A unique H3K4me2 profile marks tissue-specific gene regulation

Aleksandra Pekowska, Touati Benoukraf, Pierre Ferrier, et al.

Genome Res. published online September 14, 2010
Access the most recent version at doi:10.1101/gr.109389.110

A unique H3K4me2 profile marks tissue-specific gene regulation

Aleksandra Pekowska,[1,2,3] Touati Benoukraf,[1,2,3] Pierre Ferrier,[1,2,3,4] and Salvatore Spicuglia[1,2,3,4]

[1]*Centre d'Immunologie de Marseille-Luminy, Université Aix Marseille, Marseille 13009, France;* [2]*CNRS, UMR6102, Marseille 13009, France;* [3]*Inserm, U631, Marseille 13009, France*

Characterization of the epigenetic landscape fundamentally contributes toward deciphering the regulatory mechanisms that govern gene expression. However, despite an increasing flow of newly generated data, no clear pattern of chromatin modifications has so far been linked to specific modes of transcriptional regulation. Here, we used high-throughput genomic data from CD4$^+$ T lymphocytes to provide a comprehensive analysis of histone H3 lysine 4 dimethylation (H3K4me2) enrichment in genomic regions surrounding transcriptional start sites (TSSs). We discovered that a subgroup of genes linked to T cell functions displayed high levels of H3K4me2 within their gene body, in sharp contrast to the TSS-centered profile typical of housekeeping genes. Analysis of additional chromatin modifications and DNase I hypersensitive sites (DHSS) revealed a combinatorial chromatin signature characteristic of this subgroup. We propose that this epigenetic feature reflects the activity of an as yet unrecognized, intragenic *cis*-regulatory platform dedicated to refining tissue-specificity in gene expression.

[Supplemental material is available online at http://www.genome.org.]

Nucleosome, the basic unit of chromatin, consists of an octameric histone core surrounded by genomic DNA (Kornberg and Lorch 1999). The flexible amino termini of nucleosomal histones harbor multiple residues that can be subjected to different types of reversible post-translational modifications, including acetylation and methylation (Bernstein et al. 2007; Li et al. 2007; Mellor et al. 2008; Wang et al. 2009). Combinatorial patterns of histone modifications are expected to convey specific information potential that, in the end, impinges on gene expression (Turner 2002). At a given locus, nuclear factors recruited onto the regulatory elements likely induce changes in the pattern of histone modifications, in turn creating (or suppressing) further binding sites for transcriptional activators (or repressors) with consequences on the transcriptional behavior of the *cis*-linked gene(s) (Kouzarides 2007). Thus, characterization of the epigenetic landscape fundamentally contributes toward deciphering the regulatory mechanisms that govern gene expression.

With the development of high-throughput technologies such as next-generation DNA sequencing, combinations of chromatin modifications can now be associated with functional sequences in the genome. Indeed, distinct chromatin signatures have been linked to individual types of regulatory elements, including enhancers and promoters (Heintzman et al. 2007, 2009; Hon et al. 2009b). Numerous studies have also shown that alterations in chromatin signatures can distinguish the different functional states of a particular regulatory element. For example, active promoters are marked by H3K4me3, repressed promoters by H3K27me3, and poised promoters by both marks (Bernstein et al. 2006). More recently, using genome-wide chromatin maps, specific chromatin signatures were likewise found at gene promoters (Wang et al. 2008), enhancers (Wang et al. 2008), and even exons (Andersson et al. 2009; Kolasinska-Zwierz et al. 2009; Schwartz et al. 2009; Spies et al. 2009; Tilgner et al. 2009).

It is conceivable that genes regulated in a tissue-specific manner display a unique chromatin signature compared with those that are ubiquitously expressed. However, despite an increasing amount of available epigenetic data, no specific chromatin signature has yet been identified to mark tissue-specific gene regulation. H3K4me2 is a histone post-translational modification enriched in *cis*-regulatory regions, in particular promoters, of transcriptionally active genes as well as genes primed for future expression during cell development in higher eukaryotes (Bernstein et al. 2005; Koch et al. 2007; Orford et al. 2008). Recently, a few reports have evidenced, in addition to a quantitative correlation between H3K4me2 levels and transcript abundance (Barski et al. 2007), unexpected disparities in the overall shape of H3K4me2 signals at discrete loci in mammalian cells (Bernstein et al. 2005; Bonnet et al. 2009), with either large H3K4me2 domains extending over the promoter and downstream into the gene body or, conversely, discrete H3K4me2 peaks localized at the promoter only. We hypothesized that such differences could potentially reflect distinct regulatory mechanisms acting to finely tune gene expression.

In this study we set out to explore whether enrichment patterns of H3K4me2 could distinguish different gene categories. To address this question, we examined H3K4me2 chromatin immunoprecipitation with massively parallel sequencing (ChIP-seq) data from human CD4$^+$ T cells (Barski et al. 2007). By thorough computational analysis, we found that a distinct pattern of H3K4me2 enrichment, spanning from the transcriptional start site (TSS) to the gene body, marks a subset of tissue-specific genes. We validated and extended this initial observation using additional epigenetic data, and correlated the resulting chromatin signature with a high frequency of intronic *cis*-regulatory elements. Our study provides evidence for the existence of intragenic *cis*-regulatory platforms differing from "classical" promoters and enhancers and presumably dedicated to the control of tissue-specific gene expression.

[4]Corresponding authors.
E-mail ferrier@ciml.univ-mrs.fr.
E-mail spicuglia@ciml.univ-mrs.fr.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.109389.110.

Results

H3K4me2 profiles distinguish different classes of genes

A visual inspection of H3K4me2 ChIP-seq data from human CD4$^+$ T cells (Barski et al. 2007) reveals striking differences in the levels of H3K4me2 enrichment between individual loci expressed at similar levels (Supplemental Fig. 1). High levels of enrichment were either found localized uniquely within the TSS-surrounding region or spread over a much larger area, including the gene body. To globally visualize this phenomenon, we clustered H3K4me2 signals over regions extending from −2 kb to +8 kb with respect to the TSS, using a set sample composed of all nonoverlapping genes in the human genome (see Methods). Overall, we distinguished five main H3K4me2 profiles (Fig. 1A,B), defined as follows: (1) low levels upstream of the TSS and, unexpectedly, high levels of dispersed enrichment along the gene body (cluster 1); (2) main peak of enrichment ~0.5 kb upstream of the TSS (cluster 2); (3) main peak of enrichment ~1.5 kb downstream of the TSS (cluster 3); (4) main peak of enrichment ~0.5 kb downstream the TSS (cluster 4); and (5) no (or at most faint) enrichment (cluster 5) (examples of genes belonging to the five different clusters are shown in Fig. 2A). Therefore, our approach enabled us to distinguish previously unappreciated patterns of H3K4me2 enrichment in promoters and within genic regions (also see Supplemental Fig. 2A). To gain insight into the category of genes present within each cluster, we performed Gene Ontology (GO) term and KEGG pathway analyses (Table 1). Strikingly, genes within cluster 1 demonstrated a significant enrichment in T-cell-associated functions, including, for example, *IL7R*, *CD3E*, and *CD2* genes. In parallel, we found enrichment of GO terms associated with metabolic processes within clusters 2, 3, and 4. Finally, genes within cluster 5 were preferentially associated with a neuronal function, in agreement with previous reports (Guenther et al. 2007).

A distinct H3K4me2 profile characterizes a subset of tissue-specific genes

Transcriptome analyses of CD4$^+$ T cells revealed varying levels of gene expression between clusters, with cluster 5 displaying the lowest levels (median 7.8), clusters 2–4 intermediate levels (median 8.1), and cluster 1 the highest (median 8.5) (Fig. 1C). Since genes within cluster 1 are also primarily associated with T cell functions, we wondered whether this particular gene set is preferentially expressed in T cells. Indeed, visual inspection of tissue expression data from a recent microarray study (Su et al. 2004) of genes within distinct H3K4me2 clusters suggested that genes within cluster 1 are preferentially expressed in T cells (Fig. 2B). To address this question in a global manner, we thus performed three independent statistical analyses. First, for each H3K4me2 cluster, we performed a comparative analysis between the average levels of

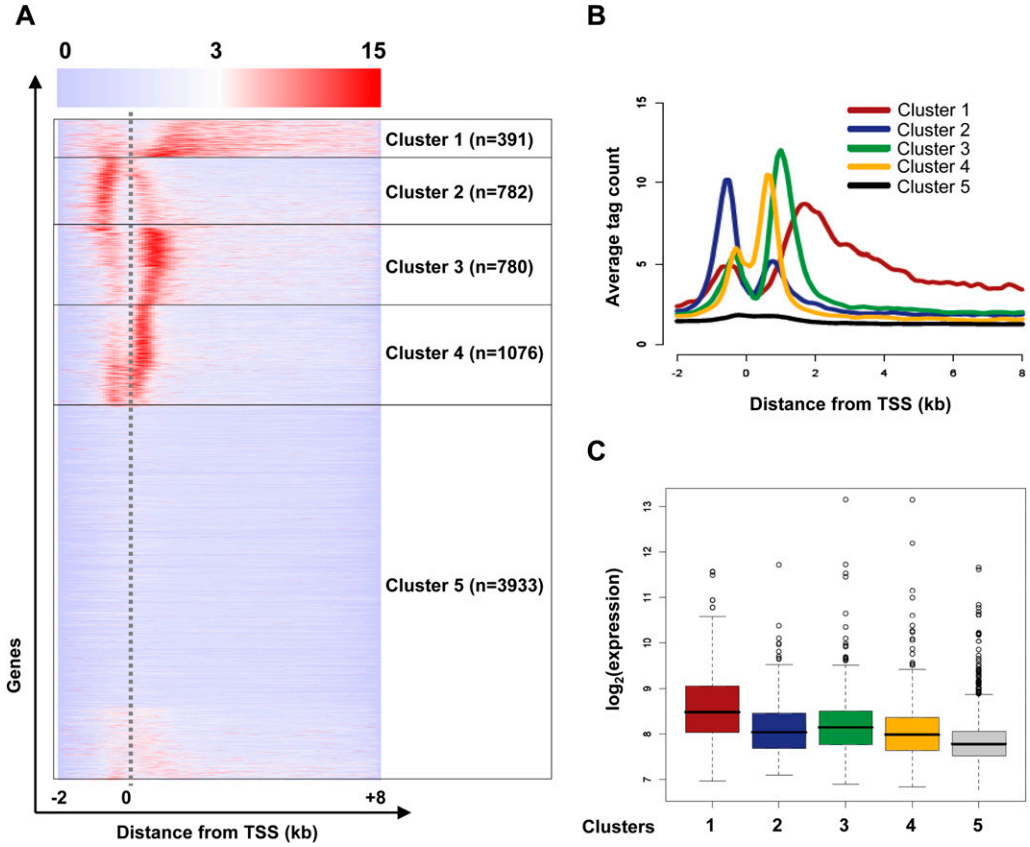

Figure 1. Clustering of H3K4me2 profiles reveals the presence of five main clusters. (*A*) *K*-means clustering of H3K4me2 profiles in regions from −2 to +8 kb around the TSS of a set of nonoverlapping genes. The Euclidean distance was used as a measure of similarity. Color scale indicates the level of enrichment. The number of genes within each cluster is indicated at the *right* of the panel (a list of all genes per cluster is provided in Supplemental Table 1). (*B*) Average profiles of H3K4me2 for genes within the five identified clusters. (*C*) Box plots showing the level of expression in CD4$^+$ T cells of genes identified in the different H3K4me2 clusters.

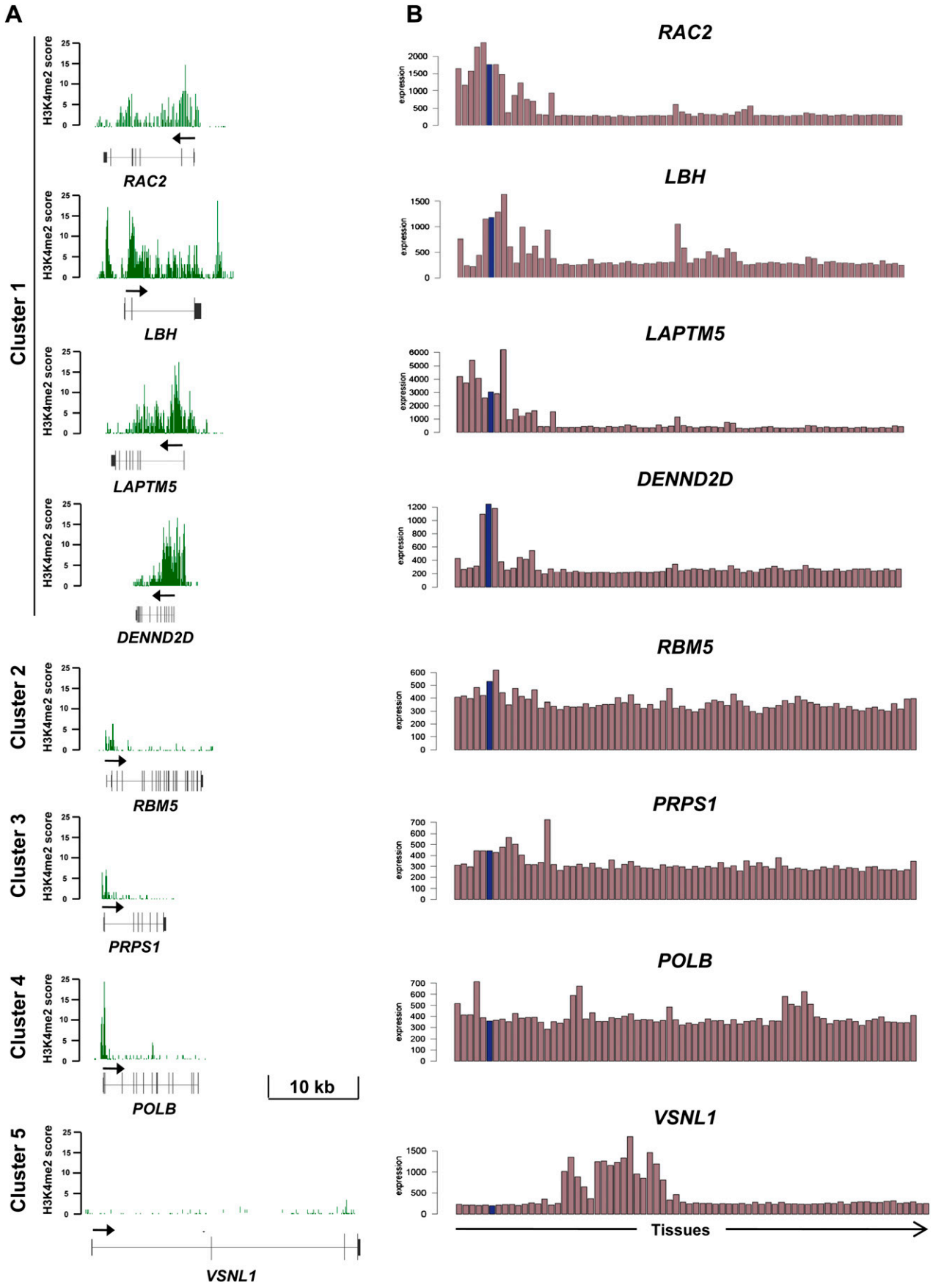

Figure 2. (Legend on next page)

Table 1. KEGG pathway and GO term analyses

Cluster	KEGG pathway	GO: Basic process	GO: Cellular component	GO: Molecular function
1	T cell receptor signaling pathway ($P = 3.2 \times 10^{-5}$)	Immune system process ($P = 6.0 \times 10^{-7}$); T cell activation ($P = 7.4 \times 10^{-6}$)	Cell ($P = 1.5 \times 10^{-5}$)	Protein binding ($P = 4.0 \times 10^{-8}$)
2	—	Cellular metabolic process ($P = 2.3 \times 10^{-4}$)	Intracellular ($P = 2.5 \times 10^{-18}$)	
3	—	Cellular protein metabolic process ($P = 5.6 \times 10^{-5}$)	Intracellular ($P = 1.4 \times 10^{-24}$); membrane-bounded organelle ($P = 1.1 \times 10^{-19}$)	Protein binding ($P = 6.4 \times 10^{-4}$)
4	—	Cellular metabolic process ($P = 2.7 \times 10^{-14}$); nucleobase, nucleoside, nucleotide, and nucleic acid metabolic process ($P = 4.3 \times 10^{-14}$)	Membrane-bounded organelle ($P = 5.7 \times 10^{-30}$)	—
5	Neuroactive ligand-receptor interaction ($P = 1.7 \times 10^{-13}$)	Multicellular organismal process ($P = 9.1 \times 10^{-31}$)	Extracellular region ($P = 8.3 \times 10^{-64}$); plasma membrane ($P = 3.2 \times 10^{-21}$)	Calcium ion binding ($P = 1.1 \times 10^{-17}$); substrate-specific channel activity ($P = 7.8 \times 10^{-11}$)

Top functional annotation terms for genes within the five H3K4me2 clusters obtained from $CD4^+$ T cell ChIP-seq data are indicated. Benjamini-Hochberg corrected *P*-values are noted in parentheses.

gene expression in $CD4^+$ T cells versus those in the remaining 71 tissues. The results demonstrated that the mean level of expression of genes within cluster 1 was statistically higher in $CD4^+$ T cells (one sided *t*-test, $\alpha = 0.05$; note however that genes within cluster 3 also generally exhibited higher levels of expression in $CD4^+$ T cells, although in a less pronounced manner compared with those in cluster 1) (Fig. 3A). Conversely, genes within clusters 2 and 4 were expressed at similar levels both in $CD4^+$ T cells and all the other tissues. Second, we compared expression data between $CD4^+$ T cells and each of the remaining tissues to retrieve differentially expressed genes. We calculated the net number of tissues for which we found each gene to be statistically differentially expressed with respect to $CD4^+$ T cells (moderated *t* statistics, Benjamini-Hochberg corrected, $P < 0.05$). We found that cluster 1 contained more genes statistically overexpressed in $CD4^+$ T cells than any other cluster (Fig. 3B). Finally, we selected T-cell-specific genes using stringent criteria (see Methods) and analyzed their distribution within H3K4me2 clusters. As shown in Figure 3C, T-cell-specific genes were preferentially associated with cluster 1. Indeed, as assessed by χ^2 test, only cluster 1 came out to be significantly enriched for T-cell-specific genes (48% of T-cell-specific genes; $P < 2.2 \times 10^{-16}$). In summary, independent analyses consistently evidenced that cluster 1 is enriched for tissue-specific genes highly expressed in $CD4^+$ T cells.

The H3K4me2 profile of genes within cluster 1 is independent of the expression level

The H3K4me2 profile observed for genes within cluster 1 generalizes our original observation of atypically high H3K4me2 levels in the gene body, and potentially reflects tissue-specific associated functions. We therefore sought to better characterize this previously unappreciated pattern of H3K4me2 enrichment. Since genes within this cluster show a relatively higher level of expression in $CD4^+$ T cells than do genes in the four remaining clusters, we asked whether the intragenic H3K4me2 signal may simply be explained by the expression level of associated genes. Therefore, we subdivided all genes included in this study and those within cluster 1 into four ranges of expression (Supplemental Fig. 3), and calculated the enrichment levels of H3K4me2 and RNA polymerase II (Pol II) (Barski et al. 2007) in the corresponding intragenic regions (from 0 to +8 kb). We observed higher levels of H3K4me2 enrichment and Pol II binding in genes within cluster 1 compared with the randomized set of control genes for all analyzed ranges of expression (Fig. 4A,B; see also Supplemental Fig. 2B). Transcriptome data reflect not only transcription rate but also RNA stability. Thus, our results could imply that genes within cluster 1 either harbor a specific chromatin structure that is independent on the expression level, or encode less stable mRNAs. To discriminate between the two hypotheses, we calculated, for genes in different ranges of expression, the enrichment level of H3K36me3, a hallmark of transcriptional elongation that positively correlates with transcription rates (Guenther et al. 2007; Li et al. 2007). Strikingly, we observed no significant differences of H3K36me3 level between cluster 1 and the control set of genes (Fig. 4C), suggesting that these two set of genes are transcribed at similar levels. Instead, substantially higher levels of H3K4me2 and Pol II in the intragenic region of genes within cluster 1 point to a specific chromatin environment and mode of transcriptional control.

Figure 2. H3K4me2 enrichment profiles at genes within the five H3K4me2 clusters. (*A*) Examples of H3K4me2 enrichment profiles are shown. Gene structure (all genes drawn at the same genomic scale) and transcriptional orientation (arrow) are indicated at the *bottom* of each panel. (*B*) Tissue expression of genes shown in *A*. The following tissues were analyzed (from *left* to *right*): whole blood, CD33 myeloid, CD4 monocytes, dendritic cells, CD56 NK cells, CD4 T cells, CD8 T cells, B cells, endothelial cells, CD34 HSC, thymus, tonsil, lymph node, fetal liver, early erythroid, bone marrow, temporal lobe, globus pallidus, cerebellum peduncles, cerebellum, caudate nucleus, whole brain, parietal lobe, medulla oblongata, amygdala, prefrontal cortex, occipital lobe, hypothalamus, brain thalamus, subthalamic nucleus, cingulate cortex, pons, spinal cord, fetal brain, adrenal gland, lung, heart, liver, kidney, prostate, uterus, thyroid, fetal thyroid, fetal lung, placenta, cardiac myocytes, smooth muscle, HBEC, cultured adypocyte, pancreas, islet cells, testis, testis leydig cell, testis germ cell, testis interstitial cells, testis seminiferous tubulae, salivary gland, trachea, adrenal cortex, ovary, appendix, skin, ciliary ganglion, trigeminal ganglion, atrioventricular node, DRG, superior cervical ganglion, skeletal muscle, uterus corpus, tongue, olfactory bulb, and pituitary. Expression levels in $CD4^+$ T cells are highlighted in blue.

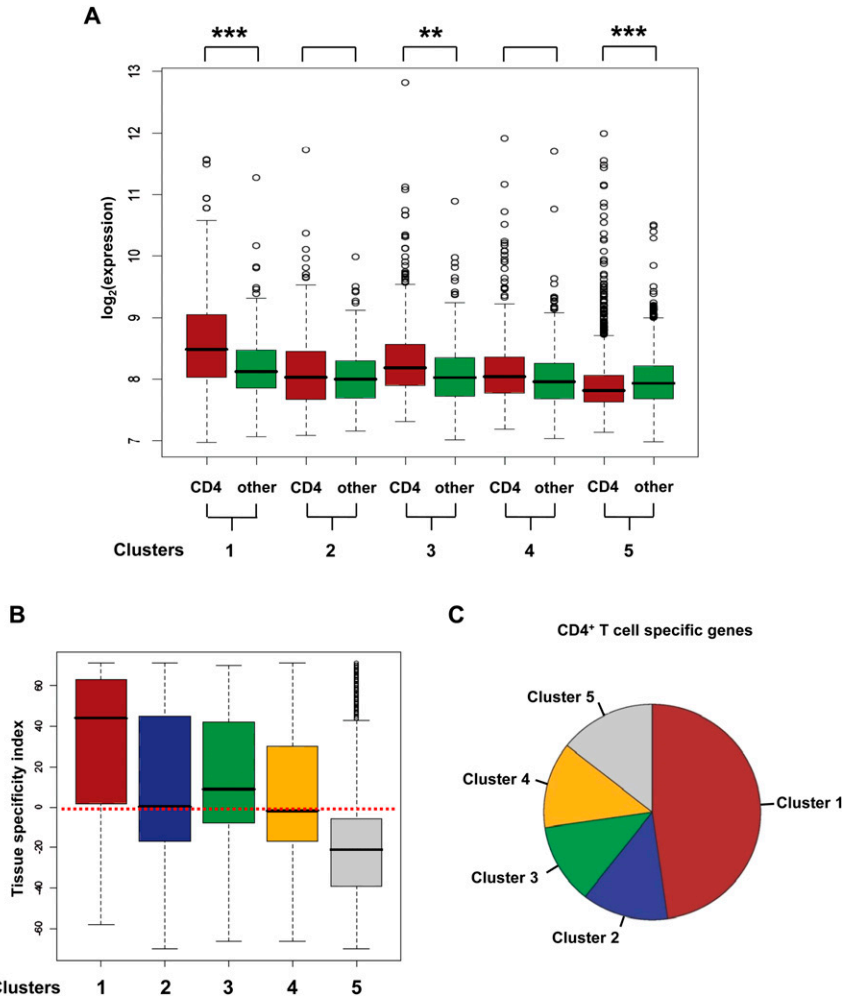

Figure 3. Genes within cluster 1 are preferentially expressed in CD4$^+$ T cells. (*A*) Box plots showing the comparison between levels of gene expression from different H3K4me2 clusters in CD4$^+$ T cells and the mean level of expression calculated across all other tissues (except CD4$^+$ T cells; $^{**}P < 0.005$, $^{***}P < 0.0005$, Mann-Whitney *U* test; other, other tissues). (*B*) Box plots showing per gene calculations of the net number of tissues per gene (i.e., T cell specificity index), for which a difference of expression (between a given tissue and CD4$^+$ T cells) was statistically significant (Benjamini-Hochberg adjusted $P < 0.05$, moderated *t*-test). (*C*) Pie plot showing the distribution of CD4$^+$ T-cell-specific genes within the five H3K4me2 clusters.

A specific chromatin signature marks genes within cluster 1

In order to define the chromatin signature associated with genes belonging to cluster 1, we analyzed available ChIP-seq data for 36 histone modifications together with histone variant H2AZ and Pol II (Barski et al. 2007; Wang et al. 2008). For each feature, we determined the average profile for genes within cluster 1 and normalized this result with respect to the average values calculated for control genes (i.e., expressed at the same level as genes within cluster 1; see Methods). Hierarchical clustering of these values revealed three groups of epigenetic modifications (Fig. 4D). Modifications in group A (mostly histone methylations) generally mimicked the H3K4me2 profile, notably displayed high enrichment throughout gene bodies. In contrast, those in group B (mostly histone acetylations) displayed a marked depletion around the TSS. Finally, modifications in group C (including H3K36me3) displayed minor changes in comparison to the control set of genes. These observations were confirmed by the analysis of individual epigenetic marks for the five H3K4me2 clusters (Supplemental Fig. 4). Given the H3K4me2-mimicking profile observed with chromatin marks that defined group A, we asked whether clustering analysis using individual marks in this group (other than H3K4me2) could also delineate a similar sort of gene classification. However, except possibly to some extent with H3K9me1, a histone modification previously shown to be enriched at transcriptionally active genes (Rosenfeld et al. 2009), these analyses did not unambiguously retrieve a tissue-specific related cluster (Supplemental Fig. 5; data not shown). Altogether, these findings support the existence of a unique chromatin signature for genes gathered in cluster 1, primarily defined by their H3K4me2 profile, general decrease of histone marks around the TSS, and enhanced methylation of specific histone moieties throughout the gene body.

High H3K4me2 levels within the gene body potentially reflect the presence and activity of intragenic *cis*-regulatory elements

How could the discrete epigenetic profile described herein be molecularly linked to tissue specificity? A potential explanation would be that this profile reflects the presence and the activity of intragenic *cis*-regulatory elements. Indeed, genes within cluster 1 had a statistically longer first intron (without displaying significant differences in gene size) and showed a significant enrichment of conserved transcription factor binding sites (TFBSs) and a substantially higher number of putative *cis*-regulatory sequences in regions comprised between the TSS and +8 kb (Supplemental Fig. 6A–D). Another argument supporting the presence of active *cis*-regulatory elements is the high level of gene-body enrichment for H3K4me1 observed for genes within cluster 1 (Fig. 5A), since H3K4me1 strongly associates with tissue-specific enhancer elements (Heintzman et al. 2007, 2009). To further validate this interpretation, we analyzed published DNase I hypersensitive sites (DHSS) of human CD4$^+$ T cells (Boyle et al. 2008). Typically, DHSS are enriched in and around transcribed genes and highlight *cis*-regulatory elements within the genome (Gross and Garrard 1988; Crawford et al. 2004). We calculated, on a per gene basis, the frequency of DHSS in regions comprised between TSS and +8 kb. We observed an accumulation of DHSS for genes within cluster 1 compared with those in the four remaining clusters or those in the control set (Fig. 5B,C). Notably, DHSS in genes within cluster 1 were preferentially localized approximately +2 kb downstream the TSS (Fig. 5D). Moreover, tissue-specific genes within cluster 1 also displayed higher number of DHSS compared with those within the remaining clusters

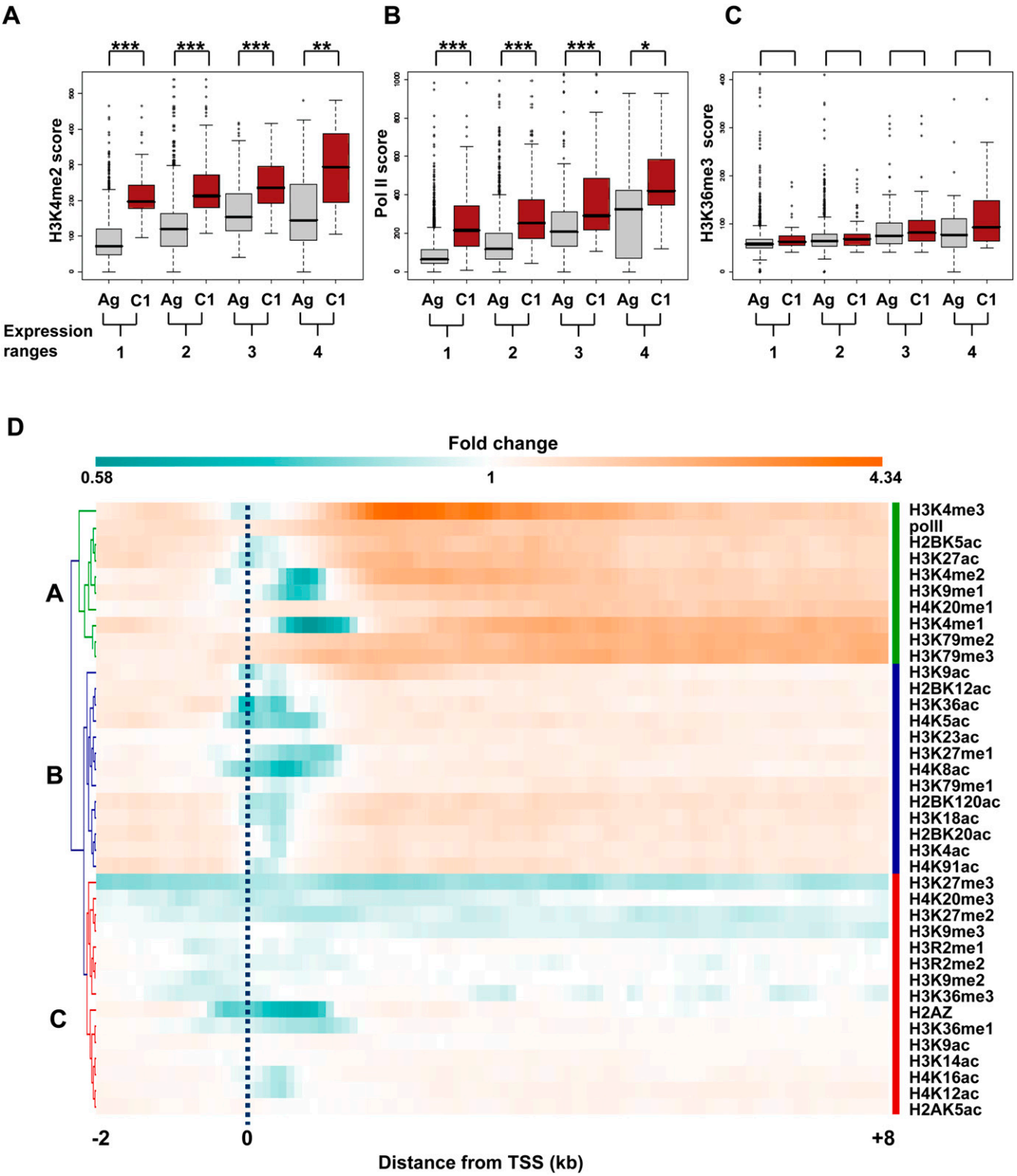

Figure 4. Genes within cluster 1 show increased levels of specific epigenetic modifications. Comparison of scores of H3K4me2 (*A*), RNA polymerase II (*B*), and H3K36me3 (C) for all nonoverlapping genes (Ag, gray box plots) or genes found within cluster 1 (C1, red box plots) in the analyzed genic regions (0–8 kb) was performed for four distinct ranges of gene expression (see text). A significant difference between values obtained for genes within cluster 1 versus the control genes was observed for each range of gene expression (*$P < 0.05$, **$P < 0.005$, ***$P < 0.0005$, Mann-Whitney *U* test). (*D*) Direct comparison of epigenetic marks and Pol II between genes within cluster 1 and a control set of genes. A normalized signal in the regions −2 to +8 kb around the TSS was calculated in each case by dividing the average signal obtained for cluster 1 by the average signal obtained for all genes expressed at the same level as genes within cluster 1. A hierarchical clustering of these values is shown. Most of the analyzed modifications displayed a marked decrease in signal levels in the regions close to the TSS.

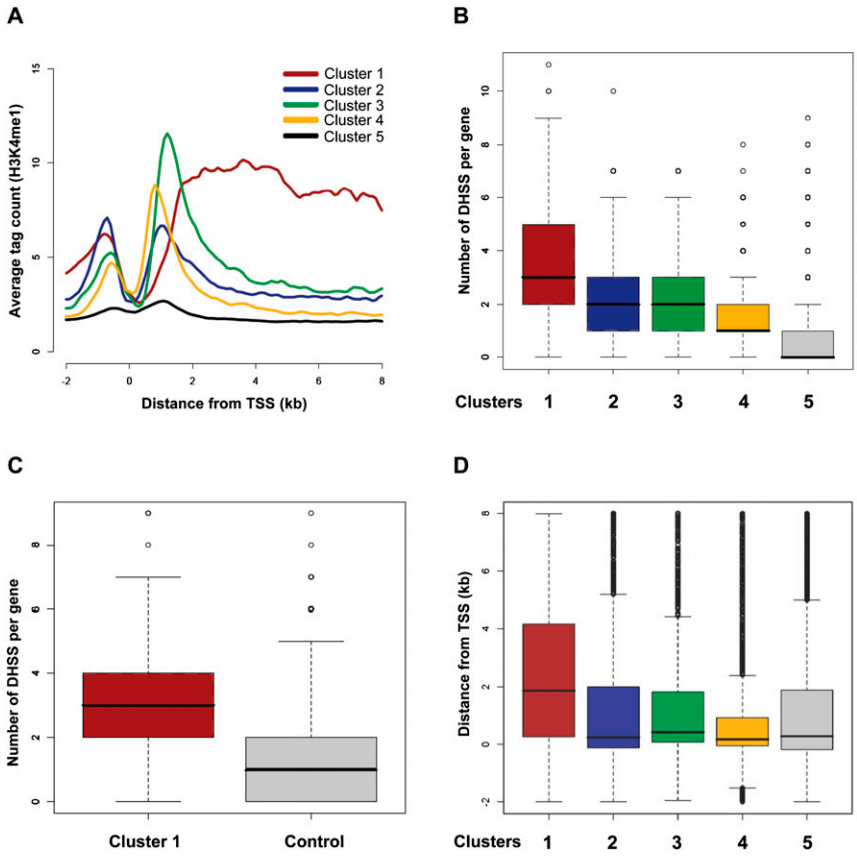

Figure 5. Epigenetic and accessibility profiles reflect different modes of transcriptional control of genes within the H3K4me2 clusters. (*A*) H3K4me1 average profiles. (*B*) The number of DHSS in regions from TSS to 8 kb downstream is shown. (*C*) Comparison between the frequency of DHSS in genes within cluster 1 and that observed for all genes expressed at the same level. (*D*) Distribution of DHSS locations with respect to the TSS for each H3K4me2 cluster.

(Supplemental Fig. 7), further supporting the specificity of the chromatin environment of these particular genes. Together with the observation that active epigenetic marks are generally diminished in the vicinity of the TSS, these analyses strongly suggest that genes in cluster 1 harbor an extended promoter region with a high frequency of *cis*-regulatory elements localized within the gene body.

H3K4me2 profiles from brain also distinguish tissue-specific genes

In order to generalize our finding that H3K4me2 chromatin profiles may allow the identification of tissue-specific genes, we analyzed H3K4me2 ChIP-seq data obtained in this case from mouse brain (Meissner et al. 2008). Remarkably, clustering of H3K4me2 signals in this situation also allowed the isolation of a gene cluster (hereafter referred to as brain-cluster 1), characterized by a high level of H3K4me2 enrichment within the gene body (Supplemental Fig. 8A–C), thus reminiscent of the $CD4^+$ T cell cluster 1 described above. Indeed, genes within brain-cluster 1 were (1) expressed at a higher level in the brain compared to other tissues (Supplemental Fig. 8D,E); (2) enriched in brain-specific genes (Supplemental Fig. 8F); and (3) significantly enriched in brain-associated functional categories (Supplemental Table 2). In conclusion, the distinctive H3K4me2 pattern characteristic of tissue-specific genes appears to be a general feature observed in distinct tissues from different species, suggesting functional conservation in mammals.

Discussion

Whether tissue-specific regulated genes carry a distinct chromatin signature, opposite to ubiquitously expressed genes, is a long-standing question. Thorough analysis of H3K4me2 signatures enabled us to distinguish several categories of genes (clusters 1–5) and to highlight a unique chromatin structure typical of a subset of tissue-specific genes (cluster 1). Importantly, we observed this property in two different tissues, $CD4^+$ T cells and brain, from two distinct species (human and mouse, respectively), suggesting a fundamental mechanism of the regulation of tissue-specific gene expression. Enrichment for H3K4me2 within the gene body is a characteristic trait of expressed genes in yeast (Bernstein et al. 2002; Li et al. 2007). In contrast, in mammals, H3K4me2 has been shown to be predominantly enriched around the TSS (Bernstein et al. 2005; Barski et al. 2007). Therefore, the fact that a subset of tissue-specific mammalian genes displays high levels of H3K4me2 within the gene body is remarkable. Instead of being linked to a general transcriptional mechanism, as it seems to be the case in yeast, this epigenetic profile in mammals may reflect the activity of a wealth of seemingly functional intragenic *cis*-regulatory elements. This is supported by the observation that genes within cluster 1 are highly enriched for histone methylations, DHSS, and conserved TFBSs within their gene bodies. Overall, our data imply that a subset of tissue-specific regulated genes harbor previously unrecognized, TSS-proximal, intragenic *cis*-regulatory elements.

Are these structures different from classical promoter and enhancer elements? They generally are larger than such regulatory elements, appear to preferentially localize within the gene body, and are associated with diminished levels of active epigenetic modifications around the TSS. The latter observation indeed suggests that the *cis*-linked core promoters would display a weaker transcriptional activity by themselves. We therefore postulate that the newly identified intragenic regions represent complementary regulatory structures or "platforms" dedicated to deal with the complex regulation of tissue-specific gene expression. This hypothesis is further supported by several recent observations, including those of the frequent binding of specific transcription factors within intragenic regions (Hatzis et al. 2008; Lupien et al. 2008; Kwon et al. 2009; Visel et al. 2009a) and of the loading of the Ldb1 regulatory complex in the first intron of CpG-rich genes activated during erythroid differentiation (Soler et al. 2010). Considering the latter observation, we found that the majority of genes within cluster 1 displayed high CpG content at their promoters (Supplemental Fig. 9). This is also interesting in light of recent

observations of a subset of CpG-rich promoters displaying high tissue specificity (Carninci et al. 2006; Saxonov et al. 2006; Akalin et al. 2009).

Developmentally and tissue regulated genes respond to signaling cues from within and outside the cell in connection to discrete functional contexts. It is obvious that such genes harbor a complex circuitry of distal and proximal regulatory elements, most likely dedicated to orchestrate spatial and temporal profiles of gene expression (West and Fraser 2005; Ochoa-Espinosa and Small 2006; Visel et al. 2009b). Akalin and colleagues have recently found that genes involved in development and differentiation, contained within genomic regulatory blocks, carry a characteristic intronic structure displaying a wide distribution of transcription initiation sites around the TSS, large CpG islands, and high density of TFBS spanning the gene body. Interestingly, these properties were proposed to be linked to long-range regulation by distal enhancers (Akalin et al. 2009; Soler et al. 2010). In this line, we found that genes within cluster 1 are associated with high numbers of distal intergenic DHSS (Supplemental Fig. 10). Thus, it is formally possible that the regulatory platform identified here for genes within cluster 1 serves as a hub for the interactions between TSS proximal and distal regulatory sequences, in order to fine-tune tissue-specific regulation of gene expression. Testing this hypothesis and the functional significance of the putative platforms within their genomic context will require further experimental investigations.

So far, to our knowledge, only a few studies have used ChIP-seq data available for $CD4^+$ T cells to investigate the relationship between histone modifications and regulation of gene expression (e.g., Hon et al. 2009a; Lim et al. 2009; She et al. 2009; Karlic et al. 2010). Here, we have demonstrated that by using this comprehensive set of data, different classes of genes can be distinguished, therefore providing insight into the regulatory mechanisms underlying tissue- and locus-specific gene expression.

Methods

Data source

ChIP-seq data (Barski et al. 2007; Wang et al. 2008) corresponding to histone methylation, acetylation, Pol II, and histone variant H2AZ in human $CD4^+$ T cells, were downloaded from the Keji Zhao laboratory server (http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx; http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellacetylation.aspx). Downloaded files display the number of uniquely mapped reads in 200-bp windows. All read start coordinates (Hg18 human genome annotation) are adjusted by 75 bp in the direction of the strand they map to before summing is done. Mouse whole-brain H3K4me2 ChIP-seq data (Meissner et al. 2008) were downloaded from the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRR002257 and aligned to the reference genome (NCBI37/mm9) using Bowtie (Langmead et al. 2009). Only uniquely mapped sequence tags were considered for further analyses (34% of all sequenced tags). We next discarded all tags with exactly the same coordinates to reduce potential bias. Sequences mapping only once to the reference genome were included for further analyses (34% of all sequenced tags). Tags with exactly the same coordinates were discarded. Reads were then elongated to 300 bp in the direction of the strand. The values were next summed up in nonoverlapping 200-bp bins along each chromosome. The summed values were finally normalized (by division) by the overall mean of values per bin. The combined set of DHSS data (Boyle et al. 2008) (isolated DHSS peaks) were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/) and converted to Hg18 genome annotation. Conserved TFBSs and the list of putative *cis*-regulatory elements were downloaded from the UCSC Genome Browser (tracks "conserved TFBS" and "regPotential7x", respectively). The raw expression data for 72 human (Su et al. 2004) and 75 mouse (Lattin et al. 2008) tissues were downloaded from the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; accession nos. GSE1133 and GSE10246, respectively).

Data analysis

All the data analysis was performed using R software (http://r-project.org/) except for H3K4me2 visualization, for which we used the igb software (http://www.affymetrix.com); ChIP-seq profile clustering analysis performed with MeV software (Saeed et al. 2006) (http://www.tm4.org/mev/); and the lift-over analysis of DHSS performed with a custom Perl script. Mouse whole-brain H3K4me2 ChIP-seq tags were annotated to the reference genome using Bowtie software. Venn Diagrams were drawn using BioVenn software (Hulsen et al. 2008; http://www.cmbi.ru.nl/cdd/biovenn/).

Statistical analysis

Differentially expressed genes were retrieved using the moderated t statistics (limma package) using as a threshold Benjamini-Hochberg adjusted *P*-values < 0.05. Statistical significance of the level of expression between $CD4^+$ T cells (or, in the case of mouse data analysis, the value obtained by averaging the expression level for a given gene in all the brain tissues covered by the analyzed study) and the mean values for the remaining tissues, as well as the comparisons of scores for H3K4me2 and Pol II for genes within cluster 1 versus all genes expressed at the same levels, were assessed by Mann-Whitney *U* test. For further details, refer to the specific subsections in Methods.

Choice of the width of analyzed regions

Based on the NCBI RefSeq annotation (build 36.1), we selected all nonoverlapping genes. The filtering criteria were as follows: (1) the extended genomic region (i.e., the region comprised between −2 kb from the TSS and the end of the transcribed region) should not overlap with any other RefSeq annotated transcribed region; (2) the gene has only one transcript annotated in the RefSeq database; and (3) the gene must be >8 kb. This resulted in a list of nonoverlapping genes, >8 kb, harboring a unique TSS. For all analyzed epigenetic modifications and Pol II binding together with histone variant H2AZ, we annotated tags and performed a linear interpolation of the data corresponding to the regions from −2 kb to +8 kb around each TSS, in order to obtain uniform reference positions of bins for every gene. The analysis of vast regions around TSS, rather than scaling the total length of all genes (and analyzing the profiles of whole transcribed regions), enabled us to directly compare epigenetic profiles that were less influenced by the gene structure (i.e., gene length, exon density). The average profiles (Supplemental Fig. 2A) for genes from the highest range of expression show an important decrease in the H3K4me2 level toward the end of the analyzed region, supporting the empirically chosen genomic analysis frame. In addition, clustering of H3K4me2 profiles in regions from −2 kb to +8 kb enabled us to maintain a large population of genes.

Clustering of H3K4me2 profiles and further analyses

We performed a *k*-means clustering on fitted H3K4me2 values with a Euclidean distance measure. We isolated five clusters for further

studies. In order to calculate a score per modification in regions surrounding TSS (±2 kb) and in included intragenic regions (TSS to + 8 kb), we summed the fitted tag count in isolated regions. We then divided the complete set of all nonoverlapping genes and genes from cluster 1 in four different ranges of expression (we did not include genes expressed at a higher level than the last range of expression because of a too small sample size) (see Supplemental Fig. 3). We next compared previously calculated scores of H3K4me2 and Pol II for all genes and genes from cluster in all four ranges of expression (Mann-Whitney *U* test).

In order to compare genes within cluster 1 to a similarly expressed set of genes we selected genes from cluster 1 and from the total set of genes (control set) for which the expression was comprised between the 25th and 75th percentile of expression values observed for genes within cluster 1. We then normalized (by dividing) the average profile obtained for a given epigenetic feature for the selected genes within cluster 1, by the average profile of the control set of genes. Finally we performed a hierarchical clustering of aforementioned normalized values with the Manhattan distance measure.

In order to annotate DHSS, putative *cis*-regulatory elements, and conserved TFBSs, we calculated the middle point for each feature and assessed its distance from the studied list of TSS. Only features included in the regions from TSS to +8 kb were retained for further data analysis.

Gene function analysis

KEGG pathway and GO term analysis were performed with DAVID 2008 (Hochberg and Benjamini 1990). We used as a cut-off false-discovery rate < 0.01 and a Benjamini-Hochberg corrected *P*-value < 0.05.

Transcriptome data analysis

The raw expression data for 72 human tissues, including $CD4^+$ T cells, and mouse tissues were VSN-normalized (Huber et al. 2002), and probe annotation to the HG18 genome assembly and NCBI37/mm9 was used for subsequent analyses. In order to compare the level of expression of genes within identified H3K4me2 clusters in $CD4^+$ T cells with the level of expression in the remaining 71 tissues, we calculated the mean level of expression of these genes in $CD4^+$ T cells and across the remaining 71 tissues. We then performed a Mann-Whitney *U* test to check for statistical significance of observed differences. In order to isolate differentially expressed genes between $CD4^+$ T cells and all the remaining tissues, we performed moderated *t* statistics (Smyth 2004) using as a cut-off the Benjamini-Hochelberg adjusted $P < 0.05$. We focused at the analysis of the differential expression rather than on comparison of the number of tissues for which a given gene is called expressed because of the threshold independence of the former approach. This gave us for each gene, a matrix of values across all comparisons ($CD4^+$ T cells vs. every tissue), with 1 encoding for genes with a statistically higher level of expression in $CD4^+$ T cells; −1, for genes for which we observed statistically lower level of expression in $CD4^+$ T cells; and 0, for no difference. We then summed these values per gene ("T cell specificity index") and visualized their distributions for different H3K4me2 clusters with box plots. In order to isolate highly $CD4^+$ T-cell-specific genes, we performed moderated *t* statistics with thresholds as follows: (1) Benjamini-Hochberg corrected $P < 0.05$; and (2) at least twofold difference in the level of expression between $CD4^+$ T cells and each tissue. We next calculated a gene-specific net number of tissues for which we found it differentially expressed. Based on this, we set a cut-off equal to 20 and retrieved the highly $CD4^+$ T-cell-specific genes. The same approach was used when analyzing mouse gene expression data. In the later case, the whole-brain transcriptome corresponded to the averaged values of expression of a given gene in all brain tissues. These values were subsequently used for gene expression analyses, assignment of a "brain specificity score," and the retrieval of highly brain-specific genes. Because of high number of brain tissues included in the gene expression data set, we set up a threshold of 55 in order to define brain-specific genes.

CpG analysis

Sequences for regions from −2.5 kb to +0.5 kb around the TSS of our total set of nonoverlapping genes >8 kb were downloaded from the UCSC Genome Browser. We then calculated CpG observed versus expected ratio in those regions (Saxonov et al. 2006). We obtained a bimodal distribution of ratios (see Supplemental Fig. 5). We next retrieved CpG observed versus expected ratios for genes within the H3K4me2 cluster 1.

Acknowledgments

We thank members of the P.F. laboratory for critical reading of this manuscript. Work in this laboratory is supported by institutional grants from INSERM and the CNRS, and by specific grants from the "Fondation Princesse Grace de Monaco," the "Association pour la Recherche sur le Cancer" (ARC), the "Agence Nationale de la Recherche" (ANR), the "Institut National du Cancer" (INCa), and the Commission of the European Communities. A.P. and T.B. were supported by a Marie Curie research training fellowship and the ANR, respectively, and are now supported by the "Fondation pour la Recherche Medicale" (FRM).

Author contributions: A.P., T.B., and S.S. made the original observations; A.P. performed computational analysis; A.P. and S.S. analyzed the data; A.P., S.S., and P.F. wrote the paper; S.S. and P.F. directed the study.

References

Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, Daub CO, Hayashizaki Y, Lenhard B. 2009. Transcriptional features of genomic regulatory blocks. *Genome Biol* **10:** R38. doi: 10.1186/gb-2009-10-4-r38.

Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. 2009. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res* **19:** 1732–1741.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, Liu JS, Kouzarides T, Schreiber SL. 2002. Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci* **99:** 8695–8700.

Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ III, Gingeras TR, et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120:** 169–181.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125:** 315–326.

Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. *Cell* **128:** 669–681.

Bonnet M, Huang F, Benoukraf T, Cabaud O, Verthuy C, Boucher A, Jaeger S, Ferrier P, Spicuglia S. 2009. Duality of enhancer functioning mode revealed in a reduced TCRβ gene enhancer knockin mouse model. *J Immunol* **183:** 7939–7948.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132:** 311–322.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38:** 626–635.

Crawford GE, Holt IE, Mullikin JC, Tai D, Blakesley R, Bouffard G, Young A, Masiello C, Green ED, Wolfsberg TG, et al. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc Natl Acad Sci* **101:** 992–997.
Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57:** 159–197.
Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88.
Hatzis P, van der Flier LG, van Driel MA, Guryev V, Nielsen F, Denissov S, Nijman IJ, Koster J, Santo EE, Welboren W, et al. 2008. Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol* **28:** 2732–2744.
Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318.
Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459:** 108–112.
Hochberg Y, Benjamini Y. 1990. More powerful procedures for multiple significance testing. *Stat Med* **9:** 811–818.
Hon G, Wang W, Ren B. 2009a. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* **5:** e1000566. doi: 10.1371/journal.pcbi.1000566.
Hon GC, Hawkins RD, Ren B. 2009b. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* **18:** R195–R201.
Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18:** S96–S104.
Hulsen T, de Vlieg J, Alkema W. 2008. BioVenn: A web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics* **9:** 488. doi: 10.1186.1471-2164-9-448.
Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. 2010. Histone modification levels are predictive for gene expression. *Proc Nat Acad Sci* **107:** 2926–2931.
Koch CM, Andrews RM, Flicek P, Dillon SC, Karaoz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P, et al. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* **17:** 691–707.
Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41:** 376–381.
Kornberg RD, Lorch Y. 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98:** 285–294.
Kouzarides T. 2007. Chromatin modifications and their function. *Cell* **128:** 693–705.
Kwon H, Thierry-Mieg D, Thierry-Mieg J, Kim HP, Oh J, Tunyaplin C, Carotta S, Donovan CE, Goldman ML, Tailor P, et al. 2009. Analysis of interleukin-21-induced Prdm1 gene regulation reveals functional cooperation of STAT3 and IRF4 transcription factors. *Immunity* **31:** 941–952.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.
Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, et al. 2008. Expression analysis of G protein–coupled receptors in mouse macrophages. *Immunome Res* **4:** 5. doi: 10.1186/1745-7580-4-5.
Li B, Carey M, Workman JL. 2007. The role of chromatin during transcription. *Cell* **128:** 707–719.
Lim PS, Hardy K, Bunting KL, Ma L, Peng K, Chen X, Shannon MF. 2009. Defining the chromatin signature of inducible genes in T cells. *Genome Biol* **10:** R107. doi: 10.1186/gb-2009-10-10-r107.
Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M. 2008. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132:** 958–970.
Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454:** 766–770.
Mellor J, Dudek P, Clynes D. 2008. A glimpse into the epigenetic landscape of gene regulation. *Curr Opin Genet Dev* **18:** 116–122.
Ochoa-Espinosa A, Small S. 2006. Developmental mechanisms and *cis*-regulatory codes. *Curr Opin Genet Dev* **16:** 165–170.
Orford K, Kharchenko P, Lai W, Dao MC, Worhunsky DJ, Ferro A, Janzen V, Park PJ, Scadden DT. 2008. Differential H3K4 methylation identifies developmentally poised hematopoietic genes. *Dev Cell* **14:** 798–809.
Rosenfeld JA, Wang Z, Schones DE, Zhao K, DeSalle R, Zhang MQ. 2009. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* **10:** 143.
Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. 2006. TM4 microarray software suite. *Methods Enzymol* **411:** 134–193.
Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci* **103:** 1412–1417.
Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol* **16:** 990–995.
She X, Rohl CA, Castle JC, Kulkarni AV, Johnson JM, Chen R. 2009. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* **10:** 269. doi: 10.1186/1471-2164-10-269.
Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3:** Article3. doi: 10.2202/1544-6115.1027.
Soler E, Andrieu-Soler C, de Boer E, Bryne JC, Thongjuea S, Stadhouders R, Palstra RJ, Stevens M, Kockx C, van Ijcken W, et al. 2010. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes Dev* **24:** 277–289.
Spies N, Nielsen CB, Padgett RA, Burge CB. 2009. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36:** 245–254.
Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci* **101:** 6062–6067.
Tilgner H, Nikolaou C, Althammer S, Sammeth M, Beato M, Valcarcel J, Guigo R. 2009. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol* **16:** 996–1001.
Turner BM. 2002. Cellular memory and the histone code. *Cell* **111:** 285–291.
Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009a. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457:** 854–858.
Visel A, Rubin EM, Pennacchio LA. 2009b. Genomic views of distant-acting enhancers. *Nature* **461:** 199–205.
Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40:** 897–903.
Wang Z, Schones DE, Zhao K. 2009. Characterization of human epigenomes. *Curr Opin Genet Dev* **19:** 127–134.
West AG, Fraser P. 2005. Remote control of gene transcription. *Hum Mol Genet* **14:** R101–R111.

Received April 20, 2010; accepted in revised form August 13, 2010.