nature genetics

Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers

Leighton J Core[1,3,4], André L Martins[2,4], Charles G Danko[2,3], Colin T Waters[1,3], Adam Siepel[2,3] & John T Lis[1]

Despite the conventional distinction between them, promoters and enhancers share many features in mammals, including divergent transcription and similar modes of transcription factor binding. Here we examine the architecture of transcription initiation through comprehensive mapping of transcription start sites (TSSs) in human lymphoblastoid B cell (GM12878) and chronic myelogenous leukemic (K562) ENCODE Tier 1 cell lines. Using a nuclear run-on protocol called GRO-cap, which captures TSSs for both stable and unstable transcripts, we conduct detailed comparisons of thousands of promoters and enhancers in human cells. These analyses identify a common architecture of initiation, including tightly spaced (110 bp apart) divergent initiation, similar frequencies of core promoter sequence elements, highly positioned flanking nucleosomes and two modes of transcription factor binding. Post-initiation transcript stability provides a more fundamental distinction between promoters and enhancers than patterns of histone modification and association of transcription factors or co-activators. These results support a unified model of transcription initiation at promoters and enhancers.

Regulation of RNA transcription is a critical process for directing cell fates during organismal development and is necessary to maintain homeostasis throughout the lifespan of all organisms. Promoters and enhancers are major control hubs for gene regulation that integrate information from a multitude of signaling pathways through the binding of signal-responsive activators and repressors. Therefore, accurately mapping and characterizing these regulatory regions is essential for defining how cell-specific transcriptomes are generated and maintained.

In mammalian cells, transcription initiation at the promoters of annotated genes is accompanied by upstream antisense transcription initiation[1–3]. The divergent TSSs are tightly spaced (<250 bp apart) and are presumed to arise from separate core promoters. The transcript representing the gene is typically stable and is thus detected by standard RNA sequencing (RNA-seq) techniques. In contrast, the upstream, antisense RNA (uaRNA) is typically short and more difficult to detect owing to a polyadenylation site (PAS)-dependent termination mechanism that rapidly targets the transcript for degradation by the exosome[4,5]. Occasionally, the uaRNA appears to be replaced with that for another mRNA[6], a long intergenic noncoding RNA (lincRNA) or a tRNA gene[7] to produce a pair of stable transcripts. Nearly 80% of active mammalian promoters display a bidirectional arrangement of initiation; thus, this back-to-back arrangement of initiation has emerged as a general feature of promoters[2].

Transcription initiation also occurs at enhancers. Although such transcription was originally identified at several canonical enhancers, more recent high-throughput sequencing methods have demonstrated enhancer transcription to be widespread[8–11]. Production of enhancer RNAs (eRNAs) is also bidirectional and is associated with chromatin modifications or binding of cofactors that are suggestive of enhancer activity (monomethylation of histone H3 at lysine H4 (H3K4me1), p300 binding and acetylation of histone H3 at lysine 27 (H3K27ac))[12–14]. The widespread existence of eRNAs and uaRNAs raises several important questions regarding how these RNAs are produced and whether they are functional. For example, is the initiation of eRNA transcription governed by the same rules as transcription at promoters? RNA polymerase II (Pol II) can operate with lower stringency when encountering naked DNA[15]; thus, it is possible that Pol II initiates at enhancers by virtue of the open chromatin environment and high local concentration of Pol II, rather than as part of a bona fide preinitiation complex. Additionally, some studies suggest that eRNAs are important for the activation of target genes[16,17], whereas others suggest that eRNA production is dispensable in constructing

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA. [2]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA. [3]Present addresses: Department of Molecular and Cell Biology, Institute for Systems Genomics, University of Connecticut, Storrs, Connecticut, USA (L.J.C.), Baker Institute for Animal Health, Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York, USA (C.G.D.), Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA (C.T.W.) and Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA (A.S.). [4]These authors contributed equally to this work. Correspondence should be addressed to J.T.L. (jtl10@cornell.edu) or A.S. (acs4@cornell.edu).

Received 20 May; accepted 15 October; published online 10 November 2014; doi:10.1038/ng.3142

a functional enhancer[9]. Furthermore, the process of transcription itself may be functional through the modification of chromatin architecture or creation of negative supercoils that enhance transcription factor binding[18].

Although divergent transcription at promoters and enhancers remains incompletely understood, it is nevertheless a characteristic signature that can be exploited in the identification of active regulatory elements[9,19,20]. The signature of divergent transcription is particularly evident when transcriptional activity is assayed using the global nuclear run-on sequencing (GRO-seq) method, owing to the high sensitivity of this method for all transcriptionally engaged RNA polymerase molecules regardless of subsequent transcript turnover rates[2,9,19]. In addition, a variation of the GRO-seq method that enriches for 5′-capped (m^7G) RNAs (GRO-cap) can greatly increase the sensitivity and specificity for detecting transcription initiation[21,22] (Online Methods). In this article, we apply this GRO-cap method to human cells and show that it efficiently and precisely maps the TSSs of coding and noncoding RNAs, regardless of the subsequent stability of the transcripts. Thus, GRO-cap provides a more complete picture of genome-wide initiation than CAGE (cap analysis of gene expression), which mainly detects TSSs resulting in stable RNAs[20,23]. Using our comprehensive, GRO-cap–based annotations of TSSs, we then report a detailed analysis of transcription initiation sites that sheds new light on the architecture of both promoters and enhancers across the human genome.

RESULTS

Identification of TSSs in human cells using GRO-cap

We prepared GRO-cap and GRO-seq libraries from human lymphoblastoid B cell (GM12878) and chronic myelogenous leukemic (K562) cell lines and PRO-seq (high-resolution GRO-seq[22]) data from K562 cells (**Supplementary Table 1**). Both cell lines are 'Tier 1' cell lines in the Encyclopedia of DNA Elements (ENCODE) Project, allowing us to take advantage of abundant publicly available functional genomic data[24]. The GRO-cap assay efficiently captured TSS information from nascent transcripts, as evidenced by a dramatic enrichment of GRO-cap signal at gene promoters and enhancers (**Fig. 1a**,**b** and **Supplementary Fig. 1a**). **Figure 1a** shows a specific example of the classic globin locus where divergent transcription is seen from active regions, including the ε-globin gene (*HBE1*) and the upstream hypersensitive sites that mark enhancers[25].

To comprehensively identify candidate TSSs using our data, we developed a hidden Markov model (HMM) that contrasted GRO-cap data with those from control experiments in which the critical cap-removing enzyme, tobacco acid pyrophosphatase (TAP), was omitted (Online Methods and **Supplementary Fig. 2a**,**b**). The HMM identified a total of ~120,000 putative TSSs in each cell line, within the range previously reported (80,000–150,000)[8,10,26,27]. The predicted TSS regions were narrow (mean of 57 bp in length, with 95% under 140 bp) but accounted for 69% of all GRO-cap TAP^+ reads and included both sharp and more dispersed TSSs (Online Methods and **Supplementary Fig. 2c**,**d**). Ninety-three percent of these TSSs were contained within enhancer or promoter regions predicted from patterns of histone modification (ChromHMM regions) in the same cell types[28]. However, our mapping of these regions was more stringent and localized, identifying ~4-fold fewer regions with ~3-fold higher resolution per site than combined ChromHMM promoter and enhancer predictions (**Supplementary Fig. 2e**,**f**).

In comparison to CAGE, GRO-cap resulted in a similar composite profile when reads were aligned to annotated gene TSSs (**Fig. 2a**). However, fewer reads mapped to introns and internal exons, indicating that GRO-cap has reduced background in comparison to CAGE (**Fig. 2a**,**b**). The decreased background for GRO-cap results in part from differences in the methodologies (cap-trapping[29] versus the oligonucleotide-capping method[30]) used to capture capped transcripts. GRO-cap has the additional strength that it is highly sensitive to rare or rapidly degraded noncoding RNAs because it

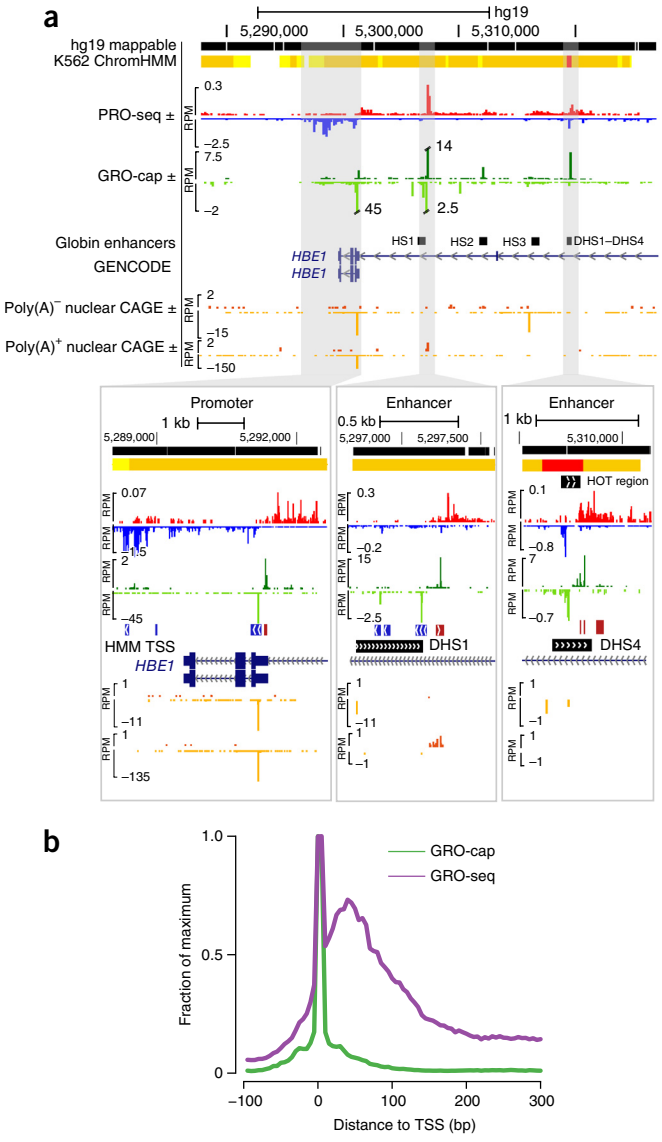

Figure 1 GRO-cap identifies TSSs in promoters and enhancers. (**a**) A UCSC Genome Browser[54] shot of the globin locus near the locus control region (LCR) using K562 cell line data sets generated or used in this study. The locus contains a portion of the β-globin locus, including the ε-globin gene (*HBE1*) and LCR enhancers. The insets are zoomed-in views of the shaded regions that show divergent GRO-cap signal (dark green, plus strand; light green, minus strand) at the *HBE1* promoter (left) and two enhancers associated with DHS1 (center) and DHS4 (right). The locations of the DHS sites are taken from probe locations in Ashe *et al.*[55]. The ChromHMM regions track is shown on top, with predicted promoters indicated in red and enhancers indicated in orange. Note that CAGE signal (dark orange, plus strand; light orange, minus strand) is at background levels in the enhancer region. RPM, reads per million. (**b**) GRO-cap dramatically enriches the signal for initiation sites in comparison to GRO-seq. Composite GRO-seq and GRO-cap reads from the cell line are plotted relative to all GENCODE TSSs.

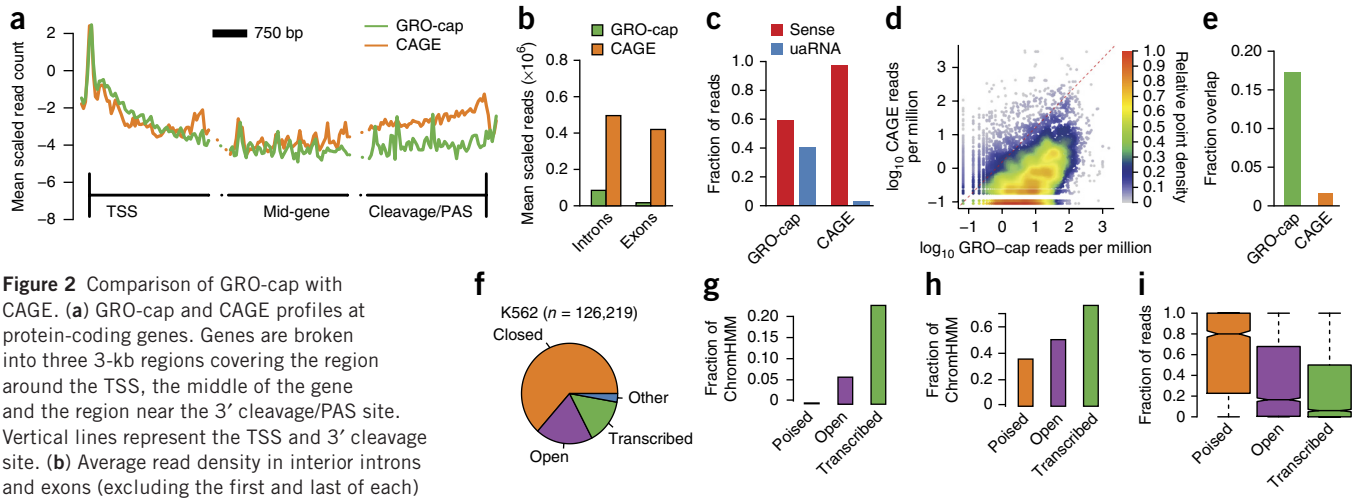

Figure 2 Comparison of GRO-cap with CAGE. (**a**) GRO-cap and CAGE profiles at protein-coding genes. Genes are broken into three 3-kb regions covering the region around the TSS, the middle of the gene and the region near the 3′ cleavage/PAS site. Vertical lines represent the TSS and 3′ cleavage site. (**b**) Average read density in interior introns and exons (excluding the first and last of each) as a measure of GRO-cap and CAGE background signals. (**c**) GRO-cap and CAGE relative fraction of reads aligned to sense and divergent (uaRNA) directions at protein-coding genes (counted within the underlying ChromHMM region). (**d**) Density scatterplot showing the signal intensity (reads per million) for GRO-cap versus CAGE surrounding distal transcription factor ChIP-seq peaks from the Hudson Alpha Institute for Biotechnology. (**e**) Fraction of ChromHMM regions containing a detectable GRO-cap (green) or CAGE (orange) TSS. (**f**) Comparing enhancer regions on the basis of chromatin marks (ChromHMM enhancers; Ernst *et al.*[28]) with DHSs (OpenChromatin Consortium) and GRO-cap reads identifies 3 main classes of enhancer regions, poised (no DHS peak or GRO-cap TSS; $n = 80{,}117$), open (DHS peak but no GRO-cap TSS; $n = 24{,}263$) and transcribed (DHS peak and GRO-cap TSS; $n = 18{,}367$), and a negligible 'other' group (no DHS peak but GRO-cap TSS; $n = 3{,}472$). (**g**–**i**) The three classes represent a progression in terms of functional activity, as measured by an increase in detectable transcription factor footprints (Wellington footprints on DHSs) (**g**), an increase in chromatin links (ChIA-PET overlap) (**h**) and a significant reduction in CpG methylation (**i**), between each transition (poised versus open, $P < 2.2 \times 10^{-16}$; open versus transcribed, $P < 3.448 \times 10^{-16}$; Mann-Whitney test). The center line of the boxplot represents the median, the boxes encompass the interquartile range and the whiskers extend to the minimum and maximum.

captures nascent RNAs as they are being made and before events that determine stability occur[4,21]. This feature also eliminates background from post-transcriptionally capped RNAs[31]. In contrast, CAGE libraries are often dominated by highly abundant and stable RNAs (for example, mRNAs), resulting in decreased sensitivity to unstable RNAs[5,23,31], such as uaRNAs at protein-coding promoters (**Figs. 1a** and **2c**). The high sensitivity and low background of GRO-cap also contribute to an increased coverage of enhancer regions predicted from histone modification patterns[28] (**Figs. 1a** and **2d**,**e**, and **Supplementary Fig. 3a**–**c**). As expected, GRO-cap signal correlated better with polymerase levels measured by PRO-seq in promoter-proximal regions than in the gene body (**Supplementary Fig. 3d**), suggesting that the signal originates primarily from nascent RNAs associated with polymerases that are paused proximal to promoters. Although this means that GRO-cap data cannot be used on its own as a measure of either initiation rates or levels of transcription elongation, GRO-cap does comprehensively map TSS locations regardless of the eventual stability of the RNA.

We characterized putative enhancers captured by GRO-cap by contrasting our TSSs that were not at annotated genes with ChromHMM enhancers and open chromatin (DNase I–hypersensitive (DHS)) regions. This three-way comparison subdivided ChromHMM enhancers into three main classes: closed (ChromHMM only), open (ChromHMM and DHS) and transcribed (ChromHMM, DHS and GRO-cap) TSSs (**Fig. 2f**). The transcribed subset, our main focus in this study, was enriched for positive regulatory activity, namely increased transcription factor binding (Wellington footprints[32]; **Fig. 2g**), distal chromatin interactions (chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[33]; **Fig. 2h**) and reduced CpG methylation[34] (**Fig. 2i**). In addition, the various histone modifications differed in expected patterns among poised, open and transcribed enhancers (**Supplementary Fig. 4**). These results suggest that our approach identifies, with high resolution, a subset of the sites identified by other methods, with theses sites appearing to be enriched for active roles in transcriptional regulation.

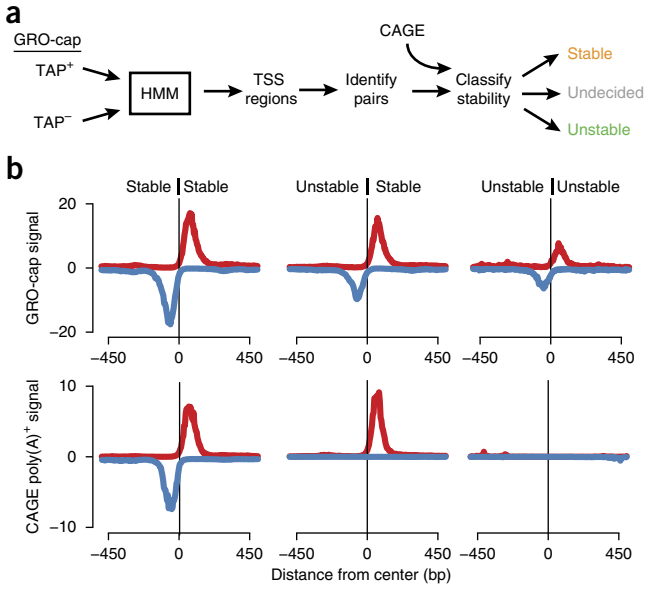

Figure 3 TSS identification and classification. (**a**) TSS regions were identified with an HMM from GRO-cap reads and control reads (GM12878, 117,613; K562, 128,471) and combined into pairs of divergent TSSs that were then classified according to the presence of CAGE signal. (**b**) Composite profiles of GRO-cap and CAGE reads aligned to the center of GRO-cap TSS pairs after classifying pairs on the basis of the stability of the transcripts produced. Profiles are stable-stable (left), unstable-stable (center) and unstable-unstable (right). The *y* axes show median read counts in 5-bp windows.

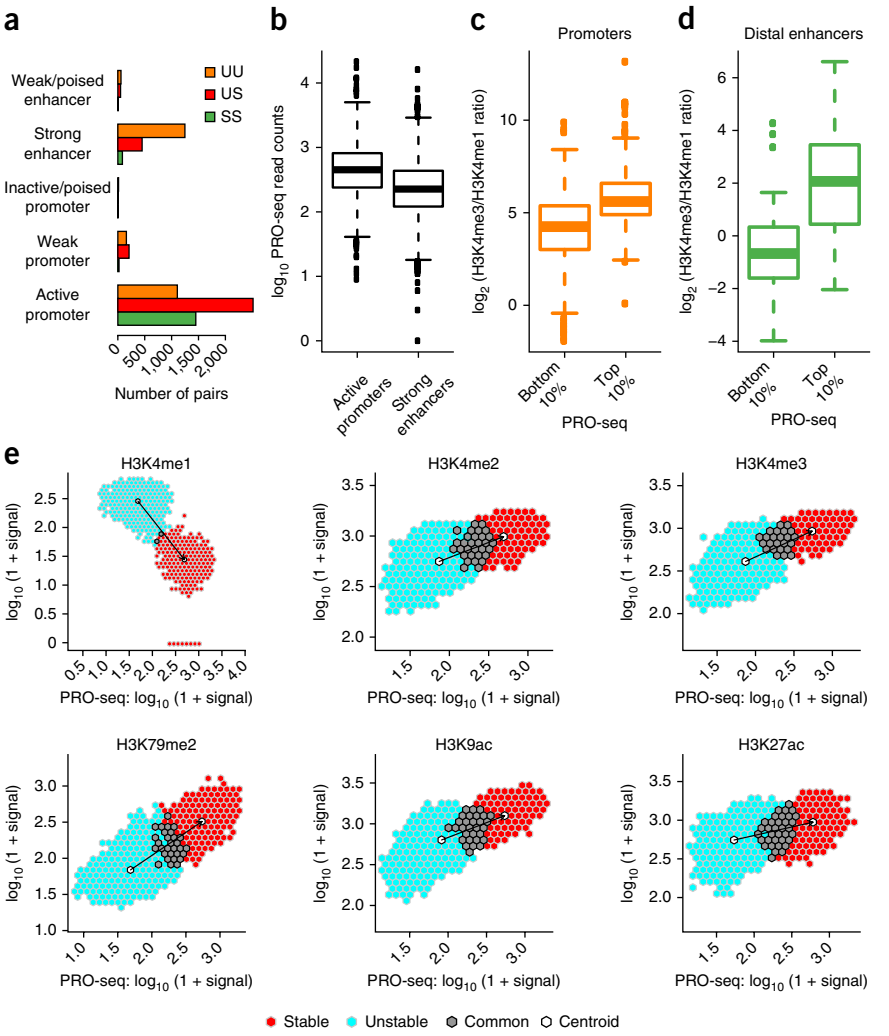

Figure 4 Histone marks at enhancers and promoters scale with Pol II intensity. (**a**) Number of TSS pairs from each stability class (UU, unstable-unstable; US, unstable-stable; SS, stable-stable) mapping to different regulatory regions as designated by ChromHMM. (**b**) Unstable-unstable pairs mapping to active promoter regions (n = 1,478) have a higher PRO-seq signal than those mapping to strong enhancer regions (n = 3,171), with active promoters and strong enhancers defined by ChromHMM. (**c**,**d**) H3K4me3/H3K4me1 ratio at the top and bottom deciles of PRO-seq signal in both promoter (n = 247 and 248 at top and bottom deciles, respectively) (**c**) and enhancer (n = 91 and 97 at top and bottom deciles, respectively) (**d**) TSS regions at distal enhancers (**e**) PRO-seq signal versus the indicated histone modifications at TSS regions. Signal is further split between TSSs classified as unstable (light blue) or stable (red) and points overlapping between the two (gray). The centroid for each subset is shown in white. H3K4me2, dimethylation of histone H3 at lysine 4; H3K9ac, acetylation of histone H3 at lysine 9.

Stable and unstable RNAs at transcription start sites

GRO-seq identified divergent transcription at promoters and enhancers[2,9], and GRO-cap had the sensitivity to detect and precisely map divergent transcription in over 90% of the TSS regions (**Supplementary Fig. 5d**). To simplify downstream analyses that compare various characteristics of initiation at promoters and enhancers, we created a set of 'divergent TSS pairs' that was filtered against cases of partially overlapping initiation pairs (Online Methods). The resulting set was composed of 22,443 TSS pairs from GM12878 cells and 24,894 pairs from K562 cells (38% and 39% of all TSSs, respectively). As both cell lines showed similar results, we will refer to GM12878 data unless otherwise stated. We then classified high-confidence GRO-cap–based TSSs into those giving rise to 'stable' transcripts (captured by CAGE and GRO-cap) and those that produce 'unstable' transcripts (captured only by GRO-cap) (**Fig. 3a**, Online Methods and **Supplementary Fig. 5**). The distinction between stable and unstable transcripts is also apparent from other RNA-based assays. For instance, stable TSSs have strong RNA-seq profiles (**Supplementary Fig. 6**), whereas unstable TSSs have very weak or non-existent RNA-seq profiles. These patterns hold for both the poly(A)$^+$ and poly(A)$^-$ versions of CAGE and RNA-seq, indicating that this difference is not simply due to differential polyadenylation.

We analyzed three classes of divergent TSS pairs: stable-stable, unstable-stable and unstable-unstable pairs (**Fig. 3a**,**b** and **Supplementary Fig. 5a**,**b**). Each of these classes covered a wide range of directional transcription preferences, suggesting that the directionality of initiation is not directly linked to RNA stability (Online Methods and **Supplementary Fig. 5d**,**e**). The stability of transcripts from individual TSSs and, by extension, classes of TSS pairs generally corresponded to distinct transcript annotation types (**Supplementary Fig. 5c**) and histone marks (**Supplementary Fig. 7**). In particular, the stable-stable and unstable-stable classes were enriched in chromatin signatures associated mainly with promoter regions (trimethylation of histone H3 at lysine 4 (H3K4me3)) and active transcription elongation (dimethylation of histone H3 at lysine 79 (H3K79me2) and trimethylation of histone H3 at lysine 36 (H3K36me3)) and corresponded to various stable transcripts such as protein-coding genes and lincRNAs (**Supplementary Fig. 5c**). In contrast, unstable-unstable pairs had enhancer-like chromatin features such as high levels of H3K4me1 and low levels of or ill-defined transcription elongation marks. Thus, our TSS pair classes generally correspond with the expected transcript annotation types, yet, by using transcript stability as the basis for our analysis, we are able to reduce TSSs to three fundamental classes in a data-driven and annotation-independent fashion.

Transcriptional level explains differences in histone modification

Although the ChromHMM distinction between promoters and enhancers was generally consistent with our TSS classes, with stable-stable and unstable-stable pairs mainly found at active promoters and unstable-unstable pairs mainly found at enhancers (**Fig. 4a**), a substantial fraction of unstable-unstable pairs were classified by ChromHMM as active promoter regions. This observation is unexpected given that active gene promoters should produce a stable transcript in at least one direction. Inspection of the unstable-unstable pairs classified as active promoters showed that they had stronger PRO-seq signals than unstable-unstable pairs classified as enhancers

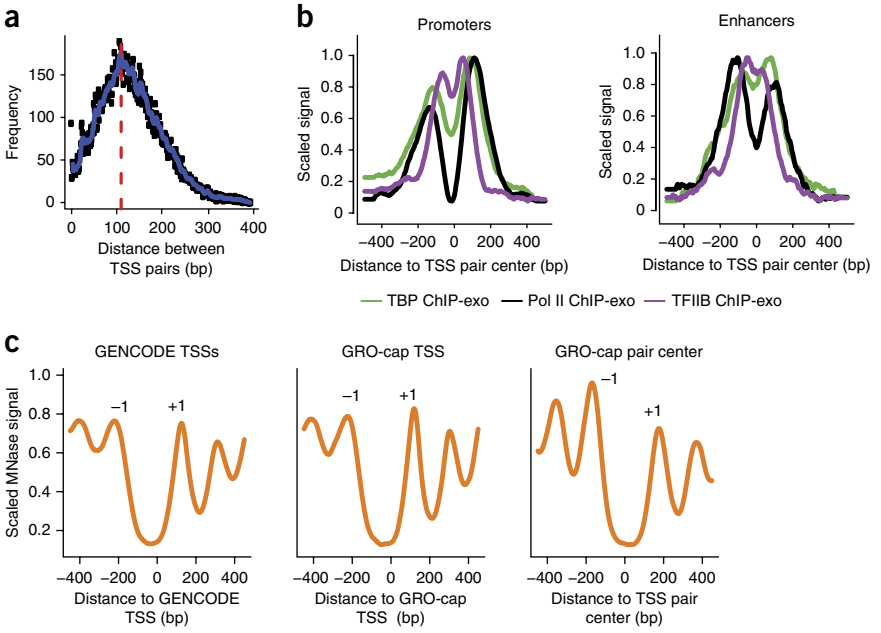

Figure 5 Architecture of TSS pairs. (**a**) Divergent TSSs are tightly packed, with an estimated 110-bp inter-TSS distance, as determined from the overall distribution of opposing-strand read distances. (**b**) ChIP-exo profile[26] for Pol II, TBP and TFIIB, centered on TSS pairs and split between promoter (left) and enhancer (right) regions (ChromHMM). (**c**) MNase-seq profiles at protein-coding promoters, aligned either by GENCODE annotations (left; also positive for GRO-cap signal), GRO-cap TSSs at GENCODE promoters (center) or GRO-cap TSS pair centers (right). Peaks corresponding to −1 and +1 nucleosomes are indicated.

(**Fig. 4b**). Thus, it is possible that these unstable-unstable pairs are actually enhancers that are misclassified as promoters owing to the presence of high levels of transcription-related histone marks (i.e., H3K4me3). A striking example occurs at the β-globin locus, where the upstream HS4-transcribed enhancer was erroneously characterized as a promoter by ChromHMM, whereas the promoter was erroneously predicted to be an enhancer (**Fig. 1a**).

To closely investigate the relationship between transcription level and histone marks at promoters and enhancers, we defined a set of stable TSSs from unstable-stable pairs proximal to annotated protein-coding genes (putative promoters) and contrasted them with TSSs identified from unstable-unstable pairs in transcription factor ChIP-seq (chromatin immunoprecipitation and sequencing) peaks that were distal to genes (putative enhancers). Although the promoters were generally more highly transcribed than the enhancers, the H3K4me3/H3K4me1 ratio at both promoters and enhancers scaled with the corresponding level of transcription (**Fig. 4c**,**d**). Expanding this analysis to all GRO-cap–identified TSSs in our TSS pairs (including both promoters and enhancers), we observed that transcription-associated histone modifications were directly related to the transcription level and that this relationship was maintained independently of transcript stability (**Fig. 4e**). That is, as the level of transcriptionally engaged Pol II increases at TSS pairs, so do the levels of H3K4me3 and other transcription-associated histone modifications.

One defining feature of mammalian promoters is a higher CpG nucleotide content than enhancers, which is thought to contribute to the transcription-independent deposition of H3K4me3. For instance, the CpG-binding protein Cfp1 has been implicated in the deposition of H3K4me3 through its recruitment of Setd1 (ref. 35). However, the DNA-binding domain of Cfp1 is dispensable for targeting H3K4me3 to active genes, suggesting that the relationship between CpG content and H3K4me3 marks might be indirect. Furthermore, we saw a clear disconnect between CpG content and histone modifications (H3K4me3 and others) at promoters and enhancers (**Supplementary Fig. 8**), suggesting that H3K4me3 levels at enhancers are not directly tied to CpG content. Thus, the difference in histone modifications at promoters and enhancers is not specific to the type of regulatory element; rather, this difference appears to be more fundamentally associated with the level of transcription.

Architecture of initiation at promoters and enhancers

To identify features of initiation regions that might distinguish promoters from enhancers, we closely examined the architecture of TSS regions. Using our high-confidence TSS pairs, we showed that divergent initiation occurred, on average, 110 bp apart (**Fig. 5a**), with relatively small variations between TSS pair classes (**Supplementary Fig. 9**). Although divergent initiation is less common in *Caenorhabditis elegans*, our estimates of the distance between divergent pairs in that species were nearly identical[21]. Despite the narrow distance, high-resolution ChIP exonuclease (ChIP-exo)[26] localization of two general transcription factors (GTFs) that bind core promoters (TBP and TFIIB) showed that an independent transcription initiation complex formed in each direction at divergent TSS pairs at promoters and enhancers (**Fig. 5b**).

Transcription initiation is often closely followed by promoter-proximal pausing. ChIP-exo data have shown that the majority of the Pol II molecules at promoters are downstream of TBP and TFIIB and are likely to be in a paused state[26]. Thus, we hypothesized that there might be some interplay between the strength and location of pausing and divergent TSS distances. Although we observed distinct pause modes (proximal focused and distal dispersed, as previously found in *Drosophila melanogaster*[22]), we found no effect of these modes on divergent initiation distances (**Supplementary Fig. 10a–c**) or the peak locations of TFIIB binding (**Supplementary Fig. 10d**). Together with the similar divergent TSS distance results from *C. elegans* (where pausing is rare), this observation suggests that pausing location does not feed back and influence the locations of divergent TSSs.

Although we find symmetric initiation and GTF binding at divergent promoter TSSs, nucleosome positioning is thought to be asymmetric at promoters. Typically, with respect to GENCODE TSSs, there is a well-positioned downstream nucleosome (+1 nucleosome), whereas the upstream nucleosome (−1 nucleosome) has more variable positioning[36] (**Fig. 5c**, top). In contrast, nucleosomes are reported to be strongly positioned at both sides of transcription factor–bound enhancers[37] (**Supplementary Fig. 11**). However, when we aligned nucleosome data from MNase-seq experiments (micrococcal nuclease digestion and sequencing) to the center of our TSS pairs, we clearly saw that both nucleosomes flanking the protein-coding unstable-stable and stable-stable TSSs were well positioned (**Fig. 5c**, bottom), with similar profiles to those at enhancers. Thus, the symmetric architecture of initiation regions applies universally to promoters and enhancers.

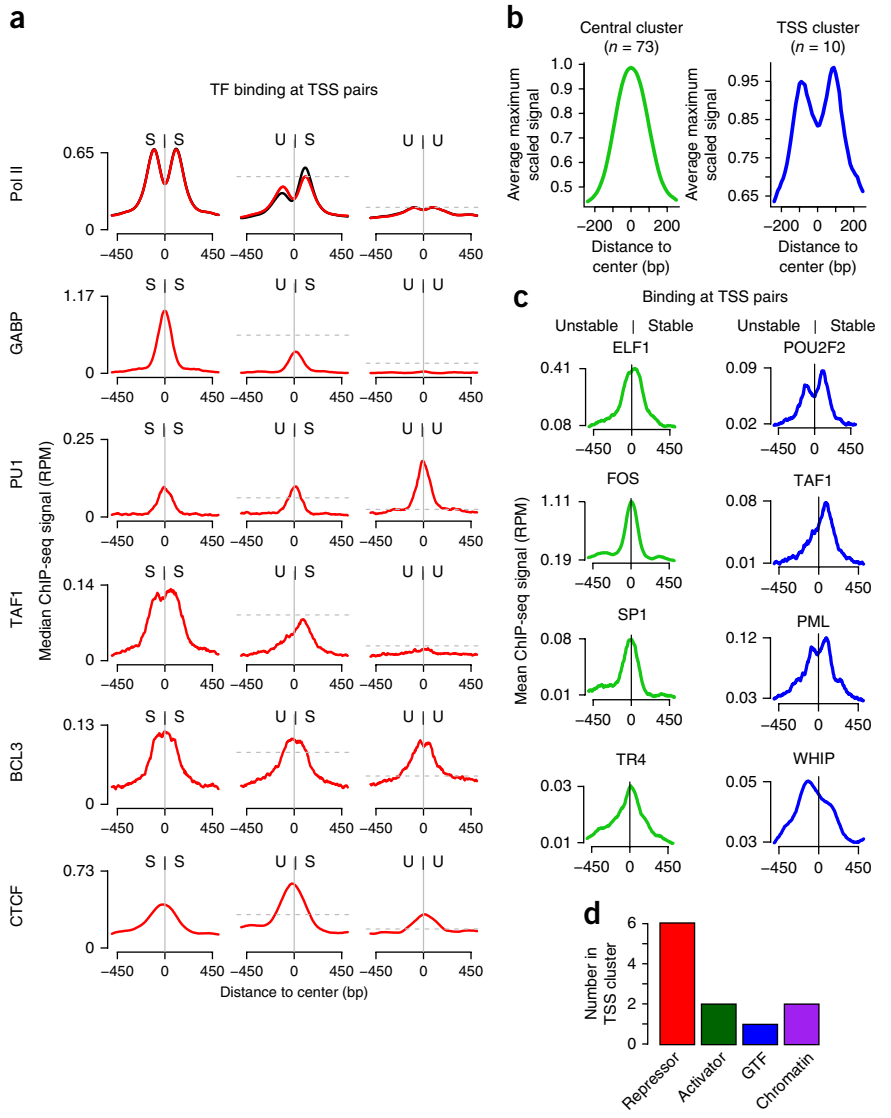

Figure 6 Modes of transcription factor binding at TSS pairs. (**a**) Representative ChIP-seq profiles of different modes of transcription factor binding at different TSS pair stability classes (S, stable; U, unstable). Signals are subject to paired subsampling to correct for Pol II signal dependency (top; Online Methods). The *y* axes show median read density in 5-bp windows. The horizontal dashed lines represent the expected peak signal if the signal followed the scaling of Pol II relative to the stable-stable panel. (**b**) ENCODE transcription factor ChIP-seq profiles, anchored on TSS pairs, cluster into two distinct groups, central binders (left) and TSS binders (right). (**c**) Examples of the two positional modes of binding at unstable-stable pairs. (**d**) Classification of factors within the TSS-binding cluster. The total number of factors is greater than the number of TSS-binding factors because factors can be part of more than one functional group (**Supplementary Table 2**).

The observed symmetries of nucleosome positioning and core promoter factors raise the question of how sequence-specific transcription factors bind within this context. Using transcription factor ChIP-seq data from ENCODE, we observed four main preferences for pair classes by transcription factors (**Fig. 6a** and **Supplementary Fig. 12**): factors that bound preferentially at stable-stable pairs (for example, GABP), factors that bound preferentially at unstable-unstable pairs (for example, PU1), factors that bound indiscriminately at all pair classes (for example, BCL3) and factors with a preference for unstable-stable pairs (for example, CTCF). In addition, we observed two clusters of transcription factors defined by the relative positions of their binding sites within divergent TSS pairs (**Fig. 6b,c**): central-binding factors (for example, SP1) and TSS-proximal binding factors (for example, PML). We were limited by the ChIP-seq sets available, but, with the data sets used (n = 84), most factors fell into the central-binding cluster (binding profile peaks in the center between the divergent TSSs; n = 73) versus the TSS-binding cluster (binding profile peaks over the TSS positions; n = 10) (**Supplementary Table 2**). Interestingly, the TSS-proximal binding cluster included both GTFs such as TAF1 and transcriptional repressors such as NRSF and PMl (**Fig. 6d**), suggesting a potential involvement of these factors in transcript stability determination or preferential targeting of these factors to stable transcripts. These results demonstrate a clear relationship between transcription factor binding and TSS structure and suggest that central-binding transcription factors and the symmetrical structure of initiating regions might be mechanistically linked.

Sequence predictors of transcript stability

Because DNA sequence is known to influence initiation, productive transcription, and RNA processing and stability, we also examined the sequence composition near our TSS pairs. In general, we found that sequence conservation and nucleotide frequency were indicative of transcript stability (**Supplementary Fig. 13a–c**). In particular, stable-stable TSSs were associated with increased proportions of cytosine and guanine nucleotides and increased proportions of CpG dinucleotides within and around the pairs. In contrast, unstable-unstable TSS pairs were depleted for cytosines, guanines and CpG sites. Unstable-stable TSS pairs displayed a combination of these two patterns. Despite these biases, we saw similar frequencies of core promoter elements (TATA and Inr sites) in the expected positions at all classes of TSS pairs (**Supplementary Fig. 14a,b**). This observation is consistent with ChIP-exo detection of GTFs at all classes of TSS pairs (**Supplementary Fig. 14c**), indicating that other mechanisms might be dictating the production of stable versus unstable transcripts. Indeed, recent work has shown that sequences that direct the binding and activity of polyadenylation-dependent termination machinery or the U1 splicing complex work antagonistically to direct unstable or stable transcription, respectively, at protein-coding genes[4,5]. In this model, 5′ splice sites (SS5) that bind U1 can suppress PAS-dependent termination, thus promoting productive elongation of protein-coding mRNAs.

To determine whether there was a direct relationship between our transcript stability classes and premature PAS-dependent termination, we scanned the regions downstream of the TSSs for matches to the PAS and SS5 motifs and observed that our stable and unstable TSS classes

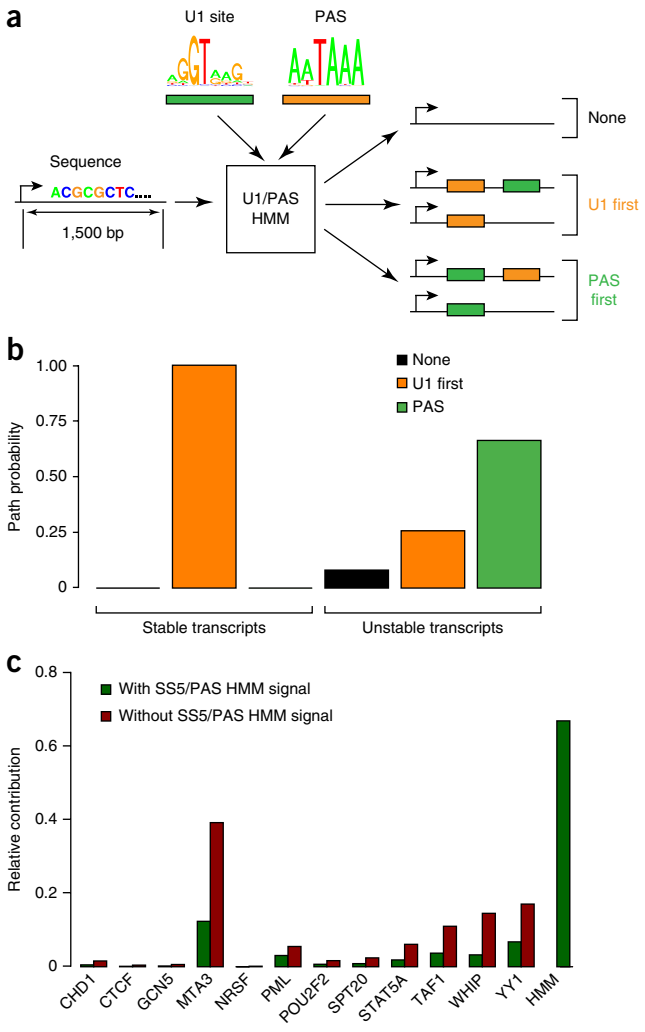

Figure 7 Determinants of RNA stability for both promoters and enhancers. (**a**) Diagram of transcript U1/PAS classification. Each transcript (first 1.5 kb) is processed through an HMM to determine the relative order and occurrence of SS5 and PAS elements. (**b**) Estimated path probabilities of alternative element occurrences (black, neither SS5 nor PAS; orange, SS5 first; green, PAS first) obtained by applying the expectation-maximization algorithm to each transcript subset (stable and unstable TSS stability classes). (**c**) Relative contribution of various transcript factors in logistic regression of the stability classes, with and without including the U1/PAS HMM–derived signal (posterior path probability of being in the unstable class).

followed a pattern consistent with the earlier reports (**Supplementary Fig. 15a**,**b**). That is, the SS5 motif was enriched downstream of TSSs for stable transcripts but depleted at TSSs for unstable transcripts, with the reverse true for the PAS motif. We devised an HMM that incorporated SS5 and PAS motif models and used it to compare the likelihoods of finding SS5 binding sites before and after a PAS site (**Fig. 7a** and **Supplementary Fig. 15c**). Our results indicate that SS5 binding sites strongly tend to precede the PAS on stable transcripts but not on unstable transcripts (**Fig. 7b**). In the case of single-exon genes (n = 105), both SS5 and PAS sites were less frequent but PAS sites were more depleted than SS5 sites (**Supplementary Fig. 15d**,**e**). These results are consistent with previous observations for protein-coding genes, and, notably, they demonstrate that these sequence predictors of elongation hold for all TSSs, including those at enhancers. Furthermore, our HMM can be used to predict transcript stability with high accuracy (63%), suggesting that these motifs and their spatial relationship are strong determinants in this process.

Finally, we used logistic regression to assess the relevance of transcription factors in the TSS-binding cluster to transcription stability. Transcription factors by themselves explained only a small fraction of the variance in stability (R^2 = 0.05). Furthermore, when the signal from the poly(A)-U1 HMM was also considered, their relative importance dropped considerably (**Fig. 7c**). These observations suggest that most of the information about stability comes from the presence or absence of early PASs and U1 splicing signals, but they do not rule out the possibility that some transcription factors might be components of the splicing pathway or contribute to feedback between splicing and transcription levels.

DISCUSSION

Several studies have documented divergent transcription at promoters and enhancers[2,3,8,9,38]; however, the nature and organization of initiation sites, their underlying DNA elements and their relationships with transcription factor binding and nucleosome positions have yet to be reconciled. In this article, we show that assaying nascent RNAs dramatically increases sensitivity for enhancer detection in comparison with methods that map accumulated RNAs. By contrasting our GRO-cap data with CAGE data, we are able to classify TSS pairs on the basis of the stability of the resulting transcripts. Unstable transcripts are those that are likely targeted for immediate degradation by the exosome and thus are unable (or less likely) to be discovered in assays that detect accumulated RNAs, such as CAGE. By contrast, stable transcripts are detectable in both nascent and accumulated RNA pools. These classifications allow us to work directly from genome-wide functional genomic assays without reliance on genomic annotations. By analyzing these annotation-free TSSs together with DNA sequences and functional genomic data, we are able to catalog the precise nature of the structure and chromatin content at initiation sites. We find that the divergent TSS pairs at both promoters and active enhancers (i) have similar frequencies of canonical core promoter elements, (ii) have distinct transcription complexes at each member of a pair, (iii) are separated by 110 bp on average, (iv) are bound by central transcription activators, (v) are flanked on both sides by positioned nucleosomes and (vi) have histone modifications typically associated with transcription initiation, present in proportion to the amount of transcription. These results suggest a unified model for the mechanisms that govern transcription initiation at both enhancers and promoters (**Fig. 8a**).

We show that divergent initiation occurs within a window of 90–120 bp, which is a surprisingly narrow interval considering that a preinitiation complex makes contacts up to 50 bp upstream and downstream of the TSS[39]. The close proximity of divergent initiation events and the evidence for bound transcription factors between them make it difficult to imagine that multiple independent polymerase complexes and transcription activators simultaneously occupy the same promoter. One possible alternative is that one polymerase initiates first and then pauses downstream, allowing enough space for a second polymerase to initiate upstream and in the opposite direction. Consistent with this hypothesis, high-resolution ChIP-exo data suggest that the majority of Pol II molecules on chromatin in human cells (K562) are paused approximately 50 bp downstream of the initiation site[26]. We also show that these independent and divergent transcription complexes have similar frequencies of well-known core promoter elements in the underlying DNA. This finding suggests that recruited Pol II is not randomly initiating at open DNA regions associated with enhancers and divergent

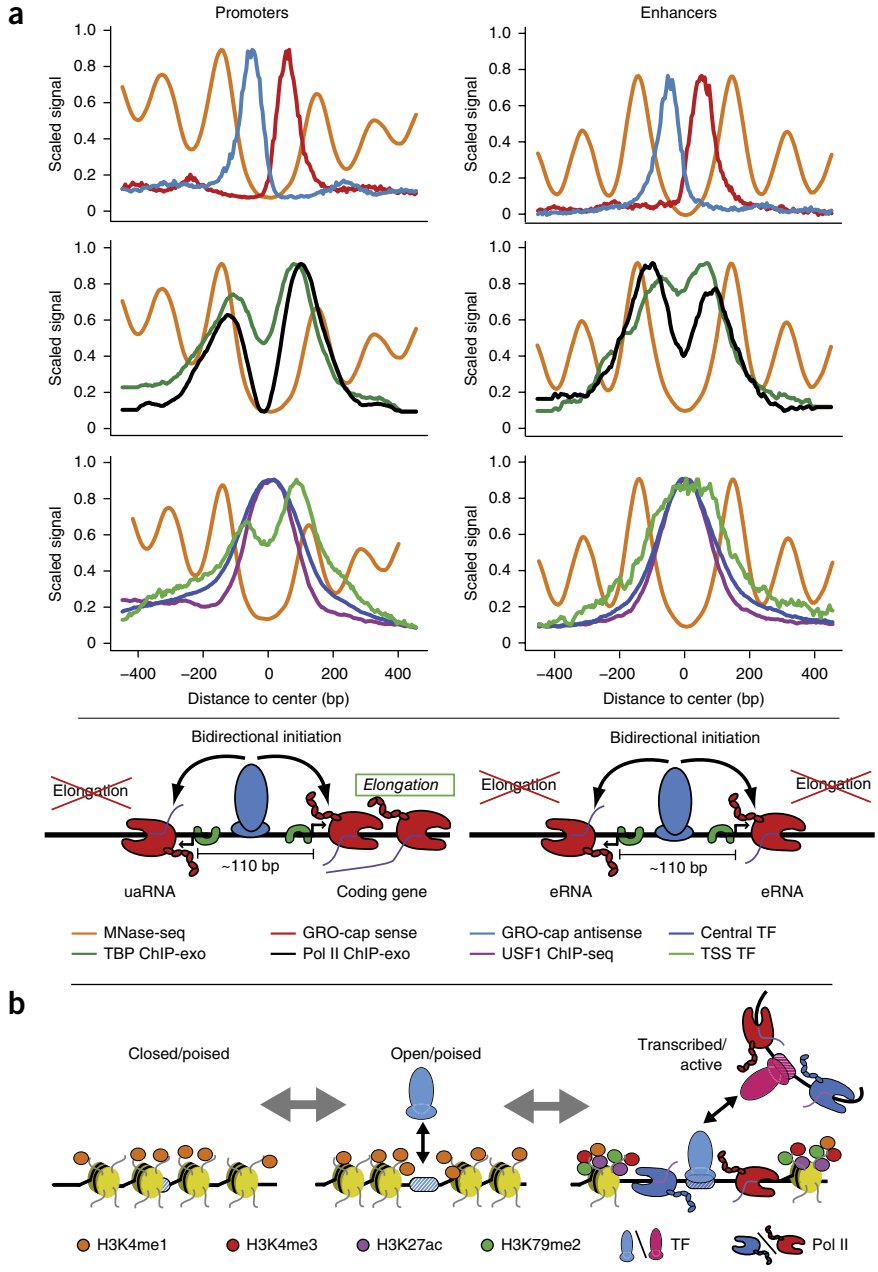

Figure 8 Summary of transcription initiation at regulatory regions. (**a**) Our analysis of TSSs identifies a common structure across all initiation regions, including promoters and enhancers. In both cases (top row), a tightly packed (110 bp apart) divergent TSS pair (red, plus strand; blue, minus strand) is surrounded by well-positioned nucleosomes (orange), with independent preinitiation complexes (separate TBP (green) and Pol II (black) ChIP-exo peaks) (second row) and sharing two distinct transcription factor cluster binding modes (green, central; blue, TSS proximal) (third row). We propose that central, activator transcription factor binding (for example, by USF1; purple), in conjunction with core promoter elements, determines the positioning of the divergent initiation sites. Finally, DNA sequence properties (not depicted here), possibly in cooperation with other factors, determine the resulting transcript type (stable/elongating, protein coding; unstable/terminating, uaRNA, eRNA, etc.). TF, transcription factor. (**b**) A model depicting the possible progression of enhancer states from chromatin-marked but largely inaccessible regions (left) to more open regions through transcription factor binding (center) and, finally, active transcription, which brings with it associated chromatin marks (in particular, H3K79me2, H3K27ac and increased methylation levels of H3K4) (right).

promoters[15]. Rather, the normal cohort of GTFs is positioned to facilitate initiation at these sites.

We also find evidence for positional modes for transcription factor binding in divergent TSS regions. Most factors bind between the two divergent TSSs (central binders), suggesting that they have a role in activation and are likely a major determinant or result of the overall architecture of initiation sites. In contrast, TSS-proximal transcription factors are primarily enriched for repressors, suggesting that certain repressors can act by preventing access of the transcription machinery to critical parts of the core promoter. The apparent tight spacing and organization of binding suggest that few factors simultaneously bind at any given initiation region. This is in agreement with evidence for a small number of identifiable sequence motifs, even when numerous factors are found in narrow regions by ChIP-seq[40]. Coinciding signals might reflect indirect binding of transcription factors or binding events that occur in a subset of cells within a population. Finally, the close relationship between transcription factor binding and initiation in our model provides a possible explanation for why protein-coding genes typically have multiple associated mRNAs with small differences in TSS location. These alternative TSSs likely result from the presence of multiple neighboring binding sites for transcription factors that compete as anchors for initiation. As a result, depending on cell type and condition, different transcription factor binding events lead to small shifts in the position of the initiation site.

Promoter regions are generally assumed to be quite broad, with promoter-associated transcription factor binding sites spanning a multiple-kilobase region near the TSS, but our results suggest that initiation regions are primarily defined by a relatively narrow 100- to 200-bp window. Part of this discrepancy can be attributed to poor or incomplete annotation of genes, but it might also indicate that multiple independent initiation regions often act as neighboring enhancers. Although we have focused here on nonoverlapping TSS pairs to simplify our analyses, we expect that overlapping TSS pairs will represent an aggregate of the local transcription factor occupancies. In the future, it will be interesting to further investigate transcription factor occupancy at these more complex regions with the help of higher-resolution assays, such as ChIP-exo[41].

Previous work suggested that enhancer chromatin undergoes a progression from a closed state to an open state that is required for transcription factor binding[14,42–44]. Our analyses of DNase I hypersensitivity and GRO-cap data at enhancers generally support the existence of and potential progression through at least three enhancer states: closed, open and transcriptionally active (**Fig. 8b**). Comparisons of these states with other functional genomics data

suggest that transcribed enhancers are the most active, whereas the closed and open classes represent a poised state. We envision that it is equally plausible to progress in either direction between states; thus, the poised states could represent enhancers that have yet to be activated or dormant enhancers that are vestiges of past activity[45]. Interestingly, poised enhancers resemble a form of preactivated promoter recently observed during developmental transitions[41], providing yet another similarity between regulation at promoters and enhancers. Although we see less evidence for transcription factor binding at open, untranscribed enhancers, these regions could arise through the binding of a small number of 'pioneering' transcription factors. Also, some poised enhancers could be in an open state simply because they have relatively poor affinity for nucleosomes owing to their underlying sequences. Alternatively, permissive chromatin could arise concomitantly with transcription factor binding and transcription[14]. In either case, the transition from the open or poised states to the transcriptionally active state is clearly related to the binding of central, activating transcription factors (**Fig. 8b**). It will require further work to determine whether all functionally active enhancers (influencing the activity of target transcripts) generate local transcription.

It is generally thought that distinct mechanisms selectively mark histones at enhancers and promoters. In particular, enhancers are typically identified as having high levels of H3K4me1 relative to H3K4me3 (refs. 12,13). However, we observe a strong positive correlation between the absolute levels of transcription and the H3K4me3/H3K4me1 ratio at active enhancers, suggesting that differences in H3K4 methylation patterns at enhancers and promoters might simply reflect differences in transcription levels. Consistent with this observation, H3K4me3 has been detected at some active enhancers[11,46] and can be deposited in a transcription-dependent manner[11,47]. Why, then, are enhancers generally observed to have less transcription initiation and, hence, less H3K4me3 than promoters? One possible explanation comes from observations of the feedback mechanisms whereby elongation of transcription positively contributes to subsequent rounds of initiation. A related possibility, consistent with our observation of a splicing-dependent difference in transcript stability at promoters and enhancers, would be feedback from the splicing machinery. Indeed, the presence of a U1 splice site can positively influence the recruitment of GTFs to promoters[48]. In addition, the GTF TAF15 has been shown to interact with the U1 small nuclear ribonuclear protein (snRNP), providing another link between splicing and initiation complexes. Therefore, splicing-dependent elongation of transcription not only distinguishes promoters from enhancers but might also help explain the different intensities of transcription initiation and, hence, histone modifications at these regions.

The original definition of an enhancer describes a genomic interval that stimulates the transcription of another locus independently of its position and orientation relative to the transcribed locus[49]. Our analyses show that the mechanisms governing chromatin content and architecture at enhancers are quite similar to those at promoters. What, then, is a proper description of an enhancer? Three-dimensional chromatin links bridging different initiation regions have been observed both between traditional enhancers and promoters and between pairs of promoters[33]. Thus, the implication is that any initiation region can function as an enhancer, through the central-binding activator, irrespective of the fate or function of the local transcripts that are generated. Conversely, it is currently not clear whether some transcription factors can enhance distal transcription activity without generating local transcription.

Our observations have implications for an intriguing potential relationship between divergent transcription and the origin of new genes. It has recently been shown that asymmetries in productive transcriptional elongation favoring the sense-coding direction at gene promoters can be explained by a disproportional tendency for promoter-proximal cleavage and polyadenylation shortly after initiation in the antisense direction, which appears to be associated with an enrichment for PASs in the upstream antisense regions of genes[4,5]. Furthermore, PASs are depleted and U1 snRNP recognition sites (SS5s) are enriched in the sense direction, consistent with observations that the U1 snRNP complex protects pre-mRNAs from cleavage and polyadenylation[50,51]. Building on these observations, Wu and Sharp recently proposed a model for the evolutionary origin of new genes whereby short, unstable uaRNAs gradually increase in length and stability as mutations eliminate PASs and create new SS5s[52]. In this way, uaRNAs or eRNAs could develop, in a stepwise fashion, first into noncoding RNAs and then into protein-coding mRNAs, perhaps acquiring splicing capabilities along the way (which, in turn, would further improve stability). This process could be encouraged by positive feedback with transcription-associated mutational asymmetries, which are biased toward guanine and thymine nucleotides[53] and therefore would favor the formation of SS5s and the abolishment of PASs. In this article, we have shown that transcription initiation occurs in a bidirectional fashion at thousands of enhancers that have fundamentally the same architecture of initiation as traditional promoters. Thus, if uaRNAs and eRNAs do indeed sometimes develop into genes, then the genome is replete with potential new genes, many of them far from existing genes. Additional studies of nascent RNAs across cell types and species may help to shed light on these evolutionary questions.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All data files are available on the Gene Expression Omnibus (GEO) under accession GSE60456. All data are also available as tracks on the UCSC Genome Browser[54] using http://compgen.bscb.cornell.edu/GROcap/. Tracks for the TSS calls and stability classifications are also available in the **Supplementary Data Set**. These files can be analyzed directly or can be uploaded onto the UCSC Genome Browser as custom tracks for viewing.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Research reported in this publication was supported by NIGMS (National Institute of General Medical Sciences) and NHGRI (National Human Genome Research Institute) grants from the US National Institutes of Health under award numbers GM25232 to J.T.L. and HG0070707 to A.S. and J.T.L., respectively. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

AUTHOR CONTRIBUTIONS

L.J.C. and J.T.L. designed the experiments. L.J.C. and C.T.W. produced the data sets. A.L.M. designed and implemented software for data analysis. A.L.M., L.J.C., A.S., J.T.L. and C.G.D. analyzed the data and interpreted the results. L.J.C., A.L.M., A.S., J.T.L. and C.G.D. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
2. Core, L.J., Waterfall, J. & Lis, J. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
3. Seila, A.C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
4. Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B. & Sharp, P.A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
5. Ntini, E. *et al.* Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nat. Struct. Mol. Biol.* **20**, 923–928 (2013).
6. Trinklein, N.D. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res.* **14**, 62–66 (2004).
7. Oler, A.J. *et al.* Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat. Struct. Mol. Biol.* **17**, 620–628 (2010).
8. Kim, T.K. *et al.* Widespread transcription at neuronal activity–regulated enhancers. *Nature* **465**, 182–187 (2010).
9. Hah, N., Murakami, S., Nagari, A., Danko, C.G. & Kraus, W.L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* **23**, 1210–1223 (2013).
10. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390–394 (2011).
11. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* **18**, 956–963 (2011).
12. 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).
13. Heintzman, N.D. & Ren, B. Finding distal regulatory elements in the human genome. *Curr. Opin. Genet. Dev.* **19**, 541–549 (2009).
14. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).
15. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* **14**, 103–105 (2007).
16. Ørom, U.A. & Shiekhattar, R. Long non-coding RNAs and enhancers. *Curr. Opin. Genet. Dev.* **21**, 194–198 (2011).
17. Lam, M.T., Li, W., Rosenfeld, M.G. & Glass, C.K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* **39**, 170–182 (2014).
18. Seila, A.C., Core, L.J., Lis, J.T. & Sharp, P.A. Divergent transcription: a new feature of active promoters. *Cell Cycle* **8**, 2557–2564 (2009).
19. Melgar, M.F., Collins, F.S. & Sethupathy, P. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.* **12**, R113 (2011).
20. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
21. Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T. & Meyer, B.J. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *Elife* **2**, e00808 (2013).
22. Kwak, H., Fuda, N.J., Core, L.J. & Lis, J.T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
23. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *bioRxiv* doi:10.1101/005447 (29 August 2014).
24. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
25. Orkin, S.H. Regulation of globin gene expression in erythroid cells. *Eur. J. Biochem.* **231**, 271–281 (1995).
26. Venters, B.J. & Pugh, B.F. Genomic organization of human transcription initiation complexes. *Nature* **502**, 53–58 (2013); retraction **513**, 444 (2014).
27. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
28. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
29. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **100**, 15776–15781 (2003).
30. Maruyama, K. & Sugano, S. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**, 171–174 (1994).
31. Affymetrix ENCODE Transcriptome Project & Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
32. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, e201 (2013).
33. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
34. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
35. Clouaire, T. *et al.* Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev.* **26**, 1714–1728 (2012).
36. Schones, D.E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
37. Gaffney, D.J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet.* **8**, e1003036 (2012).
38. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
39. Coulombe, B. & Burton, Z.F. DNA bending and wrapping around RNA polymerase: a "revolutionary" model describing transcriptional mechanisms. *Microbiol. Mol. Biol. Rev.* **63**, 457–478 (1999).
40. Foley, J.W. & Sidow, A. Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines. *BMC Genomics* **14**, 720 (2013).
41. Wamstad, J.A. *et al.* Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* **151**, 206–220 (2012).
42. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
43. Zentner, G.E., Tesar, P.J. & Scacheri, P.C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **21**, 1273–1283 (2011).
44. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
45. Stergachis, A.B. *et al.* Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888–903 (2013).
46. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
47. Shilatifard, A. The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. *Annu. Rev. Biochem.* **81**, 65–95 (2012).
48. Damgaard, C.K. *et al.* A 5′ splice site enhances the recruitment of basal transcription initiation factors *in vivo*. *Mol. Cell* **29**, 271–278 (2008).
49. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
50. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).
51. Berg, M.G. *et al.* U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**, 53–64 (2012).
52. Wu, X. & Sharp, P.A. Divergent transcription: a driving force for new gene origination? *Cell* **155**, 990–996 (2013).
53. Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33**, 514–517 (2003).
54. Kent, W.J. *et al.* in The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
55. Ashe, H.L., Monks, J., Wijgerde, M., Fraser, P. & Proudfoot, N.J. Intergenic transcription and transinduction of the human β-globin locus. *Genes Dev.* **11**, 2494–2509 (1997).

ONLINE METHODS

Preparation of GRO-cap, PRO-seq and GRO-seq libraries. GRO-cap libraries for K562 and GM12878 cells were produced precisely as described in Kruesi *et al.*[21]. We used 1×10^7 nuclei for each GRO-cap library or control. GRO-seq libraries for K562 and GM12878 cells were produced as described in Wang *et al.*[56]. PRO-seq libraries were produced as described previously[22], using TruSeq small-RNA adaptors (Illumina) and 5×10^6 nuclei. K562 cells were purchased from the American Type Culture Collection (ATCC; CCL-243). GM12878 cells were from the Coriell Institute for Medical Research (GM12878). Cells were tested for mycoplasma before use.

Mapping of sequencing data. After sequencing, GRO-seq and GRO-cap reads were trimmed to 30 bases and first mapped to a single copy of the ribosomal gene DNA (rDNA) locus to remove related transcribed sequences. Reads that did not map to rDNA were then mapped to the hg19 version of the human genome. Reads were required to be unique and have no more than two mismatches. PRO-seq reads (100 bases) were processed essentially as in Kwak *et al.*[22]. Adaptors were removed with cutadapt[57], and unique sequences of 15 bp or more in length that mapped to the hg19 genome were kept for further analysis.

Prediction of transcription start sites. *Pre-processing of GRO-cap data.* GRO-cap aligned data, normalized by total read counts, were summarized in fixed intervals of 10 bp along the reference genome to increase the signal for low-intensity initiation sites and 'smooth away' minor misalignments under TAP^+ and TAP^- conditions. Each 10-bp interval was assigned two values, one summarizing the signal difference under the TAP^+ and TAP^- conditions and the other indicating the presence of a TAP^+ 'peak'. To summarize the signal difference under TAP^+ and TAP^- conditions at each interval, we assigned the interval to one of three categories: (i) 'no signal' (interval with no reads under the TAP^+ condition), (ii) 'enriched' (interval with more reads under the TAP^+ than TAP^- condition) and (iii) 'depleted' (interval with more reads under the TAP^- than TAP^+ condition). To compute the binary peak indicator for an interval, we searched for depleted intervals (using the above definition) within ten 10-bp intervals (100 bp) in either direction of a given interval; if at least two were found, we used their mean normalized read count as an estimate of the local background level. The interval in question was then called a peak if its normalized read count was greater than twice the estimated local background level. We found that our final predictions were not very sensitive to the threshold for calling peaks, with a wide range of fold enrichments producing numbers of predictions that differed by no more than 3%.

Design of the hidden Markov model. Previous CAGE studies have shown that TSS regions can be both 'sharp' (highly peaked) and 'broad' (refs. 58–60). Therefore, we designed our HMM to have a single background state (B) and two groups of alternative states, representing non-peaked (M1) and peaked (M2) TSS regions (**Supplementary Fig. 2a**). The M1 and M2 groups were each composed of three states, and, within each group, these states shared the same multinomial emission distribution for no signal, enriched and depleted TAP^+ read counts. In addition, the states had a conditionally independent emission distribution for the peak signal, set such that only the middle state of the M2 group permitted peaked intervals. Because multiple peaks can occur in a single peaked TSS region, transitions among the states in the M2 group allowed for zero or more steps between consecutive peaks (middle state). This design enforced a distinction between sharp and broad TSS regions, while avoiding false positives due to highly local spikes in the data.

Parameter estimation and transcription start site prediction. The free parameters of the model were set as follows. Most transition probabilities were set to 0 or 1 according to the constraints of model design (**Supplementary Fig. 2a**) or were assigned values reflecting a non-informative uniform prior distribution over possible state transitions (for example, transitions out of the first and last states of the M2 group). The two exceptions to this rule were the self-transition probabilities for the background state and the middle (peak-emitting) M2 state, which were assigned high (0.99) and low (0.1) values, respectively, because we expected peaks to be sparse along the genome. The emission parameters were set approximately on the basis of empirical observations of TSS regions. In particular, we observed that background regions were mostly devoid of reads (P(no signal) = 0.9; P(enriched) = P(depleted) = 0.05). By contrast, non-peaked regions (M1 group; broad TSSs) were dense in enriched intervals (P(no signal) = 0.09; P(enriched) = 0.9; P(depleted) = 0.01). Peaked regions (M2 group; peaked TSSs) had both enriched and depleted intervals in varying proportions, but, because this group was anchored by the peaked indicator, it was not sensitive to exact emission probabilities as long as no signal was unlikely; therefore, for the these states, we used P(no signal) = 0.1, P(enriched) = 0.45 and P(depleted) = 0.45.

TSS regions were obtained by running the Viterbi algorithm[60,61] on the pre-processed GRO-cap data, which found the most likely path through the HMM given the data and the model parameters. The predicted TSS regions were then refined for further analysis as follows. First, regions of longer than 100 bp that were assigned to the M2 group were split into constituent peaked subregions such that distances of at least 30 bp were maintained between them. Second, all regions were trimmed of leading and trailing depleted intervals (with more reads under the TAP^- than TAP^+ condition). The effects of these post-processing steps can be seen in **Supplementary Figure 2b**.

TSS paired regions. A divergent TSS pair was composed of adjacent TSS regions in opposing orientations (a minus-strand TSS region followed by a plus-strand TSS region) within 150 bp of each other (nearest edges). This threshold was set empirically, after manual observation of initiation sites, to capture the observed distances between divergent TSS regions (the median nearest-edge distance was 40 bp). We further filtered TSS pairs by requiring a high GRO-cap signal (minimum number of reads above the 20% quantile) so that we could reliably scale the various signals of interest by expression level in downstream analysis.

TSS stability classification. GRO-cap–based TSSs were classified into those giving rise to stable transcripts (captured by CAGE and GRO-cap) and those that produced unstable transcripts (captured only by GRO-cap). In practice, our TSS regions were classified as unstable in the absence of CAGE reads and as stable if they contained at least eight CAGE reads. These thresholds are conservative, and the latter is above the estimated CAGE background in introns (**Supplementary Fig. 5a**). We focused on high-confidence sets of both stable and unstable transcripts by further requiring a high GRO-cap signal (minimum number of reads above the 20% quantile). Interestingly, regardless of whether they arose from regions classified as promoters or enhancers[62], GENCODE[63] lincRNAs were largely stable by our classification.

Paired subsampling. In our analysis of divergent initiation regions, we produced composite profiles for paired TSSs in a variety of ChIP-based assays. A challenge in interpreting these profiles is that the marginal distributions of transcription levels often differ significantly at members of each pair, and other signals of interest, such as ChIP-seq measures of transcription factor binding, correlate strongly with transcription levels. Thus, apparent differences in the signals of interest might simply reflect differences in overall transcription level. This is especially a problem for unstable-stable pairs because unstable TSSs tend to have substantially lower transcription levels than their stable counterparts.

To improve the interpretability of these plots, we generated composite profiles using a subsampling method that ensured that the marginal Pol II ChIP-seq distributions were the same at both TSSs in a pair. Briefly, we summarized each TSS pair by four values: the Pol II ChIP-seq values and the signal of interest, both at the left and right TSS. For convenience, the Pol II ChIP-seq values were discretized into bins. We then defined a shared 'target' distribution for Pol II by pooling the data for the left and right TSSs. Finally, we subsampled from the collection of TSS pairs (summarized by their four values) in such a way that the left and right Pol II distributions exactly matched the target distribution. This subsampling step is complicated by the dependency between the left and right Pol II distributions, but this complication could be addressed using a simple algorithm that performed a depth-first search over the possible combinations of samples from the original distribution, branches of which were terminated whenever the constraints on the subsample were violated. The induced marginal distributions of values for the signal of interest at the left and right TSSs were then compared. In this way, differences in the profiles that simply reflected differences in Pol II binding (a surrogate for transcription level) were eliminated.

Splicing signal hidden Markov model. To define the HMM for splicing signals, we started with an SS5 position weight matrix (PWM) estimated from GENCODE 16 annotations of the first exon for protein-coding genes (**Supplementary Fig. 15b**). In addition, a PWM for PASs was estimated from the sequences reported in Beaudoing *et al.*[64]. Finally, a background model was estimated from the full DNA sequences, assuming the independence of sites.

Our HMM combined these motif models in such a way that we could make inferences about the relative positioning of SS5 and PAS sequence motifs. In particular, the HMM permitted branching from an initial background state into five alternative paths. Two of these paths visited an SS5 before an optional PAS, two others visited a PAS before an optional U1 site and a final path included neither of the two motif signals. The HMM was structured such that the transition from the initial background state was taken once and only once (**Supplementary Fig. 15c**).

We applied this HMM to sequences spanning the first 1.5 kb of the TSSs in each class (stable and unstable). To estimate the relative likelihood of each path, we computed maximum-likelihood estimates of the probabilities of transition into each of the five alternative paths using the Baum-Welch algorithm[65]. Because the number of free parameters was the same for all paths, no model complexity penalty was needed for this comparison.

Additionally, the probability of each alternative path for each sequence could be estimated by setting uniform transition probabilities out of the initial background state and then computing the respective posterior probabilities. This approach enabled the use of the HMM as a sequence classifier (by thresholding the sum of the posterior over the sequence), and it was used as input for stability regression.

Stability regression. The relative contribution of individual transcription factors and the splicing signal HMM to predicting TSS class (stable or unstable) was assessed by logistic regression. Transcription factor signals corresponded to sums of ChIP-seq signals in the predicted TSS region. The relative contribution of regression weights was computed according to Johnson *et al.*[66]. Because transcription factor binding patterns are often strongly correlated with transcription levels, we applied logistic regression to subsamples of stable and unstable TSSs with matching Pol II signal distributions.

56. Wang, I.X. *et al.* RNA-DNA differences are generated in human cells within seconds after RNA exits polymerase II. *Cell. Rep.* **6**, 906–915 (2014).
57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* **17**, 10–12 (2011).
58. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
59. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* **8**, 424–436 (2007).
60. Viterbi, A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **13**, 260–269 (1967).
61. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
62. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
63. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
64. Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**, 1001–1010 (2000).
65. Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G.J. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, UK, 1998).
66. Johnson, J.W. A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behav. Res.* **35**, 1–19 (2000).

doi:10.1038/ng.3142