nature immunology

Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells

Philip Brennecke[1,2,5], Alejandro Reyes[3,5], Sheena Pinto[4,5], Kristin Rattay[4,5], Michelle Nguyen[1,2], Rita Küchler[4], Wolfgang Huber[3,6], Bruno Kyewski[4,6] & Lars M Steinmetz[1–3,6]

Expression of tissue-restricted self antigens (TRAs) in medullary thymic epithelial cells (mTECs) is essential for the induction of self-tolerance and prevents autoimmunity, with each TRA being expressed in only a few mTECs. How this process is regulated in single mTECs and is coordinated at the population level, such that the varied single-cell patterns add up to faithfully represent TRAs, is poorly understood. Here we used single-cell RNA sequencing and obtained evidence of numerous recurring TRA–co-expression patterns, each present in only a subset of mTECs. Co-expressed genes clustered in the genome and showed enhanced chromatin accessibility. Our findings characterize TRA expression in mTECs as a coordinated process that might involve local remodeling of chromatin and thus ensures a comprehensive representation of the immunological self.

Discrimination between self and non-self, including self-tolerance, is a hallmark of the adaptive immune system, and when this subtle distinction fails, various autoimmune diseases have been shown to develop[1,2]. Self-tolerance of T cells, as imposed in the thymus (i.e., central tolerance), relies on the exhaustive scanning of self antigens by maturing T cells[3]. Distinct types of thymic antigen-presenting cells display a broad range of self antigens in a partly redundant and partly complementary fashion[4]. Among the various thymic antigen-presenting cells, medullary thymic epithelial cells (mTECs) stand out due to their unique ability to ectopically express a wide range of tissue-restricted self antigens (TRAs)[5,6]. In mTECs, TRAs, whose expression outside of the thymus is tightly controlled in time and space, become accessible to developing T cells when they are still most responsive to tolerance imprinting. The induction of self-tolerance operates via two modes, either through the elimination of self-reactive T cells or by cell-fate diversion toward the regulatory T cell lineage[3,4,7–9]. Typically, each TRA protein is expressed in only 1–3% of mTECs, and thus TRA expression follows a mosaic pattern. Therefore, the availability of self antigens is a potential limiting factor during the induction of self-tolerance[4,10–12].

Many aspects of the complex molecular regulation of thymic TRA expression are poorly understood; the transcriptional regulator Aire, which is responsible for the expression of a large part of ectopically expressed TRAs in the thymus, represents a notable exception[1,13–15]. Aire targets inactive chromatin either directly, by binding to the repressive chromatin mark H3K4me0 (histone H3 not methylated at Lys4) with its PHD1 finger domain[16,17], or indirectly, through its binding partners, such as the ATF7ip-MBD1 complex[18] or the Cdh4 protein[19]. These proteins are thought to recruit Aire to methylated CpG dinucleotides at repressed promoters and polycomb-silenced chromatin, respectively. Upon being recruited to silent chromatin, Aire is believed to promote ectopic expression of TRA-encoding genes by releasing stalled polymerase II from their promoters[20]. Such studies indicate that Aire 'preferentially' targets inactive chromatin, potentially using multiple mechanisms. However, it remains unclear which underlying rules govern the patterning of thymic TRA expression at the single-cell level, such that the composite of mTECs reliably covers the combined transcriptomes of peripheral tissues. It is also unclear whether each mTEC samples a random set of TRAs or whether there are constraints on the set of TRAs that individual mTECs express. Likewise, it remains elusive how thymic TRA expression is coordinated at the intra- and intercellular levels in time and space, as well as how stable these patterns are throughout the lifetime of an individual mTEC.

Published studies have addressed some of those questions by applying bulk transcriptome analysis, single-cell multiplex PCR and single-cell RNA sequencing (scRNA-seq)[10,12,19,21]. Such studies have indicated that single mTECs express genes encoding TRAs of diverse functional categories, which challenges the proposal that thymic TRA expression mimics tissue-specific gene-expression patterns at the single-cell level. However, while multiple studies using single-cell approaches have not discerned TRA–co-expression patterns in single mouse mTECs[10,19,21], a study of human mTECs has provided evidence of the co-regulation of TRAs within single cells[12]. Identifying

[1]Department of Genetics, Stanford University, School of Medicine, California, USA. [2]Stanford Genome Technology Center, Stanford University, California, USA. [3]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [4]Division of Developmental Immunology, German Cancer Research Center, Heidelberg, Germany. [5]These authors contributed equally to this work. [6]These authors jointly directed this work. Correspondence should be addressed to L.M.S. (larsms@stanford.edu), W.H. (whuber@embl.de) or B.K. (b.kyewski@dkfz.de).

Received 10 February; accepted 8 July; published online 3 August 2015; doi:10.1038/ni.3246

the molecular mechanisms that regulate thymic TRA expression in single cells is key to understanding how the diversity of ectopically expressed self antigens, a prerequisite of self-tolerance, is generated in the mTEC compartment.

Hence, we applied scRNA-seq to mouse mTECs and studied the single-cell expression profiles of 203 mature ($MHCII^{hi}$) mTECs, as well as three mature mTEC subsets selected for their expression of particular TRAs. We focused our study on mature mTECs, as they represent the mTEC subset mainly responsible for inducing self-tolerance in developing T cells by expressing the largest diversity of TRAs. At the same time, they are fully competent antigen-presenting cells with high surface expression of major histocompatibility complex class II (MHCII) and the maturation marker CD80 (B7-1). Using this genome-wide approach, we found that the mature mTEC population at large was composed of cells with numerous distinct co-expression clusters of TRA-encoding genes. Each cluster comprised only a fraction of all genes, and individual clusters were expressed only in a small subset of mTECs. Our findings characterize thymic TRA expression as a highly regulated process that ensures representation of the full diversity of self antigens in the mTEC compartment by assembling a population composite of recurrent and complementary co-expression clusters present in individual cells.

RESULTS

Comprehensive coverage of the immunological self by mTECs

To investigate the extent of heterogeneity and patterning of thymic TRA expression in single mTECs, we performed scRNA-seq on mature $MHCII^{hi}$ mouse mTECs (called 'mature mTECs' here). We sorted single mature mTECs ($PI^-CD45^-Ly51^-EpCAM^+MHCII^{hi}$) from pooled thymic tissue of 4- to 6-week-old female C57BL/6 mice (5–20 mice) and generated 211 single-cell cDNA libraries using a modified version of the Smart-seq2 method[22,23]. After implementing data quality control, we retained 203 cells (96%) for further analysis (**Supplementary Code**). For each mTEC, we counted the protein-coding genes and TRA-encoding genes (i.e., a subset of protein-coding genes) whose expression was detected by scRNA-seq. We found that the number of TRA-encoding genes detected within a single cell was proportional to the total number of genes detected (19% ± 3.6% of genes detected were classified as TRA-encoding genes) (**Fig. 1a** and **Supplementary Fig. 1**). We did not observe evidence of cell-to-cell variation in the proportion of expressed TRA-encoding genes, as the variation in the number of TRA-encoding genes detected per mTEC could be explained by varying sequencing coverage (**Fig. 1a**). Moreover, 95% of the previously reported 3,976 TRA-encoding genes[12] were cumulatively detected in the 203 mature mTECs analyzed (**Fig. 1b**). In addition, the scRNA-seq assay cumulatively detected expression of 86% of all annotated protein-coding genes in the 203 mature mTECs analyzed (19,619 of 22,740 genes; release 75 of the Ensembl project of genome databases) (**Fig. 1b**), which indicated that nearly 90% of the protein-coding genome was sampled across a few hundred mature mTECs. These data documented a comprehensive representation of the immunological self in mature mTECs at the population level, as has been suggested before[19,24].

Next we used a published method[25] to identify genes whose expression was highly variable across the 203 single mTECs. This analysis revealed a high degree of heterogeneity in gene expression across mTECs, with 9,689 genes having a biological coefficient of variation larger than 50% (i.e., a squared coefficient of variation larger than 0.25) at a false-discovery rate (FDR) of 10% (**Fig. 1c**). This set of highly variable genes showed enrichment for TRA-encoding genes compared with the abundance of TRA-encoding genes among all protein-coding genes (odds ratio = 2.2, and $P < 2.2 \times 10^{-16}$ (Fisher's exact test)). More specifically, 26% of the highly variable genes encoded TRAs, while only 14% of the genes not detected as highly variable encoded TRAs (**Supplementary Fig. 2**). Thus, mature mTECs represented a cell type that was highly heterogeneous at the level of individual cells and yet collectively seemed to reliably express most of the genome.

TRA-encoding genes are generally expressed mosaically

Next we investigated the Aire dependence of TRA expression in single mature mTECs. For this analysis, we integrated our single-cell gene-expression data with the transcriptome atlas of 91 cell types (88 primary cell types and three cell lines) acquired by the FANTOM ('functional annotation of the mammalian genome') consortium[26] and a list of Aire-regulated genes[19]. We found that Aire-dependent genes were expressed in a smaller fraction of mTECs than were Aire-independent genes (**Fig. 1d**,**e**). Moreover, we found that genes with tissue-restricted expression patterns in the periphery of the body were expressed at a low frequency in single mTECs, regardless of Aire regulation (**Fig. 1f**,**g**). When we considered a set of 912 genes detected in at most 10 of the 91 cell types from the FANTOM data set, 522 genes were Aire dependent and 390 were Aire independent (**Fig. 1f**,**g**). Of the 522 Aire-dependent genes, 94% (492) were detected in less than 15% of our single mature mTECs (**Fig. 1f**). In a similar manner, of the 390 Aire-independent genes, 68% (265) were detected in less than 15% of mTECs (**Fig. 1g**). These results indicated that genes whose expression tends to be restricted to fewer cell types in the periphery of the body were generally expressed at a low frequency in mature mTECs, with a more pronounced effect for Aire-dependent genes.

Non-random TRA-expression patterns in single mature mTECs

Next we addressed whether TRA expression in single mTECs occurs randomly—i.e., without noticeable gene–co-expression patterns[10,19,21]—or instead is governed by rules of gene co-regulation[12]. Because the cell cycle was a potential confounding factor, due to many genes being co-regulated in a cell cycle–dependent manner, we first regressed out cell-cycle variation from the 203 mature mTEC single-cell transcriptomes by the scLVM ('single-cell latent variable model') method[27]. Next we used clustering by the *k*-medoids algorithm to group highly variable Aire-dependent genes on the basis of their level of expression across cells and assessed the statistical stability of the clustering by resampling[28] (**Supplementary Code**). We identified 11 stable gene clusters (A–K) that showed patterns of co-expression and one cluster (L) that grouped together genes for which the data provided no evidence of co-expression (**Fig. 2a**). Most of these co-expression patterns showed high expression in only a small fraction of mature mTECs (**Fig. 2b**). This was consistent with the published identification of three distinct co-expression groups at low cell frequencies in human mTECs[12]. We observed a notable exception for co-expression cluster B, which was present in a larger fraction of cells (**Fig. 2**). These results suggested the existence of co-expression patterns in single mTECs and that the regulation of TRA-encoding genes followed discernible patterns in individual mature mTECs.

TRA co-expression regardless of Aire dependence

To further evaluate the concept of co-expression patterns in single mTECs, we chose an independent *in silico* analytical approach to assess the co-expression of TRA-encoding genes within mature mTECs (203 cells). For this, we selected an Aire-dependent TRA-encoding gene, *Tspan8* (encoding tetraspanin-8), which belonged to cluster B (**Fig. 2a**). We detected *Tspan8* expression in 66 of the 203 mature

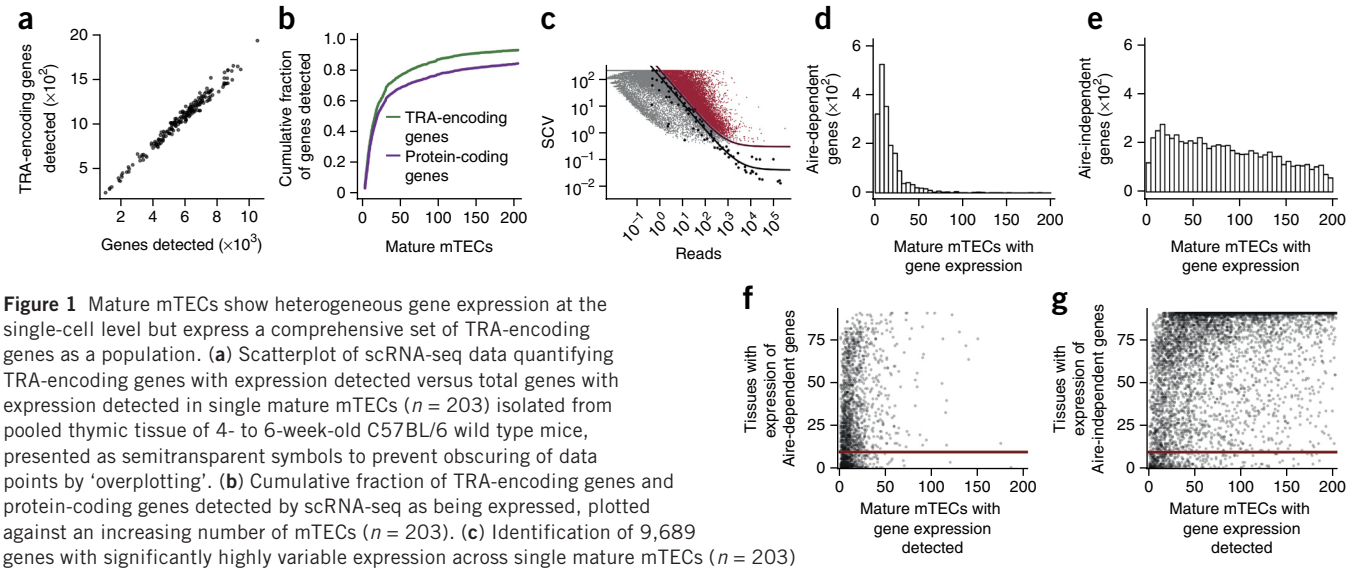

Figure 1 Mature mTECs show heterogeneous gene expression at the single-cell level but express a comprehensive set of TRA-encoding genes as a population. (**a**) Scatterplot of scRNA-seq data quantifying TRA-encoding genes with expression detected versus total genes with expression detected in single mature mTECs ($n = 203$) isolated from pooled thymic tissue of 4- to 6-week-old C57BL/6 wild type mice, presented as semitransparent symbols to prevent obscuring of data points by 'overplotting'. (**b**) Cumulative fraction of TRA-encoding genes and protein-coding genes detected by scRNA-seq as being expressed, plotted against an increasing number of mTECs ($n = 203$). (**c**) Identification of 9,689 genes with significantly highly variable expression across single mature mTECs ($n = 203$) by a published method[25]: maroon symbols indicate genes with a biological squared coefficient of variation (SCV) of >0.25 at an FDR of 10%, classified as highly variable; gray symbols indicate all other genes; black symbols indicate external control 'spike-in' RNA; solid black line indicates model fit for technical noise; purple line indicates the biological squared coefficient of variation threshold of 0.25 (i.e., 50% coefficient of variation). (**d**,**e**) Aire-dependent genes (**d**) and Aire-independent genes (**e**) as a function of the number of mature mTECs ($n = 203$) for which expression of the genes was detected. (**f**,**g**) Quantification of tissues in which expression of individual genes was detected in the FANTOM data set[26] as a function of the number of mature mTECs ($n = 203$) in which expression of the gene was detected: each data point represents one Aire-dependent gene (**f**) or Aire-independent gene (**g**); maroon horizontal line indicates the threshold value of 10. Data are representative of 203 experiments with one cell in each.

mTECs (~33%) (**Fig. 2b**). Next we assessed each of the 9,689 highly variable genes (**Fig. 1c**) to determine whether they had higher expression in the 66 cells in which we detected *Tspan8* mRNA than in the remaining 137 mTECs that lacked *Tspan8* expression. Because both Aire-dependent genes and Aire-independent genes are concomitantly upregulated upon differentiation into mature mTECs, we considered both gene sets for testing. Using this approach, we identified 595 genes as being co-expressed with *Tspan8* at an FDR of 10%; we called this the '*Tspan8*–co-expressed gene set' (**Supplementary Table 1**). This gene set consisted of 129 Aire-dependent genes and 466 Aire-independent genes (**Supplementary Table 1**). Consistent with the *k*-medoids clustering analysis (**Fig. 2a**), the 129 Aire-dependent genes showed much more overlap with the genes from cluster B than with genes of the other clusters (odds ratio = 22, $P < 2.2 \times 10^{-16}$ (Fisher's exact test); **Supplementary Fig. 3**).

We then independently confirmed the finding that the genes were indeed co-expressed with *Tspan8* by using flow cytometry to sort single mTECs expressing Tspan8 on the cell surface, by a published procedure used for human mTECs[12]. We sequenced single-cell cDNA libraries from 48 Tspan8$^+$ mature mTECs (PI$^-$CD45$^-$CDR1$^-$EpCAM$^+$MHCII^{hi}Tspan8$^+$). We found that the patterns of co-expression for both Aire-dependent genes and Aire-independent genes were highly concordant between these 48 sorted Tspan8$^+$ mTECs and the 66 unselected mature mTECs in which the expression of *Tspan8*

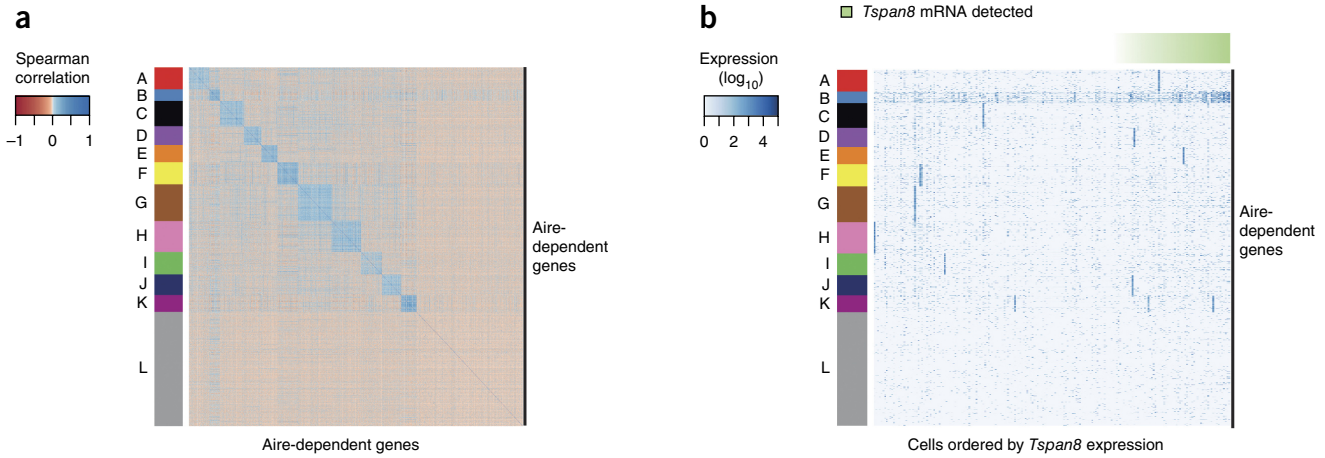

Figure 2 Single mature mTEC transcriptomes reveal numerous low-frequency sets of co-expressed genes. (**a**) Pairwise Spearman correlation matrix of the expression profiles of 2,174 highly variable Aire-dependent genes (identified in **Fig. 1c**) across mature mTECs ($n = 203$); left margin, 12 gene clusters identified by *k*-medoids clustering. (**b**) Expression of highly variable Aire-dependent genes across individual mature mTECs ($n = 203$): row order, as in **a**; columns indicate individual mature mTECs ordered by *Tspan8* expression (low (left) to high (right)). Data are representative of 203 experiments with one cell in each.

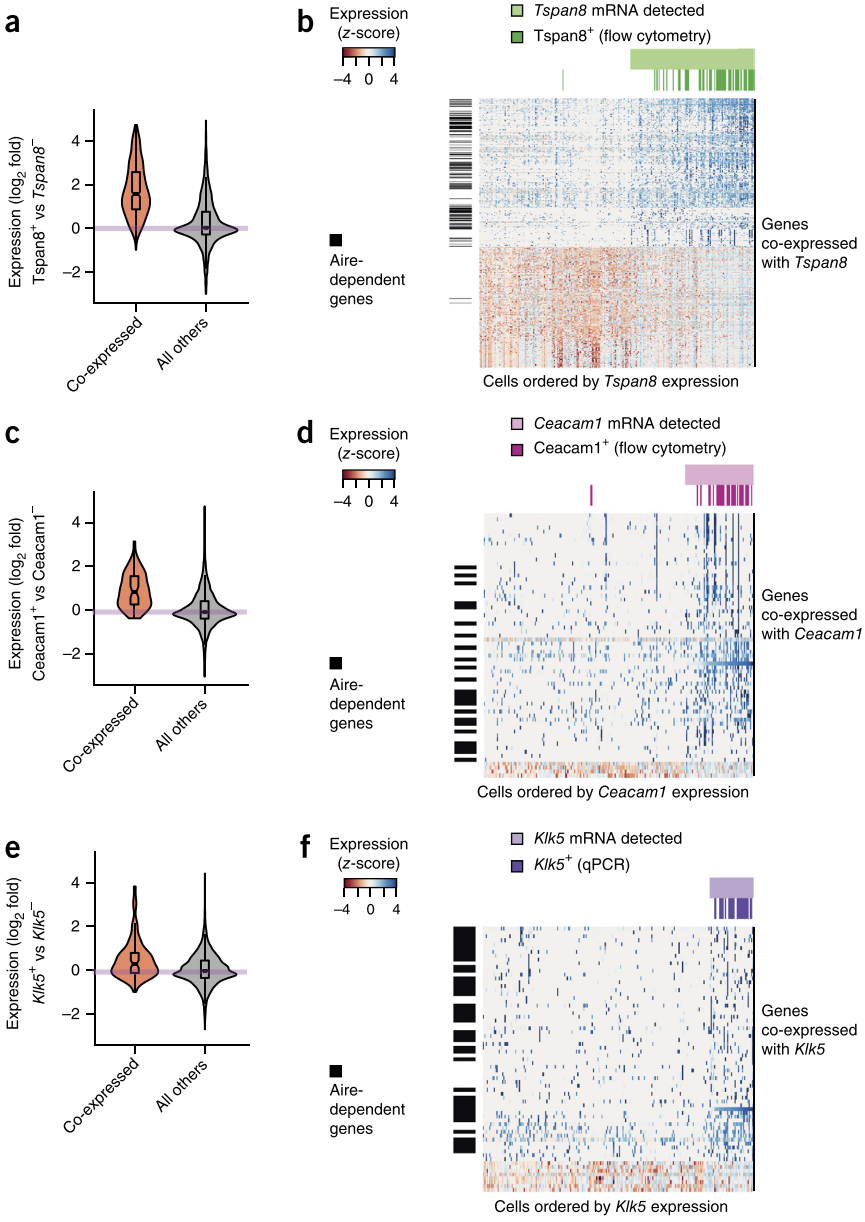

Figure 3 Confirmation of co-expression in gene sets by independent experimental approaches. (**a**) Distribution of changes in expression of the *Tspan8*–co-expressed gene set (**Supplementary Table 1**) or all other genes in the 48 Tspan8$^+$ mature mTECs selected by flow cytometry versus the 137 unselected mature mTECs for which *Tspan8* mRNA was not detected by scRNA-seq (Tspan8$^+$ vs *Tspan8*$^-$). $P < 2.2 \times 10^{-16}$ (*t*-test). (**b**) Expression of genes in the *Tspan8*–co-expressed gene set in unselected mTECs ($n = 203$) and pre-selected Tspan8$^+$ mTECs ($n = 48$): columns indicate individual cells (ordered by increasing *Tspan8* transcripts, as measured by scRNA-seq); rows indicate genes co-expressed with *Tspan8* (**Supplementary Table 1**); left margin, Aire-dependent genes. (**c**) Distribution of changes in expression (as in **a**) for the *Ceacam1*–co-expressed gene set in preselected Ceacam1$^+$ mTECs ($n = 30$) versus unselected Ceacam1$^-$ mTECs ($n = 172$) (Ceacam1$^+$ vs Ceacam1$^-$). $P = 9.8 \times 10^{-11}$ (*t*-test). (**d**) Expression of genes in the *Ceacam1*–co-expressed gene set in unselected mTECs ($n = 203$) and preselected Ceacam1$^+$ mTECs ($n = 30$), presented as in **b**. (**e**) Distribution of changes in expression (as in **a**) for the *Klk5*–co-expressed gene set in preselected *Klk5*$^+$ (with mRNA detected by quantitative PCR (qPCR)) ($n = 24$) versus unselected *Klk5*$^-$ mTECs ($n = 190$) (*Klk5*$^+$ vs *Klk5*$^-$). $P = 8.2 \times 10^{-5}$ (*t*-test). (**f**) Expression of genes in the *Klk5*–co-expressed gene set in unselected mTECs ($n = 203$) and preselected *Klk5*$^+$ mTECs ($n = 24$), presented as in **b**. Data are representative of 185 experiments (**a**) 251 experiments (**b**), 202 experiments (**c**), 233 experiments (**d**), 214 experiments (**e**) or 227 experiments (**f**) with one cell in each.

mRNA was detected initially (**Fig. 3a**,**b**). Specifically, 96% of the genes belonging to the *Tspan8*–co-expressed gene set were also upregulated in the 48 sorted Tspan8$^+$ cells (**Fig. 3a** and **Supplementary Fig. 4**; $P < 2.2 \times 10^{-16}$ (*t*-test)).

To further confirm co-expression in mature mTECs for both Aire-dependent genes and Aire-independent genes, we repeated the strategy followed for *Tspan8* for two additional TRA-encoding genes. First, we selected the gene encoding the cell-adhesion protein Ceacam1, an Aire-independent TRA-encoding gene detected as being co-expressed with *Tspan8* (**Supplementary Table 1**). As we had done for *Tspan8*, we screened the 203 mature mTECs for the presence of *Ceacam1* transcripts and detected expression of *Ceacam1* in 15% of the mature mTECs (31 of the 203 cells). We found 65 genes (23 Aire-dependent genes and 42 Aire-independent genes) that were co-expressed with *Ceacam1* at a FDR of 10%; we called this the '*Ceacam1*–co-expressed gene set' (**Supplementary Table 1**). Next we confirmed the co-expression in this gene set with *Ceacam1* by sequencing 30 single mTECs selected by flow cytometry for surface expression of Ceacam1 (PI$^-$CD45$^-$CD R1$^-$EpCAM$^+$MHCII^{hi}Ceacam1$^+$) (**Fig. 3c**,**d**). Of the 65 genes belonging to the *Ceacam1*–co-expressed gene set, 92% showed consistent upregulation in the Ceacam1$^+$ mTECs selected by flow cytometry, compared with their expression in the unselected Ceacam1$^-$ mTECs (**Fig. 3c**,**d** and **Supplementary Fig. 4**; $P = 9.8 \times 10^{-11}$ (*t*-test)).

Both *Tspan8* and *Ceacam1* were expressed relatively frequently across the mature mTEC population (33% and 15%, respectively). Thus, we also assessed a TRA-encoding gene, *Klk5*, that was expressed at a more representative frequency, and was assigned to cluster D in the *k*-medoids clustering (**Fig. 2a**). As we had defined *Tspan8* and *Ceacam1*, we defined the '*Klk5*–co-expressed gene set' on the basis of detection of *Klk5* transcripts in 13 of the 203 mature mTECs (6.4%) (**Supplementary Table 1**). The *Klk5*–co-expressed gene set consisted of 68 genes: 39 Aire-dependent genes and 29 Aire-independent genes (**Supplementary Table 1**). Consistent with the *k*-medoids clustering (**Fig. 2a**), these 39 Aire-dependent genes showed significant enrichment among the genes from cluster D compared with their abundance among the rest of the clusters (odds ratio = 4.7, $P = 8.2 \times 10^{-5}$ (Fisher's exact test); **Supplementary Fig. 5**).

We experimentally confirmed the finding that the genes were indeed co-expressed with *Klk5* by screening 562 mature mTEC cDNA libraries confirmed to be positive for the housekeeping gene *Ubc* (encoding ubiquitin C) by quantitative PCR. 28 of the 562 mTECs (5.0%) were also positive for *Klk5* expression, as determined by quantitative

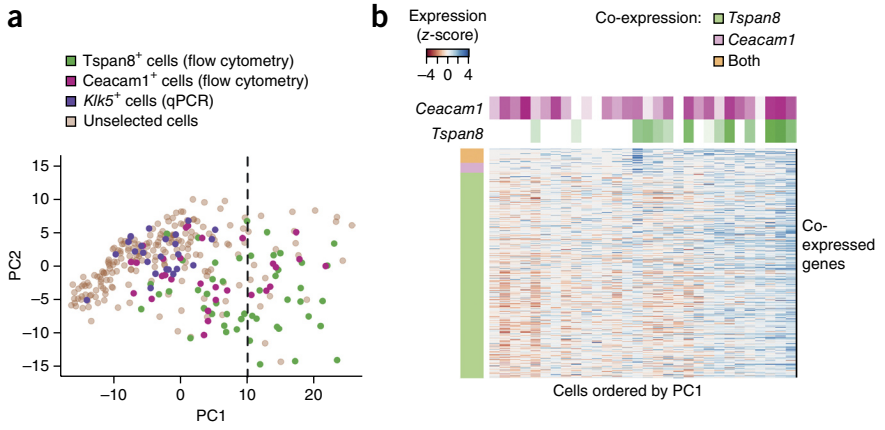

Figure 4 The *Tspan8*– or *Ceacam1*–co-expressed gene sets overlap, and corresponding mTECs are organized along a gradient of *Tspan8* expression. (**a**) Principal-component analysis of all mature mTECs sequenced ($n = 305$: 203 unselected mTECs, and 48 Tspan8$^+$ mTECs, 30 Ceacam1$^+$ mTECs and 24 Klk5$^+$ mTECs), based on expression of genes in the union of the *Tspan8*– and *Ceacam1*–co-expressed gene sets; dashed vertical line indicates the threshold of 10 along the PC1 projection. (**b**) Genes (rows) detected as being co-expressed with *Tspan8* or *Ceacam1* or both (left margin) in mature preselected Ceacam1$^+$ mTECs ($n = 30$) (columns ordered by PC1); top, expression of *Tspan8* mRNA and *Ceacam1* mRNA in individual mTECs. Data are representative of 305 experiments (**a**) or 30 experiments (**b**) with one cell in each.

PCR (data not shown). Next we sequenced the transcriptomes of 24 of the *Klk5*$^+$ mTECs. In agreement with findings obtained for the 13 unselected mature mTECs in which we detected the expression of *Klk5* transcripts, 71% of the genes from this defined *Klk5*–co-expressed gene set (**Supplementary Table 1**) showed a consistent upregulation in the *Klk5*$^+$ mature mTECs selected by quantitative PCR (**Fig. 3e**,**f** and **Supplementary Fig. 4**; $P = 8.2 \times 10^{-5}$ (*t*-test)). Notably, this concordance was particularly pronounced for the genes neighboring *Klk5* in the genome (discussed below).

In addition, while we found that the three co-expressed gene sets showed enrichment for TRA-encoding genes ($P < 2.2 \times 10^{-16}$ (*Tspan8*), $P = 7 \times 10^{-15}$ (*Ceacam1*) and $P = 1.3 \times 10^{-4}$ (*Klk5*) (Fisher's exact test)), they were not restricted to genes encoding products classified as TRAs (according to the TRA definition used in this study). Thus, we identified patterns of co-expression by initial transcriptome analysis of 203 single unselected mature mTECs and by transcriptome sequencing of subsets of mature mTECs pre-selected on the basis of surface expression of three TRAs of varying population frequency: Tspan8, Ceacam1 and Klk5.

Potential genealogies within mTEC co-expression groups

We found significant overlap of the genes in the *Ceacam1*– and *Tspan8*–co-expressed gene sets (odds ratio = 23.5, and $P < 2.2 \times 10^{-16}$ (Fisher's exact test); **Supplementary Table 1**). Specifically, 39 genes belonging to the *Ceacam1*–co-expressed gene set (i.e., 60%) were co-expressed with *Tspan8*. Despite such substantial overlap, we also identified 27 genes (40% of the *Ceacam1*–co-expressed gene set) that were co-expressed only with *Ceacam1* and 557 (93% of the *Tspan8*–co-expressed gene set) that were co-expressed only with *Tspan8*. A model in which single mTECs would sequentially shift through distinct co-expression groups throughout their lifespan has been suggested[12], which would indicate the existence of overlapping co-expression patterns in mTECs during their transition between distinct groups.

To explore that hypothesis, we visualized the interrelationships of the expression profiles of all single mature mTECs (305 cells: 203 unselected mature mTECs, 48 Tspan8$^+$ mature mTECs selected by flow cytometry, 30 Ceacam1$^+$ mature mTECs selected by flow cytometry, and 24 *Klk5*$^+$ mature mTECs selected by quantitative PCR) by principal-component analysis of the expression data of all genes co-expressed in the *Ceacam1*– and *Tspan8*–co-expressed gene sets (i.e., the union of the two co-expressed gene sets). The dominant axis of gene-expression variation, principal component 1 (PC1), distinguished the 48 Tspan8$^+$ cells and 30 Ceacam1$^+$ cells from the rest of the cells, with the Tspan8$^+$ cells being separated further than the Ceacam1$^+$ cells (**Fig. 4a**). 52% of the Tspan8$^+$ mature mTECs had a PC1 projection (position along the horizontal axis) higher than 10, compared with 27% of the Ceacam1$^+$ cells. Only 10% of the unselected mTECs and none of the *Klk5*$^+$ cells had a PC1 projection higher than 10. These results suggested that a single gene-expression program was underlying most of the observed cell-to-cell variability of the selected genes and that the Tspan8$^+$ mTECs had a more pronounced adoption of this program than did the Ceacam1$^+$ mTECs.

To further expand the findings reported above, we quantified the expression of *Tspan8* mRNA (from the scRNA-Seq analysis) in the Tspan8$^+$ and Ceacam1$^+$ mTECs. We found that *Tspan8* mRNA expression correlated with the mean expression of all genes from the union of the *Tspan8*– and *Ceacam1*–co-expressed gene sets (Spearman correlation = 0.62; **Supplementary Code**). The correlation was still present when we considered only the Ceacam1$^+$ mTECs (Spearman correlation = 0.35; **Fig. 4b**). Thus, the amount of *Tspan8* mRNA in Ceacam1$^+$ mTECs was concomitant with increased expression of the co-expressed genes and increasing similarity to Tspan8$^+$ mTECs. These data were consistent with the hypothesis that individual mTECs transition from one co-expression group to another[12].

Clustering of co-expressed genes in the genome

One possible mechanism for the generation of non-random co-expression patterns could be local chromatin configurations that would allow ectopic expression of neighboring genes regardless of their regulation in peripheral tissues[6]. Ectopic expression of gene clusters in human and mouse mTECs has been reported[10,12,15,29,30]. However, because inference of clustered gene expression from heterogeneous cell populations would be misleading due to averaging of different gene-expression patterns from individual cells, only transcriptome-wide single-cell analysis can adequately address this point. Thus, for each of the 11 co-expression clusters, we calculated the median genomic distance between each gene to its nearest co-expressed gene neighbor within the same cluster. For each of the 11 clusters, we constructed a null model that allowed us to estimate the expected median genomic distance between genes given the size of the respective cluster (**Supplementary Code**). On the basis of these null models, we found that the genes from 8 of the 11 gene clusters were located in significant genomic proximity (FDR of 10%; **Supplementary Fig. 6**). To visualize these effects, we plotted the localization of each of the 11 gene clusters resulting from the *k*-medoids clustering in a karyogram (**Supplementary Fig. 7**). Despite being dispersed across the genome, many genes

from the same gene co-expression cluster were located in close genomic proximity to each other (exemplified by co-expression cluster D; **Fig. 5a**,**b**). Some of these loci comprised gene families encompassing genes encoding structurally and functionally related products. For example, four genes in cluster D encoding products belonging to 'BPI fold–containing family B' ('bactericidal permeability-increasing protein-like 1') were located consecutively in the genome on chromosome 2 (**Supplementary Fig. 8a**), while two genes (*Gstm2* and *Gstm7*) encoding products from the 'glutathione *S*-transferase-μ' family were close neighbors in the genome on chromosome 3 (**Supplementary Fig. 8b**). Notably, we also identified groups of neighboring genes that were co-expressed but encoded products with no obvious functional relationship (**Supplementary Fig. 8c**).

The locus encoding kallikrein-related peptidases (**Fig. 5c**) represented a prominent example of a structurally and functionally related family. The locus contains 27 genes encoding products belonging to the kallikrein-related peptidase family, located in close genomic proximity on chromosome 7 (**Fig. 5c**). Nine of these genes, including *Klk5*, were assigned to cluster D (**Fig. 5c**). Moreover, we explored the gene-expression patterns of the kallikrein genomic locus in our 203 unselected mature mTECs and the 24 *Klk5*$^{+}$ mature mTECs selected by quantitative PCR. We found that *Klk5* expression served as a proxy for the expression of neighboring genes (**Fig. 5d** and **Supplementary Fig. 9**). These results showed that the expression of TRA-encoding genes in mTECs involved co-expressed groups of genes located in close proximity in the genome.

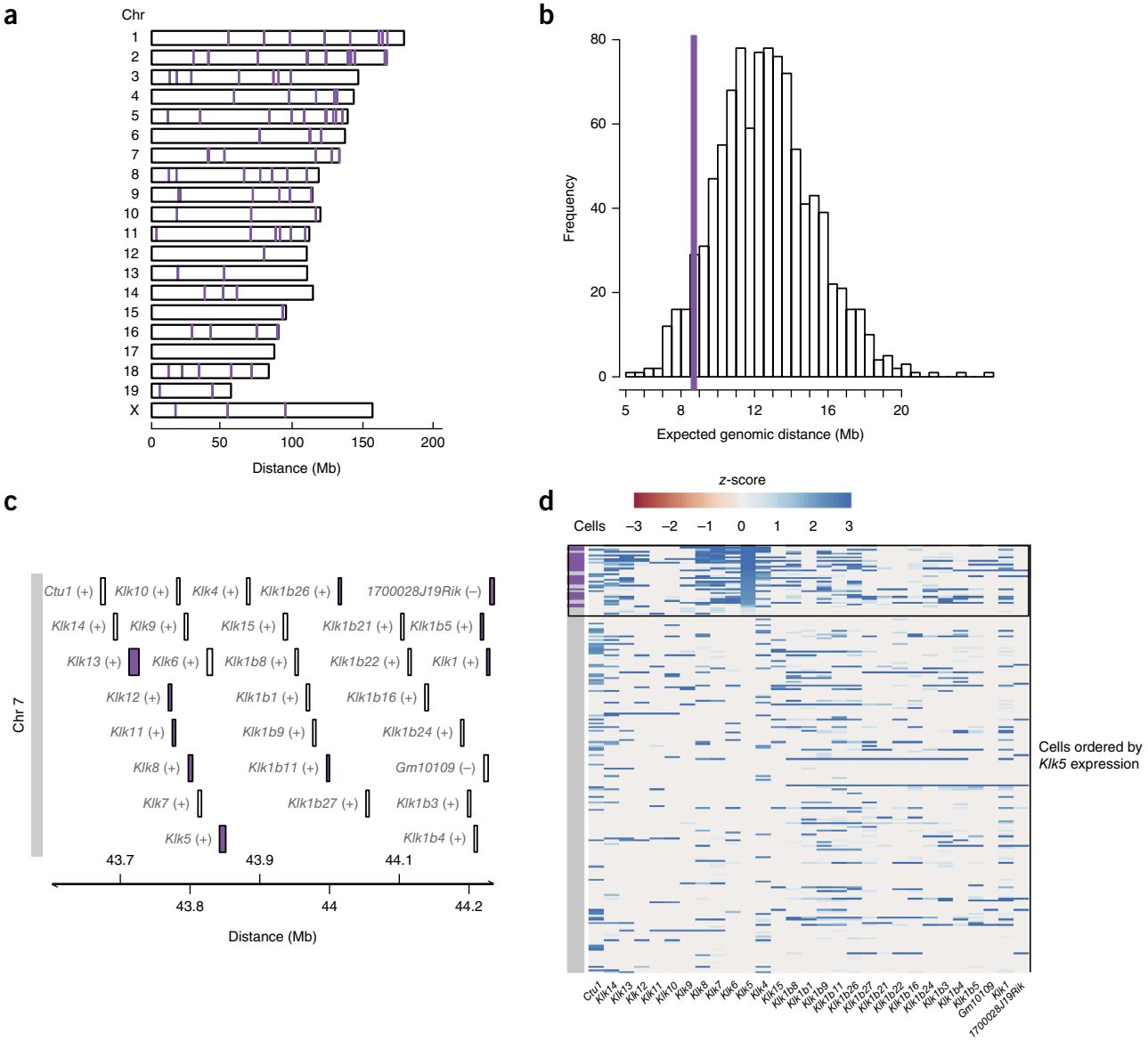

Figure 5 Co-expressed genes cluster in the genome. (**a**) Karyogram of the genomic localization of genes in co-expressed cluster D (**Fig. 2**). Chr, chromosome; Mb, megabases. (**b**) Distribution of expected median genomic distance between two genes in the genome (based on 1×10^3 permutations selecting random sets of genes of the same set size as gene set D); purple vertical line indicates median distance observed for the 115 co-expressed genes belonging to cluster D, which deviates from the null model (FDR = 10%). (**c**) Genomic region on chromosome 7 hosting genes encoding peptidases of the kallikrein (Klk) family; purple indicates genes assigned to cluster D (**2a**); (+), plus strand; (–), minus strand. (**d**) Expression profiles for genes encoding kallikrein peptidases (ordered by genomic position as in **c**) in single unselected mature mTECs ($n = 203$) and *Klk5*$^{+}$ mature mTECs ($n = 24$) selected by quantitative PCR (left margin (purple)), presented by decreasing Klk5 expression (top (highest) to bottom (lowest)); black box indicates mTECs for which *Klk5* transcripts were detected by scRNA-seq. Data are representative of 203 experiments (**a**–**c**) or 227 experiments (**d**) with one cell in each.

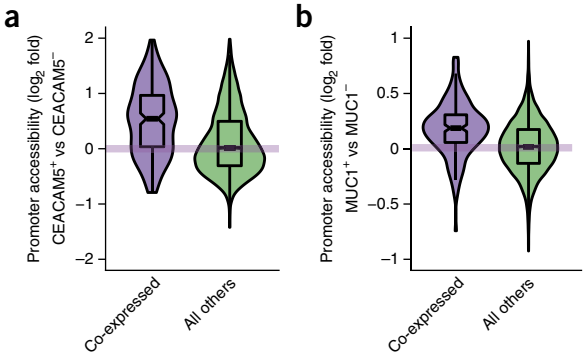

Figure 6 Promoters of co-regulated genes show increased chromatin accessibility. (**a**) Chromatin accessibility for the *CEACAM5*–co-expressed gene set (288 genes)[12] and all other protein-coding genes in CEACAM5$^+$ mTECs versus CEACAM5$^-$ mTECs ($n = 3$ donors), assayed by bulk ATAC-seq and presented as moderated logarithmic 'fold' changes calculated by the DESeq2 method[44]. $P = 1.2 \times 10^{-15}$ (*t*-test). (**b**) Chromatin accessibility for the *MUC1*–co-expressed gene set (219 genes) in MUC1$^+$ mTECs versus MUC1$^-$ mTECs, presented as in **a**. $P = 1.1 \times 10^{-14}$ (*t*-test). Data are representative of three experiments with one donor in each.

Promoters of co-expressed genes map to accessible chromatin

To directly assess the chromatin state of co-expressed genes, we assayed genome-wide DNA accessibility by the ATAC-seq method of epigenomic profiling[31], which is based on the 'preference' of the transposase TN5 to integrate into un-compacted chromatin and thus allows direct measurement of chromatin accessibility. To obtain a sufficient number of surface TRA–specific mTECs required for this assay, we used human thymic tissue and sorted cells on the basis of two published human co-expressed gene sets: the *CEACAM5* and *MUC1* gene sets[12]. We performed the ATAC-seq experiments with mTECs from the respective surface TRA–positive and TRA-negative mTEC fractions. When we accounted for all protein-coding genes, there was no difference between the TRA-positive mTECs and TRA-negative mTECs in their chromatin accessibility (**Fig. 6**). However, we observed that loci that were co-expressed with the respective TRA-positive subsets (either CEACAM5 or MUC1) were significantly more accessible in the TRA-positive mTECs than in the TRA-negative mTECs (**Fig. 6**). Thus, gene co-expression in distinct mTEC subsets accompanied enhanced chromatin accessibility at the promoter regions of the respective loci.

DISCUSSION

TRA expression in mTECs is essential for the induction of self-tolerance. However, its molecular regulation remains poorly understood. One open question relates to the regulation of TRA expression in single mTECs; i.e., to what extent the process is random or follows rules. Here, we applied scRNA-seq[22,23,32–36] and obtained evidence of numerous recurring co-expression patterns in mature mTECs. These patterns generally occurred at low cell frequencies. Co-expressed genes clustered in the genome, and their promoters displayed enhanced chromatin accessibility. Co-expressed gene sets formed mosaic patterns that faithfully 'added up' at the population level to present a comprehensive set of TRAs.

Mosaic gene-expression patterns in the thymus have been reported[10–12], and they allow a considerable diversity of antigens to be presented at the population level while limiting the number of TRA-encoding genes expressed in individual mTECs. As mTECs have a limited capacity for antigen presentation, restricting the number of ectopically expressed genes per cell seems to be crucial to ensuring epitope presentation at sufficient density to transmit a tolerogenic signal to maturing T cells.

It has been proposed that mosaic expression patterns arise by random induction of TRA-encoding genes in single mTECs[10,19,21]; this model has been challenged by the discovery that subsets of human mTECs selected by flow cytometry for the expression of particular TRAs display differential gene-expression patterns[12]. However, the preselected mTEC subsets analyzed previously represent only a narrow subset of the mTEC population, because those studies were constrained by the availability of antibodies suitable for flow cytometry. The data we have provided here substantially advance those findings, because the single-cell approach we used here addressed the issue of co-expression in a genome-wide unbiased way (i.e., no pre-selection required). The current depth of analysis allowed us to identify 11 previously unknown co-expression patterns within the mature mTEC population. As the number of mature mTECs we sequenced was limited (203 cells), we expect this number to be an underestimate.

Nevertheless, even this relatively small number of mTECs covered 95% of the reported TRA-encoding genes. Given the size of the mouse mTEC compartment ($\sim 1 \times 10^5$ cells)[10], this finding indicates that the complete TRA repertoire would be covered multiple times within the thymic medulla, even with allowance for a generous error margin in our calculations. Hence, T cells would only have to scan sub-domains of this compartment for efficient induction of self-tolerance.

Moreover, by 'zooming in' on the co-expression groups identified, we observed a positive correlation between *Tspan8* transcript levels and increased expression of genes co-expressed with *Tspan8* in both Ceacam1$^+$ cells and Tspan8$^+$ cells. This finding would be in line with the transitioning of individual cells between different co-expression groups, a concept that has been proposed in a model that postulates that individual mTECs transit between different TRA co-expression patterns and thus might express a sizeable portion of the TRA repertoire during their lifetime[12]. Such a mechanism could further reduce the minimal number of mTECs any single T cell would need to interact with to encounter the full TRA repertoire, because a given mTEC could express different TRAs when re-encountering the same T cell during its sojourn in the medulla[37].

We were able to assign 71% of the TRA-encoding genes to a co-expressed gene set on the basis of 203 single mature mTECs. The remaining TRA-encoding genes either escaped detection of co-expression due to the limited sample size or represent some features of random sampling. In addition, the extent to which mono-allelic expression versus bi-allelic expression, slippage of promoter usage resulting in truncated mRNA isoforms, and variable splicing patterns serve a role is unclear[6,21,38,39]. Those last features might extend the diversity of thymic presentation of self antigens; at the same time, they might represent pitfalls of thymic TRA expression that potentially undermine the process of tolerance induction and might lead to autoimmunity[38,39].

Our single-cell data showed that co-expressed genes tended to cluster in the genome. In conjunction with our ATAC-seq experiments, this suggests a potential mechanism for the generation of intra- and inter-chromosomal co-expression patterns. Such a mechanism would rely on local chromatin remodeling that allows neighboring genes to be co-expressed in a coordinated fashion in single mTECs, regardless of their distinct tissue-specific regulation in the periphery. Although the definition of TRAs is operational and is highly dependent on the thresholds used, our observation that co-expressed gene sets also contain genes that did not encode TRAs might indicate that TRA expression also promotes the expression of other genes adjacent to TRA-encoding genes. However, co-expressed gene sets

showed enrichment for TRA-encoding genes, which would suggest that the mechanism underlying co-expression patterns in mTECs targets mainly genes whose expression in the periphery of the body is restricted to a small number of tissues.

Chromatin remodeling can affect nearby genes on the same chromosome but also genes nearby in the three-dimensional architecture of the nucleus. Correlation between gene co-expression and co-localization in 'transcription factories' has been described for lineage-specific gene regulation[40], and this might also be the case for thymic TRA expression[12]. The finding that co-expressed gene clusters contained genes encoding products of unrelated biological function further supports our proposition that genomic positions influences thymic TRA expression.

Epigenetic signatures specifying such 'accessible' chromatin stretches in mTECs have not yet been investigated genome wide. However, a study focusing on the casein locus in mouse mTECs has shown that ectopic expression of the gene encoding casein-β correlates with marks of active transcription[41]. Thus, future studies should identify the molecular pathways that target co-expressed gene clusters and, moreover, should define the transcriptional regulators that promote transcription. In this context, spatially localized activation of gene expression by epigenetic remodeling, as proposed here for TRA expression in mTECs, has been reported for embryonic stem cells[42] and cancer cells[43].

Why mTEC-mediated tolerance induction, which presumably evolved in early vertebrates, uses coordinated co-expression patterns in single cells remains an open question. If cells were to coordinate their expression programs with each other (for example, to avoid expressing the same genes and thus ensure maximal coverage), then co-expression groups might provide a more economic means than a fully independent, cell-autonomous 'choice' of every single gene.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. ArrayExpress: sequencing data, E-MTAB-3346 and E-MTAB-3624.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank K. Hexel and S. Schmitt for single-cell sorting; S. Egle for technical help; C. Sebening and T. Loukanov (University of Heidelberg) for human thymic tissue; The Genomics Core Facility of the European Molecular Biology Laboratory for initial sequencing, and M. Miranda and E. Hopmans for support during subsequent sequencing at the Stanford Genome Technology Center; J. Buenrostro and C. Chabbert for discussions about ATAC-seq experiments and data, respectively; C. Michel and S. Anders for advice and comments on the manuscript; W. Wei and M. Sikora for help with data transfer; and The Central Animal Facility (German Cancer Research Center) for animal care. Supported by the European Union 7th Framework Programme (Health) via Project Radiant (W.H. and A.R.), The Helmholtz Center (K.R.), the Sonderforschungsbereich (DFG 938 to S.P.), the European Research Council (ERC-2012-AdG to B.K.) and the US National Institutes of Health (P01 HG000205 and R01 GM068717 to P.B., M.N. and L.M.S.).

AUTHOR CONTRIBUTIONS

P.B., S.P., B.K. and L.M.S. conceived of the project; P.B., S.P. and K.R. designed experiments; P.B. performed single-cell sequencing experiments, *Klk5* single-cell quantitative PCR confirmation experiments and ATAC-seq experiments; S.P. helped with the ATAC-seq experiments; S.P. and K.R. performed experimental mTEC preparations and flow cytometry of single and bulk mTECs; A.R. and W.H. designed analysis strategy and analyzed the data; A.R. prepared the figures; P.B., A.R., S.P., K.R., W.H., B.K. and L.M.S. interpreted the data and wrote the manuscript; M.N. and R.K. provided technical assistance; and L.M.S., B.K. and W.H. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Anderson, M.S. *et al.* Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298**, 1395–1401 (2002).
2. DeVoss, J.J. & Anderson, M.S. Lessons on immune tolerance from the monogenic disease APS1. *Curr. Opin. Genet. Dev.* **17**, 193–200 (2007).
3. Hogquist, K.A., Baldwin, T.A. & Jameson, S.C. Central tolerance: learning self-control in the thymus. *Nat. Rev. Immunol.* **5**, 772–782 (2005).
4. Klein, L., Kyewski, B., Allen, P.M. & Hogquist, K.A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
5. Derbinski, J., Schulte, A., Kyewski, B. & Klein, L. Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self. *Nat. Immunol.* **2**, 1032–1039 (2001).
6. Kyewski, B. & Klein, L. A central role for central tolerance. *Annu. Rev. Immunol.* **24**, 571–606 (2006).
7. Perry, J.S. *et al.* Distinct contributions of Aire and antigen-presenting-cell subsets to the generation of self-tolerance in the thymus. *Immunity* **41**, 414–426 (2014).
8. Yang, S., Fujikado, N., Kolodin, D., Benoist, C. & Mathis, D. Regulatory T cells generated early in life play a distinct role in maintaining self-tolerance. *Science* **348**, 589–594 (2015).
9. Malchow, S. *et al.* Aire-dependent thymic development of tumor-associated regulatory T cells. *Science* **339**, 1219–1224 (2013).
10. Derbinski, J., Pinto, S., Rosch, S., Hexel, K. & Kyewski, B. Promiscuous gene expression patterns in single medullary thymic epithelial cells argue for a stochastic mechanism. *Proc. Natl. Acad. Sci. USA* **105**, 657–662 (2008).
11. Cloosen, S. *et al.* Expression of tumor-associated differentiation antigens, MUC1 glycoforms and CEA, in human thymic epithelial cells: implications for self-tolerance and tumor therapy. *Cancer Res.* **67**, 3919–3926 (2007).
12. Pinto, S. *et al.* Overlapping gene coexpression patterns in human medullary thymic epithelial cells generate self-antigen diversity. *Proc. Natl. Acad. Sci. USA* **110**, E3497–E3505 (2013).
13. Mathis, D. & Benoist, C. Aire. *Annu. Rev. Immunol.* **27**, 287–312 (2009).
14. Abramson, J., Giraud, M., Benoist, C. & Mathis, D. Aire's partners in the molecular control of immunological tolerance. *Cell* **140**, 123–135 (2010).
15. Derbinski, J. *et al.* Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. *J. Exp. Med.* **202**, 33–45 (2005).
16. Koh, A.S. *et al.* Aire employs a histone-binding module to mediate immunological tolerance, linking chromatin regulation with organ-specific autoimmunity. *Proc. Natl. Acad. Sci. USA* **105**, 15878–15883 (2008).
17. Org, T. *et al.* The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression. *EMBO Rep.* **9**, 370–376 (2008).
18. Waterfield, M. *et al.* The transcriptional regulator Aire coopts the repressive ATF7ip-MBD1 complex for the induction of immunotolerance. *Nat. Immunol.* **15**, 258–265 (2014).
19. Sansom, S.N. *et al.* Population and single cell genomics reveal the Aire-dependency, relief from Polycomb silencing and distribution of self-antigen expression in thymic epithelia. *Genome Res.* **24**, 1918–1931 (2014).
20. Giraud, M. *et al.* Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. *Proc. Natl. Acad. Sci. USA* **109**, 535–540 (2012).
21. Villaseñor, J., Besse, W., Benoist, C. & Mathis, D. Ectopic expression of peripheral-tissue antigens in the thymic epithelium: probabilistic, monoallelic, misinitiated. *Proc. Natl. Acad. Sci. USA* **105**, 15854–15859 (2008).
22. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
23. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
24. St-Pierre, C., Brochu, S., Vanegas, J.R., Dumont-Lagace, M., Lemieux, S. & Perreault, C. Transcriptome sequencing of neonatal thymic epithelial cells. *Sci. Rep.* **3**, 1860 (2013).
25. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
26. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
27. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
28. Ohnishi, Y. *et al.* Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16**, 27–37 (2014).
29. Gotter, J., Brors, B., Hergenhahn, M. & Kyewski, B. Medullary epithelial cells of the human thymus express a highly diverse selection of tissue-specific genes colocalized in chromosomal clusters. *J. Exp. Med.* **199**, 155–166 (2004).
30. Johnnidis, J.B. *et al.* Chromosomal clustering of genes controlled by the aire transcription factor. *Proc. Natl. Acad. Sci. USA* **102**, 7233–7238 (2005).
31. 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).

32. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
33. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat. Protoc.* **5**, 516–535 (2010).
34. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
35. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
36. Islam, S. *et al.* Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828 (2012).
37. Le Borgne, M. *et al.* The impact of negative selection on thymocyte migration in the medulla. *Nat. Immunol.* **10**, 823–830 (2009).
38. Pinto, S. *et al.* Misinitiation of intrathymic MART-1 transcription and biased TCR usage explain the high frequency of MART-1-specific T cells. *Eur. J. Immunol.* **44**, 2811–2821 (2014).
39. Klein, L., Klugmann, M., Nave, K.A., Tuohy, V.K. & Kyewski, B. Shaping of the autoreactive T-cell repertoire by a splice variant of self protein expressed in thymic epithelial cells. *Nat. Med.* **6**, 56–61 (2000).
40. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61 (2010).
41. Tykocinski, L.O. *et al.* Epigenetic regulation of promiscuous gene expression in thymic medullary epithelial cells. *Proc. Natl. Acad. Sci. USA* **107**, 19426–19431 (2010).
42. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol.* **8**, 532–538 (2006).
43. Bert, S.A. *et al.* Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* **23**, 9–22 (2013).
44. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

ONLINE METHODS

Mice. C57BL/6 mice were used in this study for the isolation of mTECs. All breeding and cohort maintenance was performed in the central animal laboratory of the German Cancer Research Center (Deutsches Krebsforschungszentrum) under approved conditions in accordance with the European Convention for the Protection of Vertebrate Animals used for Experimental and other Scientific Purposes and the German Legislation.

Isolation of mouse medullary thymic epithelial cells. Mouse mTECs were isolated and purified as described[45] with pooling of cells 5–20 mice per experiment. The pre-enriched stromal cell fraction, sorted for unselected mature mTECs (n = 211 cells), was stained with the following antibodies: peridinin chlorophyll protein (PerCP)–anti-CD45 (30-F11; BD Pharmingen), Alexa Fluor 647–anti-EpCAM (G8.8; prepared in-house)[46], phycoerythrin (PE)–anti-I-A^b (16-10A1; BD Biosciences) and fluorescein isothiocyanate (FITC)–anti-Ly51 (6C3; BD Biosciences).

For the selection of mTECs by expression of the surface TRAs Tspan8 (n = 48 cells) or Ceacam1 (n = 30 cells), the following antibodies were used in the antibody mixture: peridinin chlorophyll protein (PerCP)–anti-CD45 (30-F11; BD Pharmingen), Alexa Fluor 647–anti-EpCAM (G8.8; prepared in-house)[46], FITC–anti-I-A^b (AF6-120.1; BD Pharmingen) and Pacific Blue-anti-CDR1 (CDR1 hybridoma; prepared in-house)[47], and either PE–anti-Tspan8 (657909; R&D Systems) or PE–anti-CD66a (anti-Ceacam1; CC1; eBioscience). Dead cells were excluded through the use of propidium iodide at a final concentration of 0.2 µg/ml. Cells were sorted on BD FACSAria III cell sorter (BD Biosciences) by the single-cell sorting mode as described[10]. Single mature mTECs used in all the experiments represent cells from pooled thymic tissue.

Single-cell RNA-seq. Single-cell sequencing libraries were prepared as reported[22,23] with the following modifications: 1 µl of a 1:1,000,000 dilution of ERCC Spike-In Mix (Life Technologies) in RNase-free water was included in a total volume of 5 µl lysis buffer. During analysis, sequencing reads mapping to ERCC 'spike-ins' were used for estimation of technical 'noise' levels and for 'calling' of significantly highly variable genes by a published method[25]. We used 19 cycles of initial PCR amplification and used a ratio of 0.6:1.0 (beads/total PCR volume; instead of 1.0:1.0) of Ampure XP beads (Beckman Coulter) for the first PCR purification to minimize primer dimer carryover. After the first PCR amplification, cDNA libraries were screened via quantitative PCR (we used a 1:10 dilution of purified cDNA libraries for quantitative PCR) for expression of a mouse housekeeping gene (*Ubc*), and the distribution of library size was checked on a Bioanalyzer instrument (Agilent) as reported[22,23]. Only cDNA libraries that passed both quality controls were processed further. We used 100 pg of cDNA for the 'tagmentation' (transposase-based fragmentation) reaction and applied 12 cycles for the final enrichment PCR. The final purification step was performed with a ratio of 0.8:1.0 (as above) of Ampure SPRIselect beads (Beckman Coulter). We 'multiplexed' 24 samples per Illumina HiSeq 2500 lane and used 105–base pair paired-end sequencing. A HiSeq sequencing lane typically yielded between ~150×10^6 and ~200×10^6 reads.

ATAC-seq. Human thymic tissue was obtained from children in the course of corrective cardiac surgery at the Department of Cardiac Surgery, Medical School of the University of Heidelberg. Studies of human samples were approved by the Institutional Review Board of the University of Heidelberg (367/2002), and informed consent was obtained from all patients. Human mTEC subsets ($MHCII^{hi}$ cells positive for surface TRAs and $MHCII^{hi}$ cells negative for surface TRAs) were isolated and sorted by flow cytometry as described[12]. ATAC-seq experiments were performed as reported[31] with the following modifications: 5×10^3 to 50×10^3 pooled cells (depending on mTEC subset frequency) were sorted in flow cytometry buffer (PBS containing 5% FCS) and were used for ATAC-seq experiments. We used 50% of each purified 'tagmentation' reaction for enrichment PCR (without five cycles of pre-amplification). Each enrichment PCR was monitored individually with the StepOnePlus Real-Time PCR System (Life Technologies), and the amplification reaction was stopped as soon as amplification approached saturation. After the enrichment PCR and subsequent purification of PCR products, we performed gel extraction (QIA MinElute Gel Extraction Kit; Qiagen) for removal of primer dimers. The final 'multiplexed' sequencing libraries were quantified by quantitative PCR and were sequenced on a HiSeq 2500 machine (Illumina). 105–base pair paired-end sequencing was used, and samples yielded between 16,867,055 and 40,820,441 sequenced fragments.

Confirmation of the *Klk5*–co-expressed gene set by quantitative PCR. Single-cell cDNA libraries of mature mTECs were prepared as described above. Libraries were purified after 19 cycles of PCR amplification with a ratio of 0.6:1.0 (as above) of Ampure XP beads (Beckman Coulter). Dilutions of 1:10 (in nuclease-free water) of the cDNA libraries were used for subsequent quantitative PCR pre-screening. Primers were designed with the NCBI Primer-BLAST tool. Single-cell cDNA libraries that were positive for expression of both *Klk5* and the housekeeping gene *Ubc* were processed further for Illumina sequencing. Since we used the 24-sample Illumina dual indexing kit, only 24 of the 28 *Klk5*-positive cells (instead of the 28 identified) were subjected to Illumina sequencing.

Bioinformatics. For the single-cell data, we mapped the sequenced read fragments (with the GSNAP nucleotide-alignment program, version 2014-07-04) to the mouse reference genome (ENSEMBL release 75). Only uniquely mapped sequenced fragments were considered for further analysis. For each single-cell transcriptome, we tabulated the number of sequenced fragments that overlapped with each gene through the use of the HTSeq package for data processing, and normalized for sequencing depth by a published method[48]. To account for technical variation, we used a published method[25] to identify genes whose biological coefficients of variation were larger than 50%, and we used this subset for further analysis. We used another published method[27] to 'regress out' the variation on the data explained by the cell cycle. We identified groups of co-regulated genes by the 'partitioning around medoids' (pam) method of the R package 'cluster' (software of the R project for statistical computing) and assessed their stability with the R package 'clue'. To identify genes co-expressed with TRA-encoding genes, we used the Wilcoxon test. Multiple testing corrections were done using the Benjamini-Hochberg method. The ATAC-seq data were mapped to the human reference genome (ENSEMBL release 75) with GSNAP version 2014-07-04.

Code availability. We have provide a comprehensive and reproducible workflow containing the documented R code used for the analysis of all the data, including the generation of all reported figures and summary statistics, in the **Supplementary Code**.

45. Rattay, K. *et al.* Homeodomain-interacting protein kinase 2, a novel autoimmune regulator interaction partner, modulates promiscuous gene expression in medullary thymic epithelial cells. *J. Immunol.* **194**, 921–928 (2015).
46. Farr, A., Nelson, A., Truex, J. & Hosier, S. Epithelial heterogeneity in the murine thymus: a cell surface glycoprotein expressed by subcapsular and medullary epithelium. *J. Histochem. Cytochem.* **39**, 645–653 (1991).
47. Rouse, R.V., Bolin, L.M., Bender, J.R. & Kyewski, B.A. Monoclonal antibodies reactive with subsets of mouse and human thymic epithelial cells. *J. Histochem. Cytochem.* **36**, 1511–1517 (1988).
48. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

doi:10.1038/ni.3246