Genome Biology

SOFTWARE

Open Access

Harmonizing single-cell 3D genome data with STARK and scNucleome

Wen-Jie Jiang[1†], KangWen Cai[2†], YuanChen Sun[2†], An Liu[2], HanWen Zhu[3], RuiXiang Gao[2], Chunge Zhong[4], Nana Wei[2], Futing Lai[3], Teng Fei[4], Yu-Juan Wang[2], Xiaoqi Zheng[5*], Ming Xu[1,6*] and Hua-Jun Wu[2,3,7*]

[†]Wen-Jie Jiang, KangWen Cai and YuanChen Sun contributed equally to this work.

*Correspondence: xqzheng@shsmu.edu.cn; xuminghi@bjmu.edu.cn; hjwu@pku.edu.cn

[1] Department of Cardiology and Institute of Vascular Medicine, State Key Laboratory of Vascular Homeostasis and Remodeling, NHC Key Laboratory of Cardiovascular Molecular Biology and Regulatory Peptides, Beijing Key Laboratory of Cardiovascular Receptors Research, Peking University Third Hospital, Peking University, Beijing 100191, China
[2] Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Peking University Cancer Hospital & Institute, Beijing 100142, China
[5] Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China
Full list of author information is available at the end of the article

Abstract

Single-cell three-dimensional genome sequencing (sc3DG-seq) reveals genome regulation and heterogeneity in various biological processes, but a universal analysis tool is lacking. Here we present STARK, a versatile toolkit for processing, quality control, and analysis of diverse sc3DG-seq data. Utilizing STARK, we benchmark 15 technologies, quantitatively comparing their strengths and limitations. We also develop EmptyCells to filter empty barcodes and introduce Spatial Structure Capture Efficiency (SSCE) to assess chromatin structure capture quality. Additionally, we establish scNucleome, a uniformly processed repository of sc3DG-seq datasets, to serve as a foundational resource for future 3D genome research.

Keywords: Single cell 3D genome, Single cell Hi-C, STARK, ScNucleome, Benchmark

Background

Chromosome conformation capture technology plays a crucial role in elucidating the three-dimensional (3D) genome organization, revealing the complexities of gene expression, regulation, and cellular functions [1–4]. This technique and its derivatives allow us to study the functionality of chromatin domains, chromosomal interactions, and spatial associations between genes and regulatory elements, which are critical for understanding gene regulation during cell differentiation, tissue development and disease progression [5–13]. Despite the contributions of these techniques, their applications to mixed cell populations often result in an aggregate signal that masks cell-to-cell variability and dynamic processes [14, 15]. In light of these limitations, recent technological breakthroughs in single-cell genome-wide mapping of chromatin contacts have significantly advanced our understanding of chromatin organization, particularly in elucidating the dynamics of the cell cycle and revealing the complex organization of tissues.

Currently, over ten methods of sc3DG-seq have been developed, including scHi-C [16], scHi-C$^+$ [17], sciHi-C [18], Dip-C [19], sn-m3C [20], scMethyl [21], HiRES [22], scSPRITE [23], snHi-C [14], snHi-C$^+$, scNanoHi-C [24], LiMCA [25], GAGE-seq [26],

Droplet Hi-C and Paired Hi-C [27]. These methods can be broadly categorized into three types: First, methods that apply bulk Hi-C techniques directly to individual cells post single-cell sorting; second, the use of single-cell high-throughput profiling methods that leverage cellular barcodes; and third, the integration of additional omics, including DNA methylation and gene expression, through simultaneous sequencing within the same single cell. The first type of technology adapts the standard Hi-C protocol [8, 28, 29] by performing ligation after nuclear lysis and dilution of chromatin complexes. This approach processes each cell individually, leading to a more consistent quality and total contacts across single cells. The second type of technology features high-throughput capacity by introducing cellular barcodes at various experimental steps or utilizing microfluidic systems, enabling parallel sequencing of chromatin contacts across a larger cell population. Although variations in total contacts across individual cells may be more pronounced, it significantly expands the scope of analysis for complex tissues. In this category, scSPRITE, an advancement of the SPRITE method [30], and scNanoHi-C, the first technology to use long-read sequencing in 3D genome analysis [31], are particularly noteworthy. Both technologies are adept at delineating higher-order chromatin interactions, opening new avenues for studying complex chromosomal structures within single cells. Microfluidic-based approaches such as Droplet Hi-C and Paired Hi-C employ commercial microfluidic systems to achieve scalable processing. The third type of technology, including HiRES, sn-m3C, and scMethyl, integrates sequencing from additional omics with mapping of 3D chromosome structures. This multi-omics approach demonstrates the interplay between the spatial organization of chromosomes and other molecular context, leading to a deeper insight into complex mechanisms of gene regulation within individual cells [20, 24, 32].

The diversity of sc3DG-seq has led to a challenge in developing versatile software that can accommodate the varied data formats and analysis requirements from the full spectrum of these methods. Current tools focus on specific downstream analyses, offering functionalities like quality control [33], imputation [34–37], clustering [35, 38–40], denoising [41, 42], and the identification of chromatin loops [43, 44]. With the rapid advancement of sc3DG-seq technologies, the throughput has increased dramatically, creating a demand for software capable of handling and analyzing large volumes of sequencing data, akin to the progress seen in single-cell RNA-seq. However, an efficient workflow for processing and filtering high-quality sc3DG-seq data for downstream analysis is still lacking. To fill this gap, we propose STARK—a cutting-edge toolkit for Structural Topology Analysis and a Rich Knowledge base. STARK offers a systematic workflow for preprocessing diverse sc3DG-seq data with unified, standardized procedures. It provides robust, versatile quality control measures to ensure the selection of high-quality cells and generates unified single-cell chromatin contact maps alongside the necessary downstream analysis results. STARK represents a significant step forward in addressing the analytical challenges posed by the growing complexity and volume of sc3DG-seq data.

Furthermore, different sc3DG-seq technologies each exhibit unique features and capabilities in mapping the 3D architecture of chromosomes, which can significantly impact the resolution and depth of the data obtained. For instance, variations in sequencing depth and resolution across technologies can lead to different interpretations of genome-wide regulatory interactions and insights into the dynamic changes in

chromatin topology under various biological processes [45–47]. Another critical determinant of technological superiority is the ability to prereserve chromatin contacts within individual cells [48]. These technological nuances suggest that they may be optimized for specific experimental goals, such as high-resolution mapping, multi-omics integration, or the investigation of complex chromatin interactions. Therefore, it is particularly important to evaluate the characteristics of each technology.

To thoroughly assess the performance and applicability of these technologies, we compiled a comprehensive dataset of sc3DG-seq data and employed STARK for unified preprocessing, quality control and analysis. By minimizing systematic biases, we provided an objective benchmarking of the various methods. We developed a comprehensive set of evaluation indices to quantitatively compare the strengths and limitations of each technology, providing researchers with a clear framework to understand the technological differences, streamline data processing, and inform future experimental design. Additionally, we established scNucleome, a comprehensive publicly accessible repository of uniformly processed sc3DG-seq datasets, encompassing diverse cell lines and tissues across the spectrum of available technologies, serving as a valuable resource for the research community.

Results

STARK framework

sc3DG-seq has emerged as an essential tool for understanding the variability in 3D chromatin structure among individual cells. The process encompasses cell sorting, chromosome cross-linking, digestion, ligation, and sequencing, culminating in a map of interactions across the genome. Despite the development of over ten distinct sc3DG-seq techniques, each with unique processing requirements, a unified computational framework has been lacking. This gap poses a significant challenge to the research community, hindering the effective utilization of these extensive and valuable data.

To address this challenge, we propose STARK, a unified framework that conducts preprocessing, quality control and downstream analysis for all current sc3DG-seq data types (Fig. 1). STARK encompasses three modules: Preprocess (Fig. 1a), Cell Quality Control (Fig. 1b) and Downstream Analysis (Fig. 1c).

In Preprocess module (Fig. 1a), we consolidate raw data from all techniques into a unified framework. Initially, we extract high-quality sequencing reads and perform sequence alignment using an aligner suited to the sequencing platform. Subsequently, we parse pair information for chromatin interactions, process these based on enzyme cleavage sites, and filter to obtain high-quality chromatin interactions. For techniques employing barcodes, we include a dedicated step for demultiplexing and cell allocation. Finally, we convert processed data into a binary format (.cool) and apply Hi-C correction to each cell.

In Cell Quality Control (QC) module (Fig. 1b), we introduce the EmptyCells algorithm, a newly developed approach to filtering and comprehensive quality control of sc3DG-seq data. By first eliminating cells with insufficient interactions and then employing Monte Carlo simulations to identify high-quality cells, EmptyCells ensures that only the most reliable data is used in downstream analyses. To thoroughly evaluate each single cell, we offer a suite of metrics that provide a comprehensive insight into chromatin

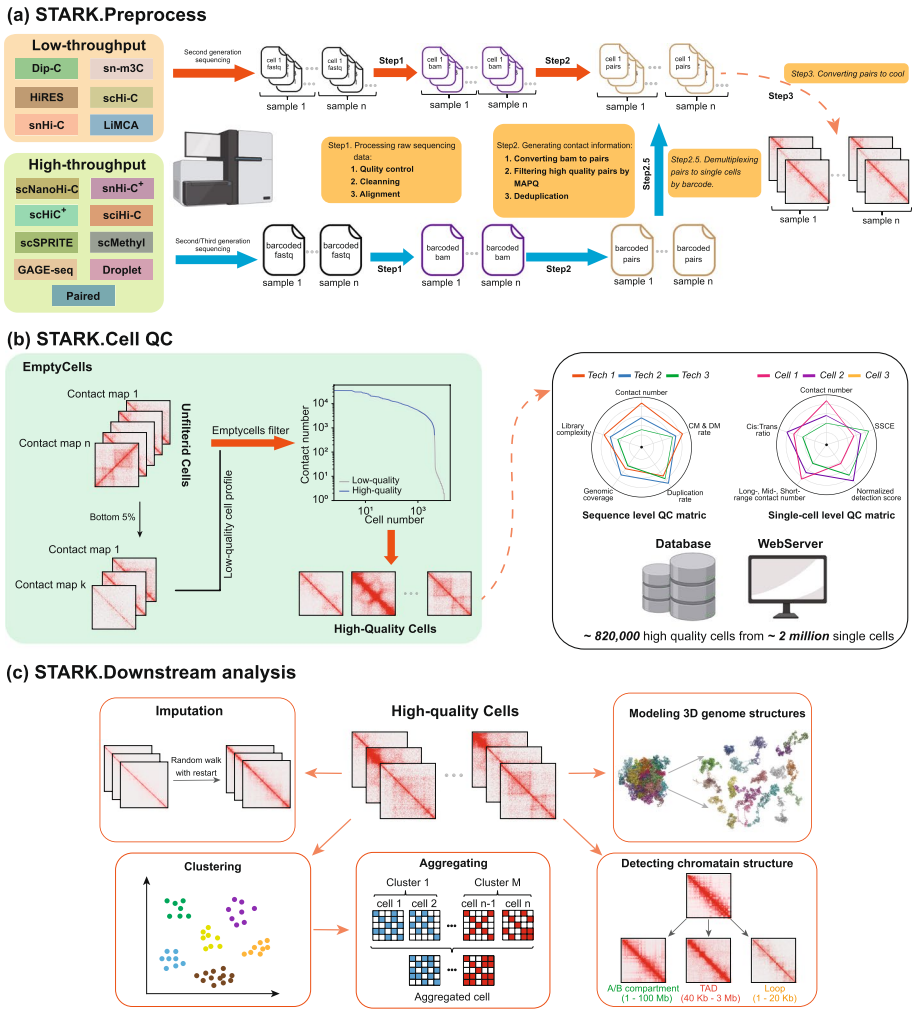

Fig. 1 An overview of the STARK workflow. STARK includes preprocess, cell quality control and downstream analysis modules, each designed to work seamlessly with the others to provide a comprehensive pipeline for the analysis of sc3DG-seq data. **a** In STARK.Preprocess module, samples undergo a series of steps including sequencing quality control, cleaning and alignment based on the sequencing platforms used. High-throughput sequencing samples are demultiplexed. The process results in the generation of single-cell contact matrices. **b** STARK.Cell QC module employs the EmptyCells algorithm to filter cells based on contact count, ensuring that only high-quality cells are retained. These cells are then evaluated using a suite of metrics, including contact number, GiniQC, and long-/mid-/short-range contact rates. **c** STARK.Downstream analysis module encompasses a range of analytical processes, including imputing contact matrix, clustering cells, aggregating single cell contact matrices, identifying A/B compartments, TADs, and DNA loops, as well as modeling 3D genome structure

interactions. These include the total number of contacts, GiniQC, short-range to long-range contact rates, and other relevant measures. Furthermore, we propose the Spatial Structure Capture Efficiency (SSCE), a new QC metric that assesses a single cell's ability to capture spatial chromatin structures. The SSCE integrates multiple topological features, enhancing the selection of cells that exhibit informative structural patterns. The SSCE is designed to ensure that cells with fewer overall contacts but notable structural patterns are not disregarded, therefore, complements existing QC metrics.

Downstream Analysis module (Fig. 1c) offers a suite of analytical tools that impute, cluster, and aggregate chromatin interactions in single cells, as well as identify key

genomic topology such as A/B compartments, topologically associating domains (TADs) and chromatin loops, and reconstruct 3D genome configurations from 2D interaction maps.

In summary, STARK is a powerful software suite capable of processing extensive datasets from various sc3DG-seq techniques. The framework implements computational optimizations including parallel processing and Monte Carlo simulation acceleration, demonstrating excellent scalability for large-scale applications (Additional file 1: Fig. S1a, b, c, d and Additional file 2: Table S1). By processing current and future sc3DG-seq data, STARK empowers the research community to conduct in-depth analyses and unlock new discoveries in the study of 3D genome organization (Table 1).

Benchmark on read-level sequencing efficiency

To thoroughly evaluate the current sc3DG-seq technologies, we analyzed each technique data uniformly preprocessed using STARK (see Methods). We began by assessing the sequencing depth and throughput of the various sc3DG-seq technologies (Fig. 2a). Low-throughput methods (w/o cell barcode) like snHi-C demonstrate greater sequencing depth, whereas high-throughput methods (with cell barcode), such as snHi-C^{+}, sequence a higher number of cells per experiment. In particular, the scSPRITE technology achieves the highest average number of contacts per cell, with its approach to chromatin contact capture being a primary factor. scSPRITE utilizes sonication to fragment the cross-linked DNA complexes, releasing DNA spatial clusters, which are then labeled with spatial barcodes [23, 54]. This method, unlike ligation-based technologies, preserves a broader range of spatially proximate contacts, leading to a unique distribution of captured interactions. The ability of snHi-C to obtain the second highest average number of contacts per cell is largely due to its use of whole-genome amplification, which increases DNA quantity without substantial sequence bias [55].

To evaluate the efficiency of various sc3DG-seq technologies in capturing chromatin contacts and the impact of ligation methods on sequencing outcomes, we calculated the rate of concordantly-mapped (CM) and discordantly-mapped (DM) read pairs (Fig. 2b). The majority of technologies demonstrate effective capture capabilities, as evidenced by decent proportion of CM read pairs. However, some methods, such as Dip-C and LiMCA, produce a higher proportion of DM read pairs, possibly due to the use of the META whole-genome amplification approach. The META method is known to reduce chimeric reads [56], and enhance the quality of CM read pairs (Additional file 1: Fig. S2a), but it may also lead to an increased generation of DM read pairs (Additional file 1: Fig. S2b), suggesting a trade-off in the amplification process. Technologies that simultaneously capture DNA methylation and 3D genome data in the same sets of single cells also show a higher proportion of DM read pairs, potentially due to the bisulfite treatment step required for methylation profiling, which can lead to a decreased alignment rate (Additional file 1: Fig. S2c). In contrast, scSPRITE and scNanoHi-C, which are designed to generate higher-order chromatin contacts, differing from the paired contacts between two genomic regions captured by other methods, are primarily evaluated on their ability to capture these complex interactions effectively, and they also maintain satisfactory alignment rates.

Table 1 Summary of sc3DG-seq datasets

Name	Source paper	Description	Accession	Species	Biotin fill-in	Ligation	Pulldown	Enzyme	Library	Product	Pre EmptyCells	Post EmptyCells
scSPRITE	Arrastia MV [23]	Multiple split-pool barcoding rounds;spatial barcode	GSE154353	Mouse	No	Spatial barcode	No	HpyCH4V	PCR	Higher-order contacts	148,871	7,456
HiRES	Liu Z [22]	In situ reverse transcription; optimized 3C	GSE223917	Mouse	No	T4 DNA	No	MboI	PCR	Hi-C; mRNA	8,129	8,129
Dip-C	Tan L [45]	META amplification; omit biotin pulldown	GSE117874	Human	No	T4 DNA	No	MboI	META	Hi-C	61	61
sciHi-C	Bonora G [46]	Two-round barcoding without cell separation; combinatorial indexing	GSE185608	Mouse	No	Bridge adaptor ligation	Yes	DpnII	PCR	Hi-C	142,736	19,754
	Ramani V [18]		GSE84920	Human				DpnII			8,489	1,054
scHi-C	Pang QY [49]	Modified dilution Hi-C	GSE201917	Human	Yes	T4 DNA	Yes	MboI	PCR	Hi-C	24	24
			GSE201918	Human				MboI			9	9
	Lambuta RA [50]		GSE222390	Human				MboI			73	73
	Nagano T [16]		GSE48262	Mouse				BglII,DpnII			15	15
scHi-C$^+$	Rappoport N [51]	Flow cytometry sorting; PCR with barcode	GSE148793	Mouse	Yes	T4 DNA	Yes	MboI	PCR with Tn5	Hi-C	3,584	3,562
	Nagano T [17]		GSE94489	Mouse				MboI			1,295,641	28,568
sn-m3C	Luo C [52]]	Bisulfite conversion after cell sorting	GSE124391	Mouse	No	T4 DNA	No	MboI	PCR with random primers	Hi-C; DNA methylation	475	475
				Human							114	144
			GSE130711	Human							4,238	4,238
snNanoHi-C	Li W [24]	Nanopore sequencing; higher-order interactions	GSE217189	Human	No	T4 DNA	No	MboI	PCR with Tn5	Higher-order contacts	1,508	1,302
				Mouse							576	5,75

Table 1 (continued)

Name	Source paper	Description	Accession	Species	Biotin fill-in	Ligation	Pulldown	Enzyme	Library	Product	Pre EmptyCells	Post EmptyCells
snHi-C^{+}	-	-	4DNESF-829JOW	Human	-	-	-	MboI	-	Hi-C	40,607	4,688
			4DNES-JQ4RXY5	human				MboI			14,747	692
snHi-C	Flyamer IM [14]	Omits biotin labeling and enrichment	GSE80006	Human; Mouse	No	T4 DNA	No	DpnII	WGA	Hi-C	269	269
	Dequeker B [51]		GSE196300	Mouse				DpnII			108	108
	Chatzidaki EE [45]		GSE171159	Mouse				DpnII			125	125
			GSE130585	Mouse				DpnII			66	66
	Silva MC [47]		GSE132264	Mouse				DpnII			130	130
	Gassler J [53]		GSE100569	Mouse				DpnII			210	210
scMethyl	Li G [21]	bisulfite conversion after cell sorting; PCR with barcode	GSE119171	Mouse	No	T4 DNA	No	MboI	PCR with random primers	Hi-C; DNA methylation	217	150
LiMCA	Honggui Wu [52]	nucleus–cytoplasm physical separation	GSE239969	Mouse	No	T4 DNA	No	NlaIII	META	Hi-C; mRNA	411	411
			GSE240114	Human						Hi-C; mRNA	389	389
GAGE-seq	Tianming Zhou [26]	combinatorial indexing;Biotin isolation	GSE238001	Human; Mouse	No	T4 DNA	Yes	Combined	PCR	Hi-C; mRNA	119,296	39,789
Droplet Hi-C	Lei Chang [27]	commercial microfluidic systems	GSE253407	Human; Mouse	No	T4 DNA	No	Combined	PCR with Tn5	Hi-C	32,959	25,542
Paired Hi-C	Lei Chang [27]	milder SDS treatment;10 × Genomics Chromium Single Cell Multiome kit	GSE253407	Human; Mouse	No	T4 DNA	No	Combined	PCR with Tn5	Hi-C; mRNA	50,874	45,424
												Sum: 192,857

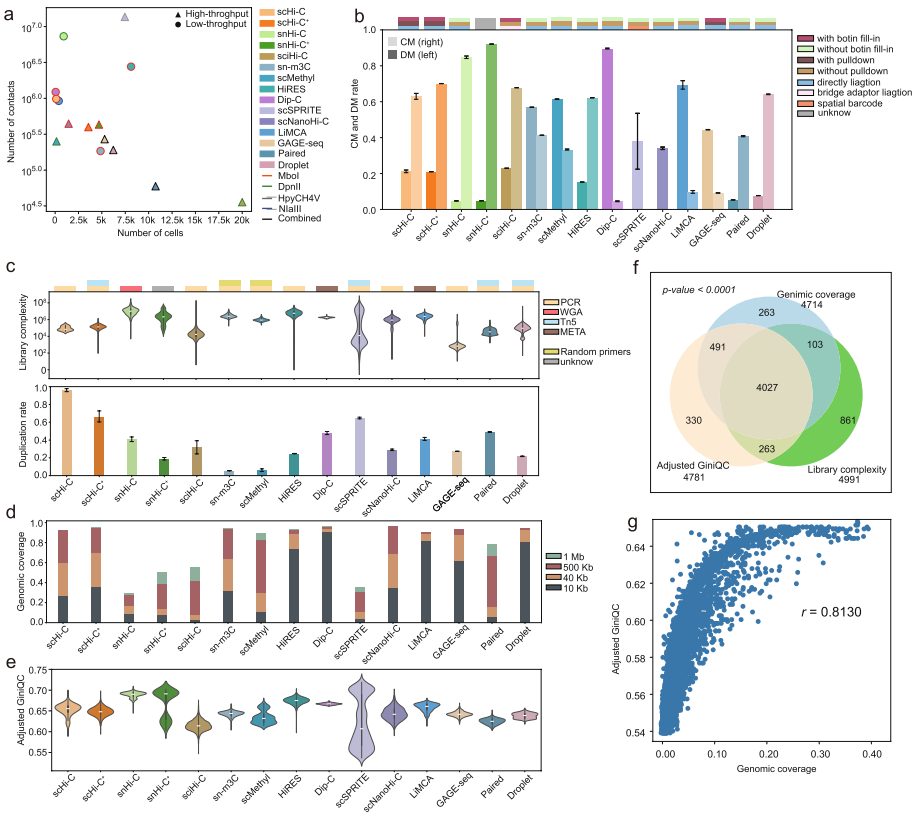

Fig. 2 Benchmark of read-level sequencing efficiency across sc3DG-seq technologies. **a** Scatter plot showing the relationship between the number of cells and the average number of contacts per cell. Fill colors of points represent technologies. Border colors of points represent restriction enzymes used. Shapes of points distinguish low- and high-throughput methods. **b** Bar plot displays the rate of concordantly-mapped (CM) and discordantly-mapped (DM) read pairs. Left bars indicate the CM read pairs and right bars indicate DM read pairs. Colors represent different technologies. The bar graph above each pair of bars represents the ligation method, whether biotin is used, and if pulldown is performed. For scSPRITE and scNanoHi-C, only the rate of CM read pairs is calculable. **c** Violin and bar plots represent the library complexity and duplication rate across cells for various technologies, with a corresponding bar above indicating the library preparation method applied. **d** Stacked bar plot demonstrates the average genome coverage across cells at resolutions of 1 Mb, 500 Kb, 40 Kb, and 10 Kb for different technologies. **e** Violin plot exhibits the distribution of GiniQC values across cells for different technologies. **f** Venn diagram illustrates the overlap among cells with low genomic coverage at 1 Mb resolution, low GiniQC and low library complexity in the scSPRITE dataset. **g** Scatter plot depicts the correlation between GiniQC and genome coverage at 1 Mb resolution across the overlapped cells in **f**

We next evaluated the library complexity and sequence duplication rate across various sc3DG-seq technologies (Fig. 2c), revealing significant variability. Low-throughput techniques, which afford more meticulous processing of individual cells, generally result in higher library complexity and lower duplication rate (Additional file 1: Fig. S2d-e). Furthermore, methods using whole-genome amplification (WGA) typically achieve higher library complexity but also come with the trade-off of higher duplication rate (Additional file 1: Fig. S2f, g).

We then employed the genomic coverage across different resolutions to evaluate the sequencing efficiency of different sc3DG-seq technologies. More than half technologies demonstrate robust performance, achieving approximately full-genome coverage

at a resolution of 1 Mb (Fig. 2d), whereas snHi-C, snHi-C$^+$, sciHi-C, and scSPRITE provide genomic coverage rate ranging from 0.3–0.6.

To assess the sequencing efficiency in an alternative way, we employed the GiniQC metric [57], which measures the unevenness of read distribution in the contact matrix as an indicator of noise levels. Across the fifteen technologies evaluated, average GiniQC values are comparable, suggesting a consistent level of noise. Notably, scSPRITE exhibits a bimodal distribution in genomic coverage (Additional file 1: Fig. S3a), GiniQC (Fig. 2e), as well as library complexity (Fig. 2c). When analyzing cells with lower genomic coverage, GiniQC, and library complexity, a significant overlap was observed (Fig. 2f), indicating a strong correlation between these metrics. The Pearson correlation coefficient between genomic coverage and GiniQC for these cell populations is 0.813 (Fig. 2g), highlighting the interdependence of genomic coverage, GiniQC and library complexity in evaluating cell quality. Similar bimodal patterns and results were obtained in scMethyl and snHi-C (Additional file 1: Fig. S3b, c).

Benchmark on the efficiency of capturing contacts at various genomic ranges

Chromatin interactions across various genomic distance scales have distinct biological implications (Fig. 3a). Short-range interactions, typically less than 20 Kb apart, often denote spatial contacts between local regulatory elements, such as proximal enhancers and promoters, thereby playing a key role in gene transcriptional regulation. Mid-range interactions, ranging from 20 Kb to 2 Mb, indicate spatial contacts between distal regulatory elements and promoters or the organization of functional gene clusters within topologically associating domains (TADs). These interactions represent the functional blocks of gene regulation and insulation, contributing to the precise control of gene transcription [58]. Long-range chromatin interactions, extending over 2 Mb, are essential for the compartment-level folding of chromosomes, influencing transcriptional activity [15, 28, 59]. While less common, inter-chromosomal interactions reflect the spatial organization of chromosomes in the nucleus, sometimes associated with chromosomal translocations and cytogenetic abnormalities [60]. Therefore, a comprehensive understanding of the chromatin interaction landscape across all genomic scales is crucial for elucidating the complexities of gene regulation and delineating the functional organization of the cell nucleus.

To assess the distance distribution characteristics of chromatin interactions captured by various sc3DG-seq methods, we initially computed the ratio of cis to trans interactions (cis:trans ratio) (Fig. 3b). Although all methods detect a higher number of cis interactions, which are interactions between genomic regions on the same chromosome, compared to trans interactions, which occur between different chromosomes, notable variations are observed. snHi-C, snHi-C$^+$, sn-m3C, scMethyl, HiRES, Paired and Droplet Hi-C yield higher cis:trans ratios than others. Interestingly, the three methods capable of multi-omics sequencing—integrating additional genomic information alongside 3D genome data—consistently demonstrate higher cis:trans ratios. This may be attributed to the more intricate experimental protocols potentially leading to a reduced detection of inter-chromosomal contacts. Additionally, the snHi-C method, which utilizes whole-genome amplification (WGA), also yields favorable cis:trans ratios. It is important

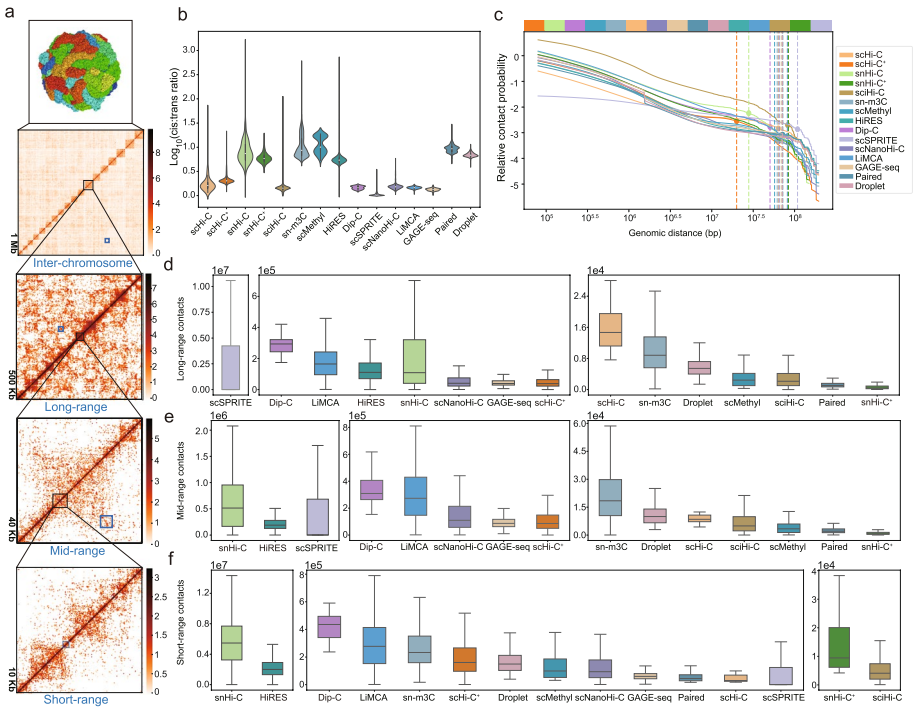

Fig. 3 Benchmark of contact capture efficiency across sc3DG-seq technologies. **a** Schematic illustration of contacts at different distance scales and resolutions. From top to bottom: the 3D structure of a cell as obtained by the Dip-C method, contact maps at 1 Mb (with a blue box indicating inter-chromosome contacts), 500 Kb (indicating long-range contacts), 40 Kb (mid-range contacts), and 10 Kb (short-range contacts) resolutions. **b** Violin plot displays the cis:trans ratios for different technologies. **c** Line plot illustrates the relative contact probability against genomic distance for the most deeply sequenced 10 cells across technologies. Points crossing the dashed lines indicates inflection points, with the color bar above the plot showing the order of these points from low to high. **d-f** Box plots depict long-range, mid-range, and short-range contacts, respectively, across technologies. Each is ordered by descending median contact number

to highlight that the average cis:trans ratio for high-throughput methods is lower than those for low-throughput methods (Additional file 1: Fig. S4a).

Among the evaluated methods, scSPRITE exhibits the lowest cis:trans ratio, which is a logical outcome given its use of spatial barcodes. This approach results in a more uniform capture of chromatin contacts across diverse genomic distances. To further investigate this point, we integrated the top 10 cells with the highest number of contacts from each method and studied the relationship between chromatin interaction frequency and genomic distance (Fig. 3c). Consistent with the well-established distance decay law of chromatin interactions in Hi-C data [8], all methods show a declining trend in interaction frequency with increasing genomic distance. Notably, scSPRITE displays the most gradual decline and a prolonged plateau phase, indicating its enhanced ability to capture interactions across a broader range of genomic scales. In contrast, sciHi-C, despite starting with the highest interaction frequency, exhibits the fastest decline, suggesting a greater propensity to capture local chromatin structures at shorter genomic distances. snHi-C, while achieving a favorable cis:trans ratio, shows an early inflection point (see Methods) in the interaction frequency curve. Coupled with its capacity to detect a large number of contacts, this suggests that snHi-C excels at capturing interactions at shorter

genomic distances. The remaining technologies demonstrate relatively uniform performance in capturing chromatin interactions as genomic distance increased.

Finally, we quantitatively compared the number of chromatin interactions captured by different methods across three genomic distance ranges: short-range (< 20 Kb), mid-range (20 Kb—2 Mb), and long-range (> 2 Mb) (Fig. 3d-f). Consistent with our aforementioned findings, scSPRITE outperforms other methods in capturing long-range chromatin contacts. Methods employing whole-genome amplification (WGA)—snHi-C, Dip-C and LiMCA—demonstrate a favorable depth in capturing chromatin contacts across all distance scales, which underscores the substantial benefit of WGA approach. HiRES, which captures both Hi-C and transcriptome, also shows a remarkable performance in capturing full range of chromatin contacts. Further analysis of the correlations between the three distance ranges revealed that most methods exhibit higher correlations in the number of captured contacts between adjacent distance ranges (Additional file 1: Fig. S4b).

Benchmark on the efficiency of capturing topological structures

Chromatin structures play a crucial role in regulating gene expression and orchestrate the functional organization of the cell nucleus. sc3DG-seq serves as a powerful tool for investigating the heterogeneity of chromatin structures among individual cells, with the resolution of detected structures varying [15]. For instance, at a resolution of 1 Mb, chromatin territories can be identified, A/B compartments are detectable at 500 Kb resolution, and topologically associating domains (TADs) become discernable at 40–50 Kb (Fig. 4a).

To evaluate the ability of different methods to detect chromatin structures, we calculated normalized Detection Scores (nDS) [23] for single cells derived from each method. To minimize potential biases and enhance computational efficiency, we have revised the original calculation approach (see Methods). Using this improved methodology, we assessed the capability of the 15 sc3DG-seq methods to reconstruct 3D chromatin structures at various resolutions, focusing on chromatin territories, A/B compartments, and TADs. The results indicate that most technologies can detect these structures to some extent (Fig. 4a-c). In the detection of chromatin territories, low-throughput methods generally yield higher nDS compared to high-throughput methods (Additional file 1: Fig. S5a). This may be due to the washing steps involved in barcoding of single cells, which can lead to the loss of detection efficiency of chromatin territories. This aligns with our cis:trans ratio analysis findings. Similarly, for A/B compartment detection, low-throughput methods outperform high-throughput methods (Additional file 1: Fig. S5b). In discerning TAD structures, high-throughput methods generally show lower nDS compared to low-throughput methods (Additional file 1: Fig. S5c). Notably, among high-throughput technologies, scSPRITE and scNanoHi-C, which are capable of capturing higher-order chromatin contacts, demonstrate superior TAD detection performance compared to other high-throughput methods (Additional file 1: Fig. S5d). We hypothesize that the cell barcoding and processing steps in most high-throughput approaches may lead to some loss in TAD-level structural information. Technologies specifically designed to capture higher-order

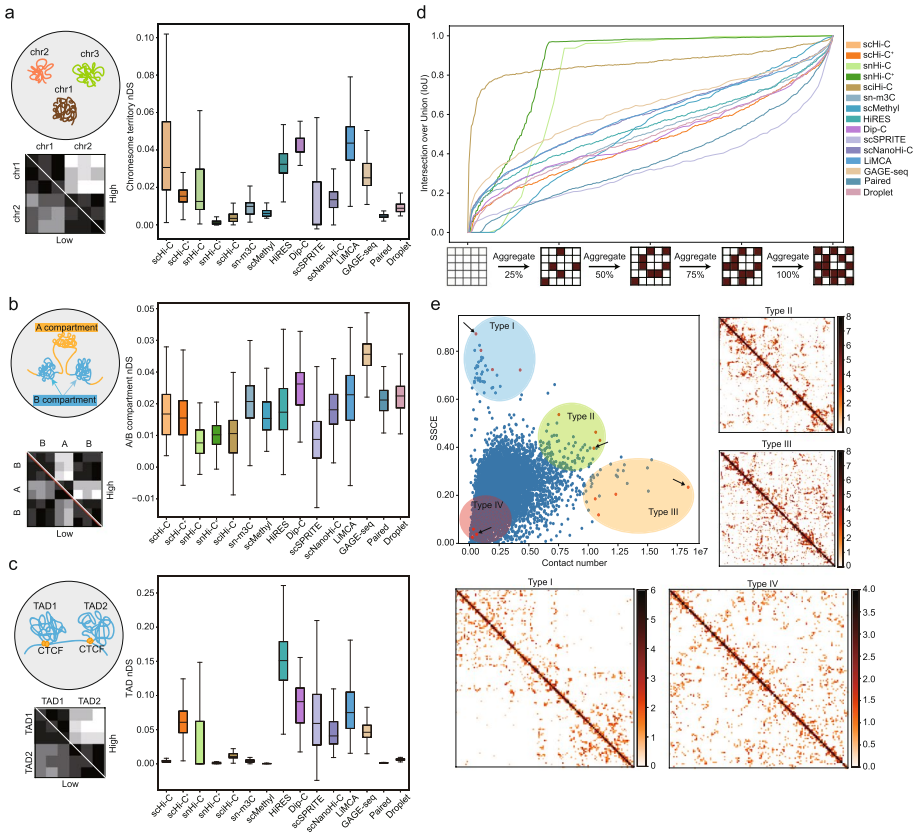

Fig. 4 Benchmark of structure detection ability across sc3DG-seq technologies. **a-c** Normalized Detection Score (nDS) of chromosome territory, A/B compartment, and TAD for the top 300 deeply sequenced cells across various technologies. An inset on the top left provides a schematic illustration of these genomic structures, and the diagram on the bottom left shows the pattern of contact maps with high and low nDSs. **d** Line plot displays the cumulative similarity of TAD boundaries between those obtained from aggregated single cells and total cells for different technologies. This analysis is performed on the top 300 deeply sequenced cells in each technology. The y-axis represents the Intersection over Union (IoU) as a measure of similarity. The x-axis indicates the percentage of cells integrated, with a schematic illustration of the integration process shown. **e** Scatter plot demonstrates the relationship between the SSCE and contact number using HiRES data as an example. Cells are categorized as type I-IV, color-coded in blue, green, orange, and red respectively. Each dot represents an individual cell, with red dots providing visualized examples. Heatmaps correspond to the cells indicated by the black arrows in the scatter plot, showcasing the contact map for the region Chr10: 2 Mb-133 Mb. Additional examples are provided in Additional file 1: Fig. S6

interactions, such as scSPRITE and scNanoHi-C, appear to partially compensate for this limitation through their enhanced ability to detect complex chromatin organization patterns that contribute to TAD boundary definition.

When examining the performance of individual technologies, scSPRITE, Dip-C, scNanoHi-C, scHi-C^{+}, snHi-C, LiMCA and GAGE-seq consistently achieve higher nDS across various genomic resolutions, with particularly notable advantages in detecting TADs. Their superior performance is likely attributed to their proficiency in capturing mid- and long-range chromatin interactions, as previously discussed in our analysis. Interestingly, multi-omics methods that simultaneously sequencing DNA methylation, which are sn-m3C and scMethyl, demonstrate elevated nDS at the A/B compartment and chromatin territory levels, but lower nDS at the TAD levels. This

observation suggests that the incorporation of methylation information may affect the detection of compact chromatin structures.

Next, we examined the ability to reconstruct bulk-like TAD boundaries by integrating an increasing number of single-cell data for different sc3DG-seq technologies (Fig. 4d). To achieve bulk-like TAD boundaries, we consolidated the top 300 most abundant single-cell data from each technology. Utilizing the Accumulative Single-cell Quality Analysis method (see Methods), we tracked the trend in similarity between these bulk-like TAD boundaries and those reconstructed from integrating an increasing number of cells. We employed the intersection over union (IoU) as a metric for consistency evaluation. Overall, the consistency scores of all technologies rise with an increasing number of cells integrated, yet there are notable differences in the rate of score increase. snHi-C, snHi-C$^+$, and sciHi-C exhibited rapid increase in consistency score. Particularly, sciHi-C reached saturation early on, suggesting its ability to approximate bulk-like characteristics with a small subset of cells. In contrast, other technologies showed a slower progression toward the fitting trend, indicating that they require a larger number of cells to reliably reconstruct a bulk-like TAD structures. These findings imply that TAD structures may be inadequately captured and can be considerably variable among individual cells, highlighting the need for further optimization of experimental protocols to enhance data stability.

Similarly, we applied accumulative analysis to chromatin loop detection using mainstream loop detection algorithms (HiCCUPS [28] and Mustache [61]). Results showed that most technologies require aggregation of at least 75 cells before detecting reliable loops, with technologies capable of capturing higher sequencing depth per cell (snHi-C$^+$, scNano-Hi-C, LiMCA, and HiRES) showing better loop detection capability compared to other high-throughput methods (Additional file 1: Fig. S5e). This analysis demonstrates that loop detection is feasible at the meta-cell level after aggregation, but not yet at the single-cell level, underscoring the current technical limitations posed by low signal-to-noise ratios in existing single-cell methods.

nDS can capture structure-specific information but require analyses at various resolutions for a comprehensive assessment. To address this, we developed the Spatial Structure Capture Efficiency (SSCE), a new QC metric that quantitatively evaluates a single cell's ability to capture spatial chromatin structures (Fig. 4e). The SSCE incorporates multiple topological information, including chromatin territories, A/B compartments and TADs, and evaluates their contribution to the total contacts within each cell. Based on the SSCE and total contacts, we categorized cells into four distinct types I—IV. Notably, Type I cells, which have a relatively high SSCE but low total contacts, are observed to exhibit more pronounced structures despite similar sequencing depth as Type IV cells (Fig. 4e and Additional file 1: Fig. S6a, b). Conversely, Type II and Type III cells, while having a relatively high sequencing depth, show lower SSCE values, and their contact maps indicate a higher, albeit structural capturing capability, noise. This holistic approach ensures that cells with fewer overall contacts but significant structural patterns are not overlooked. Traditional quality control methods that rely solely on total contacts may inadvertently exclude cells with lower sequencing depths that still capture important structures. By incorporating SSCE, we can enhance the retention of cells that capture critical structures. Evaluation demonstrated that SSCE-based quality control

improves downstream clustering performance across multiple clustering metrics (Additional file 1: Fig. S7a, b, c and Additional file 1: Fig. S8a, b, c). Therefore, SSCE serves as a valuable complement to existing quality control measures for sc3DG-seq data, aiding in the downstream selection of single cells for further analysis.

scNucleome: a harmonized sc3DG-seq database

Processing large-scale sc3DG-seq data from scratch under a unified framework remains a major challenge due to the high costs of computational resources. To fill this gap, we employed STARK to process approximately 2 million single cells from 28 sc3DG-seq datasets, encompassing over 40 sample types. This extensive effort has resulted in scNucleome, currently the most comprehensive single-cell 3D genome database (Fig. 5a).

scNucleome facilitates data search by technology, sample type, and data source, enabling precise and efficient data retrieval. For each dataset, we provide multi-dimensional data analysis results, including low-dimensional representation, cell clustering and annotation, contact map visualization by HiGlass [62], and simulated 3D organization model by Mol* Viewer [63], allowing researchers to explore datasets from various perspectives (Fig. 5b). Furthermore, scNucleome offers a wide array of quality control metrics at both the dataset and single-cell levels, including library complexity, CM/DM read pair rates, total contacts, GiniQC, nDS, SSCE, and more. These metrics are designed to aid in further experimental design, benchmarking and customizable cell filtration (Fig. 5c). We anticipate that this comprehensive platform will markedly accelerate research progress, providing a robust foundation for advancements in the field of 3D genome studies.

Discussion

In this study, we introduce STARK, a comprehensive computational framework designed to provide a unified and standardized approach for processing and analyzing a wide array of sc3DG-seq data. STARK encompasses three core modules—Preprocess, Cell QC, and Downstream Analysis – enabling unified data handling, extensive quality control, and an in-depth exploration of chromatin structures within individual cells.

A key strength of STARK lies in its adaptability to the diverse data formats and analytical requirements across the full spectrum of current sc3DG-seq technologies. This includes capabilities for processing data from both low and high-throughput technologies, as well as accommodating second and third-generation sequencing-based methods and single and multi-omics approaches. In addition, STARK introduces EmptyCells, a new quality control method tailored for high-throughput technologies, and proposes the SSCE, a new metric for evaluating the capture efficiency of chromatin structure in single cells. By systematically processing approximately 2 million single cells across 15 different sc3DG-seq techniques, STARK offers a comprehensive benchmarking of the strengths and limitations inherent to each method. This comprehensive assessment empowers researchers to make informed choices regarding the most suitable technology for their experimental objectives, whether they aim for high-through in cell number, high-resolution contact mapping, or the investigation of complex chromatin interactions.

Our benchmarking analysis reveals that different technologies exhibit complementary strengths suited to various research applications. For instance, technologies like scSPRITE and scNanoHi-C appear particularly well-suited for high-resolution structural

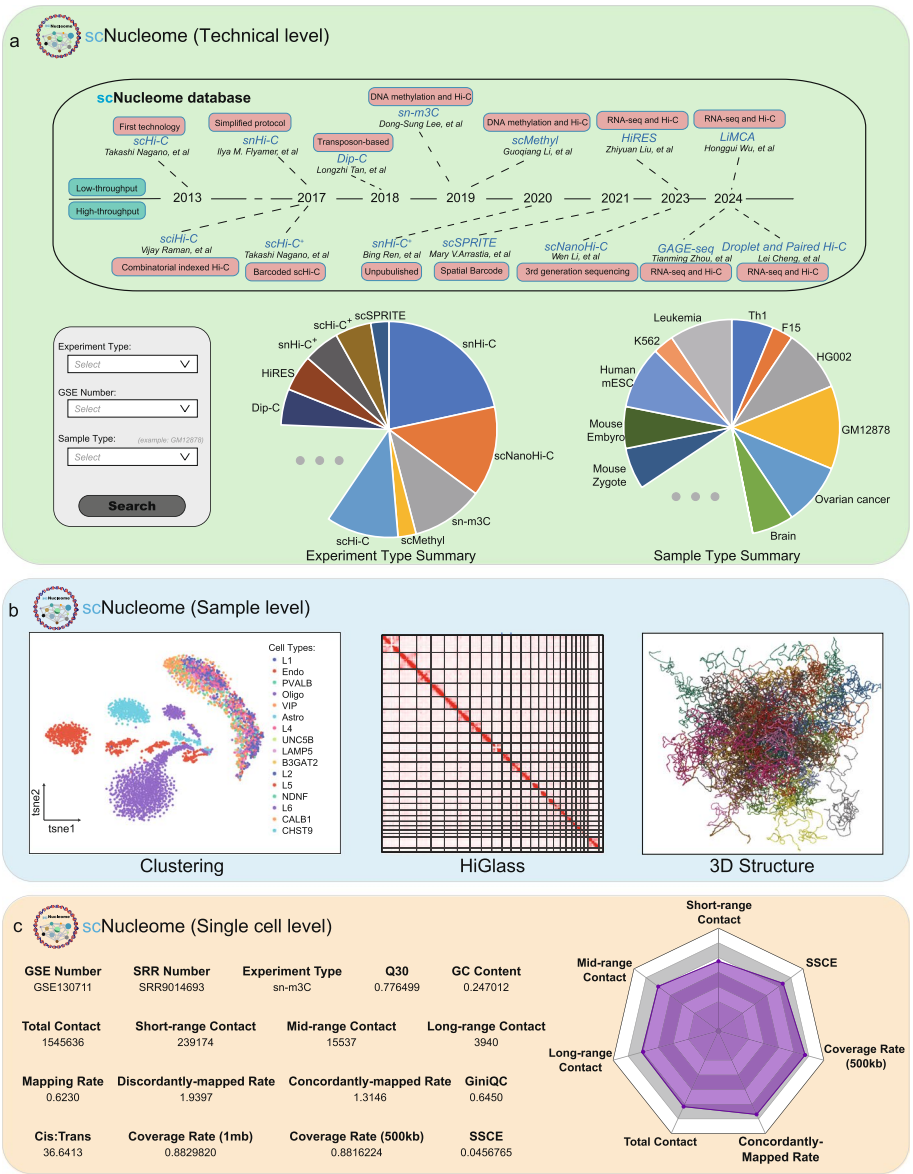

Fig. 5 Overview of the scNucleome database. **a** Milestones in the development of sc3DG-seq technologies (upper panel) and summary of the sc3DG-seq data collected in the database (lower panel). **b** Example of the analyses results and visualization choices available within the database. **c** Example of a quality control report (left) and radar visualization (right) for an individual cell available in the database, which offers a multi-dimensional view of the cell's characteristics and quality metrics

analysis and capturing higher-order chromatin interactions, while methods such as sciHi-C and snHi-C^{+} are more appropriate for high-throughput applications requiring large cell numbers. Multi-omics approaches including HiRES, sn-m3C, LiMCA, GAGE-seq, and Paired Hi-C offer the potential for integrated analysis of 3D genome organization alongside additional gene expression and DNA methylation. The scNucleome database serves as a resource to help researchers navigate these choices by providing technical specifications, performance metrics, and application examples that may inform technology selection based on specific experimental needs.

While STARK demonstrates broad compatibility with current sc3DG-seq technologies, there remain opportunities for further development. The framework currently focuses primarily on chromatin conformation data processing, and enhanced integration of multi-omics analytical capabilities could provide additional value. Similarly, expanding the suite of advanced analytical features may further enhance STARK's utility as a comprehensive analysis solution.

Nevertheless, the STARK framework represents a meaningful contribution toward addressing the analytical challenges posed by the growing complexity and volume of sc3DG-seq data. By providing a unified, standardized platform for data processing and quality control, STARK may facilitate more in-depth analyses and support continued discoveries in 3D genome organization research. The accompanying scNucleome repository, with its curated collection of uniformly processed datasets, has the potential to support collaborative efforts and contribute to progress in this evolving field.

Conclusions

This study introduces STARK, a computational framework designed for analyzing single-cell 3D genome sequencing (sc3DG-seq) data. It features three key modules—Preprocess, Cell QC, and Downstream Analysis—that provide a standardized approach for data processing, quality control, and chromatin structure analysis. STARK is versatile, supporting a wide range of data formats and sequencing methods, including both low- and high-throughput technologies. It also introduces two innovations: the EmptyCells quality control method for high-throughput data and SSCE, a metric for evaluating chromatin capture efficiency. By processing data from ~ 2 million single cells across 15 sc3DG-seq techniques, STARK helps researchers choose the most appropriate technology for their needs.

Methods

STARK framework

STARK is a software suite designed to manage the complete analysis workflow for sc3DG-seq data. It comprises three main modules, each tailored to address specific stages of the workflow: (a) STARK.Preprocess, (b) STARK.Cell QC, and (c) STARK. Downstream Analysis. The following sections are the description of the implementation:

(a) STARK.Preprocess

We initiate the preprocessing by performing sequence quality control and adaptor trimming on the raw fastq files from sc3DG-seq data using fastp [64] with default parameters. This step yields trimmed fastq files. For technologies that do not involve bisulfite conversion, the trimmed fastq files are aligned to the reference genome (hg38/mm10) using one of the aligners: BWA, Bowtie2, or minimap2 [65–67], selected based on the characteristics of the sequencing technology. For sn-m3C and scMethyl technologies, a two-round alignment approach is employed. The first-round alignment is conducted using bismark [68] to map the fastq files to the reference genome. Sequences not aligning in the first round are trimmed by 40-bp from both ends, and a second alignment

round is performed using the same settings. The alignment results are stored into BAM format and are subsequently filtered and sorted using samtools [69].

Next, we proceed to parse the BAM files to generate "pairs" files, we apply stringent filtering criteria to ensure high-quality pairs: only pairs with a mapping quality score (mapq) of 30 or higher for both fragments are retained. Subsequently, we use the STARK to generate whole-genome enzyme cutting sites for respective experimental enzymes and assign the nearest restriction sites to each alignment in a pair. For high-throughput technologies, including scNanoHi-C, snHi-C$^+$, scHi-C, scHi-C$^+$, and scSPRITE, we extract cell barcode from the raw sequencing files, linking each read to its originating single cell. Using cell barcodes, we demultiplex the contacts in the "pairs" files down to the single-cell level. The "pairs" files are then sorted, deduplicated and filtered to eliminate PCR duplicates and other artifacts. Finally, we convert the "pairs" format to the "cool" format, correct the contact matrix, and generate single-cell multi-resolution "cool" files.

(b)STARK.Cell QC

Quality control on cell level is a critical step to filter out low-quality single cells in high-throughput sequencing data analysis. We develop the "EmptyCells" algorithm to perform this task (details provided below). Initially, the algorithm removes single cells with fewer than T contacts. It then models the bottom 5% of cells with lowest contacts and applies a Monte Carlo simulation method to test each cell, identifying and removing those that are significantly low-quality. Cells that pass this filtering are considered to have high-quality single-cell 3D structures.

Next, we implement a wide range of quality control (QC) metrics to provide a comprehensive assessment for each single cell. These QC metrics include the total contact number, genomic coverage, the rates of short-range (<20 Kb), mid-range (20 kb—2 Mb), and long-range (>2 Mb) contacts, GiniQC, normalized Detection Scores (nDS), and a newly developed metric – the Spatial Structure Capture Efficiency (SSCE). The SSCE is designed for a more nuanced assessment of the structural capture efficiency in single cells. These metrics are discussed in greater detail in the corresponding section of the Methods.

(c) STARK.Downstream Analysis

1. Imputation via Random Walk with Restart (RWR)

The RWR algorithm is a variant of the traditional random walk technique. It is designed to explore the graph by randomly walking through its nodes, with the added feature that at each step, there is a probability the walk will "restart" and return to a predefined starting node. We implement this algorithm, as referenced in previous research [34], for the interpolation of single-cell 3D genome data.

First, we construct a probability transition matrix (P) by normalizing the single-cell contact matrix (A) such that each row sums to 1. The walk restarts at the starting node with a probability r, and with the probability $1-r$, it proceeds to a neighboring node. The goal is to find the steady-state π which represents the matrix given the restart probability r. The RWR is solved iteratively through the following steps:

1. Initialization: Start with an initial probability vector $\pi^0 = A$.
2. Iteration: Update π using the formula: $\pi^{t+1} = (1-r)P\pi^t + rA$. P is the transition probability matrix.
3. Convergence: Repeat the iteration until π converges, when $|| \pi^{t+1} - \pi^t || < 10^{-6}$.

Upon completion of the iterations, we obtain π^{t+1} which serves as the imputed single-cell contact matrix.

2. Dimensionality reduction and clustering

sc3DG-seq data is typically represented as a two-dimensional (2D) contact matrix for each chromosome. However, this format is not conducive to dimensionality reduction processes. To address this, we flatten each 2D matrix into a one-dimensional (1D) vector, and then concatenate these vectors to create a long vector representative of each cell's data. We proceed by normalizing the data with respect to library size to adjust for differences in sequencing depth. For dimensionality reduction, we apply Principal Component Analysis (PCA) to the normalized data. This allows us to extract the primary principal components (PCs) that capture the most variance within the dataset. Subsequently, we utilize Uniform Manifold Approximation and Projection (UMAP) [70] to transform the PCs into a two-dimensional (2D) representation. Additionally, STARK integrates advanced embedding methods including Higashi [35] and Fast-Higashi [40] which utilize hypergraph representation learning and tensor decomposition approaches to generate cell embeddings that capture complex chromatin interaction patterns. We also evaluated the embedding methods across multiple resolutions (10 kb, 50 kb, 100 kb, 500 kb, and 1 Mb) using representative datasets, with results showing that both methods achieved relatively good performance at 500 kb resolution (Additional file 1: Fig. S9-12). Visualization on this 2D space facilitates the identification of patterns and structures within the data. For the clustering analysis, we employ the Leiden community detection algorithm, a method widely recognized for its effectiveness in the analysis of single-cell RNA sequencing (scRNA-seq) data. This algorithm helps to identify clusters within the single-cell data, providing insights into the cellular heterogeneity and potential cell types or states.

3. Cell aggregation

Given the sparsity of sc3DG-seq data, an integrated analysis of data from multiple cells is often beneficial. STARK facilitates this by enabling the aggregation of data from multiple single cells for subsequent analysis, employing a "clustering-before-aggregation" strategy to preserve biologically relevant cellular heterogeneity. Specifically, STARK first uses clustering algorithms (such as scHiCluster) to identify cell subpopulations with similar chromatin conformation characteristics, then performs signal aggregation only within the same subpopulation. The aggregation is mathematically represented by the equation:

$$A(i,j) = \sum_{k-1}^{n} a_k(i,j)$$

Matrix A denotes the integrated 3D genome data from multiple cells, and a_k represents the single-cell 3D genome data from the k-th cell. The indices i and j indicate the genomic positions of interactions. This aggregation function serves to generate a pseudo-bulk dataset that represents cell types or clusters of interest. The creation of pseudo-bulk data allows for the enhancement of signal-to-noise ratio and can reveal patterns that may not be evident when analyzing single cells in isolation. This approach is particularly useful for downstream analyses such as identifying consensus chromosomal structures, studying cell type–specific interactions, and exploring the organization of the genome across different cellular states.

4. Detecting chromatin structure

Analyzing 3D genome sequencing data to identify A/B compartments, Topologically Associating Domains (TADs), and chromatin loops is crucial for understanding the spatial organization of chromatin and its role in gene expression and regulation. Below are the methods we implemented for identifying these three structural features in STARK:

A/B Compartments: We begin by conducting Principal Component Analysis (PCA) on preprocessed contact matrices. PCA helps capture interaction patterns between chromatin regions and projects them into the principal component space. The first principal component (PC1) is analyzed for positive and negative values by gene density, allowing us to distinguish A compartments, which are associated with active regions, from B compartments, which are inactive.

TAD Identification: We apply the Insulation Score algorithm [71] to the contact matrix. This score detects TAD boundaries by calculating the average change in contact frequency within sliding windows along the chromatin sequence. Significant increases in Insulation Score values indicate TAD boundaries.

Loop Identification: The process of identifying loops involves pinpointing significant hotspots in the contact matrix, which represent frequent interactions between distal genomic locations. We employ peak calling methods [71] to identify these loops, using statistical analysis to filter hotspots in the contact matrix and identify regions with significantly higher interaction frequencies compared to the background.

These methods provide a comprehensive approach to detecting key chromatin structures that are vital for elucidating the complex interactions within the genome.

5. Modeling 3D-genome structures

Reconstructing the 3D structure of chromatin from a contact matrix involves several steps. Initially, the contact matrix undergoes preprocessing to eliminate noise and adjust its resolution. Following this, distance constraints are derived from the matrix, which serve as the foundation for establishing the 3D structure. Then, Monte Carlo simulation is employed to find chromatin structures that align with these constraints. We refer to the nuc_dynamics [48] for the code implementation of this function.

EmptyCells

sc3DG-seq technologies are broadly categorized into two approaches: low-throughput non-barcoded and high-throughput barcoded methods. Non-barcoded methods involve the precise isolation of individual cells, which are then subjected to library construction and sequencing. In contrast, barcoded methods allow for the high-throughput sequencing of multiple cells simultaneously by incorporating cell barcodes. While barcoded methods enable the sequencing of a larger number of cells per experiment, it can result in a substantial variation in the number of contacts captured from each cell. Therefore, the effective filtering of low-quality cells is critical. Common quality control approach rely on total contact number thresholds to filter cells/barcodes. However, this approach have limitation, as library preparation efficiencies can vary among barcodes. Moreover, stringent thresholds may inadvertently exclude cells with less compact chromatin structures, which could be particularly problematic when analyzing cell-state transitions [17].

Inspired by EmptyDrops [72] and Cellranger (10 × Genomics, version 7.0.1), We propose a two-step method for filtering low-quality cells in sc3DG-seq data: (i) Preliminary filtering by setting a low threshold based on the total number of contacts to remove low-quality cells with insufficient sequencing or capture efficiency. (ii) Further refinement by modeling the profile of these low-quality cells to enhance the filtering process. We hypothesize that high-quality cells exhibit a diverse array of interaction types. Therefore, we incorporate the variety of interactions between different chromosomes as a key feature in our model. Mathematically, all barcodes of a sample can be represented as $A = y_{bg} \in R^{N\times\}}$, where y_{bg} denotes the number of contacts for interaction type g in barcode b, N denotes the total number of barcodes, and } denotes the interaction type. Notably, we exclude interaction types that are not present in all barcodes from our analysis to maintain consistency and accuracy in the filtering process.

In the first step, we establish a threshold T based on the total number of contacts per barcode. Barcodes that have fewer contacts than T are considered incomplete or of low quality and are therefore removed. Importantly, a barcode exceeding the threshold T does not necessarily indicate a complete cell representation; it may still only capture a subset of contacts from a cell. We define the threshold T as follows:

$$T = max\big(500, \alpha f(A_b)\big),$$

where α is a hyperparameter, which we set to 0.01, f(·) calculates the median value of the barcode contact vector A_b, which is represented as the sum of contacts across all interaction types per barcode:

$$A_b = \sum\nolimits_{g\in\}} y_{gb}.$$

Barcodes that fall below this threshold T are discarded, while the remaining barcodes proceed to the next step of the filtering process.

To construct the profile of low-quality barcodes, we selected a subset of barcodes characterized by a low number of contacts, specifically selecting the bottom 5% to form a set we designate as G. We then calculated A as the sum of contacts across all barcodes in G:

$$A_g = \sum\nolimits_{b \in G} y_{gb},$$

yielding a contact vector $\mathrm{A} = \left(A_1, \ldots, A_\}\right)$.

We apply the Simple Good-Turing algorithm to A_g to obtain the posterior expectation $\widehat{p}_g$, representing the probability of sampling a contact from an interaction type g. This method ensures that even contact types with zero observed contacts are assigned non-zero probabilities, thereby preventing undefined values in subsequent calculations [73].

For a given barcode b, the total number of contacts is denoted as t_b and can be modeled using a Dirichlet multinomial distribution. The likelihood of obtaining contacts from barcode b conditional on t_b, is defined as follows:

$$L_b = \frac{t_b!\Gamma(\alpha)}{\Gamma(t_b + \alpha)} \prod_{g=1}^{N} \frac{\Gamma\left(y_{bg} + \alpha_g\right)}{y_{bg}!\Gamma\left(\alpha_g\right)},$$

where $\alpha_g = \beta\widehat{p}_g$ β is a scaling factor used to assess the importance of $\widehat{p}_g$ in sampling from A. By modeling overdispersion in contact types through the estimation of β, our model accounts for broader range of variability in contact numbers than a simple multinomial model would allow. For example, a lower β value signifies a greater dispersion level in contact type, which could be due to variations in capture efficiency or differences in library amplification. This approach allows for more accurate p-value calculation when evaluating if a barcode deviates significantly from the low-quality set.

We employ the Monte Carlo simulation method to calculate the p-value for each barcode. For each iteration i of a given barcode, a new contact number vector is randomly sampled. The likelihood $L\prime_{bi}$ is then calculated based on the posterior expectation $\widehat{p}_g$,the corresponding total number of contacts t_b and the scaling factor β. The count R_b of likelihoods from the R round simulation that are less than L_b is utilized to compute the p-value [74]:

$$P_b = \frac{R_b + 1}{R + 1}.$$

This approach allows us to quantitatively assess the quality of each barcode and to filter out those that do not meet the predetermined quality thresholds, thereby enhancing the reliability of downstream analyses.

To validate EmptyCells' effectiveness in retaining structurally informative cells, we systematically compared its performance against traditional sequencing depth-based filtering. Comparative analysis revealed that cells retained by EmptyCells but filtered by depth thresholds still maintain distinct 3D chromatin structural features, with meta cells showing comparable correlation levels to bulk Hi-C data (Additional file 1: Fig. S14a, b). This demonstrates that EmptyCells successfully recovers structurally informative cells that would be discarded by conventional depth-based approaches, thereby maximizing the retention of high-quality single-cell data for downstream analysis.

SSCE

Spatial Structure Capture Efficiency (SSCE) is introduced as a quantitative metric designed to evaluate the efficacy of single-cell sequencing in capturing the chromatin structural information. It is crucial to recognize that while the total number of sequenced contacts can vary significantly among cells, a lower count does not necessarily indicate inferior sequencing quality. Such a result may simply reflect the capture of a specific subset of the chromatin architecture, which can still have substantial biological implications. Conversely, a high number of detected contacts does not guarantee the biological relevance of these interactions or confirm an accurate representation of the cell's chromatin spatial organization. With this in mind, we introduce a novel indicator, termed Spatial Structure Capture Efficiency (SSCE), which quantifies the proportion of intra-cellular contacts contributing to the formation of the cell's chromatin architecture. This metric is essential for a more nuanced understanding of the structural fidelity of single-cell sequencing data. The SSCE are calculated as follows:

$$\mathrm{SSCE} = \frac{\mathrm{T} + \mathrm{A}}{\mathrm{I} + \mathrm{C}},$$

where T, A, I, and B are derived from fitting the following regression:

$$C_i \sim B + I * inter_i + T * tad_i + A * ab_i.$$

In this equation, C_i denotes the aggregate count of contacts associated with chromosome i, $inter_i$ represents the number of interactions that chromosome i engages in with other chromosomes in the genome. tad_i corresponds to the count of TAD boundaries identified on chromosome i at a resolution of 40 kilobases, which is pivotal for elucidating chromatin domain organization. ab_i signifies the count of A/B compartment switches for chromosome i determined at a resolution of 100 kilobases, providing insights into the chromatin's compartmentalization. The constant B in the equation serves two critical roles. Firstly, it accounts for contacts that cannot be attributed to any specific chromosomal structure feature. Secondly, it includes contacts that may result from technical or systematic errors during the sequencing process. To ensure the accuracy of the model, any chromosomes where all variables are nullified are systematically excluded from the regression analysis.

Considering the inherent collinearity between the counts of ab_i and tad_i, we employ Ridge regression as our fitting method. The coefficients derived from the Ridge regression model are then utilized to compute the SSCE. It is logical to infer that higher values of structural features ab_i and tad_i, along with lower values of non-structural features B and I, contribute to higher SSCE scores. An elevated SSCE score, therefore, implies a more thorough and precise depiction of the single-cell chromatin structure. Additionally, we have confirmed that the loss of contacts due to various factors can result in a decrease in SSCE (Additional file 1: Fig. S4b). SSCE calculation demonstrates efficient computational performance, with processing time scaling linearly with sequencing depth as shown in the benchmarking analysis (Additional file 1: Fig. S1d).

GiniQC and genomic coverage

We employ GiniQC [57] to quantify the internal noise in sc3DG-seq data. GiniQC is also utilized within the STARK framework for data quality assessment. Genomic coverage is determined based on a deduplicated contact matrix, providing an essential measure for evaluating the distribution and comprehensiveness of the data. To thoroughly dissect the genomic interaction landscape, we compute genomic coverage at three resolutions: 10 Kb, 40 Kb, and 100 Kb. These varying resolutions allow for a multi-scale analysis, enhancing our understanding of the 3D genome organization of the genome.

Duplication rate and library complexity

The duplication rate and library complexity were assessed using 'pairtools dedup'. To effectively remove PCR duplicates, we utilized the '–max-dist 3' parameter during the deduplication process. The library complexity estimate was derived from the intrinsic estimation method implemented within pairtools. This metric provides a proxy for the number of unique molecular identifiers (UMIs) expected in a hypothetical, duplicate-free library. For the purpose of visualizing and comparing these estimates across various datasets, we applied a log-normalization transformation.

Normalized detection scores for 3D genome structures

Normalized Detection Scores (nDS) are utilized to evaluate the clarity of 3D genome structures, including chromosome territories, A/B compartments, and TADs. We have refined the approach introduced in scSPRITE, leveraging the high-throughput capabilities of single-cell sequencing while bypassing the requirement for pre-computed genome-wide 3D structural data. The Detection Scores (DS) are computed using the native contact matrix, which defines the frequency of interactions among various genomic bins. A higher DS for chromatin compartments in a cell, for instance, suggests a more pronounced preference for intra-chromosomal interactions over inter-chromosomal interactions. Detection scores were calculated for each structure in each cell as follows:

Chromosome territories

The unnormalized Detection Score (DS) for chromatin territories is calculated using the following formula:

$$\mathrm{DS}_{terr} = \frac{O_{chr}^{intra}}{E_{chr}^{intra}} - \frac{O_{chr}^{inter}}{E_{chr}^{inter}}$$

Here,O_{chr}^{intra} and E_{chr}^{intra} represent the observed and expected intra-chromosomal contact frequencies, respectively, while O_{chr}^{inter} and E_{chr}^{inter} denote the observed and expected inter-chromosomal contact frequencies. The DS of chromatin territory is calculated in different chromtain interactions, such as chr1-chr2, and does not include the same chromosome interaction, such as chr1-chr1, with a contact map resolution of 100 Kb. Thus, taking detection score of the chr1 and chr2 as an example, E_{chr}^{intra} is calculated as $N_1^c * N_1^c + N_2^c * N_2^c$,where N_1^c and N_2^c are the number of bins of chr1 and chr2. E_{chr}^{inter} is calculated as $N_1^c * N_2^c * 2$. Finally, DS_{terr} of a single cell is the mean of the different chromatin interactions' DS.

A/B compartments

In contrast to scSPRITE, which requires a prior bulk A/B compartment structural data, we leverage the high-throughput capacity of sc3DG-seq data. By merging the top 300 cells with the most contacts from each dataset or the entire dataset, we generate a pseudobulk data. This data is used to identify A/B compartments, creating a prior compartment annotation for subsequent single-cell DS calculation. This approach avoids the need for prior bulk data, which may not always be available. The unnormalized DS of A/B compartments is calculated as follows:

$$\mathrm{DS}_{AB} = \frac{O_{AB}^{intra}}{E_{AB}^{intra}} - \frac{O_{AB}^{inter}}{E_{AB}^{inter}}.$$

Here, O_{AB}^{intra} and E_{AB}^{intra} represent the observed and expected contact frequencies within compartments, respectively, while O_{AB}^{inter} and E_{AB}^{inter} denote the observed and expected contact frequencies between compartments. The DS_{AB} of a cell is obtained by calculating the A/B compartment switches, which are transitions like 'A-B-A' or 'B-A-B', derived from the pseudobulk. Thus, taking a 'A_1-B_1-A_2' with as an example, E_{AB}^{intra} is calculated as $N_1^A * N_1^A + N_1^B * N_1^B + N_2^A * N_2^A$, where N_1^A, N_1^B and N_2^A are the number of bins of A_1 compartment, B_1 compartment and A_2 compartment. E_{chr}^{inter} is calculated as $N_1^B * \left(N_1^A + N_2^A\right) * 2$. Finally, DS_{AB} of a single cell is the mean of all transition from pseudobulk. The calculation is performed on contact maps with a resolution of 100 Kb.

Topologically associating domains (TADs)

Similar to the calculation of A/B compartments, the prior TAD structure is derived from merged cells. The formula for calculating the DS of TADs is as follows:

$$\mathrm{DS}_{TAD} = \frac{O_{TAD}^{intra}}{E_{TAD}^{intra}} - \frac{O_{TAD}^{inter}}{E_{TAD}^{inter}}.$$

Here, O_{TAD}^{intra} and E_{TAD}^{intra} represent the observed and expected contact frequencies within TADs, respectively, while O_{TAD}^{inter} and E_{TAD}^{inter} denote the observed and expected contact frequencies between TADs. For the calculation of DS_{TAD}, the 10 bins before and after the boundary are taken as two TADs. Thus, the calculation of DS_{TAD} is similar to the previous mentioned calculation method. The calculation is performed on contact maps with a resolution of 40 Kb.

Finally, we normalize the detection scores. For each structure in each cell, such as chr1-chr2 when calculating DS_{terr}, 'A-B-A' or 'B-A-B' when calculating DS_{AB}, and 10 bins before and after the detected boundary when calculating DS_{TAD}, we sample the bins from the corresponding cells to construct the same structure and calcaulated the corresponding DS which repeat 100 times to get the mean of the results as the expected DS (Additional file 1: Fig. S15a, b, c). Thus, the normalized detection scores(nDS) are calculated as the DS minus the expected DS.

Insulation score and A/B compartment

To compute insulation score (IS) and annotate A/B compartments from single-cell or pseudobulk 3D genome data, we utilized the cooltools package (https://github.com/

open2c/cooltools). IS, which measures the degree of chromatin insulation, was calculated using a contact map with a 40-Kb bin resolution. We used a window size that is ten times the bin size. Specifically, the 'cooltools.insulation' function was utilized with the parameter 'ignore_diag = 3'. A/B compartments were identified using a contact map with a 100-Kb resolution. The 'cooltools.eigs_cis' function was applied. We conducted a genome-wide computation without implementing specific filters to allow for a thorough analysis of genome compartmentalization. The parameter 'n_eigs = 3' was used to capture the principal eigenvectors, which are indicative of the primary compartment patterns, with all other parameters remaining at their default settings.

Accumulative single-cell quality analysis

For a comprehensive quality analysis of single-cell data, we implemented an iterative strategy to cumulatively aggregate individual cells based on either the top 300 cells with the highest contact counts or the entire dataset. This approach allows us to evaluate the consistency of 3D genome structure detection across cells within a sc3DG-seq dataset. To measure the similarity between the aggregated cells at each iteration and the final aggregate, we calculate the Intersection over Union (IoU) of the TAD boundaries. The IoU is calculated using the following formula:

$$\mathrm{IOU}_i = \frac{A_i \cap T}{A_i \cup T}.$$

Here, A_i represents the set of TAD boundaries identified in the aggregated single-cell data at the i^{th} iteration, and T represents the set of TAD boundaries in the final aggregated cell. The IoU is a metric that quantifies the overlap between two sets of TAD boundaries. A higher IoU value signifies a greater degree of overlap, indicating higher consistency in the detected 3D structures. We applied a similar analytical framework for chromatin loop detection, where loop calling was performed using mainstream algorithms (HiCCUPS and Mustache) on incrementally aggregated cell populations. For loop detection analysis, we assessed the capability of different technologies to detect loops as a function of cell aggregation level, determining the minimum number of cells required for reliable loop identification across different sc3DG-seq methods.

Clustering evaluation metrics

To evaluate the quality of clustering results derived from cell embeddings at various resolutions, we employed a comprehensive set of external and internal validation metrics. External metrics utilize ground-truth cell type labels to assess the agreement between predicted clusters and true classifications, while internal metrics evaluate clustering coherence based on the data structure and cluster assignments. All metrics were computed using the scikit-learn library in Python, with K-means clustering applied to the embeddings, setting the number of clusters equal to the number of unique ground-truth cell types. Below, we describe each metric and provide its mathematical formulation.

External validation metrics

These metrics quantify the correspondence between the clustering output and the reference labels.

1. Adjusted Rand Index (ARI): Measures the similarity between two clusterings by considering all pairs of samples and counting pairs assigned to the same or different clusters in both the predicted and true clusterings, adjusted for chance.

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]},$$

where $\mathrm{RI} = \frac{a+b}{\binom{n}{2}}$ is the Rand Index, a is the number of pairs in the same cluster in both clusterings, b is the number of pairs in different clusters in both, $\binom{n}{2}$ is the total number of pairs, and $\mathbb{E}[\mathrm{RI}]$ is the expected Rand Index under random labeling. Values range from -1 to 1, with 1 indicating perfect agreement.

2. Normalized Mutual Information (NMI): Quantifies the mutual dependence between the true labels and predicted clusters, normalized by the average entropy of the labelings.

$$\mathrm{NMI}(U, V) = \frac{\mathrm{MI}(U, V)}{\frac{1}{2}[H(U) + H(V)]},$$

where $\mathrm{MI}(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log\left(\frac{N|U_i \cap V_j|}{|U_i||V_j|}\right)$ is the mutual information, $H(U) = -\sum_{i=1}^{|U|} \frac{|U_i|}{N} \log\left(\frac{|U_i|}{N}\right)$ is the entropy of U (and similarly for$H(V)$), U and V are the true and predicted label sets, and N is the number of samples. Values range from 0 to 1, with 1 indicating perfect correlation.

3. Adjusted Mutual Information (AMI): An adjustment of mutual information that accounts for chance, similar to ARI.

$$\mathrm{AMI} = \frac{\mathrm{MI} - \mathbb{E}[\mathrm{MI}]}{\frac{1}{2}[H(U) + H(V)] - \mathbb{E}[\mathrm{MI}]},$$

where MI, $H(U)$, and $H(V)$ are as defined for NMI, and $\mathbb{E}[\mathrm{MI}]$ is the expected mutual information computed via hypergeometric modeling of random label assignments. Values range from approximately 0 to 1.

4. Homogeneity: Assesses whether each cluster contains only samples from a single class.

$$h = 1 - \frac{H(C|K)}{H(C)},$$

where $H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{N}\log\left(\frac{n_{c,k}}{n_k}\right)$ is the conditional entropy of classes given clusters, $H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{N}\log\left(\frac{n_c}{N}\right)$ is the entropy of classes, $n_{c,k}$ is the number of samples from class c in cluster k, n_c and n_k are the sizes of class c and cluster k, and N is the total number of samples. Values range from 0 to 1.

5. Completeness: Measures whether all samples from a given class are assigned to the same cluster.

$$c = 1 - \frac{H(K|C)}{H(K)},$$

where $H(K|C)$ and $H(K)$ are defined symmetrically to homogeneity, with entropies conditioned on classes instead of clusters. Values range from 0 to 1.

6. V-measure: The harmonic mean of homogeneity and completeness, providing a balanced external metric.

$$v = 2 \cdot \frac{h \cdot c}{h + c},$$

where h is homogeneity and c is completeness. Values range from 0 to 1.

7. Fowlkes-Mallows Index (FMI): Computes the geometric mean of precision and recall for pairs of samples.

$$\text{FMI} = \frac{\text{TP}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})}},$$

where TP is the number of true positive pairs (same cluster and same class), FP is false positives (same cluster but different classes), and FN is false negatives (different clusters but same class). Values range from 0 to 1.

Internal validation metrics

These metrics evaluate clustering quality based on intrinsic data properties, without reference labels.

1. Silhouette Score: Quantifies how similar a sample is to its own cluster compared to other clusters.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i) = \frac{1}{|C_i|-1}\sum_{j\in C_i, j\neq i} d(i,j)$ is the mean intra-cluster distance for sample i in cluster C_i, $b(i) = \min_{k\neq i}\frac{1}{|C_k|}\sum_{j\in C_k} d(i,j)$ is the mean distance to the nearest cluster,

and $d(i,j)$ is the distance metric (e.g., Euclidean). The overall score is the mean $s(i)$ across all samples, ranging from −1 to 1 (higher is better).

2. Calinski-Harabasz Index (CH): Also known as the variance ratio criterion, it measures the ratio of between-cluster dispersion to within-cluster dispersion.

$$\mathrm{CH} = \frac{\mathrm{BCSS}/(k-1)}{\mathrm{WCSS}/(N-k)},$$

where BCSS is the between-cluster sum of squares, WCSS is the within-cluster sum of squares, k is the number of clusters, and N is the number of samples. Higher values indicate better clustering.

3. Davies-Bouldin Index (DB): Evaluates the average similarity ratio of each cluster to its most similar cluster.

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{S_i + S_j}{M_{ij}}\right),$$

where $S_i = \left(\frac{1}{T_i}\sum_{j=1}^{T_i} \parallel \mathbf{x}_j - \mathbf{a}_i\parallel_q^q\right)^{1/q}$ is the within-cluster scatter for cluster i (with centroid $\mathbf{a}_i$, size T_i, and norm parameter q), and $M_{ij} = \parallel \mathbf{a}_i - \mathbf{a}_j\parallel_p$ is the distance between centroids (with norm parameter p). Lower values indicate better clustering.

Overall ranking calculation

To integrate the diverse metrics into a single comparative ranking across configurations (e.g., different embedding resolutions), a normalized combined score was computed. For each metric, scores were preprocessed if necessary and normalized using min–max scaling across all configurations:

1. For the Davies-Bouldin index (where lower values are preferable), scores were inverted as:

$$\widetilde{\mathrm{DB}} = \frac{1}{1+\mathrm{DB}},$$

to align with metrics where higher is better. Infinite or NaN values were replaced with twice the maximum observed value or a large constant

2. For all metrics (post-inversion for DB), normalization was applied as:

$$\widetilde{m}_c = \frac{m_c - \min(m)}{\max(m) - \min(m)},$$

where m_c is the metric value for configuration c, and $\min(m)$ and $\max(m)$ are the minimum and maximum values of metric m across configurations. If all values for a metric were identical, a normalized value of 0.5 was assigned.

3. The combined score for each configuration was the arithmetic mean of its normalized metric values.
4. Configurations were ranked in descending order of combined scores to identify the overall best performers.

This approach ensures a balanced, multi-faceted evaluation, mitigating biases from individual metrics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-03938-x.

Additional file 1: Figures S1-S14.

Additional file 2: Table S1. STARK performance metrics.

Authors' contributions

W.-J.J. and H.-J.W. conceived and designed the study. W.-J.J., and K.C. wrote the STARK software. W.-J.J., K.C., and Y.S. collected and analyzed the sc3DG-seq data with input from C.Z., R.G., N.W., F.L., A.L. and K.C. designed and constructed the scNucleome website. H.-J.W., M.X., X.Z., T.F., and Y.-J.W. supervised the study. W.-J.J., K.C., Y.S., and H.-J.W. wrote the manuscript with input from all authors. All authors reviewed and approved the final manuscript.

Peer review information

Andrew Cosgrove and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Funding

This work was supported by the National Natural Science Foundation of China (32470662 and 32270683), the Beijing Natural Science Foundation (5242006), the Fundamental Research Funds for the Central Universities (BMU2021YJ064), the National Key R&D Program of China (2021YFC1712805), the National Natural Science Foundation of China (61572327, U20A20345 and 61972257), the CAMS Innovation Fund for Medical Sciences (2021-I2M5-003), the Natural Science Foundation of Shanghai (20JC1413800), and the National Key R&D Program of China (2020YFA0803803 and 2018YFA0900600). We gratefully acknowledge the High-performance Computing Platform of Peking University for conducting the data analyses.

Data availability

Statistical analyses were conducted using the Python 3.9.0 platform (https://www.python.org/). STARK is freely available under MIT licence at https://github.com/CCCKW/stark [75] and https://zenodo.org/records/18065403 [76]. scNucleome is accessible at http://scnucleome.com/. All sc3DG-seq datasets analyzed in this study were obtained from publicly available repositories. The datasets are organized by technology as follows: scSPRITE datasets were downloaded from GEO under accession GSE154353 [23, 77] HiRES datasets were downloaded from GEO under accession GSE223917 [22, 78]. Dip-C datasets were downloaded from GEO under accession GSE117874 [19, 79]. sciHi-C datasets were downloaded from GEO under accessions GSE185608 [46, 80] and GSE84920 [18, 81]. scHi-C datasets were downloaded from GEO under accessions GSE201917 [49, 82], GSE201918 [49, 82], GSE222390 [50, 83], and GSE48262 [16, 84]. scHi-C+ datasets were downloaded from GEO under accessions GSE148793 [85, 86] and GSE94489 [17, 87]. sn-m3C datasets were downloaded from GEO under accessions GSE124391 [20, 88] and GSE130711 [20, 89]. snNanoHi-C datasets were downloaded from GEO under accession GSE217189 [24, 90]. snHi-C+ datasets were downloaded from 4DN under accessions 4DNESF829JOW [91] and 4DNESJQ4RXY5 [91]. snHi-C datasets were downloaded from GEO under accessions GSE80006 [14, 92], GSE196300 [51, 93], GSE171159 [45, 94], GSE130585 [45, 95], GSE132264 [47, 96], and GSE100569 [53, 97]. scMethyl datasets were downloaded from GEO under accession GSE119171 [21, 98]. LiMCA datasets were downloaded from GEO under accessions GSE239969 [25, 99] and GSE240114 [25, 99]. GAGE-seq datasets were downloaded from GEO under accession GSE238001 [26, 100]. Droplet Hi-C and Paired Hi-C datasets were downloaded from GEO under accession GSE253407 [27, 101].

Declarations

Ethics approval and consent to participate

Not applicable for this study.

Competing interests

The authors declare no competing interests.

Author details
[1]Department of Cardiology and Institute of Vascular Medicine, State Key Laboratory of Vascular Homeostasis and Remodeling, NHC Key Laboratory of Cardiovascular Molecular Biology and Regulatory Peptides, Beijing Key Laboratory of Cardiovascular Receptors Research, Peking University Third Hospital, Peking University, Beijing 100191, China. [2]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Peking University Cancer Hospital & Institute, Beijing 100142, China. [3]Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China. [4]Key Laboratory of Bioresource Research and Development of Liaoning Province, College of Life and Health Sciences, Northeastern University, Shenyang 110819, China. [5]Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China. [6]Research Unit of Medical Science Research Management/Basic and Clinical Research of Metabolic Cardiovascular Diseases, Chinese Academy of Medical Sciences, Beijing 100050, China. [7]Center for Precision Medicine Multi-Omics Research, Institute of Advanced Clinical Medicine, Beijing Advanced Center of Cellular Homeostasis and Aging-Related Diseases, Peking University, Beijing 100191, China.

Received: 13 November 2024 Accepted: 7 January 2026
Published online: 21 January 2026

References

1. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17:661–78.
2. Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. Cell. 2016;164:1110–21.
3. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013;153:307–19.
4. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature. 2011;472:247.
5. Zheng H, Xie W. The role of 3D genome organization in development and cell differentiation. Nat Rev Mol Cell Biol. 2019;20:535–50.
6. Hafner A, Boettiger A. The spatial organization of transcriptional control. Nat Rev Genet. 2023;24:53–68.
7. Tan L, Ma W, Wu H, Zheng Y, Xing D, Chen R, et al. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. Cell. 2021;184:741-758. e717.
8. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
9. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, et al. Two independent modes of chromatin organization revealed by cohesin removal. Nature. 2017;551:51–6.
10. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.
11. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet. 2001;2:292–301.
12. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.
13. Achinger-Kawecka J, Stirzaker C, Portman N, Campbell E, Chia K-M, Du Q, et al. The potential of epigenetic therapy to target the 3D epigenome in endocrine-resistant breast cancer. Nat Struct Mol Biol. 2024. https://doi.org/10.1038/s41594-023-01181-7.
14. Flyamer IM, Gassler J, Imakaev M, Brandão HB, Ulianov SV, Abdennur N, et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature. 2017;544:110–4.
15. Kempfer R, Pombo A. Methods for mapping 3D chromosome architecture. Nat Rev Genet. 2020;21:207–26.
16. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502:59–64.
17. Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, Baran Y, et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature. 2017;547:61–7.
18. Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, et al. Massively multiplex single-cell Hi-C. Nat Methods. 2017;14:263–6.
19. Tan L, Xing D, Chang CH, Li H, Xie XS. Three-dimensional genome structures of single diploid human cells. Science. 2018;361:924–8.
20. Lee D-S, Luo C, Zhou J, Chandran S, Rivkin A, Bartlett A, et al. Simultaneous profiling of 3d genome structure and DNA methylation in single human cells. Nat Methods. 2019;16:999–1006.
21. Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, et al. Joint profiling of DNA methylation and chromatin architecture in single cells. Nat Methods. 2019;16:991–3.
22. Liu Z, Chen Y, Xia Q, Liu M, Xu H, Chi Y, et al. Linking genome structures to functions by simultaneous single-cell Hi-C and RNA-seq. Science. 2023;380:1070–6.
23. Arrastia MV, Jachowicz JW, Ollikainen N, Curtis MS, Lai C, Quinodoz SA, et al. Single-cell measurement of higher-order 3D genome organization with scSPRITE. Nat Biotechnol. 2022;40:64–73.
24. Li W, Lu J, Lu P, Gao Y, Bai Y, Chen K, et al. Scnanohi-c: a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. Nat Methods. 2023;20:1493–505.
25. Wu H, Zhang J, Jian F, Chen JP, Zheng Y, Tan L, et al. Simultaneous single-cell three-dimensional genome and gene expression profiling uncovers dynamic enhancer connectivity underlying olfactory receptor choice. Nat Methods. 2024;21:974–82.
26. Zhou T, Zhang R, Jia D, Doty RT, Munday AD, Gao D, et al. GAGE-seq concurrently profiles multiscale 3D genome organization and gene expression in single cells. Nat Genet. 2024;56:1701–11.

27. Chang L, Xie Y, Taylor B, Wang Z, Sun J, Armand EJ, et al. Droplet hi-c enables scalable, single-cell profiling of chromatin architecture in heterogeneous tissues. Nat Biotechnol. 2024;10. https://doi.org/10.1038/s41587-024-02447-1.
28. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
29. Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using chromosome conformation capture. Methods. 2012;58:192–203.
30. Quinodoz SA, Bhat P, Chovanec P, Jachowicz JW, Ollikainen N, Detmar E, et al. SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. Nat Protoc. 2022;17:36–75.
31. Xiao T, Zhou W. The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr. 2020;9:163–73.
32. He J, Liu D, Zhao L, Zhou D, Rong J, Zhang L, et al. Myocardial ischemia/reperfusion injury: mechanisms of injury and implications for management (review). Exp Ther Med. 2022;23:430.
33. Li X, Feng F, Pu H, Leung WY, Liu J. scHiCTools: a computational toolbox for analyzing single-cell Hi-C data. PLoS Comput Biol. 2021;17:e1008978.
34. Zhou J, Ma J, Chen Y, Cheng C, Bao B, Peng J, et al. Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation. Proc Natl Acad Sci U S A. 2019;116:14011–8.
35. Zhang R, Zhou T, Ma J. Multiscale and integrative single-cell Hi-C analysis with Higashi. Nat Biotechnol. 2022;40:254–61.
36. Ma W, Wang F, Fan Y, Hao G, Su Y, Wong K-C, et al. Deepnanohi-c: deep learning enables accurate single-cell nanopore long-read data analysis and 3D genome interpretation. Nucleic Acids Res. 2025;53:gkaf640.
37. Xie Q, Han C, Jin V, Lin S. Hicimpute: a bayesian hierarchical model for identifying structural zeros and enhancing single cell Hi-C data. PLoS Comput Biol. 2022;18:e1010129.
38. Liu J, Lin D, Yardımcı GG, Noble WS. Unsupervised embedding of single-cell Hi-C data. Bioinformatics. 2018;34:i96–104.
39. Kim H-J, Yardımcı GG, Bonora G, Ramani V, Liu J, Qiu R, 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.
40. Zhang R, Zhou T, Ma J. Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi. Cell Syst. 2022;13:798-807. e796.
41. Cheng J. HiCDiff: single-cell Hi-C data denoising with diffusion models. 2023.
42. Zheng Y, Shen S, Keleş S. Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. Genome Biol. 2022;23:222.
43. Yu M, Abnousi A, Zhang Y, Li G, Lee L, Chen Z, et al. Snaphic: a computational pipeline to identify chromatin loops from single-cell Hi-C data. Nat Methods. 2021;18:1056–9.
44. Li X, Lee L, Abnousi A, Yu M, Liu W, Huang L, et al. Snaphic2: a computationally efficient loop caller for single cell hi-c data. Comput Struct Biotechnol J. 2022;20:2778–83.
45. Chatzidaki EE, Powell S, Dequeker BJ, Gassler J, Silva MC, Tachibana K. Ovulation suppression protects against chromosomal abnormalities in mouse eggs at advanced maternal age. Curr Biol. 2021;31:4038-4051. e4037.
46. Bonora G, Ramani V, Singh R, Fang H, Jackson DL, Srivatsan S, et al. Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and X inactivation. Genome Biol. 2021;22:1–36.
47. Silva MC, Powell S, Ladstätter S, Gassler J, Stocsits R, Tedeschi A, et al. Wapl releases Scc1-cohesin and regulates chromosome structure and segregation in mouse oocytes. J Cell Biol. 2020;219:e201906100.
48. Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, et al. 3d structures of individual mammalian genomes studied by single-cell Hi-C. Nature. 2017;544:59–64.
49. Pang QY, Tan TZ, Sundararajan V, Chiu YC, Chee EYW, Chung VY, et al. 3d genome organization in the epithelial-mesenchymal transition spectrum. Genome Biol. 2022;23:121.
50. Lambuta RA, Nanni L, Liu Y, Diaz-Miyar J, Iyer A, Tavernari D, et al. Whole-genome doubling drives oncogenic loss of chromatin segregation. Nature. 2023;615:925–33.
51. Dequeker BJH, Scherr MJ, Brandao HB, Gassler J, Powell S, Gaspar I, et al. MCM complexes are barriers that restrict cohesin-mediated loop extrusion. Nature. 2022;606:197–203.
52. Lee DS, Luo C, Zhou J, Chandran S, Rivkin A, Bartlett A, et al. Simultaneous profiling of 3d genome structure and DNA methylation in single human cells. Nat Methods. 2019;16:999–1006.
53. Gassler J, Brandao HB, Imakaev M, Flyamer IM, Ladstatter S, Bickmore WA, et al. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. EMBO J. 2017;36:3600–18.
54. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, et al. Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus. Cell. 2018;174:744-757. e724.
55. Ordonez CD, Redrejo-Rodriguez M. DNA polymerases for whole genome amplification: considerations and future directions. Int J Mol Sci. 2023. https://doi.org/10.3390/ijms24119331.
56. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, et al. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). Science. 2017;356:189–94.
57. Horton CA, Alver BH, Park PJ. Giniqc: a measure for quantifying noise in single-cell Hi-C data. Bioinformatics. 2020;36:2902–4.
58. Wu H-J, Landshammer A, Stamenova EK, Bolondi A, Kretzmer H, Meissner A, et al. Topological isolation of developmental regulators in mammalian genomes. Nat Commun. 2021;12:4897.
59. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503:290–4.
60. Ye C, Paccanaro A, Gerstein M, Yan K-K. The corrected gene proximity map for analyzing the 3D genome organization using Hi-C data. BMC Bioinformatics. 2020;21:1–18.
61. Roayaei Ardakany A, Gezer HT, Lonardi S, Ay F. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. Genome Biol. 2020;21:256.

62. Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 2018;19:1–12.
63. Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, et al. Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 2021;49:W431–7.
64. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
65. Bayat A, Gaeta B, Ignjatovic A, Parameswaran S. Pairwise alignment of nucleotide sequences using maximal exact matches. BMC Bioinformatics. 2019;20:261.
66. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9:357–9.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
68. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics. 2011;27:1571–2.
69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
70. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018. https://doi.org/10.1038/nbt.4314.
71. Open2C, Abdennur N, Abraham S, Fudenberg G, Flyamer IM, Galitsyna AA, et al. Cooltools: enabling high-resolution Hi-C analysis in Python. PLoS Comput Biol. 2024;20:e1012067.
72. Lun AT, Riesenfeld S, Andrews T, Dao TP, Gomes T, Jamboree PitsHCA, et al. Empty drops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019;20:1–9.
73. Gale WA, Sampson G. Good-turing frequency estimation without tears. J Quant Linguist. 1995;2:217–37.
74. Phipson B, Smyth GK. Permutation p-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010. https://doi.org/10.2202/1544-6115.1585.
75. Jiang W, Cai, K., & Sun, Y. STARK. Github. https://github.com/CCCKW/stark (2025).
76. Jiang W, Cai K, Sun Y. 2025. STARK Zenodo. https://doi.org/10.5281/zenodo.18065403.
77. Arrastia MV JJ, Ollikainen N, Curtis MS, Lai C, Quinodoz SA, Selck DA, Guttman M, Ismagilov RF. A single-cell method to map higher-order 3D genome organization in thousands of individual cells reveals structural heterogeneity in mouse ES cells. Gene Expression Omnibus. https://identifiers.org/geo:GSE154353 (2020).
78. Liu Z CY, Xia Q. Linking genome structures to functions during lineage specification by simultaneous single-cell Hi-C and RNA-seq simultaneous single-cell Hi-C and RNA-seq. Gene Expression Omnibus. https://identifiers.org/geo:GSE223917 (2023).
79. Li H TL. Single-cell chromatin conformation capture of diploid cells by Dip-C. Gene Expression Omnibus. https://identifiers.org/geo:GSE117874 (2018).
80. Nucleome D. sci-Hi-C on mESCs differentiated to embryoid body. Gene Expression Omnibus. https://identifiers.org/geo:GSE185608 (2021).
81. V R. Massively multiplex single-cell Hi-C. Gene Expression Omnibus. https://identifiers.org/geo:GSE84920 (2017).
82. Pang QY HR. 3D genome organization in the epithelial-mesenchymal transition spectrum. Gene Expression Omnibus. https://identifiers.org/geo:GSE201917 (2022).
83. Lambuta RA NL, Liu Y, Diaz-Miyar J, Iyer A, Tavernari D, Katanayeva N, Ciriello G, Oricchio E. Whole genome doubling drives oncogenic loss of chromatin segregation. Gene Expression Omnibus. https://identifiers.org/geo:GSE222390 (2023).
84. Nagano T LY, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single cell Hi-C reveals cell-to-cell variability in chromosome structure. Gene Expression Omnibus. https://identifiers.org/geo:GSE48262 (2013).
85. Rappoport N, Chomsky E, Nagano T, Seibert C, Lubling Y, Baran Y, et al. Single cell Hi-C identifies plastic chromosome conformations underlying the gastrulation enhancer landscape. Nat Commun. 2023;14:3844.
86. Rappoport N CE, Nagano T, Seibert C, Lubling Y, Baran Y, Lifshitz A, Leung W, Mukamel Z, Shamir R, Fraser P, Tanay A. Single cell Hi-C deconvolutes proliferation and differentiation of chromosome conformation in embryonic mesoderm, ectoderm and blood. Gene Expression Omnibus. https://identifiers.org/geo:GSE148793 (2022).
87. Nagano T LY, Varnai C, Dudley C, Leung W, Baran Y, Mandelson-Cohen N, Wingett S, Fraser P, Tanay A. Cell cycle dynamics of chromosomal organisation at single-cell resolution. Gene Expression Omnibus. https://identifiers.org/geo:GSE94489 (2017).
88. Luo C LD, Zhou J, Chandran S, Rivkin A, Bartlett A, Nery JR, Fitzpatrick C, O'Connor C, Dixon JR, Ecker JR. Single-cell multi-omic profiling of chromatin conformation and DNA methylome. Gene Expression Omnibus. https://identifiers.org/geo:GSE124391 (2018).
89. Luo C LD, Zhou J, Chandran S, Rivkin A, Bartlett A, Nery JR, Fitzpatrick C, O'Connor C, Dixon JR, Ecker JR. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Gene Expression Omnibus. https://identifiers.org/geo:GSE130711 (2019).
90. Li W LJ. scNanoHi-C: a single-cell long-read concatemer sequencing method to reveal high-order structures within individual cells. Gene Expression Omnibus. https://identifiers.org/geo:GSE217189 (2023).
91. Ren B. Single nucleus Hi-C on WTC-11 with GFP tagged AAVS1. 4DN Data Portal. https://data.4dnucleome.org/experiment-set-replicates/4DNESF829JOW; https://data.4dnucleome.org/experiment-set-replicates/4DNESJQ4RXY5 (2020).
92. Flyamer IM GJ, Imakaev M, Brandao HB, Ulianov SV, Abdennur N, Razin SV, Mirny L, Tachibana-Konwalski K. Single-cell Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Gene Expression Omnibus. https://identifiers.org/geo:GSE80006 (2017).
93. Dequeker B GJ, Powell S, Gaspar I, Flyamer I, Tachibana K. MCM complexes are barriers that restrict cohesin-mediated loop extrusion. Gene Expression Omnibus. https://identifiers.org/geo:GSE196300 (2022).
94. Chatzidaki EE PS, Dequeker BJ, Gassler J, Silva MC, Tachibana K. Ovulation suppression protects against chromosomal abnormalities in mouse eggs at advanced maternal age. Gene Expression Omnibus. https://identifiers.org/geo:GSE171159 (2021).
95. Chatzidaki E GJ, Powell S, Tachibana K. Ovulation suppression prevents the increase in egg aneuploidy with maternal age in mammals. Gene Expression Omnibus. https://identifiers.org/geo:GSE130585 (2021).

96. Silva MC PS, Ladstätter S, Gassler J, Stocsits R, Tedeschi A, Peters J, Tachibana K. Wapl-mediated release of Scc1-cohesin regulates chromosome structure and segregation in mouse oocytes. Gene Expression Omnibus. https://identifiers.org/geo:GSE132264 (2020).
97. Gassler J BH, Imakaev M, Flyamer IM, Ladstätter S, Bickmore WA, Peters J, Mirny LA, Tachibana-Konwalski K. A Mechanism of Cohesin-Dependent Loop Extrusion Organizes Zygotic Genome Architecture. Gene Expression Omnibus. https://identifiers.org/geo:GSE100569 (2017).
98. Li G LY, Kellis M, Ren B. Long-range epigenetic concordance revealed by simultaneous profiling of DNA methylation and chromosomal conformation changes in bulk and single cell Methyl-HiC. Gene Expression Omnibus. https://identifiers.org/geo:GSE119171 (2018).
99. Wu H ZJ, Jian F. Simultaneous single-cell three-dimensional genome and gene expression profiling Uncovers Dynamic Enhancer Connectivity Underlying Olfactory Receptor Choice. Gene Expression Omnibus. https://identifiers.org/geo:GSE239969;https://identifiers.org/geo:GSE240114. (2024).
100. Ma J DZ, Zhou T. Concurrent profiling of multiscale 3D genome organization and gene expression in single mammalian cells. Gene Expression Omnibus. https://identifiers.org/geo:GSE238001 (2024).
101. Chang L XY, Ren B. Droplet Hi-C enables scalable, single-cell profiling of chromatin architecture in heterogeneous tissues. Gene Expression Omnibus. https://identifiers.org/geo:GSE253407 (2024).

Publisher's Note