HHS Public Access

Author manuscript

Nat Rev Genet. Author manuscript; available in PMC 2024 May 25.

Published in final edited form as:

Nat Rev Genet. 2024 February ; 25(2): 123–141. doi:10.1038/s41576-023-00638-1.

Computational methods for analyzing multiscale 3D genome organization

Yang Zhang[1], **Lorenzo Boninsegna**[2], **Muyu Yang**[1], **Tom Misteli**[3,†], **Frank Alber**[2,†], **Jian Ma**[1,†]

[1.] Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

[2.] Department of Microbiology, Immunology and Molecular Genetics and Institute for Quantitative and Computational Biosciences, University of California Los Angeles, Los Angeles, CA, USA

[3.] Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA

Abstract

Recent progress in whole-genome mapping and imaging technologies has enabled the characterization of the spatial organization and folding of the genome in the nucleus. In parallel, advanced computational methods have been developed to leverage these mapping data to reveal multiscale 3D genome features and provide a more complete view of genome structure and its connections to genome functions such as transcription and DNA replication. Here, we discuss how recently developed computational tools, including machine learning-based methods and integrative structure modelling frameworks, have led to a systematic, multiscale delineation of the connections among different scales of 3D genome organization, genomic and epigenomic features, functional nuclear components, and genome function. However, moving forwards, new approaches that more comprehensively integrate a wide variety of genomic and imaging datasets will be needed to uncover the functional role of 3D genome structure in defining cellular phenotypes in health and disease.

ToC blurb

In this Review, Zhang et al. discuss how recent advances in computational methods are helping to reveal the multiscale features involved in genome folding within the nucleus and how the resulting 3D genome organization relates to genome function.

Introduction

Nuclear genomes contain most of the genetic information needed to define the phenotype of a cell, tissue, and organism, and they are intricately organized in 3D space[1]. This 3D genome architecture is crucial for genome functions such as transcription. Genetic, molecular,

[†] jianma@cs.cmu.edu (J.M.), falber@g.ucla.edu (F.A.), mistelit@mail.nih.gov (T.M.).

Author contributions

The authors contributed equally to all aspects of this manuscript.

Competing interests

The authors declare no competing interests.

imaging, biochemical, and computational approaches over the past two decades have led to the identification of some of the basic principles that shape genome architecture. Chromatin fibers can form loops which compact the genome and bring control elements, such as distal enhancers, close to target genes for regulation[2,3]. These loops also form larger structures called topologically associating domains (TADs)[4–6], which are often established by cohesin-mediated loop extrusion **[G]**[7]. Multiple chromatin domains **[G]** further associate with each other via physical homotypic interactions **[G]** to form higher-order compartments that segregate chromosomes, referred to as compartments A and B. These compartments largely represent transcriptionally active, decondensed euchromatin and transcriptionally repressed, compacted heterochromatin, respectively[8,9]. Ultimately, the interphase chromosomes occupy a restricted volume within the nuclear space, known as chromosome territories (Box 1, Fig 1). While these multiscale genome organization features are found in most cell types and organisms, they are dynamic and show considerable single-cell heterogeneity[10–16].

In the past decades, methods for understanding genome organization have advanced greatly. Early studies often involved correlative microscopy studies of chromatin organization of single loci with limited resolution, but recent breakthroughs in biochemical approaches have enabled the mapping of genome-wide chromatin-chromatin interaction contact frequency **[G]**, through methods such as Hi-C[8,17] and Micro-C[18–20]. Notably, in the context of these mapping methods, the term ‘interaction’ is often used loosely to represent physical proximity rather than direct physical interaction and frequently refers to population averages rather than single cell behavior. The variation of 3D genome organization, including the dynamic folding patterns, can be revealed by capturing chromatin-chromatin interactions in individual cells using single-cell Hi-C techniques[21], but these methods suffer from sparse data. Complementary information on cell-to-cell heterogeneity of chromatin organization has also been generated from imaging methods, which offer sub-megabase scale resolution and direct spatial measurement of the chromatin fiber in intact single cells[12,22], and by using computational frameworks that deconvolve ensemble data into single-cell structures[23]. A combination of population-based and more affordable single-cell methods is bound to provide the most accurate description of 3D genome organization.

Computational methods are critical tools for addressing pressing challenges in the study of genome organization. Essential computational tools have a range of applications, including identifying multiscale 3D genome structural features using genome-wide mapping data, modeling local dynamic properties of chromatin and analyzing single-cell imaging data. Importantly, computational methods are crucial in multi-omics approaches for connecting multiple aspects of genome organization (such as DNA sequence, epigenomic features and 3D genome organization) and linking these features to genome functions, particularly gene activity and transcription. While methods for analyzing individual features are widely available, the integration of multiple types of data, especially sequencing-based mapping data with data generated by imaging approaches, remains a formidable challenge.

In this Review, we explore the latest computational approaches for studying 3D genome organization and highlight opportunities for creating integrated multi-omic models of genome structure and function. While many methods have been developed to handle specific types of datasets, such as those obtained through sequencing-based or imaging assays,

or those focused on single-cell or population-based data, there is a growing interest in integrative methods. These methods can combine different modalities, giving us a more comprehensive view of genome organization. We also discuss limitations of current methods and outline future needs to fully harness the tremendous potential of computational methods in elucidating genome organization in space and time.

3D features from genomic mapping data

Much of our understanding of multiscale 3D genome features comes from patterns in population-based genomic mapping data, especially Hi-C data, generated through a series of computational data analysis steps. The Hi-C analysis workflow starts with mapping paired-end sequencing reads to the reference genome, filtering, binning, and summarizing the read count into a symmetric contact frequency map **[G]** (Fig. 2a, for review see[24,25]). Normalization schemes then correct for systematic biases[25,26], either by taking bias factors as input[27] or using matrix-balancing strategies[8,28–30]. The goal of Hi-C data analysis is to identify meaningful 3D genome structural features from the normalized 2D contact frequency map. The identified features can be grouped into three categories based on their length scales: from largest to smallest, these are compartments (and sub-compartments), TADs (and sub-TADs), and loops.

Compartments and sub-compartments.

At the largest length scale, Hi-C analysis confirmed early cytological observations that chromosomes tend to occupy distinct territories in the nuclear space[31] and that chromatin segregates into A and B compartments that largely correspond to euchromatin and heterochromatin[8]. Chromatin in the same type of compartment tends to have higher contact frequencies than chromatin in different compartments, which creates characteristic checkerboard or plaid patterns in megabase-scale 2D contact frequency maps (Fig. 2b)[8]. From an algorithmic perspective, the goal of identifying these coarse-grain patterns in the 2D contact frequency maps is to classify genomic bins based on the similarity of their chromatin interactions (Fig. 2b) (for example, via eigenvector analysis). Subsequent high-coverage Hi-C experiments further refined this view, suggesting that finer-grained chromatin structures exist within A or B compartments (that is, sub-compartments, Fig. 2b), each characterized by a unique pattern of histone modifications and replication timing[8,30]. In line with this observation, predictive models have been developed to identify cell-type-specific sub-compartments based on epigenomic data[32]. However, traditional clustering-based approaches, such as the Gaussian hidden Markov model[30], requires billions of read pairs to identify sub-compartments, restricting their application to a limited number of datasets. To address this problem, recent methods have reduced high-dimensional 2D contact frequency maps into low-dimensional representations, making it easier to find patterns[33–35]. In addition, ligation-free methods, including Tyramide Signal Amplification sequencing (TSA-seq) and DNA adenine methyltransferase identification (DamID), provide information about the relative positioning of chromatin with respect to different nuclear bodies (Fig. 1c). Recent integrative modelling approaches, such as Spatial Position Inference of the Nuclear genome (SPIN), aim to reconstruct the 3D compartmentalization of chromatin by reconciling Hi-C, TSA-seq, and DamID data into a unified probabilistic model[36].

TADs and subTADs.

TADs and subTADs are smaller structural units of chromosomes that were first identified in population-level Hi-C contact frequency maps[37]. In Hi-C maps, TADs are represented as blocks along the diagonal with sizes ranging from ~100kb-2Mb, and they indicate increased interactions amongst chromatin within the domain compared to the upstream and downstream regions. Early TAD-calling methods, such as the directionality index (DI)[4] and insulation score (IS)[38,39], partitioned the genome into bins and quantified the skewness of chromatin interactions between a sliding window and flanking regions. The TAD boundaries are then determined by segmenting genomes based on DI or finding local minima of IS (Fig. 2c). Several recently developed methods aim to reconstruct the hierarchy of TADs, including subTADs and nested TADs[37,40,41], which exhibit weaker skewness than TADs, show a nested structure within TAD boundaries (Fig. 2c), and are often more cell-type specific[37]. Early methods reconstruct the nested structure of TADs by recursively testing the skewness of putative multiscale domain boundaries[42,43]. Other methods, such as 3DNetMod[44], consider calling nested TADs in a manner similar to that used to identify dense subgraphs, whereby nodes are genomic bins and edges represent contact frequencies. However, systematic evaluation studies found poor concordance among existing TAD-calling or subTAD-calling methods, particularly for determining the number and size of TADs, whereas TAD boundaries are more consistently defined among different TAD calling methods[45,46].

Chromatin loops and stripes.

At the smallest length scale, chromatin loops are pairs of genomic loci that have higher contact frequency than expected based on random interactions and their genomic distance. Two types of existing loop-calling methods differ from each other by whether the background expectation is based on a genome-wide (global) model or local context (Fig. 2d). Global background methods use the contact frequency map to fit a model relating contact frequency with 1D sequence distance. Examples of this approach are Fit-Hi-C[47] and HiC-DC[48]/HiC-DC+[49]. Local context-based methods, such as HiCCUPS[30], SIP[50], and Mustache[51], use filters or post-processing steps to distinguish direct interactions from bystander interactions **[G]**. Additionally, there have been efforts to use supervised machine learning to predict chromatin loops, such as LooPbit[52] and Peakachu[53]. Besides chromatin loops, computational methods have also been developed to detect other structure features such as stripes[54,55] (also called lines or flames), which reflect statistically increased contact frequency of a loop anchor with an extensive stretch of consecutive chromatin across a domain, likely caused by one-sided loop extrusion[56]. However, the agreement of loop identifications among methods is less consistent than it is for determining TAD boundaries[46,57].

Predicting features with machine learning

A fundamental question in the field of 3D genome organization is how chromatin interactions combine to generate the emerging higher-order genome structure. Because machine learning, particularly supervised learning, is a powerful tool for sample classification and outcome prediction based on input features, it offers a unique opportunity

to predict 3D chromatin interactions and higher order structure from population-level data. Although building a model to infer 3D genome data does not require a complete understanding of the underlying biological mechanism, a successful predictive model would potentially provide insights into how DNA sequences and epigenetic profiles contribute to 3D genome folding[58,59]. Additionally, the capability to reliably predict chromatin interactions has the potential to identify regulatory elements and genetic variants that affect chromatin folding.

Overview of machine learning approaches for predicting 3D chromatin interactions.

Among various machine learning methods, ensemble learning approaches (that is, those that combine predictions from multiple models) such as random forest **[G]** and gradient boosting **[G]**, are commonly used owing to their robustness to overfitting and their ability to capture non-linear relationships between features that encode the dependencies from 3D genome features and functions[60]. However, these methods are limited in their ability to quantitatively predict chromatin interactions from genome sequences, for which sequence context information has to be aggregated into megabase scales to reflect structure features in contact frequency maps. Deep learning models have emerged as effective approaches for predicting large-scale genome organization involving larger sequence context typically at megabase-scale. Convolutional neural networks (CNNs) **[G]**[61] use convolutional kernels **[G]** to scan and incorporate local information up to ~100kb, while recurrent neural networks (RNNs) **[G]**[61,62] are effective in modelling long-dependence in sequence data. Graph neural networks (GNNs) **[G]**[63,64] use the graph structure of data to learn the representation by incorporating information from the neighborhood. However, it is critical to note that a well-performing predictive model does not necessarily equate to understanding the chromatin folding mechanism. Effectively interpreting the ‘black box’ of machine learning remains an active area of research. The machine learning approaches used to predict chromatin structures can be grouped into three major categories: predicting pairwise chromatin interaction, predicting genome-wide contact frequency maps, and multiway chromatin interaction prediction.

Predicting pairwise chromatin interactions.

The goal of predicting pairwise chromatin interactions is to determine whether the propensity of two genomic loci to form a loop is related to their DNA sequence, relative genomic distance, and/or epigenetic marks (Fig. 3a). This task can be formulated as a supervised learning problem in which chromatin loops[65,66] or enhancer-promoter interactions[67–69] observed from population-level experimental data are used as the ground truth for training. Early predictive models use handcrafted features such as transcription factor (TF) binding site composition and the average profiles of epigenomic marks[67,70]. However, these methods have limitations owing to the quality and relevance of feature selection and the need for prior knowledge of the potentially predictive features. In recent years, deep learning-based methods have gained popularity in the genomics field, as they can automatically model sequence dependencies from multi-scale genome architecture measured by Hi-C, reducing the need for manual feature engineering[59]. Alternatively, some methods use the methods initially developed for natural language processing by encoding DNA sequences into vector representations **[G]** (also known as embeddings) in

a semantic space[65,68,71]. While a combination of sequence and epigenomic marks provides strong predictive power for chromatin interactions in matched cell types, deep learning models using DNA sequences alone can still perform well[69,72]. However, these models struggle to predict cell-type specific chromatin interactions[72]. Some methods solve this issue by integrating sequence information with chromatin accessibility data[73], which has been generated for many cell types using Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) and DNase I hypersensitive sites sequencing (DNase-seq). All these methods, however, have limited interpretability, which hinders the accurate identification of specific features that drive the formation of chromatin interactions. It is also crucial to consider that at the single-molecule level, loop formation is short-lived and dynamic, as demonstrated by live-cell imaging[74]. Thus, the process by which features predicted by machine learning methods on average observables from cell population data can be linked to actual biological processes remains unclear.

Predicting genome-wide contact frequency maps.

The methods discussed previously focus on predicting specific pairwise interactions and do not consider genome-wide chromatin contacts. This leads to the intriguing question of whether we can predict genome-wide chromatin interactions directly from 1D genomic signals of any paired loci. Recently, several methods have demonstrated that this is computationally feasible through the use of deep learning models on available Hi-C contact frequency maps that incorporate megabase-scale sequence context. These methods can be grouped into three types (Fig. 3b). The first group, including Akita[75] and Orca[76], predicts chromatin interactions between genomic bins solely from DNA sequences with a trained model using Hi-C data from a specific cell type. Akita predicts contact frequency maps for any regions located within 1Mb sequence distance, whereas Orca achieves multiscale prediction from kilobase to chromosome scale by using a cascading encoder and decoder module. The second group, represented by DeepC[77], uses a transfer learning approach consisting of two steps: It first encodes sequences using a CNN model pre-trained for epigenomic signal prediction across multiple cell types and then generates contact frequency maps using a stack of dilated CNNs **[G]** with the extracted DNA sequence features. However, both of these methods face the problem of whether the model can be generalized to make *de novo* predictions in a different cell type or species as well as in time course events such as differentiation. To address this issue, the last group of methods aims to predict cell type-specific Hi-C contact frequency maps by incorporating epigenomic marks, allowing for predictions in cell types for which Hi-C data are not available[78,79]. As methodological advancements continue, there is high potential, and an urgent need, for new methods to not only accurately predict chromatin interactions but also create deeper insights into the mechanisms of chromatin folding. Techniques aimed at improving the interpretability of machine learning models could potentially synergize with *in silico* mutagenesis and high-throughput perturbation experiments, leading to identification of genetic variants that result in reconfiguration of the 3D genome structure[75–78].

Predicting multiway chromatin interactions.

The prediction of multiway chromatin interactions is challenging. Representing higher-order chromatin interactions as hypergraphs **[G]** extends the modelling of pairwise

interactions, whereby hyperedges **[G]** can connect any number of genomic loci. However, the development of computational methods for denoising or predicting multiway interactions is still in its early stages[80–82]. The number of potential hyperedges increases exponentially as the order of interactions increases, making the prediction of higher-order interactions difficult. The algorithm MATCHA[81] aims to overcome these challenges by utilizing the hypergraph representation learning framework to analyze multiway interactions. It can denoise the observed multiway interactions data as well as predict the occurrence of multiway interactions. However, effectively predicting cell type-specific multiway interactions remains an unsolved problem.

Polymer models for physical processes

The mechanisms that lead to the observable features of chromatin and genome folding are not well understood. Computer simulations also have the capability to test potential folding mechanisms. Successful computational models for genome modelling identify appropriate degrees of freedom needed to describe the genome and their dynamic interactions at a given structural resolution. These inferences are trained and validated using independent experimental data[83,84]. An effective model is expected to generate genome structures that both accurately predict and explain the mechanistic underpinnings of experimental observations.

Mechanistic approaches to genome structure modelling assume that one or more known physical mechanisms, such as phase separation **[G]** or loop extrusion, drive the chromatin folding process. In these approaches, the chromatin fiber is represented as a polymer chain of monomers that can be divided into classes with different physicochemical properties. Interaction terms model the physical processes believed to drive folding, and model parameters are optimized to maximize agreement with experimental data. Overall, a model can be validated if a given folding hypothesis can, in principle, explain experimental observations; however, it cannot rule out other potential mechanisms that may also explain the data equally well. Simulations of excluded volume and gene tethering constraints[85,86] as well as loop extrusion and phase separation have provided insights into the mechanistic underpinnings of chromatin organization and predicted structural changes based on biological perturbations, such as sequence mutations[87,88] or protein knockdown experiments[89–91]. Additionally, unbiased comparisons of different models on benchmark datasets can allow a uniform and comprehensive evaluation of their predictive power[92].

Loop extrusion.

Active loop extrusion, whereby cohesin complexes slide along chromatin fibers in different directions to form loops, was proposed more than 20 years ago to play a role in chromatin condensation[93], but was only recently put to the test in simulations[7,93–96] and eventually observed *in vitro* in single-molecule experiments[97,98] (Fig. 4a). Computer simulations of the loop extrusion process have explained a variety of observed phenomena[99,100], including mitotic chromosome compaction and segregation[93,101–103], meiotic chromosome organization[104], interphase chromatin domains and loops[7,94], and functional interactions, such as V(D)J recombination of the immunoglobulin heavy chain[105]. The model's success

lies in the ability to predict the effects of loop extrusion perturbations[106], particularly by protein knockdown experiments[89–91]. However, loop extrusion does not explain, and even acts antagonistically to, large-scale chromatin compartmentalization[107], suggesting that other mechanisms may also be involved.

Phase separation.

Phase separation has been proposed to contribute to genome compartmentalization (such as heterochromatin compartments), transcriptional condensates, and the formation of nuclear bodies[108–111] (Fig. 4b). In phase separation, proteins and other biomolecules demix and form distinct condensates via weak, multivalent interactions. In modelling approaches, chromatin fiber segments can be partitioned into different classes with preferential homotypic interactions leading to micro-phase separation that mimics chromatin compartmentalization. Many approaches have incorporated similar models of phase separation to reproduce experimental observations[112–124]. Critical to the predictivity of such models is the correct assignment of polymer loci to different interaction classes, which usually exploits knowledge of histone modifications for each chromatin region[87,114,120,125,126]. Phase separation models can be extended by adding other folding mechanisms, such as loop extrusion, which can act antagonistically to phase separation at the level of short-range chromatin interactions[107,116,127], as well as incorporating centromere interactions[128], or chromatin interactions with nuclear bodies[129,130].

However, mechanistic approaches to modelling chromatin folding have limitations. They are not easily transferable between genome types and across different scales, and often rely heavily on knowledge of physical processes that are not fully understood. Also, multiple unknown mechanistic processes are likely to coexist and simultaneously shape chromatin architecture on a global scale and disentangling the relative importance of those remains challenging.

Data-driven structure modelling

Data-driven approaches in 3D genome structure modelling do not require prior knowledge of folding mechanisms or pre-defined chromatin classes. Instead, they use experimental data, such as contact frequencies from Hi-C or information from imaging experiments, to reconstruct 3D genome structures that statistically recapitulate that experimental data[131,132] (Fig. 5a). Structural information is modeled through spatial proximity terms between chromatin loci or nuclear landmarks: for example, a harmonic potential that constrains the positions of two chromatin regions is a viable interpretation of a single cell chromatin contact probed in experiments. Chromatin structures are then generated by minimizing one or more scoring functions that describe the discrepancy between the model and the experimental data. These structures are the result of the interplay of a multitude of data-driven interactions rather than of a single physical folding process[132,133]. These 3D structures can then be examined to derive structure-function correlations and make quantitative predictions. There are two main data-driven approaches: resampling-based and deconvolution-based methods, which differ in their interpretation of experimental data, scoring functions, and sampling strategies (Fig. 5b).

Resampling approaches.

Resampling approaches distill all the information contained in the data into a solitary scoring function, which is then sampled in independent optimizations to generate multiple structures that are all compatible with the experimental data[134–139] (Fig. 5b). However, the use of a single scoring function can be flawed when applied to ensemble data from highly variable objects such as entire genomes, since mutually conflicting information will inevitably produce structures that cannot fully reconcile all the data. In this scenario, the ensemble of structures cannot provide a realistic representation of structures and their cell-to-cell variability. This problem has been addressed in some approaches by considering only the most significant subsets of interactions, assuming the presence of a single dominant structural state, for instance when modelling the folding patterns of individual gene loci or some chromosomal regions[134–136,138,140–143]. Because of these limitations in handling highly variable objects, resampling is most suited for modelling chromosome structures from single-cell Hi-C maps[13,21,144–147] or modelling chromatin structures to augment the coverage of single cell imaging data[137].

Deconvolution approaches.

Deconvolution methods are designed to resolve cell-to-cell variability by allowing individual structures to account for only a subset of the experimental data, so that contradicting information can be resolved in different structures[132,133] (Fig. 5b). These methods divide ensemble data into individual subsets, each representing a single structure in a population. This means that the population of structures as a whole, rather than an individual model, is statistically consistent with the overall ensemble data. Some methods encode Hi-C data into tailored potential energy functions between pairs of chromatin regions, producing an ensemble of structures through molecular dynamics **[G]** or Monte Carlo **[G]** simulations. This helps reproduce experimental Hi-C contact probabilities, while also capturing their cell-to-cell variability[113,148,149]. However, the computing demand could become intractable with increasing complexity of the underlying structures. This is particularly true for methods that explore the conformational landscape by simulating the conformational changes in one genome model over time with molecular dynamics simulations, which could require prohibitive amounts of sampling for whole genome models. Another approach, called population-based modelling[131,150], deconvolves ensemble data into a large population of single-cell structures by finding optimal allocations of data to all structures in an iterative fashion. This is achieved by solving a maximum likelihood estimation problem that uses structural information of the models during the optimization process. As a result, each individual structure in the population is described by a unique scoring function, which expresses only a subset of all data and describes the deviation of the model from experimental data. The collective optimization of all scoring functions generates a population of single cell genome structures recapitulating all the experimental data. As a result, conflicting data is assigned to different structural models, which capture the cell-to-cell variability of genome structures observed between individual cells[150–152]. This approach allows for a rigorous and unbiased sampling of the conformational space[16]. Such an approach was recently used[23] to simultaneously deconvolve data from four different sources (Hi-C, lamin DamID, Split-Pool Recognition of Interactions by Tag

Extension (SPRITE), and High-throughput Imaging Position Mapping Fluorescence In Situ Hybridization (HIPMap FISH imaging)) into a population of whole diploid genome structures that not only recapitulate the input data but also accurately predict orthogonal observables from a variety of imaging and genomics experiments. These structures are highly predictive for nuclear locations of chromatin and nuclear bodies, chromatin folding patterns, spatial segregation of functionally related chromatin domains and reproduce chromosome conformations from single-cell chromosome tracing experiments, as discussed in a recent preprint[152]. Thus, these structures can provide a characterization of the nuclear microenvironment of individual genes in single cell nuclei (Fig. 5c), which can be related to their gene expression or other functional properties. For example, increased gene expression has been linked to specific nuclear microenvironments and the cell-to-cell variability of a gene's nuclear microenvironment predicted by the models has been linked to observed heterogeneity in gene expression from single-cell RNA-seq experiments, as shown in[152] and a recent preprint[153].

Overall, data-driven approaches are well-suited to uncover relationships between genome structures and function and can provide insights into the differences in nuclear organization across different cell types. However, they cannot provide information on the physical processes that drive chromatin folding.

Single-cell 3D genome analysis

Extracting 3D genome features from single cells is challenging for both single-cell Hi-C (scHi-C) data and multiplexed imaging data. For scHi-C data, the computational difficulties are caused by the sparsity and low signal-to-noise of the data (Fig. 6a). Despite yielding 20k to 80k interactions per cell, which is similar to other single-cell assays such as single-cell RNA-seq (scRNA-seq) or single-cell ATAC-seq (scATAC-seq), the fraction of missing interactions in scHi-C data is much higher (95% vs 30%)[21,154]. This is attributed to its two-dimensional structure and the technology limitations of the current scHi-C methods. Therefore, imputing missing interactions is essential for a comprehensive understanding of chromatin organization. By contrast, multiplexed imaging techniques are less affected by missing data and are capable of portraying chromatin loci in 3D, but their limited imaging resolution makes it difficult to precisely determine the spatial location of genomic loci from noisy fluorescent signals. Despite these challenges, both modalities offer a unique opportunity to study cell-to-cell variability of 3D genome features, especially in complex tissues where conventional bulk Hi-C experiments cannot distinguish cell type-specific 3D genome features. To fully unleash the potential of single-cell 3D genome mapping data, new computational methods are needed to alleviate experimental biases, identify multiscale 3D genome features in single cells, and advance our understanding of the connections between 3D genome features and cell type-specific gene regulation. Below, we discuss computational challenges of analyzing multiplexed imaging data and two important steps for scHi-C data analysis: dimensionality reduction and imputation (Fig. 6).

Chromatin tracing.

Chromatin tracing involves labeling DNA with fluorescent probes and visualizing the spatial arrangement of these DNA sequences through either super-resolution microscopy or multiplexed imaging (Box 1). However, assigning fluorescent signals to a probe with very high accuracy can be challenging owing to inefficient probe detection in individual cells and signal background, often resulting in ambiguous patterns and variable data representation amongst probes. Furthermore, variability in signal detection may also reflect spatial and/or temporal biological variations, such as copy number variation and DNA replication timing. Early computational methods relied heavily on the assumptions of a consistent number of fluorescent spots in each cell and often utilized heuristics to assign spots to the reconstructed chromatin fiber, for instance, choosing the spot with the strongest fluorescent intensity. A recent method named 'Jie'[155] addresses these issues by employing a polymer model to rank the potential of all the possible paths of fluorescent spots, with the final path selected based on the highest likelihood. Integrating this method with existing ones allows for the separation of sister chromatids and prediction of the cell's ploidy.

Dimensionality reduction.

Dimensionality reduction aims to simplify the information contained in scHi-C data by representing the chromatin interactions of each cell in a lower-dimensional feature space (Fig. 6b). This reduction of dimensionality results in a cell embedding, which can be further visualized in two dimensions using t-distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UMAP)[156,157]. Early work in scHi-C analysis adapted existing methods for other single-cell assays such as principal component analysis **[G]** (PCA)[158] and latent Dirichlet allocation **[G]** (LDA)[159] to reduce the dimensions of the scHi-C contact frequency maps, and found that the learned cell embeddings could differentiate between various cell lines[159]. The HiCRep metric, which originally aimed to measure similarity between bulk Hi-C samples, has also been used to calculate the similarity scores between scHi-C contact matrices, followed by multidimensional scaling (MDS) to capture structural patterns during cell-cycle states[159,160]. While these methods perform well in some high-coverage datasets, generating accurate embeddings from tissue datasets with complex cell types and low coverage can be a difficult task owing to the high fraction of missing interactions. More recent approaches, such as Higashi[146] and Fast-Higashi[161], solve this by combining embedding and imputation into a unified framework (see below). Another approach, demonstrated by scVI-3D[162], uses variational inference **[G]** to tackle this problem.

Data imputation and integrative modelling.

To further improve the effectiveness of embedding, imputation, and denoising strategies have been developed to address the sparsity of scHi-C data and enhance the signal-to-noise ratio[147,163] (Fig. 6b). The idea behind imputation is that missing interactions can be inferred from adjacent genomic bins because they tend to have similar contact frequency profiles. Methods such as scHiCluster[164] use linear convolution and random walk with restart **[G]** to impute missing interactions. Higashi[146] is an integrative framework that combines embedding and imputation in an end-to-end framework through the formulation

of hypergraph representation learning. In Higashi, the whole single-cell 3D genome data is represented as a hypergraph in which individual cells and genomic loci are modeled as two types of nodes in the same embedding space, with hyperedges connecting them to represent observed contact interactions in each cell. Because the whole dataset is trained together, Higashi enables cells with a similar overall 3D chromatin organization to share the latent representation of interactions to further improve imputation. A variant of Higashi, called Fast-Higashi[161] is based on tensor decomposition **[G]** and scalable to large datasets and can jointly infer chromatin meta-interactions with cell embeddings. Both Higashi and Fast-Higashi have shown that cell embeddings alone have sufficient information to distinguish cell types in complex tissues such as the brain.

In addition to dimensionality reduction and imputation, several important aspects of data processing and analysis of scHi-C data need more future exploration (Fig. 6c). Computational methods to identify 3D genome features[146,164–168], including A/B compartments, sub-compartments, TAD-like structures, and chromatin loops, are still limited and lack benchmark datasets for evaluation owing to high intrinsic cell-to-cell variation. There are only a few methods available that can effectively reconstruct 3D models from scHi-C data while also taking into account the different haplotypes[145]. Moreover, methods that can jointly model different types of single-cell epigenomic data together with scHi-C data should offer a more comprehensive view of 3D genome structure in single cells (see reviews[147,163]).

Connecting structure and function

How structural features of the genome relate to its function remains the central question in the field. Long-range chromatin interactions have long been believed to play crucial roles in various aspects of biological function, such as transcription regulation, DNA replication timing, and DNA damage repair[169]. While most conclusions regarding structure-function relationships are based on correlations, disruption of chromatin interactions can cause abnormal gene expression and lead to disease[170,171]. Predictive models, which aim to estimate the activities of biological processes from DNA sequences and epigenomic marks, have the potential to identify crucial features involved in these processes and hold promise for prioritizing non-coding genetic variants responsible for human diseases and traits. However, most current methods can only capture local dependencies of sequences up to 100kb away from a gene and do not incorporate the spatial connectivity of regulatory sequences in their model architecture[172].

To overcome this limitation, methods[173] have emerged that use graph-based neural network models to integratively model 2D chromatin interactions with 1D functional epigenomic signatures and DNA sequences in order to reveal the associations between chromatin folding and gene regulation. In these models, chromatin interactions are represented as a graph with nodes representing genomic bins and edges indicating potential regulatory relationships inferred from the chromatin interaction mapping data, such as interactions from Hi-C data. The node representations are initially assigned the sequence or epigenomic features and iteratively updated by incorporating the neighborhood information in the interaction graph, enabling the prediction of gene expression levels from both 1D and

3D features. Interpretable machine learning techniques[174,175] can be applied to reveal the important genomic bins (nodes) and chromatin interactions (edges) contributing to the predictive power of the model. Additionally, integrative analysis of genetic variants in expression quantitative trait loci (eQTL) and experiments that perturb the spatial positioning or activity of selected genomic loci[176] may validate important loci predicted by the model and interpret the functions of *cis*-determinants in gene regulation. While graph-based neural network models exhibit superior performance compared to models that do not account for long-range interactions, it is crucial to recognize the inherent stochasticity of gene regulation. Consequently, using a fixed chromatin network to model gene regulation might not precisely capture the intricacies of the underlying biological processes. In fact, recent studies utilizing live-cell imaging have demonstrated that chromatin loops are generally relatively short-lived (~20 min) and infrequent events, further highlighting the dynamic nature of gene regulation[74].

In addition to transcriptional regulation, the spatial organization of the genome has a crucial role in various other cellular processes, including DNA replication[177] and cell division[177,178]. However, the mechanisms connecting chromatin structures and these events remain unclear. A recent preprint[179] on replication timing reported the use of a deep learning model to predict replication timing profiles by extracting features from DNA sequences alone and successfully prioritized DNA elements that overlap with previously validated loci known to modulate genome-wide replication timing[180]. Integrating genomic features using a graph-based neural network to predict the formation of DNA double-strand breaks revealed specific subgraph patterns related to loop extrusion process[181]. Applying network theory to model the chromatin interactions has also been found to be effective to identify crucial network topological patterns in gene regulation[182] and cancer development[183].

Graph-based models have also been developed to integrate genomic mapping data for a more comprehensive view of large-scale genome structures. An example of this is the inference of genome compartmentalization, which can be achieved by extending the framework of the hidden Markov model into the hidden Markov random field. This extension allows for the spatial dependency among observed DamID[184] and TSA-seq[185] signals on genomic bins to be accounted for, as demonstrated in SPIN[36]. Segway-GBR is a related algorithm that uses significant Hi-C interactions to help with inference of chromatin states through a graph-based posterior regularization **[G]**[186].

Taken together, the integration of genome structure features and functional genomic data combined with DNA sequences is a promising new direction in the study of 3D genome organization. These approaches and models are expected to be crucial to shedding new light on the multiscale organization of the nuclear genome and its relationship to function.

Current limitations

Recent and ongoing advances in computational modelling approaches for probing 3D genome organization play an important part in revealing 3D genome features and generating new testable hypotheses on the interplay between genome structure and function. However, major limitations remain.

One of the most pressing challenges in 3D genome analysis is the lack of methods that can identify the full range of multiscale 3D genome features. The folding of chromatin occurs over ranges of kilobases to megabases, but existing methods designed to identify 3D genome features are often limited to a fixed set of length scales, which can make it difficult to study the relationships between the different levels of chromatin folding and cellular processes and functions. Moreover, few methods can use multiple data types (generated by different approaches such as Hi-C, ligation-free mapping methods, scHi-C, and imaging) as input[23,187]. This can limit the scope and accuracy of the analysis, as different data types could provide complementary information about the 3D genome if they were compared in the same analysis approach. Integrative frameworks that can utilize different types of data will ultimately provide a more comprehensive view of nuclear organization.

A further challenge is the analysis of scHi-C data, which exhibits a high level of complexity because single-cell 3D genome organization is intrinsically significantly variable[16]. This variability, together with data quality limitations and substantial costs for profiling large numbers of cells, can make it challenging to obtain reliable and consistent results from single-cell 3D genome data analysis. Additionally, integrative analysis with other single-cell data types and bulk data is difficult. In particular, new methods are needed to reveal, and understand the biological implications of, the differences between features observed from bulk Hi-C and scHi-C data.

It is imperative to recognize that evaluating the outcomes of computational methods should consider potential biases stemming from various input data types and preprocessing steps. For instance, the normalization strategy used in processing Hi-C data could introduce biases, such as possible underestimation of inter-chromosomal interactions. Each experimental technique produces data with its own distinct patterns and limitations, meaning computational models tested on certain data types, such as Hi-C, may not deliver accurate results for other data types, such as SPRITE. Therefore, care must be taken to choose a method appropriate for the data type and research questions. A comprehensive evaluation of methods using the same benchmark data can also be advantageous in identifying these biases and providing guidance on different application scenarios.

Deep-learning-based methods have demonstrated substantial success in predicting 3D genome features from 1D genomic features and in the integrative analysis of nuclear structure and function. However, these complex ‘black-box’ methods still have limitations in terms of interpretability and often require independent validation in unseen datasets for cell type-specific predictions. Over recent years, post-hoc methods[174,188] and attention-based methods[189] **[G]** have been developed to reveal the essential input features contributing to the model’s predictions. These methods have delivered mechanistic insights into nuclear structure and function, including the identification of sequence determinants for genome folding[77], the prediction of chromatin spatial positioning relative to nuclear bodies[190], and the recognition of chromatin interactions that are crucial for regulating gene expression[173]. Nevertheless, it is important to note that feature importance scores can vary across different methods and may be cell-type specific, potentially yielding inconsistent results[191]. Therefore, additional verification steps are necessary to ensure the robustness of findings, especially across various cell types not used during the training process. Moreover, it is

essential to acknowledge that feature importance scores do not necessarily reflect true underlying biology, particularly given the high correlation among multi-modal genomic features. To overcome these challenges, future research and collaboration between genomics and machine learning experts are paramount in designing interpretable machine learning methods specifically tailored to genomic data and in deciphering the mechanisms of the 3D genome from the computational perspective.

Future directions

The field of 3D genome research is rapidly evolving, and there are many exciting future directions ahead. One outstanding question in the field is how to design computational models that can coherently integrate observations from multimodal data, including genomics, imaging, 3D genome structural models, as well as multi-omic single-cell data. These different techniques often probe various aspects of genome architecture at different resolutions, creating a potential opportunity for an integrated framework that can simultaneously analyze complementary data to provide a complete view of various technical, biological, and physiological aspects of nuclear organization. Some of the major areas of expected progress are outlined below.

Integration between genomic and imaging data.

One future direction to obtain a comprehensive picture of genome organization is to coherently integrate various genomic data with imaging data to better understand the relationship between the 3D genome and other components in the cell nucleus. Compared with sequencing-based methods, recent multiplexed imaging based approaches have emerged as powerful tools for visualization of chromatin structure. These methods can probe multiple loci close to the nucleosome scale, as demonstrated in a recent preprint[192], or simultaneously visualize hundreds of DNA probes and RNA molecules, along with immunostaining of nuclear bodies and histone modification, all within a single cell[193]. Although the resolution of current multiplexed imaging methods is low compared to sequencing-based methods at the population level and are often restricted to pre-selected loci, their throughput has the potential to be improved through the development of experiment strategies, thereby providing a more comprehensive view of the 3D genome and its role in cellular processes and functions[23]. For instance, a recent preprint reported a new method, two-layer DNA FISH[194], which utilizes two rounds of imaging to separately resolve each 25kb bin and the 1.5Mb chromosome blocks in which these bins are located, thereby probing ~100,000 loci per cell. In another recent preprint, the method Pop-C[195] was reported as a multiplexed version of Dip-C that was designed to enhance throughput.

More realistic nuclear architecture models.

Another future direction is to develop more realistic models of nuclear architecture by integrating various -omics assays, single-cell assays, and imaging data. These models will need to accurately reflect the variability and dynamics of the 3D genome across different cells and over time[196], as compared with static models. This will enable a better understanding of the complex relationships between the 3D genome and cellular processes and functions and how chromatin organization reacts to external stimuli.

Explanation of the functional impact of 3D genome features.

Another important future direction is to gain a better understanding of the functional impact of 3D genome features. One promising avenue for future research involves using *in silico* mutagenesis to assess the impacts of sequence alterations[75]. This approach was used to identify loci with divergent 3D genome organization between humans and archaic hominins in a recent preprint[197]. Making predictions trustworthy will require the development of new methods and techniques to specifically perturb 3D genome features and assess the downstream impact on gene expression and cellular processes.

Disease relevance.

Finally, a major future direction is to understand the role of the 3D genome in disease. This will involve identifying causal 3D genome changes that lead to phenotypic abnormalities and understanding how these changes disrupt cellular processes and functions. DNA alterations such as short tandem repeats[198] and structural variations[199] have been found to shape the 3D genome organization. As the volume of Hi-C data from patients continues to increase, a crucial area of focus emerges: discerning the connections among 3D genome structure, disease-causing DNA variants, and the consequent ectopic gene expression induced by aberrant chromatin interactions[200]. This information will be critical for developing new treatments for a range of diseases, including cancer and genetic disorders.

Conclusions

Computational approaches are a central part of most 3D genome mapping methods. Their importance now reaches far beyond data analysis, with roles in integrating complex datasets into structural models and generating mechanistic and functional hypotheses. In fact, it can be argued that the study of the 3D genome currently represents the most extensive use of advanced computing methods in cell biology. There is no doubt that the development and combined use of increasingly sensitive 3D genome mapping methods and more advanced computational approaches will provide a deeper understanding of the relationships between the 3D genome and cellular processes and functions and will have important implications for human health and disease.

Acknowledgements

This work was supported in part by the National Institutes of Health Common Fund 4D Nucleome Program grant UM1HG011593 (J.M., F.A., T.M.), National Institutes of Health grants R01HG007352 (J.M.), R01HG012303 (J.M.). J.M. was additionally supported by a Guggenheim Fellowship from the John Simon Guggenheim Memorial Foundation and T.M. was supported by the Intramural Program of the NIH, the NCI Center for Cancer Research (1 ZIA BC010309-23).

Glossary

Attention-based mechanism
A machine learning technique used in neural network models to prioritize the most relevant parts of input when making decisions

Bystander interactions

An indirect effect of chromatin interactions that occurs through nearby long-range chromatin interactions

Chromatin domains
Distinct units formed by the chromatin fiber inside the cell nucleus

Cohesin-mediated loop extrusion
The process by which cohesin complexes extrude DNA into loops until they reach boundaries insulated by architectural chromatin proteins such as CTCF

Contact frequency
The probability that a pair of genomic loci in the 1D linear genome are spatially closer than a threshold value

Contact frequency map
A symmetrical square matrix filled with the estimated contact probabilities between any pair of loci

Convolutional kernel
A compact matrix or vector applied to an input DNA sequence to represent the significance of a specific pattern or feature at each position within the sequence

Convolutional neural network (CNN)
A type of artificial neural network that uses convolution layers to learn data representations by applying filters or kernels to the input signal to generate transformed output signals

Dilated CNN
A variant of convolutional neural networks (CNN) in which a ‘dilated’ or expanded convolution kernel is applied to broaden the receptive field of the network without increasing the number of parameters

Gradient boosting
A machine learning method that iteratively trains a series of models, each of which is designed to correct error from the previous model

Graph-based posterior regularization
A method to improve the prediction accuracy by incorporating a penalty term that discourages solutions that violate the dependencies between variables, as represented within a graph structure

Graph neural network
A type of neural network that is designed to handle data represented as graphs

Homotypic interactions
Binding or association of objects with similar properties, in this context chromatin segments sharing the same type of activity

Hypergraph

A generalization of a graph in which an edge can connect any number of vertices

Hyperedge
A connection between any number of vertices of a hypergraph

Latent Dirichlet allocation
A type of generative statistical model that allows sets of observations to be explained by unobserved groups or topics

Molecular Dynamics
A computational method used to simulate the evolution of a molecular system over time, by numerically integrating Newton's equations of motion

Monte Carlo
A computational method to generate configurations of a physical system by drawing samples from a probability distribution

Principal component analysis
A statistical method used to reduce the dimensionality of data by transforming it into a set of linearly uncorrelated variables, known as principal components

Phase separation
Spatial separation of different phases of matter from one homogeneous mixture

Random forest
A classic machine learning method for classification and regression tasks that works by combining the output of multiple decision trees

Recurrent neural network
An artificial neural network designed to recognize patterns in sequences of data, with each output dependent on previous computations

Random walk with restart
A stochastic process on a graph that randomly selects a starting node and then probabilistically determines the next move

Tensor decomposition
A mathematical technique used to break down a complex tensor (multi-dimensional data) into a series of simpler, more interpretable components

Variational inference
A probabilistic method in machine learning and statistics that uses optimization techniques to approximate complex, intractable posterior distributions

Vector representation
A means of representing data, usually in the form of a real-valued vector, such that the data points that are closer in the vector space are expected to have similar attributes

References

1. Misteli T The Self-Organizing Genome: Principles of Genome Architecture and Function. Cell 183, 28–45 (2020). [PubMed: 32976797] This review captures the recent state of the field and defines some of the basic principles that shape genome organization.
2. Tolhuis B, Palstra R-J, Splinter E, Grosveld F & de Laat W Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Mol. Cell 10, 1453–1465 (2002). [PubMed: 12504019]
3. Tang Z et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell 163, 1611–1627 (2015). [PubMed: 26686651]
4. Dixon JR et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012). [PubMed: 22495300]
5. Sexton T et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148, 458–472 (2012). [PubMed: 22265598]
6. Nora EP et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385 (2012). [PubMed: 22495304]
7. Fudenberg G et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 15, 2038–2049 (2016). [PubMed: 27210764]
8. Lieberman-Aiden E et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–293 (2009). [PubMed: 19815776]
9. Simonis M et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat. Genet 38, 1348–1354 (2006). [PubMed: 17033623]
10. Finn EH et al. Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. Cell 176, 1502–1515.e10 (2019). [PubMed: 30799036]
11. Wang S et al. Spatial organization of chromatin domains and compartments in single chromosomes. Science 353, 598–602 (2016). [PubMed: 27445307]
12. Bintu B et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science 362, eaau1783. (2018). [PubMed: 30361340]
13. Stevens TJ et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59–64 (2017). [PubMed: 28289288]
14. Kind J et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell 163, 134–147 (2015). [PubMed: 26365489]
15. Arrastia MV et al. Single-cell measurement of higher-order 3D genome organization with scSPRITE. Nat. Biotechnol 40, 64–73 (2021). [PubMed: 34426703]
16. Finn EH & Misteli T Molecular basis and biological function of variability in spatial genome organization. Science 365, eaaw9498 (2019). [PubMed: 31488662]
17. Duan Z et al. A three-dimensional model of the yeast genome. Nature 465, 363–367 (2010). [PubMed: 20436457]
18. Hsieh T-HS et al. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell 162, 108–119 (2015). [PubMed: 26119342]
19. Krietenstein N et al. Ultrastructural Details of Mammalian Chromosome Architecture. Mol. Cell 78, 554–565.e7 (2020). [PubMed: 32213324]
20. Hsieh T-HS et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol. Cell 78, 539–553.e8 (2020). [PubMed: 32213323]
21. Nagano T et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64 (2013). [PubMed: 24067610] This study describes the first single-cell Hi-C method and reveals single cell heterogeneity.
22. Su J-H, Zheng P, Kinrot SS, Bintu B & Zhuang X Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. Cell 182, 1641–1659.e26 (2020). [PubMed: 32822575]
23. Boninsegna L et al. Integrative genome modeling platform reveals essentiality of rare contact events in 3D genome organizations. Nat. Methods 19, 938–946 (2022). [PubMed: 35817938] This

work develops a multimodal data integration approach for generating populations of single-cell 3D genome structures that are predictive for various features of genome organization and function.

24. Ay F & Noble WS Analysis methods for studying the 3D architecture of the genome. Genome Biol. 16, 183 (2015). [PubMed: 26328929]
25. Schmitt AD, Hu M & Ren B Genome-wide mapping and analysis of chromosome architecture. Nat. Rev. Mol. Cell Biol 17, 743–755 (2016). [PubMed: 27580841]
26. Yaffe E & Tanay A Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet 43, 1059–1065 (2011). [PubMed: 22001755]
27. Hu M et al. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 28, 3131–3133 (2012). [PubMed: 23023982]
28. Knight PA & Ruiz D A fast algorithm for matrix balancing. IMA J. Numer. Anal 33, 1029–1047 (2012).
29. Imakaev M et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 9, 999–1003 (2012). [PubMed: 22941365]
30. Rao SSP et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014). [PubMed: 25497547] This work reveals 3D genome subcompartments as well as chromatin folding patterns from Hi-C maps at 1kb resolution, providing insights into the convergent orientation CTCF binding sites at loop anchors.
31. Cremer T & Cremer C Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat. Rev. Genet 2, 292–301 (2001). [PubMed: 11283701]
32. Di Pierro M, Cheng RR, Lieberman Aiden E, Wolynes PG & Onuchic JN De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. Proc. Natl. Acad. Sci. U. S. A 114, 12126–12131 (2017). [PubMed: 29087948]
33. Xiong K & Ma J Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. Nat. Commun 10, 5069 (2019). [PubMed: 31699985]
34. Ashoor H et al. Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. Nat. Commun 11, 1173 (2020). [PubMed: 32127534]
35. Liu Y et al. Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. Nat. Commun 12, 2439 (2021). [PubMed: 33972523]
36. Wang Y et al. SPIN reveals genome-wide landscape of nuclear compartmentalization. Genome Biol. 22, 36 (2021). [PubMed: 33446254]
37. Beagan JA & Phillips-Cremins JE On the existence and functionality of topologically associating domains. Nat. Genet 52, 8–16 (2020). [PubMed: 31925403]
38. Crane E et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523, 240–244 (2015). [PubMed: 26030525]
39. Giorgetti L et al. Structural organization of the inactive X chromosome in the mouse. Nature 535, 575–579 (2016). [PubMed: 27437574]
40. Bonev B et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171, 557–572.e24 (2017). [PubMed: 29053968]
41. Zhan Y et al. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Res. 27, 479–490 (2017). [PubMed: 28057745]
42. Filippova D, Patro R, Duggal G & Kingsford C Identification of alternative topological domains in chromatin. Algorithms Mol. Biol 9, 14 (2014). [PubMed: 24868242]
43. Weinreb C & Raphael BJ Identification of hierarchical chromatin domains. Bioinformatics 32, 1601–1609 (2016). [PubMed: 26315910]
44. Norton HK et al. Detecting hierarchical genome folding with network modularity. Nat. Methods 15, 119–122 (2018). [PubMed: 29334377]
45. Zufferey M, Tavernari D, Oricchio E & Ciriello G Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 19, 217 (2018). [PubMed: 30526631]

46. Forcato M et al. Comparison of computational methods for Hi-C data analysis. Nat. Methods 14, 679–685 (2017). [PubMed: 28604721]
47. Ay F, Bailey TL & Noble WS Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 24, 999–1011 (2014). [PubMed: 24501021]
48. Carty M et al. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat. Commun 8, 15454 (2017). [PubMed: 28513628]
49. Sahin M et al. HiC-DC+ enables systematic 3D interaction calls and differential analysis for Hi-C and HiChIP. Nat. Commun 12, 3366 (2021). [PubMed: 34099725]
50. Rowley MJ et al. Analysis of Hi-C data using SIP effectively identifies loops in organisms from *C. elegans* to mammals. Genome Res. 30, 447–458 (2020). [PubMed: 32127418]
51. 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. 21, 256 (2020). [PubMed: 32998764]
52. Galan S, Serra F & Marti-Renom MA Identification of chromatin loops from Hi-C interaction matrices by CTCF–CTCF topology classification. NAR Genom. Bioinform 4, lqac021 (2022). [PubMed: 35274099]
53. Salameh TJ et al. A supervised learning framework for chromatin loop detection in genome-wide contact maps. Nat. Commun 11, 3428 (2020). [PubMed: 32647330]
54. Yoon S, Chandra A & Vahedi G Stripenn detects architectural stripes from chromatin conformation data using computer vision. Nat. Commun 13, 1602 (2022). [PubMed: 35332165]
55. Gupta K et al. StripeDiff: Model-based algorithm for differential analysis of chromatin stripe. Sci. Adv 8, eabk2246 (2022). [PubMed: 36475785]
56. Vian L et al. The Energetics and Physiological Impact of Cohesin Extrusion. Cell 175, 292–294 (2018). [PubMed: 30241609]
57. Zhang Y & Blanchette M Reference panel guided topological structure annotation of Hi-C data. Nat. Commun 13, 7426 (2022). [PubMed: 36460680]
58. Avdeyev P & Zhou J Computational Approaches for Understanding Sequence Variation Effects on the 3D Genome Architecture. Annu Rev Biomed Data Sci 5, 183–204 (2022). [PubMed: 35537461]
59. Yang M & Ma J Machine Learning Methods for Exploring Sequence Determinants of 3D Genome Organization. J. Mol. Biol 434, 167666 (2022). [PubMed: 35659533]
60. Zhang S, Chasman D, Knaack S & Roy S In silico prediction of high-resolution Hi-C interaction matrices. Nat. Commun 10, 5449 (2019). [PubMed: 31811132]
61. LeCun Y, Bengio Y & Hinton G Deep learning. Nature 521, 436–444 (2015). [PubMed: 26017442]
62. Elman JL Finding structure in time. Cogn. Sci 14, 179–211 (1990).
63. Zhou J et al. Graph neural networks: A review of methods and applications. AI Open 1, 57–81 (2020).
64. Wu Z et al. A Comprehensive Survey on Graph Neural Networks. IEEE Trans Neural Netw Learn Syst 32, 4–24 (2021). [PubMed: 32217482]
65. Zhang R, Wang Y, Yang Y, Zhang Y & Ma J Predicting CTCF-mediated chromatin loops using CTCF-MP. Bioinformatics 34, i133–i141 (2018). [PubMed: 29949986]
66. Kai Y et al. Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. Nat. Commun 9, 4221 (2018). [PubMed: 30310060]
67. Whalen S, Truty RM & Pollard KS Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet 48, 488–496 (2016). [PubMed: 27064255]
68. Yang Y, Zhang R, Singh S & Ma J Exploiting sequence-based features for predicting enhancer–promoter interactions. Bioinformatics 33, i252–i260 (2017). [PubMed: 28881991]
69. Singh S, Yang Y, Póczos B & Ma J Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Biol 7, 122–137 (2019). [PubMed: 34113473]
70. Cao Q et al. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nat. Genet 49, 1428–1436 (2017). [PubMed: 28869592]
71. Zeng W, Wu M & Jiang R Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics 19, 84 (2018). [PubMed: 29764360]

72. Cao F et al. Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. Genome Biol. 22, 226 (2021). [PubMed: 34399797]
73. Li W, Wong WH & Jiang R DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res. 47, e60 (2019). [PubMed: 30869141]
74. Gabriele M et al. Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. Science 376, 496–501 (2022). [PubMed: 35420890]
75. Fudenberg G, Kelley DR & Pollard KS Predicting 3D genome folding from DNA sequence with Akita. Nature Methods vol. 17 1111–1117 Preprint at 10.1038/s41592-020-0958-x (2020). [PubMed: 33046897] This method demonstrates that using DNA sequence alone could predict chromatin contact frequency maps with high accuracy.
76. Zhou J Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. Nat. Genet 54, 725–734 (2022). [PubMed: 35551308]
77. Schwessinger R et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat. Methods 17, 1118–1124 (2020). [PubMed: 33046896]
78. Tan J et al. Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. Nat. Biotechnol (2023) doi:10.1038/s41587-022-01612-8.
79. Yang R et al. Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. Genome Biol. 24, 134 (2023). [PubMed: 37280678]
80. Kim M et al. MIA-Sig: multiplex chromatin interaction analysis by signal processing and statistical algorithms. Genome Biol. 20, 251 (2019). [PubMed: 31767038]
81. Zhang R & Ma J MATCHA: Probing multi-way chromatin interaction with hypergraph representation learning. Cell Syst. 10, 397–407.e5 (2020). [PubMed: 32550271]
82. Dotson GA et al. Deciphering multi-way interactions in the human genome. Nat. Commun 13, 5498 (2022). [PubMed: 36127324]
83. Lin D, Bonora G, Yardımcı GG & Noble WS Computational methods for analyzing and modeling genome structure and organization. Wiley Interdiscip. Rev. Syst. Biol. Med 11, e1435 (2019). [PubMed: 30022617]
84. Brackey CA, Marenduzzo D & Gilbert N Mechanistic modeling of chromatin folding to understand function. Nat. Methods 17, 767–775 (2020). [PubMed: 32514111]
85. Wong H et al. A predictive computational model of the dynamic 3D interphase yeast nucleus. Curr. Biol 22, 1881–1890 (2012). [PubMed: 22940469]
86. Tjong H, Gong K, Chen L & Alber F Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome Res. 22, 1295–1305 (2012). [PubMed: 22619363]
87. Bianco S et al. Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet 50, 662–667 (2018). [PubMed: 29662163] This study highlights the efficacy of polymer modelling-based methods in predicting chromatin structure changes with large-scale sequence alterations, such as structural variants.
88. Kragesteen BK et al. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. Nat. Genet 50, 1463–1473 (2018). [PubMed: 30262816]
89. Nora EP et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell 169, 930–944.e22 (2017). [PubMed: 28525758]
90. Haarhuis JHI et al. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell 169, 693–707.e14 (2017). [PubMed: 28475897]
91. Schwarzer W et al. Two independent modes of chromatin organization revealed by cohesin removal. Nature 551, 51–56 (2017). [PubMed: 29094699]
92. International Nucleome Consortium. 3DGenBench: a web-server to benchmark computational models for 3D Genomics. Nucleic Acids Res. 50, W4–W12 (2022). [PubMed: 35639501]
93. Alipour E & Marko JF Self-organization of domain structures by DNA-loop-extruding enzymes. Nucleic Acids Res. 40, 11202–11212 (2012). [PubMed: 23074191]

94. Sanborn AL et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U. S. A 112, E6456–65 (2015). [PubMed: 26499245]
95. Rossini R, Kumar V, Mathelier A, Rognes T & Paulsen J MoDLE: high-performance stochastic modeling of DNA loop extrusion interactions. Genome Biol. 23, 247 (2022). [PubMed: 36451166]
96. Gitchev T, Zala G, Meister P & Jost D 3DPolyS-LE: an accessible simulation framework to model the interplay between chromatin and loop extrusion. Bioinformatics 38, 5454–5456 (2022). [PubMed: 36355469]
97. Ganji M et al. Real-time imaging of DNA loop extrusion by condensin. Science 360, 102–105 (2018). [PubMed: 29472443]
98. Davidson IF et al. DNA loop extrusion by human cohesin. Science 366, 1338–1345 (2019). [PubMed: 31753851]
99. Banigan EJ & Mirny LA Loop extrusion: theory meets single-molecule experiments. Curr. Opin. Cell Biol 64, 124–138 (2020). [PubMed: 32534241]
100. Mirny LA, Imakaev M & Abdennur N Two major mechanisms of chromosome organization. Curr. Opin. Cell Biol 58, 142–152 (2019). [PubMed: 31228682]
101. Gibcus JH et al. A pathway for mitotic chromosome formation. Science 359, eaao6135 (2018). [PubMed: 29348367]
102. Goloborodko A, Imakaev MV, Marko JF & Mirny L Compaction and segregation of sister chromatids via active loop extrusion. Elife 5, e14864 (2016). [PubMed: 27192037]
103. Goloborodko A, Marko JF & Mirny LA Chromosome Compaction by Active Loop Extrusion. Biophys. J 110, 2162–2168 (2016). [PubMed: 27224481]
104. Schalbetter SA, Fudenberg G, Baxter J, Pollard KS & Neale MJ Principles of meiotic chromosome assembly revealed in S. cerevisiae. Nat. Commun 10, 4795 (2019). [PubMed: 31641121]
105. Zhang Y et al. The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. Nature 573, 600–604 (2019). [PubMed: 31511698]
106. Fudenberg G, Abdennur N, Imakaev M, Goloborodko A & Mirny LA Emerging Evidence of Chromosome Folding by Loop Extrusion. Cold Spring Harb. Symp. Quant. Biol 82, 45–55 (2017). [PubMed: 29728444]
107. Nuebler J, Fudenberg G, Imakaev M, Abdennur N & Mirny LA Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc. Natl. Acad. Sci. U. S. A 115, E6697–E6706 (2018). [PubMed: 29967174] The mechanisms proposed in this study successfully recapitulates perturbations in chromatin organization observed in knock-down experiments.
108. Hildebrand EM & Dekker J Mechanisms and Functions of Chromosome Compartmentalization. Trends Biochem. Sci 45, 385–396 (2020). [PubMed: 32311333]
109. Falk M et al. Heterochromatin drives compartmentalization of inverted and conventional nuclei. Nature 570, 395–399 (2019). [PubMed: 31168090]
110. Feric M & Misteli T Phase separation in genome organization across evolution. Trends Cell Biol. 31, 671–685 (2021). [PubMed: 33771451]
111. Sabari BR, Dall'Agnese A & Young RA Biomolecular Condensates in the Nucleus. Trends Biochem. Sci 45, 961–977 (2020). [PubMed: 32684431]
112. Jost D, Carrivain P, Cavalli G & Vaillant C Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res. 42, 9553–9561 (2014). [PubMed: 25092923]
113. Zhang B & Wolynes PG Topology, structures, and energy landscapes of human chromosomes. Proc. Natl. Acad. Sci. U. S. A 112, 6062–6067 (2015). [PubMed: 25918364]
114. Di Pierro M, Zhang B, Aiden EL, Wolynes PG & Onuchic JN Transferable model for chromosome architecture. Proc. Natl. Acad. Sci. U. S. A 113, 12168–12173 (2016). [PubMed: 27688758]
115. Brackley CA, Taylor S, Papantonis A, Cook PR & Marenduzzo D Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. Proc. Natl. Acad. Sci. U. S. A 110, E3605–11 (2013). [PubMed: 24003126]

116. Buckle A, Brackley CA, Boyle S, Marenduzzo D & Gilbert N Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci. Mol. Cell 72, 786–797.e11 (2018). [PubMed: 30344096]

117. Shi G, Liu L, Hyeon C & Thirumalai D Interphase human chromosome exhibits out of equilibrium glassy dynamics. Nat. Commun 9, 3161 (2018). [PubMed: 30089831]

118. Haddad N, Jost D & Vaillant C Perspectives: using polymer modeling to understand the formation and function of nuclear compartments. Chromosome Res. 25, 35–50 (2017). [PubMed: 28091870]

119. Barbieri M et al. Complexity of chromatin folding is captured by the strings and binders switch model. Proc. Natl. Acad. Sci. U. S. A 109, 16173–16178 (2012). [PubMed: 22988072]

120. Barbieri M et al. Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. Nat. Struct. Mol. Biol 24, 515–524 (2017). [PubMed: 28436944]

121. Bianco S et al. Computational approaches from polymer physics to investigate chromatin folding. Curr. Opin. Cell Biol 64, 10–17 (2020). [PubMed: 32045823]

122. Conte M et al. Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. Nat. Commun 11, 3289 (2020). [PubMed: 32620890]

123. Fujishiro S & Sasai M Generation of dynamic three-dimensional genome structure through phase separation of chromatin. Proc. Natl. Acad. Sci. U. S. A 119, e2109838119 (2022). [PubMed: 35617433]

124. Contessoto VG et al. Interphase chromosomes of the Aedes aegypti mosquito are liquid crystalline and can sense mechanical cues. Nat. Commun 14, 326 (2023). [PubMed: 36658127]

125. Brackley CA et al. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome Biol. 17, 59 (2016). [PubMed: 27036497]

126. Qi Y & Zhang B Predicting three-dimensional genome organization with chromatin states. PLoS Comput. Biol 15, e1007024 (2019). [PubMed: 31181064] This paper predicts chromatin structure de novo by simulating phase separation using a maximum entropy approach on ensemble Hi-C and epigenomic data.

127. Conte M et al. Loop-extrusion and polymer phase-separation can co-exist at the single-molecule level to shape chromatin folding. Nat. Commun 13, 4070 (2022). [PubMed: 35831310]

128. Qi Y et al. Data-Driven Polymer Model for Mechanistic Exploration of Diploid Genome Organization. Biophys. J 119, 1905–1916 (2020). [PubMed: 33086041]

129. Brahmachari S, Contessoto V, Di Pierro M & Onuchic JN Shaping the Genome via Lengthwise Compaction, Phase Separation, and Lamina Adhesion. Nucleic Acids Res. 50, 4258–4271 (2022). [PubMed: 35420130]

130. Kamat K et al. Compartmentalization with nuclear landmarks yields random, yet precise, genome organization. Biophys. J S0006–3495(23)00156-X (2023).

131. Kalhor R, Tjong H, Jayathilaka N, Alber F & Chen L Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol 30, 90–98 (2011). [PubMed: 22198700]

132. Yildirim A, Boninsegna L, Zhan Y & Alber F Uncovering the Principles of Genome Folding by 3D Chromatin Modeling. Cold Spring Harb. Perspect. Biol 14, a039693 (2022). [PubMed: 34400556] This comprehensive review discusses many mechanistic and data-driven computational approaches for genome structure modelling.

133. Boninsegna L, Yildirim A, Zhan Y & Alber F Integrative approaches in genome structure analysis. Structure 30, 24–36 (2022). [PubMed: 34963059]

134. Baù D et al. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat. Struct. Mol. Biol 18, 107–114 (2011). [PubMed: 21131981]

135. Serra F et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput. Biol 13, e1005665 (2017). [PubMed: 28723903]

136. Umbarger MA et al. The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. Mol. Cell 44, 252–264 (2011). [PubMed: 22017872]

137. Nir G et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS Genet. 14, e1007872 (2018). [PubMed: 30586358]

138. Paulsen J et al. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. Genome Biol. 18, 21 (2017). [PubMed: 28137286]

139. Shi G & Thirumalai D From Hi-C Contact Map to Three-Dimensional Organization of Interphase Human Chromosomes. Phys. Rev. X 11, 011051 (2021).

140. Gehlen LR et al. Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. Nucleus 3, 370–383 (2012). [PubMed: 22688649]

141. Trieu T & Cheng J Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. Nucleic Acids Res. 42, e52 (2014). [PubMed: 24465004]

142. Di Stefano M, Paulsen J, Lien TG, Hovig E & Micheletti C Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. Sci. Rep 6, 35985 (2016). [PubMed: 27786255]

143. Yildirim A & Feig M High-resolution 3D models of Caulobacter crescentus chromosome reveal genome structural variability and organization. Nucleic Acids Res. 46, 3937–3952 (2018). [PubMed: 29529244]

144. Rosenthal M et al. Bayesian Estimation of Three-Dimensional Chromosomal Structure from Single-Cell Hi-C Data. J. Comput. Biol 26, 1191–1202 (2019). [PubMed: 31211598]

145. Tan L, Xing D, Chang C-H, Li H & Sunney Xie X Three-dimensional genome structures of single diploid human cells. Science 361, 924–928 (2018). [PubMed: 30166492]

146. Zhang R, Zhou T & Ma J Multiscale and integrative single-cell Hi-C analysis with Higashi. Nat. Biotechnol 40, 254–261 (2022). [PubMed: 34635838] This work introduces a computational framework based on hypergraph representation learning to effectively achieve embedding and data imputation for scHi-C data, enabling integrative analysis of 3D genome features in individual cells.

147. Zhou T, Zhang R & Ma J The 3D Genome Structure of Single Cells. Annual Review of Biomedical Data Science 4, 21–41 (2021).

148. Giorgetti L et al. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. Cell 157, 950–963 (2014). [PubMed: 24813616]

149. Perez-Rathke A et al. CHROMATIX: computing the functional landscape of many-body chromatin interactions in transcriptionally active loci from deconvolved single cells. Genome Biol. 21, 13 (2020). [PubMed: 31948478]

150. Tjong H et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proc. Natl. Acad. Sci. U. S. A 113, E1663–72 (2016). [PubMed: 26951677]

151. Li Q et al. The three-dimensional genome organization of Drosophila melanogaster through data integration. Genome Biol. 18, 145 (2017). [PubMed: 28760140]

152. Yildirim A et al. Population-based structure modeling reveals key roles of nuclear microenviroment in gene functions. Nat. Struct. Mol. Biol. (in press) 2021.07.11.451976v2 doi:10.1101/2021.07.11.451976v2.

153. Zhan Y, Yildirim A, Boninsegna L & Alber F Conformational analysis of chromosome structures reveals vital role of chromosome morphology in gene function. bioRxiv (2023) doi:10.1101/2023.02.18.528138.

154. Stegle O, Teichmann SA & Marioni JC Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet 16, 133–145 (2015). [PubMed: 25628217]

155. Jia BB, Jussila A, Kern C, Zhu Q & Ren B A spatial genome aligner for resolving chromatin architectures from multiplexed DNA FISH. Nat. Biotechnol (2023) doi:10.1038/s41587-022-01568-9.This work presents a computational framework that effectively reconstructs chromatin structures from multiplexed DNA FISH data with high accuracy.

156. van der Maaten L & Hinton G Visualizing Data using t-SNE. J. Mach. Learn. Res 9, 2579–2605 (2008).

157. McInnes L, Healy J & Melville J UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint (2018) doi:10.48550/arXiv.1802.03426.

158. Ramani V et al. Massively multiplex single-cell Hi-C. Nat. Methods 14, 263–266 (2017). [PubMed: 28135255]

159. Kim H-J et al. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. PLoS Comput. Biol 16, e1008173 (2020). [PubMed: 32946435]

160. Liu J, Lin D, Yardimci GG & Noble WS Unsupervised embedding of single-cell Hi-C data. Bioinformatics 34, i96–i104 (2018). [PubMed: 29950005]

161. Zhang R, Zhou T & Ma J Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi. Cell Syst. 13, 798–807.e6 (2022). [PubMed: 36265466]

162. Zheng Y, Shen S & Keleş S Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. Genome Biol. 23, 222 (2022). [PubMed: 36253828]

163. Galitsyna AA & Gelfand MS Single-cell Hi-C data analysis: safety in numbers. Brief. Bioinform 22, bbab316 (2021). [PubMed: 34406348]

164. Zhou J et al. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. Proc. Natl. Acad. Sci. U. S. A 116, 14011–14018 (2019). [PubMed: 31235599]

165. Yu M et al. SnapHiC: a computational pipeline to identify chromatin loops from single-cell Hi-C data. Nat. Methods 18, 1056–1059 (2021). [PubMed: 34446921]

166. Tan L et al. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. Cell 184, 741–758.e17 (2021). [PubMed: 33484631]

167. Zhang S et al. DeepLoop robustly maps chromatin interactions from sparse allele-resolved or single-cell Hi-C data at kilobase resolution. Nat. Genet 54, 1013–1025 (2022). [PubMed: 35817982]

168. Xiong K, Zhang R & Ma J scGHOST: Identifying single-cell 3D genome subcompartments. bioRxiv (2023) doi:10.1101/2023.05.24.542032.

169. Arnould C et al. Loop extrusion as a mechanism for formation of DNA damage repair foci. Nature 590, 660–665 (2021). [PubMed: 33597753]

170. Lupiáñez DG, Spielmann M & Mundlos S Breaking TADs: How Alterations of Chromatin Domains Result in Disease. Trends Genet. 32, 225–237 (2016). [PubMed: 26862051]

171. Hnisz D et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science 351, 1454–1458 (2016). [PubMed: 26940867]

172. Avsec Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods 18, 1196–1203 (2021). [PubMed: 34608324]

173. Karbalayghareh A, Sahin M & Leslie CS Chromatin interaction–aware gene regulatory modeling with graph attention networks. Genome Res. 32, 930–944 (2022). [PubMed: 35396274] This paper presents a graph-based deep learning approach that predicts gene expression by combining chromatin interaction, 1D epigenomic data, and DNA sequence features.

174. Selvaraju RR et al. Grad-CAM: Visual explanations from deep networks via Gradient-based localization. in Proc. IEEE Int. Conf. Comput. Vis 618–626 (2016).

175. Ying R, Bourgeois D, You J, Zitnik M & Leskovec J GNNExplainer: Generating Explanations for Graph Neural Networks. Adv. Neural Inf. Process. Syst 32, 9240–9251 (2019). [PubMed: 32265580]

176. Wang H et al. CRISPR-Mediated Programmable 3D Genome Positioning and Nuclear Organization. Cell 175, 1405–1417.e14 (2018). [PubMed: 30318144]

177. Marchal C, Sima J & Gilbert DM Control of DNA replication timing in the 3D genome. Nat. Rev. Mol. Cell Biol 20, 721–737 (2019). [PubMed: 31477886]

178. Zheng H & Xie W The role of 3D genome organization in development and cell differentiation. Nat. Rev. Mol. Cell Biol 20, 535–550 (2019). [PubMed: 31197269]

179. Yang Yang, Y., Wang Y, Zhang Y & Ma J Concert: Genome-wide prediction of sequence elements that modulate DNA replication timing. bioRxiv (2022) doi:10.1101/2022.04.21.488684.

180. Sima J et al. Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. Cell 176, 816–830.e18 (2019). [PubMed: 30595451]

181. Sun Y et al. A graph neural network-based interpretable framework reveals a novel DNA fragility-associated chromatin structural unit. Genome Biol. 24, 90 (2023). [PubMed: 37095580]
182. Pancaldi V et al. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biol. 17, 152 (2016). [PubMed: 27391817]
183. Wang J et al. Characterization of network hierarchy reflects cell state specificity in genome organization. Genome Res. 33, 247–260 (2023). [PubMed: 36828586]
184. Guelen L et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948–951 (2008). [PubMed: 18463634]
185. Chen Y et al. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. J. Cell Biol 217, 4025–4048 (2018). [PubMed: 30154186]
186. Libbrecht MW et al. Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. Genome Res. 25, 544–557 (2015). [PubMed: 25677182]
187. Abbas A et al. Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes. Nat. Commun 10, 2049 (2019). [PubMed: 31053705]
188. Shrikumar A, Greenside P & Kundaje A Learning Important Features Through Propagating Activation Differences. in Proceedings of the 34th International Conference on Machine Learning (eds. Precup D & Teh YW) vol. 70 3145–3153 (PMLR, 06--11 Aug 2017).
189. Vaswani A et al. Attention is All you Need. Adv. Neural Inf. Process. Syst 30, (2017).
190. Yang M & Ma J UNADON: transformer-based model to predict genome-wide chromosome spatial position. Bioinformatics 39, i553–i562 (2023). [PubMed: 37387176]
191. Yang M & Kim B Benchmarking Attribution Methods with Relative Feature Importance. arXiv [cs.LG] (2019).
192. Sasaki HM, Kishi JY, Wu C-T, Beliveau BJ & Yin P Quantitative multiplexed imaging of chromatin ultrastructure with Decode-PAINT. bioRxiv 2022.08.01.502089 (2022) doi:10.1101/2022.08.01.502089.
193. Takei Y et al. Integrated spatial genomics reveals global architecture of single nuclei. Nature 590, 344–350 (2021). [PubMed: 33505024] This study reports an extensive multi-omics analysis relating chromatin structure to epigenetic marks and gene expression at the single cell level.
194. Takei Y et al. High-resolution spatial multi-omics reveals cell-type specific nuclear compartments. bioRxiv (2023) doi:10.1101/2023.05.07.539762.
195. Tan L et al. Cerebellar Granule Cells Develop Non-neuronal 3D Genome Architecture over the Lifespan. bioRxiv (2023) doi:10.1101/2023.02.25.530020.
196. Viana MP et al. Integrated intracellular organization and its variations in human iPS cells. Nature 613, 345–354 (2023). [PubMed: 36599983]
197. McArthur E et al. Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence. bioRxiv 2022.02.07.479462 (2022) doi:10.1101/2022.02.07.479462.
198. Sun JH et al. Disease-associated short tandem repeats co-localize with chromatin domain boundaries. Cell 175, 224–238.e15 (2018). [PubMed: 30173918]
199. Dubois F, Sidiropoulos N, Weischenfeldt J & Beroukhim R Structural variations in cancer and the 3D genome. Nat. Rev. Cancer 22, 533–546 (2022). [PubMed: 35764888]
200. Wang X et al. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. Nat. Methods 18, 661–668 (2021). [PubMed: 34092790]
201. Redolfi J et al. DamC reveals principles of chromatin folding in vivo without crosslinking and ligation. Nat. Struct. Mol. Biol 26, 471–480 (2019). [PubMed: 31133702]
202. Dekker J, Rippe K, Dekker M & Kleckner N Capturing chromosome conformation. Science 295, 1306–1311 (2002). [PubMed: 11847345] This landmark paper describes the first method to detect chromatin interactions.
203. Kempfer R & Pombo A Methods for mapping 3D chromosome architecture. Nat. Rev. Genet 21, 207–226 (2019). [PubMed: 31848476]
204. Fullwood MJ et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58–64 (2009). [PubMed: 19890323]

205. Mumbach MR et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods 13, 919–922 (2016). [PubMed: 27643841]

206. van de Werken HJG et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat. Methods 9, 969–972 (2012). [PubMed: 22961246]

207. Hughes JR et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat. Genet 46, 205–212 (2014). [PubMed: 24413732]

208. Mifsud B et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet 47, 598–606 (2015). [PubMed: 25938943]

209. Ramani V et al. Mapping 3D genome architecture through in situ DNase Hi-C. Nat. Protoc 11, 2104–2121 (2016). [PubMed: 27685100]

210. Goel VY, Huseyin MK & Hansen AS Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. Nat. Genet 55, 1048–1056 (2023). [PubMed: 37157000]

211. Allahyar A et al. Enhancer hubs and loop collisions identified from single-allele topologies. Nat. Genet 50, 1151–1160 (2018). [PubMed: 29988121]

212. Zhong J-Y et al. High-throughput Pore-C reveals the single-allele topology and cell type-specificity of 3D genome folding. Nat. Commun 14, 1250 (2023). [PubMed: 36878904]

213. Deshpande AS et al. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat. Biotechnol 40, 1488–1499 (2022). [PubMed: 35637420]

214. Beagrie RA et al. Complex multi-enhancer contacts captured by genome architecture mapping. Nature 543, 519–524 (2017). [PubMed: 28273065]

215. Quinodoz SA et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell 174, 744–757.e24 (2018). [PubMed: 29887377]

216. Zheng M et al. Multiplex chromatin interactions with single-molecule precision. Nature 566, 558–562 (2019). [PubMed: 30778195]

217. Greil F, Moorman C & van Steensel B [16] DamID: Mapping of In Vivo Protein–Genome Interactions Using Tethered DNA Adenine Methyltransferase. in Methods in Enzymology vol. 410 342–359 (Academic Press, 2006). [PubMed: 16938559]

218. Vogel MJ, Peric-Hupkes D & van Steensel B Detection of in vivo protein–DNA interactions using DamID in mammalian cells. Nat. Protoc 2, 1467–1478 (2007). [PubMed: 17545983]

219. Girelli G et al. GPSeq reveals the radial organization of chromatin in the cell nucleus. Nat. Biotechnol 38, 1184–1193 (2020). [PubMed: 32451505]

220. Nagano T et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature 547, 61–67 (2017). [PubMed: 28682332]

221. Flyamer IM et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature 544, 110–114 (2017). [PubMed: 28355183]

222. Lee D-S et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat. Methods 16, 999–1006 (2019). [PubMed: 31501549]

223. Mateo LJ et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. Nature 568, 49–54 (2019). [PubMed: 30886393]

224. Cardozo Gizzi AM et al. Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. Mol. Cell 74, 212–222.e5 (2019). [PubMed: 30795893]

225. Schnitzbauer J, Strauss MT, Schlichthaerle T, Schueder F & Jungmann R Super-resolution microscopy with DNA-PAINT. Nat. Protoc 12, 1198–1228 (2017). [PubMed: 28518172]

226. Nguyen HQ et al. 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. Nat. Methods 17, 822–832 (2020). [PubMed: 32719531]

227. Payne AC et al. In situ genome sequencing resolves DNA sequence and structure in intact biological samples. Science 371, eaay3446 (2021). [PubMed: 33384301]

Box 1 |

Overview of different mapping methods for 3D genome organization

Diverse methods for mapping chromatin structure exist, which have improved our understanding of chromosome folding inside the nucleus (Fig. 1a).

Methods to capture pairwise chromatin interactions.

While ligation-free methods such as DamC[201] exist, most 3D genome mapping methods are derivatives of the chromosome conformation capture (3C) assay[202] and involve a DNA ligation step. In these approaches, spatially proximal chromatin loci are chemically crosslinked, the DNA fragmented with restriction enzymes and then re-ligated before paired-end sequencing to revealing the identities of proximal sequences in the genome[203] (Fig. 1b). The basic 3C assay has been further improved by incorporating various modifications, such as increasing the number of captured interacting loci with ultimate 'all versus all' methods (Hi-C)[8,17,18], selective enrichment of interactions involving specific proteins (for example, Chromatin Interaction Analysis with Paired-End-Tag sequencing (ChIA-PET), Hi-C chromatin immunoprecipitation (HiChIP))[204,205], capturing interactions at regions of interest (such as gene promoters[206–208]) and using DNase or MNase for more uniform fragmentation and enrichment of local interaction between pairs of nucleosomes[18,209,210].

Methods to capture multiway interactions.

Conventional 3C-like methods identify pairwise interactions but generally fail to uncover concurrent multiway chromosomal interactions. To address this limitation, multiplexed barcoding or long-read-based sequencing has been applied to 3C libraries[211–213]. Ligation-free methods such as Genome Architecture Mapping (GAM)[214], Split-Pool Recognition of Interactions by Tag Extension (SPRITE)[215], and ChIA-Drop[216] can provide complementary views to ligation-based methods by capturing multiway interactions beyond the range of proximity ligation (Fig. 1b)[203]. Multiway interactions provide higher order information regarding 3D genome structures and can facilitate structure modelling.

Ligation-free methods to map subnuclear positioning.

Ligation-free methods have been developed to capture interactions between chromatin and nuclear bodies (such as Tyramide Signal Amplification sequencing (TSA-seq)[185] and DNA Adenine Methyltransferase Identification (DamID)[217,218]) or probe radial positioning of chromatin in the nucleus (for example Genomic Loci Positioning by Sequencing (GPSeq)[219]) (Fig. 1c).. These methods provide information about the spatial arrangement of chromatin, including spatial positioning and an approximation of the cytological distance or contact frequency to a specific type of subnuclear structure.

Single-cell 3D genome mapping approaches.

While most genome mapping methods are population-based, the averaged chromatin interaction profiles generated from these methods are not capable of uncovering the heterogeneity among individual cells or the unique patterns in rare cell types. To solve

these issues, a series of single-cell Hi-C (scHi-C) methods have emerged to probe chromatin structures in individual cells and reveal cell type-specific 3D genome features in complex tissues (Fig. 1d)[13,21,145,158,166,220–222]. However, the sparsity of scHi-C data, with typically thousands of fold lower coverage than population-based methods, poses a challenge for analysis.

Imaging-based methods.

Microscopy-based imaging provides a direct measurement of 3D chromatin structures at the single cell level (Fig. 1e). Chromosome tracing using super-resolution microscopy can detect detailed chromatin folding trajectories[12,137,223–225], although it is limited to short stretches of chromatin. Recent innovations leveraging multiplexed barcoding methods have enabled the detection of thousands of probes simultaneously, along with RNA detection and subnuclear structure analysis[22,193,226,227]. These methods are limited by their throughput and coverage, with a maximum of a few thousand loci per single cell. Computational methods designed to assign raw fluorescence signals onto chromatin fibers, while accounting for both imaging noise and intrinsic biological noise, remain in the early-stage of development[155].

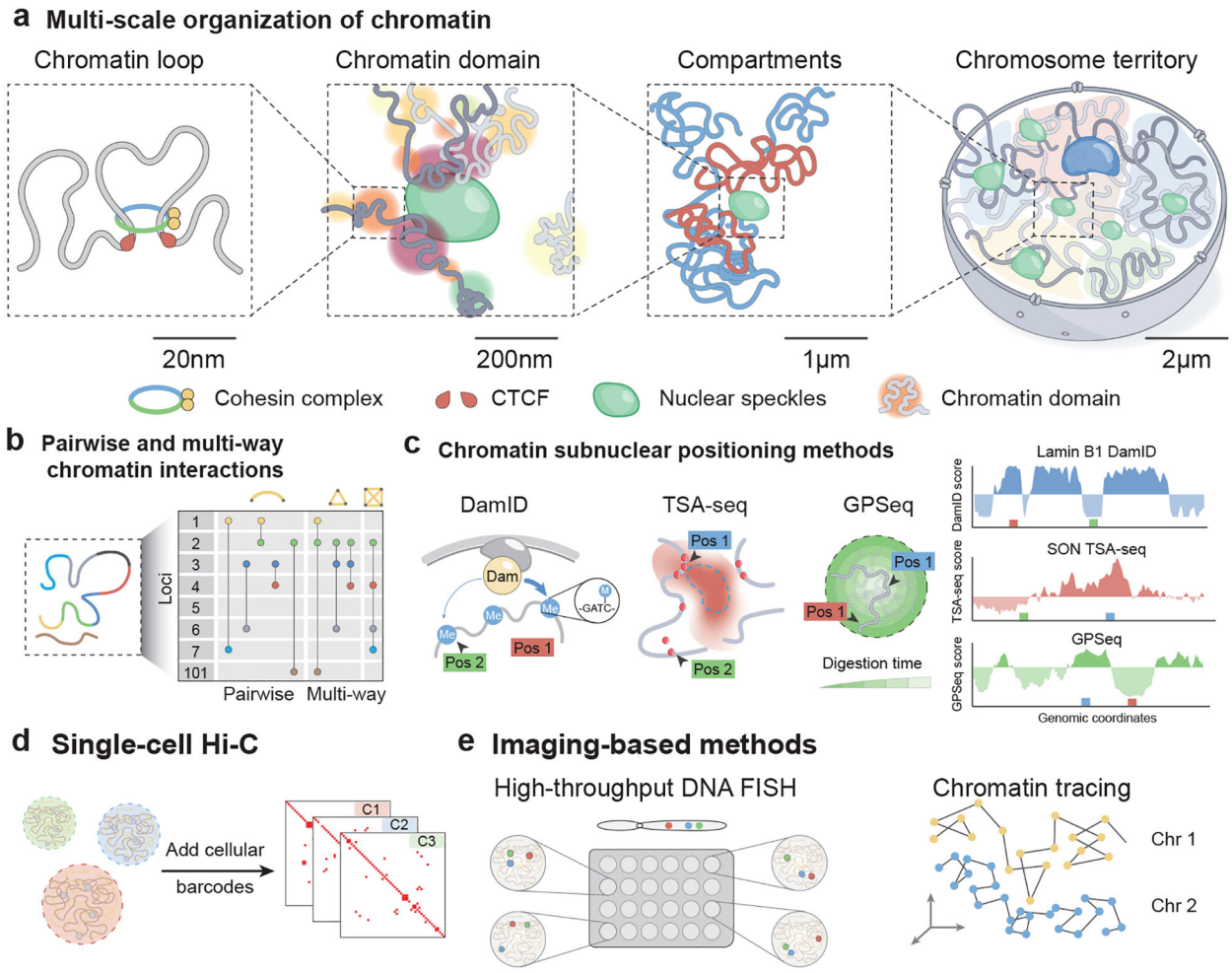

Figure 1.

Overview of multiscale 3D genome features and assays. **a.** Schematic view of multiscale 3D chromatin organization: DNA is packaged into chromosome territories, where it is intertwined with nuclear bodies such as nuclear speckles. Packaging is achieved through progressively finer resolution structural motifs such as compartments (megabase scale), chromatin domains such as TADs (100kb to a few Mb) and chromatin loops (10kb to 100kb apart, mediated by architectural proteins such as cohesin and CTCF). **b.** Experimental methods such as Hi-C can be used to capture chromatin contact frequency between pairwise loci or across multiple loci. **c.** Ligation-free methods (such as DNA Adenine Methyltransferase Identification (DamID), Tyramide Signal Amplification sequencing (TSA-seq), and genomic loci positioning by sequencing (GPSeq)) measure distance and contact frequency relative to nuclear bodies or positioning within the nucleus. **d.** Single-cell Hi-C can be used to detect variation in chromatin interactions among cells in complex tissues, for example, through adding multiplexed barcodes to individual cells. **e.** Multiplexed DNA Fluorescence In Situ Hybridization (DNA-FISH) provides direct spatial location of DNA loci and traces chromatin conformations in the nucleus.

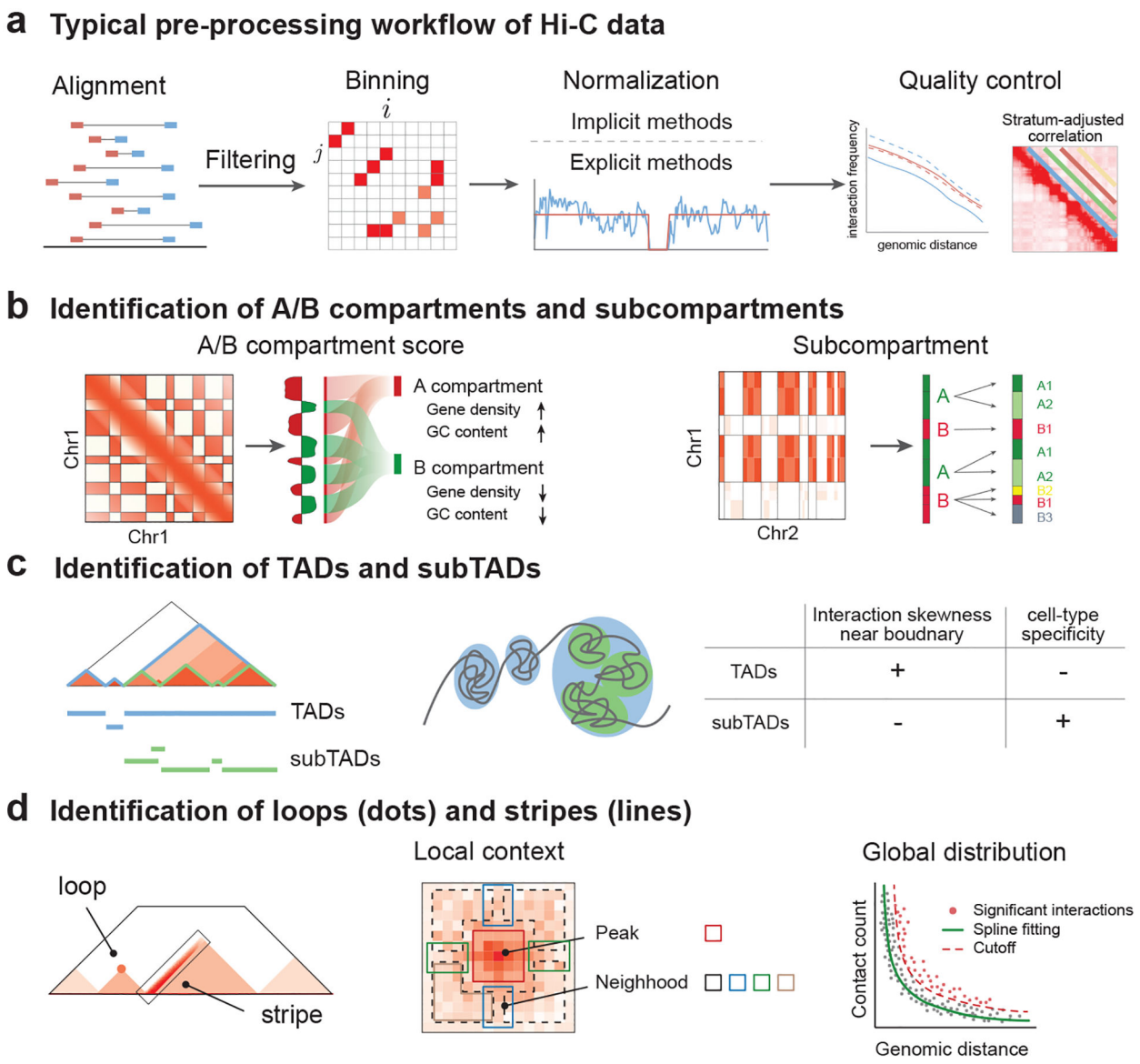

Figure 2.

Computational methods for identifying 3D genome features from Hi-C data. **a.** First, Hi-C reads are aligned onto the reference genome, filtered, and invalid read pairs are removed. Bins of equal size are created and valid interactions are assigned to generate the raw contact frequency map, which is further normalized and assessed by various quality control metrics. The quality of the Hi-C contact frequency map can be evaluated through a distance-dependent contact frequency decay curve or stratum-adjusted correlation based on distance-based stratification. **b.** A/B compartments are identified from the 2D contact frequency maps. Applying Principal Component Analysis (PCA) on the Pearson correlation matrix calculated from the observed over expected (O/E) matrix allows genomic bins to be assigned to A/B compartments, which can be further separated into sub-compartments. **c.** Topologically Associating Domains (TADs) are domain structures along the diagonal of the Hi-C contact frequency map. Some TADs have subTADs nested within a meta-TAD,

and they may show partially overlapping structures (left). The schematic chromatin structure corresponding to the contact frequency map is shown as a cartoon (middle). TAD and sub-TADs have distinct characteristics (right). **d.** Loops and stripes are fine-scale structures on contact frequency maps. Two different approaches are used to identify significant chromatin interactions (left). The first approach selects strong chromatin interactions by comparing contact frequency with neighboring bins on the contact frequency map (center). The second approach fits a global distribution between contact frequency and 1D distance and selects significant chromatin interactions as outliers based on the fitted model (right).

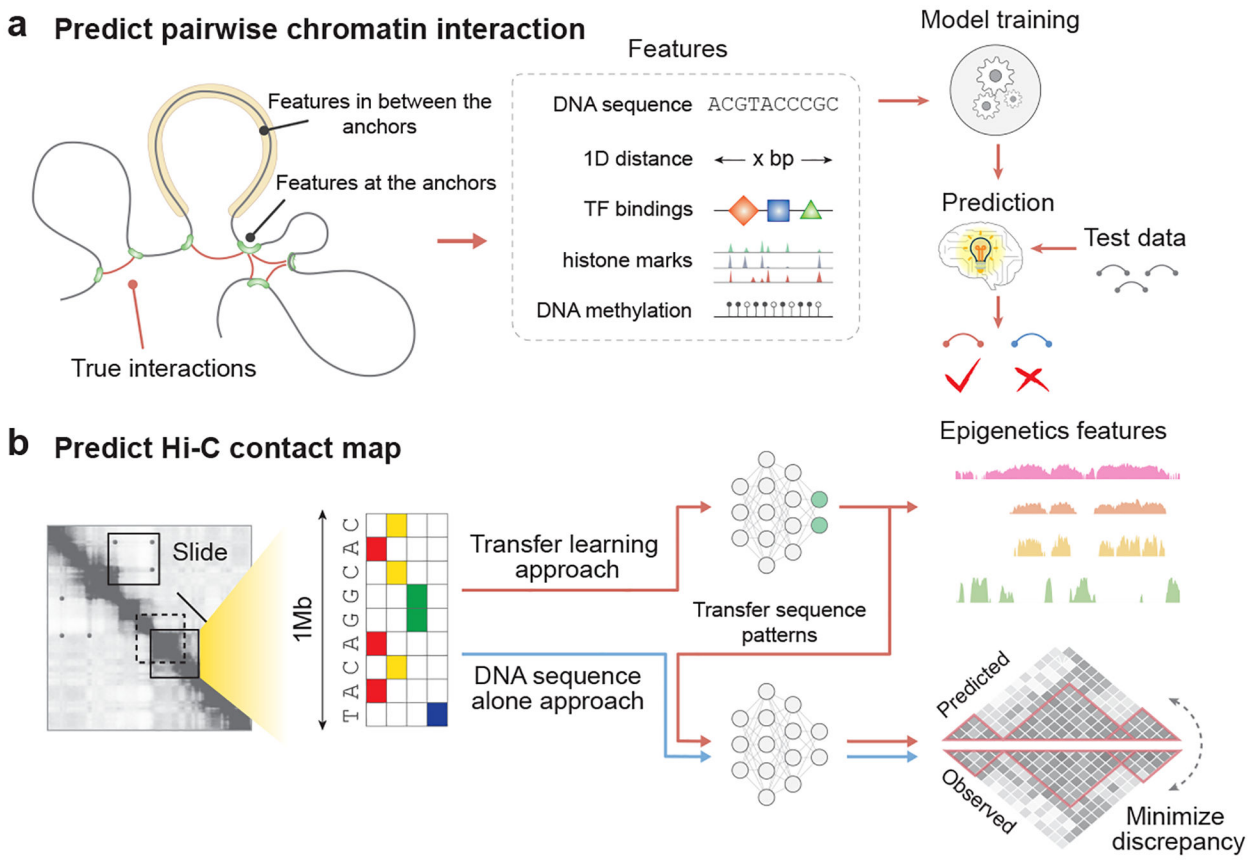

Figure 3.

Machine learning-based approaches for predicting 3D genome features. **a.** Significant chromatin interactions, such as CTCF-CTCF loops and enhancer-promoter interactions, can be predicted using machine learning methods that take both sequence and epigenomic signals at loop anchors and features between anchors as input. A supervised method is trained on a subset of true interactions and evaluated on unseen test data. TF, transcription factor. **b.** Deep neural network models predict genome-wide contact frequency between loci using large stretches of DNA sequence context as input. Some methods also pre-train the model by predicting 1D epigenetic signals from DNA sequence and then transferring the learned DNA feature representations to a second model to predict 2D contact frequency maps.

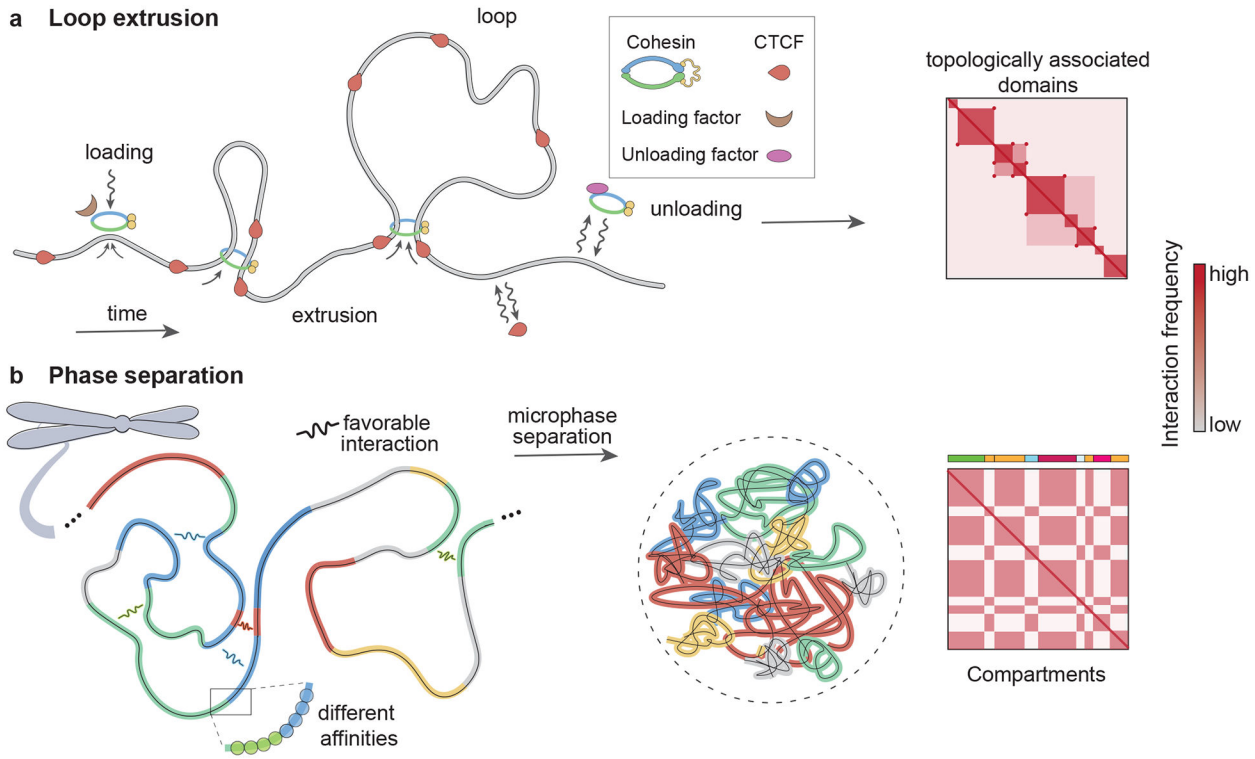

Figure 4.

Incorporation of mechanisms of genome folding in modelling approaches. **a.** The process of loop extrusion is shown, whereby a cohesin molecule attaches to the chromatin fiber and starts extruding it into a loop; the process stops when cohesin falls off or encounters another cohesin or a bound CTCF protein. Loading and unloading factors facilitate the process. Loop extrusion accounts for both loops and Topologically Associating Domains (TAD) observed in Hi-C contact frequency maps. **b.** The mechanism underlying phase separation is shown. Chromatin segments with different affinities (represented by different colors) micro-phase separate within the nucleus owing to attractive interactions between regions of the same affinity class, spatial restraints from the polymer chain, and competition with other interactions. This mechanism accounts for chromatin compartmentalization as observed in the characteristic Hi-C contact frequency map checkerboard pattern.

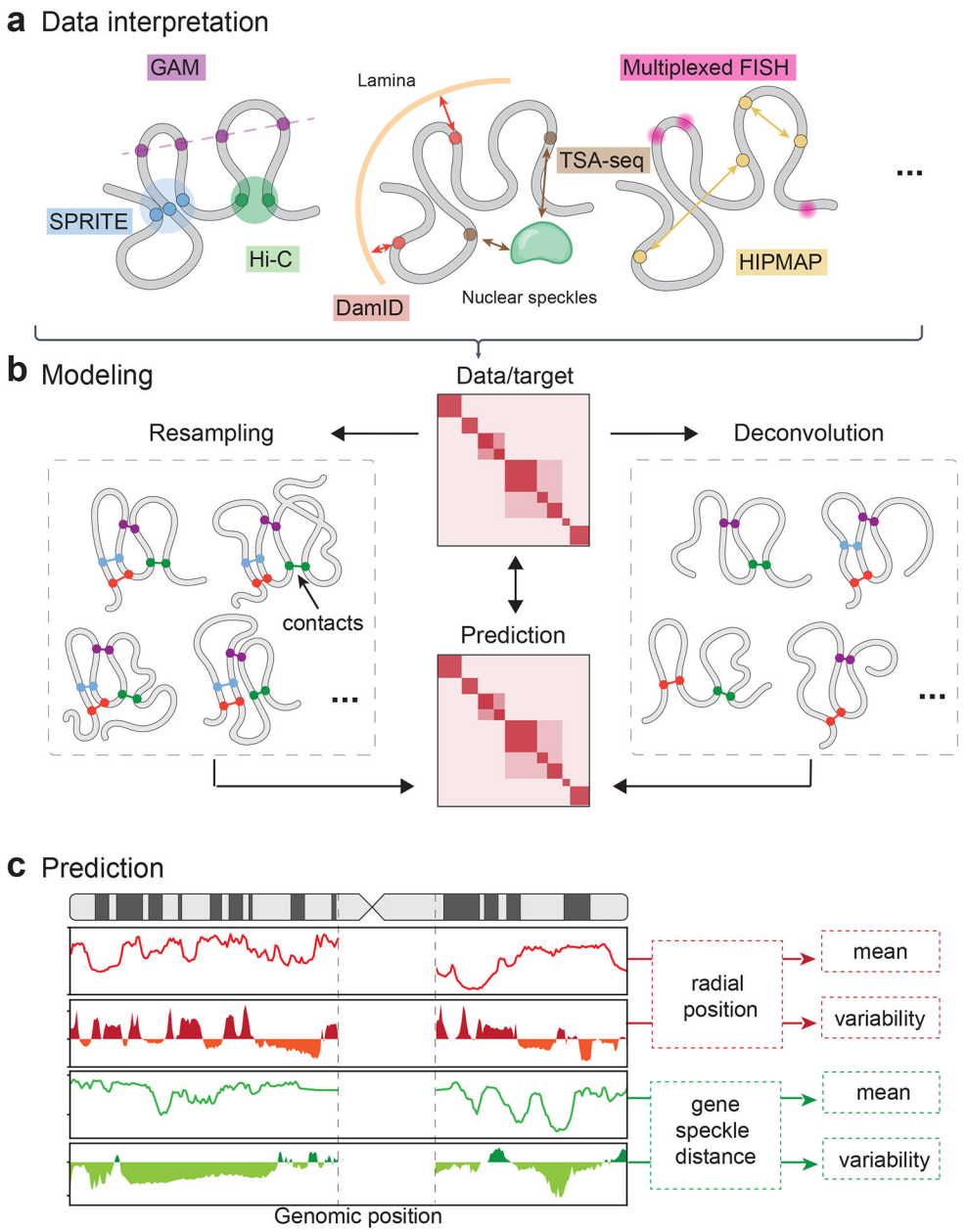

Figure 5.

Data-driven genome modelling methods. **a.** Spatial information from data generated by a variety of 3D genome mapping methods, such as Hi-C, Tyramide Signal Amplification sequencing (TSA-seq), and imaging data, is used to simulate 3D genome structures by minimizing the deviation of the model prediction from the experimental input. **b.** The rationale behind the resampling (left) and deconvolution (right) modelling methods is shown using ensemble Hi-C data as experimental input. In the resampling approach, the same set of contacts (shown as coloured dumbbells) is expressed in all the sampled structures. By contrast, in the deconvolution method, different batches of contacts are allocated into different structures. **c.** The simulated structures are used to compute 3D genome features that comprehensively characterize the local and global nuclear microenvironment

of all chromatin loci, such as the average radial position and distance to speckle of each chromatin locus and their variabilities across the ensemble of single-cell models. SPRITE, Split-Pool Recognition of Interactions by Tag Extension; GAM, Genome Architecture Mapping; DamID, DNA Adenine Methyltransferase Identification; FISH, Fluorescence In Situ Hybridization; HIPMap, High-throughput Imaging Position Mapping.

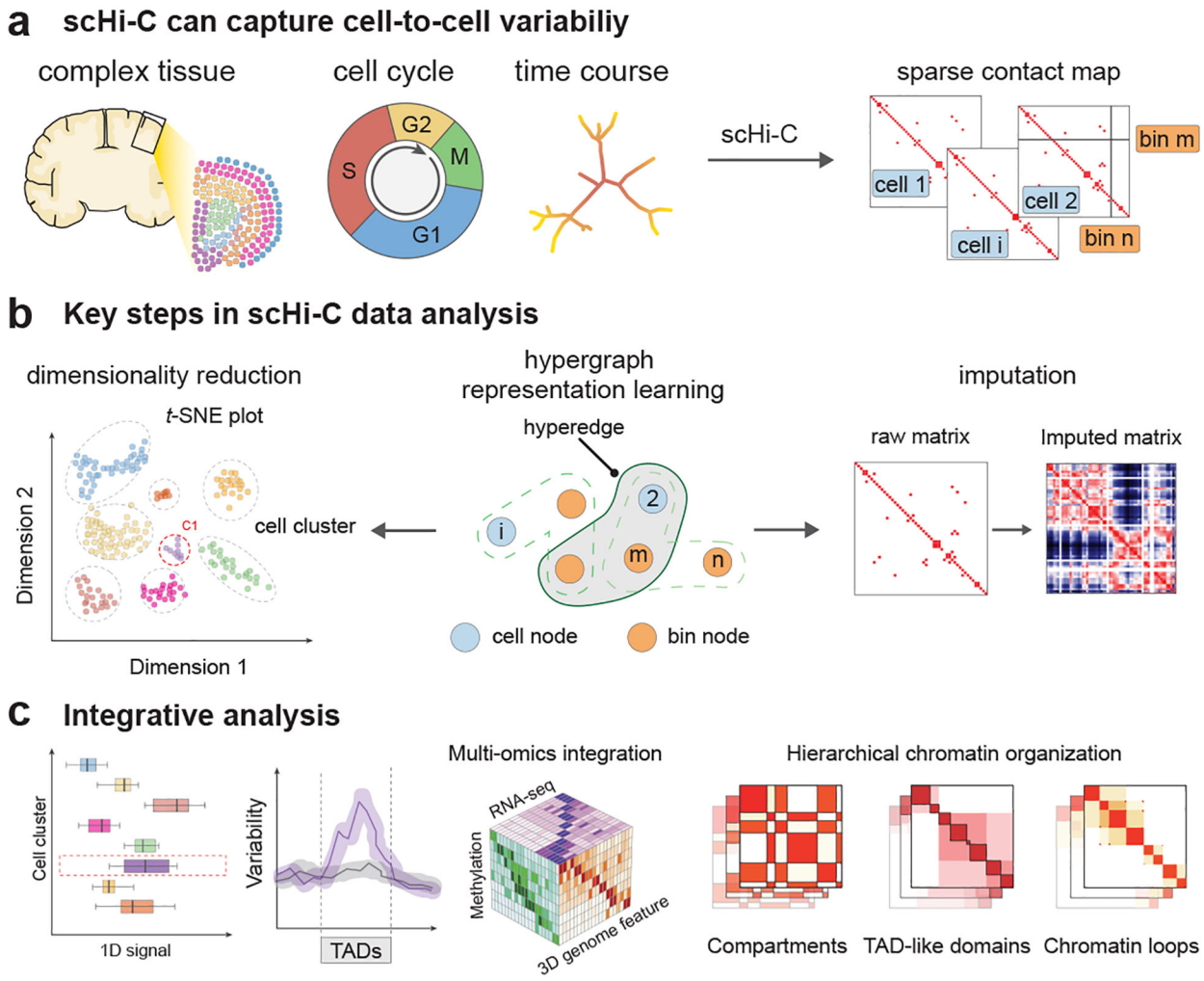

Figure 6.

A typical workflow for processing and analyzing single-cell Hi-C (scHi-C) data. **a.** scHi-C data provides insights into cell-to-cell variability and temporal cellular processes by separating sequencing reads into individual cells based on cellular barcodes. **b.** Computational methods are used to transform the contact frequency map into lower dimensional space (embeddings), impute missing values, and enhance data quality. Hypergraph representation learning (middle) can perform embedding and data imputation jointly. **c.** Using the embeddings, downstream analyses can reveal cell types (through clustering), multiscale 3D genome features, heterogeneity and variability of 3D genome organization, and association with other cellular processes such as DNA methylation. TAD, topologically associating domain