Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation

Ageliki Tsagaratou[a,1], Tarmo Äijö[a,b,1], Chan-Wang J. Lio[a], Xiaojing Yue[a], Yun Huang[a,c], Steven E. Jacobsen[d,e,f], Harri Lähdesmäki[b], and Anjana Rao[a,c,g,2]

[a]Department of Signaling and Gene Expression, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92034; [b]Department of Information and Computer Science, Aalto University School of Science, FI-00076 Aalto, Finland; [c]Sanford Consortium for Regenerative Medicine, La Jolla, CA 92037; [d]Howard Hughes Medical Institute, [e]Department of Molecular, Cell, and Developmental Biology, and [f]Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, University of California, Los Angeles, CA 90095; and [g]Department of Pharmacology and Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093

Contributed by Anjana Rao, July 3, 2014 (sent for review May 18, 2014)

The discovery of Ten Eleven Translocation proteins, enzymes that oxidize 5-methylcytosine (5mC) in DNA, has revealed novel mechanisms for the regulation of DNA methylation. We have mapped 5-hydroxymethylcytosine (5hmC) at different stages of T-cell development in the thymus and T-cell differentiation in the periphery. We show that 5hmC is enriched in the gene body of highly expressed genes at all developmental stages and that its presence correlates positively with gene expression. Further emphasizing the connection with gene expression, we find that 5hmC is enriched in active thymus-specific enhancers and that genes encoding key transcriptional regulators display high intragenic 5hmC levels in precursor cells at those developmental stages where they exert a positive effect. Our data constitute a valuable resource that will facilitate detailed analysis of the role of 5hmC in T-cell development and differentiation.

thymic development | epigenetics

Cell lineage specification is established through epigenetic modifications of histones and DNA (1, 2). Until recently, the only known modified base in DNA was 5-methylcytosine (5mC), an epigenetic mark established by DNA methyltransferases (DNMTs) (3). 5mC represents 5–8% of total cytosine in different cell types and in somatic cells is almost exclusively found in the CpG sequence context (4–6). With the exception of CpG islands and high CpG promoters, which predominantly contain unmethylated CpGs (7), almost 90% of CpGs in the genome show high levels of cytosine methylation (>50%) (6). With respect to changes in DNA methylation, promoters with intermediate CpG density appear to be the most dynamic (5). At a global level, the lowest levels of CpG methylation are observed at the promoters of the most highly expressed genes; conversely, dense CpG methylation of promoters is generally associated with decreased gene expression (8). There is also dense DNA methylation in gene bodies, but the association of gene-body CpG methylation with transcriptional regulation is less clear (4, 8).

The discovery that Ten Eleven Translocation (TET) proteins are 5-methylcytosine oxidases added additional complexity to our understanding of DNA methylation (reviewed in refs. 9, 10). TET proteins are 2-oxoglutarate- and Fe (II)-dependent dioxygenases that catalyze the hydroxylation of 5mC to 5-hydroxymethylcytosine (5hmC) in DNA (11); further oxidation of 5hmC produces 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (12, 13). These oxidized methylcytosine (oxi-mC) species promote DNA demethylation through inhibition of maintenance DNA replication mediated by the DNMT1/UHRF1 complex; they can also mediate "active" (replication-independent) DNA demethylation through a process that involves excision of 5fC and 5caC by thymine DNA glycosylase (TDG), followed by replacement with an unmethylated cytosine through base excision repair (13–15; reviewed in refs. 9, 10, 16, 17).

Recent studies have addressed DNA methylation status either at a genome-wide level (18) or by focusing on the epigenetic regulation of gene expression at specific key regulatory loci, such as that encoding ThPOK (Th-inducing POZ-Kruppel factor) (19). However, treatment of DNA with bisulfite, the gold standard method used to evaluate the distribution of 5mC, cannot distinguish 5mC and 5hmC (20). As a result, regions of 5mC that have been perceived as unchanged potentially include 5hmC as well; in other words, genes that show differences in gene expression at different developmental stages but are perceived as stable in terms of 5mC distribution could in reality be very dynamic in terms of 5hmC.

In this study we took advantage of the fact that thymic development and peripheral T-cell differentiation are tractable systems for the study of dynamic changes in epigenetic modifications during cell differentiation and lineage specification. Because the thymus contains multiple cell types at different stages of T-cell lineage specification, isolation and analysis of individual cell types provides a snapshot of the transcriptional and epigenetic changes that mark specific developmental transitions. Moreover, the transcription factors that drive these processes of lineage specification have been extensively studied (21–24). We have mapped the genome-wide distribution of 5hmC in selected thymic and peripheral T cells and have analyzed its correlation with other

Significance

5-Hydroxymethylcytosine (5hmC) is an epigenetic DNA modification produced through the enzymatic activity of TET proteins. Here we present the first genome-wide mapping of 5hmC in T cells during sequential steps of lineage commitment in the thymus and the periphery (thymic DP, CD4 SP, and CD8 SP cells; peripheral naive CD8 and CD4 T cells; and in vitro-differentiated Th1 and Th2 cells). We show that 5hmC is enriched at gene bodies and cell type-specific enhancers, that its levels in the gene body correlate strongly with gene expression and histone modifications, and that its levels change dynamically during the course of T-cell development and differentiation. Our analysis will facilitate increased understanding of the role of 5hmC in T-cell development and differentiation.

Author contributions: A.T., H.L., and A.R. designed research; A.T., T.Ä., C.-W.J.L., X.Y., and Y.H. performed research; S.E.J. contributed new reagents/analytic tools; A.T., T.Ä., H.L., and A.R. analyzed data; and A.T. and A.R. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE59213).

[1]A.T. and T.Ä. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: arao@liai.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412327111/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1412327111

epigenetic marks. We focused on 5hmC because 5fC and 5caC are considerably less abundant than 5hmC (12, 13), although their abundance increases upon depletion of TDG (25, 26). We show that 5hmC is enriched in gene bodies of developing T cells and that its levels show a strong positive correlation with gene expression, active transcription by RNA polymerase II, and the histone marks characteristic of these processes: trimethyl histone 3 lysine 4 (H3K4me3) and trimethyl histone 3 lysine 36 (H3K36me3). Moreover, 5hmC is enriched at active thymus-specific enhancers, marked by dual H3K4 monomethyl (H3K4me1) and acetylated H3K27 (H3K27Ac) modifications, again emphasizing the correlation with active gene expression. For genes that are equivalently expressed in precursor cells as well as their differentiated progeny, we find that 5hmC is higher in gene bodies of the precursor cells [$CD4^+CD8^+$ double positive (DP) thymocytes, naive CD4, or CD8 T cells] compared with differentiated T helper (Th) 1 and Th2 cells, suggesting an association of 5hmC not only with gene expression but also with precursor status in a developmental pathway. Overall, our data constitute an important resource for future studies addressing the role of DNA modifications in regulating gene expression during T-cell lineage specification in a physiological or pathological context.

Results

Capturing 5hmC Distribution in Different T-Cell Subsets During Sequential Steps of Specification. We purified DP, CD4 SP, CD8 SP, naive CD4, and naive CD8 T cells and also differentiated naive CD4 T cells into Th1 and Th2 cells in vitro (*Experimental Procedures* and *SI Appendix*, Figs. S1 and S2). We used two different methods to map 5hmC. The first involves treatment of genomic DNA with sodium bisulfite, which reacts with 5hmC to yield the highly antigenic adduct cytosine-5-methylenesulfonate (CMS); CMS-containing DNA is then immunoprecipitated with a specific antiserum developed in our laboratory (CMS-IP) (27). The second method is similar to the glucosylation, periodate oxidation, and biotinylation (GLIB) method also developed in our laboratory (27); it involves glucosylation followed by specific biotinylation of 5hmC and subsequent isolation of biotinylated DNA fragments using streptavidin beads (28). Using DP thymocytes, which are abundant, to compare these two methods, we found significant overlap (*SI Appendix*, Fig. S3 *A–C*). To map 5*h*mC-*e*nriched *r*egions of the *g*enome (HERGs) in all seven selected cell types, we used CMS-IP followed by deep sequencing (Fig. 1*A* and *SI Appendix*, Tables S1 and S2; biological replicates are shown in *SI Appendix*, Fig. S3*D*).

5hmC Is Enriched in the Gene Body of Highly Expressed Genes and Its Presence Correlates with Markers of Active Transcription. Bioinformatic analysis showed that, in all cell types studied, 5hmC was enriched at promoters and intragenic regions (exons and introns) of annotated RefSeq genes relative to the representation of these regions in the mm9 mouse genome (Fig. 1*B* and *SI Appendix*, Table S5), although there was clear depletion of 5hmC near the transcription start site (TSS) (Fig. 1 *C–E*). There

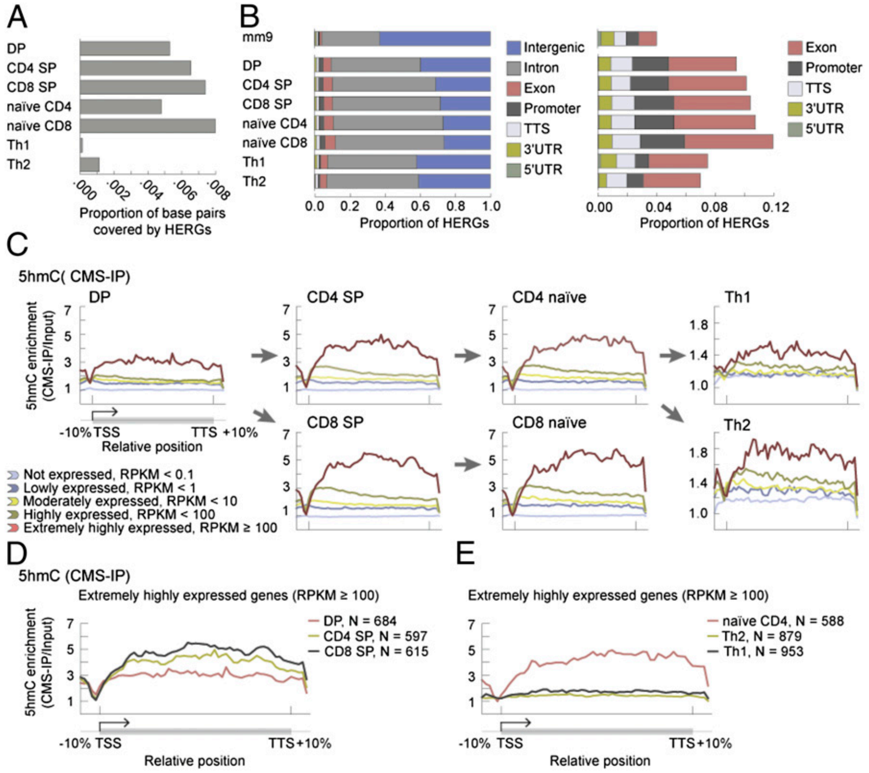

Fig. 1. 5hmC is enriched in the gene body of highly expressed genes. (*A*) Proportion of the mm9 genome covered by HERGs in each indicated cell type. (*B*) Proportions of HERGs that fall into annotated genomic regions for each indicated cell type. TTS indicates transcription termination site. A magnified view of promoters and intragenic regions (other than introns) is shown at right. Comparison with the top bar (representation of the annotated genomic regions in the mm9 reference genome) shows that 5hmC is enriched at promoters, exons, and introns. (*C*) Average 5hmC enrichment over the gene body, categorized based on gene expression levels in the indicated cell types (*SI Appendix*, Table S4). The arrows indicate differentiation pathways, from precursor to progeny cell type. (*D*) Average 5hmC enrichment over extremely highly expressed genes (RPKM ≥ 100) in DP, CD4 SP, and CD8 SP cells. Gene body 5hmC levels in highly expressed genes are higher in CD4 SP and CD8 SP cells compared with DP thymocytes. *E*, as in *D*, but here the average 5hmC enrichment profiles are calculated over extremely highly expressed genes (RPKM ≥ 100) in naive CD4, Th1, and Th2 cells. Even in highly expressed genes, gene-body 5hmC levels drop in Th1 and Th2 cells compared with their precursor naive CD4 T cells.

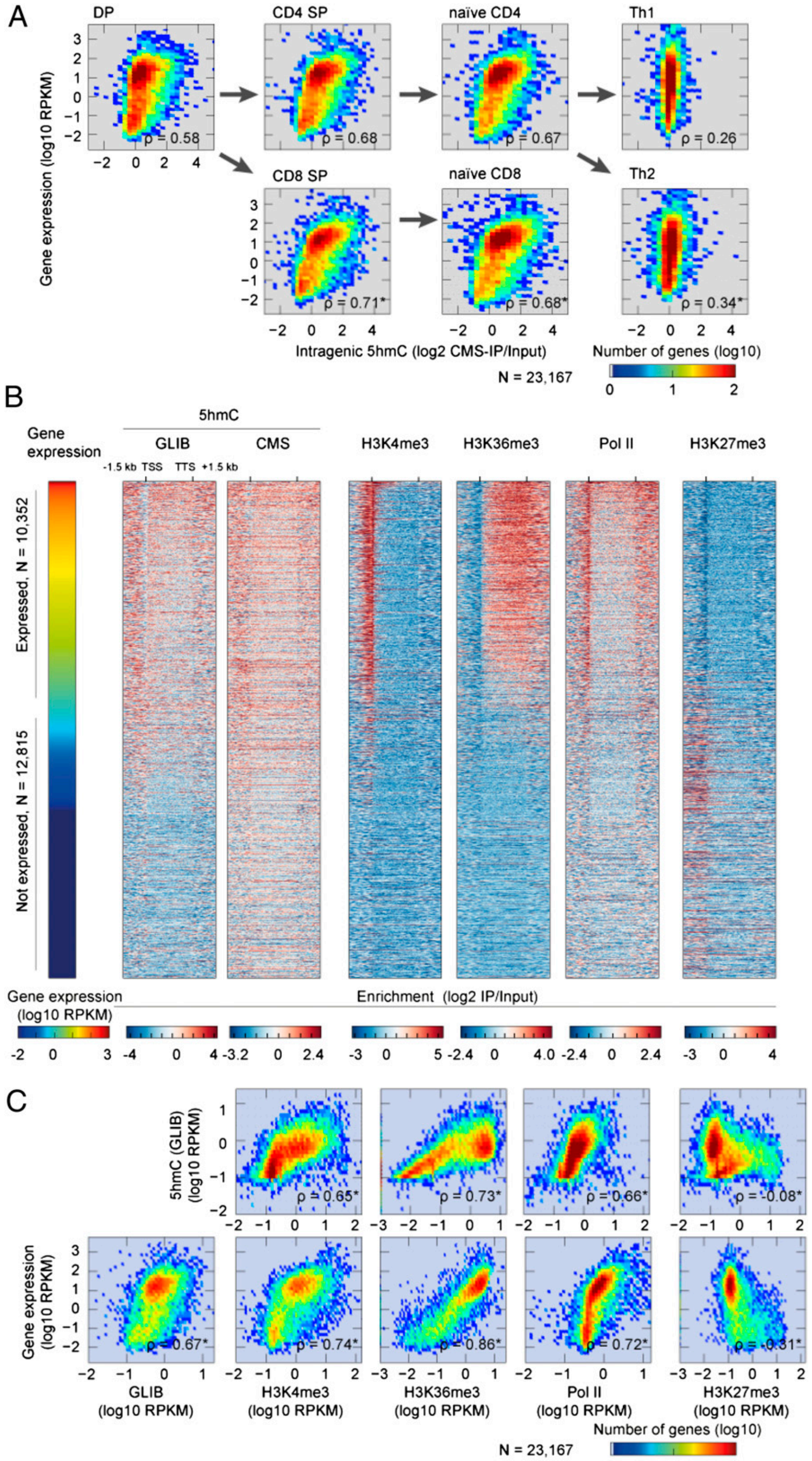

Fig. 2. Intragenic 5hmC levels correlate with RNA Pol II occupancy and with histone marks associated with active transcription. (*A*) Density plots depicting gene expression (log10 RPKM) and intragenic 5hmC (log2 CMS-IP/Input) in DP, CD4 SP, CD8 SP, naive CD4, naive CD8, and Th1 and Th2 cells. The Spearman correlation coefficient ρ is positive in all cases, reflecting the positive correlation of the two examined parameters; the weak correlation observed for Th1 and Th2 cells (*Right panels*) most likely reflects the low level of 5hmC in these cells. Arrows indicate the directions of lineage specification and differentiation. (*B*) Heat map depicting gene expression (log10 RPKM), intragenic 5hmC (log2 IP/Input) evaluated by GLIB and CMS-IP, and intragenic H3K4me3, H3K36me3, RNA polymerase II (Pol II), and H3K27me3 in DP T cells. Each row represents a gene, ordered from top to bottom by gene expression in DP cells. The cutoff RPKM of 1 was used to distinguish expressed and nonexpressed genes. (*C*) Density plots depicting the correlation of gene-body 5hmC (GLIB) (*y* axis, *Upper*) and gene expression (*y* axis, *Lower*) with gene-body H3K27me3, H3K4me3, H3K36me3 marks and RNA Pol II (Pol II). All values are represented as log10 RPKM; RPKM values are calculated from TSS to TTS. For each panel, the Spearman rank correlation coefficient ρ is shown (exact permutation test for testing the null hypothesis that there is no correlation, two tailed, $*P < 2.2 \times 10^{-16}$).

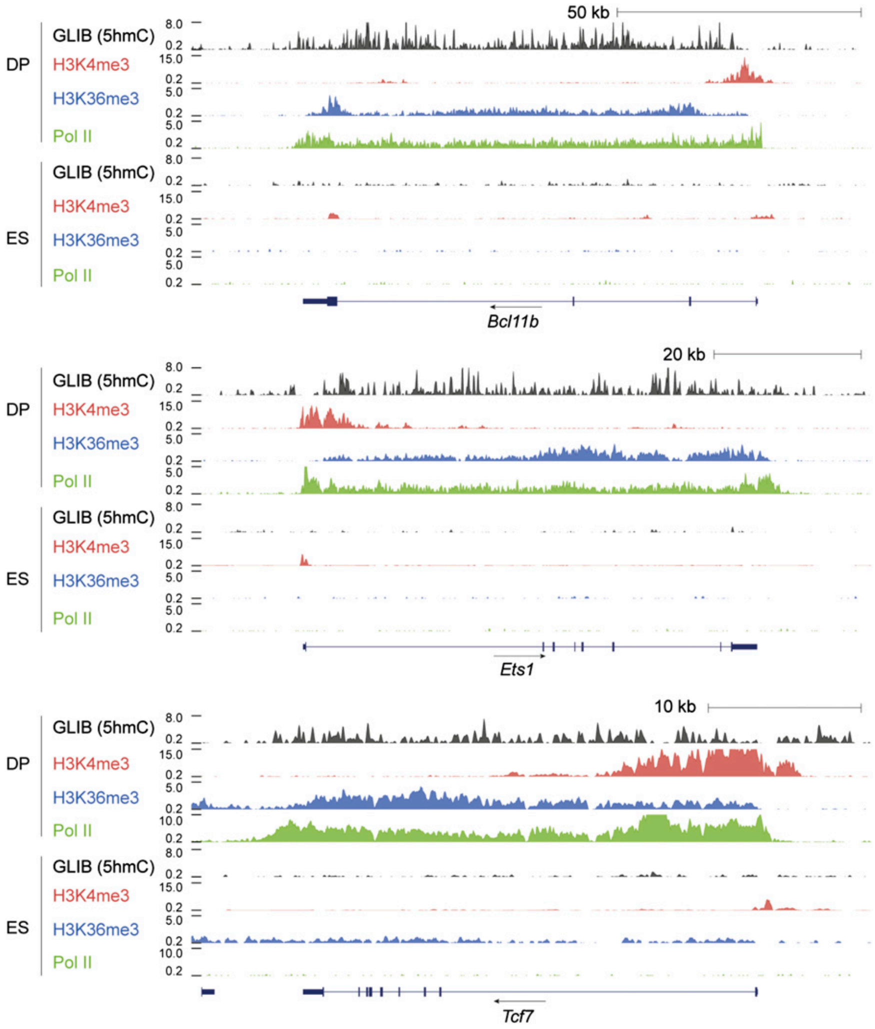

Fig. 3. Portraits of genes in DP and ES cells demonstrating the intragenic distribution of 5hmC and marks of active transcription. Genome browser views of the distribution of 5hmC (GLIB), H3K36me3, and RNA polymerase II (Pol II) around the *Bcl11b*, *Ets1*, and *Tcf7* genes in DP thymocytes and ES cells. 5hmC (GLIB) is shown in black, H3K4me3 in red, H3K36me3 in blue, and RNA polymerase II in green.

was no clear-cut correlation between gene-body 5hmC levels in a given cell type and the position of that cell type in a differentiation pathway: specifically, CD4 SP and CD8 SP T cells were enriched for gene-body 5hmC compared with their DP thymocyte precursors (Fig. 1 *C* and *D*), but terminally differentiated Th1 and Th2 cells showed greatly decreased gene-body 5hmC compared with their precursor naive CD4 T cells (Fig. 1 *C* and *E*). Interestingly, 5hmC levels remained relatively stable as judged by anti-CMS dot blot at the initial stages of differentiation (18 h after activation) but then showed progressive reduction during the expansion phase (36 h to 5 d) (*SI Appendix*, Fig. S4), suggesting a replication-dependent dilution of this modification.

There was a striking correlation between gene-body 5hmC levels and gene expression in T cells. 5hmC was particularly enriched in gene bodies of the most highly expressed genes [reads per kilobase per million mapped reads (RPKM) > 100, Fig. 1*C*], reminiscent of its distribution in neurons (29) and in embryonic stem (ES) cells (27, 30). The correlation remained strong even when all genes—both expressed and nonexpressed—were considered (Fig. 2*A*). In DP cells for which published data on histone modifications are available (31–33) (summarized in *SI Appendix*, Table S3), we compared gene-body 5hmC to the levels of RNA polymerase II (Pol II), H3K4me3, H3K27me3, and H3K36me3 (Fig. 2 *B* and *C*). Fig. 2*B* shows a heat map for individual genes with each line representing one gene (−1.5 kb upstream of the TSS to + 1.5 kb downstream of the TSS); the genes are ordered based on their expression levels. Fig. 2*C* shows the same data presented as a density plot indicating the correlation coefficient; each dot represents the averaged value for each modification at a single gene. In both representations, there is a clear positive correlation of gene-body 5hmC with Pol II, H3K4me3, and H3K36me3 [all markers of active transcription (1)] and an equally clear negative correlation with H3K27me3, a modification negatively correlated with gene expression (1) (Fig. 2 *B* and *C*).

The co-occurrence of gene-body 5hmC with Pol II and H3K36me3 across the gene body and H3K4me3 at the TSS is illustrated in Fig. 3 for *Bcl11b*, *Ets1*, and *Tcf7*, three genes highly expressed in DP T cells. In ES cells in which *Bcl11b* and *Ets1* are not expressed, these genes show greatly diminished peaks of Pol II and the epigenetic marks. In contrast, the *Tcf7* gene, which is expressed at low levels in ES cells, shows mild enrichment for 5hmC and H3K36me3 (Fig. 3).

5hmC Is Enriched in Active Tissue-Specific Enhancers. In a previous study of mouse ES cells (34), 5hmC was shown to be enriched at enhancers, which are distal DNA regulatory elements that contain transcription-factor–binding site and the majority of differentially methylated CpGs, and are important for regulating

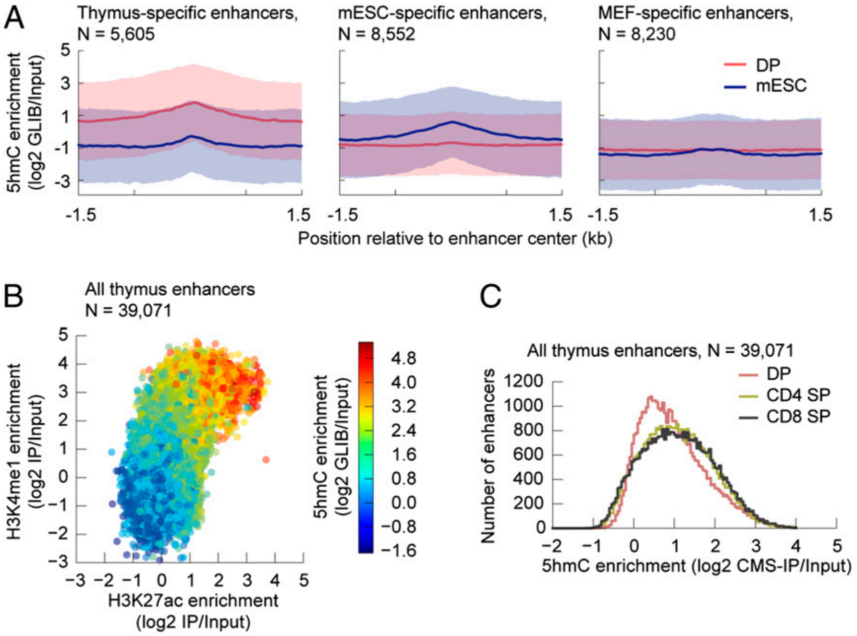

Fig. 4. 5hmC is enriched in active tissue-specific enhancers. (*A*) Average enrichment of 5hmC (log2 GLIB/Input) (*y* axis) from DP (red line) and mouse embryonic stem cells (blue line) in thymus-specific (*n* = 5,605) (*Left*), mESC-specific (*n* = 8,552) (*Center*), and MEF-specific (*n* = 8,230) (*Right*) enhancers. The enrichment is quantified ±1.5 kb of the reported enhancer center positions (35). The shaded regions depict the SDs of the means. (*B*) Scatter density plot evaluating the coenrichment of 5hmC, H3K4me1, and H3K27ac in all thymus enhancers (*n* = 39,071). The enrichments are quantified over the regions defined as ±1.5 kb of the reported enhancer center positions (35). (*C*) Histogram showing enrichment of 5hmC in thymic enhancers during the specification of DP T cells to CD4 and CD8 SP T cells. The enrichments are quantified over the regions defined as ±1.5 kb of the reported enhancer center positions (35).

tissue-specific expression patterns during mammalian development (18, 35). In DP T cells, we found that 5hmC was enriched at enhancers marked with H3K4me1 and p300 (35) that are specifically active in the thymus and depleted from enhancers specifically active in ES cells (Fig. 4*A*, *Left* and *Center*). In ES cells, the pattern was reversed: 5hmC was enriched in ES-specific enhancers and not in thymus-specific enhancers (Fig. 4*A*, *Left* and *Center*). As further verification, DP 5hmC and ES 5hmC were not enriched at mouse embryonic fibroblast (MEF)-specific enhancers (Fig. 4*A*, *Right*).

Enhancers are marked by H3K4me1 and have been classified into poised and active enhancers based on the absence or presence of acetylated H3K27 (H3K27ac) (35–39). We observed that active enhancers, defined by the co-occurrence of H3K4me1 and H3K27Ac, showed greater enrichment for 5hmC than poised enhancers marked by H3K4me1 but not H3K27Ac (Fig. 4*B*). Moreover, there is a clear enrichment of 5hmC at thymus-specific enhancers as DP thymocytes progress to the CD4 and CD8 lineages (Fig. 4*C*).

5hmC Exhibits Dynamic Changes During T-Cell Development and Lineage Specification. We compared 5hmC distribution in gene bodies in five of the seven T-cell subsets (excluding terminally differentiated Th1 and Th2 cells that have very little 5hmC). Clustering genes by gene-body 5hmC, we were able to identify several gene clusters whose patterns of 5hmC distribution changed during T-cell lineage commitment (Fig. 5*A* and *SI Appendix*, Table S6). In the heat map of Fig. 5*A*, each line represents a gene, for which intragenic 5hmC [measured from TSS to transcription termination site (TTS)] is shown during sequential stages of T-cell lineage specification. The data confirm, at a gene-by-gene level, that gene-body 5hmC levels change dynamically during the course of T-cell differentiation. An enlarged view of cluster 2 is shown in Fig. 5*A*, *Right*, and a few genes with important roles in thymic development are highlighted here and shown individually in Fig. 5*B*. Comparing cells related by a single step of differentiation (i.e., DP with CD4 SP and CD8 SP), we found that with very few exceptions, genes that increased in expression during the differentiation step also showed increased gene-body 5hmC and vice versa (Fig. 5*C* and *SI Appendix*, Table S7). As expected from this strong correlation of gene-body 5hmC with gene expression (also shown in Figs. 1–3), Gene Ontology analysis confirmed that the genes for which 5hmC levels are altered during differentiation fall into functional categories involved in immunological function (*SI Appendix*, Fig. S5 and Table S8).

The correlation between gene expression and 5hmC was also generally borne out when we examined genes encoding key transcription factors that were expressed or silenced during T-cell development (Figs. 6 and 7). For example, 5hmC was observed at highest levels across the gene body of the *Zbtb7b* gene in CD4 SP and naive CD4$^+$ T cells, the cells in which *Zbtb7b* [encoding ThPOK, a lineage-determining factor for CD4 T cells (40)] is most highly expressed (Fig. 6, *Top*). In contrast, the genes encoding the lineage-specifying factors Gata3 (GATA-binding protein 3) and Runx3 (runt-related transcription factor 3) showed a different behavior, reflecting the fact that they act not only in the thymus but also in the periphery. Gata3 is necessary for the DP-to-SP transition (40), for Th2 differentiation (22, 24) and for CD8$^+$ T-cell homeostasis and proliferation (41, 42). Levels of 5hmC over the *Gata3* gene were high, as expected, in CD4 SP cells and Th2 cells, which have high *Gata3* expression, but were also high in naive CD4 T cells, the immediate precursors of Th2 cells, and in naive CD8$^+$ T cells, both of which show much lower *Gata3* gene expression (Fig. 6, *Middle*). Finally Runx3, a lineage-specifying factor for CD8$^+$ cytolytic effector T cells (43), is highly expressed in CD8 SP cells but poorly expressed in naive CD8$^+$ T cells; however, gene-body 5hmC levels were comparably high in these two cell types (Fig. 6, *Bottom*), perhaps because naive CD8$^+$ T cells are the immediate precursors of cytolytic effector T cells that use Runx3 to control transcription of several effector proteins (43).

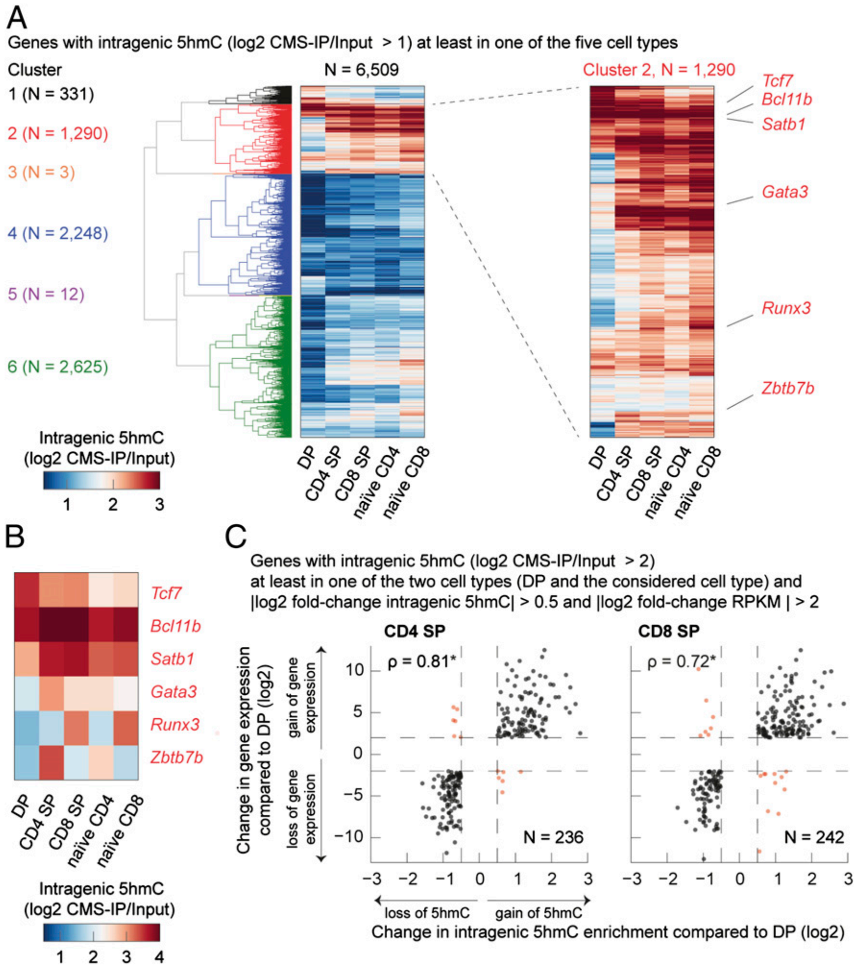

Fig. 5. 5hmC changes dynamically over the course of T-cell lineage specification. (*A*) Heat map representation of 6,509 genes hierarchically clustered according to their 5hmC patterns in gene sets (*Experimental Procedures*). Genes (≥1 kb) with intragenic 5hmC (log2 CMS-IP/Input > 1) in at least one of the five studied subsets (DP, CD4 SP, CD8 SP, naive CD4, naive CD8) were considered in the clustering analysis. The six identified clusters are numbered and differently colored, and the number of genes per cluster is shown on the left. Six transcriptional regulators in the second cluster that exert a significant role in shaping T-cell identity are indicated (*Right*). (*B*) Close-up view of the six transcriptional regulators (*Tcf7*, *Bcl11b*, *Satb1*, *Gata3*, *Runx3*, *Zbtb7b*) highlighted in *A* and clearly showing the differential intragenic distribution of 5hmC in these different T-cell subsets. (*C*) Scatter plots depicting the change in gene expression (*y* axis) versus the change in intragenic 5hmC (*x* axis) in CD4 SP and DP cells (*Left*, $n = 236$) and CD8 SP and DP cells (*Right*, $n = 242$). Genes with intragenic 5hmC (log2 CMS-IP/Input > 2) in at least one of the two cell types examined in each plot were considered; in addition, the logarithmic fold change of 5hmC (5hmC log2 fold change) greater than 0.5 and the logarithmic fold change in gene expression (RPKM log2 fold change) greater than 2 were required. Black dots depict genes that show the same direction of change in gene expression and intragenic 5hmC, whereas red dots indicate genes that show opposite changes. The depicted genes are listed in *SI Appendix*, Table S7. In each plot, the Spearman rank correlation coefficient ρ is shown (exact permutation test for testing the null hypothesis that there is no correlation, two tailed, $*P < 2.2 \times 10^{-16}$).

Notably, at many genes encoding important regulators of T-cell biology such as *Tcf7*, *Ets1*, *Satb1*, and *Itk*, the distribution of 5hmC across promoters and gene bodies seems to reflect not only current gene expression level but also regulatory history on a multiple-day timescale (Fig. 7). Specifically, there is a stronger enrichment of 5hmC at promoter-proximal and 5′-end regions surrounding the TSS in the precursor state (DP stage), but 5hmC depletion in the same region during subsequent developmental stages (Fig. 7).

We also examined genes encoding transcription factors that determine cell fate in non–T-cell lineages and therefore are silenced in T cells (*SI Appendix*, Fig. S6). The *Cebpa* gene, encoding a key factor in myeloid development, showed no hydroxymethylation in T cells (*SI Appendix*, Fig. S6, *Top*; note scale of 5hmC graph). Moreover, *Gata1* and *Pax5*, encoding key factors in erythroid and B-cell development, respectively, exhibited very low or undetectable levels of 5hmC (*SI Appendix*, Fig. S6, *Middle* and *Bottom*; note scale of 5hmC graph).

Discussion

By taking genomic-wide snapshots of 5hmC distribution in thymocytes and peripheral T cells at several developmental stages, we have effectively performed a kinetic analysis of the dynamic changes in these modifications during thymic development and peripheral Th1/Th2 differentiation. Our results make several key points: (*i*) intragenic 5hmC levels are high in gene bodies of highly expressed genes, and correlate significantly with marks of active transcription; (*ii*) 5hmC is enriched in active thymus-specific enhancers, again emphasizing the correlation with gene expression; and (*iii*) genes encoding key transcription factors that positively regulate a specific developmental transition have high 5hmC in their gene bodies, particularly in the precursor cells for that developmental stage. Together, these data provide important insights into the potential biological roles of 5hmC in regulating gene expression during normal T-cell lineage specification.

One of our most striking findings is the pronounced correlation between intragenic 5hmC levels, gene expression, Pol II

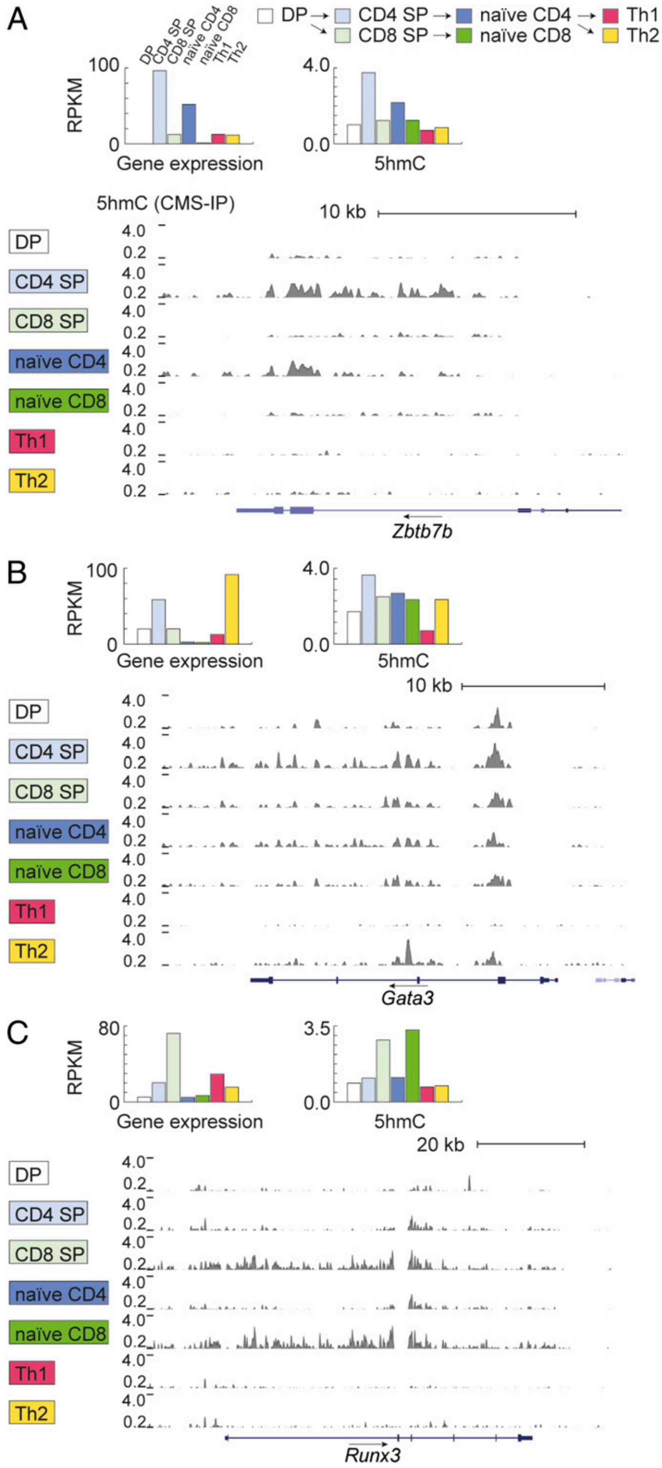

Fig. 6. Portraits of intragenic 5hmC in T-cell lineage-specifying transcription factors. (*Left*) Genome browser views of 5hmC (evaluated by CMS-IP, gray) in (*Top*) the *Zbtb7b* gene, encoding the transcription factor ThPOK that regulates the CD4 lineage; (*Middle*) *Gata3*; and (*Bottom*) *Runx3*, the lineage-determining factor of the CD8 lineage. A black arrow above the gene symbol indicates the direction of transcription. Bar graphs depict RPKM values evaluated over the gene body (*y* axis) for gene expression and 5hmC in each of seven cell subtypes (DP in pink, CD4 SP in white, CD8 SP in purple, naive CD4 in gray-blue, naive CD8 in yellow, Th1 in green, and Th2 in bright blue).

occupancy, and the H3K4me3 and H3K27me3 modifications that have an established association with active gene expression (1). Focusing on cell types related by developmental transitions, there was a clear dynamic modulation of the 5hmC epigenetic mark in gene bodies during T-cell differentiation. Gain of gene expression during lineage commitment correlated well with gain of 5hmC; conversely, decreased gene expression correlated with loss of intragenic 5hmC. This correlation of gene-body 5hmC with active gene transcription is in striking contrast to the inverse correlation established for DNA methylation at promoters, with inactive and active promoters marked by high and low levels of 5mC, respectively (8).

5hmC enrichment over the gene body of highly expressed genes has also been noted in mouse ES cells (30), neuronal cells/tissues (29), and differentiating sperm cells (44) (reviewed in ref. 9). Despite this intense research, the functional importance of gene-body 5hmC is not yet clear in any system. Does intragenic 5hmC facilitate gene expression, or is it passively deposited as a consequence of high-level gene expression, for example, through the association of TET2 or TET3 with elongating RNA Pol II via the SET/COMPASS complex (45). In mouse ES cells, Tet2 predominantly controls intragenic 5hmC levels (30); however, Tet2-deficient mice are viable and fertile and exhibit only mild immunological and hematological phenotypes (46). Because both Tet2 and Tet3 are expressed at high levels in thymocytes and T cells (47, 48), studies of T cells doubly deficient in both *Tet2* and *Tet3* will be needed to determine whether gene-body 5hmC facilitates transcriptional elongation by RNA polymerase II or is merely deposited in a passive manner during transcriptional elongation.

Examining H3K4me1-marked enhancers in five different thymic cell types, we found that 5hmC was highest at active enhancers marked by H3K4me1 as well as H3K27ac (38, 39), intermediate at "poised" enhancers marked by enrichment for H3K4me1 alone, and lowest at inactive enhancers not bearing either modification in a given cell type. Once again, these data showcase the positive correlation of 5hmC with actively transcribed genes. Analyzing cell types related by a single developmental transition, we found that 5hmC is enriched at thymus-specific enhancers during the DP → CD4 SP and DP → CD8 SP lineage commitment steps. Thus, at both gene bodies and distal regulatory elements, 5hmC enrichment is a marker of transcriptional activity and gene expression. 5hmC may facilitate long-range interactions between enhancers and other regulatory regions that are dynamically modulated during T-cell development or may be involved in recruiting or excluding transcriptional regulators that in turn modulate the expression of enhancer target genes. Another scenario is that 5hmC is passively deposited at enhancers by TET proteins that are associated with RNA polymerase II molecules engaged in transcribing enhancer RNA. Additional studies will be needed to distinguish these possibilities for 5hmC and other oxi-mC marks at distal enhancers.

Even though the oxi-mC species produced by TET proteins are intermediates in DNA demethylation, we have not mapped changes in DNA methylation in the T-cell subsets that we have studied. In part the reason is technical: the available methods for precipitating 5mC-containing DNA are very density-dependent and so cannot be compared directly with 5hmC mapping by CMS-IP. Although this problem could in theory be resolved by using single-base resolution methods for mapping 5mC (whole-genome bisulfite sequencing, or WGBS) and 5hmC (TAB-seq, oxBS-seq) (reviewed in ref. 48), these methods are very inefficient at capturing dynamic changes: only 21.8% of autosomal CpGs in ES cells showed differential methylation when 30 different human cell types and tissues were compared by WGBS (18), emphasizing the expense and inefficiency of accumulating reads at CpGs unlikely to show dynamic changes during thymic development. A second reason is that the functions of TET proteins likely extend beyond DNA demethylation to a more direct involvement in the epigenetic regulation of gene transcription. Specifically, all three oxi-mC species—5hmC, 5fC, and 5caC—are likely to function as epigenetic modifications that

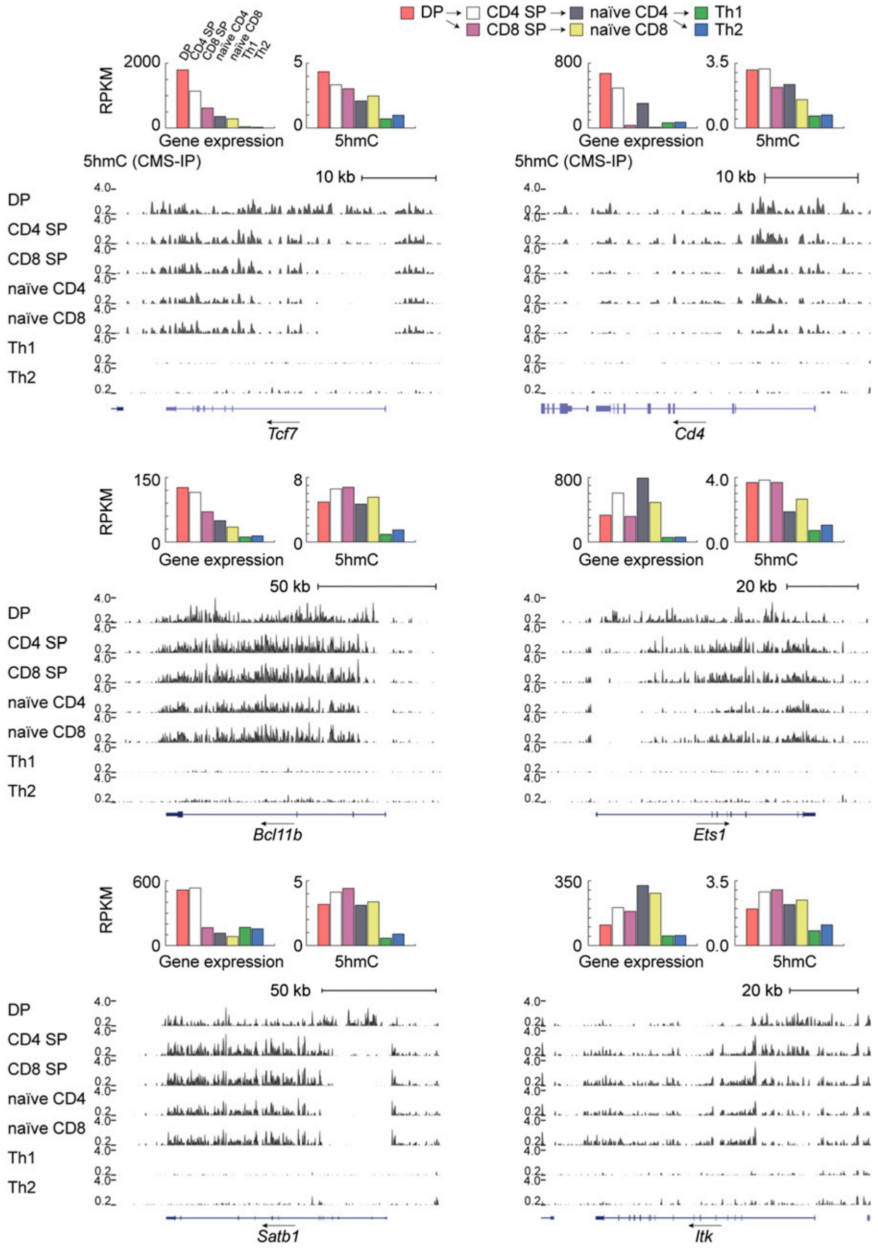

Fig. 7. Genome browser views of 5hmC distribution at additional genes with key roles in T-cell biology. Arrows show the direction of transcription. Bar graphs depict RPKM values summed over the gene body (TSS to TTS, *y* axis) for gene expression and 5hmC in each of the seven cell subtypes. (*Left*) The genes encoding the transcription factors Tcf7 and Bcl11b and the chromatin-associated protein Satb1 are most highly expressed in DP T cells, and their expression is considerably diminished in progeny CD4 and CD8 SP cells; notably, however, 5hmC levels either diminish more slowly (*Tcf7*) or increase slightly (*Bcl11b*, *Satb1*) from DP to SP T cells. (*Right*) Genes encoding proteins implicated in T-cell receptor signaling during T-cell development also show gene-body 5hmC at early but not developmental stages that is almost extinguished in terminally differentiated Th1 and Th2 despite continuing expression (*Cd4*, *Ets1*, *Itk*).

affect chromatin conformation and gene expression by recruiting "reader" proteins that recognize these modifications (29, 49, 50). 5hmC is a particularly stable and abundant mark, comprising ~5–10, 40, and ~1% of total 5mC in ES cells, Purkinje neurons, and immune cells, respectively (47, 51). In consequence, the relation between 5mC and 5hmC is not a simple one: 5mC is a substrate for 5hmC, but any single TET protein contributes only partially to the generation of 5hmC (30), and the existence of three distinct oxi-mC modifications adds to the complex relationship of 5mC to oxi-mC (reviewed in ref. 9).

In conclusion, we have mapped the distribution of 5hmC, an important epigenetic modification, in developing and differentiating T cells. We show that 5hmC is a reliable marker for actively transcribed genes and for active enhancers, especially in cells that can be considered precursors in a specific developmental transition. By providing a comprehensive map of 5hmC during the commitment and specification of DP cells toward the CD4 and CD8 lineages, our study sets the stage for a systematic dissection of the function of 5hmC in T-cell development and function, expands our understanding of the T-cell epigenome, and introduces an additional parameter that will need to be taken into consideration in future studies of gene expression in T cells and other immune cells.

Experimental Procedures

Enrichment-Based Detection of 5hmC. We used the abundant DP thymocytes to compare two methods for genome-wide 5hmC mapping: the CMS-IP method developed in our laboratory (52) and a method similar to GLIB (53) for specific biotinylation of 5hmC and subsequent isolation of biotinylated DNA fragments using streptavidin beads (28). The two methods showed significant overlap (*SI Appendix*, Fig. S3 *A–C*). For the other six cell subtypes, we chose the CMS-IP method because it can also provide insight into the methylation status of individual cytosines in the immunoprecipitated DNA fragments.

Isolation of Genomic DNA. Genomic DNA was prepared by lysing T cells, followed by treatment with RNase A (Qiagen) for 1 h at 37 °C and then by overnight treatment with proteinase K (Roche) at 55 °C with vigorous shaking as previously described (52). DNA was purified after sequential treatment with phenol, phenol/chlorophorm/isoamyl alcohol, chlorophorm/isoamyl alcohol (all obtained from Sigma), and then ethanol (Sigma) precipitation. DNA was resuspended in 10 mM Tris·Cl, pH 8.0, 0.1 mM EDTA and sheared to an average size of 165 bp by Adaptive Focused Acoustics using a Covaris S2 instrument.

Library Preparation of 5hmC-Enriched Genomic DNA. For CMS-IP samples, genomic DNA was spiked with cl857 *Sam7* λDNA (Promega) at a ratio of 200:1. End Repair was performed using the End Repair kit by Epicentre. A-tailing was performed using the large Klenow fragment enzyme by NEB. To perform the CMS-IP method, we used the TruSeq indexed adapters by Illumina because all of the Cs are methylated, making them compatible for use with bisulfite treatment. After adaptor ligation (usinq Quick Ligase by NEB), two rounds of Ampure Beads (Beckman Coulter) clean-up were performed to ensure removal of unligated adaptors. This step is crucial for this method before bisulfite conversion because accurate quantitation of the DNA is required to convert the optimum amount of DNA and ensure the successful outcome of the reaction. Therefore, less than 450 ng of genomic DNA was bisulfite-treated using the Methylcode kit by Invitrogen. In total, around 4 μg of genomic DNA were bisulfite-treated per cell subset. After bisulfite treatment, we quantified the DNA. A total of 2.5 μg was used for the CMS-IP whereas 25 ng were kept as input. The process of CMS-IP has been described in detail elsewhere (27, 52). The amplification of the IP sample as well as of the input was performed using the uracil-insensitive Kapa HiFi HotStart Uracil+. Two rounds of Ampure Beads purification were performed to ensure full removal of primer dimers. The samples were sequenced using the Illumina HiSeq platform. Using CMS-IP, we mapped HERGs in the seven selected cell types. The number of properly paired reads for both methods in each cell type is shown in *SI Appendix*, Table S1, and the overlap between biological replicates in *SI Appendix*, Fig. S3C. Deep sequencing to achieve saturation was performed (*SI Appendix*, Supplementary Methods Fig. 1). To evaluate bisulfite conversion efficiency, we spiked unmethylated λ-phage DNA into our samples and measured the extent of C > T conversion, which reflects the efficiency of conversion of unmethylated cytosine to uracil after bisulfite treatment. The conversion efficiency was >99.8% in all cases (*SI Appendix*, Table S2).

Further details about the experimental procedures and the data analysis are presented in *SI Appendix*.

ACKNOWLEDGMENTS. We thank the staff of the Flow Cytometry Facility of the La Jolla Institute; Cheryl Kim, Kurt van Gunst, Anthony Jose, and Lara Nosworthy for help with cell sorting; Dr. Suhua Feng for assistance in next-generation sequencing using the Illumina platform and for comments; and Dr. Matthew E. Pipkin, Dr. Gregory Seumois, and Jeremy Day for advice and assistance in next-generation sequencing in the La Jolla Institute for Allergy and Immunology Sequencing Facility. This study was supported by National Institutes of Health (NIH) Grant R01 AI44432; Grant RM1-01729 from the California Institute of Regenerative Medicine; Grant 6187-12 from the Leukemia and Lymphoma Society Translational Research Program (to A.R.); and a grant from the Academy of Finland Centre of Excellence in Molecular Systems Immunology and Physiology Research (to H.L.). Additional funds for next-generation sequencing were provided by NIH Grant R01 HL114093 (to A.R., Dr. Bjoern Peters, and Dr. Pandurangan Vijayanand, La Jolla Institute for Allergy and Immunology). A.T. and C.-W.J.L. are supported by Irvington Postdoctoral Fellowships from the Cancer Research Institute; Y.H. is supported by a Postdoctoral Fellowship from the Leukemia and Lymphoma Society; and T.Ä. is supported by a graduate student fellowship from the Finnish Doctoral Program in Computational Sciences. S.E.J. is an Investigator of the Howard Hughes Medical Institute.

1. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 12(1):7–18.
2. Smith ZD, Meissner A (2013) DNA methylation: Roles in mammalian development. *Nat Rev Genet* 14(3):204–220.
3. Ooi SK, O'Donnell AH, Bestor TH (2009) Mammalian cytosine methylation at a glance. *J Cell Sci* 122(Pt 16):2787–2791.
4. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.
5. Suzuki MM, Bird A (2008) DNA methylation landscapes: Provocative insights from epigenomics. *Nat Rev Genet* 9(6):465–476.
6. Stadler MB, et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480(7378):490–495.
7. Deaton AM, Bird A (2011) CpG islands and the regulation of transcription. *Genes Dev* 25(10):1010–1022.
8. Laurent L, et al. (2010) Dynamic changes in the human methylome during differentiation. *Genome Res* 20(3):320–331.
9. Pastor WA, Aravind L, Rao A (2013) TETonic shift: Biological roles of TET proteins in DNA demethylation and transcription. *Nat Rev Mol Cell Biol* 14(6):341–356.
10. Wu H, Zhang Y (2014) Reversing DNA methylation: Mechanisms, genomics, and biological functions. *Cell* 156(1-2):45–68.
11. Tahiliani M, et al. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324(5929):930–935.
12. Ito S, et al. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 333(6047):1300–1303.
13. He YF, et al. (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* 333(6047):1303–1307.
14. Maiti A, Drohat AC (2011) Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *J Biol Chem* 286(41):35334–35338.
15. Zhang L, et al. (2012) Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nat Chem Biol* 8(4):328–330.
16. Koh KP, Rao A (2013) DNA methylation and methylcytosine oxidation in cell fate decisions. *Curr Opin Cell Biol* 25(2):152–161.
17. Kohli RM, Zhang Y (2013) TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* 502(7472):472–479.
18. Ziller MJ, et al. (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500(7463):477–481.
19. Tanaka H, et al. (2013) Epigenetic Thpok silencing limits the time window to choose CD4(+) helper-lineage fate in the thymus. *EMBO J* 32(8):1183–1194.
20. Huang Y, et al. (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* 5(1):e8888.
21. Carpenter AC, Bosselut R (2010) Decision checkpoints in the thymus. *Nat Immunol* 11(8):666–673.
22. Kanno Y, Vahedi G, Hirahara K, Singleton K, O'Shea JJ (2012) Transcriptional and epigenetic control of T helper cell specification: Molecular mechanisms underlying commitment and plasticity. *Annu Rev Immunol* 30:707–731.
23. Rothenberg EV, Taghon T (2005) Molecular genetics of T cell development. *Annu Rev Immunol* 23:601–649.
24. Zhu J, Yamane H, Paul WE (2010) Differentiation of effector CD4 T cell populations (*). *Annu Rev Immunol* 28:445–489.
25. Shen L, et al. (2013) Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* 153(3):692–706.
26. Song CX, et al. (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* 153(3):678–691.
27. Pastor WA, et al. (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* 473(7347):394–397.
28. Song CX, et al. (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat Biotechnol* 29(1):68–72.
29. Mellén M, Ayata P, Dewell S, Kriaucionis S, Heintz N (2012) MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* 151(7):1417–1430.
30. Huang Y, et al. (2014) Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proc Natl Acad Sci USA* 111(4):1361–1366.
31. Wei G, et al. (2011) Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. *Immunity* 35(2):299–311.
32. Zhang JA, Mortazavi A, Williams BA, Wold BJ, Rothenberg EV (2012) Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity. *Cell* 149(2):467–482.
33. Zhang J, et al. (2012) Harnessing of the nucleosome-remodeling-deacetylase complex controls lymphocyte development and prevents leukemogenesis. *Nat Immunol* 13(1):86–94.
34. Yu M, et al. (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149(6):1368–1380.
35. Shen Y, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature* 488(7409):116–120.
36. Heintzman ND, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459(7243):108–112.
37. Heintzman ND, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39(3):311–318.
38. Creyghton MP, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 107(50):21931–21936.
39. Rada-Iglesias A, et al. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470(7333):279–283.
40. Wang L, et al. (2008) Distinct functions for the transcription factors GATA-3 and ThPOK during intrathymic differentiation of CD4(+) T cells. *Nat Immunol* 9(10):1122–1130.
41. Wang Y, et al. (2013) GATA-3 controls the maintenance and proliferation of T cells downstream of TCR and cytokine signaling. *Nat Immunol* 14(7):714–722.

42. Tai TS, Pai SY, Ho IC (2013) GATA-3 regulates the homeostasis and activation of CD8+ T cells. *J Immunol* 190(1):428–437.
43. Cruz-Guilloty F, et al. (2009) Runx3 and T-box proteins cooperate to establish the transcriptional program of effector CTLs. *J Exp Med* 206(1):51–59.
44. Gan H, et al. (2013) Dynamics of 5-hydroxymethylcytosine during mouse spermatogenesis. *Nat Commun* 4:1995.
45. Deplus R, et al. (2013) TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. *EMBO J* 35(5):645–655.
46. Ko M, et al. (2011) Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proc Natl Acad Sci USA* 108(35):14566–14571.
47. Ko M, et al. (2010) Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* 468(7325):839–843.
48. Tsagaratou A, Rao A (2013) TET proteins and 5-methylcytosine oxidation in the immune system. *Cold Spring Harb Symp Quant Biol* 78:1–10.
49. Spruijt CG, et al. (2013) Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. *Cell* 152(5):1146–1159.
50. Iurlaro M, et al. (2013) A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. *Genome Biol* 14(10):R119.
51. Kriaucionis S, Heintz N (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* 324(5929):929–930.
52. Huang Y, Pastor WA, Zepeda-Martínez JA, Rao A (2012) The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nat Protoc* 7(10):1897–1908.
53. Pastor WA, Huang Y, Henderson HR, Agarwal S, Rao A (2012) The GLIB technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nat Protoc* 7(10):1909–1917.