GeneRHi-C: 3D GENomE Reconstruction from Hi-C data

Kimberly MacKay*
Department of Computer Science,
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
kimberly.mackay@usask.ca

Mats Carlsson
RISE
Kista, Sweden
mats.carlsson@ri.se

Anthony Kusalik
Department of Computer Science,
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
kusalik@cs.usask.ca

ABSTRACT

Background: Many computational methods have been developed that leverage the results from biological experiments (such as Hi-C) to infer the 3D organization of the genome. Formally, this is referred to as the 3D genome reconstruction problem (3D-GRP). Hi-C data is now being generated at increasingly high resolutions. As this resolution increases, it has become computationally infeasible to predict a 3D genome organization with the majority of existing methods. None of the existing solution methods have utilized a non-procedural programming approach (such as integer programming) despite the established advantages and successful applications of such approaches for predicting high-resolution 3D structures of other biomolecules. Our objective was to develop a new solution to the 3D-GRP that utilizes non-procedural programming to realize the same advantages.

Results: In this paper, we present a three-step consensus method (called GeneRHi-C; pronounced "generic") for solving the 3D-GRP which utilizes both new and existing techniques. Briefly, (1) the dimensionality of the 3D-GRP is reduced by identifying a biologically plausible, ploidy-dependent subset of interactions from the Hi-C data. This is performed by modelling the task as an optimization problem and solving it efficiently with an implementation in a non-procedural programming language. The second step (2) generates a biological network (graph) that represents the subset of interactions identified in the previous step. Briefly, genomic bins are represented as nodes in the network with weighted-edges representing known and detected interactions. Finally, the third step (3) uses the ForceAtlas 3D network layout algorithm to calculate (x, y, z) coordinates for each genomic region in the contact map. The resultant predicted genome organization represents the interactions of a population-averaged consensus structure. The overall workflow was tested with Hi-C data from *Schizosaccharomyces pombe* (fission yeast). The resulting 3D structure clearly recapitulated previously established features of fission yeast 3D genome organization.

Conclusion: Overall, GeneRHi-C demonstrates the power of non-procedural programming and graph theoretic techniques for providing an efficient, generalizable solution to the 3D-GRP.

Project Homepage: https://github.com/kimmackay/GeneRHi-C

CSBio 2019, December 4–7, 2019, Nice, France

ACM ISBN 978-1-4503-7215-2/19/12…$15.00
https://doi.org/10.1145/3365953.3365962

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**; **Computational genomics**; **Molecular structural biology**; **Biological networks**.

KEYWORDS

3D Genome Reconstruction Problem, Mathematical Modelling, Declarative Programming, Integer Programming, Network Layouts

ACM Reference Format:
Kimberly MacKay, Mats Carlsson, and Anthony Kusalik. 2019. GeneRHi-C: 3D GENomE Reconstruction from Hi-C data. In *Proceedings of 10th International Conference on Computational Systems-Biology and Bioinformatics (CSBio 2019).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3365953.3365962

1 INTRODUCTION

Within the nucleus, a cell's genetic information undergoes extensive folding and reorganization throughout normal physiological processes. Just like in origami where the same piece of paper folded in different ways allows the paper to take on different forms and potential functions, it is possible that different genomic organizations are related to various nuclear functions. Until recently, it has been extremely difficult to comprehensively investigate this relationship due to the lack of high-resolution and high-throughput techniques for identifying genomic organizations. The development of a biological technique called Hi-C (based on chromosome conformation capture) [26] has made it possible to detect the complete set of genomic regions simultaneously in close physical proximity. This proximity is often referred to as an "interaction" between two genomic regions. These interactions can be categorized as either intra-chromosomal (*cis*) interactions or inter-chromosomal (*trans*) interactions (Figure 1).

Hi-C [26] is a biological technique that utilizes next generation sequencing technologies to detect regions of the genome that are interacting in 3D space. These regions may be located on different chromosomes or distally on the same chromosome. An overview of the experimental procedure is depicted in Figure 2. Briefly, (1) cells are fixed with formaldehyde in order to covalently cross-link genomic regions that are in close 3D proximity. (2) The cross-linked fragments are then digested with a restriction enzyme to remove the potentially large non-interacting interconnecting segments of DNA. (3) The sticky ends generated through the restriction digest are filled in with biotinylated nucleotides. (4) Digested fragments are ligated together. (5) The initial cross-linking is removed, resulting in DNA fragments that represent the two genomic regions that form an interaction. (6) The biotinylated products are purified using streptavidin beads allowing for the detection of fragments that were cut by restriction enzymes. (7) Paired-end sequencing is

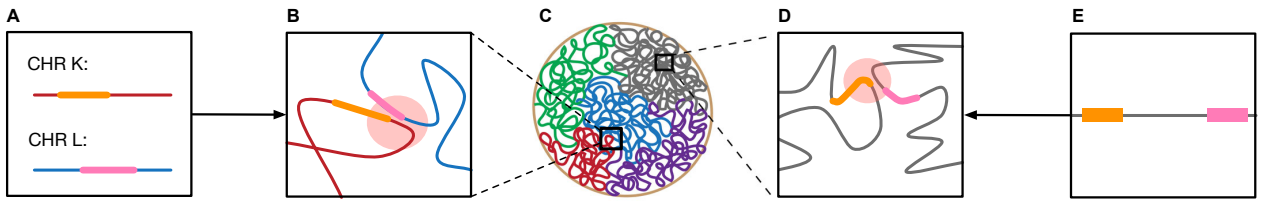


Figure 1: A representation of the DNA-DNA interactions that can occur within the 3D genome structure. Panels give the following representations. A: the linear locations of the genes undergoing a *trans*-interaction between two hypothetical chromosomes, K and L. B: a *trans*-interaction. C: a nucleus with the coloured lines representing the separate chromosomes from Babaei et al. [2]. D: a *cis*-interaction. These genes might be linearly "distant" but still have a detectable interaction in 3D space. E: the linear locations of the genes that are undergoing a 3D *cis*-interaction. The orange and pink regions in panels A, B, D and E are examples of possible gene locations. The red circles in panels B and D represent the genomic regions involved in an interaction.

then performed and the resultant reads are mapped to a reference genome using a Hi-C specific read mapper [1].

Mapping the raw data of a Hi-C experiment to a reference genome results in the generation of a $N \times N$ matrix (a whole-genome contact map) where N is the number of "bins" which represent linear regions of genomic DNA. In general, the number of genomic bins is approximately equal to the total genome size divided by the Hi-C experimental resolution. Whole-genome contact maps are characteristically sparse and symmetric along the diagonal. Each cell $(A_{i,j})$ of a contact map (**A**) records the count of how many times the genomic bin i was found to interact with the genomic bin j. These counts are often referred to as the frequency of the interaction between A_i and A_j (or interaction frequency). Inherent systematic biases within the whole-genome contact map are dampened by normalizing the interaction frequencies. Typically, an iterative correction and eigenvector decomposition (ICE) [17] or Knight-Ruiz (KR) [20, 25] normalization are/is applied to the raw data resulting in fractional interaction frequencies. It should be noted that during the normalization process interactions involving highly repetitive genomic regions such as centromere and/or telomere regions are often removed from the contact map (represented as 'NA') due to large amounts of noise and/or low signal [23].

The normalized whole-genome contact maps can be used to infer the 3D organization of the genome. The process of predicting a model of the 3D genomic organization from a contact map is known as the 3D genome reconstruction problem (3D-GRP) [36]. Typically this is done by converting the normalized interaction frequencies into a set of corresponding pairwise Euclidean distances. In general it is assumed that a pair of genomic regions with a higher interaction frequency will often be closer in 3D space than a pair of genomic regions with a lower interaction frequency [12, 16, 24]. Most computational tools for solving the 3D-GRP then take the predicted pairwise Euclidean distances as input and produce a visualization of the 3D genome by modelling the chromatin fibre as a polymer [38]. In general, most existing programs can be broadly classified as either (1) consensus or (2) ensemble methods. Consensus methods generate a single population-averaged genomic model that best represents the whole-genome contact map, while ensemble methods produce a collection of genome models that represent the inherent heterogeneity of genome organizations within a population of cells [23].

As the resolution of whole-genome contact maps increases, it is computationally infeasible to predict a 3D genome organization with many of the existing methods. One notable exception is the method miniMDS [34] which uses a divide-and-conquer approach to overcome this problem. In order to divide the overall problem into subproblems, miniMDS utilizes an algorithm for detecting chromosomal sub-domains called TADs to make the initial division. Unfortunately, this makes it inapplicable to organisms which do not have TADs like *Arabadopsis thaliana* [11, 27]. To overcome this, we have developed a generalizable, three-step consensus method for solving the 3D-GRP called GeneRHi-C (3D **Gen**om**e R**econstruction from **Hi-C** data; pronounced “generic"). Unlike other 3D-GRP solutions, GeneRHi-C does not rely on chromosomal sub-domains or organism specific constraints in the prediction process making it generalizable to any organism. Briefly, GeneRHi-C preforms the following three steps: (1) dimensionality reduction, (2) graph representation and (3) calculation of (x, y, z) coordinates. The resulting 3D genome organization represents the interactions of a population-averaged consensus structure. In order to demonstrate its utility GeneRHi-C was used to predict a 3D genome organization from an existing *Schizosaccharomyces pombe* (fission yeast) Hi-C dataset.

2 COMPUTATIONAL WORKFLOW

2.1 Step 1: Dimensionality Reduction

Under normal cellular conditions, a given genomic region can be simultaneously involved in more than one interaction within the genome [14]. In contrast, a single genomic region within an individual cell is only able to participate in one Hi-C mediated interaction due to inherent restrictions within the biochemistry of the Hi-C experimental protocol [42]. In diploid organisms (organisms with two genomic copies) single cell Hi-C reactions are only able to detect two Hi-C mediated interactions per genomic region, one for each genomic copy [32]. An analogous restriction can be assumed in haploid organisms (organisms with only one genomic copy), where a single genomic region can only be actively detected in one Hi-C mediated interaction in a single cell. Using this restriction, a

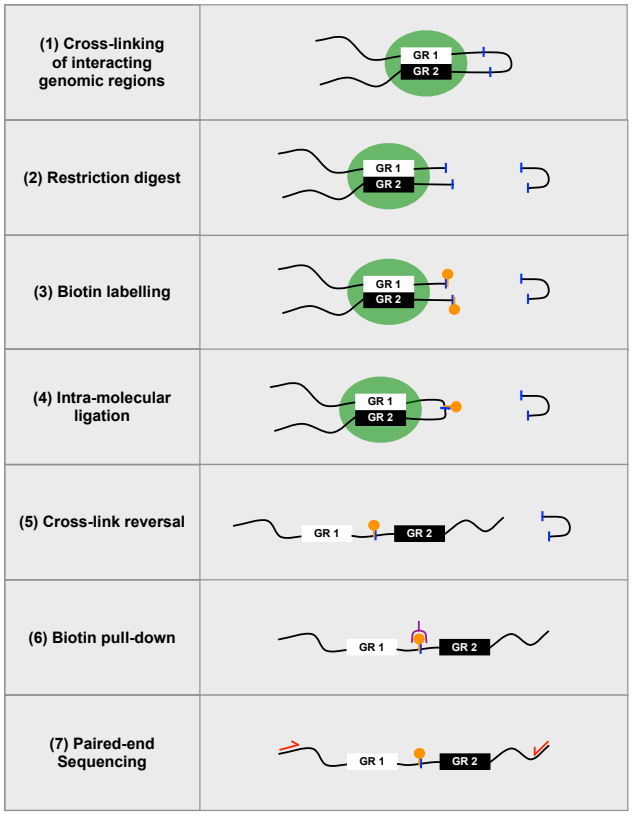


Figure 2: A simplified overview of the Hi-C protocol adapted from reference [26]. GR stands for "genomic region". The blue lines represent the location of a restriction enzyme cut site; green circles, a pair of genomic regions being chemically cross-linked together; orange circles, biotin; and red arrows, the primers that are required for paired-end sequencing. The purple symbol in step (6) represents a streptavidin bead that can be used to purify molecules with a biotin label.

model of the 3D genome organization can be constructed from a whole-genome contact map by selecting a ploidy-dependent subset of the interactions for each genomic region that maximizes the sum of the corresponding interaction frequencies. The mathematical model and corresponding implementation presented in this paper focus on modelling the 3D organization of haploid genomes but, as outlined in Section 5.3, could be extended to organisms with higher ploidies.

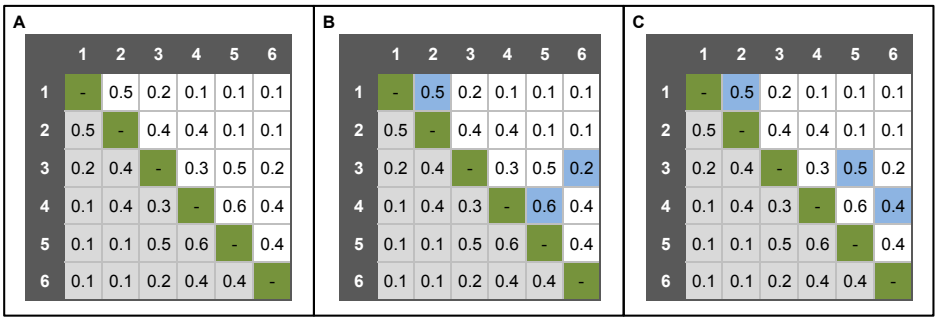
Naively, a greedy heuristic could be employed to model the 3D fission yeast genome organization using the strategy described above. Briefly, the subset of interactions representing the solution set can be chosen by sorting and selecting the interactions with the largest corresponding frequency values. This process would then be repeated, rejecting any frequency for a region of the genome that has already been selected. This heuristic will fail to take into account the situation where lower frequencies, which were rejected by selecting a higher frequency interaction, actually result in a greater overall maximum value for the sum of all selected frequencies within the solution. An example of this can be seen in Figure 3 where panel A is a hypothetical whole-genome contact map and panels B and C represent two possible solution matrices with different overall frequency sums. Specifically, Figure 3B follows the greedy heuristic described above which results in a non-optimal solution where the selected frequencies sum to 1.3. Figure 3C shows the optimal solution where the selected frequencies sum to 1.4. This type of optimization problem has been shown to be well-suited for solutions using non-procedural computational paradigms.

We have developed and tested three mathematical models (**CP**, **GM**, **IP**) to reduce the dimensionality of the 3D-GRP which describe the relationships present within the whole-genome contact map. These mathematical models differ in terms of their problem representation, underlying non-procedural implementation and overall generalizability. Briefly, the **CP** mathematical model is encoded as a set of constraints over finite domains with constraint programming [35]. The **GM** mathematical model represents the problem as a maximum-weight matching [13] and is encoded as a logic program that uses Kolmogorov's Blossom V algorithm [21]. The **IP** mathematical model is encoded with integer programming [43] and is described in more detail below. Each model takes a normalized whole-genome contact map as input. As mentioned previously, such a contact map is a $N \times N$ matrix where the genome has been partitioned into N genomic bins. For a hypothetical whole-genome contact map (**A**), each cell ($A_{i,j}$) records the normalized interaction frequency between genomic bins i and j. By construction, the contact map is symmetric ($A_{i,j} = A_{j,i}$ for all i, j), and its main diagonal elements are all zero ($A_{i,i} = 0$ for all i).

The current formulations of the **CP** and **GM** mathematical models are only valid for haploid organisms whereas the **IP** mathematical model can be applied to organisms with any ploidy through an additional parameter called m. The value of m encodes the maximal number of interactions in which a given genomic bin can be involved based on the source organism's ploidy. For instance, m would be set to the following values based on the number of chromosome copies present: $m = 1$ (haploid), $m = 2$ (diploid; common in mammals), $m = 4$ (tetraploid; common in plants). Since the **IP** mathematical model is the most general, it will be the focus for the rest of this manuscript. Additional details on the **CP** and **GM** mathematical models can be found on the project homepage at GitHub [1] as well as in the first pre-print version of this paper [28].

The mathematical model **IP** uses integer programming [43] and is valid for any value of m. It is based on introducing variables $x_{i,j}$ that assume a value of 1 if genomic bin i interacts with genomic bin j, and 0 otherwise. The goal of this model is to solve $x_{i,j}$ for all i, j. The complete model is given in Mathematical Model 1. It was implemented in SICStus Prolog [2] [5] and solved using the mixed integer programming based Gurobi Optimizer [3] [18]. The implemented program using this representation with the hypothetical whole-genome contact map depicted in Figure 3A is shown in Additional File 1 [4]. An example associated data file for this program is given in

[1] https://github.com/kimmackay/GeneRHi-C/tree/master/step1
[2] http://sicstus.sics.se
[3] http://www.gurobi.com/
[4] https://github.com/kimmackay/GeneRHi-C/blob/master/step1/IP/supplementary_files/additional_file_1.pl



A	1	2	3	4	5	6
1	-	0.5	0.2	0.1	0.1	0.1
2	0.5	-	0.4	0.4	0.1	0.1
3	0.2	0.4	-	0.3	0.5	0.2
4	0.1	0.4	0.3	-	0.6	0.4
5	0.1	0.1	0.5	0.6	-	0.4
6	0.1	0.1	0.2	0.4	0.4	-

B	1	2	3	4	5	6
1	-	0.5	0.2	0.1	0.1	0.1
2	0.5	-	0.4	0.4	0.1	0.1
3	0.2	0.4	-	0.3	0.5	0.2
4	0.1	0.4	0.3	-	0.6	0.4
5	0.1	0.1	0.5	0.6	-	0.4
6	0.1	0.1	0.2	0.4	0.4	-

C	1	2	3	4	5	6
1	-	0.5	0.2	0.1	0.1	0.1
2	0.5	-	0.4	0.4	0.1	0.1
3	0.2	0.4	-	0.3	0.5	0.2
4	0.1	0.4	0.3	-	0.6	0.4
5	0.1	0.1	0.5	0.6	-	0.4
6	0.1	0.1	0.2	0.4	0.4	-

Figure 3: An example of two (of many) possible solutions to a 3D genome reconstruction problem. For all of the panels: the symmetric lower half of the contact map is indicated in light grey, the diagonal that represents "self-self" interactions is indicated in green and the genomic bin labels are represented in dark grey. For panels B and C: the blue boxes represent the subset of frequencies that could be selected as possible solutions (for $m = 1$). Panel B is a representation of a valid, non-optimal solution from the greedy algorithm and panel C is a representation of the valid optimal solution for the contact map where the sum of the selected interaction frequencies are 1.3 and 1.4, respectively.

Additional File 2 [5] and is based on the interaction frequency values from the hypothetical whole-genome contact map depicted in Figure 3A. For the fission yeast results, all of this has been automated in a makefile that is available on the project homepage [6].

maximize

$$\sum_{(i,j)\in E} A_{i,j}x_{i,j} \tag{1}$$

subject to:

$$\sum_{(i,j)\in E} x_{i,j} + \sum_{(j,i)\in E} x_{j,i} \leq m, \ \forall i \in V \tag{2}$$

$$x_{i,j} \in \{0,1\}, \ \forall (i,j) \in E \tag{3}$$

Mathematical Model 1: The IP model, for any m. V is the set $\{1,\ldots,N\}$ representing the genomic bins. E is the set $\{(i,j) \mid i < j \wedge A_{i,j} > 0\}$ representing the interactions whose frequencies (weights) are given by A.

2.2 Step 2: Graph Representation

The reduced set of interactions is converted into an undirected graph based on the graphical representation of Hi-C data described in GrapHi-C [29]. Briefly, the nodes in the network represent the individual genomic bins of the whole-genome contact map and the edges represent either selected interactions between bins or known linear interactions between adjacent bins. Linear interactions add additional biological constraints by representing the *bonafide in vivo* linear connections between bins (i.e. the linear extent of the chromosome). Unlike GrapHi-C, each edge is weighted using either: the interaction frequency divided by a dynamics coefficient for *cis*- and *trans*- interactions (described in more detail below) or the experimental resolution for linear interactions.

[5] https://github.com/kimmackay/GeneRHi-C/blob/master/step1/IP/supplementary_files/additional_file_2.csv

[6] https://github.com/kimmackay/GeneRHi-C/tree/master/step1/IP

2.3 Step 3: Calculation of (x, y, z) Coordinates

Finally, the ForceAtlas 3D network layout algorithm provided as a Gephi plugin [7] (which is an extension of the ForceAtlas2 layout [19]) is used to calculate (x, y, z) coordinates for the centre of each node in the network. Recall that each node represents a genomic bin from the whole-genome contact map.

3 PROBLEM DECOMPOSITION

When the **IP** implementation for step 1 was run on a complete fission yeast whole-genome contact map there was only a single *trans*-chromosomal interaction within the solution set making it difficult to infer the organization of the chromosomes in relation to each other. The low number of *trans*-chromosomal interactions is likely due to the fact that *cis*-interactions are known to have higher interaction frequencies than *trans*-interactions within the genome [9, 23]. This makes it more likely for *cis*-interactions to be included in the solution set since the goal of the mathematical models described above is to select a maximal subset of interaction frequencies. Since the disparity between *cis*- and *trans*- interaction frequencies is an inherent characteristic of whole-genome contact maps, all optimization-based solutions to the 3D-GRP must use an additional strategy to overcome this.

There are a number of possible strategies to deal with the disparity between *cis*- and *trans*- interaction frequencies such as data transformation or problem decomposition. Here the original computational problem (the 3D-GRP) has been decomposed into subproblems (described in more detail below) where the *cis*-chromosomal subproblems represent the individual chromosome structure while the *trans*-chromosomal subproblems represent how the chromosomes are organized in relation to each other within the nucleus. Each subproblem has been locally solved and the results are combined to retain the selected interactions from each subproblem. This is similar to the divide-and-conquer strategy employed by miniMDS which aims to first solve local substructures and then fit the results onto a global organization [34].

[7] https://gephi.org/plugins/

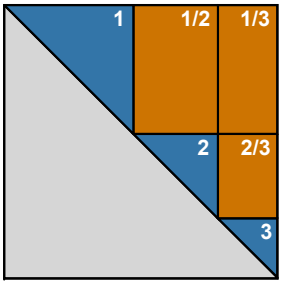


Figure 4: Identification of subproblems within the fission yeast contact map. The large grey triangle represents the portion of the contact map that does not need to be processed since all contact maps are mirrored along the diagonal. The blue triangles represent the subsections of the contact map that correspond to intra-chromosomal interactions, while the orange squares represent the subsections that correspond to the inter-chromosomal interactions. The label on the blue and orange areas represent the chromosome(s) involved in the interactions within that subsection of the contact map. In terms of the intra-chromosomal interactions, chromosome 1 contains the largest number of genomic bins while chromosomes 2 and 3 account for 34 and 80 percent fewer bins, respectively.

A single whole-genome contact map can be naturally divided into a finite, organism specific number of subproblems representing its constituent *cis*-interactions and pairwise *trans*-interactions. Each subproblem can be defined within the whole-genome contact map by specifying the range of genomic bins that correspond to the *cis*- or *trans*-interactions for each chromosome. In general, the number of subproblems for a whole-genome contact map with k chromosomes is equal to $\frac{k(k-1)}{2} + k$ where $\frac{k(k-1)}{2}$ represents the number of pairwise *trans*-interaction subproblems and k represents the number of *cis*-interaction subproblems. For example, because fission yeast has three chromosomes, its whole-genome contact map can be naturally partitioned into six subproblems (three *cis*- and three *trans*-interaction subproblems) to be solved in parallel. The location of these subproblems within a fission-yeast whole-genome contact map are depicted in Figure 4.

In order to solve the entire 3D-GRP, programs corresponding to the *cis*-interaction and pairwise *trans*-interaction subproblems can be generated and run independently. The solutions from each subproblem are then combined to reconstruct the entire 3D genomic model. This step is a heuristic which utilizes a novel metric (called the "dynamics coefficient") to account for the instances when a single genomic region participates in more than m subproblem solutions; i.e. more than m interactions. Instead of discarding interactions from subproblem solutions involving the same genomic region when this region has already been selected in m interactions, each identified interaction is maintained and associated with a region-specific dynamics coefficient to encode the mobility (or lack of mobility) of that genomic region. Briefly, the dynamics coefficient for each genomic region is calculated by scanning all of the resultant files for each subproblem and counting how many times a specific genomic bin is found across the subproblem solution sets. The more interactions a genomic region is involved in, the higher its corresponding dynamics coefficient, and *vice versa*.

In general, the dynamics coefficient is an integer value in the range of 0 to k where k is the number of chromosomes present in the genome. For example, in fission yeast ($k = 3$) if genomic bin 1 was involved in an interaction in the solution sets of the chromosome 1 *cis*-interaction subproblem and the chromosome 1/2 *trans*-interaction subproblem it would have a dynamics coefficient of 2, whereas if it was involved in an interaction in each of the relevant *trans*-interaction subproblems and the *cis*-interaction subproblem it would have an associated dynamics coefficient of 3. A higher dynamics coefficient suggests that the corresponding genomic region has been more mobile within the genome and that there is less certainty about its fixed position within the model. This is similar to the B factor (also known as the temperature factor or the Debye-Waller factor) generated with protein x-ray crystallography experiments [22]. The B factor encodes the degree of uncertainty associated with computed atomic positions in 3D space.

The dynamics coefficient is used to calculate edge weights for *cis*- and *trans*-interactions in step 2 of the GeneRHi-C workflow ($A_{i,j}/d$ where $A_{i,j}$ is the interaction frequency and d is the dynamics coefficient). These weights are then used to visualized the predicted model. Although this use causes violation of the initial ploidy restriction used to constrain each subproblem, it is still biologically valid. As mentioned previously, it is possible for a given genomic bin to be involved in more than one interaction in 3D space [6, 14], even though Hi-C is only able to detect one pairwise interaction per restriction site within a single haploid cell. Additionally, the dynamics coefficient allows the program to encode some of the mobility of genome organization into the predicted model by representing the certainty of whether an interaction is fixed within the population of cells. Overall, this decomposition approach results in a larger number of *trans*-chromosomal interactions being included in the final solution set. It also has been applied to the CP and GM mathematical models and should be utilized in future applications of GeneRHi-C.

4 RESULTS

The above workflow was tested with an existing fission yeast Hi-C dataset (GEO accession number: GSM1379427 [31]). All programs were run on a server-grade computer with sufficient main memory to represent the entire problem. All times reported in this and subsequent sections are elapsed times. Results from the **CP** and **GM** mathematical models can be found on the project homepage at GitHub [8] as well as in the first pre-print version of this paper [28].

4.1 Step 1: Dimensionality Reduction

The implementation of the **IP** model for the complete whole-genome fission yeast contact map ($m = 1$; $|V| = 1258$, $|E| = 745595$) was able to identify the optimal solution set in 294.44 seconds using the Gurobi optimizer. As mentioned above (and depicted in Figure 6A), only one *trans*-chromosomal interaction was represented in the solution set. This outcome made it difficult to infer the organization

[8] https://github.com/kimmackay/GeneRHi-C/tree/master/step1

Table 1: Subproblem sizes and elapsed times (runtime) for the IP mathematical model applied to the fission yeast dataset.

Subproblem	Number of Vertices ($\|V\|$)	Number of Edges ($\|E\|$)	Run Time (seconds)
chromosome 1 *cis*-interaction	558	148734	15.75
chromosome 2 *cis*-interaction	454	96562	9.26
chromosome 3 *cis*-interaction	246	27255	2.06
chromosome 1/2 *trans*-interaction	454	241022	4.90
chromosome 1/3 *trans*-interaction	246	128472	4.70
chromosome 2/3 *trans*-interaction	246	103550	3.40

of the chromosomes in relation to each other. In order to overcome this, the decomposition approach described above was used. Six separate subprograms were generated and run independently. The size of each subproblem in terms of V and E, as well as the time it took to identify the optimal solution, is presented in Table 1 (total summed run time of 40.07 seconds).

4.2 Step 2: Graph Representation

The results from each subproblem were combined as described above and converted into a GrapHi-C [29] representation using the `generate_gephi_input_subproblems.pl` script [9]. Please note this is an extension of the original GrapHi-C script that allows for the dynamics coefficient to be included in the edge-weight calculation. This step took less than 1 second of execution time.

4.3 Step 3: Calculation of (x, y, z) Coordinates

The graph generated in Step 2 was used as input to the ForceAtlas 3D network layout algorithm. The resulting layout was exported to a `.gexf` file and (x, y, z) coordinates were extracted using the `gexf2xyz.py` script on the project homepage [10]. This step took less than 10 seconds of execution time.

4.4 Visualization

The results were visualized in Gephi (Figure 5) [3]. Nodes were coloured according to their chromosome number (Panel A) or genomic feature (Panel B). We would like to stress that this graph-based visualization is not a polymer model of the DNA chain that is often seen in other 3D genome prediction tools. Therefore, the smoothness of the edges is not a result of any bending rigidity constraints. Instead, it is a result of the visualization tool (Gephi) and the network layout applied.

One of the most well–documented features of fission yeast genomic organization is the 3D clustering of centromeres and telomeres within the nucleus [7, 15]. In order to determine whether the yeast model predicted by GeneRHi-C was able to recapitulate these features, the genomic bins corresponding to centromeres and telomeres were coloured in the Gephi visualization (see Figure 5B). This figure provides evidence that the predicted genome model is consistent with established principles of fission yeast chromosomal organization including: (1) chromosomal organization into a hemispherical region, (2) a single centromere cluster and (3) the presence of two telomere clusters (chromosome 1/2) located near the nuclear periphery, opposite the centromere cluster [30, 31, 41]. Additionally, the clustering in the GeneRHi-C predicted model is consistent with the clustering seen in previous fission yeast 3D genome predictions [39]. This provides confidence in the accuracy that was achieved using the GeneRHi-C workflow with fission yeast data. This type of evaluation (comparison to known genome and chromosome structural features) is typical within the 3D genome community [33, 40]. Future work (described in Section 5.2) will

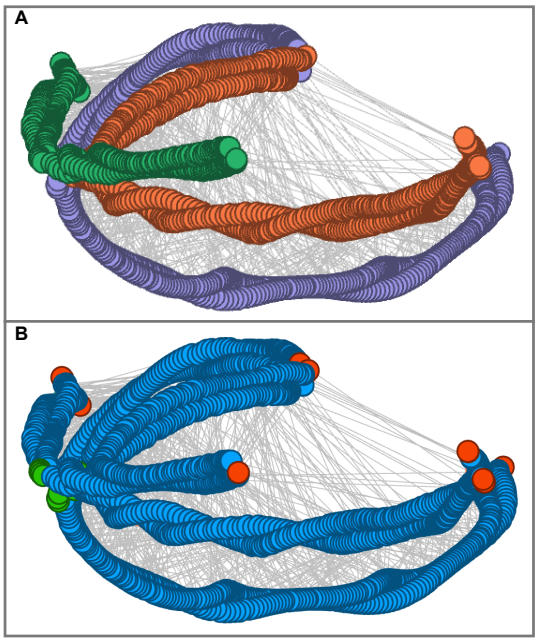


Figure 5: Visualization of the Predicted Genome Model Using the IP Model. Circles depict the genomic bins, grey lines represent *cis*- and *trans*-interaction edges selected by the IP model, and line lengths are proportional to original edge-weight calculated in step 2. In Panel A, circles are coloured according to their corresponding chromosome (CHR1: purple, CHR2: orange, CHR3: green). In Panel B, the following genomic features are highlighted: telomeres (red), centromeres (green) and nuclear DNA (blue).

[9] https://github.com/kimmackay/GeneRHi-C/tree/master/step2/scripts
[10] https://github.com/kimmackay/GeneRHi-C/tree/master/step3

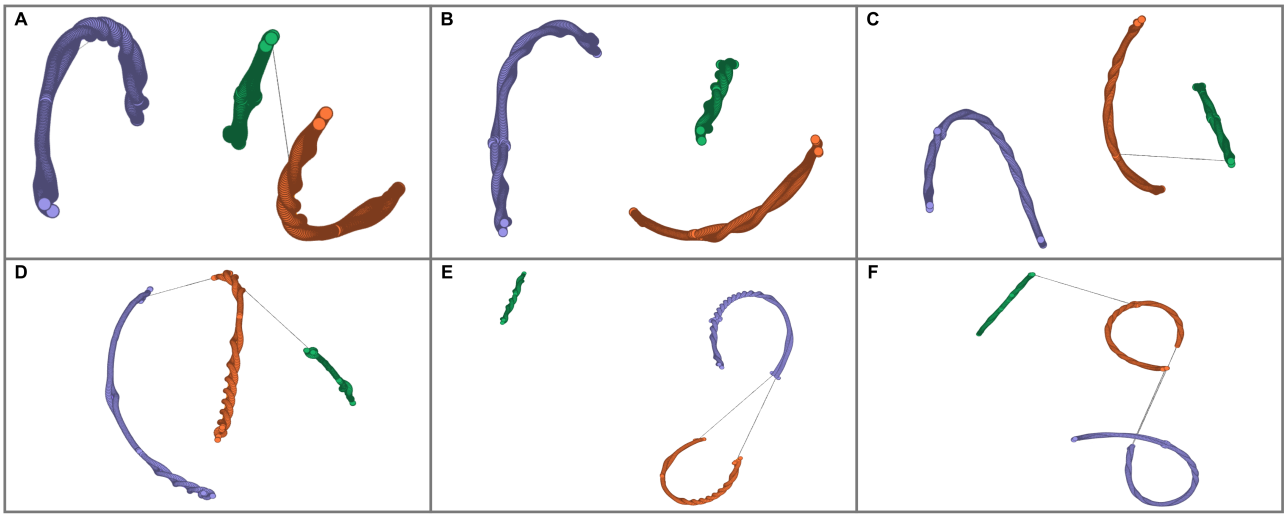


Figure 6: Visualization of the Predicted Genome Model Using the IP Model with Various m Values. Circles depict the genomic bins, grey lines represent *trans*-interaction edges selected by the IP model, and line lengths are proportional to the associated interaction frequency ($A_{i,j}$). Circles are coloured according to their corresponding chromosome (CHR1: purple, CHR2: orange, CHR3: green). The results for each m values are presented in the following panels: A ($m = 1$), B ($m = 2$), C ($m = 3$), D ($m = 4$), E ($m = 5$), F ($m = 6$).

extend this evaluation to provide a more comprehensive snapshot of GeneRHi-C's reconstruction ability.

5 DISCUSSION

5.1 Effect of m on Genome Organization in Fission Yeast

As mentioned previously, it is possible that each genomic region could be involved with more than one interaction within the genome but is restricted to m Hi-C interactions (where m is organism ploidy). To determine whether or not relaxing this restriction would result in a more comprehensive genomic model without having to use the decomposition approach described in Section 3, the GeneRHi-C workflow with the **IP** mathematical model was tested with values of m from 1 to 6 for the same fission yeast Hi-C dataset (GSM1379427 [31]). As mentioned previously, the **IP** mathematical model allows for a single genomic bin to be involved in more than one Hi-C mediated interaction in the predicted genome organization. Recall that the **CP** and **GM** mathematical models could not be used for this since they are only valid for $m = 1$.

For each value of m, the program was able to find an optimal solution in 294.44, 13.20, 104.46, 15.31, 38.79, and 16.94 seconds for $m = 1..6$, respectively. The results were visualized using Gephi (Figure 6). Nodes were coloured according to their chromosome number. The nodes in the graph represent the individual genomic bins of the whole-genome contact map and the edges represent either selected interactions between bins or known linear interactions between adjacent bins. The results presented in Figure 6 indicate that relaxing the ploidy restriction (by increasing the value of m) did not result in a more comprehensive genomic model. This is consistent with the work of Diament and Tuller [10] which suggested that as little as 5 % of the original Hi-C data is required to generate biologically accurate 3D genome models [36].

Minimal numbers of *trans*-chromosomal interactions were selected by the model regardless of what the parameter m was set to. Specifically the following number of *trans*-chromosomal interactions were observed in each solution set: 1 ($m = 1$), 0 ($m = 2$), 1 ($m = 3$), 2 ($m = 4$), 2 ($m = 5$), 3 ($m = 6$). As mentioned previously, this is likely due to the fact that *cis*-chromosomal interactions occur more frequently than *trans*-chromosomal interactions within the genome resulting in higher interaction frequency values [9]. The decomposition strategy described in Section 3 can again be applied to circumvent this issue.

5.2 Evaluation

The 3D genomics community does not have a standardized methodology for evaluating 3D genome reconstruction tools like GeneRHi-C. Ideally, 3D genome reconstruction tools would be evaluated using synthetic Hi-C datasets with known ground-truth structures. The 3D models predicted from these synthetic datasets could then be compared with their ground truth counter-parts using measures like the Spearman correlation coefficient and root-mean-square deviation. Unfortunately, a standardized dataset of synthetic 3D structures and associated Hi-C matrices for tool evaluation does not exist. We are actively working towards the creation of this type of data. Once it has been generated it will be used to evaluate GeneRHi-C's reconstruction ability using the methodology described above. This methodology could also be employed to develop a ranked-list of all existing 3D genome reconstruction tools.

5.3 Application to Organisms with Higher Ploidies and/or Larger Genomes

The **IP** mathematical model described above could be applied to organisms with higher ploidies by specifying the value of the m parameter. For instance, m could be modified in the following ways according to the number of chromosome copies present: $m = 2$ (diploid; common in mammals), $m = 4$ (tetraploid; common in plants), and so on. One issue that would need to be addressed in organisms with higher ploidies is phasing the interactions to each chromosome copy. This could potentially be solved using existing phasing tools [8] and additional biological data [4, 37].

Utilizing the decomposition approach described in Section 3 allows one to take advantage of coarse-grained parallelism ensuring the mathematical models are scaleable to organisms with larger genomes (and more chromosomes). This type of parallelism is easy to obtain on many types of computational infrastructure. As an example, this decomposition could be easily applied to a whole-genome contact map from *Homo sapiens*. This contact map would result in the generation of 276 subproblems (given $k = 23$ and the number of subproblems = $\frac{k(k-1)}{2} + k$). It should be noted that the size of these subproblems would likely be larger than what was seen in the fission yeast example (depending on the Hi-C resolution).

5.4 Future Work

Future work will focus on the validation, modification and extension of the GeneRHi-C. Specifically, an extensive biological validation of the predicted genome models will be performed with targeted chromosome conformation capture assays and advanced microscopy techniques to better characterize the biological accuracy of the developed mathematical model. Different types of data transformations will be explored to address the disproportionate numbers of *cis*- and *trans*-chromosomal interactions in the whole-genome contact map in case they result in a better alternative to the decomposition approach described in Section 3. The **IP** mathematical model will be utilized as a computational framework which will be extended and further developed to incorporate a variety of additional genomic datasets and information types into the prediction process. For example, each genomic bin could have an associated list of variables representing the genes found within that bin and their corresponding gene expression values. Constraints could then be applied to favour interactions between regions with similar expression profiles. The **IP** mathematical model will also be utilized as a starting point for predicting the 3D genomic structure of organisms with higher ploidies by applying the modifications suggested in Section 5.3.

6 CONCLUSION

This is the first time a non-procedural programming approach has been used to model the 3D genome organization from Hi-C data. Specifically, we developed a three-step consensus method (called GeneRHi-C; pronounced “generic") for solving the 3D-GRP which utilizes both new and existing techniques. Briefly, (1) the **IP** mathematical model is used to reduce the dimensionality of the 3D-GRP by identifying a biologically plausible, ploidy-dependent subset of interactions from the Hi-C data. A decomposition approach is used in this step to generate a more comprehensive 3D genome organization. $\frac{k(k-1)}{2} + k$ separate subproblems (one for each set of *cis*-chromosomal interactions and one for each set of pairwise *trans*-chromosomal interactions) are independently solved and combined. A novel coefficient is defined to aid in combining the results of each subproblem (the dynamics coefficient) which allows a level of positional uncertainty to be encoded into the predicted genomic organization. The second step (2) generates a biological network (graph) that represents the subset of interactions identified in the previous step. Briefly, genomic bins are represented as nodes in the network with weighted-edges representing known and detected interactions. Finally, the third step (3) uses the ForceAtlas 3D network layout algorithm to calculate (x, y, z) coordinates for each genomic region in the contact map. The resultant predicted genome organization represents the interactions of a population-averaged consensus structure. The GeneRHi-C workflow was tested with Hi-C data from *Schizosaccharomyces pombe* (fission yeast). This predicted 3D genome organization was then validated through literature search which verified that the GeneRHi-C prediction recapitulated key documented features of the yeast genome. Overall, GeneRHi-C demonstrates the power of non-procedural programming and graph theoretic techniques for providing an efficient, generalizable solution to the 3D-GRP and is a step towards a better understanding of the relationship between genomic structure and function.

A SUPPLEMENTAL INFORMATION

The following files are available on the project homepage. **Additional file 1:** The implemented program using the **IP** mathematical model. **Additional file 2:** An example data file (`.csv` file) depicting the interaction frequencies from the hypothetical whole-genome contact map depicted in Figure 3A utilized by the **IP** model. **Availability of data and material:** The datasets utilized in this article are available in the Gene Expression Omnibus database (accession number: GSM1379427; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1379427). **Software information: Project Name:** GeneRHi-C (pronounced "generic"); **Project Homepage:** https://github.com/kimmackay/GeneRHi-C; **Programming language(s):** SICtus Prolog (Step 1), Perl (Step 2), Python (Step 3); **Other requirements:** A local version of the Gurobi solver is required to run the **IP** model (commercial software that is freely available to academic users).

B FUNDING

This work was supported by the Natural Sciences and Engineering Research Council of Canada [RGPIN 37207 to AK, Vanier Canada Graduate Scholarship to KM].

C AUTHOR’S CONTRIBUTIONS

KM, MC, AK developed the **CP**, **GM** and **IP** mathematical models. MC implemented and tested the mathematical models. KM visualized the results and verified accuracy. KM wrote the manuscript. MC, AK edited the manuscript. AK supervised the research.

ACKNOWLEDGMENTS

We could like to thank Dr. Christopher Eskiw, Morgan W.B. Kirzinger and Conor Lazarou for their input and advice.

REFERENCES

[1] Ferhat Ay and William S. Noble. 2015. Analysis methods for studying the 3D architecture of the genome. *Genome Biology* 16, 1 (September 2015), 1–15.

[2] Sepideh Babaei, Waseem Akhtar, Johann de Jong, Marcel Reinders, and Jeroen de Ridder. 2015. 3D hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes. *Nature Communications* 6 (2015), 6381.

[3] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media* (2009).

[4] Shay Ben-Elazar, Benny Chor, and Zohar Yakhini. 2016. Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data. *Bioinformatics* 32 (2016), i559–i566.

[5] Mats Carlsson and Per Mildner. 2012. SICStus Prolog - The first 25 years. *Theory and Practice of Logic Programming* 12, 1-2 (2012), 35–66. https://doi.org/10.1017/S1471068411000482

[6] Andrea M. Chiariello, Carlo Annunziatella, Simona Bianco, Andrea Esposito, and Mario Nicodemia. 2016. Polymer physics of chromosome large-scale 3D organisation. *Scientific Reports* 6 (July 2016), 29775.

[7] Yuji Chikashige, Da-Qiao Ding, Yoshiyuki Imai, Masayuki Yamamoto, Tokuko Haraguchi, and Yasushi Hiraoka. 1997. Meiotic nuclear reorganization: switching the position of centromeres and telomeres in the fission yeast Schizosaccharomyces pombe. *The EMBO Journal* 16, 1 (1997), 193–202.

[8] Yongwook Choi, Agnes P. Chan, Ewen Kirkness, Amalio Telenti, and Nicholas J. Schork. 2018. Comparison of phasing strategies for whole human genomes. *PLOS Genetics* 14, 4 (2018), e1007308.

[9] Job Dekker and Leonid Mirny. 2016. The 3D Genome as Moderator of Chromosomal Communication. *Cell* 164, 6 (2016), 1110–1121.

[10] Alon Diament and Tamir Tuller. 2015. Improving 3D Genome Reconstructions Using Orthologous and Functional Constraints. *PLOS Computational Biology* 11, 5 (2015), e1004298.

[11] Pengfei Dong, Xiaoyu Tu, Po-Yu Chu, Peitao Lü, Ning Zhu, Donald Grierson, Baijuan Du, Pinghua Li, and Silin Zhong. 2017. 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Molecular Plant* 10, 12 (2017), 1497–1509.

[12] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C. Anthony Blau, and William S. Noble. 2010. A three-dimensional model of the yeast genome. *Nature* 465 (2010), 363–367.

[13] Jack Edmonds. 1965. Paths, trees, and flowers. *Canadian Journal of Mathematics* 17, 3 (1965), 449–467.

[14] Stephanie Fanucchi, Youtaro Shibayama, Shaun Burd, Marc S. Weinberg, and Musa M. Mhlanga. 2013. Chromosomal contact permits transcription between coregulated genes. *Cell* 115, 3 (2013), 606–620.

[15] Hironori Funabiki, Iain Hagan, Satoru Uzawa, and Mitsuhiro Yanagida. 1993. Cell Cycle-dependent Specific Positioning and Clustering of Centromeres and Telomeres in Fission Yeast. *Journal of Cell Biology* 121 (June 1993), 961–976.

[16] Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S. Liu. 2013. Bayesian inference of spatial organizations of chromosomes. *PLOS Computational Biology* 9, 1 (January 2013), e1002893.

[17] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R. Lajoie, Job Dekker, and Leonid A. Mirny. 2012. Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization. *Nature Methods* 9, 10 (2012), 999–1003.

[18] Gurobi Optimization Inc. 2014. Gurobi Optimizer Reference Manual. http://www.gurobi.com http://www.gurobi.com.

[19] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* 9, 6 (2014), e98679.

[20] Philip A. Knight and Daniel Ruiz. 2012. A fast algorithm for matrix balancing. *Journal of Numerical Analysis* 33, 3 (2012), 1029–1047.

[21] Vladimir Kolmogorov. 2009. Blossom V: a new implementation of a minimum cost perfect matching algorithm. *Mathematical Programming Computation* 1, 1 (2009), 43–67.

[22] Antonija Kuzmanic, Navraj S. Pannu, and Bojan Zagrovic. 2014. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nature Communications* 5 (2014), 3220.

[23] Bryan R Lajoie, Job Dekker, and Noam Kaplan. 2015. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* 72 (January 2015), 65–75.

[24] Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 2014. 3D genome reconstruction from chromosomal contacts. *Nature Methods* 11 (March 2014), 1141–1143.

[25] Wenyuan Li, Ke Gong, Qingjiao Li, Frank Alber, and Xianghong Jasmine Zhou. 2015. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics* 31, 6 (2015), 960–962.

[26] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. 2009. Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science* 326, 5950 (October 2009), 289–293.

[27] Chang Liu, Ying-Juan Cheng, Jia-Wei Wang, and Detlef Weigel. 2017. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nature Plants* 3, 9 (2017), 742–748.

[28] Kimberly MacKay, Mats Carlsson, and Anthony Kusalik. 2018. *SonHi-C: a set of non-procedural approaches for predicting 3D genome organization from Hi-C data.* Technical Report. bioRxiv 392407.

[29] Kimberly MacKay, Anthony Kusalik, and Christopher H. Eskiw. 2018. GrapHi-C: graph-based visualization of Hi-C datasets. *BMC Research Notes* 11, 1 (2018), 418. https://doi.org/10.1186/s13104-018-3507-2

[30] Takeshi Mizuguchi, Jemima Barrowman, and Shiv IS Grewal. 2015. Chromosome domain architecture and dynamic organization of the fission yeast genome. *FEBS Letters* 589, 20 part A (2015), 2975–2986.

[31] Takeshi Mizuguchi, Geoffrey Fudenberg, Sameet Mehta, Jon-Matthew Belton, Nitika Taneja, Hernan Diego Folco, Peter FitzGerald, Job Dekker, Leonid Mirny, Jemima Barrowman, and Shiv IS Grewal. 2014. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature* 516, 7531 (December 2014), 432–435.

[32] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502 (October 2013), 59–64.

[33] Oluwatosin Oluwadare, Max Highsmith, and Jianlin Cheng. 2019. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biological Procedures Online* 21 (2019), 7.

[34] Lila Rieber and Shaun Mahony. 2017. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* 33, 14 (2017), i261–i266.

[35] Francesca Rossi, Peter van Beek, and Toby Walsh (Eds.). 2006. *Handbook of Constraint Programming.* Elsevier, New York, NY, USA.

[36] Mark R Segal and Henrik L Bengtsson. 2015. Reconstruction of 3D genome architecture via a two-stage algorithm. *BMC Bioinformatics* 16 (November 2015), 373.

[37] Siddarth Selvaraj, Jesse R Dixon, Vikas Bansal, and Bing Ren. 2013. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology* 31, 12 (2013), 1111–1118.

[38] François Serra, Marco Di Stefano, Yannick G. Spill, Yasmina Cuartero, Michael Goodstadt, Davide Baù, and Marc A. Marti-Renom. 2015. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Letters* 589, 20 (May 2015), 2987–2995.

[39] Hideki Tanizawa, Osamu Iwasaki, Atsunari Tanaka, Joseph R Capizzi, Priyankara Wickramasinghe, Mihee Lee, Zhiyan Fu, and Ken ichi Noma. 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research* 38, 22 (October 2010), 8164–8177.

[40] Tuan Trieu and Jianlin Cheng. 2014. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research* 42, 7 (January 2014), e52.

[41] S Uzawa and M Yanagida. 1992. Visualization of centromeric and nucleolar DNA in fission yeast by fluorescence in situ hybridization. *Journal of Cell Science* 101, Pt 2 (1992), 267–275.

[42] Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* 30, 12 (2014), i26–i33.

[43] Laurence A Wolsey. 1998. *Integer Programming.* Wiley.