SCIENTIFIC REPORTS

OPEN

Hierarchical Reconstruction of High-Resolution 3D Models of Large Chromosomes

Received: 23 September 2018
Accepted: 7 March 2019
Published online: 21 March 2019

Tuan Trieu, Oluwatosin Oluwadare & Jianlin Cheng

Eukaryotic chromosomes are often composed of components organized into multiple scales, such as nucleosomes, chromatin fibers, topologically associated domains (TAD), chromosome compartments, and chromosome territories. Therefore, reconstructing detailed 3D models of chromosomes in high resolution is useful for advancing genome research. However, the task of constructing quality high-resolution 3D models is still challenging with existing methods. Hence, we designed a hierarchical algorithm, called Hierarchical3DGenome, to reconstruct 3D chromosome models at high resolution (<=5 Kilobase (KB)). The algorithm first reconstructs high-resolution 3D models at TAD level. The TAD models are then assembled to form complete high-resolution chromosomal models. The assembly of TAD models is guided by a complete low-resolution chromosome model. The algorithm is successfully used to reconstruct 3D chromosome models at 5 KB resolution for the human B-cell (GM12878). These high-resolution models satisfy Hi-C chromosomal contacts well and are consistent with models built at lower (i.e. 1 MB) resolution, and with the data of fluorescent *in situ* hybridization experiments. The Java source code of Hierarchical3DGenome and its user manual are available here https://github.com/BDM-Lab/Hierarchical3DGenome.

The architecture of chromosomes and genomes is important for cellular function[1–3]. However, the principles governing the folding of chromosomes are still poorly understood. The traditional microscopy technique - Fluorescent *in Situ* Hybridization (FISH) has been used to study chromosome architecture, but is limited by its low resolution and low throughput[4–7]. Chromosome conformation capture (3C) techniques like Hi-C[1] and TCC[2] can capture interactions between chromosomal fragments, and quantify the number of interaction frequencies (IFs) between them at a specific resolution. The bigger the interaction frequency between two fragments, the higher the probability that they are close in the three-dimensional (3D) space.

The interaction frequencies between pairs of chromosomal fragments are often summarized as a symmetric matrix, called contact matrix (or map). Contact matrices can be used to analyze the spatial organization of chromosomes or genomes. For instance, the chromosomal contact matrices have been used to confirm or identify the hallmarks of the human genome organization such as chromosome territories, chromosomal two-compartment partitions, chromatin loops, and topologically associated domains (TAD)[1,8,9]. Contact matrices can also be used to reconstruct 3D models of chromosomes and genomes to further facilitate the study of their organization. Various methods have been proposed to reconstruct 3D models of chromosomes or genomes[10–29]. On one hand, some of these methods utilize a function that approximates the inverse relationship between interaction frequencies (IFs) and spatial distances between fragments and then uses the distances as restraints to build 3D models via spatial optimization. These methods are called the optimization based method[10,14,17,24–26,29]. In the early work of Duan *et al.*[10], 3D models of a yeast genome were reconstructed to fit the Euclidian distances converted from IFs. On the other hand, some methods are designed to maximize the likelihood of a 3D model by using model-based methods that assumes that contact frequencies are related to distances via a probabilistic function. These methods use for example the Markov Chain Monte Carlo sampling technique to reconstruct 3D chromosome models by satisfying as many converted Euclidian distances as possible[11,12,27,28].

Most of the existing methods are capable of reconstructing chromosome or genome models of low resolution (e.g. 1 MB or 100 KB) from Hi-C datasets. They can also reconstruct the 3D models of a small region of a chromosome at a high resolution. For example, LorDG[17] built 3D models of 10 MB long chromosomal fragments at 10 KB resolution. LorDG solves a non-convex optimization problem to obtain coordinates of loci so that it can generate

Department of Electrical Engineering and Computer Science, University of Missouri-, Columbia, MO, 65211, USA. Correspondence and requests for materials should be addressed to J.C. (email: chengji@missouri.edu)

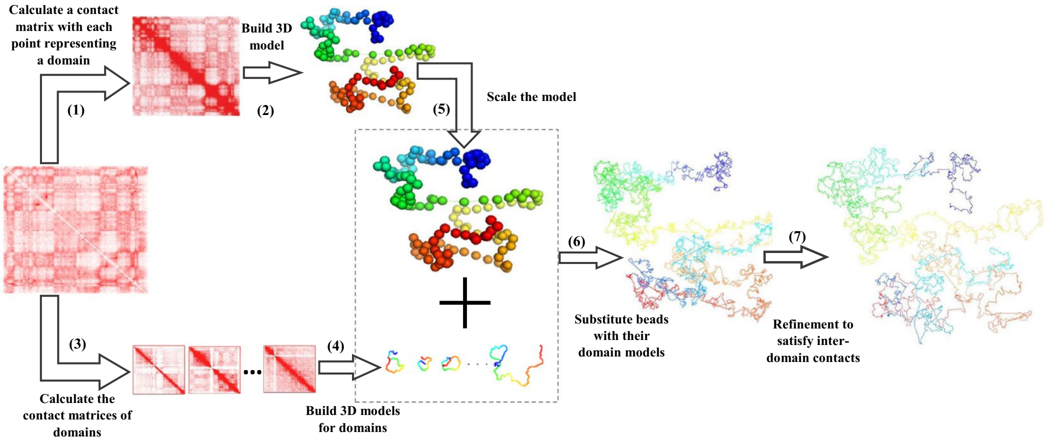
Calculate a contact matrix with each point representing a domain
(1)
Build 3D model
(2)
(5)
Scale the model
(6)
(7)
Substitute beads with their domain models
Refinement to satisfy inter-domain contacts
(3)
(4)
Calculate the contact matrices of domains
Build 3D models for domains

Figure 1. The steps of Hierarchical3DGenome algorithm to reconstruct high-resolution models of chromosomes. The seven steps that Hierarchical3DGenome takes to reconstruct high-resolution models of chromosomes are: (1) break a chromosome into chromosomal domains according to input data and represent each domain as a point or bead, (2) build a 3D model of the entire chromosome at low resolution, (3) create a contact matrix for each domain, such that for n domains there are n contact matrices, (4) build 3D model of high resolution for each domain, (5) scale the 3D Models of the entire chromosome at low resolution to match with the models of the domains at high resolution, (6) substitute beads in low-resolution models with their high-resolution domain models, and (7) refine the high-resoluiton models of the entire chromosome to satisfy inter-domain contacts.

different models corresponding with different local optimums. However, because LorDG put a high priority on satisfying contacts with high interaction frequencies, its models often satisfy a major set of contacts with high interaction frequencies and are similar to each other.

However, the ability to construct high-resolution (<=5 KB) 3D models of entire large chromosomes that are needed to study the detailed interactions between genes and regulatory elements, such as enhancers at the genome scale, are still out of reach for most, if not all, of the existing methods. More recently, Rieber, L. and Mahony, S[23], developed an approximate multidimensional scaling (MDS) algorithm called miniMDS, capable of constructing the 3D structure of chromosomes and genome at high resolution from Hi-C datasets better than most of the existing methods. This algorithm partitions a Hi-C dataset into subsets, performs high-resolution MDS separately on each subset, and then reassembles the partitions using low-resolution MDS. At the time of writing this manuscript, the miniMDS is the only known method that has attempted to build a relatively higher resolution 3D models of an entire chromosome.

High-resolution genome structure modeling has several challenges. Firstly, the structure sampling is much more intensive since the number of chromosomal fragments to be modeled is much larger in high resolution. For instance, the number of chromosomal fragments to be considered at 5 KB resolution is 20 times as many as at 100 KB. Secondly, as the resolution increases, the number of contacts between fragments gets smaller, leaving less contact data for restraining the positions of the fragments. Finally, the search space for high-resolution models is much larger than low-resolution models, making spatial optimization much more complicated. And due to the substantial increase of the model space, models with different topologies in high resolution may satisfy the same chromosomal data. One way to reduce the search space is to require that high-resolution models have a structural topology similar to that of low-resolution models whose structure can be more stably constructed due to the availability of more contact data between larger fragments. In this work, we introduce a hierarchical algorithm to build high-resolution 3D chromosome models at 5 KB resolution by using low-resolution models at 1 MB resolution to assemble high-resolution models of chromosomal domains into full high-resolution models of entire chromosomes. Our results show that the high-resolution chromosome models reconstructed by our method satisfied input chromosomal contacts well, and are consistent with the experimental FISH data.

Methods

Data source.

We used the Hi-C contact matrices datasets (GEO Accession Number: GSE63525)) of cell line GM12878[8] for all analyses.

Normalization.

In this work, we used two normalizations in two modelling processes. The first one is Knight–Ruiz normalization (KR) method[30] for building high resolution model of individual domains (step 4 in Fig. 1). The second one is iterative correction and eigenvector decomposition (ICE) normalization[31] method for building the low-resolution model of the entire chromosome (step 2 in Fig. 1), where each domain is represented by a point or bead.

Overview of the algorithm.

A chromosome is modeled as a string of beads in 3D space, where each bead denotes the midpoint of a DNA fragment at a specific resolution (e.g. 5 KB long). The position of a bead is then represented by three coordinates (*x*, *y*, *z*). Interaction frequencies between beads *i*, *j* are converted into spatial

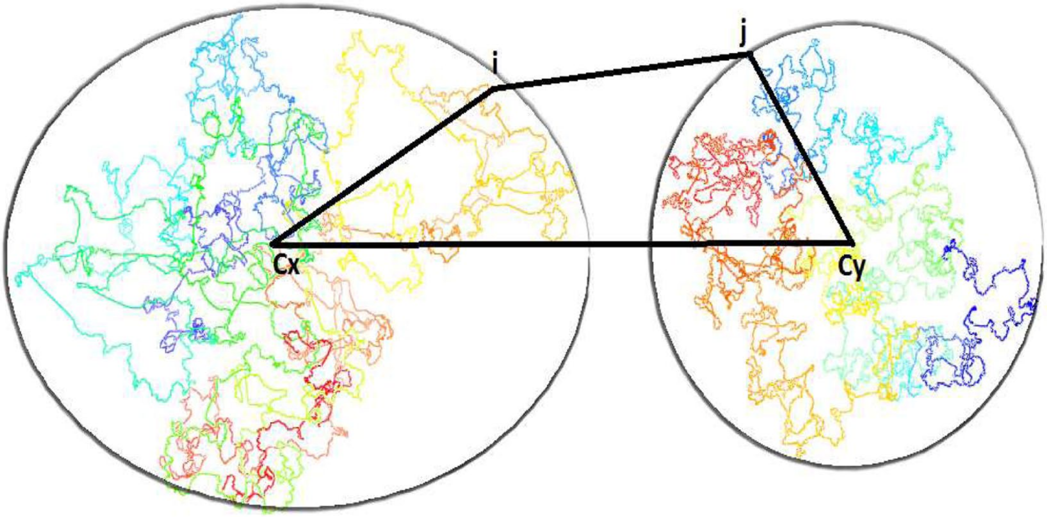

Figure 2. Estimation of the distance between the center of two domain models. An illustration of how the distance between centers (c_x, c_y) of the mass of two adjacent domain models (x, y) is estimated, where i and j are fragments in domains x and y respectively.

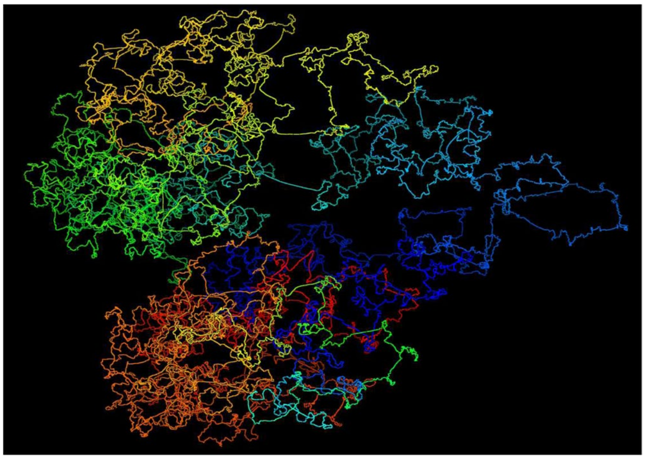

Figure 3. A 3D structure of chromosome 11 of the cell line GM12878 at 5 KB resolution. It shows a 3D model of Chromosome 11 represented by 26,065 points at 5 KB resolution for the GM12878 cell line Hi-C dataset.

distances by the function $d_{ij} = \frac{1}{IF_{ij}^{\alpha}}$, where IF_{ij} and d_{ij} are the interaction frequency and approximate distance (expected distance) between bead i and j, with α as a conversion factor. The goal is to place beads in 3D space so that their pairwise distances satisfy the expected distances converted from interaction frequencies as well as possible.

Our algorithm first reconstructs the 3D model of a chromosome at low resolution, which is used to guide the search for optimal models at high resolution. Each fragment (or point) in low-resolution models represents a contact domain, which is considered a structural unit of chromosome[8]. A chromosomal domain has substantially more contacts within itself than with other domains. Therefore, the accurate models of each chromosomal domain at high-resolution can be reconstructed individually. The models of individual domains are then assembled together according to the overall topology of full chromosomal models at low resolution.

Specifically, our hierarchical algorithm, Hierarchical3DGenome, constructs high resolution chromosome 3D models in seven steps (Fig. 1). The input is a contact matrix of a chromosome at a high-resolution (e.g. 5 KB). In Step 1, the chromosome is partitioned into contact domains (or topologically associated domains (TADs)) using the arrowhead domain algorithm[8]. When a domain contains small domains inside, only this domain is considered because the small domains have been represented by it. Then, each separate domain is represented by a point or bead and the interaction frequencies between beads are calculated to make a low-resolution contact matrix for the entire chromosome. The matrix is normalized by the iterative correction and eigenvector decomposition (ICE) normalization[31] to remove technical biases[32], biological factors[33] and the different visibility of beads due to their different lengths. This new contact matrix is used to build a low-resolution model of the entire chromosome using our in-house method LorDG[17] in Step 2. We used the default parameter settings in LorDG. The default

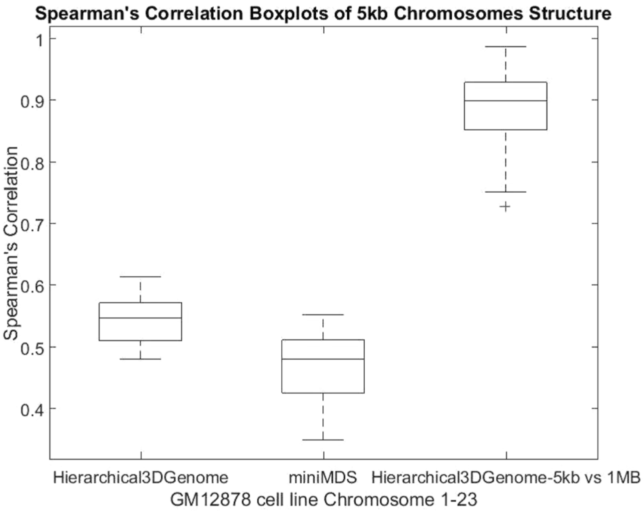

Figure 4. Spearman's correlation boxplots of 5 kb chromosome structures built by Hierarchical3DGenome and miniMDS for GM12878 cell line Chromosome 1–23. The box plots of the Spearman's correlation between reconstructed distances and expected distances for all 23 pairs of human chromosomes of Hierarchical3DGenome, miniMDS, and the consistency check of 5 KB-resolution Hierarchical3DGenome models with the models built directly from contact matrices of 1 MB resolution.

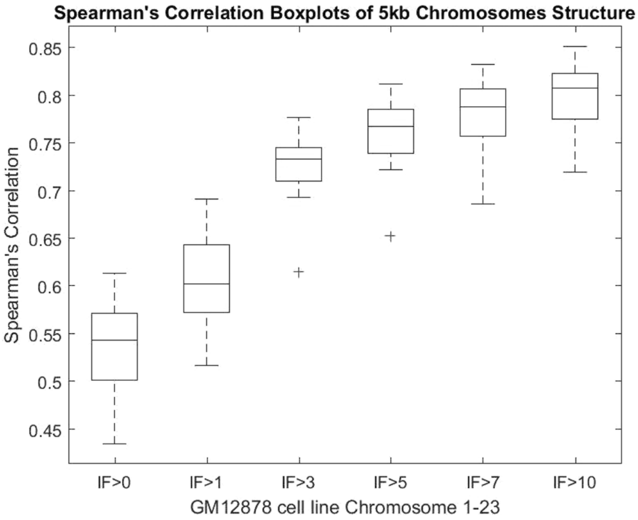

Figure 5. Models' Spearman's correlation at increasing IF cut-off thresholds. Box plots of the Spearman's correlation between the contact maps derived from chromosomal models and the input Hi-C data contact map for each chromosome at increasing IF cut-off threshold 0, 1, 3, 5, 7, 10. Contacts below each threshold are removed and not used in the Spearman correlation calculation for each chromosome. The Spearman's correlation increases with respect to the increase of the cut-off threshold.

parameter setting for LorDG algorithm allows it to search for the best conversion factor within the range [0.1, 3.0] with a step-size of 0.1 for an input contact matrix of a chromosome. The high-resolution contact matrices of individual domains are also extracted from the full high-resolution contact matrix of the chromosome in Step 3. The 3D models of each domain at the high-resolution are reconstructed individually in Step 4 (see the detailed description in the Sub Section "Construction of High-Resolution Models for each Domain").

The topology of a correct high-resolution model of a chromosome should be similar to that of its correct low-resolution model, even though it is not guaranteed that they are in the same scale. So, in Step 5, the low-resolution model of the full chromosome constructed in Step 2 is scaled by a ratio so that it can be used to guide the assembly of the high-resolution models of individual domains into a final high-resolution model of the entire chromosome. The ratio used for this scaling is estimated from the models of individual domains and the low-resolution model of the entire chromosome (see the detailed description in the Sub Section "Estimating the Scaling Ratio between High-Resolution and Low-Resolution Models").

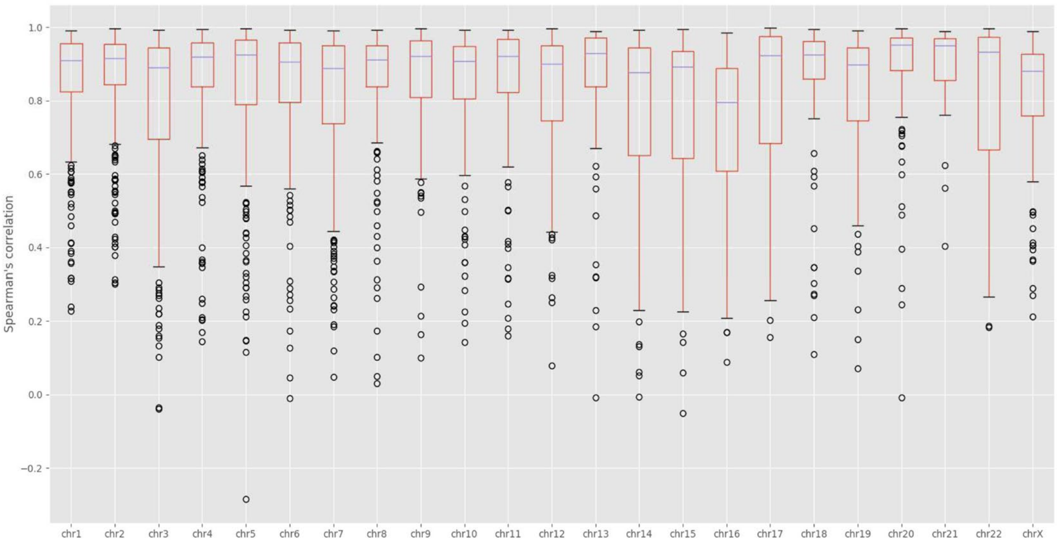

Figure 6. Boxplots of Spearman's correlations between models of two adjacent domains. The box plot of similarity scores between models of two adjacent domains reconstructed individually and those extracted from the full chromsome model at 5 KB resolution. Similarity is measured as Spearman's correlation between distances from models. The average correlations are high. The result suggests that inter-domain contacts were well satisfied.

After the low-resolution model is scaled to match with the scale of high resolution, in Step 6, each bead of the low-resolution model is substituted by a high-resolution model constructed in Step 4 for the corresponding domain that the bead represents. Finally, Step 7 is to further refine the location of the models of the domains to satisfy more inter-domain contacts (see the detailed description in the Sub Section "Model Refinement").

Construction of High-Resolution Models for Each Domain. To build high-resolution models from the contact matrix of a domain, a good conversion factor (α) to translate interaction frequencies into spatial distances is needed because the conversion function $\left(d_{ij} = \frac{1}{IF_{ij}^{\alpha}}\right)$ plays a crucial role in determining the quality of reconstructed models. We used the LorDG[17] algorithm to search for the best conversion factor within the range [0.1, 3.0] with a step-size of 0.1 for an input contact matrix. Each domain could have a different conversion factor, therefore, the median of conversion factors from all the domains (e.g. $\alpha = 0.9$) was selected as the consensus conversion factor to translate interaction frequencies into spatial distances to build high resolution models of domains with LorDG.

Estimating the Scaling Ratio between High-Resolution and Low-Resolution Models. To estimate the ratio to scale the low-resolution model to match with the high resolution model, the distance between centers (c_x, c_y) of the mass of two adjacent domain models (x, y) is estimated by the following formula: $d_{xy} = \min_{\forall i,j}\left(d_{ij} + d_{xi} + d_{yj}\right)$, where d_{ij} is the distance between fragment i in domain x and fragment j in domain y, d_{xi} is the distance between c_x and i and d_{yj} is the distance between c_y and j (Fig. 2). The rationale is that d_{xy} is always less than or equal to $d_{ij} + d_{xi} + d_{yj}$. Therefore, d_{xy} can be well approximated by $\min_{\forall i,j}\left(d_{ij} + d_{xi} + d_{yj}\right)$ given a sufficient number of fragments i and j are tested. It is worth noting that the adjacent domains are chosen because they have a high enrichment of inter-domain contacts between domains, hence, the distance estimated between them is more reliable.

Given the 3D models of domains, the distances d_{xi} and d_{yj} can be calculated from the coordinates of the centers (i.e. the average of the coordinates of the domain model (x, y)), the fragment i in domain x, and the fragment j in domain y. The distance d_{ij} can be calculated from the formula $d_{ij} = \frac{1}{IF^{\alpha}}$ using the interaction frequency (IF) between fragments i and j according to the conversion factor (α) found in Step 4.

The distance between the centers of two adjacent domains, d_{xy}, calculated above, are then divided by their corresponding distance in the low-resolution model to obtain a scaling ratio. In total, there will be $n - 1$ estimated distance ratios where n is the number of domains or beads in the low-resolution model. The final ratio used to scale the low-resolution model is the median of these estimated ratios. The centers of mass of the high-resolution domain models are placed at the locations of the corresponding beads (or points) of the low-resolution model in order to obtain an initial high-resolution model of the entire chromosome for further refinement.

Model Refinement. In the refinement step (Step 7), we used the LorDG algorithm to adjust the coordinates of all the points of the initial high-resolution models of all the domains to satisfy high-resolution chromosomal contacts. Starting from the initial, unrefined model produced in Step 6, both intra-domain and inter-domain

Chr.	L1 (position, MB)	L2 (position, MB)	L3 (position, MB)	L1-L2	L2-L3
11	130.72–130.75	130.29–130.32	129.86–129.89	0.8	7.1
13	86.37–86.40	85.46–85.49	84.55–84.58	1.9	13.9
14	71.60–71.63	72.20–72.23	72.80–72.83	1.2	7.8
17	66.76–66.79	67.22–67.25	67.68–67.71	1.2	9.7

Table 1. Distances between three fluorescence *in situ* hybridization (FISH) probes in the models of 5 KB resolution for Chromosomes 11, 13, 14 and 17. The genomic positions of three probes (L1, L2, and L3) on the four loops of four chromosomes (Chr. 11, 13, 14 and 17) and the distances between these probes in the high-resolution model. Columns 2–4 list the start/end position of each probe. Columns 5 and 6 report the distances between the three probes.

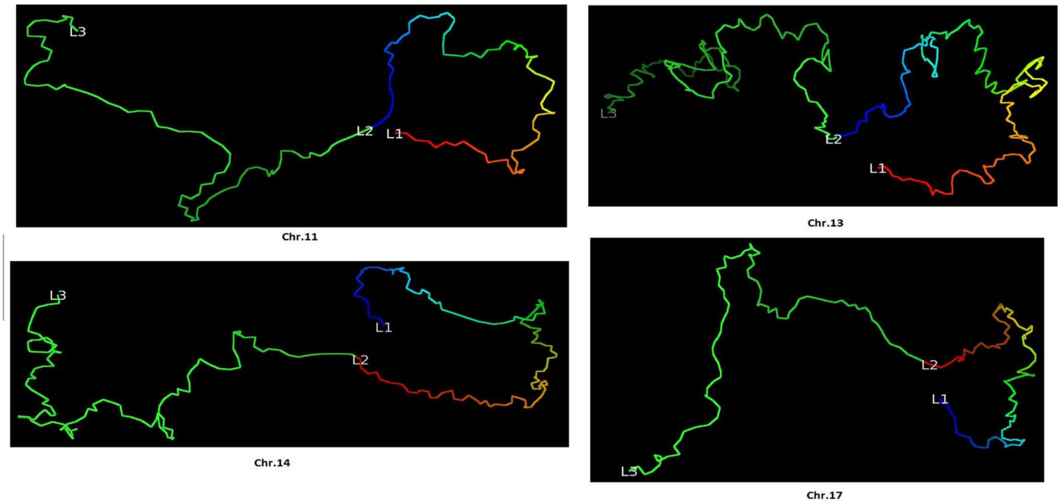

Figure 7. Visualization of the loop on a fragment of chromosome 11, 13, 14 and 17 models. The four loops validated by FISH data for four different chromosomes (Chr. 11, 13, 14 and 17 of cell line GM12878) in the models at 5 KB resolution.

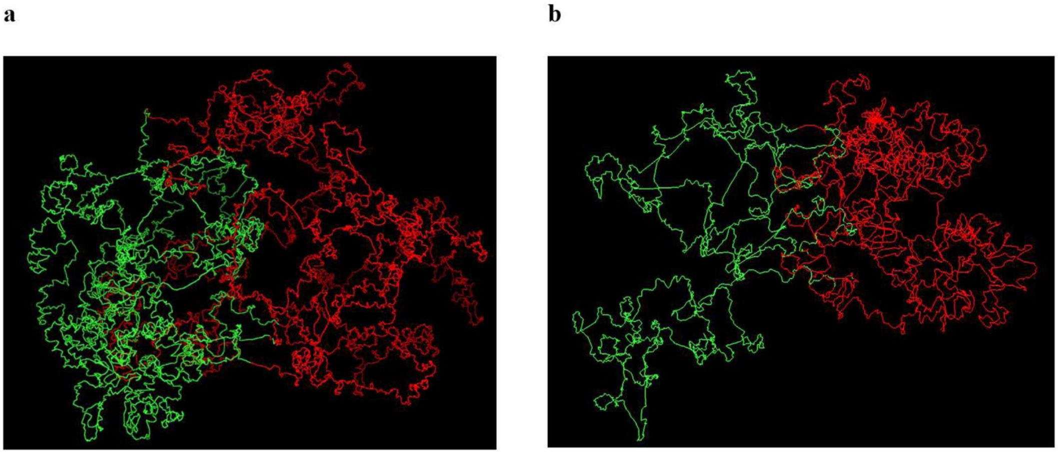

Figure 8. Two compartments in models of Chromosomes 19 and 21. The compartments were obtained from the principal component analysis performed on the contact maps and colored in red and green in the 3D models of (**a**) Chromosome 19 and (**b**) Chromosome 21.

contacts are used in the optimization to refine it. LorDG uses all contacts to adjust the model to maximize the satisfaction of the contacts. The objective function of LorDG is non-convex and its optimization converges at local optimums. Therefore, the intra-domain contacts that have higher interaction frequency than inter-domain contacts and have already been well satisfied in the initial model are mostly preserved during the optimization. The optimization in the refinement step mostly tries to satisfy more inter-domain contacts to assemble domain models together.

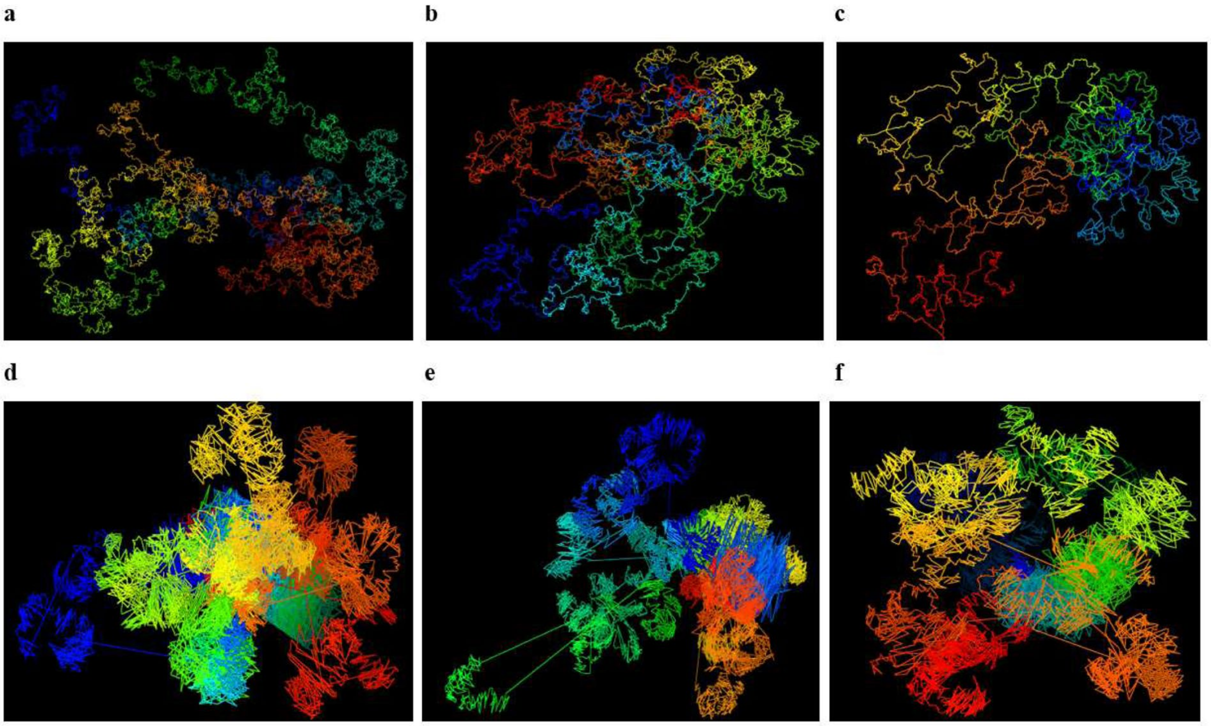

Figure 9. Comparison of 3D chromosome models of Hierarchical3DGenome and miniMDS. Models of chromosomes 16,19, and 21 of 5 KB resolution produced by Hierarchical3DGenome (**a**–**c**) and miniMDS (**d**–**f**).

Results

We used the high-resolution Hi-C dataset[8] with our method to build and evaluate the chromosome models at 5 KB resolution. Figure 3 shows a 3D model of Chromosome 11 represented by 26,065 points at 5 KB resolution, which is a 3D model of a human chromosome of the highest resolution to the best of our knowledge. We conducted five tests to evaluate the quality of these models. Firstly, we calculated the correlation between the fragment-fragment distances in the models and the expected distances calculated from contact matrices. Secondly, we checked if our 5 KB-resolution models were consistent with models reconstructed directly from contact matrices at 1 MB resolution. Thirdly, to figure out if domain models were well adjusted to satisfy inter-domain contacts, we extracted contact sub-matrices for every two adjacent domains consisting of both inter- and intra-domain contacts, reconstructed 3D models of the two adjacent domains from the matrices, and then compared them with the corresponding models of the two domains extracted from the full-chromosome models at 5 KB resolution. Fourthly, we investigated if the high resolution chromosomal models were consistent with the FISH data[8]. The comparison shows that our models exhibited the 4 loops on four different chromosomes that were identified from the FISH data. Finally, we compared our method with an existing method for high-resolution modeling.

Correlation between Reconstructed Distances and Expected Distances.

We calculated the Spearman's correlation between reconstructed fragment-fragment distances in the 3D models and their expected distances derived from the input contact matrices (Figs 4 and S1). The average and standard deviation of the correlations are 0.5357 and 0.0397, respectively. Considering the large number of expected distances to be correlated for each chromosome (e.g. 42,000,000 expected distances for Chromosome 1) and a lot of noise and inconsistency in these distances, these correlation values are good and suggest the 3D structures reconstructed models are of reasonable quality. In addition, we performed a comparison of the contact maps derived from each model with the input Hi-C data at varying contact cut-off (Figs 5 and S2). When contacts with low interaction frequencies are removed, the correlations are better. This indicates that low interaction frequencies, unreliable contacts drive the correlations down, and our models put a high priority on satisfying reliable contacts with high interaction frequencies. The correlation increases as the cut-off threshold increases, which suggests an increase in the model consistency with the Hi-C data as contacts with low interaction frequencies are removed.

Consistency between models of 5 KB resolution and 1 MB resolution.

At 1 MB resolution, because contacts often have high interaction frequencies and a fragment is in contact with more other fragments and therefore more restrained, the topology of models is generally more reliable and stable. We compare 5 KB-resolution models with the models built directly from contact matrices of 1 MB resolution to check if their topologies are consistent. To make this comparison, the 5 KB resolution models were zoomed out (reduced the resolution) to 1 MB resolution models. This was achieved by averaging coordinates of points of the same bin which are 1 MB long as in the 1 MB resolution models. We then calculated Spearman's correlation between

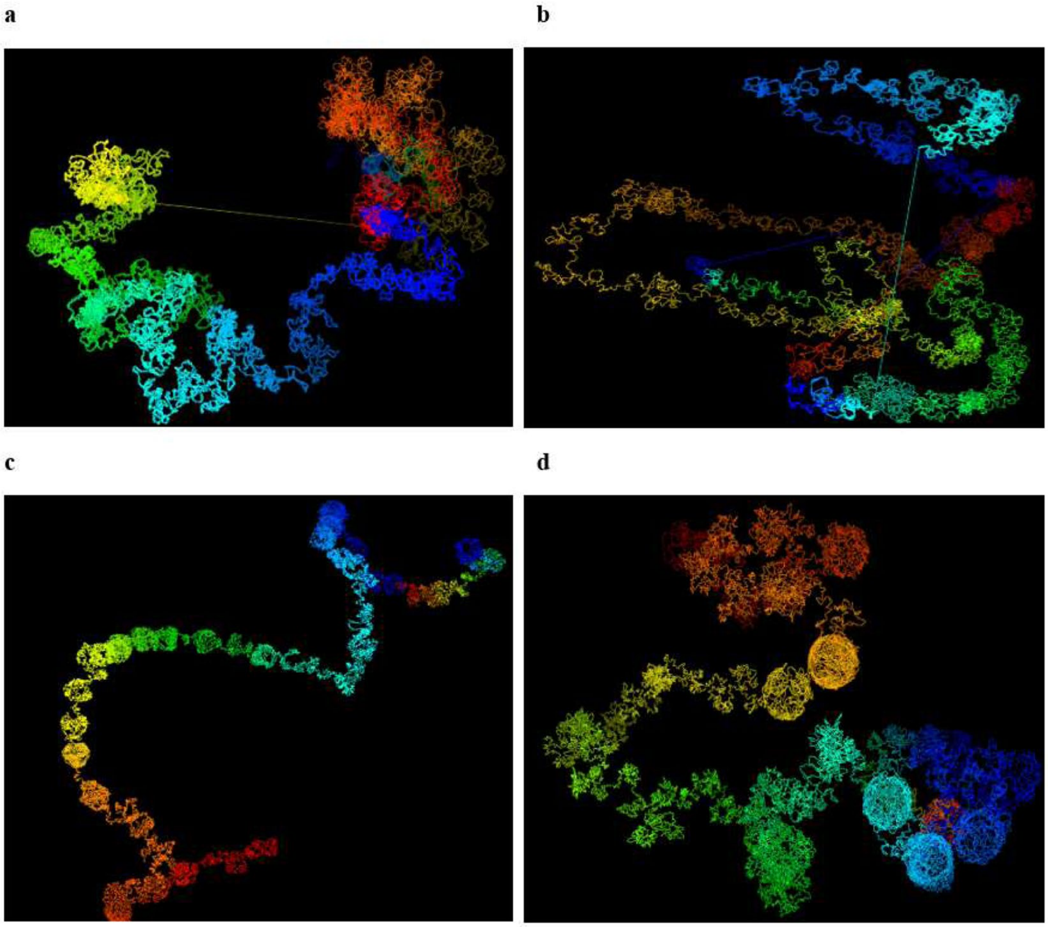

Figure 10. Visualization of 1 KB resolution models of Chromosome 1, 11, 13, and 14. 1 KB resolution structures of (**a**) Chromosome 1, (**b**) Chromosome 11, (**c**) Chromosome13, and (**d**) Chromosome14, generated by Hierarchical3DGenome.

Chr.	L1-L2 (IF)	L2-L3 (IF)	L1-L2 (Euclidean distance in models)	L2-L3 (Euclidean distance in models)
11	2.496299	0.815569	10.52133	47.190198
13	2.156871	0.924415	19.49644	121.037814
14	2.065476	0.892723	13.85555	43.714938

Table 2. The IF and distance between the three probes from the Hi-C data and 1 KB models of Chromosomes 11, 13 and 14, respectively. The genomic positions of three probes (L1, L2, and L3) on the four loops of chromosomes 11, 13, and 14, and the distances between these probes in the high-resolution 1 KB models. Columns 2 and 3 report the IF between the three probes from the Hi-C data, columns 4 and 5 report the distances between the three probes from the models. Distances in the models are consistent with IFs in Hi-C data, but they are not small enough to appear like loops.

pairwise distances of the zoomed-out model and the 1 MB resolution models (Figs 4 and S3). For all chromosomes, the correlations are >0.72, suggesting that models at 1 MB and 5 KB have the similar topologies.

Evaluation of inter-domain interactions.

To check if domain models were adjusted appropriately to satisfy inter-domain contacts, we built models of every two adjacent domains from their contact sub-matrices extracted from the full contact matrix of a chromosome. We compared these models with their counterparts extracted from 5 KB-resolution full-chromosome models. The boxplot in Fig. 6 shows Spearman's correlations between the models reconstructed individually and those extracted from the 5 KB resolution models for all chromosomes. The high average correlations suggest that the domains in the high-resolution full-chromosome models were generally well adjusted to satisfy inter-domain contacts.

Validation with FISH Data.

We validated our models against the Fluorescence *In Situ* Hybridization (FISH) data[8]. The FISH data identified four chromatin loops in four chromosomes of the cell line GM12878 (Chr.

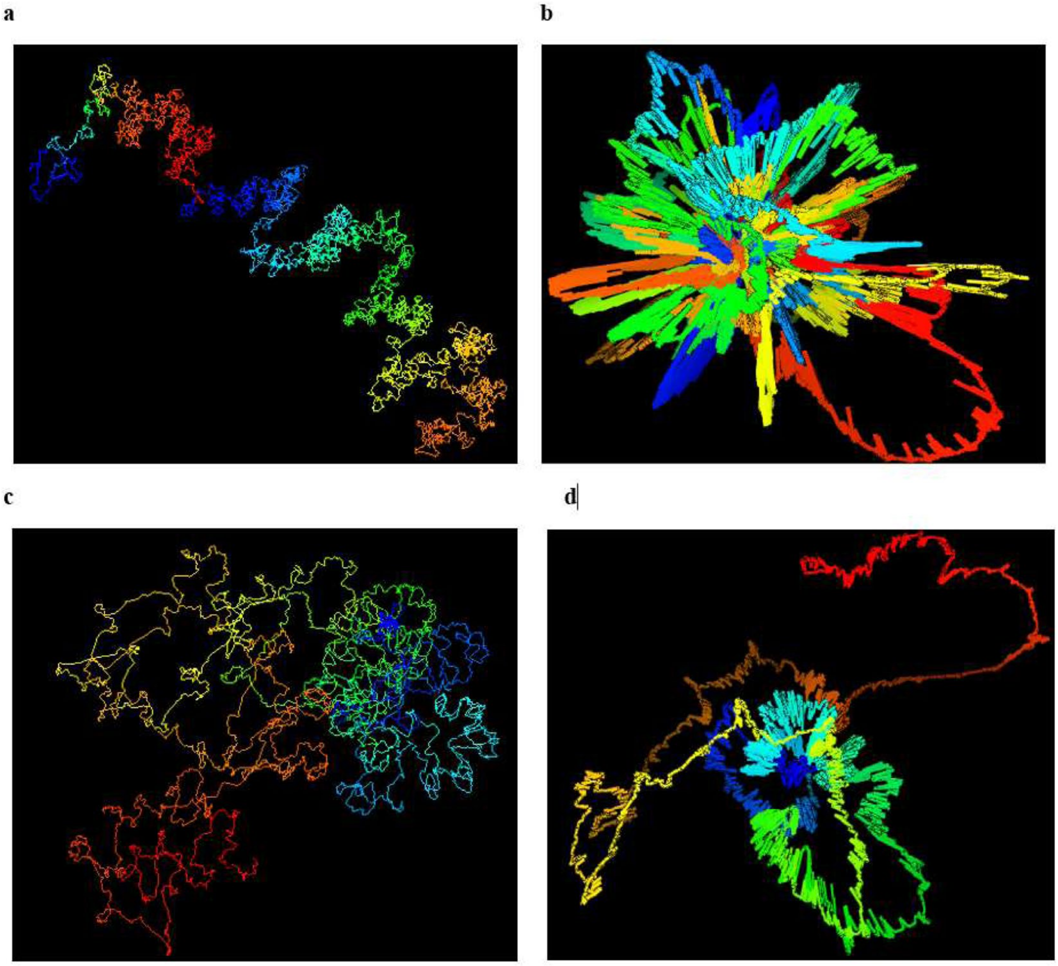

Figure 11. Comparison of the models built by Hierarchical3DGenome and LorDG without domain partitioning. A side by side comparison of the models of Chromosomes 4 and 21 of 5 KB resolution generated by Hierarchical3DGenome (**a**,**c**) and LorDG without domain partitioning (**b**,**d**).

11, 13, 14 and 17). For each loop, the distances between three probes (or loci), i.e., L1, L2 and L3 on the loop, were measured by FISH experiments. Although the three probes have the same genomic distance, the spatial distance of L1-L2 is much smaller than the spatial distance of L2-L3 according to the FISH experiment.

We calculated the spatial distances L1-L2 and L2-L3 in the 5 KB high-resolution models and confirm that the distance of L1-L2 is indeed much smaller than the distance of L2-L3, indicating that the loops involving L1 and L2 were correctly reconstructed in the 3D models. These distances between the probes are shown in Table 1. The four loops in our models are visualized and shown in Fig. 7.

Validation based on two-compartment feature of chromosomes. Previous studies[1] has shown that the chromosomes can be divided into separate compartments (euchromatin and heterochromatin). We performed the Principal Component Analysis (PCA) on the Hi-C contact matrices to divide the chromosome into two compartments. Afterwards, we annotated the identified compartments with different colors to assess how separable they are in 3D chromosomal models. As illustrated in Fig. 8, regions in the same compartments are spatially grouped together in the models. Figure 8 shows the two compartments of chromosomes 19 and 21.

Comparison with an existing high-resolution model construction method. We compared Hierarchical3DGenome with the state of the art high resolution model construction method, miniMDS[23]. We calculated the Spearman's correlation between input distances and distances inferred from the output 3D structure of each chromosome at 5 KB resolution produced by miniMDS. Only the non-missing points identified in the miniMDS output structure were used in this calculation. The default parameters of miniMDS were used. The result shows that the correlation is higher for Hierarchical3DGenome for every chromosome than miniMDS, indicating that Hierarchical3DGenome infers 3D structures that are more consistent with the input Hi-C data (Figs 4 and S4). The structure from miniMDS sometimes contains folds or cluttered points with unrecognizable chromosomal features and quite some fragments are missing. In comparison, the features in the structure of Hierarchical3DGenome are well distinguished (Fig. 9a–c versus Fig. 9d–f).

	Hierarchical3DGenome	LorDG Without Partition
1	0.898720584	0.074093704
2	0.933686968	0.025649295
3	0.938667819	0.074893363
4	0.988076596	0.013585756
5	0.875989854	−0.020093576
6	0.911444415	0.230877171
7	0.72885098	0.096554188
8	0.962162304	0.0460202
9	0.911241704	0.27890049
10	0.908920242	0.100821567
11	0.856633536	0.127690704
12	0.933434252	0.113935156
13	0.91778003	0.181649622
14	0.8897597	0.279229499
15	0.851429745	0.06654513
16	0.900103388	0.011566887
17	0.914467918	0.031162663
18	0.899153126	0.094189398
19	0.810298639	0.400262366
20	0.823568359	0.202652796
21	0.757115911	0.37725465
22	0.751419102	0.445408394

Table 3. Topology consistency check of the models produced by Hierarchical3DGenome and LorDG without domain partitioning. The Spearman's correlation between distances from models at 1 MB resolution and those at 5 KB resolution of Hierarchical3DGenome and LorDG without domain partitioning. Chromosomes models generated by Hierarchical3DGenome is much more consistent with models at 1 MB resolution that the models generated by LorDG without domain partioning.

Chromosome models at 1 KB resolution. We attempted to build chromosomal models at 1 KB resolution for Chromosomes 1, 11, 13, 14 and 17 to test the capability of our method for building models of even higher resolution (Fig. 10). These models satisfied input contacts reasonably well. Spearman's correlations between expected distances and reconstructed distances were 0.44, 0.45, 0.43, 0.48 and 0.53 respectively for these chromosomes. They were also similar to the models at 1 MB resolution, indicated by a high Spearman's correlations (>0.89) between them. However, the detailed shapes of the four loops identified by FISH experiments on chromosome 11, 13, 14 and 17 [8] in the models of 1 KB resolution did not appear as loop-like as they did in the models of 5 KB resolution. In particular, the distances between L1-L2 in the models is not small enough to appear like loops even though they are still smaller than the distances between L2-L3 in all cases as expected (Table 2). One possible reason is that the input chromosomal contact data is not dense enough to build the loops of 1 KB resolution, which were initially predicted with Hi-C data at 10 KB resolution prior to being verified by FISH experiments. Another reason could be that at 1 KB resolution, the increased level of noise and structural variance in the dataset reduced the prediction performance of the basic modeling tool (e.g. LorDG) used by the hierarchical modeling algorithm in this work. The problem in the former case can be solved if the Hi-C data of higher quality and resolution can be generated. In the latter case, the hierarchical modeling algorithm could be further improved by using a more robust, basic 3D modeler to build domain models. The two issues will be investigated in the future.

Computational requirement and performance. We assessed Hierarchical3DGenome on two server machines: a x86_64 bit Redhat-Linux server consisting of multi-core Intel(R) Xeon(R) CPU E7-L8867 @ 2.13 GHz with 120 GB RAM and a high-performance computing cluster (Lewis) with Linux. For each computational task performed on the Lewis cluster, we allocated 10 cores and 80 G memory. On average on both servers, Hierarchical3DGenome takes about four to eight hours for each chromosome model construction, even though running on the Lewis cluster is faster. To run Hierarchical3DGenome for structure reconstruction on local computers, readers may follow the instructions in the user manual.

Comparison of Hierarchical3DGenome with LorDG. We compared the accuracy and performance of Hierarchical3DGenome with LorDG without chromosome domain partitioning on the high-resolution data. Using a high-performance computing (HPC) cluster machine, we allocated 10 cores, 80G of memory, with a time limit of 2 days for each chromosome structure reconstruction task. Hierarchical3DGenome took 4–8 hours to reconstruct the 5 KB structure for each chromosome (Fig. 11(a,c)). In contrast, for LorDG without domain partitioning, no structural models can be constructed for Chromosomes 1 to 10, and 23 largely because they required more than 2 days for their structure construction to be completed. This indicates that LorDG without domain partitioning is much slower than Hierarchical3DGenome. Additionally, the structures constructed from

LorDG without domain partitioning is unrealistically condensed and cluttered, making it difficult to identify the relationship between chromosomal regions at this resolution (Fig. 11(b,d)).

Moreover, we performed a topology consistency check of the 5 KB models produced by the two methods with the 1 MB models built directly from contact matrices of 1 MB resolution (Table 3). At 1 MB resolution, most contacts have high interaction frequencies, and thus models are very reliable for serving as a reference. All 5 KB chromosome models of LorDG without domain partitioning have the spearman correlations with corresponding 1 MB resolution models less than 0.45, much lower than the correlation of >0.72 for the 5 KB models built by Hiearchical3DGenome. This result clearly demonstrates the limitation of LorDG without domain partitioning and supports the hierarchical reconstruction approach of Hierarchical3DGenome.

Discussion

The reconstruction of high-resolution 3D structures plays an important role in understanding and identifying various detailed interactions within a genome. It has been shown that the genome architecture and spatial structure is crucial for genome activity and DNA functions[34–37] because the spatial organization facilitates cellular activities such as gene expression and transcriptional regulation. Hence, modeling the chromosome 3D structures at higher resolution facilitates the identification of chromosomal regions such as chromosome territories, compartments, TADs and chromatin loops, and the relationship between these regions.

We introduced a hierarchical modeling algorithm to build high-resolution models of chromosomes by using low-resolution chromosomal models as a framework to position high-resolution models of topologically associated domains (TADs) constructed from Hi-C data. The algorithm was able to successfully reconstruct 3D models of full chromosomes of the human cell line GM12878 at 5 KB resolution. These high-resolution models were consistent with models at 1 MB resolution and FISH data. This algorithm has the potential to reconstruct 3D chromosome models at even higher resolution given Hi-C data of sufficient sequencing depth and quality.

Therefore, through the 3D structures generated by Hierarchical3DGenome researchers can study the relationship between chromosomal regions at a fine-grained scale. By using Hierarchical3DGenome, researchers can also identify structural patterns and relationships between regions related to topologically associated domains (TADs). This is possible because Hierarchical3DGenome takes TADs of chromosomes into consideration during structure construction to preserve important properties exhibited by intra-and inter-TAD contacts. Hence, Hierarchical3DGenome method is useful tool to gain detailed knowledge about chromatin organization and enable a rational interpretation of genome functions based on this organization.

Data Availability

The Java source code of Hierarchical3DGenome, sample datasets, its user manual, and the parameters for running the different analysis are available here: https://github.com/BDM-Lab/Hierarchical3DGenome.

References

1. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* **326**, 289–293 (2009).
2. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology* **30**(1), 90 (2012).
3. Bernardi, G. Chromosome architecture and genome organization. *PLoS One,* **10**(11), e0143739 (2015).
4. Markaki, Y. *et al.* The potential of 3D-FISH and super-resolution structured illumination microscopy for studies of 3D nuclear architecture: 3D structured illumination microscopy of defined chromosomal structures visualized by 3D (immuno)-FISH opens new perspectives for studies of nuclear architecture. *Bioessays* **34.5**, 412–426 (2012).
5. Cremer, T. & Cremer, C. Rise, fall and resurrection of chromosome territories: a historical perspective Part II. Fall and resurrection of chromosome territories during the 1950s to 1980s. Part III. Chromosome territories and the functional nuclear architecture: experiments and m. *European journal of histochemistry,* **50**(4), 223–272 (2006).
6. Edelmann, P., Bornfleth, H., Zink, D., Cremer, T. & Cremer, C. Morphology and dynamics of chromosome territories in living cells. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1551**(1), M29–M39 (2001).
7. Williamson, I. *et al.* Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence *in situ* hybridization. *Genes & development* **28.24**, 2778–2791 (2014).
8. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159.7**, 1665–1680 (2014).
9. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485.7398**, 376 (2012).
10. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**(7296), 363 (2010).
11. Rousseau, M., Fraser, J., Ferraiuolo, M. A., Dostie, J. & Blanchette, M. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC bioinformatics* **12**(1), 414 (2011).
12. Varoquaux, N., Ay, F., Noble, W. S. & Vert, J. P. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30**(12), i26–i33 (2014).
13. Lesne, A., Riposo, J., Roger, P., Cournac, A. & Mozziconacci, J. 3D genome reconstruction from chromosomal contacts. *Nature methods* **11**(11), 1141 (2014).
14. Adhikari, B., Trieu, T. & Cheng, J. Chromosome3D: reconstructing three-dimensional chromosomal structures from Hi-C interaction frequency data using distance geometry simulated annealing. *BMC genomics* **17**(1), 886 (2016).
15. Trieu, T. & Cheng, J. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic acids research* **42**(7), e52–e52 (2014).
16. Trieu, T. & Cheng, J. MOGEN: a tool for reconstructing 3D models of genomes from chromosomal conformation capturing data. *Bioinformatics* **32**(9), 1286–1292 (2015).
17. Trieu, T. & Cheng, J. 3D genome structure modeling by Lorentzian objective function. *Nucleic acids research* **45**(3), 1049–1058 (2016).
18. Ferraiuolo, M. A. *et al.* The three-dimensional architecture of Hox cluster silencing. *Nucleic acids research* **38.21**, 7472–7484 (2010).
19. Serra, F. *et al.* Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS letters* **589.20**, 2987–2995 (2015).
20. Carstens, S., Nilges, M. & Habeck, M. Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS computational biology* **12**(12), e1005292 (2016).

21. Segal, M. R. & Bengtsson, H. L. Reconstruction of 3D genome architecture via a two-stage algorithm. *BMC bioinformatics* **16**(1), 373 (2015).
22. Caudai, C., Salerno, E., Zoppè, M. & Tonazzini, A. Inferring 3D chromatin structure using a multiscale approach based on quaternions. *BMC bioinformatics* **16**(1), 234 (2015).
23. Rieber, L. & Mahony, S. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* **33**(14), i261–i266 (2017).
24. Baù, D. & Marti-Renom, M. A. Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. *Methods* **58**(3), 300–306 (2012).
25. Szałaj, P, et al. An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization. *Genome research*: gr-**205062** (2016).
26. Oluwadare, O., Zhang, Y. & Cheng, J. A maximum likelihood algorithm for reconstructing 3D structures of human chromosomes from chromosomal contact data. *BMC genomics* **19**(1), 161 (2018).
27. Hu, M. *et al*. Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology* **9.1**, e1002893 (2013).
28. Zou, C., Zhang, Y. & Ouyang, Z. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. *Genome biology* **17**(1), 40 (2016).
29. Zhang, Z., Li, G., Toh, K. C. & Sung, W. K. 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of computational biology* **20**(11), 831–846 (2013).
30. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* **33**(3), 1029–1047 (2013).
31. Imakaev, M. *et al*. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* **9.10**, 999 (2012).
32. Cheung, M. S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic acids research* **39**(15), e103–e103 (2011).
33. Teytelman, L. *et al*. Impact of chromatin structures on DNA processing for genomic analyses. *PloS one* **4.8**, e6700 (2009).
34. Oluwadare, O. & Cheng, J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. *BMC bioinformatics* **18.1**, 480 (2017).
35. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews genetics* **2.4**, 292 (2001).
36. Parada, L. A. & Misteli, T. Chromosome positioning in the interphase nucleus. *Trends in cell biology* **12.9**, 425–432 (2002).
37. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nature reviews Molecular cell biology* **17.12**, 743 (2016).

Acknowledgements

This work has been supported by National Science Foundation [Grant No: DBI1149224 to J.C.]. The computation for this work was performed on the high-performance computing infrastructure provided by Research Computing Support Services and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri, Columbia M.O.

Author Contributions

J.C. conceived the idea and supervised the project. T.T. designed and implemented the method. T.T. and O.O. performed the experiments and analyzed the results. T.T., O.O. and J.C. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41369-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019