Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells

Vijay Ramani[a,*], Xinxian Deng[b], Ruolan Qiu[a], Choli Lee[a], Christine M. Disteche[b,c], William S. Noble[a], Jay Shendure[a,f,*], Zhijun Duan[d,e,*]

[a] Department of Genome Sciences, University of Washington, Seattle, WA, United States
[b] Department of Pathology, University of Washington, Seattle, WA, United States
[c] Department of Medicine, University of Washington, Seattle, WA, United States
[d] Division of Hematology, University of Washington School of Medicine, Seattle, WA, United States
[e] Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA, United States
[f] Howard Hughes Medical Institute, Seattle, WA, United States

ARTICLE INFO

ABSTRACT

The highly dynamic nature of chromosome conformation and three-dimensional (3D) genome organization leads to cell-to-cell variability in chromatin interactions within a cell population, even if the cells of the population appear to be functionally homogeneous. Hence, although Hi-C is a powerful tool for mapping 3D genome organization, this heterogeneity of chromosome higher order structure among individual cells limits the interpretive power of population based bulk Hi-C assays. Moreover, single-cell studies have the potential to enable the identification and characterization of rare cell populations or cell subtypes in a heterogeneous population. However, it may require surveying relatively large numbers of single cells to achieve statistically meaningful observations in single-cell studies. By applying combinatorial cellular indexing to chromosome conformation capture, we developed single-cell combinatorial indexed Hi-C (sci-Hi-C), a high throughput method that enables mapping chromatin interactomes in large number of single cells. We demonstrated the use of sci-Hi-C data to separate cells by karytoypic and cell-cycle state differences and to identify cellular variability in mammalian chromosomal conformation. Here, we provide a detailed description of method design and step-by-step working protocols for sci-Hi-C.

1. Introduction

Recent advances in sequencing-based high-throughput genomic technologies, such as Hi-C [1], have revealed that the genome is hierarchically organized in the nucleus of eukaryotic cells [2]. In mammalian cells, this hierarchy of three-dimensional (3D) genome organization is characterized by multilevel architectural features, such as chromosome territories [3], A/B compartments [1], topologically associated domains (TADs) [4,5], and long-range chromatin loops between specific loci [6]. The existence of these commonly shared conformational signatures in individual cells has further been demonstrated by recent super-resolution microscopy studies [7,8]. Moreover, increasing evidence has demonstrated that the 3D genome organization critically impacts various nuclear, cellular and developmental events, such as transcription, DNA replication, cell cycle and cell

fate decisions, and that defects in the 3D genome organization can lead to disease phenotypes [9].

However, despite the existence of commonly shared structural features in 3D genome organization, it is important to note that, due to the stochastic nature of the chromatin fiber and the variability in the various nuclear processes (e.g., DNA replication and transcription) between individual cells, the spatial organization of the genome varies considerably, even among cells of a functionally homogeneous population [10]. For example, early microscopy studies demonstrated that, although interphase chromosomes occupy distinct and relative static territories, such chromosome territories are also dynamic during cell cycle and differentiation [11–14]. Hence, the 3D genome organization constructed from ensemble Hi-C represents only an average of millions of cells, but not the actual genome organizations in individual cells. Unlike bulk Hi-C, single-cell Hi-C enables to construct whole genome

Table 1
sci-Hi-C protocol improvements.

Experimental steps	Initial version (Ramani et al., 2017)	Current version	How does it improve the protocol
Reverse crosslinking	SDS + proteinase K	2× NEBuffer 2, no proteinase k is needed	Reduced cost
Digestion of genomic DNA	MseI + HaeIII	AluI + HaeIII	As both AluI and HaeIII produce blunt-ended DNA fragments, the downstream end-repair step therefore can be omitted
End-repair	Needed	Omitted	Simplified the protocol
Ampure beads purification	Carried out before the digestion of genomic DNA	Carried out at the end of library construction (i.e., after all the enzymatic reactions done)	1) Simplified the protocol; 2) Easier to carry out; 3) Improved the efficiency of the enzymatic reactions

structures for single cells [15,16]. To characterize the 3D genome architecture at the single-cell resolution, we recently developed sci-Hi-C, a massively multiplexed single-cell Hi-C method for mapping the 3D genome in large number of single cells in parallel without requiring physical compartmentalization of each cell or microfluidic manipulation [17]. We have demonstrated that sci-Hi-C is able to separate different cell types on the basis of single cell Hi-C maps and identified previously uncharacterized cell-to-cell heterogeneity in the conformational properties of mammalian chromosomes, enabling "in silico sorting" of mixed cell populations by cell-cycle stage[17–19]. We have since improved the initial version of our sci-Hi-C protocol, which is now greatly simplified, easier-to-adapt, and more cost-effective and have successfully applied sci-Hi-C to a diversity of mouse and human cell lines [17–19] (Table 1).

2. Method development

Sci-Hi-C [17] was developed by extending the "combinatorial indexing" paradigm, which was first employed in the single-cell ATAC-seq method to measure chromatin accessibility in thousands of single cells per experiment [20]. Similar to ensemble Hi-C and the other single-cell Hi-C protocols [15,16,21–23], the sci-Hi-C protocol involves cross-linking the cells with formaldehyde. The cells are then permeabilized with their nuclei intact. A 4-cutter restriction enzyme DpnII is used to fragment the chromatin. Nuclei are then distributed to a 96-well plate, wherein the first-round barcode is introduced through ligation of barcoded biotinylated bridge adaptors. Intact nuclei are then pooled and subjected to proximity ligation, followed by dilution and redistribution to a second 96-well plate. Importantly, no more than 25 nuclei should be distributed to each well of the second plate. Following reverse crosslinking, restriction digestion, end repair and dA tailing, the second-round barcode is then introduced through ligation of barcoded Y-adaptors g. Because the number of barcode combinations (96 × 96, when 96-well plates are used) are much bigger than the number of nuclei (96 × 25) in the second 96-well plate, each single nucleus in the second 96-well plate is therefore labeled by a unique combination of barcodes. All genomic DNA materials are once again pooled, and biotinylated DNA fragments are purified with streptavidin beads and PCR amplified to generate Illumina sequencing libraries (Fig. 1). Finally, high throughput sequencing with relatively long paired-end reads (i.e., 2 × 250 base pairs) is performed, which allows one to identify not only the chromatin interactions of conventional Hi-C, but also both rounds of the barcodes, thereby enabling decomposition of the Hi-C data into single-cell chromatin contact maps. It is worth noting that the throughput (i.e., the number of single cells surveyed per experiment) of sci-Hi-C can readily be scaled up by using 384-well plates to replace the 96-well plates during the two-rounds of barcoding.

2.1. Cell fixation and cell lysis

In this protocol, cells are crosslinked with 2% formaldehyde for 10 min at room temperature (25 °C). However, some cell types (e.g., mouse and human embryonic stem cells) may require different

crosslinking conditions and tissues also need to be homogenized. Optimization of crosslinking conditions for a new cell type can be carried out as described previously in [24,25]. To open the chromatin for enzymatic reactions, while retaining the 3D chromatin conformation and the nuclei intactness, crosslinked cells are then treated sequentially with hypotonic buffer containing IGEPAL CA-630 and buffers containing detergents SDS and Triton X-100. Different cell types may require different SDS concentration, which can be titrated as described previously [24,25].

2.2. Chromatin fragmentation

Sci-Hi-C protocol employs a four-cutter enzyme, DpnII, to fragment the crosslinked chromatin. DpnII digestion is carried out in the presence of detergents (i.e., SDS and Triton X-100) and the size of the majority of the resulting genomic DNA fragments should be less than 1 kb (Fig. 2A). It is worth pointing out that not all restriction enzymes are active under these conditions. However, if an enzyme has been used successfully in standard Hi-C protocols, it will likely work for sci-Hi-C. It is also important to note that, since a four-cutter enzyme (e.g., DpnII) has much more cut sites in the genome than a six-cutter enzyme (e.g, HindIII), using a four cutter to fragment the chromatin in principle will generate higher-resolution chromatin interaction maps than using a six cutter.

2.3. Bridge adaptor and 1st round barcode design.

In the sci-Hi-C protocol, the bridge adaptor is designed to possess several unique features (Fig. 2B). First, it is biotinylated through including a biotin-dT, allowing for enrichment of chromatin interactions with streptavidin-coupled magnetic beads. Second, the two ends of the bridge adaptor are designed to play different roles. One of its ends, through which the adaptor can be ligated to the DpnII digested chromatin fragments, possesses a 5′-GATC overhang. The other end of the adaptor, which is designed to mediate proximity ligation, is not 5′-phosphorylated and its 3′-overhang contains a biotin-dT within a palindromic sequence, which is part of the recognition sequence of EcoRI. Hence, only after the phosphorylation step (Step 37 in the Detailed protocols Section), two adaptor-tagged chromatin fragments can ligate with each other through this end (steps 38–39 in the Detailed protocols Section), which will result in both biotin-labeling and the formation of an EcoRI site) at the junction. And finally, the bridge adaptors also harbor the 1st round barcodes, which have been described in detail previously [17].

2.4. Removal of the excess adaptors after the 1st round barcoding

In the sci-Hi-C protocol, the biotin marking of chromatin fragments and the introduction of the 1st round cellular indexing is achieved simultaneously by ligation of the indexed- and biotinylated-bridge adaptors to the chromatin ends produced by DpnII digestion. It is necessary to clear out the excess free adaptors after the ligation reaction, which otherwise will severely reduce the efficiency of the downstream proximity ligation. We found that multiple rounds of washing of the
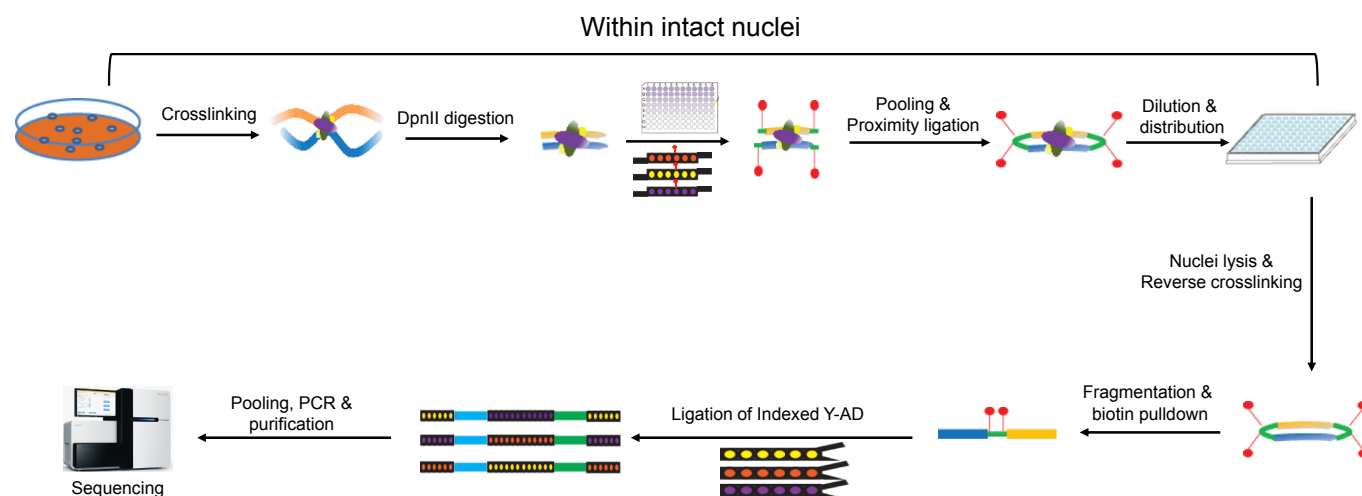
Within intact nuclei

Fig. 1. The workflow of sci-Hi-C.

nuclei with buffers containing SDS and tween-20 can lead to efficient removal of the excess adaptors.

2.5. Quality control of sci-Hi-C libraries

During the construction of sci-Hi-C libraries, quality control (QC) analysis should be carried out at several critical steps.

First, the DpnII digestion efficiency can be examined by DNA gel electrophoresis (Fig. 2A). Genomic DNA can be isolated from a small portion (5–10%) of DpnII digested sample, and the size distribution of the DNA fragments determined by running a 1% agarose gel or a precast 6% polyacrylamide DNA gel. Usually an efficient DpnII digestion

will result in the majority of DNA fragments with a size < 1 kb (Fig. 2A).

Second, the efficiency of the DNA phosphorylation and proximity ligation can also be examined by DNA gel electrophoresis (Fig. 2A). It is worth noting that, the size distribution of the genomic DNAs after proximity ligation should shift significantly towards bigger sizes compared to that of the genomic DNAs isolated from the experimental steps upstream the step of proximity ligation (Fig. 2A).

And finally, as mentioned above, the biotinylated- and indexed-bridge adaptor for marking the chromatin ends in the sci-Hi-C protocol is designed in such a way that an EcoRI enzyme site will be formed at the junction when two chromatin fragments ligate with each other
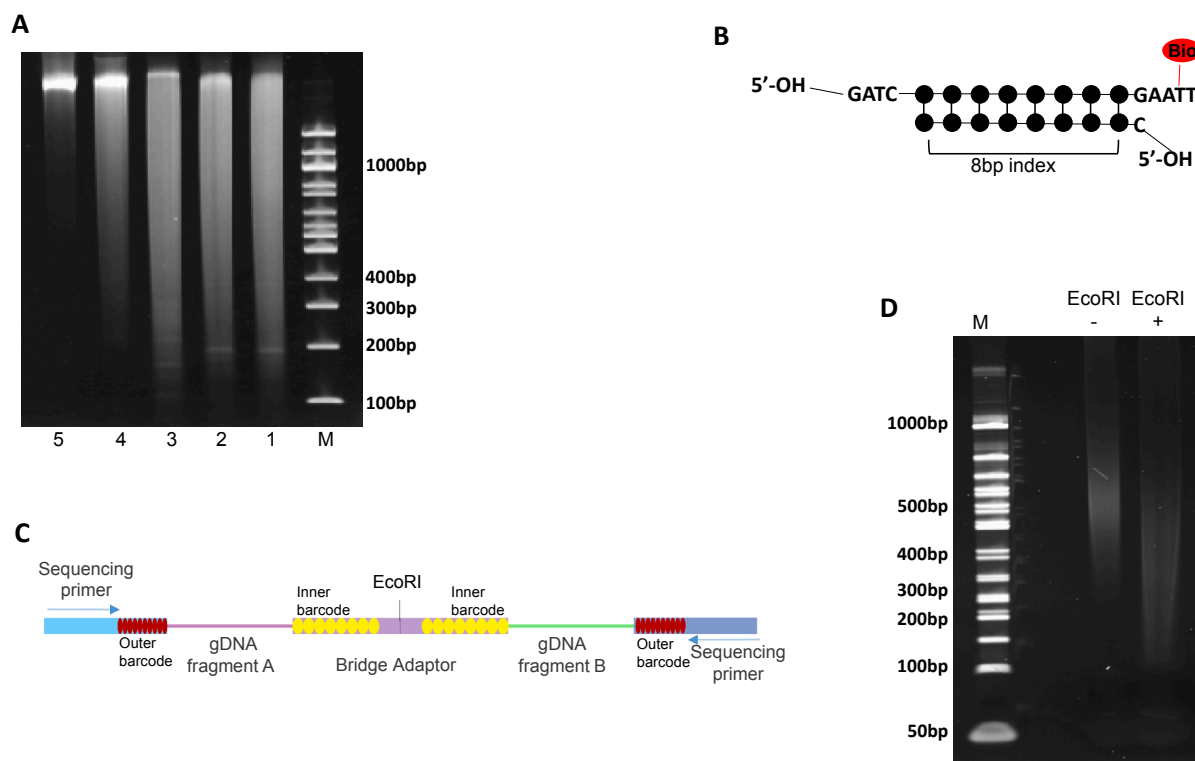
Fig. 2. Quality control (QC) of experimental steps in the sci-Hi-C protocol. (A) Example showing the size distribution of genomic DNA samples purified from the nuclei at the indicated various experimental steps of the sci-Hi-C protocol. 1–3, genomic DNA isolated from GM12878 (1), HFF (2) and NIH3T3 (3) cells following DpnII digestion (at 37 °C for 20 h), 4, genomic DNA isolated from the nuclei following the ligation of the biotynated- and indexed-bridge adaptors, 5, genomic DNA isolated from the nuclei following the proximity ligation. M, the 100 bp DNA ladder. (B) The structure of the biotinylated- and indexed- bridge adaptors designed for sci-Hi-C. (C) Schematic depiction of the structure of the chimeric DNA molecules in sci-Hi-C libraries. (D) Example of gel electrophoresis showing a sci-Hi-C library before (-) and after (+) EcoRHI digestion. M, the 50 bp DNA ladder. All reactions were run on one 6% TBE-PAGE gel.

through the bridge adaptors during the proximity ligation. Hence, in sci- Hi-C libraries, each valid DNA fragment is chimeric, consisting of two interacting genomic fragments via the bridge adaptor with an EcoRI site at the joint (Fig. 2C). Therefore, sci-Hi-C libraries can be validated by carrying out an EcoRI digestion assay, i.e., for a valid sci-Hi-C library, there should be an apparent size shift of DNA fragments after EcoRI digestion (Fig. 2D).

2.6. Library sequencing

Because the 1st round of indexes are located in the middle of the DNA fragments in a sci-Hi-C library, it requires relatively long paired-end reads to identify these indexes. Therefore, sci-Hi-C libraries are usually sequenced using 250 bp paired-end sequencing on an Illumina Hiseq 2500 instrument.

3. Detailed protocols

3.1. Adaptor preparation

The sequences of all the adaptors can be found in Ramani et al [17], and all adaptors can be ordered from Integrated DNA Technologies, Inc. (IDT).

- 1. Setup the 96-well plate for the bridge adaptors (the final concentration is 45 μM),

Fill-in mix (add in this order)	Per well
10× NEBuffer #2 (NEB, CAT# B7002S)	10 μl
Indexed Forward oligo (100 μM)	45 μl
Indexed Reverse oligo (100 μM)	45 μl

- 2. Setup the 96-well plate for the Y adaptors (the final concentration is 20 μM),

Fill-in mix (add in this order)	Per well
Water	50 μl
10× NEBuffer #2	10 μl
Indexed Forward oligo (100 μM)	20 μl
Indexed Reverse oligo (100 μM)	20 μl

- 3. Anneal the mixtures by heating them to 98 °C for 6 min in a thermocycler, and then turn off the instrument to allow the plates to naturally cool to RT for overnight. Annealed adaptors can be kept at −20 °C.

3.2. Cell lysis and chromatin digestion with DpnII

- 4. Thaw a vial of 2% formaldehyde-fixed human cells (e.g., HFF c6, GM12878, K562, MIR90, Hela S3 and Hap-1) and mouse cells (e.g., Patski, NIH3T3 and MEFs cells) at 37 °C for 10–20 s.
- 5. Resuspend in 1 ml of ice-cold cell lysis buffer (10 mM Tris-HCl pH8.0, 10 mM NaCl, 0.2% Igepal CA-630) containing protease inhibitor cocktail (Sigma-Aldrich, Cat#4693132001).
- 6. Transfer about 2.5 million of each type of cells to a new tube, respectively.
- 7. Incubate at 0 °C for 60 min.
- 8. Centrifuge for 60 s at 850×g at room temperature (RT).
- 9. Discard the supernatant and resuspend the pellet in each tube in 100 μl SDS buffer (0.5% SDS in 0.1× DpnII buffer (NEB #B7006)).
- 10. Incubate at 60 °C for 10 min.
- 11. Add 100 μl of water and 25 μl of 10% Triton X-100 to each tube, mix well.
- 12. Incubate at 37 °C for 10 min.

- 13. Add 40 μl 10× DpnII buffer, 10 μl DpnII (50units/ul, NEB R0543M) and 125 μl water to each tube, mix well.
- 14. Incubate at 37 °C overnight (18–20 h).
- 15. Centrifuge for 60 s at 850×g at RT.
- 16. Discard the supernatant and resuspend the pellet in 400 μl water and 3 μl 10 mg/ml BSA (ThermoFisher Scientific, AM2616) in each tube, mix well.
- 17. Centrifuge for 60 s at 850×g at RT.
- 18. Discard the supernatant and resuspend the pellet in 600 μl water and 6 μl 10 mg/ml BSA in each tube, mix well. (Note, the volume for cell resuspension dependent on the cell number and the number of 96-well plates used in the next step; the number listed here is for 2 96-well plates).

3.3. Ligation of the biotin-labeled, barcoded bridge adaptors

- (~25,000–100,000 cells /per well, 10 μl reaction volume, 2–4 96-well plates)
- 19. Distribute T4 DNA ligase/buffer to each well of 2–4 plates (below is for 2 plates, 3 μl per well),

Fill-in mix (add in this order)	Per well
water	1.1 μl
50% PEG-4000	0.8 μl
10× T4 ligase buffer (NEB B0202s)	1 μl
T4 DNA ligase (5 U/μl, Thermo Scientific EL0012)	0.1 μl

- 20. Distribute the Bridge adaptor to each well of a 96-well plate using a multi-channel pipette, 2 μl per well.
- 21. Distribute 5 μl of the above DpnII digested nuclei to each well as below: human cells in one 96-well plate and mouse cells in the other plate. If there are more than two cell types in each specie, the cells can be distributed into different rows of the 96-well plate.
- 22. Incubate at 16 °C overnight.
- 23. Distribute 1 μl 0.25 M EDTA to each well to stop the reaction.
- 24. On ice, pool the samples from each well to two 1.5 ml tubes.
- 25. Centrifuge at 850×g at RT for 90 s.
- 26. Resuspend the pellet in 294 μl of water, 3 μl 10%SDS, and 3 μl 10 mg/ml BSA, combine the tubes into one.
- 27. Incubate at RT for 2 min.
- 28. Centrifuge at 850×g at RT for 60 s.
- 29. Resuspend the pellet in 295.5 μl of 2 mM MgCl₂, 1.5 μl 2% tween-20, and 3 μl 10 mg/ml BSA.
- 30. Centrifuge at 850×g at RT for 60 s.
- 31. Repeat steps 29 and 30 one more time.
- 32. Resuspend the pellet in 294 μl of 2 mM MgCl₂, 3 μl 10%SDS, and 3 μl 10 mg/ml BSA.
- 33. Centrifuge at 850×g at RT for 60 s.
- 34. Repeat steps 32 and 33 one more time.
- 35. Resuspend the pellet in 294 μl of 2 mM MgCl₂, 3 μl 10%SDS and 3 μl 10 mg/ml BSA, mix well.

3.4. Phosphorylation and proximity ligation

- 36. Prepare the PNK-ligation mastermix as follows:

Fill-in mix (add in this order)	Per reaction
water	50 μl
Nuclei	20 μl
10× T4 ligase buffer	10 μl
PNK (10 U/μl, Thermo Scientific, EK0032)	15 μl
10% Triton X-100	5 μl

- 37. Incubate at 37 °C for 1 h. Three reactions may be carried out in parallel.
- 38. Add the following to the above tube,

water	172 µl
10× T4 ligase buffer	20 µl
T4 DNA ligase (5 U/µl)	5 µl
10 mg/ml BSA	3 µl

- 39. Mix well, incubate at RT for 4 h.
- 40. Add 5 µl 10% SDS to each reaction.

3.5. Serial dilution and reverse cross-linking

- 41. Carry out serial dilution as follows: transfer 100 µl solution from step 40 to 885 µl water plus 10 µl 10 mg/ml BSA and 5 µl10% SDS, mix well. Transfer 100 µl of the diluted nuclei to 885 µl water plus 10 µl 10 mg/ml BSA and 5 µl 10% SDS, mix well. Then transfer 200 µl of the diluted nuclei to 196 µl water plus 100 µl 10× NEBuffer 2 and 4 µl 10 mg/ml BSA, mix well. Now there are about 2–4 cells per 1 µl solution.
- 42. Distribute 5 µl per well to a low-DNA binding 96-well PCR plate (Bio-Rad, HSS-9601).
- 43. Incubate at 65 °C for overnight. Plates can then be stored at −20 °C for future use.

3.6. Genomic DNA digestion with AluI + HaeIII

- 44. add the following to each well of the 96-well plate from the step 43:

Fill-in mix (add in this order)	Per well
Water	4.3 µl
10% Triton X-100	0.1 µl
AluI (10 unit/µl, NEB R0137L)	0.3 µl
HaeIII (10 unit/µl, NEB R0108L)	0.3 µl

- 45. Briefly spin the plate at 1000 rpm.
- 46. Incubate at 37 °C for 24 h.
- 47. Incubate at 65 °C for 30 min

3.7. dA-tailing

- 48. Add the following dA-tailing reaction reagent to each well of the 96-well plate:

Fill-in mix (add in this order)	Per well
10× NEBuffer 2	0.5 µl
20 mM dATP (Thermo Scientific, FERR0181)	1 µl
Klenow (exo⁻) (5 U/µl, Thermo Scientific, EP0422)	0.3 µl
water	3.2 µl

- 49. Briefly spin the plate at 1000 rpm.
- 50. Incubate at 37 °C for 45 min.
- 51. Incubate at 75 °C for 20 min.

3.8. Ligation of the indexed Illumina Y adaptors

- 52. Distribute 1 µl Indexed Y adaptor (20 µM) to each well of the 96-well plate.
- 53. Add the following ligation reaction reagent (14 µl) to each well of the 96-well plate, mix well:

Fill-in mix (add in this order)	Per well
water	7.8 µl
10× T4 ligation Buffer	3 µl
50% PEG-4000	3 µl
T4 DNA ligase (5 U/µl)	0.2 µl

- 54. Briefly spin the plate at 1000 rpm.
- 55. Incubate at 16 °C overnight.
- 56. Incubate at 70 °C for 10 min

3.9. Ampure beads purification and biotin pull down

- 57. For each 96-well plate from Step 56, dilute 2.5 ml Ampure beads (Beckman Coulter, A63881) with 3.5 ml 20%PEG-8000 buffer (20% PEG-8000 (Sigma-Aldrich, 89510-250G-F) in 2.5 M NaCl), Mix well.
- 58. Add 50 µl of the diluted Ampure beads to each well of the 96-well plate from the step 56.
- 59. Combine every 8 wells (each column) into a 1.5 ml low-binding tube, and 12 tubes for a 96-well plate.
- 60. Rinse each 8-wells with 60 µl EB. Now, the volume in each tube is 700 µl.
- 61. Incubate at RT for 2 min.
- 62. Wash two times with 80% Ethanol, combine the 12 tubes into one tube.
- 63. Elute DNA from the beads with 100 µl EB (10 mM Tris-HCl pH 8.5).
- 64. Bind to 10 µl DynaBeads M-280 (Thermo Fisher Scientific, 11205D) in 200 µl 1× B&W buffer (5 mM Tris-65. HCl pH8.0, 0.5 mM EDTA, 1 M NaCl).
- 65. Incubate at RT for 20 min.
- 66. Wash 4 times with 200 µl 1× B&W buffer containing 0.1% tween-20.
- 67. Wash twice with 200 µl EB.
- 68. Resuspend the beads in 40 µl EB in each tube.

3.10. Library amplification and purification

- 69. Set up 10 PCR reactions as below,

Fill-in mix (add in this order)	Per Reaction
Water	12 µl
2× NEBNext® Ultra™ II Q5® Master Mix (NEB, M0544L)	20 µl
Forward primer (10 uM)	2 µl
Reverse primer (10 uM)	2 µl
DNA/beads	4 µl

- 70. Carry out the following PCR program:
- 71. Using the following PCR program:

Cycle number	Denature	Anneal	Extend
1	98 °C, 2 min		
4	98 °C, 20 s	60 °C, 20 s	72 °C, 1 min
8–12	98 °C, 20 s	65 °C, 20 s	72 °C, 1 min

- 72. Pool all the samples from each PCR reaction into a 1.5 ml tube.
- 73. Add 100 µl Ampure beads and 300 µl of 20% PEG-8000 buffer to get the ratio of 1:1, mix well.
- 74. Incubate for 5 min at RT.
- 75. Wash once with 1 ml 80% Ethonal.
- 76. Resuspend the beads in 100 µl water and transfer the DNA to a new tube.
- 77. Add 80 µl 20% PEG solution, mix well.
- 78. Incubate for 5 min at RT.

- 79. Wash twice with 0.4 ml 80% ethonal.
- 80. Repeat steps 76–79 one more time if necessary.
- 81. Elute DNA with 30 µl water or EB.
- 82. Determine the DNA concentration using Qubit dsDNA HS kit (ThermoFisher Scientific, CAT# Q32851) according to manufacturer's instructions.

3.11. Library validation by EcoRI digestion

- 83. Digest a small aliquot of the final sci-Hi-C library with EcoRI to estimate the portion of molecules with valid biotinylated junctions.

Reagent	Experimental	Control
10× Fast digestion buffer	1 µl	1 µl
Hi-C product	1–2 µl	1–2 µl
	(5–10 ng)	(5–10 ng)
Fast digestion EcoRI (ThermoFisher Scientific, F-D0274)	1 µl	0 µl
Water	to 10 µl	to 10 µl

- 84. Incubate at 37 °C for 20 min.
- 85. Run the entire volume of the reaction on a 2% agarose gel (Fig. 3C).

4. Data processing and expected results

4.1. Sci-Hi-C processing pipeline

We developed a computational pipeline to process sci-Hi-C data, which is available at https://github.com/VRam142/combinatorialHiC. This pipeline consists of scripts for conducting the following four types of tasks: (i) barcode identification and read trimming; (ii) read alignment, read pairing, barcode association and filtering of spurious reads; (iii) cellular demultiplexing and quality analysis; and (iv) filtration of cellular indices. The rationale behind these scripts and step-by-step instructions are described in detail in the online methods of Ramani

et al [17]. The processing workflow is also outlined in Fig. 3.

The quality of a sci-Hi-C dataset can be gauged in several ways (Fig. 4). First, the overall distribution of chromatin interactions in a sci-Hi-C dataset should resemble that in a conventional bulk Hi-C dataset (e.g., showing strong intrachromosomal interaction enrichment) (Fig. 4A–C), and thereby the aggregated chromatin contact heatmap created from assembled single cells should be highly similar to that generated from bulk Hi-C assays with the same cell population [17]. Second, in a successful sci-Hi-C dataset, at least forty percent of the read pairs are associated with appropriate barcoding combinations (i.e., cellular indices). In a sci-Hi-C library, since the 1st round barcodes are located in the middle of the valid DNA fragments, not every sequence read will cover the 1st round index sequence, even the 250 bp paired-end sequencing is used, which consequently limits the resolution (i.e., genomic coverage per single cell) of sci-Hi-C. Third, if the nuclei remain intact through the experimental steps 4–39 during a sci-Hi-C assay, which can be examined by phase-contrast microscopy, there should be minimal crosstalk between cellular indices (Fig. 4D). And fourth, each 96-well plate in a sci-Hi-C experiment can typically yield hundreds to 2000 single cells with at least 1000 unique long-range chromatin interactions (ranging from 1000 to near 1 million interactions) (Fig. 4D).

4.2. Using sci-Hi-C datasets to characterize 3D genome organization in individual cells

Like other single-cell Hi-C methods [15,16,21–23], sci-Hi-C captures snapshots of chromatin conformations from individual cells at the time of cell fixation. Despite the sparsity in genome coverage, single-cell Hi-C assays enable to interrogate 3D genome organization in individual cells. Compared to other single-cell Hi-C datasets, sciHi-C datasets are relatively more sparse yet rich in cell numbers, allowing to obtain single-cell and bulk measurement at once (the latter generated by summing large number of single cells).

We have successfully applied the improved sci-Hi-C protocol to a diversity of mouse and human tissues and cell lines, such as mouse and human embryonic stem cells, mouse brain and liver cells and human blood cells ([18] and unpublished data). Previously, we have

Sequencing reads of a sci-Hi-C library

- Identifying the outer barcodes in each read
- Trimming off the first 11 bases of each read
- Identifying the inner barcodes in each read
- Trimming off the bridge adaptor sequences in each read

Barcode association and read trimming

Read 1 | **Read 2** | **Barcode association information**

- Mapping read 1 and read 2 independently to the reference genome
- Discarding reads with MAPQ<30
- Identifying the DpnII site in each read
- Pairing read 1 with read 2 and associating with the barcode combination
- Removing PCR duplicates

Read alignment and pairing

Unique chromatin interactions associated with cellular indexing

- Generating single-cell matrices in long format
- Generating numpy matrices for visualization and further analysis
- Removing all cellular indices with fewer than 1,000 unique reads
- Removing all indices where the cis:trans ratio is lower than 1
- Removing cellular indices where less than 95% of reads aligned uniquely to either the mouse) or human genomes
- Other more stringent filtration can also be carried out

Cellular demultiplexing and filtration

Single cell chromatin interaction profiles

Fig. 3. The outline of sci-Hi-C data processing.

A

Total of raw read pairs	Reads pairs with matched 1st round barcodes	Reads pairs with matched 2nd round barcodes	total mapped read pairs with MAPQ >= 30	Distinct read pairs after removal of PCR duplicates
114,896,492	110,304,511	52,625,010	21,132,456	10,236,204

B

Interaction type	Number of interactions of each orientation			
	+/+	-/+	+/-	-/-
>= 20 kb	1,410,817	1,410,979	1,411,133	1,411,703
<= 1 kb	30,753	30,630	86,935	39,839

C

Interaction type	Count	Percent
Human_Intra-short (≤ 1kb)	139463	1.36
Human_Intra-mid (>1kb, < 20kb)	1100858	10.75
Human_Intra-long (≥ 20kb)	4660264	45.53
Human_inter	2711639	26.49
Inter-species	1292	0.01
Mouse_Intra-short (≤ 1kb)	48694	0.48
Mouse_Intra-mid (>1kb, < 20kb)	263156	2.57
Mouse_Intra-long (≥ 20kb)	984368	9.62
Mouse_inter	326470	3.19

D

Fig. 4. Examples of quality analysis of sci-Hi-C datasets. (A) Barcoding association rates of sequence reads of an example sci-Hi-C dataset produced from a 96-well plate library, containing human and mouse cells. Note, we usually only keep alignments with MAPQ > = 30, which is relatively conservative. If a cutoff of MAPQ > = 10 is used, the mapping percentage of reads will be about 5% higher. (B) Directionality breakdown of unique intra-chromosomal chromatin interactions in the sci-Hi-C dataset. Number of interactions in each orientation ($+/-$, inward; $-/+$, outward; $+/+$, both reads in forward orientation; $-/+$, both reads in reverse orientation). Note, dangling ends ($+/-$, inwards) account for a large portion of the short-range ($< = 1$ kb) interactions. (C) Chromatin interaction distribution of the sci-Hi-C library. (D) Example scatterplot showing number of reads mapping uniquely to human or mouse genome for individual barcode combinations, indicating low levels of collision between species.

demonstrated that sci-Hi-C is able to identify cell-to-cell heterogeneity in the conformational properties of mammalian chromosomes and to separate different cell types on the basis of karyotypic and cell-cycle state differences, enabling "*in silico* sorting" of mixed cell populations by cell-cycle stage [17], which remains true for the sci-Hi-C datasets generated using the improved protocol (unpublished data). Moreover, by combining HiCRep [26] with multidimensional scaling (MDS), we were able to implement an analytical tool to embed sci-Hi-C data into a low-dimensional space, which successfully separate cell subtypes from a cell population based on cell-to-cell variations in cell-cycle phase [19]. More recently, We found that topic modeling [27] provides a powerful tool that is able to capture cell type-specific differences in chromatin compartment structures from sci-Hi-C data [18].

5. Conclusions

Sci-Hi-C enables profiling of chromosome conformation in large number of single cells by employing combinatorial cellular indexing.

Acknowledgements

References

[1] E. Lieberman-Aiden, et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science 326 (5950) (2009) 289–293.

[2] V. Ramani, J. Shendure, Z. Duan, Understanding spatial genome organization: methods and insights, Genom. Proteom. Bioinf. 14 (1) (2016) 7–20.

[3] T. Cremer, M. Cremer, Chromosome territories, Cold Spring Harbor Perspect. Biol. 2 (3) (2010) a003889.

[4] J.R. Dixon, et al., Topological domains in mammalian genomes identified by analysis of chromatin interactions, Nature (2012).

[5] E.P. Nora, et al., Spatial partitioning of the regulatory landscape of the X-inactivation centre, Nature (2012).

[6] S.S. Rao, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159 (7) (2014) 1665–1680.

[7] S. Wang, et al., Spatial organization of chromatin domains and compartments in single chromosomes, Science 353 (6299) (2016) 598–602.

[8] B. Bintu, et al., Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells, Science (2018) 362(6413).

[9] A. Krumm, Z. Duan, Understanding the 3D genome: emerging impacts on human disease, Semin. Cell Dev. Biol. (2018) Epub ahead of print.

[10] E.H. Finn, et al., Extensive heterogeneity and intrinsic variation in spatial genome organization, Cell (2019).

[11] J. Borden, L. Manuelidis, Movement of the X chromosome in epilepsy, Science 242 (4886) (1988) 1687–1691.

[12] U. De Boni, The interphase nucleus as a dynamic structure, Int Rev Cytol 150 (1994) 149–171.

[13] L. Manuelidis, A view of interphase chromosomes, Science 250 (4987) (1990) 1533–1540.

[14] T. Cremer, C. Cremer, Chromosome territories, nuclear architecture and gene regulation in mammalian cells, Nat Rev Genet 2 (4) (2001) 292–301.

[15] T.J. Stevens, et al., 3D structures of individual mammalian genomes studied by single-cell Hi-C, Nature 544 (7648) (2017) 59–64.

[16] L. Tan, et al., Three-dimensional genome structures of single diploid human cells, Science 361 (6405) (2018) 924–928.

[17] V. Ramani, et al., Massively multiplex single-cell Hi-C, Nat Methods 14 (3) (2017) 263–266.

[18] H.-J. Kim, et al., Capturing cell type-specific chromatin structural patterns by applying topic modeling to single-cell Hi-C data, bioRxiv (2019).

[19] J. Liu, et al., Unsupervised embedding of single-cell Hi-C data, Bioinformatics 34 (13) (2018) i96–i104.

[20] D.A. Cusanovich, et al., Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing, Science 348 (6237) (2015) 910–914.

[21] T. Nagano, et al., Single-cell Hi-C reveals cell-to-cell variability in chromosome structure, Nature 502 (7469) (2013) 59–64.

[22] T. Nagano, et al., Cell-cycle dynamics of chromosomal organization at single-cell resolution, Nature 547 (7661) (2017) 61–67.

[23] I.M. Flyamer, et al., Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition, Nature 544 (7648) (2017) 110–114.

[24] W. Ma, et al., Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution, Methods 142 (2018) 59–73.

[25] V. Ramani, et al., Mapping 3D genome architecture through in situ DNase Hi-C, Nat. Protoc. 11 (11) (2016) 2104–2121.

[26] T. Yang, et al., HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient, Genome Res. 27 (11) (2017) 1939–1949.

[27] C. Bravo Gonzalez-Blas, et al., cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data, Nat. Methods 16 (5) (2019) 397–400.