Society for
Mathematical
Biology

ORIGINAL ARTICLE

Check for
updates

3D Genome Reconstruction from Partially Phased Hi-C Data

Diego Cifuentes[1] · Jan Draisma[2] · Oskar Henriksson[3] ·
Annachiara Korchmaros[4] · Kaie Kubjas[5]

Abstract

The 3-dimensional (3D) structure of the genome is of significant importance for many
cellular processes. In this paper, we study the problem of reconstructing the 3D struc-
ture of chromosomes from Hi-C data of diploid organisms, which poses additional
challenges compared to the better-studied haploid setting. With the help of tech-
niques from algebraic geometry, we prove that a small amount of phased data is
sufficient to ensure finite identifiability, both for noiseless and noisy data. In the light
of these results, we propose a new 3D reconstruction method based on semidefinite
programming, paired with numerical algebraic geometry and local optimization. The
performance of this method is tested on several simulated datasets under different noise
levels and with different amounts of phased data. We also apply it to a real dataset from

✉ Kaie Kubjas
kaie.kubjas@aalto.fi

Diego Cifuentes
diego.cifuentes@isye.gatech.edu

Jan Draisma
jan.draisma@math.unibe.ch

Oskar Henriksson
oskar.henriksson@math.ku.dk

Annachiara Korchmaros
annachiara.korchmaros@uni-leipzig.de

1 School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive,
   NW, Atlanta, GA 30332, USA

2 Mathematisches Institut, University of Bern, Sidlerstrasse 5, 3012 Bern, Switzerland

3 Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100
   Copenhagen, Denmark

4 Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for
   Bioinformatics, University of Leipzig, Härtelstraße 16-18, 04107 Leipzig, Germany

5 Department of Mathematics and Systems Analysis, Aalto University, P.O. Box 11100, 00076 Aalto,
   Finland

🖄 Springer

mouse X chromosomes, and we are then able to recover previously known structural features.

Keywords 3D genome organization · Diploid organisms · Hi-C · Applied algebraic geometry · Numerical algebraic geometry

Mathematics Subject Classification 92E10 · 92-08 · 13P25 · 14P05 · 65H14 · 90C90

1 Introduction

The eukaryotic chromatin has a three-dimensional (3D) structure in the cell nucleus, which has been shown to be important in regulating basic cellular functions, including gene regulation, transcription, replication, recombination, and DNA repair (Uhler and Shivashankar 2017; Wang et al. 2018). The 3D DNA organization is also associated with brain development and function; in particular, it is shown to be misregulated in schizophrenia (Rajarajan et al. 2018; Rhie et al. 2018) and Alzheimer's disease (Nott et al. 2019).

All genetic material is stored in chromosomes, which interact in the cell nucleus, and the 3D chromatin structure influences the frequencies of such interactions. A benchmark tool to measure such frequencies is high-throughput chromosome conformation capture (Hi-C) (Lafontaine et al. 2021). Hi-C first crosslinks cell genomes, which "freezes" contacts between DNA segments. Then the genome is cut into fragments, the fragments are ligated together and then are associated with equally-sized segments of the genome using high-throughput sequencing (Rao et al. 2014). These segments of the genome are called *loci*, and their size is known as *resolution* (e.g., bins of size 1Mb or 50Kb). The result of Hi-C is stored in a matrix called *contact matrix* whose elements are the *contact counts* between pairs of loci.

According to the structure they generate, computational methods for inferring the 3D chromatin structure from a contact matrix fall into two classes: ensemble and consensus methods. In a haploid setting (organisms having a single set of chromosomes), ensemble models such as MCMC5C (Rousseau et al. 2011), BACH-MIX (Hu et al. 2013) and Chrom3D (Paulsen et al. 2017), try to account for structure variations on the genome across cells by inferring a population of 3D structures. On the other hand, consensus methods aim at reconstructing one single 3D structure which may be used as a model for further analysis. In this category, probability-based methods such as PASTIS (Varoquaux et al. 2014; Cauer et al. 2019) and ASHIC (Ye and Ma 2020) model contact counts as Poisson random variables of the Euclidean distances between loci, and distance-based methods such as ChromSDE (Zhang et al. 2013) and ShRec3D (Lesne et al. 2014) model contact counts as functions of the Euclidean distances. An extensive overview of different 3D genome reconstruction techniques is given in Oluwadare et al. (2019).

Most of the methods for 3D genome reconstructions from Hi-C data are for haploid organisms. However, like most mammals, humans are diploid organisms, in which the genetic information is stored in pairs of chromosomes called homologs. Homologous chromosomes are almost identical besides some single nucleotide polymorphisms
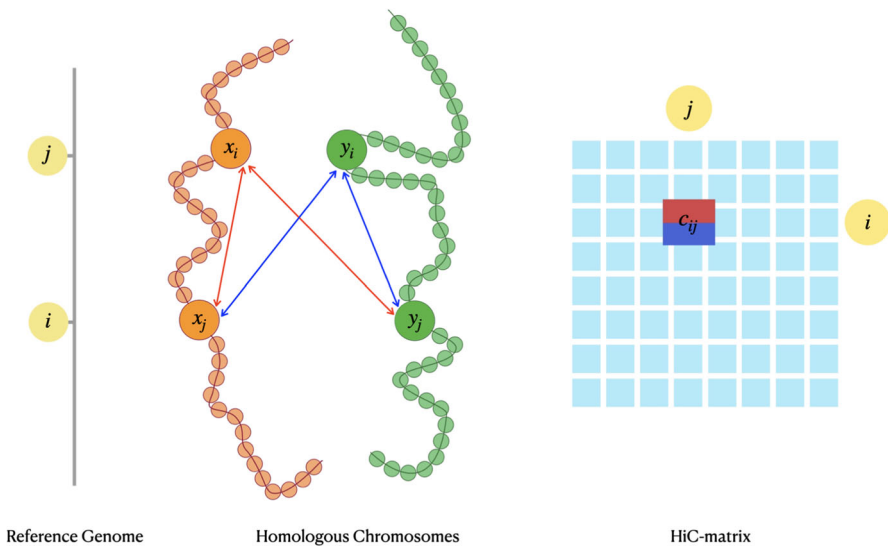
Fig. 1 Ambiguity of phased data. Each entry $c_{i,j}$ of the Hi-C matrix corresponds to four different contacts between the two pairs (x_i, y_i) for locus i and (x_j, y_j) for locus j

(SNPs) (Li et al. 2021). In the case of diploid organisms, the Hi-C data does not generally differentiate between homologous chromosomes. If we model each chromosome as a string of beads, then we associate two beads to each locus $i \in \{1, \ldots, n\}$, one bead for each homolog. Therefore, each observed contact count $c_{i,j}$ between loci i and j represents aggregated contacts of four different types of interactions, more precisely one of the two homologous beads associated to locus i gets in contact with one of the two homologous beads associated to locus j, see Fig. 1. This means that the Hi-C data is *unphased*. *Phased* Hi-C data that distinguishes contacts for homologs is rare. In our setting, we assume that the data is *partially phased*, i.e., some of the contact counts can be associated with a homolog. For example, in the (mouse) Patski (BL6xSpretus) (Deng et al. 2015; Ye and Ma 2020) cell line, 35.6% of the contact counts are phased; while this value is as low as 0.14% in the human GM12878 cell line (Rao et al. 2014; Ye and Ma 2020). Therefore, methods for inferring diploid 3D chromatin structure need to take into account the ambiguity of diploid Hi-C data to avoid inaccurate reconstructions.

Methods for 3D genome reconstruction in diploid organisms have been studied in Tan et al. (2018); Ye and Ma (2020); Cauer et al. (2019); Luo et al. (2020); Belyaeva et al. (2022); Lindsly et al. (2021); Segal (2022). One approach is to phase Hi-C data (Tan et al. 2018; Luo et al. 2020; Lindsly et al. 2021), for example by assigning haplotypes to contacts based on assignments at neighboring contacts (Tan et al. 2018; Lindsly et al. 2021). Cauer et al. (2019) and Ye and Ma (2020) model contact counts as Poisson random variables. To find the optimal 3D chromatin structure, Cauer et al. maximize the associated likelihood function combined with two structural constraints. The first constraint imposes that the distances between neighboring beads are similar, and the second one requires that homologous chromosomes are located in different

regions of the cell nucleus. On the other hand, Ye and Ma first compute the maximum likelihood estimate of model parameters for each of the homologs separately; these estimates are then refined by estimating the distance between the homologs. Belyaeva et al. (2022) show identifiability of the 3D structure when the Euclidean distances between neighboring beads and higher-order contact counts between three or more loci simultaneously are given. Under these assumptions, the 3D reconstruction is obtained by combining distance geometry with semidefinite programming. Segal (2022) applies recently developed imaging technology, in situ genome sequencing (IGS) (Payne et al. 2021), to point out issues in the assumptions made in Tan et al. (2018); Cauer et al. (2019); Belyaeva et al. (2022), and suggests as alternative assumptions that intra-homolog distances are smaller than corresponding inter-homolog distances and intra-homolog distances are similar for homologous chromosomes. IGS (Payne et al. 2021) provides yet another method for inferring the 3D structure of the genome, however, at present the resolution and availability of IGS data is limited.

Contributions In this work, we focus on a distance-based approach for partially phased Hi-C data. In particular, we assume that contacts only for some loci are phased. In the string of beads model, the locations of the pair of beads associated to i-th loci are denoted by $x_i, y_i \in \mathbb{R}^3$. Then homologs are represented by two sequences $x_1, x_2, \ldots, x_n$ and $y_1, x_2, \ldots, y_n$ in $\mathbb{R}^3$; see Fig. 1. Inferring the 3D chromatin structure corresponds to estimating the bead coordinates. Based on Lieberman-Aiden et al. (2009), we assume the power law dependency $c_{i,j} = \gamma d_{i,j}^\alpha$, where α is a negative conversion factor, between the distance $d_{i,j}$ and contact count $c_{i,j}$ of loci i and j. Following Cauer et al. (2019), we assume that a contact count between loci is given by the sum of all possible contact counts between the corresponding beads. We call a bead unambiguous if the contacts for the corresponding locus are phased; otherwise, we call a bead ambiguous.

Our first main contribution is to show that for negative rational conversion factors α, knowing the locations of six unambiguous beads ensures that there are generically finitely many possible locations for the other beads, both in the noiseless (Theorem 1) and noisy (Corollary 1) setting. Moreover, we prove finite identifiability also in the fully ambiguous setting when $\alpha = -2$ and the number of loci is at least 12 (Theorem 2). Note that the identifiability does not hold for $\alpha = 2$ as shown in Belyaeva et al. (2022).

Our second main contribution is to provide a reconstruction method when $\alpha = -2$, based on semidefinite programming combined with numerical algebraic geometry and local optimization (Sect. 4). The general idea is the following: We first estimate the coordinates of the unambiguous beads using only the unambiguous contact counts (which precisely corresponds to the haploid setting) using the SDP-based solver implemented in ChromSDE (Zhang et al. 2013). We then exploit our theoretical result on finite identifiability to estimate the coordinates of the ambiguous beads, one by one, by solving several polynomial systems numerically. These estimates are then improved by a local estimation step considering all contact counts. Finally, a clustering algorithm is used to overcome the symmetry $(x_i, y_i) \mapsto (y_i, x_i)$ in the estimation for the ambiguous beads.

The paper is organized as follows. In Sect. 2, we introduce our mathematical model for the 3D genome reconstruction problem. In Sect. 3, we recall identifiability results in the unambiguous setting (Sect. 3.1) and then prove identifiability results in the

partially ambiguous setting (Sect. 3.2) and in the fully ambiguous setting (Sect. 3.3). We describe our reconstruction method in Sect. 4. We test the performance of our method on synthetic datasets and on a real dataset from the mouse X chromosomes in Sect. 5. We conclude with a discussion about future research directions in Sect. 6.

2 Mathematical Model for 3D Genome Reconstruction

In this section we introduce the distance-based model under which we study 3D genome reconstruction. In Sect. 2.1 we give the background on contact count matrices. In Sect. 2.2 we describe a power-law between contacts and distances, which allows to translate the information about contacts into distances.

2.1 Contact Count Matrices

We model the genome as a string of $2n$ beads, corresponding to n pairs of homologous beads. The positions of the beads are recorded by a matrix

$$Z = [x_1, \ldots, x_n, y_1, \ldots, y_n]^T \in \mathbb{R}^{2n \times 3}.$$

The positions x_i and y_i correspond to homologous beads. When convenient, we use the notation $z_1 := x_1, \ldots, z_n := x_n, z_{n+1} := y_1, \ldots, z_{2n} := y_n$. In this notation,

$$Z = [z_1, \ldots, z_n, z_{n+1}, \ldots, z_{2n}]^T \in \mathbb{R}^{2n \times 3}.$$

Let U be the subset of pairs that are unambiguous, i.e., beads in the pair can be distinguished, and let A be the subset of pairs that are ambiguous, i.e., beads in the pair cannot be distinguished. The sets U and A form a partition of $[n]$.

A Hi-C matrix C is a matrix with each row and column corresponding to a genomic locus. Following Cauer et al. (2019), we call these contact counts ambiguous and denote the corresponding contact count matrix by C^A. If parental genotypes are available, then one can use SNPs to map some reads to each haplotype (Deng et al. 2015; Minajigi et al. 2015; Rao et al. 2014). If both ends of a read contains SNPs that can be associated to a single parent, then the contact count is called unambiguous and the corresponding contact count matrix is denoted by C^U. Finally, if only one of the genomic loci present in an interaction can be mapped to one of the homologous chromosomes, then the count is called partially ambiguous and the contact count matrix is denoted by C^P.

The unambiguous count matrix C^U is a $2n \times 2n$ matrix with the first n indices corresponding to $x_1, \ldots, x_n$ and the last n indices corresponding to $y_1, \ldots, y_n$. The ambiguous count matrix C^A is an $n \times n$ matrix and we assume that each ambiguous count is the sum of four unambiguous counts:

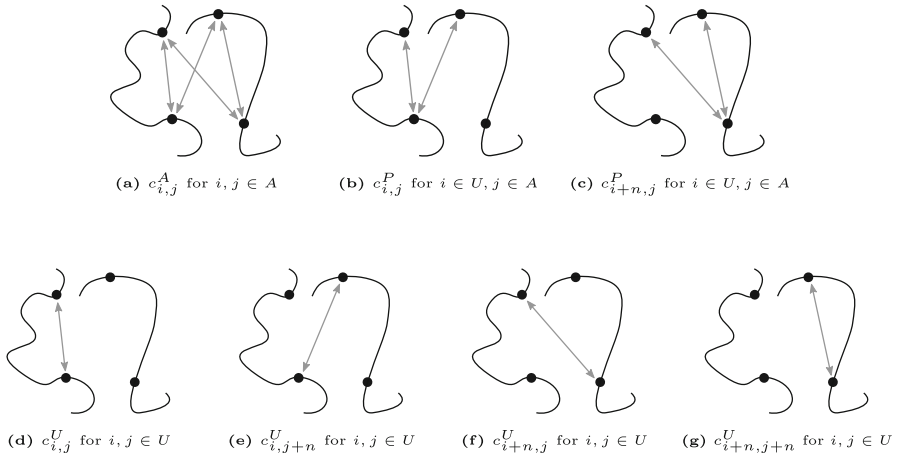$$c_{i,j}^A = c_{i,j}^U + c_{i,j+n}^U + c_{i+n,j}^U + c_{i+n,j+n}^U.$$

(a) $c_{i,j}^A$ for $i, j \in A$ (b) $c_{i,j}^P$ for $i \in U, j \in A$ (c) $c_{i+n,j}^P$ for $i \in U, j \in A$

(d) $c_{i,j}^U$ for $i, j \in U$ (e) $c_{i,j+n}^U$ for $i, j \in U$ (f) $c_{i+n,j}^U$ for $i, j \in U$ (g) $c_{i+n,j+n}^U$ for $i, j \in U$

Fig. 2 Seven different types of contacts between the ith and jth locus

The partially ambiguous count matrix C^P is a $2n \times n$ matrix and each partially ambiguous count is the sum of two unambiguous counts:

$$c_{i,j}^P = c_{i,j}^U + c_{i,j+n}^U.$$

2.2 Contacts and Distances

Denoting the distance $\|z_i - z_j\|$ between z_i and z_j by $d_{i,j}$, the power law dependency observed by Lieberman-Aiden et al. (2009) can be written as

$$c_{i,j}^U = \gamma d_{i,j}^\alpha, \tag{1}$$

where $\alpha < 0$ is a conversion factor and $\gamma > 0$ is a scaling factor. This relationship between contact counts and distances is assumed in Belyaeva et al. (2022); Zhang et al. (2013), while in Cauer et al. (2019); Varoquaux et al. (2014) the contact counts $c_{i,j}$ are modeled as Poisson random variables with the Poisson parameter being $\beta d_{i,j}^\alpha$.

In our paper, we assume that contact counts are related to distances by (1). Similarly to Belyaeva et al. (2022), we set $\gamma = 1$ and in parts of the article $\alpha = -2$. In general, the conversion factor α depends on a dataset and its estimation can be part of the reconstruction problem (Varoquaux et al. 2014; Zhang et al. 2013). Setting $\gamma = 1$ means that we recover the configuration up to a scaling factor. In practice, the configuration can be rescaled using biological knowledge, e.g., the radius of the nucleus.

Our approach to 3D genome reconstruction builds on the power law dependency between contacts and distances between unambiguous beads. We convert the empirical

contact counts to Euclidean distances and then aim to reconstruct the positions of beads from the distances. This leads us to the following system of equations:

$$
\begin{cases}
c_{i,j}^A = \|x_i - x_j\|^\alpha + \|x_i - y_j\|^\alpha + \|y_i - x_j\|^\alpha + \|y_i - y_j\|^\alpha & \forall i, j \in A \\
c_{i,j}^P = \|x_i - x_j\|^\alpha + \|x_i - y_j\|^\alpha, \ c_{i+n,j}^P = \|y_i - x_j\|^\alpha + \|y_i - y_j\|^\alpha \ \forall i \in U, j \in A \\
c_{i,j}^U = \|x_i - x_j\|^\alpha, \ c_{i,j+n}^U = \|x_i - y_j\|^\alpha, \\
c_{i+n,j}^U = \|y_i - x_j\|^\alpha, \ c_{i+n,j+n}^U = \|y_i - y_j\|^\alpha & \forall i, j \in U
\end{cases}
\tag{2}
$$

If α is an even integer, then (2) is a system of rational equations.

Determining the points x_i, y_i, where $i \in U$, is the classical Euclidean distance problem: We know the (noisy) pairwise distances between points and would like to construct the locations of points, see Sect. 3.1 for details. Hence after Sect. 3.1 we assume that we have estimated the locations of points x_i, y_i, where $i \in U$, and we would like to determine the points x_i, y_i, where $i \in A$.

3 Identifiability

In this section, we study the uniqueness of the solutions of the system (2) up to rigid transformations (translations, rotations and reflections), or in other words, the identifiability of the locations of beads. We study the unambiguous, partially ambiguous and ambiguous settings in Sects. 3.1, 3.2 and 3.3, respectively.

3.1 Unambiguous Setting and Euclidean Distance Geometry

If all pairs are unambiguous, i.e., $U = [n]$, then constructing the original points translates to a classical problem in Euclidean distance geometry. The principal task in Euclidean distance geometry is to construct original points from pairwise distances between them. In the rest of the subsection, we will recall how to solve this problem. Since pairwise distances are invariant under translations, rotations and reflections (rigid transformations), then the original points can be reconstructed up to rigid transformations. For an overview of distance geometry and Euclidean distance matrices, we refer the reader to Dokmanic et al. (2015), Krislock and Wolkowicz (2012), Liberti et al. (2014) and Mucherino et al. (2012).

The Gram matrix of the points $z_1, \ldots, z_{2n}$ is defined as

$$
G = ZZ^T = [z_1, \ldots, z_{2n}]^T \cdot [z_1, \ldots, z_{2n}] \in \mathbb{R}^{2n \times 2n}.
$$

Let $\overline{z} = \frac{1}{2n} \sum_{i=1}^{2n} z_i$ and $\tilde{z}_i = z_i - \overline{z}$ for $i = 1, \ldots, 2n$. The matrix $\tilde{Z} = [\tilde{z}_1, \ldots, \tilde{z}_{2n}]^T$ gives the locations of points after centering them around the origin. Let $\tilde{G}$ denote the Gram matrix of the centered point configuration $\tilde{z}_1, \ldots, \tilde{z}_{2n}$.

Let $D_{i,j} = \|z_i - z_j\|^2$ denote the squared Euclidean distance between the points z_i and z_j. The Euclidean distance matrix of the points $z_1, \ldots, z_{2n}$ is defined as $D = (D_{i,j})_{1 \le i,j \le 2n} \in \mathbb{R}^{2n \times 2n}$. To express the centered Gram matrix in terms of the Euclidean distance matrix, we define the geometric centering matrix

$$J = I_{2n} - \frac{1}{2n}\mathbf{1}\mathbf{1}^T,$$

where I_{2n} is the $2n \times 2n$ identity matrix and $\mathbf{1}$ is the vector of ones. The linear relationship between $\tilde{G}$ and D is given by

$$\tilde{G} = -\frac{1}{2}JDJ.$$

Therefore, given the Euclidean distance matrix, we can construct the centered Gram matrix for the points $z_1, \ldots, z_{2n}$.

The centered points up to rigid transformations are extracted from the centered Gram matrix $\tilde{G}$ using the eigendecomposition $\tilde{G} = Q\Lambda Q^{-1}$, where Q is orthonormal and Λ is a diagonal matrix with entries ordered in decreasing order $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_{2n} \ge 0$. We define $\Lambda_3^{1/2} := [\mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \sqrt{\lambda_3}), \mathbf{0}_{3 \times (2n-3)}]^T$ and set $\hat{Z} = Q\Lambda_3^{1/2}$. In the case of noiseless distance matrix D, the Gram matrix $\tilde{G}$ has rank three and the diagonal matrix Λ has precisely three non-zero entries. Hence we could obtain $\hat{Z}$ also from $Q\Lambda^{1/2}$ by truncating zero columns. Using $\Lambda_3^{1/2}$ has the advantage that it gives an approximation for the points also for a noisy distance matrix D. The uniqueness of $\hat{Z}$ up to rotations and reflections follows from Krislock ([2010](), Proposition 3.2) which states that $AA^T = BB^T$ if and only if $A = BQ$ for some orthogonal matrix Q.

The procedure that transforms the distance matrix to origin centered Gram matrix and then uses eigendecomposition for constructing original points is called classical multidimensional scaling (cMDS) (Cox and Cox [2008]()). Although cMDS is widely used in practice, it does not always find the distance matrix that minimizes the Frobenius norm to the empirical noisy distance matrix (Sonthalia et al. [2021]()). Other approaches to solving the Euclidean distance and Euclidean completion problems include non-convex (Fang and O'Leary [2012](); Mishra et al. [2011]()) as well semidefinite formulations (Alfakih et al. [1999](); Fazel et al. [2003](); Nie [2009](); Weinberger et al. [2007](); Zhang et al. [2013](); Zhou et al. [2020]()).

3.2 Partially Ambiguous Setting

The next theorem establishes the uniqueness of the solutions of the system ([2]()) in the presence of ambiguous pairs. In particular, it states that there are finitely many possible locations for beads in one ambiguous pair given the locations of six unambiguous beads. The identifiability results in this subsection hold for all negative rational numbers α. In the rest of the paper, we denote the true but unknown coordinates by x^* and the symbol x stands for a variable that we want to solve for. We write $\|\cdot\|$ for the standard inner product on $\mathbb{R}^3$.

Theorem 1 *Let α be a negative rational number. Then for $a^*, b^*, \ldots, f^*, x^*, y^* \in \mathbb{R}^3$ sufficiently general, the system of six equations*

$$\|x - t^*\|^\alpha + \|y - t^*\|^\alpha = \|x^* - t^*\|^\alpha + \|y^* - t^*\|^\alpha \text{ for } t^* = a^*, b^*, \ldots, f^* \quad (3)$$

in the six unknowns $x_1, x_2, x_3, y_1, y_2, y_3 \in \mathbb{R}$ has only finitely many solutions.

Remark 1 The proof will show that this system has only finitely many solutions over the *complex* numbers.

We believe that the theorem holds for general nonzero rational α. Indeed, our argument works, with a minor modification, also for $\alpha > 2$, but for α in the range $(0, 2]$ a refinement of the argument is needed.

Proof First write $Q(x) := x_1^2 + x_2^2 + x_3^2$, so that $\|x\| = \sqrt{Q(x)}$ for $x \in \mathbb{R}^3$. The advantage of Q over $\|x\|$ is that it is well-defined on $\mathbb{C}^3$.

Write $\frac{\alpha}{2} = \frac{m}{n}$ with m, n relatively prime integers, $m \neq 0$, and $n > 0$. Consider the affine variety $X \subseteq (\mathbb{C}^3)^8 \times (\mathbb{C}^2)^6$ consisting of all tuples

$$((a^*, \ldots, f^*, x^*, y^*), (r_{t^*}, s_{t^*})_{t^* = a^*, \ldots, f^*})$$

such that

$$Q(x^* - t^*)^m = r_{t^*}^n \neq 0 \text{ and } Q(y^* - t^*)^m = s_{t^*}^n \neq 0 \text{ for } t^* = a^*, \ldots, f^*.$$

Note that, if x^*, t^* are real, then it follows that

$$Q(x^* - t^*)^m = (\|x^* - t^*\|^\alpha)^n,$$

and similarly for $Q(y^* - t^*)$. Hence if $a^*, \ldots, y^*$ are all real, then the point

$$((a^*, \ldots, f^*, x^*, y^*), (\|x^* - t^*\|^\alpha, \|y^* - t^*\|^\alpha)_{t^*}) \quad (4)$$

is a point in X with real-valued coordinates.

The projection π from X to the open affine subset $U \subseteq (\mathbb{C}^3)^8$ where all $Q(x^* - t^*)$ and $Q(y^* - t^*)$ are nonzero is a finite morphism with fibers of cardinality n^{12}; to see this cardinality note that there are n possible choices for each of the numbers r_{t^*}, s_{t^*}. Each irreducible component of X is a smooth variety of dimension 24.

Consider the map $\psi : X \to (\mathbb{C}^3 \times \mathbb{C}^1)^6$ defined by
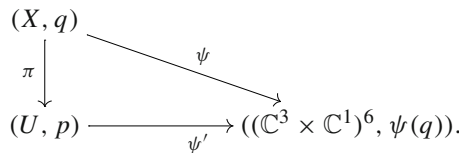
$$((a^*, \ldots, f^*, x^*, y^*), (r_{t^*}, s_{t^*})_{t^*}) \mapsto ((t^*, r_{t^*} + s_{t^*}))_{t^*}$$

We claim that for q in some open dense subset of X, the derivative $d_q \psi$ has full rank 24. For this, it suffices to find one point $p \in U$ such that $d_q \psi$ has rank 24 at each of the n^{12} points $q \in \pi^{-1}(p)$. We take a real-valued point $p := (a^*, b^*, \ldots, f^*, x^*, y^*) \in (\mathbb{R}^3)^8$ to be specified later on. Let $q \in \pi^{-1}(p)$. Then, near q, the map ψ factorises

via π and the unique algebraic map $\psi' : U \to (\mathbb{C}^3 \times \mathbb{C}^1)^6$ (defined near p) which on a neighborhood of p in $U \cap (\mathbb{R}^3)^8$ equals

$$\psi'(a, \ldots, f, x, y) = ((t, \xi_{t*} \cdot Q(x - t)^{\alpha/2} + \eta_{t*} \cdot Q(y - t)^{\alpha/2}))_{t=a,\ldots,f} \in (\mathbb{C}^3 \times \mathbb{C}^1)^6$$

where ξ_{t*} and ζ_{t*} are n-th roots of unity in $\mathbb{C}$ depending on which q is chosen among the n^{12} points in $\pi^{-1}(p)$. The situation is summarised in the following diagram:

$$\begin{array}{ccc} (X, q) & & \\ \pi \downarrow & \searrow^{\psi} & \\ (U, p) & \xrightarrow{\psi'} & ((\mathbb{C}^3 \times \mathbb{C}^1)^6, \psi(q)). \end{array}$$

Now, $d_q\psi = d_p\psi' \circ d_q\pi$, and since $d_q\pi$ is a linear isomorphism, it suffices to prove that $d_p\psi'$ is a linear isomorphism. Suppose that $(a', \ldots, f', x', y') \in \ker d_p\psi'$. Then, since the map ψ' remembers $a, \ldots, f$, it follows immediately that $a' = \ldots = f' = 0$. On the other hand, by differentiating we find that, for each $t^* \in \{a^*, \ldots, f^*\}$,

$$\xi_{t*} \cdot (\alpha/2) \cdot Q(x^* - t^*)^{\alpha/2-1} \cdot 2 \cdot \langle x', x^* - t^* \rangle$$
$$+\eta_{t*} \cdot (\alpha/2) \cdot Q(y^* - t^*)^{\alpha/2-1} \cdot 2 \cdot \langle y', y^* - t^* \rangle = 0,$$

where $\langle \cdot, \cdot \rangle$ stands for the standard bilinear form on $\mathbb{C}^3$. In other words, the vector $(x', y') \in \mathbb{C}^6$ is in the kernel of the 6×6-matrix

$$M := \begin{bmatrix} \|x^* - a^*\|^{\alpha-2} \cdot \xi_{a*} \cdot (x^* - a^*) & \|y^* - a^*\|^{\alpha-2} \cdot \eta_{a*} \cdot (y^* - a^*) \\ \vdots & \vdots \\ \|x^* - f^*\|^{\alpha-2} \cdot \xi_{f*} \cdot (x^* - f^*) & \|y^* - f^*\|^{\alpha-2} \cdot \eta_{f*} \cdot (y^* - f^*) \end{bmatrix}$$

where we have interpreted $a^*, \ldots, f^*, x^*, y^*$ as row vectors. It suffices to show that, for some specific choice of $p = (a^*, \ldots, f^*, x^*, y^*) \in (\mathbb{R}^3)^8$, this matrix is nonsingular *for all n^{12} choices of* $((\xi_{t*}, \eta_{t*}))_{t*}$.

We choose $a^*, \ldots, f^*, x^*, y^*$ as the vertices of the unit cube, as follows:

$$\begin{array}{lll} a^* = (1, 0, 0) & b^* = (0, 1, 0) & c^* = (0, 0, 1) \\ c^* = (0, 1, 1) & d^* = (1, 0, 1) & f^* = (1, 1, 0) \\ x^* = (0, 0, 0) & y^* = (1, 1, 1). \end{array}$$

Then the matrix M becomes, with $\beta = \alpha - 2$:

$$
\begin{bmatrix}
-\xi_{a*} & 0 & 0 & 0 & 2^{\frac{\beta}{2}} \cdot \eta_{a*} & 2^{\frac{\beta}{2}} \cdot \eta_{a*} \\
0 & -\xi_{b*} & 0 & 2^{\frac{\beta}{2}} \cdot \eta_{b*} & 0 & 2^{\frac{\beta}{2}} \cdot \eta_{b*} \\
0 & 0 & -\xi_{c*} & 2^{\frac{\beta}{2}} \cdot \eta_{c*} & 2^{\frac{\beta}{2}} \cdot \eta_{c*} & 0 \\
0 & -(2^{\frac{\beta}{2}} \cdot \xi_{d*}) & -(2^{\frac{\beta}{2}} \cdot \xi_{d*}) & \eta_{d*} & 0 & 0 \\
-(2^{\frac{\beta}{2}} \cdot \xi_{e*}) & 0 & -(2^{\frac{\beta}{2}} \cdot \xi_{e*}) & 0 & \eta_{e*} & 0 \\
-(2^{\frac{\beta}{2}} \cdot \xi_{f*}) & -(2^{\frac{\beta}{2}} \cdot \xi_{f*}) & 0 & 0 & 0 & \eta_{f*}
\end{bmatrix}.
$$

Now, $\det(M)$ equals

$$
-\xi_{a*} \cdot \xi_{b*} \cdot \xi_{c*} \cdot \eta_{d*} \cdot \eta_{e*} \cdot \eta_{f*} + 2^{2+3\beta} \cdot \eta_{a*} \cdot \eta_{b*} \cdot \eta_{c*} \cdot \xi_{d*} \cdot \xi_{e*} \cdot \xi_{f*} + 2^{2\beta} \cdot R
\tag{5}
$$

where R is a sum of (products of) roots of unity. Now $\alpha < 0$ implies that $\beta < -2$, so that $2 + 3\beta < 2\beta < 0$. Since roots of unity have 2-adic valuation 0, the second term in the expression above is the unique term with minimal 2-adic valuation. Hence $\det(M) \neq 0$, as desired.

It follows that ψ is a dominant morphism from each irreducible component of X into $(\mathbb{C}^3 \times \mathbb{C}^1)^6$, and hence for all q in an open dense subset of X, the fiber $\psi^{-1}(\psi(q))$ is finite. This then holds, in particular, for q in an open dense subset of the real points as in (4). This proves the theorem. □

Remark 2 If $\alpha > 2$, then $\beta > 0$, and hence the unique term with minimal 2-adic valuation in (5) is the first term. This can be used to show that the theorem holds then, as well. The only subtlety is that for positive α, solutions where x or y equal one of the points $a*, \ldots, f*$ are not automatically excluded, and these are not seen by the variety X. But a straightforward argument shows that such solutions do not exist for sufficiently general choices of $a*, \ldots, f*, x*, y*$.

We now consider the setting when we know locations of seven unambiguous beads. In the special case when $\alpha = -2$, we construct the ideal generated by the polynomials obtained from rational Eqs. (3) for seven unambiguous beads after moving all terms to one side and clearing the denominators. Based on symbolic computations in `Macaulay2` for the degree of this ideal, we conjecture that the location of a seventh unambiguous bead guarantees unique identifiability of an ambiguous pair of beads:

Conjecture 1 *Let $a*, b*, c*, d*, e*, f*, g*, x*, y* \in \mathbb{R}^3$ be sufficiently general. The system of rational equations*

$$
\frac{1}{\|t* - x*\|^2} + \frac{1}{\|t* - y*\|^2} = \frac{1}{\|t* - x\|^2} + \frac{1}{\|t* - y\|^2} \text{ for } t* = a*, b*, c*, d*, e*, f*, g*
\tag{6}
$$

has precisely two solutions $(x, y*)$ and $(y*, x*)$.*

In practice, we only have noisy estimates $a, b, \ldots, f \in \mathbb{R}^3$ of the true positions of unambiguous beads $a^*, b^*, \ldots, f^* \in \mathbb{R}^3$, and we have noisy observations c_t of the true contact counts $c_t^* := \|x^* - t^*\|^\alpha + \|y^* - t^*\|^\alpha$. We aim to find $x, y \in \mathbb{R}^3$ such that

$$\|x - t\|^\alpha + \|y - t\|^\alpha = c_t \text{ for } t = a, b, \ldots, f.$$

We may write $c_t = \|x^* - t\|^\alpha + \|y^* - t\|^\alpha + \epsilon_t$ for some ϵ_t that depends on the noise level. Hence, the above system of equations can be rephrased as

$$\|x - t\|^\alpha + \|y - t\|^\alpha = \|x^* - t\|^\alpha + \|y^* - t\|^\alpha + \epsilon_t \text{ for } t = a, b, \ldots, f. \quad (7)$$

In the following corollary we show that this system has generically finitely many solutions.

Corollary 1 *Let α be a negative rational number. Then for $a, b, \ldots, f, x^*, y^* \in \mathbb{R}^3$ and $\epsilon_a, \epsilon_b, \ldots, \epsilon_f \in \mathbb{R}$ sufficiently general, the system of six equations*

$$\|x - t\|^\alpha + \|y - t\|^\alpha = \|x^* - t\|^\alpha + \|y^* - t\|^\alpha + \epsilon_t \text{ for } t = a, b, \ldots, f \quad (8)$$

in the six unknowns $x_1, x_2, x_3, y_1, y_2, y_3 \in \mathbb{R}$ has only finitely many solutions.

Proof Recall the map $\psi : X \rightarrow (\mathbb{C}^3 \times \mathbb{C}^1)^6$ from the proof of Theorem 1 defined by

$$((a, \ldots, f, x^*, y^*), (r_{x^*,t}, s_{y^*,t})_t) \mapsto ((t, r_{x^*,t} + s_{y^*,t})_t).$$

We showed that ψ is a dominant morphism from each irreducible component of X into $(\mathbb{C}^3 \times \mathbb{C}^1)^6$, and that each irreducible component of X is 24-dimensional. Every solution to (8) is the (x, y)-component of a point in the fiber

$$\psi^{-1}((t, \|x^* - t\|^\alpha + \|y^* - t\|^\alpha + \epsilon_t))_t.$$

Since this is a fiber over a sufficiently general point, the fiber is finite. $\qquad\qquad\square$

Corollary 1 will be the basis of a numerical algebraic geometric based reconstruction method in Sect. 4.

3.3 Ambiguous Setting

Finally we consider the ambiguous setting, where one would like to reconstruct the locations of beads only from ambiguous contact counts. It is shown in Belyaeva et al. (2022) that for $\alpha = 2$, one does not have finite identifiability no matter how many pairs of ambiguous beads one considers. We show finite identifiability for the locations of beads given contact counts for 12 pairs of ambiguous beads for $\alpha = -2$ in both the noisy and noiseless setting. We believe that the result might be true for further conversion factors α's, however our proof technique does not directly generalize.

Theorem 2 *Let $\alpha = -2$. Then for $(c_{ij})_{1 \leq i < j \leq 12} \in \mathbb{R}^{66}$ sufficiently general, the system of 66 equations*

$$\|x_i - x_j\|^\alpha + \|x_i - y_j\|^\alpha + \|y_i - x_j\|^\alpha + \|y_i - y_j\|^\alpha = c_{ij} \text{ for } 1 \leq i < j \leq 12 \quad (9)$$

in the 72 unknowns $x_{1,1}, x_{1,2}, x_{1,3}, y_{1,1}, y_{1,2}, y_{1,3}, \ldots, x_{12,1}, x_{12,2}, x_{12,3}, y_{12,1}, y_{12,2}, y_{12,3} \in \mathbb{R}$ has only finitely many solutions up to rigid transformations. In particular, it holds that for sufficiently general $(x_1^*, y_1^*, \ldots, x_{12}^*, y_{12}^*) \in (\mathbb{R}^3)^{24}$, the system

$$\begin{aligned}
&\|x_i - x_j\|^\alpha + \|x_i - y_j\|^\alpha + \|y_i - x_j\|^\alpha + \|y_i - y_j\|^\alpha = \\
&\|x_i^* - x_j^*\|^\alpha + \|x_i^* - y_j^*\|^\alpha + \|y_i^* - x_j^*\|^\alpha + \|y_i^* - y_j^*\|^\alpha \text{ for } 1 \leq i < j \leq 12
\end{aligned}$$

$$(10)$$

has finitely many solutions up to rigid transformation.

Proof As before, we write $Q(x) := x_1^2 + x_2^2 + x_3^2$, so that $\|x\| = \sqrt{Q(x)}$ for $x \in \mathbb{R}^3$. Consider the affine open subset $X \subseteq (\mathbb{C}^3)^{24}$ consisting of all tuples $(x_1^*, y_1^*, \ldots, x_{12}^*, y_{12}^*)$ such that

$$Q(x_i^* - x_j^*) \neq 0, \quad Q(x_i^* - y_j^*) \neq 0, \quad Q(y_i^* - x_j^*) \neq 0 \text{ and } Q(y_i^* - y_j^*) \neq 0 \text{ for } i < j.$$

Consider also the map $\psi \colon X \to \mathbb{C}^{66}$ defined by

$$(x_1^*, \ldots, y_{12}^*) \mapsto \left(Q(x_i^* - x_j^*)^{-1} + Q(x_i^* - y_j^*)^{-1} + Q(y_i^* - x_j^*)^{-1} + Q(y_i^* - y_j^*)^{-1} \right)_{i<j}.$$

By a computer calculation (with exact arithmetic) we found that at a randomly chosen $q \in X$ with rational coordinates, the derivative $d_q\psi$ had full rank 66. It then follows that for q in some open dense subset of X, $d_q\psi$ has rank 66. Hence ψ is dominant, and for any sufficiently general $c \in \mathbb{C}^{66}$, all irreducible components of the fiber $\psi^{-1}(c)$ have dimension 6. Moreover, each such component C is preserved by the 6-dimensional connected group $G = SO(3, \mathbb{C}) \ltimes \mathbb{C}^3$.

The stabilizer in G of a sufficiently general point in X is zero-dimensional. This follows from a Lie algebra argument: if a point $(x_1^*, y_1^*, \ldots, x_{12}^*, y_{12}^*) \in X$ has a positive-dimensional stabilizer in G, then there is a nonzero element A in the Lie algebra of $SO(3, \mathbb{C})$ that maps all the differences $x_i^* - x_j^*, x_i^* - y_j^*, y_i^* - y_j^*$ to zero. Since A is a skew-symmetric matrix and hence of rank 2, it follows that all points x_i^*, y_j^* lie on a line. The variety of such collinear tuples has dimension 28, so it does not map dominantly to $\mathbb{C}^{66}$. Hence there exists a Zariski open dense subset $V \subseteq \mathbb{C}^{66}$ such that for all $c \in V$, the fiber $\psi^{-1}(c)$ contains no points with positive-dimensional stabilizers in G, and hence $\psi^{-1}(c)$ is a disjoint union of finitely many 6-dimensional G-orbits. Likewise, $\psi^{-1}(V)$ is a Zariski open dense subset of $(\mathbb{C}^3)^{24}$ such that $\psi^{-1}(\psi(q))$ consists of finitely many G-orbits for all $q \in \psi^{-1}(V)$. With this, we have proven the complex analog of the theorem.

To obtain the statement over the real numbers, we note that if $c \in V$ has real-valued coordinates, then a finite number of the G-orbits that make up $\psi^{-1}(c)$ contain a real-valued tuple. If $G \cdot q$ for $q \in (\mathbb{R}^3)^{24}$ is such an orbit, it holds that $(G \cdot q) \cap (\mathbb{R}^3)^{24} = (SO(3, \mathbb{R}) \ltimes \mathbb{R}^3) \cdot q$ whenever the 24 points that make up the tuple q are not coplanar. The set of coplanar configurations form a subset of X of dimension 51, and does therefore not map dominantly to $\mathbb{C}^{66}$. Hence, by shrinking V appropriately, we can assume that no fibers above it contain coplanar configurations. In particular, this means that the real part of the fiber over any real point in V consists of a finitely many orbits under the action of $SO(3, \mathbb{R}) \ltimes \mathbb{R}^3$, as desired.   □

Remark 3 A standard numerical algebraic geometry computation with monodromy and the certification techniques of Breiding et al. (2023), using `Homotopy Continuation.jl` (see, e.g., Sturmfels and Telen (2021)), proves that the system (8) generically has more than 1000 complex solutions up to the action of $O(3, \mathbb{C}) \ltimes \mathbb{C}^3$ and the symmetries $(x_i, y_i) \mapsto (y_i, x_i)$ for $i = 1, \ldots, 12$. This constitutes theoretical motivation for working with partially phased data, even if we, in principle, have finite identifiability already from the unphased data.

Remark 4 When $\alpha = 2$, which corresponds to the setting studied in Belyaeva et al. (2022), then computationally we found that for some special choices of $x_1^*, y_1^*, \ldots, x_{12}^*, y_{12}^* \in \mathbb{R}^3$ the rank of the Jacobian matrix in Theorem 2 is 42. This is consistent with the fact that Theorem 2 fails for $\alpha = 2$ (Belyaeva et al. 2022).

4 A New Reconstruction Method

In this section, we outline a new approach to diploid 3D genome reconstruction for partially phased data, based on the theoretical results discussed in subsection 3.2. The method consists of the following main steps:

1. Estimation of the unambiguous beads $\{x_i, y_i\}_{i \in U}$ through semidefinite programming (discussed in Sect. 4.1).
2. A preliminary estimation of the ambiguous beads using numerical algebraic geometry, based on Corollary 1 (discussed in Sect. 4.2).
3. A refinement of this estimation using local optimization (discussed in Sect. 4.3).
4. A final clustering step, where we disambiguate between the estimations (x_i, y_i) and (y_i, x_i) for each $i \in A$, based on the assumption that homolog chromosomes are separated in space (discussed in Sect. 4.4).

In what follows, we will refer to this method by the acronym SNLC (formed from the initial letters in semidefinite programming, numerical algebraic geometry, local optimization and clustering).

4.1 Estimation of the Positions of Unambiguous Beads

As discussed in Sect. 3.1, the unambiguous bead coordinates $\{x_i, y_i\}_{i \in U} = \{z_i\}_{i \in U \cup (n+U)}$ can be estimated with semidefinite programming. More specifically, we

use ChromSDE Zhang (2013, Section 2.1) for this part of our reconstruction, which relies on a specialized solver from Jiang et al. (2014), to solve an SDP relaxation of the optimization problem

$$\min_{\{z_i\}_{i \in U \cup (n+U)}} \sum_{\substack{i,j \in U \cup (n+U) \\ c_{ij}^U \neq 0}} \sqrt{c_{ij}^U} \left(\frac{1}{c_{ij}^U} - \|z_i - z_j\|^2 \right)^2 + \lambda \sum_{\substack{i,j \in U \cup (n+U) \\ c_{ij}^U = 0}} \|z_i - z_j\|^2$$

(11)

with $\lambda = 0.01$ (cf. Zhang, et al. (2013, Eq. 4)). The terms in the first sum are weighted by the square root for the corresponding contact counts, in order to account for the fact that higher counts can be assumed to be less susceptible to noise.

4.2 Preliminary Estimation Using Numerical Algebraic Geometry

To estimate the coordinates of the ambiguous beads $\{x_i, y_i\}_{i \in A}$, we will use a method based on numerical equation solving, where we estimate the ambiguous bead pairs one by one.

Let x, y be the unknown coordinates in $\mathbb{R}^3$ of a pair of ambiguous beads. We pick six unambiguous beads with already estimated coordinates $a, b, c, d, e, f \in \mathbb{R}^3$. For each $t \in \{a, \ldots, f\}$, let $c_t \in \mathbb{R}$ be the corresponding partially ambiguous counts between t and the ambiguous bead pair (x, y). Clearing the denominators in the system (8), we obtain a system of polynomial equations

$$\|x - t\|^2 + \|y - t\|^2 = c_t \|x - t\|^2 \|y - t\|^2 \text{ for } t = a, b, c, d, e, f. \quad (12)$$

By Corollary 1, this system has finitely many complex solutions both in the noiseless and noisy setting, which can be found using homotopy continuation.

We observe that the system (12) generally has 80 complex solutions, and we only expect one pair of solutions (x, y), (y, x) to correspond to an accurate estimation. Naively adding another polynomial arising from a seventh unambiguous bead (as in Conjecture 1) does not work; in the noisy setting this over-determined system typically lacks solutions. Instead, we compute an estimation based on the following two heuristic assumptions:

1. The most accurate estimation should be *approximately real*, in the sense that the max-norm of the imaginary part is below a certain tolerance (in this work, 0.15 was used for the experiments in both Sects. 5.1 and 5.2). The choice of this threshold was made based on analysing the imaginary parts of solutions to (12) for various choices of unambiguous beads, see Fig. 9.
2. The most accurate estimation should be consistent when we change the choice of six unambiguous beads.

Based on these assumptions, we apply the following strategy. We make a number $N \geq 2$, choices of sets of six unambiguous beads, and solve the corresponding N square systems of the form (12). Since larger contact counts can be expected to have

smaller relative noise, we make the choices of beads among the 20 unambiguous beads t that have highest contact count c_t to the ambiguous locus at hand. For each system, we pick out the approximately real solutions, and obtain N sets $\mathcal{S}_1, \ldots, \mathcal{S}_N \subseteq \mathbb{R}^6$ consisting of the real parts of the approximately real solutions. Up to the symmetry $(x, y) \mapsto (y, x)$, we expect these sets to have a unique "approximately common" element. We therefore compute, by an exhaustive search, the tuple $(w_1, \ldots, w_N) \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$ that minimizes the sum

$$\left\| w_1 - \frac{w_1 + \cdots + w_N}{N} \right\| + \cdots + \left\| w_N - \frac{w_1 + \cdots + w_N}{N} \right\|,$$

and use $\frac{w_1 + \cdots + w_N}{N}$ as our estimation of (x, y). For the computations presented in Sect. 5, we use $N = 5$.

To solve the systems, we use the Julia package $HomotopyContinuation.jl$ (Breiding et al. 2018), and follow the two-phase procedure described in Sommese and Wampler (2005, Sect. 7.2). For the first phase, we solve (12) with randomly chosen parameters $a^*, \ldots, f^* \in \mathbb{C}^3$ and $c_{a^*}, \ldots, c_{f^*} \in \mathbb{C}$, using a polyhedral start system (Huber and Sturmfels 1995). We trace 1280 paths in this first phase, since the Newton polytopes of the polynomials appearing in the system (12) all contain the origin, and have a mixed volume of 1280, which makes 1280 an upper bound on the number of complex solutions by Li (1996, Theorem 2.4). For the second phase, we use a straight-line homotopy in parameter space from the randomly chosen parameters $a^*, \ldots, f^* \in \mathbb{C}^3$ and $c_{a^*}, \ldots, c_{f^*} \in \mathbb{C}$, to the values $a, \ldots, f$ and $c_a, \ldots, c_f \in \mathbb{C}$ at hand. We observe that we generally find 80 complex solutions in the first phase, which means 40 orbits with respect to the symmetry $(x, y) \mapsto (y, x)$. By the discussion in Sommese, (2005, Sect. 7.6) it is enough to only trace one path per orbit, so in the end, we only trace 40 paths in the second phase.

Remark 5 If the noise levels are sufficiently high, there could be choices of six unambiguous beads for which the system lacks approximately-real solutions. If this situation is encountered, we try to redraw the six unambiguous beads until we find an approximately-real solution. If this does not succeed within a certain number of attempts (100 in the experiments conducted for this paper), we use the average of the closest neighboring unambiguous beads instead.

4.3 Local Optimization

A disadvantage of the numerical algebraic geometry based estimation discussed in the previous subsection is that it only takes into account "local" information about the interactions for one ambiguous locus at a time, which might make it more sensitive to noise. In our proposed method, we therefore refine this preliminary estimation of $\{x_i, y_i\}_{i \in A}$ further in a local optimization step that takes into account the "global" information of all available data.

The idea is to estimate $\{x_i, y_i\}_{i \in A}$ by solving the optimization problem

$$\min_{\{x_i, y_i\}_{i \in A}} \sum_{i \in U, j \in A} \left(\left(c_{i,j}^P - \frac{1}{\|x_i - x_j\|^2} - \frac{1}{\|x_i - y_j\|^2} \right)^2 + \left(c_{i+n,j}^P - \frac{1}{\|y_i - x_j\|^2} - \frac{1}{\|y_i - y_j\|^2} \right)^2 \right)$$

(13)

while keeping the estimates of $\{x_i, y_i\}_{i \in U}$ from the ChromSDE step fixed. We use the quasi-Newton method for unconstrained optimization implemented in the Matlab Optimization Toolbox for this step. The already estimated coordinates of $\{x_i, y_i\}_{i \in A}$ from the numerical algebraic geometry step are used for the initialization.

4.4 Clustering to Break Symmetry

Our objective function remains invariant if we exchange x_i and y_i for any $i \in A$. We can break symmetry by relying on the empirical observation that homologous chromosomes typically are spatially separated in different so-called compartments of the nucleus (Eagen 2018). Let $(\bar{x}_i, \bar{y}_i)_{i=1}^n$ denote the estimates from the previous steps. Our final estimations will be obtained by solving the minimization problem

$$\min_{\{x_i, y_i\}_{i \in A}} \sum_{i=1}^{n-1} g_{i,i+1}(x, y), \text{ with } g_{i,i+1}(x, y) := \left(\|x_i - x_{i+1}\|^2 + \|y_i - y_{i+1}\|^2 \right),$$

(14)

where $(x_i, y_i) = (\bar{x}_i, \bar{y}_i)$ for $i \in U$ are fixed, and $(x_i, y_i) \in \{(\bar{x}_i, \bar{y}_i), (\bar{y}_i, \bar{x}_i)\}$ for $i \in A$ are the optimization variables. The optimal solution can be computed efficiently, as explained next.

We first decompose the problem into contiguous chunks of ambiguous beads. Let $(i_1, \ldots, i_L) := U$ be the indices of the unambiguous beads and let $i_0 := 1, i_{L+1} := n$. The optimization problem can be phrased as

$$\min_{\{x_i, y_i\}_{i \in A}} \sum_{\ell=0}^{L} G_\ell(x, y), \quad \text{with} \quad G_\ell(x, y) := \sum_{i=i_\ell}^{i_{\ell+1}-1} g_{i,i+1}(x, y)$$

(15)

where there is one summand $G_\ell(x, y)$ for each contiguous chunk of ambiguous beads. Since the summands $G_\ell(x, y)$ do not share any ambiguous bead, we can minimize them independently.

We proceed to describe the optimal solution of the problem. Let

$$s_i = \begin{cases} 1, & \text{if } (x_i, y_i) = (\bar{x}_i, \bar{y}_i) \\ -1, & \text{if } (x_i, y_i) = (\bar{y}_i, \bar{x}_i) \end{cases}, \quad w_{i,i+1} = (\bar{x}_i - \bar{y}_i)^T (\bar{x}_{i+1} - \bar{y}_{i+1}).$$

The variable s_i indicates whether we keep using $(\bar{x}_i, \bar{y}_i)$ or we reverse it. Note that $s_i = 1$ for $i \in U$. The next lemma gives the optimal assignment of s_i for $i \in A$. This assignment is constructed by using inner products $w_{i,i+1}$.

Lemma 1 *The optimal solution of (14) can be constructed as follows:*

1. *For the last chunk ($\ell = L$) we have*

$$s^*_{i_\ell} = 1, \qquad s^*_{i+1} = \text{sgn}(w_{i,i+1})s^*_i \quad for \; i = i_\ell, i_\ell+1, \ldots, i_{\ell+1}-1$$

   where $\text{sgn}(\cdot)$ *is the sign function and* $\text{sgn}(0)$ *can be either 1 or* -1.
2. *For the first chunk ($\ell = 0$) we have*

$$s^*_{i_{\ell+1}} = 1, \qquad s^*_i = \text{sgn}(w_{i,i+1})s^*_{i+1} \quad for \; i = i_{\ell+1}-1, i_{\ell+1}-2, \ldots, i_\ell$$

3. *For any other chunk, let k be the index of the smallest absolute value* $|w_{k,k+1}|$, *among* $i_\ell \leq k \leq i_{\ell+1} - 1$. *The solution is*

$$s^*_{i_\ell} = 1, \qquad s^*_{i+1} = \text{sgn}(w_{i,i+1})s^*_i \quad for \; i = i_\ell, i_\ell+1, \ldots, k-1$$
$$s^*_{i_{\ell+1}} = 1, \qquad s^*_i = \text{sgn}(w_{i,i+1})s^*_{i+1} \quad for \; i = i_{\ell+1}-1, i_{\ell+1}-2, \ldots, k+1$$

Proof Denoting $\bar{u}_i := \frac{1}{2}(\bar{x}_i + \bar{y}_i)$, $\bar{v}_i := \frac{1}{2}(\bar{x}_i - \bar{y}_i)$, then $x_i = u_i + s_i v_i$, $y_i = u_i - s_i v_i$. Note that

$$\|\bar{x}_i\|^2 + \|\bar{y}_i\|^2 + \|\bar{x}_{i+1}\|^2 + \|\bar{y}_{i+1}\|^2 - g_{i,i+1}(x,y) = 2(x_i^T x_{i+1} + y_i^T y_{i+1})$$
$$= 2(\bar{u}_i + s_i\bar{v}_i)^T (\bar{u}_{i+1} + s_{i+1}\bar{v}_{i+1}) + 2(\bar{u}_i - s_i\bar{v}_i)^T (\bar{u}_{i+1} - s_{i+1}\bar{v}_{i+1})$$
$$= 4(\bar{u}_i^T \bar{u}_{i+1}) + 4(\bar{v}_i^T \bar{v}_{i+1})s_i s_{i+1}$$
$$= 4(\bar{u}_i^T \bar{u}_{i+1}) + w_{i,i+1}s_i s_{i+1}$$

Since $\bar{x}_i$, $\bar{y}_i$, $\bar{u}_i$, $\bar{v}_i$ are constants, minimizing $g_{i,i+1}(x,y)$ is equivalent to maximizing $w_{i,i+1}s_i s_{i+1}$. Then for each chunk we have to solve the optimization problem

$$\max_{s_i \in \{1,-1\}} \sum_{i=i_\ell}^{i_{\ell+1}-1} w_{i,i+1}s_i s_{i+1}, \tag{16}$$

The formulas from the first and last chunk are such that $w_{i,i+1}s^*_i s^*_{i+1} \geq 0$ for all i. This is possible because in these cases only one of the endpoints has a fixed value, and the remaining values are computed recursively starting from such a fixed point. Since all summands are nonnegative, the sum in (16) is maximized.

For the inner chunks, the two endpoints are fixed, so it may not be possible to have that $w_{i,i+1}s^*_i s^*_{i+1} \geq 0$ for all indices. In an optimal assignment we should pick at most one term to be negative, and such a term (if it exists) should be the one with the smallest absolute value $|w_{i,i+1}|$. This leads to the formula from the lemma. $\square$

5 Experiments

In this section, we apply the SNLC scheme described in Sect. 4 to synthetic and real datasets, and compare its performance with the preexisting software packages ASHIC (Ye and Ma 2020) and PASTIS (Cauer et al. 2019). We chose these two reconstruction methods for comparison because they are best suited for our setting. Also Belyaeva et al. (2022) and Tan et al. (2018) have reconstruction methods for diploid organisms, but the former method requires higher-order contact information and the latter method is targeted for single cell data.

All SNLC experiments are done using Julia 1.6.1, with ChromSDE being run in Matlab 2021a, and the Julia package `MATLAB.jl` (v0.8.3) acting as interface between Julia and Matlab. The numerical algebraic geometry part of the estimation procedure is done with `HomotopyContinuation.jl` (v2.5.5) (Breiding et al. 2018). The PASTIS experiments are run in Python 3.8.10, and the ASHIC experiments in Python 3.10.5.

For the PASTIS computations, we fix $\alpha = -2$ to ensure compatibility with the modelling assumptions made in this paper. We run PASTIS without filtering, in order to make it possible to compare RMSD values. Since PASTIS only takes integer inputs, we multiply the theoretical contact counts calculated by (2) by a factor 10^5 and round them to the nearest integer. Following the approach taken in Cauer et al. (2019), we use a coarse grid search to find the optimal coefficients for the homolog separating constraint and bead connectivity constraints. Specifically, we fix a structure simulated with the same method as used in the experiments, and compute the RMSD values for all $\lambda_1, \lambda_2 \in \{1, 10^1, 10^2, \ldots, 10^{12}\}$. In this way, we find that $\lambda_1 = 10^{11}$ and $\lambda_2 = 10^{12}$ give optimal results.

For the ASHIC computations, we use the ASHIC-ZIPM method, which has the lowest distance error rate among the ASHIC's models according to Ye (2020, Fig. 2) and models the contact counts as a zero-inflated Poisson distribution (ZIP) to account for the sparsity of the Hi-C matrix. We run ASHIC without filtering out any loci and with the setting $|$aggregate$|$ to ensure that the coordinates of all beads are estimated.

5.1 Synthetic Data

We conduct a number of experiments where we simulate a single chromosome pair (referred to as X and Y in figures) through Brownian motion with fixed step length, compute unambiguous, partially ambiguous and ambiguous contact counts according to (2), add noise, and then try to recover the structure of the chromosomes through the SNLC scheme described in Sect. 4. Following (Belyaeva et al. 2022), we model noise by multiplying each entry of C^U, C^P and C^A by a factor $1 + \delta$, where δ is sampled uniformly from the interval $(-\varepsilon, \varepsilon)$ for some chosen noise level $\varepsilon \in [0, 1]$.

As a measure of the quality of the reconstruction, we use the minimal root-mean square distance (RMSD) between, on the one hand, the true coordinates $(x_i^*, y_i^*)_{i=1}^n$, and, on the other hand, the estimated coordinates $(x_i, y_i)_{i=1}^n$ after rigid transformations and scaling, i.e., we find the minimum

$$\min_{\substack{R \in \mathrm{O}(3) \\ s > 0,\, b \in \mathbb{R}^3}} \sqrt{\frac{1}{2n} \sum_{i=1}^{n} \Big(\|(sRx_i + b) - x_i^*\|^2 + \|(sRy_i + b) - y_i^*\|^2 \Big)}.$$

This can be seen as a version of the classical Procrustes problem solved in Schönemann (1966), which is implemented in Matlab as the function `procrustes`.

Specific examples of reconstructions of the Brownian motion and helix-shaped chromosomes obtained with SNLC at varying noise levels and 50% of ambiguous beads are shown in Fig. 3. For low noise levels the reconstructions by SNLC and the original structure highly overlap. For higher noise levels the general region occupied by the reconstructions overlaps with the original structure, while the local features become less aligned. Analogous reconstructions obtained with SNLC without the local optimization step are shown in Fig. 6 in Appendix.

A comparison of how the quality of the reconstruction depends on the noise level and proportion of ambiguous beads for SNLC, ASHIC and PASTIS is done in Fig. 4. We measure the RMSD value between the reconstructed and original 3D structure for different noise levels over 20 runs. The RMSD values obtained by SNLC are consistently lower than the ones obtained by ASHIC and PASTIS. The difference is specially large for low to medium noise levels. While our method outperforms ASHIC and PASTIS in the setting considered in this paper, it is worth mentioning that ASHIC and PASTIS work also in a more general setting, where there might be contacts of all three types (ambiguous, partially ambiguous and unambiguous) between every pair of loci.

5.2 Experimentally Obtained Data

We compute SNLC reconstructions based on the real dataset explored in Cauer et al. (2019), which is obtained from Hi-C experiments on the X chromosomes in the Patski (BL6xSpretus) cell line. The data has been recorded at a resolution of 500 kb, which corresponds to 343 bead pairs in our model.

For some of these pairs, no or only very low contact counts have been recorded. Since such low contact counts are susceptible to high uncertainty and can be assumed to be a consequence of experimental errors, we exclude the 47 loci with the lowest total contact counts from the analysis. To select the cutoff, the loci are sorted according to the total contact counts (see Fig. 7a in Appendix), and the ratios between the total contact counts for consecutive loci are computed. A peak for these ratios is observed at the 47th contact count, as shown in Fig. 7b in Appendix. After applying this filter, we obtain a dataset with 296 loci. Out of these, we consider as ambiguous all loci i for which less than 40% of the total contact count comes from contacts where x_i and y_i were not distinguishable. These proportions for all loci are shown in Fig. 7c in Appendix. For the Patski dataset, we obtain 46 ambiguous loci and 250 unambiguous loci in this way.

In the Patski dataset, a locus can simultaneously participate in unambiguous, partially ambiguous and ambiguous contacts. To obtain the setting of our paper where loci are partitioned into unambiguous or ambiguous, we reassign the contacts according to
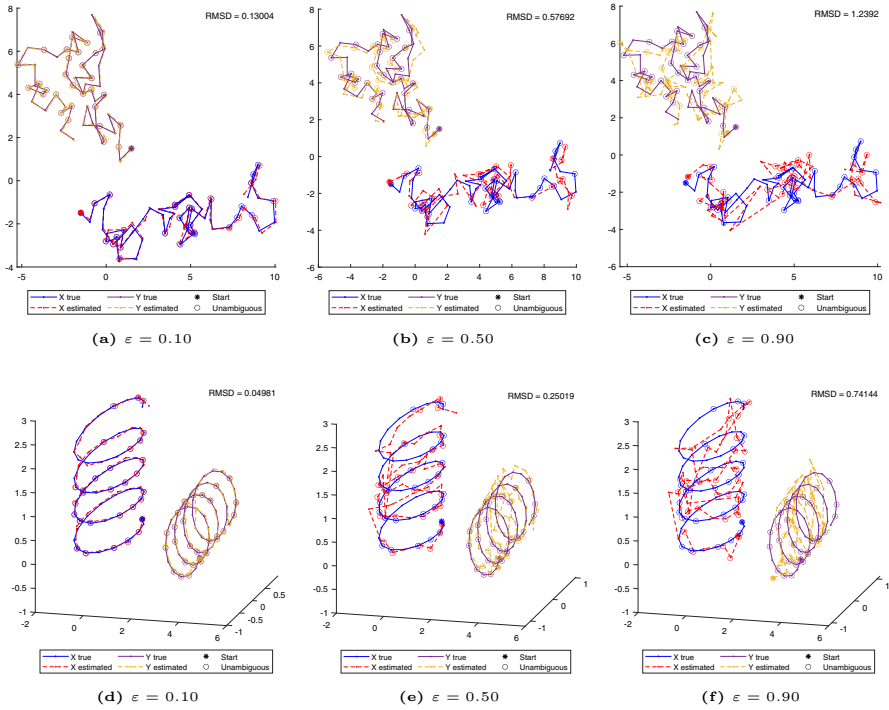
Fig. 3 Examples of reconstructions for varying noise levels, for a chromosome pair with 60 loci, out of which 50% are ambiguous. **a–c** Show chromosomes simulated with Brownian motion (projected onto the xy-plane), whereas **d–e** show helix-shaped chromosomes (color figure online)
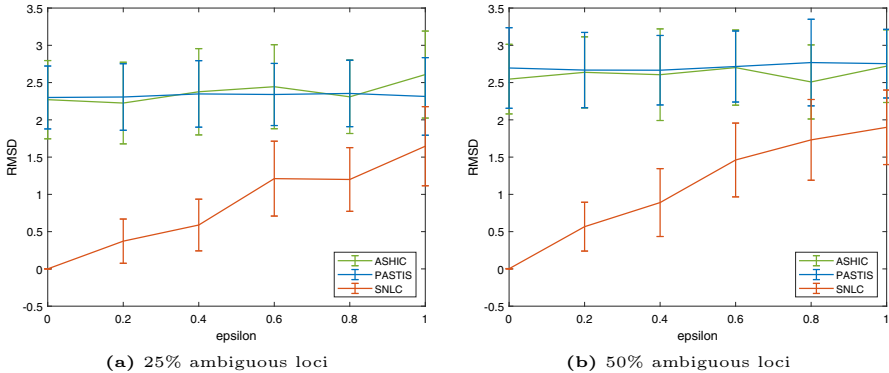

Fig. 4 Comparison between our reconstruction method, ASHIC and PASTIS. The values are the average over 20 runs, with the error bars showing the standard deviation. All experiments took place with 60 loci, with varying levels of noise, as well as varying numbers of ambiguous loci, uniformly randomly distributed over the chromosomes (color figure online)
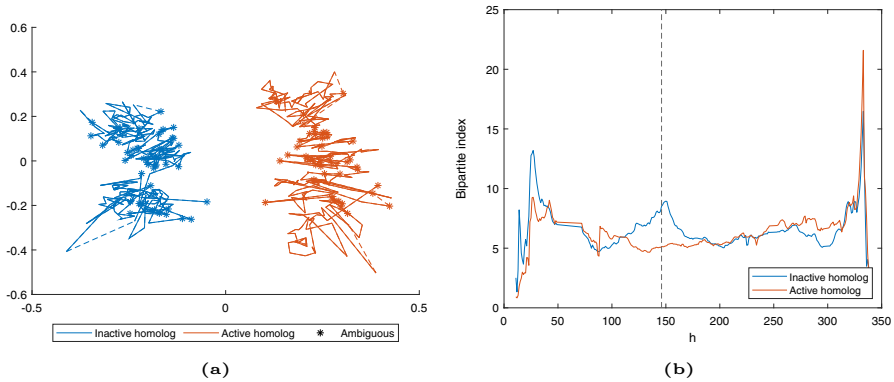
Fig. 5 **a** Reconstruction from a real dataset using our reconstruction method. A dashed line between two beads is used to indicate that there is one or more beads between them, for which we have not given an estimation (due to low contact counts). **b** Bipartite index for the reconstructed chromosomes. The dashed vertical line indicates the known hinge point at locus 146 (color figure online)

whether a locus is unambiguous or ambiguous. Our reassignment method is motivated by the assignment of haplotype to unphased Hi-C reads in Lindsly et al. (2021). The exact formulas are given in Appendix.

The reconstruction obtained via SNLC can be found in Fig. 5a. The logarithmic heatmaps for contact count matrices for original data and the SNLC reconstruction are shown in Fig. 8.

It was discovered in Deng et al. (2015) that the inactive homolog in the Patski X chromosome pair has a bipartite structure, consisting of two superdomains with frequent intra-chromosome contacts within the superdomains and a boundary region between the two superdomains. The active homolog does not exhibit the same behaviour. The boundary region on the inactive X chromosome is centered at $72.8-72.9$ MB (Deng et al. 2015) which at the 500 kB resolution corresponds to the bead 146 (Cauer et al. 2019). We show in Fig. 5b that the two chromosomes reconstructed using SNLC exhibit this structure by computing the bipartite index for the respective homologs as in Cauer et al. (2019); Deng et al. (2015). We recall that, in the setting of a single chromosome with beads $z_1, \ldots, z_n \in \mathbb{R}^3$, the bipartite index is defined as the ratio of intra-superdomain to inter-superdomain contacts in the reconstruction:

$$BI(h) = \frac{\frac{1}{h^2} \sum_{i=1}^{h} \sum_{j=1}^{h} \frac{1}{\|z_i - z_j\|^2} + \frac{1}{(n-h)^2} \sum_{i=h+1}^{n} \sum_{j=h+1}^{n} \frac{1}{\|z_i - z_j\|^2}}{\frac{2}{h(n-h)} \sum_{i=1}^{h} \sum_{j=h+1}^{n} \frac{1}{\|z_i - z_j\|^2}}.$$

6 Discussion

In this article we study the finite identifiability of 3D genome reconstruction from contact counts under the model where the distances $d_{i,j}$ and contact counts $c_{i,j}$ between two beads i and j follow the power law dependency $c_{i,j} = d_{i,j}^{\alpha}$ for a conversion factor $\alpha < 0$. We show that if at least six beads are unambiguous, then the locations of the rest of the beads can be finitely identified from partially ambiguous contact counts for rational α satisfying $\alpha < 0$ or $\alpha > 2$. In the fully ambiguous setting, we prove finite identifiability for $\alpha = -2$, given ambiguous contact counts for at least 12 pairs of beads. From Belyaeva et al. (2022) it is known that finite identifiability does not hold in the fully ambiguous setting for $\alpha = 2$. It is an open question whether finite identifiability of 3D genome reconstruction holds for other $\alpha \in \mathbb{R} \setminus \{-2, 2\}$ in the fully ambiguous setting and for rational $\alpha \in (0, 2]$ in the partially ambiguous setting. We conjecture that in the partially ambiguous setting seven unambiguous loci guarantee unique identifiability of the 3D reconstruction for rational $\alpha < 0$ or $\alpha > 2$. When $\alpha = -2$, then one approach to studying the unique identifiability might be via the degree of a parametrized family of algebraic varieties.

After establishing the identifiability, we suggest a reconstruction method for the partially ambiguous setting with $\alpha = -2$ that combines semidefinite programming, homotopy continuation in numerical algebraic geometry, local optimization and clustering. To speed up the homotopy continuation based part, we observe that the parametrized system of polynomial equations corresponding to six unambiguous beads has 40 pairs of complex solutions and we trace one path for each orbit. It is an open question to prove that for sufficiently general parameters the system has 40 pairs of complex solution. This question again reduces to studying the degree of a family of algebraic varieties. While our goal is to highlight the potential of our method, one could further regularize its output and use interpolation for the beads that are far away from the neighboring beads. A future research direction is to explore whether numerical algebraic geometry or semidefinite programming based methods can be proposed also for other conversion factors $\alpha < 0$.

Supplementary information

The code for computations and experiments is available at https://github.com/kaiekubjas/3D-genome-reconstruction-from-partially-phased-HiC-data.

Code Availability The code used for generating the synthetic data discussed in Sect. 5.1 is available in the GitHub repository https://github.com/kaiekubjas/3D-genome-reconstruction-from-partially-phased-HiC-data. This repository also contains the code used for the computations referred to in the discussion preceeding Conjecture 1, and in the proof of Theorem 2.

Appendix

In this part of the paper, we include additional details and figures for the experiments in Sect. 5.

Figure 6 shows reconstructions of the same chromosomes as displayed in Fig. 3 but without the local optimization step, indicating that semidefinite programming, numerical algebraic geometry and clustering alone can recover the main features of the 3D structure.

Figure 7 illustrates the preprocessing steps of the real dataset where loci with low contact counts are removed and the rest of the loci are partitioned into unambiguous and ambiguous. The total contact count for the ith locus is defined as the sum of all contacts where it participates:

$$T(i) = \sum_{j \in [n]} \left(c^A(i, j) + c^P(i, j) + c^P(i + n, j) \right) + \sum_{j \in [2n]} \left(c^P(j, i) + c^U(i, j) + c^U(i + n, j) \right).$$
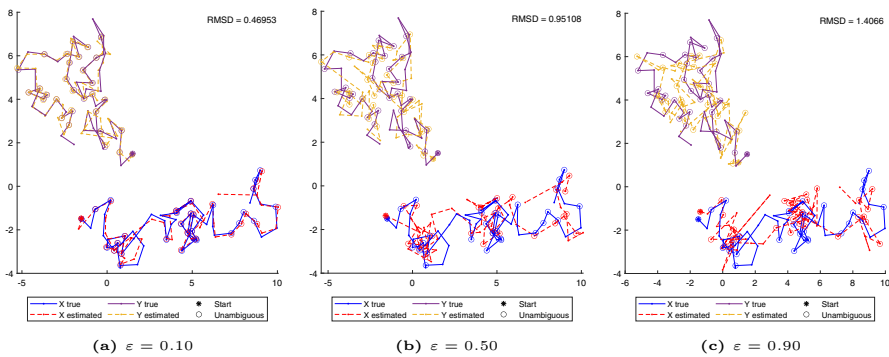

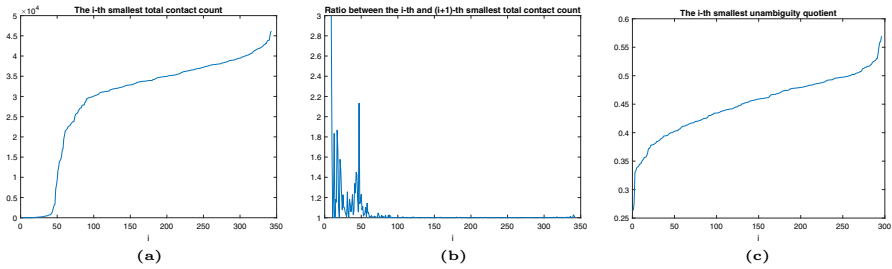**Fig. 6** SNLC reconstructions, without the local optimization step (color figure online)

Fig. 7 a Total contact counts sorted in increasing order. **b** Ratios between total contact counts. The peak corresponding to the ratio between the 48th and the 47th smallest count is used as a motivation for excluding the 47 loci with smallest total contact from the analysis. **c** Unambiguity quotients for each of the remaining 296 loci, sorted in increasing order. We consider a locus as ambiguous if this ratio is less than 0.4; otherwise, we consider it as unambiguous (color figure online)

Similarly, we define the unambiguity quotient as the proportion of $T(i)$ that consists of contacts where x_i and y_i could be distinguished:

$$
UQ(i) = \frac{1}{T(i)} \left(\sum_{j\in[n]} \left(c^P(i, j) + c^P(i+n, j) \right) + \sum_{j\in[2n]} \left(c^U(i, j) + c^U(i+n, j) \right) \right).
$$

To obtain the setting of our paper where loci are partitioned into unambiguous or ambiguous, we reassign the contact counts of $\tilde{C}^U$ $\tilde{C}^P$ and $\tilde{C}^A$ of the Patski dataset according to whether a locus is unambiguous or ambiguous. For $i, j \in U$, we define

$$
c^U_{i,j} = \tilde{c}^U_{i,j} + \tilde{c}^P_{i,j} \frac{\tilde{c}^U_{i,j}}{\tilde{c}^U_{i,j} + \tilde{c}^U_{i,j+n}} + \tilde{c}^P_{j,i} \frac{\tilde{c}^U_{i,j}}{\tilde{c}^U_{i,j} + \tilde{c}^U_{i+n,j}} + \tilde{c}^A_{i,j}
$$

$$
\frac{\tilde{c}^U_{i,j}}{\tilde{c}^U_{i,j} + \tilde{c}^U_{i,j+n} + \tilde{c}^U_{i+n,j} + \tilde{c}^U_{i+n,j+n}},
$$

$$
c^U_{i,j+n} = \tilde{c}^U_{i,j+n} + \tilde{c}^P_{i,j} \frac{\tilde{c}^U_{i,j+n}}{\tilde{c}^U_{i,j} + \tilde{c}^U_{i,j+n}} + \tilde{c}^P_{j+n,i} \frac{\tilde{c}^U_{i,j+n}}{\tilde{c}^U_{i,j+n} + \tilde{c}^U_{i+n,j+n}} +
$$

$$
+ \tilde{c}^A_{i,j} \frac{\tilde{c}^U_{i,j+n}}{\tilde{c}^U_{i,j} + \tilde{c}^U_{i,j+n} + \tilde{c}^U_{i+n,j} + \tilde{c}^U_{i+n,j+n}},
$$

$$
c^U_{i+n,j} = \tilde{c}^U_{i+n,j} + \tilde{c}^P_{i+n,j} \frac{\tilde{c}^U_{i+n,j}}{\tilde{c}^U_{i+n,j} + \tilde{c}^U_{i+n,j+n}} + \tilde{c}^P_{j,i} \frac{\tilde{c}^U_{i+n,j}}{\tilde{c}^U_{i,j} + \tilde{c}^U_{i+n,j}} +
$$

$$
+ \tilde{c}^A_{i,j} \frac{\tilde{c}^U_{i+n,j}}{\tilde{c}^U_{i,j} + \tilde{c}^U_{i,j+n} + \tilde{c}^U_{i+n,j} + \tilde{c}^U_{i+n,j+n}},
$$

$$
c^U_{i+n,j+n} = \tilde{c}^U_{i+n,j+n} + \tilde{c}^P_{i+n,j} \frac{\tilde{c}^U_{i+n,j+n}}{\tilde{c}^U_{i+n,j} + \tilde{c}^U_{i+n,j+n}} + \tilde{c}^P_{j+n,i} \frac{\tilde{c}^U_{i+n,j+n}}{\tilde{c}^U_{i,j+n} + \tilde{c}^U_{i+n,j+n}} +
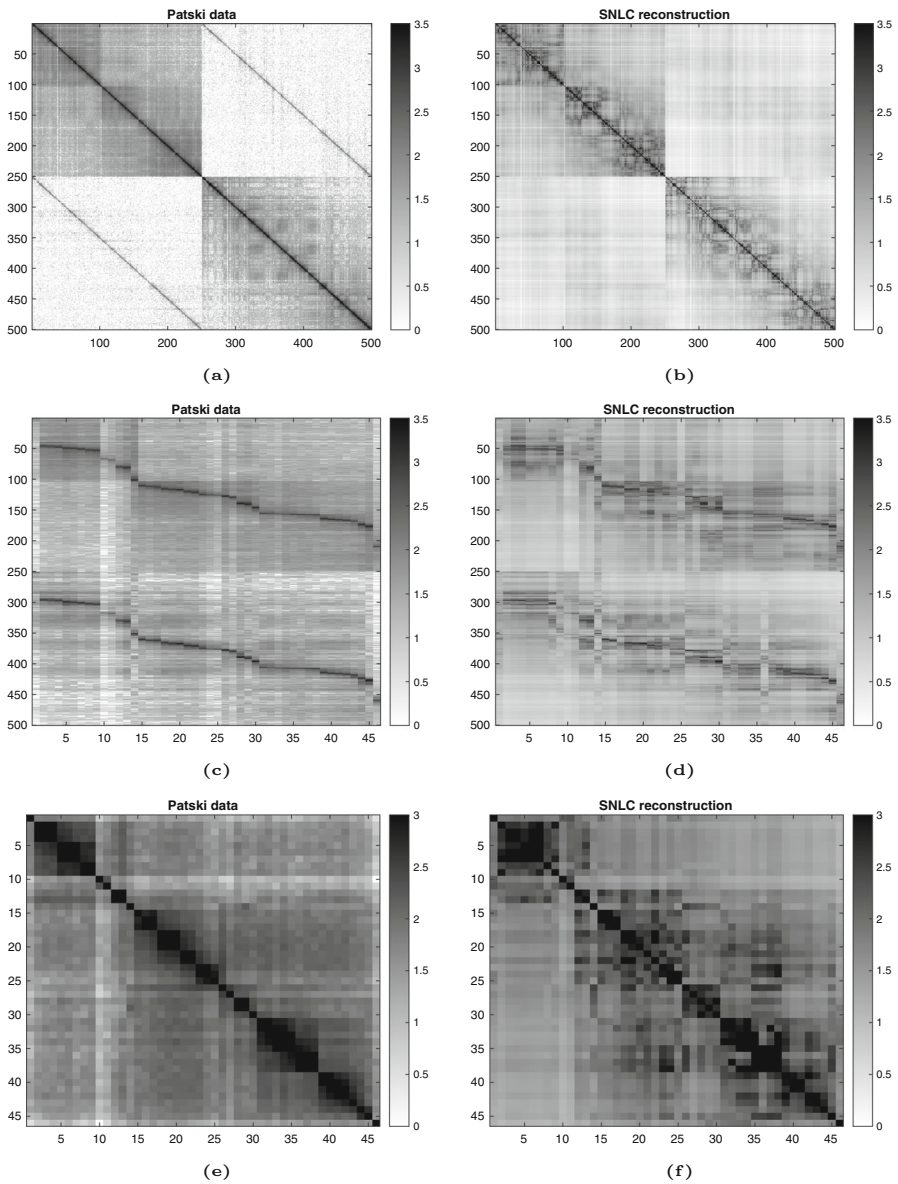$$

Fig. 8 Logarithmic heat maps for the reassigned contact count matrices obtained from the original Patski dataset and from the SNLC reconstruction: **a** and **b** C^U; **c** and **d** C^P; **e** and **f** C^A. The axis labels correspond to the 500 unambiguous beads, and the 46 ambiguous loci
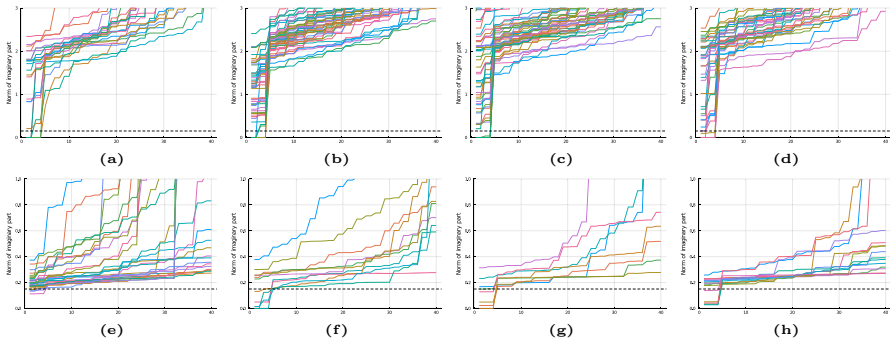
Fig. 9 Max-norm of the imaginary parts encountered in the numerical algebraic geometry estimation of various loci. Each subfigure corresponds to an ambiguous locus: **a–d** correspond to the first four loci of the synthetic dataset used in Fig. 3b; **e–h** correspond to the first four ambiguous loci of the Patski dataset. Each colored line corresponds to a specific choice of 6 unambiguous beads used in the estimation of the locus. Each line connects 40 points, that record the max-norm of the imaginary part of a solution (up to symmetry) found for the corresponding choice of 6 unambiguous beads. The dashed line at 0.15 corresponds to the choice of threshold for when a solution is considered approximately real. Similar figures for the rest of the ambiguous loci in the respective chromosome pairs can be found in the Github repository (color figure online)

$$
+ \tilde{c}_{i,j}^{A} \frac{\tilde{c}_{i+n,j+n}^{U}}{\tilde{c}_{i,j}^{U} + \tilde{c}_{i,j+n}^{U} + \tilde{c}_{i+n,j}^{U} + \tilde{c}_{i+n,j+n}^{U}}.
$$

For $i \in U$, $j \in A$, we define

$$
c_{i,j}^{P} = \tilde{c}_{i,j}^{U} + \tilde{c}_{i,j+n}^{U} + \tilde{c}_{i,j}^{P} + \tilde{c}_{j,i}^{P} \frac{\tilde{c}_{i,j}^{U}}{\tilde{c}_{i,j}^{U} + \tilde{c}_{i+n,j}^{U}} + \tilde{c}_{j+n,i}^{P} \frac{\tilde{c}_{i,j+n}^{U}}{\tilde{c}_{i,j+n}^{U} + \tilde{c}_{i+n,j+n}^{U}} +
$$

$$
+ \tilde{c}_{i,j}^{A} \frac{\tilde{c}_{i,j}^{P}}{\tilde{c}_{i,j}^{P} + \tilde{c}_{i+n,j}^{P}},
$$

$$
c_{i+n,j}^{P} = \tilde{c}_{i+n,j}^{U} + \tilde{c}_{i+n,j+n}^{U} + \tilde{c}_{i+n,j}^{P} + \tilde{c}_{j,i}^{P} \frac{\tilde{c}_{i+n,j}^{U}}{\tilde{c}_{i,j}^{U} + \tilde{c}_{i+n,j}^{U}} + \tilde{c}_{j+n,i}^{P} \frac{\tilde{c}_{i+n,j+n}^{U}}{\tilde{c}_{i,j+n}^{U} + \tilde{c}_{i+n,j+n}^{U}} +
$$

$$
+ \tilde{c}_{i,j}^{A} \frac{\tilde{c}_{i+n,j}^{P}}{\tilde{c}_{i,j}^{P} + \tilde{c}_{i+n,j}^{P}}.
$$

Finally, for $i, j \in A$, we define

$$
c_{i,j}^{A} = \tilde{c}_{i,j}^{U} + \tilde{c}_{i,j+n}^{U} + \tilde{c}_{i+n,j}^{U} + \tilde{c}_{i+n,j+n}^{U} + \tilde{c}_{i,j}^{P} + \tilde{c}_{i+n,j}^{P} + \tilde{c}_{j,i}^{P} + \tilde{c}_{j+n,i}^{P} + \tilde{c}_{i,j}^{A}.
$$

In Fig. 8 in Appendix, the experimental contact counts from the Patski dataset are compared with the contact counts from the SNLC reconstruction.

Figure 9 shows how the max-norm of the imaginary part of the solutions varies between different instances of the system (12) used for the reconstruction in Fig. 3(b), and for the reconstruction from the Patski data in Fig. 5. A complete set of figures for

these two datasets can be found in the Github repository. Taken together, the figures indicate that a max-norm of 0.15 was an appropriate threshold for approximate realness for both data sets, in the sense that it is low enough to single out solutions that have significantly smaller imaginary parts than the others, while also ensuring that it is possible to find an approximately real solution for each ambiguous locus.

References

Alfakih AY, Khandani A, Wolkowicz H (1999) Solving euclidean distance matrix completion problems via semidefinite programming. Comput Optim Appl 12(1):13–30

Belyaeva A, Kubjas K, Sun LJ, Uhler C (2022) Identifying 3D genome organization in diploid organisms via Euclidean distance geometry. SIAM J Math Data Sci 4(1):204–228

Breiding P, Rose K, Timme S (2023) Certifying zeros of polynomial systems using interval arithmetic. ACM Trans Math Softw 49(1):1–14

Breiding P, Timme S (2018) HomotopyContinuation.jl: A package for homotopy continuation in Julia. In: Davenport JH, Kauers M, Labahn G, Urban J (eds) Mathematical Software—ICMS 2018. Springer, Cham, pp 458–465

Cauer AG, Yardimci G, Vert JP, Varoquaux N, Noble WS (2019) Inferring diploid 3D chromatin structures from Hi-C data. In: 19th International workshop on algorithms in bioinformatics (WABI 2019)

Cox MA, Cox TF (2008) Multidimensional scaling. In: Handbook of data visualization. Springer, Berlin, pp 315–347

Deng X, Ma W, Ramani V, Hill A, Yang F, Ay F, Berletch JB, Blau CA, Shendure J, Duan Z (2015) Bipartite structure of the inactive mouse X chromosome. Genome Biol 16(1):1–21

Dokmanic I, Parhizkar R, Ranieri J, Vetterli M (2015) Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Signal Process Mag 32(6):12–30

Eagen KP (2018) Principles of chromosome architecture revealed by Hi-C. Trends Biochem Sci 43(6):469–478

Fang H-R, O'Leary DP (2012) Euclidean distance matrix completion problems. Optim Methods Softw 27(4–5):695–717

Fazel M, Hindi H, Boyd SP (2003) Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In: Proceedings of the 2003 American control conference, vol 3. IEEE, pp 2156–2162

Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS (2013) Bayesian inference of spatial organizations of chromosomes. PLoS Comput Biol 9(1):1002893

Huber B, Sturmfels B (1995) A polyhedral method for solving sparse polynomial systems. Math Comput 64(212):1541–1555

Jiang K, Sun D, Toh K-C (2014) A partial proximal point algorithm for nuclear norm regularized matrix least squares problems. Math Program Comput 6:1

Krislock N (2010) Semidefinite facial reduction for low-rank Euclidean distance matrix completion. PhD thesis, University of Waterloo. http://hdl.handle.net/10012/5093

Krislock N, Wolkowicz H (2012) Euclidean distance matrices and applications. Handbook on Semidefinite. Conic and Polynomial Optimization. Springer, New York, pp 879–914

Lafontaine DL, Yang L, Dekker J, Gibcus JH (2021) Hi-C 3.0: improved protocol for genome-wide chromosome conformation capture. Curr Protoc 1(7):198

Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J (2014) 3D genome reconstruction from chromosomal contacts. Nat Methods 11(11):1141–1143

Li T-Y, Wang X (1996) The BKK root count in $\mathbb{C}^n$. Math Comput 65(216):1477–1484

Li J, Lin Y, Tang Q, Li M (2021) Understanding three-dimensional chromatin organization in diploid genomes. Comput Struct Biotechnol J 19:3589

Liberti L, Lavor C, Maculan N, Mucherino A (2014) Euclidean distance geometry and applications. SIAM Rev 56(1):3–69

Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293

Lindsly S, Jia W, Chen H, Liu S, Ronquist S, Chen C, Wen X, Stansbury C, Dotson GA, Ryan C (2021) Functional organization of the maternal and paternal human 4D nucleome. IScience 24(12):103452

Luo H, Li X, Fu H, Peng C (2020) HiCHap: a package to correct and analyze the diploid hi-c data. BMC Genomics 21(1):1–13

Minajigi A, Froberg JE, Wei C, Sunwoo H, Kesner B, Colognori D, Lessing D, Payer B, Boukhali M, Haas W et al (2015) A comprehensive Xist interactome reveals Cohesin repulsion and an RNA-directed chromosome conformation. Science 349(6245):1

Mishra B, Meyer G, Sepulchre R (2011) Low-rank optimization for distance matrix completion. In: 2011 50th IEEE conference on decision and control and european control conference. IEEE, pp 4455–4460

Mucherino A, Lavor C, Liberti L, Maculan N (2012) Distance geometry: theory, methods, and applications. Springer, New York

Nie J (2009) Sum of squares method for sensor network localization. Comput Optim Appl 43(2):151–179

Nott A, Holtman IR, Coufal NG, Schlachetzki JC, Yu M, Hu R, Han CZ, Pena M, Xiao J, Wu Y (2019) Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. Science 366(6469):1134–1139

Oluwadare O, Highsmith M, Cheng J (2019) An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data. Biol Proced Online 21(1):1–20

Paulsen J, Sekelja M, Oldenburg AR, Barateau A, Briand N, Delbarre E, Shah A, Sørensen AL, Vigouroux C, Buendia B (2017) Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. Genome Biol 18(1):1–15

Payne AC, Chiang ZD, Reginato PL, Mangiameli SM, Murray EM, Yao C-C, Markoulaki S, Earl AS, Labade AS, Jaenisch R (2021) In situ genome sequencing resolves DNA sequence and structure in intact biological samples. Science 371(6532):3446

Rajarajan P, Borrman T, Liao W, Schrode N, Flaherty E, Casiño C, Powell S, Yashaswini C, LaMarca EA, Kassim B et al (2018) Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. Science 362(6420):1

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665–1680

Rhie SK, Schreiner S, Witt H, Armoskus C, Lay FD, Camarena A, Spitsyna VN, Guo Y, Berman BP, Evgrafov OV (2018) Using 3D epigenomic maps of primary olfactory neuronal cells from living individuals to understand gene regulation. Sci Adv 4(12):8550

Rousseau M, Fraser J, Ferraiuolo MA, Dostie J, Blanchette M (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. Bioinformatics 12(1):414

Schönemann PH (1966) A generalized solution of the orthogonal Procrustes problem. Psychometrika 31(1):1–10

Segal MR (2022) Can 3D diploid genome reconstruction from unphased Hi-C data be salvaged? NAR Genom Bioinf 4(2):038

Sommese AJ, Wampler CW (2005) Numerical solution of systems of polynomials arising in engineering and science. World Scientific Publishing Company, Singapore

Sonthalia R, Van Buskirk G, Raichel B, Gilbert A (2021) How can classical multidimensional scaling go wrong? Adv Neural Inf Process Syst 34:12304–12315

Sturmfels B, Telen S (2021) Likelihood equations and scattering amplitudes. Algebr Stat 12(2):167–186

Tan L, Xing D, Chang C-H, Li H, Xie XS (2018) Three-dimensional genome structures of single diploid human cells. Science 361(6405):924–928

Uhler C, Shivashankar G (2017) Regulation of genome organization and gene expression by nuclear mechanotransduction. Nat Rev Mol Cell Biol 18(12):717–727

Varoquaux N, Ay F, Noble WS, Vert J-P (2014) A statistical approach for inferring the 3D structure of the genome. Bioinformatics 30(12):26–33

Wang H, Xu X, Nguyen CM, Liu Y, Gao Y, Lin X, Daley T, Kipniss NH, La Russa M, Qi LS (2018) CRISPR-mediated programmable 3D genome positioning and nuclear organization. Cell 175(5):1405–1417

Weinberger KQ, Sha F, Zhu Q, Saul LK (2007) Graph Laplacian regularization for large-scale semidefinite programming. In: Advances in neural information processing systems, pp 1489–1496

Ye T, Ma W (2020) ASHIC: hierarchical Bayesian modeling of diploid chromatin contacts and structures. Nucl Acids Res 48(21):123–123

Zhang Z, Li G, Toh K-C, Sung W-K (2013) Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-C data. In: Annual international conference on research in computational molecular biology. Springer, pp 317–332

Zhou S, Xiu N, Qi H-D (2020) Robust Euclidean embedding via EDM optimization. Math Program Comput 12(3):337–387

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