Zhang *et al. BMC Genomics* 2019, **20**(Suppl 2):186
https://doi.org/10.1186/s12864-019-5470-2

BMC Genomics

RESEARCH Open Access

Large-scale 3D chromatin reconstruction from chromosomal contacts

Yanlin Zhang[1], Weiwei Liu[1], Yu Lin[2], Yen Kaow Ng[3] and Shuaicheng Li[1*]

From The 17th Asia Pacific Bioinformatics Conference (APBC 2019)
Wuhan, China. 14–16 January 2019

Abstract

Background: Recent advances in genome analysis have established that chromatin has preferred 3D conformations, which bring distant loci into contact. Identifying these contacts is important for us to understand possible interactions between these loci. This has motivated the creation of the Hi-C technology, which detects long-range chromosomal interactions. Distance geometry-based algorithms, such as ChromSDE and ShRec3D, have been able to utilize Hi-C data to infer 3D chromosomal structures. However, these algorithms, being matrix-based, are space- and time-consuming on very large datasets. A human genome of 100 kilobase resolution would involve ∼30,000 loci, requiring gigabytes just in storing the matrices.

Results: We propose a succinct representation of the distance matrices which tremendously reduces the space requirement. We give a complete solution, called SuperRec, for the inference of chromosomal structures from Hi-C data, through iterative solving the large-scale weighted multidimensional scaling problem.

Conclusions: SuperRec runs faster than earlier systems without compromising on result accuracy. The SuperRec package can be obtained from http://www.cs.cityu.edu.hk/~shuaicli/SuperRec.

Keywords: Hi-C, 3D chromosome structure, Multidimensional scaling, Chromosome conformation capture, 3D genome

Backgound

Genome-wide sequencing studies, such as the Human Genome Project (HGP) [1, 2], have deciphered the genomic sequences of humans. We are now in a position to reconstruct the 3D structure of the genome, that is, the conformations of the chromosomes within the nucleus. This will further our understanding of chromosomal interactions.

Recent discoveries through imaging analysis revealed that, while chromosome conformations may vary from cell to cell, they are not random [3, 4]. Hotspots of interactions and transcriptions are unevenly distributed; transcriptionally inactive segments prefer locations such as on nuclear periphery, around nucleoli, or at nuclear substructures [5–16]. These observations all point to the fact that gene expressions are highly associated with the chromatin structure.

However, since imaging techniques are not yet able to achieve high enough resolutions for genome-wide studies, researchers have sought to reconstruct the chromatin structure from knowledge of the interactions between genomic loci [17–23]. Valuable insights on gene regulations, genome translocations, copy number variations, genome stability, *etc.*, have been derived, owing to the success of these methods [24–31]. Among the technologies for capturing interaction information, one called Hi-C has been used prominently. The method produces a matrix, called a *contact map*, which stores the normalized frequencies between all pairs of genome loci (also referred to in the literature as *bins*, *regions* or *windows*) at some resolution.

*Correspondence: shuaicli@cityu.edu.hk
[1]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong SAR
Full list of author information is available at the end of the article

With these advances, it is reasonable to anticipate that the community will amass a very large collection of chromosome interaction data in the near future. These data are expected to be collected under many different conditions as well as resolutions, and for a variety of genomes, large and small. Their processing and analysis will present tremendous challenges to bioinformaticians [32].

A number of methods have been proposed for chromosome structure inferences from the contact maps. They can either infer a mean structure from the contact map [18, 19, 33–37] or solve multiple structures [28, 38–42]. Most of these methods operate on distance matrices, which consume very large amounts of memory for genomes with high resolution.

For our method, we assume the availability of high resolution data, which gives us more information at the expense of a larger problem size. We anticipate that data will be at the level of kilobase pairs (kbp). Existing strategies have difficulties with such data. In particular, distance matrix-based methods require very large amounts of memory to work; a resolution of one kbp for the human genome would require terabytes of memory. Processing time presents another issue.

The fastest method currently has a time complexity of $O\left(n^3\right)$ [36], rendering them inefficient for data of large sizes. In spite of the challenge, genome-scale 3D chromatin reconstruction has nonetheless been performed in at least two studies [43, 44]. In the study by Diament el al., a sparse contact matrix was generated with only a sampled portion of the Hi-C matrix, and the reconstruction was performed with a reduced set of constraints. This method becomes inefficient when the number of loci is increased [43]. In the study by Segal et al., [44] existing single chromosome structures were incorporated to form a whole genome structure. Since the method is dependent on existing tools such as ChromSDE for the inference of single chromosome structures, they are constrained by the efficiency of those tools.

In this work, we propose a *progressive* multidimensional scaling (MDS) approach for structure reconstruction from Hi-C data. We introduce a succinct representation of the distance matrix to reduce space consumption. The proposed approach progressively infer the coordinates to allow more flexible control of runtime. On the benchmark dataset which consists of simulated data of 100 to 30,000 loci, our approach (implemented as a program called SuperRec) performed 5 to 435 times faster compared to ShRec3D. In particular, it demonstrated a speedup of more than 400 times in reconstructing a structure of 30,000 loci, a length sufficient for us to analyze the longest human chromosome at a resolution of 10 kbp. When accessed with normalized root-mean-square deviation (RMSD) (Additional file 1: S1), Spearman's rank correlation coefficient (SRCC) and Pearson correlation coefficient (PCC), we found no loss of accuracy in the results obtained by SuperRec.

Methods

This section presents our method in detail. First, we model the sequence as a continuous linear polymer. We show how the contacts from Hi-C can be transformed and represented succinctly in the form of a distance matrix. The reconstruction then works by assigning coordinates to chromosome loci progressively. After all the coordinates are assigned, we refine and sharpen the coordinates iteratively through local search.

Structure modeling

In our algorithm, the chromosome is modelled as a continuous linear polymer, composed of many chromosomal loci. For example, human chromosome 1 can be modelled as ∼25,000 bins at a resolution (number of base pairs per bin) of 10 kbp (kilo base pairs). We use each bin to represent a locus within the chromosome. During the structure reconstruction step, the determination of each locus's location is simplified to that of determining the 3D position of the locus's centroid. Although this model omits the local structures at each locus, it is the most widely accepted representation and is considered the most accurate model achievable by the resolution of current 3C-based techniques [35, 36, 41]. The 3D positions of all the loci are referred to as a *structure* or a *configuration*. In this work, we let n denote the number of loci.

Converting contact frequency to distance

To infer the 3D structures from Hi-C experimental data, the pairwise distances between loci are first computed. Most of the Hi-C protocols are cell population-based experiments; they provide the average contact frequencies across different cells. Every pair of loci i and j are associated with a number of m replicates by accumulating different structures, and the normalized contact frequency f_{ij} can be inferred from Hi-C dataset. Many structure inference methods assume a power-law relationship between contact frequency and 3D distance, allowing a contact $\left(f_{ij}\right)$ to be converted into a corresponding distance $\left(d_{ij}\right)$ through the equation $d_{ij} = 1/f_{ij}^{\alpha}$. The power-law coefficient α varies across datasets and needs to be estimated using other techniques. The special case where $\alpha = 1$ is called an inverse frequency (IF). Given a fixed α and the inferred structure X, the goodness of fit is calculated as $\sum_{f_{ij}>0}\left(f_{ij}^{'} - f_{ij}\right)^2$ where $f_{ij}^{'} = 1/d(X)_{ij}^{1/\alpha}$. Similar to ChromSDE, we assume that $0.1 \leqslant \alpha \leqslant 3$ (a range covering most α in previous studies) and use a golden section search to find the correct α through minimizing the goodness function. We face several challenges with this approach. First, Hi-C can only capture 2.5% of the contact loci, with large variations in the captured frequencies; these contacts are moreover only reliable for the physically close loci. The power-law relation $d_{ij} = 1/f_{ij}^{\alpha}$ also results in infinite distances for low frequency contact pairs. While this can be remedied by enforcing an upper bound on the distances, a criteria for deciding the upper bound would be difficult to derive. Furthermore, the distances converted with a power-law relationship are not metric, rendering many computations that work for metric relations unusable.

Recently, Lesne el al. solved these problems elegantly using the shortest-path method in graph theory [36]. They modeled the contact matrix as a connected graph, in which a vertex represents a locus and an edge is associated with a distance as the inversion contact frequency of the corresponding locus pair. The final distance between loci i and j is modified with the shortest-path distance within the graph. Not only is the shortest-path distance metric and represents a tighter estimation of locus distance, the approach also mitigates the problem due to low frequency contact pairs.

Computing shortest-path distance is, however, both time- and space-consuming since it requires the computation of all-pairs shortest-paths. Hence we propose a method to approximate this distance. We randomly choose ℓ ($\ell \ll n$) loci as *pivots*, and denote this set of loci as P. After that, we compute the single source shortest-path distances with each pivot locus as the source to all the n loci. Denote the shortest distance from pivot p to a vertex v as $d_p(v)$. With these shortest-path distances from the pivots we approximate the remaining shortest-path distances. Given a pair of loci i, j, if i or j is a pivot, we can obtain their shortest distance from the computed shortest-path distances. Otherwise, we use $\min_{p \in P} d_p(i) + d_p(j)$ to approximate the shortest distance between i and j. By increasing the number of pivots, the approximate shortest-path distance can be made arbitrarily close to the true shortest-path distance.

To reduce the space consumption, while computing the shortest-path distances, we adopt an adjacent list representation for the distance matrix derived from Hi-C dataset. Also, the approximated distances are not stored; we merely store the ℓ sets of shortest-path distances from the pivots — the approximated distances are computed on the fly. This data structure reduces the space complexity from $O(n^2)$ to O(ln+e) (e is the number of significant Hi-C contacts) to store a distance matrix with $n \times n$ dimensions. See Additional file 1: Figure S8 for a comparision of run-time memory usage.

Assigning coordinates progressively

The structure reconstruction problem is often formulated as: Given the pairwise distances (with errors) of all loci, to find a 3D configuration X for those loci which satisfy the distance constraints. Denote the distance between loci i and j inferred from contact information as $\hat{d}_{ij}$, and denote the Euclidean distance between loci i and j in a configuration X as $d_{ij}(X)$. The problem can be solved by minimizing the following objective function:

$$\min \sum \sum\nolimits_{i \le j \le n} \left(d_{ij}(X) - \hat{d}_{ij} \right)^2 \tag{1}$$

which can be solved by multidimensional scaling (MDS) [45]. Such an approach has been utilized by several groups to reconstruct the chromatin structures [18, 24, 36, 42]. It performs well on Euclidean distance with small error. However, the distances inferred from contact information suffer from large errors, which are especially significant in the larger distances due to the underlying mechanism of Hi-C. This prompts us to adapt the formulation to a weighted one:

$$\min \sum \sum\nolimits_{i \le j \le n} w_{ij} \left(d_{ij}(X) - \hat{d}_{ij} \right)^2 \tag{2}$$

where a weight w_{ij} can be assigned to each loci pair i and j according to their distance, d_{ij}, allowing us to give higher weights to closer pairs. Similar to the earlier problem, this problem has a solution through the use of WMDS (weighted multidimensional scaling) [46].

However, the use of MDS and WMDS remains time- and space-consuming, and will not scale on problems of larger sizes. To solve this, we propose a progressive solution for the MDS and WMDS problem, namely, iMDS (iterative MDS) and sMDS (scalable MDS). Both methods relay on conducting MDS on subsets of loci.

Our proposed approaches are based on the following insight: A distance matrix of size $n \times n$ contains $\binom{n}{2}$ variables. On the other hand, the inferred structures contain $3n - 6$ free variables. That is, the distance matrix contains information that may be considered redundant, which can be potentially discarded without affecting the quality of the inferred structure. Hence, our approach will reduce these redundant values, which we expect to lower the chances of errors. Besides this, our approach will also naturally allow the assignment of larger weights to distances that are more reliable. These will become apparent in the subsequent subsections.

Iterative MDS (iMDS) for structure polish

In the same way as performed in SC-MDS [47], we randomly split the set of loci into k overlapping subsets, $s_1, \ldots, s_i, \ldots, s_k$, with small intersections I_i between s_i and s_{i+1}. Then, classical multidimensional scaling [48] is performed on each subset to obtain the local coordinates of each locus. The subsets of loci are then combined by first selecting s_1 as the reference, and then combining each s_{i+1} to s_i iteratively until all the subsets are combined. Our combination method differs from SC-MDS in that we incorporate the reflection of objects, which we now describe.

Local structures recombination To combine s_{i+1} into s_i in iMDS, we use I_i as a set of anchors. Denote the local coordinates for I_i in s_i and s_{i+1} as P_I and P_I', respectively. We want to superimpose P_I' onto P_I with a rigid transformation (a translation T and a rotation R) such as to minimize the RMSD. This problem is known to be solvable in linear time [49]. In order to solve the reflection problem, P_I' and its mirror were both superimposed to P_I. In the case that the RMSD between the mirror of P_I' and P_I is smaller than that between P_I' and P_I, the coordinates in s_{i+1} are replaced with those in the mirror of s_{i+1}.

$$RMSD = \min \sum_{i=1}^{N} \sqrt{\left\| p_i' - (Rp_i + T) \right\|^2} \tag{3}$$

After obtaining T and R, we apply them to the coordinates of s_{i+1} such that s_{i+1} and s_i would have the same frame of reference. This integrates s_{i+1} into s_i. In addition, we average P_I and the transformed coordinates of P_I' by R and T to update the coordinates of I_i.

Successive subsetting and combination As discussed in [47], grouping only neighboring loci is not as beneficial as grouping both close and distant loci. Hence, we randomly split all loci into k overlapping subsets. In order to combine two 3D sets successfully, we need to select at least 4 points not lie on the same plane (points on the same plane cannot identity a structure in 3D space), additionaly, a small number of loci may lead to a poor estimation of the rotation and translation matrix. Hence, we need to choose enough loci in I_i. In practice, we set the number of loci in I_i as 50, which is large enough for successive combinations in a three dimensional space. The performance of random subsetting and intersection is described in Additional file 1: S2.

Due to Hi-C's mechanism, the distances $\hat{d}$ measured have different degrees of reliability, which may aversely affect MDS and iMDS. This situation worsens when noise is elevated. To address this, we propose a scalable MDS to polish the structure obtained by iMDS.

Incorporating different distances

The structure from iMDS is further improved to better agree with the data from Hi-C experiments. In Eq. 2, a weight w_{ij} was introduced for each distance $\hat{d}_{ij}$ from Hi-C. Since shorter distances are more reliable than the longer ones, we set $w_{ij} = \hat{d}_{ij}^{-2}$ [42] to decrease the influence from the longer distances. We found this weighting scheme to be more robust than other schemes such as $w_{ij} = 1$ and $w_{ij} = 1/\hat{d}_{ij}$ [35, 36] (Additional file 1: S3).

However, computing WMDS for our framework remains infeasible for large data sets due to time and space complexities. We propose a scalable MDS approach here to address this issue. As far as we know, our approach is among the very few that are currently available for large-scale WMDS.

Scalable MDS

Our proposed scalable MDS is an iterative procedure which employs WMDS as a subroutine. At each iteration, we permute the loci randomly, and partition them by the permuted order into sets of size k each. Then, we apply WMDS to each resultant set, S say. Denote the loci coordinates in S before executing WMDS and after as P_S and P_S' respectively. We apply RMSD to discover the optimal superposition that maps each locus in P_S' to itself in P_S. Then, we update the coordinates for the locus to its average values from P_S and P_S' after the superposition. After we iterate through every set S, we restart the process for another round. This is repeated until the coordinates of the loci converge. The performance of random subsetting and iteration is described in Additional file 1: S2 and the sensitivity analysis of parameter settings is described in Additional file 1: S4.

WMDS is an iterative algorithm, the time complexity in each iteration is $O(n^3)$. In comparsion, the time complexity of sMDS for updating all loci once is $O(nk^2)$. Though sMDS need more iterations than WMDS (10 times more in practice), sMDS is faster.

Simulated contact maps

Binary contact map

We used known 3D structures as the basis for our simulation. For a given 3D structure, we constructed a binary matrix to store the pairwise contact information, with the top k nearest pairs of loci set to 1 and others set to 0.

Poisson distribution model

For a given 3D structure with pairwise distance d_{ij}, We generated the $M(i,j)$ entry of a simulated contact map M as a Poisson-distributed random number n_{ij}. The parameter λ_{ij} is defined as $\lambda_{ij} = \beta/d_{ij}^{\alpha}$ based on the power-law conversion of distance and contact frequency. β is a tunable parameter to control the signal coverage of the contact map.

Data preparation

In silico genome structure

The in silico nucleus with 1 and 16 chromosome(s) (Additional file 1: Table S1) used to test our software were generated using a polymer model. The coordinates were taken from the Langevin dynamics simulation after it has reached thermal equilibrium. After that, we constructed Poisson distribution-based contact maps and binary contact maps for the in silico structures. In addition to the contact matrices corresponding to in silico chromosomes, we also included contact maps generated from Poisson distribution model at different signal coverage levels of the regular helix structure used as benchmark dataset by Zou et al. [37] as well as the structure of chromosome 2 reconstructed from a real Hi-C dataset (mESC).

Hi-C experimental dataset

In this study, we used Hi-C data from two cell lines mESC and GM12878. These raw sequence data are transformed into contact map with Juicer [50]. The matrix balancing method described in [51] was used to normalize the contact matrix in order to remove biases in Hi-C dataset.

Results and discussion

We implemented our proposed approach in a package named *SuperRec*. We compared SuperRec with public publicly available softwares on a ubuntu 16.04 server equipped with two Intel(R) Xeon(R) E5-2620 CPUs, and 256 GB memory. All softwares were executed with suggested configurations, default setting were used when there is no recommended configuration.

The approximated shortest-path distances are reliable

We first assessed the quality of our approximation of the shortest-path distance. Figure 1 visualizes the Hi-C dataset SRX764938 (GM12878) [52] of human. Initially, there are many distances of small values derived from power-law conversion that are indistinguishable from each other (Fig. 1a). After using the shortest-path distances to refine the power-law converted distances, the local distances along the genome became significantly closer than the longer-range ones as expected (Fig. 1b).

An effective solution here is to use the shortest-path distance refinement approach to infer the spatial distances. However, the approach is time- and space-consuming, requiring time cubic to the number of loci; the very large number of loci to compute for renders it infeasible.

To overcome this we approximated the shortest-path distances through the use of pivots as described earlier. To assess the effects of the accuracy loss, we attempted to reconstruct the 3D configuration from the approximated distances of an in silico dataset with 10,000 loci. We repeated this for a range of different number of pivots from 1 to 10,000 to examine how having more pivots would affect the result; the case of using 10,000 pivots is the same as when the actual shortest-path distances are used.

We first assessed the accuracy loss due to approximation through two parameters: *Approximation Ratio (AR)* and *Exact Matching Rate (EMR)*. The approximation ratio is defined to be the ratio between the actual all-pair shortest-path length and the approximated all-pair shortest-path length. We obtained favorable AR values of 0.93 for 100 pivots, and 0.98 for 500 pivots. With 1000 pivots, the approximation ratio became more than 0.99 (Fig. 2a). The exact matching rate is defined to be the rate of the approximated shortest paths lengths that are equal to the corresponding actual shortest-path length. We obtained EMR of 0.85, 0.93 and 0.98 with 500, 1000, and 20,00 pivots, respectively (Fig. 2b). Additionaly, both EMR and AR show very small variances in our repeated analysis (Additional file 1: Figure S9).

We next examined if the quality of our reconstruction would be affected when the approximated shortest-path distances are used instead of the actual shortest-path distances. We conducted experiments on the same data with 10,000 loci, and with varying number of pivots from 1 to 10,000. The PCC between pairwise distance calculated from reconstructed and true structures was found to be 0.89 when 50 pivots are used. These values converge to 0.95 when more than 100 pivots are used. Similarly, the

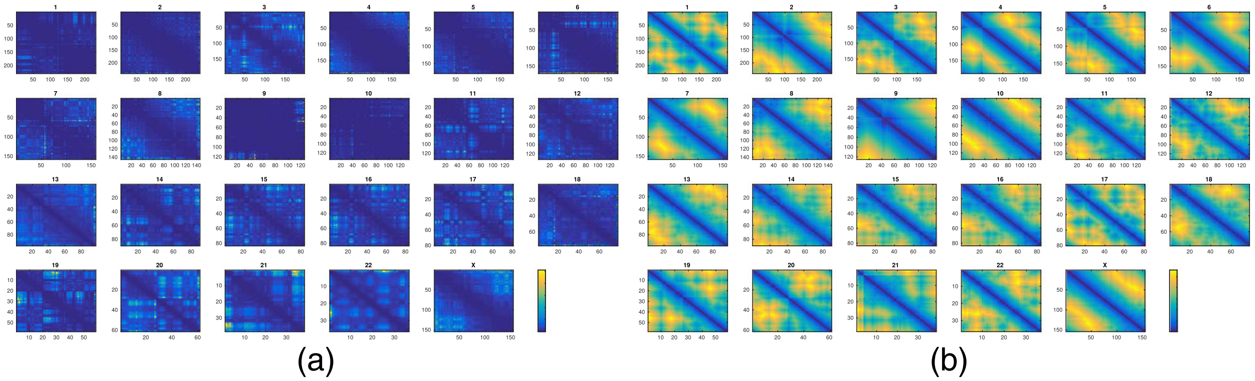

Fig. 1 Pairwise distances within each chromosome: **a** distances converted with power-law conversion, **b** power-law converted distances with the refinement of all-pair shortest-path distance

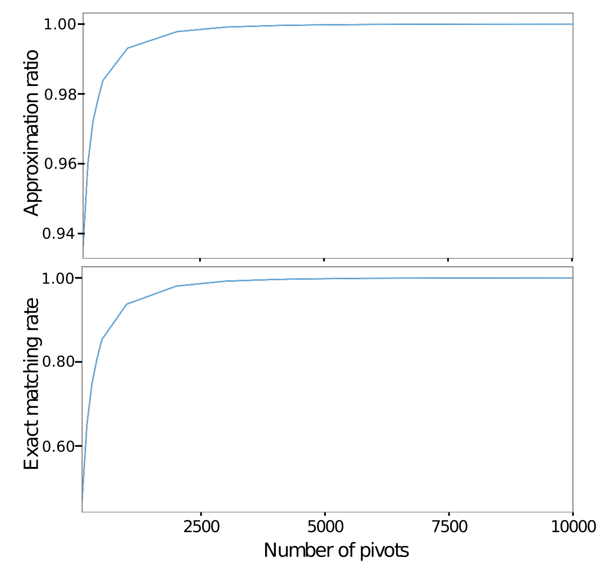

Fig. 2 Accuracy of approximate shortest-path distance: Approimation ratio calculated as the ratio of sum of all-pair shortest-path distance; Number of matched pairs of distance between approximate shortest-path distance and real shortest-path distance

SRCC between pairwise distances calculated from reconstructed and real structures converge to 0.965 when more than 100 pivots are used. The normalized RMSD between reconstructed and real structure varies from 0.17 to 0.27 throughout the experiment (Fig. 3); these values are considered small since they are the aggregate of 10,000 loci.

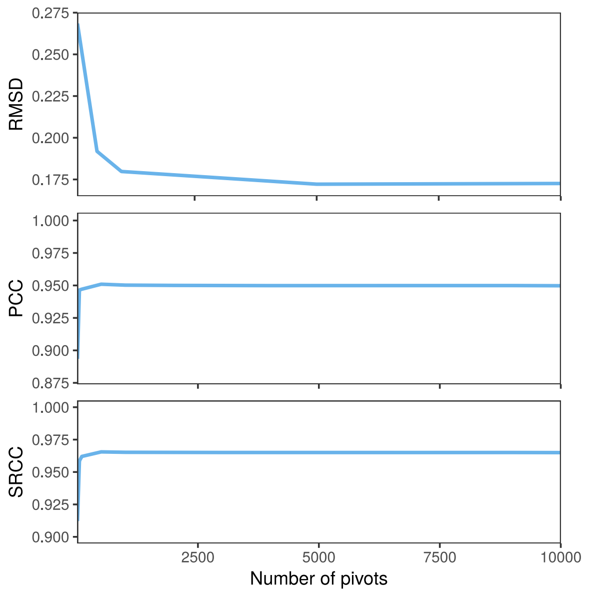

Fig. 3 Accuracy analysis of a reconstructed structure using various numbers of pivots to compute the shortest-path distance. For cases with more than 500 pivots, the difference between structures from accurate shortest-path distances and those from approximated shortest-path distances is negligible. Normalized RMSD, PCC as well as SRCC measurements are consistent

These show that there is no significant degradation of result from the use of the approximated shortest-path distances. In our experiment, SuperRec performs well when using 10% or more loci as pivots (Additional file 1: Figures S6, S7), and we suggest using at least 10% loci as pivots when using SuperRec.

SuperRec is fast

To compare the speed of SuperRec with existing methods, we simulated chromosomal structures with different numbers of loci up to 30,000 (Additional file 1: Table S1). The corresponding binary contact information were inferred with in silico structures as described in Method. These binary contact data were further analyzed with SuperRec using 10% loci as pivots as well as ChromSDE, HSA [37], and ShRec3D. Figure 4 shows the computation time plotted against the number of loci. SuperRec achieves significant improvements when handling thousands of loci. For large dataset with 2000 loci and more, SuperRec performed between 5 to 435 times faster than its alternatives. In one instance with 30,000 loci, SuperRec took only 43 min, whereas ShRec3D required more than 13 days. ChromSDE and HSA would require a few hours to complete on cases with more than 1000 loci Hence, we stopped the comparison with ChromSDE and HSA on the larger cases.

SuperRec is accurate

Whereas we have used only synthetic data in our speed benchmark test, the quality assessment of our algorithm was performed using both synthetic data and actual Hi-C experimental data. We first performed the reconstruction from the synthetic contact matrix used (Additional file 1: Table S1) using HSA, ChromSDE, ShRec3D, and SuperRec. To achieve a fair and comprehensive comparison, three different measurements were calculated: (1) normalized RMSD of each pair of reconstructed structure and original structure, (2) PCC and (3) SRCC between the original and the reconstructed distances (Fig. 5). On Poisson distribution-based contact map datasets, all algorithms achieved comparable and accurate results under all three kinds of measurement, consistently reporting correlation coefficients greater than 0.90. The corresponding normalized RMSDs reported are small compared to the number of loci, except for HSA with 400 and 500 loci. On binary contact maps datasets, SuperRec and ShRec3D achieved comparable and accurate results under all three kinds of measurement, consistently reporting correlation coefficients greater than 0.90. The corresponding normalized RMSDs reported are also small compared to the number of loci. In contrast, the performance of HSA and ChromSDE is dissatisfactory due to their inability to handle binary contact map. This shows that SuperRec is as accurate as these state-of-the-art algorithms.

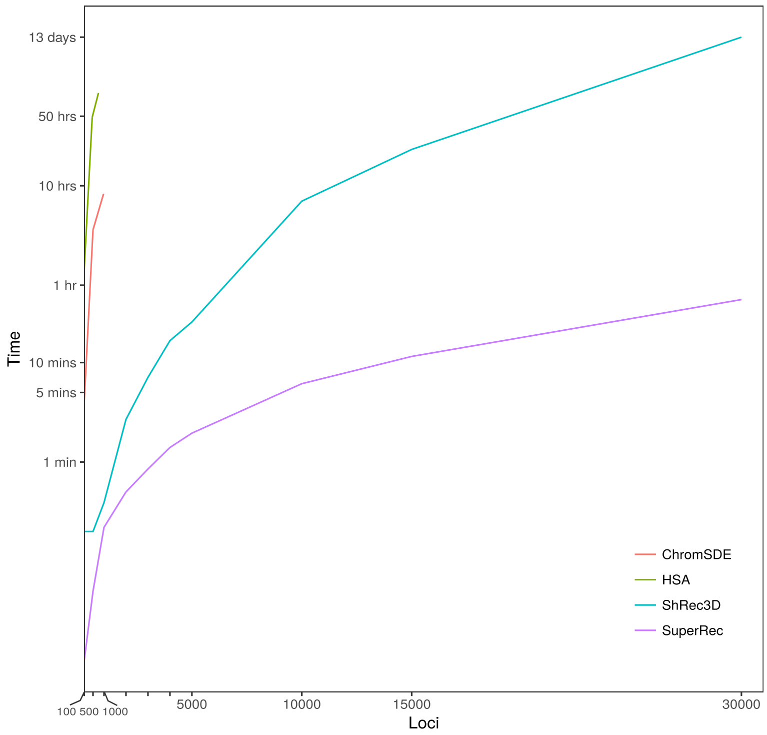

Fig. 4 Time usage for reconstructing chromosomal structures

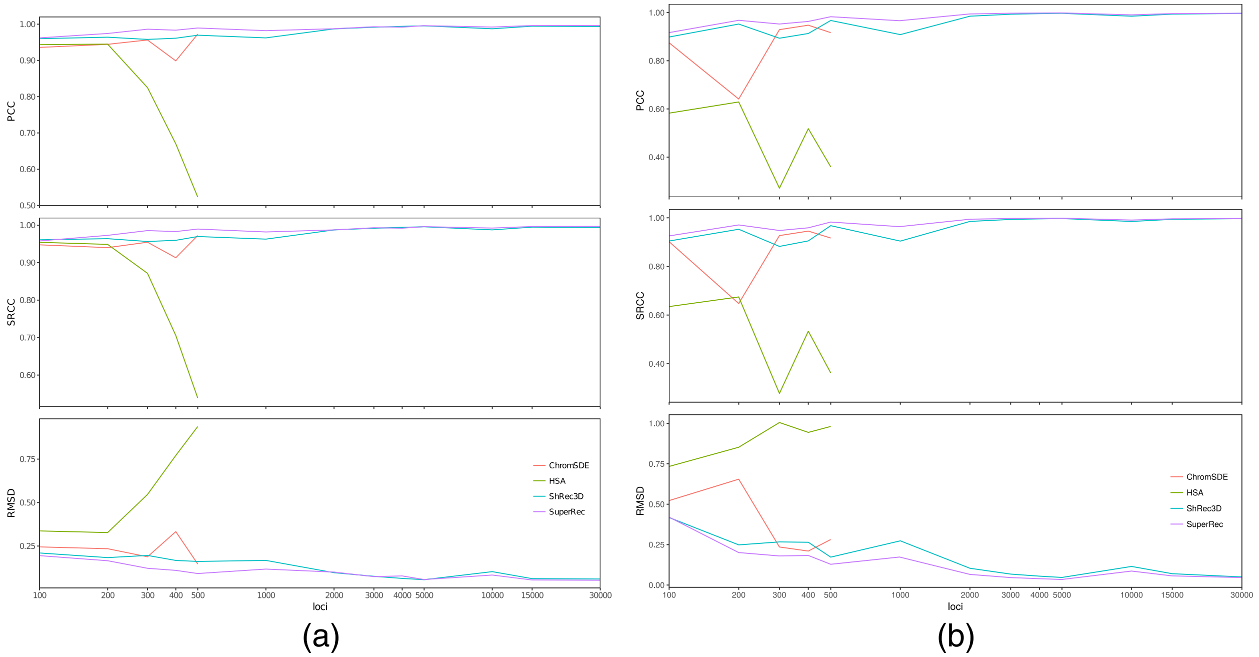

Fig. 5 Accuracy measurement by normalized RMSD, PCC and SRCC vs different numbers of loci. For structures with large number of loci, the SRCC and PCC is almost one, which indicated the reconstructed structures are close to the original structures. We stopped ChromSDE, and HSA computation beyond 500 loci due to the high runtime: **a** Poisson distribution-based contact maps; **b** Binary contact maps

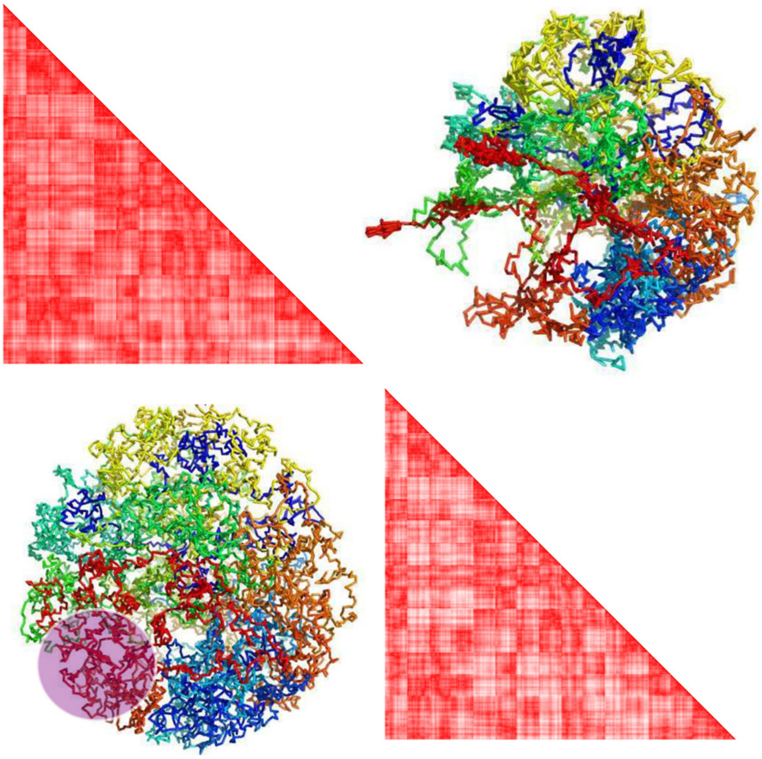

Fig. 6 The upper right and bottom left are two structures with 16 chromosomes and 10,000 loci: the upper right corresponding to a reconstructed structure by SuperRec with 1000 pivots, while the bottom left is the corresponding original structure. The upper left and bottom right heatmaps are plotted using pairwise distances inferred from reconstructed and real structures respectively

Figure 6 shows two structure: an in silico structure with 10,000 loci, and the structure reconstructed by SuperRec. The two structures are highly similar except for the purple highlighted region. The discrepancy is likely due to the relative sparsity of loci in the highlighted region, which resulted in the loci of that region to have fewer contacts with others, thus complicating the inference. On the other hand, we note that the polymer connectivity is preserved in our reconstructed structure (Fig. 7).

In addition to the comparison based on synthetic chromatin structures, we also reconstructed the structure of the regular helix from contact matrices at different signal coverage levels with HSA, ShRec3D, ChromSDE as well as SuperRec. Evaluating by the normalized RMSD between reconstructed structure and real structure, we find our method to be effective and robust (Additional file 1: Table S2).

We also performed the analysis with an *in situ* Hi-C dataset (GM12878) [52] of human at 1MB resolution. The chromosome (Fig. 8) reconstructions were performed with the usage of intra-contact matrices. Since the underlying structure of the Hi-C dataset is unknown, to evaluate the accuracy of the reconstructed structures we compared the pairwise distances from the reconstructions with the all pair shortest-path distances calculated from

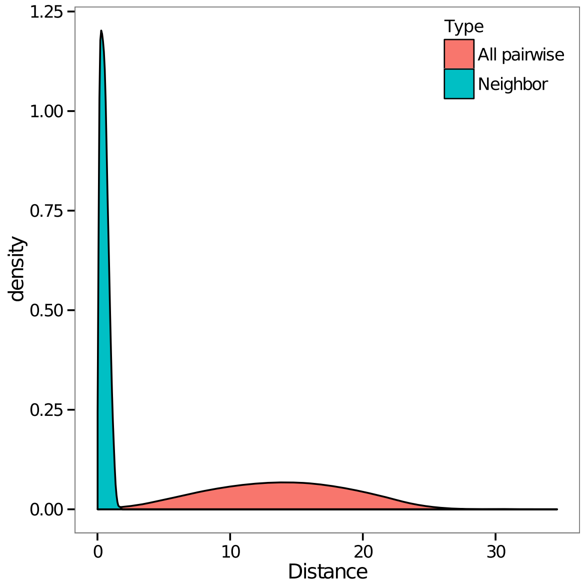

Fig. 7 Polymer connectivity: Histogram of distances between neighboring loci along chromosomal sequence (blue) and all pairwise distances (red) computed from a reconstructed configuration of 10,000 loci

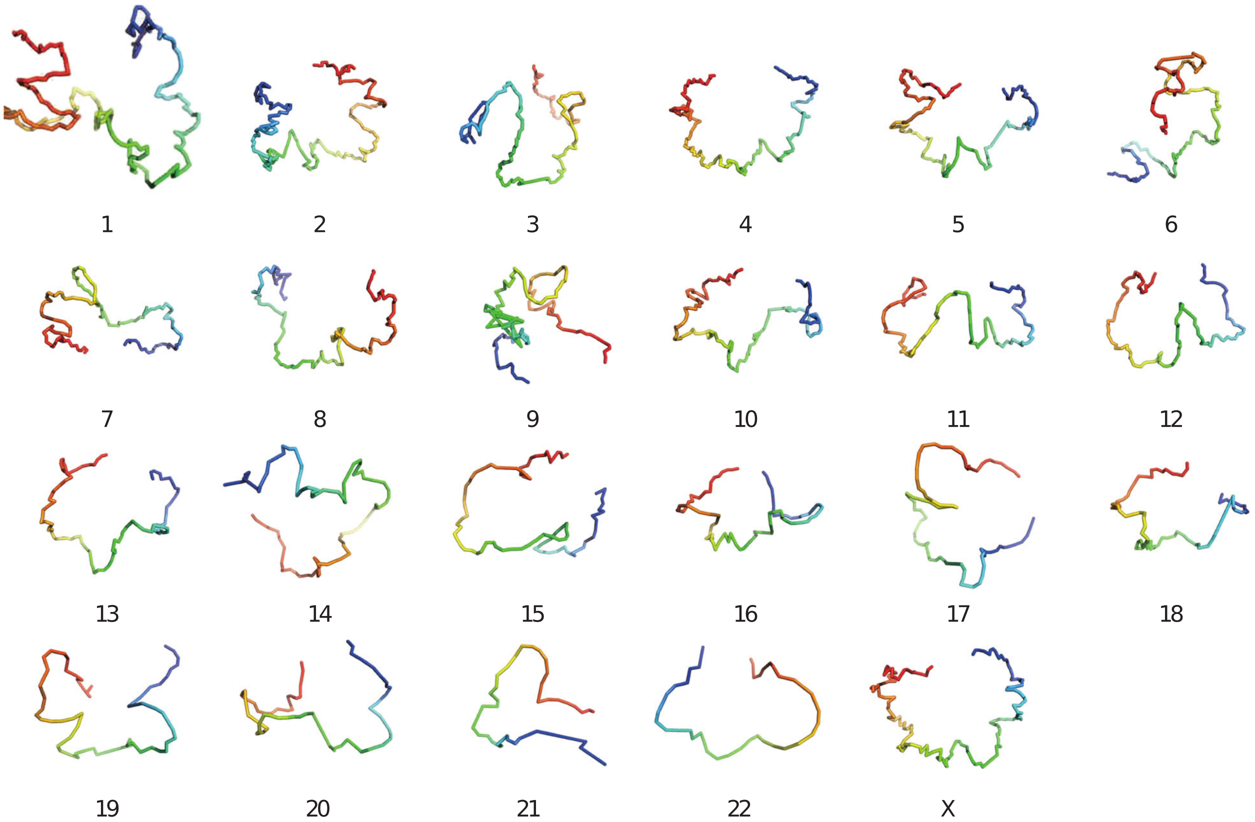

Fig. 8 3D visualization of human autosomes and X chromosome based on Hi-C data (SRX764938) [52] at 1MB resolution, only intra-contact frequency account for those reconstructions

contact frequency. The similarity is then expressed using correlation coefficients (Table 1). ShRec3D and SuperRec achieved similar performances when handling chromosome level reconstruction. At genome-wide level of number of loci, SuperRec slightly outperformed ShRec3D.

Validations and comparisons using FISH data

We also compared the methods' performance in inferring 3D chromatin structures using known distances derived from public 3D-FISH data for the cell lines mESC [53]. We selected six pairs of genomic loci from chromosome 2 or chromosome 11 for validation, with distances derived from FISH probes at 40-kb resolution. We inferred structures of chromosome 2 and chromosome 11 with HSA, BACH, ShRec3D and SuperRec with Hi-C contact data at 40-kb resolution. Single-track HSA, multi-track HSA and BACH were executed with the raw contact maps of NcoI and HindIII. ShRec3D and SuperRec were executed with the normalized contact maps of NcoI and HindIII. In addition, a modifed version of ShRec3D with a fixed α for distance conversion was included in our analysis. PCCs between the predicted distances from reconstructed structures and distance from 3D-FISH were calculated. Since each FISH locus spans two neighboring loci in the structures derived from Hi-C dataset, different combinations of neighbor loci at the two ends of FISH probed pair were used to compute the distances from 3D structures, and a range of PCCs for each FISH data set were obtained. The PCCs of SuperRec and multi-track HSA are most robust and significantly higher than those of other approaches (Fig. 9).

Application to sparse contact map

We carried out reconstructions using SuperRec on simulated contact maps of chromosome 2 (mouse) at different signal coverages (from 10 to 90%), and 10 simulated contact maps were generated at each signal coverage. SuperRec works well for both sparse and dense contact map when accessed with PCC and SRCC between distances from reconstructed structure and true structure. PCC ranges between 0.60 and 0.75, SRCC ranges between 0.65 and 0.80 at 10% signal coverage. The correlations increased with increasingly high signal coverage, with both PCC and SRCC approaching 0.9 when signal coverage reaches above 50 (Fig. 10). This demonstrates SuperRec's ability in handling Hi-C contact maps from low to high coverage.

Table 1 Accuracy measurement of SuperRec, ShRec3D and HSA based on real Hi-C data using Spearman's rank/Pearson's coefficient between distances calculated from reconstructions and all-pair shortest-path distances computed from contact information

Correlation	Method	Chromosome																							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	All
Spearman	SuperRec	0.981	0.986	0.973	0.942	0.971	0.964	0.989	0.993	0.997	0.981	0.987	0.979	0.991	0.992	0.976	0.976	0.973	0.980	0.979	0.974	0.974	0.970	0.982	0.880
	ShRec3D	0.981	0.983	0.984	0.988	0.985	0.989	0.990	0.987	0.998	0.991	0.977	0.981	0.985	0.991	0.986	0.986	0.974	0.983	0.986	0.982	0.992	0.979	0.990	0.851
	HSA	0.799	0.916	0.916	0.740	0.857	0.884	0.866	0.952	0.904	0.904	0.892	0.801	0.733	0.891	0.880	0.873	0.759	0.808	0.747	0.841	0.934	0.828	0.892	-
Pearson	SuperRec	0.982	0.985	0.974	0.947	0.974	0.960	0.987	0.993	0.998	0.982	0.989	0.983	0.989	0.992	0.980	0.980	0.978	0.985	0.982	0.972	0.970	0.978	0.981	0.896
	ShRec3D	0.981	0.978	0.983	0.976	0.977	0.985	0.986	0.988	0.998	0.983	0.983	0.980	0.984	0.990	0.984	0.984	0.974	0.984	0.985	0.980	0.991	0.977	0.979	0.857
	HSA	0.820	0.915	0.904	0.719	0.812	0.831	0.835	0.935	0.909	0.897	0.879	0.813	0.757	0.893	0.851	0.890	0.790	0.813	0.752	0.786	0.914	0.763	0.836	-

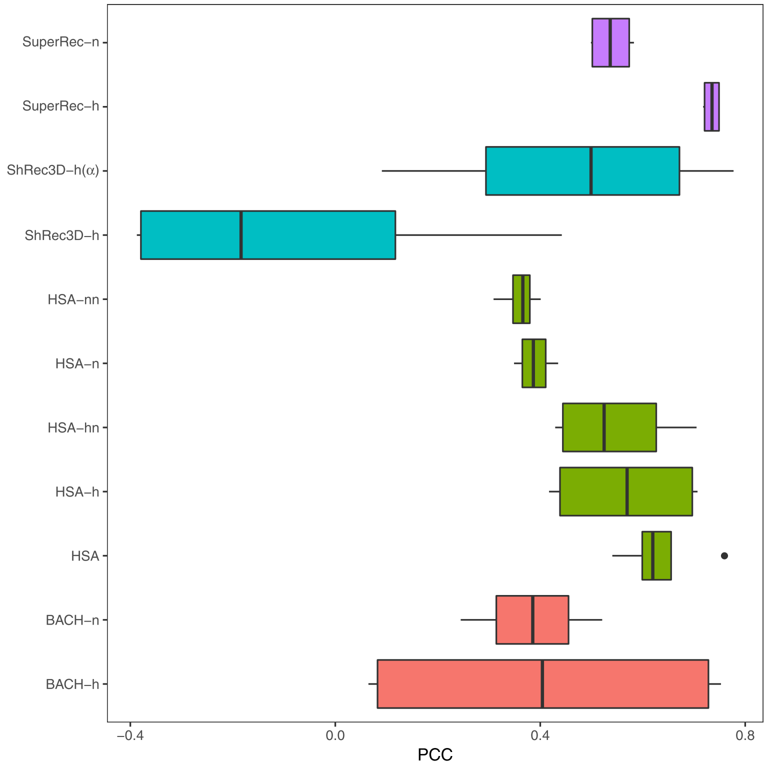

Fig. 9 Box plots of PCCs between FISH measured distance and predicted distances by different Hi-C dataset of mESC at 40-kb resolution. *BACH-h* BACH with contact map of HindIII, *BACH-n* BACH with contact map of NcoI, *HSA* multi-track HSA, *HSA-h* HSA with contact map of HindIII, *HSA-nn* HSA with normalized contact map of NcoI, *HSA-hn* HSA with normalized contact map of HindIII, *HSA-n* HSA with contact map of NcoI, *ShRec3D-h* ShRec3D with normalized contact map of HindIII, *ShRec3D-h(α)* modified ShRec3D with given α on normalized contact map of HindIII, *SuperRec-h* SuperRec with normalized contact map of HindIII, and *SuperRec-n* SuperRec with normalized contact map of NcoI

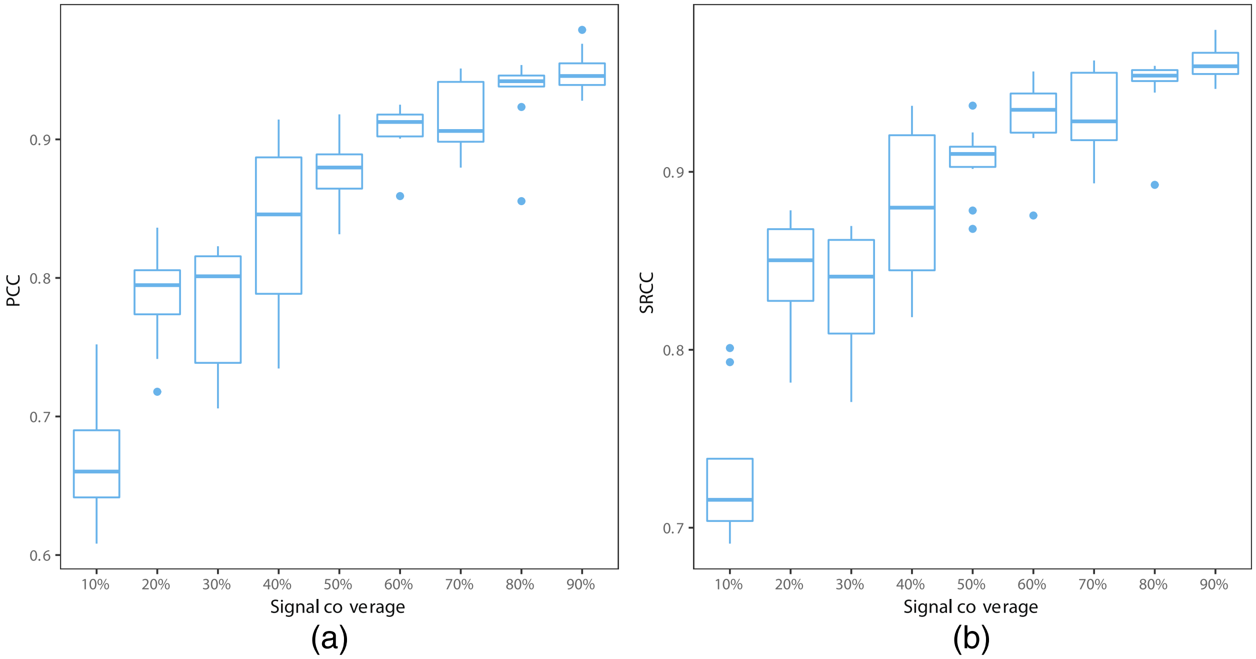

Fig. 10 Box plots of PCC (**a**) and SRCC (**b**) between true distances and predicted distances by simulated contact maps at different signal coverage

Conclusion

In this study, we devised a novel method for 3D chromatin reconstruction from chromosomal contacts, and implemented it into a complete software solution called *SuperRec*. We tested SuperRec on both synthetic and real Hi-C datasets. SuperRec achieved significant improvements in the analysis of longest human chromosome, completing the reconstruction at a resolution of 10 kbp within hours without loss of accuracy in the results.

Additional file

Additional file 1: Large Scale 3D Chromatin Reconstruction From Chromosomal Contacts - Supplementary materials. This file contains the supplementary text and figures mentioned in the text. (PDF 6926 kb)

Abbreviations

3C: Chromosome conformation capture; AR: Approximation ratio; EMR: Exact matching rate; FISH: Fluorescence in situ hybridization; IF: Inverse frequency; iMDS: Iterative multi-dimensional scaling; kbp: Kilobase pairs; MDS: Multi-dimensional scaling; PCC: Pearson correlation coefficient; RMSD: Root mean square deviation; sMDS: Scalable multi-dimensional scaling; SRCC: Spearman's rank correlation coefficient; WMDS: Weighted multi-dimensional scaling

Acknowledgements

Not applicable.

Funding

The work described in this manuscript and the publication costs for this manuscript were supported by a grant of GRF Project from RGC General Research Fund [9042181 (CityU 11203115)].

Availability of data and materials

The GM12878 dataset analysed during the current study are available at GEO repository, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. The in silico and mESC datasets used in this study can be downloaded from here: http://www.cs.cityu.edu.hk/~shuaicli/SuperRec/.

About this supplement

This article has been published as part of BMC Genomics Volume 20 Supplement 2, 2019: Selected articles from the 17th Asia Pacific Bioinformatics Conference (APBC 2019): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-20-supplement-2.

Authors' contributions

SL conceived the project. SL and YZ designed the algorithm. YZ implemented the algorithm. YZ and WL performed the analyses. YZ, WL and SL evaluated the results, and wrote the manuscript. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong SAR. [2]Research School of Computer Science, the Australian National University, Canberra, Australia. [3]Department of Computer Science, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia.

Published: 10 April 2019

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51.
3. Boyle S, Rodesch MJ, Halvensleben HA, Jeddeloh JA, Bickmore WA. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosom Res. 2011;19(7):901–9.
4. Muller I, Boyle S, Singer RH, Bickmore WA, Chubb JR. Stable morphology, but dynamic internal reorganisation, of interphase human chromosomes in living cells. PLoS ONE. 2010;5(7):11560.
5. Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G. Polycomb-dependent regulatory contacts between distant hox loci in drosophila. Cell. 2011;144(2):214–26.
6. Branco MR, Pombo A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. PLoS Biol. 2006;4(5):138.
7. Brown JM, Green J, das Neves RP, Wallace HA, Smith AJ, Hughes J, Gray N, Taylor S, Wood WG, Higgs DR, et al. Association between active genes occurs at nuclear speckles and is modulated by chromatin environment. J Cell Biol. 2008;182(6):1083–97.
8. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet. 2001;2(4):292–301.
9. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. Nature. 2007;447(7143):413–7.
10. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008;453(7197):948–51.
11. Iborra FJ, Pombo A, Jackson DA, Cook PR. Active rna polymerases are localized within discrete transcription "factories' in human nuclei. J Cell Sci. 1996;109(6):1427–36.
12. Németh A, Conesa A, Santoyo-Lopez J, Medina I, Montaner D, Péterfia B, Solovei I, Cremer T, Dopazo J, Langst G. Initial genomics of the human nucleolus. PLoS Genet. 2010;6(3):1000889.
13. Pirrotta V, Li H-B. A view of nuclear polycomb bodies. Curr Opin Genet Dev. 2012;22(2):101–9.
14. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell JA, Umlauf D, Dimitrova DS, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet. 2010;42(1):53–61.
15. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B. Interactions among polycomb domains are guided by chromosome architecture. PLoS Genet. 2011;7(3):1001343.
16. van Koningsbruggen S, Gierliński M, Schofield P, Martin D, Barton GJ, Ariyurek Y, den Dunnen JT, Lamond AI. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Mol Biol Cell. 2010;21(21):3735–48.
17. Van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. Hi-c: a method to study the three-dimensional architecture of genomes. J Vis Exp. 2010;39:e1869. https://doi.org/10.3791/1869.
18. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. A three-dimensional model of the yeast genome. Nature. 2010;465(7296):363–7.
19. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93.

20. Rodley C, Bertels F, Jones B, O'sullivan J. Global identification of yeast chromosome interactions using genome conformation capture. Fungal Genet Biol. 2009;46(11):879–86.
21. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the drosophila genome. Cell. 2012;148(3): 458–72.
22. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503(7475):290–4.
23. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. Nature. 2013;502(7469):59–64.
24. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert J-P, Noble WS, Le Roch KG. Three-dimensional modeling of the p. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. 2014;24(6): 974–88.
25. De S, Michor F. Dna replication timing and long-range dna interactions predict mutational landscapes of cancer genomes. Nat Biotechnol. 2011;29(12):1103–8.
26. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80.
27. Homouz D, Kudlicki AS. The 3d organization of the yeast genome correlates with co-expression and reflects functional relations between genes. PloS ONE. 2013;8(1):54699.
28. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 2012;30(1):90–8.
29. Lemieux JE, Kyes SA, Otto TD, Feller AI, Eastman RT, Pinches RA, Berriman M, Su X-z, Newbold CI. Genome-wide profiling of chromosome interactions in plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic variation. Mol Microbiol. 2013;90(3):519–37.
30. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, Zhang J, Schulz TC, Robins AJ, Dalton S, Gilbert DM. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. 2010;20(6):761–70.
31. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, et al. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012;488(7409):116–20.
32. Gibcus JH, Dekker J. The hierarchy of the 3d genome. Mol Cell. 2013;49(5):773–82.
33. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol. 2011;18(1):107–14.
34. Ben-Elazar S, Yakhini Z, Yanai I. Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. Nucleic Acids Res. 2013;41(4):2191–201.
35. Zhang Z, Li G, Toh K-C, Sung W-K. 3d chromosome modeling with semi-definite programming and hi-c data. J Comput Biol. 2013;20(11): 831–46.
36. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J. 3d genome reconstruction from chromosomal contacts. Nat Methods. 2014;11(11): 1141–3.
37. Zou C, Zhang Y, Ouyang Z. Hsa: integrating multi-track hi-c data for genome-scale reconstruction of 3d chromatin structure. Genome Biol. 2016;17(1):40.
38. Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS. Bayesian inference of spatial organizations of chromosomes. PLoS Comput Biol. 2013;9(1):1002893.
39. Rousseau M, Fraser J, Ferraiuolo MA, Dostie J, Blanchette M. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. BMC Bioinformatics. 2011;12(1):414.
40. Tjong H, Gong K, Chen L, Alber F. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome Res. 2012;22(7):1295–305.
41. Trieu T, Cheng J. Large-scale reconstruction of 3d structures of human chromosomes from chromosomal contact data. Nucleic Acids Res. 2014;42(7):52.
42. Varoquaux N, Ay F, Noble WS, Vert J-P. A statistical approach for inferring the 3d structure of the genome. Bioinformatics. 2014;30(12):26–33.
43. Diament A, Tuller T. Improving 3d genome reconstructions using orthologous and functional constraints. PLoS Comput Biol. 2015;11(5): 1004298.
44. Segal MR, Bengtsson HL. Reconstruction of 3d genome architecture via a two-stage algorithm. BMC Bioinformatics. 2015;16(1):373.
45. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964;29(1):1–27.
46. Borg I, Groenen P. Modern multidimensional scaling: theory and applications. J Educ Meas. 2003;40(3):277–80.
47. Tzeng J, Lu HH, Li W-H. Multidimensional scaling for large genomic data sets. BMC Bioinformatics. 2008;9(1):179.
48. Torgerson WS. Multidimensional scaling: I. theory and method. Psychometrika. 1952;17(4):401–19.
49. Arun KS, Huang TS, Blostein SD. Least-squares fitting of two 3-d point sets. Pattern Anal Mach Intell, IEEE Trans. 1987;PAMI-9(5):698–700.
50. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. Cell Syst. 2016;3(1):95–8.
51. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J Numer Anal. 2013;33(3):1029–47.
52. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
53. Eskeland R, Leeb M, Grimes GR, Kress C, Boyle S, Sproul D, Gilbert N, Fan Y, Skoultchi AI, Wutz A, et al. Ring1b compacts chromatin structure and represses gene expression independent of histone ubiquitination. Mol Cell. 2010;38(3):452–64.