL12P40
IL1RA
ENA78
GCSF
RESISTIN
EGF
IL7
PAI1
PDGFBB
RANTES
IL8
IL9
MCSF
IL27
IL31
NGF
FGFB
TGFB
FASL
BDNF
HGF
IL5
IL23
TNFB
IFNB
IL15
IL13
CD40L
VCAM1
LEPTIN
VEGFD
EOTAXIN
ICAM1
IL12P70
TGFA
IFNG
MIG
IL18
IL1A
GMCSF
GROA
IL6
MIP1B
MIP1A
IL21
SDF1A
VEGF
IL22
TRAIL
IFNA
IL17A
LIF
IL17F
IL2
IL4
SCF
d1_1_bac
d7_1_bac
d30_1_bac
d180_1_bac
d1_1_vir
d7_1_vir
d30_1_vir
d180_1_vir
C
ENA78
GCSF
IL10
IL12P40
IL1B
IL1RA
IP10
MCP1
MCP3
TNFA
Log2 MFI
bac
vir
d0 d30
Time after vaccination [d]
D
bac
vir
Log2 FC
d0
d1
d7
d30
d180
Time after vaccination [d]
E
Pearson r
markType
ac
me
ph
other
non−epi
H3K27ac
H2BK5ac
H3K14ac
freq
H3.3S31ph
PADI4
H4K5ac
H3K9ac
H2AK119Ub
H3K56ac
H3S10ph
Rme1
CleavedH3T22
gammaH2AX
Vab
H3K27me1
H3K4me2
H2BK120Ub
H4K16ac
H3K9me1
macroH2A
H3K18ac
CENPA
Rme2sym
H2A.Z
H3K36me3
H3K23ac
Rme2asy
CrotonylK
H3K4me3
H3K9me2
H3.3
H4K20me1
H4K20me2
H3K36me2
H3K79me1
H3K27me3
H4K20me3
H3R2cit
H2BS14ph
F
ENA78
GCSF
IL10
IL12P40
IL1B
IL1RA
IP10
MCP1
MCP3
TNFA
Log2 MFI cytokine
bac
vir
H3K27ac level − Classical monocytes
G
Unstim
LPS
LPS + A-485 (HAT inhib.)
LPS + Cl-Amidine (PADI, inhib)
IL-1b Pacific blue
IL-1b 4.78
TNFa 0.058
IL-1b 72.4
TNFa 69.0
IL-1b 7.25
TNFa 3.77
IL-1b 0.46
TNFa 0
TNFa PE-DAZZLE
H
LPS
R848
% IL−1b+
% TNFa+
none
a485−33
a485−200
cla−10
cla−50
cla−200
Inhibitor [μM]
cla = Cl-Amidine (PADI inhibitor)
a485 (HAT inhibitor)

Figure 2

Figure 2: **TIV-induced histone modification changes correlate with cytokine production. (A)** Schematic overview of experiment. **(B)** Heatmap showing the relative change in cytokine levels at indicated time points compared to day 0. **(C)** Cytokine levels before (d0) and after (d30) vaccination for each investigated subject. Wilcoxon signed rank test was used for hypothesis testing. * $p <= 0.05$, ** $p <= 0.01$, *** $p <= 0.001$, **** $p <= 0.0001$, $n = 18 - 19$ **(D)** Change in cytokine levels relative to day 0 for cytokines in C. Dots and lines indicate average. **(E, F)** Pearson correlation of the cytokine levels of the 10 cytokines in **C)** with histone modification levels in C monos as well as C mono frequency in PBMCs as determined by EpiTOF and sample viability. (n = 87 samples from all time points) **E)** Boxplots of correlation coefficients for each cytokine after stimulation with either viral or bacterial cocktail. **F)** Scatter plots for the indicated histone modifications and cytokines. **(G)** Gating scheme showing the production of IL-1b and TNFa in C monos after indicated treatment. **(H)** Boxplot summary of the fraction of IL-1b+ or TNFa+ cells in multiple donors. Wilcoxon rank sum test, * $p <= 0.05$, ** $p <= 0.01$, *** $p <= 0.001$, **** $p <= 0.0001$, $n = 4 - 11$. See also Figure S4.

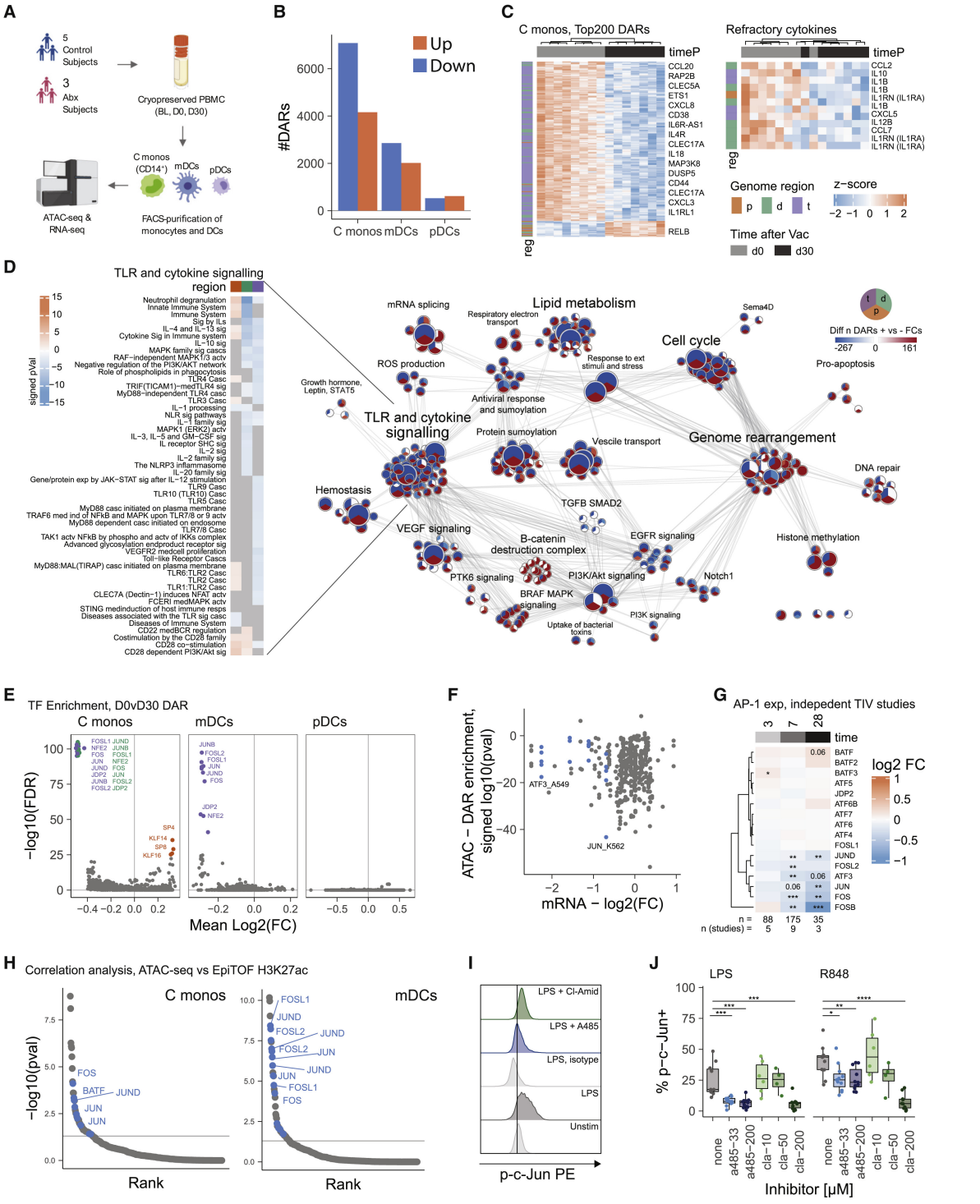

A
5 Control Subjects
3 Abx Subjects
Cryopreserved PBMC (BL, D0, D30)
C monos (CD14+)
mDCs
pDCs
ATAC-seq & RNA-seq
FACS-purification of monocytes and DCs
B
Up
Down
#DARs
C monos
mDCs
pDCs
C
C monos, Top200 DARs
timeP
Refractory cytokines
Genome region
z−score
Time after Vac
d0
d30
D
TLR and cytokine signalling region
signed pVal
TLR and cytokine signalling
Lipid metabolism
Cell cycle
Genome rearrangement
Hemostasis
VEGF signaling
B-catenin destruction complex
TGFB SMAD2
EGFR signaling
PI3K/Akt signaling
BRAF MAPK signaling
Notch1
Histone methylation
DNA repair
Pro-apoptosis
Diff n DARs + vs - FCs
E
TF Enrichment, D0vD30 DAR
C monos
mDCs
pDCs
−log10(FDR)
Mean Log2(FC)
F
ATAC − DAR enrichment, signed log10(pval)
mRNA − log2(FC)
ATF3_A549
JUN_K562
G
AP-1 exp, indepedent TIV studies
time
log2 FC
n =
n (studies) =
H
Correlation analysis, ATAC-seq vs EpiTOF H3K27ac
C monos
mDCs
−log10(pval)
Rank
I
LPS + CI-Amid
LPS + A485
LPS, isotype
LPS
Unstim
p-c-Jun PE
J
LPS
R848
% p−c−Jun+
Inhibitor [μM]
none
a485−33
a485−200
cla−10
cla−50
cla−200

Figure 3

Figure 3: **TIV induces reduced chromatin accessibility in immune response genes and AP-1 controlled regions.** **(A)** Schematic overview of the experiment. **(B)** Differentially accessible chromatin regions (DARs) at day 30 vs day 0 were identified using DESeq2. Pval <= 0.05. **(C)** Heatmap representation of the normalized accessibility at the top 200 as well as cytokine-associated DARs in C monos for each analyzed sample. p: promoter -2000 bp to +500 bp; d: distal -10kbp to +10kbp – promoter; t: trans $< -10kbp$ or $> +10kbp$. **(D)** Network representation of gene set enrichment analysis of DARs in C monos using the Reactome database. Only significantly enriched terms (p <= 0.05) are shown. Color indicates whether majority of enriched regions showed enhanced (red) or reduced (blue) accessibility. Heatmaps show signed –log10(pval) for significantly enriched terms in highlighted clusters. **(E)** Motif-based overrepresentation analysis of transcription factor binding sites in DARs at day 30 vs day 0. **(F)** Scatter plot showing the change in TF gene expression (x-axis) plotted against the enrichment in DARs for selected transcription factors in the Encode database. Blue color indicates AP-1 members with significantly reduced expression. **(G)** Change in gene expression of AP-1 family members using bulk transcriptomics data from 3-9 independent flu vaccine trials previously conducted. Heatmap indicates average log2 fold change in gene expression over all trials. N indicates subject and study number at each time point. Wilcoxon signed rank test, * $p <= 0.05$, ** $p <= 0.01$, *** $p <= 0.001$. **(H)** DARs in indicated cell type were correlated with H3K27ac levels as measured by EpiTOF and DARs with significant correlation ($p <= .05$) were analyzed for transcription factor target gene enrichment using the Encode database. Blue color indicates significantly changed AP-1 members. **(I)** Histogram showing the level of phospho-c-Jun in C monos in the indicated conditions. **(J)** Box plot summary of the fraction of phospho-c-Jun positive cells in classical monocytes. Wilcoxon rank sum test, * $p <= 0.05$, ** $p <= 0.01$, *** $p <= 0.001$, **** $p <= 0.0001$, $n = 4 - 11$ See also Figure S5

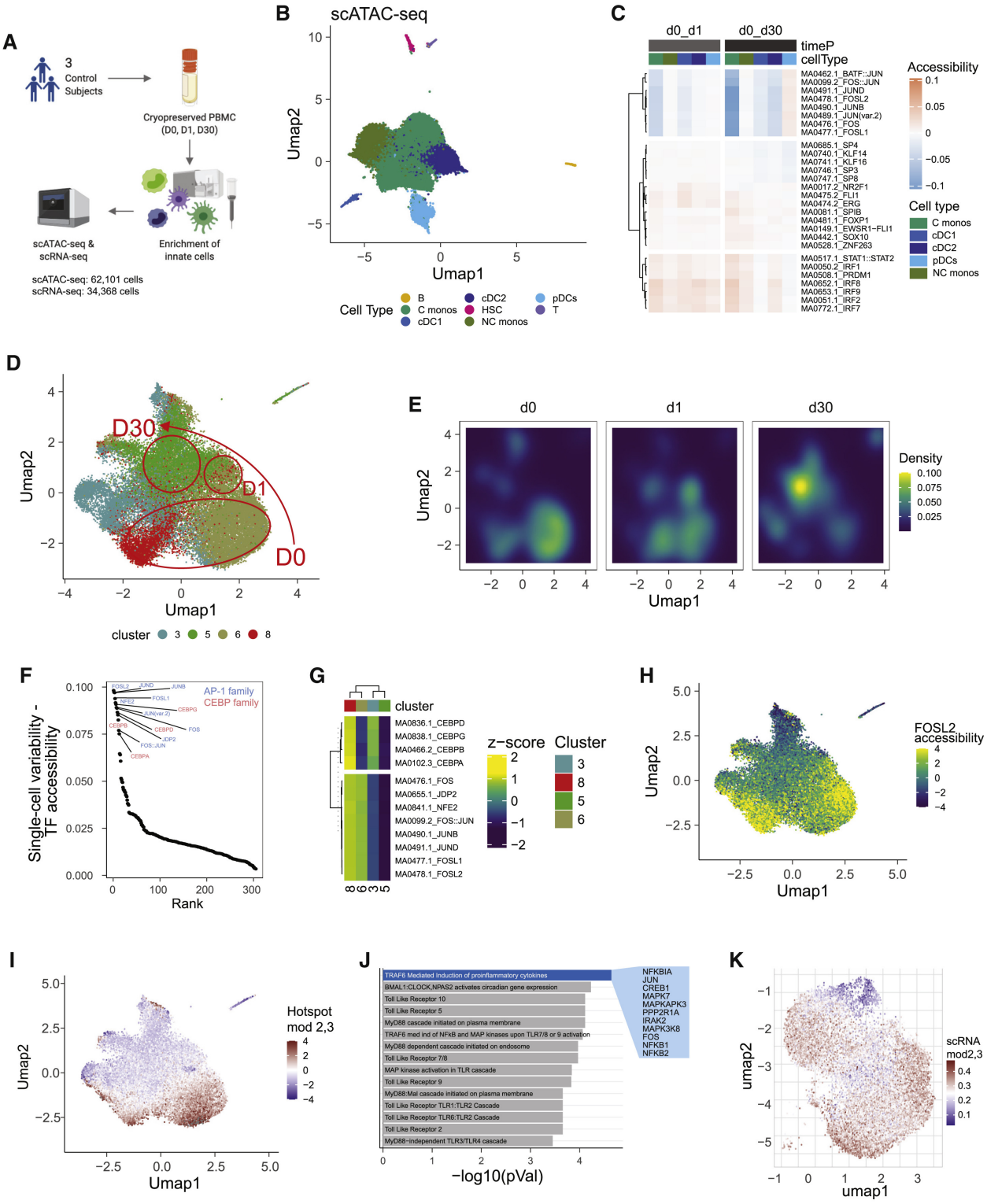

A
3
Control
Subjects
Cryopreserved PBMC
(D0, D1, D30)
scATAC-seq &
scRNA-seq
Enrichment of
innate cells
scATAC-seq: 62,101 cells
scRNA-seq: 34,368 cells
B
scATAC-seq
Umap2
Umap1
Cell Type
B
C monos
cDC1
cDC2
HSC
NC monos
pDCs
T
C
d0_d1
d0_d30
timeP
cellType
Accessibility
Cell type
C monos
cDC1
cDC2
pDCs
NC monos
D
D30
D1
D0
Umap2
Umap1
cluster
E
d0
d1
d30
Density
F
AP-1 family
CEBP family
Single-cell variability -
TF accessibility
Rank
G
cluster
z−score
Cluster
H
FOSL2
accessibility
I
Hotspot
mod 2,3
J
−log10(pVal)
K
scRNA
mod2,3
umap1
umap2

Figure 4

Figure 4: **Heterogeneity within monocyte population drives TIV induced epigenomic changes. (A)** Schematic overview of the experiment. **(B)** UMAP representation of scATAC-seq landscape after pre-processing and QC filtering. **(C)** Heatmap showing differences in chromatin accessibility at indicated time points for the top5 transcription factors per subset. **(D)** UMAP representation of epigenomic subclusters within the classical monocyte population. **(E)** Density plot showing the relative contribution of different epigenomic subclusters to the total monocyte population at a given vaccine time point. **(F)** Variability in TF accessibility within the monocyte population. Value indicates range of accessibility values in all single monocytes. **(G)** Heatmap showing differences in chromatin accessibility between monocyte subclusters subset. **(H)** UMAP representation of monocyte subclusters showing differences in AP-1 accessibility. **(I)** UMAP representation of monocyte subclusters showing difference in accessibility at Hotspot module 2,3 gene loci. **(J)** Enrichment analysis of genes associated with loci in Hotspot module 2,3. **(K)** UMAP representation of the transcriptional landscape of single monocytes. Color indicates expression of genes associated with Hotspot modules 2,3.

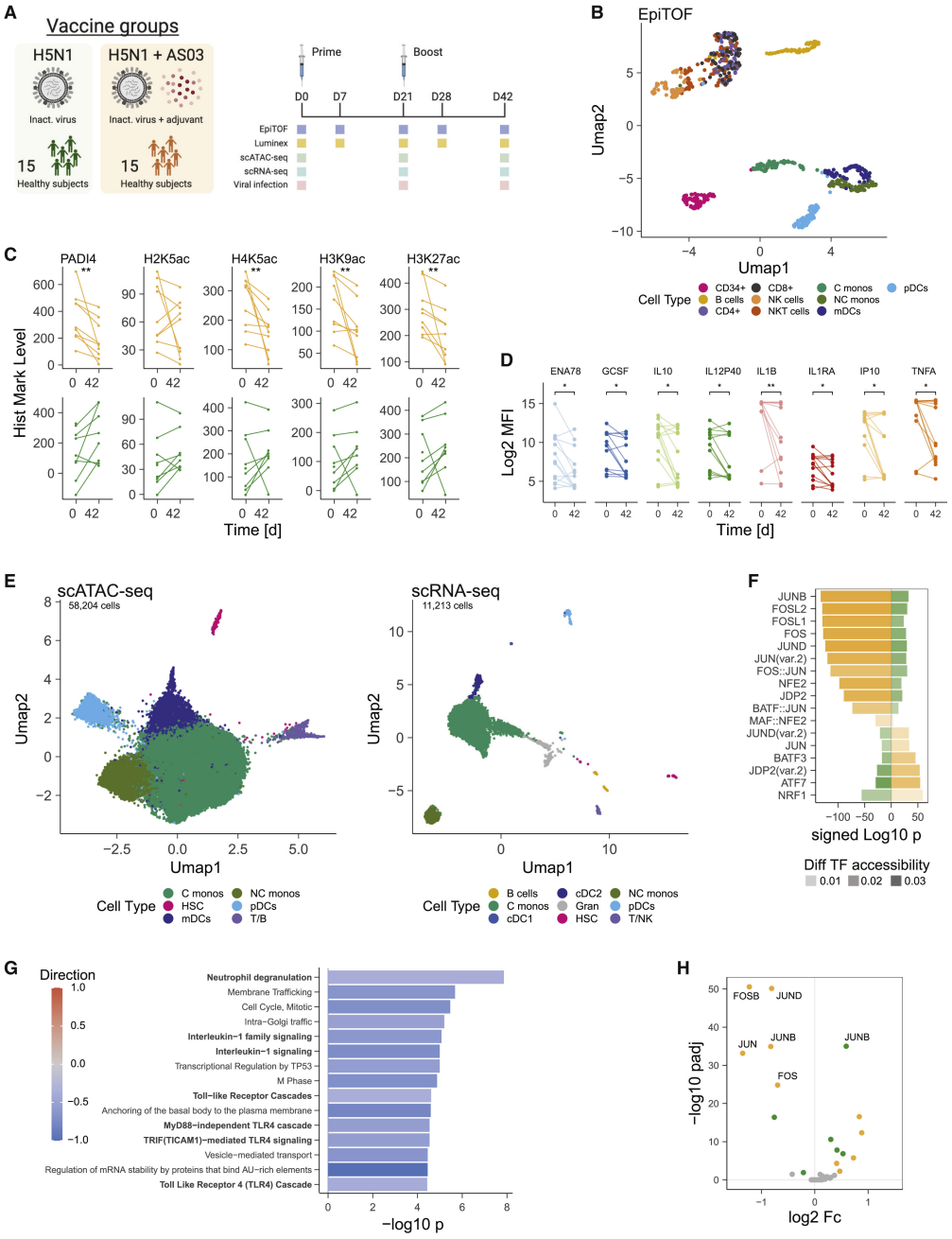

A
Vaccine groups
H5N1
Inact. virus
15
Healthy subjects
H5N1 + AS03
Inact. virus + adjuvant
15
Healthy subjects
Prime
Boost
D0
D7
D21
D28
D42
EpiTOF
Luminex
scATAC-seq
scRNA-seq
Viral infection
B
EpiTOF
Umap2
Umap1
Cell Type
CD34+
B cells
CD4+
CD8+
NK cells
NKT cells
C monos
NC monos
mDCs
pDCs
C
PADI4
H2K5ac
H4K5ac
H3K9ac
H3K27ac
Hist Mark Level
Time [d]
D
ENA78
GCSF
IL10
IL12P40
IL1B
IL1RA
IP10
TNFA
Log2 MFI
Time [d]
E
scATAC-seq
58,204 cells
Umap2
Umap1
Cell Type
C monos
HSC
mDCs
NC monos
pDCs
T/B
scRNA-seq
11,213 cells
Umap2
Umap1
Cell Type
B cells
C monos
cDC1
cDC2
Gran
HSC
NC monos
pDCs
T/NK
F
JUNB
FOSL2
FOSL1
FOS
JUND
JUN(var.2)
FOS::JUN
NFE2
JDP2
BATF::JUN
MAF::NFE2
JUND(var.2)
JUN
BATF3
JDP2(var.2)
ATF7
NRF1
signed Log10 p
Diff TF accessibility
0.01
0.02
0.03
G
Direction
Neutrophil degranulation
Membrane Trafficking
Cell Cycle, Mitotic
Intra−Golgi traffic
Interleukin−1 family signaling
Interleukin−1 signaling
Transcriptional Regulation by TP53
M Phase
Toll−like Receptor Cascades
Anchoring of the basal body to the plasma membrane
MyD88−independent TLR4 cascade
TRIF(TICAM1)−mediated TLR4 signaling
Vesicle−mediated transport
Regulation of mRNA stability by proteins that bind AU−rich elements
Toll Like Receptor 4 (TLR4) Cascade
−log10 p
H
FOSB
JUND
JUN
JUNB
JUNB
FOS
−log10 padj
log2 Fc

Figure 5

Figure 5: **H5N1+AS03 induces repressive epigenomic state akin to TIV. (A)** Schematic overview of experiment. **(B)** UMAP representation of EpiTOF landscape. **(C)** Histone modification levels in classical monocytes at day 0 vs day 42 as measured by EpiTOF. **(D)** Cytokine levels in supernatant of TLR-stimulated PBMCs at day 0 and day 42 after vaccination with H5N1+AS03. (C, D) Wilcox signed rank test; * p <= 0.05, ** p <= 0.01; EpiTOF: n = 9/9, Luminex: n = 13. **(E)** UMAP representation of scATAC-seq (left) and scRNA-seq (right) landscape after pre-processing and QC filtering. **(F)** Change in accessibility of detected AP-1 family TFs in classical monocytes. Color indicates whether cells are derived from subjects vaccinated with H5N1 (green) or H5N1+AS03 (orange). (n = 2/2) **(G)** Overrepresentation analysis of significantly different DARs in classical monocytes using the Reactome database. Color indicates whether enriched genes were predominantly up- or down-regulated. **(H)** Volcano plot showing changes in expression of AP-1 TF genes in classical monocytes at D42 compared to D0. See also Figure S6

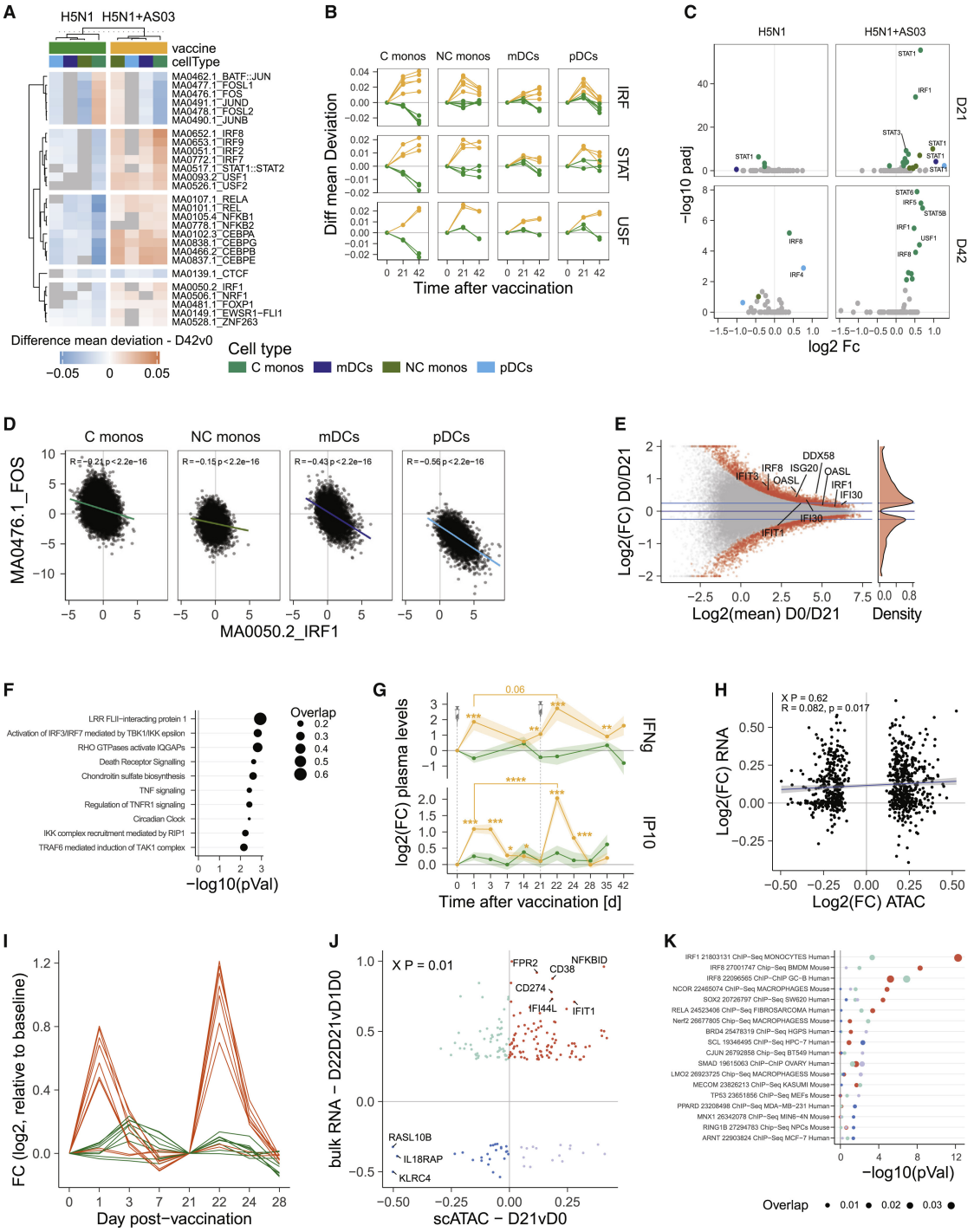

A
H5N1 H5N1+AS03
vaccine
cellType
MA0462.1_BATF::JUN
MA0477.1_FOSL1
MA0476.1_FOS
MA0491.1_JUND
MA0478.1_FOSL2
MA0490.1_JUNB
MA0652.1_IRF8
MA0653.1_IRF9
MA0051.1_IRF2
MA0772.1_IRF7
MA0517.1_STAT1::STAT2
MA0093.2_USF1
MA0526.1_USF2
MA0107.1_RELA
MA0101.1_REL
MA0105.4_NFKB1
MA0778.1_NFKB2
MA0102.3_CEBPA
MA0838.1_CEBPG
MA0466.2_CEBPB
MA0837.1_CEBPE
MA0139.1_CTCF
MA0050.2_IRF1
MA0506.1_NRF1
MA0481.1_FOXP1
MA0149.1_EWSR1−FLI1
MA0528.1_ZNF263
Difference mean deviation - D42v0
−0.05 0 0.05
Cell type
C monos mDCs NC monos pDCs
B
C monos NC monos mDCs pDCs
Diff mean Deviation
IRF
STAT
USF
Time after vaccination
C
H5N1 H5N1+AS03
D21
D42
−log10 padj
log2 Fc
D
C monos NC monos mDCs pDCs
MA0476.1_FOS
MA0050.2_IRF1
E
Log2(FC) D0/D21
Log2(mean) D0/D21
Density
DDX58
IRF8
ISG20
OASL
IFIT3
IRF1
IFI30
IFIT1
F
Overlap
LRR FLII−interacting protein 1
Activation of IRF3/IRF7 mediated by TBK1/IKK epsilon
RHO GTPases activate IQGAPs
Death Receptor Signalling
Chondroitin sulfate biosynthesis
TNF signaling
Regulation of TNFR1 signaling
Circadian Clock
IKK complex recruitment mediated by RIP1
TRAF6 mediated induction of TAK1 complex
−log10(pVal)
G
log2(FC) plasma levels
IFNg
IP10
Time after vaccination [d]
H
X P = 0.62
R = 0.082, p = 0.017
Log2(FC) RNA
Log2(FC) ATAC
I
FC (log2, relative to baseline)
Day post−vaccination
J
X P = 0.01
bulk RNA − D22D21vD1D0
scATAC − D21vD0
FPR2
CD38
NFKBID
CD274
IFI44L
IFIT1
RASL10B
IL18RAP
KLRC4
K
−log10(pVal)
Overlap

Figure 6

Figure 6: **H5N1+AS03 induces epigenomic state of enhanced antiviral immunity. (A)** Heatmap showing the change in chromatin accessibility at day 42 vs day 0 for the top5 transcription factors per subset. Color indicates the difference in accessibility, grey fields indicate non-significant changes (fdr > 0.05). **(B)** Line graph showing the difference in transcription factor (TF) accessibility during vaccination. Each line represents a separate TF within the indicated family. **(C)** Volcano plot showing the change in gene expression for IRF/STAT TF genes. **(D)** Scatter plot showing chromatin accessibility values for IRF1 (x-axis) and FOS (y-axis) in single cells. Indicated statistics are based on Pearson correlation. **(E)** MA plot showing the average accessibility and log2(FC) accessibility for genomic regions containing an IRF1 binding motif. Red color indicates regions with significantly changed accessibility (P <= 0.05). **(F)** Gene set enrichment analysis of significantly changed regions in **E)** occurring in at least 5% of C monos using the Reactome database. **(G)** Interferon gamma and IP10 levels in plasma of vaccinated subjects. Dots and lines indicate average, ribbons indicate standard error of mean. (H5N1/H5N1+AS03: IFNG, n = 7/14; IP10, n = 16/34) **(H)** Scatter plot showing changes in chromatin accessibility (x-axis) and changes in gene expression (y-axis) at day 21 vs day 0 for C monos (scATAC P <= 0.05 and occurring in at least 5% of cells). Indicated statistics are based on Pearson correlation analysis and Chi-square test. **(I)** Change in gene expression for selected antiviral and interferon-related BTMs in bulk RNA-seq analysis for subjects vaccinated with H1N1 (green) and H1N1+AS03 (orange) at indicated time points. (H1N1: n = 16, H1N1+AS03: n = 34). **(J)** Scatter plot showing the change in chromatin accessibility at day 21 vs day 0 in C monos (x-axis) and the significant change (p <= 0.05, log2(FC) > ±0.03) in vaccine-induced gene expression at the booster vaccination compared to the prime vaccination (y-axis, Day22day21 vs Day1day0). Chi-square test was used to determine whether both variables were related. **(K)** Bubble plot showing enrichment results using the Encode TF target gene database. Color indicates the origin of the analyzed genes in **J)**. See also Figure S7

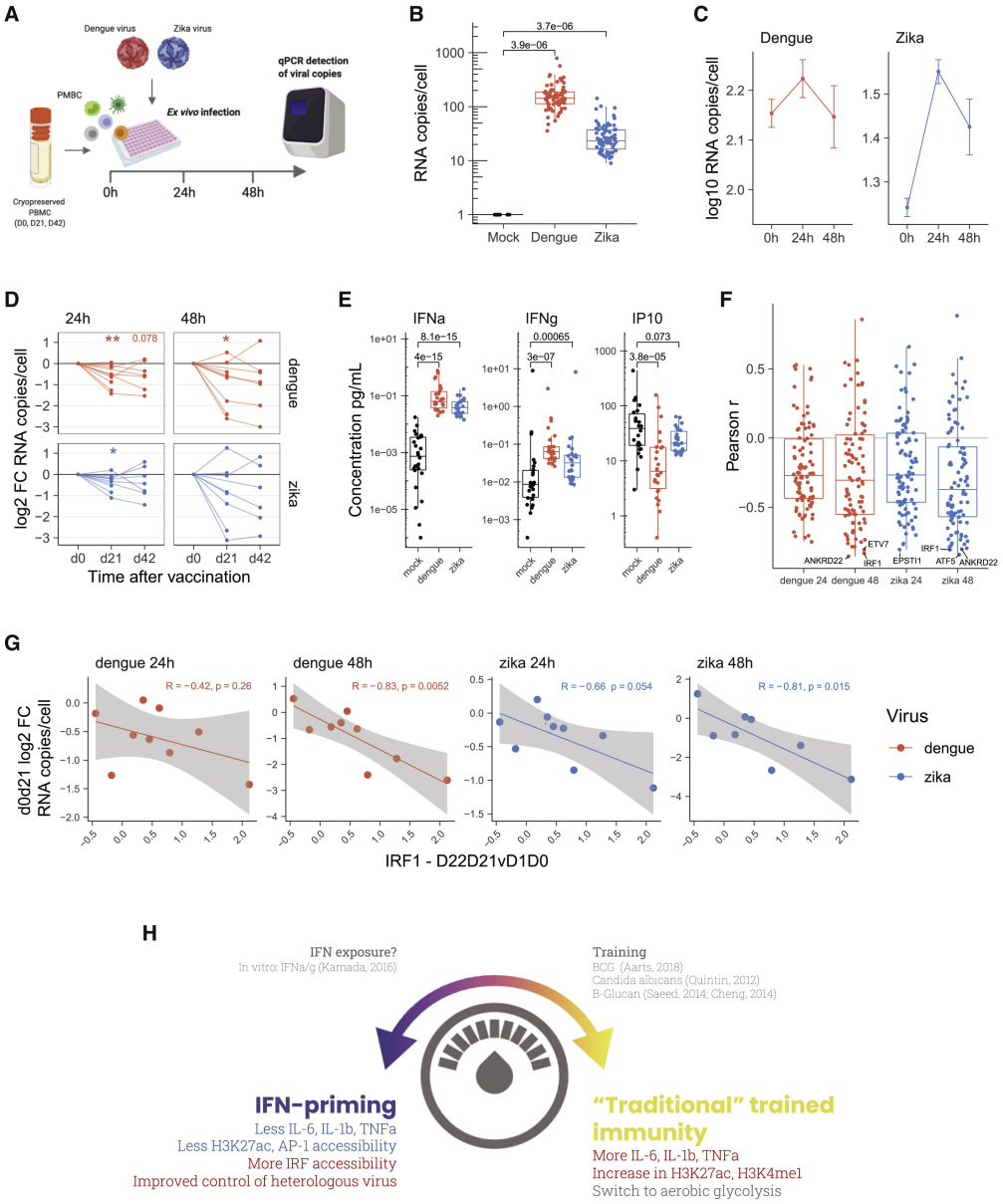


Figure 7: **H1N1+AS03 induces enhanced resistance to in-vitro infection with heterologous viruses.** **(A)** Schematic overview of the experiment. **(B)** Boxplot showing viral titers in Dengue-, Zika-, and mock-infected samples. **(C)** Line graph showing the viral growth curve for Dengue virus (red) and Zika virus (blue). Dots and lines indicate average, error bars indicate standard error of mean. $n > 21$ samples **(D)** Log2 fold change in viral titers relative to day 0 before vaccination. Wilcoxon signed rank test was used to compare changes within group; ** $p <= 0.01$, * $p <= 0.05$, $n = 8 - 9$. **(E)** Boxplot showing the concentration of IFNa, IFNg, and IP10 in Dengue-, Zika-, and mock-infected cultures at 24h after incubation. Wilcoxon rank-sum test was used to compare groups. **(F, G)** Pearson correlation analysis of the change in viral titers (d0 vs d21) with change in vaccine-induced, in-vivo expression of enhanced antiviral genes at prime (d0 vs d1) and boost (d21 vs d22) (red genes Figure 6G). **F)** Boxplot showing correlation coefficient per viral condition. **G)** Scatter plot showing change in vaccine-induced expression of IRF1 (x-axis) and viral titers (y-axis). **(H)** Model of bi-directional epigenomic reprogramming. **B, E)** Wilcoxon rank sum test was used to compare groups.

2.7 References

[1] C.D. Allis and T. Jenuwein. "The molecular hallmarks of epigenetic control". en. In: *Nat. Rev. Genet* 17 (2016), pp. 487–500.

[2] J.D. Buenrostro et al. "Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation". en. In: *Cell* 173 (2018), pp. 1535–1548 16.

[3] M Ryan Corces et al. "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution". en. In: *Nat. Genet.* 48.10 (Oct. 2016), pp. 1193–1203. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3646. URL: http://dx.doi.org/10.1038/ng.3646.

[4] M. Farlik et al. "DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation". it. In: *Cell Stem Cell* 19 (2016), pp. 808–822.

[5] R.S. Akondy et al. "Origin and differentiation of human memory CD8 T cells after vaccination". en. In: *Nature* 552 (2017), pp. 362–367.

[6] Ansuman T Satpathy et al. "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion". en. In: *Nat. Biotechnol.* 37.8 (Aug. 2019), pp. 925–936. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0206-z. URL: http://dx.doi.org/10.1038/s41587-019-0206-z.

[7] B. Youngblood et al. "Effector CD8 T cells dedifferentiate into long-lived memory cells". en. In: *Nature* 552 (2017), pp. 404–409.

[8] B.G. Barwick et al. "Plasma cell differentiation is coupled to division-dependent DNA hypomethylation and gene regulation". de. In: *Nat. Immunol* 17 (2016), pp. 1216–1225.

[9] M. Kulis et al. "Whole-genome fingerprint of the DNA methylome during human B cell differentiation". en. In: *Nat. Genet* 47 (2015), pp. 746–756.

[10] R.J.W. Arts et al. "BCG Vaccination Protects against Experimental Viral Infection in Humans through the Induction of Cytokines Associated with Trained Immunity". en. In: *Cell Host Microbe* 23 (2018), pp. 89–100 5.

[11] J. Kleinnijenhuis et al. "Ifrim". it. In: *Proc. Natl. Acad. Sci* 109 (2012). Ed. by Saeed D.C. et al., pp. 17537–17542.

[12] S. Saeed et al. "Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity". fr. In: *Science* 345 (2014), pp. 1251086–1251086.

[13] J.C. Sun et al. "NK Cells and Immune "Memory". fr. In: *J. Immunol* 186 (2011), pp. 1891–1897.

[14] M.G. Netea et al. "Defining trained immunity and its role in health and disease". en. In: *Nat. Rev. Immunol* 20 (2020), pp. 375–388.

[15] F. Wimmers and B. Pulendran. "Emerging technologies for systems vaccinology — multi-omics integration and single-cell (epi)genomic profiling". en. In: *Curr. Opin. Immunol* 65 (2020), pp. 57–64.

[16] M. Alcántara-Hernández et al. "High-Dimensional Phenotypic Mapping of Human Dendritic Cells Reveals Interindividual Variation and Tissue Specialization". en. In: *Immunity* (2017).

[17] P.S. Arunachalam et al. "Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans". fr. In: *Science* (2020).

[18] S.W. Kazer et al. "Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection". en. In: *Nat. Med* (2020).

[19] J. Schulte-Schrepping et al. en. Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. Cell S0092867420309922. 2020.

[20] P. See et al. "Mapping the human DC lineage through the integration of high-dimensional techniques". fr. In: *Science* 356 (2017), p. 3009.

[21] A.K. Shalek et al. "Single-cell RNA-seq reveals dynamic paracrine control of cellular variation". en. In: *Nature* 510 (2014), pp. 363–369.

[22] A.-C. Villani et al. "Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors". en. In: *Science* 356 (2017), p. 4573.

[23] F. Wimmers et al. "Single-cell analysis reveals that stochasticity and paracrine signaling control interferon-alpha production by plasmacytoid dendritic cells". en. In: *Nat. Commun* 9 (2018).

[24] D. Gaucher et al. "Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses". en. In: *J. Exp. Med* 205 (2008), pp. 3119–3131.

[25] T. Hagan et al. "Systems vaccinology: Enabling rational vaccine design with systems biological approaches". en. In: *Vaccine* 33 (2015), pp. 5294–5301.

[26] Y. Kotliarov et al. "Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus". en. In: *Nat. Med* 26 (2020), pp. 618–629.

[27] S. Li et al. "Metabolic Phenotypes of Response to Vaccination in Humans". en. In: *Cell* 169 (2017), pp. 862–877 17.

[28] B. Pulendran, S. Li, and H.I. Nakaya. "Systems Vaccinology". cs. In: *Immunity* 33 (2010), pp. 516–529.

[29] T.D. Querec et al. "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans". en. In: *Nat. Immunol* 10 (2009), pp. 116–125.

[30] H.-C.S.P. Team and H.-I. Consortium. "Multicohort analysis reveals baseline transcriptional predictors of influenza vaccination responses". fr. In: *Sci. Immunol* 2 (2017).

[31] J.S. Tsang et al. "Global Analyses of Human Immune Variation Reveal Baseline Predictors of Postvaccination Responses". en. In: *Cell* 157 (2014), pp. 499–513.

[32] P. Cheung et al. "Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging". en. In: *Cell* (2018).

[33] T. Hagan et al. "Antibiotics-Driven Gut Microbiome Perturbation Alters Immunity to Vaccines in Humans". it. In: *Cell* 178 (2019), pp. 1313–1328 13.

[34] W.L. Cheung et al. "Apoptotic Phosphorylation of Histone H2B Is Mediated by Mammalian Sterile Twenty Kinase". en. In: *Cell* 113 (2003), pp. 507–517.

[35] S. Solier and Y. Pommier. "The apoptotic ring: a novel entity with phosphorylated histones H2AX and H2B and activated DNA damage response kinases". en. In: *Cell Cycle Georget. Tex* 8 (2009), pp. 1853–1859.

[36] W. Wen et al. "MST1 Promotes Apoptosis through Phosphorylation of Histone H2AX". en. In: *J. Biol. Chem* 285 (2010), pp. 39108–39116.

[37] K.-M. Cho et al. "Mst1-Deficiency Induces Hyperactivation of Monocyte-Derived Dendritic Cells via Akt1/c-myc Pathway". en. In: *Front. Immunol* 10 (2019).

[38] C. Li et al. "Dendritic cell MST1 inhibits Th17 differentiation". en. In: *Nat. Commun* 8 (2017), p. 14275.

[39] W. Li et al. "STK4 regulates TLR pathways and protects against chronic inflammation–related hepatocellular carcinoma". en. In: *J. Clin. Invest* 125 (2015), pp. 4239–4254.

[40] X. Zhou et al. "YAP Aggravates Inflammatory Bowel Disease by Regulating M1/M2 Macrophage Polarization and Gut Microbial Homeostasis". en. In: *Cell Rep* 27 (2019), pp. 1176–1189 5.

[41] Y. Liu et al. "Peptidylarginine deiminases 2 and 4 modulate innate and adaptive immune responses in TLR-7–dependent lupus". en. In: *JCI Insight* 3 (2018).

[42] K. Nakashima et al. "Molecular Characterization of Peptidylarginine Deiminase in HL-60 Cells Induced by Retinoic Acid and 1a,25-Dihydroxyvitamin D3". pt. In: *J. Biol. Chem* 274 (1999), pp. 27786–27792.

[43] E.R. Vossenaar et al. "Expression and activity of citrullinating peptidylarginine deiminase enzymes in monocytes and macrophages". en. In: *Ann. Rheum. Dis* 63 (2004), pp. 373–381.

[44] B.T. Weinert et al. "Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome". en. In: *Cell* 174 (2018), pp. 231–244 12.

[45] A.K. Voss et al. "Moz and Retinoic Acid Coordinately Regulate H3K9 Acetylation, Hox Gene Expression, and Segment Identity". en. In: *Dev. Cell* 17 (2009), pp. 674–686.

[46] "BCG Vaccination in Humans Elicits Trained Immunity via the Hematopoietic Progenitor Compartment". it. In: *Cell Host Microbe* 28 (2020), pp. 322–334 5.

[47] E. Kaufmann et al. "BCG Educates Hematopoietic Stem Cells to Generate Protective Innate Immunity against Tuberculosis". it. In: *Cell* 172 (2018), pp. 176–190 19.

[48] I. Mitroulis et al. "Modulation of Myelopoiesis Progenitors Is an Integral Component of Trained Immunity". it. In: *Cell* 172 (2018), pp. 147–161 12.

[49] W.E. O'Gorman et al. "Single-cell systems-level analysis of human Toll-like receptor activation defines a chemokine signature in patients with systemic lupus erythematosus". en. In: *J. Allergy Clin. Immunol* 136 (2015), pp. 1326–1336.

[50] R. Cao and Y. Zhang. "The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3". en. In: *Curr. Opin. Genet. Dev* 14 (2004), pp. 155–164.

[51] J.D. Stender et al. "Control of Proinflammatory Gene Programs by Regulated Trimethylation and Demethylation of Histone H4K20". en. In: *Mol. Cell* 48 (2012), pp. 28–38.

[52] L.M. Lasko et al. "Discovery of a selective catalytic p300/CBP inhibitor that targets lineage-specific tumours". en. In: *Nature* 550 (2017), pp. 128–132.

[53] Y.-H. Lee et al. "Regulation of coactivator complex assembly and function by protein arginine methylation and demethylimination". en. In: *Proc. Natl. Acad. Sci. U. S. A* 102 (2005), pp. 3611–3616.

[54] T. Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". en. In: *Nucleic Acids Res* 41 (2013), pp. 991–995.

[55] S. Mohanty et al. "Prolonged Proinflammatory Cytokine Production in Monocytes Modulated by Interleukin 10 After Influenza Vaccination in Older Adults". en. In: *J. Infect. Dis* 211 (2015), pp. 1174–1184.

[56] H.I. Nakaya et al. "Systems biology of vaccination for seasonal influenza in humans". fr. In: *Nat. Immunol* 12 (2011), pp. 786–795.

[57] H.I. Nakaya et al. "Systems Analysis of Immunity to Influenza Vaccination across Multiple Years and in Diverse Populations Reveals Shared Molecular Signatures". en. In: *Immunity* 43 (2015), pp. 1186–1198.

[58] J. Thakar et al. "Aging-dependent alterations in gene expression and a mitochondrial signature of responsiveness to human influenza vaccination". en. In: *Aging* 7 (2015), pp. 38–52.

[59] J. Arias et al. "Activation of cAMP and mitogen responsive genes relies on a common nuclear factor". en. In: *Nature* 370 (1994), pp. 226–229.

[60] Y. Kamei et al. "A CBP Integrator Complex Mediates Transcriptional Activation and AP-1 Inhibition by Nuclear Receptors". en. In: *Cell* 85 (1996), pp. 403–414.

[61] K. Zanger, S. Radovick, and F.E. Wondisford. "CREB Binding Protein Recruitment to the Transcription Complex Requires Growth Factor–Dependent Phosphorylation of Its GF Box". en. In: *Mol. Cell* 7 (2001), pp. 551–558.

[62] M. Guilliams, A. Mildner, and S. Yona. "Developmental and Functional Heterogeneity of Monocytes". en. In: *Immunity* 49 (2018), pp. 595–613.

[63] D. DeTomaso and N. Yosef. *Identifying Informative Gene Modules Across Modalities of Single Cell Genomics*. it. BioRxiv, 2020.

[64] B. Pulendran et al. "Emerging concepts in the science of vaccine adjuvants". en. In: *Nat. Rev. Drug Discov* (2021), pp. 1–22.

[65] N. Garçon, D.W. Vaughn, and A.M. Didierlaurent. "Development and evaluation of AS03, an Adjuvant System containing a-tocopherol and squalene in an oil-in-water emulsion". fr. In: *Expert Rev. Vaccines* 11 (2012), pp. 349–366.

[66] S. Khurana et al. "AS03-adjuvanted H5N1 vaccine promotes antibody diversity and affinity maturation, NAI titers, cross-clade H5N1 neutralization, but not H1N1 cross-subtype neutralization". en. In: *Npj Vaccines* 3 (2018), pp. 1–12.

[67] P.S. Arunachalam et al. "Adjuvanting a subunit COVID-19 vaccine to induce protective immunity". en. In: *Nature* (2021).

[68] P.A. Goepfert et al. "Safety and immunogenicity of SARS-CoV-2 recombinant protein vaccine formulations in healthy adults: interim results of a randomised, placebo-controlled, phase 1–2, dose-ranging study". en. In: *Lancet Infect. Dis* 0 (2021).

[69] B.J. Ward et al. "Phase 1 trial of a Candidate Recombinant Virus-Like Particle Vaccine for Covid-19 Disease Produced in Plants". en. In: *MedRxiv* (2020).

[70] D. Langlais, L.B. Barreiro, and P. Gros. "The macrophage IRF8/IRF1 regulome is required for protection against infections and is associated with chronic inflammationIRF8 and IRF1 regulate macrophage gene expression". en. In: *J. Exp. Med* 213 (2016), pp. 585–603.

[71] T. Tamura et al. "The IRF Family Transcription Factors in Immunity and Oncogenesis". en. In: *Annu. Rev. Immunol* 26 (2008), pp. 535–584.

[72] Z. Kou et al. "Monocytes, but not T or B cells, are the principal target cells for dengue virus (DV) infection among human peripheral blood mononuclear cells". es. In: *J. Med. Virol* 80 (2008), pp. 134–146.

[73] D. Michlmayr et al. "CD14+CD16+ monocytes are the main target of Zika virus infection in peripheral blood mononuclear cells in a paediatric study in Nicaragua". en. In: *Nat. Microbiol* 2 (2017), pp. 1462–1470.

[74] A. Soares-Schanoski et al. "Systems analysis of subjects acutely infected with the Chikungunya virus". en. In: *PLOS Pathog* 15 (2019), p. 1007880.

[75] D.H. Phanstiel et al. "Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development". en. In: *Mol. Cell* 67 (2017), pp. 1037–1048 6.

[76] T.van der Bruggen et al. "Lipopolysaccharide-Induced Tumor Necrosis Factor Alpha Production by Human Monocytes Involves the Raf-1/MEK1-MEK2/ERK1-ERK2 Pathway". en. In: *Infect. Immun* 67 (1999), pp. 3824–3829.

[77] M. Das et al. "Induction of hepatitis by JNK-mediated expression of TNFa". en. In: *Cell* 136 (2009), pp. 249–260.

[78] M.F. Fontana et al. "JUNB Is a Key Transcriptional Modulator of Macrophage Activation". en. In: *J. Immunol* 194 (2015), pp. 177–186.

[79] S. Fujioka et al. "NF-kB and AP-1 Connection: Mechanism of NF-kB-Dependent Regulation of AP-1 Activity". cs. In: *Mol Cell Biol* 24 (2004), p. 14.

[80] N. Hannemann et al. "The AP-1 Transcription Factor c-Jun Promotes Arthritis by Regulating Cyclooxygenase-2 and Arginase-1 Expression in Macrophages". en. In: *J. Immunol* 198 (2017), pp. 3605–3614.

[81] J.-J. Ventura et al. "c-Jun NH2-Terminal Kinase Is Essential for the Regulation of AP-1 by Tumor Necrosis Factor". en. In: *Mol. Cell. Biol* 23 (2003), pp. 2871–2882.

[82] M.A. Meraz et al. "Targeted Disruption of the Stat1 Gene in Mice Reveals Unexpected Physiologic Specificity in the JAK–STAT Signaling Pathway". en. In: *Cell* 84 (1996), pp. 431–442.

[83] D. Panda et al. "IRF1 Maintains Optimal Constitutive Expression of Antiviral Genes and Regulates the Early Antiviral Response". en. In: *Front. Immunol* 10 (2019).

[84] G. Behre et al. "c-Jun Is a JNK-independent Coactivator of the PU.1 Transcription Factor". en. In: *J. Biol. Chem* 274 (1999), pp. 4939–4946.

[85] M.M. Monick, A.B. Carter, and G.W. Hunninghake. "Human Alveolar Macrophages Are Markedly Deficient in REF-1 and AP-1 DNA Binding Activity". pt. In: *J. Biol. Chem* 274 (1999), pp. 18075–18080.

[86] E.Y. Tsai et al. "A Lipopolysaccharide-Specific Enhancer Complex Involving Ets, Elk-1, Sp1, and CREB Binding Protein and p300 Is Recruited to the Tumor Necrosis Factor Alpha Promoter In Vivo". en. In: *Mol. Cell. Biol* 20 (2000), pp. 6084–6094.

[87] Beisaw Arica et al. "AP-1 Contributes to Chromatin Accessibility to Promote Sarcomere Disassembly and Cardiomyocyte Protrusion During Zebrafish Heart Regeneration". en. In: *Circ. Res* 126 (2020), pp. 1760–1778.

[88] S.C. Biddie et al. "Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding". en. In: *Mol. Cell* 43 (2011), pp. 145–155.

[89] R. Furth and Z.A. Cohn. "The Origin And Kinetics Of Mononuclear Phagocytes". en. In: *J. Exp. Med* 128 (1968), pp. 415–435.

[90] A.T. Kamath et al. "The Development, Maturation, and Turnover Rate of Mouse Spleen Dendritic Cell Populations". en. In: *J. Immunol* 165 (2000), pp. 6762–6770.

[91] S. Boettcher and M.G. Manz. "Regulation of Inflammation- and Infection-Driven Hematopoiesis". en. In: *Trends Immunol* 38 (2017), pp. 345–357.

[92] W. Burny et al. "Different Adjuvants Induce Common Innate Pathways That Are Associated with Enhanced Adaptive Responses against a Model Antigen in Humans". en. In: *Front. Immunol* 8 (2017).

[93] J.E. McElhaney et al. "AS03-adjuvanted versus non-adjuvanted inactivated trivalent influenza vaccine against seasonal influenza in elderly people: a phase 3 randomised trial". en. In: *Lancet Infect. Dis* 13 (2013), pp. 485–496.

[94] M.R. Corces et al. "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues". en. In: *Nat. Methods* 14 (2017), pp. 959–962.

[95] L.V. Hedges and I. Olkin. en. Statistical Methods for Meta-Analysis (Academic press). 2014.

[96] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". en. In: *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1 (1995), pp. 289–300. ISSN: 1369-7412, 0035-9246. URL: http://www.jstor.org/stable/2346101.

[97] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". en. In: (Feb. 2018). ArXiv180203426 Cs Stat. arXiv: 1802.03426 [stat.ML]. URL: http://arxiv.org/abs/1802.03426.

[98] "Exploration, normalization, and summaries of high density oligonucleotide array probe level data". en. In: *Biostatistics* 4 (2003), pp. 249–264.

[99] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". en. In: *Bioinformatics* 25 (2009), pp. 1754–1760.

[100] Y. Zhang et al. "Model-based Analysis of ChIP-Seq (MACS". en. In: *Genome Biol* 9 (2008), p. 137.

[101] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12 (2014), p. 550. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0550-8. URL: http://dx.doi.org/10.1186/s13059-014-0550-8.

[102] D. Merico et al. "Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation". en. In: *Plos One* 5 (2010), p. 13984.

[103] P. Shannon et al. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks". en. In: *Genome Res* 13 (2003), pp. 2498–2504.

[104] Alicia N Schep et al. "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data". en. In: *Nat. Methods* 14.10 (Oct. 2017), pp. 975–978. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4401. URL: https://doi.org/10.1038/nmeth.4401.

[105] A. Mathelier et al. "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles". en. In: *Nucleic Acids Res* 44 (2016), pp. 110–115.

[106] A. Dobin et al. "STAR: ultrafast universal RNA-seq aligner". en. In: *Bioinformatics* 29 (2013), pp. 15–21.

[107] S. Anders, P.T. Pyl, and W. Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". en. In: *Bioinformatics* 31 (2015), pp. 166–169.

[108] Rongxin Fang et al. "Comprehensive analysis of single cell ATAC-seq data with SnapATAC". en. In: *Nat. Commun.* 12.1 (Feb. 2021), p. 1337. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21583-9. URL: http://dx.doi.org/10.1038/s41467-021-21583-9.

[109] H.M. Amemiya, A. Kundaje, and A.P. Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". en. In: *Sci. Rep* 9 (2019), p. 9354.

[110] J. Baglama, L. Reichel, and B.W. Lewis. *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices.* en. 2019.

[111] Maxim V Kuleshov et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". en. In: *Nucleic Acids Res.* 44.W1 (July 2016), W90–7. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw377. URL: http://dx.doi.org/10.1093/nar/gkw377.

[112] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". en. In: *Nat. Methods* 15.12 (Dec. 2018), pp. 1053–1058. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: http://dx.doi.org/10.1038/s41592-018-0229-2.

[113] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". en. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btp616. URL: http://dx.doi.org/10.1093/bioinformatics/btp616.

[114] A. Kauffmann, R. Gentleman, and W. Huber. "arrayQualityMetrics—a bioconductor package for quality assessment of microarray data". en. In: *Bioinformatics* 25 (2009), pp. 415–416.

2.8 Supplementary Figures

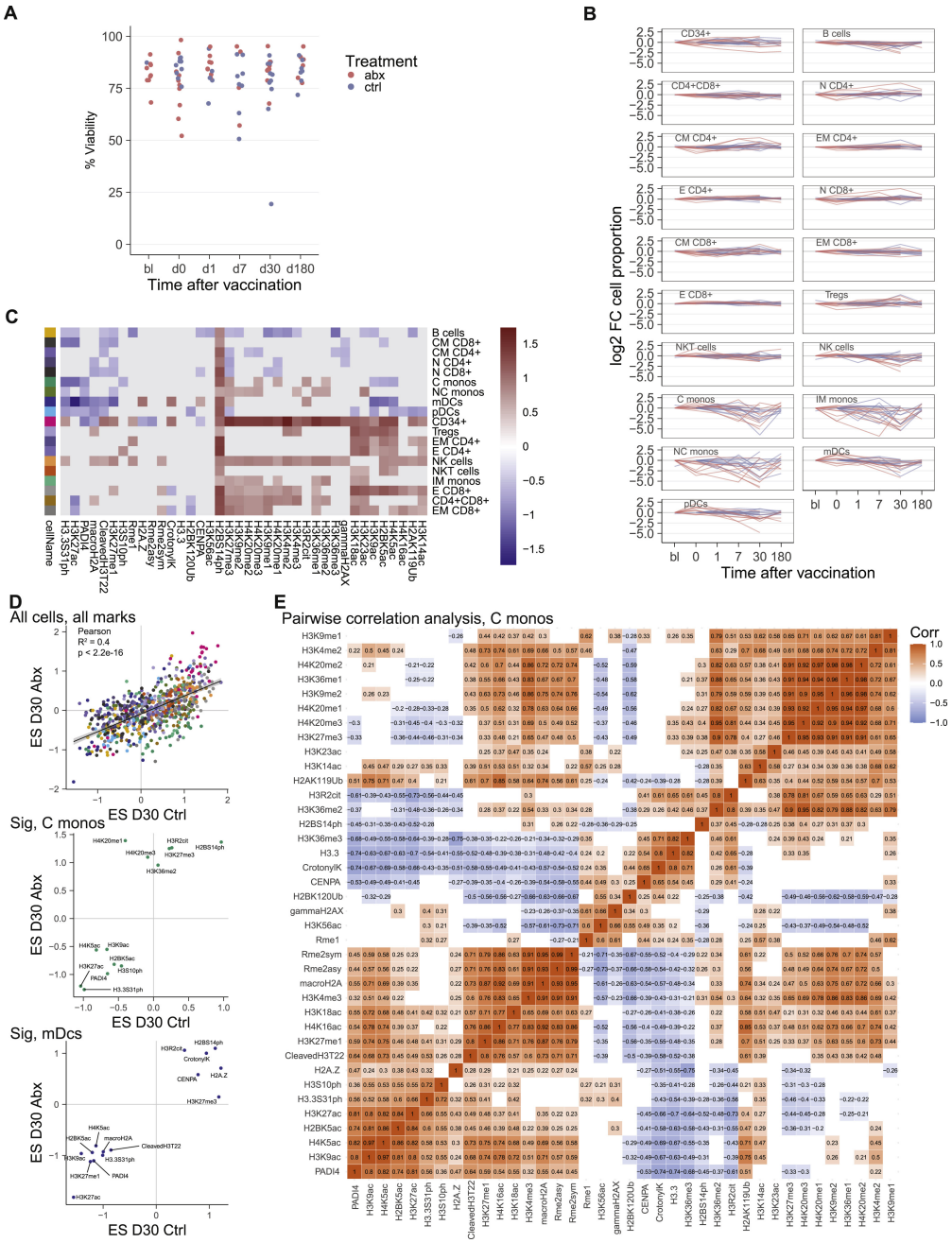


Figure S1: **Cell type abundance and vaccine induced epigenomic changes by EpiTOF, related to Figure 1. (A)** PBMC viability after thawing by vaccination time point. **(B)** Change in cell type abundance per subject. Wilcoxon signed rank test was used to compare changes at post-vaccine time points with d0 and p-values were corrected using the FDR approach. No comparison passed the threshold of $FDR <= 0.05$. **(C)** Heatmap showing histone modification changes at day 30 compared to day 0 in all detected immune cell subsets. Changes were calculated using the effect size approach. Only changes with an $FDR <= 0.2$ are shown. **(D)** Correlation of histone modification changes at day 30 compared to day 0 calculated separately for subjects in the control (x-axis) and antibiotics group (y-axis). For monocytes and mDCs, only significantly changed histone modifications are shown ($FDR <= 0.2$). **(E)** Correlation matrix showing the pair-wise correlation coefficient between all histone modification in classical monocytes. Only significant correlations ($p <= 0.05$) are shown.

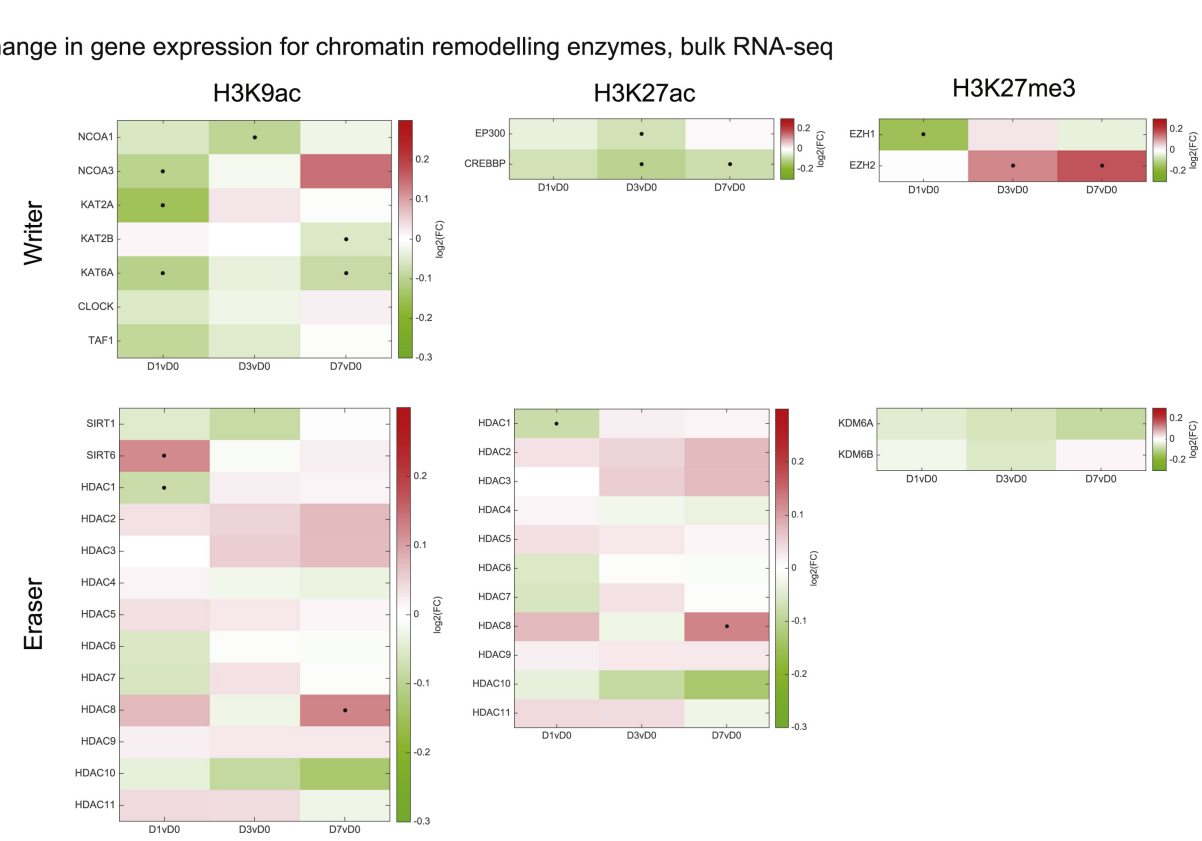


Figure S2: Analysis of vaccine-induced changes in gene expression of histone modifying enzyme by blood transcriptomics, related to Figure 1. (A) Heatmap showing the log2 fold change in gene expression relative to day 0 before vaccination. T-test was used for statistical testing. * $p <= 0.05$

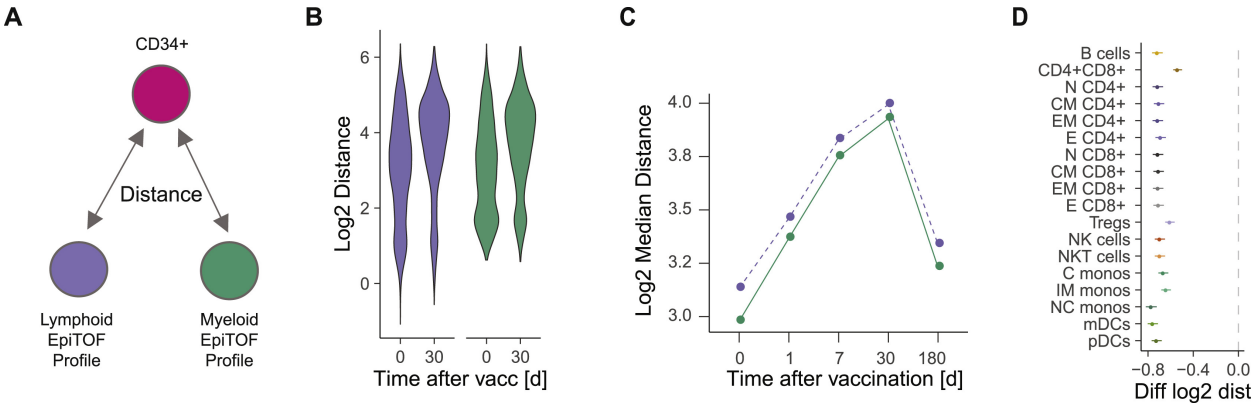


Figure S3: **Histone modification profile distance of CD34+ progenitor cells by EpiTOF, related to Figure 1. (A)** Cartoon of the analysis approach. The Euclidean distance between the histone modification profile of every single CD34+ progenitor cell to an average lymphoid or myeloid profile was calculated. **(B)** Violin plot showing the histone modification profile distance of single CD34+ progenitor cells to a common lymphoid (purple) or myeloid (turquoise) profile at the indicated time point using EpiTOF panel 2. **(C)** Median change in histone modification profile distance over time. **(D)** Change in histone modification profile distance of CD34+ progenitor cells to indicated cell types at day 30 after vaccination compared to day 0.

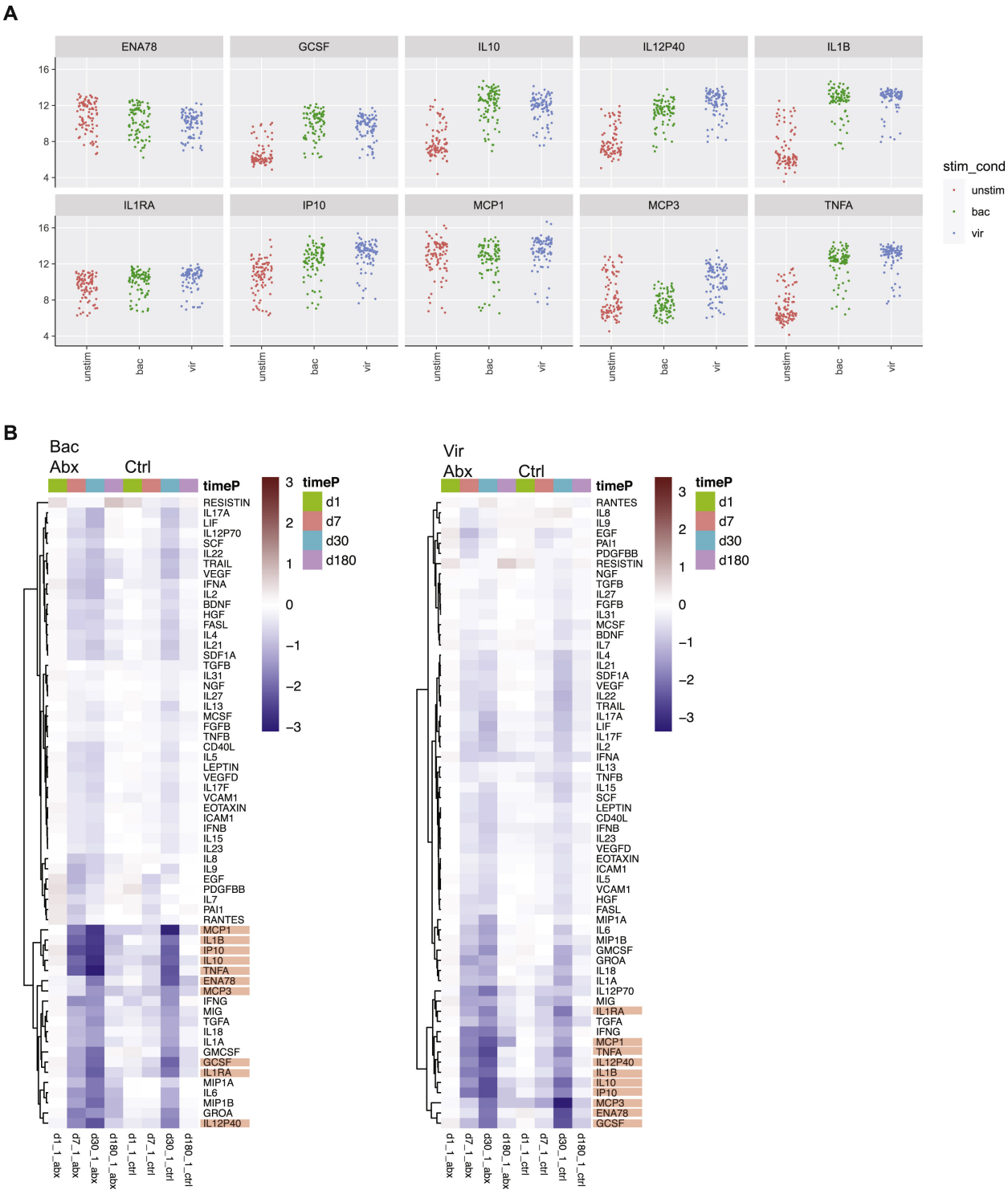
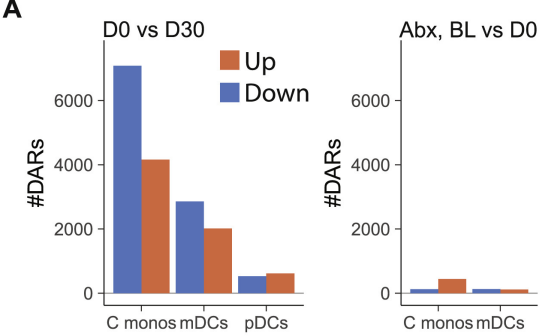


Figure S4: **Cytokine production upon TLR stimulation, related to Figure 2 (A)** Dot plot showing log2 cytokine levels in each TLR-stimulated PBMC culture by stimulation condition. **(B)** Heatmap showing the change in cytokine levels relative to day 0 separately for antibiotics-treated and control subjects.

A



B

PathID	Pathway Description	# Total genes in Pathway	# DAR in Pathway	# Up DAR in Pathway	# Down DAR in Pathway	P-value
M11.0	enriched in monocytes (II)	186	13	6 up	7 down	0.00027
M89.0	putative targets of PAX3	16	4	4 up	0 down	0.00035
M89.1	putative targets of PAX3	10	3	3 up	0 down	0.00115
M113	golgi membrane (I)	10	3	0 up	3 down	0.00115
M4.0	cell cycle and transcription	331	16	10 up	6 down	0.00306
M43.0	myeloid, dendritic cell activation via NFkB (I)	15	3	3 up	0 down	0.00401
M160	leukocyte differentiation	16	3	3 up	0 down	0.00486
M64	enriched in activated dendritic cells/monocytes	17	3	3 up	0 down	0.00580
S11	Activated (LPS) dendritic cell surface signature	37	4	4 up	0 down	0.00876
M186	TBA	9	2	1 up	1 down	0.01585

C

PathID	# Total genes in Pathway	# DAR in Pathway	# Up DAR in Pathway	# Down DAR in Pathway	P-value
RELA_GM19193_HG19	1756	80	76 up	4 down	5.27E-10
MAZ_MELCELLLINE_MM9	1758	77	69 up	8 down	6.61E-09
RELA_GM18526_HG19	717	42	40 up	2 down	1.18E-08
EGR1_ECC1_HG19	2128	87	75 up	12 down	1.54E-08
EGR1_HCT116_HG19	1602	70	63 up	7 down	3.87E-08
CTCF_MELCELLLINE_MM9	4096	137	120 up	17 down	2.06E-07
MYOD1_MYOCYTE_MM9	2759	99	87 up	12 down	7.72E-07
JUN_K562_HG19	3691	123	111 up	12 down	1.31E-06
MYC_K562_HG19	6077	183	161 up	22 down	1.44E-06
RELA_GM12891_HG19	1803	71	64 up	7 down	1.54E-06

D

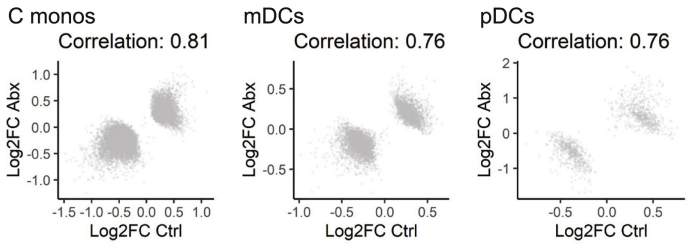


Figure S5: **Antibiotics and vaccine-induced epigenomic changes by bulk ATAC-seq, related to Figure 3.** **(A)** DARs at day 30 compared to day 0 (left) and day 0 vs baseline before antibiotics treatment (right, antibiotics subjects only). **(B)** Overrepresentation analysis of significantly different DARs at day 0 vs baseline in classical monocytes using the BTM database. **(C)** Overrepresentation analysis of significantly different DARs at day 0 vs baseline in classical monocytes using the Encode transcription factor targets database. **(D)** DARs at day 30 compared to day 0 were calculated separately for control and antibiotics subjects. Log2 FC values from peaks that were significantly changed in the combined analysis (Figure 3b) were correlated with each using Pearson.

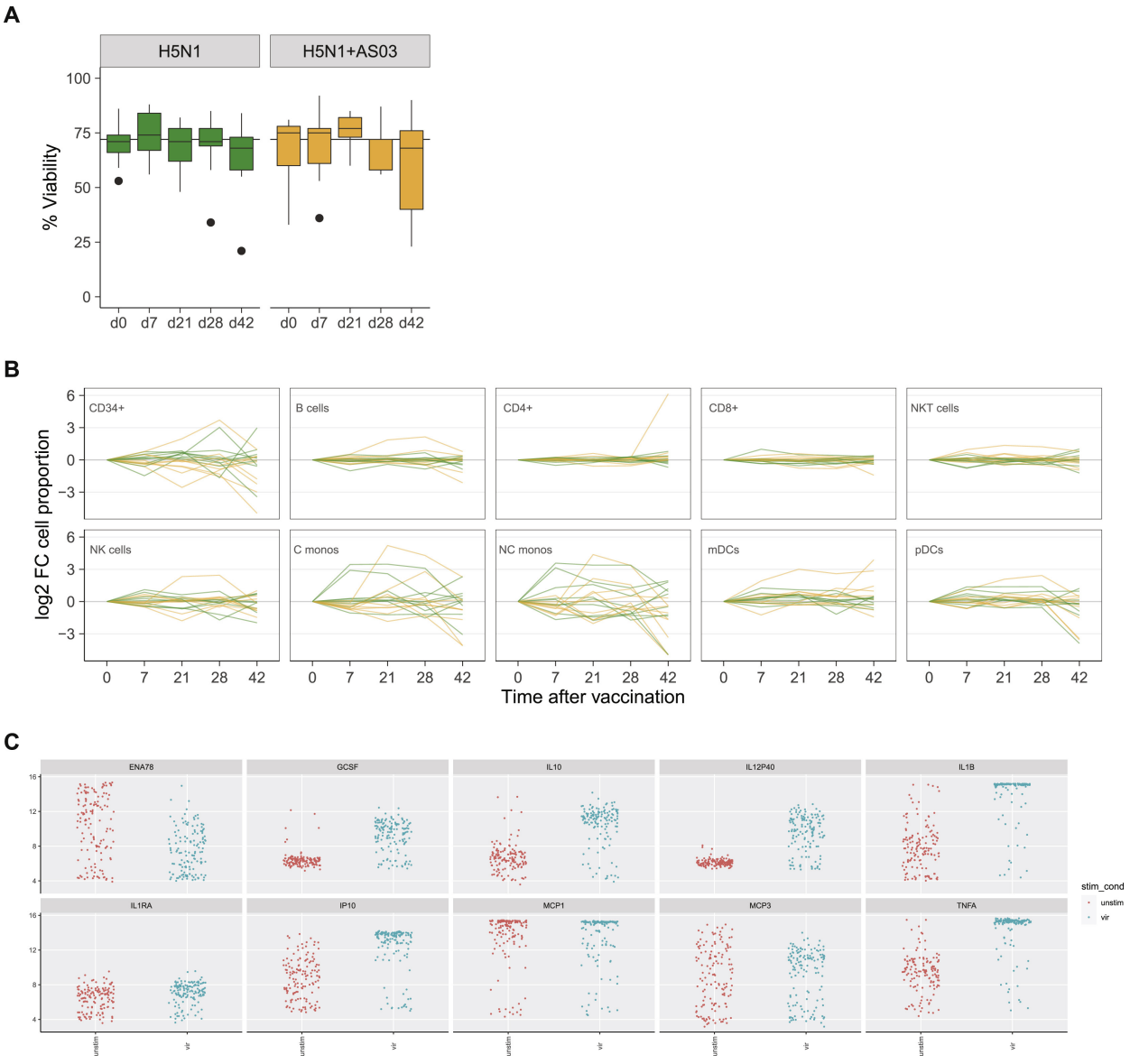
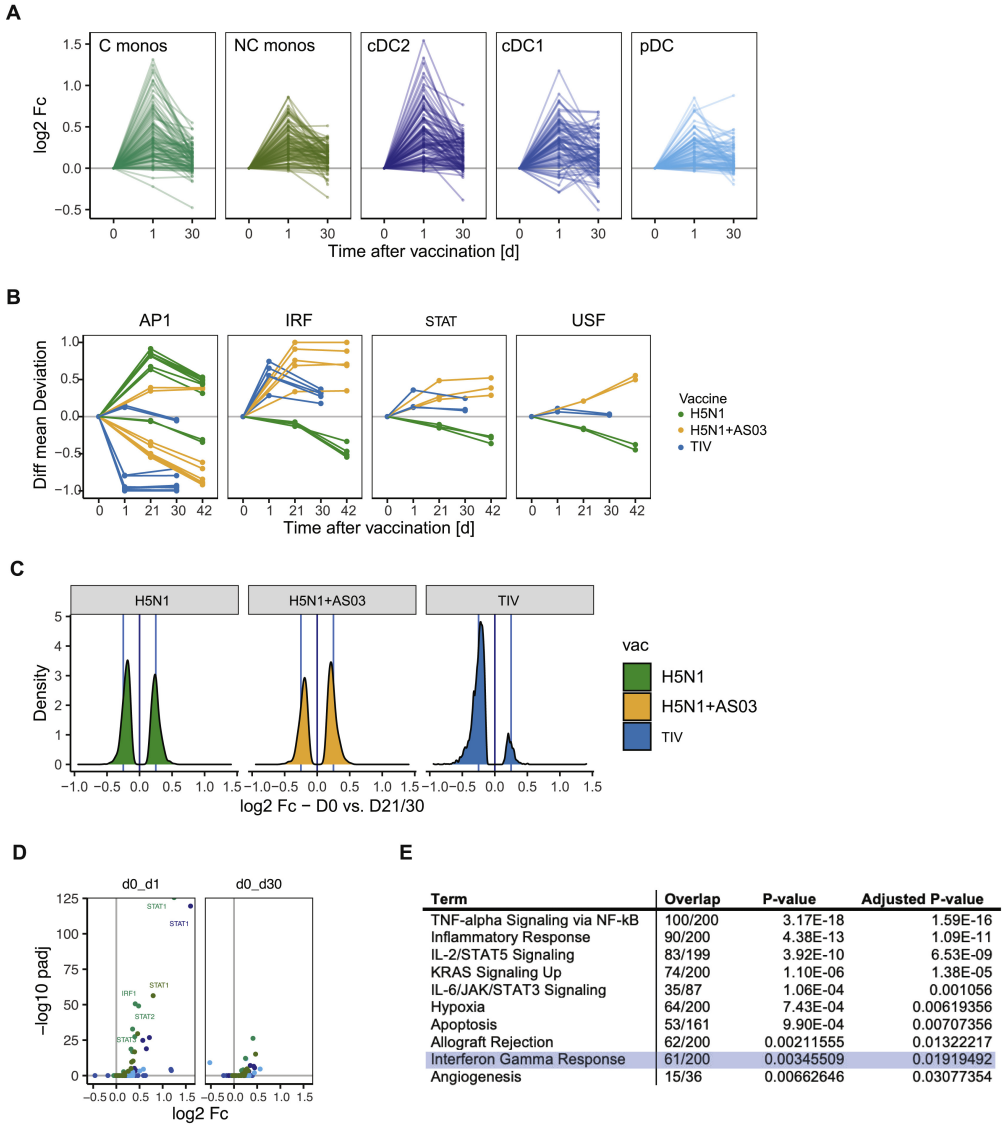


Figure S6: **Changes in cell abundance and cytokine production upon TLR stimulation, related to Figure 5. (A)** EpiTOF/Luminex PBMC viability after thawing by vaccination time point. **(B)** Change in cell type abundance per subject as measured by EPITOF. Wilcoxon signed rank test was used to compare changes at post-vaccine time points with d0 and p-values were corrected using the FDR approach. No comparison passed the threshold of $fdr <= 0.05$. **(C)** Dot plot showing log2 cytokine levels in each TLR-stimulated PBMC culture by stimulation condition.



Term	Overlap	P-value	Adjusted P-value
TNF-alpha Signaling via NF-kB	100/200	3.17E-18	1.59E-16
Inflammatory Response	90/200	4.38E-13	1.09E-11
IL-2/STAT5 Signaling	83/199	3.92E-10	6.53E-09
KRAS Signaling Up	74/200	1.10E-06	1.38E-05
IL-6/JAK/STAT3 Signaling	35/87	1.06E-04	0.001056
Hypoxia	64/200	7.43E-04	0.00619356
Apoptosis	53/161	9.90E-04	0.00707356
Allograft Rejection	62/200	0.00211555	0.01322217
Interferon Gamma Response	61/200	0.00345509	0.01919492
Angiogenesis	15/36	0.00662646	0.03077354

Figure S7: **TIV induces transient type I IFN response in innate immune cells, related to Figure 6. (A)** Line plot showing the change in expression of interferon response genes (Hallmark gene set) after vaccination with TIV as determined by scRNA-seq. **(B)** Line graph showing the difference in transcription factor (TF) accessibility during vaccination in classical monocytes as determined by scATAC-seq. To compare between different vaccines, changes in accessibility at each time point were normalized by the maximum change at that time point. Each line represents a separate TF within the indicated family. **(C)** Histogram showing the change in accessibility for genomic regions containing an IRF1 binding motif at day 21 vs day 0 (H5N1±AS03) and day 30 vs day 0 (TIV) for classical monocytes (scATAC $P <= 0.05$ and occurring in at least 5% of cells) as determined by scATAC-seq. **(D)** Volcano plot showing change in gene expression for IRF/STAT TFs after vaccination with TIV as determined by scRNA-seq. **(E)** Enrichment analysis of bulk DARs with reduced accessibility at day 30 vs day 0 after TIV using Enrichr with the MSigDB Hallmark 2020 gene set.

2.9 Supplementary Materials

All supplemental materials for this chapter are included in *Chapter2_Additional_Data.pdf*. The additional data are:

- **DataS1** Subject and vaccine information, related to Methods
- **DataS2** EpiTOF panel and gating, related to Methods
- **DataS3** Vaccine-induced epigenomic changes by bulk ATAC-seq and RNA-seq, related to Figure 3
- **DataS4** ScATAC-seq and scRNA-seq analysis of immune response to TIV, related to Figure 4
- **DataS5** ScATAC-seq and scRNA-seq analysis of immune response to H5N1/H5N1+AS03, related to Figure 5

Chapter 3

PeakVI: A Deep Generative Model for Single Cell Chromatin Accessibility Analysis

This chapter is currently under review, has been posted to bioRxiv (2021), and is reported here in the most recent form. The authors on the manuscript are:

Tal Ashuach[1,2], Daniel A. Reidenbach[2], Adam Gayoso[1,2], Nir Yosef[1,2,3,4,*]

1. Center for Computational Biology, University of California, Berkeley, CA, USA.
2. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA.
3. Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA
4. Chan Zuckerberg BioHub, San Francisco, CA, USA

∗ Corresponding author. Email: niryosef@berkeley.edu

3.1 Abstract

Single-cell ATAC sequencing (scATAC-seq) is a powerful and increasingly popular technique to explore the regulatory landscape of heterogeneous cellular populations. However, the high noise levels, degree of sparsity, and scale of the generated data make its analysis challenging. Here we present PeakVI, a probabilistic framework that leverages deep neural networks to analyze scATAC-seq data. PeakVI fits an informative latent space that preserves biological heterogeneity while correcting batch effects and accounting for technical effects such as library size and region-specific biases. Additionally, PeakVI provides a technique for identifying differential accessibility at a single region resolution, which can be used for cell-type annotation as well as identification of key cis-regulatory elements. We use public datasets to demonstrate that PeakVI is scalable, stable, robust to low-quality data, and outperforms current analysis methods on a range of critical analysis tasks. PeakVI is publicly available and implemented in the scvi-tools framework: `https://docs.scvi-tools.org/`.

3.2 Introduction

Regulatory elements in the genome tend to reside in regions of open chromatin, making the landscape of chromatin accessibility a valuable target of study. Several molecular assays have been developed to support this effort [1, 2, 3], among them ATAC-seq [4], in which accessible regions are fragmented, and the corresponding DNA fragments are sequenced and mapped back to the reference genome, accumulating in areas of open chromatin. Recent advances in sequencing technologies enable performing this assay in single cells [5], thereby allowing the study of chromatin variability at a single cell resolution. Application of Single-cell ATAC-seq (scATAC-seq) has led to promising results in discerning sources of variation, beyond those observed at the transcriptional level [6, 7] and allowed for high resolution characterization of the regulation of in continuous processes, e.g., in immunity [6].

Despite the potential of scATAC-seq, analyzing the resulting data remains challenging. scATAC-seq assays have generally limited sensitivity, detecting 5-15% of accessible regions [7], a common issue for single cell genomics. Additionally, the coverage of this data is limited a-priori since each genomic region has at most two copies in a single cell. Finally, scATAC-seq is extremely high-dimensional, often consisting of hundreds of thousands of genomic regions. These challenges require specialized processing and analysis methods that are designed to account for the specific properties of scATAC-seq data.

One common task for analyzing scATAC-seq is dimensionality reduction: transforming the data to a low-dimensional space that preserves the meaningful information in the original data. This step is crucial to make some downstream analyses, such as clustering and visualization, less noisy, more stable, and computationally tractable. Existing methods use various approaches to achieve this task. Some use methods developed for natural language processing (e.g., latent Dirichlet allocation used by cisTopic [8] and latent semantic analysis (LSA) used by ArchR [9]) that inherently handle sparse high-dimensional data, but do not

inherently account for confounding factors that do not have an analog in textual language, such as batch effects. Other methods reduce dimensionality by first aggregating individual regions in the scATAC-seq data to easily interpretable features, such as binding motif scores in the case of chromVAR [10] or gene activity scores in the case of Cicero [11], which makes the data easier to analyze but masks the fine-grain single-region resolution provided by scATAC-seq. These methods have been demonstrated to be under-powered in capturing the true heterogeneity in the original data [12]. Finally, recent methods use deep generative models (e.g., SCALE [13]), but do not account for technical factors, and suffer from model over-fitting due to the dimensionality of the data in contrast with the limited number of samples.

Another common task is differential accessibility analysis. The ability to identify chromatin regions that are preferentially accessible in one population compared with another is foundational to characterizing the chromatin remodelling between cellular identities and states. However, specialized methods to perform this task in the context of scATAC-seq data have not yet been developed. Methods that rely on aggregation of individual regions, like chromVAR and Cicero, perform differential analyses in the aggregated space, thereby losing the single-region resolution. Other methods use linear models developed for RNA-seq data [14] or standard statistical tests [9]. These approaches often suffer from numerical instability due to the sparsity of the data, and statistically overpowered due to the large sample size.

Some recent processing pipelines, like SnapATAC [14] and ArchR [9], offer comprehensive end-to-end analysis pipelines that resolve many issues with processing scATAC-seq data, such as sensitive peak-calling, promoter-enhancer association, and doublet detection. However, for the fundamental tasks mentioned above these pipelines rely on methods that were not optimized for scATAC-seq data, and can therefore be improved upon.

Here we present PeakVI, a deep generative model that learns a probabilistic low-dimensional representation of single cells from their chromatin accessibility landscape. PeakVI accounts for technical biases in the data stemming from batch effects, variation in sequence coverage, and bias due to the width of DNA regions, and creates a representation of the data that minimizes these effects. The representation is provided at two levels. One part of the model infers a representation for each cell in a latent low-dimensional space. This latent representation and the space it is embedded in can be used directly for downstream analyses: integration of data sets, identification of cellular sub-populations, and visualization. A second part of the model provides a corrected, probabilistic representation of the raw data. This high dimensional representation enables statistically robust inference of single region-level differential accessibility and cell state annotation. We demonstrate PeakVI's performance on published data and benchmark it against state-of-the-art published methods on a range of analysis tasks. We show that PeakVI is a powerful addition to the arsenal of scATAC-seq methods and provides capabilities that can help unlock the full potential of scATAC-seq data analysis. PeakVI is publicly available as part of the scvi-tools [15] suite of deep generative models for single cell genomics.

3.3 Results

PeakVI Model

PeakVI leverages variational inference with deep neural networks to model scATAC-seq data. For each cell, PeakVI estimates the probability of each chromatin region being accessible, as well as technical factors that affect the probability of an accessible region being observed. The standard output of most scATAC-seq preprocessing pipelines (including those employed here; see Methods) is a table of N cells and K genomic regions. The regions typically correspond to DNA segments with enriched accessibility that are inferred through peak-calling over cell aggregates [5, 14, 9].

The starting point of PeakVI is therefore a $N \times K$ matrix X where x_{ij} is the number of reads from cell i that map to region j. While these observations are counts, the underlying biology is mostly binary (a region is either accessible or not). Therefore, PeakVI models the observations as samples from a Bernoulli distribution $P\left(x_{ij} > 0 \mid y_{ij}, r_j, \ell_i\right)$, where y_{ij} is the probability of region j being accessible in cell i, $r_j \in [0, 1]$ is a region-specific scaling factor, and $\ell_i \in [0, 1]$ is a cell-specific scaling factor (Figure 1A). Conceptually, these components are related to the three molecular events that are required for a region to be observed as accessible: (1) the region must be accessible in the cell, which largely depends on the cell state and identity, captured by y_{ij}; (2) the accessible region must be tagmented by the transposase that underlies the ATAC-seq protocol, a process which may be skewed by region-specific factors such as width (in base pairs) and sequence biases, captured by r_j; (3) finally, the corresponding fragment must be captured and sequenced, which may also depend on library-specific factors, such as sequencing depth and efficacy of the library preparation, captured by ℓ_i.

PeakVI uses a variational autoencoder [16] (VAE) and an auxiliary neural network to estimate these factors. The VAE consists of two major components: (1) the encoder network f_z infers the distributional parameters of the d-dimensional (for $d \ll D$) latent representation z_i (also known as the variational posterior) from the observed data: $f_z\left(x_i\right) = q\left(z_i | x_i\right)$; (2) the decoder network g_z and the generative model, which takes a sample from the latent representation z_i and the batch annotations s_i and generates an estimate of the probability of each genomic region being accessible in the cell i: $\left(g_z\left(z_i, s_i\right)\right)_j = y_{ij}$. The cell-specific scaling factor ℓ_i is inferred from the observed data using an additional neural net f_ℓ, and the region-specific scaling factor $r_j \in [0, 1]$ is optimized directly as a model parameter. Finally, the probability of observing a region in a cell (i.e $p\left(x_{ij} > 0\right)$) are computed as the product

of the three probabilities: $p\left(x_{ij} > 0\right) = y_{ij} \cdot \ell_i \cdot r_j$ (Figure 1A). Formally:

$$
\begin{aligned}
\left(\mu_i, \sigma_i\right) &= f_z\left(x_i\right) && \text{Infer distributional parameters} \\
z_i &\sim \mathcal{N}\left(\mu_i, \sigma_i\right) && \text{Sample latent representation} \\
y_{ij} &= \left(g_z\left(z_i, s_i\right)\right)_j && \text{Estimate probability of accessibility} \\
\ell_i &= f_\ell\left(x_i\right) && \text{Estimate cell-specific factor} \\
x_{ij} > 0 &\sim Ber\left(y_{ij} \cdot \ell_i \cdot r_j\right) && \text{Calculate likelihood}
\end{aligned}
$$

Conditioning on batch annotations, or any other known sources of unwanted variation, encourages the encoder to capture batch-independent biological variation in the latent representation z_i, which can then be used for normalized and batch-corrected visualization, clustering, and other downstream analyses. The inferred accessibility probabilities y_{ij} are an estimate of the true chromatin landscape in each cell, while technical effects that stem from either region-specific biases or cell-specific biases are captured by the r and ℓ scaling factors, respectively. We can then estimate the probability of observing a region in each cell as the product of these factors $y_{ij} \cdot \ell_i \cdot r_j$ and compute the likelihood of the observations. During training, a lower bound of the marginal log likelihood $\log p(x_{ij} > 0)$ is then maximized using auto-encoding variational Bayes [16]. Full model architecture and training parameters are provided in the Methods section.

Benchmark Datasets

In order to evaluate the performance of PeakVI, we examined both simulated and real datasets. We found, however, that current simulation techniques [12] rely on independent sampling from distributions attained from bulk ATAC-seq data, which creates a highly-sparse covariance structure that does not realistically reflect assayed datasets (Figure S1). Our analysis therefore relies primarily on two publicly available datasets: (1) Hematopoiesis data from Satpathy et al. [6] which consists of bone marrow and blood samples that were flow-sorted for different cell subsets, as well as several batches of unsorted samples that consist of multiple cell types; (2) A dataset released by 10x Genomics of joint RNA-seq and ATAC-seq from single human peripheral blood mono-nuclear cells (PBMCs). The first dataset contains cell type specific labels that provide an established benchmark, as well as multiple batches that allow comparison of batch effect correction. The second dataset provides an orthogonal modality of data that can be used to validate scATAC-based analyses. Finally, the two datasets are generated using different protocols and are processed differently, allowing us to demonstrate the PeakVI's performance is protocol- and processing-independent.

PeakVI Captures Nuanced Effects of Technical Confounders

Since the normalization factors included in the PeakVI model, r and ℓ, are optimized by the training process, we set out to confirm that they converge on values that correspond to the empirical, technical confounders. We used the 10X PBMC data for these analyses. For the

region-specific factor r, we examined how it corresponds to the width of the genomic region, a known technical confounder. we found that PeakVI assigns the vast majority of regions with a value around 0.5, with higher values indeed being assigned to wider regions, which have a higher probability of being fragmented (Figure 1B). Notably, the overall distribution of this factor only reaches as high as roughly 0.75, well below the max value of 1. This translates to a global penalty imposed on all observations, which implicitly reflects the limited sensitivity of this assay and the resulting abundance of false-negative observations. For the cell-specific factor ℓ we examined how it corresponds to the number of reads captured in each cell. We find that the vast majority of cells have $\ell \approx 1$, and the dynamic values of ℓ indeed correspond to the empirical library size (Figure 1C). The saturation of this factor reflects an important consideration when normalizing library sizes for chromatin profiling: different cell types may have different levels of accessibility (e.g unbalanced chromatin remodeling during differentiation [17]), therefore this factor should not penalize cells states with less accessible chromatin, but rather only weigh down cells in cases where the decrease in fragments is due to technical effects. Overall we see that the normalization factors used by the model have a clear but nuanced correspondence to empirical confounders.

PeakVI is Robust to Increased Sparsity and Stable Across Hyperparameters

Limited sensitivity, which results in an abundance of missing observations, is a major problem in single-cell assays and particularly scATAC-seq. We therefore examined how PeakVI handles increasing levels of sparsity. We corrupted the 10X PBMC data by randomly replacing non-zero observations with zeros at a range of probabilities ($10 - 90\%$) and trained PeakVI on each corrupted dataset. We then used PeakVI's estimates of the probability of accessibility for these corrupted observations and compared the estimates from the models trained on corrupted data, in which these observations were 0, to the original estimates from the model trained on the full data, where these observations where non-zero. We computed the error: $\frac{1}{|C|}\sum_{ij \in C}\left(y^c_{ij} - y_{ij}\right)^2$, where C is the set of corrupted observations, y^c is the probability of accessibility estimated by peakVI when trained on the corrupted data, and y is the probability of accessibility estimated from the original, uncorrupted, data. We found that PeakVI produces highly consistent results, even in highly sparse situations: with a mean squared error of 0.06 when 10% of the observations are removed, to 0.17 when 90% of the data is removed (Figure 1D, S2). We also observed that the corrupted estimates are generally lower than the original estimates, consistent with the corruption being one-directional (introducing false negatives, not false positives). These results demonstrate that PeakVI is robust to low-quality and highly sparse data.

Since training PeakVI involved stochastic optimization of a non-convex function, the model can produce different results in different runs. We examined how stable PeakVI is to changes in architecture and training hyperparameters by training PeakVI on a variety of configurations and comparing how the different models perform on held-out data. We varied the number of hidden layers in the neural networks, the size of the mini-batch used in training, the dropout rate, and learning rate, and the weight decay. For each set of hyper-

paramters, we trained the model 3 times, and measured the likelihood the model achieves on the held-out data in each run. We found that PeakVI is highly stable, and that the default hyperparameters perform well without a need to fine-tune the model for each analysis (Supp. Table 1, Methods). Finally, to see how PeakVI stability is impacted by the sparsity of the data, we artificially corrupted data to only retain 50% and 10% of observations, and repeated the stability analysis. In both cases, while the model performance decreased compared with the full data, the model remained stable in terms of hyperparameters, indicating that the default hyperparameters perform well even in highly-sparse situations (Supp. Tables 2-3).

PeakVI Learns an Informative Batch-Corrected Latent Representation

PeakVI learns a low-dimensional representation of each cell that preserves biological heterogeneity while reducing noise, technical artifacts, and batch effects. We compared the latent space learned by PeakVI with representations from published methods. We compared to four methods: 1) latent semantic analysis (LSA), a natural language processing (NLP) technique commonly used in scATAC-seq analysis pipelines, such as Signac [18] and ArchR [9]; 2) cisTopic [8], which uses Latent Dirichlet Allocation; 3) SCALE [13], which also employs a VAE and incorporates Gaussian mixture modelling (GMM) to create a clustered latent space; 4) chromVAR [10], an algorithm that aggregates genomic regions by known binding motifs and normalizes these aggregates to motif accessibility scores. The first two methods, LSA and cisTopic, were chosen since a recent benchmark of computational analysis methods for scATAC-seq methods [12] found them to be the best performing methods. SCALE is included in our comparison due to the conceptual similarities with PeakVI. Finally, we included chromVAR since it is commonly used as both a dimensionality reduction method as well as an annotation technique.

First we used the 10X PBMC scATAC-seq data to measure how consistent each latent representation is with the gene expression profiles that are also measured from each cell. We ran all methods on the 10X PBMC data and extracted the latent representation computed by each. We then independently analyzed the paired scRNA-seq data and clustered the cells based on their gene expression profiles (Methods). We then overlaid the scRNA-based cluster labels on the scATAC-based representations (Figure 2A), and measured for each cell the fraction of its chromatin-based K nearest neighbors that are from the same RNA-based cluster for varying values of K (Figure 2B, Methods). We found that PeakVI and cisTopic outperformed all other methods, with PeakVI doing marginally better than cisTopic. We also measured how robust each method is to library size effects, by computing for each latent space the correlation of the latent representation with the empirical library size ($\log$(number of fragments)), using Geary's C [19] (Figure S3, Methods). We found that LSA and SCALE are especially sensitive to library size effects, while PeakVI and cisTopic are more robust, and chromVAR is insensitive to library size effects.

Next we looked into how each method handles a more complex experimental design, as

in the hematopoiesis dataset, which consists of multiple samples of different sizes, some cell type specific and others general. We analyzed the data with all methods. For completeness, we also included a variation of LSA used by the ArchR pipeline [9] called Iterative LSA (Methods), as well as three configurations of PeakVI: (1) "no batch", without any batch annotation; (2) "full batch", treating each sample as a separate batch; (3) "replicate batch", treating each replicate from multi-replicate conditions as a separate batch (Methods). These configurations correspond to having no batch correction, strict batch correction, or an intermediate approach, respectively. We examined how well each method preserves biological heterogeneity by measuring how separated the sorted cell populations are, using the cell type-specific fluorescence-based labels (Figure 2C, S4, S5). We also examined how well each method handles batch effects, which none of the examined methods explicitly corrects, by measuring how well-mixed are the four different batches of unsorted PBMC samples (Figure 2D, S4). For both analyses we computed an enrichment score by computing for every cell the number of neighbors out of its K-nearest neighbors that share its label, and comparing to the random expectation (Methods), for varying values of K (scores in the text are for $K = 50$) (Figure 2E). Ideally, this enrichment score would be high for biological labels and low for batch labels. We find that LSA, cisTopic, and PeakVI with no-batch configuration all achieve high separation (enrichment scores 9.1, 9.13, and 9.42, respectively) but separate the different batches as well (enrichment scores 2.33, 2.28, 2.39 respectively); conversely, chromVAR and SCALE outperform all methods in batch mixing (1.57 and 1.59, respectively), but do worse on cell type separation (5.78 and 7.03, respectively). Iterative LSA seems to underperform on both tasks. In contrast, we find that PeakVI with replicate-batch strikes a desirable balance, preserving biological heterogeneity comparably well (enrichment score 9.04) while more effectively mixing the batches (enrichment score 1.85). Finally, PeakVI with full-batch configuration also achieves a good balance (8.37 for cell type separation, 1.88 for batch mixing), but underperforms the replicate-batch configuration on both tasks. Overall these results demonstrate that PeakVI is better able to correct batch effects while preserving biological heterogeneity, reaching an overall better latent representation than all examined methods.

PeakVI performs differential accessibility analysis at a single-region resolution

Among the main promises of scATAC-seq is the ability to better identify individual genomic elements that help regulate certain biological processes. Achieving this requires the ability to identify individual regions that are differentially accessible between different groups of cells. In practice this task is challenging due to the binary nature of each observation, batch effects, and the high levels of noise and sparsity. Most differential analyses thus choose to aggregate the differential signal across different regions, either by the binding motifs they harbor (i.e the differential analysis chromVAR performs) or by aggregating the surrounding regions to each gene and creating a gene activity score [11]. While these analyses are useful, they

do not enable identification of individual regions, thereby not fully unlocking the promise of scATAC-seq data. Some differential analyses are performed in single-region resolution: ArchR [9] uses Wilcoxon rank-sum test, and Signac [18] uses a logistic regression model which models the total number of fragments to account for library size effects. Both of these approaches offer partial solutions to the noise and sparsity issues presented by scATAC-seq.

PeakVI addresses this problem by leveraging the probabilistic nature of the latent space to produce denoised and normalized estimates of accessibility, which enable a robust and accurate estimate of differential accessibility at a single-region resolution. Briefly, given a population of cells C and a region j, PeakVI samples from the area of the latent space that corresponds to C and estimates the probability of region j being accessible for each sample, then averages over the samples to get a stable estimate of accessibility: Y_{C_j} (Methods). Importantly, the representation of the latent space using random variables means that each cell in the original data can be sampled multiple times, allowing PeakVI to sample beyond the available number of observed cells. Additionally, this procedure can be conditioned on batch annotation, thereby correcting batch effects. When comparing two populations of cells, C_A and C_B, we use the absolute difference between estimates ($Y_{C_B} - Y_{C_A}$) as a measure for the extent of differential accessibility (effect size). Compared to ratio-based statistics (e.g odds-ratio), this estimate is more interpretable (representing absolute increase or decrease in binding propensity) and more stable to low-level signals. For instance, this means that an increase from 0.01 to 0.21 will be equivalent to an increase from 0.7 to 0.9 as opposed to the first being a 20-fold increase and the second being a 1.3-fold increase.

Using PeakVI estimates for differential accessibility is more sensitive and robust than using the observed data directly

To compare the estimated effect from PeakVI to the empirical effect calculated directly from the observations, we used the hematopoiesis data, and the replicate-batch PeakVI model. We define the empirical accessibility as the proportion of cells in C in which j is observed as accessible: $X_{C_j} = \sum_{i \in C} \mathbb{1}(x_{ij} > 0)$, and the empirical effect is defined equivalently to the estimated effect, as $X_{C_B} - Y_{C_A}$. We clustered the latent representations of the cells and ran a series of comparisons for each cluster. First, we ran two comparisons for each cluster: 1) a "biological" comparison, comparing all cells within the cluster to all other cells; 2) an "artifactual" comparison, comparing within each cluster cells that originated from the two large PBMC batches (replicates 1 and 2; excluding clusters with less than 5 cells in either group), (Figure 3A). The biological comparisons are a common use for differential analyses where some real differences in accessibility are expected, whereas the artifactual comparisons are used as negative controls. We ran two additional comparisons for each cluster, comparing cells within that cluster that originated from a given PBMC batch (either replicate 1 or 2) to all cells in all other clusters, which essentially provided two technical replicates of the biological analysis (denoted 'biological b1' and 'biological b2').

We first measured the correlation between the PeakVI estimated effects and the raw data (empirical) effects. We found that the effects are highly correlated in biological comparisons

(mean Pearson correlation 0.97), but less so in artifactual comparisons (mean correlation 0.52) (Figure 3B). We then used the results from "biological b1" and "biological b2" results, and found that the estimated effect is highly reproducible (mean correlation 0.95), while we see a marked decrease in reproducibility of the empirical effect (mean correlation 0.66) (Figure 3C). We also noticed that while the results were highly correlated, there was a difference in the width of the distributions between the estimated and the empirical effects (Figure 3C, S6). To investigate this effect more thoroughly, we calculated the standard deviation of the distributions for each comparison, and found that in all biological comparisons (including "biological b1" and "biological b2") the estimated effect had a wider distribution than the empirical effect, whereas in artifactual comparisons the distributions were either similarly wide or the estimated effect had a narrower distribution (Figure 3D). We additionally found that this is related to the number of cells included in the compared groups, especially in comparisons that rely on small numbers of cells: in these cases we observed the least difference in standard deviations for the biological comparisons, and the most difference for the artifactual comparisons (Figure 3E).

Taken together, these results demonstrate that PeakVI is amplifying the empirical effect when the effect corresponds to real biological difference, but silences it when it's a product of noise. When the empirical effect is more susceptible to noise (e.g., smaller number of cells included in the comparison), PeakVI is less able to amplify biological signal, but more efficient in silencing the noise. In contrast, when the empirical effect is calculated with a large number of cells, and is therefore less noisy, PeakVI has less silencing effect, but is able to amplify real differences better.

Statistical significance with PeakVI

To estimate the statistical significance of differential effects, PeakVI uses techniques described in previous methods from our group [20, 21]. Briefly, during the sampling procedure described above, PeakVI considers pairs of samples, one from each of the compared groups (y_a, y_b). PeakVI determines for each pair if the measured effect for each region j is greater than some minimal effect size δ: $h_j = \mathbb{1}(|y_{C_b} - y_{C_a}| > \delta)$ (for one-sided tests: $h_j = \mathbb{1}(y_{C_a} > y_{C_b} + \delta)$). We repeat this many times, and define the probability of differential accessibility, p^j_{DA}, as the proportion of pairs for which $h_j = 1$ (Methods). We then use a conservative multiple hypothesis correction procedure previously described by Lopez et al. [21] to identify differentially accessible regions with some nominal false discovery rate.

Established pipelines perform this analysis using generalized linear models (e.g Signac [18]) or standard statistical tests like the Wilcoxon rank-sum test or a two-sided T-test(e.g ArchR [9]). We therefore compared out differential accessibility analysis with a generalized linear model (GLM) equivalent to that used by Signac: a logistic regression with an additional covariate for the number of fragments in each cell to avoid library size effects dominating the analysis (Methods), as well as to a Wilcoxon rank-sum test used by ArchR. We performed two comparisons using all methods: (1) an artifactual comparison, using the hematopoiesis data we compared between cells from the two PBMC replicates that mapped to cluster 1, corre-

sponding to cells the NK-cell label (Figure 3G); (2) a biological comparison, comparing cells from the NK-cell sample to cells from the B-cell sample (using only cells that were FACS-sorted) (Figure 3H). We found that all approaches show a clear relationship between effect size and statistical significance in both analyses. Both GLM and Wilcoxon results revealed two common issues: i) some regions have a very large effect size but are not statistically significant, corresponding to regions that have very low detection rates in both populations; ii) the p-values were inflated due to the large sample size. In the artificial comparison, where no biological signal is expected, PeakVI correctly identified no regions as differentially accessible, compared with 910 regions identified by the GLM model and 6761 regions identified by the wilcoxon rank-sum test. In the biological comparison, PeakVI identified 11362 (16.5%) regions as differentially accessible, compared with 33679 (48.9%) identified by the GLM, and 26410 (19.7%) identified by Wilcoxon test.

We then ran an equivalent comparison between B-cells and NK-cells using bulk ATAC-seq data from Calderon et al. [22] with sorted immune cell populations, as a ground truth (Methods), and compared the results with the scATAC-seq based results from both analyses (Figure 3I). Overall results from all methods are consistent with the bulk results, but PeakVI achieves higher correlation between the effect sizes (0.74 compared with 0.48 and 0.52 for the GLM and Wilcoxon results, respectively). In terms of correctly identifying differential regions, for both PeakVI and Wilcoxon, 86% of the regions identified were also differential according to the bulk analysis, compared with 65.6% for the GLM. In terms of overlap between the regions found with bulk comparison vs. single cell, all analyses resulted in sets of regions that are over-represented at the bulk results, with PeakVI reaching an odds-ratio of 1.92, Wilcoxon reaching 1.93, and GLM reaching 1.47. Overall, these results demonstrate that PeakVI provides a well-calibrated statistical significance estimation and enables identification of differentially active regions at a single-region resolution, while minimizing false discovery and avoiding numerical issues due to low detection rates.

PeakVI supports multiple approaches for annotation and discovery of cell states

A major challenge in analyzing scATAC-seq data is the lack of region-based annotations of cell state, in contrast to the abundant resources for RNA-based annotation. Current methods therefore rely on annotations that were generated from gene expression profiles, which are useful but only provide a partial solution, since chromatin accessibility may carry information that is not discernible from gene expression alone. We therefore set out to demonstrate two different approaches for how PeakVI can be leveraged for annotation and downstream discovery. First, PeakVI's integration capabilities can be used for transfer learning, projecting annotated reference data and un-annotated query data onto a joint space, and transferring insights from the former to the latter. Importantly, this approach relies solely on the regions, without associating regions to target genes or identifying harbored motifs. Secondly, in the lack of an annotated reference, PeakVI's differential accessibility

analysis can be leveraged for de-novo annotation, associating marker regions with nearby genes and identifying enriched gene sets or known marker genes.

PeakVI can be used for transfer learning, by leveraging an annotated reference dataset to annotate a query dataset. First, the reference and query datasets need to be integrated into a joint space, which can be achieved using PeakVI in one of two ways: (i) naively, by analyzing both datasets together and conditioning on the dataset of origin; (ii) using a two-step procedure first presented in scArches [23], in which the reference data is processed in advance, and then incoming query data can be projected onto the reference-based space. The scArches procedure is particularly useful when creating a detailed atlas to be used as a reference resource. After the query and reference are in a shared space, transferring annotations from one to the other can be done using proximity based classifiers, such as KNN or cluster majority vote (which we utilized here). We demonstrate this ability using the hematopoiesis data as the reference, and a dataset of human PBMCs provided by 10X as a query (note that this dataset is different from the multiomic dataset used in previous sections). Notably, the reference data covers both bone marrow and blood, and consists of samples that were sorted to specific cell types, as well as samples that consist of the entire PBMC compartment. We therefore expect the query data to align only to the parts covered by the reference PBMC samples, and not next to cell subsets that are more abundant in the bone marrow. Furthermore, we expect technical hurdles to complicate the integration of the datasets as they were generated by different experimental protocols and processed with different computational pipelines.

We began by creating a reference model, by analyzing the hematopoiesis data using PeakVI in a scArches-compatible configuration (Methods). We then used PeakVI to project the query PBMC data onto the reference space. PeakVI was able to mix the datasets well, only mapping query cells onto areas of the space occupied by PBMCs, but not those corresponding to progenitor cells, which are absent from the query PBMC data (Figures 4A, S7). We then clustered the cells and assigned each cluster with the most abundant cell-specific FACS-based label in that cluster from the reference data. Importantly, these annotations are based on similarity of chromatin landscapes between cells in the query and reference data, without any association to other biological features or aggregation, resulting in a straightforward labelling of the query data (Figure S8).

However, this procedure requires an annotated atlas from a corresponding system, while many scenarios require de-novo annotation, which PeakVI facilitates using the differential accessibility analysis. We demonstrate this using the hematopoiesis data, by de-novo annotating the data and using the FACS-based labels as a ground-truth. We first clustered the latent space (Figure 4D), and consistent with our previous findings we found that clusters tend to consist primarily of cells that have the same label. Next, using our differential accessibility analysis, we compared each cluster to all other clusters except for the 3 most similar clusters, to avoid highly-similar clusters masking the differences (Methods). For each cluster we used a one-sided test to only identify regions that are preferentially open in the target cluster. We then used enrichr [24, 25] to associate the regions to nearby genes and leveraged the ARCHS4 [26] collection to find over-represented cell-type specific gene signatures. We

were able to confidently identify many of the cell type-specific clusters, which matched their FACS-based label (Figure4E, Methods). For instance, marker regions for clusters 13 and 17, in which labelled cells are overwhelmingly B-cells, were indeed enriched for regions associated with B-cell marker genes; Cluster 1 marker regions were enriched for NK-cell marker genes, and indeed the labelled cells in that cluster are NK-cells. Similarly signatures for CD4+ T-cells, Regulatory T-cells, and pDCs, were all highly enriched in the clusters with the corresponding FACS-based labels. Thus, using PeakVI and gene-based signatures, we are able to annotate the data and recapitulate many of the FACS-based labels.

These results are nonetheless limited by the availability of gene signatures, which may not be available for all cell types, or provide only a a high-level annotation at a limited resolution. Specifically, most progenitor cells in the hematopoiesis data could not be annotated in a similar fashion for lack of corresponding signatures, and despite clustering separately, both CD4+ naïve T-cells and CD4+ memory T-cells were annotated simply as CD4+ T-cells, since higher-resolution signatures were not available. PeakVI can therefore be used in a two-step approaches whereby cells can be stratified into broad types, using reference-based annotation, and then assigned with more high resolution labels of cell sub-types or states using de-novo analysis. As a case in point, we focused on the set of cells which were annotated as B cells in our reference-based analysis. These cells can be divided into two clusters (clusters 13 and 17). To derive a higher resolution annotation of the B cell compartment, we ran a two-sided comparison between the two clusters and identified 1043 differentially accessible regions in total, 207 preferentially accessible in cluster 13 and 836 preferentially accessible in cluster 17 (Figure 4F; Supp Table 4, Methods). Among the genes associated with regions detected for cluster 13 we found TCL1A, known to be expressed throughout B-cell differentiation up to naive B-cells but silenced in memory B-cells and plasma cells [27, 28], and YBX3, implicated in B-cell differentiation as an immature B cell marker [29]. We also found SATB1, TENT5A, and ZNF667-AS1, which along with TCL1A and YBX3, were previously found to be differentially expressed in naive B-cells compared with Memory B-cells [30]. Concordantly, genes associated with cluster 17 included known markers for memory B-cells AIM2[31] and CD80 [32], and 9 other genes previously found to be differentially expressed in memory B-cells compared with naive B-cells [30] (Figure 4G). Taken together, we concluded that cluster 13 consists of naive B-cells and cluster 17 consists of memory B-cells, therefore demonstrating that PeakVI's differential accessibility analysis can be used in conjunction with a reference-based annotation to increase the resolution of annotations and identify new targets for further study.

3.4 Discussion

PeakVI is a deep generative model for analyzing single cell chromatin accessibility data. The model is designed to explicitly account for various technical effects that mask and distort the biological signal. The latent representation learned by the model is probabilistic in nature, embedding the observed cells in a smooth variational space that preserves the

biological heterogeneity, minimizes confounding effects, and can be used directly to explore the chromatin landscape of a population of cells. Importantly, PeakVI takes as input a region-by-cell count matrix, allowing the user to integrate PeakVI with current and future preprocessing and peak-calling methods.

PeakVI improves upon previous attempts to use deep learning to analyze scATAC-seq data in several manners. First, the architecture used in the underlying neural networks scales with the size of the input data, increasing the expressiveness of the model to match with increasingly large and complex datasets (Methods). Second, PeakVI accounts for technical confounders and enables correction of batch effects, with clear benefits to downstream results. Thirdly, Since it is common for features (regions) to outnumber the samples (cells), and the observations are mostly binary and therefore contain little information, PeakVI also takes measures to successfully prevent the model from over-fitting, by holding out some of the data as a validation set, tracking the model's performance on the validation data, and halting the training process when the performance on the validation data stops improving, thus ensuring that the model is learning generalizable features. Finally, PeakVI provides extensive methods to take advantage of the learned latent space for analysis tasks beyond dimensionality reduction, visualization, and clustering. Specifically, PeakVI enables high resolution annotation of cell state, by allowing both reference-based analysis and de-novo annotation analysis. In that capacity, PeakVI enables accurate differential accessibility analysis at a single-region resolution that reduces the effect of confounders and avoids common issues with the current practices for differential accessibility, namely numerical instability and inflation of significance scores.

Since PeakVI takes as input a region-by-cell matrix, it does not offer a full end-to-end solution to all of the challenges presented by scATAC-seq, instead relying on other methods and pipelines to perform upstream tasks such as fragment alignment, peak calling and cell calling. This allows users to match PeakVI with other methods, for instance using specialized peak-callers like Lancetron [33] or AtacWorks [34], and analyzing the resulting matrix with PeakVI to benefit from superior dimensionality reduction, batch correction, differential accessibility, and annotation. Additionally, PeakVI is implemented in the scvi-tools suite[15] which provides interfaces with popular processing environments like scanpy [35] and Signac [18]. Finally, PeakVI is robust to low-quality data, easy to configure, train, and use. It can be easily incorporated in existing analysis pipelines to enhance current analyses for dimensionality reduction, batch correction, differential accessibility, and annotation.

3.5 Methods

The PeakVI Model

Let $X \in \mathbb{N}_0^{N \times K}$ be a scATAC-seq region-by-cell matrix with N cells and K regions, where $x_{ij} \in \mathbb{N}_0$ is the number of fragments from cell i that map to region j. Since PeakVI models the probability of observing a region, regardless of the number of reads supporting that obser-

vation, the observations are treated as binary: $X^* \in [0,1]^{N \times K}$, where $x^*_{ij} = \mathbb{1}(x_{ij} > 0)$. The observations are therefore generated from a Bernoulli distribution $x^* \sim Ber(q_{ij})$. PeakVI computes q_{ij} as a product of three probabilities: $q_{ij} = y_{ij} \cdot r_j \cdot \ell_i$, where y_{ij} captures the true biological heterogeneity; r_j captures region-specific biases (e.g width, sequence); ℓ_i captures cell-specific biases (e.g library size). The three probabilities are estimated jointly using deep neural networks.

The biological component y_{ij} is estimated using a VAE[cite:kingma], which is composed of two deep neural networks, the encoder f_z and decoder g_z. Briefly, the encoder $f_z : \mathbb{N}_0^K \to \left(\mathbb{R}^D, \mathbb{R}^D\right)$, computes the distributional parameters of a D-dimensional multivariate normal random variable: $Z \sim MVN\left(f_z\left(x_i\right)_1, f_z\left(x_i\right)_2\right)$. The sample is then concatenated to the batch annotation for cell i, and passed through the decoder $g_z : \left(\mathbb{R}^D, \{0,1\}^S\right) \to [0,1]^K$, for S being the dimension of the one-hot batch annotation (the number of batches). The cell-specific factor ℓ_i computed from the input data for cell i via a deep neural network $f_\ell : \mathbb{N}_0^K \to [0,1]$. Finally, the region-specific factor r_j, since it is optimized across samples, is stored as a K-dimensional tensor, used and optimized directly.

Architecture

All PeakVI neural nets are fully connected networks, composed of repeated blocks that share a basic structure. For convenience, we define a fully connected block $FC(I, O, D, A)$ as having a fully connected layer with I input nodes and O output nodes, followed by a dropout layer with a D probability of dropout, a layer-norm layer, and finally an A activation function.

The encoder f_z is constructed as follows:

$$
\begin{gathered}
FC\left(N, \sqrt{N}, 0.1, \text{leakyReLU}\right) \to \\
FC\left(\sqrt{N}, \sqrt{N}, 0.1, \text{leakyReLU}\right) \to \\
FC\left(\sqrt{N}, \sqrt{N}, 0.1, \text{leakyReLU}\right) \to \\
\left(FC\left(\sqrt{N}, \sqrt[4]{N}, 0.1, \text{Identity}\right), FC\left(\sqrt{N}, \sqrt[4]{N}, 0.1, \text{Identity}\right)\right)
\end{gathered}
$$

With $\sqrt[4]{N}$ being the default dimensionality of the latent representation. This ensures that the model architecture scales with the number of features in the data and the complexity of the representation.

The decoder g_z is constructed as follows:

$$FC\left(\sqrt[4]{N}+S,\sqrt{N},0,\text{leakyReLU}\right)\rightarrow$$
$$FC\left(\sqrt{N},\sqrt{N},0,\text{leakyReLU}\right)\rightarrow$$
$$FC\left(\sqrt{N},\sqrt{N},0,\text{leakyReLU}\right)\rightarrow$$
$$FC\left(\sqrt{N},N,0,\text{sigmoid}\right)$$

With S as the dimensionality of the batch annotations, concatenated to the latent representation.

The cell-specific factor network f_ℓ is constructed similarly:

$$FC\left(N,\sqrt{N},0,\text{leakyReLU}\right)\rightarrow$$
$$FC\left(\sqrt{N},\sqrt{N},0,\text{leakyReLU}\right)\rightarrow$$
$$FC\left(\sqrt{N},\sqrt{N},0,\text{leakyReLU}\right)\rightarrow$$
$$FC\left(\sqrt{N},1,0,\text{sigmoid}\right)$$

Training Procedure

By default, PeakVI is optimized using AdamW[cite loshchilov] with a learning rate of 0.0001, weight decay of 0.001, and minibatch size of 128. The model is trained on 90% of the data, with the remaining 10% used as a validation set. Training is performed for at most 500 epochs, with early stop: if there is no improvement in terms of the reconstruction loss on the validation set for 50 epochs, the training stops. For epochs $i \in [1, 50]$ the KL divergence term is weighed done by a factor of $i/50$. The best state throughout training, defined as the state that achieves the best reconstruction loss, is saved during the training and used as the final state. All training settings are configurable.

Differential Accessibility Analysis

For a differential accessibility analysis between two populations A and B, the analysis is performed as follows:

1) N cells are sampled from each population, with replacement (default $N = 5000$). We denote the resulting cells C_X^i for the i-th sample from population X, for $i \in [N]$ and $X \in \{A, B\}$.

2) for each cell C, we apply the inference model on the cell's chromatin accessibility profile $f_z\left(x_C\right)$ to get the variation distribution corresponding to that cell, q_C, sample from that distribution to get an estimated profile of the probability of accessibility of all regions in that cell: z_C. We then use the generative model g_z to estimate the probability of accessibility

of each region j in that cell: $(y_C)_j$. Sampling from the variational space allows us to sample the same cell multiple times and get different estimates, thereby enabling statistical power beyond the original sample size.

3) to calculate the effect size for each region, we simply take the average estimated probability of accessibility across all samples from each population, and compute the absolute difference between the averages: $\Delta_j = (\bar{y_A})_j - (\bar{y_B})_j$.

4) to calculate the statistical significance, we randomly pair samples from each population into N pairs of estimates $\left\{(y_A, y_B)^i \,|i \in [N]\right\}$, then for each region we count for how many pairs the difference between estimates was greater than some minimal δ (default 0.05): $p_{DA_j} = \frac{1}{N}\sum_i^N \mathbb{1}\left((y_A)_j^i - (y_B)_j^i > \delta\right)$. This procedure has been previously described by [21].

5) In addition to p_{DA}, we also compute the Bayes factor: $BF_j = \log \frac{p_{DA}}{1-p_{DA}}$, and perform multiple testing correction using the procedure previously described by Lopez et al[21] to get a qualitative, binary label for each region.

Benchmarking and Evaluation

Stability Analysis

To measure the stability of PeakVI to hyperparameter selection, we ran a full grid search using the 10X Genomics sample data. We held out 10% of the data as a test set and trained all models on the remaining set. We trained each model 3 times (with an independent train-validation split) and measured the likelihood on the held-out data. The full results are available in Supplemental Table TODO. The hyperparameters we varied and the values used are as follows: learning rate (1e-2, 1e-3, 1e-4); number of hidden layers (1,2,3,4); dropout rate (0.1, 0.3); minibatch size (64, 128, 256); weight decay (0.1, 0.01, 0.001).

Dataset Processing

The hematopoiesis data was downloaded from GEO (Accession GSE129785); specifically the processed peak-by-cell matrix and metadata files: scATAC-Hematopoiesis-All.cell-barcodes.txt.gz, scATAC-Hematopoiesis-All.mtx.gz, scATAC-Hematopoiesis-All.peaks.txt.gz. We then filtered the genomic region to only those that are detected in at least 0.1% of the cells in the sample, reducing the data from TODO regions to TODO regions. The sample data from 10X genomics was also downloaded as preprocessed peak-by-cell matrices, without any additional filters.

Running Published Methods

For all methods, we followed the standard recommended procedure for analyzing data. For visualization, we computed the umap[cite:mcInnes] coordinates using the python implementation from the latent space computed by the respective method (except for SCALE, see below). **cisTopic** (v0.3.0): We used the WarpLDA model fitting procedure, and chose

the best number of topics based on the second derivative, as recommended by the package documentation. For the hematopoiesis data the model used 100 topics, and 40 topics for the paired PBMC sample data from 10X Genomics. **chromVAR** (v1.12.0): We used the JASPAR2016 motif set, containing 386 motifs, and followed the standard analysis outlined in the package documentation. We used the unnormalized motif deviation scores. For dimensionality reduction, we found no clear difference between using the chromVAR scores directly and applying an additional linear procedure (i.e principle component analysis). Results described in the manuscript use the deviation scores directly. **LSA:** We used the python implementation from the Scikit-learn [TODO: cite pedregosa]. We first binarized the data, then computed the top 50 components used the TruncatedSVD method, on the tfidf-transformed data. **SCALE** (v1.0.4): we used the external script to run SCALE without a pre-determined number of clusters, using the default arguments. In all visualizations, we used the umap coordinates computed by SCALE. **ArchR**: Since ArchR doesn't provide an interface to run the dimensioanlity reduction method directly on an existing count matrix, we downloaded the raw fragment files for each sample in the Hematopoiesis dataset and reanalyzed the data using ArchR. To maintain consistency, we used the same peaks as were used in the processed data, and when possible kept the same barcodes. Of the 27 samples, only 19 were successfully read by ArchR, resulting in the ArchR benchmark being limited to fewer cell types and fewer cells. **SnapATAC**: we could not successfully run snapATAC.

Enrichment Score Calculation

Enrichment scores used to quantify cell type separation and batch mixing were computed in an identical way. Given a latent representation R, an integer k, and cell labels L, we first compute $G_{R,k}$, the K-nearest neighbor graph from R with k neighbors. We then compute for each cell the proportion of neighbors that share the same label: $s_i = \frac{1}{k} \sum_{j \in G_{R,k}(i)} \mathbb{1} (L_i = L_j)$. The overall score is the average score across all cells, $\bar{s}$, normalized by the expected score for a random sample from the distribution of labels: $E[s] = \sum_{\ell \in \{L\}} p_\ell^2$, for $\{L\}$ being the set of available labels, and p_ℓ being the proportion of each label $\ell \in \{L\}$. The enrichment score is then $\frac{\bar{s}}{E[s]}$.

Differential Accessibiliy Analyses

As a simple benchmark for differential accessibility, we constructed a standard logistic regression model to compare B-cells to NK-cells, using the design $y \sim$ number of fragments + cell type, where y is the binary detection of a genomic region. We fit the model using the *glm* function in *R*. Due to the runtime of this analysis, we limited the results to regions that are detected in at least 1% of the compared cells. For the Wilcoxon rank-sum test, we used scanpy's *rank_genes_groups* implementation.

Analysis of bulk ATAC-seq data

The bulk ATAC-seq data used as a ground truth reference for differential accessibility analysis was downloaded from GEO (accession GSE118189). We used the unstimulated samples of all B-cell and NK-cell subtypes included in the study and used DESeq2[TODO: cite Love], whichwas found to be among the ebst performing methods for differential accessibility from bulk ATAC-seq data[TODO: cite gotras] for differential accessibility between the two group. We then found regions in the hematopoiesis data that overlap with the regions in the bulk data, and used the differential signal found in the bulk data for the overlapping regions in the hematopoiesis data.

Projection of query data onto reference

Projection of query data onto a latent space learned from reference data is done using scArches[TODO cite]. First, the 10X sample PBMC data was downloaded and processed (using CellRanger v3.1.0) using the hematopoiesis peaks. We then trained a PeakVI model on the hematopoiesis data using ell covariate injection, which adds one-hot encoded batch annotation to each layer in the VAE (as opposed to only the decoder layers, which is the default behavior). We then trained the resulting model on the query data, which involves adding batch annotations corresponding to the query data, and only training the nodes in the network that interact with these additional batches. This preserves the latent representation of the reference data while projecting the query data onto the same space, while correcting batch effects between the query and data.

Cluster Annotation with differential accessibility

Differential accessibility to identify marker regions for each cluster was performed between each cluster and all other clusters except the three most similar clusters. This was in order to avoid sampling pairs of cells that are highly similar from the two groups, which would reduce the signal. We therefore calculated the 'centroid' of each cluster (the average position in the latent space of all cells in the cluster), computed the Euclidean distance matrix between all centroids, and identified for each cluster the 3 most similar clusters. We then used the identified regions (using the Bayesian FDR method described by Lopez et al. [TODO:cite]), ran them through enrichr [TODO:cite], and downloaded the enrichment results for the ARCHS4 Tissues set. For associating regions with genes, we used the bioconductor package TxDb.Hsapiens.UCSC.hg19.knownGene[TODO:cite] and considered only strict overlaps between the region and the annotated gene body or promoter.

3.6 Figures

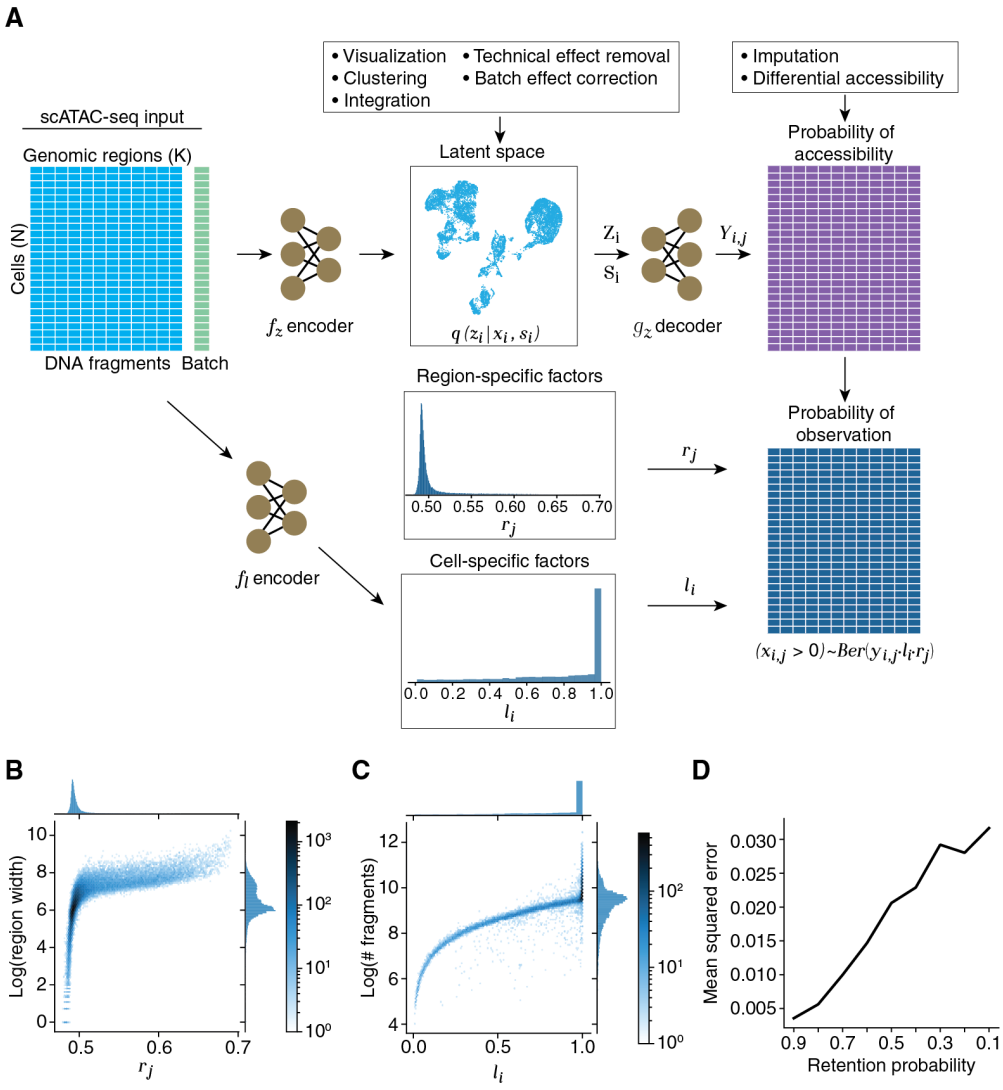


Figure 1: **PeakVI Model overview.** **(A)** conceptual model illustration. The input region-by-cell count matrix (left) is estimated as the product of region-specific effects (center top), cell-specific effects (center), and accessibility probability estimates (center bottom). The observation probability matrix (right) is used to calculate the likelihood of the data for optimization. **(B)** The region-specific factor r_j is assigned higher values for wider regions, indicating a higher probability of those regions being fragmented. **(C)** The cell-specific factor ℓ_i increases with the number of fragments up to a saturation point. Cells with sufficient fragments are not penalized even if other cells have significantly more fragments. **(D)** Random corruption of the data at increasing rates leads to a small but steady increase in the mean squared error (measured from corrupted indices).

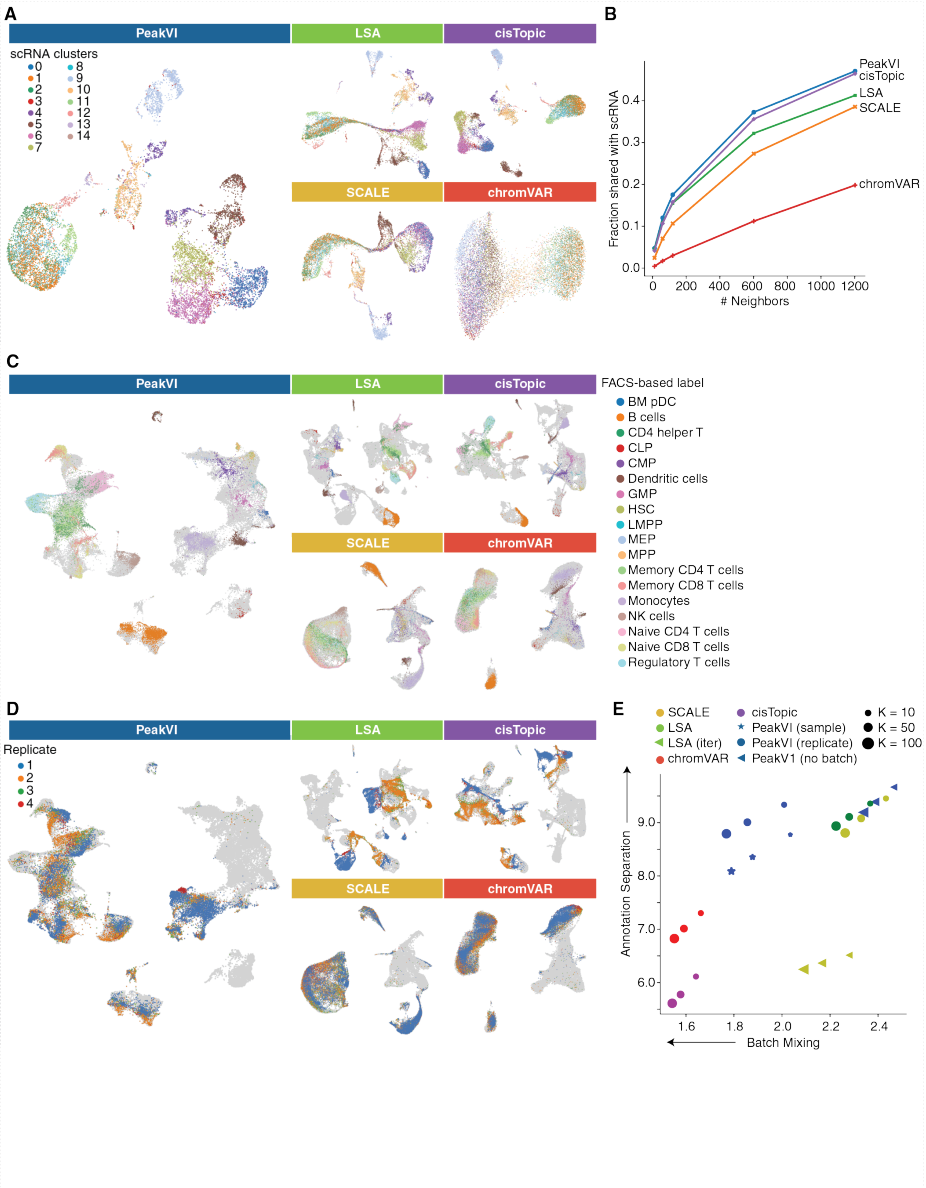


Figure 2: **UMAP visualizations of latent representations from PeakVI, LSA, cisTopic, SCALE, and chromVAR. (A)** The paired scRNA-scATAC sample PBMC dataset from 10X Genomics. Cells are colored based on the scRNA-based clustering, umaps are computed from the scATAC representations. All methods except for chromVAR are comparably consistent with the scRNA data. **(B)** Quantitative consistency of the latent representation with the scRNA data; fraction of the K nearest neighbors in the scATAC representation that are also among the K nearest neighbors in the scRNA representation, for various values of K. PeakVI marginally outperforms cisTopic, followed by LSA, SCALE, and chromVAR. **(C)** Data from Satpathy et al [6]; cells are colored using the FACS-based cell type-specific labels. Cells from unsorted samples or non-specific sorted samples are colored in light gray. PeakVI, LSA, and cisTopic all achieve good separation of cell types. **(D)** Data from Satpathy et al [6]; cells are colored using the unsorted PBMC replicates. Cells from all other samples are colored in light gray. Batch effects are reduced with PeakVI, chromVAR, and SCALE. **(E)** Enrichment of labels among the K-nearest neighbors for each cell; X-axis is the enrichment of batch labels, where lower enrichment indicates better batch mixing. Y-axis is the enrichment of cell type labels, where higher enrichment indicates better separation. PeakVI reaches a better balance of the two tasks.

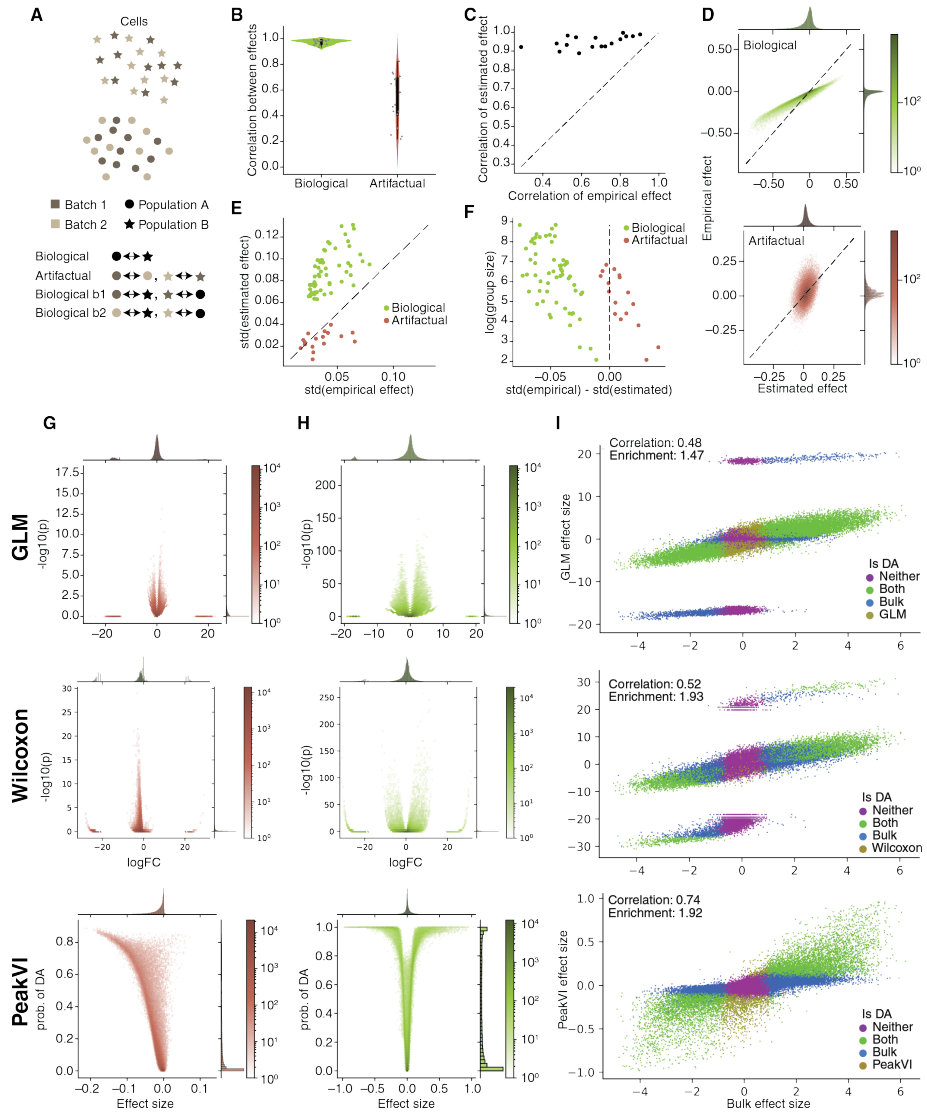


Figure 3: **Differential Accessibility Analysis with PeakVI. (A)** Illustration of the different comparisons. "real": compare cells between two population; "null": compare cells from different batches within a single population; "real b1"/"real b2": compare cells from a specific batch in a population to all cells in the other population. **(B)** Pearson correlations between the estimated and empirical effects. **(C)** correlation of effect size in 'real b1' and corresponding effect in 'real b2' comparisons. PeakVI estimated effects are far less sensitive to batch effects. **(D)** An example (using cluster 14) relationship between the PeakVI estimated effect to the empirical effect in real (top) and null (bottom) comparisons. **(E)** the width (measured by the standard deviation) of the effect distributions; PeakVI amplifies real differential effects, and silences nuisance ones. **(F)** Level of amplification/silencing depends on level of noise in the empirical effect. **(G-H)** Volcano plots for a GLM **(G)** and PeakVI **(H)** when comparing between two batches of NK-cells. **(I-J)** Volcano plots for a GLM **(I)** and PeakVI **(J)** when comparing between B-cells and NK-cells. **(K-L)** PeakVI **(L)** effect is better correlated with a bulk-ATAC based ground truth comparison, and the significant regions have a higher enrichment scores, compared with the GLM **(K)**.

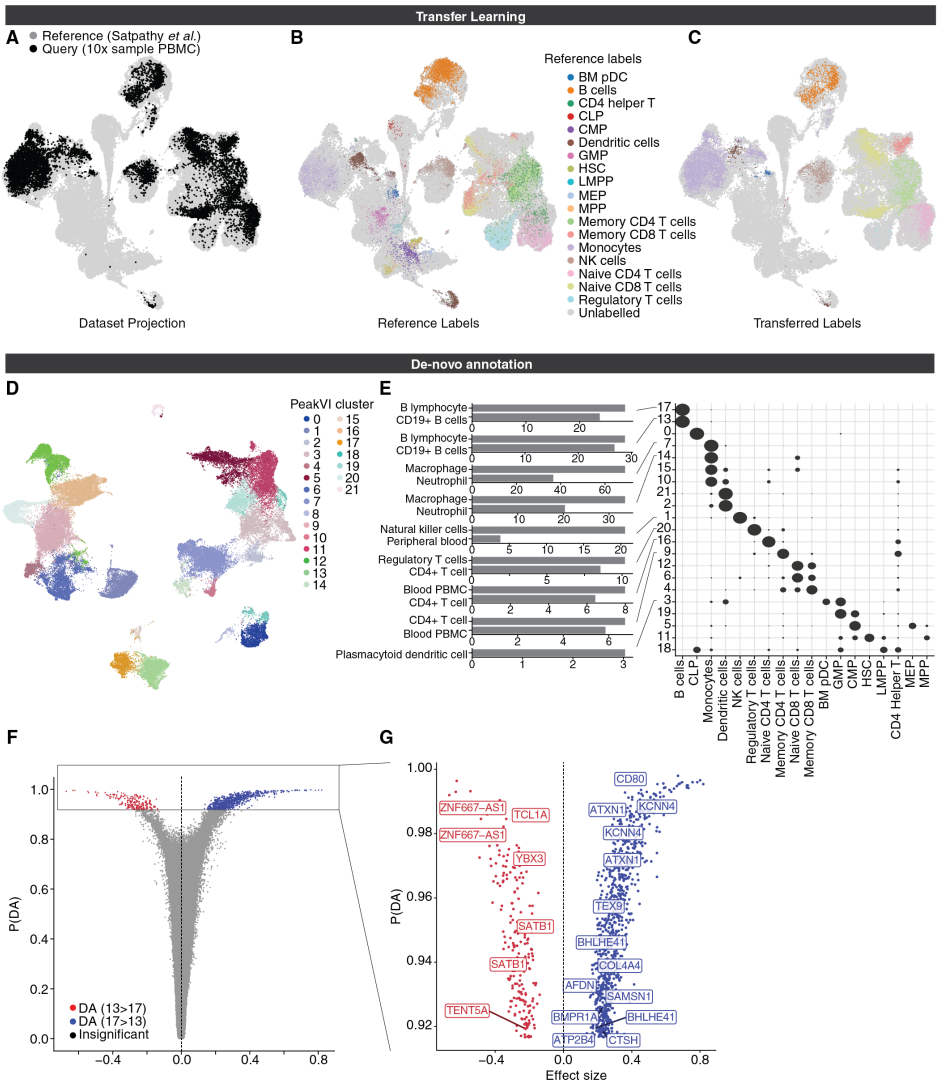


Figure 4: **PeakVI unlocks multiple paths for annotation and identification.** **(A-C)** PeakVI supports transfer learning. **(A)** Mapping of query data (Sample PBMC data from 10X Genomics) unto a reference data (from Satpathy et al[6]). PeakVI mixes the query data with the reference despite the data being generated by a different protocol and processed by a different pipeline. **(B)** The reference data, colored by FACS-based cell type-specific labels; **(C)** The query data, colored by the transferred cell type-specific labels. **(D-F)** De-novo annotation using PeakVI's differential accessibility analysis. **(D)** Hematopoiesis data colored by clusters. **(E)** Regions that are preferentially accessible in each cluster were analyzed for enriched cell-type signatures from ARCHS[26] signatures, using enrichr[24, 25]. Heatmap shows distribution of cell type-specific labels for each cluster, normalized by row. **(F)** Volcano plot for a differential accessibility analysis between the two B-cell clusters (clusters 13 and 17). **(G)** Volcano plot for only significant regions, labelled by associated genes that are implicated in Naive B-cells (red) and Memory B-cells (blue).

3.7 References

[1] Dustin E Schones et al. "Dynamic regulation of nucleosome positioning in the human genome". en. In: *Cell* 132.5 (Mar. 2008), pp. 887–898. ISSN: 0092-8674, 1097-4172.

DOI: 10.1016/j.cell.2008.02.022. URL: http://dx.doi.org/10.1016/j.cell.2008.02.022.

[2] Alan P Boyle et al. "High-resolution mapping and characterization of open chromatin across the genome". en. In: *Cell* 132.2 (Jan. 2008), pp. 311–322. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2007.12.014. URL: http://dx.doi.org/10.1016/j.cell.2007.12.014.

[3] Gregory E Crawford et al. "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)". en. In: *Genome Res.* 16.1 (Jan. 2006), pp. 123–131. ISSN: 1088-9051. DOI: 10.1101/gr.4074106. URL: http://dx.doi.org/10.1101/gr.4074106.

[4] Jason D Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". en. In: *Curr. Protoc. Mol. Biol.* 109 (Jan. 2015), pp. 21.29.1–21.29.9. ISSN: 1934-3639, 1934-3647. DOI: 10.1002/0471142727.mb2129s109. URL: http://dx.doi.org/10.1002/0471142727.mb2129s109.

[5] Jason D Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561 (June 2015), pp. 486–490. ISSN: 0028-0836. DOI: 10.1038/nature14590. URL: https://www.nature.com/articles/nature14590.

[6] Ansuman T Satpathy et al. "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion". en. In: *Nat. Biotechnol.* 37.8 (Aug. 2019), pp. 925–936. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0206-z. URL: http://dx.doi.org/10.1038/s41587-019-0206-z.

[7] Sebastian Preissl et al. "Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation". en. In: *Nat. Neurosci.* 21.3 (Mar. 2018), pp. 432–439. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-018-0079-3. URL: http://dx.doi.org/10.1038/s41593-018-0079-3.

[8] Carmen Bravo González-Blas et al. "cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data". en. In: *Nat. Methods* 16.5 (May 2019), pp. 397–400. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0367-1. URL: http://dx.doi.org/10.1038/s41592-019-0367-1.

[9] Jeffrey M Granja et al. "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis". en. In: *Nat. Genet.* 53.3 (Mar. 2021), pp. 403–411. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-021-00790-6. URL: http://dx.doi.org/10.1038/s41588-021-00790-6.

[10] Alicia N Schep et al. "chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data". en. In: *Nat. Methods* 14.10 (Oct. 2017), pp. 975–978. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4401. URL: https://doi.org/10.1038/nmeth.4401.

[11] Hannah A Pliner et al. "Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data". en. In: *Mol. Cell* 71.5 (Sept. 2018), 858–871.e8. ISSN: 1097-2765, 1097-4164. DOI: 10.1016/j.molcel.2018.06.044. URL: http://dx.doi.org/10.1016/j.molcel.2018.06.044.

[12] Huidong Chen et al. "Assessment of computational methods for the analysis of single-cell ATAC-seq data". en. In: *Genome Biol.* 20.1 (Nov. 2019), p. 241. ISSN: 1465-6906. DOI: 10.1186/s13059-019-1854-5. URL: http://dx.doi.org/10.1186/s13059-019-1854-5.

[13] Lei Xiong et al. "SCALE method for single-cell ATAC-seq analysis via latent feature extraction". en. In: *Nat. Commun.* 10.1 (Oct. 2019), p. 4576. ISSN: 2041-1723. DOI: 10.1038/s41467-019-12630-7. URL: http://dx.doi.org/10.1038/s41467-019-12630-7.

[14] Rongxin Fang et al. "Comprehensive analysis of single cell ATAC-seq data with SnapATAC". en. In: *Nat. Commun.* 12.1 (Feb. 2021), p. 1337. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21583-9. URL: http://dx.doi.org/10.1038/s41467-021-21583-9.

[15] Adam Gayoso et al. "scvi-tools: a library for deep probabilistic analysis of single-cell omics data". en. Apr. 2021. DOI: 10.1101/2021.04.28.441833. URL: https://www.biorxiv.org/content/10.1101/2021.04.28.441833v1.

[16] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (Dec. 2013). arXiv: 1312.6114v10 [stat.ML]. URL: http://arxiv.org/abs/1312.6114v10.

[17] Debattama R Sen et al. "The epigenetic landscape of T cell exhaustion". en. In: *Science* 354.6316 (Dec. 2016), pp. 1165–1169. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aae0491. URL: http://dx.doi.org/10.1126/science.aae0491.

[18] Tim Stuart et al. "Single-cell chromatin state analysis with Signac". en. In: *Nat. Methods* 18.11 (Nov. 2021), pp. 1333–1341. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-021-01282-5. URL: http://dx.doi.org/10.1038/s41592-021-01282-5.

[19] R C Geary. "The Contiguity Ratio and Statistical Mapping". In: *The Incorporated Statistician* 5.3 (1954), pp. 115–146. ISSN: 1466-9404. DOI: 10.2307/2986645. URL: http://www.jstor.org/stable/2986645.

[20] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". en. In: *Nat. Methods* 15.12 (Dec. 2018), pp. 1053–1058. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: http://dx.doi.org/10.1038/s41592-018-0229-2.

[21] Romain Lopez et al. "Decision-Making with Auto-Encoding Variational Bayes". In: (Feb. 2020). arXiv: 2002.07217 [stat.ML]. URL: http://arxiv.org/abs/2002.07217.

[22] Diego Calderon et al. "Landscape of stimulation-responsive chromatin across diverse human immune cells". en. In: *Nat. Genet.* 51.10 (Oct. 2019), pp. 1494–1505. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0505-9. URL: http://dx.doi.org/10.1038/s41588-019-0505-9.

[23] Mohammad Lotfollahi et al. "Query to reference single-cell integration with transfer learning". en. July 2020. DOI: 10.1101/2020.07.16.205997. URL: https://www.biorxiv.org/content/10.1101/2020.07.16.205997v1/.

[24] Edward Y Chen et al. "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". en. In: *BMC Bioinformatics* 14 (Apr. 2013), p. 128. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-128. URL: http://dx.doi.org/10.1186/1471-2105-14-128.

[25] Maxim V Kuleshov et al. "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". en. In: *Nucleic Acids Res.* 44.W1 (July 2016), W90–7. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw377. URL: http://dx.doi.org/10.1093/nar/gkw377.

[26] Alexander Lachmann et al. "Massive mining of publicly available RNA-seq data from human and mouse". en. In: *Nat. Commun.* 9.1 (Apr. 2018), p. 1366. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03751-6. URL: http://dx.doi.org/10.1038/s41467-018-03751-6.

[27] Michael A Teitell. "The TCL1 family of oncoproteins: co-activators of transformation". en. In: *Nat. Rev. Cancer* 5.8 (Aug. 2005), pp. 640–648. ISSN: 1474-175x. DOI: 10.1038/nrc1672. URL: http://dx.doi.org/10.1038/nrc1672.

[28] L Virgilio et al. "Identification of the TCL1 gene involved in T-cell malignancies". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 91.26 (Dec. 1994), pp. 12530–12534. ISSN: 0027-8424. DOI: 10.1073/pnas.91.26.12530. URL: http://dx.doi.org/10.1073/pnas.91.26.12530.

[29] Robin D Lee et al. "Single-cell analysis of developing B cells reveals dynamic gene expression networks that govern B cell development and transformation". en. July 2020. DOI: 10.1101/2020.06.30.178301. URL: https://www.biorxiv.org/content/10.1101/2020.06.30.178301v1.full.

[30] Nancy S Longo et al. "Analysis of somatic hypermutation in X-linked hyper-IgM syndrome shows specific deficiencies in mutational targeting". en. In: *Blood* 113.16 (Apr. 2009), pp. 3706–3715. ISSN: 0006-4971, 1528-0020. DOI: 10.1182/blood-2008-10-183632. URL: http://dx.doi.org/10.1182/blood-2008-10-183632.

[31] Alexandra Svensson et al. "Maturation-dependent expression of AIM2 in human B-cells". en. In: *PLoS One* 12.8 (Aug. 2017), e0183268. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0183268. URL: http://dx.doi.org/10.1371/journal.pone.0183268.

[32] N C Sahoo, K V S Rao, and K Natarajan. "CD80 expression is induced on activated B cells following stimulation by CD86". en. In: *Scand. J. Immunol.* 55.6 (June 2002), pp. 577–584. ISSN: 0300-9475. DOI: 10.1046/j.1365-3083.2002.01093.x. URL: http://dx.doi.org/10.1046/j.1365-3083.2002.01093.x.

[33] Lance D Hentges et al. "LanceOtron: a deep learning peak caller for ATAC-seq, ChIP-seq, and DNase-seq". en. Aug. 2021. DOI: 10.1101/2021.01.25.428108. URL: https://www.biorxiv.org/content/10.1101/2021.01.25.428108v3.

[34] Avantika Lal et al. "Deep learning-based enhancement of epigenomics data with AtacWorks". en. In: *Nat. Commun.* 12.1 (Mar. 2021), p. 1507. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21765-5. URL: http://dx.doi.org/10.1038/s41467-021-21765-5.

[35] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. "SCANPY: large-scale single-cell gene expression data analysis". en. In: *Genome Biol.* 19.1 (Feb. 2018), p. 15. ISSN: 1465-6906. DOI: 10.1186/s13059-017-1382-0. URL: http://dx.doi.org/10.1186/s13059-017-1382-0.

3.8 Supplementary Figures

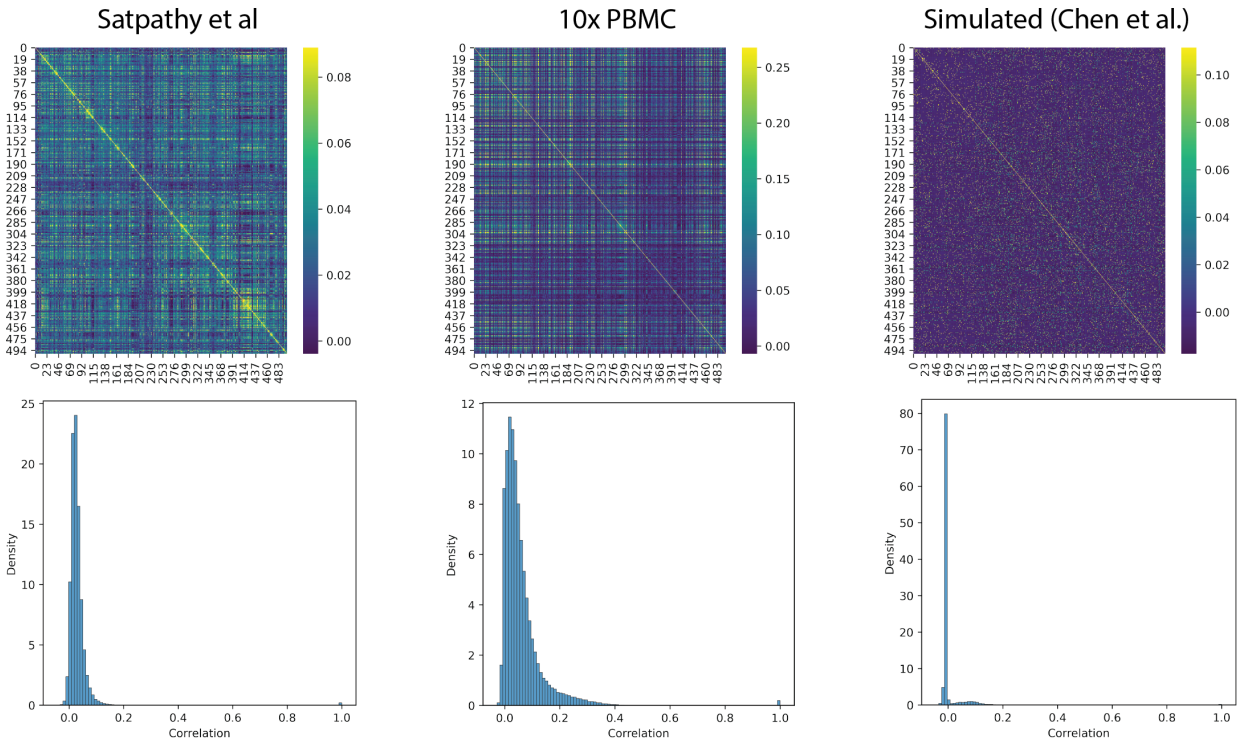


Figure S1: A Pearson correlation matrix (top) and distribution of correlation coefficients (bottom) of regions in three datasets: the immune cell dataset from Satpathy et al [6] (left); the sample multi-omics 10K cells PBMC dataset from 10x Genomics (center); and a simulated Bone Marrow dataset generated by Chen et al [12]. [cite]. For visual purposes, figures were generated using only the first 500 regions in each dataset, and across all available cells. Simulated data does not adequately represent the covariance structure of real scATAC-seq data.

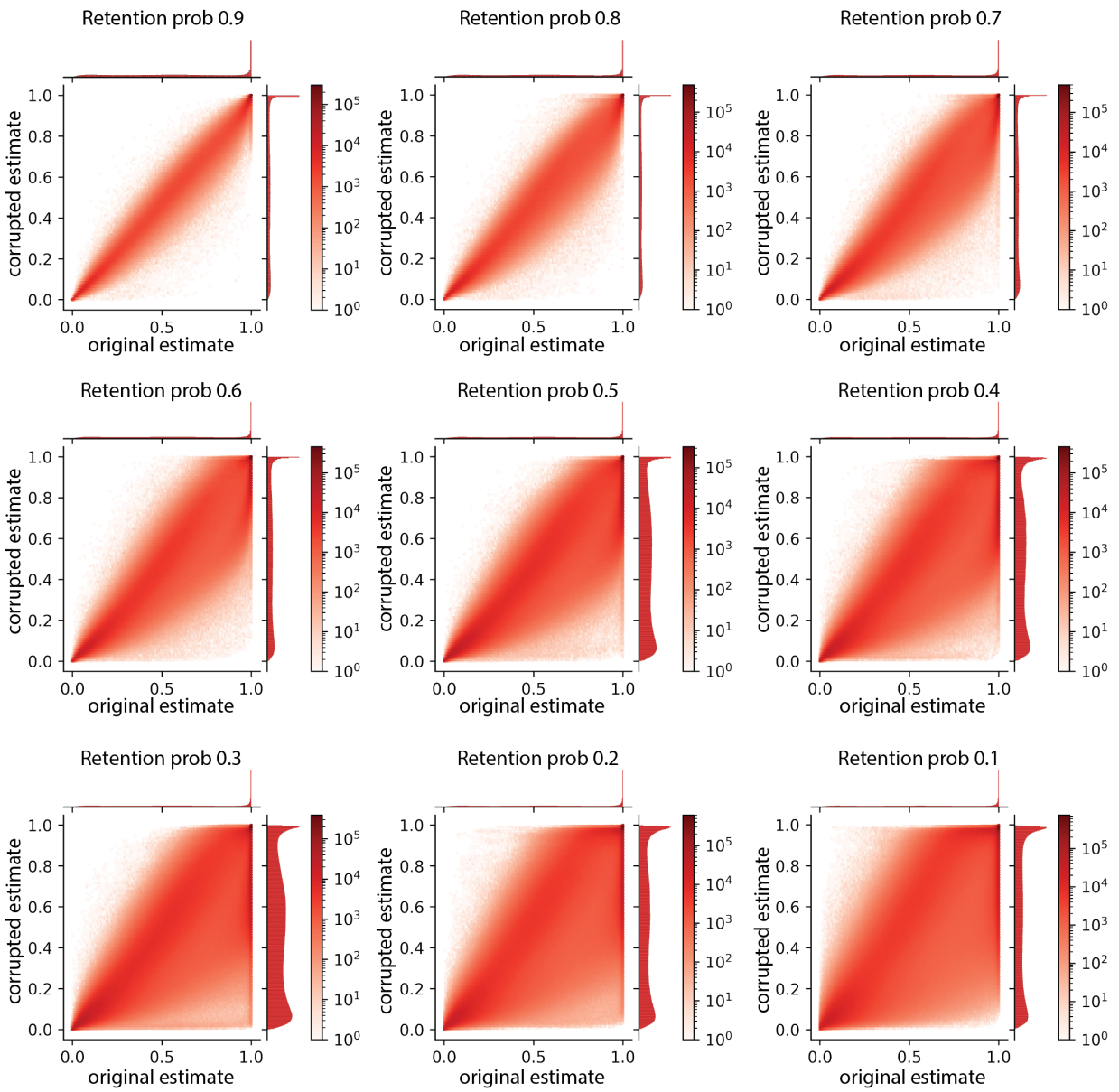


Figure S2: Corruption analysis, in which observations were randomly replaced by zeros. Visualization is limited only to corrupted indices, showing that while increased corruption destabilizes the model, PeakVI is overall highly robust to the sparsity of low quality data.

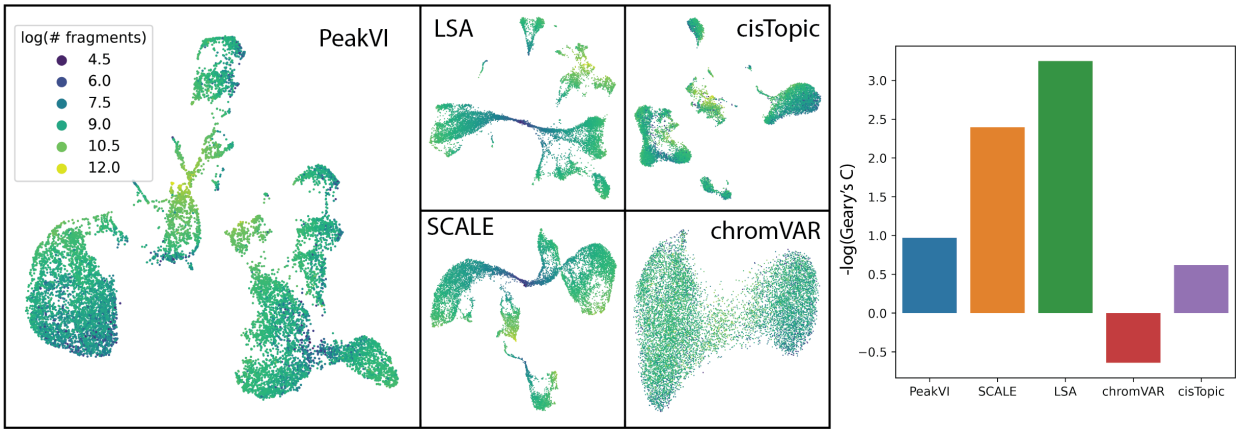


Figure S3: UMAPs of the sample paired scRNA and scATAC-seq PBMC data from 10X genomics, colored by the number of fragments mapped for each cell (left) and the spatial autocorrelation measured using Geary's C[19] (right). LSA and SCALE are most impacted by library size effects, PeakVI and cisTopic are robust, and chromVAR is negatively correlated.

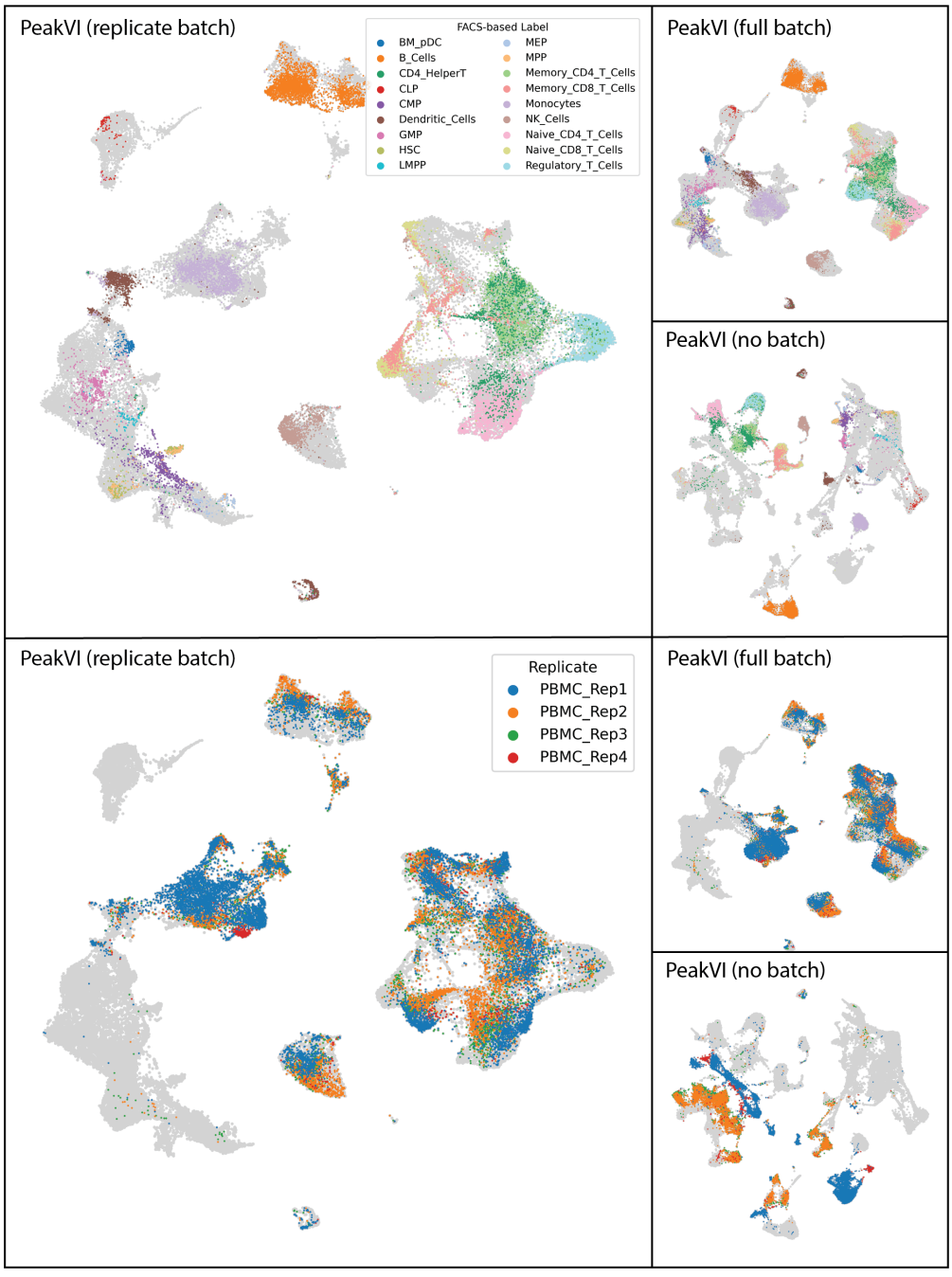


Figure S4: Visualizations of the Hematopoiesis data using three configurations of PeakVI: treating replicates of multi-replicate samples as separate batches (replcate batch); without batch correction (no batch); treating each sample as a separate batch (full batch). Colored by FACS-based labels (top) and replicates of the unsorted PBMC samples (bottom). Unlabelled cells are colored in light gray.

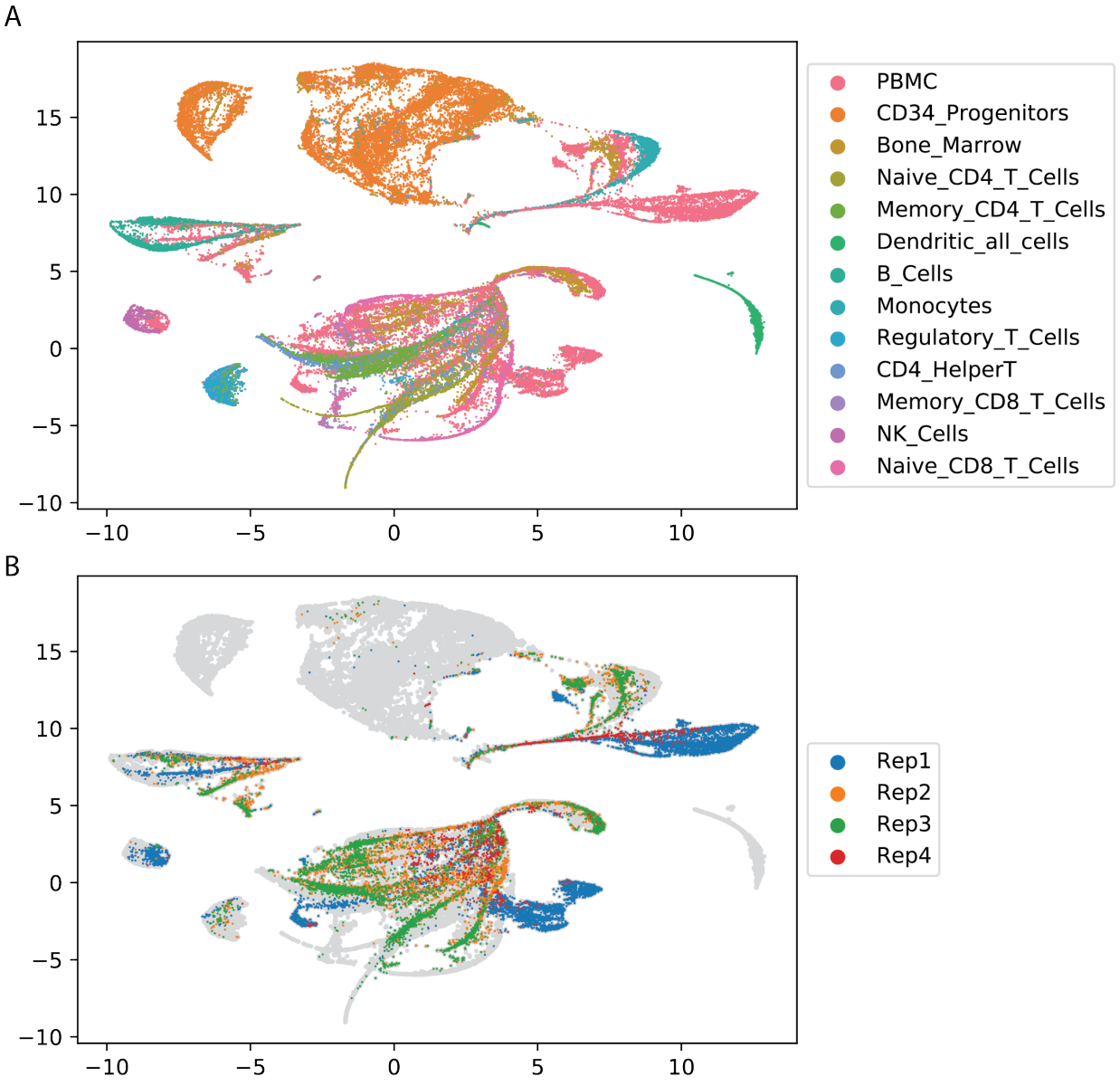


Figure S5: Visualization of the Hematopoiesis data using the ArchR dimensioanlity reduction (Iterative LSA), colored by cell type (A) and batch (B).

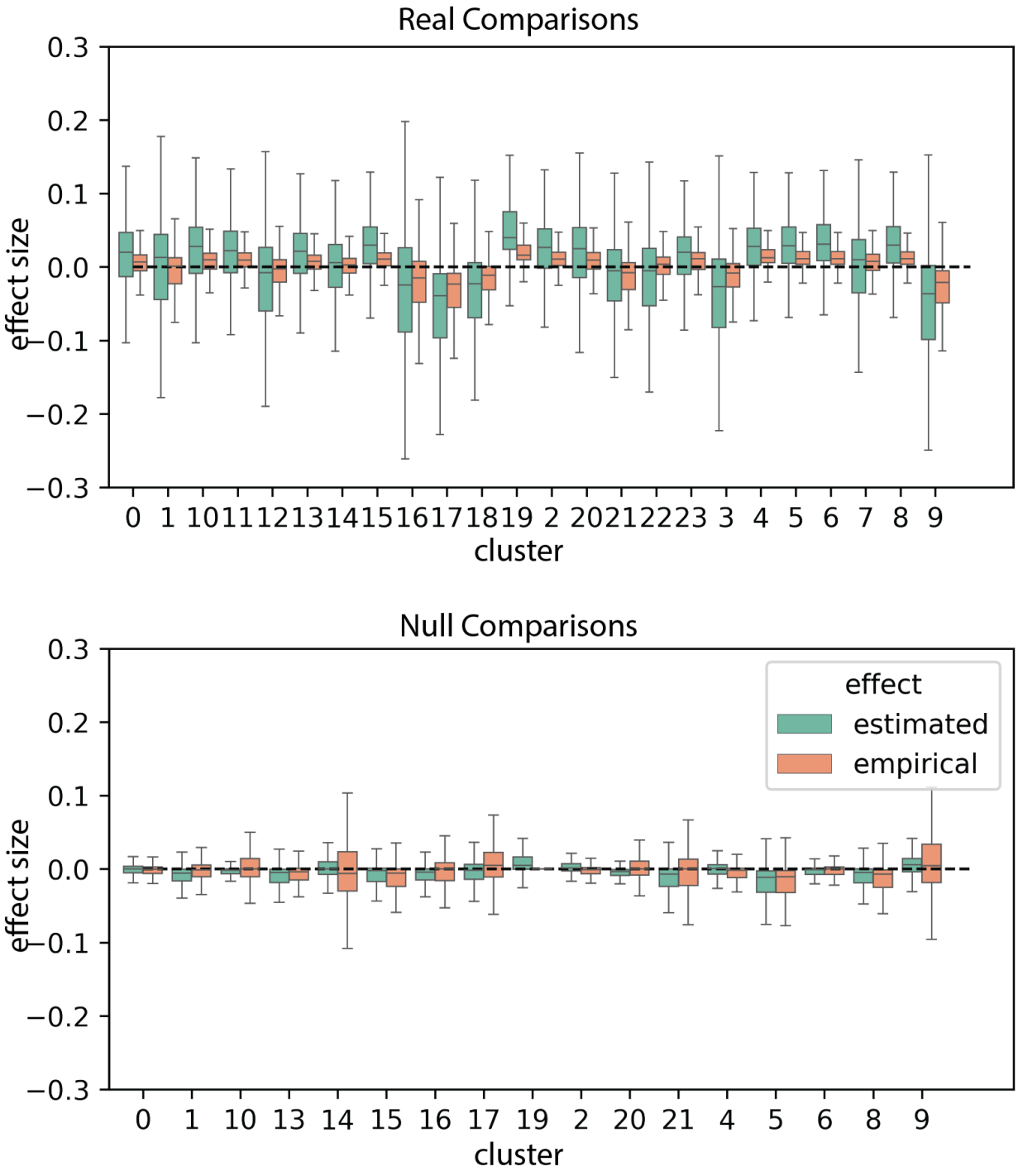


Figure S6: The effect size distribution for each real (top) and null (bottom) comparison. PeakVI estimated effects are amplified compared with the empirical effect in real comparisons, but the opposite is true for null comparisons. Overall PeakVI consistently has a better signal-to-noise ratio.

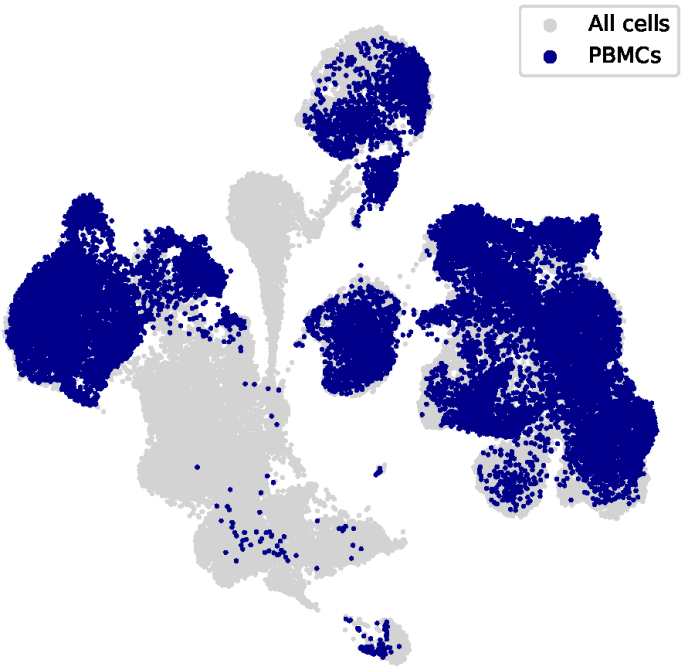


Figure S7: The low-dimensional representation of the Hematopoiesis data, trained in a scArches-compatible manner, with cells from PBMC samples in dark blue, showing how PBMCs are distributed in the space.

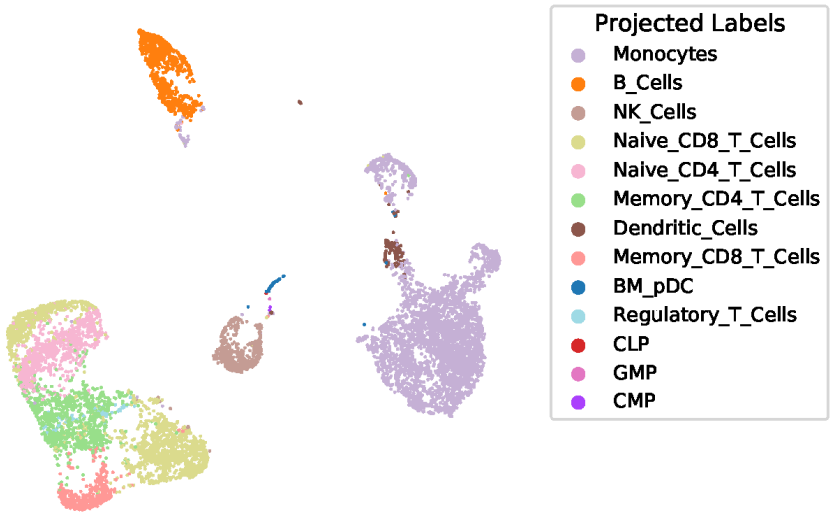


Figure S8: The low-dimensional representation of the Sample 10X PBMC data, with labels transferred from the hematopoiesis data.

3.9 Supplementary Materials

All supplemental materials for this chapter are included in *Chapter3_Additional_Files.zip*. The files are:

- **Supplemental Table 1** Stability analysis results using the full data
- **Supplemental Table 2** Stability analysis results using the data with 50% corruption
- **Supplemental Table 3** Stability analysis results using the data with 90% corruption
- **Supplemental Table 4** Full differential accessibility results

Chapter 4

MultiVI: deep generative model for the integration of multi-modal data

This chapter is currently under review, has been publicly posted on bioRxiv (2021), and is reported here in the most recent version. The authors on the manuscript are:

Tal Ashuach[1,2,*], Mariano I. Gabitto[2,3,6,*], Rohan V. Koodli[2], Michael I. Jordan[2,3], Nir Yosef[1,2,4,5,†]

1. Center for Computational Biology, University of California, Berkeley, CA, USA.
2. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA.
3. Department of Statistics, University of California, Berkeley, Berkeley, CA, USA
4. Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA
5. Chan Zuckerberg BioHub, San Francisco, CA, USA
6. Allen Institute for Brain Science, Seattle, WA, USA

∗ These authors contributed equally to the work.

† Corresponding author. Email: niryosef@berkeley.edu

4.1 Abstract

Jointly profiling the transcriptome, chromatin accessibility, and other molecular properties of single-cells offers a powerful way to study cellular diversity. Here we present MultiVI, a probabilistic model to analyze such multiomic data and leverage it to enhance single modality datasets. MultiVI creates a joint representation that allows an analysis of all modalities included in the multiomic input data, even for cells for which one or more modalities are missing. It is available at `scvi-tools.org`.

4.2 Introduction

The advent of technologies for profiling the transcriptional and chromatin accessibility landscapes at a single cell resolution has been paramount for cataloging cellular types and states, identifying important genomic regions, and linking genes to their regulatory elements [1, 2]. However most uses of single-cell RNA-seq (scRNA-seq) [3, 4] and single-cell ATAC-seq (scATAC-seq) [2, 5] have been limited such that a given cell can only be profiled by one technology. Recently, multi-modal single-cell protocols have emerged for simultaneously profiling gene expression, chromatin accessibility, and more recently, the abundance of surface protein in the same cell [6, 7]. This concomitant measurement enables a more refined categorization of cell states and, ultimately, a better understanding of the mechanisms that underlie their diversity.

The emerging area of multi-modal profiling has benefited greatly from new statistical methods that jointly account for multiple data types, most prominently gene expression and chromatin accessibility, in a range of analysis tasks [8, 9]. Another promising application of multi-modal assays, however, is to improve the way by which the more common and less costly single-modality datasets (e.g., scRNA-seq) are analyzed and interpreted. By leveraging datasets with multi-modal (paired) information, one can infer properties of the missing modalities and thus gain new insight that are otherwise difficult to achieve. To provide a comprehensive solution, such an integrative analysis should be done at two levels. First, it should generate a low-dimensional summary of the state of each cell that reflects all the molecular types in the input data (e.g., gene expression and chromatin accessibility, and if available, also protein expression), regardless of which type of information is available for that particular cell. As commonly done in other applications of single cell genomics, such a representation can facilitate the identification of sub-populations or gradients and enable a more informative data visualization [10]. A second level of analysis should generate a normalized, batch corrected view of each high-dimensional data type (e.g., accessibility of each chromatin region), either observed or inferred. Such an analysis can enable broader identification of molecular features that characterize cellular sub-populations of interest.

Here, we introduce MultiVI, a deep generative model for probabilistic and integrative analysis of multi-modal datasets with single modality datasets. Focusing on gene expression and chromatin accessibility as our main case study, we demonstrate that MultiVI provides

solutions for the two levels of analysis, with a low-dimensional summary of cell state and a normalized high-dimensional view of both modalities (measured or inferred) in each cell. MultiVI was designed to account for the general caveats of single cell genomics data, namely: batch effects, different technologies for the same modality, variability in sequencing depth, limited sensitivity, and noise. It does so while explicitly modeling the statistical properties of each modality, namely discrete signal for scRNA-seq and a largely binary signal for scATAC-seq. A key part in the design of MultiVI is its modularity, which allows for inclusion for additional data modalities. Here, we demonstrate it by adding surface protein expression with tagged antibodies as a third modality [7]. The extended model accounts for key properties of the protein data (e.g., non zero background component), and enables integration and joint analysis with single modality (RNA-, chromatin- or protein- only) datasets.

A recent method (Cobolt; available as a preprint [11]) presented an approach similar to that of MultiVI, with promising results. However, its functionality is limited to handling two modalities (gene expression and chromatin accessibility) and only at the first level of analysis (creation of a joint low-dimensional representation). As we will show, MultiVI provides a more comprehensive solution for integrating and interpreting information across modalities, studies, and technologies. In addition to showcasing its ability to derive accurate low-dimensional representations, we demonstrate several key properties of MultiVI as a way of imputing high-dimensional signals. First, we demonstrate that MultiVI provides calibrated estimates of the uncertainty in the imputed values (e.g., predicted chromatin accessibility for scRNA-seq only cells and predicted gene expression for scATAC-seq only cells), such that less accurate predictions are also less confident. Second, we demonstrate that these estimates of uncertainty give rise to accurate estimates of differential gene expression or chromatin accessibility in cells for which the respective modality was not available. Third, we show that even if a population of cells has information from only one modality, accurate imputation may still be achieved when multi-modal information is available for related populations (thus effectively performing out-of-sample prediction). MultiVI is available in scvi-tools as a continuously supported, open source software, along with detailed documentation and a usage tutorial `https://docs.scvi-tools.org/`.

4.3 Results

The MultiVI Model

MultiVI leverages our previously presented variational autoencoding (VAE [12]) models for gene expression (scVI [13]), chromatin accessibility (PeakVI [14]), and protein abundance (totalVI [15]). For clarity, we focus the discussion here on jointly modeling scRNA- and scATAC-seq data. The extension to surface protein measurements is provided in the Methods section.

Given multi-modal data from a single cell (X) from sample (or batch) S, we divide the observations into gene expression (X_R) and chromatin accessibility (X_A). Two deep neural

networks termed ''encoders" learn modality-specific, batch-corrected multivariate normal distributions that represent the latent state of the cell based on the observed data, $q(z_R|X_R, S)$ and $q(z_A|X_A, S)$, from the expression and accessibility observations, respectively. To achieve a latent space that reflects both modalities, we penalize the model so that the distance between the two latent representations is minimized and then estimate the integrative cell state $q(z|X_R, X_A, S)$ as the average of both representations. The states of cells for which only one modality is available (i.e., unpaired), are drawn directly from the representation for which data is available (i.e., z_R or z_A). This encoding part of the model can be naturally extended for handling other molecular properties (such as protein abundance), by including additional encoder networks.

In the second part of the model, observations are probabilistically generated from the latent representation using modality-specific ''decoder" neural networks. Similar to our previous generative models for gene expression (scVI) and accessibility (PeakVI), the model assumes that the RNA expression data is drawn from a negative binomial distribution, and the accessibility data is modeled with a Bernoulli distribution. The likelihood is computed from both modalities for paired (multi-modal) cells, and only from the respective modality of unpaired cells. Finally, during training, we include an adversarial component which penalizes the model if cells from different modalities are overly separated in latent space. The same framework readily extends for handling protein abundance or other types of information, by including an additional decoder network, along with an appropriate model for estimating likelihood (see Methods).

This two-part architecture enables MultiVI to achieve several goals. First, it leverages the paired data to learn a low-dimensional representation of cell state, which reflects both data types. Second, it allows cells for which only one modality is available to be represented at the same (joint) latent space. Finally, the ''decoding" part of the model provides a way to derive normalized, batch-corrected gene expression and accessibility values for both the multi-modal cells (i.e., normalizing the observed data) and for unpaired cells (i.e., imputing unobserved data; see Supplemental Figure 1 and methods).

MultiVI integrates paired and unpaired samples

To study how well MultiVI integrates paired and single-modality data into a common low-dimensional representation, we inspected the outcome of artificially unpairing a multi-modal (ATAC, and RNA-seq) dataset. Using a peripheral blood mono-nuclear cells (PBMC) dataset from 10X genomics, a randomly selected set of cells (with sizes ranging between 1% to 99% of all cells), are made unpaired such that each cell in the set appears twice: once with only gene expression data, and once with only chromatin accessibility data. This action resulted in a heterogeneous dataset containing three sets of ''cells": one set has both modalities available, a second set has only RNA-seq information, and the third set of cells has only ATAC-seq information present.

Next, we compared MultiVI to Cobolt ([11]), a model similar to MultIVI that uses products of experts to create a common latent space. To explore the performance of additional

analysis strategies and since to the best of our knowledge there are no published methods for integrating multi-modal data with single- modality data, we also added several adaptations of Seurat [16]. Specifically, we attempted to use the Seurat V4 code base with three different approaches: (1) *gene activity*: we converted the ATAC-seq data of the accessibility-only cells to gene activity scores (using the *signac* procedure), and then integrated all the cells using the gene- level data (i.e., gene scores when RNA-seq is not available, or gene expression when RNA-seq is available); (2) *imputed*: we followed the steps in (1) and then used Seurat to impute the RNA expression values for the accessibility-only cells. This is done by averaging over nearby cells in the integrated space for which RNA-seq is available (methods). The data from the accessibility-only cells was then re-integrated with the remaining cells using the imputed RNA expression values instead of the gene scores; (3) *WNN*: using weighted nearest neighbor graphs, which leverages information from both modalities to create a joint representational space, then project single-modality data onto this space (methods).

We ran all methods on the artificially unpaired datasets and compared their latent representations (with the exception of the WNN-based approach on the 99% unpaired dataset, which failed to produce results due to the low number of paired cells; Figure 1A-C, Supplemental Figure 2). We first quantified the mixing abilities of the different approaches, by calculating the local inverse Simpson's index (LISI) score described by [17]. Briefly, for each unpaired cell the fraction of neighbors among the K nearest single-modality neighbors that are of the same modality (expression or accessibility), for varying values of K, normalized by the overall fraction of that modality. This results in an enrichment score, with a score of one being perfect mixing (Figure 1D). We found that algorithms based on generative modeling (Cobolt, MultiVI) outperform the alternative approaches of gene scoring and WNN in most rates of unpaired cells. Conversely, the Seurat-based imputation approach (unlike the other two Seurat-based approaches) maintains high mixing performance across all levels of unpaired cells. This result is expected, though, since each accessibility-only cell is represented by a gene expression vector that is an average over cells for which RNA-seq is available and that have gene expression profiles that are similar to one another (i.e., a local neighborhood in a transcriptome-based space). It does not, however, indicate whether these representations are accurate.

To explore this, we next examined the accuracy of the inferred latent space. To measure how well each method captures the true biological state of a cell, we took advantage of the ground truth information contained in our artificially unpaired datasets. For the unpaired cases, we have two distinct representations of the same cell: one based solely on the expression profile and the other solely on the chromatin landscape. Ideally, the two representations would be situated closely in the latent representation, as both capture the same biological state. To measure this, we looked at the distances between the two representations of the unpaired cells in the latent space created by each method. To account for the varying scales of different latent spaces, we used the rank distance (the minimal K for which the two representations are within each other's K nearest neighbors, averaged across all cells; methods, Figure 1E). In this experiment, we found that MultiVI and Cobolt maintain the multi-modal mixing accuracy substantially better than the three alternatives, and that all

methods have a deteriorating performance as the level of unpaired cells increases.

Taken together, these results show that the deep generative modeling approach, as taken by MultiVI, efficiently integrates unpaired scRNA and scATAC data while capturing the true biological state of each cell. They also demonstrate that the alternative approaches we implemented with Seurat either mix the modalities less effectively, or mix them well but less accurately.

Integration of Independent Datasets

Our previous analyses rely on artificially unpaired data, where our model benefits from all data fundamentally being generated in a single batch and by a single technology. While allowing for more accurate benchmarks, this does not reflect real-world situations in which it is desired to integrate datasets that were generated at different batches or even different studies, while possibly using different modalities and technologies. We therefore sought to demonstrate MultiVI on a set of real-world data. We collected three distinct datasets of PBMCs: 1) Multi-modal data from the 10X dataset we used previously; 2) ATAC-seq from a subset of Hematopoeisis data generated by Satpathy et al[18], containing multiple batches of PBMCs as well as cell-type specific (FACS-sorted) samples; 3) PBMC data generated by several different technologies for single cell RNA sequencing, taken from a benchmarking study by Ding et al [19]. The datasets were processed to create a set of shared features (genes or genomic regions, when measurements are available), and annotations were collected from both the Satpathy et al and Ding et al datasets and combined into a shared set of cell type labels (methods). The resulting dataset has 47148 (53%) ATAC-only cells from Satpathy et al, 30495 (34%) RNA-only cells from Ding et al, and 12012 (13%) jointly profiled cells from 10X.

To gauge the extent of batch effects in this data, we first ran MultiVI without accounting for the study of origin of each sample or its specific technology (which varies between the RNA-seq samples from Ding et al). In this setting, we found substantial batch effects, both between different samples in the chromatin accessibility data and between technologies in the gene expression data (Supplemental Figure 3). We then reanalyzed the data, this time configuring MultiVI to correct for batch effects and technology-specific effects within each dataset (methods). The resulting, corrected, joint latent space mixes the three datasets well (Figure 1F), while accurately matching labeled populations from both datasets (Figure 1G). MultiVI achieves this while also correcting batch effects within the Satpathy data and technology-specific effects within the Ding data (Figures 1H-I). To better examine the correctness of the integration, we examined the set of labelled cells from the two single-modality datasets (FACS-based labels from Satpathy and manually annotated cells from Ding). For each cell, we examined its 100 nearest neighbors that came from the other modality, and summarized the distribution of labels of those neighbors. We find a clear agreement between the labels of each cell and the labels of its neighbors, with some mixing among related cell types (Supplemental Figure 4). Using a similar set of experiments, we also observed that the low-dimensional representation inferred by Cobolt achieves similar

levels of mixing and accuracy (data not shown). This analysis therefore demonstrates that MultiVI, and, more generally, the deep generative modeling approach, are capable of deriving biologically meaningful low-dimensional representations that effectively integrate data not only from different modalities, but also from different labs and technologies.

Probabilistic Data Imputation with Estimated Uncertainty

The generative nature of MultiVI enables several functionalities for analyzing the data in the full high-dimensional space, performing imputation of missing observations and modalities, estimation of uncertainty, and differential analysis. These functionalities are currently unique to MultiVI, and are not implemented by Cobolt or other generative models. To demonstrate MultiVI's imputation abilities, we resorted to the 10X PBMC dataset where 75% of the cells were artificially unpaired (as in Figure 1). We used MultiVI to infer the values of the missing modality for the unpaired cells and found that for both modalities, the imputation had high correspondence to the observed values (Figures 2A-C). Specifically, we observe Spearman correlation of 0.57 between the imputed expression values and the observed data (taking the raw values, scaled by library size), and an area under the precision-recall curve (PRAUC) of 0.41 for the accessibility data (taking the raw, binary signal). Since the raw data can be significantly affected by low sensitivity, we also calculated the correlation between the imputed values and a smoothed version of the data (obtained with a method different of MultiVI; methods), where the signal is averaged over similar cells (separately for ATAC-seq and RNA-seq), thus mitigating this issue. As expected, we see a higher level of correspondence between the imputed values and this corrected version of the raw data (Spearman correlations 0.8 and 0.86 for accessibility and expression respectively; Supplemental Figure 5A-B).

Next, we focus our analysis on uncertainty estimation for the imputed accessibility values. We measured the uncertainty of the model for each imputed accessibility value by sampling from MultiVI's generative model (methods) and found a strong relationship between the estimated uncertainty and the error at each data point $((\text{imputed} - \text{observed})^2)$, indicating that the model is indeed less certain of predictions that are farther from the unobserved "ground truth" values (Figure 2C). Equivalent analysis for expression imputations is hindered by the high correlation between the average expression and both the measured error and the uncertainty of the imputed results.

Interestingly, we identified a small subset of values (roughly 0.5% of observations) for which we have high confidence imputations that are associated with high error, when comparing to the unobserved raw accessibility data (Figure 2C, green square). In the case of chromatin accessibility, these high-confidence high-error imputations correspond to cases where the model confidently predicts the opposite of the actually observed value (Figure 2D). To investigate the source of these errors, we inspected the same cases when comparing the imputed values to the smoothed accessibility estimates (methods). We found that many of these regions were detected as inaccessible in the raw data, but predicted to be accessible by MultiVI, and vice versa. Interestingly, the smoothed data agrees with the

MultiVI predictions—namely, observations that were predicted as accessible tend to be open in highly-similar cells, and observations that were predicted as inaccessible tend to be close in similar cells (Figure 2E). This indicates that these high-confidence high-error values may correspond to false negatives and false positives in the raw data.

As a specific example of imputation, we highlight the T-cell marker gene CD3G. While the observed expression and the observed accessibility of the region containing the transcription start site (TSS) of the gene show high noise and sparsity, the imputed values are highly consistent and clearly mark the T-cell compartment of the latent space (Figure 2F). Overall, these results show that MultiVI is capable of imputing missing observations, and quantifying the uncertainty for each value, allowing the user to then determine which imputed values are reliable for downstream analyses and which are not.

MultiVI enables data imputation in a three modality setting

We tested the ability of MultiVI to integrate more than two experimental modalities and to impute missing data in this setting. Using a dataset of PBMCs profiled with the new DOGMA-seq protocol [7], we artificially unpaired 75% of the data , such that each cell is represented by three copies with only transcriptional, chromatin accessibility, or protein expression data (generating a dataset in which 8% of the total data is paired and the rest has only one of the three modalities). Next, we turned to evaluate the accuracy of imputing the missing modalities in the unpaired cells (Supplemental Figure 6). Again, we found high correspondence between the imputed and observed gene expression values, independent of the single modality used as input data (Spearman correlation of 0.78; using smoothed observations, as above). We observed a similar outcome in the case of chromatin accessibility, with a Spearman correlation of 0.73 and 0.76 between the smoothed observed values and the imputed ones when only RNA or protein information is available, respectively.

Our model for the protein data was designed to control for the non negligible background component in the signal (which may result of non-specific binding of antibodies). We therefore first calculated the foreground ("denoised") component of all observed protein expression values using the two component (foreground, background) model in TotalVI [15]. Since the protein imputed values in MultiVI are also generated with a similar two component model, we were able to compare the imputed foreground signals to foreground signals that were inferred from the respective hidden observations. We find that the imputed values recapitulate the observed data (with Spearman correlations of 0.53 when only chromatin accessibility data or gene expression data is available).

The inclusion of protein data into our analysis highlights the ability of MultiVI to handle additional data types and leverage them for a joint analysis with measurements of chromatin accessibility and gene expression.

Cross-modal Differential Analyses

Our results thus far demonstrate that MultiVI can be used to accurately impute missing observations of single cells, in situations where the multi-modal and the single-modality data both contain the same cellular subsets. The imputation task becomes more challenging when analyzing a population in which one of the modalities was not observed at all. However, the ability to impute values in this scenario will enable leveraging multi-modal data to analyze a wider variety of single-modality datasets, even if a fully matching multi-modal data is not available.

To explore this, we used the same 10x PBMC multiome dataset, with 75% of cells artificially unpaired, and clustered the latent space to identify distinct cellular populations (Supplemental Figure 7A). We chose the B-cell cluster, which we annotated as such using established markers (e.g., CD19, CD79A). Next, we corrupted the data further by removing all expression information (paired or unpaired) from the B-cell population, thus creating a distinct population for which only accessibility data is available to the model. In a second experiment, we removed all accessibility data from the same compartment to create a dataset for which only expression was observed for B-cells (Supplemental Figure 7B-C). We trained MultiVI separately on each of the two corrupted datasets, and used the model to perform differential analyses, comparing the B-cell population and the remainder of the cells. Specifically, we conducted differential expression analysis with the model trained without B-cell expression data (corrupted dataset 1), and differential accessibility with the model trained without B-cell accessibility data (corrupted dataset 2). Estimates of significance were done with a Bayes factor, as in previous work [13, 14, 20] (Methods). To evaluate the accuracy of this analysis, we used standard differential analyses (not using generative models) on the held-out data to create ground-truth results and compared them to our inferred results (Methods). Considering the first corrupted dataset, although no expression data was observed in the B-cell population, we found high concordance between the observed and predicted log Fold-Change values (Figure 2G, Pearson's correlation 0.57). When examining genes that are preferentially expressed in B cells (observed $logFC > 1$) this became more evident (Pearson's correlation 0.74). Similarly, with the second corrupted dataset, we found high concordance between observed and predicted differences of accessibility (Figure 2H, Pearson Correlation 0.67).

Among the top most differentially expressed genes, we found known B-cell markers, including IGLC3, IGHM, CD79A, and IGHD (Supplementary Table 1). Overall, we identified 1621 significantly differential genes (false discovery rate < 0.05), of which 75% were also identified with the held-out data at a 5% false discovery rate (FDR), thus representing a modest but significant enrichment (odds-ratio 1.22, Hypergeometric test $p < 1.9 \cdot 10^{-35}$; Supplementary Table 1). Increasing the threshold of significance (on the FDR for the standard analysis, and the Bayes Factor for the MultiVI results) increased the overlap between the sets of results, indicating that the results are more consistent for more highly significant genes (Figure 2I). Similarly, we identified 922 differentially accessible regions (FDR[20] 0.05, Supplementary Table 2), of which 86% were also identified with the held-out data at 5%

FDR (odds-ratio 1.57, $p < 1.7 \cdot 10^{-95}$). As in the expression analysis, the overlap between the inferred and observed differential accessibility analyzes increased with the significance thresholds (Figure 2J).

Finally, we predicted the expression of genes identified as preferentially expressed in B cells by the model trained without B-cell expression data. CD79A, which encodes for part of the B-cell receptor complex, and one of the top genes identified by MultiVI, was indeed found to have highly localized predicted expression in the B-cell compartment (Figure 2K, displayed using original UMAP coordinates as in Figure 2F). Another differentially expressed gene, CR2, a membrane protein found on both B cells and T cells, was predicted specifically on the corresponding compartments (Figure 2L).

Taken together, these results demonstrate that MultiVI can be used to impute missing modalities even for populations that were only identified in a single-modality dataset. This unlocks the ability for leveraging multi-modal data to reanalyze existing single-modality datasets and impute the missing modality: chromatin landscape for existing scRNA experiments, and gene expression for existing scATAC experiments, as well as perform differential analyses using these imputed values.

4.4 Discussion

MultiVI is a deep generative model, designed for integrated analysis of multi-modal and single-modality datasets of single cells. We demonstrated that MultiVI is able to handle gene expression, chromatin accessibility and protein abundance data. MultiVI uses jointly profiled data to learn a multi-modal model of data sources and to relate measurements of individual modalities on the same population of cells. The model accounts for various technical sources of noise and can correct additional sources of unwanted variation (e.g., batch effects). MultiVI learns a rich latent representation of the data coalescing information present in each individual data type, which can be used for further single-cell sequencing analysis.

Recent algorithms for the analysis of multi-modal data were developed to process paired datasets, in which both modalities have been profiled at the same cell [9, 8]. These algorithms handle multi-modal data, but lack the ability to integrate single modality datasets into the same analysis. While this task is possible to achieve with the Seurat code base [16], the respective methods we utilized here were not specifically designed to this end, and their performance was not tested for this task. Here, we have shown that use of deep generative modeling, either with MultiVI or the recently presented Cobolt [11] can effectively combine unpaired scRNA and scATAC data with multi-modal single-cell data, generating a meaningful low dimensional representation of the cells' state that captures information about both their transcriptome and epigenome. Importantly, this joint representation is achievable even when the amount of paired data is minimal, thus opening exciting opportunities for future studies in which only a small amount of paired data can be sufficient for deriving a more nuanced interpretation of single modality data. In contrast to Cobolt, we demonstrate

that MultiVI is able to integrate information from additional molecular properties of cells, exemplified through the addition of measurements of surface protein expression.

An additional key capability that is unique to MultiVI is the inference of the actual values of the missing modality. We have demonstrated that we can identify preferential gene expression in sub-populations for which only chromatin accessibility data is available and distinguishing chromatin features for sub-populations for which only gene expression data is available. Here again, we show that missing modalities can be imputed, even in the case in which we reverse the central dogma of molecular biology by imputing chromatin accessibility from protein abundance data. These results open the way for exciting future applications: first, MultiVI and similar methods have the potential to enable a reanalysis of the large compendia of available single-modality datasets (representing the majority of existing data) with relatively small additional paired data, thus potentially leading to more comprehensive characterizations of cell state. Second, it can facilitate cost-effective designs for future studies, in which only a subset of samples need to be profiled with the (more costly) multi-modal protocol.

In summary, MultiVI is able to seamlessly integrate single- and multi-modal data, process information from different labs or technologies, and create a rich joint representation (low and high dimensional) that harnesses all available information. It is implemented in the scvi-tools framework [21], making it easy to configure, train, and use.

4.5 Methods

The MultiVI Model

MultiVI inherits generative models describing chromatin accessibility and transcriptional observations from scVI [13], peakVI [14], and TotalVI [15]. Briefly, Let $X_R \in \mathbb{N}_0^{C \times G}$ be a scRNA-seq genes-by-cell matrix with C cells and G genes, where $x_R^{cg} \in \mathbb{N}_0$ is the number of reads from cell c that map to gene g. Let $X_A \in \mathbb{N}_0^{C \times J}$ be a scATAC-seq region-by-cell matrix with C cells and J regions, where $x_A^{cj} \in \mathbb{N}_0$ is the number of fragments from cell c that map to region j. Let $X_P \in \mathbb{N}_0^{C \times P}$ be a protein-by-cell matrix with C cells and P proteins, where $x_P^{cp} \in \mathbb{N}_0$ is the number of fragments from cell c that map to protein g.

MultiVI models the probability of observing x_{cj} counts in a gene by using a negative binomial distribution,

$$x_R^* \sim \text{NegBin}\left(\ell_c \rho_{cg}, \theta_g\right) \tag{4.1}$$

where ℓ_c is a scaling factor that captures cell-specific biases (e.g library size), $\rho_c g$ is a normalized gene frequency and θ_g models the per gene dispersion. The probability of observing a region as accessible is modeled with a Bernoulli distribution,

$$x_A^* \sim \text{Ber}\left(\ell_c p_{cj} r_j\right) \tag{4.2}$$

where p_{cj} captures the true biological heterogeneity; r_j captures region-specific biases (e.g width, sequence). Lastly, MultiVI models protein expression with a mixture of Negative Binomial distributions that encompass background and foreground protein expression.

$$x_P^* \sim \pi_1 \text{NegBin}\left(\ell_c \beta_{cg}^b, \theta_g^b\right) + (1-\pi_1)\text{NegBin}\left(\ell_c \alpha_{cg}^f \beta_{cg}^b, \theta_g^f\right) \tag{4.3}$$

In this model, π_1 accounts for the mixture proportion, β for the background expression level and $\alpha \geq 0$ is a value that corrects for foreground expression. In the observational models, the scaling factor the region-specific and the per gene dispersion parameters are inferred from data using deep neural networks (this is in contrast to the original implementation of scVI in which library size was modelled using a lognormal distribution).

Next, for each cell, normalized gene frequencies ρ_{cg}, accessibility biological heterogeneity p_{cj}, and background and foreground protein expression α_{cg}^f and β_{cg}^b, are estimated using a latent representation as in VAE [12]. Briefly, each modality is assign their own latent representation, a isotropic multivariate normal distribution $Z_c^A \sim \text{MVN}(0,1)$, $Z_c^R \sim \text{MVN}(0,1)$, and $Z_c^P \sim \text{MVN}(0,1)$. Then, with the purpose of bringing all representations together, they are combined by taking their average (e.g., in the case of two modalities profiled such as ATAC and RNA, we have $Z_c = \frac{Z_c^A + Z_c^R}{2}$). This merged representation is then used to decode all model parameters.

MultiVI Inference Model

We use variational inference [22] to compute posterior estimates of model parameters using the following variational approximation:

$$q(z^R, z^A, z^P, r, \ell, \theta | x_A, x_R, x_P) = q(z^R|X_R)q(z^A|X_A)q(z^P|X_P)\delta_{\ell^*}\delta_{\theta^*}\delta_{r^*}, \delta_{\pi_1^*} \tag{4.4}$$

where the delta distribution δ highlights the fact that parameters are inferred from the data as point estimates. The cell-specific factor ℓ_c is computed from the input data for cell c via a deep neural network $f_\ell : \mathbb{N}_0^K \to [0,1]$. The region-specific factor r_j, since it is optimized across samples, is stored as a K-dimensional tensor, used and optimized directly. In the case of each latent representation, encoders are computed as $h_z^{Transc} : \mathbb{N}_0^K \to \left(\mathbb{R}^D, \mathbb{R}^D\right)$, $h_z^{Chrom} : \mathbb{N}_0^K \to \left(\mathbb{R}^D, \mathbb{R}^D\right)$, $h_z^{Protein} : \mathbb{N}_0^K \to \left(\mathbb{R}^D, \mathbb{R}^D\right)$ where each of them computes the distributional parameters of a D-dimensional multivariate normal random variable: $Z \sim MVN\left(h_z\left(x_c\right)_1, h_z\left(x_c\right)_2\right)$.

Using the variational approximation, the evidence lower bound (ELBO) is computed and optimized with respect to the variational and model parameters using stochastic gradients. To enforce the similarity between chromatin and transcription latent representations, we add to the ELBO a term that penalizes the distance between representations using a symmetric Jeffrey's divergence between distributions $d(Z_c^A, Z_c^R) = \text{symmKL}(q(z_c^A), q(z_c^R)) = \text{KL}(q(z_c^A), q(z_c^R)) + \text{KL}(q(z_c^T), q(z_c^A))$. In the case of three or more distributions, we extend the penalty to match every possible set of distributions (when we include protein data, $d(Z_c^A, Z_c^R, Z_c^P) = \text{symmKL}(q(z_c^R), q(z_c^A) + \text{symmKL}(q(z_c^R), q(z_c^P) + \text{symmKL}(q(z_c^A), q(z_c^P))$.

Modeling differences between MultiVI and Cobolt

While conceptually similar, MultiVI and Cobolt have several key differences in design and implementation choices. MultiVI offers additional functionalities due to its generative model, i.e denoising, imputation, uncertainty estimation, and differential analyses, that are discussed in detail in this manuscript. In addition to those, we detail several other differences between the methods: (a) MultiVI uses a distributional average and penalization to mix the latent representations, compared with the classical product of experts calculation used by Cobolt. (b) the distributional assumptions made by the models are different: MultiVI uses tailored noise models for each modality (negative binomial for expression, Bernoulli for accessibility), and uses a deep neural network for the generative component of the model as well as the inference component. In contrast, Cobolt uses a multinomial likelihood for both modalities and uses a linear transformation as a generative model. (c) MultiVI explicitly avoids overfitting the data, in both the architecture (e.g dropout layers) and training procedure (holding out data to use for early-stop if the model overfits), whereas Cobolt does not contain such guardrails.

Benchmarking and Evaluation

Dataset Preprocessing

The 10x multiomic unsorted PBMC dataset was downloaded from the company website. For artificial unpairing analyses, the processed peak-by-cell matrix was downloaded and filtered to remove features that are detected in fewer than 1% of the cells. For the mixed-source PBMC dataset, the fragment file was downloaded and reprocessed using CellRanger-ARC (v2.0.0) with the Satpathy hg38 peaks. The Satpathy dataset was downloaded from GEO (Accession GSE129785); specifically the processed peak-by-cell matrix and metadata files: *scATAC-Hematopoiesis-All.cell-barcodes.txt.gz*,
scATAC-Hematopoiesis-All.mtx.gz, *scATAC-Hematopoiesis-All.peaks.txt.gz*. We then filtered the data to only include peaks that were detected in at least 0.1% of the data, and lifted those peaks over from the hg19 to the hg38 genome reference using the UCSC liftover utility [23]. The Ding dataset was downloaded from GEO (Accession GSE132044); specifically the pbmc data:
pbmc_hg38_count_matrix.mtx.gz, *pbmc_hg38_cell.tsv.gz*, *pbmc_hg38_gene.tsv.gz*.

Matching cell type annotation was downloaded from SCP (Accession SCP424). After preprocessing, the reanalyzed 10x dataset was combined with both single-modality datasets, and the features were filtered to remove features (either genes or peaks) that were detected in fewer than 1% of the cells.

The DOGMA-seq dataset[7], containing paired scRNA, scATAC, and surface protein abundance observations were downloaded from GEO (Accession GSE156478); specifically the four samples containing all three modalities:
*CD28_CD3_**. The ATAC observations were merged using ArchR [24] using default arguments to produce a unified set of peaks called from all four samples. For model training, we

only used features that were detected in at least 1% of cells. For the analyses included in this manuscript, only cells originating from the *DIG_CTRL* sample were used.

RNA-based Seurat integration

This integration modality disregards multiomic information and only RNA information is considered from multiome cells. Briefly, RNA information is first integrated and then, chromatin accessibility is integrated using gene activity scores (*RNA-based* method) or RNA imputed values (*RNA-based Imputed* method).

In more detail, cells were separated into three different datasets, multiomic cells (using only expression data), rna-only cells and atac-only cells. Seurat objects were created for multiome and rna-only data, and were then normalized, scaled, and the first 50 principal components are calculated. For atac-only cells, a Seurat object was created, gene activity scores were calculated, scaled, and principal components were computed. To integrate the three datasets, integration anchors (using *FindIntegrationAnchors*) were calculated and the data was then integrated (using *IntegrateData*). The *RNA-based* method uses gene activity scores as representative values from the atac-only cells. The *RNA-based Imputed* method includes an additional step in which RNA imputed values are calculated from gene activity scores by running *FindTransferAnchors* and *TransferData*. In this integration method, RNA imputed values are used as representative values from atac-only cells. Finally, integrated data was then scaled and principal components were calculated to generate the final latent space. Across these integration methods, we followed the standard recommended procedure for analyzing data with Seurat given in their tutorials [25].

WNN-based Seurat Integration

This approach aims to leverage information from both modalities (chromatin accessibility and expression values), using the newly described weighted nearest neighbors approach from Seurat V4 [16]. We first created a weighted nearest neighbor graph using multiomic information and then project chromatin and transcriptional information onto this.

We begin by separating cells in unpaired datasets into three different datasets, multiomic cells (with both expression and chromatin data), RNA-only, and ATAC-only. First, multiome latent representation is found by calculating SC transform and principal components on the expression data and latent semantic analysis (TF-IDF decomposition followed by SVD) on the chromatin data. Next, multimodal neighbors and the first 50 supervised PCA are calculated. To merge RNA only and ATAC only data to multiome representation, transfer anchors (FindTransferAnchors) are computed on RNA only data and gene activity scores on ATAC only and each datasets is integrated using IntegrateEmbeddings function. Finally, datasets and dimensionality reductions are merged and umap is visualized using the merged information.

Neighbor Rank Distance Calculation

For artificially unpaired cells, each cell has two unpaired representations in the latent space. Given cell c with representations c_a and c_b, let $S(c_a, K)$ be the set of K nearest neighbors to c_a. We then define $\delta(c_a, c_b)$ as the minimal K for which cell c_b is among the K nearest neighbors of cell c_a: $\min\{k : c_b \in S(c_a, k)\}$.

LISI Score Calculation

Enrichment scores were computed as they were in our previous work [14], and similarly to the LISI scores described in the Harmony paper [17]. Briefly, given a latent representation R, an integer k, and the modality labels (expression, or accessibility) L, we compute $G_{R,k}$ the K-nearest neighbor graph from R with k neighbors. Using $G_{R,k}$, we compute for each cell the proportion of neighbors that share the same modality: $s_i = \frac{1}{k}\sum_{j \in G_{R,k}(i)} \mathbb{1}(L_i = L_j)$. The enrichment score is the average score across all cells, $\bar{s}$, normalized by the expected score for a random sample from the distribution of labels: $E[s] = \sum_{\ell \in \{L\}} p_\ell^2$, with p_ℓ being the proportion of each modality.

Estimating Imputation Uncertainty

We estimate the uncertainty of the model for each imputed value by sampling from the latent space (n=15) and computing the standard deviation of the imputed values for each observation. More consistent predictions correspond to less uncertainty.

KNN-based Estimate of Accessibility

To estimate accessibility without using MultiVI, we computed a lower-dimensional representation of the data using Latent Semantic Analysis (LSA, top 30 components), then for each cell we computed the average accessibility profile of the 50 nearest neighbors in the LSA space. This creates a smooth estimate of accessibility using highly-similar cells, mitigating the effect of false observations.

Expression Smoothing

Expression smoothing was achieved by taking the top 30 principle components of the expression data (computed with PCA), computing the K-nearest neighbors graph (for $K = 50$) and averaging the expression values of the neighbors for each cell (scaled by library size).

Differential Analyses with Held-Out Data

To identify a distinct population of cells, we used the Leiden community detection algorithm[26], then examined the expression levels of known marker genes (CD79A, CD19) to identify the cluster of B-cells. We then unpaired the data within the cluster, once by removing all expression data from the B-cells and once by removing all accessibility data from

the clusters. Since the data was already unpaired, this resulted in several cells with no observations at all, and those were removed from the dataset.

Differential Expression using Held-Out Data

Differential expression was computed in two ways: 1) using the held-out data, values were normalized per-cell by dividing the expression levels by the total number of reads in the cell. log Fold-Change values were then computed by dividing the mean expression values in the two groups. Statistical significance was determined using Wilcoxon rank-sum test. 2) without the held-out data, using MultiVI, in a procedure described by Lopez et al [13] which samples from the latent space and uses the generative model to estimate expression profiles. Statistical significance was then determined using Bayes Factors, as well as an FDR approach described by Lopez et al [20].

Differential Accessibility using held-out data

Differential accessibility was computed equivalently to differential expression. 1) using held-out data, values were normalized using the TF-IDF transformation, differential accessibility was computed by subtracting the mean accessibility in the reference group from the same value in the target group. Statistical significance was determined using Wilcoxon rank-sum test. 2) Without the held-out data, using MultiVI, using the procedures described in our previous work [13, 14, 20].

4.6 Figures

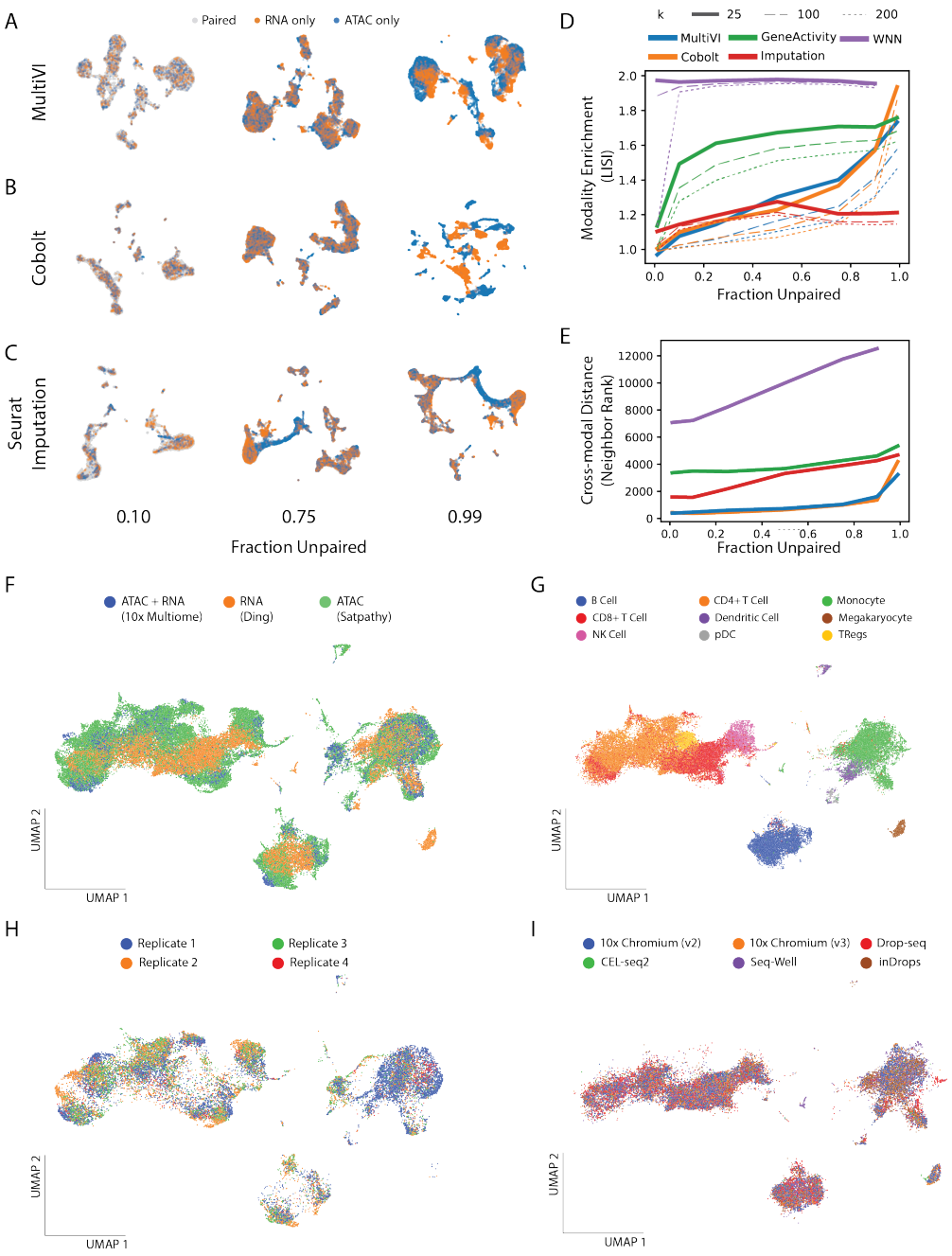


Figure 1

Figure 1: **MultiVI accurately integrates gene expression and chromatin accessibility data. A-C)** UMAP representations of the latent spaces learned by MultiVI **(A)**, Cobolt **(B)**, Seurat using the RNA-imputation based integration **(C)**, for various rates of unpaired data, colored by cell modality. **D)** Modality Enrichment (LISI score), computed as the fraction of neighbors of the K-nearest neighbors that are from the same modality, normalized by the overall fraction of the cells from that modality. **E)** The mean distance between the two representations of artificially unpaired cells, measured as the number of cells between them. **F-I)** UMAP representation computed from the latent space of MultiVI in which cells are color labeled by: **(F)** their modality; **(G)** cell type label; **(H)** scATAC-seq PBMC cells labelled by the replicate from which they were collected; **(I)** scRNA-seq cells labelled by their experimental technology.

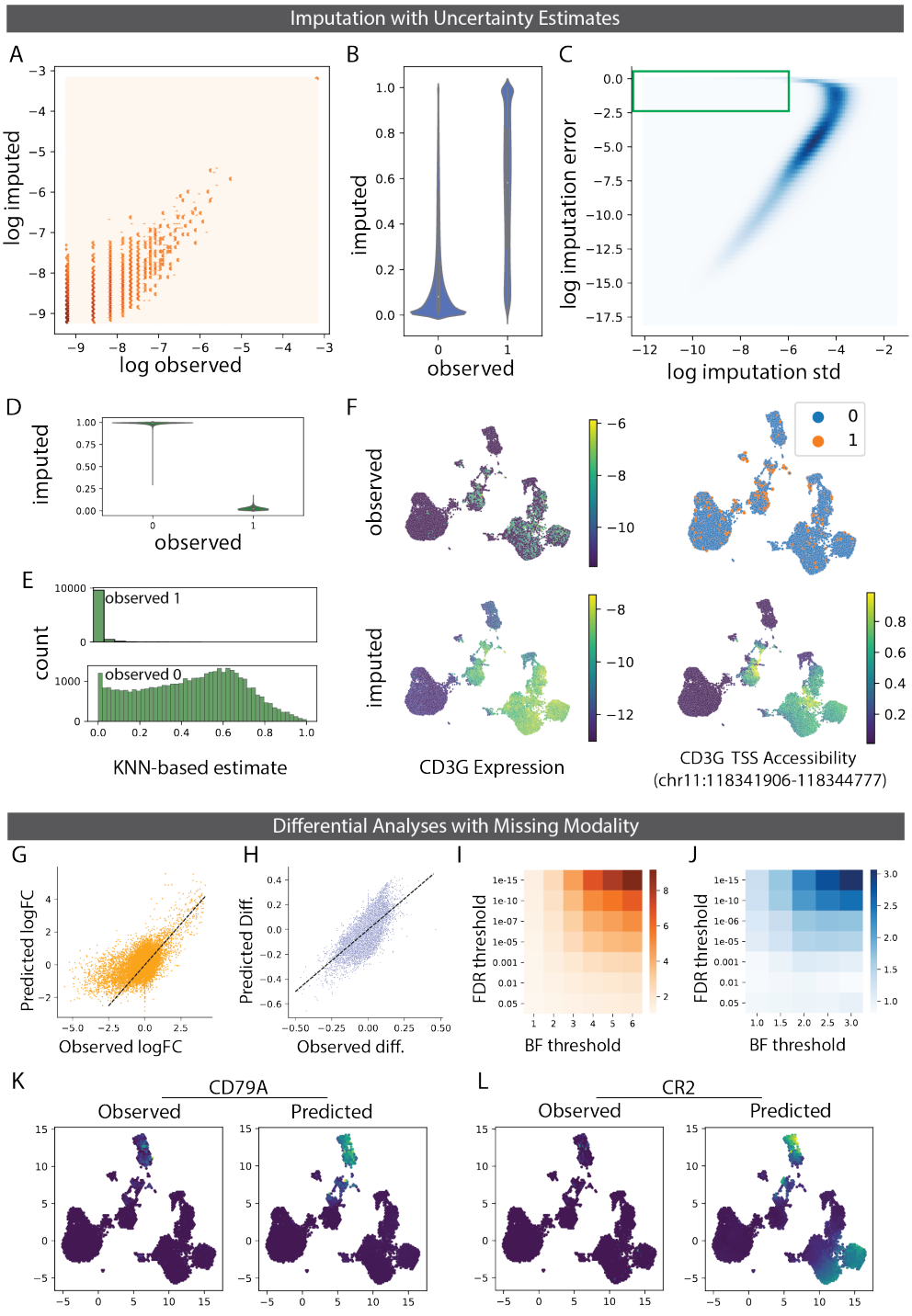

Imputation with Uncertainty Estimates
A
log imputed
log observed
B
imputed
observed
C
log imputation error
log imputation std
D
imputed
observed
E
count
observed 1
observed 0
KNN-based estimate
F
observed
imputed
CD3G Expression
CD3G TSS Accessibility
(chr11:118341906-118344777)
Differential Analyses with Missing Modality
G
Predicted logFC
Observed logFC
H
Predicted Diff.
Observed diff.
I
FDR threshold
BF threshold
J
FDR threshold
BF threshold
K
CD79A
Observed
Predicted
L
CR2
Observed
Predicted

Figure 2

Figure 2: **MultiVI imputation and uncertainty estimation. A)** normalized observed RNA counts by MultiVI-imputed RNA estimates; all values, including color intensity, are presented on a log scale ($\log(x+1e-4)$ for stability). **B)** MultiVI-imputed accessibility estimates by the observed values. **C)** the imputation error $(\text{imputed} - \text{observed})^2$ as a function of the standard deviation of the imputed accessibility estimates. Green box marks high-confidence-high-error values examined in following panels. **D)** MultiVI-imputed accessibility estimates by the observed values for high-confidence-high-error cases. **E)** smooth accessibility estimates for values observed as 1 (top) and 0 (bottom). estimates computed by averaging the accessibility profiles of the 50 nearest neighbors, in a 50-dimensional space computed using Latent Semantic Indexing. **F)** observed and imputed values for CD3G expression and CD3G TSS accessibility. Expression values are normalized per cell and displayed in log scale. **G-H)** Differential effect sizes between B-cells and the remained of the data, comparing the effects computed from the held-out expression data with those predicted by MultiVI, for differential expression **(G)** and differential accessibility **(H)**. **I-J)** Fold-enrichment of the overlap between statistically significant results for various significance thresholds for expression **(I)** and accessibility **(J)**. **K-L)** expression values for B-cell marker CD79A **(K)** and B- and T-cell marker CR2 **(L)**, observed in the held-out data (left) and predicted by MultiVI (right), displayed using latent space coordinated computed using all the available data.

4.7 References

[1] Bosiljka Tasic et al. "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics". en. In: *Nat. Neurosci.* 19.2 (Jan. 2016), pp. 335–346. ISSN: 1097-6256. DOI: 10.1038/nn.4216. URL: https://www.nature.com/articles/nn.4216.

[2] Jason D Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". en. In: *Curr. Protoc. Mol. Biol.* 109 (Jan. 2015), pp. 21.29.1–21.29.9. ISSN: 1934-3639, 1934-3647. DOI: 10.1002/0471142727.mb2129s109. URL: http://dx.doi.org/10.1002/0471142727.mb2129s109.

[3] Fuchou Tang et al. "mRNA-Seq whole-transcriptome analysis of a single cell". en. In: *Nat. Methods* 6.5 (May 2009), pp. 377–382. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1315. URL: http://dx.doi.org/10.1038/nmeth.1315.

[4] Diego Adhemar Jaitin et al. "Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types". en. In: *Science* 343.6172 (Feb. 2014), pp. 776–779. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1247651. URL: https://science.sciencemag.org/content/343/6172/776.

[5] Jason D Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561 (June 2015), pp. 486–490. ISSN: 0028-0836. DOI: 10.1038/nature14590. URL: https://www.nature.com/articles/nature14590.

[6] Junyue Cao et al. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". en. In: *Science* 361.6409 (Sept. 2018), pp. 1380–1385. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aau0730. URL: http://dx.doi.org/10.1126/science.aau0730.

[7] Eleni P Mimitou et al. "Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells". In: *Nature Biotechnology* (2021), pp. 1–13.

[8] Ricard Argelaguet et al. "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". en. In: *Genome Biol.* 21.1 (May 2020), p. 111. ISSN: 1465-6906. DOI: 10.1186/s13059-020-02015-1. URL: http://dx.doi.org/10.1186/s13059-020-02015-1.

[9] Rohit Singh et al. "Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities". en. In: *Genome Biol.* 22.1 (May 2021), p. 131. ISSN: 1465-6906. DOI: 10.1186/s13059-021-02313-2. URL: http://dx.doi.org/10.1186/s13059-021-02313-2.

[10] David DeTomaso et al. "Functional interpretation of single cell similarity maps". en. In: *Nature communications* 10.1 (2019), pp. 1–11.

[11] Boying Gong, Yun Zhou, and Elizabeth Purdom. "Cobolt: Joint analysis of multimodal single-cell sequencing data". In: *bioRxiv* (2021). DOI: 10.1101/2021.04.03.438329. eprint: https://www.biorxiv.org/content/early/2021/04/04/2021.04.03.438329.full.pdf. URL: https://www.biorxiv.org/content/early/2021/04/04/2021.04.03.438329.

[12] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (Dec. 2013). arXiv: 1312.6114v10 [stat.ML]. URL: http://arxiv.org/abs/1312.6114v10.

[13] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". en. In: *Nat. Methods* 15.12 (Dec. 2018), pp. 1053–1058. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: http://dx.doi.org/10.1038/s41592-018-0229-2.

[14] Tal Ashuach et al. "PeakVI: A Deep Generative Model for Single Cell Chromatin Accessibility Analysis". en. Apr. 2021. DOI: 10.1101/2021.04.29.442020. URL: https://www.biorxiv.org/content/10.1101/2021.04.29.442020v1.

[15] Adam Gayoso et al. "Joint probabilistic modeling of single-cell multi-omic data with totalVI". In: *Nature Methods* 18.3 (2021), pp. 272–282.

[16] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". en. In: *Cell* 0.0 (May 2021). ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2021.04.048. URL: http://www.cell.com/article/S0092867421005833/abstract.

[17] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". en. In: *Nat. Methods* 16.12 (Nov. 2019), pp. 1289–1296. ISSN: 1548-7091. DOI: 10.1038/s41592-019-0619-0. URL: https://www.nature.com/articles/s41592-019-0619-0.

[18] Ansuman T Satpathy et al. "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion". en. In: *Nat. Biotechnol.* 37.8 (Aug. 2019), pp. 925–936. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-019-0206-z. URL: http://dx.doi.org/10.1038/s41587-019-0206-z.

[19] Jiarui Ding et al. "Systematic comparison of single-cell and single-nucleus RNA-sequencing methods". en. In: *Nat. Biotechnol.* 38.6 (June 2020), pp. 737–746. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-020-0465-8. URL: http://dx.doi.org/10.1038/s41587-020-0465-8.

[20] Romain Lopez et al. "Decision-Making with Auto-Encoding Variational Bayes". In: (Feb. 2020). arXiv: 2002.07217 [stat.ML]. URL: http://arxiv.org/abs/2002.07217.

[21] Adam Gayoso et al. "scvi-tools: a library for deep probabilistic analysis of single-cell omics data". en. Apr. 2021. DOI: 10.1101/2021.04.28.441833. URL: https://www.biorxiv.org/content/10.1101/2021.04.28.441833v1.

[22] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877. DOI: 10.1080/01621459.2017.1285773. eprint: https://doi.org/10.1080/01621459.2017.1285773. URL: https://doi.org/10.1080/01621459.2017.1285773.

[23] *LiftOver Utility*. URL: https://genome.ucsc.edu/cgi-bin/hgLiftOver.

[24] Jeffrey M Granja et al. "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis". en. In: *Nat. Genet.* 53.3 (Mar. 2021), pp. 403–411. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-021-00790-6. URL: http://dx.doi.org/10.1038/s41588-021-00790-6.

[25] *Integrating scrna-seq and scatac-seq data*. URL: https://satijalab.org/seurat/articles/atacseq%5C%5Fintegration%5C%5Fvignette.html.

[26] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific reports* 9.1 (2019), pp. 1–12.

4.8 Supplementary Figures

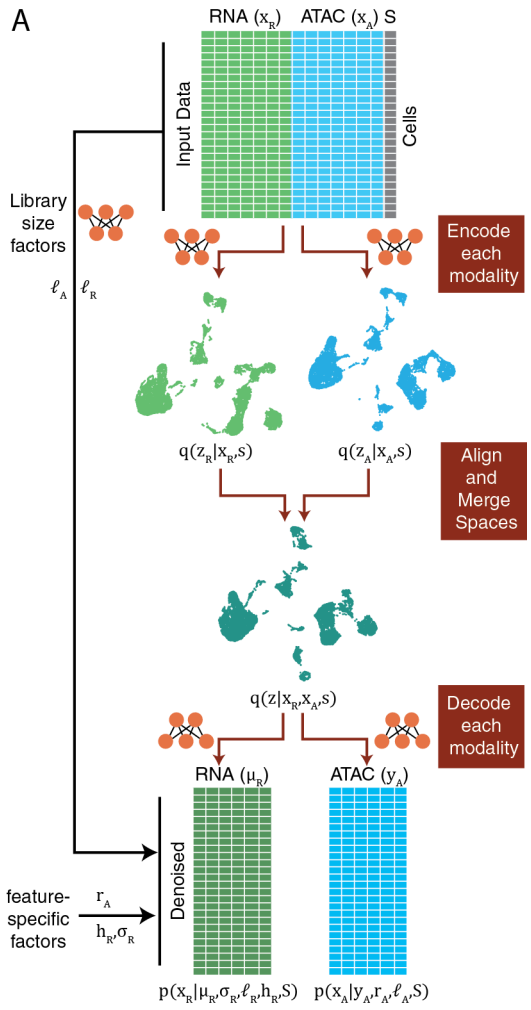


Figure S1: MultiVI Model Overview. Conceptual model illustration in which input data (top) consists of either chromatin accessibility (ATAC), gene expression (RNA) or both data types (Multiome). Variable S represents experimental covariates, such as batch or experimental condition. Each data modality is encoded into modality-independent latent representations (using neural network encoders) and then, these representations are merged into a joint latent space. The joint latent representation is used to estimate (decode) the input data together with chromatin region-specific effects (r_A), gene-specific dispersion (σ_R), cell-specific effects (ℓ_A, ℓ_R), accessibility probability estimates (Y_Z) and mean gene expression values (μ_R).

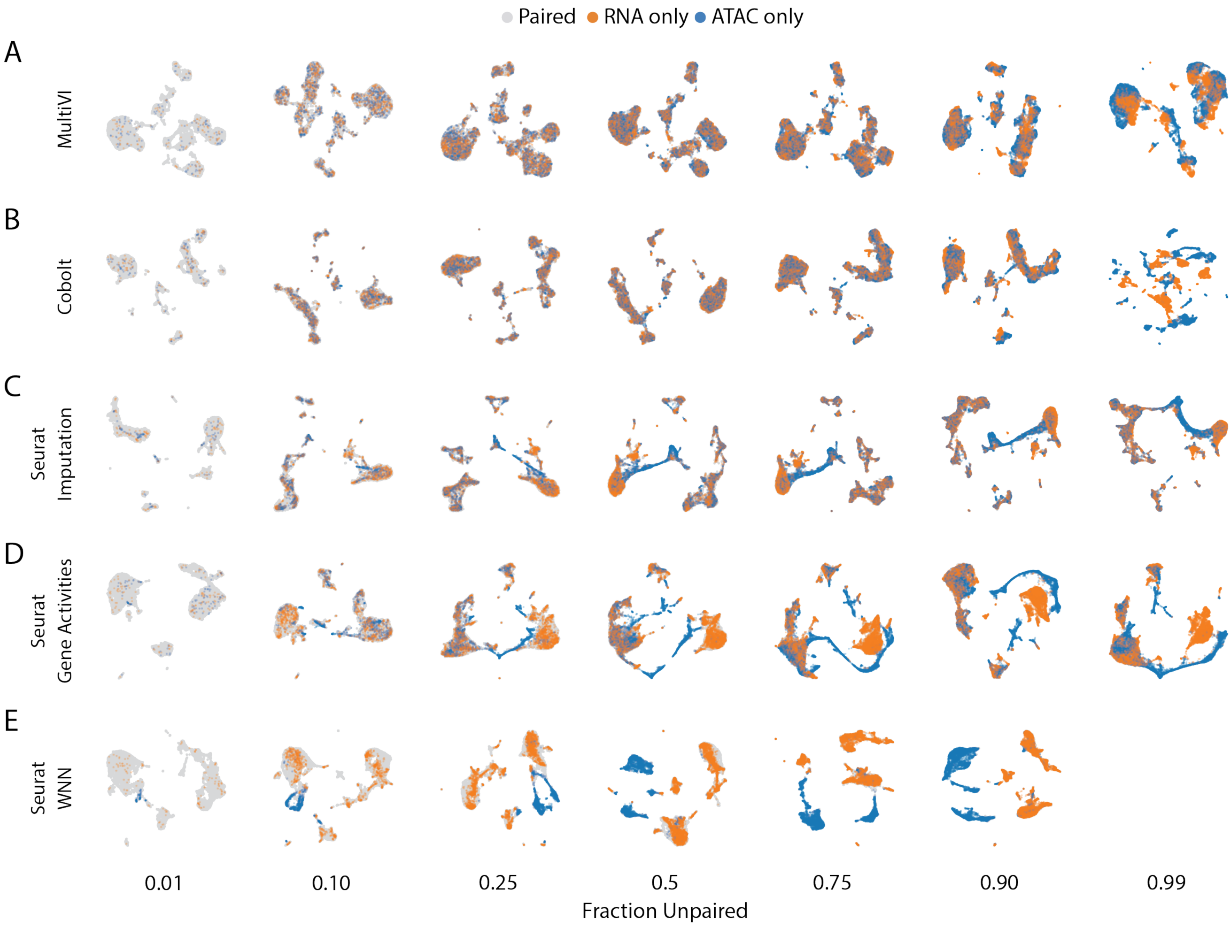


Figure S2: **Extended Integration results depicting mixing of cells in data sets with different fraction of cells unpaired.** UMAPS of latent representations for MultiVI **(A)**, Seurat imputation method **(B)**, Seurat Gene Activity Scores method **(C)**, and Seurat wKNN method **(D)**.

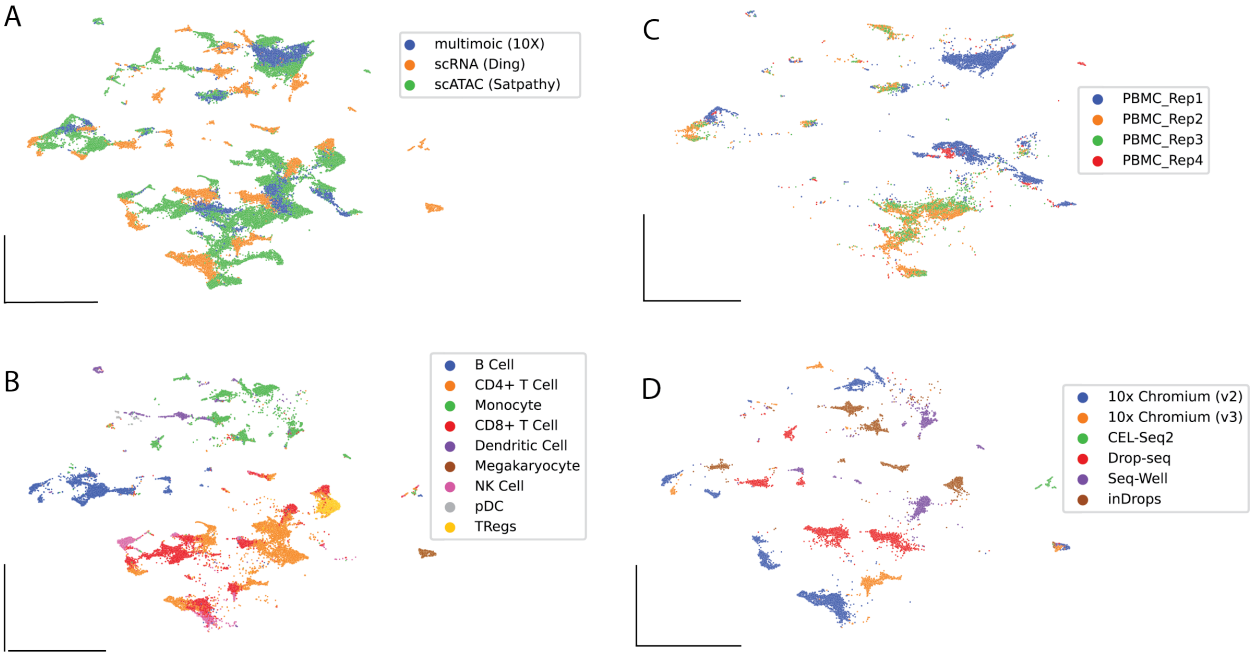


Figure S3: **Latent representation of mixed sources data sets in which no batch correction techniques have been applied.** We integrated three PBMC datasets in which only multi-modal data (10X multiome), only ATAC-seq information (Satpathy et al [18]) and only RNA-seq information (Ding et al [19]) is present without correcting for batch or modalities effects. **A)-D)** UMAP representation computed from the latent space of MultiVI in which cells are color labeled by their dataset **(A)**, their cell type **(B)** or ATAC-seq cells are labelled by the replicate in which they were collected **(C)** or RNA-seq cells are labelled by their collection experimental technology.

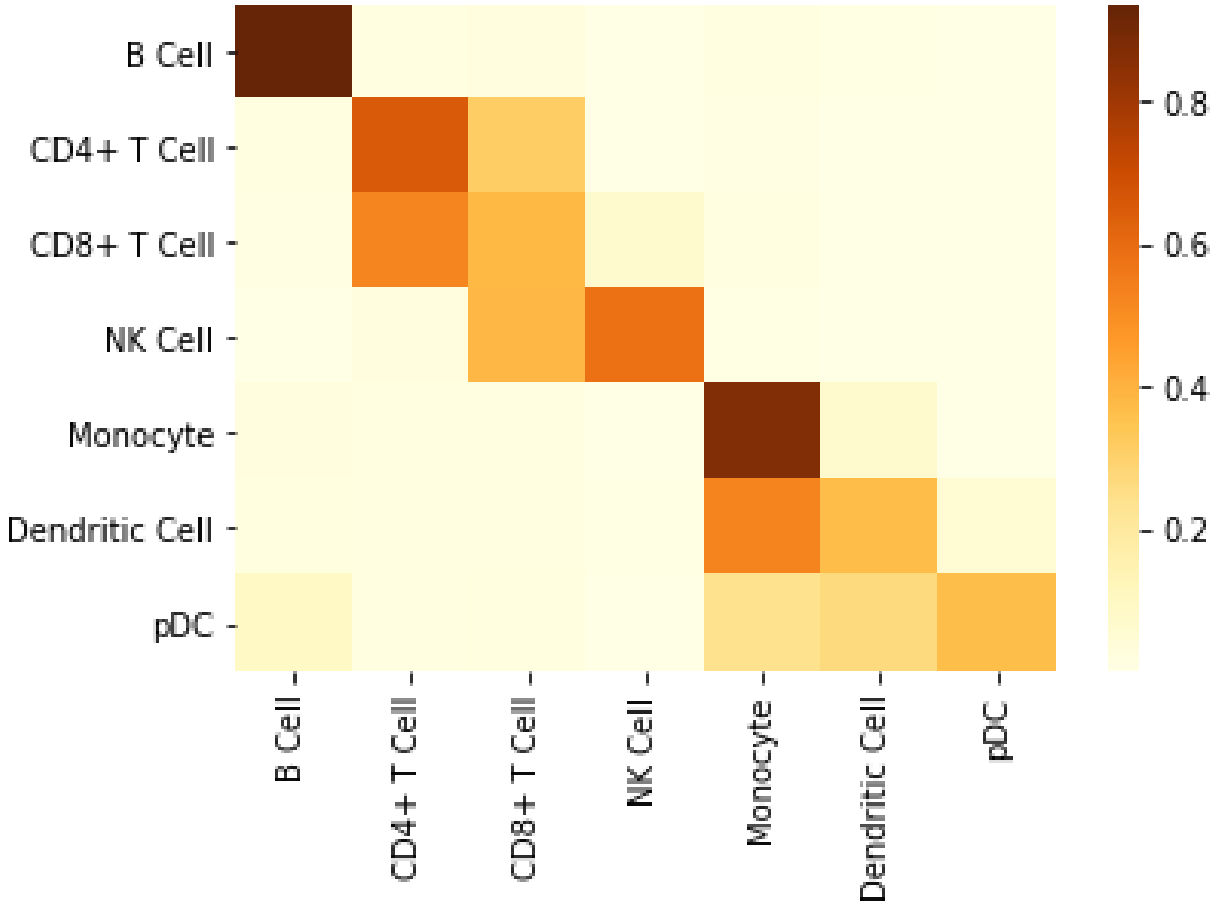


Figure S4: The distribution of labels of neighboring cells by the label of origin

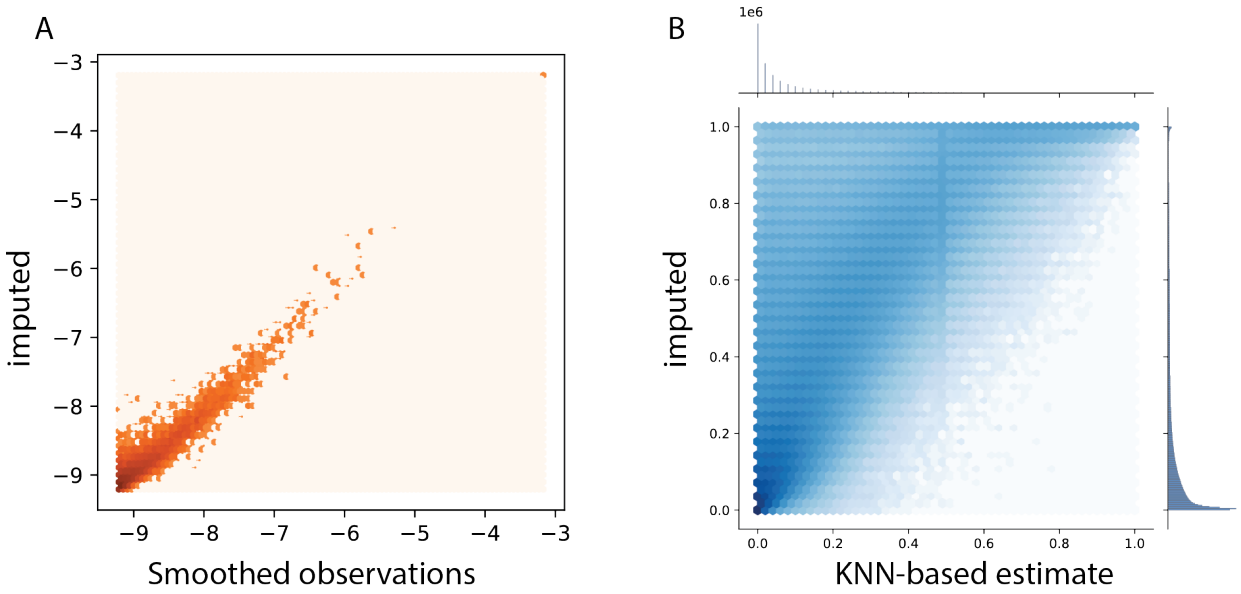


Figure S5: **Imputed values compared with smoothed observations.** Smooth averages of highly-similar cells (using 50 nearest neighbors in an independent low-dimensional space, computed separately for RNA and ATAC data) plotted against the MultiVI-imputed values for expression **(A)** and accessibility **(B)**.

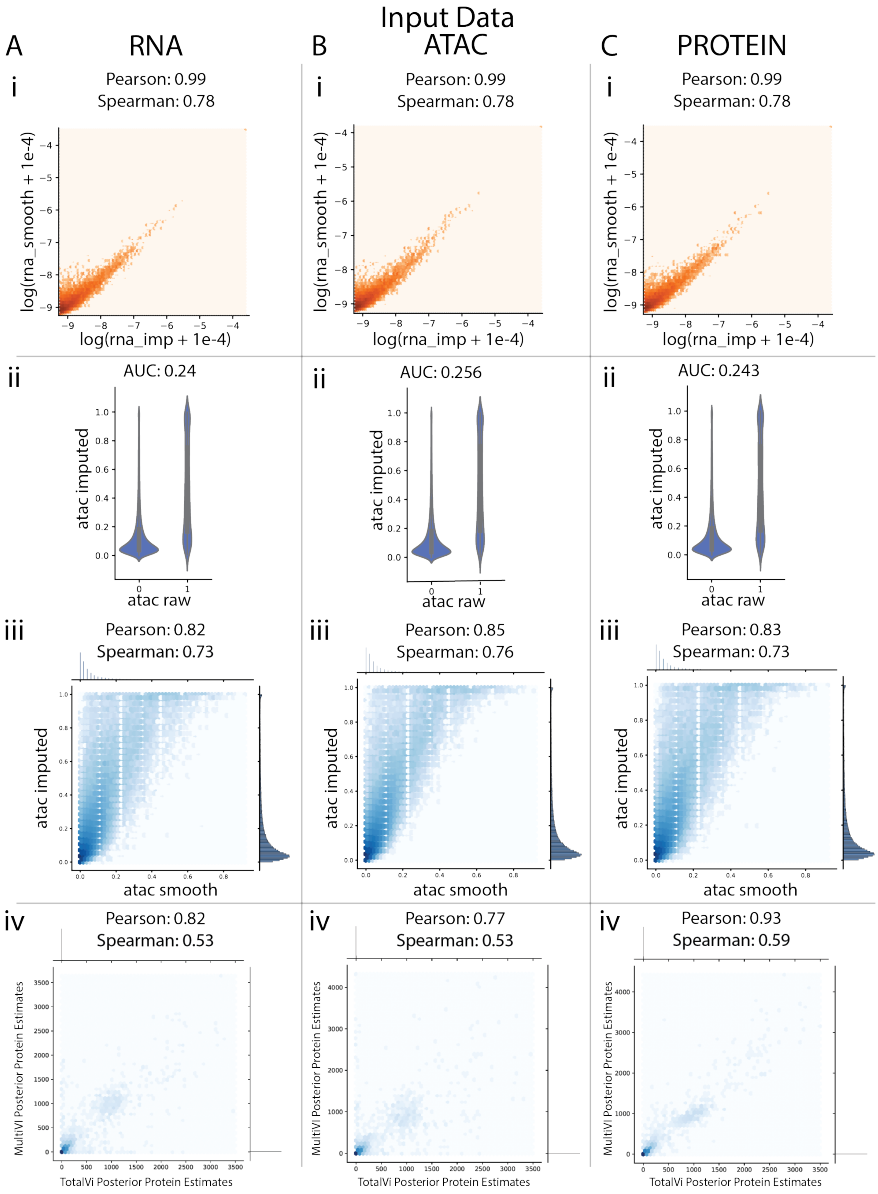


Figure S6: **MultiVI integrates transcriptional, chromatin accessibility and protein expression information, and imputes data missing in single modality data sets, generating uncertainty estimation.** MultiVI was trained using a DOGMA-seq data set in which information for RNA-seq, ATAC-seq, and CITE-seq is present for 8.3% of the cells, and only single-modality information is present for remainder (1/3 only RNA, 1/3 only ATAC, 1/3 only protein). Imputation of missing data in single modality cells is organized in three columns representing input data; i.e., RNA **A)**, Chromatin Accessibility **B)**, and Protein Expression **C)**. For each of the corresponding input modalities, we impute the values of **i)** normalized RNA expression and compare it against smooth estimates of RNA expression (presented on a log scale ($\log(x+1e-4)$ for stability); **ii)** impute accessibility estimates and visualize them by their observed values; **iii)** impute accessibility estimates and visualize them against smooth accessibility estimates computed by averaging cells using Latent Semantic Indexing; **iv)** impute normalized foreground protein expression estimates and compared them against estimates computed using TOTALVI in which RNA and protein data is used as input to the model.

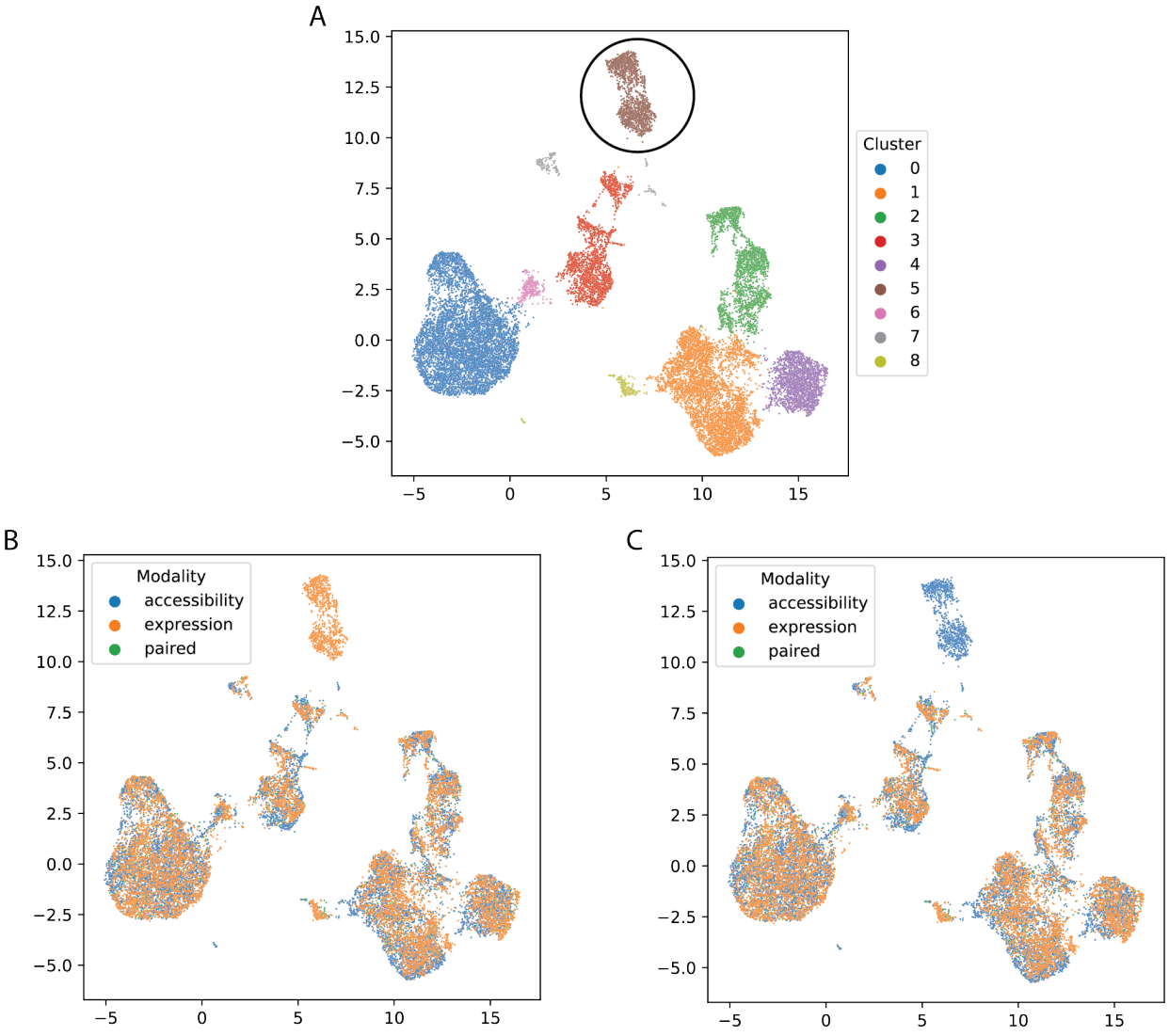


Figure S7: **UMAP visualizations of the 10x PBMC multiome dataset, with 75% of cells artificially corrupted. A)** Leiden clustering of the cells. **B-C)** modalities of the different cells after removal of all accessibility **(B)** or expression **(C)** data from the B-cell compartment (cluster 5).

4.9 Supplementary Materials

All supplemental information for this chapter is included in *Chapter4_Additional_Files.zip*. The files are:

- **Supplemental Table 1** Full differential expression results, related to Figure 2I.
- **Supplemental Table 2** Full differential accessibility results, related to Figure 2J

Part II

High-throughput Functional Characterization of Enhancer Sequences

Chapter 5

MPRAnalyze - A statistical framework for Massively Parallel Reporter Assays

This chapter was published in *Genome Biology* (2019), and is included here as published. The authors on the paper are:

Tal Ashuach[1,2,*], David Sebastian Fischer[3,4,*], Anat Kreimer[1,5,6], Nadav Ahituv[5,6], Nir Yosef[1,2,7,8,†]

1. Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California USA.
2. Center for Computational Biology, University of California, Berkeley, California USA.
3. Institute of Computational Biology, Helmholz Zentrum Munchen, 85764 Neuherberg, Germany
4. TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany
5. Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA
6. Institute for Human Genetics, University of California San Francisco, San Francisco, California, 94158, USA
7. Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA
8. Chan Zuckerberg BioHub, San Francisco, CA, USA

∗ These authors contributed equally to the work.

† Corresponding author. Email: niryosef@berkeley.edu

5.1 Abstract

Massively parallel reporter assays (MPRAs) can measure the regulatory function of thousands of DNA sequences in a single experiment. Despite growing popularity, MPRA studies are limited by a lack of a unified framework for analyzing the resulting data. Here we present MPRAnalyze: a statistical framework for analyzing MPRA count data. Our model leverages the unique structure of MPRA data to quantify the function of regulatory sequences, compare sequences' activity across different conditions, and provide necessary flexibility in an evolving field. We demonstrate the accuracy and applicability of MPRAnalyze on simulated and published data and compare it with existing methods.

5.2 Introduction

Understanding the function of the non-coding genome poses one of the most significant and outstanding challenges following the completion of the human genome project [1]. One critical function that is primarily associated with non-coding regions is to regulate the transcription of nearby genes by interaction with transcription factors and other proteins and through recruitment of the RNA polymerase complex [2, 3]. Two of the main classes of regulatory regions consist of promoters (which are proximal to the transcription start site of the respective gene) and enhancers (distal elements), both demonstrated to harbor many disease-related mutations [4, 5]. The delineation of these critical regions on a genome-wide scale has traditionally relied on chromatin-associated features that are indicative of regulatory activity, such as acetylation or methylation of certain residues along histone tails [1]. However, this approach does not provide direct evidence for regulatory activity, nor the dependence of this putative activity on the cellular context or on the presence of mutations.

Recent advances in reporter assays address this issue in a set of procedures dubbed Massively Parallel Reporter Assays (MPRAs) [6, 7]. In these assays, a synthetic DNA construct that contains a minimal transcriptional unit is introduced into cells. Each such construct is generally composed of a candidate regulatory sequence of interest, a minimal promoter, and a unique DNA "barcode" that can be transcribed. The candidate sequences are assumed to be capable of regulating the transcription of the barcode sequence similarly to how a native sequence may regulate the transcription of its target gene. The cells then undergo RNA and DNA sequencing to measure both RNA transcript counts and DNA construct counts, and the RNA-to-DNA ratio is used to estimate the transcription rate of every barcode. Relying on sequence-based reporters leverages the vast combinatorial space of unique sequences (instead of a limited set of fluorescent reporters [8]), and utilizes next generation sequencing to measure the activity of thousands of putative regulatory sequences in a single experiment. To ensure robustness, each candidate regulatory sequence is usually associated with several barcodes (< 10 to over 100, depending on the study).

MPRAs can be used to address several important questions. In classification studies, MPRAs are used to identify which putative regulatory regions are indeed inducing tran-

scription (albeit in a synthetic context) [9, 10]. In allelic comparison studies, MPRAs are used to quantify the effect that variations to the sequence of regulatory regions may have on their ability to regulate transcription. This approach is primarily utilized for studying the effect of genetic polymorphisms that are observed in humans [11, 12, 13], but also to explore more basic science questions such as the effect of perturbing the sequence content, spacing or number of transcription factor binding sites [14, 15, 13]. In comparative studies, MPRAs are used to quantify the dependence between the regulatory activity of each sequence and the cellular context, comparing tissues [16], cell lines [11], or other conditions of interest [17]. A combination of two or more study types is also possible through more complex experimental designs, for example measuring the interaction between alleles and conditions [12].

Despite growing popularity of MPRAs, most studies to date rely on analysis approaches that either discount the inherent noise in the data (e.g., taking an average ratio across all barcodes) or designed for other data modalities (such as DESeq2 [18], typically used for RNA-seq data, whose underlying assumptions may not hold true for MPRA). Other MPRA analysis methods only address some of the types of questions MPRAs can address, such as QuASAR-MPRA [19] and mpralm [20] that only perform comparative analyses, and rely on ratio-based summary statistics that limits the statistical power provided in these experiments. To address this, we have developed MPRAnalyze - a statistical framework that leverages information from multiple barcodes to ensure robust analysis of MPRA data. In the following, we demonstrate the use of MPRAnalyze for the three primary analysis tasks listed above, and compare its performance to the existing approaches using a collection of published datasets. MPRAnalyze is available as an R package through Bioconductor [21].

5.3 Results

MPRA data is produced from two parallel procedures: RNA-sequencing is used to measure the number of transcripts produced from each barcode, and DNA-sequencing is used to measure the number of construct copies of each barcode. Thus for each barcode the ratio of RNA to DNA can serve as a conceptual proxy for the transcription rate [7]. However, both DNA and RNA measurements procedures provide an approximate and noisy estimation, an issue exacerbated by the unstable nature of a ratio: minor differences in the counts themselves can result in major shifts in the ratio, especially when dealing with small numbers. This problem can be handled by associating multiple barcodes with each sequence, providing multiple replicates within a single experiment and a single sequencing library. This approach introduces an additional problem of summarizing counts from multiple barcodes to get a single transcription rate estimate for a candidate regulatory sequence, which is made difficult since the efficiency of incorporation inside cells, while theoretically uniform across the different constructs, has a significant degree of variability in practice (Figure 1A). Two commonly used techniques of addressing this issue are based on summary statistics: the aggregated ratio, which is the ratio of the sum of RNA counts across barcodes divided by the sum of DNA counts across barcodes $\left(\frac{\frac{1}{n}\sum_i^n RNA_i}{\frac{1}{m}\sum_j^m DNA_j}\right)$; and the mean ratio, which is the

mean of the observed RNA/DNA ratios across barcodes $\left(\frac{1}{n}\sum_i^n \frac{DNA_i}{RNA_i}\right)$. Although intuitive, both summary statistics have inherent limitations. The aggregated ratio loses the statistical power that multiple barcodes provide and is often dominated by a minority of barcodes with high counts, and the mean ratio is highly sensitive to noise, as recently demonstrated in a paper by Myint and colleagues [20]. A method to leverage the multiplicity of barcodes in a robust manner is therefore needed to fully fulfill the potential of these assays.

MPRAnalyze Model

We introduce MPRAnalyze, a method for the analysis of MPRA data that uses a graphical model to relate the DNA and RNA counts, account for the uncertainty in both libraries and leverage the unique structure and opportunities presented by MPRA data. Out model relies on the assumption of a linear relationship between the RNA counts and the corresponding DNA counts: $RNA = DNA \times \alpha$, similar to ratio-based approaches, with α denoting the transcription rate. Our framework comprises two nested models: the DNA model, which estimates the latent construct counts for the observed DNA counts; and the RNA model, which uses the construct count estimates from the DNA model and the observed RNA counts to estimates the rate of transcription, α (Figure 1B).

For each candidate regulatory sequence, the model requires two vectors of observations: DNA counts $\vec{d}$ and RNA counts $\vec{r}$, where each observation is the number of times a specific barcode, associated with this sequence, was observed at the DNA and RNA levels respectively. Additionally, we denote $\vec{\hat{d}}$ the vector of latent construct counts (DNA) and $\vec{\hat{r}}$ the vector of latent transcript counts (RNA). We assume that the latent construct counts, from which the observed DNA counts are sampled, are generated by a gamma distribution. Second, we assume that the conditional distribution of the RNA counts follows a Poisson distribution. Formally:

$$\vec{\hat{d}} \sim Gamma\left(k, b\right) \tag{5.1}$$

$$\vec{\hat{r}} | \vec{\hat{d}} \sim Poisson\left(\alpha \vec{\hat{d}}\right) \tag{5.2}$$

These result in a closed-form negative binomial likelihood for the RNA counts:

$$\vec{\hat{r}} \sim NB\left(\mu = \frac{\alpha \cdot k}{\beta},\ \psi = k\right) \tag{5.3}$$

The negative binomial distribution is a common approximation of sequencing data due to the observed over-dispersion [22], and indeed all datasets we examined have a quadratic relationship between the mean and the variance, which can be captured by a negative binomial. This relationship is also observed for the DNA libraries, which is expected of Gamma-distributed data if the distribution's shape parameter $k \approx 1$ (Supplementary Figures S1, S3; Methods).

Now, assume we have two conditions. In this case, each barcode is measured twice (once in each condition), and the model needs to relate these observations and account for potential

differences between them. MPRAnalyze achieves this by assuming that the effects are log-additive, and replacing the simple components of the DNA estimate ($\vec{\hat{d}}$) and the transcription rate estimate (α) with generalized linear models (GLM) that enable easy encoding of various relationships between experimental factors. The model then becomes:

$$\log\left(\vec{d}\right) = X_D\vec{\beta} + \log\left(\vec{S_D}\right) \tag{5.4}$$

$$\log\left(\vec{r}\right) = X_D\vec{\beta} + X_R\vec{\gamma} + \log\left(\vec{S_R}\right) \tag{5.5}$$

Here, S_D, S_R are external correction factors, used to account for various technical effects such as library size in the DNA and RNA data respectively. X_D, X_R are design matrices for the DNA and RNA models, which encode the experimental setup of the assay. For instance, in a two condition settings, each matrix will include a column with a 0/1 indicator corresponding to the first or second condition respectively. The respective coefficients β and γ will then capture the effect associated with the choice of condition. Notably, the DNA design matrix X_D will also usually encode the identity of the barcode, so as to enable per-barcode estimation of construct abundance. This is not necessary for the RNA design matrix X_R since we assume that the barcodes are replicates that should have a single estimate of the transcription rate. An illustrative example is provided in Figure S2 (Additional File 1) and a formal description of the model is provided in additional File 2.

The model can be further extended to encode multiple covariates, both quantitative and qualitative, and thus support the common structure of MPRA experiments, namely multiple barcodes per sequence, multiple replicates or batches, and multiple conditions analyzed simultaneously. An important aspect of this flexibility is that it supports "un-paired" datasets in which the DNA sequencing was performed on the pool of constructs, prior to incorporation into cells [12, 11, 13, 10]. In these cases there might not be separate DNA estimates for each condition being tested, in which case the conditions of interest would only be modelled in the RNA design matrix and excluded from the DNA model.

In summary, MPRAnalyze utilizes a model that accounts for barcode specific effects and leverages them for increased statistical power and robustness of estimation. Since a standard for MPRA experimental design has yet to be formed, the nested GLM construction provides flexibility and is easily adjustable to changing experimental designs. Our model is also highly interpretable, allowing for quantitative estimates of sequence activity to be easily extracted, as well as differential activity to be tested directly using established statistical tests. This framework can explicitly leverage negative controls (sequences with no expected regulatory function) when available, either to establish the null distribution in classification analyses or to correct for systemic bias in comparative analyses (Methods).

Benchmark Datasets

In the following sections, we investigate the performance of MPRAnalyze in quantifying the transcriptional activity of candidate regions, as well as in the three major analysis tasks,

namely - classification, cross-condition analysis, and allelic comparisons. Finally, we evaluate MPRAnalyze in a complex setup where we investigate both multiple conditions and multiple alleles. We compare MPRAnalyze to the current set of tools and analysis methodologies, using simulated data and a collection of public data sets. These datasets were chosen for representing a diversity of MPRA protocols (e.g., episomal or lentiviral integration, DNA sequencing pre- or post-transduction), study focus (classification, comparative analyses, allelic comparisons), and experimental design (number of barcodes per sequence, number of replicates). A summary of the data sets and their properties is provided in Table 1. Applying MPRAnalyze to these data, we found that the model is able to provide a good fit ($R^2 > 0.86$ for all datasets, figure 1C), which is consistent with our distributional assumptions (Supplementary Figure S3).

Quantification

We set out to examine the properties of the estimate of transcription rate generated by MPRAnalyze, denoted α (alpha), and compare it to the ratio-based summary statistics (i.e., mean of RNA-to-DNA ratios across all barcodes, or alternatively, the ratio of means [henceforth referred to as the *aggregate ratio*]).

Reassuringly, the three estimates are largely in agreement (Pearson's $r > 0.9$ across datasets, Supplementary Figure S4). To further examine the accuracy of the estimates, we used the negative control sequences included in some of the datasets. These are assumed to have an identical transcription rate induced by the minimal promoter included in each construct with no sequence-induced activity. We examined the variance of the estimates on these sets. In the Kwasnieski dataset, the limited number of barcodes ($n = 4$) is mitigated by high counts per barcode (Figure 1A), leading to all estimates having similarly low variance. In the barcode-rich datasets ($n \geq 90$), the mean ratio is expectedly [20] the most variable, with α being the most consistent in the Inoue-Kircher datasets and comparably consistent to the aggregated ratio in the Inoue-Kreimer dataset (Figure 2A). These results suggest that MPRAnalyze is estimating similar transcription rates across the negative controls, as expected from this collection.

We then explored the effect of the number of barcodes on the estimates' performance. Using the barcode-rich datasets, barcodes were sampled at various rates and estimates were recomputed for each sequence (3 independent samples per sequence per barcode-rate). Using the full-data estimates as the ground truth, we found that down-sampling barcodes does not result in a systemic bias in any of the estimates (Figure 2B), and all estimates showed reduced variance with increased barcodes, with the mean ratio under-performing the other two estimates, and α having a similar or lower variance than the aggregated ratio (Figure 2C).

In many cases the goal of quantifying sequence activity is to rank and compare different sequences, as in mutagenesis experiments. To compare the stability of the ordering of sequences, the Spearman correlation was computed between the estimates in each sub-sample

to the estimates of the full data. Alpha has either similar or higher correlation than both ratio-based estimates across datasets and barcode abundance (Figure 2D).

Since these analyses are limited by a lack of ground truth, MPRA data was then simulated by generating random coefficients and using the same nested GLM construction as described above to generate samples. To avoid biasing the results, samples were generated with a log-normal noise model instead of the default Gamma-Poisson convolutional model MPRAnalyze uses (methods). We generated 281 sequences with gradually increasing transcription rates spanning a range of possible values (from 0.2 to 3, in 0.01 steps), with three replicates in each simulation. The analyses above were repeated with the simulated data. We found that while the measured bias was indeed not influenced by the number of barcodes, the mean ratio is substantially more biased than both α and the aggregated ratio (Figure 2E). Similar to the real data results, we found α has lower variance than both ratio-based estimates, and higher correlation with the true transcription rates (Figure 2F-G). We also simulated data with varying number of replicates and found that increasing the number of replicates has a similar effect to increasing the number of barcodes, since both parameters increase the effective sample size. With any given number of barcodes, increasing the number of replicates improved performance - the degree of improvement decreased when more barcodes were available (Supplementary Figure S5).

Overall, we found that α performs similarly or better than both ratio-based estimators in terms of accuracy, consistency, and robustness to missing data.

Classification

A common use case for MPRA is classification of active sequences, which induce transcriptional activity. This is commonly done by comparing the ratio-based estimates of the assayed sequences to a control set of sequences [9, 10], an approach that suffers from the summary statistics' sensitivity to noise and missing data, demonstrated above, which in the context of classification leads to decreased power and accuracy. Other studies performed this analysis using DESeq2 [18], a differential expression analysis (DEA) method, by treating the DNA and RNA libraries as two conditions and looking for significant differences between the two [11]. In the following we demonstrate that overall DEA methods either lack power or are not well calibrated for MPRA data. More importantly, these methods rely on an implicit assumption that the majority of features do not display differential behavior, a valid assumption for RNA-seq that does not hold for MPRA, in which the assayed sequences are often explicitly selected for their potential activity. This assumption makes the results of DEA methods highly dependent on experimental design and sequence selection.

MPRAnalyze performs classification of active sequences by comparing the respective α estimates against the null distribution of transcription rate induced solely by the minimal promoter. The null is based on negative control sequences when available, and otherwise MPRAnalyze relies on a conservative assumption that the mode of the distribution of the α values is the mode of the null distribution, and that values lower than the mode are broadly

generated by the null. These values are therefore used to estimate the mean and variance of the null distribution.

In both scenarios, the α value of each candidate sequence is compared against the null distribution using the Median-Absolute-Deviation (MAD) - a variant of the Z-score that is less sensitive to outliers. MPRAnalyze supports either a one-sided or two-sided test, allowing for identification of inducing sequences (inducing transcription beyond the minimal promoter levels) or repressive sequences (repressing transcription to below the promoter levels). A one-sided test was used to generate all results presented in this paper.

Comparing MPRAnalyze with existing methods

To assess the performance of MPRAnalyze in classification analyses we compared six methods: MPRAnalyze with and without negative controls; empirical p-values computed using the two ratio-based estimates; and DESeq2 in either full mode (each barcode as a separate sample) or collapsed mode (each replicate as a sample, taking the sum across barcodes within each replicate; see Methods). Similarly to MPRAnalyze, DESeq2 was applied using an asymmetric mode, namely focusing on inducing sequences that have a higher signal in the RNA library than in the DNA library.

We examined the fraction of sequences that were significantly active (FDR < 0.05) in each dataset, stratified by group: negative controls, candidate sequences and positive controls when available (Figure 3A). As expected, empirical p-values from the ratio-based estimates show a clear lack of power. Both DESeq2-collapsed and MPRAnalyze without controls have inflated rates of false positives in the Kwasnieski datasets (compared with the theoretically expected 5% false discovery rate among the negative controls set). When examining the results across all datasets, we find that while MPRAnalyze and DESeq2 have overall comparable results, both modes of MPRAnalyze achieve a better balance between sensitivity (identifying candidates as active) and specificity (not identifying negative controls as active) than both modes of DESeq2 (Figure 3B).

Since the above analysis overlooks the overall statistical behavior of the methods, we examined the full p-value distribution of each method within each dataset. Considering multiple datasets, we found that both modes of MPRAnalyze, both ratio-based methods and DESeq2-full appear well calibrated, whereas DESeq2-collapsed does not follow the theoretical distribution of p-values: a mixture of uniform values (corresponding to non-active sequences that follow the null distribution) and low values (active sequences for which the null is rejected) (Supplementary Figure S6). Similar results were found when examining the distribution over negative controls only (expected to be uniform), with MPRAnalyze in the no-control mode having some inflated values (assigning more low p-values than expected), which emphasizes the importance of using negative controls in classification studies (Supplementary Figure S7). Finally, we examined the distribution over positive controls (only available in the Inoue-Kircher datasets), and found that MPRAnalyze in both modes has significantly higher statistical power, being outperformed only by ill-calibrated DESEq2-collapsed (Supplementary Figure S8). Overall, we found that despite comparable rates

of sequences found statistically significant, the MPRAnalyze model is better calibrated to MPRA data.

Caveats of using methods designed for differential expression

DESeq2 pools information across all features included in the dataset (genes for RNA-seq, candidate enhancers for MPRA), both in the library size correction and estimation of the dispersion parameter. However, unlike genome-wide assays such as RNA-seq, the set of assayed features in MPRA experiments is curated according to the specific goals and context of the study. We hypothesized that DESeq2-based classification would be highly dependent on the sequences included in the analysis. We repeated the classification analysis on the Inoue-Kreimer dataset using only the 200 negative controls sequences and 685 candidate sequences that were previously classified as active by MPRAnalyze and both modes of DESeq2. This simulated a scenario in which the data was generated in an experiment that included fewer sequences. Confirming our hypothesis, MPRAnalyze results remain unchanged with all candidate sequences significantly active, whereas DESeq2-full only classifies 161 (23.5%) of the sequences as active and DESeq2-collapsed finds no active sequences at all. This reveal an inherent limitation of using differential expression methods such as DESEq2 for analyzing MPRA data.

Comparative Studies

Another common use for MPRAs is comparative studies, looking for differential transcription induced by a putative regulatory sequence between different cell types, stimuli, or other experimental covariates [11, 16]. More complex experimental settings are also possible, e.g using MPRA to evaluate transcriptional activity over time as in the Inoue-Kreimer data [17], or the interaction between differential allele activity and the presence of a certain transcription factor, as performed by Ulirsch and colleagues [12].

Here we use the Inoue-Kircher data to demonstrate that MPRAnalyze is more statistically powerful than establish methods for analyzing comparative MPRA data, and therefore enables discovery of more neuanced biological signals, and that MPRAnalyze supports more complex experimental designs that are not supported by previous methods (e.g temporal analysis).

Performing differential activity analysis in MPRAnalyze can be done in two ways: first, since MPRAnlyze optimizes the model using likelihood maximization, any single hypothesis that can be encoded in a generalized linear model can be tested using a likelihood ratio test. This includes complex hypotheses that can be captured by interaction terms between covariates (e.g cell type and genetic background [12]). Additionally, in simple two-condition designs, or in cases where multiple contrasts are compared to a single reference (e.g multiple different stimuli compared against the unstimulated behavior), the model coefficients can be extracted from the RNA model and tested using a Wald test. While both options are

supported in the implementation of MPRAnalyze the results in this paper are based on likelihood ratio testing.

When performing comparative analysis, it is important to account for possible biases, such as those induced by overall differences in the basal transcription rate. In RNA-seq experiments this issue is usually resolved via library size correction [23], but with MPRA this is not necessarily sufficient. This is because for the library size to properly correspond to bias in the data, either the vast majority of features must be non-differential, or the differential signal must be symmetric. Neither of these assumptions necessarily hold for MPRA data, as they largely depend on the selection of the candidate sequences. For instance, MPRA can be designed with most sequences being more active in one condition than in the other, and thus most sequences are indeed differentially active. To address this issue, MPRAnalyze utilizes negative controls in the data to define the null differential behavior. This is done by fitting a separate, joint model for the controls, in which each control sequence has a distinct DNA model but they all share a single RNA model, reflecting the basal activity in each condition (Methods).

Alternative methods have been developed to address this or similar questions. QuASAR-MPRA [19] was designed specifically for allelic-comparisons and uses a beta-binomial model and mpralm [20] which is a general differential-activity tool designed for MPRA which fits a linear model. Both methods use summary statistics and do not include barcode-level information in their model. Mpralm can use either the aggregated ratio or the mean ratio as the statistic, and is therefore subject to the limitations described above. QuASAR-MPRA, similar to MPRAnalyze, models the DNA and RNA separately, but it does so using the sum of counts across all barcodes in each condition, collapsing the data into a single measurement.

Comparing MPRAnalyze with existing methods

To compare these different methods, we used the Inoue-Kreimer dataset and extended the subset of samples we used to include both the 0hr and 72hr timepoints (post neural induction of human embryonic stem cells (hESC)). We then looked for sequences whose activity differed between the two time points, using MPRAnalyze, mpralm (both aggregated ratio and mean ratio modes), and QuASAR-MPRA (Methods). The distribution of p-values (Figure 4A) shows that overall MPRAnalyze and both modes of mpralm are well calibrated, following the expected mixture of uniform values and low values among candidates, and showing slight inflation but overall uniform behavior among the negative controls. Conversely, QuASAR-MPRA is less calibrated on both candidates and negative control sequences, recapitulating the results described by Myint et al [20]. Indeed, QuASAR-MPRA only identified two candidates as significantly differential (BH-corrected p-values < 0.05).

Overall we observe that the estimates of effect size (log Fold-Change) are largely reproducible across methods (Pearson's $r > 0.84$ across all pairs). In terms of statistical power (Figure 4B), we observe that MPRAnalyze calls more sequences as significant compared to the other methods. We further note that the FDR values of MPRAnalyze are largely correlated with those of mpralm among statistically significant candidates (Spearman correlation

> 0.63 for sequences MPRAnalyze calls differential) and that the estimates of QuASAR-MPRA do not correlate with the other two methods (consistent with the results in Figure 4A). Further examination of the results excluded QuASAR-MPRA since it did not identify a sufficient number differential sequences.

We further examined the differential sequences, after filtering the results to only include candidate sequences that are classified as active in at least one of the conditions (BH-corrected $p < 0.05$, using MPRAnalyze's classification method). Interestingly, mpralm in aggregate mode finds a roughly balanced number of sequences that are increasing (99) and decreasing (91) in activity (comparing 0hr to 72hr), and in mean mode finds more decreasing (89) than increasing (49), while MPRAnalyze finds far more increasing (351) than decreasing (115) sequences (Figure 4C). However, sequences in the Inoue-Kreimer study were explicitly selected to correspond to increased activity over the course of differentiation (2037 [82%] of the assayed sequences are genomic regions selected due to their closest gene showing increased expression over differentiation). Therefore the imbalance in the results from MPRAnalyze fits to the design of the experiment.

We then explored the set of candidates that were detected by each method. To this end, we divided the set of differentially active sequences into decreasing and increasing activity (comparing 0hr to the 72hr time point), then within each set we tested for over-representation of DNA binding motifs (hypergeometric test, BH-corrected $p < 0.05$; Methods). To narrow down the results, we examined the union of top 15 most enriched transcription factor binding motifs by each method (Figures 4D, Supplementary Figure S9, Additional File 3: Table S1, Additional File 4: Table S2).

Among decreasing-activity sequences we find as expected binding sites for two of the core pluripotent factors (NANOG, POU5F1). While these are captured by all methods, we observe a higher significance with MPRAnalyze. Among increasing-activity sequences, where the methods have more profound differences, we find that MPRAnalyze generally has lower fold-enrichment scores, but compensates by a substantial increase in statistical power. Overall, mpralm in *mean* mode does not detect many of the enriched transcription factors found by the other methods, with a total of 23 (compared with 106 and 195 found by mpralm *aggregate* and MPRAnalyze, respectively), and displays diminished statistical power.

To ensure that these results are not simply explained by the higher number of differential sequences detected by MPRAnalyze, we also examined a *consensus + noise* option, where the consensus set (sequences called differential by all methods) was inflated with randomly chosen sequences (taken from the remaining population) to match the number of differential sequences called by MPRAnalyze (Methods). We find that this simulated inflation that does not reflect true biological signal does not explain the increased power displayed by MPRAnalyze.

Notably, MPRAnalyze results are enriched for binding sites for TEAD2 and NRF1, but results accourding to the other methods do not contain such enrichment. Both factors have been implicated in neurogenesis by previous studies [24, 25], and upon closer examination we found that NRF1 binding sites have comparable fold-enrichment in all methods (1.48 in MPRAnalyze, 1.39 in mpralm *aggregate* and 1.45 in mpralm *mean*), but only pass the statis-

tical threshold with MPRAnalyze. In the other direction, we found the mpralm results are enriched for binding sites of MYF5 and GSX1, but not the MPRAnalyze results. However, when examining the mRNA levels measured in the corresponding time points, we found that both factors have very low expression levels in the conditions in which MPRA was conducted (Additional File 5: Table S3). These levels are below their characteristic expression levels in tissues they are known to be active in [26], making them less attractive candidates for driving differential transcription. Overall, MPRAnalyze identifies biological signal that is consistent with the competing methods, with increased statistical power, which allows for more nuanced results.

Detecting temporal activity

Finally we note that MPRAnalyze can be used on the entire Inoue-Kreimer dataset, which consists of seven time-points, to identify sequences whose activity changes over time. MPRAnalyze performs this analysis by comparing two models: the full model, which allows for time-dependent activity; and the reduced model, in which time-point factors are excluded, thereby forcing a constant behavior across time-points (methods). This analysis cannot be performed by either of the competing methods: QuASAR-MPRA only supports two-condition comparisons, and mpralm only supports coefficient-based hypothesis testing. We ran MPRAnalyze in this fashion and after filtering sequences to only those that are active in at least one time-point (FDR < 0.05, using MPRAnalyze to perform classification analysis per time-point) MPRAnalyze finds 749 (28%) sequences that have temporal activity (methods, FDR < 0.05). Reassuringly, of the 466 sequences identified as differential between the first and last time-points, 420 (90.1%) are found to have overall temporal activity.

We found that temporal sequences broadly tend to have a smooth impulse-like activation pattern over time [27], whereas negative control sequences have less clear patterns (Supplementary Figure S10). We then clustered the temporal sequences (K-means with K=4 on α values, z-normalized for each sequence) in order to group sequences with similar temporal behavior pattern, and repeated the same binding site enrichment analysis as above (Additional File 6: Table S4) for each cluster. As evidence for the validity of our approach, we found that the sequences that are active at the early time points were indeed enriched for binding sites of core pluripotent factors (NANOG, SOX2, POUF51), and that sequences that are active later in the differentiation process were enriched for binding sites of transcription factors known to participate in neural differentiation (ATF2 [28], HES1 [29], GLI1, LEF [30]).

Allelic Comparison

Many MPRA studies deal with quantifying the effect of sequence variants on regulatory function. These studies, referred to here as allelic comparison studies, include those that compare observed genetic variants to investigate the regulatory effect of different alleles of a regulatory sequence [12], as well as studies that deliberately change a sequence to elucidate the regulatory grammar in a systemic fashion [13]. While conceptually similar to

comparative analyses, allelic comparisons require different factors to be considered. Two important differences are: (1) the compared sequences (e.g., wild type and mutant allele) come from the same sample and therefore a systemic bias is less concerning than it is when comparing different conditions, and (2) the different alleles being compared are associated with different barcodes, in contrast with condition-wise comparison in which barcodes are shared between conditions.

To demonstrate the utility of MPRAnalyze in this scenario, we used recently published data by Mattioli and colleagues [13], who measured the effects of all possible single-nucleotide deletions were examined on 31 selected promoters. To this end, an MPRA was conducted with all the deletion and corresponding wild type (WT) sequences, where each deletion was associated with 26 barcodes and each WT sequence was associated with 80 barcodes. A single sample of the pre-transduction plasmids was sequenced to produce the DNA library; The RNA samples were taken from two different tissues: eight samples from the HepG2 cell line and four samples from the K562 cell line. This asymmetrical experimental design exemplifies the diverse nature of MPRA studies, and the necessity of a flexible framework.

Using this dataset we demonstrate that MPRAnalyze is well calibrated and more statistically powerful than established methods, and supports studying the interaction of multiple conditions: in this case finding sequence variants with cell-line specific functional effects.

Comparing MPRAnalyze with existing methods

Similar to the comparative analysis described above, we compared each deletion sequence with the corresponding WT in each tissue separately, with all three methods: MPRAnalyze, mpralm (which only supports the *aggregated* mode for allelic comparisons), and QuASAR-MPRA.

When examining the P-value distribution generated by each method we find that MPRAnalyze and mpralm are both better calibrated than QuASAR-MPRA (Figure 5A-B). Consistent with our previous results, all methods have correlated estimates of biological effects (Figure 5C-F). The methods are better correlated in the HepG2 data compared with the K562 data (correlations with MPRAnalyze: Pearson's $r = 0.72$ in K562 and 0.77 in HepG2 for mpralm, and 0.78 in K562 and 0.96 in HepG2 for QuASAR), which we hypothesized is due to the higher number of replicates in the HepG2 data. When the comparison was repeated using only four replicates of the HepG2 data, the correlations between methods decreased (correlations with MPRAnalyze: Pearson's $r = 0.63$ for mpralm and 0.38 for QuASAR, Supplementary Figure S11).

We then compared the effects estimated by each method across cell lines. Overall, we find a high degree of similarity in the effects of sequence perturbation across cell lines - a finding supported by all the methods we considered (Figure 5G-I). Looking more closely, we find that mpralm and QuASAR-MPRA both find a systemic skew towards stronger effects in K562, with 72.6% and 63.1% of deletions having a more extreme log fold-change value in K562 compared with HepG2 in mpralm and QuASAR-MPRA, respectively, whereas MPRAnalyze results are more balanced, with 49.8%. When comparing statistical power, we again find

that MPRAnalyze can detect more deletions that significantly affect the rate of transcription (FDR < 0.05). In HepG2, MPRAnalyze finds 2855 (72%) deletions with a significant effect, whereas mpralm finds 2710 (68.4%), with 2071 (52.2%) of the sequences significant in both; in K562, MPRAnalyze finds 1230 (31%) significant deletions compared with 360 (9%) found by mpralm, with 272 (6.8%) significant in both. In both cell types, QuASAR-MPRA does not find any significantly functional deletion. As expected, due to the larger sample size, both MPRAnalyze and mpralm are more powerful in HepG2 compared with K562.

Identifying variants with cell-line specific effects

Since the Mattioli study performed allelic comparisons in two cell types, it can also be used for the identification of deletions that have a different effect in HepG2 cells compared with K562 cells. With MPRAnalyze it is possible to address this question directly, testing the interaction between the tissue and the allele covariates in the model. When performing this analysis, MPRAnalyze found 608 (15.3%) differential deletions that had a different effect between cell types. For example, the core promoter of the lncRNA gene DLEU1 has several functional deletions that are highly concordant between cell types, and a single differentially functional deletion in position 83, where the deletion has a significantly larger effect in HepG2 ($logFC = -0.86$) than in K562 ($logFC = -0.13$)(Supplementary Figure S12).

To examine the biological implications of our results, we followed the analysis performed by Mattioli and colleagues and identified transcription factor binding motifs that are disturbed by the single nucleotide deletions. Focusing only on functional deletions (i.e., deletions that had any effect in one or both cell lines), we looked for DNA binding motifs whose disruptive deletions are over-represented in the set of conditionally functional deletions (i.e., deletions with significantly more effect in one cell line vs. the other)(Figure 5J). Overall, we found three statistically enriched (Hypgergeometric test, $FDR < 0.05$, Methods) motifs in the cell type specific deletions (Figure 5K-L). Reassuringly, we found that K562-specific deletions were enriched for motifs of the erythroid transcription factor NF-E2. These results demonstrate the potential utility of MPRAnalyze in addressing cases of complex and possibly asymmetric experimental designs.

5.4 Discussion

Massively Parallel Reporter Assays are a powerful technique for functional characterization of enhancer activity in a high-throughput manner. MPRAs can be used to quantify the contribution made by non-coding DNA elements, such as enhancers, to transcriptional activity in nearby genes [10]. It can be further extended to evaluate differences in regulatory activity between different alleles [12], elucidate regulatory grammar via mutagenesis studies [13], and compare enhancer activity between conditions [9]. Complex experimental designs can include interaction studies, where one is interested in how sequence changes affect differen-

tial activity between cellular conditions [12], or identifying temporal patterns in time-course data [17].

Since MPRAs are still an actively-developing technology, they often vary in experimental design. While MPRAnalyze is flexible and can handle various study designs, the method benefits from certain experimental decisions that are generally recommended but not always leveraged in other analyses. First, pairing the DNA and RNA libraries by extracting DNA from the same post-transduction libraries that the RNA libraries are extracted from, avoids introducing further experimental noise into the data, and enables MPRAnalyze to better fit and relate the two models to increase accuracy of estimating nuisance factors. Additionally, as demonstrated in our results, increasing the number of available barcodes and replicates can greatly reduce the measured noise and increase performance of all methods, as seen in recent studies [31, 32]. Finally, the inclusion of negative control sequences allows explicit modeling of the null behavior and avoids relying on assumptions that may bias the results and prevent proper interpretation of them. The curated nature of MPRA datasets makes negative controls a valuable and often crucial aspect of properly interpreting the results.

MPRAnalyze offers a robust statistical framework that enables all major uses of MPRAs in a unified model. Our model avoids relying on ratio-based summary statistics, seeking to directly model the data as structured, following a similar trend in other high throughput functional assay analysis methods, such as recent methods developed for the analysis of Deep Mutational Scanning data [33, 34, 35]. MPRAnalyze models noise in both DNA and RNA libraries and uses a nested GLM design to control barcode-specific effects and leverage the multiplicity of barcodes for increased statistical power. The method is highly flexible and allows various complex study designs to be tested in a straight-forward manner, including those currently not supported by any established method. Additionally, MPRAnalyze avoids relying on population-level properties in the analysis, instead leveraging negative controls when available to establish null behaviors.

5.5 Methods

Dataset Collection and Processing

For all datasets included in this paper, we relied on the pre-processing and filtering performed by the authors of the original papers. This ensures that MPRAnalyze's performance isn't reflecting any favorable processing steps we chose.

Kwasnieski: The study [10] measured the activity of potential regulatory regions in K562 cells. Regions were selected according to ENCODE annotations of four groups: enhancers; weak enhancers; repressed enhancers; enhancers active in ESCs. The repressed and ESC-annotated enhancers were used as controls, and were excluded from the analysis after library size normalization factors were computed. In addition to control classes, each class had internal sets of scrambled sequences used as negative controls, which were used as controls in our analyses. Each sequences in this dataset was associated with 4 barcodes. The

DNA was sequenced before transduction and with a single replicate, while 4 replicates are available for the RNA. While the sample size in this data is very limited, this allowed for higher read counts to be achieved, mitigating the loss of statistical power by getting more reliable quantification of the counts.

Inoue-Kircher: The study [9] compared activity in HepG2 cells of liver enhancers that were either episomal or chromosomally integrated using a lentivirus (lentiMPRA). While the study is comparative, the comparison is not between biological conditions and the results are therefore difficult to validate or interpret. We therefore decided to use the data as two separate quantification datasets. The datasets were analyzed together to better account for batch and barcode-specific effects, and α estimates were extracted from the joint model for each condition separately. Negative control sequences were generated by scrambling candidate sequences, and positive controls were sequences that have been previously validated as having a regulatory function in these cells. Each sequence was associated with 100 unique barcodes. DNA was sequenced post-transduction. Both DNA and RNA have three replicates.

Inoue-Kreimer: The study [17] identified enhancers with temporal activity over the first 72 hours after neural induction. lentiMPRA was performed in 7 timepoints (0, 3, 6, 12, 24, 48 and 72 hours after undiction). For the purpose of our analysis, we used only the data from the first timepoint in the quantification and classification analyses, and timepoints 0 and 72 hours for the comparative analysis. Negative controls are scrambled candidate sequences. Each sequence was associated with 90 barcodes. DNA was sequenced post-transduction. Both DNA and RNA have three replicates.

Mattioli: The study [13] compared WT sequences containing core promoters of 31 long non-coding RNAs (21 sequences), enhancer RNAs (5 sequences) and messenger RNAs (5 sequences) with the same sequences with single-nucleotide deletions. Each core promoter was divided to 2 'tiles' to cover more of the putative regulatory sequence, resulting in a total of 62 WT sequences. MPRAs were performed in both K562 and HepG2 cells, with varying number of replicates (4 in K562, 8 in HepG2). DNA was sequenced pre-transduction, and in a single replicate (used for both cell types).

Computing Transcription Rate Estimates

All transcription rate estimates were computed for library size normalized MPRA data, using upper quartile normalization to compute size factors. **MPRAnalyze's** α was computed for each dataset using the quantification analysis (See supplemental methods). Across datasets, batch and barcode-level effects were modelled in the nested DNA model, but excluded from the RNA model design. This allows MPRAnalyze to model nuisance effects but asserts that all barcodes associated with a single sequence must share the same transcription rate. Both ratio-based estimates were computed using only barcodes with non-zero DNA and RNA counts. So for each sequence: $S = \{i \in [n] \,|\, R_i \neq 0, D_i \neq 0\}$. Then the **Mean Ratio** $= \frac{1}{|S|} \sum_{i \in S} \frac{R_i}{D_i}$, and the **Aggregated Ratio** $= \frac{\sum_{i \in S} R_i}{\sum_{j \in S} D_j}$.

Running alternative methods

DESeq2: DESeq2 was used as a method for classifying active enhancers, by comparing the DNA and RNA libraries as the two conditions being compared. DEseq2 was used in two modes: the full mode included each barcode as separate sample, and the collapsed mode took the sum across barcodes within each batch as a sample. In full mode, a single count was added to each observation to avoid issues with DESeq2 library normalization scheme. The model used within DESeq2 was a simple comparison model: type, where type identifies DNA and RNA observations.

QuASAR-MPRA: QuASAR MPRA was used according to the documentation provided in the package. The *betas.beta.binom* value, which is the logit transformation of the allelic skew, was used as a proxy for log fold-change.

mpralm: mpralm was used according to the documentation provided in the package. For allelic comparisons, while the package vignette uses the sum across barcodes when aggregating the counts, we used the mean across barcodes instead, since the two compared alleles did not have the same number of associated barcodes. Additionally, since the package requires manual aggregation of barcodes in this situation, only the *aggregated* mode of the model is supported for this type of analysis.

Subsampling analysis

For the subsampling analysis, barcodes were sampled down to varying levels (for Inoue-Kircher datasets: 15, 30, 45, 60, 75, 90 out of the total 100 barcodes; for Inoue-Kreimer: 15, 30, 45, 60, 75 of the total 90 barcodes). The analysis uses three independent replicates of this down-sampling process, so overall for each sequence we get a set of $3 \times K$ estimates at various numbers of available barcodes, where $K = 6$ for the Inoue-Kircher datasets and $K = 5$ for Inoue-Kreimer. The analyses were done on the entire down-sampled dataset in a single run and included the original data as well as the reduced-barcodes data, to neutralize any effect that the library size correction might have on the estimates.

Simulating MPRA data

MPRA data was simulated by generating random coefficients for the nested GLM construction that MPRAnalyze uses. The *latent (true)* DNA and RNA counts were generated directly from the model, then log-normal noise was added to the latent counts to get the *observed*

counts. Formally:

$$\begin{aligned}
\vec{\beta} &= \left[\beta_0, \vec{\beta_{batch}},\ \vec{\beta_{BC}}\right] \\
\beta_0 &\sim N\left(K, \sigma_0^2\right) \\
\vec{\beta_{batch}} &\sim N\left(0, \sigma_{batch}^2\right) \\
\vec{\beta_{BC}} &\sim N\left(0, \sigma_{BC}^2\right) \\
&\Downarrow \\
\vec{D}_{true} &= nint\left(exp\left(X_d\vec{\beta}\right)\right) \\
\vec{R}_{true} &= nint\left(exp\left(\alpha \cdot X_d\vec{\beta}\right)\right) \\
\vec{D}_{observed} &\sim nint\left(log - Normal\left(exp\left(X_d\vec{\beta}\right), \sigma_D^2\right)\right) \\
\vec{R}_{observed} &\sim nint\left(log - Normal\left(exp\left(\alpha \cdot X_d\vec{\beta}\right), \sigma_R^2\right)\right)
\end{aligned}$$

where K controls the intercept term for the construct distribution, the variance of which is σ_0^2; $\sigma_{batch}^2, \sigma_{BC}^2$ control the size of batch and barcode effects, respectively; σ_D^2, σ_R^2 determine the noise levels added to the data; *nint* is the nearest-integer function, using base R's *round* function. An implementation of this simulation process is included in the MPRAnalyze package.

Noise was generated using log-normal noise instead of Gamma/Negative Binomial to avoid generating data directly from MPRAnalyze's model, which might bias the results.

Simulated data in this manuscript was generated with 3 batches, varying numbers of barcodes, $K = 5$, and $\sigma_0 = \sigma_{batch} = \sigma_{BC} = \sigma_D = \sigma_R = 0.5$.

Transcription Factor Binding Site enrichment analysis

The transcription factor binding site enrichment analysis was performed using the binary binding matrix computed by Inoue & Kreimer et al. [17], with each entry indicating the potential for binding (motif-based binding prediction using Fimo [36], $FDR < 10^{-4}$) or overlap with transcription factor ChIP-seq peaks from publicly available data [37, 38]. Enrichment was calculated using a hypergeometric test, with all binding motifs of a each transcription factors being pooled together. A factor was deemed enriched if BH-corrected $P < 0.05$. Enrichment scores were calculated as: $\log_2\left(\frac{\text{fraction of differential sequences containing a binding site of the TF}}{\text{fraction of total sequences containing a binding st of the TF}}\right)$. For the *consensus + noise* option, for each TF we calculated the number of predicted binding sites in the consensus set and in the remaining population. We then added artificial biding sites to the consensus set, proportional to their abundance in the remaining population, to match the number of differential sequences called by MPRAnalyze.

Temporal Activity Analysis

The analysis was performed by setting the full RNA model to include both batch and time-course factors ($\sim$ *batch + time*), and the reduced model to batch factors only ($\sim$ *batch*). A Likelihood-ratio test is performed for statistical significance, and a sequence is deemed 'temporal' if BH-corrected $P < 0.05$. Heatmaps for visualization were generated using the ComplexHeatmap R package [39].

Differential Deletions Analysis

To identify differential deletions (deletions that affect the induced transcription rate in K562 differently than in HepG2) in the Mattioli dataset we used an interaction term in the RNA model design, encoding the interaction between the cell type factor and the allele factor:

$$\mathcal{H}_0 : RNA \sim Allele + CellType + Allele : CellType$$
$$\mathcal{H}_1 : RNA \sim Allele + CellType$$

Then a standard likelihood ratio test was performed to determine statistical significance. Since the DNA data has a different design (a single replicate shared across all RNA samples), that design only modeled for barcode specific effects.

Differential Deletion Motif Analysis

Once differential deletions were identified, we divided the differential deletions to those that had a greater effect in HepG2 or K562. For each cell type, we used the motif hits curated by Mattioli and collegues, which rely on FIMO-based [36] predicted binding scores, to associate each deletion with differential motifs: motifs predicted in one allele and not the other. If the deletion cause a decrease in induced transcription rate, we took the 'lost' motifs (predicted in WT, not in the deletion); and if the deletion caused an increase we took the 'gained' motifs (predicted in the deletion, not the WT). All motifs associated with the same transcription factor were pooled. Enrichment scores were calculated using a hypergeometric test, using the total set of functional deletions as background (motifs for these were acquired in the same fashion).

5.6 Figures

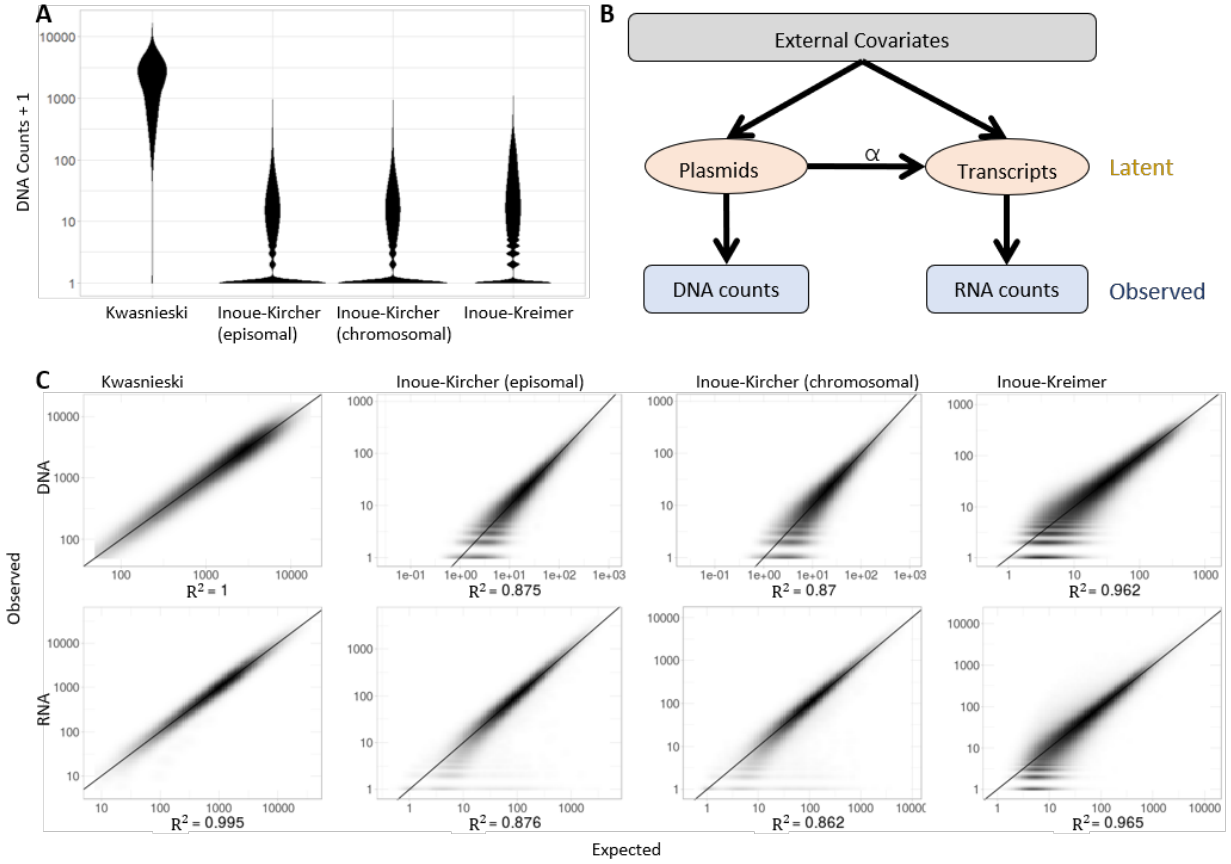


Figure 1: **MPRAnalyze model properties and fit. (A)** Distribution of construct abundances (DNA barcodes) across datasets, computed as the observed barcode count + 1 for visualization purposes. **(B)** A graphical representation of the MPRAnalyze model. External covariates (e.g conditions of interest, batch effects, barcode effects) are design-dependent; Latent construct and transcript counts are related by the transcription rate α. **(C)** Goodness of fit plots for both DNA and RNA libraries across datasets. Expected counts were extracted from the fitted GLMs. MPRAnalyze's model fits MPRA data well, with $R^2 > 0.86$ across all datasets. Since the Kwasnieski data only has one replicate in the DNA library, the DNA model is able to reach a perfect fit, in which case the DNA estimates used in the RNA model are identical to the original DNA counts.

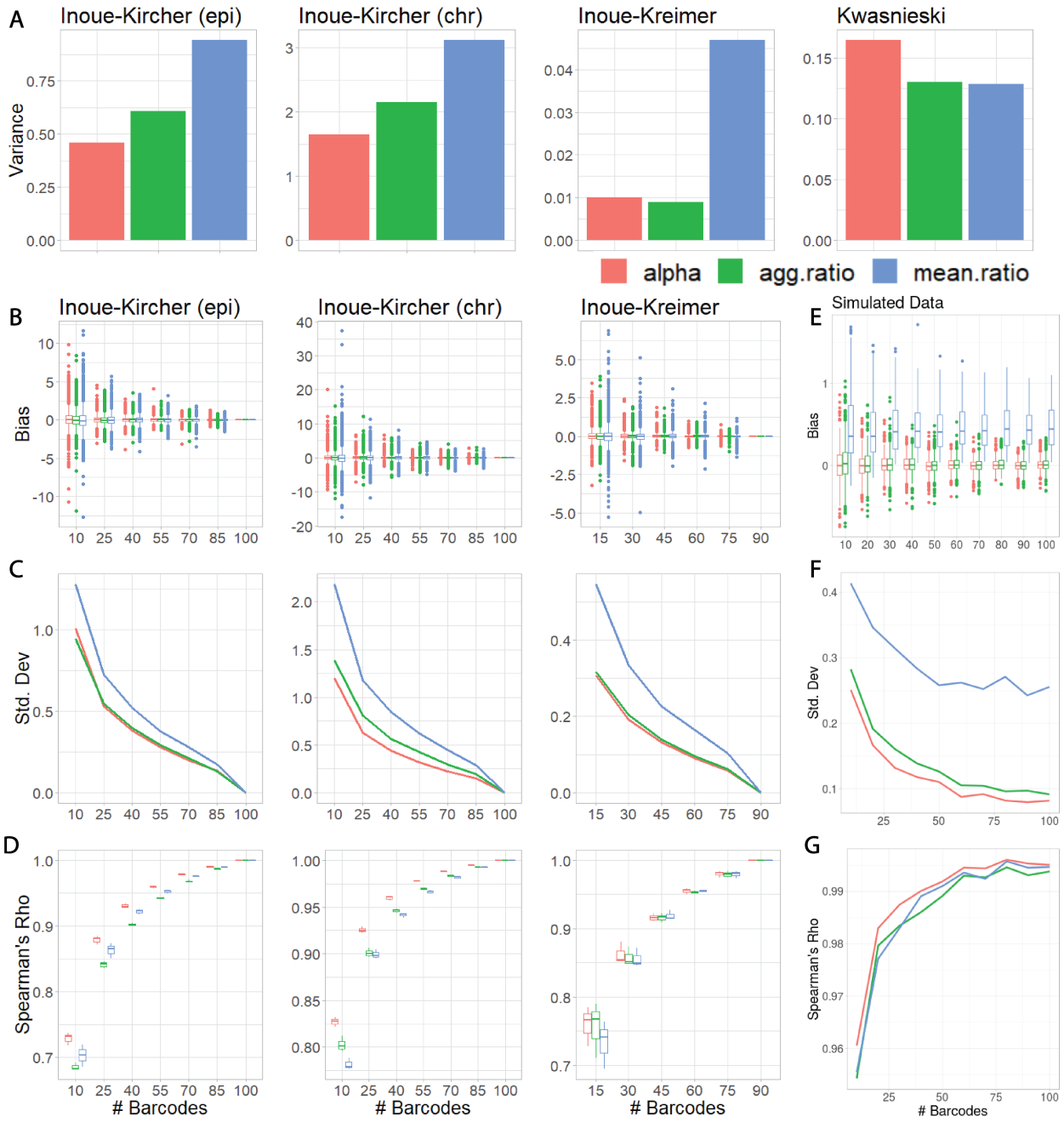


Figure 2: **Comparison of MPRAnalyze's α estimate of transcription rate with the ratio-based estimates** $\left(\text{agg.ratio: } \frac{\frac{1}{n}\sum_i^n RNA_i}{\frac{1}{m}\sum_j^m DNA_j}; \text{mean.ratio: } \frac{1}{n}\sum_i^n \frac{DNA_i}{RNA_i}\right)$ **(A)** The variance measured among estimates of negative-control sequences in each dataset (these are assumed to have an identical transcription rate). **(B-D)** Barcodes were sampled and quantification was recomputed based on the partial data to measure the effect of barcode number on estimate performance [See methods for further subsampling details]. Analyses were performed using the full-data estimate as the ground truth. **(E-G)** MPRA data was simulated to provide an actual ground truth. In each case we measured the bias $(estimate - truth)$ **(B,E)**; the standard deviation $\left(\sqrt{Var\left(estimate - truth\right)}\right)$ **(C,F)**; and the Spearman correlation between the estimates and the ground truth **(D,G.)**

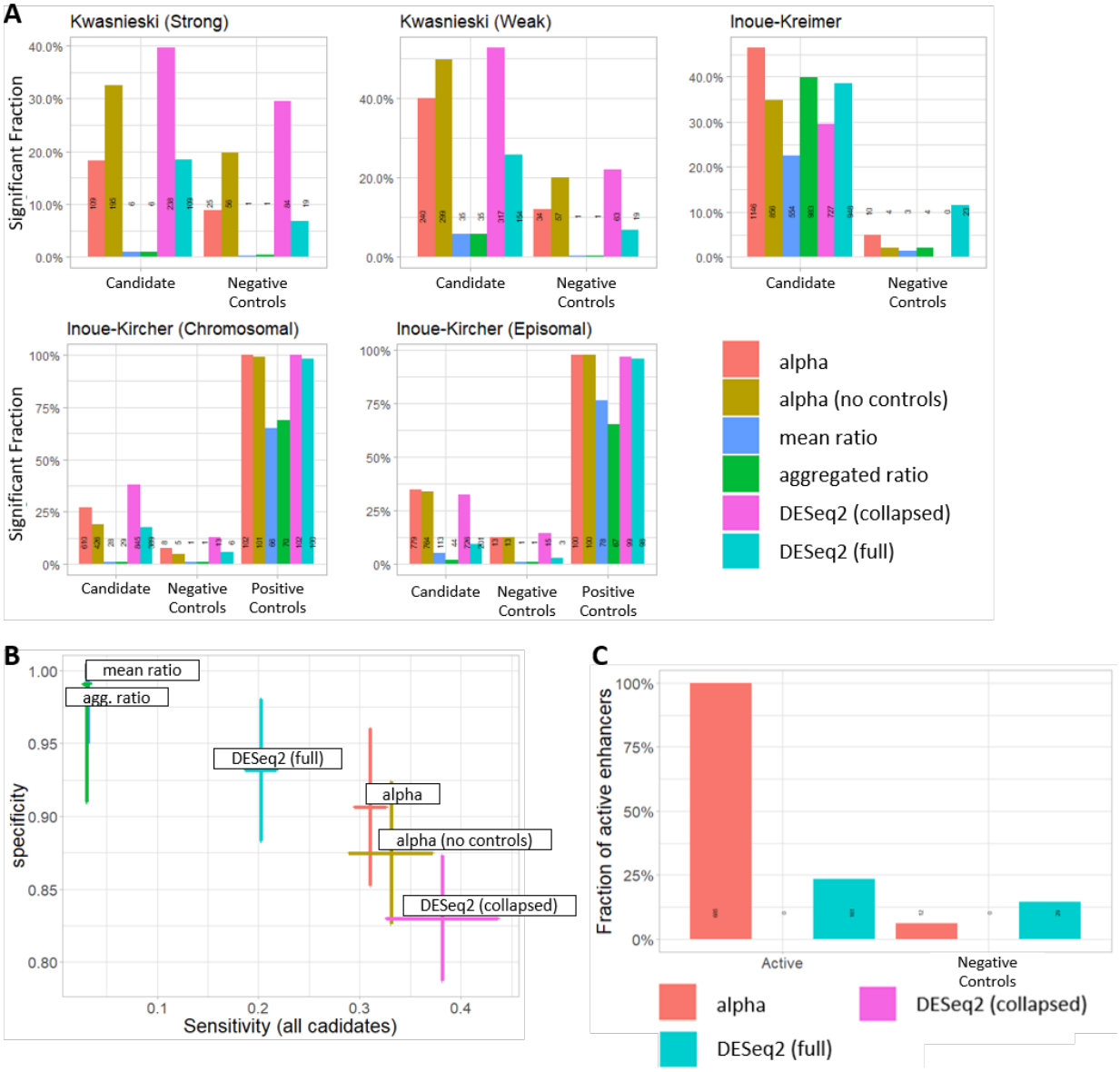


Figure 3: **Classification analysis comparisons. (A)** fraction of sequences identified as significantly active (BH-corrected $P < 0.05$) by method and class of sequence. MPRAnalyze results both in control-based (red) and no-controls (orange) modes; empirical p-values based on the mean ratio (blue) or aggregated ratio (green); DESeq2 results in collapsed mode (barcodes are summed within each batch, purple) or full mode (full data, light blue). Absolute number of active sequences is displayed on the bars. **(B)** Precision-Recall curve. Precision is based on performance on the negative controls, Recall is based on the total population of sequences, assuming all candidates are active. Error bars are $\pm$ the standard deviation of these measures across datasets. **(C)** Fraction of active sequences detected after re-running the analyses on 685 sequences from the Inoue-Kreimer dataset that were identified as active by MPRAnalyze (regular mode) and both DESeq2 modes, and the 200 controls from the same dataset. MPRAnalyze recapitulates the same results, finding that 100% of the candidates are active, whereas DESeq2 full only identifies 161 (23.5%) and DESeq2 collapsed completely fails to identify any active sequences.

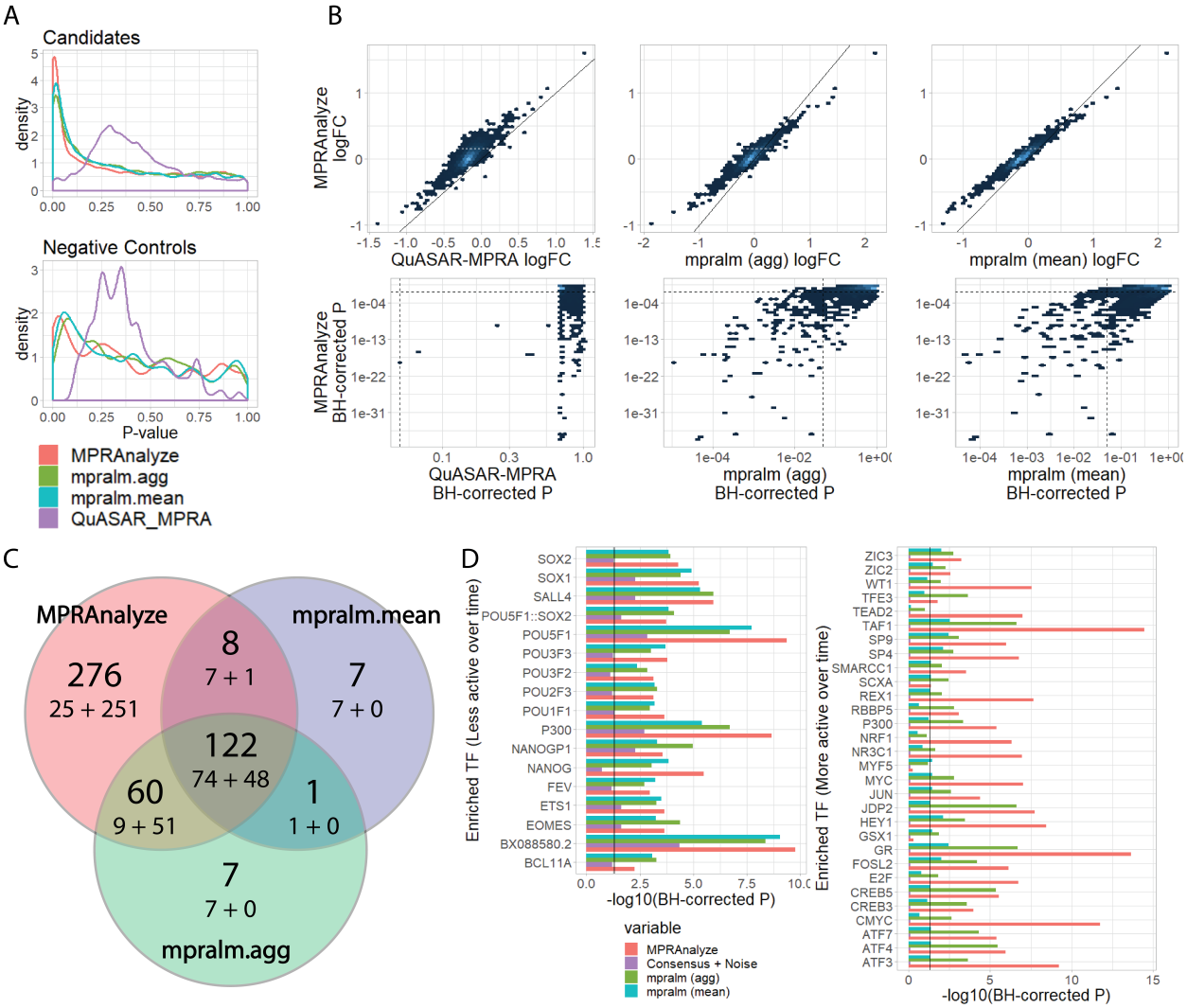


Figure 4: **Comparative analysis results of comparing timepoint 0h to 72h in the Inoue-Kreimer dataset. (A)** P-value distributions of candidates (top) and controls (bottom). QuASAR-MPRA is poorly calibrated, whereas MPRAnalyze and both mpralm modes follow the theoretical behavior (mixture of uniform and low values). **(B)** Direct comparison of MPRAnalyze to competing methods. Top panels show the biological effect size (log Fold-change); Bottom panels show the statistical significance (BH-corrected P; dotted lines are 0.05 threshold). **(C)** Venn diagram for MPRAnalyze and mpralm (both modes). The numbers in each area are (top) the total number of sequences in the area, and (bottom) the number of decreasing-activity sequences (left) + and increasing-activity sequences (right). **(D)** Enrichment of transcription factor binding sites in differentially active sequences as determined by each method. Solid line represents threshold of 0.05. (see Methods for further details).

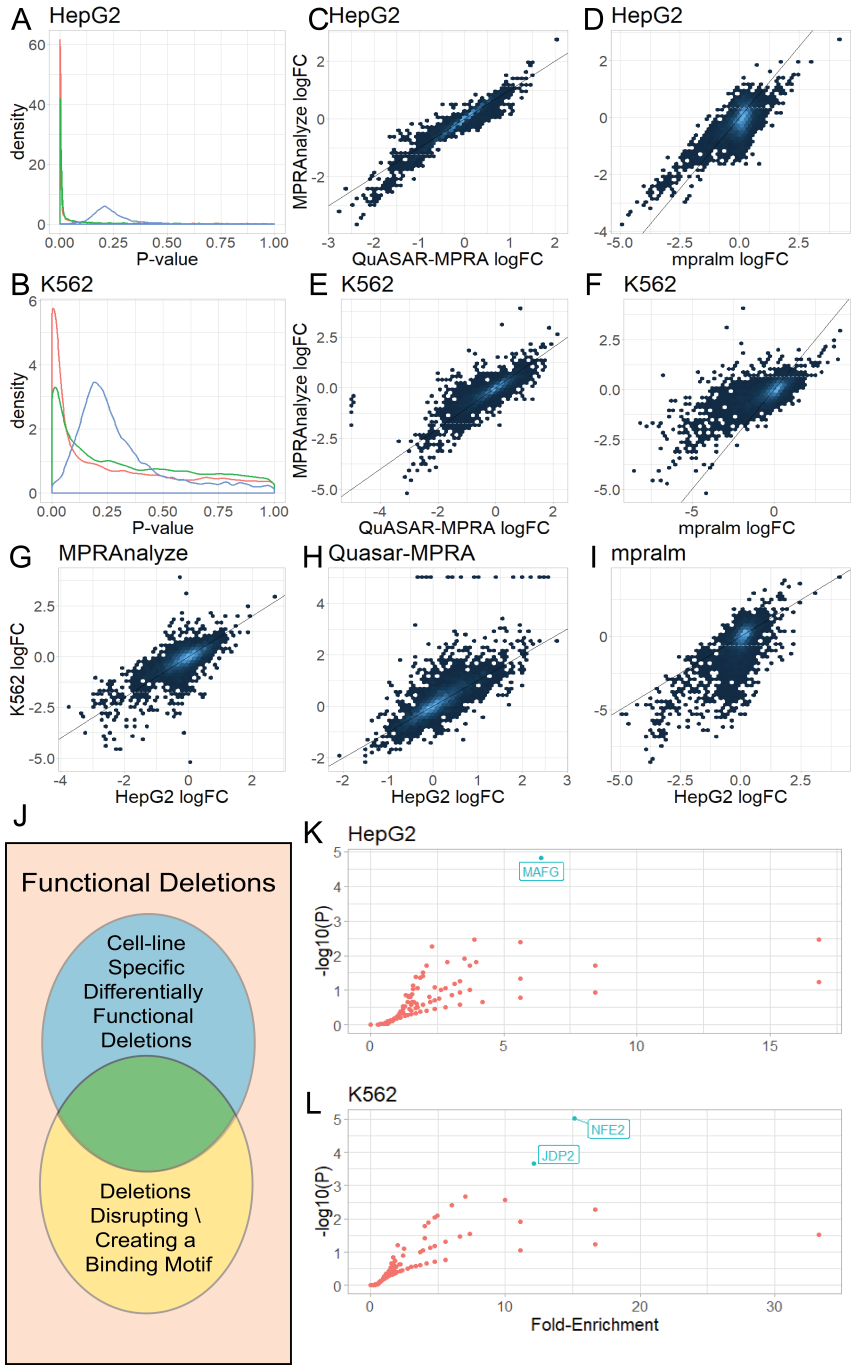


Figure 5: **Performance evaluation in allelic comparison. (A,B)** P-value density of the three evaluated methods in both cell lines. **(C-F)** logFC values between methods in each cell type shows all methods find a similar biological signal. **(G-I)** logFC values between cell types for each method. Some differences are expected, but overall values are highly correlated. **(J)** Schematic of the enrichment analysis, testing cell-line specific functional deletions for enrichment of motifs that were gained or lost by those deletions. **(K-L)** results of motif enrichment analyses. Transcription Factors with significant enrichment ($FDR < 0.05$) are labeled.

5.7 Tables

Dataset	Type of Analysis	Integration	DNA Sequencing	#Sequences	#Negative Controls	#Barcodes	#Replicates (DNA,RNA)
Kwasnieski [10]	Quantification	Episomal	Pre-transduction	1200	568	4	1,4
Inoue-Kircher [9] (epi)	Quantification	Episomal	Post-transduction	2338	102	100	3,3
Inoue-Kircher [9] (chr)	Quantification	Lentiviral	Post-transduction	2338	102	100	3,3
Inoue-Kreimer[17]	Comparative	Lentiviral	Post-transduction	2464	200	90	3,3
Mattioli[13]	Allelic Comparison	Episomal	Pre-transduction	3960	0	26/80	1, 4/8

Table 5.1: MPRA datasets used for evaluation of MPRAnalyze throughout the paper. In the Mattioli data, multiple values indicate an asymmetric design: reference alleles were associated with 80 barcodes compared with 26 barcodes for alternative alleles, and 4 replicates were available for K562 cells compared with 8 in HepG2. For further details on each datasets, see Methods

5.8 References

[1] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11247. URL: http://dx.doi.org/10.1038/nature11247.

[2] Olga I Kulaeva et al. "Distant activation of transcription: mechanisms of enhancer action". en. In: *Mol. Cell. Biol.* 32.24 (Dec. 2012), pp. 4892–4897. ISSN: 0270-7306, 1098-5549. DOI: 10.1128/mcb.01127-12. URL: http://dx.doi.org/10.1128/MCB.01127-12.

[3] Glenn A Maston, Sara K Evans, and Michael R Green. "Transcriptional regulatory elements in the human genome". en. In: *Annu. Rev. Genomics Hum. Genet.* 7 (2006), pp. 29–59. ISSN: 1527-8204. DOI: 10.1146/annurev.genom.7.080505.115623. URL: http://dx.doi.org/10.1146/annurev.genom.7.080505.115623.

[4] Lucia A Hindorff et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.23 (June 2009), pp. 9362–9367. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0903103106. URL: http://dx.doi.org/10.1073/pnas.0903103106.

[5] Sumantra Chatterjee and Nadav Ahituv. "Gene Regulatory Elements, Major Drivers of Human Disease". en. In: *Annu. Rev. Genomics Hum. Genet.* 18 (Aug. 2017), pp. 45–63. ISSN: 1527-8204, 1545-293x. DOI: 10.1146/annurev-genom-091416-035537. URL: http://dx.doi.org/10.1146/annurev-genom-091416-035537.

[6] Justin B Kinney and David M McCandlish. "Massively Parallel Assays and Quantitative Sequence-Function Relationships". en. In: *Annu. Rev. Genomics Hum. Genet.* (May 2019). ISSN: 1527-8204, 1545-293x. DOI: 10.1146/annurev-genom-083118-014845. URL: http://dx.doi.org/10.1146/annurev-genom-083118-014845.

[7] Fumitaka Inoue and Nadav Ahituv. "Decoding enhancers using massively parallel reporter assays". en. In: *Genomics* 106.3 (Sept. 2015), pp. 159–164. ISSN: 0888-7543, 1089-8646. DOI: 10.1016/j.ygeno.2015.06.005. URL: http://dx.doi.org/10.1016/j.ygeno.2015.06.005.

[8] Axel Visel et al. "VISTA Enhancer Browser–a database of tissue-specific human enhancers". en. In: *Nucleic Acids Res.* 35.Database issue (Jan. 2007), pp. D88–92. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkl822. URL: http://dx.doi.org/10.1093/nar/gkl822.

[9] Fumitaka Inoue et al. "A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity". en. In: *Genome Res.* 27.1 (Jan. 2017), pp. 38–52. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.212092.116. URL: http://dx.doi.org/10.1101/gr.212092.116.

[10] Jamie C Kwasnieski et al. "High-throughput functional testing of ENCODE segmentation predictions". en. In: *Genome Res.* 24.10 (Oct. 2014), pp. 1595–1602. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.173518.114. URL: http://dx.doi.org/10.1101/gr.173518.114.

[11] Ryan Tewhey et al. "Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay". en. In: *Cell* 165.6 (June 2016), pp. 1519–1529. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.04.027. URL: http://dx.doi.org/10.1016/j.cell.2016.04.027.

[12] Jacob C Ulirsch et al. "Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits". en. In: *Cell* 165.6 (June 2016), pp. 1530–1545. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.04.048. URL: http://dx.doi.org/10.1016/j.cell.2016.04.048.

[13] Kaia Mattioli et al. "High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity". en. In: *Genome Res.* 29.3 (Mar. 2019), pp. 344–355. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.242222.118. URL: http://dx.doi.org/10.1101/gr.242222.118.

[14] Robin P Smith et al. "Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model". en. In: *Nat. Genet.* 45.9 (Sept. 2013), pp. 1021–1028. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.2713. URL: http://dx.doi.org/10.1038/ng.2713.

[15] Shira Weingarten-Gabbay and Eran Segal. "A shared architecture for promoters and enhancers". en. In: *Nat. Genet.* 46.12 (Dec. 2014), pp. 1253–1254. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3152. URL: http://dx.doi.org/10.1038/ng.3152.

[16] Susan Q Shen et al. "Massively parallel cis-regulatory analysis in the mammalian central nervous system". en. In: *Genome Res.* 26.2 (Feb. 2016), pp. 238–255. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.193789.115. URL: http://dx.doi.org/10.1101/gr.193789.115.

[17] Fumitaka Inoue et al. "Massively parallel characterization of regulatory dynamics during neural induction". en. July 2018. DOI: 10.1101/370452. URL: https://www.biorxiv.org/content/early/2018/07/16/370452.

[18] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12 (2014), p. 550. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0550-8. URL: http://dx.doi.org/10.1186/s13059-014-0550-8.

[19] Cynthia A Kalita et al. "QuASAR-MPRA: Accurate allele-specific analysis for massively parallel reporter assays". en. In: *Bioinformatics* (Sept. 2017). ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btx598. URL: http://dx.doi.org/10.1093/bioinformatics/btx598.

[20] Leslie Myint et al. "Linear models enable powerful differential activity analysis in massively parallel reporter assays". en. Sept. 2017. DOI: 10.1101/196394. URL: https://www.biorxiv.org/content/early/2017/09/30/196394.

[21] Wolfgang Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". en. In: *Nat. Methods* 12.2 (Feb. 2015), pp. 115–121. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3252. URL: http://dx.doi.org/10.1038/nmeth.3252.

[22] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". en. In: *Bioinformatics* 26.1 (Jan. 2010), pp. 139–140. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btp616. URL: http://dx.doi.org/10.1093/bioinformatics/btp616.

[23] Farnoosh Abbas-Aghababazadeh, Qian Li, and Brooke L Fridley. "Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing". en. In: *PLoS One* 13.10 (Oct. 2018), e0206312. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0206312. URL: http://dx.doi.org/10.1371/journal.pone.0206312.

[24] Kotaro J Kaneko et al. "Transcription factor TEAD2 is involved in neural tube closure". en. In: *Genesis* 45.9 (Sept. 2007), pp. 577–587. ISSN: 1526-954x. DOI: 10.1002/dvg.20330. URL: http://dx.doi.org/10.1002/dvg.20330.

[25] Wen-Teng Chang et al. "A novel function of transcription factor alpha-Pal/NRF-1: increasing neurite outgrowth". en. In: *Biochem. Biophys. Res. Commun.* 334.1 (Aug. 2005), pp. 199–206. ISSN: 0006-291x. DOI: 10.1016/j.bbrc.2005.06.079. URL: http://dx.doi.org/10.1016/j.bbrc.2005.06.079.

[26] Gil Stelzer et al. "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses". en. In: *Curr. Protoc. Bioinformatics* 54 (June 2016), pp. 1.30.1–1.30.33. ISSN: 1934-3396, 1934-340x. DOI: 10.1002/cpbi.5. URL: http://dx.doi.org/10.1002/cpbi.5.

[27] Nir Yosef and Aviv Regev. "Impulse control: temporal dynamics in gene transcription". en. In: *Cell* 144.6 (Mar. 2011), pp. 886–896. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2011.02.015. URL: http://dx.doi.org/10.1016/j.cell.2011.02.015.

[28] Julien Ackermann et al. "Loss of ATF2 function leads to cranial motoneuron degeneration during embryonic mouse development". en. In: *PLoS One* 6.4 (Apr. 2011), e19090. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0019090. URL: http://dx.doi.org/10.1371/journal.pone.0019090.

[29] Ryoichiro Kageyama, Toshiyuki Ohtsuka, and Taeko Kobayashi. "Roles of Hes genes in neural development". en. In: *Dev. Growth Differ.* 50 Suppl 1 (June 2008), S97–103. ISSN: 0012-1592, 1440-169x. DOI: 10.1111/j.1440-169X.2008.00993.x. URL: http://dx.doi.org/10.1111/j.1440-169X.2008.00993.x.

[30] Hyun-Kyung Lee, Hyun-Shik Lee, and Sally A Moody. "Neural transcription factors: from embryos to neural stem cells". en. In: *Mol. Cells* 37.10 (Oct. 2014), pp. 705–712. ISSN: 1016-8478, 0219-1032. DOI: 10.14348/molcells.2014.0227. URL: http://dx.doi.org/10.14348/molcells.2014.0227.

[31] Jessica E Davis et al. "Multiplexed dissection of a model human transcription factor binding site architecture". en. May 2019. DOI: 10.1101/625434. URL: https://www.biorxiv.org/content/10.1101/625434v2.

[32] Joris van Arensbergen et al. "Systematic identification of human SNPs affecting regulatory element activity". en. Jan. 2019. DOI: 10.1101/460402. URL: https://www.biorxiv.org/content/10.1101/460402v2.

[33] Douglas M Fowler and Stanley Fields. "Deep mutational scanning: a new style of protein science". en. In: *Nat. Methods* 11.8 (Aug. 2014), pp. 801–807. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3027. URL: http://dx.doi.org/10.1038/nmeth.3027.

[34] Sebastian Matuszewski et al. "A Statistical Guide to the Design of Deep Mutational Scanning Experiments". en. In: *Genetics* 204.1 (Sept. 2016), pp. 77–87. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.116.190462. URL: http://dx.doi.org/10.1534/genetics.116.190462.

[35] Alan F Rubin et al. "A statistical framework for analyzing deep mutational scanning data". en. In: *Genome Biol.* 18.1 (Aug. 2017), p. 150. ISSN: 1465-6906. DOI: 10.1186/s13059-017-1272-5. URL: http://dx.doi.org/10.1186/s13059-017-1272-5.

[36] Charles E Grant, Timothy L Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif". en. In: *Bioinformatics* 27.7 (Apr. 2011), pp. 1017–1018. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btr064. URL: http://dx.doi.org/10.1093/bioinformatics/btr064.

[37] Casey A Gifford et al. "Transcriptional and epigenetic dynamics during specification of human embryonic stem cells". en. In: *Cell* 153.5 (May 2013), pp. 1149–1163. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2013.04.037. URL: http://dx.doi.org/10.1016/j.cell.2013.04.037.

[38] Alexander M Tsankov et al. "Transcription factor binding dynamics during human ES cell differentiation". en. In: *Nature* 518.7539 (Feb. 2015), pp. 344–349. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14233. URL: http://dx.doi.org/10.1038/nature14233.

[39] Zuguang Gu, Roland Eils, and Matthias Schlesner. "Complex heatmaps reveal patterns and correlations in multidimensional genomic data". en. In: *Bioinformatics* 32.18 (Sept. 2016), pp. 2847–2849. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btw313. URL: http://dx.doi.org/10.1093/bioinformatics/btw313.

5.9 Supplementary Figures

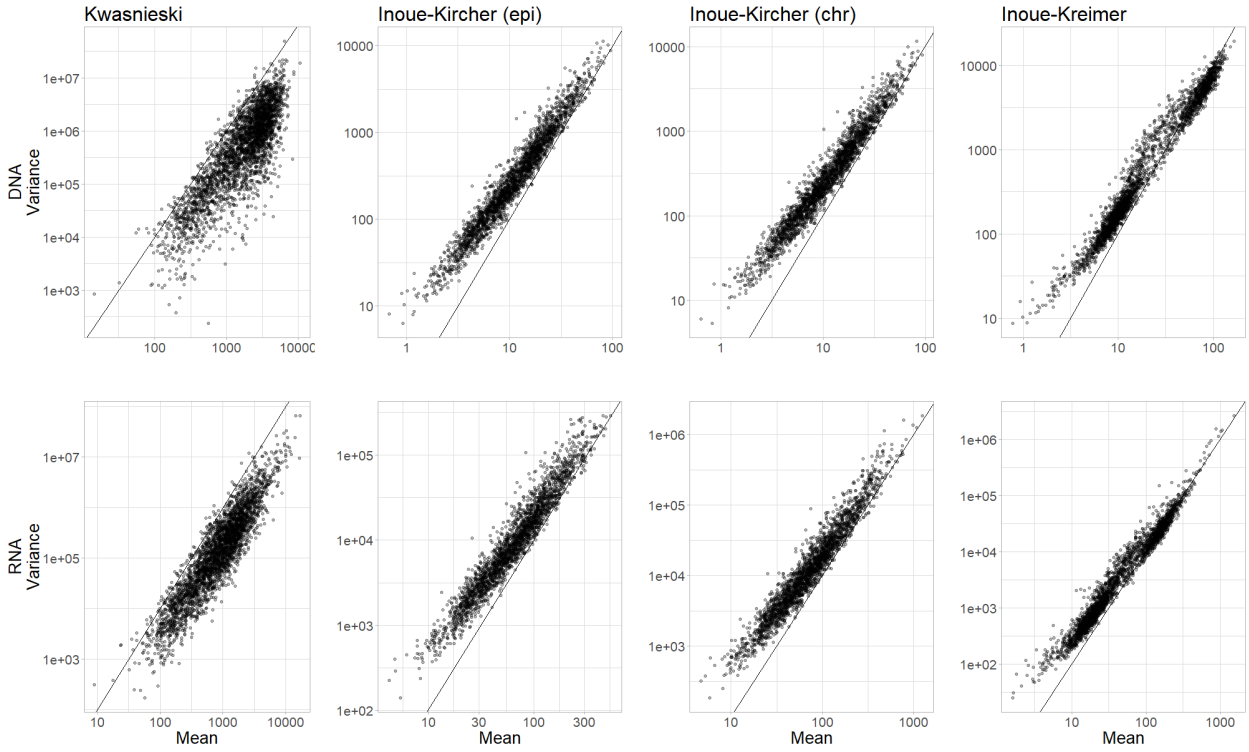


Figure S1: Relationship between the mean and variance of the counts measured for each sequence. Reference line (slope = 2) is a quadratic relationship.

Sample	Condition	Barcode
A	Ref	1
B	Ref	2
C	Ref	3
D	Contrast	1
E	Contrast	2
F	Contrast	3

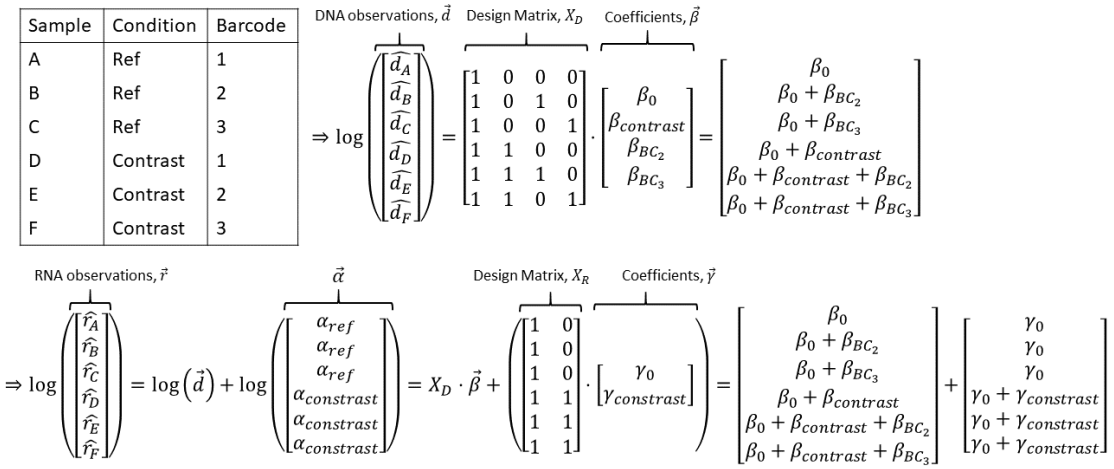


Figure S2: A simplified example of the MPRAnalyze model: two conditions are tested with three barcodes in a paired experiment (each DNA observation has a corresponding RNA observation). No replicates or external normalization factors are included in this design to maintain simplicity. The DNA's model estimation of the latent DNA count, computed as $X_D\vec{\beta}$, is included in the RNA model. The α estimates of transcription rate can be extracted from the model as: $\alpha_{\text{ref}} = e^{\gamma_0}$, $\alpha_{\text{ref}} = e^{\gamma_0+\gamma_{\text{contrast}}}$. Note that while modeling the barcodes in the RNA model is possible, the result will be a separate α estimator for each barcode, which is usually not desired. Barcode-level information is therefore only incorporated into the RNA model via the nested DNA model.

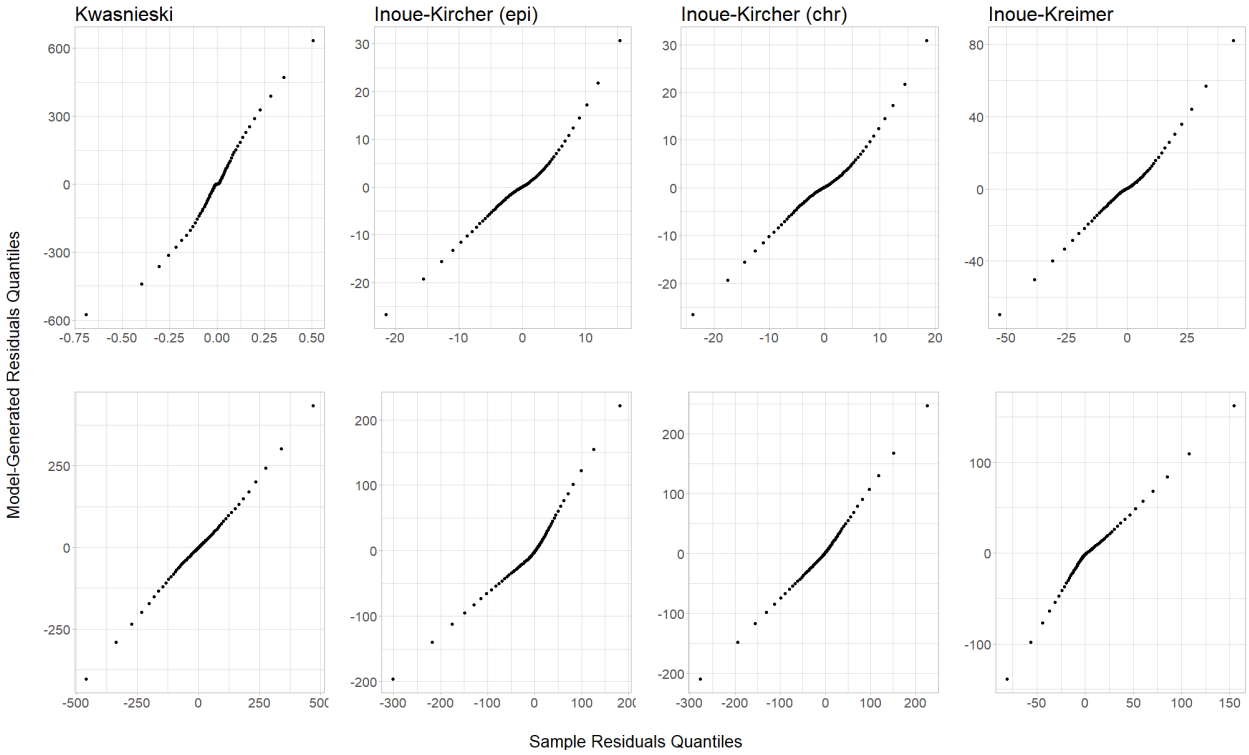


Figure S3: Comparison between model residuals from the observed counts of each dataset, and residuals from random data generated by Gamma (for DNA) and Negative Binomial (for RNA) using the model parameters. Quantile-quantile comparisons indicate that the observed noise and the generated noise follow similar distributions.

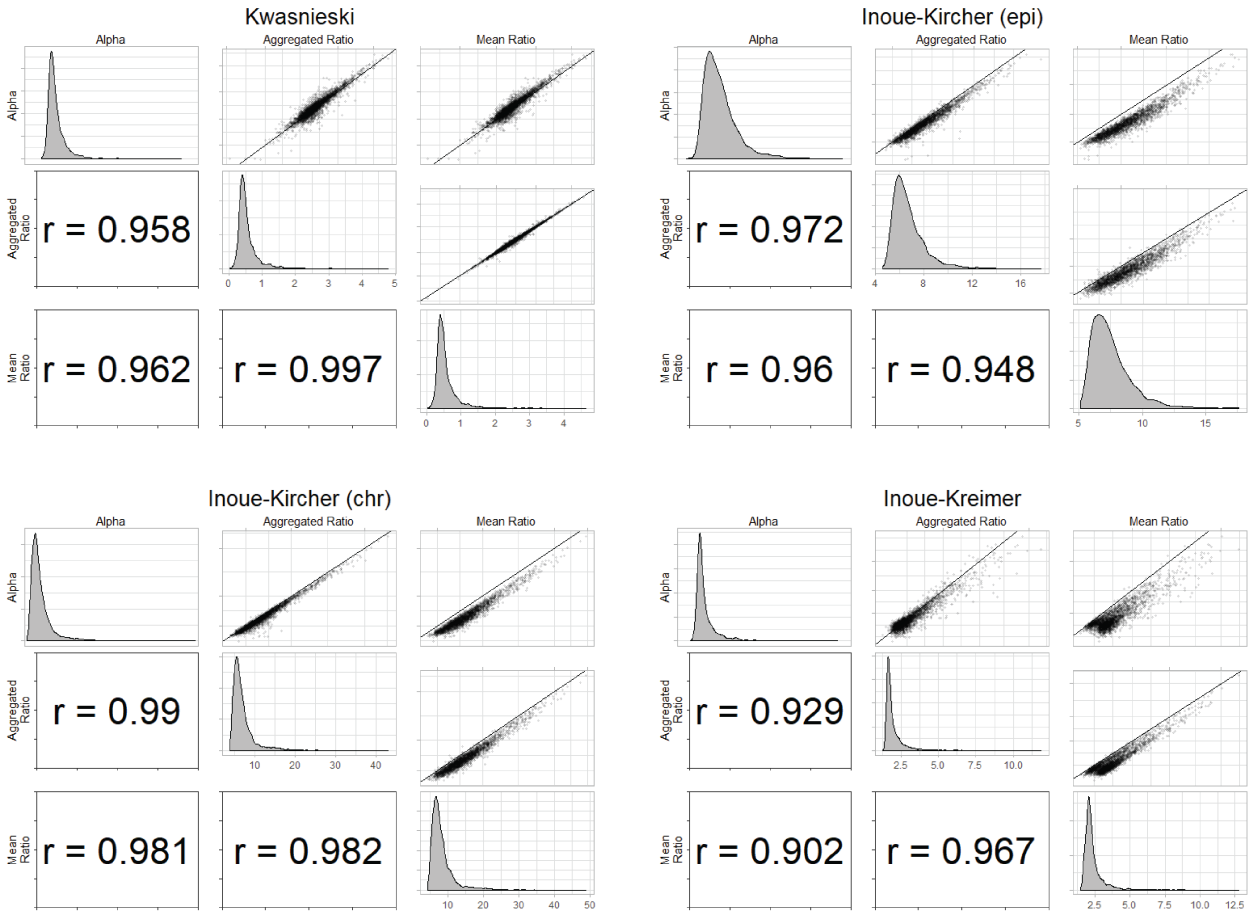


Figure S4: Correlations of MPRAnalyze's *alpha* estimate with the ratio-based estimates. Correlations are Pearson's r.

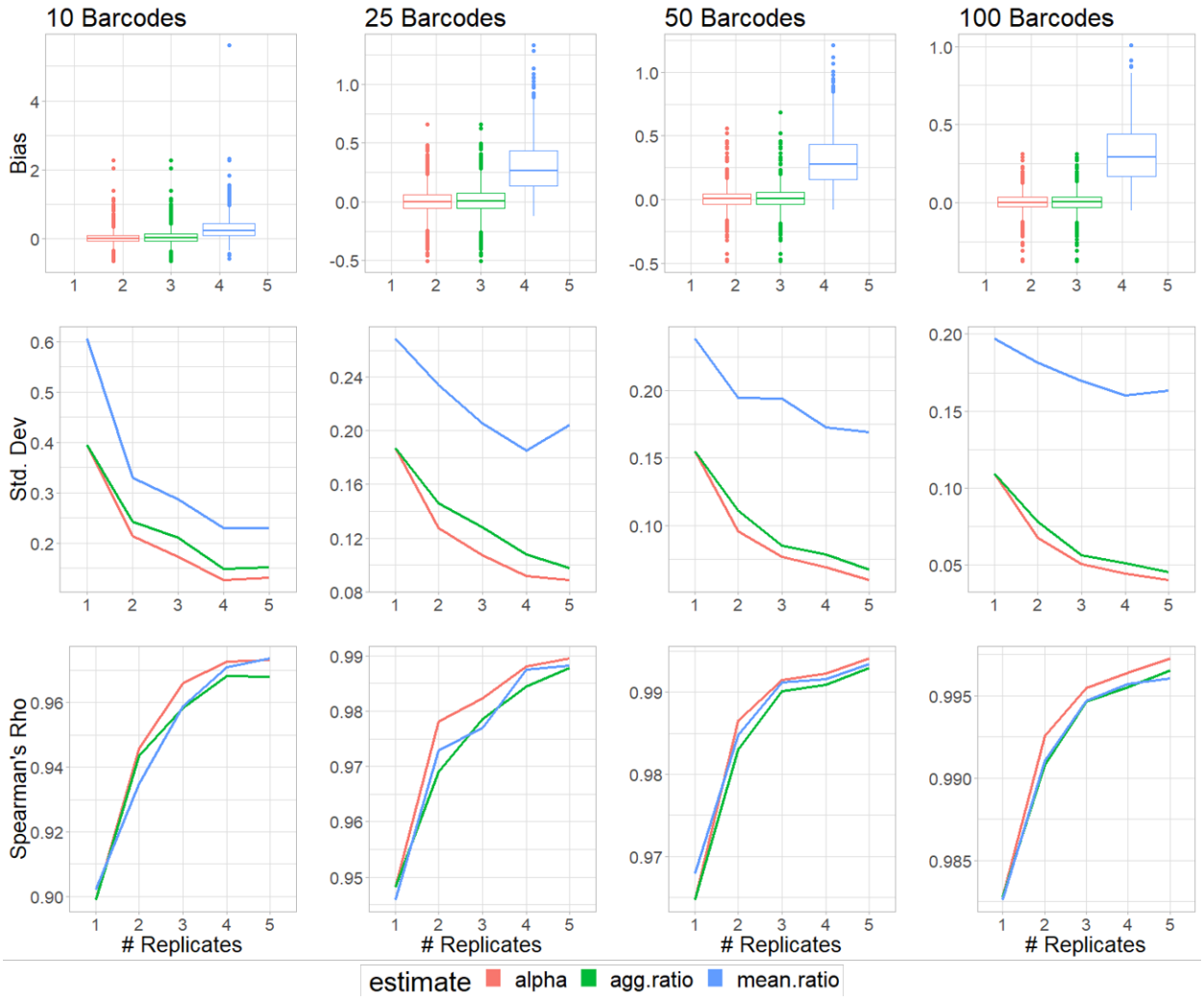


Figure S5: Performance evaluation of MPRAnalyze's α estimate and ratio-based estimates on simulated data with varying number of barcodes and replicates.

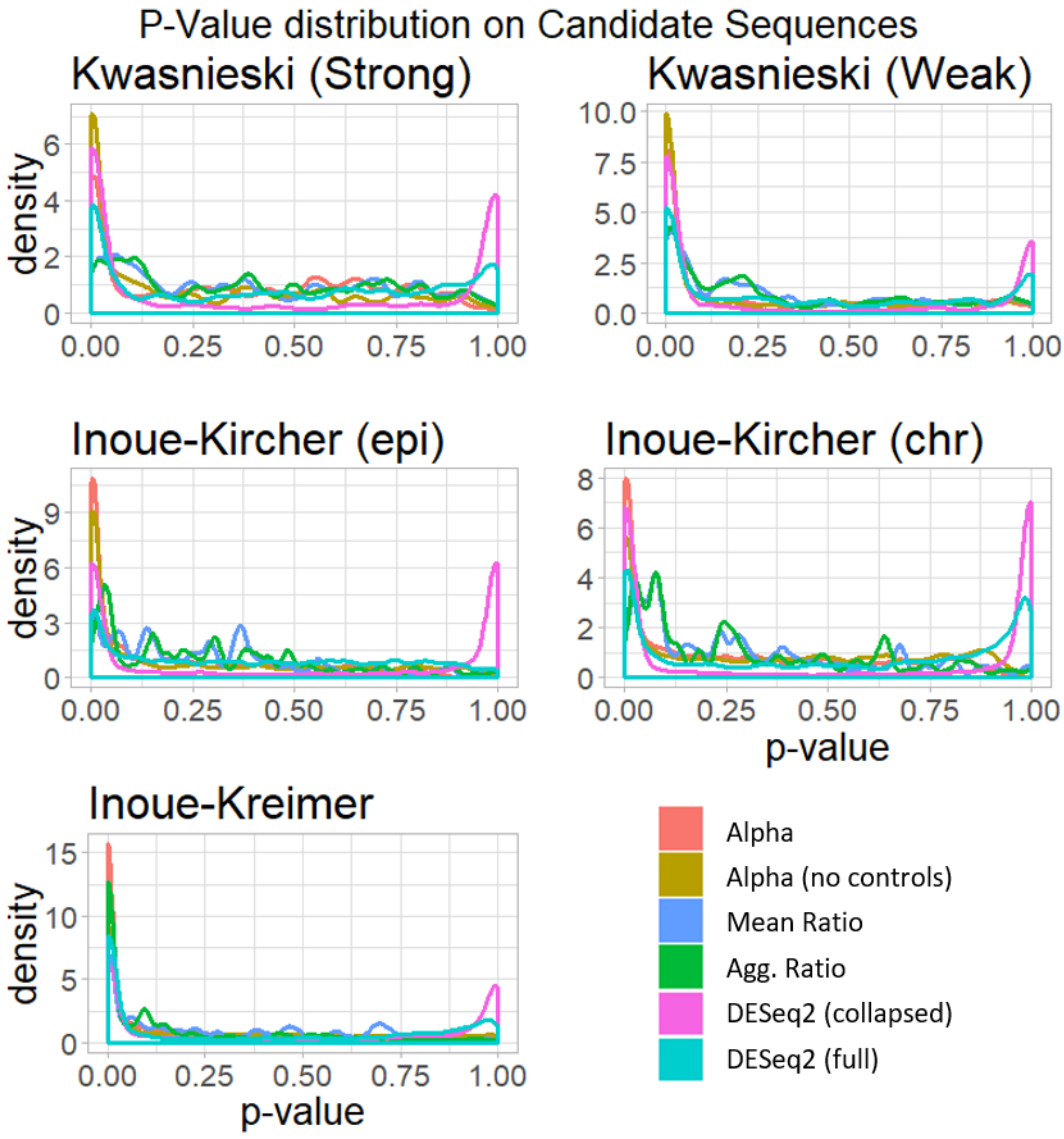


Figure S6: P-value density of classification analysis of candidate sequences for each dataset, by method of classification. Aside from DESeq2-collapsed, all methods seem to follow the theoretical distribution of a mixture of uniformly distributed values and low values.

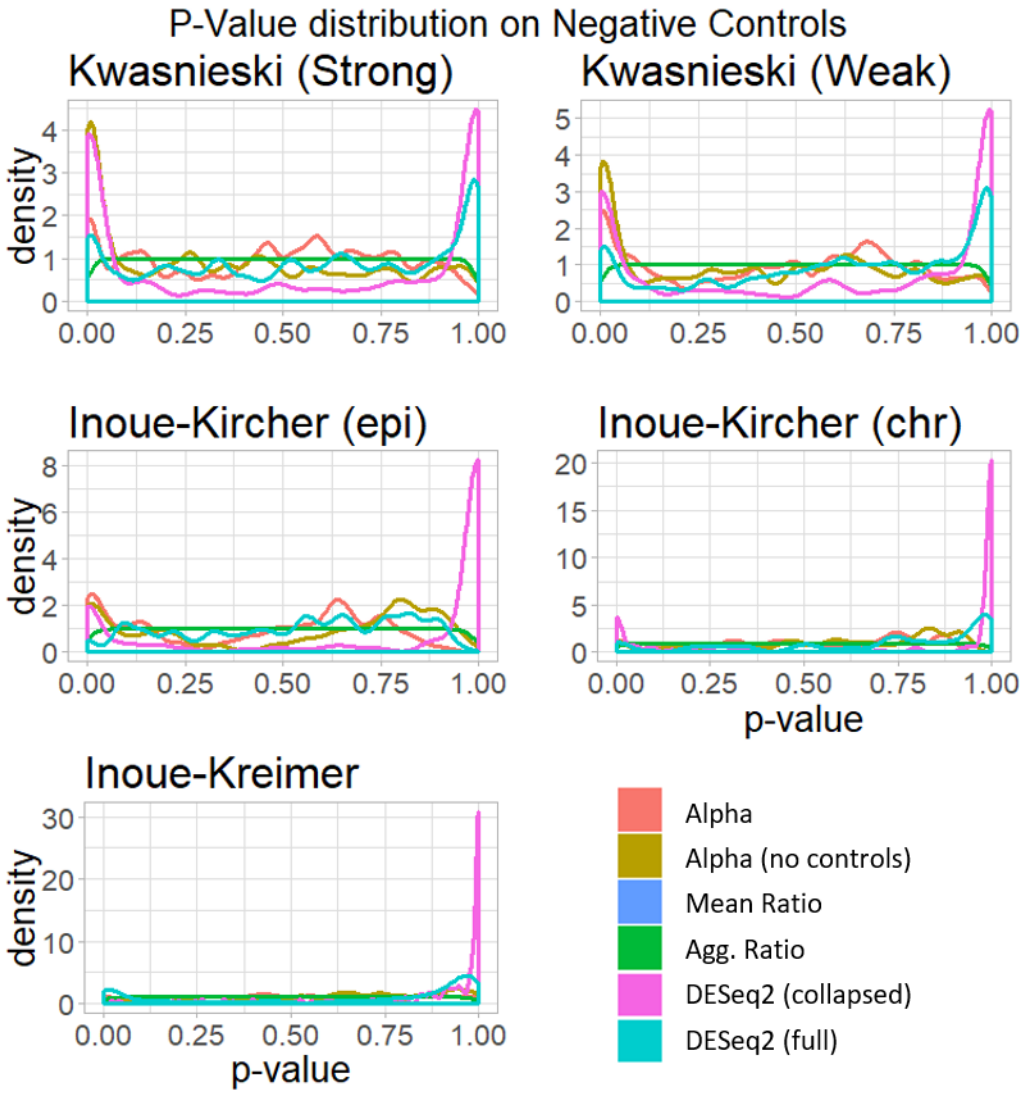


Figure S7: P-value density of classification analysis of negative control sequences for each dataset, by classification method. Empirical methods follow the theoretical uniform y definition, with the Mean Ratio line plotted behind the Agg. Ratio line. Some inflation can be observed with MPRAnalyze in no-controls mode, and DESeq2-collapsed is generally not calibrated.

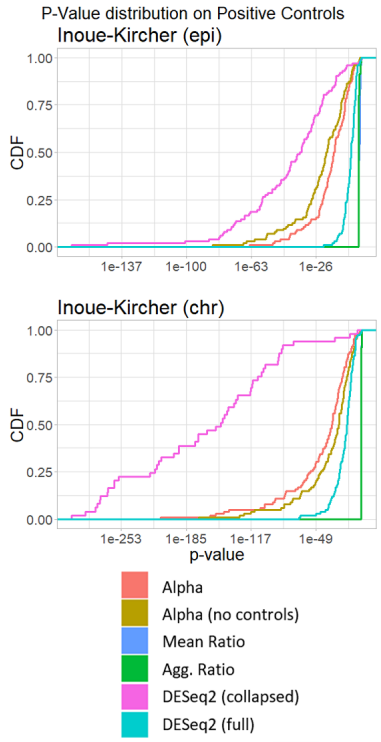


Figure S8: P-value CDF of classification analysis of positive control sequences for each dataset, by classification method, displayed in log scale for ease of visualization. Both modes of MPRAnalyze are substantially more powerful than competing methods, with the exception of DESeq2-collapsed.

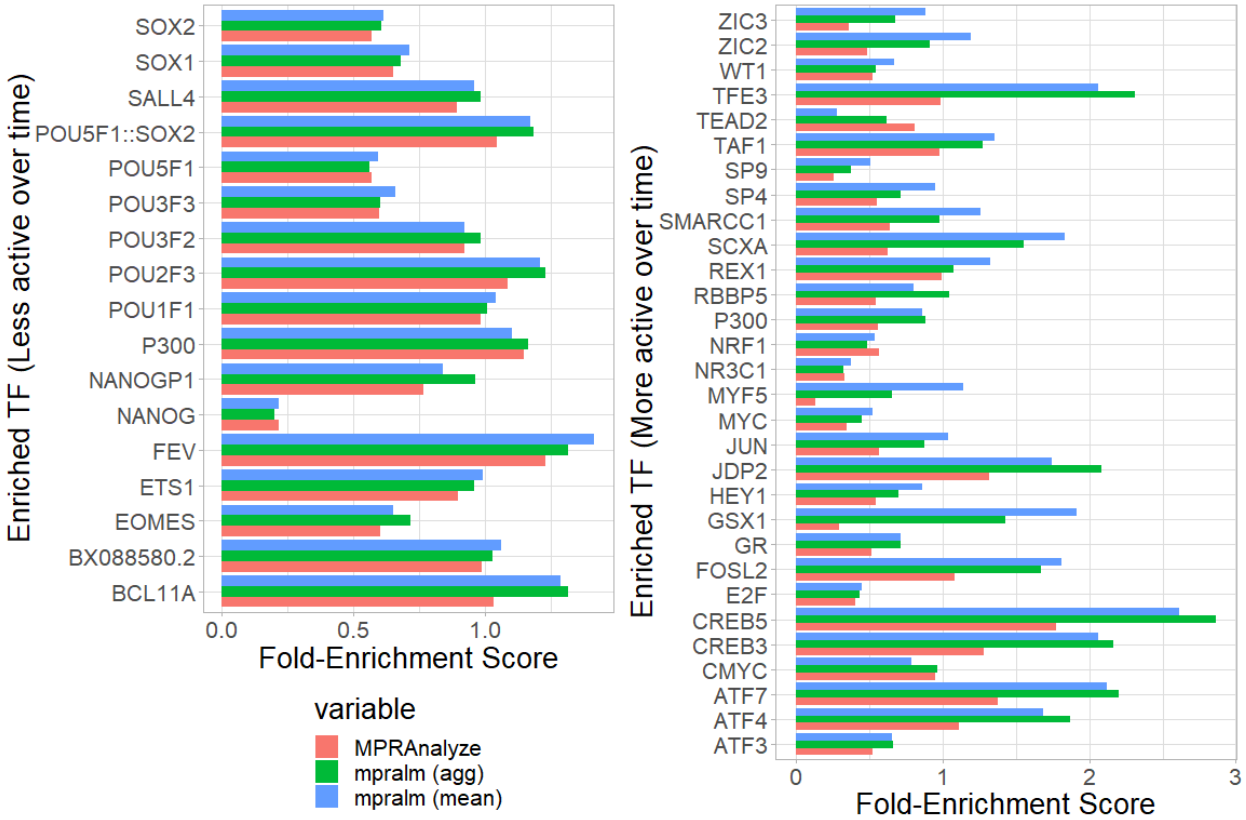


Figure S9: Enrichment score calculated for the set of top-enriched factors (union of top 15 most enriched factors for each method, ranked by statistical score). Enrichment score is calculated as the log2 of the ratio between the fraction of differential sequences that contain a binding motif of that factor, to the fraction of sequences in the entire assay that contain a binding motif for that factor.

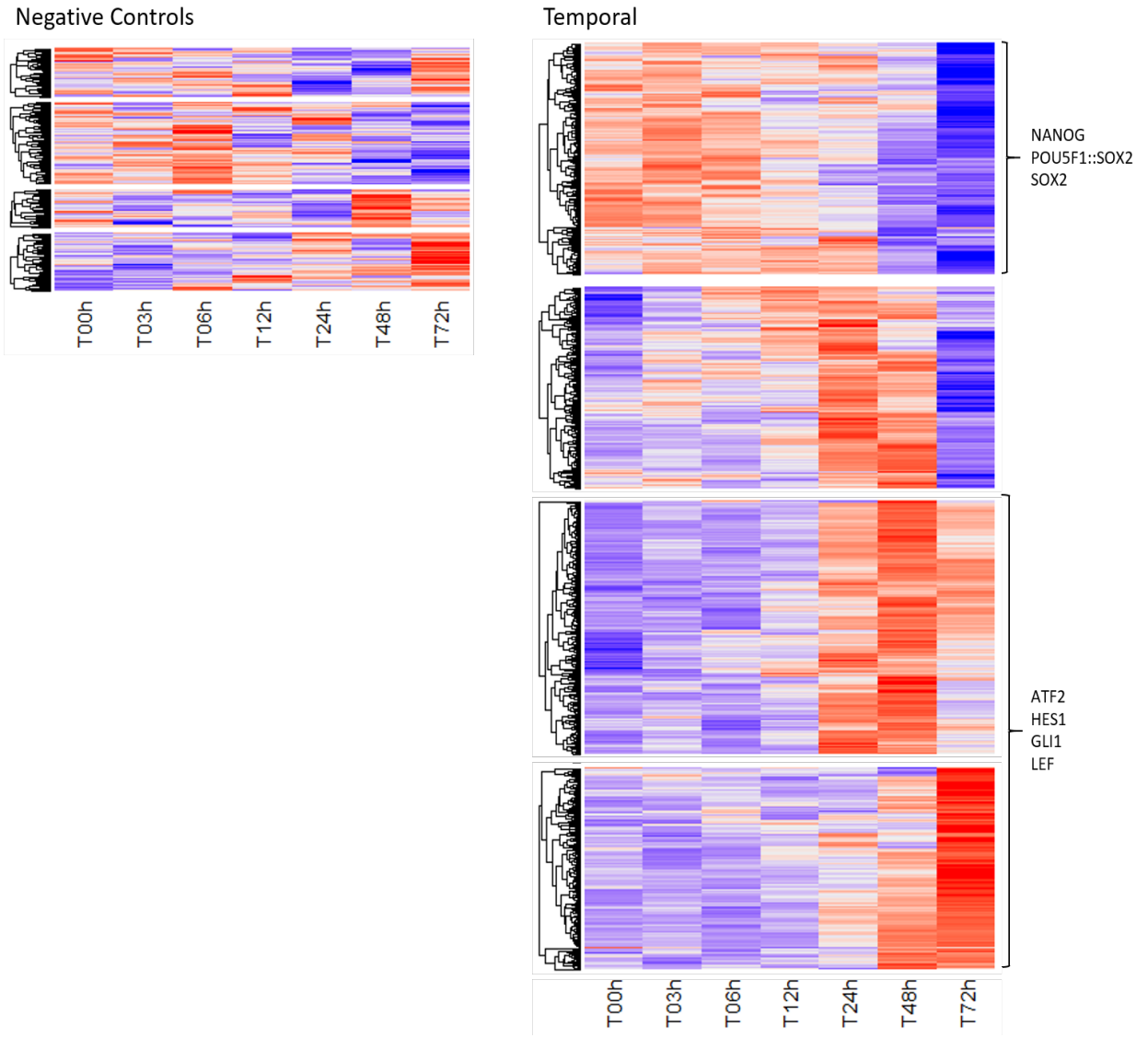


Figure S10: Heatmap of sequence activity (α values) in each time point, for the negative control sequences (left) and sequences that displayed significant temporal activity (right). Rows were z-noramlized for visualization, and each set was clustered using K-Means, K=4, before hierarchical clustering was performed for visualization (gaps indicate distinct clusters). Some enriched transcription factor binding sites are indicated.

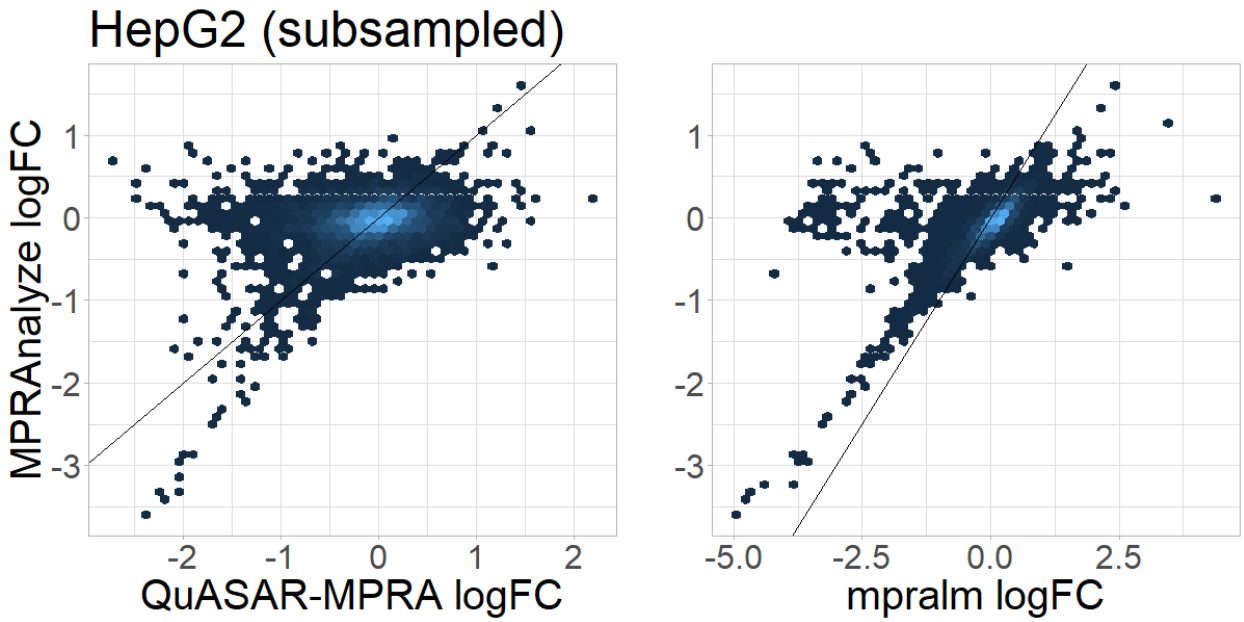


Figure S11: Comparison of log Fold-Change values from competing methods to MPRAnalyze in HepG2 data from the Mattioli study, using only the first 4 replicates.

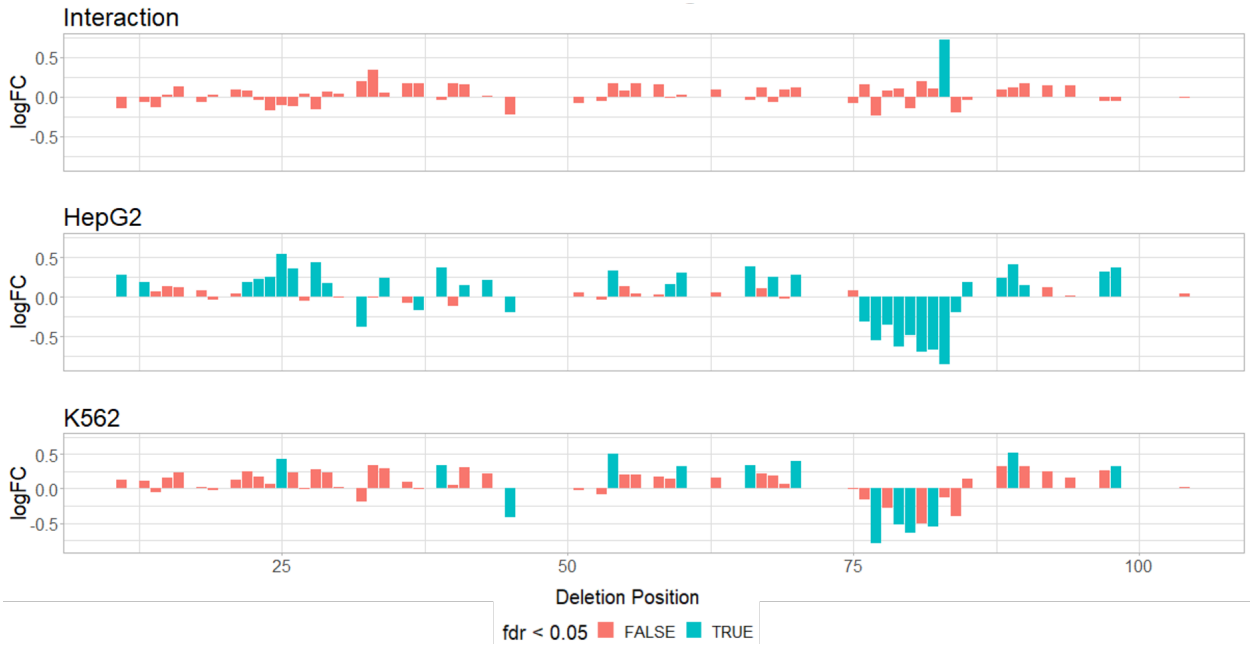


Figure S12: Effect of each single-nucleotide deletion in the core promoter of the lncRNA gene DLEU1. The differential effect size (top) captures the difference in the effect of each deletion in HepG2 (middle) compared with K562 (bottom).

5.10 Additional Files

All additional files for this chapter are included in *Chapter5_Additional_Files.zip*. The files are:

- **Additional File 2** Supplemental methods. Formal description of the MPRAnalyze model, hypothesis testing schemes and optimization details.
- **Additional File 3** Table S1. Transcription factor binding site enrichment analysis results: BH-corrected p values.
- **Additional File 4** Table S2. Transcription factor binding site enrichment analysis results: fold-enrichment scores.
- **Additional File 5** Table S3. Full RNA-seq measurements from the Inoue-Kreimer study of the timepoints T0h and T72h.
- **Additional File 6** Table S4. BH-corrected p values from the Transcription Factor finding site enrichment analysis of the temporal results.

Chapter 6

Identification and massively parallel characterization of regulatory elements driving neural induction

This chapter was published in *Cell Stem Cell* (2019), and is included here as published. The authors on the paper are:

Fumitaka Inoue[1,2,*], Anat Kreimer[1,2,3,*], Tal Ashuach[3], Nadav Ahituv[1,2,†], Nir Yosef[3,4,5,†]

1. Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, 94158, USA
2. Institute for Human Genetics, University of California San Francisco, San Francisco, California, 94158, USA.
3. Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, California USA
4. Chan-Zuckerberg Biohub, San Francisco, CA 94158, USA

∗ these authors contributed equally to the work

† Corresponding author

6.1 Abstract

Epigenomic regulation and lineage-specific gene expression act in concert to drive cellular differentiation, but the temporal interplay between these processes are largely unknown. Using neural induction from human pluripotent stem cells (hPSCs) as a paradigm, we interrogated these dynamics by performing RNA-seq, ChIP-seq, and ATAC-seq at seven time points during early neural differentiation. We found that changes in DNA accessibility precede H3K27ac, which is followed by gene expression changes. Using massively parallel reporter assays (MPRAs) to test activity of 2,464 candidate regulatory sequences at all seven time points, we show many of these sequences have temporal activity patterns that correlate with their respective cell-endogenous gene expression and chromatin changes. A prioritization method incorporating all genomic and MPRA data further identified key transcription factors involved in driving neural fate. These results provide a comprehensive resource of genes and regulatory elements that orchestrate neural induction and illuminate temporal frameworks during differentiation.

6.2 Introduction

Global changes in gene expression are an essential part of cellular differentiation [1]. To date, many genome-scale maps of epigenetic properties in progenitor and differentiated cells have been used in comparative studies, demonstrating the importance of modifications of the epigenome to the pertaining changes in gene expression and shedding light on the mechanisms involved in this process [2, 3, 4]. For instance, in human embryonic stem cells (ESCs), the regulatory regions marked by histone modifications and binding of key regulators associated with gene expression were globally reorganized in accordance with multilineage differentiation [5, 6, 7, 8]. However, the majority of these studies provide descriptive genome-wide maps without large-scale functional analyses of candidate sequences. Furthermore, while a few studies used functional validation following large scale genomic studies [9, 10, 11, 12], these studies did not focus on differentiation processes.

The differentiation of human embryonic stem cells (hESCs) into neural cells provides an exceptional model to study this. During early neural induction, the cells exhibit marked changes in gene expression as pluripotency-associated genes are rapidly downregulated and neural associated genes are induced. These changes are then maintained for a duration of several weeks, until the establishment of a neural progenitor cells (NPCs) population [13]. Several large-scale mapping efforts have characterized in a genome-wide manner the transcriptional and epigenetic landscape of hESC-derived NPCs or neural tissues and have annotated numerous genes and potential regulatory elements that could be important in neural differentiation [2, 4, 5, 14, 6, 7, 8]. However, while these studies have identified putative regulatory elements, they have not comprehensively analyzed them for their function. Furthermore, none of these genomic studies focused on the early stages of neural differentiation when neural induction takes place. Thus, the intrinsic mechanism that governs neural in-

duction remains largely unknown.

The differentiation of hESCs to neuronal cells also provides an important model system for studying the etiology of neurodevelopmental diseases. Mutations in genes and regulatory elements involved in neural induction and development have been associated with numerous human diseases. For example, dysfunction of cortical GABA neurons in schizophrenia begins during prenatal development [15]. Similarly, autism spectrum disorders (ASD) are associated with de novo mutations in developmental genes [16] and alterations in canonical Wnt signaling in developing embryos [17]. In addition, the majority of disease-risk loci discovered through genome-wide association studies (GWAS) in general and specifically for neuropsychiatric and neurodevelopmental disorders reside in noncoding regions [18, 19, 20], suggesting an important role for enhancers in disease susceptibility.

Here, we set out to generate a genomic map of the transcriptional (RNA-seq) and epigenetic landscape (H3K27ac/me3 ChIP-seq and ATAC-seq) of neural induction and then coupled these observations with comprehensive functional assays (MPRA). We integrated all of the resulting data modalities (genomics maps and MPRA) to computationally infer the activity of transcription factors (TFs) over time and characterize candidate TFs that could be important drivers of neural induction. Our work provides a comprehensive resource of genes and regulatory elements and a blueprint for the interplay between them during differentiation.

6.3 Results

The neural induction-associated transcriptome

We performed deep RNA sequencing (average of 200 million reads per replicate) on undifferentiated H1-ESCs (0 hour) and six different time points of early neural differentiation (3, 6, 12, 24, 48, and 72 hours) following dual-Smad inhibition [21]. Principal component analysis (PCA) of the RNA-seq data showed consistency between the three replicates and a clear separation between the earlier and later time points (Figure S1A). As expected, we observed neural marker genes, such as SOX1, to be upregulated after 12 hours (Figure 1A), with limited expression changes in mesendoderm (EOMES), mesoderm (T and TBX6), endoderm (SOX17 and GATA4), and neural crest markers (FOXD3 and SNAI1/2). Pluripotent markers (NANOG, POU5F1) and direct targets of TGFβ and BMP signaling (SMAD7, ID1, LEFTY2) were downregulated and immediate early genes (ATF3, FOS, FOSB and EGR1/2/3) were transiently upregulated at 3 hours, corresponding to the cell's stress response against differentiation stimuli. For a more general analysis, we used a conservative approach to identify genes whose expression differed significantly over time, using a consensus over two methods - ImpulseDE [22] and DESeq2 [23]. Altogether, we detected 2,172 genes as differentially expressed over time (henceforth referred to as temporal genes), with 85% of them being induced at some point in time (Figure 1B; the remaining genes show monotonic decrease of expression). Gene set enrichment analysis [24] of the resulting clus-

ters of temporal profiles found that genes that are more strongly expressed at the early time points (0-12 hours, False Discovery Rate (FDR)< 0.05; hypergeometric test) are enriched for regulation of multicellular organismal development, indicating an association with pluripotency. Conversely, genes induced at later time points (> 24 hours, FDR< 0.05) are enriched for neurogenesis processes, consistent with the progression of the cells toward a neural lineage fate (Table S1). Combined, our transcriptomic analyses validated the ability of the dual-Smad inhibition protocol to obtain the expected neural trajectory and provides a catalog of genes involved in neural induction.

The neural induction-associated regulome

To identify candidate enhancers involved in the differentiation process that could be driving neural induction, we performed ATAC-seq as well as ChIP-seq for the active histone mark H3K27ac and the silencing mark H3K27me3 at all seven time points. We then identified regions that are enriched (i.e., peak regions) in each of these assays ([25]; FDR< 0.05) by analyzing each time point separately and then taking the merged set of peaks over all time points. To establish peak calling quality, we compared our 0hr time point H3K27ac and ATAC-seq peaks to H1-ESC H3K27ac peaks and DNase I hypersensitive sites (DHS) from ENCODE [4] and observed a substantial overlap of 80% and 90% respectively. Overall, we identified 40,486 ATAC-seq peaks, 40,170 H3K27ac peaks and 4,446 H3K27me3 peaks that are present in at least one time point. To exclude potentially inactive regions from further analysis, we removed H3K27ac peaks that overlap a H3K27me3 peak at all the time points in which that peak was detected. This resulted in a filtered set of 40,042 H3K27ac peaks, indicating that the two chromatin marks have little overlap in our data. Conversely, we observed a substantial overlap between the H3K27ac peaks and the ATAC-seq peaks, with an overall 60% (23,294) of the H3K27ac peaks overlapping an accessible region at the same point in time. These results correspond to a similar overlap of 61% between H3K27ac peaks and DHS in H1-ESC from ENCODE [4]. Using a strict procedure, similar to the gene expression analysis [23, 22], we found 2,435 ATAC-seq and 2,024 H3K27ac peaks that were differentially enriched between time points, henceforth referred to as temporal H3K27ac or ATAC-seq peaks (Table S1). Similar analysis of H3K27me3 peaks showed weaker temporal signal (STAR Methods) with a smaller number of 248 temporal peaks (Figure S1B), possibly due to histone methylation being less dynamic than acetylation [26, 27, 28, 29].

We next set out to study the association between the temporal changes observed at the epigenome level, and those observed at the gene expression level. We clustered the two sets of temporal regions (in terms of accessibility and H3K27ac) into several prototypical patterns (Figure 1C, D) as we have done for the temporal genes (Figure 1B). Functional enrichment analysis (using GREAT [30] with FDR< 0.05) on the accessibility and H3K27ac clusters was overall consistent with the results observed with the gene expression clusters, with an enrichment for pluripotent factors and nervous system development processes in early and late response regions respectively (Table S1). Interestingly, the observed temporal changes of accessibility, histone acetylation and proximal gene expression, were highly correlated to

each other (Figures 1E, S1C; exact overlaps are displayed in Table S1). Furthermore, for a significant fraction of the genes induced at the late stages (RNA-seq clusters 4-6), chromatin accessibility was found to be acquired first (ATAC-seq clusters 4-5) or simultaneously (ATAC-seq clusters 6) with H3K27ac modification followed by an increase in mRNA expression of the nearest gene (Figure 1E), while at early stages (RNA-seq clusters 1-3) this was a less obvious trend. For example, the DNA accessibility cluster 4 that peaks at 24 hours showed the strongest overlap with H3K27ac clusters 5-6 that peak at 48-72 hours and this cluster significantly overlaps (in terms of genes; p-value$<$ 0.0014; hypergeometric test) with gene expression cluster 6 which peaks at 72 hours (Figure 1E). Specifically, examination of potential enhancers within these clusters that are located near neural marker genes, MAP2 [31] and ROR2 [32], found them to be enriched for ATAC-seq signal at 12-24 hours, H3K27ac signal at 48-72 hours and their expression to peak at 72 hours (Figure S2A, B). Combined, these results suggest that regions that are associated with changes to chromatin structure during neural induction are statistically related to changes in gene expression.

Neurological disorder associated variants are enriched in temporal H3K27ac peaks

As genes and regulatory elements involved in neural development may be associated with neurological disorders, we tested whether our neural induction regulome overlaps with disease-associated variants. We first tested whether the temporal loci (in terms of accessibility or H3K27ac) are enriched for GWAS variants associated with neurological disorders. To this end, we used the complete set of peaks (temporal and non-temporal) as background and added variants associated with height as negative controls. We observed a significant enrichment for H3K27ac (but not accessibility) temporal peaks with neurological disorders (Table S1; STAR Methods; p-value$<$ 0.05; Fisher's exact test) but not with height variants. Specifically, we observe significant enrichment when examining variants associated with a combined set of neuropsychological disorders (schizophrenia, attention-deficit hyperactivity disorder (ADHD), ASD, bipolar disorder and major depressive disorder) as well as enrichment when examining for individual disorders (i.e. Bipolar and Psychosis disorders). As the smaller size of ATAC-seq peaks might account for the lack of enrichment in ATAC-seq temporal peaks, we expanded the ATAC-seq peaks to the average size of H3K27ac peaks, but observed similar results.

Expression quantitative trait loci (eQTLs) mark variants that can be associated with modulating the regulation of nearby genes. We tested for overlap between eQTLs found in various tissues [33] and our temporal ATAC-seq or H3K27ac peaks. We found the temporal H3K27ac peaks to be significantly enriched for eQTL variants [34] in general and specifically for those from brain tissues [33] (Table S1; STAR Methods; p-value$<$ 0.05; Fisher's exact test). Similarly to GWAS variants, we did not observe an enrichment of eQTLs in temporal ATAC-seq peaks even upon their expansion. When restricting H3K27ac peaks to not overlap with H3K27ac ChIP-seq peaks obtained from different cell types (GM12878, K562 and

HepG2) via the ENCODE project [4], we observe similar results, indicating that our signal is not biased by constitutive peaks. When restricting variant enrichment analysis to H3K27ac temporal peaks and assessing the enrichment in each temporal cluster (Figure 1C), we observe that the late response cluster 6 is significantly enriched in nervous system disease and specifically with ASD-associated variants (Fisher's exact test; p-value< 0.005). Combined, these results suggest that our temporal H3K27ac regions could be functional enhancers that harbor neurological disease risk variants. They also suggest that temporal changes to the chromatin early in the differentiation process can facilitate the identification of potentially functional regions more so than data from a single time point.

LentiMPRA identifies regulatory regions that are active during neural induction

In order to test whether our candidate regulatory sequences can in fact induce temporal transcriptional response, we carried out lentiMPRA at all seven time points. Overall, we investigated 2,464 candidate sequences, covering both promoters (N=386 (15.7%)) and putative enhancers (N=2,078 (84.3%)), termed henceforth as candidate regulatory sequences (CRS). As the number of potential CRS is large, we developed a prioritization scheme to select the set of assayed regions, (Figure 2A; Table S2; STAR Methods) using the following criteria: 1) Manually curated list of enhancers that are next to genes involved in neural differentiation (N=102; Table S2); 2) Sequences that overlap a temporal H3K27ac ChIP-seq peak that also overlap an ATAC-seq peak (not necessarily temporal) and that their closest gene shows increased expression due to neural induction (N=1596); 3) Sequences that overlap non-temporal H3K27ac peaks and temporal ATAC-seq peaks and their closest gene shows increased expression due to neural induction (N=441); 4) Among the regions not included in the first three groups, we select sequences that showed the strongest difference in signal of either H3K27ac ChIP-seq, ATAC-seq or mRNA of the closest genes (N=132; comparing either 0 vs. 3 hours or 0 vs. 72 hours); and 5) Positive control sequences (N=193) that included previously reported sequences that were validated forebrain enhancers in the VISTA Enhancer Browser [35] (N=105), sequences near pluripotent factors (N=42) and commonly used positive controls from the ENCODE project (N=46) (Table S2). For negative controls, we randomly selected 200 of our candidate sequences and shuffled their nucleotides obtaining scrambled sequences. Overall, we chose 2,664 sequences using this process. As our assayed sequences were 171 base pairs (bp) long, due to oligonucleotide synthesis limitations, we chose the 171bp window within a peak of interest by maximizing the number of motifs in it [36, 37].

The selected oligonucleotides were generated and cloned upstream of a minimal promoter (mP) and EGFP reporter gene into a lentivirus-based enhancer assay vector (Figure 2B) as previously described [38]. While the sequences are assayed in the context of enhancer activity in this assay, previous work has shown that it also provides a good indication for promoter activity [39, 40, 41]. Each individual CRS was designed to be associated with 90 different

15-bp barcodes, thus allowing robust evaluation of the pertaining expression output and to correct for site of integration biases [42, 38]. In total 239,760 sequences (2,664 CRS x 90 barcodes) were included in the library (Figure 2B). The cloned library was sequenced in order to evaluate the quality of the designed oligonucleotides and the representation of individual barcodes (STAR Methods; Figure S3A-D).

hESCs were infected with the library with an average of 5-8 integrations per cell (Figure S3E), cultured for 3 days to clean out for unintegrated lentivirus and then subsequently induced into a neural lineage via dual-Smad inhibition. LentiMPRA was performed at all seven time points of neural differentiation with three replicates (two biological replicates, one of which was split into two technical replicates; Table S3). Due to the short time spans between some conditions, we collected nuclear RNA in all time points to detect their immediate expression. We observed an average of 70 barcodes out of 90 per CRS in each replicate (Table S3). By aggregating these barcodes (STAR Methods), we were able to get highly reproducible results across replicates (Figures S3F) with similar magnitude to a previously characterized lentiMPRA in another cell type [38]. We then combined replicates to produce a normalized RNA/DNA ratio for each CRS (henceforth referred to as MPRA signal). Examination of the signal observed for regions nominated by the different experimental design criteria found that temporal H3K27ac signal (criterion 2) provides the most effective predictor of functional enhancer activity, while as expected, the negative controls showed the lowest activity (Figures 3A, S3G).

LentiMPRA identifies temporal CRS

We next set out to examine whether the enhancer activity observed in our assay changes over time, and then characterize these changes with respect to the cell-endogenous temporal processes depicted in Figure 1. As a starting point, we considered each time point separately and applied MPRAnalyze [42], a method and Bioconductor package for statistical analysis of MPRA data developed in our group, to identify active enhancers, namely enhancers whose activity significantly deviates from that of the negative controls (median-based z-score; FDR$<$ 0.05; STAR Methods). Of note, the dynamic ranges we observed were comparable to a previous library generated in a similar manner [38]. From the 2,464 CRS that we tested via lentiMPRA, 1,681 (68%) were called significant in at least one of the time points and on average 1,141 (46.3%) sequences were active per each individual time point.

While we saw similar levels of activity at each time point, the respective sets of active CRS may differ greatly over time. Reassuringly, we observed substantial overlaps between the sets of active CRS at nearby time points (Figure 3B), along with a marked decrease in overlap as the distance between the respective time points increases (Figures 3C). This indicated that regulatory programs carried out by enhancers are far from fixed, but instead change over the course of neural induction. As an example, we observed that a known enhancer of NANOG as well as its promoter [43, 44] both have activity only at the early time points (Figure 3D), as expected. We also found novel enhancers near SOX1 that showed increased temporal activity at 24-48 hours, while being less active at 72 hours and further away (140kb) an enhancer

that has strong enhancer activity at 72 hours, suggesting a complex temporal regulation pattern for this gene (Figure 3E). To validate our temporal observations with lentiMPRA, we individually tested five ESC enhancers, four immediate response enhancers (12-24 hours), and four NPC enhancers (48-72 hours) using luciferase assays. We observed the expected temporal activities for these sequences, that were consistent with our MPRA results (Figure 4A-B). As an additional validation, we used CRISPR activation (CRIPSRa) [45] to target three CRS detected in our study at the SOX1, IRX3 and OTX2 loci in hESCs. We found that all three CRS upregulated the expression of their predicted target gene (SOX1, IRX3 and OTX2) following CRISPRa (Figure 4C-F), further suggesting that the enhancers we identified are functional and can affect gene expression.

We next carried out a more global analysis that aims to identify enhancers whose MPRA signal significantly changed over time [42]. This alternative approach pools together information from all time points, rather than considering each time point individually, and therefore has the potential to identify effects that may otherwise be missed. In this analysis, the temporal activity of each CRS was compared with a null temporal behavior displayed by the set of negative controls. Regions with significantly different temporal activity were called temporally active using a likelihood ratio test (FDR$<$ 0.05; [42]). We found that 1,547 sequences out of the 2,464 we tested (63%) showed temporal regulatory activity (henceforth referred to as temporal CRS). Out of these temporal CRS 1,261 (82%) were also detected by the per- time point analysis. In the following analyses, we focused on the complete set of temporal CRS. Importantly, we observed consistent results when limiting our analyses to the smaller and more stringent set of 1,261 regions.

Comparative analysis of temporal versus non-temporal CRS for differences in transcription factor (TF) binding motifs [37] found an enrichment for pluripotency-related regulators, such as POU5F1, SOX2, SALL4, NANOG and SMAD1 (largely targeting CRS with a marked decrease in activity over time), as well as the NPC-associated TF, SOX1 (largely targeting CRS with a marked increase in activity over time; Figure 3F). Of note, previous reports have shown that SOX2-POU5F1 and SOX2-POU3F2 regulate ESC and NPC genes, respectively [46], suggesting that SOX and POU motifs not only function in a pluripotent state, but also in a neural state. We also observed an enrichment for immediate early response factors (e.g., AP-1, ATF3, EGR3), corresponding to the cell's response to differentiation stimuli. Finally, we found that regulators of chromatin conformation including regulators of histone acetylation (EP300 and HDAC) and chromatin boundary and looping (CTCF), were also enriched in temporal CRS, indicating that changes in activity over time may also be mediated by a more direct regulation of the epigenome and not only by state-specific TFs.

TF binding site analyses identifies important neural induction genes

As the RNA product of MPRA is non-endogenous, it provides an effective way for directly estimating the effects of TFs on transcription. We utilized this property to pinpoint which

TFs could be driving neural induction at the different time points. To this end, we used experimental data from the public domain along with DNA binding motifs to determine the potential binding landscape of a large cohort of TFs across our tested regions. We recorded, for each temporal CRS: 1) its predicted binding sites using Fimo [36] with two sets of TF motifs [37, 47] and 2) its overlap with TF ChIP-seq peaks in hESCs [6] or in hESC-derived neuroectoderm [7]. The result of this analysis is a binary binding matrix of TFs by CRS with entries indicating either potential binding using FDR$< 10^{-}4$ for TF motifs or overlap with TF ChIP-seq peaks.

We next employed a strict enrichment analysis based on comparing the number of putative binding sites in regions within each temporal MPRA cluster versus the set of all regions in our MPRA design (FDR$<$ 0.05, Hypergeometric test; Table S5). This analysis was designed to nominate candidate TFs whose activity is specific to certain phases of the differentiation process. Accordingly, we found that motifs of pluripotent factors (e.g. NANOG, POU5F1, SOX2 [48]), were enriched in the early cluster. Furthermore, immediate early response factors (ATF, JUN, FOS), were enriched in mid-early enhancers (Table S5). These observations suggest that early- and mid-early clusters may respond to TFs that function in pluripotency maintenance and the cell's acute response, such as apoptosis [49], respectively. We also found that both mid-late and late clusters were enriched for cell fate commitment and specification factor binding. Specifically, SOX, OTX, and Class III POU factor motifs were enriched in both mid-late and late enhancers, suggesting that enhancers in these group were the direct targets of these key neural factors (Table S5).

Activity score identifies novel TFs that are important for neural induction

To narrow down the list of candidate TFs for a follow up investigation of their effect on neural induction, we defined a TF activity score, which represents the potential to affect transcription at each time point (STAR Methods). We considered two factors that influence this score at each time point: 1) The extent of deviation from the null- expected amount of active enhancers at that time point that are potentially bound by the TF [50], suggesting that this TFs may provide a parsimonious explanation for the MPRA signal [51]; and 2) An added requirement that the mRNA that codes for the TF is induced compared to previous time points, which may also suggest functional importance [52, 53, 54]. For the former, we focused our attention to enhancers in which the temporal MPRA pattern significantly overlaps with the endogenous patterns, namely – those CRS pertaining to significant entries in Figure 5E. Each of these 'consistent' entries represents a certain mode of temporal relationship between MPRA and the endogenous genome - e.g., early induction with matched timing of mRNA expression, or late induction that appears after the establishment of chromatin accessibility. While other CRS in our data can be of additional interest, we postulate that focusing on temporal regions that conform with the major patterns of overlap with the endogenous processes is desirable when integrating additional genomic readouts (TF binding potential

in this case), and may also increase the odds that the respective endogenous region is indeed functional.

The resulting activity matrix (Figure 6A; Table S5) provided a catalog of 107 TFs that could potentially function as regulators of neural induction. Repeating this analysis with the stricter set of temporal regions that were also detected by the per- time point analysis yielded largely similar results (94 out of 107 cataloged TFs were detected). Similar to previous analyses, we clustered the TF activity score to four representative patterns of activity: early, mid-early, mid-late and late response, and ranked it by the strength of induction of the respective TF's mRNA expression and the extent of overlap between TF's targets and the significant sub clusters of MPRA activity. Overall, we observed an agreement between known hESC and neural induction-associated TFs and their temporal time points. For example, in the early cluster, the pluripotent marker NANOG showed high TF activity score at 0 hour, and immediate-early gene products, ATF3, MYC and EGR1, showed high score at 3 hours, as expected [49]. TFs that had a high score at later time points (24-72 hours) included several neural TFs, such as SOX1, OTX2 and PAX6.

Overexpression and CRISPRi identify novel neural induction associated TFs

To test whether our identified TFs are indeed involved in neural induction, we selected for follow up overexpression studies 26 highly ranked TFs that were predicted as active during different time points of the induction process; top six in mid-early (FOXL2, BACH2, NR3C1, SMAD1, ELF3, HOMEZ, primarily active at 12-24 hours) and top ten in mid-late (SOX1, NFE2, OTX2, SP5, MAF, ID4, TCF7L2, IRX3, SMAD4, SOX2, 24-48 hours) and late response (DMBX1, OTX1, BARHL1, POU3F2, FOXB1, NR2F2, SOX11, LHX5, SOX5 and PAX6, 48-72 hours) clusters. In this follow up analysis, we used PAX6 as a positive control, since overexpression of PAX6a (short isoform of PAX6) is known to function as a neuroectoderm fate determinant and was previously shown to induce hESCs into a neural lineage [55]. In addition, we used EGFP as a negative control.

The chosen TFs were individually overexpressed in hESCs via lentivirus. Four days post infection, cells were harvested and examined for various lineage marker genes by RT-qPCR (Figure 6B). We found that overexpression of BARHL1, IRX3, LHX5, OTX1 and OTX2 were sufficient to induce PAX6 expression, suggesting that these TFs may play a role in neural fate specification. Overexpression of these TFs also induced other neural marker genes directly or indirectly via PAX6 (Figure 6B). Previous studies have shown that OTX2 overexpression promotes PAX6 expression in hESCs upon treatment with the TGFβ inhibitor SB431542 and FGF2 [56], and its paralogous gene OTX1 is known to function similarly [57]. It was also reported that LHX2, a paralog of LHX5, promotes PAX6 expression and neural differentiation in hESCs [58]. However, despite LHX5 being expressed in NPCs, its overexpression has yet to be associated with neural induction. The same holds true for IRX3 and BARHL1, which are known to be expressed in the neuroectoderm and central nervous

system, respectively, in the mouse embryo [59], but whose function in neural induction has not been evaluated. Consistent with these findings, analysis of the PAX6 promoter region identified binding sites of OTX, IRX3, SOX and POU that are evolutionarily conserved (Figure 6D). These results are in line with the observations that the respective region is active around 12-24 hours, when these TFs are significantly expressed (Figure 2A), and starts to gain a high TF activity score (Figure 6A) at those time points. We found several additional examples of functional neural enhancers that contain conserved OTX, SOX, IRX, and/or homeo-domain (recognized by both LHX and BARHL) binding motifs upstream from the LHX5, POU3F2 and OTX2 genes (Figure S6).

We next tested whether these five TFs could lead to a more established neural lineage by analyzing the expression of late neural marker genes (e.g. FABP7 and CDH2) nine and fourteen days after infection. We observed continuous upregulation of the late neural markers at day nine and fourteen (Figure 6C), consistent with the observation that these five factors activated the neural lineage determinant PAX6 at an early stage. Further examination of these cells at day fourteen via bright field microscopy and immunocytochemistry for the neuronal marker MAP2 indicated that the TF activated cells have acquired neuronal hallmarks (Figure 6E).

To gain a broader understanding of the changes to the transcriptional landscape following overexpression of the five TFs, we carried out RNA-seq analysis at day fourteen post infection, using three replicates per condition. As before, we used PAX6 as a positive control as well as EGFP as a negative control. As reference, we also sequenced NPCs that were induced by dual-Smad inhibition for 72 hours followed by 72h-culture in N2B27 medium supplemented with FGF and EGF, termed here as 'dSi'. Principal component analysis (PCA) of the resulting data validated the reproducibility amongst three replicates (Figure 7A). Interestingly, the first principal component captured the dichotomy between the two reference states (ESC and NPC, represented by EGFP and dSi respectively), as can be observed by marker genes and by a more systematic analysis of gene set enrichment (Table S6). The assayed TFs spanned a spectrum between the reference states, where PAX6-overexpression has the most similar effect to that of dSi as expected, and LHX5-overexpression has the least amount of similarity.

To assess cell lineage, we examined overlaps of DE genes [23] between each of TF-overexpressed cells and previously published hESC-derived mesendoderm (ME), trophoblast-like cells (TBL), mesenchymal stem cells (MSCs) and NPCs [8]. This analysis confirmed that the overlaps are most significant for NPCs (Figure 7B) than the other non-neural cells, supporting the role of all the six TFs in neural lineage specification. Gene set enrichment analysis of GO annotations [24] confirmed, for all overexpressed TFs, significant enrichment in central nervous system development and neurogenesis processes (Table S6). This analysis also validates that all overexpressed TFs lead to transcriptional changes that significantly overlap with those induced in dSi (Figure 7C; $p< 1 \cdot 10^{-}20$, hypergeometric test). Indeed, specific neural marker genes (e.g. CDH2 and FABP7) in TF-overexpressed cells were upregulated at a similar level to PAX6-overexpressed cells (Figure 7D), recapitulating the qPCR results (Figure 6C), while mesoderm and endoderm markers were expressed in a more limited

manner.

To explore potential regional characteristic of the TF-overexpressed cells, we focused on anterior-posterior brain marker genes, and found that BARHL1 induced more posterior markers (hindbrain marker GBX2 and hindbrain-spinal cord markers HOXB2 and HOXD4). While, as expected, OTX1 and OTX2 induced anterior identity (fore-midbrain markers FOXG1, SIX3, OTX1, EMX2, PAX2, EN1 and EN2) (Figure 7E). However, we should note that these TF-overexpressed cells are likely to comprise a heterogeneous regional identity.

To further validate the role of these five genes in neural induction, we set out to test whether knocking them down via CRISPR interference (CRISPRi) [45] will affect neural differentiation. We introduced dCas9-KRAB and sgRNAs that target the promoters of the five genes into hESCs, followed by neural differentiation and RT-qPCR analyses for various markers at 72 hours post neural induction. We found that CRISPRi of each of the five TFs decreased the expression of early neural genes, such as PAX6 and POUF3F1, and increased the expression of the pluripotent marker NANOG (Figure 7F). At six days post induction, later neural markers (i.e. CDH2 and FABP7) were also decreased, although other neural markers, such as MEIS2 and DLK1 showed normal expression and NANOG was downregulated. These results suggest that knockdown of these genes leads to impairment or possibly a delay in neural differentiation and therefore associates these genes as potential players in the regulation of neural differentiation.

6.4 Discussion

Genomic analyses of multiple time points during early neural induction provided several findings. We confirmed that neural induction first involves the silencing of pluripotent markers and upregulation of immediate early genes, corresponding to the cell's stress response against differentiation stimuli. This is then followed by the upregulation of genes involved in neural lineage fate specification. We also observed that this process is first controlled by chromatin accessibility or simultaneously with H3K27ac modification followed by an increase in mRNA expression. These results support previous reports about the importance of H3K27ac as an active promoter and enhancer mark that correlates with (and possibly affects) temporal changes in transcription levels, which are not captured by accessibility alone [60]. Finally, our work provides a comprehensive catalog of dynamically changing genes and regulatory elements during neural induction.

Analysis of temporal genes and DNA regions is important not only to understand the regulatory network underling neural induction, but also to dissect neurological disease. Indeed, a large body of evidence suggests that the temporal alteration of genes and regulatory elements involved in neural development can affect neurological phenotypes [61], such as cognition and brain size [62]. Fitting with these studies, we observed significant overlap between regions with induced H3K27ac histone modification and neurological disorder GWAS variants. We also observed a significant overlap between the set of loci that had temporal H3K27ac signal in our data and the set of loci found to have an eQTL in the brain and in

other tissues. Our null model for computing this statistic was the observed overlap between the set of all H3K27ac peaks in our data (regardless of how they change over time) and the eQTL hits. Finding significant enrichments beyond this baseline suggests that the temporal aspect adds important information, pointing at phenotypically important regions.

The use of lentiMPRA allowed us to functionally test thousands of CRS and identify 63% that have temporal activity. While we observed an overall strong correlation between the temporal patterns of these regions and their respective gene expression and chromatin features, it is important to note that this overlap was not obtained for all sequences. While this may result from inaccuracies in the various assays, it may also point to a biologically-driven cause (described in detail in STAR Methods).

By integrating information from across all of our large-scale assays, we proposed a scheme to identify and rank TFs based on their predicted activity during the course of development. After an initial screen, we identified BARHL1, IRX3, LHX5, OTX1 and OTX2 as important regulators of neural induction, as both overexpression and knockdown of these factors up- and down-regulated PAX6 and other neural markers, respectively. While PAX6 expression in hESCs was shown to be upregulated via OTX2 [56], this finding was novel for the other TFs. While LHX5 is a commonly used neural marker, its ability to induce neural induction was not tested. IRX3 and BARHL1 were of particular interest. In our study, we found their expression to increase at 24-72 hours, and it obtained a high activity score at these time points, suggesting an important role in neural induction. In our overexpression experiments, we observed that they could by themselves control several neural markers in the hESC culture condition, including PAX6. To our knowledge, this is the first report demonstrating a potential role for either IRX3 or BARHL1 in neural induction. Although we identified important regulators of neural induction, we also observed that both overexpression and knockdown of different TFs perturbed different sets of neural markers or even non-neural markers. This observation suggests that the orchestration of multiple TFs is necessary to fine tune neural differentiation. Assays that target the molecular function or regulatory grammar of these regulators will be necessary in order to further understand this regulatory network.

6.5 Methods

EXPERIMENTAL MODEL AND SUBJECT DETAILS

hESC culture and neural differentiation

H1 hESCs (WiCell WA-01, RRID:CVCL_9771) were cultured on Matrigel (Corning) in mTeSR1 media (STEMCELL Technologies). Medium was changed daily. For passaging, cells were dissociated using StemPro Accutase (Fisher Scientific), washed and replated on Matrigel-coated dish at a dilution of 1:5 to 1:10 in mTeSR1 media supplemented with 10 μM Y-27632 (Selleck Chemicals). For genomic assays, hESCs were allowed to expand until they were nearly confluent and harvested to obtain undifferentiated hESCs (0 hour). Neural dif-

ferentiation was performed using a dual-Smad inhibition protocol [21]. Briefly, the mTeSR1 media were replaced by neural differentiation media (Knockout DMEM; Life technologies) supplemented with knockout serum replacement (Life technologies), 2 mM L-glutamine, 1x MEM-NEAA (Life technologies), 1x beta-mercaptoethanol (Life technologies), 200 ng/mL Recombinant mouse Noggin (R&D systems), and 10 μM SB431542 (EMD Millipore), and harvested at 3, 6, 12, and 24 hours. At these time points, the cells were 50-90% confluent in 6-well plates (for RNA-seq and ATAC-seq) or 10 cm dishes (for ChIP-seq and lentiMPRA). The cells were further cultured by refreshing the neural differentiation media daily and harvested at 48 and 72 hours, when the cells were 100% confluent.

METHOD DETAILS

RNA-seq

hESCs were plated in 6-well plates and induced to neural differentiation as described above. Cells from all time points were lysed in RLT buffer (Qiagen) supplemented with beta-mercaptoethanol and stored in -20^{o}C. Total RNA were extracted using the RNeasy mini kit (Qiagen) following the manufacturer's protocol. RNA was quantified with Qubit RNA HS assay kit (Thermo Fisher Scientific). Sequencing library preparation was carried out using Illumina TruSeq Stranded Total RNA Kit. Massively parallel sequencing was performed on an Illumina NextSeq500 with 75 bp paired-end reads. RNA-seq was done with three biological replicates for each of the seven time points and sequenced deeply with an average of 200M reads per replicate.

ChIP-seq

ChIP-seq was performed using LowCell# ChIP kit (Diagenode) according to manufacturer's instruction with modifications. Briefly, cells cultured in 10 cm dishes were crosslinked in 1% formaldehyde (Thermo Fisher Scientific) for 5 minutes. Crosslinking was quenched with 125 mM Glycine. The cells were washed with PBS and precipitated with centrifugation at 6000 rpm for 5 minutes. The cell pellet was stored in -80^{o}C for each time point, so that all the samples were processed together. The pellet was lysed in 250μl of Buffer B (LowCell# ChIP kit) supplemented with complete protease inhibitor (Roche) and 20 mM Na-butyrate (Sigma). 130μl of lysed chromatin was sheared using a Covaris S2 sonicator to obtain on average 250 bp size fragments. 870μl of Buffer A (LowCell# ChIP kit) supplemented with complete protease inhibitor (Roche) and 20mM Na-butyrate (Sigma) was added to the shared chromatin. 20μl of the chromatin solution was saved as an input control. To obtain magnetic bead-antibody complexes, a mixture of 40μl of Dynabeads protein A and 40μl of Dynabeads protein G was washed twice with Buffer A (LowCell# ChIP kit) and resuspended in 800μl of Buffer A. 10 μg of H3K27ac (Abcam Cat# ab4729, RRID:AB_2118291) or H3K27me3 antibodies (Millipore Cat# 07-449, RRID:AB_310624) were added to the washed beads, and gently agitated at 4^{o}C for 2 hours. The beads-antibody complex was precipitated

with a magnet and the supernatant was removed. $800\mu l$ of shared chromatin was added to the beads-antibody complex and rotated at 4^{o}C overnight. The immobilized chromatin was then washed with Buffer A three times and Buffer C once, and eluted in $100\mu l$ of IPure elution buffer (IPure kit; Diagenode). In addition, $80\mu l$ of IPure elution buffer was added to the $20\mu l$ input that were saved before immunoprecipitation, and purified using the IPure kit. Purified DNA was sheared using a Covaris S2 sonicator once again to obtain on average 250 bp fragments. Sequencing libraries were generated using ThruPLEX DNA-seq kit (Rubicon Genomics) according to manufacturer's protocol. The DNA was size-selected using SPRIselect (Beckman Coulter). 0.7x and 0.9x volume of SPRIselect was used for right side and left side selection, respectively. DNA was quantified with Qubit DNA HS assay kit and Bioanalyzer using the DNA High Sensitivity kit (Agilent). Massively parallel sequencing was performed on an Illumina HiSeq4000 with 50 bp single-end read. ChIP-seq was done with two biological replicates for each time point.

ATAC-seq

ATAC-seq was performed according to previously described protocol [63] with modifications. Briefly, 50,000 cells were dissociated using Accutase and precipitated with centrifugation at 500g for 5 minutes. The cell pellet was washed with PBS, resuspended in $50\mu l$ lysis buffer (10 mM Tris·Cl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% Igepal CA-630), and precipitated with centrifugation at 500g for 10 minutes. The nuclei pellet was resuspended in $50\mu l$ transposition reaction mixture which includes $25\mu l$ Tagment DNA buffer (Nextera DNA sample preparation kit; Illumina), $2.5\mu l$ Tagment DNA enzyme (Nextera DNA sample preparation kit; Illumina), and $22.5\mu l$ nuclease-free water, and incubated at 37^{o}C for 30 minutes. Tagmented DNA was purified with MinElute reaction cleanup kit (Qiagen). The DNA was size-selected using SPRIselect (Beckman Coulter) according to the manufacturer's protocol. 0.6x and 1.5x volume of SPRIselect was used for right and left side selection, respectively. Library amplification was performed as previously described [63]. Amplified library was further purified with SPRIselect as described above. DNA was quantified on a Bioanalyzer using the DNA High Sensitivity kit (Agilent). Massively parallel sequencing was performed on an Illumina HiSeq2500 or HiSeq4000 with PE150. ATAC-seq was done in 2 biological replicates for each time point.

lentiMPRA library generation

The lentiMPRA plasmid library was constructed as previously described [38] with minor modifications. Briefly, array-synthesized oligos were amplified with two sets of adaptor primers mentioned previously (pLSmP-AG-f01/r02 and pLSmP-AG-f03/r04, Table S7, sheet 1). The amplified fragments were cloned into pLS-mP vector (Addgene_81225, RRID:Addgene_81225) following its digestion with SbfI and EcoRI using In-Fusion HD cloning kit (Takara). The reaction products were transformed into electrocompetent cells (NEB C3020). The pre-library was purified using Plasmid plus midi kit (Qiagen) and tested for its quality via sequencing on

a MiSeq (see below section). The minimal promoter and EGFP fragment (mP-EGFP) was inserted into the SbfI and EcoRI restriction sites contained between the enhancer and barcode sequence in the pre-library using T4 DNA Ligase (NEB M0202). The ligation products were then transformed and midi-prepped as mentioned above to obtain the final lentiMPRA library.

Before inserting the mP-EGFP, the plasmid pre-library was examined for the quality of the designed oligos and the representation of individual barcodes via sequencing as previously described [38]. CRS-barcode fragments were amplified using pLSmP-ass-F and pLSmP-ass-R-i# primers (Table S7, sheet 1), and purified using MinElute PCR cleanup kit (Qiagen). The DNA was sequenced with MiSeq (PE150). Two sets of sequencing primers (pLSmP-AG-seqR1 and pLSmP-AG-seqR1_2 for read 1, pLSmP-AG-seqR2 and pLSmP-AG-seqR2_2 for read 2, and pLSmP-AG-seqIndx and pLSmP-AG-seqIndx_2 for index read) were mixed at 1:1 ratio and used for the sequencing. We sequenced the CRS, spacer, and barcode sequences from both read ends and called a consensus sequence from the two reads using PEAR [64, 65]. We obtained 16.4 million paired-end consensus sequences from this sequencing experiment, 43% of which had the expected length, 30% of sequences were 1 bp short, and 13% were 2 bp short (summing up to 86%), similar to previously reported results [38]. Only 0.9% of sequences showed an insertion of 1 bp (Figure S3A). These results are in line with expected dominance of small deletion errors in oligo synthesis. We aligned all consensus sequences back to all designed sequences using BWA MEM [66] with parameters penalizing soft-clipping of alignment ends (-L 80). We consensus called reads aligning with the same outer alignment coordinates and SAM-format CIGAR string to reduce the effects of sequencing errors. We analyzed all those consensus sequences based on at least three sequences with a mapping quality above 0. Figure S3B shows the distribution of alignment differences (as a proxy for synthesis errors) along the designed oligo sequences. Errors are distributed evenly along the designed sequence, with deletions dominating the observed differences, similar to previous libraries generated in a similar manner [38]. We characterized the abundance of oligos further by focusing only on the barcode sequences. Barcode sequences were identified from the respective alignment positions of the alignments created above. To match the RNA/DNA count data analysis (see below), we only considered barcodes of 15-bp length. The number of barcodes per CRS are shown in Figure S3C. The distribution of the abundance of barcodes is available in Figure S3D.

The lentiMPRA library was packaged into lentivirus using Lenti-Pac HIV expression packaging kit (GeneCopoeia) and the lentivirus was concentrated using Lenti-Pac lentivirus concentration solution (GeneCopoeia) according to manufacturer's protocol. The lentivirus was titrated as described previously [38]. In brief, H1-hESCs were plated at 1-2 $\times$ 105 cells/well in 24-well plates and incubated for 24 hours. Serial volume (0, 2, 4, 8, 16, 32 μl) of the lentivirus was added with $8\mu g/mL$ polybrene. The infected cells were cultured for 3 days and washed with PBS three times. Genomic DNA was extracted using the Wizard SV genomic DNA purification kit (Promega). Copy number of viral particle per cell was measured by qPCR as previously described [38].

Lentiviral infection and DNA and RNA extraction

H1 hESCs cultured in a 10 cm dish at 80-90% confluency were split at 1:4 ratio and re-plated on Matrigel-coated 10 cm dishes in mTeSR1 media supplemented with 10 uM Y-27632. After 24 hours, the cells were infected with the lentivirus library with a multiplicity of infection (MOI) of 5-8 along with $8\mu g/mL$ polybrene (Sigma) and incubated for 3 days with a daily change of the media. Three independent replicate cultures were infected. The infected cells were harvested right before differentiation (0 hours), or differentiated into neural lineage as described previously until appropriate time points (3, 6, 12, 24, 48, and 72 hours). In order to distinguish the barcode expression level between short time gaps (i.e. 0 vs 3 hours, and 3 vs 6 hours), we collected nuclear RNA from all time points and analyzed the nascent state of barcode RNA expression as below. The cells were washed with PBS three times and dissociated with Accutase. The cells were then precipitated with centrifugation at 500g for 3 minutes and washed with PBS. To isolate cell nuclei, the pellet was lysed in $500\mu l$ lysis buffer [(10 mM Tris-HCL, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% (v/v) Igepal CA-630, $1U/\mu l$ RiboLock RNase inhibitor Thermo Fisher Scientific)]. The cell nuclei were precipitated with centrifugation at 500g for 10 minutes, and lysed in RLT plus lysis buffer (Qiagen) supplemented with 2-mercaptoethanol. Genomic DNA and nuclear RNA was purified using an AllPrep DNA/RNA mini kit (Qiagen). Copy number of viral particle per cell was confirmed by qPCR and shown in Figure S3E. RNA was treated with Turbo DNase (Thermo Fisher Scientific) to remove contaminating DNA.

RT-PCR, amplification, and sequencing of RNA/DNA

Sequencing libraries were prepared as previously described [38]. Briefly, $25\mu g$ nuclear RNA was used for reverse transcription with SuperScript II (Invitrogen) using a primer downstream from the barcode (pLSmP-ass-R-UMI-i#, Table S7, sheet 1), which contained a sample index, unique molecular identifier (UMI), and a P7 Illumina adaptor sequence. Barcode sequence was amplified with NEBNext high-fidelity 2X PCR master mix (New England Biolabs) for three cycles using this same reverse primer paired with a forward primer complementary to the 3' end of EGFP with a P5 adaptor sequence (BARCODE_ lentiF_v4, Table S7, sheet 1). PCR products were purified with 1.8x volume of SPRIselect and underwent a second round of amplification for 22 cycles with the same forward primer and a P7 primer. The PCR products were gel-extracted and purified with MinElute reaction cleanup kit (Qiagen). To amplify barcode sequence integrated into the genome, $4\mu g$ of genomic DNA was used for PCR amplification as the RNA. The amplified DNA was quantified on a Bioanalyzer (Agilent) using the DNA High Sensitivity kit, and sequenced with an Illumina HiSeq4000 with 100 bp paired-end read. The BARCODE-SEQ-R1-v4 primer was used for read 1 (Table S7, sheet 1). The same primers as pre-library sequencing with MiSeq were used for read 2 and index reads.

The forward and reverse reads on this run each sequenced the designed 15-bp barcodes as well as an adjacent sequence to correctly trim and consensus call barcodes. We obtained a

total of 398.9, 406.5 and 415.1 million reads for replicate 1, 2 and 3 respectively; full statistics of read counts is presented in Table S3, sheet1. Across replicates and time points, 95% of barcodes were of the correct length of 15 bp when matched against the designed barcode; full counts and barcode statistics is presented in Table S3, sheet 1-3.

Luciferase assays

To analyze enhancer temporal changes via luciferase assays, we engineered a lentiviral reporter vector pLS-mP-Luc (Addgene, #106253, RRID:Addgene_106253) to have a destabilized form of the luciferase gene by placing a degeneration sequence, hPEST, downstream of the luciferase gene. The hPEST sequence (including 3' partial sequence of the luciferase gene) was amplified from pGL4.11 (Promega) using the following primers (forward, TTCGAGGCTAAGGTGGTGGA; reverse TACGAAGTTATTAGGTCCCTCGACGAATTCTTAGACGTTGATCCTGGCGC), and inserted into AgeI and EcoRI sites of the pLS-mP-Luc. OTX2 ESC (chr14:57385639-57386304; hg19), NANOG ESC (chr12:7940151-7940848; hg19), ENSA ESC (chr1:150613721-150614409; hg19), EDNRB ESC (chr13:78427691-78428404; hg19), TMEM132D ESC (chr12:130255696-130256309; hg19), PAX6 IR (chr11:31832507-31832981; hg19), PARD3 IR (chr10:34716118-34717054; hg19), CRIM1 IR (chr2:36688268-36688907; hg19), PCLO IR (chr7:82434703-82435321; hg19), OTX2 NPC (chr14:57313599-57314370; hg19), CTNNA2 NPC (chr2:80235184-80235754; hg19), DLK1 NPC (chr14:101306429-101307069; hg19), MYB NPC (chr6:135571820-135572468; hg19) enhancers were amplified from the human genome and inserted into XbaI site of the vector. Primers used for cloning are shown in Table S7, sheet 2. The empty pLS-mP-Luc vector was used as a negative control. The plasmids were individually packaged into lentivirus together with Renilla luciferase vector, pLS-SV40-mP-Rluc (Addgene, #106292, RRID:Addgene_106292) at 1:1 molar ratio using Lenti-Pac HIV expression packaging kit (GeneCopoeia) and the lentivirus was concentrated using Lenti-X concentrator (Takara) according to the manufacturer's protocol. H1 hESCs seeded 24 hours before were infected with the lentivirus along with $8\mu g/mL$ polybrene (Sigma). Three independent replicate cultures were infected. After 48 hours, the cells were induced into a neural lineage using the dual-Smad inhibition method described above. The cells were lysed in buffer PLB (Promega) at 0 (before neural induction), 12 and 72 hours after neural induction. Firefly and renilla luciferase activities were measured on a Glomax microplate reader (Promega) using the Dual-Luciferase Reporter Assay System (Promega). Enhancer activity was calculated as the fold change of each plasmid's firefly luciferase activity normalized to Renilla luciferase activity.

CRISPRa

To generate a H1 hESC line that stably expresses dCas9-VP64, the lenti dCAS-VP64_Blast vector (Addgene, #61425, RRID:Addgene_61425) was transduced into H1 hESCs via lentivirus at a MOI of 0.2 along with $8\mu g/mL$ polybrene (Sigma) and incubated for 2 days to allow genomic integration. The cells were further cultured for 5 days in a media supplemented with

$2\mu g/mL$ blasticidin for selection. Individual colonies were isolated to obtain clonal cell populations and expanded for 2 weeks in blasticidin media. sgRNA sequences for SOX1, IRX3, OTX2 enhancers, PAX6 promoter and non-targeting negative control sequence were amplified as a part of PCR primers (Table S7, sheet 3) using the pLG1 plasmid (gift from Prof. Jonathan Weissman) as a template and cloned into XhoI and BstXI site of the pLG1. The sgRNA plasmids were transduced into dCas9-VP64 ESCs via lentivirus at a MOI of 5 along with $8\mu g/mL$ polybrene (Sigma) and incubated for 2 days to allow genomic integration. The cells were further cultured for 2 days in a media supplemented with $2\mu g/mL$ puromycin for selection. Total RNA was collected using RNeasy mini kit (Qiagen). Reverse-transcription was carried out using SuperScript III first-strand synthesis system (Invitrogen). qPCR was performed using SsoFast EvaGreen supermix (Bio Rad) according to the manufacturer's protocol. Primer sequences used for qPCR are shown in Table S7, sheet 5.

Immunocytochemistry

TF-overexpressed cells were fixed using 4% paraformaldehyde (Thermo Fisher Scientific) for 10 minutes and washed three times with PBS. Blocking was performed using blocking/staining solution (0.05% sodium azide, 0.1% NP40, 0.4% BSA, 4% normal goat serum in PBS) for 1 hour. Mouse anti-MAP2 antibody (Thermo Fisher Scientific, catalog# 13-1500, RRID: AB2533001) and Donkey anti-Mouse IgG conjugated with Alexa Fluor 488 (Thermo Fisher Scientific, catalog# R37114, RRID:AB2556542) were used for immunostaining.

Overexpression and RT-qPCR

Total RNA was collected from neural progenitor cells differentiated from hESCs by dual-Smad inhibition as described above. The total RNA was reverse-transcribed using SuperScript III first-strand synthesis system (Invitrogen) according to manufacturer's protocol. cDNA of ELF3, FOXB1, HOMEZ, ID4, IRX3, LHX5, OTX2, PAX6a, SMAD1, SMAD4 and SOX1 were amplified. BARHL1 (catalog#, MHS6278-213245170), MAF (catalog#, MHS6278-202806268), NR2F2 (catalog#, MHS6278-202800802), NR3C1 (catalog#, MHS6278-202832263), POU3F2 (catalog#, OHS6271-213587035) and SOX2 (catalog#, MHS6278-202826163) cDNA clones were obtained from Dharmacon. SOX11 (clone ID, OHu15579D) and SP5 (clone ID, OHu03497D) cDNA clones were obtained from Genscript. BACH2, DMBX1, FOSL2, NFE2, OTX1, SOX5 and TCF7L2 cDNA sequences were synthesized by Twist Bioscience. The EGFP gene (negative control) and T2A fragment were also amplified using the pJA291 vector (Addgene #74487, RRID:Addgene_74487) as a template. Sequences synthesized by Twist Bioscience and primers used for the cloning are shown in Table S7, sheet 4. The gene's cDNA fragment and T2A fragment were assembled into pJA291 vector that had been digested with EcoRI and XcmI to generate overexpression vectors that expresses PuroR-mCherry-T2A-cDNA under the control of EF1-alpha promoter. For BACH2, SOX5 and TCF7L2, as their cDNA are quite long (i.e. 3 kb), 5' and 3' parts of the sequences that overlap each other were separately synthesized by Twist Biosciences and assembled when

cloned into the vector. The sequences of cloned cDNA were confirmed by Sanger sequencing. The overexpression vectors were individually packaged into lentivirus using Lenti-Pac HIV expression packaging kit (GeneCopoeia) and the lentivirus was concentrated using Lenti-X concentrator (Takara) according to the manufacturer's protocol. The lentivirus were titrated with H1-ESCs by qPCR, as described above. H1-ESCs cultured in a 24-well plate for 24 hours were infected with the lentivirus with a MOI of 5 along with $8\mu g/mL$ polybrene (Sigma) and incubated for 2 days to allow genomic integration. The cells were further cultured for 2-7 days in a media supplemented with $2\mu g/mL$ puromycin for selection. Three independent replicate cultures were infected. Total RNA was collected using RNeasy mini kit (Qiagen). Reverse-transcribed using SuperScript III first-strand synthesis system (Invitrogen). qPCR was performed using SsoFast EvaGreen supermix (Bio Rad) according to manufacturer's instruction. Primer sequences used for qPCR are shown in Table S7, sheet 5.

CRISPRi

sgRNA sequences for BARHL1, IRX3, LHX5, OTX1, OTX2, and PAX6 promoters were cloned into pLG1 as described above. sgRNA plasmids and pHR-SFFV-KRAB-dCas9-P2A-mCherry (Addgene, #60954, RRID:Addgene_60954) were co-packaged and transduced into H1 hESCs via lentivirus at a MOI of 5 along with $8\mu g/mL$ polybrene (Sigma). Cells were incubated for 2 days to allow genomic integration and further cultured for 2 days in mTeSR media supplemented with $2\mu g/mL$ puromycin for selection. At day 4 after infection, the cells were replated and cultured in mTeSR supplemented with puromycin. At day six, cells were induced into a neural lineage by dual-Smad inhibition. At day 9 and 12 (72 hours and 6 days post neural induction), total RNA was collected using RNeasy mini kit (Qiagen) and reverse-transcribed using SuperScript III first-strand synthesis system (Invitrogen). qPCR was performed using SsoFast EvaGreen supermix (Bio Rad) according to the manufacturer's protocol. sgRNA sequences and primers used for the plasmid construction and RT-qPCR are shown in Table S7, sheet 5.

QUANTIFICATION AND STATISTICAL ANALYSIS

Computational pipeline for RNA-seq, ChIP-seq and ATAC-seq

For RNA-seq, reads were aligned to the hg19 human genome assembly with Tophat2 (Version 2.1.1) [67], and low quality reads were trimmed or removed with Trimmomatic (Version 0.3.2) [68]. Reads that aligned to more than one gene as well as chimeric fragments were excluded. We also removed genes that failed to be quantified in at least one sample by Cufflinks [69]. We implemented a quality control (QC) pipeline that computes an extensive set of quality metrics, relying in part on FASTQC (Version 0.3.2; Babraham Bioinformatics) and the PICARD suite of alignment metrics (Version 2.5.0 with samtools 1.3.1). Transcript levels were determined using RefSeq transcript annotations, and counting the number of

reads aligning to every gene (defined as the union of all splice forms) with featureCounts (Version 1.5.0-p3) [70].

For both ChIP-seq and ATAC-seq, we used the FASTQC pipeline (Version 0.3.2; Babraham Bioinformatics) on our reads, and aligned them to the reference genome (hg19) with bowtie version 1.1.1 [71] (for ChIP-seq) and bowtie2 version 2.2.9 [72] for ATAC-seq, retaining only reads that mapped to a unique position in the genome ["–m 1"]. We marked duplicate reads in the bam files using PICARD and checked for contamination of primer sequences using Trimmomatic (Version 0.3.2) [68].

For each of our H3K27ac and H3K27me3 ChIP-seq replicate pairs per time points peaks were called using MACS2 version 2.1.0 [73], with the relevant control input file with default parameters (setting the FDR to 0.05 and default hg19 human genome size). For each mark in each time point, we intersected the peaks from the two replicates. We then took the union of all of these peaks from all time points per mark while merging regions with maximum distance of 1,000 bp using bedtools [64]. This resulted in 40,170 H3K27ac peaks and 4,446 H3K27me3 peaks (Table S1, sheet 1). For ATAC-seq, after alignment of the reads to the reference genome, reads aligned to the positive strand were moved +4 bp, and reads aligning the negative strand were moved -5bp. For each of our two replicate pairs per time point, we called peaks using MACS2 version 2.1.0 [73] with default parameters (setting the FDR to 0.05 and default hg19 human genome size). For each time point, we intersected the narrow peaks from the two replicates. To generate our final universe of peaks for the LentiMPRA experimental design, we took the union of all peaks from all time points while merging regions with maximum distance of 100 bp using bedtools [64], resulting in 40,486 peaks (Table S1, sheet 1).

Differential activity analyses

We used the 40,042 H3K27ac peaks that did not intersect with H3K27me3 peaks (excluding intersecting peaks per time point) to test for differential activity. We counted the number of H3K27ac reads that fell inside each peak region for each time point, for each of the two replicates. We extracted the shifted ATAC-seq cut sites from our data and counted the number of cut sites that fell inside each of the 40,486 ATAC-seq peak regions for each time point, for each of the two replicates. We used the read count for each transcript across time points and replicates for the RNA-seq. We then used DESeq2 [23] for all three assays to identify the differential abundance of reads and provide normalized reads (by a scaling factor) matrices. We performed all pairwise comparisons of the seven time points and recorded the FDR for each such comparison for every region/gene.

As an additional analysis of differential expression/ activity over time, we used ImpulseDE [22], a package that fits impulse like functions to temporal data and reports differential signals across a time course by assigning an FDR value to each region/gene. To call differential H3K27ac, ATAC-seq or RNA-seq signal we used a cutoff of FDR < 0.01 from ImpulseDE and FDR < 0.05 for DESeq2, while taking a maximum of 500 top regions/genes per every two time point comparison. This resulted in 2,435 H3K27ac regions, 2,024 ATAC-

seq regions (for the 7 time points experiment) and 2,172 genes that showed differential and temporal activity (Table S1, sheet 2). To call differential H3K27me3 signal, we used a more relaxed cutoff of FDR ¡ 0.1 from ImpulseDE and FDR < 0.05 for DESeq2. This resulted in 248 H3K27me3 regions that showed differential and temporal activity.

ChIP-seq, ATAC-seq and RNA-seq clustering

Considering those regions/genes defined as differential in the previous section, we created for each assay, a matrix with the number of reads in each peak region or gene scaled according to the DESeq2 scaling factor and averaged between the two replicates for each time point. We clustered these matrices using the K-means clustering algorithm with 6 clusters (Figures 1B-D, S1B, Table S1, sheet 3). We also computed for each cluster the significance of its intersection with a cluster from a different assay (Figure 1E) using the hypergeometric test with Bonferroni correction for the p-value. For the intersection between H3K27ac and ATAC-seq peaks, we used bedtools [64] to determine if two peaks share at least 1 bp. For the intersection between H3K27ac/ATAC-seq peak region and a gene, we assigned the closest gene to a region (up to 1MB) using GREAT [30].

Enrichment of genomic variants

We compared enrichment of variant groups in H3K27ac and ATAC-seq temporal regions to their respective full set of peaks using Fisher's exact test. Variant groups included: the full GWAS catalog as downloaded in February 2018 [74], relevant disorder subsets of the catalog: alcohol, alzheimer, anxiety, autism, bipolar, borderline, brain volume, cognitive, depression, epilepsy, major depression, OCD, psychosis, schizophrenia, a combined list of disorders (schizophrenia, attention deficit disorder (ADHD), autism, bipolar and major depressive disorder), nervous system disorders obtained using the Experimental Factor Ontology (EFO) [75] and negative control height variants. For nervous system disorders and height variants we extracted all variants in linkage disequilibrium (r2$>$ 0.8) using the SNP Annotation and Proxy Search (SNAP) tool Version 2.2 [76]. We also examined eQTLs from different studies [34] and eQTLs from brain tissues [33]. All variants were converted to hg19 genomic location using the LiftOver tool available on the human genome browser [77].

lentiMPRA library design

We devised five criteria to nominate a set of CRS to be tested for their function during neural induction. For criterion 1 we selected sequences that are next to genes involved in neural differentiation or known enhancers that were validated (Table S2, sources of manually curated enhancers were shown in the column "References"). Criteria 2 and 3 require the closest gene to be induced upon neural induction; to satisfy this, we require the gene to be included in one of clusters 2 to 6 in Figure 1B. For criterion 4, we selected the most significant 20 (sorted by FDR) of each of the following tests: 1) RNA-seq differential expression over time (using ImpulseDE); differential signal in one of the following: 2) ATAC-seq 3hr vs. 0hr

(using Deseq2); 3) ATAC-seq 72hr vs. 0hr (Deseq2); 4) H3K27ac 3hr vs. 0hr (Deseq2); 5) H3K27ac 72hr vs. 0hr (Deseq2); 6) RNA-seq of nearest gene 3hr vs. 0hr (Deseq2); 7) RNA-seq of nearest gene 72hr vs. 0hr (Deseq2). The criteria were applied sequentially (in the order in which they were described), and the respective sets of candidate enhancers are mutually exclusive. To focus on neural induction, we excluded regions that are adjacent to the pluripotent factors SOX2, KLF4, MYC, NANOG, and POU5F1.

Notably, all selection criteria use a subset of our ATAC-seq data (0, 3, and 72 hours), which was available during the design of the library. Furthermore, we excluded from the design sequences that overlap with regions in the hg19/ENCODE blacklist `https://sites.google.com/site/anshulkundaje/projects/blacklists` (Table S2).

Due to limitations of the procedure of oligonucleotide synthesis, the assayed sequence are required to be 171 bp long. If a selected candidate region is shorter than 171 bp, we extended it equally from each side. If it is longer than 171 bp - we record all 171 bp options with a sliding window of 1 bp. For each such 171 bp sequence candidate we recorded motif hits using Fimo [36] with FDR $< 10^{-4}$ cutoff using the motif list from ENCODE [37] and chose the candidate sequence that has the maximal number of hits and satisfies the following criteria: 1) The sequence should not contain EcoRI (GAATTC) and SbfI (CCTGCAGG), because these sites were later used for inserting the minimal promoter (mP) and EGFP gene between the candidate regulatory sequence and barcode; 2) We discarded sequences with homopolymers longer than 8bp, since homopolymers can affect oligo synthesis; 3) There should be no more than 25% overlap (of the 171bp) with simpleRepeats regions from ENCODE (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/simpleRepeat.txt.gz).

We added to these 171 bp sequences a 5' primer sequence (AGGACCGGATCAACT), along with a 14 bp spacer sequence (CCTGCAGGGAATTC) that contains two restriction enzyme sites (SbfI and EcoRI), to allow for the subsequent insertion of the minimal promoter and EGFP gene followed by a 15 bp designed barcode sequences and a 3' primer sequence (CATTGCGTGAACCGA) (Figure 2B) [38]. In our final array design, we included 2,664 different target sequences (2,271 – sequences after filtering, 193 controls, 200 scrambled sequences), each with 90 different barcodes to provide a robust readout (Figure 2B). Barcode sequences of 15 bp length were designed to have at least two substitutions and one 1 bp insertion distance to each other. Homopolymers of length 3 bp and longer were excluded in the design of these sequences, and so were ACA/CAC and GTG/TGT trinucleotides (bases excited with the same laser during Illumina sequencing). More than 556,000 such barcodes were designed using a greedy approach. The barcodes were then checked for the creation of SbfI and EcoRI restriction sites when adding the spacer and 3' flanking sequences. From the remaining sequences, a random subset of 239,760 barcodes was chosen for the design. The final designed oligo sequences are available in Data File S1.

Replicates, normalization and RNA/DNA ratios

We used both the forward and reverse reads to sequence the 15bp reporter barcodes and obtain consensus sequences. We matched the observed barcodes against the designed barcodes,

and noticed that across replicates and sample types, $\sim 95\%$ of barcodes had the correct 15bp length. Only correct size barcodes that are observed at least once in both RNA and DNA of the same sample were subsequently used for analysis (assuming basal levels of transcription through the minimal promoter). To estimate the RNA to DNA ratio per barcode in each replicate, we first scaled the RNA and DNA read counts using the number of reads as scaling factor. DNA/RNA ratio per barcode: $\frac{\text{RNA reads}}{\text{sum RNA reads}} / \frac{\text{DNA reads}}{\text{sum DNA reads}}$.

Although the DNA and RNA counts of individual barcodes are highly correlated between experiments (Table S3, sheet 4), the noise of each measure results in a poor correlation of RNA/DNA ratios (Table S3, sheet 4). However, there are on average 68-72 barcodes per CRS in each replicate (out of 90 barcodes programmed on the array; Table S3, sheet 5). To reduce noise, we aggregated the RNA or DNA counts across all associated barcodes for each CRS.

To estimate the abundance of DNA or RNA per CRS and for each replicate (in order to compare replicates and time point, we use a simple averaging scheme: D(R)NA per CRS = $\frac{10^6 \cdot \sum_{i=1}^{\#BC} D(R)NA_i}{\#BC \cdot (\text{sum D(R)NA reads})}$ where $D(R)NA_i$ denotes the reads of a specific barcode i among the #BC barcodes that belong to the respective CRS.

To determine the RNA/DNA ratios per CRS and for each replicate we used two strategies:

$$\text{Ratio of sums}: \frac{\sum_{i=1}^{\#BC} \frac{\text{RNA}_i}{(\text{sum RNA reads})}}{\sum_{i=1}^{\#BC} \frac{\text{DNA}_i}{(\text{sum DNA reads})}}$$

$$\text{Sum of ratios}: \frac{\sum_{i=1}^{\#BC} \left(\frac{\text{RNA}_i}{(\text{sum RNA reads})} / \frac{\text{DNA}_i}{(\text{sum DNA reads})} \right)}{\#BC}$$

We added a pseudo count of 1 to the numerator and denominator to stabilize signal from CRS with low numbers of reads. Table S3, sheet 6 shows DNA or RNA abundance and RNA/DNA ratios per CRS between every two replicates per time point. Table S3, sheet 7 shows DNA or RNA abundance and RNA/DNA ratio (using the two schemes) per CRS for each replicate, comparing every two time points. Notably, the two schemes are largely consistent. In the remainder of this study, we used the second scheme (sum of ratios). We also compared between DNA, RNA and ratio per time point for each replicate. We observed low correlation between RNA/DNA ratios and DNA counts, indicating that enhancer activity was not influenced by the number of DNA integrations (Table S3, sheet 8).

Although normalized individually, the three replicates do not seem to be on the exact same scale (Table S3). To combine replicates, we therefore first divided the RNA/DNA ratios observed in each sample (time point/ replicate) by the median ratio and then obtained the final RNA/DNA ratio by averaging the normalized values across replicates.

Determining differential and temporal CRS activity using MPRAnalyze

MPRAnalyze is a statistical framework for analyzing MPRA data [42], using a parametric graphical model to infer the enhancer induced transcription rate. The model assumes a

linear relationship between the latent plasmid (DNA) and transcript (RNA) counts, relating them through scaling by the transcription rate α. The plasmid counts are assumed to follow a log-Normal distribution, and the RNA counts are assumed to follow a Negative Binomial distribution. This model incorporates external covariates, such as batch effect, barcode-specific effect and conditions of interest, by fitting two nested generalized linear models, one fitting the latent plasmid counts from the DNA counts, and the other fitting the transcription rate from the latent plasmid counts and the observed transcript counts. This model was designed to leverage the statistical power of multiple barcodes. More details are provided in Ashuach et al. [42].

Quantification and classification of active enhancers: to classify active CRS, estimates of α were extracted for each time point from the model described above. The α values corresponding to control enhancers are used as the baseline, and a modified z-score is computed for each CRS. The scores are computed as the distance from the median of the control α values, normalized by the median absolute deviation (MAD): $\text{score}_i = \frac{\alpha_i - median(\alpha_C^{\sim})}{K \cdot MAD(\alpha_C^{\sim})}$, where the constant K is set to ensure that the scores behave asymptotically normal, and aC is the vector of values corresponding to control enhancers. P-values are produced based on these scores compared with the standard normal distribution.

Identifying temporal CRS: To test for temporal activity, we incorporate control enhancers to define the null temporal behavior and use a likelihood ratio testing to detect significant temporal behavior. For a given CRS, the null assumption is that it behaves according to the null temporal behavior. We evaluate this assumption by fitting a joint model for the time course data of this enhancers, together with the set of negative controls. In the alternative model, the CRS has a temporal profile that is different form the null. To evaluate it, we fit a separate model for the controls and the CRS. In this scheme, a CRS with temporal behavior that significantly deviates from the null will have a clear benefit to the likelihood under the alternative model. The score is therefore computed by a likelihood ratio test between the two models.

Clustering MPRA data and association with other genomic assays

We clustered temporal MPRA regions into four rough patterns of expression, namely early, mid-early, mid-late and late response (Table S4, sheet 1). Using the genomic location of each region, we retrieved the normalized number of reads using DESeq2 [23] from an overlapping H3K27ac and ATAC-seq peaks (if any), as well as the expression of the nearest gene. We clustered each genomic assay separately to four clusters (similar to MPRA signal clustering). We then compared MPRA temporal profile to that of each genomic assay by measuring the overlap between the resulting clusters using a hypergeometric test and Bonferroni corrected FDR$<$ 0.05 (Table S4, sheet 2).

TF activity score computation

To compute the activity score of each TF (represented by a motif or a ChIP-seq experiment) at each time point, we look for consistent sub-clusters that peak during that time point (in terms of MPRA signal) and that significantly overlap with the putative target regions of the TF (p-value$<$ 0.005, Hypergeometric test). We then count the number of putative target regions that appear in at least one significantly overlapping sub-cluster. The final score is defined by the number of regions found at each time point divided by the total number of regions found across all time points. As an additional constraint, we only consider time points in which the mRNA that encodes for the TF is highly expressed (6th or higher quantile of expressed genes) and significantly induced compared to the preceding time point [p-value $< 10^{-5}$; for the first time point (0 hour), we compare to the subsequent time point (3 hours) [23].

TF activity score ranking

The TF activity score was ranked per each one of the 4 clusters with an unbiased approach that is based only on the data produced for this paper. It uses two components: (i) the p-value of the overlap between the TF's targets and significant sub clusters of MPRA activity (Figure 5E; Table S5) – taking the minimum p-value. (ii) log fold of the TF's mRNA induction according to cluster: 0-12hr for cluster 2, 0-48hr for cluster 3, 0-72hr for cluster 4. We rank (i) and (ii) per cluster and use their average as the final ranking score.

RNA-seq following TF overexpression

RNA-seq was performed by Novogene for the eight cell populations across three replicates, including: overexpression of BARHL1, IRX3, LHX5, OTX1, OTX2 and PAX6, negative control EGFP (corresponding to hESC state) and dSi, which were induced from hESCs via dual-Smad inhibition for 72 hours followed by further 72h-culture in N2B27 medium supplemented with 20 ng/mL bFGF (R&D systems) and 20 ng/mL EGF (MilliporeSigma). The RNA-sequencing data was processed similarly to the procedure described above. PCA analysis of RNA-sequencing these eight cell populations across three replicates, is based on the 1000 most variable genes (Figure 7A). For the overlap between DE genes (EGFP, dSi) and (EGFP, $factor$) or (EGFP, $factor_i$) and (EGFP, $factor_j$) we used a jaccrad score of:

$$\frac{|\bigcap \text{upregulated genes}| + |\bigcap \text{downregulated genes}|}{|\bigcup \text{upregulated genes}| + |\bigcup \text{downregulated genes}|}$$

The hypergeometric test of the overlap used a background of all genes with TMP$>$ 1, resulting in p-value=0 for all the tests (Figure 7C). We used DESeq 2 [23] for differential expression (DE) analysis for comparing each of the six overexpressed TFs to controls (EGFP and dSi) and comparing EGFP to dSi. Upregulated and downregulated genes were defined based on the cutoff of FDR$<$ 0.05; $|logFC| > 1$ (Figure 7D-E). For the cell lineage analysis,

we used data on lineage-restricted genes of four hESC-derived cell types (i.e. trophoblast-like cells (TBL), mesendoderm (ME), mesenchymal stem cells (MSCs) and neural progenitor cells (NPCs)), and restricted the lineage-restricted genes to have FPKM>1 only in that lineage based on Table S1 in Xie at al. [8]. We examined their overlaps with our DE genes (EGFP, $factor_i$) in a similar way to the jaccard score described above (Figure 7B).

Characterizing MPRA and chromatin/mRNA inconsistencies

To investigate the inconsistency phenomenon at the chromatin level, we turned to the cluster-level analysis (Figures 5E, S6). This analysis was designed to identify cases where regions that exhibit a certain temporal pattern with MPRA are likely to exhibit another pattern in their accessibility or H3K27 acetylation (adjusted p-value<0.05). As a general trend, the results indicate that 'inconsistent' temporal regions tend to become induced (when assayed by MPRA) after the occurrence of chromatin changes in their respective endogenous loci. For instance, we observe a significant overlap between the set of regions that become induced after 24 hours when examined by MPRA (MPRA cluster 3; Figure 5A), and the set of regions that become (or remain) accessible during the preceding time points (ATAC-seq cluster 1; Figure 5D). Furthermore, this pattern of delay is observed more often with chromatin accessibility, compared with H3K27ac (Figure S4). These results could potentially be explained by our previous observations that DNA accessibility precedes H3K27ac during neural induction, which is followed by gene expression changes (Figure 1E), and that the temporal H3K27ac signal is a stronger indicator for MPRA enhancer activity (Figure 3A). In addition to inconsistency with the chromatin readouts, we also observe temporal CRS that show inconsistency with their postulated target genes.

Specifically, we observe temporal CRS that were active several time points before their postulated target genes (Figure S4B) and the opposite, where genes were active before the CRS (Figure S4C). The pattern of MPRA induction before the endogenous mRNA can be rationalized by additional constraints that may exist in the endogenous regions, but not necessarily in the (random) integration sites such as dependence on a wider chromatin context, which may be required to enable transcription. Additional technical factors of the assay, including the length of the assayed sequence (171 bp), may also underlie these discrepancies. It is also worth noting that our assays only find potential enhancers but not their target gene/s. Conversely, the latter pattern (mRNA before MPRA) is harder to rationalize and is more likely a result of the assay's inaccuracy.

We investigated the cases in which the MPRA data of temporal CRS and the mRNA data of their respective genes did not match. To account for cases of CRS- gene miss-assignment, we removed from this analysis cases where there was another nearby gene (looking at the closest four) that was more correlated with the MPRA but showed inconsistency with the closest gene mRNA signal. To this end we first separately clustered each of the two sets (closest genes and the most correlated neighboring genes) to four temporal clusters. We declared two clusters c1 (from the set of closest genes) and c2 (from the set of most correlated genes) as sufficiently matching if the median of the Pearson correlation coefficient across all

pairwise comparisons of the respective genes was larger than 0.5. We considered a region for further analysis if the clusters that contain its closest gene and its most correlated neighboring gene are sufficiently matching. Counting the number of occurrences of each of the two patterns, we find that the second one (mRNA before MPRA) is of a substantially lower abundance (137 vs. 358 enhancers), and that its size is in fact at the level of overlap between random sets (adjusted Hypergeometric p-value$>$ 0.05) (Figure 5E). The resulting regions are depicted in Figure S4.

6.6 Figures

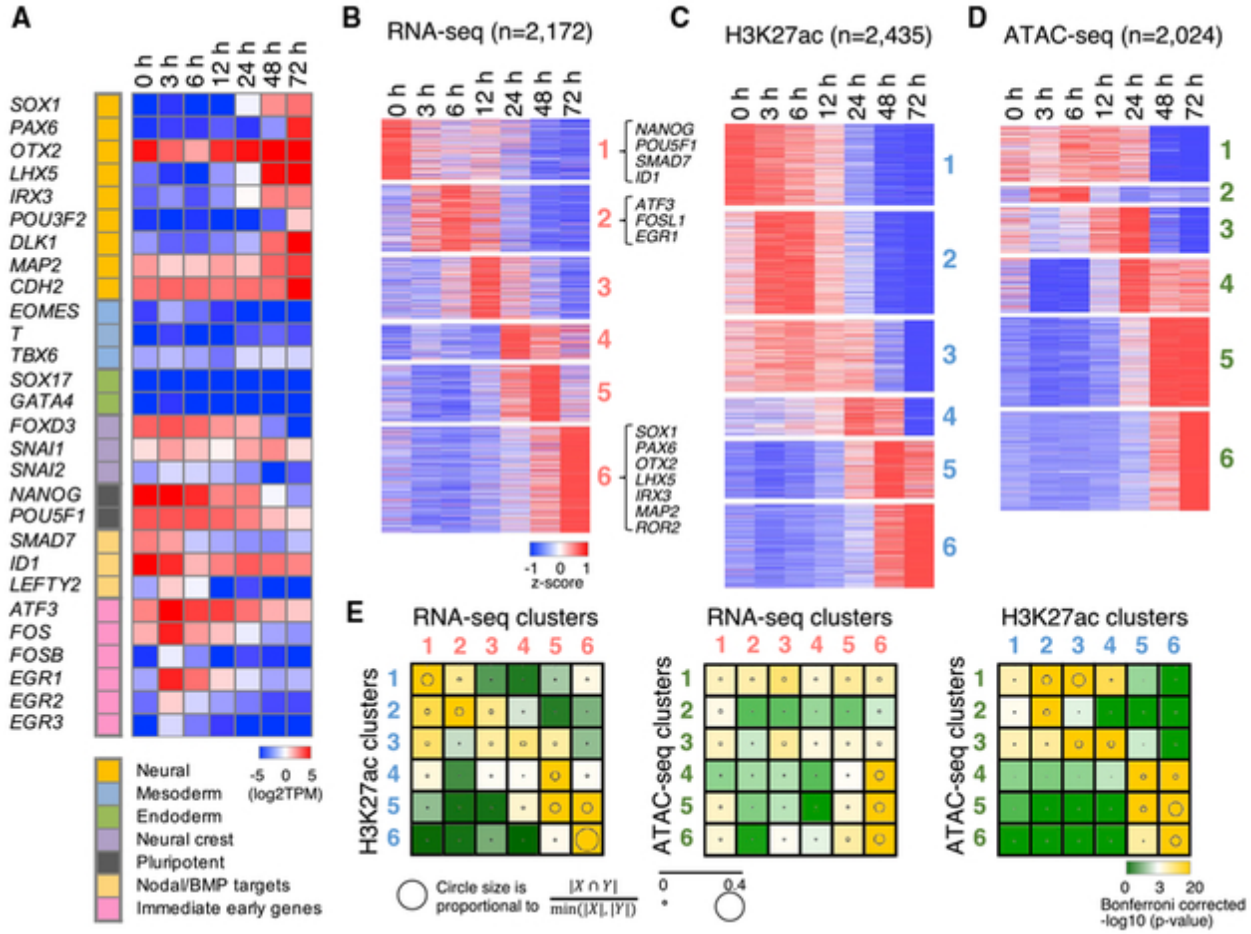


Figure 1: **The Dynamic Changes of ATAC-seq, ChIP-seq, and RNA-seq Peaks Are Sequentially Correlated. (A)** Transcripts per million (TPM) (log2, averaging over three biological replicates) per time point of marker genes (neural, mesoderm, endoderm, neural crest, pluripotent, nodal/BMP targets, and immediate early genes). **(B–D)** Heatmap of scaled read counts (log2, averaged over three biological replicates and standardized per row) of temporal genes and genomic regions, showing data from RNA-seq **(B)**, H3K27ac ChIP-seq **(C)**, and ATAC-seq **(D)**. The loci in each assay were clustered into six groups based on their temporal patterns. **(E)** Overlap between the temporal clusters in the three data modalities (Bonferroni-corrected p values of a hypergeometric test). Circle sizes represent the proportion of overlap between every two clusters. The overlap is computed either at the region level (ATAC-seq versus ChIP-seq) or at the gene level (ATAC/ChIP-seq versus RNA-seq; regions in the former assays are represented by their nearest gene).

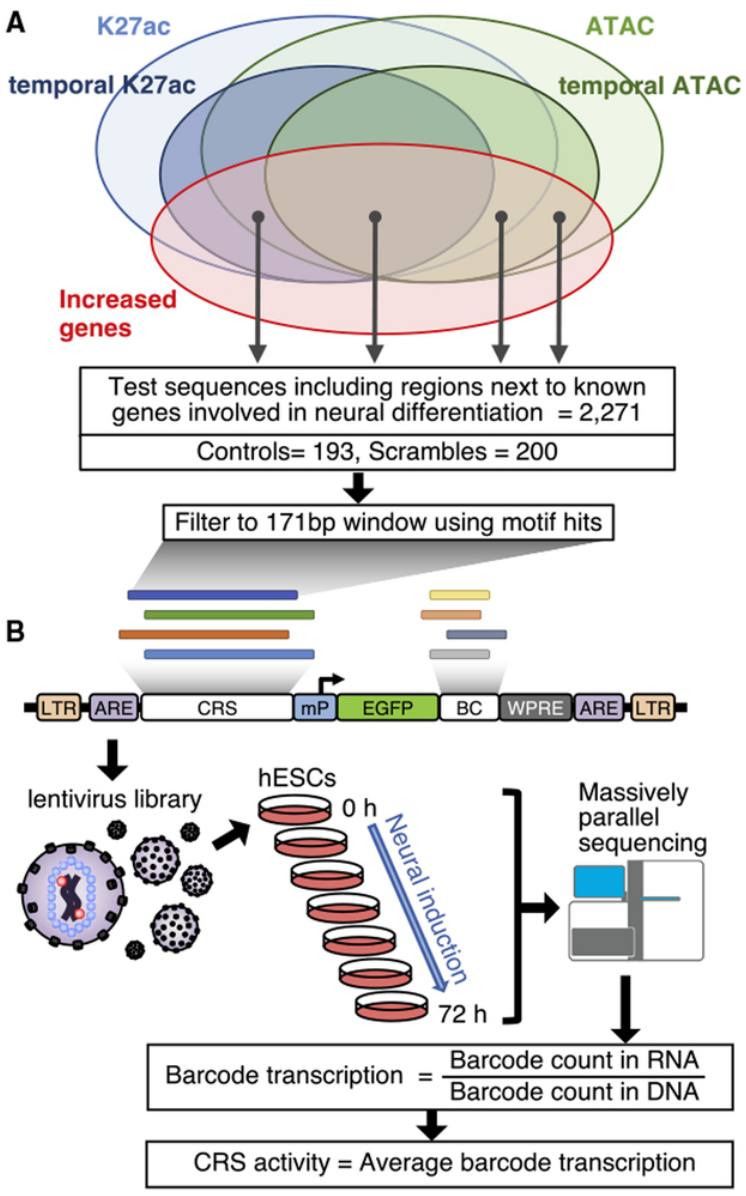


Figure 2: **Experimental Design of lentiMPRA. (A)** Sequence selection for lentiMPRA. 2,271 candidate regulatory regions (CRSs) were selected based on RNA-seq, H3K27ac ChIP-seq, and ATAC-seq data. Curated known enhancers (Table S2), 193 positive control regions, and 200 negative controls were included as well. **(B)** Schematic showing lentiMPRA design. CRSs along with 15-bp barcodes were synthesized on a custom array and cloned into a lentiMPRA vector. The library was packaged into lentivirus and infected into hESCs. The infected cells were cultured for 3 days to allow genomic integration. DNA and nuclear RNA were extracted at seven time points (0, 3, 6, 12, 24, 48, and 72 h) and subjected to sequencing followed by estimation of transcriptional activity. ARE, antirepressor element; BC, barcode; LTR, long terminal repeat; mP, minimal promoter; WPRE, woodchuck hepatitis virus posttranscriptional regulatory element.

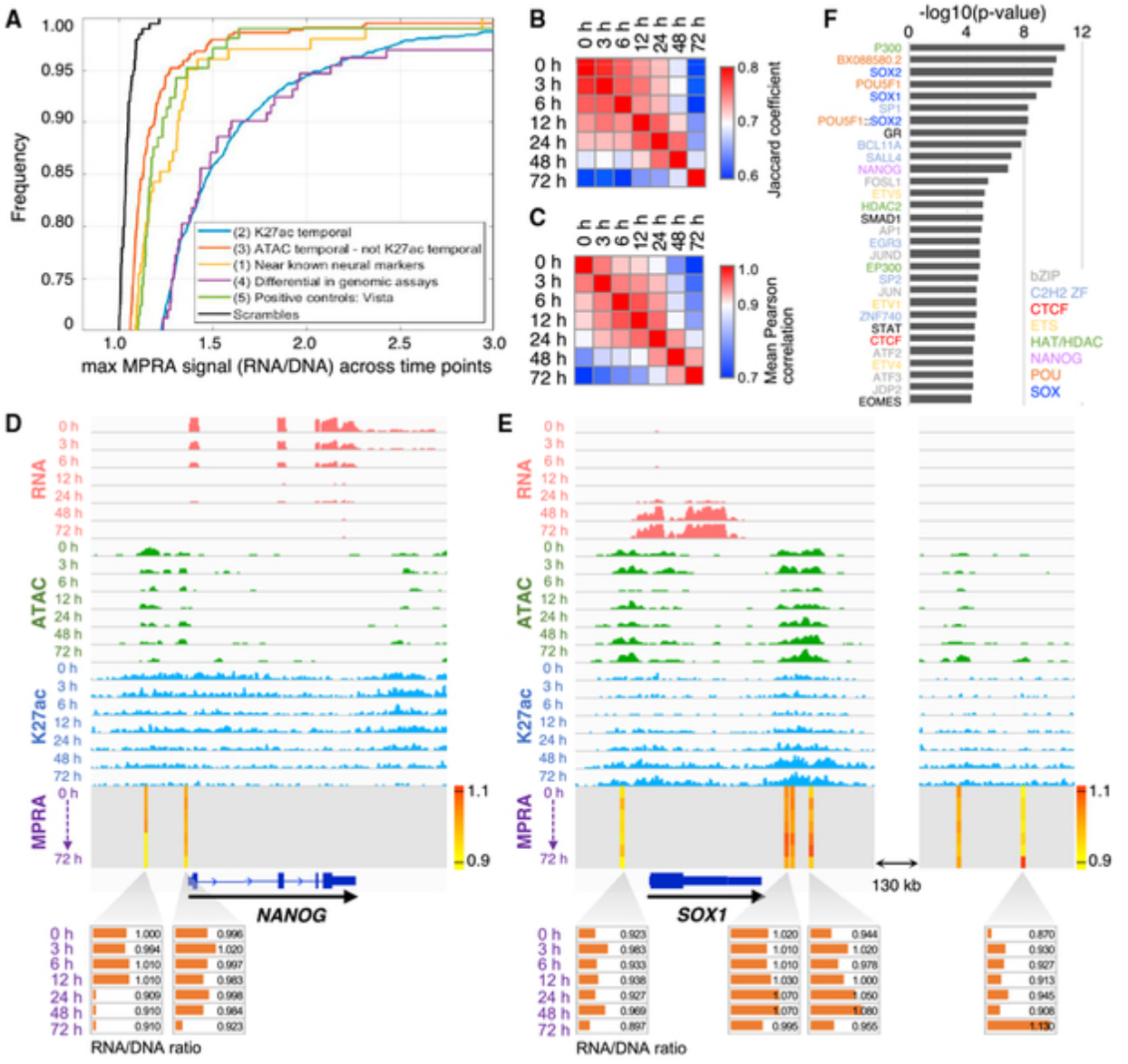


Figure 3: **lentiMPRA Signal for Different Enhancer Types. (A)** Cumulative distribution function indicating the frequency (y axis) of sequences with a specific MPRA signal (x axis; taking the maximum signal over time). Design criterion number (1–5) is indicated per each tested group of CRSs. **(B and C)** Similarity between the MPRA signal measured at different time points, using either the intersection of the sets of significantly active regions (Jaccard coefficient; **B**) or the correlation of the signals (Pearson correlation; **C**). **(D and E)** RNA-seq (red), ATAC-seq (green), H3K27ac ChIP-seq (blue), and MPRA (RNA/DNA ratio heatmap) tracks around NANOG **(D)** and SOX1 **(E)**. RNA/DNA ratio at each time point is shown as bar charts at the bottom. **(F)** Enrichment of predicted TF binding sites in temporal CRS. Top 30 differentially enriched TF binding sites when comparing temporal and non-temporal CRSs are shown (Fisher's exact test). TF categories are indicated on the right side.

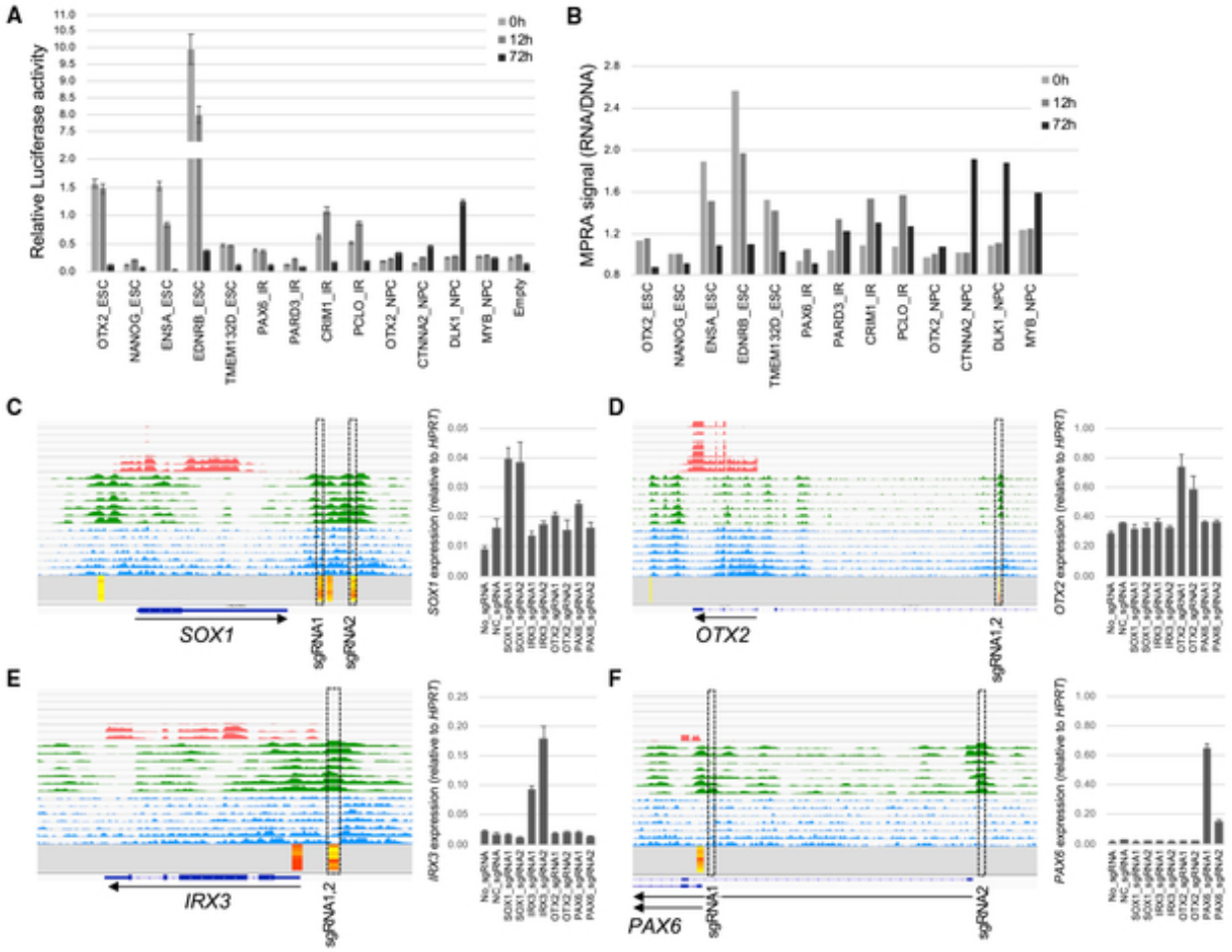


Figure 4: **Validation of Enhancers by Luciferase Assays and CRISPRa. (A)** Relative luciferase activity for each enhancer compared to Renilla luciferase activity at 0, 12, and 72 h post-neural induction. Five ESC enhancers, four immediate response (IR) enhancers, four NPC enhancers, and empty pLS-mP-Luc vector (negative control) were tested. **(B)** MPRA signal (RNA/DNA ratio) at 0, 12, and 72 h post-neural induction. **(C–F)** Functional validation of enhancers by CRISPRa. sgRNAs that target enhancers nearby SOX1 **(C)**, OTX2 **(D)**, IRX3 **(E)**, and PAX6 alternative promoters **(F)** or negative control sgRNA (NC_sgRNA) were infected into hESCs that stably express dCas9-VP64. Upregulation of respective genes relative to HPRT were examined by qPCR and shown as bar charts on the right. Data are presented as means ± SD of three independent experiments.

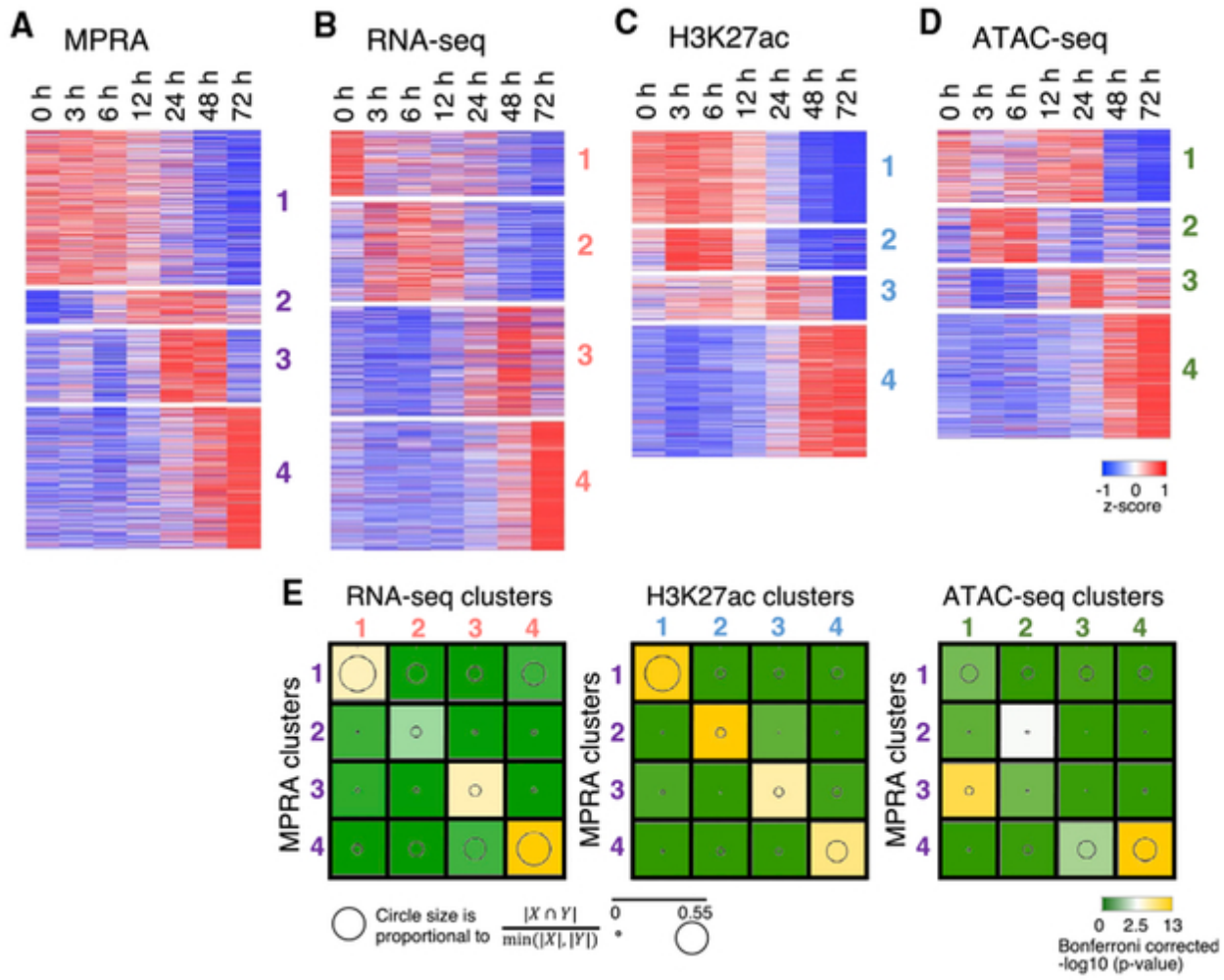


Figure 5: **Activity of Temporal CRS: Comparing lentiMPRA to the Endogenous Signals (A–D)** Temporal MPRA signal (RNA/DNA ratio; **A**), normalized read count of the closest gene detected by RNA-seq **(B)**, H3K27ac ChIP-seq **(C)**, and ATAC-seq **(D)**, clustered into four temporal groups separately. Rows are standardized. **(E)** Overlap between the lentiMPRA clusters and the three genomic data modalities. Shown are Bonferroni-corrected p values of a hypergeometric test. Circle sizes represent the proportion of overlap between every two clusters. The overlap is computed either at the region level (lentiMPRA versus ATAC-seq or ChIP-seq) or at the gene level (lentiMPRA versus RNA-seq; using the nearest gene to represent each lentiMPRA region).

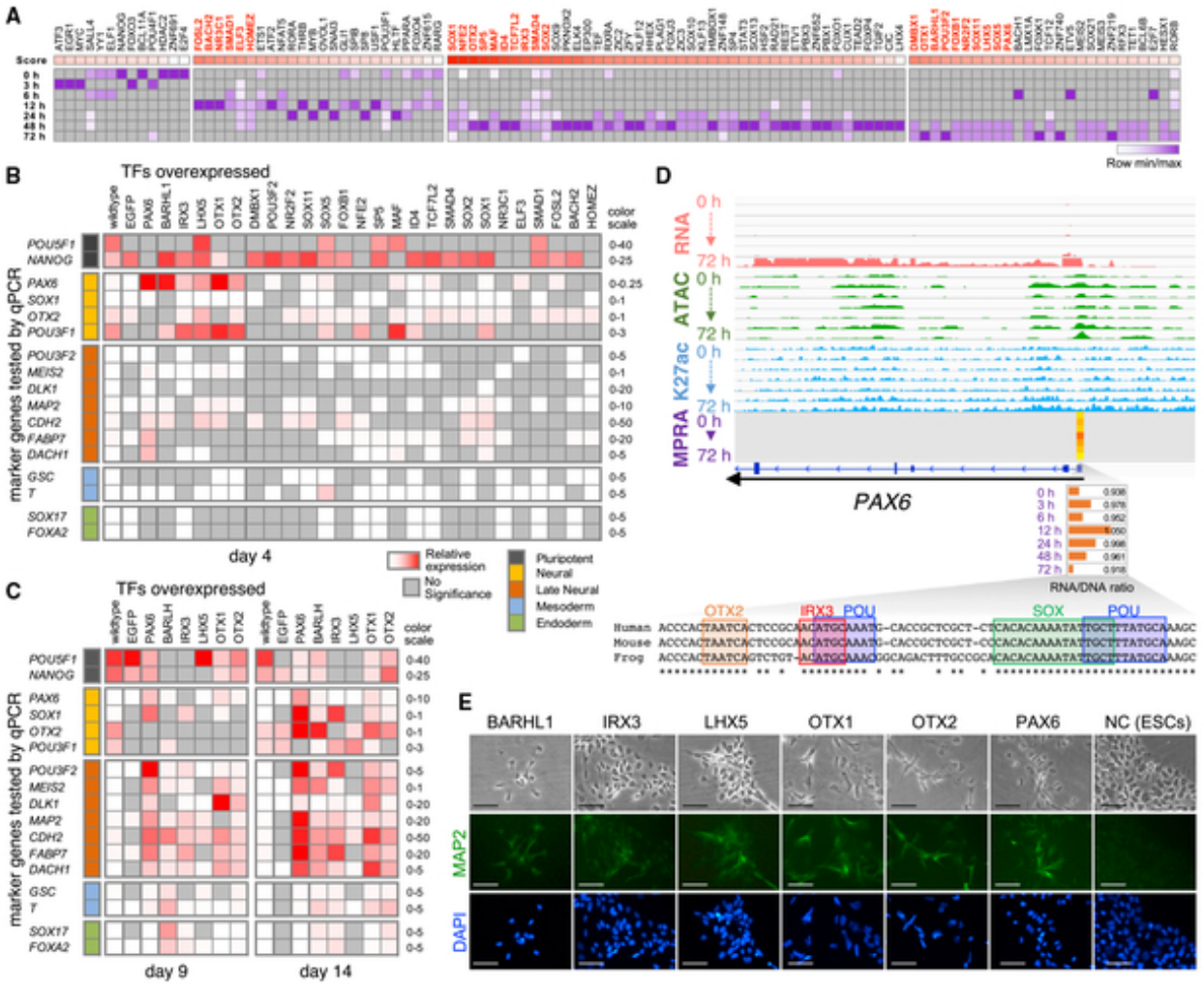


Figure 6: **Activity Score Identifies Novel TFs Involved in Neural Induction (A)** Heatmap of activity scores per TF per time point. Values are normalized (minimum to maximum) per each row and sorted by considering both the induction of the TF's mRNA expression and the overlap of the TF's targets with significant sub-clusters of MPRA activity for each cluster. The 26 TFs used for overexpression are marked in red font. **(B and C)** TF overexpression. Marker gene expressions (pluripotent, mesoderm, endoderm, and neural) are examined by qRT-PCR at early (**B**; day 4) and late (**C**; days 9 and 14) time points post-vector transduction. Relative expression compared to the HPRT gene is shown as a heatmap with the scale on the right side. Grey entries indicate no significant changes (Student's t test; $p > 0.05$). **(D)** TF analyses of the PAX6 promoter region show binding sites for OTX2, IRX3, POU, and SOX that are evolutionally conserved between human, mouse, and frog (Xenopus tropicalis). **(E)** MAP2 immunocytochemistry. hESCs overexpressing BARHL1, IRX3, LHX5, OTX1, OTX2, PAX6, and negative control (NC) were stained with MAP2. Bright field (top), MAP2 (middle), and DAPI (bottom) are shown. Scale bars represent 200 μm.

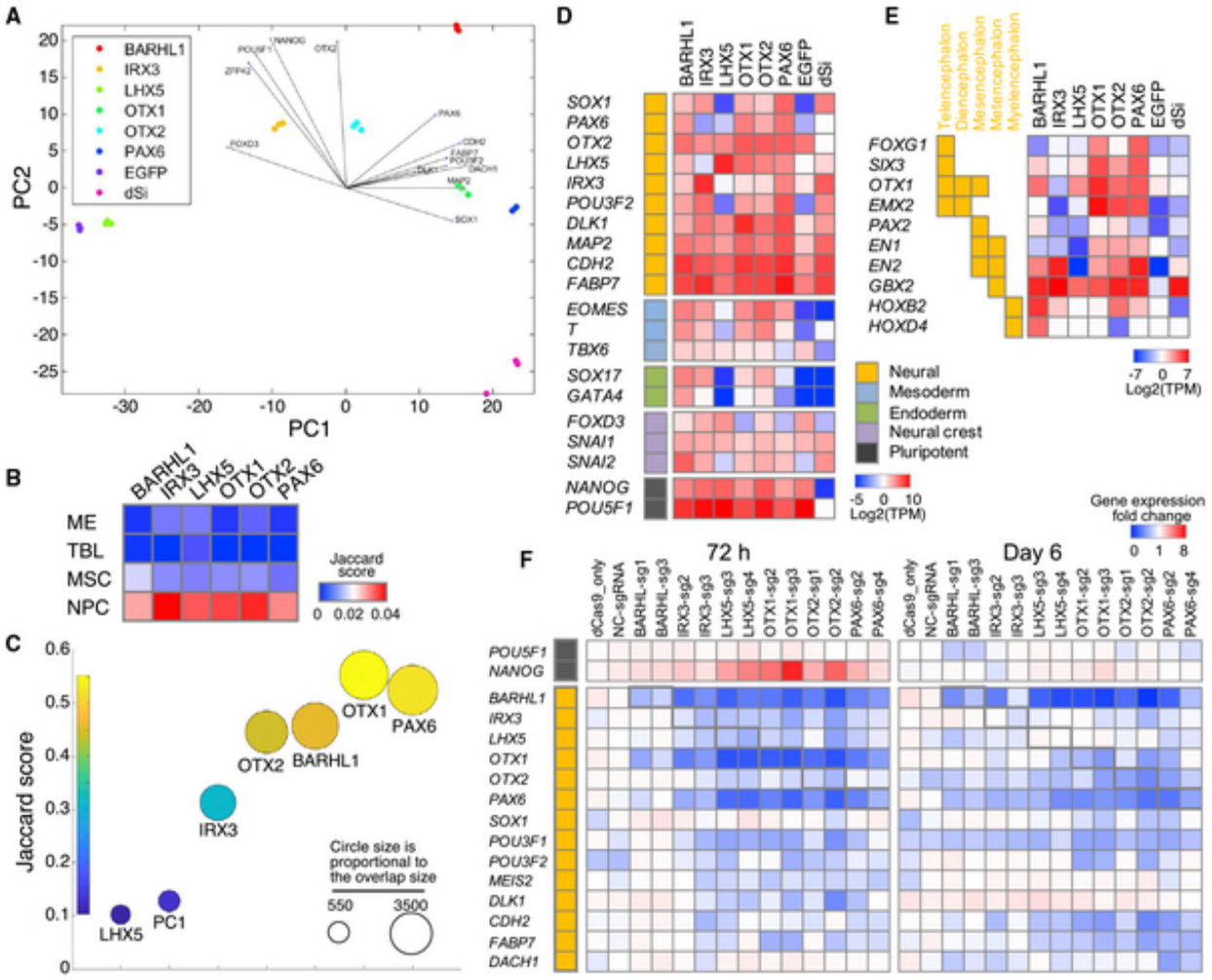


Figure 7: **RNA-seq for Cell Overexpressing TFs and CRISPRi (A)** PCA analysis of RNA-seq for overexpression of BARHL1, IRX3, LHX5, OTX1, OTX2, PAX6, and EGFP (negative control) and dSi (positive control) across three replicates, based on the 1,000 most variable genes. x axis PC1; y axis PC2. **(B)** Jaccard score for the overlap between lineage-restricted genes of four hESC-derived cell types (ME [mesendoderm]; MSCs [mesenchymal stem cells]; NPCs [neural progenitor cells]; TBLs, [trophoblast-like cells]) [8] and our DE genes (EGFP and factor). **(C)** Overlap between genes that are differentially expressed between the reference conditions (EGFP and dSi) and genes that are differentially expressed after overexpression using a Jaccard score (intersection over union; note that only genes that had consistent direction of change [upregulated in both or down-regulated in both] were considered to be a part of the intersection set). **(D and E)** TPM (log2, averaging over three biological replicates) for selected cell lineage markers (neural, mesoderm, endoderm, neural crest, and pluripotent; **D**) and brain regional markers (telencephalon, diencephalon, mesencephalon, metencephalon, and myelencephalon; **E**). **(F)** TF knockdown by CRISPRi. sgRNAs that target promoters of BARHL1, IRX3, LHX5, OTX1, OTX2, and PAX6 or negative control sgRNA (NC_sgRNA) were infected into hESCs along with dCas9-KRAB. Cells infected only with dCas9-KRAB (dCas9 only) were used as a negative control. Marker gene expression relative to HPRT was examined by qPCR at 72 h and 6 days after neural induction. Upregulation (red) or downregulation (blue) comparing to non-treated wild-type hESCs is shown as heatmap matrices.

6.7 References

[1] N. Yosef and A. Regev. "Writ large: Genomic dissection of the effect of cellular environment on immune response". fr. In: *Science* 354 (2016), pp. 64–68.

[2] R. Andersson et al. "An atlas of active enhancers across human cell types and tissues". en. In: *Nature* 507 (2014), pp. 455–461.

[3] E. Arner et al. "Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells". en. In: *Science* 80-.). 347 (2015), pp. 1010–1014.

[4] B.E. Bernstein et al. "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489 (2012), pp. 57–74.

[5] J.R. Dixon et al. "Chromatin architecture reorganization during stem cell differentiation". en. In: *Nature* 518 (2015), pp. 331–336.

[6] Casey A Gifford et al. "Transcriptional and epigenetic dynamics during specification of human embryonic stem cells". en. In: *Cell* 153.5 (May 2013), pp. 1149–1163. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2013.04.037. URL: http://dx.doi.org/10.1016/j.cell.2013.04.037.

[7] Alexander M Tsankov et al. "Transcription factor binding dynamics during human ES cell differentiation". en. In: *Nature* 518.7539 (Feb. 2015), pp. 344–349. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14233. URL: http://dx.doi.org/10.1038/nature14233.

[8] W. Xie et al. "Epigenomic analysis of multilineage differentiation of human embryonic stem cells". it. In: *Cell* 153 (2013), pp. 1134–1148.

[9] P. Kheradpour et al. "Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay". en. In: *Genome Res* 23.5 (May 2013), pp. 800–811. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.144899.112. URL: http://dx.doi.org/10.1101/gr.144899.112.

[10] Jamie C Kwasnieski et al. "High-throughput functional testing of ENCODE segmentation predictions". en. In: *Genome Res.* 24.10 (Oct. 2014), pp. 1595–1602. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.173518.114. URL: http://dx.doi.org/10.1101/gr.173518.114.

[11] Jacob C Ulirsch et al. "Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits". en. In: *Cell* 165.6 (June 2016), pp. 1530–1545. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.04.048. URL: http://dx.doi.org/10.1016/j.cell.2016.04.048.

[12] X. Wang et al. "High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human". en. In: *Nat. Commun* 9.1 (Dec. 2018), p. 5380. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07746-1. URL: http://dx.doi.org/10.1038/s41467-018-07746-1.

[13] M.J. Ziller et al. "Dissecting neural differentiation regulatory networks through epigenetic footprinting". en. In: *Nature* 518.7539 (Feb. 2015), pp. 355–359. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13990. URL: http://dx.doi.org/10.1038/nature13990.

[14] A. Fort et al. "Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance". en. In: *Nat. Genet* 46 (2014), pp. 558–566.

[15] D.W. Volk and D.A. Lewis. "Prenatal ontogeny as a susceptibility period for cortical GABA neuron disturbances in schizophrenia". es. In: *Neuroscience* 248 (2013), pp. 154–164.

[16] "A framework for the interpretation of de novo mutation in human disease". en. In: *Nat. Genet* 46 (2014), pp. 944–950.

[17] H. Kalkman. "A review of the evidence for the canonical Wnt pathway in autism spectrum disorders". en. In: *Mol. Autism* 3 (2012), p. 10.

[18] Lucia A Hindorff et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 106.23 (June 2009), pp. 9362–9367. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0903103106. URL: http://dx.doi.org/10.1073/pnas.0903103106.

[19] M.T. Maurano et al. "Systematic localization of common disease-associated variation in regulatory DNA". en. In: *Science* 337.6099 (2012), pp. 1190–1195. DOI: 10.1126/science.1222794. URL: http://dx.doi.org/10.1126/science.1222794.

[20] S.J. Sanders et al. "Whole genome sequencing in psychiatric disorders: the WGSPD consortium". it. In: *Nat. Neurosci* 20 (2017), pp. 1661–1668.

[21] S.M. Chambers et al. "Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling". en. In: *Nat. Biotechnol* 27.3 (2009), pp. 275–280. DOI: 10.1038/nbt.1529. URL: http://dx.doi.org/10.1038/nbt.1529.

[22] J. Sander, J.L. Schultze, and N. Yosef. "ImpulseDE: detection of differentially expressed genes in time series data using impulse models". en. In: *Bioinformatics* 33 (2017), pp. 757–759.

[23] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12 (2014), p. 550. ISSN: 1465-6906. DOI: 10.1186/s13059-014-0550-8. URL: http://dx.doi.org/10.1186/s13059-014-0550-8.

[24] A. Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". en. In: *Proc. Natl. Acad. Sci. U. S. A* 102.43 (Oct. 2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102. URL: http://dx.doi.org/10.1073/pnas.0506580102.

[25] J. Feng, T. Liu, and Y. Zhang. "Using MACS to Identify Peaks from ChIP-Seq Data". pt. In: *Curr. Protoc. Bioinforma* 34 (2011), pp. 2 14 1–2 14 14.

[26] E. Donnard et al. "Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon". en. In: *Cell Syst* 6 (2018), pp. 381–394 7.

[27] M. Garber et al. "A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals". en. In: *Mol. Cell* 47 (2012), pp. 810–822.

[28] M.R. Luizon et al. "Genomic Characterization of Metformin Hepatic Response". en. In: *PLoS Genet* 12 (2016).

[29] R.P. Smith et al. "Genome-wide discovery of drug-dependent human liver regulatory elements". en. In: *PLoS Genet* 10 (2014), p. 1004648.

[30] C.Y. McLean et al. "GREAT improves functional interpretation of cis-regulatory regions". en. In: *Nat. Biotechnol* 28 (2010), pp. 495–501.

[31] W. Herzog and K. Weber. "Fractionation of brain microtubule-associated proteins. Isolation of two different proteins which stimulate tubulin polymerization in vitro". en. In: *Eur. J. Biochem* 92 (1978), pp. 1–8.

[32] M. Endo et al. "Ror family receptor tyrosine kinases regulate the maintenance of neural progenitor cells in the developing neocortex". en. In: *J. Cell Sci* 125 (2012), pp. 2017–2029.

[33] K.G. GTEx Consortium et al. "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans". en. In: *Science* 348 (2015), pp. 648–660.

[34] R. Leslie, C.J. O'Donnell, and A.D. Johnson. "GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database". en. In: *Bioinformatics* 30 (2014), pp. 185–94.

[35] Axel Visel et al. "ChIP-seq accurately predicts tissue-specific activity of enhancers". en. In: *Nature* 457.7231 (Feb. 2009), pp. 854–858. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07730. URL: http://dx.doi.org/10.1038/nature07730.

[36] Charles E Grant, Timothy L Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif". en. In: *Bioinformatics* 27.7 (Apr. 2011), pp. 1017–1018. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btr064. URL: http://dx.doi.org/10.1093/bioinformatics/btr064.

[37] P. Kheradpour and M. Kellis. "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments". en. In: *Nucleic Acids Res* 42.5 (2014), pp. 2976–2987. DOI: 10.1093/nar/gkt1249. URL: http://dx.doi.org/10.1093/nar/gkt1249.

[38] Fumitaka Inoue et al. "A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity". en. In: *Genome Res.* 27.1 (Jan. 2017), pp. 38–52. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.212092.116. URL: http://dx.doi.org/10.1101/gr.212092.116.

[39] Anat Kreimer et al. "Predicting gene expression in massively parallel reporter assays: A comparative study". en. In: *Hum. Mutat.* 38.9 (Sept. 2017), pp. 1240–1250. ISSN: 1059-7794, 1098-1004. DOI: 10.1002/humu.23197. URL: http://dx.doi.org/10.1002/humu.23197.

[40] A. Kreimer et al. "Meta-analysis of massively parallel reporter assays enables prediction of regulatory function across cell types". en. In: *Hum. Mutat* umu.23820 (2019).

[41] A. Melnikov et al. *Massively parallel reporter assays in cultured mammalian cells*. fr. J Vis Exp, 2014.

[42] Tal Ashuach et al. "MPRAnalyze: statistical framework for massively parallel reporter assays". en. In: *Genome Biol.* 20.1 (Sept. 2019), pp. 1–17. ISSN: 1465-6906. DOI: 10.1186/s13059-019-1787-z. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1787-z.

[43] D.J. Rodda et al. "Transcriptional regulation of nanog by OCT4 and SOX2". en. In: *J. Biol. Chem* 280 (2005), pp. 24731–24737.

[44] Q. Wu et al. "Sall4 interacts with Nanog and co-occupies Nanog genomic sites in embryonic stem cells". en. In: *J. Biol. Chem* 281 (2006), pp. 24090–24094.

[45] L.A. Gilbert et al. "CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes". en. In: *Cell* 154 (2013), pp. 442–451.

[46] M.A. Lodato et al. "SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state". en. In: *PLoS Genet* 9.2 (Feb. 2013), p. 1003288. ISSN: 1553-7390, 1553-7404. DOI: 10.1371/journal.pgen.1003288. URL: http://dx.doi.org/10.1371/journal.pgen.1003288.

[47] M.T. Weirauch et al. "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity". en. In: *Cell* 158.6 (Sept. 2014), pp. 1431–1443. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2014.08.009. URL: http://dx.doi.org/10.1016/j.cell.2014.08.009.

[48] L.A. Boyer et al. "Core transcriptional regulatory circuitry in human embryonic stem cells". en. In: *Cell* 122 (2005), pp. 947–956.

[49] H.R. Herschman. "Primary Response Genes Induced by Growth Factors and Tumor Promoters". en. In: *Annu. Rev. Biochem* 60 (1991), pp. 281–319.

[50] S.R. Grossman et al. "Systematic dissection of genomic features determining transcription factor binding and enhancer function". en. In: *Proc. Natl. Acad. Sci. U. S. A* 114.7 (Feb. 2017), pp. 1291–1300. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1621150114. URL: http://dx.doi.org/10.1073/pnas.1621150114.

[51] N. Kashtan and U. Alon. "Spontaneous evolution of modularity and network motifs". en. In: *Proc. Natl. Acad. Sci. U. S. A* 102 (2005), pp. 13773–13778.

[52] N. Rosenfeld et al. "Gene regulation at the single-cell level". en. In: *Science* 307 (2005), pp. 1962–1965.

[53] Y. Setty et al. "Detailed map of a cis-regulatory input function". en. In: *Proc. Natl. Acad. Sci. U. S. A* 100 (2003), pp. 7702–7707.

[54] N. Yosef et al. "Dynamic regulatory network controlling TH17 cell differentiation". en. In: *Nature* 496.7446 (Apr. 2013), pp. 461–468. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11981. URL: http://dx.doi.org/10.1038/nature11981.

[55] X. Zhang et al. "Pax6 is a human neuroectoderm cell fate determinant". it. In: *Cell Stem Cell* 7.1 (July 2010), pp. 90–100. ISSN: 1934-5909, 1875-9777. DOI: 10.1016/j.stem.2010.04.017. URL: http://dx.doi.org/10.1016/j.stem.2010.04.017.

[56] "FGF signalling inhibits neural induction in human embryonic stem cells". en. In: *Embo J* 30 (2011), pp. 4874–4884.

[57] D. Acampora et al. "OTX1 compensates for OTX2 requirement in regionalisation of anterior neuroectoderm". fr. In: *Gene Expr Patterns* 3 (2003), pp. 497–501.

[58] P.-S. Hou et al. "LHX2 regulates the neural differentiation of human embryonic stem cells via transcriptional modulation of PAX6 and CER1". en. In: *Nucleic Acids Res* 41.16 (Sept. 2013), pp. 7753–7770. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt567. URL: http://dx.doi.org/10.1093/nar/gkt567.

[59] A. Bosse et al. "Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system". en. In: *Mech. Dev* 69 (1997), pp. 169–181.

[60] Nathaniel D Heintzman et al. "Histone modifications at human enhancers reflect global cell-type-specific gene expression". en. In: *Nature* 459.7243 (May 2009), pp. 108–112. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07829. URL: http://dx.doi.org/10.1038/nature07829.

[61] J. Grove et al. "Identification of common genetic risk variants for autism spectrum disorder". en. In: *Nat. Genet* 51 (2019), pp. 431–444.

[62] L. Torre-Ubieta et al. "The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis". en. In: *Cell* 172 (2018), pp. 289–304 18.

[63] J.D. Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". en. In: *Nat. Methods* 10 (2013), pp. 1213–1218.

[64] A.R. Quinlan and I.M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". en. In: *Bioinformatics* 26 (2010), pp. 841–842.

[65] J. Zhang et al. "PEAR: a fast and accurate Illumina Paired-End reAd mergeR". en. In: *Bioinformatics* 30 (2014), pp. 614–620.

[66] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows-Wheeler transform". en. In: *Bioinformatics* 25 (2009), pp. 1754–1760.

[67] D. Kim et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". en. In: *Genome Biol* 14 (2013), p. 36.

[68] A.M. Bolger, M. Lohse, and B. Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". fr. In: *Bioinformatics* 30 (2014), pp. 2114–2120.

[69] C. Trapnell et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". en. In: *Nat. Biotechnol* 28 (2010), pp. 511–515.

[70] Y. Liao, G.K. Smyth, and W. Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". en. In: *Bioinformatics* 30 (2014), pp. 923–930.

[71] B. Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". en. In: *Genome Biol* 10 (2009), p. 25.

[72] B. Langmead and S.L. Salzberg. "Fast gapped-read alignment with Bowtie 2". en. In: *Nat. Methods* 9 (2012), pp. 357–359.

[73] Y. Zhang et al. "Model-based Analysis of ChIP-Seq (MACS". en. In: *Genome Biol* 9 (2008), p. 137.

[74] J. MacArthur et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog". et. In: *Nucleic Acids Res* 45 (2017), pp. 896–901.

[75] J. Malone et al. "Modeling sample variables with an Experimental Factor Ontology". pt. In: *Bioinformatics* 26 (2010), pp. 1112–1118.

[76] A.D. Johnson et al. "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap". en. In: *Bioinformatics* 24 (2008), pp. 2938–2939.

[77] W.J. Kent et al. "The human genome browser at UCSC". en. In: *Genome Res* 12 (2002), pp. 996–1006.

[78] "Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium". en. In: *Mol. Autism* 8 (2017), p. 21.

6.8 Supplementary Figures

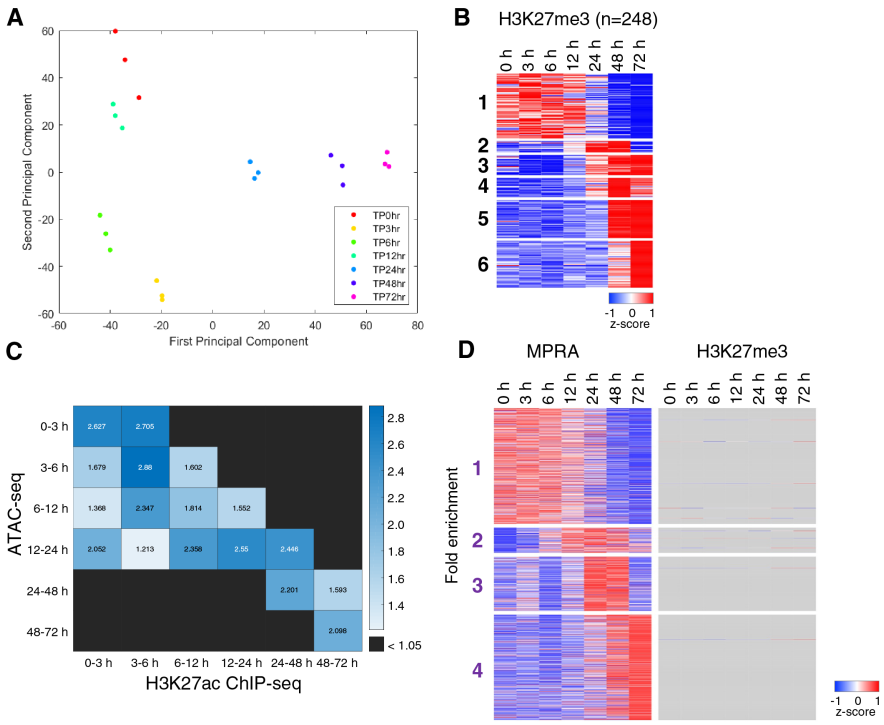


Figure S1: **PCA of RNA-seq data, temporal H3K27me3 ChIP-seq peaks and temporal ordering of ATACseq/H3K27ac.** Related to Figures 1 and 5. **(A)** Principal component analysis of the RNA-seq data, including seven time points and three replicates, is based on the top 5,000 variable genes. X-axis: first principal component. Y-axis: second principal component. **(B)** Heat map of scaled read counts (log2, averaged over three biological replicates and standardized per row) of H3K27me3 temporal peaks clustered into six groups. Red and blue indicate high and low signals, respectively. **(C)** Heatmap depicting fold enrichment of the overlap between the profiles of ATAC-seq and H3K27ac temporal regions (considering only regions that were called significantly temporal in both data modalities). Rows/ columns represent time segments [0-3 hours (h), 3-6 h, 6-12 h, 12-24 h, 24-48 h, 48-72 h] in which each region reaches is maximal expression in accessibility (rows) or H3K27 acetylation (columns). The maximal time segment is derived by averaging the values in each pair of subsequent time points and taking the maximum pair. For two sets S1 and S2 within a background set N, the fold enrichment is defined as $\left(\frac{|S1 \cap S2|}{|S1|}\right) / \left(\frac{|S2|}{|N|}\right)$. In this test, S1 and S2 correspond to the sets of regions that peak at the respective time segments; N is the set of all regions that are temporal in terms of both H3K27 acetylation and chromatin accessibility. **(D)** The dynamic changes of MPRA signal for the 1,547 temporal MPRA regions and their corresponding signal from an overlapping peak of H3K27me3 ChIP-seq. MPRA signal was clustered into 4 temporal groups separately. Grey entries indicate there was no intersecting peak with the MPRA region.

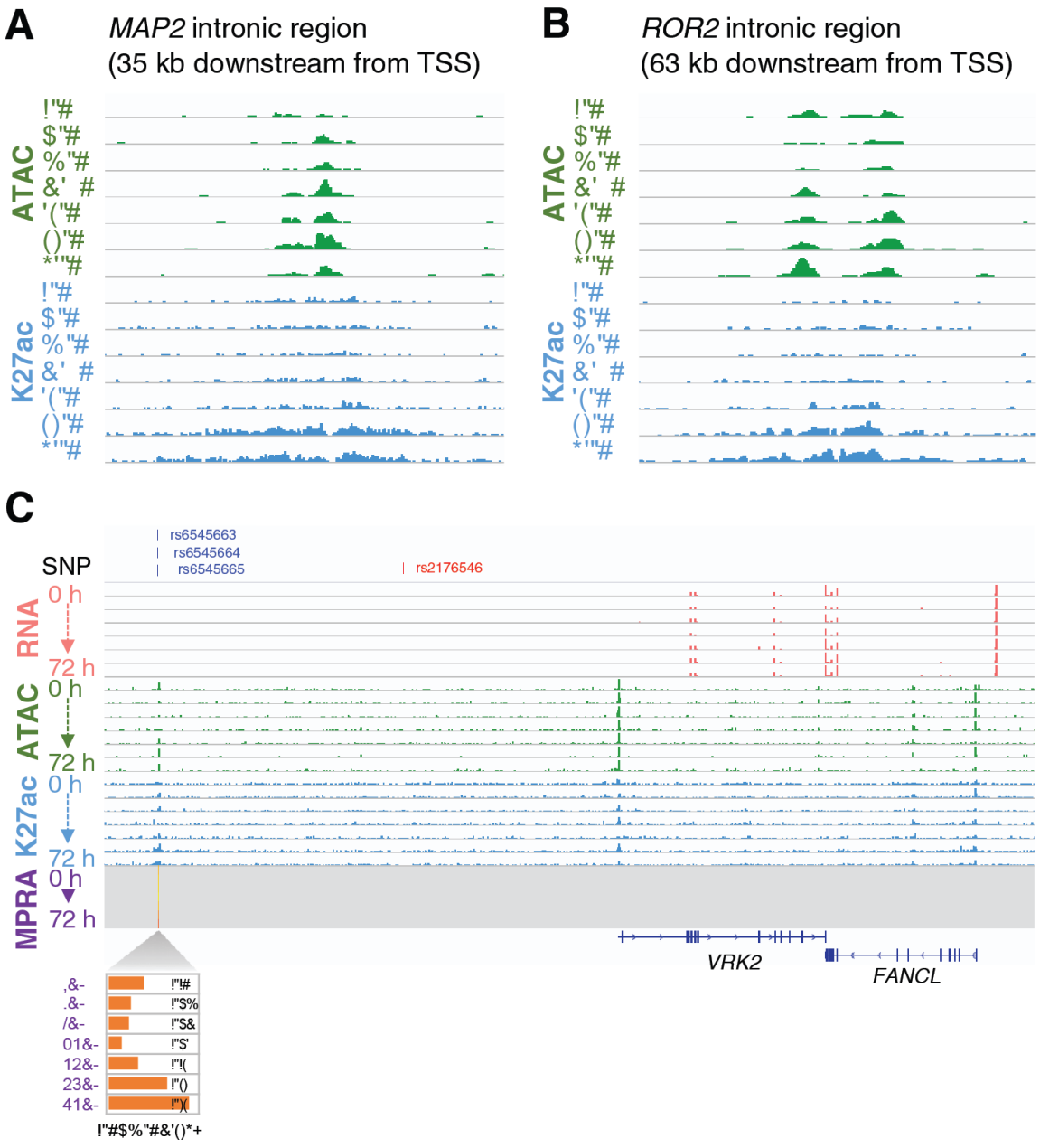


Figure S2: **Temporal ATAC-seq and H3K27ac signals at MAP2, ROR2, and VRK2/FANCL loci.** Related to Figures 1 and 3. **(A)** Potential enhancer region located within a MAP2 intron (chr2:210479842-210480012; hg19). **(B)** Potential enhancer regions located within a ROR2 intron (chr9:94647461-94647631 and chr9:94649070-94649240; hg19). **(C)** Enhancer region (chr2:58023768-58023938; hg19) located near VRK2 and FANCL. The SNPs track shows the ASD-associated lead SNP [78] in red and in addition SNPs that are in linkage disequilibrium with this SNP that reside within the H3K27ac peak in blue.

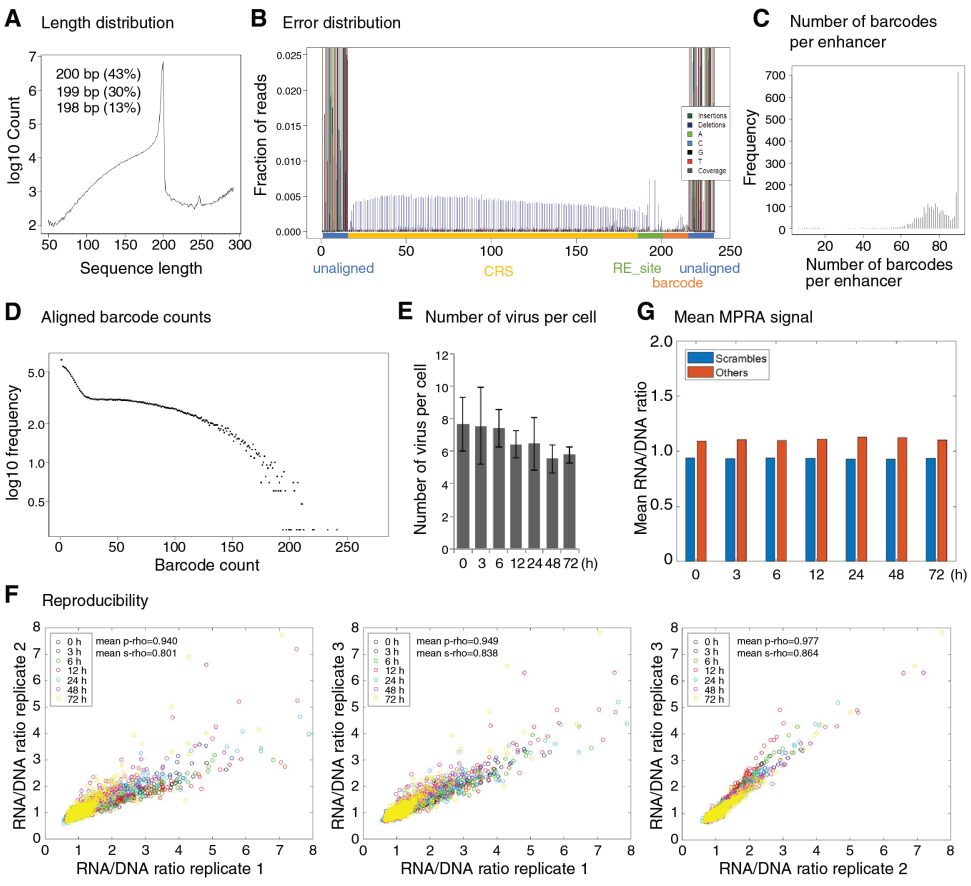


Figure S3: **lentiMPRA statistics and reproducibility.** Related to Figures 2 and 3. **(A)** Length distribution of the cloned sequence showing that the majority of the sequences were at the designed length (200 bp). **(B)** Position of errors/differences to the designed oligo sequences as observed from consensus-called BWA alignments. Bottom thick lines indicate the different oligo parts: candidate regulatory sequence (CRS), yellow; spacer, green; barcode, orange). **(C)** Number of barcodes associated per CRS. On average different barcodes were associated with each CRS in the cloned lentiMPRA library. **(D)** Number of times designed barcodes are observed. The frequency axis (y) has been log-transformed to show an over dispersion effect in the library, where a minority of barcodes contribute to many of the observations. **(E)** Copy number of viral particles per cell at each time point. Copy number of viral particles per cell was determined by qPCR with primers against WPRE compared to genomic primers for the intronic region of the LIPC gene. Data are presented as means ± SD of three independent experiments. **(F)** Reproducibility between biological and technical replicates and enhancer activity across time points. Enhancer activities measured by lentiMPRA were highly reproducible between biological replicates 1 vs. 2 (A, p-rho=0.94, s-rho=0.8) and 1 vs. 3 (B, p-rho=0.95, s-rho=0.84) (separate lentivirus prep) as well as between technical replicates 2 vs. 3 (C, p-rho=0.97, s-rho=0.86) (separate cell populations). **(G)** MPRA signal across time points of CRS vs. scrambled controls. Mean MPRA signal per time point for 2,464 CRS (red) vs. 200 scrambled sequences controls (blue).

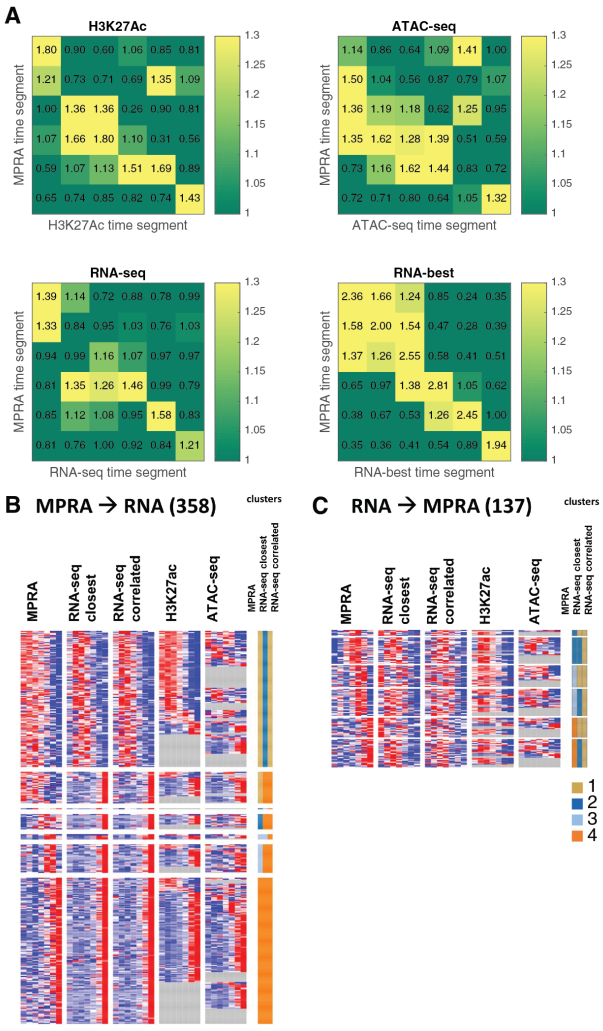


Figure S4: **Association between MPRA and endogenous temporal profiles.** Related to Figure 5. **(A)** Heatmaps depicting fold enrichment of the overlap between MPRA and endogenous temporal profiles. Rows/columns in each heatmap represent time segments [0-3 hours (h), 3-6 h, 6-12 h, 12-24 h, 24-48 h, 48-72 h] in which each region reaches is maximal expression in MPRA measurements (rows) or genomic signal (columns). The values for H3K27ac and ATAC-seq are computed based on the signal in the genomic coordinates that correspond to each MPRA region. The values in RNA-seq are based on the closest gene. The values in RNA-seq best are based on the most correlated gene, out of the four closest genes. **(B-C)** Heatmap of all signal measurements (MPRA, H3K27ac, ATAC-seq, RNA-seq of closest gene and RNA-seq of most correlated nearby gene) for the cases where the MPRA signal precedes the RNA signal **(B)** or RNA signal precedes the MPRA signal **(C)**. Each row corresponds to an MPRA region. Values are standardized per row per data source.

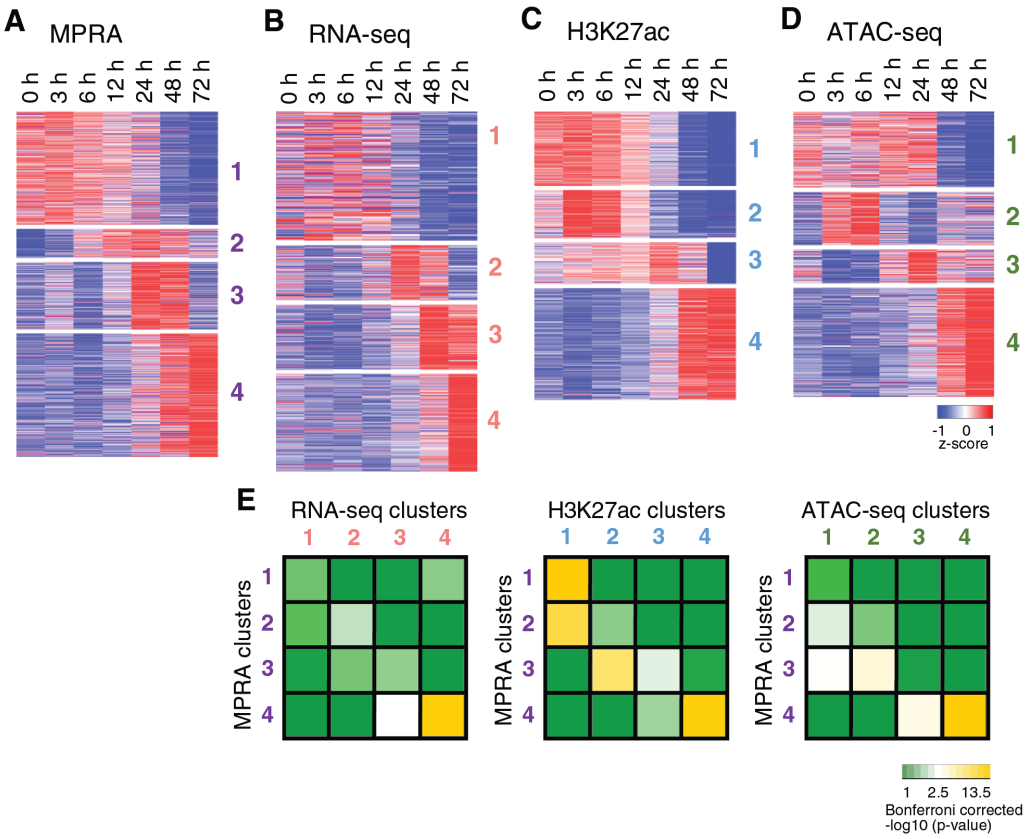


Figure S5: **Comparing lentiMPRA to the endogenous signals using a more restricted set of CRS.** Related to Figure 5. This figure is similar to Figure 5, where instead of the complete set of 1,547 temporal CRS, we include a more stringent set of 1,261 temporal CRS that are also detected by the per- time point analysis. **(A-D)** Temporal signal of MPRA activity **(A)** and the closest gene signal detected by RNA-seq **(B)** and H3K27ac ChIP-seq **(C)** and the corresponding signal from an overlapping peak of ATAC-seq **(D)** clustered into four temporal groups separately. Values shown are RNA/DNA ratio (for lentiMPRA) and normalized read counts (for all others). Rows are standardized. **(E)** Overlap between the lentiMPRA clusters and the three genomic data modalities. Shown are Bonferroni corrected p-values of a hypergeometric test. The overlap is computed either at the region level (lentiMPRA vs. ATAC-seq or ChIP-seq) or at the gene level (lentiMPRA vs. RNA-seq; using the nearest gene to represent each lentiMPRA region).

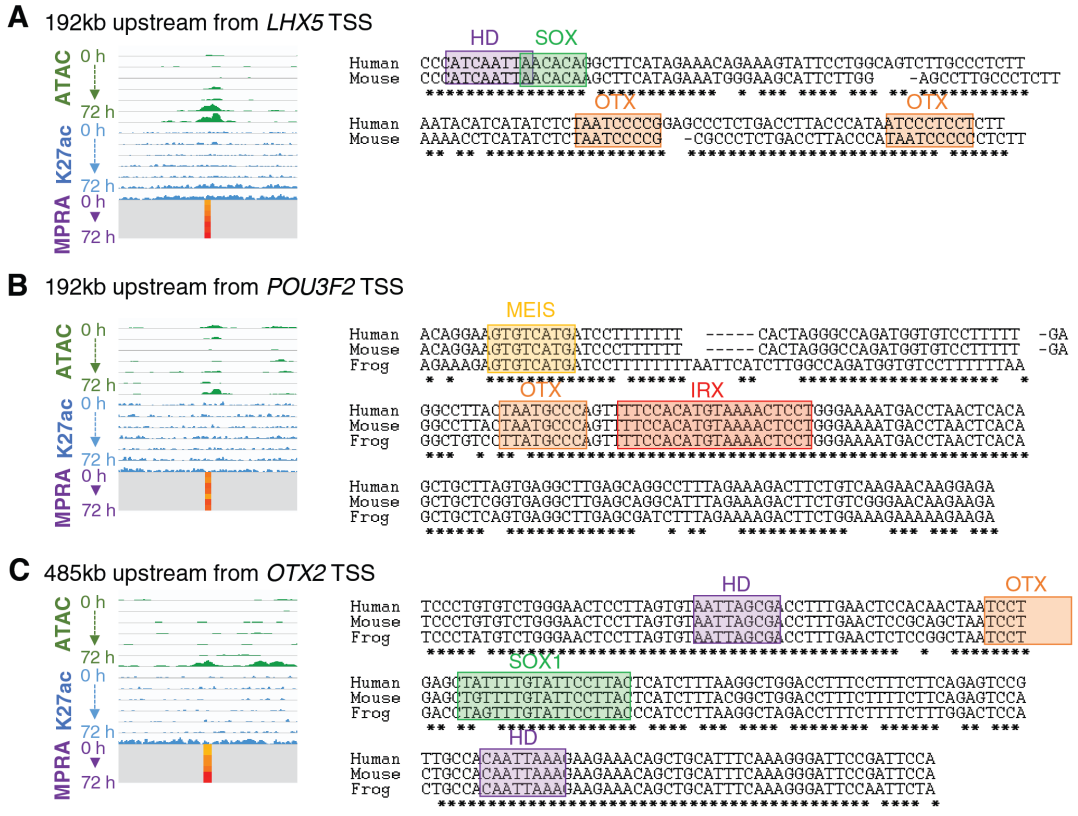


Figure S6: **Functional enhancers with conserved TF motifs near LHX5, POU3F2, and OTX2.** Related to Figure 6. **(A)** Enhancer region (chr12:114101961-114102131; hg19) located near LHX5 contains Homeodomain (HD), SOX, and two OTX motifs that are conserved between human and mouse. **(B)** Enhancer region (chr6:99090064-99090234; hg19) located near POU3F2 contains MEIS, OTX, and IRX motifs that are conserved between human, mouse and frog (Xenopus tropicalis). **(C)** Enhancer region (chr14:57757595-57757765; hg19) located near OTX2 contains two HD motifs, OTX, and SOX1 motifs that are conserved between human, mouse and frog (Xenopus tropicalis).

6.9 Supplemental Information

All supplemental information for this chapter is included in *Chapter6_Additional_Files.zip*. The files are:

- **Data S1** lentiMPRA Array Design Sequences, Related to Figure 2.
- **Table S1** Peak Calling, Clustering, and Enrichment Analysis for ChIP-seq, ATAC-seq, and RNA-seq Data, Related to Figure 1.
- **Table S2** Design of lentiMPRA Library, Related to Figure 2.
- **Table S3** lentiMPRA Experimental Statistics, Related to Figures 2 and 3.
- **Table S4** MPRA Clusters, Related to Figure 5.
- **Table S5** Transcription Factor Enrichment and Activity Score, Related to Figure 6.
- **Table S6** Enrichment in GO Annotations for Differentially Expressed Genes from RNA-seq Datasets Generated from Transcription Factor Overexpression, Related to Figure 7.
- **Table S7** Primer Sequences, Related to Methods.

Chapter 7

Massively parallel reporter perturbation assay uncovers temporal regulatory architecture during neural differentiation

This chapter has been accepted for publication in *Nature Communications*. It was previously posted on bioRxiv (2021), and is included here in the most recent form. The authors on the paper are:

Anat Kreimer[1,2,3,4,*,†], Tal Ashuach[3,*], Fumitaka Inoue[1,2,5,*], Alex Khodaverdian[3], Nir Yosef[3,6,7,†], Nadav Ahituv[1,2,†]

1. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA
2. Institute for Human Genetics, University of California, San Francisco, San Francisco, CA 94158, USA
3. Department of Electrical Engineering and Computer Sciences and Center for Computational Biology, University of California, Berkeley, CA 94720, USA
4. Department of Biochemistry and Molecular Biology, Center for Advanced Biotechnology and Medicine, Rutgers–Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA
5. Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, 606-8501, Japan
6. Chan-Zuckerberg Biohub, San Francisco, CA 94158, USA
7. Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA 02139, USA

∗ these authors contributed equally to the work

† Corresponding author

7.1 Abstract

Gene regulatory elements play a key role in orchestrating gene expression during cellular differentiation, but what determines their function over time remains largely unknown. Here, we performed perturbation-based massively parallel reporter assays at seven early time points of neural differentiation to systematically characterize how regulatory elements and motifs within them guide cellular differentiation. By perturbing over 2,000 putative DNA binding motifs in active regulatory regions, we delineated four categories of functional elements, and observed that activity direction is mostly determined by the sequence itself, while the magnitude of effect depends on the cellular environment. We also find that fine-tuning transcription rates is often achieved by a combined activity of adjacent activating and repressing elements. Our work provides a blueprint for the sequence components needed to induce different transcriptional patterns in general and specifically during neural differentiation.

7.2 Introduction

Enhancers are DNA sequences containing clustered recognition sites (i.e. motifs) for transcription factors (TFs) that play a pivotal role in transcriptional regulation of gene expression during numerous biological processes, including cellular differentiation [1]. This is evident by the abundance of disease associated variants discovered through genome-wide association studies (GWAS) and expression quantitative trait loci (eQTLs) residing in noncoding regions [2]. Despite their importance, our understanding of the regulatory grammar of enhancers, namely the manner by which their DNA sequences pertain to their function remains largely unknown, thus limiting our ability to infer how changes in these sequences affect their functionality and lead to higher- level consequences.

Various biochemical assays (e.g. ChIP-seq, DNase-seq, ATAC-seq) have enabled genome-wide identification and characterization of candidate regulatory sequences such as enhancers, across different cell types [3], providing descriptive maps of the human genome. Complementary studies use genome modification approaches, such as CRISPR-Cas9, to functionally characterize enhancer elements by targeting their locations in the genome [4]. Such assays capture both direct and indirect causal relationships between the tested regulatory elements and cellular phenotype (e.g. gene expression) and in many cases target regions that are bound by specific transcription factors of interest [5]. Massively parallel reporter assays (MPRAs) provide an alternative approach that enables simultaneously testing the regulatory activity of thousands of regulatory sequences and their variants. In MPRA, a copy of any sequence of interest is synthesized in front of a transcribed barcode. There are many variants to this technology [6], including one that utilizes lentivirus to integrate into the genome (hereafter we refer to this as lentiMPRA; [7]) used for these assays. The ratio between the abundance of a transcribed barcode (read with RNA-seq) and the number of coding sequences (evaluated with DNA- sequencing) provides a quantitative readout for the regulatory activity of the assayed sequence [6, 8, 9, 10, 11, 12].

Approaches to understanding the roles of TFs in determining the activity of a given enhancer and the interplay between TFs in an enhancer [13] are generally limited by the number of causal relationships they can study directly (e.g., via gene knockdown), primarily due to cost and availability of efficient perturbing agents. Therefore large scale studies often use correlational inference, e.g. associating TF binding with changes in gene expression based on motif- gene association [14]. These, however, are confounded by a slew of observations whereby only a small fraction of potential TF-binding sites (TFBSs) are actually occupied in any given cell type, and these sites vary substantially across cell types and conditions [15, 16, 17, 18]. Another caveat of perturbing endogenous factors that affect gene expression (e.g., TFs, enhancer regions) is the abundance of indirect effects, which are difficult to discern from the direct ones. These two issues are mitigated by MPRAs, as it provides a cost-effective approach to investigate thousands of candidate enhancer sequences along with variants of these sequences in which certain DNA binding motifs are perturbed. The concern for indirect effects is mitigated to some extent as well due to the synthetic nature of the assay (i.e., the transcribed barcode is non-functional). Previous approaches to perturbation MPRA for sequence motifs were limited to several factors and a specific cell-type or condition. For example, a previous study [19, 8] explored the activity of five activator motifs and two repressor motifs in K562 and HepG2 cells by introducing different variations to the motif sequence and another study [19] disrupted a single motif (PPARγ) in mouse adipocytes. Altogether, they focused on a specific time point and not a temporal course or developmental process.

The differentiation of stem cells into the neural lineage provides an exemplary model for studying how gradual and non-reversible changes to the cell's phenotype may be transcriptionally regulated. During this process, stem cells rapidly differentiate both on a molecular and physiological level to generate neurons. We previously characterized the temporal dynamics of gene expression (RNA-seq) and gene regulation (ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and lentiMPRA) at seven time points (0-72 hours) during the early parts of this process [20]. Using lentiMPRA, we identified numerous endogenous sequences that had temporal enhancer activity (i.e. the expression of their target barcode was well over the background levels and significantly changed over time). This activity tended to correlate with cell-endogenous changes to the expression of their target gene and to the structure of their surrounding chromatin. In addition, the genomic positions of the validated temporal sequences significantly overlapped with loci that have been associated with neurodevelopmental disorders, in particular autism spectrum disorder (ASD). Combining all our genomic data, we developed a prioritization method to select TFs that are putatively involved in driving a neural fate, and validated the role of several candidates with direct genetic perturbations. This study, however, was still limited to validations of a handful of TFs and lacks in understanding of the way by which these TFs may drive changes in transcription over time.

To more comprehensively identify DNA binding motifs that may affect transcription and characterize the timing in which they carry out their effect, we utilized a 'perturbation MPRA' approach. Based on our previous data, we compiled a list of 591 regulatory sequences whose activity differed over time (considering different temporal patterns) as well

as a selected set of 255 motifs within those regions. We then prioritized for testing 2,144 instances of the selected motifs in the selected regions. We used leniMPRA to perturb, via three different approaches, the selected instances, and evaluated their effect over at the same seven time points (0-72 hours) during the neural differentiation process. Using this approach, we found that 27% (598) of the perturbations had a significant effect on the transcription of the reporter gene. We divided these motif instances into several subtypes based on the direction (suppressing or inducing transcription) and magnitude (fold change, compared to the unperturbed and negative control sequences) of their effect. We observed that the magnitude of the effects often varied over time (indicating that it depends on the cellular environment), while the direction of the effect is independent of time and is broadly determined by the DNA sequence (i.e the combination of the perturbed motif and the surrounding region). Furthermore, we observed cases of activating and repressing motif instances that are harbored within the same regulatory region, suggesting that in those cases, fine-tuning of transcription levels may be achieved by a combination of opposing effects. Finally, by perturbing pairs of motifs in a select set of sequences, we found evidence for different patterns of cooperation between motifs, and that both fundamental models, namely the 'enhanceosome' model of an all-or-nothing machinery, and the 'billboard' model of independent contribution [21, 17] are supported by our data. Overall, our findings suggest that the regulatory grammar of enhancers that changed in gene expression in our system is an amalgam of a wide variety of different mechanisms. It also helps establish perturbation MPRA as a powerful approach for high-throughput investigation of such mechanisms in different cellular contexts.

7.3 Results

Selection of regions and motifs for perturbation MPRA

To characterize the effect of DNA binding motifs on gene expression over time, we first set out to choose a set of regulatory regions that showed temporal activity during early neural differentiation using lentiMPRA data from our previous study (0, 3, 6, 12, 24, 48, and 72 hours post induction; [20]). Our initial candidate set consisted of 1,547 171bp sequences that were identified as temporally active (i.e. the expression of their target barcode varied significantly, both over time and in comparison to a control sequence; Methods; [22]). We then used FIMO [23] to computationally identify occurences of known DNA binding motifs in each sequence (using motifs identified by Kheradpour and Kellis [24] and Weirauch et al. [25]).

Following these analyses, we chose specific regions and motifs for perturbation lentiMPRA. As we are limited by the number of sequences that can be included in a single lentiMPRA library due to low integration rate in ESCs, we developed an optimization framework to select the combination of regions and motifs that maximizes the representation of relevant genomic properties (Fig. 1a-b; Methods). To this end, we wanted to include regions and motifs that are associated with different temporal patterns of chromatin and gene expression

signals, derived from our previous analysis of H3K27ac ChIP-seq, ATAC-seq and RNA-seq data in the same time points [20]. We made sure to include a sufficient number of regions in which H3K27ac is induced early in the differentiation, as well as regions that gain this mark later on, and closer to the neural progenitor (NP) phase. Similarly, we selected a minimal number of motifs whose corresponding TFs are induced early in the differentiation process, as well as motifs associated with late- induced TFs. We also chose to provide explicit preference for a curated list of regions and TFs that have been previously associated with neural induction pathways. Finally, we required that every selected motif will be perturbed in at least 20 regions (thus allowing us to observe the motif in multiple contexts), and every selected region will have at least two different perturbations (for two different motifs). With these considerations taken together, the respective experimental design problem can be represented as an optimization problem: selecting the minimal number of motif instances [(region x motif) pairs] while satisfying all of our design constraints above. In the methods section, we describe how we represent this as a connectivity problem in graphs and how we derive a solution for it using Integer Linear Programming. Applying this scheme to our data resulted in a selection of 2,144 motif instances over 591 regions and 255 motifs (Fig. 1a-c).

We considered the 2,144 motif instances both in their wild type (WT) and in a perturbed form (PERT) where the sequence of the motif instance is modified in order to estimate its effect. For 100 of our genomic regions, chosen by the motifs they harbor and their importance for neural differentiation ([20]; see Methods), we also perturbed pairs of motifs (including two appearances of the same motif in the sequence or two different motifs), to analyze cooperative effects (Fig. 1c; Methods). We perturbed each of the selected motifs using three different designs that rely on (Fig. 1d), two approaches: In designs 1 and 2 we identified two fixed "non-motif" sequences (i.e sequences with minimal number of predicted motif hits - details in Methods) and replaced the motif with the prefix of these sequences, adjusting to the motif length. In the third design we randomly shuffled the nucleotides of the motif (Methods). We also included two sets of negative controls: 1) scrambled sequences (SCRAM) – where we shuffle all nucleotides of each of the 591 WT sequences; 2) random sequence alterations (RAND) – where we randomly shuffled a small region (length of the median motif size) at a random location in each region. In total, 10,041 sequences were included in our lentiMPRA library (Fig. 1c-d).

lentiMPRA perturbation

The designed sequences were synthesized and cloned upstream of a minimal promoter (mP) into the lentiMPRA vector (Fig. 1d; Methods). During the cloning process, 15-bp random barcodes were placed in the 5'UTR of the EGFP reporter gene [26]. The association between the cloned sequences and barcodes was determined via DNA-seq (Methods). Lentivirus was generated and human embryonic stem cells (hESCs) were infected with the library (Fig. 1e). Following three days, to allow for viral integration and degradation of unintegrated virus, the hESC were differentiated to a neural lineage using the dual-Smad inhibition protocol [27]. Integrated DNA barcodes and transcribed RNA barcodes were quantified by DNA-seq and

RNA-seq, respectively, at seven time points of neural differentiation (0, 3, 6, 12, 24, 48, and 72 hours) (Fig. 1f). The library infections were carried out using three biological replicates (two replicates were infected with the same lentivirus batch, while the other replicate was infected with another lentivirus batch).

Using a computational pipeline developed in our group, MPRAflow [28], we took a stringent approach to associate barcodes with the cloned sequences. For each barcode, we required at least 80% of the reads associated with the barcode to map it to a single sequence, and a minimum of 3 reads supporting that assignment, resulting in over 1.4 million confidently assigned barcodes, and averaging 139 barcodes per sequence (Methods). We then analyzed the barcodes from the lentiMPRA infected cells and matched them with the confidently assigned barcodes of the library. Across biological replicates, we were able to confidently assign an average of 61.6% of the barcodes (Methods and Supplementary Fig. 1). Considering only confidently assigned barcodes that have a representation both in RNA and DNA from infected cells, we observed an average of 134.4 barcodes per sequence in each replicate (Supplementary Fig. 2), corresponding to 9,948 out of the 10,041 designed sequences (2,082, 2,086, and 2,114 sequences for perturbation methods 1-3 respectively (Supplementary Table 1)). We then used MPRAnalyze [22] to aggregate the barcodes and quantify the transcription rate induced by each tested sequence (dubbed 'alpha'). We observed reproducible results between replicates (average Pearson correlation 0.98) in every time-point (Supplementary Fig. 3), and results were highly concordant with our previously characterized lentiMPRA in the same system [20] (mean Pearson correlation 0.79, Supplementary Fig. 4). Comparing the four categories of sequences that we tested, we observe as expected, that overall, the scrambled negative controls (SCRAM) had the lowest transcriptional activity, while the unperturbed sequences (WT) had the highest (Supplementary Fig. 5). We also observed that sequences with a perturbed binding site (PERT) had a generally lower level of activity than sequences with a perturbation of random sites (RAND), confirming that perturbing known motifs has an effect larger than expected by chance. We next quantified the magnitude of deviation between PERT and WT transcription rates (Log(WT/PERT)) and compared the results between all three perturbation methods. Overall, we observed correlated results between the three methods, both in terms of the estimated transcription rate of the perturbed sequences (average Pearson correlation 0.81) and the differential activity between the perturbed sequences and their corresponding WT sequence (Log(FC), average Pearson correlation 0.71) (Supplementary Fig. 6).

Identification of functional TF motifs

We next set out to identify which of the DNA binding motifs we assayed is a functional site, i.e a site that causes a significant change in regulatory activity when perturbed. To this end, we initially focused on sequences with a single perturbed site (rather than deletion of two sites) and used MPRAnalyze [22] to apply a set of four filters (illustrated in Fig. 2a; Supplementary Table 1a), requiring that each tested sequence passes all four filters: 1) the PERT sequence activity significantly deviates from that of the WT sequence in at least one

time point (likelihood ratio test (LRT); FDR< 0.05; Methods) ; 2) the time course of PERT activity significantly deviates from that of the WT sequence (LRT; FDR< 0.05; Methods); 3) either the PERT (in at least one time point) or WT (in all the time points) sequences are significantly more active than the SCRAM negative controls (MAD-based z-test; FDR< 0.05; Methods); 4) either the PERT or the WT sequence temporal activity significantly deviate from the temporal activity observed among the SCRAM negative control sequences (LRT; FDR< 0.05; Methods). Overall, these filters will include sites that when perturbed cause a significant change (compared to WT) in regulatory activity in at least one time point (filter 1) and across the temporal pattern (filter 2). Additionally, sequences that are potentially activating or repressing in at least one time point or across time will be included using filters 3 and 4. Our analysis will not remove constitutive sequences as long as their temporal activity is significantly different from the WT. We applied these filters to each perturbation method separately, which resulted in 747, 775, and 749 sequences in perturbation methods 1, 2 and 3 respectively (Fig. 2b; Supplementary Fig. 7 and Supplementary Table 1a). Across the 3 perturbation methods, we observe that most of the sequences pass all 4 filters and less than 10% of the sequences pass no filter, indicating that our experimental design mainly consists of functional regulatory sequences across these time points of neural induction (Supplementary Fig. 7). Comparison analysis of temporal properties from our previous work [20] confirmed that the signal of H3K27ac, ATAC-seq, MPRA and mRNA of both the closest gene and the motif's associated TF were significantly lower (all time points combined, Wilcoxon p-value$< 10^{-10}$) in removed vs. retained sequences, in each of the perturbation methods, supporting our filtering approach (Supplementary Table 2). We observed an overall similar level of concordance between the different methods, with an average overlap of $\tilde{7}0\%$ between the three methods (Fig. 2b; Supplementary Fig. 6 and Supplementary Table 1).

In the subsequent analysis, we took a conservative approach to aggregate the evidence from the three ways of perturbing motif instances. We focused on instances that had strong evidence from both approaches for perturbing a motif (i.e., random shuffle or replacement by a fixed "non-motif" sequence). To this end, we consider only instances that passed all four filters above in perturbation method 3 and in either perturbation methods 1 or 2. We also require that the direction of effect (increasing or decreasing expression) to be consistent between the different methods. This resulted in 598 motif instances that had a significant and consistent effect. We refer to this set as functional regulatory sites (FRSs).

We examined the FRSs by conducting our analysis in three different axes: (i) the FRS level, i.e. perturbation of a specific motif in a specific region; (ii) the motif level, across different regions the motif appears in; (iii) the region level, taking into account the various functional sites that appear in it. For each axis we also examined how the perturbation effect may change over the different time points. While our analysis is based on the consensus set of 598 motif instances, we repeated it based on sites found by each of our three perturbation methods individually, where we observe largely consistent results (Fig. 2b; Supplementary Tables 1-2;4-7; Methods).

Delineating major categories of functional regulatory sites

We first analyzed the general effect of our perturbations in all 598 FRSs. Comparing the MPRA signal of WT to PERT sequences in each time point, we generally observed a reduction in activity, indicating that perturbing the predicted motif disrupts the function of an activating TF. For a smaller portion of the sites we observed the opposite effect, i.e. increased activity, indicating that these sequences harbor binding sites with a repressive function (Supplementary Fig. 8). Importantly, these elements do not lower the baseline transcription rate of the reporter gene, and are not transcriptional repressors, but rather reduce the expression to levels comparable to the baseline of the control sequences (SCRAM), but not below it. To avoid confusion with transcriptional repressors we refer to these elements as dampeners, as they dampen the activity of the enhancer. We thus divided the perturbation effects into two main categories (Fig. 2c): 1) activators, identified by perturbations resulting in reduced transcription (WT ¿ PERT); 2) dampeners, identified by perturbations resulting in increased transcription (PERT ¿ WT).

Out of the 598 FRSs, we observed 526 (87.9%) that had activating effects in at least one time point (and non-significant effects in the rest of the time points), and 70 (11.7%) that had dampening effects in at least one time point (and non-significant effects in the rest of the time points) (Fig. 2d, Supplementary Tables 3-4), with only two FRSs alternating between activating and dampening effects at different time points (DMRTA2 motif DMRTA2_M0629_1.02 and Interferon Regulatory Factor 4 motif IRF4_M5573_1.02; Supplementary Tables 2-3). This suggests that the direction of the effect (activating or dampening) of an FRS primarily depends on DNA sequence, and less so on the protein milieu or on other epigenetic properties that change during differentiation. Of note, as lentivirus randomly integrates into the genome, our results consider a cumulative signal from different integration locations in many cells, which essentially controls for the effects of local chromatin properties that may be present around the FRS.

To gain a better understanding of perturbation effects, we further divided our sites into four subcategories (Fig. 2c-d, Methods): 1) Essential: activating sites that when perturbed, reduced the expression level to that of the controls (SCRAM) sequences; 2) Contributing: activating sites that upon perturbation reduce the expression but not to baseline levels; 3) Inhibiting: sites that when perturbed lead to increased activity suggesting that they encompass dampening sites that fine-tune transcription levels; 4) Silencing: dampening sites that block a sequence from regulating transcription, i.e. WT levels are similar to control (SCRAM) and when perturbed make the sequence active. (Fig. 2c; Methods).

Considering this refined division, we found that 159 and 367 out of the 526 activating FRSs, correspond to categories essential and contributing respectively (Fig. 2d, Supplementary Table 2). Out of 72 dampening FRSs, we find 9 silencers and 63 inhibitors (Fig. 2d; Supplementary Table 1). These results represent the distribution of FRSs categories in our dataset. Notably, these FRSs are not a comprehensive list of all functional sites in the selected regions. For instance, we found several regions in which only dampening FRSs were identified (Fig. 2d). Since dampening FRSs only reduce the overall activating function of

the region, dampener-only regions must contain additional unknown activating FRSs that were not included in our design.

We next examined how the strength of the mutation effects caused by perturbing activator sites (WT - PERT) depends on the strength of the expression generated from their respective unperturbed sequence (WT). We found that these effects scale linearly with the WT activity levels (WT - PERT a + b* WT for some constants a, b) across time. While this is trivial for essential FRSs, we found that this linear relationship still holds among contributing activators as well (median R-squared 0.95, methods, Supplementary Fig. 9a-c). When examining fold-change values, this linear relationship translates to: FC = PERT / WT = (1-a) - (b / WT) for the same constants. This relationship saturates and approaches a constant (1-a) for sufficiently high levels of unperturbed (WT) expression (Methods; Supplementary Fig. 9d-e). These constants therefore capture the activation dynamics of each element: a determines the saturated value, and b determines the rate of saturation. We observed that different FRSs within a given region often have different constants, and the same motif has different constants when harbored in different regions, suggesting that the dynamics are not context- or factor-specific, but rather a combination of both. Overall, while the relationship between WT activity and the effect of perturbation is linear, our results show that both depend on the sequence content and the specific cellular context in which it is being assayed.

Characterization of activating and dampening motif effects

Overall, our 598 FRSs include 147 unique motifs. Out of these 147, we observed 68 motifs that are strictly activators, 16 motifs that are strictly dampeners and 63 motifs that show either activating or dampening effects in different genomic contexts (Fig. 2d; Supplementary Figs. 10-11; Supplementary Tables 2-3). When examining the distribution of motif effects across regions (Supplementary Fig. 10; Supplementary Table 3), we observe that related motifs tend to appear in the same regions and importantly - that motifs have different, in many times opposing, effects in different regions. This is also supported by a per motif visualization showing the distribution of categories per motif (Supplementary Fig. 11; Supplementary Table 3). Additionally, there are groups of similar regions that contain the same motifs (Supplementary Fig. 10; Supplementary Table 3). We note that most of the motifs in our dataset ($\tilde{7}5\%$ Supplementary Fig. 10; Supplementary Table 3) appear in five or less regions. Constraining the analysis to motifs that appear in more than five regions shows that 16 out of 35 such motifs ($\tilde{4}5\%$) are strictly activators and all of them have mixed effects depending on the region.

We set out to examine the aforementioned subcategories of specific motifs (Supplementary Fig. 11). Within the activating FRSs, we observed that motifs associated with the SRY-Box Transcription Factor SOX1 are the only motifs that are enriched in the set of essential FRSs (i.e over-representation that is unlikely to occur by chance; hypergeometric test, FDR$<$ 0.05; Supplementary Fig. 11). Both SOX1 and its homolog SOX2 are thought to function as pioneer factors that enable subsequent binding by other TFs [29]. This is in line with our observation that the enhancer activity is completely disrupted when these

motifs are perturbed. Among the motifs that were enriched in the second category of having a contributing binding effect, we observed ZIC factors, which play important roles in neuroectoderm cell development [30].

Among the transcription factors whose motifs are associated with a silencing effect is the Neuronal Differentiation factor NEUROD2. Perturbing a NEUROD2 binding site in a late-response regulatory element (chr15:75409661-75409832 (hg19); Fig 3a) increases the transcription induced by that sequence at the later time points (48-72hr) (Fig. 3a). While NEUROD2 is thought to be a transcriptional activator, our results accord with its previously reported role as a repressor of REELIN gene expression in primary cortical neurons by interacting with CTCF that is known to function as transcriptional repressor in a context dependent manner [31].

Considering the set of Inhibitor motifs, which could fine-tune regulatory activity by partially reducing it, we saw enrichment for the P53-Like Transcription Factor TP73. For example, perturbing a TP73 binding site in region chr6:167854597-167854768 (hg19) substantially increases the activity of that enhancer across all time points. Notably, this region also contains two functional binding sites that activate transcription, and harbor NANOG (NANOG_disc2) and SOX1 (SOX1_M3910_1.02) binding motifs (Fig. 3b). Interestingly, we also found six instances where TP73 binding motifs function as activators (Fig. 3c, Supplementary Table 1). TP73 has been shown to regulate NPC proliferation in the developing and adult mouse central nervous system [32, 33] and is known to interact via its subdomains with many different partner proteins, including POU [34] which has corresponding motifs in this region and YAP1, which is known to function as both an activator or repressor [35] in a context dependent manner [36]. These instances demonstrate that FRSs can achieve their desired transcriptional rate by combining both activating and repressive motif sites, and that using our perturbation MPRA approach allowed us to distinguish the functionality in each specific context. When examining the distribution of sub-categories effects across motifs, we observed 84 (57%) motifs that appear in only one subcategory and 63 (43%) motifs with mixed effects (Fig. 2d; Supplementary Table 1). For most of the motifs the effects are mixed (Supplementary Fig. 11). These results indicate that enhancer activity is influenced both by the motif sequence and the surrounding sequence of the region harboring the motif (Supplementary Figs. 10-11).

Focusing on motifs that are consistently associated primarily with one direction of effect (activating or repressing), we next set out to analyze the effects of motifs on transcription during our time course, by aggregating the results from all their respective instances. We summarized the signals of motifs that show activating or repressing cumulative effect (Fig. 4). Among the TFs associated with activator motifs, we observe the neural markers SOX, LHX, ZIC, and FOX families [29, 30, 37, 38, 20, 39, 40, 41, 42, 43, 44] (Fig. 4a), as well as motifs associated with factors known to be involved in neural induction, such as OTX2 [45, 46] and PAX6 [47, 43]. Consistent with our recent characterization of neural induction associated TFs [20], we also identified Iroquois Homeobox Protein 3 (IRX3) to be one of the strongest activating motifs. Among the TFs associated with repressive activity (Fig. 4b), we observed factors from the HOXD gene family, which are thought to function as repressors

when bound in monomeric form [48]. We also found an enrichment for a SIN3A motif, which is generally known to interact [37] with histone deacetylase (HDAC) and function as a transcriptional co-repressor [49]. It was also reported that the SIN3A/HDAC co-repressor complex was involved in the maintenance of ESC pluripotency [49, 50].

To examine how the effects of motifs change over time, we clustered the signal of all activating and repressing motifs. We observed that the magnitude of effects often changes over time in a manner proportional to the unperturbed expression level (Supplementary Fig. 9). These effects range from perturbations that are effective only at the ESC stage to those that influence late induced regions (Fig. 4). Enrichment analysis of the TFs (both activating and repressing) in the early cluster (Fig. 4a-b) indicated their involvement in processes related to cell differentiation, cell fate commitment and regulation of development for the top 10 categories, whereas enrichment of late response TFs (Fig. 4a-b) indicated, more specifically, categories related to neurogenesis and nervous system development [51]. These results support the functionality of these clusters in earlier and later stages of neural differentiation. For example, enhancers that have OTX2 binding sites reach their peak activity during the neural progenitor cell (NPC) stage. When the OTX2 sites are perturbed, the activity at later time points (48-72hr) was decreased (Figs. 4a,c). Similarly, NPC enhancers harboring IRX2/3 (Fig. 4a) or BARHL1 (Fig. 4d) motifs decreased in activity when the binding sequences were mutated. Correspondingly, we observe that OTX2, IRX2/3 and BARHL1 mRNA levels peak at later time points (48-72hr) (based on data published in [20]). When HOXD sites (HOXD12_M5560_1.02, HOXD9_2) were mutated, the activity at later time points (48-72hr) (Fig. 4b,e; Supplementary Tables 2-3) was increased. These findings indicate that these binding sites have different levels of induced activity at distinct time points of neural differentiation. This suggests that the abundance of the binding TF (i.e. the TF's mRNA levels) at a given time point and the abundance of additional cell-state specific factors (e.g. expression of other TFs) play a significant role in proper enhancer activity.

Interestingly, we also observed TFs whose corresponding motifs show both activating and repressing effects in different regions (Fig. 4a,b; Supplementary Table 3; Supplementary Figs 10-11). For example, different motifs for the Zinc finger protein (ZIC) family have repressing and activating effects across different regions (ZIC2 and ZIC3). Members of the ZIC family are involved in neurogenesis and are known to function as both transcriptional activators and repressors in a context-dependent manner during embryogenesis [52]. Additionally, we observed both effects for the ZEB1 motif in different regions (Fig. 4a,b; Supplementary Table 3) in concordance with the role of ZEB1, acting as both a transcriptional activator and repressor during neurogenesis [53, 54]. We saw similar effects for the RARG motif (Fig. 4a,b; Supplementary Table 3). RARG is a retinoic acid receptor (RAR), a family of factors that plays a role in developmental processes and acts as a ligand-dependent transcriptional regulator. When bound to ligands, RARs activate transcription, whereas in their unbound form they repress transcription of their target genes [55].

Perturbation of motif pairs identifies different modes of motif interaction

We next examined the activity of the assayed regions as composite functional units consisting of multiple FRSs. Our 598 FRSs include 254 unique genomic regions. We observe complexity in these regions in terms of having sites with different direction of effect and different sub-categorization (Supplementary Fig. 10). Specifically, when examining the set of significant perturbation effects in those regions, we observed 141 cases ($\tilde{5}6\%$) with only activating effects (Fig. 2d; Supplementary Table 2, Supplementary Fig. 10), which is consistent with our analysis being focused on regions that were previously identified as enhancers during neural induction [20]. We found 86 regions (30%) that harbor both activating and repressing motif instances. This suggests that regulatory activity within these enhancers can be achieved by fine tuning of binding effects, including both activating and repressing motifs to achieve the desired regulatory function. This phenomenon of context-dependent repression by transcriptional activators is consistent with what was previously reported in yeast [56], drosophila [57] and mammalian cells [58]. Regions with multiple essential FRSs, all required for regulatory activity, offer support to the 'enhanceosome model' of a specific combination of factors being required in an all-or-nothing machinery [17]. In contrast, regions with multiple contributing FRSs are evidence of the 'billboard model', of a flexible modular machinery that fine tunes the induced transcription levels by having independently contributing factors [59]. These results demonstrate that different regulatory sequences may be governed by either the enhanceosome or the billboard model, and some appear to be governed by a combination of both.

We wanted to further examine how pairs of motifs interact in regulatory sequences. To that end, we examined the results of perturbing pairs of motifs, both individually and in combination, to determine how different binding sites interact in a single region (Fig. 5a; Supplementary Table 8). We considered the FRSs to have independent effects (following the billboard model) if the effects were log-additive: perturbing both sites was equivalent to multiplying the effects of perturbing each site separately. We used MPRAnalyze [22] to test this hypothesis for each assayed pair in each perturbation method, by including an interaction term in the model that captures the effect of perturbing both sites while accounting for the effect of perturbing both sites individually (Methods). We considered pairs to have significant interaction if the size of the interaction term was larger than 0.5 and the test was statistically significant (BH-corrected $p< 0.05$). We then defined interaction as "consistent" if the pair were labeled the same (either significant or non-significant) in perturbation methods 3 and either 1 or 2, and removed inconsistent pairs from the analysis. We then also removed pairs in which none of the perturbations are functional, by requiring that at least one of the single perturbations pass the filtering scheme we described above. Finally, to make interpreting the results easier, we excluded pairs in which the assayed sites overlap since overlapping sites cannot be conclusively independent. Overall, out of 149 examined pairs, 24 pairs remained, of which 13 were log-additive, consistent with a billboard model of cooperation, and 11 had significant non-additive interactions (Fig. 5b). While the small number of functional pairs in

our results does not allow for extensive or systemic analyses, we do find anecdotal evidence of different cooperation models operating in different regions.

Among the billboard-consistent pairs we found chr10:100206539-100206710 (hg19), residing in an intron of the HPS1 gene, contains two FRSs each containing a motif instance of ELF1 (ELF1_known3), a transcription factor known for its binding near prefrontal cortex splicing QTL SNPs [60] and for its role in brain development [61]. Both FRSs are activators, but do not have an identical effect: with one driving down transcription to SCRAM levels when perturbed (essential), and the other having a milder effect (contributing). Perturbing both FRSs in this region further reduces the expression to levels significantly below the SCRAM baseline (Fig. 5c). Additionally, we find that additive contribution can also apply to cooperating activators and dampeners, as in chr4:152405951-152406122 (hg19), an intronic region in the FAM106A1 gene body, which contains an FRS with a SOX1 motif that has a contributing effect and an FRS with a ZIC2 motif that has an inhibiting effect. Perturbing both sites results in an additive effect: transcription levels that are lower than WT, but higher than those obtained when perturbing the SOX1 motif alone (Fig. 5d). In the non-additive regions, we found both enhanceosome and composite examples. In the all-or-nothing enhanceosome model, different elements act in a fully-dependent manner. For example, chr8:62736150-62736321 (hg19) contains two essential functional sites: a SOX1 (SOX1_M6129_1.02) and a POU3F1 (POU3F1_2) motif, both necessary for activity. Perturbing either one, and concordingly both, reduces induced transcription to SCRAM levels (Fig. 5e). Both factors are known to have a key role in determining neural fate [62]. In a combination of the billboard and enhanceosome models, some factors are required for any activity while others are independently contributing. For instance, chr11:130016427-130016598 (hg19), downstream of the APLP2 gene which is involved in neural differentiation [63] contains two FRSs: a dampening site with a motif for neural factor MEIS2, and an essential FRS harboring a SOX1 motif. Perturbing both sites results in a reduction of activity to the SCRAM levels, indicating that the SOX1 FRS is required for the overall activity of the region, whereas the dampening MEIS2 FRS is only functional in the presence of a functional activator (Fig. 5f).

Additionally, we found regions that follow neither the billboard or enhanceosome models. In chr16:51185391-51185562 (hg19), upstream of the promoter of neurogenesis regulator SALL1 [64], we find two binding sites of TRIM28. When perturbed individually, one site has no effect, while the other has a mild dampening effect. However, when both are perturbed the effect is a significant decrease in activity. This potentially demonstrates a redundancy mechanism, whereby either binding site is sufficient for the WT activity, and both need to be perturbed in order to disrupt it (Fig. 5g). Overall our results demonstrate the power and potential of perturbation MPRA in uncovering a variety of different patterns of interaction and elucidating the complex regulatory grammar governing these behaviors.

7.4 Discussion

Regulatory elements play a major role in cell-type specific response to environmental conditions and perturbations. Teasing out the regulatory rules and sequences responsible for these responses could lead to a better understanding of how variations in these sequences alter their activity, and allow the accurate design or targeting of specific sequences for therapeutic purposes. Here, we used perturbation MPRA across seven time points of neural differentiation to characterize the regulatory grammar during early stages of neural induction. Our work allowed us to evaluate the effect of intact motif instances over time and annotate these instances into four major categories (essential, contributing, inhibiting or silencing). We observe that generally a FRS either has an activating or repressive effect across all time points, suggesting that the binding motif and surrounding region largely determine the direction of effect, and that the magnitude of this effect changes over time, in a manner proportional to the activity of the WT sequence, in different cellular environments, indicating earlier and later functional motifs in this process. Finally, by carrying out two motif perturbations in a single sequence, we observed different modes of interaction between pairs of motifs.

Several studies have utilized MPRAs to characterize how TF binding may affect regulatory activity. However, these studies examined a small number of TF motifs and assessed their functional effects in a limited number of conditions or cell types. For example, placing TFBSs at different numbers, order, spacing and orientation on 'neutral' background sequences allowed the dissection of regulatory grammar in a human hepatocellular carcinoma cell line [59]. One common finding is that the number of TFBSs (i.e. homotypic clusters of TFBSs [65]) largely determines expression and this relationship follows a non-linear increase with an eventual plateauing of expression [66, 67, 59, 56, 68]. Grossman et al. [19] used both synthetic and endogenous sequences to specifically test the effect of PPARγ binding motifs and show that distinct sets of features govern PPARγ binding vs. enhancer activity. Specifically, they found that PPARγ binding is largely governed by the affinity of the specific motif binding site while the enhancer activity of PPARγ binding sites depends on varying contributions from dozens of TFs in the immediate vicinity, including interactions between combinations of these TFs. Kheradpour et al. [8] examined five predicted TF activators and two predicted repressors and measured effects of their motif disruption in regulatory elements using MPRA. Their findings indicate that disrupting predicted activator motifs abolishes enhancer function, while changes in repressors maintain enhancer activity. They point to evolutionary conservation, nucleosome exclusion, binding of other factors, and motif affinity, as being predictive features of enhancer activity.

Here, we analyzed the effect of over 250 motifs with three different perturbations using two approaches. In the first approach, we replaced the motif with two different 'non-motif' sequences and in the second approach, we scrambled the motif's nucleotides. All these perturbations showed high reproducibility between replicates (r¿0.95). Analyzing and comparing the three perturbation methods, we observed a similar level of overlap between the different methods, but we do not observe more consistency between perturbation methods

1 and 2 than either one is with perturbation method 3 (Fig. 2b,2d; Supplementary Fig. 6; Supplementary Tables 1;4-7). This may indicate that at least one of the fixed-sequence perturbation methods potentially introduces bias that separates it from the other, e.g by forming de novo binding sites with endogenous sequences adjacent to the perturbed sites. Since methods 1 and 2 insert a fixed sequence, this introduced bias could be systemic across the assayed regions and skew downstream results. For future experimental designs, we suggest using a more robust perturbation approach that randomly shuffles the nucleotides of the perturbed site and is less likely to introduce systemic biases.

We cataloged the function of 598 FRSs representing 254 unique endogenous regions and 147 unique motifs. Approximately 90% of FRSs act as activators with 30% of them as essential and the rest as contributors. This finding is also in line with a saturation based MPRA that analyzed ten disease-associated promoters and enhancers, finding that the majority of mutations lead to a reduction in activity (i.e. act as activators that when mutated reduce activity) [69]. Additionally, while our data does not contain FRSs that repress transcription below the baseline rate, we found many instances of binding sites that have a repressive effect on the function of the enhancer itself: reducing the level of induced transcription, or even completely blocking the enhancer's activity. These instances suggest that enhancers can be kept in a pseudo-poised state: residing in open chromatin but being blocked from activity by TF binding, and that repressive factors are often bound to functional enhancers as a mechanism for fine-tuning transcription levels.

Finally, a smaller subset of sequences were perturbed in two locations, where we perturbed two single motifs separately and jointly to assess their interaction, as a proof of concept (Fig. 5). To model these interactions, we used the billboard model of independent contribution as our null hypothesis, by assuming that the effect of each individual contribution is log-additive [19]. We tested this hypothesis using MPRAnalyze [22] for each pair in each perturbation method (Methods). Only pairs which showed consistency (in perturbation methods 3 and either 1 or 2) in the significance of their interaction term (determined by magnitude and p-value; Methods), where the single motifs were not overlapping, and at least one of the single perturbation is a FRS, were considered further in our analysis. Overall, out of 149 examined pairs, 24 pairs remained, of which 13 were log-additive (Fig. 5b-d), consistent with a billboard model of cooperation, and 11 had significant non-additive interactions (Fig. 5b). In the latter category, we observed different TF cooperation models, including the 'enhanceosome model' in which a strict composition of TFs are required for an enhancer's function (Fig. 5e), a hybrid of billboard and enhanceosome models (Fig. 5f) in the same region and instances that do not fall under any of these categories (Fig. 5g). Notably, for FRSs containing two instances of the same motif, the single perturbations did not have identical effects, consistent with the growing body of work showing that the function of an enhancer depends on the specific locations and distances between binding sites, and not only of their presence [66, 67, 59, 56, 68]. Albeit being underpowered in the number of functional pairs does not allow for systematic conclusions, our anecdotal examples demonstrate the complexity of different TF cooperation models.

Examining whether we can gain a better understanding on the determinants of time-point

specific regulatory activity using this model system, revealed complex results, suggesting that motif sequence alone is less likely to determine temporality without the context of the surrounding region and other bound factors (Supplementary Note 1, Supplementary Fig. 12, Supplementary Table 9). Therefore, future challenges following our work will include developing strategies to further understand regulatory logic and its determinants across different conditions. For example, using endogenous manipulations via CRISPR to examine the function of specific motifs and their combinations across different cellular conditions.

To address whether temporal activity of the functional regulatory sites (FRSs) are consistent with TF temporal binding using the following 3 strategies: First, we used RNA-seq data from [20, 70] to compare the timing of motif importance with the respective TF expression. Testing this correlation did not show conclusive results. We speculate that this is due to the nature of our analysis which is motif-based, and since similar sequence motifs are not independent, it is likely that the annotation of the FRSs suffers from misclassification of the binding factor. Additionally, even if the exact factor was known, it is not established in current literature that the magnitude of TF gene expression is directly correlated with its regulatory effect, so a strong correlation is not necessarily expected. Second, we examined the overlap of ChIP-seq peaks of different TFs in hESC-derived neuroectoderm [71] with regions where SOX1 motifs were perturbed, for sufficient statistical power. We observe significant overlap (fisher exact test FDR$<$ 0.05) of ChIP-seq peaks of OTX2 and SOX2 factors for FRSs compared to regions that were filtered out using the 4 filters described previously. This indicates that the signal we are observing using perturbation MPRA is consistent with endogenous binding of the key transcription factors that play pivotal roles in ES-to-neural differentiation [40, 70]. Finally, we utilized the data we collected in our previous work [20, 70] of RNA-seq following overexpression of theseTFs: BARHL1, IRX3, LHX5, OTX1/2, PAX6. For the FRSs that contain motifs of these factors, we observe that 85% of their closest genes are differentially expressed (compared to hESC; FDR$<$ 0.05). This serves as an additional support of the endogenous functionality of motifs of these factors in these regions. Comparing the number of differentially expressed genes that are closest to the FRS to the distribution of the total number of differentially expressed genes, for each overexpressed TF, yielded a statistically significant result for PAX6 (Fisher exact test p-value$<$ 0.02). However a larger number of tested FRSs will be needed to make more rigorous conclusions.

During early neural induction, pluripotency-associated genes are rapidly downregulated and neural associated genes are induced by a variety of factors [27, 43]. As such, the rapid differentiation of hESCs into neural cells provides an exceptional model to study motif effects and how they change across developmental time points. Using this model, we previously interrogated [20] the temporal dynamics of gene expression (RNA-seq) and gene regulation (ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and LentiMPRA) at seven time points during early neural differentiation. Our current work further validated the novel motifs and TFs identified in our previous report to have temporal effects across neural induction. For example, we find that FRSs harboring BARHL1 and IRX3 motifs exhibit time point specific activating effects and show changes in magnitude over time, with higher signal at the NPC state - supporting their suggested role in neural induction (Fig. 4).

Overall, our results provide an atlas of motif function across early time points of neural differentiation by directly testing hundreds of regulatory regions for the function of the motifs they harbor. To the best of our knowledge this provides the first comprehensive perturbation MPRA study across a developmental time course, showing clear changes in regulatory activity over time. This system provides a model for how perturbation MPRA can be leveraged to identify and characterize in a high throughput manner the functional effects of regulatory sequences across different cellular conditions/perturbations.

7.5 Methods

Computational Analyses

Choosing region and motif combinations

General Description Our previous analysis [20] points to a large number of regulatory regions of interest as well as multiple motif hits within those regions. Our goal is to select the most informative set of [region x motif] combinations (each corresponding to a motif instance) so as to fit within a single MPRA design. To address this, we developed a selection scheme to represent various biological aspects of our system and account for experimental limitations for the number of assayed sequences.

To do this, we formalize the information that we have about the motifs and regions as a tripartite graph, with one layer of nodes corresponding to DNA regions, another layer of nodes that represent motifs and a third layer of nodes, each representing a different property of motifs or regions (Supplementary Fig. 13). The region layer consists of the 1,547 genomic regions we identified in our previous work [20] that show temporal activity when tested using lentiMPRA in the same seven time points. The motif layer consists of motif hits found in those regions computationally (using Fimo [23] (p-value$< 10^{-5}$) with two sets of TF motifs [24, 25]). Edges between the first two layers connect every motif with the regions in which it occurs. Each node in the third layer corresponds to a property of interest which characterizes a subset of the motifs and regions that are represented in the first two layers. These properties are based on genomics assays from our previous work [20] (based on ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and RNA-seq data from these 7 time points). For instance, we identified several temporal patterns associated with each data modality and designated each of these patterns as a node (e.g., a node for "regions that have a transient peak in H3K27ac 48 hour post induction"). We then connect a region to a node if that respective pattern is observed in that region in the endogenous genome. Similarly we connect a motif node to a property node. For example, a node for "motifs with associated TF that is expressed 24 hour post induction"). We then connect a motif to a node if that respective pattern is observed for that motif in the endogenous genome. We describe the "property layer" and its edges with the "motif" and "region" layers in greater detail below.

Altogether our graph now has 1547 region nodes, 4393 motif nodes and 68 property nodes. These nodes are connected by a total of 99165 edges. Our goal now becomes to find

the minimum number of [region x motif] combinations (each representing a specific motif instance, or - equivalently- an edge in our graph) that will guarantee a sufficient coverage of each property. In other words, we want to select a minimal number of motif-region pairs such that every "property node" in our third layer is connected by an edge to a sufficient number of motifs and regions (as detailed below). Having staged our data in a tripartite graph helps us re-state our goal as a constrained optimization problem - guaranteeing minimal level of connectivity for the third layer, while minimizing the number of selected nodes and edges in the first two layers. Since this problem is NP- hard, we followed the common practice and formulated it as an integer linear program (ILP), which can be solved efficiently through a range of heuristic with available solvers. With this ILP, we were able to select 591 regulatory regions and 255 motifs that are organized into 2,144 region-motif pairs. Below, we provide a more in-depth description of this process.

Defining the property layer We composed a list of biological properties based on published literature and on ATAC-seq, H3K27ac and H3K27me3 ChIP-seq and RNA-seq data we produced and analyzed in our previous paper [20]. The biological properties of TFs associated with motifs and regions include: (i)TF/region is induced/active at a specific time point. (ii) TF/region binds/belongs to significantly overlapping sub-clusters (as defined in Inoue et al.[20]) of temporal MPRA and H3K27ac/ATAC-seq/RNA-seq signals (iii) The TF/the proximal gene for the region is a known neural factor or belongs to one of the pathways defined below. Known neural factors: POU3F1, MYT1L, SOX2, POU3F2, LHX2, PAX6, ASCL1, SOX1, OTX2, ZNF521, NEUROG1, NEUROG2, NEUROG3, NEUROD1, NEUROD2. Pathways taken from KEGG [72]: FGF/MAPK signaling pathway hsa04010, IGF-1/mTOR signaling pathway hsa04150, Wnt/Ca+/PCP signaling pathway hsa04310, Sonic Hedgehog signaling pathway hsa04340. (iv) Hand picked TFs (POU3F1, POU3F2, SOX2, SOX1, PAX6, OTX2, LHX2, NEUROG1, NEUROG2, NEUROD2, SP8, IRX3, SOX10, PKNOX2, HHEX, LMX1A, BARHL1, LHX5, NR2F2, DMBX1, MEIS2, OTX1, SOX21, FOXB1, SOX5, MEIS3, HOMEZ, TCF3, TCF4, ZIC1, ZIC2, ZIC3, ZIC4, ZIC5), including factors known to have a role in neural differentiation based on previous literature [38, 20, 39, 40, 41, 42, 43, 44], or based on their expression in neuroectoderm in mouse embryo, or show high 'TF activity score' in the relevant time points in our data [20]. The direct edges from motifs and regions to properties, represent the biological properties a region or a motif satisfies as described above.

The optimization program minimize $\left(\sum_{r\in R}\theta_r\right)+3\cdot\left(\sum_{(t,r)\in E;t\in T;r\in R}e_{t,r}\right)$

1. $\sum_{(t,p')\in E;t\in T}\theta_t\geq 12 \qquad \forall p'\in P$
2. $\sum_{(r,p')\in E;r\in R}\theta_r\geq \min\{17,\deg_R(p')\} \qquad \forall p'\in P$
3. $\sum_{(t,r')\in E;t\in T}\theta_t\geq \theta_{r'}\min\{3,\deg_T(r')\} \qquad \forall r'\in R$
4. $\sum_{(t',r)\in E;r\in R}\theta_r\geq \theta_{t'}\min\{20,deg_R(t')\} \qquad \forall t'\in T$

5. $e_{t,r} \geq \theta_t + \theta_r - 1 \qquad \forall (t,r) \in E; t \in T; r \in R$
6. $\sum_{t \in T_i} \theta_t \leq 2 \qquad \forall T_i$
7. $\sum_{t \in T_i} \theta_t \geq 1 \qquad \forall T_i \in$ Hand Picked
8. $\sum_{r \in R} \theta_R \geq 0.4 \cdot |R|$
9. $\sum_{(t,r) \in E; t \in T; r \in R} e_{t,r} \geq 5 \cdot \sum (t,r) \in E_p; t \in T; r \in R e_{t,r}$
10. $\sum_{t \in T_B} \theta_t \geq 1.5 \cdot \sum_{t \in T_S} \theta_t$
11. $\theta_t, \theta_r, e_{t,r} \in 0, 1$

the decision variables represent the following: θ_t is a binary variable that indicates whether we chose the motif t; $\theta_r$is a binary variable that represents whether the region r was selected. $e_{t,r}$ is a binary variable that denotes whether a motif x region pair (t and r) has been selected. Parameters include: P: properties; R: regions; T: motifs; $\deg_R(p)$: the number of edges connecting property p to regions; $\deg_R(t)$: the number of edges connecting motif t to regions; $\deg_T(r)$: the number of edges connecting region r to motifs; T_i is a subset of T that contains all the motifs corresponding to TF i; E_p as a subset of the edges with lower confidence (i.e edges that connect to properties representing non significantly overlapping sub-clusters of temporal MPRA and H3K27ac/ATAC-seq/RNA-seq signals); We define T_B as the subset of motifs connected to at least 5 regions, and T_S as the subset of motifs connected to fewer than 5 regions.

The constraints described in the equations above ensure that: (1) Each property is connected to at least 12 motifs. (2) Each property is connected to at least 17 regions (or all regions if it's below 17). (3) Each region is connected to at least 3 motifs. (4) Each motif is connected to at least 20 regions. (5) An edge is active if both nodes of the edge are active. (6) For each TF, no more than two motifs are chosen. (7) All hand picked TFs are used at least once. (8) At least 40% of all regions are used. (9) At most 1/6 of the total edges used are low confidence edges. (10) At least 60% of motifs chosen are motifs connected with many regions (TB), s.t. the solver does not bias towards lowly connected motifs. (11) All variables are binary.

For each $T_i \in$ Hand Picked, one representative motif must be in the solution.

Our objective is to minimize the overall number of MPRA sequences to design. It is a sum that accounts for corresponding to the number of unperturbed (WT) regions plus the number of perturbations (i,e, regions and motif combinations). We multiply by 3 since we have 3 perturbation methods (i.e., we need 3 MPRA sequences for every pair).

Different categories of sequences designed on the array

Overall, the solver picked 591 regions, 255 unique motifs which correspond to 166 unique TFs. We used the combinations of region and motifs chosen by the solver to represent the following sequence categories on the array (Supplementary file 1):

1. One motif is perturbed in the sequence. For combinations of regions and motifs where the motif is detected once in the sequence (hit1; N=1620).

2. Two motifs of the same motif are perturbed in the sequence. For combinations of regions and motif s where the motif is detected twice in the sequence: if the +/- strand carry exactly the same motif we only replace the motif one time in the + strand (hit2; N=62), otherwise (hit2diff; N=90) we perturbed each motif separately and then both of them - starting with the + strand. If 3 or more hits of the same motif are observed – we discard those region-motif combinations (N=52).

Additionally to the combinations picked by the solver, we considered the 591 WT regions and added more combinations (not chosen by the solver) that contain motifs of the following 11 TFs. These TFs were chosen (LHX5, MEIS2, PAX6, FOXB1, SOX1, IRX3, OTX2, ZIC2, SP8, POU3F1, HOMEZ) based on their high 'TF activity score' in the relevant time points in our data [20] and their mRNA expression in neuroectoderm in mouse embryo.

3 One motif is perturbed in the sequence. For combinations of region and motif where the motif is detected once in the sequence (Overexpressed_hit1 N=221 and Overexpressed_permutation N=58).

4 Two motifs of the same motif are perturbed in the sequence. For combinations of regions and motifs where the motif is detected twice in the sequence: if the +/- strand carry exactly the same motif we only replace the motif one time in the + strand (Overexpressed_hit2 N=3), otherwise (Overexpressed_hit2diff N=1) we perturbed each motif separately and then both of them - starting with the + strand.

5 Combinations of two or more motifs are perturbed in the sequence. For combinations of regions and motifs where we observe two or more different motifs in the sequence (Overexpressed_permutation N=125). We examined combinations of motif hits of these 11 TFs in our regions.

Overall, most of the data includes a single motif perturbation per region (N=2,144) and a smaller part with two or more motif perturbations per region (N=216 out of those: N=154 two motifs; N=62 more than two motifs) comprising a total of 2,360 designed region and motif sequences. We also assayed WT and control sequences:

1. We assayed 591 WT sequences. WT sequences are the endogenous 171bp sequences.

2. We assayed 591 scrambled sequences (SCRAM). Scrambled sequences are based on WT sequences with shuffled nucleotides, creating a set of negative controls.

3. We assayed 591 sequences with random alterations (RAND) – where we randomly chose a location in the region and perturbed the median motif size (12bp) starting in that location, creating an additional set of negative controls.

We perturbed predicted motifs within each genomic region (2,360 combinations) using three perturbation approaches: the first two replace the predicted binding site with a "non-motif" sequence whereas the third one shuffles the nucleotides of the predicted binding site described in the next section. For the RAND sequence category, we used the same 3 perturbation approaches.

Different motif scrambling (perturbation) approaches

Approach 1 - create "non-motif" sequences following these steps:

1. Use all the 2,464 MPRA sequences we designed in our previous work [20] based on their potential to be active during neural differentiation.
2. Count #di-nucleotides and calculate their percentage of appearance in those sequences.
3. Create a di-nucleotide scrambled sequence in the length of the maximal motif, i.e. "scrambled motif".
4. Create 1,000 maximal length "scrambled motifs".
5. Run these 1,000 "scrambled motifs" through Fimo [23] with the two sets of TF motifs [24, 25] and choose the ones with the lowest number of motif hits (p-val$< 10^{-4}$) - 13 "scrambled motifs" had 0 hits.
6. In each chosen combination of region and motif (described in the previous section) - replace the motif appearance with the prefix of the "scrambled motif" (adjusting to each motif length) using these two strategies that avoid motifs creation in the edges of the sequences: (1) use 3bp downstream and upstream of the motif in the original sequence (2) use the original sequence. Repeat this 13 times using each one of the "scrambled motifs".
7. Run the sequences created using the two strategies: (1) 3bp "scrambled motif prefix" 3bp; (2) original_sequence_start "scrambled motif prefix" original_sequence_end, through Fimo [23] with the two sets of TF motifs [24, 25] (p-val $< 10^{-4}$).
8. Choose the 2 "scrambled motifs" that result in the lowest number of motif hits indicated by the median rank across the two strategies, i.e. "non-motif sequences" that would be used on the array.

Approach 2 - shuffle the motif: in each chosen combination of region and motif - scramble the motif by shuffling its nucleotides.

Library processing: replicates, association, barcode count, ratio

Association Reads from the association library were aligned to the reference set of sequences using bowtie2 [73] with the –very-sensitive preset parameters for maximal accuracy. A barcode was confidently assigned to a sequence if at least 3 unique UMIs supported that assignment and at least 80% of the UMIs associated with that barcode were aligned to the sequence. Barcodes that were not confidently assigned were considered ambiguous and discarded from downstream analyses. Overall, 7,004,354 barcodes were observed, of which 1,447,874 (20%) were confidently assigned, averaging 139.2 barcodes per sequence Supplementary Figs. 1-2). To make sure that our results are robust to the association thresholds, we repeated our analysis with a 99% threshold for confident association, which resulted in highly consistent activity estimates (Pearson's correlation 0.97).

MPRA Barcode Counting Reads from the MPRA libraries were processed against the set of confidently assigned barcodes, requiring a perfect match. Of the barcodes observed in the MPRA libraries, an average of 61.6% were confidently assigned, 37.4% were ambiguous (observed in the association library but were not confidently assigned), and 0.9% were unobserved in the association library (Supplementary Fig. 2). Only barcodes that appeared in at least two corresponding libraries (DNA and RNA libraries from the same time point and replicate) were included in downstream analyses, resulting in an average of 134.4 barcodes per sequence.

Quantification of Induced Transcription Rate with MPRAnalyze Quantification of induced transcriptional rates ('alpha' values) was performed using MPRAnalyze [22]. Briefly, MPRAnalyze fits two nested generalized linear models (GLMs): the first estimates the latent construct counts from the observed DNA counts, and the second estimates the latent rate of transcription from the latent construct estimates and observed RNA counts. The models are optimized using likelihood maximization, with a gamma likelihood for the DNA counts and a negative binomial likelihood for the RNA counts. MPRAnalyze includes library-size normalization factors, which were computed once using the entire dataset and then used across all analyses, including per-time-point analyses, to maintain consistency. For quantification of alpha values, the full experimental design was included in the design matrix for the DNA model ($\sim$ timepoint + replicate + barcode), and an alpha value was extracted for each time point and replicate (RNA model: $\sim$ timepoint + replicate).

Classification of active sequences with MPRAnalyze Classification of active sequences was performed using the standard MPRAnalyze classification analysis, in which alpha values are mad-normalized (a median-based variant of z-normalization) and tested each value against the null distribution, estimated from the alpha values from the negative control scrambled sequences.

Comparative Analyses with MPRAnalyze The GLM structure of MPRAnalyze allows for a flexible framework to perform comparative analyses by using various design matrices for the different analyses (detailed below). Since the models are optimized using likelihood maximization, a likelihood ratio testing can be used for statistical significance, and was used throughout all analyses in the manuscript. P-values were computed for each comparison and corrected within each analysis using Benjamini-Hochberg FDR correction [74].

For the per time point comparative analyses, each PERT and RAND sequence was compared with the corresponding WT sequence within each time point (DNA design: $\sim$ replicate + barcode + sequence; Full RNA design: $\sim$ sequence; reduced RNA design: ~ 1). The resulting p-values were corrected jointly across all timepoints.

For temporal analyses, aimed at determining which sequences had temporal activity, we set the null behavior to be the temporal behavior exhibited by the scrambled sequences, by fitting a joint model to all SCRAM sequences and using the model coefficients as normalization factors for the comparative models (DNA design: $\sim$ timepoint + replicate + barcode; Full RNA design: $\sim$ timepoint; reduced RNA design: ~ 1).

For the comparative temporal analyses, we compared the temporal activity of each PERT or RAND sequence with the corresponding WT sequence, using an interaction term in the design (DNA design: $\sim$ timepoint + replicate + barcode; full RNA design: $\sim$ time * sequence; reduced RNA design: $\sim$ time). Note that the barcode covariate in the allele-comparative analyses (per-time point comparative analysis and temporal comparative analysis) is sequence-specific, so the barcode factor is confounded by the sequence variable.

Interaction analyses for multiple-perturbations using MPRAnalyze The distribution of the joint perturbation design is as follows: for the same PWM joint perturbations (91): we have 19 that appear in one region, 6 in two regions, 4 in three regions, 2 in four regions, 2 in five regions, 2 in six regions, 1 in seven regions and 1 in eleven regions. For the different PWM joint perturbations (63): 5 appear in two regions, 1 in three regions, 2 in four regions, 1 in eight regions and 1 in nine regions. All the rest appear in one region.

We used MPRAnalyze to characterize the interactions between pairs of motifs by testing the hypothesis that corresponds to the billboard model of independent contribution, by assuming that the effects of each individual contribution is log-additive. We therefore included two binary covariates in the model: Pert1 indicated whether the observation comes from a sequence that contained the first Perturbation, Pert2 indicated whether it contained the second perturbation. In the full model we then included an interaction term between these two covariates (y $\sim$ time + Pert1 * Pert2), which we excluded from the reduced model (y $\sim$ time + Pert1 + Pert2), so the effects will be independent. We then used a Likelihood Ratio Test to determine statistical significance, and the interaction coefficient was used as the interaction effect size.

Calculating RNA/DNA ratios The calculation of RNA to DNA ratio is explained in detail in our previous work [20, 59]. Briefly, to estimate the abundance of DNA or RNA

per sequence and for each replicate (in order to compare replicates and time point), we use a simple averaging scheme: D(R)NA per sequence $= \frac{10^6 \cdot \sum_{i=1}^{\#\text{BC}} \text{D(R)NA}_i}{\#\text{BC} \cdot \text{sum(D(R)NA reads}}$ where $D(R)NA_i$ denotes the reads of a specific barcode i among the #BC barcodes that belong to the respective sequence.

To determine the RNA/DNA ratios per sequence and for each replicate we the sum of ratios: $\frac{1}{\#\text{BC}} \cdot \sum_{i=1}^{\#\text{BC}} \left(\frac{RNA_i}{\text{sum RNA reads}} / \frac{DNA_i}{\text{sum DNA reads}} \right)$ We added a pseudo count of 1 to the numerator and denominator to stabilize the signal from sequences with low numbers of reads. To combine replicates, we first divided the RNA/DNA ratios observed in each sample (time point/replicate) by the median ratio and then obtained the final RNA/DNA ratio by averaging the normalized values across replicates. We use the ratio calculation to compare the MPRA signal in this work to our previous work [20] (Supplementary Fig 4.).

Filtering Sequences

We use MPRAnalyze to determine differential activity (explained in the previous section), for each perturbation method and each time point, comparing the following: (PERT,WT), (RAND,WT), (PERT,RAND), (WT,SCRAM) and (PERT,SCRAM). We use the following filters:

Filtering sequences per time point

1. We consider only sequences where WT (at each of the 7 time points) or PERT have significantly different (MAD-score) regulatory activity than the null (SCRAM) (filter 3): length(FDR(WT, SCRAM) < 0.05)==nof_TPs it FDR(PERT, SCRAM) < 0.05

2. We only consider sequences where PERT has significantly different regulatory activity than its matching WT (filter 1): FDR(PERT, WT)< 0.05

1008, 1042, 998 out of (2082, 2086, 2114) sequences for perturbation methods 1, 2 and 3 respectively, pass these filters in at least one time point.

Filtering sequences across time

3 We consider only sequences where WT or PERT have significantly temporally different regulatory activity than the null (SCRAM) (filter 4). FDR(temporal(PERT,SCRAM))$<$ 0.05 *or* FDR(temporal(WT,SCRAM))< 0.05)

4 PERT has significantly temporally different regulatory activity than its matching WT (filter 2) FDR(temporal(PERT,WT))< 0.05

1189, 1224, 1354 out of (2082, 2086, 2114) sequences pass the temporal filtering for perturbation methods 1, 2 and 3 respectively.

We consider only sequences that are significant (pass all filtering steps per time point) in at least one time point and follow the temporal constraints, after filtering for duplicates,

resulting in overall 747, 775, 749 sequences for perturbation methods 1, 2 and 3 respectively. Duplicates, i.e. sequences with motifs perturbed in the exact same locations (corresponding to different PWMs), were filtered, by picking the sequence with the lowest temporal FDR. FRSs are defined as sequences that pass all 4 filters and belong to the same main category (as described in the next section) in perturbation method 3 and either perturbation methods 1 or 2.

Motif effect - main and sub categories

Activators: when this motif is perturbed in a region, the regulatory activity of PERT compared to WT is significantly reduced in at least one time point.

i **Essential:** this motif is essential for the regulatory activity of the region - i.e. scrambling this motif reduces the regulatory activity to null (SCRAM) *or* for all time points - the regulatory activity of PERT is similar to SCRAM. FDR(temporal(PERT,SCRAM))> 0.05 *or* length(FDR_MAD(PERT,SCRAM)> 0.05)) == nof_TP

ii **Contributing:** this motif is contributing to the regulatory activity of the region - i.e. if we scramble this motif, the region is still regulatory active and its activity is different from null (SCRAM). If a motif is not essential, it is deemed contributing.

Repressors: when this motif is perturbed in a region, the regulatory activity of PERT compared to WT is significantly increased in at least one time point.

iii **Silencing:** this motif has a silencing effect on the regulatory activity of the region - i.e. the regulatory activity of the WT region is not temporarily different from SCRAM or for all time points - the regulatory activity of WT is similar to SCRAM.

FDR(temporal(WT,SCRAM))> 0.05

or

length(FDR_MAD(WT,SCRAM) > FDR_thresh)==nof_TP

iv **Inhibiting:** this motif is reducing the regulatory activity of the region. If a motif is not silencing, it is deemed inhibiting.

Activation dynamics analysis

To examine the activation dynamics of activating FRSs, we looks at activators that are active in all 7 time points and fit a linear regression line to each FRS, modeling the absolute effect (WT - PERT) as a function of the WT activity level (delta $\sim$ wt), using the lm function in R. the model parameters were then extracted and used for the extrapolation in Supplementary Fig. 9e).

Experimental procedures

LentiMPRA library cloning and sequence-barcode association

The lentiMPRA library construction was performed as previously described [28]. In brief, array-synthesized oligo pool was amplified by 5-cycle PCR using forward primer (5BC-AG-f01, Supplementary Table 10) and reverse primer (5BC-AG-r01, Supplementary Table 10) that adds mP and spacer sequences downstream of the sequence. The amplified fragments were purified with 1.8x AMPure XP (Beckman coulter), and proceeded to second round 11-cycle PCR using the same forward primer (5BC-AG-f01) and reverse primer (5BC-AG-r02, Table Supplementary Table 10) to add 15-nt random sequence that serves as a barcode. The amplified fragments were then inserted into SbfI/AgeI site of the pLS-SceI vector (Addgene, 137725) using NEBuilder HiFi DNA Assembly mix (NEB), followed by transformation into 10beta competent cells (NEB, C3020) using the Gemini X2 machine (BTX). We note that there is not a typical polyA signal downstream of the WPRE in our lentiviral vector, as it was reported that an internal polyA signal can decrease virus titer [75]. Colonies were allowed to grow up overnight on Carbenicillin plates and midiprepped (Qiagen, 12945). We collected approximately 1 million colonies, so that on average 100 barcodes were associated with each sequence. To determine the sequences of the random barcodes and their association to each sequence, the sequence-mP-barcodes fragment was amplified from the plasmid library using primers that contain flowcell adapters (P7-pLSmP-ass-gfp and P5-pLSmP-ass-i#, Supplementary Table 10). The fragment was then sequenced with a NextSeq 150PE kit using custom primers (R1, pLSmP-ass-seq-R1; R2 (index read), pLSmP-ass-seq-ind1; R3, pLSmP-ass-seq-R2, Supplementary Table 10) to obtain approximately 50M total reads.

Lentiviral infection and barcode sequencing

Lentivirus was produced in twelve 15cm dishes of 293T cells using Lenti-Pac HIV expression packaging kit following the manufacture's protocol (GeneCopoeia, LT002). Lentivirus was filtered through a 0.45um PES filter system (Thermo Scientific, 165-0045) and concentrated by Lenti-X concentrator (Takara Bio, 631232). Titration of the lentiMPRA library was conducted on hESCs as described previously [28]. Lentiviral infection, DNA/RNA extraction, and barcodes sequencing were all performed as previously described [20]. Briefly, approximately 8 million cells (three 10 cm dishes) per time point were infected with the lentivirus library with a multiplicity of infection (MOI) of 5-8 along with 8 µg/mL polybrene (Sigma). Three independent replicate cultures were infected. To normalize technical bias of lentivirus preps, two of these replicates were infected with the same lentivirus batch, while the other replicate was infected with another lentivirus batch. The cells were incubated for 3 days with a daily change of the media. The infected cells were induced into neural lineage using dual-Smad inhibition and harvested at 0 (right before differentiation), 3, 6, 12, 24, 48, and 72 hours. DNA and RNA were purified using an AllPrep DNA/RNA mini kit (Qiagen). RNA was treated with Turbo DNase (Thermo Fisher Scientific) to remove contaminating DNA, and reverse-transcribed with SuperScript II (Invitrogen, 18064022) us-

ing barcodes specific primer (P7-pLSmp-assUMI-gfp, Supplementary Table 10), which has a unique molecular identifier (UMI). Barcode DNA/cDNA from each replicate of each time point were amplified with 3-cycle PCR using specific primers (P7-pLSmp-assUMI-gfp and P5-pLSmP-5bc-i#, Supplementary Table 10) to add sample index and UMI. A second round of PCR was performed for 19 cycles using P5 and P7 primers (P5, P7, Supplementary Table 10). The fragments were purified and further sequenced with NextSeq 15PE with 10-cycle dual index reads, using custom primers (R1, pLSmP-ass-seq-ind1; R2 (index read1 for UMI), pLSmP-UMI-seq; R3, pLSmP-bc-seq; R4 (index read2 for sample index), pLSmP-5bc-seqR2, Supplementary Table 10).

7.6 Figures

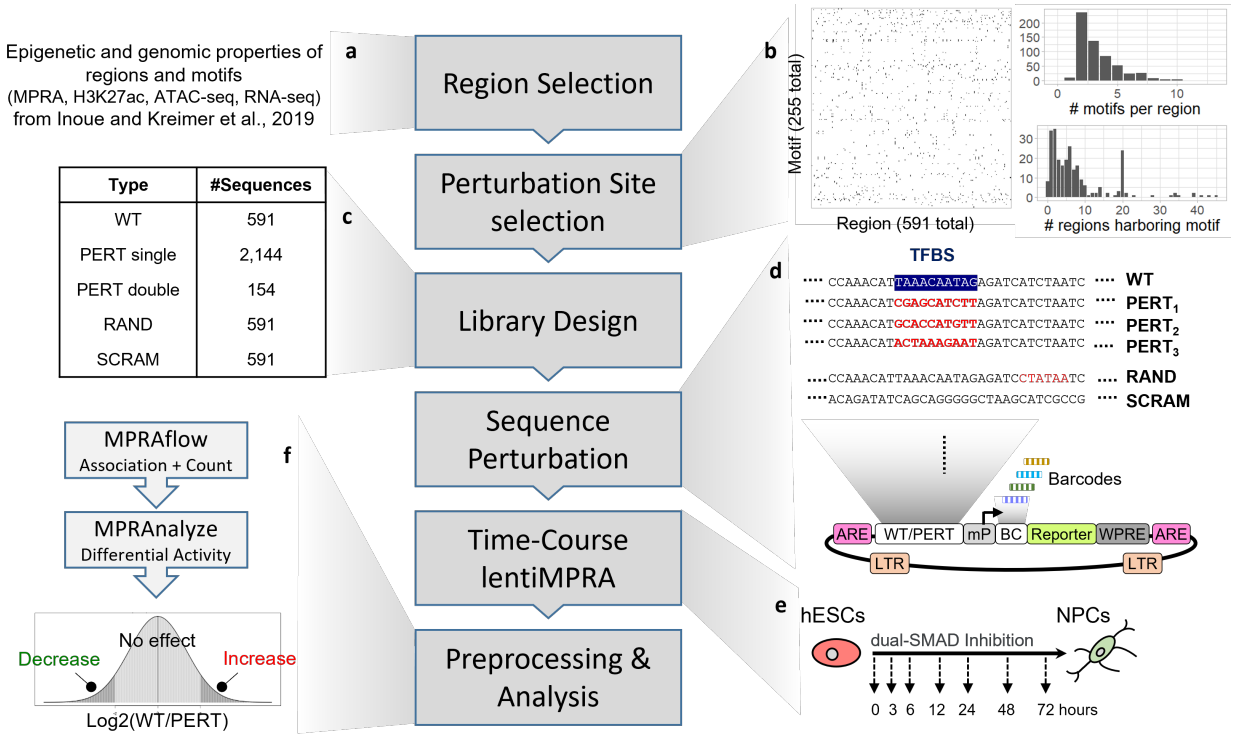


Figure 1: **Experimental design.** **a**, Computational framework to select regions and perturbation sites. **b**, hitmap of motif instances in the assayed regions (left); distribution of the number of motifs perturbed in each region (top right); distribution of the number of regions harboring each motif (bottom right). **c**, Library design; selected regions were included in their wild type (WT) form, selected motifs were perturbed (by altering the sequence in the predicted motif site) using three perturbation methods individually (PERT single) as well as in combination with other perturbations in selected cases (PERT double). Random sites were perturbed (RAND) and the entire WT sequence was scrambled as negative controls (SCRAM) for each WT sequence. **d**, The designed sequences were synthesized and cloned into the lentiMPRA vector and associated with 15-bp barcodes. Abbreviations: ARE, antirepressor element. BC, barcode. Reporter, EGFP, enhanced green fluorescent protein. LTR, long terminal repeat. mP, minimal promoter. WPRE, Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element. **e**, lentiMPRA libraries were infected into hESCs and following three days, we induced neural differentiation via dual-SMAD inhibition and obtained DNA and RNA at seven time points (0, 3, 6, 12, 24, 48, and 72 hours). **f**, Association between barcodes and designed sequences, and the number of barcodes observed in DNA and RNA sequencing was determined using MPRAflow [28]. Differential analysis between WT and PERT activity to determine motif regulatory effect over time was assessed using MPRAnalyze [22].

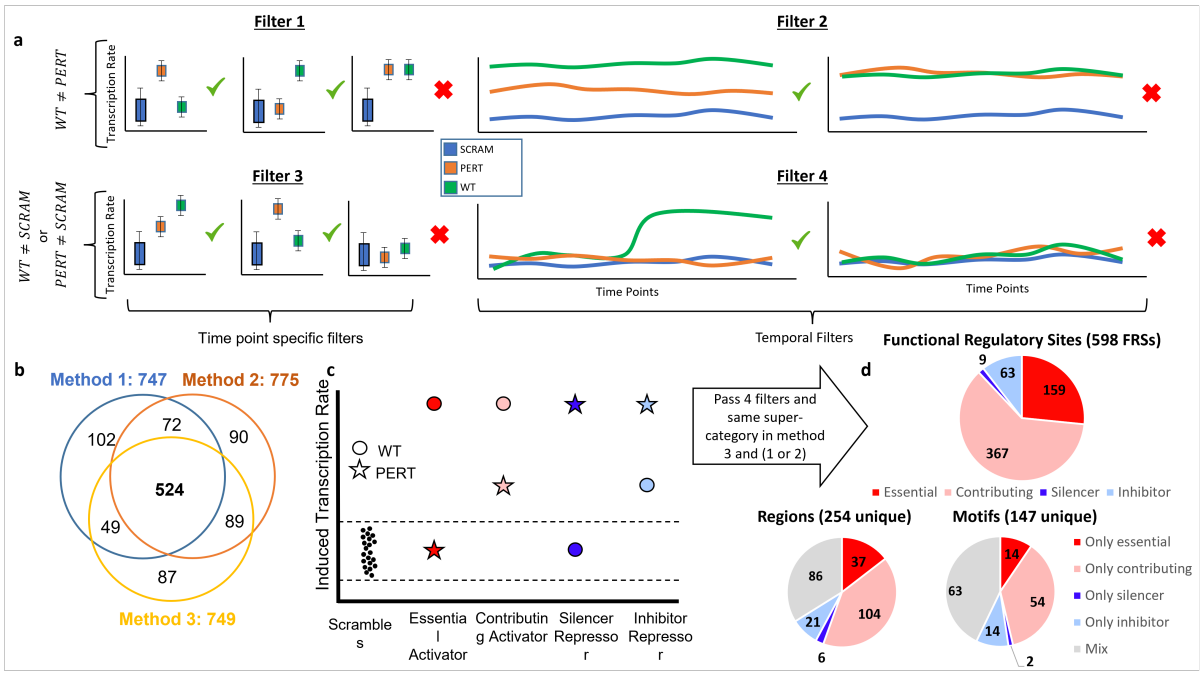


Figure 2: **Preprocessing, consistency, and categorization of FRSs.** **a**, Four filters applied to perturbed sequences to remove inactive and non-functional sites, both at each time-point and across time-points (Methods). **b**, number of sequences that passed all three filters for each perturbation method. **c**, Definition of main and sub-categories of motif binding effects based on their effect on transcription. **d**, Distribution of categories across FRSs that pass the 4 filters and are under the same main-category (activating or dampening) in perturbation method 3 and at least one of the perturbation methods 1 or 2. The distribution is shown across FRSs (top) and across the unique motifs and regions composing the FRSs in this study (bottom).

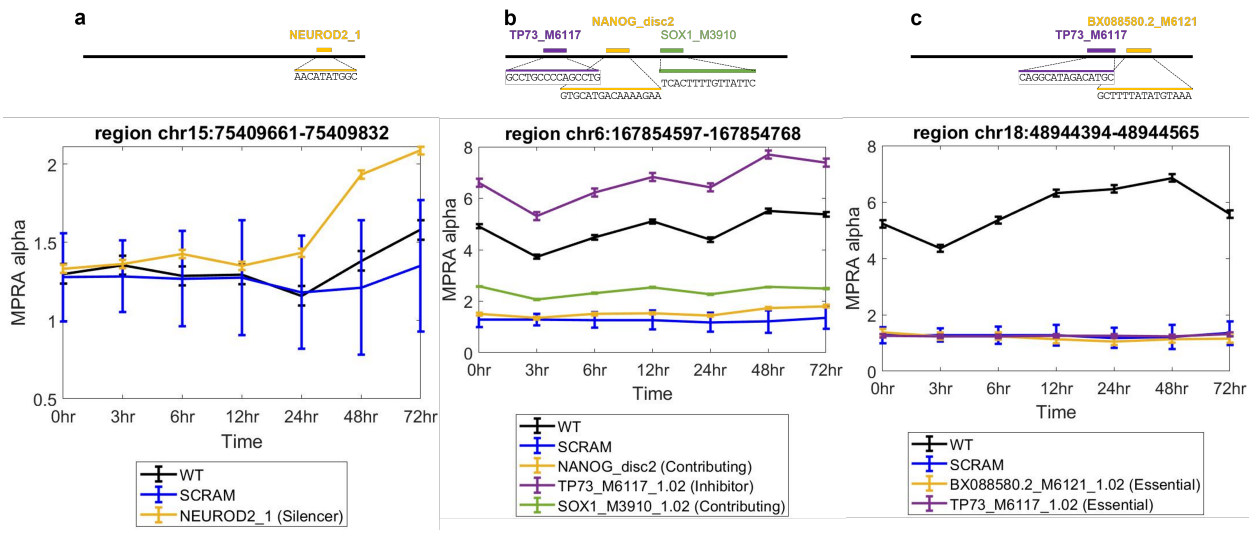


Figure 3: **Examples of the 4 subcategories of FRSs.** In each figure the WT sequence is indicated in black including error bars of ±1SD across the 3 replicates, and SCRAM in blue including error bars of ±1SD across all scrambled sequences. Each motif is plotted in a different color including error bars of ±1SD across the 3 replicates and its perturbation effect in the regions is indicated in the text box. **a.** NEOROD2 has a silencer effect. **b.** NANOG, TP73, SOX1 have contributing, inhibiting, contributing effects concordantly. **c.** BX088580.2, TP73 both have essential effects. All genomic coordinates are hg19.

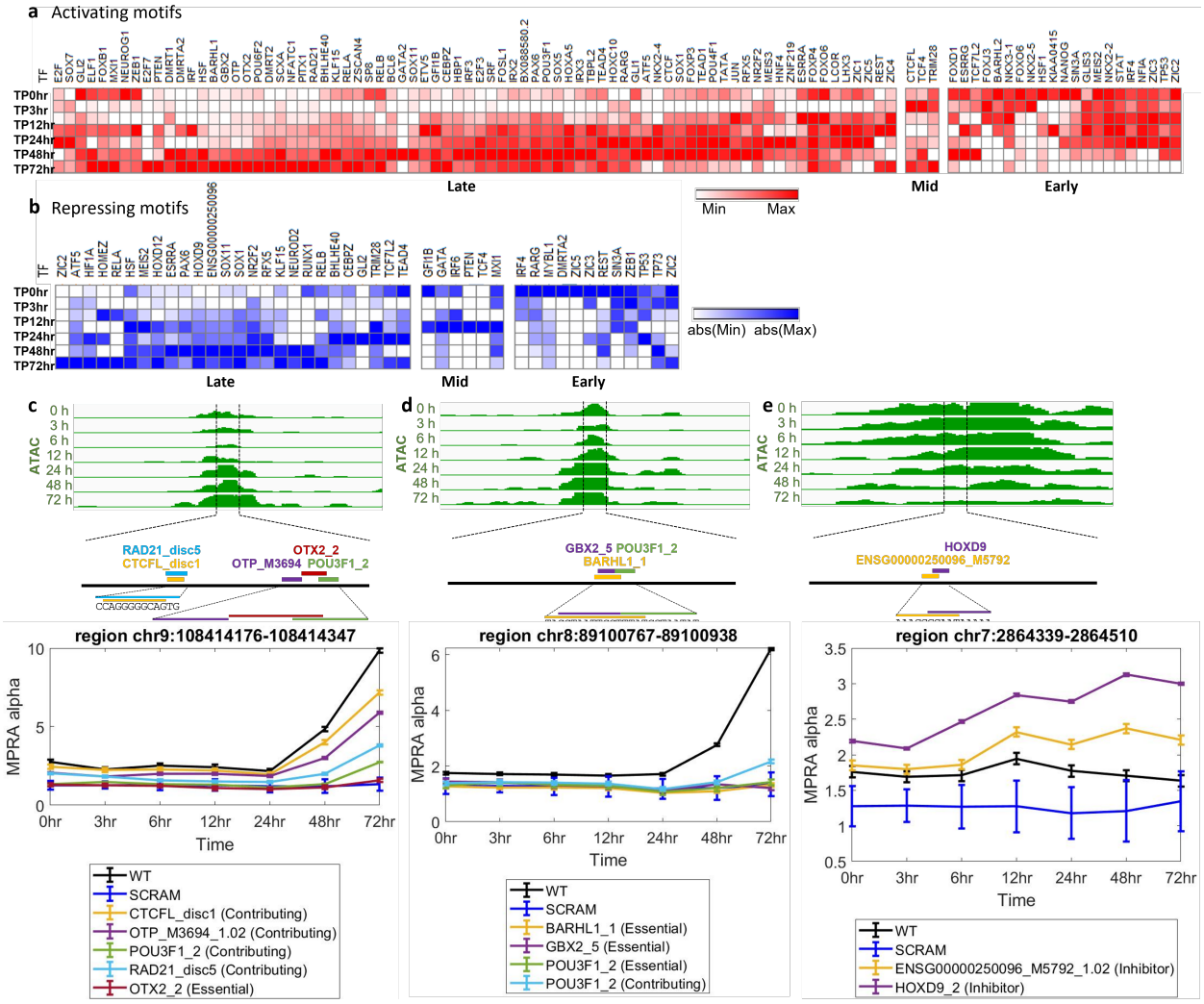


Figure 4: **Temporal motif effects.** The **a.** activating or **b.** repressing motifs in at least one time point. red - activator motifs, blue - repressor motifs. Color scale indicates the average of the perturbation signal (LogFC(WT/PERT)) across all significant instances of motifs for a specific TF (row normalized). Data is organized using hierarchical clustering and early, mid and late clusters are indicated. Genome browser snapshots of assayed sequences near predicted sequence motifs that are associated with **c.** OTX2 and **d.** BARHL1 TFs, showing the motifs that were perturbed and their effect on activity across time points. **e.** HOXD9 repressor motif example. Line plots similar to Fig. 3. All coordinates are hg19.

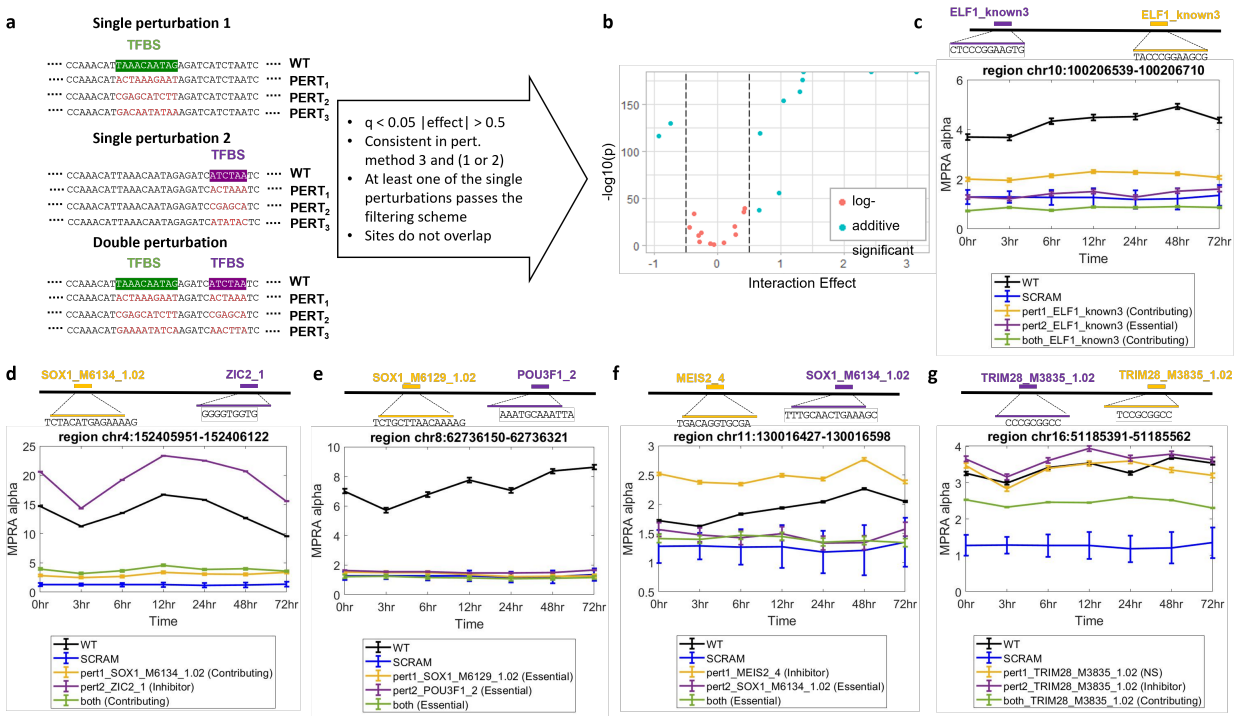


Figure 5: Double perturbation scheme. **a.** Experimental design for perturbing two single motifs separately and then a double perturbation of both simultaneously, and the requirements for being included in downstream analysis. **b.** Volcano plot for the model testing for log-additivity of the individual effects. **c-g.** Examples of double perturbation results demonstrating different patterns of cooperation: log-additive effects consistent with a billboard model **(c)**; log-additive effects of one dampening and one activating element **(d)**; fully-dependent cooperation consistent with the enhanceosome model **(e)**; a billboard-enhanceosome hybrid with one required element and on with a contributing effect **(f)**; a redundancy example, perturbing either motif has negligible effect, but perturbing both has a substantial effect **(g)**. All coordinates are hg19.

7.7 References

[1] Stefan Schoenfelder and Peter Fraser. *Long-range enhancer–promoter contacts in gene expression control*. 2019. DOI: 10.1038/s41576-019-0128-0. URL: http://dx.doi.org/10.1038/s41576-019-0128-0.

[2] M.T. Maurano et al. "Systematic localization of common disease-associated variation in regulatory DNA". en. In: *Science* 337.6099 (2012), pp. 1190–1195. DOI: 10.1126/science.1222794. URL: http://dx.doi.org/10.1126/science.1222794.

[3] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11247. URL: http://dx.doi.org/10.1038/nature11247.

[4] Nicola A Kearns et al. "Functional annotation of native enhancers with a Cas9-histone demethylase fusion". en. In: *Nat. Methods* 12.5 (May 2015), pp. 401–403. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3325. URL: http://dx.doi.org/10.1038/nmeth.3325.

[5] Gozde Korkmaz et al. "Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9". en. In: *Nat. Biotechnol.* 34.2 (Feb. 2016), pp. 192–198. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3450. URL: http://dx.doi.org/10.1038/nbt.3450.

[6] Fumitaka Inoue and Nadav Ahituv. "Decoding enhancers using massively parallel reporter assays". en. In: *Genomics* 106.3 (Sept. 2015), pp. 159–164. ISSN: 0888-7543, 1089-8646. DOI: 10.1016/j.ygeno.2015.06.005. URL: http://dx.doi.org/10.1016/j.ygeno.2015.06.005.

[7] Fumitaka Inoue et al. "A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity". en. In: *Genome Res.* 27.1 (Jan. 2017), pp. 38–52. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.212092.116. URL: http://dx.doi.org/10.1101/gr.212092.116.

[8] P. Kheradpour et al. "Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay". en. In: *Genome Res* 23.5 (May 2013), pp. 800–811. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.144899.112. URL: http://dx.doi.org/10.1101/gr.144899.112.

[9] Jamie C Kwasnieski et al. "High-throughput functional testing of ENCODE segmentation predictions". en. In: *Genome Res.* 24.10 (Oct. 2014), pp. 1595–1602. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.173518.114. URL: http://dx.doi.org/10.1101/gr.173518.114.

[10] Jacob C Ulirsch et al. "Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits". en. In: *Cell* 165.6 (June 2016), pp. 1530–1545. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.04.048. URL: http://dx.doi.org/10.1016/j.cell.2016.04.048.

[11] X. Wang et al. "High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human". en. In: *Nat. Commun* 9.1 (Dec. 2018), p. 5380. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07746-1. URL: http://dx.doi.org/10.1038/s41467-018-07746-1.

[12] Michael A White et al. "Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.29 (July 2013), pp. 11952–11957. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1307449110. URL: http://dx.doi.org/10.1073/pnas.1307449110.

[13] Abdenour Soufi, Greg Donahue, and Kenneth S Zaret. "Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome". en. In: *Cell* 151.5 (Nov. 2012), pp. 994–1004. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2012.09.045. URL: http://dx.doi.org/10.1016/j.cell.2012.09.045.

[14] N. Yosef et al. "Dynamic regulatory network controlling TH17 cell differentiation". en. In: *Nature* 496.7446 (Apr. 2013), pp. 461–468. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11981. URL: http://dx.doi.org/10.1038/nature11981.

[15] Mark D Biggin. "Animal transcription networks as highly connected, quantitative continua". en. In: *Dev. Cell* 21.4 (Oct. 2011), pp. 611–626. ISSN: 1534-5807, 1878-1551. DOI: 10.1016/j.devcel.2011.09.008. URL: http://dx.doi.org/10.1016/j.devcel.2011.09.008.

[16] Jane M Landolin et al. "Sequence features that drive human promoter function and tissue specificity". en. In: *Genome Res.* 20.7 (July 2010), pp. 890–898. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.100370.109. URL: http://dx.doi.org/10.1101/gr.100370.109.

[17] François Spitz and Eileen E M Furlong. "Transcription factors: from enhancer binding to developmental control". en. In: *Nat. Rev. Genet.* 13.9 (Sept. 2012), pp. 613–626. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3207. URL: http://dx.doi.org/10.1038/nrg3207.

[18] Troy W Whitfield et al. "Functional analysis of transcription factor binding sites in human promoters". en. In: *Genome Biol.* 13.9 (Sept. 2012), R50. ISSN: 1465-6906. DOI: 10.1186/gb-2012-13-9-r50. URL: http://dx.doi.org/10.1186/gb-2012-13-9-r50.

[19] S.R. Grossman et al. "Systematic dissection of genomic features determining transcription factor binding and enhancer function". en. In: *Proc. Natl. Acad. Sci. U. S. A* 114.7 (Feb. 2017), pp. 1291–1300. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1621150114. URL: http://dx.doi.org/10.1073/pnas.1621150114.

[20] Fumitaka Inoue et al. "Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction". en. In: *Cell Stem Cell* 25.5 (Nov. 2019), 713–727.e10. ISSN: 1934-5909, 1875-9777. DOI: 10.1016/j.stem.2019.09.010. URL: http://dx.doi.org/10.1016/j.stem.2019.09.010.

[21] Hannah K Long, Sara L Prescott, and Joanna Wysocka. "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution". en. In: *Cell* 167.5 (Nov. 2016), pp. 1170–1187. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.09.018. URL: http://dx.doi.org/10.1016/j.cell.2016.09.018.

[22] Tal Ashuach et al. "MPRAnalyze: statistical framework for massively parallel reporter assays". en. In: *Genome Biol.* 20.1 (Sept. 2019), pp. 1–17. ISSN: 1465-6906. DOI: 10.1186/s13059-019-1787-z. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1787-z.

[23] Charles E Grant, Timothy L Bailey, and William Stafford Noble. "FIMO: scanning for occurrences of a given motif". en. In: *Bioinformatics* 27.7 (Apr. 2011), pp. 1017–1018. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btr064. URL: http://dx.doi.org/10.1093/bioinformatics/btr064.

[24] P. Kheradpour and M. Kellis. "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments". en. In: *Nucleic Acids Res* 42.5 (2014), pp. 2976–2987. DOI: 10.1093/nar/gkt1249. URL: http://dx.doi.org/10.1093/nar/gkt1249.

[25] M.T. Weirauch et al. "Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity". en. In: *Cell* 158.6 (Sept. 2014), pp. 1431–1443. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2014.08.009. URL: http://dx.doi.org/10.1016/j.cell.2014.08.009.

[26] D Klein et al. "Accurate estimation of transduction efficiency necessitates a multiplex real-time PCR". en. In: *Gene Ther.* 7.6 (Mar. 2000), pp. 458–463. ISSN: 0969-7128. DOI: 10.1038/sj.gt.3301112. URL: http://dx.doi.org/10.1038/sj.gt.3301112.

[27] S.M. Chambers et al. "Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling". en. In: *Nat. Biotechnol* 27.3 (2009), pp. 275–280. DOI: 10.1038/nbt.1529. URL: http://dx.doi.org/10.1038/nbt.1529.

[28] M Grace Gordon et al. "lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements". en. In: *Nat. Protoc.* 15.8 (Aug. 2020), pp. 2387–2412. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/s41596-020-0333-5. URL: http://dx.doi.org/10.1038/s41596-020-0333-5.

[29] Makiko Iwafuchi-Doi and Kenneth S Zaret. "Cell fate control by pioneer transcription factors". en. In: *Development* 143.11 (June 2016), pp. 1833–1837. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.133900. URL: http://dx.doi.org/10.1242/dev.133900.

[30] Jun Aruga. "The role of Zic genes in neural development". en. In: *Mol. Cell. Neurosci.* 26.2 (June 2004), pp. 205–221. ISSN: 1044-7431. DOI: 10.1016/j.mcn.2004.01.004. URL: http://dx.doi.org/10.1016/j.mcn.2004.01.004.

[31] Gizem Guzelsoy et al. "Terminal neuron localization to the upper cortical plate is controlled by the transcription factor NEUROD2". en. In: *Sci. Rep.* 9.1 (Dec. 2019), p. 19697. ISSN: 2045-2322. DOI: 10.1038/s41598-019-56171-x. URL: http://dx.doi.org/10.1038/s41598-019-56171-x.

[32] Massimiliano Agostini et al. "p73 regulates maintenance of neural stem cell". en. In: *Biochem. Biophys. Res. Commun.* 403.1 (Dec. 2010), pp. 13–17. ISSN: 0006-291x, 1090-2104. DOI: 10.1016/j.bbrc.2010.10.087. URL: http://dx.doi.org/10.1016/j.bbrc.2010.10.087.

[33] F Talos et al. "p73 is an essential regulator of neural stem cell maintenance in embryonal and adult CNS neurogenesis". en. In: *Cell Death Differ.* 17.12 (Dec. 2010), pp. 1816–1829. ISSN: 1350-9047, 1476-5403. DOI: 10.1038/cdd.2010.131. URL: http://dx.doi.org/10.1038/cdd.2010.131.

[34] C D Hudson et al. "Brn-3a/POU4F1 interacts with and differentially affects p73-mediated transcription". en. In: *Cell Death Differ.* 15.8 (Aug. 2008), pp. 1266–1278. ISSN: 1350-9047. DOI: 10.1038/cdd.2008.45. URL: http://dx.doi.org/10.1038/cdd.2008.45.

[35] Minchul Kim et al. "Transcriptional co-repressor function of the hippo pathway transducers YAP and TAZ". en. In: *Cell Rep.* 11.2 (Apr. 2015), pp. 270–282. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.03.015. URL: http://dx.doi.org/10.1016/j.celrep.2015.03.015.

[36] Toshinori Ozaki, Natsumi Kubo, and Akira Nakagawara. "p73-Binding Partners and Their Functional Significance". en. In: *Int. J. Proteomics* 2010 (2010), p. 283863. ISSN: 2090-2166, 2090-2174. DOI: 10.1155/2010/283863. URL: http://dx.doi.org/10.1155/2010/283863.

[37] Alan W Leung et al. "Pre-Border Gene Foxb1 Regulates the Differentiation Timing and Autonomic Neuronal Potential of Human Neural Crest Cells". en. June 2019. DOI: 10.1101/646026. URL: https://www.biorxiv.org/content/10.1101/646026v1.

[38] P.-S. Hou et al. "LHX2 regulates the neural differentiation of human embryonic stem cells via transcriptional modulation of PAX6 and CER1". en. In: *Nucleic Acids Res* 41.16 (Sept. 2013), pp. 7753–7770. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkt567. URL: http://dx.doi.org/10.1093/nar/gkt567.

[39] Daisuke Kamiya et al. "Intrinsic transition of embryonic stem-cell differentiation into neural progenitors". en. In: *Nature* 470.7335 (Feb. 2011), pp. 503–509. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09726. URL: http://dx.doi.org/10.1038/nature09726.

[40] M.A. Lodato et al. "SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state". en. In: *PLoS Genet* 9.2 (Feb. 2013), p. 1003288. ISSN: 1553-7390, 1553-7404. DOI: 10.1371/journal.pgen.1003288. URL: http://dx.doi.org/10.1371/journal.pgen.1003288.

[41] Misako Matsushita et al. "Neural differentiation of human embryonic stem cells induced by the transgene-mediated overexpression of single transcription factors". en. In: *Biochem. Biophys. Res. Commun.* 490.2 (Aug. 2017), pp. 296–301. ISSN: 0006-291x, 1090-2104. DOI: 10.1016/j.bbrc.2017.06.039. URL: http://dx.doi.org/10.1016/j.bbrc.2017.06.039.

[42] Thomas Vierbuchen et al. "Direct conversion of fibroblasts to functional neurons by defined factors". en. In: *Nature* 463.7284 (Feb. 2010), pp. 1035–1041. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08797. URL: http://dx.doi.org/10.1038/nature08797.

[43] X. Zhang et al. "Pax6 is a human neuroectoderm cell fate determinant". it. In: *Cell Stem Cell* 7.1 (July 2010), pp. 90–100. ISSN: 1934-5909, 1875-9777. DOI: 10.1016/j.stem.2010.04.017. URL: http://dx.doi.org/10.1016/j.stem.2010.04.017.

[44] Yingsha Zhang et al. "Rapid single-step induction of functional neurons from human pluripotent stem cells". en. In: *Neuron* 78.5 (June 2013), pp. 785–798. ISSN: 0896-6273, 1097-4199. DOI: 10.1016/j.neuron.2013.05.029. URL: http://dx.doi.org/10.1016/j.neuron.2013.05.029.

[45] G D Frantz et al. "Otx1 and Otx2 define layers and regions in developing cerebral cortex and cerebellum". en. In: *J. Neurosci.* 14.10 (Oct. 1994), pp. 5725–5740. ISSN: 0270-6474. URL: https://www.ncbi.nlm.nih.gov/pubmed/7931541.

[46] Zhenghui Su et al. "Antagonism between the transcription factors NANOG and OTX2 specifies rostral or caudal cell fate during neural patterning transition". en. In: *J. Biol. Chem.* 293.12 (Mar. 2018), pp. 4445–4455. ISSN: 0021-9258, 1083-351x. DOI: 10.1074/jbc.M117.815449. URL: http://dx.doi.org/10.1074/jbc.M117.815449.

[47] Stephen N Sansom et al. "The level of the transcription factor Pax6 is essential for controlling the balance between neural stem cell self-renewal and neurogenesis". en. In: *PLoS Genet.* 5.6 (June 2009), e1000511. ISSN: 1553-7390, 1553-7404. DOI: 10.1371/journal.pgen.1000511. URL: http://dx.doi.org/10.1371/journal.pgen.1000511.

[48] J Pinsonneault et al. "A model for extradenticle function as a switch that changes HOX proteins from repressors to activators". en. In: *Embo J.* 16.8 (Apr. 1997), pp. 2032–2042. ISSN: 0261-4189. DOI: 10.1093/emboj/16.8.2032. URL: http://dx.doi.org/10.1093/emboj/16.8.2032.

[49] Rama Kadamb et al. "Sin3: insight into its transcription regulatory functions". en. In: *Eur. J. Cell Biol.* 92.8-9 (Aug. 2013), pp. 237–246. ISSN: 0171-9335, 1618-1298. DOI: 10.1016/j.ejcb.2013.09.001. URL: http://dx.doi.org/10.1016/j.ejcb.2013.09.001.

[50] Arven Saunders et al. "The SIN3A/HDAC Corepressor Complex Functionally Cooperates with NANOG to Promote Pluripotency". en. In: *Cell Rep.* 18.7 (Feb. 2017), pp. 1713–1726. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.01.055. URL: http://dx.doi.org/10.1016/j.celrep.2017.01.055.

[51] A. Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". en. In: *Proc. Natl. Acad. Sci. U. S. A* 102.43 (Oct. 2005), pp. 15545–15550. ISSN: 0027-8424. DOI: 10.1073/pnas.0506580102. URL: http://dx.doi.org/10.1073/pnas.0506580102.

[52] Jehangir N Ahmed et al. "Systematized reporter assays reveal ZIC protein regulatory abilities are Subclass-specific and dependent upon transcription factor binding site context". en. In: *Sci. Rep.* 10.1 (Aug. 2020), p. 13130. ISSN: 2045-2322. DOI: 10.1038/s41598-020-69917-9. URL: http://dx.doi.org/10.1038/s41598-020-69917-9.

[53] Waltraut Lehmann et al. "ZEB1 turns into a transcriptional activator by interacting with YAP1 in aggressive cancer types". en. In: *Nat. Commun.* 7 (Feb. 2016), p. 10498. ISSN: 2041-1723. DOI: 10.1038/ncomms10498. URL: http://dx.doi.org/10.1038/ncomms10498.

[54] Shalini Singh et al. "Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition". en. In: *Elife* 5 (May 2016). ISSN: 2050-084x. DOI: 10.7554/eLife.12717. URL: http://dx.doi.org/10.7554/eLife.12717.

[55] Alessandra di Masi et al. "Retinoic acid receptors: from molecular mechanisms to cancer therapy". en. In: *Mol. Aspects Med.* 41 (Feb. 2015), pp. 1–115. ISSN: 0098-2997, 1872-9452. DOI: 10.1016/j.mam.2014.12.003. URL: http://dx.doi.org/10.1016/j.mam.2014.12.003.

[56] David van Dijk et al. "Large-scale mapping of gene regulatory logic reveals context-dependent repression by transcriptional activators". en. In: *Genome Res.* 27.1 (Jan. 2017), pp. 87–94. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.212316.116. URL: http://dx.doi.org/10.1101/gr.212316.116.

[57] Gerald Stampfel et al. "Transcriptional regulators form diverse groups with context-dependent regulatory functions". en. In: *Nature* 528.7580 (Dec. 2015), pp. 147–151. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature15545. URL: http://dx.doi.org/10.1038/nature15545.

[58] Michael A White et al. "A Simple Grammar Defines Activating and Repressing cis-Regulatory Elements in Photoreceptors". en. In: *Cell Rep.* 17.5 (Oct. 2016), pp. 1247–1254. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2016.09.066. URL: http://dx.doi.org/10.1016/j.celrep.2016.09.066.

[59] Robin P Smith et al. "Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model". en. In: *Nat. Genet.* 45.9 (Sept. 2013), pp. 1021–1028. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.2713. URL: http://dx.doi.org/10.1038/ng.2713.

[60] Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. "Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci". en. In: *Nat. Commun.* 8 (Feb. 2017), p. 14519. ISSN: 2041-1723. DOI: 10.1038/ncomms14519. URL: http://dx.doi.org/10.1038/ncomms14519.

[61] P P Gao et al. "Regulation of topographic projection in the brain: Elf-1 in the hippocamposeptal system". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 93.20 (Oct. 1996), pp. 11161–11166. ISSN: 0027-8424. DOI: 10.1073/pnas.93.20.11161. URL: http://dx.doi.org/10.1073/pnas.93.20.11161.

[62] Makiko Iwafuchi-Doi et al. "Transcriptional regulatory networks in epiblast cells and during anterior neural plate development as modeled in epiblast stem cells". en. In: *Development* 139.21 (Nov. 2012), pp. 3926–3937. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.085936. URL: http://dx.doi.org/10.1242/dev.085936.

[63] S Ali M Shariati et al. "APLP2 regulates neuronal stem cell differentiation during cortical development". en. In: *J. Cell Sci.* 126.Pt 5 (Mar. 2013), pp. 1268–1277. ISSN: 0021-9533, 1477-9137. DOI: 10.1242/jcs.122440. URL: http://dx.doi.org/10.1242/jcs.122440.

[64] Susan J Harrison et al. "Sall1 regulates cortical neurogenesis and laminar fate specification in mice: implications for neural abnormalities in Townes-Brocks syndrome". en. In: *Dis. Model. Mech.* 5.3 (May 2012), pp. 351–365. ISSN: 1754-8403, 1754-8411. DOI: 10.1242/dmm.002873. URL: http://dx.doi.org/10.1242/dmm.002873.

[65] Valer Gotea et al. "Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers". en. In: *Genome Res.* 20.5 (May 2010), pp. 565–577. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.104471.109. URL: http://dx.doi.org/10.1101/gr.104471.109.

[66] Jessica E Davis et al. "Dissection of c-AMP Response Element Architecture by Using Genomic and Episomal Massively Parallel Reporter Assays". en. In: *Cell Syst* 11.1 (July 2020), 75–85.e7. ISSN: 2405-4720, 2405-4712. DOI: 10.1016/j.cels.2020.05.011. URL: http://dx.doi.org/10.1016/j.cels.2020.05.011.

[67] Eilon Sharon et al. "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters". en. In: *Nat. Biotechnol.* 30.6 (May 2012), pp. 521–530. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2205. URL: http://dx.doi.org/10.1038/nbt.2205.

[68] Shira Weingarten-Gabbay et al. "Systematic interrogation of human promoters". en. In: *Genome Res.* 29.2 (Feb. 2019), pp. 171–183. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.236075.118. URL: http://dx.doi.org/10.1101/gr.236075.118.

[69] M Kircher et al. "Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution". In: *Nat. Commun.* 10.1 (2019), 3583. doi: 10.1038/s41467–019–11526–w. ISSN: 2041-1723.

[70] M.J. Ziller et al. "Dissecting neural differentiation regulatory networks through epigenetic footprinting". en. In: *Nature* 518.7539 (Feb. 2015), pp. 355–359. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13990. URL: http://dx.doi.org/10.1038/nature13990.

[71] Alexander M Tsankov et al. "Transcription factor binding dynamics during human ES cell differentiation". en. In: *Nature* 518.7539 (Feb. 2015), pp. 344–349. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14233. URL: http://dx.doi.org/10.1038/nature14233.

[72] M Kanehisa and S Goto. "KEGG: kyoto encyclopedia of genes and genomes". en. In: *Nucleic Acids Res.* 28.1 (Jan. 2000), pp. 27–30. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27. URL: http://dx.doi.org/10.1093/nar/28.1.27.

[73] B Langmead and S L Salzberg. "Langmead. 2013. Bowtie2". In: *Nat. Methods* 9 (2013), pp. 357–359. ISSN: 1548-7091.

[74] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". en. In: *J. R. Stat. Soc. Series B Stat. Methodol.* 57.1 (1995), pp. 289–300. ISSN: 1369-7412, 0035-9246. URL: http://www.jstor.org/stable/2346101.

[75] Stefanie Hager et al. *An Internal Polyadenylation Signal Substantially Increases Expression Levels of Lentivirus-Delivered Transgenes but Has the Potential to Reduce Viral Titer in a Promoter-Dependent Manner*. 2008. DOI: 10.1089/hum.2007.165. URL: http://dx.doi.org/10.1089/hum.2007.165.

7.8 Supplementary Figures

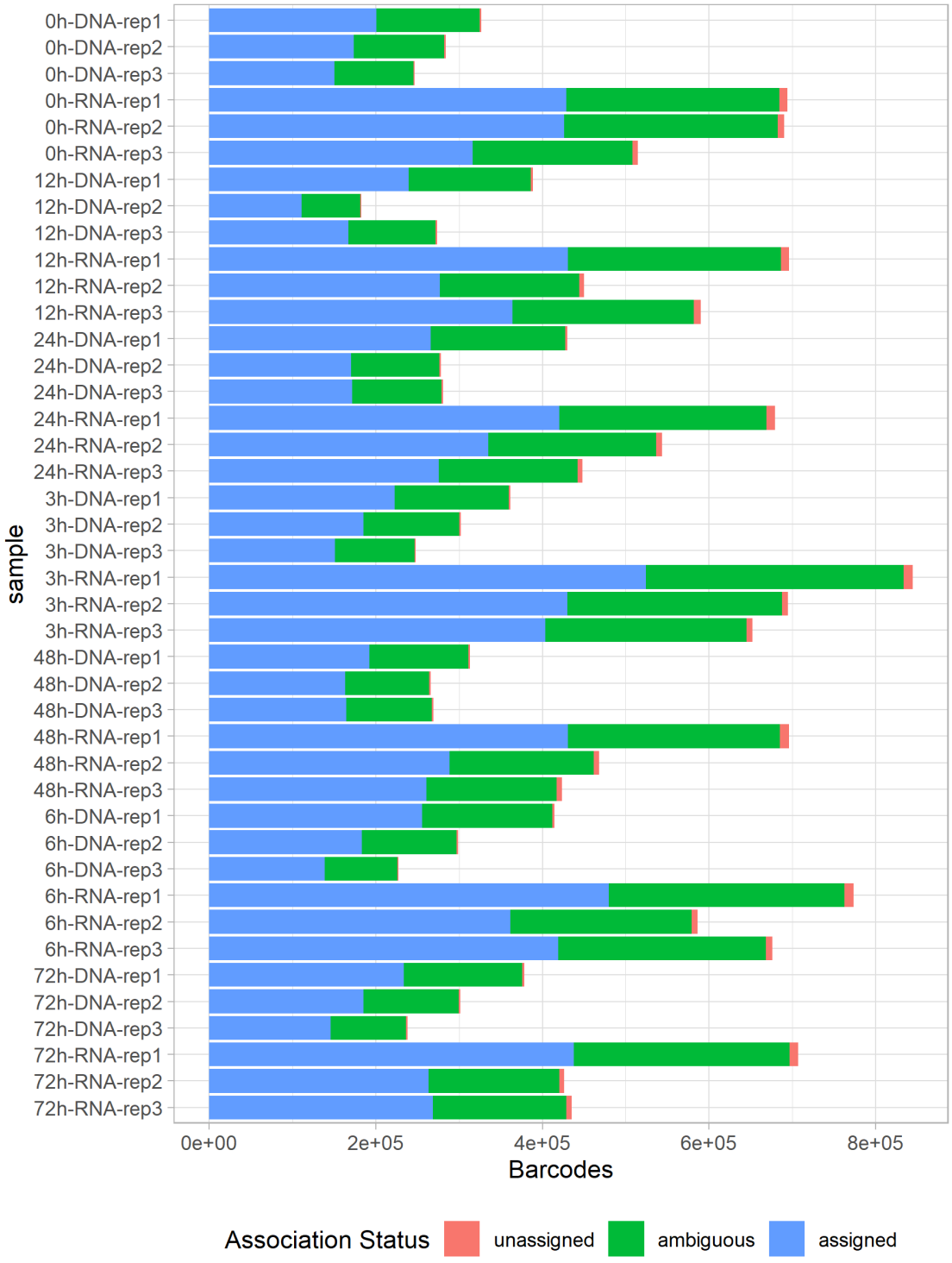


Figure S1: Barcode association status per sample.

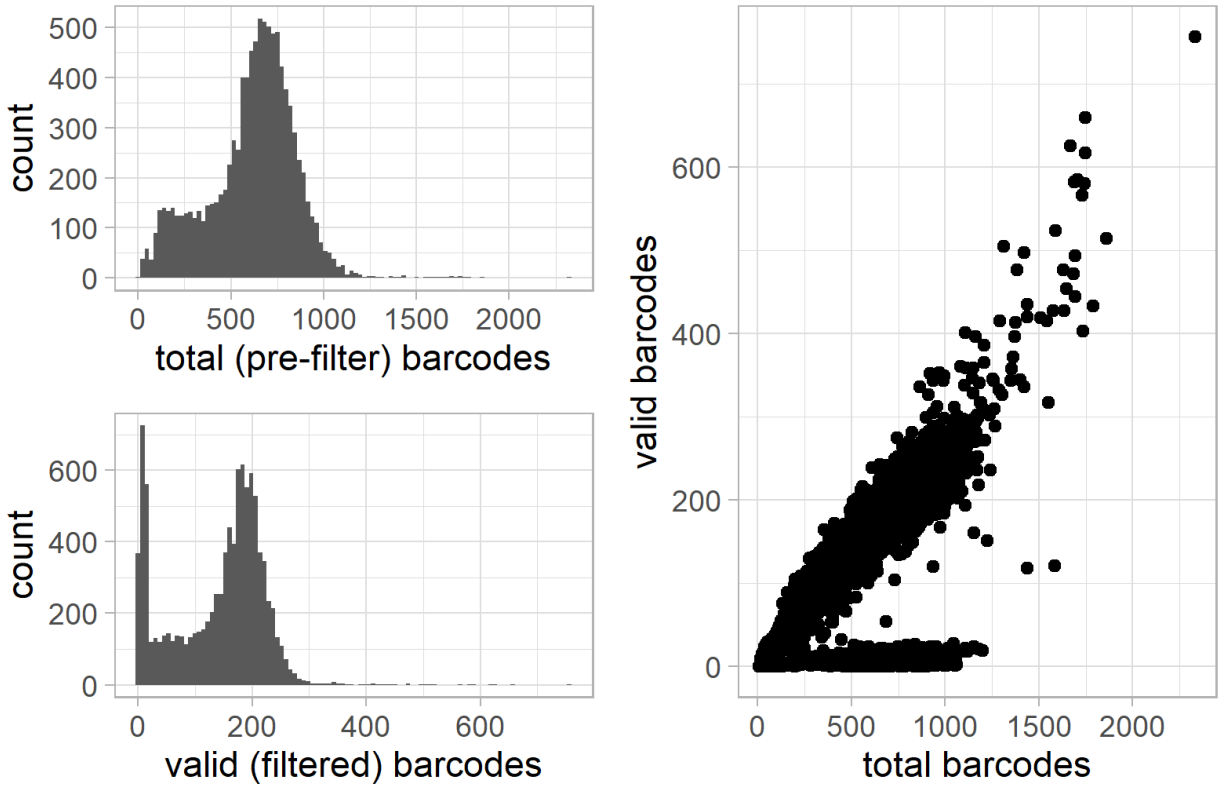


Figure S2: Barcodes per sequence per replicate – pre- and post-filtering.

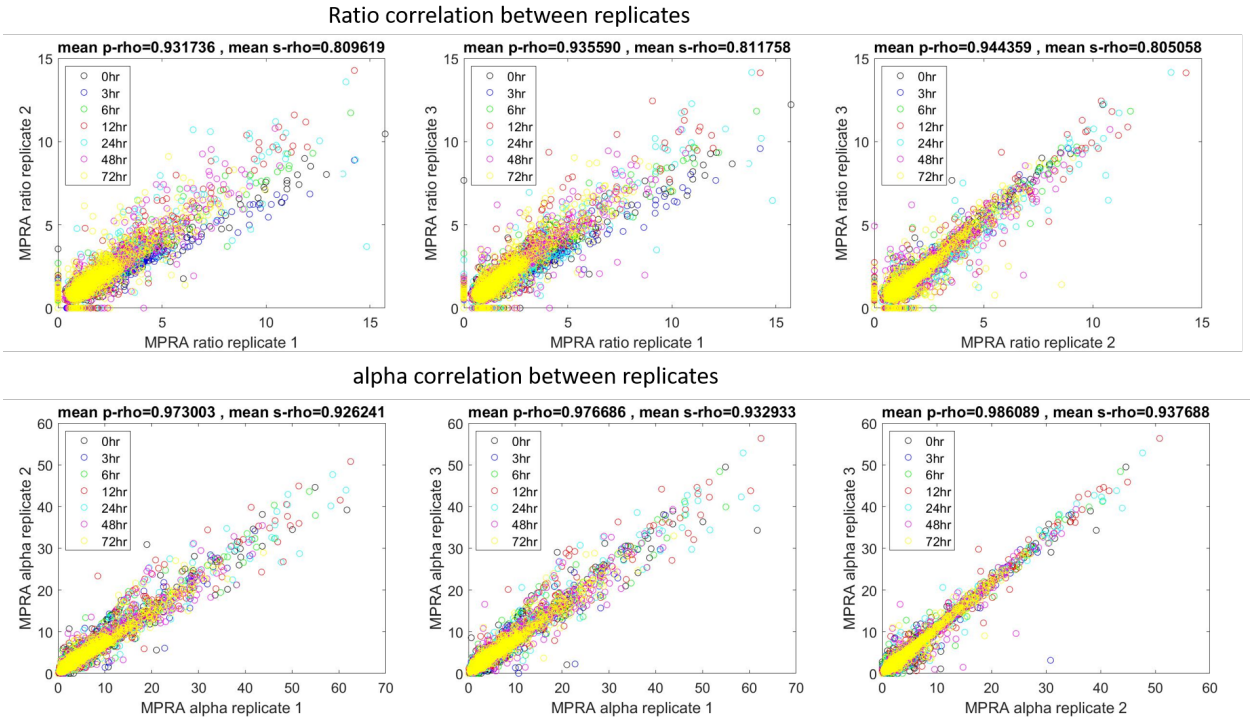


Figure S3: Correlation between replicates signal - ratio (upper panel) and alpha (bottom panel).

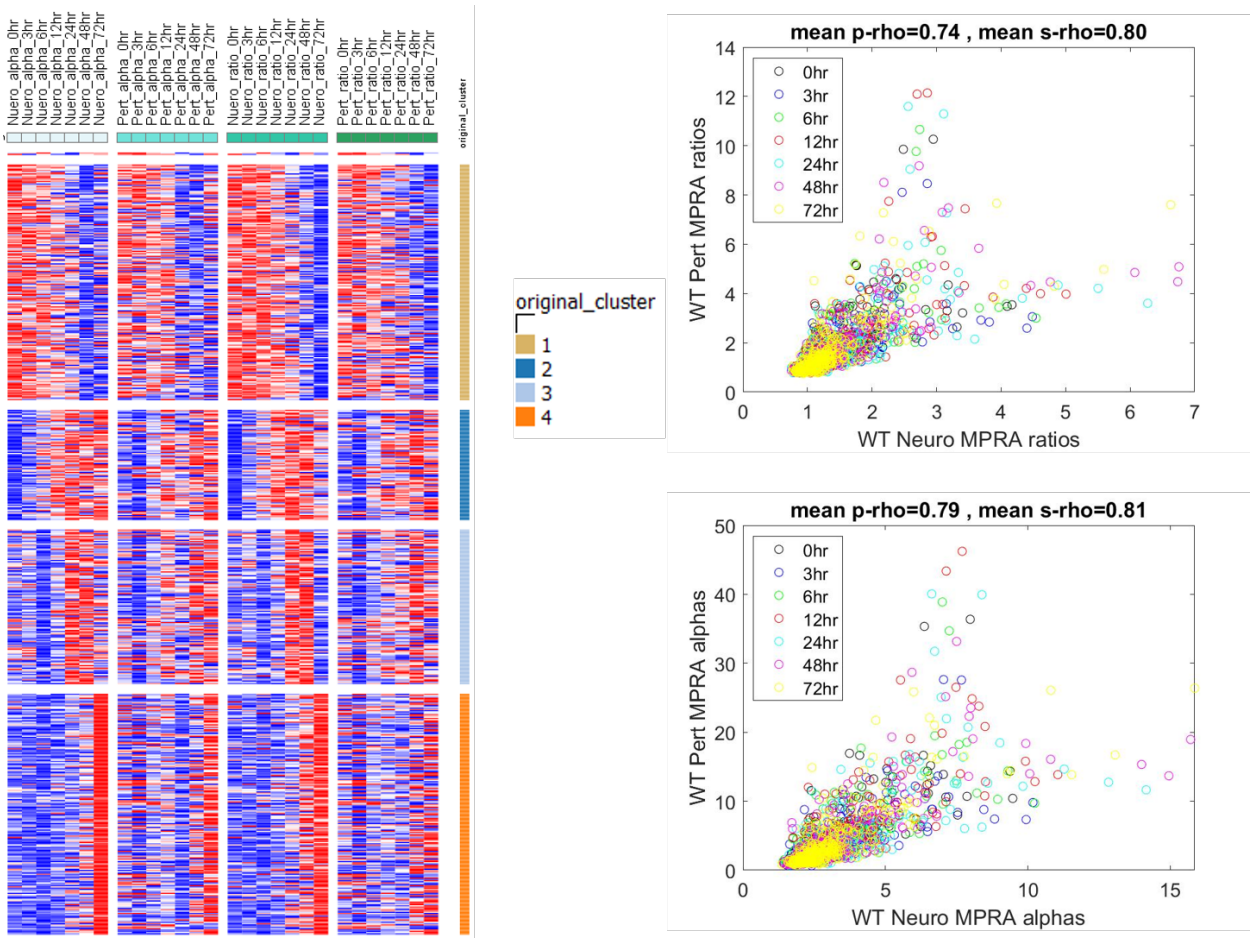


Figure S4: Reproducibility with Inoue et al. 2019. The left panel shows alpha (from MPRAnalyze) and ratio (RNA/DNA) for WT regions – comparing the original data in Inoue et al. 2019 (labeled 'Neuro') vs. the current paper (labeled 'Pert'). Each row shows normalized values – ranged from the lowest (blue) to the highest (red). The data is clustered based on the original MPRA clusters from Inoue et al. 2019. The right panel shows Pearson and Spearman correlations in each of the seven time points for both ratio (upper) and alpha (bottom) for WT regions between the two experiments.

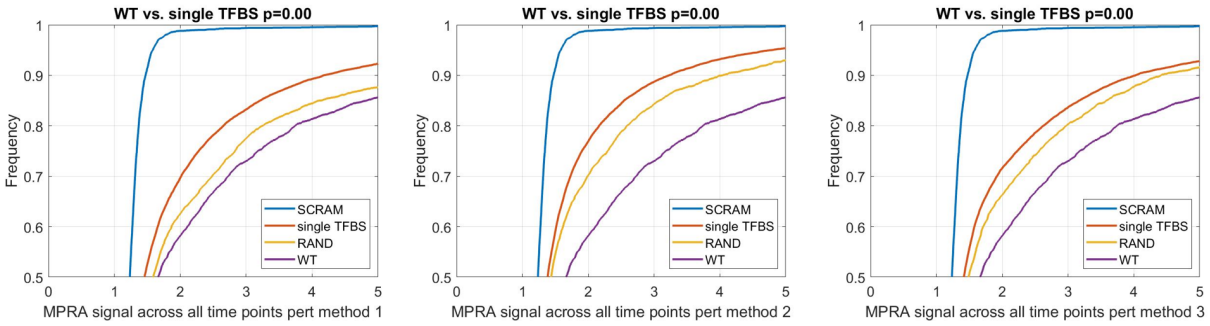


Figure S5: Cumulative distribution of the MPRA signal (alpha) across different categories for the three perturbation methods.

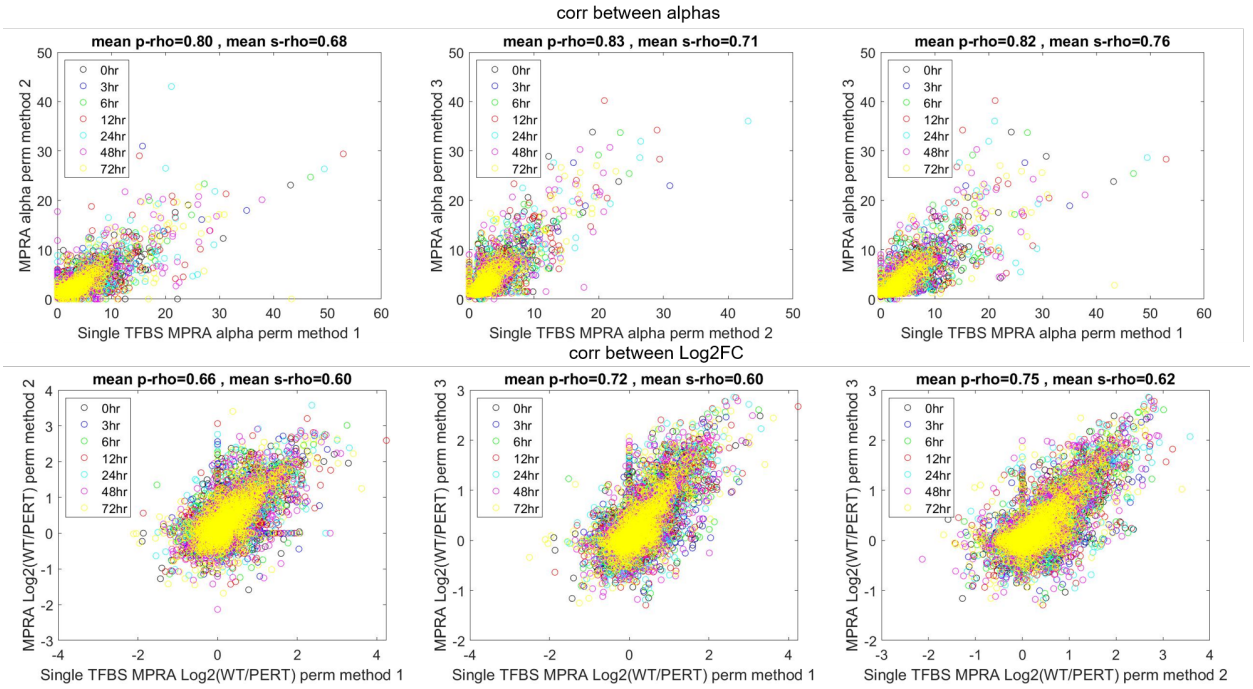


Figure S6: Correlation of alpha (upper panel) and Log2FC (bottom panel) per time point, between the different perturbation methods.

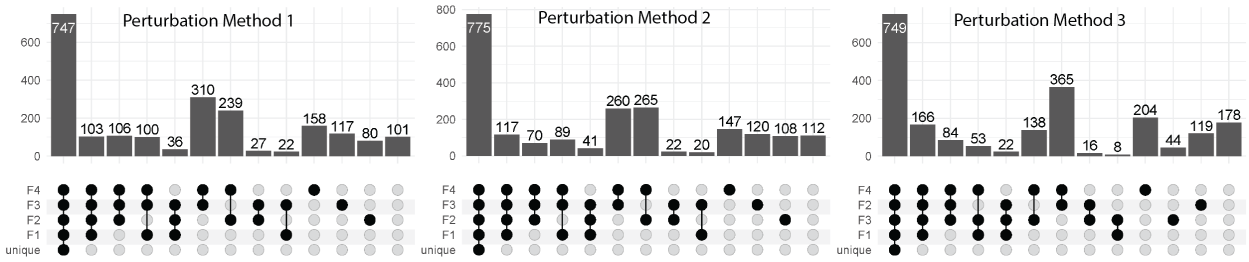


Figure S7: Plots showing the number of tested sequences that pass each filter, with each perturbation method. Histograms show number of sequences counted for each combination of filters, bottom panels describe each combination (filled dot indicates passing the filter, empty otherwise).

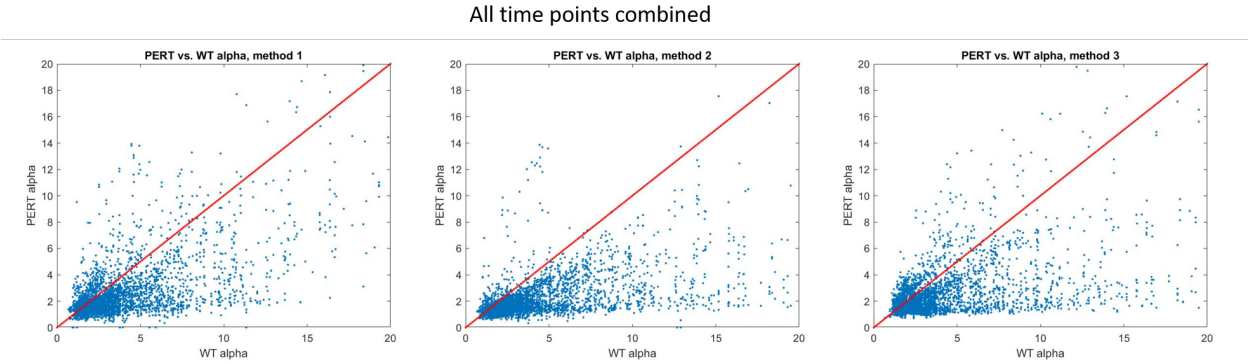


Figure S8: PERT alpha vs. WT alpha for all 3 perturbation methods.

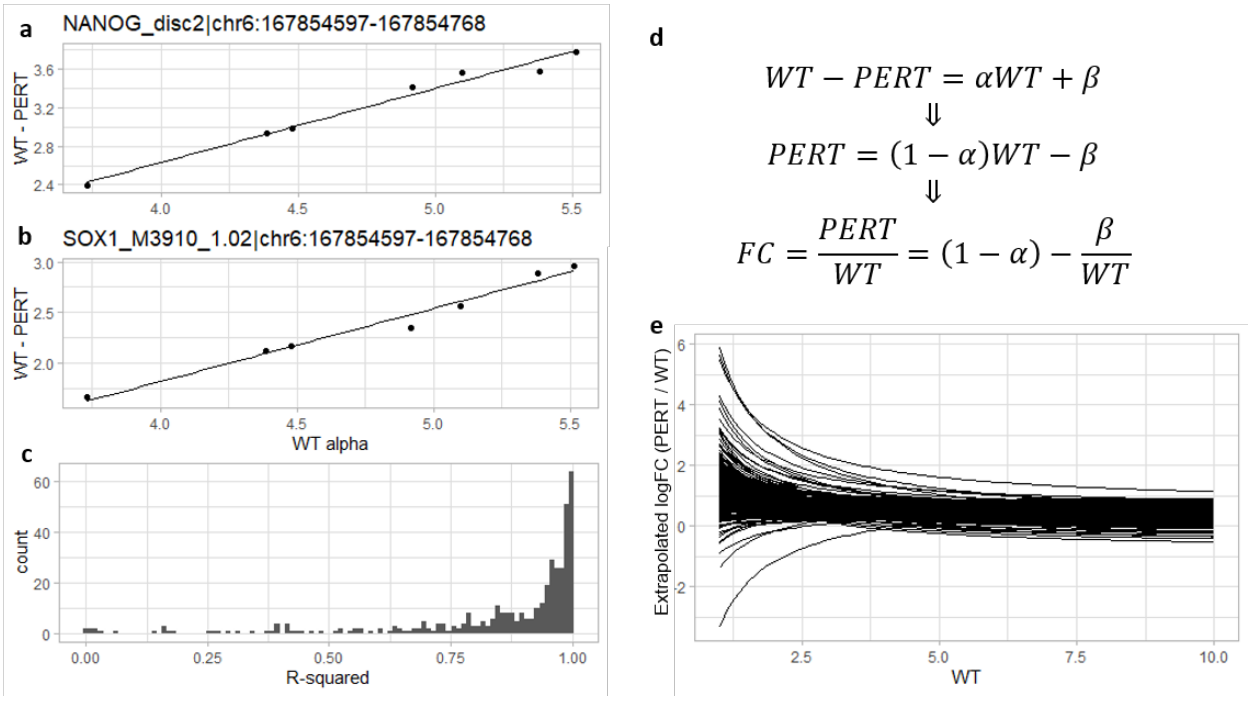


Figure S9: **A linear relationship between the absolute perturbation effect and the WT activity levels across time. a-b.** examples of a fitted linear regression line (WT delta, where delta = WT – PERT) for two contributing FRSs in the same region. Both display a clear linear relationship with a different slope. **c.** R^2 values for all fitted models shows that the relationship is overwhelmingly linear across FRSs. **d.** transition from absolute effect as a linear function of the WT activity to the fold-change values as a function of WT activity. **e.** using the model parameters from the models fitted for each FRSs to extrapolate the log Fold-change values for that FRS as a function of WT activity. The FC decays to a constant in sufficiently high activity levels.

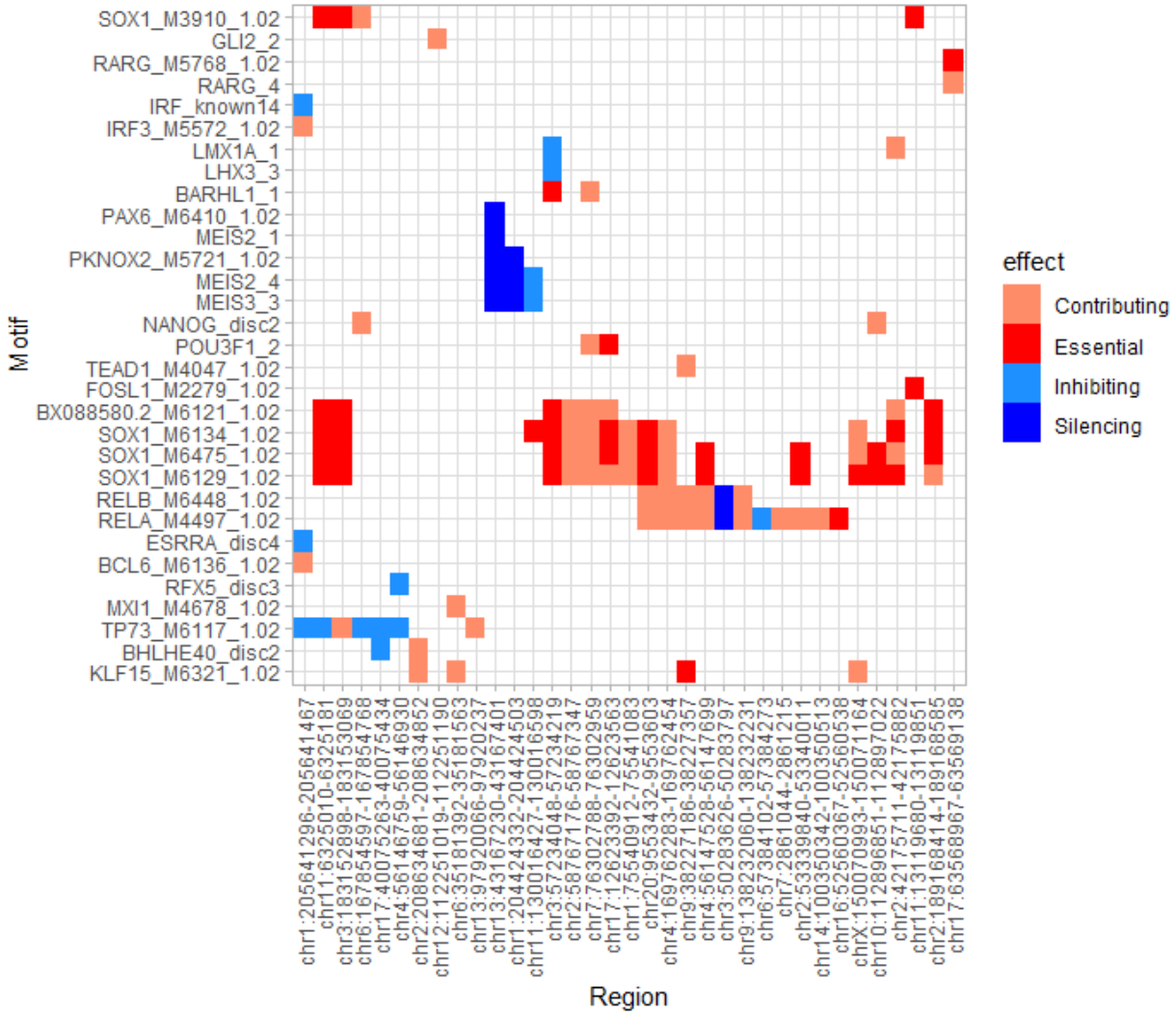


Figure S10: Functional FRSs heatmap, showing the perturbed motif (Y-axis) and the genomic region (X-axis). Colors correspond to the FRS category. Randomly Selected 50 motifs and 50 regions were selected for visualization.

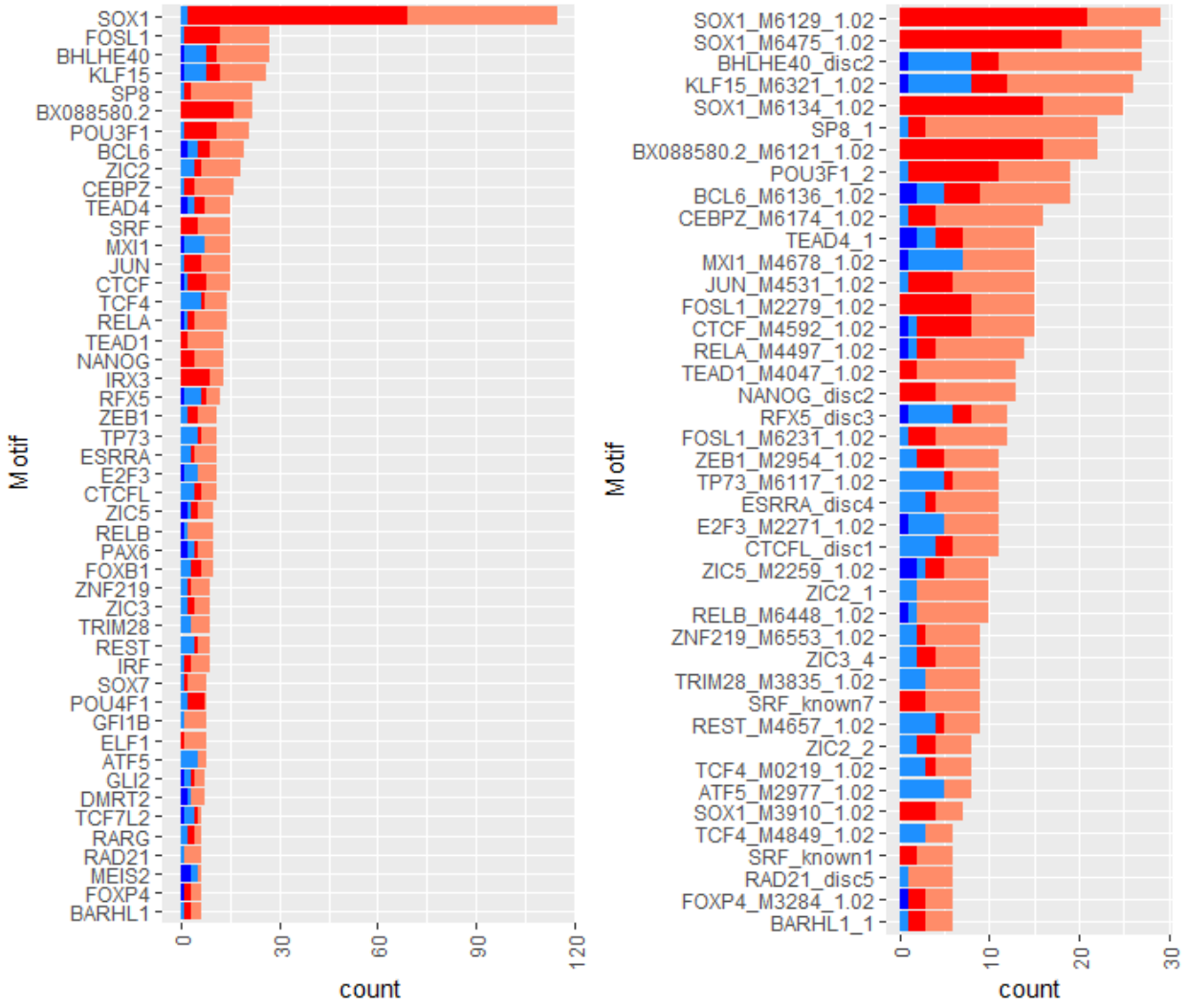


Figure S11: Bar plots illustrating the composition of categories for each motif (left) and aggregated across motifs for each transcription factor (right). Rare features were removed for visual vlarity.

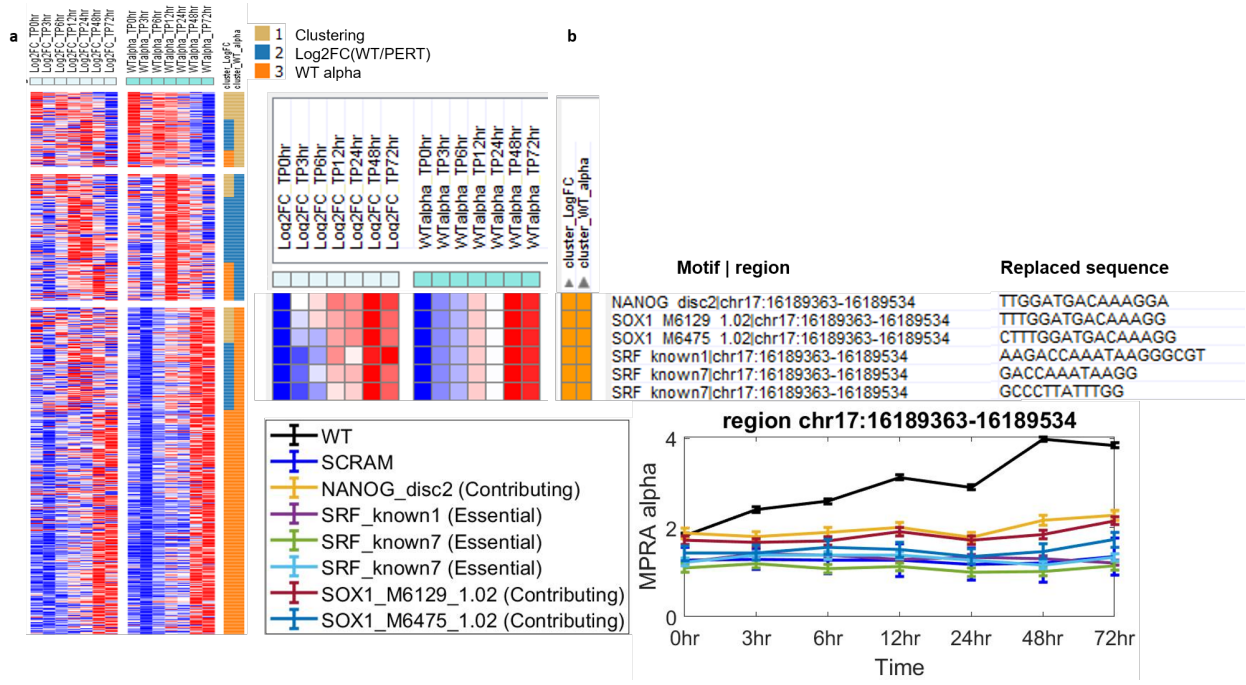


Figure S12: **(a)** Temporal clustering of FRSs signal (WT, Log2FC), clustered by the WT alpha signal and then by the Log2FC signal. Each row shows normalized values – ranged from the lowest (blue) to the highest (red). **(b)** Sequence specific temporal effects.

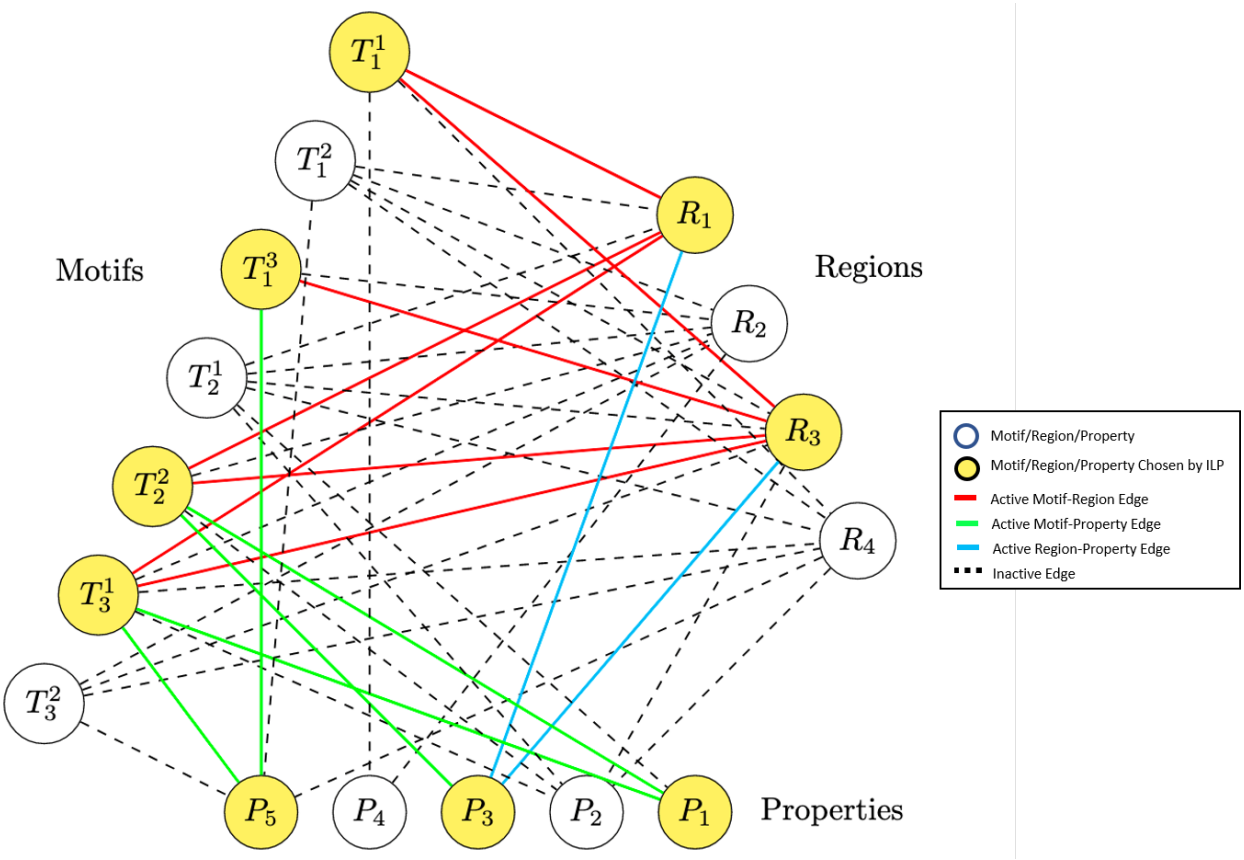


Figure S13: Motifs and regions selection scheme represented by a tri-partite graph.

7.9 Supplementary Materials

All supplemental information for this chapter is included in *Chapter7_Additional_Files.zip*. The files are:

- **Table S1_4-7** supplementary tables S1, S4, S5, S6, S7. **S1** Filtering statistics and overlap between perturbation methods. **S4** CRS categories (activators/repressors) statistics and Overlap of categories between perturbation methods. **S5** Motif categories (repressive/active) statistics and Overlap of categories between perturbation methods. **S6** Region categories (repressive/active) statistics and Overlap of categories between perturbation methods. **S7** CRS categories (Activators: essential, contributing. Repressors: silencers, inhibitors) statistics.
- **Table S2** single TFBS perturbation effects
- **Table S3** single TFBS active/repressive

- **Table S8** Candidate CRSs for driving state specific activity
- **Table S9** double TFBS perturbation effects