Inferring whole-genome 3D chromatin structures from diploid Hi-C data

Gesine Cauer

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
William S. Noble, Chair
Maitreya Dunham
Brian Beliveau

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Inferring whole-genome 3D chromatin structures from diploid Hi-C data

Gesine Cauer

Chair of the Supervisory Committee:
Professor William S. Noble
Department of Genome Sciences

The three-dimensional organization of the genome plays an important part in regulating numerous basic cellular functions, including gene regulation, differentiation, the cell cycle, DNA replication, and DNA repair. Assays like Hi-C measure DNA-DNA contacts in a high-throughput fashion, and inferring accurate 3D models of chromosomes can yield insights hidden in the raw data. For example, structural inference can account for noise in the data, disambiguate the distinct structures of homologous chromosomes, orient genomic regions relative to nuclear landmarks, and serve as a framework for integrating other data types. Accordingly, many methods have been developed to infer 3D structures from Hi-C data.

However, many challenges remain. Importantly, although many methods exist to infer the 3D structure of haploid genomes, accurately inferring a diploid structure from Hi-C data is still an open problem. Indeed, the diploid case is very challenging, because Hi-C data does not typically distinguish between homologous chromosomes. Inference is also complicated in the setting of low-coverage or high-resolution data, which can lead to poor performance and high computational costs.

This work describes two methods for inferring 3D diploid chromatin structures from Hi-C data. The first approach extends a previously published haploid method and enables diploid inference via the addition of two constraints. We demonstrate the accuracy of this method on simulated data, and we also use the method to infer 3D structures for mouse chromosome X, confirming that the

inactive homolog exhibits a bipartite structure, whereas the active homolog does not.

Our second method addresses the difficulties presented by low-coverage or high-resolution data via multiscale optimization, an optimization strategy that solves a large optimization problem by building upon the solutions to smaller versions of the problem. Similar approaches have been successfully employed in the context of haploid structural inference methods. However, because many organisms of interest are diploid, we sought to develop a multiscale optimization approach that infers the structure of diploid genomes. We use simulations to show that integrating multiscale optimization into our first method significantly improves the accuracy of inferred structures.

TABLE OF CONTENTS

Page

List of Figures iii

Chapter 1: Introduction 1
1.1 Experimental techniques used to assess chromatin structure 1
1.2 The utility of working with 3D chromatin structures 2
1.3 Methods to infer 3D chromatin structure from contact count matrices 3
1.4 Challenges in chromatin structural inference 4
1.5 Organization of dissertation 5

Chapter 2: Inferring diploid 3D chromatin structures from Hi-C data 6
2.1 Introduction 6
2.2 Method 8
2.3 Results 15
2.4 Discussion 19

Chapter 3: A multi-resolution optimization strategy for inferring 3D genome architecture from Hi-C data 21
3.1 Introduction 22
3.2 Approach 26
3.3 Methods 38
3.4 Results 45

Chapter 4: Conclusion 50
4.1 Significance of contributions 50
4.2 Further validating inferred homologs 51
4.3 Limitations of our approach 51
4.4 Potential improvements on our method 53

4.5 Inferring single-cell chromatin structures 54
4.6 Comparing and integrating our method with microscopy data 55

Appendix A: Appendix to "Inferring diploid 3D chromatin structures from Hi-C data" . . . 68

Appendix B: Appendix to "A multi-resolution optimization strategy for inferring 3D genome architecture from Hi-C data" 69

LIST OF FIGURES

Figure Number Page

2.1 **Inferring 3D structure using ambiguous diploid data.** Each observed count (left) corresponds to a sum of four pairs of genomic loci (right). The Poisson model must be adjusted to account for this ambiguity. 10

2.2 **Constraints improve ambiguous inference.** The simulated data consists of a single diploid chromosome with 9.3×10^6 reads and 343 beads, the size of which corresponds to mouse chromosome X at 500 kb resolution. (A) The quality of the inferred structure, as measured by three different error scores (y-axis), improves upon application of the bead-chain connectivity constraint ($\lambda_1 = 10^8$) and the homolog separation constraint ($\lambda_2 = 10^10$). Best results are seen when both constraints are applied simultaneously. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. "Null" indicates inference performed without the Poisson model. Corresponding *p*-values are in Supplementary Table A.1. (B) A simulated chromosome is shown alongside inferred versions of the same chromosomes using various strategies. Each panel also lists the RMSD and distance error associated with the given structure, relative to the true structure. 16

2.3 **Inference with ambiguous and disambiguated data.** The simulated data consists of a single chromosome with 9.3^6 reads and 343 beads, the size of which corresponds to mouse chromosome X at 500 kb resolution. The quality of the inferred structure, as measured by three different error scores (y-axis), improves when one or both ends of a contact are disambiguated, and best results are seen in the latter case. "A" indicates ambiguous data, "U" indicates unambiguous data, and "P" indicates partially ambiguous data. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Corresponding *p*-values are in Supplementary Table A.2. 17

2.4 **Bipartite structure of the mouse inactive X chromosome.** The bipartite index (y-axis) at each genomic distance bin (x-axis) for the active (orange) and inactive (blue) homologs of the mouse X chromosome. The black line corresponds to the known boundary between superdomains of the inactive homolog at bin 146. 18

3.1 **Multi-resolution optimization improves diploid inference on simulated ambiguous data.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred via three different inference methods (x-axis): without multi-resolution optimization, multi-resolution optimization with the naive approach, and multi-resolution optimization with our novel negative binomial model. The α parameter was either fixed at the same value used during simulation **(A)** or jointly inferred alongside the structure **(B)**. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, Bonferroni corrected p-value < 0.05). There were significant improvements in the combined error score (y-axis, Section 3.3.3) when multi-resolution optimization was applied with the negative binomial model, relative to multi-resolution optimization with the naive approach or inference without multi-resolution optimization. Each of the ten individual datasets shows the best results with the negative binomial multi-resolution model. For results on each constituent of the combined error score, see Supplementary Figures B.1 and B.2. 49

3.2 **Novel constraints improve diploid inference on simulated ambiguous data.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred via two different inference methods: without multi-resolution optimization **(A)** and multi-resolution optimization with our novel negative binomial model **(B)**. Inference was performed using either the two constraints described in Cauer *et al.*, 2019 [18] or our two novel constraints (x-axis). The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, p-value < 0.05). The combined error score (y-axis, Section 3.3.3) is significantly better when the novel constraints were applied, and the magnitude of difference is most pronounced in the context of multi-resolution optimization. However, in both cases, each of the ten individual datasets shows the best results with the new constraints. For results on each constituent of the combined error score, see Supplementary Figures B.3 and B.4. 49

B.1 **When simulated diploid structures are inferred from ambiguous data with α fixed at the true value, multi-resolution optimization improves accuracy along multiple measures.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred via three different inference methods (x-axis): without multi-resolution optimization, multi-resolution optimization with the naive approach, and multi-resolution optimization with our novel negative binomial model. The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, Bonferroni corrected p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The negative binomial-based multi-resolution model significantly outperforms single-resolution inference on all six measures. When comparing between the two multi-resolution optimization strategies, the negative binomial model yields significantly better intra-molecular similarity scores as well as significantly better intra-chromosomal inter-homolog distance error **(F)**. 70

B.2 **When jointly inferring simulated diploid structures and α from ambiguous data, multi-resolution optimization improves accuracy along multiple measures.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred via three different inference methods (x-axis): without multi-resolution optimization, multi-resolution optimization with the naive approach, and multi-resolution optimization with our novel negative binomial model. The value of α was jointly inferred alongside the structure. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, Bonferroni corrected p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The negative binomial-based multi-resolution model significantly outperforms single-resolution inference as well as inference with the naive multi-resolution model on all six similarity scores. Each of the ten individual datasets shows improvement on all measures with the negative binomial multi-resolution model. 71

B.3 **Novel constraints improve single-resolution diploid inference accuracy along multiple measures in the context of simulated ambiguous data.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred without multi-resolution optimization. Inference was performed using either the two constraints described in Cauer *et al*., 2019 [18] or our two novel constraints (x-axis). The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The novel constraints significantly outperform those of Cauer *et al*. with regard to intra-molecular both inter-molecular scores **(E, F)**. However, the Cauer *et al*. constraints significantly outperform ours on intra-molecular RMSD and distance error **(A, B)**. No significant differences are seen when comparing intra-molecular TM-score or GDT **(C, D)**. 72

B.4 **Novel constraints improve multi-resolution diploid inference accuracy along multiple measures in the context of simulated ambiguous data.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred via multi-resolution optimization with our novel negative binomial model. Inference was performed using either the two constraints described in Cauer *et al*., 2019 [18] or our two novel constraints (x-axis). The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The novel constraints significantly outperform those of Cauer *et al*. with regard to intra-molecular TM-score and GDT **(C, D)** as well as both inter-molecular scores **(E, F)**, but no significant differences are seen when comparing intra-molecular RMSD or distance error **(A, B)**. 73

ACKNOWLEDGMENTS

Bill Noble has been an incredible mentor and PI during my time in the lab. Thank you for your helpful guidance on all aspects of my work, I couldn't have wished for a better PI.

I would also like to thank my collaborators, who have invested a lot of time in my projects and provided me with truly invaluable advice and ideas over the years. Thanks to Jean-Philippe Vert for helping me develop the multi-resolution inference approach as well as teaching me about optimization and statistics, and thanks to Nelle Varoquaux for sharing her extensive knowledge about chromatin structural inference and best practices in software development.

To the members of my committee, thank you for the rich insight and advice you've provided me over the years. I also appreciated your guidance on structuring the timeline of my PhD.

In addition, I would like to thank all current and past members of the Noble lab for their perspectives and suggestions, and for providing a cheerful work environment. In particular, I would like to thank Gürkan Yardımcı, for his mentorship and guidance throughout the project described in Chapter 1, as well as Mozes Jacobs, for his hard work in improving functionality of and access to the PASTIS software.

Lastly, to my partner, thank you for your unwavering support throughout this long process.

DEDICATION

To my partner, Erik.

Chapter 1
INTRODUCTION

The three-dimensional organization of the genome plays an important part in regulating numerous basic cellular functions, including gene regulation [68, 80], differentiation [45, 24], the cell cycle [60], DNA replication, and DNA repair [51]. Numerous lines of evidence also link changes in diverse components of 3D genome architecture to many different diseases, including cancer, laminopathies, cohesinopathies, and limb malformation diseases [46]. Furthermore, chromatin has been shown to exhibit a hierarchy of 3D architectures, which undergo dynamic rearrangement during normal development. However, relatively little is known about the large-scale 3D structure of chromatin.

1.1 Experimental techniques used to assess chromatin structure

The three-dimensional organization of the genome can be experimentally assessed by both sequence- and imaging-based techniques.

Most sequence-based approaches rely on crosslinking of spacially adjacent chromatin fragments, followed by isolation and sequencing of the paired sequences. Many of these assays are based on chromatin conformation capture (3C) [22], an early technique that probed pairwise interactions between select loci. 3C and its derivatives use ligation to maintain connectivity between the crosslinked genomic fragments [50, 66, 37, 29, 59, 26, 39, 57, 60, 67]. These methods output interactions between pairs of loci. Ligation-independent techniques have also been developed [65, 5, 94, 20, 78, 69], many of which are able to assess multiway contacts between three or more simultaneously interacting loci in addition to pairwise contacts [65, 5, 94]. Some sequence-based methods enrich for specific proteins [29, 59, 26, 20, 78, 94] or genomic regions [39, 57, 69] of inter-

est, whereas others yield genome-wide interactions [50, 66, 37, 65, 5]. While these techniques were originally developed for bulk data, in which an entire population of cells is measured together, many have since been extended to work with data from single cells [60, 67, 44, 1]. Of these methods, none have been adopted as widely as Hi-C [50], an early genome-wide ligation-based technique.

Recent advances in super-resolution microscopy [32, 73, 33, 36], microfluidics, and high-throughput oligonucleotide-based FISH protocols [8, 6] have enabled assessment of genome-wide chromatin conformation with unprecedented resolution and throughput. Briefly, these techniques involve hybridizing different probes to sequential segments of each chromosome to enable super-resolution tracing of chromatin folding. Multiplexing is used to achieve massive throughput. The methods include multiplex FISH imaging [63, 86], ORCA [54], Hi-M [31], OligoFISSEQ [62], and OligoSTORM [7]. Due to the nature of these techniques, they provide data on the single-cell level. However, these methods have not been adopted as broadly as sequence-based techniques.

1.2 The utility of working with 3D chromatin structures

While direct analyses of contact counts can address many questions of interest, inferred 3D structures can provide additional value over the original counts matrices. For example, 3D models of genome architecture can facilitate interpretation, account for noise, and enable integration of other data types [14]. Structural inference can also orient multiple genomic regions relative to each other or to nuclear landmarks [3]. Such higher-order interactions arise from accurately inferring 3D structures from two-dimensional data. Chromatin structures also have potential to be used in ways not yet explored. For example, it might be interesting to determine whether the 3D structures themselves contain recurring features and patterns beyond those visible in contact counts matrices.

For diploid organisms, structural inference also has the potential to distinguish between homologous chromosomes [83, 64, 88, 9, 18, 21], as raw contact count data does not discriminate intra-homologous interactions from inter-homologous interactions. Phased heterozygous single nucleotide variant (SNV) data, if available for the sample being assayed, can be used to phase a small portion of the read pairs. Such data is typically derived from parental genotype information, although it can also be computationally extracted from the contact count itself using methods such

as HaploSeq [75]. Either way, the contact counts matrix composed of reads that are phased on each end is generally very sparse, thus limiting interpretation. Some researchers augment this matrix by assuming all contacts phased on only one end to be derived from intra-homologous interactions. However, this assumption is clearly not correct, and the resulting inaccuracies likely lead to problems when this data is used as input to subsequent analyses.

Diploid contact counts that are unphased or improperly phased cannot be used to compare between homologous chromosomes or assess their relative orientations. Such analyses are especially interesting in the context of chromosome X inactivation [53, 11], allele-specific expression [77, 70] and homolog pairing [43] Reliably phased contact counts data would also enable other potentially interesting avenues for research, such as capitalizing on the mostly shared genetic and environmental conditions between homologous sequences to assess the impact of heterozygous SNVs on chromatin architecture.

1.3 Methods to infer 3D chromatin structure from contact count matrices

Numerous methods have been developed to infer three-dimensional structures from contact count matrices (reviewed by [51]). While many of these methods could plausibly apply to the output of any sequence-based assay of chromatin architecture, nearly all have been developed with Hi-C in mind and subsequently validated with Hi-C data. These methods can be broadly divided into two categories based on whether or not multiple structures are inferred from an individual Hi-C matrix. "Consensus" methods infer a single 3D structure per contact counts matrix [25, 22, 81, 84, 3, 49, 34]. These approaches can be applied to bulk or single-cell Hi-C data. On the other hand, "ensemble" approaches infer entire populations of 3D structures that jointly explain the observed contact counts [64, 83, 16, 38, 42, 82, 30, 55, 87, 90, 41]. Ensemble approaches are designed to be applied specifically to bulk data, and claim to mimic the biological heterogeneity in a population. While this is theoretically a more accurate way to assess the cells assayed by bulk Hi-C experiments, ensemble approaches are inherently underdetermined, as the number of unknown parameters is very large. These methods are also typically much more computationally costly than consensus approaches, which likely limits their widespread adoption by biologists. However, while consensus approaches

are a natural choice for analysis of single-cell Hi-C data, the structure produced may not resemble any individual cell in the population. Consensus approaches also rely on local optimization techniques that may miss the global optimum, whereas most ensemble approaches use alternative techniques, such as Markov chain Monte Carlo sampling [16].

The consensus approaches can be further categorized based on whether they use a probabilistic loss function. Many consensus methods first convert contact counts to a estimated distance matrix using some simple heuristic, then apply multidimensional scaling (MDS) to convert this distance matrix into a 3D structure. Other approaches use statistical models of contact counts to better account for noise.

1.4 Challenges in chromatin structural inference

Despite the abundance of 3D structural inference methods, many challenges remain. In general, accuracy suffers as coverage, and thus the ratio of signal to noise, decreases. While probabilistic methods may be better suited to noisy data, they typically incur higher computational costs than MDS-based approaches. Low coverage can also be managed by decreasing the resolution of the inferred structure, but this is not desirable because some features are only visible at high resolution. Some recently published datasets have sufficient coverage to enable high-resolution analyses [68], but computational requirements increase exponentially as resolution increases. Additionally, there is great interest in inferring structures for single-cell datasets, which are inherently low-coverage.

To our knowledge, only six existing methods are capable of inferring diploid chromatin structures from bulk Hi-C data. Two of these methods are "ensemble" methods that infer a population of structures from a given Hi-C experiment [83, 64]. The other four are "consensus" methods, meaning that they infer a single structure per counts matrix [88, 9, 18, 21]. The method introduced by Belyaeva *et al*. infers chromatin structures via a MDS-based method [9]. It only accepts unphased counts as input, and any available phasing information must be obscured. In order to resolve diploid structures, this method requires data from experiments that measure multiway chromatin interactions between three or more simultaneously interacting loci in addition to the pairwise counts provided by Hi-C. Another method, ASHIC solves the diploidy problem via an expecta-

tion–maximization algorithm that iterates between predicting the homolog assignments of unphased Hi-C data and inferring a diploid 3D structure via a probabilistic loss function [88]. However, it can only infer one chromosome at a time, and extending the method to solve whole-genome structures would not be trivial. Additionally, ASHIC requires a portion of the input counts to be phased *a priori*. Another diploid inference method, SNLC [21], also requires phased data as input. Additionally, a large portion of the data is discarded. If less than a certain percentage of the total reads associated with given locus are phased on both ends, that locus is labeled as "unphased" and any phased reads mapped to this locus are discarded. Otherwise, the locus is labeled as "phased" and all unphased reads mapping to this are discarded. Reads that are phased on one end but not another are also discarded. SNLC first infers the spatial coordinates of "phased" loci via semidefinite programming, then fills in the positions of "unphased" loci via numerical algebraic geometry, local optimization, and clustering. The fourth approach, PASTIS, relies on a Poisson model of contact counts and disambiguates homologous chromosomes with the help of additional constraints [18]. PASTIS is the only diploid structural inference method that is simultaneously capable of inferring structures in the absence of phased data and is also able to accept any available contacts that are phased on one or both ends.

1.5 Organization of dissertation

The remaining chapters in this thesis discuss two methods for inferring consensus diploid chromatin structures from bulk Hi-C data, the latter of which builds upon the former. Chapter 3 describes PASTIS, the first published method to achieve this task [18]. It is itself an extension of a previous probabilistic method for inferring the chromatin structures of haploid genomes [84]. The modified approach outlined in the subsequent chapter substantially improves the quality of inference of diploid genomes. It utilizes a technique called multiscale optimization, an optimization strategy that solves a large optimization problem by building upon the solutions to smaller versions of the problem [58]. Lastly, concluding thoughts are provided in Chapter 4.

Chapter 2

INFERRING DIPLOID 3D CHROMATIN STRUCTURES FROM HI-C DATA

This chapter is adapted from the following work:

A. G. Cauer, G. Yardimci, J.-P. Vert, N. Varoquaux, and W. S. Noble. Inferring diploid 3D chromatin structures from Hi-C data. In Katharina T. Huber and Dan Gusfield, editors, *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:13, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik

2.1 Introduction

The 3D organization of the genome plays an important role in regulating basic cellular functions, including gene regulation [68, 80], differentiation [45, 24], and the cell cycle [61]. Chromosome conformation capture techniques such as Hi-C measure the frequency of interactions between pairs of loci, thereby allowing a systematic analysis of genome structure. Although Hi-C contact matrices yield valuable insights, modeling and visualizing genome structures in 3D can unveil relationships and higher-order structural patterns that are not apparent in the raw data [25, 83, 61, 51] by providing a humanly interpretable 3D structure, orienting genomic regions relative to various nuclear landmarks, and serving as a framework for integrating other data types [12]. Embedding contact count data in a 3D Euclidean space can also reduce noise in the underlying Hi-C data.

Previous methods to inferring chromatin structure from population Hi-C data fall into one of two broad categories. "Ensemble" approaches create populations of 3D structures that jointly explain the observed Hi-C data [64, 83, 16, 38, 42, 82, 30, 55, 87, 90, 41]. Theoretically, structural ensembles can mimic the heterogeneity of cells in a population. However, these methods are fre-

quently underdetermined because there are often more parameters to estimate for a large population of cells than data points. Ensemble models can also be difficult to validate and interpret. "Consensus" approaches, on the other hand, make the assumption that bulk Hi-C data can be accurately summarized in a single, consensus 3D structure [25, 22, 81, 84, 3, 49, 34]. Modeling a single structure tends to be less computationally demanding than modeling an entire population of structures. Furthermore, the resulting model has the advantage of relatively straightforward visualization and interpretation.

For either ensemble or consensus approaches, a particular challenge is presented by Hi-C data derived from diploid organisms. As in most high-throughput sequencing experiments, a typical Hi-C experiment does not produced phased data; that is, the data does not distinguish between allelic copies. Thus, an observation of a single Hi-C contact between loci i and j corresponds to one of four possible events: either copy of locus i coming into contact with either copy of locus j. Any 3D inference method that aims to model diploid genomes must accurately account for this allelic uncertainty.

A variety of strategies have been developed to account for diploidy in Hi-C 3D models. In general, ensemble models face less of a challenge on this front, since the two allelic copies can be treated like additional members of the ensemble. Among consensus methods, by far the most common approach is to assume that the two homologous copies of a given chromosome share the same 3D structure [84, 49, 92] and then to model each chromosome separately.

We are aware of only three previous attempts to model diploidy in non-ensemble methods. Previously, we described an extension of our PASTIS software to handle the near-haploid cell line KBM7 [4]. We proposed to infer jointly the distribution of contact counts between homologs and the 3D structures by maximizing a constrained and relaxed likelihood. However, this relaxation is unsatisfying, as it yields non-integer counts modeled as random Poisson variables. More recently, two separate research groups have developed methods for modeling diploid genomes from single-cell data [15, 79]. However, these methods cannot be directly applied to bulk Hi-C data, which is much more widely available.

In this work, we propose a method to infer diploid consensus 3D models from Hi-C data. Our

approach builds upon PASTIS [84], which infers 3D models by using a Poisson model of Hi-C counts coupled with a simple biophysical model of polymer packing. The key idea of extending PASTIS to infer diploid genomes is to explicitly model the uncertainty of allelic assignments for each observed read. We consider two distinct settings: the more challenging setting where the data is fully ambiguous, and the setting where a subset of the reads can be mapped to a single parental allele. To assist in inference, we incorporate several constraints into our objective function, reflecting our prior knowledge of genome architecture. Through extensive simulations, we demonstrate that our approach can successfully model two distinct homologous chromosome structures, given a sufficient number of reads, even when the data is fully ambiguous. We also apply our approach to real Hi-C data derived from a first generation (F1) cross of two divergent mouse strains (F121 and *Castaneus*). The resulting diploid model of the X chromosome exhibits the expected "superdomain" structure [23], and is quite distinct from the inferred structure of the inactive X.

2.2 Method

Hi-C experiments involve sequencing pairs of interacting DNA fragments. Specifically, cells are cross-linked, DNA is digested using a restriction enzyme, and interacting fragments are then ligated together. Fragments are subsequently sequenced through paired-end sequencing, and each mate is associated with one interacting locus. Hi-C data can then be summarized in a symmetric $n \times n$ contact count matrix C, where each row and column corresponds to a genomic locus and each matrix entry c_{ij} to the number of time those two loci have been observed to interact.

For diploid organisms, reads from homologous chromosomes cannot be distinguished from one another, and the resulting Hi-C matrix aggregates contact counts from homologous chromosomes into a single Hi-C matrix (Figure 2.1). The challenge of inferring diploid structures from Hi-C data lies in disambiguating the contact counts from the two homolog chromosomes. We call these aggregated counts "ambiguous," and denote by C^A the corresponding contact count matrix. If the parental genomes are known *a priori*, then a small proportion of reads can be mapped to each haplotype: contact counts from the two homolog chromosomes can be disambiguated based on heterozygous positions, yielding a single-allele Hi-C count matrix [68, 23]. We refer to these counts

as "unambiguous" and denote the corresponding matrix by C^U. On the other hand, if only one mate can be mapped uniquely to one of the homologous chromosome, then the contact count is only partially disambiguated between the two homologs. We refer to these as "partially ambiguous" contact counts, and we denote the corresponding matrix by C^P.

We model chromosomes as m evenly-spaced beads, and we denote by $\mathbf{X} = (x_1, \cdots, x_m) \in \mathbb{R}^{3\times m}$ the coordinate matrix of the structure. The variable m denotes the total number of beads in the genome, and $x_\ell \in \mathbb{R}^3$ corresponds to the 3D coordinate of the ℓth bead. In the case of a haploid structure, the number of beads corresponds to the number of rows and columns in the contact count matrix C: $n = m$.

2.2.1 Inferring haploid structures with a Poisson model

Before we turn to inferring diploid structures, let us first review the approach proposed by PASTIS [84] to infer haploid 3D structures from a bulk Hi-C contact map $\mathbf{C}$. PASTIS models the interaction frequency between genomic loci i and j as a random independent Poisson variable, where the intensity of the Poisson distribution is a decreasing function f of the Euclidean distance between the two beads (d_{ij}). Leveraging relationships found from studying biophysical properties of DNA as a polymer, PASTIS sets this function as follows: $f(d_{ij}) \sim d_{ij}^{\alpha}$, $\alpha < 0$. The α parameter can be set using prior knowledge (e.g., $\alpha = -3$), or inferred jointly with the 3D structure. Inference is thus performed by maximizing the likelihood of the following Poisson model:

$$c_{ij} \sim \text{Poisson}\left(b_i b_j \beta d_{ij}^{\alpha}\right), \tag{2.1}$$

where β scales for the total number of contacts in the matrix ("coverage"), and b_i and b_j are locus-specific biases that are estimated using a standard procedure [40].

Our strategy to infer diploid structures builds upon this approach. Note that inferring a diploid structure from "unambiguous" contact counts C^U is similar to inferring a haploid structure from a classic Hi-C experiment, with the only difference concerning the biases, which are computed using all contact counts available per locus.

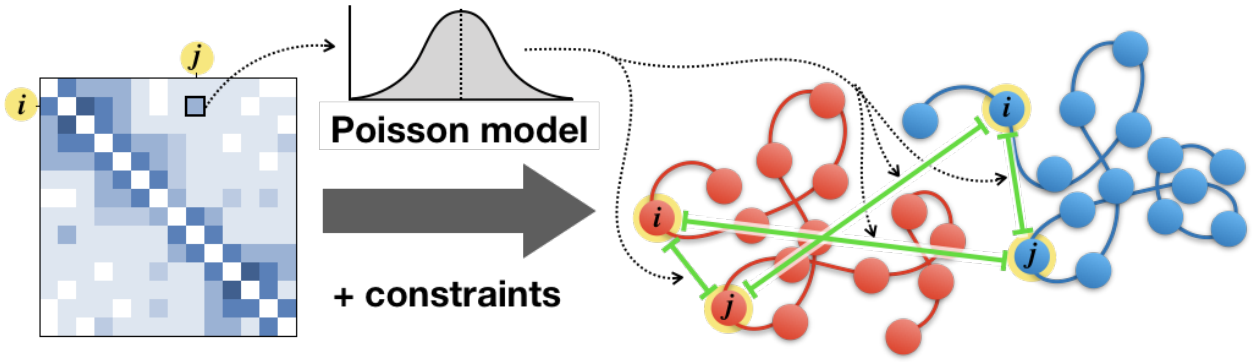

Figure 2.1: **Inferring 3D structure using ambiguous diploid data.** Each observed count (left) corresponds to a sum of four pairs of genomic loci (right). The Poisson model must be adjusted to account for this ambiguity.

2.2.2 Modeling contact counts of diploid structures with a Poisson model

We propose to extend PASTIS to diploid genomes by leveraging the properties of each type of Hi-C contact map: ambiguous, partially ambiguous, and unambiguous. Let us first take a closer look at the common scenario, where the data is fully ambiguous.

For a given ambiguous contact count matrix C^A, each observed contact count c^A_{ij} between a given pair of loci (i, j) corresponds to the sum of four different unambiguous contact counts (Figure 2.1):

$$c^A_{ij} = \sum_{\ell : \Phi(\ell)=i} \sum_{p : \Phi(p)=j} c^U_{\ell p}, \tag{2.2}$$

where $\Phi : [1, n] \to [1, m]$ is the mapping that associates bead ℓ with locus i. Leveraging the property that the sum of i independent Poisson random variables of intensities λ_i is a Poisson variable of intensity $\sum_i \lambda_i$, we model the interaction count as

$$c^A_{ij} \sim \text{Poisson}\left(b_i b_j \beta^A \sum_{\ell : \Phi(\ell)=i} \sum_{p : \Phi(p)=j} d^{\alpha}_{\ell p} \right), \tag{2.3}$$

where m is the number of loci, n is the number of beads, $d_{\ell,p}$ is the Euclidean distance between beads ℓ and p, and β^A is a scaling factor determined by the coverage of the ambiguous contact count matrix.

Similarly, for a given partially ambiguous contact count matrix C^P, each observed contact count c_{ij}^P between a given pair of loci corresponds to the sum of two unambiguous contact counts, and is modeled by the interaction frequency of two pairs of loci.

$$c_{ij}^P \sim \text{Poisson}\left(b_i b_j \beta^P \sum_{\ell : \Phi(\ell)=i} d_{\ell j}^{\alpha} \right) \tag{2.4}$$

β^P is a scaling factor determined by coverage of the partially ambiguous contact count matrix.

We can thus cast the 3D structure inference as maximizing the log-likelihood

$$\begin{aligned}
\max_{\mathbf{X}} \quad \mathcal{L}(X) &= \mathcal{L}_U(X) + \mathcal{L}_P(X) + \mathcal{L}_N(X) \\
&= \sum_{1 \le i < j \le m} c_{ij}^{\mathrm{U}} \log(b_i b_j \beta^{\mathrm{U}} d_{ij}^{\alpha}) - b_i b_j \beta^{\mathrm{U}} d_{ij}^{\alpha} + \\
&\quad \sum_{1 \le i \le n} \sum_{1 \le j \le n, i \ne j} c_{ij}^{\mathrm{P}} \log\left(b_i b_j \beta^{\mathrm{P}} \sum_{\ell : \Phi(\ell)=i} d_{\ell j}^{\alpha} \right) - b_i b_j \beta^{\mathrm{P}} \sum_{\ell : \Phi(\ell)=i} d_{\ell j}^{\alpha} + \\
&\quad \sum_{1 \le i < j \le n} c_{ij}^{\mathrm{A}} \log\left(b_i b_j \beta^{\mathrm{A}} \sum_{\ell : \Phi(\ell)=ip : \Phi(p)=j} d_{\ell p}^{\alpha} \right) - b_i b_j \beta^{\mathrm{A}} \sum_{\ell : \Phi(\ell)=ip : \Phi(p)=j} d_{\ell p}^{\alpha}
\end{aligned} \tag{2.5}$$

Note that this approach holds for polyploid genomes in addition to diploid genomes.

2.2.3 Incorporating prior knowledge

Because the resulting optimization is challenging, we add two constraints that reflect our prior knowledge about chromatin 3D structure: two neighboring beads should not be too far apart from one another, and homologs of most organisms occupy distinct territories [79, 76, 10, 63].

The first constraint maintains bead-chain connectivity by minimizing the variance in the distance between neighboring beads:

$$h_1(\mathbf{X}) = |\omega| \frac{\sum_{\ell \in \omega} d_{\ell,\ell+1}^2}{\left(\sum_{\ell \in \omega} d_{\ell,\ell+1} \right)^2} - 1 \,, \tag{2.6}$$

where ω includes all beads except for cases where the given bead (ℓ) and the subsequent bead ($\ell + 1$) lie on different molecules. This ensures that only intra-molecular distances are included in the constraint. This type of constraint has been used previously in Simba3D [72].

The second constraint aims to disentangle the structures of the two homologs, and operates on the distance between homolog centers of mass. Specifically, it compares the expected distances between homolog centers of mass for chromosome ψ (r_ψ), as estimated prior to diploid inference, with the actual distances between homolog centers of mass in the current inferred structure (r'_ψ), as defined below:

$$r'_\psi = \left\| \frac{1}{|♂_\psi|} \sum_{j \in ♂_\psi} \mathbf{X}_\ell - \frac{1}{|♀_\psi|} \sum_{p \in ♀_\psi} \mathbf{X}_p \right\|, \tag{2.7}$$

where ψ denotes the chromosome, and $♂_\psi$ and $♀_\psi$ denote the set of beads associated with the two homologs of chromosome ψ. The constraint penalizes structures in which the distance between homolog centers of mass is less than the distance expected for that chromosome. It takes the form

$$h_2(\mathbf{X}) = \sum_\psi \max\left(0, r_\psi - r'_\psi\right)^2. \tag{2.8}$$

We note that such a penalty may be interpreted as a log-prior in a Bayesian setting, where the distance between homolog centers of mass of chromosome c is *a priori* normally distributed with mean r_ψ.

When unambiguous data is available, the values of r_ψ may be estimated via the distances between homolog centers of mass in an extremely coarse-grained structure inferred from unambiguous data alone. Alternatively, when unambiguous data is not available, r_ψ may be estimated as the mean distance between chromosome centers of mass in a coarse-grained structure inferred from ambiguous data, since this distance is expected to be similar to that between homologs.

We penalize the likelihood in Equation 2.5 and solve the following optimization problem via L-BFGS-B, a widely used quasi-Newton method[13]:

$$\max_{\mathbf{X}} \quad \mathcal{L}(X) + \lambda_1 h_1(X) + \lambda_2 h_2(X), \tag{2.9}$$

where λ_1 and λ_2 are penalization parameters, the values of which were chosen via a grid search. A version of PASTIS that implements the diploid inference approach is available at `https://github.com/hiclib/pastis`.

2.2.4 Data

Simulated Hi-C data

To validate our approach, we generated 10 simulated genomes with coverage, number of beads, and ratios of disambiguated contact counts corresponding to those of Hi-C data from the mouse Patski cell line (described in Section 2.2.4) at 500 kb resolution. We also generated additional sets of 10 simulated genomes with the same number of beads, varying the proportion of ambiguous, unambiguous, and partially ambiguous contact counts.

To simulate "true" structures, we applied a random walk algorithm. This algorithm places beads successively along each chromosome, constraining each bead to lie within a given distance of the previous bead, provided the new bead does not overlap with any of the previously placed beads and that the entire homolog fits within a sphere of a predefined radius. We then derive unambiguous counts using the following model:

$$c_{ij} = \text{Poisson}\left(\beta d_{ij}^{\alpha}\right), \tag{2.10}$$

where $\alpha = -3$, corresponding to a previously used theoretical exponent for the contact-to-distance transfer function [84]. To convert unambiguous counts to ambiguous or partially ambiguous counts, we summed contacts from the appropriate pairs of loci. In all experiments, we simulated a 343-bead chromosome with 9.3×10^6 reads, which corresponds to the number of beads and reads in the real data we examined. All simulated Hi-C data used for this project is available at available at `https://noble.gs.washington.edu/proj/diploid-pastis/`.

Real Hi-C data

We applied our method to publicly available in situ DNAse Hi-C of Patski fibroblast mouse kidney cells [23]. This line was derived from F1 female embryos, obtained by mating a BL6 female with a *Spretus* male. The BL6 female had an *Hprt* mutation, so hypoxanthine-aminopterin-thymidine medium was used to select for cells with X chromosome inactivation on the maternal allele. All real Hi-C data used for this project is available at available at `https://noble.gs.washington.edu/proj/diploid-pastis/`.

2.2.5 Structure similarity measures

We use the following quantitative measures of similarity between 3D structures to determine the quality of structures inferred from simulated data and assess the stability of chromatin structures across biological replicates.

Root mean square deviation (RMSD) is a common way of comparing two three dimensional structures described by their coordinates $\mathbf{X}, \mathbf{X}' \in R^{3\times m}$. RMSD is defined as

$$RMSD = \min_{X^*} \sqrt{\frac{\sum_{i=1}^{m}(\mathbf{X}_i - \mathbf{X}_i^*)^2}{m}}\ , \tag{2.11}$$

where $\mathbf{X}^*$ is obtained by translating, rotating, and rescaling $\mathbf{X}'$ ($\mathbf{X}^* = s\mathbf{R}\mathbf{X}' - \mathbf{t}$ where $\mathbf{R} \in R^{3\times3}$ is a rotation matrix, $\mathbf{t} \in R^3$ is a translation vector, and s is a scaling factor). RMSD values are computed independently on each homolog of each chromosome and summed.

Distance error [84] assesses the similarity between two distance matrices. This measure assigns more weight to long distances than RMSD. It is given by

$$distError = \min_{X^*} \sqrt{\frac{\sum_{i\in\gamma}(d_i(\mathbf{X}) - d_i(\mathbf{X}^*))^2}{m}}, \tag{2.12}$$

where γ is a set of distances of interest (e.g., intra-chromosomal distances). The structure $\mathbf{X}^*$ is obtained by rescaling $\mathbf{X}'$ ($\mathbf{X}^* = s\mathbf{X}'$ where s is a scaling factor). To distinguish discrepancies in

intra-chromosomal structure from those affecting the relative orientation of each pair of homologs or the relative orientation of different chromosome pairs, we compute distance error in two ways. Intra-chromosomal distance error is computed separately for each homolog of each chromosome, and γ encompasses distances between all beads of the given homolog. Inter-homolog distance error is computed separately for chromosome pair, and γ encompasses distances connecting all beads of two different homologs of a given chromosome. For both measures, values are summed for all chromosomes.

2.3 Results

2.3.1 Constraints improve ambiguous inference.

First, we assessed the accuracy of our method on simulated datasets (Section 2.2.4) using ambiguous data alone, with and without our proposed constraints. Because of the lack of disambiguated contact counts, we expected this inference task to be difficult. Our results demonstrated that the two sets of constraints—bead-chain connectivity and homolog separation—are necessary for successful inference. In the absence of the constraints, ambiguous inference performed poorly (Figure 2.2). Specifically, inferred homolog structures overlapped one another, and adjacent beads sometimes had large gaps between one another. The homolog separation constraint (Equation 2.8) and the bead-chain connectivity constraint (Equation 2.6) were specifically designed to address these problems. Therefore, we repeated the inference with each constraint individually and the two constraints in combination. In this experiment, we compared results generated with and without each constraint at the optimal λ values ($\lambda_1 = 10^8$ and $\lambda_2 = 10^10$, respectively). The results showed that RMSD and distance error are lowest when both constraints were incorporated (Figure 2.2), and error scores obtained from structures inferred with both constraints were significantly lower than those obtained from structures inferred without constraints (pairwise t-test, Bonferroni corrected p-value <0.05, Supplementary Table A.1). 3D structures produced with the constraints had fewer large gaps between neighboring beads and exhibited distinct territories for the two homologs. With both constraints, optimization on a heterogeneous CPU cluster running at 1.90-2.4 GHz took an

average of two hours to converge (averaged over 50 jobs).

As an additional control for the previous experiment, we sought to confirm that the Poisson model for ambiguous diploid contact counts improved inference above what could be attained by the constraints alone. Accordingly, we compared results generated with simulated ambiguous data to "null" structures, which were inferred with the same initialization and constraints but without the Poisson model. Both measures of intra-homolog similarity showed a clear improvement when the Poisson model was incorporated in inference (Figure 2.2). On the other hand, the inter-homolog distance error did not improve with the addition of the Poisson model, suggesting that the constraints are the primary influence in orienting the homologs relative to one another.

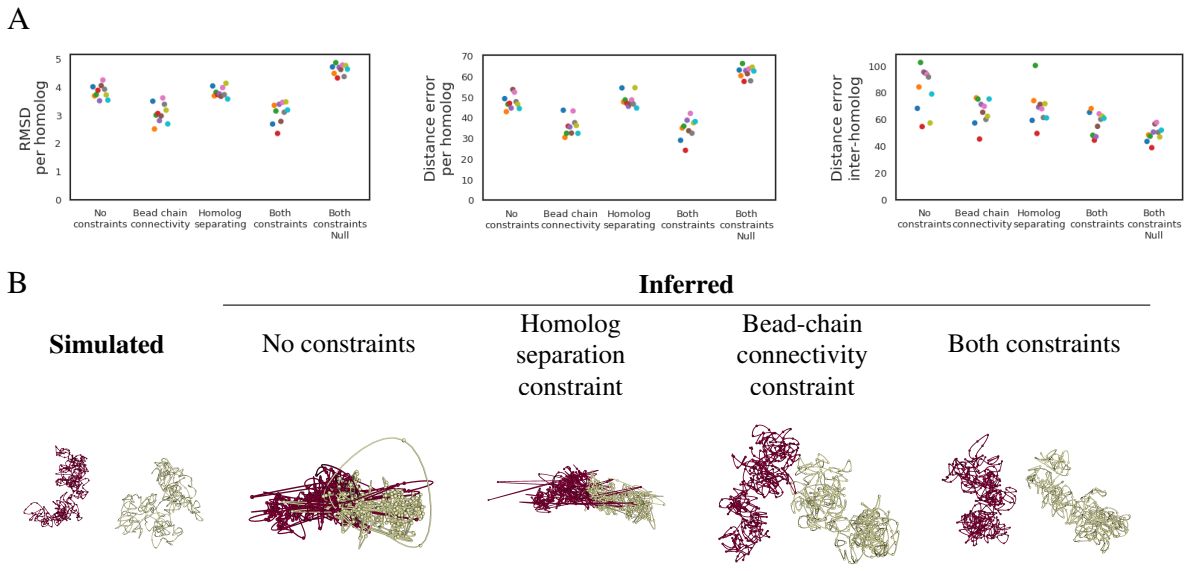

Figure 2.2: **Constraints improve ambiguous inference.** The simulated data consists of a single diploid chromosome with 9.3×10^6 reads and 343 beads, the size of which corresponds to mouse chromosome X at 500 kb resolution. (A) The quality of the inferred structure, as measured by three different error scores (y-axis), improves upon application of the bead-chain connectivity constraint ($\lambda_1 = 10^8$) and the homolog separation constraint ($\lambda_2 = 10^10$). Best results are seen when both constraints are applied simultaneously. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. "Null" indicates inference performed without the Poisson model. Corresponding p-values are in Supplementary Table A.1. (B) A simulated chromosome is shown alongside inferred versions of the same chromosomes using various strategies. Each panel also lists the RMSD and distance error associated with the given structure, relative to the true structure.

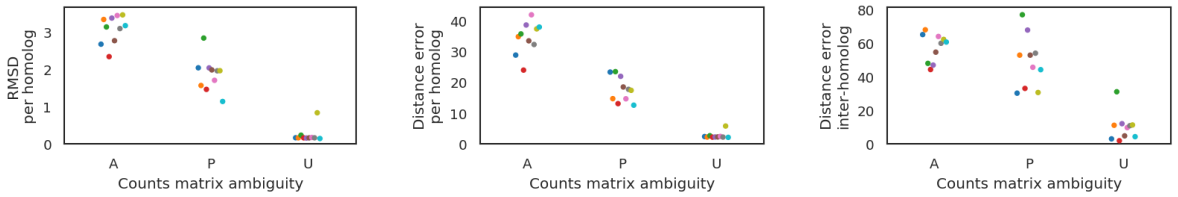

Figure 2.3: **Inference with ambiguous and disambiguated data.** The simulated data consists of a single chromosome with 9.3^6 reads and 343 beads, the size of which corresponds to mouse chromosome X at 500 kb resolution. The quality of the inferred structure, as measured by three different error scores (y-axis), improves when one or both ends of a contact are disambiguated, and best results are seen in the latter case. "A" indicates ambiguous data, "U" indicates unambiguous data, and "P" indicates partially ambiguous data. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Corresponding *p*-values are in Supplementary Table A.2.

2.3.2 Best results obtained by incorporation of disambiguated data.

We expected that more accurate structure inference could be achieved using data where one or more ends of each contact count was disambiguated, relative to fully ambiguous data. We also expected that unambiguous data, in which both ends of each contact count are disambiguated, would yield better models than partially ambiguous data, in which only one end of each contact count is disambiguated. To test these hypotheses, we simulated partially ambiguous data and unambiguous data. Across all similarity measures, inference with unambiguous data performed best, and inference with ambiguous data performed worst, as expected (Figure 2.3). Partially ambiguous contacts seem especially beneficial in inference of intra-homolog structure, since intra-homolog RMSD and distance error of structures inferred with partially ambiguous counts was significantly lower than intra-homolog RMSD and distance error of structures inferred with ambiguous counts (pairwise t-test, Bonferroni corrected p-value <0.05, Supplementary Table A.2).

2.3.3 Inference successfully identifies the superdomain structure of the inactive X chromosome.

Deng et al. [23] previously showed that inactive X chromosome adopts a bipartite structure with two large superdomains, whereas the active homolog does not. We sought to validate our approach by inferring the mouse X chromosome structure and examining the degree to which each homolog

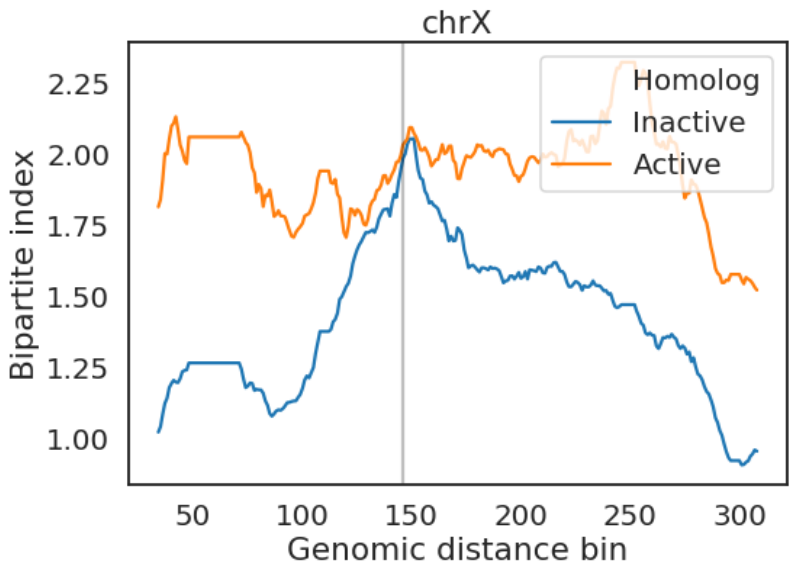

Figure 2.4: **Bipartite structure of the mouse inactive X chromosome.** The bipartite index (y-axis) at each genomic distance bin (x-axis) for the active (orange) and inactive (blue) homologs of the mouse X chromosome. The black line corresponds to the known boundary between superdomains of the inactive homolog at bin 146.

exhibits a bipartite structure. Bipartite structure was assessed via the "bipartite index," which refers to the ratio of the frequency of counts within each superdomain to those between superdomains [23]. To determine the bipartite index of an inferred 3D structure, we induced counts by applying the biophysical model used during inference (Equation 2.10) to the distances between beads.

We inferred a 3D structure for the mouse X chromosome at 500 kb resolution and computed the bipartite index at each bin along the chromosome. The boundary between superdomains of the inactive X chromosome has been shown to center at position 72.8–72.9 Mb (mm9, corresponding to bead 146 in our structure) [23]. In our analysis, the bipartite index of the inferred inactive homolog exhibited a prominent peak around position 75 Mb (corresponding to bead 150), whereas the active homolog only had a relatively small peak at this position (Figure 2.4). This observation suggests that the inference method has successfully recovered this known feature of the mouse inactive X chromosome.

2.4 Discussion

Three-dimensional structural inference of diploid genomes is a challenging problem because most Hi-C data is inherently ambiguous and does not discriminate between contact counts from the two homologs of a given chromosome. Even in the rare cases when parental genotype information is available, only a minority of reads can be disambiguated. As a consequence, many inference methods have modeled a single structure per diploid chromosome [84, 49, 92]. Such an approach assumes that the two homologous copies of a given chromosome have the same 3D structure and prevents structural inference of more than one chromosome at a time. Because of these limitations, the degree of structural similarity between homologous autosomes is not currently well understood.

In this work, we show how to carry out true diploid structural inference by modifying the objective function of PASTIS, a previously published haploid inference method [84]. PASTIS models each contact count via a Poisson distribution of a biophysical model between pairwise distances connecting the corresponding beads. In this work, we model each diploid contact count as the sum of biophysical models between all possible distances between the corresponding bead on each homolog. We combine this modified Poisson model with two constraints that limit the scope of possible solutions to more realistic structures. One constraint enforces even spacing of beads along the chain of the chromosome, and the other serves to spatially separate homologs. Using simulations, we show that the most accurate structures are obtained by inferring with the Poisson model in conjunction with both constraints. We note that the homologs of our simulated structures occupy distinct territories. While this is the case for many organisms, there are some exceptions [56, 79, 76, 10, 63]; therefore, the weight assigned to the homolog separation constraint should be tuned for each organism based on prior knowledge. These analyses were performed at the relatively coarse resolution of 500 kb, and the relationship between resolution, coverage, computational cost, and accuracy of this method remains unexplored.

A limitation to this method involves the distribution of contact count data, which may be better fit by a negative binomial model than a Poisson model [17]. Unfortunately, our method of diploid inference relies on a specific property of Poisson models, namely, that the sum of multiple Poisson

variables is also a Poisson variable. Another caveat involves the biophysical model used during inference (Equation 2.10), which may not accurately capture the relationship between contact counts and pairwise distances in all situations. For example, this relationship may vary depending on the organism, resolution, genomic distance range, and cell cycle status [93, 2, 3, 48, 28]. We also note that in the completely ambiguous case, it is possible that the inferred homologs represent different subpopulations within the sample rather than separating the two haplotypes.

We envision several ways in which diploid PASTIS could be further improved. First, diploid PASTIS could allowing for joint estimation of the α parameter of the biophysical model alongside the 3D structure, as is possible for haploid PASTIS. Second, results could potentially be improved by incorporating a multiscale optimization strategy, in which a high-resolution structures is inferred in a stepwise fashion through multiple rounds of inference with gradually increasing resolution. Similarly, inference of the whole genome may be improved by a stepwise approach where each chromosome is first inferred individually before being placed in the context of the whole genome.

Chapter 3

A MULTI-RESOLUTION OPTIMIZATION STRATEGY FOR INFERRING 3D GENOME ARCHITECTURE FROM HI-C DATA

3.1 Introduction

The three-dimensional organization of the genome plays an important part in regulating numerous basic cellular functions, including gene regulation [68, 80], differentiation [45, 24], the cell cycle [60], DNA replication, and DNA repair [51]. Numerous lines of evidence also link changes in diverse components of 3D genome architecture to many different diseases, including cancer, laminopathies, cohesinopathies, and limb malformation diseases [46]. However, there is still much to discover about the large-scale 3D structure of chromatin. Chromatin conformation capture assays, such as Hi-C [50], probe chromatin structure by quantifying interactions between pairs of loci. While direct analyses of contact counts can address many questions of interest, inferred 3D structures can provide additional value over the original counts matrices. For example, 3D models of genome architecture can facilitate interpretation, account for noise, enable integration of other data types [14], orient genomic regions relative to each other or to nuclear landmarks [3], and discriminate between the homologs of a diploid organism [18].

Numerous methods have been developed to infer three-dimensional structures from Hi-C contact count matrices (reviewed by [51]). These methods can be broadly divided into two categories based on whether or not multiple structures are inferred from an individual bulk Hi-C matrix. "Consensus" approaches assume that bulk Hi-C data can be represented by a single 3D structure [25, 22, 81, 84, 3, 49, 34], whereas "ensemble" approaches infer populations of 3D structures that jointly explain the observed contact counts [64, 83, 16, 38, 42, 82, 30, 55, 87, 90, 41]. The consensus approaches can be further categorized based on whether they use a probabilistic loss function. Many consensus methods first convert contact counts to a estimated distance matrix using some simple heuristic, then apply multidimensional scaling (MDS) to convert this distance matrix into a 3D structure. Other approaches use statistical models of contact counts to better account for noise.

Despite the abundance of methods, many challenges remain. In general, accuracy suffers as coverage, and thus the ratio of signal to noise, decreases. While probabilistic methods may be better suited to noisy data, they typically incur higher computational costs than MDS-based approaches. Low coverage can also be managed by decreasing resolution, but this is not desirable because some

features are only visible at high resolution. Some recently published datasets have sufficient coverage to enable high-resolution analyses [68], but computational requirements increase significantly as resolution increases. Additionally, there is great interest in inferring structures for single-cell datasets, which are inherently low-coverage.

To facilitate inference in difficult settings, previous methods have employed multiscale optimization, an optimization strategy that solves a large optimization problem by building upon the solutions to smaller versions of the problem [58]. Multiscale-based approaches have a number of theoretical benefits. For example, multiscale optimization may help navigate complex and non-convex loss functions, and lead to final solutions with smaller energies than the local optima found by non-multiscale methods. This is because the smaller problems that multiscale optimization solves first may have relatively low-dimensional search spaces, and the landscape of potential solutions to the full problem is constrained by the solutions obtained by these smaller problems. Additionally, multiscale approaches may reduce computational cost, depending on the relative costs of the full problem and its various sub-problems.

We are aware of three previous methods that use multiscale optimization in the context of chromatin structural inference from bulk or single-cell Hi-C data, all of which inferred one consensus structure for a given counts matrix. Two of these methods, Rieber *et al*. and Segal *et al*., first performed structural inference on partitions of Hi-C data and then assembled the resulting structures relative to one another [71, 74]. The third method, Rosenthal *et al*., gradually increased resolution in successive rounds of optimization [72]. All three methods were reported to improve the accuracy of inferred structures and reduce computational cost.

Despite these benefits, each method also had significant drawbacks. For one, the methods Rieber *et al*. and Segal *et al*. do not use probabilistic models. The code from Segal *et al*. is also not publicly available. Furthermore, the Rieber *et al*. and Rosenthal *et al*. methods have substantial limitations related to the function they use to convert between counts and distances. All three methods employ a commonly used counts-to-distance transfer function derived from a biophysical model of polymer packing [50]. Given an appropriate exponent, this function defines a realistic relationship between counts and Euclidean distances. Despite its advantages, the exponent used in this transfer function

depend on cell cycle [48] and organism [28]. Accordingly, in order to avoid systematic biases in the inferred structure, the exponent must be tuned to each dataset. However, Rieber *et al*. and Rosenthal *et al*. do not offer a way to tune the exponent. Furthermore, while the transfer function is also highly dependent on resolution [93, 92, 84], Rieber *et al*. and Rosenthal *et al*. apply the same transfer function to different resolutions of the counts matrix. This limits the accuracy of the structures inferred by these methods.

Lastly, like most structural inference methods, the three multiscale-based approaches are not capable of interring chromatin structures from diploid Hi-C data that is unphased, as nearly all diploid data is. Unphased diploid data does not discriminate between homologs of a diploid chromosome. Although phased heterozygous SNVs, if available, can be used to phase a portion of the reads, the resulting matrix of counts that are phased on both ends is typically extremely sparse. Thus, structural inference of diploid genomes is an inherently difficult task. When used on unphased diploid data, haploid inference methods can only be applied one chromosome at a time, and they make the unrealistic assumption that both copies of the target chromosome adopt identical 3D conformations.

To our knowledge, only six existing methods are capable of inferring diploid chromatin structures from bulk Hi-C data. Two of these methods are "ensemble" methods that infer a population of structures from a given Hi-C experiment [83, 64]. The other four are "consensus" methods, meaning that they infer a single structure per counts matrix [88, 9, 18, 21]. The method introduced by Belyaeva *et al*. infers chromatin structures via an MDS-based method [9]. It only accepts unphased counts as input, and any available phasing information must be obscured. In order to resolve diploid structures, this method requires data from experiments that measure multiway chromatin interactions between three or more simultaneously interacting loci in addition to the pairwise counts provided by Hi-C. Another method, ASHIC solves the diploidy problem via an expectation–maximization algorithm that iterates between predicting the homolog assignments of unphased Hi-C data and inferring a diploid 3D structure via a probabilistic loss function [88]. ASHIC can only infer one chromosome at a time, and extending the method to solve whole-genome structures would not be trivial. Additionally, ASHIC requires a portion of the input counts to be phased *a priori*. Another diploid inference method, SNLC [21], also requires phased data as input. Addi-

tionally, a large portion of the data is discarded. If less than a certain percentage of the total reads associated with a given locus are phased on both ends, then that locus is labeled as "unphased" and any phased reads mapped to this locus are discarded. Otherwise, the locus is labeled as "phased" and all unphased reads mapping to this locus are discarded. Reads that are phased on one end but not another are also discarded. SNLC first infers the spatial coordinates of "phased" loci via semidefinite programming, then fills in the positions of "unphased" loci via numerical algebraic geometry, local optimization, and clustering. The fourth approach, PASTIS, relies on a Poisson model of contact counts and disambiguates homologous chromosomes with the help of additional constraints [18]. PASTIS is the only diploid structural inference method that is simultaneously capable of inferring structures in the absence of phased data and is also able to accept any available contacts that are phased on one or both ends. However, none of these diploid methods utilize multiscale optimization.

Accordingly, we have developed a diploid multiscale optimization approach for chromatin structural inference that rectifies the limitations of previous multiscale methods. Our method, based on PASTIS, uses probabilistic models of contact counts and tunes the counts-to-distance relationship based on the given dataset. Importantly, our updated version of PASTIS also resolves the strict dependence of the counts-to-distance relationship on resolution by approximating the relationship between high-resolution contact counts and low-resolution distances via a novel statistical model. Because many organisms of interest are diploid, we designed our multiscale optimization approach to be capable of inferring the structures of both haploid and diploid genomes. Our approach starts by inferring low resolution structures, then slowly increases resolution in subsequent rounds of inference. This multi-resolution version of PASTIS is compatible with multiscale methods that first infer high-resolution substructures before assembling them into a whole-genome structure, and could be extended to incorporate such a strategy. We use simulations and analysis of real Hi-C data to show that this approach dramatically improves the accuracy of inferred structures.

3.2 Approach

3.2.1 Inferring chromatin structure with PASTIS

Inferring haploid structures

We begin by briefly summarizing the PASTIS method for haploid and diploid chromatin structural inference, previously described in Varoquaux *et al*. and Cauer *et al*. [84, 18]. In PASTIS, chromosomes are represented by evenly-spaced beads on a string. The coordinate matrix of the inferred structure is denoted by $\mathbf{Z} = (x_1, \cdots, x_m) \in \mathbb{R}^{3\times m}$, where m is the total number of beads in the genome and $z_\ell \in \mathbb{R}^3$ corresponds to the 3D coordinate of the ℓ-th bead. PASTIS models the frequency of interaction between two genomic loci, i and j, as a Poisson variable, where the intensity of the Poisson distribution is a decreasing function $f(\cdot)$ of the Euclidean distance d_{ij} between the two beads. Based on a simple biophysical model of polymer packing [50], PASTIS uses the counts-to-distance transfer function

$$f(d_{ij}) \sim d_{ij}^{\alpha},\ \alpha < 0. \tag{3.1}$$

Because the relationship between counts and distances can vary between datasets, the α parameter can optionally be jointly inferred alongside the 3D structure. Thus, optimization maximizes the log-likelihood of the following Poisson model of the counts between two loci (c_{ij}):

$$c_{ij} \sim \text{Poisson}\left(b_i b_j \beta d_{ij}^{\alpha}\right), \tag{3.2}$$

where β scales for the total number of counts in the matrix ("coverage"), and b_i and b_j indicate locus-specific biases [40]. The optimization problem is solved using L-BFGS-B, a commonly used quasi-Newton method [13].

Inferring diploid structures

Inference of diploid organisms presents additional challenges because reads from homologous chromosomes cannot typically be distinguished from one another unless the parental genomes are known

a priori or inferred from Hi-C reads [75]. Parental genotype information allows a small subset of reads to be mapped to a given haplotype based on phased heterozygous positions [68, 23]. We call counts in which it is clear which homologs are being connected "unambiguous" and refer to contacts that are unphased on both ends as "ambiguous." Contacts which are phased on one end and unphased on the other are described as "partially ambiguous."

When inferring chromatin structures for diploid organisms, Cauer *et al*. makes use of all available contacts, regardless of whether one or both ends of a contact are phased, and is even capable of inferring a 3D structure when all reads are ambiguous. Inference with unambiguous counts can be performed using a model analogous to the one used in haploid inference. However, ambiguous contact counts must be disambiguated during inference. Each observed ambiguous contact count c_{ij}^{A} between a given pair of loci (i, j) corresponds to the sum of four different unambiguous contact counts $c_{\ell p}^{U}$:

$$c_{ij}^{A} = \sum_{\ell\,:\,\Phi(\ell)=i} \sum_{p\,:\,\Phi(p)=j} c_{\ell p}^{U}, \tag{3.3}$$

where n is the number of loci, m is the number of beads, and $\Phi : [1, m] \to [1, n]$ is the mapping that associates bead ℓ with locus i. The same relationship also applies to the counts-to-distance transfer function (Equation 3.1):

$$d_{ij}^{\alpha} = \sum_{\ell\,:\,\Phi(\ell)=i} \sum_{p\,:\,\Phi(p)=j} d_{\ell p}^{\alpha}. \tag{3.4}$$

Therefore, each ambiguous interaction count is modeled as

$$c_{ij}^{A} \sim \text{Poisson}\left(b_i b_j \beta^{A} \sum_{\ell\,:\,\Phi(\ell)=i} \sum_{p\,:\,\Phi(p)=j} d_{\ell p}^{\alpha} \right), \tag{3.5}$$

where β^{A} is a scaling factor determined by the coverage of the ambiguous contact count matrix and $d_{\ell,p}$ is the Euclidean distance between beads ℓ and p.

Due to the challenge of inferring distinct homolog structures from primarily unphased data, Cauer *et al*. also incorporated two constraints reflecting prior knowledge about chromatin 3D struc-

ture. The "bead-chain connectivity constraint" discourages large gaps between neighboring beads, thus enforcing the expectation that genomically adjacent regions co-localize in 3D. In addition, the "homolog separation constraint" disentangles the structures of the two homologs by ensuring that intra-homolog distances are smaller than inter-homolog distances [18].

3.2.2 Multi-resolution optimization

Our multiscale optimization approach enables accurate, full-resolution inference by building a series of models of increasing resolution, a strategy we refer to as "multi-resolution optimization." In this approach, we first infer a coarse resolution structure, starting with randomly initialized bead coordinates. The inferred coarse-resolution structure is then linearly interpolated and used to initialize inference of a higher-resolution structure. Subsequent optimizations continue to increase resolution until the desired level of detail is obtained. In our method, each optimization doubles the resolution over that of the previous optimization. This basic strategy is similar to that taken by Rosenthal *et al*. [72].

One difficulty that arises in the context of this multi-resolution strategy is to define the relationship between counts and distances across different resolutions. At any one resolution, the relationship between counts and distances can be approximated by a counts-to-distance transfer function, as is done by PASTIS and many other chromatin structural inference methods [84, 72, 71, 74] (Equation 3.1). However, the relationship between counts and distances is highly dependent on resolution [93, 92, 84]. Although two previous chromatin structural inference methods have attempted to utilize low-resolution structures to improve high-resolution inference [72, 71], neither accounted for the impact of resolution on the counts-to-distance relationship. For multi-resolution optimization to succeed, the counts-to-distance transfer functions used at each resolution should be in agreement. We resolve this challenge by approximating the relationship between high-resolution counts and low-resolution distances as follows.

We begin by defining some important notation. Let $X = \{x_1, \ldots, x_{m_X}\}$ and $Y = \{y_1, \ldots, y_{m_Y}\}$ represent sets of consecutive high-resolution beads, each of which corresponds to a single low-

resolution bead, defined by $\overline{x} = \frac{1}{m_X}\sum_{\ell=1}^{m_X} x_\ell$ and $\overline{y} = \frac{1}{m_Y}\sum_{p=1}^{m_Y} y_p$, respectively. We define the difference between these low-resolution beads as $\Delta_{XY} = \overline{x} - \overline{y}$, and model the difference between sets of high-resolution beads, $x_\ell - y_p$ (for $\ell = 1, \ldots, m_X$, $p = 1, \ldots, m_Y$), as a random variable with mean of Δ_{XY}. The Euclidean distances between high-resolution beads ℓ and p are denoted by $d_{\ell p} = \|x_\ell - y_p\|$, and c_{ij} denotes the high-resolution counts associated with these distances. As in Varoquaux *et al.* and Cauer *et al.*, c_{ij} is modeled as a Poisson random variable, parameterized as described above (Section 3.2.1) [84, 18].

To infer a low-resolution structure from c_{ij}, we must derive a tractable expression for $P(c_{ij} \mid \|\Delta_{XY}\|)$ that is parameterized by distances between low-resolution beads ($\|\Delta_{XY}\|$). To accomplish this, we approximate the law of the random parameter d_{ij}^{α} by a gamma distribution. Then, the law of c_{ij} becomes a negative binomial. In order to formulate the gamma distribution in terms of distances between low-resolution beads, we first estimate the mean and variance of d_{ij}^{α} as a function of these low-resolution distances and the transfer function parameter α. We then use moment matching to assign the parameters of the gamma distribution.

To facilitate approximation of the mean and variance of d_{ij}^{α}, we assume that the difference between high-resolution beads (x_ℓ and y_p) is related to the difference between low resolution beads (Δ_{XY}) by

$$x_\ell - y_p = \Delta_{XY} + \varepsilon\kappa_{\ell p}\,, \tag{3.6}$$

where $\kappa_{\ell p}$ is an independent standard normal vector ($\kappa_{\ell p} \sim \mathcal{N}(0, I_3)$). Thus, ε determines the difference between $x_\ell - y_p$ and Δ_{XY} for each of the three dimensions in 3D space. Consequently, ε scales how similar high-resolution distances are to their corresponding low-resolution distance.

Estimation of the mean and variance of d_{ij}^{α} can be further simplified by disentangling the moments of $d_{\ell p}^{\alpha}$ from the low-resolution distance, $\|\Delta_{XY}\|$. To illustrate, we rewrite $d_{\ell p}^{\alpha}$ as follows:

$$d_{\ell p}^{\alpha} = \|\Delta_{XY} + \varepsilon\kappa_{\ell p}\|^{\alpha} = \|\Delta_{XY}\|^{\alpha} \times \left\| \frac{\Delta_{XY}}{\|\Delta_{XY}\|} + \frac{\varepsilon}{\|\Delta_{XY}\|}\kappa_{\ell p} \right\|^{\alpha}. \tag{3.7}$$

This allows us to obtain the mean and variance of $d_{\ell p}^{\alpha}$ in terms of the mean and variance of Λ, a

system where the low-resolution beads are spatially separated by a unit of 1. We denote the mean and variance of Λ as $m(\alpha, \varepsilon)$ and $v(\alpha, \varepsilon)$, respectively. It follows that

$$\mathbb{E}[d_{\ell p}^{\alpha}] = \|\Delta_{XY}\|^{\alpha} \cdot m(\alpha, \varepsilon) \tag{3.8}$$

$$\mathrm{Var}[d_{\ell p}^{\alpha}] = \|\Delta_{XY}\|^{2\alpha} \cdot v(\alpha, \varepsilon)\,, \tag{3.9}$$

where $\mathbb{E}$ indicates the mean and Var is the variance.

We approximate $m(\alpha, \varepsilon)$ and $v(\alpha, \varepsilon)$ by fitting a polynomial function to a simple simulation prior to optimization. Using the assumption described in Equation 3.6, we simulate Λ across a range of ε and α values ($-8 \leq \log \varepsilon \leq 4$, step 0.1 and $-4 \leq \alpha \leq -1$, step 0.1). Specifically, we model a system of two low-resolution beads that are separated by a unit of 1, each of which will be associated with $10,000$ high-resolution beads. We utilize Equation 3.6 to simulate $10,000^2$ distances between these sets of high-resolution beads for the given value of ε. Because we do not expect high-resolution beads of a biological structure to overlap, simulated high-resolution distances below a given cutoff are set to that cutoff. Enforcing a lower limit on the high-resolution distances is also necessary to prevent $m(\alpha, \varepsilon)$ and $v(\alpha, \varepsilon)$ from being infinite. We arbitrarily set this minimum distance cutoff to 0.5, a value at which no more than 5% of high-resolution distances are thresholded for any given value of ε. The resulting high-resolution distances are then raised to the power of α to yield Λ. We next fit polynomial functions to the mean and variance of the simulated Λ. The polynomial coefficients are subsequently used to approximate $m(\alpha, \varepsilon)$ and $v(\alpha, \varepsilon)$ during optimization, and ε and α are jointly inferred alongside the low-resolution structure.

In the diploid case, we sum the means and variances of $d_{\ell p}^{\alpha}$ across homologs.

$$\mathbb{E}[d_{ij}^{\alpha}] = \sum_{\ell\,:\,\Phi(\ell)=i} \sum_{p\,:\,\Phi(p)=j} \mathbb{E}[d_{\ell p}^{\alpha}] \tag{3.10}$$

$$\mathrm{Var}[d_{ij}^{\alpha}] = \sum_{\ell\,:\,\Phi(\ell)=i} \sum_{p\,:\,\Phi(p)=j} \mathrm{Var}[d_{\ell p}^{\alpha}] \tag{3.11}$$

This strategy mirrors the approach by which Cauer *et al*. achieves diploid inference in the absence

of multiscale optimization [18].

Once the mean and variance of d_{ij}^{α} have been estimated, we approximate the law of d_{ij}^{α} by a Gamma distribution using moment matching (shape $k_{XY} = \frac{\mathbb{E}[d_{ij}^{\alpha}]^2}{\mathrm{Var}[d_{ij}^{\alpha}]}$, scale $\theta_{XY} = \frac{\mathrm{Var}[d_{ij}^{\alpha}]}{\mathbb{E}[d_{ij}^{\alpha}]}$). This yields the following negative binomial likelihood:

$$P(c_{ij} \mid \|\Delta_{XY}\|) = \frac{\Gamma\left(c_{ij} + k_{XY}\right) \theta_{XY}^{c_{ij}}}{\Gamma\left(k_{XY}\right) c_{ij}! \, (\theta_{XY} + 1)^{c_{ij} + k_{XY}}}, \tag{3.12}$$

where Γ denotes the gamma function.

To improve numerical stability during calculation of the log likelihood, we use Stirling's approximation of the log gamma function:

$$\log \Gamma(x) \approx x \log x - x + \frac{1}{2} \log \frac{2\pi}{x} + B_N(x), \tag{3.13}$$

where

$$B_N(x) = \sum_{n=1}^{N} \frac{B_{2n}}{2n(2n-1)x^{2n-1}}, \tag{3.14}$$

and B_n are Bernoulli numbers. This results in the following log likelihood for a given low-resolution

distance and its associated high-resolution counts:

$$\begin{aligned}
\mathcal{L}(\|\Delta_{XY}\|, \alpha, \varepsilon \mid c_{ij}) \;=\; & -\tfrac{1}{m_X m_Y} \textstyle\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} k_{XY} \log(1 + \theta_{XY}\beta b_i b_j) - B_N(k_{XY}+1) \\
& + \tfrac{1}{m_X m_Y} \textstyle\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} B_N(c_{ij} + k_{XY} + 1) \\
& + \tfrac{1}{m_X m_Y} \left(\textstyle\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} c_{ij} \right) \left(\log \theta_{XY}\beta + \log(k_{XY}+1) - 1 \right) \\
& - \tfrac{1}{m_X m_Y} \textstyle\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} c_{ij} \log(1 + \theta_{XY}\beta b_i b_j) \\
& + \tfrac{1}{m_X m_Y} \textstyle\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} (c_{ij} + k_{XY} + 0.5) \log\left(\tfrac{c_{ij}}{k_{XY}+1} + 1 \right) \\
& + \tfrac{1}{m_X m_Y} \textstyle\sum_{i=1}^{m_X} \sum_{j=1}^{m_Y} \log\left(\tfrac{c_{ij}}{k_{XY}} + 1 \right)
\end{aligned} \tag{3.15}$$

This term is then summed for all low-res distances to yield the final log likelihood for a given counts matrix ($\boldsymbol{C}$):

$$\mathcal{L}(\mathbf{Z}, \alpha, \varepsilon \mid \boldsymbol{C}) = \sum_X \sum_Y \mathcal{L}\left(\|\Delta_{XY}\|, \alpha, \varepsilon \mid c_{ij} \right) \tag{3.16}$$

As in Cauer *et al*., the likelihoods for ambiguous ($\boldsymbol{C}^A$), unambiguous ($\boldsymbol{C}^U$), and partially ambiguous ($\boldsymbol{C}^P$) counts matrices are then summed as follows:

$$\mathcal{L}(\mathbf{Z}, \alpha, \varepsilon \mid \boldsymbol{C}^A, \boldsymbol{C}^U, \boldsymbol{C}^P) = \mathcal{L}(\mathbf{Z}, \alpha, \varepsilon \mid \boldsymbol{C}^A) + \mathcal{L}(\mathbf{Z}, \alpha, \varepsilon \mid \boldsymbol{C}^U) + \mathcal{L}(\mathbf{Z}, \alpha, \varepsilon \mid \boldsymbol{C}^P) \tag{3.17}$$

3.2.3 Incorporating prior knowledge during diploid multi-resolution inference

Due to the inherent difficulty of diploid inference, Cauer *et al*. introduced two constraints that limit the range of solutions to realistic results by incorporating prior knowledge [18]. While these constraints successfully improved results during standard diploid inference, both constraints have limitations, some of which may pose a challenge during multi-resolution optimization or when applied

to data from certain organisms. Consequently, we reformulated the constraints to be more widely applicable.

Maintaining continuity of consecutive beads in diploid structures

We expect that the 3D structure of each homolog of each chromosome should not exhibit large gaps between genomically adjacent beads. Cauer *et al*. [18] maintained this continuity between neighboring beads by minimizing the variance in the distances between neighboring beads - a strategy previously employed by Rosenthal *et al*. [72]:

$$h_1(\mathbf{Z}) = |\omega| \frac{\sum_{\ell \in \omega} d_{\ell,\ell+1}^2}{\left(\sum_{\ell \in \omega} d_{\ell,\ell+1}\right)^2} - 1\,, \tag{3.18}$$

where ω includes all beads except for cases where the given bead (ℓ) and the subsequent bead ($\ell + 1$) lie on different molecules. This ensures that only intra-molecular distances are included in the constraint.

Unfortunately, the variance in the distances between neighboring low-resolution beads may differ from the variance in the distances between neighboring high-resolution beads. Therefore, the ideal penalty for this constraint is dependent on resolution, which is impractical in the context of multi-resolution inference. The difficulty arises because the function by which the ideal constraint penalty changes across different resolutions is unknown - it depends on the relative orientation of the high-resolution beads in the particular structure being inferred. Therefore, the penalty would have to be tuned experimentally at every resolution, which is computationally costly, time-consuming, and inconvenient for the user.

To resolve this issue, we maintain bead-chain connectivity via a different approach, in which the mean and variance of the distances between neighboring beads is derived from the counts data. To achieve this, we make two simplifying assumptions. First, we assume that the distances between neighboring beads do not depend on parent of origin, and that consequently the counts corresponding to these distances do not differ between the maternal and paternal homologs. Second, we assume that inter-homolog distances are a negligible contributor to the ambiguous counts associated with

the distances between neighboring beads. Therefore, for the purposes of this constraint, we hide any available phasing information and treat all counts as ambiguous. We refer to the ambiguated contact counts corresponding to distances between neighboring beads as $c^*_{i,i+1}$, where and i and $i+1$ are on the same chromosome. At high resolution $c^*_{i,i+1}$ is a single interaction bin. However, during low-resolution inference, it represents a set of high-resolution bins, all of which correspond to the given low-resolution i and $i+1$.

We then relate these counts to the distances between neighboring beads via the log likelihoods described in Sections 3.2.1–3.2.2. The bead-chain connectivity constraint then takes the form

$$h_1(\mathbf{Z}, \alpha, \varepsilon) = \frac{1}{2|\omega^*|}\left(\sum_{i\in\omega^*} \mathcal{L}(2d^{♂}_{i,i+1} \mid c^*_{i,i+1}) + \sum_{i\in\omega^*} \mathcal{L}(2d^{♀}_{i,i+1} \mid c^*_{i,i+1})\right), \tag{3.19}$$

where ω^* includes all loci in the ambiguated counts matrix except for cases where the given locus (i) and the subsequent locus ($i+1$) lie on different chromosomes. This ensures that only intra-chromosomal counts are included in the constraint. Thus, $d^{♂}_{i,i+1}$ and $d^{♀}_{i,i+1}$ are the distances between neighboring beads on the paternal and maternal homologs of each chromosome, respectively. When inferring high-resolution structures, $\mathcal{L}$ refers to the log likelihood of the Poisson model (Section 3.2.1). At low resolution, $\mathcal{L}$ refers to the log likelihood of the multi-resolution negative binomial model (Section 3.2.2).

In order to maintain continuity of the beads in each molecule, this constraint must be applied to all genomically neighboring beads. However, genomic loci with very few counts across all of their interaction bins, such as those that are poorly mappable or contain high repeat content, are typically masked from inference to avoid artifacts in the 3D structure. Therefore, we applied this constraint to $d_{i,i+1}$ associated with such loci using the mean of $c^*_{i,i+1}$ across all ω^*.

Enforcing separation of homologous chromosomes

For a given chromosome, we anticipate that intra-homolog distances are, on average, shorter than inter-homolog distances. However, this prior belief is not enforced by our statistical model of contact counts or by the bead-chain connectivity constraint. Without an additional constraint to ensure

an appropriate distance between homologous chromosomes, inference on unphased data produces structures in which homologous chromosomes overlap more than would be expected [18]. Therefore, Cauer *et al*. added a second constraint to maintain a minimum distance between the centers of mass of homologous chromosomes [18]. This constraint compares the expected distances between homolog centers of mass for chromosome ψ (r_ψ), as estimated prior to diploid inference, with the distances between homolog centers of mass in the current inferred structure (r'_ψ). The constraint penalizes structures in which the distance between homolog centers of mass is less than the distance expected for that chromosome. The expected distances between homolog centers of mass are estimated prior to inference based on an extremely course-grained draft structure. In the absence of phased data, the draft structure is inferred as if the data were haploid, with only one molecule inferred for each chromosome, and r_ψ is taken as the mean distance between chromosome centers of mass.

Although Cauer *et al*. demonstrated that this constraint performed well with simulated and mouse data, it has noteworthy limitations. First, because the constraint operates on molecule centers of mass, it is only suitable for organisms whose chromosomes form distinct territories, which is not the case for many species [35]. Second, the inference of draft structures required to estimate r_ψ adds to computational requirements, and the inferred values of r_ψ may be inaccurate. Lastly, this constraint does not enforce an upper limit on r'_ψ, which could potentially lead to homologs being too far apart.

To resolve these issues, we formulated a constraint that separates homologs by working with the distribution of all inter-homolog distances, rather than operating on the distances between homolog centers of mass. We expect that the distribution of distances between beads on different homologs of each chromosome is similar to the distribution of distances between beads on different chromosomes, the latter of which is directly related to the distribution of inter-chromosomal counts. We enforce this expectation with a constraint that uses Kullback–Leibler (KL) divergences [47] between distributions. Specifically, we compare the distribution of high-resolution inter-chromosomal ambiguated counts ($c_{ij}^{*\text{inter}-\text{chrom}}$, where i and j are on different chromosomes) with a negative binomial distribution that is parameterized by high-resolution inter-homolog $(d_{\ell p}^{\alpha})^{\text{inter}-\text{hmlg}}$ (where ℓ and p

are beads on different homologs of the same chromosome).

In order to use the similarity between these two distributions to separate homologous chromosomes, we must make a simplifying assumption about the nature of inter-chromosomal distances. In general, the ambiguous counts between loci i and j correspond to d_{ij}^{α}, which is itself a sum of $d_{\ell p}^{\alpha}$ across all four combinations of homologs (see Equation 3.4). Our primary objective function, $\mathcal{L}(\mathbf{Z}, \alpha, \varepsilon)$ (see Sections 3.2.1 and 3.2.2), does not make any assumptions about the relationship between the terms in this sum. However, for the purposes of this constraint, we assume that the distribution of high-resolution distances between a given pair of molecules is the same regardless of whether those molecules share the same parental origin and regardless of whether they are homologs of the same chromosome. We thus anticipate the following relationship between the mean of high-resolution inter-chromosomal ambiguated counts and the mean of high-resolution inter-homolog $(d_{\ell p}^{\alpha})^{\text{inter-hmlg}}$:

$$\mathbb{E}\left[c_{ij}^{\text{inter-chrom}}\right] \approx \mathbb{E}\left[4(d_{\ell p}^{\alpha})^{\text{inter-hmlg}}\right] \tag{3.20}$$

Our goal now is to use the values of $4(d_{\ell p}^{\alpha})^{\text{inter-hmlg}}$ to parameterize a negative binomial distribution with a variance that is approximately equal to $\text{Var}\left[c_{ij}^{\text{inter-chrom}}\right]$.

This negative binomial distribution is derived as follows. We first fit a gamma distribution to the set of high-resolution inter-homolog $4(d_{\ell p}^{\alpha})^{\text{inter-hmlg}}$. When inferring a high-resolution structure, the parameters of this gamma distribution, $k^{\text{inter-hmlg}}$ and $\theta^{\text{inter-hmlg}}$, are obtained via moment matching, using the mean and variance of $4(d_{\ell p}^{\alpha})^{\text{inter-hmlg}}$. When inferring a low-resolution structure, this gamma distribution is derived as a mixture of the gamma distributions associated with every inter-homolog low-resolution distance bin, each of which is parametrized by k_{XY} and θ_{XY} (Section 3.2.2). This gamma mixture model is then multiplied by 4 to yield the desired gamma distribution. We next compound this gamma distribution with a Poisson, which results in a negative binomial distribution. The choice of a Poisson distribution reflects our Poisson model of high-resolution counts (Equation 3.2).

The constraint then minimizes the difference between this negative binomial distribution (defined as Q) and the distribution of high-resolution inter-chromosomal ambiguated counts (defined

as P) by minimizing the KL divergence between the two:

$$h_2(\mathbf{Z}, \alpha, \varepsilon) = -\frac{1}{|P|} D_{\mathrm{KL}}\left(P \parallel Q\right) . \tag{3.21}$$

3.2.4 Full objective

The log likelihood of counts is combined with the constraints to produce the following objective function, which we solve using L-BFGS-B [13]:

$$f(\mathbf{Z}, \alpha, \varepsilon) = \max_{\mathbf{Z}, \alpha, \varepsilon} \quad \left(\frac{1}{\tau}\right) \mathcal{L}(\mathbf{Z}, \alpha, \varepsilon) + \lambda_1 h_1(\mathbf{Z}, \alpha, \varepsilon) + \lambda_2 h_2(\mathbf{Z}, \alpha, \varepsilon) , \tag{3.22}$$

where λ_1 and λ_2 are penalization parameters for each constraint and τ is the total number of counts bins across all given Hi-C matrices. Note that this formulation differs slightly from the objective used by Cauer *et al*., which does not divide by τ [18].

Tuning the counts-to-distance transfer function

Because the relationship between counts and distances depends on a variety of factors, including species [28] and stage of the cell cycle [48], we allow for α to be inferred jointly alongside the structure. Varoquaux *et al*. described an approach for inferring the structure and α together using a coordinate descent algorithm that iterates between maximizing $f(\mathbf{Z}, \alpha, \varepsilon)$ with respect to α and maximizing $f(\mathbf{Z}, \alpha, \varepsilon)$ with respect to $\mathbf{Z}$ [84]. We apply a similar approach to infer α.

However, inference of α in the multi-resolution and diploid settings presents additional challenges. Diploidy complicates the optimization landscape for both single- and multi-resolution inference. Furthermore, due to the approximate nature of the low-resolution negative binomial model and the flexibility introduced by inferring ε, inferring α during multiscale optimization is unlikely to yield satisfactory results. Therefore, we perform α inference at high resolution using a modified coordinate descent approach. To limit computational complexity, we only use intra-chromosomal counts during this process. And to guard against the possibility that the imperfect assumptions made by our homolog separation constraint have an exaggerated impact on the value of α, this constraint

is omitted during the process of updating α given the current structure (but included when updating the structure given α). Once the modified coordinate descent process converges, a whole-genome structure is inferred using the final value of α. The whole-genome structure may be inferred with or without multiscale optimization.

3.3 Methods

3.3.1 Phased diploid yeast Hi-C data

We applied our method to publicly available Hi-C of diploid yeast generated by Kim *et al*. [43]. The datasets we used were generated from the diploid ILY456 line created by Kim *et al*. This line is a hybrid between the *Saccharomyces cerevisiae* and *Saccharomyces uvarum* strains of budding yeast. Although these strains have maintained nearly complete synteny [27], they are also diverged enough that nearly all of the Hi-C contacts can be phased on both ends [43]. We used a total of five datasets, which were generated using four experimental conditions: exponentially growing cultures, saturated cultures, galactose-induced cultures, and cultures arrested by the anti-mitotic agent nocodazole. Kim *et al*. generated two experimental replicates for the exponentially growing cultures, and all other conditions had one replicate. The five datasets used here are publicly available at the NCBI Sequence Read Archive (accession numbers: SRR4433970, SRR4433972, SRR4433973, SRR4433974, SRR4433975).

Processing of Hi-C reads was performed similarly to Kim *et al*. [43]. Briefly, reads were mapped to a combined reference containing genomes for both species, and subsequently binned into fixed-width fragments [43]. As in Kim *et al*., reads that did not uniquely map to either *S. cerevisiae* or *S. uvarum* were discarded. Although Kim *et al*. used a bin size of 32 kb, we used bin sizes of 8 kb and 16 kb to better assess the multi-resolution approach. In all other respects, our initial processing and binning of the reads was as described in Kim *et al*.

After binning the reads, we performed an additional processing step, not present in Kim *et al*., to establish a direct correspondence between loci on the maternal and paternal homologs, since such a correspondence is necessary for PASTIS to infer diploid genomes. Although the genomes

of *S. cerevisiae* and *S. uvarum* are highly similar, the genomic coordinates of a given sequence may differ between the two. Therefore, we remapped the binned reads onto a shared genomic coordinate system.

This shared genomic coordinate system was determined on a chromosome-by-chromosome basis. For each chromosome, the mapping between all loci on *S. cerevisiae* and all loci on *S. uvarum* was determined via the "bin offset" of the centromeric loci, where the bin offset between locus i on *S. cerevisiae* and locus j on *S. uvarum* is defined as $i - j$. Using the same bin offset for all loci of a given chromosome ensured that the spacing between loci on the *S. cerevisiae* homolog of the shared genomic coordinate system matched the spacing between loci on the *S. uvarum* homolog. For a given chromosome, the shared genomic coordinate system was based on the sequence of the species in which that chromosome was smallest.

In order to avoid artifacts in the remapped counts matrices, we next excluded regions suspected to be non-syntenic between *S. cerevisiae* and *S. uvarum*. To facilitate this process, we first established homology between the binned genomic coordinates of each species by counting the number of homologous gene annotations in each locus—-an approach described by Kim *et al*. for the purposes of assessing proximity between homologs loci. As in Kim *et al*., we excluded isolated homologous interaction bins, which may arise from repetitive sequences. We subsequently restricted the homology assignments to be one-to-one by selecting the locus pairings with the greatest number of shared homologous genes. Chromosome labels were swapped in cases where the majority of loci on a given *S. cerevisiae* chromosome associated with a different chromosome on *S. uvarum*. All loci with inter-chromosomal homology assignments were then masked from the shared genomic coordinate system. Loci at the telomeres for which homology was unknown were also masked.

Once the Hi-C counts were assigned to the genomic coordinate system, the data were filtered to remove poorly mappable regions as well as loci where the mapping between species was suspected to be incorrect. Both filtering steps were performed using the summed counts from all five datasets, and each individual dataset was subsequently updated to mask the loci that were filtered out. We first removed loci whose total genome-wide interactions were below the 3rd percentile. Next, because we do not expect large differences between the homologs with regard to the counts

between genomically neighboring loci, we removed loci for which the ratio of such counts was above the 94th percentile. Lastly, any masked regions at the beginning or end of each chromosome were removed from the dataset, so that those beads would not be inferred.

3.3.2 Simulated Hi-C data

Simulation approach

Simulations were performed as in Cauer *et al*. [18]. We simulated "true" consensus chromatin structures via a random walk algorithm. This algorithm ensures that each bead is situated near its genomic neighbors. It also prevents overlap of beads and confines the entire structure of every molecule to a sphere of predefined radius.

Contact counts were derived from these true structures via the following model, which mirrors the PASTIS counts-to-distance transfer function:

$$c_{ij} = \text{Poisson}\left(\beta d_{ij}^{\alpha}\right), \tag{3.23}$$

where β determines coverage and $\alpha = -3$, corresponding to a previously used theoretical exponent for the contact-to-distance transfer function [84, 18]. To create ambiguous counts, we summed unambiguous contacts from the appropriate pairs of loci. While this simple approach to simulating contact counts would not be suitable for comparing methods that use this counts-to-distance transfer function to methods that related distances and counts via a different approach, all of the inference methods we evaluated with these simulations rely on the same PASTIS transfer function, rendering the comparison fair.

Simulated genome structure and counts

To validate our approach, we generated a set of 10 simulated diploid genome structures, with number of chromosomes and beads per chromosome corresponding to the 16 kb diploid yeast hybrid between *S. cerevisiae* and *S. uvarum* from Kim *et al*. (described in Section 3.3.1) [43]. In order

to evaluate multi-resolution optimization with and without the complicating factor of unphased diploid data, we simulated both ambiguous and unambiguous counts matrices. We also varied coverage of the simulated reads, with simulated read depths chosen to emulate low-, medium-, and high-coverage experimental data (NUMBER-OF-READS , NUMBER-OF-READS , and NUMBER-OF-READS reads, respectively). The medium-coverage datasets contain approximately as many reads as obtained for *S. cerevisiae* x *S. uvarum* hybrid Hi-C datasets in Kim *et al*. [43]. In order to ensure a realistic relationship between different molecules, we verified that the percentage of inter-molecular counts was consistent with the *S. cerevisiae* x *S. uvarum* Hi-C data from Kim *et al*. The genomic bins that were masked in Section 3.3.1 were also masked in the simulated counts matrices.

3.3.3 Structure similarity measures

We used a variety of structural similarity scores to evaluate inference results against expected 3D structures and examine the stability of inferred structures. Each similarity measure is applied to a pair of three-dimensional structures described by their coordinates $\mathbf{S}, \mathbf{S}' \in \mathbb{R}^{3\times m_S}$, where $\mathbf{S}$ indicates the target structure and $\mathbf{S}'$ indicates the predicted structure.

Root mean square deviation

Root mean square deviation (RMSD), a common measure of similarity between three-dimensional structures, is defined as follows:

$$\mathrm{RMSD} = \sqrt{\frac{1}{\sum_{\gamma_1}|\gamma_1|}\sum_{\gamma_1}\min_{\mathbf{S}^*}\left[\sum_{i\in\gamma_1}\left(s_i - s_i^*\right)^2\right]}\,, \tag{3.24}$$

where γ_1 is a set containing beads of interest (e.g., beads in a given molecule). The structure $\mathbf{S}^*$ is obtained by optimally translating, rotating, and reflecting $\mathbf{S}'$ via Procrustes transformation ($\mathbf{S}^* = \mathbf{RS}' - \mathbf{t}$ where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is an improper rotation matrix, $\mathbf{t} \in \mathbb{R}^3$ is a translation vector).

We use RMSD to assess the similarity of each individual molecule in the genome. Here, γ_1 denotes the set of beads in a given molecule. This allows for each molecule to be aligned separately.

An RMSD score of 0 would indicate that all molecules in $\mathbf{S}$ and $\mathbf{S}'$ are identical, regardless of the relative orientation of these molecules.

Distance error

We also compared structures by computing the root mean squared error between their distance matrices, a measure we refer to as "distance error" [84, 18]. Distance error assigns more weight to long distances than RMSD. Distance error also enables us to compare the relative orientation of two of more molecules because the distances connecting between molecules can be assessed separately from the distances within each molecule. It is defined as

$$\text{distError} = \sqrt{\frac{1}{|\gamma_2|} \sum_{(i,j)\in\gamma_2} \left(\|s_i - s_j\| - \|s'_i - s'_j\| \right)^2}, \tag{3.25}$$

where γ_2 is a set containing pairs of beads that correspond to distances of interest (e.g., intra-molecular distances).

We use distance error for three purposes. First, we evaluate the similarity of intra-molecular distances. Second, we compare all inter-molecular distances together. Lastly, we separately examine the relative orientation of each pair of homologous chromosomes via inter-homolog intra-chromosomal distances. A distance error score of 0 would indicate that the set of distances being compared are identical.

Global Distance Test

The Global Distance Test (GDT) [89] is commonly used to evaluate local similarity of protein structures, but it can be applied to any pair of structures for which there is a direct correspondence between the beads. It is useful to examine local structural similarity because even reasonably similar structures may differ in the relative orientation of compact domains. In such cases, the regions of similarity may be overlooked by other structural similarity measures. GDT aligns small fragments of $\mathbf{S}$ and $\mathbf{S}'$, then iteratively extends these fragments to determine the maximum number of beads on $\mathbf{S}$ that fall within a specified distance cutoff of their corresponding bead on $\mathbf{S}'$. The final GDT

score for a given distance cutoff is taken as the maximum number of beads within this cutoff at convergence.

Because the use of a single distance cutoff would not fully describe the similarity between two structures, GDT is typically reported as a sum across a variety of distance cutoffs—the GDT "total score" (GDT-TS). When used for the analysis of protein structures, these cutoffs are 1, 2, 4, and 8 Å. When applying GDT-TS to chromatin structures, we first rescale both structures such that the mean distance between neighboring beads in each molecule of $\mathbf{S}$ is equal to the mean distance between consecutive $C\alpha$ atoms of protein structures (3.8 Å) [19]. This allows us to assess the GDT-TS of chromatin structures using the same distance cutoffs used for proteins. GDT-TS outputs values in $(0, 1]$, with higher values indicating greater local similarity.

We use GDT-TS to assess local structural similarities for each individual molecule. We then take the mean of these results, weighted by the number of beads in each molecule.

Template modeling score

In order to specifically assess global structural similarities, we also compared structures using the template modeling score (TM-score) [91]. Compared to RMSD and distance error, in which a few large local errors can have a substantial impact on the final score, the TM-score gives a more global estimate of similarity because it weighs small errors more strongly than large ones. TM-score also differs from GDT-TS, which is a measure of local structural similarity. It is defined as follows:

$$\text{TM-score} = \frac{1}{\sum_{\gamma_1} |\gamma_1|} \sum_{\gamma_1} \max_{\mathbf{S}^*} \left[\sum_{i \in \gamma_1} \frac{1}{1 + \upsilon_{\gamma_1}^2 \left(\frac{s_i - s_i^*}{d_0(|\gamma_1|)} \right)^2} \right], \tag{3.26}$$

where γ_1 is a set containing beads of interest (e.g., beads in a given molecule) and $d_0(|\gamma_1|) = 1.24\sqrt[3]{|\gamma_1| - 15} - 1.8$ normalizes for the number of beads assessed. The structure $\mathbf{S}^*$ is obtained by optimally translating, rotating, and reflecting $\mathbf{S}'$ via an iterative search algorithm, as described in Zhang *et al.* [91] ($\mathbf{S}^* = \mathbf{R}\mathbf{S}' - \mathbf{t}$ where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is a Kabsch's rotation matrix, $\mathbf{t} \in \mathbb{R}^3$ is a translation vector). Because TM-score is dependent on structure size, we rescaled $\mathbf{S}$ and $\mathbf{S}'$ as in

Section 3.3.3, and v_{γ_1} indicates the scaling factor for each set of beads being evaluated.

We use the TM-score to provide the most global view of how similar two individual DNA molecules are. Accordingly, γ_1 indicates the set of beads in each molecule. Like GDT-TS, TM-score outputs values in $(0, 1]$, with higher values indicating greater global similarity.

Combined score

Although evaluating each similarity measure individually can give detailed information about the ways in which two structures are similar, there is also utility in defining a single score to indicate overall structural similarity. We compute such a score by taking the weighted mean of normalized similarity scores. A total of 70% of this "combined score" accounts for intra-molecular similarity, with each intra-molecular score (RMSD, intra-molecular distance error, GDT-TS, and TM-score) contributing equally to this total. The inter-molecular distance error contributes 20% to the combined score, and the distance error of inter-homolog intra-chromosomal distances contributes 10%.

To ensure that the components of the combined score are on the same scale, we must normalize them prior to taking the weighted mean. We normalize each score between the target and predicted structures, $e(S, S')$, against $e(S, S^\circ)$, where S° is a "baseline" structure inferred without constraints and without multi-resolution optimization. The normalized score, $e^{\text{norm}}(S, S')$, is defined as follows:

$$e^{\text{norm}}(S, S') = \begin{cases} 1 - \frac{1-e(S,S')}{1-e(S,S^\circ)}, & \text{if } e \in \{\text{TM-score, GDT-TS}\} \\ 1 - \frac{e(S,S')}{e(S,S^\circ)}, & \text{otherwise} \end{cases} \tag{3.27}$$

This normalization results in $e^{\text{norm}}(S, S')$ values in $[-\infty, 1]$, with higher values indicating greater similarity and a value of 0 indicating that performance is equivalent with that of the baseline.

3.3.4 Naive approach to multi-resolution inference

In addition to the approach outlined in Section 3.2.2, we also experimented with a simpler approach, similar to that proposed by Rosenthal *et al*. [72]. In general, to infer any low-resolution structure, contact counts must be formulated in terms of distances between low-resolution beads.

The naive approach uses the same Poisson model and transfer function for the relationship between high-resolution counts and high-resolution distances as it does for the relationship between low-resolution counts and low-resolution distances.

3.3.5 Tuning the constraints

Optimal constraint penalties (λ_1 and λ_2) were determined via grid search. Separate grid searches were performed for each inference setting. During multi-resolution optimization, constraints were tuned based on performance at the final resolution. The combination of penalties that maximized the combined error score between target and predicted structures (Section 3.3.3) was used for subsequent analyses. Constraints were not used when inferring structures from entirely unambiguous counts data because they are not necessary in this setting.

3.4 Results

3.4.1 Multi-resolution optimization improves accuracy of structures inferred from simulated data

We assessed the performance of multi-resolution optimization using simulated datasets, which have the advantage of involving a predefined "true" structure. In general, we expect structures inferred with multi-resolution optimization to be more similar to the true structure than those inferred without a multi-resolution strategy. In addition to the negative binomial model of multi-resolution optimization, we also examined the results of a "naive" multi-resolution approach, in which the same transfer function is applied at high and low resolution [72]. Because the relationship between counts and distances is not consistent across high and low resolution [93], we expect that applying the same transfer function across resolutions will result in overdispersion of the Poisson at low resolutions. We therefore anticipate that the naive multi-resolution approach will be less accurate than the negative binomial multiscale model.

We evaluated the performance of our multi-resolution optimization approach on simulated diploid Hi-C data (Section 3.3.2). The number of beads per chromosome in the simulated datasets corresponds to the diploid yeast genome at 16 kb, resulting in a total of $1,374$ beads. We generated

10 datasets, each with 10^7 simulated ambiguous reads. Structures were inferred using the modified bead-chain connectivity constraint (Section 3.2.3) as well as the modified homolog separation constraint(Section 3.2.3), with optimal constraint penalties (λ_1, λ_2) determined via grid search (Section 3.3.5). In all cases where multi-resolution optimization was applied, we inferred high-resolution structures via 5 rounds of optimization at progressively increasing resolutions, each of which doubled the resolution above what was obtained in the previous optimization round. Inference began with a 145 bead structure at the lowest resolution.

As anticipated, the simulation results suggest that our negative binomial model of multi-resolution optimization improves accuracy of the inferred structures above what can be attained with the naive multi-resolution approach or with single-resolution inference. Per the combined structural similarity score (Section 3.3.3), the negative binomial model of multi-resolution optimization significantly outperforms both other inference strategies (Figure 3.1). Improvements are seen when inference is performs using the true value of α (pairwise t-test, Bonferroni corrected p-values < 0.001 for comparison with single-resolution inference and 0.006 for comparison with the naive model) as well as when α is inferred jointly alongside the structure (pairwise t-test, Bonferroni corrected p-values 0.001 for comparison with single-resolution inference and 0.004 for comparison with the naive model). The negative binomial multi-resolution approach also significantly outperformed single-resolution inference on each constituent of the combined structural similarity score. We additionally saw improved performance of the negative binomial model when compared to the naive multi-resolution model on five of the six individual similarity scores (Supplementary Figures B.1 and B.2). The improved results of the negative binomial model over the naive model suggest that accurately adapting the counts-to-distance transfer function across resolutions is important in obtaining the best results.

3.4.2 New constraints improve inference with multi-resolution optimization without interfering with single-resolution inference

Cauer *et al*. [18] demonstrated that successful diploid inference with PASTIS requires constraints that reflect prior knowledge of the structure. In particular, genomically adjacent beads are expected

to colocalize in 3D space, as enforced by the bead-chain connectivity constraint. Additionally, we expect some amount of separation between the homologs of a chromosome, which is formally maintained by the homolog separation constraint. Although these constraints successfully enabled single-resolution inference in Cauer *et al*., each constraint was associated with a theoretical weakness, either in general or specifically in the context of multi-resolution optimization. Therefore, we reformulate both constraints as described here (Section 3.2.3).

In brief, we were concerned that the previously published [18] bead-chain connectivity constraint, which minimizes the variance in the distance between genomically neighboring beads, would need to be applied with different penalties at different resolutions. We consequently reworked this constraint such that resolution does not impact how strongly the constraint must be applied (Section 3.2.3). With regard to the homolog separation constraint, the equation in Cauer *et al*. operated on the distance between homolog centers of mass, which may not be relevant for organisms with a Rabl-like genome. Therefore, we adjusted the constraint to work with all inter-homologous distances instead (Section 3.2.3). Our goal was to improve results obtained with multi-resolution optimization, ideally without impairing canonical single-resolution inference. We also aim to demonstrate that our modified homolog separation constraint does not impair structural inference for genomes that do not follow a Rabl-like orientation.

To evaluate the reformulated constraints, we compared performance of the old and new constraints on simulated ambiguous diploid data (Section 3.3.2). As before, the number of chromosomes and chromosome sizes mirror the structure of the diploid yeast genome at 16 kb, with each of the 10 datasets containing 10^7 simulated reads. For each simulated dataset, both the percentage of counts between different chromosomes and the percentage of counts between homologs of a given chromosome were consistent with the Hi-C data from Kim *et al*. [43]. Beads in each molecule were confined to a spherical territory, and simulated structures do not appear to have a Rabl-like orientation (Section 3.3.2). For each pair of constraints, optimal constraint penalties (λ_1, λ_2) were determined via grid search (Section 3.3.5).

The simulation results show that structures inferred with the new constraints are more accurate than those inferred with the Cauer *et al*. constraints. Per the combined structural similarity score

(Section 3.3.3), our novel constraints yield significantly better results than the constraints from Cauer *et al*. These improvements are seen for both single-resolution inference and multi-resolution inference with the negative binomial model, although the relative improvement is greater in the context of multi-resolution inference (Figure 3.2, pairwise t-test, p-values 0.001 for single-resolution and 0 for multi-resolution inference). For both single-resolution and multi-resolution inference, the new constraints significantly improved accuracy on both inter-molecular structural similarity scores (Supplementary Figures B.3 and B.4, pairwise t-test, p-values < 0.05). This demonstrates that the novel homolog separating constraint is able improve the relative orientation of homologous chromosomes, even when the structure does not have a Rabl-like orientation. Results with intra-molecular structural similarity scores were more complicated, but show the expected superiority of the new constraints in the context of multi-resolution optimization. The Cauer *et al*. constraints yielded significantly better RMSD and distance error when used with single-resolution inference, and the novel constraints significantly improved TM-score and GDT when used with multi-resolution inference (Supplementary Figures B.3 and B.4). Because we anticipated that the ideal penalty of the Cauer *et al*. bead-chain connectivity constraint would vary based on resolution, and these results were generated with a single penalty across all resolutions, we suspect that the novel bead-chain connectivity constraint improved intra-molecular similarity in the context of multi-resolution inference because it is not dependent on resolution.

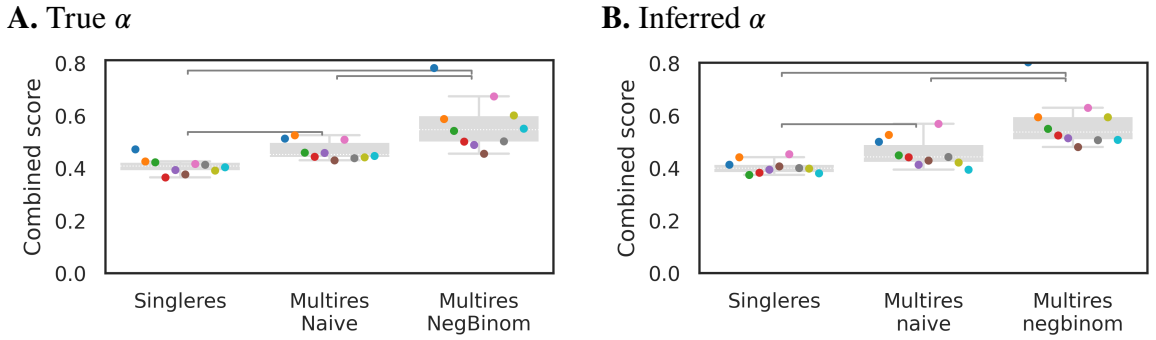

Figure 3.1: **Multi-resolution optimization improves diploid inference on simulated ambiguous data.** The simulated data consists of a diploid genome with a total of 1,374 beads and 10^7 ambiguous reads. Structures were inferred via three different inference methods (x-axis): without multi-resolution optimization, multi-resolution optimization with the naive approach, and multi-resolution optimization with our novel negative binomial model. The α parameter was either fixed at the same value used during simulation **(A)** or jointly inferred alongside the structure **(B)**. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, Bonferroni corrected p-value < 0.05). There were significant improvements in the combined error score (y-axis, Section 3.3.3) when multi-resolution optimization was applied with the negative binomial model, relative to multi-resolution optimization with the naive approach or inference without multi-resolution optimization. Each of the ten individual datasets shows the best results with the negative binomial multi-resolution model. For results on each constituent of the combined error score, see Supplementary Figures B.1 and B.2.

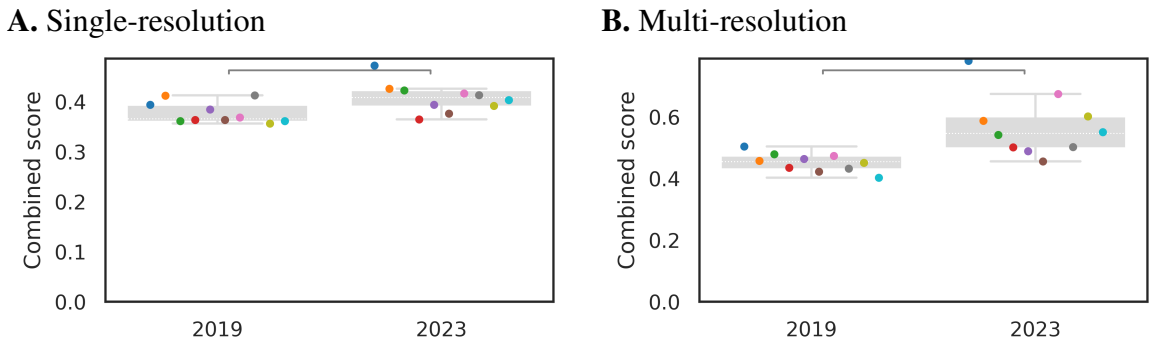

Figure 3.2: **Novel constraints improve diploid inference on simulated ambiguous data.** The simulated data consists of a diploid genome with a total of 1,374 beads and 10^7 ambiguous reads. Structures were inferred via two different inference methods: without multi-resolution optimization **(A)** and multi-resolution optimization with our novel negative binomial model **(B)**. Inference was performed using either the two constraints described in Cauer *et al.*, 2019 [18] or our two novel constraints (x-axis). The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, p-value < 0.05). The combined error score (y-axis, Section 3.3.3) is significantly better when the novel constraints were applied, and the magnitude of difference is most pronounced in the context of multi-resolution optimization. However, in both cases, each of the ten individual datasets shows the best results with the new constraints. For results on each constituent of the combined error score, see Supplementary Figures B.3 and B.4.

Chapter 4

CONCLUSION

4.1 Significance of contributions

The 3D organization of the genome plays an important part in regulating a broad array of basic cellular functions [68, 80, 45, 24, 60]. Additionally, numerous lines of evidence now link changes in the diverse components of 3D genome architecture to many different human diseases. The analysis methods and software produced by this thesis can help shed light on the specific mechanisms by which 3D genome derangement causes disease phenotypes.

Chapter 2 describes the first method for inferring consensus diploid chromatin structures from bulk Hi-C data. The optimization process used to achieve this goal is also novel. For instance, ambiguous contact counts are modeled as the sum of the interactions between all pairs of unambiguous contact counts for a given pair of genomic loci. This approach enables inference to use unphased data, which makes up the majority of the Hi-C reads produced by nearly all samples and is often the only type of data available. The method also accepts contacts phased on one or both ends, if available, and the inclusion of such data further improves inference. Inference also relies on two constraints, one of which is novel.

Chapter 3 introduces a multi-resolution optimization strategy that substantially improves the quality of structures inferred from diploid data. This method is the first to accurately model the counts-to-distance relationship at different resolutions. In order to improve low-resolution inference and more accurately model the genomic structures of organisms that do not form chromosome territories, both of the constraints introduced in Chapter 2 were replaced with novel constraints in Chapter 3.

4.2 Further validating inferred homologs

In the future, it would be useful to further demonstrate that the 3D structures inferred from ambiguous Hi-C data alone are meaningful. Resolving accurate diploid chromatin structures is a difficult problem, and a robust characterization of the solutions provided by diploid structural inference methods such as PASTIS would provide additional confidence in these approaches. The following experiments have not been conducted for any of the previously published diploid-compatible consensus chromatin structural inference methods, and it would be interesting to use these experiments to compare among diploid approaches.

First, we could validate structures using microscopy data from super-resolution chromatin tracing experiments [7, 63, 86, 54, 31, 62]. Since this comparison is between a structure derived from bulk data and the structures of thousands of single cells, we would assess similarity in the distances between loci rather than comparing the 3D structures themselves. This would allow us to conveniently compare the inferred distances with the mean or median of the microscopy-based distances.

It would also be useful to compare inferred and expected structures with regard to the structural similarity between homologs of a given chromosome. For simulated data, we could assess this directly by comparing inferred structures against the true structure. We could also assess the yeast data generated by Kim *et al*. [43] in this manner, as discussed in Sections 3. Lastly, we could evaluate the bipartite index (Section 2.3.3) of structures inferred from the mouse X chromosome. Unlike the results described in Section 2, we would hide all available phasing information and treat all counts as ambiguous.

4.3 Limitations of our approach

We wish to highlight three significant limitations of our method. The first two limitations also apply to the multitude of chromatin structural inference methods that use the same counts-to-distance transfer function as PASTIS [72, 88, 71, 74] The final limitation is only relevant to the comparatively small number of of 3D inference methods that rely on a Poisson model of contact counts [72, 88] Resolution of these limitations has the potential to improve inference with either bulk or

single-cell data in the haploid or diploid setting.

First, the counts-to-distance transfer function ($f(d_{ij}) \sim d_{ij}^{\alpha},\ \alpha < 0$) may not be appropriate for inter-molecular distances. This transfer function is based on a simple biophysical model of polymer packing [50], and provides a convenient and well-supported method for relating intra-molecular counts to distances. While this counts-to-distance transfer function is commonly used for both haploid and diploid methods [72, 88, 71, 74] we are unaware of any approach that uses an alternative transfer function for inter-molecular distances. Although establishing the correct counts-to-distance relationship for inter-molecular data is not trivial, such an alternative transfer function would not only benefit PASTIS but many other structural inference as well.

There is also room for improvement in the way this counts-to-distance transfer function is applied to intra-molecular distances. Both PASTIS and most other methods that utilize this transfer function use the same value of α for all interactions. Segal *et al*. infer a separate α for each chromosome [74]. Because their approach infers individual chromosomes separately before assembling them into a whole-genome structure, this strategy is necessary for their method. However, it may be beneficial to explore whether using a different α for each chromosome is generally advantageous. The value of α could also be modified based on other genomic features. For example, the distances-to-counts relationship may be affected by chromatin compaction and allowing the transfer function to differ between regions of open and closed chromatin may improve results.

The second limitation involves the Poisson model used to model high-resolution contact counts. As evidenced by recent work from Varoquaux *et al*., the Poisson distribution is unlikely be the best fit for the Hi-C data. Varoquaux *et al*. extended the haploid inference capabilities of PASTIS to model counts as a negative binomial distribution, rather than a Poisson [85]. The authors demonstrate that the data has higher variance than is predicted by the Poisson model, a phenomenon known as "overdispersion." The use of a negative binomial model uncouples the mean from the variance, resolving the issue of overdispersion. This approach was shown to be especially advantageous when structures were inferred from low-coverage bulk Hi-C datasets. Furthermore, accurate modeling of the mean-variance relationship is likely to be a critical component of single-cell inference.

However, accounting for the overdispersion of high-resolution counts data is complicated in the

context of our diploid structural inference method. To model the ambiguous and partially ambiguous interaction counts, our method relies on the property that the sum of i independent Poisson random variables of intensities λ_i is a Poisson variable of intensity $\sum_i \lambda_i$. There is no analogous property for the sum of negative binomial random variables. The gamma distribution also lacks this property; however, moment matching provides a very good approximation for the sum of gamma-distributed random variables, a feature that we apply to allow for multi-resolution diploid inference in Section 3. It may be possible to solve the problem of overdispersion by modeling high-resolution counts as a gamma distribution. Because the gamma distribution is continuous, the integer-based counts could be transformed by normalization and scaling prior to inference. In the context on multi-resolution optimization, a gamma-based model of high-resolution counts would result in a gamma-gamma compound distribution, also known as the beta prime distribution, at low resolution.

4.4 Potential improvements on our method

There are also other strategies, aside from rectification of the above limitations, that may benefit chromatin structural inference. These strategies could be applied to the inference of either bulk or single-cell structures with PASTIS or other methods.

First, PASTIS and any other multi-resolution inference methods [72] may yield superior results when combined with one of the multiscale optimization approaches that assembles a whole-genome structure in a piecewise manner [74, 71]. These methods infer individual chromosomes in isolation before incorporating them into a whole-genome structure. The approach described by Segal *et al*. [74] seems especially promising, since the relative orientations of any compact subunits in the chromosomal structures are allowed to shift during the process of assembling the chromosomes into a whole-genome structure.

Another approach that might benefit structural inference, particularly in the context of noisy data, involves utilizing counts from significantly interacting loci, as determined by methods such as FitHiC [2]. These significant interaction loci could be weighted more strongly during inference, or inference could simply be limited to these significant interactions. The latter strategy has

successfully been employed in a previously published structural inference method [64]. While it may be impossible to distinguish significantly interacting loci from high-resolution single-cell Hi-C data, which tends to be extremely sparse, significantly interacting loci could still be identified from low-resolution data and subsequently used during a multi-resolution inference approach.

4.5 Inferring single-cell chromatin structures

The variability between cells presents an additional barrier to progress in the field. A consensus structure inferred from population data may not resemble any individual cell, but ensemble methods that operate on population data are extremely underdetermined. Therefore, to assess variability between cells and more accurately determine chromatin structure, it is essential to use single-cell Hi-C data. There would be great utility in methods that can address these barriers and simplify optimization, validation, and analysis of single-cell chromatin structures. The following two strategies could be applied to any preexisting single-cell method and may also be helpful in adapting a consensus approach previously used on bulk Hi-C data for use with single-cell datasets.

While inferring structures from single-cell Hi-C experiments is necessary to address the variability between cells, single-cell data presents additional difficulties for 3D structural inference. First, inferring the structure of each cell is computationally expensive. Second, the data for any particular cell is incredibly sparse, since only a subset of all chromosomal contacts are measured. One could resolve the former challenge by only building models for a set of representative cells. Since we expect the structures of certain cells to be redundant, a well-chosen set of cells should be able to represent most of the information in the full population. Limiting inference to representative cells would also aid interpretability of the results. The latter challenge could be addressed by aggregating counts from each representative cell with data from other highly similar cells using weights that reflect degrees of similarity.

If sparsity is still a problem, even after aggregating counts from similar cells, it may be beneficial to add a restraint that prevents pairs of loci that lack contacts in bulk Hi-C from coming into contact in any of the single-cell models. The rationale is that the absence of bulk Hi-C contacts between pairs of loci can be assumed universal to all cells, whereas we wouldn't want to constrain by the

presence of contacts in bulk Hi-C data because it isn't clear which individual cells gave rise to these contacts.

4.6 Comparing and integrating our method with microscopy data

While most chromatin structural inference methods use only the counts produced by chromatin conformation capture assay as input, inference techniques that also include orthogonal datasets from the same cell population could enable a more accurate and detailed picture of subnuclear 3D architecture. For example, it may be interesting to incorporate data from microscopy of chromatin-bound nuclear proteins or whole-genome chromosome tracing techniques [7, 63, 86, 54, 31, 62]. Because the limitations and biases inherent in chromatin conformation capture assays may not necessarily overlap with those of the orthogonal assays, the techniques used could compensate for each other's weaknesses. Furthermore, superposition of this orthogonal data on the inferred structure could potentially yield insights into the spatial relationships between DNA and various subnuclear structures and compartments (speckles, pores). It could also shed light on the variability of 3D architecture across otherwise homogeneous populations of cells.

However, joint inference of single-cell Hi-C data and orthogonal single-cell datasets, such as those that result from microscopy, poses a challenge: it is difficult to determine which cell in the orthogonal dataset best corresponds to each cell in the Hi-C data. While it may be possible to solve this problem by evaluating structural similarity scores between all microscopy-based structures and all structures inferred from single-cell Hi-C data alone, this process would be computationally costly and may yield sub-optimal results. The inferred structures would need to be rescaled and aligned to match the size and orientation of the microscopy-based structures, and the structural similarity score used during alignment would need to capture the most important global features of chromatin structures. When working with microscopy of nuclear proteins, these tasks are particularly challenging because the images do not reveal the relative ordering of the genomic loci to which proteins are bound. Alternatively, the output of multiple assays could be integrated using an "*in silico* co-assay" method such as MMD-MA [52], which embeds cells profiled by different experiments into a learned latent space.

BIBLIOGRAPHY

[1] Mary V Arrastia, Joanna W Jachowicz, Noah Ollikainen, Matthew S Curtis, Charlotte Lai, Sofia A Quinodoz, David A Selck, Mitchell Guttman, and Rustem F Ismagilov. A single-cell method to map higher-order 3d genome organization in thousands of individual cells reveals structural heterogeneity in mouse es cells. *bioRxiv*, pages 2020–08, 2020.

[2] F. Ay, T. L. Bailey, and W. S. Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research*, 24:999–1011, 2014. PMC4032863.

[3] F. Ay, E. M. Bunnik, N. Varoquaux, S. M. Bol, J. Prudhomme, J.-P. Vert, W. S. Noble, and K. G. Le Roch. Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Research*, 24:974–988, 2014.

[4] F. Ay, T. H. Vu, M. J. Zeitz, N. Varoquaux, J. E. Carette, J.-P. Vert, A. R. Hoffman, and W. S. Noble. Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C. *BMC Genomics*, 16(121), 2015.

[5] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. Kraemer, M. Chotalia, S. Q. Xie, M. Barbieri, I. de Santiago, L. M. Lavitas, M. R. Branco, J. Fraser, J. Dostie, L. Game, N. Dillon, P. A. Edwards, M. Nicodemi, and A. Pombo. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, 2017.

[6] Brian J Beliveau, Alistair N Boettiger, Maier S Avendaño, Ralf Jungmann, Ruth B McCole, Eric F Joyce, Caroline Kim-Kiselak, Frédéric Bantignies, Chamith Y Fonseka, Jelena Erceg, et al. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using oligopaint fish probes. *Nature communications*, 6(1):7147, 2015.

[7] Brian J Beliveau, Alistair N Boettiger, Guy Nir, Bogdan Bintu, Peng Yin, Xiaowei Zhuang, and C-ting Wu. In situ super-resolution imaging of genomic dna with oligostorm and oligodna-paint. *Super-resolution microscopy: methods and protocols*, pages 231–252, 2017.

[8] Brian J Beliveau, Eric F Joyce, Nicholas Apostolopoulos, Feyza Yilmaz, Chamith Y Fonseka, Ruth B McCole, Yiming Chang, Jin Billy Li, Tharanga Niroshini Senaratne, Benjamin R Williams, et al. Versatile design and synthesis platform for visualizing genomes with oligopaint fish probes. *Proceedings of the National Academy of Sciences*, 109(52):21301–21306, 2012.

[9] Anastasiya Belyaeva, Kaie Kubjas, Lawrence J Sun, and Caroline Uhler. Identifying 3d genome organization in diploid organisms via euclidean distance geometry. *SIAM Journal on Mathematics of Data Science*, 4(1):204–228, 2022.

[10] A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, and T. Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLOS Biology*, 3(5):e157, 2005.

[11] G. Bonora, V. Ramani, R. Singh, H. Fang, D. Jackson, S. Srivatsan, R. Qiu, C. Lee, C. Trapnell, J. Shendure, et al. Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and x inactivation. *bioRxiv*, 2020.

[12] E. M. Bunnik, K. B. Cook, N. Varoquaux, G. Batugedara, J. Prudhomme, A. Cort, L. Shi, C. Andolina, L. S. Ross, D. Brady, D. A. Fidock, F. Nosten, R. Tewari, P. Sinnis, F. Ay, J.-P. Vert, W. S. Noble, and K. G. Le Roch. Changes in genome organization of parasite-specific gene families during the *Plasmodium* transmission stages. *Nature Communications*, 15(9):1910, 2018.

[13] R. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[14] Daniel Capurso and Mark R Segal. Distance-based assessment of the localization of functional annotations in 3D genome reconstructions. *BMC Genomics*, 15:992, 18 2014.

[15] S Carstens, M Nilges, and M Habeck. Inferential structure determination of chromosomes from single-cell Hi-C data. *PLOS Computational Biology*, 12(12):e1005292, 2016.

[16] S Carstens, M Nilges, and M Habeck. Bayesian inference of chromatin structure ensembles from population Hi-C data. *bioRxiv*, page 493676, 2018.

[17] M. Carty, L. Zamparo, M. Sahin, A. Gonzalez, R. Pelosoof, O. Elemento, and C. S. Leslie. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nature Communications*, 8:15454, 2017.

[18] A. G. Cauer, G. Yardimci, J.-P. Vert, N. Varoquaux, and W. S. Noble. Inferring diploid 3D chromatin structures from Hi-C data. In Katharina T. Huber and Dan Gusfield, editors, *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:13, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[19] Sandeep Chakraborty, Ravindra Venkatramani, Basuthkar J Rao, Bjarni Asgeirsson, and Abhaya M Dandekar. Protein structure quality assessment based on the distance profiles of consecutive backbone cα atoms. *F1000Research*, 2:Article–ID, 2013.

[20] Yu Chen, Yang Zhang, Yuchuan Wang, Liguo Zhang, Eva K Brinkman, Stephen A Adam, Robert Goldman, Bas Van Steensel, Jian Ma, and Andrew S Belmont. Mapping 3d genome organization relative to nuclear compartments using tsa-seq as a cytological ruler. *J Cell Biol*, 217(11):4025–4048, 2018.

[21] Diego Cifuentes, Jan Draisma, Oskar Henriksson, Annachiara Korchmaros, and Kaie Kubjas. 3d genome reconstruction from partially phased hi-c data. *arXiv preprint arXiv:2301.11764*, 2023.

[22] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.

[23] X. Deng, W. Ma, V. Ramani, A. Hill, F. Yang, F. Ay, J. B. Berletch, C. A. Blau, J. Shendure, Z. Duan, W. S. Noble, and C. M. Disteche. Bipartite structure of the inactive mouse X chromosome. *Genome Biology*, 16:152, 2015.

[24] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331, 2015.

[25] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble. A three-dimensional model of the yeast genome. *Nature*, 465:363–367, 2010.

[26] R. Fang, M. Yu, G. Li, S. Chee, T. Liu, A. D. Schmitt, and B. Ren. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Research*, 26(12):1345–1348, 2016.

[27] G Fischer, SA James, IN Roberts, SG Oliver, and EJ Louis. Chromosomal evolution in saccharomyces. *Nature*, 405(6785):451–454, 2000.

[28] G. Fudenberg and L. A. Mirny. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev.*, 22(2):115–124, 2012.

[29] Melissa J Fullwood, Chia-Lin Wei, Edison T Liu, and Yijun Ruan. Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses. *Genome research*, 19(4):521–532, 2009.

[30] L Giorgetti, R Galupa, E P Nora, T Piolot, F Lam, J Dekker, G Tiana, and E Heard. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950–963, 2014.

[31] Andrés M Cardozo Gizzi, Diego I Cattoni, Jean-Bernard Fiche, Sergio M Espinola, Julian Gurgo, Olivier Messina, Christophe Houbron, Yuki Ogiyama, Giorgio L Papadopoulos, Giacomo Cavalli, et al. Microscopy-based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Molecular cell*, 74(1):212–222, 2019.

[32] Rainer Heintzmann and Christoph G Cremer. Laterally modulated excitation microscopy: improvement of resolution by using a diffraction grating. In *Optical biopsies and microscopic techniques III*, volume 3568, pages 185–196. SPIE, 1999.

[33] Stefan W Hell and Jan Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics letters*, 19(11):780–782, 1994.

[34] Y Hirata, A Oda, K Ohta, and K Aihara. Three-dimensional reconstruction of single-cell chromosome structure using recurrence plots. *Scientific reports*, 6:34982, 2016.

[35] Claire Hoencamp, Olga Dudchenko, Ahmed MO Elbatsh, Sumitabha Brahmachari, Jonne A Raaijmakers, Tom van Schaik, Ángela Sedeño Cacciatore, Vinícius G Contessoto, Roy GHP van Heesbeen, Bram van den Broek, et al. 3d genomics across the tree of life reveals condensin ii as a determinant of architecture type. *Science*, 372(6545):984–989, 2021.

[36] Michael Hofmann, Christian Eggeling, Stefan Jakobs, and Stefan W Hell. Breaking the diffraction barrier in fluorescence microscopy at low light intensities by using reversibly photoswitchable proteins. *Proceedings of the National Academy of Sciences*, 102(49):17565–17569, 2005.

[37] T. S. Hsieh, A. Weiner, B. Lajoie, J. Dekker, N. Friedman, and O. J. Rando. Mapping nucleosome resolution chromosome folding in yeast by Micro-C. *Cell*, 162(1):108–119, 2015.

[38] M. Hu, K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, B. Ren, and J. S. Liu. Bayesian inference of spatial organizations of chromosomes. *PLOS Comput Biol*, 9(1):e1002893, 2013.

[39] J. R. Hughes, N. Roberts, S. McGowan, D. Hay, E. Giannoulatou, M. Lynch, M. De Gobbi, S. Taylor, R. Gibbons, and D. R. Higgs. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics*, 46(2):205–212, 2014.

[40] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9:999–1003, 2012.

[41] I. Junier, R. K. Dale, C. Hou, F. Kepes, and A. Dean. CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the α-globin locus. *Nucleic Acids Research*, 40(16):7718–7727, 2012.

[42] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1):90–98, 2011.

[43] S. Kim, I. Liachko, D. G. Brickner, K. Cook, W. S. Noble, J. H. Brickner, J. Shendure, and M. J. Dunham. The dynamic three-dimensional organization of the diploid yeast genome. *eLife*, 6, 2017.

[44] Jop Kind, Ludo Pagie, Sandra S de Vries, Leila Nahidiazar, Siddharth S Dey, Magda Bienko, Ye Zhan, Bryan Lajoie, Carolyn A de Graaf, Mario Amendola, et al. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, 163(1):134–147, 2015.

[45] P H L Krijger, B Di Stefano, E de Wit, F Limone, C Van Oevelen, W De Laat, and T Graf. Cell-of-origin-specific 3D genome structure acquired during somatic cell reprogramming. *Cell Stem Cell*, 18(5):597–610, 2016.

[46] T. Krumm and Z. Duan. Understanding the 3D genome: Emerging impacts on human disease. *Seminars in Cell & Developmental Biology*, 90:62–77, 2019.

[47] Solomon Kullback. *Information Theory And Statistics*. Dover, 1968.

[48] T. B. K. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, 342(6159):731–734, 2013.

[49] A. Lesne, J. Riposo, P. Roger, A. Cournac, and J. Mozziconacci. 3D genome reconstruction from chromosomal contacts. *Nature Methods*, 11(11):1141–1143, 2014.

[50] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[51] D Lin, G Bonora, G G Yardımcı, and W S Noble. Computational methods for analyzing and modeling genome structure and organization. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 11(1):e1435, 2019.

[52] J. Liu, Y. Huang, R. Singh, J.-P. Vert, and W. S. Noble. Jointly embedding multiple single-cell omics measurements. In Katharina T. Huber and Dan Gusfield, editors, *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 10:1–10:13, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. PMC8496402.

[53] Wenxiu Ma, Giancarlo Bonora, Joel B Berletch, Xinxian Deng, William S Noble, and Christine M Disteche. X-chromosome inactivation and escape from x inactivation in mouse. *X-Chromosome Inactivation: Methods and Protocols*, pages 205–219, 2018.

[54] Leslie J Mateo, Sedona E Murphy, Antonina Hafner, Isaac S Cinquini, Carly A Walker, and Alistair N Boettiger. Visualizing dna folding and rna in embryos at single-cell resolution. *Nature*, 568(7750):49–54, 2019.

[55] D. Meluzzi and G. Arya. Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Res.*, 41(1):63–75, Jan 2013.

[56] C W Metz. Chromosome studies on the diptera. ii. the paired association of chromosomes in the diptera, and its significance. *Journal of Experimental Zoology*, 21(2):213–279, 1916.

[57] B. Mifsud, F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, S. W. Wingett, S. Andrews, W. Grey, P. A. Ewels, B. Herman, S. Happe, A. Higgs, E. LeProust, G. A. Follows, P. Fraser, N. M. Luscombe, and C. S. Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, 47:596–606, 2015.

[58] Eric Mjolsness, Charles D Garrett, and Willard L Miranker. Multiscale optimization in neural nets. *IEEE Transactions on Neural Networks*, 2(2):263–274, 1991.

[59] M. R. Mumbach, A. J. Rubin, R. A. Flynn, C. Dai, P. A. Khavari, W. J. Greenleaf, and H. Y. Chang. Hichip: Efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, 13:919–922, 2016.

[60] T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[61] T. Nagano, Y. Lubling, C. Várnai, C. Dudley, W. Leung, Y. Baran, N. M. Cohen, S. Wingett, P. Fraser, and A. Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547:61–67, 2017.

[62] Huy Q Nguyen, Shyamtanu Chattoraj, David Castillo, Son C Nguyen, Guy Nir, Antonios Lioutas, Elliot A Hershberg, Nuno MC Martins, Paul L Reginato, Mohammed Hannan, et al. 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nature methods*, 17(8):822–832, 2020.

[63] G Nir, I Farabella, C P Estrada, C G Ebeling, B J Beliveau, H M Sasaki, S H Lee, S C Nguyen, R B McCole, S Chattoraj, et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLOS genetics*, 14(12):e1007872, 2018.

[64] J Paulsen, M Sekelja, A R Oldenburg, A Barateau, N Briand, E Delbarre, A Shah, A L Sørensen, C Vigouroux, B Buendia, et al. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome biology*, 18(1):21, 2017.

[65] S. A. Quinodoz, N. Ollikainen, B. Tabak, A. Palla, J. M. Schmidt, E. Detmar, M. M. Lai, A. A. Shishkin, P. Bhat, Y. Takei, V. Trinh, E. Aznauryan, P. Russell, C. Cheng, M. Jovanovic, A. Chow, L. Cai, P. McDonel, M. Garber, and M. Guttman. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, 174(3):744–757, 2018.

[66] V. Ramani, D.A. Cusanovich, R.J. Hause, W. Ma, R. Qiu, X. Deng, C.A. Blau, C.M. Disteche, W.S. Noble, J. Shendure, et al. Mapping 3D genome architecture through in situ DNase Hi-C. *Nature protocols*, 11(11):2104, 2016.

[67] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, 2017.

[68] S. S. P. Rao, M. H. Huntley, N. Durand, C. Neva, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 59(7):1665–1680, 2014.

[69] Josef Redolfi, Yinxiu Zhan, Christian Valdes-Quezada, Mariya Kryzhanovska, Isabel Guerreiro, Vytautas Iesmantavicius, Tim Pollex, Ralph S Grand, Eskeatnaf Mulugeta, Jop Kind, et al. Damc reveals principles of chromatin folding in vivo without crosslinking and ligation. *Nature structural & molecular biology*, 26(6):471–480, 2019.

[70] Stephen Richer, Yuan Tian, Stefan Schoenfelder, Laurence Hurst, Adele Murrell, and Giuseppina Pisignano. Widespread allele-specific topological domains in the human genome are not confined to imprinted gene clusters. *Genome Biology*, 24(1):40, 2023.

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

[72] Michael Rosenthal, Darshan Bryner, Fred Huffer, Shane Evans, Anuj Srivastava, and Nicola Neretti. Bayesian estimation of three-dimensional chromosomal structure from single-cell hi-c data. *Journal of Computational Biology*, 26(11):1191–1202, 2019.

[73] Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm). *Nature methods*, 3(10):793–796, 2006.

[74] M R Segal and H L Bengtsson. Reconstruction of 3d genome architecture via a two-stage algorithm. *BMC bioinformatics*, 16(1):373, 2015.

[75] S. Selvaraj, R. Dixon J, V. Bansal, and B. Ren. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology*, 31(12):1111–1118, 2013.

[76] S Shah, Y Takei, W Zhou, E Lubeck, J Yun, C Linus Eng, N Koulena, C Cronin, C Karp, E J Liaw, et al. Dynamics and spatial genomics of the nascent transcriptome by intron seqfish. *Cell*, 2018.

[77] Lin Shao, Feng Xing, Conghao Xu, Qinghua Zhang, Jian Che, Xianmeng Wang, Jiaming Song, Xianghua Li, Jinghua Xiao, Ling-Ling Chen, et al. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proceedings of the National Academy of Sciences*, 116(12):5653–5658, 2019.

[78] Bas van Steensel and Steven Henikoff. Identification of in vivo dna targets of chromatin proteins using tethered dam methyltransferase. *Nature biotechnology*, 18(4):424–428, 2000.

[79] L Tan, D Xing, C Chang, H Li, and X S Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, 2018.

[80] Z Tang, O J Luo, X Li, M Zheng, Jacqueline J Zhu, P Szalaj, P Trzaskoma, A Magalska, J Wlodarczyk, B Ruszczycki, et al. CTCF-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell*, 163(7):1611–1627, 2015.

[81] H. Tanizawa, O. Iwasaki, A. tanaka, J. R. Capizzi, P. Wickramasignhe, M. Lee, Z. Fu, and K. Noma. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research*, 38(22):8164–8177, 2010.

[82] H. Tjong, K. Gong, L. Chen, and F. Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Research*, 22(7):1295–1305, 2012.

[83] H Tjong, Wenyuan Li, R Kalhor, C Dai, S Hao, K Gong, Y Zhou, Haochen Li, Xianghong J Z, M A Le Gros, et al. Population-based 3d genome structure analysis reveals driving forces in spatial genome organization. *Proceedings of the National Academy of Sciences*, 113(12):E1663–E1672, 2016.

[84] N. Varoquaux, F. Ay, W. S. Noble, and J.-P. Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

[85] N. Varoquaux, W. S. Noble, and J.-P. Vert. Inference of 3D genome architecture by modeling overdispersion of Hi-C data. *Bioinformatics*, 39(1):btac838, 2023. 2021.02.04.429864.

[86] S. Wang, J. H. Su, B. J. Beliveau, B. Bintu, J. R. Moffitt, C. T. Wu, and X. Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, 2016.

[87] S Wang, J Xu, and J Zeng. Inferential modeling of 3D chromatin structure. *Nucleic Acids Research*, 43(8):e54, 2015.

[88] Tiantian Ye and Wenxiu Ma. Ashic: hierarchical bayesian modeling of diploid chromatin contacts and structures. *Nucleic acids research*, 48(21):e123–e123, 2020.

[89] A. Zemla. LGA – a method for finding 3d similarities in protein structures. *Nucleic Acids Research*, 31:3370–3374, 2003.

[90] B. Zhang and P. G. Wolynes. Topology, structures, and energy landscapes of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 112(19):6062–6067, 2015.

[91] Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57:702–710, 2004.

[92] Z Zhang, G Li, K-C Toh, and W-K Sung. 3D chromosome modeling with semi-definite programming and Hi-C data. *Journal of Computational Biology*, 20(11):831–846, 2013.

[93] Z. Zhang, G. Li, K.-C. Toh, and W.-K. Sung. Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-C data. In *Proceedings of the 17th International Conference on Research in Computational Molecular Biology*, volume 7821 of *Lecture Notes in Computer Science*, pages 317–332, Berlin, Heidelberg, 2013. Springer-Verlag.

[94] Meizhen Zheng, Simon Zhongyuan Tian, Daniel Capurso, Minji Kim, Rahul Maurya, Byoungkoo Lee, Emaly Piecuch, Liang Gong, Jacqueline Jufen Zhu, Zhihui Li, et al. Multiplex chromatin interactions with single-molecule precision. *Nature*, 566(7745):558–562, 2019.

Appendix A

APPENDIX TO "INFERRING DIPLOID 3D CHROMATIN STRUCTURES FROM HI-C DATA"

Table A.1: **Constraints improve ambiguous inference.** Each entry is a Bonferroni adjusted *p*-value for a *t*-test applied to the specified pair of methods. Values <0.05 are in boldface.

		RMSD per homolog	Distance error per homolog	Distance error, inter-homolog
No constraints	Bead-chain connectivity	**0.0458**	**0.0104**	0.275
No constraints	Homolog separation	1.3	0.222	4.14
No constraints	Both constraints	**0.00309**	**0.00667**	**0.0179**
No constraints	Both constraints + Null	**0.0000404**	**0.00000773**	**0.0000883**
Bead-chain connectivity	Homolog separation	**0.00731**	**0.00588**	1.62
Bead-chain connectivity	Both constraints	4.89	8.74	0.159
Bead-chain connectivity	Both constraints + Null	**0.000018**	**0.00000054**	**0.000263**
Homolog separation	Both constraints	**0.0018**	**0.00713**	0.183
Homolog separation	Both constraints + Null	**0.0000397**	0.152	0.103
Both constraints	Both constraints + Null	**0.0000179**	**0.00000387**	0.981

Table A.2: **Inference with ambiguous and disambiguated data.** Each entry is a Bonferroni adjusted *p*-value for a *t*-test applied to the specified pair of methods. Values <0.05 are in boldface.

		RMSD per homolog	Distance error per homolog	Distance error, inter-homolog
Ambiguous	Partially ambiguous	**0.000231**	**0.0000669**	0.654
Ambiguous	Unambiguous	**3.36E-09**	**2.88E-08**	**0.00000369**
Partially ambiguous	Unambiguous	**0.00000462**	**0.00000345**	**0.00000342**

Appendix B

APPENDIX TO "A MULTI-RESOLUTION OPTIMIZATION STRATEGY FOR INFERRING 3D GENOME ARCHITECTURE FROM HI-C DATA"

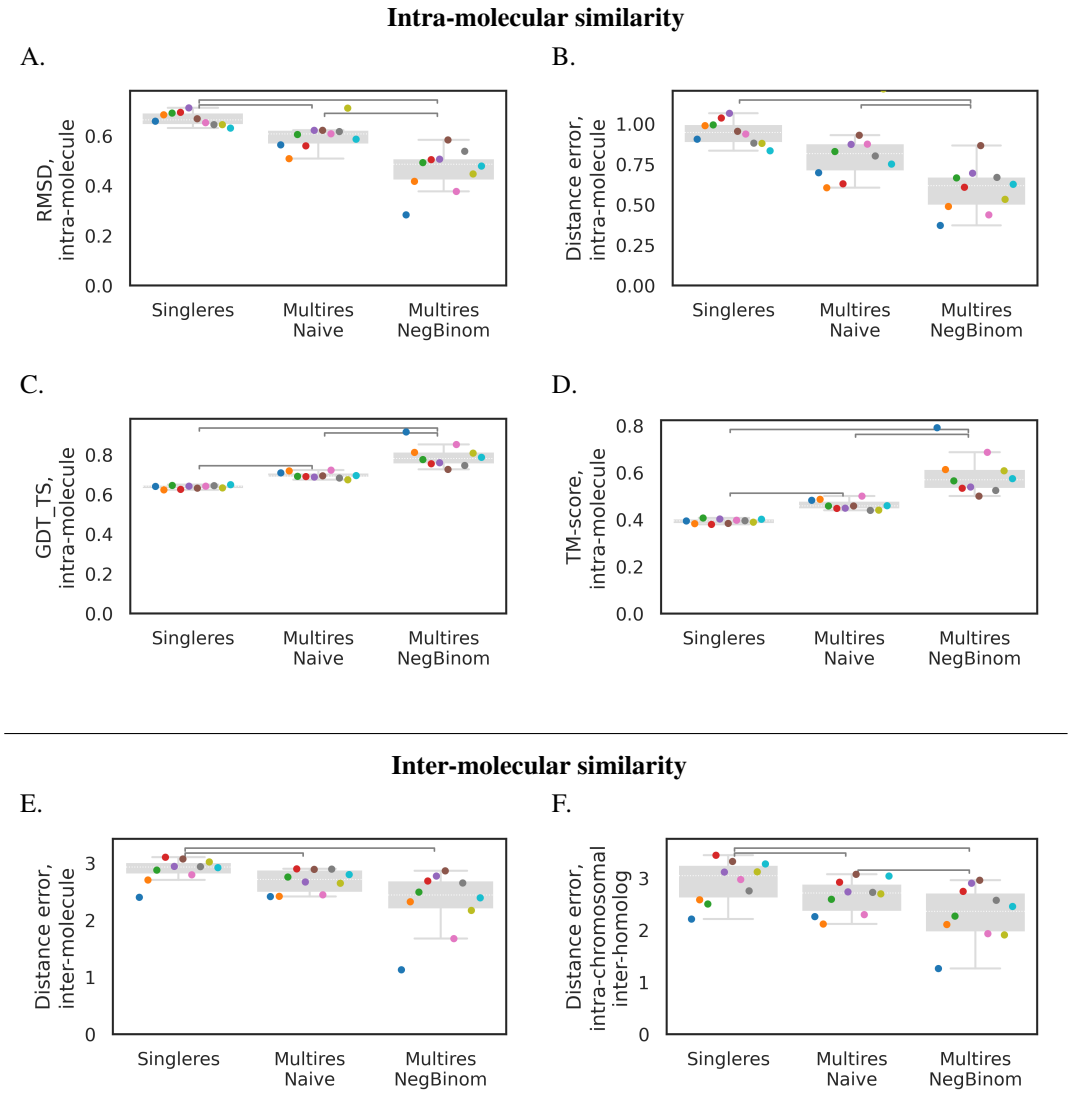

Figure B.1: **When simulated diploid structures are inferred from ambiguous data with α fixed at the true value, multi-resolution optimization improves accuracy along multiple measures.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred via three different inference methods (x-axis): without multi-resolution optimization, multi-resolution optimization with the naive approach, and multi-resolution optimization with our novel negative binomial model. The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, Bonferroni corrected p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The negative binomial-based multi-resolution model significantly outperforms single-resolution inference on all six measures. When comparing between the two multi-resolution optimization strategies, the negative binomial model yields significantly better intra-molecular similarity scores as well as significantly better intra-chromosomal inter-homolog distance error **(F)**.

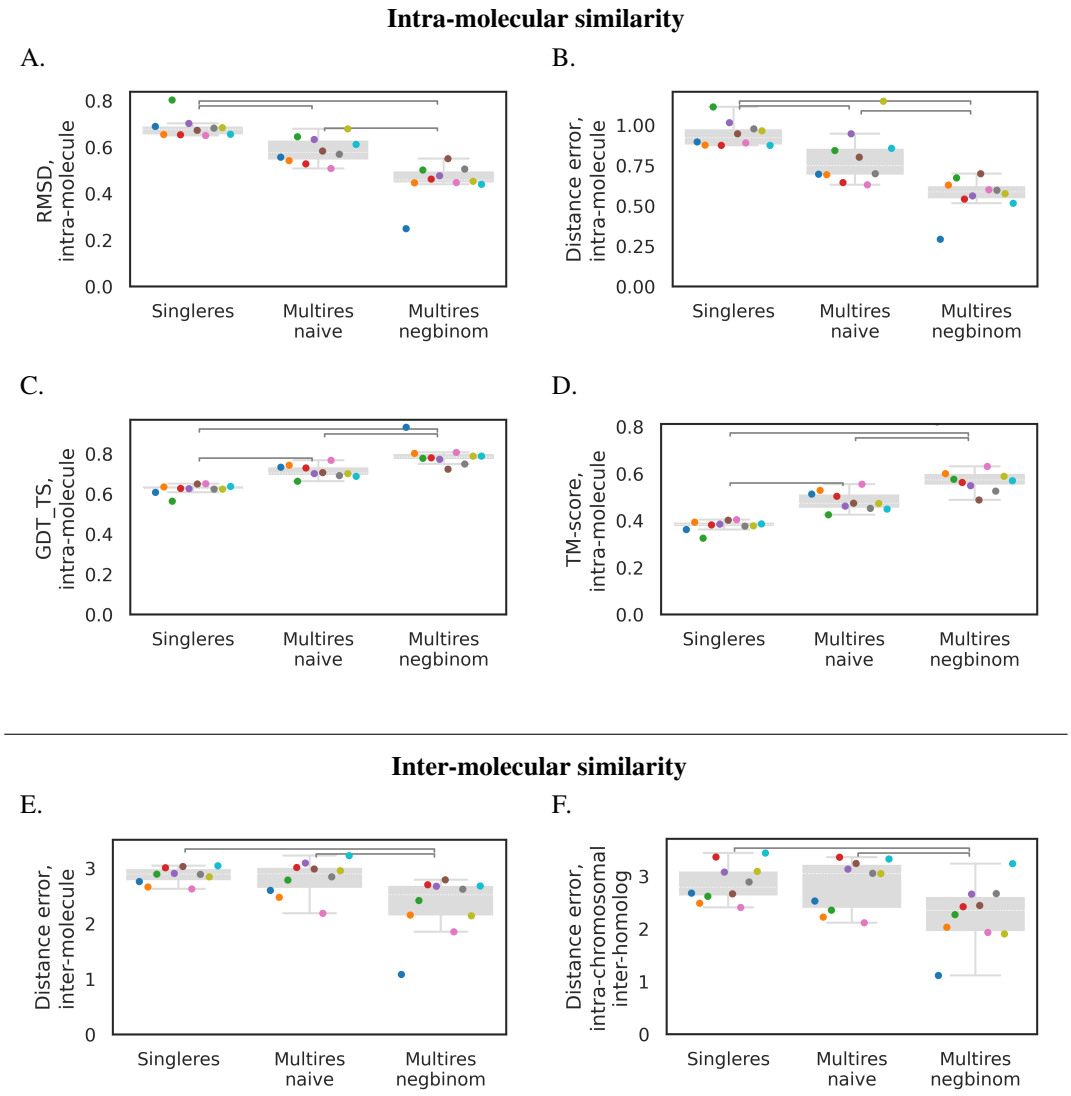

Figure B.2: **When jointly inferring simulated diploid structures and α from ambiguous data, multi-resolution optimization improves accuracy along multiple measures.** The simulated data consists of a diploid genome with a total of $1,374$ beads and 10^7 ambiguous reads. Structures were inferred via three different inference methods (x-axis): without multi-resolution optimization, multi-resolution optimization with the naive approach, and multi-resolution optimization with our novel negative binomial model. The value of α was jointly inferred alongside the structure. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, Bonferroni corrected p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The negative binomial-based multi-resolution model significantly outperforms single-resolution inference as well as inference with the naive multi-resolution model on all six similarity scores. Each of the ten individual datasets shows improvement on all measures with the negative binomial multi-resolution model.

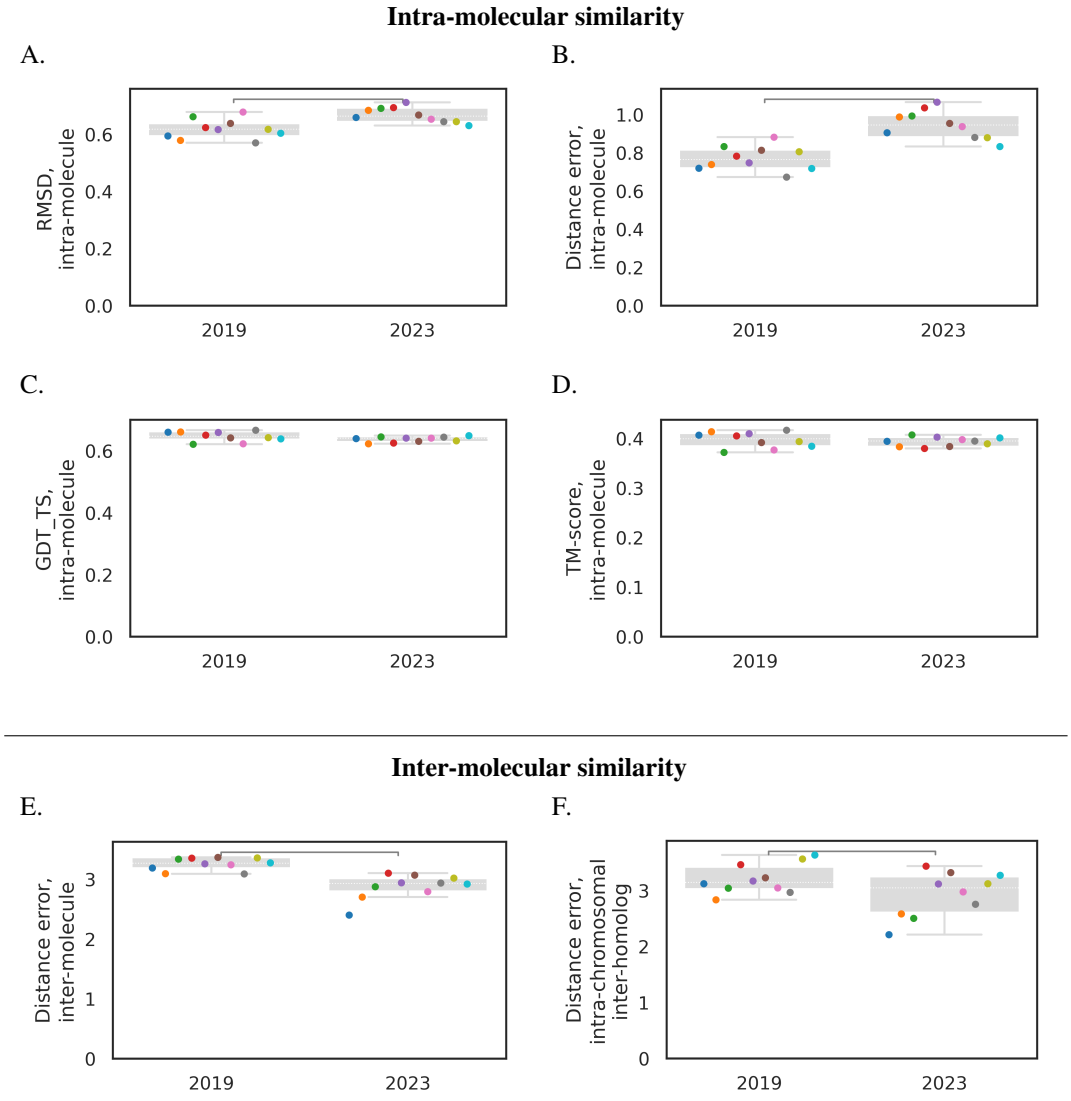

Figure B.3: **Novel constraints improve single-resolution diploid inference accuracy along multiple measures in the context of simulated ambiguous data.** The simulated data consists of a diploid genome with a total of 1,374 beads and 10^7 ambiguous reads. Structures were inferred without multi-resolution optimization. Inference was performed using either the two constraints described in Cauer *et al.*, 2019 [18] or our two novel constraints (x-axis). The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The novel constraints significantly outperform those of Cauer *et al.* with regard to intra-molecular both inter-molecular scores **(E, F)**. However, the Cauer *et al.* constraints significantly outperform ours on intra-molecular RMSD and distance error **(A, B)**. No significant differences are seen when comparing intra-molecular TM-score or GDT **(C, D)**.

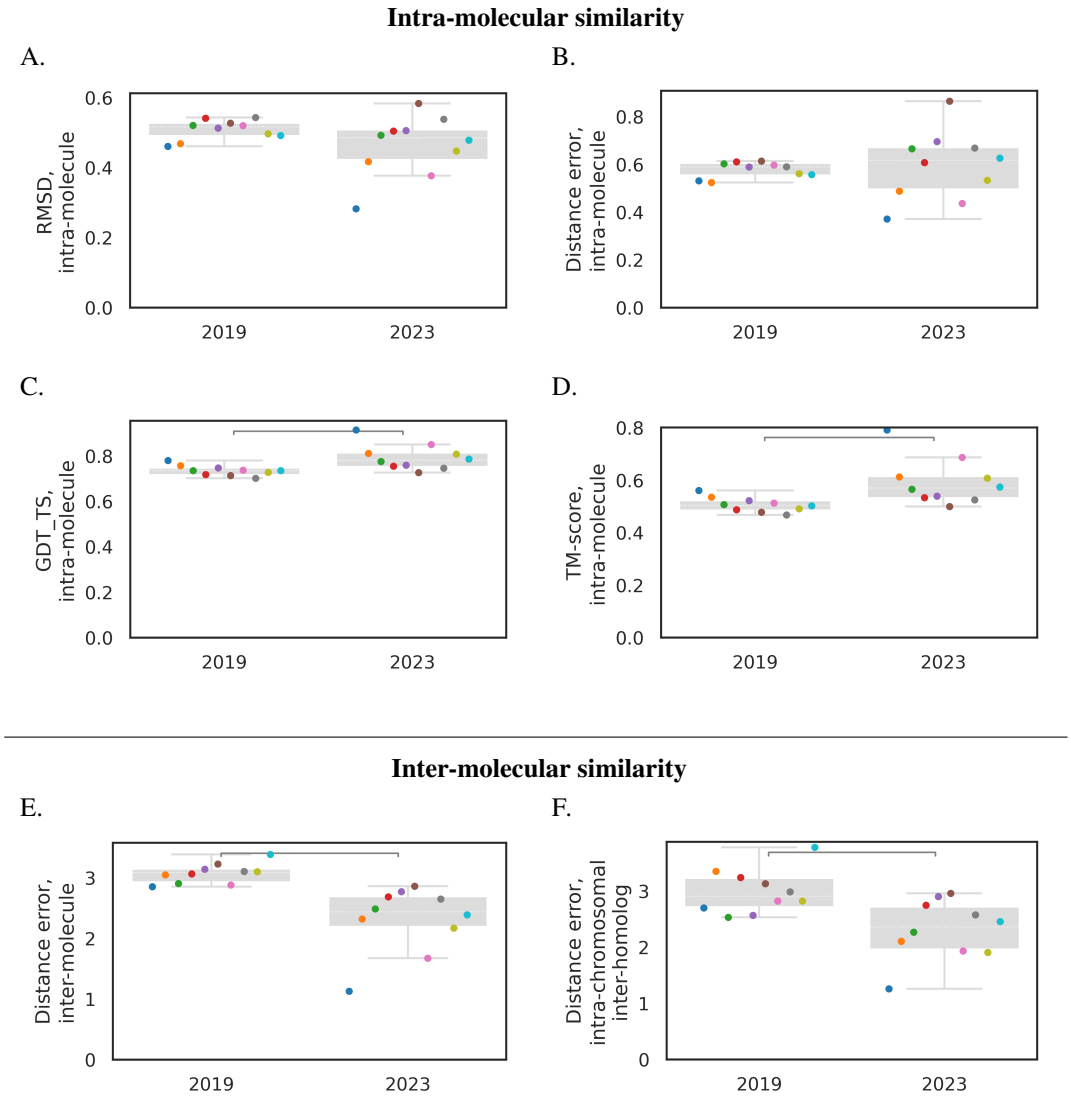

Figure B.4: **Novel constraints improve multi-resolution diploid inference accuracy along multiple measures in the context of simulated ambiguous data.** The simulated data consists of a diploid genome with a total of 1,374 beads and 10^7 ambiguous reads. Structures were inferred via multi-resolution optimization with our novel negative binomial model. Inference was performed using either the two constraints described in Cauer *et al*., 2019 [18] or our two novel constraints (x-axis). The α parameter was fixed at the same value used during simulation. Each point corresponds to a single inferred structure, and colors indicate the simulated true structure from which counts were derived. Significant differences are indicated (pairwise t-test, p-value < 0.05). Various intra-molecular **(A, B, C, D)** and inter-molecular **(E, F)** structural similarity scores are shown (y-axis). The novel constraints significantly outperform those of Cauer *et al*. with regard to intra-molecular TM-score and GDT **(C, D)** as well as both inter-molecular scores **(E, F)**, but no significant differences are seen when comparing intra-molecular RMSD or distance error **(A, B)**.

Table B.1: Notation

Variable	Definition
$\mathbf{Z}$	3D structure
m	Number of beads in Z
z_ℓ	3D coordinate of the ℓth bead in Z
C	Haploid contact counts matrix
C^A, C^U, and C^P	Diploid contact counts matrices: ambiguous (A), unambiguous (U), and partially ambiguous (P)
n	Number of loci in a given counts matrix
c_{ij}	High-resolution haploid contact counts between loci i and j
c^A_{ij}	High-resolution ambiguous diploid contact counts between loci i and j
$c^U_{\ell p}$	High-resolution unambiguous diploid contact counts between loci ℓ and p
c^P_{ij}	High-resolution partially ambiguous diploid contact counts between loci i and j.
c^*_{ij}	High-resolution ambiguated diploid contact counts between loci i and j, where any available phasing information is hidden
β	Coverage parameter for a given haploid counts matrix
β^A	Coverage parameter for a given ambiguous diploid counts matrix
β^*	Coverage parameter corresponding to a given ambiguated diploid contact counts matrix
b_i and b_j	Hi-C bias associated with loci i and j, respectively
$d_{\ell p}$	Euclidean distance between beads ℓ and p
α	Exponent in counts-to-distance transfer function

$\Phi : [1, m] \to [1, n]$	Mapping from bead ℓ to locus i. For haploid or unambiguous data, this is a one-to-one mapping, and $\ell = i$. Otherwise, it is a many-to-one mapping.
d_{ij}^{α}	$d_{ij}^{\alpha} = \sum_{\ell:\Phi(\ell)=i} \sum_{p:\Phi(p)=j} d_{\ell p}^{\alpha}$, where $d_{\ell p}^{\alpha}$is the counts-to-distance transfer function applied to high-resolution distances between beads ℓ and p. For haploid or unambiguous counts, $d_{ij}^{\alpha} = d_{\ell p}^{\alpha}$.
X and Y	Sets of consecutive high-resolution beads, each of which corresponds to a single low-resolution bead ($\overline{x}$ and $\overline{y}$, respectively)
m_X and m_Y	Number of beads in X and Y, respectively
x_{ℓ}	3D coordinate of the ℓth bead in X
y_p	3D coordinate of the pth bead in Y
$\overline{x}$ and $\overline{y}$	Low-resolution beads, corresponding to the center of mass of X and Y, respectively.
Δ_{XY}	Difference between the low-resolution beads $\overline{x}$ and $\overline{y}$
$\kappa_{\ell p}$	Independent standard normal vector ($\kappa_{\ell p} \sim \mathcal{N}(0, I_3)$)
ε	Determines the difference between $x_{\ell} - y_p$ and Δ_{XY} for each of the three dimensions in 3D space. Consequently, ε scales how similar high-resolution distances are to their corresponding low-resolution distance.
Λ	A set of high-resolution $d_{\ell p}^{\alpha}$ associated with two low-resolution beads, where the low-resolution beads are separated by a unit of 1
$m(\alpha, \varepsilon)$	Mean of Λ
$v(\alpha, \varepsilon)$	Variance of Λ
k_{XY}	Shape parameter of the gamma distribution that is associated with distances between X and Y

θ_{XY}	Scale parameter of the gamma distribution that is associated with distances between X and Y
$\Gamma(\cdot)$	The gamma function
$B_N(\cdot)$	Component of the Stirling approximation to $log\Gamma(\cdot)$
B_n	Bernoulli numbers
$\lvert\cdot\rvert$	Cardinality of a given set - the number of items in the set.
ω	The set of all bead indices such that $\ell \in \omega$ and $\ell + 1$ are on the same molecule
ω^*	The set of indices in the ambiguated counts matrix such that $i \in \omega^*$ and $i + 1$ are on the same chromosome
$d^{\female}_{i,i+1}$, $d^{\male}_{i,i+1}$	The distances between neighboring beads on the paternal and maternal homologs of each chromosome, respectively
$\mathcal{L}$	The log likelihood of the Poisson model (at high resolution) or the multi-resolution negative binomial model (at low resolution)
r_ψ	The expected distances between homolog centers of mass for chromosome ψ, as estimated prior to diploid inference
r'_ψ	The distances between homolog centers of mass for chromosome ψ in the current inferred structure
$c^{*\text{inter}-\text{chrom}}_{ij}$	High-resolution inter-chromosomal ambiguated counts between loci i and j, where i and j are on different chromosomes
$(d^{\alpha}_{\ell p})^{\text{inter}-\text{hmlg}}$	Counts-to-distance transfer function applied to high-resolution distances between beads ℓ and p, where ℓ and p are on different homologs.
$k^{\text{inter}-\text{hmlg}}$, $\theta^{\text{inter}-\text{hmlg}}$	Shape and scale parameter of the gamma distribution of $(d^{\alpha}_{\ell p})^{\text{inter}-\text{hmlg}}$
P	The distribution of high-resolution inter-chromosomal ambiguated counts

Q	The negative binomial distribution that is parameterized by high-resolution inter-homolog $(d_{\ell p}^{\alpha})^{\text{inter-hmlg}}$, where ℓ and p are beads on different homologs.
$h_1(\mathbf{Z})$	The bead-chain connectivity constraint
$h_2(\mathbf{Z})$	The homolog separation constraint
λ_1	Penalization parameter for the bead-chain connectivity constraint
λ_2	Penalization parameter for the homolog separation constraint
τ	The total number of counts bins across all inputted Hi-C matrices
$f(\mathbf{Z}, \alpha, \varepsilon)$	The objective function, including the log likelihood of the counts and any constraints
$\mathbb{E}[\cdot]$	Mean
Var[·]	Variance
$\mathbf{S}$	Target 3D structure, in similarity score calculation
$\mathbf{S}'$	Predicted 3D structure, in similarity score calculation
$\mathbf{S}^*$	3D structure obtained by optimally translating, rotating, and reflecting $\mathbf{S}'$
S°	A baseline 3D structure inferred without constraints and without multi-resolution optimization
s_i	The ith bead of S
s_i^*	The ith bead of S^*
γ_1	A set containing beads of interest
γ_2	A set containing pairs of beads that correspond to distances of interest
υ_{γ_1}	For TM-score, the scaling factor for each set of beads being evaluated
$e^{\text{norm}}(S, S')$	The normalized error score e between $\mathbf{S}$ and $\mathbf{S}'$