SCIENTIFIC REPORTS

OPEN

Global dynamics of stage-specific transcription factor binding during thymocyte development

Received: 13 December 2017
Accepted: 20 March 2018
Published online: 04 April 2018

Tomonori Hosoya[1], Ricardo D'Oliveira Albanus[2], John Hensley[2], Greggory Myers[1], Yasuhiro Kyono[2,3], Jacob Kitzman[2,3], Stephen C. J. Parker[2,3] & James Douglas Engel[1]

In vertebrates, multiple transcription factors (TFs) bind to gene regulatory elements (promoters, enhancers, and silencers) to execute developmental expression changes. ChIP experiments are often used to identify where TFs bind to regulatory elements in the genome, but the requirement of TF-specific antibodies hampers analyses of tens of TFs at multiple loci. Here we tested whether TF binding predictions using ATAC-seq can be used to infer the identity of TFs that bind to functionally validated enhancers of the *Cd4*, *Cd8*, and *Gata3* genes in thymocytes. We performed ATAC-seq at four distinct stages of development in mouse thymus, probing the chromatin accessibility landscape in double negative (DN), double positive (DP), CD4 single positive (SP4) and CD8 SP (SP8) thymocytes. Integration of chromatin accessibility with TF motifs genome-wide allowed us to infer stage-specific occupied TF binding sites within known and potentially novel regulatory elements. Our results provide genome-wide stage-specific T cell open chromatin profiles, and allow the identification of candidate TFs that drive thymocyte differentiation at each developmental stage.

T cells develop in the thymus, where biologically distinct events driven by the interplay of multiple transcription factors (TFs) acting in coordination take place at each thymocyte stage. After migration of thymic seeding progenitors from the bone marrow and their occupation of supportive niches in the thymic medulla, early thymic progenitors (ETP) develop through immature double negative (DN; $CD4^{-}CD8^{-}$) cells to the double positive (DP; $CD4^{+}CD8^{+}$) stage, and then mature into either CD4 single-positive (SP4) helper or CD8 SP (SP8) killer T cells. While ETP retain multi-lineage differentiation capacity, they gradually lose the potential to become non-T lineage cells and become increasingly restricted to a T lineage fate[1–4]. During the DN stages, committed, developing T cells undergo immune system-specific DNA recombination, and must successfully recombine a *Trb* gene allele (encoding the T cell β receptor, TCRβ) to then pass the β selection checkpoint (when the formation of a pre-TCR complex is assessed)[5,6] in order to survive. At the next DP stage (where both CD4 and CD8 are expressed on the cell surface), the TCRα receptor rearranges, and only cells expressing a functional cell surface TCR complex (TCRα plus TCRβ) that is able to bind with appropriate affinity to the major histocompatibility complex (MHC) survive positive selection[7]. DP cells recognizing MHC class I can then mature into SP8 T cells, while DP cells recognizing MHC class II mature into SP4 T cells. Finally, negative selection eliminates by apoptosis cells that bind to self-peptides presented by the MHC, and only cells that do not exhibit high affinity to self-peptides survive[7].

Although T cell developmental stage-specific gene expression profiling has been previously described[8,9], the mechanisms that regulate those spatial and temporal expression patterns are far less well understood for all but a handful of genes. DNA-binding TFs play a central role governing gene expression in each cell, often eliciting transcriptional responses through specialized regulatory elements, including promoters, enhancers, and silencers. A widely accepted model for gene expression is that multiple transcription factors bind to an enhancer, assemble an enhanceosome, and then recruit co-activators and chromatin-remodeling proteins to the promoter[10,11]. Given the limitations of ChIP-seq to detect a single TF per assay, an alternative approach for detecting TF binding is using open chromatin assays, such ATAC-seq[12,13]. The genome is highly compact except within transcribed genes and regulatory elements, where chromatin is open and sensitive to cleavage by DNaseI[14–16] or transposition by Tn5 transposase[17]. The binding of TFs to DNA affects DNase/transposase cleavage in the vicinity of the bound

[1]Department of Cell and Developmental Biology, Ann Arbor, USA. [2]Department of Computational Medicine and Bioinformatics, Ann Arbor, USA. [3]Department of Human Genetics, University of Michigan, 3035 BSRB, 109 Zina Pitcher Place, Ann Arbor, Michigan, 48109-2200, USA. Tomonori Hosoya and Ricardo D'Oliveira Albanus contributed equally to this work. Correspondence and requests for materials should be addressed to J.D.E. (email: engel@umich.edu)

site, allowing for TF occupancy to be predicted from the chromatin accessibility data[12,13,18]. Thus DNase/ATAC footprinting can be used to identify TF binding motif sequences within regulatory elements.

To generate genome-wide profiles of stage-specific chromatin accessibility and TF binding during thymocyte development, we performed ATAC-seq at four different stages of adult thymocyte development: DN, DP, SP4 and SP8 stages. The open chromatin regions identified by ATAC-seq highlighted both known, biologically validated regulatory elements, as well as many novel potential regulatory elements. Furthermore, footprinting analysis[12,13] of those open chromatin regions revealed the high-resolution landscape of predicted TF-bound motifs within those sequences. Our ATAC-seq data enabled the discovery of both stage-independent and stage-specific domains of open chromatin, and the TF footprinting data revealed 10–20 novel protein bound sequences within the previously validated enhancers of the *Cd4*, *Cd8, Trb* and *Gata3* genes. Furthermore, enrichment analyses of TF binding in stage-specific open chromatin allowed the identification of TF motifs potentially driving each stage of thymocyte development. These data demonstrate that stage-specific changes in open chromatin are highly dynamic as thymocytes develop and provide deep insight into how the stage-specific binding of multiple TFs orchestrate transcriptional regulatory networks.

Results

T cell developmental stage-specific genome-wide mapping of accessible chromatin. To gain insight into developing T cell stage-specific chromatin opening, DN, DP, SP4 and SP8 cells were isolated from adult thymi by flow cytometry (Supplementary Fig. S1). 50,000–100,000 cells were processed for ATAC-library preparation as described[17]. The ATAC-seq reads (Supplementary Table S1) were then mapped to mouse reference genome mm10 using BWA[19] and peaks were called using MACS2[20]. ATAC-seq signals depicted in the IGB browser[21] were reproducible in thymocytes recovered from 4 individual animals (Supplementary Fig. S2), and all peaks were highly correlated across biological and technical replicates (median Spearman correlations: DN = 0.89, DP = 0.87, SP4 = 0.88, SP8 = 0.90; Supplementary Fig. S3). ATAC-seq signals at the DP stage (which comprises approximately 85% of total thymocytes, Supplementary Fig. S1) reflected profiles that were similar to DNase-seq peaks of total adult thymocytes[22] (Supplementary Fig. S2), as anticipated. On a global scale, DP ATAC-seq peak signals were highly correlated with DNase-seq peak signals of total thymocytes (median Spearman correlation = 0.70 to 0.79 Supplementary Fig. S3). Based on these results, we concluded that ATAC-seq provides a biologically reliable strategy to attain deeper insights into T cell stage-specific chromatin accessibility and transcription factor binding.

We identified 150,139 (DN), 107,110 (DP), 115,074 (SP4) and 104,411 (SP8) genomic open chromatin peaks at 5% FDR (Supplementary Fig. S4). These open chromatin domains correspond to 1.63% (DN), 1.22% (DP), 1.32% (DP) and 1.26% (SP4) of the mouse genome. 73,177 peaks were present at all four stages of thymocyte development, while the others were stage-specific. 20% (DN), 27% (DP), 24% (SP4) and 26% (SP8) of the ATAC peaks overlapped with promoter regions (defined as 200 bp upstream of a gene transcriptional start site). 10% (DN), 9% (DP), 8% (SP4) and 9% (SP8) of the ATAC peaks overlapped with an exon, but not with a promoter. 73% (DN), 63% (DP), 68% (SP4) and 65% (SP8) of the ATAC peaks overlapped with neither an exon nor a promoter (Supplementary Fig. S4).

We next sought to quantify the full spectrum (from specific to ubiquitous) of patterns of chromatin accessibility across the analyzed thymocyte developmental stages in an unbiased manner. We performed *k*-means clustering using the ATAC-seq signal. This analysis yielded 6 clusters of accessible regions: four that were specific for each stage (DN, DP, SP4, SP8), one that was ubiquitous, and one that was a combination of DN and ubiquitous (Fig. 1a, Supplementary Fig. S5). The ubiquitous cluster covered more genomic territory than any of the stage-specific clusters, while the DN-specific cluster covered more territory than the other stage-specific clusters (Fig. 1b), which is consistent with the previous conclusion that in general differentiated cells maintain a more compact chromatin architecture than their immature counterparts[23].

We next measured the distance of each peak in the four clusters to the nearest TSS and found that the ubiquitous cluster was significantly closer to TSS than the other clusters ($p < 10^{-3}$ [24], pairwise Kolmogorov-Smirnov tests with Bonferroni correction), suggestive of it being more associated with promoters and housekeeping genes than cell-identity features (Fig. 1c). Supporting this hypothesis, we found that SP4- and SP8-specific clusters were the most enriched for T cell related GO terms using ChIP-Enrich[24] (Supplementary Fig. S6). The DP-specific cluster also had high enrichment for terms related to T cell differentiation, but to a lesser extent. Conversely, the DN-specific and ubiquitous clusters were strongly enriched for non-specific developmental terms, suggesting that these might regulate more general functions. These results form a comprehensive map of developmental dynamics in the open chromatin landscape across thymocyte maturation.

TF binding identification by ATAC-seq footprinting. In order to achieve greater insights into genomic DNA sequences that are bound by TFs, we performed TF footprinting predictions using CENTIPEDE[25]. To validate the performance of CENTIPEDE footprint calls in our thymocyte data, we first compared our results with GATA3 ChIP-seq data in DN and DP thymocytes (GSE20898)[26] and CTCF ChIP-seq in total thymocytes (ENCODE, ENCSR000CDZ)[22]. We used the Genomic Annotation Tester (GAT) tool[27] to statistically evaluate the overlap between footprint calls and ChIP-seq bound motifs, while controlling for genome and feature sizes, as well as mapability issues (see Methods for details). Tests on both datasets showed significant overlap between ChIP-seq and footprint data ($p < 10^{-3}$), indicating that the footprint predictions recapitulate actual protein binding events detected by ChIP-seq for the corresponding TF (Supplementary Table S2). These data demonstrate the effectiveness of the footprint calls from these deeply sequenced ATAC-seq data in order to generate a high confidence catalogue of putative TF-bound sequences in a thymocyte stage-specific manner.

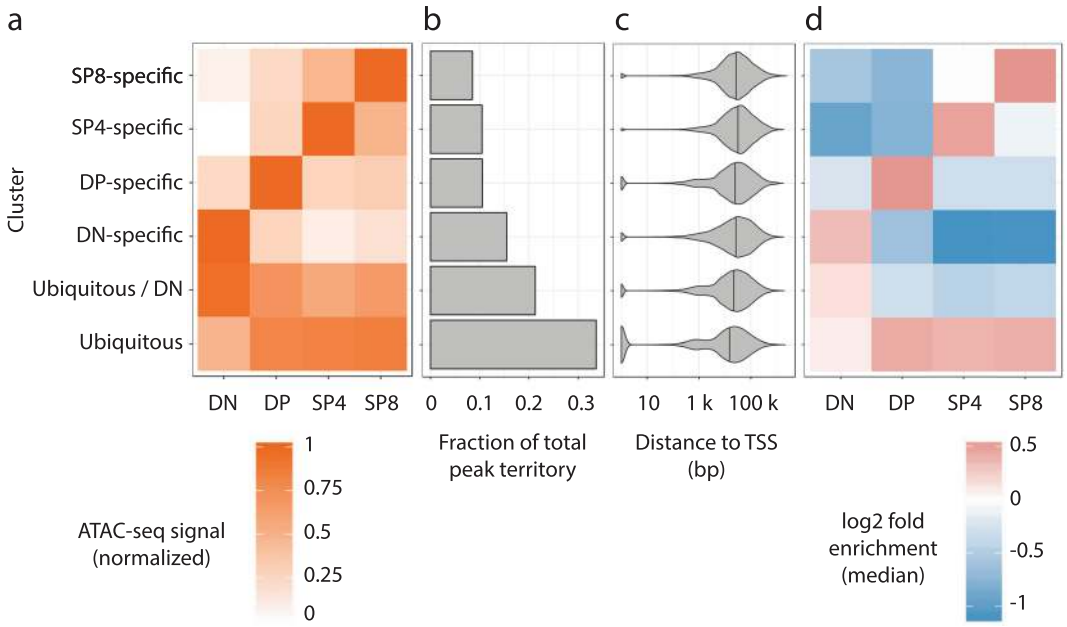

Figure 1. Analysis of stage-specific and ubiquitous ATAC-seq clusters. (**a**) *k*-means clustering results for the ATAC-seq signal in the four samples. Colors indicate the mean ATAC-seq signal for each sample in the respective cluster (*i.e.* the cluster centers). (**b**) Fraction of total peaks territory covered by each of the clusters. (**c**) TSS distance distribution (in $\log_{10}$ scale) for each of the clusters. Vertical bars in the violin plots correspond to the median of the dataset. (**d**) Median GAT footprint enrichment of all motifs for each dataset in the *k*-means clusters. GAT footprint enrichment heatmaps for each motif are shown in Supplementary Fig. S7.

We next focused on TF binding motifs for T cell activators and repressors that were predicted by Jojic *et al.*[28] from stage-specific gene expression profiling. We predicted the binding for the Jojic factors and their families. As expected, we found that the footprints called in each sample were enriched both in its own specific cluster and the ubiquitous cluster, but depleted in the other sample-specific clusters using GAT, indicating that CENTIPEDE did not detect bound TF binding sites in regions that were not active in the sample being analyzed (Fig. 1d and Supplementary Fig. S7).

We next focused on footprint calls within functionally validated enhancers. The classic definition of enhancer requires that it must be functionally validated by tests for both sufficiency and necessity in regulating its specific target gene expression, but to date only few T cell enhancers have been tested for both *in vivo*. The ATAC-seq data identified open chromatin regions within functionally validated regulatory elements for the *Cd4* (Fig. 2 and Supplementary Fig. S8), *Cd8* (Fig. 3 and Supplementary Fig. S9), *Trb* (Supplementary Fig. S10) and *Gata3* (Fig. 4) genes. The fact that the ATAC footprints recapitulated TF binding to the previously characterized motifs within these regulatory elements underscore the robust nature of the footprint approach employed in this study. Furthermore, our footprint data unveiled 8–20 novel sequences that were predicted to be bound within each of these regulatory elements (Figs 2–4 and Supplementary Figs S8–10). Based on these data, we propose that TFs bind to these sequences to assemble an active structural element that initiates and/or maintains the activity of each of these regulatory modules.

Changes in global TF binding during thymocyte development. We next sought to identify higher-resolution differences in predicted TF binding across samples by measuring the pattern of chromatin accessibility anchored on footprint motifs. We found striking differences for the footprint motifs across the samples and clusters in which they were active (Fig. 5 and Supplementary Fig. S11). CTCF had strong detectable binding patterns only in the ubiquitous cluster, and a similar pattern was observed for EGR3. TCF7 (aka TCF-1) had significant binding in all clusters. TCF4, on the other hand, was detected more strongly in the DP and DN clusters, was mostly absent in the common clusters, and almost undetectable in SP4 and SP8 thymocytes, even though it was one of the most significantly enriched motifs in these two stages ($p = 0.001$). RUNX patterns were visible in the common and DN clusters, but not in the more differentiated stages. Although GATA footprints were enriched in all clusters, we could not detect strong binding patterns, which is suggestive that it may have weaker interactions with DNA[29]. Interestingly, we did not find any SP4- or SP8-specific occupancy patterns, even though some motifs, such as TCF3 and ID4, had higher enrichment values in the SP4- and SP8-specific clusters than in the ubiquitous cluster (see next and Supplementary Table. S3 for the enrichment values). These different patterns between stages for TCF3 and ID4 suggest that the availability (expression or protein levels) of these TFs changes or that different TFs recognize these motifs at each stage.

We finally asked which footprints were enriched in each of the stage-specific open chromatin clusters defined in Fig. 1a. Each cluster showed enrichment of different TF motif footprints (Fig. 6 and Supplementary Table S3). Of note, we independently performed motif enrichment in the ATAC-seq clusters using HOMER (Supplementary

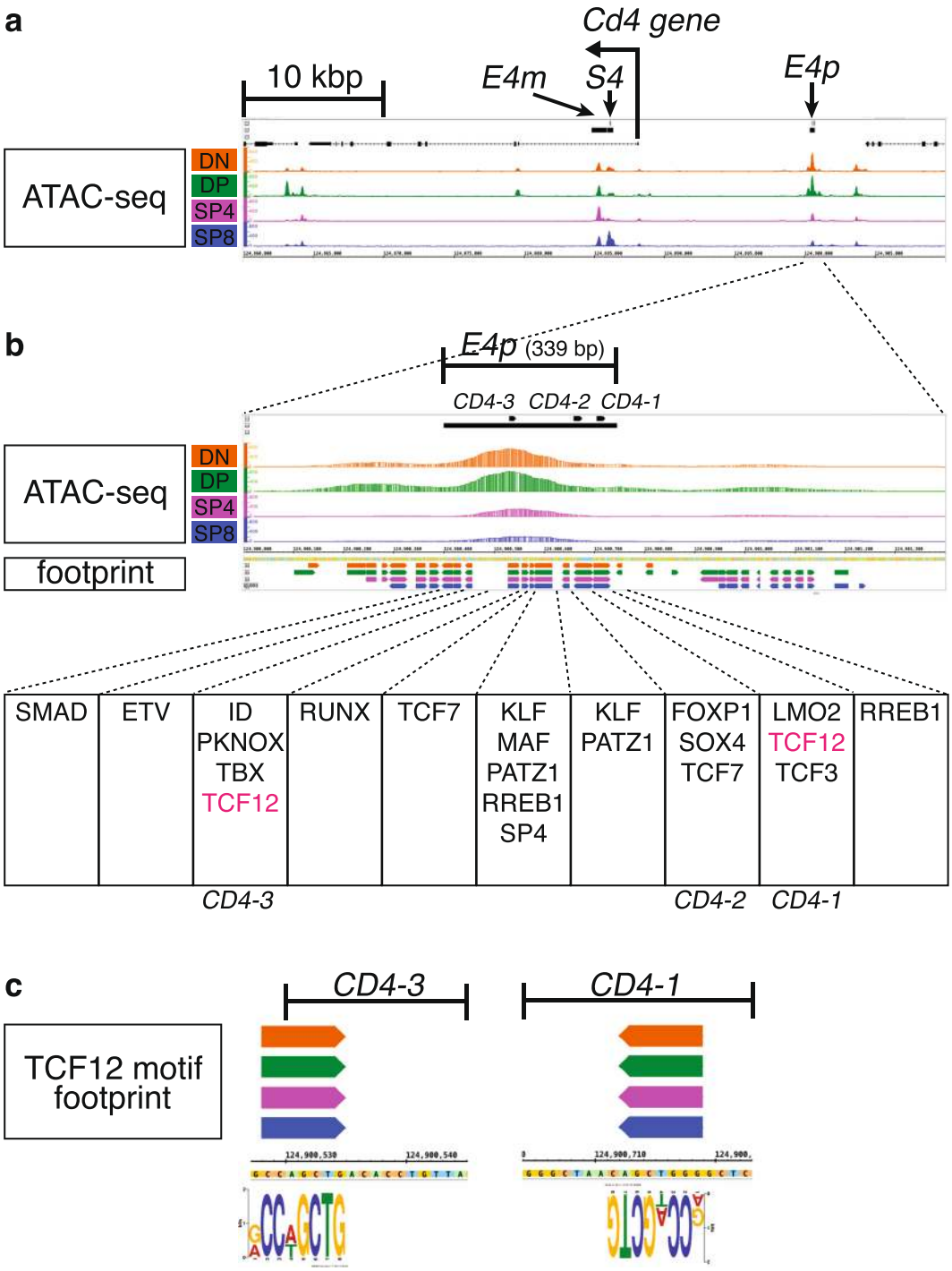

Figure 2. ATAC-seq signal and CENTIPEDE footprint calls around the functionally validated *E4p Cd4* gene enhancer. (**a**) ATAC-seq signals are shown on the IGB browser within around 50 kbp of the *Cd4* locus; mm10, chr6:124,860,001–124,910,000. The positions of the *E4p*, *E4m* enhancers and *S4* silencer[40–42] are shown at the top. (**b**) ATAC signal and footprint calls around *E4p* are depicted. *CD4-1, CD4-2* and *CD4-3* sequences were first identified by DNaseI footprinting in the SL3B T cell line[40]. (**c**) TCF12 (aka HEB) motif footprints in the *CD4-1* and *CD4-3* sequences. Footprint calls within *S4* and *E4m* are shown in Supplementary Fig. S8.

Fig. S12), but this approach did not capture the nuanced enrichments we detected with the footprinting approach, as HOMER only takes the motif occurrences and not the ATAC-seq signal into account. These data support the concept that many TFs bind to specific transcriptional regulatory elements at each developmental stage to achieve stage-specific gene expression patterns, and that the binding of these individual factors is reflected in the dynamic changes in transcriptional networks that must accompany thymocyte developmental progression from one stage to the next.

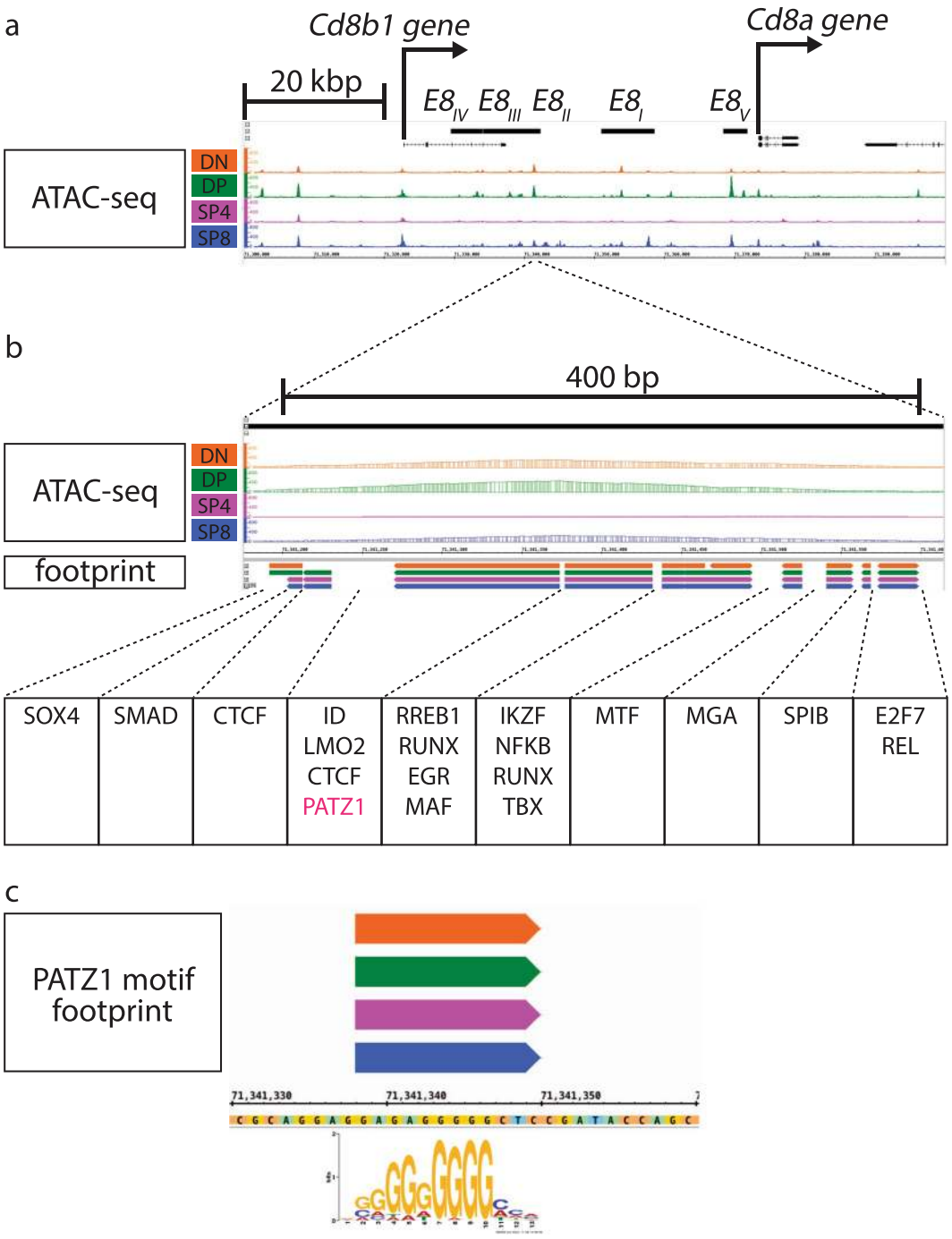

Figure 3. ATAC-seq signal and footprint calls within functionally validated enhancers for the *Cd8* gene. (**a**) ATAC-seq signals are shown on the IGB browser within around 100 kbp of the *Cd8* locus; mm10, chr6:71,300,001-71,400,000. The positions of the $E8_I$ - $E8_V$ enhancers[43–45] are depicted at the top (**b**) ATAC signals and footprint calls at an ATAC peak identified in $E8_{II}$ are shown. (**c**) A PATZ1 motif footprint within $E8_{II}$ is shown. Footprint calls within $E8_I$ and $E8_V$ are shown in Supplementary Fig. S9.

These footprinting data identified potential stage-specific regulators. Out of the 34 SOX family TF motifs tested, 10 and 9, respectively, were within the top 20 enrichment scores for DN and DP, but not in in SP4 and SP8 (Supplementary Table S3), suggesting that a SOX family TF(s) is important for DN- and DP-specific gene expression. Of 3 PBX family TF motifs tested, 2 were within the top 20 fold-enrichment scores for the DP-specific cluster (Supplementary Table S3). These data suggest a role for PBX family TFs in DP-specific gene expression. Out of 4 PKNOX family TF motifs tested, 3 were within the top 20 enrichment scores for the SP4 and SP8 clusters (Supplementary Table S3), indicating that PKNOX family TFs contribute to SP4- and SP8-specific gene transcription. Finally, out of 7 MAF family TF motifs tested, 1 and 3 were within the top 20 fold-enrichment scores for the

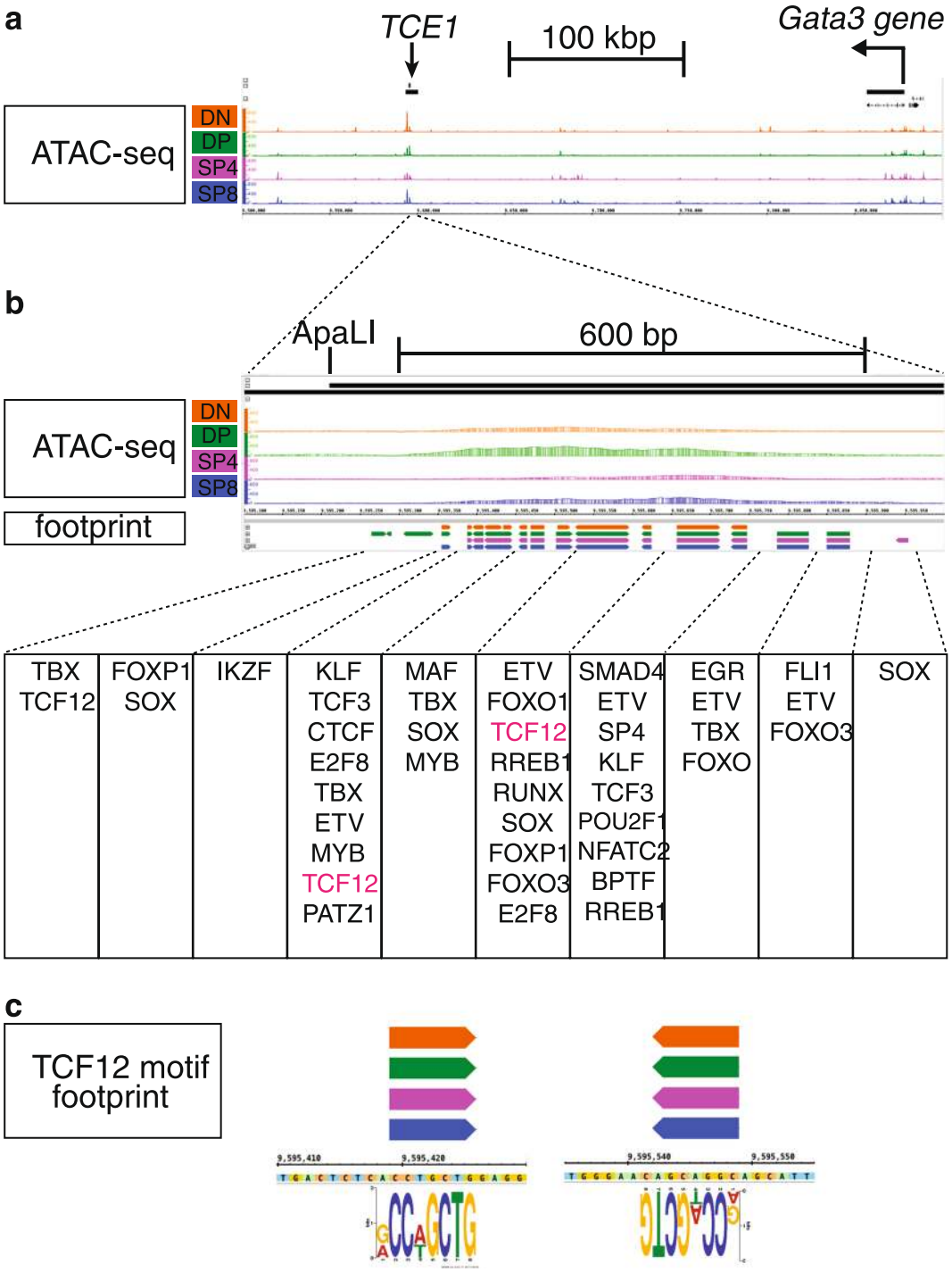

Figure 4. ATAC-seq signal and footprint calls within the functionally validated *TCE1 Gata3* enhancer. (**a**) ATAC-seq signals are shown on the IGB browser around 400 kbp of the *Gata3* gene; mm10, chr2: 9,500,001–9,900,000. (**b**) ATAC peak and TF footprint calls at an ATAC peak found in *TCE1* core[33,34]. (**c**) TCF12 (aka HEB) motif footprints.

SP4 and SP8 clusters, respectively (Supplementary Table S3), supporting the hypothesis that MAF family TFs are important for SP8-specific gene expression.

Discussion

We performed ATAC-seq experiments and footprint analyses at four major stages of thymocyte development in order to compile a catalogue of stage-specific accessible chromatin sequences as well as to identify specific sequences bound by TFs. We identified ubiquitous and stage-specific open chromatin regions, recapitulating the identity of functionally validated regulatory elements, as well as revealing novel regulatory loci. The ATAC-seq

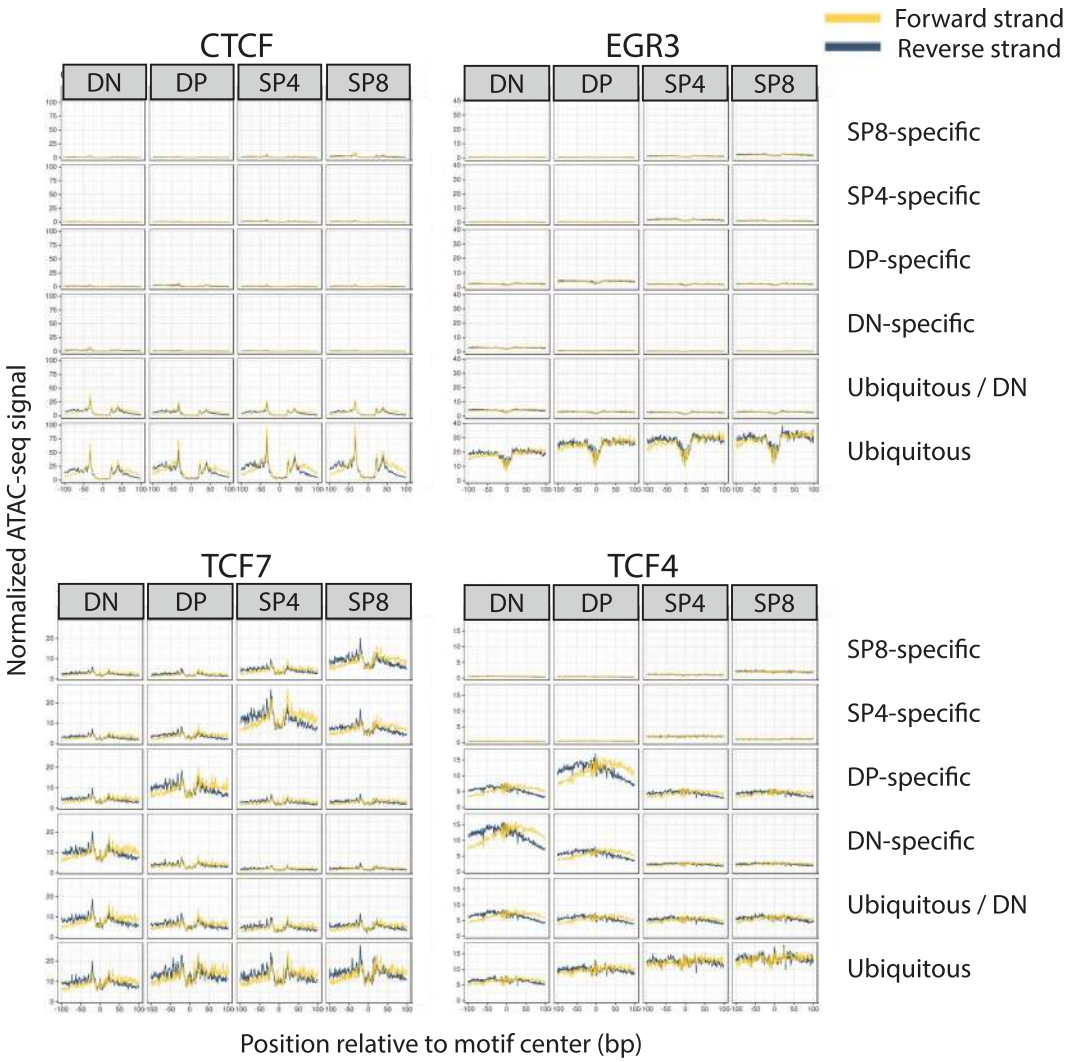

Figure 5. Footprint occupancies across samples and clusters. Normalized occupancy signals (see Methods) at ±100 bp of motif center for CTCF, EGR3, TCF7 and TCF4. Horizontal facets correspond to the ATAC-seq samples, and vertical facets correspond to the *k*-means clusters.

footprinting data for predicted αβT cell activators and repressors highlighted TF-bound motifs within those regulatory regions, as well as bound motifs that were enriched in each of the thymocyte stage-specific accessible chromatin clusters, providing an in-depth view into the inner regulatory workings of thymocyte development. We identified between 8 and 20 novel sequences that were predicted to be bound by proteins within previously identified regulatory elements for the *Cd4*, *Cd8*, *Trb* and *Gata3* genes, which supports the idea that an approximately 8–20 TFs bind to an enhancer in order to form a TF complex/enhanceosome that is capable of supporting the initiation and/or activation of enhancer activity. Thus one future goal is to investigate the ability of individually bound sequences to contribute to enhancer activity, which can be tested by *in vivo* ablation or mutation of specific TF motifs. The genome-wide footprinting approach detailed here is an alternative to ChIP experiments, but the two are complementary. It is well known and has been documented that several different proteins can bind to a given sequence motif (*e.g.* all six vertebrate GATA factors bind with reasonably high affinity to the AGATAA sequence motif, so identification of a given *cis* element in the absence of data regarding the tissue specificity of a given family of factors may only be marginally informative). ChIP experiments, in contrast, can capture indirect binding by virtue of protein-protein interactions that occur in larger complexes formed with a specific DNA binding protein[30,31], potentially complicating assignment of which factor is genuinely bound to DNA at any given site.

The thymocyte stage-specific open chromatin regions identified here by ATAC-seq followed by *k*-means clustering approach highlighted the positions for thousands of potentially novel developmental stage-specific regulatory elements. The ATAC peaks provided evidence for the previously predicted closed- or open-chromatin status in both the *Cd4* and *Cd8* loci during thymocyte development[32]. Furthermore, the identification of two major ATAC peaks within the 7.1 kbp that originally defined the *Gata3* enhancer, *TCE1*[33,34], suggests that one or both of these two open chromatin domains (of approximately 600 bp and 500 bp) play a major role in the enhancer activity of *TCE1*. In agreement with this hypothesis, one of these ATAC-seq peaks aligns perfectly with a 1.2 kbp "core"

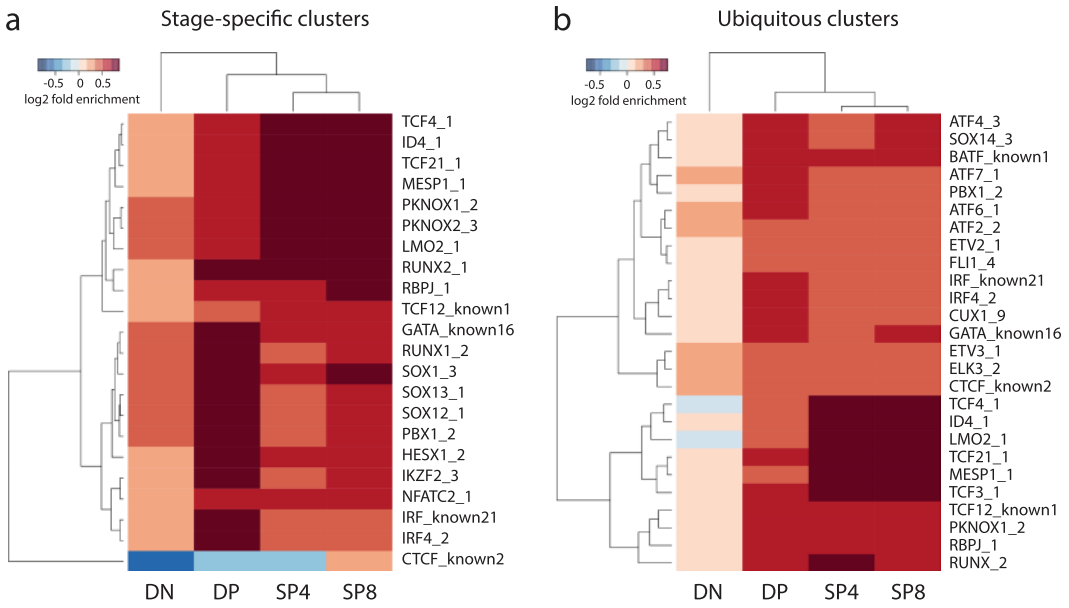

Figure 6. Individual footprint enrichments in each of the samples. (**a**) GAT enrichment scores for the top factors that maximize the variance between the stage-specific clusters of each of the ATAC-seq samples. (**b**) Similar to (**a**), but comparing the enrichments for each ATAC-seq sample in the ubiquitous cluster. Darker red colors correspond to stronger enrichments, and blue colors correspond to depletions.

sequence that exerts similar reporter gene activation in thymocytes of transgenic mice that is roughly equivalent to the whole 7.1 kbp *TCE1* sequence[34].

Enrichment of footprints in the stage-specific open chromatin clusters highlighted TF families binding to the motif as potential stage-specific regulators. These data provide an additional layer of information to the αβ T cell factors[28] predicted from lineage specific gene expression profiles. The most immediate future plans following these identifications are to investigate whether or not each bound sequence is necessary for any specific enhancer/silencer activity, which can be tested by *in vivo* genomic DNA mutation of the TF motif. In summary, the genome-wide view of open chromatin presented here as well as the identification of the sequence motifs bound by TFs at four different stages of thymocyte development is a useful point from which to begin to assemble precise models for transcriptional regulation of T cell stage-specific gene expression.

Methods

ATAC-seq. ATAC libraries were prepared as described previously[17]. In brief, 50,000 to 100,000 DN, DP, CD4 SP and CD8 SP thymocytes were isolated by flow cytometry (Supplementary Fig. S1). Cells were processed for ATAC reaction, and then the ATAC libraries were PCR amplified with barcoded primers. The ATAC-seq libraries were paired end 75 bp sequenced on a HiSeq 4000 at the UM Sequencing Core. Raw reads were trimmed for barcodes and aligned to the mm10 reference genome using BWA[19]; duplicates were removed with Picard and then filtered for high quality (mapq $\geq$ 30), properly paired alignments and uniquely mapped as described in our previous study[12].

Peak calling. In order to account for sequencing depth differences between each library, we down-sampled reads (keeping read pairs intact) to the median depth of all libraries after the pruning steps described above. This ensures that sequencing depth would not confound the analysis. After this step, we combined all replicates from each stage into a single BAM file to increase sequencing depth (ranging from 120 to 134 million reads per stage) and called peaks using MACS2[20] with options *–nomodel –shift -50 –extsize 100 -B –keep-dup all*. For testing the reproducibility between samples, we generated a set of regions that were called (narrow) peaks in at least one of the merged samples, retrieved the number of fragments mapping to these regions in each replicate and calculated the pairwise Pearson correlations between all replicates of the same stage.

***k*-means clustering and functional enrichments.** To perform k-means clustering, we generated a set of genomic regions that were called peaks in at least one of the samples (master peaks list) by using bedtools merge in the combined MACS2 output for all samples. For each sample, we calculated the FPKM in each of the master peaks regions, and normalized the signal by dividing the values by the TSS enrichment of the sample, which accounts for the signal-to-noise ratio, and then applied robust IQR scaling $\left(X_{scaled} = \frac{x_i - median(X)}{IQR(X)}\right)$, where IQR is the distance between the 1st and 3rd quartiles, to make the values comparable across samples. This signal was then row-wise normalized by the maximum of every sample $\left(Y_{normalized} = \frac{y_i}{\max(Y)}\right)$. Using this matrix of genomic coordinates per samples, we ran the k-means implementation available in R 3.3.1 for $k = 1, 2, \ldots, 15$ k values and determined that $k = 6$ was suitable for our analyses. Increasing k to higher values only marginally decreased variance and yielded repetitive clusters patterns, with 1,000 random starts for robustness. We analyzed the within

cluster variances for all (Supplementary Fig. S5). In order to perform functional annotation of the clusters, we used the ChIP-Enrich R package[24], which allow us to directly compare the enrichment scores and *p* values for the same GO terms across samples.

PWM scans and ATAC-seq footprints. In the current study we focused on TF binding motifs for T cell activators and repressors that were predicted by Jojic *et al.*[28] from stage-specific gene expression profiling (171 αβT cell factors, Supplementary Table 10 in ref.[28]). Position weight matrix (PWMs) for each motif was obtained from ENCODE[35], JASPAR[36] and TF pairs identified by Jolma *et al.*[37]. Total 417 binding motifs for 67 out of 171 Jojic αβ T cell factors were derived from these databases.

We scanned the mm10 genome for the PWMs for the 417 motifs using FIMO[38] with the G-C content background frequency for mm10 (41.7%), and used the default 10^{-4} P value threshold, also filtering for motif occurrences intersecting regions with known mapability issues (blacklisted regions). CENTIPEDE[25] was used to call footprints from the ATAC-seq data as we have done previously[12,13]. Briefly, for each PWM scan result we generated a strand-specific (relative to the motif orientation) single base pair resolution matrix encoding the number of Tn5 transposase integration events in a region ±100 bp from each motif occurrence. A motif occurrence was considered bound if the CENTIPEDE posterior probability was higher than 0.99 and its coordinates were entirely contained by an ATAC-seq peak. To generate the motif occupancy plots for each factor, we aggregated the signal used as input for CENTIPEDE for all the predicted bound motifs, as well as an equal number of motifs with posteriors less than or equal to 0.5 and not intersecting ATAC-seq peaks in that sample. The normalized signal plotted was obtained by dividing the bound signal by the unbound.

Overlap of ATAC-seq footprint and ChIP-seq. In order to test the correspondence between footprint calls and ChIP-seq data for GATA and CTCF, we used GAT[27] with the workspace set as all the GATA or CTCF motif matches in the mm10 genome, the respective ChIP-seq peaks as the segments, and the respective CENTIPEDE footprint calls as the annotation. By limiting the workspace only to the specific motifs, the data stringently delimit the space for genomic interval overlap testing. The footprint enrichments in the ATAC-seq clusters were performed separately for each sample and for each motif. We used as workspace all the motif occurrences within the master peaks regions (see *k*-means clustering above) for the individual motif being analyzed. As annotations, we used the cluster designations from the *k*-means analysis. The segments were all the footprints for that motif in that sample. Additionally, we used the option -n to 1,000 in order to increase statistical robustness. This resulted in a table with the GAT results for every motif in each cluster and in each sample.

Data Availability. ATAC-seq and footprint data have been deposited in GEO database[39] and accessible through accession number GSE107076.

References

1. Scripture-Adams, D. D. *et al*. GATA-3 dose-dependent checkpoints in early T cell commitment. *J Immunol* **193**, 3470–3491 (2014).
2. Allman, D. *et al*. Thymopoiesis independent of common lymphoid progenitors. *Nat Immunol* **4**, 168–174 (2003).
3. Bell, J. J. & Bhandoola, A. The earliest thymic progenitors for T cells possess myeloid lineage potential. *Nature* **452**, 764–767 (2008).
4. Van de Walle, I. *et al*. GATA3 induces human T-cell commitment by restraining Notch activity and repressing NK-cell fate. *Nat Commun* **7**, 11171 (2016).
5. Raulet, D. H., Garman, R. D., Saito, H. & Tonegawa, S. Developmental regulation of T-cell receptor gene expression. *Nature* **314**, 103–107 (1985).
6. Pardoll, D. M. *et al*. Differential expression of two distinct T-cell receptors during thymocyte development. *Nature* **326**, 79–81 (1987).
7. Zerrahn, J., Held, W. & Raulet, D. H. The MHC reactivity of the T cell repertoire prior to positive and negative selection. *Cell* **88**, 627–636 (1997).
8. Heng, T. S. P., Painter, M. W. & Immunological Genome Project Consortium The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* **9**, 1091–1094 (2008).
9. Mingueneau, M. *et al*. The transcriptional landscape of αβ T cell differentiation. *Nat Immunol* **14**, 619–632 (2013).
10. Thanos, D. & Maniatis, T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).
11. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111–1123 (2007).
12. Scott, L. J. *et al*. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat Commun* **7**, 11764 (2016).
13. Varshney, A. *et al*. Genetic regulatory signatures underlying islet gene expression and type 2 diabetes. *Proc Natl Acad Sci USA* **114**, 2301–2306 (2017).
14. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science* **193**, 848–856 (1976).
15. Stalder, J., Groudine, M., Dodgson, J. B., Engel, J. D. & Weintraub, H. Hb switching in chickens. *Cell* **19**, 973–980 (1980).
16. Stalder, J. *et al*. Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. *Cell* **20**, 451–460 (1980).
17. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).
18. Wall, L., deBoer, E. & Grosveld, F. The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev* **2**, 1089–1100 (1988).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
20. Zhang, Y. *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
21. Freese, N. H., Norris, D. C. & Loraine, A. E. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* **32**, 2089–2095 (2016).
22. Yue, F. *et al*. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
23. Gaspar-Maia, A., Alajem, A., Meshorer, E. & Ramalho-Santos, M. Open chromatin in pluripotency and reprogramming. *Nat Rev Mol Cell Biol* **12**, 36–47 (2011).
24. Welch, R. P. *et al*. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Res* **42**, e105 (2014).
25. Pique-Regi, R. *et al*. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21**, 447–455 (2011).
26. Wei, G. *et al*. Genome-wide Analyses of Transcription Factor GATA3-Mediated Gene Regulation in Distinct T Cell Types. *Immunity* **35**, 299–311 (2011).

27. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046–2048 (2013).
28. Jojic, V. *et al*. Identification of transcriptional regulators in the mouse immune system. *Nat Immunol* **14**, 633–643 (2013).
29. Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V. & Zaret, K. S. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev* **23**, 804–809 (2009).
30. Porcher, C., Liao, E. C., Fujiwara, Y., Zon, L. I. & Orkin, S. H. Specification of hematopoietic and vascular development by the bHLH transcription factor SCL without direct DNA binding. *Development* **126**, 4603–4615 (1999).
31. Wright, C. W. & Duckett, C. S. The aryl hydrocarbon nuclear translocator alters CD30-mediated NF-kappaB-dependent transcription. *Science* **323**, 251–255 (2009).
32. Kioussis, D. & Ellmeier, W. Chromatin and CD4, CD8A and CD8B gene expression during thymic differentiation. *Nat Rev Immunol* **2**, 909–919 (2002).
33. Hosoya-Ohmura, S. *et al*. An NK and T cell enhancer lies 280 kilobase pairs 3′ to the gata3 structural gene. *Mol Cell Biol* **31**, 1894–1904 (2011).
34. Ohmura, S. *et al*. Lineage-affiliated transcription factors bind the Gata3 Tce1 enhancer to mediate lineage-specific programs. *J Clin Invest* **126**, 865–878 (2016).
35. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976–2987 (2014).
36. Mathelier, A. *et al*. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110–D115 (2016).
37. Jolma, A. *et al*. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
38. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, W39–W49 (2015).
39. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).
40. Sawada, S. & Littman, D. R. Identification and characterization of a T-cell-specific enhancer adjacent to the murine CD4 gene. *Mol Cell Biol* **11**, 5506–5515 (1991).
41. Siu, G., Wurster, A. L., Duncan, D. D., Soliman, T. M. & Hedrick, S. M. A transcriptional silencer controls the developmental expression of the CD4 gene. *EMBO J* **13**, 3570–3579 (1994).
42. Wurster, A. L., Siu, G., Leiden, J. M. & Hedrick, S. M. Elf-1 binds to a critical element in a second CD4 enhancer. *Mol Cell Biol* **14**, 6452–6463 (1994).
43. Hostert, A. *et al*. A region in the CD8 gene locus that directs expression to the mature CD8 T cell subset in transgenic mice. *Immunity* **7**, 525–536 (1997).
44. Hostert, A. *et al*. A CD8 genomic fragment that directs subset-specific expression of CD8 in transgenic mice. *J Immunol* **158**, 4270–4281 (1997).
45. Ellmeier, W., Sunshine, M. J., Losos, K., Hatam, F. & Littman, D. R. An enhancer that directs lineage-specific expression of CD8 in positively selected thymocytes and mature T cells. *Immunity* **7**, 537–547 (1997).

Acknowledgements

This work was supported by National Institute of Health Grant AI094642 (to T.H. and J.D.E). The research was also supported in part by the National Institute of Health through the University of Michigan Comprehensive Cancer Center Support Grant (P30 CA046592) for use of the Flow Cytometry and the Sequencing Cores at the University of Michigan.

Author Contributions

T.H. designed the study, performed experiments, analyzed the data, performed bioinformatics analyses and wrote the manuscript. R.D.A. designed the study, performed bioinformatics analyses and wrote the manuscript. G.M. assisted with experiments. J.H. performed bioinformatics analyses and edited the manuscript. Y.K. and J.K. provided essential reagents and protocols for ATAC-seq and edited the manuscript. S.P. designed the study and edited the manuscript. J.D.E. designed the study and wrote the paper.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-23774-9.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.