

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Natural and synthetic genetic regulation in stimulated CD4+ T cells

Permalink

<https://escholarship.org/uc/item/2n09q75r>

Author

Gate, Rachel E

Publication Date

2019

Supplemental Material

<https://escholarship.org/uc/item/2n09q75r#supplemental>

Peer reviewed|Thesis/dissertation

Natural and synthetic genetic regulation in stimulated CD4+ T cells

by
Rachel Gate

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

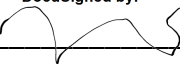
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:


DocuSigned by:

7CBD3BB0EE2B418... Jimmie Ye
Chair

DocuSigned by:

Michael T McManus

DocuSigned by:

Katherine Pollard

DocuSigned by:

Jonathan Weissman
A16A9953185F4C9...

Committee Members

Dedication and Acknowledgments

I would like to dedicate my thesis to my family, friends, and UCSF community who have supported me throughout this journey. First and foremost, I'd like to thank my mentor, Dr. Jimmie Ye, with whom none of this would be possible. My thesis committee, Drs. Katherine Pollard, Jonathan Weissman, and Michael McManus, who have been instrumental in my scientific development. I would also like to thank my parents, Karen and Michael Gate, who have been the exemplary role models of perseverance and grit. Finally, I'd like to thank Gregory Gate, William Edwards, and Meena Subramaniam, who have been an amazing support system and continue to challenge me to be a better person and scientist.

The text of this thesis is a reprint of the material as it appears in Nature Genetics and Biorxiv. The chapter entitled "Genetic determinants of co-accessible chromatin regions in activated T cells across humans" was published in Nature Genetics in 2018 (PMID: 29988122; doi: 10.1038/s41588-018-0156-2). The chapter entitled "Mapping gene regulatory networks of primary CD4⁺ T cells using single-cell genomics and genome engineering" was published in a preprint journal (doi: <https://doi.org/10.1101/678060>). The coauthors listed in these publications directed and supervised the research that forms the basis for the thesis. Dr. Jimmie Ye clarifies that Rachel E. Gate preformed the computational analysis, figure generation, and paper writing in "Genetic determinants of co-accessible chromatin regions in activated T cells across humans" and additionally carried out the experimental design in "Mapping gene regulatory networks of primary CD4⁺ T cells using single-cell genomics and genome engineering".

Natural and synthetic genetic regulation in stimulated CD4⁺ T cells

Rachel Elena Gate

Abstract

Over 90% of genetic variants associated with complex human traits map to non-coding regions, but little is understood about how they modulate gene regulation in health and disease. One possible mechanism is that genetic variants affect the activity of one or more *cis*-regulatory elements leading to gene expression variation in specific cell types. To identify such cases, we analyzed ATAC-seq and RNA-seq profiles from stimulated primary CD4⁺ T cells in up to 105 healthy donors. We found regions of accessible chromatin (ATAC-peaks) are co-accessible at kilobase and megabase resolution, consistent with the 3D chromatin organization measured by *in situ* Hi-C in T cells. 15% of genetic variants located within ATAC-peaks affected the accessibility of the corresponding peak (ATAC-QTLs). ATAC-QTLs have the largest effects on co-accessible peaks, are associated with gene expression, and are enriched for autoimmune disease variants. Our results provide insights into how natural genetic variants modulate *cis*-regulatory elements, in isolation or in concert, to influence gene expression.

To elucidate the *trans* regulatory network governing activation and polarization of CD4⁺ T cells we sequenced the transcriptomes of ~160k CD4⁺ T cells from 9 donors following pooled CRISPR perturbation targeting 140 regulators. We identified 134 regulators that affect T cell functionalization, including *IRF2* as a positive regulator of Th₂ polarization. Leveraging correlation patterns between cells, we mapped 194 pairs of interacting regulators, including known (e.g. *BATF* and *JUN*) and novel interactions (e.g. *ETS1* and *STAT6*). Finally, we identified 80 natural genetic variants with effects on gene expression, 48 of which are modified by a

perturbation. In CD4⁺ T cells, CRISPR perturbations can influence *in vitro* polarization and modify the effects of *trans* and *cis* regulatory elements on gene expression.

Table of Contents

<i>Chapter 1: Introduction</i>	1
References	5
<i>Chapter 2: Genetic determinants of co-accessible chromatin regions in activated T cells across humans</i>	9
Introduction.....	11
Results	12
Discussion.....	23
References	26
Figures	32
Materials and Methods.....	40
Materials and Methods References.....	53
Supplementary Figures.....	55
References:.....	77
<i>Chapter 3: Mapping gene regulatory networks of primary CD4⁺ T cells using single-cell genomics and genome engineering</i>	78
Introduction.....	80
Results	81
Discussions.....	94
References	97

Figures	110
Materials and Methods.....	119
Supplementary Figures.....	135

List of Figures

<i>Figure 2.1. Changes in chromatin state in human T cell activation.....</i>	<i>32</i>
<i>Figure 2.2. Changes in transcription factor enrichment in response to T cell activation.....</i>	<i>33</i>
<i>Figure 2.3. Inter-individual chromatin co-accessibility.</i>	<i>34</i>
<i>Figure 2.4. Genetic variants that affect chromatin states in human T cell activation.....</i>	<i>36</i>
<i>Figure 2.5. Genetic determinants of co-accessible peaks.....</i>	<i>37</i>
<i>Figure 2.6. Association of chromatin accessibility and gene expression.....</i>	<i>39</i>
<i>Figure 3.1: CRISPR perturbation screen in activated CD4⁺ T cells across donors</i>	<i>110</i>
<i>Figure 3.2: Heterogeneity of activated CD4⁺ T cells</i>	<i>112</i>
<i>Figure 3.3. Regulator perturbations drive T cell polarization and maintenance</i>	<i>114</i>
<i>Figure 3.4. Perturbations and single cell analysis reveal transcription factor interactions.....</i>	<i>116</i>
<i>Figure 3.5: CRISPR perturbation modifies genetic effects on gene expression</i>	<i>117</i>

List of Tables

Table S2.1 stimulation reponse

Table S2.2 hic

Table S2.3 covars mismatches

Table S2.4 pc exp cov correlation

Table S2.5 coaccessibility

Table 2.S6 atacqtl

Table S2.7 atac h analysis

Table S2.8 eqtl

Table S2.9 expression h

Table S2.10 stimulation response

Table S3.1 sgRNA info

Table S3.2 single cell metadata supp table

Table S3.3 sgRNA proportions and CE

Table S3.4 sgrna clusterprop

Table S3.5 tf interaction supp table

Table S3.6 indiv

Table S3.7 eqtl

Chapter 1: Introduction

The adaptive immune system is the defense system against infections, which is often slower to respond than the innate immune system yet has the ability to unleash targeted defenses and get stronger with each infection. At the center of the adaptive immune system are CD4⁺ T cells, which develop in the thymus as CD45RA⁺/CD45RO⁻/CCR7⁺/CD62L⁺/CD27⁺ naive CD4⁺ T cells that then migrate to the peripheral blood. Then, moving in between the periphery and tissue they are poised to activate and localize to the affected tissue¹. Once activated, CD4⁺ T cells can differentiate into a myriad of lineages, including cytotoxic and helper cells^{2,3}. Although cytotoxic responses have historically been attributed to CD8⁺ cells, more recently they have been shown to have MHC class II receptor activity, giving rise to cytotoxic CD4⁺ T cells⁴⁻⁷. In contrast to direct killing, helper CD4⁺ T cells support the immune system by guiding it in a targeted response^{8,9}. Initially in 1986 there were two helper CD4⁺ cells subtypes identified, IFNG⁺ Th₁ guiding bacterial infections and IFNG⁻ Th₂ that direct allergy responses, but further subtyping has since identified, but not limited to, RORγ⁺ Th₁₇, induced FOXP3⁺ Treg, and CXCR13⁺ Tfh subtypes⁹⁻¹¹.

As geneticist, phenotypes include molecular and biochemical readouts from functional genomics. Recent advances in next generation sequencing have improved cell phenotypic annotations in eukaryotic cells, including assaying the chromatin state through chromatin accessible regions, DNase I hypersensitive sites, histone modifications, and transcription factor binding sites (TFBS)¹²⁻¹⁵. Variation in chromatin state has been implicated in disease, such as Sjorgen's, where regions of hypomethylations in patients are related to T cell activation^{16,17,18}. To that end, only 1% of our genome encodes for protein and over 90% of genome wide association study (GWAS)

variants reside in non-coding regions^{19,20}. Unlike coding regions, which have an obvious impact on a gene by potentially changing the amino acid code, it is unclear how genetic variants in non-coding regions influence gene expression. This is primarily due to the fact that gene expression is a complex trait, governed by an intricate network of *cis*-regulatory elements and *trans* factors. More pointedly, the *cis* and *trans* regulatory networks governing activated CD4⁺ T cells has yet to be elucidated.

Cis means proximal to in Latin, and correspondingly *cis* regulation encompasses regulatory regions that are nearby the respective gene that consists of enhancers and promoters^{21,22,23–25}. Unlike promoters, enhancers are much more context specific and can influence genes up to 1 Mb away, with multiple enhancers governing one gene²⁶. Genetic analysis of single nucleotide polymorphisms (SNPs) in chromatin regions that vary in their accessibility has proven to be a powerful way to assign functional interpretation of genetic variations^{27–29}. Previous work in lymphoblastoid cell lines (LCLs) studying DNase I hypersensitivity sites²⁷ and histone modifications^{30–32} have laid the groundwork for chromatin state profiling. However, we have yet to elucidate CD4⁺ T cell context specific *cis* regulatory regions. Additionally, the contribution of interindividual variation to *cis* regulatory variation has yet to be elucidated, largely due to inhibitory labor time, cost, and sample requirement.

With the advent of ATAC-seq in 2015, we profiled chromatin accessible regions in CD4⁺ T cells in 105 donors. Here, we present work that identifies the chromatin state of stimulated and unstimulated CD4⁺ T cells, and their associated genetic variants. We also identified regions of co-

accessibility, which are regions whose variability are correlated, corresponding to mostly enhancer-enhancer associations, as well as identified their genetic determinants.

Trans regulation on the other hand corresponds to regulation that occurs far away from the respective gene. *Trans* regulation, governed by *trans* factors, includes transcription factors (TFs), chromatin remodelers, and RNA binding proteins. However, *trans* associations are notoriously difficult to detect due to a high multiple testing burden and potential for false positives, leading to unreproducible results³³. With the advent of CRISPR editing in primary T cells³⁴, followed by genome wide CRISPR screens, we began to dissect *trans* associations genome wide. For example, previous work mapping *trans* factors in Th₂ identified *Trappc12*, *Mpv17l2*, and *Pou6f1* as *trans* regulators in a bulk, negative selection screen³⁵. However, these bulk sequencing techniques are limited to specific phenotype screening, do not capture the heterogeneity of cell states, and are often limited to a few donors due to labor and batch effects. Therefore, the extent of the variability of the *trans* networks is incomplete, this is compounded by the lack of interindividual variation associations.

Droplet-based single cell sequencing (dscRNA-seq) has enabled transcriptomic profiling at unparalleled resolution, capturing a high resolution snapshot of the immune system³⁶. DscRNA-seq data allows for unbiased annotation of CD4⁺ T cells and a continuous phenotype, which captures the true heterogeneity of CD4⁺ T cells. Additionally, dscRNA-seq allows us to move beyond just capturing average gene expression, to understanding the variance.

In 2016, Dixit et al 2016, Adamson et al. 2016, and Jaitin et al. 2016 coupled dscRNA-seq and CRISPR screening to perturbatively profile thousands of cells by barcoding each perturbation³⁷⁻³⁹. More recently, Datlinger et al. 2017 developed CROP-seq, which is a dscRNA-seq CRISPR screening method that removes the need for a barcode to reduce barcoding mismatching⁴⁰.

Out of necessity to bring down the cost of dscRNA-seq to perform large scale studies, our lab developed an experimental framework, mux-seq, and a corresponding computational algorithm, demuxlet, to reduce the overall cost of dscRNA-seq by 10-fold⁴¹. In short, experimentally we pool all genetically distinct donors into one well of dscRNA-seq and then computationally using the unique combination of SNPs belonging to each donor we can identify which cell came from which donor in the pool. In general, we need 20 SNPs per cell to uniquely identify the donor of origin.

Therefore, by combing CROP-seq and mux-seq, we were able to profile hundreds of *trans* factors at an unprecedented resolution in multiple donors, while minimizing batch effects. We identified CD4⁺ T cell subsets, as well as sgRNA knockout (KO) conditions that were enriched in each of these subsets, where each sgRNA KO corresponds to a *trans* factor. Specifically, we found novel *IRF2* regulation of Th₂ cells and tested for regulator interactions, both regulator - regulator and regulator - SNP to for the first time systematically identify *cis-trans* epistatic interactions.

References

1. van den Broek, T., Borghans, J. A. M. & van Wijk, F. The full spectrum of human naive T cells. *Nat. Rev. Immunol.* 18, 363–373 (2018).
2. Alberts, B. et al. *The Adaptive Immune System.* (Garland Science, 2002).
3. Koch, U. & Radtke, F. Mechanisms of T cell development and transformation. *Annu. Rev. Cell Dev. Biol.* 27, 539–562 (2011).
4. Taniuchi, I. CD4 Helper and CD8 Cytotoxic T Cell Differentiation. *Annu. Rev. Immunol.* 36, 579–601 (2018).
5. Takeuchi, A. & Saito, T. CD4 CTL, a Cytotoxic Subset of CD4+ T Cells, Their Differentiation and Function. *Front. Immunol.* 8, 194 (2017).
6. Wagner, H., Starzinski-Powitz, A., Jung, H. & Röllinghoff, M. Induction of I region-restricted hapten-specific cytotoxic T lymphocytes. *J. Immunol.* 119, 1365–1368 (1977).
7. Billings, P., Burakoff, S., Dorf, M. E. & Benacerraf, B. Cytotoxic T lymphocytes specific for I region determinants do not require interactions with H-2K or D gene products. *J. Exp. Med.* 145, 1387–1392 (1977).
8. Zhu, J. & Paul, W. E. CD4 T cells: fates, functions, and faults. *Blood* 112, 1557–1569 (2008).
9. Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A. & Coffman, R. L. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J. Immunol.* 136, 2348–2357 (1986).
10. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu. Rev. Immunol.* 28, 445–489 (2010).
11. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*

- 21.29.1–21.29.9 (2015). doi:10.1002/0471142727.mb2129s109
12. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57 (2009).
 13. Raha, D., Hong, M. & Snyder, M. ChIP-Seq: a method for global identification of regulatory elements in the genome. *Curr. Protoc. Mol. Biol.* Chapter 21, Unit 21.19.1–14 (2010).
 14. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* 2010, db.prot5384 (2010).
 15. Durek, P. et al. Epigenomic Profiling of Human CD4⁺ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. *Immunity* 45, 1148–1161 (2016).
 16. Moskowitz, D. M. et al. Epigenomics of human CD8 T cell differentiation and aging. *Science immunology* 2, (2017).
 17. Altorok, N. et al. Genome-wide DNA methylation patterns in naive CD4⁺ T cells from patients with primary Sjögren’s syndrome. *Arthritis & rheumatology* 66, 731–739 (2014).
 18. Zhao, R. F. ENCODE: Deciphering function in the human genome. (2012).
 19. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195 (2012).
 20. Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. *Transcription: an overview of gene regulation in eukaryotes.* (W. H. Freeman, 2000).
 21. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69 (2011).
 22. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.*

8, 206–216 (2007).

23. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
24. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100 (2012).
25. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90 (2012).
26. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343 (2015).
27. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394 (2012).
28. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* 489, 75–82 (2012).
29. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
30. Kasowski, M. et al. Extensive variation in chromatin states across humans. *Science* 342, 750–752 (2013).
31. McVicker, G. et al. Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–749 (2013).
32. Kilpinen, H. et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 342, 744–747 (2013).
33. Ashis Saha, A. B. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res.* 7, (2018).

34. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620 (2015).
35. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14, 297 (2017).
36. Adamson, B. et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* 167, 1867–1882.e21 (2016).
37. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853–1866.e17 (2016).
38. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94 (2018).

Chapter 2: Genetic determinants of co-accessible chromatin regions in activated T cells across humans

Rachel E. Gate^{1,2,21}, Christine S. Cheng^{3,4,21,22}, Aviva P. Aiden^{5,6}, Atsede Siba³, Marcin Tabaka³, Dmytro Lituiev¹, Ido Machol⁵, M. Grace Gordon², Meena Subramaniam^{1,2}, Muhammad Shamim^{5,7}, Kendrick L. Hougen⁸, Ivo Wortman³, Su-Chen Huang⁵, Neva C. Durand⁵, Ting Feng⁹, Philip L. De Jager^{3,10,11}, Howard Y. Chang¹², Erez Lieberman Aiden^{5,7,13-15}, Christophe Benoist⁹, Michael A. Beer^{8,16}, Chun J. Ye^{1,17-19,22}, Aviv Regev^{3,20,22}

¹Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA

²Biological and Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, California, USA

³Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

⁴Department of Biology, Boston University, Boston, Massachusetts, USA

⁵Department of Molecular and Human Genetics, the Center for Genome Architecture, Baylor College of Medicine, Houston, Texas, USA

⁶Department of Bioengineering, Rice University, Houston, Texas, USA

⁷Medical Scientist Training Program, Baylor College of Medicine, Houston, Texas, USA

⁸Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA

⁹Division of Immunology, Department of Microbiology and Immunology, Harvard Medical School, Boston, Massachusetts, USA

¹⁰Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology and Psychiatry, Division of Genetics, Department of Medicine, Brigham

and Women's Hospital, Boston, Massachusetts, USA

¹¹Harvard Medical School, Boston, Massachusetts, USA

¹²Center for Personal Dynamic Regulomes, Stanford University, Stanford, California, USA

¹³Department of Computer Science, Rice University, Houston, Texas, USA

¹⁴Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA

¹⁵Center for Theoretical Biological Physics, Rice University, Houston, Texas, USA

¹⁶McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, USA

¹⁷Institute of Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA

¹⁸Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, California, USA

¹⁹Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, USA

²⁰Howard Hughes Medical Institute, Koch Institute of Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

²¹These authors contributed equally to this work.

²²Corresponding authors. Email: aregev@broadinstitute.org (A.R); jimmie.ye@ucsf.edu (C.J.Y.); chcheng@bu.edu (C.S.C)

Introduction

The vast majority of disease-associated loci identified through genome-wide association studies (GWAS)¹⁻³ are located in non-coding regions of the genome, often distant from the nearest gene⁴. Quantitative trait loci (QTL) studies that associate genetic variants with molecular traits provide a framework for assessing the gene regulatory potential of disease-associated variants. For example, a statistically significant number of GWAS loci are associated with gene expression (expression QTLs – eQTLs) across diverse cell types and states⁵⁻¹⁰, implicating gene regulation in determining disease risk^{11,12}.

Genetic analysis of variation in chromatin state¹³⁻¹⁷ is a powerful approach for identifying single nucleotide polymorphisms (SNPs) that directly affect *cis*-regulatory activity¹⁸. In lymphoblastoid cell lines, thousands of SNPs have been associated with DNase I hypersensitivity (measured by DNase-seq)¹⁹ and histone tail modifications (measured by ChIP-seq)²⁰⁻²². Similarly, SNPs have been associated with variation in DNA methylation and histone tail modifications in resting primary immune cell types (neutrophils, monocytes and CD4⁺/CD45RA⁺ effector memory T cells)¹². Most of the associated SNPs in these studies were also associated with nearby transcript abundance, suggesting the genetic perturbation of *cis*-regulatory activity as a determinant of gene expression variability^{11,12,19,21,23}.

These studies have provided foundational resources for understanding the genetic basis of gene regulation in resting cells, but many disease states are associated with immune cell activation^{24,25}. In particular, dysregulation of T cell homeostasis and activation are known to play a role in autoimmunity^{26,27}, cancer^{28,29} and infectious diseases³⁰, and hundreds of SNPs have been associated with gene expression during T cell activation and polarization^{10,31}. Moreover, because

both DNase-seq and ChIP-seq are laborious and require large cell numbers, it remains challenging to apply them to primary human cells at the scale required for genetic association.

Here, we optimized and performed Assay for Transposase Accessible Chromatin sequencing (ATAC-seq) on stimulated CD4⁺ T cells from 105 healthy individuals to characterize the extent of natural variability in chromatin state, identify its genetic basis, and assess its influence on gene expression. We further leveraged the variability between individuals to identify co-accessible chromatin regions and to relate those to genetic variation and 3D genome organization. Our work helps lay the foundation for the critical tasks of annotating *cis*-regulatory elements in primary human T cells and characterizing how genetic variation contribute to variability in gene regulation between individuals.

Results

Changes in T cell chromatin state in response to activation

We used ATAC-seq³² to assay CD4⁺ T cells in two different conditions: either unstimulated (Th) or stimulated *in vitro* using tetrameric antibodies against CD3 and CD28 for 48 hours (Th_{stim}) (**Fig. 2.1a**). Aligned reads from six samples (five donors, one pair of donors were replicates) were pooled for each condition, yielding a total of 209 million reads for Th_{stim} and 58 million for Th cells (**Methods**). Of these five donors, two are of East Asian, two are of African American and one is of European decent (**Supplementary Fig. 2.1**). There was a global increase in chromatin accessibility in response to stimulation, with 52,154 chromatin accessible peaks detected in Th_{stim} (average width: 483 bp +/- 344 bp) and 36,487 in Th cells (average width of 520 bp +/- 319 bp) (MACS2, FDR < 0.05, **Fig. 2.1a, b**). Down sampling each Th_{stim} sample to the same number of reads as the matching Th sample yielded a similar trend (24,665 Th_{stim} vs. 17,313 Th peaks)

suggesting the increased accessibility is not due to differences in sequencing depth. Of the 63,763 peaks identified in at least one condition, 27,446 are similarly accessible between the conditions (shared peaks), 28,017 are more accessible in Th_{stim} cells (Th_{stim}-specific peaks) (FDR < 0.05), and 8,298 are more accessible in Th cells (Th-specific peaks) (FDR < 0.05) (**Fig. 2.1a, b**, and **Supplementary Table 2.1**).

Peaks of accessible chromatin are associated with distinctive genomic features and enriched for SNPs associated with autoimmune diseases. Compared to Th-specific peaks, Th_{stim}-specific peaks overlap a higher percentage of enhancers defined using H3K27Ac marks¹⁸ in aCD3/aCD28- (Th₀, 6.9% vs. 2.6%) and phorbol myristate acetate (PMA)-stimulated CD4⁺ T cells (Th_{stim}, 7.2% vs. 3.6%); and a lower percentage of enhancers in regulatory (T_{reg}, 1.4% vs. 4.0%), naïve (T_{naive}, 1.2% vs. 4.9%) and IL17 producing CD4⁺ T cells (Th₁₇, 3.2% vs. 4.6%) (**Fig. 2.1c**)¹⁸. Th_{stim}-specific and shared peaks also overlap a higher percentage of SNPs associated with autoimmune diseases, including inflammatory bowel disease (IBD) (32% and 41% vs. 20%) and rheumatoid arthritis (21% and 27% vs. 13%) (**Fig. 2.1d**), highlighting the importance of profiling cells under stimulation to identify disease-relevant *cis*-regulatory elements.

Analyzing peaks of accessible chromatin in aggregate provides estimates of the frequencies and single-nucleotide resolution footprints of transcription factor (TF) binding³². Th_{stim}-specific peaks are enriched for genomic locations bound by TFs important for CD4⁺ T cell activation or differentiation, including members of the AP-1 super family (*e.g.*, 36% contain a BATF binding site) and interferon regulatory factors (*e.g.*, 15% contain a IRF4 binding site)³³⁻³⁵ (**Fig. 2.2a, b**). Th_{stim}-specific peaks overlapping regions bound by both BATF and IRF4 (17.4% of peaks)³⁴ reveal

a different footprint compared to those overlapping regions bound by only one of the TFs (**Fig. 2.2b**, left). Conversely, shared peaks are enriched for regions bound by CTCF and BORIS (encoded by CTCFL), two transcriptional repressors known to maintain chromatin state independent of cell type and state³³⁻³⁵ (**Fig. 2.2a**), and their binding footprints are invariant of condition (**Fig. 2.2b**, right). ETS1 binding sites overlapping shared, Th_{stim}-, and Th-specific peaks have distinct footprints and binding motifs: we observed the canonical ETS1 motif (5'-CACTTCCTGT-3') in shared peaks, a 3' extended motif (5'-CACTTCCTGTCA-3') in Th-specific peaks, and a T/G to T (5'-CACTTCCTGT-3') substitution at the eighth position in Th_{stim}-specific peaks, consistent with sequence motifs found at distal ETS1 binding sites (**Fig. 2.2c**)³⁶. Th-specific peaks are more likely to overlap ETS/RUNX binding sites than shared or Th_{stim}-specific peaks (OR = 2.7 and 3.9; Fisher's exact test, P -value $< 2.2 \times 10^{-16}$ and P -value $< 2.2 \times 10^{-16}$, respectively) (**Fig. 2.2d**), which could be due to an enrichment of Th-specific peaks for T_{reg} enhancers known to be bound by the ETS/RUNX complex^{37,38}. An additional 6,102 Th_{stim}-specific (6.6% of intergenic regions) and 4,118 shared peaks (4.5% of intergenic regions) were located in non-coding regions, previously not annotated by H3K27Ac^{18,39}, of which 53.5% and 35.6% overlap known binding sites for TFs in the AP-1 super family and IRF family, respectively. Thus, regions of accessible chromatin overlap both known enhancers and TF binding sites important for polarization-independent activation of T cells, consistent with our stimulation protocol, and in aggregate reveal high-resolution footprints distinguishing condition specific and combinatorial transcription factor binding.

Chromatin co-accessibility at multiple genomic scales

Because Th_{stim}-peaks, including shared and Th_{stim}-specific peaks, better overlap known T cell *cis*-

regulatory elements and autoimmune disease loci, we next characterized the inter-individual variability of chromatin accessibility only in stimulated T cells. We optimized the ATAC-seq protocol to profile stimulated CD4⁺ T cells (**Supplementary Fig. 2.2; Methods**) from 105 healthy donors in the ImmVar Consortium¹⁰, all of European descent (**Fig. 2.3a** and **Supplementary Fig. 2.1**). We obtained a median of 37 million (MAD +/-13 million) reads per sample, from highly complex libraries with low mitochondrial DNA (mtDNA) contamination (average contamination < 3%, **Supplementary Fig. 2.3**). Using a pool of 4.2 billion merged reads from all 105 individuals, we jointly called 167,140 peaks of accessible chromatin (hereto after, ATAC-peaks) (MACS2, FDR < 0.05, **Fig. 2.3a, b**). These included 85.1% of the 52,154 Th_{stim} peaks identified in the initial set of six samples from five individuals with similar enrichment for GWAS loci (Pearson R = 0.65) and enhancer elements (Pearson R = 0.88) (**Supplementary Fig. 2.4**).

Leveraging the variability in ATAC-peaks across 105 individuals, we found patterns of co-accessibility (defined as correlation between individual or sets of ATAC-peaks) at multiple genomic scales, recapitulating the 3D chromatin organization, as determined by domain-resolution *in situ* Hi-C⁴⁰ of stimulated CD4⁺ T cells pooled from another five donors (**Supplementary Table 2.2**, and **Supplementary Fig. 2.5**). At the resolution of 1 Mb bins, we observed significant intra-chromosomal co-accessibility, as measured by correlation of total counts of ATAC-peaks within each bin (Chr1: **Fig. 2.3c**, other chromosomes: **Supplementary Fig. 2.6**). These pairwise correlations are qualitatively similar to and quantitatively consistent with (Pearson R = 0.66) Hi-C interaction frequencies at the same resolution (**Fig. 2.3d** and **Supplementary Fig. 2.6**), likely reflecting variability in the signal (regions of accessible chromatin) to noise (regions of inaccessible chromatin) ratio across samples similar to observations in single cells³². At 100 kb

resolution, pairwise correlations are also consistent with Hi-C interaction frequencies (Pearson R = 0.52, **Supplementary Fig. 2.7**).

We next characterized the co-accessibility between pairs of ATAC-peaks within each 1.5 Mb bin across the genome by linear regression (**Fig. 2.3b**, dashed black line, left). After accounting for sources of variation (**Supplementary Tables 2.3 and 2.4**), we found 2,158 pairs of co-accessible peaks enriched for those in close proximity (on average 514 kb apart), encompassing 2% (3,204/167,140) of ATAC-peaks (permutation FDR < 0.05, **Fig. 2.3e**, **Supplementary Table 2.5**, and **Supplementary Fig. 2.8**). The sequencing coverage of co-accessible peaks is similar to that of all ATAC-peaks (**Supplementary Fig. 2.9a**), but they are individually more likely to overlap $T_{\text{naïve}}$, T_{stim} , and T_{H17} enhancers (**Supplementary Fig. 2.10**) and binding sites for three pioneering factors: NRF, NFY, and STAF (FDR < 0.05, **Supplementary Fig. 2.11**). Pairs of co-accessible peaks were more correlated when both peaks reside in the same contact domain (estimated from Hi-C interactions, **Fig. 2.3f**) and 80% consisted of peaks overlapping pairs of *cis*-regulatory annotations (*e.g.* enhancer/enhancer, enhancer/promoter, super enhancer/promoter; **Fig. 2.3g**). Finally, co-accessible peaks were enriched in annotated T_{stim} super-enhancer regions⁴¹ (**Fig. 2.3h**, **Methods**)^{41,42}. These results suggest that chromatin co-accessibility may be determined by the 3D conformation of the genome and may correspond to coordinated regulation of multiple *cis*-regulatory elements, including known T cell enhancers and regions bound by pioneering factors.

Genetic variants associated with chromatin accessibility

We next defined the genetic basis of chromatin accessibility by associating ATAC-peaks with common SNPs (minor allele frequency > 0.05) across the 105 individuals. To maximize statistical

power, we analyzed only the 64,188 SNP-containing ATAC-peaks (**Fig. 2.3b**) and found 3,318 that were significantly associated with at least one SNP (RASQUAL⁴³, P -value $< 2.91 \times 10^{-3}$, permutation FDR < 0.05) (**Fig. 2.4a**, **Supplementary Fig. 2.12** and **Supplementary Table 2.6**). Each best-associated SNP we term a local ATAC quantitative trait locus (*local*-ATAC-QTL) and the corresponding peak a *local*-ATAC-peak (**Fig. 2.3b**, middle). We estimate that 15% of the 64,188 peaks are associated with at least one *local*-ATAC-QTL using a method to estimate the proportion of null hypotheses while accounting for incomplete power⁴⁴. Sequencing coverage of *local*-ATAC-peaks was similar to all ATAC-peaks (**Supplementary Fig. 2.9b**) and the estimated effects of *local*-ATAC-QTLs are correlated with their effects on H3K27AC ChIP-seq peaks in similar cell types¹² (**Supplementary Fig. 2.13**).

Several lines of evidence support a model where *local*-ATAC-QTLs affect accessibility by perturbing *cis*-regulatory elements active in stimulated T cells. First, for the 1,428/3,318 heritable *local*-ATAC-peaks determined by fitting a linear mixed model over SNPs +/- 500 kb of each peak⁴⁵ (mean $h^2 = 44\%$, GCTA FDR < 0.05), 81% of the heritability is explained by the corresponding *local*-ATAC-QTLs (**Fig. 2.4b** and **Supplementary Tables 2.6** and **2.7**; **Methods**). This suggests a genetic architecture where a single SNP is responsible for the majority of heritable variation. Second, compared to SNP-containing ATAC-peaks, *local*-ATAC-peaks are preferentially located near transcription start and termination sites (**Fig. 2.4c**), are more enriched for T cell enhancers (P -value $< 9.23 \times 10^{-63}$, hypergeometric test; **Supplementary Figure 2.14**), and are more enriched for genomic regions bound by TFs involved in T cell development and activation (*e.g.* BATF, AP-1 and IRF) (**Supplementary Figure 2.15**). Applying deltaSVM^{46,47} to predict the effects of SNPs on TF binding for 903 ATAC-QTLs located within 300 bp of the middle

of the corresponding peaks, we found that almost half (45%) are predicted to strongly disrupt bindings for one of six (BATF, ETS1, IRF, RUNX1, SP1 and CTCF) TF binding sites (**Fig. 2.4d**). The effect sizes of *local*-ATAC-QTLs are correlated with SNP motif disruption scores obtained by deltaSVM⁴⁸ (Pearson R = 0.627, P -value < 2.33×10^{-98} , **Fig. 2.4e; Methods**). For *local*-ATAC-peaks that overlap BATF, ETS1 and CTCF binding sites, differential accessibility between genotypes in the core motifs were observed at single nucleotide resolution, even though only 5% of the corresponding *local*-ATAC-QTLs directly alter the core motif sequences. This suggests that the genetic perturbation of TF binding – either directly by disrupting their sites, or more likely indirectly by first disrupting binding by other factors in the same *cis*-regulatory element – may be a major driver for the observed variation in chromatin accessibility across individuals (**Fig. 2.4f** and **Supplementary Fig. 2.16**). Note that the relation between the accessibility of *local*-ATAC-peaks and 3D chromatin organization is similar to that observed for SNP-containing ATAC-peaks in general (**Fig. 2.4g**). Both *local*-ATAC-peaks and SNP-containing ATAC-peaks overlapping BATF and ETS motifs are enriched within Hi-C contact domains, whereas those overlapping CTCF motifs are enriched at the contact domain boundaries (**Fig. 2.4g**). These results are consistent with previous reports of CTCF enrichment at contact domain boundaries^{40,49-51}.

Local-ATAC-peaks are more likely to overlap GWAS SNPs from autoimmune diseases than other SNP-containing ATAC-peaks (**Supplementary Figure 2.17**), providing a functional context for interpreting disease associations. Even though *local*-ATAC-peaks consist of only ~5% of the SNP-containing ATAC-peaks, they overlap a much larger percentage of the loci associated with autoimmune diseases including Celiac's disease (28%), Crohn's disease (22%), and rheumatoid arthritis (12%), an 8-fold (hypergeometric P -value < 4.34×10^{-7}), 6-fold (hypergeometric P -value

$< 8.58 \times 10^{-17}$), and 5-fold (hypergeometric P -value $< 6.18 \times 10^{-8}$) enrichment, respectively. To corroborate this, we performed partitioned heritability analysis to test for enrichment of *local*-ATAC-QTLs in explaining the heritability of 28 common diseases where summary statistics are available. While *local*-ATAC-QTLs are highly enriched for disease associated variants (i.e. Celiac's disease: 6%, enrichment = 51x and Crohn's disease: 7%, enrichment = 63x), the results are not statistically significant after multiple testing correction. However, by relaxing the FDR thresholding for defining *local*-ATAC-QTLs, we found a general trend of increased proportion of heritability explained and statistical significance, and a decrease in enrichment, especially in autoimmune diseases (**Fig. 2.4h; Methods**). In fact, all SNP-containing ATAC-peaks (corresponding to $FDR < 1$) account for a statistically significant proportion of the heritability for all autoimmune diseases ($> 22\%$, Bonferroni-corrected P -value $< 1.3 \times 10^{-2}$, **Fig. 2.4h; Methods**). For example, rs17293632 (NC_000015.10:g.67442596C>T) has been associated with Crohn's disease and IBD⁵² and is located in the first intron of *SMAD3*, a gene that encodes for a transcription factor involved in the TGF- β signaling pathway that regulates T cell activation and metabolism⁵³. This SNP disrupts a consensus BATF binding site at a conserved position ($\Delta SVM = -12.72$), and results in decreased chromatin accessibility in individuals that carry the alternate allele (**Fig. 2.4i**).

Together, these results suggest that when the accessibility of ATAC-peaks is affected by *local*-ATAC-QTLs residing within peaks, this often involves the disruption of TF binding, even though the SNPs almost always reside outside of the core TF binding site. Moreover, *local*-ATAC-QTLs in stimulated CD4⁺ cells are enriched for autoimmune disease loci, both in the number of overlaps and proportion of heritability explained.

Genetic determinants of chromatin co-accessibility

We next tested if the effect of each *local*-ATAC-QTL could be propagated to co-accessible peaks, for example through 3D chromatin organization, and thus have effects on multiple *cis*-regulatory elements simultaneously. We first estimated the heritability of ATAC-peaks using SNPs +/- 500 kb of each peak. As expected, *local*-ATAC-peaks (2,444/3,318 that converged) were more heritable (mean $h^2 = 0.22$) than all ATAC-peaks (**Fig. 2.5a**). Co-accessible peaks were also more heritable than all ATAC-peaks, both those containing SNPs (mean $h^2 = 0.44$ vs. mean $h^2 = 0.04$) and those that do not (mean $h^2 = 0.10$ vs. mean $h^2 = 0.04$). Excluding the 3,318 *local*-ATAC-peaks, we identified 382 ATAC-peaks that were associated with a *local*-ATAC-QTL (RASQUAL, P -value $< 1.27 \times 10^{-4}$, permutation FDR < 0.05) located +/- 500 kb from the peak. We term each associated SNP a *distal*-ATAC-QTL and each associated peak a *distal*-ATAC-peak (**Fig. 2.2b**). Consistent with the heritability analysis, *distal*-ATAC-QTLs imparted the strongest effects on co-accessible peak (**Fig. 2.5b** and **Supplementary Tables 2.2, 3**).

Co-accessible peaks and co-accessible *distal*-ATAC-peaks are both more likely to overlap Th_{stim} super enhancers than randomly shuffled super enhancers⁴¹. The effect is stronger in co-accessible *distal*-ATAC-peaks (6-fold vs. 4-fold) (**Fig. 2.5c**). In an example, rs10882660 (NC_000010.10:g.97517949A>G) is simultaneously a *local*- and *distal*-ATAC-QTL for a pair of co-accessible peaks residing in the 1st and 2nd introns of ectonucleoside triphosphate diphosphohydrolase I (*ENTPDI*) and a Hi-C contact domain (**Fig. 2.5d**). *ENTPDI* encodes a protein that is one of the dominant drivers of hydrolysis of ATP and ADP in T_{regs} cells, whose expression can lead to tumor growth in mouse models⁵⁴⁻⁵⁷. These results and example suggest a

model where *local*-ATAC-QTLs residing within peaks could also distally affect co-accessible peaks likely reflecting shared genetic effects on pairs of *cis*--regulatory elements.

Linking variation in chromatin state and gene expression

We hypothesized that variants affecting chromatin accessibility (*local*-ATAC-QTLs) would – in some cases – also impact the transcription of the genes controlled through these regulatory regions, and thus provide an important link between variant and target.

To test this hypothesis, we assessed if *local*-ATAC-QTLs are also associated with gene expression in stimulated CD4⁺ T cells, measured by RNA-seq from 95 donors (92 from an aliquot of the same cells with matching ATAC-seq data). After accounting for covariates and principal components for expression heterogeneity (**Supplementary Tables 2.3, 4**), we identified 424 genes significantly associated with at least one of 6,903 *local*-ATAC-QTLs located +/- 500 kb from the center of each gene (RASQUAL, P -value $< 1.65 \times 10^{-3}$, permutation FDR < 0.05 , **Fig. 2.6a**, and **Supplementary Table 2.8**). The 383 best-associated SNPs are eQTLs, and we term the corresponding 424 genes eGenes (**Fig. 2.2b**, right). We estimate that 30% of *local*-ATAC-QTLs are also eQTLs (with a procedure to estimate the proportion of null hypotheses; **Methods**), consistent with previous reports in lymphoblastoid cell lines^{19,21}. Considering all genetic variants located +/- 500 kb from the center of each eGene, we found 191/424 genes to be significantly heritable (GCTA FDR < 0.05), with the eQTL explaining on average 68% of the heritability (**Fig. 2.6b** and **Supplementary Table 2.9**). The lower estimates of explained heritability than *local*-ATAC-peaks suggests that the genetic control of gene expression may involve more than one SNP and *cis*-regulatory element in some cases.

We next examined the sharing of genetic effects between *local*-ATAC-peaks and eGenes using a bivariate linear mixed model⁴⁵ and mediation analysis⁵⁸. Among the 383 SNPs that are simultaneously associated with chromatin accessibility (as *local*-ATAC-QTLs) and gene expression (as eQTLs), 286 have effect sizes in the same direction (Spearman $r = 0.73$) indicative of activating effects, while 138 have effect sizes in the opposite direction indicative of repressive effects (Spearman $r = -0.69$) (**Fig. 2.6c**). Because of limited sample size, measuring the genetic correlation for individual pairs of *local*-ATAC-peaks and eGenes is likely under powered. However, the distribution of genetic correlations for 161 pairs of *local*-ATAC-peaks and eGenes that converged (inverse variance weighted average of 0.66) was significantly higher than both randomly sampled (inverse variance weighted average of 0.23, Kolmogorov-Smirnov P -value $< 4.32 \times 10^{-10}$) and permuted ATAC-peaks (inverse variance weighted average of 0.07, Kolmogorov-Smirnov P -value $< 1.68 \times 10^{-10}$) (**Supplementary Fig. 2.18** and **Supplementary Table 2.10**). This is corroborated by mediation analysis where the genetic effects on 21/424 eGenes were significantly mediated by the corresponding *local*-ATAC-peaks (FDR < 0.1 , **Fig. 2.6d**) and the high correlation of the mediation effects and the inverse variance weighted genetic correlation (Pearson $R = 0.52$, P -value $< 1.2 \times 10^{-12}$, **Supplementary Fig. 2.19**). For example, consider the locus spanning *FADS1* and *FADS2*, genes that encode two fatty acid desaturases (FADS) that regulate inflammation, promote cancer development, and impact dermal and intestinal ulcerations (in *FADS2* knockout mice)⁵⁹⁻⁶². Before conditioning on rs174575 (NC_000011.10:g.61602003C>G), an eQTL for *FADS2* and a *local*-ATAC-QTL for chr11:61,601,708-61,602,451, *FADS2* expression and accessibility of ATAC-peak chr11:61,601,708-61,602,451 are correlated ($R^2 = 0.31$, P -value $< 8.25 \times 10^{-9}$) and after

conditioning there is no longer a correlation ($R^2 = 0.08$, P -value $< 6.1 \times 10^{-3}$) (**Fig. 2.6e**). Similarly, after conditioning on rs174561 (NC_000011.10:g.61582708T>C), *FADS1* expression is no longer correlated with accessibility of ATAC-peak chr11:61,582,207-61,584,717 (before conditioning: $R^2 = 0.2$, P -value $< 8.74 \times 10^{-6}$; after conditioning: $R^2 = 0.01$, P -value < 0.3) (**Fig. 2.6f**). Notably, rs174561 is an eQTL for *FADS1*, a *local*-ATAC-QTL associated with a pair of co-accessible peaks and has been previously associated with Crohn's disease. It is also in LD with rs174537 (NC_000011.10:g.61552680T>G, $r^2 = 0.82$, $D' = 0.99$), a SNP previously identified as an eQTL in blood^{63,64}. The associated co-accessible peaks span the promoters of *FADS1* and *FADS2* (**Fig. 2.6g, h**). These results suggest that 30% of the time, genetic variants associated with chromatin accessibility are also associated with gene expression, and in some cases, such as the *FADS1* and *FADS2* loci previously associated with Crohn's disease, can be directionally linked through mediation analysis.

Discussion

Although variability in gene expression has been extensively characterized, variability in chromatin state has been challenging to study in primary cells. To this end, we analyzed ATAC-seq profiles in primary CD4⁺ T cells from five individuals at rest and in response to stimulation. We found global remodeling of accessible chromatin after stimulation, with a significantly higher number of accessible regions overlapping a large proportion (12 - 28%) of SNPs associated with autoimmune diseases and different T cell enhancer subsets (*e.g.* T_{regs}, Th₁₇, etc).

Due to these initial observations, we dissected the relationship between genetic variation and variability in chromatin accessibility in a physiologically-relevant system, stimulated CD4⁺ cells. Variation across 105 individuals highlights four inter-related phenomena. **First**, accessible regions

co-vary across the genome of an individual (co-accessibility), reflective of the 3D structure of the genome. At individual peak resolution, ~2% of ATAC-peaks are co-accessible, especially if they are within the same Hi-C contact domain, and these are more likely to overlap T cell enhancers, pioneering factors, and “pairs” of regulatory regions, including super-enhancers. These results suggest that co-accessibility between pairs of peaks may be determined by the 3D conformation of the genome and may correspond to coordinated regulation of multiple *cis*-regulatory elements. **Second**, combining genetic variation with variation in individual peak accessibility, we identified *local*-ATAC-QTLs. Even though only a minority (5%) of *local*-ATAC-QTLs directly reside within the core binding sites of TFs, nearly half (45%) are predicted to dramatically disrupt binding at TF binding sites. Moreover, even though *local*-ATAC-peaks are only 5% of SNP-containing ATAC-peaks, they overlap ~10-30% of the previously reported loci for several common autoimmune diseases and explain 1-7% of the disease heritability. The overwhelming enrichment for autoimmune disease loci among *local*-ATAC-peaks could be the result of both the increased number of features tracking cell state and the propensity for disease-causing variants to perturb *cis*-regulatory elements containing key TFs active in specific cell types or states. **Third**, we found that *local*-ATAC-QTLs can further act *distally* on additional peaks in a 1 Mb window, with the strongest effects on ATAC-peaks that are co-accessible, which substantially increase their mechanistic and functional impact. **Fourth**, considering *local*-ATAC-QTLs in the context of variation in gene expression (by RNA-seq; 92 overlapping individuals), we estimated that 30% of *local*-ATAC-QTLs are also eQTLs, with bivariate and mediation analyses suggesting there may be mechanistic directionality between these functional phenotypes.

In a manner consistent with known modes of transcriptional regulation, our approach for a staged

analysis, testing the effects of *local*-ATAC-QTLs on *distal*-ATAC-peaks and gene expression, allowed us to overcome power limitations from the sample size and the technical and biological variability in the assays to detect hundreds of genes associated with *local*-ATAC-QTLs. Despite this, there was limited power for bivariate analysis to quantify the shared genetic effects and establish causality for the observed association to both chromatin state and gene expression. These limitations will likely be overcome in future studies with larger sample sizes and higher sequencing depth.

Our findings, derived from large scale genetic association of quantitative chromatin and gene expression traits in primary human cells implicated in many diseases, provide a molecular framework for how disease-causing variants could alter local chromatin structure to modulate gene expression. With the recent advancement of single cell epigenomic⁶⁵ and transcriptomic^{66,67,68} profiling, it should be possible to more directly detect context-specific genetic effects in a heterogeneous cell population. Future studies that use other disease-relevant primary cells and tissues will help pinpoint causal disease variants and understand the regulatory mechanism underlying common disease.

References

1. McCarthy, M.I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9, 356-69 (2008).
2. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* 90, 7-24 (2012).
3. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6, 95-108 (2005).
4. Maurano, M.T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190-5 (2012).
5. Stranger, B.E. et al. Population genomics of human gene expression. *Nat Genet* 39, 1217-24 (2007).
6. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506-11 (2013).
7. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24, 14-24 (2014).
8. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* 344, 519-23 (2014).
9. Lee, M.N. et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343, 1246980 (2014).
10. Ye, C.J. et al. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* 345, 1254665 (2014).
11. Astle, W.J. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 167, 1415-1429 e19 (2016).
12. Chen, L. et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human

Immune Cells. *Cell* 167, 1398-1414 e24 (2016).

13. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74 (2012).

14. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91-100 (2012).

15. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83-90 (2012).

16. Thurman, R.E. et al. The accessible chromatin landscape of the human genome. *Nature* 489, 75-82 (2012).

17. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317-30 (2015).

18. Farh, K.K. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337-43 (2015).

19. Degner, J.F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390-4 (2012).

20. Kasowski, M. et al. Extensive variation in chromatin states across humans. *Science* 342, 750-2 (2013).

21. McVicker, G. et al. Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747-9 (2013).

22. Kilpinen, H. et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 342, 744-7 (2013).

23. Waszak, S.M. et al. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell* 162, 1039-50 (2015).

24. Elinav, E. et al. Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nat Rev Cancer* 13, 759-71 (2013).
25. Donath, M.Y. & Shoelson, S.E. Type 2 diabetes as an inflammatory disease. *Nat Rev Immunol* 11, 98-107 (2011).
26. Ohashi, P.S. T-cell signalling and autoimmunity: molecular mechanisms of disease. *Nat Rev Immunol* 2, 427-38 (2002).
27. Kronenberg, M. & Rudensky, A. Regulation of immunity by self-reactive T cells. *Nature* 435, 598-604 (2005).
28. Speiser, D.E., Ho, P.C. & Verdeil, G. Regulatory circuits of T cell function in cancer. *Nat Rev Immunol* 16, 599-611 (2016).
29. Restifo, N.P., Dudley, M.E. & Rosenberg, S.A. Adoptive immunotherapy for cancer: harnessing the T cell response. *Nat Rev Immunol* 12, 269-81 (2012).
30. Belkaid, Y. & Rouse, B.T. Natural regulatory T cells in infectious disease. *Nat Immunol* 6, 353-60 (2005).
31. Feuerer, M., Hill, J.A., Mathis, D. & Benoist, C. Foxp3⁺ regulatory T cells: differentiation, specification, subphenotypes. *Nat Immunol* 10, 689-95 (2009).
32. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 1213-8 (2013).
33. Kurachi, M. et al. The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8⁺ T cells. *Nat Immunol* 15, 373-83 (2014).
34. Li, P. et al. BATF-JUN is critical for IRF4-mediated transcription in T cells. *Nature* 490, 543-6 (2012).

35. Murphy, T.L., Tussiwand, R. & Murphy, K.M. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nat Rev Immunol* 13, 499-509 (2013).
36. Cauchy, P. et al. Dynamic recruitment of Ets1 to both nucleosome-occupied and -depleted enhancer regions mediates a transcriptional program switch during early T-cell differentiation. *Nucleic Acids Res* 44, 3567-85 (2016).
37. Samstein, R.M. et al. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* 151, 153-66 (2012).
38. Hollenhorst, P.C. et al. DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet* 5, e1000778 (2009).
39. Chen, X. et al. ATAC-seq reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat Methods* 13, 1013-1020 (2016).
40. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665-80 (2014).
41. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* 155, 934-47 (2013).
42. Whyte, W.A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153, 307-19 (2013).
43. Kumasaka, N., Knights, A.J. & Gaffney, D.J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* 48, 206-13 (2016).
44. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100, 9440-5 (2003).
45. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88, 76-82 (2011).

46. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M.A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* 10, e1003711 (2014).
47. Ghandi M, M.-N.M., Ghareghani N, Lee D, Garraway L, Beer MA. gkmSVM, an R package for gapped-kmer SVM. *Bioinformatics*. Apr 19(2016).
48. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* 47, 955-61 (2015).
49. Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc Natl Acad Sci U S A* 105, 20398-403 (2008).
50. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* 137, 1194-211 (2009).
51. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* 20, 2349-54 (2006).
52. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 42, 1118-25 (2010).
53. Delisle, J.S. et al. The TGF-beta-Smad3 pathway inhibits CD28-dependent cell growth and proliferation of CD4 T cells. *Genes Immun* 14, 115-26 (2013).
54. Enjyoji, K. et al. Targeted disruption of cd39/ATP diphosphohydrolase results in disordered hemostasis and thromboregulation. *Nat Med* 5, 1010-7 (1999).
55. Deaglio, S. et al. Adenosine generation catalyzed by CD39 and CD73 expressed on regulatory T cells mediates immune suppression. *J Exp Med* 204, 1257-65 (2007).
56. Plesner, L. Ecto-ATPases: identities and functions. *Int Rev Cytol* 158, 141-214 (1995).
57. Sun, X. et al. CD39/ENTPD1 expression by CD4⁺Foxp3⁺ regulatory T cells promotes hepatic metastatic tumor growth in mice. *Gastroenterology* 139, 1030-40 (2010).

58. Hicks, R., & Tingley, D. Causal Mediation Analysis. Vol. 4 (Stata Journal 2011).
59. Fan, Y.Y. et al. Characterization of an arachidonic acid-deficient (Fads1 knockout) mouse model. *J Lipid Res* 53, 1287-95 (2012).
60. Barrie, A. et al. Prostaglandin E2 and IL-23 plus IL-1beta differentially regulate the Th1/Th17 immune response of human CD161(+) CD4(+) memory T cells. *Clin Transl Sci* 4, 268-73 (2011).
61. Sakata, D., Yao, C. & Narumiya, S. Prostaglandin E2, an immunoactivator. *J Pharmacol Sci* 112, 1-5 (2010).
62. Stroud, C.K. et al. Disruption of FADS2 gene in mice impairs male reproduction and causes dermal and intestinal ulceration. *J Lipid Res* 50, 1870-80 (2009).
63. Schmidt, E.M. et al. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* 31, 2601-6 (2015).
64. Marigorta, U.M. et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat Genet* 49, 1517-1521 (2017).
65. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486-90 (2015).
66. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187-201 (2015).
67. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202-14 (2015).
68. Kang, H.M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* 36, 89-94 (2018).

Figures

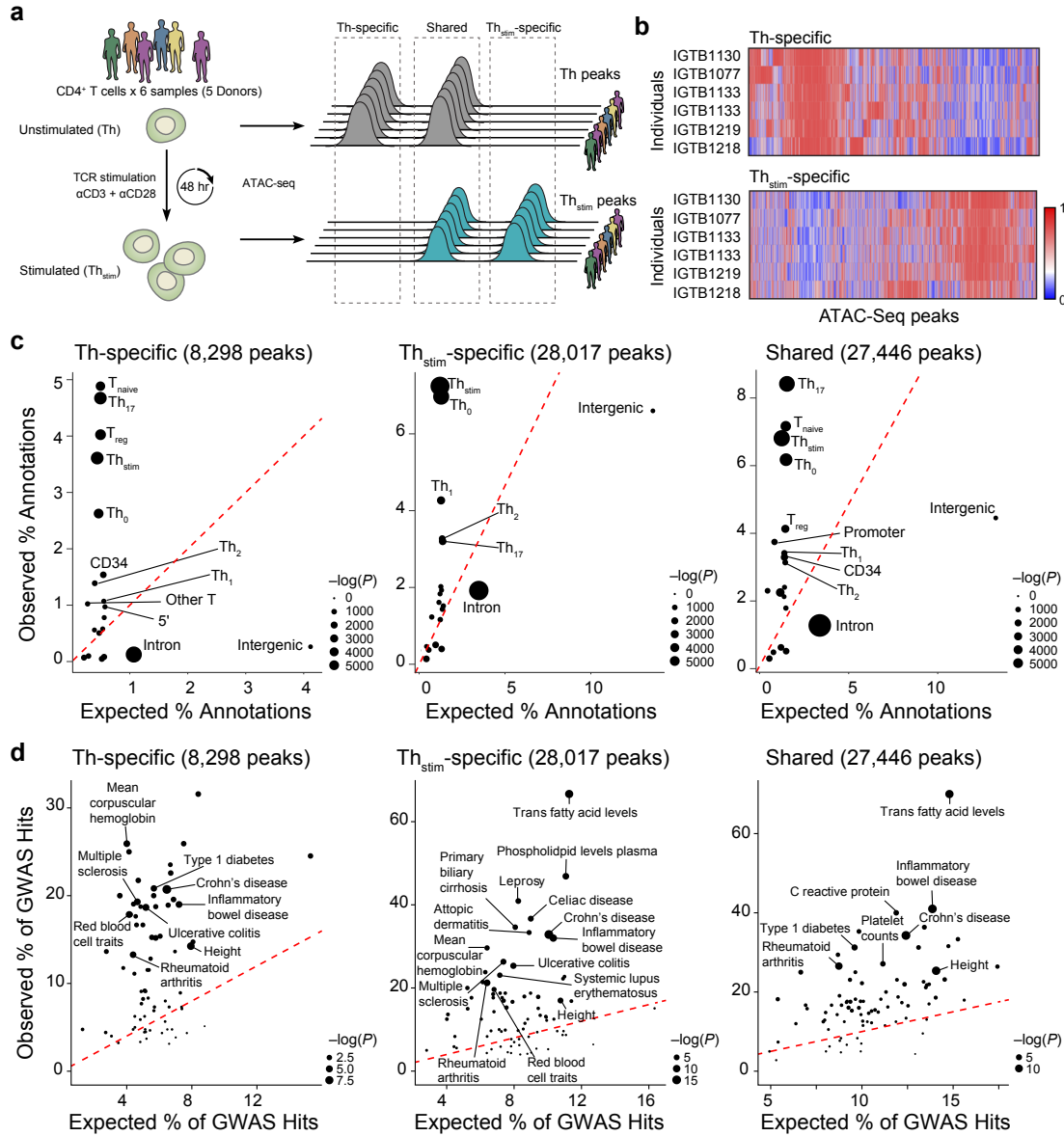


Figure 2.1. Changes in chromatin state in human T cell activation.

(a) Experimental overview (left) and schematic of nomenclature (right). (b) Differential chromatin accessibility. Regions of open chromatin (columns) in six samples (rows) before (top, Th-specific) and 48hr after (bottom, Th_{stim}-specific) activation of primary T cells with anti-CD3/CD28 antibodies. (c) Overlap with previously annotated T cell enhancers. For each annotation, expected (x-axis) vs. observed (y-axis) percentages of annotated features overlapping Th-specific (left), Th_{stim}-specific (center) and shared peaks (right). (d) Overlap with GWAS variants. For each phenotype or disease, expected (x-axis) vs. observed (y-axis) percentages of GWAS loci overlapping Th-specific (left), Th_{stim}-specific (center), or shared (right) peaks.

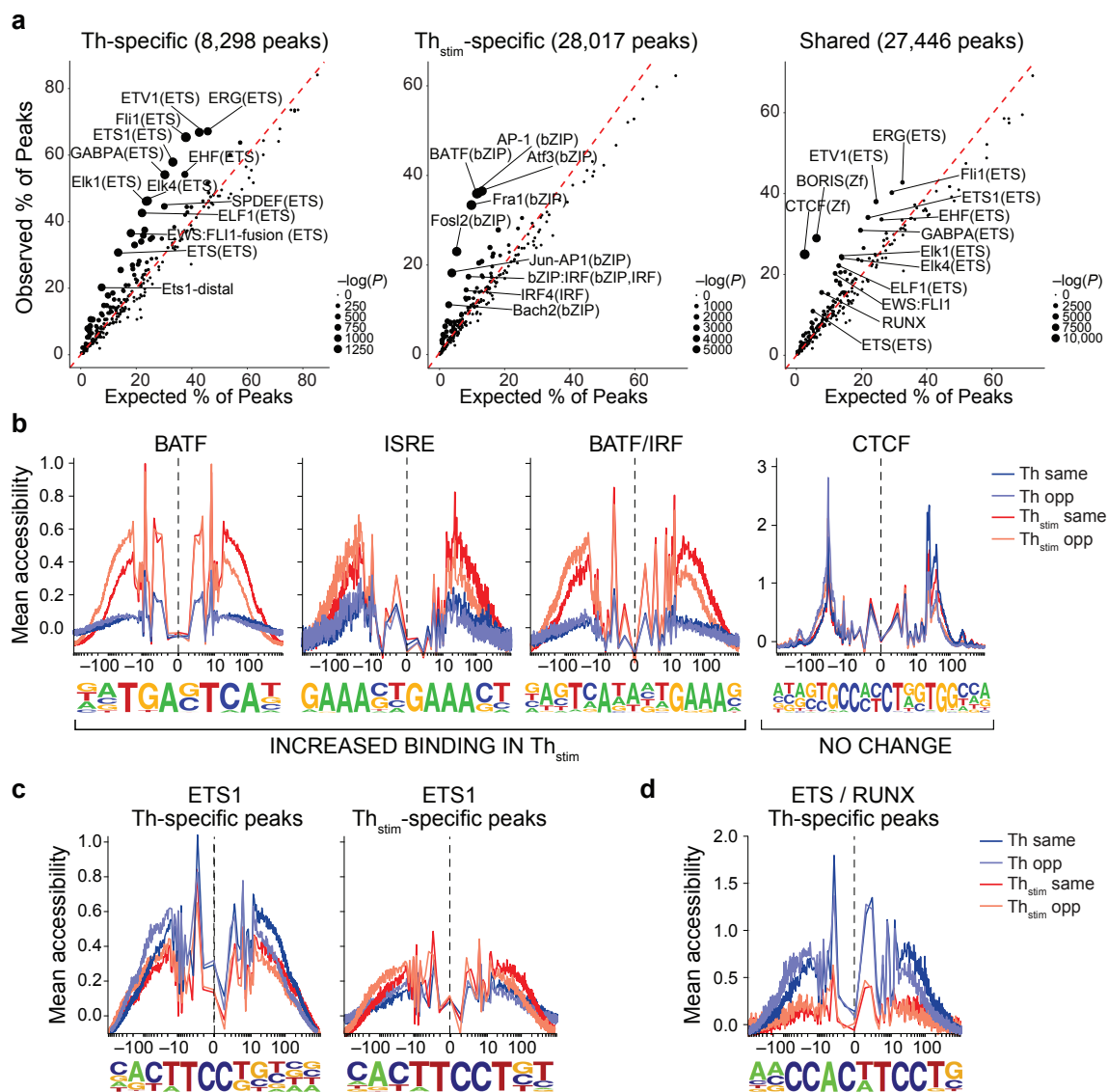


Figure 2.2. Changes in transcription factor enrichment in response to T cell activation.

(a) Transcription factor motif enrichment. Expected (x-axis) vs. observed (y-axis) percentages of Th-specific (left), Th_{stim}-specific (center), or shared (right) peaks overlapping each TF binding site annotation. (b-d) TF footprinting. For each TF motif (as defined in ENCODE⁶³), nucleotide resolution average chromatin accessibility (y-axis) in Th (purple) or Th_{stim} (red) cells along the TF binding site (x-axis; log(bp from center of each TF motif)). Aggregated locations are defined as (b) Th_{stim}-specific peaks overlapping BATF, ISRE, and BATF/IRF motifs (three left panels) and shared peaks overlapping CTCF binding sites (right panel), (c) Th-specific (left) and Th_{stim}-specific (right) peaks overlapping ETS1 binding sites, and (d) Th-specific peaks overlapping ETS1/RUNX combinatorial binding sites.

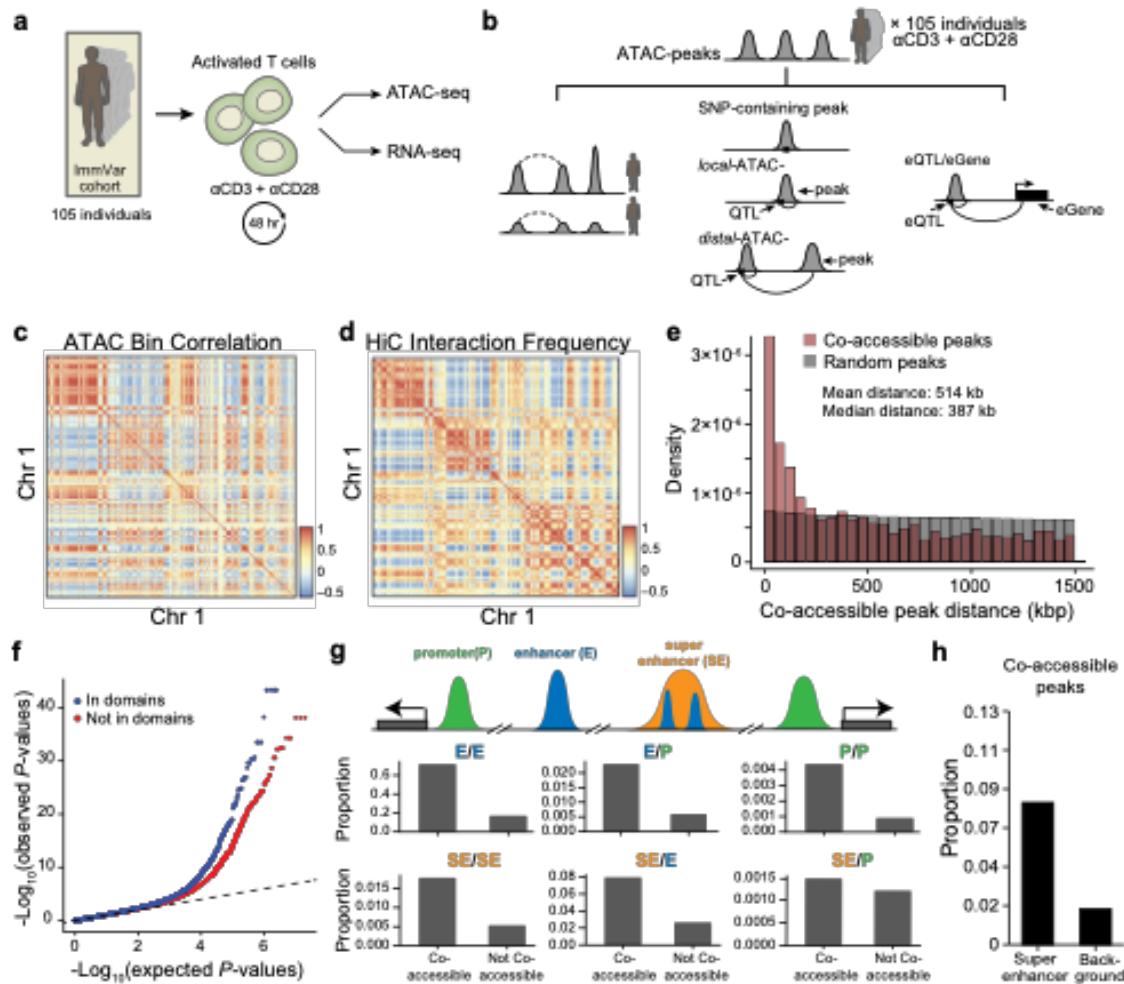


Figure 2.3. Inter-individual chromatin co-accessibility.

(a) Overview of T cell activation for 105 ATAC-seq and 95 RNA-seq samples. (b) Schematic of nomenclature for co-accessible peaks, SNP-containing peaks, *local*-ATAC-QTLs, *distal*-ATAC-QTLs, and eQTLs. Dashed lines denote a correlation between co-accessible peaks and solid lines denote a genetic association. (c) Megabase scale correlation of chromatin accessibility across 105 individuals. Heat map shows the pairwise Pearson correlation of chromatin accessibility between 1 Mb bins (row, column) for Chr 1. (d) Pearson correlation of Hi-C interactions at 1 Mb resolution for Chr 1. (e) Histogram of distances between significantly co-accessible peaks (pink) and random permuted peaks (grey). (f) Co-accessible peaks overlap with Hi-C domains. Q-Q plot of linear regression P -values for pairs of peaks residing in (blue) or out (red) of the same Hi-C domain. (g) Pairs of co-accessible peaks overlapping with multiple *cis*-regulatory regions. A cartoon depiction (top) of co-accessible peaks in promoters (green), enhancers (blue), and super enhancers (orange). Proportion (y-axis) of pairs of co-accessible peaks and non-co-accessible peaks overlapping pairs of annotated *cis*-regulatory elements (right). (h) Proportion of co-accessible peaks overlapping super-enhancers or randomly shuffled background.

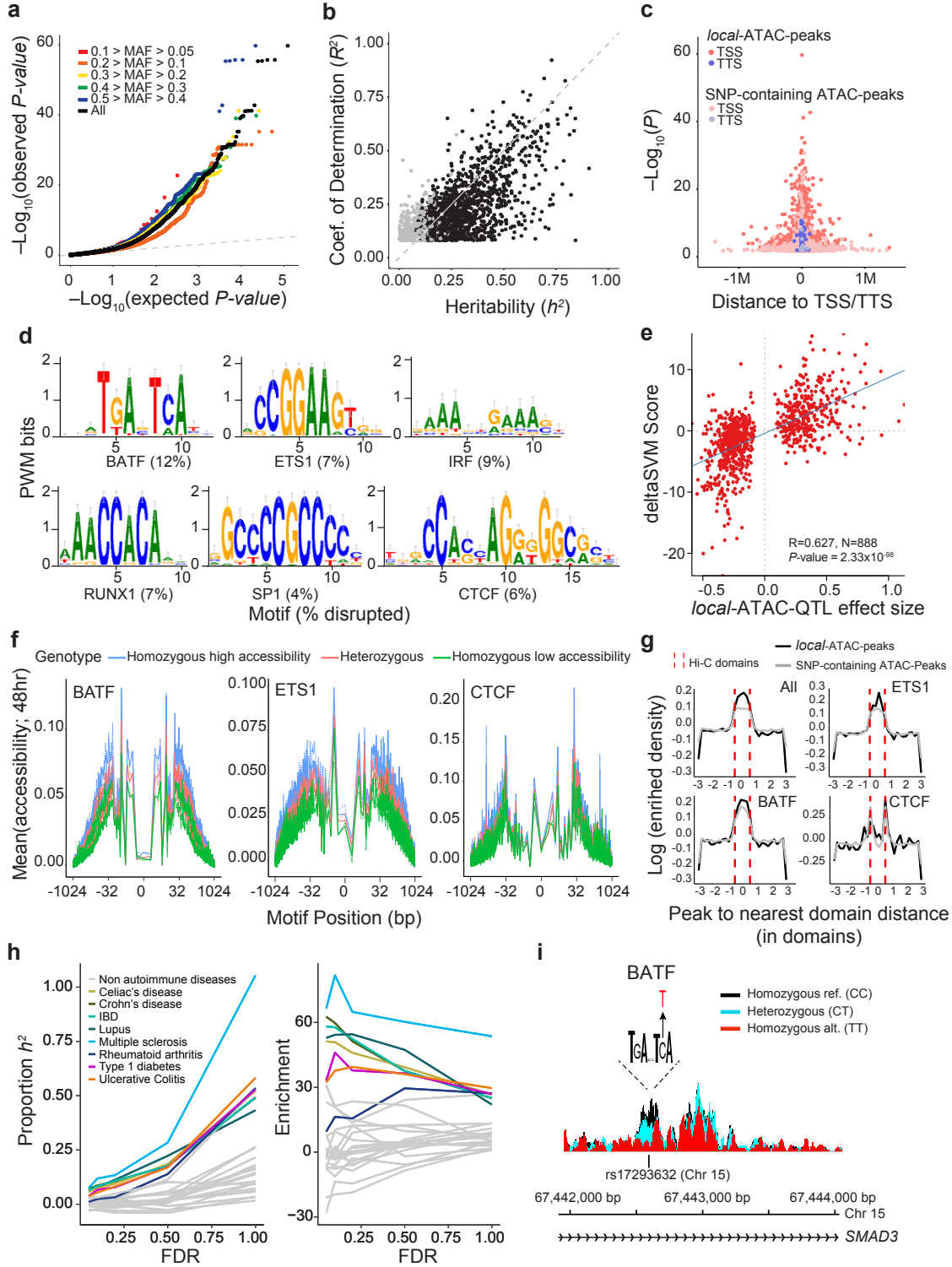


Figure 2.4. Genetic variants that affect chromatin states in human T cell activation.

(a) Q-Q plot of linear regression P -values for all *local*-ATAC-QTLs (black) and *local*-ATAC-QTLs binned by minor allele frequency: $0.1 > \text{MAF} > 0.05$ (red), $0.2 > \text{MAF} > 0.1$ (orange), $0.3 > \text{MAF} > 0.2$ (yellow), $0.4 > \text{MAF} > 0.3$ (green), $0.5 > \text{MAF} > 0.4$ (blue), and $\text{MAF} > 0.05$ (black). (b) Heritability of chromatin accessibility. For each of 1,428 *local*-ATAC-peaks, coefficient of determination (R^2) of the best associated *local*-ATAC-QTL (y-axis) vs. *cis* heritability (h^2) estimated based on all genotypes +/- 500 kb of each peak (x-axis). Black points: significantly heritable peaks ($\text{FDR} < 0.05$). (c) Enrichment of *local*-ATAC-peaks in TSS and TTS. 3,318 *local*-ATAC-peaks (dark pink and purple) vs. 3,318 randomly sampled SNP-containing ATAC-peaks (light pink and purple). (d-f) Disruption of TF binding sites by *local*-ATAC-QTLs. (d) Unsupervised TF binding site analysis of *local*-ATAC-peaks. Motifs for six TFs associated with most of the large gkmSVM weights, and the percentage of the overall disruption (%; bottom) explained by *local*-ATAC-QTLs. (e) Correlation of effect sizes of *local*-ATAC-QTLs (x-axis) vs. deltaSVM scores (y-axis). (f) Allele specificity of *local*-ATAC-QTLs. For BATF, ETS1 and CTCF motifs (as identified in ENCODE⁶³), aggregated plots of mean chromatin accessibility (y-axis) of *local*-ATAC-peaks along the TF binding site (x-axis; log(bp from center of the TF motif)) for samples heterozygous (pink), homozygous for the high (blue) or low (green) *local*-ATAC-QTL alleles (g) Relation between contact domains and SNP-containing ATAC-peaks or *local*-ATAC-peaks. For ATAC-peaks or *local*-ATAC-peaks overlapping ETS1, CTCF, or BATF binding sites, enrichment density (y-axis) vs. distance (number of domains) of peak to nearest domain (x-axis). Hi-C contact domain boundaries are indicated (dotted red lines). (h) Partitioned heritability estimates. The proportion of the heritability for 28 diseases explained (proportion: left, y-axis; enrichment: right, y-axis) captured by *local*-ATAC-QTLs called at different FDR thresholds (x-axis). (i) Effects of *local*-ATAC-QTL rs17293632 on the accessibility of the corresponding BATF containing *local*-ATAC-peak on chromosome 15. ATAC-seq profiles were aggregated per rs17293632 genotypes (black: homozygous major allele, light blue: heterozygous, red: homozygous minor allele).

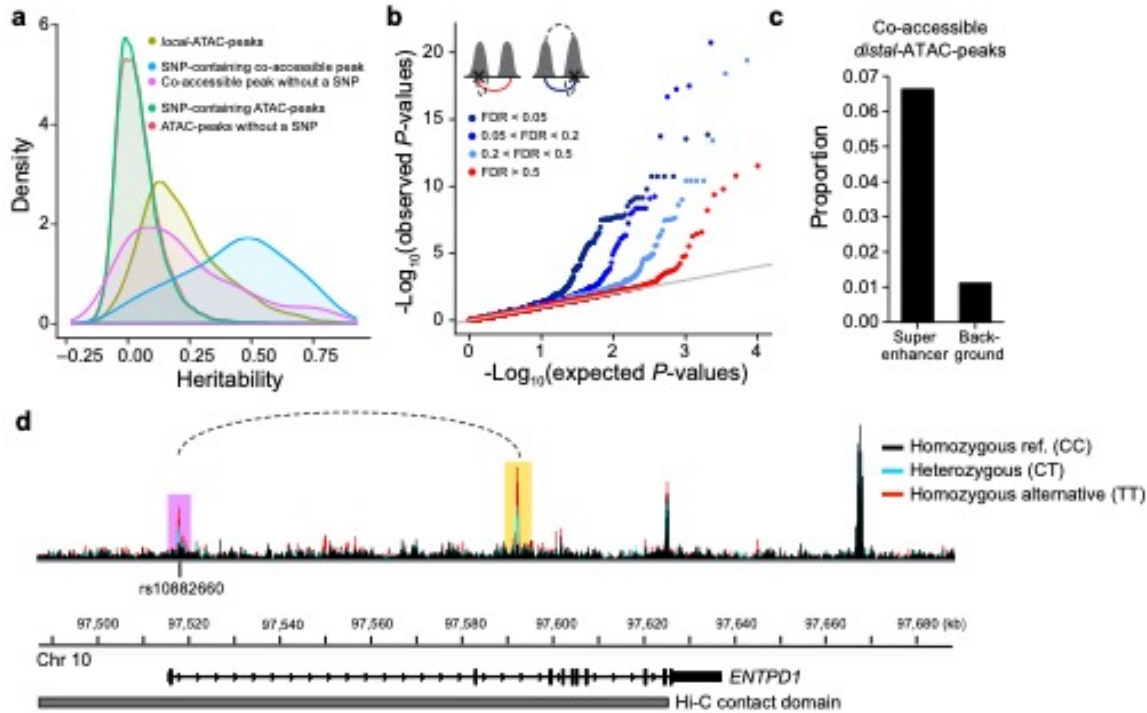


Figure 2.5. Genetic determinants of co-accessible peaks.

(a) Distribution of the heritability explained by SNPs +/- 500 kb of ATAC-peaks. *Local*-ATAC-peaks (olive). SNP-containing co-accessible peaks (blue). Co-accessible peaks without a SNP (purple). SNP-containing ATAC-peaks (green). ATAC-peaks without a SNP (red). (b) Q-Q plots of the linear regression P -values of *distal*-ATAC-peaks that are single peaks (red: co-accessibility FDR > 0.5), or co-accessible peaks called at various significance cutoffs (light blue: $0.2 < \text{FDR} < 0.5$, medium blue: $0.05 < \text{FDR} < 0.2$, dark blue: $\text{FDR} < 0.05$). The cartoons (upper left corner) depict the *distal*-ATAC-QTL association for single peaks (left cartoon; red line is the association plotted) and *distal*-ATAC-QTL association for co-accessible peaks (right cartoon; blue line is the association plotted; upper dashed line is the co-accessible peak at various significance cutoffs). (c) Proportion of co-accessible *distal*-ATAC-peaks overlapping super-enhancers or randomly shuffled background. (d) An example of a genetic variant (rs10882660) residing in the first intron in *ENTPD1*, associated *locally* (in purple) and *distally* (in yellow) to ATAC-peaks. The *local* and *distal*-ATAC-peaks are co-accessible (dotted line) and reside in a Hi-C contact domain (grey). ATAC-seq profiles were aggregated for individuals of different rs10882660 genotypes (black: homozygous major allele, light blue: heterozygous, red: homozygous minor allele).

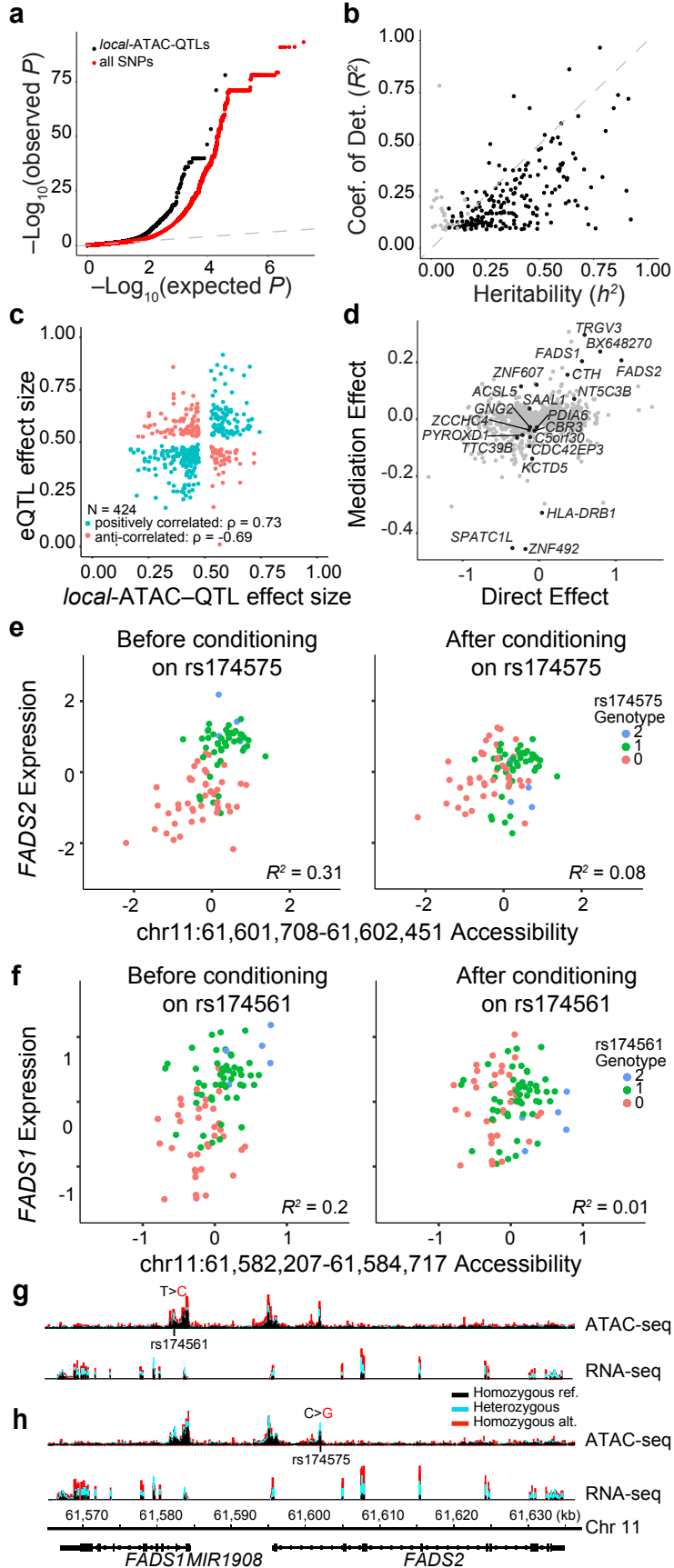


Figure 2.6. Association of chromatin accessibility and gene expression.

(a) eQTLs. Q-Q plot of associations between *local*-ATAC-QTLs (black) or all SNPs (red) and expression of genes +/- 500kb. (b) Heritability of gene expression. For each of 191 eGenes, coefficient of determination (R^2) of the best associated eQTL (y-axis) vs. heritability (h^2) of all variants +/- 500 kb of each gene (x-axis). Black points: significantly heritable peaks (FDR < 0.05). (c) Correlation of effect sizes between *local*-ATAC-QTLs (x-axis) and eQTLs (y-axis). (d) Mediation of eGenes. Average causal mediation effect estimates (y-axis) and average direct effect estimates (x-axis) for *local*-ATAC-peaks (mediator) and eGenes (outcome variable) sharing a SNP (instrument variable). FDR < 0.1 *local*-ATAC-peaks are colored in black. (e,f) Examples of gene expression conditioned on chromatin accessibility. (e) *FADS2* expression (y-axis) vs. chromatin accessibility at chr11:61,601,708-61,602,451 (x-axis) before (left) and after (right) conditioning on rs174575. (f) *FADS1* expression (y-axis) vs. chromatin accessibility at chr11:61,582,207-61,584,717 (x-axis) before (left) and after (right) conditioning, colored by rs174561 genotypes. (g,h) ATAC-seq (top) and RNA-seq (bottom) profiles were aggregated for individuals of different (g) rs174561 and (h) rs174575 genotypes (black: homozygous major allele, light blue: heterozygous, red: homozygous minor allele).

Materials and Methods

Study subjects and genotyping

Healthy subjects between the ages of 18 to 56 (avg. 29.9) enrolled in the PhenoGenetic study⁸ were recruited from the Greater Boston Area and gave written informed consent for the studies. Individuals were excluded if they had a history of inflammatory disease, autoimmune disease, chronic metabolic disorders or chronic infectious disorders. Genotyping using the Illumina Infinium Human OmniExpress Exome BeadChips (704,808 SNPs are common variants [MAF > 0.01] and 246,229 are part of the exomes; Illumina Inc., San Diego, CA) has been previously described¹⁸. The genotyping success rate was at least 97%. We applied rigorous subject and SNP quality control (QC) that includes: (1) gender misidentification; (2) subject relatedness; (3) Hardy-Weinberg Equilibrium testing; (4) use concordance to infer SNP quality; (5) genotype call rate; (6) heterozygosity outlier; and (7) mismatch detection using SNP overlapping reads from ATAC-seq and RNA-seq. We excluded 1,987 SNPs with a call rate < 95%, 459 SNPs with Hardy-Weinberg equilibrium P -value < 10^{-6} , and 63,781 SNPs with MAF < 1% from the 704,808 common SNPs (a total of 66,461 SNPs excluded). Principal component analysis of genotypes from all individuals used in the study are shown in **Supplementary Figure S2.6**.

We used the IMPUTE2 software (version: 2.3.2) to impute the post-QC genotyped markers from the entire ImmVar cohort (N = 688) using reference haplotype panels from the 1000 Genomes Project (The 1000 Genomes Project Consortium Phase III) that contain a total of 37.9 Million SNPs in 2,504 individuals with ancestries from West Africa, East Asia, and Europe. After

genotype imputation, we extracted the genotypes for 105 individuals assayed for chromatin accessibility and gene expression. Additional removal of SNPs with MAF < 0.05 in our cohort resulted in 4,558,693 and 4,421,936 common variants tested for chromatin accessibility and gene expression assays, respectively.

Preparation and activation of primary human CD4⁺ T cells

CD4⁺ T cells were isolated and stimulated as previously described¹⁰. Briefly, CD4⁺ T cells were isolated from whole blood by negative selection using RosetteSep human CD4⁺ T cell enrichment cocktail (STEMCELL Technologies Inc., Vancouver, BC) and RosetteSep density medium gradient centrifugation. Isolated CD4⁺ T cells were placed in freezing container at -80°C for overnight, and then moved into a liquid nitrogen tank for long-term storage. On the day of activation, CD4⁺ T cells were thawed in a 37°C water bath, counted and resuspended in RPMI-1640 supplemented with 10% FCS, and plated at 50,000 cells per well in a 96 well round-bottom plate. Cells were either left untreated or stimulated with beads conjugated with anti-CD3 and anti-CD28 antibodies (Dynabeads, Invitrogen #11131D, Life Technologies) at a cell:bead ratio of 1:1 for 48 hours, a time point we previously found to maximize the gene expression response in CD4⁺ T cells. At each time point, cells were further purified by a second step positive selection with CD4⁺ Dynabeads (Invitrogen #11145D, Life Technologies).

ATAC-seq profiling

ATAC-seq profiles were collected for 139 individuals (**Supplementary Table 2.4**). We performed ATAC-seq as previously described³², with a modification in the lysis buffer to reduce mitochondrial DNA contamination. 200,000 purified CD4⁺ T cells were lysed with cold lysis

buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl₂ and 0.03% tween20). Immediately after lysis, nuclei were spun at 500g for 8 minutes at 4°C. After pelleting the nuclei, we carefully removed the supernatant and resuspended the nuclei with Tn5 transposase reaction mix (25 ul 2X TD buffer, 2.5 ul Tn5 transposase, and 22.5 ul nuclease-free water) (Illumina Inc). The transposition reaction was performed at 37°C for 30 minutes. Immediately after the transposition reaction, DNA was purified using a Qiagen MinElute kit. Libraries were sequenced on an Illumina HiSeq 2500 sequencer to an average read depth of 42 million (+/- 38 million) per sample (**Supplementary Fig. 2.S2**), with low mtDNA contamination (0.30 – 5.39%, 1.96% on average), low rates of multiply mapped reads (6.7 – 56%, 19% on average) and a relatively high percentage of usable nuclear reads (60 – 92%, 79% on average).

RNA-seq profiling

RNA-seq profiles were collected for 95 individuals, of which 92 have matching ATAC-seq profiles (**Supplementary Table 2.4**). RNA was isolated using Qiagen RNeasy Plus Mini Kit and RNA integrity was quantified by Agilent RNA 6000 Nano Kit using the Agilent Bioanalyzer. Purified RNA were converted to RNA-seq libraries using a previously published protocol⁶⁹, where reverse transcription was carried out based on the SMART template switching method and the resulting cDNA was further tagmented and PCR amplified using Nextera XT DNA Sample kit (Illumina) to add the Illumina sequencing adaptors. Samples were sequenced on Illumina HiSeq 2500 to an average depth of 16.9 million reads per sample (+/- 8.7 million).

***In situ* Hi-C**

CD4⁺ T cells were isolated from commercially available fresh blood of healthy individuals

(Research Blood Components). CD4⁺ T cells were stimulated for 48 hours with beads conjugated with anti-CD3 and anti-CD28 antibodies. *In situ* Hi-C was performed on the pool of donors as previously described⁴⁰. Cells were crosslinked with 1% formaldehyde for 10 min at room temperature. After nuclei permeabilization, DNA was digested with MboI and digested fragments were labeled using biotinylated d-ATP and ligated. After reverse crosslinking, ligated DNA was purified and sheared to ~400 bp. Biotin labeled DNA fragments were then pulled down with streptavidin beads and prepped for Illumina sequencing⁴⁰. The final libraries were sequenced using Illumina HiSeq and NextSeq to produce ~3.5 billion 100bp paired-end reads.

Alignment of ATAC-seq reads

25bp ATAC-seq reads were aligned to the human genome assembly (hg19) with the Burrows-Wheeler Aligner-MEM (version: 0.7.12)⁷⁰. For each sample, multiply-mapped reads were filtered using Samtools “view”⁷¹ with option “-F 4” and mitochondrial reads were filtered out using BEDtools (function intersectBed)⁷². After filtering, we had a median of 37 million (MAD +/- 13 million) reads per sample.

ATAC-seq peak identification

Filtered ATAC-seq reads from six matched samples (five individuals, of which one individual was repeated) for Th and Th_{stim} cells were merged (separately for Th and Th_{stim} cells) using the Samtools function “merge”⁷¹. Peaks were called on the respective Th and Th_{stim} merged bam files using MACS2 –callpeak (with parameters: --nomodel, --extsize 200, and --shift 100), resulting in 36,486 Th peaks with an average width of 520 bp (+/- 319 bp) and 52,154 Th_{stim} peaks with an average width of 483 bp (+/- 344 bp) (Benjamini-Hochberg FDR < 0.05)⁷³. The Th and Th_{stim}

peaks were further merged (using the BEDtools “merge” function), to a total of 63,763 jointly called peaks. BEDtools “coverage”⁷² was used to create a 63,763 (peaks) x 12 (6 samples by 2 conditions) input matrix used for detecting differentially accessible peaks. Differentially accessible peaks between Th and Th_{stim} cells were identified using the DESeq2 R package (version 3.2)⁷⁴, with 8,298 Th-specific peaks (FDR < 0.05, more accessibility in Th cells), 28,017 Th_{stim}-specific peaks (FDR < 0.05, more accessibility in Th_{stim}), and 27,446 shared peaks (FDR > 0.05).

For the co-accessibility and genetic analyses, 4.2 billion filtered ATAC-seq reads from 105 Th_{stim} samples were merged to call 167,140 peaks (FDR < 0.05) using the same parameters as previously described, at an average peak size of 642 bp (+/- 512 bp). Coverage for each peak over all 105 samples was computed.

Percentage of peaks overlapping transcription factor binding motifs

Percentages of MACS2 called peaks overlapping TF binding motifs were computed using the default setting in the Homer suite⁷⁵ function findMotifsGenome.pl (with genome reference hg19, option `-size` given). For co-accessible peaks and *local*-ATAC-peaks, background overlap percentages were computed using randomly sampled genomic regions preserving the width of each peak to assess the expected TF motif enrichment.

Transcription factor footprinting

Using the Homer suite tool annotatePeaks⁷⁵, and options `-m` and `-mbed`, we found all instances of BATF, ISRE, BATF/IRF, ETS1, and CTCF motifs in shared, Th-specific and Th_{stim}-specific peaks. Next, we determined the per-base coverage +/- 1 kb around the center of the motif using

BEDtools “coverage”, only counting reads that begin at a given location in order to prevent smoothing of the chromatin accessibility signal, and splitting the reads into those that map to the same or opposite strand as the motif to account for biases in ATAC-seq requiring two transposases (i.e. one at a TF binding site and another at a nucleosome accessible region). For each TF footprint, we generated a matrix with the number of rows equal to the number of instances of the motif by 4,000 columns quantifying coverage: +/- 1kb from the same and opposite strand and as the motif. Final TF footprints were derived from median normalized reads.

Outlier analysis and sample mix-up analysis

ATAC-seq samples were further filter if the samples contained a minimum of 8 million QC-passed reads (median of 37 million, MAD +/-13 million) and were highly correlated with other samples (mean Pearson R > 0.68). ATAC-seq and RNA-seq profiles from the 105 individuals were further filtered to identify sample mix-ups. We used the software VerifyBamID⁷⁶ to match each ATAC-seq and RNA-seq sample with the genotyping profile with the highest fIBD score. Samples with designated labels not matching the VerifyBamID predicted genotyping labels were flagged as sample mix-ups. We switched the designated label to the predicted label for cases where the fIBD > 90%. 15 out of the 139 total ATAC-seq samples were re-labeled and four out of the 110 total RNA-seq samples were re-labeled. For the ATAC-seq samples: 18 do not have genotypes, three are outliers, one did not match anyone. For the 110 RNA-seq samples: eight samples do not have genotypes, five are outliers, one did not match anyone. 111 ATAC-seq samples and 96 RNA-seq samples were used in the final analysis after filtering. In the response to activation study, there were five people total, 1 person was repeated for a total of six samples, none were genotyped.

Genetic association analysis of ATAC-peaks

Genetic association analysis was performed on 105 samples of European descent (**Supplementary Fig. 2.S6**) by running RASQUAL⁴³ on the 167,140 peaks identified in Th_{stim} cells and 4,558,693 imputed genetic variants, testing variants within a 1 Mb window of each ATAC-peak, and filtering for a minor allele frequency of greater than 5% using uniquely mapped nuclear reads per individual. Sex and ten principal components (**Supplementary Table 2.4**) were included as covariates to minimize the effects of confounding factors. Using the RASQUAL “-r” option, 10 permutations were generated for each ATAC-peak. For *local*-ATAC-peak analysis, association statistics for 158,613 peak-SNP pairs where the SNP resides within the peak are compared. For *distal*-ATAC-peak analysis, association statistics for peak-SNP pairs where the SNP does not reside within the peak are compared. In each case, empirical *P*-values and the corresponding false discovery rates were computed using the R qvalue⁴⁴ package to detect a total of 3,318 *local* - ATAC-peaks (FDR < 0.05) and 382 *distal*-ATAC-peaks (FDR < 0.05).

Hi-C data analysis

The sequenced reads were analyzed using the Juicer pipeline⁷⁷. We sequenced 2,940,433,604 Hi-C read pairs in stimulated T cells. Loci were assigned to A and B compartments at 500 kB resolution. Contact domains were annotated using the Arrowhead algorithm with default Juicer parameters at 5kB for stimulated T cells. This yielded a list of 4,008 domains in stimulated T cells at MAPQ > 30. We also ran Arrowhead with these same respective parameters on MAPQ > 0 Hi-C maps, which yielded a list of 4,419 domains in stimulated T cells. The Hi-C maps and feature annotations were visualized using the Juicebox software⁷⁷.

Determination of distance from ATAC-peak to contact domains

We determined the distance from each SNP-containing ATAC-peak to the middle of the closest contact domain. We analyzed the following features: (1) all SNP-containing ATAC-peaks; (2) *local*-ATAC-peaks; and all SNP-containing ATAC-peaks and *local*-ATAC-peaks containing (3) BATF, (4) ETS1, or (5) CTCF motifs. Homer annotatePeaks ‘-mbed’⁷⁵ option was used to identify SNP-containing ATAC-peaks and *local*-ATAC-peaks that contain BATF, ETS1, and CTCF motifs, as previously described. We normalized the distances from each peak to the closest domain by the length of the domain. In order to determine that the distribution of the distance between a given peak and a contact domain is different than the null distribution, we kept the length of each contact domain constant and shuffled the positions of the contact domain. The distances from each peak to the contact domain were binned into 30 bins and divided by the binned distances between a given peak and the shuffled contact domains to determine enrichment at each position.

Co-accessible peak analysis

To identify co-accessible peaks, we computed the correlation between every pair of 167,140 ATAC-peaks within 1.5 Mb of each other using a linear regression model implemented by Matrix eQTL⁷⁸. We first normalized the ATAC-peaks by (1) removing sequencing depth bias using median normalization, (2) standardizing the matrix by subtracting out the mean and dividing by the standard deviation for each peak; and (3) quantile normalizing the matrix⁷⁹. Adjusting for sex and 15 principal components, we used Matrix eQTL to identify 2,158 pairs of co-accessible peaks (1,809 unique ATAC-peaks, FDR < 0.05-). We reran the analysis using 10 permuted datasets generated by shuffling the peak counts for an individual to obtain a distribution of permuted *P*-

values. The qvalue package was used to obtain empirical P -values and false discovery rates⁴⁴.

RNA-seq analysis

25bp paired end RNA-seq reads were aligned to the hg19 using UCSC transcriptome annotations. Expression levels (expected counts) were determined using RSEM⁷⁹. We applied trimmed mean of M-values normalization method (TMM) to the expected counts using the edgeR package and kept genes that had TMM count > 1 in at least 75% of the samples. For the mapping of eQTLs, we inputted expected counts for filtered genes into RASQUAL⁴³. For the heritability analyses, we used log-transformed TMM counts of filtered genes in order to fit linear mixed models.

Percentage of GWAS loci overlapping

The GREGOR suite⁶³ was used for calculating the percentage of GWAS loci in features of interest: (1) peaks differentially accessible in Th and Th_{stim} cells, (2) co-accessible peaks, (3) SNP-containing peaks, and (4) *local*-ATAC-peaks. GWAS loci in the National Human Genome Research Institute GWAS catalogue as of November 2016 were overlapped. For *local*-ATAC-peaks, peaks were randomly permuted, while retaining the width of each peak to assess the expected GWAS enrichment.

Partitioned heritability analysis

Partitioned heritability analysis was performed using LD Score⁸⁰. Summary statistics for all SNPs for 28 GWAS (Alzheimer, anorexia, autism, bipolar disorder, BMI, celiac, coronary artery disease, Crohn's disease, DS, ever smoked fasting glucose, HDL, IBD, LDL, lupus, multiple sclerosis, neuroticism, primary biliary cirrhosis, rheumatoid arthritis, schizophrenia, SWB, triglycerides,

type 1 diabetes, type 2 diabetes, ulcerative colitis, years of education 1, and years of education 2) phenotypes were downloaded from the Broad Institute (see URLs). *Local*-ATAC-QTLs were thresholded at FDR < 0.05, FDR < 0.1, FDR < 0.2, FDR < 0.5, and all tested SNPs. Using SNPs at each FDR threshold, annotation files and LD Scores were estimated for all 28 GWAS phenotypes using ‘ldsc.py -l2’. Finally, to calculate the partitioned heritability across each phenotype, including our *local*-ATAC-QTLs at each FDR threshold, respectively, ‘ldsc.py -h2’ was run.

Percentage overlapping T cell annotations

Using the Homer suite `annotatePeaks.pl` with the `-genomeOntology` option⁷⁵, we calculated how many of the Th_{stim}-specific peaks, co-accessible peaks, SNP-containing peaks, and *local*-ATAC-peaks fall T cell enhancers¹⁸. For co-accessible peaks, SNP-containing peaks, and *local*-ATAC-peaks peak subsets, background overlap was calculated using randomly sampled genomic regions preserving the width of each peak to assess the expected T cell enrichment.

Proportion in super-enhancer regions

Using the BEDtools “intersect” function, we calculated how many of the co-accessible peaks and co-accessible *local*-ATAC-peaks are also in stimulated Th super-enhancers (as reported in Hinsz et al.⁴¹). Background proportions were computed using randomly sampled genomic regions preserving the length of each super enhancer.

Proportion co-accessible peaks in known regulatory elements

Using the BEDtools “intersect” function⁷², we annotated each peak in our unique pairs of co-

accessible peaks as residing in a known Th_{stim} super enhancer (as reported in Hinsz et al.⁴¹), promoter (as reported in Fahr et al.¹⁸), and T cell promoter (as reported in Fahr et al.¹⁸). We determined if each peak in a pair of peaks resided in a promoter and a promoter, promoter and a super enhancer, a promoter and an enhancer, an enhancer and an enhancer, an enhancer and a super enhancer, and a super enhancer and a super enhancer. As background ('non co-accessible peak'), we used pairs of ATAC-peaks with P -value > 0.9 , sampled to the same number as co-accessible peaks, and performed the same analysis.

Gkm-SVM and deltaSVM

We ran gkm-SVM^{46,47} on 24,745 300bp ATAC-peaks centered on MACS summits using default parameters and an equal size GC matched negative set, excluding from training any region containing a SNP to be scored by deltaSVM, and repeated with 5 independent negative sets, and averaged the deltaSVM predictions, as previously described⁴⁸. We then calculated deltaSVM for each SNP in a *local*-ATAC-peak, scoring 903 SNPs in 888 loci. We find a Pearson correlation of $R=0.627$ between ATAC-QTL beta and the largest deltaSVM SNP. 777 of the peak P -value SNPs had the largest deltaSVM, but 111 flanking SNPs scored more highly than the peak P -value SNP and disrupt immune associated TF binding sites. While the gkm-SVM weights fully specify the deltaSVM score, for interpretation we associated the large gkm-SVM weights with the most similar TF PWM from a catalog of JASPAR, Transfac, Uniprobe, and Homer motifs.

Heritability of gene expression and ATAC-peaks

For the univariate analyses, restricted maximum likelihood heritability (h^2) estimates were calculated using GCTA software⁴⁵ with algorithm 1 and no constraints on heritability (*i.e.*, h^2 can

be less than 0), while the bivariate analysis was run constrained. For the gene expression heritability analysis, where gene expression was residualized for 12 principal components and sex, and ATAC-peak heritability analysis, where ATAC-peaks were residualized for 10 principal components and sex, we used genotypes +/- 500 kb from the transcription start site of the gene and center of each ATAC-peak, respectively. Of the 64,188 SNP-containing ATAC-peaks and 3,318 *local*-ATAC-peaks, 32,317 and 2,444 converged respectively. The bivariate GCTA analysis used genotypes +/- 500kb from the transcription start site of the gene. Randomly sampled ATAC-peaks (non *local*-ATAC-peaks) and permuted ATAC-peaks were plotted as background at the same number of the tested *local*-ATAC-peaks (N=161, standard errors < 1).

Mediation of eGenes by *local*-ATAC-peaks

Pairs of *local*-ATAC-peak and eGenes were matched through their shared eQTL. Our normalized ATAC-peak matrix, as previously described (**Methods**), was further adjusted for gender and 10 principal components was used as input *local*-ATAC-peaks. Normalized gene expression matrix, as previously described (**Methods**), was further adjusted for gender and 12 principal components used as input as input for our eGenes. For each eQTL a 92 x three matrix was formatted. Each row in the matrix corresponded to an individual and each column corresponded to (1) eQTL genotype, (2) normalized *local*-ATAC-peak, and (3) normalized eGene. First, we regressed *local*-ATAC-peak ~ eQTL. Second, we regressed eGene ~ eQTL + *local*-ATAC-peak. To test for statistically significant mediation effects, the mediator package⁵⁸ ‘mediate’ function was called using both regression models as input.

Conditioning on eQTL

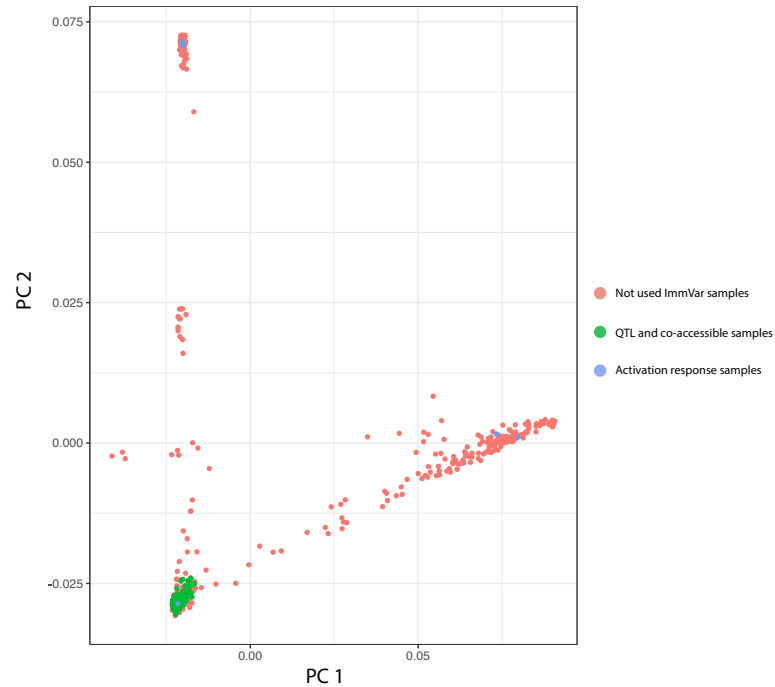
For each eQTL, the 92 x three matrix that was previously described (**Methods**) was used as input. To capture the effects of the eQTL, we regressed the eGene \sim *local*-ATAC-peak. To capture the effects after conditioning on the eQTL, we regressed the residuals of eGene \sim eQTL to the *local*-ATAC-peak.

Materials and Methods References

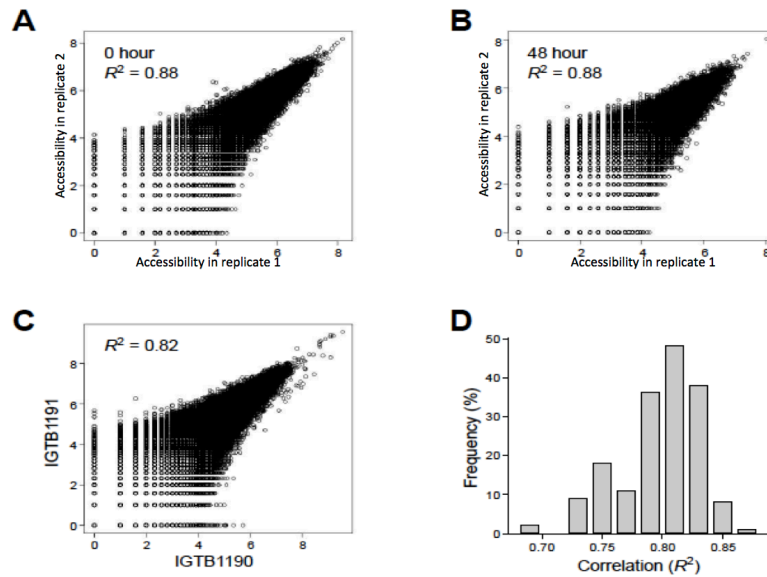
69. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 33, 495-502 (2015).
70. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589-95 (2010).
71. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-9 (2009).
72. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-2 (2010).
73. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137 (2008).
74. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).
75. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38, 576-89 (2010).
76. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 91, 839-48 (2012).
77. Durand, N.C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* 3, 95-8 (2016).
78. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353-8 (2012).
79. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323 (2011).
80. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-wide

association summary statistics. Nat Genet 47, 1228-35 (2015).

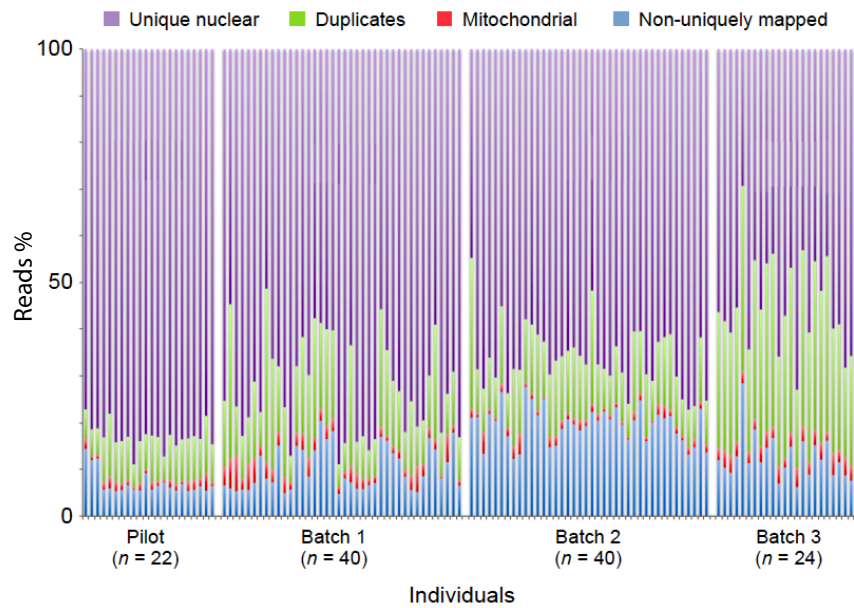
Supplementary Figures



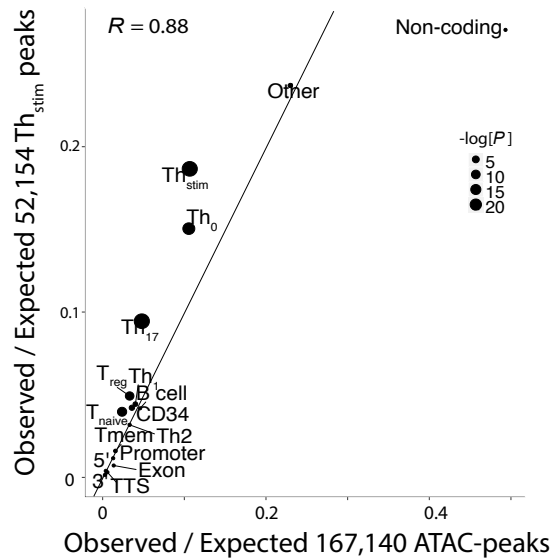
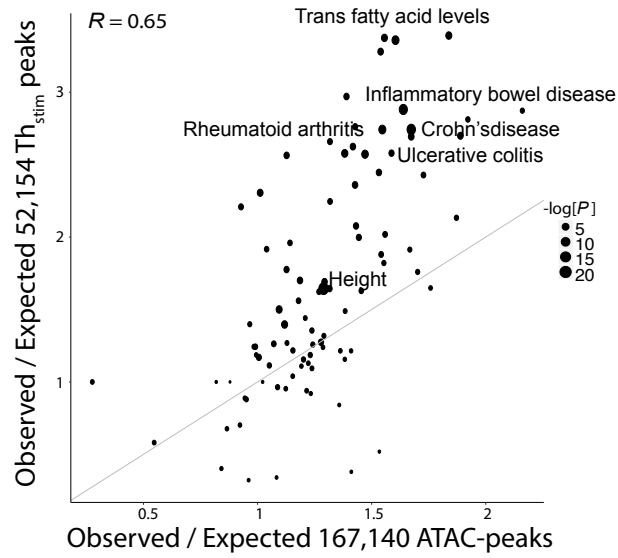
Supplementary Figure 2.1. PCA of the genetic relationships between individuals in the ImmVar cohort. Shown are the scores for each of 688 individuals in the ImmVar Consortium along the first two principal components (PCs, x and y axes) in a PCA of the genetic relationship matrix. Individuals used in this study are highlighted in blue (activation response analysis) and green (QTL and co-accessibility analyses).



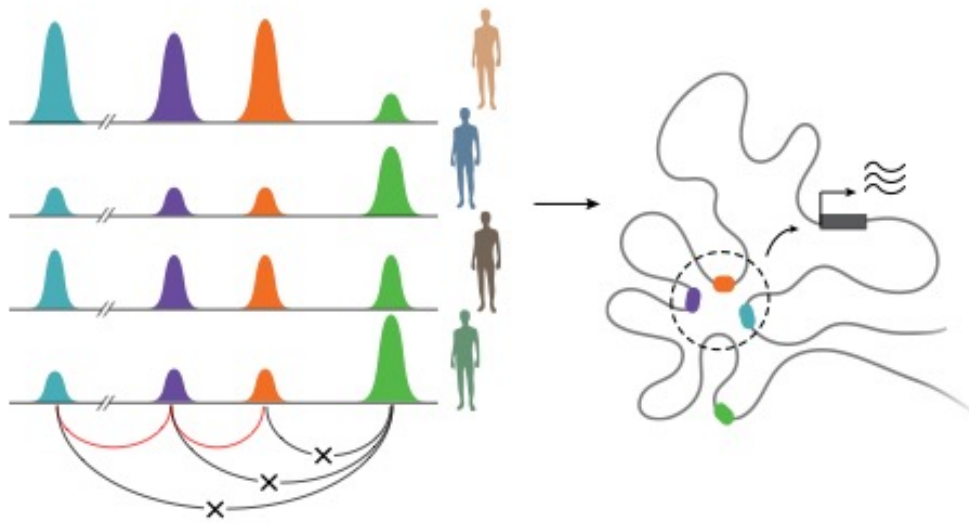
Supplementary Figure 2.2. ATAC-seq reproducibility. (a, b) Technical reproducibility. Scatter plots of chromatin accessibility (ATAC-seq signal, x and y axes) for two replicate experiments of either unstimulated (a; 36,486 Th peaks) or activated (b; 52,154 Th_{stim} peaks) T cells. (c, d) Reproducibility between individuals. (c) Chromatin accessibility for activated T cells from individuals IGTB1191 (y axis) and IGTB1190 (x axis) (d) and histogram of correlations between every pairs of individuals for the 52,154 Th_{stim} peaks.



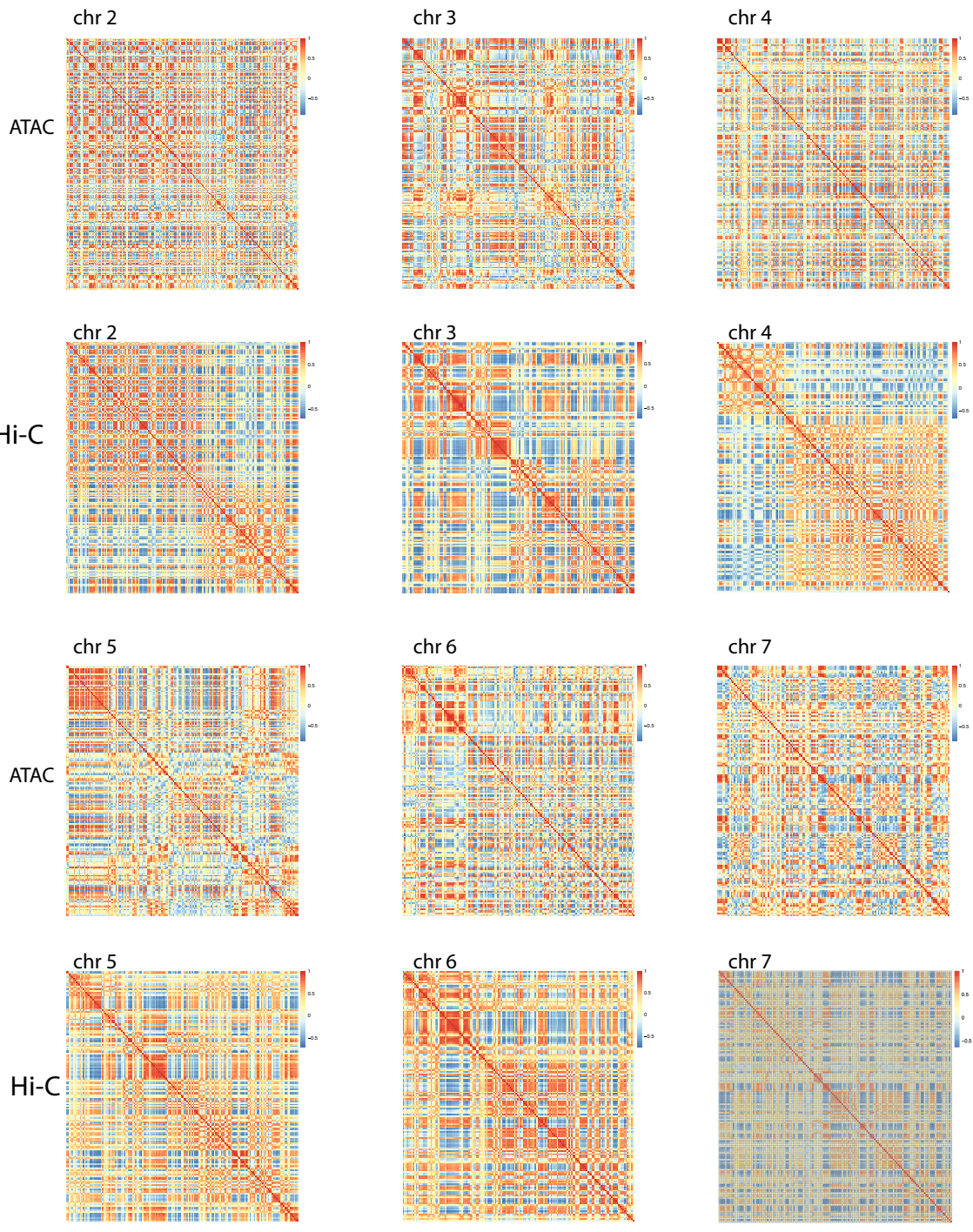
Supplementary Figure 2.3. ATAC-seq alignment rates. Each sample's (*x* axis) percentage of reads (*y* axis) that are non-uniquely mapped (blue), uniquely mitochondrial DNA (red), uniquely nuclear genomic duplicates (green), and uniquely nuclear genomic non-duplicates (purple).

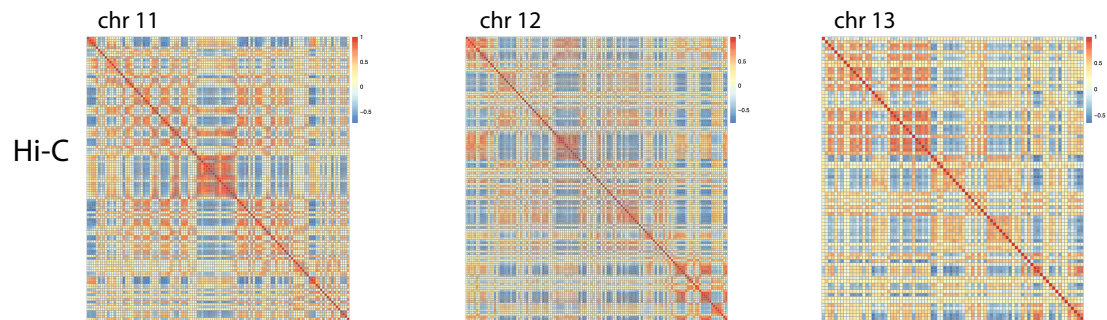
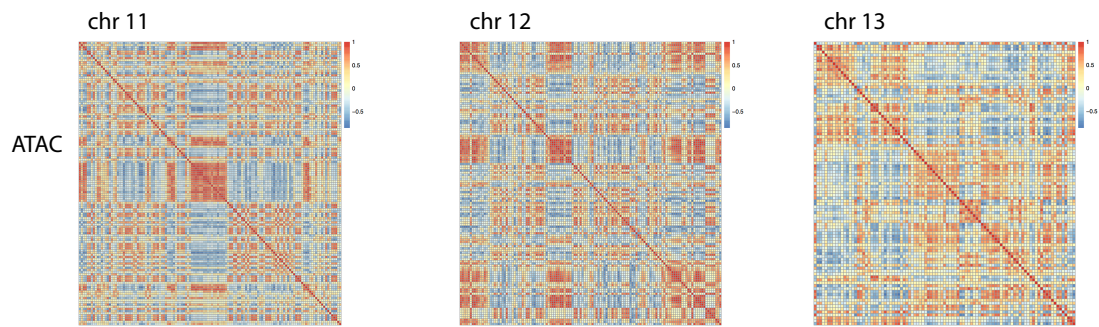
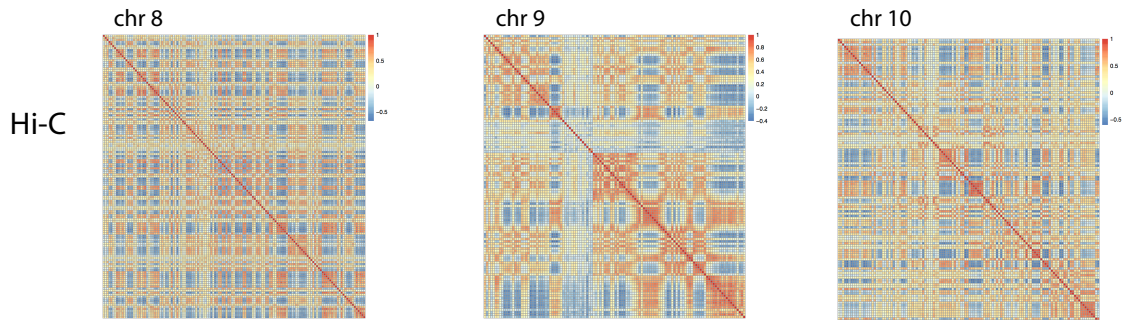
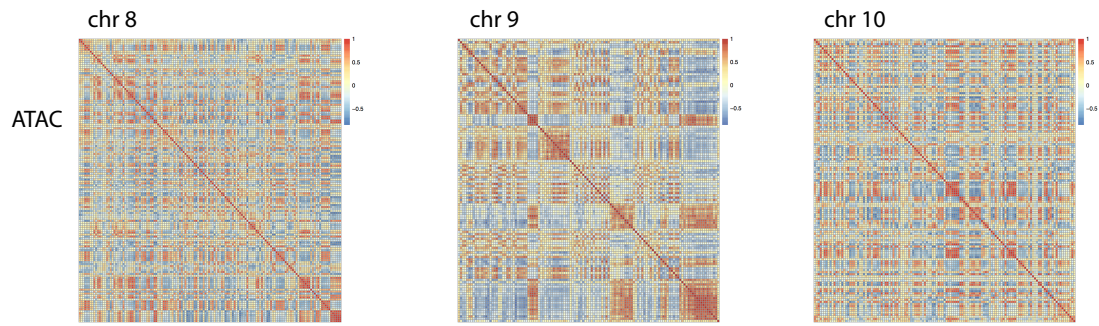


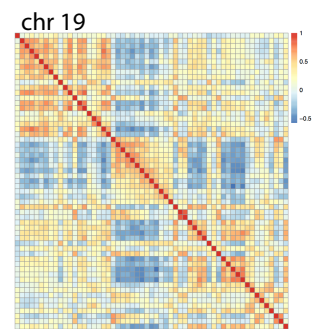
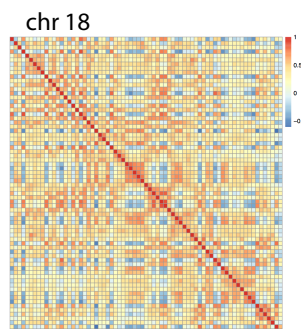
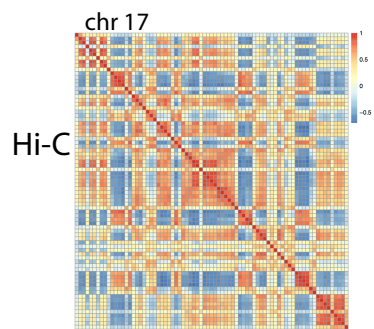
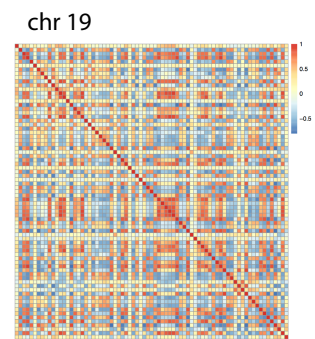
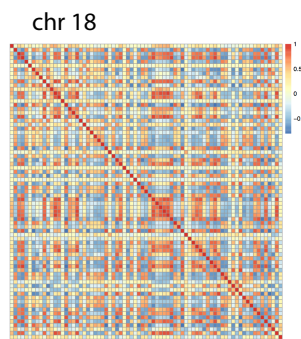
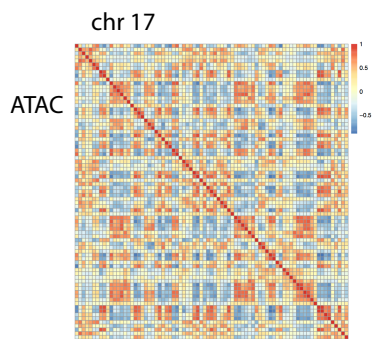
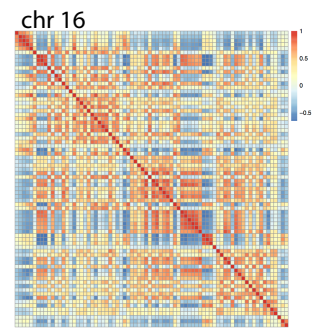
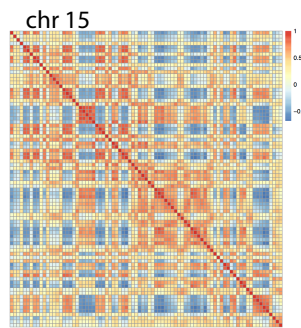
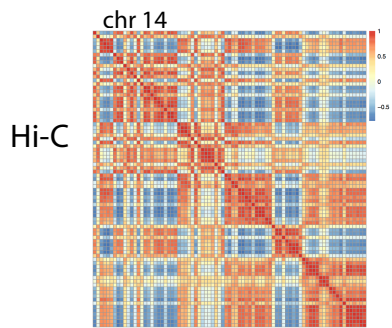
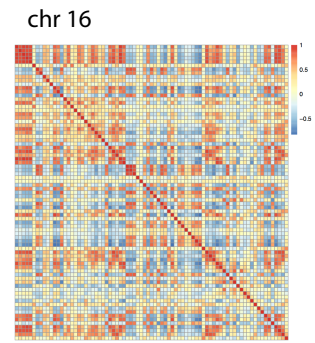
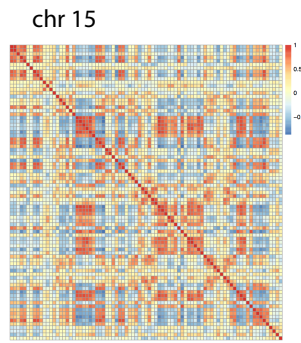
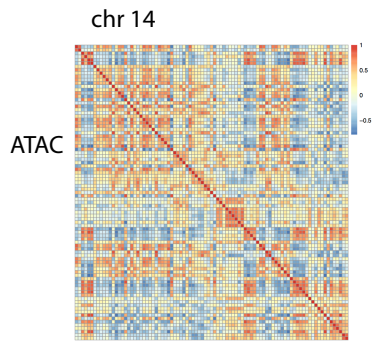
Supplementary Figure 2.4. Comparison of peak annotations in activated CD4⁺ T cells. For each set of GWAS loci (a) or enhancer (H3k27Ac marks, b) features, shown is their observed over expected enrichment of proportions in 52,154 Th_{stim} peaks called from pooling five individuals (y axis; as in **Fig. 2.1**) or in 167,140 ATAC-peaks called from 105 individuals (x axis, as in **Fig. 2.2**). Point size is scaled to associated significance (hypergeometric *P* value).

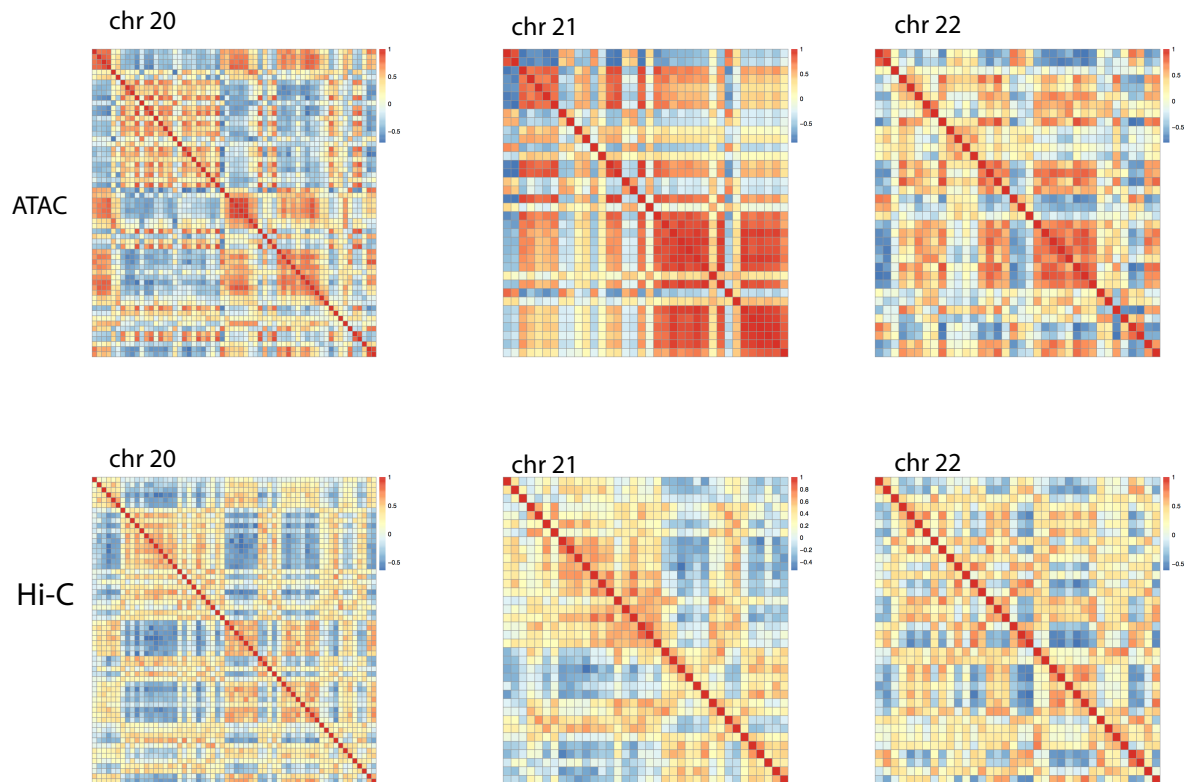


Supplementary Figure 2.5. Cartoon of co-accessible regions of accessible chromatin. Cartoon of the multiscale relationship between co-accessible regions across individuals and 3D genome structure.

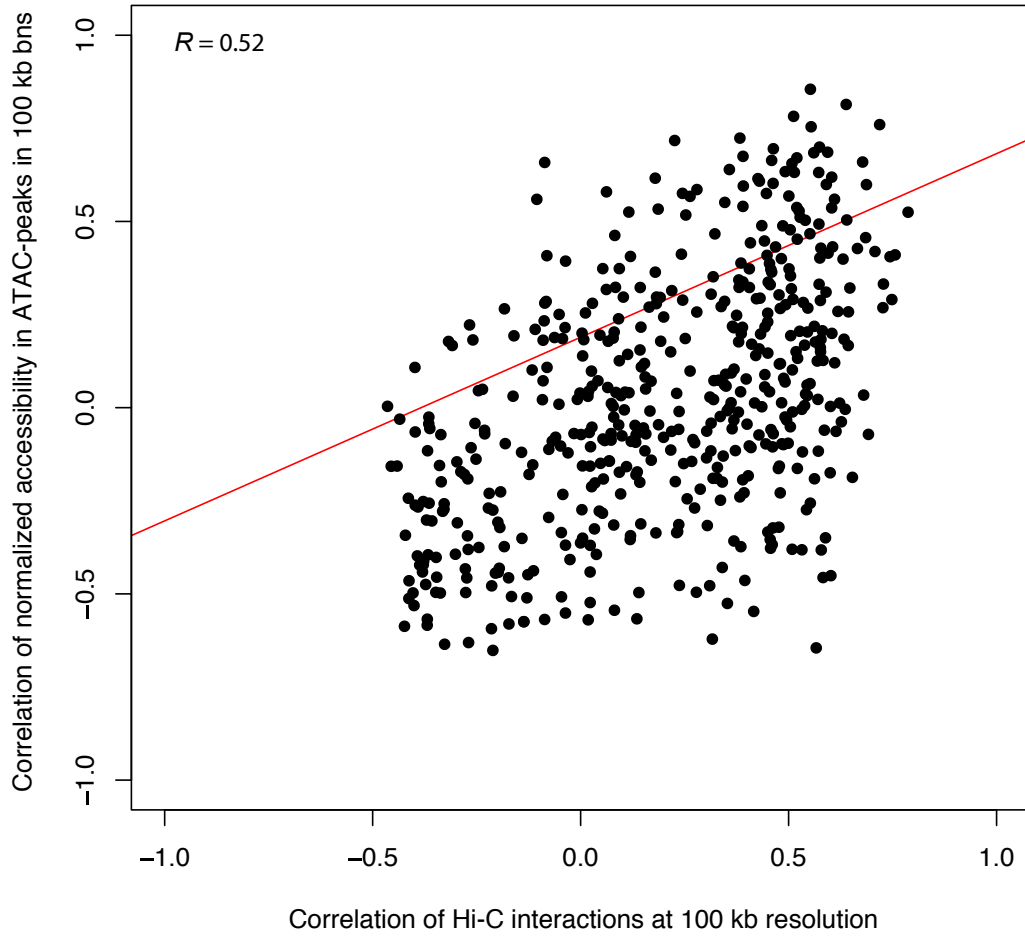






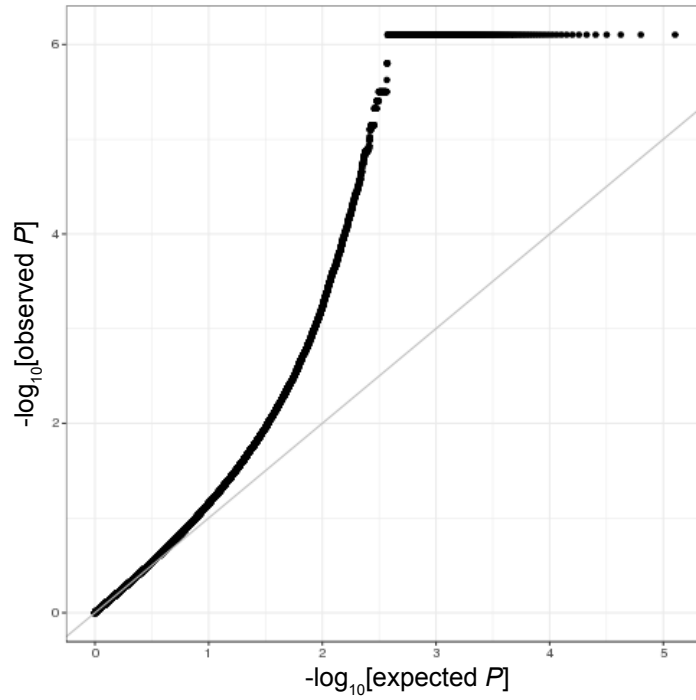


Supplementary Figure 2.6. Megabase scale inter-individual co-accessibility (by ATAC-peak) or physical interactions (by Hi-C). For each chromosome matching heat maps show the pairwise Pearson correlation in chromatin accessibility across 105 ATAC-seq profiles in ATAC-peaks binned into 1 Mb windows (top panel) and correlation of Hi-C interactions at 1 Mb resolution (bottom panel).

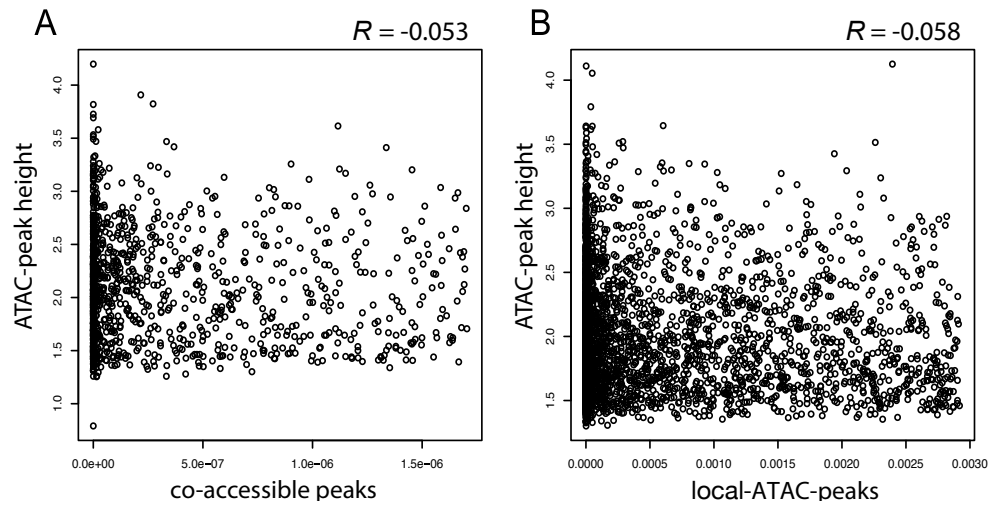


Supplementary Figure 2.7. Comparison of co-accessibility and Hi-C interactions.

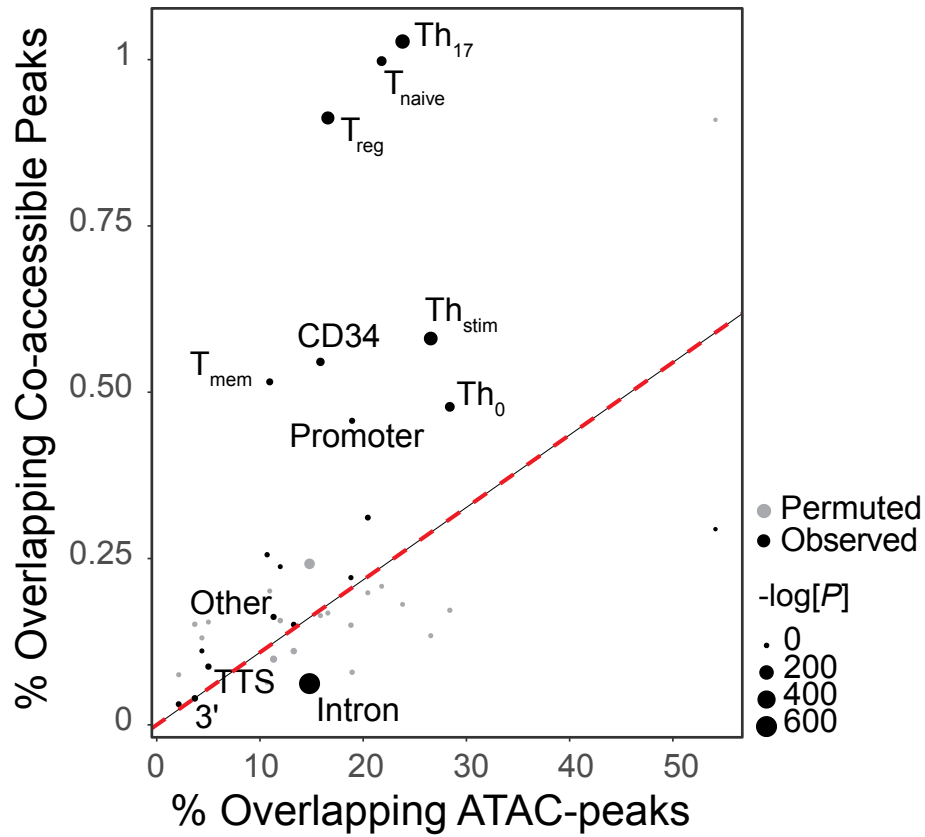
For each pair of 100 kb windows (shown are 500 randomly sampled pairs) along chromosome 22 are the Hi-C interaction score (x axis) and the correlation between ATAC-peaks (y axis) (Pearson $R = 0.52$).



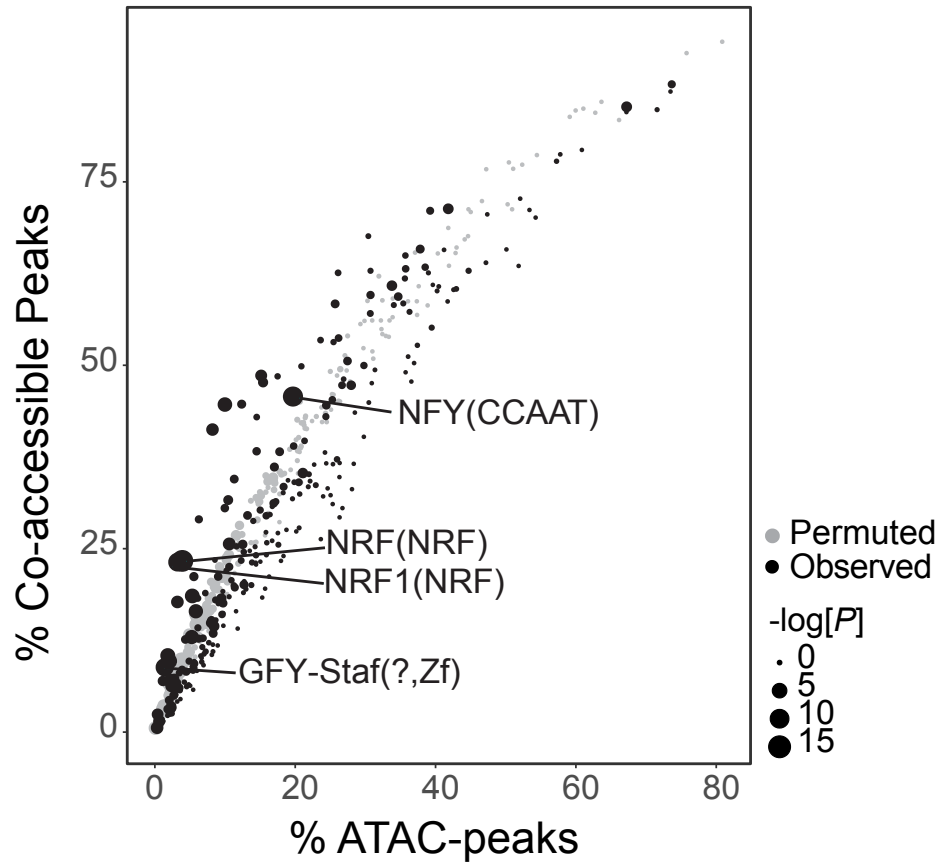
Supplementary Figure 2.8. Co-accessible peaks. Q-Q plot for all tests of correlation between co-accessible peaks within 1.5 Mb region around a target ATAC-peak.



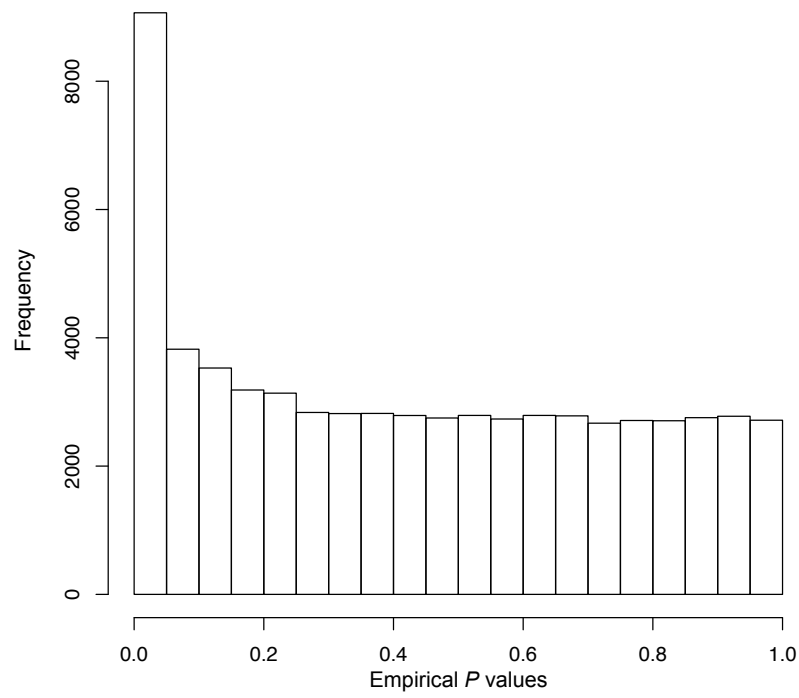
Supplementary Figure 2.9. Relation of ATAC-peak height and statistical significance of local-ATAC-peaks and co-accessible peaks. Relationship between ATAC-peak height (*y* axis) and statistical significance (*x* axis) of co-accessible peaks (a) and local-ATAC-peaks (b). Pearson *R* is marked on top, showing little relationship between the significance of local-association or co-accessibility and the strength of the peak.



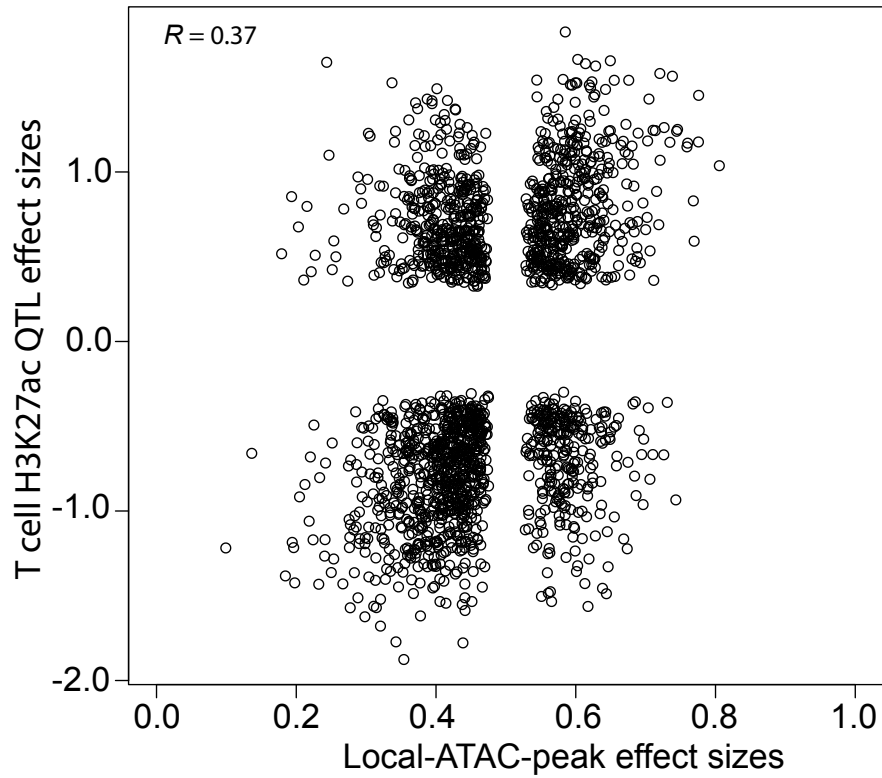
Supplementary Figure 2.10. Co-accessible peak overlap with Th cell enhancers. Percentages of enhancer annotations overlapping all ATAC-peaks (x axis) versus co-accessible peaks (y axis). Real peaks (black) and permuted peaks (gray).



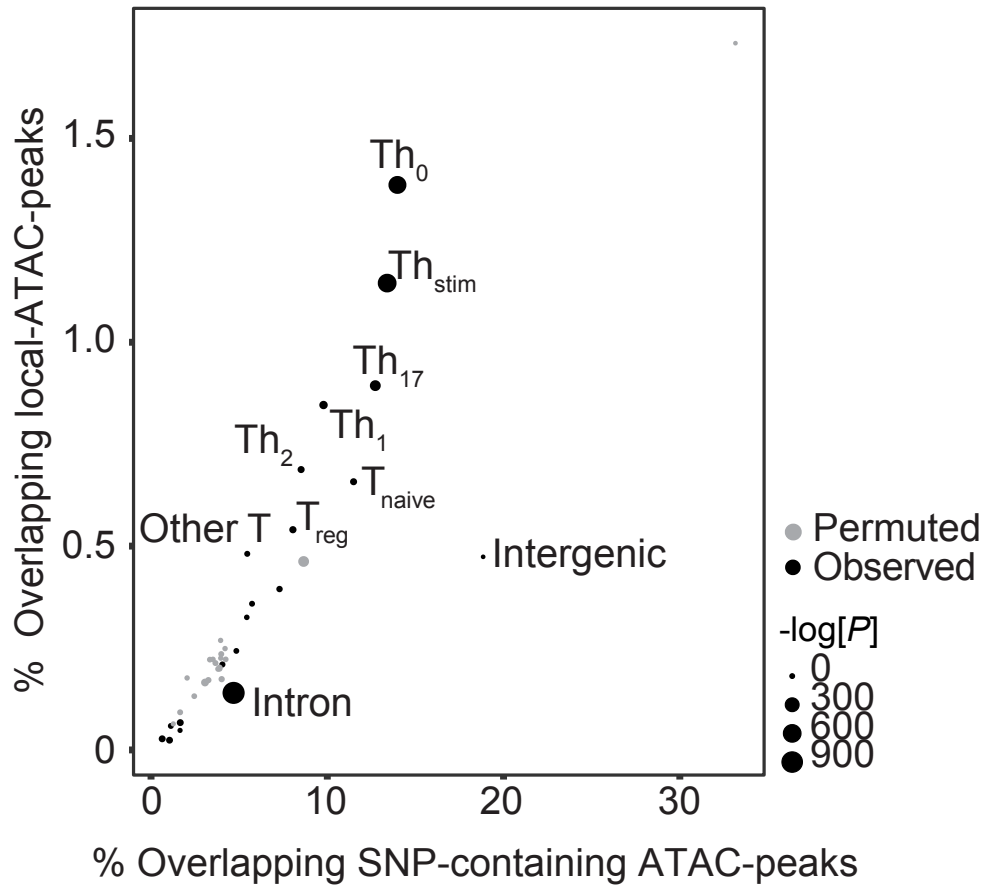
Supplementary Figure 2.11. Co-accessible peak TF motif enrichment. Percentage of ATAC-peaks (x axis) versus percentage of co-accessible peaks (y axis) overlapping TF binding sites. Real peaks (black) and permuted peaks (gray).



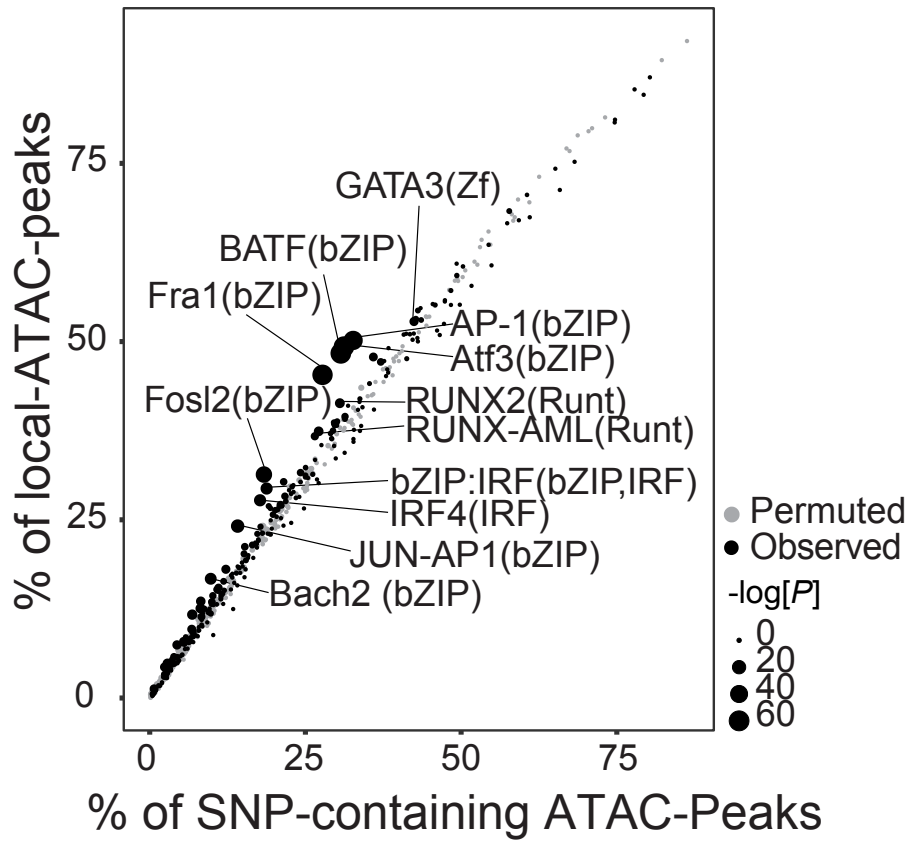
Supplementary Figure 2.12. Local-ATAC-QTL P values. Distribution of the empirical P values for the minimum statistical association per ATAC-peak containing a SNP.



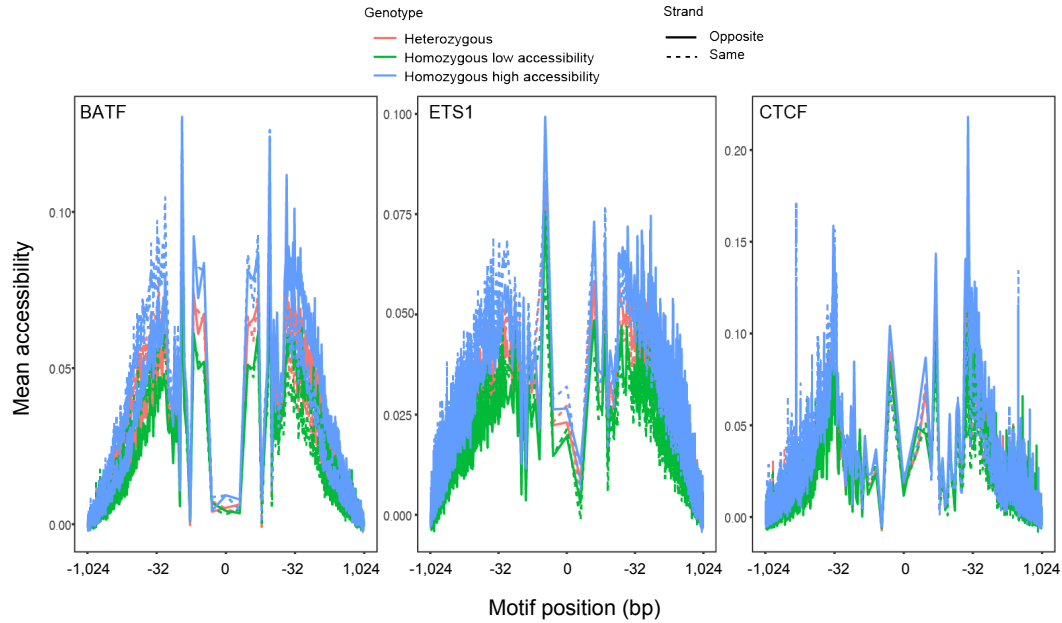
Supplementary Figure 2.13. Correlation to H3K27ac QTLs identified in immune cells in the BLUEPRINT epigenome project. Correlation of effect sizes for each of 2,015 loci identified as both T cell H3K27ac histone mark QTLs in the BLUEPRINT epigenome project (y axis) (I) and as local-ATAC-peaks (x axis) in our analysis.



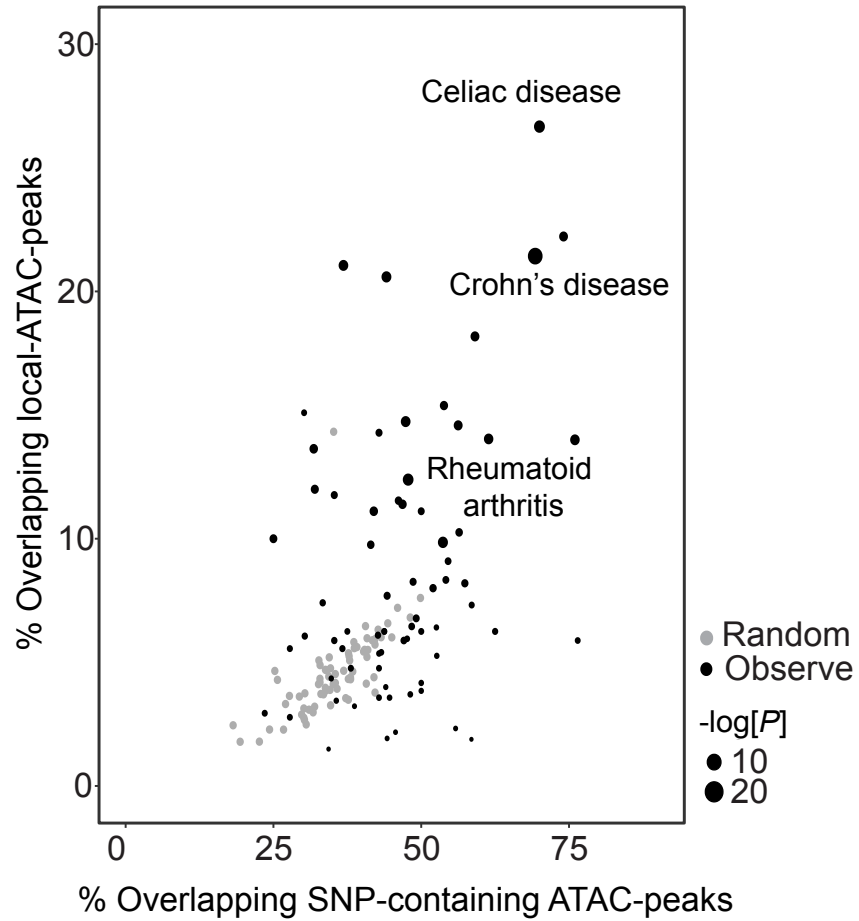
Supplementary Figure 2.14. Overlap of local-ATAC-peaks with T cell enhancers. Percentage of annotations overlapping SNP-containing ATAC-peaks (*x* axis) versus local-ATAC-peaks (*y* axis). Real peaks (black) and permuted peaks (gray).



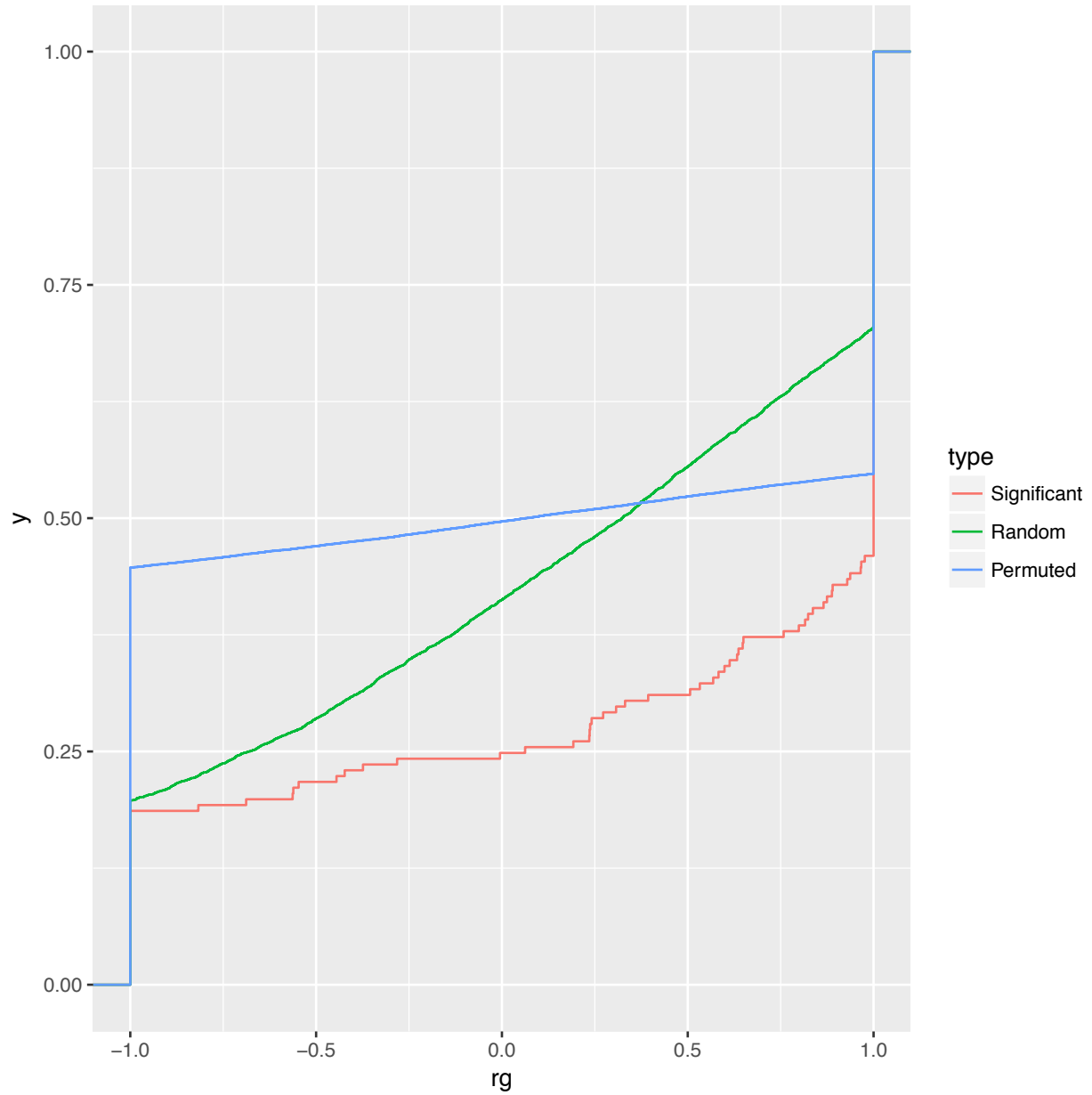
Supplementary Figure 2.15. TF motif enrichment of local-ATAC-peaks. Percentage of SNP-containing ATAC-peaks (x axis) versus percentage of local-ATAC-peaks (y axis) overlapping each TF motif. Real peaks (black) and permuted peaks (gray).



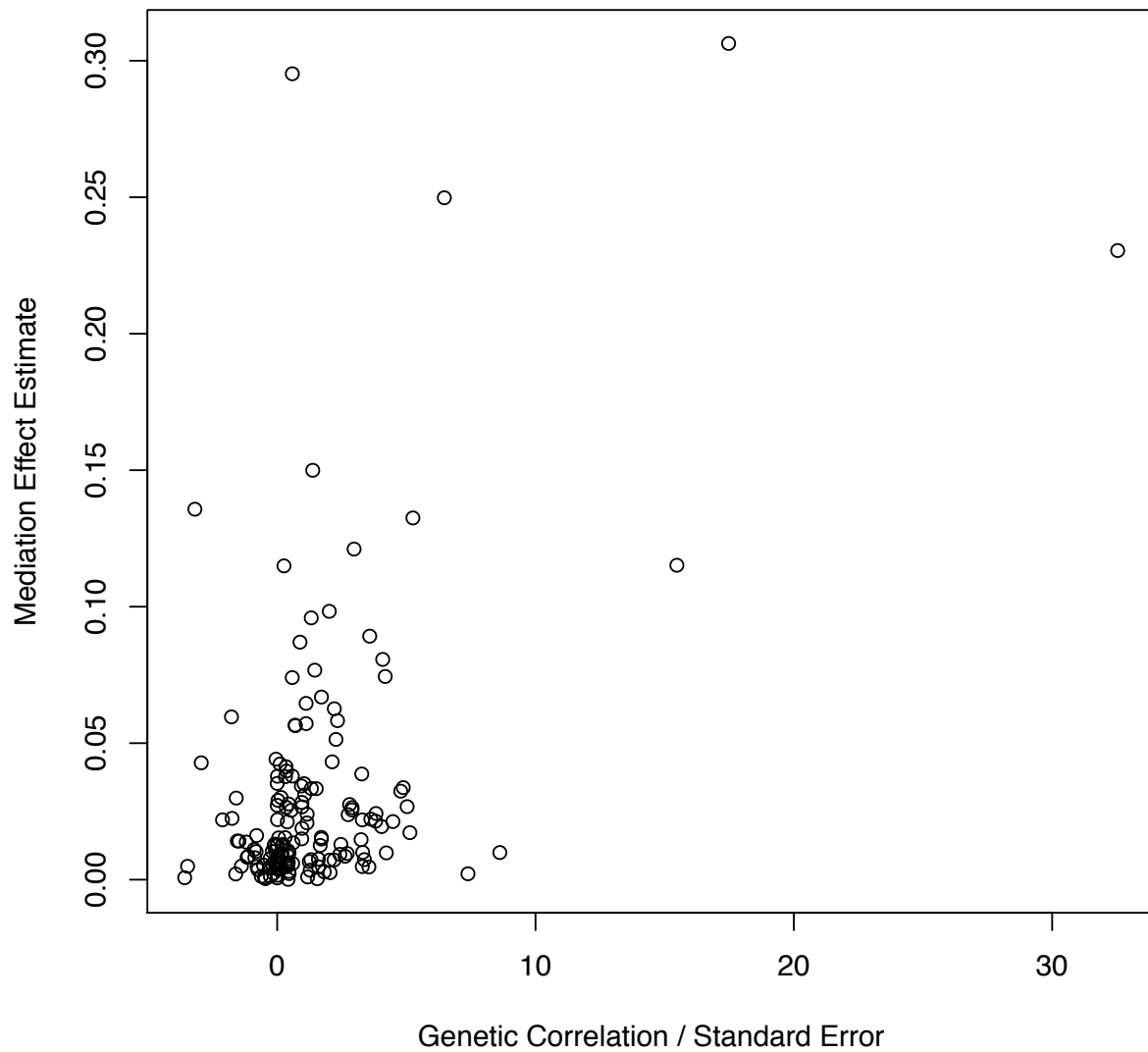
Supplementary Figure 2.16. Control for TF allele specific footprinting with local-ATAC-peaks. Aggregated chromatin accessibility (mean ATAC-seq signal, y axis) in local-ATAC-peaks around BATF, ETS1, and CTCF binding sites (x axis, log[bp from center of each TF motif]) previously identified by CHIP-seq (2) for the heterozygote (pink), homozygous with high ATAC-seq signal (blue) and homozygous with low ATAC-seq signal (green). Strand with the motif in dotted lines and the complementary strand with the motif in solid lines.



Supplementary Figure 2.17. Overlap of local-ATAC-peaks with GWAS loci. Percentages of GWAS loci overlapping with SNP-containing ATAC-peaks (*x* axis) versus local-ATAC-peaks (*y* axis). Real peaks (black) and permuted peaks (gray).



Supplementary Figure 2.18. Cumulative Distribution Function of genetic correlation. Colored by significant (red), randomly sampled (green) or permuted (blue) pairs of local-ATAC-peaks.



Supplementary Figure 2.19. Genetic correlation versus Mediation Effect Estimate. Genetic correlation (weighted by the standard error) (x axis) versus the mediation effect estimate (y axis) for 161 pairs of local-ATAC-peaks and eGenes that converged.

References:

1. L. Chen et al., Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 167, 1398-1414 e1324 (2016).
2. P. Kheradpour, M. Kellis, Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* 42, 2976-2987 (2014).

Chapter 3: Mapping gene regulatory networks of primary CD4⁺ T cells using single-cell genomics and genome engineering

Rachel E. Gate^{1,2†}, Min Cheol Kim^{1,3†}, Andrew Lu¹, David Lee¹, Eric Shifrut⁴⁻⁶, Meena Subramaniam^{1,2}, Alexander Marson⁴⁻⁹, Chun J. Ye^{1,9-12*}

¹Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA

²Biological and Medical Informatics Graduate Program, University of California, San Francisco, San Francisco, California, USA

³Medical Scientist Training Program, University of California, San Francisco, San Francisco, California, USA

⁴Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94143, USA

⁵Diabetes Center, University of California, San Francisco, San Francisco, CA 94143, USA

⁶Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA 94720, USA

⁷UCSF Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA 94158, USA

⁸Parker Institute for Cancer Immunotherapy, San Francisco, CA 94129, USA

⁹Department of Medicine, University of California, San Francisco, San Francisco, CA 94143, USA

¹⁰Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

¹¹Institute of Computational Health Sciences, University of California, San Francisco, San Francisco, CA, USA.

¹²Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA.

† These authors contributed equally to this work.

*Corresponding author. Email: jimmie.ye@ucsf.edu (C.J.Y.)

Introduction

CD4⁺ T cells display an incredible degree of functional diversity during adaptive immune responses classically characterized by the expression of canonical cytokines that create systemic inflammatory responses (e.g. Th₁₇)^{42–45}, signal and recruit B cells (e.g. Th₁ and Th₂)^{46–52}, and induce tolerance in the tissue microenvironment (e.g. T_{reg})^{53,54}. Recently, the ability of CD4⁺ T cells to directly kill infected and tumor cells have also received renewed attention^{5,55,56}. While the functionalization of CD4⁺ T cells is predominantly determined by the polarization of naive T cells (T_{naive}), recent results have suggested a high degree of variation within and plasticity between canonical subtypes⁵⁷. Indeed, we^{58,59} and others^{60–62} have shown that human circulating CD4⁺ T cells are composed of a mixture of canonical and non-canonical populations with significant interindividual variability in both the proportion and gene expression of each population⁵⁸.

Key regulators of CD4⁺ T cell differentiation and polarization have been mapped in mice utilizing pooled and arrayed genetic screens. For example, pooled knockdown screens with RNA interference have identified *Ppp2r2d* as a key regulator of T cell proliferation⁶³ and mapped two self-reinforcing, mutually antagonistic modules of regulators that drive Th₁₇ differentiation⁶⁴. More recently, a pooled genome wide CRISPR screen paired with bulk RNA sequencing identified *Trappc12*, *Mpv17l2*, and *Pou6f1* as regulators of both Th₂ activation and differentiation³⁵. Despite these insights, mapping gene regulatory programs that underlie human CD4⁺ T cell state transitions is incomplete and the intra- and inter-individual variability in these programs remain largely unknown.

Recent advances in the integration of droplet-based single cell RNA-sequencing (dscRNA-seq) and CRISPR/Cas9-mediated genome engineering has created new opportunities to assess the functional consequences of genetic perturbations in primary human T cells at an unprecedented molecular and cellular resolution ^{37,38,40,37,38,40}. Here, we integrate single guide RNA (sgRNA) lentiviral infection with Cas9 protein electroporation (SLICE) and multiplexed dscRNA-seq (mux-seq) to screen the effects of 140 regulators in primary CD4⁺ T cells across nine donors. By linking each sgRNA with the transcriptomes of hundreds of heterogeneous CD4⁺ cells, we map regulators that affect the activation and polarization of specific T cell subsets. By further leveraging the coexpression patterns across single cells, we define novel gene regulatory relationships between pairs of regulators and their downstream targets. Finally, by incorporating donor genetics, we identify instances where genetic effect on gene expression is modified by CRISPR perturbations. Our work demonstrates that systematic analyses using multiplexed single-cell genomics and genome engineering is a powerful approach to map the gene regulatory networks that govern the functionalization of primary human T cells and characterize the intra- and inter-individual variability in these networks.

Results

CRISPR perturbation screen in activated CD4⁺ T cells across donors

To map gene regulatory programs that underlie the polarization and activation of human CD4⁺ T cells, we performed sgRNA lentiviral infection with Cas9 protein electroporation (SLICE) ^{40,40} followed by multiplexed single-cell RNA-sequencing (mux-seq). Primary CD4⁺ T cells were isolated from peripheral blood mononuclear cells (PBMCs) and activated *in vitro* as previously described ^{41,65,41,65} (**Fig. 3.1A; Methods**). Activated cells were transfected with 280 sgRNAs

targeting 140 regulators that were either highly expressed (top quartile from bulk RNA-seq) or have binding sites that were differentially accessible (from bulk ATAC-seq) in activated CD4⁺ T cells⁵⁸⁵⁸ (**Fig. 3.1B; Table 3.1; Methods**). Following Cas9 electroporation and multiple rounds of selection and proliferation, activated CD4⁺ T cells³⁴³⁴ were pooled across 9 donors and profiled using the 10X Chromium platform in 16 wells resulting in 320,708 cell-containing droplets and 16,750 reads/droplet (**Fig. 3.S1-3; Methods**). To maximize the probability of detecting sgRNAs in cells, we further amplified and sequenced the sgRNA transcripts from the resulting 10X cDNA library to near saturation as previously described⁶⁶⁶⁶ (98% compared to 63% in the 3' tagged library; **Fig. 3.1C; Methods**).

After filtering for doublets using exonic SNPs (77,046 - 24%) and cells with ambiguous sgRNA assignments (79,037 cells - 25%; **Methods**), 164,623 cells were kept for subsequent analyses corresponding to 18,291±4,571 cells per donor and 882±522 cells per sgRNA (**Fig. 3.1C, D; Table 3.2; Methods**). Of these, 105,664, 40,960, 12,675, and 5,324 cells contained one, two, three, or four sgRNAs, resulting in an estimated multiplicity of infection (MOI) of 1 (**Fig. 3.1E, S3.4**).

To assess the cutting efficiency of each sgRNA, we sequenced the sgRNA pool and DNA of the edited cells from each donor by targeted amplification of 268/280 loci (**Methods**). The insertion and deletion (indel) frequencies at each targeted locus and coverage of the corresponding sgRNA in the pool are expected correlated ($R = 0.36$, $P < 1.68 \times 10^{-8}$, **Fig. 3.1F**) and the average ratio between these two quantities - defined as the cutting efficiency - is 21%±15% (**Fig. 3.1F inset**). We defined 14 sgRNAs as uncut negative controls (WT) where the ratios between the cutting efficiencies and proportion of cells containing each sgRNA are 1.645 standard deviations below

the mean (z-score < -1.645, $P < 0.05$) (**Fig. 3.1G, S3.5; Methods**). In all, the integration of SLICE and mux-seq is an efficient and cost-effective strategy for pooled screening and profiling of primary human T cells across many donors.

Heterogeneity of activated CD4⁺ T cells

Because CD4⁺ T cells dynamically migrate to and egress from tissues through circulation, cells isolated from PBMCs likely represent a functionally diverse population of cells reflective of the specific immunological state of an individual. This is supported by previous functional genomic analyses of activated primary CD4⁺ cells demonstrating marked heterogeneity within and variability between individuals in the chromatin and expression profiles that overlap signatures from multiple sorted populations including Th₁S, Th₂S, and Th₁₇S^{26,58,6226,58,62}.

To assess the heterogeneity of activated CD4⁺ T cells using dscRNA-seq, we clustered 164,623 cells into 10 Leiden clusters⁶⁷⁶⁷ with each cluster containing on average 16,462 cells (maximum: 65,720; minimum: 127; **Fig. 3.2A, B**). We identified 2,189 differentially expressed genes in at least one cluster (624±691 per cluster) and annotated each cluster based on the most highly expressed markers. We found a CD27⁺/CCR7⁺ naive T cell population (T_{naive})^{68,6968,69} (65,720 cells, 40%), and three effector populations including IL5⁺/IL17RB⁺/GATA3⁺ Th₂S⁷⁰⁻⁷⁸⁷⁰⁻⁷⁸ (1,969, 1.1%), IFNG⁺ Th₁S^{75,79,8075,79,80} (8,022 cells, 4.8%), and PRFI⁺/GNLY⁺/NKG7⁺ cytotoxic cells (T_{cyto})⁵⁶⁵⁶ (37,960, 23%) (**Fig. 3.2C, D**). We also identified three activated populations (Th_{stim}) defined by the expression of HMGB2 and STMN1^{81,8281,82} and distinguished from each other by the expression of histone markers (Th_{stim,histone}, 19,920 cells, 12%), cell cycling markers PTTG1^{83,8483,84} and KIAA0101⁸⁵⁻⁸⁷⁸⁵⁻⁸⁷ (Th_{stim,cycling}, 16,905 cells, 10%), and naive markers CCNB1⁸⁸⁸⁸

($T_{stim,naive}$, 12,345 cells, 7.4%). We also found a proliferating population (T_{prolif}) that expressed genes associated with tumor progression, including *FXYD5*⁸⁹⁻⁹¹⁸⁹⁻⁹¹, *LIMD2*^{92,9392,93}, and *PFDN5*⁹⁴⁹⁴ (903 cells, 0.05%). Finally, we identified two small clusters likely to be transitional, as they are intermediates in lineage trajectory⁹⁵⁹⁵ space either between naive and cytotoxic cells ($T_{naive \rightarrow cyto}$ 752 cells, 0.04%) or between naive and stimulated cells ($T_{naive \rightarrow stim}$: 172 cells, 0.01%; **Fig. 3.S6**). To validate these annotations, for each cluster, we correlated the average log fold change in expression of upregulated genes (with respect to all other clusters) to the bulk RNA-seq expression profiles across 45 reference circulating immune populations⁶²⁶² (**Methods**). This approach assigned 8/10 clusters to their expected reference population (**Fig. 3.2E, Methods**). The frequency of each cluster was generally consistent across donors (average pairwise $R = 0.94 \pm 0.04$; **Fig. 3.2F, S3.7**). These results demonstrate that multiplexed single-cell RNA-sequencing of activated CD4⁺ T cells recapitulate the expected T cell subpopulations obtained from sorted PBMCs and enables estimates of donor variability in T cell composition.

Regulator perturbations drive T cell polarization

Regulators that control the activation and polarization of specific CD4⁺ T cell subsets have been mapped in mice and humans using either pooled CRISPR/Cas9 screens sorting for specific cell surface markers or RNA-interference (RNAi) followed by bulk transcriptomic profiling^{64,9664,96}. These assays trade off perturbation throughput (low in RNAi, high in CRISPR/Cas9) and phenotypic resolution (low by cell sorting, high in bulk transcriptomic profiling). Here, we leverage the ability to link CRISPR perturbations to the transcriptomes of single cells using SLICE followed by mux-seq to enable high throughput (hundreds of loci) and high phenotypic resolution (transcriptome wide) mapping of regulators during human CD4⁺ activation and polarization. We

first demonstrate the robustness and performance of our strategy by the following two quality control assessments. One, comparing cells expressing each knockout sgRNA (KO cells) to cells expressing the wild type sgRNAs (WT cells), the expression fold change for the targeted regulator was lower than random genes ($FC_{\text{targeted}} = 0.56$ vs $FC_{\text{random}} = 0$; KS test; $P < 2.26 \times 10^{-16}$, **Fig. 3.3A**). Second, the transcriptomes of cells expressing sgRNAs targeting the same regulator or WT sgRNAs are more correlated on average ($R_{KO} = 0.44$, $R_{WT} = 0.50$) than cells expressing two random sgRNAs ($R_{\text{random}} = 0$; KS test $P < 2.2 \times 10^{-16}$) (**Fig. 3.3B**). These two results suggest that sgRNAs targeting the same gene have similar downstream transcriptomic effects despite a modest (albeit statistically significant) change in overall fold change of the targeted genes.

We next assessed the effects of KO sgRNAs on T cell states. Compared to WT cells, the proportion of KO cells is statistically enriched in activated or polarized subsets (e.g. $T_{\text{stim,naive}}$ and T_{H1}) and depleted in the inactivated subsets (e.g. T_{naive} cells and transitional $T_{\text{naive} \rightarrow \text{cyto}}$; hypergeometric test, $FDR < 0.05$; **Fig. 3.3C**). In order to quantify the effect of each sgRNA on cell state, we computed the proportion of cells in each cluster that contained a particular sgRNA and identified those sgRNAs significantly enriched or depleted using a Z test (**Fig. 3.3D**; **Table 3.4**; **Methods**). Each cluster had on average 13 sgRNAs depleted (z-score < -1.5 , $P < 0.1$) and 25 sgRNAs enriched (z-score > 1.5 , $P < 0.1$; **Fig. 3.S8**). For example, the sgRNA (cutsite: chr2:96551631, cutting efficiency: 0.47) targeting the RNA-binding protein AT-Rich Interactive Domain-Containing Protein 5A (*ARID5A*) was enriched in the T_{H2} cluster (z-score > 1.5 , $P = 9.0 \times 10^{-3}$; **Fig. 3.3E,F**, top panels) and slightly depleted, although not statistically significantly, in the T_{naive} cluster (z-score < -0.4 , $P = 0.33$). The second *ARID5A*-targeting sgRNA (cutsite: chr2:96550280, cutting efficiency: 0.097) showed consistent patterns of enrichment in T_{naive} and T_{H2} cells (T_{naive} : z-score

$< -1.5, P < 0.1$; Th₂: z-score $> 0.2, P = 0.35$) (**Fig. 3.S9**). In contrast, the sgRNA targeting interferon response factor 2 (*IRF2*; cutsite: chr4:184418577, cutting efficiency: 0.25) was depleted in Th₂ cells (z-score $< -1.5, P < 0.1$; **Fig. 3.3E, F**, right panels). The second *IRF2*-targeting sgRNA (cutsite: chr4:184418667) had a cutting efficiency of 0.04 and was thus considered a WT sgRNA and is not enriched or depleted in any cluster (**Fig. 3.S5**).

We next estimated the trajectory of polarization from T_{naive} to Th₂ cells using diffusion pseudotime (DPT) and quantified the distribution of *ARID5A*-targeting (cutsite: chr2:96551631) and *IRF2*-targeting (cutsite: chr4:184418577) sgRNAs along the trajectory (**Methods**). The shape of the cumulative distribution function along the DPT is informative of enrichment or depletion of cells along the polarization trajectory. The steeper the initial rise in the cumulative percentage, the more likely a group of cells are to be naive, residing at an earlier “time-point”. Compared to all cells, cells containing the *IRF2*-targeting (cutsite: chr4:184418577) sgRNA are less likely to be Th₂-like as shown by the 98.1% cumulative percentage at ~ 0.1 DPT (**Fig. 3.3G**), suggesting that *IRF2* could be important for the polarization of T_{naive} cells to Th₂ cells or the maintenance of already polarized Th₂ cells. In contrast, cells containing *ARID5A*-targeting (cutsite: chr2:96551631) sgRNA had a 90.6% cumulative percentage at ~ 0.1 DPT, exemplifying a greater enrichment at a later pseudotime (more Th₂-like). This suggests that *ARID5A* may play a role in maintaining the T_{naive} cell phenotype, which is consistent with previous reports⁹⁷⁹⁷ (**Fig. 3.3G**).

To validate the *IRF2*- and *ARID5A*-targeting phenotypes, we used the Cas9 ribonucleoprotein (RNP) system to knockout *GATA3*, *ARID5A* and *IRF2* each with two sgRNAs and two non-targeting negative controls under general activation (anti-CD3/CD28) and Th₂ (IL-4, anti-IFNG,

anti-IL-12) polarization conditions (**Methods**). After two weeks of culturing, we used fluorescence-activated cell sorting (FACS) to sort for Th₂ (CD62L⁺) and Th₁ (T-bet⁺) cells and extracted DNA to assess cutting efficiency for each sample. In activated cells containing *IRF2*-targeting sgRNAs, the proportion of Th₂s (CD62L⁺) was lower (3.81% and 2.02%) compared to non-targeting controls (3.46% and 7.71%) but higher compared to *GATA3*-targeting cells (1.26% and 0.892%; **Fig. 3.3H, S3.10**), consistent with the pooled screen results. Further, the proportion of Th₁s (T-bet⁺) was higher in *IRF2*-targeting cells (10.5% and 8.02%) compared to non-targeting controls (0.87% and 1.93%) and *GATA3*-targeting cells (6.42% and 7.42%; **Fig. 3.3H, S3.10**). Interestingly, in Th₂-polarized cells, there was not a change in the proportion of Th₂ cells (**Fig. 3.S11**). In contrast, in activated and Th₂ polarized cells containing *ARID5A*-targeting sgRNAs, the proportion of Th₂ cells are slightly higher (10.2% and 6.61%) and the proportion of Th₁ cells remain unchanged (1.83% and 1.27%) (**Fig. 3.3H, S3.11**). These results recapitulate the pooled screen with *IRF2* acting as a positive regulator and *ARID5A* as a negative regulator of Th₂s. Using a multiplexed pooled screening framework, we were able to harness the high-resolution transcriptomic data to help elucidate and validate novel cell state regulators.

Regulators interact to alter gene expression

Activation and polarization of T cells is known to involve the genetic interaction of regulators through direct physical cooperation, competition, and feedback and feedforward regulation of gene expression^{98–10198–101}. For example, BATF and JUN physically interact to regulate transcription in dendritic cells (DCs), T cells, and B cells by jointly interacting with IRFs to bind compound-binding AP-1–IRF consensus elements (AICEs)¹⁰²¹⁰². However, a map of genetic interactions

between regulators that specify T cell function remain uncharted, primarily due to insufficient scalable methods to test for genetic interactions in primary T cells.

RNA interference or CRISPR perturbations followed by bulk RNA-seq allows us to study how genetic perturbations change gene expression on average across a population of cells. Detecting genetic interactions between regulators in this setting would require perturbing multiple regulators and observing non-additive changes in average expression, which is both experimentally and statistically intractable beyond a few regulators. By leveraging the ability to link genetic perturbations with their effects in many cells, we used SLICE followed by mux-seq to estimate the effects of CRISPR perturbations on the correlation between genes across cells to detect genetic interactions between regulators.

Specifically, we sought to map genetic interactions by identifying mutually mediating pairs of regulators, defined as one regulator modifying the correlation between another regulator with a downstream gene. As an example, if knocking out R_1 modifies the correlation between R_2 and G , then directionality is established as R_1 *mediates* the effect of R_2 on G (**Fig. 3.4A**) and vice versa. If R_1 and R_2 mutually mediate each other's effects on G , we call R_1 and R_2 a genetic *interaction* and R_1 , R_2 , and G as a regulator (R) pair - gene triplet. To statistically detect mediation, we performed a likelihood ratio test between two linear mixed effect models, testing for the interaction term of R_1 (presence of sgRNA targeting R_1) and $R_{2,exp}$ (R_2 expression) (**Methods**). A significant change in correlation between expression of $R_{2,exp}$ and a downstream gene suggests R_1 mediates the effect of R_2 on the downstream gene.

We identify four different types of regulator interactions: 1) cooperative activation, where regulators and the target gene are positively correlated in WT cells and uncorrelated in KO cells; 2) cooperative repression, where regulators and the target gene are negatively correlated in WT cells and uncorrelated in KO cells; 3) competitive activation, where regulators and the target gene are uncorrelated in WT cells and are positively correlated in KO cells; and 4) competitive repression, where regulators and target gene are uncorrelated in WT cells and are negatively correlated in KO cells (**Fig. 3.4A**).

We tested 37/140 of the most variably expressed regulators corresponding to total of 666 possible regulator pairs and 1,456,542 possible R pair - gene triplets (**Methods**). For each regulator, we first identified a set of downstream genes whose correlations with the regulator were affected when the regulator was perturbed using the same linear mixed effect model (**Fig. 3.4A; Methods**). For 33 of the 37 regulators (4 regulators each had a WT sgRNA), sgRNAs targeting the same regulator were more likely to identify the same downstream targets ($P < 0.005$, Mann-Whitney U test, **Fig. S3.12**) with similar changes in correlation ($R = 0.31$, $P = 3.2 \times 10^{-78}$; **Fig. S3.13**) compared to random ($R = -0.004$, $P=0.56$).

To identify R pair - gene triplets, we intersected downstream genes for each regulator pair (FDR < 0.1) and tested for mutual mediation. We identified 310 R pair - gene triplets (FDR < 0.05) where the regulators mutually mediated each other's effect on the downstream gene, comprised of 194 unique regulator pairs (**Fig. 3.4B; Table 3.5**). 24 of the regulator pairs identified were among the 48 regulator pairs previously known to interact (hypergeometric test, $P < 0.005$)¹⁰³¹⁰³. Combining all candidate genetic interactions reveals a core gene regulatory network in the functionalization

of primary T cell (**Fig. 3.4C**) that suggests *JUN*, *MYC*, *XBPI*, and *STAT3/6* forming a central hub with many overlapping interacting partners. To validate the predicted regulator interactions and their downstream targets, we searched for transcription factor binding sites (TFBSs) of the 31 transcription factor (TF) pairs upstream and downstream of each predicted downstream gene's transcription start site (TSS) that exist in the Homer database^{104,104} (**Methods**). We found a greater proportion of downstream gene TSSs containing TFBSs for both TFs compared to random sampling of TF pair - gene triplets ($P < 0.05$, Kolmogorov-Smirnov test across all windows, **Fig. 3.4D**).

We detected both known and novel interactions between key regulators for T cell functionalization. We identified two possible targets of the previously mentioned BATF-JUN interaction, *ATG14* and *TMEM204*. In addition, we identified an ETS1-STAT6 interaction, which is known to modulate cytokine responses in keratinocytes^{102,105,102,105}. While ETS1 has been shown to interact with numerous genes, in particular those in the STAT family, involved in the development and function of T cells^{106,106}, our result specifically suggests STAT6 as an interacting partner.

Amongst the candidate regulator interactions, we identified 23 pairs of cooperative activators, 29 pairs of competitive activators, 109 pairs of competitive repressors, and 149 pairs of competitive activators (**Fig. 3.4E**). In one example of competitive interaction, compared to WT cells, *GTF3A* is more correlated with genes in cells expressing the *CREM*-targeting (cutsite: chr10:35179264) sgRNA and *CREM* is more correlated with genes in cells expressing the *GTF3A*-targeting (cutsite: chr13:27424803) sgRNA (**Fig. 3.4F**). *CREM* and *GTF3A* is an example of a competitive repressor

pair that regulates the expression of *CLUAPI*, where *CLUAPI* is negatively correlated with each regulator in KO cells but uncorrelated in WT cells (**Fig. 3.4G**). In another example, MYC and NFATC3 cooperatively interact to activate *XRNI* (**Fig. 3.4H**) where *XRNI* is positively correlated with each regulator in WT cells but not correlated in KO cells (FDR < 0.05). These results suggest that when a regulator is perturbed, downstream effects of other regulators become more prominent and this change can be harnessed to detect subtle interactions, often competitively activating interactions, between regulators.

CRISPR perturbation modifies genetic effects on gene expression

While the contribution of interindividual variability to the composition, expression and activation of CD4⁺ T cells *ex vivo* has been described by us ^{58,59,58,59} and others ^{27,60,107–111,27,60,107–111}, little is known about the interindividual variability in CRISPR perturbed cells. Using a linear mixed model, we analyzed cells expressing each sgRNA to identify 125 genes whose expression were variable between individuals (interindividual genes, FDR < 0.2) across 79 sgRNAs (interindividual sgRNAs; **Fig. 3.5A, B; Table 3.7; Methods**).

Interindividual variability can be attributed to genetic, environmental and technical confounding effects. To identify the genetic contribution, we performed an expression quantitative trait loci (eQTL) analysis using a linear mixed model that includes a genetic covariate term (**Fig. 3.5A; Methods**). Because of the limited number of samples, we significantly reduced the multiple testing burden by only testing for SNPs +/- 100 kb around a TSS with a minor allele frequency > 0.4 and only highly expressed genes per sgRNA (on average 1,891 genes). We found a total of 88 *cis*-eQTLs across cells expressing all sgRNAs (permutation FDR < 0.2) corresponding to 84 genes

(eGene) (**Fig. 3.5B**). To assess the robustness of these results, we performed two quality check analyses. First, genes previously reported to have an eQTL in activated CD4⁺ T cells⁵⁸⁵⁸ and with significant interindividual variability were more statistically significant than those that were not (**Fig. 3.5B**). Second, the variance explained by genetics was correlated with and expectedly less than the variance due to interindividual variation amongst eGenes ($R = 0.19$, $P = 0.03$, **Fig. 3.5C**).

We next assessed eQTLs detected in KO versus WT cells. Overall, eQTL effect sizes were more correlated between cells expressing sgRNAs targeting the same regulator ($R=0.24$, $P=0.003$) than between cells expressing a random pair of sgRNAs ($R=0.09$, $P=0.2$), suggesting genetic effects specific to each knockout (**Fig. 3.5D**). For the regulators that exist in the Homer database¹⁰⁴¹⁰⁴, we found that genes harboring binding sites for 10 regulators are more likely to be eQTLs (as indicated by more significant P) (**Fig. 3.5E**). On average, genetics explained 64% and 56% of the variability in KO and WT cells respectively (**Fig. 3.5F**). To ensure that this observation is not confounded by the limited number of WT sgRNAs, we bootstrapped the KO cells and found that in 92% of the bootstraps, genetic variants on average explained more variance in KO cells compared to WT cells (binomial $P < 2.2e-16$; **Fig. S3.14**). Finally, we found eQTLs were more likely to be detected in KO cells (22% of KO vs. 14% of WT sgRNAs; hypergeometric $P = 0$) (**Fig. S3.15**). These results support that *in vitro* genetic perturbations by CRISPR/Cas9 can uncover effects of natural genetic variation undetectable in unperturbed cells.

The increased power to detect genetic effects on gene expression only in KO cells could be due to a change in the *trans* environment or regulator-genetic interactions (*cis* x *trans* epistatic interaction) (**Fig. 3.5G**). If the activity of a regulator has an additive effect on gene expression,

then ablating the regulator will decrease the overall variance of gene expression thereby increasing the genetic contribution to gene expression (**Fig. 3.5F, S3.16-18**). Supporting this model, 80 out of the 88 eQTLs had lower standard errors in KO cells compared to WT cells (**Fig. S3.17**). If the activity of a regulator multiplicatively interacts with genetic variants (epistasis), then ablating the regulator should change the genetic effect on gene expression. To identify instances of epistasis, we fit a linear mixed model testing for *cis* (SNP) x *trans* (sgRNA presence or absence) interactions (**Fig. 3.5G**). We found statistical evidence for epistasis for 48 out of 88 eQTLs (FDR < 0.05), where the sgRNA is more likely to interact with the eQTL than a random SNP (**Fig. 3.5H**). These results suggest that CRISPR/Cas9 ablation of a regulator can uncover both additive and epistatic effects from standing genetic variation on gene expression.

As an illustrative example, we found evidence for an epistatic interaction between *IRF1*-targeting (cutsite: chr5:132487047) sgRNA and genetic variant (rs1885125) to regulate *MCM9* expression (**Fig. 3.5H**). *IRF1* (also known as interferon regulatory factor 1) has previously been shown to regulate T cell activation^{112–114},^{112–114}, particularly driving Th₁ polarization¹¹⁵,¹¹⁵. Two independent epigenetic analyses suggest IRF1 binding at the *MCM9* promoter (**Fig. S3.19**). First, in K562 cells, there is an IRF1 ChIP-seq peak 142 bp upstream from rs1885125, containing four SNPs in LD ($D' > 0.97$). Second, the Homer database¹⁰⁴,¹⁰⁴ predicted an IRF1 binding site 595 bp upstream from rs1885125 using an IRF1 ChIP-seq in peripheral blood mononuclear cells, which is flanked by rs4946371 ($D' > 0.98$). These results support the interaction between a genetic variant in a *cis* regulatory element of *MCM9* and the *trans* factor, IRF1, to account for 47% of *MCM9* expression variability. All together, these results suggest that CRISPR perturbations can uncover

interindividual variation in gene expression and in some cases, epistatically interact with natural standing variation to modulate the variability of gene expression.

Discussions

CD4⁺ T cells serve diverse roles in the adaptive immune system by dynamically responding to extracellular signals in their microenvironment. While the gene regulatory networks governing these responses have been extensively studied in mice, the topology and parameters of these networks, including how they vary across individuals, have not been mapped in humans. To address these gaps, we present the first large scale, multiplexed single cell RNA-seq study of activated primary CD4⁺ T cells isolated from 9 donors across CRISPR perturbations targeting hundreds of candidate regulators.

Activated CD4⁺ T cells are heterogeneous, capturing cell states reminiscent of canonical helper subtypes (e.g. Th₁, Th₂, etc), cytotoxic phenotypes, and broad activation or cell cycle. We find that cells expressing WT sgRNAs are more likely to be T_{naive} cells while those expressing KO sgRNAs are more likely to promote polarization into a differentiated state (e.g. Th_{1/2} cells), exemplified by the identification *ARID5A* as a negative regulator of Th₂ polarization. We expect that the application of our approach to cells differentiated or polarized under specific conditions (e.g. toward a Th₂ phenotype) rather than broad activation (e.g. anti-CD3/CD28 activation) can further refine the mapping of gene regulatory programs that control T cell differentiation, polarization and maintenance.

The identification of gene gene interactions is experimentally and combinatorically challenging in primary cells. By exploiting coexpression patterns between single cells, we devised a new approach based on differential correlation analysis to detect interactions between pairs of regulator during T cell activation and polarization. Using this approach, we reconstructed a gene regulatory network for T cell activation that includes known interactions (e.g. *JUN*, *MYC*, *XBPI*, *STAT*) and previously unreported interactions such as between *ETSI* and *STAT6*, which may be involved in the propagation of T cell cytokine signaling. By increasing the number of cells profiled using multiplexed workflows ^{41,11641,116} and the number of genetic perturbations through higher multiplicity of infection ¹¹⁷¹¹⁷, future integration of genome engineering and single cell transcriptomics would allow for refined mapping and causal reconstruction ¹¹⁸¹¹⁸ of gene regulatory networks in specific low frequency T cell subsets (e.g., Th_{1/2}, T_{naive}).¹¹⁸¹¹⁸

Although epistatic interactions involving naturally segregating variants have been identified in model organisms, there has been limited success in identifying these interactions using observational studies in humans due to limited power ¹¹⁹⁻¹²¹¹¹⁹⁻¹²¹. Our genetic-multiplexed approach allowed us to identify genes that are interindividual variable in CRISPR perturbed primary human cells and in some cases, pinpoint the genetic variants that likely mediate the variability. Akin to reducing the *trans* contribution of gene expression through *in vitro* perturbations^{58-60,12258-60,122} or computational adjustments ¹²³¹²³, we provide evidence of decreased gene expression variance in CRISPR perturbed cells thus increasing the ability to detect *cis* genetic effects. Surprisingly, we also found that some CRISPR perturbations can modify the effects of genetic variants on gene expression epistatically reminiscent of gene by environment effects detected by *in vitro* perturbation of cells ^{58-60,12258-60,122}. Thus, a comprehensive perturbative-QTL

analysis using CRISPR/Cas9 is a compelling alternative strategy to large observational studies for mapping genetic interactions that involve standing genetic variants in primary human cells.

Our work provides the first view into the heterogeneity of activated CD4⁺ T cells at the single cell resolution across pooled CRISPR perturbations and individuals. We identify candidate regulators of T cell polarization and two classes of genetic interactions. By harnessing natural and CRISPR genome engineering, we can begin to efficiently dissect gene regulatory networks and identify genetic interactions in primary human cells.

References

1. van den Broek, T., Borghans, J. A. M. & van Wijk, F. The full spectrum of human naive T cells. *Nat. Rev. Immunol.* 18, 363–373 (2018).
2. Alberts, B. et al. *The Adaptive Immune System.* (Garland Science, 2002).
3. Koch, U. & Radtke, F. Mechanisms of T cell development and transformation. *Annu. Rev. Cell Dev. Biol.* 27, 539–562 (2011).
4. Taniuchi, I. CD4 Helper and CD8 Cytotoxic T Cell Differentiation. *Annu. Rev. Immunol.* 36, 579–601 (2018).
5. Takeuchi, A. & Saito, T. CD4 CTL, a Cytotoxic Subset of CD4+ T Cells, Their Differentiation and Function. *Front. Immunol.* 8, 194 (2017).
6. Wagner, H., Starzinski-Powitz, A., Jung, H. & Röllinghoff, M. Induction of I region-restricted hapten-specific cytotoxic T lymphocytes. *J. Immunol.* 119, 1365–1368 (1977).
7. Billings, P., Burakoff, S., Dorf, M. E. & Benacerraf, B. Cytotoxic T lymphocytes specific for I region determinants do not require interactions with H-2K or D gene products. *J. Exp. Med.* 145, 1387–1392 (1977).
8. Zhu, J. & Paul, W. E. CD4 T cells: fates, functions, and faults. *Blood* 112, 1557–1569 (2008).
9. Mosmann, T. R., Cherwinski, H., Bond, M. W., Giedlin, M. A. & Coffman, R. L. Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins. *J. Immunol.* 136, 2348–2357 (1986).
10. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu. Rev. Immunol.* 28, 445–489 (2010).
11. Laurent, C., Fazilleau, N. & Brousset, P. A novel subset of T-helper cells: follicular T-

helper cells and their markers. *Haematologica* 95, 356–358 (2010).

12. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology* 21.29.1–21.29.9 (2015). doi:10.1002/0471142727.mb2129s109
13. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57 (2009).
14. Raha, D., Hong, M. & Snyder, M. ChIP-Seq: a method for global identification of regulatory elements in the genome. *Curr. Protoc. Mol. Biol.* Chapter 21, Unit 21.19.1–14 (2010).
15. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* 2010, db.prot5384 (2010).
16. Durek, P. et al. Epigenomic Profiling of Human CD4⁺ T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development. *Immunity* 45, 1148–1161 (2016).
17. Moskowitz, D. M. et al. Epigenomics of human CD8 T cell differentiation and aging. *Science immunology* 2, (2017).
18. Altorok, N. et al. Genome-wide DNA methylation patterns in naive CD4⁺ T cells from patients with primary Sjögren’s syndrome. *Arthritis & rheumatology* 66, 731–739 (2014).
19. Zhao, R. F. ENCODE: Deciphering function in the human genome. (2012).
20. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195 (2012).
21. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69 (2011).

22. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8, 206–216 (2007).
23. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
24. Gerstein, M. B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100 (2012).
25. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90 (2012).
26. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343 (2015).
27. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394 (2012).
28. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* 489, 75–82 (2012).
29. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
30. Kasowski, M. et al. Extensive variation in chromatin states across humans. *Science* 342, 750–752 (2013).
31. McVicker, G. et al. Identification of genetic variants that affect histone modifications in human cells. *Science* 342, 747–749 (2013).
32. Kilpinen, H. et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* 342, 744–747 (2013).
33. Ashis Saha, A. B. False positives in trans-eQTL and co-expression analyses arising from

- RNA-sequencing alignment errors. *F1000Res.* 7, (2018).
34. Schumann, K. et al. Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. *Proceedings of the National Academy of Sciences* 112, 10437–10442 (2015).
 35. Henriksson, J. et al. Genome-wide CRISPR Screens in T Helper Cells Reveal Pervasive Crosstalk between Activation and Differentiation. *Cell* 176, 882–896.e18 (2019).
 36. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620 (2015).
 37. Adamson, B. et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* 167, 1867–1882.e21 (2016).
 38. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* 167, 1853–1866.e17 (2016).
 39. Jaitin, D. A. et al. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* 167, 1883–1896.e15 (2016).
 40. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* 14, 297–301 (2017).
 41. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* 36, 89–94 (2018).
 42. Guglani, L. & Khader, S. A. Th17 cytokines in mucosal immunity and inflammation. *Curr. Opin. HIV AIDS* 5, 120–127 (2010).
 43. Kimura, A. & Kishimoto, T. Th17 cells in inflammation. *Int. Immunopharmacol.* 11, 319–322 (2011).
 44. Leung, S. et al. The cytokine milieu in the interplay of pathogenic Th1/Th17 cells and regulatory T cells in autoimmune disease. *Cell. Mol. Immunol.* 7, 182–189 (2010).

45. Liu, D. et al. IL-25 attenuates rheumatoid arthritis through suppression of Th17 immune responses in an IL-13-dependent manner. *Sci. Rep.* 6, 36002 (2016).
46. Rao, D. A. T Cells That Help B Cells in Chronically Inflamed Tissues. *Front. Immunol.* 9, 1924 (2018).
47. Petersone, L. et al. T Cell/B Cell Collaboration and Autoimmunity: An Intimate Relationship. *Front. Immunol.* 9, 1941 (2018).
48. Kuchen, S. et al. Essential role of IL-21 in B cell activation, expansion, and plasma cell generation during CD4⁺ T cell-B cell collaboration. *J. Immunol.* 179, 5886–5896 (2007).
49. Kerfoot, S. M. et al. Germinal center B cell and T follicular helper cell development initiates in the interfollicular zone. *Immunity* 34, 947–960 (2011).
50. Goodman, M. T cell regulation of polyclonal B cell responsiveness. III. Overt T helper and latent T suppressor activities from distinct subpopulations of unstimulated splenic T cells. *Journal of Experimental Medicine* 153, 844–856 (1981).
51. Hodes, R. J. T helper cell-b cell interaction: the roles of direct Th-B cell contact and cell-free mediators. *Semin. Immunol.* 1, 33–42 (1989).
52. Parker, D. C. T cell-dependent B cell activation. *Annu. Rev. Immunol.* 11, 331–360 (1993).
53. Legoux, F. P. et al. CD4⁺ T Cell Tolerance to Tissue-Restricted Self Antigens Is Mediated by Antigen-Specific Regulatory T Cells Rather Than Deletion. *Immunity* 43, 896–908 (2015).
54. Abbas, A. K., Lohr, J., Knoechel, B. & Nagabhushanam, V. T cell tolerance and autoimmunity. *Autoimmunity Reviews* 3, 471–475 (2004).
55. van Bergen, J. et al. Phenotypic and functional characterization of CD4 T cells expressing

- killer Ig-like receptors. *J. Immunol.* 173, 6719–6726 (2004).
56. Zaunders, J. J. et al. Identification of circulating antigen-specific CD4⁺ T lymphocytes with a CCR5⁺, cytotoxic phenotype in an HIV-1 long-term nonprogressor and in CMV infection. *Blood* 103, 2238–2247 (2004).
57. Zhou, X. et al. Instability of the transcription factor Foxp3 leads to the generation of pathogenic memory T cells in vivo. *Nat. Immunol.* 10, 1000–1007 (2009).
58. Gate, R. E. et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* 50, 1140–1150 (2018).
59. Ye, C. J. et al. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* 345, 1254665 (2014).
60. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* 344, 519–523 (2014).
61. Kasela, S. et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4⁺ versus CD8⁺ T cells. *PLoS Genet.* 13, e1006643 (2017).
62. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. *bioRxiv* 409722 (2018). doi:10.1101/409722
63. Zhou, P. et al. In vivo discovery of immunotherapy targets in the tumour microenvironment. *Nature* 506, 52–57 (2014).
64. Yosef, N. et al. Dynamic regulatory network controlling TH17 cell differentiation. *Nature* 496, 461–468 (2013).
65. Shifrut, E. et al. Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. *Cell* 175, 1958–1971.e15 (2018).
66. Hill, A. J. et al. On the design of CRISPR-based single-cell molecular screens. *Nat.*

Methods 15, 271–274 (2018).

67. Traag, V., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. (2018).

68. Campbell, J. J. et al. CCR7 expression and memory T cell diversity in humans. *J. Immunol.* 166, 877–884 (2001).

69. Worbs, T. & Förster, R. A key role for CCR7 in establishing central and peripheral tolerance. *Trends Immunol.* 28, 274–280 (2007).

70. Kanhere, A. et al. T-bet and GATA3 orchestrate Th1 and Th2 differentiation through lineage-specific targeting of distal regulatory elements. *Nature Communications* 3, (2012).

71. Zhu, J., Yamane, H., Cote-Sierra, J., Guo, L. & Paul, W. E. GATA-3 promotes Th2 responses through three different mechanisms: induction of Th2 cytokine production, selective growth of Th2 cells and inhibition of Th1 cell-specific factors. *Cell Research* 16, 3–10 (2006).

72. O’Garra, A. & Gabryšová, L. Transcription Factors Directing Th2 Differentiation: Gata-3 Plays a Dominant Role. *The Journal of Immunology* 196, 4423–4425 (2016).

73. Jakiela, B., Szczeklik, W. & Plutecka, H. Increased production of IL-5 and dominant Th2-type response in airways of Churg–Strauss syndrome patients. (2012).

74. Greenfeder, S., Umland, S. P., Cuss, F. M., Chapman, R. W. & Egan, R. W. Th2 cytokines and asthma. The role of interleukin-5 in allergic eosinophilic disease. *Respir. Res.* 2, 71–79 (2001).

75. Mosmann, T. R. & Coffman, R. L. TH1 and TH2 cells: different patterns of lymphokine secretion lead to different functional properties. *Annu. Rev. Immunol.* 7, 145–173 (1989).

76. Seumois, G. et al. Transcriptional Profiling of Th2 Cells Identifies Pathogenic Features Associated with Asthma. *J. Immunol.* 197, 655–664 (2016).

77. Angkasekwinai, P. et al. Interleukin 25 promotes the initiation of proallergic type 2 responses. *J. Exp. Med.* 204, 1509–1517 (2007).
78. Weathington, N. M. et al. IL-4 Induces IL17Rb Gene Transcription in Monocytic Cells with Coordinate Autocrine IL-25 Signaling. *Am. J. Respir. Cell Mol. Biol.* 57, 346–354 (2017).
79. Bradley, L. M., Dalton, D. K. & Croft, M. A direct role for IFN-gamma in regulation of Th1 cell development. *J. Immunol.* 157, 1350–1358 (1996).
80. Smeltz, R. B., Chen, J., Ehrhardt, R. & Shevach, E. M. Role of IFN- in Th1 Differentiation: IFN- Regulates IL-18R Expression by Preventing the Negative Effects of IL-4 and by Inducing/Maintaining IL-12 Receptor 2 Expression. *The Journal of Immunology* 168, 6165–6172 (2002).
81. Filbert, E. L., Le Borgne, M., Lin, J., Heuser, J. E. & Shaw, A. S. Stathmin regulates microtubule dynamics and microtubule organizing center polarization in activated T cells. *J. Immunol.* 188, 5421–5427 (2012).
82. Best, J. A. et al. Transcriptional insights into the CD8(+) T cell response to infection and memory T cell formation. *Nat. Immunol.* 14, 404–412 (2013).
83. Noll, J. E. et al. PTTG1 expression is associated with hyperproliferative disease and poor prognosis in multiple myeloma. *J. Hematol. Oncol.* 8, 106 (2015).
84. Wu, D. et al. Impact of PTTG1 downregulation on cell proliferation, cell cycle and cell invasion of osteosarcoma and related molecular mechanisms. *Zhonghua bing li xue za zhi= Chinese journal of pathology* 43, 695–698 (2014).
85. Fan, S., Li, X., Tie, L., Pan, Y. & Li, X. KIAA0101 is associated with human renal cell carcinoma proliferation and migration induced by erythropoietin. *Oncotarget* 7, 13520–13537 (2016).

86. Yuan, R.-H. et al. Overexpression of KIAA0101 predicts high stage, early tumor recurrence, and poor prognosis of hepatocellular carcinoma. *Clin. Cancer Res.* 13, 5368–5376 (2007).
87. Jain, M., Zhang, L., Patterson, E. E. & Kebebew, E. KIAA0101 is overexpressed, and promotes growth and invasion in adrenal cancer. *PLoS One* 6, e26866 (2011).
88. Chevaleyre, C. et al. The Tumor Antigen Cyclin B1 Hosts Multiple CD4 T Cell Epitopes Differently Recognized by Pre-Existing Naive and Memory Cells in Both Healthy and Cancer Donors. *J. Immunol.* 195, 1891–1901 (2015).
89. Brazee, P. L. et al. FXYD5 Is an Essential Mediator of the Inflammatory Response during Lung Injury. *Front. Immunol.* 8, 623 (2017).
90. Gotliv, I. L. FXYD5: Na⁺/K⁺-ATPase Regulator in Health and Disease. *Frontiers in Cell and Developmental Biology* 4, (2016).
91. Lubarski-Gotliv, I., Asher, C., Dada, L. A. & Garty, H. FXYD5 Protein Has a Pro-inflammatory Role in Epithelial Cells. *J. Biol. Chem.* 291, 11072–11082 (2016).
92. Peng, H. et al. LIMD2 Is a Small LIM-Only Protein Overexpressed in Metastatic Lesions That Regulates Cell Motility and Tumor Progression by Directly Binding to and Activating the Integrin-Linked Kinase. *Cancer Research* 74, 1390–1403 (2014).
93. Wang, F. et al. LIMD2 targeted by miR-34a promotes the proliferation and invasion of non-small cell lung cancer cells. *Mol. Med. Rep.* 18, 4760–4766 (2018).
94. Wang, D. et al. Prefoldin 1 promotes EMT and lung cancer progression by suppressing cyclin A expression. *Oncogene* 36, 885–898 (2017).
95. Wolf, F. A., Hamey, F., Plass, M., Solana, J. & Dahlin, J. S. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv*

(2018).

96. Ciofani, M. et al. A validated regulatory network for Th17 cell specification. *Cell* 151, 289–303 (2012).
97. Masuda, K. et al. Arid5a regulates naive CD4⁺ T cell fate through selective stabilization of Stat3 mRNA. *J. Exp. Med.* 213, 605–619 (2016).
98. Ravasi, T. et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140, 744–752 (2010).
99. Hao Yuan Kueh, E. V. R. Regulatory gene network circuits underlying T-cell development from multipotent progenitors. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 4, 79 (2012).
100. Hernandez-Munain, C., Roberts, J. L. & Krangel, M. S. Cooperation among Multiple Transcription Factors Is Required for Access to Minimal T-Cell Receptor α -Enhancer Chromatin In Vivo. *Mol. Cell. Biol.* 18, 3223 (1998).
101. Kröger, A. IRFs as competing pioneers in T-cell differentiation. *Cellular and Molecular Immunology* 14, 649 (2017).
102. Murphy, T. L., Tussiwand, R. & Murphy, K. M. Specificity through cooperation: BATF–IRF interactions control immune-regulatory networks. *Nat. Rev. Immunol.* 13, 499 (2013).
103. von Mering, C. et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33, D433–7 (2005).
104. Heinz, S. et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* 38, 576–589 (2010).
105. Travagli, J., Letourneur, M., Bertoglio, J. & Pierre, J. STAT6 and Ets-1 Form a Stable

Complex That Modulates Socs-1 Expression by Interleukin-4 in Keratinocytes. *J. Biol. Chem.* 279, 35183–35192 (2004).

106. Nguyen, H. V. et al. The Ets-1 transcription factor is required for Stat1-mediated T-bet expression and IgG2a class switching in mouse B cells. *Blood* 119, 4174–4181 (2012).

107. Storey, J. D. et al. Gene-Expression Variation Within and Among Human Populations. *Am. J. Hum. Genet.* 80, 502–509 (2007).

108. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511 (2013).

109. Chen, L. et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* 167, 1398–1414.e24 (2016).

110. Zhang, W. et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* 50, 613–620 (2018).

111. Grubert, F. et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* 162, 1051–1065 (2015).

112. Giang, S. & La Cava, A. IRF1 and BATF: key drivers of type 1 regulatory T-cell differentiation. *Cellular & molecular immunology* 14, 652–654 (2017).

113. Ohteki, T. et al. The Transcription Factor Interferon Regulatory Factor 1 (IRF-1) Is Important during the Maturation of Natural Killer 1.1 T Cell Receptor- α/β (NK1 T) Cells, Natural Killer Cells, and Intestinal Intraepithelial T Cells. *The Journal of Experimental Medicine* 187, 967–972 (1998).

114. Brien, J. D. et al. Interferon Regulatory Factor-1 (IRF-1) Shapes Both Innate and CD8⁺ T Cell Immune Responses against West Nile Virus Infection. *PLoS Pathog.* 7, e1002230 (2011).

115. Kano, S.-I. et al. The contribution of transcription factor IRF1 to the interferon-gamma-

- interleukin 12 signaling axis and TH1 versus TH-17 differentiation of CD4⁺ T cells. *Nat. Immunol.* 9, 34–41 (2008).
116. McGinnis, C. S. et al. MULTI-seq: Scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. doi:10.1101/387241
117. Gasperini, M. et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* 176, 1516 (2019).
118. Yang, K., Katcoff, A. & Uhler, C. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. in *International Conference on Machine Learning* 5541–5550 (2018).
119. Tong, A. H. Y. Global Mapping of the Yeast Genetic Interaction Network. *Science* 303, 808–813 (2004).
120. Davierwala, A. P. et al. The synthetic genetic interaction spectrum of essential genes. *Nat. Genet.* 37, 1147–1152 (2005).
121. Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A. G. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat. Genet.* 38, 896–903 (2006).
122. Lee, M. N. et al. Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. *Science* 343, 1246980–1246980 (2014).
123. Lee, D., Cheng, A., Lawlor, N., Bolisetty, M. & Ucar, D. Detection of correlated hidden factors from single cell transcriptomes using Iteratively Adjusted-SVA (IA-SVA). *Sci. Rep.* 8, 17040 (2018).
124. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* 34, 184–191 (2016).

125. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595 (2010).
126. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079 (2009).
127. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993 (2011).
128. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15 (2018).
129. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
130. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420 (2018).
131. Dewey, M. *metap: Meta-analysis of significance values*. R Package Version 0. 7. Available online: <https://cran.r-project.org/package=metap> (accessed on 20 October 2017) (2017).
132. Haghverdi, L., Büttner, M., Wolf, F. A., Büttner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848 (2016).
133. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, (2015).
134. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–9445 (2003).

Figures

Figure 1

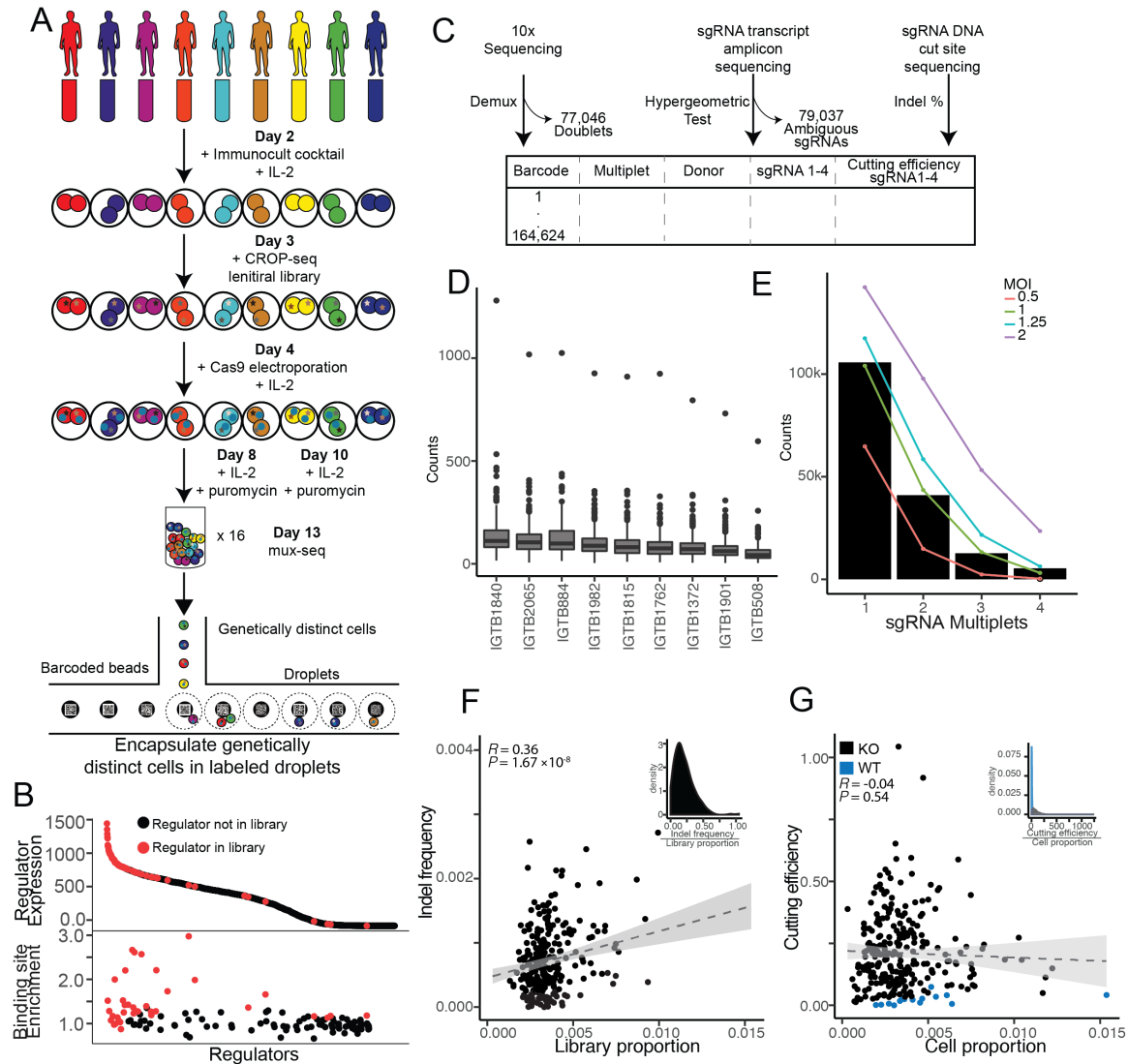


Figure 3.1: CRISPR perturbation screen in activated CD4⁺ T cells across donors

(A) Experiment overview. (B) Unbiased identification of candidate regulators from transcript abundance (top) and accessibility of binding sites (bottom) in activated CD4⁺ T cells⁵⁸⁵⁸. Targeted regulators are in red and all other regulators in the human genome are in black. (C) Data processing overview of 10X single-cell RNA-sequencing, sgRNA amplicon sequencing, and target loci DNA sequencing. (D) Total number of cells expressing each sgRNA per donor. (E) Observed distribution of cells with 1-4 sgRNA (black bars) and expected Poisson distributions at a MOI of 0.5 (pink), 1 (green), 1.25 (blue), and 2 (purple). (F) For each sgRNA, cutting efficiency (inset) is estimated as the ratio of indel frequency at the targeted locus (y-axis) and sgRNA frequency in the pool (x-axis). (G) For each wildtype (WT: blue) and knockout (KO: black) sgRNA, ratio (inset) of cutting efficiency (y-axis) and the proportion of cells expressing the sgRNA (x-axis).

Figure 2

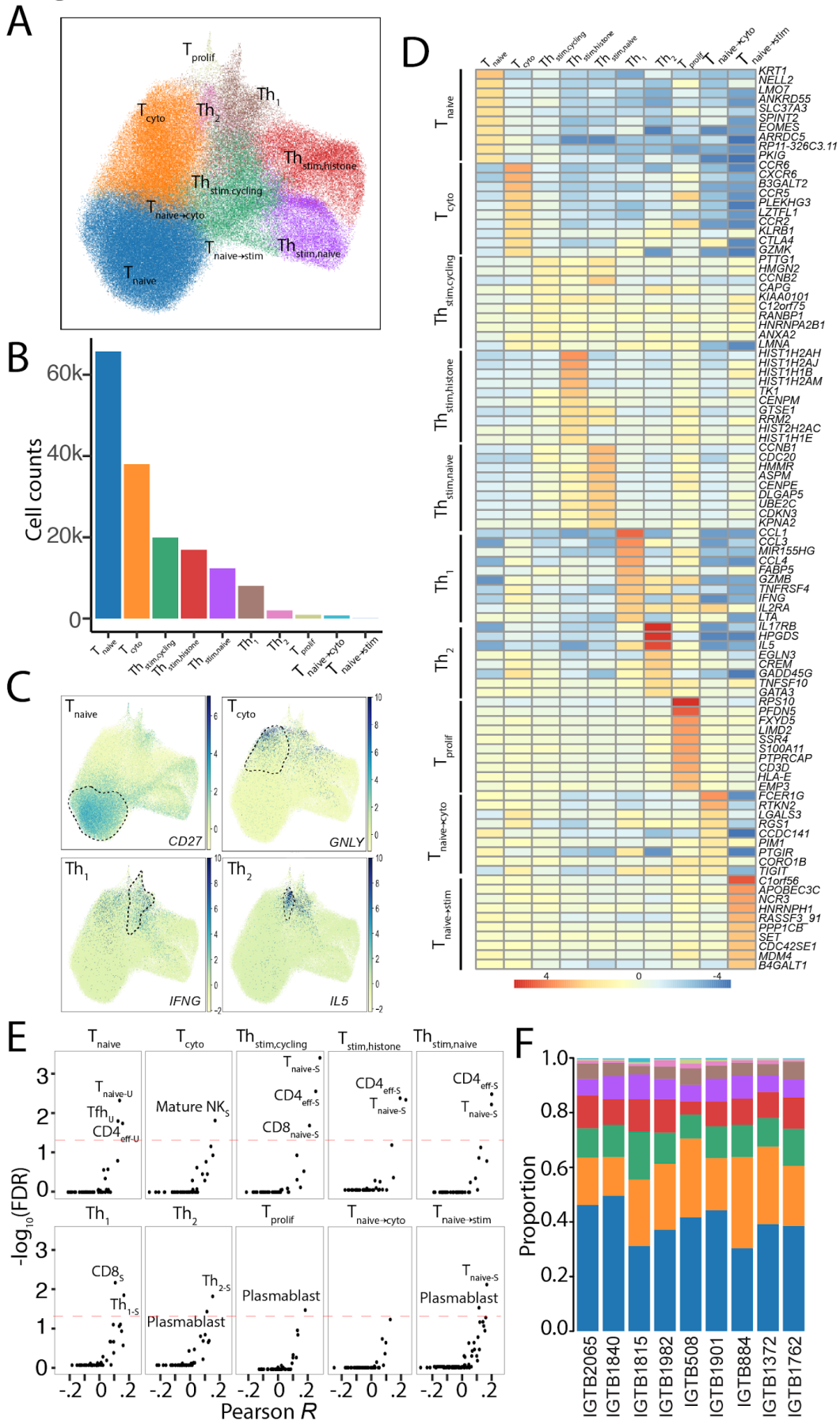


Figure 3.2: Heterogeneity of activated CD4⁺ T cells

(A) UMAP of activated CD4⁺ T cells. Each color represents an identified Leiden cluster. **(B)** Number of cells per cluster, colors correspond to the populations in (A). **(C)** Feature plots of normalized expression in UMAP coordinates of *CD27* (naive T; top left), *GNLY* (T_{cyto}; top right), *IFNG* (Th₁; bottom left), and *IL5* (Th₂; bottom right). **(D)** Log fold-change (with respect to all other clusters) of top 10 positively differentially expressed (DE) genes (row) per cluster (column). **(E)** For each cluster, correlation of average log fold-change of DE genes to sorted bulk RNA sequencing transcriptomes⁶²⁶² (x-axis) versus $-\log_{10}(\text{FDR})$ (y-axis). **(F)** Cluster proportions (y-axis) across nine donors (x-axis). Each color corresponds to the population in (A).

Figure 3

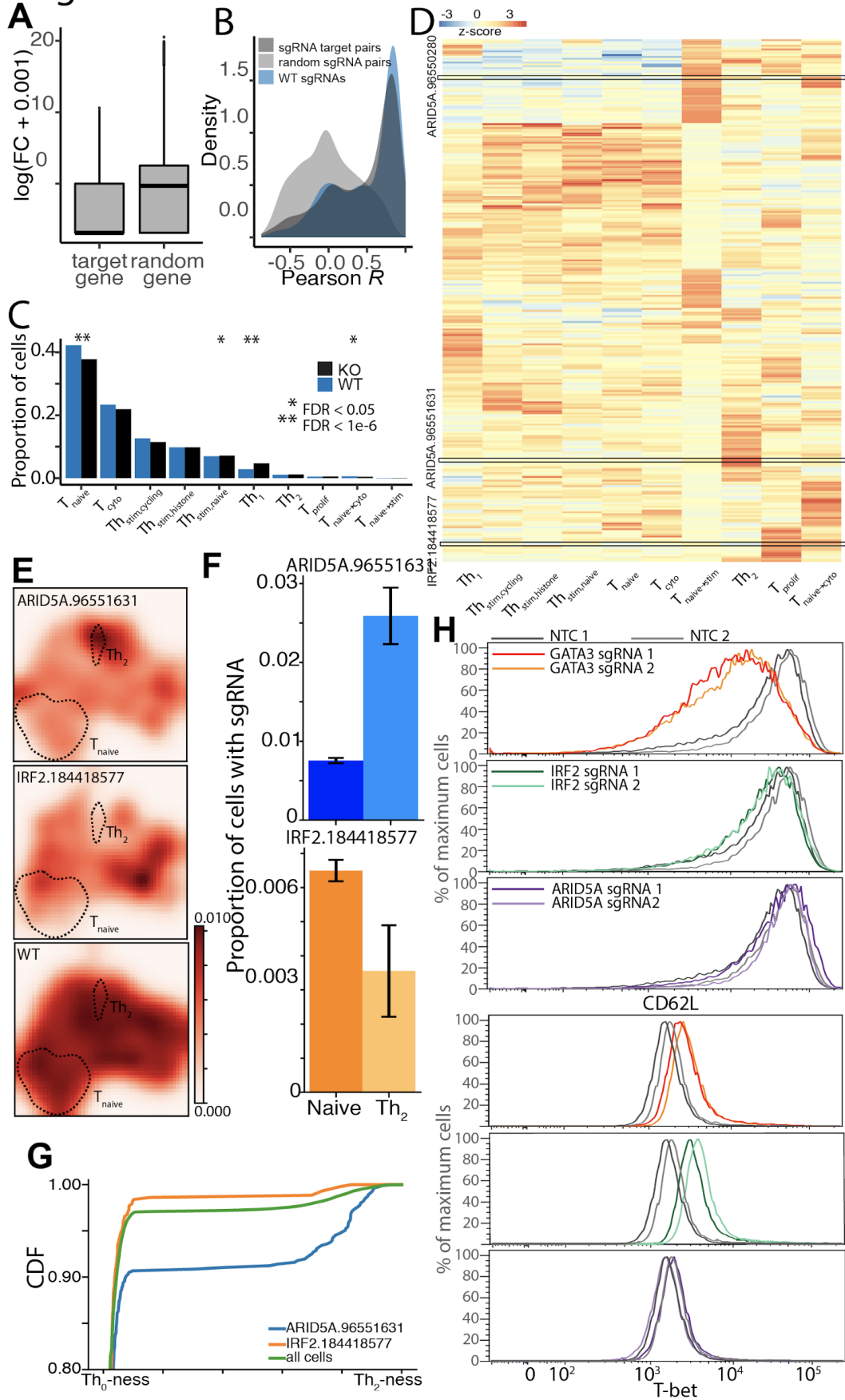


Figure 3.3. Regulator perturbations drive T cell polarization and maintenance

(A) Boxplot of fold change in expression of regulator targeted by sgRNA (left) and random gene (right) in cells expressing each sgRNA. (B) Distribution of transcriptomic correlations between cells expressing KO sgRNAs targeting the same gene (dark grey), WT sgRNAs (blue) and random sgRNA (light grey). (C) Proportion of KO (black) and WT (blue) cells per cluster. * indicates FDR < 0.05 and ** indicates FDR < 1e-6. (D) Clustered heatmap of sgRNA enrichment or depletion (z-score) across clusters. Red indicates a positive z-score and blue indicates a negative z-score. (E) Density of KO cells expressing sgRNA targeting *ARID5A* *ARID5A*-targeting (cutsite: chr2:96551631); top), *IRF2* (*IRF2*-targeting (cutsite: chr4:184418577); middle), and WT (top) sgRNAs in UMAP space. (F) Proportion of cells expressing *ARID5A.96551631* (top) or *IRF2.184418577* (bottom) in T_{naive} and Th₂ clusters. (G) Empirical cumulative distribution function (ECDF) of the estimated diffusion pseudotime of cells expressing sgRNA *IRF2.184418577* (orange), cells expressing sgRNA *ARID5A.96551631* (blue), and all cells (green). The shape of the ECDF reflects the enrichment of the guide along the pseudotime axis. (H) FACS validation. Distribution of cells expressing Th₂ marker CD62L⁺ (top 3 panels) and Th₁ marker T-bet⁺ (bottom 3 panels) electroporated with non-targeting control sgRNAs (grey), *GATA3*-targeting sgRNAs (red and orange), *IRF2*-targeting sgRNAs (light and dark green), or *ARID5A*-targeting sgRNAs (light and dark purple).

Figure 3.4. Perturbations and single cell analysis reveal transcription factor interactions

(A) Cartoon of detecting genetic interactions between regulators by comparing magnitude of correlation between KO and WT cells. In cooperative activation, magnitude of positive correlation decreases; in cooperative repression, magnitude of negative correlation decreases; in competitive activation, magnitude of positive correlation increases; in competitive repression, magnitude of negative correlation increases. (B) Number of genes downstream of each interacting regulator pair. (C) Network of interaction regulators known to affect T cell function. Solid line indicates known interactions; dashed indicates predicted interactions; dotted indicates known but undetected interactions. The colored edges indicate the number of downstream genes. (D) Ratio of identified target genes with both predicted binding sites to those without (y-axis) within a window size around the TSS (x-axis). (E) Distribution of regulator interactions found by subtype. (F) On top, distribution of magnitude of correlations between *GTF3A* and downstream genes for WT cells (blue) and *CREM* KO cells (orange). On the bottom, distribution of magnitude of correlations between *CREM* and downstream genes for WT cells (blue) and *GTF3A* KO cells (green). (G) *CLUAP1* expression versus *CREM* (top) and *GTF3A* (bottom) expression, illustrating an example of competitive repression. (H) *XRNI* expression versus *MYC* (top) and *NFATC3* expression (bottom), illustrating an example of cooperative activation. For (G) and (H), trend lines reflect the coefficients fitted by the linear mixed effects model.

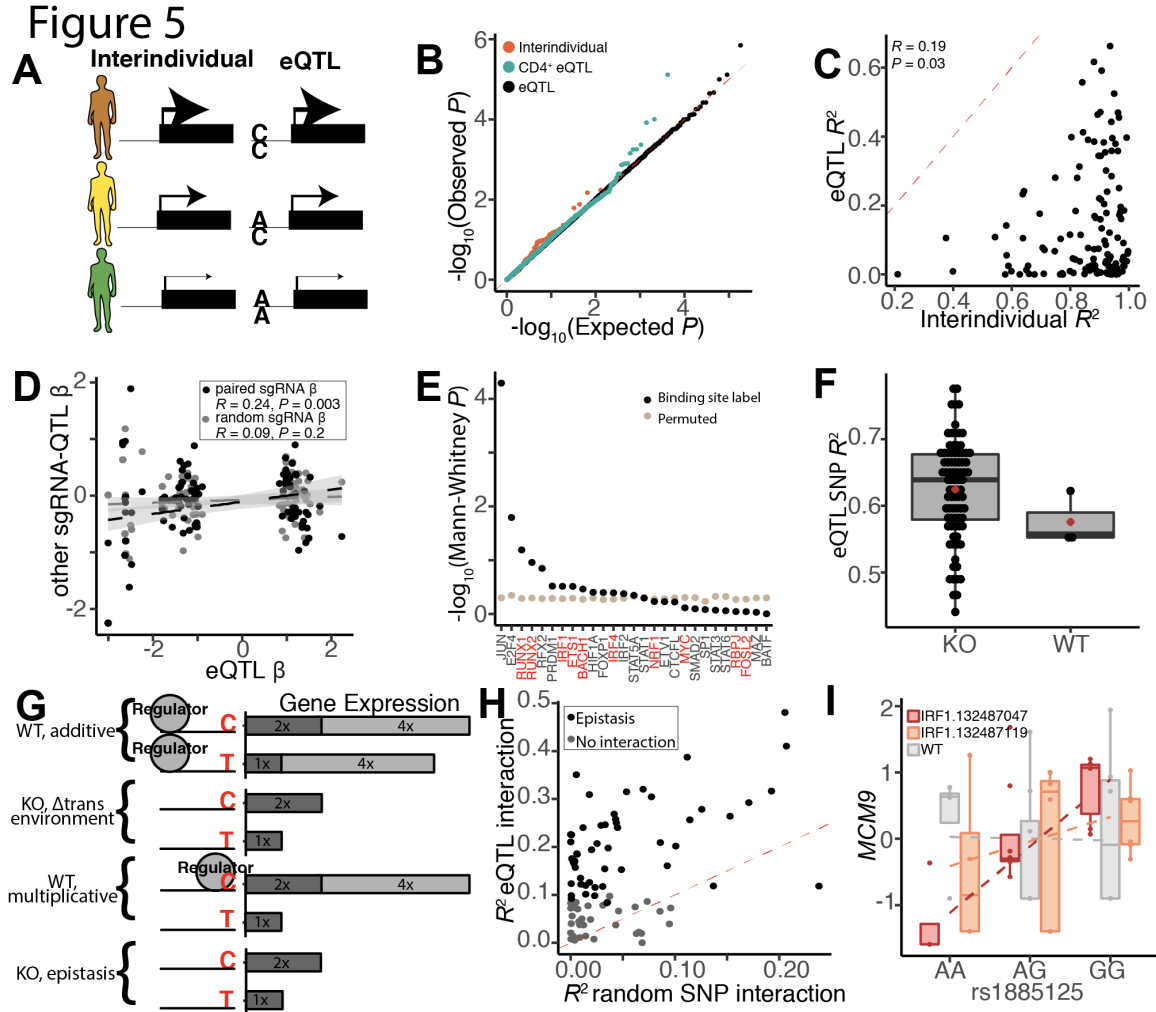


Figure 3.5: CRISPR perturbation modifies genetic effects on gene expression

(A) Cartoon of interindividual variation of gene expression (left) and possible genetic causes due to a SNP located in the cis regulatory region (right). Size of the arrow corresponding to the amount of gene expression for each donor. (B) eQTL QQ-plot. Each point represents an eGene empirical P value across sgRNAs (black), and those that are also interindividual genes (orange) and previously identified $CD4^+$ eGenes (teal). X-axis: expected P values. Y-axis: observed P values. Red dashed line is null. (C) Scatter plot of variance explained by interindividual (x-axis) and genetic (y-axis) variation, per eGene. (D) Scatter plot of eQTL effect sizes between pairs of sgRNAs targeting the same gene (black) or not (gray). (E) $-\log_{10}$ Mann-Whitney P (y-axis) of observing the ranked order of genes harboring binding sites for each regulator. Black is observed and tan is for permuted binding sites, and regulators in red have an eQTLs. (F) Genetic variance explained of significant eQTLs (y-axis) in KO and WT cells (x-axis). (G) Cartoon depicting genetic ablation impact on the effects of a donor with a C and another with a T allele on gene expression. If the regulator has an additive effect on gene expression in WT (first cartoon), then regulator absence changes the *trans* environment (second cartoon). If regulator interacts with a *cis* regulatory element to have a multiplicative effect (third cartoon), then regulator absence changes

the effect of the SNP (fourth cartoon). **(H)** Variance explained by *cis* x *trans* interaction for our eQTLs (y-axis) and a random SNP interaction (x-axis). In black are significant interactions. **(I)** Normalized *MCM9* expression (y-axis) is subsetted by donor genotype (x-axis) at rs1885125 in *IRF1*-targeting (chr5:132487047, red), *IRF1*-targeting (chr5:132487119, orange), and WT (grey) cells.

Materials and Methods

Study subjects and genotyping

Our samples were enrolled in PhenoGenetic study (age 18 to 56, average 29.9), as part of the Immvar cohort⁶⁰⁶⁰, which were recruited in the Greater Boston Area. Each donor gave written consent to participate and were healthy, without any history of inflammatory disease, autoimmune disease, chronic metabolic disorders or chronic infectious disorders. We genotyped 56 Caucasian samples on the OmniExpressExome54 chip, and excluded 2080 SNPs with a call rate <90% (0.22% of total), 1521 SNPs with Hardy Weinberg $P < 0.0001$ (0.16%) and 259,860 SNPs with MAF < 0.01 (27.04%) out of the total 960,919 SNPs profiled. The Michigan Imputation Server was used to impute these genotypes with the Haplotype Reference Consortium Panel Version r1.1. After genotype imputation had 5,324,560 SNPs, which were then subsetted for our nine donors.

Regulator target identification

Our library contained targeted 140 regulators (transcription factors and RNA-binding proteins) with 2 sgRNAs each. Each regulator was unbiasedly chosen using gene expression and accessibility data from activated CD4⁺ T cells in 95 and 105 healthy donors⁵⁸⁵⁸. To get the highly expressed regulators using RNA-seq data, we performed a TMM normalization and took the upper quartile of highly expressed genes and subsetted those that were regulators. To get the regulators with highly accessible binding sites using ATAC-seq data, we enriched for all binding sites on the HOMER database¹⁰⁴¹⁰⁴ in activated accessible chromatin regions. We took the union of the highly expressed regulators and accessible binding sites, for a total of 140 regulators (**Fig. 3.1B**).

CROP-seq library generation

The backbone plasmid used to clone the CROP-Seq library was CROPseq-Guide-Puro⁴⁰⁴⁰, purchased from Addgene (Addgene. Plasmid #86708). We used two sgRNAs oligo sequences from the Brunello library¹²⁴¹²⁴

for each of our chosen 140 regulators. Oligos for the sgRNA library were purchased from Integrated DNA Technologies (IDT) and cloned into the CROPseq plasmid backbone using the methods described by Datlinger et al. (2017)⁴⁰⁴⁰. Lentivirus was produced using the UCSF ViraCore.

SLICE experiment and sequencing

Primary human CD4⁺ T cells were isolated from peripheral blood mononuclear cells (PBMCs) by magnetic negative selection using the EasySep Human CD4⁺ T Cell Isolation Kit (STEMCELL, Cat #17952). Cells were cultured in X-Vivo media, consisting of X-Vivo15 medium (Lonza, Cat #04- 418Q) with 5% Fetal Calf Serum, 50mM 2-mercaptoethanol, and 10mM N-Acetyl L-Cysteine. On the day of isolation (Day 1), cells were rested in media without stimulation for 24 hours. The day after isolation (Day 2), cells were stimulated with ImmunoCult Human CD3/CD28 T Cell Activator (STEMCELL, Cat #10971) and IL-2 at 50U/mL. 24 hours post stimulation (Day 3), 1 uL of lentivirus was added directly to cultured T cells and gently mixed. Following 24 hours (Day 4), cells were collected, pelleted, and washed in PBS twice. Then, cells were resuspended in Lonza electroporation buffer P3 (Lonza, Cat #V4XP-3032). Cas9 protein (MacroLab, Berkeley, 40mM stock) was added to the cell suspension at a 1:10 v/v ratio. Cells were transferred to a 96 well electroporation cuvette plate (Lonza, cat #VVPA-1002) for nucleofection using the Lonza Nucleofector 96-well Shuttle System and pulse code EH115 (Lonza, cat #VVPA-1002). Immediately after electroporation, pre-warmed media was added to each electroporation well, and 96-well plate was placed at 37 degrees for 20 minutes. Cells were then transferred to culture vessels in X-Vivo media containing 50U/mL IL-2 at 1e6 cells /mL in appropriate tissue culture vessels. Two days later, 1.5ug/mL Puromycin was added in culture media for selection. Cells were expanded every two days, adding fresh media with IL-2 at 50U/mL. Cells were maintained at a cell density of 1e6 cells /mL. On the final day (Day 13) of the experiment, cells from each of the nine donors were counted using Vi-CELL XR and pooled at equal numbers to obtain a final 180,000 cells in 60 uL of PBS (**Fig. 2.1A**). The pooled cells were then processed by UCSF Institute for Human Genetics (IHG) Genomics Core using 16 wells of 10X Chromium

Single Cell v2 (PN-120237), as per manufacturer’s protocol, with each well being separately index. The final library was sequenced on two lanes on the Nova-seq for a total of 6.7B reads.

10x transcriptome alignment

Each 10x well was separately aligned to our personalized reference, which contained the hg19 transcriptome and our 280 sgRNA sequences using cellranger “count” function. Our reference sgRNA sequences contained the U6 plasmid promoter on the 5’ and sgRNA scaffold on the 3’, such that our sgRNA reference sequences were as follows:

```
TATGCTTACCGTAACTTGAAAGTATTTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACG
AAACACCG          -          20          bp          gRNA          -
GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGG
CACCGAGT. Using the cellranger “aggr” function, we aggregated all 16 wells into one combined dataset.
```

Demuxlet

Per well, demuxlet⁴¹⁴¹ was run on all 737,280 raw error corrected barcodes with 10x cellranger bam file, using a 1% genotype error rate, 0.5 alpha, 255 minimum mapping quality, 0 minimum distance to tail. We combined all 16 demuxlet runs, then we used the “BEST” column to identify droplet multiplet and the “SNG.1ST” column to identify the donor of origin.

sgRNA amplicon sequencing and analysis

For the sgRNA amplicon sequencing, donors were separately amplified and barcoded using a two-step PCR protocol. First, each donor was divided into 8 PCR reactions with 0.1ng template of cDNA. Each 25mL reaction consisted of 1.25mL P5 forward primer, 1.25mL Nextera Read 2 reverse primer, priming to the U6 promoter to enrich for guides, 12.5mL NEBNext Ultra II Q5 Master Mix (NEB, cat #M0544L), 0.1ng template, and water to 25mL. The PCR cycling conditions were: 3 minutes at 98C, followed by 10 s at 98C, 10 s at 62C, 25 s at 72C, for 10 cycles, and a final 2 minute extension at 72C. After the PCR, all reactions

were pooled for each donor and purified using Agencourt AMPure XP SPRI beads (cat #A63881) per the manufacturer's protocol. Next, 1mL was taken from each purified PCR product to go into a second PCR for sequencing index. Each reaction included 1mL of PCR product, 12.5 mL NEBNext Ultra II Q5 Master Mix (NEB, cat #M0544L), 1.25mL P5 forward primer, 1.25mL Illumina i7 primer, and water to 25mL. The PCR cycling conditions were: 3 minutes at 98C, followed by 10 s at 98C, 10 s at 62C, 25 s at 72C, for 10 cycles, and a final 2 minute extension at 72C. After the PCR, all reactions were SPRI purified and quantified using the Qubit dsDNA high sensitivity assay kit (Thermo Fisher Scientific, cat# Q32854) and run on a gel to confirm 500bp size.

The sgRNA amplicon sequencing library was sequenced paired-end on one lane HiSeq 4000, resulting in 171 million mapped paired-end reads. This library was similarly aligned to our personalized reference containing our 280 sgRNA sequences using 10x "count" function. We generated a read count matrix of 320,708 barcodes x 280 sgRNA matrix.

sgRNA identification

In order to identify the sgRNAs present in each cell, we performed a series of binomial test of enrichment for the top 4 sgRNAs present each of our 320,708 cells. Let n be the total number of reads mapping to any sgRNA in a given cell. Let c_1, c_2, c_3, c_4 be the number of reads of the top four sgRNA in a given cell, with c_1 representing the sgRNA with the most reads. We performed a binomial test with c_1 successes from trials $= n - c_2 - c_3 - c_4$, with $1/280$ probability of success. This test for the enrichment of the sgRNA with c_1 counts, disregarding the other three of the top four sgRNAs. We performed a similar test with the guide with c_2 read count, performing a binomial test of c_2 successes from trials $= n - c_1 - c_3 - c_4$ trials, with $1/280$ probability of success. The binomial tests statistics were Bonferroni corrected $P < 0.05$ (**Table S3.2**).

Cell identification

To identify our cell barcodes, we used our combined transcriptomic dataset and our combined demuxlet output, as previously mentioned. We filtered out cell-containing droplets that contained less than 10 SNPs from our demuxlet results, totaling 320,708 cells. 10 SNPs was the threshold at which the number of cell-containing barcodes and then number of cell-containing barcodes with at least 1000 UMIs no longer increased (**Fig. S3.1-3**). Cells were filtered for demuxlet identified singlets, which resulted in a 24% doublet rate, and cells where we could not identify the sgRNA (25%), resulting in 164,624 cells (**Fig. 3.1C, Table 3.2**).

sgRNA cutting efficiencies

Genomic DNA was isolated from cell pellets using the Promega Wizard Genomic DNA Purification Kit (cat #A1120). Amplification was performed as described by Paragon User Guide (CleanPlex Custom NGS Panel). Briefly, gDNA template was added to a multiplex PCR reaction with 2 uL of 5X Paragon PCR Mix, 2 uL of 5X Primer Pool, gDNA and Nuclease-Free Water. Each sgRNA amplicon was designed to be ~200bp, centering around the cutsite(**Table S3.1**). The PCR cycling conditions were: 10 minutes at 95C, followed by 15 cycles of 15 seconds at 98C and 5 minutes at 60C. Following the PCR, amplified DNA was purified using Magnetic Beads from Paragon and subjected to a digestion reaction (CP Reagent Buffer, CP Digestion Reagent) to remove nonspecific PCR products. A post-digestion purification was performed, followed by a second PCR reaction to amplify and index libraries. Second PCR reaction contained 5X Second PCR Reaction Master Mix, purified DNA from previous step, and i5/i7 Indexed PCR Primers for Illumina. The PCR cycling conditions were: 10 minutes at 95C, followed by 10 cycles of 15 seconds at 98C and 75 seconds at 60C. Finally, DNA was purified using magnetic beads from Paragon and subjected to next-generation sequencing.

Paragon sgRNA amplicon DNA sequencing analysis

Per donor, our sgRNA DNA sequencing was aligned to hg19 using `bwa -mem`^{125,125}. Using the “mpileup” function from the samtools suite of tools^{126,127,126,127} we estimated the number of indels and reads per basepair +/-200bp around each sgRNA cutsite(**Table S3**). Per sgRNA and per donor, we estimated indel frequency as the maximal number reads with indels +/- 5 target cutsite divided by total reads covering a cutsite. Cutting efficiencies for each sgRNA were then estimated as the indel frequency x the proportion in the original sgRNA library.

sgRNA WTs

Using our sgRNA amplicon DNA sequencing we identified WT sgRNAs by calculating a z-score for each sgRNA. Let i be a sgRNA, and c is the cutting efficiency of sgRNA _{i} , and p is the proportion of cells with sgRNA _{i} , then WT sgRNAs were identified as z-scores of p / c with $P < 0.05$. In total we identified 14 WT sgRNAs, which have the maximum cutting efficiencies was $< 5\%$ with at least 484 cells. To estimate WT proportions (**Fig. 2.2E**), we calculated a hypergeometric P comparing the number of WT cells and KO cells per cluster as a function of all total WT and KO cells.

Single cell normalization

We normalized our 164,623 cells using the scanpy^{128,128} suite of tools. Working with our 32,739 genes x 164,1623 cell matrix, we calculated the percentage of mitochondrial contamination, filtered out genes using “filter_genes” with options “min_counts=1” and then normalized using “normalize_per_cell”, which normalizes each gene cell by the total counts for that cell. We further filtered our gene list for variable genes, which we identified by subsetting the cells from one well, then calling “filter_genes_dispersion” with options “min_mean=0.0125, max_mean=3, min_disp=0.5”. Using only one well safe guarded us from variable genes due to well to well batch effects. Subsetting our gene x cell matrix for our 2,189 variables genes, we re-normalized our cells for total sequencing, log transformed, and then regressed out mitochondrial contamination (as previously calculated), the 10x well, and total UMIs. Finally, we scaled

the data to have mean = 0 and variance = 1. To reduce the data down two dimensions for visualization, we ran UMAP¹²⁹¹²⁹ through scanpy¹²⁸¹²⁸ using our 2,189 variable genes across our 164,623 cells(Fig. 3.3A).

Leiden clustering

We performed unbiased cluster detection on our 164,623 cells using our 2,189 variable genes, using the scanpy¹²⁸¹²⁸ suite of tools. First, we calculated a neighborhood graph¹²⁹¹²⁹, second, performed leiden clustering⁶⁷⁶⁷ at 0.68 resolution. We compared our leiden clusters at 0.5, 0.6, 0.68, 0.75, and 1 resolutions and found that our clusters called at a 0.68 resolution were qualitatively the most similar to the gene expression patterns of T cell subtype markers (Th₂: IL5, Th₁:IFNG, naive T: CD27).

Differential analysis

For cluster differential expression, we used our 2,189 variable gene x 164,1623 cell normalized matrix (as previously described), which was subsetted by donor. Using the “FindAllMarkers” function from the Seurat package¹³⁰¹³⁰ using options “min.pct=0, logfc.threshold=0, min.cells.gene=0, min.diff.pct=0, return.thresh=1” we calculated the log fold-change for all 2,189 genes per donor. We then performed a meta-analysis to estimate a meta *P* per gene across all nine donors using the metap package in ¹³¹¹³¹ and we averaged the log fold-changes across all nine donors to get an average log fold-change.

For sgRNA differential expression, we again started with our 2,189 variable gene x 164,1623 cell normalized matrix, which was subsetted by donor and by sgRNA. Using the “FindMarkers” function from the Seurat package¹³⁰¹³⁰ using options “min.pct=0, logfc.threshold=0, min.cells.gene=0, min.diff.pct=0, return.thresh=1” we compared cells containing each KO sgRNA to our WT cells, per donor. Again, we calculated a meta *P* per gene and log fold-changes were averaged across the nine donors per gene per sgRNA.

Comparison to sorted, bulk T cell subtypes

Using the RNA-seq dataset from Calderon et al. 2018⁶²¹³², we averaged the gene expressions across all donors for each cell type and to normalize it, we log transformed the expression counts, subtracted out the median count per sample, and then standardized by the gene (mean = 0, variance = 1). For each cell type we took the top 300 most highly expressed genes, and then correlated those 300 genes to their respective log average fold-changes from the cluster differential expression analysis. We calculated a *P*, which was then FDR adjusted.

Th₂ validation experiment

PBMCs were sourced from anonymized female Caucasian donors and were purified from whole blood by Ficoll gradient. Cells were frozen in 10% DMSO in FBS in a cryostorage vessel for one day at -80°C before being moved into a liquid nitrogen tank. Frozen PBMCs were quickly thawed in a 37°C water bath and slowly diluted with RPMI1640(Sigma, R0883) supplemented with 10% FBS(HI-FBS; Invitrogen, catalog 10438026), 1 mM GlutaMAX(Invitrogen; catalog 35050061), 100 U/ml penicillin and 100 mg/ml streptomycin(Invitrogen; catalog 15140122). Cells were pelleted at 300 xg for 5 minutes before being washed with SepMate Buffer(Stem Cell; catalog 20144) for naïve CD4⁺ isolation.

Naïve CD4⁺ T cells were isolated using an EasySep™ Human Naïve CD4⁺ T Cell Isolation Kit II(catalog 17555) according to the manufacturer's protocol. Harvested naïve CD4 cells($1-3 \times 10^6$) were plated on a 24 well plate with 1 ml of supplemented RPMI1640 with 50 ng/ml IL-2(R&D; catalog 202-IL-010) and 25 ul/ml of ImmunoCult™ Human CD3/CD28 T Cell Activator(StemCell Technologies, catalog 10971) for 24 hours. For differentiation modulation, activated T cells were split into 16 groups($5 \times 10^4-2 \times 10^5$) per donor (n=7) for 8 guides and two polarizing conditions(Th₂, activated CD4⁺). For RNP electroporation, 4 ul of 160 uM of tracr RNA (Dharmacon, catalog U-002005-50) was incubated with 4 ul of 160uM sgRNA (Dharmacon) for 30 minutes at 37°C. Following the annealing of sgRNAs, 8 ul of 40 uM Cas9-NLS(MacroLab, Berkeley, 40 uM stock) was added to each sgRNA mixture for 15 minutes at 37°C. 3 ul

of each complete RNP was added to a 96 U-bottom plate(Genesee Scientific; catalog 25-221) alongside 1 ul of ssODN Alt-R Cas9 Electroporation Enhancer(IDT: catalog 1075916; 100 uM). Cells were then pelleted at 90 x g for 5 minutes at room temperature. Cells were resuspended in 20 ul of P3 buffer(Lonza; catalog V4SP-3096) and transferred to the aliquoted RNP mixtures. 20 ul of the cell-RNP mixture was added to a 96-well electroporation plate(Lonza, V4SP-3096) and electroporated on a 4D nucleofector system(Lonza) with program EH-115. Cells were quickly rescued by adding 100 ul of supplemented RPMI with 50 ng/ml IL-2 dropwise to electroporated cells. Cells were incubated for 10 minutes in a 37°C incubator with 5% CO₂ before being transferred into a 96 well U-bottom plate and brought up to a total volume of 200 ul with 50 ng/ml IL-2 for activated CD4⁺ groups and 10 ng/ml IL-4(R&D, catalog 204-IL-010), 2 ug/ml anti-IL-12 antibody(R&D, catalog MAB219-500), and 2 ug/ml anti-IFN-G antibody(R&D, catalog MAB285-500) for Th₂ and incubated for 24 hours. Media with appropriate cytokines were refreshed every 2-3 days and cell density was adjusted to 1x10⁶ cells/well with every media change. After a total of 14 days, cells were harvested for flow cytometry.

For maintenance modulation, naïve T cells were isolated as described. Cells were incubated with supplemented RPMI1640 with 50 ng/ml IL-2 and 25 ul/ml Immunocult for 72 hours in a 96 well U-bottom plate in 200 ul. Cells were then split into a Th₂ and activated CD4⁺ conditions as described and cultured for 1 week, with cytokine supplemented media every 2-3 days. Cells were then split into 16 groups and electroporated as previously described for each guide. Cells were harvested 7 days after electroporation for flow cytometry.

For flow cytometry, cells were stimulated with PMA and ionomycin with Brefeldin A (Leukocyte activation cocktail with BD Golgiplug; BD Bioscience; catalog 550583) for five hours before staining. Cells were then washed twice with 1% BSA in PBS by pelleting cells at 300 x g for 5 minutes at 4°C and resuspended in staining buffer (Biolegend cell staining buffer; Biolegend; catalog 420201) with 5 ul of Trustain FCX(Biolegend; 422302) and incubated for 5 minutes on ice. Cells were then stained with 5 ul of each

extracellular antibody(CD4-FITC, BV785-CD62L, AF700-CD127, BV711-CRTH2, BV510-CCR5; Biolegend) for 30 minutes on ice in the dark in a total reaction volume of 100 ul. After staining, cells were washed twice with 1% BSA in PBS. Cells were fixed and permeabilized using an eBioscience FOXP3/Transcriptional factor staining kit(eBioscience, 00-5523-00) per manufacturer's protocol. Permeabilized cells were stained with 5 ul of each intracellular marker(BV421-GATA-3, APC-T-bet, BV605-IL-4, PE-IFN-G; Biolegend) and incubated for 30 minutes on ice in the dark for a total reaction volume of 100 ul. Cells were washed twice with 1% BSA in PBS and resuspended in 200 ul of PBS before flow analysis. BD LSRII(Parnassus Flow Core, Grover) was used for flow acquisition and FlowJo 9 was used for analysis.

For TIDE validation, 10^4 cells were placed into 50 ul of Quickextract(Lucigen, catalog QE09050) and vortexed for 15 seconds. The cell solution was then incubated at 65C for 6 minutes and vortexed for another 15 seconds. The cell solution was then placed into a heat block at 98C for 2 minutes. The extracted gDNA was then stored at -20°C until amplification. Primers for targeted genes were designed to create a 700 bp amplicon, starting from 350 bp upstream of the cut site for the guide. Primers were designed using Primer-Blast(NCBI). Sequencing primers designed to be 200 bp upstream of the cut site were designed using the same tool. For amplification, 1 ul of gDNA solution was amplified using KAPA Hotstart HIFI Readymix (Kapa Biosystems, catalog KK2602). Generated amplicons were sent for sequencing with designed sequencing primers. Analysis of provided chromatograms were done using a TIDE web tool(<https://tide.deskgen.com>) and cutting efficiency was determined.

Target gene expression

We created a pseudobulked matrix, per sgRNA, per donor, and per cluster for a 32,739 genes x 17,845 samples, which was then filtered for our targeted regulators genes. Per sample, let i be each of our 140 targeted regulators, n is the total counts per sample, and r is counts for regulator $_i$. Now, let w counts for regulator $_i$ WT sample and m is total counts for the WT sample, then the fold-change $_i = (r_i / c) / (w_i / m)$.

As background, we randomly sampled a regulator for each sample and performed the same calculations (**Fig. 2.3A**).

Correlation of sgRNAs

We pseudobulked by sgRNA, by donor, and by cluster, for a total of 2,189 variables gene 17,845 samples x matrix. We normalized our matrix using a log2 transformed median normalization and then standardized across a gene(mean=0, variance=1). For every sample, we averaged the gene expression across our 9 donors and clusters, and then correlated the normalized, averaged transcriptome for every pair of sgRNAs targeting the same regulator. As background, we correlated 280 unpaired sgRNA (**Fig. 3.3B**).

sgRNA cluster enrichment/depletion

We subsetted our cells by those that contained KO sgRNAs and calculated the sgRNA proportion per cluster. Then, we calculated a z-score per cluster across all KO sgRNAs. KO sgRNAs that had a z-score > 1.5 were considered enriched in that cluster and those that had z-scores < -1.5 were considered depleted in that cluster (**Fig. 2.3C**). To visualize sgRNA enrichment and depletion, the UMAP space (as previously described) was partitioned into a grid of 50x50 rectangular pixels, and the density of cells with a specific guide was computed in each rectangle. Gaussian blurring with sigma 3 was applied to the UMAP density image (**Fig. 2.3D**).

Lineage trajectory

First, we estimated diffusion pseudotime from the naive T cells to the Th₂ cluster using the scanpy¹²⁸¹²⁸ implementation of Haghverdi et al. 2016¹³²¹³³. Using our normalized 2,189 gene x 164,623 cell matrix (as previously described), we filtered out cells that were not in either the naive T cell or the Th₂ cluster. PCA, neighborhood construction (500 neighbors, 40 PCs), and UMAP were re-run prior to applying the diffusion pseudotime (DPT) algorithm, all with default parameters. DPT is a random-walk-based distance that is computed based on simple Euclidian distances in the 'diffusion map space'. The diffusion map is a nonlinear

method for recovering the low-dimensional structure underlying high-dimensional observations¹³²¹³³. This algorithm assigns a single number to each cell, corresponding to the “time” that each cell has passed from a root cell. An empirical cumulative density plot was created using these estimated pseudotimes to detect and visualize distinct DPT profiles of cells containing different sgRNA.

Second, we used the scanpy implementation of the partition-based graph abstraction (PAGA) algorithm to quantify the connectivity of all of our cell clusters, approximating the overall cellular trajectory manifold. The default parameters for the PAGA algorithm was used, and connectivity > 0.3 was used for visualization.

Regulator - Regulator interaction model

We created ten technical replicates of pseudobulks by sgRNA and by donors, for a 2,189 variable gene x 5,040 sample matrix, which was normalized by estimating the proportion of total reads per gene for each sample, which was then multiplied by the median total reads across samples. Then, the data was log transformed and standardized(mean=0, variance=1). Out of the 140 regulators considered in our study, 37 were included when we selected for genes with highly variable expression using scanpy¹²⁸¹²⁸. To identify potential downstream genes for each regulator, we used a linear mixed model $\text{exp}(G) \sim \text{exp}(\text{regulator}) + \text{regulator_KO} + \text{exp}(\text{regulator}):\text{regulator_KO} + \text{intercept}$ and donor as a random effect, testing for the addition of the interaction term $\text{exp}(\text{regulator}):\text{regulator_KO}$ via a likelihood ratio test. Once potential downstream genes were identified for each regulator, pairs of regulators were formed and candidate regulator-pair, gene triplets were formed based on the intersection of potential downstream genes of those two regulators (regulator1 and regulator2). For every candidate regulator1, regulator2, G triplet, we test the interaction term in the following linear mixed models: $\text{exp}(G) \sim \text{exp}(\text{regulator1}) + \text{regulator2_KO} + \text{exp}(\text{regulator1}):\text{regulator2_KO} + \text{intercept}$, and $\text{exp}(G) \sim \text{exp}(\text{regulator2}) + \text{regulator1_KO} + \text{exp}(\text{regulator2}):\text{regulator1_KO} + \text{intercept}$, both using donor as a random effect. If the interaction terms are significant in both of the linear mixed models via likelihood ratio test, we call this regulator1, regulator2, G triplet an *interaction*.

Regulator - Regulator interaction binding site validation

Of the 37 regulators considered for the regulator interaction analysis, 18 had binding sites in the HOMER database. 31 interactions of regulator-pair gene triplets were identified for this validation, where both regulators are in the HOMER database. For each regulator-pair gene triplet in this set, we searched for binding sites of regulator1 and regulator2 upstream and downstream of gene G's TSS at various window length by using HOMER's `annotatePeaks.pl` program¹⁰⁴¹⁰⁴. Window lengths were ranged from +/- 1 to 5 kbps in intervals of 250 bps. For each window, the ratio of number of interactions with both binding sites in the window to the number of interactions without both binding sites in the window was calculated. The background was generated by taking all variable genes and randomly assigning pairs of those 18 regulators and applying the same procedure of looking for both binding sites in within the window around the TSS of each gene.

Interindividual variation analysis

We created two technical replicates of pseudobulks by sgRNA and by donors, for a 32,739 gene x 5,040 sample matrix. We filtered for genes with at least 10 counts and then further filtered for genes with a SNP +/- 100kb from the TSS with a minor allele frequency > 0.4, for a final 2,095 tested genes. We normalized each sgRNA separately, such that we subsetting our matrix 2,095 genes x 18 samples, where we calculated the percentage of total reads per gene and multiplied by the median total counts for all samples. Then, we log transformed the data and standardized it (mean=0, variance=1). For each sgRNA we also created a covariate file, containing donor and cutting efficiency, which was also standardized (mean=0, variance=1).

To test for interindividual variation, we used a linear mixed model, using the "Lme4" package¹³³¹³⁴. Per sgRNA we tested two models, our alternative model: $\text{expr} \sim \text{cutting efficiency} + (1|\text{donor})$, and our null model: $\text{expr} \sim \text{cutting efficiency}$. As Storey, et al. 2007¹⁰⁷¹⁰⁷ noted, including donor as a random effect properly accounts for donor variation, rather than a fixed effect. We calculated a likelihood ratio P to

determine if donor was significant. Using the package “r.squaredGLMM” function from the MuMIn R package we calculated the R^2 for each model, where the variance explained due to interindividual variation was calculated as the difference between the alternative and the null (R^2 interindividual variation = R^2 alternative - R^2 null).

We calculated empirical P -values per gene per sgRNA. We permuted the donors, while maintaining donor pairs, 1000 times per gene, for a total of 527,282,000 permutations. To calculate empirical P -values first, we filtered interindividual variation associations by those that converged, and then filtered our permuted P -values for duplicate P -values. The former filtering step was performed because we wanted to reduce our multiple testing burden and therefore did not want to include tests that did not converge. The latter filtering step was performed because if permuting the donors caused that specific model to not converge, and if that occurred multiple times, then that could inflate out statistics. Using the remaining, unique list of P -values we calculated empirical P -values using “empPvals” function from the q value package¹³⁴¹³⁵ with the option “pool=T”. Finally, we FDR adjusted our empirical P -values to determine significant interindividual associations.

eQTL analysis

Using the normalized expression matrices and covariate files for each sgRNA from our interindividual variation analysis we associated each gene to a genetic variation. As previously mentioned, we only tested variants that had a minor allele frequency > 0.4 and were +/-100kb around a TSS of a tested gene.

To detect eQTLs, we fit a linear mixed model, per sgRNA, where we fit two models, alternative: $\text{expr} \sim \text{SNP} + \text{cutting efficiency} + (1|\text{donor})$, and our null: $\text{expr} \sim \text{cutting efficiency} + (1|\text{donor})$. We performed a likelihood ratio and determine if the genetic statistical significance. Similarly, to our interindividual association analysis, we used the function “r.squaredGLMM” function from the MuMIn R package to calculate R^2 for each model. The variance explained due to genetics SNP $R^2 = R^2$ alternative - R^2 null.

To calculate empirical P -values, per sgRNA, we permuted our genotypes 1000 times per gene - SNP test. We performed 1000 permutations tests per sgRNA per gene, in total we performed 301,020,000 permutations. To calculate well calibrated empirical P -values, per sgRNA we pooled all P -values and calculated empirical P -values using the “empPvals” function from the q value package¹³⁴¹³⁵ with the “pool=T” option, and then FDR adjusted. To test for sgRNA specific eQTLs, we recalculated FDRs per gene across all 268 sgRNAs, filtering for sgRNAs that did not test that gene.

Binding site enrichment in eGenes

27 out of our 140 regulators are in the Homer database¹⁰⁴¹⁰⁴, therefore we parsed each gene for our 27 regulator binding site +/-100 kb around it's TSS. Per tested regulator, we ranked our eQTL associations by R^2 and compared our ranked list of genes that did and did not contain a binding site using a Mann-Whitney test. For each gene - regulator pair, we permuted the labels of genes that did and did not have the binding site 100 times, calculating a Mann-Whitney P per test, taking the average of the permuted P (**Fig. 3.5I**).

Bootstrapping variance explained

To overcome our unbalanced sample sizes between KO and WT sgRNAs, we performed sampled each KO eQTL (with replacement) to the depth of our WT eQTLs (three eQTLs), 100 times. Per bootstrap, we estimated the mean and standard deviation of the variance explained across the 3 sampled KO eQTLs (**Fig. 3.5K**).

Epistasis analysis

Using the normalized expression matrices and covariate files for each sgRNA from our interindividual variation and eQTL analysis we performed an interaction test for the 88 eQTLs. For each eQTL we

compared the KO sgRNA to a randomly sampled WT sgRNA condition. We fit two models, alternative: $\text{expr} \sim \text{SNP} * \text{sgRNA} + \text{cutting efficiency} + (1|\text{donor})$, and our null: $\text{expr} \sim \text{SNP} + \text{sgRNA} + \text{cutting efficiency} + (1|\text{donor})$. We performed a likelihood ratio and determine interaction statistical significance. We used the function “r.squaredGLMM” function from the MuMIn R package to calculate R^2 for each model. The variance explained due to the interaction is $R^2 = R^2 \text{ alternative} - R^2 \text{ null}$.

eQTL standard error simulations

Using an effect size of 0.5 and minor allele frequency of 0.5, we first simulated WT expression as the sum of the genetic effect, regulator effect, and an independent noise term ($\text{WT} = g * \text{beta} + \text{tf_expr} + \text{noise}$) and KO expression as the genetic effect and an independent noise term ($\text{KO} = g * \text{beta} + \text{noise}$). Next, we performed a linear regression on the WT and KO conditions and calculated the effect sizes, standard errors, and p-values for 1000 iterations.

Supplementary Figures

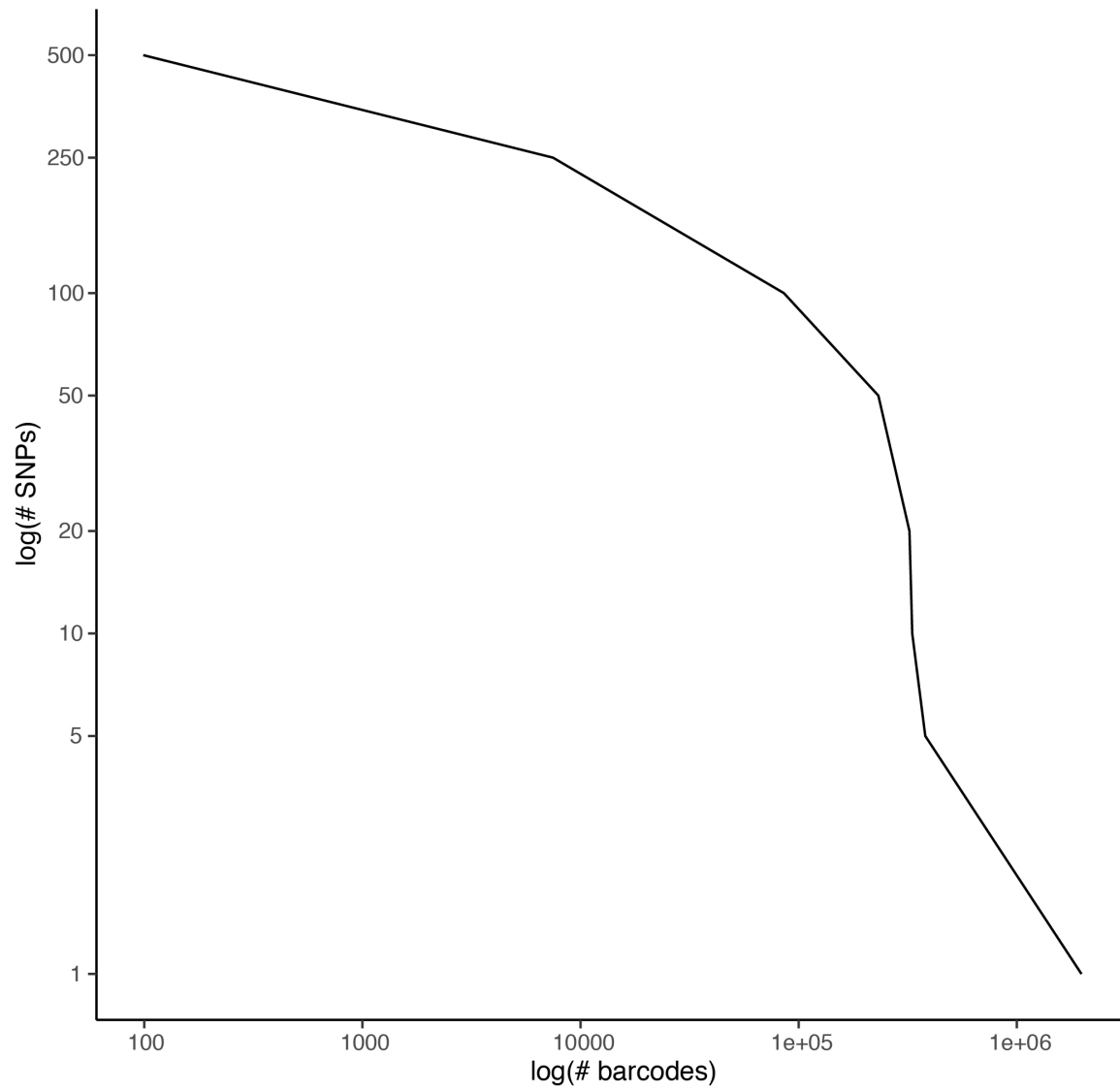


Figure S3.1. Number of cell-containing barcodes per number of SNPs.

We estimated the number of cell-containing droplets (x-axis) by the number of SNPs used by demuxlet to identify the donor of origin (y-axis).

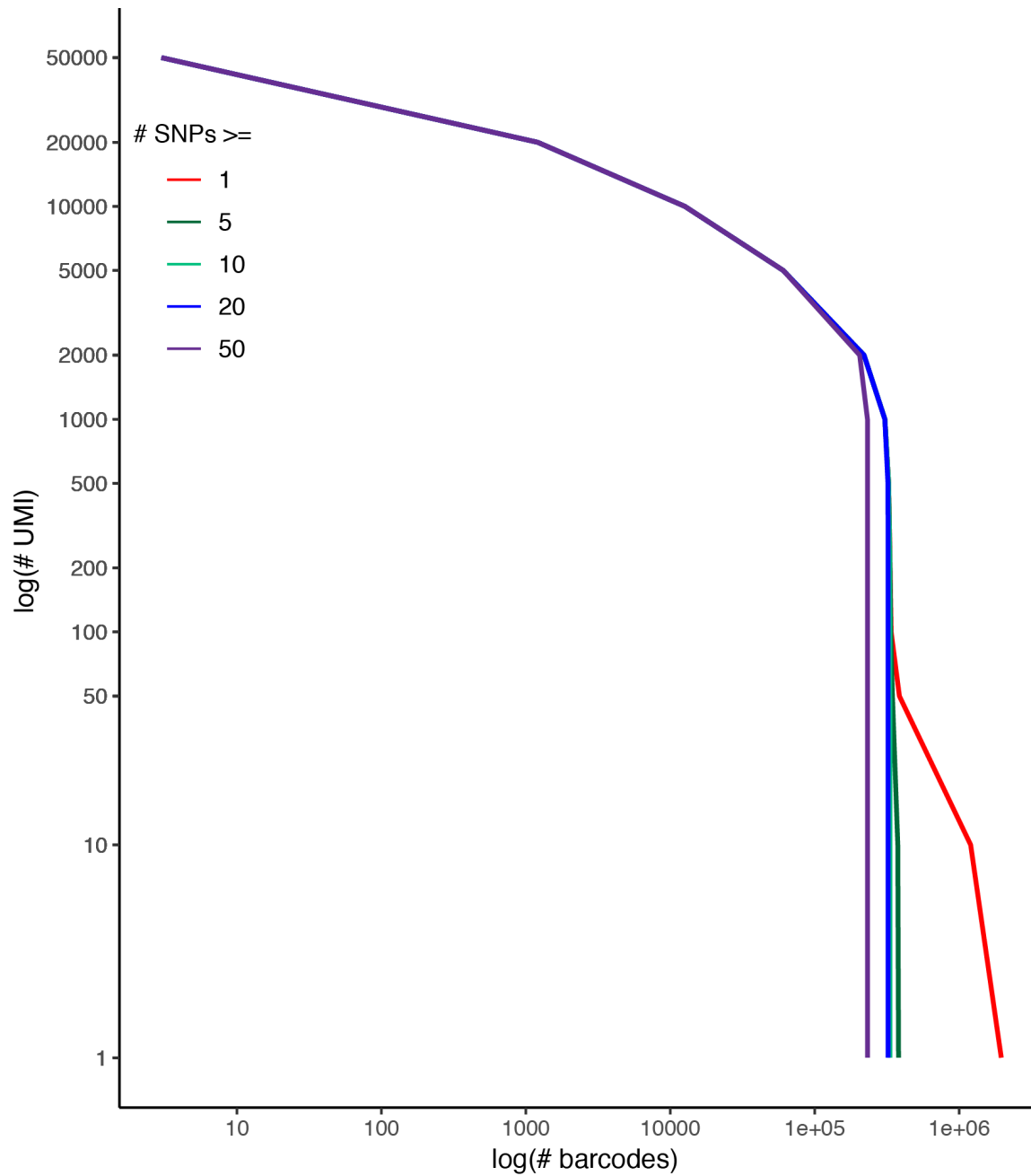


Figure S3.2. Number of cell-containing barcodes per number of UMIs.

We estimated the number of cell-containing droplets (x-axis) by the number of UMIs filtered for one (red), five (green), 10 (blue), and 50 (purple) SNPs used by demuxlet to identify the donor of origin (y-axis).

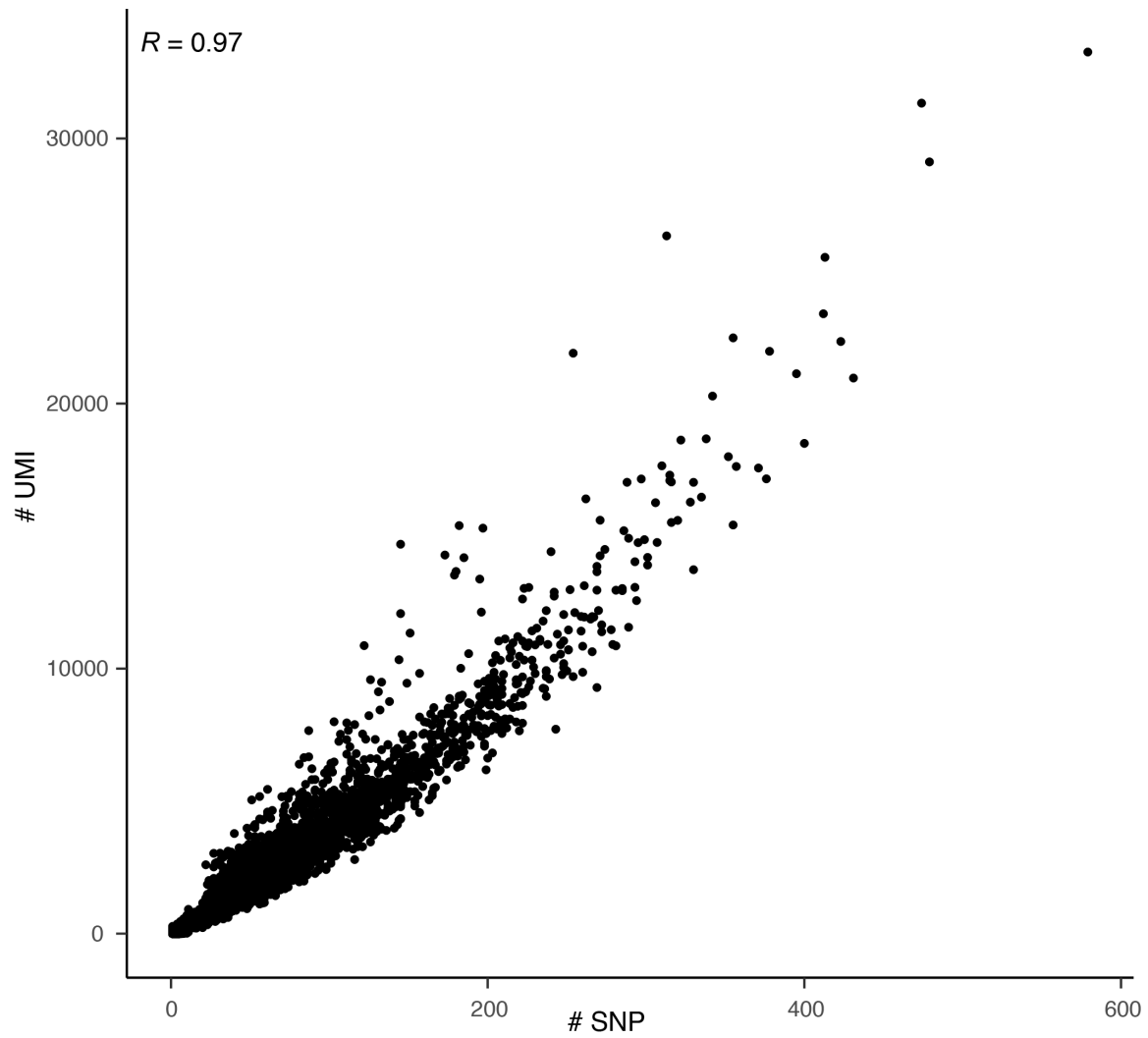


Figure S3.3. Correlation of number of SNP and nUMIs.

The number of SNPs (x-axis) and the number of UMIs (y-axis) estimated from demuxlet (Pearson $R=0.97$). Each point represented a cell-containing droplets, sampled to 20,000 cell-containing droplets.

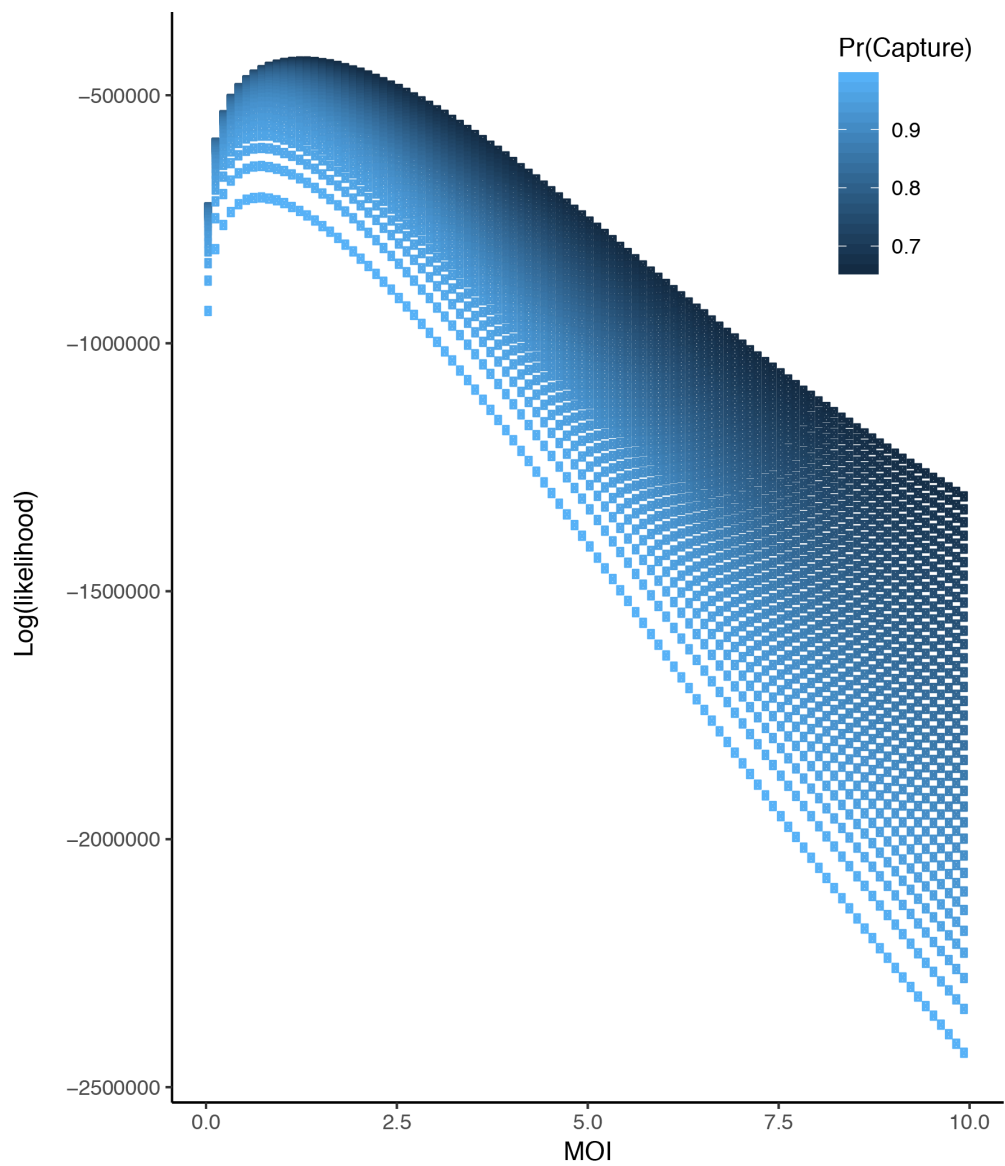


Figure S3.4. MOI probability.

Given a capture rate, we estimated the likelihood (y-axis) of seeing the transcript at a given MOI (x-axis).

Figure S3.5. sgRNA KO efficiency

We estimated the average sgRNA KO efficiency (x-axis) per sgRNA (y-axis). Each point represents the average KO efficiency and error bars are the standard deviations across donors.

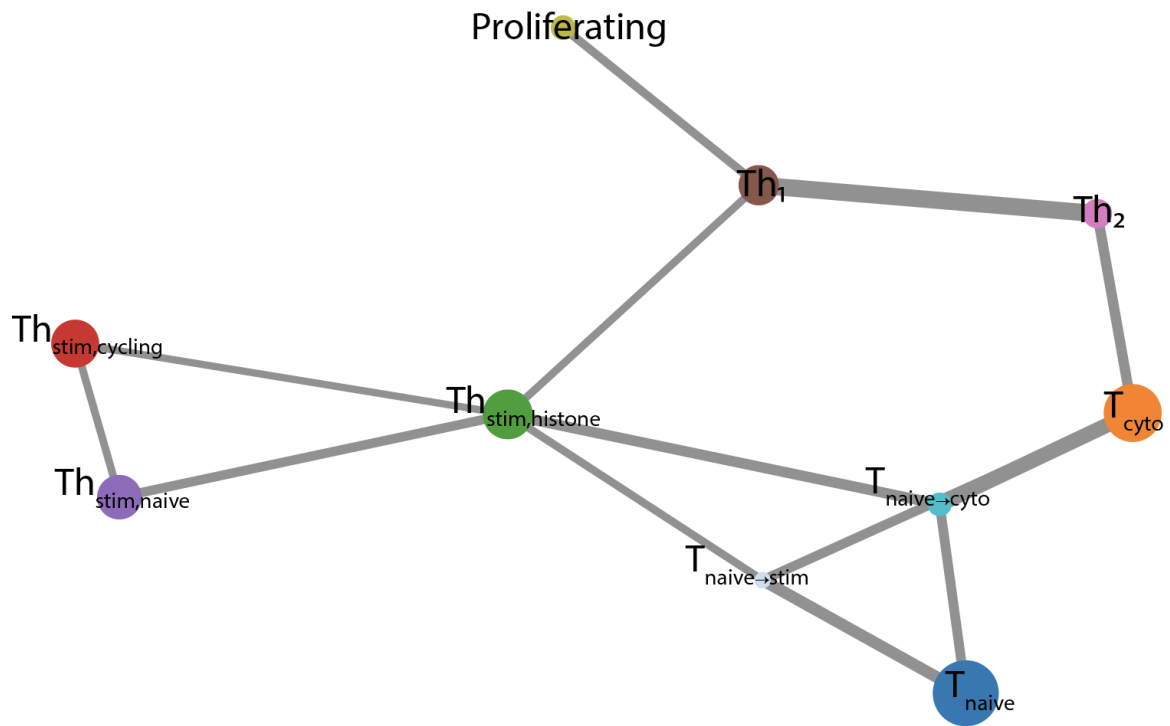


Figure S3.6. Cell state trajectory.

Cell state lineage trajectory using PAGA⁹⁵, where every node is a cell population and the size of the node corresponds to the size of the population. The width of each edge is the strength of connection between the nodes. The color of the point corresponds to the population in Fig. 3.2A.

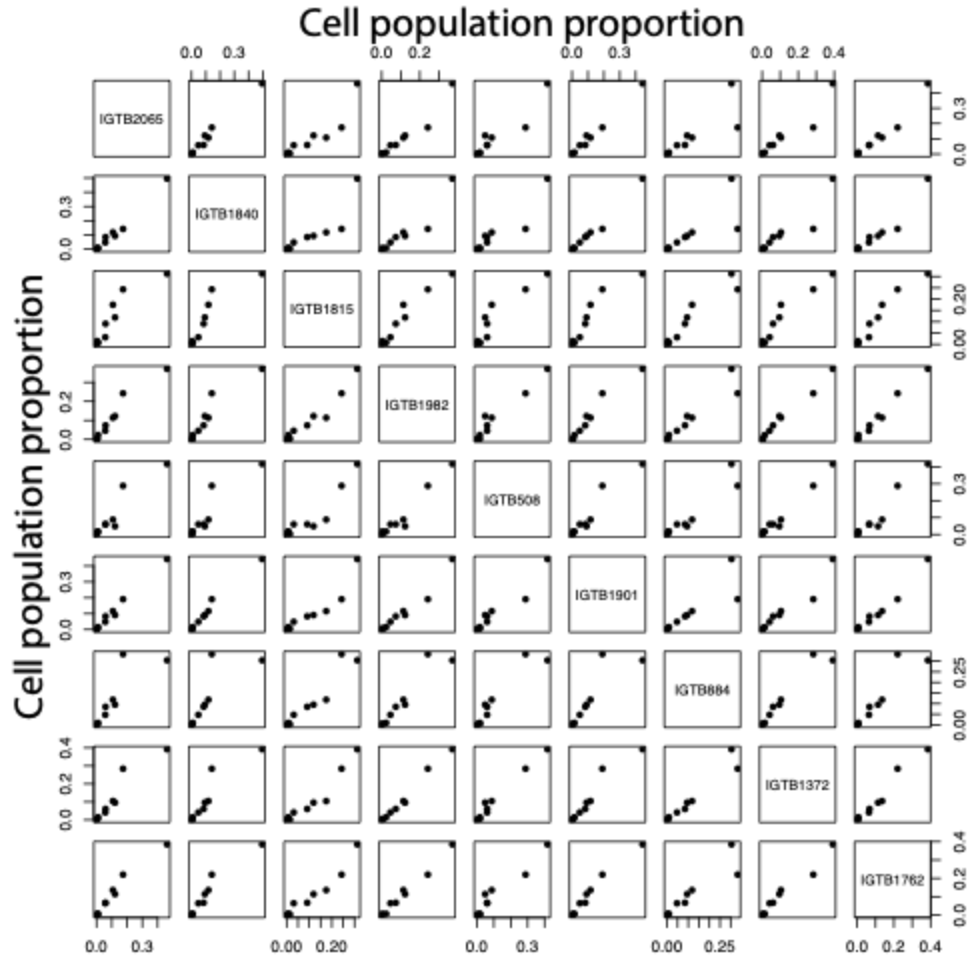


Figure S3.7. Cell proportion across donors.

Correlation matrix (Pearson R) of cell type proportions across donors. Each point represents a cell type.

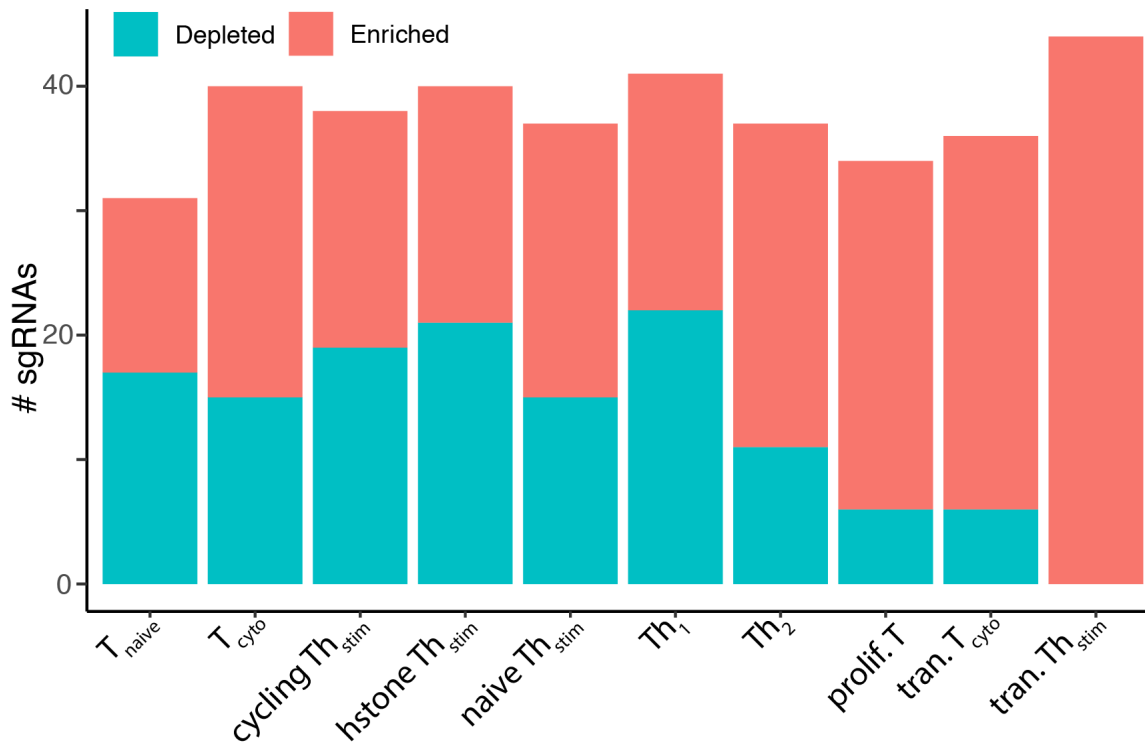


Figure S3.8. sgRNA enrichment and depletion in a cluster.

For each cluster, we calculated if the proportion of cells belonging to sgRNA is enriched (z-score > 1.5) or depleted (z-score < -1.5) in each cluster as compared to the proportion of all cells belonging to that cell state.



Figure S3.9. ARID5A cell state enrichment and depletion

For each cluster, we calculated if the proportion of cells belonging to both *ARID5A*-targeting sgRNAs (ARID5A, cutsite: chr2:96551631 in pink and chr2:96550280 in blue) and calculated a z-score (y-axis) for each cell state (x-axis).

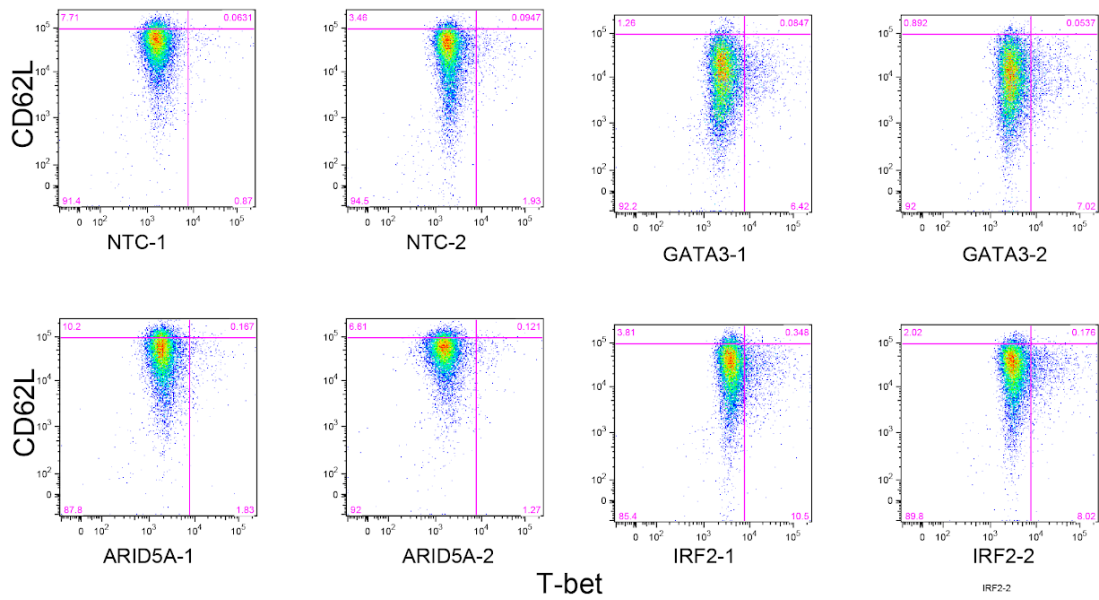


Figure S3.10. Th₁ vs. Th₂ validation

Using FACS, we sorted for CD62L⁺ (Th₂ marker, y-axis) and T-bet⁺ (Th₁ marker, x-axis) cells in our activated CD4⁺ stimulation across 2 non-targeting controls (top left), *GATA3-targeting* (top right), *ARID5A-targeting* (bottom left), *IRF2-targeting* (bottom right) sgRNAs.

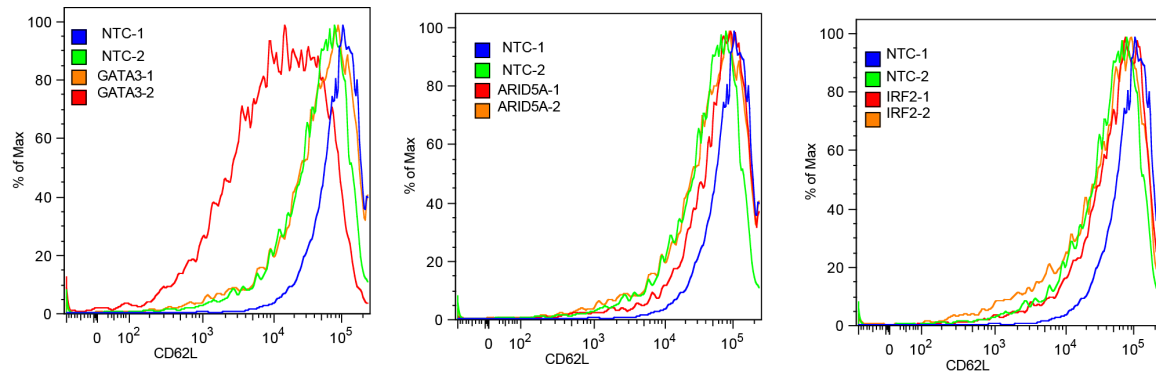


Figure S3.11. Th₂ polarization

Under Th₂ stimulation condition, we sorted for proportion of CD62L⁺ (Th₂ marker, x-axis) across our eight sgRNA conditions. Green and blue are non-targeting controls compared to our red and orange knockout sgRNAs (*GATA3*-targeting sgRNAs left panel, *ARID5A*-targeting sgRNAs in middle panel, *IRF2*-targeting sgRNAs right panel)

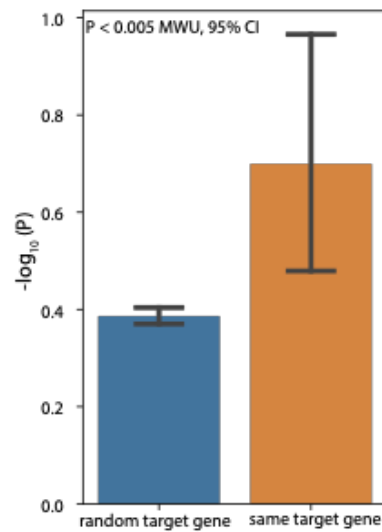


Figure S3.12. Number of TF interaction downstream genes per TF.

Difference between average $-\log_{10}(P)$ for agreement of downstream genes for randomly paired sgRNAs (right) and sgRNAs targeting the same regulator. Agreement of downstream genes for a pair of sgRNAs was tested using a Chi-squared test, and the difference in the $-\log_{10}(P)$ distribution was tested using Mann-Whitney U-test.

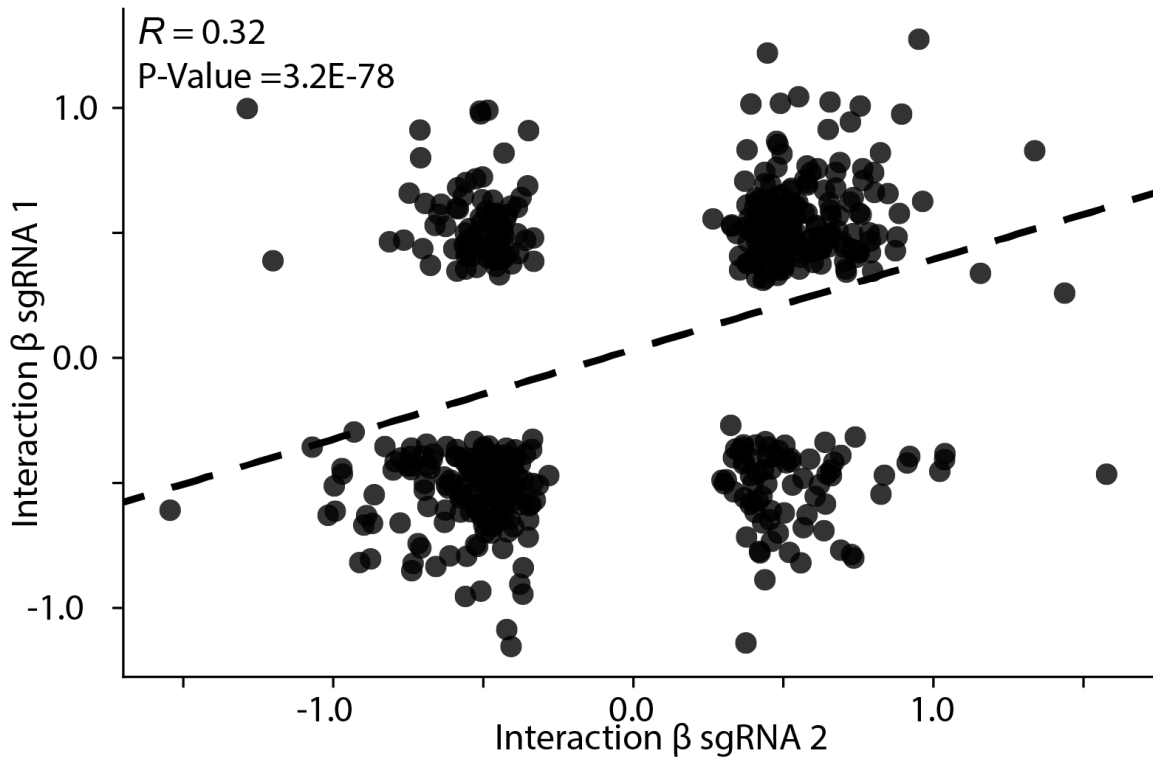


Figure S3.13. Interaction effect size correlation.

The correlation of interaction effect sizes for both sgRNAs targeting the same gene, sampled to 500 points. Each point is a sgRNA pair, with sgRNA 1 on the y-axis and sgRNA 2 on the x-axis. The dashed line is the trend line.

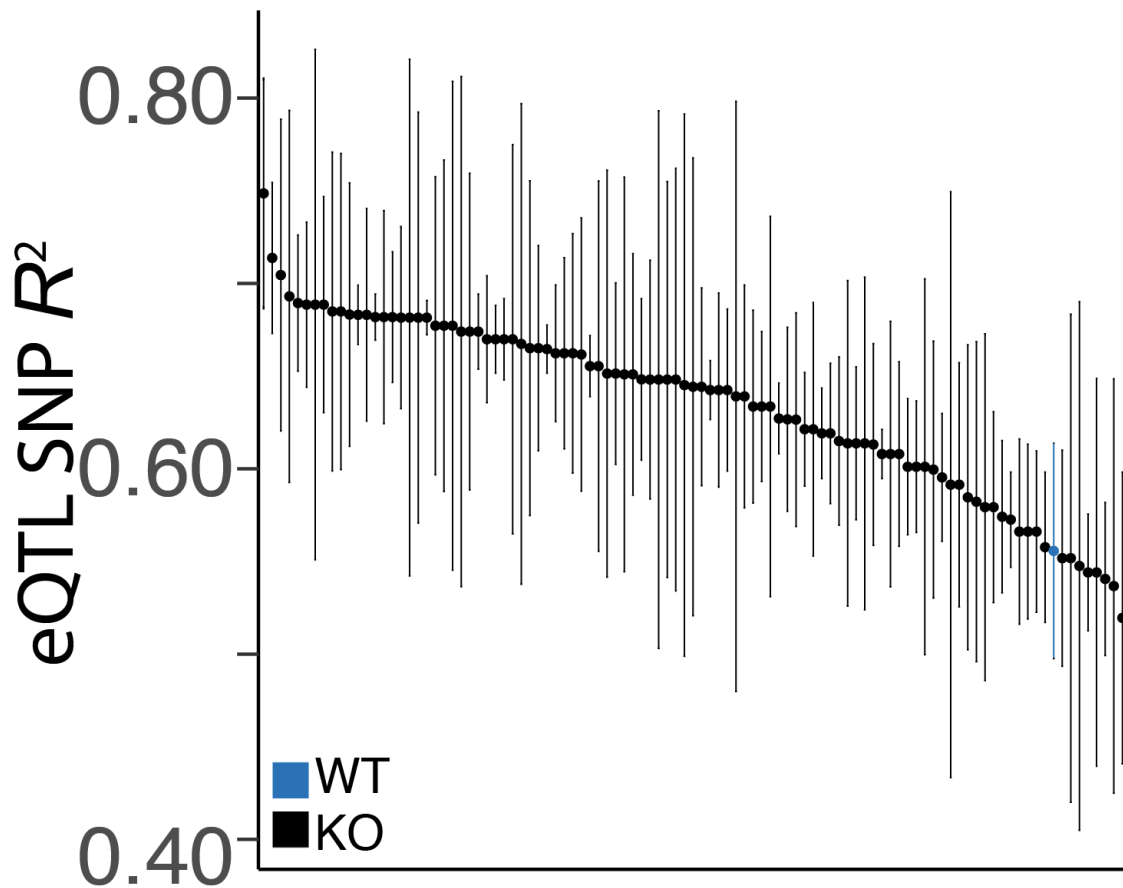


Figure S3.14. Bootstrapped KO R^2 .

Mean and standard deviation of variance explained due to genetics of sampled eQTLs from KO (x-axis, black) and WT (blue) cells after 100 bootstraps.

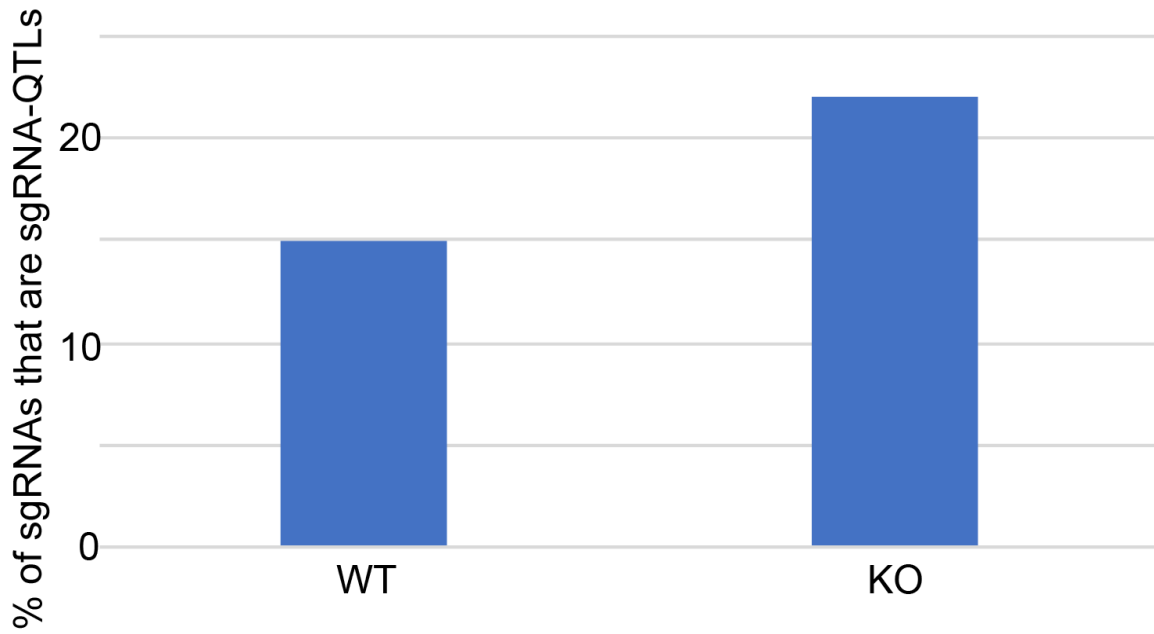


Figure S3.15. Percentage of sgRNAs that have an eQTL.

Percentage (y-axis) of total WT (14) and KO (244) sgRNAs (x-axis) that have an eQTL.

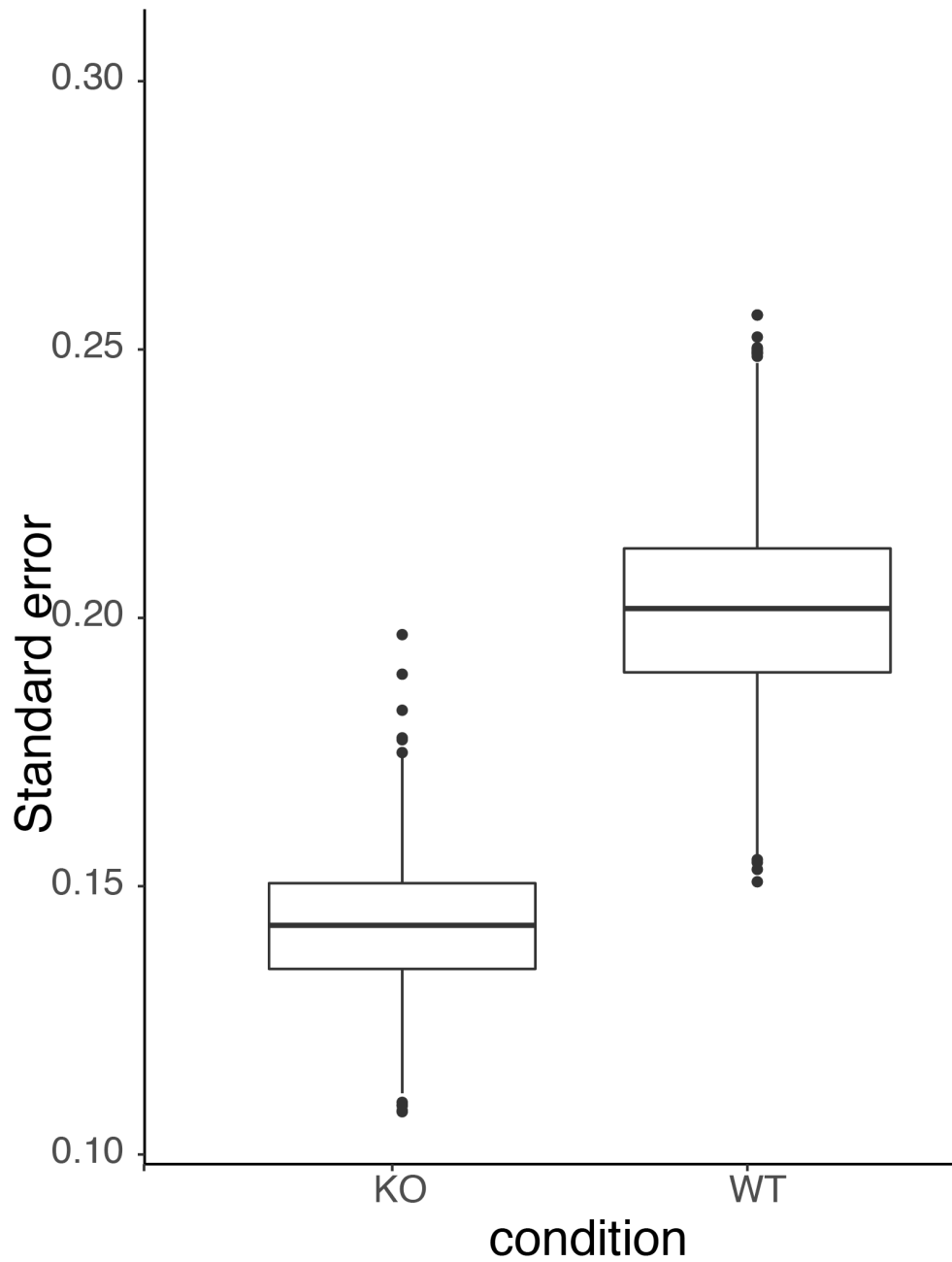


Figure S3.16. Standard error simulation.

The standard errors on 1,000 simulations with an effect size = 0.5, including a regulator variable (WT) and excluding the regulator variable (KO).

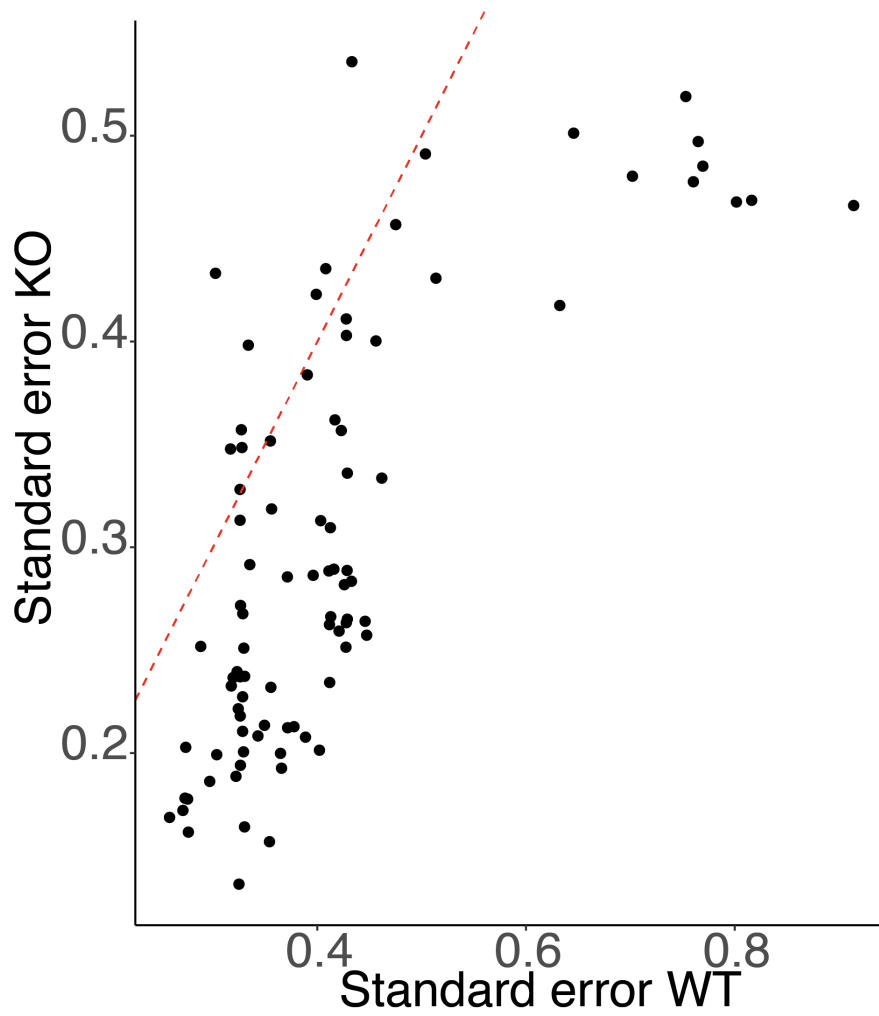


Figure S3.17. Standard error of eQTLs.

Standard error of eQTL (y-axis) compared to the standard error of the association in a WT sgRNA (x-axis). The red dashed line is an abline(slope=1, intercept=0)

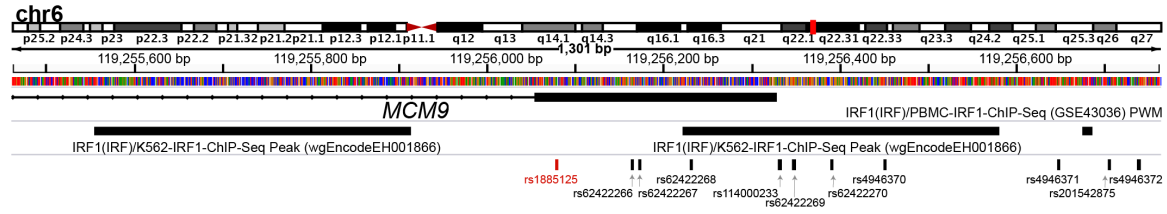


Figure S3.19. Track of *MCM9* locus.

Shown is a 1.3 kb window around the first *MCM9* exon, with three epigenetic annotations surrounding the region. There are two IRF1 ChIP-Seq peaks in K562 (IRF1(IRF1)/K562-ChIP-Seq Peak (wgEncodeEH001866)) and one from IRF1 position weight matrix calculated from a peripheral blood mononuclear cell ChIP-Seq (IRF1(IRF1)/PBMC-IRF1-ChIP-Seq (GSE43036) PWM). Below are SNPs in LD ($D' \geq 0.97$) with our eQTL (in red).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

A handwritten signature in black ink, appearing to be 'P. H. L.', written over a horizontal line.

Author Signature _____

Date August 28, 2019