nature structural & molecular biology

Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters

Frederic Koch[1–3,10], Romain Fenouil[1–3,10], Marta Gut[4–6,10], Pierre Cauchy[1–3,7], Thomas K Albert[8], Joaquin Zacarias-Cabeza[1–3], Salvatore Spicuglia[1–3], Albane Lamy de la Chapelle[1–3], Martin Heidemann[9], Corinna Hintermair[9], Dirk Eick[9], Ivo Gut[4,6], Pierre Ferrier[1–3] & Jean-Christophe Andrau[1–3]

Recent work has shown that RNA polymerase (Pol) II can be recruited to and transcribe distal regulatory regions. Here we analyzed transcription initiation and elongation through genome-wide localization of Pol II, general transcription factors (GTFs) and active chromatin in developing T cells. We show that Pol II and GTFs are recruited to known T cell–specific enhancers. We extend this observation to many new putative enhancers, a majority of which can be transcribed with or without polyadenylation. Importantly, we also identify genomic features called transcriptional initiation platforms (TIPs) that are characterized by large areas of Pol II and GTF recruitment at promoters, intergenic and intragenic regions. TIPs show variable widths (0.4–10 kb) and correlate with high CpG content and increased tissue specificity at promoters. Finally, we also report differential recruitment of TFIID and other GTFs at promoters and enhancers. Overall, we propose that TIPs represent important new regulatory hallmarks of the genome.

Transcription initiation at promoters requires the stepwise assembly of GTFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH) and Pol II (ref. 1) and correlates with phosphorylation of the C-terminal domain (CTD) of Pol II on Ser5 residues (Ser5P). Transition to elongation further requires phosphorylation of Ser2 residues (Ser2P)[2]. Initiation can be a loose process, as most genes harbor multiple transcription start sites (TSSs). It was shown that genes with high usage of alternative TSSs weakly correlate with a lack of TATA-boxes and with high CpG content[3].

Transcriptional activation often involves a complex interplay between proximal and distal regulatory regions such as enhancers[4]. Known features of enhancers include their distant location from TSSs, enrichment for transcription factor binding sites (TFBS) and combinations of epigenetic marks such as histone H3 Lys4 monomethylation (H3K4me1) and relatively lower, though variable, levels of trimethylation (H3K4me3)[5–8]. They are also often remodeled by specific histone acetyltransferases such as CBP[6,9,10]. How enhancers function and communicate with promoters, however, remains largely elusive. Several models, including the well-studied globin genes[11,12], propose that enhancers recruit Pol II and GTFs before promoter translocation[4,13]. Recent genome-wide studies show Pol II recruitment to, and subsequent transcription of, enhancer-like regions in both macrophages[14] and neurons[15].

We set out to characterize the general transcriptional machinery in developing mouse (*Mus musculus*) CD4$^+$ CD8$^+$ double-positive thymocytes by conducting genome-wide ChIP-seq experiments for GTFs (TFIID including TBP and TAF1, TFIIA, B, E, F and H), as well as for total, initiating (Ser5P) and elongating (Ser2P) Pol II[2] (**Supplementary Fig. 1**). Active chromatin marks (H3K4me1, H3K4me3, H3K36me3), CBP and formaldehyde-assisted isolation of regulatory elements (FAIRE)[16] were analyzed to highlight open chromatin regions. Our results show that many tissue-specific enhancers of genes expressed at the double-positive stage recruit Pol II and GTFs. We further isolated new putative enhancers with similar histone modification patterns and provide evidence that the majority of known and putative enhancers can be transcribed with or without polyadenylation. Based on TBP- and Ser5P-bound regions, we also characterized new genomic elements corresponding to large platforms of Pol II and GTF recruitment and/or initiation at promoters, intragenic or intergenic regions (IGRs). Transcription initiates primarily on either boundary of these transcription initiation platforms, delimited by nucleosomal positioning. Our analysis further indicates that TIPs generally overlap with high CpG content as well as TFBS on promoters. The TIP size at promoters correlates with tissue specificity. Finally, our data also suggest differential recruitment levels of TFIID, TFIIA and Pol II compared to remaining GTFs on promoters and IGRs. We propose a model in which promoters and at least a subset of enhancers can recruit GTFs and Pol II and initiate transcription but differ in nucleosomal composition and TFBS or CpG content.

[1]Centre d'Immunologie de Marseille-Luminy, Université Aix-Marseille, Campus de Luminy, Marseille, France. [2]Centre National de la Recherche Scientifique, UMR6102, Marseille, France. [3]Institut National de la Santé et de la Recherche Médicale, U631, Marseille, France. [4]Centre National de Génotypage, Commissariat à l'Energie Atomique, Evry, France. [5]Fondation Jean Dausset—Centre d'Etude du Polymorphisme Humain, Paris, France. [6]Centre Nacional D'Anàlisi Genòmica, Parc Científic de Barcelona, Baldiri i Reixac, Barcelona, Spain. [7]Techniques Avancées pour le Génome et la Clinique, Marseille, France. [8]Institute of Molecular Tumor Biology, Medical Faculty of the Westfälische Wilhelms-Universität, Münster, Germany. [9]Department of Molecular Epigenetics, Helmholtz Center Munich, Center of Integrated Protein Science, Munich, Germany. [10]These authors contributed equally to this work. Correspondence should be addressed to J.C.A. (andrau@ciml.univ-mrs.fr), I.G. (igut@pcb.ub.es) or P.F. (ferrier@ciml.univ-mrs.fr).

Received 30 June 2010; accepted 12 May 2011; published online 17 July 2011; doi:10.1038/nsmb.2085

RESULTS

Pol II and GTFs at T cell–specific and putative enhancers

Many genes essential during T-lymphocyte ontogenesis are controlled by enhancers[17,18]. Inspection of enhancers or regulatory elements controlling T-specific genes in our ChIP-seq dataset indicated enrichments for Pol II, TAF1 and TBP, TFIIA, TFIIB, TFIIE, TFIIF and TFIIH (**Fig. 1a–c** and **Supplementary Fig. 2a–c**). This was the case for the *Cd4* double-positive stage-specific proximal enhancer[19] (**Fig. 1a** and **Supplementary Fig. 2a**), several DNase hypersensitivity sites in both the *Dntt*[20] and *Ikaros* (*Ikzf1*)[21] loci as well as the well-studied enhancers of the *Tcrα* (*Tcra*) and *Tcrβ* (*Tcrb*) genes (**Supplementary Fig. 2d–g**). At the *Cd3* (*Cd3d*) locus, we observed binding at the δ enhancer[22] as well as at the antisilencer element (ASE)[23] of the *Rag1* and *Rag2* loci, both T cell–specific regulatory elements (**Supplementary Fig. 2h–i**). Although we observed only initiating Pol II (Ser5P) in most double-positive model regulatory regions, $E8_I$ and $E8_{II}$ controlling the expression of the Cd8 co-receptor subunits *Cd8a* and *Cd8b1* (refs. 24,25) were slightly enriched for the Ser2P and H3K36me3 elongation marks. We also observed Pol II and GTF recruitment in areas controlling inactive genes such as *Il2ra*[26], which is expressed before and after the double-positive stage of differentiation[27] (**Fig. 1c** and **Supplementary Fig. 2c**). Globally, these known regulatory regions showed enhancer hallmarks such as enrichment for H3K4me1 and CBP, high chromatin accessibility, H3K36me3 depletion and a low but detectable H3K4me3 level.

We noticed similar features in IGRs flanking many known genes expressed in T cells (**Supplementary Fig. 2j–l**), so we investigated Pol II–bound putative enhancers genome wide. We selected TBP and Ser5P common peaks outside of any gene annotation in IGRs and compared them to promoters (**Supplementary Fig. 3a**). Consistent with the profiles on model T-cell enhancers, we found both H3K4me1 and H3K4me3 marks at most TBP and Ser5P promoters and IGRs (**Supplementary Fig. 3b**), but H3K4me1:H3K4me3 ratios were clearly higher in the latter (**Supplementary Fig. 3c**). In total, we isolated 708 IGRs and 2,864 promoter peaks (corresponding to 2,539 genes). Average profiling of the GTFs indicated similar TAF1 and TFIIA but higher TFIIB, TFIIE, TFIIF and TFIIH levels in IGRs compared to promoters. In addition to Pol II and GTFs, we also observed increased CBP and FAIRE and decreased signals for elongation marks at IGRs (**Fig. 2a** and **Supplementary Fig. 3d,e**), features more specific to regulatory elements. Altogether, these findings and the observed differences with promoters strongly support the idea that TBP and Ser5P IGRs are distant regulatory features rather than unannotated promoters, and they suggest a differential composition of the general machinery at promoters and IGRs. Furthermore, the observation that most GTFs show more average binding at IGRs also argues against indirect cross-linking originating from promoters.

Tissue specificity and TFBS at putative enhancers

We further analyzed genes adjacent to promoters and IGRs for tissue-specific expression and TFBS content. Our analysis indicated that these genes show significantly higher expression in T cells, primarily at the double-positive stage ($P = 1.39 \times 10^{-105}$ and 1.39×10^{-38}, respectively; **Fig. 2b** and **Supplementary Fig. 4a,b**). IGR-associated genes showed greater differential expression levels between double-positive (or hematopoietic) cells and remaining tissues compared to the promoter set (**Fig. 2c** and **Supplementary Fig. 4c**). This increased T cell–restricted expression pattern suggests that the selected putative enhancers are highly tissue specific. To assess the accuracy of our putative enhancers in IGRs, we compared it to selections based on CBP recruitment and either H3K4me1 alone or in combination with H3K4me3. This comparison indicated slightly improved tissue selectivity of expression of TBP and Ser5P selections; however, their associated genes were only partially overlapping (**Fig. 2c** and **Supplementary**

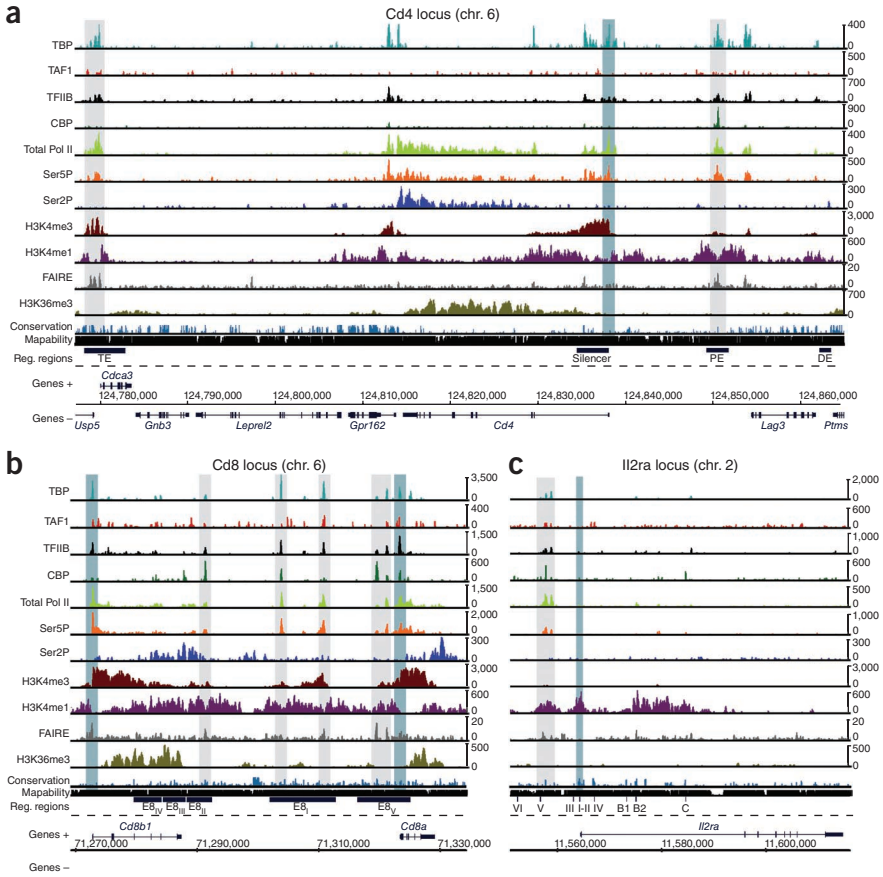


Figure 1 Pol II and GTF recruitment to T-cell stage-specific enhancers of active loci or genes poised for activation. (**a–c**) ChIP-seq binding profiles for GTFs, CBP, total, initiating (Ser5P) and elongating (Ser2P) Pol II, active chromatin marks and FAIRE (accessible DNA regions). Light gray vertical bands show previously annotated or characterized enhancer regions, and dark gray bands indicate promoters. Normalized ChIP-seq signals for each experiment are shown on the right. Conservation, mappability track, regulatory elements and genes on positive (+) or negative (–) strand are indicated below the ChIP-seq lanes. In **a** and **b** at the active *Cd4* and *Cd8* loci, GTFs and initiating (Ser5P) Pol II are detected at proximal enhancer (PE) and thymocyte enhancer (TE), as opposed to the distal enhancer (DE), elements (*Cd4*) and E_I, E_{II} and E_V (*Cd8*). At the *Il2ra* inactive locus (**c**), poised for transcription and activated either before or after the double-positive differentiation stage, Ser5P and GTFs are detected at a previously characterized enhancer (element V) region. The binding profiles of the remaining GTFs are shown in **Supplementary Figure 2a–c**.

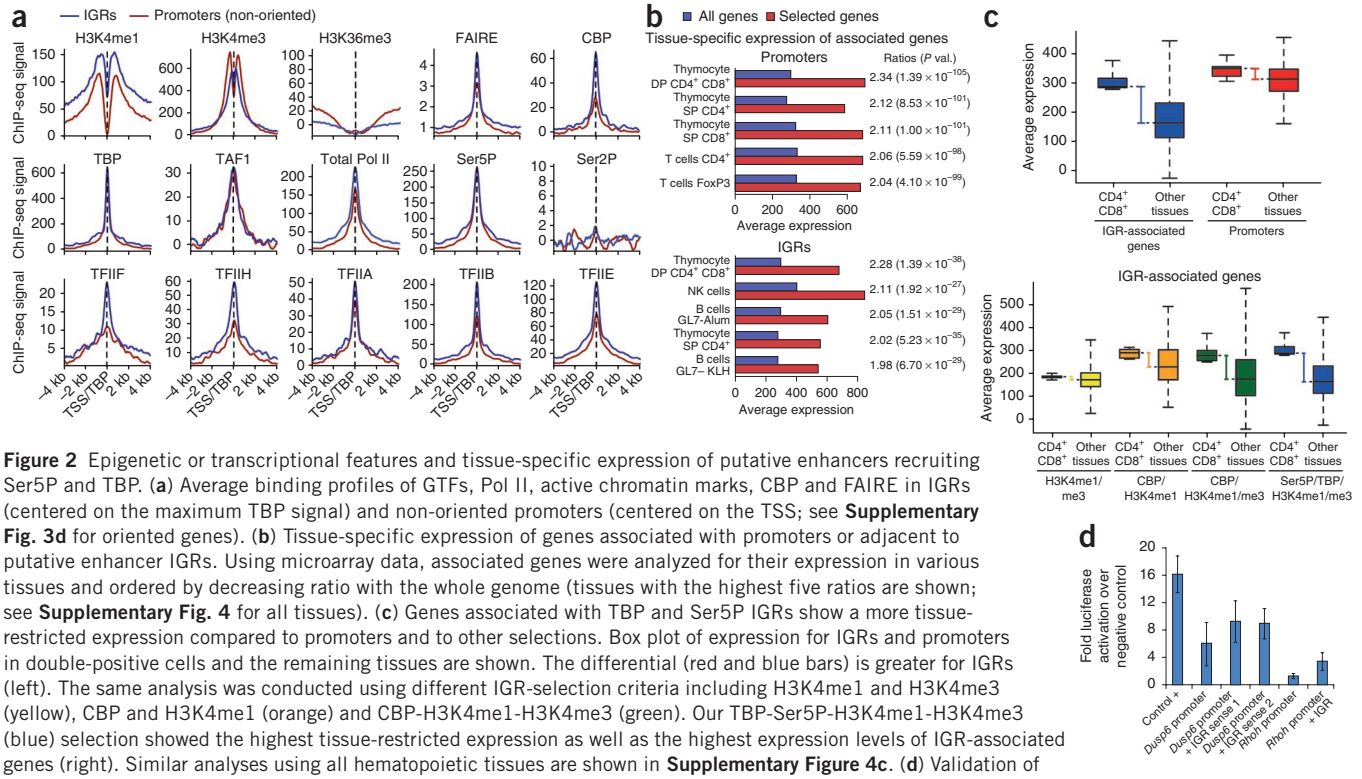


Figure 2 Epigenetic or transcriptional features and tissue-specific expression of putative enhancers recruiting Ser5P and TBP. (**a**) Average binding profiles of GTFs, Pol II, active chromatin marks, CBP and FAIRE in IGRs (centered on the maximum TBP signal) and non-oriented promoters (centered on the TSS; see **Supplementary Fig. 3d** for oriented genes). (**b**) Tissue-specific expression of genes associated with promoters or adjacent to putative enhancer IGRs. Using microarray data, associated genes were analyzed for their expression in various tissues and ordered by decreasing ratio with the whole genome (tissues with the highest five ratios are shown; see **Supplementary Fig. 4** for all tissues). (**c**) Genes associated with TBP and Ser5P IGRs show a more tissue-restricted expression compared to promoters and to other selections. Box plot of expression for IGRs and promoters in double-positive cells and the remaining tissues are shown. The differential (red and blue bars) is greater for IGRs (left). The same analysis was conducted using different IGR-selection criteria including H3K4me1 and H3K4me3 (yellow), CBP and H3K4me1 (orange) and CBP-H3K4me1-H3K4me3 (green). Our TBP-Ser5P-H3K4me1-H3K4me3 (blue) selection showed the highest tissue-restricted expression as well as the highest expression levels of IGR-associated genes (right). Similar analyses using all hematopoietic tissues are shown in **Supplementary Figure 4c**. (**d**) Validation of enhancer activity of two TBP and Ser5P IGRs in a promoter-dependent luciferase reporter assay. The *Dusp6* and *Rhoh* promoters and IGRs were cloned in a pGL3 vector and transfected in a T-cell line (EL4). The *Dusp6* IGR was cloned into both orientations and retained its ability to enhance promoter-driven expression. Error bars represent s.e.m. from two independent transfections. The complete experiment is presented in **Supplementary Figure 11c**.

Fig. 4c,**d**). Finally, we validated two putative enhancers in a promoter-dependent reporter assay for their activity (**Fig. 2d**).

To get insights into the TFBS composition of the selected IGRs, we divided TBP and Ser5P common peaks into ten ranks of TBP:Ser5P ratios. We found TATA boxes[28] enriched in the first two ranks at promoters but not at IGRs (**Supplementary Fig. 5a**). Independently, a *de novo* TFBS discovery on each rank showed the canonical TATA box as the most significant motif in promoters within the first rank (E-value = 5.3×10^{-54}) but not in IGRs (**Supplementary Fig. 5b**,**c**). Further analysis with all ranks indicated more specific TFBS, such as those of the Ets family, in both promoters and IGRs. To investigate Ets-related factors in double-positive cells, we conducted ChIP-seq for ETS1, an essential transcription factor during T-cell ontogenesis[29] and found its recruitment more pronounced at IGRs (46%) compared to promoters (9.6%) (**Supplementary Fig. 5d**). Finally, IGRs showed a marked enrichment for T cell–specific TFBS compared to non–T cell TFBS and increased conservation, though less pronounced compared to promoters[30] (**Supplementary Fig. 5e**). Overall, we showed that IGRs have different properties in core-promoter elements, TFBS composition and conservation, suggesting a different mode of recruitment of GTFs at promoters and enhancers.

Transcription with or without polyadenylation at enhancers

In ChIP assays, cross-linking artifacts—freezing of long-distance interactions between promoters and enhancers—might lead to an indirect signal of the general transcription machinery at these distal regulatory regions originating from promoters, whereas local transcription would argue for direct recruitment. We therefore assayed transcription at the *Tcrβ* and *Cd8* enhancers using QPCR and observed detectable levels of RNA transcripts (**Supplementary Fig. 6**). To extend this observation genome wide, we conducted strand-specific RNA-seq for total or polyadenylated (poly(A)) RNA. We used these data to combine the 708 TBP and Ser5P common peaks into 472 oriented transcribed areas (see **Supplementary Methods**). Our analysis revealed that most of the selected TBP and Ser5P IGRs showed detectable RNA signal (418 out of all 472). We further divided these regions into two distinct categories: 62% of the transcribed IGRs contained poly(A) RNA (for example, across the *Cd8* $E8_I$), whereas the remaining 38% showed only signal for total RNA (for example, on the *Cd4* proximal enhancer) (**Fig. 3a**). In some cases, we observed weak but appreciable RNA signal along the area linking the enhancers to the promoters and suggesting tracking of Pol II toward the gene (**Fig. 3a** and **Supplementary Fig. 7a**).

Average profiles of RNA-seq data in these regions showed that poly(A) RNAs originating from IGRs were mostly unidirectional—a feature comparable to what is seen at promoters. At IGRs without poly(A) RNA, however, we observed a substantial increase in the bidirectionality of transcription (**Fig. 3b** and **Supplementary Fig. 7b**,**c**). To further characterize differences between these groups, we conducted average profiling at IGRs centered on the main TBP peak and compared them to promoters centered on the TSS (**Fig. 4** and **Supplementary Fig. 7d**). As expected, H3K4me1:H3K4me3 ratios and binding levels of CBP and ETS1 were higher at IGRs compared to promoters, irrespective of RNA presence. Overall, however, these trends were more pronounced at IGRs without detectable poly(A) RNA. H3K36me3 levels were highest at genes and essentially absent at IGRs showing either no or only total RNA; IGRs containing poly(A) RNA showed intermediate levels. This result implies that polyadenylation of intergenic RNAs is either dependent on, or results in, the

deposition of transcription elongation marks outside of annotated genes. The sizes of detectable transcripts also discriminated IGR-associated RNAs, as the poly(A) RNAs reached at least 4 kb and the non-poly(A) RNAs 2–3 kb on average (data not shown). We validated enhancer activity in seven poly(A)- and three non-poly(A)-associated regions (**Supplementary Fig. 8a**,**b**), representative from our selection, in a promoter-independent luciferase assay. Activities were enhanced on average between 1.5- and 3.5-fold compared to levels with the SV40 promoter only (**Supplementary Fig. 8c**; error bars represent s.e.m. from two independent transfections). We find it interesting that the poly(A)-selected areas showed slightly higher activity than the non-poly(A)-selected areas.

Taken together, our analyses allow us to define three classes of putative enhancers or IGRs in T cells: transcribed with polyadenylation, transcribed without polyadenylation and untranscribed. These classes are characterized by enhanced ETS1 and CBP recruitment, by H3K4me1:H3K4me3 ratios as well as by a relative directionality of RNA, the latter being less pronounced at IGRs for the non-poly(A) population.

Genome-wide transcription initiation platforms

We observed that many locations recruiting TBP and Ser5P, including some of the previously selected IGRs, promoters and intragenic regions, are organized within wide arrays of variable size, representing large platforms of transcription initiation (**Fig. 5a**). Although it was previously shown that TSSs are often spread over areas of more than 100 bp[3], the TBP and Ser5P arrays we observed were up to several kilobases. To further characterize these areas, we first isolated 6,337 TBP-bound regions of varying sizes above 400 bp. We found that Ser5P also bound 65% of these (**Supplementary Fig. 9a**), and we used this subset, after removal of large gaps and divergent transcription units, to define transcriptional platforms for further analyses. This selection allowed for a stringent analysis, removing ambiguous areas with several platforms in close proximity as well as background signals. We found 71% of the platforms to be located in promoters (15% of them in divergent units), 12.5% in IGRs and 16.5% inside gene bodies. All categories showed comparable size distributions.

By sorting these areas according to width, we observed a large diversity of sizes ranging essentially from 0.45 kb to 10 kb (**Fig. 5b** and **Supplementary Fig. 9b**). Distribution analysis indicated that ~20% of these were >2 kb, and only very few were >10 kb. GTFs were also clearly enriched in these regions, whereas Ser2P and H3K36me3, marking transcription elongation, were essentially absent or depleted, indicating that Pol II was present in its initiating form. Most of these regions were associated with both H3K4me1 and H3K4me3, but the relative ratios of these two marks were inverted at IGRs and promoters (**Supplementary Fig. 9b**). In promoters, we mostly found H3K4me1 on the boundaries and outside, whereas most of H3K4me3 was located inside TBP and Ser5P platforms. This trend varied at IGRs, with more H3K4me1 signal on the inside. GTFs showed a localization pattern similar to the TBP and Ser5P selection. Moreover, transcripts could be detected surrounding platforms in both directions, starting from either boundary (**Fig. 5b** and **Supplementary Fig. 9c**), thereby demonstrating the importance

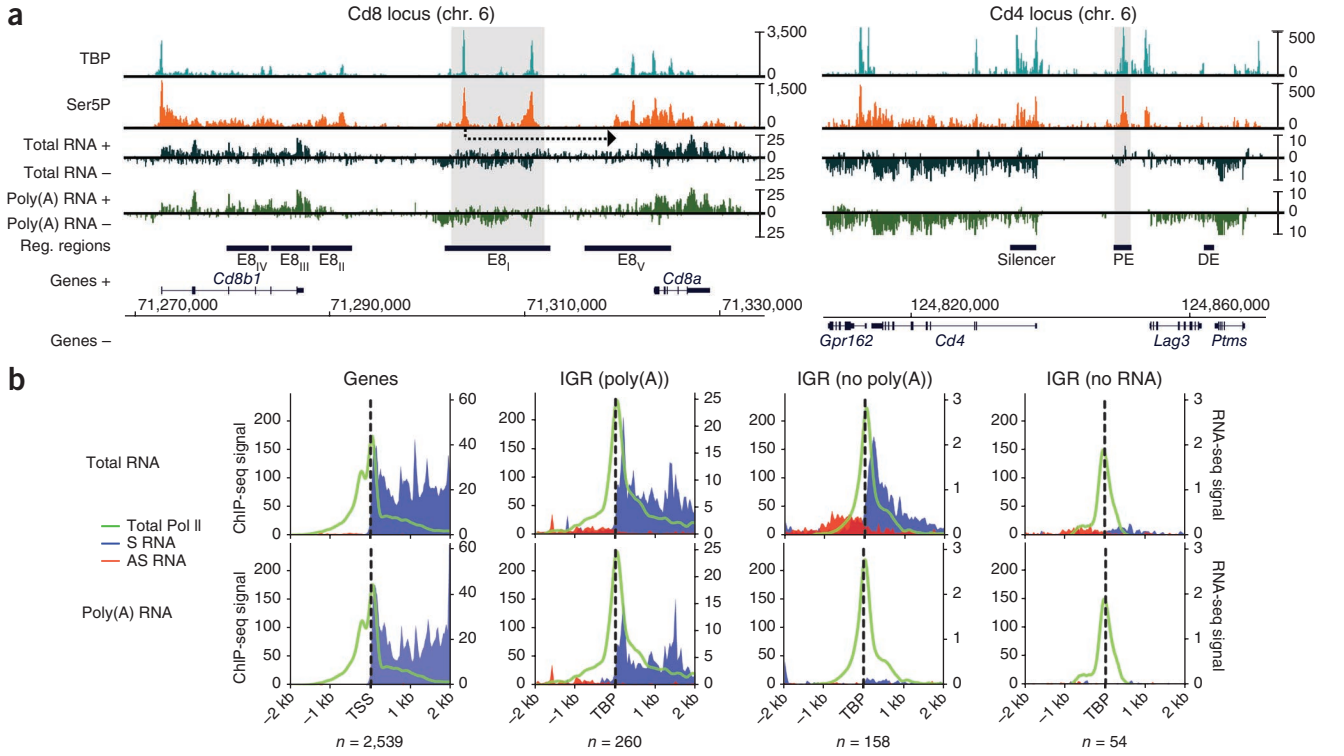


Figure 3 TBP and Ser5P enhancers are transcribed with or without polyadenylation. (**a**) Examples of known enhancers transcribed in the presence ($E8_I$ element of the *Cd8* locus, left) or absence (PE element of the *Cd4* locus, right) of polyadenylation signal. Total and poly(A) RNAs signals are represented below the TBP and Ser5P ChIP-seq lanes as $\log_2$ signals of the directional RNA-seq experiments. RNA strand orientation is indicated on the left. Transcribed enhancer elements are indicated by a light gray vertical bands. Additional examples of IGR transcription are shown in **Supplementary Figure 7**. Possible tracking of Pol II toward *Cd8a* is indicated by the dotted arrow below the Ser5P lane. (**b**) Pol II ChIP-seq and oriented RNA-seq average profiling on TBP and Ser5P enhancer IGRs or gene promoters (left panels) for total (top panels) or poly(A) RNAs (bottom panels). Selected TBP and Ser5P IGRs were divided into three populations associated with either poly(A), no poly(A) or no RNA, and orientation of the IGRs was established based on the RNA levels. Signals are centered on the TSS of genes or on the main TBP peak of IGRs.

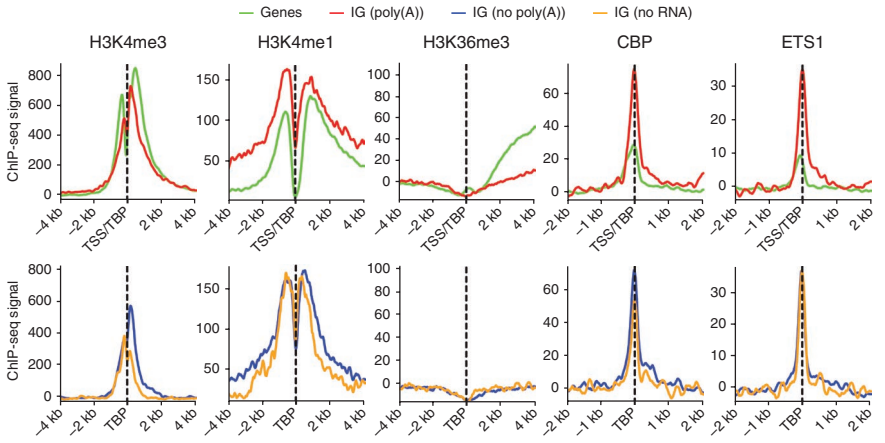


Figure 4 Poly(A) and non-poly(A) IGR subpopulations show distinct chromatin signatures between each other and genes. Comparison of active chromatin marks, CBP and ETS1 on oriented IGRs and promoters. ChIP-seq average binding profiles of the IGR populations described in **Figure 3b** are shown for genes and poly(A) (upper panels), no poly(A) and no RNA IGRs (lower panels). The profiles for the remaining factors described in this study are shown in **Supplementary Figure 7d**.

of these areas as structures that can delimit the start of transcription. Some platforms showed bidirectional transcription with limited overlap between strands (16–28%, **Supplementary Fig. 9d**). Therefore, although transcription initiation occurs primarily in a unidirectional fashion, bidirectionality can be detected at promoters and IGRs. In summary, we provide evidence for new genomic elements at promoters, IGRs and intragenic regions, which we define as transcription initiation platforms (TIPs), that allow extended recruitment of Pol II and GTFs, with a general preference for directionality of transcription.

TIPs are high in CpG and enriched at tissue-specific genes

Most promoters are known to be CpG rich[31], but because we observed TIPs at promoters as well as other genomic regions, we assayed each TIP for its CpG content. Noticeably, the isolated TIP areas overlapped with high CpG density, although the largest ones (>5 kb) showed more discontinuity (**Fig. 6a**). Even though overlap was observed at all TIP locations, it was greatest on promoters. This result indicates that high CpG content might favor recruitment and/or spreading of the transcriptional machinery over relatively large regions, irrespective of their location. Similarly, TFBS densities correlated with the presence of platforms at promoters and, to a lower extent, at intragenic or IGRs (**Fig. 6b** and **Supplementary Fig. 9e**).

We next analyzed the tissue specificity of genes associated with TIPs. Compared to our previous findings (**Fig. 2b**), we observed an increased double-positive–specific transcription profile on promoters (**Fig. 6c** and **Supplementary Fig. 10a**). We next divided promoter TIPs into size groups to investigate the possible effect on tissue specificity. We found it interesting that the tissue specificity, as measured by the rank of double-positive–expressed genes appearing within all tissues, increased with platform size (**Fig. 6d** and **Supplementary Fig. 10b**). The width of TIPs at promoters also showed a weak correlation with the expression levels of associated genes ($r = 0.33$, **Fig. 6e** and **Supplementary Fig. 10c**).

Investigation of Pol II ChIP-seq data from other laboratories showed that some TIPs isolated in double-positive T cells are conserved in various tissues and cell lines (**Supplementary Fig. 11a**). We further examined two promoter TIPs (controlling *Dusp6* and *Rhoh* genes) in a reporter assay for the importance of their orientation in driving transcription and their tissue specificity (**Supplementary Fig. 11b–d**). The *Rhoh* promoter TIP shows an enhancer-dependent activation only in a T-cell line (EL4), consistent with expression and available ChIP-seq data. Notably, *Dusp6*, which is expressed in both T cells and fibroblasts (NIH3T3), also activates the reporter in both cell lines, but the enhancer increases transcription only in T cells. This result is consistent with the absence of a Pol II-bound enhancer in fibroblasts. Although the *Dusp6* promoter can transcribe bidirectionally, it is only active in the sense orientation, suggesting that promoter TIPs are directional even in the presence of antisense transcription. Altogether, these analyses reveal a correlation between platform boundaries and CpG content that is irrespective of location, as well as a correlation with TFBS occurrence at promoters. Our findings also establish a link between platform size at promoters and tissue specificity, and they show the importance of directionality at the promoter despite the presence of bidirectional transcription.

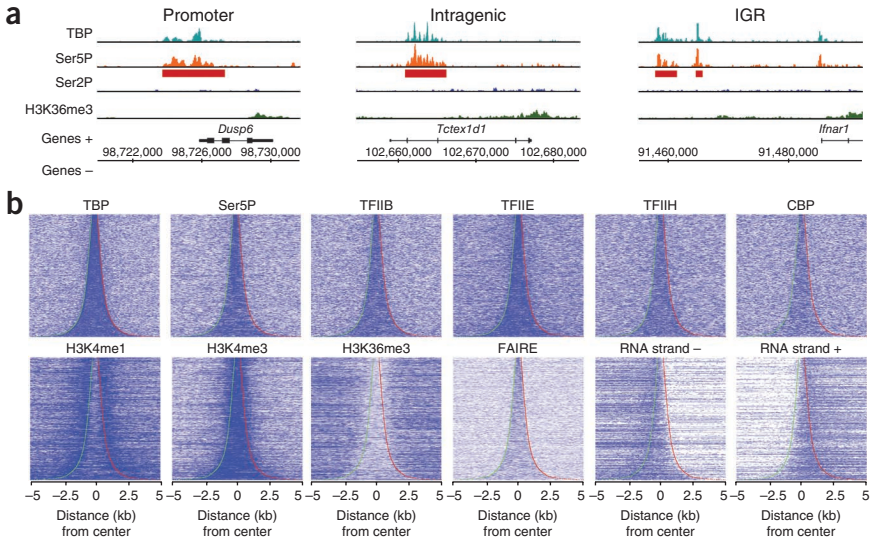


Figure 5 Pol II and GTFs transcription initiation platforms. (**a**) Examples of transcriptional initiation (TBP, Ser5P) or elongation (Ser2P, H3K36me3) hallmarks on TIPs at promoters, intragenic or IGR locations from left to right, respectively. The TBP and Ser5P TIPs isolated using a systematic approach (see Online Methods) are indicated by a red horizontal bar below TBP and Ser5P ChIP-seq signals. (**b**) Heatmaps of TIPs sorted by size and anchored on their center. TIPs (based on TBP and Ser5P selection) are shown for TBP, Ser5P and for the corresponding profiles for GTFs, active chromatin marks, FAIRE and the RNA-seq signal for positive and negative strands. TIP boundaries are represented by a red 5′ (right side) and green (left side) 3′ line. Heatmaps for all ChIP-seq and RNA-seq experiments at promoters, intragenic and IGR locations (as well as for input or mock Ig controls) are included in **Supplementary Figure 9b,c**.

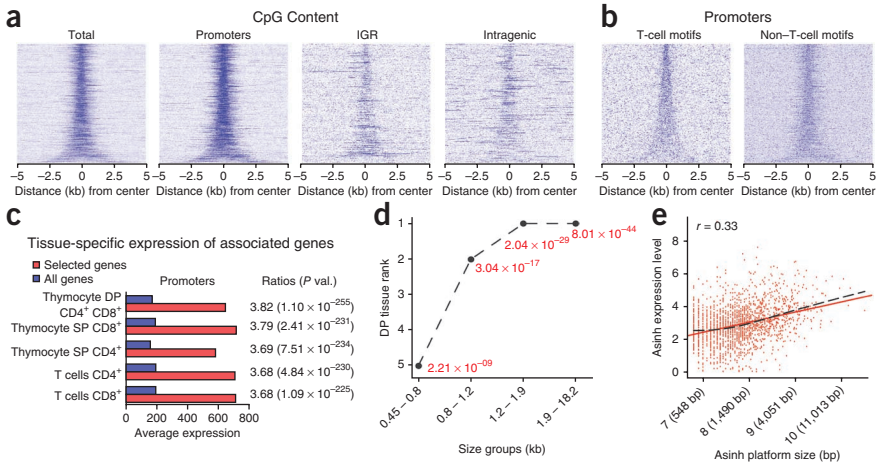
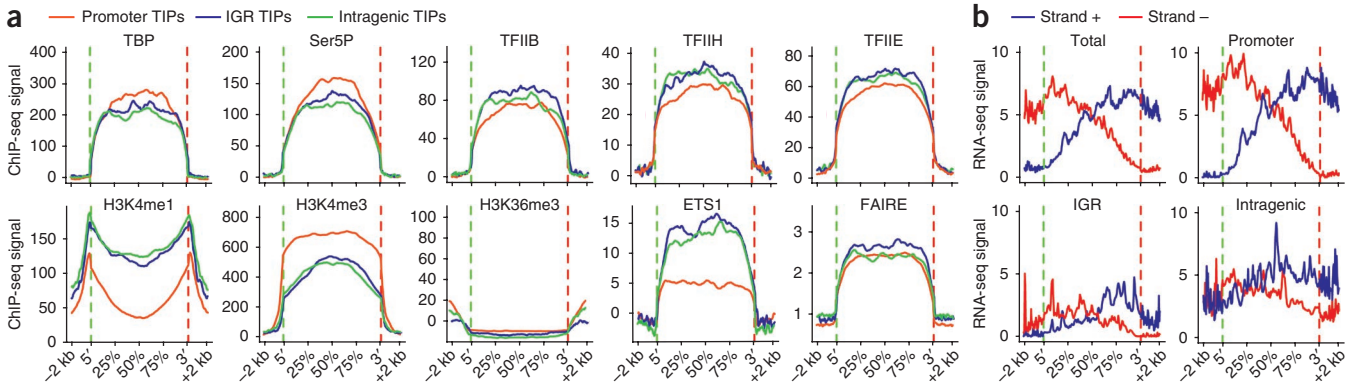


Figure 6 TIPs correlate with CpG content and tissue-specific expression at promoters. (**a**) CpG content across TIPs anchored on their center, similarly to **Figure 5b**. Promoters clearly show the highest CpG content. Similar trends are also visible in IGRs and, to a lesser extent, in intragenic regions. (**b**) Clustering of T-cell and non–T-cell transcription factor motifs at TIPs around promoters. Most putative TFBS overlap with the high CpG content from **a**. (**c**) Analysis of tissue specificity of expression for genes associated to promoter TIPs, similarly to **Figure 2b**. The associated genes show a more pronounced double-positive T-cell gene-expression pattern, indicating an increased tissue specificity at TIPs. The remaining genomic regions are analyzed in **Supplementary Figure 10a**. (**d**) Genes associated with promoter TIPs were classed into four equally sized groups (quartiles) with increasing platform size. The tissue specificity of the expression pattern increases with platform size, as indicated by the increasing rank and decreasing associated *P* values. The complete ranks are presented in **Supplementary Figure 10b**. (**e**) Correlation of TIP size to absolute gene expression levels at promoters ($r = 0.33$). The global and local fitted curves are represented by solid red and dashed black lines, repectively. Similar graphs for IGRs and intragenic regions are shown in **Supplementary Figure 10c**. TIP size and expression values were transformed using a hyperbolic arcsine (Asinh) function.

Differential recruitment and epigenetic features at TIPs

To further address the transcriptional and epigenetic patterns of TIPs, we conducted additional profiling across selected regions (**Supplementary Fig. 12a**). As expected from our previous results (**Fig. 2a**), all factors or marks associated with transcription initiation (GTFs, Pol II, H3K4me3) were enriched along the TIPs. The CBP and FAIRE signal followed similar trends, suggesting not only the presence of recruitment platforms but also of open chromatin areas.



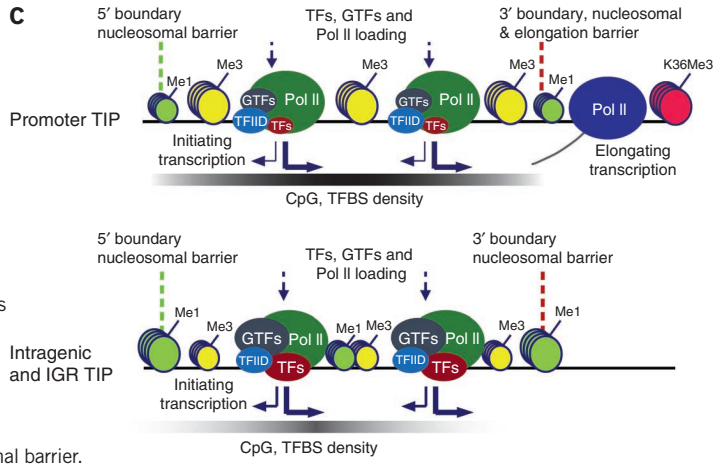


Figure 7 Average profiles of TIPs and model summarizing their features at distinct genomic locations. (**a**) Average profiles of TBP, Ser5P, TFIIB, TFIIH, TFIIE, active chromatin marks, ETS1 and FAIRE across all resized TIPs. Regions were divided into promoter (red), IGR (blue) and intragenic (green) locations (see **Supplementary Fig. 12a** for total profiles). In general, Ser5P, GTFs, ETS1, FAIRE and H3K4me3 are largely enriched throughout the platforms, H3K4me1 peaks just after the boundaries and H3K36me3 is depleted. IGR and intragenic profiles show mostly similar patterns, including lower TBP, Ser5P and H3K4me3 as well as higher TFIIH, TFIIE, H3K4me1 and ETS1 levels, compared to promoters. The remaining profiles are shown in **Supplementary Figure 12b**. (**b**) Total RNA signal across the different classes of TIPs. RNAs from the positive (blue) and negative (red) strands are shown. Transcription appears to start at either border, increases toward the opposite boundary and decreases again afterwards, indicative of a transcriptional barrier possibly imposed by H3K4me1. (**c**) Promoter, IGRs and intragenic TIPs are all characterized by open chromatin regions and are delimited by an enrichment of the H3K4me1 (Me1, green circles) histone mark (to a lesser extent at promoter), possibly reflecting a nucleosomal barrier. These areas have transcription initiation hallmarks such as Ser5P Pol II (green), H3K4me3 (Me3, yellow circles)—though less pronounced in IGRs and intragenic regions—and GTF recruitment in common. They differ in their relative proportions of ETS1 (designated here as TF) and GTFs at IGRs, compared to promoters. Promoters fundamentally differ from other TIPs in their ability to allow Pol II to enter elongation (blue), although Ser2P is not detected in the immediate proximity of TIPs (most likely because of higher elongation rate and less Ser2P accumulation at the 5′ ends). CpG and TFBS are more prominent at promoters, although TFs, such as ETS1, are more often recruited to enhancers. Bidirectional transcription is present at both promoter[37,38] and IGRs (**Fig. 3b** and **Supplementary Fig. 9d**).

In contrast, the H3K4me1 profile remained largely exclusive. Although we observed relatively low enrichment across the whole TIP, the highest levels—and, hence, possible nucleosome barriers—were observed just outside the boundaries, arguing for a delimitation of TIPs by this histone modification. H3K36me3 and Ser2P, associated with elongation, were absent or depleted from the TIPs. The depleted H3K36me3 profile showed an anti-correlation with CpG content, which is in agreement with a recent report showing that CpG islands recruit an H3K36 demethylase[32].

We next investigated the TIP distribution at promoters, intragenic regions and IGRs (**Fig. 7a** and **Supplementary Fig. 12b**). Consistent with our earlier observations (**Fig. 2a**), H3K4me1 levels were overall higher relative to those of H3K4me3 and more diffused on IGR TIPs, compared to promoters. Although TBP, TAF1, TFIIA and Ser5P levels were found to be equivalent or higher on promoters compared to IGR TIPs, other GTFs, ETS1 and CBP were more pronounced on IGRs, suggesting important differential properties and recruitment patterns between the two categories. All intragenic patterns followed those of the IGRs, with the exception of Ser2P, indicating that intragenic TIPs are most likely related to enhancer structures inside genes.

Finally, total RNA profiling confirmed that transcription initiation primarily takes place on either boundary (**Fig. 7b**). Transcript levels progressively increased toward the 3′ end, reached a plateau and decreased after the boundaries, suggesting a block of transcription by adjacent nucleosomes. Overall, TIP combinatorial binding and RNA initiation patterns together with CpG and TFBS content, suggest that these regions are the primary site of Pol II recruitment. They also reveal differential epigenetic, transcription factor and GTF recruitment between promoter and intragenic or IGR TIPs.

DISCUSSION

We now show that both Pol II and GTFs are recruited not only to known T cell–specific enhancers but also to numerous newly discovered putative regulatory regions. These findings extend previous observations on a few model enhancers[11,12,33,34] as well as two recent genome-wide investigations describing Pol II recruitment to enhancers in activated neurons and macrophages[14,15]. Intergenic Pol II recruitment also resembles a situation previously described in quiescent yeast[35,36]. Most of our selected putative enhancer regions were associated with transcripts, 60% of which were polyadenylated and 40% non-polyadenylated. This is in contrast with previous studies, where enhancer associated RNAs fell only into either category[14,15]. Our results show that poly(A) IGR RNAs are essentially directional, whereas non-poly(A) IGR RNAs are often bidirectional but to a lesser extent than in neurons. As our RNA-seq procedure was not optimized for the detection of short abortive transcripts, we speculate that most of the IGRs are transcribed bidirectionally in a paused state, as previously described for promoters[37,38]. Our results also suggest that Pol II and GTFs might be preloaded on tissue-specific enhancers before promoter translocation on a genome-wide scale, and in some cases evidence for Pol II tracking could be observed. Although TAF1 was previously found to bind a subset of ENCODE IGRs[6], this study is the first that introduces recruitment of all GTFs genome-wide to regulatory regions. Important questions regarding the role of intergenic enhancer transcription remain. In accordance with the work on macrophages[14], a subset of enhancers were transcribed and polyadenylated, including the $E8_I$ *Cd8* enhancer[24,25], and might correspond to noncoding genes associated with regulatory or other unknown functions. Further functional and sequence analyses might reveal a role for such transcripts. Recent work has described a population of new RNAs that themselves show enhancer function[39]. It is possible that at least a fraction of our TBP and Ser5P IGRs fall into this category.

We also found that initiating Pol II and TBP are often within large areas, suggesting loading and initiation platforms for the transcriptional machinery. TIPs restrict regions of transcription initiation at their boundaries, delimited by H3K4me1 nucleosomal barriers. We propose that initial transcription through the platforms results in open chromatin conformation, allowing for further recruitment of the transcriptional machinery. We provide evidence that TIPs are enriched for CpG, irrespective of their genomic location. This finding indicates that intrinsic DNA sequences might play a role in recruiting or propagating GTFs and Pol II at these locations. These features might allow for open chromatin conformation or could thermodynamically favor transcription-factor binding or transcription initiation, consistent with the overlap of TFBS. We observed that ETS1, a critical transcription factor in T cells, is more prominently recruited to IGR TIPs, although TFBS (including ETS1 sites) densities are more pronounced on promoters. Furthermore, we also observed differential recruitment patterns of TFIID, TFIIA and Pol II as compared to other GTFs on promoter and IGRs or intragenic TIPs. Altogether, promoters and enhancers share common (TFIID, TFIIA and Pol II recruitment and open chromatin) and distinct (TFs, GTFs recruitment and epigenetic marks) characteristics (**Fig. 7c**) that might reflect mechanistic differences in transcription initiation and could be used for the predictive discrimination of regulatory elements.

We showed that promoters containing TIPs regulate highly tissue-specific genes and that this trend increases with platform width. This contrasts with previous findings where alternative TSS usage was attributed to ubiquitously expressed genes[3]. Although we cannot conclude whether the presence of these platforms is a cause or a consequence of tissue specificity, we speculate that they represent genomic elements that overcome the rate-limiting step of Pol II recruitment in order to sustain high levels of transcription. Overall, our observations introduce the concept of specific and tightly regulated TIPs as genomic hallmarks for regulation and maintenance of tissue-specific gene activity.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Accession codes. Gene Expression Omnibus: GSE29362 for all ChIP-seq, RNA-seq and FAIRE data.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

Work in the P.F. laboratory is supported by institutional grants from Institut National de la Santé et de la Recherche Médicale and the Centre National de la Recherche Scientifique (CNRS), and by specific grants from the Fondation Princesse Grace de Monaco, the Agence Nationale de la Recherche (ANR), the Institut National du Cancer (INCa) and the Commission of the European Communities. F.K. was supported by grants from Chromatin Plasticity, Marie Curie Research Training Network and Association pour la Recherche sur le Cancer, R.F. by Genopole and CNRS, and P.C. by grants from INCa and Fondation pour la Recherche Médicale. The work was also supported by a Regulome grant from the ANR. D.E. was supported by Deutsche Forschungsgemeinschaft, Transregio-5. We are grateful to B. Escaliere for useful advice on the statistical analyses, to J.J. Waterfall and J.L. Core from the Lis lab (Cornell University, Ithaca, USA) for help in the generation of the mappability track, to G. Natoli (European Institute of Oncology) for the gift of plasmids used in preliminary experiments for reporter assays, to E. Soucie and V. Cauchy for critical reading of the manuscript, to J. Blanc for technical assistance, to Y. Duffourd from the Centre National de Génotypage-Commissariat à l'Energie Atomique lab for sequencing quality controls and to members of the P.F. lab for help and advice. We dedicate this work to the memory of distinguished colleague Vanessa Ranc-Rongere, who left us too early.

AUTHOR CONTRIBUTIONS

J.-C.A., F.K., T.K.A., P.F. and I.G. conceived the framework of the study. J.-C.A. and F.K. designed the experiments. R.F., P.C. and F.K. carried out the bioinformatic analyses and data treatment. D.E., C.H. and M.H. produced and provided the Ser2P and Ser5P antibodies as well as other antibodies that were not presented in this study. All ChIP-seq and RNA-seq materials were prepared by F.K. with the exception of ETS1 ChIP-seq, which was prepared by P.C., M.G. and I.G. conducted all ChIP-seq and RNA sequencing experiments. J.Z.-C. and S.S. did the FAIRE experiment. F.K. did the cloning and luciferase experiments and A.L.d.l.C. participated and provided technical assistance. J.-C.A. wrote the manuscript, and F.K., R.F. and P.C. participated in its preparation. All authors reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Sikorski, T.W. & Buratowski, S. The basal initiation machinery: beyond the general transcription factors. *Curr. Opin. Cell Biol.* **21**, 344–351 (2009).
2. Buratowski, S. Progression through the RNA polymerase II CTD cycle. *Mol. Cell* **36**, 541–546 (2009).
3. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
4. Koch, F., Jourquin, F., Ferrier, P. & Andrau, J.C. Genome-wide RNA polymerase II: not genes only!. *Trends Biochem. Sci.* **33**, 265–273 (2008).
5. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
6. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
7. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40**, 897–903 (2008).
8. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
9. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
10. Xi, H. *et al.* Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.* **3**, e136 (2007).
11. Higgs, D.R., Vernimmen, D. & Wood, B. Long-range regulation of alpha-globin gene expression. *Adv. Genet.* **61**, 143–173 (2008).
12. Fromm, G. & Bulger, M. A spectrum of gene regulatory phenomena at mammalian beta-globin gene loci. *Biochem. Cell Biol.* **87**, 781–790 (2009).
13. Szutorisz, H., Dillon, N. & Tora, L. The role of enhancers as centres for general transcription factor recruitment. *Trends Biochem. Sci.* **30**, 593–599 (2005).
14. De Santa, F. *et al.* A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol.* **8**, e1000384 (2010).
15. Kim, T.K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
16. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
17. Bosselut, R. CD4/CD8-lineage differentiation in the thymus: from nuclear effectors to membrane signals. *Nat. Rev. Immunol.* **4**, 529–540 (2004).
18. Anderson, M.K. At the crossroads: diverse roles of early thymocyte transcriptional regulators. *Immunol. Rev.* **209**, 191–211 (2006).
19. Sawada, S. & Littman, D.R. Identification and characterization of a T-cell-specific enhancer adjacent to the murine CD4 gene. *Mol. Cell Biol.* **11**, 5506–5515 (1991).
20. Cherrier, M., D'Andon, M.F., Rougeon, F. & Doyen, N. Identification of a new cis-regulatory element of the terminal deoxynucleotidyl transferase gene in the 5′ region of the murine locus. *Mol. Immunol.* **45**, 1009–1017 (2008).
21. Kaufmann, C. *et al.* A complex network of regulatory elements in Ikaros and their activity during hemo-lymphopoiesis. *EMBO J.* **22**, 2211–2223 (2003).
22. Georgopoulos, K., van den Elsen, P., Bier, E., Maxam, A. & Terhorst, C.A. T cell-specific enhancer is located in a DNase I-hypersensitive area at the 3′ end of the CD3-delta gene. *EMBO J.* **7**, 2401–2407 (1988).
23. Yannoutsos, N. *et al.* A cis element in the recombination activating gene locus regulates gene expression by counteracting a distant silencer. *Nat. Immunol.* **5**, 443–450 (2004).
24. Hostert, A. *et al.* Hierarchical interactions of control elements determine CD8alpha gene expression in subsets of thymocytes and peripheral T cells. *Immunity* **9**, 497–508 (1998).
25. Ellmeier, W., Sunshine, M.J., Losos, K. & Littman, D.R. Multiple developmental stage-specific enhancers regulate CD8 expression in developing thymocytes and in thymus-independent T cells. *Immunity* **9**, 485–496 (1998).
26. Schmidl, C. *et al.* Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res.* **19**, 1165–1174 (2009).
27. Schorle, H., Holtschke, T., Hunig, T., Schimpl, A. & Horak, I. Development and function of T cells in mice rendered interleukin-2 deficient by gene targeting. *Nature* **352**, 621–624 (1991).
28. Shi, W. & Zhou, W. Frequency distribution of TATA Box and extension sequences on human promoters. *BMC Bioinformatics* **7** (suppl. 4), S2 (2006).
29. Eyquem, S., Chemin, K., Fasseu, M. & Bories, J.C. The Ets-1 transcription factor is required for complete pre-T cell receptor function and allelic exclusion at the T cell receptor beta locus. *Proc. Natl. Acad. Sci. USA* **101**, 15712–15717 (2004).
30. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
31. Saxonov, S., Berg, P. & Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA* **103**, 1412–1417 (2006).
32. Blackledge, N.P. *et al.* CpG islands recruit a histone H3 lysine 36 demethylase. *Mol. Cell* **38**, 179–190 (2010).
33. Spicuglia, S. *et al.* Promoter activation by enhancer-dependent and -independent loading of activator and coactivator complexes. *Mol. Cell* **10**, 1479–1487 (2002).
34. Ho, Y., Elefant, F., Cooke, N. & Liebhaber, S. A defined locus control region determinant links chromatin domain acetylation with long-range gene activation. *Mol. Cell* **9**, 291–302 (2002).
35. Andrau, J.C. *et al.* Genome-wide location of the coactivator mediator: Binding without activation and transient Cdk8 interaction on DNA. *Mol. Cell* **22**, 179–192 (2006).
36. Radonjic, M. *et al.* Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon *S. cerevisiae* stationary phase exit. *Mol. Cell* **18**, 171–183 (2005).
37. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
38. Seila, A.C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
39. Ørom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).

ONLINE METHODS

Oligonucleotides. All oligonucleotides used in this study for ChIP- and RT-QPCR, as well as for cloning , are summarized in **Supplementary Table 1**.

Cell sorting. We isolated thymuses from 5–6-week-old C57BL/6 wild-type mice. After homogenization, we sorted for $CD4^+$ $CD8^+$ double-positive cells using the AutoMACS cell sorter (Miltenyi) using subsequent CD8 and CD4 positive selections. In brief, cells were stained with CD8-R-phycoerythrin antibodies and sorted using anti-phycoerythrin multisort beads (Miltenyi). After release from the beads, cells were sorted again by positive selection using CD4 beads (Miltenyi). For biological replicates, litter mates of the same age were used. The purity of the sorted double-positive population was assessed using FACS analysis (see **Supplementary Fig. 1b**).

ChIP-seq and FAIRE. ChIPs were essentially carried out as previously described[40]. All antibodies used in this study and their ChIP conditions are shown in **Supplementary Table 2**. Phosphatase inhibitors (Roche, France) were added to a final concentration of 1× to all buffers for phosphoserine ChIPs. Sonication was conducted using a Misonix 4000 (Misonix) sonicator for 10 cycles (30 s on, 30 s off, amplitude 40), resulting in sheared DNA between 100 bp and 400 bp with the bulk at ~250 bp (see **Supplementary Fig. 1c**). Sample preparation for FAIRE was carried out essentially as previously described[16]. Eluted DNA was quantified either by Picogreen (Invitrogen) or using DNA High Sensitivity chips on a Bioanalyzer (Agilent). The sequencing procedure was conducted using at least 1 ng of starting material and run on a Genome Analyzer II (Illumina), according to manufacturer's instructions. The computational processing and analysis pipeline is described in **Supplementary Methods**. Tag numbers, extension sizes and replicate correlations of each experiment are shown in **Supplementary Table 2**. All data were treated and further analyzed as 50-bp window-averaged wiggle files. The Integrated Genome Browser (IGB)[41] was used to visualize the data and export screenshots.

Strand-specific RNA-seq. Total RNA was isolated from sorted cells using Trizol (Invitrogen) according to manufacturer's instructions. For total RNA sequencing, the ribosomal RNA of 8 μg of total RNA was depleted using the eukaryotic RiboMinus kit (Invitrogen). For poly(A) RNA sequencing, RNA was purified using the Illumina poly(A) purification kit. Both samples were fragmented to 150 bp using RNase III (Ambion) and processed using the Illumina small RNA kit with some modifications (see **Supplementary Methods**). The resulting complementary DNA (cDNA) was sequenced on an Illumina Genome Analyzer II, using the small RNA sequencing kit.

Peak detection. Peak detection was carried out with CoCAS[42], using data averaged into 10 bp windows and converted to general feature format (gff). All peak detection parameters and number of peaks obtained are summarized in **Supplementary Table 2**. Peaks −5 kb to +5 kb outside of known annotations (refSeq, miRNA, rRNA, scRNA, snRNA, snoRNA and tRNA) were considered to be intergenic. Peaks within −2 kb to +1 kb from respective transcription start sites were considered to be inside promoters.

Statistical analysis. To analyze the tissue-specific expression of genes isolated in **Figures 2b**,**c** and **6c**, we compared their expression levels to all genes in every tissue (**Supplementary Fig. 4**). Using the bioGPS website (http://biogps.org/), we obtained the GeneAtlas–averaged mouse dataset containing normalized genome-wide expression values in 96 tissues[43]. This reference file provides Affymetrix probe values in every available tissue based on the MOE430 2.0 array design. As several Affymetrix probes can refer to the same gene annotation, an average of the probes was assigned to each gene. Two different gene sets were built from this expression dataset, each providing information on gene expression in every tissue: (i) a whole-genome control set and (ii) the promoter- and IGR-associated gene sets. A nonparametric statistical Kruskal-Wallis test was conducted to estimate the significance of the difference in average expression levels between whole-genome sets and selected genes. Bars of expression levels in tissues were then sorted by their expression-level ratios between the selected datasets and by their *P* values estimating the significance of their differential expression (see **Supplementary Fig. 4**).

Motif discovery and density. We used MEME (Multiple Em for Motif Elicitation)[44] to conduct an unbiased *de novo* motif search within repeat-masked regions of interest. Motifs were identified using the Jaspar[45] database. Similarly, we used DNA-pattern of Regulatory Sequence Analysis Tools (RSAT)[46] to scan these regions for the canonical (TATAWAAG) and degenerate (TATAW) TATA-motifs. Motif densities were obtained using Matrix-Scan of RSAT for either 18 T cell–specific (expressed either in thymus or mature T lymphocytes) or 110 non-T cell–specific motifs from the Jaspar database (http://jaspar.genereg.net/).

TIP isolation. TIPs were isolated using peak detection of the 200-bp binned TBP signal, as previously described. This data allowed us to score for large, enriched regions by using a relatively low extension threshold (see **Supplementary Methods**). 50-bp binned wiggle files of the TBP signal were used to adjust the boundary coordinates of the TIPs. We also removed regions smaller than 400 bp from selection, and a final filtering step was applied to the isolated platforms in order to remove the ones showing large gaps in the TBP signal (more than eight contiguous bins with no signal).

Luciferase reporter assay. Luciferase assays were conducted in either T-cell (EL4), macrophage (RAW 264.7) or fibroblast (NIH3T3) cell lines. Promoters and IGRs were cloned together into pGL3 basic vectors (Promega), or IGRs only into pGL3 promoter vectors, including different orientations. We co-transfected pRL *Renilla* luciferase vectors that we used as internal controls and expression was calculated as a fold enrichment over a normalized negative control (empty basic vector). When indicated, transfections were carried out in duplicate.

40. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).
41. Nicol, J.W., Helt, G.A., Blanchard, S.G. Jr., Raja, A. & Loraine, A.E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* **25**, 2730–2731 (2009).
42. Benoukraf, T. *et al.* CoCAS: a ChIP-on-chip analysis suite. *Bioinformatics* **25**, 954–955 (2009).
43. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10**, R130 (2009).
44. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
45. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
46. Thomas-Chollier, M. *et al.* RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.* **36**, W119–W127 (2008).