Title

Generalization of the sci-L3 method to achieve high-throughput linear amplification for replication template strand sequencing, genome conformation capture, and the joint profiling of RNA and chromatin accessibility.

Journal

Nucleic Acids Research (NAR), 53(4)

Authors

Chovanec, Peter
Yin, Yi

Publication Date

2025-02-08

DOI

10.1093/nar/gkaf101

Peer reviewed

Nucleic Acids Research, 2025, **53**, gkaf101
https://doi.org/10.1093/nar/gkaf101
Methods

Generalization of the sci-L3 method to achieve high-throughput linear amplification for replication template strand sequencing, genome conformation capture, and the joint profiling of RNA and chromatin accessibility

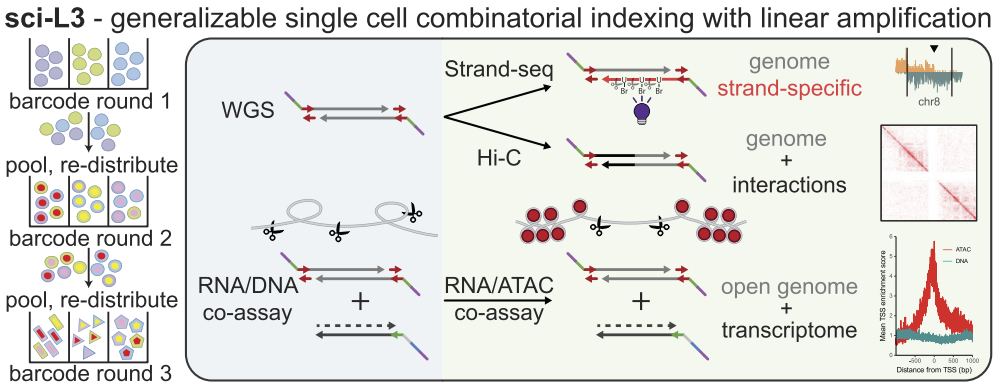
Peter Chovanec and Yi Yin*

Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, United States
*To whom correspondence should be addressed. Email: yiyin@mednet.ucla.edu

Abstract

Single-cell combinatorial indexing (sci) methods have addressed major limitations of throughput and cost for many single-cell modalities. With the incorporation of linear amplification and three-level barcoding in our suite of methods called sci-L3, we further addressed the limitations of uniformity in single-cell genome amplification. Here, we build on the generalizability of sci-L3 by extending it to template strand sequencing (sci-L3-Strand-seq), genome conformation capture (sci-L3-Hi-C), and the joint profiling of RNA and chromatin accessibility (sci-L3-RNA/ATAC). We demonstrate the ease of adapting sci-L3 to these new modalities by only requiring a single-step modification of the original protocol. As a proof of principle, we show our ability to detect sister chromatid exchanges, genome compartmentalization, and cell state-specific features in thousands of single cells. We anticipate sci-L3 to be compatible with additional modalities, including DNA methylation (sci-MET) and chromatin-associated factors (CUT&Tag), and ultimately enable a multi-omics readout of them.

Graphical abstract



Introduction

The development and utilization of single-cell genomic assays has transformed our understanding of rare and heterogeneous (diverse) events within biological systems. In the area of single-cell whole genome sequencing (scWGS), methods have commonly utilized bias-prone amplification approaches with limited scalability. To address these challenges, we have previously described the sci-L3 suite of methods for single-cell combinatorial indexing (sci) with linear amplification (L) and three-level barcoding [1] that includes sci-L3-WGS, sci-L3-target-seq, and sci-L3-RNA/DNA-seq [2].

As a review of the sci-L3 workflow (Fig. 1), (i) millions of fixed nuclei undergo nucleosome depletion to enable efficient *in situ* enzymatic chromosome fragmentation and (ii) the nuclei are then split into pools of tens of thousands of nuclei. Each pool undergoes subsequent "tagmentation," i.e. Tn5 transposome-mediated DNA fragmentation, while the fragment ends are tagged with unique first rounds of DNA

Received: October 30, 2024. **Revised:** December 28, 2024. **Editorial Decision:** January 28, 2025. **Accepted:** February 5, 2025

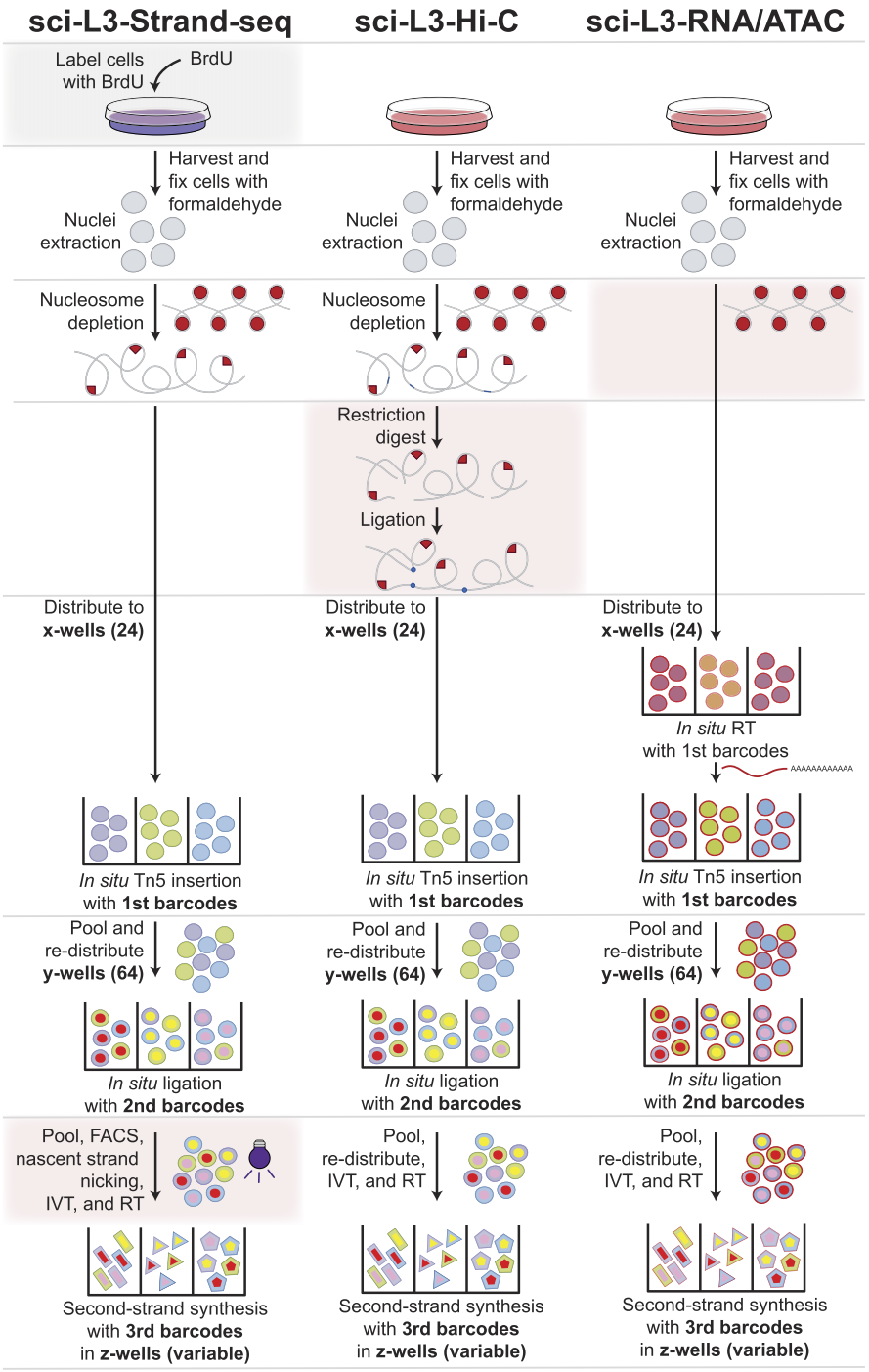


Figure 1. Overview of the sci-L3 method extensions. Each extension has only a single-step modification from the original method, shaded in its respective section along with the preparation step at the beginning (details in text for an overview of the sci-L3 key steps). The cells are fixed with formaldehyde, and then undergo nuclei extraction and nucleosome depletion. The subsequent three levels of barcodes are introduced with a split and pool scheme using Tn5 insertion, ligation, and second-strand synthesis. IVT, *in vitro* transcription; RT, reverse transcription; UDG, uracil DNA glycosylase; EndoVIII, endonuclease VIII; re-distribution can by done by fluorescence-activated cell sorting (FACS) or by dilution except for sci-L3-Strand-seq.

barcodes, typically 24–96 barcodes; (iii) tagmented nuclei are pooled and split into new pools; second-round barcodes are directionally ligated upstream of the first round of the barcode and downstream of a T7 RNA polymerase promoter for *in vitro* transcription (IVT)-based linear amplification; and (iv) cells are then pooled again and sorted into new wells for linear amplification. The amplified RNA molecules are converted to double-stranded complementary DNA (cDNA) with unique third-round barcodes for each amplification pool of cells, compatible with various library preparation workflows.

The advantages of sci-L3 are three-fold: (i) the three successive rounds of split and pool barcoding significantly increase throughput over conventional methods and yield a low per-cell cost; (ii) linear amplification through IVT maintains coverage uniformity by avoiding polymerase chain reaction (PCR)- or multiple displacement amplification (MDA)-based expo-

nential amplification bias; and (iii) the generalizable scheme can be applied to various modalities, such as WGS or targeted sequencing, and to the simultaneous readout of multiple modalities, such as transcriptome and WGS from the same single cell. In principle, sci-L3 can be applied to other genomic assays beyond scWGS [3]. Here we describe the extension of sci-L3 to template strand sequencing (Strand-seq), genome conformation capture (Hi-C), and the joint profiling of RNA and chromatin accessibility (RNA/ATAC).

Strand-seq is a technique that enables the identification of sister chromatid exchanges (SCEs) and structural variations (SV), and allows the phasing of heterozygous single nucleotide variants (SNVs) in diploid genomes [1, 4–7]. This is achieved by only sequencing the template strand of DNA replication for a given chromosome (Watson or Crick, "W" or "C") in the subsequent cell division. As a result, to adapt sci-L3 to Strand-seq requires that our method specifically sequences the template strand and maintains its strand directionality throughout. sci-L3 uses the Tn5 transposome for genome fragmentation and the introduction of the first combinatorial barcode. The Tn5 transposases catalyze single-stranded strand transfer and thus the original directionality of the replication template strand is preserved in the first round of cell barcoding [8]. Similarly, the following molecular features of sci-L3 retain strand directionality given the unidirectional nature of the following steps: sticky-ended ligation of the second-round barcodes and the T7 promoter, the downstream linear amplification by IVT, reverse transcription (RT) with fold-back primer and/or annealed RT primer, and primer-directed second-strand synthesis with the third-round barcodes. Altogether, this in theory allows the compatibility of the sci-L3 chemistry with Strand-seq. With each extension of the sci-L3 method, only a single-step modification from the original method is required. For sci-L3-Strand-seq, the introduction of nascent strand nicking before IVT ensures that only the unnicked template strand is used for amplification (Fig. 1). We additionally added an enzymatic nicking step for the nascent strand and thus improved on the original Strand-seq chemistry. We show that sci-L3-Strand-seq generates strand-specific libraries for thousands of single cells at low cost and without requiring specialized equipment [9].

Chromosome conformation capture techniques and its derivatives, such as Hi-C, have revealed the different scales of genome organization ranging from compartments, self-interacting domains, down to individual loops [10–12]. The principle behind conformation capture is proximity ligation, where cut chromatin ends in close physical proximity are ligated together to form hybrid molecules. For the adaptation of the sci-L3 protocol to single-cell Hi-C, we omitted the ligation junction enrichment step conventionally used [11, 13], leading us to obtain a combination of whole genome sequencing (WGS) and interaction (Hi-C) data for each single cell (Fig. 1). While increasing the sequencing burden per cell, this design maximizes the information recovered for each single cell for applications such as structural variation detection, for which both modalities are informative. The first two split and pool barcoding steps in sci-L3 are performed in nucleus, permitting the introduction of restriction enzyme digestion and ligation before the initial Tn5 barcoding and fragmentation [14]. We show that sci-L3-Hi-C captures features of genome organization for thousands of single cells with equivalent performance to other methods.

Single-cell multi-omics are desirable for advancing our understanding of cellular diversity and fundamental biological mechanisms [15], one example being the ability to link *cis*- and *trans*-regulatory elements with gene transcription using chromatin accessibility and transcriptome co-assays [16–18]. We have previously shown that sci-L3 can integrate multiple modalities with the sci-L3-RNA/DNA co-assay [2]. Here, we have extended the sci-L3 co-assay to RNA and ATAC. The chromatin accessibility readout was obtained by omitting the nucleosome depletion step before Tn5 barcoding and fragmentation (Fig. 1). The linear amplification nature of sci-L3, in particular, enabled better recovery of accessible chromatin. Ultimately, we show that sci-L3-RNA/ATAC captures distinguishing features of cell identity from both modalities and is capable of scaling to thousands of single cells.

Materials and methods

Cell culture

BJ-5ta (CRL-4001, ATCC), HEK239T (CRL-3216, ATCC), NIH/3T3 (CRL-1658, ATCC), and Patski (gift from Disteche lab) cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1× Pen–Strep. HAP1 cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM) supplemented 10% FBS and 1× Pen–Strep. GM12878 and CH12 cells were cultured in RPMI 1640 medium supplemented with 15% FBS and 1× Pen–Strep. All cells were cultured at 37°C, 5% CO_2. For sci-L3-Strand-seq, cells were cultured with bromodeoxyuridine (BrdU) at 40 μM final concentration for 24 h prior to fixation. For sci-L3-Strand-seq, we did not mix human and mouse cells prior to fixation as the combinatorial indexing steps are exactly the same as sci-L3-WGS developed previously.

sci-L3 library generation

sci-L3-Strand-seq

Fixation and nucleosome depletion were performed as previously described in the "Methods and molecular design of sci-L3-WGS and sci-L3-target-seq," subsections "Single cell preparation and nucleosome depletion" and "Tagmentation (first-round barcodes) and ligation (second-round barcodes)" [2]. Notably, cells were trypsinized and fixed with 37% formaldehyde (final 1%–1.5% concentration) in 1× phosphate buffered saline (PBS) at a cell density of 1 million/ml for 10 min at room temperature with gentle tube inversion. For first-round barcoding, Patski nuclei were distributed into wells with barcodes 1–12, while HAP1 nuclei were distributed into wells with barcodes 13–24. The remaining nuclei were stained with 4',6-diamidino-2-phenylindole (DAPI) at a final concentration of 5 mg/ml and used for FACS as described in Fig. 2A. After the ligation reaction was stopped by the addition of the stop solution [lysis buffer (LB: 60 mM Tris–Ac, pH 8.3, 2 mM ethylenediaminetetraacetic acid (EDTA), pH 8.0, 15 mM dithiothreitol (DTT)) with 0.1% Triton X-100 (LBT)], nuclei were pooled and stained with Hoechst 33258 to a final concentration of 10 ng/μl, and the quenched population was sorted (100–300 nuclei per well) into a 96-well plate containing 3 μl of LB. The major differences of sci-L3-Strand-seq arise in the "Cell lysis, gap extension, and linear amplification by *in vitro* transcription IVT" subsection of the protocol [2].

Right after gap extension with the Bst WarmStart 2.0 polymerase at 68°C for 5 min and its inactivation with 80°C for 10 min, Hoechst 33258 was added at a final concentration of 10 ng/μl and incubated at room temperature for 10 min in the dark. A 1- or 5-min exposure in a Bio-Rad Gel Doc with a 365 nm UV bulb, or a dose of 27–4000 mJ/cm^2 was administered with a UVP crosslinker (CL-3000L). Note that we did not observe any differences in library quality between 1- and 5-min exposures, or in doses above 270 mJ/cm^2. At this stage, each well had a volume of around 7.7 μl, to which 0.3 μl of USER enzyme (a mix of uracil DNA glycosylase and endonuclease VIII, referred to as UDG + EndoVIII in the text, NEB) was added and incubated at 37°C for 15 min. Next, 0.9 μl of uracil glycosylase inhibitor (UGI; NEB) was added with a further incubation at 37°C for 10 min. Finally, the T7 IVT system was assembled as previously described by adding 2 μl H_2O, 2 μl T7 Pol mix, and 10 μl rNMP mix (NEB, HiScribe T7 Quick High Yield RNA Synthesis Kit) and incubated at 37°C for 10–16 h. The remaining section, "RNA purification, RT, and SSS," was performed exactly as previously described [2].

sci-L3-Hi-C

The major differences of sci-L3-Hi-C arise in the first "Single cell preparation and nucleosome depletion" subsection of the protocol [2]. After 2% formaldehyde fixation and quenching as done for sci-L3-Strand-seq, the proximity ligation was performed largely as previously described for Dip-C [19]. The pelleted cells after fixation were washed with ice-cold 1× PBS, resuspended in 500 μl Hi-C lysis buffer [10 mM Tris, pH 8.0, 10 mM NaCl, 0.2% IGEPAL with 100 μl protease inhibitors (PI; Sigma P8340)], and incubated on ice for at least 15 min. The nuclei were pelleted at 2500 × *g* for 5 min at 4°C and washed with ice-cold Hi-C lysis buffer. The pellet was resuspended in 50 μl of 0.5% sodium dodecyl sulfate (SDS; diluted with H_2O) and incubated for 10 min at 62°C. The SDS was quenched with the addition of Triton X-100 (145 μl H_2O, 25 μl of 10% Triton X-100) and with a further incubation of 15 min at 37°C. Next, 25 μl of NEBuffer2 was added, followed by 20 μl of 25 U/μl MboI (NEB, R0147M) and left overnight at 37°C. The nuclei were pelleted at 2500 × *g* for 5 min at 4°C and washed with 1 ml of ligation buffer (1× T4 DNA ligase buffer, NEB B0202S, with 0.1 mg/ml BSA, NEB B9000S). Ligation was performed with 1 ml ligation buffer and 10 μl of 1 U/μl T4 DNA ligase (Life Tech 15224-025) at 16°C for 4 h. Nuclei were passed through a 35 μm cell strainer, pelleted at 2500 × *g* for 5 min at 4°C, washed with 1 ml lysis buffer (LB: 60 mM Tris–Ac, pH 8.3, 2 mM EDTA, pH 8.0, 15 mM DTT), and resuspended with LB at a concentration of 20 000 nuclei per μl. The remaining steps were performed exactly as previously described in subsection "Tagmentation (first-round barcodes) and ligation (second-round barcodes)" onward [2].

sci-L3-RNA/ATAC

The major differences of sci-L3-RNA/ATAC arise in the nucleosome depletion step previously described in the "Methods and molecular design of sci-L3-RNA/DNA co-assay" subsection of the protocol [2]. Cells were trypsinized, combined together (HEK293T, BJ-5ta, NIH/3T3), and fixed with 2% PFA in 1× PBS at room temperature for 10 min at a density of 1 million/ml. The subsequent quenching (with glycine), washing, and nuclei isolation (with 0.1% IGEPAL) steps are identical with sci-L3-WGS, except the addition of 1% Superase-In to all the LB and 1× NEBuffer 2.1 buffers. After the isolation of nuclei, the pellet was split in half. One half was subjected to nucleosome depletion with the addition of 776 μl of 1× NEBuffer 2.1, 24 μl of 10% SDS, and an incubation at 42°C for 15 min. The SDS was quenched with 180 μl of 10% Triton X-100 with 10 μl Superase-In, and further incubated at 42°C for 10 min. The depleted nuclei were pelleted, washed with 1 ml LB with 1% Superase-In, and resuspended in LB with 1% Superase-In at a 20 000 nuclei per μl concentration. These nuclei represent the sci-L3-RNA/DNA sample. For the other half, the nuclei were washed with 200 μl of LB with 1% Superase-In and 0.1% Triton X-100 and resuspended in LB with 1% Superase-In at a 20 000 nuclei per μl concentration. These nuclei represent the sci-L3-RNA/ATAC sample. For the first-round barcoding, the nucleosome-depleted nuclei were distributed into wells with barcodes 1–5 and 11–15, while the non-nucleosome-depleted nuclei were distributed into wells with barcodes 6–10 and 15–20. The subsequent steps were performed exactly as previously described in the sci-L3-RNA/DNA co-assay subsections "RT and Tagmentation, Ligation, FACS, and Cell Lysis," "Gap Extension and Linear Amplification by In Vitro Transcription," and "RNA Purification, RT, and SSS" [2].

sci-L3 read processing and alignment

The first step of processing sci-L3 sequencing data is the extraction and consolidation of the combinatorial barcodes that uniquely identify each single cell. The barcode extraction and processing has been previously described in [2] and subsequently implemented as an easy to use snakemake pipeline (sciL3pipe; available at https://github.com/recombinationlab/sciL3Pipe) [3, 20]. Briefly, the processing steps (i) first orient read pairs such that the combinatorial barcodes are always within read 1 (R1). The orientation is identified by the presence of either the RT primer sequence (allowing up to three mismatches using Levenshtein distance) or the third-round barcode; (ii) identify the third-round barcode at the start of R1 (SSS, 6 nt, no mismatches allowed) and write it into the read name together with the *in vitro* transcription unique molecular identifier (4 nt; ivt_UMI). Reads without a matching SSS are discarded, while matching reads are written into individual SSS fastq files that enable the division of subsequent steps among multiple processors; (iii) clean up reads by trimming the Tn5 Mosaic End double-stranded (MEDS) sequence from R1 (5′ adaptor: AGATGTGTATAAGAGACAG; maximum error rate: 0.2, minimum overlap: 19) and R2 (5′ adaptor: AGATGTGTATAAGAGACAG; 3′ adaptor: CTGTCTCTTATACACATCT; maximum error rate: 0.2, minimum overlap: 13) using cutadapt [21]; (iv) identify the first-round (Tn5 or RT, 8 nt, 1 mismatch allowed; Levenshtein distance) and second-round (ligation, 7 nt, 1 mismatch allowed; Levenshtein distance) barcodes from the MEDS adjacent sequence (bc2 7nt | 4nt spacer | bc1 8nt | MEDS) and write both barcodes into the read name. Additionally, for co-assays, the RNA is distinguished from the DNA modality based on the first-round barcodes. For RNA, a reverse transcription UMI (6 nt; rt_UMI) is extracted and added to the read name, while a placeholder "GGGGGG" is added for DNA (used for distinguish RNA from DNA downstream); (v) align read pairs with all three barcodes identified to hs37d5 or a hybrid reference of hs37d5 and mm10 using bwa-mem [22] and converted to sorted BAM using samtools [23].

sci-L3-Strand-seq processing and data analysis

Flow cytometry data were collected on a BD Fortessa at the BSCRC flow cytometry core and analyzed using BD FACSDiva (8.0.1) and FlowJo (10.10).

After alignment, filtering was performed for proper read pairs (FR), a maximum insert size of 2000, and a maximum soft-clipping ratio of 0.5 (defined as the number of soft-clipped bases over the number of matched bases) to remove excessively soft-clipped reads, which if retained contribute significantly to the background levels. Following filtering, reads were split by species of origin (hs37d5 or mm10) and cell barcodes. For each cell, strand breakpoints and background estimates were obtained using breakpointR (1.18.0) [24]. For Patski, a diploid cell line, we expect breakpoints in the form of WC to WW or CC transitions, while for HAP1, a haploid cell line, we expect WW to CC (or CC to WW) transitions. Irrespective of ploidy, breakpointR assigns strand state to breakpoint-segmented regions. Previously described blacklisted regions for hg19 and mm10 (version 2) [25], in addition to a high signal region we consistently found enriched on chr4-89540413:89 544 084 in mm10, were excluded from all analysis. BreakpointR breakpoints were filtered to eliminate false calls using a set of distance and strand state-based criteria (details in companion sci-L3-Strand-seq paper).

breakpointR run command used:

```
breakpointr(inputfolder = <>, outputfolder = <>, windowsize = 5000000, binMethod = "size", pairedEndReads = TRUE, pair2frgm = FALSE, min.mapq = 20, filtAlt = TRUE, peakTh = 0.33, trim = 10, background = 0.05, minReads = 50, maskRegions=<>)
```

After breakpoint filtering, the strand state of each breakpoint and segment was re-evaluated with the `background` and `minReads` set to 0.1 and 10, respectively. These final segment strand states were used to evaluate the percentage of WC and "uncategorized (?)" regions as designated by breakpointR within each cell. We refer to these regions as "strand-neutral" regions. Cells with a WC and "uncategorized (?)" percentage of more than (=>) 15% for haploid and 75% for diploid, together with a background estimate of 0 or more than (>) 0.08, were filtered out.

The relative chromosome coverage was calculated as the coverage per chromosome divided by the total coverage in the cell, normalized by the chromosome mappable size divided by the total size of the mappable genome. The mappable size was defined as the proportion of the chromosome or genome that does not contain ambiguous bases (*N*) over the total number of bases.

The complexity analysis that extrapolates the percent of genomic coverage obtained with additional sequencing effort was performed using preseqR [26].

We performed phasing as previously described using StrandPhaseR (v0.99) [5, 27]. Using the known variants from B6/SPRET [28], we calculated phasing coverage for all heterozygous SNV sites and phasing accuracy by comparing to the ground truth in the reference VCF.

sci-L3-HiC processing and data analysis

Aligned ensemble sci-L3-HiC data were filtered first by removing read pairs with cell barcodes with <10 000 counts (not enough reads to be considered as a single cell) and second by MAPQ 30 scores. For the high sequencing depth library, the barcode count cutoff was increased to 30 000. Filtered BAMs were processed using the 4DN Docker image (v43; https://github.com/4dn-dcic/docker-4dn-hic) that contains wrapper scripts for running pairtools [29], pairix [30], and cooler [31]. First, the filtered BAMs were converted to the pairsam format and sorted (run-pairsam-parse-sort.sh). Second, PCR duplicates were marked (DD) within the pairsam files (run-pairsam-markasdup.sh) and subsequently filtered out by only retaining valid pairs (UU, unique–unique; UR, unique–rescued; RU, rescued–unique) (run-pairsam-filter.sh). We also added the restriction fragment (RF) (MboI) positions to the filtered pair files (run-addfrag2pairs.sh), which were obtained using Juicer [32] (https://github.com/aidenlab/juicer/blob/main/misc/generate_site_positions.py). Interaction pairs with a distance of 1 kb or less were filtered out to remove Hi-C by-products such as dangling ends and self-circles [29]. Lastly, the filtered valid pairs were converted to 1 kb resolution cools (run-cooler.sh) and subsequently to multi-resolution cools (run-cool2multirescool.sh) for the generation of ICE normalized interaction matrix plots [31, 33]. For plotting of single-cell raw interaction counts, interaction pairs were filtered by barcodes within the read name in R, exported in BEDPE format, and converted into cool files using cooler [31, 34]. A/B compartment calls from the ensemble data were made using cooltools *cis* eigenvector decomposition at a 500 kb resolution [35]. ChIP-seq of H3K4me3 performed in GM12878 cells (ENCODE: ENCFF818GNV) [36–38] was used as the phasing track to select and orient the eigenvectors most correlated with active regions. Smoothed $P(s)$ curves for normalized intra-arm chromosomal contacts were calculated from contact matices at 10 kb resolution and aggregated using cooltools [35].

For A/B compartment comparisons, we downloaded and processed bulk Hi-C data from GM12878 cells using the same pipeline as described above, with the exception of omitting the interaction pair distance filter [39].

To compare sci-L3-Hi-C with other single-cell Hi-C methods, we downloaded processed and/or supplementary data for sci-Hi-C [13], s3-GCC [40], and droplet Hi-C [41]. For the combinatorial indexing method, sci-Hi-C, the ML3 valid read pairs provided by the authors were used [13]. Since the ML3 library contained K562, GM12878, Patski, and primary mouse embryonic fibroblast (MEF) cells, only valid pairs corresponding to the GM12878 cell line were used (GSM2254217). For the GM12878 cells, a bimodal distribution of valid pairs per cell barcode was observed. Consequently, any barcodes with <100 valid pairs were excluded. For generating interaction matrices, the same interaction distance filter of 1 kb was also applied as used for sci-L3-Hi-C data. The final set of valid pairs was exported in BEDPE format and converted into cool files using cooler [31, 34]. Additional data on sci-Hi-C performance were obtained from table 1 in [42]. For s3-GCC, the pairix file provided by the authors (GSE174226) was parsed and contacts at various distances (*cis* <1 kb, *cis* >1 kb, *trans*) were counted [40]. The counts for each cell barcode were annotated using Supplementary Table S8 and only cells matching PDAC1 and PDAC2 samples were used. For droplet Hi-C, data provided in Supplementary Table S2 were used [41]. For normalization of sequencing effort between methods with or without ligation junction enrichment, we used raw read number of the ML3 library in sci-Hi-C [13] (21 808 993 read pairs) and the proportion of GM12878 reads (37.56%

based on GSM2254217_ML3.percentages.txt.gz), and estimated 8 191 458 raw reads for GM12878 cells. Alternatively, we also extrapolated the number of Hi-C-specific reads in sci-L3-Hi-C by calculating the proportions of *cis* >1000 and *trans* reads: 5.8% and 6.3%, respectively, for high-depth and low/medium-depth cells. After excluding the 5% mouse reads, we estimated that we sequenced 86.4 × 95% × 5.8% million read pairs for 96 high-depth cells and 55.7 × 95% × 6.3% million read pairs for 929 low/medium-depth cells.

sci-L3-RNA/ATAC and sci-L3-RNA/DNA processing and data analysis

DNA/ATAC and RNA count matrices were generated from sorted BAMs using a MAPQ 10 cutoff. RNA reads were distinguished from DNA/ATAC based on the rt_UMI in the read name ("GGGGGG" for DNA/ATAC and "NNNNNN" for RNA). For DNA/ATAC, the count matrices consist of the R1 raw alignment counts (r1s) per cell barcode and split by species (mouse, m_r1s; human, h_r1s). For RNA, instead of raw alignment, the counts are based on a unique molecular identifier (UMI) that consists of the species, overlapping gene name, and the rt_UMI. An initial filtering of barcodes with <300 combined counts of DNA r1s and RNA UMIs (non-cells) was performed for both mouse and human. To identify and remove barcode collisions, >90% of DNA/ATAC counts and >80% of RNA UMIs were required to originate from either human or mouse, with an agreement between both modalities as a final filter. Non-cell barcodes (e.g. debris) are expected to have a higher rate of collisions compared to cells and overall contain a lot fewer reads. We used these expectations to further filter barcodes for high- and low-depth libraries independently. For high-depth libraries, a DNA/ATAC r1s count threshold of $>10^{3.8}$ (~6309.57) was chosen, along with a threshold of >1000 UMIs for RNA (Supplementary Fig. S3A). For low-depth libraries, a threshold of 1000 DNA/ATAC r1s and 20 RNA UMIs was chosen as a compromise between number of cells recovered and the informational content per cell (Supplementary Fig. S3A). Barnyard plots were plotted from the resulting filtered barcodes.

To establish a ground truth for cell identity, BJ SNVs called from bulk WGS data (SRA: SRP102259; [43]) and previously published HEK293T SNVs (http://bioinformatics.psb.ugent.be/downloads/genomeview/hek293/SNP/293T_RTG.vcf.gz) [44] were used to obtain a list of SNVs private to each cell line using BCFtools [23]. The HEK293T VCF was lifted over from hg18 to hg19 and filtered for SNVs using BCFtools (1.11). BJ SNV calling was performed with BCFtools using the following run command:

```
bcftools mpileup -O z –skip-indels –ignore-RG –redo-BAQ –min-BQ 13 –per-sample-mF -a 'AD,ADF,ADR,DP,SP,SCR' -f < genome.fa> <combined.bam> | bcftools call –multiallelic-caller –variants-only -O z -o <out.vcf.gz>
```

The number of reads containing a private SNV were normalized by the total number of reads overlapping with private SNV positions and scaled to be within a 0–1 range. This cell line relative total of SNVs was further normalized by the sum of both cell line relative totals [e.g. BJ/(BJ + HEK)] and is referred to as the rate of cell line-specific SNVs. A rate of >70% was required for each cell to be assigned a cell line identity, otherwise it was classified as a SNV collision and excluded from downstream analysis.

The RNA and ATAC modalities were analyzed separately using Seurat (4.4.0) and Signac (1.13.0), respectively [45, 46]. For ATAC-specific analysis, DNA reads were isolated based on the rt_UMI filtered for proper read pairs (FR), a maximum insert size of 2000, and a maximum soft-clipping ratio of 0.5. A CB tag was added to the BAM files based on the cell barcodes present in the read name. Using the CB tag, BAMs were converted into fragment files using sinto (0.10.0), and imported into Signac (1.13.0) [45]. A feature matrix of peaks was created in Signac using the built-in wrapper function for MACS2 [47] and used for downstream clustering and visualization. A 5-kb bin-based feature matrix was additionally created to calculate the fraction of reads overlapping mitochondria. Transcriptional start site (TSS) enrichment was performed based on the Ensembl gene annotations from the EnsDb.Mmusculus.v79 and EnsDb.Hsapiens.v75 Bioconductor packages. Frequency inverse document frequency (TF-IDF) normalization with the latent semantic indexing (LSI) dimensional reduction was performed [45]. The first 10 dimensions from the LSI reduction were used as input for the low-dimensionality visualization with uniform manifold approximation and projection (UMAP) [48], with the exclusion of one of the dimensions that strongly correlated with sequencing depth.

For RNA-specific analysis, the matrix of UMI counts over genes (gencode annotation v19 [49]) was used to create a Seurat object, retaining features that are shared between at least two cells and cells with at least one feature. The percentage of mitochondrial reads was added as a metadata feature. Default parameters were used for normalization, selection of the top 3000 most variable features, and scaling and centering of the data, with an exception of selecting the mitochondrial read percentage as the variable to be regressed out. The first 10 components from the principal component analysis (PCA) were used as input for the low-dimensionality visualization with UMAP, while excluding the dimension representing sequencing depth as done in the ATAC analysis.

Cluster identity was assigned based on the dominant ground truth (SNV identity) cell type within each cluster. Cell identity based on clusters derived from each modality was overlaid onto the clusters of the opposing modality (modality identity) to determine the agreement between RNA and ATAC. Examining the distribution of low and high sequencing depth cells showed similar distributions of both within the clusters, validating depth was not driving clustering.

Results

sci-L3-Strand-seq

Sequencing the template strands of DNA replication requires the newly synthesized strands to be labeled with BrdU within a single cell division [4]. The level of labeling and the phase of the cell cycle contribute to the quality of the strand information. Under-labeling results in interspersed regions of both Watson and Crick reads per chromosome as no nascent strands are ablated, while double labeling in the next S-phase results in regions with no reads as both strands contain BrdU and are thus ablated. Input cells for sci-L3-Strand-seq do not need to be cell cycle synchronized, enabling potential *in vivo* applications. The sci-L3 framework contains a flow sorting

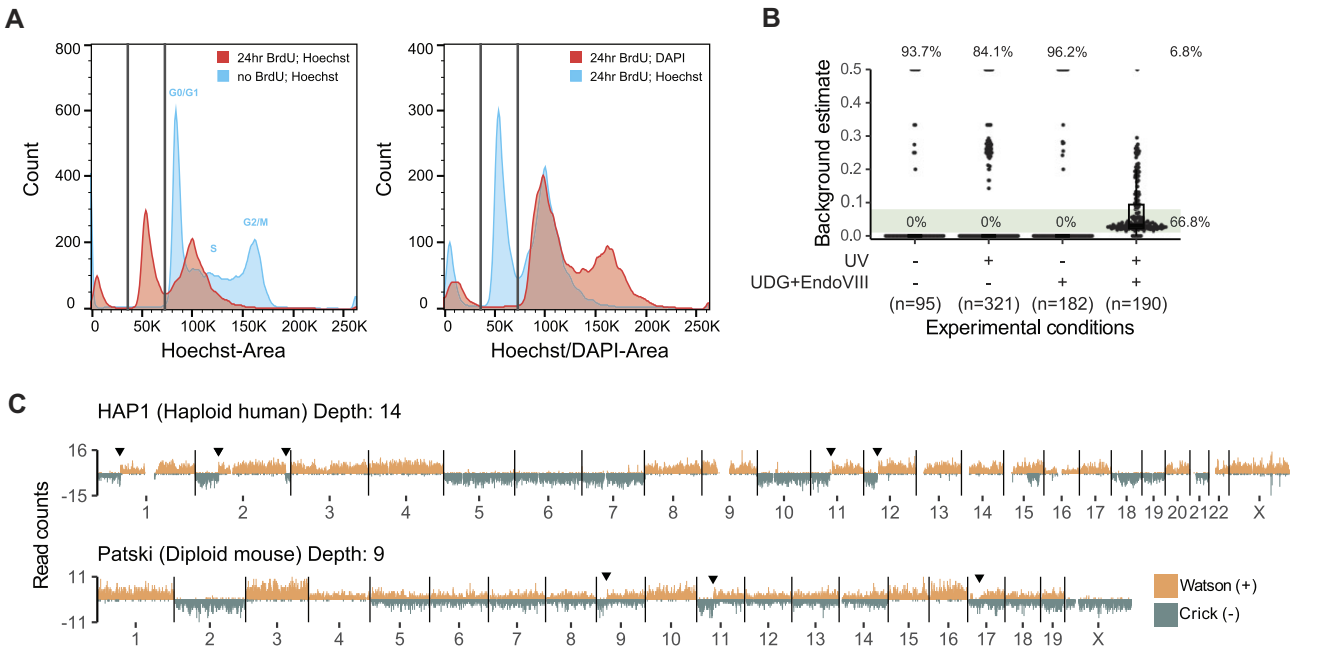


Figure 2. DNA template Strand-seq with sci-L3-Strand-seq. (**A**) Example sorting gate for BrdU-labeled cells. The sorting gates are demarcated by two vertical lines. The histogram on the left shows quenching of the Hoechst signal in cells labeled with BrdU for 24 h (red) compared with cells without BrdU (blue), both stained with Hoechst. The histogram on the right shows 24-h BrdU-labeled cells stained with DAPI (red; no quenching control) and Hoechst (blue). The desired cells quenched by Hoechst stand out at roughly half of the signal of cells stained with DAPI. Peaks to the left of the sorting gates are predominantly debris. (**B**) Cell background estimates for combinations of UV and UDG + EndoVIII conditions. The highlighted background estimate filter threshold of 0> and < 0.08 is used for identifying cells with template strand information. Number of cells within each experimental condition (*n*) is shown below the plot. Percentages within the plot are cells passing filter (within green strip) and cells with 0 or 0.5 background (top, more detailed breakdown in Supplementary Table S1), which indicates that they are completely strand-unbiased WGS libraries. Note that UV treatment alone only slightly increased the strand bias of WGS libraries, while UDG + EndoVIII treatment alone did not. Both are required to generate high-quality sci-L3-Strand-seq libraries (details in text). (**C**) Example sci-L3-Strand-seq profiles of a haploid and a diploid cell. The distribution of reads in a haploid cell (top) shows either Watson (+, forward strand, W) or Crick (−, reverse strand, C) template strand orientation, except for the known disomic region on chromosome 15 that harbors both W and C reads. For diploid cells (bottom), the inheritance of two template strands from either parent creates a mixture of WW, WC, or CC reads. Triangles highlight SCEs. Depth represents the median number of reads per Mb, excluding reads falling into blacklisted regions.

step after two rounds of cell barcoding that can be leveraged to enrich for the useful cell population, i.e. cells in the immediate subsequent G1 cycle after BrdU labeling. We thus only require a proportion of the input cells to survive until the next G1, which is compatible with experimental designs where cells are perturbed in a pooled manner, naturally growing, and dividing *in vivo*, and/or simply for the ease of the experiment. It is worth noting that the sci-L3 methods can be performed using dilution instead of FACS, and we routinely do so for quality control; however, for sci-L3-Strand-seq, the need to enrich for BrdU-labeled cells means fewer cells with template strand information would be recovered using dilution.

We fix BrdU-labeled cells along a spectrum of cell cycles (Fig. 2A, left, Hoechst or alternatively DAPI to stain for ploidy and cell cycle stage without BrdU) and subject them to the sci-L3 workflow that includes nucleosome depletion, tagmentation, and ligation. After these first and second rounds of *in situ* barcoding, nuclei are sorted by flow into separate wells for amplification and third-round barcoding [2]. In this flow sorting step, the incorporated BrdU quenches the Hoechst 33258 stain [50], shifting the cell cycle profile to the left (Fig. 2A, left, Hoechst with BrdU). In the sci-L3 workflow, we typically use DAPI to stain the nucleosome-depleted, non-tagmented control nuclei to separate debris and nuclei. Such DAPI-stained control nuclei can serve as a negative control for BrdU quenching as the incorporated BrdU does not quench the DAPI signal. In the original development of Strand-seq, a no BrdU control (unlabeled cells) is recommended to detect Hoechst quenching by BrdU. Since the DAPI and Hoechst stain for cell cycle and ploidy largely overlap without BrdU (Supplementary Fig. S1A), the use of the DAPI control sample avoids the need to additionally culture and process unlabeled cells just for setting the FACS sorting gates. We sorted hundreds of cells per well from the BrdU-quenched, G1 population to test various nascent strand ablation conditions for sci-L3-Strand-seq (Fig. 2B and Supplementary Table S1). In real applications, one can increase the number of first and second round of barcodes as described in sci-L3 [2] such that thousands of cells can be sorted per well to further increase throughput. Overall, we recovered an average of 93% of the total sorted cells (789 cells recovered out of 850 sorted).

The Strand-seq protocol utilizes a 270 mJ/cm^2 total dose of 365 nm UV in the presence of Hoechst 33258 to induce single-stranded nicks at the site of BrdU incorporation [1]. We applied the same procedure to sci-L3 barcoded and BrdU incorporated nuclei with UV doses ranging from 27 to 4000 mJ/cm^2 and assessed the quality of the resulting strand information using a background estimation calculation [24]. Essentially, regions where both template strands were inherited in the same orientation will contain a few reads with the opposite orientation that represent the background noise. Using this metric, we found that despite adequate levels of BrdU incor-

poration (Fig. 2A) and irrespective of UV dose, we consistently obtained cells without any strand information and essentially only produced single-cell WGS libraries (Fig. 2B, background of 0 and >0.08). We postulated that 365 nm UV should catalyze debromination effectively [51]; however, nicking of the DNA strand next to the U base may not necessarily occur, leading us to test the addition of the uracil DNA glycosylase (UDG) and the DNA glycosylase-lyase endonuclease VIII (EndoVIII), subsequently referred to as "UDG + EndoVIII." The combination of UV treatment and UDG + EndoVIII generated strand information in over 66% of cells (Fig. 2B and C).

We assayed both haploid and diploid cells, validating that we obtained single-strand information and were also able to detect SCEs [Fig. 2C and Supplementary Fig. S1B, 4 (±2.7) SCEs in 71 HAP1 cells and 7 (±2.8) SCEs in 38 Patski cells]. We further selected cells based on the overall proportion of regional strand states, allowing the combination of "strand-neutral" [the chromosome can be Watson–Crick (WC) and/or not biased enough to be assigned to Watson–Watson (WW) or Crick–Crick (CC)] regions to account for at most 15% in haploid and 75% in diploid cells (Supplementary Fig. S1C; see the "Materials and methods" section). In the end, we obtained over 58% of sorted cells with strand information passing QC (Supplementary Table S1) with an average 0.22% genome coverage (Supplementary Fig. S1D). With limited variation between cells in their relative chromosome coverage, we can identify obvious copy number variation (CNV) in Patski cells for chromosomes 3, 4, 5, 9, 12, and 19 (Supplementary Fig. S1E). Strand-seq is a powerful tool for phasing, i.e. assigning parent of origin to heterozygous variants [5]. In WC regions, SNVs from one parent should map solely in the W (or C) direction, and SNVs from the other parent should map in the opposite direction. In theory, even a single diploid cell sequenced by Strand-seq could provide phasing information, regardless of how small the coverage is. We thus attempted phasing with the limited set of 40 Patski cells, where we have the ground truth for haplotype information. We achieved an average phasing accuracy of 79.5% for 2.2% of heterozygous variants (2.4% collective coverage in WC regions among these cells), noting that subclonal aneuploidy greatly affects phasing accuracy. For comparison, we picked 40 cells with top coverage among a bigger set of 931 Patski cells from a larger experiment. We were able to phase 21.5% variants (23.6% coverage in WC regions among these cells) with 85% accuracy.

sci-L3-Hi-C

We next extended sci-L3 to examine the 3D genome organization in single cells. With the initial steps of sci-L3 being shared with the nucleosome depletion used in chromosome conformation capture, it became possible to incorporate MboI digestion (4-base cutter) and ligation without altering the remainder of the assay (Fig. 1) [52]. We tested our scheme with over 1000 cells, consisting of a mix of GM12878 and Ch12 cells before fixation. This small spike-in of 5% mouse cells (Ch12) allowed us to assess barcode collision, showing that we recover the expected number of cells with both human and mouse reads. Overall, we recovered 48% of the expected sorted cells with a sequencing depth of 19 000 read pairs per cell (under-sequenced), and 91% of cells with a higher sequencing depth of 68 000 read pairs per cell (Supplementary Table S2). We then deep sequenced a subset of cells to saturation at a depth of over 800 000 read pairs per cell. In total, 96 cells passing QC were recovered out of the 100 sorted (Supplementary Table S2). After examining the number of unique valid interaction pairs for each library (see the "Materials and methods" section), we retained 90.8% (35M/38M) of the initial low-pass aligned reads and 68.8% (43M/61M) of the high-depth aligned reads (7.6% and 29.8% are duplicates, respectively, with <1.5% invalid pairs, Supplementary Table S3). Note that sci-L3-Hi-C does not have the biotin incorporation step to enrich ligation junctions, and thus works best for profiling both the whole genome sequence and its 3D structure. Therefore, the vast majority of the valid interaction pairs were at a distance of <1 kb and showed a forward–reverse read orientation bias (dangling ends) consistent with whole genome sequencing (Supplementary Fig. S2A and B) [29, 53]. After filtering pairs with interaction distance of <1 kb, we obtained a comparable distribution of contact frequencies across genomic distances to bulk Hi-C and to a previously published combinatorial indexing single-cell Hi-C method, sci-Hi-C (Supplementary Fig. S2B) [39, 13]. However, the three-level indexing increases throughput and the linear amplification counteracted the loss of chromatin interactions typical of adding more rounds of split and pool, which is an expected advantage of the sci-L3 scheme.

We subsequently examined the ensemble interactions of all cells to validate that we can detect features typical of Hi-C (Fig. 3B and C). With clear enrichment of intrachromosomal contacts across all chromosomes (Supplementary Fig. S2D), we focused on individual chromosomes displaying the canonical plaid pattern of A/B compartments [11]. Using eigenvalue decomposition [35], we observed a high correlation ($R = 0.87$, P-value $<2.2 \times 10^{-16}$, Spearman correlation test) between the compartment eigenvalues in bulk Hi-C and our single-cells ensemble (Fig. 3B and Supplementary Fig. S2E). Despite substantial under-sequencing, with only 48% of the expected number of sorted cells being recovered, the interaction map of the 929 lowly sequenced cells was highly similar to that of the 96 highly sequenced cells (Supplementary Fig. S2F). This suggests that large numbers of cells at lower sequencing can provide equivalent ensemble insights into genome organization, opening the door to creating pseudo-bulk interaction maps from *in silico* grouped cells that could further our understanding of complex tissues and organisms.

Finally, we wanted to examine how our method compared to other single-cell Hi-C methods, namely the two-level combinatorial indexing methods, sci-Hi-C [13] and s3-GCC [40], as well as a 10× Genomics method, droplet Hi-C [41]. With sci-Hi-C being the only method that performed ligation junction enrichment, s3-GCC and droplet Hi-C cells contain the same high proportion of interaction pairs at a distance of <1 kb as our sci-L3-Hi-C (Supplementary Fig. S2A). sci-L3-Hi-C (high depth) obtained more total read pairs per cell than droplet Hi-C, but had fewer *cis* pairs at a distance of >1 kb, which represent informative chromatin interaction pairs (Fig. 3A, Supplementary Fig. S2A, and Supplementary Table S3). One reason for this lower proportion of ligation junction reads could be that sci-L3-Hi-C uses the MboI restriction enzyme with around 7M cut sites, which is 44% fewer than s3-GCC with 13M cut sites by AluI, and 76% fewer than droplet Hi-C with 30M cut sites by a combination of DpnII, MboII, and NlaIII.

Both sci-Hi-C and sci-L3-Hi-C use MboI, and thus have the same theoretical complexity. We compared the two methods

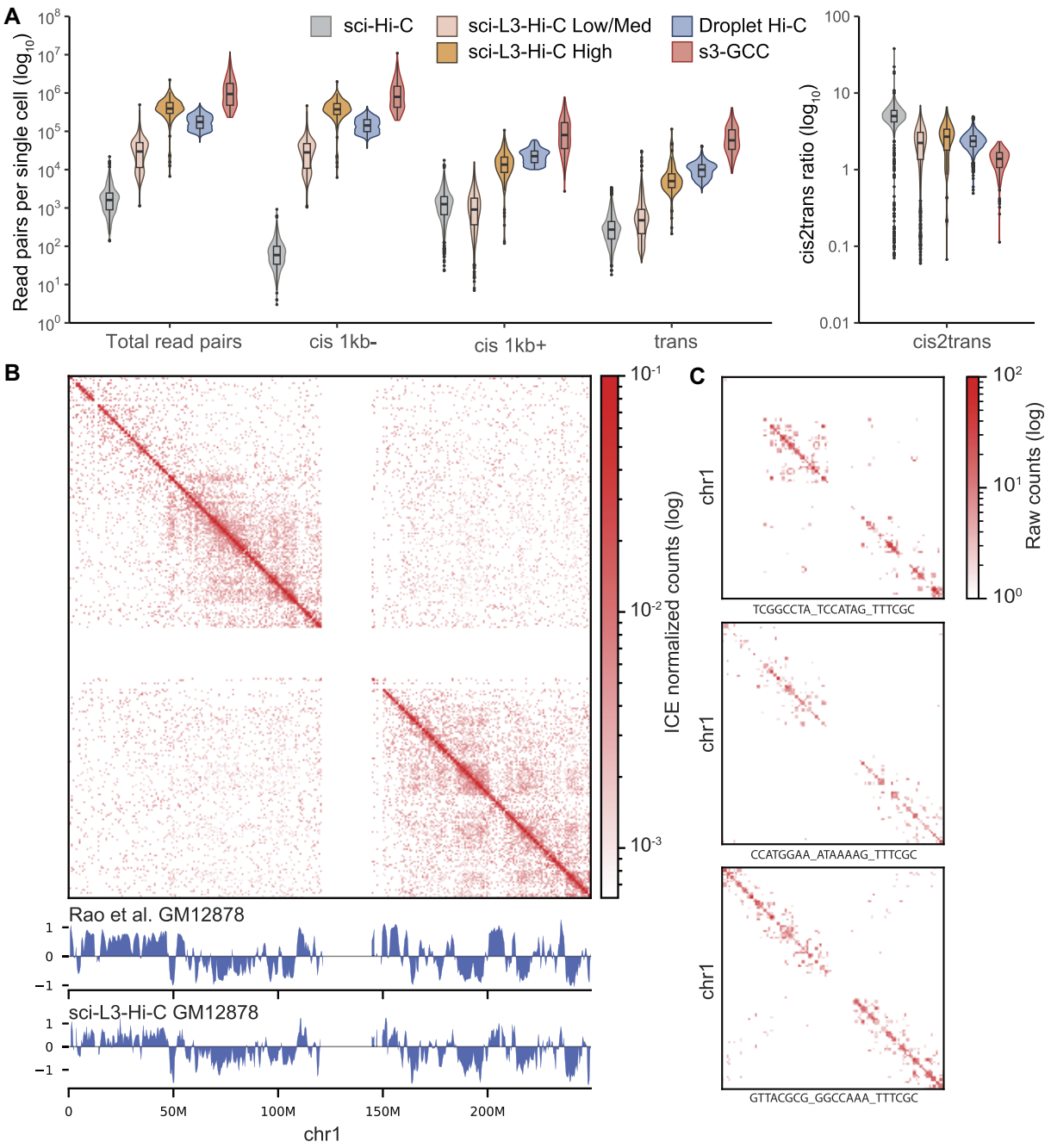


Figure 3. Single-cell genome conformation with sci-L3-Hi-C. (**A**) Comparison of total read pairs and pairs at different contact distances per single cell between four single-cell Hi-C methods: $n = 910$ (sci-Hi-C), 929 (sci-L3-Hi-C low and medium coverage), 96 (sci-L3-Hi-C high coverage), 6235 (droplet Hi-C), and 202 (s3-GCC). *Cis* to *trans* ratio uses *cis* pairs with contact distance of >1 kb. (**B**) Ensemble ICE normalized contact map of chromosome 1. The contact map was plotted at 500 kb resolution from both low- and high-coverage cells, excluding any pairs with interaction distance of <1 kb. Below the main heatmaps are tracks of the first eigenvector showing A (+) and B (−) compartments from bulk Hi-C data and from the ensemble sci-L3-Hi-C-seq data. (**C**) Example chromosome 1 contact maps from three single cells. Raw count contact maps were plotted at 2.5 Mb resolution, excluding any pairs with interaction distance of <1 kb. Single-cell barcodes are shown below each plot.

in more detail. Applying the same 1-kb interaction distance filter as in sci-L3-Hi-C to the published sci-Hi-C GM12878 data and removing any cells with <100 interaction pairs, we obtained 910 cells with a median of 1547 interaction pairs (Supplementary Fig. S2G and Supplementary Table S4). In contrast, our highly sequenced cells contained a median of 19 385 interaction pairs. However, the RNA-dependent RNA polymerase activity of the T7 RNAP used during IVT permits transcription from self-primed or cross-primed RNA that can result in truncated ends [2, 43]. The variation in ends prevents the complete removal of IVT duplicates with conventional deduplication methods. To avoid inflating the number of pairs per cell due to IVT duplicates in our data, only interactions between unique combinations of MboI fragments were counted. As a result, the median unique fragment pairs per cell for sci-Hi-C became 1390 (from 1547), while the high depth sci-L3-Hi-C cells now contained a median of 7549 (from 19 385) (Supplementary Fig. S2H). The unique MboI fragment counts

of our high-depth cells are on par with the median read counts obtained with other higher depth sci-Hi-C libraries, despite sci-L3-Hi-C omitting ligation junction enrichment [13].

We further attempted to normalize the two methods with and without enrichment of ligation junctions to estimate sequencing effort required to obtain similar numbers of unique chromatin interactions. First, we estimated the raw sequencing depth for the GM12878 data in sci-Hi-C. Without enrichment of ligation junctions, we require 6.4× the amount of sequencing effort to obtain the same 1500 informative chromatin interactions per cell. However, at higher sequencing depth, sci-L3-Hi-C is not necessarily less cost-effective. Kim *et al.* [42] performed sci-Hi-C at saturation (only ∼10% unique valid pairs) but valid pairs plateaued at 5000 per cell, which is substantially less than the 7500 unique RFs recovered with sci-L3-Hi-C (Supplementary Table S4). Meanwhile, the non-interacting reads in sci-L3-Hi-C are useful WGS data for profiling other aspects of the genome. Alternatively, we estimated Hi-C-specific read counts (>1000 distance and *trans*) representing chromatin interactions in sci-L3-Hi-C to compare with sci-Hi-C. sci-L3-Hi-C requires 49 000 read pairs per cell (high depth) for 7500 unique RFs and 3500 read pairs per cell (low or medium depth) for 1300 unique RFs, while sci-Hi-C requires 9000 read pairs per cell for 1400 unique RFs. This validated that sci-L3-Hi-C provides more complex libraries of chromatin interactions.

We additionally noticed that sci-L3-Hi-C cells contained a lower proportion of *trans* reads compared to s3-GCC and droplet Hi-C, while sci-Hi-C has the lowest proportion as reflected by its high *cis* to *trans* ratio (Fig. 3A and Supplementary Table S4). The higher background noise compared to sci-Hi-C may limit applications requiring fine resolution. Overall, we show that sci-L3 can be extended to examine genome organization in thousands of single cells.

sci-L3-RNA/ATAC

With the potential for significant biological insights offered by multi-omics assays [15], we lastly set out to extend sci-L3 for the joint profiling of chromatin accessibility (ATAC) and RNA. sci-L3-ATAC/RNA builds upon our previously described single-cell WGS and RNA co-assay, which relied on nucleosome depletion for the uniform genome-wide insertion of barcodes by the Tn5 transposome complex [2]. By omitting nucleosome depletion, we can instead obtain profiles of chromatin accessibility together with RNA in a single protocol from fixed nuclei (Fig. 1).

As proof of principle, we applied sci-L3-RNA/ATAC with half of the input cells for a mixture of human BJ-5ta, HEK293T, and mouse NIH/3T3 cells. Alongside, we also performed sci-L3-RNA/DNA on the other half, combining cells from the two methods after the first-round barcode step. Altogether, we sequenced 2700 (1512 for RNA/ATAC and 1188 for RNA/DNA) cells with 1200 (672 for RNA/ATAC and 528 for RNA/DNA) cells sequenced at around 11× higher depth than the remaining 1500 (840 for RNA/ATAC and 660 for RNA/DNA) cells. After applying initial filters with depth-specific cutoffs to identify barcodes belonging to cells (Supplementary Fig. S3A), we obtained 1265 cells containing both chromatin accessibility and transcriptome data. Excluding a 5%–8% of doublet cells that contain both human and mouse reads (Supplementary Fig. S3B and C), we recovered a total of 1164/1265 (92%) RNA/ATAC cells without collisions. We next confirmed that both the ATAC and RNA modalities were assigned to the same species for all the 1164 single cells (Supplementary Table S5). Overall, at the lower sequencing depth we recovered 66.8% of the sci-L3-RNA/ATAC sorted cells (561 passing QC out of 840 sorted, 149 human and 412 mouse cells) at a median of 2396 ATAC reads and 97 RNA UMIs per cell (low coverage) (Supplementary Table S5 and Supplementary Fig. S4A–C). At the higher sequencing depth (around 329 000 reads per cell), we recovered 89.7% (603 passing QC out of 672 sorted, 241 human and 362 to mouse cells) at a median of 25 456 ATAC reads (around 6700 unique fragments, 5× compared to sci-CAR [17]) and 1292 RNA UMIs per cell (high coverage). The method significantly improves recovery of ATAC reads per single cell somewhat at the cost of RNA recovery. Similar cell recovery numbers were also obtained for sci-L3-RNA/DNA as expected from previous development of the sci-L3 RNA/DNA co-assay (Supplementary Table S5). To evaluate the concordance of sci-L3-RNA/ATAC with open chromatin, we first examined TSS enrichment and the fraction of reads within peaks (FRiP) (Fig. 4A). Both measures showed significantly higher values for ATAC compared to our DNA libraries (TSS enrichment *P*-value 2.12×10^{-274}, FRiP *P*-value 3.52×10^{-271}; Mann–Whitney *U*), a difference consistent with open chromatin insertion. Similarly to previous reports of ATAC-seq from fixed nuclei, we did not observe a high sub-nucleosomal score in our ATAC (Supplementary Fig. S4E) [54]. We suspect that the removal of the SDS treatment step allows residual mitochondria to remain attached to the nucleus, meaning aside from open chromatin, sci-L3-RNA/ATAC also captures mitochondrial reads (Supplementary Fig. S4D). Although generally undesired in ATAC-seq, mitochondrial reads have the potential to be used in somatic mutation detection and clonal lineage reconstruction [55]. We subsequently only focused on the performance of the sci-L3-RNA/ATAC subset of data to examine whether the enrichment of TSS and peaks in general is informative of cell types in agreement with the transcriptome.

Focusing on the BJ-5ta and HEK293T human cell mixing experiment, we wanted to determine whether the ATAC or RNA profiles contained sufficient information to distinguish the two cell types. We first examined the SNV private to each cell line to obtain a ground truth cell identity, obtaining 383/390 cells (233 BJ-5ta, 150 HEK293T) with unambiguous assignment (Supplementary Fig. S5A). We first examined both high- and low-coverage cells together. After performing normalization and dimensional reduction with either LSI or PCA for ATAC or RNA, respectively, we found that the first dimension for both modalities correlated with sequencing depth and was therefore excluded from downstream analysis [45, 56]. Applying UMAP separated cells into two clusters by both ATAC or RNA alone (Supplementary Fig. S5B). Based on these modality clusters, we observed that 92.6% of cells (355/383) were correctly clustered by RNA and 98.6% by ATAC (378/383) (Supplementary Fig. S5B and Supplementary Table S6). We validated that sequencing depth was not driving clustering by observing that both high- and low-depth cells were interspersed within both clusters (Supplementary Fig. S5B). Examining the 7.4% (28/383) by RNA and 1.4% (5/383) by ATAC of cells with discordant assignment revealed that 87.9% (29/33—intersect of

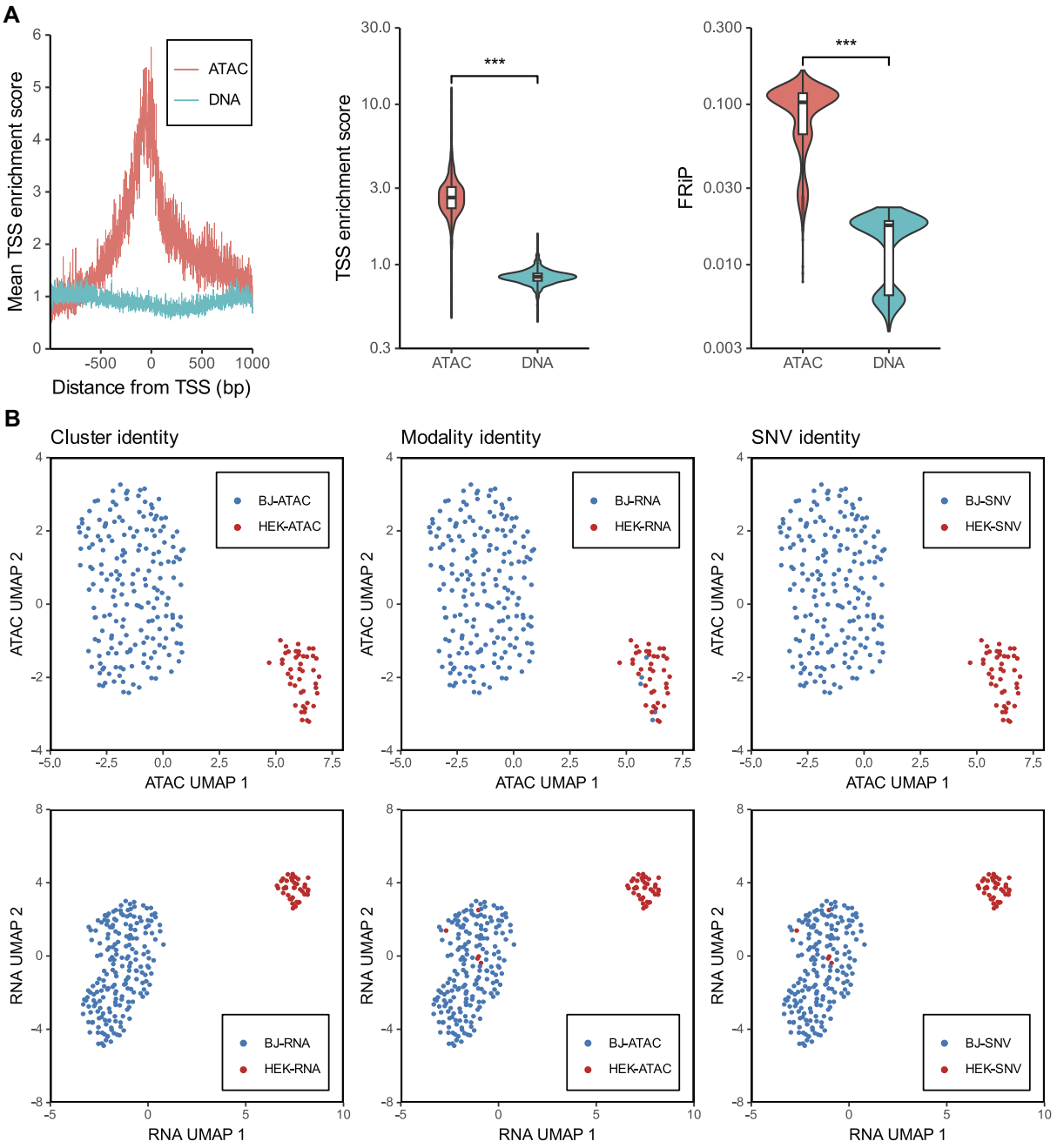


Figure 4. Cell identity assignment from RNA and open chromatin modalities with sci-L3-RNA/ATAC-seq. (**A**) ATAC signal is enriched over TSS and within peaks. ATAC signal enrichment centered over the TSS ± 1 kb, with the DNA signal from sci-L3-RNA/DNA-seq plotted as control (left plot). Significant enrichment of TSS signal, tested with Mann–Whitney *U* (middle plot) (*** ≤ .001; *P*-value: 2.12×10^{-274}; ATAC: median 2.62, MAD 0.58; DNA: median 0.84, MAD 0.06). Fraction of reads within peaks (FRiP) is shown for all MACS2 called peaks, independently called for ATAC and DNA showing significantly higher fraction in ATAC, tested with Mann–Whitney *U* (right plot) (*** ≤ .001; *P*-value: 3.52×10^{-271}; ATAC: median 0.103, MAD 0.029; DNA: median 0.018, MAD 0.003). (**B**) UMAP projection using ATAC and RNA modalities separates cells by identity. Separation of cells into individual populations with UMAP was used to label cells as either BJ or HEK for RNA and ATAC (left panel). Using these initial labels, the cell identities from ATAC were overlaid onto the RNA UMAP projection and vice versa (middle panel). As ground truth, the cell identities were assigned based on private SNVs (right panel). The number of cells misassigned based on cluster and modality is shown in Supplementary Table S6. Only high-coverage cells are plotted.

RNA and ATAC) were in the low sequencing depth group (Supplementary Table S6). Limiting our analysis to only high sequencing depth cells resulted in 97.8% (233/238) of cells with concordant assignment by RNA and 100% (238/238) assignment by ATAC (Fig. 4B). Overall, we show that sci-L3-RNA/ATAC captures cell state features by both modalities that enable cell type assignment.

Discussion

Here we have described the extension of sci-L3 to three new methods that together demonstrate the versatility of the toolkit and its potential to scale up low-throughput assays with combinatorial indexing and linear amplification. Empirically under the sci framework, adding additional rounds of cell barcoding typically reduces the number of recovered sin-

gle cells expected by the number of cells sorted, as a result of loss of per cell coverage. One general advantage of sci-L3 is the high sorted cell recovery rate of around 90% with three levels of barcoding, consistent across all extensions of this framework, which for other two-level sci methods is usually around 60% and even lower with additional rounds of barcoding. We think that this is due to the improved uniformity and coverage by linear amplification. Below we summarize features and limitations for the three extended methods.

We establish that sci-L3-Strand-seq generates high-quality libraries with low background for more than half of the sequenced cells. We also find that adding UDG + EndoVIII after UV treatment is not only helpful but also necessary to obtain a Strand-seq library. Note that although our assay uses IVT-based linear amplification, which is distinct from PCR polymerase used in the original development of the Strand-seq [1, 4], the T7 RNA polymerase cannot bypass single-stranded nick or 1-nt gaps; therefore, we think that the necessity of UDG + EndoVIII is unlikely to be IVT-specific. The ease to leverage the inherent cell sorting step in sci-L3 to enrich for BrdU-labeled cells of the right cell cycle, the increased throughput, and the improved uniformity of whole-genome amplification substantially advance the Strand-seq applicability. We foresee that this assay will be vastly useful for profiling mitotic crossovers including the genetically silent sister chromatid exchange, as well as other types of genome rearrangement [3].

With sci-L3-Hi-C, we captured previously described features of genome organization such as AB compartments and showed that our three-level barcoding has comparable performance to the previous two-level barcoding of sci-Hi-C [13]. Without ligation junction enrichment, the combination of WGS and chromatin conformation from the same single cell in one experiment provides a higher sequencing coverage per cell to aid in the annotation of structural and copy number variation. Other single-cell Hi-C methods without ligation junction enrichment utilize more frequent cutters by 2–4×, obtaining a higher capture of *cis* interactions, which provides an easy way to improve future implementations of sci-L3-Hi-C.

Lastly, we show that sci-L3-RNA/ATAC captures distinguishing features of cell identity from both modalities based on the high agreement between the cell type assignment from RNA and ATAC. We establish the capture of open chromatin with a 5× enrichment of reads over TSS as compared to sci-L3-WGS, and a significant improvement by five-fold of the single-cell ATAC performance in such co-assays with two-level indexing [17]. However, a limitation of the current approach is the low recovery of RNA reads per cell. We speculate that RT enzymes may also use DNA as a template and generate double-stranded cDNA in reverse transcription. They could then be inserted by Tn5 in a single-ended fashion, preventing further amplification. Direct RNA library preparation without reverse transcription [57] has the potential to significantly improve the performance of RNA/ATAC co-assay. Nevertheless, as small amounts of RNA reads typically inform cell types quite well, the sci-L3-ATAC/RNA co-assay is still useful in applications that aim at exploring new biology on the open chromatin aspect and only require crude cell type information from the transcriptome. Single-cell transcriptome analysis has seen an explosion of applications. Where single-cell RNA atlases are already available, one can easily use a reference-based approach to assign single cells to an annotated mini-bulk or lineages for a large number of cells, while obtaining high-quality accompanying single-cell ATAC data from the same cells.

Overall, the ability to use the same reagents and barcoding scheme across a multitude of different techniques makes the sci-L3 suite of methods an attractive option for a wide variety of biological questions. While we have shown three new methods here, we envision sci-L3 can be further extended to DNA methylation [58], chromatin-associated factors [59, 60], and combinations thereof such as RNA/Hi-C [61] and RNA/Strand-seq co-assays.

Acknowledgements

We thank Jay Shendure and Shendure lab members Junyue Cao, Anh Leith, Choli (Charlie) Lee, and Leonid Kruglyak and Kruglyak lab members Heriberto Marquez and Joshua Bloom for help with sequencing and helpful advice, Kathrin Plath and Plath lab member Amanda J. Collier and the UCLA Broad Stem Cell Research Center Flow Cytometry Core for flow analysis, David Porubsky for advice on initial running of the breakpointR, and Trevor Ridgley for his contribution to phasing Strand-seq data and helpful discussion.

Author contributions: Conceptualization: P.C., Y.Y.; Data curation: P.C, Y.Y.; Formal analysis: P.C.; Funding acquisition: Y.Y.; Investigation: P.C., Y.Y.; Methodology: P.C., Y.Y.; Supervision: Y.Y.; Writing - original draft: P.C., Y.Y.; Writing - review & editing: P.C., Y.Y.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was funded by the NIH (National Institute of General Medical Sciences, R35GM142511 to Y.Y.) and the Damon Runyon Cancer Research Foundation (Damon Runyon–Dale F. Frey Award for Breakthrough Scientists to Y.Y., DFS-43-20). Funding to pay the Open Access publication charges for this article was provided by NIGMS 5R35GM142511.

Data availability

Sequencing data are publicly available in GEO under accession GSE281238. Processed data and analysis scripts are available on Zenodo under DOI: 10.5281/zenodo.14009810.

References

1. Sanders AD, Falconer E, Hills M *et al.* Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat Protoc* 2017;**12**:1151–76. https://doi.org/10.1038/nprot.2017.029
2. Yin Y, Jiang Y, Lam K-WG *et al.* High-throughput single-cell sequencing with linear amplification. *Mol Cell* 2019;**76**:676–90.e10. https://doi.org/10.1016/j.molcel.2019.08.002

3. Chovanec P, Yin Y. A mapping platform for mitotic crossover by single-cell multi-omics. *Methods Enzymol* 2021;**661**:183–204. https://doi.org/10.1016/bs.mie.2021.08.017
4. Falconer E, Hills M, Naumann U *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods* 2012;**9**:1107–12. https://doi.org/10.1038/nmeth.2206
5. Porubsky D, Garg S, Sanders AD *et al.* Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun* 2017;**8**:1293. https://doi.org/10.1038/s41467-017-01389-4
6. Sanders AD, Hills M, Porubský D *et al.* Characterizing polymorphic inversions in human genomes by single-cell sequencing. *Genome Res.* 2016;**26**:1575–87. https://doi.org/10.1101/gr.201160.115
7. Sanders AD, Meiers S, Ghareghani M *et al.* Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat Biotechnol* 2020;**38**:343–54. https://doi.org/10.1038/s41587-019-0366-x
8. Adey AC. Tagmentation-based single-cell genomics. *Genome Res* 2021;**31**:1693–705. https://doi.org/10.1101/gr.275223.121
9. Hanlon VCT, Chan DD, Hamadeh Z *et al.* Construction of Strand-seq libraries in open nanoliter arrays. *Cell Rep Methods* 2022;**2**:100150. https://doi.org/10.1016/j.crmeth.2021.100150
10. Dekker J, Rippe K, Dekker M *et al.* Capturing chromosome conformation. *Science* 2002;**295**:1306–11. https://doi.org/10.1126/science.1067799
11. Lieberman-Aiden E, van Berkum NL, Williams L *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93. https://doi.org/10.1126/science.1181369
12. Eagen KP. Principles of chromosome architecture revealed by Hi-C. *Trends Biochem Sci* 2018;**43**:469–78. https://doi.org/10.1016/j.tibs.2018.03.006
13. Ramani V, Deng X, Qiu R *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* 2017;**14**:263–6. https://doi.org/10.1038/nmeth.4155
14. Nagano T, Várnai C, Schoenfelder S *et al.* Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol* 2015;**16**:175. https://doi.org/10.1186/s13059-015-0753-7
15. Vandereyken K, Sifrim A, Thienpont B *et al.* Methods and applications for single-cell and spatial multi-omics. *Nat Rev Genet* 2023;**24**:494–515. https://doi.org/10.1038/s41576-023-00580-2
16. Clark SJ, Argelaguet R, Kapourani C-A *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun* 2018;**9**:781. https://doi.org/10.1038/s41467-018-03149-4
17. Cao J, Cusanovich DA, Ramani V *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 2018;**361**:1380–5. https://doi.org/10.1126/science.aau0730
18. Ma S, Zhang B, LaFave LM *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* 2020;**183**:1103–16. https://doi.org/10.1016/j.cell.2020.09.056
19. Tan L, Xing D, Chang C-H *et al.* Three-dimensional genome structures of single diploid human cells. *Science* 2018;**361**:924–8. https://doi.org/10.1126/science.aat5641
20. Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with Snakemake. *F1000Research* 2021;**10**:33. https://doi.org/10.12688/f1000research.29032.2
21. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**:10. https://doi.org/10.14806/ej.17.1.200
22. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 2009;**25**:1754–60. https://doi.org/10.1093/bioinformatics/btp324
23. Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008
24. Porubsky D, Sanders AD, Taudt A *et al.* breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics* 2020;**36**:1260–1. https://doi.org/10.1093/bioinformatics/btz681
25. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. *Sci Rep* 2019;**9**:9354. https://doi.org/10.1038/s41598-019-45839-z
26. Deng C, Daley T, Smith AD. Applications of species accumulation curves in large-scale biological data analysis. *Quant Biol* 2015;**3**:135–44. https://doi.org/10.1007/s40484-015-0049-7
27. Hanlon VCT, Porubsky D, Lansdorp PM. Chromosome-length haplotypes with StrandPhaseR and Strand-seq. *Methods Mol Biol* 2023;**2590**:183–200. https://doi.org/10.1007/978-1-0716-2819-5_12
28. Keane TM, Goodstadt L, Danecek P *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 2011;**477**:289–94. https://doi.org/10.1038/nature10413
29. Open2C, Abdennur N, Fudenberg G, Flyamer IM *et al.* Pairtools: From sequencing data to chromosome contacts. *PLoS Comput Biol* 2024;**20**:e1012164. https://doi.org/10.1371/journal.pcbi.1012164
30. Lee S, Bakker CR, Vitzthum C *et al.* Pairs and Pairix: a file format and a tool for efficient storage and retrieval for Hi-C read pairs. *Bioinformatics* 2022;**38**:1729–31. https://doi.org/10.1093/bioinformatics/btab870
31. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* 2020;**36**:311–6. https://doi.org/10.1093/bioinformatics/btz540
32. Durand NC, Shamim MS, Machol I *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* 2016;**3**:95–8. https://doi.org/10.1016/j.cels.2016.07.002
33. Imakaev M, Fudenberg G, McCord RP *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 2012;**9**:999–1003. https://doi.org/10.1038/nmeth.2148
34. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 2009;**25**:1841–2. https://doi.org/10.1093/bioinformatics/btp328
35. Open 2C, Abdennur N, Abraham S, Fudenberg G *et al.* Cooltools: Enabling high-resolution Hi-C analysis in Python. *PLoS Comput Biol* 2024;**20**:e1012067. https://doi.org/10.1371/journal.pcbi.1012067
36. Bernstein B. ENCSR057BWO. *ENCODE Datasets* 2016. https://doi.org/10.17989/encsr057bwo
37. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247
38. Luo Y, Hitz BC, Gabdank I *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* 2020;**48**:D882–9. https://doi.org/10.1093/nar/gkz1062
39. Rao SSP, Huntley MH, Durand NC *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014;**159**:1665–80. https://doi.org/10.1016/j.cell.2014.11.021
40. Mulqueen RM, Pokholok D, O'Connell BL *et al.* High-content single-cell combinatorial indexing. *Nat Biotechnol* 2021;**39**:1574–80. https://doi.org/10.1038/s41587-021-00962-z
41. Chang L, Xie Y, Taylor B *et al.* Droplet Hi-C enables scalable, single-cell profiling of chromatin architecture in heterogeneous tissues. *Nat Biotechnol* 2024;https://doi.org/10.1038/s41587-024-02447-1
42. Kim H-J, Yardımcı GG, Bonora G *et al.* Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput Biol* 2020;**16**:e1008173. https://doi.org/10.1371/journal.pcbi.1008173
43. Chen C, Xing D, Tan L *et al.* Single-cell whole-genome analyses by Linear Amplification via Transposon insertion (LIANTI). *Science* 2017;**356**:189–94. https://doi.org/10.1126/science.aak9787
44. Lin Y-C, Boone M, Meuris L *et al.* Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology

manipulations. *Nat Commun* 2014;**5**:4767. https://doi.org/10.1038/ncomms5767
45. Stuart T, Srivastava A, Madad S *et al.* Single-cell chromatin state analysis with Signac. *Nat Methods* 2021;**18**:1333–41. https://doi.org/10.1038/s41592-021-01282-5
46. Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87. https://doi.org/10.1016/j.cell.2021.04.048
47. Zhang Y, Liu T, Meyer CA *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137. https://doi.org/10.1186/gb-2008-9-9-r137
48. Melville J. The uniform manifold approximation and projection (UMAP) method for dimensionality reduction. R package uwot version 0.2.2, 2024. https://doi.org/10.32614/CRAN.package.uwot
49. Frankish A, Diekhans M, Ferreira A-M *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2019;**47**:D766–73. https://doi.org/10.1093/nar/gky955
50. Perry P, Wolff S. New Giemsa method for the differential staining of sister chromatids. *Nature* 1974;**251**:156–8. https://doi.org/10.1038/251156a0
51. Hutchinson F. The lesions produced by ultraviolet light in DNA containing 5-bromouracil. *Q Rev Biophys* 1973;**6**:201–46. https://doi.org/10.1017/S0033583500001141
52. Nagano T, Lubling Y, Stevens TJ *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**:59–64. https://doi.org/10.1038/nature12593
53. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 2015;**72**:65–75. https://doi.org/10.1016/j.ymeth.2014.10.031
54. Zhang H, Rice ME, Alvin JW *et al.* Extensive evaluation of ATAC-seq protocols for native or formaldehyde-fixed nuclei. *BMC Genomics* 2022;**23**:214. https://doi.org/10.1186/s12864-021-08266-x
55. Lareau CA, Ludwig LS, Muus C *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat Biotechnol* 2021;**39**:451–61. https://doi.org/10.1038/s41587-020-0645-6
56. Butler A, Hoffman P, Smibert P *et al.* Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**:411–20. https://doi.org/10.1038/nbt.4096
57. Lyu J, Chen C. LAST-seq: single-cell RNA sequencing by direct amplification of single-stranded RNA without prior reverse transcription and second-strand synthesis. *Genome Biol* 2023;**24**:184. https://doi.org/10.1186/s13059-023-03025-5
58. Mulqueen RM, Pokholok D, Norberg SJ *et al.* Highly scalable generation of DNA methylation profiles in single cells. *Nat Biotechnol* 2018;**36**:428–31. https://doi.org/10.1038/nbt.4112
59. Wang Q, Xiong H, Ai S *et al.* CoBATCH for high-throughput single-cell epigenomic profiling. *Mol Cell* 2019;**76**:206–16. https://doi.org/10.1016/j.molcel.2019.07.015
60. Janssens DH, Greene JE, Wu SJ *et al.* Scalable single-cell profiling of chromatin modifications with sciCUT&Tag. *Nat Protoc* 2024;**19**:83–112. https://doi.org/10.1038/s41596-023-00905-9
61. Zhou T, Zhang R, Jia D *et al.* GAGE-seq concurrently profiles multiscale 3D genome organization and gene expression in single cells. *Nat Genet* 2024;**56**:1701–11. https://doi.org/10.1038/s41588-024-01745-3

Received: October 30, 2024. **Revised:** December 28, 2024. **Editorial Decision:** January 28, 2025. **Accepted:** February 5, 2025