Mapping the simultaneously accessible and ssDNA-containing genome with KAS-ATAC

Georgi K. Marinov[1,*,#] and William J. Greenleaf[1,2,3,4,5]

[1] *Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA*
[2] *Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[3] *Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, 94305, USA*
[4] *Chan Zuckerberg Biohub, San Francisco, California, USA*
[5] *Stanford University, Arc Institute*
[#] *Corresponding author*

Abstract

The KAS-ATAC assay provides a method to capture genomic DNA fragments that are simultaneously physically accessible and contain single-stranded DNA (ssDNA) bubbles. These are characteristic features of two of the key processes involved in regulating and expressing genes – on one hand, the activity of *cis*-regulatory elements (cREs), which are typically devoid of nucleosomes when active and occupied by transcription factors, and on the other, the association of RNA polymerases with DNA, which results in the presence of ssDNA structures. Here, we present a detailed protocol for carrying out KAS-ATAC and basic processing of KAS-ATAC datasets, and discuss the key considerations for its successful application.

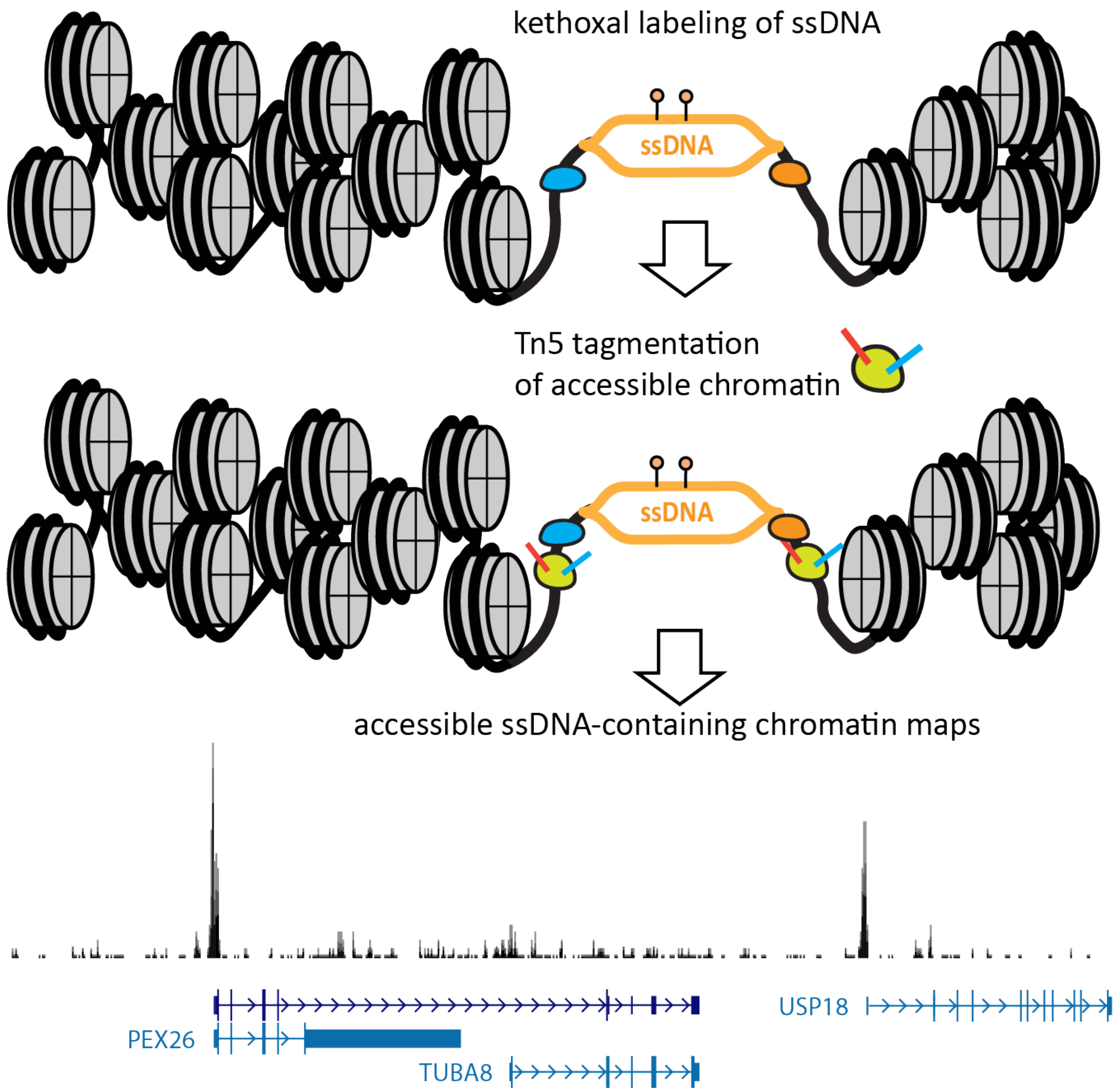
Key features

- Allows mapping of simultaneously accessible and ssDNA-containing DNA fragments.
- Describes the execution of N_3-kethoxal labeling and transposition of native chromatin
- Describes the pulldown of biotin-labeled DNA fragments and library generation
- Describes basic KAS-ATAC data processing steps

Keywords: KAS-seq, ATAC-seq, KAS-ATAC, chromatin, chromatin accessibility, transcription, ssDNA.

This protocol is used in: Genome Res (2024) [1] doi: 10.1101/gr.279621.124.

Graphical overview



Background

The major steps in process of gene expression in most eukaryotes are the act of transcription itself and its activation/repression by the combined action of transcription factors (TFs) on *cis*-regulatory elements, i.e. promoters, enhancers and insulators. The action of cREs can be highly complex – e.g. the input of on average ≥10 individual cREs per gene is integrated into the regulation of promoter activity in mammalian genomes [2, 3] – therefore mapping the location and activation status of cREs genome-wide has been a key tool towards the comprehensive charting of regulatory networks.

To this end, a key property of active cREs – that they tend to be devoid of nucleosomes [4–6] – has been the basis for technological development, due to the fact that the absence of nucleosomal protection renders active cREs accessible to enzymatic action, whether cleavage (e.g. by DNase I [7–10]) or modification (e.g. by methyltransferases [11–13]). The enzyme with most utility for mapping open chromatin genome-wide has turned out to be the hyperactive transposase Tn5 [14, 15], which allows both the labeling of accessible regions in the genome and their tagging with readily PCR-amplifiable sequencing adapters, thus greatly simplifying experimental protocols and minimizing input amount requirements, in the form of the ATAC-seq assay [15] (**A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing) and its many variations.

ATAC-seq datasets also contain highly useful finer-grained information in addition to the mere location of open chromation regions. They can be used to map nucleosome positioning around cREs [16] as well as the footprints of individual TFs associated with DNA, which also protect DNA from cleavage, even if to a lesser extent than nucleosomes [15].

The integrated action of cREs results in the modulation of transcriptional activity at promoters, with enhancers themselves also being transcribed [17]. Thus, mapping active transcription (as opposed to steady-state RNA levels, which are the result of both active transcription and the subsequent effects of intrinsic RNA stability and post-transcriptional regulation) has been another key tool for understanding the regulatory genome. To this end, multiple adaptations of nuclear run-on techniques, such as GRO-seq [18], PRO-seq [19], and others [20] have been developed. However, these are generally fairly complex protocols, and as they measure RNA rather than DNA molecules, they do not allow the simultaneous recording of both transcriptional activity and the physical state of the genome. A more recent alternative tool that does resolve these issues is the KAS-seq [21] assay (**K**ethoxal-**A**ssisted **s**sDNA **seq**uencing), based on the highly specific covalent labeling of unpaired guanine residues by N_3-kethoxal. N_3-kethoxal adducts can then be subjected to a click chemistry-mediated biotinylation and specifically enriched and amplified. As most ssDNA in the genome is the result of RNA polymerase bubbles, both elongating and paused, with the rest coming from active replication and some secondary structures, KAS-seq is a convenient tool for mapping polymerase engagement with the genome.

The nature of the KAS method – specifically, the fact that it labels the genome with permanent covalent tags – allows its extension to single-molecule multiomics readouts that capture additional modalities on the same genomic fragments. The KAS-ATAC method that we recently developed [1] enables the mapping of genomic fragments that are both physically accessible and contain ssDNA, together with nucleosome positions and TF footprints within them and their vicinity, by combining the ATAC-seq and KAS-seq assays (Figure 1). This is accomplished by first quickly incubating live cells with N_3-kethoxal, then washing the kethoxal away, and proceeding immediately to the native chromatin transposition step of the ATAC-seq protocol. The resulting Tn5-tagged accessible DNA fragments are purified and biotinylated via a click chemistry reaction, then ssDNA-containing fragments are specifically enriched using a streptavidin pull-down. Finally, PCR amplification is carried out on beads to generate final Illumina-compatible libraries.

This protocol describes the step-by-step execution of the KAS-ATAC assay and the basic computational processing of the resulting datasets.

Materials and reagents

Reagents

1. Tn5 transposase. It can be obtained from the Nextera XT DNA Library Preparation Kit offered by Illumina (cat # FC-131-1024), and also from several other commercial vendors. It can also be made locally following previously published protocols [14], which is the most cost-effective approach, especially for large-scale projects. The oligo sequences needed for transposome assembly are the following:

   Tn5MErev:

   `/5Phos/CTGTCTCTTATACACATCT`

   Tn5ME-A:

   `TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG`

   Tn5ME-B:

   `GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG`

   If homemade Tn5 is used, its activity should be carefully characterized relative to standard enzymatic formulations before production-scale use.

2. Sequencing primers/adapters. PCR and indexing primers/adapters can be obtained from the Nextera XT DNA Library Preparation Kit, or equivalent kits. Alternatively, custom-designed and synthesized oligos can also be used. The i7 primer sequence is:

   `5'-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3'`

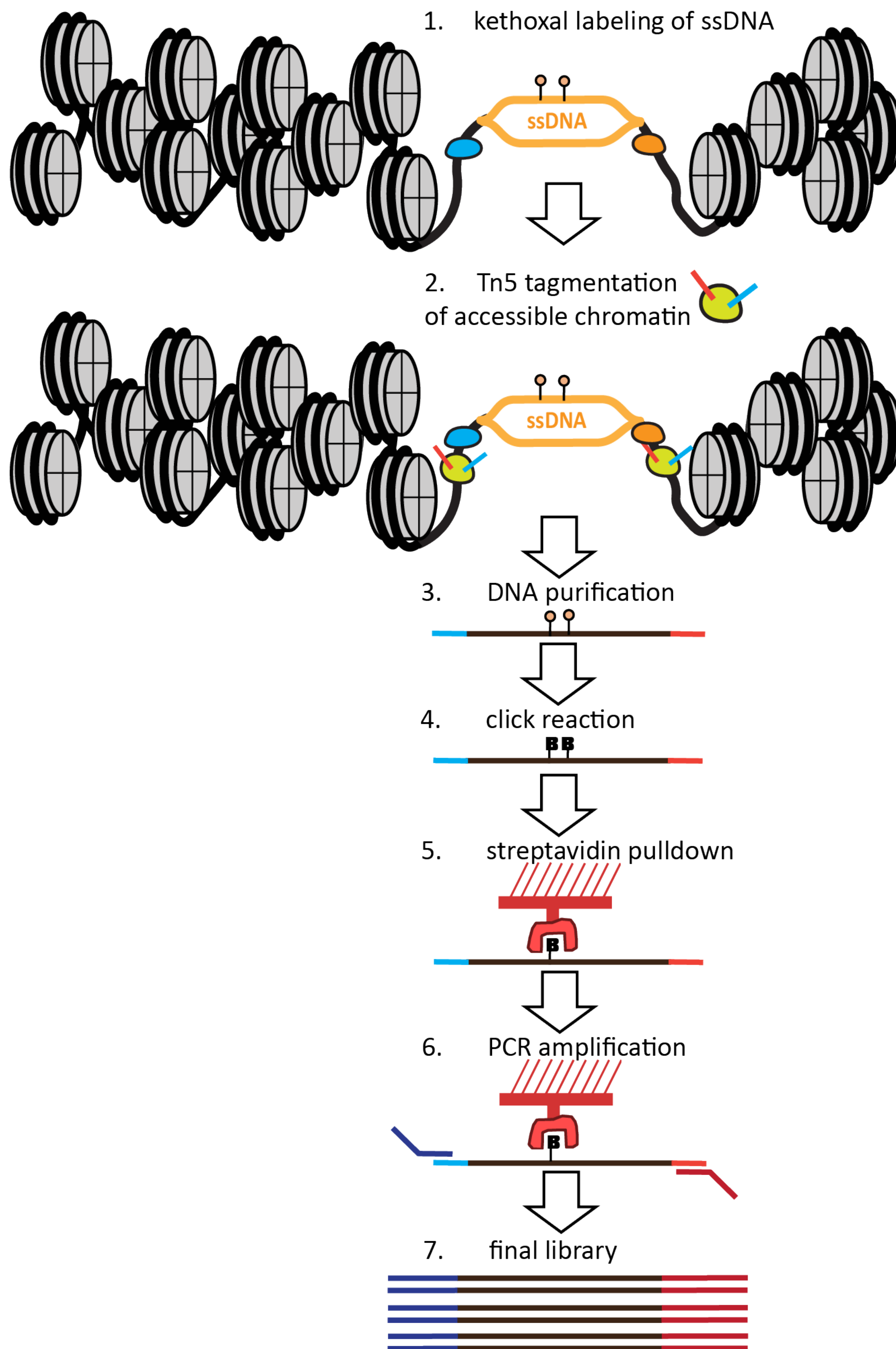


Figure 1: Outline of the KAS-ATAC assay. First, ssDNA is labeled using a quick treatment with $N_3$3-kethoxal. Next, the kethoxal is washed away, nuclei are isolated and native transposition is carried out to label accessible chromatin. DNA is then isolated and subjected to a click reaction to attach biotin to kethoxal adducts. Biotin-labeled DNA fragments are then specifically pulled down using streptavidin and PCR-amplified on beads. Final libraries are sequenced on an Illumina instrument.

The i5 sequence is:

`5'AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC-3'`

Where `[i7]` and `[i5]` are the index sequences (typically 8-bp long). Typical `[i7]` and `[i5]` index sequences are the following.

For `[i7]`:

```
701 TAAGGCGA
702 CGTACTAG
703 AGGCAGAA
704 TCCTGAGC
705 GGACTCCT
706 TAGGCATG
707 CTCTCTAC
708 CAGAGAGG
709 GCTACGCT
710 CGAGGCTG
711 AAGAGGCA
712 GTAGAGGA
713 GTCGTGAT
714 ACCACTGT
715 TGGATCTG
716 CCGTTTGT
717 TGCTGGGT
718 GAGGGGTT
719 AGGTTGGG
720 GTGTGGTG
721 TGGGTTTC
722 TGGTCACA
723 TTGACCCT
724 CCACTCCT
```

For `[i5]`:

```
501 TAGATCGC
502 CTCTCTAT
503 TATCCTCT
504 AGAGTAGA
505 GTAAGGAG
506 ACTGCATA
507 AAGGAGTA
508 CTAAGCCT
```

Dissolve and dilute to 25 μM.

3. N_3-kethoxal (ApexBio, Cat# A8793)
4. Dimethyl Formamide (Sigma, Cat# 227056-100ML)
5. 1× PBS buffer solution pH 7.4 (ThermoFisher Scientific, Cat# 10010031)
6. 10× PBS buffer solution pH 7.4 (ThermoFisher Scientific, Cat# 70011044)
7. 1M Tris-HCl pH 7.4 (ThermoFisher Scientific, Cat# J60202.K2)
8. 5M NaCl (ThermoFisher Scientific, Cat# A57006)
9. 1M $MgCl_2$ (ThermoFisher Scientific, Cat# J61014.AK)
10. IGEPAL CA-630 detergent (Sigma, Cat# 11332465001; supplied as a 10% solution)
11. Tween-20 detergent (Sigma, Cat# 11332465001, supplied as a 10% solution; store at 4 °C)
12. Digitonin detergent (Promega, Cat# G9441, supplied as a 2% solution in DMSO; store at -20 °C))
13. K_3BO_3 (Sigma, Cat# B6768; prepare a carefully pH-adjusted pH 7.0 250 mM solution)
14. 0.5M EDTA, pH 8.0 (ThermoFisher Scientific, Cat# 15575020)
15. DBCO-PEG4-biotin (Sigma, Cat# 760749; dissolve in DMSO to 20 mM)
16. Dynabeads MyOne Streptavidin T1 beads (ThermoFisher Scientific, Cat# 65601)
17. NEBNext High-Fidelity 2× PCR Master Mix (NEB, Cat# M0541S)
18. QuBit 1X dsDNA HS Assay Kit (ThermoFisher Scientific Cat# Q33231)
19. MinElute PCR Purification Kit (Qiagen Cat# 28004/28006), Zymo DNA Clean and Concentrator Kit (Zymo Cat# D4013/D4014), or equivalent
20. Nuclease-free H_2O

Solutions

1. ATAC-RSB buffer (master stock)
2. ATAC-RSB-Lysis buffer (prepare immediately before use)
3. ATAC-RSB-Wash buffer (prepare immediately before use)
4. 2× TD buffer
5. Tween Washing Buffer (TWB)
6. 2× Binding Buffer
7. Transposition Mix (prepare immediately before use)

Recipes

1. ATAC-RSB buffer (master stock)
   - 10 mM Tris-HCl pH 7.4
   - 10 mM NaCl
   - 3 mM $MgCl_2$
2. ATAC-RSB-Lysis buffer
   - 10 mM Tris-HCl pH 7.4
   - 10 mM NaCl
   - 3 mM $MgCl_2$
   - 0.1% IGEPAL CA-630
   - 0.1% Tween-20
   - 0.01% Digitonin
3. Lysis Wash Buffer (ATAC-RSB-wash)
   - 10 mM Tris-HCl pH 7.4
   - 10 mM NaCl
   - 3 mM $MgCl_2$
   - 0.1% Tween-20

4. 2× TD buffer
   20 mM Tris-HCl pH 7.6
   10 mM $MgCl_2$
   20% Dimethyl Formamide
5. 1× TWB
   5 mM Tris-HCl pH 7.4
   0.5 mM EDTA
   1 M NaCl
   0.05% Tween 20
6. 2× Binding Buffer
   10 mM Tris-HCl pH 7.4
   1 mM EDTA
   2 M NaCl
7. Transposition Mix (50 μL volume)
   25 μL 2× TD buffer
   22.5 μL H_2O
   2.5 μL Tn5

Laboratory supplies

1. 1.5-mL microcentrifuge tubes, preferably low protein and DNA binding.
2. 15-mL and 50-mL tubes
3. QuBit-compatible 200-μL PCR tubes/strips
4. TapeStation D1000 or HS D1000 tape and reagents (Agilent) or equivalent, e.g. BioAnalyzer (Agilent))
5. Pipette tips (1,000, 200, and 20 μL)
6. Micropipettes

Equipment

1. Incubator (37 °C), or a Thermomixer.
2. Tabletop centrifuge
3. Thermal cycler
4. Qubit fluorometer or equivalent
5. TapeStation (Agilent) or equivalent, e.g. BioAnalyzer (Agilent).
6. Magnetic stand for 1.5-mL tubes
7. Magnetic stand for 200-μL PCR tubes/strips
8. Rotator for 1.5 mL tubes

Software and datasets

1. `Bowtie` [22] (`http://bowtie-bio.sourceforge.net/index.shtml`) or `Bowtie2` [23] (`http://bowtie-bio.sourceforge.net/bowtie2/index.shtml`).
2. `samtools` [24]: `http://www.htslib.org/`
3. `PicardTools` `https://broadinstitute.github.io/picard/`
4. UCSC Genome Browser [25, 26] utilities: `http://hgdownload.cse.ucsc.edu/admin/exe/`
5. Python (version 2.7 or higher) `https://www.python.org/`
6. Custom Python scripts (`https://github.com/georgimarinov/GeorgiScripts`)

Procedure

Because the KAS-ATAC assay captures a relatively rare population of DNA molecules, i.e. the ones that are both physically accessible and contain ssDNA, it is necessary to start with more cells than used for regular ATAC-seq as input. The typical ATAC-seq protocol of mammalian cells uses 50,000 cells as input [15, 27, 28]; for KAS-ATAC we recommend combining at least 10×50,000-cell reactions into each KAS-ATAC library, as described below.

Part I. Cell culture

Grow cells or generate cell suspensions following the appropriate protocols for the cell line/tissue system one is studying.

Part II. Kethoxal treatment

The first step in the KAS-ATAC protocol involves kethoxal-labeling of ssDNA. This step is done on live cells, and should ideally be carried out maximally quickly in order to achieve as close coupling of the ssDNA and chromatin accessibility measurements on individual DNA molecules as possible. The kethoxal is immediately washed away in order to avoid non-specific labeling throughout subsequent steps.

1. Pellet 500,000 to 1M cells at room temperature by centrifuging at 500 *g* for 5 minutes
2. Resuspend cells in 500 μL of media supplemented with 5 mM N_3-kethoxal (final concentration)
3. Incubate for 5 minutes at 37 °C with shaking at 600 rpm in a Thermomixer
4. Pellet cells at room temperature by centrifuging at 500 *g* for 5 minutes
5. Resuspend cells in 500 μL 1× PBS
6. Pellet cells at room temperature by centrifuging at 500 *g* for 5 minutes

Part III. Cell lysis and transposition

Proceed immediately to the ATAC step of the protocol.

1. Resuspend cells in 50 μL ATAC-RSB-Lysis buffer
2. Incubate cells on ice for 3 minutes.
3. Add 1 mL ATAC-RSB-Wash buffer
4. Centrifuge at 500 g for 5 min at 4 °C
5. Resuspend in 50 μL Transposition Mix per 50,000 cells
6. Redistribute into separate tubes so that each contains 50 μL transposition reaction (i.e. 50,000 cells).
7. Incubate at 37 °C for 30 minutes in a Thermomixer at 1000 rpm
8. Stop each reaction by adding 250 μL PB buffer from the Qiagen MinElute kit
9. Pool the reactions and proceed to purify them using the Qiagen MinElute kit. Load 800 μL of the transposition reaction plus PB buffer mixture onto a MinElute column
10. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge, and discard the flowthrough
11. Repeat this step using the same column for as many times as necessary until the full pooled volume has been loaded onto the column
12. Add 750 μL Qiagen MinElute PE Wash Buffer onto the column
13. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge and discard the flowthrough
14. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge to remove residual wash buffer
15. Elute in 87.5 μL 25 mM K_3BO_3 solution

Part IV. Click reaction

Proceed immediately to the click reaction step. Note that the click reaction product should not be purified using AMPure XP/SPRI beads as is done for the regular KAS-seq protocol as this will result in the loss of the critically important short subnucleosomal fragments corresponding to transcription factor footprints. Use column clean up to make sure those fragments are retained.

1. Mix 87.5 μL of transposed DNA with 2.5 μL 20 mM DBCO-PEG4-biotin and 10 μL 10× PBS for a final volume of 100 μL
2. Incubate at 37 °C for 90 minutes in a Thermomixer
3. Add 500 μL MinElute PB buffer to the reaction.
4. Load onto a onto a MinElute column.
5. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge, and discard the flowthrough
6. Add 750 μL Qiagen MinElute PE Wash Buffer onto the column
7. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge and discard the flowthrough
8. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge to remove residual wash buffer
9. Elute in 300 μL 25 mM K_3BO_3 solution

Part V. Biotin pulldown

In this step biotinylated kethoxal-labeled transposition fragments are specifically pulled down using magnetic streptavidin beads

1. Mix 10 μL of 10 mg/mL Dynabeads MyOne Streptavidin T1 beads with 300 μL TWB
2. Separate beads on a magnetic stand
3. Resuspend beads in 300 μL of 2× Binding Buffer
4. Add 300 μL 25 mM K_3BO_3 solution
5. Incubate at room temperature on a rotator for ≥15 minutes
6. Separate beads on a magnetic stand
7. Add 300 μL of 1× TWB
8. Incubate at 55 °C in a Thermomixer with shaking at 1,000 rpm for 2 minutes
9. Separate beads on a magnetic stand
10. Add 300 μL of 1× TWB again
11. Incubate again at 55 °C in a Thermomixer with shaking at 1,000 rpm for 2 minutes
12. Separate beads on a magnetic stand
13. Resuspend in 20 μL Qiagen EB buffer (MinElute kit)

Part VI. Library generation

Final libraries are generated using the beads with the KAS-ATAC DNA bound to them as input and otherwise typical ATAC-seq settings.

1. Set up a PCR reaction as follows:
   - 20 μL KAS-ATAC DNA-bound beads
   - 2.5 μL of i5 PCR primer
   - 2.5 μL of i7 PCR primer
   - 25 μL NEBNext High-Fidelity 2× PCR Master Mix
2. Amplify DNA for 10 cycles as follows:
   - 72 °C for 3 minutes
   - 98 °C for 30 seconds
   - 10 cycles of:
     - 98 °C for 10 seconds
     - 63 °C for 30 seconds
     - 72 °C for 30 seconds
   - Hold at 4 °C

3. Separate beads on a magnetic stand
4. Add 250 μL MinElute PB buffer to the supernatant
5. Load onto a MinElute column
6. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge, and discard the flowthrough
7. Add 750 μL Qiagen MinElute PE Wash Buffer onto the column
8. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge and discard the flowthrough
9. Centrifuge at 14,000 rpm for 1 minutes in a tabletop centrifuge to remove residual wash buffer
10. Elute in 25 EB Buffer
11. Measure final library concentration using Qubit
12. Evaluate fragment distribution in the final libraries using TapeStation (D1000 or HSD1000 kits)

Part VII. Sequencing

Sequence KAS-ATAC libraries using any available Illumina instrument in a paired-end mode. Reads of length 2 *times* 36 bp are generally sufficient given that ATAC-seq libraries are often dominated by subnucleosomal fragments $\leq$100 bp in length [28].

Data analysis

The processing of KAS-ATAC datasets does not differ much from that of standard ATAC-seq sequencing data. Readers are referred to our previous detailed treatment of the latter subject (Marinov & Shipony 2021 [29]) for all details regarding peak calling, differential accessibility analysis, transcription factor footprinting, and other downstream tasks. Here we describe the basic processing steps from raw FASTQ files to alignment BAM files and genome browser tracks.

Part I. Preparation of genomic files

Here we describe processing human KAS-ATAC data against the `hg38` version of the human genome assembly. The commands listed below are to be changed accordingly if other species are studied and/or genome assemblies are used.

1. Download and unzip genome reference files:

```
wget https://www.encodeproject.org/files/
GRCh38_no_alt_analysis_set_GCA_000001405.15/
@@download/
GRCh38_no_alt_analysis_set_GCA_000001405.15.fasta.gz
 -O hg38_no_alt.fasta.gz

gunzip hg38_no_alt.fasta.gz
```

   Note that it is important to use a version of the assembly that does not include alternative haplotypes. The presence of alternative haplotypes in the index makes the homologous sequences on the main chromosomes no longer uniquely mappable and thus artificially "invisible" to downstream analysis.

2. Generate a `bowtie` genome index file:

```
mkdir genomes/hg38/bowtie-indexes
cd genomes/hg38/bowtie-indexes
ln ../hg38_no_alt.fa
bowtie-build -f hg38_no_alt.fa hg38_no_alt
```

3. Generate a `bowtie` "chrM" index file using only the mitochondrial genome as input:

```
mkdir genomes/hg38/bowtie-indexes
cd genomes/hg38/bowtie-indexes
ln ../chrM.fa
bowtie-build -f chrM.fa chrM
```

4. Generate a chromosome sizes info (`chrom.sizes`) file:

```
python makeChromSizesFromFasta.py
              hg38_no_alt.fa hg38_no_alt.chrom.sizes
```

   Chromosome-size files identify the end points of chromosomes/contigs and are used at multiple steps in high-throughput sequencing data processing. They consist of one line per chromosome as follows:

```
chr <tab> chromosome_size
```

Part II. Read mapping and filtering

ATAC-seq sequencing data, and thus also KAS-ATAC data is recommended to be processed in a 2×36mers format. This is so because of the aforementioned abundance of short fragments and in order to maintain uniformity across all datasets we work with in terms of guarding against possible mappability and alignment biases.

For certain purposes (e.g. examining the effect of sequence variants) it can be beneficial to use the full length of fragments. In such cases reads have to be trimmed of adapters, which can be done using Trimmomatic [30] or TrimGalore/Cutadapt [31].

For Bowtie mapping, use the following commands.

1. Map 2×36mer reads to whole genome:

```
python PEFastqToTabDelimited.py
  SAMPLE.end1.fastq.gz SAMPLE.end2.fastq.gz |
  bowtie hg38/bowtie-indexes/hg38_no_alt
  -p 16 -v 2 -k 2 -m 1   -t --best --strata
  -q --sam-nh -X 1000 --sam --12 -
  | egrep -v chrM |
  samtools view -F 4 -bT
  hg38/sequence/hg38_no_alt.fa - |
  samtools sort - SAMPLE.2x36mers.unique.nochrM
```

This retains uniquely mapping read pairs with up to 2 mismatches relative to the reference and filters out unaligned reads as well as reads mapping to the mitochondrial genome. Alignments are stored in the BAM format (a binary version of the SAM format [24]).

Change the `egrep -v chrM` part if the name of the mitochondrial contig is different, or other such highly accessible contigs (e.g. plastids if working with plants) are to be also filtered out. It is important to remove these alignments from the final BAM files because the fraction of mitochondrial reads can vary substantially between libraries and leaving these reads in can affects global normalization factors in undesirable ways.

2. Map 2×36mer reads to the mitochondrial genome. This step is necessary for the purpose of evaluating the fraction of mitochondrial reads present in the sequenced libraries, and is a standard part of ATAC-seq processing. Note that calculating this fraction from an alignment against the joint nuclear and mitochondrial index does not estimate it accurately because the mitochondrial genome is not uniquely mappable in several stretches due to the presence of mitochondrial insertions in the nuclear genome (so called NUMTs [32]).

```
python PEFastqToTabDelimited.py
  SAMPLE.end1.fastq.gz SAMPLE.end2.fastq.gz |
  bowtie hg38/bowtie-indexes/chrM -p 16 -v 2
  -a -t --best --strata -q --sam-nh -X 1000
  --sam --12 - | samtools view -F 4 -bT
  hg38/sequence/hg38_no_alt.fa - |
  samtools sort - SAMPLE.2x36mers.chrM
```

3. Index BAM files with `samtools`:

```
samtools index SAMPLE.2x36mers.unique.bam
samtools index SAMPLE.2x36mers.chrM.bam
```

4. Remove duplicate alignments. Because both ATAC-seq and KAS-ATAC capture only a limited initial population of original molecules, and because they are sequenced in a paired-end format, fragments with exactly the same coordinates are likely to represent PCR duplicates. They are thus by default removed, using the `MarkDuplicates` program in the `PicardTools` suite, as follows:

```
module load java; java -Xmx4G -jar
picard-tools-1.99/MarkDuplicates.jar
INPUT=SAMPLE.2x36mers.unique.nochrM.bam
OUTPUT=SAMPLE.2x36mers.unique.nochrM.dedup.bam
METRICS_FILE=SAMPLE.2x36mers.unique.nochrM.dedup.metrics
VALIDATION_STRINGENCY=LENIENT
ASSUME_SORTED=true REMOVE_DUPLICATES=true
```

5. Index the resulting deduplicated BAM file:

```
samtools index SAMPLE.2x36mers.unique.nochrM.dedup.bam
```

Part III. Generating genome browser visualization tracks

Two types of tracks can be generated for ATAC-seq/KAS-ATAC dataa: a "coverage" track assigning a score for each base in the genome that a mapped fragment covers, and "5′" tracks, which only represent Tn5 insertion points. The latter are used for transcription factor footprinting analysis, and need to be shifted by ±4 bp depending on which strand they map to as the transposase itself has a footprint of 8-9 base pairs [15, 33].

It is optimal to normalize the tracks relative to the total set of mapped and dedupped reads in RPM (Reads Per Million mapped reads) units.

1. Generate RPM-normalized coverage tracks (using the `bamCoverage` program in `deepTools`):

```
python makewigglefromBAM-NH.py track_title
SAMPLE.2x36mers.unique.nochrM.dedup.bam
hg38.chrom.sizes
SAMPLE.2x36mers.unique.nochrM.dedup.bam.wig
-uniqueBAM -notitle -RPM
```

2. Generate RPM-normalized "5′" tracks:

```
python make5primeWigglefromBAM-NH.py track_title
SAMPLE.2x36mers.unique.nochrM.dedup.bam
hg38.chrom.sizes
SAMPLE.2x36mers.unique.nochrM.dedup.bam.5p.wig
-uniqueBAM -notitle -RPM -shift 4
```

3. Convert `bedGraph` files to `bigWig`:

```
UCSC-utils/wigToBigWig
SAMPLE.2x36mers.unique.nochrM.dedup.bam.wig
hg38.chrom.sizes
SAMPLE.2x36mers.unique.nochrM.dedup.bam.bigWig
```

```
UCSC-utils/wigToBigWig
SAMPLE.2x36mers.unique.nochrM.dedup.bam.5p.wig
hg38.chrom.sizes
SAMPLE.2x36mers.unique.nochrM.dedup.bam.5p.bigWig
```

Part IV. Calculating mapping statistics and evaluating enrichment

It is important to evaluate the extent of enrichment observed in sequenced libraries as well as to collect general mapping and library complexity statistics.

1. Count raw reads:

```
zcat SAMPLE.fastq.gz | wc -l
```

Divide by 4 to get the number of reads (as each read is represented by 4 lines in a FASTQ file).

2. Calculate mapping statistics for the chrM-mapping BAM file:

```
python SAMstats.py
SAMPLE.2x36mers.chrM.bam
SAMstats-SAMPLE.2x36mers.chrM
-bam hg38.chrom.sizes samtools -paired
```

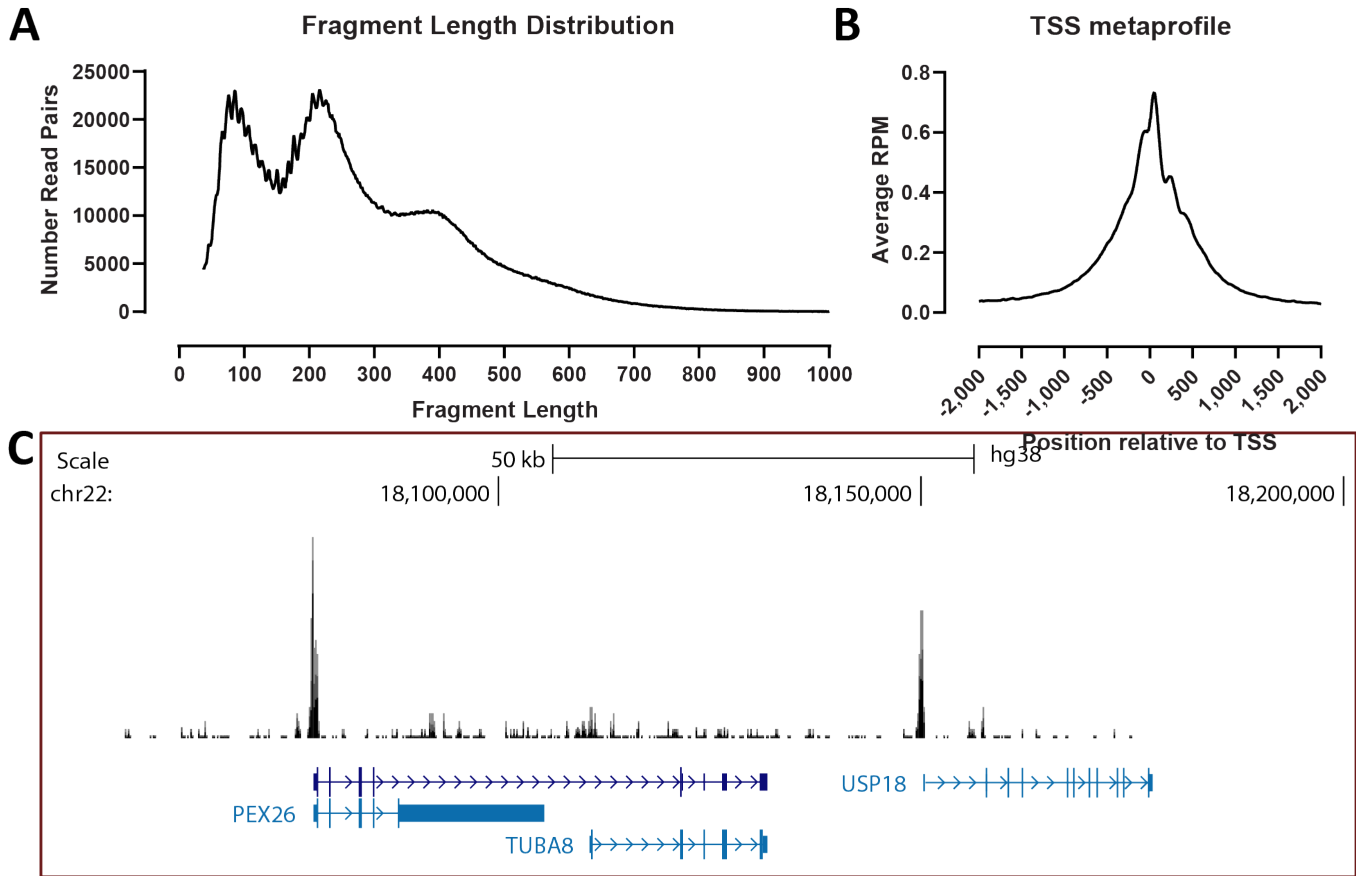


Figure 2: Typical KAS-ATAC results. A) Fragment length distribution; B) TSS metaprofile C) Representative genome browser snapshot.

The total number of reads mapping to the mitochondrion ($|R_M|$) can be obtained from this step.

3. Calculate mapping statistics for the chrM-filtered pre-deduplication BAM file:

```
python SAMstats.py
SAMPLE.2x36mers.unique.nochrM.bam
SAMstats-SAMPLE.2x36mers.unique.nochrM
-bam hg38.chrom.sizes samtools -paired
```

The total number of reads mapping to the nuclear genome ($|R_N|$) can be obtained from this file.

This step also provides the Non-Redundant read Fraction NRF estimate of library, defined as follows [34]:

$$NRF = U_P/U_R \quad (1)$$

Where U_P is the set of genomic positions to which $5'$ ends of reads map uniquely and U_R is the total number of uniquely mapped reads.

4. Calculate mapping statistics for the chrM-filtered post-deduplication BAM file:

```
python SAMstats.py
SAMPLE.2x36mers.unique.nochrM.dedup.bam
SAMstats-SAMPLE.2x36mers.unique.nochrM.dedup
-bam hg38.chrom.sizes samtools -paired
```

5. Calculate the extent of mitochondrial contamination.

The fraction of mitochondrial reads is calculated as follows:

$$MRF = \frac{|R_M|}{|R_M| + |R_N|} \quad (2)$$

Where R_M and R_N are as defined above.

6. Estimate the fragment length distribution based on the chrM-filtered post-deduplication BAM file. Note that including the chrM-mapping fragments can result in misleading results as mitochondria lack nucleosomes.

```
python PEInsertDistFromBAM.py
SAMPLE.2x36mers.unique.nochrM.dedup.bam
hg38.chrom.sizes
SAMPLE.2x36mers.unique.nochrM.dedup.InsLen
-uniqueBAM -normalize
```

7. Evaluate the degree of enrichment for open chromatin using the TSS (transcription start site) enrichment TSS_E metric as described previously [29]. This metric is based on creating an aggregate-profile curve around transcription start sites of protein coding genes, then calculating the ratio of the average signal in small (e.g. 100-bp radius) windows around the TSS versus the combined average signal in the two 100-bp long windows on the TSS flanks at a distance of 2 kbp:

$$TSS_E = \frac{|R \in [TSS \pm 100]|}{|R \in [TSS - 2050, TSS - 1950]| + |R \in [TSS + 1950, TSS + 2050]|} \quad (3)$$

8. As a one-time step, create a TSS 0-radius BED file using a genome annotation GTF file::

```
python TSS_bed_FromGTF.py annotation.gtf
0 0 annotation.TSS-0bp.bed
```

9. Generate an average profile around TSSs:

```
python signalAroundCoordinate-BW.py
refSeq.TSS-0bp.bed 0 1 3 4000
SAMPLE.2x36mers.unique.nochrM.dedup.coverage.bigWig
SAMPLE.2x36mers.unique.nochrM.dedup.coverage.TSS_profile
 -normalize
```

10. Then calculate TSS_E values:

```
python ATACTSSscore.py
SAMPLE.2x36mers.unique.nochrM.dedup.coverage.TSS_profile
 100 2000 >> ATACTSSscore.txt
```

Representative KAS-ATAC fragment length distributions, TSS metaprofiles, and genome browser coverage tracks are shown in Figure 2A, B, and C, respectively.

Validation of protocol

This protocol was developed and used in the following research article(s):

- Kim SH, Marinov GK, Greenleaf W [1]. KAS-ATAC reveals the genome-wide single-stranded accessible chromatin landscape of the human genome. *Genome Res.*

KAS-ATAC experiments were carried out on GM12878 and HEK283 cells in two replicates, and benchmarked against matching ATAC-seq and KAS-seq datasets, as well as ATAC-seq datasets generated using a biotinylated Tn5 transposome and subsequent streptavidin pull down.

General notes and troubleshooting

General notes

1. Make sure the Tn5 you are using is well characterized for activity if homemade, by carrying out ATAC-seq experiments on standard cell lines and calculating TSS enrichment QC metrics.

2. Also make sure you are using non-degraded N_3-kethoxal, i.e. preferably fresh recently ordered (within 3-4 months) stocks.

3. Make sure the input material, whether nuclei isolated from tissues or cell lines, is in optimal condition, i.e. as few dead cells as possible.

Troubleshooting

Problem 1: Low library complexity is observed.

Possible cause A: Too few cells were used.

Solution A: Repeat the experiment with at least 10×50,000 cells as recommended above. It is, however, possible, that in some other cell types and systems the abundance of accessible and ssDNA containing DNA regions is lower than in the human cell lines we have benchmarked KAS-ATAC on, and thus the size of the input needs to be increased.

Possible cause B: Inefficient kethoxal labeling.

Solution B: Redo experiments with a fresh N_3-kethoxal batch.

Possible cause C: Inefficient click reaction and/or degradation of kethoxal adducts during storage.

Solution C: Redo experiments with fresh and carefully pH-calibrated K_3BO_3 stocks.

Problem 2: High abundance of mitochondrial fragments.

Possible cause A: Poor separation of nuclei from cytoplasm.

Solution A: Redo experiments while being extremely careful to remove all of the supernatant after cell lysis.

Problem 3: Poor enrichment observed

Possible cause A: Poor-quality Tn5

Solution A: Redo experiments with freshly prepared and well characterized Tn5 transposome stocks.

Possible cause B: Too many dead cells resulting in too much naked and broken up DNA present.

Solution B: Redo experiments with cells the viability of which has been measured and assured to be high. A pre-treatment with DNase I of the cell culture as described in Corces et al. 2017 [27] might also help.

Acknowledgments

This work was supported by NIH grants (P50HG007735, RO1 HG008140, U19AI057266, UM1HG009442 and 1UM1HG009436 to W.J.G., the Rita Allen Foundation (to W.J.G.), the Baxter Foundation Faculty Scholar Grant, and the Human Frontiers Science Program grant RGY006S (to W.J.G). W.J.G. also acknowledges support by grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative.

The authors would like to thank Samuel Kim for his contributions during the development of the KAS-ATAC assay, and members of the Greenleaf, and Kundaje labs for helpful discussions and suggestions.

Competing interests

The authors declare no competing interests.

References

1. Kim SH, Marinov GK, Greenleaf W. 2024. KAS-ATAC reveals the genome-wide single-stranded accessible chromatin landscape of the human genome. *Genome Res* gr.279621.124. doi: 10.1101/gr.279621.124
2. ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, Khatun J, Lajoie BR, Landt SG, Lee BK, Pauli F, Rosenbloom KR, Sabo P, Safi A, Sanyal A, Shoresh N, Simon JM, Song L, Trinklein ND, Altshuler RC, Birney E, Brown JB, Cheng C, Djebali S, Dong X, Dunham I, Ernst J, Furey TS, Gerstein M, Giardine B, Greven M, Hardison RC, Harris RS, Herrero J, Hoffman MM, Iyer S, Kelllis M, Khatun J, Kheradpour P, Kundaje A, Lassman T, Li Q, Lin X, Marinov GK, Merkel A, Mortazavi A, Parker SC, Reddy TE, Rozowsky J, Schlesinger F, Thurman RE, Wang J, Ward LD, Whitfield TW, Wilder SP, Wu W, Xi HS, Yip KY, Zhuang J, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M, Pazin MJ, Lowdon RF, Dillon LA, Adams LB, Kelly CJ, Zhang J, Wexler JR, Green ED, Good PJ, Feingold EA, Bernstein BE, Birney E, Crawford GE, Dekker J, Elinitski L, Farnham PJ, Gerstein M, Giddings MC, Gingeras TR, Green ED, Guigó R, Hardison RC, Hubbard TJ, Kellis M, Kent WJ, Lieb JD, Margulies EH, Myers RM, Snyder M, Starnatoyannopoulos JA, Tennebaum SA, Weng Z, White KP, Wold B, Khatun J, Yu Y, Wrobel J, Risk BA, Gunawardena HP, Kuiper HC, Maier CW, Xie L, Chen X, Giddings MC, Bernstein BE, Epstein CB, Shoresh N, Ernst J, Kheradpour P, Mikkelsen TS, Gillespie S, Goren A, Ram O, Zhang X, Wang L, Issner R, Coyne MJ, Durham T, Ku M, Truong T, Ward LD, Altshuler RC, Eaton ML, Kellis M, Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Batut P, Bell I, Bell K, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena HP, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Li G, Luo OJ, Park E, Preall JB, Presaud K, Ribeca P, Risk BA, Robyr D, Ruan X, Sammeth M, Sandu KS, Schaeffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Hayashizaki Y, Harrow J, Gerstein M, Hubbard TJ, Reymond A, Antonarakis SE, Hannon GJ, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR, Rosenbloom KR, Sloan CA, Learned K, Malladi VS, Wong MC, Barber GP, Cline MS, Dreszer TR, Heitner SG, Karolchik D, Kent WJ, Kirkup VM, Meyer LR, Long JC, Maddren M, Raney BJ, Furey TS, Song L, Grasfeder LL, Giresi PG, Lee BK, Battenhouse A, Sheffield NC, Simon JM, Showers KA, Safi A, London D, Bhinge AA, Shestak C, Schaner MR, Kim SK, Zhang ZZ, Mieczkowski PA, Mieczkowska JO, Liu Z, McDaniell RM, Ni Y, Rashid NU, Kim MJ, Adar S, Zhang Z, Wang T, Winter D, Keefe D, Birney E, Iyer VR, Lieb JD, Crawford GE, Li G, Sandhu KS, Zheng M, Wang P, Luo OJ, Shahab A, Fullwood MJ, Ruan X, Ruan Y, Myers RM, Pauli F, Williams BA, Gertz J, Marinov GK, Reddy TE, Vielmetter J, Partridge EC, Trout D, Varley KE, Gasper C, Bansal A, Pepke S, Jain P, Amrhein H, Bowling KM, Anaya M, Cross MK, King B, Muratet MA, Antoshechkin I, Newberry KM, McCue K, Nesmith AS, Fisher-Aylor KI, Pusey B, DeSalvo G, Parker SL, Balasubramanian S, Davis NS, Meadows SK, Eggleston T, Gunter C, Newberry JS, Levy SE, Absher DM, Mortazavi A, Wong WH, Wold B, Blow MJ, Visel A, Pennachio LA, Elnitski L, Margulies EH, Parker SC, Petrykowska HM, Abyzov A, Aken B, Barrell D, Barson G, Berry A, Bignell A, Boychenko V, Bussotti G, Chrast J, Davidson C, Derrien T, Despacio-Reyes G, Diekhans M, Ezkurdia I, Frankish A, Gilbert J, Gonzalez JM, Griffiths E, Harte R, Hendrix DA, Howald C, Hunt T, Jungreis I, Kay M, Khurana E, Kokocinski F, Leng J, Lin MF, Loveland J, Lu Z, Manthravadi D, Mariotti M, Mudge J, Mukherjee G, Notredame C, Pei B, Rodriguez JM, Saunders G, Sboner A, Searle S, Sisu C, Snow C, Steward C, Tanzer A, Tapanan E, Tress ML, van Baren MJ, Walters N, Washieti S, Wilming L, Zadissa A, Zhengdong Z, Brent M, Haussler D, Kellis M, Valencia A, Gerstein M, Raymond A, Guigó R, Harrow J, Hubbard TJ, Landt SG, Frietze S, Abyzov A, Addleman N, Alexander RP, Auerbach RK, Balasubramanian S, Bettinger K, Bhardwaj N, Boyle

AP, Cao AR, Cayting P, Charos A, Cheng Y, Cheng C, Eastman C, Euskirchen G, Fleming JD, Grubert F, Habegger L, Hariharan M, Harmanci A, Iyenger S, Jin VX, Karczewski KJ, Kasowski M, Lacroute P, Lam H, Larnarre-Vincent N, Leng J, Lian J, Lindahl-Allen M, Min R, Miotto B, Monahan H, Moqtaderi Z, Mu XJ, O'Geen H, Ouyang Z, Patacsil D, Pei B, Raha D, Ramirez L, Reed B, Rozowsky J, Sboner A, Shi M, Sisu C, Slifer T, Witt H, Wu L, Xu X, Yan KK, Yang X, Yip KY, Zhang Z, Struhl K, Weissman SM, Gerstein M, Farnham PJ, Snyder M, Tenebaum SA, Penalva LO, Doyle F, Karmakar S, Landt SG, Bhanvadia RR, Choudhury A, Domanus M, Ma L, Moran J, Patacsil D, Slifer T, Victorsen A, Yang X, Snyder M, White KP, Auer T, Centarin L, Eichenlaub M, Gruhl F, Heerman S, Hoeckendorf B, Inoue D, Kellner T, Kirchmaier S, Mueller C, Reinhardt R, Schertel L, Schneider S, Sinn R, Wittbrodt B, Wittbrodt J, Weng Z, Whitfield TW, Wang J, Collins PJ, Aldred SF, Trinklein ND, Partridge EC, Myers RM, Dekker J, Jain G, Lajoie BR, Sanyal A, Balasundaram G, Bates DL, Byron R, Canfield TK, Diegel MJ, Dunn D, Ebersol AK, Ebersol AK, Frum T, Garg K, Gist E, Hansen RS, Boatman L, Haugen E, Humbert R, Jain G, Johnson AK, Johnson EM, Kutyavin TM, Lajoie BR, Lee K, Lotakis D, Maurano MT, Neph SJ, Neri FV, Nguyen ED, Qu H, Reynolds AP, Roach V, Rynes E, Sabo P, Sanchez ME, Sandstrom RS, Sanyal A, Shafer AO, Stergachis AB, Thomas S, Thurman RE, Vernot B, Vierstra J, Vong S, Wang H, Weaver MA, Yan Y, Zhang M, Akey JA, Bender M, Dorschner MO, Groudine M, MacCoss MJ, Navas P, Stamatoyannopoulos G, Kaul R, Dekker J, Stamatoyannopoulos JA, Dunham I, Beal K, Brazma A, Flicek P, Herrero J, Johnson N, Keefe D, Lukk M, Luscombe NM, Sobral D, Vaquerizas JM, Wilder SP, Batzoglou S, Sidow A, Hussami N, Kyriazopoulou-Panagiotopoulou S, Libbrecht MW, Schaub MA, Kundaje A, Hardison RC, Miller W, Giardine B, Harris RS, Wu W, Bickel PJ, Banfai B, Boley NP, Brown JB, Huang H, Li Q, Li JJ, Noble WS, Bilmes JA, Buske OJ, Hoffman MM, Sahu AO, Kharchenko PV, Park PJ, Baker D, Taylor J, Weng Z, Iyer S, Dong X, Greven M, Lin X, Wang J, Xi HS, Zhuang J, Gerstein M, Alexander RP, Balasubramanian S, Cheng C, Harmanci A, Lochovsky L, Min R, Mu XJ, Rozowsky J, Yan KK, Yip KY, Birney E. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414):57–74.

3. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, Kawli T, Davis CA, Dobin A, Kaul R, Halow J, Van Nostrand EL, Freese P, Gorkin DU, Shen Y, He Y, Mackiewicz M, Pauli-Behn F, Williams BA, Mortazavi A, Keller CA, Zhang XO, Elhajjajy SI, Huey J, Dickel DE, Snetkova V, Wei X, Wang X, Rivera-Mulia JC, Rozowsky J, Zhang J, Chhetri SB, Zhang J, Victorsen A, White KP, Visel A, Yeo GW, Burge CB, Lécuyer E, Gilbert DM, Dekker J, Rinn J, Mendenhall EM, Ecker JR, Kellis M, Klein RJ, Noble WS, Kundaje A, Guigó R, Farnham PJ, Cherry JM, Myers RM, Ren B, Graveley BR, Gerstein MB, Pennacchio LA, Snyder MP, Bernstein BE, Wold B, Hardison RC, Gingeras TR, Stamatoyannopoulos JA, Weng Z. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**(7818):699–710.
4. Wu C. 1980. The 5′ ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**(5776):854–860.
5. Keene MA, Corces V, Lowenhaupt K, Elgin SC. 1981. DNase I hypersensitive sites in Drosophila chromatin occur at the 5′ ends of regions of transcription. *Proc Natl Acad Sci U S A* **78**(1):143–146.
6. McGhee JD, Wood WI, Dolan M, Engel JD, Felsenfeld G. 1981. A 200 base pair region at the 5′ end of the chicken adult β-globin gene is accessible to nuclease digestion. *Cell* **27**(1 Pt 2):45–55.
7. Dorschner MO, Hawrylycz M, Humbert R, Wallace JC, Shafer A, Kawamoto J, Mack J, Hall R, Goldy J, Sabo PJ, Kohli A, Li Q, McArthur M, Stamatoyannopoulos JA. 2004. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods* **1**(3):219–225.
8. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, Cao H, Yu M, Rosenzweig E, Goldy J, Haydock A, Weaver M, Shafer A, Lee K, Neri F, Humbert R, Singer MA, Richmond TA, Dorschner MO, McArthur M, Hawrylycz M, Green RD, Navas PA, Noble WS, Stamatoyannopoulos JA. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat Methods* **3**(7):511–518.
9. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, Zhou D, Luo S, Vasicek TJ, Daly MJ, Wolfsberg TG, Collins FS. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**(1):123–131.
10. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2):311–322.
11. Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, Jones PA. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res* **22**(12):2497–2506.
12. Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. 2017. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell* **67**(3):411–422.e4.
13. Shipony Z, Marinov GK, Swaffer MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, Green-

leaf WJ. 2020. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat Methods* **17**(3):319–327.

14. Picelli S, Björklund AK, Reinius B, Sagasser S, Winberg G, Sandberg R. 2014. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* **24**(12):2033–2040.
15. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**(12):1213–1218.
16. Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. 2015. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* **25**(11):1757–1770.
17. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**(7295):182–187
18. Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**(5909):1845–1848.
19. Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**(6122):950–953.
20. Tome JM, Tippens ND, Lis JT. 2018. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat Genet* 50(11):1533-1541
21. Wu T, Lyu R, You Q, He C. 2020. Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity *in situ*. *Nat Methods* **17**(5):515–523.
22. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25.
23. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**(4):357–359.
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16):2078–2079.
25. Kuhn RM, Haussler D, Kent WJ (2013) The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144–161.
26. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**(17):2204–2207.
27. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, Kathiria A, Cho SW, Mumbach MR, Carter AC, Kasowski M, Orloff LA, Risca VI, Kundaje A, Khavari PA, Montine TJ, Greenleaf WJ, Chang HY. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**(10):959–962.
28. Marinov GK, Shipony Z, Kundaje A, Greenleaf WJ. 2023. Genome-Wide Mapping of Active Regulatory Elements Using ATAC-seq. *Methods Mol Biol* **2611**:3–19.
29. Marinov GK, Shipony Z. 2021. Interrogating the Accessible Chromatin Landscape of Eukaryote Genomes Using ATAC-seq. *Methods Mol Biol* **2243**:183–226.
30. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15):2114–2120.
31. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**(1):10–12.
32. Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* **6**(2):e1000834.
33. Pampari A, Shcherbina A, Kvon E, Kosicki M, Nair S, Kundu S, Kathiria AS, Risca VI, Kuningas K, Alasoo K, Greenleaf WJ, Pennacchio LA, Kundaje A. 2024. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal *cis*-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv* 2024.12.25.630221
34. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**(9):1813–1831.