nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-021-00647-6

Check for updates

KAS-seq: genome-wide sequencing of single-stranded DNA by N_3-kethoxal–assisted labeling

Ruitu Lyu[1,2,7], Tong Wu[1,2,7], Allen C. Zhu[1,2,3], Diana C. West-Szymanski[4], Xiaocheng Weng[5], Mengjie Chen[4,6] and Chuan He[1,2]✉

Transcription and its dynamics are crucial for gene expression regulation. However, very few methods can directly read out transcriptional activity with low-input material and high temporal resolution. This protocol describes KAS-seq, a robust and sensitive approach for capturing genome-wide single-stranded DNA (ssDNA) profiles using N_3-kethoxal–assisted labeling. We developed N_3-kethoxal, an azido derivative of kethoxal that reacts with deoxyguanosine bases of ssDNA in live cells within 5–10 min at 37 °C, allowing the capture of dynamic changes. Downstream biotinylation of labeled DNA occurs via copper-free click chemistry. Altogether, the KAS-seq procedure involves N_3-kethoxal labeling, DNA isolation, biotinylation, fragmentation, affinity pull-down, library preparation, sequencing and bioinformatics analysis. The pre-library construction labeling and enrichment can be completed in as little as 3–4 h and is applicable to both animal tissue and as few as 1,000 cultured cells. Our recent study shows that ssDNA signals measured by KAS-seq simultaneously reveal the dynamics of transcriptionally engaged RNA polymerase (Pol) II, transcribing enhancers, RNA Pol I and Pol III activities and potentially non-canonical DNA structures with high analytical sensitivity. In addition to the experimental protocol, we also introduce here KAS-pipe, a user-friendly integrative data analysis pipeline for KAS-seq.

Introduction

Formation of single-stranded DNA (ssDNA) occurs in nearly all cellular processes involving the genome, including but not limited to transcription[1], DNA replication[2], DNA repair[3] and recombination[4]. For example, during transcription, RNA polymerases and other protein co-factors resolve DNA double helices and generate ssDNA bubbles[5]. Mapping the ssDNA generated at these transcribing regions can therefore provide a precise readout of the transcriptional activity. Similarly, polymerase (Pol) II–engaged enhancers can be readily mapped through measuring ssDNA as well.

Apart from transcription bubbles, ssDNAs are also present at non-B-form DNA structures, such as left-handed double-helical Z-DNA, extruded cruciforms, triple-helical H-DNA, G-quadruplexes and I-motifs[6–9]. Generally, these non-B-form ssDNA structures can be flexible. They can result in genome instability that may induce specific transcriptional responses[10,11]. R-loops constitute another type of ssDNA-containing structure. Formed when RNA anneals to its template DNA with the non-template DNA remaining single stranded, R-loops are reported to be associated with genome instability and certain human diseases[12,13]. R-loops also occur in telomere and centromere regions of chromosomes[14,15].

Therefore, genome-wide mapping of ssDNA is of broad interest to researchers in multiple fields because of its widespread formation in many fundamental cellular processes. However, methods for ssDNA mapping have been hampered by low signal intensity and extensive DNA decay during labeling[6,16]. Moreover, many ssDNA structures can be transient, making them more difficult to be effectively captured. Thus, the development of robust methods for sensitive and comprehensive analysis of ssDNA is desirable.

To precisely map ssDNA in cells and animal tissues, we designed and synthesized the small molecule N_3-kethoxal and developed N_3-kethoxal-assisted ssDNA sequencing (KAS-seq). N_3-kethoxal modifies guanines in single-stranded nucleic acids in live cells within 10 min[17,18].

[1]Department of Chemistry, Department of Biochemistry and Molecular Biology, Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL, USA. [2]Howard Hughes Medical Institute, The University of Chicago, Chicago, IL, USA. [3]Medical Scientist Training Program, The University of Chicago, Chicago, IL, USA. [4]Department of Medicine, The University of Chicago, Chicago, IL, USA. [5]College of Chemistry and Molecular Sciences, Hubei Province Key Laboratory of Allergy and Immunology, Wuhan University, Wuhan, Hubei, China. [6]Department of Human Genetics, The University of Chicago, Chicago, IL, USA. [7]These authors contributed equally: Ruitu Lyu, Tong Wu. ✉e-mail: chuanhe@uchicago.edu

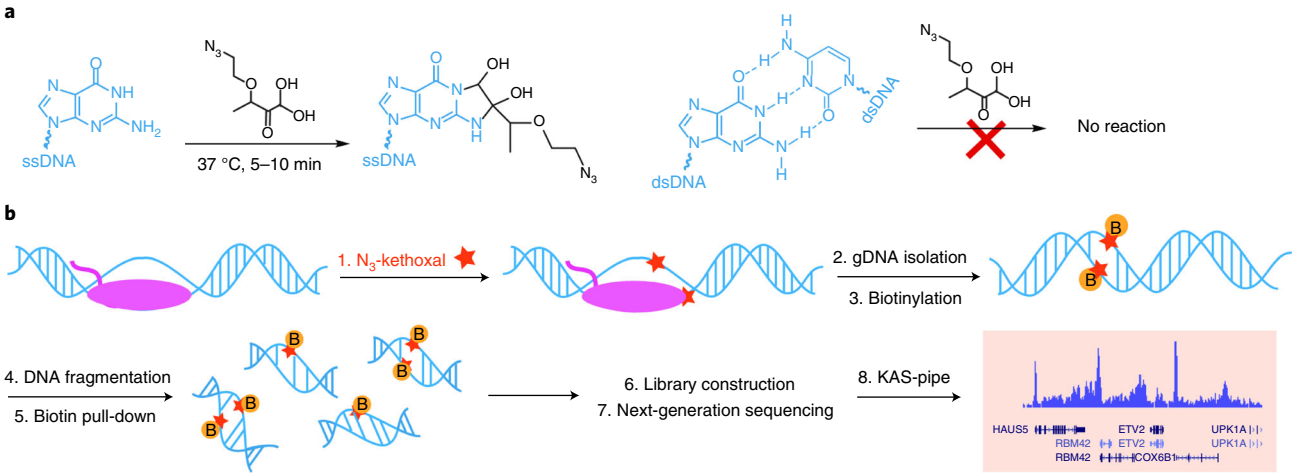

Fig. 1 | The chemical principles and workflow of KAS-seq. a, The selectivity of N_3-kethoxal for ssDNA. Left: the chemical reaction between N_3-kethoxal and exposed amine groups on guanine in ssDNA. Right: the hydrogen bonding interactions of the Watson-Crick base pairing between guanine and cytosine in double-stranded DNA (dsDNA) block the reaction between guanine and N_3-kethoxal. **b**, A schematic diagram of the KAS-seq workflow. N_3-kethoxal (red stars) modifies ssDNA in live cells (Steps 1 and 2). Most of these ssDNA regions are generated by RNA polymerases (pink ovals). Genomic DNA (gDNA) is subsequently isolated (Step 3) and biotinylated through 'click' cycloaddition with a biotin-conjugated alkyne (orange circles) (Steps 4–7). The resulting biotinylated DNA fragments can be fragmented (Step 8) and enriched by affinity capture on streptavidin beads (Steps 9–13). After library construction (Steps 14 and 15) and sequencing (Step 16), KAS-pipe is used to analyze KAS-seq data (Steps 17–26).

The azide moiety in N_3-kethoxal can then participate in an azide-alkyne click cycloaddition reaction to effectively biotinylate ssDNA regions, enabling them to be enriched through affinity pull-down. Next-generation sequencing (NGS) libraries are constructed from enriched ssDNA fragments for sequencing. Thus, KAS-seq enables rapid, sensitive and genome-wide sequencing of the dynamics of transcription mediated by RNA polymerases and enhancers containing Pol II–mediated transcription bubbles simultaneously, and KAS-seq also has the potential to map non-canonical DNA structures. KAS-seq has been optimized to work with low-input material and tissue samples. Here, we provide a detailed KAS-seq protocol for genome-wide ssDNA mapping and focus on its application on transcription activity profiling. To facilitate KAS-seq data analysis, we also introduce KAS-pipe, a user-friendly and integrative KAS-seq data analysis pipeline, in this protocol[19–26].

Overview of the KAS-seq procedure

The entire KAS-seq procedure is summarized in Fig. 1 and consists of the following stages: 1) ssDNA labeling in cells with N_3-kethoxal, 2) isolation of genomic DNA, 3) biotinylation and purification of biotinylated DNA, 4) fragmentation of genomic DNA (gDNA) by sonication or enzyme digestion, 5) enrichment of biotinylated DNA fragments using affinity capture with streptavidin beads, 6) DNA library construction using the enriched DNA fragments, 7) NGS and 8) KAS-seq data analysis using KAS-pipe[26].

The small molecule N_3-kethoxal

Kethoxal was first reported to react with and inactivate RNA viruses in the 1950s[27,28]. Since then, it has been widely used for nucleic acid footprinting[29–31]. Because kethoxal specifically reacts with single-stranded guanines at the N1 and N2 positions of the Watson-Crick interface, we reasoned that we could utilize it as a tool for single-stranded nucleic acid capture and sequencing. To accomplish this, we designed a kethoxal derivative with a bio-orthogonal azide handle to allow for downstream biotin modification and subsequent enrichment. The high labeling efficiency of N_3-kethoxal coupled with the reversibility of the kethoxal-guanine reaction under alkaline or heating conditions led us to develop a new method, called Keth-seq, for RNA secondary structure probing under mild conditions[17]. We reported both our synthetic scheme to prepare the azide-modified kethoxal (N_3-kethoxal) along with Keth-seq specific labeling of single-stranded RNA together[17]. Next, with the understanding that formation of Watson-Crick base-pairing in double-stranded DNA (dsDNA) blocks kethoxal labeling, we took advantage of the similarly rapid reaction between N_3-kethoxal and guanines in ssDNA to develop KAS-seq for genome-wide ssDNA capture and sequencing in situ[18].

Applications of KAS-seq

The KAS-seq protocol for bulk cells enables capture of ssDNA genome-wide using 1–5 million live or frozen cells, and a low-input version of the KAS-seq protocol has been developed for as few as 1,000 cultured live cells (Box 1). In addition, the bulk KAS-seq protocol is also optimized for animal tissues (Box 2). As we discuss below, KAS-seq is less technically demanding and time consuming and requires less starting material than other previously published methods, making it feasible to study many ssDNA-involved processes. As a DNA-based method to study gene transcription, KAS-seq does not require handling of degradation-prone RNA analytes, making it a suitable complement to RNA-seq. KAS-seq is particularly powerful for probing global transcription and enhancer activities simultaneously, thereby enabling sensitive detection of transient transcription dynamics in any genome of interest. In addition, KAS-seq allows researchers to detect both transient effects caused by chemical or environmental stimulation and alterations in steady-state transcription caused by genetic perturbations.

Advantages of KAS-seq and comparison with other methods

Much effort has been put into developing approaches to profile ssDNA formation genome-wide within live cells. To characterize ssDNA profiles during meiotic recombination initiation and repair processes, several laboratories developed RPA chromatin immunoprecipitation sequencing (RPA-seq)[32], RAD51 chromatin immunoprecipitation sequencing (RAD51-seq), single-stranded DNA sequencing (SSDS) and SPO11-oligo-seq[33–36]. Most of these approaches are based on chromatin immunoprecipitation sequencing (ChIP-seq) of particular ssDNA-associated proteins (RAD51, DMC1 and SPO11) at double-strand breaks, but these methods cannot map all ssDNA regions across the whole genome[35]. Similarly, ssDNA-associated protein Rad52 ChIP-seq was demonstrated to map genomic hotspots of DNA damage[37]. Remarkably, KAS-seq provides more comprehensive genome-wide ssDNA mapping data than any of the aforementioned immunoprecipitation-based methods.

Potassium permanganate ($KMnO_4$) was previously reported to preferentially oxidize single-stranded thymidine residues. In 2013, Kouzine et al. applied $KMnO_4$ in combination with nuclease S1 to map genome-wide sites of ssDNA in a method termed ssDNA-seq and revealed Pol-II–induced promoter melting during lymphocyte activation[16]. This method detects strong ssDNA signals enriched around transcriptional start sites (TSSs) but shows low sensitivity in detecting relatively weak ssDNA signals (Fig. 2a–c). Compared to the number of cells needed to perform KAS-seq (1×10^3–5×10^6 million cells), the permanganate ssDNA-seq method requires tens of millions of cells. In addition, because the permanganate method has a relatively low signal-to-noise ratio for weak ssDNA signals, its application in many biological systems is limited (Fig. 2d).

KAS-seq is especially powerful in transcriptional activity profiling. Transcriptional activity profiling has traditionally been achieved by nuclear run-on assays[38], metabolic labeling of nascent RNA[39,40] and ChIP-based methods[41]. Nuclear run-on assays are usually high resolution and can identify transcriptional directionality, but the procedure can be labor intensive and time consuming. These methods also require a large amount of starting material. Sequencing of metabolically labeled RNAs can indicate transcriptional activity, but the levels of RNA transcripts can also be affected by other steps of RNA metabolism, such as RNA decay. Pol II ChIP-seq usually requires millions of cells and can reflect only the binding of Pol II, not its activity (Fig. 3a–e). In 2013, the assay for transposase-accessible chromatin using sequencing (ATAC-seq) was developed to assess genome-wide chromatin accessibility and transcription factor binding[42]. Although extremely powerful in revealing chromatin state, chromatin accessibility does not always necessarily indicate active transcription. Moreover, ATAC-seq signals are mainly enriched at promoters and enhancers and do not reflect the activity of Pol II elongation (Fig. 3f–i).

In contrast to the aforementioned methods, KAS-seq relies on a rapid N_3-kethoxal–guanine reaction under mild conditions, which can map ssDNA genome-wide in an antibody-independent manner. Because of the high reaction efficiency between guanine and N_3-kethoxal, KAS-seq has higher sensitivity even at regions with weak and broad ssDNA signals, such as at gene bodies and transcription terminal regions (Fig. 2b–d). The whole procedure is simple and can also be applied to low-input materials and tissues.

Limitations of KAS-seq

Lack of differentiation between RNA polymerase–induced ssDNA and other types of ssDNA formation

ssDNA is known to be formed in many fundamental cellular processes. Because N_3-kethoxal can sensitively react with deoxyguanosine bases of ssDNA in a genome-wide fashion, distinguishing the

Box 1 | Low-input KAS-seq

Procedure

N3-kethoxal labeling and DNA isolation ● Timing 1 h

1 Label 1,000; 5,000; or 10,000 cells with N_3-kethoxal using the same labeling protocol as that for bulk mammalian cell cultures (Step 1 in the main Procedure).

2 Isolate gDNA from cells using the Quick-DNA Microprep Plus kit (Zymo, cat. no. D4074). Follow the manufacturer's protocol with the following exception: elute DNA with 45 μl of 25 mM K_3BO_3 (pH 7.0).

▲ **CRITICAL STEP** Avoid overexposure of DNA to guanidium-containing buffers for longer than 30 min during the purification step.

Biotinylation and purification ● Timing 1–2 h

3 Resuspend gDNA in 42.5 μl of 25 mM K_3BO_3 (pH 7.0) and add 5 μl of 10× PBS and 2.5 μl of 20 mM DBCO-PEG_4-biotin DMSO solution. Incubate the mixture at 37 °C for 1.5 h with shaking at 500 rpm to facilitate the click cycloaddition reaction.

4 Add 2.5 μl of RNase A (Thermo, cat. no. 12091039) to the reaction mixture. Incubate the mixture at 37 °C for 5 min.

5 Purify DNA from the reaction mixture by DNA Clean & Concentrator kit (Zymo, cat. no. D4043). Follow the manufacturer's protocol with the following exception: elute DNA with 25 μl of 25 mM K_3BO_3 (pH 7.0).

▲ **CRITICAL STEP** Avoid overexposure of DNA to guanidium-containing buffers for longer than 30 min during the purification step.

Tagmentation ● Timing 1 h

6 Perform tagmentation on biotinylated DNA using the Illumina Tagmentation DNA enzyme and buffer kit (Illumina, cat. no. 20034197). The exact amount of enzyme varies depending on the number of cells and should be determined in advance by users. Successful tagmentation should result in sufficient DNA fragments smaller than 1 kb as determined by gel electrophoresis or fragment analyzer. Our suggested amounts of enzyme are as follows: 1.5 μl of enzyme for 1,000 cells; 2 μl of enzyme for 5,000 cells; 5 μl of enzyme for 10,000 cells.

7 Incubate the tagmentation mixture at 37 °C for 30 min with shaking at 500 rpm.

▲ **CRITICAL STEP** If users would like to use homemade tagmentation enzymes or enzymes from a different commercial source, the tagmentation protocol will need to be optimized accordingly by determining fragmented DNA size using gel electrophoresis or a fragment analyzer.

? TROUBLESHOOTING

8 Purify DNA from the reaction mixture with the DNA Clean & Concentrator kit (Zymo, cat. no. D4043). Follow the manufacturer's protocol with the following exception: elute DNA with 60 μl of 25 mM K_3BO_3 (pH 7.0).

▲ **CRITICAL STEP** Avoid overexposure of DNA to guanidium-containing buffers for longer than 30 min during the purification step.

9 Save 5 μl of the tagmented DNA on ice as an input control. Use the remaining 50 μl for enrichment.

Enrichment of N_3-kethoxal labeled DNA ● Timing 1–2 h

10 Wash 5 μl of Dynabeads MyOne streptavidin C1 three times with 25 μl of 1× B&W buffer. Re-suspend prewashed beads in 50 μl of 2× B&W buffer.

11 Mix the beads with 50 μl of tagmented DNA from Step 9. Incubate the mixture at room temperature for 15 min with gentle rotation. After incubation, place the mixture on a magnetic rack to remove the supernatant and wash beads five times with 100 μl of 1× B&W buffer.

PCR amplification and library sequencing ● Timing 3 h (for PCR amplification)

12 Adjust the volume of input to 20 μl by adding 15 μl of nuclease-free water to the 5-μL input saved in Step 9. Re-suspend the washed beads in 20 μl of nuclease-free water.

13 Prepare the PCR reaction mix as follows:

Reagent	Amount (μl)
Input DNA or bead suspension that includes enriched DNA	20
i5 index primer (Illumina, cat. no. 20027213)	2.5
i7 index primer (Illumina, cat. no. 20027213)	2.5
NEBNext Ultra II Q5 Master Mix (NEB, cat. no. M0544S)	25

14 Perform the PCR reaction in a thermocycler as follows:

5 min, 72 °C
10 min, 95 °C
15 cycles: 10 s, 98 °C; 30 s, 60 °C; 1 min, 72 °C
4 °C hold

15 Remove the beads from the enriched samples using a magnetic rack. Purify the library using a MinElute PCR purification kit (Qiagen, cat. no. 28804). The library size should range from 200 to 600 bp.

▲ **CRITICAL STEP** Avoid overexposure of DNA to guanidium-containing buffers for longer than 30 min during the purification step.

? TROUBLESHOOTING

16 Sequence libraries on Illumina platforms using single-end or paired-end mode with 40 million reads per sample. Process the KAS-seq data using KAS-pipe as described in the 'Data processing and analysis' section of the main Procedure (Steps 17–26).

various sources of ssDNA formation is difficult with KAS-seq. Apart from the ssDNA bubbles induced by three RNA polymerases, ssDNA hotspots detected by KAS-seq may also be associated with DNA damage, non-canonical ssDNA structures, R-loops and DNA replication. We indeed

Box 2 | KAS-seq with mouse liver tissue

Procedure

N3-kethoxal labeling and DNA isolation ● **Timing 1.5 h**

1 Homogenize 5 mg of mouse liver tissue to a cell suspension in ice-cold PBS using a Dounce homogenizer or a pellet pestle.
2 Centrifuge the cell suspension at 100*g* for 15 s to sediment and separate potential large tissue pieces at the bottom of the tube. Transfer the supernatant to a clean microcentrifuge tube.
3 Centrifuge the cell suspension (supernatant from step 2) at 800*g* for 5 min. Remove the supernatant and save the cell pellet at the bottom of the tube for labeling.
4 Count and suspend 1–5 million cells in 1 ml of labeling medium for 10 min at 37 °C (refer to Step 1 in the main Procedure).
5 Isolate total DNA from cells using a PureLink Genomic DNA Mini kit. Elute DNA with 50 µl of 25 mM K_3BO_3 (pH 7.0).
▲ **CRITICAL STEP** Avoid overexposure of DNA to guanidium-containing buffers for longer than 30 min during the purification step.
■ **PAUSE POINT** The eluted DNA can be stored at −20 °C for future use.
6 Perform biotinylation and purification, enrichment of N_3-kethoxal–modified DNA, library preparation and sequencing and data processing and analysis according to Steps 4–26 in the main Procedure.

observed KAS-seq peaks overlap with a number of predicted non-canonical ssDNA structures, and additional experiments (e.g., validation based on a plasmid system with design-in non-canonical ssDNA structures, including cruciform, quadruplex, H-DNA, Z-DNA and hairpin structures[9]) need to be done to confirm this observation in the future.

In addition, KAS-seq is expected to efficiently detect R-loop structures by labeling the exposed ssDNA region of the DNA/RNA hybrid triplex[43]. Although many R-loops are co-transcriptionally mediated by nascent RNA and correlate with transcription level, R-loop–associated ssDNA structures may complicate the quantification of transcription activity, especially at regions near transcription termination sites. Strand-specific KAS-seq may distinguish transcription bubbles and R-loops in the genome.

Other factors may also contribute to the observed KAS-seq profile. For instance, although N_3-kethoxal is very small in size and should penetrate most protein-DNA interfaces, it is still possible that certain proteins that bind to ssDNA very tightly can block the N_3-kethoxal labeling reaction. The kinetics of Pol II binding and movement may also affect the relative strength of KAS-seq signals, because stable ssDNA structures may result in more N_3-kethoxal labeling than transient ones. We did not observe evident GC content bias in KAS-seq data, but since N_3-kethoxal specifically labels guanine, the GC content of ssDNA should also be considered when developing KAS-seq with a higher resolution.

N_3-Kethoxal labeling is sensitive to high temperature and guanine analogs

The covalent reaction between guanine and N_3-kethoxal is reversible, which allows the removal of N_3-kethoxal from DNA fragments after enrichment to facilitate PCR amplification. However, N_3-kethoxal–labeled DNA should be treated cautiously with regard to temperature and buffer conditions to prevent the undesired dissociation of the N_3-kethoxal–guanine adduct during the assay.

First, the adduct is unstable at high temperatures. N_3-kethoxal modification can be almost completely removed in 8 h at 37 °C or in 10 min at 95 °C. Therefore, N_3-kethoxal–labeled DNA samples should be kept on ice at all times to minimize dissociation and kept at −20 °C for long-term storage. Second, the kethoxal-guanine adduct is reported to be unstable under alkaline conditions and is best stabilized in borate buffer as previously reported[44]. Third, guanine analogs, such as GTP and guanidinium, promote the dissociation reaction by trapping dissociated N_3-kethoxal and preventing it from reacting with ssDNA guanines. Therefore, samples should avoid long-term exposure to buffers containing guanine analogs and bases. N_3-kethoxal is known to react slowly with arginine residues in proteins[45]. To address this, inclusion of a DNA purification step with proteinase digestion is important to prevent the pull-down of DNA fragments bound by N_3-kethoxal–labeled proteins.

Lastly, even though N_3-kethoxal labeling can be done within 5–10 min, we cannot rule out the possibility that longer N_3-kethoxal treatment may induce cellular stresses. We therefore do not recommend performing N_3-kethoxal treatment for >15 min.

Future improvements

KAS-seq is a robust and sensitive approach for mapping genome-wide ssDNA and captures transcription activity under different physiological conditions. However, as mentioned above, the method can be further improved. First, the coupling of KAS-seq with Pol II ChIP-seq or similar methods such

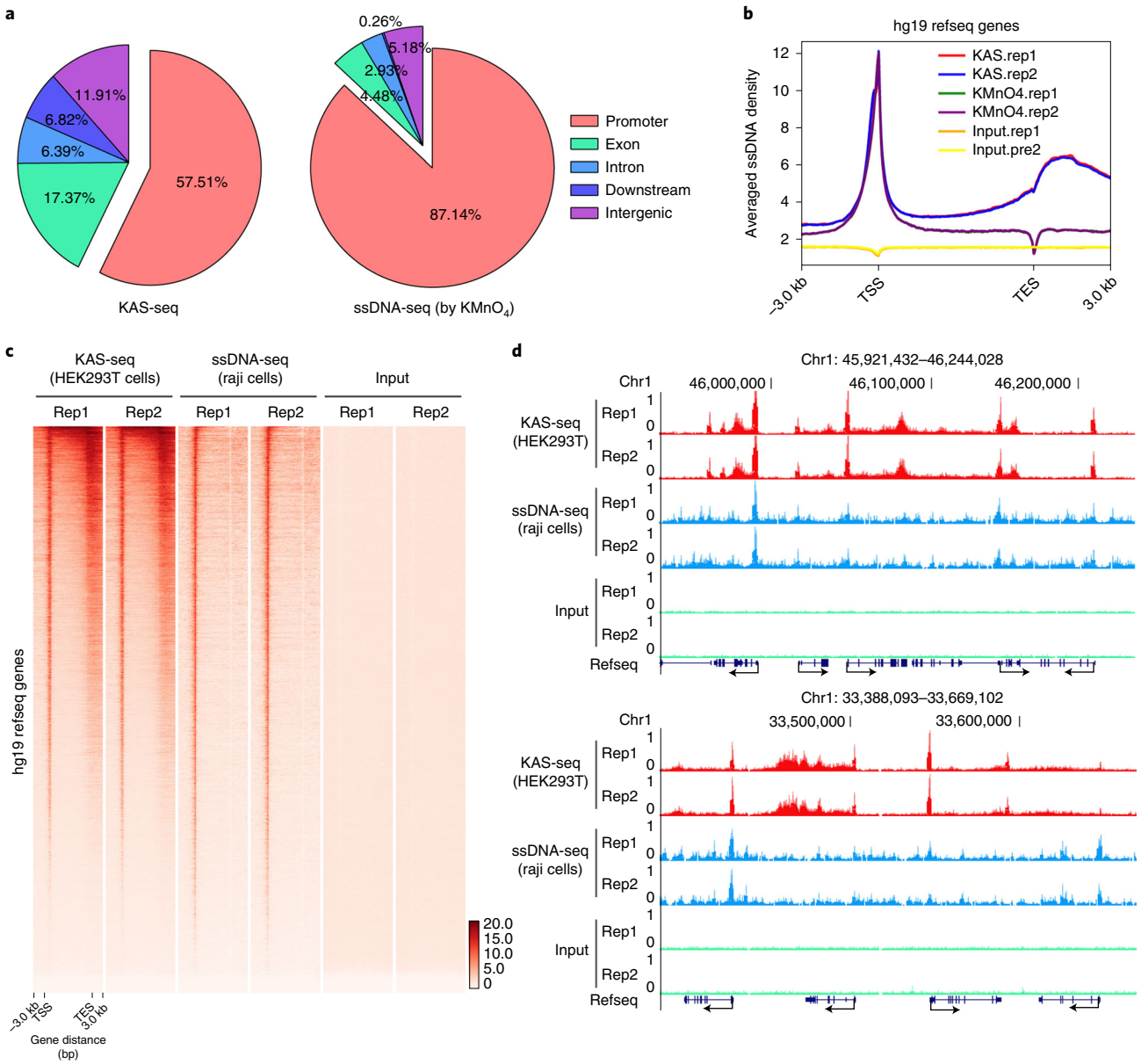

Fig. 2 | Comparison of KAS-seq with a potassium permanganate ($KMnO_4$)-based method. a, Genome-wide distribution of bins (1 kb) with significant ssDNA signals measured by KAS-seq and KMnO4-based ssDNA-seq as indicated by genomic features when mapped to the Refseq hg19 genome. Promoter regions were defined as 2 kb upstream and downstream of the transcription start site (TSS). KAS-seq data were generated in HEK293T cells, and KMnO4-based ssDNA-seq data were generated in Raji cells. The percentage of each genomic feature is shown on the pie chart. **b**, Metagene profile of averaged ssDNA density at gene-coding regions; regions from 3 kb upstream of the TSS to 3 kb downstream of the transcription end site (TES) are shown. **c**, Heatmap showing the ssDNA density on gene coding regions; regions from 3 kb upstream of the TSS to 3 kb downstream of TES are shown. **d**, A snapshot of tracks from UCSC Genome Browser showing the pattern of ssDNA peaks detected by KAS-seq and KMnO4-based ssDNA-seq on two example regions. chr, chromosome; rep, replicate.

as CUT&Tag should enable us to differentiate transcription-mediated ssDNA signals from other ssDNA structures[46]. Similar approaches can be applied to other proteins to study R-loops or ssDNA-binding proteins.

Second, the resolution of KAS-seq can be improved. KAS-seq has a resolution (200–500 bp) similar to ChIP-seq and cannot precisely locate short ssDNA species such as some non-canonical DNA structures and DNA damage sites. Fragmentating the DNA to a smaller size may lead to a higher resolution. Moreover, the resolution of KAS-seq can be further improved by detecting the exact sites of N_3-kethoxal modifications, which may be achieved biochemically by using polymerases

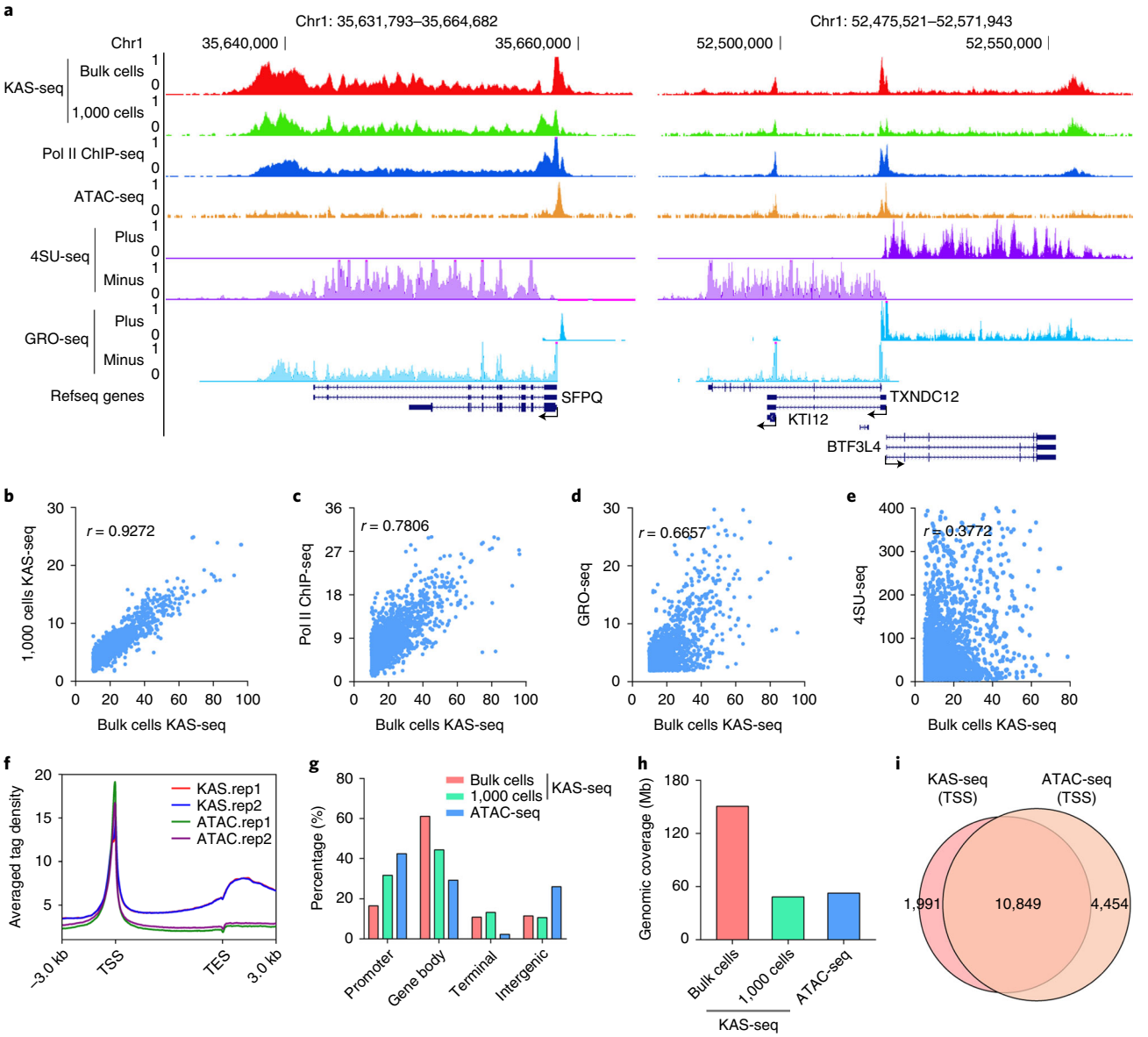

Fig. 3 | Comparison between KAS-seq and other transcriptional activity-profiling methods. a, A snapshot of tracks from the UCSC Genome Browser showing the pattern of KAS-seq (bulk cells and 1,000 cells), Pol II ChIP-seq, ATAC-seq, 4SU nascent RNA-seq and GRO-seq data on two example regions (chr1:35,631,793–35,664,682 and chr1:52,475,521–52,571,943) in HEK293T cells. **b**–**e**, Scatterplots showing the Pearson correlation between KAS-seq data in bulk and 1,000 HEK293T cell (**b**), Pol II ChIP-seq (**c**), GRO-seq (**d**) or 4SU nascent RNA-seq (**e**) data on expressed genes (n = 10,066) in HEK293T cells. The Pearson correlation coefficient r values were calculated as a two-tailed probability. **f**, Metagene profile plot showing the distribution of KAS-seq signals at gene-coding regions; regions from 3 kb upstream of the TSS to 3 kb downstream of the TES were shown. **g**, Bar plot showing the distribution of KAS-seq and ATAC-seq peaks on different genomic features. Promoter regions were defined as 2 kb upstream and downstream of the TSS. **h**, Genomic coverage of KAS-seq peaks of bulk cells and 1,000 cells, and ATAC-seq in HEK293T cells. **i**, Venn diagram showing the overlap between KAS-seq peaks and ATAC-seq peaks enriched around the TSS (+/−2 kb). The number of merged peaks between two replicates was used in each case.

that could read potentially modified Gs as mutations[47] or by using the third-generation sequencing technology PacBio single-molecule real-time sequencing[48].

Experimental design

Labeling of cells or tissues with N_3-kethoxal

When planning a KAS-seq experiment, cells can be labeled in their own culture medium supplemented with N_3-kethoxal. The medium should be pre-warmed to facilitate the dissolution of N_3-kethoxal, because the molecule may precipitate at low temperatures such as 4 °C. Commonly used

buffers with a neutral pH (such as PBS) can also be used to replace the cell culture medium if needed. Conditioned medium can also be used if special experimental conditions require it. A 10-min incubation at 37 °C is usually sufficient for labeling, and a longer incubation (>30 min) may cause cell death. To detect transient transcriptional changes, the incubation can be as short as 5 min.

The KAS-seq protocol described below has been tested with mouse liver tissues. The exact labeling condition for other tissues may need to be optimized by altering the labeling time, concentrations of N3-kethoxal and the labeling temperature. The amount of tissue should be enough for getting 2–3 μg of gDNA for biotinylation.

DNA biotinylation by click chemistry

Isolated DNA labeled by N_3-kethoxal is then biotinylated with the azide-alkyne click cycloaddition reaction by using the strained alkyne (dibenzocyclooctyne, DBCO) conjugated to a biotin molecule. The reaction is performed at 37 °C for 1.5 h in a neutral buffer supplemented with K_3BO_3. Borates have been reported to stabilize compounds with pinacol structures such as in this case, the N_3-kethoxal–guanine adduct[44]. An additional RNase treatment step is included after the reaction to remove potential RNA contamination that may affect DNA quantification and enrichment. The efficiency of N_3-kethoxal labeling and biotinylation can be readily evaluated, if needed, by dot blot using horseradish peroxidase (HRP)-conjugated streptavidin and detection using a chemiluminescent HRP substrate. The biotinylated gDNA is then fragmented to ~300 bp by sonication or enzymatic fragmentation.

Enrichment of biotinylated DNA fragments using streptavidin beads

Biotinylated DNA fragments are enriched using streptavidin-coated magnetic beads. After binding, the beads are usually washed five times with a Tween-20–containing buffer to eliminate non-specific binding. After pull-down, an incubation of 95 °C for >10 min is used to elute the ssDNA fragments from the streptavidin beads and dissociate the N_3-kethoxal adduct at the same time. If N_3-kethoxal is left covalently bound to DNA, the adduct may affect the processivity of DNA polymerases, which may cause potential bias during library construction.

Library construction

Because all the DNA species are denatured during the 95 °C incubation during DNA elution from the streptavidin beads, library construction methods that are compatible with ssDNA analytes are necessary. For KAS-seq with low-input starting material (Box 1), a transposase-based library construction method is applied to decrease potential loss during purification steps and increase the overall efficiency. In the low-input KAS-seq protocol, the adapters are introduced to the DNA fragments before DNA enrichment. Therefore, PCR amplification can be performed on the beads directly after enrichment.

Data analysis

The data analysis pipeline for KAS-seq is very similar to pipelines for other popular chromatin profiling methods, such as ChIP-seq, ATAC-seq and CUT&RUN. Widely used data processing tools including Trim Galore for adapter sequence removal[19], Bowtie 2 for alignment[20], samtools for data format conversion[21], deeptools for basic plots[25], DESeq2 for differential analysis[24] and other popular tools are also recommended for KAS-seq data analysis. To allow users to better interpret their KAS-seq data and generate publication-quality figures, we developed KAS-pipe, a user-friendly integrative data analysis pipeline in this protocol[26]. The input samples, if applicable, should be processed and analyzed in parallel with the enriched DNA.

Materials

Biological materials

- HEK293T cells (American Type Culture Collection, cat. no. CRL-11268; RRID: CVCL_1926)
- Murine embryonic stem cells (American Type Culture Collection, cat. no. CRL-1821; RRID: CVCL_9108) **! CAUTION** The cell lines used in your research should be regularly checked to ensure that they are authentic and are not infected with mycoplasma. The cell lines used in this protocol were checked regularly to ensure their authenticity and mycoplasma-negative status.
- Male B6 mice (Jackson Laboratory, C57BL/6J; RRID: IMSR_JAX:006362) **! CAUTION** Any experiments involving animals must conform to relevant institutional and national regulations. All animal care and experiments described in this protocol were approved by the University of Chicago Institutional Animal Care and Use Committee.

Reagents

- N_3-kethoxal (synthesized as described previously[17]; AccuraDX, cat. no. ADX-NK-Soln)
- Cell culture medium appropriate for the chosen cell line
- Dimethyl sulfoxide (DMSO; Sigma, cat. no. 276855)
- DBCO-PEG4-biotin (Sigma, cat. no. 760749)
- Amersham Hybond-N+ membrane (GE Healthcare, cat. no. RPN119B)
- Super Signal West Pico PLUS chemiluminescent substrate (Thermo, cat. no. 34577)
- Streptavidin-HRP (Thermo Fisher, cat. no. S-911)
- 25 mM potassium borate (K_3BO_3) buffer pH 7.0–8.0 (AccuraDX, cat. no. ADX-BB-25)
- Alternative to pre-made potassium borate buffer: boric acid (Sigma, cat. no. B6768) and potassium hydroxide (Sigma, cat. no. P5958) **! CAUTION** Boric acid is hazardous, and potassium hydroxide is corrosive. Wear gloves and handle with care.
- Tween-20 (Sigma, cat. no. P9416)
- PBS (Thermo, cat. no. 10010001)
- 10× PBS (Thermo, cat. no. AM9624)
- 1 M Tris-HCl buffer (Thermo, cat. no. 15567027)
- 5 M sodium chloride solution (Thermo, cat. no. AM9759)
- 0.5 M EDTA solution (Thermo, cat. no. AM9206G)
- Proteinase K (Thermo Fisher Scientific, cat. no. EO0492)
- RNase A (Thermo, cat. no. 12091039)
- Dynabeads MyOne streptavidin C1 (Thermo, cat. no. 65001)
- PureLink GenomicDNA Mini kit (Thermo, cat. no. K182001) or any equivalent alternatives
- Quick-DNA Microprep Plus kit (Zymo, cat. no. D4074) or any equivalent alternatives
- Qubit dsDNA High Sensitivity (HS) Kit (Life Technologies, cat. no. Q32851)
- MinElute PCR purification kit (Qiagen, cat. no. 28804) or any equivalent alternatives
- DNA Clean & Concentrator kit (Zymo, cat. no. D4043) or any equivalent alternatives
- Tagmentation DNA TDE1 enzyme and buffer kit (Illumina, cat. no. 20034197)
- i5 index primer (Illumina, cat. no. 20027213)
- i7 index primer (Illumina, cat. no. 20027213)
- NEBNext Ultra II Q5 Master Mix (NEB, cat. no. M0544S)
- Accel-NGS Methyl-Seq DNA library kit (Swift, cat. no. 30024) or any equivalent alternatives

Equipment

- 0.2-ml PCR tubes
- 1.5-ml microcentrifuge tubes
- Micro Bio-Spin P-6 gel columns (Bio-Rad, cat. no. 7326222)
- Tube rotator
- Vortex mixer
- Nanodrop, Qubit or equivalent alternatives
- Microcentrifuge
- Incubated mixer
- Thermal cycler
- Magnetic racks for 1.5-ml microcentrifuge tubes and 0.2-ml PCR tubes
- Bioruptor Pico (Diagenode) and Microtubes (Diagenode, cat. no. C30010011) or equivalent alternatives
- pH meter

Software

- FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- Cutadapt (http://cutadapt.readthedocs.io/en/stable/index.html)
- Bowtie 2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
- SAMtools (http://samtools.sourceforge.net/)
- bedtools (http://bedtools.readthedocs.io/en/latest/)
- Picard Tools (http://broadinstitute.github.io/picard)
- MACS v.2.0 (https://github.com/taoliu/MACS)
- UCSC Genome Browser (https://genome.ucsc.edu/)
- Deeptools (https://deeptools.readthedocs.io/en/develop/)

- Alternatively, use KAS-pipe, a user-friendly integrative data analysis pipeline for KAS-seq. KAS-pipe is used for quality control, producing alignments, measures of enrichment and differential and clustering analysis. All KAS-pipe codes are available on GitHub[26] (https://github.com/Ruitulyu/KAS-pipe).

Reagent setup

N_3-kethoxal stock solution

Prepare a 500 mM N_3-kethoxal stock solution. Vortex thoroughly and dissolve the N_3-kethoxal in DMSO. Store the N_3-kethoxal stock solution at −20 °C for 6 months. To ensure compound stability, divide into smaller-volume aliquots for freezing and avoid multiple freeze-thaws.

20 mM DBCO-PEG_4-biotin stock solution in DMSO

Prepare 20 mM DBCO-PEG_4-biotin stock solution by dissolving the purchased DBCO-PEG_4-biotin compound in DMSO. Store this solution at −20 °C. To ensure compound stability, divide into smaller-volume aliquots for freezing and avoid multiple freeze-thaws.

25 mM K_3BO_3 solution

If potassium borate buffer is not purchased commercially, make 500 ml of 25 mM boric acid solution by dissolving 773 mg of boric acid into 500 ml of nuclease-free water. Then, add potassium hydroxide pellets while monitoring pH with a pH meter until the pH reaches 7.0. This buffer can be stored at room temperature (~25 °C) for ≥6 months.

1× and 2× binding and wash buffer (B&W buffer)

To prepare 100 ml of 2× B&W buffer, add 1 ml of 1 M Tris-HCl (pH 7.4), 200 µl of 0.5 M EDTA, 40 ml of 5 M NaCl and 100 µl of Tween-20 to 58.7 ml of nuclease-free water. To prepare 1× B&W buffer, add 50 ml of nuclease-free water to 50 ml of 2× B&W buffer. These buffers can be stored at room temperature for ≥6 months.

Labeling medium (for immediate use)

Pre-warm fresh cell culture medium or PBS at 37 °C. Dissolve the N_3-kethoxal stock solution (500 mM) into pre-warmed cell culture medium or PBS at a 1:100 dilution. The final N_3-kethoxal solution concentration should be 5 mM. The labeling medium should be made fresh and used immediately. Conditioned medium should be used instead of fresh medium for systems that are known to be sensitive to medium changes (such as heat shock–related studies[49]). Other chemicals that are known to react with N_3-kethoxal (such as 1,2- and 1,3-diamines) should not be added in the labeling medium, to avoid reducing the efficiency of KAS-seq.

Procedure

▲ **CRITICAL** The bulk KAS-seq procedure uses 1–5 × 10^6 cultured mammalian cells. All centrifugation steps are performed at room temperature, unless otherwise specified.

N_3-kethoxal labeling and DNA isolation ● Timing 1–2 h

1 Prepare a 500 mM N_3-kethoxal stock solution using DMSO as the solvent, as described in Reagent setup. Then, prepare the labeling medium by diluting the N_3-kethoxal solution into pre-warmed (37 °C) cell culture medium to a final concentration of 5 mM.
▲ **CRITICAL STEP** Before preparing the N_3-kethoxal stock solution, vortex the N_3-kethoxal thoroughly until it appears homogeneous with no precipitation present at the bottom of the tube. To facilitate N_3-kethoxal dissolution, we recommend preparing the labeling medium by diluting the N_3-kethoxal stock solution into pre-warmed (37 °C) cell culture medium instead of adding 500 mM N_3-kethoxal stock solution into the dish directly.

2 Incubate 1–5 million cells in the labeling medium for 10 min at 37 °C. For adherent cells, remove the existing medium, apply the labeling medium to the cell culture directly in dishes and incubate cells for 10 min at 37 °C. For suspension cells, suspend cells in the labeling medium and incubate them for 10 min at 37 °C. To label cells efficiently, the labeling medium should completely cover the surface of the cell culture plates for adherent cells.

3 Harvest and pellet cells by centrifugation at 500*g* for 5 min after the incubation. Isolate total gDNA from cells using the PureLink Genomic DNA Mini kit (Thermo, cat. no. K182001, or any

equivalent kit) per the manufacturer's protocol with one exception: elute DNA using 50 µl of 25 mM K_3BO_3 (pH 7.0).

▲ **CRITICAL STEP** We recently found that guanidinium, the cation of common chaotropic salts for silica column–based DNA isolation kits, may lead to the dissociation of the N_3-kethoxal–ssDNA adduct if the samples are exposed for >30 min to the guanidinium-containing buffers. To avoid overexposing samples, we recommend working quickly through the DNA extraction and purification kits.

■ **PAUSE POINT** The eluted DNA can be stored at −20 °C for future use; avoid multiple freeze-thaws.

Biotinylation and purification ● Timing 1–2 h

4 Quantitate the purified DNA using NanoDrop, Qubit or equivalent.

▲ **CRITICAL STEP** If <2–3 µg of gDNA is recovered from Step 4, return to Steps 1–3 to obtain additional gDNA for a total of 2–3 µg.

5 Prepare 2–3 µg of gDNA in 85 µl of 25 mM K_3BO_3 (pH 7.0) and add 10 µl of 10× PBS and 5 µl of 20 mM DBCO-PEG_4-biotin DMSO solution. Incubate the mixture at 37 °C for 1.5 h with shaking at 500 rpm to facilitate the click cycloaddition reaction.

6 Add 5 µl of RNase A to the reaction mixture. Incubate the mixture at 37 °C for 15 min with shaking at 500 rpm.

7 Purify DNA from the reaction mixture by DNA Clean & Concentrator kit (Zymo) following the manufacturer's protocol for gDNA with one exception: elute DNA using 50 µl of 25 mM K_3BO_3 (pH 7.0).

▲ **CRITICAL STEP** Similar to the guanidinium warning above, avoid overexposure of DNA to guanidium-containing buffers for longer than 30 min during the purification step.

■ **PAUSE POINT** The eluted DNA can be stored at −20 °C for future use.

Enrichment of N_3-kethoxal–labeled DNA ● Timing 1–2 h

8 Quantitate DNA from Step 7 and dilute 1 µg of biotinylated DNA into 100 µl of 25 mM K_3BO_3 (pH 7.0). Fragment the DNA to 150–350 bp by sonication. For the Bioruptor Pico Sonicator, a 30 s on/30s off setting for 30 cycles at 4 °C with DNA in the manufacturer's 0.5-ml sonication tubes achieves this fragment size range.

▲ **CRITICAL STEP** If <1 µg of biotinylated DNA is recovered from Step 7, repeat Steps 5–7 to obtain additional biotinylated DNA for a total of 1 µg.

? TROUBLESHOOTING

9 Save 5 µl of the fragmented DNA as an input control on ice. Use the remaining 95 µl for enrichment.

10 Wash 10 µl of Dynabeads MyOne streptavidin C1 three times with 50 µl of 1× B&W buffer (5 mM Tris-HCl (pH 7.4), 0.5 mM EDTA, 1 M NaCl and 0.05% Tween-20), using a magnetic rack to separate the beads in between washes. Resuspend prewashed beads in 95 µl of 2× B&W buffer (10 mM Tris-HCl (pH 7.4), 1 mM EDTA, 2 M NaCl and 0.1% Tween-20).

11 Mix the beads with the 95 µl of fragmented DNA from Step 9. Incubate the mixture at room temperature for 15 min with gentle rotation. After incubation, place the mixture on a magnetic rack to remove the supernatant and wash beads five times with 100 µl of 1× B&W buffer.

12 Resuspend the beads in 15 µl of nuclease-free water and heat at 95 °C for 10 min to elute enriched DNA. Then, place the mixture on a magnetic rack and transfer the supernatant to a new labeled microcentrifuge tube.

13 In parallel with the previous step, heat the saved input from Step 9 at 95 °C for 10 min. Then, put the samples on ice immediately.

▲ **CRITICAL STEP** Step 13 is essential for removing the N_3-kethoxal modification from the DNA. Incomplete removal of N_3-kethoxal modification may affect PCR amplification during library preparation.

■ **PAUSE POINT** The eluted DNA can be stored at −20 °C for future use.

Library preparation and sequencing ● Timing 3–5 h (for library preparation only)

14 Perform library construction using the Accel-NGS Methyl-Seq DNA library kit (Swift, cat. no. 30024, or any equivalent kit). The exact PCR cycles should be determined by qPCR or equivalent approaches. As an approximate guide in the models we tested, PCR amplification with 7–8 cycles for input samples and 12–14 cycles for enriched samples should yield a sufficient quantity of library for sequencing. The library size should range from 200 to 500 bp.

■ **PAUSE POINT** Libraries can be stored at −20 °C if needed.

? TROUBLESHOOTING

Table 1 | Introduction of provided shell scripts in KAS-pipe

Name of shell script	Function of shell script
install_conda_env.sh	Install the conda environment and packages needed for KAS-pipe
update_conda_env.sh	Update the conda environments
uninstall_conda_env.sh	Uninstall the conda environments
download_reference_genome.sh	Download the reference genome used for KAS-seq data mapping
build_reference_genome.sh	Build the Bowtie 2 index for the downloaded reference genome
trim_adapter.sh	Remove low-quality reads and trim adapter sequences from raw KAS-seq data
map_KAS-seq.sh	Map trimmed KAS-seq data to the reference genome and perform de-duplication and data format conversion (sam to bam, bed and bg)
call_KAS-seq_peaks.sh	Call KAS-seq peaks at regular and broad modes
normalize_KAS-seq.sh	Normalize KAS-seq data on the basis of library size
make_BigWig_files.sh	Convert the normalized files to bigwig files
diff_KAS-seq.sh	Perform differential analysis for KAS-seq data on the basis of DESeq2. Generate volcano plot
make_UCSC_files.sh	Make files for upload and visualization on UCSC Genome Browser
plotSummary.sh	Generate the metagene profile plot and heatmap on user-defined regions, promoter, gene body and transcription termination regions
plotCorrelation.sh	Perform correlation analysis; generate circle and heatmap plot for pairwise correlation coefficients
plotFingerprint.sh	Generate the fingerprint plot
plotPCA.sh	Perform PCA
define_single-stranded_enhancers.sh	Define the single-stranded active enhancers (SSEs)

15 Perform single-end or paired-end sequencing on the libraries constructed in Step 14 on Illumina sequencing platforms or another massively parallel DNA sequencer by following the manufacturer's instructions.
▲ **CRITICAL STEP** The optimal sequencing depth varies on the basis of the size of the reference genome. For studies of human or mouse samples, we recommend >40 million unique mapped reads per sample for effective genome-wide ssDNA detection and differential analysis.

16 Process the KAS-seq data using KAS-pipe as described below in Data processing and analysis.

Data processing and analysis ● Timing ~3 h

▲ **CRITICAL** Considering the fact that experimental biologists have varying bioinformatics experience for NGS data processing, providing a facile means for data analysis is as crucial as an experimental protocol. To allow users to better interpret their KAS-seq data and generate publication-quality figures, we developed KAS-pipe, a user-friendly integrative data analysis pipeline for KAS-seq data[26]. In KAS-pipe, we have provided 17 shell scripts and 4 R scripts, which can be used to perform a very comprehensive analysis for KAS-seq data. This includes: building the reference genome, quality control, adapter trimming, producing alignments, normalization, differential analysis, correlation analysis, principle component analysis (PCA), creating heatmap and metagene profile plots, defining single-stranded enhancers[18] and visualization of KAS-seq signals (Table 1). In the data-processing part of this protocol, we describe only the commands used to invoke KAS-pipe. For the specific tools used in KAS-seq, parameters and related tutorials, the shell scripts deposited on GitHub can be accessed so that KAS-pipe can be installed onto a cluster or workstation computer.

17 Perform a quality control check on KAS-seq raw sequence data using the popular quality control tool FastQC. Remove the low-quality reads and trim adapter sequences off of reads using the Trim Galore package at single-end or paired-end mode on the basis of the layout of fastq files. Reads shorter than 50 bp will be removed.
Single-end:

```
> nohup trim_adapter.sh illumina 30 10 single KAS-seq.fastq.gz &
```

Paired-end:

```
> nohup trim_adapter.sh illumina 30 10 paired KAS-seq.R1.fastq.gz
KAS-seq.R2.fastq.gz &
```

18 Align the trimmed reads to the reference genome using the default parameters of Bowtie 2. Convert the resulting sam file to a bam file and sort using samtools sort (v1.9). For single-end KAS-seq data, remove PCR duplicates using samtools rmdup (v1.9) to filter out the uniquely mapped reads. Extend unique mapped reads to 150 bp to match the average length of insert DNA fragments in KAS-seq libraries. Convert the extended bed files to bedGraph files by using bedtools genomecov.

```
> nohup map_KAS-seq.sh /Genome/hg19_Bowtie2Index/hg19 10
KAS-seq_SE_output hg19 single KAS-seq.trimmed.fastq.gz &
```

For paired-end KAS-seq data, remove PCR duplicates using the Picard MarkDuplicates tool to filter out the uniquely mapped reads. Unlike the operation of extending uniquely mapped reads to match the average length of DNA insert fragments, the length of DNA insert fragments for paired-end reads is calculated by merging two mapped ends using SAMtoBED. Merged bed files are converted to bedGraph files using bedtools genomecov, similarly to single-end read data.

```
> nohup map_KAS-seq.sh /Genome/hg19_Bowtie2Index/hg19
10 KAS-seq_PE_output hg19 paired KAS-seq.trimmed.R1.fastq.gz KAS-seq.
trimmed.R2.fastq.gz &
```

? TROUBLESHOOTING

19 Call KAS-seq peaks using MACS2 in broad mode under qvalue=0.01.

```
> nohup call_KAS-seq_peaks.sh KAS_seq_bed_files.txt Input_bed_files.txt
broad KAS-seq_peaks hg &
```

20 Create the fingerprint plot to check that the KAS-seq signal can be separated from the background signal.

```
> nohup plotFingerprint.sh KAS_seq_files.txt labels.txt 10 hg19
KAS-seq &
```

? TROUBLESHOOTING

21 Normalize KAS-seq data based on the number of uniquely mapped reads.

```
> nohup normalize_KAS-seq.sh KAS_seq_file.txt ratio.txt &
```

22 Create a heatmap or metagene profile for KAS-seq signal enriched on promoters, gene body or any regions of interest using plotHeatmap or plotProfile.
KAS-seq heatmap on gene body:

```
> nohup plotSummary.sh KAS_seq_files.txt labels.txt hg19 10
KAS-seq_on_genebody genebody heatmap colors.txt &
```

23 Perform differential analysis of KAS-seq data using the DESeq2 package to identify regions (bins, peaks, promoter, gene body or terminator) with differential KAS-seq signal.
Bins (1 kb):

```
> nohup diff_KAS-seq.sh KAS_seq_files.txt labels.txt bins hg19
KAS-seq_treat_vs_DMSO 10 diff_condition.txt &
```

Gene body:

```
> nohup diff_KAS-seq.sh KAS_seq_files.txt labels.txt promoter hg19
KAS-seq_treat_vs_DMSO 10 diff_condition.txt &
```

Peak regions:

```
> nohup diff_KAS-seq.sh KAS_seq_files.txt labels.txt peaks hg19
KAS-seq_treat_vs_DMSO 10 diff_condition.txt peaks_list.bed &
```

24 Perform PCA and correlation analysis.
PCA:

```
> nohup plotPCA.sh KAS_seq_files.txt labels.txt colors.txt bins KAS-seq 10 &
```

Correlation analysis:

```
> nohup plotCorrelation.sh KAS_seq_files.txt labels.txt bins hg19
KAS-seq 10 heatmap &
```

25 Define two types of single-stranded enhancers (SSEs) from distal H3K27ac or ATAC-seq peaks. We previously found that the transcribed SSEs can be separated into two types on the basis of the pattern of KAS-seq signal on H3K27ac- or ATAC-seq–defined enhancers, entire SSEs and middle SSEs.

```
> nohup define_single-stranded_enhancers.sh H3K27ac/ATAC-seq_enhan-
cers.bed KAS-seq_peaks.bed hg19 &
```

26 *Visualization*. Prepare the bedGraph files with track definition lines as custom tracks of the UCSC Genome Browser for visualization.

```
> nohup make_UCSC_files.sh trackname.txt KAS_seq_bedGraph_files.txt &
```

For more details and other shell scripts included in KAS-pipe, please check the KAS-pipe GitHub page or Table 1.

Troubleshooting

Troubleshooting advice can be found in Table 2.

Table 2 | Troubleshooting table

Step	Problem	Possible reason	Solution
8 (or 7 in Box 1)	The size of DNA is too large	Inefficient sonication or non-optimal sonication parameters	The DNA size after fragmentation can be examined by gel electrophoresis. Optimize sonication parameters for your DNA
14 (or 15 in Box 1)	The concentration of library is lower than the optimal concentration suggested by Illumina sequencing	Low library construction efficiency	Strictly follow the protocol of the library construction kit
		Insufficient number of PCR cycles	Run a test PCR to determine the optimal number of PCR cycles
18	Error: unsupported reference genome	No provided reference genome of interest	Download the reference from a public database and use bowtie2-build to generate the Bowtie 2 index
	High duplication ratio	Too many PCR cycles	Decrease the number of PCR cycles
		Low labeling efficiency	Strictly follow the N_3-kethoxal labeling protocol
			After the biotinylation step, use a dot blot assay[17] to check the labeling efficiency before proceeding to the next step
			Examine the exposure of labeled DNA to guanidium buffers and reduce the exposure time
20	Weak ssDNA enrichment	Low labeling efficiency	Same as above
		Incomplete removal of N_3-kethoxal labeling before PCR amplification	Extend the heating time in Steps 12 and 13 of the bulk KAS-seq protocol

Timing

~1–3 d are needed for generating KAS-seq libraries from cultured cells or isolated tissues depending on the number of samples.

KAS-seq with bulk cultured mammalian cells

Steps 1–3, N_3-kethoxal labeling and DNA isolation: 1–2 h

Steps 4–7, biotinylation and purification: 1–2 h

Steps 8–13, enrichment of N_3-kethoxal–labeled DNA: 1–2 h

Steps 14 and 15, library preparation: 3–5 h (variable, depending on library construction kit)

Step 16, sequencing: variable

Steps 17–26, data processing and analysis: ~3 h (variable, depending on sequencing depth, available computing power and users' familiarity with bioinformatics)

Box 1: low-input KAS-seq

Steps 1 and 2, N_3-kethoxal labeling and DNA isolation: 1 h

Steps 3–5, biotinylation and purification: 1–2 h

Steps 6–9, tagmentation: 1 h

Steps 10 and 11, enrichment of N_3-kethoxal–labeled DNA: 1–2 h

Steps 12–15, PCR amplification and library clean-up: 3 h

Step 16, sequencing: variable

Box 2: KAS-seq with mouse liver tissue

Steps 1–5, N_3-kethoxal labeling and DNA isolation: 1.5 h

Step 6, the timing for each activity is the same as shown above under 'KAS-seq with bulk cultured mammalian cells'

Anticipated results

Quality control of KAS-seq data

Quality control metrics provide the best definitive measure of whether the KAS-seq experiment was successful. Common analysis questions include the following. What is the minimal sequencing depth for a regular KAS-seq experiment? What is a normal PCR duplication ratio? Did the KAS-seq

Table 3 | Summary of KAS-seq data processing in HEK293T cells

KAS-seq data	Raw reads	Trimmed reads	Mapped reads	Unique reads	Mapping ratio (%)	Duplication ratio (%)
KAS-seq rep1	39,249,037	39,194,455	38,789,121	32,336,611	98.97	16.63
KAS-seq rep2	37,235,447	37,195,072	36,798,642	31,099,474	98.93	15.49
Input rep1	45,182,939	45,162,826	44,661,618	42,241,435	98.89	5.42
Input rep2	39,911,067	39,886,186	39,386,077	37,011,673	98.75	6.03

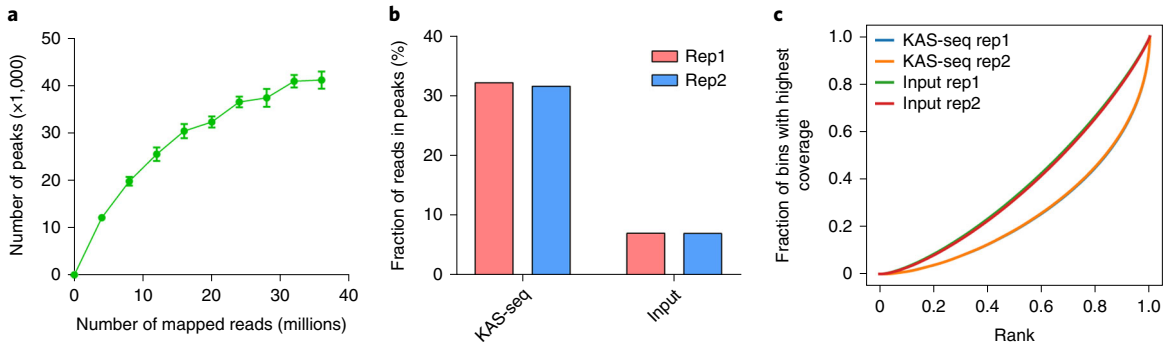

Fig. 4 | Quality control of KAS-seq data in bulk HEK293T cells. a, Saturation analysis of KAS-seq data showing the minimal number of unique mapped KAS-seq reads needed to perform effective KAS-seq data analysis. **b**, Fraction of reads in peaks score showing the fraction of unique mapped reads on KAS-seq peaks. **c**, Fingerprint plot of KAS-seq and corresponding input showing the effective enrichment of ssDNA signals by KAS-seq.

experiment effectively label and enrich ssDNA? Thus, performing quality control is an important step to help judge and evaluate KAS-seq data, as well as troubleshoot any potential problems from the KAS-seq experiment. As an example of a successful KAS-seq experiment, here, we report the KAS-seq data generated in HEK293T cells as a benchmark. Shown in Table 3, the mapping ratio is typically over 95% and the duplication ratio is ~20% or less. Saturation analysis of KAS-seq data shows that ≥40 million uniquely mapped reads per sample are needed to capture most of the ssDNA formed in the genome (Fig. 4a). When MACS2 is used to call broad peaks on ssDNA-enriched domains genome-wide, we find that the fraction of reads in peaks is >30%, which is significantly higher than that for regular transcription factor ChIP-seq at similar sequencing depth (Fig. 4b). Fingerprint

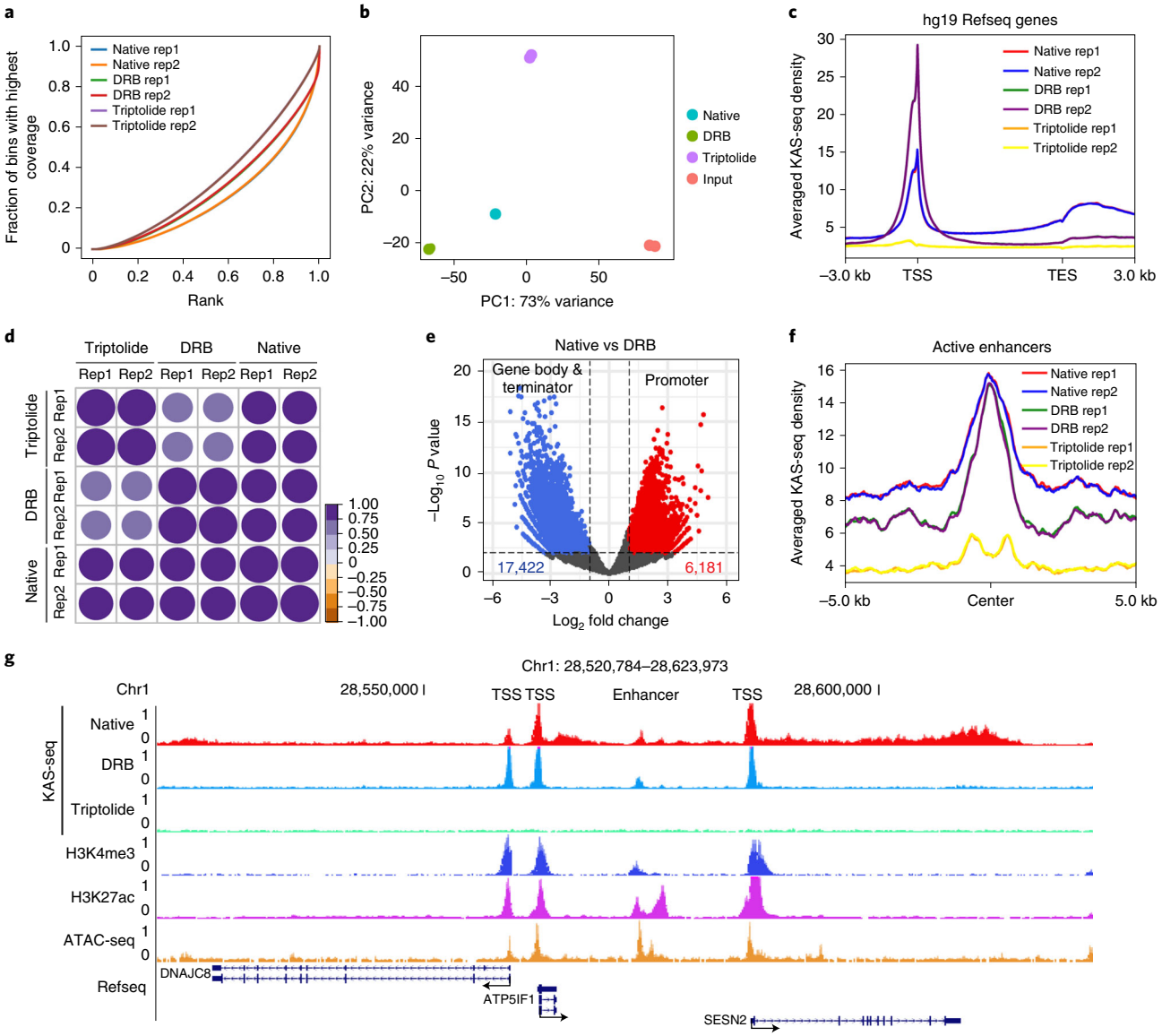

Fig. 5 | Anticipated results using the integrative KAS-seq data analysis pipeline (KAS-pipe). a, Fingerprint plot of KAS-seq data in HEK293T cells under native, DRB treatment and triptolide treatment conditions. **b**, PCA plot of KAS-seq and corresponding input data in HEK293T cells under native, DRB treatment and triptolide treatment conditions. **c**, Metagene profile of averaged KAS-seq density at gene-coding regions; regions from 3 kb upstream of the TSS to 3 kb downstream of the TES are shown. **d**, Circle plot showing correlation analysis for KAS-seq data in HEK293T cells under native, DRB treatment and triptolide treatment conditions. The colors and sizes of circles indicate the pairwise correlation coefficients. **e**, Volcano plot showing differential analysis of KAS-seq data between native and DRB conditions in HEK293T cells. Red dots indicate the bins with significantly higher KAS-seq signal under the DRB treatment conditions (n = 6,181). Blue dots indicate the bins with significantly higher KAS-seq signal under native conditions (n = 17,422). **f**, Metagene profile of averaged KAS-seq density at active enhancers defined by H3K27ac ChIP-seq data in HEK293T cells. **g**, A snapshot of tracks from the UCSC Genome Browser showing the pattern of ssDNA peaks detected by KAS-seq in HEK293T cells under native, DRB treatment or triptolide treatment conditions. H3K4me3, H3K27ac and ATAC-seq peaks in HEK293T cells are also shown.

analysis consistently shows that the ssDNA signal in these KAS-seq samples can be significantly differentiated from the background distribution of reads in the input samples (Fig. 4c). In addition, the ssDNA pattern at gene-coding regions is very similar and typical in most of the cell lines studied in our laboratory, which shows a strong and sharp peak around the transcription start site (TSS), with a weak and broad enrichment pattern along gene bodies and terminal regions (Fig. 2b–d); these trends can also be used as criteria to judge the success of a KAS-seq experiment.

Measuring the dynamics of transcriptionally engaged Pol II

To test the ability of KAS-seq and KAS-pipe to measure the dynamics of transcriptionally engaged Pol II and detecting transcribing enhancers, we used the raw data from our previously published KAS-seq experiments in which HEK293T cells were treated with DMSO (native), 5,6-dichlorobenzimidazole 1-β-D-ribofuranoside (DRB) and triptolide, respectively[18]. Both DRB and triptolide have previously been reported to inhibit transcription, albeit with different mechanisms. DRB inhibits Pol II release from pausing at TSS, whereas triptolide inhibits recruitment and loading of Pol II to promoters[50]. Processing and interpretation of the example data were performed using KAS-pipe (Table 1), and a portion of the typically expected outputs of KAS-pipe are shown in Fig. 5.

The fingerprint plot demonstrates that KAS-seq experiments yielded high enrichment efficiency in native, DRB-treated and triptolide-treated HEK293T cells (Fig. 5a). Meanwhile, the PCA plot reveals that the ssDNA profile in native HEK293T cells can be significantly differentiated from those in the DRB- and triptolide-treated HEK293T cells (Fig. 5b). As expected, the metagene profile of averaged KAS-seq density shows the typical ssDNA pattern at gene-coding regions in native HEK293T cells (Fig. 5c). In contrast, under DRB treatment, ssDNA signals increased around the TSS but almost disappeared at gene body and transcription termination regions; triptolide treatment almost completely erased all signals at gene-coding regions (Fig. 5c), which suggests that KAS-seq is highly sensitive to changes in Pol II–mediated transcription activity regulation. A circle plot of correlation analysis for KAS-seq data between replicates and conditions shows that KAS-seq is very robust and reproducible (Fig. 5d). We also performed differential analysis of KAS-seq data between native and DRB conditions in HEK293T cells. Consistent with the metagene profile, we identified 6,181 bins with significantly higher KAS-seq signal in the DRB-treated condition, which are mainly enriched around TSS, and identified 17,422 bins with significant higher KAS-seq signal in the native condition, which are mainly enriched at gene bodies and terminal regions (Fig. 5e). The metagene profile of average KAS-seq density on active enhancers shows that KAS-seq can be used to define a group of single-stranded active enhancers (Fig. 5f). The screenshot represents a 100-kb region with KAS-seq peaks identified around TSS, gene bodies, terminal regions and enhancers. KAS-seq data in HEK293T cells under native, DRB and triptolide conditions are shown, and enhancers are defined by distal H3K4me3, H3K27ac and ATAC-seq peaks (Fig. 5g).

Data availability

KAS-seq data in HEK293T and mouse embryonic stem cell lines are available at the National Center for Biotechnology Information Gene Expression Omnibus repository under the accession number GSE97072. Global run-on sequencing (GRO-seq) data in HEK293T cells are available under the accession number GSE92375. Pol II ChIP-seq and 4-thiouridine (4SU) nascent RNA-seq data in HEK293T cells are available under the accession number GSE112608.

Code availability

All the KAS-pipe code used in this study is available at https://github.com/Ruitulyu/KAS-pipe[26].

References

1. Chen, F. X., Smith, E. R. & Shilatifard, A. Born to run: control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **19**, 464–478 (2018).
2. Bell, S. P. & Dutta, A. DNA replication in eukaryotic cells. *Annu. Rev. Biochem.* **71**, 333–374 (2002).
3. Hustedt, N. & Durocher, D. The control of DNA repair by the cell cycle. *Nat. Cell Biol.* **19**, 1–9 (2017).
4. Li, X. & Heyer, W.-D. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res.* **18**, 99–113 (2008).
5. Fuchs, G. et al. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol.* **15**, R69 (2014).
6. Kouzine, F. et al. Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.* **4**, 344–356.e7 (2017).

7. Huppert, J. L. & Balasubramanian, S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.* **35**, 406–413 (2007).
8. Zeraati, M. et al. I-motif DNA structures are formed in the nuclei of human cells. *Nat. Chem.* **10**, 631–637 (2018).
9. Cer, R. Z. et al. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41**, D94–D100 (2012).
10. McIntosh, D. B., Duggan, G., Gouil, Q. & Saleh, O. A. Sequence-dependent elasticity and electrostatics of single-stranded DNA: signatures of base-stacking. *Biophys. J.* **106**, 659–666 (2014).
11. Murphy, M., Rasnik, I., Cheng, W., Lohman, T. M. & Ha, T. Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy. *Biophys. J.* **86**, 2530–2537 (2004).
12. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chédin, F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell* **45**, 814–825 (2012).
13. Sollier, J. et al. Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Mol. Cell* **56**, 777–785 (2014).
14. Lei, M., Podell, E. R. & Cech, T. R. Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nat. Struct. Mol. Biol.* **11**, 1223–1229 (2004).
15. Zeitlin, S. G. et al. Double-strand DNA breaks recruit the centromeric histone CENP-A. *Proc. Natl Acad. Sci. USA* **106**, 15762–15767 (2009).
16. Kouzine, F. et al. Global regulation of promoter melting in naive lymphocytes. *Cell* **153**, 988–999 (2013).
17. Weng, X. et al. Keth-seq for transcriptome-wide RNA structure mapping. *Nat. Chem. Biol.* **16**, 489–492 (2020).
18. 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).
19. Krueger, F. Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2015).
20. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
21. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
22. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
23. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
24. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
25. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
26. Lyu, R. Ruitulyu/KAS-pipe: first release of KAS-pipe for KAS-seq data analysis (1.0.0). *Zenodo* https://doi.org/10.5281/zenodo.4941764 (2021).
27. Shapiro, R., Cohen, B. I., Shiuey, S.-J. & Maurer, H. Reaction of guanine with glyoxal, pyruvaldehyde, and kethoxal, and the structure of the acylguanines. Synthesis of N2-alkylguanines. *Biochemistry* **8**, 238–245 (1969).
28. Staehelin, M. Inactivation of virus nucleic acid with glyoxal derivatives. *Biochim. Biophys. Acta* **31**, 448–454 (1959).
29. Litt, M. & Hancock, V. Kethoxal—a potentially useful reagent for the determination of nucleotide sequences in single-stranded regions of transfer ribonucleic acid. *Biochemistry* **6**, 1848–1854 (1967).
30. Noller, H. F. & Chaires, J. B. Functional modification of 16S ribosomal RNA by kethoxal. *Proc. Natl Acad. Sci. USA* **69**, 3115–3118 (1972).
31. LaGrandeur, T. E., Hüttenhofer, A., Noller, H. F. & Pace, N. R. Phylogenetic comparative chemical footprint analysis of the interaction between ribonuclease P RNA and tRNA. *EMBO J.* **13**, 3945–3952 (1994).
32. Yamane, A. et al. RPA accumulation during class switch recombination represents 5′–3′ DNA-end resection during the S–G2/M phase of the cell cycle. *Cell Rep.* **3**, 138–147 (2013).
33. Lange, J. et al. The landscape of mouse meiotic double-strand break formation, processing, and repair. *Cell* **167**, 695–708.e16 (2016).
34. Paiano, J. et al. ATM and PRDM9 regulate SPO11-bound recombination intermediates during meiosis. *Nat. Commun.* **11**, 857 (2020).
35. Hinch, A. G. et al. The configuration of RPA, RAD51, and DMC1 binding in meiosis reveals the nature of critical recombination intermediates. *Mol. Cell* **79**, 689–701.e10 (2020).
36. 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).
37. 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).
38. 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).
39. Paulsen, M. T. et al. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc. Natl Acad. Sci. USA* **110**, 2240–2245 (2013).
40. Fuchs, G. et al. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol.* **15**, R69 (2014).

41. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
42. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat. Methods* **10**, 1213–1218 (2013).
43. Chédin, F. Nascent connections: R-loops and chromatin patterning. *Trends Genet.* **32**, 828–838 (2016).
44. Herschlag, D. *Biophysical, Chemical, and Functional Probes of RNA Structure, Interactions and Folding: Part A* (Academic Press, 2009).
45. Akinsiku, O. T., Yu, E. T. & Fabris, D. Mass spectrometric investigation of protein alkylation by the RNA footprinting probe kethoxal. *J. Mass Spectrom.* **40**, 1372–1381 (2005).
46. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
47. Laos, R., Thomson, J. M. & Benner, S. A. DNA polymerases engineered by directed evolution to incorporate non-standard nucleotides. *Front. Microbiol.* **5**, 565 (2014).
48. Schadt, E. E. et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.* **23**, 129–141 (2013).
49. Mahat, D. B. & Lis, J. T. Use of conditioned media is critical for studies of regulation in response to rapid heat shock. *Cell Stress Chaperones* **22**, 155–162 (2017).
50. Cramer, P. Organization and regulation of gene transcription. *Nature* **573**, 45–54 (2019).

Acknowledgements

We thank all He laboratory members for discussion. We thank the Functional Genomics Facility at The University of Chicago for performing high-throughput sequencing (P30 CA014599). This work was supported by the US National Institutes of Health (R01 HG006827, RM1 HG008935 and P01 NS097206 to C.H.). C.H. is an investigator of the Howard Hughes Medical Institute.

Author contributions

T.W. developed experimental procedures and performed most experiments. D.C.W.-S. and X.W. validated the whole protocol. R.L. performed data analysis and developed the data analysis pipeline with suggestions from A.C.Z. M.C., R.L. and C.H. wrote the manuscript with input and edits 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 AccuraDX, Inc., as well as a shareholder of Epican Genetech. T.W. and D.C.W.-S. are shareholders of AccuraDX, Inc.

Additional information

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

Peer review information *Nature Protocols* thanks Fedor Kouzine, Jian Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work

Received: 9 April 2021; Accepted: 22 September 2021;
Published online: 10 January 2022

Related links
Key references using this protocol
Weng, X. et al. *Nat. Chem. Biol.* **16**, 489–492 (2020): https://doi.org/10.1038/s41589-019-0459-3
Wu, T. et al. *Nat. Methods* **17**, 515–523 (2020): https://doi.org/10.1038/s41592-020-0797-9