Title

Human promoters are intrinsically directional.

Permalink
https://escholarship.org/uc/item/3s8164v9

Journal
Molecular cell, 57(4)

ISSN
1097-2765

Authors
Duttke, Sascha HC
Lacadie, Scott A
Ibrahim, Mahmoud M
et al.

Publication Date
2015-02-01

DOI
10.1016/j.molcel.2014.12.029

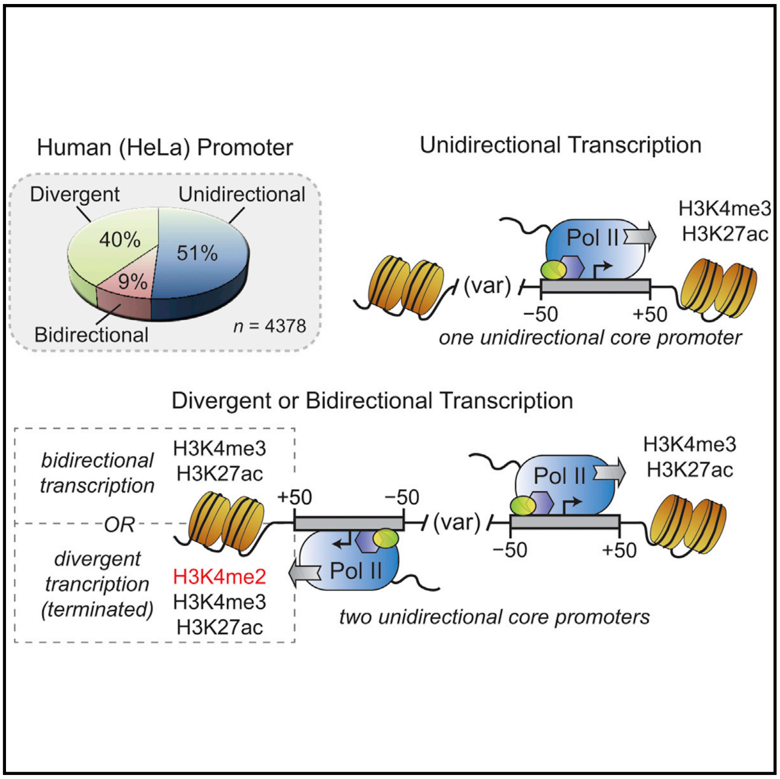
Peer reviewed

Article

Molecular Cell

Human Promoters Are Intrinsically Directional

Graphical Abstract


Human (HeLa) Promoter
Divergent
Unidirectional
40%
51%
9%
Bidirectional
n = 4378
Unidirectional Transcription
Pol II
H3K4me3
H3K27ac
(var)
−50
+50
one unidirectional core promoter
Divergent or Bidirectional Transcription
bidirectional transcription
OR
divergent trancription (terminated)
H3K4me3
H3K27ac
+50
−50
(var)
Pol II
H3K4me3
H3K27ac
−50
+50
H3K4me2
H3K4me3
H3K27ac
Pol II
two unidirectional core promoters

Authors

Sascha H.C. Duttke, Scott A. Lacadie, ..., James T. Kadonaga, Uwe Ohler

Correspondence

jkadonaga@ucsd.edu (J.T.K.),
uwe.ohler@mdc-berlin.de (U.O.)

In Brief

Duttke et al. show that the human basal RNA polymerase II machinery and core promoter are unidirectional and that reverse-oriented transcripts originate from separate reverse-directed core promoters. About half of active HeLa promoters are found as unidirectional, depleted at their upstream edges of core promoter sequences and associated chromatin features.

Highlights

- Basal RNA polymerase II machinery and core promoters are inherently unidirectional
- Divergent transcripts arise from their own core promoters at edges of open chromatin
- Unidirectional promoters are frequent and depleted of reverse core promoter sequences
- Reverse-directed core promoters are associated with a unique chromatin signature

Accession Numbers

GSE63872

Duttke et al., 2015, Molecular Cell *57*, 674–684
February 19, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.molcel.2014.12.029

Human Promoters Are Intrinsically Directional

Sascha H.C. Duttke,[1,8] Scott A. Lacadie,[5,8] Mahmoud M. Ibrahim,[5,6] Christopher K. Glass,[2,3] David L. Corcoran,[7] Christopher Benner,[4] Sven Heinz,[2,4] James T. Kadonaga,[1,*] and Uwe Ohler[5,6,*]
[1]Section of Molecular Biology
[2]Department of Cellular and Molecular Medicine
[3]Department of Medicine
University of California, San Diego, La Jolla, CA 92093, USA
[4]Salk Institute for Biological Studies, La Jolla, CA 92037, USA
[5]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany
[6]Department of Biology, Humboldt University, 10115 Berlin, Germany
[7]Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, USA
[8]Co-first Authors
*Correspondence: jkadonaga@ucsd.edu (J.T.K.), uwe.ohler@mdc-berlin.de (U.O.)
http://dx.doi.org/10.1016/j.molcel.2014.12.029

SUMMARY

Divergent transcription, in which reverse-oriented transcripts occur upstream of eukaryotic promoters in regions devoid of annotated genes, has been suggested to be a general property of active promoters. Here we show that the human basal RNA polymerase II transcriptional machinery and core promoter are inherently unidirectional and that reverse-oriented transcripts originate from their own cognate reverse-directed core promoters. In vitro transcription analysis and mapping of nascent transcripts in HeLa cells revealed that sequences at reverse start sites are similar to those of their forward counterparts. The use of DNase I accessibility to define proximal promoter borders revealed that about half of promoters are unidirectional and that unidirectional promoters are depleted at their upstream edges of reverse core promoter sequences and their associated chromatin features. Divergent transcription is thus not an inherent property of the transcription process but rather the consequence of the presence of both forward- and reverse-directed core promoters.

INTRODUCTION

Bidirectional transcription of oppositely oriented pairs of genes, each of which appears to be expressed from its own core promoter, is commonly observed, especially in compact genomes of model organisms such as yeast (Adachi and Lieber, 2002; Wakano et al., 2012). In mammals, recent studies have also revealed reverse direction transcription initiating upstream of many promoters, largely in the absence of an annotated gene in the reverse orientation. This phenomenon is termed divergent transcription (Core et al., 2008; Preker et al., 2008; Seila et al., 2008), and the resulting transcripts have sometimes been included in annotations of long-noncoding RNA (lncRNA) (Sigova et al., 2013). While divergent transcription has been suggested to be a general feature of eukaryotic promoters, its definition often relies upon arbitrary distance cutoffs, yielding numbers that inevitably increase as longer distances are considered. Furthermore, the near absence of divergent transcription in *Drosophila melanogaster* (Core et al., 2012), which shares many features of transcriptional regulation with other eukaryotes, argues strongly against divergent initiation as being an inherent property of the eukaryotic transcription process in general.

Divergent transcripts are terminated quickly and are subjected to rapid decay through a mechanism involving cleavage/polyadenylation and the nuclear exosome (Almada et al., 2013; Brannan et al., 2012; Ntini et al., 2013), which has been shown to be driven by Nrd1 in yeast (Arigo et al., 2006; Schulz et al., 2013). The process of reverse transcription initiation, on the other hand, remains to be clarified, and many mechanisms have been proposed (Seila et al., 2009). A current model suggests that the presence of CpG islands (CGIs), possibly combined with weak, forward-directed motifs (such as the TATA box), leads to transcription in both directions (Core et al., 2012; Grzechnik et al., 2014; Lepoivre et al., 2013). While this model could potentially explain the lack of divergent transcription in *Drosophila melanogaster* (Core et al., 2012), the sequence and chromatin features that mediate the initiation of divergent transcripts have remained largely speculative.

The core promoter is a fundamental regulator of gene expression. These sequences, which encompass the region that is approximately ±50 bp around the transcription start site (TSS), contain motifs such as the TATA box, Initiator, and downstream core promoter element (DPE) that are recognized by the basal transcription machinery (Butler and Kadonaga, 2002). While a substantial fraction of the extragenic mammalian genome is transcribed at least at minimal levels (Birney et al., 2007; Carninci et al., 2005; Kapranov et al., 2007; Katayama et al., 2005), it is not known if such transcription is mediated by distinct core promoter sequence elements. Hints at such regulation have recently been described by cap analysis gene expression (CAGE) in enhancer regions, where eRNA start sites show some sequence similarities to those in promoter regions (Andersson et al., 2014), and by ChIP-exo for basal transcription factors in yeast where two distinct PICs were detected at divergent promoters (Rhee and Pugh, 2012).

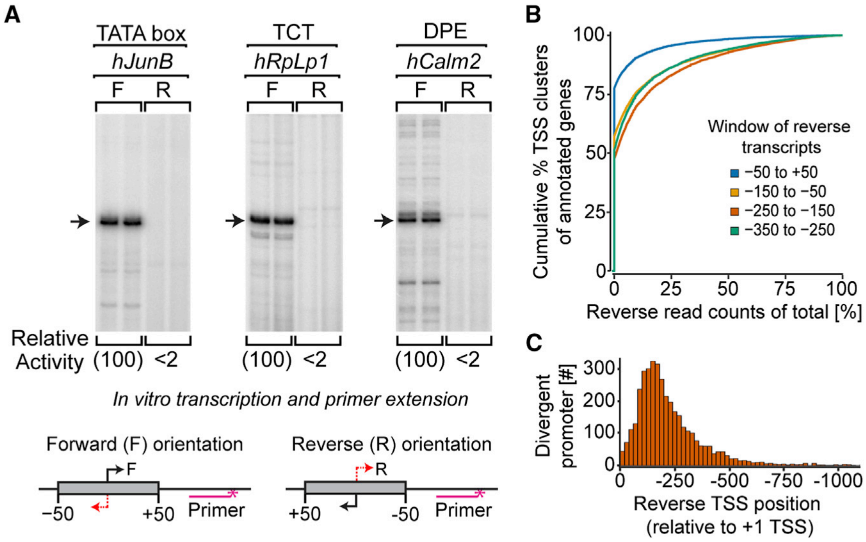


Figure 1. Directional Transcription of Core Promoters

(A) Unidirectional transcription of diverse types of core promoters, ±50 bp in respect to the +1 TSS (marked by the arrow) in vitro. See also Figure S1.

(B) Directionality of the core promoter and promoter regions (n = 15,474) as mapped by 5′-GRO-Seq in HeLa S3 cells, plotted as percent antisense activity (5′ end read counts in a given antisense window divided by that number plus the counts in the forward TSS cluster) for different windows upstream of forward TSS (see Experimental Procedures). Blue = −50 to +50, orange = −150 to −50, red = −250 to −150, and green = −350 to −250.

(C) Distribution of distances between divergent pairs of 5′GRO-Seq-defined TSSs (n = 3,865). Reverse transcription start sites were mapped relative to their corresponding +1 forward start site.

The formation of chromatin structure that facilitates the function of *trans*-regulators is thought to be an important step in gene regulation (Thurman et al., 2012). TSSs occur within nucleosome-free regions (NFRs), which can be detected by their sensitivity to DNase I cleavage and display a large range of lengths (Boyle et al., 2008; Natarajan et al., 2012). At the downstream edges of promoter-associated NFRs, histone H3 that is trimethylated at lysine 4 (H3K4me3) within well-positioned +1 nucleosomes has been shown to stimulate PIC formation (Lauberth et al., 2013). Furthermore, nucleosome positioning and histone modification states can be used to classify promoters associated with different types of transcription initiation patterns (Lenhard et al., 2012; Rach et al., 2011). However, while many histone marks show bimodal chromatin immunoprecipitation sequencing (ChIP-seq) signal patterns around TSSs, these patterns can change depending on RNA polymerase II activity (Bonn et al., 2012). Moreover, the relationship between −1 nucleosome modification and divergent transcription remains to be clarified.

In this study, we show that the basal transcription machinery and the vast majority of core promoters are inherently unidirectional both in vitro and in cells. Maps of nascent RNA 5′ ends, which were obtained by using 5′-GRO-seq (Lam et al., 2013), revealed that divergent transcripts initiate from their own distinct core promoters adjacent to the edges of NFRs, which contain sequences that are similar to those of their forward counterparts. We used DNase I hypersensitivity to define NFRs and thus the borders of proximal promoters and showed that roughly half of active promoters are intrinsically unidirectional and depleted at their upstream edges for such reverse-directed core promoter sequences. A high-resolution hidden Markov model (HMM) of promoter-associated chromatin marks revealed that divergent promoters show enrichment around the −1 nucleosome of a chromatin state containing H3K4me2, H3K4me3, and H3K27ac, a state that is enriched further downstream in the forward direction. In contrast, while all active promoters are flanked by well-positioned nucleosomes, unidirectional promoters have no preferred chromatin state in their upstream regions. These findings suggest that divergent transcription is the consequence of the presence of both forward- and reverse-directed core promoters that are located at the edges of NFRs.

RESULTS

Inherent Unidirectionality of the Basal Transcription Machinery and Core Promoters

To investigate the mechanism of divergent transcription, we tested the inherent directionality of the human core promoter and the basal transcription machinery. The core promoter is the minimal DNA sequence that is required for the accurate initiation of transcription and is typically defined as the region that is about ±50 bp relative to the TSS. Within the core promoter, DNA sequence elements, such as the TATA box, Inr, DPE, and TCT motifs, interact with the basal transcription factors to recruit RNA polymerase II. To determine the directionality of the human core transcription machinery, different types of core promoters (i.e., TATA-, DPE-, and TCT-dependent promoters) were subjected to in vitro transcription analysis with HeLa S3 nuclear extracts. Accurate transcription initiation was observed in the forward direction but not in the reverse direction (Figure 1A). In an exceptional case, divergent initiation was observed from a core promoter with a symmetric TATA element and an Inr in both directions (Figure S1A). Otherwise, these findings indicate that human core promoters and basal transcription machinery can be intrinsically unidirectional in nature.

These biochemical observations were corroborated genome wide by mapping HeLa S3 cell TSSs via 5′ end-selected global run-on followed by sequencing (5′-GRO-seq), which captures initiation events of nascent transcripts at single-nucleotide resolution, irrespective of transcript stability (Kruesi et al., 2013; Lam et al., 2013). After clustering the resulting genome-wide sequence tags, 77.4% (11,985 out of 15,474) of the TSS clusters of annotated genes did not exhibit any reverse direction transcription in the core promoter region (−50 to +50; Figure 1B; blue

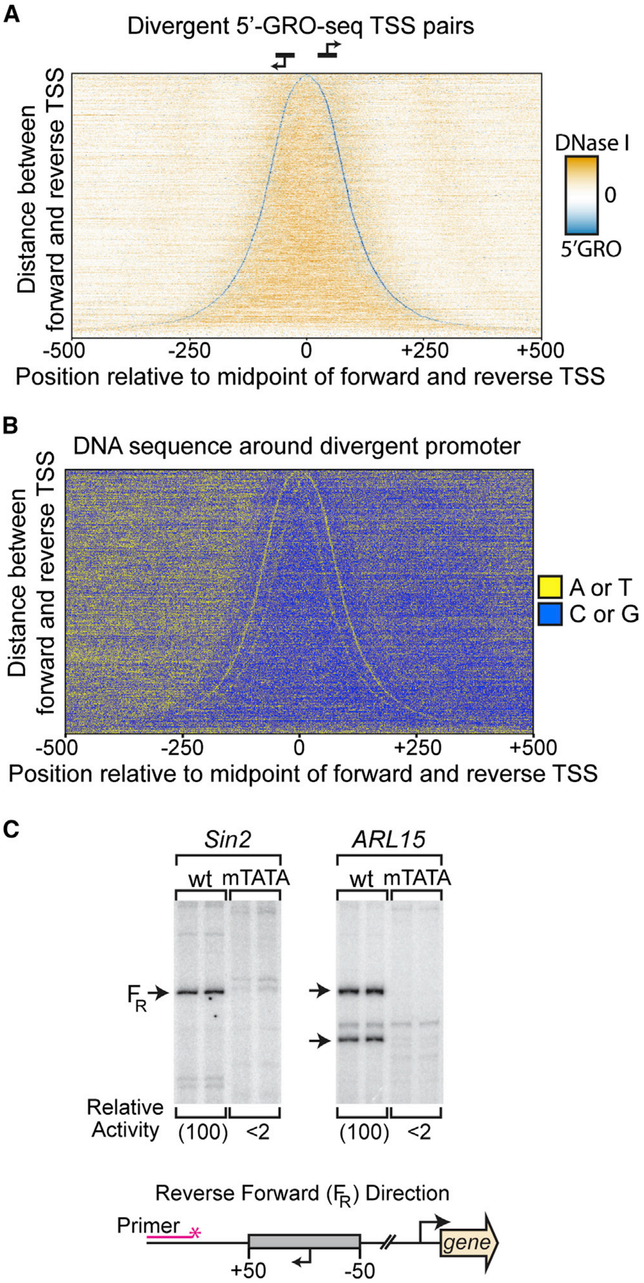


Figure 2. Transcription Initiation from Divergent Core Promoters Occurs at Edges of Open Chromatin

(A) 5′-GRO-Seq (blue) read 5′ end counts and DNaseI-seq (orange) read 5′ end counts in bins of 10, ±0.5 kb from the center point of divergent TSS pairs (n = 3,865; see Experimental Procedures), ranked from top to bottom by increasing distance between pairs.

(B) Genomic DNA sequence of divergent TSS pairs centered and ranked as in '(A).' Bases "A" and "T" are yellow; bases "C" and "G" are blue. See also Figures S2 and S3.

(C) TATA-sensitive in vitro transcription of reverse-directed core promoters. +1 TSS is marked by the arrow.

line). These data thus indicate that most human core promoters are inherently unidirectional both in vitro and in cells.

The locations of reverse direction transcription initiation were found to occur at variable distances from the forward TSSs (Figure 1B). We identified 3,865 divergent promoter pairs, each of which comprises two TSS clusters (each with density signal over background and at least 10 reads; see Experimental Procedures), one associated with an annotated gene (in the forward direction) and one reverse directed in a non-annotated 5 kb upstream genomic region. The preferred distance between divergent TSS pairs is approximately 200 bp (Figure 1C), which is clearly outside of the forward core promoter region.

Reverse Direction Initiation from Distinct Core Promoters at NFR Edges

To examine the potential relationship between reverse direction promoters and chromatin structure, we overlaid our 5′-GRO-seq TSS data with the genome-wide HeLa S3 DNase I-seq data from ENCODE (Bernstein et al., 2012; Thurman et al., 2012).We anchored promoters at the midpoint between the forward and reverse TSSs and ordered them by the paired inter-cluster distance and observed a striking pattern in which the variable distances between the two divergent peaks of 5′-GRO-seq signal are entirely filled by DNase I hypersensitive DNA (Figure 2A). Hence, in divergent promoters, initiating RNA polymerase II flanks the borders of NFRs, consistent with locations where engaged RNA polymerase II (as measured by GRO-seq or TSSa-RNAs) and yeast pre-initiation complexes were shown to accumulate (Core et al., 2008; Rhee and Pugh, 2012; Seila et al., 2008).

The high resolution of the 5′-GRO-seq assay enabled the identification of the most utilized nucleotide (mode) within each TSS cluster (Ni et al., 2010). We reasoned that anchoring plots with respect to these modes might allow us to visualize single-nucleotide promoter sequence preferences and thus gain insights into the specific features of forward and reverse TSSs (Figure 2B). Three immediate observations are apparent. First, it is evident that the increasing width of center-enriched GC content directly corresponds to the NFR (Figure 2B), which is consistent with previous reports showing a direct relationship between GC content and nucleosome positioning (Fenouil et al., 2012). Second, there is a large domain of AT enrichment upstream but not downstream of the NFR, likely reflecting depletion of coding sequence and the recently reported asymmetry of 5′ splice site/cleavage site ratios upstream and downstream of divergent promoters (Almada et al., 2013; Ntini et al., 2013). Third, two symmetric and parallel arches of enriched AT content correspond to the TSS mode (outer arch) and −30 (relative to the TSS mode; inner arch) regions for both the forward and reverse 5′-GRO-seq TSS clusters. Position-specific three-mer frequencies, as well as motif scans for TATA-box and Initiator, suggest that these arches contain initiator-like and TATA-box-like sequences, respectively (Figure S2).

To investigate the core promoter activity of the region that encompasses the reverse direction start sites, DNA sequences from −50 to +50 (relative to the "+1" reverse direction start sites, which we term F_R for "forward" transcripts in the reverse direction) were cloned and subjected to in vitro transcription analysis (Figure 2C). Strong, unidirectional (data not shown), and TATA

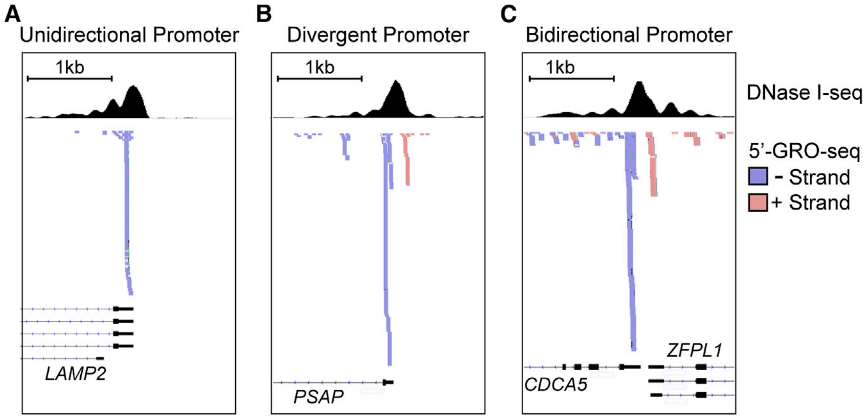


Figure 3. Examples of Divergent, Unidirectional, and Bidirectional Transcription

(A–C) Browser snapshots of examples displaying genes where divergent transcription is absent ([A]; unidirectional), present ([B]; divergent), or occurring at annotated bidirectional genes ([C]; bidirectional). Shown is DNaseI-seq signal as generated by JAMM in black and 5′GRO-seq reads in red for the + strand and blue for the − strand.

box-dependent transcription was detected. These observations, together with high scores from a computational position-specific TSS sequence model (Frith et al., 2008) (Figure S3), show that the reverse direction, non-annotated TSS clusters result from distinct reverse direction core promoters with DNA sequence elements that are similar to those of their forward, TSS-annotated counterparts. Both the forward and reverse core promoters have their own non-overlapping sequences that are enriched within the edges of open chromatin. This two-core-promoter model is consistent with general transcription factor ChIP-exo studies in yeast (Rhee and Pugh, 2012) and is distinct from previously proposed passive models of mammalian divergent initiation such as the nonspecific transcription resulting from the presence of open chromatin, the enhancement of transcription via the carboxy terminal domain of the forward direction polymerase, or the stimulation of reverse transcription via the accumulation of negative supercoiling due to forward direction transcription (Seila et al., 2008, 2009).

Approximately Half of Active HeLa Promoters Are Intrinsically Unidirectional

Visual inspection of raw 5′GRO-seq data together with DNase I-seq made it clear that NFRs also have the potential to harbor promoter regions where transcription occurs in one direction only (Figure 3A). Such examples, together with the observed relationship between divergent TSS pair distances and DNase I sensitivity (Figures 2A and 3B), provided an opportunity to switch from arbitrary distance cutoffs to a concrete definition of promoter regions (comprising a proximal promoter and its associated TSS clusters) based on functional genome-wide data. To this end, we subjected the DNase I-seq data to peak calling using the recent JAMM algorithm (Ibrahim et al., 2015) (Figures S4A–S4C) and intersected these highly accurate DHSs with the 5′GRO-seq TSS data. This approach defined 4,378 promoters containing exactly one TSS-annotated 5′-GRO-seq cluster in the forward direction (Table S3; see Experimental Procedures). Of these 4,378 promoters, 400 (9.1%) had an additional reverse cluster aligning to another annotated TSS (annotated bidirectional; example shown in Figure 3C), 1,741(39.8%) were found to be divergent (i.e., contain a second upstream reverse TSS falling into a non-annotated region), and 2,237 (51.1%) were unidirectional with only one TSS cluster orientated toward the gene.

Our DHS-based promoter region definition thus enabled a comparative analysis of reverse regions—whether or not they are transcribed—for both unidirectional and divergent promoters. We anchored the divergent (Figure 4A) and unidirectional (Figure 4B) promoters based on their DHS centers and ordered them by their DHS widths. Since border proximity of start sites within the DHS was not part of our selection criteria, these plots revealed the symmetric enrichment of 5′-GRO-seq signal around both edges of the divergent promoter DHSs as suggested in Figure 2A and an asymmetric forward edge-only enrichment for the unidirectional promoters. The experimentally/computationally defined DHS edges extend a consistent ~70 bp average distance downstream of all three 5′-GRO-seq TSS cluster groups (divergent forward and reverse and unidirectional forward), suggesting that the TSSs are directly adjacent to either the −1 or +1 nucleosomes (Figures S4D and S4E). ChIP-seq reads for TAF1 or H2AZ verify the lack of transcription initiation upstream of the unidirectional promoters and show strong signal for the +1 and −1 nucleosomes of both promoter groups (Figure 4C). Also, in agreement with a depletion of divergent transcription, the unidirectional promoter DHSs show reduced average signal for TAF1 ChIP-seq, TBP ChIP-seq, and traditional GRO-seq in their reverse regions (Figures S5A–S5C). Importantly, these data suggest that about half of the HeLa expressed, DHS-defined promoters regions are unidirectional and are in stark contrast to the theory that divergent transcription is a general feature of eukaryotic promoters (Neil et al., 2009; Sigova et al., 2013).

Why do DHSs associated with unidirectional promoters lack reverse transcription from their upstream edges? Given our previous observations (Figure 2), we wondered, in particular, whether unidirectional promoters contain functional core promoters in the reverse direction. To test this idea, we turned to the position-specific TSS sequence model (Frith et al., 2008). After training the model on different subsets of core promoter sequences ±50 bp around the forward TSSs of the divergent promoters (see Experimental Procedures), the model was used to scan the upstream and downstream DHS edges of independent unidirectional and divergent promoters (Figure 4D). The model reported high scores about 70 bp upstream of both edges of the divergent promoter DHSs and the forward edge of unidirectional promoter DHSs, consistent with the relative location of 5′-GRO-seq clusters (Figure S4E; see above). In contrast, the upstream edge of the unidirectional promoter DHSs shows an altogether different pattern of lower scores that are more evenly distributed throughout the window (Figure 4D, orange). Thus,

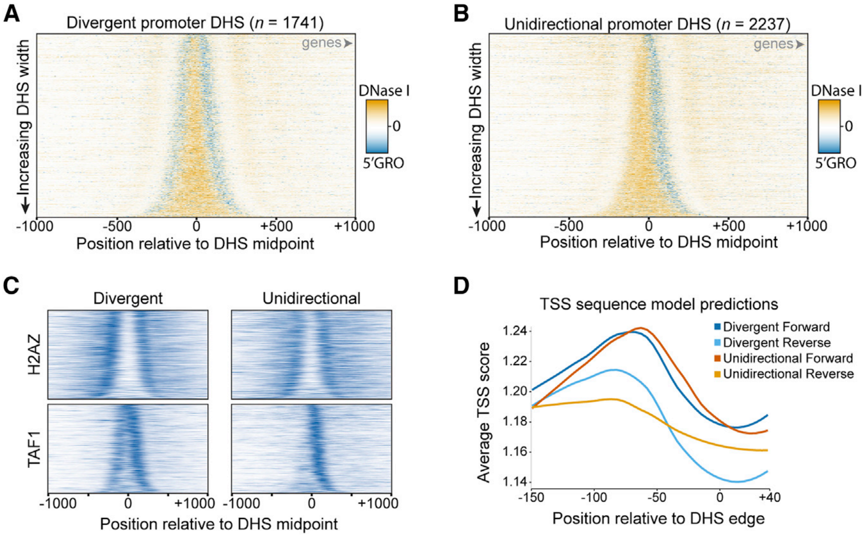


Figure 4. Many Promoter DHSs Lack Core Promoter Sequences Necessary for Divergent Transcription

(A and B) Normalized 5′-GRO-Seq (blue) read 5′ end counts and DNaseI-seq (orange) read 5′ end counts in bins of 10 bp, ±1 kb from the center point of divergent (a; n = 1,741) and unidirectional (b; n = 2,237) promoter DHSs (see Experimental Procedures) ranked from top to bottom by increasing DHS width. See also Figure S4.

(C) Normalized, fragment-extended H2AZ (top) and TAF1 (bottom) ChIP-seq read counts in bins of 10 bp for divergent (left) and unidirectional (right) DHSs centered and ranked as above. See also Figure S5.

(D) Predicted TSS scores around corresponding DHS edges resulting from a position-specific Markov chain model (see Experimental Procedures) trained on ±50 bp around divergent forward TSS. Blue = divergent forward, light blue = divergent reverse, and red = unidirectional forward, orange = unidirectional reverse. See also Tables S1 and S2.

these findings suggest that the unidirectionality of these promoters is due to the lack of a reverse direction core promoter at the upstream edge of the NFR.

Characteristics of Divergent and Unidirectional Promoters

Divergent and unidirectional promoter DHSs show similar frequencies of previously described initiation patterns (Ni et al., 2010) in the forward direction as well as the reverse directions, an observation that suggests mechanistic similarities between forward- and reverse-directed initiation (Figure 5A). In contrast to previously proposed models (Core et al., 2012; Lepoivre et al., 2013), divergent and unidirectional promoters exhibit comparable CGI content (Figure 5B). Furthermore, forward start sites of divergent promoters exhibit a lower percentage of canonical TATA boxes but higher levels of in vivo TBP recruitment than unidirectional promoters (Figures 5C and 5D). While divergent and unidirectional promoter DHSs show some subtle differences in their size, their overall similarities in DHS width and histone ChIP-seq signal bimodality in HeLa cells (Figures 4C, 5E, and S6) suggest that reverse direction transcriptional activity is not necessary for positioning of the −1 nucleosome as previously postulated (Seila et al., 2009). Of note, the reverse TSSs from the divergent group show fewer TATA-like sequences than the forward TSSs (Figures 5C and 5D), consistent with their lower scores in the TSS prediction model (Figure S3D). Divergent promoters also show higher expression levels in the forward direction than unidirectional promoters as measured by ENCODE whole-cell polyA+ CAGE data (Bernstein et al., 2012) (Figure S5D).

Unique Chromatin Environment of Divergent Promoters

It has been previously proposed that divergent transcription could explain the bimodal distribution of many histone modifications around TSSs (Core et al., 2008; Seila et al., 2008). In this regard, our concrete definitions of both divergent and unidirectional promoters enabled us to ask two questions: first, are there differences between the histone modifications of the +1 versus −1 nucleosomes at divergent promoters; second, are the modifications of the unidirectional −1 nucleosome different from those of the divergent −1 nucleosome? To address these questions, we employed a HMM framework for a high-resolution, unsupervised clustering of histone modifications in the HeLa genome based on ENCODE HeLa S3 H3K4me1-3 and H3K27ac ChIP-seq data sets (Bernstein et al., 2012) (see Experimental Procedures). We identified eight genome-wide chromatin states: four proximal promoter states described below; an “inactive enhancer” state characterized by H3K4me1 enrichment; an “active enhancer” state with H3K4me1/H3K27ac; a “transcribed enhancer” state with H3K4me1/H3K4me2/H3K27ac, which intersects strongly with enhancer-RNA-based definitions made by the FANTOM5 Consortium (Andersson et al., 2014) (Figure S6A); and a background state that does not show enrichment for any of the analyzed histone modifications (Figure 6A). A meta-analysis of divergent and unidirectional promoters (Figures 6B and 6C) displays a clear cascade of chromatin state enrichments in the forward directions of both groups, where H3K4me3 and H3K27ac are found together at the +1 nucleosome location (“promoter state1”), followed by the gain of H3K4me2 (“promoter state 2”), then the loss of H3K27ac (“promoter state 3”), and finally the loss of H3K4me3 (“promoter state 4”). Indeed, the same cascade can be observed in both directions of annotated bidirectional promoter DHSs (Figure 6D).

It is particularly notable, however, that in the reverse direction of divergent promoters there is an enrichment of promoter state 2 (H3K4me2-3 and H3K27ac; Figure S6) immediately downstream of the F_R TSS at the −1 nucleosome location (Figure 6B). Promoter state 2 is enriched in the forward direction after promoter state 1, a state clearly absent in the reverse direction. Of note, the preference of promoter state 2 is absent at intergenic transcribed enhancers, which are characterized by high levels of H3K4me1-2 and H3K27ac (Figure 6E). There is also a slight enrichment of the transcribed enhancer state (H3K4me1-2 and H3K27ac; Figure S6) around the −2 nucleosome of the divergent

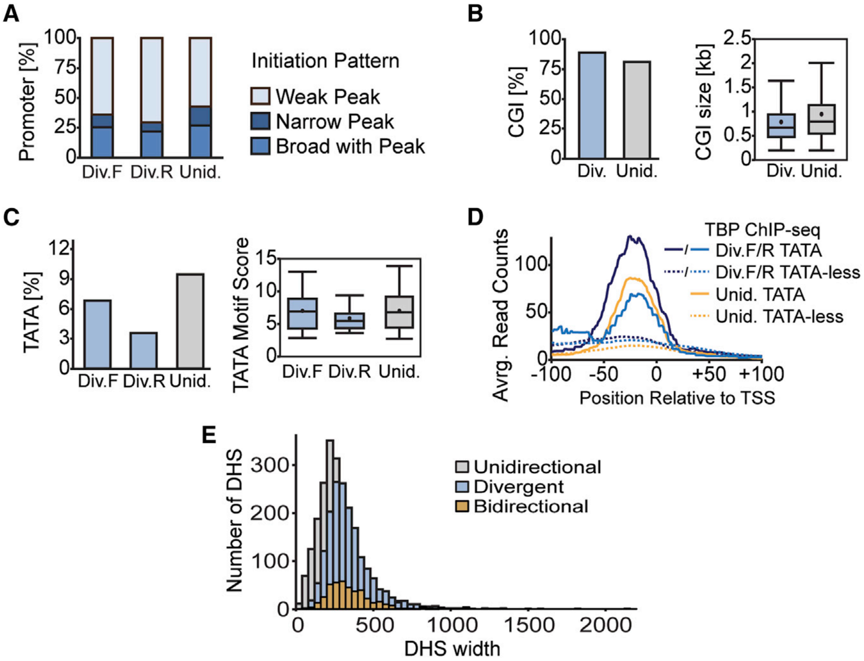


Figure 5. Characteristics of Divergent and Unidirectional Promoters

(A) Percentage of initiation patterns defined by Ni et al (2010) for forward 5′-GRO-seq clusters of the DHS-defined divergent and unidirectional promoter groups and reverse clusters of divergent group.

(B) Percentage of divergent or unidirectional promoter DHSs intersecting an annotated CGI (left; see Experimental Procedures). Size distributions of CGIs that intersect divergent or unidirectional promoter DHSs (right; see Experimental Procedures).

(C) Percent of forward direction 5′-GRO-seq clusters containing a TATA motif match from −35 to −25 relative to the cluster mode for forward 5′-GRO-seq clusters of divergent and unidirectional promoters and reverse clusters of the divergent group (left; see Experimental Procedures) and the distributions of the corresponding scores (right).

(B and C) Whiskers are set according to the default for ggplot2 and extend to the most extreme values within 1.5 times the interquartile distance from the top and bottom of the box. Data points beyond whiskers are outliers and are removed from plots.

(D) Positional average fragment-extended ChIP-seq read counts within TBP peak summits as called by SISSRS in bins of 10 bp for TATA-containing and TATA-less forward and reverse core promoter subsets of divergent or forward only for unidirectional promoters (see Experimental Procedures).

(E) Distributions of DHS widths for unidirectional, divergent, and bidirectional promoter groups. See also Figures S4 and S5 and Table S2.

promoters; this may be the result of the overlap of promoter states 2 and 3 in that region. The lack of chromatin state preference on the reverse side of the unidirectional DHSs (Figures 6C and S6B), despite detectable average signal for all marks (Figures S6C–S6F), suggests that the act of reverse transcription leads to the co-occurrence of the modifications in promoter state 2 (H3K4me2, H3K4me3, and H3K27ac) at the −1 nucleosome.

DISCUSSION

We have proposed a delineation of promoter regions based on DNase I hypersensitivity that reflects the local chromatin environment, is based on functional genomic data, and is independent of a selected distance cutoff. The basal RNA polymerase II transcription machinery initiates unidirectionally from core promoter sequences enriched at one (unidirectional) or both (divergent/bidirectional) edges of such regions. This model is in contrast to the hypothesis that a large majority of human promoters are transcribed in both directions. Reverse-directed core promoters are necessary for divergent transcription, with which they stimulate a unique chromatin signature, but not for −1 nucleosome positioning (Figure 7). While other mammalian studies have shown accumulation of RNA polymerase II activity at the edges of the NFR (Core et al., 2008; Seila et al., 2008), the initiating locations were not known since neither traditional GRO-seq nor TSS-associated small RNAs (TSSa-RNAs) detect actual start sites. The model proposed herein is consistent with ChIP-exo based studies in yeast describing pre-initiation complex formation around both edges of the NFR with a corresponding enrichment of core promoter sequence elements (Rhee and Pugh, 2012). The in vitro studies presented above go a step further and validate the capabilities of such reverse core promoter sequences in basal transcription initiation (Figure 2C).

Our higher estimate of unidirectional promoters compared to previous studies is most likely due to the anchor points and windows considered for measurement. First, while some studies consider windows both upstream and downstream of annotated TSSs to measure “divergent” transcription, this is only necessary to counteract the inaccuracies of such annotations. Since we use data to define our start sites (5′GRO-seq), with genome annotation only serving as a rough guide, we can be confident that our upstream antisense signal is truly “divergent” and excludes downstream “convergent” or “antisense” events. Second, the common practice of using a uniform window size (i.e., ± 1 or 2 kb around TSSs) for all promoter regions is likely to overestimate divergent activity due to the relationship between reverse TSSs and the NFR (Figure 2A), the typical NFR width of ~250 bp (Figure 5E), and the potential for other independent transcribed proximal regulatory elements and/or RNA gene loci. If applied equally to forward- and reverse-directed transcription, changes in background cutoffs or sequencing depth is likely to change the number of both unidirectional and divergent promoters while keeping the ratio relatively constant. We tested this idea on previously published exosome-knockdown CAGE data (Ntini et al., 2013), a technique also capable of mapping TSSs of rapidly degraded transcripts (Table S1), and observed a comparable percentage of unidirectional DHSs (47% by CAGE versus 51% by 5′GRO-seq). It is, however, possible that HeLa cells display a higher percentage of unidirectional promoters than other cell types and that some genes

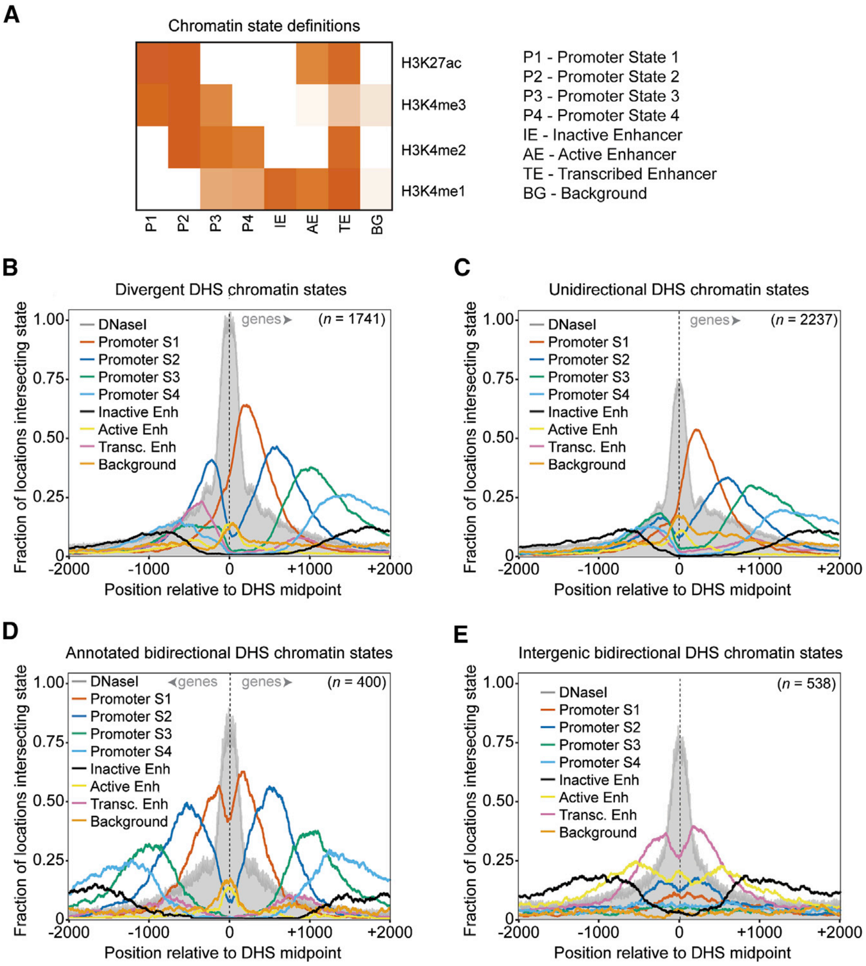


Figure 6. Distinct Chromatin Environment at Unidirectional, Bidirectional, and Divergent Promoter DHSs

(A) Chromatin state definitions based on HMM clustering of histone modification ChIP-seq signal at 10-bp resolution (see Experimental Procedures). Each state is a multivariate Gaussian distribution. Shown are the distribution mean vectors representing scaled, normalized ChIP-seq signal.

(B–E) Chromatin state coverage ±2 kb around the center of divergent promoter DHS (B), unidirectional promoter DHS (C), bidirectional promoter DHS (D), and divergent intergenic DHS (E) at single nucleotide resolution. Grey = DNaseI-seq read 5′ end counts, red = Promoter State1, blue = Promoter State 2, green = Promoter State 3, light blue = Promoter State 4, black = Inactive Enhancer, yellow = Active Enhancer, pink = Transcribed Enhancer, and orange = Background. See also Figure S6.

may be regulated by unidirectional and divergent alternative promoters.

A small group of core promoters may be intrinsically bidirectional (Figures 1B and S1A). In these cases, it is likely that the core promoter element configuration permits bidirectionality (as evidenced by the "crossing" of the AT-rich sequences in the −30 region at the top of Figure 2B). However, the vast majority of core promoters are inherently unidirectional: of 3,865 divergent TSS pairs, ~95% initiate greater than 50 bp upstream of the forward TSS. While the sequences of reverse-directed core promoters are very similar to those of their forward counterparts, there are differences as evidenced by reduced enrichment of AT content in the −30 region (Figures S2A and S2B) and lower scores in the TSS prediction model (Figures S3C, S3D, and 4D). This difference may be reflected in the overall lower levels of basal transcription factor recruitment and transcription from reverse-directed core promoters (Figure S4A–S4C) as well as the lower levels of −1 nucleosome histone modification, compared to the +1 nucleosome, in divergent promoters (Figure S5C–S5F). Indeed, while forward TSS prediction scores and H3K27ac ChIP-seq signal both correlate with forward 5′-GRO-seq signal, these correlations are slightly reduced on the reverse side of the divergent DHSs (Table S2).

Divergent transcripts are now known to often be terminated quickly and subjected to rapid decay (Almada et al., 2013; Brannan et al., 2012; Ntini et al., 2013). The reported enrichment of AT-rich cleavage/polyadenylation sequences upstream of divergent TSSs, which lead to this termination/decay mechanism, is reflected in the increased "yellow" color (AT content) on the left side of Figure 2B. Connecting these observations with the relative shift of chromatin promoter state 2 upstream versus downstream in divergent promoters (Figure 6B) suggests that the location of histone modifications is sensitive to extended transcription elongation and/or nuclear RNA decay rates. Alternatively, this shift in chromatin states could reflect the subtle differences in the core promoter sequences between the forward and reverse directions (Figures S2A–S2C, S3C, and S3D). Histone H3K4 methyltransferases are known to be associated with the carboxyl terminal domain of RNA polymerase II (Greer and Shi, 2012; Hsin and Manley, 2012), supporting the observation that the –I nucleosomes of unidirectional promoters lack coordinated histone modification (Figures 6C and S6).

The frequent but not universal presence of reverse-directed core promoters poises them as candidate regulators of forward transcription. Such a regulation may be reflected in the overall higher expression of both bidirectional and divergent promoters, compared to unidirectional promoters, as measured by basal transcription factor recruitment, histone modification levels, and whole-cell CAGE tag counts (Figures S5 and S6). This points to a possible mechanism whereby reverse-directed core promoters within the upstream edges of divergent or bidirectional promoter DHSs increase local concentrations of initiation machinery, resulting in increased expression of the forward gene (Figure S5D). In this study, we have been able to clarify the functional similarities and differences between unidirectional and

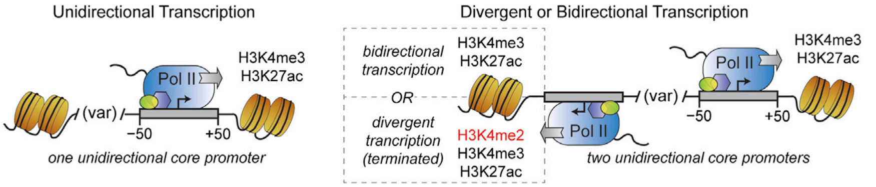

Unidirectional Transcription
Pol II
H3K4me3
H3K27ac
(var)
−50
+50
one unidirectional core promoter
Divergent or Bidirectional Transcription
bidirectional transcription
OR
divergent trancription (terminated)
H3K4me3
H3K27ac
H3K4me2
H3K4me3
H3K27ac
Pol II
(var)
−50
+50
Pol II
H3K4me3
H3K27ac
two unidirectional core promoters

Figure 7. Model of Divergent, Bidirectional, and Unidirectional Promoters

For each type of promoter, +1 and −1 nucleosome positions occur at variable spacing from each other, forward gene transcription initiates just inside the downstream edge of the NFR, and the +1 nucleosome is modified with H3K4me3 and H3K27ac. When transcription initiation occurs from the upstream NFR edge on the opposite strand from the forward gene, the −1 nucleosome gets similarly modified when stable, annotated transcripts are present (for bidirectional promoters), or is enriched for H3K4me2, in addition to H3K4me3 and H3K27ac, when divergent transcription occurs (i.e., when unstable non-coding transcripts are generated).

divergent promoters, but the underlying reasons, if any should exist, why promoters are unidirectional or divergent remain to be illuminated.

EXPERIMENTAL PROCEDURES

Cell Culture Conditions

HeLa S3 cells were grown at 37°C in DMEM (Cellgro) supplemented with 10% FBS (GIBCO), 50 U penicillin, and 50 μg streptomycin per ml (GIBCO).

In Vitro Transcription Assays

Core promoter sequences, ±50 bp in respect to the +1 TSS, were cloned, and transcription reactions were carried out as described previously (Duttke, 2014). Transcripts were subjected to primer extension analysis and separated by urea-polyacrylamide gel electrophoresis.

5′GRO-Seq and GRO-Seq Preparation

5′GRO-seq was performed as described previously (Lam et al., 2013). Briefly, about 10^7 HeLa S3 nuclei were used for run-on with BrU-labeled NTPs. Fragmented transcripts were incubated with polynucleotide kinase (PNK, NEB) to remove 3′ phosphates. BrU-labeled nascent transcripts were subsequently immunoprecipitated with anti-BrdU agarose beads (Santa Cruz Biotech). For 5′GRO-seq, immunoprecipitated RNA was dephosphorylated with calf intestinal phosphatase (NEB). Then 5′ capped fragments were de-capped with tobacco acid pyrophosphatase (Epicenter). Illumina TruSeq adapters were ligated to the RNA 3′ and 5′ ends with truncated mutant RNA ligase 2 (K227Q) and RNA ligase 1 (NEB), respectively. Reverse transcription was performed with Superscript III (Invitrogen) followed by PCR amplification for 12 cycles. Final libraries were size selected on PAGE/TBE gels to 175–225 bp.

GRO-seq was essentially performed as 5′GRO-seq, but the immunoprecipitated RNA was directly de-capped with tobacco acid pyrophosphatase (Epicenter) and subsequently kinased with PNK (NEB) prior to adaptor ligation.

5′-GRO-Seq and GRO-Seq Analysis

Two replicates of 5′ end sequenced reads from the 5′-GRO-seq or traditional GRO-seq protocols were trimmed for adapters using cutadapt (Martin, 2011) and mapped together to the hg19 human genome using Bowtie2 with default settings (Langmead and Salzberg, 2012). Reads that did not map uniquely and reads overlapping rRNA loci were removed, yielding 27,512,149 5′-GRO-seq reads and 21,765,842 traditional GRO-seq reads. Clusters were identified according to the strategy described in Ni et al. (2010). To annotate the identified clusters, the Genomic Features R package (Lawrence et al., 2013) was used with the UCSC knownGenes table.

DNase-Seq and ChIP-Seq Analysis

All five data sets of ENCODE-mapped DNase-seq reads for HeLa-S3 cells were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012) and concatenated before peak calling with JAMM v1.0.6 (Ibrahim et al., 2015) (http://code.google.com/p/jamm-peak-finder/, settings: -m narrow -f 1). HeLa-S3 cell TAF1, TBP, and histone modification ChIP-seq raw fastq files were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012). Reads were aligned to hg19 genome using Bowtie2 (Langmead and Salzberg, 2012) with default parameters and then filtered for those that did not align uniquely or had more than two mismatches. For TAF1 and TBP, replicate BED files were then concatenated before peaks were called using SISSRS (Narlikar and Jothi, 2012), while JAMM was used for histone modification peak calls that served as input to the HMM (see below and Supplemental Experimental Procedures).

Closest Upstream Antisense Pair Assignments

In order to define a set of 5′-GRO-seq cluster pairs that were reciprocally the closest upstream antisense of each other, a combination of BEDTools and custom scripts was used. BEDTools *closest* command (Quinlan and Hall, 2010) (settings: -S -id -D “a”) was run on the modes of 5′-GRO-seq clusters (the cluster position with the highest read count) using the same file for both inputs. Custom Perl scripts were then used to parse the BEDTools output for only those cluster pairs where both modes were called as closest upstream antisense of each other.

DHS-Defined Promoter Borders

BEDTools (Quinlan and Hall, 2010) intersect command was used to find overlaps between DNaseI-seq peak calls (defining DHSs) and 5′-GRO-seq cluster modes, both described above. DHSs with exactly one intersecting TSS cluster mode were considered unidirectional. DHS with exactly two intersecting 5′-GRO-seq cluster modes, for which the two modes were upstream and antisense of each other, one annotating as TSS and the other as intergenic, were considered divergent. DHSs with more than one intersecting 5′-GRO-seq cluster modes on any one DNA strand, or with two 5′-GRO-seq cluster modes on opposite strands but downstream of each other, were removed from further analysis. Unidirectional classified DHSs intersecting reverse-side annotated TSSs (yet having no 5′GRO-seq clusters) or containing exactly one TSS-annotating cluster mode that was also part of divergent or bidirectional reciprocal closest upstream antisense selections (described above) were considered ambiguous and removed from further analysis.

Plotting

All plots were made using the ggplot2 R package (Wickham, 2009). Anchor points were set by calculating the center point between the 5′-GRO-seq cluster modes of the paired forward/reverse TSS clusters or the center point of DHSs. Strand assignments were made according to the forward gene for divergent cluster pairs or unidirectional promoter DHS. For TSS-TSS or intergenic-intergenic 5′-GRO-seq cluster pairs, the cluster with higher read counts was used for strand assignment. For heat maps, windows were ranked from top to bottom by increasing distance between forward and reverse TSSs, or by DHS width. Browser snapshots were taken using the integrative genomics viewer (Thorvaldsdóttir et al., 2013).

Heat Map- and Meta-Analyses

For 5′-GRO-seq and DNaseI-seq heat maps, the number of reads whose 5′ end mapped to each position were counted independent of strand and scaled so that the minimum value for each window is 0 and the maximum value is 1. For TAF1 and H2AZ ChIP-seq heat maps, reads were extended by the fragment size calculated by JAMM, and the number of extended reads falling in 10 bp bins was plotted as above.

For sequence heat maps, BEDTools getfasta command (Quinlan and Hall, 2010) was used to retrieve the sequence corresponding to each window.

TSS Initiation Pattern Analysis

NarrowPeak, BroadPeak, and WeakPeak initiation patterns as defined previously (Ni et al., 2010) were determined from the 5′-GRO-seq clusters with at least 25 read counts.

Probabilistic Model of TSSs

We estimated parameters for a previously published position-specific Markov chain TSS model (PSMM) (Frith et al., 2008) using a first-order setting. A 10-fold cross validation scheme of the PSMM (see Supplemental Experimental Procedures) was implemented. Receiver operator characteristic and precision recall curves were generated by defining true positives as the modes of 5′-GRO-seq clusters and true negatives as every other nucleotide in the tested windows, the results plotted for each of the ten models from the closest upstream antisense selection using the R package ROCR (Sing et al., 2005).

Motif Scanning

The TRANSFAC TATA-box binding protein (M00252) or JASPAR Initiator position weight matrices were used with the Scanner Toolset (Megraw et al., 2009) to scan sequences −35 to −25 upstream for TATA and ± 5 for initiator around the forward or reverse TSS modes of the divergent and unidirectional promoter groups (see Supplemental Experimental Procedures).

CGI Analysis

Genomic coordinates of CGI were taken from the UCSC table browser (Kuhn et al., 2013), reportedly calculated according to the criteria of Gardiner-Garden and Frommer (1987). Either divergent or unidirectional DHSs were intersected with these coordinates using BEDTools intersect (Quinlan and Hall, 2010).

Chromatin State Segmentation

We employed a continuous HMM, in which state emissions are represented by a multivariate Gaussian distribution fully defined by its means vector, corresponding to the signals' means of the histone modification tracks (see Supplemental Experimental Procedures), and its co-variance matrix. To learn the emission and transition parameters of the HMM, we employ the Baum-Welch algorithm (Bilmes, 1997; Taramasco and Bauer, 2013), initialized via k means, on "semi-binarized" signal tracks of chromosome 1 at 10-bp resolution (see Supplemental Experimental Procedures). The mean vector for each state defines the average ChIP-Seq signals of the histone modification tracks in the corresponding state. We 0-to-1 scale the means across each histone modification to define the prototypical chromatin states shown in Figure 6A. Finally, we employ the Viterbi decoding algorithm (Taramasco and Bauer, 2013; Viterbi, 1967) to assign a chromatin state to each 10-bp bin in the genome that had a peak in at least one of the histone modification tracks based on the HMM model learned by the Baum-Welch algorithm. Locations that did not have a peak in any histone modification track are not assigned a state. Book-ended bins that have the same state are merged. The output of this process is genome segmentation into variable-width non-overlapping chromatin states similar to Segway (Hoffman et al., 2012) and ChromHMM (Ernst and Kellis, 2012).

Chromatin State Analysis

Chromatin state coverage plots were calculated by intersecting the promoter regions with state assignments at single-basepair resolution using BEDTools (Quinlan and Hall, 2010) intersect command and plotting the fraction of each position across promoters for each state.

ACCESSION NUMBERS

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE63872.

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, three tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2014.12.029.

AUTHOR CONTRIBUTIONS

S.H.C.D. performed all in vitro assays. S.H.C.D and S.H. performed 5′-GRO-seq and GRO-seq assays with guidance and support from C.B. and C.K.G. D.L.C. wrote the program for TSS peak calling. M.M.I. performed all peak calling and generated the HMM and its associated analyses. S.A.L. performed all other analyses. S.H.C.D. and J.T.K. initiated the study and S.A.L., M.M.I., and U.O. designed the overall analysis framework. S.A.L., S.H.C.D., M.M.I., J.T.K., and U.O. wrote the manuscript.

ACKNOWLEDGMENTS

We are grateful to Dr. Carmelo Ferrai for critical reading of the manuscript and to Dr. Neelanjan Mukherjee for useful discussions. J.T.K. is the Amylin Chair in the Life Sciences. S.H.C.D. is the recipient of the University of California at San Diego Molecular Biology/Cancer Center Fellowship. This work was supported by a grant from the National Institute of Health to J.T.K. (R01 GM041249). M.I. is supported by the MDC-NYU graduate program.

Received: October 1, 2014
Revised: November 19, 2014
Accepted: December 16, 2014
Published: January 29, 2015

REFERENCES

Adachi, N., and Lieber, M.R. (2002). Bidirectional gene organization: a common architectural feature of the human genome. Cell *109*, 807–809.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360–363.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al.; FANTOM Consortium (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol. Cell *23*, 841–851.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bilmes, J. (1997). A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models. International Computer Science Insitute, ICSI-TR 97.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., and Furlong, E.E.M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat. Genet. *44*, 148–156.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311–322.

Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., Kiemele, L., Hansen, K., Davis, R., Lykke-Andersen, J., and Bentley, D.L. (2012). mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. Mol. Cell *46*, 311–324.

Butler, J.E.F., and Kadonaga, J.T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. *16*, 2583–2592.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559–1563.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep. *2*, 1025–1035.

Duttke, S.H.C. (2014). RNA polymerase III accurately initiates transcription from RNA polymerase II promoters in vitro. J. Biol. Chem. *289*, 20396–20404.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., and Andrau, J.C. (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. *22*, 2399–2408.

Frith, M.C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., and Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. Genome Res. *18*, 1–12.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Greer, E.L., and Shi, Y. (2012). Histone methylation: a dynamic mark in health, disease and inheritance. Nat. Rev. Genet. *13*, 343–357.

Grzechnik, P., Tan-Wong, S.M., and Proudfoot, N.J. (2014). Terminate and make a loop: regulation of transcriptional directionality. Trends Biochem. Sci. *39*, 319–327.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods *9*, 473–476.

Hsin, J.-P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev. *26*, 2119–2137.

Ibrahim, M.M., Lacadie, S.A., and Ohler, U. (2015). JAMM: A Peak Finder for Joint Analysis of NGS Replicates. Bioinformatics *31*, 48–55.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermüller, J., Hofacker, I.L., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484–1488.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., et al.; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium (2005). Antisense transcription in the mammalian transcriptome. Science *309*, 1564–1566.

Kruesi, W.S., Core, L.J., Waters, C.T., Lis, J.T., and Meyer, B.J. (2013). Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. eLife *2*, e00808.

Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Brief. Bioinform. *14*, 144–161.

Lam, M.T.Y., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M., et al. (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511–515.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lauberth, S.M., Nakayama, T., Wu, X., Ferris, A.L., Tang, Z., Hughes, S.H., and Roeder, R.G. (2013). H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. Cell *152*, 1021–1036.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat. Rev. Genet. *13*, 233–245.

Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., Zacarias-Cabeza, J., Garibal, M.-A., Koch, F., Maqbool, M.A., et al. (2013). Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics *14*, 914.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. Journal. http://dx.doi.org/10.14806/ej.17.1.200.

Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., and Hatzigeorgiou, A.G. (2009). A transcription factor affinity-based code for mammalian transcription initiation. Genome Res. *19*, 644–656.

Narlikar, L., and Jothi, R. (2012). ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder. Methods Mol. Biol. *802*, 305–322.

Natarajan, A., Yardimci, G.G., Sheffield, N.C., Crawford, G.E., and Ohler, U. (2012). Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res. *22*, 1711–1722.

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature *457*, 1038–1042.

Ni, T., Corcoran, D.L., Rach, E.A., Song, S., Spana, E.P., Gao, Y., Ohler, U., and Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat. Methods *7*, 521–527.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nat. Struct. Mol. Biol. *20*, 923–928.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851–1854.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rach, E.A., Winter, D.R., Benjamin, A.M., Corcoran, D.L., Ni, T., Zhu, J., and Ohler, U. (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. PLoS Genet. *7*, e1001274.

Rhee, H.S., and Pugh, B.F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature *483*, 295–301.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. Cell *155*, 1075–1087.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849–1851.

Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle *8*, 2557–2564.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., and Young, R.A. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. USA *110*, 2876–2881.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940–3941.

Taramasco, O., and Bauer, S. (2013). RHmm: Hidden Markov Models simulations and estimations. https://r-forge.r-project.org/R/?group_id=85.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory *13*, 260–269.

Wakano, C., Byun, J.S., Di, L.-J., and Gardner, K. (2012). The dual lives of bidirectional promoters. Biochim. Biophys. Acta *1819*, 688–693.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. (New York: Springer).

Note Added in Proof

The authors wish to point the reader to a concurrent study addressing divergent transcription with similar methods in other cell types: Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320.

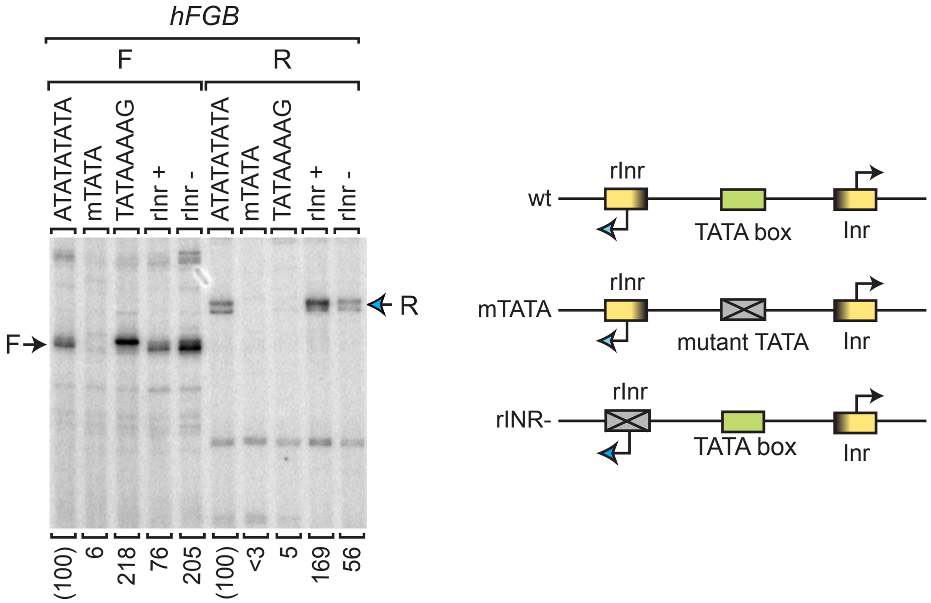
Molecular Cell, Volume *57*

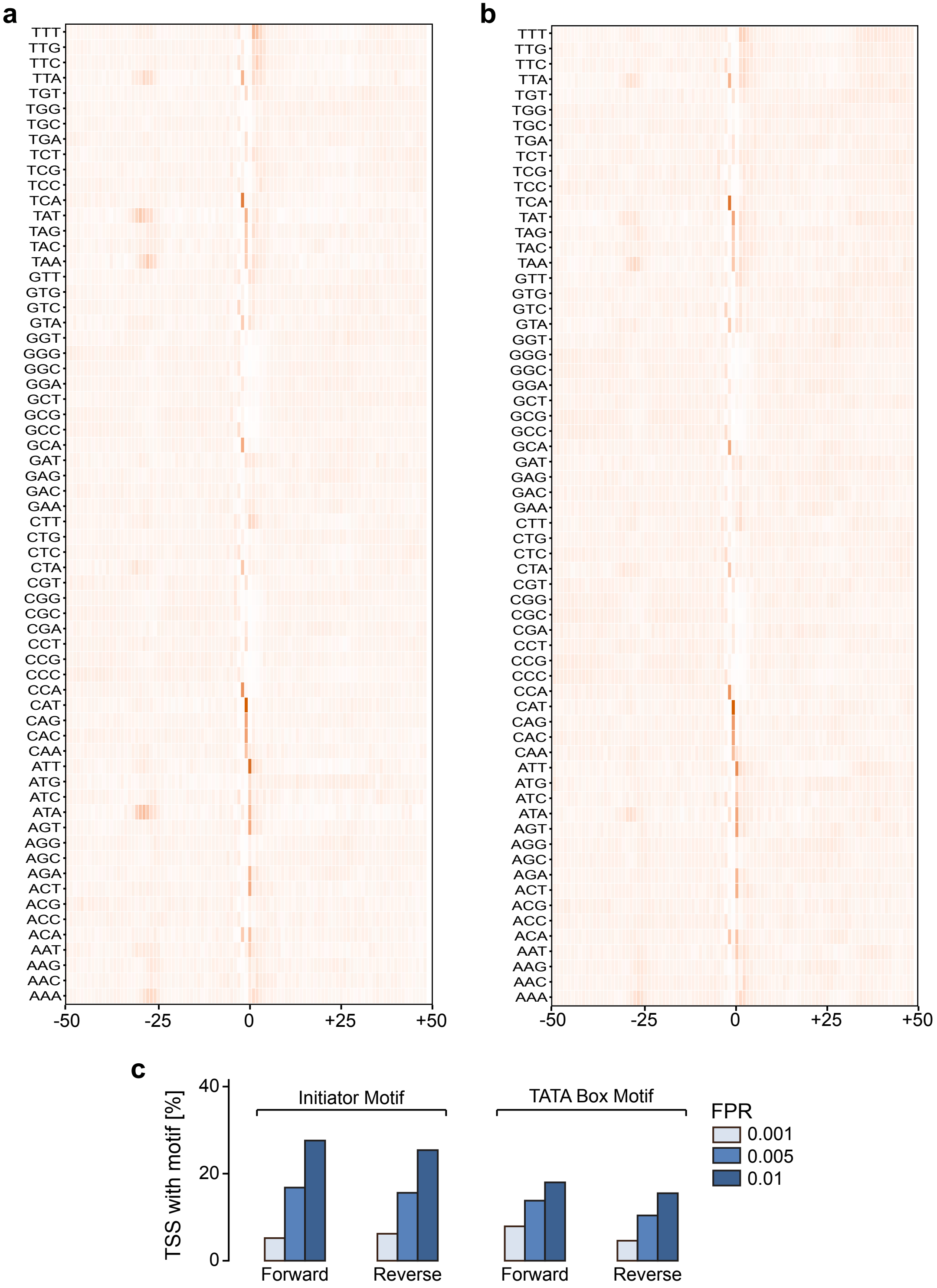
Supplemental Information

Human Promoters Are Intrinsically Directional

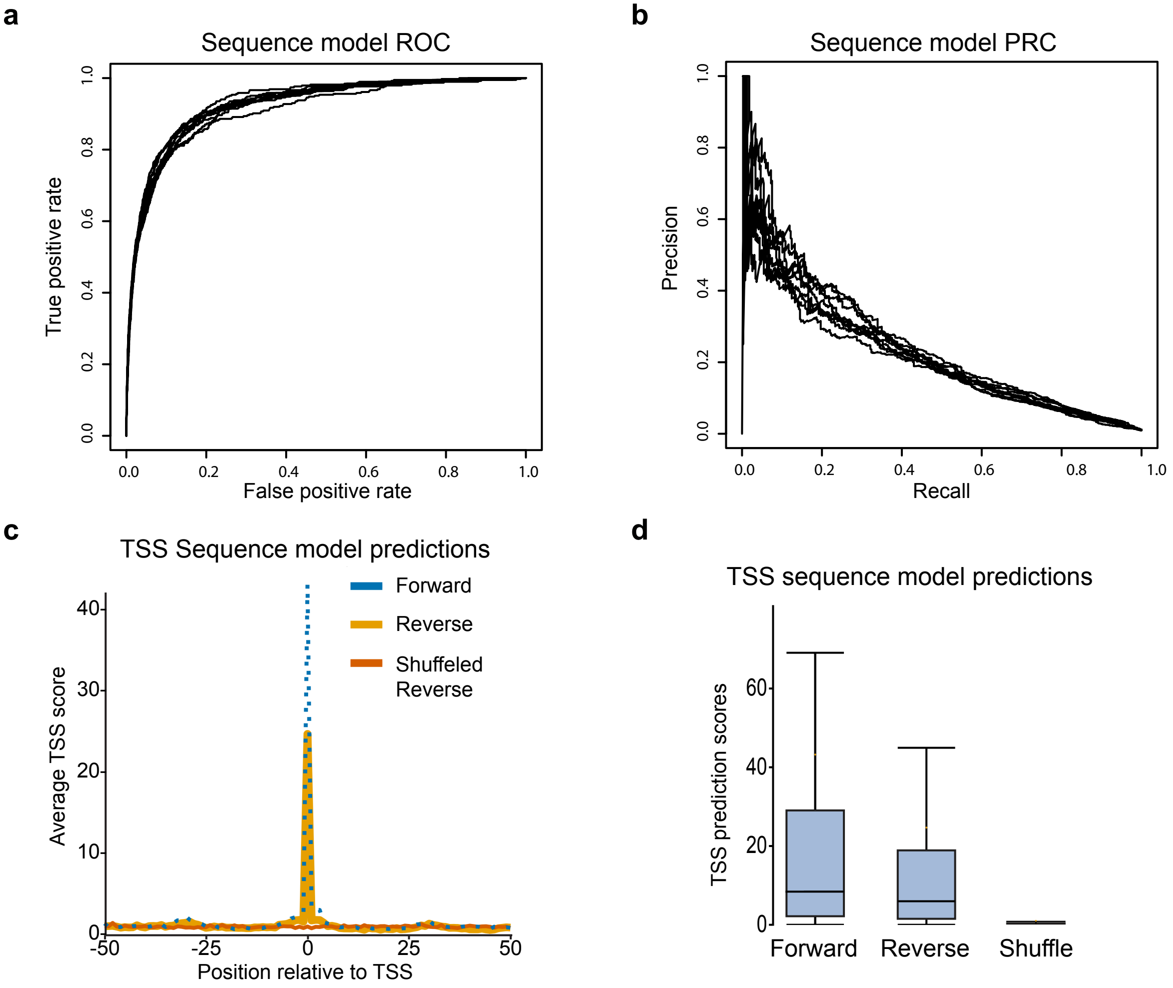
Sascha H.C. Duttke, Scott A. Lacadie, Mahmoud M. Ibrahim, Christopher K. Glass,
David L. Corcoran, Christopher Benner, Sven Heinz, James T. Kadonaga, and Uwe Ohler

Supp. Figure-1 (Ohler)

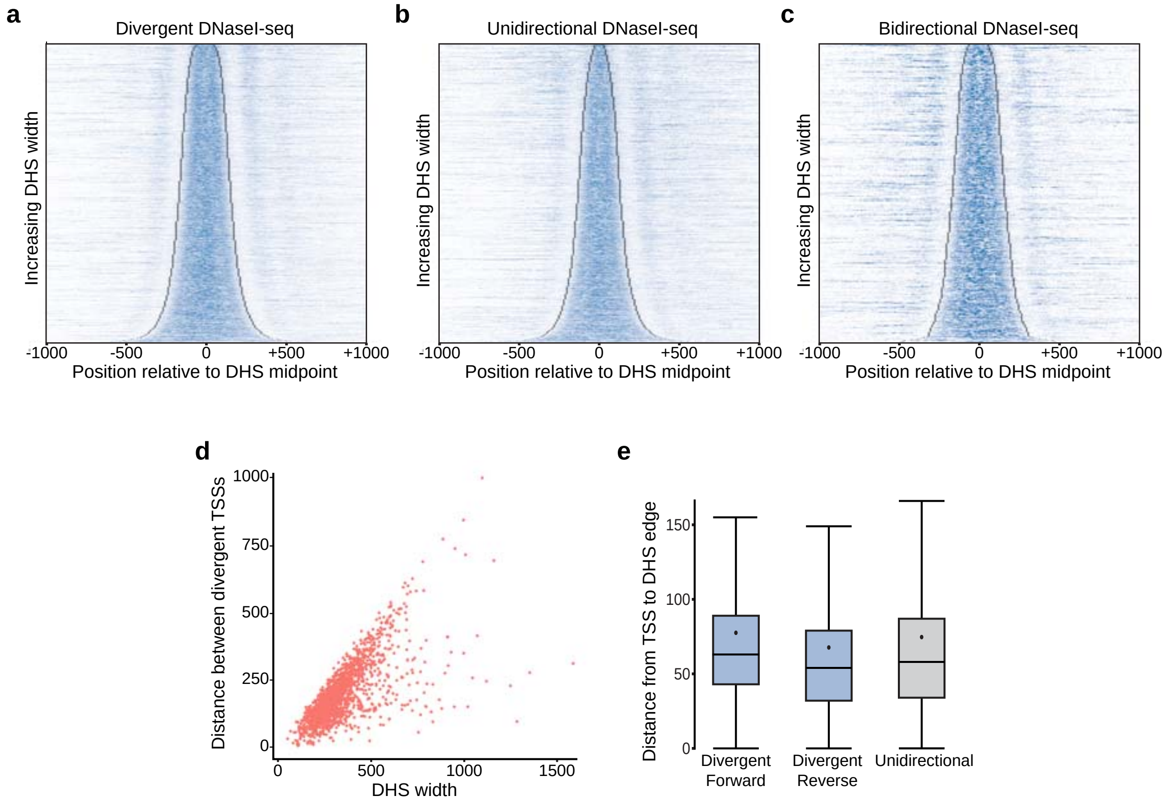



a
TTT
TTG
TTC
TTA
TGT
TGG
TGC
TGA
TCT
TCG
TCC
TCA
TAT
TAG
TAC
TAA
GTT
GTG
GTC
GTA
GGT
GGG
GGC
GGA
GCT
GCG
GCC
GCA
GAT
GAG
GAC
GAA
CTT
CTG
CTC
CTA
CGT
CGG
CGC
CGA
CCT
CCG
CCC
CCA
CAT
CAG
CAC
CAA
ATT
ATG
ATC
ATA
AGT
AGG
AGC
AGA
ACT
ACG
ACC
ACA
AAT
AAG
AAC
AAA
-50
-25
0
+25
+50
b
TTT
TTG
TTC
TTA
TGT
TGG
TGC
TGA
TCT
TCG
TCC
TCA
TAT
TAG
TAC
TAA
GTT
GTG
GTC
GTA
GGT
GGG
GGC
GGA
GCT
GCG
GCC
GCA
GAT
GAG
GAC
GAA
CTT
CTG
CTC
CTA
CGT
CGG
CGC
CGA
CCT
CCG
CCC
CCA
CAT
CAG
CAC
CAA
ATT
ATG
ATC
ATA
AGT
AGG
AGC
AGA
ACT
ACG
ACC
ACA
AAT
AAG
AAC
AAA
-50
-25
0
+25
+50
c
TSS with motif [%]
40
20
0
Initiator Motif
TATA Box Motif
Forward
Reverse
Forward
Reverse
FPR
0.001
0.005
0.01

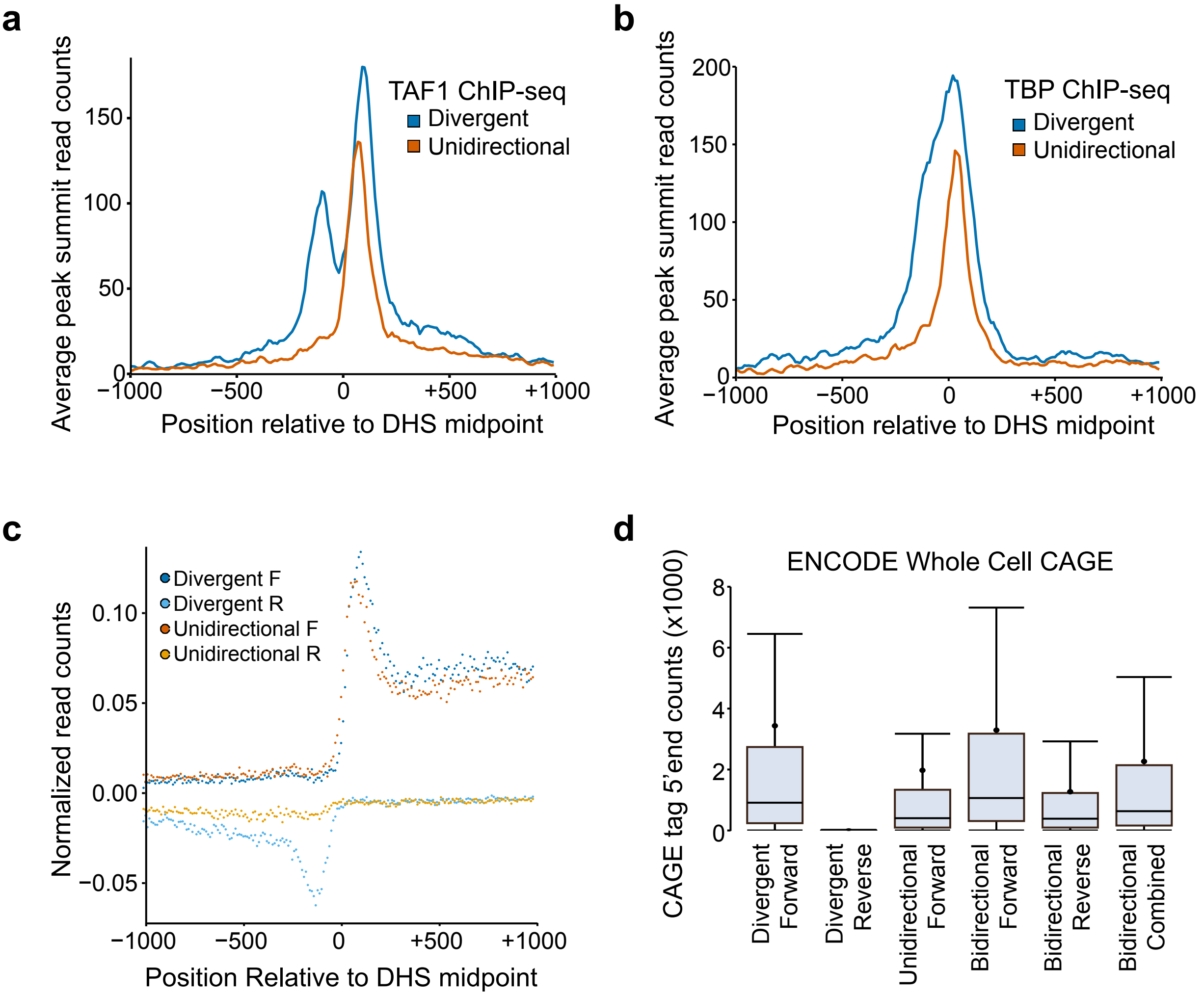
Supp. Figure-3 (Ohler)



Supp. Figure-4 (Ohler)



Supp. Figure-5 (Ohler)



Supp. Figure-6 (Ohler)

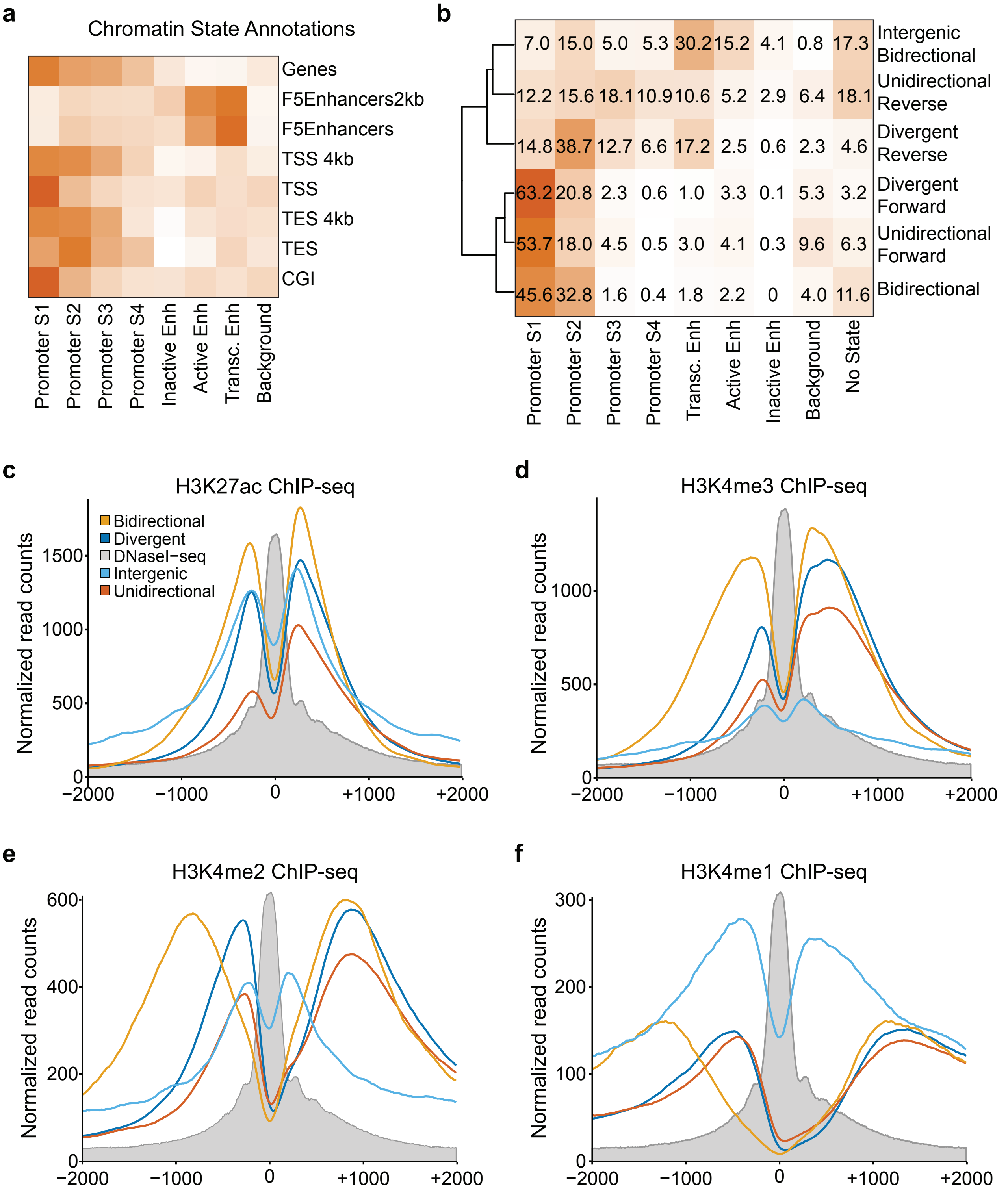


Figure S1 | Transcription from a divergent core promoter, related to Figure 1. Human FGB as an example of a divergent core promoter. The polarity depends on the DNA sequence. Promoters were cloned from +50 to -125 (relative to the +1 transcription start site) to allow reverse initiation within the natural sequence. The reverse Inr (rInr) sequence “TCAGAA” was substituted with “TCGGTC” (rInr-) or a consensus Inr “TCAGTC”(rInr+).

Figure S2 | Sequence content of forward and reverse TSSs, related to Figure 2.
a,b, Position-specific threemer counts normalized to total threemer frequencies for forward (a) and reverse (b) direction core promoters -50 to +50 bp around the 5’-GRO-seq cluster modes. **c,** Percent of forward or reverse TSSs that show motif matches to either initiator (left) or TATA-box (right) in the -35 to -25 or -5 to +5 regions, respectively, from the 5’-GRO-seq cluster modes. Different colors represent different false positive rate (FPR) cutoffs.

Figure S3 | Performance and results of TSS sequence model, related to Figures 2 and 4. a,b, Receiver operator characteristic (a) and precision-recall (b) curves for the sequence model described in Frith et al, 2008, trained and tested with a 10-fold cross validation +/- 50 bp around the mode of the forward TSSs from the divergent promoter pairs described in Figure 2 (see Experimental Procedures). **c,** Average predicted TSS scores per position for sequences +/- 50 bp around the mode of the corresponding TSSs from the divergent promoter pairs described in Figure 2, or its shuffled control, from the model trained as in “a” and “b” (see Supplemental Experimental Procedures). **d,** Distributions of 5’GRO-seq cluster mode TSS prediction scores for forward and reverse TSSs.

Figure S4 | DHS peak call accuracy and characteristics, related to Figures 4 and 5. a,b,c, Heat maps of normalized DNaseI-seq read 5’end counts (blue) anchored on each DHS midpoint and ranked by increasing DHS width together with the location of JAMM-called peak edges (black) for divergent (a), unidirectional (b), and bidirectional (c) promoter DHSs. **d,** Scatter plot of DHS width versus distance between forward and reverse 5’-GRO-seq cluster modes of divergent promoters. **e,** Boxplots of distance between 5’-GRO-seq cluster modes and corresponding DHS edges, dot = mean.

Figure S5 | Unidirectional promoters lack upstream hallmarks of divergent transcription, related to Figures 4 and 5. a,b, Positional average fragment-extended ChIP-seq read counts within Taf1 (a) and Tbp (b) peak summits as called by SISSRS in bins of 10 nucleotides (see Supplemental Experimental Procedures). **c,** Positional average of normalized read 5’end counts of traditional GRO-seq for the forward (red and blue) or reverse (orange and light blue) directions of the divergent (red and orange) or unidirectional (blue and light blue) promoters (“normalized counts” refers to 0-to-1 scaling of read counts for every DHS window, see

Supplemental Experimental Procedures). **d,** Distributions of whole HeLa cell, polyA-plus CAGE tag 5'end counts from ENCODE intersecting designated 5'GRO-seq clusters.

Figure S6 | Histone modifications HMM characteristics and analysis, related to Figure 6. a, Chromatin state – Genome Annotation enrichment map (see Supplemental Experimental Procedures). "Genes" are entire UCSC gene lengths, "TSS" are UCSC known gene transcription start sites, "TES" are UCSC known gene transcription end sites, "TSS 4kb" and "TES 4kb" are windows centered around UCSC TSSs and TESs respectively going 2kb upstream and downstream, "F5 Enhancers" are enhancers identified by the Fantom5 consortium for the hg19 genome build, "F5 Enhancers 2k" are windows centered around the midpoints of F5 Enhancers going 1kb downstream and 1kb upstream, "CGI" are UCSC "CpG" islands. **b,** Percentage of chromatin state intersections at 75 bp downstream of the NFR edges. "No State" refers to those locations that did not intersect any chromatin state. **c,d,e,f** Average fragment-extended read counts of H3K27ac (c), H3K4me3 (d), H3K4me2 (e), and H3K4me1(f) ChIP-seq in bins of 10 nucleotides for divergent (blue), unidirectional (red), bidirectional (green), and intergenic (light blue) 5'-GRO-seq-containing DHSs (see Experimental Procedures). grey = average DNaseI-seq read 5'end counts for DHSs from all four groups combined.

Table S1 | Comparison of 5'GRO-seq and exosome KD CAGE analyses, related to Figure 4. The same analyses were performed on both datasets using the same DHS peaks calls as described in the Supplemental Experimental Procedures. Margin numbers indicate the number of DHSs that were identified in each group from each dataset. Table numbers indicate the overlap between DHS classes between the two datasets. The most conservative estimate for percentage of unidirectional promoters is 34% (1196/3499) when only considering DHSs with forward gene evidence in both datasets, from which unidirectional DHSs are consistently classified in both datasets and divergent/bidirectional DHSs identified in at least one dataset. It is likely that many of the forward TSS-containing DHSs (unidirectional, divergent, or bidirectional) identified in only one of the two datasets are true; when these are included, we estimate that the true percentage of unidirectional promoters is closer to 44% (3394/7707).

Table S2 | Correlations between 5'GRO-seq and TSS prediction score or H3K27ac ChIP-seq, related to Figure 4.

Spearman Rho correlation values are shown with corresponding p values between 5'GRO-seq read 5'end counts within called clusters (top) and either the TSS prediction score (left top) or H3K27ac ChIP-seq fragment-extended read counts intersecting a window 148 bp downstream of the appropriate DHS peak edge (left bottom).

Table S3 | Final_Cluster Sets.xlsx, related to Figure 1.
5’GRO-seq cluster calls as identified using the strategy described in Ni *et al.* (Ni et al., 2010) and Supplemental Experimental Procedures.

Supplementary Tables

Table S1.

		5’-GRO-seq (n = 4378)		
		Divergent (1741)	Unidirectional (2237)	Bidirectional (400)
Exosome KD **CAGE (n = 6828)**	Divergent (2890)	1134	490	4
	Unidirectional (3188)	343	1196	1
	Bidirectional (750)	0	1	330

Table S2.

		5’GRO-seq Cluster Read Counts	
		Forward	Reverse
Sequence Model	Forward	0.22 (p < 0.0001)	-0.026 (p = 0.29)
	Reverse	-0.04 (p = 0.096)	0.16 (p < 0.0001)
H3K27ac	Forward	0.39 (p < 0.0001)	0.0001 (p = 0.09)
	Reverse	0.04 (p = 0.09)	0.25 (p < 0.0001)

Supplemental Experimental Procedures

Cell culture conditions

HeLa S3 cells were grown at 37°C in DMEM (Cellgro) supplemented with 10% FBS (Gibco), 50 U Penicillin and 50 µg Streptomycin per mL (Gibco).

In vitro transcription assays

Core promoter sequences, ±50 bp in respect to the +1 TSS, were cloned into pUC119 (F/F_R) or pUC118 (R) containing a Pol III specific terminator (Duttke, 2014) using XbaI and PstI. A spacer was further inserted into pUC118 to match the distance of the XbaI and PstI cloning sites to the reverse M13 primer site of pUC119. When indicated, the TATA-box was substituted with "ACGTCCGT" (mTATA).

Transcription reactions were carried out as described previously (Duttke, 2014). Briefly, 7 µL of 13 mg/mL human nuclear extract (HSK) were preincubated with 500 ng DNA template in a total volume of 46 µL with a final concentration of 20 mM HEPES-K^+ (pH 7.6); 50 mM KCl; 6 mM $MgCl_2$; 2.5% (w/v) polyvinyl glycol (compound); 0.5 mM DTT; 3 mM ATP; 0.02 mM EDTA and 2% glycerol at 30°C for 75 minutes. Transcription was started by addition of 4 µL NTPs (5 mM each), carried out for 20 minutes and stopped by addition of 100 µL STOP buffer [20 mM EDTA; 200 mM NaCl; 1% SDS, 0.3 mg/mL glycogen]. After mixing, 12.5 µg Proteinase K was added and reactions were incubated at room temperature (~21°C) for 15 minutes. Nucleic acids were subsequently extracted by standard phenol/chloroform purification followed by ethanol precipitation. Transcripts were subjected to primer extension analysis using 5′- ^{32}P-labeled M13 reverse sequencing primer [5'-AGCGGATAACAATTTCACACAGGA] and separated by urea-polyacrylamide gel electrophoresis. Gels were exposed to a phosphor imager plate and reverse transcription products visualized and quantified with a Typhoon imager (GE Health Sciences).

5'GRO-seq and GRO-seq library generation and sequencing

5'GRO-seq was performed as described previously (Lam et al., 2013). Briefly, about 10^7 HeLa S3 nuclei were used for run-on with BrU-labelled NTPs. Reactions were stopped by addition of 450 µL Trizol LS reagent (Invitrogen). After RNA extraction and treatment with Turbo DNase (Ambion), both according to the manufacturer's instructions, RNA was hydrolyzed by Zn^{2+} fragmentation (Ambion). The fragmented transcripts were incubated for 2 h at 37°C with polynucleotide kinase (PNK, NEB) at pH 5.5 to remove 3' phosphates. BrU-labelled nascent transcripts were subsequently immunoprecipitated with anti-BrdU agarose beads (Santa Cruz Biotech). For 5'GRO-seq, immunoprecipitated RNA was dephosphorylated with calf intestinal phosphatase (NEB). Then 5′ capped fragments were de-capped with tobacco acid pyrophosphatase (Epicentre). Illumina TruSeq adapters were ligated to the RNA 3′ and 5′ ends with truncated mutant RNA ligase 2 (K227Q) and RNA ligase 1 (NEB), respectively. Reverse transcription was performed with Superscript III (Invitrogen) followed by PCR amplification for 12 cycles. Final libraries were size selected on PAGE/TBE gels to 175–225 bp.

GRO-seq was essentially performed as 5'GRO-seq but the immunoprecipitated RNA was directly de-capped with tobacco acid pyrophosphatase (Epicentre) and subsequently kinased with PNK (NEB) prior to adapter ligation.

5'-GRO-seq and GRO-seq read processing, cluster calls, and annotation

Two replicates of 5'end sequenced reads from the 5'-GRO-seq or traditional GRO-seq protocols were trimmed for adapters using cutadapt (Martin, 2011), mapped together to the hg19 human genome using Bowtie2 with default settings(Langmead and Salzberg, 2012). Reads that did not map uniquely and reads overlapping rRNA loci were removed, yielding 27,512,149 5’-GRO-seq reads and 21,765,842 traditional GRO-seq reads. Clusters were identified according to the strategy described in Ni *et al.* (Ni et al., 2010). Briefly, a kernel density estimate (KDE) of the 5’ end positions of the mapped reads was calculated across the genome. Any region exceeding the genome-wide average KDE that contained at least 10 reads was identified as a cluster and used in subsequent analysis. To annotate the identified clusters, the Genomic Features(Lawrence et al., 2013) R package was used to generate BED files for 5’utr, 3’utr, intron, coding exon, non-coding exon, and promoter (-250 upstream of annotated transcription start sites) regions according to the UCSC knownGenes table. BEDTools (Quinlan and Hall, 2010) intersect was used to perform a prioritized intersection between the 5’-GRO-seq cluster calls and these annotation bed files with the following priorities: transcription start site (TSS) > coding exon > 3’utr > non-coding exon > intron. Clusters intersecting either promoter or 5’utr locations were considered TSS-annotating clusters. Clusters not intersecting any of these locations were considered intergenic-annotating clusters. This strategy resulted in exactly one annotation per 5’GRO-seq cluster. Following downstream analyses of cluster pair calling (either closest upstream or DNase-seq based; described below), regions containing clusters annotated as TSS but that overlapped annotated tRNA loci were removed from subsequent analysis.

DNase-seq and ChIP-seq read processing and peak calling

All 5 datasets of ENCODE-mapped DNase-seq reads for HeLa-S3 cells were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012). PCR duplicates from each file were removed using SAMTools (Li et al., 2009). The resulting files were converted to BED using BEDTools (Quinlan and Hall, 2010) and concatenated before peak calling with JAMM v1.0.6

(Ibrahim et al., 2014) (http://code.google.com/p/jamm-peak-finder/, settings: -m narrow -f 1). HeLa-S3 cell, Broad Institute histone modification ChIP-seq raw fastq files were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012). Reads were aligned to hg19 genome using Bowtie2 (Langmead and Salzberg, 2012) with default parameters and then filtered for those that did not align uniquely or had more than two mismatches. PCR duplicates were removed after alignment using SAMTools (Li et al., 2009) and converted to standard BED format using BEDTools (Quinlan and Hall, 2010). Histone modification peaks were called using JAMM v1.0.4rev1 (Ibrahim et al., 2014) with default settings while maintaining all replicates separate. The filtered peak lists produced by JAMM were considered for further analysis. Raw ENCODE HeLa-S3 ChIP-Seq fastq files for TAF1 and TBP (Bernstein et al., 2012) were processed in the same way as ENCODE histone modification datasets. Replicate BED files were then concatenated before peaks were called using SISSRS (Narlikar and Jothi, 2012) , which can resolve ChIP-Seq peak summits at high resolution (settings: -s 3095693983).

CAGE read processing

Fastq files from Ntini *et al.*(Ntini et al., 2013) (SRR922110.sra and SRR922111.sra) were obtained from the Gene Expresssion Omnibus (GEO) website. Reads were trimmed according to authors methods (Ntini et al., 2013) using Flexbar (Dodt et al., 2012) and mapped to the hg19 human genome using Bowtie2 with default settings (Langmead and Salzberg, 2012). Reads that did not map uniquely were removed. Mapped .bam files for Hela whole-cell, polyA-plus CAGE were downloaded from the UCSC ENCODE ftp server (Bernstein et al., 2012). CAGE reads were corrected for the 5'end nucleotide bias using the CAGEr R package (http://bioconductor.org/packages/release/bioc/html/CAGEr.html).

Closest upstream antisense pair assignments

In order to define a set of 5’-GRO-seq cluster pairs that were reciprocally the closest upstream antisense of each other, a combination of BEDTools and custom scripts was used. BEDTools *closest* command (Quinlan and Hall, 2010) (settings: -S -id -D “a”) was run on the modes of 5’-GRO-seq clusters (the position with the highest read count within a cluster) using the same file for both inputs. Custom Perl scripts were then used to parse the BEDTools output for only those cluster pairs where both modes were called as closest upstream antisense of each other.

DHS-based divergent and unidirectional promoter definitions

In order to define promoter DNase-I HyperSensitive regions (DHSs) as divergent or unidirectional, BEDTools (Quinlan and Hall, 2010) *intersect* command was used to find overlaps between DNaseI-seq peak calls (defining DHSs) and 5’-GRO-seq cluster modes, both described above. The output from BEDTools was then parsed with custom Perl scripts into different DHS categories. DHSs with exactly one intersecting TSS cluster mode were considered unidirectional. DHS with exactly two intersecting 5’-GRO-seq cluster modes where the two modes were upstream and antisense of each other, one annotating as TSS and the other as intergenic, were considered divergent. DHSs with more than one intersecting 5’-GRO-seq cluster modes on any one DNA strand, or with two 5’-GRO-seq cluster modes on opposite strands but downstream of each other, were removed from further analysis. For an increased-confidence unidirectional group, unidirectional classified DHSs intersecting reverse-side annotated TSSs (yet having no 5’GRO-seq clusters) or containing exactly one TSS-annotating cluster mode that was also part of the divergent or bidirectional reciprocal closest upstream antisense selection (described above) were considered ambiguous and removed from further analysis..

Heat map and meta-analysis plots

5’-GRO-seq and DNaseI-seq heat maps were made by calculating the center point between the 5’-GRO-seq cluster modes of the paired forward/reverse TSS clusters or the center point of DHSs. Windows were then taken around these center points and strand assignments (important for plotting orientation) made according to the forward, annotated, gene for divergent cluster pairs or unidirectional promoter DHS. For TSS-TSS or intergenic-intergenic 5’-GRO-seq cluster pairs, the cluster with higher read counts was used for strand assignment since there is no clear definition for sense/antisense in these situations. Genomic coordinates were then grouped in bins of 10 and the number of reads whose 5’end mapped to each bin were counted independent of strand and scaled so that the minimum value for each window is 0 and the maximum value is 1. Windows were sorted according to the distance between cluster pairs or the width of the DNaseI-seq peaks and plotted using the ggplot2 R package (Wickham, 2009).

For sequence heat maps, center positions, windows, strand and ranking were determined as above. BEDTools *getfasta* command (Quinlan and Hall, 2010) was used to retrieve the sequence corresponding to each window and ggplot2 (Wickham, 2009) was used for the plotting.

For TAF1 and Tbp ChIP-seq meta-analysis plots, sequence reads were extended by the fragment length calculated by SISSRS (Narlikar and Jothi, 2012). Center points, windows, and strands were determined as described above. For each window, genomic positions were grouped in bins of 10. If a bin overlapped a SISSRS summit (Narlikar and Jothi, 2012) (see above), then the number of extended-reads covering that bin were counted. If no peak summits overlapped a bin, it was assigned a 0. The per-bin means across all promoter locations were plotted using the ggplot2 R package (Wickham, 2009).

For GRO-seq metaplots center points, windows, and strands were determined as described above. For each window, genomic positions were grouped in bins of 10 and the number of sequence tag 5'ends counted per bin in a strand sensitive manner. The two resulting vectors of binned counts (one for each strand per window) were scaled together so that the minimum value for each window is 0 and the maximum value is 1. The per-bin means of these strand-sensitive, scaled, vectors across all promoter locations were plotted using the ggplot2 R package (Wickham, 2009).

The number of ENCODE CAGE sequence tag 5'ends were counted that intersected each 5'-GRO-seq cluster and the distribution of such counts per group were plotted as boxplots using the ggplot2 R package (Wickham, 2009).

Histone modification metaplot center points, windows, and strands were determined as described above. Reads were extended by the fragment sizes calculated within JAMM (Ibrahim et al., 2014). Genomic coordinates were then grouped in bins of 10 and the number of extended reads per million mapped overlapping each bin were counted independent of strand. The per-bin means across all promoter locations were plotted using the ggplot2 R package (Wickham, 2009).

TSS initiation pattern analysis

NarrowPeak, BroadPeak, and WeakPeak initiation patterns as defined previously(Ni et al., 2010) were determined for the specified groups from the 5'-GRO-seq clusters with at least 25 read counts.

Position-specific threemer counts

Position-specific threemer counts were determined with custom Perl scripts. After counting the instances of each threemer at each position, this value was divided by the total occurrence of that threemer in that sequence group. These values were plotted using the ggplot2 R package (Wickham, 2009).

Probabilistic model of transcription start sites

In order to compare the sequence composition of reverse direction core promoters and to scan DHS regions for transcription start site sequences, we employed a previously published position-specific Markov chain model (Frith et al., 2008) (PSMM). We used the first-order setting which will calculate the probability of a given di-nucleotide at a given position relative to that position's mono-nucleotide frequency, normalized for the di- and mono-nucleotide frequencies in the training set independent of position. Since the program reports log2 scores, all the values in our plots are 2^S, S being the log2 score output by the program.

A 10-fold cross validation scheme of the PSMM was implemented as follows. To train the model, the list of forward, TSS annotating, core promoters from either the closest upstream antisense selection, or DHS-based selection strategies, were split into 10 equal-size, non-overlapping, groups. These were designated as 10 unique "test" sets. For each test set, a corresponding training set was composed of the regions in the complete set that did not overlap the test set. The PSMM was then trained 10 separate times, once for each training set, on sequences +/- 50 bp around the TSS cluster modes. Each of the 10 models was then run on its corresponding, non-overlapping, test set. For the closest upstream antisense selection strategy, the test sequences were +/- 50 around the modes of the 5'-GRO-seq clusters. For the DHS-based selection strategy, the test sequences were -150 to +50 around the appropriate DHS edge corresponding to the 5’-GRO-seq clusters of that test set.

In addition to the test group subsets, the 10 models were each run on the complete set of other sequences in question. For the closest upstream antisense selection strategy, these other sequences were +/- 50 around the 5’-GRO-seq cluster modes. Means were calculated for each position across the promoters of each list, resulting in 10 vectors of position means, one for each trained model. The mean at each position across these 10 vectors was plotted using ggplot2 (Wickham, 2009). For the DHS-based selection strategy, the sequences were -150 to + 50 around the appropriate DHS edge. Negative scores where the background model was higher than the TSS model were set to zero. The sequence positions were grouped in bins of 10 and the average score from each bin was calculated, then the mean average score was calculated for each binned position across all promoters of the list, resulting in 10 vectors of average score means at each position, one for each trained model. Shuffled control sequences were generated using the shuffleseq algorithm with default settings from the EMBOSS suite (Rice et al., 2000). The mean at each binned position across these 10 vectors was smoothed and plotted using ggplot2 (Wickham, 2009). For divergent pair scores in Supplemental Figure 3d, the scores for each 5-GRO-seq cluster mode were combined for each of the 10 cross validation runs and plotted as boxplots using ggplot2 (Wickham, 2009).

Receiver operator characteristic and precision recall curves were generated by defining true positives as the modes of 5’-GRO-seq clusters and true negatives as every other nucleotide in the tested windows, the results plotted for each of the 10 models from the closest upstream antisense selection using the R package ROCR (Sing et al., 2005).

Motif scanning

The TRANSFAC TATA-box binding protein or JASPAR Initiator position weight matrices (M00252; pwm) were used with the Scanner Toolset (Megraw et al., 2009) to scan sequences -35 to -25 upstream for TATA and +/- 5 for initiator around the forward TSS modes of the divergent and unidirectional promoter groups. A fixed first order Markov background was used for each list calculated from sequences +/- 50 around the forward TSS modes. Thresholds for fixed background scans were determined with a false positive rate cutoff of 0.001 as described in Megraw *et al.* (Megraw et al., 2009). For score distributions, highest scores were taken when locations contained multiple hits in the region scanned.

CpG island (CGI) analysis

Genomic coordinates of CGI were taken from the UCSC table browser (Kuhn et al., 2013). Either divergent or unidirectional DHSs were intersected with these coordinates using BEDTools intersect (Quinlan and Hall, 2010), either with the –u setting for counting the number of DHSs that intersect a CGI or the –wa –wb setting for determining size distributions of CGIs that intersect DHSs.

Chromatin State Segmentation

Similar to previous approaches (Ernst and Kellis, 2012; Hoffman et al., 2012), we employed a Hidden Markov Model (Taramasco and Bauer, 2013) (HMM) for unsupervised genome-wide clustering of histone modification ChIP-Seq read counts. We chose a multivariate Gaussian distribution for the HMM state emissions. Each chromatin state is a multivariate Gaussian distribution fully defined by its means vector, corresponding to the signals' means of the histone modification tracks, and its co-variance matrix.

In a pre-processing step, we define relevant locations for each histone modification (positions intersecting a ChIP-Seq peak) separately across the whole genome at 10-basepair resolution. The signal at relevant locations is defined as background-normalized, smoothed, extended-read counts (ie. ChIP-Seq signal). Peaks were identified using JAMM (Ibrahim et al., 2014), as described above. For each histone modification dataset, we extracted the corresponding ChIP-Seq signal for each peak at single-basepair resolution, using the SignalGenerator pipeline provided with JAMM (Ibrahim et al., 2014). JAMM's SignalGenerator output is then aligned to the genome in 10-basepair bins using the BEDOps (Neph et al., 2012) *bedmap* command (settings: --mean). Bins that did not intersect ChIP-Seq peaks are assigned a signal of zero. ChIP-Seq signal for each histone modification track is then scaled so that the minimum value is zero and the maximum value is 1000 and converted to log-space.

The resulting 10-basepair binned signal tracks for all histone modifications are matched up and bins that have a zero ChIP-Seq signal in all tracks are discarded. Bins that have a zero ChIP-Seq signal in one or more histone modification track(s) but not the other(s) are assigned a simulated normally-distributed background signal with a mean equal to the lowest bin signal value in the corresponding histone modification track and a variance of 0.1.

To learn the emission and transition parameters of the HMM, we employ the Baum-Welch algorithm (Bilmes, 1997; Taramasco and Bauer, 2013), initialized via k-means, on the signal tracks of chromosome 1. This learning process results in distinct chromatin states, each represented as a multivariate Gaussian distribution. The mean vector for each state defines the average ChIP-Seq signals of the histone modification tracks in the corresponding state. We 0-to-1 scale the means across each histone modification to define the prototypical chromatin states shown in Fig. 6a.

Finally, we employ the Viterbi decoding algorithm (Taramasco and Bauer, 2013; Viterbi, 1967) to assign a chromatin state to each 10-basepair bin in the genome that had a peak in at least one of the histone modification tracks. Locations that did not have a peak in any histone modification track (no relevant features, zero signal in all tracks) are not assigned a state. Book-ended bins that have the same state are merged. The output of this process is genome segmentation into variable-width non-overlapping chromatin states similar to Segway (Hoffman et al., 2012) and ChromHMM (Ernst and Kellis, 2012).

The main advantage of our chromatin state genome segmentation pipeline is that it allows for chromatin state assignment at high-resolution using "semi-binarized" signal, as opposed to using fully binarized (enriched / not-enriched) information at 200 bp resolution utilized in the ChromHMM approach (Ernst and Kellis, 2012). Our semi-binarized signal is the smoothed-extended ChIP-Seq read counts for relevant locations in the genome (ChIP-Seq peaks) and zeros elsewhere. This allows to account for information about the co-variance of the histone modifications' signals, but without suffering from noise over-representation, and thus has the potential to lead to more meaningful clustering of the histone modification signals compared to previous approaches(Ernst and Kellis, 2012; Hoffman et al., 2012). Finally, we do not analyze the entire genome, but only locations which had ChIP-Seq peaks in at least one histone modification dataset. Therefore, we can assign chromatin states at high-resolution 10 bp bins, close to the single-basepair resolution of Segway (Hoffman et al., 2012) but without its expensive computational resources requirement. Segway (Hoffman et al., 2012) can only run on high-performance computing clusters whereas our pipeline runs on typical desktop machines.

Chromatin State Analysis

To produce chromatin state coverage plots, we started with windows defined around the midpoints of DHSs as described above. Chromatin states were intersected with DHS-based windows using BEDTools (Quinlan and Hall, 2010) *intersect* command.

Chromatin state enrichment for different categories of 5'GRO cluster annotations were based on intersection of chromatin states with single-nucleotide locations that are 75-basepair downstream of the corresponding DHS edge.

Chromatin state enrichment with different genome-wide annotations were done using ChromHMM (Ernst and Kellis, 2012) *overlapEnrichment* command (settings: -b 10) using annotations based on hg19 UCSC knownGenes table (Kuhn et al., 2013) and hg19 Fantom5 enhancer list (Andersson et al., 2014).

Supplemental References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bilmes, J. (1997). A gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models. Tech. Rep., International Computer Science Insitute *ICSI-TR 97*.

Dodt, M., Roehr, J.T., Ahmed, R., and Dieterich, C. (2012). FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology *1*, 895–905.

Duttke, S.H.C. (2014). RNA Polymerase III Accurately Initiates Transcription from RNA Polymerase II Promoters in Vitro. The Journal of Biological Chemistry *289*, 20396–20404.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nature Methods *9*, 215–216.

Frith, M.C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., and Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. Genome Research *18*, 1–12.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature Methods *9*, 473–476.

Ibrahim, M.M., Lacadie, S.A., and Ohler, U. (2014). JAMM: A Peak Finder for Joint Analysis of NGS Replicates. Bioinformatics (Oxford, England). doi: 10.1093/bioinformatics/btu568

Kuhn, R.M., Haussler, D., and Kent, W.J. (2013). The UCSC genome browser and associated tools. Briefings in Bioinformatics *14*, 144–161.

Lam, M.T.Y., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M., et al. (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. Nature *498*, 511–515.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Computational Biology *9*, e1003118.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078–2079.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. Journal. http://journal.embnet.org/index.php/embnetjournal/article/view/200.

Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., and Hatzigeorgiou, A.G. (2009). A transcription factor affinity-based code for mammalian transcription initiation. Genome Research *19*, 644–656.

Narlikar, L., and Jothi, R. (2012). ChIP-Seq data analysis: identification of protein-DNA binding sites with SISSRs peak-finder. Methods in Molecular Biology (Clifton, N.J.) *802*, 305–322.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics (Oxford, England) *28*, 1919–1920.

Ni, T., Corcoran, D.L., Rach, E. a, Song, S., Spana, E.P., Gao, Y., Ohler, U., and Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nature Methods *7*, 521–527.

Ntini, E., Järvelin, A.I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P.R., et al. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. Nature Structural & Molecular Biology *20*, 923–928

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841–842.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics : TIG *16*, 276–277.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics (Oxford, England) *21*, 3940–3941.

Taramasco, O., and Bauer, S. (2013). RHmm: Hidden Markov Models simulations and estimations. R package version 2.0.3 https://r-forge.r-project.org/R/?group_id=85.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory *13*, 260–269.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer Science & Business Media).