nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-022-01461-y

Check for updates

NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells

Amy F. Chen[1], Benjamin Parks[1,2], Arwa S. Kathiria[1], Benjamin Ober-Reynolds[1], Jorg J. Goronzy[3,4] and William J. Greenleaf[1,5,6,7] ✉

In this work, we describe NEAT-seq (sequencing of nuclear protein epitope abundance, chromatin accessibility and the transcriptome in single cells), enabling interrogation of regulatory mechanisms spanning the central dogma. We apply this technique to profile CD4 memory T cells using a panel of master transcription factors (TFs) that drive T cell subsets and identify examples of TFs with regulatory activity gated by transcription, translation and regulation of chromatin binding. We also link a noncoding genome-wide association study single-nucleotide polymorphism (SNP) within a GATA motif to a putative target gene, using NEAT-seq data to internally validate SNP impact on GATA3 regulation.

Multimodal single-cell technologies have revolutionized our ability to characterize cell states and identify gene regulatory programs across various cell types. For example, methods pairing assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) and RNA-seq in single cells have allowed association of epigenetic status with transcriptional output, enabling identification of putative target genes of regulatory elements[1]. Antibodies linked to barcoded oligonucleotides have enabled surface protein measurements using a sequencing read-out, and when combined with RNA- or ATAC-seq in single cells have been particularly informative for profiling immune populations traditionally isolated based on surface protein markers[2,3]. Recently, this approach was extended to measuring intracellular proteins[4–8]. However, quantification of nuclear gene regulatory proteins along with chromatin accessibility profiling and RNA-seq has not been demonstrated.

Since transcription factors (TFs) can bind directly to enhancer elements to modulate target gene expression[9], measuring their abundance provides critical insight into how these proteins drive gene regulatory states. Protein abundance is often challenging to estimate via single-cell RNA-seq data due to posttranscriptional regulation and relatively low RNA capture rates[10,11]. Directly measuring single-cell protein levels of TFs, which are often much more abundant than their encoding transcripts[12], can link individual TFs to regulated enhancers and target genes by correlating changes in TF protein levels to changes in regulatory element accessibility and expression of nearby genes. Such analyses can help distinguish direct target genes from secondary effects, reveal cooperative and antagonistic effects of multiple TFs on gene regulation, and enable more accurate identification of TF-mediated gene regulatory networks underpinning cell fate. Here, we develop NEAT-seq (sequencing of nuclear protein epitope abundance, chromatin accessibility and the transcriptome in single cells), a method that enables quantification of nuclear proteins along with ATAC-seq and RNA-seq in single cells. We use NEAT-seq to profile CD4 memory T cells and illustrate its use for interrogating the relationship between master TF abundance, chromatin accessibility and gene expression.

While multiple groups have demonstrated sequencing-based surface protein quantification using barcoded antibodies[2,3,13,14], application to nuclear proteins has been challenging due to high levels of background oligo-antibody staining in the nucleus[8,15]. One approach to reduce nonspecific staining is to saturate cells with single stranded nucleic acids or other negatively charged polymers in an attempt to block cellular components that bind nonspecifically to single-stranded DNA[5,7,8,15]. We hypothesized that directly blocking the charge of the antibody oligo might further improve signal over background and that *E. coli* ssDNA binding protein (EcoSSB) would be an attractive candidate for this purpose (Supplementary Note 1).

To optimize staining with EcoSSB-bound oligo-conjugated antibodies, we conjugated an anti-green fluorescent protein (-GFP) antibody with a Cy5 labeled 80-bp ssDNA oligo, allowing us to compare oligo-antibody staining levels (via Cy5 fluorescence) to GFP levels within a cell. We then used this antibody to stain human embryonic kidney 293 (HEK293) cells expressing a nuclear-localized GFP. Preincubating the oligo-antibody with EcoSSB dramatically improved correlation between Cy5 antibody signal and GFP levels within the nucleus (Fig. 1a). We further confirmed that quantification of the conjugated oligo reflected GFP levels within the nucleus by sorting nuclei into three populations of increasing GFP abundance for quantitative PCR targeting the conjugated oligo (Extended Data Fig. 1a,b). To determine whether these staining conditions would be sufficiently sensitive to detect endogenous protein levels, we stained for GATA1 in K562 cells, using embryonic stem cells (ESCs) as a negative control. We observed a marked increase in GATA1 staining in K562 relative to ESCs using a GATA1 antibody linked to the Cy5-modified oligo (Extended Data Fig. 1c). EcoSSB similarly improved specificity of cytosolic protein staining (Extended Data Fig. 1d). In comparison with the inCITE-seq (intra-nuclear cellular indexing of transcriptomes and epitopes) protocol[8], our approach showed a moderate

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Computer Science, Stanford University School of Engineering, Stanford, CA, USA. [3]Department of Medicine, Palo Alto Veterans Administration Healthcare System, Palo Alto, CA, USA. [4]Division of Immunology and Rheumatology, Department of Medicine, Stanford University, Stanford, CA, USA. [5]Center for Personal Dynamic Regulomes, Stanford University School of Medicine, Stanford, CA, USA. [6]Department of Applied Physics, Stanford University, Stanford, CA, USA. [7]Chan-Zuckerberg Biohub, San Francisco, CA, USA. ✉e-mail: wjg@stanford.edu

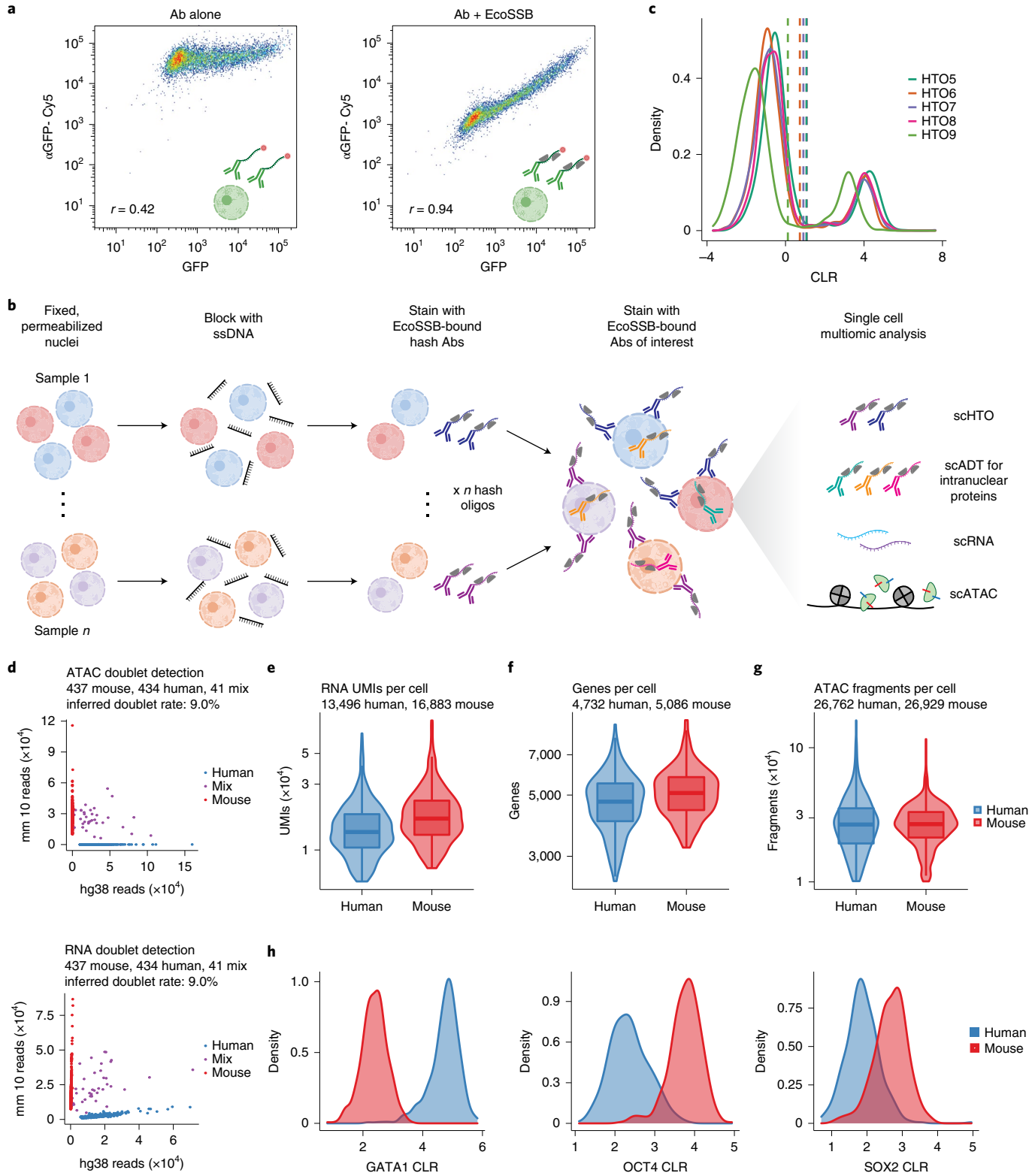


Fig. 1 | An intra-nuclear staining protocol using oligo-antibodies enables simultaneous profiling of nuclear proteins, chromatin accessibility and RNA transcripts in single cells. a, Flow cytometry plot of HEK293T cells expressing nuclear-localized GFP and stained with an anti-GFP antibody linked to an 80 bp ssDNA oligo with 3′-Cy5 modification. Spearman correlation is shown. **b**, Schematic of NEAT-seq workflow. **c**, Distribution of CLR transformed counts of HTOs from anti-NPC antibodies (HTO5-9). **d**, Scatterplot of number of reads mapping to the human versus mouse genome in each cell after removing HTO doublets, with each cell colored by its classification as a human cell, mouse cell or mixed species doublet. **e**–**g**, Distribution of RNA UMIs (**e**), genes (**f**) and ATAC-seq fragments (**g**) per cell. Boxplots show median with bounds of the box representing the 25th and 75th percentiles and the whiskers extending to the value closest to but not exceeding 1.5 times the interquartile range. Outliers extending beyond the whiskers are not shown. **h**, CLR-transformed counts of ADTs corresponding to GATA1, OCT4 and SOX2 in cells classified as human or mouse cells based on mapping of ATAC-seq and RNA-seq reads to each genome.

improvement in staining that was most pronounced at lower protein levels (Supplementary Note 2 and Extended Data Fig. 1e).

We next sought to combine nuclear protein quantification with scATAC-seq and scRNA-seq using the 10X Genomics Multiome kit (Fig. 1b). Because cell fixation increases the doublet rate in droplet-based single-cell technologies[4], we assayed a mixture of human K562 and mouse ESCs stained using antibodies against mutually exclusive TFs and antibody-linked hashtag oligos[14] (HTOs) to assess both doublet identification and specificity of nuclear protein abundance measurements. HTOs are barcoded oligos used to label distinct samples so they can be distinguished after pooling. Clear separation between positive and negative HTO staining allowed us to set stringent cutoffs and identify droplets that were positive for more than one hashing oligo, enabling in silico removal of almost 50% of doublets (Fig. 1c,d and Extended Data Fig. 2a,d). This hashing procedure can similarly be applied for straightforward sample multiplexing (Fig. 1b).

For human and mouse singlets, we detected a median of 13,496 and 16,883 RNA unique molecular identifiers (UMIs); 4,732 and 5,086 genes; and 26,762 and 26,929 ATAC fragments per cell, respectively (Fig. 1e–g). The fragment length distribution and average transcription start site (TSS) enrichment were also comparable to bulk ATAC-seq libraries (Extended Data Fig. 2e,f). Tn5 insertions in peaks were highly correlated with bulk data ($r=0.90$) and RNA-seq data were moderately correlated ($r=0.67$), with a similar RNA correlation value observed for standard 10X Multiome data (Extended Data Fig. 2g,h). We observed a greater than four-fold enrichment of GATA1 and OCT4 antibody-derived tag (ADT) counts and greater than two-fold enrichment of SOX2 ADT counts in their respective cell types (Fig. 1h). Together, these results demonstrate simultaneous quantification of endogenous nuclear protein abundance, chromatin accessibility and gene expression in fixed single cells.

We next applied NEAT-seq to profile primary human CD4 memory T cells composed of distinct T cell subsets driven by known master TFs, providing a diverse system for dissecting the regulatory mechanisms upstream and downstream of these TFs[16]. Our antibody panel targeted TFs that drive Th1 (Tbet), Th2 (GATA3), Th17 (RORγT) and Treg (FOXP3 and Helios) cell fate[17]. In this dataset, we observed a small reduction in unique ATAC and RNA reads detected relative to a standard 10X Multiome experiment, which was similarly observed in other fixed single-cell assays (Supplementary Tables 1–3).

We identified seven clusters in the population using scATAC-seq, which largely corresponded to clusters identified using scRNA-seq (Fig. 2a and Extended Data Fig. 3a). We annotated the Th1, Th2, Th17 and Treg clusters based on master TF RNA and protein abundance, genome-wide accessibility of the TF binding motif, as well as canonical surface marker expression[16] (Fig. 2b,c and Extended Data Fig. 3b). These clusters also exhibited high chromatin accessibility at functionally relevant cytokine gene loci, but low or undetectable RNA expression (Extended Data Fig. 3c,d). This observation is suggestive of epigenetic priming, where transcription is absent but the gene locus is accessible and poised for transcriptional activation, and is consistent with the primed status of memory T cells[18]. Remaining clusters were similarly annotated based on marker expression and TF motif accessibility (Supplementary Note 3 and Extended Data Fig. 3e,f).

Our antibody-based protein measurements for each TF showed clear enrichment in the cell type that the TF is known to drive and provided more robust detection of target TFs compared to our RNA data (Fig. 2c and Extended Data Fig. 4a–c): smoothing of signal across neighboring cells in the uniform manifold approximation and projection (UMAP) was necessary for identification of cell types using RNA-seq data due to high dropout rates, while unsmoothed ADT data were sufficient to clearly label cell types (Extended Data Fig. 4a,d,e).

By comparing the TF gene locus chromatin accessibility, RNA expression, protein abundance and genome-wide TF binding motif accessibility across cells for each TF assayed, our data indicate three distinct modes of regulation in our TF panel. Accessibility at the *RORγT* and *Tbet* genes were strongly correlated with measurements of downstream regulatory events, suggesting that these TFs are regulated transcriptionally (Fig. 2c and Extended Data Fig. 5a,b). FOXP3 and Helios exhibited strong correlation between gene accessibility, RNA and protein abundance but had differing patterns of motif accessibility, suggesting that their expression is regulated transcriptionally but presence of the protein does not result in increased chromatin accessibility at motif sites. The lack of concordance between FOXP3 expression and motif accessibility is consistent with previous studies showing that FOXP3 binds to pre-existing accessible enhancers to drive Treg fate[19]. For Helios, binding may result in chromatin compaction rather than accessibility, as was recently observed in mouse hematopoietic progenitor cells[20]. The uncoupling of TF protein expression and motif accessibility highlights the caveats of using motif accessibility alone to infer TF activity.

The final TF in our panel, GATA3, showed clear discordance between RNA expression and protein levels across cells (Fig. 2c). The ADT levels, but not RNA levels, were correlated with global changes in GATA3 motif accessibility. These observations are consistent with posttranscriptional regulatory mechanisms restricting GATA3 protein expression in memory T cells, which could only be uncovered with the addition of protein quantification.

Our paired RNA and protein measurements also allowed us to identify candidate posttranscriptional regulators of GATA3 by performing differential expression analysis between cells expressing high GATA3 RNA but low protein versus cells expressing both high GATA3 RNA and protein (Fig. 2d). Among the top upregulated genes were several core translation regulators, including the elongation factor *EEF1G*, large ribosome subunit *RPL18* and poly-A binding protein *PABPC4*, along with more indirect regulators such as *GAB2* and *NIBAN1* (ref. [21]) (Fig. 2e and Extended Data Fig. 6b). GATA3 translation is regulated by PI3K signaling through mTOR[22] that, like NIBAN1, phosphorylates EIF4EBP1 to allow assembly of the initiation complex[23], while GAB2 is a direct activator of PI3K. These results indicate upregulation of genes that promote translation may play a role in driving GATA3 protein production in the Th2 subset of memory T cells. Together, our results indicate three regulatory mechanisms used to modulate activity of the TFs in our panel: transcriptional regulation, as demonstrated by concordant RNA, protein and motif accessibility patterns (RORγT and Tbet); transcriptional regulation of expression but requirement of other TFs for chromatin binding (FOXP3) and translational regulation (GATA3).

In addition to using multimodal measurements to interrogate regulation of the TF itself, we can use this information to uncover downstream enhancer and gene targets of a TF by correlating protein abundance of the TF with changes in regulatory element accessibility and gene expression across cells. For RORγT, Tbet and GATA3, the correlated scATAC-seq peaks were enriched for the corresponding TF motif (Extended Data Fig. 7a). For FOXP3 and Helios, motifs were not enriched in correlated peaks, consistent with our earlier observations that these TFs are not correlated with global changes in motif accessibility (Fig. 2c). We similarly identified genes with RNA expression significantly correlated (adjusted $P<0.05$) with protein levels of each TF (Extended Data Fig. 7b). Within these correlated gene sets were genes known to be enriched or functionally important in the memory T cell subset driven by the TF in question, such as IL4R for GATA3 and CTLA4 for both Helios and FOXP3 (ref. [24]).

To identify candidate genes directly regulated by each TF through a TF-associated enhancer (that is, TF-peak–gene linkages), we overlapped the top TF ADT-correlated genes with top TF

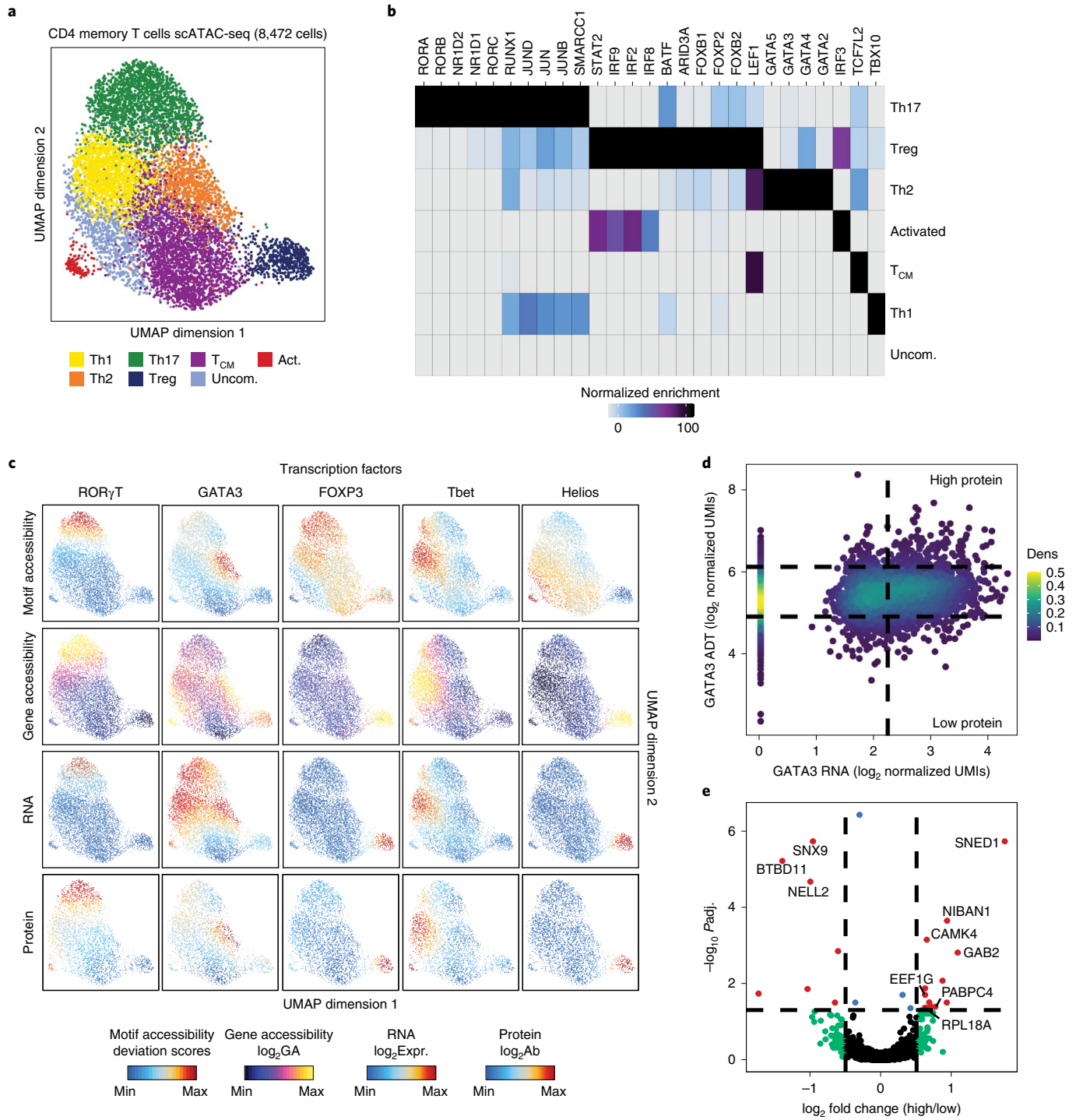


Fig. 2 | Profiling of CD4 memory T cells using NEAT-seq reveals modes of regulation across the central dogma. a, scATAC-seq UMAP of CD4 memory T cells with cell type classifications. Annotations with cell numbers in parentheses: Th1 (1562), Th2 (939), Th17 (1,855) and Treg (583); T_{CM}, central memory (2,512); Act., recently activated cells (116) and Uncom., uncommitted memory cells (905). **b**, Top enriched motifs in peaks that are more accessible in each cluster. **c**, Plots on scATAC UMAP of TF chromVAR deviations (motif accessibility), accessibility surrounding the TF gene locus (gene accessibility), RNA levels and protein levels as measured by ADTs for the indicated TFs. $n = 8,472$ cells in each plot except for ADTs, where $n = 3,841$ cells. GA, gene accessibility; Expr., RNA expression; Ab, antibody-based protein counts. **d**, Scatterplot of $\log_2$-transformed, normalized RNA versus ADT counts for GATA3 with cutoffs shown for high RNA, high protein and low protein indicated. Dens., density of cells. **e**, Differentially expressed genes between cells with high RNA and high protein versus high RNA and low protein for GATA3 based on a two-sided Wilcoxon rank sum test. Adjusted *P* values indicate Benjamini–Hochberg corrected values. Points in red represent genes with adjusted *P* value <0.05 and $\log_2$ fold change >0.5.

ADT-correlated scATAC-seq peaks containing the corresponding TF motif that were within 100 kb of the gene promoter and filtered for significant peak–gene linkages (adjusted $P < 0.05$; Fig. 3a). We performed this analysis for the TFs that showed correlation between TF abundance and motif accessibility (Fig. 3b). The candidate direct target genes were significantly enriched (adjusted $P < 0.05$) for Gene

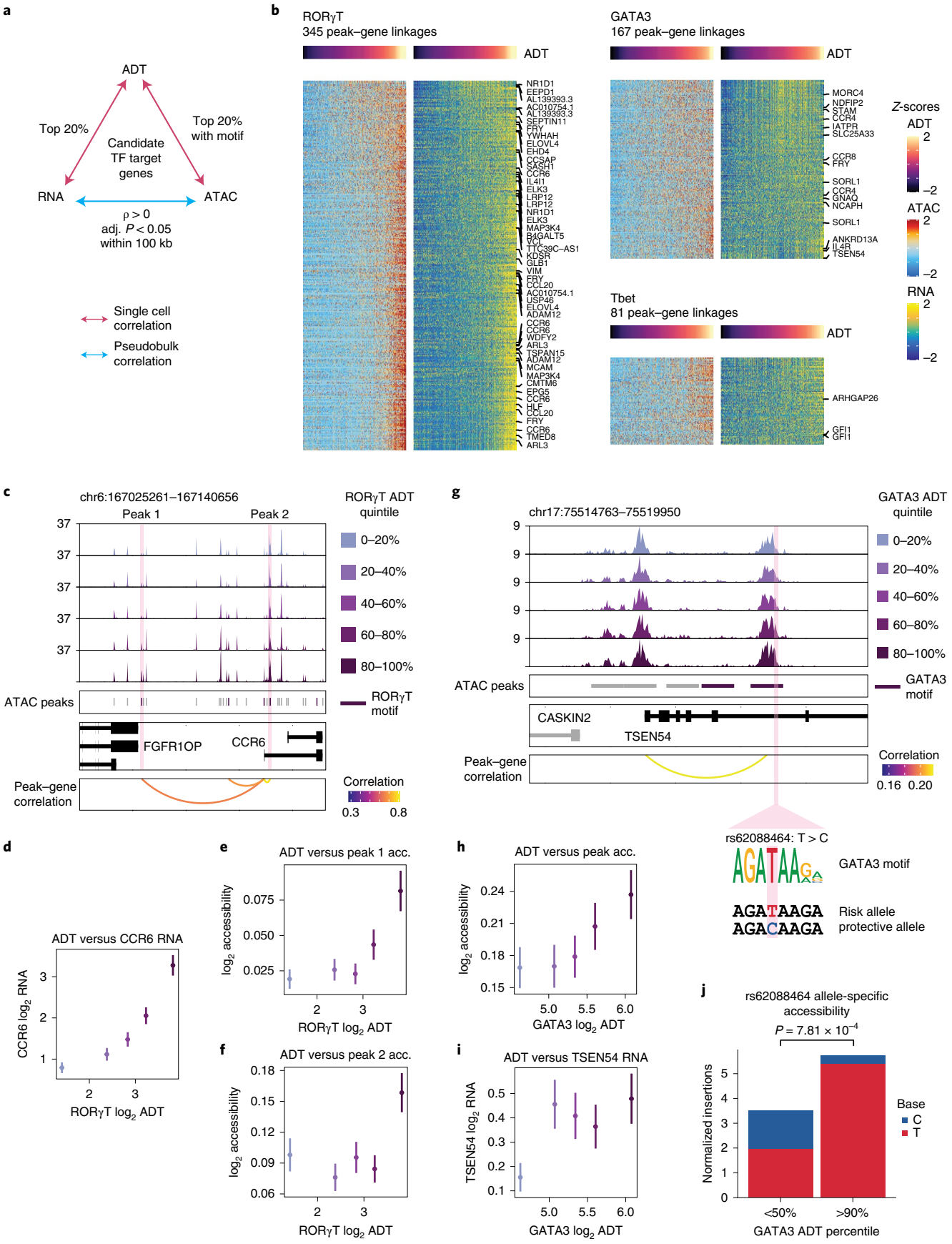

a
ADT
Top 20%
Top 20% with motif
Candidate TF target genes
RNA
ATAC
ρ > 0
adj. P < 0.05
within 100 kb
Single cell correlation
Pseudobulk correlation
b
RORγT
345 peak–gene linkages
ADT
GATA3
167 peak–gene linkages
ADT
Tbet
81 peak–gene linkages
ADT
Z-scores
ADT
ATAC
RNA
c
chr6:167025261–167140656
Peak 1
Peak 2
RORγT ADT quintile
0–20%
20–40%
40–60%
60–80%
80–100%
ATAC peaks
RORγT motif
FGFR1OP
CCR6
Peak–gene correlation
Correlation
0.3
0.8
g
chr17:75514763–75519950
GATA3 ADT quintile
ATAC peaks
GATA3 motif
CASKIN2
TSEN54
Peak–gene correlation
Correlation
0.16
0.20
rs62088464: T > C
GATA3 motif
Risk allele
protective allele
d
ADT versus CCR6 RNA
CCR6 log2 RNA
RORγT log2 ADT
e
ADT versus peak 1 acc.
log2 accessibility
RORγT log2 ADT
f
ADT versus peak 2 acc.
log2 accessibility
RORγT log2 ADT
h
ADT versus peak acc.
log2 accessibility
GATA3 log2 ADT
i
ADT versus TSEN54 RNA
TSEN54 log2 RNA
GATA3 log2 ADT
j
rs62088464 allele-specific accessibility
P = 7.81 × 10−4
Normalized insertions
Base
C
T
<50%
>90%
GATA3 ADT percentile

Fig. 3 | Identification of peak–gene linkages associated with master TF protein expression. a, Diagram showing criteria used to identify candidate TF-peak–gene linkages for each TF. **b**, Heatmaps of peak–gene linkages correlated with abundance of the indicated TF identified based on criteria in **a**. Genes that are in the top 200 genes significantly enriched in the T cell subset driven by the corresponding TF based on data by Schmiedel et al.[24] are labeled. **c**, CCR6 ATAC-seq tracks in CD4 memory cells separated into quintiles by RORγT ADT levels, along with significantly correlated peak–gene linkages (adjusted $P < 0.05$). Spearman correlations are shown. Peaks containing a RORγT motif are indicated. **d**–**f**, CCR6 RNA expression (**d**) and accessibility (acc.) at the highlighted RORγT motif-containing peaks (1, **e** and 2, **f**) as a function of RORγT ADT levels. Mean is shown with standard error of the mean of $n = 768 \pm 1$ cells per group. **g**, TSEN54 ATAC-seq tracks as in **c**, but for GATA3 ADT, with the indicated GWAS SNP location highlighted. **h**,**i**, Accessibility at the highlighted SNP-containing peak (**h**) and TSEN54 RNA expression (**i**) as a function of GATA3 ADT levels. Mean is shown with standard error of the mean, *n* means approximately 768 cells per group. **j**, Read-normalized Tn5 insertions per cell per million reads mapping to the risk versus protective allele in cells above the 90th percentile versus cells below the 50th percentile of GATA3 ADT levels. *P* value is calculated using a one-sided binomial test on T:C ratio of insertion counts.

Ontology terms related to T cell function and included canonical surface markers for the corresponding cell type. Among the GATA3 targets were Th2 markers CCR4, CCR8 and IL4R, and among RORγT targets was the Th17 marker, CCR6 (Fig. 3c–f and Extended Data Figs. 7c–e and 8).

We also reasoned that these TF-peak–gene linkages could be used to interpret the effects of noncoding genome-wide association study (GWAS) single-nucleotide polymorphisms (SNPs) on TF activity and connect the SNPs to putative target genes. We overlapped peaks in our TF-peak–gene linkages with candidate causal GWAS SNPs[25] and identified rs62088464 located within a GATA motif sequence in a GATA3 ADT-associated peak. The risk allele, which preserves the GATA motif, is associated with decreased pulmonary function[26], which can result from inflammatory lung diseases associated with Th2 immune responses[27]. The gene linked to the peak containing this SNP encodes *TSEN54*, a gene with significantly enriched expression in the sputum of patients with type-2 airway inflammation[28,29] (Fig. 3g–i). Since our T cell donor was heterozygous for this SNP, we tested whether the risk allele is regulated by GATA3 by examining whether the risk allele is more accessible than the protective allele in cells with high GATA3 protein levels. Indeed, we observed that most ATAC-seq reads in cells expressing high GATA3 ADT levels mapped to the risk allele, while little difference was observed in cells with lower GATA3 ADT levels (Fig. 3j). Similarly, the risk allele is associated with increased *TSEN54* expression in GTEx data and *TSEN54* was the gene most strongly associated with the risk allele in various tissues (Extended Data Fig. 9). Together, these results indicate that GATA3 binds the risk allele sequence to activate the regulatory element and drive expression of *TSEN54* and that this binding is disrupted with the protective allele.

NEAT-seq provides a new avenue for studying the quantitative effects of epigenetic regulator abundance on both chromatin and gene expression state in primary human samples. Whereas previous studies investigating dosage-dependent effects of TFs often required building cell lines with a combination of hypomorphic and null alleles[30,31] or inducible expression systems[32], we demonstrate that NEAT-seq can measure the molecular consequences of continuous changes in TF levels in a biologically relevant setting for a panel of proteins simultaneously. Since nuclear proteins encompass many proteins involved in gene regulation, the capacity to link nuclear protein levels to epigenetic and transcriptional status provides a powerful approach for studying mechanisms of gene regulation. Incorporating additional modalities such as cytoplasmic and cell surface proteins, CRISPR guide RNA sequencing and T cell receptor sequencing will enable measurement of the effects of cellular perturbations and signaling pathways on cell state, providing an even more comprehensive picture of cellular programs.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01461-y.

Received: 11 July 2021; Accepted: 20 March 2022;
Published online: 2 May 2022

References

1. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
2. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
3. Swanson, E. et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *eLife* https://doi.org/10.7554/eLife.63632 (2021).
4. Mimitou, E. P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-00927-2 (2021).
5. Gerlach, J. P. et al. Combined quantification of intracellular (phospho-) proteins and transcriptomics from fixed single cells. *Sci. Rep.* **9**, 1469 (2019).
6. Reimegård, J. et al. A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun. Biol.* **4**, 624 (2021).
7. Rivello, F. et al. Single-cell intracellular epitope and transcript detection revealing signal transduction dynamics. *Cell. Rep. Meth.* **1**, 100070 (2021).
8. Chung, H. et al. Joint single-cell measurements of nuclear proteins and RNA in vivo. *Nat. Methods* **18**, 1204–1212 (2021).
9. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
10. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
11. Marinov, G. K. et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496–510 (2014).
12. Gillespie, M. A. et al. Absolute quantification of transcription factors reveals principles of gene regulation in erythropoiesis. *Mol. Cell* **78**, 960–974.e11 (2020).
13. Mimitou, E. P. et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
14. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
15. Wang, Y. et al. Multiplexed in situ protein imaging using DNA-barcoded antibodies with extended hybridization chain reactions. Preprint at *bioRxiv* https://doi.org/10.1101/274456 (2021).
16. Sallusto, F. & Lanzavecchia, A. Heterogeneity of CD4$^+$ memory T cells: functional modules for tailored immunity. *Eur. J. Immunol.* **39**, 2076–2082 (2009).
17. Fang, D. & Zhu, J. Dynamic balance between master transcription factors determines the fates and functions of CD4 T cell and innate lymphoid cell subsets. *J. Exp. Med.* **214**, 1861–1876 (2017).
18. Barski, A. et al. Rapid recall ability of memory T cells is encoded in their epigenome. *Sci. Rep.* **7**, 39785 (2017).
19. Samstein, R. M. et al. Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153–166 (2012).
20. Cova, G. et al. Helios represses megakaryocyte priming in hematopoietic stem and progenitor cells. *J. Exp. Med.* **218**, e20202317 (2021).
21. Sun, G. D. et al. The endoplasmic reticulum stress-inducible protein Niban regulates eIF2alpha and S6K1/4E-BP1 phosphorylation. *Biochem. Biophys. Res. Commun.* **360**, 181–187 (2007).

22. Cook, K. D. & Miller, J. TCR-dependent translational control of GATA-3 enhances Th2 differentiation. *J. Immunol.* **185**, 3209–3216 (2010).
23. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731–745 (2009).
24. Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715.e16 (2018).
25. Taylor, K. E., Mark Ansel, K., Marson, A., Criswell, L. A. & Farh, K. K.-H. PICS2: next-generation fine mapping via probabilistic identification of causal SNPs. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab122 (2021).
26. Alkes Group, UKBB summary statistics, https://alkesgroup.broadinstitute.org/ (2020).
27. Gieseck, R. L., Wilson, M. S. & Wynn, T. A. Type 2 immunity in tissue repair and fibrosis. *Nat. Rev. Immunol.* **18**, 62–76 (2018).
28. Peters, M. C. et al. A transcriptomic method to determine airway immune dysfunction in T2-high and T2-low asthma. *Am. J. Respir. Crit. Care Med.* **199**, 465–477 (2019).
29. Singh, D. et al. COPD patients with chronic bronchitis and higher sputum eosinophil counts show increased type-2 and PDE4 gene expression in sputum. *J. Cell. Mol. Med.* **25**, 905–918 (2021).
30. Affar, E. B. et al. Essential dosage-dependent functions of the transcription factor yin yang 1 in late embryonic development and cell cycle progression. *Mol. Cell. Biol.* **26**, 3565–3581 (2006).
31. Takeuchi, J. K. et al. Chromatin remodelling complex dosage modulates transcription factor function in heart development. *Nat. Commun.* **2**, 187 (2011).
32. Sokolik, C. et al. Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. *Cell Syst.* **1**, 117–129 (2015).

Methods

Cell culture. Mouse V6.5 ESCs were obtained from Novus Biologicals (NBP1-41162) and cultured on gelatin-coated plates in Knockout DMEM (Thermo Fisher no. 10829018) supplemented with 7.5% ES-qualified serum (Applied Stem Cell no. ASM-5017), 7.5% Knockout Serum replacement (Thermo Fisher no. 10828-028), 2 mM L-glutamine (Gibco no. 35050061), 10 mM HEPES (Gibco no. 15630080), 100 units per ml penicillin/streptomycin (Gibco no. 151401222), 0.1 mM nonessential amino acids (Gibco no. 11140050), 0.1 mM beta-mercaptoethanol (Gibco no. 21985023) and leukemia inhibitory factor. The human chronic myeloid leukemia cell line, K562, was purchased from ATCC and cultured in RPMI 1640 medium (Gibco no. 11875119) containing 15% FBS and 100 units per ml penicillin/streptomycin. HEK293T cells were cultured in DMEM with GlutaMAX (Gibco no. 10566024) containing 10% FBS and 100 units per ml penicillin/streptomycin. Frozen vials of primary human $CD4^+CD45RO^+$ memory T cells were purchased from STEMCELL Technologies (catalog no. 70031) that were obtained from donors using IRB-approved consent forms and protocols.

GFP transfection, staining and sorting. On day 0, HEK293T cells were seeded at 4 million cells per 10 cm plate. On day 1, cells were transfected with 6 µg nuclear-localized GFP construct (Addgene no. 67652) using Fugene HD transfection reagent (Promega). Cells were gathered and stained using anti-GFP antibody (Biolegend 338002) linked to an 80 or 100 bp ssDNA oligo with 3′ Cy5 fluorophore as described in the oligo-antibody staining methods section, except without RNase inhibitor or DTT. A control stain was performed with the oligo-antibody in the absence of SSB. Stained cells were resuspended in PBS and analyzed on an LSRII flow cytometer or sorted on a BD FACS Aria II. FlowJo v.10.7.1 was used for analysis of flow cytometry data.

Antibody conjugation. Antibodies were conjugated with streptavidin using the Lightning-Link Streptavidin Conjugation Kit from Abcam (ab102921) according to the manufacturer's instructions. NaCl and Tween were added to the conjugated antibody mixture to a final concentration of 0.5 M NaCl and 0.01% Tween and mixed with biotinylated oligos (purchased from IDT) at equimolar ratio. The mixture was incubated overnight at room temperature and unbound oligo was removed using Amicon 100-KDa centrifugal filters (UFC510008). Antibody conjugates were eluted and stored in PBS. Antibodies conjugated with streptavidin were GATA1 (Abcam ab241393), OCT4 (R&D AF1759), SOX2 (R&D MAB2018), nuclear pore complex (Biolegend 902901) and GFP (Biolegend 338002). Due to a low observed enrichment of SOX2 in mESCs versus K562, we tested specificity of the conjugated SOX2 antibody by western blot and observed specific, if relatively weak, SOX2 binding (Extended Data Fig. 2i). Antibodies in the TF panel for CD4 memory T cells were directly conjugated to oligos by BD Biosciences. The antibodies in the panel were the following clones from BD Biosciences: GATA3 (L50-823), Tbet (4B10), RORγT (Q21-559), FOXP3 (259D/C7) and Helios (22F6). Due to discordance between GATA3 protein and RNA levels, we verified specificity of our GATA3 antibody on GATA3-overexpressing cells (Extended Data Fig. 6a).

Binding of ssDNA binding protein to oligo-antibodies. To bind EcoSSB (Promega M3011) to the antibody oligos, we incubated the antibody and EcoSSB in 50 µl of 1× NEBuffer 4 for 30 min at 37 °C. We then added a final concentration of 3% BSA, 1× PBS and 1 U µl^{-1} RNase inhibitor directly to the antibody-EcoSSB mix (without any purification) in a final volume of 100 µl for staining cells. To calculate the amount of EcoSSB needed to saturate binding sites on the antibody oligos, we estimated that each antibody was conjugated to an average of two oligos of 95 bp, and each EcoSSB tetramer would bind with a roughly 35-bp footprint[33,34], requiring six EcoSSB tetramers per antibody. Based on the concentration of antibody being used and reported K_d of EcoSSB (in the roughly 2 nM range)[35], we can then estimate the amount of EcoSSB necessary to bind a given fraction of oligos (aiming for >0.9) using the following equation:

$$\text{desired fraction of oligo bound} = \frac{\left([\text{EcoSSB}]_{\text{tot}}+[\text{oligo}]_{\text{tot}}+K_d\right)-\sqrt{\left([\text{EcoSSB}]_{\text{tot}}+[\text{oligo}]_{\text{tot}}+K_d\right)^2-4\times[\text{EcoSSB}]_{\text{tot}}\times[\text{oligo}]_{\text{tot}}}}{2\times[\text{oligo}]_{\text{tot}}}$$

where $[\text{oligo}]_{\text{tot}}$ = antibody concentration × 2 oligos × 3 EcoSSB binding sites per oligo.

Oligo-antibody staining. Cells were fixed in 1.6% formaldehyde in PBS for 2 min at room temperature, then quenched with 0.25 M glycine for 5 min on ice and spun down at 600*g* for 5 min. Cells were washed twice with PBS and then resuspended in lysis/permeabilization buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 3 mM $MgCl_2$, 0.5% NP40, 0.1% Tween-20, 0.01% digitonin, 1 U µl^{-1} RNase inhibitor, 1 mM DTT). For staining of cytosolic GFP, 0.1% NP40 was used in the buffer instead. Cells were incubated on ice for 10 min, pelleted at 600*g* for 5 min and washed twice with wash buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20, 1 U µl^{-1} RNase inhibitor, 1 mM DTT). Cells were incubated in staining buffer (PBS with 3% BSA, 1 U µl^{-1} RNase inhibitor) with 1 mM DTT and 1 mg ml^{-1} of ssDNA for 30 min at room temperature, pipetting often to resuspend cells. For the flow cytometry experiments involving GFP staining, salmon sperm DNA was used for the ssDNA block. However, due to significant amounts of annealing to form double stranded DNA that would result in contaminating reads in ATAC-seq data, we switched to using either a mixture of random 30-mers or a 30 bp ssDNA oligo sequence with no complementarity to the mouse or human genome for multiome experiments. To ensure no priming would occur with these oligos, they were modified with a terminal dideoxy cytosine.

After blocking with ssDNA, Tween was added to a final concentration of 0.1% and cells were pelleted and washed once with staining buffer + 0.1% Tween. Cells were then split into five tubes and each tube of cells was incubated with an anti-NPC antibody linked to a distinct HTO (prebound with SSB) for 30 min at room temperature. Cells were washed twice with staining buffer + 0.1% Tween, repooled and incubated with TF antibody mix for 30 min at room temperature. For the CD4 memory T cell experiment, cells were split into two tubes before incubating with two concentrations of the TF antibody mix. A distinct hashing antibody was also added to the two TF antibody mixes to identify the concentration of antibody that each cell was stained with. Cells were then washed twice with staining buffer + 0.1% Tween, and cells incubated with different concentrations of TF antibody were pooled. Cells were washed once more with PBS containing 1% BSA and 1 U µl^{-1} RNase inhibitor, then resuspended in 1× Nuclei buffer containing 1 U µl^{-1} RNase inhibitor from the 10X Genomics Multiome kit. The cell suspension was then filtered through a 40-µm Flowmi strainer 2–3 times until nuclei clusters were removed.

inCITE-seq staining conditions were performed as described in Chung et al.[8]. For NEAT-seq fixation and permeabilization followed by staining using inCITE-seq staining conditions, we performed fixation and permeabilization as described above and then proceeded with the dextran sulfate blocking and staining conditions (1:100 FcX (BioLegend 156604) + 1% BSA + 0.05% Dextran Sulfate) used by inCITE-seq.

Antibody concentrations. The NPC, GATA1, SOX2 and OCT4 antibodies were all used at 0.3 µg in 100 µl of staining buffer (3 µg ml^{-1}). The two antibody concentrations for TF antibodies used in the CD4 memory T cell experiment are indicated below:

Antibody	Concentration 1 (µg ml^{-1}; marked by NPC1)	Concentration 2 (µg ml^{-1}; marked by NPC2)
RORγT	0.39075	1.95375
Foxp3	2.5	5
GATA3	3.125	15.625
Helios	0.39075	1.95375
Tbet	3.125	15.625

Both antibody concentrations showed specific staining of the targeted TF in the appropriate cell type, as shown in Extended Data Fig. 4b,c. We chose concentration 2 for follow-up analyses since it provided slightly better enrichment over background for some antibodies.

Single-cell library preparation and sequencing. Antibody-stained cells in 1× Nuclei buffer were processed using the 10X Genomics Multiome kit as indicated in the standard protocol (Rev A) to generate ATAC-seq and RNA-seq libraries. For the barnyard experiment, 1,500 cells were targeted in one lane of the chip. For the CD4 memory T cell experiment, 6,000 cells were targeted per lane and two lanes were used. During the preamplification step, Truseq read 2 (CAGACGTGTGCTCTTCCGATC) and Nextera read 2 (GGCTCGGAGATGTGTATAAGAGACAG) primers were spiked in at 0.2 µM final concentration to amplify ADT and HTO oligos. To generate ADT and HTO libraries, 35 µl of preamplification product from step 4.3p was amplified with indexing primers using 2× NEB Next High-Fidelity PCR Master Mix (M0541). A double-sided solid-phase reversible immobilization (SPRI) bead clean up was performed using 0.6× SPRI beads (retaining supernatant) and then adding additional SPRI beads to a final concentration of 1.2×, washing with 80% ethanol and eluting ADT or HTO libraries from beads using EB buffer. Libraries were quantified by PCR using a PhiX control v.3 (Illumina FC-110–3001) standard curve. scATAC-seq libraries were sequenced alone on a NextSeq 550 sequencer and ADT libraries were either sequenced alone on a MiSeq (for the barnyard experiment) or together with scRNA-seq libraries on a NextSeq 550 (for the CD4 memory T cell experiment). Recommended sequencing read configurations for 10X Multiome libraries were used for scATAC- and scRNA-seq libraries. For sequencing of the ADT libraries from the barnyard experiment, the read configuration was 28 bp Read 1, 48 bp Read 2 and 8 bp for Index 1 and 2. We sequenced approximately 300,000 read pairs per cell for both scATAC-seq and scRNA-seq libraries and 7,000 read pairs per cell for ADT libraries in the barnyard experiment. We sequenced approximately 40,000 read pairs per cell for scATAC-seq, 35,000 read pairs per cell for scRNA-seq libraries and 5,000 read pairs per cell for both the ADT and HTO libraries in the CD4 memory T cell experiment.

Antibody oligo sequences. ADT oligos and HTO oligos from the barnyard experiment had a partial Truseq read 2 sequence followed by 12 bp UMI, 36-bp antibody-specific barcode and 25-bp poly-A tail as follows:

CAGACGTGTGCTCTTCCGATCT[12 bp UMI][36 bp Barcode] AAAAAAAAAAAAAAAAAAAAAAAAA

HTOs for the CD4 memory T cell experiment were similarly designed, except they instead had a partial Nextera read 2 sequence to allow separate amplification of TF antibody oligos from HTOs, which often stain at higher levels:

GGCTCGGAGATGTGTATAAGAGACAG[12 bp UMI][36 bp Barcode] AAAAAAAAAAAAAAAAAAAAAAAAA

Note that the hashing antibody used together with the TF antibody panel for marking the two antibody concentrations tested in CD4 memory T cells was linked to an ADT oligo with a partial Truseq read 2 sequence so that it would be amplified with the TF ADTs and could be used to normalize TF ADT counts. All antibody barcode sequences are provided in Supplementary Table 4.

Analytical methods. *ADT and HTO processing.* Raw sequencing data were converted to fastq format using bcl2fastq (Illumina). ADTs and HTOs were then assigned to individual cells and antibodies using the matcha barcode matching tool[36]. Cell barcodes were matched based on exact matches, and up to three mismatches were allowed in antibody barcode sequences. Counts for each antibody were tabulated by counting UMIs. In the barnyard experiment, cells with fewer than 200 ADT + HTO UMIs were excluded from analysis. In the CD4 T cell experiment, cells with fewer than 75 HTO UMIs or 100 ADT UMIs were excluded. All HTO count data and TF ADT count data from the barnyard experiment were normalized using a centered log ratio (CLR) transformation as previously described[2]. For the CD4 memory T cell experiment, TF ADT counts were normalized to HTO counts from the anti-NPC HTO that was added to distinguish two different concentrations of the TF antibody panel used to stain cells, since we expected that levels of the nuclear pore complex should be relatively constant across cells. We observed very similar results when normalizing to total ADT counts or just using raw ADT counts (Extended Data Fig 4a). We then multiplied by 250 (that is, roughly the median number of NPC counts per cell), added one pseudocount and $\log_2$-transformed counts. We chose the NPC normalization method because it was more robust than CLR transformation in cases where cells are primarily positive for only one antibody in the panel, as was the case for the CD4 memory T cells.

Doublet detection using HTOs. For doublet detection in the barnyard experiment, we filtered for cells with at least 400 total ADT counts, and performed CLR transformation on HTO counts only. CLR cutoffs for positive staining of each HTO was performed automatically in a similar manner to Seurat's HTODemux function[37]. Cells were *k*-means clustered based on CLR-normalized HTO counts, with *k* equal to the number of hashing oligos. This serves as a rough HTO assignment that can be used to infer background staining distributions. The cutoff value for each HTO was determined by taking the 99th percentile of a normal distribution fit to the CLR-normalized HTO counts in the $k-1$ clusters with the lowest mean value for the given HTO. This differs from Seurat's HTODemux by using a normal distribution on CLR-normalized counts rather than a negative binomial distribution on raw counts and by using the bottom $k-1$ clusters to fit the background distribution rather than the bottom 1 cluster. After computing cutoffs for each HTO, we removed cells that were not positive for exactly 1 HTO, annotating the cells positive for >1 HTO as doublets.

For doublet detection in the CD4 memory T cell experiment, we filtered for cells with at least 75 HTO counts per cell and performed CLR transformation on HTO counts only. We set CLR cutoffs for positive staining of each HTO individually based on the bimodal distribution for each HTO and only cells positive for exactly one HTO were retained. Since we also incorporated two hashing oligos in the TF staining step to distinguish between two antibody concentrations used, we also annotated doublets using these HTOs and removed them from analysis.

Barnyard experiment species analysis. Raw sequencing data were converted to fastq format, and aligned to a chimeric hg38 and mm10 reference genome using cellranger-ARC v.1.0.1 from 10X Genomics. First, we filtered droplets for high-quality cells based on >7,500 RNA UMIs, >10,000 unique ATAC fragments and TSS enrichment >10. TSS enrichment was calculated using the combined set of mouse + human TSS coordinates and the default parameters of ArchR's TSS enrichment. Next, we annotated species based on the fraction of reads aligning to either the mouse genome or the human genome. For ATAC-seq reads, this cutoff was manually set to >95% of reads aligning to a single species. For RNA-seq reads we observed greater cross-cell read contamination, particularly from mouse transcripts that had high abundance in noncell droplets. As a result, we set a cutoff of >70% reads aligning to the human genome, or >95% reads aligning to the mouse genome. For our main doublet analysis, we considered cells to be mouse-human doublets if they did not pass the species cutoff for both their ATAC-seq and RNA-seq reads.

Inferred doublet rates were calculated by dividing the observed doublet rate by the fraction of cell pairings expected to be between mouse and human cells ($\text{inferred doublet rate} = \frac{\text{mix}}{\text{mouse}+\text{human}+\text{mix}} \frac{(\text{mouse}+\text{human})^2}{2\times\text{mouse}\times\text{human}}$). For perfectly even mixtures of mouse and human cells, the inferred doublet rate will be twice the observed doublet rate, and deviations from even mixtures will increase the inferred doublet rate relative to the observed doublet rate.

Comparison of NEAT-seq with bulk ATAC-seq and RNA-seq data. For comparison with bulk data, ATAC-seq or RNA-seq reads from all K562 cells in the NEAT-seq mixing experiment were combined to calculate bulk metrics, then log transformed before calculating Pearson correlation. For ATAC-seq, insertions per peak were calculated for K562 NEAT-seq data and for bulk K562 ATAC-seq alignments (Encyclopedia of DNA Elements (ENCODE) accession ENCFF512VEZ) using a peak set derived from the bulk K562 data (ENCODE accession ENCFF558BLC). Peaks with no reads in either the bulk or NEAT-seq data were filtered, and Pearson correlation of $\log_{10}(1 + \text{insertion count})$ was calculated for bulk relative to NEAT-seq. For RNA-seq, transcripts per million (TPM) reads for each gene were calculated from all K562 cells in the NEAT-seq mixing experiment and then compared to FPKM (fragments per kilobase of exon per million mapped fragments) from ENCODE K562 RNA-seq data (ENCODE accession ENCFF501IXI). GM12878 10X Multiome data were processed with cellranger-ARC v.1.0.1 from 10X Genomics, and all filtered cells from the cellranger outputs were combined to calculate TPM. Bulk RNA-seq data in GM12878 cells from ENCODE were used for comparison (ENCODE accession ENCFF387YXX). To compare pseudobulk RNA to bulk RNA, genes were filtered to only those detected by both assays and the Pearson correlation of $\log_{10}(1 + \text{TPM})$ (single cell) with $\log_{10}(1 + \text{FPKM})$ (bulk) was calculated for both K562 and GM12878 data.

scATAC-seq analysis. Raw sequencing data were converted to fastq format and aligned to the hg38 reference genome using cellranger-ARC v.1.0.1 from 10X Genomics. Cellranger output summaries are provided in Supplementary Table 1. Fragment files were then loaded into ArchR (v.1.0.2) using the createArrowFiles function. Cells with a TSS enrichment <10 or fewer than 1,000 unique fragments per cell were removed from analysis along with HTO-annotated doublets. Remaining cells were projected onto a reference dataset of hematopoietic cells[38], using a liftover of the published hg19 peak coordinates to hg38 and the published local science instrument (LSI) loadings for each peak. Cell type annotations were transferred as the most common cell type from the ten nearest neighbors, and contaminating CD8 memory T cells were removed from further analysis. We next computed an iterative LSI dimensionality reduction using the addIterativeLSI function with the default tile matrix (insertion counts in 500 bp bins across the genome) and four iterations. Clustering was then performed using the addClusters function and a UMAP was generated using addUMAP, both with default parameters.

To call peaks, we first generated insertion coverage files from pseudobulk replicates grouped by cluster using addGroupCoverages and then called peaks with macs2 using addReproduciblePeakSet with default parameters. We then generated a matrix of insertion counts for each peak across all cells using addPeakMatrix. To aid in cluster identification, we identified marker peaks unique to each cluster and identified TF motifs enriched in these peaks using getMarkerFeatures (useMatrix = "PeakMatrix") and peakAnnoEnrichment. Results were plotted using plotEnrichHeatmap(enrichMotifs, $n = 5$, transpose = TRUE, cutOff = 5). We can also predict TF activity by measuring differences in TF motif accessibility across cells using chromVAR[39]. We first determined which peaks contain a motif of interest for motifs in the CISBP database[40] using addMotifAnnotations with the option motifSet = "cisbp". We then added a background peak set with similar GC content and number of fragments and computed motif deviations for all motifs using addBgdPeaks and addDeviationsMatrix, respectively.

To further help with cluster identification using ATAC-seq data, we can predict gene expression or epigenetic priming of a locus by calculating gene activity scores for each gene based on accessibility in the region surrounding the gene locus. These scores were calculated in ArchR during Arrow file creation with the option addGeneScoreMat = TRUE.

scRNA-seq analysis. Raw sequencing data were converted to fastq format and aligned to the reference genome using cellranger-ARC v.1.0.1 from 10X Genomics. For each lane, the gene expression matrix from the filtered_feature_bc_matrix was used to create a Seurat object using Seurat v.3.2.1. The two lanes of CD4 memory T cell data were then merged into one Seurat object and filtered for cells used in the scATAC-seq analysis. Data were normalized with NormalizeData (normalization.method = "LogNormalize" and scale.factor = 10,000). For principal component analysis, we identified the top 2,000 variable genes using FindVariableFeatures (selection.method = "vst") and RunPCA was performed on scaled data using these variable features. We then clustered cells using FindNeighbors with dimensions 1:15 and FindClusters with resolution 0.6. The RNA UMAP was generated with RunUMAP using dimensions 1:15. FindAllMarkers was used to identify marker genes enriched in each cluster.

To identify candidate regulators of GATA3 translation, we added ADT data to our Seurat object using CreateAssayObject. We first filtered for cells expressing high GATA3 RNA (natural log-normalized counts >2.25) and then identified cells expressing high GATA3 ADT ($\log_2$ NPC-normalized counts >6.12) or low GATA3 ADT ($\log_2$ NPC-normalized counts <4.9116 to match number of cells in

high GATA3 ADT subset). To identify differentially expressed genes between these two subsets, we ran FindMarkers. We converted the natural log-based fold change values output from Seurat v.3 to $\log_2$ fold changes and calculated adjusted *P* values using the Benjamini–Hochberg correction.

Data visualization. Unless otherwise indicated in the text, visualization of TF motif deviation *Z*-scores, gene activity scores, RNA and ADTs on the ATAC UMAP embedding was done by plotting imputed values using ArchR's plotEmbedding function. Ridge plots of normalized ADT counts and scatterplots with marginal histograms of normalized ADT versus RNA counts were generated using ArchR's plotGroups (plotAs = "ridges") and ggpubr's ggscatterhist, respectively. Normalized ADT counts were calculated as $\log_2(250 \times (\text{TF ADT counts/NPC HTO counts}) + 1)$. Normalized RNA counts were calculated as $\log_2(10{,}000 \times (\text{TF RNA counts/total UMI counts}) + 1)$.

Identifying peaks and genes correlated with TF abundance. To identify peaks and genes with changes that correlate with TF ADT levels, Spearman correlation values were calculated between normalized ADT counts for each TF and either normalized Tn5 insertion counts or normalized RNA counts for all peaks and genes with >10 observed reads across single cells. Raw *P* values for correlations were calculated in the same manner as R cor.test, namely using a two-sided *t*-test with $n-2$ degrees of freedom where $t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$ and n is the number of cells. *P* values were multiple-hypothesis corrected for each ADT using the 'BH' method of R's p.adjust, and significant correlations were defined as adjusted $P < 0.05$. TF motif enrichment in significantly correlated peaks was calculated using a hypergeometric test.

Identification of correlated peaks and genes. To identify peaks and genes where peak accessibility correlated with gene expression, we formed 500 aggregates of 100 cells each using the 99 nearest neighbors of randomly selected cells in LSI coordinates. These aggregates were constrained to have a maximum pairwise overlap of 80% of cells. Gene expression and peak accessibility for each aggregate was calculated by averaging the normalized accessibility or expression values across all cells in the aggregate. For all peak–gene pairs within 100 kb of each other, we calculated Spearman correlation and significance using a two-sided *t*-test as for our peak–TF and gene–TF correlations.

Identifying TF-peak–gene linkages. To identify candidate direct target genes of a TF, we identified TF ADT-correlated genes that had a TF ADT-correlated peak nearby containing the TF sequence motif. Specifically, we overlapped the top 20% of ADT-correlated genes with the top 20% of ADT-correlated peaks containing the corresponding TF motif, sorted by Spearman correlation calculated across single cells. For the overlap, we required that the peak–gene distance be less than 100 kb and that accessibility of the peak and expression of the linked gene be significantly correlated (adjusted $P < 0.05$ for Spearman correlation, as described above). To identify Gene Ontology terms enriched in these genes, we used the enrichGO function in the clusterProfiler R package (v.3.12)[41], using all genes with at least one RNA count across all cells in our dataset as the background gene list.

Analysis of fine-mapped GWAS variants. To identify candidate causal SNPs regulated by a TF and link the SNP to a putative target gene, we obtained a comprehensive list of fine-mapped GWAS SNPs (https://pics2.ucsf.edu/PICS2.html) and overlapped these with peaks from our identified GATA3 TF-peak–gene linkages. We focused on rs62088464, a SNP located within a GATA motif site and for which our donor was heterozygous for the risk allele. To determine allele-specific differences in accessibility at this SNP, we identified all reads overlapping this SNP with mapq >30 using pysam's pileup method[42]. To stratify cells by GATA3 expression, we *z*-score transformed the CLR-normalized GATA3 expression levels for each of the two antibody titration levels to ensure they were on comparable scales, then performed smoothing using the ArchR version of the MAGIC algorithm to reduce noise. Cells were divided based on their rank in the smoothed GATA3 vector. Allele-specific accessibility was determined using a one-sided binomial test, comparing the allele frequency in the top 10% of GATA3 cells using the bottom 50% as a null hypothesis. The eQTL data and analysis shown were obtained from the GTEx Portal release v.8.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed sequencing data generated in this study are available through GEO (GSE178707). Published bone marrow and peripheral blood single-cell ATAC-seq and RNA-seq data were obtained from GSE139369. The CISBP database is available at http://cisbp.ccbr.utoronto.ca/. The Transfac database is available at https://genexplain.com/transfac/. Source data are provided with this paper.

Code availability

Code used for analysis and figures are available at https://github.com/GreenleafLab/NEAT-seq_reproducibility.

References

33. Bujalowski, W. & Lohman, T. M. *Escherichia coli* single-strand binding protein forms multiple, distinct complexes with single-stranded DNA. *Biochemistry* **25**, 7799–7802 (1986).
34. Lohman, T. M. & Overman, L. B. Two binding modes in *Escherichia coli* single strand binding protein-single stranded DNA complexes. Modulation by NaCl concentration. *J. Biol. Chem.* **260**, 3594–3603 (1985).
35. Reddy, M. S., Guhan, N. & Muniyappa, K. Characterization of single-stranded DNA-binding proteins from Mycobacteria. The carboxyl-terminal of domain of SSB is essential for stable association with its cognate RecA protein. *J. Biol. Chem.* **276**, 45959–45968 (2001).
36. Parks, B. GreenleafLab, *matcha* (GitHub, 2022); https://github.com/GreenleafLab/matcha
37. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
38. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
39. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
40. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
41. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A J. Integr. Biol.* **16**, 284–287 (2012).
42. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Acknowledgements

We thank R. Baskar, S. Pierce, S. Klemm, Y. Kim and members of the Greenleaf laboratory for helpful discussions and suggestions. We also thank the Stanford Fluorescence Activated Cell Sorting facility and Stanford Functional Genomics Facility for technical support. Figure schematics were created with BioRender.com. This work was supported by funding from the National Institutes of Health (NIH) (grant nos. P50HG007735, R01HG008140, R01HG009909, UM1HG011972, U01MH116529, U54HG010426, U01DK127419, U01HG011762, U19AI057266 and UM1HG009442), the Rita Allen Foundation, the Baxter Foundation Faculty Scholar Grant and the Human Frontiers Science Program (grant no. RGY006S) to W.J.G. W.J.G. is a Chan Zuckerberg Biohub investigator (grant nos. 2017-174468 and 2018-182817). Fellowship support was provided by the Stanford School of Medicine Dean's Fellowship and NIH (grant no. F32GM135996) to A.F.C. and a training grant from the National Institute of Standards and Technology to B.P.

Author contributions

A.F.C. and W.J.G. conceived the project and designed the experiments with input from all authors. A.F.C. led method development and performed experiments with help from A.S.K. A.F.C. and B.P. performed bioinformatic analysis, visualization and interpretation. J.J.G. provided input on the CD4 memory T cell experiment and interpretation of results. A.F.C. and W.J.G. drafted the manuscript. A.F.C., B.P., B.O.-R. and W.J.G. revised and edited the manuscript with input from all authors.

Competing interests

A.F.C. and W.J.G. are listed as coinventors on a patent related to this work. 10X Genomics holds the license to other patents in which W.J.G. is listed as an inventor. W.J.G. is an equity holder of 10X genomics and a co-founder of Protillion Biosciences. W.J.G. consults for Guardant Health, Quantapore, Protillion Biosciences, and Ultima Genomics. The remaining authors declare no competing interests.

Additional information

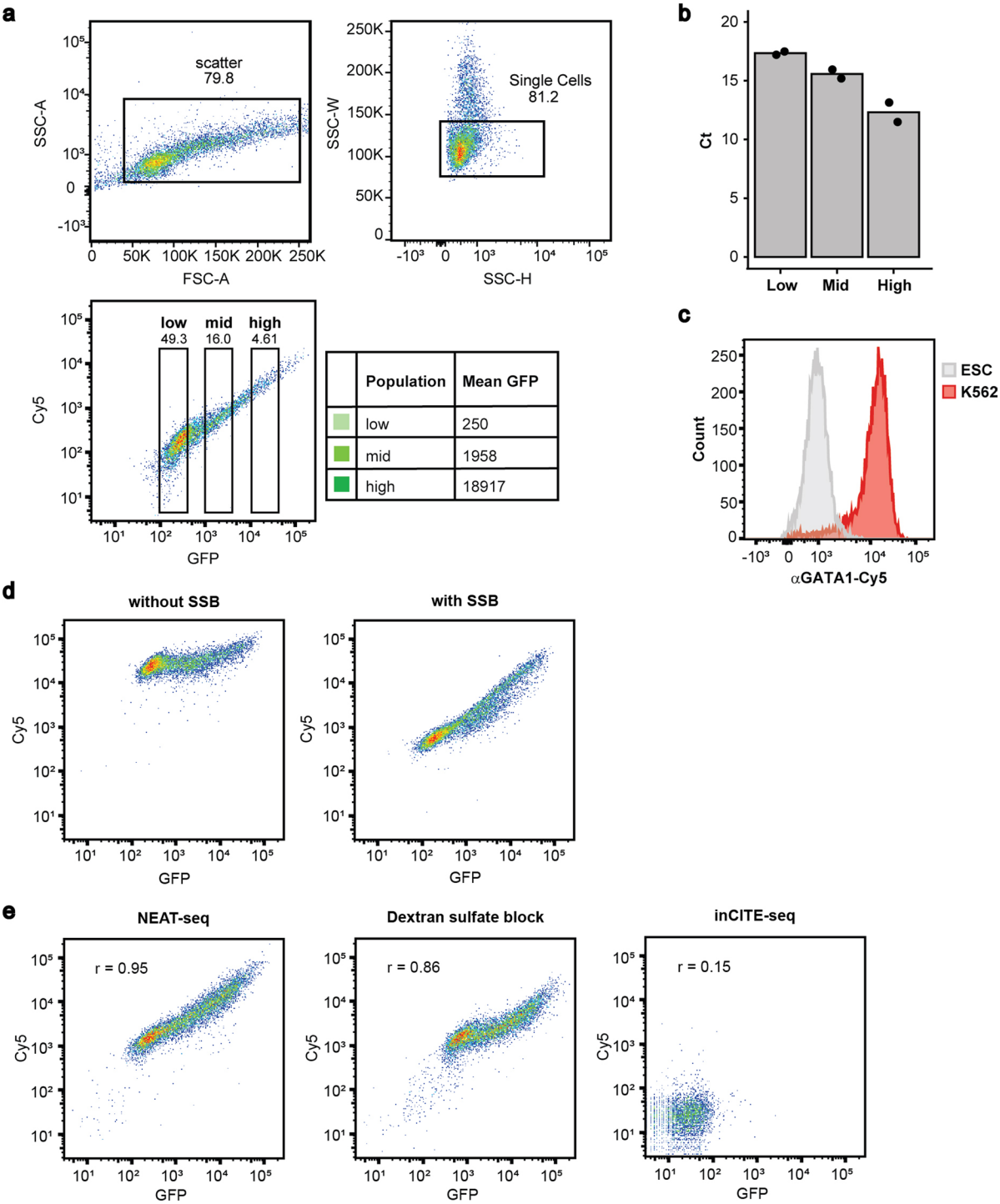
Extended data are available for this paper at https://doi.org/10.1038/s41592-022-01461-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01461-y.

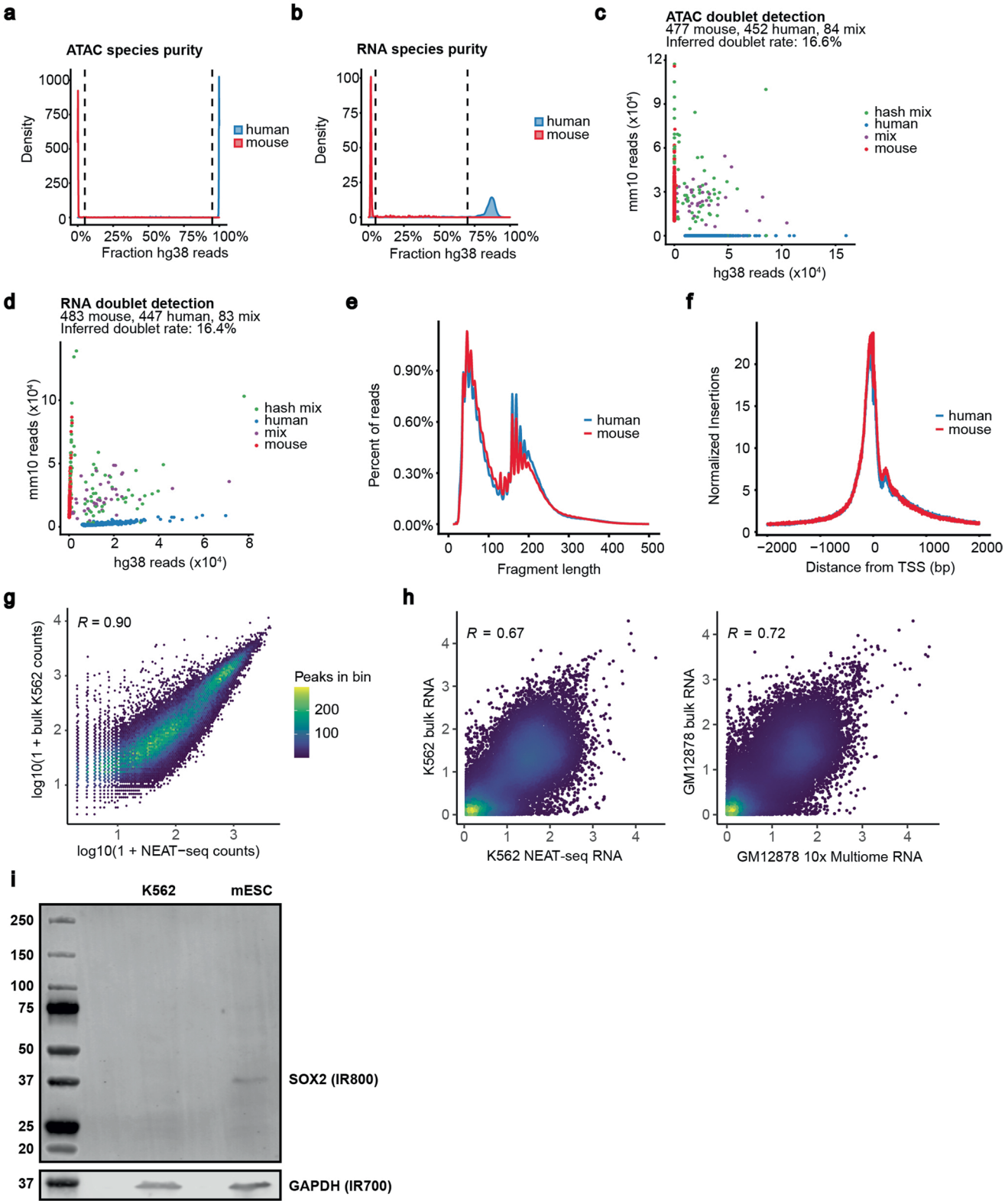
Correspondence and requests for materials should be addressed to William J. Greenleaf.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

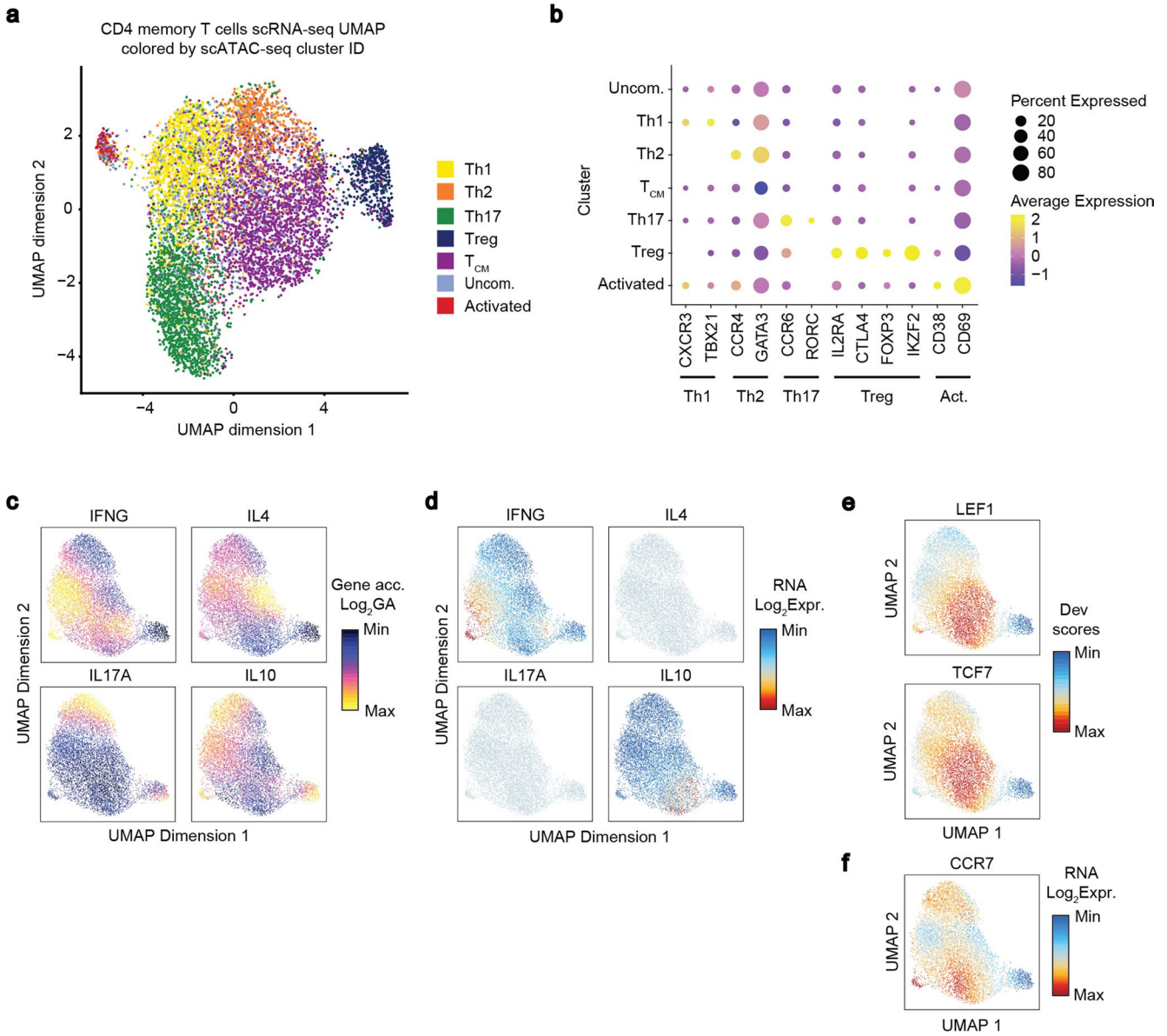


Extended Data Fig. 1 | Staining of nuclear and cytosolic proteins using oligo-antibodies blocked with EcoSSB. a) Sorting of cells expressing low, mid, or high levels of nuclear GFP that have been stained with an anti-GFP oligo-conjugated antibody. b) Quantitative PCR for the conjugated oligo from equal cell numbers of sorted populations in (a) for n = 2 technical replicates. c) Staining of K562 cells and mouse ESCs for endogenous GATA1 protein using an anti-GATA1 antibody linked to an 80 bp oligo with 3′-Cy5. d) Flow cytometry plot of HEK293T cells expressing cytosolic GFP and stained with an anti-GFP antibody linked to a 100 bp single stranded DNA oligo with 3′-Cy5 modification. e) Flow cytometry plot of nuclear GFP-expressing HEK293T cells with a GFP antibody linked to a Cy5-modified ssDNA oligo using the conditions indicated. "NEAT-seq": NEAT-seq fixation, permeabilization, and staining conditions using oligo-antibodies pre-incubated with EcoSSB. "Dextran sulfate block": NEAT-seq fixation and permeabilization conditions with inCITE-seq staining conditions (i.e with dextran sulfate blocking agent). "inCITE-seq": inCITE-seq fixation, permeabilization, and staining conditions. Spearman correlation is shown.

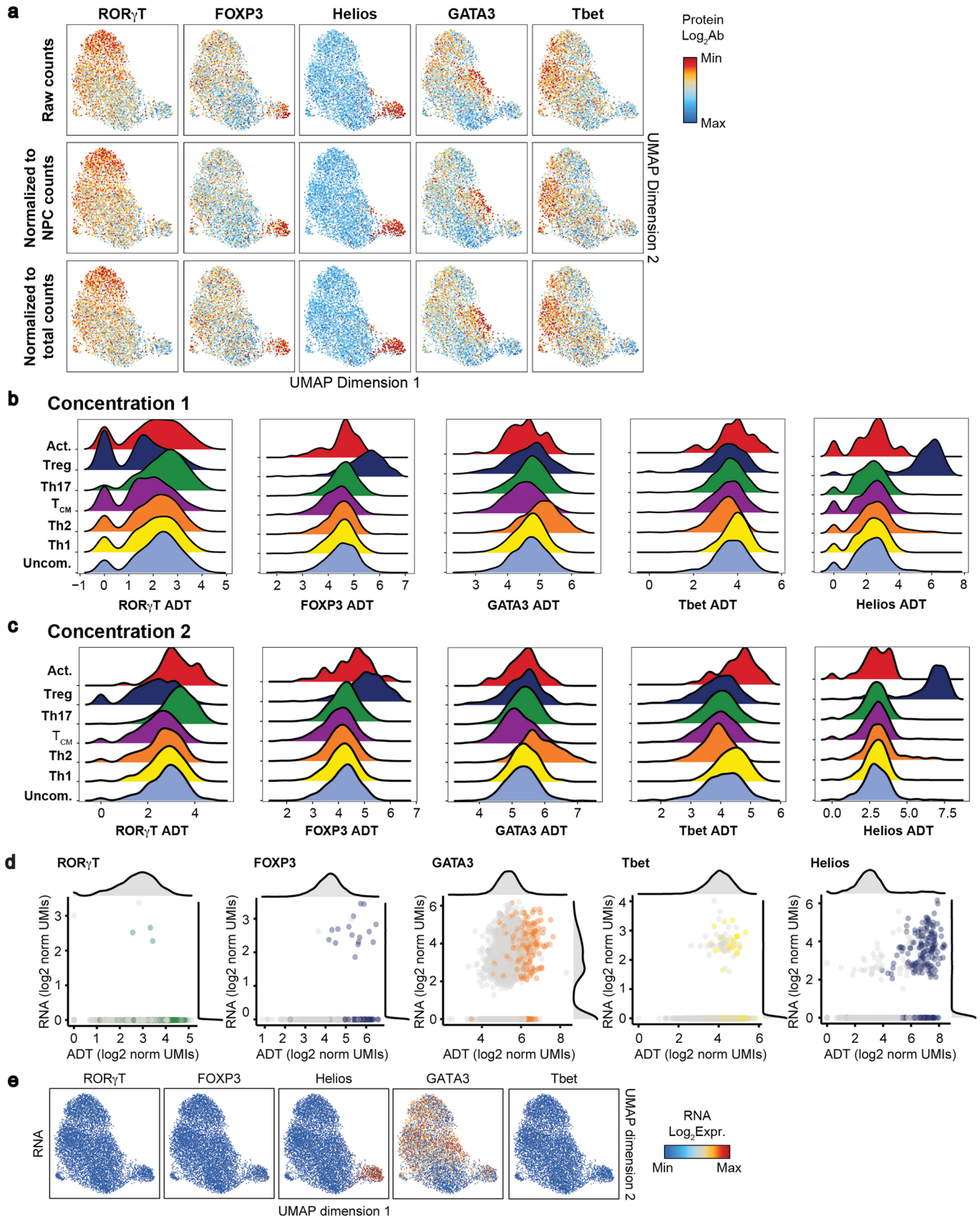


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | NEAT-seq species mixing experiment data quality. a-b) Cutoffs for annotating a cell as human, mouse, or mixed based on percentage of reads in a cell mapping to the human genome for ATAC-seq (a) and RNA-seq (b) data. c) Scatterplot of number of ATAC-seq reads mapping to the human vs mouse genome in each cell prior to removing HTO doublets, with each cell colored by its classification as a human cell, mouse cell, mixed species doublet, or an HTO doublet. d) Same as (c) but for RNA-seq reads. e) Fragment length distribution of ATAC-seq data generated using NEAT-seq. f) Average Tn5 insertions across transcriptional start sites normalized to the flanking region +/− 2 kb from the start site (i.e TSS enrichment) from scATAC-seq data generated using NEAT-seq. g) Log-transformed Tn5 insertions in ATAC-seq peaks for NEAT-seq data vs bulk ATAC-seq data in K562 cells. Pearson correlation is shown. h) A comparison of RNA-seq counts from bulk data vs NEAT-seq in K562 cells (left) or standard 10X Multiome data in GM12878 cells (right). Values are log-transformed TPM (for single cell data) or FPKM (for bulk data). Pearson correlation is shown. i) Western blot of mESC and K562 cell lysate using oligo-conjugated SOX2 antibody pre-incubated with EcoSSB and detected with IR800 secondary antibody (Licor). GAPDH was also probed as a loading control with IR700 secondary antibody. Imaging was performed on a Licor Odyssey imaging system. Images shown are representative results of 2 independent experiments.

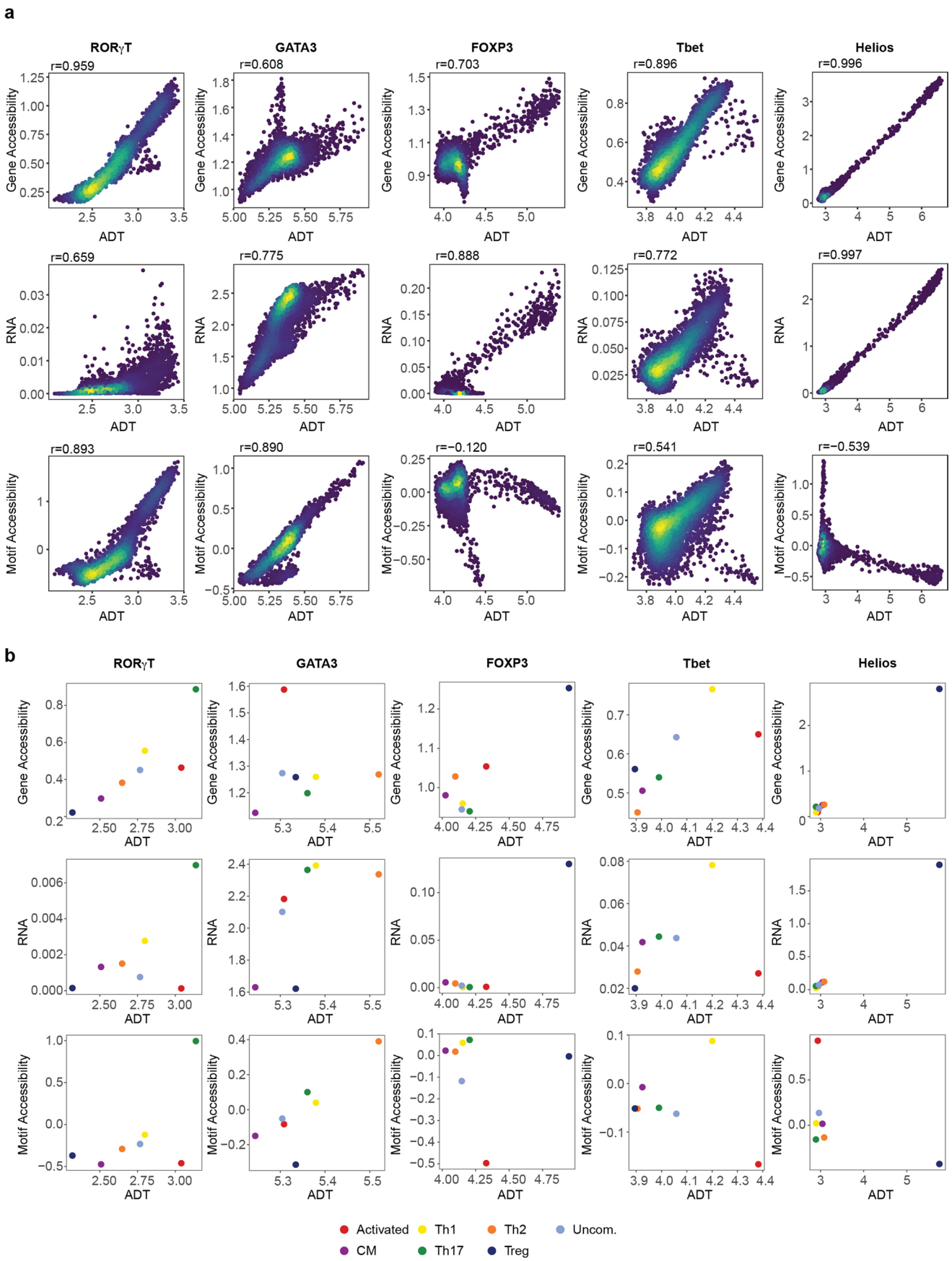


Extended Data Fig. 3 | CD4 memory T cell subset annotations. a) scRNA-seq UMAP of CD4 memory T cells colored with scATAC-seq cluster IDs. T_{CM} = central memory, Act. = recently activated cells, Uncom. = uncommitted memory cells. b) RNA expression of master TF drivers and canonical cell surface markers of CD4 memory cell subsets in each scATAC-seq cluster. TBX21 = Tbet transcript, RORC = RORγT transcript, IKZF2 = Helios transcript. c) Gene accessibility for cytokines induced in different CD4 T cell subsets overlayed on the scATAC-seq UMAP. d) RNA levels for the cytokines in (c). e) chromVAR deviation scores for the naïve and CM T cell TFs, LEF1 and TCF7, overlayed on scATAC-seq UMAP. f) RNA expression of the CM marker, CCR7, overlayed on the scATAC-seq UMAP.

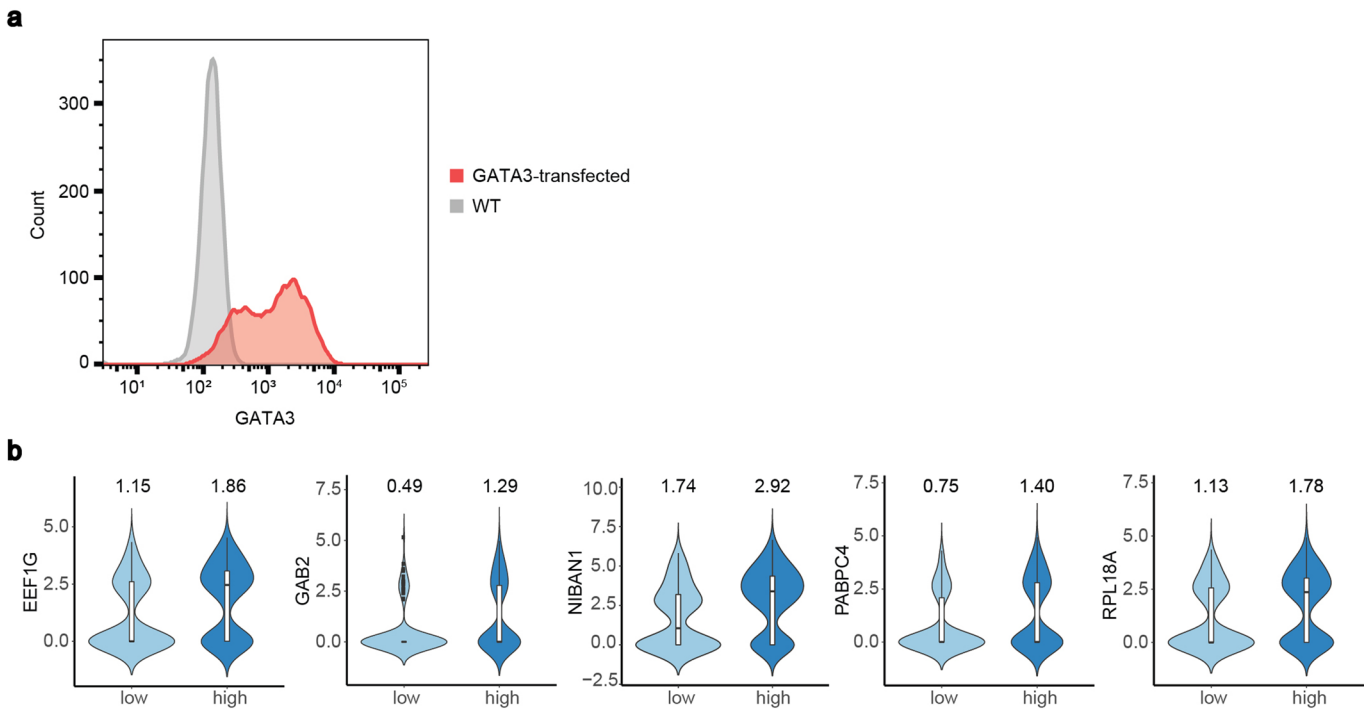


Extended Data Fig. 4 | See next page for caption.

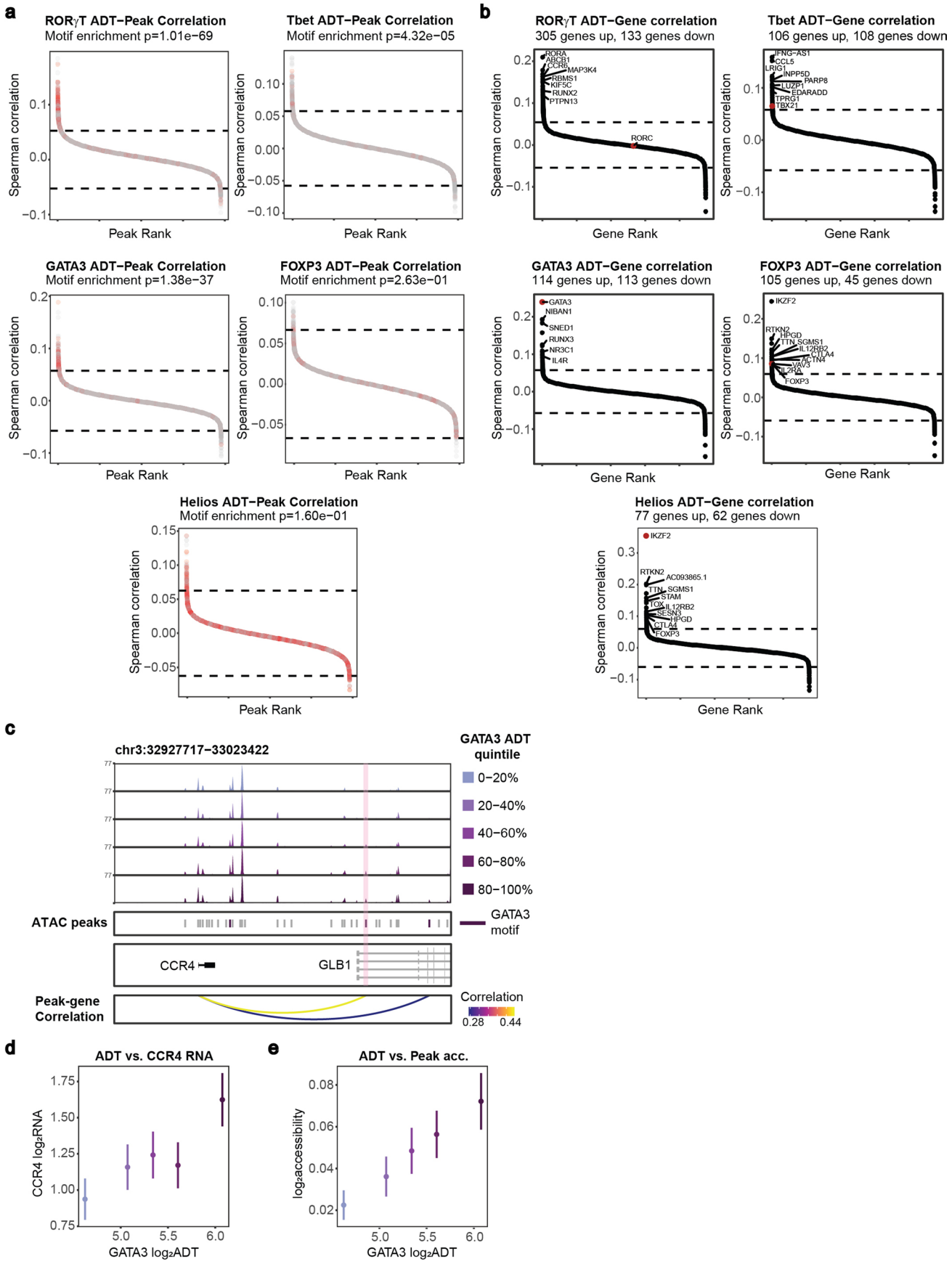
Extended Data Fig. 4 | Enrichment of TF ADTs in the T cell subtype driven by the TF. a) Unsmoothed ADT counts for each TF overlayed on the scATAC-seq UMAP after the indicated normalization method (raw counts, NPC normalization, and total ADT counts normalization). b) Log_2-transformed, NPC-normalized ADT counts for each TF separated by scATAC-seq cluster for cells stained with antibody concentration 1 (see methods). c) Same as (b) but for antibody concentration 2. d) Scatterplots with marginal histograms of log2-transformed read-normalized RNA vs log2-transformed NPC-normalized ADT counts for each TF. Colored data points represent cells belonging to the scATAC-seq cluster most enriched in expression of the indicated TF. e) Unsmoothed, normalized RNA counts of the indicated TFs overlayed on the scATAC-seq UMAP.



Extended Data Fig. 5 | Correlation of ADT levels with gene locus accessibility, RNA, and motif accessibility for each TF. a) Correlation across all cells for each measurement. Values were first smoothed across neighboring cells using MAGIC imputation to account for dropouts. Pearson correlations are shown. b) The data in (a) but averaged across cells within each scATAC-seq cluster.

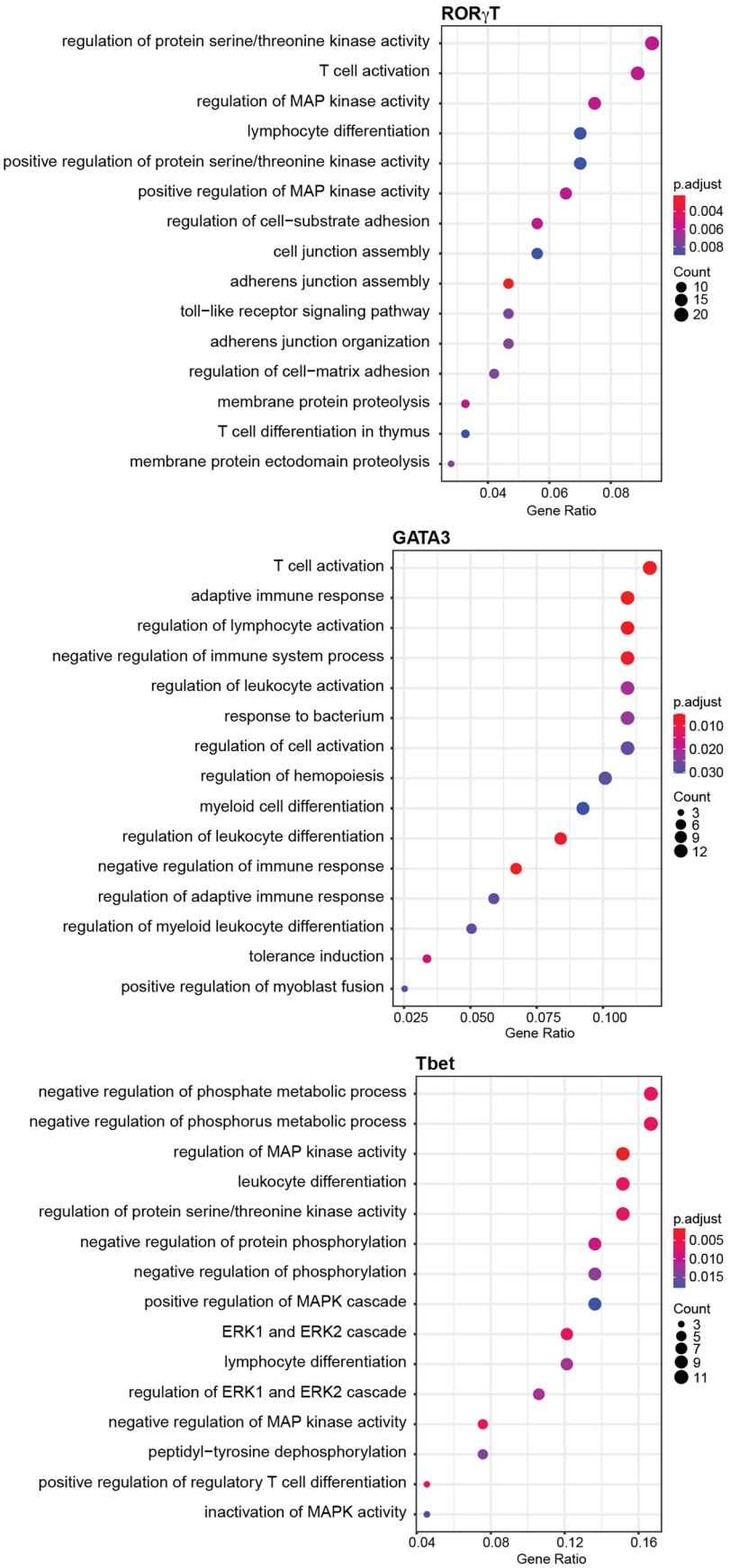


Extended Data Fig. 6 | Post-transcriptional regulation of GATA3. a) Flow cytometry plots of GATA3-transfected and WT HEK293T cells stained with an EcoSSB-bound oligo-conjugated GATA3 antibody, using a fluorescent secondary antibody for detection. b) Log2-transformed RNA levels of the indicated gene in cells expressing high RNA and low protein ("low") vs high RNA and high protein for GATA3 ("high"). The mean RNA expression for each group is shown above the violin plot. N = 140 cells examined over one independent experiment for both "high" and "low" populations. Boxplots show median with bounds of the box representing the 25th and 75th percentiles and the whiskers extending to the value closest to but not exceeding 1.5 times the interquartile range. Data extending beyond the whiskers are plotted individually as outliers.

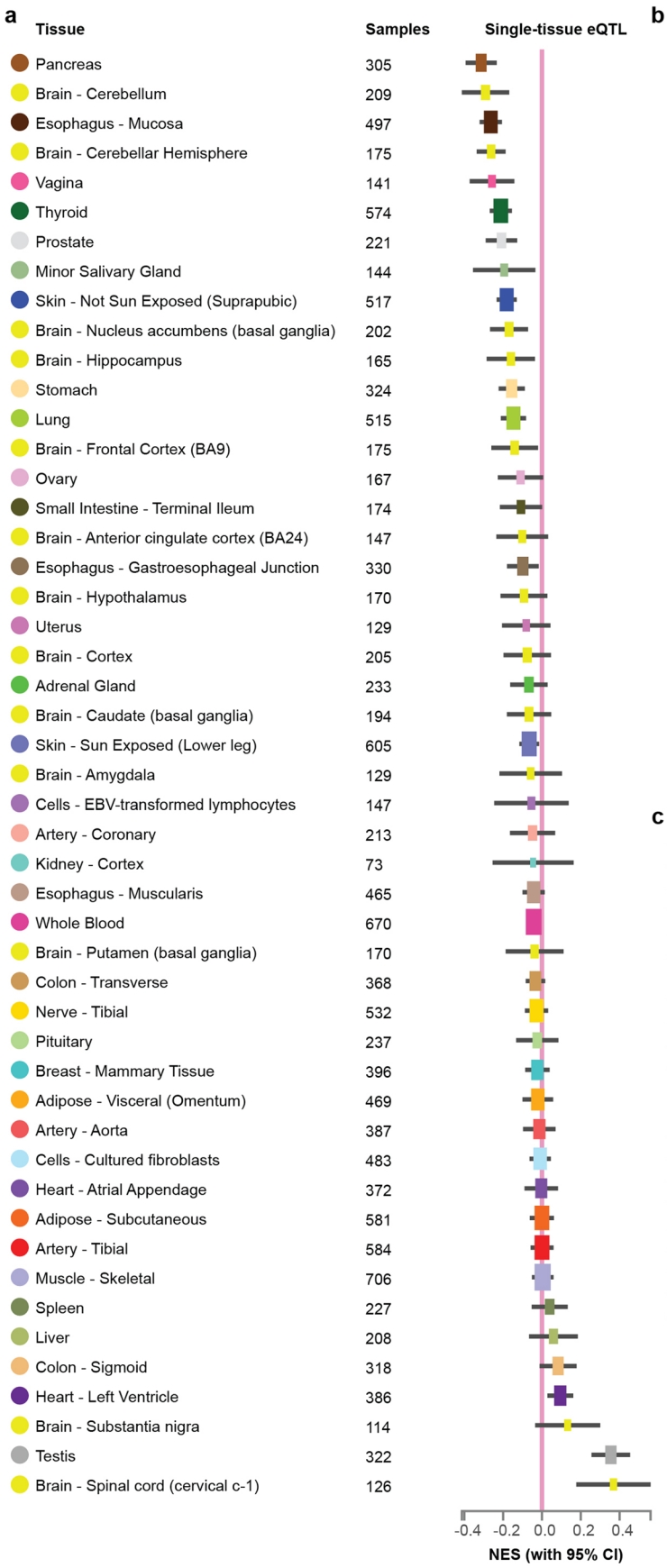


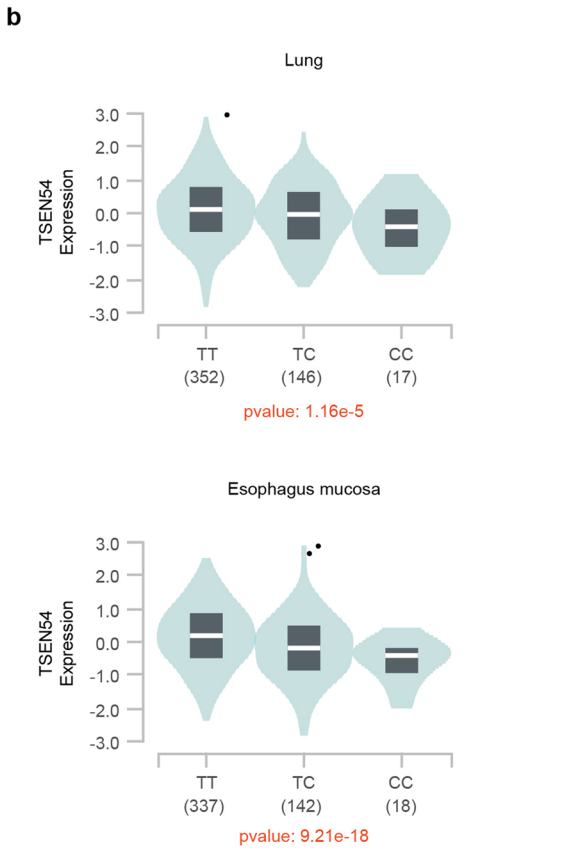
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Identification of DNA regulatory elements and genes correlated with master TF protein expression. a) Spearman correlations between ATAC-seq peak accessibility and NPC-normalized TF ADT counts across single cells. Cutoffs for significant correlations are indicated by dashed lines (see Methods). Points in red indicate peaks containing a binding motif for the TF. TF motif enrichment in significantly correlated peaks was calculated using a hypergeometric test. b) Spearman correlations between read-normalized RNA counts and NPC-normalized TF ADT counts across single cells. Cutoffs for significant correlations are indicated by dashed lines (see Methods). Significantly correlated genes known to be enriched or play a functional role in the relevant T cell subset are labeled. c) CCR4 ATAC-seq tracks in CD4 memory cells separated into quintiles by GATA3 ADT levels, along with significantly correlated peak-gene linkages (adj. $p < 0.05$). Spearman correlations are shown. Peaks containing a GATA3 motif are indicated. d-e) CCR4 RNA expression (d) and accessibility at the highlighted GATA3 motif-containing peak (e) as a function of GATA3 ADT levels. Mean is shown with standard error of the mean of $n = 768 +/- 1$ cells per group.



Extended Data Fig. 8 | GO term enrichment for candidate target genes in TF-driven peak-gene linkages. Enriched GO terms in the target gene list were identified by hypergeometric test using enrichGO in the clusterProfiler R package, using all genes with at least one RNA count in the dataset as a background gene list. Adjusted p-values were calculated using the Benjamini-Hochberg procedure.





c

Gene	P-Value	NES	Tissue
TSEN54	3.80E-13	-0.31	Pancreas
TSEN54	6.10E-06	-0.29	Brain - Cerebellum
TSEN54	9.20E-18	-0.26	Esophagus - Mucosa
TSEN54	1.70E-10	-0.26	Brain - Cerebellar Hemisphere
TSEN54	1.30E-12	-0.21	Thyroid
TSEN54	1.10E-06	-0.21	Prostate
TSEN54	1.90E-11	-0.18	Skin - Not Sun Exposed (Suprapubic)
MRPS7	2.70E-06	-0.16	Artery - Aorta
TSEN54	8.70E-06	-0.16	Stomach
TSEN54	1.20E-05	-0.15	Lung
MRPS7	2.70E-05	-0.15	Skin - Sun Exposed (Lower leg)
SLC25A19	2.90E-05	-0.15	Lung
MIF4GD	6.60E-06	-0.14	Whole Blood
MRPS7	7.60E-06	-0.12	Thyroid
MRPS7	1.80E-05	-0.12	Heart - Left Ventricle
MRPS7	3.80E-05	-0.12	Nerve - Tibial
MIF4GD	8.30E-05	-0.11	Skin - Sun Exposed (Lower leg)
MRPS7	9.90E-05	-0.11	Whole Blood
MRPS7	4.70E-05	-0.095	Adipose - Subcutaneous
MRPS7	6.60E-05	-0.085	Lung

Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Single-tissue eQTL analysis for the rs62088464-TSEN54 variant-gene pair from GTEx portal. a) Normalized effect size (NES) across various tissues for the protective (G) vs risk (A) allele on TSEN54 expression. The risk allele preserves the GATA3 motif. A negative NES value indicates a gene with expression that is associated with the risk allele. Error bars indicate 95% confidence intervals. b) Normalized TSEN54 expression grouped by rs62088464 genotype for lung and esophagus mucosa. c) Genes ordered by NES across tissues for rs62088464. P-values for (b) and (c) are outputs from the GTEx portal (release v8) and are calculated from a two-sided t-test comparing the observed NES in a tissue to a null NES of 0.

nature portfolio

Corresponding author(s): William J Greenleaf

Last updated by author(s): Feb 17, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	cellranger-ARC v.1.0.1 for collecting and processing 10x Genomics single cell data
Data analysis	ArchR v1.0.2, Seurat v3.2.1 for single cell analysis, FlowJo v10.7.1 for flow cytometry analysis, bcl2fastq v2.18, matcha 0.0.1, clusterProfiler v3.12

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw and processed sequencing data generated in this study are available through GEO (GSE178707). Published bone marrow and peripheral blood single cell ATAC-seq and RNA-seq data were obtained from GSE139369. The CISBP database is available at http://cisbp.ccbr.utoronto.ca/. The Transfac database is available at https://genexplain.com/transfac/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed. Samples were chosen mainly to validate and illustrate the capabilities of the technique and we processed a sufficient number of cells to enable identification of known subpopulations within the samples.
Data exclusions	No data were excluded from analysis
Replication	We generated two independent single cell libraries from two lanes of a 10x chip (i.e technical replicates) for the CD4 T cell experiment and no significant differences were observed between the two lanes. Staining of nuclear GFP using oligo-antibodies in the presence or absence of EcoSSB was performed at least 3 times in independent experiments and validated in all cases. Comparison of staining using NEAT-seq vs inCITE-seq methods was performed twice with similar results. Western blot validation of the conjugated SOX2 antibody was performed twice with similar results. Flow cytometry validation of conjugated GATA3 antibody was performed once.
Randomization	Not applicable since we did not have experimental or control groups in this study.
Blinding	Blinding was not relevant to this study since we did not compare experimental or control groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	GATA1 (Abcam ab241393), OCT4 (R&D AF1759), SOX2 (R&D MAB2018), nuclear pore complex (Biolegend 902901), GFP (Biolegend 338002), GAPDH (Cell Signaling Technologies 5174T), IRDye 680RD Goat anti-Rabbit IgG (Licor 925-68071), IRDye 800CW Goat anti-Mouse IgG (Licor 925-32210). The following antibody clones were custom conjugated by BD Biosciences, with catalog numbers for purified versions noted where possible: GATA3 (L50-823, #558686), Tbet (4B10, #561262), RORγT (Q21-559), FOXP3 (259D/C7, 560044), and Helios (22F6).
Validation	Validation of GATA1, OCT4, and SOX2 were performed on positive and negative control cell lines using flow cytometry. Specifically, GATA1 staining was enriched in the K562 cell line relative to mESCs, while SOX2 and OCT4 staining were enriched in mESCs relative to K562 cells. These antibodies were also validated by Western blot in these cell types by the manufacturing companies. The GAPDH antibody was also validated for Western blot usage by the manufacturing company. The GFP antibody from Biolegend and all BD Biosciences antibodies were validated for flow cytometry usage by the manufacturing company. All antibodies were validated by the seller in the relevant species for our application.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	K562 and HEK293T cells were obtained from ATCC. V6.5 ESCs were obtained from Novus Biologicals

Authentication	Cell lines were not authenticated.
Mycoplasma contamination	Cell lines were not tested for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	On day 0, HEK 293T cells were seeded at 4 million cells per 10cm plate. On day 1, cells were transfected with 6ug nuclear-localized GFP construct (Addgene #67652) using Fugene HD transfection reagent (Promega). Cells were harvested and stained using anti-GFP antibody (Biolegend 338002) linked to an 80 bp ssDNA oligo with 3' Cy5 fluorophore as described in the oligo-antibody staining methods section, except without RNase inhibitor or DTT. A control stain was performed with the oligo-antibody in the absence of SSB. Stained cells were resuspended in PBS for flow cytometry
Instrument	Cells were analyzed on an LSRII flow cytometer or sorted on a BD FACS Aria II
Software	FlowJo v.10.7.1 was used for analysis
Cell population abundance	7000 cells were sorted from low, mid, and high GFP populations for qPCR of conjugated antibody oligo.
Gating strategy	We gated cells as shown in Fig S1a, aiming to obtain populations with roughly 10-fold differences in GFP expression.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.