Briefings in Bioinformatics, 22(6), 2021, 1–13

https://doi.org/10.1093/bib/bbab316
Review

Single-cell Hi-C data analysis: safety in numbers

Aleksandra A. Galitsyna and Mikhail S. Gelfand

Corresponding author: Mikhail S. Gelfand, Skolkovo Institute of Science and Technology, Skolkovo, Russia. Tel: +7 (495) 280-14-81; E-mail: M.Gelfand@skoltech.ru

Abstract

Over the past decade, genome-wide assays for chromatin interactions in single cells have enabled the study of individual nuclei at unprecedented resolution and throughput. Current chromosome conformation capture techniques survey contacts for up to tens of thousands of individual cells, improving our understanding of genome function in 3D. However, these methods recover a small fraction of all contacts in single cells, requiring specialised processing of sparse interactome data. In this review, we highlight recent advances in methods for the interpretation of single-cell genomic contacts. After discussing the strengths and limitations of these methods, we outline frontiers for future development in this rapidly moving field.

Key words: single cell; chromatin; single-cell Hi-C; conformation capture; single-cell sequencing

Introduction

Detecting specific DNA positioning in single cells was first proposed over half a century ago [44, 72]. Deriving statistically reliable general patterns of chromatin folding in single cells, however, has been challenging [5]. Improvements towards this goal have included: increasing the number of analysed cells, studying more loci (up to the complete genome), reducing the size of the interacting regions and improving discriminative power for detection of contacts at a broader scale of spatial distances. There are two main approaches: *microscopy based* and *capture based*. These two types of methods, despite their limitations, provide complementary views on the chromatin structure of single cells [92].

Targeted microscopy approaches measure spatial distances between genomic regions in individual cells using labelled probes. These typically involve complicated probe design, which can be overcome with a new *in situ sequencing* technique [73] but remains challenging to implement. With any microscopy approach, trade-offs have to be considered: which cells are analysed (fixed or living), number of targeted regions, time dynamics and resolution of obtained images. For an extended discussion, we refer the reader to recent reviews [5, 8].

Chromosome conformation capture uses crosslinking, digestion and proximity ligation to detect genomic regions located close to each other in 3D space. It was originally designed for inputs of millions of cells and had higher statistical power than microscopy [23]. An explosion of conformation-based techniques, including the high-throughput sequencing-based Hi-C [64], has paved the way for new discoveries expanding our general understanding of DNA folding in eukaryotic cells [34], bacterial cells [19] and even viruses [9]. For eukaryotes, these patterns include *topologically associating domains (TADs)*, promoter-enhancer and architectural *loops* and *compartments* (reviewed in-depth by [6, 21, 22, 84]).

A long-standing impediment to our interpretation and understanding of structure formation principles is that chromatin features in individual cells are not equivalent to the average features in a population of cells [31]. To address this problem, the first *single-cell chromosome conformation capture assay (scHi-C)* reduced the scale of the traditional Hi-C protocol to one cell per reaction tube [68]. Then, scHi-C was extended by

Aleksandra Galitsyna is a PhD student in Prof. Gelfand's group at Skolkovo Institute of Science and Technology. Her scientific agenda includes various topic in the broad field of chromatin research focused on biology and bioinformatics of single cells.
Mikhail Gelfand is a bioinformatician, professor at the Center of Life Sciences and vice president for biomedical research at the Skolkovo Institute of Science and Technology, Moscow, Russia. His scientific interests include a broad range of problems, ranging from bacterial genomics and evolution to transcriptomics, splicing and mRNA editing in eukaryotes, with chromatin structural analysis being one of them.
Submitted: 31 May 2021; **Received (in revised form):** 09 July 2021

introduction of sorting into multi-well plates and tagmentation followed by polymerase chain reaction (PCR) [69]. A similar approach, *single-nucleus Hi-C (snHi-C)* substituted traditional PCR with whole-genome amplification and cut out the biotin fill-in step. This came, however, at the cost of larger sequencing volumes and data processing [28, 32]. *Diploid chromatin conformation capture (Dip-C)* has adapted tagmentation-based strategies [86, 87], simplifying the experimental protocol [85]. Single-cell combinatorial indexed Hi-C (sciHi-C) is yet another powerful technique based on several rounds of combinatorial barcoding of diluted samples without isolation of individual cells [49, 76]. scHi-C can be combined with other assays to investigate the methylome, such as *Methyl-3C* and *sn-m3C-seq* [55, 57]. For the sake of simplicity, we will refer to all the family of methods a scHi-C throughout this review.

Alongside scHi-C, there is a growing family of many-body interaction capture methods, including *MC-3C* [88], *PORE-C* [91], *Nano-C* [13]. These methods recover up to several dozens of pairwise contacts from individual cells but cannot yet compete with scHi-C in genome-wide searches for architectural features. *Single-cell SPRITE* is a ligation-free method that generates 30 times more contacts but captures interacting complexes instead of pairs [2].

The main challenge of analysing scHi-C data is extreme data sparsity. On average, up to 700 000 interactions are captured in any given cell (for mouse [55]). Thus, the power of scHi-C manifests itself when data for multiple cells are available. Firstly, it makes the detection of chromatin patterns of individual cells statistically reliable. Twenty cells may already be sufficient to assess the presence of TADs, compartments and loops at the level of individual cells of *Drosophila* [93]. Secondly, multiple cells may be clustered into groups of similar types and pooled *in silico*. Such pseudo-bulk Hi-C of scHi-C-guided groups is a better alternative to bulk Hi-C, where the contacts formed in different cell types are indistinguishable [69, 85, 87]. To analyse such data, one needs specialised tools and computational pipelines, which are currently designed *ad hoc* and are rarely re-used or cross-tested. Here, we describe the diversity of recent scHi-C studies and summarise computational approaches to single-cell interactome data (for a recent review of similar topics, see [102]).

Overview of single-cell Hi-C techniques

Like traditional bulk Hi-C, single-cell Hi-C includes chromatin crosslinking, cells permeabilisation, DNA digestion, proximity ligation and library preparation. A crucial step of scHi-C, however, is either *isolation* or *barcoding* of individual cells. To separate contacts from each nucleus, a typical approach is to isolate cells or nuclei into individual reaction mixtures and perform subsequent steps separately. The isolation can be done following crosslinking of cells [28, 82], after ligation [85–87] or right before de-crosslinking [16, 68, 69]. Technically, this is performed by manual placement of each nucleus into a single tube [28, 32] or fluorescence-activated cell/nucleus sorting (FACS/FANS) into individual wells of a plate [16, 82, 87]. Right before or during sorting, optional steps can be included, such as imaging [52, 82] or bisulfite conversion [55, 57]. Isolation-free technique *single-cell combinatorial indexed Hi-C (sciHi-C)* involves several rounds of combinatorial barcoding of the diluted cells [49, 76]. Isolation-free sciHi-C requires demultiplexing as one of the first data processing steps, while the isolation approach may [69] or may not include this step. A more comprehensive overview of the scHi-C experimental technique can be found [92], but we will highlight aspects of different protocols that are particularly relevant for data processing (Figure 1).

The initial step of the scHi-C protocol is to crosslink cells with formaldehyde, resulting in the fixation of DNA-DNA interactions. Next, cell membranes are lysed to guarantee the delivery of reagents into the nucleus. Then, DNA is digested by a restriction enzyme such as DpnII that cuts at the four-letter palindromic motif GATC (Figure 1A). This produces free ends of restriction fragments, which are then ligated either directly [28, 32, 85–87], after biotin fill-in [68, 69, 82] or after ligation of a biotinylated bridge adaptor [49, 76] (Figure 1B). Ligation junctions containing biotin-labelled nucleotides are pulled down using streptavidin. This pulldown is omitted in some scHi-C variants because it results in a loss of meaningful contacts [28, 32, 85–87]. Regardless of the ligation procedure, properly formed junctions are expected to contain specific sequences (restriction sites with or without a bridge, Figure 1), which can be used to computationally select real contacts [93]. The final step of scHi-C is to extract DNA and prepare it for sequencing. Multiple library preparation strategies were probed with scHi-C (Figure 1C), including whole-genome amplification (Illustra WGA in [28], META WGA in [87]), tagmentation followed by PCR [69], digestion with a restriction enzyme followed by primers ligation and PCR [68], barcoding and PCR [76] or PCR with random primers [57]. While tagmentation and restriction enzyme digestion generate fixed-point cuts in the DNA resulting in simple rules for computational deduplication of the pairs with coinciding mapping positions, this is not the case for whole-genome amplification and PCR with random primers, for which other deduplication schemes should be used. Finally, amplified DNA is purified and sequenced in the paired-end mode.

Data processing workflow

The data processing workflow (Figure 2A and B) consists of general steps shared with typical Hi-C: optional pre-processing of reads (trimming, demultiplexing, etc.), read mapping, optional restriction fragment assignment, filtration of contacts and deduplication and binning with generation of single-cell Hi-C maps. The cells are typically filtered by the quality and/or the number of contacts.

Mapping of reads

As with any other conformation capture, scHi-C generates chimeric DNA molecules (Figure 2C), making the mapping of these discontinuous reads to multiple genomic locations non-trivial [51]. Standard mappers, such as BOWTIE2 [54], cannot reliably map such reads. There are four main approaches to treat scHi-C chimaeras, three of them transferred from traditional bulk Hi-C: split read alignment, iterative mapping and read clipping. The fourth approach is one-read-based mapping (ORBITA), a special case of the split read alignment [93], which attempts to find only those contact pairs that are directly ligated (Figure 2c). In the *split read alignment* strategy, specialised mappers like BWA MEM [58] detect multiple sequential alignments in each read. Of these, only the representative alignments are retained (typically, the alignments at 5′-end). Some studies use the information about 3′-end alignments to specify the endpoints of contacting fragments [86]. *Iterative mapping* is a method of analysing chimeric reads initially used for traditional Hi-C [42] and adapted for single-cells [28, 32]: short 5′ sequences of increasing size are iteratively selected on both forward and reverse reads until the mapping of the pair

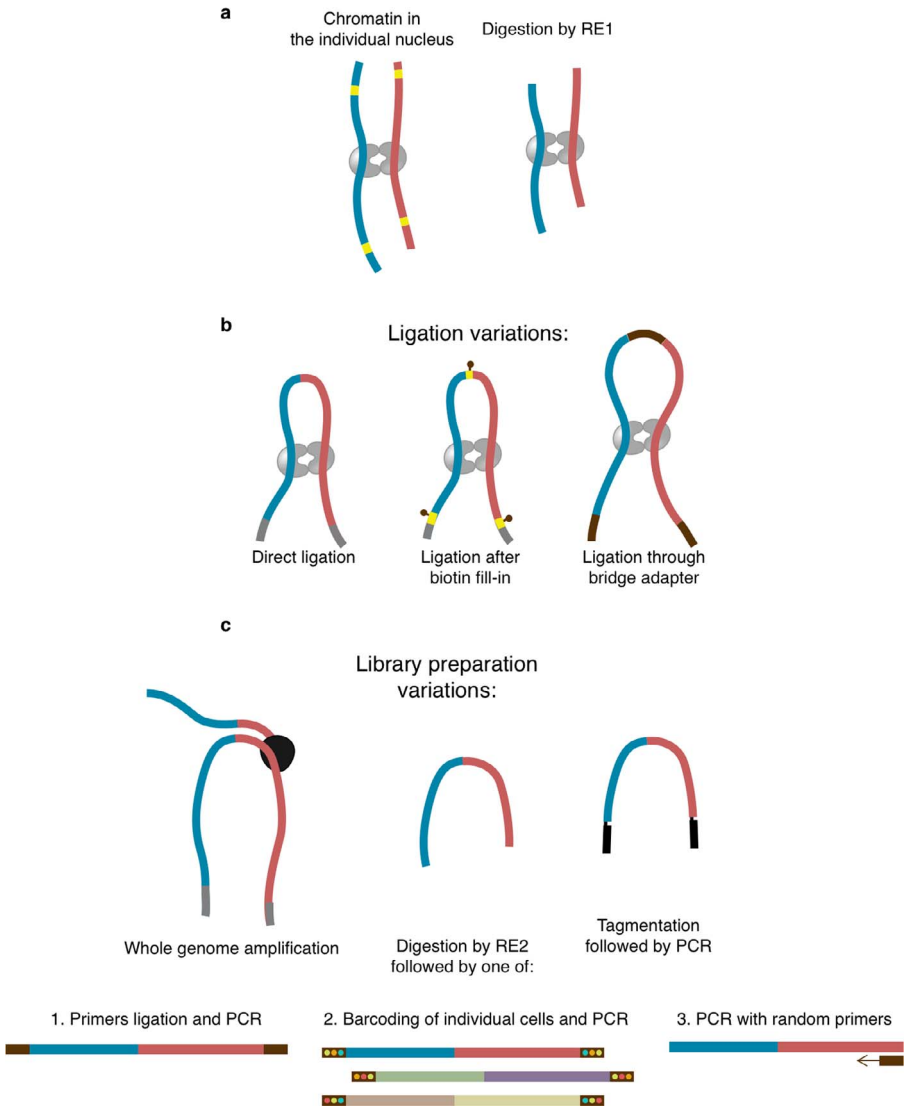


Figure 1. Overview of variations in scHi-C protocols relevant for data processing. **A**. Cross-linking and digestion, used in any scHi-C. **B**. Variations of the ligation step. **C**. Variations of the library preparation. RE1 and RE2 denote restriction enzymes selected for corresponding stages.

(or coverage of the full read length) is achieved [51]. In *read clipping*, reads are scanned for the restriction site [69, 82] or bridge adapter [76], and all the 3′ sequences after the match are removed. Two resulting paired sequences (one for forward and one for reverse read) are mapped independently and form a contact pair if the mapping was successful. However, only *one-read-based interactions annotation* utilises the information on chimeric parts to guarantees that the observed pair is a direct ligation junction of DNA fragments (Figure 2c). This approach reduces erroneous contacts in scHi-C data [93].

Another problem during scHi-C read mapping is genetic variation. Some regions of the genome of the studied cells differ from the reference hampering the mappability. Moreover, the cells are not guaranteed to descend from a single clone [86] and may have intrinsic variation, such as single-nucleotide polymorphisms (SNPs). Thus, some studies [82] ignore genomic locations with SNPs and prohibit mapping mismatches. On the other hand, SNPs can be a powerful source of information to help distinguish haplotype alleles [16, 69, 86] and impute the contacts of the maternal and paternal chromosomes [86].

Filtering of contacts

After mapping, the scHi-C maps are vastly populated with *amplification duplicates, contacts of promiscuous genomic regions* and *artifactual contacts*, which can be detected and filtered out.

Amplification duplicates are identical or nearly identical copies of the same contact pairs generated during library preparation. Depending on the experimental protocol, the scHi-C duplicates do not necessarily have the same mapping positions in the genome. Whole-genome amplification and PCR with random primers produce DNA fragments that may originate at random locations close to actual ligation position. Thus, if a group of contact pairs has the same restriction fragments [76] or their termini [69, 93], these contacts are likely to have been duplicated and should be merged into a single contact. Alternatively, contacts of the same 500 bp-bins [28] or contacts located closer than 1 kb [86] may be merged directly [28] or iteratively [86].

The genome coverage in conformation capture is affected by multiple factors, including replication, DNA accessibility, GC-content and active chromatin state [42, 78, 97]. In bulk Hi-C,

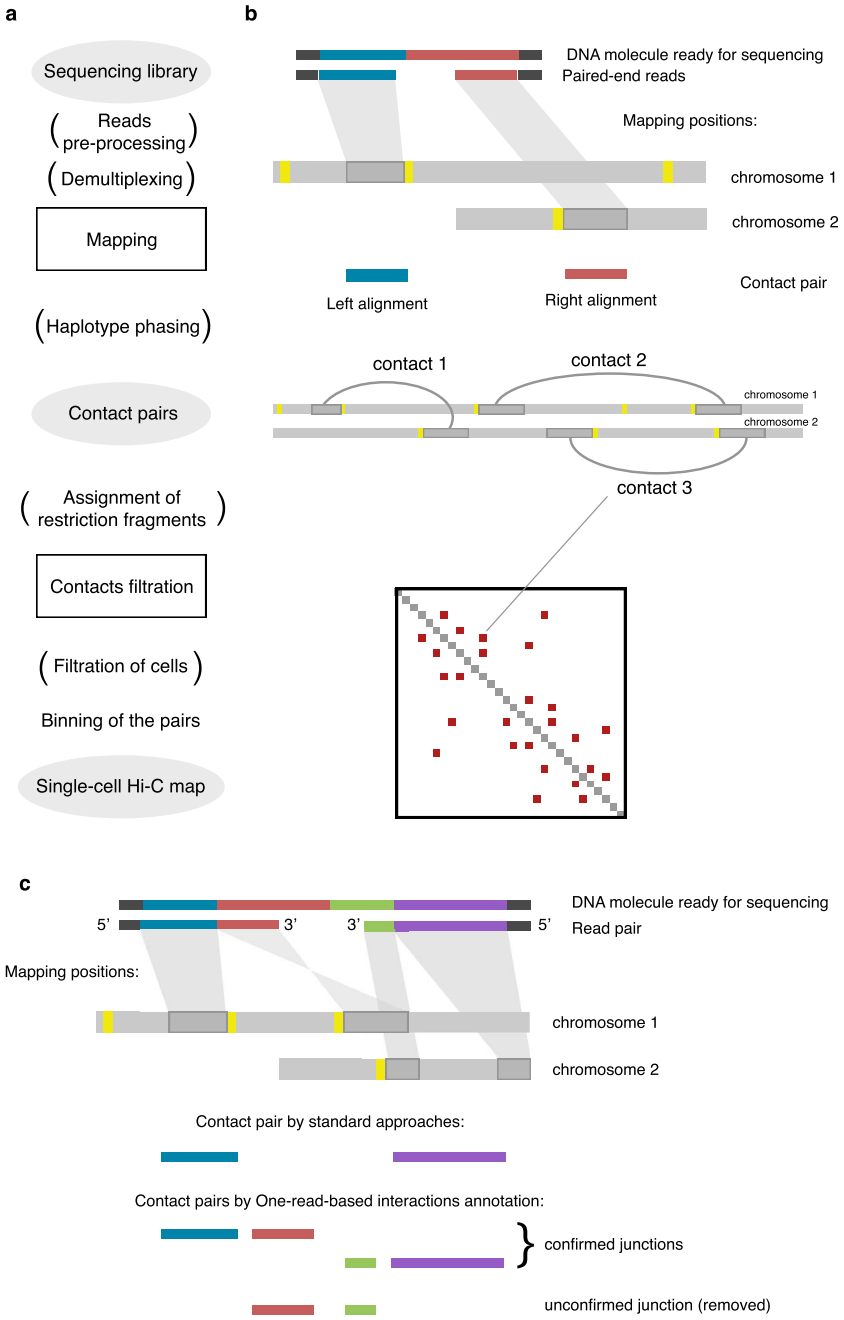


Figure 2. Outline of single-cell Hi-C data processing. The steps in brackets are optional, depending on the scHi-C protocol and the pipeline specifics.

this is mitigated by correction, such as iterative balancing [42]. However, due to data sparsity, this step is not recommended for scHi-C (although proposed as intermediary step of quality assessment [40]), and little research has been devoted to scHi-C correction alternatives [59, 66]. In the absence of data correction, scHi-C may bear intrinsic biases, such as larger numbers of contacts formed by active regions [93] and early replication domains [69]. Larger numbers of contacts have been suggested for regions with genomic rearrangements [69], e.g. Stevens *et al.* [82] detected trisomy by the increased number of contacts for the whole chromosome. As a partial remedy, one can remove *contacts of promiscuous genomic regions* [68, 69, 93], e.g. 1 Kb regions that have more than ten contacts in a given cell [86].

Artifactual contacts are random contacts happening at various stages of scHi-C sample preparation and data processing, typically not representative of the real 3D conformation of chromatin and impairing downstream analysis. First of all, properly formed and mapped pairs should be located close to the restriction sites. scHi-C protocols using Phi29 phage polymerase can generate switch templates during WGA that are devoid of this feature and should be discarded [93]. The original scHi-C protocol generates a number of *spurious ligations*, likely represented by the pairs supported by a single read [68]. Frequent artefacts are *sequencing pairing mismatches*, having a global rate of 0.1% for Illumina [69, 76], as assessed by admixture of phiX174 DNA to mouse cells [69]. Stevens *et al.* [82] suggested a general scheme for filtering a broad

range of scHi-C artefacts, which is based on the assumption that the regions in close spatial proximity have the neighbouring genomic regions located nearby, also forming a contact. Thus, if the contact is isolated (e.g. is not supported by neighbours within 2 Mb distance [82]), it is likely to represent an artefact and should be removed [82, 86].

Filtering of cells

Data from some cells should entirely be discarded due to the failure of the protocol in those cells. Multiple criteria to identify such problematic cells were proposed: *robustness to downsampling* [93], *fraction of read-pairs sequenced only once* [68] and *fraction of non-digested DNA* [69]. The most commonly used criterion is *cell coverage*, that is, the total number of detected contacts per cell [69]. For example, the cell coverage in sciHi-C follows the bimodal distribution, with low-coverage cells likely representing in-solution DNA noise [76]. Yet, another popular criterion is cumulative contacts properties, such as *cis-to-trans ratio* [68, 69, 76], defined as the ratio of the intrachromosomal contacts to the interchromosomal contacts. Typically, interchromosomal contacts in the chromatin occur with a lower probability than intrachromosomal ones, a phenomenon called *chromosome territoriality* [17]. Artifactual contacts are less likely to depend on the 3D distance between corresponding genomic positions and, thus, a deviating cis-to-trans ratio for a cell might signify excessive spurious ligation. Similar assumptions are used to filtrate the cells by *distance decay properties* of contacts [69] and *cross-species ligation frequency* [69, 76]. Another notion guiding the choice of high-quality cells is that scHi-C contacts tend to be found in clusters. Based on this observation, *GiniQC* measures the level of unevenness of inter-chromosomal scHi-C maps [40].

Data structure

The scHi-C contact data are typically represented as a matrix, similar to the standard Hi-C [68]. Each cell in this matrix corresponds to a pair of genomic bins, and the value in a cell is the absolute number of interactions between these bins. A set of experiments is stored as a set of matrices, while specialised file formats exist to store matrices for a number of cells, such as scool [96]. Hypergraphs [99] and 'topics' [49] are representations for a set of cells used for specialised applications, such as prediction of contacts using machine learning [99] and data decomposition [49]. For special applications, scHi-C can be represented as a vector, for example, when scRNA-Seq methods are transferred to 2D data [37]. The 3D model is a popular representation, although it requires substantial preprocessing of the data and is not necessarily back-convertible to the set of initial contacts [68, 82, 86].

Graph representation

Graph representation [10, 100] is a popular representation that can be used to *upper bound for the number of pairwise contacts in scHi-C maps* [93] (Figure 3A). This upper bound can be defined for scHi-C but not bulk because a single cell with defined DNA content is used in the experiment. It depends on the number of restriction fragments that can potentially form contacts, which in each cell depends on the restriction site frequency, the organism's genome size and the number of DNA copies in a particular cell type. For example, a single copy of the mouse genome mm10 [15] contains 6.6 million DpnII restriction sites (Figure 3B). In theory, if both ends of each restriction fragment were ligated to the ends of other restriction fragments and all ends are ligated, then the fragments form a circle graph. Thus, the number of contacts that could be detected would equal to the number of restriction fragments (Figure 3A). If two copies of the mouse genome are present (in a diploid cell), the number of possible contacts will be around 13.2 million. This number may be higher for cells during mitosis, S or G2 phase of the cell cycle, when the genomic content, and hence the number of restriction fragments, is completely or partially doubled. Although non-realistic to achieve in the working conditions of scHi-C, this number can serve as a theoretical upper bound to the possible number of pairwise contacts in a single nucleus. Notably, the largest number of contacts per cell obtained to date for mammals [57, 85] is already larger than the theoretical limit for the haploid genome of *Drosophila melanogaster* (Figure 3C), suggesting that the complete recovery of contacts of small genomes is possible with scHi-C.

The upper bound estimate can serve as a normalisation factor for contacts recovery in scHi-C studies (Figure 3D). The best standard scHi-C [93] has 17% contacts recovery and the joint assay with methylation, sn-m3C-seq, almost reaches 25% [55]. It is important to note that for an ideal scHi-C with 100% recovery, we still cannot expect more than 2.4 interactions per 1 Kb of the genome (for haploid mm10 genome). This number is two orders of magnitude lower than bulk Hi-C (around 1700 contacts per 1Kb or genome in neural progenitor cells [7]). Thus, even if the theoretical limit is reached, scHi-C remains profoundly sparse and specialised software is required for its downstream analysis.

Data analysis

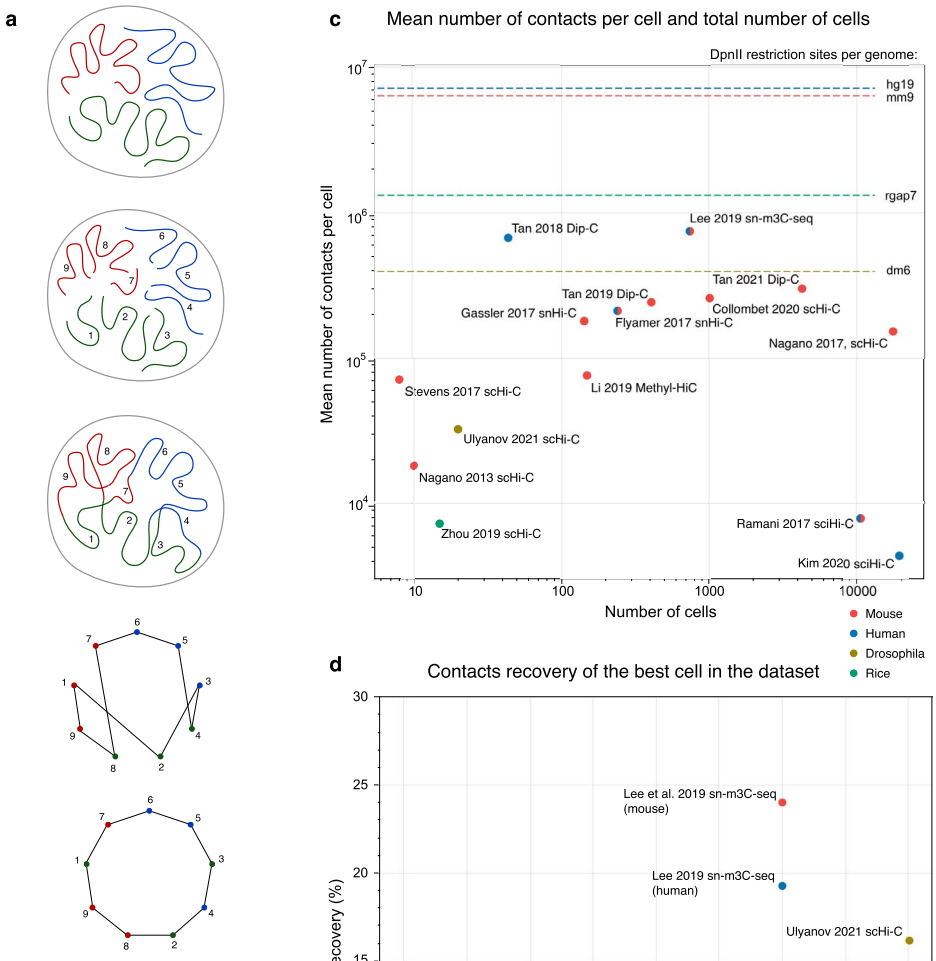
There are two general approaches to the scHi-C data analysis, depending on the solution to the problem of low statistical power of scHi-C data sparsity. In the first one, every single cell is processed independently. It includes building its 3D model, data imputation, aggregation analysis and features calling. In the second approach, single-cell maps are analysed together, then grouped and pooled to produce pseudo-bulk Hi-C maps.

Structure reconstruction

A typical approach for the 3D structure reconstruction is to build a *beads-on-string model* restrained by molecular dynamics with simulated annealing [68]. Each bead corresponds to a genomic bin of a given size (ranging from 10 Kb [93] to 1 Mb [69]), while each bond is either a polymer backbone or an observed scHi-C contact. The simulation starts from a random initial conformation, where the beads involved in observed scHi-C interactions might be overstretched. The beads connected by bonds are attracted to each other, forcing a rearrangement of the structure so that connected beads are located in close spatial proximity. Some bonds do not balance and remain overstretched; thus, they can be removed [82, 93] as potential experimental artefacts [53]. Other proposed solutions include Bayesian inference [11], recurrence plots [39] and lattice models [103]. All these methods remain data driven and do not account for the actual mechanisms of chromatin structure formation [43].

Imputation of missing data

Due to contacts sparsity, applications of bulk Hi-C analysis tools to scHi-C are restricted [59]. To mitigate this effect, imputation techniques bring the numbers of scHi-C contacts closer to bulk [102]. Zhou *et al.* [100] populate the map with contacts



Genome	Number of DpnII restriction sites
Human (hg38)	7199481
Human (hg19)	7227576
Mouse (mm10)	6684538
Mouse (mm9)	6415222
Drosophila (dm3)	332951
Rice (rgap7)	1325057

Figure 3. **A**. Illustrative upper bound estimation of the possible number of pairwise contacts per single cell. The theoretical genome has nine restriction fragments that form a circle graph after ideal ligation (nodes are restriction fragments with the valency of 2, edges denote ligation of their ends). **B**. Total numbers of DpnII restriction sites for the single copies of popular genomes. **C**. Descriptive statistics of published scHi-C studies. The lines represent the upper bounds for the possible number of contacts per single cell from (B). Colour indicate species. **D**. The best cells for some of the published scHi-C datasets as a function of the publication time. For C and D, we use the numbers reported in the supplementary materials of the original studies, when possible. For each study, we indicate the first author and the names of scHi-C techniques self-reported by the authors. For [49] and [76], the mean is calculated based on the median count per dataset. For [86], we used the cleaned contacts after removal of damaged cells. For [55], the calculated mean is based on the numbers reported for 741 cells in the supplementary table.

generated by a random walk, making the scHi-C graph closer to a complete clique. Stevens *et al.* [82] and Ulianov *et al.* [93] use the maps imputed by polymer models. Notably, both TADs and compartments can also be readily assessed from model-imputed maps [82, 93], with TADs similar to those in original scHi-C data [93]. As a substantial breakthrough in scHi-C data imputation, inter-cellular patterns of contacts can be accounted for by the hypergraph neural network [99]. Some studies test the technical possibility to transfer dropout imputation algorithms for single-cell RNA-Seq, although lacking theoretical support [37].

Contacts aggregation and features calling

Two approaches have been suggested to study TADs, loops and compartments in scHi-C maps, aggregation analysis and *features calling* (Figure 5). During *aggregation*, the statistics of contacts is accumulated over predefined regions of the genome (e.g. CTCF

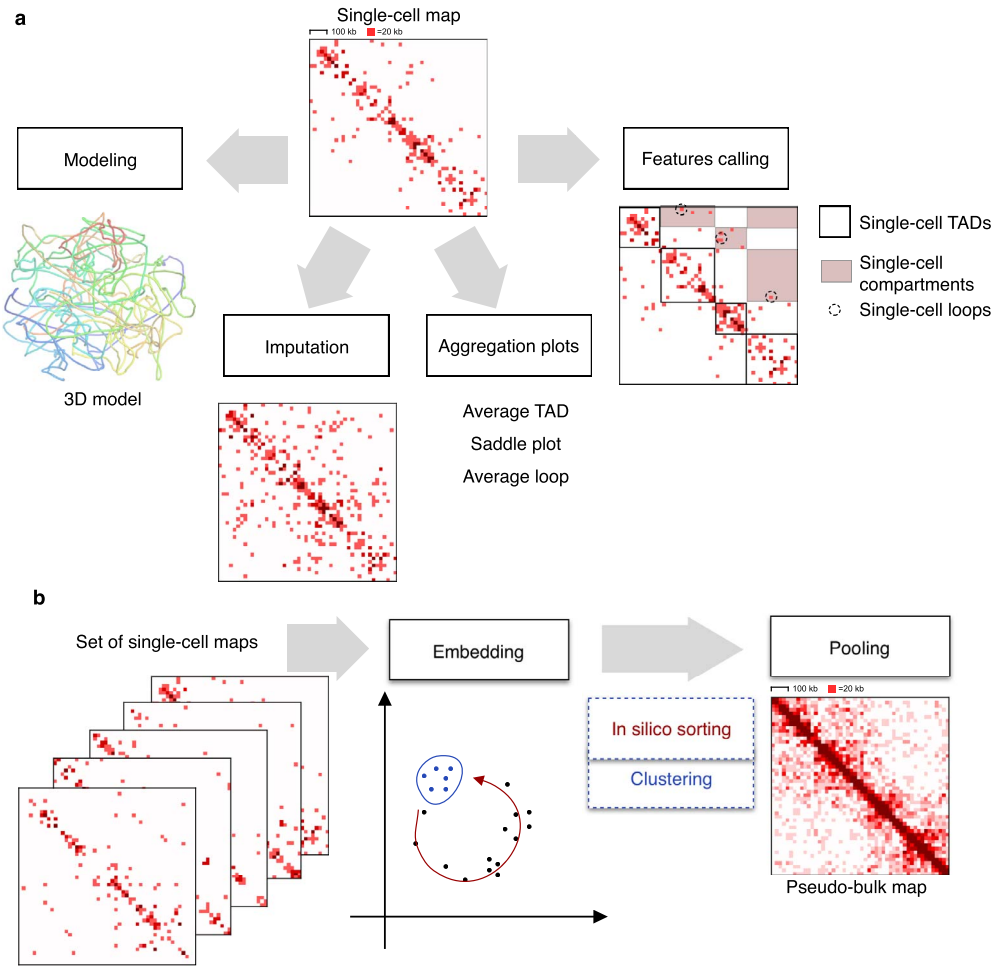


Figure 4. Approaches to studying a single scHi-C map (**A**) and a set of scHi-C maps (**B**). Single-cell Hi-C maps from [28] for the region chr1:9000000-1000000.

binding positions to assess loops; bulk TADs or bulk compartments). Aggregation confirms the presence of these chromatin features in individual cells [32], and there is specialised software for this purpose [29]. With *features calling*, the positions of individual loops [82], TADs [28, 60, 75, 93] and compartments [75, 86] are found directly in the scHi-C map, demanding high-quality scHi-C maps and providing insight into variability between individual cells. For example, the positions of TADs in individual cells demonstrated higher stability of TAD boundaries between individual cells of *Drosophila* than between mouse oocytes [93].

scHi-C embedding

scHi-C data are multidimensional ($\sim N^2$ contacts measurements for N genomic regions) and can be projected into a space of lower dimension for visualisation, clustering and sorting. Typical visualisation is a scatter plot where each dot is a cell, and the axes correspond to some characteristics of the cells. The values on the axes can be derived from some additional measurement, such as the levels of the DNA replication marker geminin and DNA content from FACS [69] or the level of DNA methylation [55, 57]. Alternatively, the axes can represent some explicitly calculated interpretable characteristic of the scHi-C maps, such as the total number of contacts, the cis-to-trans ratio [16, 69] and the percentage of local/mitotic contacts [16, 69]. Tan *et al.* [86] characterise the 3D models instead, plotting the strength of the Rabl configuration, the centrality of telomeres, the number of interchromosomal neighbours, the average CpG content of the neighbours and the probability of cell-type-specific loops.

Finally, the axes might not readily correspond to any known biological characteristics—scHi-C maps can be transformed and subjected to *dimensionality reduction* by the *principal component analysis (PCA)* or other techniques (see Table 1 for comparison). For example, Ramani *et al.* [76] apply PCA to matrices of interchromosomal interactions and find that the first component explains a large fraction of the variance (52.1%) and strongly correlates with the coverage. The combination of the first and second (1.07% explained variance) components distinguishes cell types. Nagano *et al.* [69] observe the cell cycle-dependent embedding of scHi-C by calculating the pairwise symmetric Kullback–Liebler divergence on vectors of distance decays and subsequent spectral embedding. Collombet *et al.* [16] apply *uniform manifold approximation and projection (UMAP)* to vectors of TAD contact profiles; Li *et al.* [60] perform PCA on pairwise similarities of TAD profiles; Tan *et al.* [87] calculate the compartment score profiles for each cell, take 20 principal components and then visualise it with *t-distributed stochastic neighbour embedding (t-SNE)*. One of the most generalised approaches is HiCRep [100], which calculates a similarity matrix between each pair of individual cells, taking the *stratum-adjusted correlation coefficient (SCC)* measure of similarity. HiCRep with *subsequent multidimensional scaling (MDS)* has proved to be one of the best approaches to study embedded

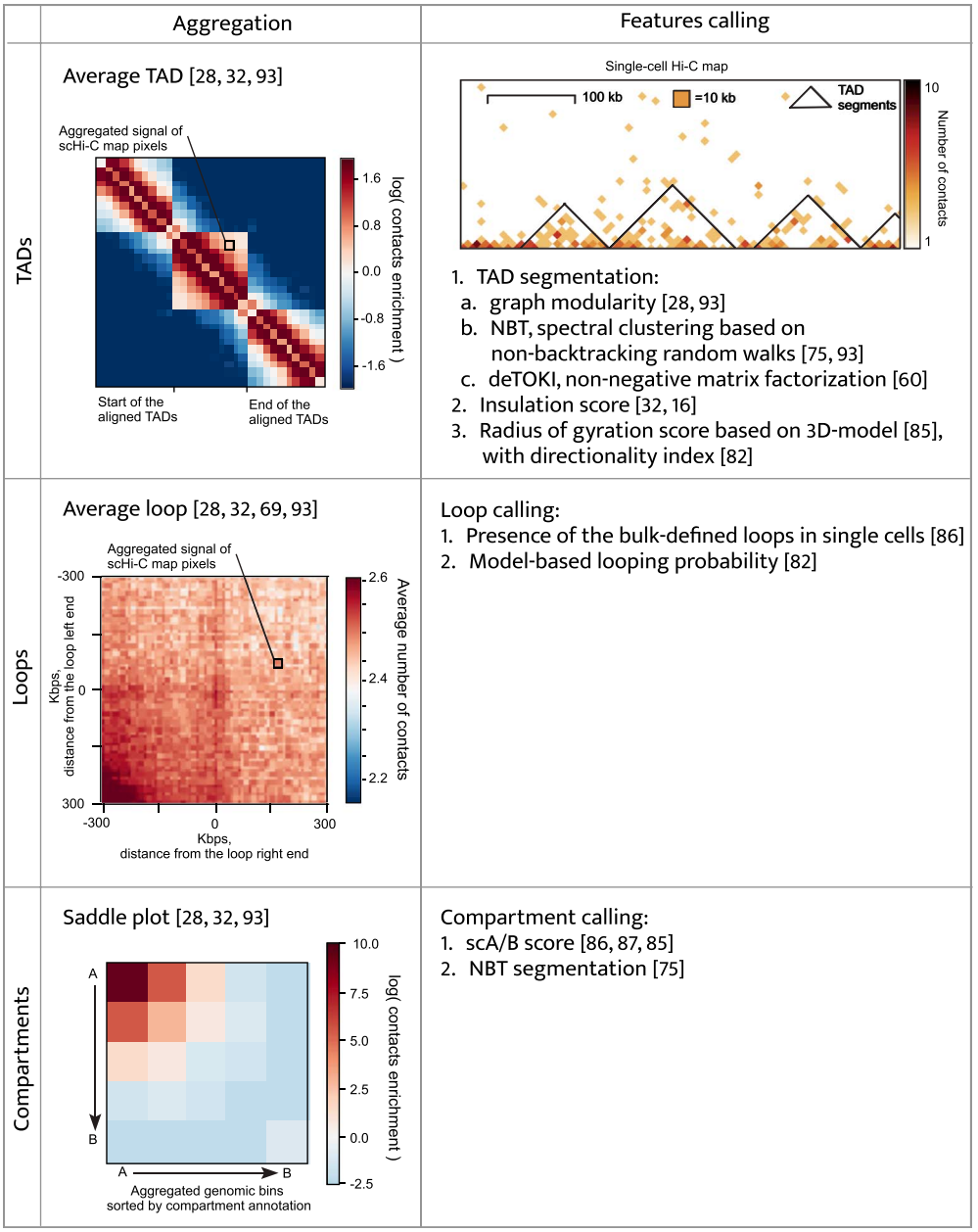


Figure 5. Comparison of aggregation of contacts and features calling for TADs, loops and compartments. All the examples are for the *Drosophila* scHi-C map of Cell 1 from [93]. Average TAD and saddle plot are for bulk TADs and compartments, while average loop is for the top 1000 regions with the highest content of RED chromatin state from [48].

scHi-C datasets [65]. In this approach, Zhou *et al.* [100] propose to impute potential dropouts before the embedding to increase the cluster separation. The imputation was further supplemented it with scRNA-Seq dropout correction methods [37] (but see the discussion above).

An alternative, scHiCExplorer [96], implements an approximate nearest neighbour method with a local sensitive hash function, MinHash. Finally, some approaches suggest using the co-occurrence of contacts in individual cells to base the embedding on meaningful single-cell patterns. For example, Kim *et al.* [49] applied *latent Dirichlet allocation* to factorise the scHi-C dataset into a set of documents, words and topics, and Zhang *et al.* [99] used a *hyper-graph neural network*. In all these studies, the axes created by *in silico* approaches are rarely interpreted, and it might be of interest to correlate them with various scHi-C characteristics such as the contact coverage, distance decay, strength of TADs, loops and compartments.

A more exotic approach is to describe scHi-C space in terms of topological data analysis [10]. Finally, joint assays of the methylome and interactome [55, 57] allow for independent embeddings of scHi-C and single-cell methylation patterns and subsequent comparison of resulting embeddings.

To date, no comprehensive studies on embedding all existing scHi-C datasets have been published. Moreover, there have been no attempts to embed datasets originating from different species, although scHi-C data for human [28, 49, 76, 86], mouse

Table 1. Summary of major scHi-C embedding techniques

Family of embedding methods	Linearity	Primary reference	Special scHi-C pre-processing	Special measure of similarity/difference between cells	Explicit usage of contacts co-occurrence patterns
PCA	Linear	[100]	Raw binned matrix	–	No
		[76]	Interchromosomal interactions profile	–	No
		[60]	TAD profile	–	No
		[87]	Compartment score profile	–	No
t-SNE	Non-linear	[85]	20 PCs of compartment score profiles	–	No
Spectral embedding	Non-linear	[69]	Distance decays	Symmetric KL	No
MDS	Non-linear	[65]	Distance decay	Jensen–Shannon divergence	No
		[100]	scHi-C binned matrix after smoothing and random-walk imputation	SCC	No
UMAP	Non-linear	[16]	TAD contact profiles	–	No
		[49]	Cell-topic matrix after LDA	–	Yes
		[99]	Hypergraph embedding	–	Yes

MDS indicates multidimensional scaling; SCC, stratum-adjusted correlation coefficient; UMAP, uniform manifold approximation and projection.

[16, 68, 69, 76, 82, 85, 87], *Drosophila* [93] and rice [101] are available. This might identify species-specific patterns in genomic interactions and their variability.

While both linear and non-linear embeddings of scHi-C have been proposed, advanced *manifold learning* techniques are yet to be developed for scHi-C, analogous to the outbreak of embedding methods for single-cell RNA-Seq data (reviewed in [67]). At that, multiple, diverse formalisations of scHi-C as matrices, graphs and vectors allow for a broad field of embedding techniques to be studied on these datasets.

In silico sorting, clustering and pooling

Based on the position in the embedding space, scHi-C data can be *in silico* sorted [69] or clustered [85]. Nagano *et al.* [69] observed the ordering of the cells by the position in the cell cycle, while Tan *et al.* [85] derived subtypes of mouse brain cells using k-means. Collombet *et al.* [16] relied on outliers in the embedding space to filter out cells undergoing mitosis and retain only interphase embryonic cells.

Specialised approaches, including the ones based on machine learning, have been designed for scHi-C data clustering. Typically, these applications require embedding (see below). The quality of clustering is tested on datasets with known ground truth (e.g. types of pronuclei in the mouse zygote [28] or types of cells forming the dataset [76]). Each cluster, or group of cells, is assigned with a particular cell type and the quality is usually assessed by normalised mutual information [62] or adjusted rand score [62, 100].

The resulting groups of cells can be pooled by simple summation of single-cell Hi-C maps, resulting in *ensemble*, or *pseudo-bulk*, Hi-C and analysed as typical bulk Hi-C [16, 69, 85]. Pseudo-bulk scHi-C maps are a powerful technique for detection of cell-type specific differences in the chromatin architecture. For example, pseudo-bulk mitotic cells lack the TAD and compartment structure [69], while subtypes of brain cells have differences in regions of cell-type specific genes [85].

The long-studied field is the reverse of the pooling, namely deconvolution of bulk interaction maps into a set on single cells [46]. Such approaches aim to construct a population of genome structures with a total set of genomic interactions approximating (or equal to) a set observed in a population of nuclei. Several advanced techniques including machine learning have been suggested, such as *maximum likelihood* [89], *Bayesian inference* [12], *fractal Monte Carlo weight enrichment with Bayesian deconvolution* [74], *Monte Carlo with bag of little bootstraps* for the generation of bootstrap structures [83] and, most recently, stochastic embedding [36]. However, these approaches are limited by the number of models that approximate bulk datasets (up to several tens of thousands), although around 5–10 million structures contribute to the typical bulk Hi-C map. Nevertheless, it might be interesting to demonstrate the reversibility of the pooling of a low number of single-cell maps by applying some of these methods to pseudo-bulk datasets. Guarnera *et al.* [36] assessed the variability of polymers after deconvolution, which might be interesting to compare with results obtained from embeddings of real scHi-C.

Design of scHi-C controls

Due to the complex nature of scHi-C data, a good practice is to design scHi-C controls to validate the hypotheses. These include *sampled, shuffled* or *de novo generated randomised scHi-C maps*, which typically have the same number of contacts as real cells. *Sampled maps* are populated by contacts randomly selected from

bulk [93] or ensemble [68, 76] datasets. However, it creates maps less sparse and heterogeneous than real scHi-C maps [100]. Thus, an effective number of sampled contacts can be increased or additional artificial noise can be introduced [100]. *Shuffled maps* are single-cell maps with randomly permuted pairs of contacts [68]. This procedure retains coverage by contacts but removes any information on the spatial structure, including distance decay. Sampling and shuffling can be combined together: bulk Hi-C maps first randomised, preserving the coverage and distance decay, and then sampled [69]. *De novo generative models* do not rely directly on the observed contact maps while preserving the meaningful properties of scHi-C maps. For example, thresholding the distance between genomic regions in polymer models [93] produces control maps with meaningful distance decays. A more advanced alternative, stepwise generation of single-cell Hi-C-like maps, preserves both distance decay and observed coverage by contacts [93].

Controls like this allow differentiating the technical and biological properties of the single-cell contact maps for features calling (such as TADs) and aggregation analysis [93]. They provided the baseline for assessing the general quality of modelling by the number of violated constraints [68]. Further, they demonstrated that scHi-C maps are non-random [82] and chromatin features of the modelled cells are similar to that of the real cells [69, 82, 93]. Yet another important observation is that real scHi-C data are more variable and sparse than bulk subsamples [100]. Although randomised scHi-C control is a powerful method, it is sporadically used in scHi-C studies. This will improve with the development of specialised tools for this task and the emergence of theoretical studies on the statistical properties of single-cell contacts.

Outlook and challenges

Single-cell Hi-C is a young and rapidly developing field in chromatin biology. Due to its extreme data sparsity and complicated experimental protocol, the quality of the datasets has been a limiting factor. However, with the emergence of simplified and cheaper protocols [76, 86], we anticipate continued growth of both coverage of scHi-C and number of cells analysed, leading to improved data resolution and statistical reliability of the biological results. This will also stimulate the development of new data processing and analysis methods. However, as we demonstrated here, scHi-C data have a natural upper bound for the possible number of recovered single-cell interactions; thus, data sparsity will remain a challenge for the field.

Despite the substantial efforts to work with sparse data, the computational analysis of scHi-C has not reached maturity yet. For example, a recent re-analysis of datasets from three studies demonstrated that inappropriate contacts mapping may result in the accumulation of experimental artefacts and overestimation of the number of recovered contacts [93]. However, if the data from multiple studies were processed uniformly, it demonstrated that TAD boundaries in *Drosophila* are more conserved than in mouse. Similar comparative analysis of scHi-C results will further shed light on reproducible chromatin features in individual cells in an unbiased way.

Machine learning has a growing impact on our understanding of biological systems (reviewed in [27, 63]) and 3D genomics [4, 30, 77, 79, 94, 95]. For single-cell chromatin research, imputation and embedding are already driven by neural networks [99] and other advanced machine learning methods will emerge. Importantly, features calling from single-cell data will be improved.

Next, an important direction is improving structural reconstruction approaches. To date, scHi-C structure reconstruction does not account for a specific mechanism of structure formation. Alternative *de novo* modelling assumes the particular mechanism but does not incorporate scHi-C contacts [28, 30]. These approaches can be, in theory, united to open intriguing perspectives. For example, can we simulate loop extrusion [31] that will produce the contact maps similar to those observed in scHi-C? Can we infer the cohesin loading sites in individual cells based on observed contacts? Finally, can we differentiate the cohesin-dependent contacts in single cells from compartmental ones [32] and study them independently?

These challenges are not the only ones that will require computational solutions. An important direction will be the design of new assays, as well as tools for their data processing. For example, currently, restriction enzymes digest chromatin into relatively large restriction fragments, which dictates the strict upper bound for the total number of pairwise contacts recoverable from a single cell. If micrococcal nuclease is used instead, it will allow for up to 15 million contacts of individual nucleosomes in the haploid human genome [1], increasing the theoretical upper bound at least twice.

Joint assays, other than Methyl-3C and sc-me3C, will unravel the interplay of chromatin architecture with other cellular mechanisms. For example, measuring single-cell lamina-associating domains (LADs) alongside scHi-C will shed light on the lamina association of individual TADs. Indeed, bulk TADs do not entirely correspond to either bulk [91] and single-cell LADs [50]. However, it is possible that single-cell TADs are elementary units of interaction with lamina if there is a one-to-one correspondence between TADs and LADs observed in the same cell. Next, measuring chromatin openness and/or transcriptional activity will accelerate the research on interplay and causality between regulation, chromatin folding and gene expression [24]. On the computational side, having more than one type of measurement in single cells is a unique opportunity to develop joint embedding [56] methods, which use both interaction graphs and single-cell features to create meaningful low-dimensionality representation. Also, having several types of measurements will help to develop and benchmark standard scHi-C embedding techniques.

Single-cell RNA-DNA contacts will help distinguish RNA-mediated interactions from the rest and depict the single-cell pattern of regulatory RNA functioning. However, the resolution of bulk RNA-DNA interaction capture techniques is relatively low [3, 33, 61, 81], which will remain a major impediment for single-cell RNA-DNA interactions as well.

Currently, scHi-C requires vast sequencing with relatively low meaningful output (e.g. Ramani *et al.* [76] sequenced over 170 mln reads per dataset on average, only 11% of them resulting in unique contacts). However, studying biological mechanisms of chromatin compaction and regulation frequently requires engineering and targeting of individual regions of the genome limited in size. Thus, it might be beneficial to develop single-cell Hi-C with enrichment for targets. Target enrichment for a genomic region is already well developed for bulk chromosome capture approaches [20, 25, 35]. Adaptation of these approaches for the single-cell level will allow for specific enrichment of single-cell interactions of regulatory regions that might undergo the specific architectural changes in a cell population.

As both wet-lab and computational scHi-C methods improve, it will lead to breakthroughs in understanding biological systems currently restricted by bulk Hi-C. For example, chromatin transitions during mouse embryogenesis were studied by low-input Hi-C [26, 47], which accommodates the limited number

of embryos available but does not distinguish individual cells. Starting from the zygote and up until the gastrulation (stage E7.5), chromatin features gradually emerge. At stage E7.5, the embryo has approximately 15000 cells, some differentiated into progenitors of diverse tissues and organs [90]. Their variability can be recovered only by scHi-C. Indeed, scHi-C demonstrated cell- and allele-specific patterns of chromosomes folding in mouse embryos but only up to a much earlier stage of 64 cells [16, 28, 32]. Given the fact that existing scHi-C assay several tens of thousands of cells [49], a whole-embryo single-cell chromatin structure study is a realistic short-term goal. This opens an intriguing perspective to answer fundamental questions about chromatin dynamics in development. What paths do chromosomes follow in individual nuclei during tissue differentiation and organogenesis? Can we track the lineages of cells based on their chromatin, as we do for single-cell transcription [80]? Finally, what are the rules governing chromatin transitions in individual cells during the development of other species studied by bulk Hi-C, including human [14], *Xenopus tropicalis* [71], Medaka fish [70], *Danio rerio* [45] and *Drosophila melanogaster* [41]?

Next, scHi-C will uncover the diversity of chromatin architecture within cancer cell, contributing to the clonal analysis of solid and liquid tumours currently done with genomic and transcriptomic methods. Finally, single-cell atlases of chromatin architecture for cell types of different organs will expand our knowledge on chromatin structural diversity. Their proper association with single-cell atlases of transcription [38] and chromatin openness [18, 98] will unravel the interplay between epigenetics, chromatin structure and gene expression.

Key Points

- Single-cell Hi-C is a powerful and rapidly developing technology to study chromatin architecture, with computational analysis playing a crucial role in extracting biological meaning from its sparse readouts.
- The number of scHi-C pairwise genomic contacts is limited by the number of genomic fragments in the nucleus requiring special approaches for sparse interactome data analysis, including structure reconstruction, imputation of interactions, aggregation of contacts and feature calling for a single map and embedding, sorting, clustering and pooling for a set of maps.
- We anticipate improvements in scHi-C data quality and computational analysis to lead to the expansion of scHi-C applications, eventually resulting in breakthroughs in our understanding of cell function comparable with those achieved by scRNA-seq and scATAC-seq.

Author contributions statement

A.A.G. wrote the manuscript and analysed the data. M.S.G. conceived the idea, wrote the manuscript and supervised the work.

Funding

This study was supported by grants from Russian Foundation for Basic Research (19-34-90136 to A.G. and 18-29-13011 to M.S.G.).

References

1. Alberts B, Johnson A, Lewis J, *et al.*. Molecular biology of the cell 5th edition. *Garland Science* 2008.
2. Arrastia MV, Jachowicz JW, Ollikainen N, *et al.* A single-cell method to map higher-order 3D genome organization in thousands of individual cells reveals structural heterogeneity in mouse ES cells. *bioRxiv* 2020. Preprint biorxiv:2020.08.11.242081.
3. Bell JC, Jukam D, Teran NA, *et al.* Chromatin-associated RNA sequencing (chAR-seq) maps genome-wide RNA-to-DNA contacts. *Elife* 2018;**7**:e27024.
4. Belokopytova PS, Nuriddinov MA, Mozheiko EA, *et al.* Quantitative prediction of enhancer–promoter interactions. *Genome Res* 2020;**30**(1):72–84.
5. Boettiger A, Murphy S. Advances in chromatin imaging at kilobase-scale resolution. *Trends Genet* 2020;**36**(4):273–87.
6. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet* 2016;**17**(11):661–78.
7. Bonev B, Cohen NM, Szabo Q, *et al.* Multiscale 3D genome rewiring during mouse neural development. *Cell* 2017;**171**(3):557–72.
8. Brandao HB, Gabriele M, Hansen AS. Tracking and interpreting long-range chromatin interactions with super-resolution live-cell imaging. *Curr Opin Cell Biol* 2021;**70**: 18–26.
9. Campbell M, Watanabe T, Nakano K, *et al.* KSHV episomes reveal dynamic chromatin loop formation with domain-specific gene regulation. *Nat Commun* 2018;**9**(1):1–14.
10. Carriere M, Rabadan R. Topological data analysis of single-cell Hi-C contact maps. *Abel Symp* 2020;**15**:147–62.
11. Carstens S, Nilges M, Habeck M. Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS Comput Biol* 2016;**12**(12):e1005292.
12. Carstens S, Nilges M, Habeck M. Bayesian inference of chromatin structure ensembles from population-averaged contact data. *Proc Natl Acad Sci U S A* 2020;**117**(14): 7824–30.
13. Chang L-H, Ghosh S, Papale A, *et al.* A complex CTCF binding code defines TAD boundary structure and function. *bioRxiv* 2021. Preprint biorxiv:2021.04.15.440007.
14. Chen X, Ke Y, Wu K, *et al.* Key role for CTCF in establishing chromatin structure in human embryos. *Nature* 2019;**576**(7786):306–10.
15. Church DM, Schneider VA, Graves T, *et al.* Modernizing reference genome assemblies. *PLoS Biol* 2011;**9**(7):1–5.
16. Collombet S, Ranisavljevic N, Nagano T, *et al.* Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature* 2020;**580**(7801):142–6.
17. Cremer T, Cremer M. Chromosome territories. *Cold Spring Harb Perspect Biol* 2010;**2**(3):1–22.
18. Cusanovich DA, Hill AJ, Aghamirzaie D, *et al.* A single-cell atlas of *in vivo* mammalian chromatin accessibility. *Cell* 2018;**174**(5):1309–24.
19. Dame RT, Rashid FZM, Grainger DC. Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat Rev Genet* 2020;**21**(4):227–42.
20. Davies JOJ, Telenius JM, McGowan SJ, *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat Methods* 2015;**13**(1):74–80.
21. de Wit E. TADs as the caller calls them. *J Mol Biol* 2020; **432**(3):638–42.
22. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting

chromatin interaction data. *Nat Rev Genet* 2013;**14**(6): 390–403.
23. Dekker J, Rippe K, Dekker M, *et al*. Capturing chromosome conformation. *Science* 2002;**295**(5558):1306–11.
24. Delaneau O, Zazhytska M, Borel C, *et al*. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* 2019;**364**(6439):1044–5.
25. Dostie J, Richmond TA, Arnaout RA, *et al*. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res* 2006;**16**(10):1299–309.
26. Du Z, Zheng H, Huang B, *et al*. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* 2017;**547**(7662):232–5.
27. Eraslan G, Avsec Z, Gagneur J, *et al*. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**(7):389–403.
28. Flyamer IM, Gassler J, Imakaev M, *et al*. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* 2017;**544**(7648):110–4.
29. Flyamer IM, Illingworth RS, Bickmore WA. Coolpup.py: versatile pile-up analysis of Hi-C data. *Bioinformatics* 2020;**36**(10):2980–5.
30. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. *Nat Methods* 2020;**17**(11):1111–7.
31. Fudenberg G, Imakaev M, Lu C, *et al*. Formation of chromosomal domains by loop extrusion. *Cell Rep* 2016;**15**(9): 2038–49.
32. Gassler J, Brandao HB, Imakaev M, *et al*. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J* 2017;**36**(24):3600–18.
33. Gavrilov AA, Zharikova AA, Galitsyna AA, *et al*. Studying RNA-DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res* 2020;**48**(12): 6699–714.
34. Goel VY, Hansen AS. The macro and micro of chromosome conformation capture. *Wiley Interdiscip Rev Dev Biol* 2020;e395. https://pubmed.ncbi.nlm.nih.gov/32987449/.
35. Golov AK, Ulianov SV, Luzhin AV, *et al*. C-TALE, a new cost-effective method for targeted enrichment of Hi-C/3C-seq libraries. *Methods* 2020;**170**(2019):48–60.
36. Guarnera E, Tan ZW, Berezovsky IN. Three-dimensional chromatin ensemble reconstruction via stochastic embedding. *Structure* 2021;1–13.
37. Han C, Xie Q, Lin S. Are dropout imputation methods for scRNA-seq effective for scHi-C data? *Brief Bioinform* 2020; **22**:1–12.
38. He S, Wang L-H, Liu Y, *et al*. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol* 2020;**21**(1):1–34.
39. Hirata Y, Oda A, Ohta K, *et al*. Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. *Sci Rep* 2016;**6**:3–8.
40. Horton CA, Alver BH, Park PJ. GiniQC: a measure for quantifying noise in single-cell Hi-C data. *Bioinformatics* 2020;**36**(9):2902–4.
41. Hug CB, Grimaldi AG, Kruse K, *et al*. Chromatin architecture emerges during zygotic genome activation independent of transcription article chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* 2017;**169**(2):216–28.
42. Imakaev M, Fudenberg G, McCord RP, *et al*. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods* 2012;**9**(10): 999–1003.
43. Imakaev MV, Fudenberg G, Mirny LA. Modeling chromosomes: beyond pretty pictures. *FEBS Lett* 2015;**589**(20): 3031–6.
44. John HA, Birnstiel ML, Jones KW. RNA-DNA hybrids at the cytological level. *Nature* 1969;**223**(5206):582–7.
45. Kaaij LJT, van der Weide RH, Ketting RF, *et al*. Systemic loss and gain of chromatin architecture throughout zebrafish development. *Cell Rep* 2018;**24**(1):1–10.
46. Kalhor R, Tjong H, Jayathilaka N, *et al*. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* 2012;**30**(1):90–8.
47. Ke Y, Xu Y, Chen X, *et al*. 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. *Cell* 2017;**170**(2):367–81.
48. Kharchenko PV, Alekseyenko AA, Schwartz YB, *et al*. Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 2011;**471**(7339):480–6.
49. Kim HJ, Yardimci GG, Bonora G, *et al*. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLoS Comput Biol* 2020;**16**(9):1–19.
50. Kind J, Pagie L, De Vries SS, *et al*. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* 2015;**163**(1):134–47.
51. Lajoie BR, Dekker J, Kaplan N. The hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 2015;**72**:65–75.
52. Lando D, Basu S, Stevens TJ, *et al*. Combining fluorescence imaging with Hi-C to study 3D genome architecture of the same single cell. *Nat Protoc* 2018;**13**(5):1034–61.
53. Lando D, Stevens TJ, Basu S, *et al*. Calculation of 3D genome structures for comparison of chromosome conformation capture experiments with microscopy: an evaluation of single-cell Hi-C protocols. *Nucleus* 2018;**9**(1):190–201.
54. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**(4):357–9.
55. Lee DS, Luo C, Zhou J, *et al*. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat Methods* 2019;**16**(10):999–1006.
56. Lerique S, Abitbol JL, Karsai M. Joint embedding of structure and features via graph convolutional networks. *Appl Netw Sci* 2020;**5**(1):1–24.
57. Li G, Liu Y, Zhang Y, *et al*. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat Methods* 2019;**16**(10):991–3.
58. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. Preprint arXiv:1303.3997.
59. Li X, An Z, Zhang Z. Comparison of computational methods for 3D genome analysis at single-cell Hi-C level. *Methods* 2020;**181–182**:52–61.
60. Li X, Zhang Z. DeTOKI identifies and characterizes the dynamics of chromatin topologically associating domains in a single cell. *bioRxiv* 2021. Preprint biorxiv: 2021.02.23.432401.
61. Li X, Zhou B, Chen L, *et al*. GRID-seq reveals the global RNA-chromatin interactome. *Nat Biotechnol* 2017;**35**(10):940–50.
62. Li X, Feng F, Hongxi P, *et al*. A computational toolbox for analyzing single-cell Hi-C data. *PLoS Comput Biol* 2021;**17**(5):e1008978.
63. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015;**16**(6):321–32.
64. Lieberman-Aiden E, Van Berkum NL, Williams L, *et al*. Comprehensive mapping of long-range interactions

reveals folding principles of the human genome. *Science* 2009;**326**(5950):289–93.
65. Liu J, Lin D, Yardlmcl GG, *et al*. Unsupervised embedding of single-cell Hi-C data. *Bioinformatics* 2018;**34**(13):i96–i104.
66. Liu T, Zheng W. ScHiCNorm: a software package to eliminate systematic biases in single-cell Hi-C data. *Bioinformatics* 2018;**34**(6):1046–7.
67. Moon KR, Stanley JS, Burkhardt D, *et al*. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr Opin Syst Biol* 2018;**7**:36–46.
68. Nagano T, Lubling Y, Stevens TJ, *et al*. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 2013;**502**(7469):59–64.
69. Nagano T, Lubling Y, Várnai C, *et al*. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 2017;**547**(7661):61–7.
70. Nakamura R, Motai Y, Kumagai M, *et al*. CTCF looping is established during gastrulation in medaka embryos. *Genome Res* 2021;**31**(6):968–80.
71. Niu L, Shen W, Shi Z, *et al*. Systematic chromatin architecture analysis in xenopus tropicalis reveals conserved three-dimensional folding principles of vertebrate genomes. *bioRxiv* 2020. Preprint biorxiv:2020.04.02.021378.
72. Pardue ML, Gall JG. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. *Proc Natl Acad Sci U S A* 1969;**64**(2):600–4.
73. Payne AC, Chiang ZD, Reginato PL, *et al*. *In situ* genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* 2021;**371**(6532): eaay3446.
74. Perez-Rathke A, Sun Q, Wang B, *et al*. Chromatix: computing the functional landscape of many-body chromatin interactions in transcriptionally active loci from deconvolved single cells. *Genome Biol* 2020;**21**(1):1–17.
75. Polovnikov K, Gorsky A, Nechaev S, *et al*. Non-backtracking walks reveal compartments in sparse chromatin interaction networks. *Scientific Reports* 2020;**10**(1):1–1.
76. Ramani V, Deng X, Qiu R, *et al*. Massively multiplex single-cell Hi-C. *Nat Methods* 2017;**14**(3):263–6.
77. Rozenwald MB, Galitsyna AA, Sapunov GV, *et al*. A machine learning framework for the prediction of chromatin folding in *Drosophila* using epigenetic features. *PeerJ Comput Sci* 2020;**6**:2–21.
78. Samborskaia MD, Galitsyna A, Pletenev I, *et al*. Cumulative contact frequency of a chromatin region is an intrinsic property linked to its function. *PeerJ* 2020;**8**:1–15.
79. Schwessinger R, Gosden M, Downes D, *et al*. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* 2020;**17**(11):1118–24.
80. Soldatov R, Kaucka M, Kastriti ME, *et al*. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* 2019;**364**(6444):eaas9536.
81. Sridhar B, Rivas-Astroza M, Nguyen TC, *et al*. Systematic mapping of RNA-chromatin interactions *in vivo*. *Curr Biol* 2017;**27**(4):602–9.
82. Stevens TJ, Lando D, Basu S, *et al*. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 2017;**544**(7648):59–64.
83. Sun Q, Perez-Rathke A, Czajkowsky DM, *et al*. High-resolution single-cell 3D-models of chromatin ensembles during *Drosophila* embryogenesis. *Nat Commun* 2021;**12**(1): 1–12.
84. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv* 2019;**5**(4):eaaw1668.
85. Tan L, Ma W, Wu H, *et al*. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell* 2021;**184**(3):741–58.
86. Tan L, Xing D, Chang CH, *et al*. Three-dimensional genome structures of single diploid human cells. *Science* 2018;**361**(6405):924–8.
87. Tan L, Xing D, Daley N, *et al*. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nat Struct Mol Biol* 2019;**26**(4): 297–307.
88. Tavares-Cadete F, Norouzi D, Dekker B, *et al*. Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nat Struct Mol Biol* 2020;**27**(12): 1105–14.
89. Tjong H, Li W, Kalhor R, *et al*. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci U S A* 2016;**113**(12): E1663–72.
90. Tzouanacou E, Wegener A, Wymeersch FJ, *et al*. Redefining the progression of lineage segregations during mammalian embryogenesis by clonal analysis. *Dev Cell* 2009;**17**(3): 365–76.
91. Ulahannan N, Pendleton M, Deshpande A, *et al*. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. *bioRxiv* 2019. Preprint biorxiv:833590.
92. Ulianov SV, Tachibana-Konwalski K, Razin SV. Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. *Bioessays* 2017;**39**(10):1–8.
93. Ulianov SV, Zakharova VV, Galitsyna AA, *et al*. Order and stochasticity in the folding of individual drosophila genomes. *Nat Commun* 2021;**12**(1):1–17.
94. Vanhaeren T, Divina F, García-Torres M, *et al*. A comparative study of supervised machine learning algorithms for the prediction of long-range chromatin interactions. *Genes* 2020;**11**(9):1–17.
95. Whalen S, Truty RM, Pollard KS. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 2016;**48**(5):488–96.
96. Wolff J, Abdennur N, Backofen R, *et al*. Scool: a new data storage format for single-cell Hi-C data. *Bioinformatics* 2020;**37**(9):1337.
97. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 2011;**43**(11): 1059–65.
98. Zhang K, Hocker JD, Miller M, *et al*. A cell atlas of chromatin accessibility across 25 adult human tissues. *bioRxiv* 2021. Preprint biorxiv:2021.02.17.431699.
99. Zhang R, Zhou T, Ma J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *bioRxiv* 2020. Preprint biorxiv:2020.12.13.422537.
100. Zhou J, Ma J, Chen Y, *et al*. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proc Natl Acad Sci U S A* 2019;**116**(28):14011–8.
101. Zhou S, Jiang W, Zhao Y, *et al*. Single-cell three-dimensional genome structures of rice gametes and unicellular zygotes. *Nat Plants* 2019;**5**(8):795–800.
102. Zhou T, Zhang R, Ma J. The 3D genome structure of single cells. *Annu Rev Biomed Data Sci* 2021;**4**:21–41.
103. Zhu H, Zheng W. SCL: a lattice-based approach to infer 3D chromosome structures from single-cell Hi-C data. *Bioinformatics* 2019;**35**(20):3981–8.