Dynamic Transformations of Genome-wide Epigenetic Marking and Transcriptional Control Establish T Cell Identity

Jingli A. Zhang,1 Ali Mortazavi,1,2,3 Brian A. Williams,1,2 Barbara J. Wold,1,2 and Ellen V. Rothenberg1,*

1Division of Biology 156-29
2Beckman Institute
California Institute of Technology, Pasadena, CA 91125, USA
3Present address: Developmental and Cell Biology Department, 2218 Biological Sciences III, University of California, Irvine, Irvine, CA 92697-2300, USA
*Correspondence: evroth@its.caltech.edu
DOI 10.1016/j.cell.2012.01.056

SUMMARY

T cell development comprises a stepwise process of commitment from a multipotent precursor. To define molecular mechanisms controlling this progression, we probed five stages spanning the commitment process using RNA-seq and ChIP-seq to track genome-wide shifts in transcription, cohorts of active transcription factor genes, histone modifications at diverse classes of cis-regulatory elements, and binding repertoire of GATA-3 and PU.1, transcription factors with complementary roles in T cell development. The results highlight potential promoter-distal cis-regulatory elements in play and reveal both activation sites and diverse mechanisms of repression that silence genes used in alternative lineages. Histone marking is dynamic and reversible, and though permissive marks anticipate, repressive marks often lag behind changes in transcription. In vivo binding of PU.1 and GATA-3 relative to epigenetic marking reveals distinctive factor-specific rules for recruitment of these crucial transcription factors to different subsets of their potential sites, dependent on dose and developmental context.

INTRODUCTION

T lymphocyte development illuminates the stepwise process of cell fate choice for descendants of multipotent stem cells. Notch pathway signaling in the thymus causes hematopoietic precursors to become committed to the T cell fate, while mobilizing a T cell gene expression program that prepares the cells for T cell antigen receptor (TCR) expression, TCR-based repertoire selection, and long, versatile careers as immune effectors. Sequential events that exclude alternative lineages occur at phenotypically well-defined stages within the thymus, providing a revealing model for the kinds of events needed to channel multipotent stem cells into a single developmental path (Rothenberg, 2011; Yang et al., 2010). However, major questions about the molecular mechanisms involved in this process have remained.

One question is how commitment works. Regulatory genes that promote access to alternative fates are either expressed or inducible in the precursors entering the thymus but end up not only repressed but irreversibly silenced as a result of commitment. The mechanisms responsible for these regulatory changes have been unknown.

Another question has been how the T cell program is deployed. Notch signaling initiates and sustains differentiation. T cell development also depends on additional transcription factors, including E2A and HEB, TCF-1 and LEF-1, GATA-3, Myb, Runx1, Ikaros, and Gfi1 (reviewed in Rothenberg et al., 2008). However, it is not clear if this list is complete, and how these factors work remains murky because so few T cell-specific cis-regulatory elements have been identified. Almost none have been functionally dissected in enough detail to explain fully the expression of the genes they control.

In other hematopoietic cell types, key cis-regulatory sequences of developmental genes have been identified through the collaborative binding of factors known to confer cell type identity. For example, combined binding sites of E2A, EBF1, and/or Pax5 predict cis-regulatory elements in developing B cells (Lin et al., 2010; Schebesta et al., 2007). In contrast, no formula known a priori has been useful to define T lineage-specific cis-regulatory elements. However, if all the cis-regulatory elements that are “in play” at crucial transitions of T cell development could be defined, then the motifs enriched in these elements could be matched with the cognate transcription factors that also change at those stages (Novershtern et al., 2011), thus narrowing the search for the key factors in commitment.

Here, we identify the dynamic transformations in transcription and epigenetic marking that occur across the genome through five stages of T cell differentiation that span lineage commitment. The results provide a genome-wide view of a lineage choice process in unusually fine resolution. To test the functional relevance of the histone marking patterns at potential cis-regulatory elements, we also track in vivo binding of GATA-3 and PU.1, two transcription factors with complementary roles in early T cell development (Rothenberg and Scripture-Adams, 2008).
Recruitment rules for these two factors are revealed to be context dependent but differently affected by dose. The results also reveal how an initial regulatory phase dominated by stem/progenitor-cell regulatory genes first overlaps with Notch signaling and then is dismantled to establish T cell identity.

RESULTS

Capturing Commitment
Our goals were first to map comprehensively the genes that undergo transcriptional change during T lineage choice, especially genes encoding transcription factors, and, second, to locate likely cis-regulatory sites mediating these gene expression changes by defining regions where histone marks are altered at each step of the process.

In the first major stage of T cell development, “early T cell precursors” or Kit+ DN1 cells, pass through the DN2a stage to the DN2b stage, when they undergo T lineage commitment. Postcommitment, they accumulate in the DN3 stage, during which they rearrange the TCR genes. Only cells that successfully express TCR proteins ever proliferate again, differentiating to the DP (CD4+ CD8+ double-positive) stage in a process called “β-selection” (reviewed in Rothenberg et al., 2008). Cells are selected after this based on their TCR recognition specificity, and further differentiation refines their mature immunological roles.

To obtain enough of the earliest cells for genomic analysis, we used an in vitro differentiation system that generates copious yields of early T cell precursors from fetal liver (FL)-derived hematopoietic progenitor cells. These precursors are cocultured with lymphoid-primissive cytokines and OP9 stromal cells expressing a Delta-like Notch ligand (OP9-DL1). In these conditions a cohort of FLDN1 and FLDN2a cells is generated by day 4.5 of culture, mostly progressing to FLDN2b cells by day 8.5 (Figure S1A available online). For an in vivo counterpart, we purified slightly more advanced DN3 stage cells from freshly isolated adult mouse thymus, and to show the effects of β-selection, DP thymocytes were also purified (ThyDN3, ThyDP) (Figures S1B and S1C; see Extended Experimental Procedures).

In vitro differentiated FL-derived DN1 and DN2 cells showed gene expression well matched to that of normal in vivo thymocyte counterparts (Yui et al., 2010; David-Fung et al., 2009; http://www.immgen.org; Heng et al., 2008). Their lineage commitment status was also in good agreement with that of in vivo counterparts (Rothenberg, 2011; Yui et al., 2010), as shown by shifting cells to non-T conditions (Table S1), despite some minor differences (Table S1 legend). As in adult thymus in vivo, cells became committed from DN2a to DN2b.

Global Gene Expression Analysis: Selective Changes during Early T Cell Development
We used RNA-seq (Mortazavi et al., 2008) to identify when major changes in gene expression occurred along the pathway from early T cell precursor to DP stages, using two to three independent biological replicates each of FLDN1, FLDN2a, FLDN2b, ThyDN3, and ThyDP cells (Pearson r > 0.97 for independent replicates of the same stages, Figure S1D). About 10,000 of the 20,861 Refseq genes were detectably expressed (≥1 reads per kilobase exon model per million reads [RPKM]) in each population; of these, ~50% changed significantly in expression (p < 0.001) between at least one pair of stages and ~40% changed from FLDN1 to ThyDP (Figure 1A).

Figure 1B shows hierarchical clustering of the expression patterns of the 3,697 genes that change expression by 2-fold or more between any stages. Between DN1 and DN2b key T cell-specific genes involved in pre-TCR expression and function were induced from a low or undetectable level (Table S2), i.e., genes encoding TCR complex components Cd3g, Cd3d, Cd3e, Cd2z (Cd247), T cell-specific signaling components Itk and Lat, recombinase Rag1, mutagenic DNA polymerase Dntt, and the surrogate α chain (pTα) Ptda. In addition, a conspicuous group of genes were repressed or silenced during these transitions. They included progenitor cell-specific growth factor receptor genes Kit, Flt3, and Csf2rb and a set of transcription factor genes described below.

The DN1 to DN2 transition is the first definitive sign of T lineage entry induced by Notch signaling. However, a much larger difference was seen between precommitment FLDN2a and postcommitment FLDN2b cells (2,429 genes different, Figures 1A and 1B) than between FLDN1 and FLDN2a (<900 genes different), also seen by hierarchical clustering. Conversely, despite their different origins and manipulation, the newly committed FLDN2b and ThyDN3 populations were more similar to each other as well (Figures 1A and 1B). Thus, the major genome-wide transcriptomic changes leading to T lineage identity do not occur in the DN1 to DN2a transition, but rather in transition to the DN2b or DN3 stages, linked with commitment.

Transcription Factor Expression Dynamics in T Lineage Commitment
Genes likely to encode transcriptional regulators (Table S3A and Extended Experimental Procedures) included 379 that changed expression by 2-fold or more (Table S3B). Hierarchical clustering of their patterns of expression (Figure 1C) again showed similarities between FLDN1 and FLDN2a and between FLDN2b and ThyDN3, while the precommitment FLDN2a cells were more different from the newly committed FLDN2b (Figure 1C); the ThyDP cells were the most different from all. Thus, the two major transitions in regulatory gene expression occur at commitment and at β-selection.

From FLDN1 to FDN2b, the most strongly upregulated “regulatory” loci in the whole genome were found to be Lef1 and Bcl11b (>75-fold increased), Pou6f1, SpiB, Ikzf3, and Ets1 among others also increased >8-fold, with weaker increases for Id3, Tcf12, Gfi1, Tcf7, Hes1, and Gata3 (Table S3B). However, many regulatory genes sharply decreased in expression between FLDN1 and FLDN2b, including genes with known, important functions in hematopoietic progenitors, e.g., Gfi1b, Lmo2, Mef2c, Hoxa9, Spi1 (PU.1), Gata2, Mycn (N-Myc), Cebpb, Bcl11a, Hhex, Nfe2, Lyl1, and several Irf factors. A major regulatory shift, with broad repression of progenitor cell transcription factor genes, thus accompanies T lineage commitment.

Dynamic Histone Modification Changes Identify Developmentally Regulated Promoters and Distal cis-Elements
The specific cis-regulatory elements affected by changing transcription factor action during commitment should be sites of
of marking at each stage as correlated with RNA expression of the nearest genes. Figure S3 presents these comprehensive results for TSSs and non-TSSs for all Refseq loci in hierarchically clustered heat maps.

Consistent with previous reports, H3Ac and H3K4me2 were overwhelmingly seen at promoters of expressed genes. Silent genes fell into two classes, with only ~35% showing H3K27me3 at their promoters. H3K27me3 was more common at silent regulatory genes than at other silent loci (Figures 2A and 2B, right). However, > 25% of the silent genes, including many with H3K27me3, were also marked with H3K4me2 (Figures S2C and S3), consistent with at least three kinds of repressed states (Filion et al., 2010). The cumulative frequency plots in Figure 2C show that genes with H3K4me2 but not H3Ac at their promoters in a given stage (red tracks) (+H3K27me3) were most likely to be newly repressed or poised for upregulation in the next stage. Thus, H3K4me2 without H3Ac marks developmentally labile promoters (Koche et al., 2011; Orford et al., 2008).

Histone modification at promoters was relatively stable across development, more than levels of corresponding RNAs (Figure S3A). However, distal elements were more dynamically marked: more than one-third of all regions marked with H3K4me2 or H3K27me3 in one stage lacked those marks in at least one other stage (Figure S3B). Thus regulatory shifts occur in development most sensitively affect histone marking at nonpromoter elements (Heintzman et al., 2010; Lin et al., 2010). For many enhancers, H3K4me2 provides more precise localization than H3K4me1 (Koche et al., 2011). The results from independent biological replicates again showed excellent correlation (Figure S1E).

The 42,000 regions with marks were distributed near and distal to the annotated TSS of expressed and silent genes (Figures 2A and S2B); Tables S2A (TSS) and S2B (non-TSS) report intensities of marking at each stage as correlated with RNA expression of the nearest genes. Figure S3 presents these comprehensive results for TSSs and non-TSSs for all Refseq loci in hierarchically clustered heat maps.

Consistent with previous reports, H3Ac and H3K4me2 were overwhelmingly seen at promoters of expressed genes. Silent genes fell into two classes, with only ~35% showing H3K27me3 at their promoters. H3K27me3 was more common at silent regulatory genes than at other silent loci (Figures 2A and 2B, right). However, > 25% of the silent genes, including many with H3K27me3, were also marked with H3K4me2 (Figures S2C and S3), consistent with at least three kinds of repressed states (Filion et al., 2010). The cumulative frequency plots in Figure 2C show that genes with H3K4me2 but not H3Ac at their promoters in a given stage (red tracks) (+H3K27me3) were most likely to be newly repressed or poised for upregulation in the next stage. Thus, H3K4me2 without H3Ac marks developmentally labile promoters (Koche et al., 2011; Orford et al., 2008).

Histone modification at promoters was relatively stable across development, more than levels of corresponding RNAs (Figure S3A). However, distal elements were more dynamically marked: more than one-third of all regions marked with H3K4me2 or H3K27me3 in one stage lacked those marks in at least one other stage (Figure S3B). Thus regulatory shifts occurring in development most sensitively affect histone marking at nonpromoter elements (Heintzman et al., 2010; Lin et al., 2010). H3K4me2-marked distal elements in fact included a number of previously noted regulatory elements (Figures S4A and S4B): the DP-specific Rag1-Rag2 gene antisilencer 71–75 kb 5′ of the Rag2 gene (Yannoutsos et al., 2004), and the DN2b/3-specific E1a promoter for the Notch1 gene (Gómez-del Arco et al., 2010). Both of these discrete cis-elements acquired H3K4me2 specifically at stages when they contribute to gene
Figure 2. Distinct Gene Expression Patterns Are Associated with Characteristic Histone Modifications

(A) Gene expression and histone modifications at the TSS of 20,861 genes. Expressed (≥1 RPKM) and silent (<1 RPKM) genes defined by RNA-seq. Ac: H3Ac, me2: H3K4me2, me3: H3K27me3. Also see Figures S2 and S3, Table S2.

(B) Gene expression and histone modifications at the TSS of 1,646 genes encoding DNA-binding proteins or transcription factors. Also see Table S3.
regulation (Figures S4A and S4B). Other non-TSS regions with developmentally dynamic marking may thus locate stage-specific cis-regulatory elements as well.

**Timing of TSS Epigenetic Changes Relative to Transcriptional Changes**

To relate the timing of changes in TSS marks with changes in RNA expression during T cell commitment, we focused on 3,697 differentially regulated genes. First, these were subdivided by K-means clustering into 25 clusters based on expression pattern (Figure S5; genes listed in Table S4A). Figure 2D tracks histone marks from stage to stage at the TSS of genes undergoing upregulation (clusters 1, 2, 6), downregulation (clusters 7, 9, 23), and transient decreases (cluster 12) or increases in expression (clusters 17, 19). H3Ac modification (Figure 2D, first group of columns) was tightly coordinated with presence of RNA (last columns), but H3K4me2 was often present before and after expression (second columns, e.g., clusters 1 and 6). Though H3K27me3 was inversely correlated with expression, only a fraction of repressed genes ever acquired this mark (Figure 2D, third columns), as also seen at promoters genome wide (Figure S3A).

Most relevant to the regulatory decisions in T lineage commitment (Rothenberg, 2011; Yang et al., 2010) are effects on genes needed for other hematopoietic cell fates, options that are shed in an ordered process. We identified 389 key hematopoietic genes by Gene Ontology (Table S4B), including “signature” regulators of erythroid cells (Gata1, Nfe2, Epor, myeloid cells (Sfpi1, Cebpa, Cebpe, Csf1r), B cells (Pax5, Ebf1), NK cells (Eomes, Il2rb), and stem cells (Gata2, Tal1, Lmo2), and tracked their expression from FLDN1 to DP in parallel with the status of histone marks at their promoters (Figure 3). Full results are shown in Table S4B and the Figure 3 master panel, while zoom-in panels allow individual genes to be identified. Again, H3Ac modification at promoters was tightly correlated with transcription, while H3K4me2 marking also preceded and persisted after transcription.

To explain alternative lineage exclusion in T cell commitment, either one or diverse mechanisms of silencing of non-T regulatory genes might be used. In fact, H3K27me3 use at these functionally relevant loci was both variable and dynamic. Some genes were silent throughout T cell specification, and many had strong H3K27me3 marks at the promoter, either apparently with H3K4me2 (e.g., Group e, Epor, Irf4, Ebf1, and Eomes) or without (e.g., Group e, Pax5). Other regulatory genes were turned off during development, often gaining H3K27me3 while they lost H3Ac (Group d). Some genes poised for early silencing already had some H3K27me3 at the TSS from FLDN1 stage (e.g., Cebpa in Group e, Gata2, Lmo1, Tal1 in Group d), suggesting repression already underway in at least part of the population. However, H3K27me3 did not mandate future silencing, for some T cell genes like Lef1 were strongly activated during commitment despite initially strong H3K27me3 marking (Group b). Furthermore, other genes stayed silent from FLDN1 to ThyDP without any H3K27me3 at the TSS (e.g., Cebpe, Cx3cr1, Zbtb32, Cd79a, and VpreB1; Group c). Unexpectedly, these variations in H3K27me3 marking cut across myeloid, erythroid, NK cell, and B cell program boundaries.

**Most Epigenetic Change in T Cell Development Occurs from DN1 to DP**

Both the foreshadowing of future expression by H3K4me2 marking of promoters and the ability of some genes to be repressed without appearance of H3K27me3 (Figures 2D and 3) raised the question of whether we might be missing changes in promoter status either before DN1, or after DP stage. We therefore compared our results with H3Ac, H3K4me2, and H3K27me3 ChIP-seq data for a prethymic lymphoid precursor population, “PPB” [EBF−/− pre-pro B cells (Lin et al., 2010; Heinz et al., 2010)] and H3K27me3 data for postthymic naive CD4 T cells, “CD4” (Wei et al., 2009), shown in flanking columns in Figures 2D and 3. The prethymic lymphoid precursor data in the great majority of genes affected by T cell specification. Furthermore, repressed genes that lacked H3K27me3 marks by the DP stage in our samples also remained silent without H3K27me3 marks in the mature T cells (the uniquely regulated Rag genes were an exception). The FLDN1 to ThyDP interval thus encompasses the crucial epigenetic changes for the great majority of genes affected by T cell specification.

**Distinct Mechanisms Control Key Developmental Genes**

Changes in modification at distal sites (compiled in Table S2B) as well as TSSs (Table S2A) often appeared implicated in gene regulation, as shown for key genes in Figure 4. Figures 4A and 4B profile two highly T cell-specific loci activated in parallel from DN2a to DN2b, the Cd3gde gene cluster (A) and Bc111b (B). These genes initially lack RNA transcripts (black tracks) and H3Ac marks (blue tracks) in FLDN1 cells, but then are strongly upregulated and kept on thereafter.

For the Cd3 genes, there was no H3K4me2 (red tracks) at the promoters and light H3K27me3 marking across the locus (Figure 4A, green tracks) during the initial silence, but the classic enhancer elements at the 3′ ends of Cd3e and Cd3d (Georgopoulos et al., 1988; van de Wetering et al., 1991) were already marked by focal H3K4me2. These enhancers were already accessible to transcription factor binding even in the FLDN1
stage, as shown by binding of the factor GATA-3 (Figure S4C; see below). Marking of these H3K4me2 sites intensified while H3Ac and H3K4me2 were recruited to the promoters of the genes during the DN2a/2b stages, when transcription began. A similar pattern for activation without initial promoter marking was seen for Il2ra (Figure S4D).

In contrast to the Cd3gde cluster, the Bcl11b gene (Figure 4B) began with substantial H3K27me3 (green tracks) over its promoter and the whole gene body at FLDN1 stage. However, its TSS had a cryptic positive cis-regulatory element marked by H3K4me2. Bcl11b then was activated from FLDN2a to FLDN2b stage through a process that swept back the H3K27me3 repressive marks off the promoter, while expanding the H3K4me2 marks into the first intron and creating a new H3K4me2 marked region in the third intron.

The changes in histone marks at these loci contrast with the precisely positioned but virtually unchanging H3K27me3, H3Ac, and H3K4me2 marks that characterized the Gata3 gene (Figure 4C). Despite a block of H3K27me3 close to the major promoter, this gene was already activated by the time of the FLDN1 stage, and underwent only a fewfold increase in expression after that.

Figure 3. Histone Modifications and Gene Expression Profiles of Genes Characterizing Hematopoiesis

Results for 379 “hematopoietic” genes are processed and displayed as in Figure 2D. Master panel: results for all 379 genes. (a)–(e): zoom in to indicated cluster regions of master panel to allow individual genes to be seen. Also see Figure S3, Tables S2 and S4B.

Repression of essential B cell regulatory factors, myeloid-cell regulatory factors, and stem or progenitor-cell regulatory factors is central to T lineage commitment. This clearly entailed a variety of distinct mechanisms (Figures 4D–4H). The Pax5 gene, crucial for the B cell program, had no H3Ac modified regions at any stages (Figure 4D; cf. neighboring Zcchc7 TSS mark). Small peaks of H3K4me2 marking were seen in intronic regions, one of them corresponding to a known hematopoietic enhancer (Decker et al., 2009). However, the gene was buried in H3K27me3 at all stages. Ebf1 (Figure S4E) was also repressed from FLDN1 on despite H3K4me2 at several sites.

Hhex and Bcl11a (Figures 4E and 4F), in contrast, were expressed strongly in FLDN1 cells but then downregulated sharply by the FLDN2b/ThyDN3 stages, showing evidence of distinct modulating roles for distal and TSS elements. For Hhex (Figure 4E), the TSS and two H3K4me2-marked distal regions lost activation marks as expression decreased, while H3K27me3 appeared focally at the TSS and then spread. A similar pattern was seen for Flt3 and the Zbtb7a (Thopk) gene, both active in FLDN1 and then repressed (Figures S4F and S4I). For Bcl11a (Figure 4F), H3Ac persisted at the promoter while RNA expression declined during commitment, reflecting a tail of low-level expression.
through DN3. H3K27me3 marks only appeared at the last stage of silencing in the DP stage. However, the H3K4me2 modification just downstream of the last exon decreased sharply between FLDN2a and FLDN2b, in parallel with RNA expression, suggesting a potential regulatory role for a distal element here.

The myeloid and progenitor-cell transcription factor gene Sfpi1 (encoding PU.1), silenced in parallel with Hhex and never re-expressed in most T cell lineages, used a different mechanism of repression (Figure 4G). H3Ac disappeared from the promoter while H3K4me2 marks in the upstream cis-regulatory elements of the gene (Rosenbauer et al., 2006; Zamegar et al., 2010) narrowed as transcription declined (Figure 4G). Yet minimal H3K27me3 was ever seen.

Not only were H3K27me3 marks dispensable for repression; they were also labile. Figure 4H shows that dense H3K27me3 marks on Mpxl2 (same as Eva1) diminished during a spike of RNA expression in the DN2b and DN3 stages, then returned during re-silencing in DP stage. Conversely, despite silencing during commitment, Zbtb7b (Figure S4I) is later activated for CD4+ cell positive selection.

**Early T Cell-Specific Sites for PU.1: A Positive Role**

The significance of epigenetic marks depends on their impact on transcription factor access and their own emplacement via transcription factor binding. We therefore correlated chromatin marks with binding of GATA-3 and PU.1 (encoded by Sfpi1), two factors needed for early T cell development, which play contrasting roles in the context of Notch signals (reviewed by Rothenberg and Scripture-Adams, 2008; Hosoya et al., 2010). PU.1 is one of the progenitor-associated transcription factors in early pro-T cells, but is even more critical for B, dendritic and myeloid cell development. A key question is whether it has distinct T lineage target genes or simply carries over a multipotent state.

PU.1 bound to ~34,000 sites in DN T cells, comparable to B and myeloid cells (Heinz et al., 2010). Although PU.1 RNA and protein levels decline sharply during T lineage commitment (Figure 5A; Yui et al., 2010), PU.1 site binding preferences remained consistent from stage to stage. We compared FLDN1 and FLDN2a cells; FLDN2b cells, where PU.1 is four to five times downregulated; and DP cells, where PU.1 is absent. Although PU.1 binding intensity per site was approximately four to five times lower in the DN2b cells, its site choices remained correlated with those in the earlier stages (r = 0.65–0.66) (Figure 5B).

Even so, the PU.1 binding sites in FLDN1 and FLDN2a cells were distinct from those reported in B cells, macrophages or even E2A+/− pre-pro B cells [representing prethymic lymphoid progenitors (Heinz et al., 2010; Figure 5C). Although the sites bound in the pre-pro B cells were most related, key PU.1 target sites occupied in pre-pro B cells were not bound by PU.1 in FLDN1 cells, e.g., the intronic enhancer of Pax5 (Figure 5D). De novo motif analysis showed that PU.1 target sites in FLDN1 cells had a different hierarchy of preferred sequences than in pre-pro B cells (Figure 5E). Thus, the consistent site choices of PU.1 from DN1 stage through commitment include a distinct T lineage-specific component.

PU.1 is needed to generate T cell precursors, but at high levels it inhibits expression of many T cell-specific genes, particularly if Notch signaling is interrupted (Franco et al., 2006). To test whether the T lineage-specific sites of PU.1 may be repressive, delaying T cell gene activation in early stages, we asked whether PU.1 binding specific to early T cells was linked to genes that are active or silent, as compared with sites bound by PU.1 in pre-pro B cells but empty in T cells. In fact, the sites occupied by PU.1 in FLDN1 cells, including FLDN1-specific sites, were mostly associated with “positive” marks (H3Ac and/or H3K4me2) and completely uncorrelated with H3K27me3 marking, more than sites bound by PU.1 only in E2A−/− pre-pro B cells (Figure 5F).

In the aggregate, genes with sites of PU.1 binding in FLDN1 cells were also more likely to show strong expression, with higher expression the more sites bound, including T lineage-specific sites (Figure 5G). Thus, PU.1 binding globally correlates with target gene expression in FLDN1 cells.

**PU.1 Binding Dynamics and Temporal Control of Target Gene Expression**

Experimental perturbation analyses have shown many specific genes in pro-T cells that are activated or repressed by manipulations of PU.1 level (Franco et al., 2006; A. Champekak, M.M. Del Real, and E.V.R., unpublished data). However, with so many binding sites for PU.1, binding alone clearly could not define genes that depend on PU.1 for positive or negative regulation. Most PU.1 binding sites were linked to genes expressed stably in all stages whether PU.1 is present or not, like the majority of genes expressed in T cell development overall. PU.1 could thus be opportunistically recruited to many active genes where it has no required role. To define properties of likely functional sites, we used the dynamics of PU.1 expression itself to filter the binding sites identified genome-wide.

Because PU.1 expression declines, functionally important PU.1 sites should be enriched near genes that themselves change in RNA expression, up or down, during development as a function of PU.1 binding. We therefore compared changes in local PU.1 occupancy from FLDN1 to FLDN2b (Figure 6A) with the direction and magnitude of changes in RNA expression of the linked genes, in two complementary ways. First, focusing only on the PU.1 site-linked genes that change expression from DN1 to DN2b (Figure 6B), we grouped them according to whether their linked sites all lost PU.1 occupancy faster (Figures 6B and 6C, blue) or slower (red) than the ~4-fold average global decrease (green = genes with both kinds of sites). Cumulative frequency plots were used to test if changes in PU.1 occupancy predicted the direction of changes in gene expression (Figure 6C), i.e., whether genes losing PU.1 earliest all turn off like PU.1 itself, or become activated as PU.1 repression might be relieved. Such analyses could detect both activated and repressed subgroups within a group as well as general trends. However, genes that lost PU.1 binding most rapidly (blue curve) were uniformly more downregulated and less upregulated than those with mixed sites. Conversely, almost 80% of genes with sites that retained PU.1 best (red curve) increased their expression from FLDN1 to FLDN2b.

Second, reciprocally, we classified individual PU.1 sites according to whether their linked genes were upregulated, downregulated, stably expressed, or silent across the DN1 to DN2b interval, and then assessed whether these sites near
developmentally regulated genes tended to lose PU.1 faster or slower than open but nonregulated sites, i.e., those linked to stably expressed genes (Figure S6A). Downregulated, upregulated, and stably expressed genes relinquished their PU.1 binding very differently, and sites linked to upregulated genes retained their PU.1 even better than fully “accessible,” stably expressed ones, again arguing against a repressive role.

At candidate target genes identified both by coregulation with PU.1 and by PU.1 perturbation effects, PU.1 typically occupied multiple regions, implying that full PU.1 regulatory function is commonly mediated through combinations of binding complexes. At Tal1, both PU.1 binding and local H3K4me2 were lost jointly from three regions, as transcription also declined (Figure 6D). At the TSS and intragenic regions, PU.1 loss appeared to open the way for H3K27me3 deposition. Similar patterns were seen at the TSS of the known PU.1 target Flt3, and at a downstream element and a known intronic enhancer of Hhex (Donaldson et al., 2005; Figures S4F and S4G). Other genes with binding sites that lose PU.1 early include Lmo1 and Bcl11a as well as Igta, which decrease naturally from FLDN1 to FLDN2b; all are sharply downregulated in FLDN2 cells if Stpi1 is deleted (data not shown; A. Champhekar and E.V.R., unpublished data). A PU.1 target with a different pattern of expression but a similar relation to PU.1 binding was Il7r, which is upregulated from FLDN1 to FLDN2b. Despite the decreasing level of PU.1 protein, the Il7r gene retained PU.1 through the FLDN2b stage, both at a known TSS positive regulatory site (DeKoter et al., 2002; Xue et al., 2004), and at another putative cis-element within a silent neighboring gene, Capsl (Figure 6E). Here too PU.1 regions have a positive link to expression, even for this gene integral for the T cell program.

Histone Marking and Modes of Action
PU.1 binding can recruit histone methyltransferases and create locally “open” chromatin states (Ghisletti et al., 2010; Heinz et al., 2010), and in early T lineage cells as in non-T cells, PU.1 occupancy was dynamically linked with local H3K4me2 modification. Due to the developmental stability of H3K4me2 modification at TSSs, the association was clearest at distal PU.1 binding regions, where H3K4me2 modification usually melted away as PU.1 binding decreased (Figures S6B, bottom, and S6D). In contrast to H3K4me2, PU.1 binding had little overlap with H3K27me3, even at silent genes (Figures S6C1, S6C12, and S6D). PU.1 binding-linked H3K4me2 was not simply an effect of general “accessibility” or expression level of the linked gene (Figure S6D, “silent” versus “E2A−/−” sites). Thus, PU.1 occupancy-linked changes in H3K4me2 could be used to screen candidate distal cis-regulatory regions with PU.1-dependent activity.

Globally, with or without distal binding sites, PU.1 binding near the TSS appeared most tightly correlated with a positive role (Figure 6F). Genes from diverse expression pattern clusters (see Figure 2D; Figure S5) could all harbor PU.1 binding either within the body of the gene or in flanking regions, but differed sharply in frequencies of genes with PU.1 binding at the TSS (Figure 6G; Table S6). Genes coregulated with PU.1 itself (cluster 7; also 3, 9, and 23; blue bars) were more likely to have PU.1 binding at the TSS than genes regulated divergently from it (e.g., clusters 1, 2, 6; red bars) (Table S6B). In contrast, many T cell genes that can be downregulated by high-level PU.1 (Franco et al., 2006) either had no PU.1 binding in early T cells or had binding only in the body or flanking regions of the genes. Genes with particularly low expression in DN1 stage were most impoverished for PU.1 binding at the TSS (Table S6A, $\chi^2$ test $p < 0.0001$). Thus, PU.1 binding at the promoter may provide, or indicate, specific antisilencing functions that maintain key stem and progenitor cell genes in early FLDN1 and FLDN2a stages.

Developmentally Plastic Deployment of GATA-3 Binding
GATA-3 is needed repeatedly in T cell stages from ETP/DN1 onward and is crucial for T lineage commitment, but capable of paradoxical effects at high doses (Taghon et al., 2007). Unlike PU.1, it is expressed almost stably across all the stages analyzed (Figure 4C; D.D. Scripture-Adams and E.V.R., unpublished data). We therefore asked whether it controls the same targets in the distinct regulatory states of FLDN1, FLDN2b, and ThyDP cells.

GATA-3 detectably bound only ~1,500 regions (Table S7). In accord with its recurrent T cell roles, these GATA-3 sites were enriched for cis-elements of T lineage genes including Cd3d, Tcf7, Zbtb7b, and the DP-specific Rag1-Rag2 distal enhancer (Figures S4B, S4C, S4H, and S4I). Occupancy patterns in our ThyDP samples were broadly consistent with those in DP CD3+ samples published elsewhere (Wei et al., 2011) ($r = 0.60$; Figure S7A). Yet progenitor-specific genes like Lyt1 and Erg (Figures 7A and 7B) as well as later-expressed T cell genes like Ets2 and Itk (Figures 7C and 7D) also harbored GATA-3 sites.

Changes in GATA-3 binding were strongly positively correlated with expression trajectories of linked genes both from DN1 to DN2b and from DN2b to DP (Figure 7E), and also correlated with H3K4me2 modification changes (Figure 7F). GATA-3 binding was even more likely to be a site of H3K4me2 enrichment than PU.1 binding (Figure S7B). Yet GATA-3 also bound regions linked to silent and active genes alike at early stages, and dramatically differed from PU.1 in its ability to bind regions with H3K27me3 (Figure S7C). For example, it remained bound to Zbtb7b even as it became silenced with H3K27me3 (Figure S4I). Intriguingly, GATA-3 occupancy also preceded full cis-element activation for Cd3d and the Rag enhancer (Figures S4B and S4C), suggesting a possible “pioneering” role.

Most unlike PU.1, the distribution of GATA-3 occupancies among different regions was strikingly different in FLDN1,
Figure 5. Lineage-Specific PU.1 DNA Binding Is Associated with Lineage-Specific Histone Modifications and Gene Expression.

(A) Mean RNA-seq level of PU.1 (Sfpi1) at each stage of early T cell development.

(B) (Left) Comparisons of PU.1 DNA binding site distributions between FLDN1 and FLDN2a or FLDN2b, with Pearson’s correlation coefficients (r). (Right) Comparisons of PU.1 DNA binding-associated H3K4me2 enrichment between FLDN1 and FLDN2a or FLDN2b. H3K4me2 signal densities were from ±1 kb of the summit of a PU.1 bound region. See Table S5 for all sites.

476 Cell 149, 467–482, April 13, 2012 ©2012 Elsevier Inc.
FLDN2b, and ThyDP. This was despite nearly constant protein availability, as shown by similar global occupancy levels and peak heights at stably occupied sites such as the Tcf3 (Tcf3b) promoter (Figure 7D) and the Tcrb 3’ enhancer (data not shown) in all stages. The most common motifs at regions of occupancy were a classic GATA site and an Ets family-like site (Figure 7G), in FLDN1, FLDN2b, and ThyDP alike. But from FLDN1 to ThyDP, GATA-3 occupancy increased sharply in some regions (e.g., Ets2 in DN2b, Itk promoter in DP; Figures 7B and 7C), while disappearing from others entirely (e.g., Lyt1, Erg, Itk introns; Figures 7A–7C). Overall, whereas regions occupied in FLDN1 and FLDN2b stages were moderately well correlated (Pearson r = 0.61), sites in FLDN2b and ThyDP were poorly correlated and those in FLDN1 and ThyDP entirely uncorrelated (r = –0.0064, Figure 7H).

These results locate elements in T and non-T genes where the crucial T cell factor GATA-3 can be contributing to regulation, from the FLDN1 stage on (Table S7). Nevertheless, they also reveal that a target gene for GATA-3 at one stage of T cell development may not normally receive input from GATA-3 at another stage, despite similar GATA-3 availability. In contrast to PU.1, GATA-3’s physiological deposition at any given stage depends not only on its own availability but also on a specific developmental regulatory context.

**DISCUSSION**

Our results provide a view of the inner workings of T cell development, an instructive reference case for regulatory epigenomics, and potentially powerful elements needed for explanation of this complex developmental process. Our global, base-resolution timecourse of chromatin and transcriptome changes sheds light on the finely defined stages of T cell specification. The RNA-seq data not only quantify RNA levels but also provide detailed information about promoter and exon choice that may affect gene regulation as well as function. Note that these data also reveal noncoding transcripts that may be important in various regulatory roles. Here, we focused on the ~400 regulatory gene loci that themselves are developmentally regulated during this process, and also the subset of candidate cis-regulatory genomic sites that undergo developmental changes in histone modifications, revealing changes in local regulatory inputs. These are the most likely trans and cis components of nodes in the gene network that causally drive successive steps of the T cell program.

Most powerfully, the results reveal the subcomponent processes out of which T cell specification is built. Relatively few regulatory genes are strongly activated during lineage commitment itself, and the list is now likely to be complete. Furthermore, a major feature of commitment is specific, marked downregulation of progenitor-cell genes, through an unexpectedly complex process. Many important non-T hematopoietic regulatory genes are still expressed in the precursors we examine through 4 days of consistent Notch pathway signaling, and many persist even into the DN2a stage before they are shut off. Importantly, the histone marking status of the promoters and linked cis-elements provides an independent line of evidence about the timing and mechanism of regulatory changes. Loss of H3Ac rules out an artifact of the decay kinetics of old mRNA persisting after transcription has ceased. Our results also show that the repression of progenitor-cell genes is not due to a single switch. The diverse histone mark transformations that are applied to different repression targets imply that a variety of biochemically and temporally distinct silencing mechanisms must be used. Notably, this rules out Notch signaling itself as a common mechanism of repression and implies that the T lineage program requires multiple distinct repressor functions to establish T cell identity.

Dynamically regulated transcriptional repression during this process is often separable from “epigenetic silencing.” Repeatedly, deposition of H3K27me3 histone marks follows RNA downregulation, more likely as an effect and stabilizer of repression than an initial cause of repression. De novo H3K27me3 marking appears in two distinct, major patterns. One can be by lateral invasion from a neighboring patch of pre-existing “closed” chromatin, another is by tight focal deposition at a previously active TSS or enhancer, followed by spreading. In other cases repression does not involve H3K27me3 at all, possibly due to the nature of the repressor: e.g., at Sphi1 and Cd4, two key genes known to be repressed by Runx factors in DN3 cells. Even when H3K27me3 is used, it is readily and precisely reversible. Repression via DNA methylation was not studied here but is also reversible, as shown recently by the cell type-specific demethylation of CpGs in DN2-DN3 cells at loci that include Tc77 and Bcl11b (Ji et al., 2010) (http://charm.jhmi.edu/hsc/). These examples show that transcriptional repressors must act first to trigger chromatin closing, while transcriptional activators retain power to undo it.

Our results also shed light on the positive regulation of the T cell program. Despite known essential roles, finding specific cis-regulatory targets for Notch, GATA-3, and TCF-1 has been achieved.
Figure 6. Functional and Stage-Dependent PU.1 Binding in Early T Cell Development

(A) Stage-specific and non-stage-specific (shared) PU.1 binding sites: stage-specific binding defined by four or more times difference in signal densities between FLDN1 and FLDN2b.

(B) Differential expression of PU.1 binding linked genes. Top: of 13,335 PU.1 binding linked genes in DN cells, the numbers expressed in FLDN1 (blue circle) and FLDN2b cells (red circle) are shown (7,244 stably expressed, 1,045 differentially expressed two or more times change). To test whether PU.1 occupancy correlated with positive or negative regulation, all differentially expressed genes were split among three subgroups based on changes in PU.1 binding to linked sites (see A): genes with FLDN1-specific sites only (Loss of PU.1 binding in FLDN2b, blue), those retaining PU.1 binding at all sites in FLDN2b (red), and genes that rapidly lose PU.1 binding from some sites but not others (Mixed, green).

(C) Cumulative Probability

(D) Differential expression of PU.1 binding linked genes. Top: of 13,335 PU.1 binding linked genes in DN cells, the numbers expressed in FLDN1 (blue circle) and FLDN2b cells (red circle) are shown (7,244 stably expressed, 1,045 differentially expressed two or more times change). To test whether PU.1 occupancy correlated with positive or negative regulation, all differentially expressed genes were split among three subgroups based on changes in PU.1 binding to linked sites (see A): genes with FLDN1-specific sites only (Loss of PU.1 binding in FLDN2b, blue), those retaining PU.1 binding at all sites in FLDN2b (red), and genes that rapidly lose PU.1 binding from some sites but not others (Mixed, green).
slow. The identification of a battery of cis-regulatory elements activated de novo from DN1 to DN2b yields important clues to clarify these links. GATA-3 effects in early T cells have been especially difficult to dissect, in part due to the profound loss of viability when GATA-3 dose is reduced (Hosoya et al., 2009), and in part due to lineage-inappropriate effects of GATA-3 in gain of function experiments (Taghon et al., 2007). Identification of potential GATA-3 regulatory inputs into Tcf7 as well as Tcf62a (Tcf3) from the earliest stages suggests a level of regulatory interlinkage, which could explain the acuteness of the GATA-3 requirement. At least in DP cells, data from Wei et al. (2011) suggest that the GATA-3 sites we see may positively regulate Tcf7, Cd3d, and Zfpml, and could negatively regulate Tcf62a. Our results may also help to explain GATA-3’s lineage infidelity in gain of function experiments by showing that its recruitment to legitimate target sites, even at a constant level of expression, is intensely stage specific. Altered dosages could thus override the mechanisms that must provide appropriate targeting specificity.

The ordered alternative lineage exclusion events in T lineage commitment are an ideal context to test whether developmental relatedness is preserved in a hierarchy of epigenetic chromatin changes. Clearly, separable events mediate repression of different alternative lineages. The B cell regulatory genes Pax5 and Ebf1 are silenced by H3K27me3 and rendered inaccessible to PU.1 binding from the start, whereas the myeloid regulatory gene Cebpα is bivalently marked. The myeloid and progenitor regulatory gene Spi1 (PU.1), initially fully activated, appears to play a regulatory role even within the DN2b stage, and is silenced only when T cell gene expression is under way. However, there is no simple mapping of developmental lineage exclusion order with a particular molecular class of repression mechanism. Drivers of the most “distant” fate in developmental terms, the erythroid genes, can be repressed via H3K27me3, or without it (Gata1), as can genes associated with the “closest,” NK-cell fate (Eomes and Il2rb, respectively). In an interesting additional case, many multipotent progenitor-cell regulatory genes are expressed throughout the early stages like Spi1 and only shut off during commitment itself. Some may be sustained by a common progenitor-cell positive regulator, Lmo2 (McCor-mack et al., 2010), and our results suggest that many receive input from PU.1 itself. The progenitor-associated genes may thus constitute a discrete early subcircuit within the T lineage specification network.

Finally, our multistage analysis shows that many mouse hematopoietic genes are each likely controlled by different constellations of cis-regulatory elements at one stage of development versus another, even within the same cell lineage. In this light, the quest for single, minimal sufficient regulatory elements for such genes seems naïve, as it would a priori sacrifice the full range of developmental control. The roles of the candidate cis-elements and their rules for engagement with promoters should be greatly clarified by future extensions of this analysis, to detect specific chromatin looping events, enhancer activation states mapped by association with p300 and H3K4me3, and latent enhancers using H3K4me1 at transcription factor binding sites. Mechanisms of repression could be clarified when effects on a broader range of nonactivating cis-elements are mapped based on DNase hypersensitivity and DNA methylation. The mapping of developmentally dynamic histone modification sites provides a guide to locate the sites in cis-regulatory DNA that process distinct inputs for crucial regulatory genes. In this collection of regulatory domains lie the answers to how cells are driven to T lineage commitment.

EXPERIMENTAL PROCEDURES

For full materials, see Extended Experimental Procedures. Briefly, in vitro developing CD4−CD8−TCR “double-negative” populations were generated from fetal liver hematopoietic precursors (Taghon et al., 2007) sorted as ETP/DN1 (Kit+CD44+CD25−), “DN2a” (Kit+CD44+CD25+), and “DN2b” (Kit+CD44+CD25+). “DN3” (Kit+CD44+CD25+) and “DP” (CD25+CD4+CD8+) cells were sorted from thymus. All mice were maintained and used under protocols approved by the Caltech Institutional Animal Care and Use Committee. ChIP-seq and RNA-seq were carried out and analyzed as previously reported (Pepke et al., 2009; Mortazavi et al., 2008; Johnson et al., 2007). The programs ERANGE and DEGSeq were used to compare samples.

ACCESSION NUMBERS

All data are deposited under GEO: GSE31235.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and seven tables and can be found with this article online at doi:10.1016/j.cell.2012.01.056.

ACKNOWLEDGMENTS

We thank Lorian Schaeffer and Vijaya Kumar for library preparation and sequencing, Henry Amrhein and Diane Trout for data curation, Igor Antoshechkin for sequencing facility management, Diana Perez, Josh Verceles, and Rochelle Diamond for cell sorting and advice, Tian Ling, Georgi Marinov, and Hao Yuan Kueh for statistical advice and programming, Rothenberg group members for sharing advice and unpublished data, and Robert Butler, Lorena Sandoval, and Scott Washburn for care of the mice. Support was from NIH grants R33HL089123, R01CA090233, R01CA090233-08S1, and R01CA129406-05S1.

(C) Relationship between PU.1 occupancy changes and mRNA expression changes between FLDN2b and FLDN1: cumulative distributions of expression changes for three groups of genes depicted in (B). The number of genes in each group and p values (K-S tests for comparisons with “Mixed”) are indicated next to the plots. Also see Figure S6.

(D and E) Developmentally distinct PU.1 binding patterns at the Tal1 (D) and Il7ra (E) loci in FLDN1, FLDN2a, FLDN2b, and E2A−pre-pro B cells, compared with H3K4me2, H3K27me3 and mRNA in all five immature T-populations. Also see Figure S4.

(F) Distribution of PU.1 occupancy relative to TSSs in expressed and silent genes at individual stages.

(G) Location of PU.1 sites in potential target genes according to expression pattern. Clusters of genes with different developmental trajectories (Figure S5) were scored by the number of genes they include with PU.1 binding sites ± 1 kb from the TSS (proximal) or further from the TSS (distant). □, ○ (left axis): % of genes in a cluster with proximal (□) or distal (○) PU.1 binding. Bar graphs (right axis): (number of genes with TSSs)/(number of genes with distal sites). Colors of bars relate expression pattern of each cluster to endogenous PU.1 expression (most similar: blue, inverse: red). See Figure 2D; Figure S6, and Table S6.
Figure 7. Developmental Plasticity of GATA-3 DNA Binding and Associated Epigenetic Marking

(A–D) Stage-specific GATA-3 binding (brown) in Ly1, Ets2-Erg, Itk, and Tcfe2a loci of FLDN1, FLDN2b, and ThyDP cells, shown with binding associated H3K4me2 (red) and H3K27me3 (green) enrichment and mRNA (black) expression in all five immature T-populations. Also see Figure S4.

(B) Cumulative distributions of changes in GATA-3 occupancy between FLDN2b and FLDN1 (top) and between ThyDP and FLDN2b (bottom), for genes differentially regulated across the same intervals. GATA-3 binding sites were divided into four subgroups, based on linkage to downregulated genes (blue), upregulated genes (red), stably expressed genes (<2-fold change in expression, green) and silent gene sites (<1 RPKM in both stages, black). P values are from K-S tests between stably expressed gene sites and each of the other three subgroups (n = no. of sites). See Figure S7A, Table S7.

(F) Cumulative distributions of changes in H3K4me2 marks associated with GATA-3 binding between FLDN2b and FLDN1 (top) and between ThyDP and FLDN2b (bottom) stages. H3K4me2 signal densities were calculated within −/+1 kb of the summit of a given GATA-3 bound region (depicted in Figure 7H), p values calculated as in E. See Figures S7B and S7C.
REFERENCES


(G) Most highly enriched sequence motifs in GATA-3 binding regions (see panel H). The percentages of regions containing ≥ 1 instance of each motif are indicated beneath each sequence logo, with the expected frequency of the motif in random regions in parentheses.

(h) Scatter plots depicting the comparisons in GATA-3 binding between FLDN1, FLDN2b, and ThyDP. Pearson correlation coefficients are shown for each comparison.
genes involved in B cell signaling, adhesion, migration, and immune function. Immunity 27, 49–63.


**EXTENDED EXPERIMENTAL PROCEDURES**

**Cell Culture**

Fetal liver (FL) cells from embryonic day 13.5 to 14 (E13.5-E14) C57BL/6 mouse embryos were first depleted of Gr-1\(^+\), F4/80\(^+\), Ter119\(^+\), and CD19\(^+\) ("Lin cocktail 1") cells using streptavidin-coupled magnetic microbeads (Miltenyi Biotec) against biotin-conjugated antibodies. Lin-c-Kit\(^+\)CD27\(^+\) multilineage precursors were then sorted from lineage-depleted FL cells by FACS and co-cultured with OP9-DL1 stromal cells as described previously (Taghon et al. 2005). 50-100 \times 10^5 Lin-c-Kit\(^+\)CD27\(^+\) FL cells were plated on OP9-DL1 monolayers in 10 cm plates in the presence of 5 ng/ml Flt3-L and 5 ng/ml IL-7 (both from Peprotech). After 4.5 d of culture, half of the cells were harvested and sorted to isolate DN1 (FLDN1, Lin-c-Kit\(^+\)CD45\(^-\)CD44\(^+\)CD25\(^+\)) and DN2a cells (FLDN2a, Lin-c-Kit\(^+\)CD45\(^-\)CD44\(^+\)CD25\(^-\)) (using "Lin cocktail 2" = antibodies to Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11c, TCR\(\gamma\delta\), TCR\(\beta\), CD3\(\epsilon\), CD8\(\alpha\)). After 8.5 d of culture, the rest of the cells were harvested and sorted for FLDN2a and DN2b (FLDN2b, Lin-c-Kit\(^+\)CD45\(^-\)CD44\(^+\)CD25\(^-\)). The FLDN2a samples used for analysis were each pools of day 4.5 and day 8.5 DN2a cells in approximately 2:1 ratio. For subsets from adult (4-6 weeks old) thymus, wild-type C57BL/6 mouse thymi were first depleted with antibodies to Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11b, CD11c, TCR\(\gamma\delta\), TCR\(\beta\), CD3\(\epsilon\), CD4 and CD8\(\alpha\) ("Lin cocktail 3"). Thymic DN3 (ThyDN3) cells were then sorted from lineage-depleted thymocytes as Lin-c-Kit\(^+\)CD44\(^+\)CD25\(^-\). Finally, to prepare ThyDP populations free of contaminating cells in early stages of TCR-dependent positive selection, while maintaining viability of these fragile cells, ThyDP cells were collected from TCR\(\alpha\beta\)-thymi (4-6 weeks old. The Jackson Laboratory, B6.129S2-Tcratm1Mom/J) by simple streptavidin-coupled magnetic microbead depletion with biotinylated antibodies against Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11b, Lin-c-Kit, CD44 and CD25. After this depletion, 90-95% of cells were CD4\(^+\)CD8\(^+\).

Antibodies used were from eBioscience and Biolegend, including anti-CD4 (JK1.5; biotin), anti-CD8\(\alpha\) (53-6.7; biotin), anti-CD11b (M1/70; biotin), anti-CD11c (N418; biotin), anti-CD19 (eBio1D3; biotin), anti-CD122 (SH4; biotin), anti-Gr1 (RB6-8C5; biotin), anti-F4/80 (BM8; biotin), anti-TCR\(\beta\) (H57-597; biotin), anti-TCR\(\gamma\beta\) (eBioGL3; biotin), anti-NK1.1 (PK136; biotin), anti-Ter119 (Ter-119; biotin), anti-c-Kit (2B8; PE, APC, biotin), anti-CD27 (LG.7F9; APC), anti-CD25 (PC61.5; APC-Alexa 750, APC-Alexa 780, biotin), anti-CD44 (1M7; Pacific Blue, eFluor 450, biotin), anti-CD45 (30-F11; PerCp-cy5.5), anti-CD3\(\epsilon\) (145-2c11, PerCp-cy5.5). For detection of biotinylated antibodies, streptavidin-Percp-cy5.5 was used.

**Lineage Commitment Assay**

Samples of 25 FLDN1, FLDN2a or FLDN2b cells were each sorted into 96 well plates coated with either OP9-DL1 or OP9-Mig (control) monolayers in 10 cm plates in the presence of 5 ng/ml Flt3-L and 5 ng/ml IL-7 (both from Peprotech). After 4.5 d of culture, the rest of the cells were harvested and sorted for FLDN2a and DN2b (FLDN2b, Lin-c-Kit\(^+\)CD45\(^-\)CD44\(^+\)CD25\(^-\)). The FLDN2a samples used for analysis were each pools of day 4.5 and day 8.5 DN2a cells in approximately 2:1 ratio. For subsets from adult (4-6 weeks old) thymus, wild-type C57BL/6 mouse thymi were first depleted with antibodies to Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11b, CD11c, TCR\(\gamma\delta\), TCR\(\beta\), CD3\(\epsilon\), CD4 and CD8\(\alpha\) ("Lin cocktail 3"). Thymic DN3 (ThyDN3) cells were then sorted from lineage-depleted thymocytes as Lin-c-Kit\(^+\)CD44\(^+\)CD25\(^-\). Finally, to prepare ThyDP populations free of contaminating cells in early stages of TCR-dependent positive selection, while maintaining viability of these fragile cells, ThyDP cells were collected from TCR\(\alpha\beta\)-thymi (4-6 weeks old. The Jackson Laboratory, B6.129S2-Tcratm1Mom/J) by simple streptavidin-coupled magnetic microbead depletion with biotinylated antibodies against Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11b, Lin-c-Kit, CD44 and CD25. After this depletion, 90-95% of cells were CD4\(^+\)CD8\(^+\).

**Chromatin Immunoprecipitation**

Each histone modification ChIP was generated using 5 million cells and 20 \(\mu\)g of each of the following antibodies: H3K(9,14)Ac (Millipore 06-599), H3K4me2 (Millipore 07-030) and H3K27me3 (Millipore 07-449). PU.1 ChIP was generated using 7.5 \(\times\) 10^6 cells. Independent biological replicates were generated for histone modification ChIP of FLDN1, FLDN2b, ThyDN3 and ThyDP. Different batches of histone modification antibodies (Millipore 06-599, Millipore 07-030, Millipore 07-449) were used among biological replicates. To obtain high-quality chromatin from ThyDP, cells were collected from TCR\(\alpha\beta\)-thymi (4-6 weeks old. The Jackson Laboratory, B6.129S2-Tcratm1Mom/J) by simple streptavidin-coupled magnetic microbead depletion with biotinylated antibodies against Ter-119, CD19, F4/80, Gr-1, NK1.1, CD122, CD11b, Lin-c-Kit, CD44 and CD25. After this depletion, 90-95% of cells were CD4\(^+\)CD8\(^+\).

**mRNA Purification and cDNA Library Building**

Total RNA was extracted from 2.5-20 million cells using Trizol (Invitrogen), and then subjected to two rounds of selection using Oligo-dT coupled magnetic beads (Dynabeads) according to the manufacturer’s protocol. About 100 ng polyadenylated mRNA per sample was obtained after double selection. Independent biological replicates were generated for all five populations (triplicates for FLDN2b), cDNA library building was performed as described (Mortazavi et al., 2008). Briefly, RNA was fragmented to an average length of 200 bp by Mg\(^{2+}\)-catalyzed hydrolysis and then converted into cDNA by random priming. cDNA was then subjected to end repairing, adaptor ligation, PCR amplification, size selection by gel electrophoresis (200-300 bp, insert plus adaptor and PCR primer sequences) and a second round of PCR amplification to generate each ChIP DNA library as described (Johnson et al., 2007) (Illumina ChIP-seq sample preparation kit #IP-102-1001).

**High-Throughput Sequencing**

Each ChIP DNA library or cDNA library was sequenced with the Illumina Genome Analyzer II or IIX following the manufacturer’s protocols (http://www.illumina.com; Johnson et al., 2007, Mortazavi et al., 2008).
**RNA-Seq Data Analysis**

Sequence reads from each cDNA library (38 bp, single-read) were trimmed to 32 bp long and mapped onto the mouse genome build NCBI37/mm9 using Bowtie (bowtie-0.12.1, http://bowtie-bio.sourceforge.net/index.shtml) with setting ‘-v 2 -k 11 -m 10 -t–best–strata’. The mappable reads were then processed by the ERANGE v. 3.3 RNA-seq analysis program (Mortazavi et al., 2008). Assuming total transcriptional activity is comparable between different cell types, the obtained data (data units in RPKM, reads per kilobase exon model per million mapped reads) were first log2 transformed and linearly normalized between individual samples, then averaged among biological replicates or triplicates. At the same time, in order to find genes that were changed in expression between two populations to a statistically significant degree, ERANGE processed data were analyzed by the Bioconductor DEGseq program (Wang et al., 2010) (http://www.bioconductor.org/packages/2.6/bioc/html/DEGseq.html) (data units in RPM, reads per million mapped reads, method = “MARS,” p < 0.001) (Figure 1A). This analysis yielded 3,697 DEGseq positive genes that had more than a 2-fold change in RNA-seq reads (after normalization and averaging), either between any two successive stages or between FLDN1 and ThyDP, and these were defined as differentially expressed genes. To identify differentially regulated transcription factors, we did Gene Ontology analysis of this set with key term “DNA-dependent regulation of transcription” (GO:0006350), and the resulting list used for alignment against our DEGseq set is presented as Table S3 part A.

**Hierarchical Clustering**

To determine the overall tendencies of change in gene expression and the connection between different populations, we hierarchically clustered RNA-seq data of these 3,697 selected genes from all 11 samples (using normalized data, biological replicates and triplicates were treated independently, Figure 1B). Hierarchical clustering was performed along both dimensions with sample similarities clustered first, and then genes. Euclidean distance and complete linkage were used (MATLAB 7.10.0). Separately, two-dimensional hierarchical clustering was also performed on 379 differentially expressed transcription factors (Figure 1C).

**K-Means Clustering**

To profile and categorize the behavior of clusters of similarly regulated genes during early T cell development, we first normalized individual mRNA data for the 3,697 selected genes by the corresponding geometric mean of five stages, and then performed K-means clustering analysis on the results after log2 transformation (Figure S5). K was set at 25 and squared Euclidean distance was used (MATLAB 7.10.0).

**ChIP-Seq Data Analysis**

**Histone Modification ChIP-seq**

DNA sequence reads from each ChIP-seq library (single-read) were trimmed and mapped onto NCBI37/mm9 using the same setting as for RNA-seq data, and uniquely mapped reads were used for further analysis. The data were processed by the ERANGE v. 3.3 findall peak finder (Johnson et al., 2007) to identify enriched genomic regions. We used a stringent setting of ‘-spacing 100 -minimum 4 -ratio 4 -minPeak 0.5 -shift learn’ for H3Ac and H3K4me2 ChIP-seq data, and a relatively less stringent setting of ‘-notrim -nodirectionality -spacing 100 -minimum 2 -ratio 4 -minPeak 0.25 -shift learn’ for H3K27me3. The sequence data of the input DNA from the same cell were used as background control. Since on average the total amount of mappable DNA reads of each H3K27me3 ChIP-seq data was about two times of that of each H3Ac and H3K4me2 ChIP-seq data, the minimum total DNA reads for called regions were comparable for all three histone modifications (that is, about minimum 60 to 80 enriched DNA reads per region).

All called regions (from all 27 samples) were pooled and merged if overlapping. Only resulting regions of at least 200 bp were considered for further analysis. This conservative approach treats any local change in peak height or spreading of histone modification as effects on a single region, thus providing a minimum estimate of the number of centers of regulatory change. Thus, for example, the change in shape factor of H3K4me2 commonly observed at active promoters is not considered to change the number of marked regions. We considered the positive regions overlapping ± 1 kb from the TSSs of UCSC known genes (mm9, NCBI v.37) as promoter-proximal regions, and the rest as promoter-distal regions. Individual regions were then assigned to the nearest genes using ERANGE (200 kb as the maximum radius). Signal densities (number of DNA reads) were calculated using ERANGE v. 3.3 regionCounts, for each region of every histone modification data set. For global histone modification status of promoters regions, we expanded every transcriptional starting site (TSS) of UCSC known genes to a window of ± 1 kb, and calculated signal densities of each TSS regions using ERANGE. Assuming that total DNA enrichment of the same histone marker is comparable among different cell types, we linearly normalized the read number (after log2 transformation) between samples from the same histone marker (i.e., based on slopes of correlation plots in Figure S1B). The mean for biological replicates was used for analysis. Since our RNA-seq data cannot accurately distinguish among isoforms, for genes that have multiple alternative promoters we selected one promoter that had the highest H3K4me2 level (or H3Ac if all had the same level of H3K4me2). Regions (both distal and promoter regions) that had more than 4 RPM in either H3Ac or H3K4me2, or more than 2 RPM in H3K27me3 were considered as positive for the particular histone modification(s) (Figure 2 and S2). The processed data were plotted and visualized in MATLAB.

All RNA-seq and ChIP-seq sequencing tracks were generated in WIG file format and uploaded onto the UCSC genome browser for visualization. Publicly available data used in this study (Lin et al., 2010; Heinz et al., 2010; Wei et al., 2009) were downloaded as raw sequence data (http://www.ncbi.nlm.nih.gov/geo) and remapped onto NCBI37/mm9 using the same settings.
**PU.1 and GATA-3 ChIP-Seq**

Since PU.1 ChIP enriched genomic regions were in general narrower than histone modification enriched regions, we used a setting of “-spacing 50 -minimum 2 -ratio 4 -minPeak 0.5 -shift learn -listPeak” for the ERANGE findall peak finder. The sequence data of the input DNA from the same cell type was used as background control. Publicly available PU.1 ChIP-seq and input data from E2A<sup>−/−</sup> prepro B cells, mature B cells and macrophages (Heinz et al., 2010) were downloaded as raw sequence data (http://www.ncbi.nlm.nih.gov/geo) and remapped using the same setting.

Called regions were pooled and merged if overlapping from each pair-wise or three-way comparison (from E2A<sup>−/−</sup> pre-pro B versus FLDN1, B cell versus FLDN1, macrophage versus FLDN1, or FLDN1 versus FLDN2a versus FLDN2b). Individual regions were calculated for PU.1 enriched signal densities and then assigned to the nearest genes (200 kb as the radius) using ERANGE. We next aligned the summits of all positive regions and calculated histone modification signal densities in a window of ± 1 kb. All histone modification data were linearly normalized (using the parameters generated from global histone modification analysis). The mean for biological replicates was used for analysis. Scatter plots were generated and visualized in MATLAB.

To compare differential PU.1 binding with associated differential gene expression and H3K4me2 enrichment during early T cell development, we divided PU.1 binding linked genes into four subgroups: upregulated and downregulated genes (selected from the differentially expressed genes group and having more than 2-fold change in expression from FLDN1 to FLDN2b; see “RNA-seq Data Analysis”), stably expressed genes (less than 2-fold change in expression between FLDN1 and FLDN2b), and silent genes (<1 RPKM in both stages). The changes in PU.1 occupancy and in H3K4me2 enrichment (within ± 1 kb of binding summits) between FLDN2b and FLDN1 were calculated and plotted separately as cumulative distribution for each group (Figures S6A and S6B). To determine whether PU.1 binding sites linked to upregulated or downregulated genes were more likely differentially bound by PU.1 and enriched by H3K4me2 compared to sites linked to stably expressed genes, two-sample Kolmogorov-Smirnov test was performed between stably expressed gene sites and each of the other three subgroups (Figures S6A and S6B).

To visualize histone modifications and degree of PU.1 occupancy surrounding the summits, we further expanded positive regions to a window of ± 2 kb, and divided each window into 50 bins (80 bp each). Histone modification and PU.1 enrichment were calculated for each bin using the same method mentioned above. The data obtained were aligned with RNA-seq data of associated genes, and then hierarchically clustered (one dimensional clustering of binding regions; using Euclidean distance and Ward linkage) and visualized as heat maps in MATLAB as shown in Figures S6C and S6D.

GATA-3 ChIP-seq data was processed similarly to PU.1 ChIP-seq data. To compare our findings with published results, raw sequence data for “CD3lo DP” cell samples (Wei et al., 2011) were downloaded from Gene Expression Omnibus and remapped using the same settings as used for our data, as described for comparing PU.1 results (Heinz et al., 2010) above.

**De Novo Motif Analysis**

We selected the top 1,000 PU.1 enriched peaks from each of the three subgroups (E2A<sup>−/−</sup> pre-pro B cells high, shared, and FLDN1 high), and performed MEME analysis on regions ± 50 bp from the peaks by ERANGE v. 3.3 using the default setting to generate the position specific frequency matrix (PSFM) representation of the motifs. The PSFMs were mapped separately back to the three enriched regions subgroups at 85% match (Johnson et al., 2007).

All 1,652 enriched GATA-3 regions (pooled from FLDN1, FLDN2b and ThyDP) were subjected to MEME analysis. Since the consensus sequence motifs of GATA-3 binding sites were shorter than the ones of PU.1 binding sites, the PSFMs were mapped back to the 1,652 enriched GATA-3 regions and 1,652 random genomic regions at 90% match instead. Random genomic regions were comparable to the GATA-3 binding regions in both length and chromosomal distribution.

**SUPPLEMENTAL REFERENCES**


Figure S1. Cell Purification and Biological Reproducibility, Related to Figure 1

(A) Lin-c-Kit+CD27+ fetal liver precursors were sorted from E13.5 to 14 C57BL/6 mouse embryos and co-cultured with OP9-DL1 stromal cells in the presence of IL-7 and Flt3-L. At day 4.5, FLDN1 cells were sorted from the coculture as Lin-CD45+cKithiCD44+CD25-. FLDN2a as Lin-CD45+cKithiCD44+CD25+. At day 8.5, FLDN2b cells were sorted as Lin-CD45+cKitloCD44intCD25+, and FLDN2a cells were also collected (see Experimental Procedures).

(B) ThyDN3 cells were sorted from lineage-depleted thymocytes as Lin-c-Kit-CD44-CD25+ (see Experimental Procedures).

(C) ThyDP cells were collected from B6 background thymi by depletion with Ter-119, CD19, F4/80, Gr-1, NK1.1, CD11c, c-Kit and CD44. After depletion, 90%–95% of cells were CD4+CD8+ (see Experimental Procedures). Cells lacking TCRα were used in order to enable DP cells to be generated normally, a process dependent only on TCRβ, but leaving them without the capacity to undergo positive selection to any later stages of T cell development.

(D) Global comparisons of RNA-seq between biological duplicates or triplicates (log2 transformed) for individual cell populations (including FLDN1 sample I versus II, FLDN2a sample I versus II, FLDN2b sample I versus II, FLDN2b sample II versus III, ThyDN3 sample I versus II, and ThyDP sample I versus II). The Pearson’s correlation coefficient for each comparison is indicated in the insert.

(E) Global comparisons of histone modifications ChIP-seq experiments between biological duplicates (after log2 transformed) for individual cell populations (including FLDN1 experiment A versus B, FLDN2b experiment A versus B, ThyDN2 experiment A versus B, and ThyDP experiment A versus B). From top row to bottom are H3Ac, H3K4me2 and H3K27me3. The Pearson’s correlation coefficient for each comparison is indicated in the insert.
The percentage of promoters enriched with the specified histone modification or combination of histone modifications within the respective subgroup is given in the table. Each number indicates the total amount of promoters enriched with the specified histone modification or combination of histone modifications. The percentage of promoters enriched with the specified histone modification or combination of histone modifications within the respective subgroup is given in the parenthesis.

Figure S2. Global Distribution of the Three Histone Modifications, Related to Figure 2
(A) Length distribution of discrete genomic regions enriched with the specified histone modification in at least one cell population. First, more regions were enriched with H3K4me2 in at least one cell population than H3Ac or H3K27me3. Second, while the majority of enriched regions were within 500–1,000 bp (>95% for H3Ac and H3K4me2; 81.5% for H3K27me3), H3K27me3 enriched regions were relatively more diverse in length, including some regions of much greater extent.

(B) Signal density plots of histone modifications at individual promoter-proximal or distal regions in each population. Signal densities were calculated for every discrete genomic region enriched with at least one histone modification in at least one population, then log2 transformed and normalized (see Experimental Procedures). x axis indicates H3K27me3 value and y axis indicates H3K4me2 value, while the color coding specifies the signal density of H3Ac in each element. In each stage, total regions were separated into four subgroups, based on the enrichment of H3K4me2 (≥4 RPM as positive) and H3K27me3 (≥2 RPM as positive). The percentage of regions for each group is indicated in the insert.

(C) Distribution of histone modifications at the promoters with respect to transcriptional activity. Based on the transcriptional activity, total Refseq annotated genes were divided into subgroups of expressed genes (in red table, normalized RNA-seq value ≥ 1 RPMK) and silent genes (in blue table, normalized RNA-seq value < 1 RPMK). Each number indicates the total amount of promoters enriched with the specified histone modification or combination of histone modifications. The percentage of promoters enriched with the specified histone modification or combination of histone modifications within the respective subgroup is given in the parenthesis.

<table>
<thead>
<tr>
<th>Table</th>
<th>FLDN1</th>
<th>FLDN2a</th>
<th>FLDN2b</th>
<th>ThyDN3</th>
<th>ThyDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>H3Ac+</td>
<td>8,469</td>
<td>8,406</td>
<td>8,603</td>
<td>8,357</td>
<td>7,948</td>
</tr>
<tr>
<td>H3K4me2+</td>
<td>9,010</td>
<td>8,913</td>
<td>9,002</td>
<td>8,963</td>
<td>8,758</td>
</tr>
<tr>
<td>H3K27me3+</td>
<td>366</td>
<td>296</td>
<td>437</td>
<td>371</td>
<td>494</td>
</tr>
<tr>
<td>H3Ac-</td>
<td>121</td>
<td>92</td>
<td>106</td>
<td>90(95%)</td>
<td>121(1.3%)</td>
</tr>
<tr>
<td>H3K4me2-</td>
<td>137(1.4%)</td>
<td>105(1.1%)</td>
<td>106(1.1%)</td>
<td>92(95%)</td>
<td>121(1.3%)</td>
</tr>
<tr>
<td>H3K27me3-</td>
<td>8,261</td>
<td>8,234</td>
<td>8,274</td>
<td>8,094</td>
<td>7,612</td>
</tr>
<tr>
<td>H3Ac+</td>
<td>57(1.6%)</td>
<td>56(1.6%)</td>
<td>50(1.6%)</td>
<td>57(1.6%)</td>
<td>51(1.6%)</td>
</tr>
<tr>
<td>H3K4me2+</td>
<td>179(1.8%)</td>
<td>159(1.7%)</td>
<td>274(2.8%)</td>
<td>256(2.6%)</td>
<td>332(3.5%)</td>
</tr>
<tr>
<td>H3K27me3+</td>
<td>695(7.0%)</td>
<td>573(6.0%)</td>
<td>761(7.7%)</td>
<td>712(7.3%)</td>
<td>699(7.4%)</td>
</tr>
</tbody>
</table>

* Total number of expressed genes (mRNA ≥ 1 RPKM) in each stage

* Total number of silent genes (mRNA < 1 RPKM) in each stage

Figure S2. Global Distribution of the Three Histone Modifications, Related to Figure 2
(A) Length distribution of discrete genomic regions enriched with the specified histone modification in at least one cell population. First, more regions were enriched with H3K4me2 in at least one cell population than H3Ac or H3K27me3. Second, while the majority of enriched regions were within 500–5,000 bp (>95% for H3Ac and H3K4me2; 81.5% for H3K27me3), H3K27me3 enriched regions were relatively more diverse in length, including some regions of much greater extent.

(B) Signal density plots of histone modifications at individual promoter-proximal or distal regions in each population. Signal densities were calculated for every discrete genomic region enriched with at least one histone modification in at least one population, then log2 transformed and normalized (see Experimental Procedures). x axis indicates H3K27me3 value and y axis indicates H3K4me2 value, while the color coding specifies the signal density of H3Ac in each element. In each stage, total regions were separated into four subgroups, based on the enrichment of H3K4me2 (≥4 RPM as positive) and H3K27me3 (≥2 RPM as positive). The percentage of regions for each group is indicated in the insert.

(C) Distribution of histone modifications at the promoters with respect to transcriptional activity. Based on the transcriptional activity, total Refseq annotated genes were divided into subgroups of expressed genes (in red table, normalized RNA-seq value ≥ 1 RPMK) and silent genes (in blue table, normalized RNA-seq value < 1 RPMK). Each number indicates the total amount of promoters enriched with the specified histone modification or combination of histone modifications. The percentage of promoters enriched with the specified histone modification or combination of histone modifications within the respective subgroup is given in the parenthesis.
Figure S3. Global Histone Modifications and Gene Expression Profiles at Both Promoter Regions and Distal Regions, Related to Figures 2 and 3

(A) The normalized signal densities of histone modifications at the promoter region were aligned with the normalized mRNA data for each of 20,861 Refseq annotated genes, and one-dimensional (along genes) hierarchical clustering was performed using Ward linkage and Euclidean distance.

(B) Each distal region was assigned to the closest gene, and one-dimensional hierarchical clustering, along genes, was performed as in Figure S3A.
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start (bp)</th>
<th>End (bp)</th>
<th>Log2 Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr9</td>
<td>44,769,390</td>
<td>44,853,731</td>
<td>0.02</td>
</tr>
<tr>
<td>chr2</td>
<td>11,534,869</td>
<td>11,616,723</td>
<td>0.02</td>
</tr>
<tr>
<td>chr5</td>
<td>148,102,020</td>
<td>148,236,170</td>
<td>0.02</td>
</tr>
<tr>
<td>chr11</td>
<td>52,044,043</td>
<td>52,138,103</td>
<td>0.02</td>
</tr>
</tbody>
</table>

**Figure Legends**

- **A**: mRNA expression levels for Notch1 gene.
- **B**: mRNA expression levels for Prhoxnb gene.
- **C**: mRNA expression levels for GATA-3 gene.
- **D**: mRNA expression levels for Zbtb7b gene.
- **E**: mRNA expression levels for Tcf7 gene.
- **F**: mRNA expression levels for E2A-/ gene.
- **G**: mRNA expression levels for DP gene.
- **H**: mRNA expression levels for AK136654 gene.
Figure S4. Dynamic Histone Modifications and Transcription Factor Binding Are Linked to Differentially Regulated Gene Expression, Related to Figures 4, 6, and 7

(A) UCSC genome browser tracks depicting H3K4me2 and gene expression at Notch1 locus. A Notch1 distal alternative promoter or cis-regulatory element (black arrow, most 5’ component of H3K4me2 peak) activity was upregulated at DN2b and DN3 stages.

(B) UCSC genome browser tracks depicting GATA-3 binding, together with binding associated H3K4me2 and gene expression, at Rag1/2 loci. Black arrows indicate the DP-specific Rag1/2 enhancer, where GATA-3 occupancy was observed at a low level in DN2b stage and sharply increased in DP stages, as did binding associated H3K4me2. Due to the strong upregulation of gene expression of Rag1 and Rag2 from DN3 to DP, note that in this panel the range of RNA-seq signal densities of Rag1 and Rag2 for DP (0.02 to 100 in red) is different from the one for other stages (0.02 to 16).

(C) UCSC genome browser tracks showing GATA-3 binding, together with binding associated H3K4me2 at Cd3e/d/g loci.

(D and E) UCSC genome browser tracks depicting histone modifications and gene expression at Il2ra (D) and Ebf1 (E) loci.

(F and G) UCSC genome browser tracks depicting PU.1 binding, together with binding associated histone modification(s) and gene expression (of Flt3), at Flt3 (F) and Hhex (G) loci. PU.1 occupancies, as well as binding associated H3K4me2, at the promoters of Flt3 and Hhex (black arrows) decreased from DN1 to DN2a stages, and completely disappeared in DN2b, in parallel with the gene expression pattern of two genes (Hhex expression and histone modification patterns are depicted in Figure 4E).

(H and I) UCSC genome browser tracks depicting GATA-3 binding, together with binding associated histone modifications and gene expression, at Tcf7 (H) and Zbtb7b (I) loci. GATA-3 occupancy at an upstream distal region of Tcf7 (black arrow) increased from DN1 to DN2b, as gene expression of Tcf7 was upregulated from DN1 to DN2b. Although repressed by H3K27me3 in DN2b and DP (pre-positive selection), Zbtb7b upstream distal region was constantly bound by GATA-3 (black arrow) in both stages. Note that the range of RNA-seq signal densities of Tcf7 for DP (0.02 to 80 in red) is different from the one for other stages (0.02 to 40).
Figure S5. K-Means Clustering for Differentially Expressed Genes, Related to Figures 2 and 6
Differentially expressed genes (3,697) were subjected to K-means clustering analysis that inferred to 25 differentially expressed patterns (see Experimental Procedures). Error bar represents the standard deviation of biological replicates or triplicates of individual genes at each stage. The genes in each cluster are listed in order in Table S4A.
**A**

Proximal Sites

- Downregulated Gene Sites
  - n = 1,034, p = 1.1e-07
- Upregulated Gene Sites
  - n = 9,695, p = 2.1e-09
- Silent Gene Sites
  - n = 1,589, p = 1.3e-17
- Downregulated Gene Sites
  - n = 1,034, p = 1.4e-19

**B**

Distant Sites

- Downregulated Gene Sites
  - n = 1,034, p = 2.6e-04
- Upregulated Gene Sites
  - n = 9,695, p = 2.1e-56
- Silent Gene Sites
  - n = 1,589, p = 2.6e-10
- Downregulated Gene Sites
  - n = 1,034, p = 4.9e-04

**C1**

FLDN2b vs. FLDN1 (log2(fold-change))

<table>
<thead>
<tr>
<th>log2(RPMK)</th>
<th>DN1</th>
<th>DN2a</th>
<th>DN2b</th>
<th>DN3</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.5</td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>3.5</td>
<td>0.7</td>
<td>0.9</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**C2**

PU.1 Binding Associated H3K4me2

Proximal Sites

- E2A-/- Specific
  - (1,120 sites)
- Stably Expressed Gene Sites
  - (6,207 sites)
- Upregulated Gene Sites
  - (9,695 sites)

**D**

Stably Expressed Gene Sites

- (1,426 sites)

ª 2012 Elsevier Inc.
Figure S6. PU.1 Occupancy Associated with Epigenetic Modifications and Gene Expression during Early T Cell Development, Related to Figures 5 and 6

(A) Cumulative distributions of changes in PU.1 occupancy between FLDN2b and FLDN1 among promoter-proximal sites (top) and promoter-distal sites (bottom). The total PU.1 binding sites in DN cells (as depicted in Figure 5B) were separated in two groups, promoter-proximal sites (≤1 kb from nearby TSS) and promoter-distal sites (>1 kb from nearby TSS). Based on the expression patterns of binding linked genes from FLDN1 to FLDN2b, each group was then divided into 4 subgroups: downregulated gene sites (linked to genes ≥ 2x downregulated, blue), upregulated gene sites (linked to genes ≥ 2x upregulated, red), stably expressed gene sites (linked to genes with < 2x change in expression, green) and silent gene sites (linked to genes with < 1 RPKM in both stages, black). Binding sites linked to downregulated genes, both distal and proximal, tend to lose PU.1 occupancy more rapidly than other groups of sites. K-S test was performed between stably expressed gene sites and each of the other three subgroups. The number of sites in each group and p value for each comparison are indicated in parentheses.

(B) Cumulative distributions of changes in PU.1 binding associated H3K4me2 between FLDN2b and FLDN1 among promoter-proximal sites (top) and promoter-distal sites (bottom). H3K4me2 signal densities were calculated within ±1 kb of the summit of a given PU.1 bound region (as depicted in Figure 5B). K-S tests were performed between stably expressed gene sites and each of the other three subgroups as described in S6A. The number of sites in each group and p value for each comparison are indicated in parentheses.

(C) 1 and 2. Heat maps of PU.1 occupancy and distribution of H3Ac, H3K4me2 and H3K27me3 surrounding ±2 kb of the binding summits for promoter-proximal regions (see Experimental Procedures). The PU.1 binding sites in DN cells were divided into four subgroups based on linked gene expression patterns as described in S6A. As comparison, a separate group of heat maps for promoter-proximal regions that were selected for much greater PU.1 binding in E2A⁻/⁻ cells than in early T lineage cells (as in Figure 5) are included.

(D) Heat maps of PU.1 occupancy at promoter-distal regions are shown as in C1&2, correlated with distribution of H3Ac, H3K4me2 and H3K27me3 surrounding ±2 kb of the binding summits (see Experimental Procedures). RNA expression heat maps refer to the nearest linked gene and a single gene can be represented by more than one PU.1-bound distal region. The PU.1 binding sites in DN cells were divided into four subgroups as described in panel A. As comparison, a separate group of heat maps for promoter-distal regions that were specific for PU.1 binding in E2A⁻/⁻ cells (as in Figure 5) are included.
Figure S7. Characterization of Sites of GATA-3 Binding in Early Developing T Cells, Related to Figure 7
(A) Comparisons of GATA-3 DNA binding site distributions in ThyDP (Tcrα−/− DP, using Santa Cruz sc-268 antibody) and CD3lo DP (using BD biosciences #558686 antibody) (Wei et al., 2011). Pearson’s correlation coefficient (r) is indicated.
(B) Cumulative distributions of H3K4me2 enrichment over genomic regions within ± 1 kb of PU.1 binding sites and GATA-3 binding sites in FLDN1 (top panels) and FLDN2b (bottom panels) cells. In each stage, positive binding sites (≥ 2 RPM of PU.1 or GATA-3 enrichment, see Tables S5 and S7, respectively) were divided into two groups, promoter-distal sites and promoter-proximal sites. Each group of binding sites was further divided into two subgroups based on the expression level of binding associated genes (expressed and silent). Since PU.1 tends to bind at multiple sites of a single gene locus, it is possible that binding sites with low or no histone modifications of a particular gene locus are nonfunctional. PU.1 binding sites with the highest H3K4me2 enrichment at each gene locus were selected (sites in promoter-distal regions and promoter-proximal regions were selected separately), and plotted accordingly.
(C) Cumulative distributions of H3K27me3 enrichment over genomic regions within ± 1 kb of PU.1 binding sites and GATA-3 binding sites in FLDN1 (top panels) and FLDN2b (bottom panels) cells. In each stage, positive binding sites were divided similarly as in Figure S7B.