

Quantitative analysis of cis-regulatory elements in transcription with KAS-ATAC-seq

Received: 6 March 2024

Accepted: 18 July 2024

Published online: 10 August 2024

 Check for updatesRuitu Lyu^{1,2,3,5} , Yun Gao^{1,2,3,5}, Tong Wu^{1,2,3,5}, Chang Ye^{1,2,3},
Pingluan Wang^{1,2,3} & Chuan He^{1,2,3,4} 

Cis-regulatory elements (CREs) are pivotal in orchestrating gene expression throughout diverse biological systems. Accurate identification and in-depth characterization of functional CREs are crucial for decoding gene regulation networks during cellular processes. In this study, we develop Kethoxal-Assisted Single-stranded DNA Assay for Transposase-Accessible Chromatin with Sequencing (KAS-ATAC-seq) to quantitatively analyze the transcriptional activity of CREs. A main advantage of KAS-ATAC-seq lies in its precise measurement of ssDNA levels within both proximal and distal ATAC-seq peaks, enabling the identification of transcriptional regulatory sequences. This feature is particularly adept at defining Single-Stranded Transcribing Enhancers (SSTEs). SSTEs are highly enriched with nascent RNAs and specific transcription factors (TFs) binding sites that define cellular identity. Moreover, KAS-ATAC-seq provides a detailed characterization and functional implications of various SSTE subtypes. Our analysis of CREs during mouse neural differentiation demonstrates that KAS-ATAC-seq can effectively identify immediately activated CREs in response to retinoic acid (RA) treatment. Our findings indicate that KAS-ATAC-seq provides more precise annotation of functional CREs in transcription. Future applications of KAS-ATAC-seq would help elucidate the intricate dynamics of gene regulation in diverse biological processes.

Gene expression regulation is largely mediated by cis-regulatory elements (CREs), which play a critical role in modulating gene functions across various biological processes^{1–3}. CREs generally contain specific binding sites of transcription factors (TFs)¹. DNA segments bound by TFs are often depleted of nucleosomes and are flanked by active histone marks⁴. Distal CREs, notably enhancers, engage in physical interactions with their target promoters, sometimes in a multilateral fashion^{5–8}. Despite their crucial roles in creating cell type-specific transcriptomes, the precise

mechanisms underlying the dynamic activation and precise looping of CREs have not been fully elucidated.

Distal CREs can be transcribed into both stable or unstable RNA transcripts, known as enhancer RNAs (eRNAs)^{9,10}. Only a subset of distal CREs is capable of enhancing gene transcription¹¹. Transcribed CREs typically demonstrate a stronger correlation with transcription activation and possess more functional relevance than those identified solely based on active histone marks, accessible chromatin regions, or DNase I hypersensitive sites (DHSs)^{12–19}. The precise identification and

¹Department of Chemistry, The University of Chicago, Chicago, IL 60637, USA. ²Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, IL 60637, USA. ³Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL 60637, USA. ⁴Howard Hughes Medical Institute, The University of Chicago, Chicago, IL 60637, USA. ⁵These authors contributed equally: Ruitu Lyu, Yun Gao, Tong Wu. ✉e-mail: lvruitu@uchicago.edu; chuanhe@uchicago.edu

in-depth characterization of these transcribed CREs are critical, especially in the context of cell differentiation, where subtle variations in gene regulation can result in significant phenotypic differences²⁰. A variety of nascent RNA based approaches have been developed to study the dynamics of CREs transcription, including global run-on sequencing (GRO-seq)^{21,22}, precision run-on sequencing (PRO-seq) and its variant PRO-CAP^{23,24}, cap analysis of gene expression (CAGE) and NET-CAGE^{25,26}, metabolic labeling with 4-thiouridine (4sU RNA)²⁷, and mammalian native elongating transcript sequencing (mNET-seq)²⁸. In addition, ATAC-seq has been developed for profiling chromatin accessibility²⁹, and KAS-seq has been recognized for its rapid and sensitive detection of genome-wide single-stranded DNA (ssDNA) produced by transcriptionally active RNA polymerases in situ^{30,31}, both serve as proxy of CRE activities.

These genomic methods all fill critical gaps, yet they also have limitations in terms of accuracy, sensitivity, or input material requirements. Nascent RNA-based methods are capable of directly detecting enhancer RNA (eRNA) but lack sufficient sensitivity due to the inherent instability of eRNA. Additionally, these methods are ineffective with limited starting materials. ATAC-seq and active histone marks are instrumental in defining enhancers, but they frequently can't reflect the transcriptional activity, and a substantial number of distal CREs also function as poised enhancers and insulators. KAS-seq offers a promising alternative by efficiently detecting ssDNA on enhancers and gene transcription units, which is indicative of active transcription. However, it faces its own set of challenges, particularly in distinguishing enhancer-associated ssDNA from other ssDNA signals across the genome. These limitations highlight the need for a more refined approach that can overcome these shortcomings of existing methods, ensuring a more accurate and comprehensive understanding of enhancer dynamics and their impact on gene regulation.

In this study, we develop an optimized KAS-seq (*Opti-KAS-seq*) protocol that significantly enhances the efficiency of capturing ssDNA. *Opti-KAS-seq* offers broader genomic coverage and higher signal-to-background ratio that works across a wide range of applications and sample types. By integrating the sensitive *Opti-KAS-seq* with ATAC-seq, we further introduce Kethoxal-Assisted Single-stranded DNA Assay for Transposase-Accessible Chromatin with Sequencing (KAS-ATAC-seq) with the dual capability to simultaneously uncover chromatin accessibility and transcriptional activity of CREs. A major advantage of KAS-ATAC-seq lies in its precise measurement of ssDNA levels within CREs, enabling the de novo identification of ssDNA promoter and Single-Stranded Transcribing Enhancers (SSTEs) as a subset of CREs without relying on eRNA or active histone marks ChIP-seq data. Additionally, we apply KAS-ATAC-seq to examine the transcriptional dynamics of CREs during the neural differentiation of mESCs into neural progenitor cells (NPCs). This analysis uncovers the involvement of specific transcription factors (TFs), including ETS and YY1, in the regulation of immediate-early activated promoters and SSTEs in response to RA treatment. These findings demonstrate the capability of KAS-ATAC-seq as a powerful genomic method for precisely exploring and understanding the gene regulatory mechanisms by CREs.

Results

Enhancing ssDNA capture efficiency with optimized KAS-seq procedure

Transcription is a multifaceted and dynamic process that generates single-stranded DNA (ssDNA) regions in the genome, commonly referred as 'transcription bubbles'³². In our previous work, we developed KAS-seq to map transcriptional activities by sensitively capturing and sequencing genome-wide ssDNA through the N₃-kethoxal-assisted labeling. Although current KAS-seq approach has proven to be effective in many contexts^{33–38}, we and others have

noticed compromised sensitivity of KAS-seq when using certain tissue samples and primary cells obtained using fluorescence-activated cell sorting (FACS). Our investigations suggested that this compromised efficiency in ssDNA capture might be due to the limited diffusion of N₃-kethoxal through the cell membrane of these primary cells and tissues. We therefore modified the cell labeling procedure of KAS-seq by adding a cell permeabilization step, which allows N₃-kethoxal to enter cells and label ssDNA more efficiently (Fig. 1a).

To confirm the effectiveness of the optimized KAS-seq (*Opti-KAS-seq*) protocol, we first tested it with HEK293T cells and conducted a thorough comparison between KAS-seq and *Opti-KAS-seq* under an equal number of uniquely mapped reads. Our quality control assessment revealed that the reproducibility, consistency, and robustness of *Opti-KAS-seq* match those of the conventional KAS-seq protocol (Supplementary Fig. 1a–f). A detailed exploration of KAS-seq peaks in HEK293T cells indicated that *Opti-KAS-seq* substantially elevates ssDNA detection sensitivity across promoters, distal enhancers, and other genomic regions (Fig. 1b–d). Comparative analysis between *Opti-KAS-seq* and KAS-seq, including peak overlaps (Fig. 1c), fingerprint plots (Supplementary Fig. 1g), and gene-coding enrichment (Fig. 1e), confirmed the expanded genomic coverage and elevated signal intensity achieved by *Opti-KAS-seq* (Fig. 1d and Supplementary Fig. 1h,i). Moreover, *Opti-KAS-seq* mapped a larger fraction of sequencing reads to promoters, distal elements, and other genomic features than KAS-seq (Supplementary Fig. 1j). To validate our findings from HEK293T cells, we extended our analyses to E14-mESCs. The results consistently demonstrated the superior efficacy of *Opti-KAS-seq* in capturing ssDNA across the genome (Supplementary Fig. 2). We next applied *Opti-KAS-seq* to a variety of mouse tissues, including mouse heart, lung, and spleen, which were challenging for conventional KAS-seq. In these tissues, *Opti-KAS-seq* exhibited high ssDNA capture efficiency (Fig. 1f, g and Supplementary Fig. 3). Taken together, these results highlight the advantages of *Opti-KAS-seq* over the conventional KAS-seq, particularly in improving ssDNA capture efficiency and expanding its applicability to previously challenging sample types.

KAS-ATAC-seq simultaneously reveals chromatin accessibility and transcriptional activity of CREs

ATAC-seq detects accessible chromatin loci but it does not reveal transcription activity^{39,40}. We envisioned that integrating ATAC-seq with KAS-seq would enable us to selectively capture ssDNA from CREs, thereby reflecting active transcription. This strategy aims to streamline the categorization of CREs and exclude ssDNA signals associated with non-regulatory regions commonly observed in KAS-seq data³⁰. Our previous attempts have led to modest signal intensities³⁰. By taking advantage of the enhanced ssDNA capture activity of *Opti-KAS-seq*, we further developed KAS-ATAC-seq that enables a comprehensive assessment of transcriptional activity of CREs in accessible chromatin regions. The integration of optimized N₃-kethoxal-assisted ssDNA labeling with Tn5 transposase-mediated accessible chromatin detection enables intricate probing of transcriptional activities within CREs by capturing ssDNA at ATAC-seq peaks (Fig. 2a). The tagmentation step also simplifies library construction and allows application of KAS-ATAC-seq to samples with limited DNA input.

KAS-ATAC-seq in mouse embryonic stem cells (mESCs) demonstrated high reproducibility between replicates, particularly in the characterization of proximal ($n = 11,522$, $R = 0.81$, $p < 2.2e-16$) and distal ($n = 25,561$, $R = 0.78$, $p < 2.2e-16$) KAS-ATAC-seq peaks (Fig. 2b). Through comparative analysis with ATAC-seq, our scatterplot investigations uncovered a detailed landscape of CREs. The majority of KAS-ATAC-seq peaks closely mirrored those detected in ATAC-seq. Furthermore, certain ATAC-seq peaks notably lacked enrichment of ssDNA signals (Fig. 2c), emphasizing the refined specificity of KAS-ATAC-seq in identifying CREs in transcription. We observed three

distinct patterns: (1) CREs showing consistent signal intensities across both methods (fully ssDNA); (2) CREs with reduced KAS-ATAC-seq read densities compared to those of ATAC-seq (partially ssDNA at accessible CREs); and (3) CREs with complete absence of KAS-ATAC-seq peaks but clear ATAC-seq peaks (fully dsDNA at accessible CREs). These observations underscore the capability of KAS-ATAC-seq to offer a more nuanced perspective on transcription across CREs that are accessible based on ATAC-seq alone (Fig. 2d).

In our investigation of Tn5 transposase-accessible chromatin, we observed a marked difference in fragment size distribution between ATAC-seq and KAS-ATAC-seq libraries. Specifically, ATAC-seq captures more DNA fragments that contain mono-nucleosomes (~200 bp), whereas KAS-ATAC-seq captures a greater proportion of ssDNA

fragments in nucleosome-free regions (<100 bp), offering insights into the dynamics between transcription initiation and nucleosome positioning (Fig. 2e). KAS-ATAC-seq signals align mostly within the ATAC-seq spectrum, but with lower average read density than that of ATAC-seq peaks in E14-mESCs (Fig. 2f). Additionally, a substantial portion (78.5%, 11,206/14,277) of proximal ATAC-seq peaks intersect with KAS-ATAC-seq peaks, in contrast to only 45.8% (18,783/41,007) for distal ATAC-seq peaks. This indicates that KAS-ATAC-seq captures a more pronounced ssDNA presence in proximal ATAC-seq peaks compared to distal ATAC-seq peaks (Fig. 2g, h). We further extended our KAS-ATAC-seq protocol to HEK293T cells. The KAS-ATAC-seq read intensities across proximal and distal regions are consistent with the results in E14-mESCs, supporting the robustness of KAS-ATAC-seq

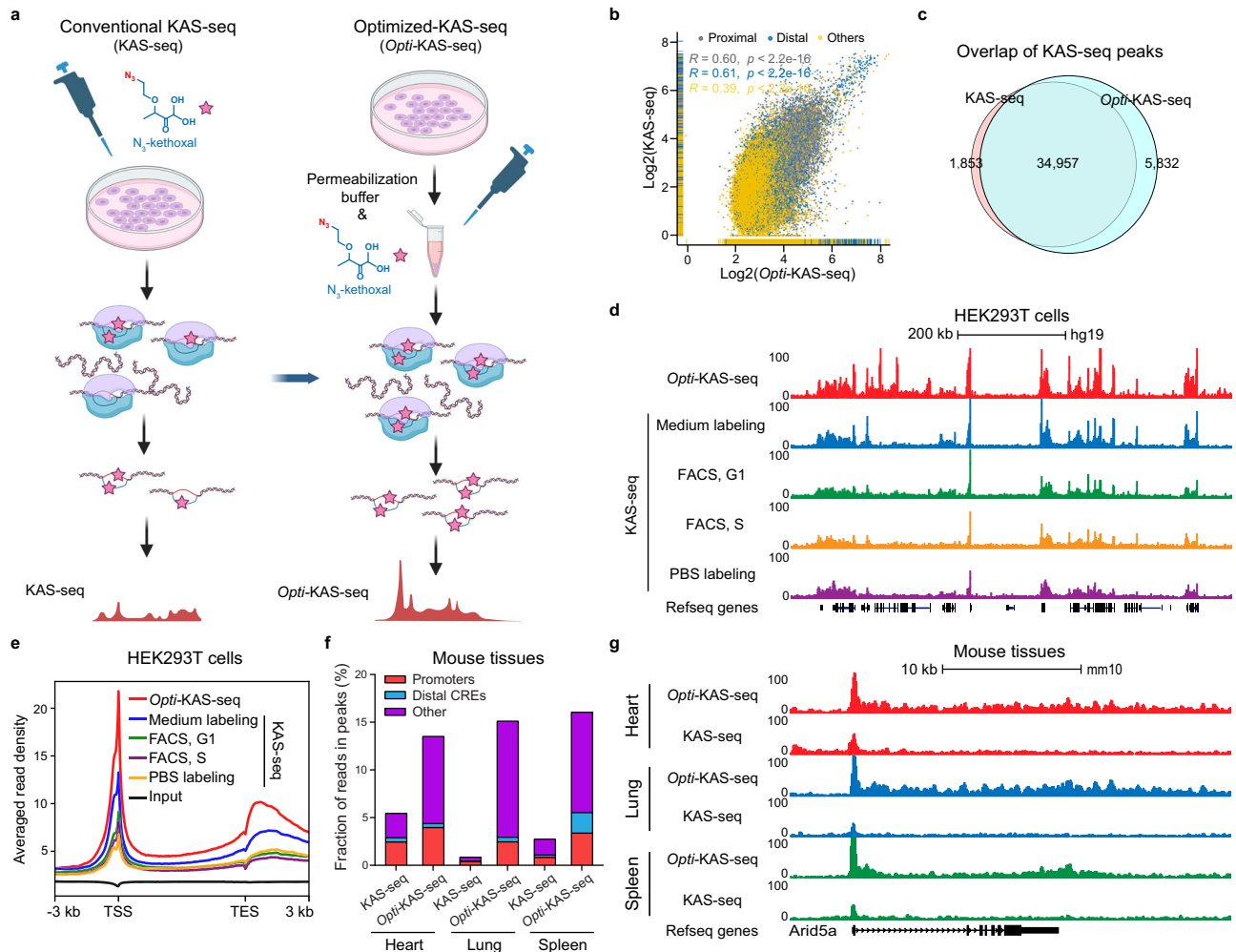


Fig. 1 | Optimization of the conventional KAS-seq protocol. **a** Schematic comparing conventional KAS-seq (KAS-seq) and optimized KAS-seq (*Opti*-KAS-seq) protocols. KAS-seq involves N_3 -kethoxal labeling of cells directly in the culture dish medium, whereas *Opti*-KAS-seq first harvests and permeabilizes the cells, then performs N_3 -kethoxal labeling in a 1.5 ml centrifuge tube. Figure 1a created with BioRender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license (<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>). **b** Scatterplot comparing KAS-seq and *Opti*-KAS-seq data in HEK293T cells across 1 kb genomic bins on merged KAS-seq and *Opti*-KAS-seq peaks. Purple dots denote proximal genomic bins ($n = 13,308$), blue dots denote distal genomic bins ($n = 6609$), and yellow dots denote other genomic bins (as labeled) ($n = 36,467$). Pearson statistical test is used to calculate the p -values. **c** Venn diagram illustrating the overlap of peaks identified using KAS-seq and *Opti*-KAS-seq data in HEK293T cells. **d** Snapshot of UCSC genome browser tracks displaying the KAS-seq data generated using the *Opti*-KAS-seq and conventional KAS-seq methods

for different types of applications across a representative region (chr1:23,572,909–24,358,036), including labeling of HEK293T cells directly in the culture dish medium (medium labeling), labeling of FACS-sorted cells at G1 (FACS, G1) and S (FACS, S) cell cycle in PBS, and labeling of harvested cells in PBS (PBS labeling). **e** Metagene profile showing the distribution of KAS-seq signals from different protocols at gene-coding regions ($n = 36,231$) in HEK293T cells, with 3 kb upstream of TSS and 3 kb downstream of TES shown. **f** Stacked bar plot showing the proportion of reads in peaks that map to promoters (± 500 bp from TSS), distal cis-regulatory elements (CREs) (> 500 bp from TSS) and other regions from KAS-seq and *Opti*-KAS-seq datasets obtained from mouse heart, lung, and spleen tissues. These values were calculated based on 30 million randomly aligned deduplicated reads. **g** Snapshot of UCSC genome browser tracks displaying KAS-seq and *Opti*-KAS-seq datasets generated in mouse heart, lung, and spleen tissues across a representative region (chr1:36,306,107–36,331,699).

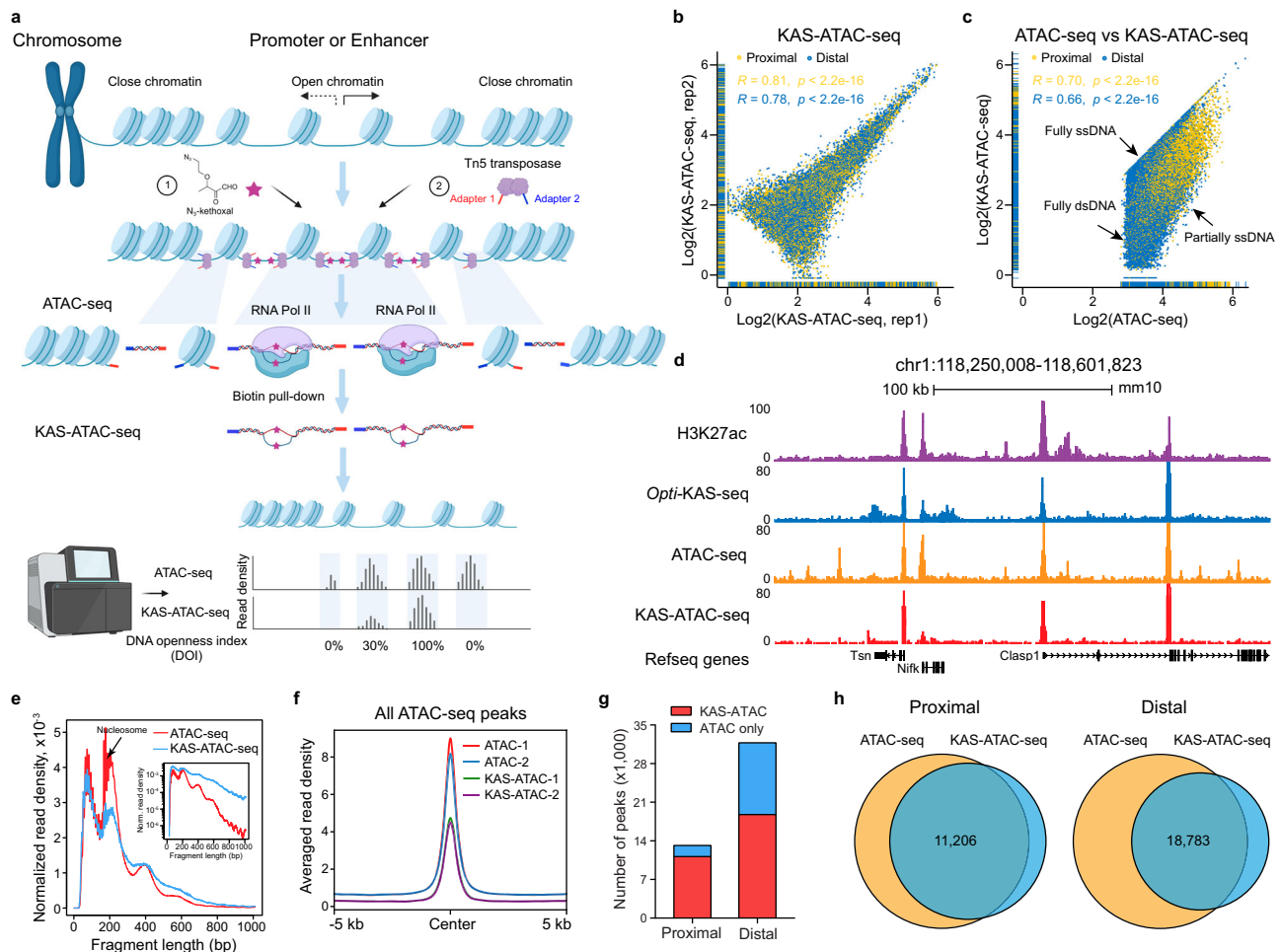


Fig. 2 | Development of KAS-ATAC-seq. **a** Schematic of library preparation workflow in the KAS-ATAC-seq protocol. Figure 2a created with BioRender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license (<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>). **b** Scatterplot illustrating the Pearson correlation between two replicates of KAS-ATAC-seq data for proximal ($n = 11,522$) and distal ($n = 25,561$) peaks in fresh mESCs. Yellow dots represent proximal peaks, blue dots represent distal peaks. Both the Pearson correlation coefficients and significance (two-sided p -values) for proximal and distal peaks are provided. **c** Scatterplot illustrating the Pearson correlation between ATAC-seq and KAS-ATAC-seq data on proximal ($n = 13,158$) and distal ATAC-seq peaks ($n = 31,768$) in fresh mESCs. Yellow dots represent proximal ATAC-seq peaks, blue dots represent distal ATAC-seq peaks. Both the Pearson correlation coefficients and significance (two-sided p -values) for proximal and distal peaks are provided. Peaks exhibiting higher signals in KAS-ATAC-seq relative

to ATAC-seq are normalized to the ATAC-seq signals. **d** Snapshot of UCSC genome browser tracks displaying H3K27ac ChIP-seq, *Opti*-KAS-seq, ATAC-seq, and KAS-ATAC-seq data on a representative region (chr1:118,250,008-118,601,823) featuring both proximal and distal KAS-ATAC-seq peaks. **e** Density plot comparing the fragment sizes of ATAC-seq and KAS-ATAC-seq data generated in fresh mESCs. The inset, a log-transformed histogram, shows clear periodicity persists to nucleosomes. **f** Metagenome profile showing the distribution of averaged ATAC-seq and KAS-ATAC-seq read density across all ATAC-seq peaks identified in mESCs, with 5 kb upstream and 5 kb downstream from the center of ATAC-seq peaks shown. **g** Stacked bar plot showing the numbers of ATAC-seq peaks, depicted by their overlap with KAS-ATAC-seq peaks in mESCs. Proximal and distal ATAC-seq peaks are displayed separately. **h** Venn diagrams showing the overlap between ATAC-seq and KAS-ATAC-seq peaks identified in mESCs, with proximal and distal ATAC-seq and KAS-ATAC-seq peaks displaying separately.

across different cell types (Supplementary Fig. 4a–c). Similar patterns in mESCs were also observed in HEK293T cells when inspecting the genomic footprints and loci-specific interactions between KAS-ATAC-seq and ATAC-seq (Supplementary Fig. 4d, e). Therefore, we establish KAS-ATAC-seq as a tool ideally suited to elucidate transcriptional activities through ssDNA capture within CREs delineated by ATAC-seq peaks.

Quantitative analysis of CRE activity using the DNA Openness Index (DOI)

We devised the DNA Openness Index (DOI), a metric specifically designed to evaluate the openness of double-stranded DNA (dsDNA). This is achieved by calculating the ratio of KAS-ATAC-seq to ATAC-seq signals across both proximal and distal CREs. The DOI thus offers insights into the transcriptional activity within these regulatory sequences, serving as a quantitative indicator of DNA transcriptional

engagement (Fig. 3a). Interestingly, a higher proportion of distal CREs (19.3%, 7897/41,007) were observed as fully ssDNA (DOI:100%) in comparison to proximal CREs (12.2%, 1736/14,277) (Fig. 3a). In the meantime, proximal CREs typically display elevated DOI values relative to distal CREs in E14-mESCs, with this difference being particularly notable in partially ssDNA CREs (Fig. 3b). This implies a more active role for proximal CREs in regulating gene expression compared to distal CREs in E14-mESCs, possibly due to their closer proximity to transcription start site (TSS). DOI values for both proximal and distal CREs remained consistent across different chromosomes (Supplementary Fig. 5a). Categorizing CREs in E14-mESCs based on their DOI values (high, medium, low, and zero) revealed that the variation of DOI is primarily dependent on KAS-ATAC-seq signals rather than ATAC-seq signals (Fig. 3c). This observation suggests that CREs with comparable levels of chromatin accessibility can exhibit markedly different transcriptional activities. Additionally, we noticed an abundance of

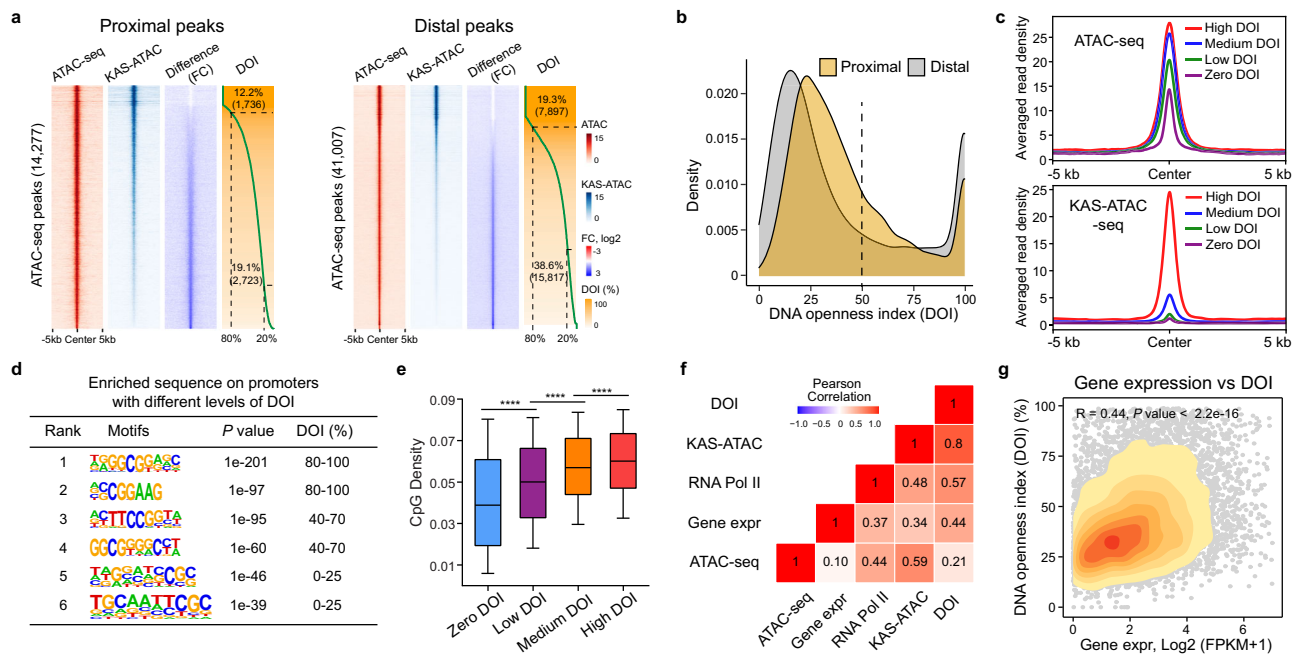


Fig. 3 | KAS-ATAC-seq quantitatively estimates the dsDNA openness in ATAC-seq peaks. **a** Heatmap showing the ATAC-seq, KAS-ATAC-seq, log₂ fold changes between KAS-ATAC-seq and ATAC-seq, and the DNA openness index (DOI) for proximal (left panel, $n = 14,277$) and distal (right panel, $n = 41,007$) ATAC-seq peaks. The DOI is determined as the proportion of KAS-ATAC-seq signal relative to the ATAC-seq signal at a specific CRE. Regions spanning 5 kb upstream and 5 kb downstream from the center of proximal and distal ATAC-seq peak were shown. **b** Density plot comparing the DNA openness index (DOI) of proximal ($n = 14,277$) and distal ($n = 41,007$) ATAC-seq peaks. **c** Metagenes profiles showing the distribution of averaged ATAC-seq (upper panel) and KAS-ATAC-seq (lower panel) read density across ATAC-seq peaks categorized by high, medium, low, zero DNA openness index (DOI) in mESCs. Regions spanning 5 kb upstream and 5 kb downstream from the center of ATAC-seq peaks are displayed. **d** Table presenting the enriched sequence features on cis-regulatory elements (CREs) associated with varying levels of the DNA openness index (DOI). The hypergeometric test was used to calculate the p -values. **e** Boxplot illustrating the CpG density within proximal

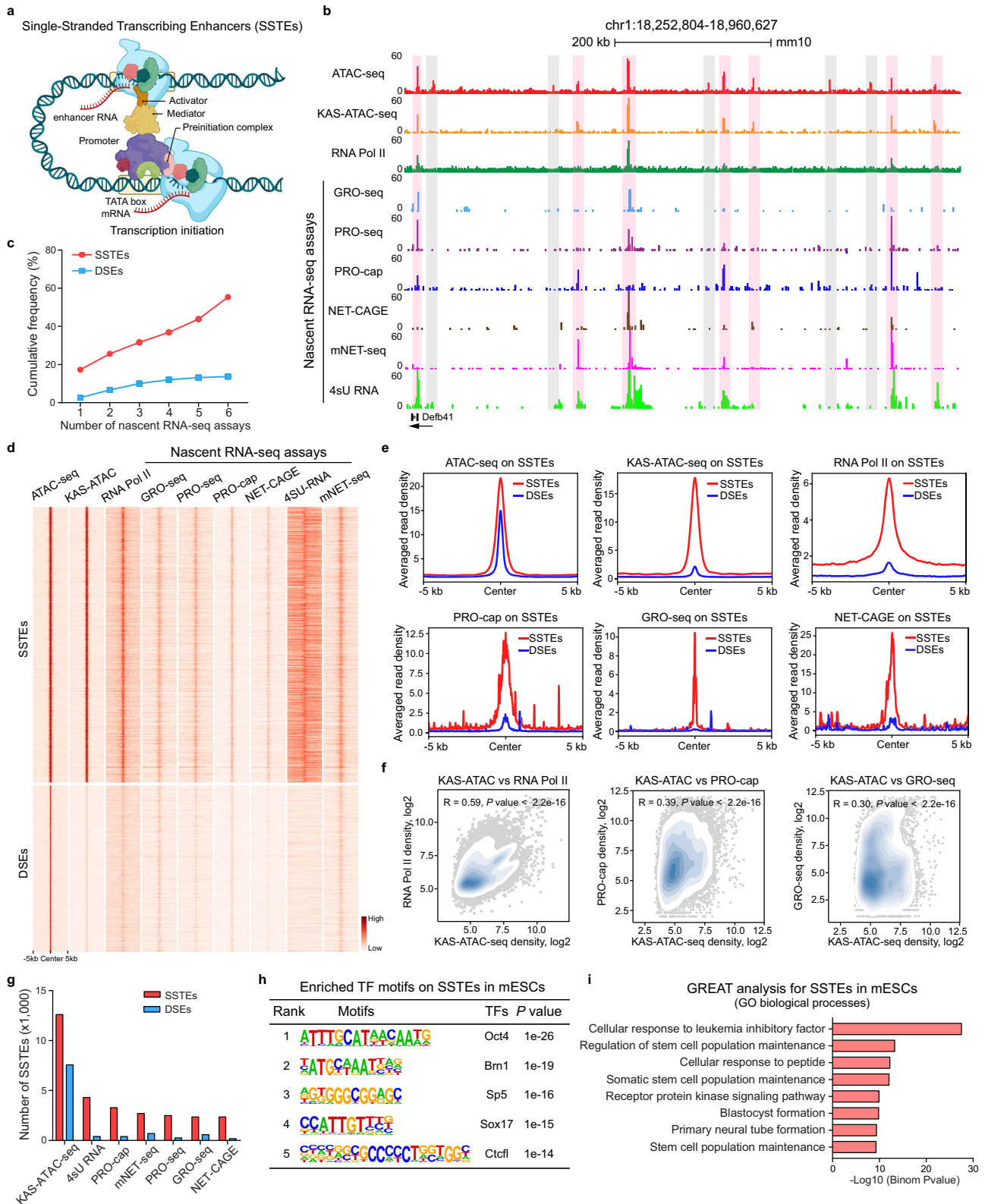
ATAC-seq peaks, categorized by high ($n = 4386$), medium ($n = 4386$), low ($n = 4386$), and zero ($n = 2001$) DNA openness index (DOI) in mESCs. The p -values were calculated using two-sided Student's t -test. P -values < 0.0001 was shown as ****. The box shows 1st quartile, median and 3rd quartile, respectively. Whiskers extend to show the range from the 10th to the 90th percentile, with data points outside this range not shown. **f** Correlation heatmap showing Pearson correlation coefficients between the DNA openness index (DOI), ATAC-seq, KAS-ATAC-seq, RNA Pol II binding, and gene expression levels of genes with ATAC-seq peaks ($n = 10,185$). The gene expression levels of proximal ATAC-seq peaks target genes are determined using bulk cell RNA-seq data. The Pearson correlation coefficients were labeled on the correlation heatmap. **g** Scatterplot showing the Pearson correlation between DNA openness index (DOI) and gene expression (Gene expr, bulk cell RNA-seq) on genes with ATAC-seq peaks ($n = 10,185$). The Pearson correlation coefficient (R) and its associated two-sided p -value are displayed at the top of the plot. Points are color-coded in orange to indicate the gene density.

CpG-rich sequence motifs specifically enriched on CREs with high-DOI values (Fig. 3d). This pattern is further supported by the finding that CREs with high-DOI values generally exhibit a higher density of CpG sites (Fig. 3e).

Our joint analysis of TF ChIP-seq and KAS-ATAC-seq datasets also identified proteins with chromatin binding affinities that show strong correlations with DOI in E14-mESCs, including the zinc finger protein (E2f1), components of the mediator complex (Med1), cyclin-dependent kinases (Cdk7), and other specific TFs such as TBP, Brd4, Chd2, and Taf3 (Supplementary Fig. 5b-c). This suggests that these TFs are more actively involved in promoting RNA Pol II-mediated transcription and play significant roles in regulating CRE transcription compared to other TFs. In addition, we generated a correlation heatmap to depict the relationships among DOI, ATAC-seq, KAS-ATAC-seq, RNA Pol II binding, and gene transcription as determined by RNA sequencing. The heatmap clearly reveals a stronger correlation between DOI metrics or KAS-ATAC-seq with gene transcription, in contrast to the relatively weak correlation observed between ATAC-seq data or active histone marks (H3K27ac and H3K4me3) and gene transcription (Fig. 3f, g and Supplementary Fig. 5d-h). This highlights the enhanced predictive power of DOI and KAS-ATAC-seq for capturing gene expression dynamics compared to ATAC-seq. In summary, our study introduces the DOI as a metric for evaluating dsDNA openness across CREs in transcription. Moreover, both KAS-ATAC-seq and DOI emerge as more accurate indicators of gene expression.

De novo mapping of Single-Stranded Transcribing Enhancers (SSTEs) using KAS-ATAC-seq

Numerous studies have established that only a subset of active enhancers is capable of enhancing gene transcription, and this capability is closely associated with the level of eRNA present on transcribed enhancers^{11,19}. Building upon our prior work that highlighted the effectiveness of KAS-seq in detecting transcriptionally active enhancers through profiling ssDNA produced by RNA polymerases³¹, we annotated a specific group of distal CREs using KAS-ATAC-seq without relying on nascent RNA-seq or active histone marks. These elements, termed as Single-Stranded Transcribing Enhancers (SSTEs) (Fig. 4a), are characterized as KAS-ATAC-seq peaks that are frequently associated with RNA Pol II binding (57.4%, 10,774/18,783) and exhibit RNA transcription (Fig. 4b). Our cumulative frequency analysis revealed that approximately 60% of SSTEs exhibit detectable eRNA transcripts in at least one of six different nascent RNA-seq assays from mESCs, including GRO-seq, PRO-seq, PRO-cap, NET-CAGE, mNET-seq, and 4sU RNA-seq, all targeting newly synthesized nascent RNA transcripts (Fig. 4c). In contrast, Double-Stranded Elements (DSEs), defined as ATAC-seq peaks not overlapping with KAS-ATAC-seq peaks, predominantly lack detectable ssDNA and eRNA signals (Fig. 4d, e and Supplementary Fig. 6a). Notably, KAS-ATAC-seq signals exhibit a stronger correlation with nascent RNA transcription than ATAC-seq signals on SSTEs (Supplementary Fig. 6b). These SSTEs revealed by KAS-ATAC-seq could therefore more accurately identify



transcriptionally active enhancers compared to traditional nascent RNA-seq assays and ATAC-seq (Fig. 4g).

Active histone marks are significantly abundant on SSTEs. Specifically, H3K4me3 and H3K36me3 are exclusively enriched on SSTEs. However, both SSTEs and DSEs exhibit substantial enrichment of H3K27ac and H3K4me1 (Supplementary Fig. 6c–e). STARR-seq (Self-Transcribing Active Regulatory Region sequencing) is a powerful

technique developed to quantify enhancer activity across the genome⁴¹. Interestingly, we found that KAS-ATAC-seq signals on SSTEs align more closely with STARR-seq data than ATAC-seq profiles (Supplementary Fig. 6f, g), indicating that KAS-ATAC-seq effectively identifies functional CREs and reflects their activities. Consensus sequence motif analysis revealed enrichment of specific transcription factors such as Oct4, Brn1, Sp5, and Sox17 on SSTEs (Fig. 4h). The GREAT

Fig. 4 | Prominent nascent RNA transcription is evident on Single-Stranded Transcribing Enhancers (SSTEs) identified by KAS-ATAC-seq. **a** Schematic illustrating transcription initiation at promoters and enhancers, as well as the long-range interactions between them. Figure 4a created with BioRender.com released under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International license (<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>). **b** Snapshot of UCSC genome browser tracks displaying ATAC-seq, KAS-ATAC-seq, RNA Pol II, and various nascent RNA-seq datasets on a representative region featuring cis-regulatory elements (CREs) defined by ATAC-seq and KAS-ATAC-seq peaks. SSTEs are defined as ATAC-seq peaks that overlap with KAS-ATAC-seq peaks and are highlighted in pink. DSEs are defined as ATAC-seq peaks without KAS-ATAC-seq overlap and are highlighted in gray. **c** Line graph depicting the cumulative frequency of SSTEs ($n = 12,601$) and DSEs ($n = 7574$) with significant nascent RNA transcription detected by various nascent RNA-seq datasets (GRO-seq, PRO-seq, PRO-cap, NET-CAGE, mNET-seq, and 4sU RNA-seq). The red dotted line represents SSTEs, and the green dotted line represents DSEs. **d** Heatmap showing the enrichment of ATAC-seq, KAS-ATAC-seq, RNA Pol II, and various nascent RNA

transcription signals (GRO-seq, PRO-seq, PRO-cap, NET-CAGE, mNET-seq, and 4sU RNA-seq) on intergenic SSTEs ($n = 12,601$, top panel) and DSEs ($n = 7574$, bottom panel). Regions spanning 5 kb upstream and 5 kb downstream from the center of CREs are shown. **e** Metagene profiles showing the averaged read density of various nascent RNA transcription datasets across SSTEs (red) and DSEs (blue). **f** Scatterplot showing the Pearson correlation between KAS-ATAC-seq data and RNA Pol II (left), PRO-cap (middle), and GRO-seq (right) data on intergenic SSTEs ($n = 12,601$). The Pearson correlation coefficient (R) and its associated two-sided p -value are displayed at the top of the plot. Points are color-coded in light blue to indicate SSTE density. **g** Grouped bar plot displaying the number of SSTEs identified by KAS-ATAC-seq data alongside various nascent RNA transcription datasets, including GRO-seq, PRO-seq, PRO-cap, NET-CAGE, mNET-seq, and 4sU RNA-seq. **h** Table presenting the enriched motifs of transcription factors (TFs) identified on SSTEs in mESCs. The hypergeometric test was used to calculate the p -values. **i** Horizontal bar plot illustrating the Gene Ontology (GO) biological processes derived from the GREAT analysis of SSTEs in mESCs. The significance of each process is calculated using the binomial test.

analysis further revealed that SSTEs are closely associated with biological processes involved in stem cell maintenance in mESCs (Fig. 4i). Conversely, DSEs, despite being enriched with ATAC-seq peaks and active histone marks, are primarily linked to signal transduction and cell differentiation (Supplementary Fig. 6h). This distinction highlights the functional importance of SSTEs in determining stem cell identity of mESCs (Fig. 4i). Additionally, SSTEs exhibit lower evolutionary conservation scores compared to DSEs and promoters, suggesting a higher level of cell-type specificity and potentially greater sequence variability (Supplementary Fig. 6i–l). Collectively, KAS-ATAC-seq enables de novo annotation of SSTEs as a subset of distal CREs. These SSTEs are distinguished by active RNA transcription, chromatin features, and specific TFs binding with functional importance.

Transcriptional dynamics of promoters and SSTEs during neural differentiation from mouse embryonic stem cells

To assess the effectiveness of KAS-ATAC-seq in characterizing CREs during a continuous differentiation process, we examined the transcriptional dynamics of CREs by conducting ATAC-seq and KAS-ATAC-seq throughout the neural differentiation process from mESCs to embryoid bodies (EBs) and neural progenitor cells (NPCs). Following the established protocol by Xiang et al.⁴², we conducted a time-course analysis employing both ATAC-seq and KAS-ATAC-seq. The neural differentiation was initiated by the removal of leukemia inhibitory factor (LIF) at Day 0, followed by treatment with retinoic acid (RA) at Day 2, resulting in the formation of NPCs by Day 8 (Fig. 5a, b). Our analysis revealed a gradual inactivation of SSTEs associated with pluripotency genes (*Oct4* and *Nanog*) and an activation of SSTEs near early neural marker genes (*Pax6* and *Neurog2*), confirming the reliability of our time-course ATAC-seq and KAS-ATAC-seq data (Supplementary Fig. 7a, b). Furthermore, we found that KAS-ATAC-seq profiles offer a more distinct separation than ATAC-seq profiles of the regulome across various stages of neural differentiation (Fig. 5c). This highlights the capability of KAS-ATAC-seq in detecting transcriptional dynamics of promoters and SSTEs during the neural differentiation of mESCs into NPCs.

Utilizing KAS-ATAC-seq, we identified 5550 ssDNA promoters and 13,110 SSTEs that undergo dynamic changes at least at one time point during the process of neural differentiation (Supplementary Fig. 7c, d). GREAT analysis of these dynamic CREs revealed that up-regulated ssDNA promoters and SSTEs are predominantly associated with neurogenesis and nervous system development (Supplementary Fig. 7e), whereas down-regulated ssDNA promoters and SSTEs are linked to stem cell maintenance and embryonic pattern specification (Supplementary Fig. 7f), reflecting the cell identity transition from mESCs to NPCs. In contrast, dynamic DSEs and non-ssDNA promoters, which showed significant changes in ATAC-seq signals but not in KAS-ATAC-

seq signals (Supplementary Fig. 7g, h), did not exhibit these specific associations, underscoring the regulatory roles of dynamic SSTEs in guiding the neural differentiation pathway.

Intriguingly, we found that a subset of promoters (C3:1140; C4:1863) and SSTEs (C4: 1151), up-regulated by retinoic acid (RA) treatment (Supplementary Data 1), exhibited an earlier peak activation in KAS-ATAC-seq at Day 4, compared to ATAC-seq at Day 6 (Fig. 5d–g and Supplementary Fig. 7i, j). This is likely due to that the KAS-ATAC-seq signal reflects real-time transcription levels by detecting ssDNA, whereas the ATAC-seq signal indicates chromatin accessibility on CREs. Additionally, we found that the binding motifs of ETS subfamily TFs and YY1 are particularly abundant in these early activated CREs (Fig. 5h). The ChIP-seq analysis of Elk4 and YY1 confirmed their significant binding enrichment on both early activated ssDNA promoters and SSTEs compared to other CREs (Fig. 5i and Supplementary Fig. 7k–m), suggesting that ETS family TFs and YY1 act as key drivers in initiating neural differentiation, especially during the critical period induced by RA treatment from Day 2 to Day 4. In summary, our findings underscore the capability of KAS-ATAC-seq to profile the transcriptional dynamics of promoters and SSTEs with higher temporal resolution than ATAC-seq during neural differentiation.

Arner et al. found that activation of enhancers occurred earlier than that of promoters during successive waves of transcriptional changes in cellular differentiation, as revealed by cap analysis of gene expression (CAGE)⁴³. In contrast, Hirabayashi et al., using native elongating transcript-cap analysis of gene expression (NET-CAGE), found that enhancers and promoters were activated simultaneously during cellular stimulation with HRG growth factor in MCF7 cells²⁶. The discrepancy could be due to CAGE being performed on total RNAs, whereas long-lived mRNAs take longer to accumulate and peak than unstable eRNAs²⁶. To explore the activation timing of promoters and SSTEs during neural differentiation induced by RA treatment, we defined activated promoters and SSTEs on Days 4 and 6 using KAS-ATAC-seq and ATAC-seq data, respectively. Our analysis indicates that promoters and SSTEs were activated simultaneously during neural differentiation induced by RA treatment, as shown by KAS-ATAC-seq results (Supplementary Fig. 8a). This finding is further supported by ATAC-seq data, which showed early simultaneous activation of promoters and SSTEs on Days 4 (Supplementary Fig. 8b). Specifically, in *Gli2* and *Ier2* genes, both KAS-ATAC-seq and ATAC-seq data showed simultaneous activation of promoters and SSTEs at Day 4 (Supplementary Fig. 8c).

To investigate the temporal relationship between dynamic CREs and gene expression during mouse neural differentiation, we integrated our KAS-ATAC-seq and ATAC-seq profiles with existing RNA sequencing (RNA-seq) data from mESCs (D0) to EBs (D2) and NPCs (D8). We focused on genes with dynamic promoters and enhancers

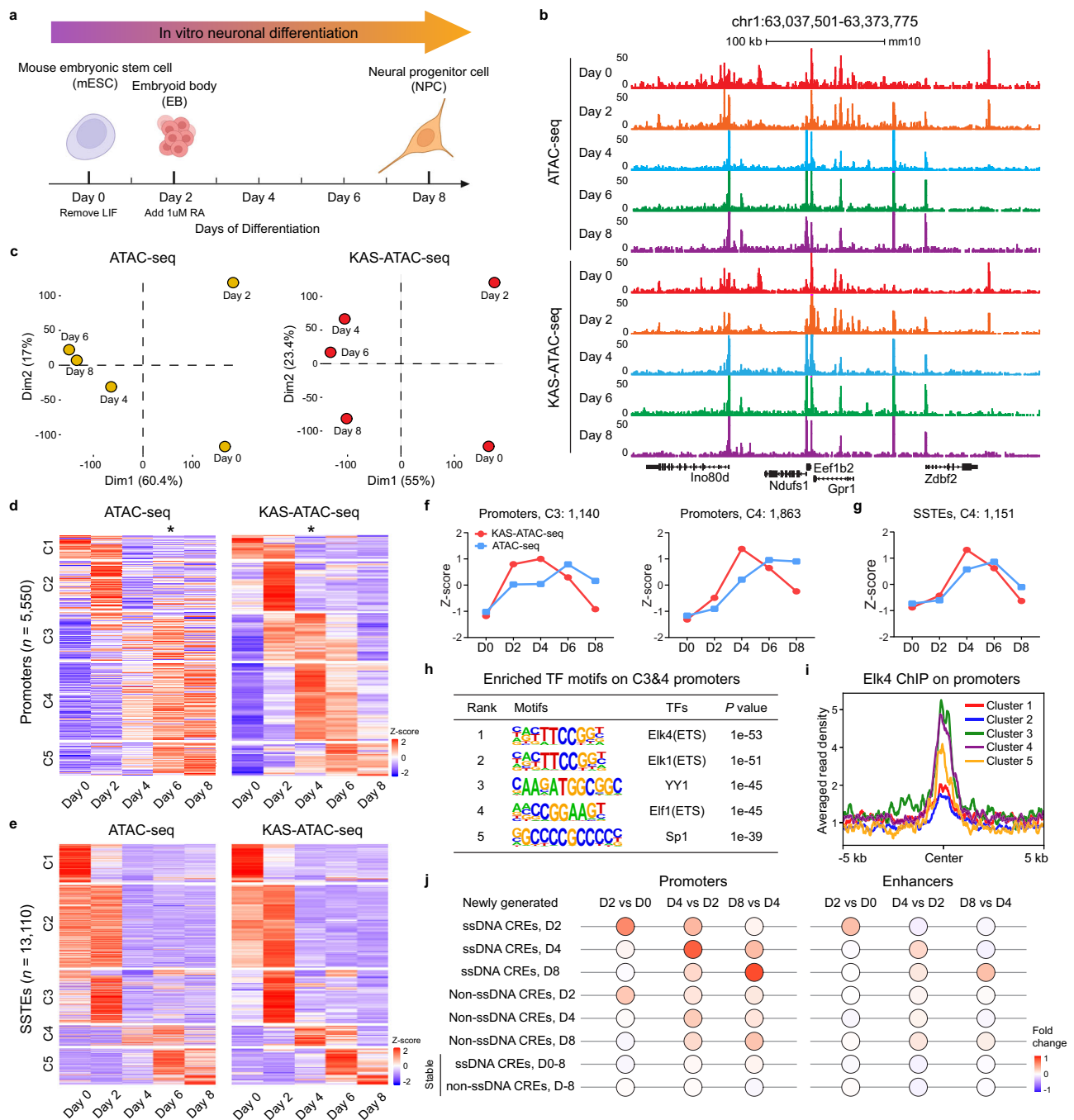


Fig. 5 | KAS-ATAC-seq accurately captures the transcriptional dynamics of ssDNA promoters and SSTEs during neural differentiation. **a** Schematic of the in vitro neuronal differentiation procedure from mouse embryonic stem cells (mESCs) to neural progenitor cells (NPCs). **b** Snapshot of UCSC genome browser tracks displaying the ATAC-seq and KAS-ATAC-seq data for a representative region. **c** Principal component analysis (PCA) plot of ATAC-seq (left panel) and KAS-ATAC-seq (right panel) data generated using cells harvested at different time points indicated during the in vitro neuronal differentiation process. **d, e** Heatmaps of time-course ATAC-seq and KAS-ATAC-seq profiles during the neural differentiation of mESCs into NPCs, using significantly dynamic CREs (up- and down-regulated), clustered on the basis of the KAS-ATAC-seq data. Each row represents a promoter (**d**) and SSTE (**e**) and each column represents a time point. The row order is the same between ATAC-seq (left panel) and KAS-ATAC-seq (right panel). Asterisks represent peak time points of promoters in the heatmaps of ATAC-seq and KAS-ATAC-seq data. **f** Line graph depicting the z-scores of C3 ($n = 1140$) and C4

($n = 1863$) promoters in the clustered heatmap of dynamic promoters (**d**). The z-scores were calculated using ATAC-seq data (blue) and KAS-ATAC-seq data (red) at various stages of the mouse neural differentiation. **g** Line graph depicting the z-scores of C4 SSTEs ($n = 1151$) in the clustered heatmap of dynamic SSTEs (**e**). **h** Table presenting the enriched motifs of TFs identified on cluster 3&4 (C3&4) immediate-early activated promoters in the clustered heatmap (**d**) during the mouse neural differentiation. The hypergeometric test was used to calculate the p -values. **i** Metagene profile showing the averaged read density of Elk4 ChIP-seq data over different promoter clusters in the clustered heatmap (**d**) during the mouse neural differentiation. **j** Bubble plots showing the fold changes of gene expression levels between consecutive stages for genes with newly generated and stable ssDNA and non-ssDNA CREs defined by KAS-ATAC-seq and ATAC-seq data during the mouse neural differentiation from mESCs to EBs and NPCs (D0 to D8). The color key, ranging from blue to red, indicates the median of fold changes of gene expression levels from low to high, respectively.

that show different ATAC-seq or KAS-ATAC-seq signals at early and late time points of neural differentiation. The normalized fold changes in expression levels of these genes were calculated across consecutive developmental stages (Fig. 5j and Supplementary Fig. 9a). We found that the expression level of genes with newly generated ssDNA promoters and SSTE in EBs and NPCs were up-regulated in these stages (Fig. 5j). Conversely, genes losing ssDNA promoters and SSTE in EBs and NPCs demonstrated down-regulation in gene expression within these stages (Supplementary Fig. 9a). Notably, ssDNA promoters displayed a more significant temporal relationship with gene expression compared to SSTE. Genes associated with dynamic promoters and enhancers lacking ssDNA showed a weak temporal regulation pattern, while genes with stable promoters and enhancers without ssDNA changes exhibited no discernible temporal regulation pattern (Fig. 5j and Supplementary Fig. 9a). Furthermore, we observed that ssDNA promoters and SSTE undergo distinct transitions during the later stages of neural differentiation, potentially transitioning into other CRE subtypes, including non-ssDNA promoters, DSEs, and beyond CRE classification (no CREs) (Supplementary Fig. 9b). By categorizing ssDNA promoters and SSTE based on their transition into other CRE subtypes, we discovered that those which did not transition into other CRE types displayed much higher KAS-ATAC-seq signal intensities (Supplementary Fig. 9c). However, this difference is less pronounced in ATAC-seq and H3K27ac profiles (Supplementary Fig. 9d–e). This suggests that the presence of ssDNA could serve as a predictive marker for the stability of CREs during neural differentiation, highlighting the potential of KAS-ATAC-seq in providing insights into the dynamic regulatory landscape of cell differentiation.

Characterization and functional implications of various SSTE subtypes

In our analysis of SSTE, we observed that H3K27ac peaks in a subset of SSTE notably extends beyond the boundaries of KAS-ATAC-seq and ATAC-seq peaks (Fig. 6a, b and Supplementary Fig. 10a–c). We hypothesize that the broadness of H3K27ac peaks could suggest two distinct subtypes of SSTE (Fig. 6a, b and Supplementary Fig. 10a–c). We found that SSTE with broadened H3K27ac peaks generally exhibit higher eRNA signals compared to those with H3K27ac peaks of comparable broadness (Fig. 6c and Supplementary Fig. 10d–g). The nuclear exosome targeting (NEXT) complex^{44,45}, essential for degrading non-coding nuclear RNA, plays a significant role in this context⁴⁶. We discovered that Zcchc8 and Rbm7, two core components of the NEXT complex⁴⁵, are enriched only in SSTE with H3K27ac peaks of comparable length (Fig. 6b, d, e). These findings suggest that eRNA transcripts from SSTE with extended H3K27ac peaks tend to be more stable, whereas eRNA transcripts from SSTE with H3K27ac peaks of comparable length appear to be more dynamic and are prone to be degraded by the NEXT complex. We have thus termed SSTE with extended H3K27ac peaks as stable-SSTE (S-SSTE, see Methods) and those with comparable H3K27ac peaks as dynamic-SSTE (D-SSTE, see Methods). S-SSTE are defined as KAS-ATAC-seq peaks that cover over half the length of H3K27ac peaks at distal CREs (Fig. 6a, b and Supplementary Fig. 10a–c), whereas D-SSTE cover less than half of these H3K27ac peaks (Fig. 6a, b and Supplementary Fig. 10a–c).

In mESCs, we identified 3247 S-SSTE, 15,536 D-SSTE, and 12,999 DSEs (Supplementary Data 2). A higher proportion of S-SSTE (76.95%, 2499/3247) exhibit pronounced eRNA signals compared to D-SSTE (38.85%, 6036/15,536) (Fig. 6c). Interestingly, ATAC-seq signals are similar in both subtypes (Fig. 6f). However, S-SSTE exhibit higher KAS-ATAC-seq read densities and a greater degree of DOI values compared to D-SSTE (Fig. 6g and Supplementary Fig. 10h). Notably, a substantial proportion of S-SSTE (60.0%, 1947/3247) overlap with super enhancers (SEs) (Fig. 6h, i and Supplementary Fig. 10i), indicating their stronger capacity for gene activation. S-SSTE also exhibit stronger STARR-seq signals and are associated with higher gene expression

levels than D-SSTE and DSEs (Fig. 6j, k). Our consensus sequence motifs analysis identified a specific set of TFs enriched on S-SSTE, including Oct4, Sox2, Nanog, YY1, and c-Myc (Fig. 6l), which are known to play crucial roles in maintaining the self-renewal and pluripotency of mESCs^{47,48}. Further analysis of ChIP-seq data for these TFs reveals markedly stronger binding signals on S-SSTE compared to D-SSTE and DSEs (Supplementary Fig. 10l), confirming the motif analysis results. In contrast, D-SSTE displayed enrichment for a distinct set of TFs (Supplementary Fig. 10j). YY1, in particular, exhibits differential binding affinities on S-SSTE, D-SSTE, and DSEs (Fig. 6m). Following transcription inhibition using Actinomycin D in mESCs⁴⁹, we found that YY1 binding on S-SSTE was noticeably reduced but less affected on D-SSTE and DSEs (Supplementary Fig. 11a–c). This decrease in YY1 binding is significantly correlated with KAS-ATAC-seq and eRNA signals on SSTE (Fig. 6n, o), indicating that eRNA transcripts and ssDNA play important roles in enhancing the chromatin binding affinity of TFs, particular YY1, on S-SSTE.

Difference in topological connectivity preference between SSTE and DSEs

Topologically Associating Domains (TADs) are fundamental in the higher-order organization of the genome and play crucial roles in transcription regulation⁵⁰. In mESCs, we noticed differences in enrichment of CTCF binding motifs between SSTE and DSEs. Specifically, CTCF is notably abundant in DSEs, which differs from the TF pattern observed in S- and D-SSTE (Fig. 7a). In our analysis of ENCODE-defined candidate CREs, including DNase I digestion sites, H3K4me3, distal enhancers, proximal enhancers, promoters, and CTCF binding sites, we discovered that a significant proportion (38.7%, 5032/12,999) of DSEs overlaps with CTCF binding sites (Fig. 7b). Our exploration of the spatial distribution of three CRE subtypes revealed significant differences in topological enrichment preference between S-SSTE and DSEs. Notably, S-SSTE are predominantly localized within TADs, while DSEs display a strong localization towards TAD boundaries (Fig. 7c). This distributional preference for DSEs is further shown by their significant enrichment of CTCF and Cohesin binding sequences, proteins indicative of insulators and TAD boundary demarcation (Fig. 7d–f and Supplementary Fig. 12a–c). While previous studies have suggested that transcriptional signals and transcription start sites (TSS) are enriched around topological boundaries⁵¹, our analysis revealed that CREs within TADs show significantly higher signals in KAS-ATAC-seq and GRO-seq compared to those located at TAD boundaries (Fig. 7e and Supplementary Fig. 12d). Interestingly, CREs located within TADs and at TAD boundaries exhibited comparable levels of ATAC-seq signals (Supplementary Fig. 12e).

To further elucidate the differences in topological connectivity preference between SSTE and DSEs, we analyzed their intersections with long-range chromatin interaction loops, as defined by RNA Pol II and CTCF ChIA-PET data^{7,52}. Our findings indicated that S-SSTE and promoters are closely associated with RNA Pol II- and YY1-mediated long-range interactions (Fig. 7g and Supplementary Fig. 12f), whereas DSEs primarily align with CTCF- and Cohesin-mediated long-range interactions typically found in insulator regions that delineate TAD boundaries (Fig. 7h and Supplementary Fig. 12g). Taken together, our findings elucidate different spatial distribution of CREs throughout the genome. This distinction is marked by S-SSTE being more prevalent in intra-TAD regions, while DSEs are inclined towards TAD boundaries. These observations suggest a distinct role in genome organization and gene regulation for S-SSTE.

Discussion

In this study, we introduce KAS-ATAC-seq, a method that combines *Opti*-KAS-seq with ATAC-seq. This approach provides a refined perspective on the dynamic transcriptional landscape of CREs and establishes connections between transcriptional change and function.

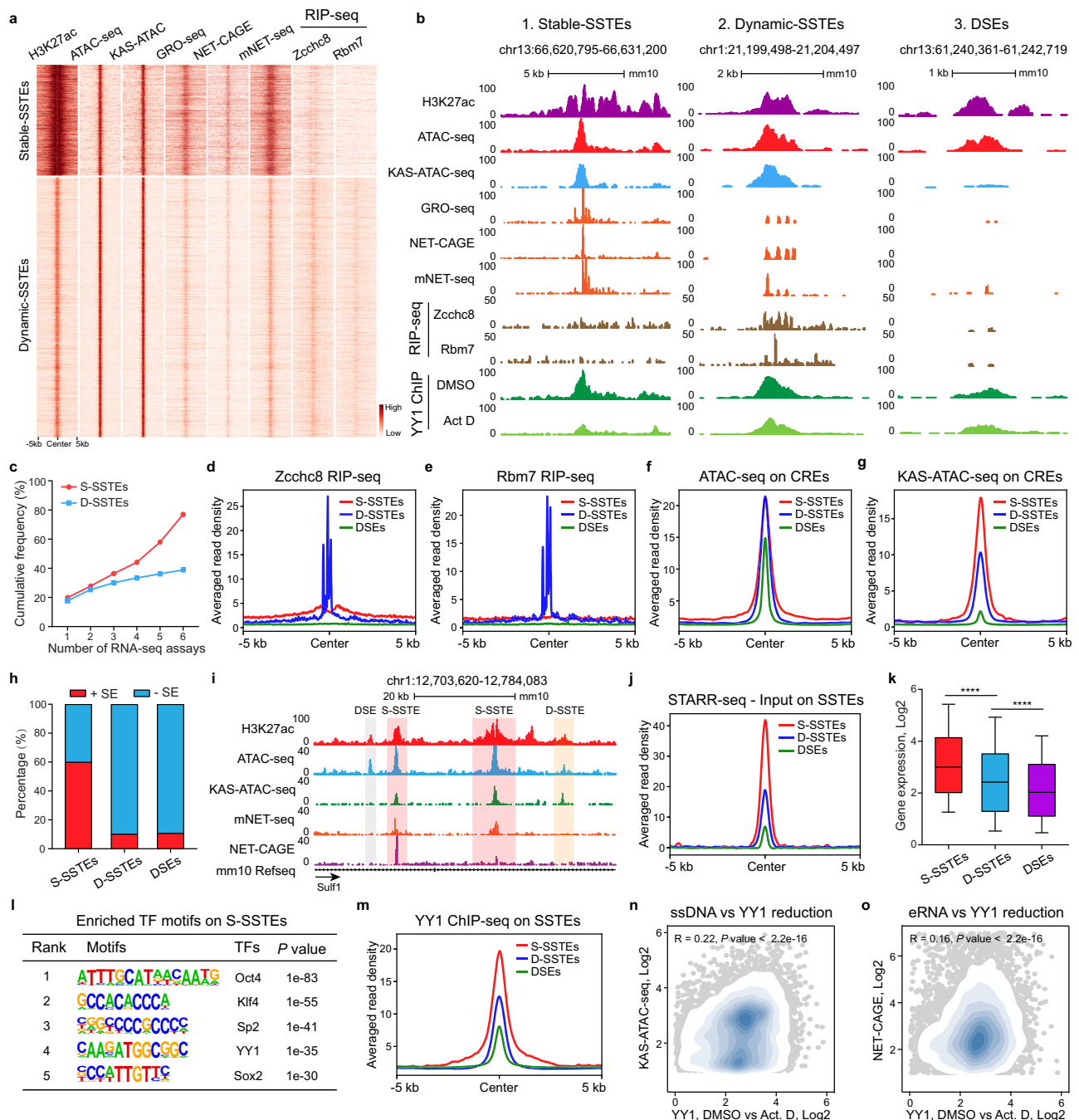


Fig. 6 | Characteristics and functional implications of two SSTE types: stable-SSTEs and dynamic-SSTEs. **a** Heatmap showing the enrichment of H3K27ac ChIP-seq, ATAC-seq, KAS-ATAC-seq, GRO-seq, NET-CAGE, mNET-seq, Zcchc8 RIP-seq, and Rbm7 RIP-seq data on intergenic stable-SSTEs ($n = 1961$) and dynamic-SSTEs ($n = 11,013$) in mESCs. **b** Snapshot of UCSC genome browser tracks displaying H3K27ac ChIP-seq, ATAC-seq, KAS-ATAC-seq, GRO-seq, NET-CAGE, mNET-seq, Zcchc8 RIP-seq, Rbm7 RIP-seq, and YY1 ChIP-seq on S-SSTEs, D-SSTEs, and DSEs. YY1 ChIP-seq data from DMSO and Actinomycin D (Act D) treated mESCs are displayed. **c** Line graph depicting the cumulative frequency of intergenic S-SSTEs ($n = 1961$) and D-SSTEs ($n = 11,013$) with detectable nascent RNAs (FPKM ≥ 0.5) detected by nascent RNA-seq assays. **d, e** Metagene profiles showing the Zcchc8 RIP-seq (**d**) and Rbm7 RIP-seq (**e**) read densities across S-SSTEs, D-SSTEs, and DSEs in mESCs. **f, g** Metagene profiles showing the ATAC-seq (**f**) and KAS-ATAC-seq (**g**) read densities across S-SSTEs, D-SSTEs, and DSEs in mESCs. **h** Stacked bar plot showing the percentages of S-SSTEs, D-SSTEs, and DSEs that overlap with and without super enhancers (SEs). **i** Snapshot of UCSC genome browser tracks

displaying H3K27ac ChIP-seq, ATAC-seq, KAS-ATAC-seq, mNET-seq, and NET-CAGE data in mESCs on a representative SE. **j** Metagene profile showing STARR-seq signals, with input subtraction, across S-SSTEs, D-SSTEs, and DSEs in mESCs. **k** Boxplot comparing the transcriptional levels of genes associated with randomly selected 3000 S-SSTEs, D-SSTEs, and DSEs. The two-sided p -values were calculated using the unpaired Student's t -test. P -values < 0.0001 was shown as ****. The box shows 1st quartile, median and 3rd quartile, respectively. Whiskers show the range from the 10th to the 90th percentile, with data points outside this range not shown. **l** Table presenting the enriched transcription factors (TFs) motifs identified on S-SSTEs in mESCs. The hypergeometric test was used to calculate p -values. **m** Metagene profiles showing YY1 ChIP-seq reads density across S-SSTEs, D-SSTEs, and DSEs in mESCs. **n, o** Scatterplots showing the Pearson correlation between KAS-ATAC-seq (**n**), NET-CAGE (**o**), and YY1 binding density reduction caused by Actinomycin D (Act D) treatment on intergenic SSTE ($n = 12,601$) in mESCs. The Pearson correlation coefficients and associated two-sided p -values are displayed. Points densities are color-coded in light blue.

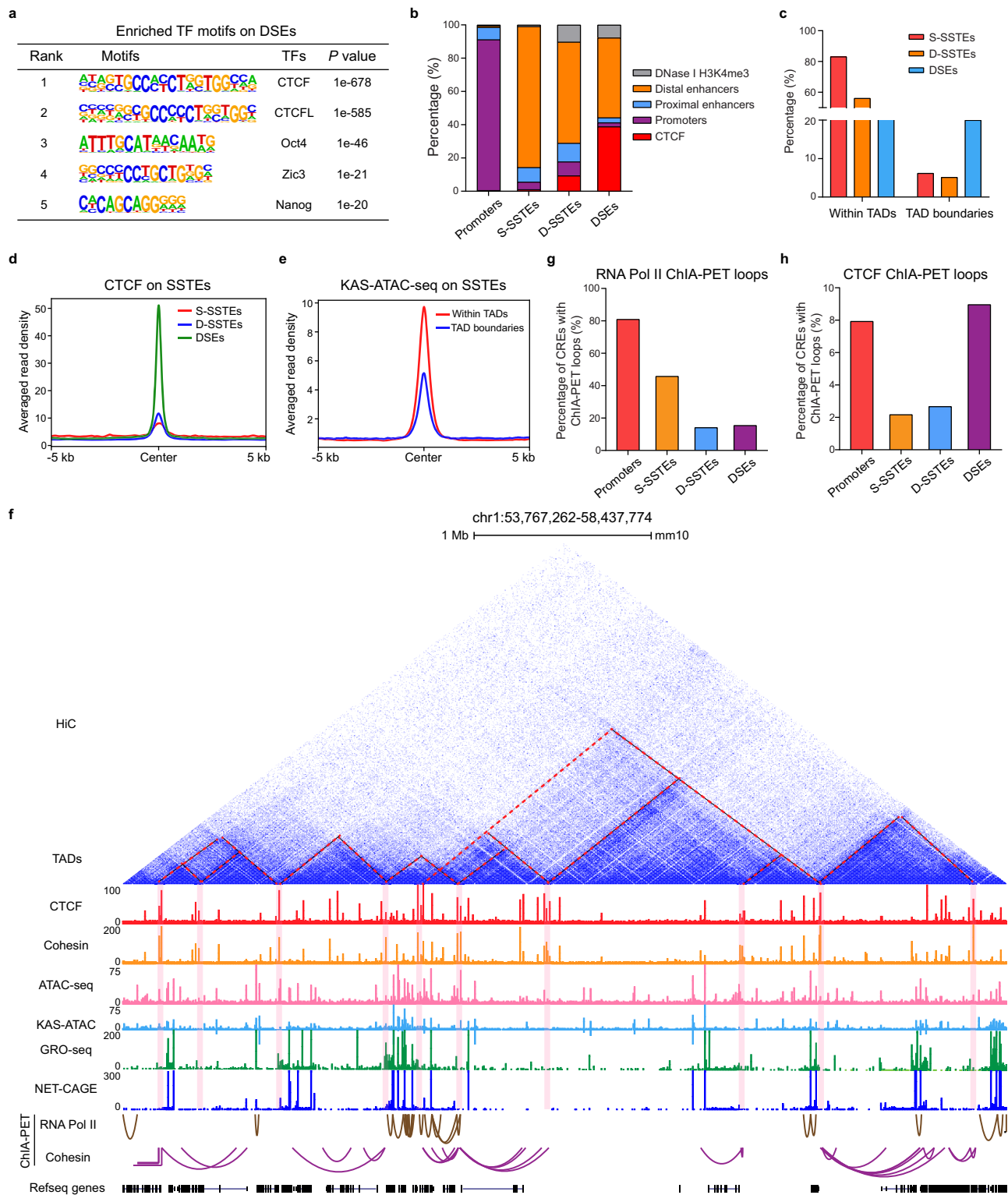


Fig. 7 | SSTEs and DSEs exhibit distinct distribution preference within higher-order chromatin structures. **a** Table presenting the enriched transcription factors (TFs) motifs identified on DSEs in mESCs. The hypergeometric test was used to calculate the *p*-values. **b** Stacked bar plot showing the distribution of different ENCODE-defined candidate cis-regulatory elements (CREs) across M-SSTEs, W-SSTEs, and DSEs in mESCs. The ENCODE-defined candidate CREs include DNase I H3K4me3, distal enhancers, proximal enhancers, promoters, and CTCF binding sites. **c** Vertical bar plot showing the percentage of S-SSTEs, D-SSTEs, and DSEs located within TADs or at TAD boundaries. **d** Metagene profiles showing the averaged read density of CTCF ChIP-seq data across S-SSTEs ($n = 3247$), D-SSTEs

($n = 15,536$), and DSEs ($n = 12,999$) in mESCs. **e** Metagene profiles showing the averaged read density of KAS-ATAC-seq data across SSTEs within TADs or at TAD boundaries in mESCs. **f**, Snapshot of UCSC genome browser tracks showing the Hi-C-seq, CTCF ChIP-seq, Cohesin ChIP-seq, ATAC-seq, KAS-ATAC-seq, GRO-seq, NET-CAGE, RNA Pol II and Cohesin ChIA-PET data in mESCs across a representative region (chr1:53,767,262-58,437,774) with prominent topologically associating domains (TADs). TADs are marked by dashed red lines and their boundaries are highlighted in pink. **g, h** Vertical bar plot illustrating the percentages of promoters, S-SSTEs, D-SSTEs, and DSEs linked to long-range interaction loops defined by RNA Pol II (**g**) and CTCF (**h**) ChIA-PET data in mESCs.

Our method improves existing methods in terms of accuracy, sensitivity, or input material requirements and enables a more comprehensive exploration of the regulatory genome.

KAS-ATAC-seq stands out for its dual capability to simultaneously uncover chromatin accessibility and transcriptional activity of CREs. It enables the evaluation of dsDNA openness across ATAC-seq peaks through devised DNA Openness Index (DOI). Importantly, KAS-ATAC-seq and the DOI have proven to be more accurate indicators of gene expression compared to ATAC-seq and active histone marks. Furthermore, KAS-ATAC-seq facilitates the de novo annotation of Single-Stranded Transcribing Enhancers (SSTEs) as a subset of distal CREs delineated by ATAC-seq peaks. Our findings reveal that SSTEs are highly enriched with nascent RNA transcription and specific transcription factors (TFs) binding sites, playing a pivotal role in defining stem cell identity.

KAS-ATAC-seq identifies more SSTEs and exhibits greater sensitivity in defining transcribed CREs compared to nascent RNA-based assays. This is partly due to that the eRNAs on transcribed CREs can be quickly degraded. Notably, KAS-ATAC-seq also offers a substantial advantage in characterizing transcriptionally active and functional enhancers over ATAC-seq and active histone marks, which tends to identify a considerable number of CREs as insulators or other non-transcribed CREs lacking enhancer activity. This distinction is underscored by the marked enrichment of CTCF and Cohesin binding on DSEs as opposed to SSTEs. Furthermore, SSTEs may offer enhanced sensitivity in predicting functional promoter-enhancer interaction loops compared to ATAC-seq and ChIP-seq peaks of active histone marks (H3K27ac and H3K4me1). Applying KAS-ATAC-seq to the neural differentiation of mESCs into NPCs has revealed intricate details of the transcriptional dynamics of CREs involved in this process. We discovered that KAS-ATAC-seq is more capable of detecting early transcriptional changes of CREs compared to conventional ATAC-seq. The enhanced resolution offered by KAS-ATAC-seq is likely attributable to its dual capability to simultaneously reveal chromatin accessibility and transcriptional activity of CREs. The quicker the transcriptional activation of CREs and their corresponding genes, the greater the time lag between the peaks observed in KAS-ATAC-seq and ATAC-seq. This also aligns with findings from published studies, which indicate that integrating RNA-seq with ATAC-seq resolves differences better than ATAC-seq alone⁵³.

Accurate identification of TAD boundaries presents several challenges, including the need to distinguish true boundaries from noise, which requires sophisticated computational methods and careful interpretation. We believe integrating KAS-ATAC-seq and ATAC-seq data with Hi-C data offers one solution for TAD boundary identification by eliminating false positives. Additionally, the combination of these datasets further enriches our understanding of genomic architecture and its impact on transcriptional regulation.

KAS-ATAC-seq also has limitations. For example, the resolution of KAS-ATAC-seq is inherently linked to the quality of chromatin accessibility and ssDNA labeling, which might be influenced by the chromatin state and cellular context. Potential biases introduced during these processes could affect the accuracy and reproducibility of the results. It is also crucial to consider that while KAS-ATAC-seq offers enhanced sensitivity and specificity, it may not capture all aspects of chromatin dynamics, and thus should ideally be used in conjunction with other methods to provide a comprehensive view. Despite these potential challenges, the ability of KAS-ATAC-seq to provide a detailed and quantitative view of transcriptional regulation opens up possibilities for understanding complex biological processes.

Methods

Ethical statement

Male B6 mice were purchased from the Jackson Laboratory (catalog no. C57BL/6J). All mice were used at 6–12 weeks of age. Mice were housed

under pathogen-free conditions per the National Institutes of Health (NIH) Guide for the Care and Use of Laboratory Animals. Mice were maintained a 12-hour light/12-hour dark cycle, with an ambient temperature of 22 °C (± 2 °C) and relative humidity of 50–60%. These conditions were carefully monitored to ensure a consistent and suitable environment for the animals throughout the duration of the study. All animal care and experiments were approved by the University of Chicago Institutional Animal Care and Use Committee (IACUC) and are compliant with all relevant ethical regulations regarding animal research.

Cell culture

HEK293T cells were purchased from ATCC (CRL11268) 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% CO₂. 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), 1 mM L-glutamine (Gibco), 0.1 mM β -mercaptoethanol (Gibco), 1% (v/v) nonessential amino acid stock (100 \times , Gibco), 1% penicillin/streptomycin stock (100 \times , Gibco), and 1000 U/mL LIF (Millipore). Cell lines used in this study were examined for mycoplasma contamination test using LookOut Mycoplasma PCR Kit (Sigma, MP0035).

Opti-KAS-seq

N₃-kethoxal was synthesized according to an established protocol. Cells harvested from the culture dish or Fluorescence-activated cell sorting (FACS) cells were washed with DPBS and subsequently resuspended in 50 μ L of ATAC-Resuspension Buffer (RSB, 10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl₂) containing 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin. After gentle pipetting, the suspension was incubated on ice for 3 min. The cells were then treated with 1 mL of cold ATAC-RSB containing 0.1% Tween-20 and centrifuged at 500 g for 5 min at 4 °C. Following a wash with 1 mL DPBS, the cells underwent N₃-kethoxal labeling at 37 °C with continuous agitation for 15 min. Genomic DNA was subsequently extracted using the PureLink genomic DNA mini kit. To perform the click reaction, 1 μ g genomic DNA was suspended in 100 μ L reaction mixture supplemented with 1 mM DBCO-PEG4-biotin (DMSO solution, Sigma, 760749), 25 mM K₃BO₃, and incubated at 37 °C for 1.5 h while being gently shaken at 500 rpm. Next, 5 μ L RNase A (Thermo, 12091039) 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 μ L H₂O and was fragmented to 150–350 bp size by using Bioruptor Pico at 30s-on/30s-off setting for 30 cycles; 5% of the fragmented DNA was saved as input, and the remaining 95% was used to enrich biotin-tagged DNA by incubation with 10 μ L pre-washed Dynabeads MyOne Streptavidin C1 (Thermo, 65001) at room temperature for 15 min. The beads were washed 5 times, and DNA was eluted by heating the beads in 15 μ L H₂O at 95 °C for 10 min. Eluted DNA and its corresponding input were used for library construction by using Accel-NGS Methyl-seq DNA library kit (Swift, 30024).

Opti-KAS-seq using mice tissues

Male B6 mice were purchased from the Jackson Laboratory (catalog no. C57BL/6J). All mice were used at 6–12 weeks of age. Homogenize mouse heart, lung, and spleen tissue to a cell suspension in ice-cold PBS by using a dounce homogenizer or a pellet pestle. Spin the cell suspension at 100 g for 15 s to sediment and remove potential large tissue pieces at the bottom of the tube. Spin the cell suspension at 500 g for 5 min. Remove the supernatant and save the cell pellet at the bottom of the tube for labeling. Suspend 5 million cells in 50 μ L of ATAC-Resuspension Buffer (RSB) containing 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin. After gentle pipetting, the suspension was incubated on ice for 3 min. The cells were then treated with 1 mL of

cold ATAC-RSB containing 0.1% Tween-20 and centrifuged at 500 g for 5 min at 4 °C. Following a wash with 1 mL DPBS, the cells underwent N₃-kethoxal labeling at 37 °C with continuous agitation for 15 min. Isolate total DNA from cells by using PureLink genomic DNA mini kit (Thermo K182001). Elute DNA by using 50 µL 25 mM K3BO3 (pH 7.0). Perform biotinylation and purification, enrichment of N₃-kethoxal-modified DNA, library preparation and sequencing according to the protocol for mammalian live cells.

KAS-ATAC-seq

To execute the KAS-ATAC-seq protocol, start by preparing the ATAC-Resuspension Buffer (RSB) by combining 500 µL of 1 M Tris-HCl (pH 7.4), 100 µL of 5 M NaCl, 150 µL of 1 M MgCl₂, and then add sterile water up to 50 mL. Initiate the procedure by collecting 50,000 viable cells, wash the collected cells with DPBS and centrifuge at 500 g for 5 min at 4 °C. Resuspend the cell pellet in 50 µL of ATAC-RSB containing 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin. Mix thoroughly by pipetting up and down three times and then incubate on ice for 3 min. Subsequently, dilute this lysed solution with 1 mL of cold ATAC-RSB that has 0.1% Tween-20 (but devoid of NP40 or digitonin). Mix by inverting the tube five times. Centrifuge this mixture again at 500 g for 5 min at 4 °C and discard the supernatant. Resuspend the pellet with 100 µL of 10 mM N₃-Kethoxal in PBS and then incubate at 37 °C with 500 rpm shaking for 10 min. After incubation, wash the nuclei by adding 1 mL of DPBS. Centrifuge the suspension at 500 g for 5 min at 4 °C, and subsequently remove the supernatant. Resuspend the nuclei pellet in 50 µL of the transposition mix. For the transposition mix designated for each 50 µL sample, mix 10 µL of 5x TTBL buffer, 5 µL of TTE Mix V50 transposase and 35 µL of H₂O (Vazyme TD501). Achieve a uniform mixture by pipetting up and down several times. Incubate the mixture at 37 °C for 30 min. Immediately terminate the transposition reaction by adding 250 µL (five volumes) of DNA Binding Buffer from the DNA Clean and Concentrator-5 Kit (Zymo) and mix well by pipetting or inversion. Mix thoroughly and followed by brief centrifugation to collect the contents at the bottom of the tubes. Purify the transposed DNA using the kit following the manufacturer's instructions. Next, biotinylate by click reaction, purify, and enrich the transposed DNA using Dynabeads™ MyOne™ Streptavidin C1. The DNA was first heated by 72 °C for 5 min for gap filling and then 95 °C for 10 min to remove the N₃-Kethoxal labeling, and 12 cycles of amplification (98 °C, 15 s, 60 °C, 30 s and 72 °C, 30 s) modified from the established ATAC-seq protocol⁵⁴.

ATAC-seq

The ATAC-seq experiments were conducted following the established ATAC-seq protocol (Omni-ATAC-seq) for chromatin accessibility profiling⁵⁴. Briefly, after the cells were permeabilized in ATAC-RSB with 0.1% NP40, 0.1% Tween-20, and 0.01% Digitonin for 3 min on ice, the cells were treated with 50 µL of the transposition mix containing 1x TTBL buffer, 5 µL of TTE Mix V50 transposase (Vazyme TD501) for 30 min at 37 °C. After DNA purification by DNA Clean and Concentrator-5 Kit (Zymo), the library was gap filled at 72 °C for 5 min and amplified by 8 cycles of 98 °C, 15 s, 60 °C, 30 s and 72 °C, 30 s using KAPA HiFi HotStart ReadyMix (Roche).

Induced neural differentiation of mESCs into NPCs in vitro

To initiate neural differentiation of mouse embryonic stem cells (mESCs), start by enzymatically digesting the mESCs with 0.05% trypsin to obtain a suspension of single cells. Neutralize the reaction in ESC culture medium without Leukemia Inhibitory Factor (LIF) and centrifuge to remove the supernatant. Resuspend the cells in basal differentiation medium (DMEM supplemented with 15% (v/v) stem cell qualified fetal bovine serum (Gibco), 1 mM GlutaMAX (Gibco), 0.1 mM β-mercaptoethanol (Gibco), 1% (v/v) nonessential amino acid stock (100x, Gibco), 1% penicillin/streptomycin stock) to a density of 1.5 × 10⁶

cells/mL and cultured at 37 °C for 2 days with 5% CO₂. On Day 2, transfer the embryoid bodies to a new dish using basal differentiation medium, allowing them to settle for 2 min before discarding the supernatant. Resuspend the embryoid bodies using basal differentiation medium with 1 µM Retinoic Acid (RA). On Day 4, Day 6 and Day 8, wash the cells with DPBS change the medium using basal differentiation medium containing 1 µM RA. After 8-day culture in these conditions, mESCs was differentiated into the neural progenitor cells (NPCs)^{42,55}.

Opti-KAS-seq and KAS-seq data processing

To align *Opti*-KAS-seq and KAS-seq data to the reference genome of interest, we utilized the `trim_galore` package to trim off low-quality sequence, adapter sequence, and primer sequence from single-end or paired-end raw fastq files⁵⁶. Subsequently, we employed `bowtie2` to perform the read alignment of *Opti*-KAS-seq and KAS-seq data⁵⁷. Mapped reads in sam files from the aligners are sorted and converted to bam files using `'samtools sort'`⁵⁸, which are subsequently deduplicated using `'picard MarkDuplicates'` (pair-end data) or `'samtools rmdup'` (single-end data)⁵⁸. For single-end *Opti*-KAS-seq or KAS-seq data, mapped reads were extended to 150 bp as default, irrespective of the initial sequencing data's read length. For paired-end KAS-seq data, the KAS-Analyzer incorporates a Python script, facilitating the merging of "properly paired" mapped reads into a singular interval⁵⁹. Peak calling for broad KAS-seq was executed using `epic2`⁶⁰, which identified broad peaks maintaining a false discovery rate (FDR) of 0.05 and a 1.5-fold change relative to the Input. Comprehensive quality control measures, such as library complexity metrics, inter-replicate correlation analyzes, fingerprint plots, saturation analyzes, genomic distribution of KAS-seq peaks, enrichment within gene-coding regions, and the Fraction of Reads in Peaks (FRiP) metric, were executed with the KAS-Analyzer package. Deduplicated mapped KAS-seq reads were converted to bedGraph and bigWig using `deeptools bamCoverage` and `UCSC tools`^{61,62}. Intersection analysis between *Opti*-KAS-seq and KAS-seq peaks were performed using `bedtools intersect` tools. Overlap analysis between *Opti*-KAS-seq and KAS-seq peaks was executed with the `bedtools intersect` tool⁶³. To obtain optimal ssDNA enrichment in practical applications of *Opti*-KAS-seq, we recommend a minimum sequencing depth of 40 million uniquely mapped reads.

KAS-ATAC-seq and ATAC-seq data processing

All data from KAS-ATAC-seq and ATAC-seq were produced in paired-end mode. The preprocessing, read alignment, and read alignments format conversion procedures for KAS-ATAC-seq and ATAC-seq data mirrored those for *Opti*-KAS-seq and KAS-seq. Peak calling for KAS-ATAC-seq and ATAC-seq was accomplished via `MACS2`⁶⁴, targeting sharp peaks with a false discovery rate (FDR) threshold of 0.05. Deduplicated KAS-seq reads were then transformed into bedGraph and bigWig formats using the `deeptools bamCoverage` and `UCSC tools`. Fragment size comparisons within KAS-ATAC-seq and ATAC-seq libraries were facilitated by the `ATACseqQC` Bioconductor package⁶⁵. Overlap analysis between KAS-ATAC-seq and ATAC-seq peaks was executed with the `bedtools intersect` tool. Additionally, the distribution of SSTes on the ENCODE Registry of candidate CREs in the mouse genome was also assessed with the `bedtools intersect` tool^{63,66}.

Hi-C data processing

In this study, we analyzed published Hi-C data (GSM2977176 and GSM2977177) in mESCs to elucidate the 3D chromatin organization, employing the `Juicer` toolkit for comprehensive data processing⁶⁶. Initially, paired-end Hi-C raw data were trimmed to remove adapters and low-quality sequences using `trim_galore` package⁵⁶, followed by alignment to the mouse reference genome (mm10) using the `BWA-MEM` aligner integrated within `Juicer`^{66,67}. Subsequent to alignment, duplicate reads were filtered out, and the data were binned into contact matrices to generate a multi-resolution.hic file. MboI restriction

sites was used. We then employed HICCUPS (java -Xmx10g -jar juicer_tools_1.22.01.jar hiccups -r 5000,10000 /content/mESCs_30.hic mESCs_hiccups) and Arrowhead algorithms (java -Xmx20g -jar juicer_tools_1.22.01.jar arrowhead -r 10000 -c 20 /content/mESCs_30.hic mESCs_arrowhead) on google colab, also part of the Juicer toolkit, to call significant chromatin loops and Topologically Associating Domains (TADs), respectively. The mESCs.hic files were used to visualize contact matrices in mESCs on the UCSC genome browser⁶².

Calculation of DNA Openness Index (DOI)

To quantitatively evaluate the dsDNA openness and the associated transcriptional activity throughout the regulome, we formulated the DNA Openness Index (DOI) metric, which is derived by determining the ratio of KAS-ATAC-seq signals to ATAC-seq signals within promoter and distal ATAC-seq peaks. In instances where ATAC-seq peaks exhibited more KAS-ATAC-seq signals than ATAC-seq signals, the ATAC-seq signals were normalized to match the KAS-ATAC-seq signal levels. Specifically, the KAS-ATAC-seq and ATAC-seq signals within ATAC-seq peaks were defined as the count of uniquely mapped reads that overlapped with ATAC-seq peaks by at least 50%.

The DOI can be mathematically calculated as follows:

Where:

$$\text{DOI} = \frac{\text{Signal}_{\text{KAS-ATAC-seq}}}{\text{Signal}_{\text{ATAC-seq}}}$$

- $\text{Signal}_{\text{KAS-ATAC-seq}}$ is the quantified read intensity or read count from KAS-ATAC-seq at a specific CRE.
- $\text{Signal}_{\text{ATAC-seq}}$ is the quantified read intensity or read count from ATAC-seq at the same CRE.

Identification of single-stranded transcribing enhancers (SSTEs)

We utilized the KAS-ATAC-seq technique to capture ssDNA generated in accessible chromatin regions. Upon successfully implementing KAS-ATAC-seq in HEK293T and mES cells, we identified KAS-ATAC-seq peaks using MACS2 with default settings. More precisely, we characterized those distal KAS-ATAC-seq peaks that intersected with ATAC-seq peaks as Single-Stranded Transcribing Enhancers (SSTEs). Our findings reveal that SSTEs are distinctively marked by detectable nascent RNA transcription. Notably, we only plot nascent RNA transcription signals on intergenic SSTEs and DSEs to avoid the potential elongation-related transcription signals.

Classification of SSTEs into stable-SSTEs and dynamic-SSTEs

H3K27ac ChIP-seq peaks exhibited more pronounced length variations compared to KAS-ATAC-seq and ATAC-seq peaks on SSTEs. Based on these observed differences in peak lengths, we categorized SSTEs into two distinct subtypes: stable-SSTEs (S-SSTEs) and dynamic-SSTEs (D-SSTEs). S-SSTEs are defined as SSTEs that cover over 50% of the corresponding H3K27ac peak length on SSTEs. Conversely, D-SSTEs are defined as SSTEs that span less than 50% of the H3K27ac peak length. S-SSTEs typically show significant depletion in RNA transcripts linked to Zcchc8 and Rbm7, whereas D-SSTEs display significant enrichment in RNA transcripts associated with Zcchc8 and Rbm7. Zcchc8 and Rbm7 are integral components of the Nuclear Exosome Targeting (NEXT) complex, which is crucial for the degradation of non-coding nuclear RNA^{45,46}. Notably, we only plot nascent RNA transcription signals on intergenic S-SSTEs and D-SSTEs to avoid the potential elongation-related transcription signals.

Motif analysis for cis-regulatory elements (CREs)

We analyzed consensus sequences and transcription factor binding motifs enriched in various Single-Stranded Transcribing Enhancers

(SSTEs) and DSEs using the HOMER findMotifsGenome.pl tool⁶⁸, employing parameters “-len 6,10,13,16 -p 20 -size given -mask”. Figures display *p*-values corresponding to the ‘corrected *P*’ from the output results.

Gene Ontology (GO) biological functions using GREAT analysis for SSTEs and DSEs

We employed the Genomic Regions Enrichment of Annotations Tool (GREAT) analysis to predict the potential biological functions associated with SSTEs and DSEs⁶⁹. Bed files containing genomic coordinates of SSTEs and DSEs were uploaded to the GREAT web tool (<http://great.stanford.edu/>). The tool was set to the default ‘Whole Genome’ background with mm10 genome assembly chosen based on the original sequencing data. Association rule settings were kept as default, which assigns genomic regions to nearby genes in a biologically meaningful manner. The output consists of statistically significant annotations from the Gene Ontology (GO) terms were shown.

Visualization analysis of the TFs binding, histone marks, and nascent RNA transcriptional levels on SSTEs

For the visualized evaluation of the enrichment of transcription factor (TF) binding, histone modifications, and nascent RNA transcription across different subtypes of SSTEs, we employed the deeptools suite to provide a detailed visualization⁶¹. The plotProfile function was utilized to generate metagene profiles, offering a comprehensive overview of their averaged distribution pattern on SSTEs. In parallel, the plotHeatmap function was engaged to generate heatmap plots, providing an intricate depiction of TF binding intensities, histone marks, and nascent RNA distributions across each specific SSTEs. Furthermore, the UCSC genome browser served as a cloud-based platform, facilitating the visualization of these datasets over selected representative regions. Collectively, these tools provided a profound and cohesive view of the interplay among TF binding, histone modifications, and nascent RNA transcription across diverse SSTEs subtypes.

Sequence conservation analysis for SSTEs

To investigate the sequence conservation across various SSTEs types in mESC, we employed the evolutionary conservation scores derived from phyloP (phylogenetic *p*-values) available in the PHAST package (<http://compugen.bscb.cornell.edu/phast/>). These scores are based on the multiple alignments of 59 vertebrate genomes with the mouse genome. In addition, three alternate sets of scores tailored for specific subsets of species, including Glires, Euarchontoglires, and placental mammals, were also considered. These datasets were sourced from the UCSC genome browser (<http://hgdownload.soe.ucsc.edu/goldenPath/mm10/phyloP60way/>). Subsequently, to compare the sequence conservation across different SSTEs types in mESC, we employed the plotProfile function within the deeptools suite.

Statistics & reproducibility

The specific statistical test methods used for data analysis are described in the figure legends. No statistical methods were used to pre-determine sample size. Instead, sample sizes were chosen based on standard practices and previous literature within our field. When comparing KAS-ATAC-seq with nascent RNA-seq, CREs on the gene body were excluded to avoid potential transcription signals related to elongation in nascent RNA-seq data. All KAS-seq, *Opti*-KAS-seq, ATAC-seq, and KAS-ATAC-seq experiments in this study were conducted with at least two biological replicates, unless specified otherwise.

Reporting summary

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

Data availability

The raw and processed data of *Opti*-KAS-seq, KAS-seq, KAS-ATAC-seq, and ATAC-seq experiments performed using HEK293T cells and mouse embryonic stem cells have been deposited in the Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) under the accession number: [GSE256232](https://www.ncbi.nlm.nih.gov/geo/acc/show/GSE256232). All published datasets reanalyzed in this study were summarized in Supplementary Data 3. The raw data generated in this study are provided in the Source Data file. Source data are provided with this paper.

Code availability

All the bioinformatic scripts used in this study are available at <https://github.com/Ruitulyu/KAS-Analyzer>^{59,70}. Bar plots were created with GraphPad Prism 7. Correlation heatmaps were generated using the `corrplot` package in R (version 3.6.3). Scatter plots were made using the `ggpubr` package in R (version 3.6.3), which facilitates the creation of `ggplot2`-based graphs.

References

- Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2012).
- Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).
- Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
- Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat. Rev. Genet.* **10**, 161–172 (2009).
- Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).
- Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
- Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature* **504**, 306–310 (2013).
- Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
- Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
- Lam, M. T., Li, W., Rosenfeld, M. G. & Glass, C. K. Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.* **39**, 170–182 (2014).
- Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
- 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).
- Li, W. et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516–520 (2013).
- Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).
- Natoli, G. & Andrau, J.-C. Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* **46**, 1–19 (2012).
- Boyle, A. P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
- Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci.* **107**, 21931–21936 (2010).
- Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. cell* **49**, 825–837 (2013).
- Heintzman, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
- Yao, L. et al. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nat. Biotechnol.* **40**, 1056–1065 (2022).
- Danko, C. G. et al. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. methods* **12**, 433–438 (2015).
- 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).
- Mahat, D. B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
- Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
- Kanamori-Katayama, M. et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res.* **21**, 1150–1159 (2011).
- Hirabayashi, S. et al. NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat. Genet.* **51**, 1369–1379 (2019).
- Windhager, L. et al. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res.* **22**, 2031–2042 (2012).
- Nojima, T. et al. Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).
- 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).
- Wu, T., Lyu, R., You, Q. & He, C. Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ. *Nat. methods* **17**, 515–523 (2020).
- Lyu, R. et al. KAS-seq: genome-wide sequencing of single-stranded DNA by N3-kethoxal-assisted labeling. *Nat. Protoc.* **17**, 402–420 (2022).
- Chen, X. et al. Structural visualization of transcription initiation in action. *Science* **382**, eadi5120 (2023).
- Wu, T., Lyu, R. & He, C. spKAS-seq reveals R-loop dynamics using low-input materials by detecting single-stranded DNA with strand specificity. *Sci. Adv.* **8**, eabq2166 (2022).
- Xu, C. et al. R-loop-dependent promoter-proximal termination ensures genome stability. *Nature* **621**, 610–619 (2023).
- Dou, X. et al. RBFox2 recognizes N 6-methyladenosine to suppress transcription and block myeloid leukaemia differentiation. *Nat. cell Biol.* **25**, 1359–1368 (2023).
- Espah Borujeni, A., Zhang, J., Doosthosseini, H., Nielsen, A. A. & Voigt, C. A. Genetic circuit characterization by inferring RNA polymerase movement and ribosome usage. *Nature. Communications* **11**, 5001 (2020).
- Sun, S. et al. Znhit1 controls meiotic initiation in male germ cells by coordinating with Stra8 to activate meiotic gene expression. *Dev. Cell* **57**, 901–913.e904 (2022).
- Fan, H. et al. Trans-vaccenic acid reprograms CD8+ T cells and anti-tumour immunity. *Nature* **623**, 1034–1043 (2023).
- Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17**, 1518–1552 (2022).
- Kiani, K., Sanford, E. M., Goyal, Y. & Raj, A. Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. *Mol. Syst. Biol.* **18**, e10979 (2022).

41. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
42. Mao, X. & Zhao, S. Neuronal differentiation from mouse embryonic stem cells in vitro. *JoVE (Journal of Visualized Experiments)*, e61190 (2020).
43. Arner, E. et al. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014 (2015).
44. Kilchert, C., Wittmann, S. & Vasiljeva, L. The regulation and functions of the nuclear RNA exosome complex. *Nat. Rev. Mol. Cell Biol.* **17**, 227–239 (2016).
45. Lubas, M. et al. Interaction profiling identifies the human nuclear exosome targeting complex. *Mol. cell* **43**, 624–637 (2011).
46. Wu, Y. et al. Nuclear exosome targeting complex core factor Zcchc8 regulates the degradation of LINE1 RNA in early embryos and embryonic stem cells. *Cell Rep.* **29**, 2461–2472.e2466 (2019).
47. Varlakhanova, N. V. et al. Myc maintains embryonic stem cell pluripotency and self-renewal. *Differentiation* **80**, 9–19 (2010).
48. Filipczyk, A. et al. Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nat. cell Biol.* **17**, 1235–1246 (2015).
49. Sigova, A. A. et al. Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978–981 (2015).
50. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat. Genet.* **52**, 8–16 (2020).
51. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
52. Handoko, L. et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* **43**, 630–638 (2011).
53. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e1120 (2020).
54. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. methods* **14**, 959–962 (2017).
55. Li, Y. et al. An optimized method for neuronal differentiation of embryonic stem cells in vitro. *J. Neurosci. Methods* **330**, 108486 (2020).
56. Krueger, F. Trim Galore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files. *GitHub*, <https://doi.org/10.5281/zenodo.7598955> (2015).
57. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. methods* **9**, 357–359 (2012).
58. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
59. Lyu, R. et al. KAS-Analyzer: a novel computational framework for exploring KAS-seq data. *Bioinforma. Adv.* **3**, vbad121 (2023).
60. Stovner, E. B. & Sætrom, P. epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* **35**, 4392–4393 (2019).
61. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids Res.* **44**, W160 (2016).
62. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinforma.* **14**, 144–161 (2013).
63. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
64. Gaspar, J. M. Improved peak-calling with MACS2. *BioRxiv*, 496521 (2018).
65. Ou, J. et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genom.* **19**, 1–13 (2018).
66. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
68. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. cell* **38**, 576–589 (2010).
69. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
70. Lyu, R. KAS-Analyzer: a new computational framework for exploring KAS-seq data. *GitHub*, <https://doi.org/10.5281/zenodo.6519166> (2024).

Acknowledgements

We thank the functional genomics facility at the University of Chicago for performing high-throughput sequencing. This work was supported by the US National Institutes of Health (R01HG012780 and RM1HG008935 to C.H.). C.H. is an investigator of the Howard Hughes Medical Institute. This article is subject to HHMI's Open Access to Publications policy. HHMI laboratory heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication. Additionally, the schematic diagrams depicting the *Opti*-KAS-seq and KAS-ATAC-seq experimental procedures were illustrated using the BioRender software.

Author contributions

C.H. and R.L. conceived the project. R.L. and T.W. developed the *Opti*-KAS-seq and KAS-ATAC-seq experimental procedures and performed KAS-seq, *Opti*-KAS-seq and KAS-ATAC-seq experiments with mESCs and HEK293T cells. Y.G. conducted *Opti*-KAS-seq of mouse tissues samples including liver, heart and spleen and the time-course ATAC-seq and KAS-ATAC-seq experiments for the neural differentiation from mouse embryonic stem cells (mESCs) to neural progenitor cells (NPCs) in vitro. P.W. synthesized the N₃-kethoxal compound to develop *Opti*-KAS-seq and KAS-ATAC-seq. C.Y. performed high-throughput sequencing on KAS-ATAC-seq and ATAC-seq libraries during neural differentiation. R.L. performed the bioinformatic analysis for all the high-throughput sequencing data with suggestions from C.Y. R.L. and C.H. wrote the manuscript with input and edits from all authors.

Competing interests

C.H. serves as a scientific co-founder and a member of the scientific advisory board for Aferna Green, Inc. and AccuaDX Inc. Additionally, C.H. holds equity in both Aferna Green, Inc. and Accent Therapeutics, Inc. T.W. is a shareholder of AccuraDX Inc. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-50680-8>.

Correspondence and requests for materials should be addressed to Ruitu Lyu or Chuan He.

Peer review information *Nature Communications* thanks Chenxu Zhu, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024