Supporting Information

Giraud et al. 10.1073/pnas.1119351109

SI Materials and Methods

Mice. *Aire*-deficient mice on the C57BL/6 (B6) and nonobese diabetic genetic backgrounds were derived and genotyped as previously described (1). Mice were housed at the Center for Animal Resources and Comparative Medicine at Harvard Medical School under Institutional Animal Care and Use Committee-approved procedures.

Thymus Digestion and Medullary Epithelial Cell (MEC) Sorting. Thymi of 4-wk-old individual WT or *Aire* knockout (KO) mice were dissected and trimmed of fat and connective tissue. Thymic lobes were cut into small pieces with a pair of fine scissors and agitated in RPMI to release thymocytes. The resulting fragments were digested 30 min at 37 °C in RPMI medium containing collagenase D (1 mg/mL final) (Roche) and DNase I (1 mg/mL final) (Sigma) and were further agitated with a 1-mL pipette to free more thymocytes. Enzyme mixtures with released thymocytes were removed after fragments had settled. RPMI containing a collagenase/dispase mixture (2 mg/mL final) (Roche) and DNase I (2 mg/mL final) was added, and the mixture incubated at 37 °C. Every 5 min, cells were agitated using a Pasteur pipette until a single-cell suspension was obtained. Cells were then passed through 70-μm mesh, spun down, and resuspended in staining buffer (PBS containing 1% FBS and 5 mM EDTA). Fluorophore-labeled antibodies CD45-PerCPCy5.5 (1:50) (Biolegend), Ly51-PE (1:800) (Biolegend), and I-A/E-APC (1:1,200) (eBioscience) were added to the samples and were incubated with resuspended cells for 20 min at 4 °C. Cells were washed, then resuspended in 400 μL of staining buffer. Sorting of MECs (CD45-Ly51-I-A/E^{high}) or GFP+MECs for lentigenics was performed on the Aria cell sorter (BD Bioscience).

Gene Expression Profiling. Total RNA was prepared from sorted MECs using TRIzol. Single-stranded DNA in the sense orientation was synthesized from total RNA with random hexamer priming using the GeneChip WT cDNA Synthesis and Amplification Kit (Affymetrix) and following the 100 ng total RNA protocol. The DNA was subsequently purified, fragmented, and terminally labeled using the GeneChip WT Terminal Labeling Kit (Affymetrix) incorporating biotinylated ribonucleotides into the DNA. The labeled DNA was then hybridized to Mouse Gene (MoGene) ST1.0 microarrays (Affymetrix), washed, stained, and scanned.

Feature-Level Microarray Analyses. We processed the raw probe-level data files (.CEL) from the Mouse Gene ST1.0 microarrays that we generated, as well as ST1.0 datasets deposited in the National Center for Biotechnology Information's (NCBI's) Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) public database. Datasets for the factors Klf4 (GSE17447), Foxl2 (GSE16853), and Stat3 (GSE17841) were generated on Mouse Gene ST1.0, those for LMX1B (GSE12008) and CDK8 (GSE19199) on Human Gene ST1.0. Raw probe-level data (.CEL files) were normalized by the robust multiarray average algorithm (2) and summarized using the R-package aroma.affymetrix (http://www.aroma-project.org/). The outcome files were used to define lists of probesets that are up-regulated, down-regulated, or neutral to Aire effect. Names of RefSeq transcripts corresponding to the probesets were retrieved according to up-to-date information provided on the Affymetrix NetAffx Web site (http://www.affymetrix.com/analysis/index.affx). Genome-wide location of the exons of the RefSeq transcripts on mm9 (mouse) or hg19 (human) builds was extracted from the University of California, Santa Cruz (UCSC) database using the Galaxy tools (http://galaxy.psu.edu/). A feature-level expression file was generated just before the summarization step by aroma.affymetrix, and features having expression values in the background for every replicate of each condition (with or without the perturbed factor) were removed from the analysis (10% of all of the features). The remaining features had their mm9 or hg19 genome-wide location retrieved from the Affymetrix NetAffx Web site and were mapped to the exons of the selected RefSeq transcripts. Raw expression values of the features, along with their exonic localization were used to perform the feature-level analysis using our R-implementation of PLATA (See *SI Text*) (3). As in the original script, exons bearing at least two features were tested for imbalanced expression: the expression of a given exon was called imbalanced (*i*) if the ratio between exon WT/KO fold change to transcript WT/KO fold change was >2 or <0.5, and (*ii*) a *t* test *P* value <0.05 was found for a WT vs. KO comparison of expression values of all features mapping to the exon, normalized to whole-transcript expression. For genes thus flagged for exon1 imbalance, expression levels of the features were plotted against their distance to the transcriptional start sites (TSSs) (from UCSC). An index was then computed as the ratio of the medians of expression values in promoter-proximal (TSS to 200 bp) vs. distal (>200 bp) features of the microarray.

ChIP-seq Analysis. Images acquired from the Illumina/Solexa sequencer were processed through the bundled Solexa image extraction pipeline, which identified polony positions, performed base-calling, and generated quality control statistics. Sequences were aligned using Bowtie software to NCBI Build 36 (UCSC hg18) of the human genome, and peaks were called by MACS (1.4.0rc2) (4). Refer to Table S2 for a list of the total number of mapped reads used for analysis of each ChIP-seq dataset. Analysis methods followed the previously published methods (5).

For comparison of Pol-II occupancy on Aire and control-transfected 293 cells, quantile normalization was used to normalize the datasets to be compared.

The complete set of RefGene based on hg18 was downloaded from the UCSC Web site (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database) in November 2010 for the annotation of genes. Within all genes listed in the RefGene, we picked up the genes that have probe(s) on the Human Genome U133 Plus 2.0 Array (Affymetrix) and are annotated to a position on the genome without any overlapping with other genes. Genes that are longer than 2.5 Kbp are applied for the analysis of ChIP-seq data (9,089 genes).

Gene expression profiles by the Human Genome U133 Plus 2.0 Array (Affymetrix) on Aire and control transfected 293 cells are based on previously published data (6), and the mean values were calculated for genes with multiple probesets. We took genes that have fold induction (293-Aire/293-control) <0.5 for Aire-depressing genes and >2 for Aire-induced genes. For one Aire-induced gene, five expression matched genes, which have fold induction between 0.7 and 1.3 and of which expression value is 70–130% of each Aire-induced gene, are picked up randomly to build the Aire-neutral gene group.

1. Anderson MS, et al. (2002) Projection of an immunological self shadow within the thymus by the aire protein. *Science* 298:1395–1401.

2. Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.

3. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 320:1643–1647.
4. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9:R137.
5. Rahl PB, et al. (2010) c-Myc regulates transcriptional pause release. *Cell* 141:432–445.
6. Abramson J, Giraud M, Benoist C, Mathis D (2010) Aire's partners in the molecular control of immunological tolerance. *Cell* 140:123–135.

Table S1. List of the genes induced or repressed in WT vs. *Aire* KO MECs

Table S1 (XLS)

The "Probeset ID" column lists the Affymetrix MoGene probesets with expression value fold changes >2 ("Aire induced genes" tab) or <0.5 ("Aire repressed genes" tab). A Refseq transcript is assigned to each probeset, and its specific expression value calculated by our exon imbalance R-script is shown in the "WT" and "Aire KO" columns. Exon1 imbalance and its associated *P* value (uncorrected) are also shown in the two last columns.

Table S2. Summary and primary data for ChIP-seq experiments

Table S2 (XLS)

The first page presents summary metrics from the genome mapping and peak calling (peaks were considered significant if passing a *P* value of 10^{-5} or better by Poisson distribution). The following sheets display significant tags pileups according to genomic assignments for each sample.

Table S3. Pol2 density at gene TSS in control and Aire-transfected cells

Table S3 (XLS)

First two columns: gene ID and symbols. Next three columns: gene expression in control and Aire-transfected cells, on U133 microarrays (1). Last four columns: Pol2 density in ChIP-seq experiments (Exp2 was normalized by quantile normalization).

1. Abramson J, Giraud M, Benoist C, Mathis D (2010) Aire's partners in the molecular control of immunological tolerance. *Cell* 140:123–135.

Other Supporting Information Files

Dataset S1 (PDF)