Belyaeva, Anastasiya; Kubjas, Kaie; Sun, Lawrence J.; Uhler, Caroline

Identifying 3D Genome Organization in Diploid Organisms via Euclidean Distance Geometry

Published in:
SIAM Journal on Mathematics of Data Science

DOI:
10.1137/21M1390372

Published: 28/02/2022

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Belyaeva, A., Kubjas, K., Sun, L. J., & Uhler, C. (2022). Identifying 3D Genome Organization in Diploid Organisms via Euclidean Distance Geometry. *SIAM Journal on Mathematics of Data Science*, *4*(1), 204–228. https://doi.org/10.1137/21M1390372

SIAM J. MATH. DATA SCI.
Vol. 4, No. 1, pp. 204–228

Identifying 3D Genome Organization in Diploid Organisms via Euclidean Distance Geometry*

Anastasiya Belyaeva†, Kaie Kubjas‡, Lawrence J. Sun†, and Caroline Uhler§

Abstract. The spatial organization of the genome in the cell nucleus plays an important role for gene regulation, replication of the deoxyribonucleic acid (DNA), and genomic integrity. Through the development of chromosome conformation capture experiments (such as 3C, 4C, and Hi-C) it is now possible to obtain the contact frequencies of the DNA at the whole-genome level. In this paper, we study the problem of reconstructing the three-dimensional (3D) organization of the genome from such whole-genome contact frequencies. A standard approach is to transform the contact frequencies into noisy distance measurements and then apply semidefinite programming formulations to obtain the 3D configuration. However, neglected in such reconstructions is the fact that most eukaryotes including humans are diploid and therefore contain two copies of each genomic locus. We prove that the 3D organization of the DNA is not identifiable from the distance measurements derived from contact frequencies in diploid organisms. In fact, there are infinitely many solutions even in the noise-free setting. We then discuss various additional biologically relevant and experimentally measurable constraints (including distances between neighboring genomic loci and higher-order interactions) and prove identifiability under these conditions. Furthermore, we provide semidefinite programming formulations for computing the 3D embedding of the DNA with these additional constraints and show that we can recover the true 3D embedding with high accuracy from both noiseless and noisy measurements. Finally, we apply our algorithm to real pairwise and higher-order contact frequency data and show that we can recover known genome organization patterns.

Key words. 3D genome organization, diploid organisms, semidefinite programming, Hi-C, Euclidean distance geometry, systems of polynomial equations

AMS subject classifications. 51K05, 92E10, 90C22, 52C25, 14P05

DOI. 10.1137/21M1390372

*Received by the editors January 6, 2021; accepted for publication (in revised form) September 23, 2021; published electronically February 28, 2022.

https://doi.org/10.1137/21M1390372

Funding: The first author was supported by an NSF Graduate Research Fellowship (1122374), the Abdul Latif Jameel World Water and Food Security Lab (J-WAFS) at MIT, and the MIT J-Clinic for Machine Learning and Health. The second author was supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement 748354 at MIT and Sorbonne University. The fourth author was partially supported by the NSF (DMS-1651995), the ONR (N00014-17-1-2147 and N00014-18-1-2765), the MIT-IBM Watson AI lab, the Eric and Wendy Schmidt Center at the Broad Institute, and a Simons Investigator Award.

†Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, and Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA 02138 USA (belyaeva@mit.edu, sunl@mit.edu).

‡Department of Mathematics and Systems Analysis, Aalto University, Aalto, FI-00076 Finland (kaie.kubjas@aalto.fi).

§Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Institute for Data, Systems and Society, Massachusetts Institute of Technology, Cambridge, MA 02138 USA, and Broad Institute of MIT and Harvard, Cambridge, MA 02142 USA (cuhler@mit.edu).

1. Introduction. It is now well established that the spatial organization of the genome in the cell nucleus plays an important role for cellular processes including gene regulation, replication of the deoxyribonucleic acid (DNA), and the maintenance of genomic integrity [12, 47, 48], and its misfolding is associated with disease [35]. Notably, a recent study [52] showed a causal link between three-dimensional (3D) genome organization and gene regulation, where gene repositioning was induced and subsequent changes in gene expression were observed. This motivates the development of methods to reconstruct the 3D structure of the genome to study its functions.

The genetic information in cells is contained in the DNA, which is organized into chromosomes and packed into the cell nucleus. Chromosome confirmation capture techniques such as chromosome conformation capture-on-chip (4C), genome-wide chromosome conformation capture (Hi-C), and Capture-C have enabled the interrogation of the contact frequencies between pairs of genomic loci at the whole-genome scale [13, 44, 28, 24]. In Hi-C, for example, interacting chromosome regions are cross-linked (i.e., frozen); the DNA is then fragmented, the cross-linked fragments are ligated, and paired-end sequencing is applied to the ligation products and mapped to a reference genome [28]. By binning the genome and ascribing each read pair into the corresponding bin, one obtains a contact frequency matrix between genomic loci that is commonly of the size $10^6 \times 10^6$.

Different computational approaches for reconstructing the 3D genome organization from contact frequency data have been considered. Distance-based approaches convert contact frequencies F_{ij} into spatial distances D_{ij} and find a Euclidean embedding of the points in three dimensions [15, 56, 27, 41]. Ensemble methods such as Markov chain Monte Carlo 5C (MCMC5C) and the Bayesian 3D constructor for Hi-C data (BACH) [42, 23] learn a set of possible 3D structures by defining a probabilistic model for contact frequencies and generating an ensemble of structures via Markov chain Monte Carlo sampling. Other ensemble methods include molecular dynamics simulations that model DNA as a polymer and output an ensemble of 3D structures [28, 30, 14, 38]. Finally, statistical methods have also been proposed that directly model contact counts instead of distances, using, for example, the Poisson distribution [50], and maximize the log-likelihood of the data to infer the 3D genome organization. Maximizing the log-likelihood of the data under the constraint that the 3D solution is a smooth one-dimensional space curve combines statistical methods with distance-based approaches [46].

Almost all existing methods make the simplifying assumption that the genome is haploid when in fact most organisms of interest including humans are diploid; i.e., there are two copies of each chromosome known as *homologous chromosomes*. For example, human cells contain two copies of 23 chromosomes each. The challenge is that the contact frequency data from chromosome conformation capture experiments is generally *unphased*, meaning that the copies of each chromosome cannot be distinguished. As a result, if the DNA is modeled as a string of beads containing two copies of each bead i for $1 \leq i \leq n$, then the measured contact frequencies result in an $n \times n$ matrix, from which we would like to infer the 3D embedding of $2n$ points. This problem cannot be solved by the classical 3D genome reconstruction methods such as those mentioned above. With significant experimental efforts, phased data can be obtained [8, 9] and used in order to reconstruct the 3D genome organization [5]. However, such data are rare and costly. Methods for inferring the full diploid 3D structure from unphased data (and optionally phased data) are beginning to be explored; for example, the authors

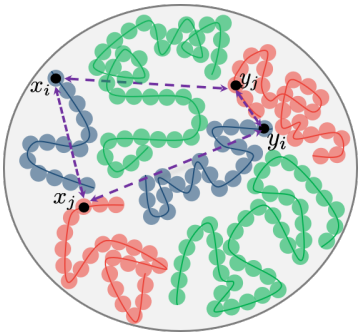


Figure 1. *Schematic of the diploid genome. Nucleus with green, blue, and red curves depicts three homologous pairs of chromosomes. In the unphased setting, the measured distance between loci i and j corresponds to the sum of the four distances (denoted in purple) between the two pairs of homologous loci x_i, y_i and x_j, y_j.*

in [5] model the contact frequency data from diploid structures using a Poisson model and maximize the log-likelihood of the model to infer the 3D diploid structure. However, it is unclear whether the true diploid structures are even identifiable.

In this paper, we provide a computational method for inferring the 3D diploid organization of the genome without relying on phased data and provide conditions under which the problem is identifiable. In particular, we consider a distance-based approach and use Euclidean distance geometry to obtain the 3D diploid structure of the genome. Similar to other distance-based approaches, we rely on prior knowledge of the mapping function between frequencies and distances. While some biological experiments have been done to quantify this relationship [53], identifying the mapping function is an important concern in 3D genome structure prediction since it can impact the performance of reconstruction algorithms, and it is a topic of interest in current research [19, 51].

The precise mathematical problem considered in this paper is as follows and is illustrated in Figure 1. DNA is modeled as a string of beads that contains two copies of each bead i for $1 \leq i \leq n$. We would like to infer the location of the two copies of each bead, which we denote by $x_i \in \mathbb{R}^3$ and $y_i \in \mathbb{R}^3$. Since for unphased data, the two copies of each bead cannot be distinguished, the problem is to identify the 3D configuration ($2n \times 3$ matrix), i.e., $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}^3$ (up to translation, rotation, and reflection), from the composite distance measurements D_{ij}, $1 \leq i \neq j \leq n$ ($n \times n$ matrix), corresponding to the sum of the distances between either copy of bead i and j, i.e.,

$$D_{ij} = \|x_i - x_j\|^2 + \|x_i - y_j\|^2 + \|y_i - x_j\|^2 + \|y_i - y_j\|^2.$$

In the haploid or phased setting, this problem boils down to the standard Euclidean distance geometry problem. This problem has a long history: in the classical setting with no missing values, this problem can be solved via the classical multidimensional scaling algorithm that is based on spectral decomposition followed by dimensionality reduction; see [10] for an overview. Other approaches for the Euclidean embedding and completion problems, including in the presence of missing values, are nonconvex formulations [16, 31] as well as semidefinite relaxations [1, 17, 6, 29, 54, 33, 55].

A naive approach in the unphased diploid setting is to assume that the four distances that make up our measured composite distance D_{ij} are equal and solve the corresponding

Euclidean embedding problem. However, it is evident from single-cell imaging studies that the four distances in D_{ij} can be wildly different [4, 34]. Hence this approach cannot provide realistic embeddings. While a simple dimension argument ($6n$ variables versus $\binom{n}{2}$ constraints) suggests that the 3D genome configuration is uniquely identifiable, one of the main results of our paper is that the 3D diploid genome configuration is not identifiable from unphased data. In fact, we show that there are infinitely many configurations that satisfy the constraints imposed by D_{ij}, even in the noiseless setting (section 2, Theorem 2.1).

We therefore consider additional biologically relevant and experimentally measurable constraints and study identifiability of the 3D diploid structure under these constraints. First, we take into account the distances between neighboring beads, i.e., $\|x_i - x_{i+1}\|$ and $\|y_i - y_{i+1}\|$, on each chromosome. While we show that this yields unique identifiability for configurations in two dimensions, there are still infinitely many configurations in three dimensions, which is of primary interest for genome modeling (section 3, Proposition 3.1, Proposition 3.2). To obtain identifiability in three dimensions, we consider adding constraints based on the contact frequencies between three or more loci simultaneously. The measurement of such higher-order contact frequencies has recently been enabled by experimental assays such as the split-pool recognition of interactions of tag extension (SPRITE) [39], chromosomal walks (C-walks) [36], and genome architecture mapping (GAM) [3]. We prove that this information can be used to uniquely identify the 3D genome organization from unphased data in the noiseless setting (section 4, Theorem 4.1).

Finally, we provide a semidefinite programming (SDP) formulation for obtaining the 3D diploid configuration from noisy measurements (section 5) and show based on simulated data that our algorithm has good performance and that it is able to recover known genome organization patterns when applied to real contact frequency data collected from human lymphoblastoid cells (section 6).

2. Unidentifiability from pairwise distance constraints.

In the remainder of the paper we denote the true but unknown coordinates of the homologous loci by x_i^* and y_i^* and the corresponding noiseless distances by D_{ij}^*, while the symbols x_i and y_i denote the variables that we want to solve for. While from a biological perspective the relevant setting is when $x_i, y_i \in \mathbb{R}^3$, results that hold more generally will be stated in $\mathbb{R}^d$. The main result of this section is Theorem 2.1, which characterizes the set of solutions given by the constraints D_{ij}^* in dimension $d \leq 3$. In particular, it establishes nonidentifiability of the 3D genome structure from pairwise distance measurements in the diploid unphased setting.

Theorem 2.1. *Let $d \leq 3$ and $n \geq 2d+3$. Then $(x_1, \dots, x_n, y_1, \dots, y_n) \in (\mathbb{R}^d)^{2n}$ satisfies*

$$D_{ij}^* = \|x_i - x_j\|^2 + \|x_i - y_j\|^2 + \|y_i - x_j\|^2 + \|y_i - y_j\|^2 \quad \text{for all } 1 \leq i \neq j \leq n \tag{2.1}$$

if and only if it satisfies

$$x_i + y_i = x_i^* + y_i^* \quad \text{and} \quad \|x_i\|^2 + \|y_i\|^2 = \|x_i^*\|^2 + \|y_i^*\|^2 \quad \text{for all } 1 \leq i \leq n \tag{2.2}$$

up to translations, rotations, and reflections in $\mathbb{R}^d$ and permutations of x_i and y_i.

As a consequence, the measurements D_{ij}^* identify the location of each pair of homologous loci (x_i, y_i) up to a sphere with center $(x_i^* + y_i^*)/2$ and radius $\|x_i^* - y_i^*\|/2$. Namely, the points

x_i, y_i lie opposite to each other anywhere on this sphere. Unless $x_i^* = y_i^*$ for all i, i.e., all spheres have radius 0, this set is infinite in dimensions $d > 1$, and hence the configuration is unidentifiable.

In the remainder of this section, we will prove Theorem 2.1. The two inclusions in Theorem 2.1 are proven in Lemma 2.2 and Lemma 2.4. In Lemma 2.3 it is shown that the distance $\|x_i - y_i\|$ within each homologous pair is fixed given the pairwise distances D_{ij}^*. This result is used to prove Lemma 2.4.

Lemma 2.2. *Let* $(x_1, \dots, x_n, y_1, \dots, y_n) \in (\mathbb{R}^d)^{2n}$ *satisfy*

$$x_i + y_i = x_i^* + y_i^* \text{ and } \|x_i\|^2 + \|y_i\|^2 = \|x_i^*\|^2 + \|y_i^*\|^2 \text{ for all } 1 \le i \le n. \tag{2.3}$$

Then

$$\|x_i - x_j\|^2 + \|x_i - y_j\|^2 + \|y_i - x_j\|^2 + \|y_i - y_j\|^2 = D_{ij}^* \text{ for all } 1 \le i \ne j \le n.$$

Proof. Observe that for each pair x_i, y_i satisfying (2.3), it holds that

$$\begin{aligned} D_{ij}^* &= 2 \cdot (\|x_i^*\|^2 + \|y_i^*\|^2) + 2 \cdot (\|x_j^*\|^2 + \|y_j^*\|^2) - 2(x_i^* + y_i^*) \cdot (x_j^* + y_j^*) \\ &= 2 \cdot (\|x_i\|^2 + \|y_i\|^2) + 2 \cdot (\|x_j\|^2 + \|y_j\|^2) - 2(x_i + y_i) \cdot (x_j + y_j) \\ &= \|x_i - x_j\|^2 + \|x_i - y_j\|^2 + \|y_i - x_j\|^2 + \|y_i - y_j\|^2. \end{aligned}$$

This completes the proof. ∎

Next we will show that the distance between homologous pairs is uniquely determined by the D_{ij}^*.

Lemma 2.3. *Let* $d \le 3$ *and* $n \ge 2d + 3$. *Then for each* $1 \le i \le n$ *the quantity* $\|x_i - y_i\|$ *is identifiable from the constraints imposed by the* D_{ij}^**; i.e., for any solution* $(x_1, \dots, x_n, y_1, \dots, y_n) \in (\mathbb{R}^d)^{2n}$ *to the equations defined by the* D_{ij}^* *in* (2.1), *the quantity* $\|x_i - y_i\|$ *is constant.*

The constraint $d \le 3$ is due to our proof technique. The condition $n \ge 2d+3$ is necessary for unique identifiability of the distance between homologous pairs of loci.

Proof. Without loss of generality we assume that $i = 1$ and show that $\|x_1 - y_1\|$ is equal to some constant. First, we perform a shift on the solution so that $x_1 = -y_1 = v$. Since shifts preserve distances, they in particular preserve the equality constraints (2.1). Hence,

$$D_{1j}^* = \|v - x_j\|^2 + \|v - y_j\|^2 + \| - v - x_j\|^2 + \| - v - y_j\|^2.$$

Expanding this out into dot products and simplifying yields

$$D_{1j}^* = 4\|v\|^2 + 2(\|x_j\|^2 + \|y_j\|^2).$$

Let $j \ne k$ be both not equal to 1. Then substituting the above leads to

$$D_{1j}^* + D_{1k}^* - D_{jk}^* = 8\|v\|^2 + 2(x_j + y_j) \cdot (x_k + y_k).$$

Defining $T_{jk} := D_{1j}^* + D_{1k}^* - D_{jk}^*$ and $s_j := \sqrt{2}(x_j + y_j)$, this is equivalent to

$$T_{jk} - 8\|v\|^2 = s_j \cdot s_k.$$

Let T' be the $(d+1)\times(d+1)$ submatrix of T satisfying $T'_{ij} = T_{i+1,j+d+2}$; i.e., the rows of T' correspond to the rows $2,3,\dots,d+2$ of T, and the columns of T' correspond to the columns $d+3, d+4,\dots,2d+3$ of T. We now show that for generic configurations $\det(T') \neq 0$. Since $\det(T')$ can be written as a polynomial in the coordinates x_i and y_i, then $\det(T') \neq 0$ for generic configurations as long as it does not identically vanish. Hence it suffices to present one configuration where $\det(T')$ is nonzero. For $d \leq 3$ we can check this using random configurations.

Since T' has full rank, then the matrix determinant lemma implies that

$$\det(T' - 8J\|v\|^2) = (1 - 8\|v\|^2 \mathbb{1}^T (T')^{-1}\mathbb{1}) \det(T'), \tag{2.4}$$

where $\mathbb{1}$ denotes the all-ones vector. Note that the scalar $\mathbb{1}^T T'^{-1}\mathbb{1}$ is fixed and $(\det T') \neq 0$. Furthermore, since $T' - 8J\|v\|^2$ is formed from the dot products between d-dimensional vectors, it has rank at most d, and therefore $\det(T' - 8J\|v\|^2) = 0$ due to $T' - 8J\|v\|^2$ being a $(d+1)\times(d+1)$ matrix. Hence, $(1 - 8\|v\|^2\mathbb{1}^T(T')^{-1}\mathbb{1})\det(T') = 0$, which is a linear equation in terms of $\|v\|^2$. As a consequence, it has a unique solution for $\|v\|^2$, and thus the distance between the homologous pair x_1, y_1 is fixed as long as $n \geq 2d+3$. ∎

We next characterize all solutions to the constraints imposed by the D^*_{ij}.

Lemma 2.4. *Let $d \leq 3$ and $n \geq 2d+3$. Let $(x_1,\dots,x_n,y_1,\dots,y_n) \in (\mathbb{R}^d)^{2n}$ be a solution to*

$$\|x_i - x_j\|^2 + \|x_i - y_j\|^2 + \|y_i - x_j\|^2 + \|y_i - y_j\|^2 = D^*_{ij} \quad \text{for all } 1 \leq i \neq j \leq n.$$

Then

$$x_i + y_i = x^*_i + y^*_i \text{ and } \|x_i\|^2 + \|y_i\|^2 = \|x^*_i\|^2 + \|y^*_i\|^2 \quad \text{for all } 1 \leq i \leq n$$

up to translations, rotations, and reflections in $\mathbb{R}^d$ and permutations of x_i and y_i.

Proof. Without loss of generality we perform a translation on the solution such that $x_1 = -y_1 = v$ for some vector v. By Lemma 2.3 the quantity $\|x_k - y_k\|$ is constant for each $1 \leq k \leq n$, and thus also $\|v\|$ is constant. Since for any $j \neq 1$ it holds that $D^*_{1j} = 4\|v\|^2 + 2(\|x_j\|^2 + \|y_j\|^2)$, $\|x_j\|^2 + \|y_j\|^2$ is also constant, and hence $\|x_i\|^2 + \|y_i\|^2 = \|x^*_i\|^2 + \|y^*_i\|^2$ for all $1 \leq i \leq n$.

Similar to the proof of Lemma 2.3, if we define $T_{jk} = D^*_{1j} + D^*_{1k} - D^*_{jk}$ and $s_j = \sqrt{2}(x_j + y_j)$, we find that

$$T_{jk} - 8\|v\|^2 = s_j \cdot s_k.$$

Because we have access to the diagonal constraints now, this relationship holds for all j,k and not just $j \neq k$. Thus $T - 8J\|v\|^2$ is a symmetric $(n-1)\times(n-1)$ matrix admitting a rank d factorization. Let S be the matrix formed with the vectors s_j. We then have $T - 8J\|v\|^2 = SS^T$. There is a result on rank factorizations of symmetric matrices that any other factorization $T - 8J\|v\|^2 = S'S'^T$ satisfies $S = S'Q$ for some orthogonal matrix Q [26, Proposition 3.2]. Thus for any other solution s'_j, we have $s_j = s'_j Q$, implying all solutions are simply orthogonal transformations of each other (rotations, reflections, etc.).

In summary, we have shown that once we have fixed $x_1 + y_1 = 0$ via translation, then the quantities $x_j + y_j$ are unique up to orthogonal transformations, and the quantities $\|x_j\|^2 + \|y_j\|^2$ are unique. ∎

3. Distance constraints between neighboring loci. In section 2, we showed that the 3D genome configuration is not identifiable from pairwise distance constraints available from typical (unphased) contact frequency maps. In order to gain identifiability, we next consider adding other biological constraints to the problem formulation that are generally available or can be measured. In particular, since DNA can be viewed as a string of connected beads, we use the distance between adjacent beads as an additional constraint. The distance between neighboring beads can be derived empirically, for example, from imaging studies [32, 25]; see also our experimental results in section 6. The additional mathematical constraints are

$$\|x_i - x_{i+1}\| = \|x_i^* - x_{i+1}^*\| \text{ and } \|y_i - y_{i+1}\| = \|y_i^* - y_{i+1}^*\| \text{ for } 1 \le i \le n-1,$$

where $x_1^*, x_2^*, \dots, x_n^*$ and $y_1^*, y_2^*, \dots, y_n^*$ correspond to consecutive beads on homologous chromosomes; see Figure 2.

In this section we show the following results: under the additional distance constraints between neighboring loci, we prove that identifiability can be obtained in the two-dimensional (2D) setting (Proposition 3.1). However, in the 3D setting we prove that there are still infinitely many 3D configurations even with these additional distance constraints (Proposition 3.2).

For the proofs of Proposition 3.1 and Proposition 3.2 we recall from Theorem 2.1 that (x_i, y_i) and (x_i^*, y_i^*) are diametrically opposite points on the same sphere. Denote the ith sphere by S_i, and let it have center c_i and radius r_i. Then $\|c_i - x_i\| = r_i$, and $2c_i - x_i = y_i$.

Proposition 3.1. *For $n \ge 3$ and generic $(x_1^*, \dots, x_n^*, y_1^*, \dots, y_n^*) \in (\mathbb{R}^2)^{2n}$, there is a unique point $(x_1, \dots, x_n, y_1, \dots, y_n) \in (\mathbb{R}^2)^{2n}$ satisfying the equations*

$$\begin{aligned} x_i + y_i = x_i^* + y_i^* \text{ and } \|x_i\|^2 + \|y_i\|^2 = \|x_i^*\|^2 + \|y_i^*\|^2 \text{ for } 1 \le i \le n, \\ \|x_i - x_{i+1}\| = \|x_i^* - x_{i+1}^*\| \text{ and } \|y_i - y_{i+1}\| = \|y_i^* - y_{i+1}^*\| \text{ for } 1 \le i \le n-1. \end{aligned} \tag{3.1}$$

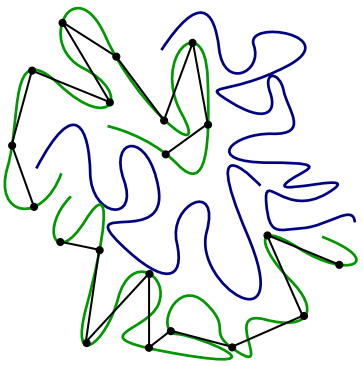


Figure 2. *Distance constraints between neighboring beads. Green and blue curves depict two homologous pairs of chromosomes. For the green curves the distances between neighboring genomic regions are shown by black lines.*

Proof. We have $y_1 = 2c_1 - x_1$ and $y_2 = 2c_2 - x_2$. Plugging this into $\|y_1 - y_2\| = \|y_1^* - y_2^*\|$ gives

$$\begin{aligned}\|y_1^* - y_2^*\| &= \|(2c_1 - x_1) - (2c_2 - x_2)\|^2 \\ &= \|(2c_1 - 2c_2) - (x_1 - x_2)\|^2 \\ &= \|2c_1 - 2c_2\|^2 + \|x_1 - x_2\|^2 - 2(2c_1 - 2c_2) \cdot (x_1 - x_2).\end{aligned}$$

The quantities $\|2c_1 - 2c_2\|^2$ and $\|x_1 - x_2\|^2$ are fixed. This implies that the quantity $(2c_1 - 2c_2) \cdot (x_1 - x_2)$ is fixed. Since we know $\|x_1 - x_2\|$ and $c_1 \neq c_2$ holds by genericness, then there are two possible angles for $x_1 - x_2$ (this is where we use the 2D constraint), and thus there are two possible solutions for $x_1 - x_2$.

Because x_1, x_2 are constrained to lie on circles, the solutions for x_1 are the intersection points of the first circle and the second circle translated by $x_1 - x_2$, and the solutions for x_2 are the intersection points of the second circle and the first circle translated by $x_2 - x_1$. Hence each solution for $x_1 - x_2$ leads to at most two possible solutions for (x_1, x_2). In turn this implies there are at most four solutions for x_2.

We now investigate the four solutions. The first two solutions are obtained by translating the circle centered at c_1 by $x_2^* - x_1^*$ and intersecting it with the circle centered at c_2; see Figure 3. One of the two solutions is x_2^*. The other two solutions are reflections of these two solutions over the line from c_1 to c_2.

Let x_1^*, x_2^*, c_1, and c_2 be fixed. They determine four possible solutions for x_2. We will show that these four solutions are different from the four solutions we get from considering x_2^*, x_3^*, c_2, and c_3 for generic x_3^* and c_3 (apart from x_2^*).

If either of the reflected solutions over the line from c_2 to c_3 coincides with one of the four original solutions, then we can perturb c_3 away from the line from c_2 to c_3 to change these solutions. If the solution that is the intersection point of the circle centered at c_2 and the translation by $x_2^* - x_3^*$ of the circle centered at c_3 (different from x_2^*) coincides with one of the four original solutions, then we can perturb x_3^*. This changes $x_2^* - x_3^*$ and hence the second intersection point of the circle centered at c_2 and the translation by $x_2^* - x_3^*$ of the circle centered at c_3.

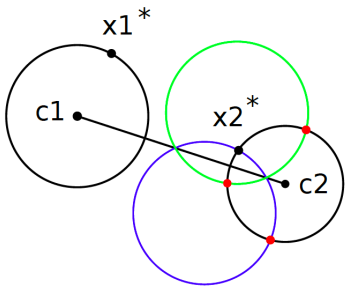


Figure 3. *Identifiability in the 2D setting with neighboring distance constraints. Two solutions for x_2 are obtained by translating the circle centered at c_1 by $x_2^* - x_1^*$ (this new circle is colored blue) and intersecting it with the circle centered at c_2. The other two solutions are obtained by reflecting the blue circle over the line through c_1 and c_2 (this new circle is colored green) and intersecting it with the circle centered at c_2. The true solution for x_2 is colored black, and the three alternative solutions for x_2 are colored red.*

A similar argument can be used to show that $x_3, \dots, x_{n-1}$ have unique solutions. Given a unique solution for x_2, there are two solutions for x_1 if and only if x_2^* lies on the line from c_1 to c_2. This is, however, not a generic configuration. A similar argument applies for x_n. ∎

Despite having uniqueness in two dimensions, we do not have uniqueness in three dimensions as shown in the following proposition.

Proposition 3.2. *Let $n \in \mathbb{N}$. For generic $(x_1^*, \dots, x_n^*, y_1^*, \dots, y_n^*) \in (\mathbb{R}^3)^{2n}$, there are infinitely many points $(x_1, \dots, x_n, y_1, \dots, y_n) \in (\mathbb{R}^3)^{2n}$ satisfying (3.1).*

Proof. If $n = 1$, then x_1^*, y_1^* can be chosen randomly with the constraint that $x_1^* \neq y_1^*$. Then x_1 and y_1 can be any points on the sphere S_1 defined by x_1^*, y_1^*. Now assume that $n \geq 2$. Fix any x_1^*, y_1^* such that $x_1^* \neq y_1^*$. Choose two circles C_1 and C_1' on the sphere S_1 defined by x_1^*, y_1^* that intersect at two points, where one is x_1^*. The circle C_1 is the intersection of S_1 and another sphere T_1. Let x_2^* be the center of the sphere T_1. Let C_1'' be the circle on S_1 that consists of points antipodal to C_1'. Then C_1'' is also an intersection of S_1 and another sphere T_2''. Let y_2^* be the center of the sphere T_2''. We use the same procedure to construct x_3^* and y_3^* from x_2^* and y_2^*, x_4^* and y_4^* from x_3^* and y_3^*, etc.

The only condition on x_1^* and y_1^* is $x_1^* \neq y_1^*$; hence (x_1^*, y_1^*) is a generic point in $\mathbb{R}^3 \times \mathbb{R}^3$. The condition that C_1 is a circle on the sphere S_1 containing x_1^* is equivalent to x_2^* being any point in $\mathbb{R}^3$ outside the line through x_1^* and y_1^*. Similarly, the condition that C_1' is a circle on the sphere S_1 containing x_1^* is equivalent to y_2^* being any point in $\mathbb{R}^3$ outside the line through x_1^* and y_1^*. The condition that C_1 and C_1' intersect at two different points of S_1 is equivalent to the normal vector of the tangent plane of S_1 at x_1^* and the normal vectors of the planes defined by C_1 and C_1' being linearly independent. Hence $(x_1^*, x_2^*, y_1^*, y_2^*)$ is a generic point in $(\mathbb{R}^3)^4$. Similar arguments can be used to show that $(x_1^*, x_2^*, \dots, x_n^*, y_1^*, y_2^*, \dots, y_n^*)$ is a generic point in $(\mathbb{R}^3)^{2n}$.

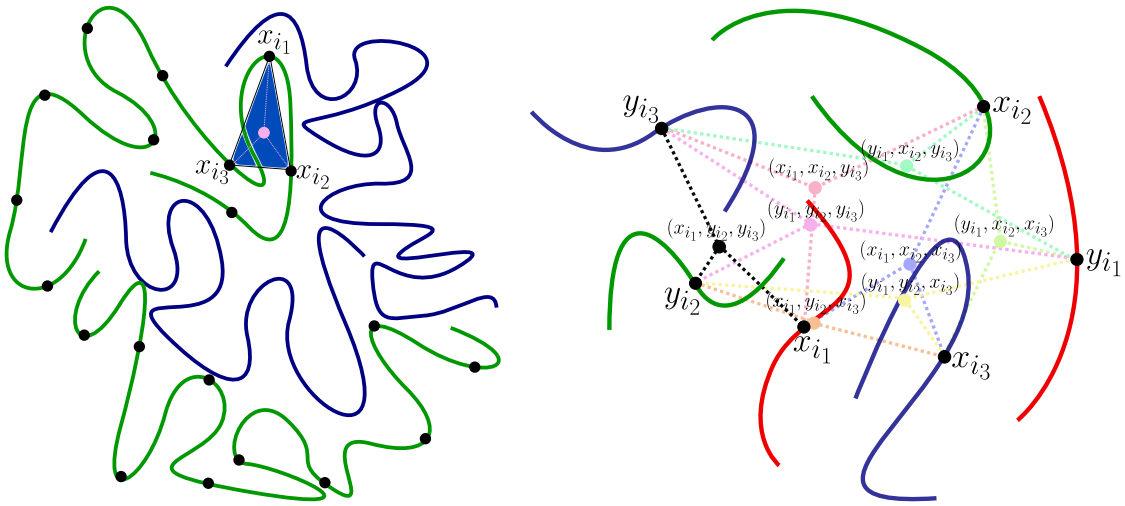
Now consider points x_n and y_n in an ε-neighborhood of x_n^* and y_n^*. Consider the spheres that are centered at x_n and y_n and have radii $\|x_{n-1}^* - x_n^*\|$ and $\|y_{n-1}^* - y_n^*\|$. The intersections of these spheres with S_{n-1} give circles $\tilde{C}_{n-1}$ and $\tilde{C}_{n-1}''$ that are perturbations of circles C_{n-1} and C_{n-1}''. In particular, the intersection of the circle $\tilde{C}_{n-1}$ and the circle $\tilde{C}_{n-1}'$ that consists of points antipodal to $\tilde{C}_{n-1}''$ consists of two points for ε small enough. Choosing x_{n-1} to be the intersection point corresponding to x_{n-1}^* and y_{n-1} its antipodal gives points x_{n-1}, y_{n-1} satisfying $\|x_{n-1} - x_n\| = \|x_{n-1}^* - x_n^*\|$ and $\|y_{n-1} - y_n\| = \|y_{n-1}^* - y_n^*\|$.

Assuming that ε is small enough, then x_{n-1} and y_{n-1} are in small neighborhoods of x_{n-1}^* and y_{n-1}^*, and we can continue the same procedure to find x_{n-2} and y_{n-2} from x_{n-1} and y_{n-1}, x_{n-3} and y_{n-3} from x_{n-2} and y_{n-2}, etc. In particular, we can find $x_1, \dots, x_{n-1}, y_1, \dots, y_{n-1}$ satisfying (3.1) for every x_n and y_n in an ε-neighborhood of x_n^* and y_n^*. ∎

The previous proposition suggests that there are two degrees of freedom for choosing $x_1, \dots, x_n, y_1, \dots, y_n$ on each homologous pair and thus that finite identifiability requires two additional algebraically independent constraints per homologous pair. Similarly this suggests that unique identifiability requires three additional algebraically independent constraints per homologous pair, where each endpoint of a chromosome needs to be included in at least one of the additional constraints.

4. Identifiability from higher-order contact constraints. In section 3, we showed that considering distances between neighboring beads only yields identifiability in two dimensions but not in three dimensions. In the following, we consider adding further constraints that are becoming widely available from experimental data, namely, higher-order contact frequencies between three or more loci as measured by experimental assays such as SPRITE [39], C-walks [36], and GAM [3]. We express these constraints mathematically by letting $F \in \mathbb{R}^{m\times m\times\cdots\times m}$ be a contact frequency tensor, where $F_{x_{i_1},x_{i_2},\ldots,x_{i_k}}$ measures the contact frequency between loci $i_1, i_2, \ldots, i_k$ with coordinates $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$. In the unphased setting, we can only measure a combination of contact frequencies over the homologous loci $\{x_{i_1}, y_{i_1}\} \times \{x_{i_2}, y_{i_2}\} \times \cdots \times \{x_{i_k}, y_{i_k}\}$, which we denote by $F_{i_1 i_2 \ldots i_k}$. In addition, as for 2-way interactions, we turn contact frequencies $F_{i_1 i_2 \ldots i_k}$ into squared "distances" $D_{i_1 i_2 \ldots i_k}$.

In the following, we provide our interpretation of distances in the higher-order setting. For simplicity, we first describe the higher-order distances for three loci in the phased setting. Since $F_{x_{i_1}x_{i_2}x_{i_3}}$ counts how often the three loci come together, we interpret $D_{x_{i_1}x_{i_2}x_{i_3}}$ as the sum of the squared distances of the three loci x_{i_1}, x_{i_2}, and x_{i_3} to their centroid (Figure 4(a)). We next provide a generalization to the unphased setting. For three homologous loci $(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2})$, and (x_{i_3}, y_{i_3}), their contact frequency can be formed by 8 possible triples, namely, $(x_{i_1}, x_{i_2}, x_{i_3})$, $(x_{i_1}, x_{i_2}, y_{i_3})$, $(x_{i_1}, y_{i_2}, x_{i_3})$, $(y_{i_1}, x_{i_2}, x_{i_3})$, $(x_{i_1}, y_{i_2}, y_{i_3})$, $(y_{i_1}, y_{i_2}, x_{i_3})$, $(y_{i_1}, x_{i_2}, y_{i_3})$, and $(y_{i_1}, y_{i_2}, y_{i_3})$. We will assume that one of the triples constitutes



(a) (b)

Figure 4. *Higher-order distance constraints.* (a) *Three loci x_{i_1}, x_{i_2}, and x_{i_3}, located on the same chromosome, are depicted. In the phased setting, the higher-order distance $D_{x_{i_1}x_{i_2}x_{i_3}}$ is defined as the sum of the squared distances (pink dashed lines) of the three loci x_{i_1}, x_{i_2}, and x_{i_3} to their centroid (pink point). Green and blue curves depict two different chromosomes.* (b) *Illustration of the definition of $D_{i_1 i_2 i_3}$ in the unphased setting. Green, blue, and red curves depict neighborhoods around three homologous loci $(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2})$, and (x_{i_3}, y_{i_3}). From these homologous loci 8 possible higher-order distances can be defined (colored dashed lines) based on the 8 centroids depicted in the figure. The higher-order distance $D_{i_1 i_2 i_3}$ is defined as the minimum of these 8 distances (achieved here by the three black dashed line segments).*

the majority of the observed contact frequency count, and hence we define $D_{i_1 i_2 i_3}$ as the minimum over all 8 higher-order distances. This is illustrated in Figure 4(b). Generalizing from three to k loci, our (squared) higher-order distance definition then becomes

$$D_{i_1 i_2 \dots i_k} = \min_{z_{i_j} \in \{x_{i_j}, y_{i_j}\}} \left(\sum_{j=1}^{k} \| z_{i_j} - (z_{i_1} + \cdots + z_{i_k})/k \|^2 \right). \tag{4.1}$$

For $k = 2$, the square root of the expression inside the brackets in (4.1) is equal to the Euclidean distance scaled by $1/\sqrt{2}$.

In the following, we prove our main result; namely, we show that the distance constraints of order 3 (3-way distances) together with the previously considered pairwise distance constraints and distance constraints among consecutive beads result in unique identifiability of the 3D genome configuration (Theorem 4.1). In fact, only very few order 3 distance constraints are required for unique identifiability. As we show in Theorem 4.1 it is sufficient that the first and last bead of each chromosome be contained in an order 3 distance constraint. This is a reasonable constraint given that methods such as SPRITE, C-walks, and GAM measure higher-order interactions over the whole genome. These insights are of interest experimentally since they suggest that the methods can restrict the measurement of such higher-order constraints to the first and last beads of each chromosome, known as telomeres.

For our biological application of interest, in the following theorem m denotes the number of chromosome pairs, $n_1, n_2, \dots, n_m$ denotes the number of domains on chromosomes $1, 2, \dots, m$, and $n = n_1 + n_2 + \cdots + n_m$ is the total number of domains. The labels of domains at the beginning and at the end of each chromosome are $1, n_1, n_1 + 1, n_1 + n_2, \dots, n_1 + n_2 + \cdots + n_{m-1} + 1, n$.

Theorem 4.1. *Let $m \in \mathbb{N}$, let $n_1, n_2, \dots, n_m \in \mathbb{N}$, and define $n := n_1 + n_2 + \cdots + n_m$. Let $I \subseteq [n] \times [n] \times [n]$ be such that each of $1, n_1, n_1 + 1, n_1 + n_2, \dots, n_1 + n_2 + \cdots + n_{m-1} + 1, n$ is contained in at least one triple in I. Let $x_1^*, \dots, x_n^*, y_1^*, \dots, y_n^* \in \mathbb{R}^3$ be fixed such that*

$$\min_{z_i^* \in \{x_i^*, y_i^*\} \text{ for } i = k_1, k_2, k_3} \left(\sum_{j \in \{k_1, k_2, k_3\}} \| z_j^* - (z_{k_1}^* + z_{k_2}^* + z_{k_3}^*)/3 \|^2 \right) = 0$$
$$\text{for } (k_1, k_2, k_3) \in I.$$

Consider the polynomial system:

$$\begin{gathered} x_i + y_i = x_i^* + y_i^* \text{ and } \|x_i\|^2 + \|y_i\|^2 = \|x_i^*\|^2 + \|y_i^*\|^2 \text{ for } 1 \le i \le n, \\ \|x_i - x_{i+1}\| = \|x_i^* - x_{i+1}^*\|, \text{ and } \|y_i - y_{i+1}\| = \|y_i^* - y_{i+1}^*\| \\ \text{for } i \in [n] \backslash \{n_1, n_1 + n_2, \dots, n\}, \\ \min_{z_i \in \{x_i, y_i\} \text{ for } i = k_1, k_2, k_3} \left(\sum_{j \in \{k_1, k_2, k_3\}} \| z_j - (z_{k_1} + z_{k_2} + z_{k_3})/3 \|^2 \right) = 0 \\ \text{for } (k_1, k_2, k_3) \in I. \end{gathered} \tag{4.2}$$

Then for generic $x_1^*, \dots, x_n^*, y_1^*, \dots, y_n^*$, this system has a unique solution in $(\mathbb{R}^3)^{2n}$.

To prove Theorem 4.1, we will need two lemmas. Lemma 4.2 states that for a fixed solution (x_1^*, y_1^*) on a sphere S_1 and given distances between solutions on S_1 and S_2, there are finitely many solutions (x_2, y_2) on the sphere S_2. Lemma 4.3 is an extension of Lemma 4.2. It states that if one has finitely many solutions on a sphere S_i, then given distances between neighboring beads, there are finitely many solutions on any sphere connected to S_i.

Lemma 4.2. *Let $x_1^*, x_2^*, y_1^*, y_2^* \in \mathbb{R}^3$ be fixed. Consider the polynomial system:*

$$\begin{aligned} x_2 + y_2 = x_2^* + y_2^*, \|x_2\|^2 + \|y_2\|^2 = \|x_2^*\|^2 + \|y_2^*\|^2, \\ \|x_1^* - x_2\| = \|x_1^* - x_2^*\|, \text{ and } \|y_1^* - y_2\| = \|y_1^* - y_2^*\|. \end{aligned} \tag{4.3}$$

For generic $x_1^*, x_2^*, y_1^*, y_2^*$, this system has finitely many solutions in $(\mathbb{R}^3)^{2n}$.

Proof. The first two equations of (4.3) say that x_2, y_2 and x_2^*, y_2^* are pairs of antipodal points on the same sphere. We denote this sphere by S_2. The third equation says that x_2 is the same distance from x_1^* as x_2^* is from x_1^*. Hence x_2 must lie on the circle C_{x_2} that is the intersection of S_2 and the sphere centered at x_1^* and with radius $\|x_1^* - x_2^*\|$. The last equation says that y_2 must lie on the circle C_{y_2} that is the intersection of S_2 and the sphere centered at y_1^* with radius $\|y_1^* - y_2^*\|$. We consider the circle C'_{x_2} that consists of antipodal points to the circle C_{y_2} on the sphere S_2. The intersection of the circles C_{x_2} and C'_{x_2} gives the solutions for x_2. Unless the two circles are equal, they intersect at at most two points. Since y_2 is antipodal to x_2, then for each x_2 there is a unique y_2. The circles coincide if and only if x_1^*, y_1^*, and the center of S_2 are collinear. ∎

Lemma 4.3. *Let $x_1^*, \dots, x_n^*, y_1^*, \dots, y_n^* \in \mathbb{R}^3$ be fixed. Consider the polynomial system:*

$$\begin{aligned} x_i + y_i = x_i^* + y_i^* \text{ and } \|x_i\|^2 + \|y_i\|^2 = \|x_i^*\|^2 + \|y_i^*\|^2 \text{ for } 2 \le i \le n, \\ \|x_1^* - x_2\| = \|x_1^* - x_2^*\|, \|y_1^* - y_2\| = \|y_1^* - y_2^*\|, \\ \|x_i - x_{i+1}\| = \|x_i^* - x_{i+1}^*\|, \text{ and } \|y_i - y_{i+1}\| = \|y_i^* - y_{i+1}^*\| \text{ for } 2 \le i \le n-1. \end{aligned}$$

For generic $x_1^*, \dots, x_n^*, y_1^*, \dots, y_n^*$, this system has finitely many solutions in $(\mathbb{R}^3)^{2n-2}$.

Proof. By Lemma 4.2, there are finitely many antipodal pairs $(x_2, y_2) \in \mathbb{R}^3 \times \mathbb{R}^3$ on S_2 such that $\|x_1^* - x_2\| = \|x_1^* - x_2^*\|$ and $\|y_1^* - y_2\| = \|y_1^* - y_2^*\|$. Similarly, for each of these antipodal pairs $(x_2, y_2) \in \mathbb{R}^3 \times \mathbb{R}^3$ on S_2, there are finitely many antipodal pairs $(x_3, y_3) \in \mathbb{R}^3 \times \mathbb{R}^3$ on S_3 satisfying $\|x_2 - x_3\| = \|x_2^* - x_3^*\|$ and $\|y_2 - y_3\| = \|y_2^* - y_3^*\|$, etc. ∎

Proof of Theorem 4.1. We recall that the first line of the polynomial system (4.2) gives that x_i, y_i are antipodal points on a sphere S_i. Consider a triple $(k_1, k_2, k_3) \in I$ that contains 1 and the equation on the last line of the polynomial system (4.2) corresponding to this triple. This equation gives that z_{k_1}, z_{k_2}, and z_{k_3}, where $z_i \in \{x_i, y_i\}$, coincide. Hence z_{k_1}, z_{k_2}, and z_{k_3} lie on the intersection of S_{k_1}, S_{k_2}, and S_{k_3}. Generically, if the intersection of three spheres is nonempty in $\mathbb{R}^3$, then it consists of two points P and P'. This gives four possible solutions for x_1, y_1: the points P, P' and their antipodals on S_1. By Lemma 4.3, there are finitely many solutions for $x_2, \dots, x_{n_1}, y_2, \dots, y_{n_1}$ given these fixed solutions x_1, y_1 on S_1. In the next two paragraphs we will show that generically these finitely many solutions do not contain antipodal points on any of the spheres $S_2, \dots, S_{n_1}$.

If there are two antipodal solutions on S_i, then we may assume that they come either from the same solution on S_1 or the antipodal solutions on S_1 because we can perturb S_{k_1}, S_{k_2}, and S_{k_3} slightly to change the other pair of solutions. First we will show that generically a solution for x_i on S_i does not give a pair of antipodal solutions for x_{i+1} on S_{i+1}. If this was the case, then both the solution for x_i and its antipodal would have to lie on the plane that is perpendicular to the line through the antipodal pair of solutions for x_{i+1} on S_{i+1}. This plane contains the centers of S_i and S_{i+1}. Hence for a solution for x_i, there is only one antipodal pair of solutions on S_{i+1}. Thus for a generic distance between the solutions on S_i and S_{i+1}, a solution on S_i does not give an antipodal pair of solutions on S_{i+1}.

Secondly, suppose that two different solutions on S_i give a pair of antipodal solutions on S_{i+1}. We will show that when we perturb the distance between solutions on S_i and S_{i+1}, we do not get an antipodal pair anymore. Let x_i and x_i' be two different solutions on S_i that give solutions x_{i+1} and $2c_{i+1} - x_{i+1}$ on S_{i+1}. Hence $\|2c_{i+1} - x_{i+1} - x_i'\|^2 = \|x_{i+1} - x_i\|^2$. We want to show that generically

$$\|2c_{i+1} - (x_{i+1} + \epsilon) - x_i'\|^2 \neq \|x_{i+1} + \epsilon - x_i\|^2,$$

where $x_{i+1} + \epsilon$ is the perturbed solution. Indeed, using the identity $\|x_{i+1} - x_i\|^2 = \|2c_{i+1} - x_{i+1} - x_i'\|^2$ gives

$$\|2c_{i+1} - (x_{i+1} + \epsilon) - x_i'\|^2 - \|x_{i+1} + \epsilon - x_i\|^2 = 2\epsilon(x_i + x_i' - 2c_{i+1}).$$

This quantity is equal to zero if and only if $\epsilon = 0$ or c_{i+1} is the middle point of the line segment from x_i to x_i'. This is generically not the case.

Using a triple $(k_1', k_2', k_3') \in I$ containing n_1 and the equation for this triple, we get four possible solutions for x_{n_1}, y_{n_1}. Generically, only one of them coincides with the finitely many solutions on S_{n_1} that we get from the solutions on S_1 because perturbing the spheres slightly (while keeping the coinciding points fixed) perturbs the second intersection point of the three spheres, and we know that generically the finitely many points do not contain antipodal points.

The unique solution on S_{n_1} comes from one solution on each of the spheres $S_1, \ldots, S_{n_1-1}$: If this was not the case, then two different solutions on S_i give the same solution on S_{i+1}. By the proof of Proposition 3.1, the dot product $(c_i - c_{i+1}) \cdot (x_i - x_{i+1})$ is fixed. Hence for a fixed x_{i+1}, all possible solutions for x_i lie on a hyperplane, and this hyperplane is perpendicular to $c_i - c_{i+1}$. Therefore, if two solutions on S_i give the same solution on S_{i+1}, then they lie on a hyperplane perpendicular to $c_i - c_{i+1}$. By slightly perturbing the sphere S_{i+1}, this is not the case anymore, and hence generically a solution on S_{i+1} comes from a unique solution on S_i. ∎

5. Algorithms and implementation. So far, we derived a theoretical framework to establish when we have unique and finite identifiability of the 3D configuration in the noiseless setting. However, a unique solution does not necessarily mean that we can find it efficiently, as in many cases finding the solution may be NP-hard. In addition, we have so far not yet considered the noisy setting. In this section, we show how to construct an optimization formulation to determine the 3D configuration efficiently.

We frame the 3D reconstruction problem as a Euclidean embedding problem, where the coordinates $x_1, \ldots, x_n, y_1, \ldots, y_n \in \mathbb{R}^3$ are inferred from distances. Similar to chromosome

semidefinite embedding (ChromSDE) [56], we formulate all distances in terms of entries in the Gram matrix G, which tracks the dot products between the $2n$ genomic regions, namely, letting the column/row i of G correspond to x_i and the column/row $n+i$ correspond to its homologous locus y_i; then the squared distances are given by $\|x_i - x_j\|^2 = G_{i,i} + G_{j,j} - 2G_{i,j}$, $\|x_i - y_j\|^2 = G_{i,i} + G_{n+j,n+j} - 2G_{i,n+j}$, and $\|y_i - y_j\|^2 = G_{n+i,n+i} + G_{n+j,n+j} - 2G_{n+i,n+j}$. It is natural to work with the Gram matrix G since it is rotation invariant. By imposing the constraint $\sum_{i,j} G_{i,j} = 0$ we can also fix the translational axis. Also the additional distance constraints that we introduced in the previous sections (Theorem 2.1, Proposition 3.1, Theorem 4.1) can be represented as linear constraints in terms of entries in G as follows:

- pairwise distances (the sum of four squared distances):

$$g_{ij}(G) := G_{i,i} + G_{j,j} + G_{n+i,n+i} + G_{n+j,n+j} - G_{i,j} - G_{n+i,j} - G_{i,n+j} - G_{n+i,n+j};$$

- (squared) distances between homologous pairs:

$$g_{ii}(G) := G_{i,i} + G_{n+i,n+i} - 2G_{i,n+i};$$

- (squared) distances between neighboring beads:

$$g_{i+}(G) := G_{i,i} + G_{i+1,i+1} - 2G_{i,i+1};$$

- (squared) distances of order 3 (can be generalized to higher orders):

$$g_{ijk}(G) := \min_l (g_{ijkl} : l = 1, \ldots, 8),$$

where

$$\begin{aligned}
g_{ijk1}(G) &:= G_{i,i} + G_{j,j} + G_{k,k} - G_{i,j} - G_{i,k} - G_{j,k},\\
g_{ijk2}(G) &:= G_{i,i} + G_{j,j} + G_{n+k,n+k} - G_{i,j} - G_{i,n+k} - G_{j,n+k},\\
g_{ijk3}(G) &:= G_{i,i} + G_{n+j,n+j} + G_{k,k} - G_{i,n+j} - G_{i,k} - G_{n+j,k},\\
g_{ijk4}(G) &:= G_{i,i} + G_{n+j,n+j} + G_{n+k,n+k} - G_{i,n+j} - G_{i,n+k} - G_{n+j,n+k},\\
g_{ijk5}(G) &:= G_{n+i,n+i} + G_{j,j} + G_{k,k} - G_{n+i,j} - G_{n+i,k} - G_{j,k},\\
g_{ijk6}(G) &:= G_{n+i,n+i} + G_{j,j} + G_{n+k,n+k} - G_{n+i,j} - G_{n+i,n+k} - G_{j,n+k},\\
g_{ijk7}(G) &:= G_{n+i,n+i} + G_{n+j,n+j} + G_{k,k} - G_{n+i,n+j} - G_{n+i,k} - G_{n+j,k},\\
g_{ijk8}(G) &:= G_{n+i,n+i} + G_{n+j,n+j} + G_{n+k,n+k}\\
&\qquad - G_{n+i,n+j} - G_{n+i,n+k} - G_{n+j,n+k}.
\end{aligned}$$

Our objective is to determine a rank 3 solution of G, satisfying the above constraints. However, this optimization problem is nonconvex due to the rank constraint, and we instead consider the standard relaxation: we minimize the trace of the Gram matrix as an approximation to the matrix rank [18]. The resulting optimization problem then becomes the following semidefinite program:

$$
\begin{aligned}
\underset{G}{\text{minimize}} \quad & \operatorname{tr}(G) \\
\text{subject to} \quad & g_{ii}(G) = D^*_{ii},\ 1 \le i \le n, \\
& g_{ij}(G) = D^*_{ij},\ 1 \le i < j \le n, \\
& g_{i+}(G) = D^*_{i+},\ i \in \Omega_1, \\
& g_{ijk}(G) = D^*_{ijk},\ (i,j,k) \in \Omega_2, \\
& \sum_{1 \le i,j \le 2n} G_{i,j} = 0, \\
& G \succeq 0.
\end{aligned} \tag{5.1}
$$

Here, D^*_{ii} denotes the squared distances between homologous pairs computed from the pairwise distances using Lemma 2.3, D^*_{ij} denotes the pairwise distances, D^*_{i+} denotes the squared distances between neighboring beads, and D^*_{ijk} denotes the squared distances between three loci. While one could also consider 4- or higher-order distance constraints, in our implementation we only used 3-way distance constraints since higher-order contacts are extremely sparse. The index set $\Omega_1 = [2n]\backslash\{n_1, n_1+n_2, \dots, n, n+n_1, n+n_1+n_2, \dots, 2n\}$ corresponds to all beads that are not the last bead on a chromosome. The index set $\Omega_2 \subseteq [n]^3$ corresponds to all triples of beads with nonzero contact frequencies.

In the noisy setting, which is relevant for biological data, we replace the equality constraints with penalties in the loss function. Namely, using D^* for the noiseless and D for the noisy distances, we replace the equality constraints of the form $g(G) = D^*$ by adding $(g(G) - D)^2$ to the objective function. For the higher-order distance constraints of the form $D^*_{ijk} = \min(g_{ijk1}(G), \dots, g_{ijk8}(G))$ for $(i,j,k) \in \Omega_2$ we use slack variables and a convex relaxation using an atomic norm that combines the ℓ_2- and ℓ_1-norms. More precisely, we propose the use of the following transformation in the noisy setting:

$$D_{ijk} + \lambda_{ijkl} = g_{ijkl}(G) + s_{ijkl} \text{ for } l = 1, 2, \dots, 8,$$

where $\lambda_{ijkl}, s_{ijkl} \ge 0$ for all i, j, k, and l act as slack variables. In general, for each triple (i,j,k) we want one of the λ_{ijkl} to be close to 0 and the sum over all s_{ijkl} to be small. Naively this can be done by placing $\sum s_{ijkl} + \sum \lambda_{ijkl}$ into the objective function. However, this would not enforce for each (i,j,k) that at least one λ_{ijkl} to be close to 0. Instead we propose to use

$$\sum_{(i,j,k)\in\Omega_2, 1\le l\le 8} s_{ijkl} + \sqrt{\sum_{(i,j,k)\in\Omega_2} \left(\sum_{1\le l\le 8} \lambda_{ijkl} \right)^2}.$$

The ℓ_2-norm will push down the $\sum_l \lambda_{ijkl}$ for each (i,j,k), while the ℓ_1-norm will drive at least one of the λ_{ijkl} to zero, which is precisely the desired behavior. The quantity $\sqrt{\sum_{i,j,k} (\sum_l \lambda_{ijkl})^2}$ is an atomic norm as defined in [7] with the set of atoms

$$\mathcal{A} = \Bigg\{ (\lambda_{ijkl}) : \sum_{i,j,k} \left(\sum_l \lambda_{ijkl} \right)^2 = 1 \text{ and } \sum_{i,j,k} \lambda^2_{ijkl_{ijk}} = 1 \text{ for } l_{ijk} = 1, \dots, 8, (i,j,k) \in \Omega_2 \Bigg\}.$$

Then the optimization problem in the noisy setting becomes

$$
\begin{aligned}
\underset{G,s,\lambda}{\text{minimize}} \quad & \rho \operatorname{tr}(G) + \sum_{1\le i\le n} (g_{ii}(G) - D_{ii})^2 + \sum_{1\le i<j\le n} (g_{ij}(G) - D_{ij})^2 \\
& + \sum_{i\in\Omega_1} (g_{i+}(G) - D_{i+})^2 + \sum_{(i,j,k)\in\Omega_2, 1\le l\le 8} s_{ijkl} \\
& + \sqrt{\sum_{(i,j,k)\in\Omega_2} \left(\sum_{1\le l\le 8} \lambda_{ijkl} \right)^2} \qquad (5.2) \\
\text{subject to} \quad & D_{ijk} + \lambda_{ijkl} = g_{ijkl}(G) + s_{ijkl},\ (i,j,k) \in \Omega_2, 1 \le l \le 8, \\
& s_{ijkl} \ge 0,\ (i,j,k) \in \Omega_2, 1 \le l \le 8, \\
& \lambda_{ijkl} \ge 0,\ (i,j,k) \in \Omega_2, 1 \le l \le 8, \\
& \sum_{1\le i,j\le 2n} G_{i,j} = 0, \\
& G \succeq 0.
\end{aligned}
$$

We use a tuning parameter ρ for the trace in the objective function, which can be used to balance obtaining a low-rank solution versus satisfying the constraints. The tuning parameter ρ can be chosen using cross-validation or by selecting it so that the resulting solution has small $(d+1)^{th}$ eigenvalue. As shown in section SM3 and section SM11, we observe on synthetic and real data that the solution is robust to the choice of ρ.

The theoretical results from Lemma 2.3 allow us to compute the distances between homologous pairs from the pairwise distances D_{ij}. We recall that we need to compute $\|v\|^2$ such that

$$\det(T' - 8J\|v\|^2) = 0,$$

where T' is an invertible matrix constructed from the pairwise distance matrix by selecting a set of $2d+2$ indices. One step of computing $\|v\|$ involves inverting T'. Even if the error in the measurements is small, noise can propagate and severely impact this computation. In order to obtain a robust estimate of homolog-homolog distances, for each locus i, we sample 100 T' matrices and obtain 100 solutions to the equation for $\|v\|^2$. We then take the median of the solutions to be the homolog-homolog distance for locus i and use these homolog-homolog distances for the evaluation of our algorithms on synthetic and real data in the following section.

To solve the two convex optimization problems presented in this section for the noiseless and noisy setting, we make use of the solver MOSEK [2] implemented in CVX within MATLAB [20, 21]. This results in the Gram matrix. In order to reconstruct the coordinates of the genomic regions from the Gram matrix, we use an eigenvector decomposition as also done in [56], namely, letting $\gamma_1, \ldots, \gamma_d$ be the top d eigenvalues and $\nu_1, \ldots, \nu_d$ be the corresponding eigenvectors of G; then

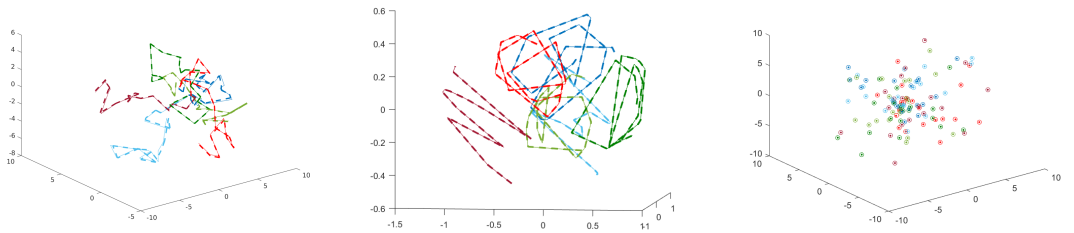
$$x_i = (\sqrt{\gamma_1} \cdot \nu_{1,i}, \ldots, \sqrt{\gamma_d} \cdot \nu_{d,i}), \text{ and } y_i = (\sqrt{\gamma_1} \cdot \nu_{1,n+i}, \ldots, \sqrt{\gamma_d} \cdot \nu_{d,n+i}) \text{ for } i = 1, \ldots, n.$$

Since we are interested in recovering the genome configuration in three dimensions, we use $d = 3$, thereby obtaining the desired 3D diploid configuration. We provide the code for our algorithm at https://github.com/uhlerlab/diploid-3D-reconstruction.

6. Evaluation on synthetic and real data.

6.1. Synthetic data. We start by testing our method on simulated data. For this we construct three different types of 3D structures: (a) a Brownian motion model using a standard normal distribution to generate successive points; (b) points sampled uniformly along a spiral with random translations sampled uniformly within $(0, 0.5)$ range and orientations sampled uniformly within $(-\frac{\pi}{4}, \frac{\pi}{4})$; (c) points sampled uniformly in a unit sphere.

Performance of our method in the noiseless setting. For the one-dimensional setting we deduced in section 2 that the pairwise distance constraints by themselves are sufficient to identify the underlying 3D configuration. For the 2D setting we proved in section 3 that knowing additionally the distances between neighboring beads leads to uniqueness. We here perform simulations in three dimensions since this is the biologically relevant setting. These results are depicted in Figure 5 with additional examples in Figure SM1. The inputs to our algorithm are the pairwise distances (which are summed over homologs), all 3-way distances, the distances between homologous loci, and the distances between neighboring beads. In the noiseless setting considered here we solve the SDP formulation in (5.1). Figure 5 and Figure SM1 show that the true and reconstructed structures highly overlap, thereby indicating that our optimization formulation is able to recover the 3D structure of the full diploid genome in the noiseless setting. When the 3-way distance constraints are removed, the reconstructions are less aligned with the true structures. This is shown in Figure 6, where we measure the root-mean-square deviation (RMSD) between true and reconstructed 3D coordinates over 20 trials. Figure SM2 provides examples of 3D structures reconstructed without tensor constraints, showing that in this case the 3D coordinates are consistently mispredicted. In line with our theoretical results, these experimental results in the noiseless setting indicate the importance of higher-order contact frequencies for recovering the 3D diploid configuration, especially when the number of chromosomes is larger. The RMSD



(a) (b) (c)

Figure 5. *Examples of true and reconstructed points on simulated data.* (a) *Brownian motion model.* (b) *Spirals.* (c) *Random points in a sphere. We generate six chromosomes with a total of* 120 *domains corresponding to three homologous pairs with* 20 *domains per chromosome in the noiseless setting. Solid lines/points correspond to true* 3*D coordinates and dashed lines/unfilled points correspond to reconstructions via our method. Each color represents a different chromosome.*

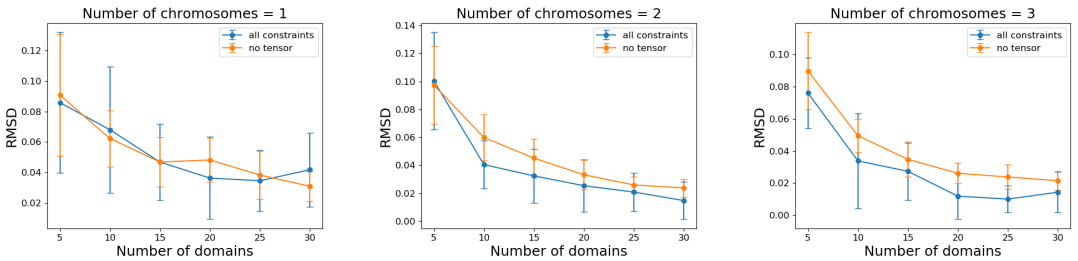


Figure 6. *Performance of our method in the noiseless setting. Root-mean-square deviation (RMSD) between the true and reconstructed structures computed with and without higher-order distance constraints. Simulated data were generated using a Brownian motion model with* (a) *one,* (b) *two, and* (c) *three chromosomes. Mean and standard deviation over* 20 *trials are shown.*

being nonzero implies that the recovery is not exact even in the presence of higher-order distance constraints. We expect this to be a consequence of considering a convex relaxation of the exact optimization problem where we minimize the trace instead of the rank of the Gram matrix.

Performance of our method in the noisy setting. Next, we consider noisy distance observations $D_{ij} = D^*_{ij}(1+\delta)$ and noisy 3-way distance observations $D_{i_1 i_2 \ldots i_k} = D^*_{i_1 i_2 \ldots i_k}(1+\delta)$ by sampling δ uniformly within $(-\epsilon, \epsilon)$ as in [56], where ϵ is a given noise level. For our simulations we sample a maximum of 1000 3-way distance constraints. As shown in Figure SM3, we observe that the number of constraints does not have a major effect on the reconstruction accuracy. While for all simulations shown in this section we set the tuning parameter $\rho = 0.000001$, Figure SM4 shows that the performance is not significantly different when using different choices of ρ. In addition, it has been observed that maximizing the trace instead of minimizing the trace of the Gram matrix may sometimes lead to better reconstructions in Euclidean distance geometry problems. Figure SM5 and Figure SM6 show that for our problem the performance does not differ significantly when $\mathrm{tr}(G)$ is minimized or maximized.

In Figure 7 we numerically assess the accuracy of our predicted structure for the Brownian motion model for different numbers of chromosomes (one, two, or three) and different numbers of domains per chromosome (10 or 20) by computing the Spearman correlation between reconstructed and true pairwise distances, similar to [56]. As expected, Figure 7 shows that when the noise level increases, then the Spearman correlation between the original and reconstructed configuration decreases. For the simulations with one chromosome, the Spearman correlation is higher for 20 domains than 10 domains. The corresponding RMSD plots are shown in Figure SM7. Finally, we compare our method to the one of [5] on the same synthetic data in the noiseless and the noisy setting; as shown in Figure SM8 and Figure SM9, our method outperforms the method in [5] in both settings.

6.2. Application to 3D diploid genome reconstruction.

We apply our algorithm to the problem of reconstructing the diploid genome from contact frequency data derived from experiments. We obtain pairwise and 3-way contact frequencies collected via SPRITE in human lymphoblastoid cells from [39]. Since we aim to reconstruct the whole diploid genome, which

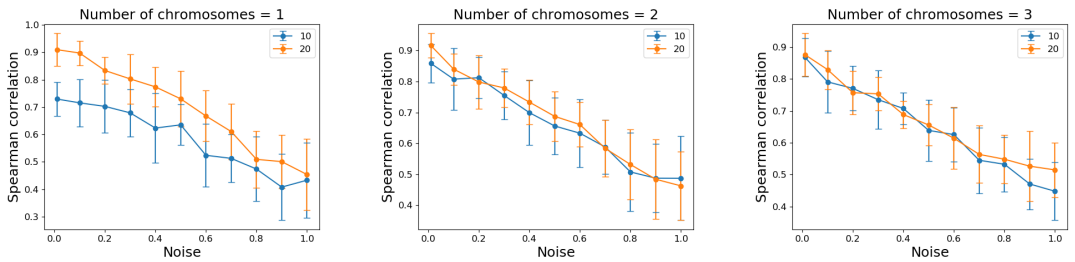


Figure 7. *Performance of our method in the noisy setting. Spearman correlation under different noise levels for* (a) *one,* (b) *two, and* (c) *three chromosomes. Simulated data were generated using a Brownian motion model, where each chromosome has* 10 *or* 20 *domains. Mean and standard deviation over* 20 *trials are shown.*

consists of approximately 6 billion base pairs, for computational reasons we bin the contact frequencies in the SPRITE data set into 10 Mega-base pair (Mb) regions. While some previous studies considered higher resolutions, the majority of the studies [5, 23, 42, 50, 56] did not attempt to reconstruct the whole diploid genome and focused only on reconstructing one chromosome, thus enabling them to consider higher resolutions.

After filtering out regions with a small number of total contacts, we obtain 514 unphased points on the chromosomes. We convert the pairwise contact frequencies to pairwise (squared) distances using the previously observed relationship $D_{ij} = (F_{ij}^{-\alpha})^2$, where $\alpha = 1/2$ [42], and use Lemma 2.3 to obtain the distances between homologous pairs from this data. We use the same conversion $D_{ijk} = (F_{ijk}^{-\alpha})^2$, where $\alpha = 1/2$, for 3-way contact frequencies and 3-way (squared) distances to be consistent with the pairwise setting. Since the conversion factor between frequencies and distances is an important parameter, especially for applications to real data, we explore additional parameters and evaluate their fit. Figure SM10 shows that the chosen conversion factor $\alpha = 1/2$ provides a good fit (low error between predicted and actual pairwise frequencies). As in our simulations in the noisy setting, we randomly sample 1000 3-way distance constraints from all nonzero 3-way contact frequencies (for the transformation from 3-way contact frequencies to 3-way distances, see section 4). Finally, we obtain the distances between neighboring 10 Mb beads by empirically evaluating the 3D reconstructions under different input distances; see section SM8, Figure SM11, and Figure SM12.

Using the pairwise constraints, homolog-homolog constraints, neighboring bead constraints, and 3-way distance constraints, we solve the SDP problem in (5.2) for the noisy setting and analyze the corresponding 3D coordinates. Our diploid reconstruction is shown in Figure 8(a), with the reconstruction of chromosome 2 illustrated in Figure 8(b) and the remaining chromosomes shown in Figure SM14. We compare this diploid genome reconstruction to the 3D structure obtained via ChromSDE, shown in Figure 8(c) obtained under the assumption that the observed contact frequencies and the corresponding distances are a sum of four equal quantities; i.e., $\|x_i - x_j\|^2, \|x_i - y_j\|^2, \|y_i - x_j\|^2$, and $\|y_i - y_j\|^2$ are equal. In Figure SM13, we show that the reconstruction obtained using ChromSDE with equal distances does not recapitulate known biology as described in the following paragraphs.

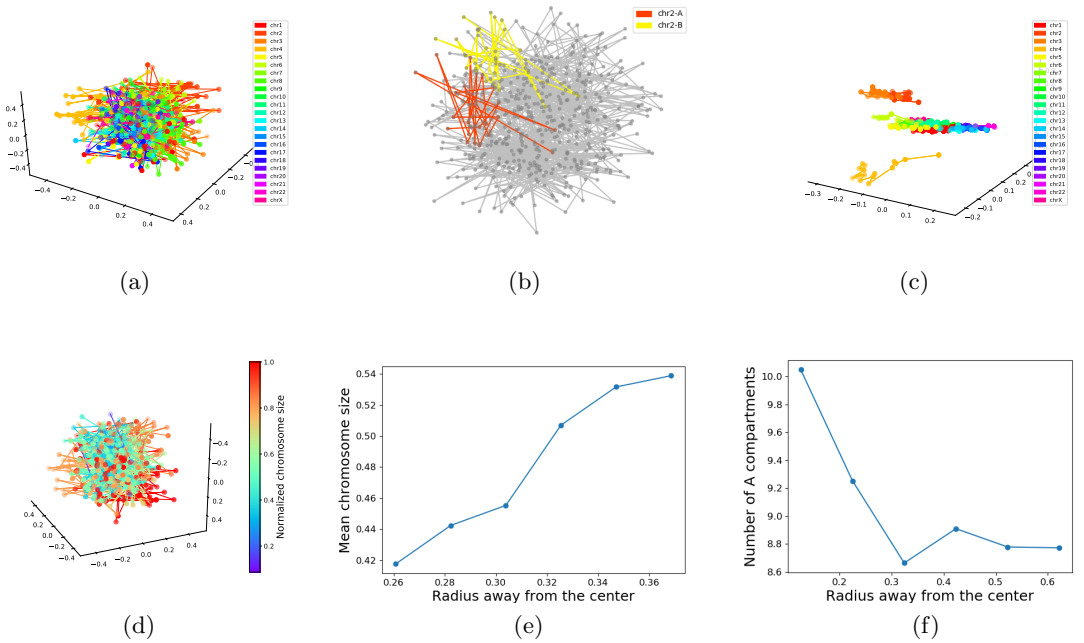


Figure 8. *3D diploid genome reconstruction. Estimated 3D positions of all chromosomes and their corresponding homologs at 10 Mb resolution. 3D positions obtained using* (a) *our method,* (b) *with chromosome 2 highlighted, and* (c) *using ChromSDE with chromosomes colored according to chromosome number.* (d) *Whole diploid organization obtained via our method, colored by chromosome size.* (e) *Mean chromosome size as the distance from the center increases.* (f) *The mean number of A compartments as the distance from the center increases.*

Experimental (imaging) studies have shown that chromosomes are organized by size within the nucleus, with small chromosomes in the interior and larger chromosomes on the periphery [4]. We colored each chromosome according to its size and computed the mean chromosome size versus distance away from the center. The results of the 3D configuration obtained using our method are shown in Figure 8(b) and 8(e) and recapitulate prior studies: smaller chromosomes are preferentially located in the center, whereas larger chromosomes are preferentially on the periphery; see also Figure SM15. This is especially apparent for chromosomes 2 and 4, which are some of the largest chromosomes, and in our reconstruction they are located on the periphery as expected.

Experimental studies on the spatial organization of the genome have also shown that the center of the nucleus is enriched in active compartments (known as A compartments), while the periphery contains inactive compartments (known as B compartments) [45]. From previously published data on the location of A and B compartments along the genome in human lymphoblastoid cells [40], we counted the number of A compartments per 10 Mb bin. Then dividing our 3D reconstruction into concentric circles of increasing radius away from the center, we found the mean number of A compartments in each concentric circle. Figure 8(f) shows that with increasing distance away from the center, the number of A compartments decreases; see also Figure SM16. Thus, our reconstruction recovers the experimentally observed trend for A compartments to be preferentially located near the nucleus center. As shown

in Figures SM17 and SM18, we note that our results are robust to the choice of the tuning parameter ρ resulting in biologically plausible configurations independent of the choice of ρ.

Currently, many studies such as [41] simply ignore the fact that the genome is diploid and infer the 3D genome organization as if the data were collected from a haploid organism, assuming that the homologous loci have the same 3D structure. However, we show in Figure SM19 that the haploid distance matrices, computed by including only one copy of each of the homologous loci, are different between the two copies with a mean Spearman correlation of only 0.08. This shows that modeling the diploid aspect of the genome provides valuable information regarding the 3D structure of each of the homologs, which may be substantially different.

7. Discussion. In this paper, we proved that for a diploid organism the 3D genome structure is not identifiable from pairwise distance measurements alone. This implies that applying any algorithm for the reconstruction of the 3D genome structure from typical chromosome conformation capture data for a diploid organism can result in any of the infinitely many configurations with the same pairwise contact frequencies. We showed that unique idenfiability is obtained using distance constraints between neighboring genomic loci as well as 3-way distance constraints in addition to the pairwise distance constraints that can be obtained from typical contact frequency data. Distances between neighboring genomic loci can be obtained empirically, e.g., from imaging studies, while 3-way distance constraints can be obtained from the most recently developed sequencing-based methods for obtaining contact frequencies such as SPRITE [39], C-walks [36], and GAM [3]. We also presented SDP formulations for determining the 3D genome reconstruction both in the noiseless and the noisy setting. Finally, we applied our algorithm to contact frequency data from human lymphoblastoid cells collected using SPRITE and showed that our results recapitulate known biological trends; in particular, in the 3D configuration identified using our method, the small chromosomes are preferentially situated in the interior of the cell nucleus, while the larger chromosomes are preferentially situated at the periphery of the cell nucleus. In addition, in the 3D configuration identified using our method the number of A domains is higher in the interior versus the periphery, which is in line with experimental results. While given the available experimental data we were only able to validate our reconstructions based on known biological trends; experimental measurements of the full 3D diploid structure of the genome are beginning to emerge [37], thereby representing an interesting area for comparison with our method as data from different cell types (such as human lymphoblastoid cells used here) become available.

Our work shows the importance of higher-order contact frequencies that can be measured using SPRITE [39], C-walks [36], and GAM [3] for obtaining the 3D organization of the genome in diploid organisms. This is particularly relevant for the reconstruction of cancer genomes, where copy number variations are frequent, and hence the genome may contain even more than two copies of each locus. We conjecture that identifiability of the 3D genome structure can also be achieved by replacing the higher-order contact constraints with distance constraints to the center of the cell nucleus. Such constraints are also biologically relevant since these distances can be measured via imaging experiments or inferred by measuring whether a particular locus is in a lamin-associated domain or a telomere, both of which tend to lie at the boundary of the cell nucleus [11, 22, 49]. Another future research direction is the development of specialized

solvers to enable the reconstruction of the genome at higher resolution. In this study we used a 10 Mbp resolution due to the computational constraints imposed by SDP solvers. Another approach for enabling the reconstruction of the genome at higher resolution is to use 2-step procedures like the one in [43]. It would also be interesting to consider other convex relaxations for the reconstruction, e.g., based on the sum of squares method [33]. Finally, the theoretical results in this paper build on the assumption that distances are inverses of square roots of pairwise and higher-order contact frequencies. An interesting future research direction is to develop a method for estimating the map between higher-order contact frequencies and distances and then to prove identifiability as well as build reconstruction algorithms for these different maps.

Acknowledgments. We thank Mohab Safey El Din for helpful discussions as well as Mark Segal and two anonymous reviewers for their insightful comments, which helped improve our paper.

REFERENCES

[1] A. Y. Alfakih, A. Khandani, and H. Wolkowicz, *Solving Euclidean distance matrix completion problems via semidefinite programming*, Comput. Optim. Appl., 12 (1999), pp. 13–30.
[2] M. ApS, *The MOSEK Optimization Toolbox for MATLAB Manual*, version 9.0, http://docs.mosek.com/9.0/toolbox/index.html, 2019.
[3] R. A. Beagrie, A. Scialdone, M. Schueler, D. C. A. 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. W. Edwards, M. Nicodemi, and A. Pombo, *Complex multi-enhancer contacts captured by genome architecture mapping*, Nature, 543 (2017), pp. 519–524.
[4] 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 Biol., 3 (2005), e157.
[5] A. G. Cauer, G. Yardimci, J.-P. Vert, N. Varoquaux, and W. S. Noble, *Inferring Diploid 3D Chromatin Structures From Hi-C Data*, preprint, bioRxiv:644294, 2019.
[6] L. Cayton and S. Dasgupta, *Robust Euclidean embedding*, in Proceedings of the 23rd International Conference on Machine Learning, ACM, New York, 2006, pp. 169–176.
[7] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The convex geometry of linear inverse problems*, Found. Comput. Math., 12 (2012), pp. 805–849.
[8] The 1,000 Genome Project Consortium, *An integrated map of genetic variation from* 1,092 *human genomes*, Nature, 491 (2012), pp. 56–65.
[9] The 1,000 Genome Project Consortium, *A global reference for human genetic variation*, Nature, 526 (2015), pp. 68–74.
[10] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, CRC Press, New York, 2000.
[11] L. Crabbe, A. J. Cesare, J. M. Kasuboski, J. A. J. Fitzpatrick, and J. Karlseder, *Human telomeres are tethered to the nuclear envelope during postmitotic nuclear assembly*, Cell Rep., 2 (2012), pp. 1521–1529.
[12] J. Dekker, *Gene regulation in the third dimension*, Science, 319 (2008), pp. 1793–1794.
[13] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, *Capturing chromosome conformation*, Science, 295 (2002), pp. 1306–1311.
[14] M. Di Pierro, B. Zhang, E. L. Aiden, P. G. Wolynes, and J. N. Onuchic, *Transferable model for chromosome architecture*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. 12168–12173.
[15] 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 (2010), pp. 363–367, http://dx.doi.org/10.1038/nature08973.

[16] H. Fang and D. P. O'Leary, *Euclidean distance matrix completion problems*, Optim. Methods Softw., 27 (2012), pp. 695–717.
[17] M. Fazel, H. Hindi, and S. P. Boyd, *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, Piscataway, NJ, 2003, pp. 2156–2162.
[18] M. Fazel, H. Hindi, and S. P. Boyd, *A rank minimization heuristic with application to minimum order system approximation*, in Proceedings of the 2001 American Control Conference, vol. 6, IEEE, Piscataway, NJ, 2001, pp. 4734–4739.
[19] G. Fudenberg and M. Imakaev, *FISH-ing for captured contacts: Towards reconciling FISH and 3C*, Nat. Methods, 14 (2017), pp. 673–678.
[20] M. Grant and S. Boyd, *Graph implementations for nonsmooth convex programs*, in Recent Advances in Learning and Control, Lect. Notes Control Inf. Sci., V. Blondel, S. Boyd, and H. Kimura, eds., Springer-Verlag Limited, Berlin, Heidelberg, 2008, pp. 95–110.
[21] M. Grant and S. Boyd, *CVX: Matlab Software for Disciplined Convex Programming*, version 2.1, http://cvxr.com/cvx, 2014.
[22] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel, *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions*, Nature, 453 (2008), pp. 948–951.
[23] 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 (2013), e1002893.
[24] 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 Genet., 46 (2014), pp. 205–212.
[25] R. Jungmann, M. S. Avendaño, J. B. Woehrstein, M. Dai, W. M. Shih, and P. Yin, *Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT*, Nat. Methods, 11 (2014), pp. 313–338.
[26] N. Krislock, *Semidefinite Facial Reduction for Low-Rank Euclidean Distance Matrix Completion*, Ph.D. thesis, University of Waterloo, 2010, http://hdl.handle.net/10012/5093.
[27] A. Lesne, J. Riposo, P. Roger, A. Cournac, and J. Mozziconacci, *3D genome reconstruction from chromosomal contacts*, Nat. Methods, 11 (2014), pp. 1141–1143.
[28] 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 (2009), pp. 289–293.
[29] F. Lu, S. Keleş, S. J. Wright, and G. Wahba, *Framework for kernel regularization with application to protein clustering*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 12332–12337.
[30] L. A. Mirny, *The fractal globule as a model of chromatin architecture in the cell*, Chromosome Res., 19 (2011), pp. 37–51.
[31] B. Mishra, G. Meyer, and R. Sepulchre, *Low-rank optimization for distance matrix completion*, in Proceedings of the 2011 50th IEEE Conference on Decision and Control and European Control Conference, IEEE, Piscataway, NJ, 2011, pp. 4455–4460.
[32] I. Müller, S. Boyle, R. H. Singer, W. A. Bickmore, and J. R. Chubb, *Stable morphology, but dynamic internal reorganisation, of interphase human chromosomes in living cells*, PloS One, 5 (2010), e11560.
[33] J. Nie, *Sum of squares method for sensor network localization*, Comput. Optim. Appl., 43 (2009), pp. 151–179.
[34] G. Nir, I. Farabella, C. P. Estrada, C. G. Ebeling, B. J. Beliveau, H. M. Sasaki, S. Dean Lee, S. C. Nguyen, R. B. McCole, S. Chattoraj, J. Ercej, J. Alhaj Abed, N. M. C. Martins, H. Q. Nguyen, M. A. Hannan, S. Russell, N. C. Durand, S. S. P. Rao, J. Y. Kishi, P. Soler-Vila, M. Di Pierro, J. N. Onuchic, S. P. Callahan, J. M. Schreiner, J. A. Stuckey, P. Yin, E. L. Aiden, M. A. Marti-Renom, and C.-T. Wu, *Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling*, PLoS Genet., 14 (2018), e1007872.

[35] H. K. Norton and J. E. Phillips-Cremins, *Crossed wires: 3D genome misfolding in human disease*, J. Cell Biol., 216 (2017), pp. 3441–3452.
[36] P. Olivares-Chauvet, Z. Mukamel, A. Lifshitz, O. Schwartzman, N. O. Elkayam, Y. Lubling, G. Deikus, R. P. Sebra, and A. Tanay, *Capturing pairwise and multi-way chromosomal conformations using chromosomal walks*, Nature, 540 (2016), pp. 296–300.
[37] A. C. Payne, Z. D. Chiang, P. L. Reginato, S. M. Mangiameli, E. M. Murray, C.-C. Yao, S. Markoulaki, A. S. Earl, A. S. Labade, R. Jaenisch, G. M. Church, E. S. Boyden, J. D. Buenrostro, and F. Chen, *In situ genome sequencing resolves DNA sequence and structure in intact biological samples*, Science, 371 (2021), eaay3446.
[38] Y. Qi and B. Zhang, *Predicting three-dimensional genome organization with chromatin states*, PLoS Comput. Biol., 15 (2019), e1007024.
[39] 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. MeDonel, M. Garber, and M. Guttman, *Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus*, Cell, 174 (2018), pp. 744–757.
[40] S. S. P. Rao, M. H. Huntley, N. C. Durand, 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, 159 (2014), pp. 1665–1680.
[41] L. Rieber and S. Mahony, *miniMDS: 3D structural inference from high-resolution Hi-C data*, Bioinformatics, 33 (2017), pp. i261–i266.
[42] M. Rousseau, J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette, *Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling*, BMC Bioinformatics, 12 (2011), article number 414.
[43] M. R. Segal and H. L. Bengtsson, *Reconstruction of 3D genome architecture via a two-stage algorithm*, BMC Bioinformatics, 16 (2015), pp. 1–10.
[44] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C)*, Nature Genet., 38 (2006), pp. 1348–1354.
[45] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O'Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sansó, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klenerman, and E. D. Laue, *3D structures of individual mammalian genomes studied by single-cell Hi-C*, Nature, 544 (2017), pp. 59–64.
[46] E. Tuzhilina, T. J. Hastie, and M. R. Segal, *Principal curve approaches for inferring 3D chromatin architecture*, Biostatistics, 2020, kxaa046.
[47] C. Uhler and G. V. Shivashankar, *Chromosome intermingling: Mechanical hotspots for genome regulation*, Trends Cell Biol., 27 (2017), pp. 810–819.
[48] C. Uhler and G. V. Shivashankar, *The regulation of genome organization and gene expression by nuclear mechanotransduction*, Nat. Rev. Mol. Cell Biol., 18 (2017), pp. 717–727.
[49] B. Van Steensel and A. S. Belmont, *Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression*, Cell, 169 (2017), pp. 780–791.
[50] N. Varoquaux, F. Ay, W. S. Noble, and J.-P. Vert, *A statistical approach for inferring the 3D structure of the genome*, Bioinformatics, 30 (2014), pp. i26–i33.
[51] N. Varoquaux, W. S. Noble, and J.-P. Vert, *Inference of Genome 3D Architecture by Modeling Overdispersion of Hi-C Data*, preprint bioRxiv 2021.02.04.429864, 2021.
[52] H. Wang, X. Xu, C. M. Nguyen, Y. Liu, Y. Gao, X. Lin, T. Daley, N. H. Kipniss, M. La Russa, and L. S. Qi, *CRISPR-mediated programmable 3D genome positioning and nuclear organization*, Cell, 175 (2018), pp. 1405–1417.
[53] 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 (2016), pp. 598–602.
[54] K. Q. Weinberger, F. Sha, Q. Zhu, and L. K. Saul, *Graph Laplacian regularization for large-scale semidefinite programming*, in Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, MA, 2007, pp. 1489–1496.

[55] L. Zhang, G. Wahba, and M. Yuan, *Distance shrinkage and Euclidean embedding via regularized kernel estimation*, J. R. Stat. Soc. Ser. B Stat. Methodol., 78 (2016), pp. 849–867.
[56] Z. Zhang, G. Li, K.-C. Toh, and W.-K. Sung, 3*D chromosome modeling with semi-definite programming and Hi-C data*, J. Comput. Biol., 20 (2013), pp. 831–846.