nature immunology

Aire controls gene expression in the thymic epithelium with ordered stochasticity

Matthew Meredith, David Zemmour, Diane Mathis & Christophe Benoist

The transcription factor Aire controls immunological tolerance by inducing the ectopic thymic expression of many tissue-specific genes, acting broadly by removing stops on the transcriptional machinery. To better understand Aire's specificity, we performed single-cell RNA-seq and DNA-methylation analysis of *Aire*-sufficient and *Aire*-deficient medullary epithelial cells (mTECs). Each of Aire's target genes was induced in only a minority of mTECs, independently of DNA-methylation patterns, as small inter-chromosomal gene clusters activated in concert in a proportion of mTECs. These microclusters differed between individual mice. Thus, our results suggest an organization of the DNA or of the epigenome that results from stochastic determinism but is 'bookmarked' and stable through mTEC divisions, which ensures more effective presentation of self antigens and favors diversity of self-tolerance between individuals.

Aire is a fascinating transcription factor, with a unique function in promoting immunological tolerance of differentiating thymocytes[1]. First, it induces ectopic expression in medullary epithelial cells (mTECs) of a large set of genes whose products are typically associated with fully differentiated parenchymal cells (so-called 'peripheral tissue antigens' (PTAs))[2]. In addition, Aire controls the expression of factors that modulate the presentation of peptides derived from these PTAs by major histocompatibility complex (MHC) molecules at the mTEC surface, or their cross-presentation by dendritic cells[3]. These peptides mold the T cell repertoire by inducing negative selection of self-reactive specificities[4,5] or by promoting positive selection of regulatory T cells[6]. The physiological consequences of Aire's activity are profound, as humans and mice with loss-of-function mutations in loci encoding Aire develop multi-organ autoimmunity[1].

Even if its structural domains are shared with conventional motif-specific transcription factors, Aire is a very unusual transcription factor. It affects a large number of genes and generally allows the transcription of genes that would not be expected to be expressed in a given cell-type. Aire contains a SAND domain typically involved in DNA binding, but it does not have a distinct DNA-binding motif, although it has been suggested to recognize methylated CpG residues in association with the methylated CpG–binding factor MBD1 (ref. 7). Instead, its transcriptional activity seems to depend on the recognition of nonspecific markers of chromatin with low activity, such as the hypomethylated amino-terminal tail of histone H3 (refs. 8,9) or transcriptional start sites (TSSs) with a surfeit of paused polymerases[10]. Aire also interacts with a variety of non-specific elements of the transcriptional and splicing machinery[11,12]. Indeed, data derived from a variety of experimental approaches indicate that Aire's main *modus operandi* is to release pausing of promoter-proximal RNA polymerase II (refs. 10,13,14).

Aire's action has an element of stochasticity. Single-cell PCR analysis suggests that individual mTECs, otherwise indistinguishable, express distinct patterns of PTAs[15–18]. Gene-expression profiling of mTECs from individual mice also suggests that inter-individual 'noise' in gene expression between genetically identical mice is higher for genes that are targets of Aire than for the bulk of transcripts[19]. Despite such clues, a coherent framework that explains Aire's action in individual cells has remained elusive.

Single-cell transcriptome sequencing (scRNA-seq) has opened completely new vistas on the analysis of gene expression[20] by combining the global quality of genome-wide transcriptome profiling with the unique granularity brought by single-cell technologies such as flow cytometry. It can reveal unrecognized subpopulation structure and avoid erroneous averaging and can provide information on the fluctuations ('noise') in gene expression[21,22] in an otherwise homogeneous population of cells[23–25]. Some of this noise can result from transcriptional bursting[26], but it may also reveal coordinated activation of specific transcriptional programs that can be important in determining cellular differentiation or responses. Technical innovations have made scRNA-seq more performant and robust, with cell 'multiplexing', molecular 'barcoding' and microfluidic devices[27]. The analysis of scRNA-seq data remains challenging, however. First, with molecular-conversion efficacies of 20% at best, the portion of the transcriptome with low expression is unreliably assessed in any one cell. Second, because real replicates are innately impossible to assess by single-cell analysis, estimation of technical variance remains uncertain. Finally, the data must be interpreted in the context of sampling statistics, which makes analysis less intuitive than conventional profiling data and necessitates complex statistical models[24,25,28].

scRNA-seq seemed to provide a good opportunity to explore the distribution of PTA expression in individual mTECs. This perspective,

Division of Immunology, Department of Microbiology and Immunobiology, Harvard Medical School, and Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts, USA. Correspondence should be addressed to D.M. (cbdm@hms.harvard.edu) or C.B. (cbdm@hms.harvard.edu).

Received 13 February; accepted 9 July; published online 3 August 2015; doi:10.1038/ni.3247

much broader than achieved earlier by PCR[15,17], allowed us to investigate how frequently individual genes that are targets of Aire are expressed in mTECs and whether Aire changes the frequency of cells expressing particular transcripts or instead boosts the intensity of transcript expression in cells in which they are already present. Although Aire-induced gene expression proved to be very noisy, affecting genes with a low frequency of expression, we identified unexpected order in this chaos, detecting a large number of Aire-induced transcripts whose expression clustered in small groups of mTECs, with no apparent logic, and varied between individual mice. Our observations have direct implications for the efficiency of tolerance induction and individual susceptibility to autoimmune deviation.

RESULTS

Range of Aire-induced gene expression

As a prelude to single-cell analysis, we performed standard RNA-seq analysis of bulk-sorted mTECs. We prepared CD45$^-$ Ly51^{lo}MHCII^{hi}GFP^{hi} cells from mice that express green fluorescent protein (GFP) driven on a bacterial artificial chromosome transgene encoding Aire[29], which we crossed with mice carrying the *Aire*-knockout mutation[2] to generate *Aire*-sufficient (called 'wild-type' here) and *Aire*-deficient littermates. In the libraries generated (11.8×10^6 to 31.3×10^6 mapped reads per sample), we observed a biased and very deep effect of Aire: of the 19,772 genes expressed (at a threshold of 1 FKPM (fragments per kilobase of exon per million mapped reads) in these data sets, 2,995 were 'Aire-induced' genes and 766 'Aire-repressed' genes (at an arbitrary change in expression of over twofold) (**Fig. 1a**). These results were consistent with published microarray analyses[19] and subsequent RNA-seq data[30] showing that Aire regulates a large fraction of the transcriptome. The sets of Aire-induced genes and 'Aire-neutral' genes (neither induced nor repressed by Aire) defined here were those tracked in the scRNA-seq analyses below.

In addition to the consequences on entire transcripts reported above, our RNA-seq data showed that Aire further exerted more subtle effects on the use of individual exons within genes. Several transcripts whose overall levels were affected little in mTECs by the absence of Aire showed Aire-dependent inclusion of particular exons (**Fig. 1b**). A more complete analysis of this phenomenon revealed 3,219 such exons with Aire-dependent expression, in contrast to the majority of exons whose representation correlated with that of the gene as a whole (175,216 exons; **Fig. 1c**). Alternative splicing has long been known to affect tolerance, as initially recognized for autoimmune responses to the product of the PTA-encoding gene *Plp1* (refs. 31,32). As has also been speculated before[33], Aire may help maximize exposure to genome-encoded peptides by enhancing exon inclusion, a property consistent with the splicing factors with which it interacts[11]. Conversely, our analysis also revealed the presence of a set of exons whose abundance remained invariant in the presence or absence of Aire, while the whole transcript was induced (**Fig. 1c**). These exons were particularly prevalent at the beginning of the transcripts (**Fig. 1d**), consistent with the demonstration that the representation of the first exon shows comparatively little change in the absence of Aire[10], reflective of polymerases that transcribe a short portion of the gene before stalling in Aire's absence. The match between the degree of induction of genes and exons by Aire increased progressively along the transcript (**Fig. 1d**), which suggested that this effect might actually extend quite some distance from the TSS.

Overall diversity in transcriptomes of individual mTECs

With the reference data above in hand, we proceeded with analyzing Aire-controlled gene expression in individual mTECs through scRNA-seq. We used index sorting of single cells into wells of microtiter plates, such that we could relate the RNA-seq profiles to the marker characteristics of the cells (**Fig. 2a**). We generated sequencing libraries from 360 single mTECs from two pairs of wild-type and *Aire*-deficient mice using a protocol modified from the original CEL-Seq technique[34]. This protocol includes oligo(dT) priming with 'barcodes' to allow attribution of each sequence read to its cell of origin, as well as unique molecular identifiers for tagging of each original molecule to avoid artifacts from over-amplification of small numbers of initial molecules[35]. Although most single-cell libraries yielded high-quality data, for robustness we restricted our further analysis to 201 cells that generated at least 1×10^4 unique mappable reads per cell (**Fig. 2b**). Several findings confirmed our single-cell data. First, there

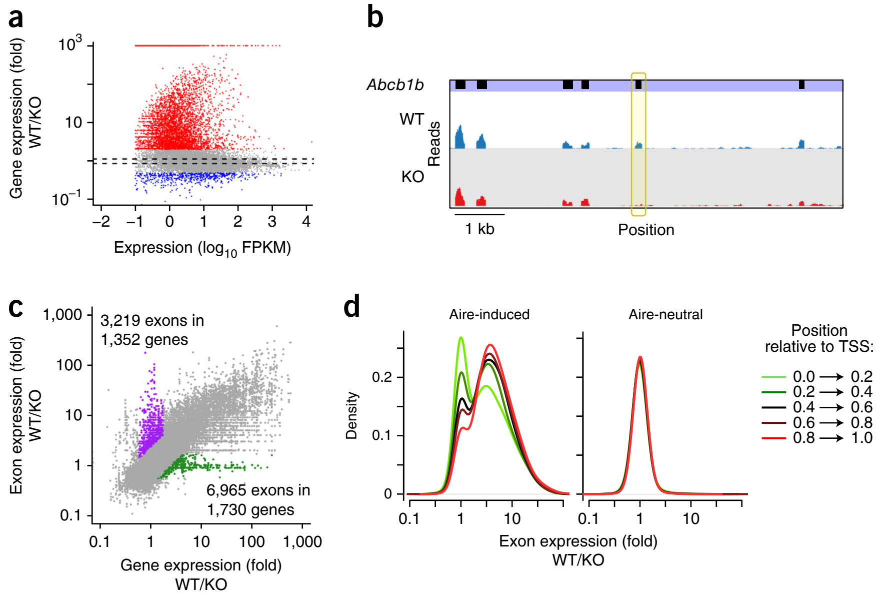

Figure 1 Aire increases the repertoire and diversity of mTEC transcriptome. (**a**) Read counts (as FPKM) versus change in expression (wild-type/Aire-deficient (WT/KO)) of all genes in RNA-seq libraries generated from whole-mTEC RNA of Aire-deficient mice and their wild-type littermates, showing genes upregulated (red; Aire-induced genes) or downregulated (blue; Aire-repressed genes) by twofold or more, or with a change in expression of less than 1.1-fold (between dashed lines; Aire-neutral genes). (**b**) Read 'pileups' (peaks) in exons (black boxes, top) in the Aire-neutral gene *Abcb1b* (defined as in **a**) in wild-type and *Aire*-deficient samples, identifying an Aire-induced (differentially spliced) exon (yellow). (**c**) Change in expression (as in **a**) of exons versus genes for mTEC RNA-seq data from **a**, showing exons upregulated more than twofold at the gene level but less than 1.1-fold at the exon level (green) or vice versa (purple). (**d**) Distribution of change in expression (as in **a**) of exons in Aire-induced genes (left) and Aire-neutral genes (right), presented according to relative position within the gene (key); 'density' (vertical axis) indicates the number of genes. Data are representative of two experiments with results pooled from two mice per group (**a**; mean; biological duplicates) or two experiments (**b**–**d**).

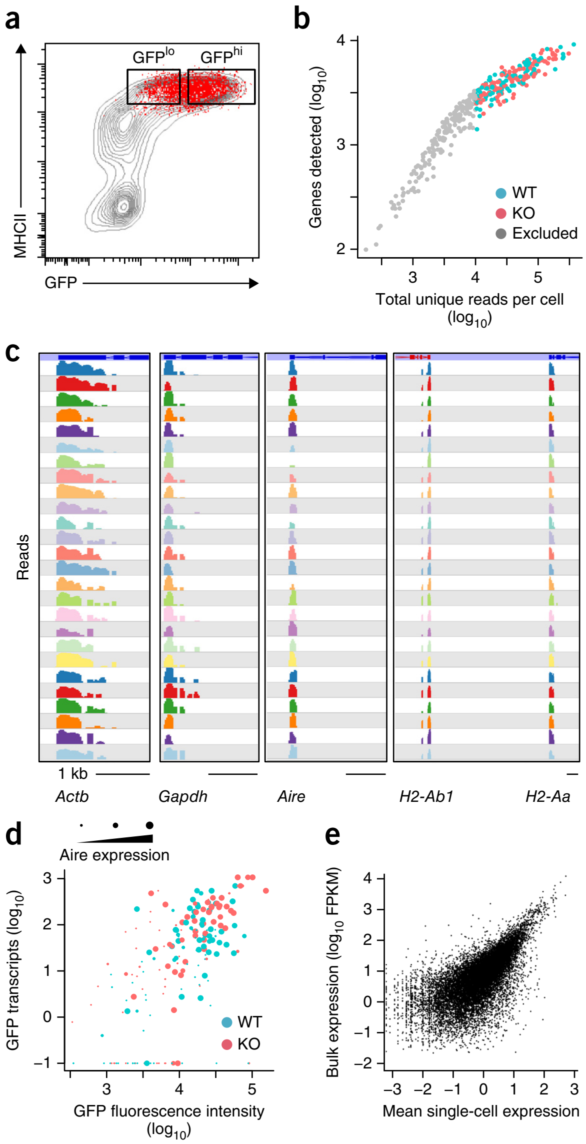

Figure 2 scRNA-seq analysis of mTECs. (**a**) Sorting of single mTECs (red) from wild-type mice for scRNA-seq analysis. (**b**) Quantification of unique mappable reads versus genes detected for each wild-type cell (WT) or *Aire*-deficient cell (KO) in the scRNA-seq data sets; gray (Excluded) indicates cells omitted from further analysis. (**c**) Read 'pileups' for five illustrative genes (columns) in the scRNA-seq data sets), in 26 representative cells (rows); only the carboxy-terminal exon is detected because the scRNA-seq technique tags only sequences adjacent to poly(A). Top, mRNA: blue (sense) or red (antisense). (**d**) Correlation between GFP fluorescence intensity during sorting (as in **a**) versus GFP mRNA reads observed in each cell (Pearson $r = 0.56$): dot size indicates the number of reads from the *Aire* transcript (wedge). (**e**) Mean single-cell read counts per gene versus bulk read counts of those same genes, in wild-type cells (Pearson $r = 0.72$). Data are pooled from two independent experiments with two mice per genotype (**a**) or two independent experiments (**b**–**e**).

was good representation of the transcripts encoding MHC class II and housekeeping transcripts expected in mTECs (**Fig. 2c**). Second, the intensity of the Aire-GFP fluorescence, as detected by flow cytometry, matched the counts of transcripts encoding GFP and Aire in each cell (**Fig. 2d**; the *Aire*-deficient mutation abolishes function but not the transcript). Third, the total number of reads per gene, obtained by aggregation of all the scRNA-seq data sets, recapitulated well the data at the population level (**Fig. 2e**).

We then analyzed computationally the expression of Aire targets in these scRNA-seq data. We determined presence or absence of individual transcripts in each cell (**Fig. 3**); this revealed each of the following points, which we substantiated and confirmed (**Figs. 4**–**6**). First, Aire targeted mainly transcripts expressed at a low frequency (Aire-induced transcripts were more sparse than Aire-neutral transcripts). Second, Aire increased that frequency, more in wild-type cells than in *Aire*-deficient cells. Third, discrete clusters of Aire-induced genes showed coordinated expression. Fourth, in accordance with that point, there were groups of mTECs with comparable expression of small gene clusters. Fifth, mTEC clusters were different in individual mice.

Aire targets mainly transcripts expressed at low frequency

We generated density plots of the frequency of mTECs expressing individual genes for Aire-induced, Aire-neutral or Aire-repressed mRNAs (transcripts matched for expression in bulk RNA-seq). This analysis showed that most Aire-induced genes were active in only 5–20% of the *Aire*-deficient mTECs sampled (**Fig. 4a**). Aire-neutral and Aire-repressed genes were more frequently expressed in these mTECs, and the difference was significant across the three expression levels (**Fig. 4a**). Two points indicated that this low frequency of Aire-induced transcripts was not merely a consequence of statistical sampling, which can be a concern for scRNA-seq. First, the frequency of false-negative results ('dropouts') from sampling is directly related to the intensity of expression, and these dropouts would be expected at the same frequency for expression-matched Aire-induced or Aire-neutral transcripts, which was not the case here (**Fig. 4a**). Second, we plotted the probability that cells with no reads for a given gene were statistical dropouts. Most Aire-induced transcripts had a very low probability of being a false-negative result (68.1%, with a nominal *P* value of <0.05; **Fig. 4b**).

Aire increases mainly the intensity of target-gene expression

The higher expression observed in wild-type mTECs than in *Aire*-deficient mTECs for a given Aire-induced gene in bulk population profiling might have resulted from an increase either in the amount of transcript per cell or in the proportion of cells expressing the transcript. When we compared the changes in mean expression intensity in wild-type mTECs versus *Aire*-deficient mTECs positive for a given transcript, as well as the change in the frequency of cells expressing this transcript, we found that Aire expression seemed to increase both (**Fig. 4c**). Curiously, we also observed a limited but significant shift between *Aire*-deficient mTECs and wild-type mTECs for transcripts in the Aire-neutral category (**Fig. 4c**), which indicated that Aire subtly activated the majority of genes in the cell. A potential confounder of this analysis is that a higher read number per cell leads to more frequent detection of positive cells simply because higher intensities favor lower dropout rates. To test for such bias, we plotted the frequency of cells expressing Aire-induced or Aire-neutral genes versus the mean intensity of expression in mTECs that did express those genes, in wild-type and *Aire*-deficient mTECs. This analysis showed that the presence of Aire resulted in a predominant shift in the distribution toward higher per-cell intensities, a shift that did not merely follow the main intensity-frequency relationship (**Fig. 4d**). Indeed, we found that the shift in expression intensity in Aire's presence led to less increase in the expression frequency of its targets than predicted from the dropout distribution of gene pairs randomly drawn from the Aire-neutral distribution (**Supplementary Fig. 1**). Thus, genes that are targets of Aire remained less frequently expressed than the genome-wide norm, even after transcriptional activation by Aire.

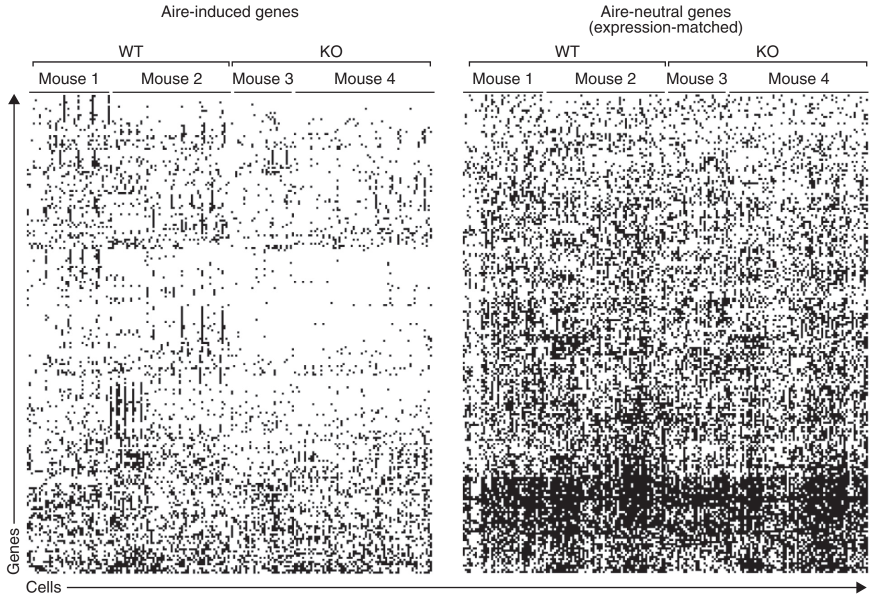

Figure 3 Summary of single-cell expression results. Presence (black) or absence (white) of individual transcripts in wild-type and *Aire*-deficient mTECs, for Aire-induced transcripts (left) and a set of Aire-neutral genes matched for mean expression in transcript-positive cells (right) from the scRNA-seq data (**Fig. 2**): genes are arranged in rows by hierarchical clustering; cells are arranged in columns according to genotype and mouse. The weighted probability of expression of each transcript in each cell was calculated by a published Bayesian approach[28]; black squares indicate transcript presence regardless of intensity, and white squares indicate no expression (most at high confidence of not being dropouts by analysis with the SCDE ('single-cell differential expression') software package, as in **Fig. 4**). Data are pooled from two independent experiments.

Coordinate expression of discrete clusters of Aire-induced genes

The results presented above (short vertical streaks, **Fig. 3**) suggested that subsets of genes were expressed in concert. To better investigate such structures in the data, we calculated gene-by-gene correlations for all Aire-induced genes (on the basis of a weighted expression matrix[24]) and performed a partition clustering using an affinity-propagation algorithm[36]. We found a high degree of structure in the scRNA-seq data sets from wild-type mTECs, as 51% of Aire-induced transcripts grouped into 19 clusters with an internal mean correlation of >0.75 (**Fig. 5a**); these clusters were small (33–114 transcripts; median, 57) and were largely distinct from each other. We verified the significance of these clusters by permutation (randomly

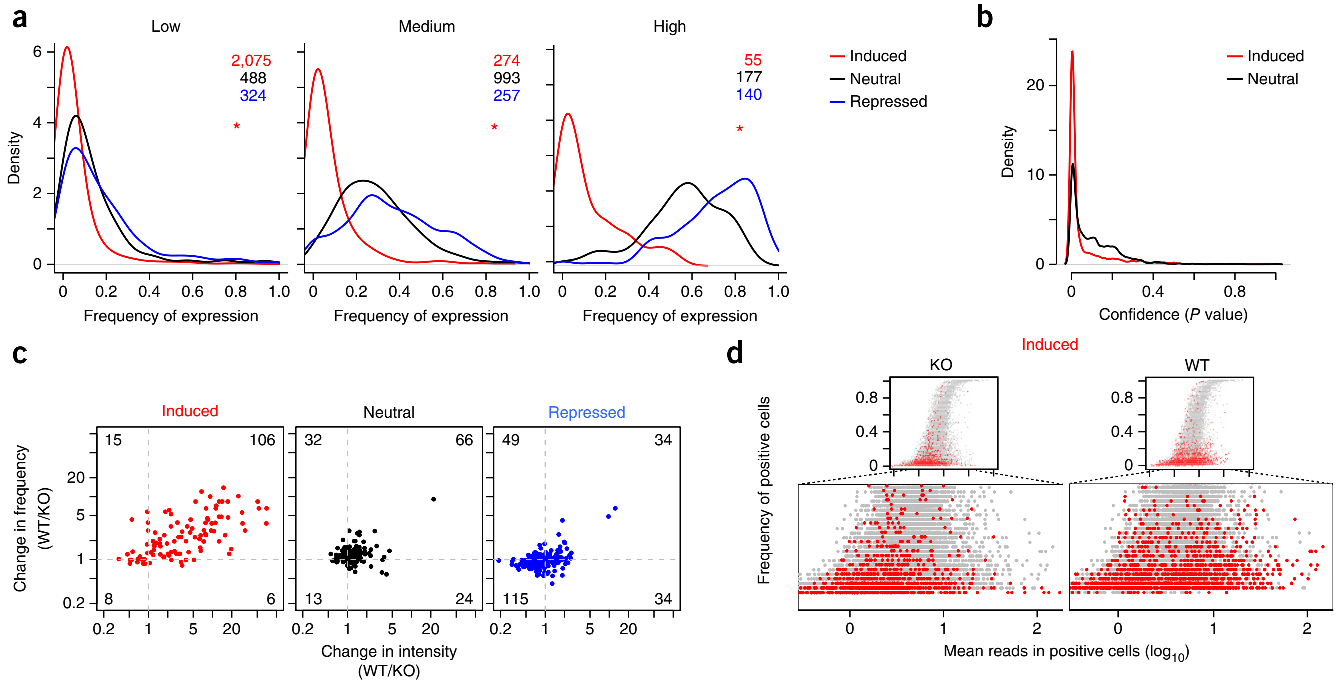

Figure 4 Aire increases the intensity and frequency of otherwise rare transcripts. (**a**) Distribution of the frequency of expression in single cells (scRNA-seq data of **Fig. 2**) from Aire-deficient mice ($n = 2$) for gene sets matched by expression in bulk RNA-seq data as low (1–5 FPKM), medium (10–25 FPKM) or high (50–100 FPKM). $*P < 10^{-15}$ (Wilcoxon test). (**b**) Bayesian confidence[28] (P value) that genes considered unexpressed a given cell were not sampling dropouts, for expression-matched Aire-induced and Aire-neutral genes (identified as in **Fig. 1**) in Aire-deficient mice ($n = 2$); 'density' (vertical axis) indicates the number of genes. (**c**) Change in expression intensity (mean count per gene in cells expressing the gene; wild-type/*Aire*-deficient) versus change in the frequency of expression (frequency of expressing cells; wild-type/*Aire*-deficient), calculated for expression-matched transcripts (window of 25–50 counts per gene) of Aire-induced, Aire-neutral or Aire-repressed genes (identified as in **Fig. 1**) in cells from wild-type and Aire-deficient mice ($n = 2$ per genotype). $P < 10^{-15}$, wild-type versus *Aire*-deficient, and $P < 10^{-7}$, Aire-neutral genes (χ^2 test). (**d**) Mean counts in transcript-positive cells versus frequency of expression for Aire-induced genes (red) relative to genome-wide distribution (gray), in cells from wild-type and Aire-deficient mice ($n = 2$ per genotype); below, expansion of results above to focus on the shift in Aire-induced genes (identified as in **Fig. 1**). Data are pooled from two independent experiments.

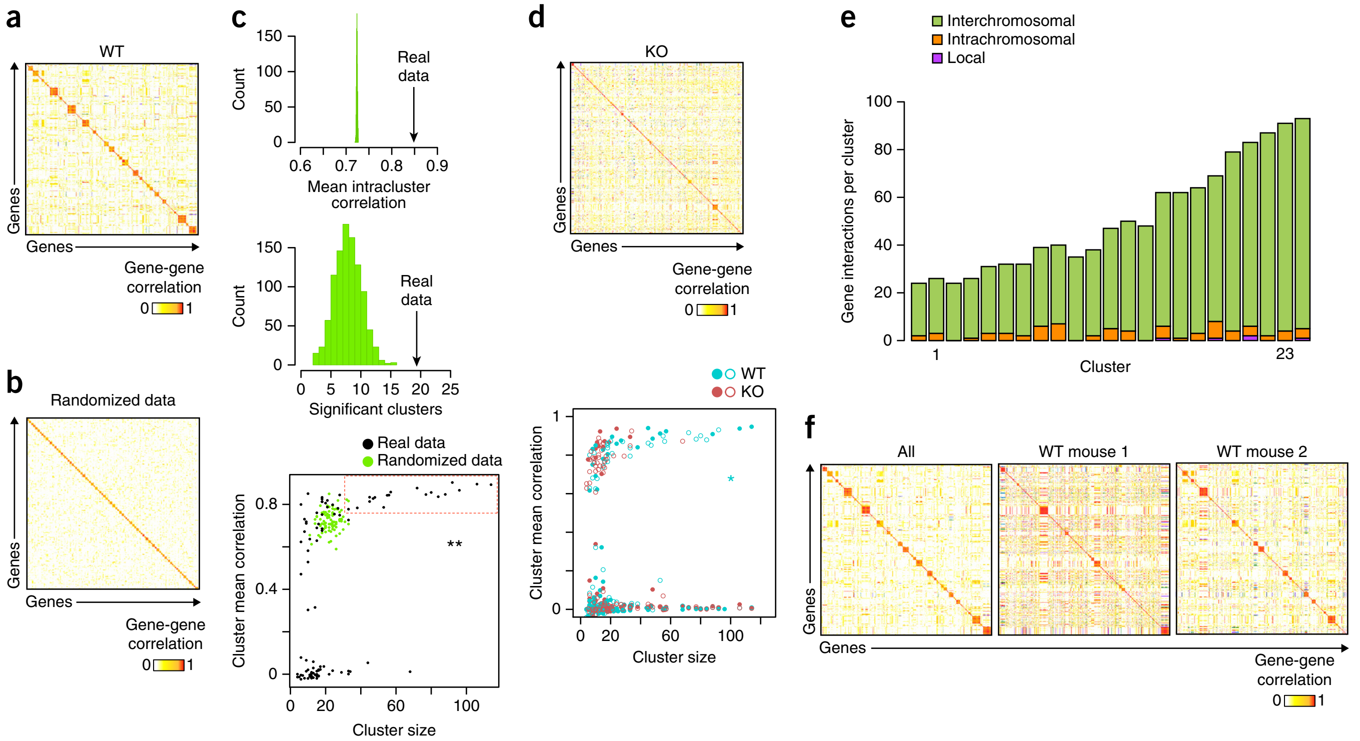

Figure 5 Aire coordinates discrete interchromosomal gene networks. (**a**) Gene-by-gene Pearson correlations calculated from the weighted expression matrix for Aire-induced genes (identified as in **Fig. 1**) in all mTECs from wild-type mice ($n = 2$); genes are ordered according to affinity-propagation clustering[36], with no preset number of clusters. (**b**) Correlations as in **a**, but in a control data matrix generated by random permutation of gene-expression values. (**c**) Results of 1,000 random permutations and affinity-propagation clustering of the scRNA-seq data for wild-type cells, presented as mean within-cluster correlation and quantification of significant correlations in each iteration (with significant clusters defined as clusters with more than 30 genes and a mean correlation of >0.75). (**d**) Correlations as in **a**, but for Aire-deficient mTECs (pooled from $n = 2$ mice) (left), and size and internal correlation of clusters in wild-type and Aire-deficient mTECs (right). (**e**) Quantification of significant gene-gene correlations defined as local (distance of less than 1 Mb on the same chromosome), intrachromosomal (distance of more than 1 Mb on same chromosome) or interchromosomal (different chromosomes) in the 23 largest clusters of the wild-type scRNA-seq data sets in **a**. (**f**) Gene-gene correlations between Aire-induced transcripts (identified as in **Fig. 1**) calculated as in **a** for all wild-type mTECs (left), and calculated independently in mTECs from each wild-type mouse (middle and right). $^*P = 0.002$ and $^{**}P = 0.001$ (Wilcoxon test). Data are pooled from two independent experiments (**a**–**c**,**e**,**f**) or are from one experiment with 1,000 permutations (**d**).

shuffling expression levels per gene between cells), which did not reproduce the same degree of cluster structure (**Fig. 5b**); comparable cluster sizes and internal correlations were not achieved in 1,000 random permutations (**Fig. 5c**). We detected fewer such clusters with the scRNA-seq data sets from Aire-deficient mTECs (**Fig. 5d**) or when we calculated results from expression-matched Aire-neutral genes ($P = 8 \times 10^{-5}$ (Wilcoxon test); **Supplementary Fig. 2**), which further substantiated the significance of these results and indicated that Aire was required for the appearance of these co-expressed clusters.

It has been reported that Aire-induced PTA-encoding genes tend to partition into local gene clusters[37,38]. We observed such co-regulated activity of local genomic segments, as illustrated for the *Sprr* and *Mup* loci (**Supplementary Fig. 3**). However, these localized co-regulation events contributed little to the overall gene clusters, most of which 'rested' on correlations across chromosomes (**Fig. 5e**). Therefore, mTECs co-expressed networks of discrete, interchromosomal genes. We searched for commonalities between the transcripts that would form these small clusters. Gene-ontology or pathway analysis (by the Molecular Signatures Database or the PANTHER ('protein analysis through evolutionary relationships') classification system) failed to reveal any function or pathway shared by products encoded by genes in any of these clusters; nor did cluster members share specificity of expression when analyzed across the GNF (Genomics Institute of the Novartis Research Foundation) compendium of gene expression[39]; nor did the promoter regions of cluster members show enrichment for binding motifs for a particular transcription factor (data not shown). Therefore, these co-expressed gene clusters seemed to be unrelated in terms of genomic position, biological function or transcriptional regulation.

Aire-induced gene networks define distinct mTEC subgroups

Given the small clusters of Aire-induced genes identified above, we sought to determine how their expression demarcated individual mTECs. Correlation analysis based on probability values for Aire-induced genes showed that wild-type mTECs partitioned into discrete groups (**Fig. 6**). These groupings were based on inter-chromosomal gene networks, because the cell-to-cell correlation maps calculated on the basis of transcripts from one chromosome were reproduced, for the most part, when calculated with transcripts from other chromosomes (**Fig. 6a**, right). For a broader perspective on mTEC heterogeneity in wild-type and *Aire*-deficient thymi, we analyzed the cell-to-cell correlation matrix with t-SNE[40], a dimensionality-reduction algorithm that computes the probability that two cells are similar and displays the best fit in two-dimensional space. *Aire*-deficient mTECs tended to group close together at the center of the t-SNE plot (**Fig. 6b**). Wild-type mTECs, although distributed around the same center, radiated further ($P < 10^{-3}$ (Wilcoxon test)) and were more distant from each other ($P < 10^{-60}$ (Wilcoxon test)) than were

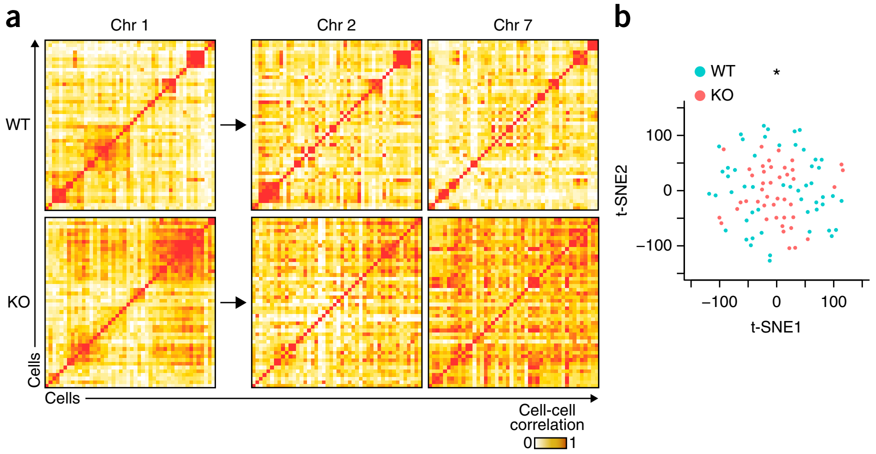

Figure 6 Aire-dependent interchromosomal gene networks generate diverse and distinct mTEC subsets. (**a**) Cell-by-cell Pearson correlation between individual mTECs (scRNA-seq data of **Fig. 2**) among mTECs from wild-type mice (top) and Aire-deficient mice (bottom) ($n = 2$ per genotype per group), calculated for Aire-induced genes on different chromosomes. Clustering was determined by affinity propagation for genes on chromosome (Chr) 1 (left), and the same cell order was applied for correlation values calculated with Aire-induced genes from chromosome 2 or 7 (right). (**b**) Distribution of mTECs (calculated by t-SNE) from wild-type and *Aire*-deficient mice ($n = 2$ per genotype), based on the expression of Aire-induced genes (identified as in **Fig. 1**). $*P < 10^{-3}$ (Wilcoxon test). Data are pooled from two independent experiments.

Aire-deficient mTECs (**Fig. 6b**). Predictably, these small t-SNE groups coincided with the cell clusters identified above (**Fig. 6a**). Thus, Aire diversified gene expression, not in a completely random fashion but with some degree of coordination between cells.

Difference in mTEC clusters in individual mice

It was already apparent that the small gene clusters expressed in mTECs were not shared by different mice (**Fig. 3**). Indeed, when we calculated gene-gene correlations independently from scRNA-seq mTEC data sets from each wild-type mouse, correlations within a cluster applied for mTECs of only one mouse, but not those of the other mouse (**Fig. 5f**). Thus, these gene networks were most probably established by stochastic events and not by 'hardwired' molecular cues.

DNA methylation in mTECs does not account for Aire specificity

Epigenetic regulatory mechanisms are likely candidates for two prominent characteristics of Aire transcriptional specificity: the 'predilection' to activate infrequently expressed genes, and the interchromosomal clusters that were coordinately expressed in small groups of mTECs. The methylation of DNA at CpG dinucleotides is one such candidate, as variable but heritable methylation patterns could be at the root of this. Indeed, it has been proposed that Aire associates with the methylated CpG–binding protein MBD1 and uses this factor's ability to 'preferentially' recognize methylation at the TCGCA motif for 'preferential' targeting of PTA-encoding genes[7]. In addition, analysis of DNA methylation by reduced representation bisulfite sequencing (RRBS) is inherently a single-cell methodology that measures the frequency of DNA-methylation marks at specific locations and was thus a good complement to the single-cell analyses reported above.

To determine their DNA-methylation status, we sorted mTECs as described above and processed their DNA for RRBS[41]. The distribution of CpG methylation at various positions did not differ markedly between mTECs from wild-type mice and those from *Aire*-deficient mice, for Aire-induced and Aire-neutral loci (**Fig. 7a**). CpG positions in upstream enhancer elements were similarly represented in both sets of genes, and the region surrounding the TSS was uniformly unmethylated in both cases (**Fig. 7a**). This observation held for MBD1 sites in particular, which were uniformly unmethylated in all loci (**Supplementary Fig. 4**). In fact, the frequency of TCGCA sites in Aire-target genes was the same in Aire-induced TSSs and Aire-neutral TSSs, and reanalysis of published expression data[7] showed a limited

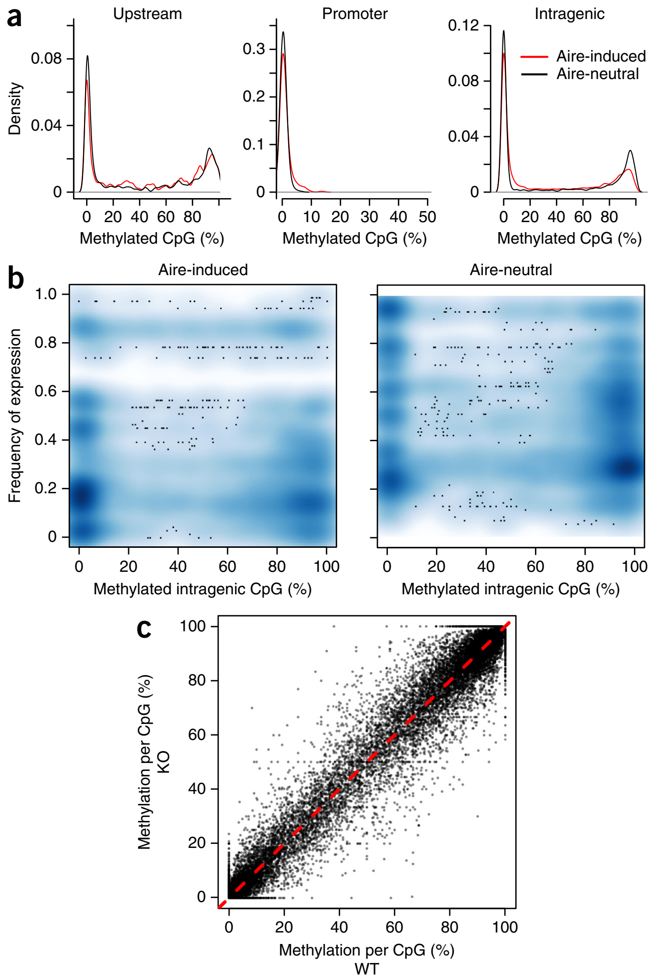

Figure 7 Little or no difference in the amount of CpG methylation at Aire-induced genes versus that at Aire-neutral genes. (**a**) Frequency of methylated CpG residues ('density') in upstream regions (–1 to –50 kilobases (kb) from the TSS), promoter regions (–100 to –1 kb from the TSS) or intragenic regions (25% or more of gene length beyond the TSS) of Aire-induced or Aire-neutral genes (key), assessed in RRBS methylation libraries from mTECs of Aire-deficient mice (pooled from $n = 2$ mice). (**b**) Relationship between mean methylation frequency at intragenic CpGs and frequency of expression for Aire-induced genes, in Aire-deficient mTECs. (**c**) Methylation frequency at each CpG position in DNA from wild-type mTECs versus Aire-deficient mTECs (data from **a**). Data are pooled from two independent experiments.

transcriptional effect of MBD1 on mTECs, which overlapped very little with Aire's transcriptional signature (**Supplementary Fig. 4c**). This indicated that MBD1 had little or no role in the Aire-dependent expression of PTA-encoding genes.

CpG methylation increases in frequency at intragenic positions[42], and this trend was slightly less pronounced for Aire-induced loci than for expression-matched Aire-neutral loci (**Fig. 7a**). We sought to determine whether the frequency of intragenic CpG methylation might relate to the frequency of expression of corresponding Aire-induced genes in wild-type mTECs. The majority of intragenic CpGs were either not methylated or highly methylated, and both of these methylation statuses were associated with a range of expression frequencies (**Fig. 7b**). Finally these methylation profiles, including the sites of variable methylation, were not Aire dependent, as shown by the high degree of correlation between wild-type mTECs and Aire-deficient mTECs (**Fig. 7c**). Therefore, Aire itself did not alter the DNA-methylation profiles in mTECs, and methylation patterns did not provide any obvious clue as to the frequency distribution of genes that are targets of Aire.

DISCUSSION

Our study has revealed several novel aspects of Aire's function as a transcription factor, which probably have direct consequences on its function in the induction of central tolerance. First, Aire increased the diversity of the mTEC transcriptome by inducing thousands of Aire-dependent transcripts, as well as Aire-dependent exons in otherwise Aire-neutral genes. This observation was consistent with the prediction that ectopic PTA expression involves splicing patterns different from those in the tissues in which PTAs are 'normally' expressed[33]. Alternative splicing is known to have important consequences on autoimmune responses[31,32], but Aire's role in this process has not been recognized thus far, to our knowledge. It seems likely that Aire's effect on differential exon inclusion is tied to its close interactions with splicing factors of the transcriptional machinery and its 'preferential' effect on spliced transcripts in cultured cells[11,14]. This broad effect on the mTEC transcriptome maximizes the representation of peptides presented to developing thymocytes.

Second, Aire seemed to 'preferentially' target genes expressed in a minority of cells, and increased the intensity of expression of its target genes. This is consistent with the notion that Aire recognizes generic features of gene and chromatin organization, such as histone H3 with no methylation at Lys4 or promoters with a surfeit of paused polymerases[8,10,13,43]. Thus, Aire had no particular specificity for PTAs but instead 'keyed' on these general features of poorly expressed genes. However, in terms of the regulation of gene expression, what does an 'infrequently expressed' gene really mean, for a rather homogenous primary cell population like the Aire$^+$ mTECs analyzed here? It is a notion far removed from the deterministic gene-expression programs usually envisaged for lineage differentiation but is instead related to notions of 'noisy' gene expression. There are several sources of noise in gene expression that can be important in allowing progression through differentiation or cellular adaptation processes[44]. Infrequent gene expression can correspond to a 'burst' of transcription, whereby any given gene is actively transcribed only a small fraction of the time[26] and produces relatively short-lived transcripts. Then, a low frequency of cells positive for the expression of a specific gene can simply indicate the odds of 'catching' a cell during such a transcriptional burst. Alternatively, low-frequency expression can result from a particular organization of the DNA or epigenetic modifications, which are set stochastically in every cell but are then stable for some period of time. Aire is found mainly in tight nuclear speckles[45] thought to be sites of active transcription, and it is possible that the set of genes ectopically expressed by an mTEC are those that have been 'threaded' into these Aire-containing speckles.

Clues to the basis of the low expression frequency of Aire-controlled genes might be found in the small clusters of co-regulated genes whose expression was shared by discrete groups of mTECs. The existence of these discrete microclusters of expression is not easily compatible with a 'burst' model, because it is unlikely that genes would 'burst' at the same time in different cells; instead, the discrete microclusters of expression are compatible with a model in which infrequent expression results from stable organization of the genome or epigenome. Because the genes within these expression clusters did not share discernable sequence motifs or chromosomal locations that might explain their coordinated transcription, their co-expression in a fraction of the mTECs is perhaps most easily interpreted in terms of clonal relationship. mTECs that share PTA clusters could plausibly be daughters of the same epithelial cell progenitor[46], which would indicated that the selection of Aire targets within an mTEC clone is 'bookmarked' across cell division. 'Bookmarking' (the recovery of gene-expression programs after mitosis[47]) can be explained for conventional transcription by the persistence of networks of specific transcription factors, but it is puzzling for a mode of regulation that does not depend on the transcription factors that normally activate specific PTA-encoding genes[15]. Epigenetic cues such as DNA modifications, albeit probably not methylation, or remanent histone marking might be involved in 'bookmarking' PTA expression. Of note, Brd4, which binds acetylated histones and promotes the release of RNA polymerase II stalled at the promoter, is involved *in trans* in mitotic bookmarking[47] and is an essential Aire cofactor (H. Yoshida *et al.*, personal communication); it is possible that Brd4, together with Aire and other cofactors, forms trans-mitotically stable complexes with fixed DNA regions. Thus, one proposal might be that an inherently stochastic mechanism initially selects and marks groups of loci, whose co-expression is then bookmarked and transmissible. This scenario parallels the stochastically determined repertoire of activating and inhibitory receptors in natural killer cells.

In terms of tolerance induction, the low expression frequency of Aire-target genes and the existence of expression microclusters would indicate that mTECs are 'splitting the burden' of PTA expression and that there is a higher local concentration of any gene product than if all PTAs were uniformly expressed in all mTECs. Since immature thymocytes scan the thymic medulla and negative selection is effective with small pockets of antigen-positive presenting cells[48,49], negative selection should be more effective than with lower but widespread amounts of PTA expression in mTECs.

Notably, the co-expressed gene clusters were not the same in the two genetically identical wild-type mice whose mTECs we analyzed by scRNA-seq; this has important implications for the inter-individual variation in tolerance within a species. On the basis of microarray profiling data, Aire-induced transcripts have shown significantly greater inter-individual variability than have Aire-independent transcripts[19]. Our data have now provided a cellular explanation for this observation. One caveat of our study here, however, is that we cannot formally know if these co-regulated clusters persist and reflect constant inter-individual differences, or if they fluctuate and represent the state of the mTEC pool at the time of cell preparation for scRNA-seq. However, since the expressed clusters of Aire-induced genes were not shared at the time of the experiment, the two mice analyzed exposed their immature thymocytes to slightly different sets of self peptides and thereby generated T cell repertoires with slightly different autoreactivities to peripheral tissues. Such diversity may be favorable at the level of the species in ensuring a diversity of potential responses to pathogens without uniform 'holes' in the repertoire, albeit at the price of susceptibility to autoimmune diseases.

METHODS

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

Accession codes. SRA: RNA-seq and methylation data: SRR2038194, SRR2038195, SRR2038196, SRR2038197, SRR2038206, SRR2038210, SRR2038212 and SRR2038213; GEO: single-cell transcriptomic analysis, GSE70798.

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

ACKNOWLEDGMENTS

We thank S. Mostafavi for advice on computational analysis; M. Anderson (University of California, San Francisco) for the *Aire*-GFP line; and K. Hattori, G. Buruzula, and K. Waraska for help with mice, sorting and sequencing. Supported by the US National Institutes of Health (DK060027; and a Children's Hospital in Pediatric Gastroenterology training grant for M.M.) and Boehringer Ingelheim Fonds (D.Z.).

AUTHOR CONTRIBUTIONS

M.M., data collection, data analysis and manuscript writing; D.Z., data analysis and manuscript writing; D.M., manuscript writing; C.B. data analysis and manuscript writing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Peterson, P., Org, T. & Rebane, A. Transcriptional regulation by AIRE: molecular mechanisms of central tolerance. *Nat. Rev. Immunol.* **8**, 948–957 (2008).
2. Anderson, M.S. *et al.* Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298**, 1395–1401 (2002).
3. Hubert, F.X. *et al.* Aire regulates the transfer of antigen from mTECs to dendritic cells for induction of thymic tolerance. *Blood* **118**, 2462–2472 (2011).
4. Liston, A. *et al.* Aire regulates negative selection of organ-specific T cells. *Nat. Immunol.* **4**, 350–354 (2003).
5. Anderson, M.S. *et al.* The cellular mechanism of Aire control of T cell tolerance. *Immunity* **23**, 227–239 (2005).
6. Malchow, S. *et al.* Aire-dependent thymic development of tumor-associated regulatory T cells. *Science* **339**, 1219–1224 (2013).
7. 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).
8. 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).
9. 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).
10. 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).
11. Abramson, J., Giraud, M., Benoist, C. & Mathis, D. Aire's partners in the molecular control of immunological tolerance. *Cell* **140**, 123–135 (2010).
12. Gaetani, M. *et al.* AIRE-PHD fingers are structural hubs to maintain the integrity of chromatin-associated interactome. *Nucleic Acids Res.* **40**, 11756–11768 (2012).
13. Oven, I. *et al.* AIRE recruits P-TEFb for transcriptional elongation of target genes in medullary thymic epithelial cells. *Mol. Cell. Biol.* **27**, 8815–8823 (2007).
14. Giraud, M. *et al.* An RNAi screen for Aire cofactors reveals a role for Hnrnpl in polymerase release and Aire-activated ectopic transcription. *Proc. Natl. Acad. Sci. USA* **111**, 1491–1496 (2014).
15. 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).
16. Taubert, R., Schwendemann, J. & Kyewski, B. Highly variable expression of tissue-restricted self-antigens in human thymus: implications for self-tolerance and autoimmunity. *Eur. J. Immunol.* **37**, 838–848 (2007).
17. Derbinski, J. *et al.* Promiscuous gene expression patterns in single medullary thymic epithelial cells argue for a stochastic mechanism. *Proc. Natl. Acad. Sci. USA* **105**, 657–662 (2008).
18. 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).
19. Venanzi, E.S., Melamed, R., Mathis, D. & Benoist, C. The variable immunological self: genetic variation and nongenetic noise in Aire-regulated transcription. *Proc. Natl. Acad. Sci. USA* **105**, 15860–15865 (2008).
20. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
21. Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
22. Ozbudak, E.M. *et al.* Regulation of noise in the expression of a single gene. *Nat. Genet.* **31**, 69–73 (2002).
23. Kalmar, T. *et al.* Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol.* **7**, e1000149 (2009).
24. Shalek, A.K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).
25. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
26. Ross, I.L., Browne, C.M. & Hume, D.A. Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. *Immunol. Cell Biol.* **72**, 177–185 (1994).
27. Wu, A.R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
28. Kharchenko, P.V., Silberstein, L. & Scadden, D.T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
29. Gardner, J.M. *et al.* Deletional tolerance mediated by extrathymic Aire-expressing cells. *Science* **321**, 843–847 (2008).
30. 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).
31. Klein, L. *et al.* 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).
32. Anderson, A.C. *et al.* High frequency of autoreactive myelin proteolipid protein-specific T cells in the periphery of naive mice: mechanisms of selection of the self-reactive repertoire. *J. Exp. Med.* **191**, 761–770 (2000).
33. Keane, P., Ceredig, R. & Seoighe, C. Promiscuous mRNA splicing under the control of AIRE in medullary thymic epithelial cells. *Bioinformatics* **31**, 986–990 (2015).
34. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
35. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
36. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27**, 2463–2464 (2011).
37. Johnnidis, J.B. *et al.* Chromosomal clustering of genes controlled by the aire transcription factor. *Proc. Natl. Acad. Sci. USA* **102**, 7233–7238 (2005).
38. Derbinski, J. *et al.* Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. *J. Exp. Med.* **202**, 33–45 (2005).
39. Su, A.I. *et al.* Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. USA* **99**, 4465–4470 (2002).
40. van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
41. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **6**, 468–481 (2011).
42. Ball, M.P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009).
43. Org, T. *et al.* AIRE activated tissue specific genes have histone modifications associated with inactive chromatin. *Hum. Mol. Genet.* **18**, 4699–4710 (2009).
44. Raser, J.M. & O'Shea, E.K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
45. Tao, Y. *et al.* AIRE recruits multiple transcriptional components to specific genomic regions through tethering to nuclear matrix. *Mol. Immunol.* **43**, 335–345 (2006).
46. Gill, J., Malin, M., Hollander, G.A. & Boyd, R. Generation of a complete thymic microenvironment by MTS24^{+} thymic epithelial cells. *Nat. Immunol.* **3**, 635–642 (2002).
47. Zhao, R. *et al.* Gene bookmarking accelerates the kinetics of post-mitotic transcriptional re-activation. *Nat. Cell Biol.* **13**, 1295–1304 (2011).
48. Le Borgne, M. *et al.* The impact of negative selection on thymocyte migration in the medulla. *Nat. Immunol.* **10**, 823–830 (2009).
49. Merkenschlager, M., Benoist, C. & Mathis, D. Evidence for a single-niche model of positive selection. *Proc. Natl. Acad. Sci. USA* **91**, 11694–11698 (1994).

ONLINE METHODS

Mice. All mice were housed and bred under specific-pathogen–free conditions at the Harvard Medical School Center for Animal Resources and Comparative Medicine (Institutional Animal Care and Use Committee protocol 2954). Mice with *Aire*-driven expression of Igrp-GFP (sequence encoding GFP fused to the gene encoding islet-specific glucose-6-phosphate-related protein; *Adig* mice*)*[29] were provided by M. Anderson.

Isolation of thymic epithelial cells. Thymus tissue was dissociated in RPMI medium and was digested for 30 min, with agitation every 10 min, with 0.5 mg/ml collagenase-dispase (Roche) and 0.2 mg/ml DNase (Sigma) in RPMI medium. Following staining with primary antibodies (allophycocyanin-conjugated antibody to (anti-) MHC class II (I-A–I-E) (M5/114.15.2), phycoerythrin-conjugated anti-Ly51 (6C3) and phycoerythrin– indodicarbocyanine (Cy5)–conjugated anti-CD45 (30-F11); all from BioLegend), samples underwent were depletion of $CD45^+$ cells by magnetic-activated cell separation with anti-PE beads (Miltenyi). $DAPI^-CD45^-Ly51^{lo}MHCII^{hi}GFP^{hi}$ mTECs (5×10^4 to 10×10^4 per mouse) were sorted on a MoFlo (Cytomation) into Trizol for RNA preparation (for TruSeq library preparation) or into RPMI medium (Gibco) for RRBS library preparation. For scRNA-seq, similar gating, which also included GFP^{lo} mTECs, was used for sorting at a density of one cell per well of a 96-well plate on a FACSAria sorter (BD).

TruSeq library preparation and analysis. Bulk RNA-seq libraries were prepared with TruSeq following the manufacturer's protocol (Illumina) from 5×10^4 to 10×10^4 sorted mTECs (one mouse) per sample. Sequencing (single-end, 50bp) was performed on a HiSeq2000 (Illumina), and reads were aligned to the mm10 assembly of the mouse genome with the TopHat2 splice-junction mapper. Duplicated reads were filtered out from further analysis. Normalized counts per transcript (FPKM) were calculated with the Cufflinks suite of tools for RNA-Seq analyses. Exon expression was calculated with the SeqMonk program for visualization and analysis of mapped sequence data (Babraham Institute).

RRBS library preparation and analysis. RRBS libraries were prepared as described[41] from 5×10^4 to 10×10^4 sorted mTECs (one mouse), except that EZ DNA Methylation-Direct (Zymo) was used for bisulfite conversion. Sequencing (single-end, 50–base pair (bp)) was performed on a HiSeq2000. Prior to alignment, reads were trimmed to remove adaptor sequences with the RRBS option of the TrimGalore! automated wrapper script (Babraham Institute). Trimmed reads were aligned to mm10 with the Bismark mapping tool[50] (Babraham Institute), and methylation calls per CpG were calculated using SeqMonk (Babraham Institute). Only those CpG sites covered by at least 20 reads were considered for subsequent analysis. Relating of CpG positions to the closest genes was determined by SeqMonk relative to mm10.

scRNA-seq library construction. Single-cell RNA sequencing libraries were generated with a modified CEL-Seq protocol[34]. First, single cells were index-sorted with a BD FACSAria II in 96-well hard-shell PCR plates (HSP9631; BioRad) filled with 4.4 µl of lysis buffer containing 0.125 µl of RNAseOut (40 U per 1 µl stock; 10777-019; Invitrogen), 0.25 µl of reverse-transcription primer (25 ng per 1 µl stock) and 4 µl of RNAse-free water (AM9932; Ambion). Each primer contained a T7 promoter, the 5′ TruSeq Illumina adaptor, unique molecular 'barcodes' (4–9 bp), a single cell DNA barcode (8–16 bp) and oligo(dT) sequence (24 bp). Each of three wells was filled with a different mixture to process the carrier RNA and into which single cells were not sorted. These wells contained 0.5 µl HeLa total RNA (1 µg/µl; AM7852; Ambion), 0.25 µl of a T7-oligo(dT) primer without the Illumina adaptor or barcodes (initially provided in the MessageAmpTM II aRNA Amplification Kit; AM1751; Ambion), 0.25 µl of RNAseOut and 3.5 µl of RNAse-free water. After cell sorting, the plates were quickly covered with an aluminum seal (AlumaSeal 96; F96100; Excel Scientific), then were vortexed for 10 s, centrifuged for 1 min at maximum speed (>2250*g* at 4 °C), frozen on dry ice and kept at −80 °C for up to 3 weeks.

RNA was denaturated by incubation of the plates for 3 min at 70 °C (thermocycler lid temperature, 80 °C) in a thermocycler. 2 µl of First Strand Reverse Transcription mix was then added to each well (ArrayScript Reverse Transcriptase; AM2048; Ambion) containing 1 µl of dNTP mix (10 mM each stock; 18427-013; Invitrogen), 0.5 µl of 10× First Strand Buffer, 0.25 µl of ArrayScript (200 U per 1 µl of stock) and 0.25 µl of RNAse Inhibitor (40 U per 1 µl stock; AM2682; Ambion). The plates were then incubated for 2 h at 42 °C (thermocycler lid temperature, 50 °C).

Second-strand reverse transcription was performed with the mRNA Second strand synthesis module (E6111L; NEBNext). 15 µl of the Second Strand Reverse Transcription containing 12 µl of RNAse-free water, 2 µl of 10× Second Strand Synthesis Reaction Buffer and 1 µl of the Second Strand Synthesis Enzyme Mix was added to each well of the plates. The plates were then incubated for 2.5 h at 16 °C (with thermocycler lid open). cDNA clean-up and size selection were then performed using the Agencourt RNAClean XP beads (A63987; Beckman Coulter). First, single-cell cDNA libraries containing different barcodes were pooled into one tube with a Hela carrier cDNA library. In our experiments, 30 single-cell cDNA libraries were pooled together with a carrier HeLa cDNA library (three pools per plate). Each pool was mixed with 0.8× volume of Agencourt RNAClean XP beads, followed by incubation for 15 min at room temperature, then incubation for 5 min on the magnet. The supernatant was carefully removed and the beads were washed with fresh 70% ethanol twice while still on the magnet. Beads were dried for 15 min before elution was carried out in 50 µl of RNAse-free water. A second bead purification was performed similarly with 1× volume of RNA AMPure XP Beads and with elution in 6 µl of RNase-free water.

In vitro transcription was then conducted with a MEGAshortscript T7 transcription kit (AM1354; Ambion). 10.4 µl of a mix containing 1.6 µl of ATP (75 mM stock), 1.6 µl of UTP (75 mM stock), 1.6 µl of GTP (75 mM stock), 1.6 µl of CTP (75 mM stock), 1.6 µl of T7 10× Reaction Buffer, 1.6 µl of T7 Enzyme Mix and 0.8 µl of RNAseOut was added to each 6-µl cDNA pool, followed by incubation for 14 h at 37 °C (thermocycler lid temperature, 70 °C). Illumina libraries were then constructed as follows. First, the amplified RNA (aRNA) was fragmented with the Magnesium RNA Fragmentation Module (NEBNext E6150S). 4 µl of the fragmentation mix, containing 2 µl of RNAse-free water and 2 µl of the RNA Fragmentation Buffer (10×), was added to 16 µl of aRNA. Samples were immediately incubated for 2 min at 94 °C (thermocycler lid temperature, 105 °C), then were immediately transferred onto ice, and the reaction was stopped by the rapid addition of 2 µl of 10× RNA Fragmentation Stop Solution. The fragmented aRNA was then cleaned with an RNeasy MinElute Cleanup Kit (74204; Qiagen) according to the manufacturer's protocol, followed by elution, twice, in 10 µl of RNAse-free water.

The size distribution and quantity of fragmented aRNA was then assessed by analysis of 1 µl of each sample in a BioAnalyzer with an Agilent RNA 6000 pico Kit (5067-1513; Agilent). The samples were then treated as follows. First, the 5′ end of the aRNA was dephosphorylated by the addition of 4 µl of a mix containing 2 µl of 10× Antarctic Phosphatase Reaction Buffer, 1 µl of Antartic phosphatase (5U per 1 µl stock; M0289; NEB) and 1 µl of RNAseOut to 16 µl of each aRNA pool, followed incubation for 30 min at 37 °C and for 5 min at 65 °C. Then, the RNA was dephosphorylated at the 3′ end and phosphorylated at the 5′ end by the addition of 30 µl of a mixture containing 21.5 µl of RNAse-free water, 5 µl of 10× Antarctic Phosphatase Reaction Buffer (M0289; NEB), 0.5 µl of ATP (100 mM stock, ATP Tris buffered; R1441; Thermo Scientific), 1 µl of RNAseOut and 2 µl of T4 PolyNucleotide Kinase (10 U/µl; M0201S; NEB), followed by incubation for 1 h at 37 °C. RNA treated with phosphatase and T4 polynucleotide kinase was then purified with an RNeasy MinElute Cleanup Kit according to the manufacturer's protocol (Qiagen) and was eluted in 14 µl of RNAse-free water. The samples were then dried down to a volume of 5 µl with a vacuum concentrator (5–7 min at 55 °C). The 3′ Illumina adaptor (RA3) was then ligated to the treated RNA with T4 RNA Ligase 2, truncated (L6070L; Enzymatics). 3 µl of a mixture containing 1 µl of 10× truncated T4 RNA Ligase 2 buffer, 1 µl of DMSO (D9170; Sigma) and 1 µl of the 3′ adaptor (10 µM stock) was added to 5 µl of the treated RNA. The samples were incubated for 2 min at 70 °C (thermocycler lid temperature, 80 °C), then were placed immediately in ice, and 2 µl of a mixture containing 0.5 µl of RNAse Inhibitor (40 U/µl; Y9240L; Enzymatics) and 1.5 µl of truncated T4 RNA Ligase 2 (5 U per 1 µl of stock) was added. The ligation was performed for 1 h at 22 °C (open thermocycler lid). The ligated RNA was then reverse-transcribed with SuperScript II (18064-014; Invitrogen). 8.5 µl of a mixture

doi:10.1038/ni.3247

containing 2 µl of RNA reverse-transcription primer (RTP primer; Illumina); 10 µM stock) and 6.5 µl of RNAse-free water was added to 10 µl of the ligated RNA. The samples were then incubated for 2 min at 70 °C (thermocycler lid temperature, 80 °C) and then placed immediately on ice, and 10.5 µl of a mixture containing 4 µl of 5× First Strand Buffer, 0.5 µl of dNTP (25 mM mix), 2 µl of DTT (100 mM stock), 2 µl of RNAseOut and 2 µl of SuperScript II (200 U per 1 µl of stock) was added.

Reverse transcription was performed for 1 h at 50 °C (thermocycler lid temperature, 70 °C), and the library was then amplified with Kapa HotStart ReadyMix (KK2602; Kapa). 71 µl of the following mixture was added to each reverse-transcription reaction: 17 µl RNAse-free water, 50 µl of Kapa 2× HotStart ReadyMix and 4 µl of P5_Rd1_Primer_F (10 µM stock). To each reaction 4 µl of a uniquely indexed P7_Rd2_Primer_idxN_R (10 µM stock) was added, and PCR cycles were performed as follows: 95 °C for 3 min; 18 cycles of 20 s at 98 °C, 30 s at 60 °C and 30 s at 72 °C; and 5 min at 72 °C. The PCR product was then cleaned up and selected by size with two rounds of treatment with Agencourt AMPure XP Beads (A63880), as described above, with the following modifications. The first purification used 1× volume of beads and elution was performed in 32 µl of water; the second purification used 1.2× volume and 12 µl of elution water.

The size distribution and quantity of the library was assessed by analysis of 1 µl of each sample in a BioAnalyzer with a Agilent High Sensitivity DNA Kit (5067-4626; Agilent). Samples were pooled for sequencing on a MiSeq ('nano kits') and HiSeq 2500 (rapid mode). Paired-end 50-bp sequencing was performed with custom primers (100 µM stock in water) as follows: 75 bp for Read 1 (custom_Read_1_seq), 7 bp for index sequencing (custom_i7_seq), and 25 bp for Read 2 (custom_Read_2_seq). Read 1 reads through the transcript sequence. Read 2 reads through the single-cell barcode and unique molecular identifiers.

scRNA-seq data processing. Raw data were processed with custom scripts. Raw reads were first trimmed with the FASTX toolkit, version 0.0.13 (fastx_trimmer –Q 33). Read 2 was trimmed for extraction of the single-cell barcode (8 bp) and the unique molecular identifiers (4–8 bp), and Read 1 was trimmed to 30 bp get rid of a potential oligo(dT) sequence. After merging of the different parts (barcode, unique molecular identifiers and transcript sequence), reads were filtered for quality (more than 80% of the sequence had a Sanger Phred+33 quality score of >33) with the command fastq_quality_filter -v -Q 33 -q 20 -p 80. Then the reads were assigned to each single cell through use of the 8-bp barcode and the fastx_barcode_splitter.pl tool script for a maximum of two mismatches. Reads assigned to each single cell were then trimmed again to retrieve the transcript sequence with the command fastx_trimmer.

Mapping was performed using TopHat2 to the mm10 mouse transcriptome and strand information was kept with the following options: tophat -p 2–library-type fr-firststrand–read-mismatches 5–read-gap-length 5–read-edit-dist 5–no-coverage-search–segment-length 15–transcriptome-index. Duplicated mapping reads were filtered out with the unique molecular barcodes as follows. First, duplicated mapped reads were marked via the command picard-tools-1.79/MarkDuplicates.jar. Then, the genomic position of the duplicated reads were extracted, and for each of these positions, only reads with unique molecular identifiers were then kept. Reads that mapped to multiple positions were filtered out via the parameter flag 256 of the SAMTools format (sequence alignment map) for storing large nucleotide sequence alignments and utilities for manipulating alignments. Finally, reads were assigned to genes through the use of htseq-count software and the biomart_mm10_gene.gff reference transcriptome with the following options: -s yes -m intersection-nonempty. The script was modified to assign reads that overlapped in several genes to the one closest to a 3′ end.

Counts were normalized between cells by quantile normalization using the normalize.quantiles function in preprocessCore collection of preprocessing functions (in software of the R project for statistical computing) to account for differences between cells in read depths. However, inherent sampling biases can occur during analysis of single-cell RNA-seq data that can cause some transcripts, particularly those with low expression, to remain undetected. These undetected events are known as 'dropouts.' Therefore, to account for these sampling biases, we calculated probability that a given transcript was unsampled (versus genuinely unexpressed) with the scde.failure. probability function of the SCDE ('single-cell differential expression') package[28]. 'Confidence probabilities' were calculated as 1 – SCDE dropout probability. An event with a confidence *P* value of less than 0.05 was considered a genuinely unexpressed event. Conversely, events with confidence values of greater than 0.95 were considered significantly confident.

To determine how frequently individual genes were expressed in the mTEC population, we calculated frequency of expression per gene as the number of mTECs expressing a given gene (specifically, if >0 reads were detected for a given transcript in a given cell) divided by the total number of mTECs; wild-type and *Aire* Aire-deficient frequencies were calculated independently. Similarly, we calculated mean counts from expressing cells to determine the transcriptional output of individual genes when that gene is expressed; therefore, we simply averaged nonzero counts per gene for wild-type mTECs and *Aire*-deficient mTECs, separately.

Gene-set definitions. Aire-induced and Aire-neutral gene lists used in many of our analyses were defined above (**Fig. 1**). Specifically, Aire-induced genes were those whose expression was at least twofold higher in wild-type mTECs than in *Aire*-deficient mTECs, at the population level. Aire-neutral genes were defined as those whose expression did not differ more than 1.1-fold in wild-type mTECS versus *Aire*-deficient mTECs.

To control for unrelated effects that could result simply from different levels of transcriptional output from different loci, we used expression-matched gene sets defined by scRNA-seq data in many of our analyses. For this, we selected genes in the expression windows for second-highest maximum read counts (that is, the number of counts per gene that was the second highest among all cells in that group) among our single-cell data. We specifically did not use maximum read counts to avoid confounding, outlier events.

Simulation of intensity-frequency joined distributions. We aimed at testing the change, between *Aire*-deficient mTECs and wild-type mTECs, in gene-expression frequency versus the change in mean expression for Aire-induced genes. To take into account the higher dropout probability observed for low expressed genes, we derived a null distribution for Aire-neutral genes of the changes in frequency that might result from changes in mean intensity in positive cells for Aire-induced gene: For each Aire-induced gene ('Gi'), we randomly sampled 50 random Aire-neutral genes expressed at the same level as Gi in Aire-deficient cells, and another 50 genes expressed at the same level as Gi in wild-type mTECs. We then calculated the average change in frequency for these 50 random pairs, and plotted it against their mean difference in expression (**Supplementary Fig. 1**).

Correlation and clustering analyses. We used a row-standardized expression matrix weighted by the confidence of expression (1 – dropout) as in published studies[24]. Specifically, expression levels per gene were standardized among all cells using the scale function in R. For zero-read events in the raw count data, the standardized expression value was multiplied by the expression confidence value (1 – dropout) to correct for dropout biases.

Gene-gene and cell-cell Pearson correlations were performed with the 'cor' function in R. To identify co-expressed gene networks and highly similar cell subsets, we clustered our expression data by affinity propagation based on Pearson correlations with the 'corSimMat' function in the apcluster package[36]). Affinity propagation was useful in this case, as it does not require a known number of clusters *a priori*.

To test the validity of the gene clusters observed in the wild-type data set, we shuffled our real data by randomly redistributing read counts per gene among wild-type cells with the sample function in R per row of the data matrix. We shuffled the data and ran apcluster with the wild-type data for 1,000 permutations, storing cluster size and mean correlation per cluster for all permutations, with a custom script.

For cell clusters, affinity propagation with apcluster was used to determine cell groups on the basis of the expression of Aire-induced genes located on chromosome 1. To determine whether the same cell groups were still highly similar on the basis of the expression of Aire-induced genes from other chromosomes, we maintained the same order determined by our initial analysis using genes on chromosome 1 and calculated cell-cell Pearson correlations on the basis of Aire-induced genes from chromosomes 2 and 7.

doi:10.1038/ni.3247

We used t-SNE computation to visualize the cell-cell heterogeneity we observed in our wild-type and *Aire*-deficient mTECs as a simple two-dimensional representation. We calculated t-SNE components on the basis of Pearson correlations of Aire-induced gene confidence probabilities with the t-SNE package[40].

Gene cluster chromosomal distances. To determine what genomic distances the components of the gene clusters spanned, we matched each gene per cluster with its most highly correlated partner (using the 'cor' function in R). On the basis of the TSS positions of the genes, each gene was designated as inter-chromosomal (located on different chromosomes), intrachromosomal (same chromosome, but >1 Mb away), or local (same chromosome and <1 Mb away) on the basis of the distance to that gene's most highly correlated partner.

50. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

doi:10.1038/ni.3247