Title
Short-Sequence Approach to Uncovering Regulatory Mechanisms in the Human Immune System

Author
Afik, Shaked David

Publication Date
2020

Peer reviewed|Thesis/dissertation

Short-Sequence Approach to Uncovering Regulatory Mechanisms in the Human Immune System

By

Shaked Afik

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

In

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nir Yosef, Chair
Professor Lisa F. Barcellos
Professor Lexin Li
Professor Daniel A. Portnoy

Spring 2020

Abstract

Short-Sequence Approach to Uncovering Regulatory Mechanisms in the Human Immune System

by

Shaked Afik

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Nir Yosef, Chair

Short DNA sequences play an important role in the immune response to pathogens. As part of the non-coding regions of the genome, short DNA sequence motifs regulate cell activation and maturation by binding chromatin modifiers and transcription factors. They also determine the ability of each cell in the adaptive immune system to respond to a specific pathogen by forming the antigen-recognizing region of their receptors. This dissertation outlines computational tools I developed for utilizing and integrating high-throughput sequencing data to study the functions of short DNA sequences in the human immune system. I focus on two main aspects of short DNA sequences: (1) As components of the regulatory landscape that control the activation of dendritic cells (DCs) in response to lipopolysaccharide (LPS), and (2) as the determinants of the specificity of T cells and B cells.

The first part of my dissertation investigates the regulatory landscape of DC activation following LPS stimulation. In chapter two I present a model which predicts gene induction based on sequence motif occurrences in the regulatory regions of each gene and show that this regulatory logic is conserved between human and mouse. Chapter three describes a supervised learning pipeline I devised to study the contribution of short sequence motifs to temporal epigenetic changes in human DCs. The second part of my dissertation describes my work on determining the specificity of T and B cells from single-cell RNA-sequencing data. Chapter four presents software I developed to reconstruct the full sequence of T cell receptors from short read single-cell RNA-sequencing. An application of the software links the length of the antigen-recognizing region of the receptor to the state of the cell, demonstrating the importance of such combined analysis in studying the immune response to viral infections. Chapter five describes an extension of the software to reconstruct B cells receptor sequences.

To my parents

Acknowledgements

My greatest gratitude goes to my advisor, Nir Yosef, for his immense support and for not only teaching me how to think about science and how to be a scientist, but for doing so with kindness and patience. I am honored to have been part of the lab from (almost) its inception and to see it grow under his leadership into an incredible team. I also owe a great deal of gratitude to my former mentors and advisors. Nir Friedman and Sebastian Kadener at the Hebrew University introduced me to computational biology and to the exciting world of research. I feel lucky to have been mentored by Manuel Garber from UMass Medical School, my former advisor and current collaborator, who I enjoyed working with and learning from. If it was not for his guidance, I would never have thought about pursuing a PhD. I would also like to thank the members of my qualifying exam and dissertation committee for their valuable feedback: Lisa Barcellos, Lexin Li, Nick Ingolia and Dan Portnoy.

I wish to thank the members of the Yosef lab for endless discussions and assistance. Special thanks to Jim Kaminski and Romain Lopaz (best office mates one could ask for), David DeTomaso, Anat Kreimer, Michael Cole, Alyssa Morrow and Chenling Xu. I would also like to thank my past and current co-first-authors and collaborators: Kathleen Yates, Kevin Bi and Nick Haining, Pranitha Vangala, Elisa Donnard and Manuel Garber, and Gabe Raulet. Their hard work and enthusiasm was an inspiration.

I have the privilege of being a member of the second cohort of the UC Berkeley computational biology PhD program. Thank you to my fellow cohort member David DeTomaso and our "adoptive" cohort: Jim Kaminski, Brooke Rhead, Jeff Spence, Amy Ko and Rob Tunney. Their friendships have made my time here not only less stressful but also so enjoyable. Many Thanks to Kate Chase and Xuan Quach for their support to the program in general and to me personally.

I would also like to thank all the great friends I gained here from the comp bio program, UC Berkeley and the Bay Area. You have made Berkeley feel like home and I will forever be grateful.

Special thanks to my family, especially to my mom and dad. I could not have endured this time away without their never-ending love and support.

This work is dedicated to my partner, Noam. For his spirit, thoughtfulness and nonstop encouragement that made me push myself further every day for the last 14 years, even when we were 10 time zones away. Thanks to him, my life is completely different than I could have ever imagined and I am so happy for it.

As I am writing this, I am only a few weeks away from the birth of my first child. There are no words to describe how remarkably better the last two years have been just knowing I will get to meet you at the end of it. I cannot wait to discover what I will learn from you.

Table of Contents

Chapter 1 - Introduction

The role of the immune system is to defend against harmful pathogens such as viruses and bacteria. It can be broadly divided into two lines of defense. The first line is the innate immune system, which detects the pathogen and reacts with a rapid yet general response. Then, the adaptive immune system, namely T and B cells, provides a slower response which is tailored to the specific pathogen. Those processes include many cell types that undergo vast molecular changes as cells differentiate and mature in response to the pathogen. Profiling the molecular basis of the human immune response is of great importance, as it uncovers the mechanisms underlying the body's response to vaccination, infections and other diseases, which in turn leads to developments of drugs and methods for cancer therapy (Jiang, 2017; Villani et al., 2018). Moreover, it allows us to gain a better understanding of basic principles in molecular biology (Pope and Medzhitov, 2018). In my dissertation, I was interested in the role that short DNA sequences play in various stages of the immune response to pathogens. I focused on two main functions of these sequences: As regulators of transcriptional changes in cells from the innate immune system, and as the major genomic component which determines the specificity of cells from the adaptive immune system.

Transcriptional regulation

Changes to cell state, e.g. in response to environmental stimuli, are controlled by changes in expression of thousands of genes. Those changes are mediated by a complex regulatory network consisting of non-coding DNA sequences, chromatin structure and a class of proteins called transcription factors (TFs) (Lelli et al., 2012). Each TF can bind a set of similar short DNA sequences, collectively represented as a DNA binding "motif". TFs promote or inhibit gene expression by binding either at promoter regions (directly upstream of the transcription start site), or at more distal genomic regions termed enhancers. However, even when the sequence motif of a TF is known, predicting the exact sequences in the genome that will be bound by that TF is a challenging task since TFs can bind low affinity sequences and some TFs bind several distinct motifs (Siggers and Gordân, 2014). In addition, for a given motif only a small portion of the motif instances in the genome are bound in each cell type. This selective binding can occur due to competition between TFs with overlapping motifs, a requirement for a TF to interact with another bound TF in order to bind a specific instance, or since many genomic regions are inaccessible because of their local chromatin conformation (Spitz and Furlong, 2012). DNA can be wrapped around proteins termed histones, making it less accessible to recognition by TFs and more inactive compared to "open" DNA which is not bound by histones. Chemical modifications to the histone proteins play a part in the opening of the chromatin and recruitment of regulatory factors. For example, methylation of lysine 4 of histone H3 is associated with transcriptional regulation - Monomethylation (H3K4me1) is linked to enhancer regions and trimethylation (H3K4me3) is associated with promoters. Enhancers and promoters are susceptible to both transcriptional repression and activation, where a histone marked with acetylation of lysine 27 (H3K27ac) is associated with an active regulatory region.

A valuable technology for studying transcriptional regulation in the immune system is high-throughput sequencing (Yosef and Regev, 2016). In addition to characterizing the transcriptome via RNA-sequencing, we can uncover histone modifications with chromatin immunoprecipitation coupled with high-throughput sequencing (ChIP-seq). Moreover, ChIP-seq can be applied to find specific TF-DNA interactions, however each such experiment is limited to one TF. Assays for genome-wide chromatin accessibility such as DNase-seq and Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq; Buenrostro et al., 2013) enable detection of potential regulatory regions by finding open regions in the genome. Another important feature of chromatin accessibility assays is that a DNA motif bound by a TF creates a footprint-like region with low number of aligned reads, surrounded by regions with a high number of alignments. This feature is used for computational estimation of TF binding within accessible regions for many TFs from a single experiment (Gusmao et al., 2016). Overall, with this range of assays we can characterize the state of immune cells under different conditions. However, computational methods are required in order to link the epigenetic landscape with the transcriptome and gain a comprehensive understanding of the regulatory mechanisms that are taking place in the cell.

Uncovering the regulatory landscape of Dendritic cells following LPS stimulation

Dendritic cells (DCs) are antigen-presenting cells that are part of the innate immune system and are essential for the initiation of the adaptive immune response in both mouse and human (Haniffa et al., 2013). DCs exist in an immature state and a mature state (Reis e Sousa, 2006). The immature state is a steady state condition in which DCs sample the environment and take up antigen. Immature DCs possess receptors that detect molecular features unique to microbes. Bacterial lipopolysaccharide (LPS), for example, activates innate immune signaling pathways in immature DCs via the TLR4 receptor complex. Activation of these receptors initiates a developmental switch that results in DC maturation. In contrast to immature DCs, mature DCs elicit potent T cell responses that target specific pathogens by turning antigen-specific naive T cells into effector T cells. The nature of the T cell response that ensues is determined by the maturation status of the antigen-presenting DC and the array of signals that the DC provides to T cells. The maturation status of DCs is determined by environmental cues transmitted by the receptors to the DC nucleus by signal transduction cascades. Given the role of DCs as the orchestrators of the adaptive immune response, the transition from immature DC to mature DC is an important developmental switch that occurs within the human immune system.

Previous work had studied the activation of mouse DCs following stimulation with LPS (Amit et al., 2009; Garber et al., 2012). Those studies revealed a large and coordinated transcriptional response that occurs within the span of a few hours and is highly synchronized, involving thousands of genes expressed in various temporal patterns. The rapid response and large expression changes makes the DC response to LPS an ideal experimental system for modeling the genomics of gene regulation.

Chapter 2 of my dissertation describes our work comparing the transcriptional response of DCs to LPS between human and mouse and investigating which short DNA sequences are important

regulatory features for gene induction. In chapter 3 we focus on the regulatory response to LPS in human Dendritic cells. I present a supervised-learning pipeline designed to detect short DNA motifs which are functional at various times up to 24 hours after LPS stimulation and provide a comprehensive map of regulatory interactions during DC activation.

Heterogeneity and specificity of the adaptive immune system

The adaptive immune system produces a strong, targeted response against pathogens. To provide such effective immunity, the T cell and B cell compartments must contain diversity in both their cell state and their ability to recognize a wide range of antigens. T cells and B cells express a surface receptor called T cell receptor (TCR) or B cell receptor (BCR), respectively. The sequence of the receptor determines to which antigen the cell can bind and consequently react to. The main antigen-recognizing part of the receptor is a short region named the complementarity determining region 3 (CDR3). During T and B cell development a unique CDR3 sequence is generated for each cell by random sequence mutations, insertions and deletions. This tightly regulated yet random process results in a different genomic sequence of the receptor for each cell, providing the needed diversity (Nikolich-Zugich et al., 2004).

Utilizing high-throughput sequencing to perform a combined TCR/BCR-transcriptome analysis can allow us to gain more insights to the molecular basis of the response to viral infections, autoimmune diseases and help in vaccine design (Venturi and Thomas, 2018). Because of the highly variable nature of the CDR3 sequence, standard population-level protocols for RNA-sequencing and subsequent analysis methods are unable to map the CDR3-originating sequences to the genome or separate the CDR3 sequences of different cells. While other protocols exist to sequence the receptors from populations of cells, they cannot be combined with transcriptome information from the same population. In recent years, many high-throughput sequencing methods have been adapted to extract information at the single-cell level (Papalexi and Satija, 2018; Stuart and Satija, 2019). Single-cell RNA-sequencing allows for simultaneous measurement of TCR/BCR sequences and global transcriptional profiles from single cells, enabling us to study how differences in TCR and BCR contribute to heterogeneity in cell state. However, similarly to population-level studies, standard transcriptome analysis methods cannot map CDR3-originating sequences, thus computational tools to perform receptor sequence reconstruction in individual cells are needed.

To this end, the second part of my dissertation describes the computational software I developed for receptor sequence reconstruction in single cells. Chapter 4 presents TRAPeS, a method for TCR reconstruction. By combining TCR sequence with transcriptomic data, we discover a link between the length of the CDR3 and the transcriptional cell state of human Yellow Fever Virus-specific T cells. In Chapter 5 I extend TRAPeS to create BRAPeS, a reconstruction method suited for B cell receptor reconstruction.

References

Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science *326*, 257–263.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Gusmao, E.G., Allhoff, M., Zenke, M., and Costa, I.G. (2016). Analysis of computational footprinting methods for DNase sequencing experiments. Nat. Methods *13*, 303–309.

Haniffa, M., Collin, M., and Ginhoux, F. (2013). Ontogeny and functional specialization of dendritic cells in human and mouse. Adv. Immunol. *120*, 1–49.

Jiang, N. (2017). Immune engineering: from systems immunology to engineering immunity. Curr Opin Biomed Eng *1*, 54–62.

Lelli, K.M., Slattery, M., and Mann, R.S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. Annu. Rev. Genet. *46*, 43–68.

Nikolich-Zugich, J., Slifka, M.K., and Messaoudi, I. (2004). The many important facets of T-cell repertoire diversity. Nat. Rev. Immunol. *4*, 123–132.

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. *18*, 35–45.

Pope, S.D., and Medzhitov, R. (2018). Emerging Principles of Gene Expression Programs and Their Regulation. Mol. Cell *71*, 389–397.

Reis e Sousa, C. (2006). Dendritic cells in a mature age. Nat. Rev. Immunol. *6*, 476–483.

Siggers, T., and Gordân, R. (2014). Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res. *42*, 2099–2111.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. Nat. Rev. Genet. *20*, 257–272.

Venturi, V., and Thomas, P.G. (2018). The expanding role of systems immunology in decoding

the T cell receptor repertoire. Curr Opin Syst Biol *12*, 37–45.

Villani, A.-C., Sarkizova, S., and Hacohen, N. (2018). Systems Immunology: Learning the Rules of the Immune System. Annu. Rev. Immunol. *36*, 813–842.

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

Chapter 2 - Comparative analysis of immune cells reveals a conserved regulatory lexicon

This chapter describes a comparative study of the transcriptional response of Dendritic cells to LPS in human and mouse. My co-authors and I show that while most enhancers are not conserved, genes with a conserved temporal activity are enriched in conserved enhancers. In addition, I built a random forest classifier to predict gene induction using a set of conserved short sequence motifs as features. My model successfully predicts gene induction in both human and mouse, demonstrating that the regulatory logic of DC activation is conserved.

This work was published in *Cell Systems* in 2018 (Donnard et al. 2018), and I am reporting it as it was published. The authors on the paper are:

Elisa Donnard[1,7], Pranitha Vangala[1,7], Shaked Afik[2,7], Sean McCauley[3], Anetta Nowosielska[3], Alper Kucukural[3,4], Barbara Tabak[1], Xiaopeng Zhu[1], William Diehl[3], Patrick McDonel[1,3], Nir Yosef[2,5], Jeremy Luban[3*], Manuel Garber[1,3,4,6*]

1. Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA-01605, USA
2. Center for Computational Biology, University of California, Berkeley, Berkeley, CA-94720, USA
3. Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA-01605, USA
4. Bioinformatics Core, University of Massachusetts Medical School, Worcester, MA-01605, USA
5. Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA-94720, USA
6. Lead Contact
7. These authors contributed equally
*Correspondence: Manuel.Garber@umassmed.edu (M.G.), Jeremy.Luban@umassmed.edu (J.L.)

Summary

Most well-characterized enhancers are deeply conserved. In contrast, genome-wide comparative studies of steady-state systems showed that only a small fraction of active enhancers are conserved. To better understand conservation of enhancer activity, we used a comparative genomics approach that integrates temporal expression and epigenetic profiles in an innate immune system. We found that gene expression programs diverge among mildly induced genes, while being highly conserved for strongly induced genes. The fraction of conserved enhancers varies greatly across gene expression programs, with induced genes and early-response genes, in particular, being regulated by a higher fraction of conserved enhancers. Clustering of conserved accessible DNA sequences within enhancers resulted in over 60 sequence motifs including motifs for known factors, as well as many with unknown function. We further show that the number of instances of these motifs is a strong predictor of the responsiveness of a gene to pathogen detection.

Introduction

Enhancers act over long chromosomal distances to control gene expression in a cell type-specific fashion (Ong and Corces, 2011). Recent advances in genomic methods have revealed hundreds of thousands of enhancers defined by biochemical signatures that include p300 binding, H3K27ac and H3K4me1 modifications (Heintzman et al., 2007; Rada-Iglesias et al., 2011; Visel et al., 2009). These studies have shown that the vast majority of regulatory elements are species-specific. Furthermore, gain or loss of species-specific enhancers across phylogeny is not concomitant with gain or loss of genomic sequence. Instead, the majority of species-specific enhancers are composed of ancestral sequences that gain enhancer activity in a species-specific manner (Ballester et al., 2014; Kunarso et al., 2010; Mikkelsen et al., 2010; Odom et al., 2007; Schmidt et al., 2010; Villar et al., 2015).

Rapid turnover of species-specific enhancers stands in stark contrast to the highly conserved nature of well-known enhancers that play essential roles in development (Chew et al., 2005; Crocker and Erives, 2008; Lettice et al., 2003), metabolism (Claussnitzer et al., 2015) and viral defense (Panne et al., 2007). Comparative sequence analysis revealed millions of conserved non-coding elements in the human genome that are likely to act as functional enhancers *in-vivo* (Pennacchio et al., 2006). Given the general expectation that most functional elements are under purifying selection, there is currently a disconnect between enhancers that are defined by biochemical activity and those defined by evolutionary conservation.

Several arguments have been proposed to reconcile this apparent contradiction between the high turnover rate of biochemical signatures of enhancers observed in comparative studies and the high conservation of a handful of well-characterized examples. One proposed explanation is that typical enhancer elements are redundant, with shadow enhancers that can compensate for the loss of another enhancer (Dunipace et al., 2011; He et al., 2011; Perry et al., 2010). However, redundant enhancers show no relaxation of sequence constraint compared to non-redundant enhancers (Cannavò et al., 2016). Another proposal is that genetic drift may sometimes yield new transcription factor binding sites, eventually leading to novel regulatory elements that make old ones redundant (Ludwig et al., 2000). Accordingly, individual binding sites within enhancers may be shuffled over time and even be replaced by sites occurring on different enhancers. Although both arguments would explain the reduced selective pressure on typical enhancers, they do not explain the apparent strong purifying selection of functionally important enhancers.

An alternative explanation is that most of the biochemically defined enhancers might not be critical in controlling conserved gene regulatory programs. Instead, conserved gene regulatory programs are controlled by a small subset of conserved enhancers. Here we revisited the question of enhancer conservation by studying the transcriptional regulation of genes that respond to Lipopolysaccharide (LPS). LPS is a cell wall component of gram negative bacteria, that is detected by the TLR4-MD-2 complex (Park et al., 2009). This is a well-defined inducible response in both human and mouse dendritic cells (Amit et al., 2009; Garber et al., 2012; Parnas et al., 2015), which involves hundreds of genes and, in its early stages, offers a virtually synchronous response that is mostly transcriptionally controlled (Rabani et al., 2011). Focusing on LPS-responsive genes reduces many confounding factors such as the role of

post-transcriptional regulation that make steady state analysis more complex. We focused on the evolutionary profile of enhancers that are associated with both species-specific and shared LPS-responsive genes. Our results reconcile the biochemical and conservation-based definitions of enhancers and demonstrate the importance of evolutionary selection of enhancers in controlling conserved transcriptional programs.

Results

Transcriptional dynamics of human and mouse DCs in response to LPS

We generated dendritic cells (DCs) from the bone marrow of two C57BL/6 mice and from human peripheral blood mononuclear cells (PBMCs) from two donors. We stimulated each set of DCs with LPS and collected cells at 0, 1, 2, 4, and 6 hours post-stimulation. We measured genome-wide gene expression by RNA sequencing (RNA-Seq), chromatin accessibility by ATAC-Seq (Buenrostro et al., 2013) and enhancer activity by chromatin immunoprecipitation of H3K27ac followed by sequencing (ChIP-Seq).

To compare human and mouse response to LPS we focused on genes that could be mapped unambiguously between human and mouse (one-to-one homologs). Immature mouse and human DCs have similar transcriptional profiles with 72% (6,370) of all one-to-one homologous genes detected in at least one species being expressed in both. Among the 3,642 genes that are LPS-responsive in at least one species only 740 have similar expression kinetics (Figure 1A, STAR Methods). However, induced genes with similar patterns showed greater induction levels (3.7-fold higher on average, Figure S1A), and were enriched in effectors (cytokines and chemokines $p < 10^{-5}$, hypergeometric test) and transcription factors (TFs, $p < 0.0001$, hypergeometric test) compared to genes induced in only one species. Overall, the bulk of the differences between mouse and human DCs involve small fold changes and genes that are not critical to the LPS response. There are, however, interesting exceptions of highly induced genes that are species-specific. A well-known example, Nitric Oxide Synthase 2 (NOS2), has an important role in the mouse immune response to microbes but is not induced by LPS in human innate immune cells (Bogdan, 2001; Mestas and Hughes, 2004). Conversely, we find that the T-Cell effector Indoleamine 2,3‑dioxygenase (IDO1) gene is highly induced in the human DCs (Mellor and Munn, 2004), but is not induced in mouse DCs.

We next clustered the genes that were responsive in both human and mouse DCs (Figure 1B, STAR Methods). We observed three broad shared expression trends: genes that were downregulated in both species (clusters D1 and D2), genes that were induced within 1h after LPS stimulation (early-induced genes, clusters I1 and I2), and genes that were induced at least 2h after LPS stimulation (clusters I3, I4 and I5). These different clusters showed broad similar expression trends while also reflecting subtle differences in species-specific timing of peak expression. Shared early-induced genes were enriched for cytokines and TFs (adjusted $p < 10^{-5}$, hypergeometric test). Cluster I1 specifically, was 5.4-fold enriched in TFs ($p < 10^{-7}$, hypergeometric test), including immediate-early genes such as JUN and FOSB. Shared late-induced genes included the TFs STAT1 and IRF9 (Figure 1C), which are involved in autocrine signals from IFNβ and TNFα resulting from LPS detection (Toshchakov et al., 2002).

Although most species-specific genes were induced at relatively low levels, these differences may result from either changes in *cis*-regulatory elements or from differences in TF expression. We first focused on differences in TF expression. Overall, 530 TFs were expressed in at least one species, of which most (70%) were expressed in both species (Figure S1B), and most TFs detected only in one species had significant lower expression (Figure S1C, $p < 10^{-15}$ Wilcoxon rank-sum test). Further, most TFs that respond to LPS have well conserved kinetics (STAR Methods, Figure S1D) and although we find specific TFs having diverging expression patterns, in most cases other members of the same family (defined by TF Class, Wingender et al., 2013) show similar kinetics. For only 15 TFs we found no evidence of compensatory changes, most of these cases involved TFs with a low peak expression or induction (Figure S1E). These results suggest that TF expression is conserved between mouse and human DCs. Two interesting exceptions are the AP1 factors ATF5 and ATF4, which are highly expressed and induced only in human DCs (Figure 1D). These two TFs respond to a variety of other stress stimuli, such as amino acid starvation, heat shock and oxidative stress (Harding et al., 2003; Wang et al., 2007a; Watatani et al., 2007), suggesting a human-specific role for cellular stress response in DC response to LPS. We next turned to *cis*-regulatory elements to further determine the source of changes in expression profiles.

The epigenetic landscape of regulatory elements in human and mouse DC response to LPS
To define the regulatory landscape of mouse and human DCs we followed a two-step process. First, we mapped candidate enhancer regions using ChIP of histone marks that are typical of transcriptionally active regions (Heintzman et al., 2007; Rada-Iglesias et al., 2011; Shlyueva et al., 2014). We then used ATAC-Seq signal to identify accessible regions within our H3K27ac-defined regions (Buenrostro et al., 2013) (STAR Methods, Figure 2A).

As in previous studies (Cheng et al., 2014; Vierstra et al., 2014; Villar et al., 2015), we defined Enhancers with Conserved Activity (ECAs) as enhancers whose sequence could be uniquely mapped across species and which also had H3K27ac signal in both species. We defined Enhancers with SPecies-specific Activity (ESPAs) to include both species-specific sequences with H3K27ac signal and homologous sequences with species-specific H3K27ac signal. Consistent with previous studies (Villar et al., 2015), for the majority of the enhancers and promoters found in one species it was possible to unambiguously identify homologous sequences in the other species (Figure 2A,B, S2A and STAR Methods). However, as observed in other systems (Mikkelsen et al., 2010; Schmidt et al., 2010), conservation of H3K27ac signal paints a different picture: While 77% of mouse DC promoters mapped to human sequence with H3K27ac signal, for mouse DC enhancers this fraction is only 25% (Figure 2B, S2A). Among transposase-accessible regions within mouse enhancers, only 19% of homologous regions are transposase-accessible in human (Figure S2B, S2C). However, among enhancer sequences with conserved H3K27ac signal, 59% also had conserved accessibility in both species. This shows that accessible regions within enhancers and hence TF binding is maintained across evolutionary time whenever the activity of the larger region is also conserved. Overall, the fraction of ECAs (25%) observed in DC enhancers was similar to the one observed between mouse and human liver enhancers (Villar et al., 2015). Thus, in spite of the strong positive selection acting on innate immune cells, the regulatory landscape has not diverged much further than in liver, likely owing to the critical nature of this response for the organism's survival. Since TF expression is

well conserved while *cis*-regulatory elements have drastically diverged, it appears that most differences in LPS-responsive expression between human and mouse are the result of *cis*-regulatory changes rather than differences in *trans*-regulators.

We observed a stronger H3K27ac and ATAC signal in enhancers and promoters that are active in both species, compared to species-specific regions (Figure 2C, S2D). This observation could result from a threshold bias to define conserved active loci, with one species having a lower signal that fails to meet the enrichment threshold. However, the H3K27ac signal on the homologous regions of ESPAs was indistinguishable from background (black lines, Figure 2C, S2D). Thus, our classification of an active regulatory region as species-specific is not influenced by differing signal intensity.

Enhancers that are active in progenitor cells are more conserved but are not involved in the response to LPS

Mouse DCs are derived from bone marrow (mBM), whereas human DCs are derived from monocytes. We therefore hypothesized that observed differences in enhancer activity in these cells could be the result of prior activity in progenitor cells. To identify such enhancers we relied on H3K27ac ChIP-Seq data from mBMs (Yue et al., 2014) and generated similar data for human monocytes. Although the fraction of pre-established active enhancers is different in mouse (23% in bone marrow) and human (55% in monocytes), enhancers that are pre-established are more conserved than those that are DC-specific (Figure 2D, S2E). Consequently, pre-established active enhancers are not likely to explain the differences we observed in the transcriptional response to LPS in human and mouse DCs.

The higher degree of conservation among enhancers that are active in progenitors may indicate that they belong to a family of ubiquitous enhancers that have been shown to be more conserved in evolution (Cheng et al., 2014). Consistent with this, nearly half (40%) of the enhancers that are pre-established in mouse bone marrow are also active in liver. Further, we found that pre-established enhancers constitute 39% of all enhancers for genes with rapid downregulation in both species (Cluster D2, Figure 1B), compared to 23% for all genes. This indicates that ubiquitous enhancers, albeit being more highly conserved than cell type specific enhancers, are not involved in response to stimulus, and are not likely to play an important role in the regulation of LPS response.

Regulation of early LPS-induced genes is both complex and conserved

Previous comparative analyses have shown that conserved enhancers are associated with genes involved in specific biological processes (Ballester et al., 2014; Kunarso et al., 2010; Mikkelsen et al., 2010; Schmidt et al., 2010). While there is a slight increase in the fraction of ECAs among shared induced genes compared to enhancers of non-induced or species-specific induced genes, the largest increase (40%, almost double than for non-induced genes) is found on enhancers associated with shared early-induced genes ($p < 10^{-12}$, Fisher exact test) (Figure 3A, S3A). This shows that selection does not act uniformly across all enhancers but rather, that it depends on the particular transcriptional program in which the enhancers function.

Visual inspection of highly induced genes after LPS stimulation such as NFKBIZ, IL6 and PRDM1 (Figure 2A), suggested that these genes were associated with a high number of enhancers and with super enhancers (Whyte et al., 2013). Such regulatory complexity was previously observed in genes that have a cell type specific regulation during lineage commitment (González et al., 2015). Interestingly, genes with high regulatory complexity (having four or more enhancers) were highly enriched in LPS-responsive genes and particularly, in early-induced genes (Figure 3B, S3B). Consistent with our initial observation, genes in the top regulatory complexity tier reached higher maximal expression after induction (Figure S3C). Enhancers that regulate highly induced early genes were also more likely to be conserved. Indeed, on average 2/5 of the enhancers are conserved for shared early response genes with complex regulatory loci, compared to only 1/5 for species-specific early response genes that also have complex regulatory loci (Figure 3C, S3D). In general, genes with shared temporal patterns constitute the core of LPS response, and accordingly, their regulation is under strong purifying selection.

Conserved lexicon within accessible regions

Chromatin accessibility is widely considered critical for transcription factor binding (John et al., 2011; Wang et al., 2012), and we confirmed the strong preference of TF binding on accessible regions using our previous transcription factor occupancy maps (Garber et al., 2012) (Figure S4A). As such, DNA accessible regions hold key information related to regulatory activity. Therefore, we next sought to establish the degree to which DNA accessible regions within ECAs are under purifying selection. To this end, we estimated the substitution rate of DNA accessible regions at 10-base resolution (Garber et al., 2009), using a multiple sequence alignment that included 41 mammalian genomes and 2 vertebrate genomes (STAR Methods). Comparison of the substitution rate between DNA accessible regions within ECAs and ESPAs showed a marked reduction in substitution rate (p-value $< 10^{-15}$, KS-Test, Figure 4A, S4B). Therefore, ECAs are not only preserved in their activity but there are clear marks of purifying selection in the chromatin accessible sequence within, which is most amenable to TF binding.

To identify sequence elements at the core of ECA function, we clustered conserved 10-mers within ECAs (STAR Methods). Clustering resulted in 66 distinct conserved sequence motifs which we represent by conserved position weight matrices (cPWMs). 31 cPWMs have a clear match to a known transcription factor motif and include all major regulators of TLR4 signaling (STAT, AP1, NFKB, ETV, Figure 4B, Table S2). In addition, we identified 35 cPWMs with no clear similarity to any reported motif in public databases (STAR Methods).

Analysis of both known and unidentified cPWMs showed enrichment for genes with specific temporal expression patterns and, in particular, genes with shared response (Figure 4C, S4C). Importantly, the enrichment of motifs on induced genes was consistent with the expression kinetics of TFs that have affinity for these motifs and recapitulated previous reports (Garber et al., 2012; Medzhitov and Horng, 2009).

To measure the contribution of this conserved lexicon to gene regulation we next trained a random forest classifier to predict if a gene would be strongly induced (> 4-fold) or maintain constant expression following LPS stimulation (STAR Methods). The classifier performed well,

achieving a mean area under the curve (AUC) value of 0.75 of the receiver operating characteristic curve (ROC) and a mean AUC value of 0.74 for the precision recall (PR) curve in 10-fold cross-validation (Figure 4D). This confirms the ability of cPWMs to predict gene induction, but also suggests that cPWM instances alone are not sufficient predictors.

Importantly, when we applied the model we trained in mouse to predict expression induction in human, it performed with similar accuracy and precision, achieving an AUROC of 0.68 and an AUC value of 0.63 for the PR curve (Figure 4D). Motifs of the key regulators such as NFKB, AP1, STAT and EGR along with several novel GC rich motifs are amongst the top classifying features (Figure 4E).

Enhanceosomes in conserved innate immune responses

Enhancers are thought to function in two broadly different mechanisms (Arnosti and Kulkarni, 2005). In enhanceosomes, TFs act cooperatively and their binding results in an on/off signal, where loss of even one TF binding site profoundly disrupts the function of the enhanceosome. Billboards on the other hand, are modular enhancers where the binding of each TF is not necessary for enhancer activity but rather has an additive or synergistic effect. The prototypical enhanceosome is the IFNβ proximal enhancer, which requires the assembly of 6 TFs to induce IFNβ expression (Thanos and Maniatis, 1995). Mutations that disrupt a single binding site disrupt the enhancer and are highly deleterious. Consistent with this, the IFNβ enhanceosome sequence is more highly constrained than the protein coding sequence of IFNβ, the gene it regulates (Figure S5). Since the effect of mutations on enhanceosomes can be highly penetrant, we sought to identify and catalog enhancers that have characteristics typical of enhanceosomes and that may help prioritize non-coding mutations associated with immune disease.

We scanned for candidate enhanceosome regions in chromatin accessible regions within ECAs that were 1) Bound by at least six TFs, based on our previous binding maps of 14 TFs and 2) Had a large portion (> 30%) of their sequence conserved. Our scan identified 80 chromatin accessible regions (Figure 5 for example & Table S3) that resemble enhanceosomes, such as the IFNβ proximal enhancer (Figure S5). Consistent with their innate immune specific function, genes associated with these conserved, highly bound regions tend to have similar temporal induction in both human and mouse ($p < 0.01$ Fisher's exact test) and are highly enriched in IRF1, RELA (also known as p65) and RUNX1 binding ($p < 10^{-10}$, Fisher's exact test). The high evolutionary sequence constraint that we required to define enhanceosome candidates translates to low variation across the human population. Indeed, human regulatory regions with similar evolutionary constraint are depleted of SNPs, having an average of only 25 SNPs compared to an average of 400 (and a minimum of 369) in similarly sized genomic regions.

Regulatory regions with conserved activity and temporal patterns regulate highly induced genes with shared kinetics

Response to LPS affects both the acetylation and chromatin accessibility of thousands of enhancers (Figure 6A, S6A-C). Although the chromatin state of most enhancers (72%) is unaffected by LPS, enhancers that show temporal kinetics tend to associate with genes having similar transcriptional kinetics. Indeed, regions whose DNA accessibility increases upon LPS stimulation are associated with induced genes (1.6-fold enrichment) while regions that close over

time are associated with downregulated genes (2-fold enrichment, Figure 6B). We further observed a clear enrichment of cPWMs, including NFKB, STAT and AP1 motifs, on DNA accessible regions that show increased ATAC signal after LPS stimulation. On the other hand, cPWMs associated with ETV and STAT transcription factor families are enriched in accessible DNA regions that become less accessible in response to LPS. Enrichment of ETV and STAT motifs on regions that lose availability is consistent with their reported repressive function (Icardi et al., 2012; Mavrothalassitis and Ghysdael, 2000) (Figure 6C). It is interesting that STAT motifs are enriched in both down and upregulated elements. These motifs may recruit different members of the STAT family or attract complexes involving different TFs that modulate the STAT TF function. Our previously generated mouse binding data for STAT1 and STAT2 shows that these proteins bind mostly to regions that become increasingly accessible upon LPS stimulation. This suggests that motifs in regions whose DNA availability decreases after LPS stimulation are likely bound by different STAT TFs or other factors that can bind this motif.

To further determine the importance of cPWMs in regulating the LPS response, we proceeded to build a random forest classifier as above, but this time we associated each cPWM with three features per gene: the number of cPWMs in regulatory regions with increased, diminished or unchanged DNA accessibility upon LPS stimulation. This dramatically improved the model performance which now showed an average AUROC of 0.82 in mouse in a 10-fold cross-validation and an AUROC of 0.78 when applied to human (Figure 6D). This highlights the importance of the chromatin context and helps explain the weaker performance of a model that was trained on sequence alone.

Given that regions with LPS-responsive chromatin dynamics were important when evaluating sequence features, we next investigated the conservation of DNA accessibility dynamics. Interestingly, although regions with LPS-induced DNA accessibility are present in both human and mouse (28% and 30%, respectively), very few are LPS-responsive in both. By simultaneously clustering ATAC-Seq peaks from ECAs that had significant LPS-induced signal changes in at least one species (Figure S6D), we found that only 500 such regions (13%) are responsive in both mouse and human DCs (Figure 6E).

These 500 regions are associated with 325 genes, of which 57% have similar expression kinetics in human and mouse, while only 21% of all the expressed genes have similar expression patterns in both species ($p < 10^{-20}$, Fisher-exact test, Figure 6F). Genes associated with these regions have much higher induction levels and reached higher maximal expression than other genes with no difference in baseline expression (Figure 6G-I, 6G: $p < 2.2e-16$ Wilcox-rank test, 6H: $p < 2.2e-16$ Wilcox-rank test, 6I: not significant Fisher-exact test). They include cytokines (e.g IL1B, IL6) and key transcription factors (e.g. REL, NFKB1, BCL2, NFKBIZ) (Figure 6J, p-adjusted < 0.004). Regions with conserved dynamics are enriched near genes with similar temporal dynamics and have maintained enhancer activity since the rodent/primate divergence. This suggests that they are crucial elements regulating this set of genes.

Transposable elements are enriched in cis-regulatory regions of LPS-induced genes

Most *cis*-regulatory elements are composed of ancestral sequence (Cheng et al., 2014; Villar et al., 2015) (Figure 2B). Therefore turnover of ancestral activity rather than sequence seems to be

the major force reshaping regulatory regions. Sequence changes can still be an important source of difference between the human and mouse response. Since lineage specific transposable elements (TEs) have been shown to significantly modify transcriptional networks (Lowe and Haussler, 2012; Wang et al., 2007b), we next sought to determine whether TEs have contributed to regulatory sequence involved in the LPS response. We identified 25 families of TEs in mouse and 15 in human that are enriched in regulatory regions (enhancers or promoters) of induced genes (Figure 7A). These enriched TE families fall into two categories: those that were actively mobile prior to the human-mouse divergence, and newer elements that have only been active in either the mouse or human lineage. The majority belong to one of the ancestral TE families of Mammalian-wide interspersed repeats (MIRs), with MIR3 elements being the most enriched (Figure 7B) and having the largest number of elements within regulatory regions. MIR elements are some of the oldest (Smit and Riggs, 1995) and most conserved families of mobile elements (Jjingo et al., 2014), and have been reported to contribute to the regulation of cell type specific expression (Jjingo et al., 2014). Our data further suggests that MIRs, and MIR3 in particular, have been co-opted into regulation of innate immune responses prior to the euarchontoglires ancestor. As one might expect for important regulatory sequences, we observed that MIRs have been under clear purifying selection (Figure S7A).

Lineage specific TEs enriched in DC regulatory regions include mainly endogenous retroviral Long Terminal Repeat (LTR) elements. We found that elements from these families (ORR1E in mouse and THE1A and THE1C in human) tend to be positioned at the most accessible regions within enhancers, possibly indicating a role in creating or facilitating opening of chromatin that is more favorable to transcription factor binding and more likely to function as a regulatory element (Figure 7C).

Discussion

Massive parallel sequencing has revealed hundreds of thousands of active non-coding regions, most of which are classified by their chromatin signatures as enhancers or long noncoding RNAs (lncRNAs). Comparative analyses of enhancers and lncRNAs have shown that although the majority are encoded by ancestral sequence, their activity is generally species-specific (Chen et al., 2016; Cheng et al., 2014; Kutter et al., 2012; Necsulea et al., 2014; Ponjavic et al., 2007; Ulitsky, 2016; Vierstra et al., 2014; Villar et al., 2015; Washietl et al., 2014). Here we showed that a higher fraction of enhancers that regulate specific pathways tend to be conserved over longer evolutionary time.

As opposed to previous studies, we used a dynamic system and focused on temporal expression patterns rather than steady state expression. In this system, changes in mRNA levels in early time points are mostly the result of transcription rather than post-transcriptional processes (Rabani et al., 2011); this helps isolating and measuring the contribution of *cis*-regulatory elements to expression changes. Temporal analysis also allowed us to study different regulatory programs individually rather than analyzing all regulatory programs together or by broad functional classes (Figure 1B). As a result, we were able to find that regulatory element conservation is not homogeneous across all enhancers, but rather that it differs across programs. We find that regulatory elements associated with shared early-induced genes are conserved at twice the rate

than those associated with other expressed genes. Not only regulatory element activity is conserved, but also the underlying sequence is under purifying selection. This allowed us to use comparative sequence analysis to identify a large set of constrained sequence motifs within active enhancers. Functional validation of these enhancers as well as the novel motifs we found will be critical, but this study provides a clear path towards the goal of functionally characterizing a well-defined set of regulatory regions involved in well-understood cellular processes.

It is interesting that, besides enhancers associated with shared induced genes, the other set of enhancers preserved since the euarchontoglires ancestor are ubiquitous or active in progenitor cells but are not associated with genes induced by TLR4 signaling. Instead, these enhancers tend to lose active marks following LPS stimulation. This is consistent with previous observations that basic cellular processes are passively downregulated upon induction of a large transcriptional program (Cheng et al., 2009; Garber et al., 2012), perhaps due to a shift of limited resources towards the response to immune challenge.

The greater conservation of enhancers associated with early-induced genes is surprising, with conserved enhancers accounting for 40% of all enhancers associated with these genes. This raises an interesting question: why are the regulatory elements of early-induced genes under stronger selection? It is reasonable to argue that this initial wave of transcription triggers a program that, although necessary for immune defense, is deleterious to the individual when misregulated. Tight control of the initiation of the program may be critical to avoid unwanted harm. It is also interesting that in our previous analysis of mouse DC enhancers we observed a low degree of sequence constraint of most enhancers, and concluded that early-induced genes were regulated by a highly redundant regulatory architecture that functioned by recruiting many different TFs in a nonspecific fashion. Our comparison with human DCs paints a more nuanced picture. Early-induced genes are regulated by a mix of highly constrained enhancers that have been preserved over hundreds of millions of years and newly evolved species-specific enhancers. The ECAs have clear signatures of undergoing purifying selection and may be necessary for induction. Nonetheless, the majority of enhancers is species-specific and may play redundant, subtler roles or have no impact on gene expression. Further functional studies will be needed to determine how different enhancers function and how they interact to produce reproducible, precise patterns of expression.

Our study sheds some light on the long-standing question of how selection acts on gene expression (Gilad et al., 2006). Although our study was not designed to answer this, we find two very clear modes of selection. On one hand, highly induced genes tend to have shared induction and are regulated by conserved regulatory elements. These observations are consistent with strong stabilizing selection. On the other hand, there is great divergence among genes with mild induction, which is consistent with neutral selection (Gilad et al., 2006). We reason that, while mutations that disrupt the level and timing of highly induced genes may have strong deleterious effect, for genes that are mildly induced, changes are tolerated.

Our comparative map provides a unique resource for future studies of *in-vitro*-derived DCs. It provides a reference map of the genomic elements that can be mapped and translated from a

mouse model to human biology. Further, recent reports on underlying differences in the cell types obtained in mouse and human DC *in vitro* cultures (Helft et al., 2015) highlights the need to compare these two systems at the molecular level. In this work, we focused on understanding both the similarities and differences between the two. Given the overall similarity in TF expression, this system offers a deep platform to understand the impact of *cis*-regulatory changes on expression.

Author Contributions

P.V. and E.D designed and performed the data analysis. S.A. and N.Y. designed and implemented the induced gene classifier. P.V. performed the mouse DC experiments and constructed the high-throughput sequencing libraries. B.T. developed the ATAC-Seq processing pipeline and advised in data analysis. S.M. performed the human DC experiments. A.N. constructed the high-throughput sequencing libraries for human DCs. A.K. helped implement the data processing pipelines and managed sample metadata. X.Z. designed and implemented the gene expression spectral clustering algorithm. W.D. and E.D. performed the TE analysis. P.M. Supervised, developed protocols and planned all high-throughput sequencing experiments. M.G., J.L., and N.Y. conceived the project, advised on the analysis and data collection and supervised the research. E.D., P.V. and M.G. wrote the paper with input from all authors.

Acknowledgments

We want to thank Mitch Guttman, Jenny Chen, Ido Amit, Zhiping Weng, Scott Wolfe and members of the Garber Lab for valuable discussions and comments on the manuscript. We thank Idan Gabdank for help managing our data submission and to Sigrid Knemeyer for assistance with figures. This project was supported by the NHGRI U01 HG007910 (M.G., J.L., N.Y.), NIDA DP1DA034990 (J.L.), NIAID RO1AI111809 (M.G., J.L) and NCATS UL1 TR001453-02 (M.G.).

Declaration of Interests

The authors declare no competing interests.

Figures

Figure 1

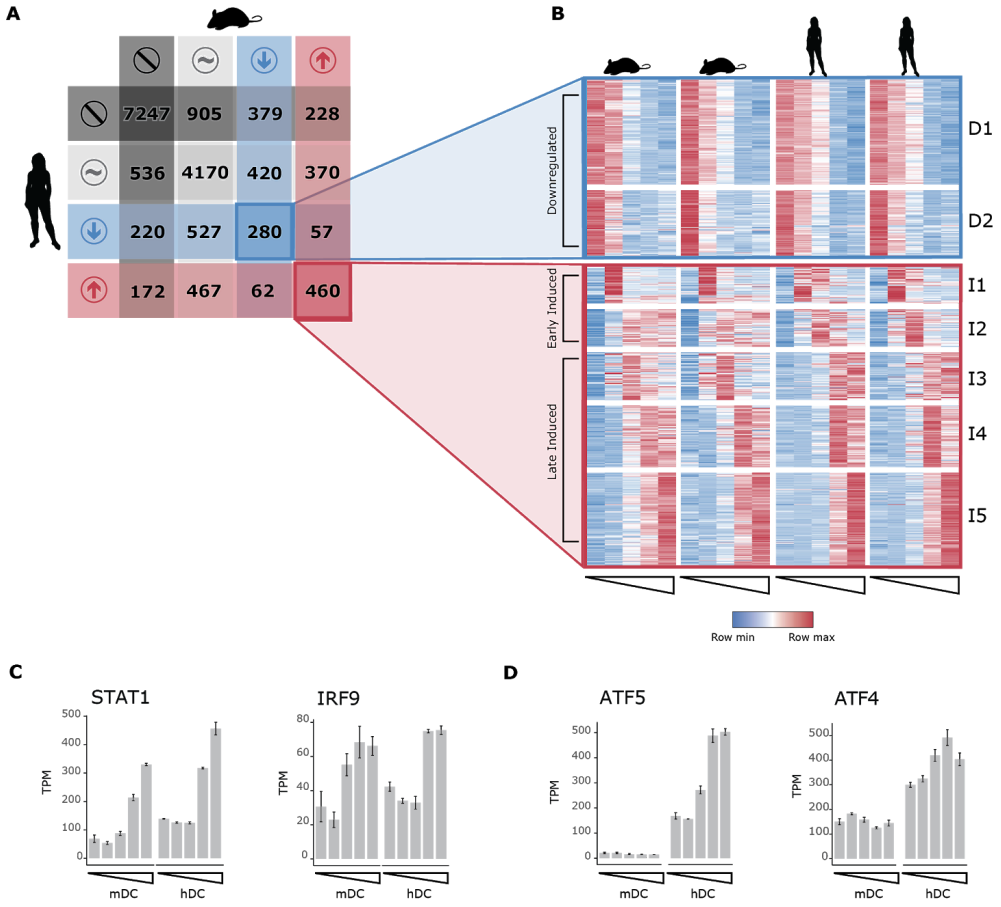


Figure 1: Highly induced LPS-responsive genes have similar expression kinetics in human and mouse dendritic cells. A) Classification of 16,500 homolog genes in mouse and human as not expressed (dark grey), expressed without significant change after LPS stimulation (light grey), downregulated (blue) or induced (red) B) Heatmap showing normalized expression values for genes with shared response to LPS across five timepoints (Unstimulated, 1h, 2h, 4h and 6h post-LPS) in DCs derived from two different C57BL/6 mouse (left) and two human donors (right). Genes were grouped by spectral clustering into two clusters of shared downregulated genes (D1 and D2, top), and five clusters of shared induced genes (I1-I5, bottom). Induced gene clusters can be classified as early (I1 and I2) or late (I3, I4 and I5). C) Average normalized expression (TPM) for two shared late-induced transcription factors (TFs), Stat1 and Irf9. D) Average normalized expression (TPM) for ATF family TFs with species-specific response.

Figure 2

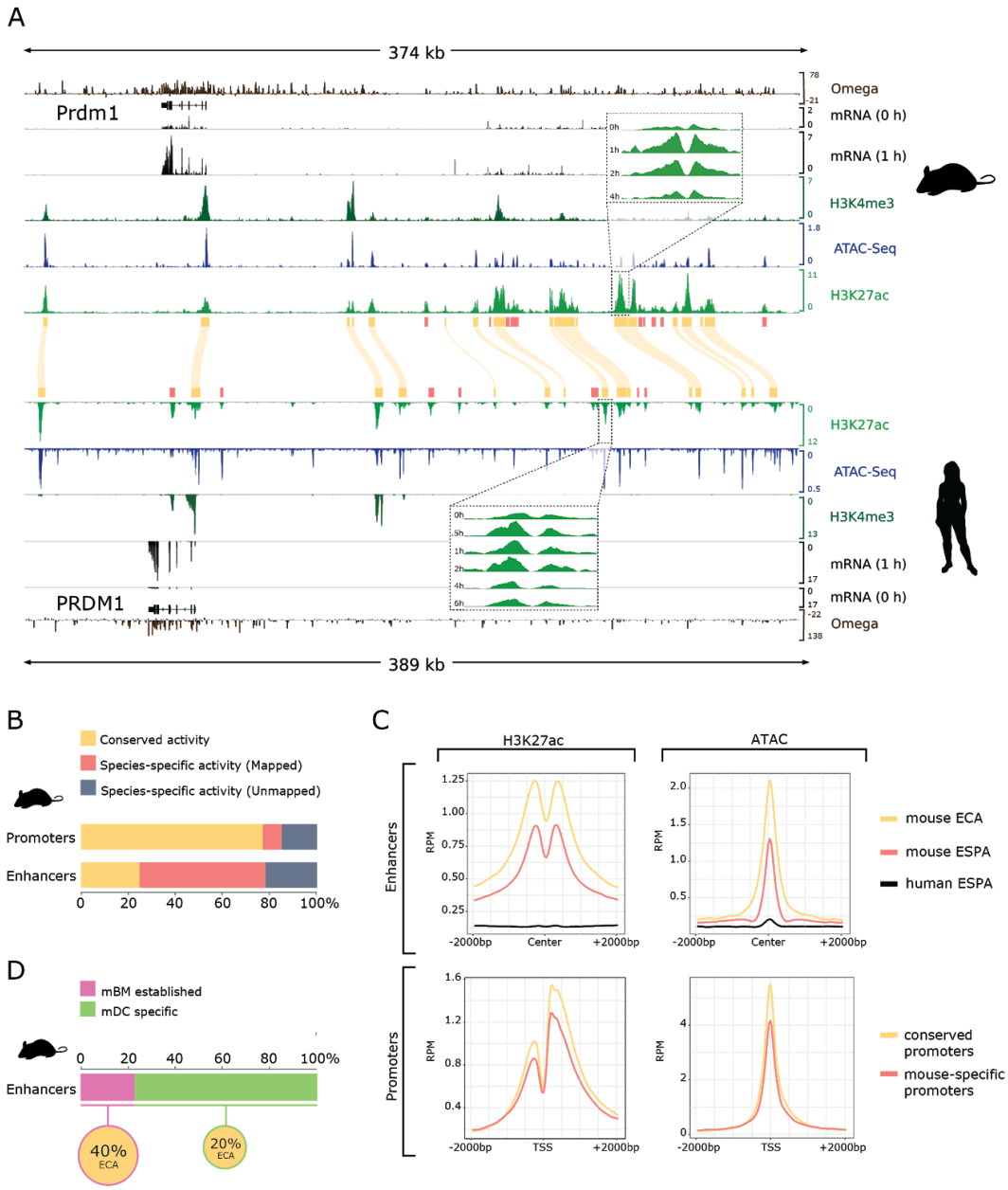


Figure 2: **Rapid turnover of enhancer elements**. A) Integrative Genome Viewer diagram of the PRDM1 regulatory region in both mouse (top) and human (bottom) displaying the data used in this study. Tracks display from top to bottom: sequence conservation as estimated by SiPhy (Omega), RefSeq gene annotations, RNA-Seq coverage for unstimulated and one hour post-LPS, overlaid H3K4Me3, ATAC and H3K27ac coverage. Human data in reverse orientation, yellow boxes and curved lines indicate conserved H3K27ac peaks (regulatory regions with conserved activity: promoters or ECAs). Inlets show individual tracks for H3K27ac time course after LPS stimulation. Red boxes indicate H3K27ac peaks with species-specific activity. B) Proportion of regulatory regions with conserved activity: conserved promoters or ECAs, mouse-specific with clear human homologous sequence (mapped promoters or ESPA) and mouse-specific with no clear homologous sequence in human (unmapped promoters or ESPA) C) Average signal for

mouse H3K27ac (left) and ATAC-Seq (right) signal over regulatory elements. Enhancer (top) H3K27ac or ATAC-Seq signal is centered in open regions, defined by ATAC-Seq peaks. Promoter (bottom) H3K27ac or ATAC-Seq signal is centered in the TSS. Data is shown for conserved enhancers and promoters (yellow), mouse-specific enhancers and promoters (red) and all other mouse genome coordinates for mapped human-specific enhancers and promoters (black). RPM = reads per million mapped reads D) Fraction of mouse enhancers that are active (pre-established) in bone marrow (mBM) cells and enhancers that are mDC specific, and fraction of mBM pre-established or mDC specific enhancers that are conserved (ECA).

Figure 3

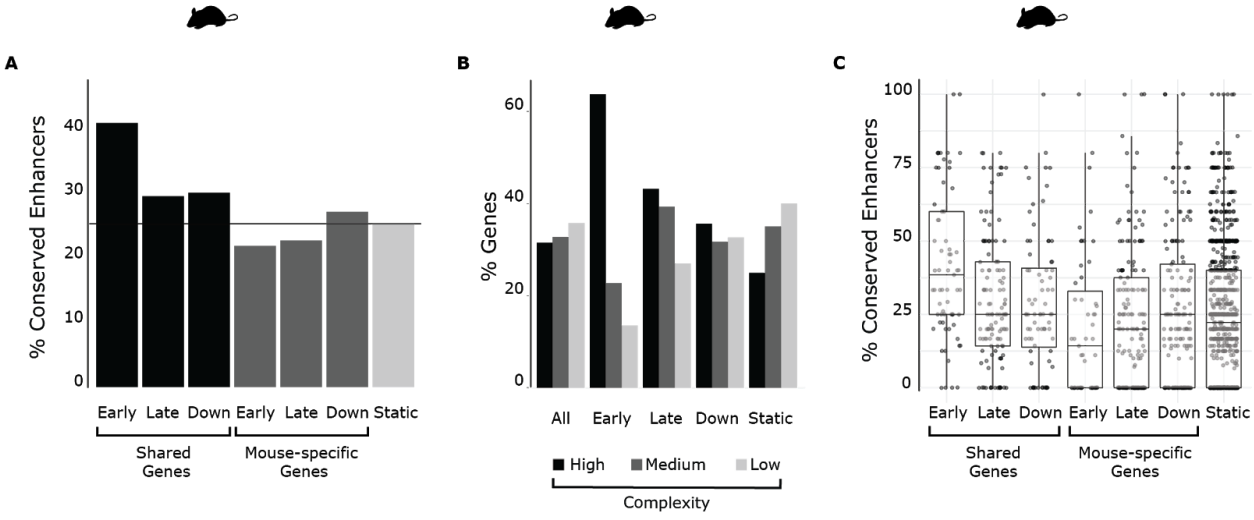


Figure 3: Genes with shared transcriptional response to LPS have complex regulatory loci and a higher conservation of enhancer activity. A) Fraction of ECAs that are associated to genes that are early-induced, late-induced or downregulated upon stimulation with LPS in mouse DCs. The black horizontal line shows the average enhancer conservation for all genes B) Fraction of genes in temporal clusters that are associated to high-, medium- or low-complexity enhancer loci. C) Fraction of ECAs in high complexity genes that have shared or species-specific response. The response patterns are: early-induced, late-induced, downregulated or unchanged.

Figure 4

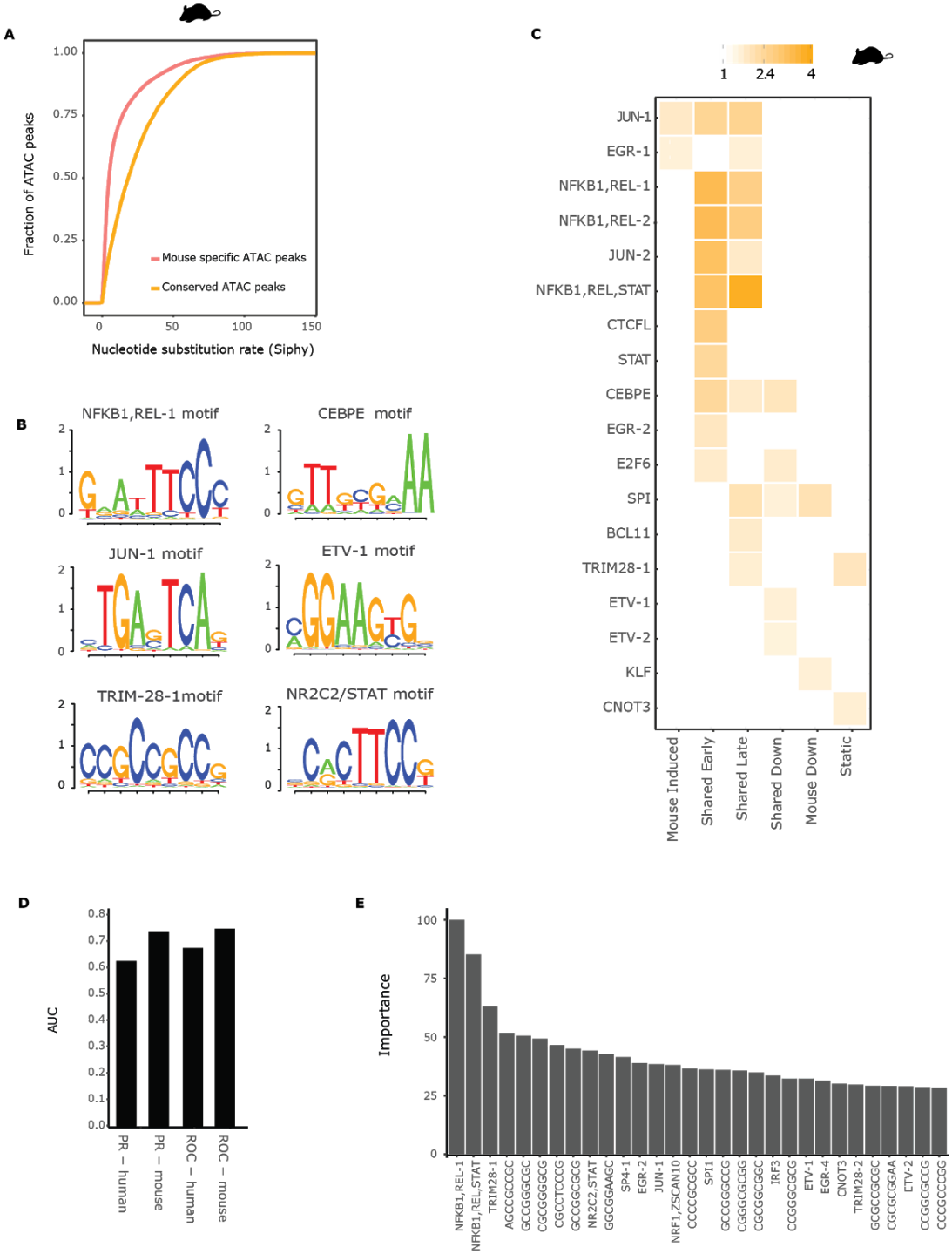


Figure 4: Enhancers with conserved activity contain a conserved lexicon. A) Distribution of SiPhy omega log-odds scores for 200bp regions around the summits of ATAC-seq peaks that have conserved signal (yellow) and species-specific signal (red) in mouse DCs. B) Examples of sequence logos of the clusters of kmers obtained after clustering the sequences in ATAC regions with conserved signal that have a log-odds score greater than 30. C) Enrichment heatmap

showing the observed over expected values for each motif in ATAC-seq peaks with conserved signal associated to the gene groups defined in Figure 1. D) AUC of the PR and ROC curves of a random forest model, predicting whether a gene will be induced or maintain constant expression following LPS stimulation. The features were the number of instances of each cPWM across all regulatory regions of a gene. E) Feature importance of the classifier, defined as the difference in mean accuracy across all trees between the model and the model after permuting the feature. The importance values were then scaled to span the range of 0 to 100. The 30 features with the highest importance values are shown.

Figure 5

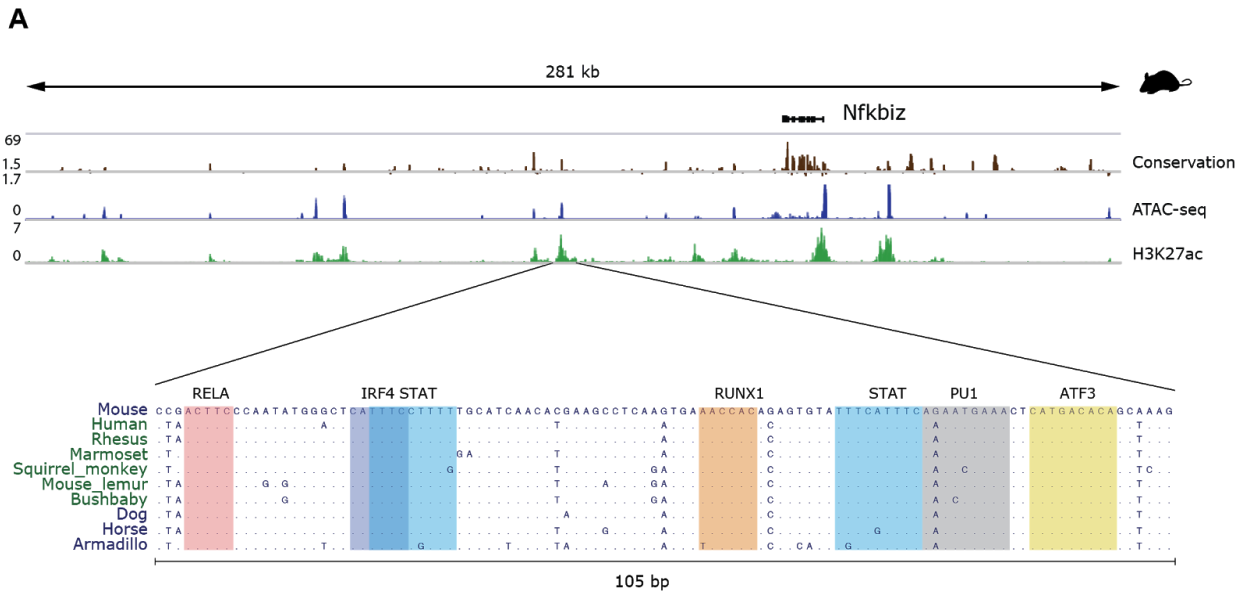


Figure 5: Candidate enhanceosome regions are highly conserved and bound by multiple TFs. A) Example of an enhanceosome-like regulatory element in the NFKBIZ locus in mouse (top panel) showing the multiple sequence alignment of the conserved DNA accessible region.

Figure 6

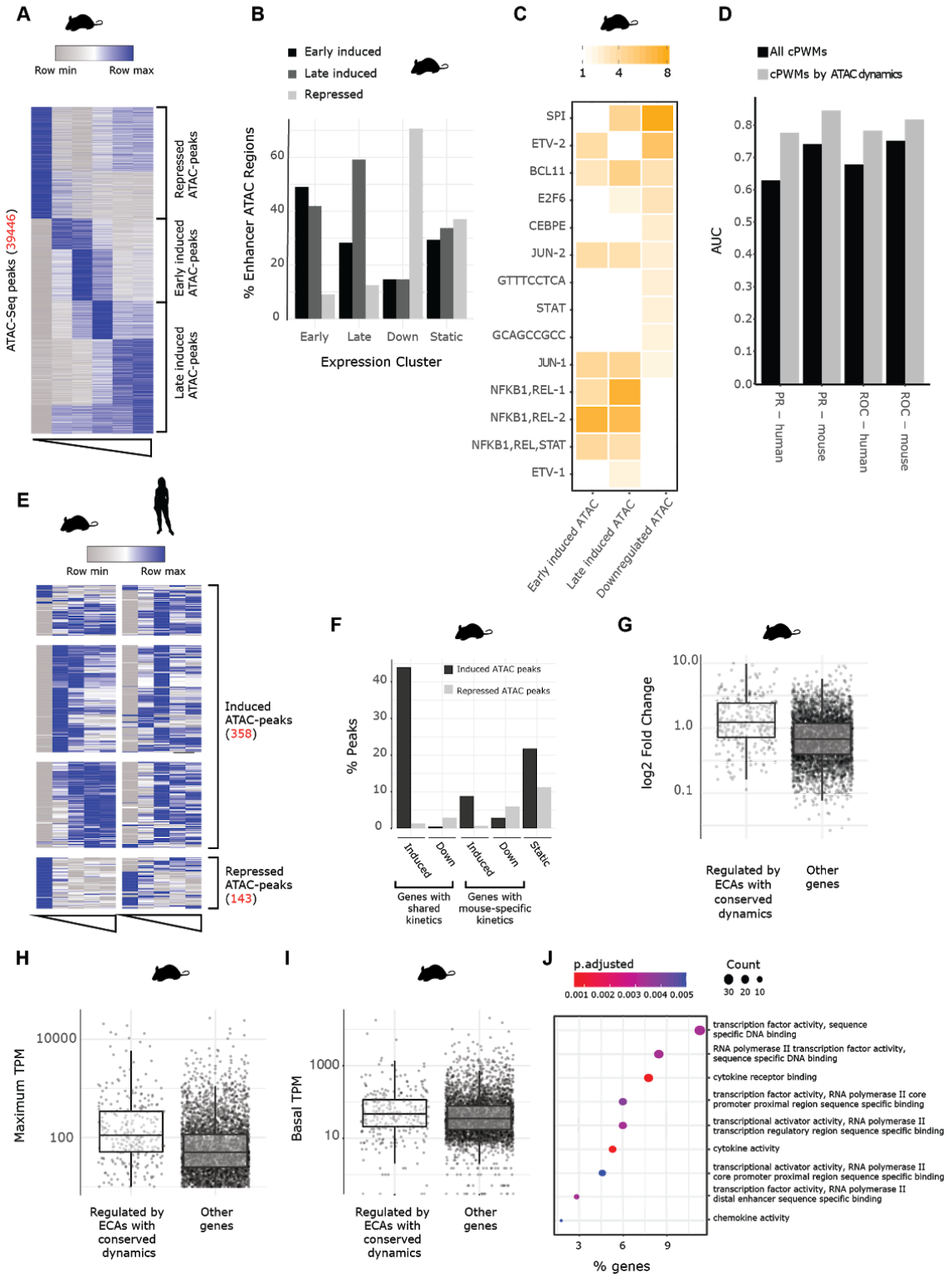


Figure 6: Regulatory regions with conserved activity and conserved kinetics regulate genes with shared induction kinetics. A) Heatmap showing k-means clustering of temporal patterns

of mean signal per bp for ATAC-Seq peaks (at enhancer or promoter regions) with dynamic response to LPS in mouse DCs (Unstimulated, 30 minutes, 1 hours, 2 hours, 4 hours and 6 hours). Regions were classified as repressed, early-induced or late-induced. B) Fraction of early-induced, late-induced, downregulated or non-changing genes that are associated to dynamic ATAC peaks. C) Enrichment of cPWMs in ATAC peaks that are under purifying selection (Fig 4A) clustered into temporal groups. D) AUC of the PR and ROC curves of a random forest model, predicting whether a gene will be induced or maintain constant expression following LPS stimulation. The features for each model were the number of instances of each cPWM across all regulatory regions of a gene (black bars), or all instances separated by the temporal pattern of the regulatory element (grey bars) E) Heatmap showing the temporal patterns of ATAC-seq peaks with conserved signal that are dynamic in both mouse and human. F) Enrichment of ATAC-seq peaks with conserved signal associated to genes that are induced in both mouse and human DCs, induced only in mouse DCs, downregulated in both mouse and human DCs, downregulated only in mouse DCs and not responsive to LPS in mouse DCs. G-I) The maximum absolute fold change, maximum tpm and baseline tpm of genes that are associated with ATAC-seq peaks with conserved signal that have same temporal response in both mouse and human versus all other genes J) Gene ontology analysis of genes associated with regulatory regions with conserved LPS response kinetics.

Figure 7

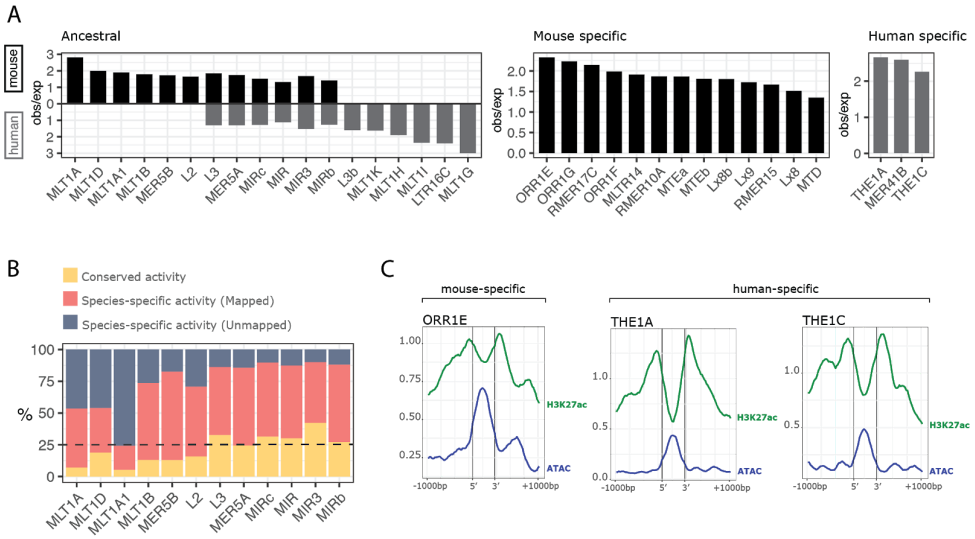


Figure 7: Mobile elements of ancestral and recent origin have reshaped response to environmental stimulus. A) Families of transposable elements (TEs) enriched in regulatory regions of induced genes in mouse and human. Observed over expected (obs/exp) values are shown for each TE only when the enrichment is significant in that species (p value < 0.004, permutation test; adjusted p < 0.05). Panels show families of TEs that have instances in the mouse and human genomes (Ancestral, Left), only in mouse (Mouse specific, Center), or only in human (Human specific, Right). B) Conservation rate of the enhancer regions that overlap each ancestral TE. C) Average aggregation signal of H3K27ac and ATAC-Seq over TE instances that overlap regulatory elements. Region is centered in each TE instance, delimited by the vertical bars, and the 2kb surrounding region is shown.

References

Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science *326*, 257–263.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. *94*, 890–898.

Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A.J., Funnell, A.P.W., Goncalves, A., et al. (2014). Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. Elife *3*, e02626.

Blashfield, R.K. (1991). Finding groups in data-an introduction to cluster-analysis-Kaufman, L, Rousseeuw, PJ.

Bogdan, C. (2001). Nitric oxide and the immune response. Nat. Immunol. *2*, 907–916.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1–9.

Cannavò, E., Khoueiry, P., Garfield, D.A., Geeleher, P., Zichner, T., Gustafson, E.H., Ciglar, L., Korbel, J.O., and Furlong, E.E.M. (2016). Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr. Biol. *26*, 38–51.

Chen, J., Shishkin, A.A., Zhu, X., Kadri, S., Maza, I., Guttman, M., Hanna, J.H., Regev, A., and Garber, M. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol. *17*, 19.

Cheng, Y., Wu, W., Kumar, S.A., Yu, D., Deng, W., Tripic, T., King, D.C., Chen, K.-B., Zhang, Y., Drautz, D., et al. (2009). Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. Genome Res. *19*, 2172–2184.

Cheng, Y., Ma, Z., Kim, B.-H., Wu, W., Cayting, P., Boyle, A.P., Sundaram, V., Xing, X., Dogan, N., Li, J., et al. (2014). Principles of regulatory information conservation between mouse and human. Nature *515*, 371–375.

Chew, J.-L., Loh, Y.-H., Zhang, W., Chen, X., Tam, W.-L., Yeap, L.-S., Li, P., Ang, Y.-S., Lim, B., Robson, P., et al. (2005). Reciprocal transcriptional regulation of Pou5f1 and Sox2 via the Oct4/Sox2 complex in embryonic stem cells. Mol. Cell. Biol. *25*, 6031–6046.

Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.

Crocker, J., and Erives, A. (2008). A closer look at the eve stripe 2 enhancers of Drosophila and Themira. PLoS Genet. *4*, e1000276.

Dunipace, L., Ozdemir, A., and Stathopoulos, A. (2011). Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. Development *138*, 4075–4084.

Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics *25*, i54–i62.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M.A. (2016). gkmSVM: an R package for gapped-kmer SVM. Bioinformatics *32*, 2205–2207.

Gilad, Y., Oshlack, A., and Rifkin, S.A. (2006). Natural selection on gene expression. Trends Genet. *22*, 456–461.

González, A.J., Setty, M., and Leslie, C.S. (2015). Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. Nature Publishing Group *47*, 1249–1259.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

Harding, H.P., Zhang, Y., Zeng, H., Novoa, I., Lu, P.D., Calfon, M., Sadri, N., Yun, C., Popko, B., Paules, R., et al. (2003). An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. Mol. Cell *11*, 619–633.

He, Q., Bardet, A.F., Patton, B., Purvis, J., Johnston, J., Paulson, A., Gogol, M., Stark, A., and Zeitlinger, J. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. Nat. Genet. *43*, 414–420.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Helft, J., Böttcher, J., Chakravarty, P., Zelenay, S., Huotari, J., Schraml, B.U., Goubau, D., and Reis e Sousa, C. (2015). GM-CSF Mouse Bone Marrow Cultures Comprise a Heterogeneous Population of CD11c+MHCII+ Macrophages and Dendritic Cells. Immunity *42*, 1197–1211.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598.

Icardi, L., Mori, R., Gesellchen, V., Eyckerman, S., De Cauwer, L., Verhelst, J., Vercauteren, K., Saelens, X., Meuleman, P., Leroux-Roels, G., et al. (2012). The Sin3a repressor complex is a master regulator of STAT transcriptional activity. Proc. Natl. Acad. Sci. U. S. A. *109*, 12058–12063.

Jjingo, D., Conley, A.B., Wang, J., Mariño-Ramírez, L., Lunyak, V.V., and Jordan, I.K. (2014). Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. Mob. DNA *5*, 14.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.-H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264–268.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., and Others caret: Classification and regression training, 2011. R Package Version *4*.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. *42*, 631–634.

Kutter, C., Watt, S., Stefflova, K., Wilson, M.D., Goncalves, A., Ponting, C.P., Odom, D.T., and Marques, A.C. (2012). Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. *8*, e1002841.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the

developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liaw, A., Wiener, M., and Others (2002). Classification and regression by randomForest. R News *2*, 18–22.

Love, M., Anders, S., and Huber, W. (2014). Differential analysis of count data--the DESeq2 package. Genome Biol. *15*, 550.

Lowe, C.B., and Haussler, D. (2012). 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. PLoS One *7*, e43128.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. Nature *403*, 564–567.

von Luxburg, U. (2007). A tutorial on spectral clustering. Stat. Comput. *17*, 395–416.

Mavrothalassitis, G., and Ghysdael, J. (2000). Proteins of the ETS family with transcriptional repressor activity. Oncogene *19*, 6524–6532.

Medzhitov, R., and Horng, T. (2009). Transcriptional control of the inflammatory response. Nat. Rev. Immunol. *9*, 692–703.

Mellor, A.L., and Munn, D.H. (2004). IDO expression by dendritic cells: tolerance and tryptophan catabolism. Nat. Rev. Immunol. *4*, 762–774.

Mestas, J., and Hughes, C.C.W. (2004). Of mice and not men: differences between mouse and human immunology. J. Immunol. *172*, 2731–2738.

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., and Rosen, E.D. (2010). Comparative epigenomic analysis of murine and human adipogenesis. Cell *143*, 156–169.

NCBI Resource Coordinators (2016). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *44*, D7–D19.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat. Genet. *39*, 730–732.

Ong, C.-T., and Corces, V.G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat. Rev. Genet. *12*, 283–293.

Panne, D., Maniatis, T., and Harrison, S.C. (2007). An Atomic Model of the Interferon-β Enhanceosome. Cell *129*, 1111–1123.

Parnas, O., Jovanovic, M., Eisenhaure, T.M., Herbst, R.H., Dixit, A., Ye, C.J., Przybylski, D., Platt, R.J., Tirosh, I., Sanjana, N.E., et al. (2015). A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. Cell *162*, 675–686.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. *20*, 1562–1567.

Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res. *17*, 556–565.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. *40*, D130–D135.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Reinhard, C., Bottinelli, D., Kim, B., and Luban, J. (2014). Vpx rescue of HIV-1 from the antiviral state in mature dendritic cells is independent of the intracellular deoxynucleotide concentration. Retrovirology *11*, 12.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science *328*, 1036–1040.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics *15*, 284.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Smit, A.F., and Riggs, A.D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. *23*, 98–102.

Smit, A., Hubley, R., and Green, P. (2004). RepeatMasker Open-3.0. 2004. Seattle (WA): Institute for Systems Biology.

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell *83*, 1091–1100.

Toshchakov, V., Jones, B.W., Perera, P.-Y., Thomas, K., Cody, M.J., Zhang, S., Williams, B.R.G., Major, J., Hamilton, T.A., Fenton, M.J., et al. (2002). TLR4, but not TLR2, mediates IFN-beta-induced STAT1alpha/beta-dependent gene expression in macrophages. Nat. Immunol. *3*, 392–398.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat. Rev. Genet. *17*, 601–614.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R.S., Stehling-Sun, S., Sabo, P.J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 1246426.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Wang, H., Lin, G., and Zhang, Z. (2007a). ATF5 promotes cell survival through transcriptional activation of Hsp27 in H9c2 cells. Cell Biol. Int. *31*, 1309–1315.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812.

Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. (2007b). Species-specific endogenous retroviruses shape the

transcriptional network of the human tumor suppressor protein p53. Proceedings of the National Academy of Sciences *104*, 18613–18618.

Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res.

Watatani, Y., Kimura, N., Shimizu, Y.I., Akiyama, I., Tonaki, D., Hirose, H., Takahashi, S., and Takahashi, Y. (2007). Amino acid limitation induces expression of ATF5 mRNA at the post-transcriptional level. Life Sci. *80*, 879–885.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. Nucleic Acids Res. *41*, D165–D170.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al. (2014). A comparative encyclopedia of DNA elements in the mouse genome. Nature *515*, 355–364.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Supplementary information
Supplementary information, methods and figures are found below. Supplementary tables can be found in the following link:
https://www.sciencedirect.com/science/article/pii/S2405471218300024?via%3Dihub#appsec2

STAR Methods

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
List of 147 data-sets used in this study	This paper	Table S4
Human 10-mers substitution rates	This paper	http://garberlab.umassmed.edu/data/conservation/hg19/omega/
Mouse 10-mers substitution rates	This paper	http://garberlab.umassmed.edu/data/conservation/mm10/mm10.omega
Software or Algorithms		
gkm-SVM	(Ghandi et al., 2016)	v1.3
Spectral clustering	This paper	https://github.com/nimezhu/ClsViz
Trimmomatic	(Bolger et al., 2014)	V0.32
Bowtie2	(Langmead and Salzberg, 2012)	v2.2.23
Samtools	(Li et al., 2009)	v0.1.19
DESeq2	(Love et al., 2014)	v1.10.1

Bedtools	(Quinlan and Hall, 2010)	V2.25.0
MACS2	(Zhang et al., 2008)	V2
IGVtools	(Robinson et al., 2011)	V2.3.31
RSEM	(Li and Dewey, 2011)	v1.2.28
SiPhy	(Garber et al., 2009)	https://github.com/garber-lab/siphy
Antibodies & Reagents		
H3K27ac	Diagenode	C15410196
H3K4me3	EMD Millipore	05-745R
Ovation Human FFPE RNA-Seq Library System	NuGen	0340
Ovation mouse RNA-Seq Library System	NuGen	0348
RNeasy mini plus kit	Qiagen	74134
Nextra TDE-1 transposase,	Illumina	FC-121-1030
Covaris tru-ChIP Chromatin Shearing and Reagent Kit	Covaris	520154
Agencourt AMPure XP	Beckman Coulter	A63880
GMCSF	Miltenyi	130-095-735

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Manuel Garber (Manuel.Garber@umassmed.edu).

Experimental Model and Subject Details

Human Subjects:

Anonymous, healthy donor leukopaks (New York Biologics, Southampton, NY), were used in accordance with UMMS-IRB protocol ID #H00004971

Mice:

All mice were housed in specific pathogen-free condition in accordance with the Institutional Animal Care and Use Committee of the University of Massachusetts Medical School. C57BL6 female mice were euthanized at 6-8 weeks of age to harvest bone marrow.

Method Details

Cell culture

All cells were maintained at 37°C in 5% CO2 humidified incubators.

Human monocyte derived dendritic cells

Human dendritic cells were derived from peripheral blood mononuclear cells (PBMCs) isolated from de-identified, healthy donor leukopaks (New York Biologics, Southampton, NY), in accordance with UMMS-IRB protocol ID #H00004971. Mononuclear leukocytes were isolated by gradient centrifugation on Histopaque-1077 (Sigma-Aldrich, St. Louis, MO). CD14+ mononuclear cells were enriched via positive selection using anti-CD14 antibody MicroBead conjugates (Miltenyi, San Diego, CA), according to the manufacturer's protocol. CD14+ cells were then plated at a density of 1 to 2 x 10^6 cells/ml in RPMI-1640 supplemented with 5% heat inactivated human AB+ serum (Omega Scientific, Tarzana, CA), 20 mM L-glutamine (ThermoFisher, Waltham, MA), 25 mM HEPES pH 7.2 (Sigma-Aldrich), 1 mM sodium pyruvate (ThermoFisher), and 1 x MEM non-essential amino acids (ThermoFisher). Differentiation of the CD14+ monocytes into dendritic cells (human DCs) was promoted by addition of recombinant human GM-CSF and human IL-4; cytokines were produced from HEK293 cells stably transduced with pAIP-hGMCSF-co or pAIP-hIL4-co, respectively, as previously described (Reinhard et al., 2014), with each cytokine supernatant added at a dilution of 1:100.

Mouse bone marrow derived dendritic cells

Mouse dendritic cells were derived from bone marrow harvested from 6-8 week old female C57BL6 mice. Bone marrow was then dissociated into single cells and filtered through 70um cell strainer. The cells were then incubated with red blood cell lysis buffer for 5 minutes. To differentiate bone marrow to dendritic cells, bone marrow cells were plated at 200,000 cells/mL in non-tissue culture treated plates. These cells were supplemented with media on day 2 and day 7. On day 5 cells were harvested and resuspended in fresh media. On day 8 all the floating cells were collected as mouse bone marrow derived dendritic cells. The media used for culturing and differentiating contains RPMI (Gibco) supplemented with 10% heat inactivated FBS (Gibco),

β-mercaptoethanol (50uM, Gibco), MEM non-essential amino acids (1X, Gibco), sodium pyruvate (1mM, Gibco), and GM-CSF (20 ng/ml; Miltenyi).

Library preparation and Sequencing

RNA-seq

Total RNA was isolated from frozen dendritic cell pellets using the RNeasy mini plus kit (QIAGEN). The RNAs were additionally treated with RNase-free DNase I for 15 minutes at room temperature to eliminate most genomic DNA. RNA-Seq libraries were prepared from 70 ng of starting RNA using the Ovation Human FFPE RNA-Seq Library System (NuGEN) or Ovation mouse RNA-Seq Library System (NuGEN), according to the manufacturer's protocol. Fragmentation of the cDNA was achieved by sonication using the M220 sonicator (Covaris) with the following conditions: sonication time = 350 seconds; temp = 20°C; peak power = 50; duty factor = 20; cycles/burst = 200. The quality of the isolated RNA, as well as of the final libraries, was assessed using the 2100 Bioanalyzer (Agilent) and Qubit (Invitrogen). The libraries were pooled according to donor in equimolar ratios and denatured. Pooled libraries were sequenced for 2 x 100 cycles to obtain paired end reads, using a HiSeq 2000 (Illumina) for human DCs and 2 x 75 cycles using Nextseq 500 for mouse DCs.

ATAC-Seq

For each time point, 5 x 10^5 scraped DC's were collected by centrifugation 500 x g for 5 min. and lysed for ATAC-seq following the protocol described in (Buenrostro et al., 2015). Each sample was tagmented using 12.5 ul Nextera TDE-1 transposase (Illumina) for 30 minutes at 37, then quenched by addition of 5 volumes DNA Binding Buffer (Zymo Research) and cleaned using Zymo Research DNA Clean and Concentrator-5 columns according to the supplied protocol. Tagmented DNA was PCR-amplified using indexed primers as described in (Buenrostro et al., 2015), using total cycle numbers for enrichment as determined empirically by qPCR to minimize PCR duplicates. The resulting libraries were purified twice by Zymo Research DNA Clean and Concentrator-5 columns using a ratio of 5:1 DNA Binding Buffer:Sample, and quantified by Qubit HS-DNA Assay (Thermo Fisher Scientific) and Bioanalyzer High-Sensitivity DNA (Agilent Technologies). Final ATAC-seq libraries were pooled (equimolar) and sequenced on an Illumina Nextseq 500.

ChIP-Seq

Harvest and Formaldehyde crosslinking. For each timepoint and donor, 5-7 x 10^6 unstimulated or LPS-stimulated hDCs were harvested by scraping in medium and centrifugation at 500 x g for 5 minutes. Each cell pellet was washed once with 2 mL PBS and gentle flicking of the tube, followed by centrifugation at 500 x g for 5 min. Cells were uniformly resuspended in 1 mL 1X Fixing Buffer A from the Covaris tru-ChIP Chromatin Shearing and Reagent Kit and fixed by adding 1 mL 2% methanol-free formaldehyde (Thermo Fisher Scientific) diluted in 1X Fixing Buffer A (1% formaldehyde final, 2.5-3.5x10^6 cells/mL) and rotated end-over end for 5 min. at room temperature. Fixation was quenched by adding 240 mL Quenching Buffer E (Covaris tru-ChIP kit) and rotating for an additional 5 min. Purified BSA was then added to 0.5% w/v final to prevent cell adherence to the tube, and crosslinked cells were harvested by centrifugation, 500 x g for 5 min. at 4°C. Crosslinked cells were washed twice in 2 mL ice-cold

PBS + 0.5% BSA with centrifugation as above, and aliquoted evenly into 3 fresh 1.5 mL tubes during the second wash. Cells were finally pelleted by centrifugation at 16,000 x g, flash-frozen as dry pellets in liquid nitrogen, and stored at -80°C.

Lysis, Shearing, and Quantification. Individual crosslinked cell pellets (1.5-2 x 10^6 cells each) were lysed according to the Covaris tru-ChIP Chromatin Shearing and Reagent Kit instructions. Following lysis, nuclei were resuspended in 130 mL ice-cold Shearing Buffer D3 and transferred to 1.5 mL BioRupter Pico Microtubes (Diagenode) on ice. Chromatin was sheared to uniform fragment lengths (150-400 bp) by sonication at 4°C in a BioRupter Pico (Diagenode) set to 6 cycles of 30s ON and 30s OFF. Sheared chromatin was diluted in 10 volumes of ChRIPA buffer (1X PBS, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.5% sodium deoxycholate, 1% Igepal CA-630, 0.1% SDS, 1X Roche cOmplete Protease Inhibitor Cocktail) and insoluble material was removed by centrifugation >15,000 x g for 10 minutes. Lysate was pre-cleared against 60 mL Dynabeads Protein A (Thermo Fisher Scientific) per 10^6 cells for 2h at 4°C with end-over-end rotation followed by two rounds of magnetic bead removal and transfer to fresh tubes. 2% of pre-cleared lysate was removed for DNA quantification and the remaining lysate was either flash-frozen in liquid nitrogen and stored at -80°C, or stored overnight at 4°C for use in immunoprecipitation. For quantification, 2% pre-cleared lysate was treated with 10 mg RNase A (Thermo Fisher Scientific) for 30 min. at 37°C, followed by addition of 100 mg Proteinase K (New England Biolabs) and crosslink reversal overnight at 65°C. DNA was purified using DNA Clean and Concentrator-5 columns (Zymo Research). Average sheared DNA fragment sizes were determined by agarose gel and chromatin yield was estimated by Qubit HS-DNA Assay. 50-100 ng purified DNA was saved as Input.

Chromatin Immunoprecipitation. Antibodies used for ChIP were rabbit anti-H3K27ac (Diagenode C15410196) and rabbit anti-H3K4me3 (EMD Millipore 05-745R). 1 mg antibody was added to 0.5 mg (anti-H3K27ac) or 1 mg (anti-H3K4me3) pre-cleared crosslinked lysate and incubated overnight with continuous mixing at 4°C. IgG/chromatin complexes were captured for 1h at room temperature on 25 mL Dynabeads Protein A that were pre-blocked for at least 1h with Blocking Buffer (1X PBS, 0.5% BSA, 0.5% Tween-20). Complexed beads were washed 5 times with ice-cold ChRIPA Buffer, twice with room temperature RIPA-500 Buffer (10 mM Tris pH 8.0, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS), twice with ice-cold LiCl Wash Buffer (10 mM Tris pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Igepal CA-630, 0.5% sodium deoxycholate), and twice with ice-cold TE buffer. Each chromatin sample was eluted from beads using 50 ul Direct Elution Buffer (10 mM Tris pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.5% SDS) and supplemented with 20 mg RNase A, incubating for 30 min. at 37°C. 20 mg glycogen was added to each bead/eluate suspension, and crosslinks were reversed by addition of 50 mg Proteinase K and incubation at 37°C for an additional 2h, followed by overnight at 65°C. Dynabeads were removed by magnet capture, and the supernatant was mixed thoroughly with 2.3 volumes of Agencourt AMPure XP (Beckman Coulter) bead suspension and incubated for 10 minutes at room temperature prior to bead capture and washing. Purified DNA was eluted in 10 mM Tris pH 8.0.

Library Preparation and Sequencing. Sequencing libraries were prepared from half of each ChIP sample and 50 ng Input DNA using the Ovation Ultralow System V2 kit (NuGEN) according to supplier's instructions, with the total numbers of enrichment PCR cycles determined empirically for each sample by qPCR to minimize PCR duplication rates. Barcoded libraries were quantified

using Qubit HS-DNA Assay, qualified using Agilent Bioanalyzer High-Sensitivity DNA, and pooled for sequencing on Illumina Nextseq 500.

Quantification and Statistical Analysis

Alignment and processing of reads

RNA-Seq: Trimmomatic-0.32 (Bolger et al., 2014) was used to remove 5' or 3' stretches of bases having an average quality of less than 20 in a window size of 10. Only reads longer than 36 bases were kept for further analysis. Reads were then aligned to human or mouse ribosomal RNA using Bowtie2 v2.2.3 (Langmead and Salzberg, 2012) with parameters -p 2 -N 1 --no-unal. All reads mapped to rRNA were discarded from further analysis. RSEM v1.2.28 (Li and Dewey, 2011) was used to estimate gene expression in Transcripts per Million (TPM), with parameters -p 4 --bowtie-e 70 --bowtie-chunkmbs 100 --strand-specific. RSEM is configured to use Bowtie v0.12.9. Quantification was run against the transcriptome (RefSeq v69 downloaded from UCSC Table Browser (Pruitt et al., 2012). Genes with more than 10 TPM in any time point were considered expressed, and genes that did not achieve this threshold were removed from further analysis. Moderate batch effects were observed between samples from different mice and between the two human donors. We used the log transformed TPM normalized expression values as input to ComBat (package sva version 3.18.0) (Johnson et al., 2007; Leek et al., 2012) with default parameters and a model that specified different donors or mice as batches. Corrected TPM values were transformed back to read counts using the expected size of each transcript informed by RSEM. We only considered genes with at least 10 TPMs in at least one replicate at any time point.

ATAC-Seq: Paired-end reads were trimmed to remove adapter sequence using Cutadapt version 1.3, and then aligned with Bowtie2, version 2.1.0, parameter –X 2000. Reference genome hg19 was used for human samples and mm10 for mouse samples. The alignments were then filtered using Samtools (Li et al., 2009), version 0.0.19, to remove (i) PCR duplicates, as identified by Picard's MarkDuplicates, and (ii) aligned reads with mapping quality below 4. While the reads were aligned as paired-end to optimize the alignment accuracy, the alignments were then further processed as if they were aligned single-end sequence data, so that each aligned read corresponded to a Tn5 cut-site.
Peak Calling: Each aligned read was first trimmed to the 9-bases at the 5'-end, the region where the Tn5 transposase cuts the DNA, and then extended 10-bases upstream and down, for smoothing. Peaks were called using these adjusted 29-base aligned reads with MACS2 (Zhang et al., 2008), parameters --bw 29 --tsize 29 and --qvalue 0.0001. For visualization, the adjusted aligned reads were converted to tdf files using IGVTools, version 2.3.31 (Robinson et al., 2011) (IGVtools count –w 5).
Quality Control: Following the standard practice (Buenrostro et al., 2015), for each sample, we examined the fragment length distribution, as well as a comparison of the aggregate nucleosome signal to the aggregate nucleosome-free signal over transcription start sites for those genes found to be expressed for at least one time point in our RNA-Seq time series. Signal-to-noise ratios were computed for the peaks as f/(1 –f) where f is the fraction of reads overlapping peaks.

ChIP-Seq: Along with in house generated data we also analyzed publicly available data for mouse bone-marrow progenitors generated by the Encode consortium (Accession: GSM1000108). Paired-end reads were trimmed to remove sequencing adapters and leading and trailing bases with quality scores less than 5. Reads that were longer than 36 bases after trimming were kept for further analysis. The reads were then aligned to human reference genome hg19 or mouse genome mm10 using Bowtie2 with options -k 1 --un-conc to filter out reads that map to multiple locations in the genome and that align un-concordantly. Duplicated reads were filtered out using picard-tools-1.131 MarkDuplicates function. Peaks were then called using MACS2 with --bw=230 --tsize=75 and --qvalue 0.0001. Alignment files were also converted to tdf format using IGVtools count function using -w 5 --pairs options for visualizing. H3K27ac ChIP-Seq peaks were filtered to retain only the peaks that are two-fold enriched over input.

Gene classification and clustering

Homologs: All our analysis were restricted to genes that had homologous pairs between human and mouse defined in the Homologene release 68 (NCBI Resource Coordinators, 2016), resulting in a list of 16,500 one to one homologous gene pairs.

Gene Classification: The expressed gene list was filtered to include only genes with homologs as defined by the previous step. We used the batch corrected (see above) counts per gene to identify differentially expressed genes by at least 2 fold between unstimulated cells (time 0) and any time point following LPS stimulation whose change in expression was significant (p-adjusted < 0.05) according to the package DESeq2 (v1.10.1) (Love et al., 2014) in R (v3.3.1). Due to the large transcriptional changes observed in this system, we turned off the fold change shrinkage in DESeq2 with betaPrior=FALSE and we added a pseudocount of 32 to all timepoints to avoid spurious large fold change estimates from lowly abundant genes. Genes were then classified based on their response to LPS stimulation in each species (induced, downregulated or non-responsive).

Clustering expression patterns: For genes expressed in both species and presenting similar response following LPS stimulation (induced in both species or downregulated in both), we applied a spectral clustering approach (von Luxburg, 2007) to identify genes with conserved expression patterns in mouse and human. Briefly, let {g1,g2,g3,…,gn} represent the set of response genes, and let E_{Mi} and E_{Hi},$1 < i < n$, represent the expression time courses in TPM for gene g_i in mouse and human respectively. Further, let $\rho_M = [\rho_{Mij}]$, $1< i,j< n$ represent the Pearson correlation coefficient matrix, where ρ_{Mij} is the coefficient of correlation of E_{Mi} with E_{Mj}. The human correlation coefficient matrix, ρ_H is defined similarly. We define similarity matrices $[s_{Mij}]$ and $[s_{Hij}]$, for mouse and human respectively, where $s_{Mij} = \exp(-(\sin(\cos^{-1}(\rho_{Mij})/2)^2)$, and $s_{Hij} = \exp(-(\sin(\cos^{-1}(\rho_{Hij})/2)^2)$. Then the matrix $W = [w_{ij}] = [s_{Mij}s_{Hij}]$ defines a similarity matrix for {g1, g2, … , gn} and can be viewed as an adjacency matrix for a weighted graph, where each gene represents a node in the graph. We associate to W its graph Laplacian L = D - W, where D is the diagonal degree matrix with entries $dii = \mathbf{\Sigma}^n_{j=1} w_{ij}$. L is positive, semi-definite and therefore has n real non-negative eigenvalues, λ_i, $1< i < n$, which we list in descending order, $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. We select k, the number of clusters, to be the smallest positive integer such that ($\lambda_1 + \lambda_2 + \ldots$

+ λ_k)/tr(L) > 0.95, where tr(L) is the trace of L. We then construct a matrix with columns set to the first k eigenvalues of L and apply k-means clustering to the rows of this matrix to cluster the genes into k distinct clusters. The python script used for spectral clustering is available on https://github.com/nimezhu/ClsViz.
We analyzed enrichments for specific Gene Ontology categories using *clusterProfiler* (Yu et al., 2012).

Transcription Factor network

We sought to first determine the extent to which the TF network in response to LPS is conserved between human and mouse DCs. To systematically explore core changes in the regulatory network, we compared the overall trends of the 258 transcription factors that responded to LPS-stimulation in at least one of the two species (Figure S1D). We calculated the Pearson correlation between the expression patterns across all timepoints for TFs with response to LPS per species. The resulting distance matrix was hierarchically clustered and displayed as a heat map. We chose the number of groups in each clustering by visual inspection of the dendrogram and selection of a threshold. Membership in each cluster was then compared across species to identify the corresponding groups.

Transcription Factor Network Overview: There are 3 large co-regulated groups of transcription factors with no major changes between the species, and a fourth cluster in mouse composed of only 8 TFs (Table S1) with very small changes in expression in mouse (< 2 fold), that are scattered across all three human clusters. The largest cluster in mouse contained 115 genes that were downregulated following LPS treatment. Further, 73% of the factors that were also expressed in human remained in the same cluster and showed a similar transcriptional downregulation pattern in human (Figure S1D, top right). Similarly, the vast majority (77%) of induced transcription factors were induced in both species, with 17 factors (19%) having different induction timing in each species (Table S1). The largest of the induced clusters (pink cluster, Figure S1D), contained mostly TFs with conserved kinetics (66% in mouse and 57% in human, Figure S1D, bottom right). This group included members of the NFKB, IRF, and STAT families (Figure 1C). The smaller cluster of induced transcription factors also contained important rapidly upregulated TFs (blue cluster, Figure S1D, middle right), including members of the FOS and JUN families, as well as MAFF, PRDM1, and EGR3, all of which show a conserved pattern in the human response. 17 mouse-specific and 12 human-specific TFs were induced by LPS. Interestingly, to the best of our knowledge, none of the species-specific factors have been studied in the context of innate immune signaling. Two mouse-specific TFs, ID1 and SIX1, are highly induced in mouse, although not detectable in human. Similarly, MSC is highly induced in human DCs but has no detectable expression in mouse DCs. Outliers such as these however, are rare, and most TFs with different responses in mouse and human DCs have moderate induction compared to genes with conserved response.

Substitution rate scan
We used SiPhy (Garber et al., 2009) to compute the substitution rate (ω) for every 10-mer in the mouse and human genomes. For human we used the vertebrate multiple sequence alignment

available from the UCSC genome browser for the hg19 assembly. We removed the vertebrates danRer6, petMar1, oryLat2, gasAcu1, fr2, tetNig2 which left us with the following phylogeny:

```
((((((((((((((((hg19:0.006653,panTro2:0.006688):0.002482,gorGor1:0.008783):0.009697,p
onAbe2:0.018183):0.040003,rheMac2:0.008812):0.002489,papHam1:0.008723):0.045139,calJac
1:0.066437):0.057049,tarSyr1:0.137822):0.010992,(micMur1:0.092888,otoGar1:0.1295):0.03
5423):0.015348,tupBel1:0.186424):0.004886,(((((mm9:0.084505,rn4:0.091627):0.197835,dip
Ord1:0.211666):0.022945,cavPor3:0.225634):0.010077,speTri1:0.148511):0.025643,(oryCun2
:0.114421,ochPri2:0.201003):0.101624):0.015291):0.020683,(((vicPac1:0.107267,(turTru1:
0.064676,bosTau4:0.123573):0.025145):0.040411,((equCab2:0.109311,(felCat3:0.098636,can
Fam2:0.102486):0.049838):0.006202,(myoLuc1:0.14262,pteVam1:0.113246):0.033792):0.00445
6):0.011576,(eriEur1:0.221758,sorAra1:0.269694):0.056557):0.021228):0.023628,(((loxAfr
3:0.082165,proCap1:0.155353):0.026774,echTel1:0.246266):0.049887,(dasNov2:0.116609,cho
Hof1:0.096318):0.053052):0.006229):0.399651,macEug1:0.133617):0.002474,monDom5:0.15092
1):0.199105,ornAna1:0.461732):0.116917,((galGal3:0.164668,taeGut1:0.172833):0.200238,a
noCar1:0.48763):0.10284):0.186338,xenTro2:0.834181):0.324842
```

Spanning 8.44 substitutions per site. We excluded 10-mers that after removing species with no alignable sequence due to either alignment gaps or missing sequence had a total branch length of less than 0.75. Data is available from

http://garberlab.umassmed.edu/data/conservation/hg19/omega/

For mouse we used the vertebrate multiple sequence alignment available from the UCSC genome browser for the mm10 assembly. We removed petMar1, gadMor1, oryLat2, gasAcu1, oreNil2, fr3, tetNig2, latCha1, xenTro3, chrPic1, anoCar2, melUnd1, taeGut1, melGal1, ornAna1, macEug2, sarHar1 vertebrate assemblies which left us with the following phylogeny:

```
((((((((((mm10:0.0861604,rn5:0.0923189):0.20235,dipOrd1:0.210872):0.0258938,(hetGla2:
0.0916296,cavPor3:0.136929):0.0994423):0.00913482,speTri2:0.145406):0.0275377,(oryCun2
:0.10975,ochPri2:0.200956):0.102105):0.0142197,(((((((((hg19:0.00672748,panTro4:0.0069
0586):0.00329132,gorGor3:0.00918574):0.00952813,ponAbe2:0.019182):0.00354391,nomLeu2:0
.0218123):0.0117068,(rheMac3:0.00815625,papHam1:0.00799922):0.0289552):0.0208613,(calJ
ac3:0.0342486,saiBol1:0.0333278):0.0358206):0.0593959,tarSyr1:0.137561):0.0111487,(mic
Mur1:0.0919295,otoGar3:0.127188):0.0351183):0.0153325,tupBel1:0.188903):0.0042042):0.0
215023,((susScr3:0.121671,(vicPac1:0.10979,(turTru2:0.0635601,(oviAri1:0.0392014,bosTa
u7:0.0315737):0.0939007):0.0204197):0.00365643):0.0444426,((((felCat5:0.0897916,(canFa
m3:0.0888559,ailMel1:0.0767967):0.0218058):0.050101,equCab2:0.109329):0.00604713,(myoL
uc2:0.137323,pteVam1:0.113957):0.0339856):0.00384687,(eriEur1:0.227177,sorAra1:0.27056
4):0.0629454):0.00322051):0.0291201):0.0231348,((((loxAfr3:0.0788116,proCap1:0.160315)
:0.00818092,echTel1:0.266806):0.00328658,triMan1:0.068537):0.0736006,(dasNov3:0.112113
,choHof1:0.0974595):0.0536232):0.00734155):0.246266,monDom5:0.35412299999999997):0.212
5305,galGal4:0.5622546999999999):0.6482475,danRer7:0.871611):0.49907
```

Spanning 8.21 substitutions per site. We excluded 10-mers that after removing species with no alignable sequence due to either alignment gaps or missing sequence had a total branch length of less than 0.5. Data is available from

http://garberlab.umassmed.edu/data/conservation/mm10/mm10.omega

The models used were downloaded directly from UCSC and correspond to the alignments used.

Enhancer and promoter definition and conservation analysis

Enhancers and promoters were defined by H3K27ac peaks. We then merged all peaks from each time point located within 200bp from each other. Our maps consist of 28,142 and 29,273

H3K27ac regions (signal peaks) in mouse and human, respectively. We calculated the distance from each peak to the nearest transcription start site (TSS) of the highest expressed isoform for each gene using bedtools closest -D ref -t all (Quinlan and Hall, 2010). We classified all H3K27ac peaks that had a distance smaller than 500 bp to the nearest TSS as promoters, and the remaining peaks were considered enhancers. Enhancers were assigned to the nearest gene based on the same TSS distances as above. Unlike promoters, which were associated to the gene with the overlapping TSS independent of expression, enhancers were only associated to the closest expressed gene within 300 kb (Garber et al., 2012; González et al., 2015). This assignment of enhancers to nearby genes will misassign enhancers that either interact with more than one gene or interact with no adjacent genes. However, the majority of enhancers have been reported to interact with the neighboring gene (González et al., 2015). Overall, 2/3 of the peaks were annotated as enhancers in each species, consistent with previous studies (Villar et al., 2015). We filtered ATAC peaks to include only peaks that overlapped with a H3K27ac region. We classified ATAC peaks as enhancers or promoters based on the H3K27ac peak definition, and maintained the association to genes defined for H3K27ac peaks. To determine the conservation of mouse enhancer and promoters in human, peaks were mapped to the human genome corresponding locations using liftOver -minMatch=0.1 -multiple (Hinrichs et al., 2006). We filtered out peaks that mapped to more than 3 locations and used the remaining peak locations to intersect with the human enhancer and promoter coordinates to determine if that region was also active in the human dendritic cells. To generate aggregation plots of the H3K27ac and ATAC-Seq signal, we used the center position of ATAC peaks for enhancers and the TSS for the genes associated to the peaks as coordinates for input to ngs.plot (Shen et al., 2014). The coverage was calculated for a 4kb region surrounding the center position (-L 2000). We selected the regions corresponding to each group of interest from the output matrix and calculated the mean signal per group.

ATAC and H3K27ac dynamics

The mean signal across each ATAC-seq or H3K27ac peak was calculated by averaging the number of reads per base pair. The average signal across the libraries are normalized to the depth of each library using DESeq2 (v1.10.1) in R (v3.3.1). ATAC-seq or H3K27ac peaks were considered dynamic in response to LPS if they have greater than two fold-change in their mean signal compared to unstimulated state. The dynamic ATAC-seq or H3K27ac peaks identified are clustered using k-means algorithm to identify groups of ATAC-seq H3K27ac peaks that are induced or repressed following LPS stimulation.

Motif analysis

Motif analysis was done on 200 bp regions around the summits of the ATAC-seq peaks. The log-odds substitution rate for each 10 base-pair window across the summits of ECAs and ESPAs ATAC-seq peaks was calculated using SiPhy (Garber et al., 2009). The value of log-odd substitution score at the top ten percentile of a given peak was assigned as the conservation score for each peak. The kmers that intersected the ATAC-seq summits and which had log-odds score greater than 30 were considered for building cPWMs. To get a background set, we shuffled these 200bp ATAC-seq peaks within the enclosing H3K27ac peaks and considered all the kmers with

log-odds score greater than 30. To identify kmers that distinguish the conserved ATAC peaks from background, we used the string kernel built-in gkm-svm R package (Ghandi et al., 2016) with 5 fold cross validation which resulted in 4500 unique kmers as features for conserved ATAC peaks. These kmers were clustered into 66 PWMs using k-medoids clustering algorithm with Euclidean distance, within the clara function in the cluster package in R (Blashfield, 1991). The cPWMs were then matched to the known motifs from CIS-BP database (Weirauch et al., 2014) using Tomtom (Gupta et al., 2007). Multiple motifs matched to the same TF are identified by numbers. For example JUN-1 and JUN-2. To find the cPWMs enriched in temporal gene groups or temporal ATAC peaks we used the Fisher exact test and all cPWMs with p value < 0.05 were considered enriched.

All cPWMs identified are available from

http://garberlab.umassmed.edu/publications/conserved_lexicon_Dec_2017/cPWMs.motifcPWMs.motif

Transposable element analysis

We used the transposable element annotation by RepeatMasker (Smit et al., 2004) to identify TE instances in each genome that overlapped at least 10% with the regulatory regions (enhancers and promoters) associated to induced genes. As a background, we shuffled these cis-regulatory regions in the genome inside boundaries defined by the regulatory regions associated to expressed genes with no response to LPS, expanded by 10kb in each direction. We then identified the number of instances for each TE family that overlapped at least 10% with these shuffled peaks. We performed this shuffling process 1000 times and compared the initial counts obtained for each TE family to this null distribution. We computed a p-value for this permutation and corrected it using the Benjamini Hochberg method. All TE families with adjusted p-value under 0.05 were considered to be overrepresented in the regulatory regions of induced genes. For each instance of these elements in induced genes, we identified the corresponding region in the other species' genome through liftOver as described above. We then evaluated if the region that can be identified in the other genome also overlaps a H3K27ac peak, classifying it as an ECA. H3K27ac and ATAC-Seq signal aggregation plots were generated as described above, with the TE start and end genomic coordinates as the target region, flanked by 1kb on each side.

Predictive model of gene induction from cPWM instances

Feature selection: For the selected set of 66 cPWMs, all instances were detected across all ATAC peaks (promoters and enhancers) using fimo (Grant et al., 2011), with a q-value threshold of 1e-4. We tested the models using two representations of the cPWMs as features: 1. All cPWM instances together - For each gene and each cPWM, we counted the number of instances across all regulatory elements of the gene. 2. All cPWM instances, separated by ATAC temporal pattern - each cPWM was separated to three features - the number of instances in LPS-induced regions (based on ATAC-seq data), number of instances in repressed regions and number of instances in unchanging regions.

Gene filtering: To build an informative model and to reduce noise from lowly expressed genes, we focused on highly expressed genes by taking only genes that were in the top 30% of expressed genes in at least one time point. Furthermore, to clearly distinguish induced from non-induced genes, we classified genes with a log2 fold change > 2 as induced, and genes with a log2 fold change between -0.3 and 0.3 as not induced, and discarded all the rest. Next, to create a balanced set of induced and non-induced genes, we downsampled the number of non-induced genes. This resulted in a total of 676 genes (338 induced and 338 non-induced) in mouse and 748 genes in human.

Model evaluation: All model training and evaluations was done in R, using the caret (v6.0.77) (Kuhn et al.) and randomForest (v4.6.12) (Liaw et al., 2002) packages. For each feature set, we evaluated the accuracy of the model on the mouse data with 10-fold cross validation. For each one of the training data in the cross validation, hyperparameters tuning was performed using 10-fold inner cross validation with the "train" command, using the following parameters: tuneLength = 20, metric = "ROC". To evaluate how well the model predicts induction on the human data, we trained a model on the full mouse data (again using 10-fold cross validation for hyperparameters selection) and applied the selected model on the human data.

Feature Importance: Importance measurement for each feature was computed with the "varImp" command, defined as the difference in mean accuracy across all trees between the model and the model after permuting the feature. The importance values were then scaled to span the range of 0 to 100.

Data and Software Availability

All samples generated for this work were submitted to NCBI as part of the Genomics of Gene Regulation Project, under accession number PRJNA356880. A list of samples used is specified on Table S4.

Supplementary Figures

Figure S1

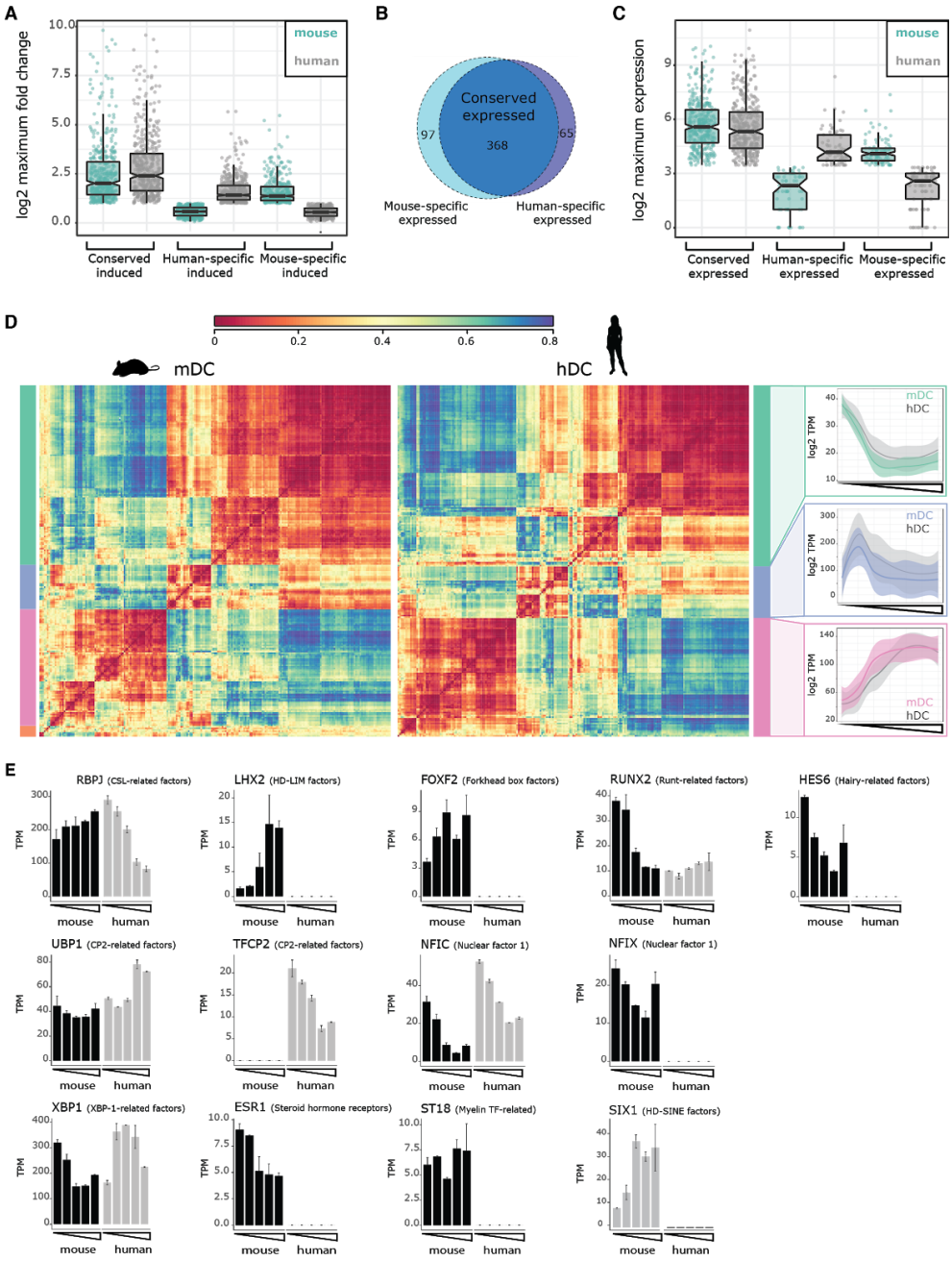


Figure S1 - Related to Figure 1 and STAR Methods A) Box plots displaying log2 maximum fold change per gene for shared and species-specific induced genes post-LPS stimulation. B) Venn diagram of TFs expressed in each species. C) Box plots displaying log2 maximum expression (TPM) for shared and species-specific expressed TFs D) Heatmap showing hierarchical clustering of the correlation of expression across time for all LPS-responsive TFs. Left: mouse factors (n=228); Center: human factors (n=224); Right: average expression of the factors in each cluster that show a shared pattern between species. E) Expression patterns of TFs

that belong to families with only species-specific response to LPS. Error bars show standard deviation from the mean

Figure S2

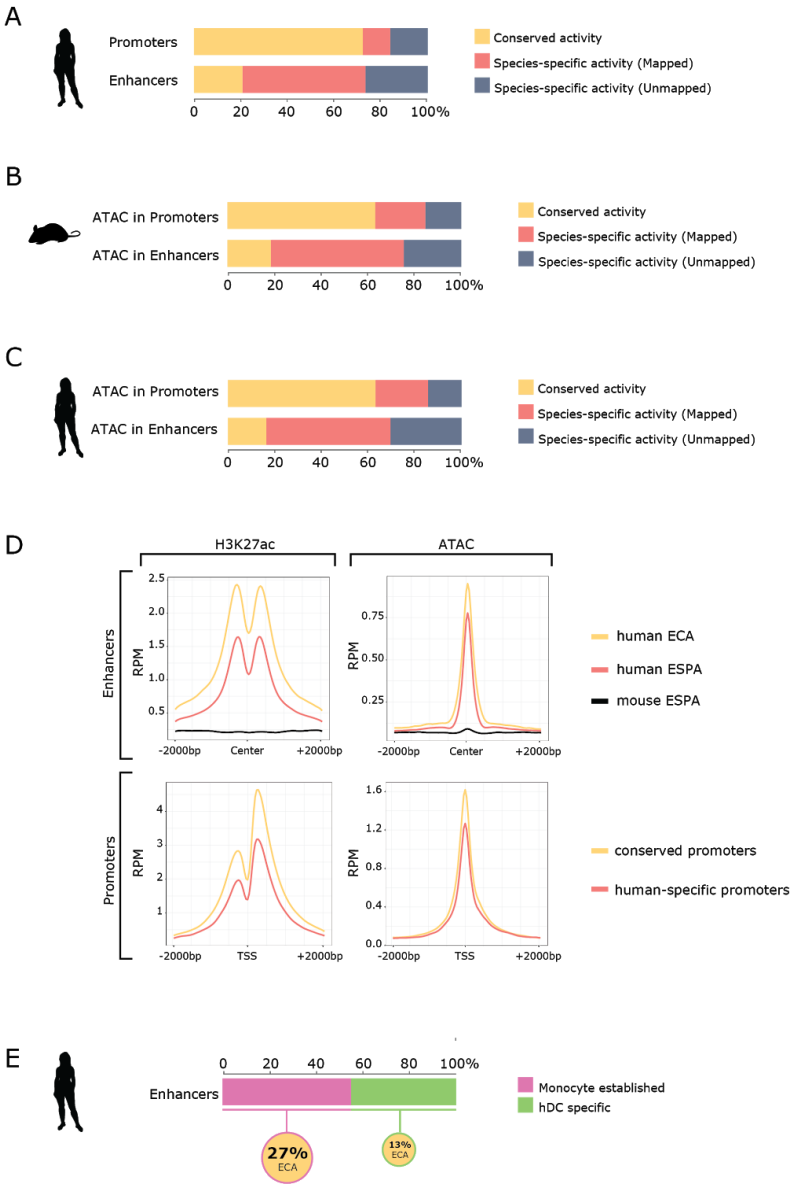


Figure S2 - Related to Figure 2 A) Overall conservation of promoter and enhancer regions in human DCs. B) Overall conservation of ATAC peaks in promoter and enhancer regions in mouse DCs. C) Overall conservation of ATAC peaks in promoter and enhancer regions in human DCs. D) Average signal aggregation plots for human H3K27ac (left) and ATAC-Seq (right) signal over regulatory elements. Enhancer (top) H3K27ac signal is centered in open regions, defined by ATAC-Seq peaks. Promoter (bottom) H3K27ac is centered in the TSS. ATAC-Seq signal for both enhancers and promoters is centered in open regions. Data is shown for conserved regulatory regions (promoters or ECAs, yellow), human specific regulatory

regions (promoters and ESPA, red) and all other human genome coordinates for mapped mouse-specific promoters or ESPAs. E) Fraction of human enhancers that are already active (pre-established) in monocytes (MONO) and enhancers that are hDC specific, and fraction of MONO pre-established or hDC specific enhancers that are conserved (ECA).

Figure S3

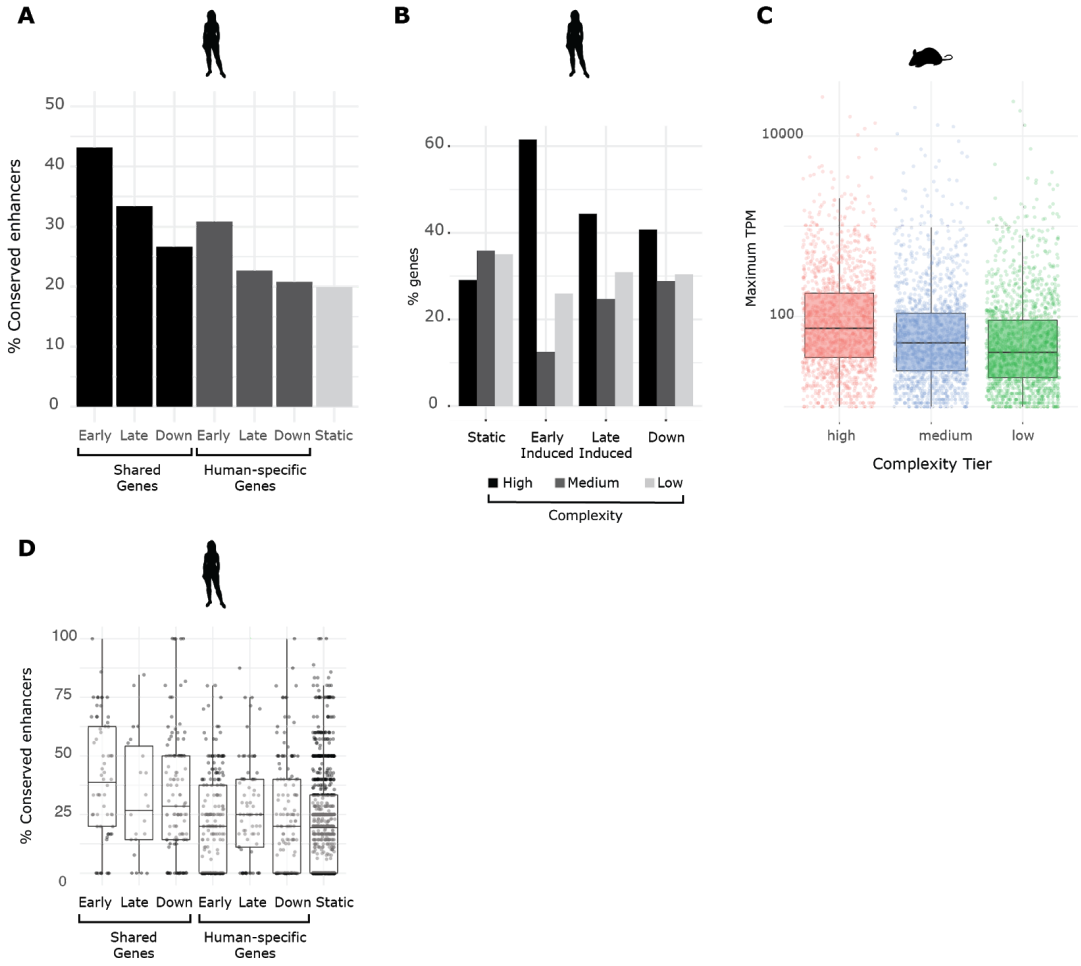


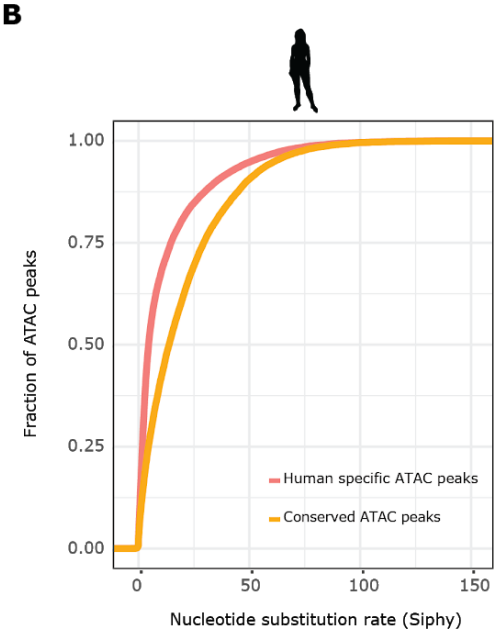
Figure S3 - Related to Figure 3 A) Fraction of enhancers that are ECAs associated to genes that have shared or species-specific response: early-induced, late-induced or downregulated upon stimulation with LPS in human DCs. B) Fraction of genes in temporal gene clusters of human DCs that are associated to high-, medium- or low-complexity enhancer loci C) Maximum expression, measured in transcripts per million (TPM) for genes in each complexity tier D) Fraction of enhancers that are ECAs in high complexity shared or species-specific response genes which are early-induced, late-induced, downregulated or have no change in response to LPS in human DCs.

Figure S4

A

	Number of peaks in ATAC	Total peaks in TF ChIP	Percent peaks in ATAC
Atf3	1497	1618	93%
Cebpb	23690	27348	87%
E2f1	2703	3690	73%
E2f4	925	1258	74%
Irf1	15375	17033	90%
Irf4	6222	6530	95%
JunB	7648	7843	98%
PU1	50576	64833	78%
Rela	19301	20966	92%
RelB	1475	1813	81%
Runx1	4098	4501	91%
Stat1	6085	6419	95%
Stat2	640	799	80%

B



C

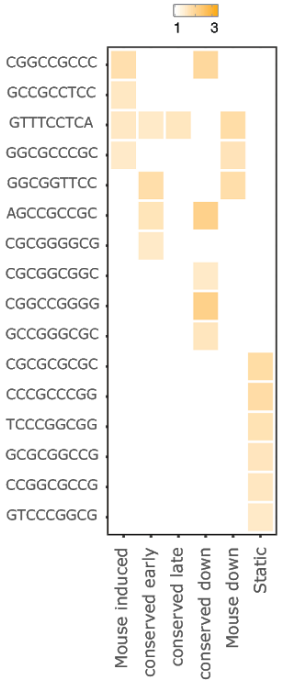


Figure S4 - Related to Figure 4 A) Table showing the number (column 1) and fraction (column 3) of TF ChIP peaks that are in ATAC-seq peaks B) Distribution of SiPhy omega log-odds scores in ATAC-seq peaks with conserved signal (yellow) and species-specific signal (red) for human DCs. C) Enrichment of cPWMs that are novel in the gene clusters.

Figure S5

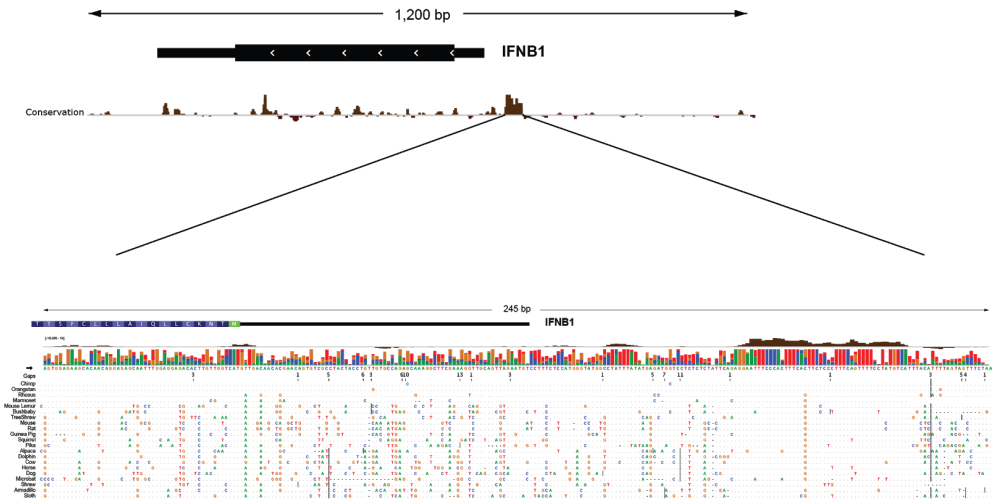


Figure S5 - Related to Figure 5. IFNβ locus showing the multiple sequence alignment of the IFNβ enhanceosome and IFNβ gene. The top half of the figure shows the locus and conservation score, bottom half shows the multiple sequence alignment of the IFNβ enhanceosome. Dots are the nucleotides that haven't changed from the mouse sequence.

Figure S6

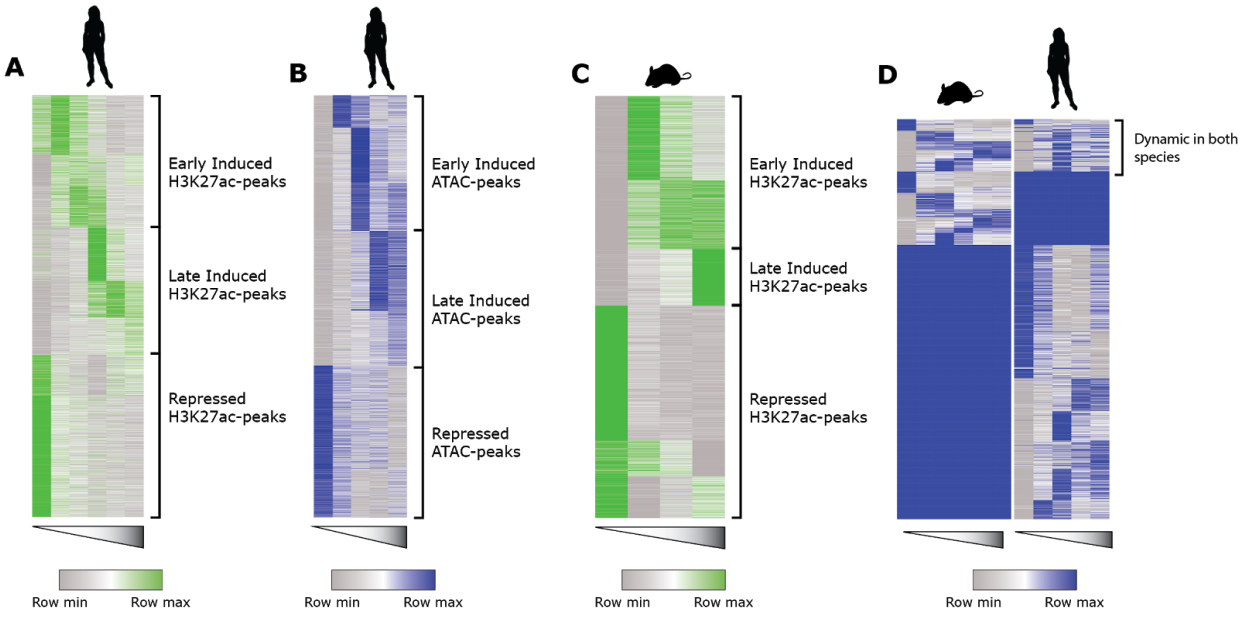


Figure S6 - Related to Figure 6. A) Heatmap showing the temporal patterns of H3K27ac peaks in response to LPS in human DCs (Unstimulated, 30 minutes, 1 hour, 2 hours, 4 hours and 6 hours) which are annotated as promoters or enhancers. B) Heatmap showing the temporal patterns of ATAC-seq signal associated with regions annotated as promoters or enhancers in human DCs (Unstimulated, 30 minutes, 2 hours, 4 hours and 6 hours) C) Heatmap showing the

temporal patterns of H3K27ac peaks in response to LPS in mouse DCs (Unstimulated, 30 minutes, 1 hour and 2 hours) D) Temporal patterns of ATAC-seq peaks that are dynamic in at least one of the species when stimulated with LPS.

Figure S7

A

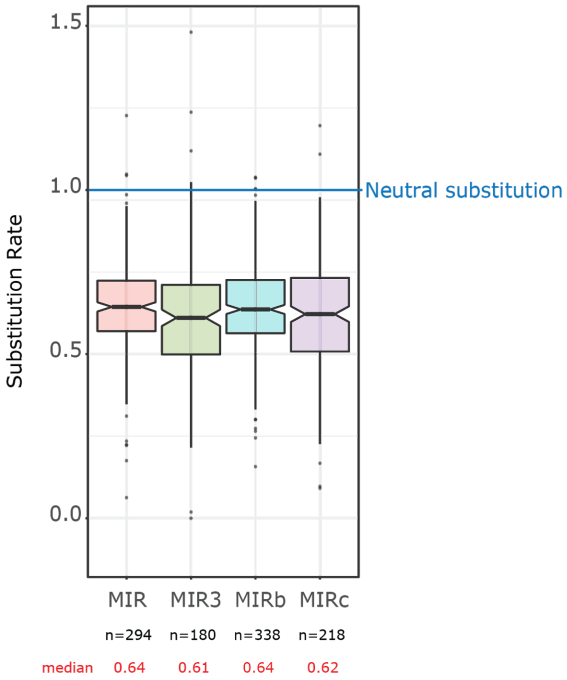


Figure S7 - Related to Figure 7. A) Distribution of the nucleotide substitution rates across 41 mammals for TEs from the MIR element families that overlap with regulatory regions of induced genes. One value for the substitution rate per element instance is shown, which corresponds to the value at the 90th percentile.

Chapter 3 - Uncovering the DNA sequence motifs which control the epigenetic landscape of Dendritic cells maturation

In this chapter, I devised a supervised learning pipeline to determine which short DNA sequence motifs may be functional within a given subset of regulatory regions. I apply this pipeline to study temporal activation patterns of regulatory regions in DCs up to 24 hours after LPS stimulation. This work resulted in a comprehensive map of TF binding motifs that are active at various times during DC maturation. This includes several factors that were previously unknown to be a part of the DC response, and our collaborators are currently in the process of validating them experimentally. This chapter is a draft of the manuscript that will be completed after we conduct the validation experiments. The tentative list of authors includes:

Shaked Afik[1,7], Pranitha Vangala[1,7], Elisa Donnard[2], Sean McCauley[3], Anetta Nowosielska[3], Alper Kucukural[3,4], Barbara Tabak[2], Patrick McDonel[2,3], Jeremy Luban[3], Manuel Garber[2,3,4], Nir Yosef[1,5,6]

1. Center for Computational Biology, University of California, Berkeley, Berkeley, CA-94720, USA
2. Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA-01605, USA
3. Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA-01605, USA
4. Bioinformatics Core, University of Massachusetts Medical School, Worcester, MA-01605, USA
5. Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA-94720, USA
6. Chan Zuckerberg Biohub, San Francisco, CA-94158, USA
7. These authors contributed equally

Abstract

Epigenetic changes are a crucial step in the cellular response to environmental stimuli, and involve interactions between chromatin, non-coding DNA regions, histone modifiers and transcription factors (TF). To date, most methods that link chromatin accessibility and TF binding only provide genome-wide TF binding prediction without functional context, or focus on gene expression prediction. Here, we present a generalized framework to detect which DNA motifs are associated with any given process in the cell, allowing a functional interpretation of TF binding that extends not only to transcriptional regulation. To test our method, we applied it to study epigenetic changes in human Dendritic cells stimulated with lipopolysaccharide (LPS). This resulted in a comprehensive map of TF binding motifs which are functional in several temporal activation patterns of regulatory regions up to 24 hours after LPS stimulation. Our results include known regulators of the LPS response, as well as TFs which interact with histone acetyltransferases and deacetylases that were previously unknown to be involved in Dendritic cells' maturation. Moreover, our computational method is modular, generalizable and can be easily applied to study many other biological systems.

Introduction

Changes to cell state involve activation and repression of many genes. These changes are often mediated by changes to chromatin accessibility and histone modifications at regulatory DNA regions that facilitate the binding of transcription factor proteins to short sequence motifs. Despite many advances in characterizing the epigenetic landscape of cells, uncovering the way all these factors interact to activate a specific cellular process remains a challenging task.

A common way to detect binding of a given transcription factor (TF) in regulatory regions is by ChIP-seq. However, this is a laborious process that is limited to one TF per experiment. Methods to evaluate genome-wide chromatin accessibility such as ATAC-seq (Buenrostro et al., 2013) provide a genome-wide view of the accessible regions, which in turn opens the way for computationally predicting where any given TF binds. To this end, computational analysis of ATAC-seq data can reveal accessible genomic regions by detecting regions that include many of the aligned ATAC-seq reads (also known as peak regions). Further analysis of minor changes to accessibility within peak regions can reveal short DNA motifs bound by a TF since the binding sites will be protected from enzymatic cleavage. Combining these read alignment patterns - also known as genomic footprints - with previous knowledge of the TF binding sites provides simultaneous predictions for many TFs bound across the genome.

Recent computational pipelines attempt to infer genomic locations bound by specific TFs from genome-wide chromatin accessibility data (Gusmao et al., 2016; Xu et al., 2018). These methods vary in their algorithmic approach as well as the features used for prediction, but can be broadly divided into two categories: (1) motif-centric algorithms, which for a given set of TF motif instances in the genome will output a per-site binding prediction (Pique-Regi et al., 2011; Quach and Furey, 2017), or (2) algorithms which provide a binding prediction for the complete genome either with no DNA motif information (Li et al., 2019) or with the motif information as one of the features used for prediction (Keilwagen et al., 2019). The main focus of those methods has been in providing per-site prediction across the genome for a TF. Recently, several methods were developed to predict differential TF binding (Baek et al., 2017; Li et al., 2019; Tripodi et al., 2018), however, they have been tested on different cell types or under different experimental conditions, which usually includes many changes to the genomic landscape.

In this work, we present a computational pipeline that extends the scope of genomic footprint algorithms to go beyond genome-wide prediction of TF binding. Instead of trying to predict which sites are bound by a TF, our goal is to detect which motifs are functional within a set of genomic regions associated with a specific process. Such an approach can be applied to study which DNA motifs are regulators of transcriptional changes to nearby genes (Natarajan et al. 2012; González, Setty, and Leslie 2015; Schmidt et al. 2017; Donnard et al. 2018), but more generally be used to study any cellular process such as chemical modifications of histones.

We define this functional prediction problem as a supervised learning problem which attempts to discern between the chromatin state of motif instances in genomic regions of the process of

interest and a negative set of open chromatin regions. This generalized approach provides functional context to changes in DNA motifs and detects which motifs are drivers for specific processes in the cell. Moreover, our pipeline does not assume prior knowledge about the structure of a functional chromatin state (i.e. no prior assumption of a reduction in cut sites in the binding site compared to the flanking regions), which is important as some TF do not exhibit a strong genomic footprint (Sung et al., 2014). Thus, our method can be easily extended to study motifs which do not necessarily act as a TF binding site.

We applied our method to uncover the factors driving changes to putative active regulatory elements of human Dendritic cells in response to lipopolysaccharide (LPS). This response involves various temporal transcriptional and epigenetic changes to thousands of genes and regulatory regions in both humans and mice (Amit et al., 2009; Donnard et al., 2018; Garber et al., 2012; Rabani et al., 2014; Vandenbon et al., 2018). Our method discovers various DNA sequences that are predictive of epigenetic changes in the hours following LPS stimulation, including many binding motifs of TFs that were known to interact with histone modifiers but were not previously described as part of the Dendritic cells' response to LPS.

Results

Supervised learning approach to detect functional motifs

We devised a motif-centric computational pipeline to detect which short sequence motifs are functional in a subset of genomic regions. For example, given a set of regulatory regions that are involved in the cellular response to stimulation or state change, we wish to detect the short sequence motifs within those regions that function as binding sites for transcription factors. It is important to note that the strategy we built can be readily applied to any biological system where there is a state change and it is possible to define a positive and negative set of regions based on your question of interest. A summary of the pipeline is provided below and in Figure 1a, a full detailed description of the pipeline can be found in the methods section.

First, we start with a complete set of TF binding motifs for TFs of interest within the accessible regions defined by ATAC-seq. Each ATAC-seq peak is assigned to a positive or negative class based on the underlying question. For example, the label can be positive if this peak is a putative regulatory region in a specific cellular response. Next, for each TF binding motif, we extract the local chromatin features for each motif instance based on ATAC-seq cut sites 128bp upstream and downstream of the motif (Methods). To optimize performance we compute the cut sites only from nucleosome-free fragments (Li et al., 2019) and correct the cut sites count to account for enzymatic sequence bias (Martins et al., 2018). Instead of using the number of corrected cut sites in each base around the motif, our features are ratios between the sum cut sites of segments around the motifs at various lengths, similar to the transformation performed by msCentipede (Raj et al., 2015) (Figure 1b, methods). With this transformation, we capture the spatial structure of the chromatin, without limiting the algorithm to a predefined shape. These features are then used as the input for a random forest classifier, where a motif instance is labeled as part of the positive or negative set based on the label of the ATAC-seq peak in which it is found.

A high area under the precision-recall curve (AUC PR) value indicates that the genomic footprint in the positive set is distinguishable from the chromatin features around instances in the negative set. This provides an association between this motif and the specific set of active regions. We then run this pipeline for all motifs to get a complete evaluation of the regulatory motifs which are predictive of the positive regions. A natural interpretation for a high AUC value is that changes in chromatin shape correspond to differential TF binding. However, we note that there could be other interpretations such as changes in co-binding which result in different chromatin features.

Detecting functional motifs for TF binding

Our pipeline is designed to detect (for each motif) which of its motif instances are associated with a property of interest, out of all motif instances that fall in open chromatin regions. To validate the generality of our approach, we tested the ability of our computational framework to detect one of the more well studied properties. Namely, which motifs instances within open chromatin regions are bound by transcription factors? Of note, there is a large body of work dedicated to this problem. Comparing our method to these published methods provides us with a way to validate and benchmark our computational approach before moving to other, less well studied properties.

To that end, we ran our method on the publicly available chromatin accessibility data (Buenrostro et al., 2013) from the GM12878 cell line. We took a set of 66 TF binding motifs, for which there exists TF ChIP-seq from the ENCODE project (ENCODE Project Consortium, 2012) (Supplementary Table 1). For each motif, our positive set was the set of motif instances within open regions that overlap a TF ChIP-seq peak, while our negative set was defined as the motif instances within open regions which do not overlap a TF ChIP-seq peak. We then computed the mean AUC PR from 5-fold cross-validation runs. Limiting our analysis to only motif instances in open chromatin regions can be challenging, as motifs from the negative set are more prone to spurious binding compared to a randomly chosen negative set of motif instances across the genome. Despite this challenge, we are able to achieve overall high classification rates when applying a random forest classifier to the transformed cut sites (Figure 1c). We are also able to achieve high classification rate, albeit slightly lower on average, when taking into account all fragment lengths, when using the cut sites prior to transformation as features, or without correcting for enzymatic bias (Supplementary Figures S1a-c)

To further benchmark our approach, we compared the accuracy of our method to previously published algorithms for detecting TF binding (Figure 1c). As each method requires different input and has different parameters, we made the runs of all methods as similar as possible to our pipeline (Methods). We tested DeFCoM, an SVM based method for TF binding prediction, as well as a simple footprint depth score which describes the average cut sites in the motif compared to its surrounding region, adapted from (Baek et al., 2017). In addition, we tested Catchitt (Keilwagen et al., 2019), which outputs a prediction for TF binding in windows of 50bp across the complete genome and was one of the winners of the ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge. Our approach, DeFCoM, and Catchitt all exhibit high classification rates, with no method significantly outperforming the other methods (ks test p-value > 0.84 for all pairwise comparison). All methods outperform the more simplistic

footprint depth score (ks test p-value < $2*10^{-6}$). We also evaluated the performance of another genome scanning method, HINT-ATAC (Li et al., 2019), however, it achieved lower classification rates (Supplementary Figure S1d). This is perhaps due to the fact that the default model provided by the software was designed to work on omni-ATAC, a different experimental protocol with a very high signal-to-noise ratio compared to the original ATAC-seq protocol.

Uncovering the TFs that are associated with changes to histone acetylation during DC activation

We applied our method to generate a comprehensive map of the TFs involved in temporal changes to the active regulatory landscape of human Monocyte-derived DC following LPS stimulation. To this end, we collected Monocyte-derived DCs from 5 human donors and stimulated the cells with LPS. To define the set of accessible regulatory regions, we generated ATAC-seq data before stimulation (0h) and at 30min, 2h, 4h, and 24h after stimulation (Methods). To catalog the changes in active regulatory elements post LPS we collected H3K27ac ChIP-seq data before stimulation (0h) and at 1h, 2h, 4h, 6h, 12h and 24h after stimulation. We first defined the complete set of accessible regions by finding peaks of open chromatin and combining the set of peaks from each donor in each time point to a total of 193,922 regions. Next, to generate the labels for the regulatory regions we computed the number of H3K27ac reads around each accessible region. Differential expression analysis (Methods) revealed 8,620 regions that show a significant change in H3K27ac signal across time. We then clustered those regions into 5 temporal activation patterns based on time of peak activation (Figure 2a).

We focused on the regulatory landscape of three of the temporal patterns - two early activated sets of regions, including regions which peak at 1 hour post-stimulation ("immediate-early regions") and regions which peak at 2-4 hours post-stimulation ("early regions"), as well as the set of regions which are only activated 24 hours post-stimulation ("late-24"). We used these regions that showed a significant change in H3K27ac level post LPS as a positive set. For each case, the negative set was a randomly chosen set of motif instances from regulatory regions that show no significant change in H3K27ac levels compared to pre-stimulation at any time point (Methods). For each set of regions, we ran our classification algorithm on all HOCOMOCO motifs (Kulakovskiy et al., 2018) for TFs that are expressed in at least one time point (a total of 279 motifs, Methods). For each set of regions and each motif, we ran our pipeline several times: Using the ATAC cut sites from the time points of peak activation, and using the ATAC cut sites from the closest previous time point.

Our pipeline resulted in a mean AUC PR value from 5-fold cross-validation for each motif in each set of regions in each time point. In addition, we wanted to filter motifs that did not have a better predictive value than expected by chance. To this end, we ran the pipeline on randomly assigned labels (i.e. each motif instance was randomly assigned to the positive or negative set) and computed the AUC PR (Methods). We pooled the results from 3,146 randomized runs to create a null distribution and generated empirical p-values for each AUC value of the original runs.

We detected 162 motifs which showed significant (FDR-adjusted p-value < 0.05) changes in the chromatin in at least one of the "activation time points" (i.e. 30m or 2h for the immediate-early

cluster, 2h or 4h for the early cluster and 24h for the late-24h cluster) (Figure 2b). We also observe an increase in expression for many of the TFs associated with these significant motifs. For each set of regions, the change in expression from time point 0h to peak activation time is greater for TFs associated with significant motifs compared to TFs which binds the motifs for which we do not see significant chromatin changes (one-sided ks-test p-value < 0.02 for all sets of regions, Supplementary Figure S2a-c).

Upon inspecting the TFs that are predictive of the various H3K27ac temporal responses, we find that the majority (75, 47%) of the TFs are exclusively predictive for immediate-early regions. (Figure 2b, S2d). Many (42/75; 56%) of these TFs also change expression in response to LPS (fold change >=2 & pval <=0.01 by DESeq2 (Love et al., 2014)). This set of predictive motifs includes the chromatin remodeler BPTF (Frey et al., 2017), as well as TFs that are associated with early transcriptional response to LPS such as IRF7, FOS, JUN, PU.1 (SPI1), CEBPD and STAT5 (Donnard et al., 2018; Garber et al., 2012; Ko et al., 2015; Yamaoka et al., 1998) (Figure 2c-d). In addition, we see changes in chromatin in both the immediate-early and early regions for motifs of TFs previously associated with the transcriptional response to LPS such as REL, STAT1 and STAT2, IRF1 and IRF2 and NFKB complex (Figure 2c-d). Our results are also consistent with the recently published study which found PRDM1 and RARA as regulators of maturation of monocyte-derived DCs in response to HIV-1 infection (Johnson et al., 2020).

Interestingly, 33% of the motifs that are predictive in the immediate-early and early regions are also predictive in late-24h regions. These include REL, STAT1, IRF1/2, NFKB, PRDM1 and RUNX1 which are known to interact with histone modifying enzymes and are LPS induced (Barutcu et al., 2016; Hoogenkamp et al., 2009; Minnich et al., 2016). These results suggest another role for the TFs involved in the early activation of the cells at a much later time point. Along with these previously known TFs, our model predicts additional factors to be associated with an active chromatin state that, to our knowledge, have not been implicated as part of the LPS response in Dendritic cells. The FOXO1 binding motif is predictive of immediate-early regions, consistent with the role of FOXO1 as a regulator of TLR4 signaling in macrophages (Fan et al., 2010). We find the motif of the Hypoxia-inducible factor 1-alpha (HIF-1A) predictive in the early and late-24h regions. HIF-1A have been previously described as having a crucial role in the inflammatory response of macrophages (Cramer et al., 2003). In the early regions we find p63 as a predictive motif, which was shown to interact with histone deacetylases (Ramsey et al., 2011). Finally, we also observe factors which are predictive only at the late-24h regions, including CREB1 and FOXQ1. CREB1 interacts with histone acetyltransferases and can induce an antiapoptotic survival signal in monocytes and macrophages (Wen et al., 2010; Yuan and Gambee, 2001). FOXQ1 was shown to increase pro-inflammatory potential in monocytes and involved in monocyte migration (Ovsiy et al., 2017). Of note, we also see predictive motifs who are repressors of the LPS response in macrophages such as the anti-inflammatory regulator NR3C1 (GR) (Chinenov et al., 2013, 2014) and NR1D1, which represses TLR4 expression and mediate temporal gating of proinflammatory cytokine responses (Fontaine et al., 2008; Gibbs et al., 2012). The association of these factors to chromatin-modifying enzymes or LPS stimulation and their role as activators or repressors in Dendritic cells needs to be experimentally determined.

To test the sensitivity of our method, we compared our results to previous methods that detect differential TF activity (Methods). First, we ran the software DAStk (Tripodi et al., 2018), which relies on changes in motif occurrences between two sets of regions (Figure S3). We also adapted the algorithm developed by Bagfoot (Baek et al., 2017), which detects TF occupancy changes based on differences in footprint depth and motif-flanking accessibility (Figure S4, Methods). While both methods are able to detect several of the main TFs involved in LPS stimulation, overall they show a lower sensitivity (Figures S3 and S4). This is possibly due to the low number of regulatory regions and motif instances that are used as input which limit the sensitivity of other methods that were designed to explore genome-wide differences between different experimental conditions.

Genetic variation reveals transcription factors associated with H3K27ac signal strength
So far, our pipeline predicts TF motifs which are functional in temporally activated regions, based on a discrete classification of the H3K27ac signal. We next sought out to examine whether we can find TFs associated with the strength of the H3K27ac signal by taking advantage of the genetic variation between our samples. To this end, we called SNPs and indels using the ATAC-seq and ChIP-seq data sets generated for the 5 donors (Methods) and found 584 immediate-early regions and 438 early regions with a genetic variant in exactly one donor. For each one of those regions, we computed a z-score of the H3K27ac signal of the donor with the variant based on the H3K27ac signal distribution from the other 4 donors (Methods). Since one of the H3K27ac ChIP-seq samples at 24h had a low signal-to-noise ratio we excluded it from further analysis and thus decided not to test for variation between donors at 24h using only the remaining four donors. For each set of regions, we computed a motif enrichment score - a modification of the GSEA score (Subramanian et al., 2005) - to associate motifs with regions where a genetic variant resulted in a large change to the H3K27ac signal (Figure 3, Methods).

We find that at immediate-early regions, many of the motifs associated with changes to the H3K27ac signal are also predictive of the immediate-early temporal pattern during activation times, with 40% of associated motifs predictive at 30min and 64% predictive at 2h (Figure 3a). Motifs associated with the H3K27ac signal include main regulators of the LPS response such as IRF1, IRF2, and RELB. Interestingly, we also observe an association between H3K27ac strength and the binding of FOXO1 as well as CXXC1, a member of the SET1 H3K4 methyltransferase complex and a regulator of macrophage phagocytosis (Hui et al., 2018; Lee and Skalnik, 2005). In addition we also find an association between H3K27ac signal and NR1D1 and MAFK which are able to interact with histone deacetylases and acetyltransferases, respectively (Hwang et al., 2013; Yin and Lazar, 2005).

Surprisingly, we do not see any motifs associated with signal strength that are also predictive of the temporal H3K27ac pattern of the early regions up until 4h (Figure 3b). During 4h we see an association for the known LPS-response regulators STAT1, STAT2, and IRF1. We also find NFIL3 associated with signal strength, which can interact with histone deacetylases (Keniry et al., 2013) as well as NFIC, which was shown to be recruited to the *C-FOS* promoter by acetylated histones (O'Donnell et al., 2008). We also see a few motifs associated with H3K27ac strength with a low AUC PR value. Those motifs include many TFs such as SP1, SP2 and NR2C2 (TR4) that can recruit histone deacetylases, and KLF16 which recruits both histone

acetyltransferases and deacetylases (Cui et al., 2011; Daftary et al., 2012; Doetzlhofer et al., 1999; Phan et al., 2004). Since our classifier predicts changes in chromatin state between induced and constant regions, it will not detect TFs that are bound genome-wide. Thus, we can hypothesise that while the LPS regulators are associated with H3K27ac signal changes only in early activated regions, we also find SP1, SP2, KLF16 as potential factors that control deacetylation at 4h post LPS-stimulation genome-wide. Another possible hypothesis is that those factors require co-binding for changes in acetylation, as it was previously shown that communication between the NFKB complex and SP1 effect histone acetylation in the promoter region of the *MCP-1* gene (Boekhoudt et al., 2003). We highlight that due to the low number of regions and donors we are limited in the statistical power for this analysis, and we only present an association and not a causal effect. Nevertheless, our study suggests potential factors that control the strength of H3K27ac signal following LPS stimulation.

Discussion

Computational methods to detect TF binding from open chromatin regions have provided valuable insights and is a great improvement over TF ChIP-seq as it saves time, requires fewer cells and is under more flexible experimental conditions. In this work, we aimed to expand the scope of binding prediction methods and design a pipeline built for prediction of context-dependent chromatin changes, allowing us to detect changes only in a subset of genomic regions of interest. This is designed as a highly modular framework that can be applied to any system with state change to understand the predictability of genomic footprints to either gene expression changes, chromatin state changes or any other label as long as it is possible to define a positive set of activated regions. Our pipeline can be used in conjunction with several different software achieving high classification results, allowing for many researchers to adapt easily with their existing pipelines. In addition, our method has no prior assumption about the expected shape of the chromatin around motifs from the positive set, thus this framework is easily extendable to test the importance of short regulatory motifs which are not known TF binding motifs. It should be noted that according to the motif classification suggested by HOCOMOCO, certain motifs are low confidence and can be found only in a small number of regions. We need to be cautious and not over interpret the results from these motifs as they can be due to technical artifacts.

Changes to the epigenome landscape are an important component of the cellular response of DCs to pathogens (Boukhaled et al., 2019). Here, we aim to gain a greater understanding of the factors involved in histone modifications during DC maturation in response to LPS up to 24 hours post stimulation. Applying our framework, we identified many TFs that could potentially be involved with chromatin-modifying enzymes to establish signatures of active chromatin (H3K27ac). Of the TFs we predict to be important for activating regulatory regions, many were previously shown to interact with chromatin-modifying enzymes, some interact with acetyltransferases while some interact with deacetylases. We highlight that our current data does not allow us to claim any causal relations between the TFs and the active regions. Thus, we cannot determine which of the TFs actively modify the chromatin and which bind to these regions because of their active state. Determining the exact interactions as well as which

chromatin-modifying enzyme the TFs interact with has to be done in future validations. Nevertheless, our work resulted in a valuable map of temporal TF-DNA interactions of the human response to pathogens, and provides an easily extendable framework to be used to answer many other biological questions.

Methods

Human Subjects

Anonymous, healthy donor leukopaks (New York Biologics, Southampton, NY), were used in accordance with UMMS-IRB protocol ID #H00004971.

Cell culture

All cells were maintained at 37°C in 5% CO2 humidified incubators.

Human monocyte-derived dendritic cells

Human dendritic cells were derived from peripheral blood mononuclear cells (PBMCs) isolated from de-identified, healthy donor leukopaks (New York Biologics, Southampton, NY), in accordance with UMMS-IRB protocol ID #H00004971. Mononuclear leukocytes were isolated by gradient centrifugation on Histopaque-1077 (Sigma-Aldrich, St. Louis, MO). CD14+ mononuclear cells were enriched via positive selection using anti-CD14 antibody MicroBead conjugates (Miltenyi, San Diego, CA), according to the manufacturer's protocol. CD14+ cells were then plated at a density of 1 to 2 x 10^6 cells/ml in RPMI-1640 supplemented with 5% heat-inactivated human AB+ serum (Omega Scientific, Tarzana, CA), 20 mM L-glutamine (ThermoFisher, Waltham, MA), 25 mM HEPES pH 7.2 (Sigma-Aldrich), 1 mM sodium pyruvate (ThermoFisher), and 1 x MEM non-essential amino acids (ThermoFisher). Differentiation of the CD14+ monocytes into dendritic cells (human DCs) was promoted by addition of recombinant human GM-CSF and human IL-4; cytokines were produced from HEK293 cells stably transduced with pAIP-hGMCSF-co or pAIP-hIL4-co, respectively, as previously described (Reinhard et al., 2014), with each cytokine supernatant added at a dilution of 1:100. The cells were then stimulated with 100ng/mL LPS for specified time.

Library preparation and Sequencing

ATAC-Seq

For each time point, 5 x 10^5 scraped DC's were collected by centrifugation 500 x g for 5 min. and lysed for ATAC-seq following the protocol described in (Buenrostro et al., 2015). Each sample was tagmented using 12.5 ul Nextera TDE-1 transposase (Illumina) for 30 minutes at 37°C, then quenched by the addition of 5 volumes DNA Binding Buffer (Zymo Research) and cleaned using Zymo Research DNA Clean and Concentrator-5 columns according to the supplied protocol. Tagmented DNA was PCR-amplified using indexed primers as described in (Buenrostro et al., 2015), using total cycle numbers for enrichment as determined empirically by qPCR to minimize PCR duplicates. The resulting libraries were purified twice by Zymo Research DNA Clean and Concentrator-5 columns using a ratio of 5:1 DNA Binding Buffer: Sample, and quantified by Qubit HS-DNA Assay (Thermo Fisher Scientific) and Bioanalyzer

High-Sensitivity DNA (Agilent Technologies). Final ATAC-seq libraries were pooled (equimolar) and sequenced on an Illumina Nextseq 500.

ChIP-Seq

Harvest and Formaldehyde crosslinking: For each timepoint and donor, 5-7 x 10^6 unstimulated or LPS-stimulated dendritic were harvested by scraping in medium and centrifugation at 500 x g for 5 minutes. Each cell pellet was washed once with 2 mL PBS and gentle flicking of the tube, followed by centrifugation at 500 x g for 5 min. Cells were uniformly resuspended in 1 mL 1X Fixing Buffer A from the Covaris tru-ChIP Chromatin Shearing and Reagent Kit and fixed by adding 1 mL 2% methanol-free formaldehyde (Thermo Fisher Scientific) diluted in 1X Fixing Buffer A (1% formaldehyde final, 2.5-3.5x10^6 cells/mL) and rotated end-over-end for 5 min. at room temperature. Fixation was quenched by adding 240 mL Quenching Buffer E (Covaris tru-ChIP kit) and rotating for an additional 5 min. Purified BSA was then added to 0.5% w/v final to prevent cell adherence to the tube, and crosslinked cells were harvested by centrifugation, 500 x g for 5 min. at 4°C. Crosslinked cells were washed twice in 2 mL ice-cold PBS + 0.5% BSA with centrifugation as above, and aliquoted evenly into 3 fresh 1.5 mL tubes during the second wash. Cells were finally pelleted by centrifugation at 16,000 x g, flash-frozen as dry pellets in liquid nitrogen, and stored at -80°C.

Lysis, Shearing, and Quantification: Individual crosslinked cell pellets (1.5-2 x 10^6 cells each) were lysed according to the Covaris tru-ChIP Chromatin Shearing and Reagent Kit instructions. Following lysis, nuclei were resuspended in 130 mL ice-cold Shearing Buffer D3 and transferred to 1.5 mL BioRupter Pico Microtubes (Diagenode) on ice. Chromatin was sheared to uniform fragment lengths (150-400 bp) by sonication at 4°C in a BioRupter Pico (Diagenode) set to 6 cycles of 30s ON and 30s OFF. Sheared chromatin was diluted in 10 volumes of ChRIPA buffer (1X PBS, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.5% sodium deoxycholate, 1% Igepal CA-630, 0.1% SDS, 1X Roche cOmplete Protease Inhibitor Cocktail) and insoluble material was removed by centrifugation >15,000 x g for 10 minutes. Lysate was pre-cleared against 60 mL Dynabeads Protein A (Thermo Fisher Scientific) per 10^6 cells for 2h at 4℃ with end-over-end rotation followed by two rounds of magnetic bead removal and transfer to fresh tubes. 2% of pre-cleared lysate was removed for DNA quantification and the remaining lysate was either flash-frozen in liquid nitrogen and stored at -80°C, or stored overnight at 4℃ for use in immunoprecipitation. For quantification, 2% pre-cleared lysate was treated with 10 mg RNase A (Thermo Fisher Scientific) for 30 min. at 37°C, followed by addition of 100 mg Proteinase K (New England Biolabs) and crosslink reversal overnight at 65°C. DNA was purified using DNA Clean and Concentrator-5 columns (Zymo Research). Average sheared DNA fragment sizes were determined by agarose gel and chromatin yield was estimated by Qubit HS-DNA Assay. 50-100 ng purified DNA was saved as Input.

Chromatin Immunoprecipitation: Antibodies used for ChIP were rabbit anti-H3K27ac (Diagenode C15410196). 1 mg antibody was added to 0.5 mg (anti-H3K27ac) pre-cleared crosslinked lysate and incubated overnight with continuous mixing at 4°C. IgG/chromatin complexes were captured for 1h at room temperature on 25 mL Dynabeads Protein A that were pre-blocked for at least 1h with Blocking Buffer (1X PBS, 0.5% BSA, 0.5% Tween-20). Complexed beads were washed 5 times with ice-cold ChRIPA Buffer, twice with room

temperature RIPA-500 Buffer (10 mM Tris pH 8.0, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS), twice with ice-cold LiCl Wash Buffer (10 mM Tris pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Igepal CA-630, 0.5% sodium deoxycholate), and twice with ice-cold TE buffer. Each chromatin sample was eluted from beads using 50 ul Direct Elution Buffer (10 mM Tris pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.5% SDS) and supplemented with 20 mg RNase A, incubating for 30 min. at 37°C. 20 mg glycogen was added to each bead/eluate suspension, and crosslinks were reversed by the addition of 50 mg Proteinase K and incubation at 37°C for an additional 2h, followed by overnight at 65°C. Dynabeads were removed by magnet capture, and the supernatant was mixed thoroughly with 2.3 volumes of Agencourt AMPure XP (Beckman Coulter) bead suspension and incubated for 10 minutes at room temperature prior to bead capture and washing. Purified DNA was eluted in 10 mM Tris pH 8.0.

Library Preparation and Sequencing: Sequencing libraries were prepared from half of each ChIP sample and 50 ng Input DNA using the Ovation Ultralow System V2 kit (NuGEN) according to supplier's instructions, with the total numbers of enrichment PCR cycles determined empirically for each sample by qPCR to minimize PCR duplication rates. Barcoded libraries were quantified using Qubit HS-DNA Assay, qualified using Agilent Bioanalyzer High-Sensitivity DNA, and pooled for sequencing on Illumina Nextseq 500.

Alignment and processing of reads

Genome reference: All the data generated and used for this paper is aligned to human reference genome hg19

ATAC-Seq

Paired-end reads were trimmed to remove adapter sequence using Cutadapt version 1.3, and then aligned to the reference genome with Bowtie2, version 2.1.0, parameter –X 2000 (Langmead and Salzberg, 2012). The alignments were then filtered using Samtools (Li et al., 2009), version 0.0.19, to remove (i) PCR duplicates, as identified by Picard's MarkDuplicates, and (ii) aligned reads with mapping quality below 4. While the reads were aligned as paired-end to optimize the alignment accuracy, the alignments were then further processed as if they were single-end sequence data, so that each aligned read corresponded to a Tn5 cut-site.

Peak Calling: Each aligned read was first trimmed to the 9-bases at the 5'-end, the region where the Tn5 transposase cuts the DNA, and then extended 10-bases upstream and down, for smoothing. Peaks were called using these adjusted 29-base aligned reads with MACS2 (Zhang et al., 2008), parameters --bw 29 --tsize 29 and --qvalue 0.0001.

Quality Control: Following the standard practice (Buenrostro et al., 2015), for each sample we examined the fragment length distribution, as well as a comparison of the aggregate nucleosome signal to the aggregate nucleosome-free signal over transcription start sites for those genes found to be expressed in at least one time point in our RNA-Seq time series. Signal-to-noise ratios were computed for the peaks as f/(1 –f) where f is the fraction of reads overlapping peaks.

ChIP-Seq
Paired-end reads were trimmed to remove sequencing adapters and leading and trailing bases with quality scores less than 5. Reads that were longer than 36 bases after trimming were kept for further analysis. The reads were then aligned to the reference genome using Bowtie2 with options -k 1 --un-conc to filter out reads that map to multiple locations in the genome and that align un-concordantly. Duplicated reads were filtered out using picard-tools-1.131 MarkDuplicates function. Peaks were then called using MACS2 with --bw=230 --tsize=75 and --qvalue 0.0001.

ATAC normalization
To include only nucleosome-free fragments, we filtered out fragments longer than 180 bp with the alignmentSieve command from DeepTools (Ramírez et al., 2016). Next, to correct sequence bias due to enzymatic sequence preferences we ran seqOutBias (Martins et al., 2018) to get a per-base estimate of read counts. We ran the correction on the length-filtered reads, performing the correction on each strand separately with the parameters "--read-size=35 --shift-counts" and the k-mer mask as recommended by the seqOutBias tutorial:
plus_mask=NXNXXXCXXNNXNNNXXN
minus_mask=NXXNNNXNNXXCXXXNXN

Classifying ATAC peaks based on H3K27 signal
For each ATAC peak, we extracted the number of H3K27ac ChIP-seq reads that align within the peak in each sample, extended by 1000bp on both sides to include bordering histones and merging peaks that were overlapping due to that extension. We removed one sample (donor F33 at 24h after LPS stimulation) due to very low read alignment across all peaks.
To detect peaks with temporal changes, we performed pairwise differential expression using DEseq2 (Love et al., 2014), by comparing all the time points with time point 0h (prior to LPS stimulation) and adding the batch as a covariate to the model. All peaks with an adjusted p-value < 0.05 and an absolute log fold change > 2 were considered as temporal peaks. In addition, we also searched for peaks which show a continuous temporal change with ImpulseDE2 (Fischer et al., 2018), and added to our set of temporal peaks regions with ImpulseDE2 adjusted p-value $<= 0.05$. Then, we used k-means clustering with $k = 5$ to cluster the temporal regions into different temporal clusters. The rest of the regions were classified as constant regions, excluding regions with a low number of aligned ChIP-seq reads (Total normalized count across all samples < 30).

RNA-seq analysis
Reads were aligned to the transcriptome with RSEM. To detect differentially expressed genes we ran DEseq2 for each time point. We took only genes with an average TPM of 10 in at least one time point as our set of expressed genes.

Motif scanning
The full set of TF binding motifs was downloaded from HOCOMOCO v11 (Kulakovskiy et al., 2018). We focused only on motifs of TFs which are expressed in our system, leaving a total of 279 motifs for TFs with at least an average of 10 TPM at any of the time points. We used

PWMscan (Ambrosini et al., 2018) to find motif instances across the genome with the "pwm_mscan_wrapper" command and parameter -e 0.00001

Supervised learning algorithm for detecting H3K27ac patterns

We repeated the following pipeline for each one of the 279 TF binding motifs:

Labels:

For a given temporal cluster, we define our positive set by finding all the motif instances that fall within that set of ATAC peaks with BedTools (Quinlan and Hall, 2010). The negative set is then selected out of all motif instances that fall within the set of constant peaks, downsampled so that the two sets are of equal size. We only ran our pipeline on motifs that had at least 10 motif instances within the positive set.

Features:

We use the ATAC-seq data as the features for our classifier. Every time point is separate, thus for a given motif and a given positive set, we run several classifiers, one for each ATAC-seq time point. We compute the normalized number of ATAC cut sites in each base at a region of 256 bp centered on the motif, where we combine the normalized cut site count from both strands oriented around the motif (e.g. cut site count 5 bp downstream on the plus strand was combined with the cut site count 5 bp upstream on the negative strand).

We represented the cut site features as ratios between regions around the motif, at different levels. The first level is the sum of cut sites around the motif, corrected for library size with the sample-specific DEseq scaling factor. The second level includes the sum of reads of the first half of the window (positions 1-128) divided by the total number of cut sites. The third level includes the first quarter divided by the first half and third quarter divided by the second half. We continue in a similar fashion until the last level in which each odd-numbered position is divided by the sum of cut sites of its position and the subsequent one. The total number of features is identical to the number of cut sites. Each donor was considered separate, i.e. each motif instance translates into 5 samples for the classifier (one for each donor).

Classification:

We ran a random forest classifier using 5-fold cross-validation with the R caret package (Kuhn, 2008). We divide our samples into training and testing based on genomic position, thus for each motif instance data from all 5 donors is either all part of the train set or all part of the test set. In each run, we use the inner 5-fold CV to tune hyper-parameters and apply the model on the test set to compute the area under the curve of the PR curve with the PRROC package (Grau et al., 2015). The final AUC value is determined as the average of the 5 runs.

Testing the significance of AUC PR values

For a given condition (i.e. a combination of ATAC time point and peak label) we collected the set of motifs that achieve an AUC value of at least 0.5. To generate a random distribution of AUC PR values, we randomly select a motif from that set and run the same pipeline as before, except after downsampling we randomly re-assign each motif instance to the positive or negative

set, keeping all biological repeats as all positive or all negative. We repeated this process for 8 different conditions (same conditions as in the original, non-random runs) 200-400 times. As each condition showed a similar distribution of randomized AUC PR values (t-test p-value > 0.4113 for all pairwise comparisons), we collected all results across all conditions to form the null distribution with a total of 3,146 AUC PR values. We then computed an empirical p-value for each AUC PR value from our non-random runs and performed FDR correction for all motifs in each condition separately.

Data preprocessing for GM12878 data

We downloaded previously published ATAC-seq data on GM12878 (Buenrostro et al., 2013). Reads were aligned to hg19 using bowtie2 (Langmead and Salzberg, 2012). We removed low-quality alignments (MAPQ < 10) and reads without a unique alignment, as well as discordant reads and reads mapping to chrM or the ENCODE "blacklist" regions.

For peak calling, reads aligned to the positive strand were shifted +4bp, and reads aligning to the negative strand were shifted -5bp. We called peaks using MACS2 on the cut sites, merging peaks that were less than 10bp apart, leaving a total of 203,977 peaks.

For footprint method evaluation, we removed reads with fragment length > 180bp for all methods except HINT-ATAC, as HINT incorporates the fragment length as part of its model. To account for sequence bias we normalized the data with seqOutBias as described in their tutorial.

TF ChIP-seq peak files for GM12878 were downloaded from the ENCODE portal (Davis et al., 2018). A full list of the TF bed files used is provided as supplementary table 1

Our prediction pipeline for GM12878 data

For each TF binding motif, our positive set was defined as the set of motif instances within an ATAC peak that overlap the corresponding TF ChIP-seq peak, while the negative set is the motif instances within ATAC peak that do not overlap the TF ChIP-seq peaks. We downsampled the set to have equal sizes and removed motifs that had less than 100 total instances after downsampling. The rest of our pipeline was performed as described above.

Footprint depth score for GM12878 data

The footprint depth score was adapted from (Baek et al., 2017). In each motif instance, we define the footprint depth score as the 10% trimmed mean normalized cut site within the motif (extended 2 bp from the motif boundary). From this value, we subtract the mean normalized cut site in the regions flanking the motif, up to a window of 256bp (same window used for the random forest classifier). We multiplied the score in -1 so that a more positive score will be associated with greater footprint depth. Using that score, AUC PR was computed to each one of the five testing sets used for the random forest classifier.

Running DeFCoM

DeFCoM requires a BAM file as input, thus we ran DeFCoM on the processed BAM file after removing long fragments and read shifting, with the default parameters described in the example config file from the DeFCoM website. In each iteration of the 5-fold cross-validation we ran DeFCoM with the same test and train sets as the random forest classifier.

Running Catchitt

Labels for each 50bp window across the genome were computed with Catchitt's "labels" commend, taking the ENCODE ChIP-seq peak file as input. Chromatin accessibility was computed with the "access" command, providing the processed BAM file as input. The "motif" command provided motif scores for each window. Since training and testing are performed on entire chromosomes we implemented a greedy algorithm to ensure a balanced 5-fold training set. We ranked the chromosomes based on the number of ChIP-seq peaks found in it from highest to lowest. We then sorted the chromosomes into five sets, each time adding the remaining chromosome with the highest number of ChIP-seq peaks into the bin with the lowest total number of ChIP-seq peaks. In each iteration, one of the sets was used for training while another set was used for testing. For the testing chromosomes, we computed the AUC PR of windows that overlap motif instances that fall within an ATAC-seq peak, with the instance label determined by the ChIP-seq and downsampling the positive and negative set to be of equal sizes. In case a motif instance spanned the edges of two windows, the score of the instance was the mean of the score for the two windows.

Running HINT-ATAC

We ran HINT-ATAC on the full set of peaks and on all aligned reads with the following parameters: "--atac-seq --paired-end --organism=hg19".

To compute AUC values, for each set of test data used by the random forest classifier we assigned each motif the HINT score that overlaps it, or zero if it didn't overlap any HINT footprints. We used those values to compute the AUC score for each test set.

Running DAStk on DC data

For each set of regions, we ran DAStk using default parameters, comparing the set of induced regions to the set of regions classified as constant. ΔMD score was computed as the difference in the MD scores between the induced and the constant regions from the output of the "differential_md_score" command.

Running BagFoot on DC data

We adapted the Bagfoot algorithm as described by (Baek et al., 2017) to run on our normalized data. For each motif instance, we computed the footprint depth score and the flanking accessibility score. The footprint depth score is calculated as above, except we take a window of 200bp. The flanking accessibility score is the mean normalized cut site count of the 200bp centered around the motif. The footprint depth difference is taken as the difference between the mean footprint depth score of motif instances in the induced region and the mean footprint depth score of motif instances in the constant regions, and similarly for the normalized cut count difference. The Bagplot and p-value and adjusted p-value computation were done with the code of the gen_bagplot_chisq function, with minor modifications to work on our input data.

SNP and variant calling

We used GATK with the steps described in the best practices guide published by the GATK developers: https://software.broadinstitute.org/gatk/best-practices. We performed read grouping, base quality score recalibration (BQSR) on the BAM files for which we used the default parameters dbSNP-147 VCF file. Raw variants identified by the genotyping tool were

recalibrated using dbSNP-147 VCF file and default parameters. The variants are then refined using VariantFiltration, with parameters QualByDepth, FisherStrand, StrandOddsRatio, RMSMappingQuality, MQRankSum, and ReadPosRankSum were <2.0, <40.0, >60.0, >3.0, <12.5 and <−8.0 as recommended, respectively.

Association between H3K27ac signal and motif abundance

For this analysis, we only considered regions classified as immediate-early or early regions with a variant in only one donor. For each peak, we computed the z-score of the normalized H3K27ac signal for the donor with the variant based on the mean and standard deviation of the signal from the other 4 donors. For a given set of regions and a given time point, we performed an enrichment test for whether a motif is enriched in peaks with high z-scores: for each motif, we counted the number of motif instances in each region. We computed the enrichment score as described in (Subramanian et al., 2005), where the number of motif instances in the peak was used as the magnitude of increment in each step and the absolute value of the z-score was the weight of each region, normalized to sum up to one. We used the absolute z-score since we wanted to test the association between a TF motif and the magnitude of the effect of a variant on the H3K27ac signal. However, we don't observe a significant change in the results when taking the signed z-score (supplementary figure 5) or when the magnitude of increment was one if the region had a motif instance (instead of the number of motif instances, supplementary figure 6). For the significance of the enrichment score, we shuffled the z-scores between the peaks 10,000 times and computed an empirical p-value.

Figures

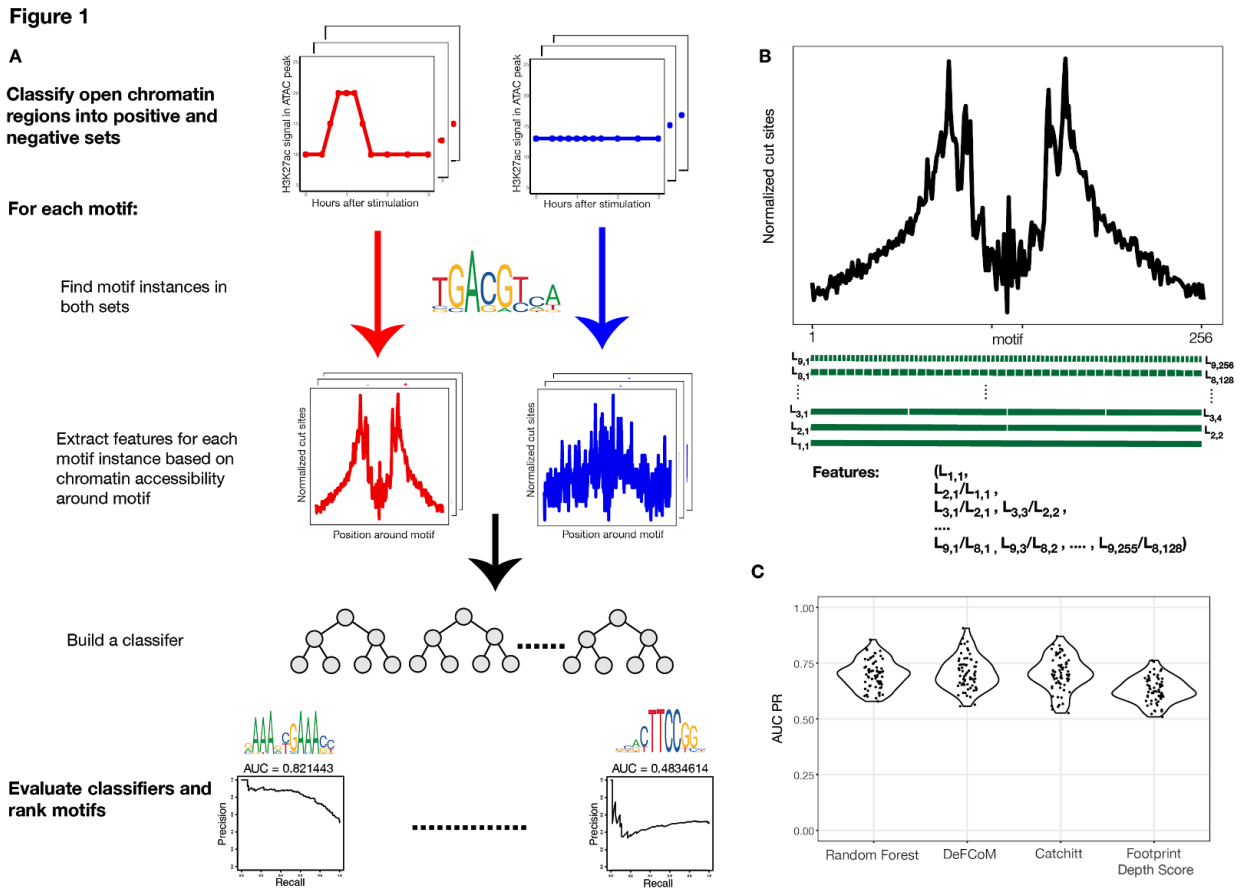


Figure 1: A supervised learning approach to detect functional motifs sequences. A) Illustration of the pipeline. First, regulatory regions are divided into positive and negative sets based on some features of their activity (e.g. activation time). Then, for each sequence motif, we build a random forest classifier using the cut sites from the chromatin accessibility assay as our features. We then evaluate the success for each classifier. High scoring motifs indicate a difference in chromatin structure due to the functionality of the motif. B) Illustration of the cut site transformation used as features for the classifier. C) Evaluation and comparison of the random forest classifier. AUC PR values of the classifier when used to predict TF binding in 66 TF binding motifs. Results are compared to the AUC PR values of two previously published methods for TF binding prediction - DeFCoM and Catchitt, as well as a simple method of predicting binding with the footprint depth score. AUC PR values are mean across 5-fold cross-validation

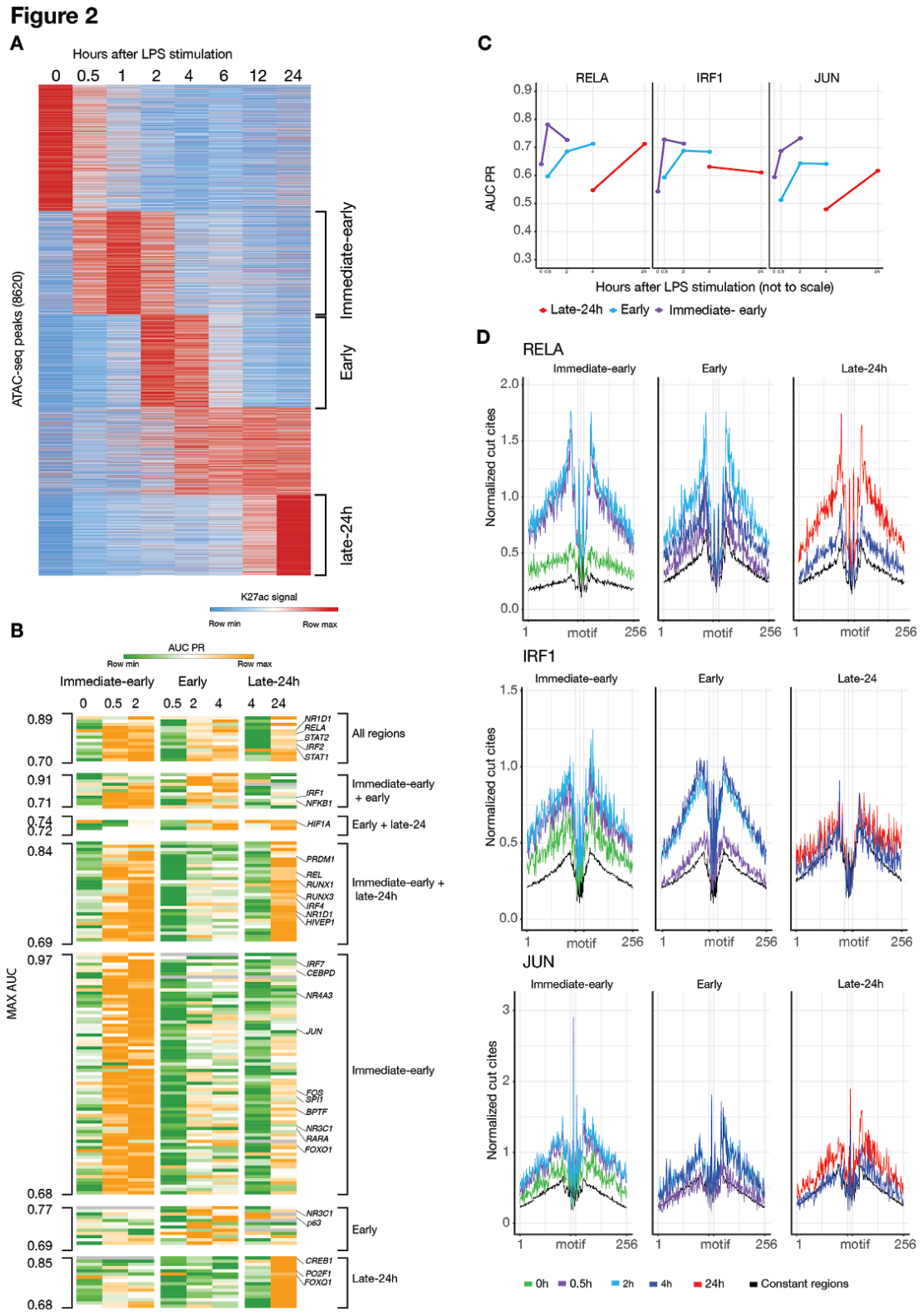


Figure 2: Comprehensive map of predictive TF binding motifs in temporally-activated regulatory regions. A) Mean normalized H3K27ac signal from 5 donors in all regulatory

regions which exhibit temporal changes in H3K27ac signal across time after LPS stimulation. B) Heatmap summarizing the AUC PR values of all motifs that were found to be significant (FDR < 0.05) in at least one set of regions in one peak activation time point. Column names are the region's temporal cluster and time after LPS stimulation. Rows are grouped by the set of regions in which the motif was significant, and each group is ordered by the max AUC PR value of each row. C) Plots highlighting the AUC PR values from B for 3 motifs of TFs known to be involved in the DC response to LPS. D) Mean normalized cut sites for each of the motifs from C in all sets of regions before and during peak H3K27ac signal. Cut site counts from the constant regions (black lines) were computed at the earliest time point in each plot.

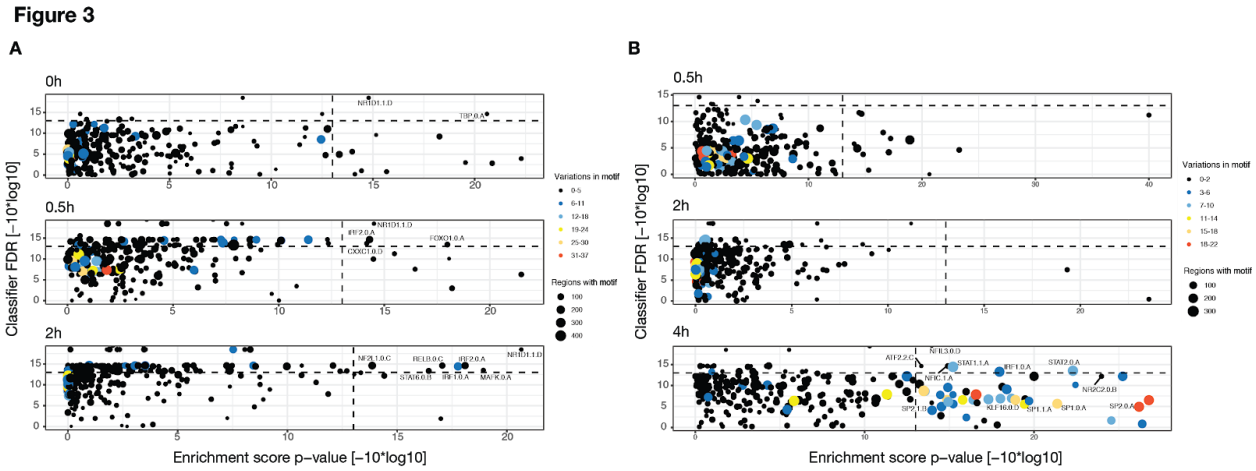


Figure 3: Association between TF binding motifs and H3K27ac signal strength. A) Enrichment of motifs in immediate-early regions that exhibit a strong change in their H3K27ac signal in donors with a genetic variant. The x-axis shows the p-value of the association test between each motif and the z-scores of the H3K27ac signal (Methods) at three different time points. Y-axis shows the FDR corrected p-value of our classifier at each time point. Motifs with an enrichment p-value < 0.05 and a classification FDR corrected p-value < 0.05 are named in the plot. The size of each point represents the number of immediate-early regions with at least one motif instance. Each point is colored based on the number of regions in which the genetic variants overlap the motif instance. B) Same as A, except for early induced regions. For the 4h plot, in addition to motifs named as in A, we also named a few motifs that only have an enrichment p-value < 0.05 and are known to be associated with histone acetylation and deacetylation.

References

Ambrosini, G., Groux, R., and Bucher, P. (2018). PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics *34*, 2483–2484.

Amit, I., Garber, M., Chevrier, N., Leite, A.P., Donner, Y., Eisenhaure, T., Guttman, M., Grenier, J.K., Li, W., Zuk, O., et al. (2009). Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science *326*, 257–263.

Baek, S., Goldstein, I., and Hager, G.L. (2017). Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. Cell Rep. *19*, 1710–1722.

Barutcu, A.R., Hong, D., Lajoie, B.R., McCord, R.P., van Wijnen, A.J., Lian, J.B., Stein, J.L., Dekker, J., Imbalzano, A.N., and Stein, G.S. (2016). RUNX1 contributes to higher-order chromatin organization and gene regulation in breast cancer cells. Biochim. Biophys. Acta *1859*, 1389–1397.

Boekhoudt, G.H., Guo, Z., Beresford, G.W., and Boss, J.M. (2003). Communication between NF-kappa B and Sp1 controls histone acetylation within the proximal promoter of the monocyte chemoattractant protein 1 gene. J. Immunol. *170*, 4139–4147.

Boukhaled, G.M., Corrado, M., Guak, H., and Krawczyk, C.M. (2019). Chromatin Architecture as an Essential Determinant of Dendritic Cell Function. Front. Immunol. *10*, 1119.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1–9.

Chinenov, Y., Gupte, R., and Rogatsky, I. (2013). Nuclear receptors in inflammation control: repression by GR and beyond. Mol. Cell. Endocrinol. *380*, 55–64.

Chinenov, Y., Coppo, M., Gupte, R., Sacta, M.A., and Rogatsky, I. (2014). Glucocorticoid receptor coordinates transcription factor-dominated regulatory network in macrophages. BMC Genomics *15*, 656.

Cramer, T., Yamanishi, Y., Clausen, B.E., Förster, I., Pawlinski, R., Mackman, N., Haase, V.H., Jaenisch, R., Corr, M., Nizet, V., et al. (2003). HIF-1alpha is essential for myeloid cell-mediated inflammation. Cell *112*, 645–657.

Cui, S., Kolodziej, K.E., Obara, N., Amaral-Psarris, A., Demmers, J., Shi, L., Engel, J.D., Grosveld, F., Strouboulis, J., and Tanabe, O. (2011). Nuclear receptors TR2 and TR4 recruit multiple epigenetic transcriptional corepressors that associate specifically with the embryonic β-type globin promoters in differentiated adult erythroid cells. Mol. Cell. Biol. *31*, 3298–3311.

Daftary, G.S., Lomberk, G.A., Buttar, N.S., Allen, T.W., Grzenda, A., Zhang, J., Zheng, Y., Mathison, A.J., Gada, R.P., Calvo, E., et al. (2012). Detailed structural-functional analysis of the Krüppel-like factor 16 (KLF16) transcription factor reveals novel mechanisms for silencing Sp/KLF sites involved in metabolism and endocrinology. J. Biol. Chem. *287*, 7010–7025.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Doetzlhofer, A., Rotheneder, H., Lagger, G., Koranda, M., Kurtev, V., Brosch, G., Wintersberger, E., and Seiser, C. (1999). Histone deacetylase 1 can repress transcription by binding to Sp1. Mol. Cell. Biol. *19*, 5504–5511.

Donnard, E., Vangala, P., Afik, S., McCauley, S., Nowosielska, A., Kucukural, A., Tabak, B., Zhu, X., Diehl, W., McDonel, P., et al. (2018). Comparative Analysis of Immune Cells Reveals a Conserved Regulatory Lexicon. Cell Syst *6*, 381–394.e7.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fan, W., Morinaga, H., Kim, J.J., Bae, E., Spann, N.J., Heinz, S., Glass, C.K., and Olefsky, J.M. (2010). FoxO1 regulates Tlr4 inflammatory pathway signalling in macrophages. EMBO J. *29*, 4223–4236.

Fischer, D.S., Theis, F.J., and Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. Nucleic Acids Res. *46*, e119.

Fontaine, C., Rigamonti, E., Pourcet, B., Duez, H., Duhem, C., Fruchart, J.-C., Chinetti-Gbaguidi, G., and Staels, B. (2008). The nuclear receptor Rev-erbalpha is a liver X receptor (LXR) target gene driving a negative feedback loop on select LXR-induced pathways in human macrophages. Mol. Endocrinol. *22*, 1797–1811.

Frey, W.D., Chaudhry, A., Slepicka, P.F., Ouellette, A.M., Kirberger, S.E., Pomerantz, W.C.K., Hannon, G.J., and Dos Santos, C.O. (2017). BPTF Maintains Chromatin Accessibility and the Self-Renewal Capacity of Mammary Gland Stem Cells. Stem Cell Reports *9*, 23–31.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Gibbs, J.E., Blaikley, J., Beesley, S., Matthews, L., Simpson, K.D., Boyce, S.H., Farrow, S.N., Else, K.J., Singh, D., Ray, D.W., et al. (2012). The nuclear receptor REV-ERBα mediates circadian regulation of innate immunity through selective regulation of inflammatory cytokines. Proc. Natl. Acad. Sci. U. S. A. *109*, 582–587.

González, A.J., Setty, M., and Leslie, C.S. (2015). Early enhancer establishment and regulatory

locus complexity shape transcriptional programs in hematopoietic differentiation. Nat. Genet. *47*, 1249–1259.

Grau, J., Grosse, I., and Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics *31*, 2595–2597.

Gusmao, E.G., Allhoff, M., Zenke, M., and Costa, I.G. (2016). Analysis of computational footprinting methods for DNase sequencing experiments. Nat. Methods *13*, 303–309.

Hoogenkamp, M., Lichtinger, M., Krysinska, H., Lancrin, C., Clarke, D., Williamson, A., Mazzarella, L., Ingram, R., Jorgensen, H., Fisher, A., et al. (2009). Early chromatin unfolding by RUNX1: a molecular explanation for differential requirements during specification versus maintenance of the hematopoietic gene expression program. Blood *114*, 299–309.

Hui, Z., Zhou, L., Xue, Z., Zhou, L., Luo, Y., Lin, F., Liu, X., Hong, S., Li, W., Wang, D., et al. (2018). Cxxc Finger Protein 1 Positively Regulates GM-CSF-Derived Macrophage Phagocytosis Through Csf2rα-Mediated Signaling. Front. Immunol. *9*, 1885.

Hwang, Y.-J., Lee, E.-W., Song, J., Kim, H.-R., Jun, Y.-C., and Hwang, K.-A. (2013). MafK positively regulates NF-κB activity by enhancing CBP-mediated p65 acetylation. Sci. Rep. *3*, 3242.

Johnson, J.S., De Veaux, N., Rives, A.W., Lahaye, X., Lucas, S.Y., Perot, B.P., Luka, M., Garcia-Paredes, V., Amon, L.M., Watters, A., et al. (2020). A Comprehensive Map of the Monocyte-Derived Dendritic Cell Transcriptional Network Engaged upon Innate Sensing of HIV. Cell Rep. *30*, 914–931.e9.

Keilwagen, J., Posch, S., and Grau, J. (2019). Accurate prediction of cell type-specific transcription factor binding. Genome Biol. *20*, 9.

Keniry, M., Pires, M.M., Mense, S., Lefebvre, C., Gan, B., Justiano, K., Lau, Y.-K.I., Hopkins, B., Hodakoski, C., Koujak, S., et al. (2013). Survival factor NFIL3 restricts FOXO-induced gene expression in cancer. Genes Dev. *27*, 916–927.

Ko, C.-Y., Chang, W.-C., and Wang, J.-M. (2015). Biological roles of CCAAT/Enhancer-binding protein delta during inflammation. J. Biomed. Sci. *22*, 6.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Articles *28*, 1–26.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat.

Methods *9*, 357–359.

Lee, J.-H., and Skalnik, D.G. (2005). CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. J. Biol. Chem. *280*, 41725–41731.

Li, Z., Schulz, M.H., Look, T., Begemann, M., Zenke, M., and Costa, I.G. (2019). Identification of transcription factor binding sites using ATAC-seq. Genome Biol. *20*, 45.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Martins, A.L., Walavalkar, N.M., Anderson, W.D., Zang, C., and Guertin, M.J. (2018). Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. Nucleic Acids Res. *46*, e9.

Minnich, M., Tagoh, H., Bönelt, P., Axelsson, E., Fischer, M., Cebolla, B., Tarakhovsky, A., Nutt, S.L., Jaritz, M., and Busslinger, M. (2016). Multifunctional role of the transcription factor Blimp-1 in coordinating plasma cell differentiation. Nat. Immunol. *17*, 331–343.

Natarajan, A., Yardımcı, G.G., Sheffield, N.C., Crawford, G.E., and Ohler, U. (2012). Predicting cell-type–specific gene expression from regions of open chromatin. Genome Res. *22*, 1711–1722.

O'Donnell, A., Yang, S.-H., and Sharrocks, A.D. (2008). MAP kinase-mediated c-fos regulation relies on a histone acetylation relay switch. Mol. Cell *29*, 780–785.

Ovsiy, I., Riabov, V., Manousaridis, I., Michel, J., Moganti, K., Yin, S., Liu, T., Sticht, C., Kremmer, E., Harmsen, M.C., et al. (2017). IL-4 driven transcription factor FoxQ1 is expressed by monocytes in atopic dermatitis and stimulates monocyte migration. Sci. Rep. *7*, 16847.

Phan, D., Cheng, C.-J., Galfione, M., Vakar-Lopez, F., Tunstead, J., Thompson, N.E., Burgess, R.R., Najjar, S.M., Yu-Lee, L.-Y., and Lin, S.-H. (2004). Identification of Sp2 as a transcriptional repressor of carcinoembryonic antigen-related cell adhesion molecule 1 in tumorigenesis. Cancer Res. *64*, 3072–3078.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447–455.

Quach, B., and Furey, T.S. (2017). DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. Bioinformatics *33*, 956–963.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D.J., Pauli, A., Hacohen,

N., Schier, A.F., Blackshear, P.J., Friedman, N., et al. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. Cell *159*, 1698–1710.

Raj, A., Shim, H., Gilad, Y., Pritchard, J.K., and Stephens, M. (2015). msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding. PLoS One *10*, e0138030.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.

Ramsey, M.R., He, L., Forster, N., Ory, B., and Ellisen, L.W. (2011). Physical association of HDAC1 and HDAC2 with p63 mediates transcriptional repression and tumor maintenance in squamous cell carcinoma. Cancer Res. *71*, 4373–4379.

Reinhard, C., Bottinelli, D., Kim, B., and Luban, J. (2014). Vpx rescue of HIV-1 from the antiviral state in mature dendritic cells is independent of the intracellular deoxynucleotide concentration. Retrovirology *11*, 12.

Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J.K., Ebert, P., Nordström, K., Barann, M., Sinha, A., et al. (2017). Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. Nucleic Acids Res. *45*, 54–66.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Sung, M.-H., Guertin, M.J., Baek, S., and Hager, G.L. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol. Cell *56*, 275–285.

Tripodi, I.J., Allen, M.A., and Dowell, R.D. (2018). Detecting Differential Transcription Factor Activity from ATAC-Seq Data. Molecules *23*.

Vandenbon, A., Kumagai, Y., Lin, M., Suzuki, Y., and Nakai, K. (2018). Waves of chromatin modifications in mouse dendritic cells in response to LPS stimulation. Genome Biol. *19*, 138.

Wen, A.Y., Sakamoto, K.M., and Miller, L.S. (2010). The role of the transcription factor CREB in immune function. J. Immunol. *185*, 6413–6419.

Xu, T., Zheng, X., Li, B., Jin, P., Qin, Z., and Wu, H. (2018). A comprehensive review of computational prediction of genome-wide features. Brief. Bioinform.

Yamaoka, K., Otsuka, T., Niiro, H., Arinobu, Y., Niho, Y., Hamasaki, N., and Izuhara, K. (1998). Activation of STAT5 by lipopolysaccharide through granulocyte-macrophage

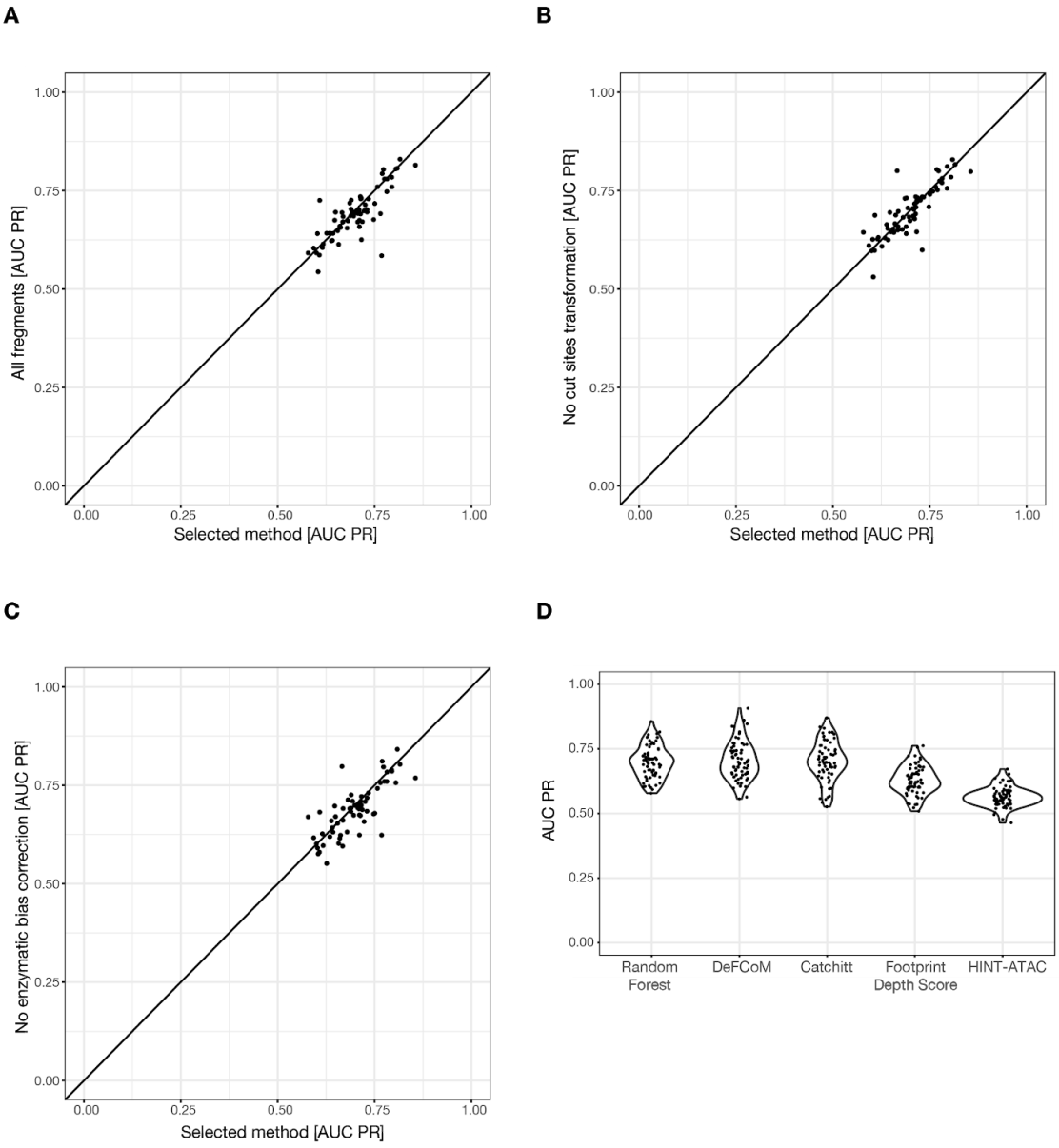
colony-stimulating factor production in human monocytes. J. Immunol. *160*, 838–845.

Yin, L., and Lazar, M.A. (2005). The orphan nuclear receptor Rev-erbalpha recruits the N-CoR/histone deacetylase 3 corepressor to regulate the circadian Bmal1 gene. Mol. Endocrinol. *19*, 1452–1459.

Yuan, L.W., and Gambee, J.E. (2001). Histone acetylation by p300 is involved in CREB-mediated transcription on chromatin. Biochim. Biophys. Acta *1541*, 161–169.

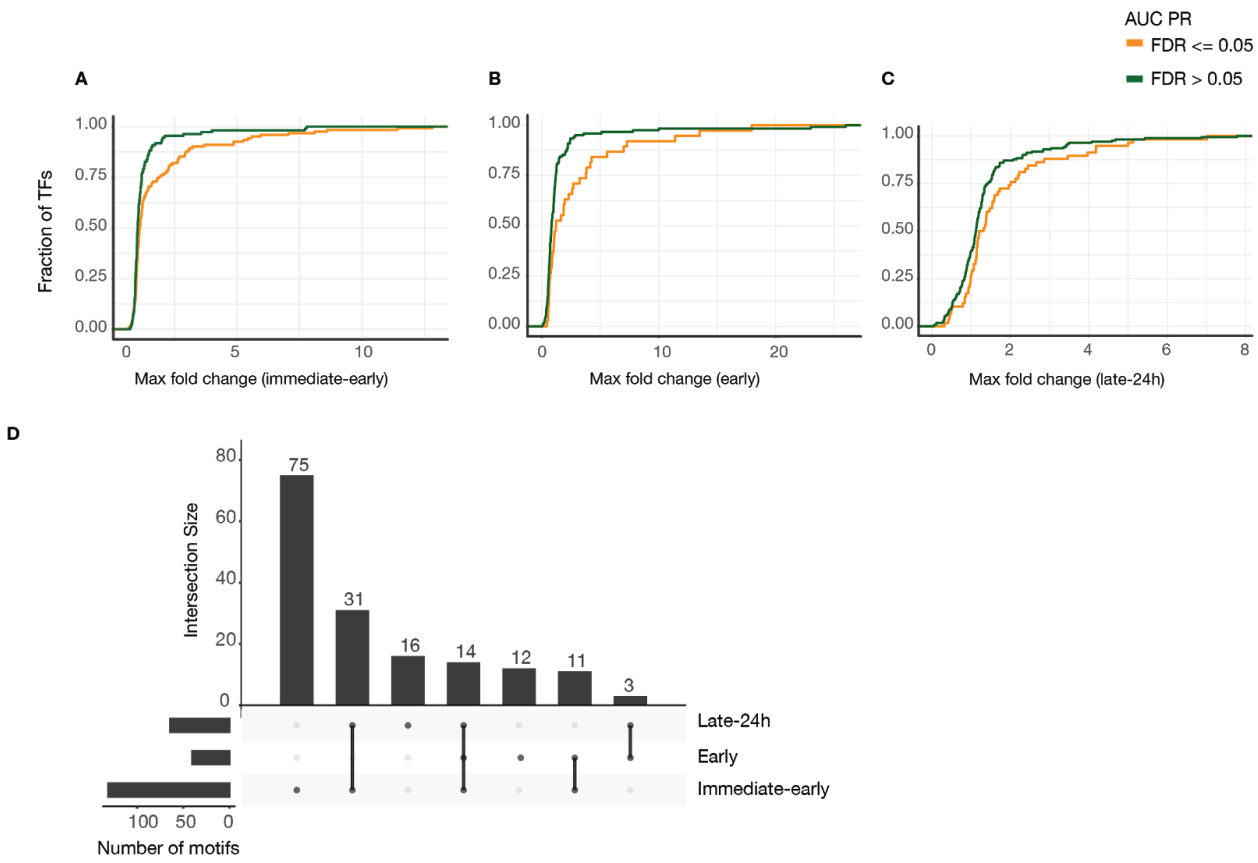
Supplementary Figures

Supplementary Figure 1



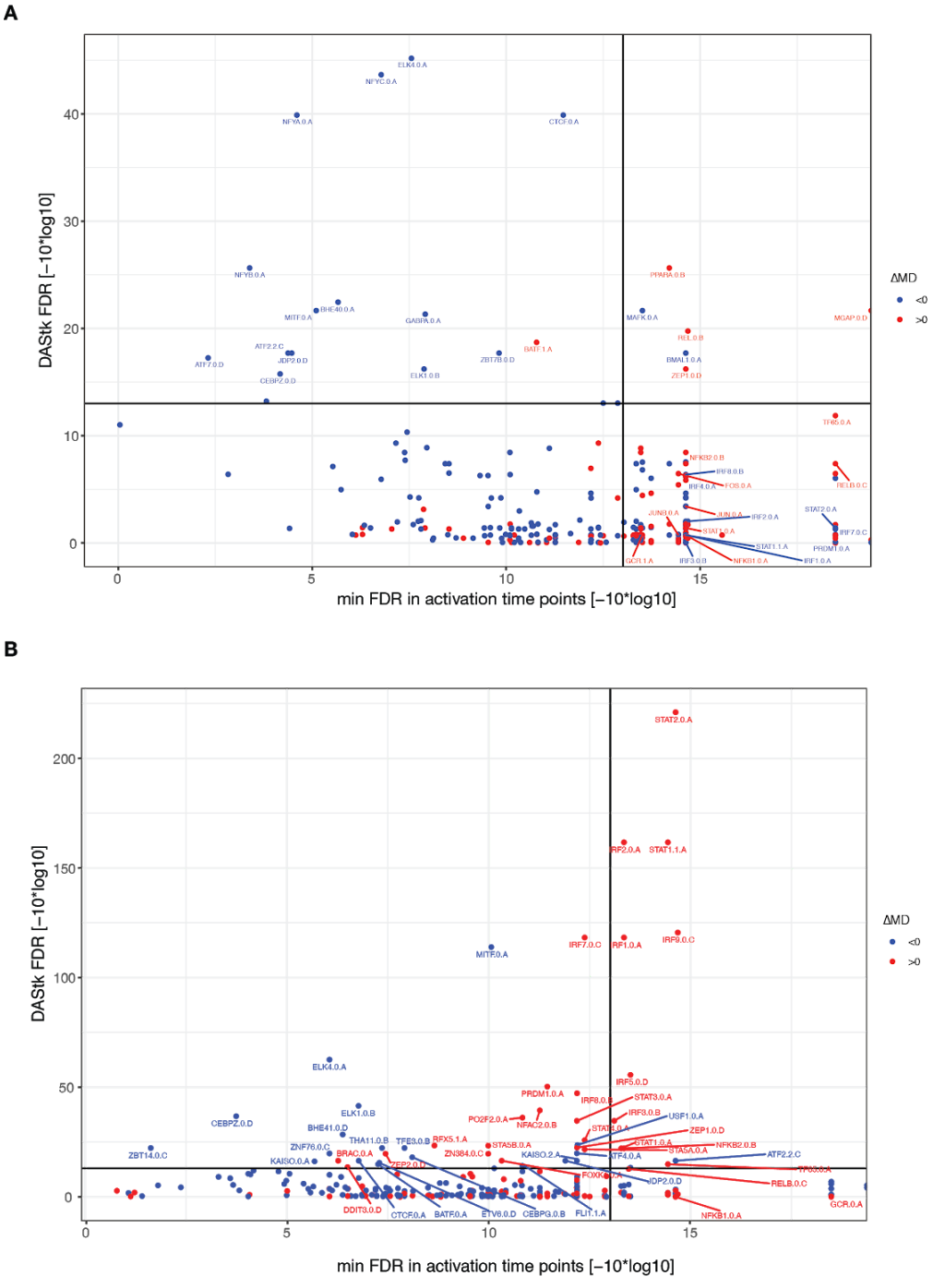
Supplementary figure 1: A) Mean AUC PR values across 5-fold cross-validation of our selected method (x-axis), against the same pipeline but including all read fragment lengths instead of only short fragments (y-axis). B) Similar to A, except the y-axis depicts the same pipeline as our selected method but using the cut site counts in each position as features, instead of the transformed cut sites. C) Similar to A, except the y-axis depicts the same pipeline as our selected method without correcting for enzymatic bias. D) Evaluation and comparison of the random forest classifier. Same as figure 1c, but including also the results of HINT-ATAC.

Supplementary Figure 2



Supplementary figure 2: (A-C) TFs of significant motifs in temporally-activated regulatory regions show increase in expression following LPS stimulation. Cumulative distribution function of the maximum TPM fold change from peak activation time points to time point 0 for the immediate-early (A), early (B) and late (C) regions. TFs whose binding motifs were found to be significant (FDR adjusted p-value <= 0.05) in peak activation time points are in orange, while TFs with an FDR adjusted p-value > 0.05 are in green. For time points 30m and 2h, if a motif was significant in one set of regions (e.g. immediate-early) but was not significant at the other set (e.g. early), it was discarded from the analysis for the set of regions in which it was not significant. D) Plot summarizing the number of significant motifs in each set of regions.

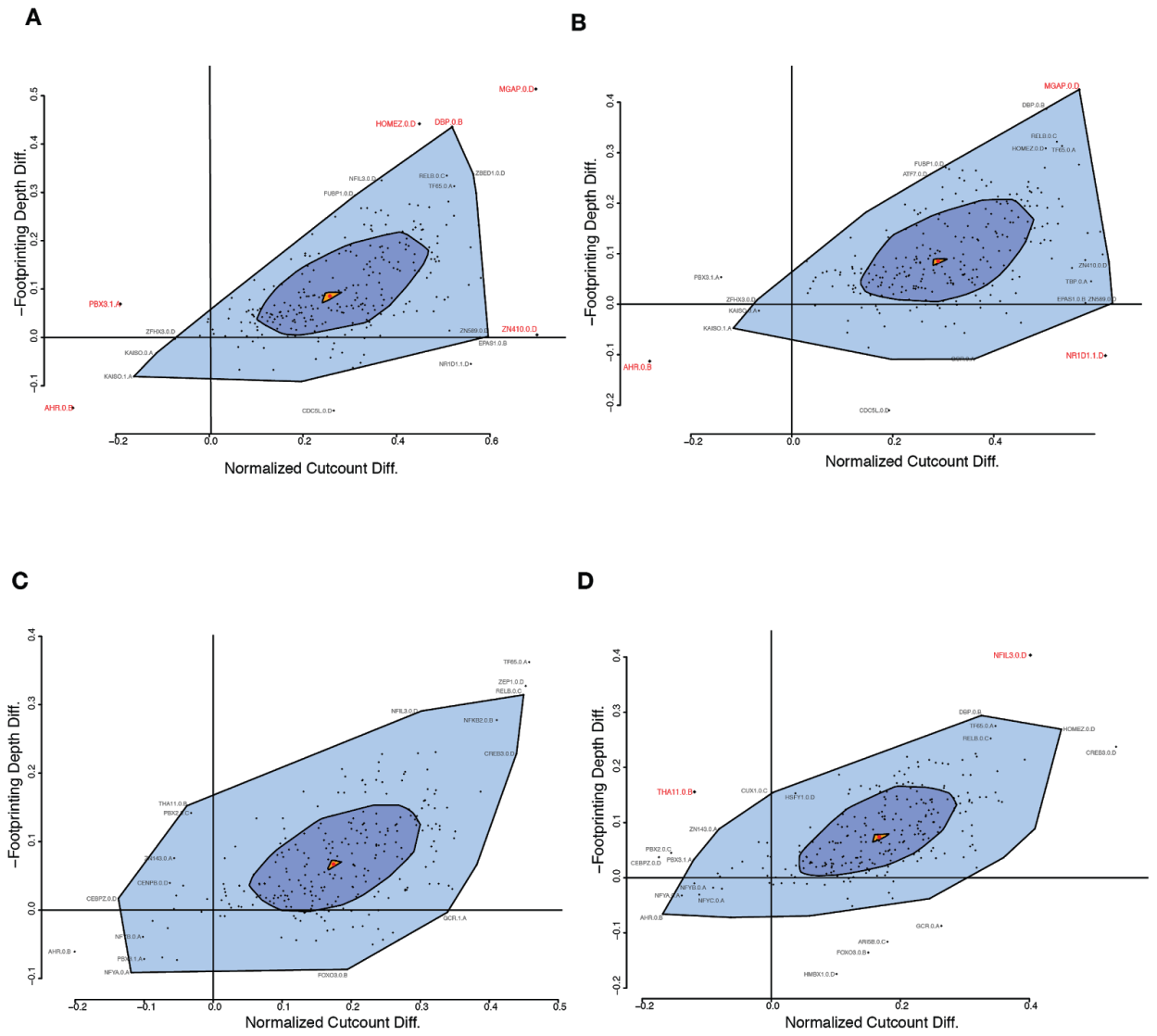
Supplementary Figure 3



Supplementary figure 3: Comparison to DAStk. DAStk results for the set of immediate-early regions (A) and early regions (B). Y-axis shows the FDR adjusted p-value of DAStk, while the x-axis shows the minimum of FDR adjusted p-value of our classifier in the activation time points (30m and 2h for the immediate-early regions, 2h and 4h for the early regions). Motifs are colored based on the ΔMD score, where a positive value indicates that this motif is enriched in the

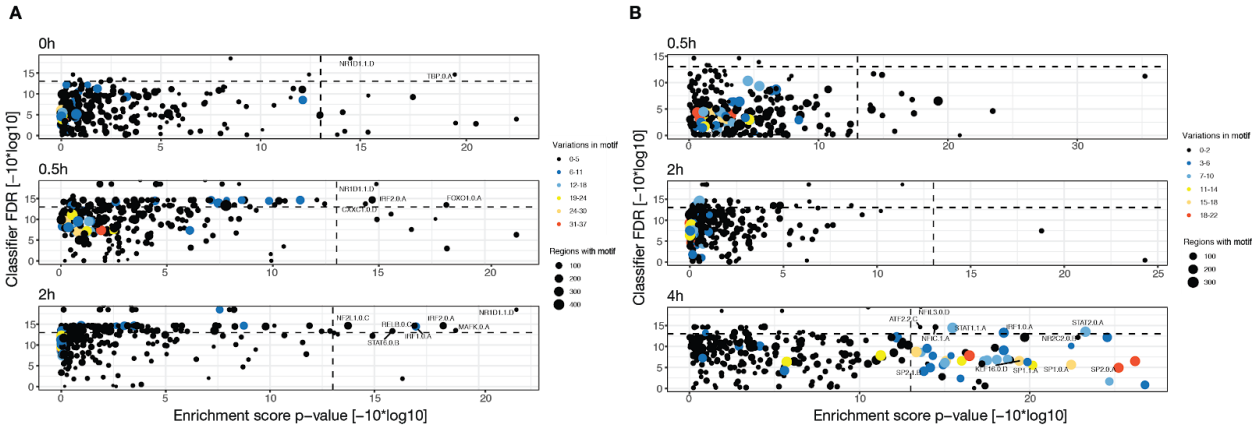
induced regions. Horizontal and vertical lines show an FDR value of 0.05. Motifs with an FDR < 0.05 by DAStk and a few other selected motifs are named in the plot.

Supplementary Figure 4



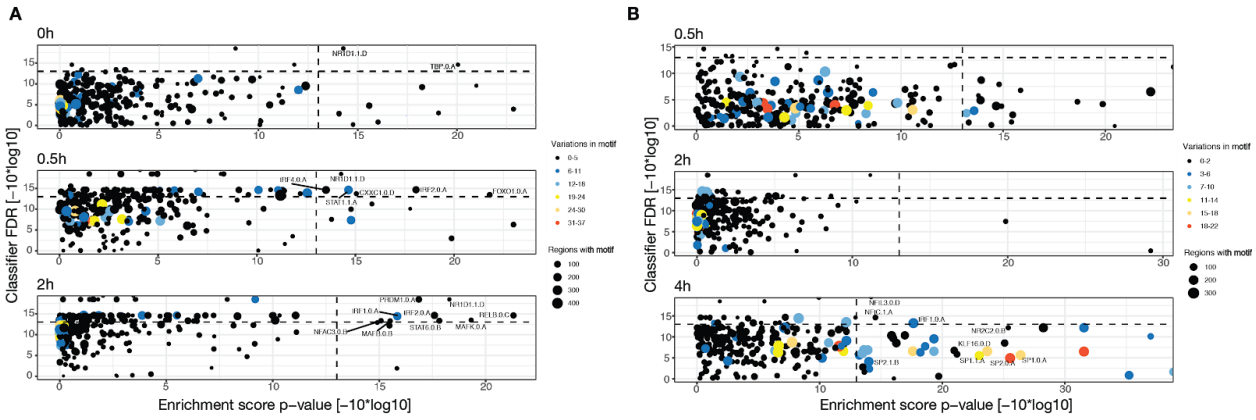
Supplementary figure 4: **Results of the Bagfoot algorithm**. Bag plot depicting the difference in footprint depth and flanking accessibility of each motif in the immediate early regions at 30m (A) and 2h (B), and bag plots for the early regions at 2h (C) and 4h (D). Motifs with a BH corrected p-value <= 0.05 are named in red, while motifs with a BH corrected p-value > 0.05 but a p-value <= 0.05 are named in black.

Supplementary Figure 5



Supplementary Figure 5: **Association between TF binding motifs and H3K27ac signal strength using the signed z-score as region weights.** Enrichment of motifs in immediate-early regions (A) and early regions (B) that exhibit a strong change in their H3K27ac signal in donors with a genetic variant. This plot is similar to Figure 3, except when computing the enrichment score for each motif, the weight of each region was the signed z-score, instead of the absolute value of the z-score.

Supplementary Figure 6



Supplementary Figure 6: **Association between TF binding motifs and H3K27ac signal strength without the number of motifs in each region as magnitude of increment.** Enrichment of motifs in immediate-early regions (A) and early regions (B) that exhibit a strong change in their H3K27ac signal in donors with a genetic variant. This plot is similar to Figure 3, except that the magnitude of increment is one for each region, instead of the number of motif instances within the region.

Supplementary Tables

Supplementary Table 1: List of TF ChIP files from ENCODE used for method evaluations

Motif ID (HOCOMOCO)	TF name	TF ChIP Bed ENCODE ID
ARNT_HUMAN.H11MO.0.B	ARNT	ENCFF794KET
ATF2_HUMAN.H11MO.0.B	ATF2	ENCFF133GHG
ATF2_HUMAN.H11MO.1.B	ATF2	ENCFF133GHG
ATF2_HUMAN.H11MO.2.C	ATF2	ENCFF133GHG
BACH1_HUMAN.H11MO.0.A	BACH1	ENCFF748WOQ
BATF_HUMAN.H11MO.0.A	BATF	ENCFF482FJT
BATF_HUMAN.H11MO.1.A	BATF	ENCFF482FJT
CEBPB_HUMAN.H11MO.0.A	CEBPB	ENCFF701HMB
CEBPZ_HUMAN.H11MO.0.D	CEBPZ	ENCFF235AEB
COE1_HUMAN.H11MO.0.A	EBF1	ENCFF382VEJ
CREM_HUMAN.H11MO.0.C	CREM	ENCFF642JEY
CTCF_HUMAN.H11MO.0.A	CTCF	ENCFF096AKZ
E2F4_HUMAN.H11MO.0.A	E2F4	ENCFF850MAC
E2F4_HUMAN.H11MO.1.A	E2F4	ENCFF850MAC
EGR1_HUMAN.H11MO.0.A	EGR1	ENCFF002CGW
ELF1_HUMAN.H11MO.0.A	ELF1	ENCFF880NTF
ELK1_HUMAN.H11MO.0.B	ELK1	ENCFF434DKI
ERR1_HUMAN.H11MO.0.A	ESRRA	ENCFF077VXQ
ETS1_HUMAN.H11MO.0.A	ETS1	ENCFF565SXH
GABPA_HUMAN.H11MO.0.A	GABPA	ENCFF627POZ
HSF1_HUMAN.H11MO.0.A	HSF1	ENCFF662JYS
HTF4_HUMAN.H11MO.0.A	TCF12	ENCFF237IPT

IRF3_HUMAN.H11MO.0.B	IRF3	ENCFF880CYV
IRF4_HUMAN.H11MO.0.A	IRF4	ENCFF708VKT
JUNB_HUMAN.H11MO.0.A	JUNB	ENCFF784PEF
JUND_HUMAN.H11MO.0.A	JUND	ENCFF321KTX
MAFK_HUMAN.H11MO.0.A	MAFK	ENCFF112CKJ
MAFK_HUMAN.H11MO.1.A	MAFK	ENCFF112CKJ
MAX_HUMAN.H11MO.0.A	MAX	ENCFF407JNK
MAZ_HUMAN.H11MO.0.A	MAZ	ENCFF288RYL
MAZ_HUMAN.H11MO.1.A	MAZ	ENCFF288RYL
MEF2A_HUMAN.H11MO.0.A	MEF2A	ENCFF811FYS
MEF2C_HUMAN.H11MO.0.A	MEF2C	ENCFF138CXP
MXI1_HUMAN.H11MO.0.A	MXI1	ENCFF861YUL
MXI1_HUMAN.H11MO.1.A	MXI1	ENCFF861YUL
MYB_HUMAN.H11MO.0.A	MYB	ENCFF173YZN
MYC_HUMAN.H11MO.0.A	MYC	ENCFF002DAI
NFIC_HUMAN.H11MO.0.A	NFIC	ENCFF269LZJ
NFIC_HUMAN.H11MO.1.A	NFIC	ENCFF269LZJ
NFYA_HUMAN.H11MO.0.A	NFYA	ENCFF414JLN
NFYB_HUMAN.H11MO.0.A	NFYB	ENCFF363BLT
NR2C1_HUMAN.H11MO.0.C	NR2C1	ENCFF538XDH
NR2C2_HUMAN.H11MO.0.B	NR2C2	ENCFF208TMB
NRF1_HUMAN.H11MO.0.A	NRF1	ENCFF931XAL
PAX5_HUMAN.H11MO.0.A	PAX5	ENCFF309VXL
PBX3_HUMAN.H11MO.0.A	PBX3	ENCFF511YXY
PBX3_HUMAN.H11MO.1.A	PBX3	ENCFF511YXY
RELB_HUMAN.H11MO.0.C	RELB	ENCFF739VBA

REST_HUMAN.H11MO.0.A	REST	ENCFF936XYD
RFX5_HUMAN.H11MO.0.A	RFX5	ENCFF968KDX
RFX5_HUMAN.H11MO.1.A	RFX5	ENCFF968KDX
RUNX3_HUMAN.H11MO.0.A	RUNX3	ENCFF147DQK
RXRA_HUMAN.H11MO.0.A	RXRA	ENCFF299YDM
SP1_HUMAN.H11MO.1.A	SP1	ENCFF002CHV
SPI1_HUMAN.H11MO.0.A	SPI1	ENCFF002CHQ
SRF_HUMAN.H11MO.0.A	SRF	ENCFF703TFD
STAT1_HUMAN.H11MO.0.A	STAT1	ENCFF680DVR
STAT1_HUMAN.H11MO.1.A	STAT1	ENCFF680DVR
TAF1_HUMAN.H11MO.0.A	TAF1	ENCFF325FCK
TBX21_HUMAN.H11MO.0.A	TBX21	ENCFF515HWO
TCF7_HUMAN.H11MO.0.A	TCF7	ENCFF817AOQ
TF65_HUMAN.H11MO.0.A	RELA	ENCFF002CPA
USF1_HUMAN.H11MO.0.A	USF1	ENCFF859GUL
USF2_HUMAN.H11MO.0.A	USF2	ENCFF372DRC
ZEB1_HUMAN.H11MO.0.A	ZEB1	ENCFF621OAS
ZN143_HUMAN.H11MO.0.A	ZNF143	ENCFF631JFD

Chapter 4 - Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state

In this chapter I present TRAPeS, software I developed to reconstruct T cell receptor (TCR) sequences from short read single-cell RNA-sequencing. Specifically, the software outputs the CDR3 sequence of each TCR, which is a short DNA sequence that is the main determinant of T cell specificity. In addition, my co-authors and I show a link between CDR3 length and the cell state for Yellow Fever Virus-specific T cells, demonstrating the utility of TRAPeS and the advantages of a combined TCR-transcriptome analysis with single cell RNA-sequencing.

This work was published in *Nucleic Acid Research* in 2017 (Afik et al. 2017), and I am reporting it as it was published. The authors on the paper are:

Shaked Afik[1]*, Kathleen B. Yates[2]*, Kevin Bi[2]*, Samuel Darko[3], Jernej Godec[2,4,5], Ulrike Gerdemann[2], Leo Swadling[6], Daniel C. Douek[3], Paul Klenerman[6,7], Eleanor J. Barnes[6,7], Arlene H. Sharpe[4,5], W. Nicholas Haining[2,8,9,13,‖] and Nir Yosef[10,11,12,13‖]

[1] Computational Biology Graduate Group, UC Berkeley, Berkeley, CA, USA
[2] Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA
[3] Human Immunology Section, Vaccine Research Center, NIAID, NIH, Bethesda, MD, USA
[4] Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, USA
[5] Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA
[6] Translational Gastroenterology Unit, Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK
[7] NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK
[8] Broad Institute of MIT and Harvard, Cambridge, MA, USA
[9] Division of Hematology/Oncology, Children's Hospital, Harvard Medical School, Boston, MA, USA
[10] Department of Electrical Engineering and Computer Science and Center for Computational Biology, UC Berkeley, Berkeley, CA, USA
[11] Ragon Institute of Massachusetts General Hospital, MIT and Harvard, Cambridge, MA, USA
[12] Chan Zuckerberg Biohub Investigator
[13] Corresponding authors. Email: Nicholas_Haining@dfci.harvard.edu, niryosef@berkeley.edu

* These authors contributed equally
‖ These authors contributed equally

Abstract
The T cell compartment must contain diversity in both T cell receptor (TCR) repertoire and cell state to provide effective immunity against pathogens. However, it remains unclear how differences in the TCR contribute to heterogeneity in T cell state. Single cell RNA-sequencing (scRNA-seq) can allow simultaneous measurement of TCR sequence and global transcriptional profile from single cells. However, current methods for TCR inference from scRNA-seq are limited in their sensitivity and require long sequencing reads, thus increasing the cost and decreasing the number of cells that can be feasibly analyzed. Here we present TRAPeS, a publicly available tool that can efficiently extract TCR sequence information from short-read scRNA-seq libraries. We apply it to investigate heterogeneity in the $CD8^+$ T cell response in humans and mice, and show that it is accurate and more sensitive than existing approaches. Coupling TRAPeS with transcriptome analysis of $CD8^+$ T cells specific for a single epitope from Yellow Fever Virus (YFV), we show that the recently described "naive-like" memory population have significantly longer CDR3 regions and greater divergence from germline sequence than do effector-memory phenotype cells. This suggests that TCR usage is associated with the differentiation state of the $CD8^+$ T cell response to YFV.

Introduction
The population of antigen-specific $CD8^+$ T cells formed in response to infection or vaccination is highly heterogeneous in terms of function and phenotype (Appay et al., 2002; Newell et al., 2012). Efforts to deconvolve this cellular heterogeneity have used flow cytometry, mass spectrometry, and more recently, single-cell RNA-sequencing (Chattopadhyay and Roederer, 2015). These approaches have identified a reliable set of phenotypic markers that can classify antigen-specific T cells into a large number of subsets, and distinguish them from antigen-naive T cells. However, recent work also suggests that some antigen-experienced $CD8^+$ T cells can have a naive-like phenotype, meaning that despite their potential to effectively respond to an antigen, they show transcriptomic and surface marker similarities to antigen-naïve T cells (Fuertes Marraco et al., 2015a; Pulko et al., 2016; Swadling et al., 2014). The cellular heterogeneity in the T cell compartment is thought to arise from different exposure to differentiation cues such as antigen dose, duration of contact, and cytokines. How the T cell receptor (TCR) sequence expressed by each T cell contributes to that cellular heterogeneity is not fully understood.

The T cell receptor is a heterodimer of two chains - alpha and beta, each consisting of three types of genomic segments - variable (V), joining (J) and constant (C) (the beta chain includes an additional short diversity (D*)* segment; Methods) (Venturi et al., 2011). The V and J segments are selected out of a pool of several dozen loci encoded in the germline genome, through a recombination process. The diversity of the TCR repertoire (estimated at $\sim 10^7$ in humans (Venturi et al., 2011)) is further enhanced by random insertions and deletions into the complementarity determining region 3 (CDR3) – the junction between the V and J segments, which largely determines the ability of the cell to recognize specific antigens. However despite this diversity, some T cell responses can include TCRs that are identical between individuals - known as "public" clonotypes, while other T cell responses use TCRs that are unique to each individual ("private" clonotypes). Previous studies have shown that these public clonotypes tend

to appear at a higher frequency and have a shorter CDR3 region, possibly as a result of a more efficient recombination process (Robins et al., 2010; Venturi et al., 2006, 2008, 2011).

Unlike analysis of the cell state, the clonal diversity of the TCR repertoire has to date been studied mostly in aggregated samples from pools of T cells rather than individual cells (Ji et al., 2015; Li et al., 2016; Venturi et al., 2011). This approach has two significant limitations: (1) since each chain of the TCR (alpha, beta) is a separate transcript, it cannot determine which chains are co-expressed in the same cell, leading to a partial view of the TCR identity; (2) the sequence of the TCR and the global transcriptional state of the cell that expresses it cannot be simultaneously determined. Previous studies have profiled TCR use in single cells, but these studies were limited in the number of transcripts that were quantified (Han et al., 2014; Ji et al., 2015).

Single cell RNA-seq can generate full-length sequence information for many transcripts in individual cells including the alpha and beta chains of the TCR. However, standard methods to map sequence fragments to the genome (Li and Dewey, 2011) cannot be directly used for reconstructing and estimating the abundance of TCRs because of the highly variable nature of the CDR3 regions. One approach to address this challenge is to rely on scRNA-seq with long sequencing reads (>100 bp), which can cover the entire CDR3 region along with the flanking V and J sub-segments (Stubbington et al., 2016). The underlying TCR (along with the junctional diversification events) can then be reconstructed using methods similar to TCR-seq population repertoire analysis (Venturi et al., 2011; Yu et al., 2014). However, sequencing with long reads is costly and time consuming, thus a method to successfully reconstruct TCRs from shorter, paired-end reads is desirable. Another approach (Eltahla et al., 2016; Redmond et al., 2016; Stubbington et al., 2016) relies on previous methods for *de-novo* transcriptome or genome assembly to reconstruct the CDR3 region (Boetzer and Pirovano, 2012; Grabherr et al., 2011). In general, *de-novo* assemblers were designed with a very large input data set and long reads in mind, and use the concept of de-bruijn graphs to achieve high efficiency. Indeed the TCR reconstruction methods that use this approach have mainly been tested on long RNA-seq libraries (except scTCRseq which was also tested on simulated short reads (Redmond et al., 2016)). However, more accurate yet possibly more computationally intensive algorithms are feasible and may be more appropriate for the smaller target of reconstructing only the TCR.

To address this, we have developed "TCR Reconstruction Algorithm for Paired-End Single cell" (TRAPeS), a software capable of accurately reconstructing TCRs from paired-end sequencing libraries of single cells, even at short (25bp) read length. Unlike the previous methods, TRAPeS does not reduce the input sequences into *k*-mers, but rather works on the original reads - leading to increased sensitivity. We benchmarked TRAPeS on a diverse set of viral stimulations, and then demonstrate how simultaneous analysis of TCR properties and global expression profiling in individual cells helps relate specific TCR properties such as CDR3 length to heterogeneity of T cell state among $CD8^+$ T cells that respond to YFV. TRAPeS is publicly available, and can be readily used to investigate the relationship between the TCR repertoire and cellular phenotype.

Materials and Methods

TRAPeS

The TRAPeS algorithm has 4 main steps, each applied separately to the alpha and beta chains:

1. **Identifying putative pairs of V and J segments**. In order to recognize the V and J segments of the TCR, TRAPeS takes as input the alignment of the RNA-seq reads to the genome. TRAPeS searches for a paired-end read where one mate maps to a V segment while the other mate is mapped to a J segment, and takes those V-J pairs as putative candidates for the CDR3 reconstruction. In a case where there are no such pairings, TRAPeS takes all possible V-J combinations of V and J segments that have V-C and J-C pairing (i.e. reads where one mate maps to V or J and the other mate maps to the C segments). We note that reads are not successfully aligned to D segments of the beta chain due to their short length. Thus, for the beta chain, reconstruction of the CDR3 includes reconstruction of the D segment sequence. In addition, TRAPeS allows the user to specify the maximum number of reconstructions per chain. If the number of possible V-J pairs exceeds this number, TRAPeS ranks the pairs based on the number of reads initially mapped to them, and only attempts to reconstruct the top pairs.

2. **Collecting putative CDR3-originating reads**. TRAPeS finds the putative CDR3-originating reads by taking all the unmapped reads whose mates map to the V/J/C segments. In addition, since the first step of the CDR3 reconstruction includes alignment to the genomic V/J sequences (see below), TRAPeS also collects the reads that map to the V and J segments.

3. **Reconstructing the CDR3**. Using an iterative dynamic programming algorithm, TRAPeS extends the V and J regions. TRAPeS takes only the bases at the ends of the V and J segments closest to the CDR3 (3' of the V segment and 5' of the J segment). The number of initial bases is a parameter that can be tuned, set by default to min(length(V), length(J)). If the specified length is longer than the J segment, TRAPeS concatenates the J sequence to the beginning of the C sequence and uses this extended segment as the initial J segment. In each iteration, we align all the reads to the V and J segments separately with the Needleman-Wunsch algorithm, using the following scoring scheme: +1 for a match, -1 for a mismatch, -20 for gap opening and -4 for gap extension. In addition, we don't penalize for having the read "flank" the V and J toward their 3' and 5', respectively.
Next, we take all the reads that aligned to the V and J segments above a certain score threshold, and build the "extended" V and J sequences based on the reads. For each position, we take the base that appears in most reads as the chosen sequence for this position. This way, we extend the V and J regions in each iteration and also correct for mutations or SNPs in the known genomic V and J segments. TRAPeS repeats this step until the extended V and extended J overlap, or until TRAPeS reaches a number of predefined iterations. If no overlap is found, TRAPeS also offers an optional "one-sided" mode, where it will attempt to determine the productivity (see below) of only the extended V segment. For this work, we used a threshold score of 21 for the alignment of the reads. However, in some cases a lower threshold was required, thus if no sequence was reconstructed we ran TRAPeS with a scoring threshold of 15.

4. **Separating similar TCRs and determining chain productivity**. Since some V and J segments have similar sequence, reads can be mapped to several segments, creating few similar putative V-J pairs. In addition, two alpha or beta chains can be created within a single cell. TRAPeS takes all possible pairing and attempts to reconstruct the CDR3 region for all pairs. After reconstruction, full-length TCR sequences are created by extending the reconstructed region with the known reference sequences. Then, TRAPeS runs RSEM (Li and Dewey, 2011) on all reconstructed TCRs and the set of reads used as input (and their mates) in order to rank the TCRs based on the relative abundance. Next, TRAPeS determines if the TCR is productive: V and J segments are in the same reading frame and the CDR3 does not contain a stop codon. TRAPeS outputs a file with a summary of all possible reconstructions (see Table S1 for example) for all cells, as well as separate files for each cell with the full-length TCR sequences. For this paper we used the productive chain with the highest expression as the TCR sequence for each cell.

TRAPeS is implemented in python. To increase performance, the CDR3 reconstruction using the dynamic programming algorithm is implemented in C++, and uses the seqan package (Döring et al., 2008). TRAPeS is freely available and can be downloaded in the following link: https://github.com/YosefLab/TRAPeS

TRAPeS can be easily extended to work with single-end data. The reconstruction algorithm only requires the paired-end information for the recognition of V/J segments and CDR3-originating reads, which can be easily done in single-end reads by searching for partial alignment of the read edges to the V/J segments. This feature will be available in the next TRAPeS version.

Single cell sorting

Mouse LCMV Experiments: Female C57BL/6 mice (The Jackson Laboratory), aged 7 weeks, were infected with $2x10^5$ plaque forming units (PFU) LCMV Armstrong intraperitoneally i.p. or $4x10^6$ PFU LCMV Clone 13 i.v. LCMV viruses were a generous gift from Dr. E John Wherry (University of Pennsylvania, Perelman School of Medicine). Peripheral blood was obtained from the mice at day 7 post infection (p.i.) and lymphocytes were enriched using LSM density centrifugation. Cells were prestained with a near-IR fixable live/dead marker (Life Technologies, cat# L34976) and an APC-conjugated dextramer reagent for gp33 (Immudex, cat# A2160-APC) according to manufacturer recommendations. The cells were then stained with the following antibodies: FITC 2B4 (BioLegend, cat# 133504), PerCP-Cy5.5 CD44 (BioLegend, cat# 103032), PE KLRG1 (BioLegend, cat# 138408), PE-Cy7 PD1 (BioLegend, cat# 135215), BV421 CD127 (BioLegend, cat# 135024), BV510 CD8A (BioLegend, cat# 100752).

Human CMV Experiments (Donor 1): Blood samples were obtained from a donor with detectable NLV-specific $CD8^+$ T cell response. Lymphocytes were enriched via Ficoll gradient and prestained with a near-IR fixable live/dead marker (Life Technologies, cat# L34976) and an APC-conjugated dextramer reagent (Immudex, cat# WB2132-APC). The cells were then stained with the following antibodies: FITC CD8A (BioLegend, cat# 300906), PerCP-Cy5.5 CCR7 (BioLegend, cat# 353220), PE CD3 (BioLegend, cat# 317308), BV605 CD45RA (BioLegend, cat# 304133).

Human YFV Experiment (Donor 2): A healthy volunteer was vaccinated with a single dose (0.5 ml containing at least 10^5 PFU) of 17D live-attenuated yellow fever vaccine strain administered subcutaneously. Seroconversion after vaccination was confirmed by assaying the neutralizing antibody titers for YF-17D (data not shown). A whole blood sample was obtained 9 months post-vaccination and lymphocytes were enriched from whole blood via Ficoll gradient centrifugation and a CD8 negative selection magnetic bead kit (Miltenyi Biotec). Cells were prestained with a live/dead marker (Life Technologies, cat# L34976) and an APC-labeled tetramer reagent (NS4B 214–222 LLWNGPMAV, kindly provided by Dr. Rafi Ahmed). The cells were then stained with the following antibodies: FITC CD8A (BioLegend, cat# 300906), PE CXCR3 (BioLegend, cat# 353705), PE-Cy7 CCR7 (BioLegend, cat# 353226), BV421 IL2Rb (BioLegend, cat# 339009), BV510 CD3 (BioLegend, cat# 317332), BV605 CD95 (BioLegend, cat# 305627), BV780 CD45RA (BioLegend, cat# 304140).

Human Hepatitis C Experiment (Donor 3): Patient 355 (59yr old Male, infected with genotype 1a HCV, baseline viral load 467,000 IU/ml) received a prime vaccination of ChAd3-NSmut ($2.5x10^{10}$ viral particles) and an MVA-NSmut ($2x10^8$ plaque forming units) boost vaccination 8 weeks later. PBMC were collected 14 weeks post-boost vaccination for assessment of single cell gene expression (Swadling et al., 2016). PBMC were thawed and prestained with a live/dead marker (Life Technologies, cat# L34976) and a PE-conjugated pentamer reagent (PE-labeled HCV NS31406–1415 (KLSALGINAV; HLA-A*0201)). The cells were then stained with the following antibodies: FITC 2B4 (BioLegend, cat# 329505), PerCP-eFluor 710 LAG3 (eBioscience, cat# 46-2239), PE-Cy7 CCR7 (BioLegend, cat# 329919), APC CD39 (BioLegend, cat# 328209), BV421 PD1 (BioLegend, cat# 329919), BV510 CD3 (BioLegend, cat# 317332), BV605 CD8A (BioLegend, cat# 301040), BV780 CD45RA (BioLegend, cat# 304140).

The relevant institutional review boards approved all human subject protocols, and all subjects provided written consent before enrollment.

Single cell sorts: All single cell sorts were performed on a BD Aria II with a 70um nozzle. Cells were sorted into 5µL of Qiagen TCL Buffer plus 1% beta-mercaptoethanol v/v. Immediately following sorting, plates were sealed, vortexed on high for 30 seconds, and spun at 400g for 1 minute prior to flash freezing on dry ice. Samples were stored at -80°C until library preparation.

RNA sequencing

Single cell lysates were converted to cDNA following capture with Agencourt RNA Clean beads using the SmartSeq2 protocol as previously described (Trombetta et al., 2014). The cDNA was amplified using 22-24 PCR enrichment cycles prior to quantification and dual-index barcoding with the Illumina Nextera XT kit. The libraries were enriched with 12 cycles of PCR, then combined in equal volumes prior to final bead cleanup and sequencing. All libraries were sequenced on an Illumina HiSeq 2500 or NextSeq by either single-end 150bp reads or short paired-end reads using the following read lengths: Mouse samples - 30bp, Human donor 1 - 26bp for read 1 and 25bp for read 2, Human donor 2 - 30bp, human donor 3 - 26bp. Donor 1 and Donor 2 were sequenced using two batches, where every batch had cells from all of the donor's population (i.e. Donor 1 batch 1 had both naive and CMV-specific cells, same for batch 2. Donor

2 batch 1 had YFV-specific, naive and effector memory cells, same for batch 2). Donor 3's entire sample was sequenced on a single batch, and the LCMV samples from both mice were combined and sequenced on a single batch (Table S2).

Preprocessing and Normalization of scRNA-seq data

Low quality bases were trimmed with trimmomatic (Bolger et al., 2014) using the following parameters: LEADING:15, TRAILING:15, SLIDINGWINDOW:4:15, MINLEN:16. Trimmed reads were then aligned to the genome (hg38 or mm10 for human or mouse samples, respectively) with TopHat2 (Kim et al., 2013) for TCR reconstruction, and aligned to the transcriptome with RSEM (Li and Dewey, 2011) for transcriptome quantification.

For transcriptome analysis of the human CMV and YFV donors (donors 1 and 2), low quality cells were filtered out prior to normalization. Cells were filtered out if their read depth was less than 1 million pairs or if the cell expressed less than 20% of all expressed transcript, where a transcript was considered expressed if it had a Transcripts Per Million (TPM) value of >10 in at least 10% of cells, leaving 353 out of 378 cells for further analysis.

Normalization of TPM values was done with our newly developed normalization framework SCONE (https://niryosef.wordpress.com/tools/scone/). SCONE considers a large number of unsupervised normalization pipelines (i.e. without using any prior biological information about samples' origin), applying different ways to scale the data (e.g., full quantile, upper quantile) and perform factor analysis to eliminate unwanted variation. SCONE then uses a number of quality metrics to choose the best normalization, which reduces technical variation and maintains prior biological knowledge. In our study, the chosen normalization first scaled each sample with the DEseq (Anders and Huber, 2010) scaling factor to account for differences in sequencing depth. Then, we ran RUVg (Risso et al., 2014) with k=1. In order to run RUVg, a list of genes that are constant across conditions should be provided. To find constant genes across the specific conditions that were tested in this paper, we also sequenced bulk populations of naive $CD8^+$ T cells from donor 1 and CMV-specific effector memory $CD8^+$ T cells, as well as populations of 50 cells of naive $CD8^+$ T cells from donor 2 and YFV-specific effector memory $CD8^+$ T cells. We ran DESeq2 (Love et al., 2014) on those samples and defined the set of constant genes as the genes that showed no change (FDR-adjusted p-value > 0.98 and absolute log fold change < 0.2) across all pairwise comparisons (naive vs. all effector memory cells, naive vs. CMV-specific effector memory, naive vs. YFV-specific effector memory and CMV-specific effector memory vs. YFV-specific effector memory), resulting in a total of 373 genes.

Dimensionality reduction with PCA on samples from each donor after normalization revealed that the normalization process maintain biological information, while reducing the correlation between the data and technical variables such as batch, number of expressed genes in each cell, and the values of the first PC of the quality matrix (where the quality matrix includes for each cell technical information as previously described (Gaublomme et al., 2015) (Figure S10)).

Reconstructing TCR sequence from long reads

Detection of CDR3 sequence using long (150bp) reads was performed similar to Venturi et al (Venturi et al., 2011). In short, reads were aligned against the set of known V and J segments using blastn (Altschul et al., 1990). Reads with V and J segments aligning to their edges were selected, extracting the CDR3 sequence in each read. In case where more than one productive

CDR3 sequence was discovered in a cell, the sequence with the highest number of supporting reads was selected.

Reconstructing TCR sequence from short paired-end reads using Trinity

Trinity (Grabherr et al., 2011) was run on each cell with the following parameters: --max_memory 10G, --min_contig_length 50. In addition, using the --KMER_SIZE parameter Trinity was run with 4 different k-mer sizes - 13, 15, 17 and 19. For each k-mer size we ran Trinity twice: once in single-end mode, using the set of reads used by TRAPeS for CDR3 reconstruction, and once in paired-end mode, taking all the mapped and unmapped reads along with their pairs. Then, for each k-mer we combined the final Trinity output from both runs (paired-end and single-end) for each cell. To determine whether or not a transcript is productive and to annotate the CDR3 sequence, all possible reconstructed transcripts were run through IMGT/HighV-QUEST (Alamyar et al., 2012; Li et al., 2013). We considered each productive chain output by IMGT as a successful reconstruction.

Comparing TRAPeS to TraCeR

TraCeR was run using default parameters. To compare TRAPeS to TraCeR on the benchmark data used by TraCeR (Stubbington et al., 2016), raw single cell RNA-seq data was downloaded as fastq files from ArrayExpress (accession number E-MTAB-3857). While the original data consisted of 100bp paired-end reads, we converted it to that equivalent of short-read sequences by trimming each fragment to leave only the outer 25 or 30bp of each read. We also ran TRAPeS on the original 100bp paired-end data with the following parameters: -score 80 -bases 150 -top 15 -byExp -oneSide

Comparing TRAPeS to scTCRseq and VDJPuzzle

VDJPuzzle was run using the default parameters. For scTCRseq, since running the software with the default parameters resulted in no alignments for human TRBV segments, we ran the software using the parameters -e 1e-7 -c 2. In addition, since scTCRseq does not summarize the data, we collected the fasta sequences of scTCRseq final results (*.gapfilled.final.vdj.fa files) and ran them through IMGT to annotate the junction sequence in each cell, taking only productive CDR3 with a complete reconstruction (no missing amino acids) as successful reconstructions. To compare TRAPeS and scTCRseq on the benchmark data used by scTCRseq (Mahata et al., 2014), raw single cell RNA-seq data was downloaded as fastq files from ArrayExpress (accession number E-MTAB-2512) and trimmed from 75bp paired-end into 25 or 30bp paired-end. We also ran TRAPeS on the original 75bp paired-end data with the following parameters: -score 65 -top 10 -bases 100

Gini coefficient calculation:

For each population, cells were considered from the same clone if they had identical CDR3 sequences of both alpha and beta chains. Cells with only one reconstructed chain were excluded from this analysis. The number of cells for each clone was counted and the Gini coefficient was calculated by using the Gini command in R from the “ineq” package.

Inference of cell clusters, visualization and differential expression analysis

For cluster inference in the YFV + CMV human data, we defined an expression matrix consisting of normalized TPM values of 353 cells by 10827 transcripts (expressed at a level of >= 5 TPM in at least 1% of cells; Table S11). We applied the SC3 software (Kiselev et al., 2017) for clustering the cells in this matrix using default parameters.

To visualize the data, we first used the jackStraw package (Chung and Storey, 2015) to reduce the dimensionality of the data and retain only principal components (PC) that are statistically significant (p-value<10^{-4}) in terms of the respective percent of explained variance. This analysis retained the first three PCs. We then applied t-SNE (Maaten and Hinton, 2008) with default parameters and 2000 iterations to these significant PCs, further reducing the data for visualization in two dimensions.

We used the DESeq2 package (Love et al., 2014) to identify genes that are differentially expressed (DE) between the different clusters. In this application, each cluster was compared to the other two clusters, looking for genes that are differentially expressed. Genes were called as differentially expressed using an FDR-adjusted p-value cutoff of 0.05. The heatmap in Figure 3B was populated with $\log_2$(TPM) values for genes identified as uniquely up- or down- regulated in each of three major phenotypic groups. We also see similar results of DE genes using the scRNA-seq analysis package Seurat (McDavid et al., 2013; Satija et al., 2015). Enrichment of DE genes with respect to immunological pathways was determined using a Fisher exact test (FDR-adjusted p-value<10^{-3}) quantifying the significance of overlap between differential genes and signatures from the ImmuneSigDB database (Godec et al., 2016).

Gene enrichment by signature analysis

We used FastProject (DeTomaso and Yosef, 2016) together with large collection of transcriptional signatures from ImmuneSigDB (Godec et al., 2016) to characterize the phenotype of our single cells. In short, each transcriptional signature is comprised of genes that are either over-expressed or under-expressed between two cell states of interest (e.g. using published bulk RNA-seq data from naive vs. memory cells). For each single cell, the signature score is computed as:

$$R_S(j) = \mathbf{\Sigma} \text{sign}_s(i) \cdot X'_{ij} \cdot \omega_{ij} / \mathbf{\Sigma}_{i \in S} \omega_{ij}$$

Where s is the signature, j is the cell, sign(i) = -1 for genes under-expressed in this signature and +1 for over-expressed genes, X'_{ij} is the standardized (Z-normalized across all cells) log expression level of gene i in cell j, and ω_{ij} is the estimated false-negative weight for gene i in cell j. To identify transcriptional signatures that are associated with an scRNA-seq data set of interest, FastProject looks for consistency between signatures and low-dimensional projections of the data. To this end, FastProject first computes a wide range of 2-dimensional projections (e.g. PCA, ICA, spectral embedding, tSNE), each capturing (possibly different) key axes of variation in the data. For each transcriptional signature and each projection it then computes a consistency score, which reflects the extent to which cells that have a similar signature score reside close to each other in the projection (thus extending our previous work (Gaublomme et al., 2015) and facilitating the analysis of non-linear projections). The significance of the consistency score is evaluated by random shuffling.

To include only relevant signatures, we analyzed only signatures with a significant consistency score (FDR-adjusted p-value<0.05) in at least one projection. In addition, only signatures that

include ‘CD8’ in their name were used for further analysis, leaving a total of 95 signatures for the YFV + CMV human data and 154 signatures for the YFV-specific analysis.

Characterization of TCR properties of YFV-specific cells

TCR expression

To compute the expression of each reconstructed TCR, we added the reconstructed sequences to the transcriptome and ran RSEM on the complete extended transcriptome, using the original sequencing results (the complete fasta files) as input. This was performed for each cell separately, i.e. for each cell only its TCR sequences were added to the transcriptome. In cases where a cell had more than one reconstructed alpha or beta chain (by having two productive chains or having one productive and one unproductive chain) they were both added to the transcriptome.

Germline score

Classification of each base in the CDR3 as germline (originating from the V, D, J regions) or added nucleotide was done by running the reconstructed TCR sequences thorough IMGT/V-Quest (Brochet et al., 2008; Giudicelli et al., 2011). The germline score was calculated by dividing the number of nucleotides encoded by V, D, J segments by the length of the CDR3 (Yu et al., 2014).

Comparing transcriptomic signatures with TCR length

Identification of gene signatures associated with TCR length was done with the PARIS algorithm (Cowley et al., 2014), a module in GenePattern (Reich et al., 2006). PARIS describes the association between each signature score and TCR length by estimating their differential mutual information. For each signature, the mutual information is computed between the TCR length and the signature, and then normalized using the joint entropy. This score is rescaled with the mean of the score of the TCR length against itself and the score of the signature against itself, resulting in a rescaled normalized mutual information (RNMI) matching score. The significance of the score is evaluated by a permutation test (performed on the TCR length) and then FDR correction.

Hydrophobicity

The mean hydrophobicity of each CDR3 was computed using the Kyte-Doolittle (Kyte and Doolittle, 1982) numeric hydrophobicity scale. In order to account for CDR3 length, we also computed mean hydrophobicity for each CDR3 using a sliding window (of both size 3 and 5), taking the mean across all windows. However, the sliding window also didn’t result in significant differences between YFV-specific naive-like and YFV-specific effector memory-like cells (K-S test p-value > 0.1, data not shown).

Normalized tetramer binding intensity

Normalized tetramer binding intensity was defined based on flow cytometry data acquired at the time of sorting. The tetramer binding was measured with the APC-labeled tetramer reagent. To correct for baseline expression of CD3, we divided the APC-labeled tetramer measurement by the expression of CD3 surface molecules.

Results

TRAPeS reconstructs TCR sequences using short (25-30bp) scRNA-seq

TRAPeS starts by recognizing putative pairs of V and J segments that flank the CDR3 region, using genome alignment (Trapnell et al., 2009) (Figure 1, top; see Methods for a complete description of the algorithm). It then identifies the set of unaligned reads that may have originated from the CDR3 region, taking the unmapped mates of reads aligned to the putative V-J segments or to the constant (C) segment (Figure 1, middle). Next, it uses an iterative dynamic programming scheme to piece together the putative CDR3 reads, gradually extending the CDR3 reconstruction on both ends (V and J) until convergence (Figure 1, bottom). Finally, after the TCR chain has been reconstructed, TRAPeS determines whether it is productive (i.e., has an in-frame CDR3 without a stop codon) and determines its exact CDR3 sequence, based on the criteria established by the international ImMunoGeneTics information system (IMGT) (Lefranc et al., 2005). For each cell, TRAPeS outputs a set of reconstructed TCR transcripts (from both chains), along with their complete sequence, an indication of whether or not they are productive, and the number of reads mapped to them. In some cases multiple reconstructions can be generated for the same cell. This may happen when more than one chain is produced in the cell (a phenomenon that have been previously reported (Eltahla et al., 2016; Redmond et al., 2016; Stubbington et al., 2016)), or when sequence similarity between some V or J segments results in several possible V-J pairs with an identical CDR3 reconstruction. In such cases, we report all V-J pairs, while ranking the putative TCR transcripts in accordance to their estimated expression levels (Table S1). The average running time of TRAPeS on a Human single cell library with an average two million reads per cell is less than two minutes per cell on a standard machine (Figure S1).

TRAPeS is accurate and more sensitive than previous methods using short reads and comparable to previous methods using long reads

We applied and tested TRAPeS to scRNA-seq data from a range of $CD8^+$ T cell responses (Methods, Figure 2A). These data sets were selected to include both mouse and human $CD8^+$ T cells as well as those expected to have a range of TCR complexities (Figure S2). In mice, we used the lymphocytic choriomeningitis virus (LCMV) infection model, and profiled $CD8^+$ T cells responding to either acute or chronic infection (using the Armstrong and Clone 13 strains of LCMV, respectively). In healthy human subjects we profiled naive $CD8^+$ T cells, effector memory $CD8^+$ T cells, and antigen-specific $CD8^+$ T cells elicited by CMV infection; vaccination with the live attenuated yellow fever virus infection (YFV-17D) (Akondy et al., 2015); or by vaccination with adenoviral and modified vaccinia Ankara vectors encoding HCV proteins (Swadling et al., 2014, 2016). We sorted up to 128 single $CD8^+$ T cells from each dataset to a total of 565 cells, and generated scRNA-seq libraries with short (25-30bp) paired-end reads as previously described (Picelli et al., 2014; Trombetta et al., 2014) and observed good quality metrics using previously used measures (Gaublomme et al., 2015) (Table S2, Methods). To test TRAPeS, we applied cell quality filtering scheme similar to the criteria used by others (Stubbington et al., 2016), removing samples with less than 2000 genes or with more than 10% of reads mapping to mitochondrial genes, resulting in a total of 513 high quality cells (Figure 2A). Importantly, our results below remain consistent also when cell filtering is not applied (Figure S3).

To evaluate the accuracy of TRAPeS, we compared its output with that of directly sequencing the TCR sequence using long reads (in which reconstruction is not required, Methods). To that end, we sequenced libraries of epitope-specific cells for Clone 13, Armstrong and CMV, and naive T cells from the CMV donor with both short (25-30bp) paired-end and 150bp single-end sequence reads (Figure 2A). TCR sequences identified by TRAPeS were almost perfectly consistent with those produced based on the long read data (Methods; Figure 2B-C), indicating a high level of specificity.

We compared TRAPeS to previously published methods for TCR reconstruction in single cells. First, we compared TRAPeS to TraCeR (Stubbington et al., 2016) - a TCR reconstruction software that is built upon Trinity (Grabherr et al., 2011), a *de-novo* transcriptome assembly tool. We found that the sensitivity of TRAPeS was markedly higher (Figure 2A-C). On average (across all data sets), TRAPeS successfully reconstructed productive alpha chains from 66% of the cells and productive beta chains from 80% of the cells, using the short (25-30bp) libraries. In contrast, TraCeR resulted in no reconstruction for the 25bp paired-end libraries, and was able, for the 30bp libraries, to reconstruct CDR3 regions in an average of 43% and 15% of the cells for alpha and beta chains respectively.

Next, we considered two additional recently published methods - VDJPuzzle (Eltahla et al., 2016) and scTCRseq (Redmond et al., 2016), both based on *de-novo* assembly algorithms (Trinity and GapFiller (Boetzer and Pirovano, 2012), respectively). As above, we observe substantially higher sensitivity with TRAPeS (Figure 2A-C, Methods). VDJPuzzle was also unable to reconstruct any productive chains in the 25bp data and, for the 30bp libraries, reconstructed CDR3 regions in an average of 40% and 63% of the cells for alpha and beta chains, respectively. scTCRseq, which is built upon GapFiller (Boetzer and Pirovano, 2012), managed to successfully reconstruct CDR3 regions in an average of 50% and 60% of the cells for alpha and beta chains, respectively. While scTCRseq achieves better results compared with Trinity-based methods, TRAPeS clearly outperforms all methods in terms of specificity and sensitivity (Figure 2A-C).

The low success rate of Trinity-based methods TraCeR and VDJPuzzle is likely due to its requirement for seed *k-mer* length (25nt) that is unsuitable for short reads. Thus, we also directly ran Trinity on our set of CDR3-originating reads, using a *k-mer* length of 13 (Methods). This resulted in an increased sensitivity for the 30bp libraries compared to TraCeR and VDJPuzzle, but did not improve the reconstruction rates for 25bp libraries (Figure 2A-C). Running Trinity with several other *k-mer* lengths (15, 17 and 19) did not significantly change the results (Figure S4).

Notably, the average rate of successful reconstruction of TRAPeS in our mouse libraries is 93.7% (with 30bp reads), which is higher than that achieved by TraCeR with the mouse libraries used by Stubbington et al (86.3% with 100bp reads) (Stubbington et al., 2016). To further substantiate this result, we applied TRAPeS and TraCeR on a trimmed version of this published data. We found that discarding 70-75% of the information (i.e., taking only 25 or 30bp out of each 100bp read) substantially hurts the performance of TraCeR, while TRAPeS is able to

maintain rates of successful TCR reconstructions that are similar to those achieved in the original paper (Stubbington et al., 2016) (Figure S5). Running TRAPeS on the original long read data is also comparable to the success rates obtained by TraCeR, demonstrating the ability of TRAPeS to be applied on long reads as well (Figure S5). In addition, running TRAPeS on short or long reads is comparable to running scTCRseq using long reads, as evident by running TRAPeS on the original and a trimmed version of the data used to benchmark scTCRseq (Mahata et al., 2014; Redmond et al., 2016) (Figure S6).

TRAPeS captures various clonality levels

We investigated the clonality of the TCR repertoire measured by TRAPeS among the human $CD8^+$ T cells (Figure 2D, Table S3), using the Gini Index, a clonality measure (Qi et al., 2014) ranging from zero (i.e. no two cells share the same TCR) to one (i.e. all cells are from the same clone; Methods). As expected, the naïve population had a Gini index of zero, indicating that each naive $CD8^+$ T cell expressed a unique TCR. The CMV-specific $CD8^+$ T cell population had a high Gini index (with 83% of CMV-specific $CD8^+$ T cells with reconstructed alpha and beta chains originated from a single clone), indicating a high degree of oligoclonality as previously described (Trautmann et al., 2005; Weekes et al., 1999). In contrast, $CD8^+$ T cells elicited by YFV or HCV vaccines showed much greater heterogeneity in TCR repertoire, consistent with a more limited, rather than persistent, exposure to antigen (Barnes et al., 2012; Bolinger et al., 2015; DeWitt et al., 2015; Miles et al., 2011; Swadling et al., 2014). This demonstrates the ability of TRAPeS to capture cells from the same clone even with relatively small number of antigen-specific cells, assuming a clonal response.

Single-cell transcriptome analysis detects subpopulations of YFV cells

In order to determine the relationship between TCR use and $CD8^+$ T cell state, we focused on $CD8^+$ T cells from two healthy donors (YFV and CMV peptide-specific, as well as naive and effector memory cells without sorting for peptide specificity; Methods) to avoid introducing additional complexity from chronic infection. To identify groups of cells with similar expression profiles, we used SC3 (Kiselev et al., 2017), a robust clustering method for sparse datasets, to identify subpopulations of cells (Figure S7, Table S4, Methods) which we then visualized using t-SNE (Maaten and Hinton, 2008) (Figure 3A). We found three clusters of cells: one that contained all CMV-specific cells (Figure 3A, purple symbols); one that contained all effector memory cells (blue symbols); and one that contained all naive $CD8^+$ T cells (green symbols). In contrast to these discrete groupings, we observed that YFV-specific $CD8^+$ T cells were split between two clusters: one containing effector memory $CD8^+$ T cells and one containing naive $CD8^+$ T cells.

Differential gene expression analysis between cell clusters revealed transcripts consistent with the known patterns of gene expression in antigen-experienced or naive $CD8^+$ T cells (Tables S5-S7, Figure 3B, Methods). CMV-specific $CD8^+$ T cells expressed effector molecules and transcription factors characteristic of antigen experienced cells (e.g., Granzyme B, *PRDM1*), which were not detected in naive cells. Naive $CD8^+$ T cells expressed canonical markers of the naive state (*CCR7*, *SATB1*, *LEF1*) that were absent in CMV-specific and effector memory $CD8^+$ T cells. The expression of these genes in YFV-specific $CD8^+$ T cells was consistent with the

cluster in which they were associated, with those in the naive cluster expressing minimal Granzyme B or *PRDM1*, but showing robust expression of *CCR7*, *SATB1*, and *LEF1* (Figure S8).

To identify broader patterns of transcriptional signatures, we applied FastProject (DeTomaso and Yosef, 2016) - a software tool that enables the expression of gene sets of interest to be quantified in transcriptional profiles of single cells (Methods). We surveyed the enrichment of a collection of gene sets, from the C7 (ImmuneSigDB) (Godec et al., 2016) collection of MSigDB (Liberzon et al., 2011) corresponding to cell states and perturbations of $CD8^+$ T cells. We found significant up-regulation of multiple gene sets corresponding to naive $CD8^+$ T cells (K-S test FDR-adjusted p-value<0.01) in the naive cluster (cluster 3) compared to the other two clusters. Consistent with this, we found significantly greater up-regulation of effector signatures in clusters 1 and 2 compared with the other clusters (FDR-adjusted p-value<0.01; Figure 3C and Table S8).

To confirm these patterns of transcript abundance at the protein level, we compared flow cytometry data for a set of surface markers acquired at the time of sorting (Methods) with transcript abundance in the same cell (Figure 3D). Consistent with the gene expression profiles, we observed that YFV-specific $CD8^+$ T cells in the naive-like cluster (open symbols) showed higher protein levels of CCR7 and CD45RA than those in the effector memory cluster (purple symbols). Thus, single-cell analysis shows that $CD8^+$ T cells specific for the same peptide epitope from YFV are heterogeneous and includes both effector-memory and naive-like gene expression profiles, as has been reported previously for cells analyzed at the bulk level (Fuertes Marraco et al., 2015a, 2015b; Pulko et al., 2016).

Combined TCR-transcriptome analysis reveals longer CDR3 regions for naive-like YFV-specific cells

We reasoned that differences in TCR might contribute to the heterogeneous differentiation of $CD8^+$ T cells following YFV vaccination. To that end, we evaluated a number of properties to characterize each reconstructed TCR - CDR3 specific properties such as length, hydrophobicity and germline score as well as TCR expression. In addition, we measured the normalized tetramer staining intensity per cell (Table S9, Methods). We then asked whether any of those properties differed between naive-like and effector memory-like YFV-specific $CD8^+$ T cells. Naive-like and effector memory-like YFV-specific $CD8^+$ T cells were indistinguishable (p-value>0.05, FDR-adjusted p-value>0.1) in terms of TCR transcript expression, hydrophobicity of the CDR3 region and normalized tetramer staining intensity (Methods). However, we found that the CDR3 sequence was significantly longer in YFV-specific $CD8^+$ T cells with a naive-like state compared with those with an effector memory profile for both alpha and beta chains (Figure 4A, K-S test p-value 0.038 and 0.027 for alpha and beta chains respectively, FDR-adjusted p-value 0.084 for both alpha and beta chains).

We next evaluated the germline score of CDR3 regions in YFV-specific $CD8^+$ T cells, a measure of the contribution of germline nucleotides to the CDR3 region. The germline score is defined as the ratio between the number of nucleotides in the CDR3 that originate from the germline (V, D, J segments) to the total number of nucleotides in the CDR3 (Yu et al., 2014) (Methods). Consistent with the differences in the CDR3 length, we found that naive-like YFV-specific $CD8^+$

T cells had a significantly lower germline score in both alpha and beta chains than did effector memory-like cells (Figure 4B, K-S test p-value of 0.034 and 0.029 for alpha and beta chains respectively, FDR-adjusted p-value 0.084 for both alpha and beta chains), suggesting that generating the CDR3 region of these TCRs involved a greater degree of nucleotide addition/subtraction.

To further characterize the relationship between CDR3 length and cellular state in YFV-specific $CD8^+$ T cells, we identified $CD8^+$ transcriptional signatures (extracted from ImmuneSigDB (Godec et al., 2016) and scored with FastProject (DeTomaso and Yosef, 2016), as above) that correlated with CDR3 length across all YFV-specific $CD8^+$ T cells (Table S10, Methods). Of all signatures evaluated, we found that only naive $CD8^+$ T cell signatures showed a significant positive correlation with CDR3 length (FDR-adjusted p-value<0.1; Figures 4C-D). Previous work has suggested that YFV-specific $CD8^+$ T cells with a naive-like phenotype include those with a stem-cell memory (Tstem-memory) differentiation state. We found that Tstem-memory signatures were more enriched in naive-like YFV-specific $CD8^+$ T cells than in effector memory YFV-specific $CD8^+$ T cells (Figure S9). However, the enrichment for these signatures was equivalent between naive-like YFV-specific and phenotypically naive $CD8^+$ T cells, making it difficult to discern whether these cells manifest a specific stem-cell-like state. Our results, however, show that heterogeneity in the differentiation state of $CD8^+$ T cells responding to a single epitope of YFV is strongly associated with the CDR3 length.

Discussion

TRAPeS enables the analysis of TCR clonality in scRNA-seq profiles using short sequence reads. Other methods of direct TCR sequencing (Venturi et al., 2011) or reconstruction (Eltahla et al., 2016; Redmond et al., 2016; Stubbington et al., 2016) have lower rate of successful TCR reconstruction or requires long sequence reads, which substantially increase the per-cell cost of single cell profiling. As single-cell RNA-seq technologies move towards massively parallel scale, long-read sequencing is likely to become unfeasibly expensive, making approaches such as TRAPeS critical for studies of TCR use in single cells.

We applied TRAPeS to short-read sequencing data from human $CD8^+$ T cells and were able to discover a new association between the differentiation state of $CD8^+$ T cells specific to a single YFV antigen and the CDR3 length of the TCRs that they express. Long CDR3 lengths have been associated with private clonotypes, which in turn may reflect low precursor frequency within the naive T cell pool (Robins et al., 2010; Venturi et al., 2006, 2008, 2011). We therefore speculate that within a population of naive T cells capable of recognizing a specific antigen, those that exist at low frequency may enter the T cell response later than more abundant precursors, resulting in an altered differentiation state compared to those that existed at a higher precursor frequency. Alternatively, a greater degree of cross-reactivity in T cells with short CDR3 regions may result in more repeated TCR stimulation, leading to the difference in T cell phenotype we observe. While in this case the phenotype could be validated with protein surface markers, this is

not true for many other phenotypes, highlighting the importance of transcriptome analysis using scRNA-seq.

More generally, we anticipate that TRAPeS will facilitate broad efforts to determine the relationship between T cell state and TCR sequence in the immune response. TRAPeS can be applied to further basic biological understanding of the relationship between TCR avidity and T cell differentiation. Being able to identify alpha and beta chains allows cloning of TCRs into experimental systems to study their binding properties, which will help determine how TCR properties are related to TCR avidity and T cell biology. This is highly relevant for studying vaccine responses and for thymic development. Moreover, linking the CDR3 sequence to T cell transcriptome can help identify biological similarities in clonal populations of T cells. For instance, in tumors where the identities of T cells responding to the tumors are not known, identifying clonal expansion can be used to infer tumor-specificities both for analyzing gene expression profiles and cloning both alpha and beta chains of the same TCR for clinical use. Additionally, we recently applied TRAPeS to study the clonality of $CD4^+$ and HLA class II-restricted $CD8^+$ T cells in HIV-infected individuals (Ranasinghe et al., 2016), demonstrating the wide use for a combined analysis of transcriptome and TCR sequence at the same cell.

Availability
TRPAeS is publicly available and can be found in the following link:
https://github.com/YosefLab/TRAPeS.

Accession Number
scRNA-seq data have been deposited in the Gene Expression Omnibus (GEO) under accession number GSE96993.

Funding
L.S and E.B are funded by the Medical Research Council UK (L.S as an MRC CASE studentship). N.Y was supported by the National Institute of Health [grant number 5U19AI090023-07]. This work was supported by US National Institute of Health [grant numbers AI090023, AI057266 and AI082630].

Acknowledgments
The authors would like to thank Rama Akondy and Rafi Ahmed for providing the YFV vaccine samples and related reagents; members of the Haining and Yosef lab for input; and subjects for their participation in the studies.

Conflict of Interest
The authors declare no conflict of interest.

Figures

Figure 1

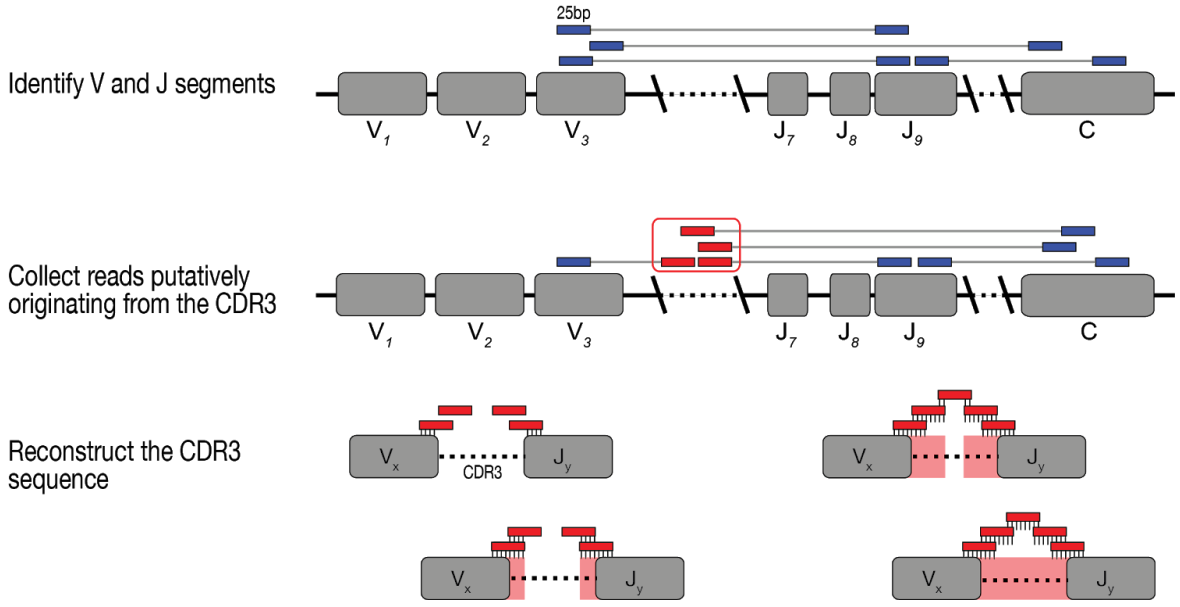


Figure 1: TRAPeS - An algorithm for TCR reconstruction in single cell RNA-seq
Illustration of the TRAPeS algorithm. First, the V and J segment are identified by searching for paired reads with one read mapping to the V segment and its mate mapping to the J segment. Then, a set of putative CDR3-originating reads is identified as the set of unmapped reads whose mates map to the V, J and C segments. Finally, an iterative dynamic programming algorithm is used to reconstruct the CDR3 region.

Figure 2

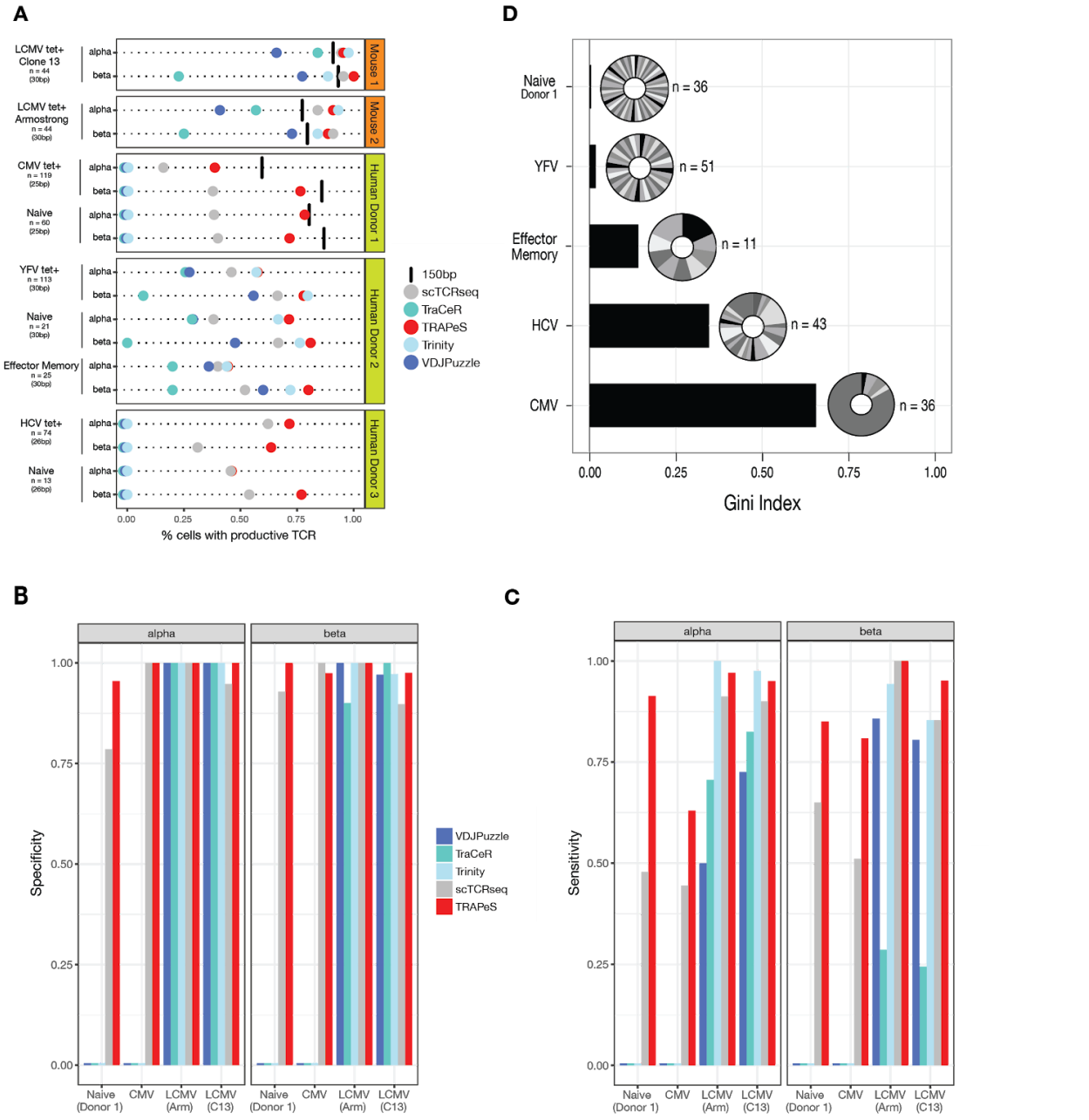


Figure 2: Validation of TRAPeS and comparison to other methods

a) Success rates for reconstruction of productive CDR3 in various CD8$^+$ T cell data sets. Each line depicts the fraction of cells with a productive alpha or beta chain in a given data set with each one of the following methods - 150bp sequencing (black line), short paired-end data reconstructed using TRAPeS (red), TraCeR (turquoise), scTCRseq (gray), VDJPuzzle (dark blue) or Trinity (light blue). **b)** Specificity of TRAPeS. Fraction of cells with identical CDR3 sequence between 150bp data and the 25-30bp data reconstructed either by TRAPeS, TraCeR, scTCRseq, VDJPuzzle or Trinity. This was calculated as the fraction out of cells with a productive chain in both 150 and 25-30bp data. **c)** Sensitivity of TRAPeS. Same as b, except the

fraction of cells is calculated out of the total number of cells that had a successful reconstruction using 150bp sequencing only. **d)** Single cell RNA-sequencing captures a variety of clonal responses. Bars represent the Gini coefficient of each human $CD8^+$ T cell data set. The Gini coefficient can range from zero (a complete heterogeneous population) to one (a complete homogenous population). Pie charts represent the distribution of clones in each population, n represents the number of cells with a successful reconstruction of both alpha and beta chains.

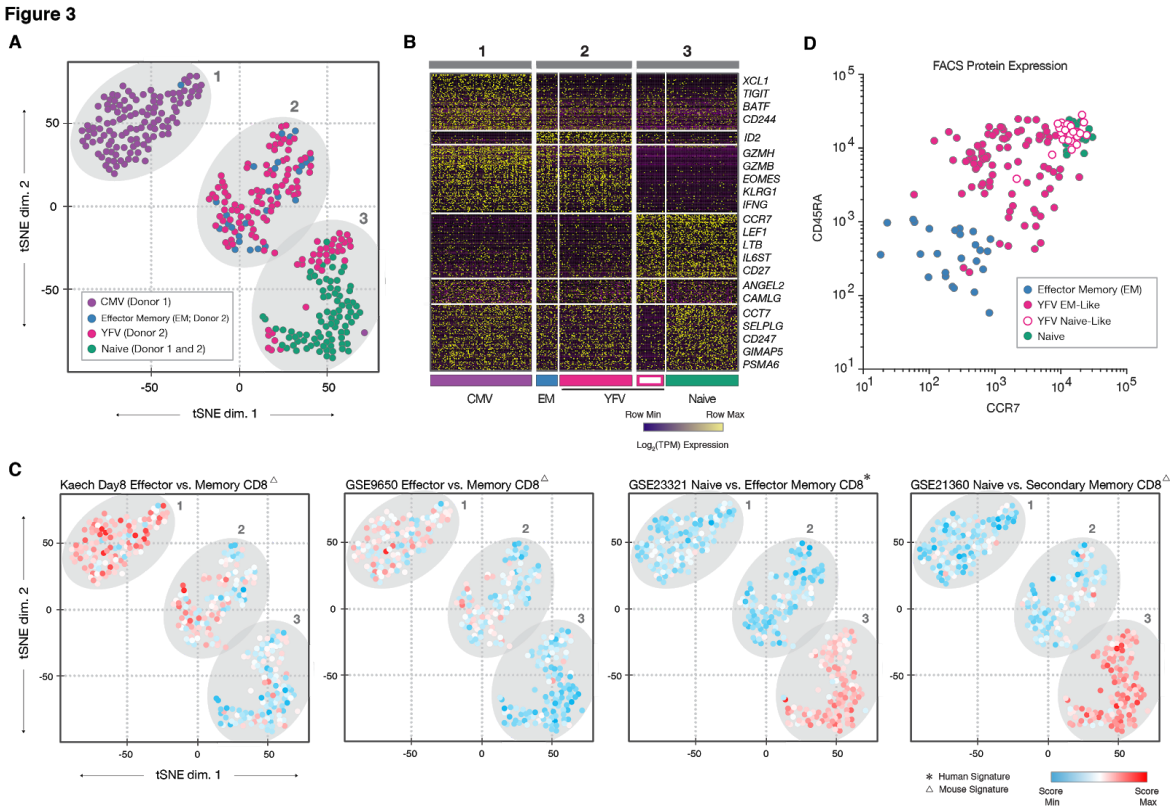


Figure 3: Transcriptome analysis reveals distinct subpopulation of YFV-specific cells exhibiting a naive-like profile

a) t-SNE projection of 353 CMV-specific, Effector Memory, YFV-specific, and Naïve cells, using normalized Transcripts Per Million (TPM) values of 10827 transcripts. Ellipses indicate three distinct spatial clusters. A discrete subset of YFV-specific cells cluster with Naive. **b)** Genes differentially expressed between relevant phenotypic groups. YFV-specific cells were classified as effector memory-like or naive-like using SC3, a non-spatial consensus clustering approach (Figure S7). **c)** t-SNE projections, each cell colored by relative signature score. Shown are two signatures from the ImmuneSigDB distinguishing CMV-specific from YFV-specific cells, and two signatures distinguishing Naïve or YFV-specific naive-like cells from Effector memory, CMV-specific and YFV-specific effector memory-like populations. **d)** FACS protein expression of CCR7 and CD45RA surface molecules from index sort of Effector Memory, YFV-specific effector memory-like, YFV-specific naive-like, and Naïve cells.

Figure 4

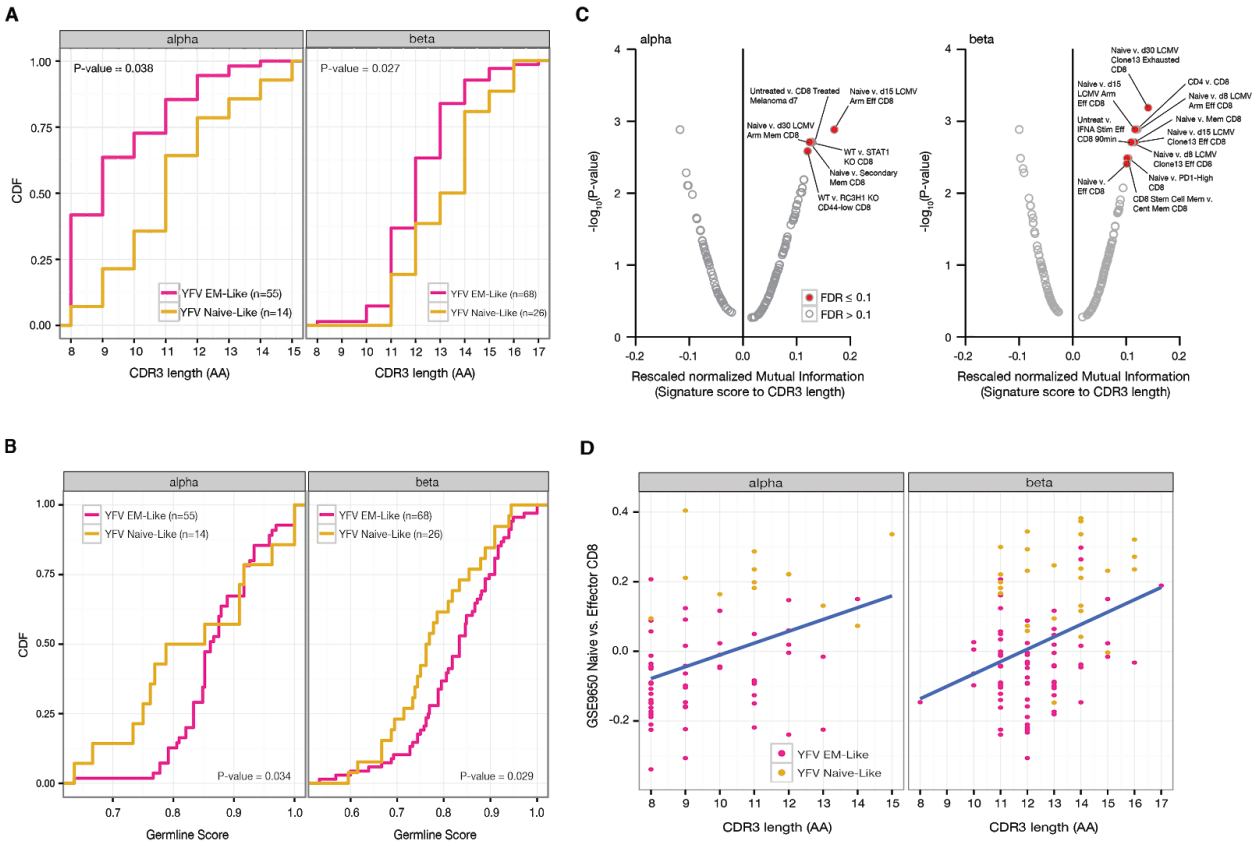


Figure 4: YFV-specific subpopulations display different TCR structure

a) YFV-specific naive-like cells tend to have longer CDR3. Distribution of the YFV-specific effector memory-like and naive-like CDR3 lengths in both alpha (left) and beta (right) chains. P-values were calculated with K-S test. **b)** Differences between naive-like and effector memory-like CDR3 lengths are due to added nucleotides. Distribution of the YFV-specific effector memory-like and naive-like CDR3 germline scores, defined as the number of nucleotides in the CDR3 encoded by the V, D or J segments divided by the total number of nucleotides in the CDR3, for both alpha (left) and beta (right) chains. P-values were calculated with K-S test. **c**) Signature analysis reveals significant correlation between CDR3 length and cell state. The plot depicts the rescaled normalized mutual information score between CDR3 length and transcriptional signatures of $CD8^+$ T cells from ImmuneSigDB. Signatures identified as statistically significant using a permutation test (FDR-adjusted p-value<0.1) are highlighted in red. **d)** YFV-specific cells with long CDR3 tend to have a higher transcriptomic naive signature than cells with short CDR3. Plot represents the score of each cell for a transcriptional signature of a naive vs. effector $CD8^+$ T cell state. A high signature score means that a cell has higher expression of naive signature genes compared to effector signature genes.

References

Akondy, R.S., Johnson, P.L.F., Nakaya, H.I., Edupuganti, S., Mulligan, M.J., Lawson, B., Miller, J.D., Pulendran, B., Antia, R., and Ahmed, R. (2015). Initial viral load determines the magnitude of the human CD8 T cell response to yellow fever vaccination. Proc. Natl. Acad. Sci. U. S. A. *112*, 3050–3055.

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., and Lefranc, M.-P. (2012). IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res. *8*, 26.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, 1–12.

Appay, V., Dunbar, P.R., Callan, M., Klenerman, P., Gillespie, G.M.A., Papagno, L., Ogg, G.S., King, A., Lechner, F., Spina, C.A., et al. (2002). Memory CD8+ T cells vary in differentiation phenotype in different persistent virus infections. Nat. Med. *8*, 379–385.

Barnes, E., Folgori, A., Capone, S., Swadling, L., Aston, S., Kurioka, A., Meyer, J., Huddart, R., Smith, K., Townsend, R., et al. (2012). Novel adenovirus-based vaccines induce broad and sustained T cell responses to HCV in man. Sci. Transl. Med. *4*, 115ra1.

Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. Genome Biol. *13*, R56.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bolinger, B., Sims, S., Swadling, L., O'Hara, G., de Lara, C., Baban, D., Saghal, N., Lee, L.N., Marchi, E., Davis, M., et al. (2015). Adenoviral Vector Vaccination Induces a Conserved Program of CD8(+) T Cell Memory Differentiation in Mouse and Man. Cell Rep. *13*, 1578–1588.

Brochet, X., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res. *36*, W503–W508.

Chattopadhyay, P.K., and Roederer, M. (2015). A mine is a terrible thing to waste: high content, single cell technologies for comprehensive immune analysis. Am. J. Transplant *15*, 1155–1161.

Chung, N.C., and Storey, J.D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics *31*, 545–554.

Cowley, G.S., Weir, B.A., Vazquez, F., Tamayo, P., Scott, J.A., Rusin, S., East-Seletsky, A., Ali,

L.D., Gerath, W.F.J., Pantel, S.E., et al. (2014). Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Scientific Data *1*, 140035.

DeTomaso, D., and Yosef, N. (2016). FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. BMC Bioinformatics *17*, 315.

DeWitt, W.S., Emerson, R.O., Lindau, P., Vignali, M., Snyder, T.M., Desmarais, C., Sanders, C., Utsugi, H., Warren, E.H., McElrath, J., et al. (2015). Dynamics of the cytotoxic T cell response to a model of acute viral infection. J. Virol. *89*, 4517–4526.

Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics *9*, 11.

Eltahla, A.A., Rizzetto, S., Pirozyan, M.R., Betz-Stablein, B.D., Venturi, V., Kedzierska, K., Lloyd, A.R., Bull, R.A., and Luciani, F. (2016). Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. Immunol. Cell Biol. *94*, 604–611.

Fuertes Marraco, S.A., Soneson, C., Cagnon, L., Gannon, P.O., Allard, M., Maillard, S.A., Montandon, N., Rufer, N., Waldvogel, S., Delorenzi, M., et al. (2015a). Long-lasting stem cell like memory CD8 T cells with a naïve-like profile upon yellow fever vaccination. Science Translational Medicine.

Fuertes Marraco, S.A., Soneson, C., Delorenzi, M., and Speiser, D.E. (2015b). Genome-wide RNA profiling of long-lasting stem cell-like memory CD8 T cells induced by Yellow Fever vaccination in humans. Genom Data *5*, 297–301.

Gaublomme, J.T., Yosef, N., Lee, Y., Gertner, R.S., Yang, L.V., Wu, C., Pandolfi, P.P., Mak, T., Satija, R., Shalek, A.K., et al. (2015). Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. Cell *163*, 1400–1412.

Giudicelli, V., Brochet, X., and Lefranc, M.-P. (2011). IMGT/V-QUEST: IMGT Standardized Analysis of the Immunoglobulin (IG) and T Cell Receptor (TR) Nucleotide Sequences. Cold Spring Harb. Protoc. *2011*, db.prot5633.

Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A.J., Mesirov, J.P., and Haining, W.N. (2016). Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. Immunity *44*, 194–206.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Han, A., Glanville, J., Hansmann, L., and Davis, M.M. (2014). Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat. Biotechnol. *32*, 684–692.

Ji, X., Lyu, S.-C., Spindler, M., Bacchetta, R., Goncharov, I., Han, A., Glanville, J., Wang, W.,

Roncarolo, M., Meyer, E., et al. (2015). Deep profiling of single T cell receptor repertoire and phenotype with targeted RNA-seq (TECH2P. 927). The Journal of Immunology *194*, 206–237.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods *14*, 483–486.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clément, O., Chaume, D., and Lefranc, G. (2005). IMGT, the international ImMunoGeneTics information system®. Nucleic Acids Res. *33*, D593–D597.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, B., Li, T., Pignon, J.-C., Wang, B., Wang, J., Shukla, S.A., Dou, R., Chen, Q., Hodi, F.S., Choueiri, T.K., et al. (2016). Landscape of tumor-infiltrating T cell repertoire of human cancers. Nat. Genet. *48*, 725–732.

Li, S., Lefranc, M.-P., Miles, J.J., Alamyar, E., Giudicelli, V., Duroux, P., Freeman, J.D., Corbin, V.D.A., Scheerlinck, J.-P., Frohman, M.A., et al. (2013). IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat. Commun. *4*, 2333.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Love, M., Anders, S., and Huber, W. (2014). Differential analysis of count data--the DESeq2 package. Genome Biol. *15*, 550.

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Mahata, B., Zhang, X., Kolodziejczyk, A.A., Proserpio, V., Haim-Vilmovsky, L., Taylor, A.E., Hebenstreit, D., Dingler, F.A., Moignard, V., Göttgens, B., et al. (2014). Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. Cell Rep. *7*, 1130–1142.

McDavid, A., Finak, G., Chattopadyay, P.K., Dominguez, M., Lamoreaux, L., Ma, S.S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell

qPCR-based gene expression experiments. Bioinformatics *29*, 461–467.

Miles, J.J., Thammanichanond, D., Moneer, S., Nivarthi, U.K., Kjer-Nielsen, L., Tracy, S.L., Aitken, C.K., Brennan, R.M., Zeng, W., Marquart, L., et al. (2011). Antigen-driven patterns of TCR bias are shared across diverse outcomes of human hepatitis C virus infection. J. Immunol. *186*, 901–912.

Newell, E.W., Sigal, N., Bendall, S.C., Nolan, G.P., and Davis, M.M. (2012). Cytometry by Time-of-Flight Shows Combinatorial Cytokine Expression and Virus-Specific Cell Niches within a Continuum of CD8+ T Cell Phenotypes. Immunity *36*.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Pulko, V., Davies, J.S., Martinez, C., Lanteri, M.C., Busch, M.P., Diamond, M.S., Knox, K., Bush, E.C., Sims, P.A., Sinari, S., et al. (2016). Human memory T cells with a naive phenotype accumulate with aging and respond to persistent viruses. Nat. Immunol.

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., Olshen, R.A., Weyand, C.M., Boyd, S.D., and Goronzy, J.J. (2014). Diversity and clonal selection in the human T-cell repertoire. Proc. Natl. Acad. Sci. U. S. A. *111*, 13139–13144.

Ranasinghe, S., Lamothe, P.A., Soghoian, D.Z., Kazer, S.W., Cole, M.B., Shalek, A.K., Yosef, N., Jones, R.B., Donaghey, F., Nwonu, C., et al. (2016). Antiviral CD8(+) T Cells Restricted by Human Leukocyte Antigen Class II Exist during Natural HIV Infection and Exhibit Clonal Expansion. Immunity *45*, 917–930.

Redmond, D., Poran, A., and Elemento, O. (2016). Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. Genome Med. *8*, 80.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat. Genet. *38*, 500–501.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

Robins, H.S., Srivastava, S.K., Campregher, P.V., Turtle, C.J., Andriesen, J., Riddell, S.R., Carlson, C.S., and Warren, E.H. (2010). Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci. Transl. Med. *2*, 47ra64.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Stubbington, M.J.T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A.O., Dougan, G., and Teichmann, S.A. (2016). T cell fate and clonality inference from single-cell transcriptomes. Nat.

Methods.

Swadling, L., Capone, S., Antrobus, R.D., Brown, A., Richardson, R., Newell, E.W., Halliday, J., Kelly, C., Bowen, D., Fergusson, J., et al. (2014). A human vaccine strategy based on chimpanzee adenoviral and MVA vectors that primes, boosts, and sustains functional HCV-specific T cell memory. Sci. Transl. Med. *6*, 261ra153.

Swadling, L., Halliday, J., Kelly, C., Brown, A., Capone, S., Ansari, M.A., Bonsall, D., Richardson, R., Hartnell, F., Collier, J., et al. (2016). Highly-Immunogenic Virally-Vectored T-cell Vaccines Cannot Overcome Subversion of the T-cell Response by HCV during Chronic Infection. Vaccines *4*, 27.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Trautmann, L., Rimbert, M., Echasserieau, K., Saulquin, X., Neveu, B., Dechanet, J., Cerundolo, V., and Bonneville, M. (2005). Selection of T Cell Clones Expressing High-Affinity Public TCRs within Human Cytomegalovirus-Specific CD8 T Cell Responses. The Journal of Immunology *175*, 6123–6132.

Trombetta, J.J., Gennert, D., Lu, D., Satija, R., Shalek, A.K., and Regev, A. (2014). Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. Curr. Protoc. Mol. Biol. *107*, 4.22.1–17.

Venturi, V., Kedzierska, K., Price, D.A., Doherty, P.C., Douek, D.C., Turner, S.J., and Davenport, M.P. (2006). Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. Proc. Natl. Acad. Sci. U. S. A. *103*, 18691–18696.

Venturi, V., Price, D.A., Douek, D.C., and Davenport, M.P. (2008). The molecular basis for public T-cell responses? Nat. Rev. Immunol. *8*, 231–238.

Venturi, V., Quigley, M.F., Greenaway, H.Y., Ng, P.C., Ende, Z.S., McIntosh, T., Asher, T.E., Almeida, J.R., Levy, S., Price, D.A., et al. (2011). A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J. Immunol. *186*, 4285–4294.

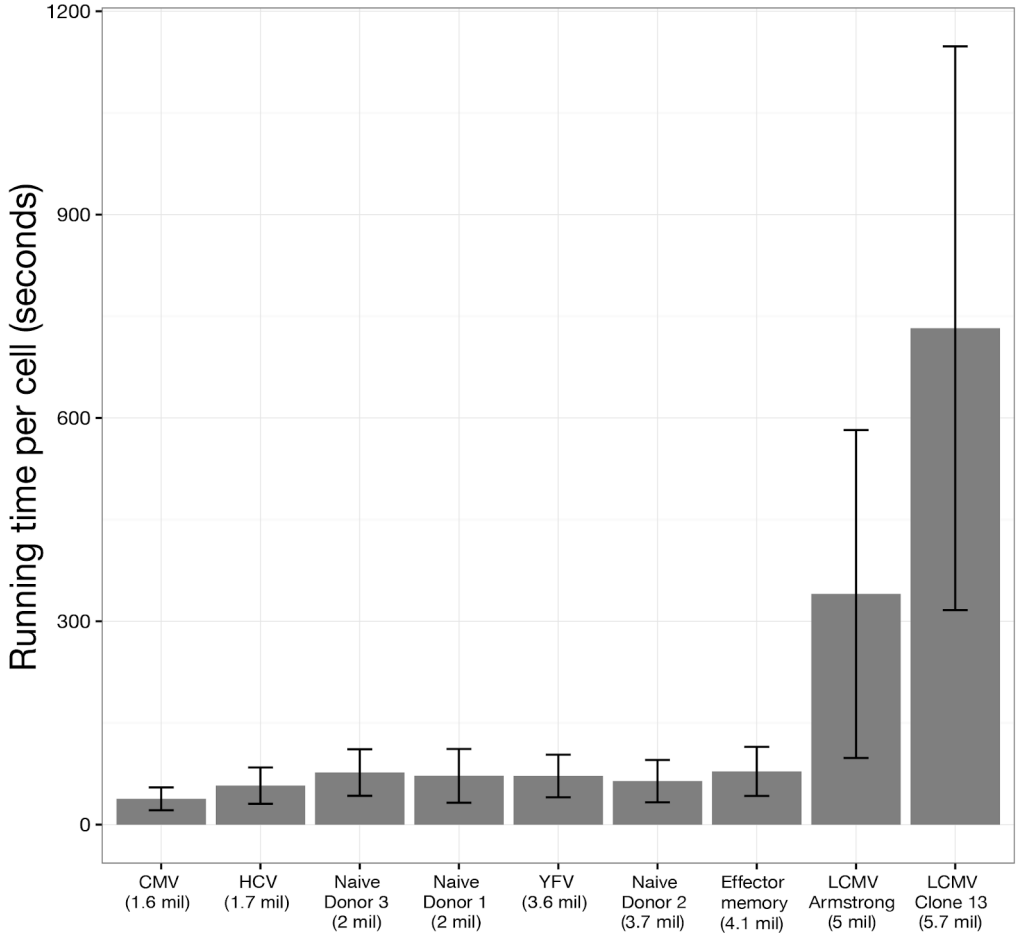
Weekes, M.P., Wills, M.R., Mynard, K., Carmichael, A.J., and Sissons, J.G.P. (1999). The memory cytotoxic T-lymphocyte (CTL) response to human cytomegalovirus infection contains individual peptide-specific CTL clones that have undergone extensive expansion in vivo. J. Virol. *73*, 2099–2108.

Yu, X., Almeida, J.R., Darko, S., van der Burg, M., DeRavin, S.S., Malech, H., Gennery, A., Chinn, I., Markert, M.L., Douek, D.C., et al. (2014). Human syndromes of immunodeficiency and dysregulation are characterized by distinct defects in T-cell receptor repertoire development. J. Allergy Clin. Immunol. *133*, 1109–1115.

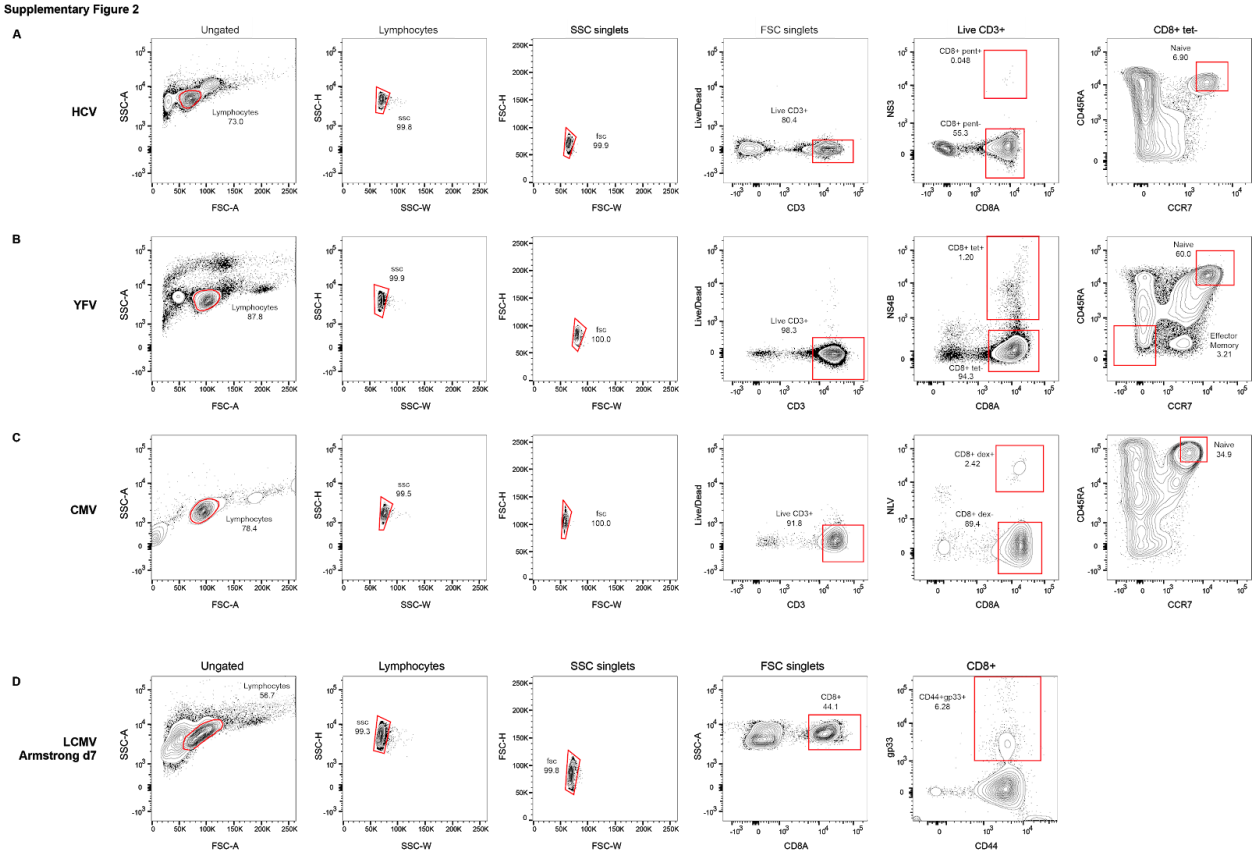
Supplementary information
Supplementary figures and table legend found below. Supplementary tables can be found in the following link: https://academic.oup.com/nar/article/45/16/e148/3976466#119462923

Supplementary Figures

Supplementary Figure 1

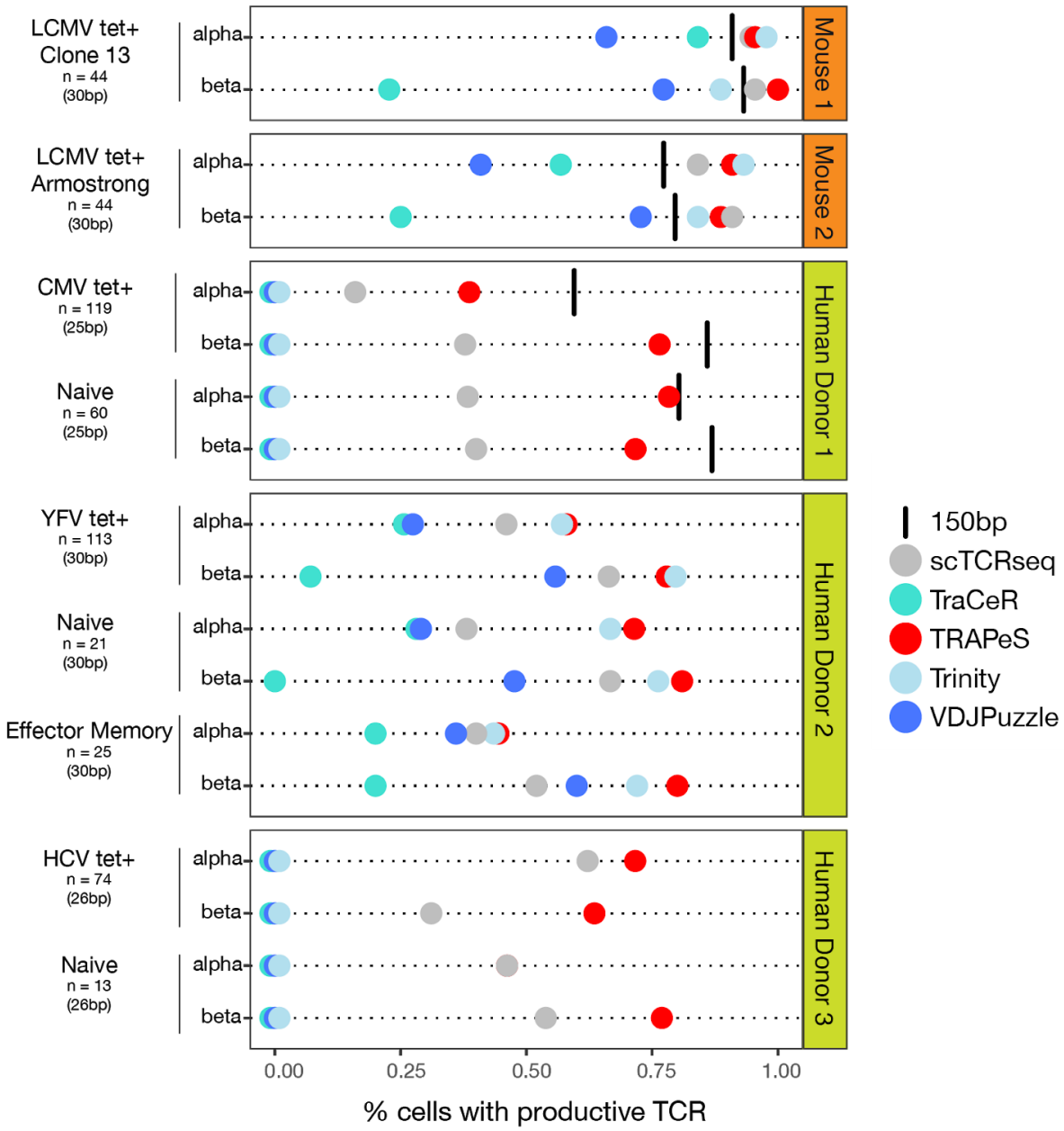


1. TRAPeS run times - Average running time (in seconds) of TRAPeS per cell on all $CD8^+$ T cell datasets analyzed in this study, using a single GHz processor and 8 threads. Error bars represent the standard deviation. For each dataset the average number of reads per cell is mentioned in parentheses. The increased running time in mouse samples is due to the larger sequencing depth but also due to the number of similar V and J segments which resulted in a larger number of possible V-J pairs, increasing running time.



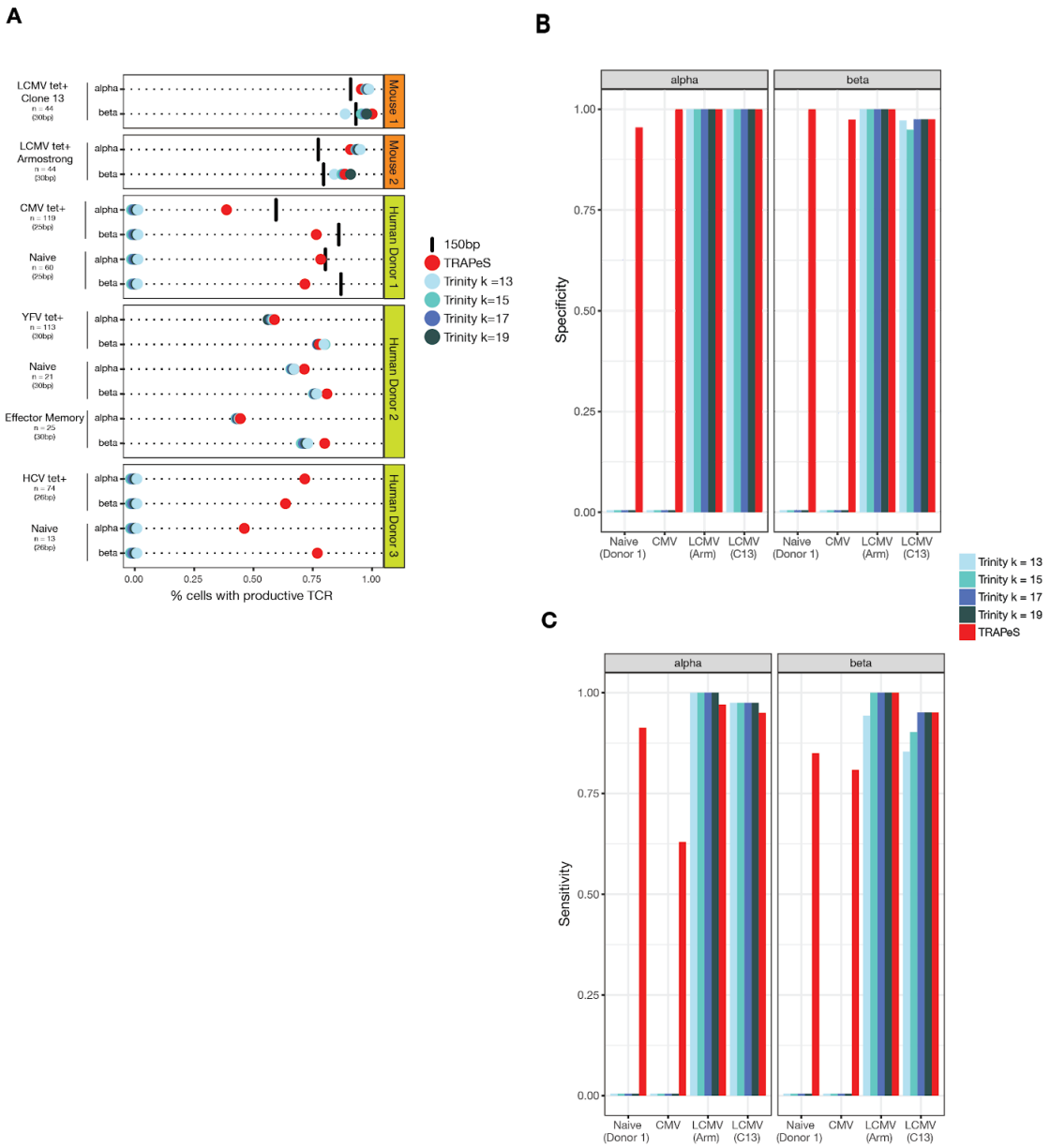
2. Gating strategy for CD8$^+$ T lymphocytes for scRNAseq a) Human lymphocytes were gated based on forward-scatter (FSC) and side-scatter (SSC) characteristics, then singlets were selected from SSC and FSC projections, and Live/Dead-negative CD3$^+$ cells, then CD8$^+$ HCV NS3 (1406-1415; KLSALGINAV; HLA-A*0201)$^+$ cells were selected for HCV-specificity. CD8$^+$ HCV NS3- cells were gated and then CCR7$^+$ and CD45RA$^+$ cells were selected to represent bulk naive CD8$^+$. **b)** Lymphocytes were gated based on FSC and SSC characteristics, then singlets were selected from SSC and FSC projections, and Live/Dead-negative CD3$^+$ cells, then CD8$^+$ YFV NS4 (LLWNGPMAV)$^+$ cells were selected for YFV-specificity. CD8$^+$ YFV NS4- cells were gated and then CCR7$^+$CD45RA$^+$ and CCR7$^-$CD45RA$^-$ cells were selected to represent bulk naive and effector memory CD8$^+$, respectively. **c)** Lymphocytes were gated based on FSC and SSC characteristics, then singlets were selected from SSC and FSC projections, and Live/Dead-negative CD3$^+$ cells, then CD8$^+$ CMV NLV$^+$ cells were selected for CMV-specificity. CD8$^+$ CMV NLV$^-$ cells were gated and then CCR7$^+$CD45RA$^+$ cells were selected to represent bulk naive CD8$^+$. **d)** Mouse lymphocytes were gated based on FSC and SSC characteristics, then singlets were selected from SSC and FSC projections, from which CD8$^+$ cells, then CD44$^+$ gp33$^+$ cells were selected for LCMV-specificity.

Supplementary figure 3



3. Success rates for reconstruction of productive CDR3 in the various CD8$^+$ T cell data sets described in figure 2a, without applying cell quality filtering. Results are based on the same dataset as in Figure 2a, with the exception of not applying any cell quality filtering before TCR reconstruction. Each line depicts the fraction of cells with a productive alpha or beta chain in a given data set with each one of the following methods - 150bp sequencing (black line), short paired-end data reconstructed using TRAPeS (red), TraCeR (turquoise), scTCRseq (gray), VDJPuzzle (dark blue) or Trinity (light blue).

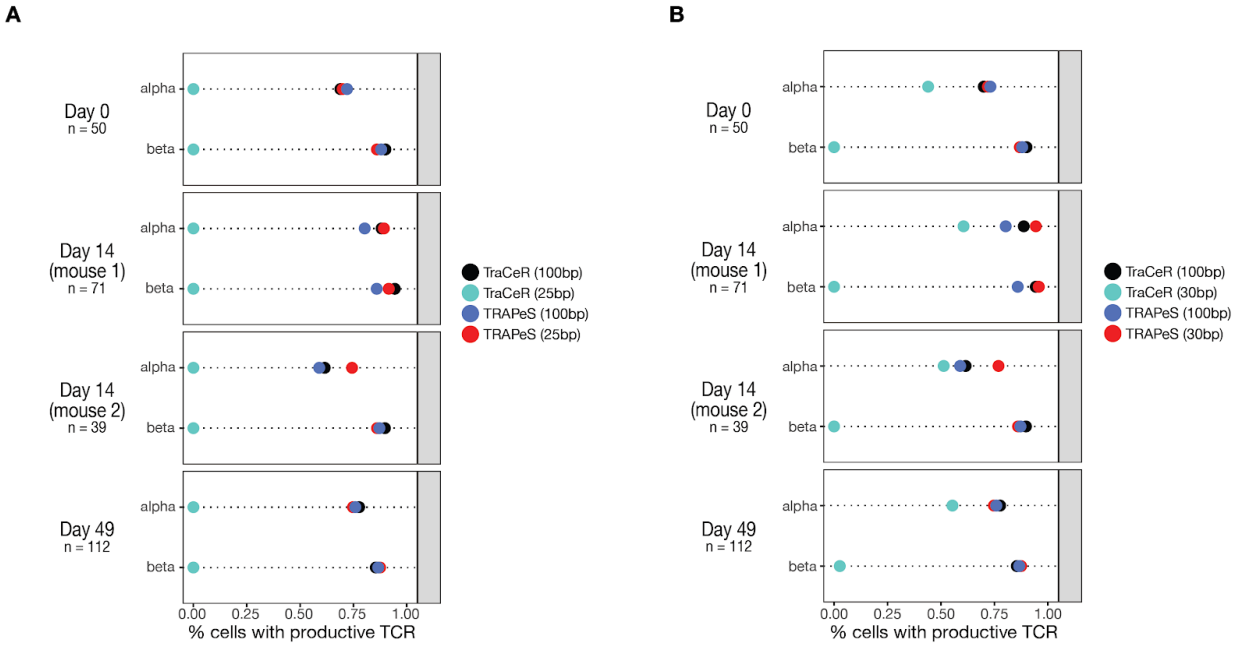
Supplementary figure 4



4. Comparison of success rates for Trinity with various choices of *k-mer* length

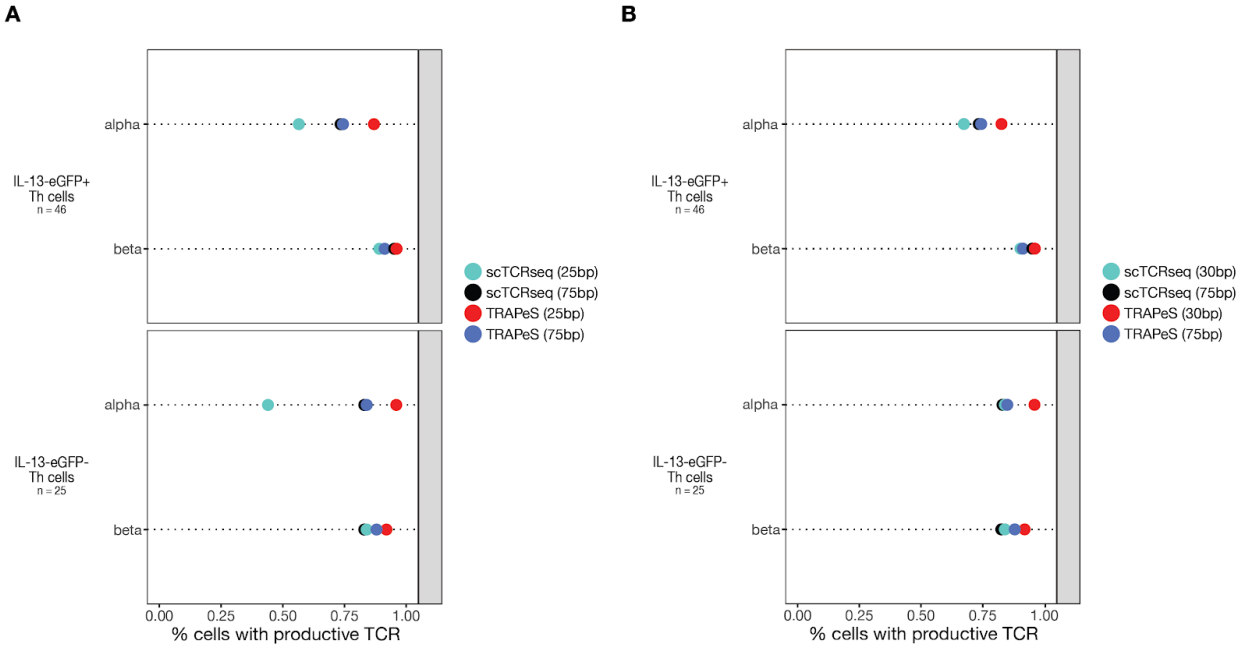
a) Success rates for reconstruction of productive CDR3 in various $CD8^+$ T cell data sets. Each line depicts the fraction of cells with a productive alpha or beta chain in a given data set with each one of the following methods - 150bp sequencing (black line), short paired-end data reconstructed using TRAPeS (red), or Trinity with *k-mer* length of 13 (light blue), 15 (turquoise), 17 (dark blue) or 19 (dark gray). **b)** Specificity of TRAPeS and Trinity. Fraction of cells with identical CDR3 sequence between 150bp data and the 25-30bp data reconstructed either by TRAPeS or Trinity using various *k-mer* lengths. This was calculated as the fraction out of cells with a productive chain in both 150 and 25-30bp data. **c)** Sensitivity of TRAPeS and Trinity. Same as b, except the fraction of cells is calculated out of the total number of cells that had a successful reconstruction using 150bp sequencing only.

Supplementary figure 5



5. Success rates for reconstruction of productive CDR3 in the benchmark data sets used by TraCeR - Data includes 272 $CD4^+$ T cells from an uninfected mouse, two mice with Salmonella typhimurium infection at day 14 and one mouse at day 49 post-infection. **a)** Each line depicts the fraction of cells with a productive alpha or beta chain in the original data (100bp paired-end) with TRAPeS (dark blue) or TraCeR (black), and in the trimmed data (taking only the outer 25bp of each read) with TRAPeS (red) or TraCeR (turquoise). Reconstruction rates for TraCeR on the original data were calculated based on supplementary table 2 from Stubbington et al., counting the number of cells with a reported CDR3 sequence that was annotated as productive. **b)** Similar to figure a, except data was trimmed to include only the outer 30bp (instead of 25bp).

Supplementary figure 6

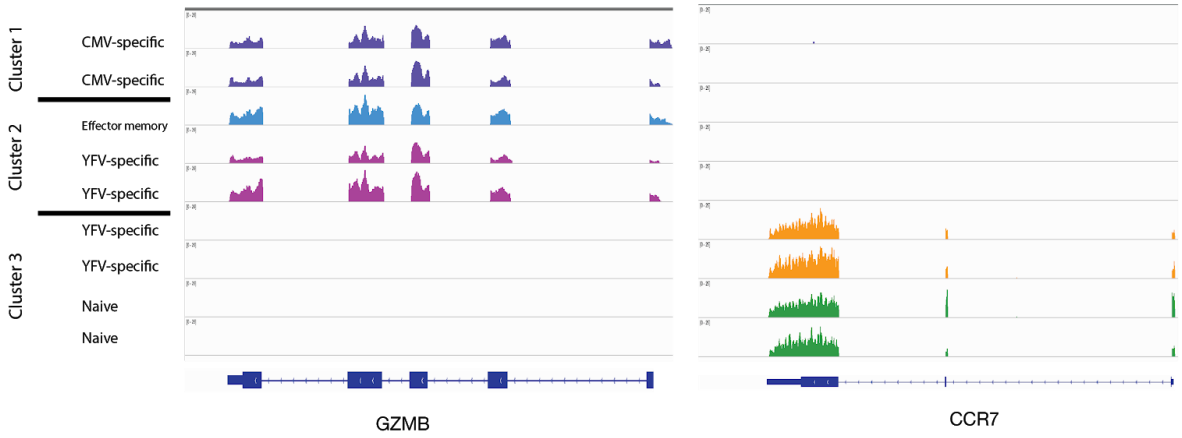


6. Success rates for reconstruction of productive CDR3 in the benchmark data sets used by scTCRseq - Data includes 71 $CD4^+$ T cells from a single mouse. Naive cells were activated under conditions inducing Th2 differentiation, and single cell RNA-sequencing was performed on cells from two populations: IL-13-$eGFP^+$ and IL-13-$eGFP^-$. **a)** Each line depicts the fraction of cells with a productive alpha or beta chain in the original data (75bp paired-end) with TRAPes (dark blue) or scTCRseq (black), and in the trimmed data (taking only the outer 25bp of each read) with TRAPeS (red) or scTCRseq (turquoise). Reconstruction rates for scTCRseq on the original data were calculated based on supplementary table 5 from Redmond et al., counting the number of cells with a complete productive CDR3 (no missing amino acids). **b)** Similar to figure a, except data was trimmed to include only the outer 30bp (instead of 25bp).

Supplementary Figure 7

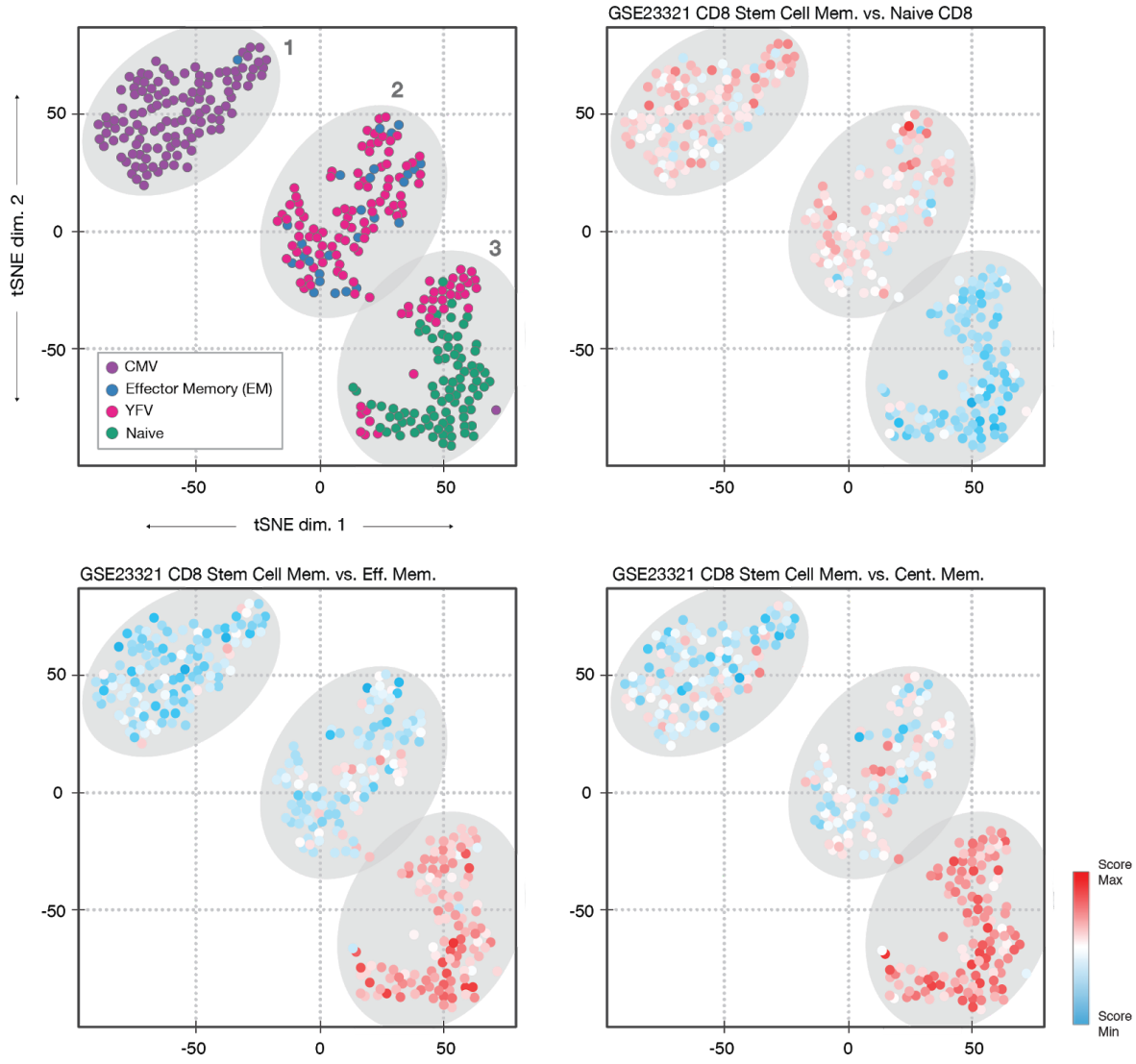
7. SC3 consensus clustering matrix - Heatmap of pairwise reclustering frequencies between 353 single cells, indicating three unsupervised clusters. Color bars indicate cellular phenotypes. YFV-specific cells co-clustering with Effector Memory cells were subsequently annotated as Effector Memory-like, and YFV-specific cells co-clustering with true naive cells were annotated as Naive-like (Table S4).

Supplementary Figure 8



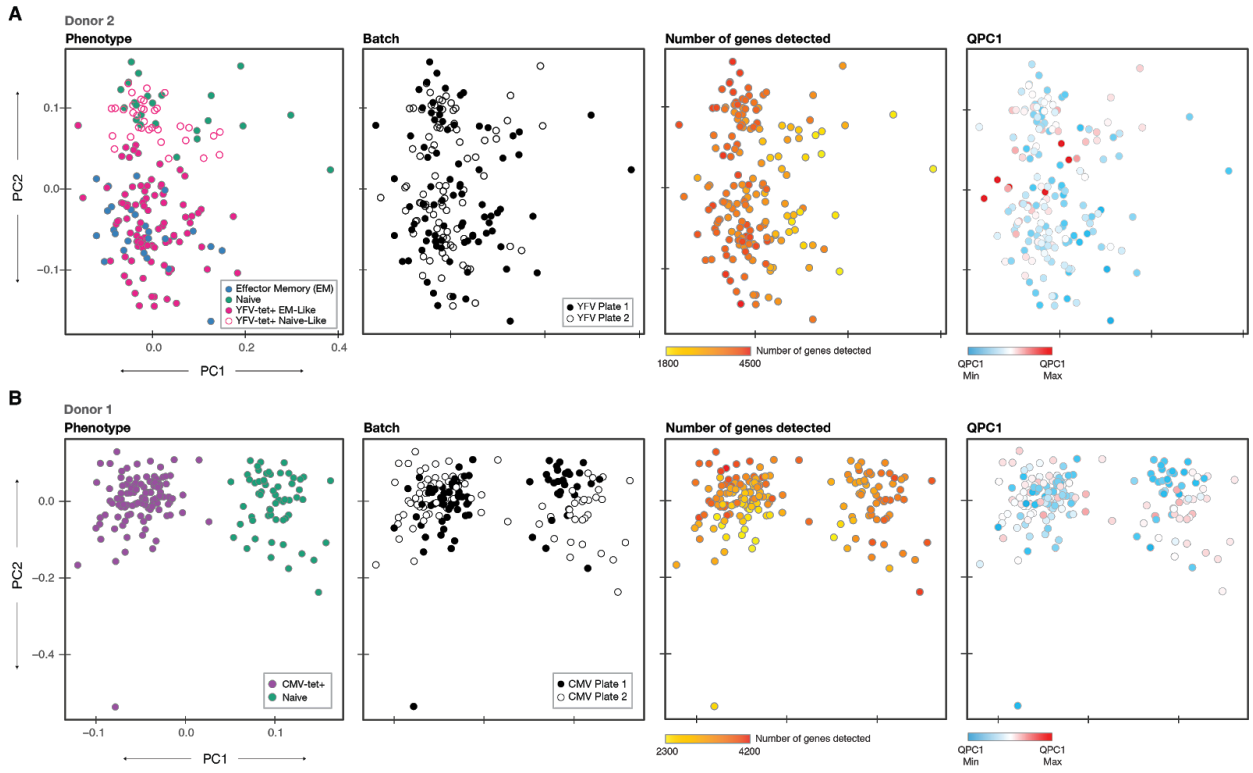
8. Gene expression changes across clusters - An Integrative Genome Viewer plot of selected cells from all clusters of genes highly expressed in naive cells (CCR7, right) and antigen-experienced cells **(**Granzyme B, left).

Supplementary Figure 9



9. Stem cell signatures - t-SNE projections, cells colored by relative signature score for a CD8 Stem Cell Memory vs. Naive, CD8 Stem Cell Memory vs. Effector Memory, and CD8 Stem Cell Memory vs. Central Memory signature from ImmuneSigDB.

Supplementary Figure 10



10. Normalization reduces technical effect - PCA projections of donor 2 (top) and donor 1 (bottom) after normalization. Cells colored by (from left to right): Phenotype, sequencing batch, number of genes detected, value of the first PC of the quality matrix (see Methods).

Supplementary Table Legends

1. TRAPeS output for all data sets - the output of the TRAPeS software for all the data sets presented in this study. In addition to the standard output, to each entry the following information was added: The threshold score of the alignment used in the reconstruction and whether or not this was the TCR sequence used for further analysis.

2. Sample information - For each data set, the table includes batch, total number of sequenced cells, number of cells after quality filtering that were used for TCR reconstruction and for transcriptome analysis (Methods), read lengths, average number of sequencing depth per cell, average percent of reads that were aligned to the genome as well as number of reads used for transcriptome quantification and TCR reconstruction. In addition, the table mentions to which data sets exist long (150bp) RNA-sequencing, used for TCR reconstruction validation.

3. Clones detected in each data set - CDR3 sequences of the alpha and beta chains of the clones detected in all data sets.

4. Clustering of CMV and YFV donors cells - Single cells annotated by phenotype. YFV-specific cells were determined to be Effector Memory-Like or Naive-Like based on SC3 co-clustering with either Effector Memory or Naive cells.

5. Differentially expressed genes among clusters - DESeq2 results of differential expression analysis among clusters. Figure 3B was populated with genes significantly up-regulated (FDR-adjusted P-value < 0.05) in cluster 1 vs. clusters 2 and 3, cluster 2 vs. clusters 1 and 3, cluster 3 vs. clusters 1 and 2 and YFV Naive-Like vs. Naive. Genes are ordered by $\log_2$(Fold Change) in the given comparison, and each comparison was allowed to contribute a maximum of 100 genes to the heatmap.

6. Overlap of differentially expressed genes with ImmuneSigDB signatures - Top genes differentially up-regulated in one cluster relative to all other clusters (i.e. cluster 1 vs. clusters 2 and 3, cluster 2 vs. clusters 1 and 3, and cluster 3 vs. clusters 1 and 2) or in YFV Naive-Like cells relative to Naive, were tested for significance of overlap with ImmuneSigDB $CD8^+$ signatures (up-regulated and down-regulated genes) using a Fisher Exact test. Top differentially expressed genes were defined as having an FDR-adjusted P-value < 0.05 and a $\log_2$(Fold Change) > 1.00, in a given DESeq2 comparison.

7. Differentially expressed genes among clusters using Seurat - Results of differential expression analysis among clusters using a likelihood ratio test based on bi-modally distributed, zero-inflated data. FDR-adjustment was performed using the Benjamini and Hoschberg method. Reported are all genes differentially up-regulated in one cluster relative to all other clusters, or in YFV Naive-Like cells relative to Naive, at an FDR-adjusted P-value < 0.05.

8. Gene Signature analysis of cell clusters - FastProject signatures scores for each cell and statistical analysis of the signatures across clusters. The statistical analysis was performed with a

K-S test of the signature scores between one cluster and the other two. The table includes the K-S test p-value as well as FDR-adjusted p-value.

9. TCR properties of YFV-specific cells - TCR properties of the YFV-specific cells, including CDR3 length, germline score, transcript expression and hydrophobicity for alpha and beta chains, as well as normalized tetramer staining intensity of each cell (Methods). The table also includes the results of the K-S test used to compare all properties above in the naive-like vs. the effector memory-like cells.

10. Association of YFV-specific gene signature and TCR length - FastProject signatures scores for each YFV-specific cell and statistical analysis of the association of the signatures with TCR length using mutual information (Methods).

11. YFV and CMV donors' expression matrix - Normalized TPM values of cells from YFV and CMV donors (Methods). The TPM matrix was collapsed from transcripts to genes by highest mean expression across all cells. We then applied a gene filter, which retained only genes expressed at a minimum of 5 TPM in at least 1% of cells, resulting in a matrix of 10827 genes by 353 cells.

Chapter 5 - Reconstructing B cell receptor sequences from short-read single cell RNA-sequencing with BRAPeS

In this chapter I describe BRAPeS, an extension of the TRAPeS software to reconstruct B cell receptor (BCR) sequences from short read single-cell RNA-sequencing. BRAPeS is modified from the TRAPeS algorithm to account for somatic hypermutations and isotype switching, biological processes that improve the specificity and function of the receptor and are unique to B cells.

This work was published in *Life Science Alliance* in 2019 (Afik et al. 2019), and I am reporting it as it was published. The authors on the paper are:

Shaked Afik[1,+], Gabriel Raulet[2,+] and Nir Yosef[1,3, 4, 5,*]
1. Center for Computational Biology, University of California, Berkeley, Berkeley, CA, 94720, USA
2. Department of Computer Science, University of California, Davis, CA, 95616, USA
3. Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, 94720, USA
4. Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA
5. Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

[+] both authors contributed equally
* Corresponding author

Abstract

RNA-sequencing of single B cells provides simultaneous measurements of the cell state and its antigen specificity as determined by the B cell receptor (BCR). However, in order to uncover the latter further reconstruction of the BCR sequence is needed. We present BRAPeS, an algorithm for reconstructing BCRs from short-read paired-end single cell RNA-sequencing. BRAPeS is accurate and achieves a high success rate even at very short (25bp) read length, which can decrease the cost and increase the number of cells that can be analyzed compared to long reads. BRAPeS is publicly available at the following link: https://github.com/YosefLab/BRAPeS.

Introduction

B cells play a significant role in the adaptive immune system, providing protection against a wide range of pathogens. This diversity is due to the B cell receptor (BCR), which enables different cells to bind different pathogens (Imkeller and Wardemann, 2018). Single cell RNA-sequencing (scRNA-seq) has emerged as one of the leading technologies to characterize and study heterogeneity in the immune system across cell types, development and dynamic processes (Papalexi and Satija, 2018; Villani et al., 2018). Combining transcriptome analysis with BCR reconstruction in single cells can provide valuable insights to the relation between BCR and cell state, as was demonstrated by similar studies in T cells (Afik et al., 2017; Eltahla et al., 2016; Stubbington et al., 2016).

The BCR is comprised of two chains, a heavy chain and a light chain (either a kappa or lambda chain). Each chain is encoded in the germline by multiple segments of three types - variable (V), joining (J) and constant (C) segments (the heavy chain also includes a diversity (D) segment, see Materials and Methods). The specificity of the BCRs comes from the V(D)J recombination process, in which for each chain one variable (V) and one joining (J) segment are recombined in a process which introduces insertions and deletions into the junction region between the segments, called the complementarity determining region 3 (CDR3) (Tonegawa, 1983). The resulting sequence is the main determinant of the cell's ability to recognize a specific antigen. Following B cell activation, somatic hypermutations are introduced to the BCR and the constant region may be replaced in a process termed isotype switching (Di Noia and Neuberger, 2007). The random mutations make BCR reconstruction a challenging task. While methods to reconstruct BCR sequences from full length scRNA-seq are available (Canzar et al., 2017; Lindeman et al., 2018; Rizzetto et al., 2018) (as well as single cell V(D)J enriched libraries from 10x Genomics: https://www.10xgenomics.com/solutions/vdj/), they were only tested on long reads (150bp and 50bp). The ability to reconstruct BCR sequences from short (25-30bp) reads is important, as it can decrease cost which can, in turn, increase the number of cells which could be feasibly analyzed.

We introduce BRAPeS ("BCR Reconstruction Algorithm for Paired-end Single cells"), an algorithm and software for BCR reconstruction. Conversely to other methods, BRAPeS was designed to work with short (25-30bp) reads, and indeed we demonstrate that under these settings it performs better than other methods. Furthermore, we show that the performance of BRAPeS when provided with short reads is similar to what can be achieved with much longer (50-150bp) reads from the same cells, suggesting that BCR reconstruction does not necessitate costly sequencing with many cycles.

Results

BRAPes is an extension of the TCR reconstruction software TRAPeS (Afik et al., 2017), with significant modifications added to address the processes of isotype switching and somatic hypermutations which are specific to B cells (Figure 1, see Materials and Methods for full description of the algorithm). Briefly, BRAPeS takes as input the alignment of the reads to the reference genome. BRAPeS first recognizes the possible V and J segments by finding reads with

one mate mapping to a V segment and the other mate mapping to a J segment. All unmapped reads whose mates were mapped to the V/J/C segments are then collected, assuming that most CDR3-originating reads will be unmapped when aligning to the reference genome. Then, the CDR3 region is reconstructed with an iterative dynamic programming algorithm. At each step, BRAPeS aligns the unmapped reads to the edges of the V and J segments, using the sequence of the aligned reads to extend the V and J sequences until convergence. Next, the BCR isotype is determined by appending all possible constant segments to the reconstructed sequence and taking the most likely complete transcript based on transcriptomic alignment with RSEM (Li and Dewey, 2011). Finally, BRAPeS corrects for somatic hypermutations by collecting all reads aligning to the genomic regions of the CDR1, CDR2 and the framework regions (FRs) and aligning these reads against each other to obtain a reconstruction of the consensus sequence. The CDR3 sequences and their productivity are determined based on the criteria established by the international ImMunoGeneTic information system (IMGT) (Lefranc, 2014; Lefranc et al., 2015) (Methods).

We evaluated BRAPeS' performance on 374 cells from two previously published data sets - 174 human B cells and 200 mouse B cells (Materials and Methods, Supplementary Table S1) (Canzar et al., 2017; Wu et al., 2016). To evaluate BRAPeS, we first trimmed the original reads (50bp for the human data and 150bp for the mouse data) and kept only the outer 25 or 30 bases. We compared BRAPeS' performance on the trimmed data to two other previously published methods - BASIC (Canzar et al., 2017) and VDJPuzzle (Rizzetto et al., 2018) applied either on the trimmed data or the original long reads.

When applied to 30bp reads, BRAPeS' success rates are similar to other methods for the light chain, but are higher for heavy chain reconstruction (Figure 2a, Supplementary Table S2). BRAPeS reconstructs productive heavy chains in a total of 348 cells, 93% of the cells across both datasets, and reconstructs productive light chains in 370 cells (98.9% of the cells). These results are in line with the success rates of BASIC and VDJPuzzle on the original long reads: BASIC reconstructs productive heavy and light chains in 353 (94.4%) and 364 (97.3%) cells, respectively, and VDJPuzzle reconstructs heavy chains in 346 (92.5%) cells and light chains in 368 (98.4%) cells. On 30bp reads, BASIC and VDJPuzzle achieve similar reconstruction rates for the light chain (362 (96.8%) cells and 370 (98.9%) cells with a productive light chain in BASIC and VDJPuzzle, respectively). However, BASIC and VDJPuzzle see a decline in success rates for the heavy chain, reconstructing a productive heavy chain in only 273 (73%) cells for BASIC and 242 (64.7%) cells for VDJPuzzle (Figure 2a, Supplementary Table S2).

BRAPeS is also able to maintain a high success rate on 25bp reads, reconstructing heavy chains in 328 (87.7%) cells and light chains in 370 (98.9%) cells (Figure 2b and Supplementary Table S3). Yet, we observe a substantial decrease in the results of other methods. VDJPuzzle is unable to reconstruct any chains with 25bp reads. This is likely due to its use of the *de-novo* assembler Trinity (Grabherr et al., 2011) which requires a seed k-mer length of 25bp that is unsuitable for very short reads. Similarly to 30bp, BASIC is able to maintain a high reconstruction rate for light chains, with productive reconstructions in 363 (97.1%) cells, but is only able to reconstruct productive heavy chains in 204 (54.5%) cells (Figure 2b, Supplementary Table S3). Moreover, BASIC only outputs fasta sequences, thus requiring further processing to annotate the BCR.

We next turn to evaluate the accuracy of the short-read based CDR3 reconstructions, by comparing the resulting sequences to those obtained with long reads (Figure 3, Materials and Methods). We use the long-read based reconstruction of BASIC as a reference (we achieve similar results with VDJPuzzle on the long-read data; see Supplementary Figure S1) and evaluate the accuracy in terms of sensitivity (how many of the CDR3 sequences in the full length data have an identical reconstruction with the short reads) and specificity (how many of the CDR3 sequences in the short-read data have an identical long-read reconstruction). In general, all methods show a high level of specificity, having almost all CDR3 sequences identical to the sequences reconstructed on long reads, whenever both read lengths produce a productive reconstruction (Figure 3a-b). In accordance with the higher success rate, BRAPeS shows a high sensitivity, with a rate of 0.96 for 30bp data and 0.92 for 25bp data (Figure 3c-d). This is in line with the agreement of different methods on the original data, as VDJPuzzle on long reads has a sensitivity rate of 0.96. On the trimmed data, BASIC and VDJPuzzle show a lower sensitivity rate - BASIC achieves sensitivity rates of 0.87 and 0.78 for 30bp and 25bp respectively, and VDJPuzzle has a sensitivity rate of 0.83 for 30bp. These results also hold if we only take the top reconstruction of BRAPeS, as more than 97.5% of the identical CDR3 sequences between BRPAeS and BASIC are the highest ranked CDR3 sequences for both 25bp and 30bp (Supplementary Figures S2 and S3).

BRAPeS' correction of somatic hypermutations is also accurate across the various regions of the transcript (Figure 3). Besides a slight decrease in specificity for CDR2 and FR1 reconstruction, BRAPeS maintains a very high level of specificity across all regions in line with the other methods. We note that BASIC achieves lower specificity rates for FR1 reconstructions for short reads mostly due to partial reconstructions. Overall, BRAPeS has a high sensitivity rate across all regions (0.92-0.97 for 30bp and 0.88-0.94 for 25bp), comparable to the sensitivity of VDJPuzzle on long reads (0.87-0.95). Similar to the CDR3 results, the high sensitivity and specificity results hold when comparing only the top-ranking reconstruction, as 96.8%-99.9% of identical regions are the top-ranking regions for 30bp, and 95.4%-100% of the top ranking regions for 25bp (supplementary figures S2 and S3).

Discussion

Coupling BCR reconstruction with transcriptome analysis in single cells can provide valuable information about the effect of antigen specificity and isotype to cellular heterogeneity. Despite an increase in technical noise in transcriptome analysis compared to longer reads (Chhangawala et al., 2015; Rizzetto et al., 2017), short-read sequencing is still widely used as it can reduce sequencing costs by hundreds to thousands of dollars per run, depending on the sequencing platform and desired total number of reads. However, current methods do not provide a sufficient solution for reconstructing immune cell receptors from short reads (Rizzetto et al., 2017). To this end we provide BRAPeS, a software for BCR reconstruction tailored to work on short-read scRNA-seq. BRAPeS is accurate and has a success rate on short reads similar to other methods applied to long reads, demonstrating that BCR reconstruction can be achieved at a much lower cost. BRAPeS is publicly available at https://github.com/YosefLab/BRAPeS

Materials and Methods

The BRAPeS algorithm

The input given to BRAPeS is a directory where each subdirectory includes genomic alignments of a single cell.
The BRAPeS algorithm has several steps, performed separately for each chain in each cell:

1. **Identifying possible pairs of V and J segments**: BRAPeS searches for reads where one mate of the pair is mapped to a V segment and the other mate is mapped to a J segment. BRAPeS collects all possible V-J pairs and attempts to reconstruct complete BCRs from all possible pairs. Since the D segment is very short, reads do not align to it, thus as part of the reconstruction step (step 3) the sequence of the D segment is also reconstructed. If no V-J pairs are found, BRAPeS will look for V-C and J-C pairs and will take all possible V/J pairing of the found V and J segments.
   In case of many possible V-J pairs (which can occur due to the similarity among the segments), the user can limit the number of V-J pairs to attempt reconstruction on. BRAPeS will rank the V-J pairs based on the number of reads mapped to them and take only the top few pairs (the exact number is a parameter controlled by the user).
2. **Collecting the set of putative CDR3-originating reads**: BRAPeS collects the set of reads that are likely to originate from the CDR3 region. Those are the reads that are unmapped to the reference genome, but their mates are mapped to the V/J/C segments. In addition, since the first step of CDR3 reconstruction includes alignment to the ends of the genomic V and J sequences, reads mapping to the V and J segments are also collected.
3. **Reconstructing the CDR3 region**: For each V-J pair, the edges of the V and J segments are extended with an iterative dynamic programming algorithm. In each iteration, BRAPeS tries to align all the unmapped reads to the V and J sequences separately with the Needleman-Wunsch algorithm with the following scoring scheme: +1 for match, -1 for mismatch, -20 for gap opening and -4 for gap extension. In addition, BRAPeS does not penalize having a read "flank" the genomic segment. All reads that passed a user defined threshold are considered successful alignments. BRAPeS then builds the extended V and J segments by taking for each position the base which appears in most reads. This process repeats for a given number of iterations or until the V and J segments overlap. Since the purpose of this step is to reconstruct only the CDR3 region, in order to reduce running time the alignment is performed only on a predetermined number of bases leading to the ends of the V and J segment (3' end of the V segment and 5' end of the J segment). The number of bases taken from the end of each segment is a parameter controlled by the user, set by default to the length of the J segment. BRAPeS can also run a "one-sided" mode, where if an overlap was not found (e.g. due to assigning the wrong V segment), BRAPeS will attempt to determine the productivity of only the extended V and only of the extended J segment.
4. **Isotype determination**: To find the BCR isotype, for each V-J pair with a reconstructed CDR3 BRAPeS concatenates the full sequences of all possible constant segments. Then, BRAPeS runs RSEM (Li and Dewey, 2011) on all sequences using all paired-end reads with at least one mate mapped to the genomic V/J/C segments as input. For each V-J pair

the constant region with the highest expected count is taken as the chosen constant segment.

5. **Somatic hypermutation correction**: All the reads from step 4 are aligned against the genomic CDR1, CDR2 and framework sequences obtained from IMGT using the SeqAn package (Döring et al., 2008). Reads are chosen as candidates for reconstruction if the percentage of mutations in the aligned sequence is below a given input threshold, set by default to 0.35 for CDRs and 0.2 for FRs. Separate thresholds are used for framework and complementary determining regions to account for higher rates of somatic hypermutations in the CDRs. When reads align across flanking CDR-framework regions, the rate of mutation is calculated separately for the aligned framework segment and the aligned CDR segment. If both score below their given thresholds, the read is saved for reconstruction. Once all putative reads have been collected, they are first aligned based on the coordinates obtained from the genomic alignments. Then, to correct for possible misalignments, the consensus alignment algorithm in the Seqan package is run using these approximate positions as guides. Finally, the reconstructed sequence is obtained by aligning the genomic sequences against the consensus sequence to find their start and end coordinates.
6. **Separating similar BCRs and determining chain productivity**: After selecting the top isotype for each V-J pair and correcting for somatic hypermutations, BRAPeS determines if the reconstructed sequence is productive (i.e. the V and J are in the same reading frame with no stop codon in the CDR3) and annotates the CDR3 junction. If more than one V-J pair produces a CDR3 sequence (either due to having more than one recombined chain in the cell or due to similar V-J segments resulting in the same CDR3 sequence reconstruction), the various productive reconstructions are ranked based on their expression values as determined by RSEM.

The output for BRAPeS is the full ranked list of reconstructed chains, including the CDR3 sequences, V/J/C annotations and the number of reads mapped to each segment, as well as a summary file of the success rates across all cells. In addition, for each cell the output is the full sequence of each reconstructed BCR, as well as a file detailing the sequences of the CDR1, CDR2 and framework regions, a file with the read count for each isotype and a file with the read count for each productive BCR.

BRAPeS is implemented in python. To increase performance, the dynamic programming algorithm and the somatic hypermutation correction algorithm is implemented in C++ using the SeqAn package (Döring et al., 2008). Moreover, to decrease running time for deeply sequenced cells, BRAPeS has the option to randomly downsample the number of reads for CDR3 reconstruction to 10,000 and the number of reads for somatic hypermutation correction to 40,000. BRAPeS is publicly available and can be downloaded at the following link: https://github.com/YosefLab/BRAPeS

Data availability and preprocessing

Raw fastq files of mouse B cells were downloaded from Wu et al. (ArrayExpress E-MTAB-4825) (Wu et al., 2016). All analysis was performed on the 200 cells that were

available through ArrayExpress. Raw fastq files for the human data from Canzar et al. (Canzar et al., 2017) were provided by the author. We excluded single-end cells and cells filtered out in the original study, leaving a total of 174 cells. Next, reads were trimmed to be 25 or 30bp paired-end with trimmomatic (Bolger et al., 2014), keeping only the outer bases.
For BRAPeS, low quality reads were trimmed using trimmomatic with the following parameters: LEADING:15, TRAILING:15, SLIDINGWINDOW:4:15, MINLEN:16. The remaining reads were aligned to the genome (hg38 or mm10) using Tophat2 (Kim et al., 2013). Running VDJPuzzle and BASIC on the trimmed reads resulted in no reconstructions for VDJPuzzle and a slight decrease in reconstruction rates for BASIC, thus the results presented in the paper for VDJPuzzle and BASIC are for the raw reads.

Running BRAPeS

For this study, BRAPeS was run using the following parameters for the human data: "-score 15 -top 6 -byExp -iterations 6 -downsample -oneSide". The "score" is the minimal alignment score for the CDR3 reconstruction step and "iterations" limits the number of times BRAPeS attempts to extend the V and J segments. The parameters "top" and "byExp" determine the maximal number of V-J pairs per chain on which reconstruction is attempted, by ranking the pairs based on their number of aligned reads and sampling from the pairs with the highest read count. The "downsample" parameter reduces the number of reads used for CDR3 reconstruction and somatic hypermutation correction.
For the mouse data BRAPeS was run with the following parameters: "-score 15 -oneSide -byExp -top 10". In addition, as some cells required a higher alignment score threshold, we ran BRAPeS with a scoring threshold of 21 for chains without a productive reconstruction.

Running VDJPuzzle and BASIC

We ran VDJPuzzle using default parameters, providing VDJPuzzle with the hg38 genome and GRCh38.p2 annotation for human, and mm10 genome with the GRCm38.p4 annotation for mouse. We then considered only reconstructions with a complete CDR3 (no missing bases) which appeared in the "summary_corrected" folder as valid productive reconstructions.
BASIC was ran with default parameters. After running BASIC we collected all the output fasta files and ran them through IMGT/HighV-Quest (Alamyar et al., 2012; Li et al., 2013). Only sequences that resulted in productive CDR3 according to IMGT were considered successful reconstructions.

Comparison of sensitivity and specificity

To determine the accuracy of the methods, we compared the reconstructed CDR3 nucleotide sequences to the reconstruction produced by running BASIC or VDJPuzzle on long reads. Only CDR3s with sequences identical to the sequences reconstructed on the long-read data were considered accurate. In case of more than one reconstructed CDR3 sequence, if both methods had at least one identical CDR3 sequence it was considered an accurate reconstruction, except for supplementary figures S2 and S3 for which we only compared the highest ranking reconstruction. We used the same criteria of a perfect match to estimate the reconstruction accuracy of CDR1, CDR2, FR1, FR2 and FR3 regions. The annotated FR4 VDJPuzzle output was much longer compared to BASIC, thus when comparing to BASIC we considered the FR4 sequence accurate if the FR4 prefix was identical to the full BASIC FR4 reconstruction.

Author Contributions

S Afik: conceptualization, software, formal analysis, methodology, and writing — original draft, review, and editing.
G Raulet: software, formal analysis, and methodology.
N Yosef: conceptualization, supervision, funding acquisition, methodology, and writing — original draft, review, and editing.

Conflict of Interest

The authors declare they have no conflict of interest

Figures

Figure 1

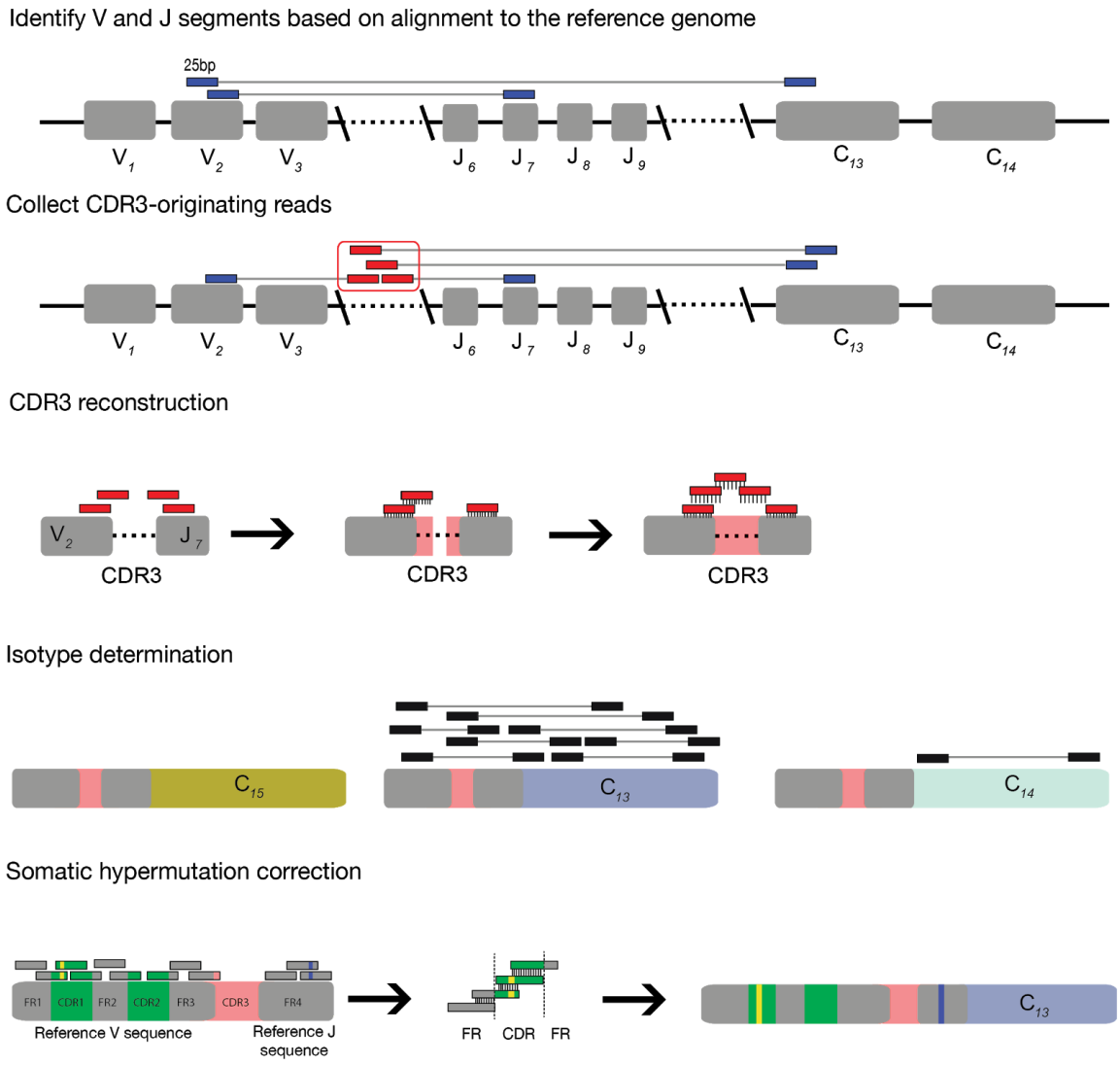


Figure 1: The BRAPeS algorithm. First, the V and J segments are selected based on the initial alignment to the reference genome by searching for paired reads with one read mapping to a V segment and its mate mapping to a J segment. Next, putative CDR3-originating reads are identified as the unmapped reads whose mates map to the V/J/C segments. BRAPeS runs an iterative dynamic programming algorithm to align the CDR3-originating reads to the V and J segments and extend them until they overlap. BCR isotype is then determined by running RSEM on all possible full BCR transcripts (the reconstructed V-J segments combined with all possible constant segments). Finally, BRAPeS corrects for somatic hypermutations by building a consensus sequence of the reads aligning to the CDR1, CDR2 and framework regions.

Figure 2

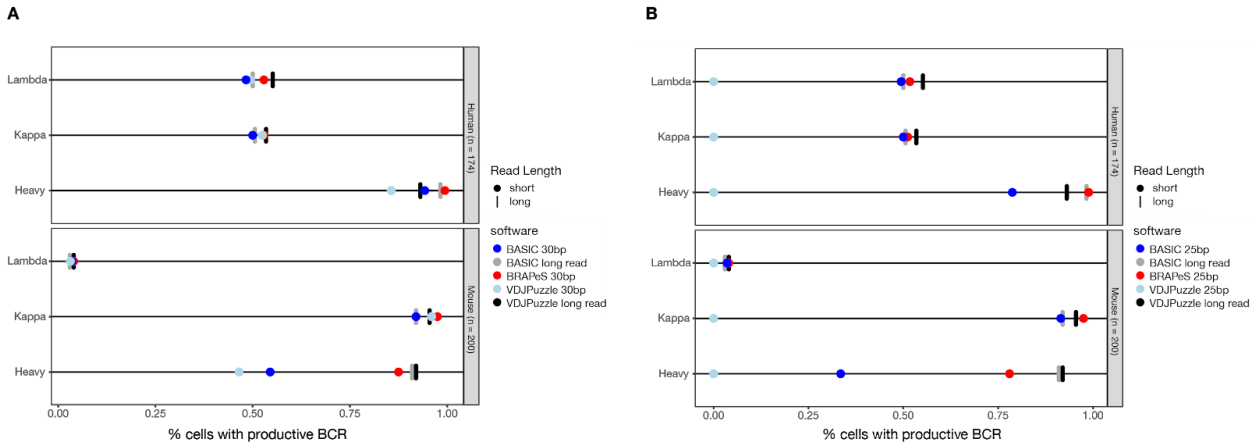


Figure 2: BRAPeS success rates. A) Fraction of cells with a successful reconstruction of a productive CDR3 in human and mouse B cells using the following methods: VDJPuzzle applied to the original, long-read data (black line) and the trimmed version of the data, trimmed to 30bp (light blue circle). BASIC applied to the long-read (grey line) and the trimmed data (dark blue circle), and BRAPeS applied to the trimmed data (red circle). **B)** Same as A, but the trimmed version of the data was trimmed down to include only the outer 25bp, instead of 30bp.

Figure 3

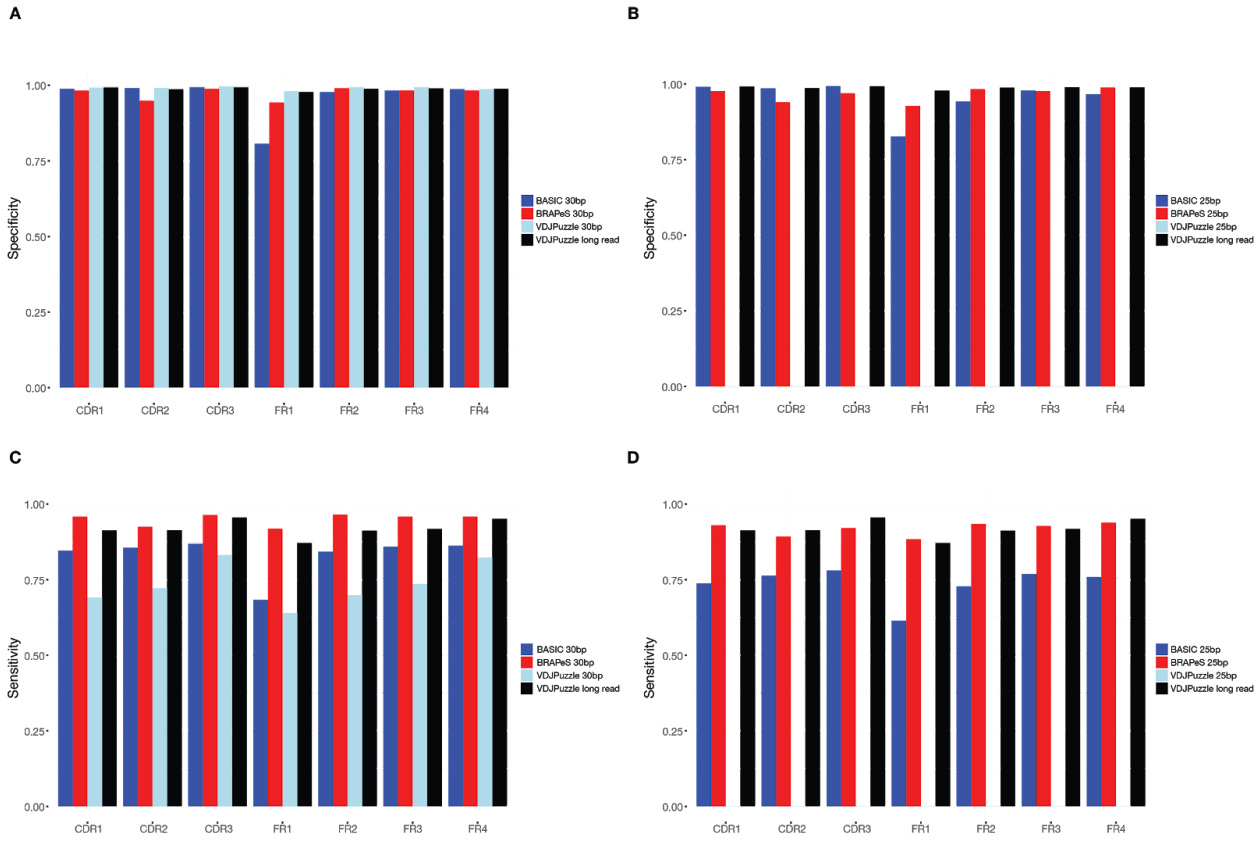


Figure 3: Sensitivity and specificity of BRAPeS. A) Specificity of BRAPeS for 30bp for each CDR and framework region. The fraction of chains with a sequence identical to the sequence reconstructed by BASIC on the long-read data for each region, using the following methods: VDJPuzzle when applied to the long-read data (black), BRAPeS (red), BASIC (dark blue) and VDJPuzzle (light blue) applied to a version of the data trimmed to 30bp. The fraction is calculated only for chains that had a productive reconstruction in both the long-read BASIC results and the other method. **B)** Specificity of BRAPeS for 25bp. Same as A, except the short-read version of the data was trimmed to include only the outer 25bp, instead of 30bp. **C)** Sensitivity of BRAPeS for 30bp for each CDR and framework region. Same as A, except the fraction is calculated out of all the chains that had a productive reconstruction when running BASIC on the long-read data. **D)** Sensitivity of BRAPeS for 25bp. Same as B, except the fraction is calculated out of all the chains that had a productive reconstruction when running BASIC on the long-read data.

References

Afik, S., Yates, K.B., Bi, K., Darko, S., Godec, J., Gerdemann, U., Swadling, L., Douek, D.C., Klenerman, P., Barnes, E.J., et al. (2017). Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. Nucleic Acids Res. *45*, e148.

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., and Lefranc, M.-P. (2012). IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. Immunome Res. *8*, 26.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Canzar, S., Neu, K.E., Tang, Q., Wilson, P.C., Khan, A.A., and Hancock, J. (2017). BASIC: BCR assembly from single cells. Bioinformatics *33*, 425–427.

Chhangawala, S., Rudy, G., Mason, C.E., and Rosenfeld, J.A. (2015). The impact of read length on quantification of differentially expressed genes and splice junction detection. Genome Biol. *16*, 131.

Di Noia, J.M., and Neuberger, M.S. (2007). Molecular mechanisms of antibody somatic hypermutation. Annu. Rev. Biochem. *76*, 1–22.

Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). SeqAn An efficient, generic C++ library for sequence analysis. BMC Bioinformatics *9*, 11.

Eltahla, A.A., Rizzetto, S., Pirozyan, M.R., Betz-Stablein, B.D., Venturi, V., Kedzierska, K., Lloyd, A.R., Bull, R.A., and Luciani, F. (2016). Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. Immunol. Cell Biol. *94*, 604–611.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Imkeller, K., and Wardemann, H. (2018). Assessing human B cell repertoire diversity and convergence. Immunol. Rev. *284*, 51–66.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Lefranc, M.-P. (2014). Immunoglobulin and T Cell Receptor Genes: IMGT(®) and the Birth and Rise of Immunoinformatics. Front. Immunol. *5*, 22.

Lefranc, M.-P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., Carillon, E., Duvergey, H., Houles, A., Paysan-Lafosse, T., et al. (2015). IMGT®, the international

ImMunoGeneTics information system® 25 years on. Nucleic Acids Res. *43*, D413–D422.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, S., Lefranc, M.-P., Miles, J.J., Alamyar, E., Giudicelli, V., Duroux, P., Freeman, J.D., Corbin, V.D.A., Scheerlinck, J.-P., Frohman, M.A., et al. (2013). IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat. Commun. *4*, 2333.

Lindeman, I., Emerton, G., Mamanova, L., Snir, O., Polanski, K., Qiao, S.-W., Sollid, L.M., Teichmann, S.A., and Stubbington, M.J.T. (2018). BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat. Methods *15*, 563–565.

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. *18*, 35–45.

Rizzetto, S., Eltahla, A.A., Lin, P., Bull, R., Lloyd, A.R., Ho, J.W.K., Venturi, V., and Luciani, F. (2017). Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. Sci. Rep. *7*, 12781.

Rizzetto, S., Koppstein, D.N.P., Samir, J., Singh, M., Reed, J.H., Cai, C.H., Lloyd, A.R., Eltahla, A.A., Goodnow, C.C., and Luciani, F. (2018). B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. Bioinformatics.

Stubbington, M.J.T., Lönnberg, T., Proserpio, V., Clare, S., Speak, A.O., Dougan, G., and Teichmann, S.A. (2016). T cell fate and clonality inference from single-cell transcriptomes. Nat. Methods *13*, 329–332.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575.

Villani, A.-C., Sarkizova, S., and Hacohen, N. (2018). Systems Immunology: Learning the Rules of the Immune System. Annu. Rev. Immunol. *36*, 813–842.

Wu, Y.L., Stubbington, M.J.T., Daly, M., Teichmann, S.A., and Rada, C. (2016). Intrinsic transcriptional heterogeneity in B cells controls early class switching to IgE. J. Exp. Med. jem.20161056.

Supplementary Information

Supplementary figures are found below. Supplementary tables can be found in the following link:
https://www.life-science-alliance.org/content/2/4/e201900371/tab-figures-data#fig-data-supplementary-materials

Supplementary figures

Figure S1

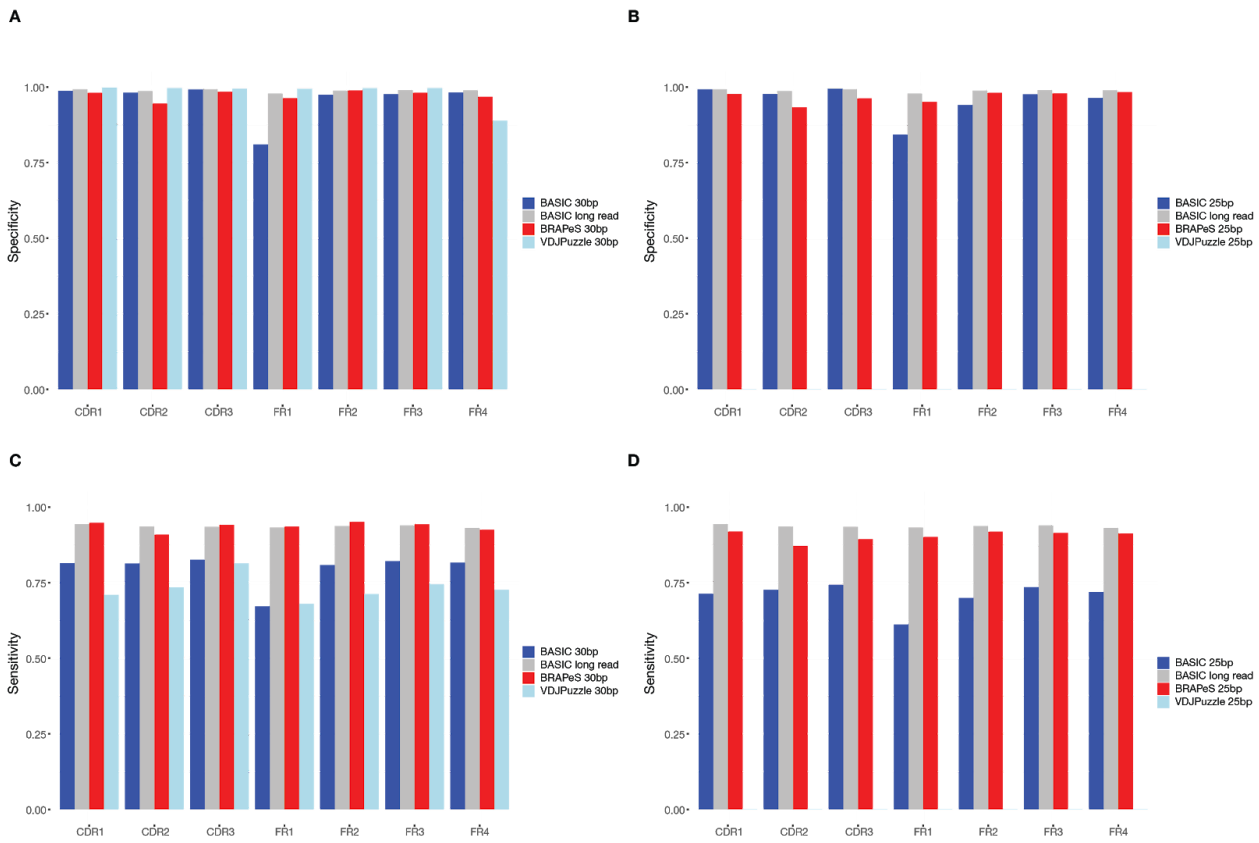


Figure S1: Sensitivity and specificity of BRAPeS compared to VDJPuzzle reconstructions on long-read data. **A**) Specificity of BRAPeS for 30bp for each CDR and framework region. The fraction of chains with a sequence identical to the sequence reconstructed by VDJPuzzle on the long-read data for each region, using the following methods: BASIC when applied to the long-read data (grey), BRAPeS (red), BASIC (dark blue) and VDJPuzzle (light blue) applied to a version of the data trimmed to 30bp. The fraction is calculated only for chains that had a productive reconstruction in both the long-read VDJPuzzle results and the other method. **B**) Specificity of BRAPeS for 25bp. Same as A, except the short-read version of the data was trimmed to include only the outer 25bp, instead of 30bp. **C**) Sensitivity of BRAPeS for 30bp for each CDR and framework region. Same as A, except the fraction is calculated out of all the chains that had a productive reconstruction when running VDJPuzzle on the long-read data. **D**)

Sensitivity of BRAPeS for 25bp. Same as B, except the fraction is calculated out of all the chains that had a productive reconstruction when running VDJPuzzle on the long-read data.

Figure S2

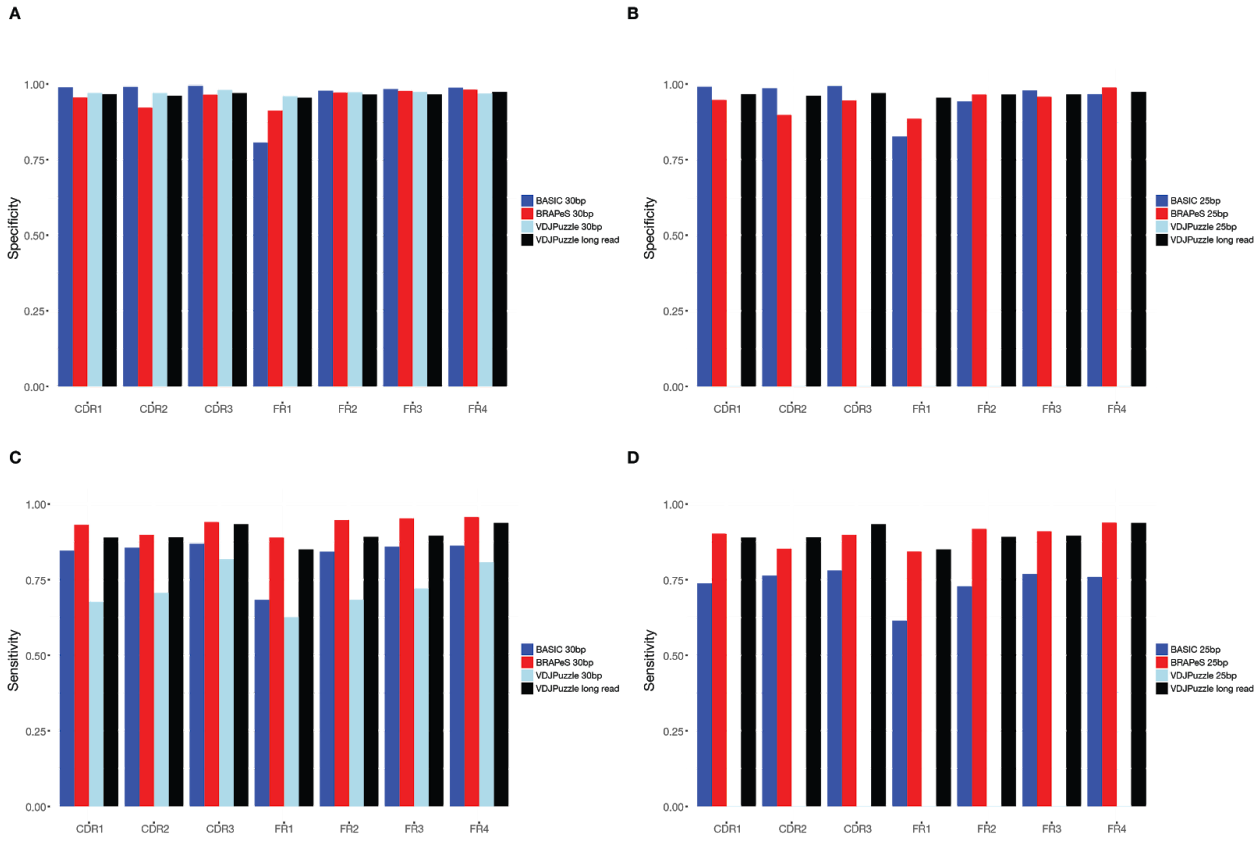


Figure S2: Sensitivity and specificity of the top-ranking reconstruction. A) Specificity of the top-ranking BRAPeS reconstruction for 30bp for each CDR and framework region. The fraction of chains where the top-ranking sequence is identical to the sequence reconstructed by BASIC on the long-read data for each region, using the following methods: VDJPuzzle when applied to the long-read data (black), BRAPeS (red), BASIC (dark blue) and VDJPuzzle (light blue) applied to a version of the data trimmed to 30bp. The fraction is calculated only for chains that had a productive reconstruction in both the long-read BASIC results and the other method. **B**) Specificity of top-ranking BRAPeS for 25bp. Same as A, except the short-read version of the data was trimmed to include only the outer 25bp, instead of 30bp. **C**) Sensitivity of the top-ranking BRAPeS reconstruction for 30bp for each CDR and framework region. Same as A, except the fraction is calculated out of all the chains that had a productive reconstruction when running BASIC on the long-read data. **D**) Sensitivity of BRAPeS for 25bp. Same as B, except the fraction is calculated out of all the chains that had a productive reconstruction when running BASIC on the long-read data.

Figure S3

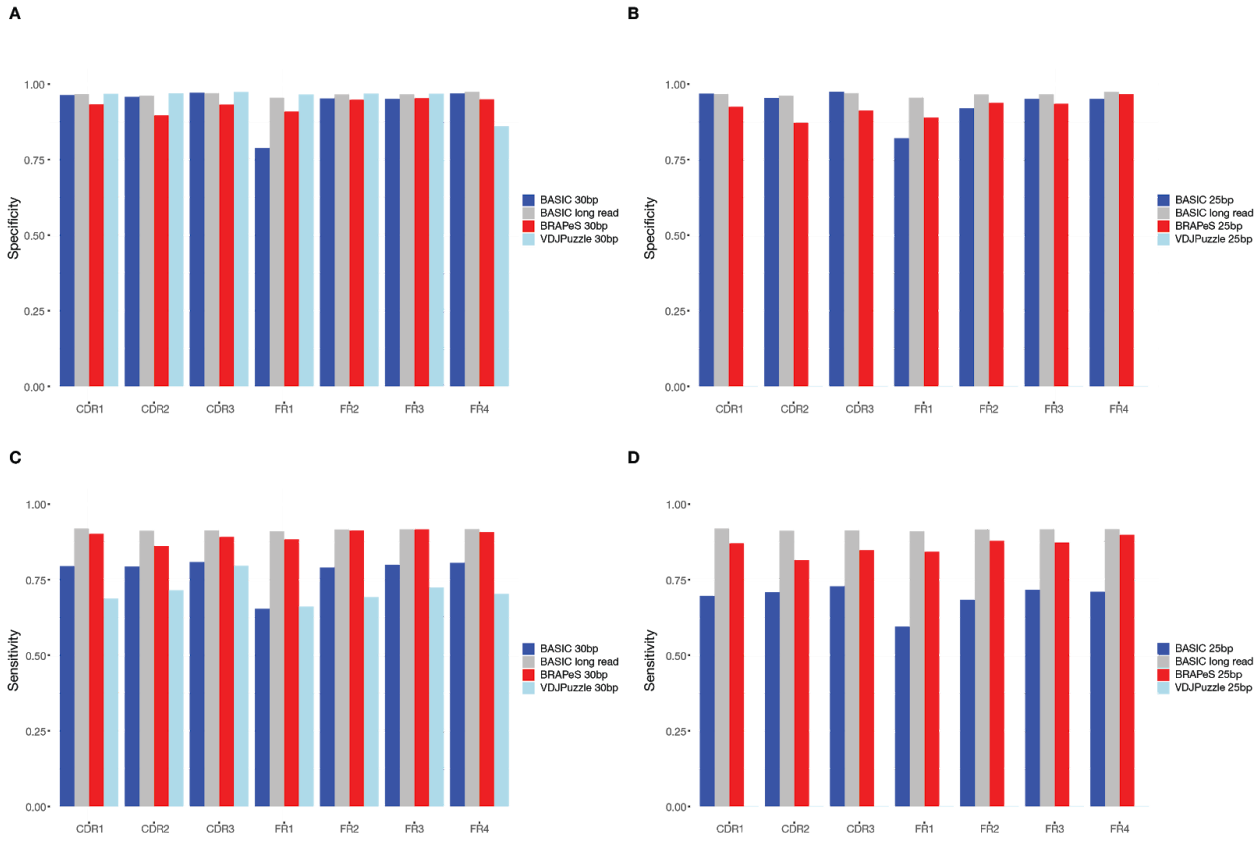


Figure S3: Sensitivity and specificity of the top-ranking reconstruction compared to the top-ranking VDJPuzzle reconstruction on long-read data. **A**) Specificity of the top-ranking BRAPeS reconstruction for 30bp for each CDR and framework region. The fraction of chains where the top-ranking sequence is identical to the top-ranking sequence reconstructed by VDJPuzzle on the long-read data for each region, using the following methods: BASIC when applied to the long read data (grey), BRAPeS (red), BASIC (dark blue) and VDJPuzzle (light blue) applied to a version of the data trimmed to 30bp. The fraction is calculated only for chains that had a productive reconstruction in both the long-read VDJPuzzle results and the other method. **B**) Specificity of the top-ranking BRAPeS reconstruction for 25bp. Same as A, except the short-read version of the data was trimmed to include only the outer 25bp, instead of 30bp. **C**) Sensitivity of the top-ranking BRAPeS reconstruction for 30bp for each CDR and framework region. Same as A, except the fraction is calculated out of all the chains that had a productive reconstruction when running VDJPuzzle on the long-read data. **D**) Sensitivity of BRAPeS for 25bp. Same as B, except the fraction is calculated out of all the chains that had a productive reconstruction when running VDJPuzzle on the long-read data.