Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ

Tong Wu, Ruitu Lyu, Qiancheng You and Chuan He

Transcription is a highly dynamic process that generates single-stranded DNA (ssDNA) in the genome as ‘transcription bubbles’. Here we describe a kethoxal-assisted single-stranded DNA sequencing (KAS-seq) approach, based on the fast and specific reaction between N₃-kethoxal and guanines in ssDNA. KAS-seq allows rapid (within 5 min), sensitive and genome-wide capture and mapping of ssDNA produced by transcriptionally active RNA polymerases or other processes in situ using as few as 1,000 cells. KAS-seq enables definition of a group of enhancers that are single-stranded and enrich unique sequence motifs. These enhancers are associated with specific transcription-factor binding and exhibit more enhancer-promoter interactions than typical enhancers do. Under conditions that inhibit protein condensation, KAS-seq uncovers a rapid release of RNA polymerase II (Pol II) from a group of promoters. KAS-seq thus facilitates fast and accurate analysis of transcription dynamics and enhancer activities simultaneously in both low-input and high-throughput manner.

Transcription and its regulation determine cell fate and physiological functions. Dysfunctions in transcriptional regulation are often associated with human diseases. To understand global transcription regulation, genome-wide sequencing approaches have been developed to analyze the occupancy of RNA polymerases (ChIP-seq), or detect the presence and level of nascent RNA. Nascent RNA analysis is usually based on run-on assays, metabolic enrichment and Pol II ChIP-seq. Although powerful, these methods have limitations. Run-on-based methods and Pol II-associated RNA enrichment typically require millions of cells as starting material. Pol II ChIP-seq could not distinguish whether RNA polymerases are simply bound or are actively engaged in transcription. Metabolic labeling may not be able to accurately measure transient and low-abundant RNA species, such as enhancer RNAs (eRNAs), especially when using limited materials with modest sequencing depth. As most RNAs undergo post-transcriptional processing, their levels are indirect readouts that may not accurately reflect transcription dynamics in situ.

Transcriptionally engaged RNA polymerases resolve DNA double helices and generate single-stranded DNA bubbles. Therefore, we envision that mapping ssDNA throughout the genome provides a readout of the activity and dynamics of transcriptionally engaged RNA polymerases. Permanganate was previously reported to preferentially oxidize single-stranded thymidine residues and was subsequently used to reveal Pol II-induced promoter melting in both loci-specific and genome-wide manners. The combination of permanganate treatment and S1 nuclease digestion allows genome-wide identification of non-B-form DNA structures. However, this method requires tens of millions of cells, and shows low sensitivity when detecting relatively weak and broad signals derived from Pol II elongation at gene bodies.

Here we describe a rapid and sensitive labeling of ssDNA for sequencing (KAS-seq), based on a kethoxal–guanine reaction. We show that KAS-seq simultaneously measures the dynamics of transcriptionally engaged Pol II, transcribing enhancers, Pol I and Pol III activities and non-canonical DNA structures involving ssDNA in situ, by using as few as 1,000 cells or mice tissues. Furthermore, we demonstrate that KAS-seq detects transcription dynamics during transient physiological environment changes, such as inhibition of protein condensation.

Results

Genome-wide profiling of single-stranded DNA using N₃-kethoxal-based labeling. Kethoxal (1,1-dihydroxy-3-ethoxy-2-butane) was previously reported to react with the N1 and N2 positions of guanines in single-stranded DNAs and RNAs under physiological conditions. We recently developed a synthesis for an azide-tagged kethoxal (N₃-kethoxal), which not only preserves its high reactivity and specificity to guanines in single-stranded nucleic acids, but also offers a bio-orthogonal handle that can be readily modified with a biotin or other functional groups. We showed that this reagent provides an effective way to map RNA secondary structures by labeling guanines in single-stranded RNAs under mild conditions in live cells. On the basis of this initial success, we reasoned that N₃-kethoxal should also enable specific ssDNA labeling and profiling, because the formation of Watson–Crick base-pairing in double-stranded DNA blocks the labeling reaction.

We first performed an in vitro labeling assay using a synthetic DNA oligonucleotide probe containing four deoxyguanosine bases. After incubating the oligonucleotide with N₃-kethoxal at 37 °C for 5 min, all four deoxyguanosine bases on all oligonucleotide molecules were labeled (Extended Data Fig. 1a), suggesting a high labeling reactivity of N₃-kethoxal on ssDNA in vitro. While N₃-kethoxal reacts with deoxyguanosine bases under neutral conditions within 2 min, very few 1-arginine could be labeled within 10 min under the same conditions (Extended Data Fig. 1b), indicating that protein labeling could be minimized under the labeling conditions of KAS-seq. After labeling live cells with N₃-kethoxal, genomic DNA
KAS-seq signals mark active transcription. We performed KAS-seq starting from one million live HEK293T cells and mouse embryonic stem cells (mESCs). N3-kethoxal labeling does not affect gDNA isolation yield and purity (Extended Data Fig. 1c,d). KAS-seq performed in the absence of N3-kethoxal or the biotinylation reagent (biotin-DBCO) resulted in negligible biotin signals shown by dot blot (Extended Data Fig. 1e), nor sufficiently enriched DNA for library construction (Extended Data Fig. 1f), suggesting minimum background of KAS-seq.

KAS-seq is robust and reproducible, showing high enrichment efficiency along with high correlation (r=0.99) and peak overlap between replicates (Extended Data Fig. 2a–c). KAS-seq signals exhibit a similar distribution pattern as Pol II ChIP-seq signals along regions with different G/C contents (Extended Data Fig. 2d), suggesting that the G-specific labeling does not notably induce bias, although the G/C content effect should be considered for more specified applications of KAS-seq. KAS-seq reads are considerably enriched at gene-coding regions, especially at gene promoters and transcription termination areas, while depleted at intergenic regions (Fig. 2a). KAS-seq profiles on gene-coding regions revealed a strong and sharp peak around the transcription start site (TSS), relatively weak and broad signals that cover the entire gene body and a strong but broad peak starting from transcription end site (TES) to its downstream regions (Fig. 2b and Extended Data Fig. 2e). KAS-seq signals show positive correlations with histone modifications that mark active transcription, such as H3K4me3, H3K27ac and H3K36me3, and are negatively correlated with inactive chromatin markers such as H3K27me3 and H3K9me3 (Fig. 2c). Notably, KAS-seq signals correlate better with H3K36me3 than ATAC-seq results do, indicating that while ATAC-seq serves as a powerful tool to probe chromatin accessibility18, KAS-seq directly measures transcription activities. KAS-seq signals at TSS overlap with H3K4me3 and H3K27ac, and those at gene body overlap with H3K36me3 (Fig. 2d and Extended Data Fig. 2f). These results collectively suggest that KAS-seq signals are derived from the Pol-II-mediated transcription. We also compared KAS-seq with permanganate/S1 footprinting. Both methods show similar sensitivity on detecting the strong ‘promoter melting’ signals, but KAS-seq is much more sensitive on detecting the weak and broad ssDNA signals on the gene bodies and terminal regions (Extended Data Fig. 2g).

Because of the high guanine labeling reactivity of N3-kethoxal and the high affinity between biotin and streptavidin, KAS-seq is expected to maintain its sensitivity when using low-input starting materials or primary tissue samples. Indeed, the distribution of KAS-seq signals at gene-coding regions and the overlap with histone modifications remain unchanged when using 10,000, 5,000 or even 1,000 HEK293T cells (Fig. 2d and Extended Data Fig. 3a,b). KAS-seq results with low-input cells showed similar enrichment efficiency and captured similar numbers of peaks compared with KAS-seq libraries generated from 1 million cells (Extended Data Fig. 3c,d). Using mouse liver tissue, KAS-seq also shows strong signals at TSS, with weakened signals on the gene bodies and at TES regions (Extended Data Fig. 3e). Thus, KAS-seq is suitable for studying rare cell samples and clinical samples in the future.

KAS-seq reveals the dynamics of transcriptionally engaged Pol II. We next compared KAS-seq with GRO-seq and Pol II ChIP-seq in HEK293T cells. KAS-seq results correlate well with results from
these assays (Fig. 3a). In mESCs, ~95% of KAS-seq peaks on promoters overlap with Pol II ChIP-seq peaks (Extended Data Fig. 4a). Reads density of KAS-seq and Pol II ChIP-seq on the gene bodies show a strong positive correlation (Pearson’s r = 0.81, Extended Data Fig. 4b). We then ranked all genes into four groups according to their expression levels based on RNA-seq data (Extended Data Fig. 4c), and showed that the strength of KAS-seq signals drop notably in genes with low expression levels (Fig. 3b).

To further validate that transcriptionally engaged Pol II is the primary source of detected ssDNA signals, we treated HEK293T cells with 5,6-dichlorobenzimidazole 1-β-d-ribofuranoside (DRB) and triptolide, respectively, before performing KAS-seq. DRB inhibits Pol II release from pausing at TSS, and triptolide inhibits recruitment and loading of Pol II to promoters15. While the majority of peaks overlap with those at the native state, after DRB and triptolide treatment, KAS-seq peak numbers decreased by 57% and 93%, respectively (Fig. 3c). As expected, DRB severely diminished ssDNA signals at gene body and transcription termination regions with increased signals at TSS; triptolide almost completely erased all signals at the entire gene-coding regions (Fig. 3d and Extended Data Fig. 4d,e). These observations confirm that the strong and sharp KAS-seq peaks on gene promoters reflect transcription initiation and pausing of Pol II near the TSS,13, and that KAS-seq signals at gene bodies are derived from transcription elongation.

Comparing KAS-seq signals at promoter-proximal and gene-body regions enabled us to sort genes into four classes with distinct transcription states: class I, paused and active; class II, paused and inactive; class III, not paused and active; and class IV, not paused and inactive (Fig. 3e and Extended Data Fig. 4f). In HEK293T cells, 60% (11,715 out of 19,279) also showed active Pol II elongation, which is consistent with results obtained previously from GRO-seq.14

Apart from signals on promoters and gene bodies, we also found KAS-seq signals considerably enriched at transcription termination regions (Fig. 2a,b). These signals were removed by DRB treatment.
Fig. 3 | KAS-seq reveals Pol II dynamics and defines gene transcription states in HEK293T cells. a, Genome-wide Pearson correlation heat map between KAS-seq, Pol II ChIP-seq, GRO-seq and nascent RNA-seq (4SU-seq) reads density on gene-coding regions. Pairwise correlation coefficients are noted in each square (n = 839,684 1-kb bins in the hg19 genome). b, KAS-seq reads density at gene-coding regions of genes with different expression levels (defined by RNA-seq). c, Venn diagram showing overlap of KAS-seq peaks in HEK293T cells under native, DRB treatment and triptolide treatment conditions. The number of common peaks between two replicates was used in each case. d, Heatmap showing KAS-seq signal distribution at gene-coding regions under native, DRB treatment and triptolide treatment conditions. Regions 3 kb upstream of TSS and 3 kb downstream of TES were shown. e, Defining four groups of genes with different transcription states on the basis of KAS-seq results. In each group, one gene is shown as an example by using the snapshot of KAS-seq signals under native and DRB-treated conditions.

(Extended Data Fig. 5a), indicating that they are derived from Pol II elongation (and pausing) at the termination window. We sorted all genes with KAS-seq signals at this region into three groups according to the length of their termination signals (Extended Data Fig. 5b). We then analyzed the averaged KAS-seq reads density on the entire terminal region of these three groups, without observing notable differences (Extended Data Fig. 5c), suggesting that KAS-seq does not exhibit length-dependent bias. We calculated the ‘termination index’ as the ratio of reads density at TES-downstream regions relative to the density in the promoter-proximal regions.
KAS-seq detects Pol-I- and Pol-III-mediated transcription events and non-B form ssDNA structures in the same assay. RNA polymerase I (Pol I) transcribes 5.8S, 18S and 28S ribosomal RNAs (rRNAs); RNA polymerase III (Pol III) synthesizes 5S rRNAs, transfer RNAs (tRNAs) and certain small RNAs\(^{21,22}\). As expected, apart from detecting Pol II activities, KAS-seq simultaneously detects transcription events mediated by Pol I and Pol III, which do not respond to DRB and triptolide (Extended Data Fig. 6a–c). Note that only a portion of tRNAs are actively transcribed (411/606) (Extended Data Fig. 6b), which may suggest a transcription level regulation of codon usage. KAS-seq can thus monitor the transcription activity dynamics of all RNA polymerases in one assay.

We also noticed many KAS-seq peaks that are not derived from Pol I or Pol III-mediated transcription under triptolide-treatment conditions; these peaks could be derived from other DNA forms and telomeric DNAs. We followed a previous method\(^{22}\) to predict potential genomic locations of different non-B form DNA species, including cruciform, quadruplex, H-DNA, Z-DNA, and hairpin structures. Under triptolide-treatment conditions, we found a number of KAS-seq signals overlapping with and enriching these non-B DNA and telomere regions (Extended Data Fig. 6d–f), suggesting potential applications of KAS-seq to study other ssDNA-involved biological processes.

Many enhancer regions are single-stranded, which correlate with higher enhancer activity. Pol II is known to bind at certain enhancers and generate enhancer RNAs bidirectionally\(^1\). KAS-seq can, therefore, be used to identify enhancers that are being transcribed by Pol II. We defined enhancers with KAS-seq peaks as ssDNA-containing enhancers (SSEs). We used the KAS-seq data under DRB-treatment conditions to annotate SSEs, because some enhancers are located at gene bodies that can show ssDNA signals derived from transcription elongation. Only around 25% of all annotated enhancers were defined as SSEs in mESCs, with the majority of enhancers showing no KAS-seq signal (Fig. 4a,b). Note that the cutoff we used for peak-calling filters off some weak KAS-seq signals, which may appear in the defined double-stranded enhancers.

ssDNA-containing enhancers include two subtypes, with one type showing KAS-seq signals spanning the entire enhancer, and the other showing KAS-seq signals more localized when compared with H3K27ac signals (Fig. 4b). KAS-seq signals at ssDNA-containing enhancers tend to increase upon DRB treatment (Fig. 4c), supporting the presence of enhancer transcription pausing and elongation\(^{19}\). ssDNA-containing enhancers include 94% of super-enhancers\(^{36}\), suggesting that most of the super-enhancers are actively transcribed (Fig. 4d). Genes associated with SSEs show higher expression levels (Fig. 4c), and these enhancers possess much more long-range interactions mediated by both CTCF and Pol II (Fig. 4f), indicating that these transcribing enhancers may possess a stronger capability to activate their target genes.

ssDNA-containing enhancers appear to enrich unique sequence motifs (Fig. 4g), suggesting that they have distinct sequence features and their potential binding by specific transcription factors (TFs). To compare SSEs with enhancers that simply possess high TF-binding signals, we sorted all ATAC-seq-positive enhancers into two groups according to whether they possessed KAS-seq signals or not. We found 50% of ATAC-seq-positive enhancers showed no (or very weak) KAS-seq signals in mESCs (Extended Data Fig. 7a). The integrated intensities of ATAC-seq signals on these two groups were similar (Extended Data Fig. 7b), but genes associated with the KAS-seq-positive group showed a higher expression level (Extended Data Fig. 7c). Sequence motifs enriched in ATAC-seq-positive but KAS-seq-negative enhancers were different than those in SSEs (Extended Data Fig. 7d).

We then examined the occupancy of Pol II, histone modifications and other transcription regulatory proteins on the ssDNA-containing enhancers. Consistent with them being transcribed, the occupancies of Pol II, H3K4me3, H3K27ac, Med1, Cdk8 and Cdk9 on these enhancers are considerably higher than those on double-stranded enhancers (Fig. 4h). Moreover, while the binding of Oct4, Nanog and Sox2 showed no notable difference in SSEs compared with double-stranded ones (Extended Data Fig. 7e), Brd4 is considerably enriched in SSEs (Fig. 4h). Brd4 was previously observed to regulate the expression of pluripotency factors such as Pou5f1 (Oct4) and Nanog in mESCs and mouse embryos\(^{26,27}\), indicating potential roles of these transcribing enhancers on regulating mESC differentiation. Moreover, gene ontology\(^\ast\) analysis revealed critical biological processes enriched in genes regulated by ssDNA-containing enhancers, including regulation of stem-cell population maintenance, differentiation and embryo implantation (Fig. 4i).

In HEK293T cells, though the ratio of SSEs to total enhancers is lower than that in mESCs (Extended Data Fig. 8a), SSEs maintain their high overlap with super-enhancers, response to DRB treatment, and high correlations with Pol II, H3K4me3 and H3K27ac signals (Extended Data Fig. 8b–c). Chromatin regulatory factors such as CTCF and YY1, as well as transcription factors such as SP1, SP2, MAZ, NCAPH2, KLF8 and KLF9, showed high occupancy on these ssDNA-containing enhancers (Extended Data Fig. 9a), with their binding motifs enriched at these regions (Extended Data Fig. 9b). Several other zinc-finger-domain-containing TFs were also shown enriched on these SSEs (Extended Data Fig. 9a). Messenger RNA processing, translation regulation and several other essential pathways are enriched in genes regulated by these enhancers (Extended Data Fig. 9c). Enriched TFs and gene sets in HEK293T cells are different from those in mESCs, suggesting potential regulatory functions by these transcribing enhancers in cell-type-specific manners.

Collectively, KAS-seq is able to detect SSEs as transcribing enhancers, which appear to possess distinct genomic features and unique TF-binding footprints. Consistent with previous observations\(^{26,27}\), these enhancers are associated with higher enhancer activity and can be cell-type specific.

ssDNA dynamics upon the inhibition of protein condensates. Considering the fast reaction kinetics between N3-kethoxal and ssDNA as well as the high sensitivity of KAS-seq, we speculated that KAS-seq can detect transcription dynamics in transient events. Protein condensates are highly dynamic structures formed through interactions between mediators, TFs and other transcription coactivators, and were shown to incorporate Pol II to activate transcription\(^{14,15}\). 1,6-hexanediol is widely used to dissociate these condensates in vivo, reducing the occupancy of BRD4, MED1 and Pol II on many genes and enhancers\(^{36}\). However, how transcription (Pol II) is perturbed dynamically during this process has not been fully elucidated.

To probe protein-condensation dynamics, by taking advantage of the superb sensitivity of KAS-seq, we performed KAS-seq in HEK293T cells treated with 1.5% 1,6-hexanediol for 10 min (no treatment), 5 min, 15 min, 30 min and 60 min, respectively. PCA analysis showed that KAS-seq profiles at each time point are distinct from the others (Extended Data Fig. 10a), indicating dynamic transcription changes happening from 5 min to 60 min. Consistent with previous results\(^{36}\), total KAS-seq signals on the gene body gradually decrease from 15 min to 60 min (Extended Data Fig. 10b), supporting a role of protein condensate formation on transcription activation. However, after 5 min treatment, we observed a previously unnoticed increase of ssDNA clustered in a ~4-kb window...
around the TSS, which resulted in a slightly increased ssDNA signal on the gene body, accompanied by a decreased ssDNA signal at the TSS (Fig. 5a–c and Extended Data Fig. 10b). These ssDNA clusters form at both directions of TSS at bidirectional promoters (Fig. 5a), while they were only observed at TSS downstream regions for unidirectionally transcribed genes (Fig. 5b). As time went by,
Fig. 5 | KAS-seq reveals transcription dynamics upon inhibition of protein condensation. **a, b.** KAS-seq read densities around TSS on unidirectional (a) and bidirectional (b) transcribed genes after HEK293T cells were treated with 1,6-hexanediol for denoted time intervals. Arrows indicate the upstream and downstream ‘released’ KAS-seq signals at the 5 min time point. **c.** Snapshots of KAS-seq and Pol II ChIP-seq signals on the BAIAP2 gene after cells were treated with 1,6-hexanediol for denoted time intervals. Snapshots at different time points for each data set are staggered to clearly show differences. The autoscale setting was used for all tracks. The genomic coordinates and the Refseq tracks are aligned to the 60 min time point.

**d.** Pol II ChIP-seq read densities around TSS after cells were treated with 1,6-hexanediol for denoted time intervals. **e.** Box plot showing the calculated release index, the ratio of KAS-seq reads density at 0.5–2.5 kb downstream TSS at 5 min versus that under native state, and defined 4,510 genes as ‘fast-responsive genes’, 4,510 genes as ‘medium responsive genes’, 4,510 genes as ‘low responsive genes’, and 4,510 genes as ‘nonresponsive genes’. For **e and g**, the 10th to 90th percentile of data points is shown, with the center line showing the median, and the box limits showing the upper and lower quartiles. *P* values were calculated using two-sided unpaired Student’s *t* test.

These clustered ssDNA signals moved continuously towards TES and gradually diminished, accompanied with increased ssDNA signals at promoter-proximal regions (Fig. 5a–c and Extended Data Fig. 10c).

We next performed Pol II ChIP-seq at corresponding time points to validate the observations revealed by KAS-seq. The change of Pol II binding generally followed the changes observed by KAS-seq, with a portion of clustered Pol II released from TSS and subsequently moved towards TES at a similar speed as ssDNA clusters (Fig. 5c, d, Extended Data Fig. 10c).

Notably, the moving speed of these released Pol II is much slower (~40 kb per hour, Fig. 5c, Extended Data Fig. 10c) than the rate of Pol II elongation under native condition (>200 kb per hour), perhaps due to a lack of certain regulatory components under 1,6-hexanediol treatment.

We defined and ranked genes with the aforementioned ‘release’ feature by calculating a ‘release index’, the ratio of KAS-seq reads density at 0.5–2.5 kb downstream TSS at 5 min versus that under native state, and defined 4,510 genes as ‘fast-responsive genes’, with significant ssDNA cluster formation at this region (Fig. 5e).
We then performed similar analysis by using Pol II ChIP-seq. 75% (2,020/2,685) of fast-responsive genes defined by Pol II ChIP-seq overlap with those detected by KAS-seq, but KAS-seq detected considerably more genes (Extended Data Fig. 10d). This number difference and the metagene profiles (Fig. 5a,b,d) showed that KAS-seq exhibits higher sensitivity than Pol II ChIP-seq on revealing transcription dynamics during the early stage of inhibition. The extent of Pol II release correlates with the Pol II CTD serine-5 phosphorylation (SSP) level at TSS at the native state (Fig. 5f,g), supporting Pol II phosphorylation as a mechanism to tune transcription regulated through condensate formation20,29.

Discussion

KAS-seq simultaneously detects the dynamics of transcriptionally engaged Pol II, transcribing enhancers, potential non-B-form ssDNA structures and the activities of Pol I and Pol III with high sensitivity and low input material. ssDNA hotspots may form during DNA damage36, DNA replication and meiotic/mitotic double-strand break22,42. Although we focused on transcription in this work, the robust and tissue-friendly nature coupled with the feasibility of low input material make KAS-seq a method that can be broadly applied to profile transcription dynamics and other ssDNA-involving processes in rare samples, such as primary cells and patient samples.

ssDNA-containing enhancers show unique sequence features, correlate with more active transcription of downstream genes and enrich certain functions. Although we observed two different types of ssDNA-containing enhancers, our current analysis did not distinguish these two types. Current KAS-seq has a similar resolution as ChIP-seq, which is commonly used to study and define enhancers. Other techniques with higher resolution, or a high-resolution version of KAS-seq, may be applied to differentiate the two types of enhancers and to study their unique properties.

KAS-seq has revealed a previously unnoticed phosphorylation-dependent Pol II released from promoters to elongation at an early stage of protein-condensation inhibition, suggesting that protein condensates at promoters may store pre-phosphorylated Pol II20 and facilitate fast initiation—elongation transition. Released Pol IIs move continuously at a relatively slow speed upon condensate inhibition, while new Pol IIs recruited to the promoter are not subjected to release, potentially owing to the dissociation of a series of key TFs, coactivators and kinases required for elongation. A similar process may exist during cell response to other stresses. The nature of the Pol II complexes that are released from the promoter and elongated at a slow rate is unclear at this moment, nor is its potential physiological relevance or functional roles. Future characterization of this process and the complexes involved may reveal new insights into transcription regulation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0797-9.

Received: 30 October 2019; Accepted: 2 March 2020; Published online: 6 April 2020

References

1. Schwatzhauser, R. et al. Global quantification of mammalian gene expression control. Nature 473, 337–342 (2011).
2. Kim, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182–187 (2010).
3. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322, 1845–1848 (2008).
4. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science 339, 953–955 (2013).
5. Fuchs, G. et al. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. Genome Biol. 15, R69 (2014).
6. Schwalb, B. et al. TT-seq maps the human transient transcriptome. Science 352, 1225–1228 (2016).
7. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 469, 368–373 (2011).
8. Christopher, M. W., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. Mol. Cell 53, 819–830 (2014).
9. Nojima, T. et al. Mammalian NET-Seq reveals genomewide nascent transcription coupled to RNA processing. Cell 161, 526–540 (2015).
10. Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell 161, 541–554 (2015).
11. Hirabayashi, S. et al. NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. Nat. Genet. 51, 1369–1379 (2019).
12. Mirkovicth, J. & Darnell, J. E. Mapping of RNA polymerase on mammalian genes in cells and nuclei. Mol. Biol. Cell 3, 1085–1094 (1992).
13. Muse, G. W. et al. RNA polymerase is poised for activation across the genome. Nat. Genet. 39, 1507–1511 (2007).
14. Kouzine, F. et al. Global regulation of promoter melting in naive lymphocytes. Cell 153, 988–999 (2013).
15. Kouzine, F. et al. Permanganate/S1 nuclelease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. Cell Syst. 4, 344–356.e347 (2017).
16. Shaprio, R. & Hachmann, J. The reaction of guanine derivatives with 1,2-dicarbonyl compounds*. Biochemistry 5, 2799–2807 (1966).
17. Weng, X. et al. Keth-seq for transcriptome-wide RNA structure mapping. Nat. Chem. Biol. https://doi.org/10.1038/nchembio.1904 (2020).
18. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).
19. Cramer, P. Organization and regulation of gene transcription. Nature 573, 45–54 (2019).
20. Paule, M. R. Transcription by RNA polymerases I and III. Nucleic Acids Res. 28, 1283–1298 (2000).
21. Borchert, G. M., Lanier, W. & Davidson, B. L. RNA polymerase III transcribes human microRNAs. Nat. Struct. Mol. Biol. 13, 1097–1101 (2006).
22. Cer, R. Z. et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. Proc. Natl Acad. Sci. USA 111, 15941–15946 (2013).
23. Henriques, T. et al. Widespread transcriptional pausing and elongation control at enhancers. Genes Dev. 32, 26–41 (2018).
24. Warren et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319 (2013).
25. Raffaella et al. Control of embryonic stem cell identity by BRD4-dependent transcriptional elongation of super-enhancer-associated pluripotency genes. Cell Rep. 9, 234–247 (2014).
26. Liu, W. et al. BRD4 regulates Nanog expression in mouse embryonic stem cells and preimplantation embryos. Cell Death Differ. 21, 1950–1960 (2014).
27. Wu, T., Kamikawa, Y. F. & Donoohoe, M. E. Bdr4's bromodomains mediate histone H3 acetylation and chromatin remodeling in pluripotent cells through P300 and Brg1. Cell Rep. 25, 1756–1771 (2018).
28. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501 (2010).
29. Wang, D. et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 474, 390–394 (2011).
30. Li, W. et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498, 516–520 (2013).
31. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
32. Arner, E. et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science 347, 1010–1014 (2015).
33. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. Nat. Rev. Genet. 17, 207–223 (2016).
34. Huins, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for control of transcription. Cell 169, 13–23 (2017).
35. Roja, A. et al. Transcription factors activate genes through the phase-separation capacity of their activation domains. Cell 175, 1842–1855. e1816 (2018).
36. Sabari, B. R. et al. Coactivator condensation at super-enhancers links phase separation and gene control. Science 361, eaar3958 (2018).
37. Cho, W.-K. et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science 361, 412–415 (2018).
38. Chong, S. et al. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science 361, eaar2555 (2018).
39. Guo, Y. E. et al. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature 572, 543–548 (2019).
40. Zhou, Z. X. et al. Mapping genomic hotspots of DNA damage by a single-strand-DNA-compatible and strand-specific ChIP-seq method. Genome Res. 23, 705–715 (2013).
41. Khil, P. P., Smagulova, F., Brick, K. M., Camerini-Otero, R. D. & Petukhova, G. V. Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. Genome Res. 22, 957–965 (2012).
42. Lydall, D., Nikolsky, Y., Bishop, D. K. & Weinert, T. A meiotic recombination checkpoint controlled by mitotic checkpoint genes. Nature 383, 840–843 (1996).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020
Methods

Labeling DNA oligos with N\textsubscript{3}-kethoxal in vitro. We mixed 1 µL 100 µM synthetic DNA oligo (IDT) with 5 µL nucleic-acid-free water, 2 µL 5x reaction buffer (0.5 M sodium cacodylate, 50 mM MgCl\textsubscript{2}, pH 7.0) and 2 µL 500 mM N\textsubscript{3}-kethoxal (DMSO solution). The mixture was incubated at 37 °C for 30 min. The reaction product was purified by Micro Bio-Spin P-6 Gel Columns (Biorad, 7326222) and then used for matrix-assisted laser desorption/ionization-time-of-flight (MALDI-TOF) analysis directly. 2.4'-6’-trihydroxyacetophenone (10 mg mL\textsuperscript{-1} in 50% CH\textsubscript{3}CN/H\textsubscript{2}O) and ammonium citrate (50 mM L-arginine). Then, 1 µL mixed purified reaction product with 1 µL matrix on the MALDI sample plate and analyzed by Bruker UltraFlextreme MALDI-TOF.

Dot blot. We loaded 1 µL DNA onto the Amersham Hybond-N+ membrane (GE Healthcare, RN1198). Membranes were air-dried and were crosslinked by ultraviolet (UV) stratalinker 2400 at 150 mJ/cm\textsuperscript{2} twice. The membranes were then blocked overnight in 5% fatty-acid free BSA in PBST (0.1% Tween 20). The second day, the membrane was washed and incubated in streptavidin-HRP (Thermo, S-911) in PBS supplemented with 3% fatty-acid free BSA. The membrane was washed in PBS for five times before developed by Super Signal West Pico PLUS Chemiluminescent Substrate (Substrate, 34577).

Comparing the labeling reactivity of N\textsubscript{3}-kethoxal on deoxyguanosine and t-arginine. 2 mM deoxyguanosine or 2 mM L-arginine were mixed with 4 mM N\textsubscript{3}-kethoxal in neutral reaction buffer (0.1 M sodium cacodylate, 10 mM MgCl\textsubscript{2}, pH 7.0), respectively, at 37 °C and the reactions were monitored by thin-layer chromatography (TLC). The reaction between N\textsubscript{3}-kethoxal and deoxyguanosine was developed in 2:1 (v/v) ratio of dichloromethane and methanol, and was visualized by 254 nm UV light. The reaction between N\textsubscript{3}-kethoxal and t-arginine was developed in 1:1 (v/v) ratio of acetonitrile and ammonium hydroxide, and was visualized by ninhydrin staining.

Cell culture. HEK293T cells were purchased from ATCC (CRL12686) and were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), 1% penicillin and streptomycin (Gibco) and grown at 37 °C with 5% CO\textsubscript{2}. Murine embryonic stem (ES) cells were purchased from ATCC (CRL-1821) and were cultured in DMEM (Gibco 11995) supplemented with 10% (v/v) fetal bovine serum (Gibco), 0.1 mM M-mercaptoethanol (Gibco), 1% (v/v) nonessential amino acid stock (100x, Gibco), 1% penicillin/streptomycin stock (100x, Gibco), and 1,000 U/mL LIF (Millipore).

KAS-seq. N\textsubscript{3}-kethoxal was synthesized according to a previous protocol\textsuperscript{1}. Cells were incubated in completed culture medium containing 5 mM N\textsubscript{3}-kethoxal and for 5–10 min at 37 °C, 5% CO\textsubscript{2}. For transcription-inhibition experiments, cells were treated for 2 h under 100 µM DRB (Sigma, D1916) or 1 µM triptolide (Sigma, T3652) before incubated in the N\textsubscript{3}-kethoxal-containing medium. For 1.6-hexanediol treatment experiments, cells were treated with 1.5% (v/v) 1.6-hexanediol (Sigma, 240117) in the culture medium for 0 min, 5 min, 15 min, 30 min and 60 min before being subjected to N\textsubscript{3}-kethoxal labeling. Cells were collected and genomic DNA (gDNA) was isolated from cells by PureLink genomic DNA mini kit (Thermo, K182002). Then, 1 µg genomic DNA was suspended in 95 µL DNA elution buffer supplemented with 5 µL DMSO-PEG4-biotin (Biotin, DMSO solution, Sigma, 760749), 25 mM K\textsubscript{3}BO\textsubscript{4}, and incubated at 37 °C for 1.5 h while being gently shaken. Next, 5 µL RNase A (Thermo, 12091309) was added into the reaction mixture followed by incubation at 37 °C for 5 min. Biotinylated gDNA was then recovered by DNA Clean & Concentrator-5 kit (Zymo, D4013). gDNA was suspended into 100 µL water and was fragmented to 150–350 bp size by using Bioruptor Pico at 300 W for 45 min. The fragmented DNA was purified twice with 0.1% SDS lysis buffer, high salt buffer (50 mM HEPES, pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 1% SDS, 150 mM NaCl) and incubated on ice for 10 min, and then were sheared using Bioruptor Pico at the 30s-on/30s-off setting for 20 cycles. Of the sheared chromatin, 5% was saved as input and the rest was subjected to pre-clear and then mixed with 30 µL protein A/G beads coated with 5–10 µg antibodies. Immunoprecipitation was performed overnight and the resulting pellets were resuspended twice with 0.1% SDS lysis buffer, high salt buffer (50 mM HEPES, pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 1% SDS, 350 mM NaCl), LiCl wash buffer (100 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.5% NP-40, 0.5% sodium deoxycholate, 250 mM LiCl) and once with TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.2% Triton X-100). Enriched chromatin was eluted from the beads by incubating with elution buffer (50 mM NaHCO\textsubscript{3}, 10 mM EDTA, 1% SDS) at room temperature for 1 h. The eluent and the input were subjected to reverse crosslink and protease K digestion before DNA were purified from the mixture by using DNA Clean & Concentrator-5 kit (Zymo, D4013). Recovered DNA was used for library construction by Kapa HyperPlus kit (Kapa, KEB8515).

KAS-seq data processing and peak calling. Low-quality and adapter-containing reads were trimmed from KAS-seq raw data using trim-galore\textsuperscript{44} package in single-end mode. Reads shorter than 50 bp were removed. Trimmed reads were aligned to the reference genome (hg19 for HEK293T cells or mm10 for mESC) using bowtie2 (v2.3.3.1)\textsuperscript{45} under default parameters. Mapped sam files were subsequently converted to bedGraph from UCSC pre-compiled utilities. BedGraph files were then converted to bigWig files using bedGraph to bigWig from UCSC pre-compiled utilities. BigWig files were then subjected to similar overlap analysis with exons, introns and terminal regions. Peaks that did not have overlap with these genomic features are regarded as intergenic peaks.

ChIP-seq. Cells were crosslinked in 1% formaldehyde diluted in culture medium for 10 min and then quenched with 125 mM glycine for 5 min. Five million cells were used for all ChIP reactions. Crosslinked cells were resuspended in ice-cold lysis buffer (50 mM HEPES, pH 7.5, 9.5 mM MgCl\textsubscript{2}, 0.2% Triton X-100, 20% glycerol, 300 mM NaCl) and incubated on ice for 10 min before being centrifuged at 5000 g for 5 min. The pellets were resuspended in 0.1% SDS lysing buffer (50 mM Tris-HCl, pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.5% SDS, 150 mM NaCl) and incubated on ice for 10 min, and then were sheared using Bioruptor Pico at the 30s-on/30s-off setting for 20 cycles. Of the sheared chromatin, 5% was saved as input and the rest was subjected to pre-clear and then mixed with 30 µL protein A/G beads coated with 5–10 µg antibodies. Immunoprecipitation was performed overnight and the resulting pellets were resuspended twice with 0.1% SDS lysis buffer, high salt buffer (50 mM HEPES, pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 1% SDS, 350 mM NaCl), LiCl wash buffer (100 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.5% NP-40, 0.5% sodium deoxycholate, 250 mM LiCl) and once with TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.2% Triton X-100). Enriched chromatin was eluted from the beads by incubating with elution buffer (50 mM NaHCO\textsubscript{3}, 10 mM EDTA, 1% SDS) at room temperature for 1 h. The eluent and the input were subjected to reverse crosslink and proteinase K digestion before DNA were purified from the mixture by using DNA Clean & Concentrator-5 kit (Zymo, D4013). Recovered DNA was used for library construction by Kapa HyperPlus kit (Kapa, KEB8515).

RNA-seq. Total RNA was extracted from cells by using Trizol reagent (Thermo, 15596026). Total RNA were subjected to polyA selection by using Dynabeads mRNA purification kit (Thermo, 61006). Library construction was done with 20 ng polyA RNA using SMARTer Stranded RNA-seq kit (Takara, 634839).

Genome-wide distribution of KAS-seq peaks. A group of regions that have the same number and length of KAS-seq peaks were randomly generated from the hg19 genome by using bedtools shuffle. These random regions and real KAS-seq peaks in genomic regions are broad, MACS2 was run using broad peaks call mode under default parameters, except for ‘-broad-cutoff = 0.1’ and ‘--qvalue = 0.01’.
shorter than 50 bp were removed. The remaining trimmed sequences were mapped to the reference genome (hg19) with hisat2\textsuperscript{42} under default settings. The expression level of each gene was quantified with normalized fragments per kilobase of exon model per million reads mapped (FPKM) value with FPKM\_count.pl in the RSeQC\textsuperscript{43} software. Genes FPKM value higher than 0.5 were defined as expressed genes. In Fig. 3b, expressed genes were ranked and sorted into 3 groups based on their FPKM values, with the top 2,000 defined as high FPKM, 2,000 genes in the middle defined as medium FPKM and the bottom 2,000 defined as low FPKM. 2,000 genes with FPKM lower than 0.5 were randomly selected and defined as silent genes.

ChIP-seq data processing and peak calling. ChIP-seq data processing and peak calling generally follow the procedure used for KAS-seq data processing and peak calling.

**Correlation analysis.** Correlation calculations between KAS-seq, histone modification ChIP-seq and ATAC-seq were performed using deeptools\textsuperscript{44} package. First, multiBigwigSummary was used to calculate averaged read coverage within equally sized 100 kb bins of the entire genome. Regions in the human genome blacklist were excluded from the read coverage calculation. PlotCorrelation was subsequently used to calculate pairwise Pearson correlation coefficients with the output of multiBigwigSummary. Outliers were defined using the median absolute deviation (MAD) method by applying a threshold of 200, and were removed for correlation analysis. Heatmaps were generated with pairwise Pearson correlation coefficients depicted by varying color intensities, and were clustered using a hierarchical clustering. Correlation calculations between KAS-seq, Pol II ChIP-seq, GRO-seq and 4SU-seq were performed using a similar approach but were based on gene-coding regions.

**Definition of four transcription states.** To define transcription states, we calculated the KAS-seq tag density on the promoters (from −200 bp to +400 bp from TSS) and gene bodies (from +400 bp downstream TSS to TES) of protein-coding genes. Gene promoters with KAS-seq tag densities more than 20x as the density on average were considered to be paused. Similarly, gene bodies with KAS-seq tag densities more than 10x as the density on average were considered to be actively transcribed. A list of genes at four different transcription states can be found in supplemented Source Data.

**Defining genes with long, medium and short-terminal regions.** Genes with terminal regions that do not overlap with other genes within a 10-kb range downstream of the TES were used for analysis. The 10-kb region downstream TES was divided into 20 bins of the same length. We calculated the averaged KAS-seq reads density on each bin, and bins with averaged KAS-seq reads density equal to or greater than five were defined as positive bins. We ranked all genes according to their number of positive bins, from highest to lowest. Genes ranked among top third were defined as long-terminal genes; genes ranked among bottom third were defined as short-terminal genes; the rest were defined as medium-length terminal genes.

**Calculation of the termination index.** We calculated the termination index for KAS-seq, Pol II ChIP-seq, and GRO-seq as the log2 ratio of reads density on terminal regions (from TES to +2 kb from TES) over that around TSS (from −200 bp to +400 bp from TSS). Only genes with KAS-seq tag density on promoters more than 50x as the density on average were included in the calculation.

**Identification of predicted non-B form DNA with KAS-seq peaks.** The positions of all the non-B-form DNA motifs in this study are downloaded from non-B D B v2.0 (ref. 22). To obviate the effect of Pol-II-induced KAS-seq signals, we used KAS-seq peaks identified in triploid-treated HEK293T cells. KAS-seq peaks related to TRNA, RNA, small NS80-associated RNAs and U6 spicosomal RNA, which are transcribed by Pol I and Pol III, were excluded from the analysis. Enrichment of KAS-seq signals on non-B-form DNAs were determined by calculating log(IP/reads density/input reads density) on each KAS-seq-positive non-B-form DNA region, with the distribution of enrichment for each non-B-form DNA type shown in box plots, compared with the same number of regions randomly found in the genome. To calculate the enrichment of KAS-seq signal on telomeres, we used the KAS-seq signals the 15 kb rightmost and 15 kb leftmost regions of all chromosomes on the hg38 reference genome.

**Definition of ssDNA-containing enhancers and super-enhancers.** We used H3K27ac and H3K4m1 peaks distal from genes promoters (based on mm10 and hg19 on NCBI RefSeq) to define active and poised enhancers. H3K27ac-enriched regions were defined as active enhancers, regions with enriched H3K4m1 but not H3K27ac were defined as poised enhancers. We found that very few poised enhancers are single-stranded, so only active enhancers with KAS-seq peaks were defined as single-strand-DNA-containing enhancers. In addition, some active enhancers are located on the gene body. Thus KAS-seq signals on these enhancers may derive from Pol II elongation. Therefore, KAS-seq under DRB treatment was used to define single-stranded-DNA-containing enhancers. Active enhancers with KAS-seq peaks observed in both DRB replicates were defined as single-stranded-DNA-containing enhancers. Super-enhancers were defined using the ROSE package as previously described\textsuperscript{45}.

**Motif analysis.** Sequence motifs enriched by ssDNA-containing enhancers and ATAC-seq-positive but KAS-seq negative enhancers were analyzed by using HOMER\textsuperscript{46}.

The sequences of ssDNA-containing enhancers were extracted and used as input for TRAP\textsuperscript{47} using TRANSFAC vertebrates as the comparison library, promoter sequences as the background and Benjamini–Hochberg as the correction. P values are displayed in figures corresponding to the ‘corrected’ P in the output.

**Assigning enhancers to their regulated genes.** We assigned enhancers to their regulated genes based on the NCBI ReSeq gene annotations. We calculated the distance from the center of the enhancer to the TSS of each gene, and the gene closest to the enhancer and with the distance less than 50 kb is assigned as the gene regulated by this enhancer.

**Pol II and CTCF ChIA-pet data were used to define the long-range interactions.**

**Calculate the release index to define genes responding to protein-condensation inhibition.** We calculated Pol II or ssDNA release index as the log ratio of Pol II or ssDNA reads density in a region from +0.5 kb to +2.5 kb downstream TSS at 5 min versus that with no treatment (0 min). As some genes have very short gene bodies, only genes with gene bodies longer than 5 kb were included in the calculation. Genes with Pol II or ssDNA release index higher than 0.5 were defined as genes responding to protein-condensation inhibition, which were sorted into high-, medium- and low-responsive genes groups, with the number of genes in each group the same. Genes with Pol II or ssDNA release index lower than 0.2 were defined as non-responsive genes.

**Definition of bidirectional and unidirectional promoters.** We defined bidirectional and unidirectional promoters by analyzing published NET-seq data in HEK293 cells by following a previous method\textsuperscript{48}. Promoter-proximal regions were carefully defined to ensure minimal signal contamination from genes nearby. Genes shorter than 5 kb were excluded from the analysis. Genes with TSS located within 2.5 kb upstream of the TSS of another gene, or 2.5 kb downstream of the polyA cleavage site of another gene, were excluded from the analysis. In cases of conflicting isoform annotations, the most upstream annotated TSS and the most downstream annotated polyA cleavage sites were used. Within a 4 kb region around TSS, promoters with more than 40 NET-seq signals covering both sense and antisense directions were defined as bidirectional. In contrast, promoters with 40 NET-seq signals covering only sense but not antisense direction were defined as unidirectional.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**

All sequencing data are available at NCBI Gene Expression Omnibus with the accession number: GSE139420. Other data that support the findings of this study are available from the corresponding author upon request.

**References**

43. Wu, T., Lyu, R., He, C. Kethoxal-assisted single-stranded DNA sequencing (KAS-seq) for capturing transcription dynamics and enhancer activity. Protoc. Exch. https://doi.org/10.21203/rs.3.pex-835/v1 (2020).

44. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17, 10 (2011).

45. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).

46. Li, H. et al. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

47. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008).

48. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with high sensitivity. Bioinformatics 34, W187–W191 (2018).

51. Heinz, S. et al. Simple combinations of lineage-determining transcription factors are sufficient to model mesoderm and endoderm in zebrafish. Mol. Cell 38, 576–589 (2010).
52. Thomas-Chollier, M. et al. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat. Protoc.* 6, 1860–1869 (2011).

**Acknowledgements**
We thank all He lab members for discussion. We thank B. Harada for helpful comments on the manuscript. We thank Genomics Facility at the University of Chicago for performing high-throughput sequencing (P30 CA014599). This work was supported by US National Institutes of Health (R01 HG006827, R01 HG008935 and P01 NS097206 to C.H.). C. H. is an investigator of the Howard Hughes Medical Institute.

**Author contributions**
All authors designed experiments and interpreted the data. T.W. performed the experiments with suggestions from Q.Y.R.L. performed the bioinformatics analysis. T.W. and C.H. wrote the paper with input from all authors.

**Competing interests**
The University of Chicago has filed a patent application on KAS-seq. C.H. is a scientific founder and a member of the scientific advisory board of Accent Therapeutics, Inc., and a shareholder of Epican Genetech.

**Additional information**
Extended data is available for this paper at https://doi.org/10.1038/s41592-020-0797-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0797-9.

Correspondence and requests for materials should be addressed to C.H.
Peer review information Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.
Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Characterization of the N3-kethoxal-based labeling. a, MALDI-TOF analysis of the reaction between a 16-mer DNA oligo and N3-kethoxal. The experiment was performed in duplicates with similar results obtained. b, TLC analysis of the reaction between N3-kethoxal and deoxyguanosine (dG, left) or L-arginine (L-Arg, right) after different time intervals. The N3-kethoxal-dG results were visualized by 254 nm UV light. The N3-kethoxal-L-Arg results were visualized by ninhydrin staining. The experiment was performed in duplicates with similar results obtained. c-d, The DNA yield (c) and the A260/280 ratio (d) of gDNA isolated from N3-kethoxal-treated and control cells. P values were calculated by using two-sided unpaired Student’s t-test (n = 3 independent experiments). e, Dot blot showing biotin signals of the DNA after the biotinylation reaction in the presence or absence of N3-kethoxal or biotin-DBCO. Results from two replicates were shown for each condition. The experiment was performed in duplicates with similar results obtained. f, Agarose gel image showing the profile of libraries constructed by using input and enriched DNA samples made in the presence or absence of N3-kethoxal or biotin-DBCO. Results from two replicates were shown for each condition. The experiment was performed in duplicates with similar results obtained.
Extended Data Fig. 2 | KAS-seq validation and an overview of the KAS-seq profile. a, Fingerprint plot of KAS-seq libraries and the corresponding inputs in HEK293T cells. b, Pearson correlation scatterplot between two independent KAS-seq replicates (r = 0.99) in HEK293T cells (n = 287,970 10 Kb bins in the hg19 genome). c, Peak overlaps between two independent KAS-seq replicates in HEK293T cells. The p value was calculated using two-sided Fisher’s exact test. d, Reads distributions of KAS-seq (left) and Pol II ChIP-seq (right) signals respect to different GC fractions. e, Heatmap showing reads distribution of two independent KAS-seq replicates at gene-coding regions in mESCs. f, The distribution of KAS-seq signals, ATAC-seq signals, and selected histone modifications at gene-coding regions in HEK293T cells. g, Heatmap showing the reads distribution of two KMnO₄/S1 footprinting replicates (activated mouse B cells) at gene-coding regions.
Extended Data Fig. 3 | KAS-seq using low input cells and mouse liver. KAS-seq signal distribution at gene-coding regions revealed by using different numbers of HEK293T cells (n = 26,910 genes). a, Profiles of KAS-seq data at gene-coding regions using different numbers of HEK293T cells. b, Fingerprint plot of low-input KAS-seq libraries. c, Numbers of KAS-seq peaks detected by using different amounts of HEK293T cells. e, Heatmap showing reads distribution of two independent KAS-seq replicates at gene-coding regions generated by using livers from two mice. 1M: 1 million; 10K: 10 thousand; 5K: 5 thousand; 1K: 1 thousand.
Extended Data Fig. 4 | Correlation between KAS-seq signals, gene expression levels, Pol II dynamics, and gene transcription states. a, Venn diagram showing the overlap between KAS-seq peaks and Pol II ChIP-seq peaks at promotor in mESCs. The p value was calculated using two-sided Fisher’s exact test. b, Pearson correlation scatterplot (n=24,359 genes) between KAS-seq and Pol II ChIP-seq at gene bodies in mESCs. The r value was calculated as two-tailed probability. c, Genes were grouped according to different expression levels based on RNA-seq. 10–90 percentile of data points are shown, with the center line showing the median, and the box limits showing the upper and lower quartiles. d, Metagene profile of KAS-seq signals at gene-coding regions under control, DRB treatment, and triptolide treatment conditions. e, a snapshot of KAS-seq profiles from UCSC Genome Browser under control, DRB treatment, and triptolide treatment conditions. f, Heatmaps showing KAS-seq, Pol II ChIP-seq, and GRO-seq signals on genes with four different transcription states defined by using KAS-seq.
Extended Data Fig. 5 | KAS-seq shows no significant length-dependent bias and yields strong signals around TES regions. a, A snapshot from UCSC Genome Browser showing KAS-seq and Pol II ChIP-seq profiles at the native state, and KAS-seq profile at the DRB-treated state, indicating that KAS-seq signals around TES are derived from Pol II. Autoscale setting is used for all tracks. b, KAS-seq reads densities of three groups of genes with different lengths of termination signals. c, Averaged KAS-seq reads density in the entire terminal regions in the three groups of genes defined in (b). n = 660 genes for all three groups. d, Termination index for each gene was calculated as the ratio of KAS-seq reads density on TES to its downstream 2 kb region, versus reads density on the -200 bp to +400 bp region around TSS. e, The distribution of termination index for all genes in KAS-seq, GRO-seq, and Pol II ChIP-seq (n = 29,160 genes). For c and e, 10 - 90 percentile of data points are shown, with the center line showing the median, and the box limits showing the upper and lower quartiles. P values were calculated using two-sided unpaired Student’s t-test.
Extended Data Fig. 6 | KAS-seq detects Pol I and Pol III-mediated transcription events, as well as other non-B form DNA structures and telomeric DNA regions. a–c, Snapshots of KAS-seq signals at selected small RNA, tRNA, and rRNA loci in HEK293T cells under native, DRB treatment, and triptolide treatment conditions. d, A summary of different types of non-B form DNA structures and the number of KAS-seq peaks (under triptolide-treatment condition) detected at each type of predicted non-B form DNA regions. e, Snapshots from UCSC genome browser showing examples of KAS-seq signals under native, DRB, and triptolide-treatment conditions at different non-B form DNA regions and telomeric DNA regions. f, Enrichment of KAS-seq signals at different non-B form DNA and telomeric DNA regions showed in (d). n = 715 regions for hairpin, n = 1,643 regions for cruciform, n = 730 regions for H-DNA, n = 356 regions for quadruplex, n = 256 regions for Z-DNA, n = 29 regions for telomere.
Extended Data Fig. 7 | Features of ssDNA-containing enhancers in mESCs. a, All ATAC-seq-positive enhancers were sorted into two groups based on whether they are KAS-seq-positive or not. Heatmaps of KAS-seq, ATAC-seq, and Pol II ChIP-seq signals on these two groups are shown. b, A metagene profile showing ATAC-seq reads density on the two groups of enhancers defined in (a). c, Expression levels of genes associated with KAS-seq positive (\(n = 3,080\) genes) and KAS-seq negative (\(n = 1,544\) genes) enhancers defined in (a). 10 - 90 percentile of data points are shown, with the centerline showing the median, and the box limits showing the upper and lower quartiles. The p value was calculated using two-sided unpaired Student's t-test. d, Sequence motifs enriched in ATAC-seq-positive but KAS-seq-negative enhancers from mESCs (\(n = 6,082\) enhancers). The p values were calculated by two-sided binomial test. e, Metagene profiles of Nanog, Oct4 and Sox2 ChIP-seq read densities at denoted enhancers in mESCs. Regions within 10 kb around the enhancer centers are shown.
Extended Data Fig. 8 | ssDNA-containing enhancers in HEK293T cells. a, A group of enhancers are single-stranded in HEK293T cells. Heatmap of KAS-seq reads densities at all enhancer regions in HEK293T cells. Active and poised enhancer regions are defined by distal H3K27ac and H3K4me1 signals. Active enhancers are sub-grouped into SSEs and DSAEs. b, Distribution of H3K27ac ChIP-seq signal across all HEK293T enhancers. Super-enhancers are defined as containing exceptionally high amounts of H3K27ac. c, The number of ssDNA-containing enhancers and super-enhancers in HEK293T cells and the overlap. The p value was calculated by two-sided Fisher’s exact test. d, KAS-seq reads densities on SSEs in HEK293T cells under native and DRB-treatment conditions. e, Metagene profiles of KAS-seq, Pol II, H3K4me3, and H3K27ac ChIP-seq reads densities at denoted enhancers in HEK293T cells. Regions within 10 kb around the enhancer centers are shown. SSE: ssDNA-containing enhancers; DSAE: double-stranded active enhancers; PE: poised enhancers.
Extended Data Fig. 9 | Transcription factors that preferentially bind at ssDNA-containing enhancers in HEK293T cells. a, Metagene profiles of CTCF, YY1, SP1, SP2, MAZ, NCAPH2, KLF8, KLF9, ZNF335, ZNF341, ZBTB20, and ZBTB26 ChIP-seq reads densities at denoted enhancers in HEK293T cells. Regions within 10 kb around the enhancer centers are shown. b, Transcription factor binding motifs enriched at ssDNA-containing enhancers (n = 1,969 enhancers) in HEK293T cells with corresponding p values by using the genome as background. Only TFs with motif information in the TRANSFAC vertebrates library were analyzed. P values were calculated by two-sided binomial test. c, GREAT analysis of genes regulated by ssDNA-containing enhancers (n = 1,969 enhancers) in HEK293T cells. P values were calculated by two-sided binomial test. SSE: ssDNA-containing enhancers; DSAE: double-stranded active enhancers; PE: poised enhancers.
Extended Data Fig. 10 | KAS-seq and Pol II ChIP-seq signals in response to protein condensation inhibition. 

a, PCA analysis of KAS-seq data at different time points after 1,6-hexanediol treatment (n = 3,122,843 1 kb bins). 

b, Box plots showing normalized KAS-seq reads densities on gene bodies (from 0.5 kb downstream TSS to TES) of the genes defined as responsive to 1,6-hexanediol treatment. 10–90 percentile of data points are shown, with the center line showing the median, and the box limits showing the upper and lower quartiles. P values were calculated by using two-sided unpaired Student’s t-test. 

c, Heat map showing the release and movement of KAS-seq signals (left) and Pol II clusters (right) from 0 min to 60 min after 1,6-hexanediol treatment. 

d, Numbers of fast responsive genes defined by KAS-seq and Pol II ChIP-seq, and the overlap. The p value was calculated by two-sided Fisher’s exact test.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed |
|-----|-----------|
|     |           |

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted.
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated
- Clearly defined error bars
  State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

No software was used.

Data analysis

Bowtie2(v2.3.4.3); samtools(v1.9); bedtools(v2.28.0); Trim-galore(v0.6.2); bedgraphtobigwig; MACS2(v2.1.1.20160309); hisat2(v2.1.0); RSeQC(v3.0.1); Deeptools(v2.0); R(v3.6.2); TRAP(v3.05);

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data have been deposited in the NCBI Gene Expression Omnibus (GEO) and are accessible through GEO series accession number GSE139420.
### Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

**Sample size**
No statistical methods were used to predetermine sample size. For sequencing data, sample size were determined based on our prior experience on similar experiments and literature reports. We performed all biochemistry experiments in duplicates or triplicates to make sure results are consistent. For cell-based assays, samples were collected till we have sufficient number to obtain reliable statistics.

**Data exclusions**
For correlation analysis, we removed outliers using the median absolute deviation (MAD) method by applying a threshold of 200, which is commonly used.

The information about blacklisted genomic regions for functional genomics analysis were provided by the ENCODE project.

To define unidirectional and bidirectional enhancers, short genes and genes very close to each other were excluded from the analysis to avoid signal contaminations from nearby genes.

**Replication**
Results were confirmed in two biological replicates for each experiment unless otherwise stated. All attempts to replicate data are successful.

**Randomization**
The experiments were not randomized. Controlling for covariates was unnecessary because all assays were performed in pairs.

**Blinding**
The investigators were not blinded to allocation during experiments and outcome assessment due to feasibility.

### Reporting for specific materials, systems and methods

#### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
|     | Unique biological materials |
| x   | Antibodies |
| x   | Eukaryotic cell lines |
|     | Palaeontology |
|     | Animals and other organisms |
|     | Human research participants |

#### Methods

| n/a | Involved in the study |
|-----|-----------------------|
|     | ChIP-seq |
| x   | Flow cytometry |
| x   | MRI-based neuroimaging |

#### Antibodies

**Antibodies used**
SP1 antibody (polyclonal), Abcam, catalog No: ab13370, lot No: GR252118-39. 
Pol II antibody (clone No: 8WG16), Biolegend, catalog No: 64912, lot No: B281694.

**Validation**
Validation statements of the SP1 antibody for human samples is available on the manufacturer’s websites.
Validation statements of the Pol II antibody for human samples is available on the manufacturer’s websites.

#### Eukaryotic cell lines

**Policy information about cell lines**

**Cell line source(s)**
HEK293T cells are purchased from ATCC (catalog No: CRL1286). 
mESCs are purchased from ATCC (Catalog No: CRL-1821).

**Authentication**
None of the cell lines used were authenticated.

**Mycoplasma contamination**
All cell lines used in this study were tested negative of mycoplasma contamination.
Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

**Laboratory animals**

Male B6 mice were purchased from The Jackson Laboratory (catalog No: C57BL/6J). All mice were used at 6-12 weeks of age. All mice were maintained under specific pathogen-free conditions and used in accordance with the animal experimental guidelines set by the Institute of Animal Care and Use Committee.

**Wild animals**

This study did not involve wild animals.

**Field-collected samples**

This study did not involve samples collected from field.

ChIP-seq

**Data deposition**

- Confirm that both raw and final processed data have been deposited in a public database such as GEO.
- Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

**Data access links**

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139420;elwnagouzzoxrkb.

**Files in database submission**

- GSM4141457 HEK293_SP1_Input.rep1
- GSM4141458 HEK293_SP1_Input.rep2
- GSM4141459 HEK293_SP1.rep1
- GSM4141460 HEK293_SP1.rep2
- GSM4141461 HEK293_Phase_Inhibition_0min_Pol2_ChIP-seq.rep1
- GSM4141462 HEK293_Phase_Inhibition_0min_Pol2_ChIP-seq.rep2
- GSM4141463 HEK293_Phase_Inhibition_0min_Pol2_Input.rep1
- GSM4141464 HEK293_Phase_Inhibition_0min_Pol2_Input.rep2
- GSM4141465 HEK293_Phase_Inhibition_5min_Pol2_ChIP-seq.rep1
- GSM4141466 HEK293_Phase_Inhibition_5min_Pol2_ChIP-seq.rep2
- GSM4141467 HEK293_Phase_Inhibition_5min_Pol2_Input.rep1
- GSM4141468 HEK293_Phase_Inhibition_5min_Pol2_Input.rep2
- GSM4141469 HEK293_Phase_Inhibition_15min_Pol2_ChIP-seq.rep1
- GSM4141470 HEK293_Phase_Inhibition_15min_Pol2_ChIP-seq.rep2
- GSM4141471 HEK293_Phase_Inhibition_15min_Pol2_Input.rep1
- GSM4141472 HEK293_Phase_Inhibition_15min_Pol2_Input.rep2
- GSM4141473 HEK293_Phase_Inhibition_30min_Pol2_ChIP-seq.rep1
- GSM4141474 HEK293_Phase_Inhibition_30min_Pol2_ChIP-seq.rep2
- GSM4141475 HEK293_Phase_Inhibition_30min_Pol2_Input.rep1
- GSM4141476 HEK293_Phase_Inhibition_30min_Pol2_Input.rep2
- GSM4141477 HEK293_Phase_Inhibition_60min_Pol2_ChIP-seq.rep1
- GSM4141478 HEK293_Phase_Inhibition_60min_Pol2_ChIP-seq.rep2
- GSM4141479 HEK293_Phase_Inhibition_60min_Pol2_Input.rep1
- GSM4141480 HEK293_Phase_Inhibition_60min_Pol2_Input.rep2

**Genome browser session**

http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr10:60033-60183&hgsid=775660251_qWL8HOG8P2u6IS8gz4T1k2mHXIA

**Methodology**

**Replicates**

We generated two biological replicates for all ChIP-seq experiments.

**Sequencing depth**

Sample, Total number of reads, Uniquely mapped reads, Length of reads, Paired- or single-end

| Sample | Total number of reads | Uniquely mapped reads | Length of reads | Paired-end |
|--------|-----------------------|-----------------------|----------------|------------|
| HEK293_Phase_Inhibition_0min_Pol2_ChIP-seq.rep1 | 23740696 | 18129270 | 106 | single-end |
| HEK293_Phase_Inhibition_0min_Pol2_ChIP-seq.rep2 | 26908310 | 19678312 | 106 | single-end |
| HEK293_Phase_Inhibition_5min_Pol2_ChIP-seq.rep1 | 20442485 | 15700754 | 106 | single-end |
| HEK293_Phase_Inhibition_5min_Pol2_ChIP-seq.rep2 | 23301890 | 18255541 | 106 | single-end |
| HEK293_Phase_Inhibition_15min_Pol2_ChIP-seq.rep1 | 19733902 | 16003346 | 106 | single-end |
| HEK293_Phase_Inhibition_15min_Pol2_ChIP-seq.rep2 | 24489814 | 19493644 | 106 | single-end |
| HEK293_Phase_Inhibition_30min_Pol2_ChIP-seq.rep1 | 20986742 | 17247544 | 106 | single-end |
| HEK293_Phase_Inhibition_30min_Pol2_ChIP-seq.rep2 | 25091056 | 20549824 | 106 | single-end |
| HEK293_Phase_Inhibition_60min_Pol2_ChIP-seq.rep1 | 23513679 | 18168194 | 106 | single-end |
| HEK293_Phase_Inhibition_60min_Pol2_ChIP-seq.rep2 | 289100723 | 24891947 | 106 | single-end |
### Antibodies

| Sample                     | Nature ID | Start Position | End Position | Length | Library Type | Notes  |
|----------------------------|-----------|----------------|--------------|--------|--------------|--------|
| HEK293_Phase_Inhibition_30min_Pol2_ChIP-seq.rep2 | 21533810 | 17141137 | 106 | single-end |
| HEK293_Phase_Inhibition_30min_Pol2_Input.rep1  | 21579406 | 18844620 | 106 | single-end |
| HEK293_Phase_Inhibition_30min_Pol2_Input.rep2  | 23767453 | 20820210 | 106 | single-end |
| HEK293_Phase_Inhibition_60min_Pol2_ChIP-seq.rep1 | 20069322 | 15282611 | 106 | single-end |
| HEK293_Phase_Inhibition_60min_Pol2_ChIP-seq.rep2 | 21135666 | 15182909 | 106 | single-end |
| HEK293_Phase_Inhibition_60min_Pol2_Input.rep1  | 10005733 | 8634317 | 106 | single-end |
| HEK293_Phase_Inhibition_60min_Pol2_Input.rep2  | 17360718 | 15080039 | 106 | single-end |
| HEK293_SP1_Input.rep1  | 40947324 | 31727477 | 51 | single-end |
| HEK293_SP1_Input.rep2  | 40567692 | 32042965 | 51 | single-end |
| HEK293_SP1.rep1  | 30656408 | 23775835 | 51 | single-end |
| HEK293_SP1.rep2  | 35703045 | 27324189 | 51 | single-end |

### Antibodies

- SP1 antibody (polyclonal), Abcam, catalog No: ab13370, lot No: GR252118-39.
- Pol II antibody (clone No: 8WG16), Biolegend, catalog No: 64912, lot No: B281694.

### Peak calling parameters

- macs14/macs2; pvalue=1e-5(macs14); --broad -g hs --broad-cutoff 0.01 --qvalue 0.01.

### Data quality

- We used fastqc to check the sequencing quality of ChIP-seq sequencing data, then use cutadapt to remove reads with adapter and low-quality sequences.

### Software

- Trim-galore;
- Bowtie2(v2.3.4.3);
- samtools(v1.9);
- macs(v2.1.1.20160309);
- Deeptools(v2.0);
- bedgraphbtobigwig;
- bedtools(v2.28.0);
- TRAP.