Title

Reproducibility of 3D chromatin configuration reconstructions

Permalink

https://escholarship.org/uc/item/2f707966

Journal

Biostatistics, 15(3)

ISSN

1465-4644

Authors

Segal, Mark R

Xiong, Hao

Capurso, Daniel

et al.

Publication Date

2014-07-01

DOI

10.1093/biostatistics/kxu003

Biostatistics (2014), **15**, 3, *pp.* 442–456
doi:10.1093/biostatistics/kxu003
Advance Access publication on February 11, 2014

Reproducibility of 3D chromatin configuration reconstructions

MARK R. SEGAL*, HAO XIONG, DANIEL CAPURSO, MARIEL VAZQUEZ, JAVIER ARSUAGA

Department of Epidemiology and Biostatistics, Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco, CA 94143, USA
Department of Mathematics, San Francisco State University, San Francisco, CA 94132, USA
mark@biostat.ucsf.edu

Summary

It is widely recognized that the three-dimensional (3D) architecture of eukaryotic chromatin plays an important role in processes such as gene regulation and cancer-driving gene fusions. Observing or inferring this 3D structure at even modest resolutions had been problematic, since genomes are highly condensed and traditional assays are coarse. However, recently devised high-throughput molecular techniques have changed this situation. Notably, the development of a suite of chromatin conformation capture (CCC) assays has enabled elicitation of *contacts*—spatially close chromosomal loci—which have provided insights into chromatin architecture. Most analysis of CCC data has focused on the contact level, with less effort directed toward obtaining 3D reconstructions and evaluating the accuracy and reproducibility thereof. While questions of accuracy must be addressed experimentally, questions of reproducibility can be addressed statistically—the purpose of this paper. We use a constrained optimization technique to reconstruct chromatin configurations for a number of closely related yeast datasets and assess reproducibility using four metrics that measure the distance between 3D configurations. The first of these, Procrustes fitting, measures configuration closeness after applying reflection, rotation, translation, and scaling-based alignment of the structures. The others base comparisons on the within-configuration inter-point distance matrix. Inferential results for these metrics rely on suitable permutation approaches. Results indicate that distance matrix-based approaches are preferable to Procrustes analysis, not because of the metrics *per se* but rather on account of the ability to customize permutation schemes to handle within-chromosome contiguity. It has recently been emphasized that the use of constrained optimization approaches to 3D architecture reconstruction are prone to being trapped in local minima. Our methods of reproducibility assessment provide a means for comparing 3D reconstruction solutions so that we can discern between local and global optima by contrasting solutions under perturbed inputs.

Keywords: Chromatin conformation; Distance matrix; Genome architecture, Procrustes analysis.

1. Introduction

The three-dimensional (3D) architecture of eukaryotic chromatin is receiving increasing attention on account of the numerous critical roles it plays in nuclear and cellular function. In particular, gene regulation

*To whom correspondence should be addressed.

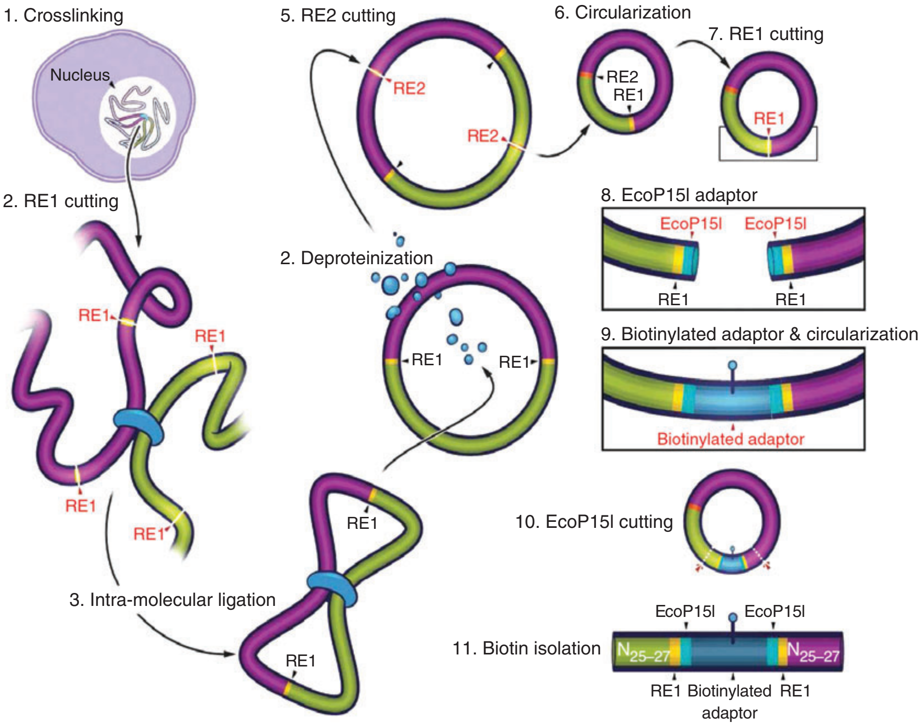

Fig. 1. Generation of CCC data. Reprinted by permission from Macmillan Publishers Ltd: Nature, Duan *and others* (2010) copyright 2010.

(Yip *and others*, 2012), genome stability, and cancer-driving gene fusions (Misteli, 2007; Mitelman *and others*, 2007) are believed to be strongly influenced by the 3D organization of the genome. Until recently observing, or inferring, 3D structure at even modest resolutions was problematic due to the high degree of genome condensation and the low-throughput, labor-intensive nature of traditional methods, such as fluorescence *in situ* hybridization (FISH), for determining spatial configuration. However, the development of next generation sequencing-based techniques, such as the suite of chromatin conformation capture assays, hereafter termed CCC and surveyed in van Steensel and Dekker (2010) and Marti-Renom and Mirny (2011), has enabled elicitation of chromatin contacts or interactions—spatially close chromosomal loci—based on the frequency of cross-linking between (pairwise) physically proximal sites (which may be genomically distal). Figure 1 provides a schematic of one such protocol (Duan *and others*, 2010).

Most analyses of CCC data to date have focused on the contact level with appreciably less effort directed toward using the data to derive (let alone assess) 3D reconstructions, as illustrated by Figure 3. While much insight has been gained from contact-level information, notably in terms of assessing colocalization of genomic or functional elements (e.g. Witten and Noble, 2012), there is potentially added value to be gleaned from a 3D perspective as we note in Section 5.

There have been three classes of approaches advanced to obtain such 3D genome configuration reconstructions and these are briefly described in the next section. The main objective of this paper is to develop statistical methods whereby the *reproducibility* of 3D reconstructions can be evaluated, this aspect having received scant attention. Because CCC data can have poor signal-to-noise ratios (Dekker, 2006; Kalhor *and others*, 2011), reproducibility has been extensively investigated at the contact level (Duan *and others*, 2010; Tjong *and others*, 2010; Dixon *and others*, 2012), but with no follow-through to reproducibility on the 3D level.

Variation between 3D reconstructions can arise due to data perturbation, reconstruction algorithm stochasticity, changes to algorithm inputs (tuning parameters, constraint formulations), and combinations thereof. It has recently been contended by Hu *and others* (2013) that optimization-based reconstructions may become trapped in local minima due to high parameter space dimensionality. Such possibilities could be assessed by comparing solutions obtained under differing starts and/or input/data perturbations. However, this requires methodology for comparing solutions, as we subsequently develop. Further, there is a fundamental distinction between reproducibility and *accuracy*. Assessing accuracy of 3D reconstructions is challenging as there are no gold standards and chromatin conformations are dynamic and cell/condition/tissue specific.

The paper is organized as follows. Section 2 describes approaches to 3D reconstruction, in particular the constrained optimization method of Duan *and others* (2010), whose data and algorithms are central to our subsequent development. Section 3 presents competing approaches and issues surrounding evaluations of reproducibility, while Section 4 showcases application of these techniques to specific yeast datasets. Section 5 provides concluding discussion.

2. Genome configuration reconstruction from CCC data

2.1 *CCC data generation*

CCC technologies enable discernment of long-range chromatin interactions, at high resolution, on a genome-wide scale. Common to CCC experimental protocols, as illustrated in Figure 1, is treatment of cells with formaldehyde resulting in cross-linking of physically proximal chromatin, followed by various restriction enzyme (RE) fragmentation, ligation, labeling, purification, and sequencing steps. Two genomic loci that are ligated, captured, sequenced together, and mapped back to the reference genome are called *contacts*. The number of times the two loci are sequenced together is their contact frequency, which is inversely proportional to their physical proximity.

Preprocessing approaches applied to these raw data include model-based false discovery rate (FDR) filtering (Duan *and others*, 2010), corrections for biases induced by GC content, fragment length and mappability (Yaffe and Tanay, 2011), and normalization methods (Lieberman-Aiden *and others*, 2009; Kalhor *and others*, 2011). As our focus is on the reproducibility (not accuracy) of reconstructions, we only explore FDR filtering.

2.2 *Reconstruction approach issues*

Hu *and others* (2013) differentiate between two classes of approach to obtaining 3D reconstructions from CCC data: optimization and probabilistic. Under optimization approaches, the observed contact frequencies between genomic loci are translated into spatial distances and 3D points configured so as to best conform to these distances, subject to biological constraints. A variety of translation strategies, constraint formulations, and optimization algorithms have been deployed (Duan *and others*, 2010; Tjong *and others*, 2010; Kalhor *and others*, 2011; Nagano *and others*, 2013). Below we elaborate on the first of these, which uses an interior point algorithm for optimization. The remainder use simulated annealing (SA), which, like probabilistic approaches, readily produces a large ensemble of solutions. Probabilistic methods (Rousseau *and others*, 2011; Hu *and others*, 2013) are distinct from optimization methods in that they are *generative*: they prescribe parametric models with attendant distributional assumptions and use sophisticated sampling schemes to obtain solutions. An alternate approach uses contact data indirectly: a theory-based 3D genome configuration is generated according to a hypothesized topology (e.g. equilibrium or fractal globules, Lieberman-Aiden *and others*, 2009), and summaries thereof are contrasted with corresponding summaries derived from the contact data.

We offer some brief comments on attributes of the differing reconstruction techniques. Hu *and others* (2013) similarly proffer assessments, noting limitations of existing methods. First, they note that no approach accounts for systematic biases (e.g. GC content, fragment length) in performing reconstruction. However, algorithms exist for effecting corresponding bias correction (Yaffe and Tanay, 2011), and so this can be treated as a preprocessing rather than modeling task. Secondly, they contend that optimization-based methods are prone to being trapped in local optima due to the high dimensionality of the optimization problem. This, along with the companion issue of non- or slow convergence to any solution, is a concern. As noted in Section 1, there is a role of reproducibility assessment here: given means for comparing 3D reconstruction solutions, we can attempt to discern between local and global optima by contrasting solutions under perturbed inputs. A further limitation identified is the focus of existing methods on providing a consensus reconstruction without consideration of structural variation therefrom. But, this is not so for SA-based optimization methods that, in yielding a large ensemble of solutions, facilitate exploration of variation, and admit differing modes of summarization beyond positional (coordinate-wise) consensus (Tjong *and others*, 2010; Kalhor *and others*, 2011).

A notable shortcoming of current probabilistic models is that they are restricted to providing reconstructions one chromosome at a time. As such, they fail to utilize much of the data, there being 5–8-fold more inter- than intra-chromosomal contacts, albeit at lower frequencies. Relative chromosomal positioning is also lost. Further, notions that either probabilistic or optimization-derived ensembles can dissect variation that reflects chromatin dynamics or between-cell differences are aspirational: without single-cell and/or time-course data (but instead relying on cell population averages) it is impossible to distinguish such variation from other factors. The emergence of single-cell assays (Nagano *and others*, 2013) can help address these issues.

2.3 *Duan and others*: *interior point algorithms*

Using CCC data obtained from the yeast *Schizosaccharomyces cerevisiae*, hereafter *S. cerevisiae*, Duan *and others* (2010) generated a 3D genome reconstruction (Figure 3, panel 1) using constrained optimization. The key ideas are as follows. We assume that a contact with frequency f has the same distance between its concomitant loci as an *intra*-chromosomal contact that has frequency f due to polymer packing ($\approx$ 130 bp of packed chromatin corresponds to $\approx$ 1 nm). This makes it possible to convert genomic distance to physical distance as illustrated in Figure 2, which depicts (smoothed) contact frequency: genomic distance relationships for the 16 *S. cerevisiae* chromosomes. Now, represent each chromosome as a series of equi-spaced beads, where the (predefined) spacing will determine the *resolution* of the solved 3D configuration. Using genomic coordinates, we map each contact locus to its closest bead, the mapping potentially being many-to-one. Let $p_i = (x_i, y_i, z_i)$ denote the (unknown) 3D coordinates of the ith bead and let $d(p_i, p_j)$ be the Euclidean distance between beads i and j. Let δ_{ij} be the inferred 1D physical distance between loci corresponding to these beads, based on the above contact frequency conversion. To obtain a 3D configuration (i.e. solve for unknown bead coordinates x_i, y_i, z_i), we minimize an objective function that attempts to place interacting loci at their expected distance apart:

$$\min_p \sum_{i<j} (d(p_i, p_j) - \delta_{ij})^2, \qquad (2.1)$$

where $\sum_{i<j}$ represents the double sum over all bead pairs. In order to obtain biologically meaningful solutions, it is essential to impose biological constraints. For *S. cerevisiae*, these include the following: (i) all beads lie within a sphere of radius 1 μm corresponding to the shape and dimension of the *S. cerevisiae*

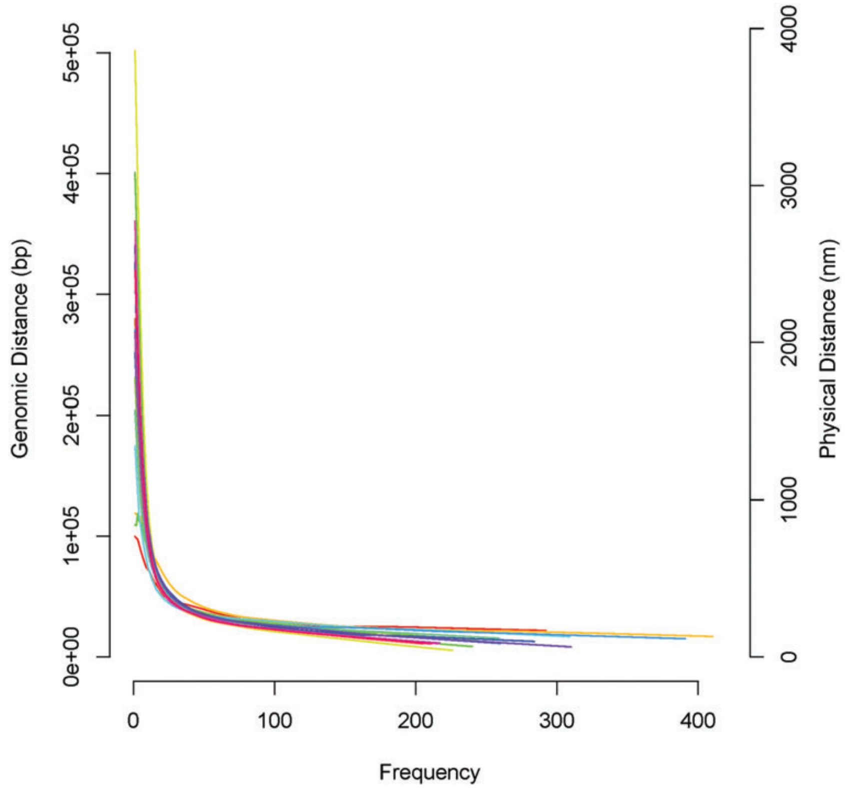

Fig. 2. Conversion of genomic to physical distance. The underlying contact data derive from HindIII (RE1) and MseI (RE2) REs, exclude loci within 20 kb, and do not employ FDR-based filtering.

nucleus; (ii) adjacent beads lie within a given range so as to capture contiguity; (iii) no two beads on the same chromosome can be closer than 30 nm (motivated by the thickness of the chromatin fiber); (iv) to preclude inter-chromosomal crossings between segments connecting adjacent beads, beads on different chromosomes are at least 75 nm apart; (v) rDNA repeats are localized within the nucleolus, which is assigned a predefined position and extent; and (vi) the chromosome XII centromere is positioned opposite the nucleolus. Some of these constraints are highly *S. cerevisiae* specific; others reflect broad chromatin fiber properties.

Solving (2.1), which represents a constrained multi-dimensional scaling problem, is challenging since the nature of the constraints makes for a high-dimensional, non-convex problem: for *S. cerevisiae* with beads spaced every 10 kbp there are $\sim 4 \times 10^3$ parameters and 10^6 constraints. Duan *and others* (2010) use interior point optimization (IPO), using the Ipopt library (ver. 3.10.0), to handle these challenges. We have implemented their approach and obtained a series of 3D *S. cerevisiae* reconstructions corresponding to differing inputs (filtering extent, RE library, physical distance imputation).

3. Assessing reproducibility

Evaluating agreement between 3D genome reconstructions falls under the rubric of comparing spatial point patterns or shapes. On one hand, the problem is simplified since *registration*—matching points (beads) across patterns (reconstructions), often a substantial preliminary task—is immediate since the genomic position of all beads is known. On the other hand, the problem is complicated by the existence of *sub-shapes*, as defined by individual chromosomes, and their inherent contiguity properties. Two components need to be specified to effect reproducibility assessment. First, we need a statistic that provides a measure of closeness for the respective reconstructions; four are outlined below. Secondly, we need a referent distribution for evaluating statistic significance. For the statistics we employ, referents have previously been obtained by permutation (Legendre and Legendre, 2012) or closed-form approximations thereof (Minas *and others*, 2013). As we demonstrate, accommodating sub-shapes can impact both statistic formulation and permutation strategy. Alternatively, referents have been simulated based on random walk models (Hu *and others*, 2013) that, in turn, are based on FISH data (Sachs *and others*, 1995). However, the random walk model is incompatible with several aspects of genome organization including absence of "giant loop" components. More importantly, this approach pertains to individual chromosomes and affords no means for relative positioning of multiple chromosomes.

Our methods for assessing 3D reproducibility operate strictly on 3D configurations; we do not consider approaches that gauge reproducibility on the 1D contact level. However, in Section 5, we describe limitations of 1D methods that further motivate utilization of 3D methods. It is notable that even in instances where solution ensembles are obtained, reproducibility has been assessed using 1D measures rather than the reconstructions themselves (Tjong *and others*, 2010).

3.1 *Procrustes analysis*

There are many sources describing Procrustes analysis (e.g. Dryden and Mardia, 1988) which provides methods for assessing correspondences between shapes, according to differing definitions thereof (Kent and Mardia, 2012). In comparing genome configurations, we are interested in *reflection similarity shape*, under which two configurations that only differ by a reflection, rotation, translation, and scaling are deemed equivalent. Let G_s, G_t be $n \times 3$ matrices with rows the 3D coordinates for the n (common) beads resulting from two differing reconstructions. The closeness of G_s and G_t can be measured by how far apart corresponding points are, after optimizing for the allowed transformations. Initially, ignoring scaling, this gives rise to the criterion

$$\min_{\mu, Z} \|G_t - (G_s Z + 1\mu^{\mathrm{T}})\|_{\mathrm{F}}, \tag{3.1}$$

where Z is a 3×3 orthogonal matrix and μ is a 3-vector of translation coordinates. Closeness is measured by the Frobenius norm: $\|X\|_{\mathrm{F}}^2 = \mathrm{trace}(X^{\mathrm{T}}X) = \sum_{ij} x_{ij}^2$. Let $\bar{g}^s$, $\bar{g}^t$ be the column means of G_s, G_t, respectively, with $\tilde{G}^s$, $\tilde{G}^t$ centered versions obtained by subtracting column means. Let the singular value decomposition of $(\tilde{G}^s)^{\mathrm{T}}\tilde{G}^t = UDV^{\mathrm{T}}$. Then the solution to (3.1) is

$$\hat{Z} = UV^{\mathrm{T}}, \tag{3.2}$$

$$\hat{\mu} = \bar{g}^t - \hat{Z}\bar{g}^s. \tag{3.3}$$

Based on the form of the solution (3.2) and (3.3), we can work with $\tilde{G}^s$, $\tilde{G}^t$ and disregard location. Then, after re-introducing scaling, we arrive at our Procrustes distance reproducibility criterion:

$$\phi_{\mathrm{P}}(\tilde{G}^s, \tilde{G}^t) = \min_{\beta, Z} \|\tilde{G}^t - \beta\tilde{G}^s Z\|_{\mathrm{F}}, \tag{3.4}$$

with solutions $\hat{Z}$ as in (3.2) and $\hat{\beta} = \mathrm{trace}(D)/\|\tilde{G}^s\|_{\mathrm{F}}^2$.

Obtaining a referent distribution for $\phi_{\mathrm{P}}(\tilde{G}^s, \tilde{G}^t)$ is challenging due to high dimensionality and structural complexity. Inference for $\phi_{\mathrm{P}}(\tilde{G}^s, \tilde{G}^t)$ in other settings has made recourse to permutation testing. However, this makes implicit exchangeability assumptions and violation of these can result in poor performance (Amaral *and others*, 2007). Here, where permutation involves scrambling 3D points, exchangeability is problematic in view of within-chromosome contiguity. Nonetheless, in part due to lack of alternatives, we apply permutation testing, both overall and within chromosome, so as to demonstrate this poor performance.

3.2 *Within-structure distance-based methods*

Shared patterns in paired multivariate data, as constituted by G_s, G_t, can be assessed using statistics computed on *distance matrices* rather than the original data (Mantel, 1967; Legendre and Lapointe, 2004; Minas *and others*, 2013). This makes the problem of obtaining null referent distributions in the face of within-chromosome contiguity more approachable. We detail three such statistics and describe permutation-based inference.

3.2.1 *Mantel's test/congruence among distance matrices.* The Mantel test (Mantel, 1967), and its generalization from pairwise to multi-way comparisons, the *congruence among distance matrices* test (Legendre and Lapointe, 2004), are widely used in bioinformatics (Minas *and others*, 2013) and phylogenetics (Campbell *and others*, 2011). A version of the latter based on Kendall's coefficient of concordance is as follows. For each pair of genome configurations G_s, G_t obtain the $n \times n$ distance matrix of inter-point distances: $D_s = (d(p_i^s, p_j^s))$, where $d(p_i^s, p_j^s)$ is the Euclidean distance between positions p_i^s, p_j^s of rows (beads) i, j of G_s; and similarly for D_t. Vectorize the upper triangle of each distance matrix (excluding the main diagonal) and apply the rank transform to each vector. Compute r_j for $j = 1, \ldots, m = n(n-1)/2$ as the sum of ranks for the jth position over the two vectors. Then Kendall's coefficient of concordance is defined as

$$\phi_{\mathrm{W}}(G_s, G_t) = \frac{12\sum_{j=1}^{m} r_j^2 - 3q^2 m(m+1)^2}{q^2(m^3 - m) - qC}, \tag{3.5}$$

where q is the number of genome configurations being compared (here $q = 2$) and C is a correction for any tied ranks: $C = \sum_{k=1}^{K}(c_k^3 - c_k)$ with K being the number of groups of ties and c_k being the number of tied ranks in the kth group. Computation of $\phi_{\mathrm{W}}(G_s, G_t)$ is appreciably slower than for the other statistics considered due to the ranking required with m large.

A simple transformation of ϕ_{W} gives Friedman's χ^2 statistic for two-way analysis of variance using ranks. However, use of either a referent χ^2 distribution or permutation of the $m = n(n-1)/2$ distances for inference is misplaced on account of dimensionality (n vs. m) and dependency concerns (Legendre and Lapointe, 2004). Rather, permutation operates on the distance matrices D_s, D_t, with the same permutation applied to rows or columns, separately for each matrix. By restricting contributions to $\phi_{\mathrm{W}}(G_s, G_t)$ to distances corresponding to *inter*-chromosomal comparisons, and similarly for permuted versions thereof, we eliminate the exchangeability concerns due to *intra*-chromosomal contiguity raised in Section 3.1. It is this feature that makes distance matrix approaches attractive. It is not possible to gauge reasonableness of the restricted permutation approach in terms of associated induced 3D structures. While it is immediate to map from 3D coordinates to distance matrices, the inverse mapping is NP-hard. That we were able to solve an analogous problem, on a multi-chromosomal level, in Section 2.3 using criterion (2.1) was contingent on the constraints. But, after intra-chromosomal exclusions and distance permutation have been applied, we can no longer frame meaningful biological constraints. Moreover, even if this inverse problem could be solved, there is little basis for assessing the plausibility of the attendant structure as a null referent.

3.2.2 *Distance differencing.* In the context of generative models of individual chromosome configurations Rousseau *and others* (2011) are critical of Procrustes methods due to the computational burdens incurred by rotation, reflection, and translation alignment, although this has not been an issue in our applications. We term the statistic they use *distance differencing*:

$$\phi_{\mathrm{D}}(G_s, G_t) = \sqrt{\sum_{i<j}(d(p_i^s, p_j^s) - d(p_i^t, p_j^t))^2}. \tag{3.6}$$

As noted by a referee, $\phi_{\mathrm{D}}(G_s, G_t)$ is not scale invariant. But we can readily attain scale invariance by substituting $d^*(p_i^s, p_j^s) = d(p_i^s, p_j^s)/\sum_{i<j} d(p_i^s, p_j^s)$ for $d(p_i^s, p_j^s)$ in (3.6) and similarly for $d(p_i^t, p_j^t)$. It is this invariant version, still named ϕ_{D}, that we use subsequently. Either version is amenable to distributional evaluation using the same permutation scheme as for $\phi_{\mathrm{W}}(G_s, G_t)$.

Weighted versions, using predefined weights, for each of the statistics $\phi_{\mathrm{P}}(G_s, G_t)$, $\phi_{\mathrm{W}}(G_s, G_t)$, and $\phi_{\mathrm{D}}(G_s, G_t)$ can be readily formulated so as to up/down weight inter-bead distances corresponding to *domains* of the structure. For such versions inference would proceed as previously, weighting carrying over to the permutation scheme. Also ensemble, as opposed to pairwise, comparisons are possible for each statistic wherein multiple structures are compared with some consensus structure/null. Indeed, the formulation given for $\phi_{\mathrm{W}}(G_s, G_t)$ explicitly allows for this via $q > 2$, while details for $\phi_{\mathrm{P}}(\tilde{G}^s, \tilde{G}^t)$ are given in Hastie *and others* (2009).

3.2.3 *Generalized RV test.* The recently proposed generalized RV test (GRV, Minas *and others*, 2013) can handle a variety of data types and distance measures and offers improved power over Mantel's test in many settings. The precursor RV statistic is developed as a matrix extension of Pearson's correlation:

$$\phi_{\mathrm{R}}(G_s, G_t) = \mathrm{RV}(G_s, G_t) = \frac{\mathrm{tr}(G_s^{\mathrm{T}} G_t G_t^{\mathrm{T}} G_s)}{\|G_s^{\mathrm{T}} G_s\|_{\mathrm{F}} \|G_t^{\mathrm{T}} G_t\|_{\mathrm{F}}} = \frac{\mathrm{tr}(G_s G_s^{\mathrm{T}} G_t G_t^{\mathrm{T}})}{\|G_s G_s^{\mathrm{T}}\|_{\mathrm{F}} \|G_t^{\mathrm{T}} G_t^{\mathrm{T}}\|_{\mathrm{F}}}. \tag{3.7}$$

Since $G_s G_s^{\mathrm{T}} = -\frac{1}{2} A D_s^2 A$ where $A = (I_n - J_n/n)$ with I_n the $n \times n$ identity matrix and J_n the $n \times n$ matrix of ones, and similarly for $G_t G_t^{\mathrm{T}}$, ϕ_{R} is completely determined by the (Euclidean) distance matrices D_s, D_t. The GRV simply replaces the underlying Euclidean distances with any distance measure. For our spatial applications, we do not consider non-Euclidean distances.

The relationship of ϕ_{R} to Mantel's test is detailed in Minas *and others* (2013) who also detail means for calculating closed-form p-value approximations and the considerable benefits these bestow. They are derived by moment matching the exact null distribution as obtained by using all $n!$ distance matrix row and column permutations to a continuous distribution. In particular, the first three moments of the null are matched to a Pearson type III distribution which has been shown to capture appropriate skewness characteristics. Analytical results enable these moments to be readily computed. However, a critical issue here is that these results pertain to the exact null based on *all* permutations. So, use of closed-form p-values for ϕ_{R} requires inclusion of intra-chromosal permutations in contrast to the scheme proposed for ϕ_{W} and ϕ_{D} and hence, as for ϕ_{P}, making questionable exchangeability assumptions.

4. Results

4.1 *3D genome reconstructions using interior point algorithms*

Using the methods of Duan *and others* (Section 2.3), we obtained 12 distinct *S. cerevisiae* genome reconstructions corresponding to differing data inputs (Table 1). We deliberately focus on assessing reproducibility across structures obtained from differing datasets, as opposed to perturbed constraints, since the latter

Table 1. *Attributes of the reconstructions shown in Figure 3*

Panel	RE1	RE2	FDR%	Physical distance	Iterations
1	HindIII	MseI ∪ MspI	0.01	Original	Unknown
2	HindIII	MseI ∪ MspI	0.01	Original	1936
3	HindIII	MseI ∪ MspI	0.10	Original	1881
4	HindIII	MseI ∪ MspI	1.00	Original	1818
5	HindIII	MseI	—	Original	2591
6	HindIII	MspI	—	Original	2106
7	HindIII	MseI ∪ MspI	0.01	Re-computed	2207
8	HindIII	MseI ∪ MspI	0.01	Re-computed	4937
9	EcoRI ∩ HindIII	MseI ∪ MspI	0.01	Original	1990
10	EcoRI	MseI ∪ MspI	0.01	Original	2406
11	EcoRI	MseI	—	Original	1591
12	EcoRI	MspI	—	Original	1993
13	Linear combination of panels 1 (0.25) and 7 (0.75)				
14	Linear combination of panels 1 (0.25) and 8 (0.75)				
15	Linear combination of panels 2 (0.25) and 7 (0.75)				
16	Linear combination of panels 2 (0.25) and 8 (0.75)				

constitutes an open-ended range of possibilities, and we have no basis for departing from the original specifications. Further, we also made a point of including some replicate re-runs. Obtaining reconstructions is slow with each requiring a wall-clock time of ~ 3.0 days on an Intel Xeon 23 running at 3.00 GHz with 32 GB of memory. An additional 4 structures were obtained as linear combinations from among the 12 for illustrative purposes.

The inputs examined represent an exhaustive treatment of available *S.cerevisiae* contact data (http://noble.gs.washington.edu/proj/yeast-architecture/sup.html). These vary according to (i) RE1 and RE2 (Figure 1) choice, there being two possibilities for each and (ii) FDR contact filtering extent with levels 0.01%, 0.1%, or 1.0%, or none. RE1 is used to capture the actual interaction and turn it into circularized DNA, while RE2 makes the plasmid shorter for sequencing. Only select combinations of RE1, RE2, and FDR are provided. In addition, we explored recomputing imputed physical distances (cf. Figure 2) so that they reflect data being used for a given reconstruction. The logic is that if it is assumed that FDR filtering at the 0.01% level is required to obtain reliable contact data for reconstruction purposes, then the same filtering ought to pertain to physical distance imputation. The final column of Table 1 gives the number of IPO algorithm iterations until convergence under default criteria. Those datasets for which RE2 is designated as MseI ∪ MspI correspond to combining interaction data for the two designated REs, whereas the dataset for which RE1 is designated as EcoRI ∩ HindIII corresponds to utilizing overlapping interactions. Figure 3 shows snapshots of all reconstructions with panels labeled as per Table 1.

4.2 *Reconstruction reproducibility*

Figure 4 shows inter-relationships among the metrics $\phi_{\mathrm{D}}, \phi_{\mathrm{W}}, \phi_{\mathrm{P}}, \phi_{\mathrm{R}}$ for the $\binom{16}{2} = 120$ pairwise comparisons of the 16 structures. Agreement between the metrics is strikingly good as indicated by the large absolute correlations. This is not surprising for $\phi_{\mathrm{W}}, \phi_{\mathrm{R}}$ in particular, which differ only in the correlation method (Spearman, Pearson, respectively) used, but is notable for ϕ_{P} with the others. The inverse relationship between $\phi_{\mathrm{W}}, \phi_{\mathrm{R}}$ and $\phi_{\mathrm{D}}, \phi_{\mathrm{P}}$ simply reflects that closeness for the correlation-based metrics corresponds to large values, whereas closeness for the distance-based metrics ($\phi_{\mathrm{D}}, \phi_{\mathrm{P}}$) corresponds to small values.

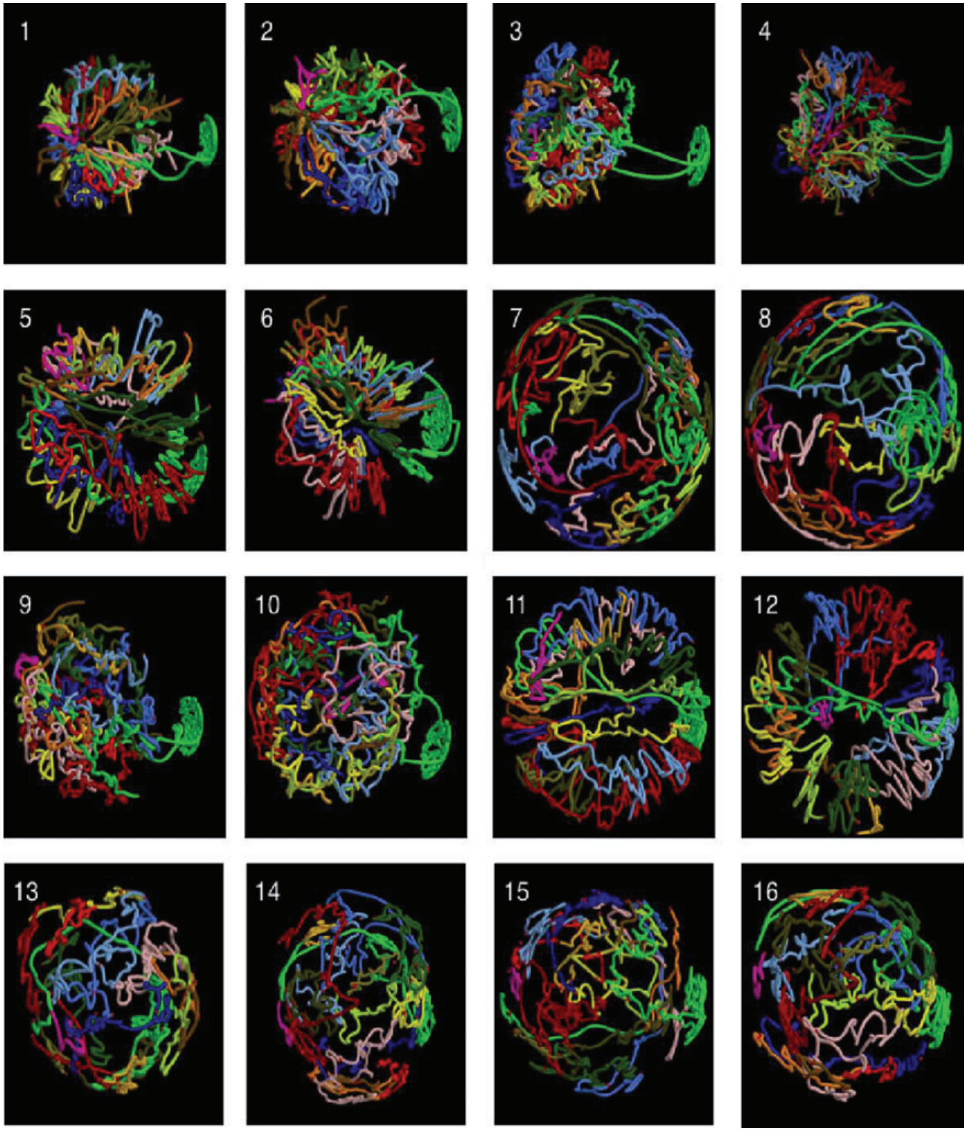

Fig. 3. Differing *S.cerevisiae* genome reconstructions according to the specifications in Table 1 with panel numbers corresponding to rows therein. Colors correspond to the 16 chromosomes. Centromeres and telomeres are depicted by red dots.

Table 2 contains p-values based on 1000 permutations for distance differencing ϕ_D (above the diagonal) and Procrustes distance ϕ_P (below the diagonal; based on permuting coordinates within chromosomes) for the 16 structures examined. Concordances and contrasts with ϕ_W and ϕ_R are described below. For the actual reconstructions (panels 1–12), results are extreme with most p-values <0.001 or >0.999, hereafter 0 or 1, respectively. Panels 13–16 are included to demonstrate that permutation-based inference does not necessarily yield extreme p-values with intermediary structures obtained by linear combination of distinct parent structures yielding intermediary values when compared with their parents. Consider panels 1 and

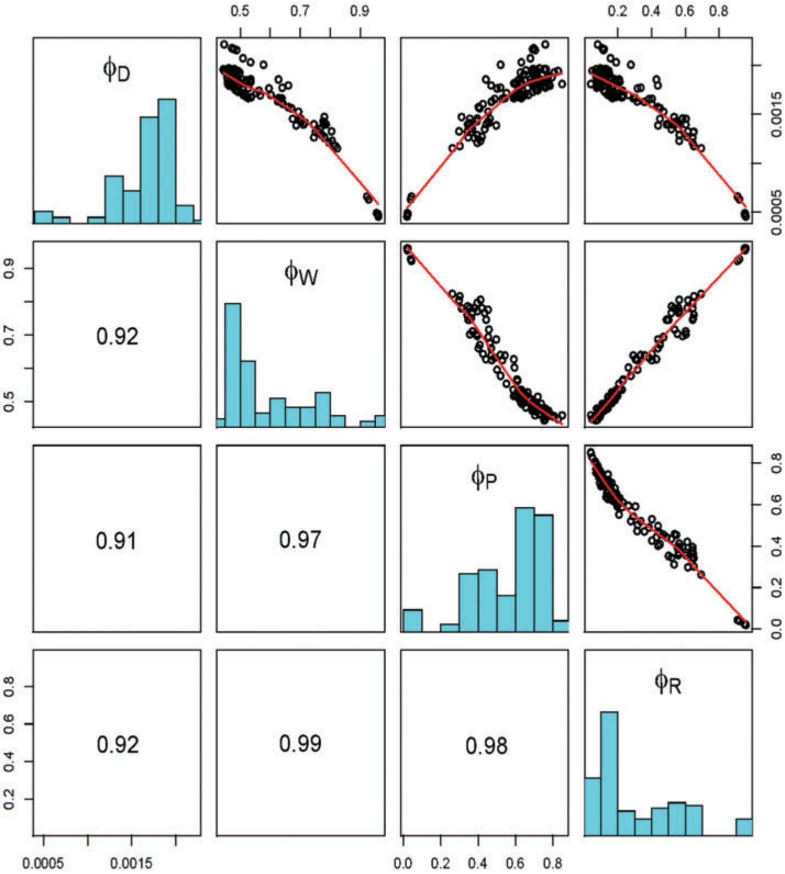

Fig. 4. Pairwise comparison of $\phi_{\rm D}$, $\phi_{\rm W}$, $\phi_{\rm P}$, $\phi_{\rm R}$ values for the reconstructions of Figure 3. Upper triangle: smoothed pairwise scatterplots; lower triangle: pairwise absolute correlations; diagonal: histograms of metric values.

2 that constitute replicate runs. The p-value of 1 obtained for all four metrics means that the two original reconstructions are more similar than any of their 1000 permuted counterparts. Conversely, for the visually dissimilar panels 1 and 7, we obtain a p-value of 0 for $\phi_{\rm D}$ and $\phi_{\rm W}$, meaning that the original reconstructions are more dissimilar than any of their 1000 permuted counterparts, whereas a p-value of 1 is obtained for $\phi_{\rm P}$ and $\phi_{\rm R}$. Moreover, if we modify $\phi_{\rm D}$ and $\phi_{\rm W}$ so as to allow inclusion of intra-chromosomal distances in both metric and associated permutation computations, the p-values flip to 1.

This illustrates the key finding: permutation schemes need to exclude intra-chromosomal contributions so as to overcome contiguity constraints. This is impossible for $\phi_{\rm P}$ and the closed-form p-value approximations of $\phi_{\rm R}$. The results given for $\phi_{\rm P}$ are based on permuting coordinates within chromosomes; if we permute all coordinates, p-values are uniformly 1. This is also the result for $\phi_{\rm R}$ when using *all* distance matrix permutations, as required by the closed-form approximations.

Owing to the absence of gold standards, we have deliberately not addressed accuracy. However, in broad terms we can distinguish between two classes of reconstruction: those obtained using recomputed physical distances (replicate panels 7 and 8) versus the remainder of the original 12. In terms of large-scale attributes such as compaction and colocalization of centromeres, these are less credible than the

Table 2. *p-values for the* 16 *reconstructions for measures* ϕ_{D} (*upper triangle*) *and* ϕ_{P} (*lower triangle*)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	—	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.99	0.01	1.00	0.98
2	1.00	—	1.00	1.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.77	0.11	1.00	1.00
3	1.00	1.00	—	1.00	1.00	1.00	0.00	0.04	1.00	1.00	1.00	1.00	0.96	0.02	1.00	1.00
4	1.00	1.00	1.00	—	1.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.98	0.00	0.95	0.02
5	1.00	1.00	1.00	1.00	—	1.00	0.00	0.00	1.00	1.00	1.00	1.00	0.03	0.42	0.97	0.72
6	1.00	1.00	1.00	1.00	1.00	—	0.00	0.00	1.00	1.00	1.00	1.00	0.08	0.02	1.00	0.99
7	1.00	1.00	1.00	0.64	1.00	1.00	—	0.95	0.78	0.01	0.00	0.00	1.00	0.94	1.00	0.83
8	0.98	1.00	1.00	0.00	0.78	1.00	1.00	—	0.00	0.00	0.00	0.00	0.95	1.00	0.81	1.00
9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	—	1.00	1.00	1.00	1.00	0.01	1.00	0.73
10	1.00	1.00	1.00	0.78	1.00	1.00	1.00	0.15	1.00	—	1.00	1.00	0.00	0.00	1.00	0.00
11	1.00	1.00	1.00	0.00	1.00	1.00	0.41	0.40	1.00	1.00	—	1.00	0.00	0.10	0.16	0.89
12	1.00	1.00	1.00	1.00	1.00	1.00	0.03	0.36	1.00	1.00	1.00	—	0.00	0.04	0.00	0.00
13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	0.92	0.02	—	0.83	1.00	0.86
14	1.00	1.00	0.56	1.00	1.00	1.00	0.96	1.00	1.00	0.97	0.99	1.00	0.99	—	0.57	1.00
15	1.00	1.00	1.00	1.00	1.00	1.00	0.03	1.00	1.00	1.00	1.00	0.47	1.00	1.00	—	0.85
16	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00	—

Results for ϕ_{W} are similar to those for ϕ_{D}. p-value determination is based on 1000 simulations: entries of 0.00 and 1.00 should be interpreted as <0.001 and >0.999, respectively.

other reconstructions. So, we prefer metrics ϕ_{D}, ϕ_{W}, which better discriminate between these classes than ϕ_{P}, ϕ_{R}, consistent with the above exchangeability considerations.

5. Discussion

The task of reconstructing 3D genome configurations from CCC contact data is formidable and assessing the accuracy of putative solutions is problematic. This makes gauging the merits of any given reconstruction difficult. In this paper, we have tackled the lesser task of assessing agreement between candidate reconstructions. One of the forefront criticisms of optimization-based reconstruction approaches is their potential to be trapped in local optima due to the high dimensionality of the parameter space (Hu *and others*, 2013). Being able to measure agreement between differing solutions, obtained under perturbed data inputs, constraint specifications, starting conditions or even just re-runs for algorithms with stochastic components, provides a means for distinguishing global from local optima.

The reproducibility measures we consider are highly concordant as evidenced by the correlations exhibited in Figure 4. The distance-based measures utilize inter-bead distances either directly (ϕ_{D}, ϕ_{R}), or after rank transformation (ϕ_{W}). An alternate measure used in this setting, the overlap index (Kalhor *and others*, 2011; Tjong *and others*, 2010), instead creates indicators based on dichotomizing distances according to a prescribed threshold, and then determining the extent to which the indicators for the two structures intersect. Aside from this representing a coarse treatment of distances, results will be highly sensitive to the threshold employed. In view of the high metric correlations, the extent to which they produce differing reproducibility p-values can be ascribed to differing permutation schemes. For distance-based approaches, this involves permuting the attendant distance matrices, while, for the Procrustes approach, 3D coordinates are permuted. The former provides a partial finesse: we can overcome concerns about chromosome contiguity simply by evaluating the metrics, and permuted versions thereof, only for between-chromosome interpoint distances. The impact of this finesse was demonstrated by the profound and undesirable p-value shifts that result from including within-chromosome inter-point distances. Efforts at more sophisticated domain-level control (Paulsen *and others*, 2013) have been limited to 1D analyses, while model-based approaches to null referent generation (Hu *and others*, 2013) have been limited to individual chromosomes and so do not permit between-chromosome positioning, critical in assessing genome architecture.

Considerable effort has been invested in assessing reproducibility on the 1D contact level. This has ranged from comparing correlations between contact frequency maps arising from differing RE digests (Duan *and others*, 2010), preprocessing steps (Yaffe and Tanay, 2011), and summaries of ensemble components (Kalhor *and others*, 2011). In this latter application, an extreme and highly significant correlation resulted ($\rho = 0.999$) and was used to conclude that the contact maps were highly reproducible. However, a dominant contributor to this measure is the vast number of non-contacting inter-chromosomal loci. This illustrates the importance of assessing structure reproducibility on the 3D level and not just the 1D level.

3D reconstructions also provide added value over 1D analyses in assessing *colocalization*. Efforts at deriving biological insight from CCC data have naturally focused on colocalization of genomic functional elements and ontological categories, since it is such proximity information that the assays provide. Examples for *S. cerevisiae* include claimed colocalizations of tRNAs, early origins of DNA replication, chromosomal breakpoints, and numerous transcription factors (Duan *and others*, 2010; Tjong *and others*, 2010; Dai and Dai, 2012; Witten and Noble, 2012). To date all colocalization analyses are based on 1D contacts rather than 3D reconstructions. There are compelling reasons for using the latter. First, structure-based analyses can evaluate whether functional groups are more highly *dispersed*. This is a problem for contact-level approaches due to filtering/reliability/missingness of *low* frequency contacts. Secondly, contact-level approaches are inherently pairwise and thus fail to capture 3D chromatin structure. Identifying functional groups that are significantly colocalized under structural but not contact-based analyses may illuminate facets of chromatin architecture and help address conjectures; for example, the random structure of chromatin after accommodation of landmark tethering in yeast as advanced by Tjong *and others* (2010). Of course, these advantages are contingent on obtaining accurate 3D reconstructions. To better investigate questions of accuracy and reproducibility of inferred genome architectures from CCC assays, it is necessary that reconstruction algorithms be fast and stable—improving these aspects is the subject of ongoing work.

Acknowledgments

The authors thank Mirela Andronescu for help with regard to implementing the interior point optimization algorithm and two referees for helpful comments. *Conflict of Interest*: None declared.

Funding

This work was supported by the National Institutes of Health [grant number GM109457].

References

Amaral, G. J. A., Dryden, I. L. and Wood, A. T. A. (2007). Pivotal bootstrap methods for *k*-sample problems in directional statistics and shape analysis. *Journal of the American Statistical Association* **102**, 695–707.

Campbell, V., Legendre, P. and Lapointe, F.-J. (2011). The performance of the congruence among distance matrices test in phylogenetic analysis. *BMC Evolutionary Biology* **11**, 64.

Dai, Z. and Dai, X. (2012). Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Research* **40**, 27–36.

DEKKER, J. (2006). The three 'C's of chromosome conformation capture: controls, controls, controls. *Nature Methods* **3**, 17–21.

DIXON, J. R., SELVARAJ, S., YUE, F., KIM, A., LI, Y., SHEN, Y., HU, M., LIU, J. S. AND REN, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin contacts. *Nature* **485**, 376–380.

DRYDEN, I. L. AND MARDIA, K. V. (1998). *Statistical Shape Analysis*. Chichester: Wiley.

DUAN, Z., ANDRONESCU, M., SCHUTZ, K., MCILWAIN, S., KIM, Y. J., LEE, C., SHENDURE, J., FIELDS, S., BLAU, C. A. AND NOBLE, W. S. (2010). A three-dimensional model of the yeast genome. *Nature* **465**, 363–367.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

HU, M., DENG, K., QIN, Z., DIXON, J., SELVARAJ, S., FANG, J., REN, B. AND LIU, J. S. (2013). Bayesian inference of spatial organizations of chromosomes. *PLoS Computational Biology* **9**(1), e1002893.

KALHOR, R., TJONG, H., JAYATHILAKA, N., ALBER, F. AND CHEN, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology* **30**, 90–98.

KENT, J. T. AND MARDIA, K. V. (2012). A geometric approach to projective shape and the cross ratio. *Biometrika* **99**, 833–849.

LEGENDRE, P. AND LAPOINTE, F.-J. (2004). Assessing congruence among distance matrices: single malt Scotch whiskies revisited. *Australian and New Zealand Journal of Statistics* **46**, 615–629.

LEGENDRE, P. AND LEGENDRE, L. (2012). *Numerical Ecology*. Amsterdam: Elsevier.

LIEBERMAN-AIDEN, E., VAN BERKUM, N. L., WILLIAMS, L., IMAKAEV, M., RAGOCZY, T., TELLING, A., AMIT, I., LAJOIE, B. R., SABO, P. J., DORSCHNER, M. O. *and others*. (2009). Comprehensive mapping of long-range contacts reveals folding principles of the human genome. *Science* **326**, 289–293.

MANTEL, N. A. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 209–220.

MARTI-RENOM, M. A. AND MIRNY, L. A. (2011). Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Computational Biology* **7**, e1002125.

MINAS, C., CURRY, E. AND MONTANA, G. (2013). A distance-based test of association between paired heterogeneous genomic data. *Bioinformatics* **29**, 2555–2563.

MISTELI, T. (2007). Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787–800.

MITELMAN, F., JOHANSSON, B. AND MERTENS, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer* **7**, 233–245.

NAGANO, T., LUBLING, Y., STEVENS, T. J., SCHOENFELDER, S., YAFFE, E., DEAN, W., LAUE, E. D., TANAY, A. AND FRASER, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64.

PAULSEN, J., LIEN, T. G., SANDVE, G. K., HOLDEN, L., BORGAN, O., GLAD, I. K. AND HOVIG, E. (2013). Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Research* **41**, 5164–5174.

ROUSSEAU, M., FRASER, J., FERRAIUOLO, M. A., DOSTIE, J. AND BLANCHETTE, M. (2011). Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics* **12**, 414.

SACHS, R. K., VAN DEN ENGH, G., TRASK, B., YOKOTA, H. AND HEARST, J. E. (1995). A random- walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Science* **92**, 2710–2714.

TJONG, H., GONG, K., CHEN, L. AND ALBER, F. (2012). Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Research* **22**, 1295–1305.

van Steensel, B. and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. *Nature Biotechnology* **28**, 1089–1095.

Witten, D. M. and Noble, W. S. (2012). On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Research* **40**, 3849–3855.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics* **43**, 1059–1065.

Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P. *and others*. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology* **13**, R48.

[*Received May 19, 2013; revised January 9, 2014; accepted for publication January 10, 2014*]