Title
Assessing stationary distributions derived from chromatin contact maps

Journal
BMC Bioinformatics, 21(1)

ISSN
1471-2105

Authors
Segal, Mark R
Fletez-Brant, Kipper

Publication Date
2020-12-01

DOI
10.1186/s12859-020-3424-y

Peer reviewed

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Assessing stationary distributions derived from chromatin contact maps

Mark R. Segal[1*] and Kipper Fletez-Brant[2]

Abstract

Background: The spatial configuration of chromosomes is essential to various cellular processes, notably gene regulation, while architecture related alterations, such as translocations and gene fusions, are often cancer drivers. Thus, eliciting chromatin conformation is important, yet challenging due to compaction, dynamics and scale. However, a variety of recent assays, in particular Hi-C, have generated new details of chromatin structure, spawning a number of novel biological findings. Many findings have resulted from analyses on the level of native contact data as generated by the assays. Alternatively, reconstruction based approaches often proceed by first converting contact frequencies into distances, then generating a three dimensional (3D) chromatin configuration that best recapitulates these distances. Subsequent analyses can enrich contact level analyses via superposition of genomic attributes on the reconstruction. But, such advantages depend on the accuracy of the reconstruction which, absent gold standards, is inherently difficult to assess. Attempts at accuracy evaluation have relied on simulation and/or FISH imaging that typically features a handful of low resolution probes. While newly advanced multiplexed FISH imaging offers possibilities for refined 3D reconstruction accuracy evaluation, availability of such data is limited due to assay complexity and the resolution thereof is appreciably lower than the reconstructions being assessed. Accordingly, there is demand for new methods of reconstruction accuracy appraisal.

Results: Here we explore the potential of recently proposed stationary distributions, hereafter StatDns, derived from Hi-C contact matrices, to serve as a basis for reconstruction accuracy assessment. Current usage of such StatDns has focussed on the identification of highly interactive regions (HIRs): computationally defined regions of the genome purportedly involved in numerous long-range intra-chromosomal contacts. Consistent identification of HIRs would be informative with respect to inferred 3D architecture since the corresponding regions of the reconstruction would have an elevated number of k nearest neighbors (kNNs). More generally, we anticipate a monotone decreasing relationship between StatDn values and kNN distances. After initially evaluating the reproducibility of StatDns across replicate Hi-C data sets, we use this implied StatDn - kNN relationship to gauge the utility of StatDns for reconstruction validation, making recourse to both real and simulated examples.

Conclusions: Our analyses demonstrate that, as constructed, StatDns do *not* provide a suitable measure for assessing the accuracy of 3D genome reconstructions. Whether this is attributable to specific choices surrounding normalization in defining StatDns or to the logic underlying their very formulation remains to be determined.

Keywords: Chromatin conformation capture, Transition probability matrix, Nearest neighbors, 3D genome reconstruction, Normalization

*Correspondence: mark.segal@ucsf.edu
[1]Division of Bioinformatics, Department of Epidemiology and Biostatistics, UCSF, 550 16th Street, 94158 San Francisco, CA, USA
Full list of author information is available at the end of the article

Background

The spatial configuration of chromosomes is essential to various cellular processes, notably gene regulation. Conversely, architecture related alterations, such as translocations and gene fusions, are often cancer drivers. Accordingly, eliciting chromatin conformation is important. Such elicitation had been challenging due to chromatin compaction, dynamics and scale. However, the emergence of the suite of chromatin conformation capture assays, in particular Hi-C, generated new details of chromatin structure and spawned a number of subsequent biological findings [2, 9, 10, 18, 23]. Many of these findings have directly resulted from analyses of interaction or contact level data generated by Hi-C assays. Such data, usually obtained from bulk cell populations, record the frequency with which pairs of genomic loci (or bins thereof) are cross-linked, indicating spatial proximity of those loci within the nucleus. A less common Hi-C analysis paradigm proceeds by first converting these contact frequencies into distances, this transformation often invoking inverse power-laws [2, 13, 29, 35, 41]), and then generating a putative three dimensional (3D) reconstruction of the associated chromatin configuration via variants of multi-dimensional scaling (MDS). Such 3D reconstruction has been shown to enrich analyses based solely on the underlying contact map, these deriving, in part, from superposing genomic features. Examples include identifying co-localized genomic landmarks such as early replication origins [6, 37], expression gradients and co-localization of virulence genes in the malaria parasite *Plasmodium falciparum* [2], the impact of spatial organization on double strand break repair [14], and elucidation of '3D hotspots' corresponding to overlaid ChIP-Seq transcription factor maxima, revealing novel regulatory interactions [7].

But, any potential added value in analyses based on 3D reconstruction is conditional on the accuracy of the corresponding reconstruction and, appropriately, many concerns have been expressed regarding such accuracy. Firstly, the very notion of *a* single reconstruction being representative of the large ($\sim 10^6$) cell populations characterizing Hi-C assays is highly simplistic [19]. This issue has prompted reconstruction approaches [13, 33] that produce an ensemble of solutions, intended to capture inter-cell variation. However, whether these collections capture biologic, as opposed to algorithmic, variation is unclear [26, 35]. The recent development of high-throughput *single-cell* Hi-C assays [22, 31] provides an opportunity for systematic investigation of structural variation. Secondly, even at the single-cell level, genome conformation is dynamic with, for instance, obvious changes over the course of the cell cycle, as well as cell type specific. Finally, the lack of 3D chromatin structure gold standards makes accuracy assessment inherently problematic. To address this obstacle several authors have appealed to simulation [16, 20, 34, 35, 41, 42]. In order to deploy real data referents many of the same reconstruction algorithm developers have made recourse to fluorescence *in situ* hybridization (FISH) imaging as a means for gauging the accuracy of competing algorithms and/or tuning parameter settings. This approach proceeds by comparing measured distances between imaged probes with corresponding distances obtained from 3D reconstruction algorithms. These standard FISH-based methods, however, are tenuous due to the limited number of imaged probes ($\sim 2-6$, [18, 20, 29]) and the poor resolution thereof, many straddling over 1 megabase.

To improve on these accuracy assessment shortcomings we previously devised methods that centered on two newly devised biotechnologies [28]: (i) multiplex FISH [36] which provides an order of magnitude more probes, each at higher resolution, and hence two orders of magnitude more distances than conventional FISH, and (ii) a proximity-based ligation-free method, genome architecture mapping [3], predicated on sequencing DNA from a large collection of randomly-oriented, thin nuclear cryosections which enables determination of an internal measure of accuracy by evaluating how well the reconstruction conforms to the underlying collection of planar nuclear cryosections. However, these approaches to accuracy assessment have their own limitations. The primary drawback is that each biotechnology is experimentally intensive and, accordingly, has had minimal uptake. The resultant dearth of associated public data profoundly restricts the extent to which these approaches can be applied. Additionally, there is a resolution disparity, with Hi-C data being available at higher resolutions, mandating a coarsening of reconstructions prior to accuracy assessment.

In seeking to devise a more broadly applicable means for reconstruction accuracy assessment we were drawn to the recently proposed (Sobhy et al., [30], hereafter SKLLS) stationary distribution (hereafter StatDn(s)) of a Hi-C matrix and associated highly interactive regions (HIRs): computationally defined regions of the genome purportedly involved in numerous long-range intra-chromosomal contacts. Consistent identification of HIRs would be informative with respect to inferred 3D architecture since the corresponding regions of the reconstruction would have an elevated number of k nearest neighbors (kNNs) compared with non-highly interacting regions. More generally, we would anticipate a monotone decreasing relationship between StatDn values and kNN distances for fixed values of k. This posited relationship provides one means for evaluating the potential utility of StatDns, ithe objective of this paper, which is organized as follows. Under Methods we first recapitulate how StatDns are derived, highlighting normalization and interpretation

issues, and then detail data sources to be used in the evaluation thereof. The "Results" section showcases StatDn findings with respect to reproducibility across replicate Hi-C data sets, effects of normalization scheme, and performance for 3D reconstruction validation, via assessment of the above monotonicity between StatDn values and kNN distances, based on real and simulated examples. The Discussion frames conclusions based on the foregoing findings.

Methods

Stationary distributions from Hi-C contact matrices

Given a (possibly normalized – see below) symmetric, non-negative $n \times n$ observed contact matrix $O = [o_{ij}]$ the associated StatDn is generated as follows. First, O is standardized by dividing every entry by its row sum. This enables the key step: treating the resultant matrix, W, as a transition probability matrix (TPM), with entry w_{ij} interpreted as the probability of 'jumping' from node i to node j where 'nodes' denote a rebranding of the underlying Hi-C bins or loci, thereby allowing an overlay of graph / network concepts. The fact that, due to row sum based standardization, W is not symmetric complicates this interpretation since the original 'proximities' as measured via Hi-C are symmetric: $o_{ij} = o_{ji}$. SKLLS proceed by prescribing a Markov model with TPM W. Let $p_i(t)$ be the probability of occupying node i at time t and $p(t) = (p_1(t), p_2(t), \ldots, p_n(t))$ be the corresponding probability distribution. Then, under the Markov assumption, transitions occur according to

$$p(t+1) = p(t)W \tag{1}$$

The limiting ($t \to \infty$) StatDn, designated $p(\infty)$, satisfies $p(\infty) = p(\infty)W$, and is given by the (left) eigenvector corresponding to the (largest) eigenvalue one, the non-negative entries of $p(\infty)$ being normalized to sum to one. We use the R package `RSpectra` [21] to perform the requisite spectral decomposition.

SKLLS categorize StatDns, at 30^{th}, 50^{th}, 80^{th} and 90^{th} percentiles, and deploy the resultant ordered categories in downstream analyses, with an emphasis on HIRs corresponding to the latter upper decile. In contrast, we utilize StatDns in their native, continuous form obviating the need for thresholding. As a check, we extracted SKLLS-defined categories and reprised select analyses with concordant findings.

Normalization and interpretation issues

There has been extensive discussion surrounding normalization issues for Hi-C data and development of companion corrective methods [8, 11, 12, 17, 38]. Much of this effort pertains to mitigating systematic biases affecting observed o_{ij} values deriving from factors such as fragment length, GC content and mappability. A distinct aspect of some normalization strategies concerns removing 'expected' contact counts from the observed values so as to adjust for contiguity and thereby emphasize features of interest such as loops. In this context expected values are often computed as a function of genomic distance [2, 10]. This equates to applying a common correction within each diagonal of O, elements thereof being equispaced with respect to genomic distance, presuming equal sized contact matrix bins as is standard. It is this approach that is considered by SKLLS.

Specifically, for each of the n diagonals of O, the median of the corresponding entries is obtained. An $n \times n$ expectation matrix E with constant diagonals is then created, the constants being the respective medians. In addition to obtaining StatDns (as detailed above) from (unnormalized) O, they are also generated from $O - E$ and O/E. To satisfy the non-negativity requirement of a TPM any negative values arising post normalization are replaced with a small positive constant. For $O - E$ normalization, with E based on diagonal medians, this means that approximately half the entries will be replaced by this constant. The ramifications, both interpretive and performance-wise, of such wholesale substitution are unclear.

In order to decide between the competing normalization schemes SKLLS assert that $O - E$ normalization produces StatDns with a larger 'dynamic range' than O or O/E approaches, and is accordingly preferred. Presuming dynamic range is defined as the difference between maximum and minimum StatDn values, the rationale for its selection as a normalization criterion is obscure. Moreover, it will be susceptible to the influence of outliers as can arise from extreme (normalized) contact matrix row sums. The supporting evidence presented for choosing $O - E$ consists of visually comparing StatDns from the three schemes over a limited range of a single chromosome. Further, it is claimed that, in using O directly, the inclusion of both short- and long-range contacts attenuates dynamic range but the basis for this is unclear.

It is pertinent to consider StatDns, as operationalized above, arising from specific patterned matrices. For a compound symmetric (exchangeable) matrix the StatDn is constant ($p_i(\infty) = 1/n \ \forall i$) irrespective of the value of the off-diagonal entries, with this same StatDn resulting from a tri-diagonal matrix, again independent of the value of the off-diagonal entries [25]. While these patterns don't reflect $O, O - E, O/E$ matrices arising in practice, the lack of StatDn discrimination between such appreciably different matrices raises interpretative concerns about the proposed approach, at least from the perspective of evaluating 3D reconstructions, and potentially beyond.

Data sources and simulated 3D structures

Hi-C data [23] for GM12878 cells was obtained from the Gene Expression Omnibus (GEO) with accession

GSE63525. Contact matrices deriving from several series of experiments were grouped (by the original authors) into 'primary' and 'replicate' datasets and we utilize these to assess reproducibility, as has been done previously [28]. Hi-C data [9] for IMR90 cells was obtained from the Gene Expression Omnibus (GEO) with accession GSE35156. For both cell types analyses were restricted to reads with alignment mapping quality scores $\geq$ 30 and conducted with contact matrices at 25kb resolution since this corresponds to the resolution of SKLLS defined HIRs.

Noised-up versions of simulated chain-like and topologically associated domain (TAD)-like structures and attendant contact maps obtained under differing regimes have been used to evaluate 3D reconstruction algorithms in settings intended to recapitulate practice [34, 42]. Similarly, simulated helical and random walk structures have been used for this purpose [42]. Here we follow an analogous agenda by (i) computing StatDns from the contact matrices provided using each of the normalization schemes described above, and (ii) comparing these to the corresponding structures using k nearest neighbors as described subsequently.

As an illustration of how such synthetic data is obtained we present a brief overview of the formulation used for helical structures following Zou et al., [42]. O_{ij}, the $(i,j)^{th}$ entry of the observed contact matrix O, is generated as a random Poisson variate with rate parameter λ_{ij}. In turn, this parameter is set using the abovementioned inverse power-law transformation: $\lambda_{ij} = c/d_{ij}^{\alpha}$. Here d_{ij} corresponds to the distance between the i^{th} and j^{th} points on the helix, α is fixed at 1.5, and c varies so as to govern the signal coverage – the percentage of non-zero entries in the contact matrix. For the results presented subsequently we obtain 100 points on a helix defined by coordinate functions

$$x(t) = 2\sin(t/3); \quad y(t) = 2\cos(t/3);$$
$$z(t) = t/20; \quad t = 1, \ldots, 100.$$

and set c to yield 25% signal coverage, with similar findings at 90% coverage.

Obtaining 3D genome reconstructions from Hi-C data

Use of simulated 3D architectures and associated contact maps, as above, in evaluating StatDns as a validation tool has the advantage of eliminating uncertainties inherent in the reconstruction process. Nonetheless, it is purposeful to assess StatDns using real data reconstructions, reflecting use in practice.

Multi-dimensional scaling

As noted in the Background, there are numerous approaches for generating 3D reconstructions from Hi-C contact maps and, in turn, most of these feature several tuning parameters. In order not to obscure our purpose of appraising StatDns we showcase findings from a simple, minimal-assumption approach to reconstruction: multi-dimensional scaling, fit using the R package `smacof` [15]. MDS is an established approach to finding configurations that recapitulate dissimilarity measures which, in turn, can be obtained from Hi-C contacts, by power-law transformation for example. Accordingly, MDS-based approaches have been widely used in the context of genome reconstruction [2, 4, 16, 24, 27, 29, 32, 35, 41].

Under MDS we seek a 3D configuration $X = \{\vec{x}_1, \ldots, \vec{x}_n\}; \vec{x}_j \in R^3$ that best fits the dissimilarity matrix D according to:

$$\min_{\{\vec{x}_1,\ldots,\vec{x}_n | \sum \vec{x}_i = 0\}} \sum_{\{i,j | D_{ij} < \infty\}} \omega_{ij} \cdot (\|\vec{x}_i - \vec{x}_j\| - D_{ij})^2 \qquad (2)$$

Though confining our attention to MDS, we explored a variety of schemes within this framework, using both metric and non-metric scaling, and varying dissimilarity weights ω_{ij} whereby downweighting of imprecise contact counts can be accommodated, and power-law indices for transforming O to D. We note that irrespective of MDS reconstruction method examined results were largely similar.

Hamiltonian simulated annealing

In order for findings not to be solely reliant on a single (MDS) reconstruction strategy – although, as noted, a range of MDS specifications were examined – we additionally applied the Hamiltonian simulated annealing (HSA, [42]) algorithm. HSA has a number of compelling attributes: (i) it can simultaneously handle multiple data tracks allowing for integration of Hi-C contact data from differing restriction enzyme digests; (ii) it can adaptively estimate the power-law index whereby contacts are transformed to distances, the importance of which has been previously emphasized [41]; and (iii) by using simulated annealing combined with Hamiltonian dynamics it can effectively optimize over for the high dimensional space representing the genomic loci's 3D coordinates.

Analogous to other 3D reconstruction algorithms [20, 35], HSA models (normalized) contact counts, n, via Poisson regression:

$$n_{i_k j_k} \sim Poi(\mu_{i_k j_k}), \qquad k = 1, \ldots, K \qquad (3)$$
$$\ln(\mu_{i_k j_k}) = \beta_{k0} + \beta_{k1} \ln(d_{i_k j_k}) \qquad (4)$$
$$d_{i_k j_k} = ||X_{i_k} - X_{j_k}||_2 \qquad (5)$$

where in (3) k indexes track and $n_{i_k j_k}$ is the count for genomic loci i_k, j_k. The parameters β_{k1} are (track specific) power-law indices relating expected counts (μ) to Euclidean distances (d). Covariates such as GC content and fragment length can be included in (4) in order to

facilitate in-line normalization. The $X_{i_k} = (x_{i_k}, y_{i_k}, z_{i_k})$ and $X_{j_k} = (x_{j_k}, y_{j_k}, z_{j_k})$ in (5) are the 3D coordinates for loci i_k, j_k and constitute the unknown parameters providing the reconstruction. These are subject to constraints designed to capture the local contiguity of chromatin, represented by induced dependencies of a hidden Gaussian Markov chain. The full log-likelihood for β, X is then

$$\ln(L(\beta, X|\mu, i_k, j_k) \propto \sum_k \sum_{i_k, j_k} \left[-\exp(\ln(\mu_{i_k j_k}) + n_{i_k j_k}(\ln(\mu_{i_k j_k}))) \right] \tag{6}$$

to which a penalty term controlling local smoothness is added. Note that (constrained) X enters (6) through μ and d from (4) and (5) respectively. The resulting penalized likelihood is optimized by iterating between generalized linear model (GLM, *cf* Poisson regression) fitting to obtain estimates $\hat{\beta}$ and simulated annealing to obtain estimates of the 3D coordinates $\hat{X} = (\hat{x}, \hat{y}, \hat{z})$. Several tuning parameters control the simulated annealing search and we used default values, as established by the authors' for their custom R scripts.

Stationary distribution reproducibility

We assessed the reproducibility – between primary and replicate data series – of StatDns obtained under the differing normalization schemes – using scatterplot smoothing and associated correlations. We contrast these correlations with stratum-adjusted correlation coefficients (SCCs) of the corresponding Hi-C data. SCCs, described below, are custom correlation measures developed for Hi-C contact matrices that reflects the same constant diagonal expected counts described above which, on average, decreases substantially as genomic distance increases [39].

The SCC is based on the generalized Cochran-Mantel-Haenszel statistic, M^2, which is used for testing whether two variables are associated while being stratified by a third variable [1]. Since the magnitude of M^2 depends on sample size it does not provide a direct measure of association strength. In the unstratified setting we have the relationship $\rho^2 = M^2/(n-1)$ where ρ is the Pearson correlation coefficient and n is the number of observations. This relationship underscores the derivation of the SCC to measure association in the presence of stratification. Let (X, Y) denote a pair of samples (here contact matrices) with n observations stratified into K strata (here diagonal bands corresponding to equal genomic distances), each having n_k observations so that $\sum_{k=1}^{K} n_k = n$. Let the observations in stratum k be $(x_{i_k}, y_{i_k}); i = 1, \ldots, K$ with associated random variables (X_k, Y_k).

The Pearson correlation coefficient ρ_k for the k^{th} stratum is $\rho_k = r_{1k}/r_{2k}$, where

$$\begin{aligned} r_{1k} &= E(X_k Y_k) - E(X_k)E(Y_k) \\ &= \frac{\sum_{i=1}^{n_k} x_{i_k} y_{i_k}}{n_k} - \frac{\sum_{i=1}^{n_k} x_{i_k} \sum_{j=1}^{n_k} y_{j_k}}{n_k^2} \\ r_{2k}^2 &= Var(X_k) Var(Y_k) \\ &= \left[\frac{\sum_{i=1}^{n_k} x_{i_k}^2}{n_k} - \left(\frac{\sum_{i=1}^{n_k} x_{i_k}}{n_k} \right)^2 \right] \left[\frac{\sum_{i=1}^{n_k} y_{i_k}^2}{n_k} - \left(\frac{\sum_{i=1}^{n_k} y_{i_k}}{n_k} \right)^2 \right] \end{aligned}$$

It is straightforward to represent M^2 in terms of a weighted sum of the ρ_k which gives rise to the SCC defined as

$$\rho_s = \sum_{k=1}^{K} \left(\frac{n_k r_{2k}}{\sum_{k=1}^{K} n_k r_{2k}} \right) \rho_k. \tag{7}$$

Further aspects of SCCs, including obtaining the variance of ρ_s, deploying variance stabilizing weights in computing ρ_s, guidelines for determining the number of strata K are detailed in Yang et al., [39], with fitting making recourse to R package `hicrep` [40].

Comparing stationary distributions and 3D genome reconstructions

For each locus of a 3D structure, either simulated or obtained via reconstruction, we compute the distance to its k^{th} nearest neighbor (kNN) in the structure, for $k \in \Omega = \{5, 15, 25\}$, using the R package `FNN` [5]. Since kNN distances are monotone in k it suffices to consider a few select values. We plot these kNN distances against StatDn values obtained from the corresponding contact matrix. We again use scatterplot smoothing (R function `lowess`) to highlight relationships, with a monotone decreasing association anticipated if StatDn identification of highly (and remotely) interacting loci are supported by the structure. To appreciate the basis for this monotone decreasing relationship consider the antithesis of a HIR, namely a minimally interacting region, characterized by low StatDn values. By virtue of its minimal interactions nearest neighbor distances for given $k \in \Omega$ will be large. The converse holds for HIRs and the underlying high StatDn values leading to the monotone decreasing relationship between StatDns and kNN distances.

Results

Our findings are presented largely by way of figures. These are constructed so that comparisons between $O, O - E, O/E$ normalizations are highlighted. But, more important than these internal contrasts are overall assessments of StatDns for the stated objective of appraising 3D reconstructions. In most of the settings considered the overall performance is such that StatDns cannot be endorsed

as a 3D reconstruction evaluation technique since the abovementioned monotone decreasing relationship with kNN distances fails to hold. Moreover, examples wherein anomalous behavior of StatDns is exhibited are showcased.

We report results for GM12878 chromosome 9 since this exhibits the highest density (per base) of HIRs as defined by SKLLS. We also present results for GM12878 chromosome 4 which is relatively sparse with respect to HIRs. However, similar trends were consistently observed across all chromosomes examined (not shown). Additionally, findings from select IMR90 cells are illustrated, revealing instances of StatDn breakdown.

Stationary distribution reproducibility

In Fig. 1 we compare the StatDns of GM12878 cells chromosome 9 primary and replicate series corresponding to respective normalizations $O, O-E, O/E$. The respective correlations are 0.962, 0.937 and 0.977 whereas the SCC between the primary and replicate contact matrices is

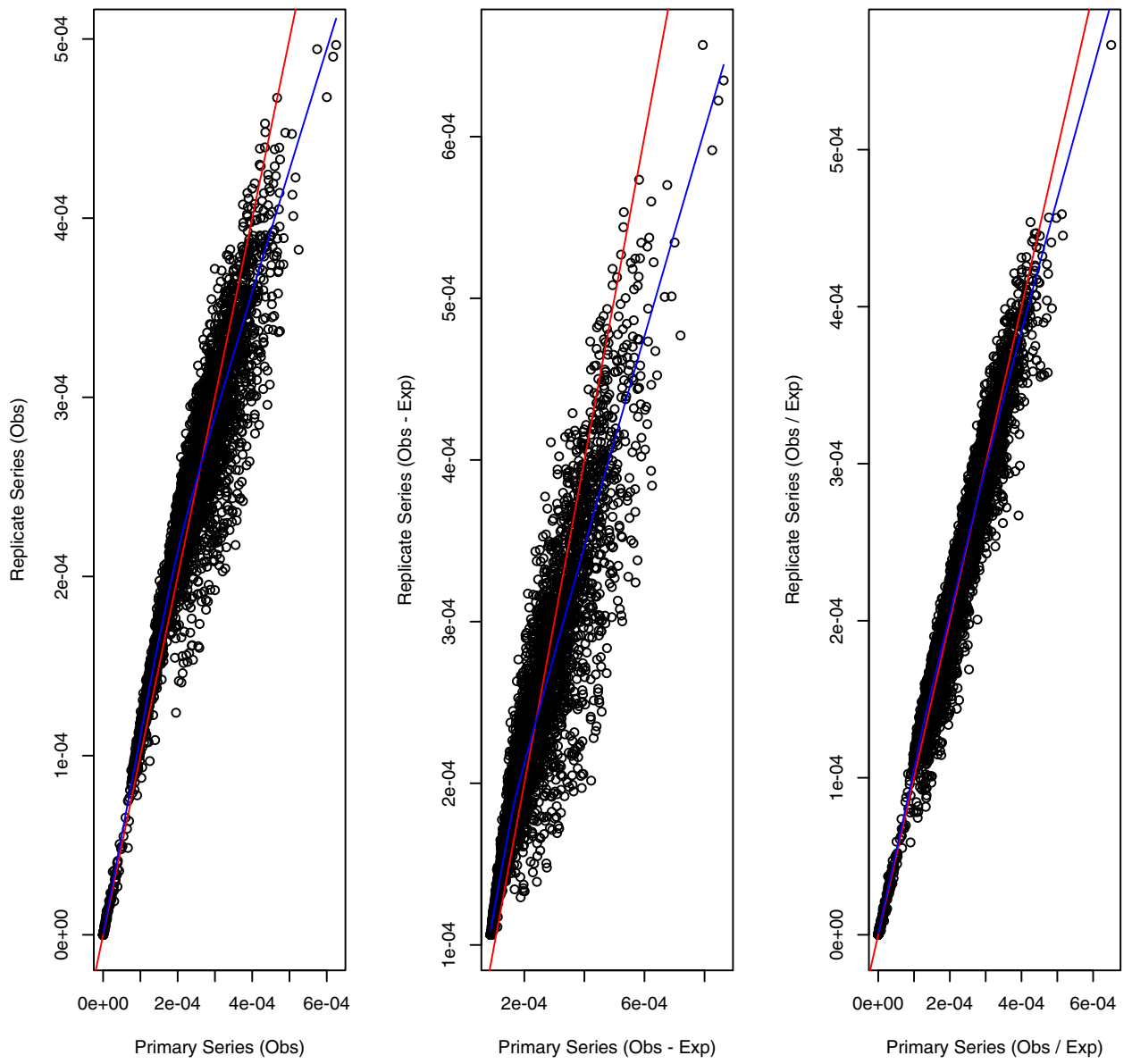

Fig. 1 StatDn reproducibility for GM12878 Chromosome 9. Agreement between Stationary Distributions obtained from primary and replicate series Hi-C data at 25kb resolution [23]. StatDn normalization schemes are O (left panel), $O-E$ (middle) and O/E (right). In each panel the identity line is in red and the lowess smooth is in blue

0.966. Thus, reproducibility for the $O - E$ normalization chosen by SKLLS is furthest removed from the correlation between the underlying contact matrices.

More interesting findings emerge when we similarly assess reproducibility for IMR90 cells. Figure 2 displays the StatDns for IMR90 chromosome 21 primary and replicate series, again corresponding to respective normalizations $O, O - E, O/E$. The corresponding correlations are 0.935, 0.936 and 0.966, whereas the SCC between the primary and replicate contact matrices is 0.808. Thus, the StatDn correlations appreciably exceed the SCC between the underlying contact matrices, indicative of possible problems with StatDns in view of the careful and contact map customized construction of SCCs [39].

Also apparent in Fig. 2 are StatDn outliers, for both O and the chosen $O - E$ normalizations, which result from (relatively) extreme contact matrix row sums, indicating possible normalization breakdown for such instances. An even more dramatic example of anomalous StatDn values is shown below with respect to reconstruction (Fig. 8).

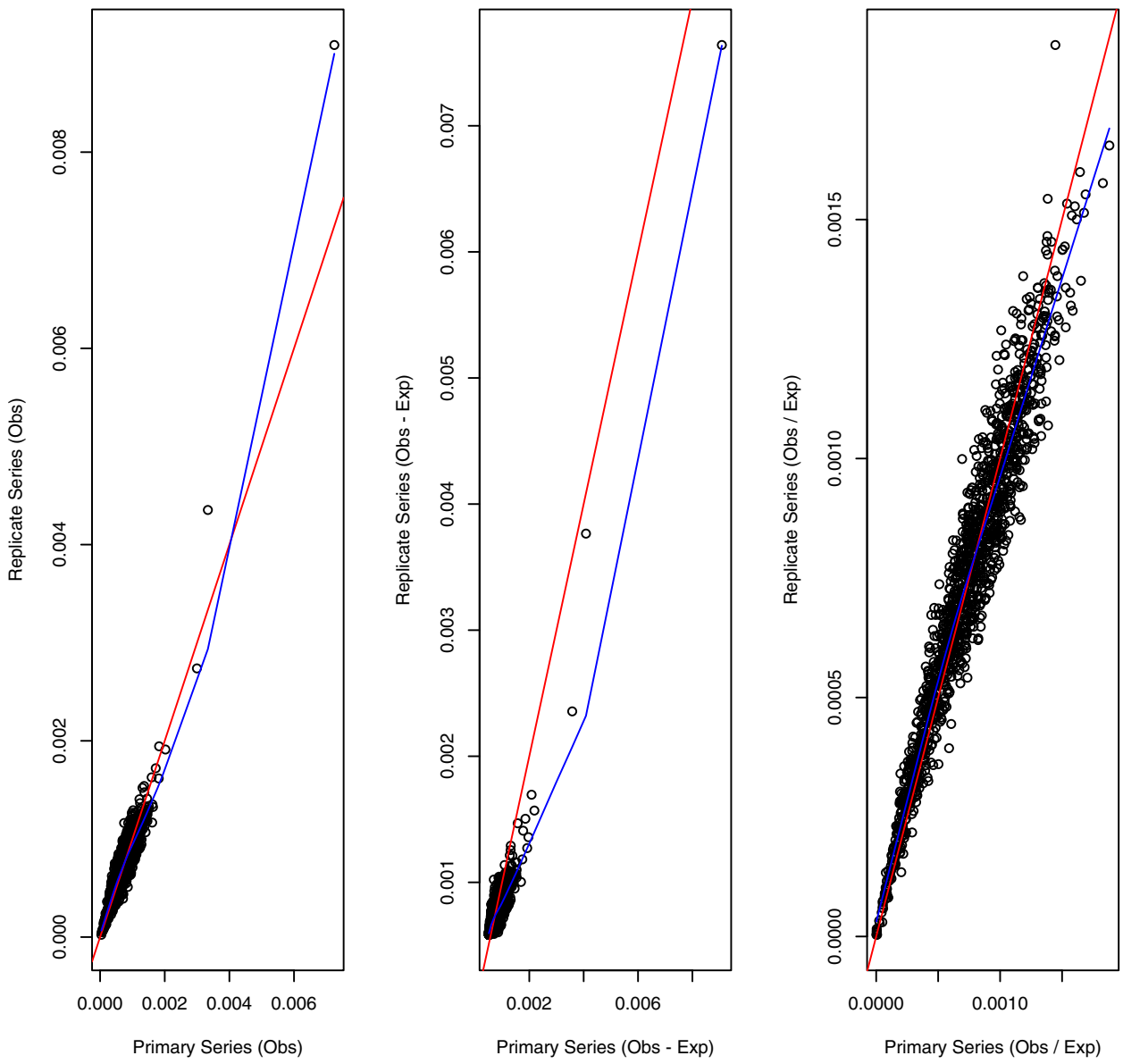

Fig. 2 StatDn reproducibility for IMR90 chromosome 21. Agreement between Stationary Distributions obtained from primary and replicate series Hi-C data at 25kb resolution [9]. StatDn normalization schemes are O (left panel), $O - E$ (middle) and O/E (right). In each panel the identity line is in red and the lowess smooth is in blue

Relating stationary distributions to 3D structures

The simulated helical and random walk structures previously used for 3D reconstruction evaluation [42] include instances varying according to the extent of signal coverage, defined as the percentage of non-zero entries in the contact matrix derived from the generated structure. Here we illustrate results for the lowest levels of signal coverage: 25% and 10% for the helix and random walk respectively. Findings at higher levels of signal coverage are similar (not shown) although the helical structure with 90% signal coverage does not display a monotone decreasing relationship between kNN distances and StatDns with O/E normalization.

Results for the simulated helical structure, based on 100 loci, are presented in Fig. 3. The quantal nature of the kNN distances (we display results for $k = 5, 15$) – for example, there are only three distinct 5 nearest neighbor distances – reflects the regularity of the helical configuration. The left and right panels, corresponding to O and O/E normalization, exhibit decreasing trends: the higher the StatDn value, nominally corresponding to loci with greater numbers of interactions, the smaller the kNN distance in the

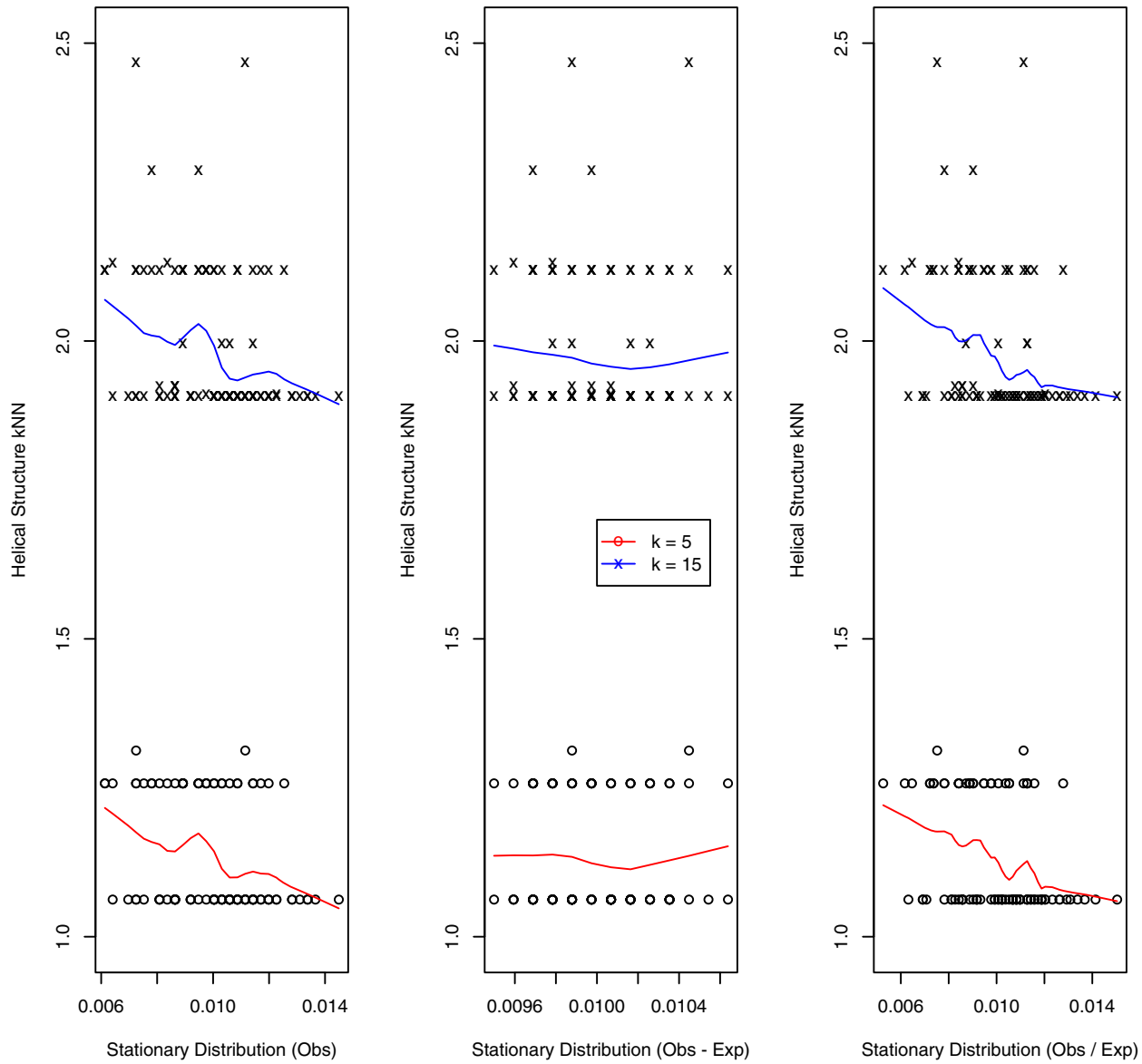

Fig. 3 Helical structure: kNNs versus StatDns. Relationships between k nearest neighbors and StatDns for $k = 5$ (o, red lowess smooth) and $k = 15$ (x, blue lowess smooth) for the simulated helical structure generated to have 25% signal coverage (percentage of non-zero contact matrix entries) per [42]. StatDn normalization schemes are O (left panel), $O - E$ (middle) and O/E (right)

structure, as would be expected. However, for the middle panel, corresponding to $O-E$ normalization, no such relationship is evident. Further, by virtue of the manner whereby $O-E$ normalization handles non-positive values, there is substantial duplication of StatDn values: 47 uniques versus 97 for $O, O/E$. Results for the random walk structure are presented in Fig. 4. Here we see very similar performance across normalization schemes with the anticipated decreasing relationship exhibited for each.

A comprehensive effort to generate structures and attendant contact matrices that more realistically reflect chromatin architecture has been undertaken by Trussart et al., [34]. Here we focus on two such structures, TAD-like and chain-like, each generated with mid-level noise and structural variability corresponding to Trussart et al., parameter settings of $\alpha = 100$ and $\Delta t = 10^3$ respectively. Results for the TAD-like structure are presented in Fig. 5 and for the chain-like structure in Fig. 6. For both structures we observe StatDns displaying an *increasing* relationship with *k*NN distances, this being strongest for $O-E$ normalization.

Results from StatDn evaluation of a reconstruction for GM12878 chromosome 9 via unweighted metric MDS are depicted in Fig. 7. While the left and

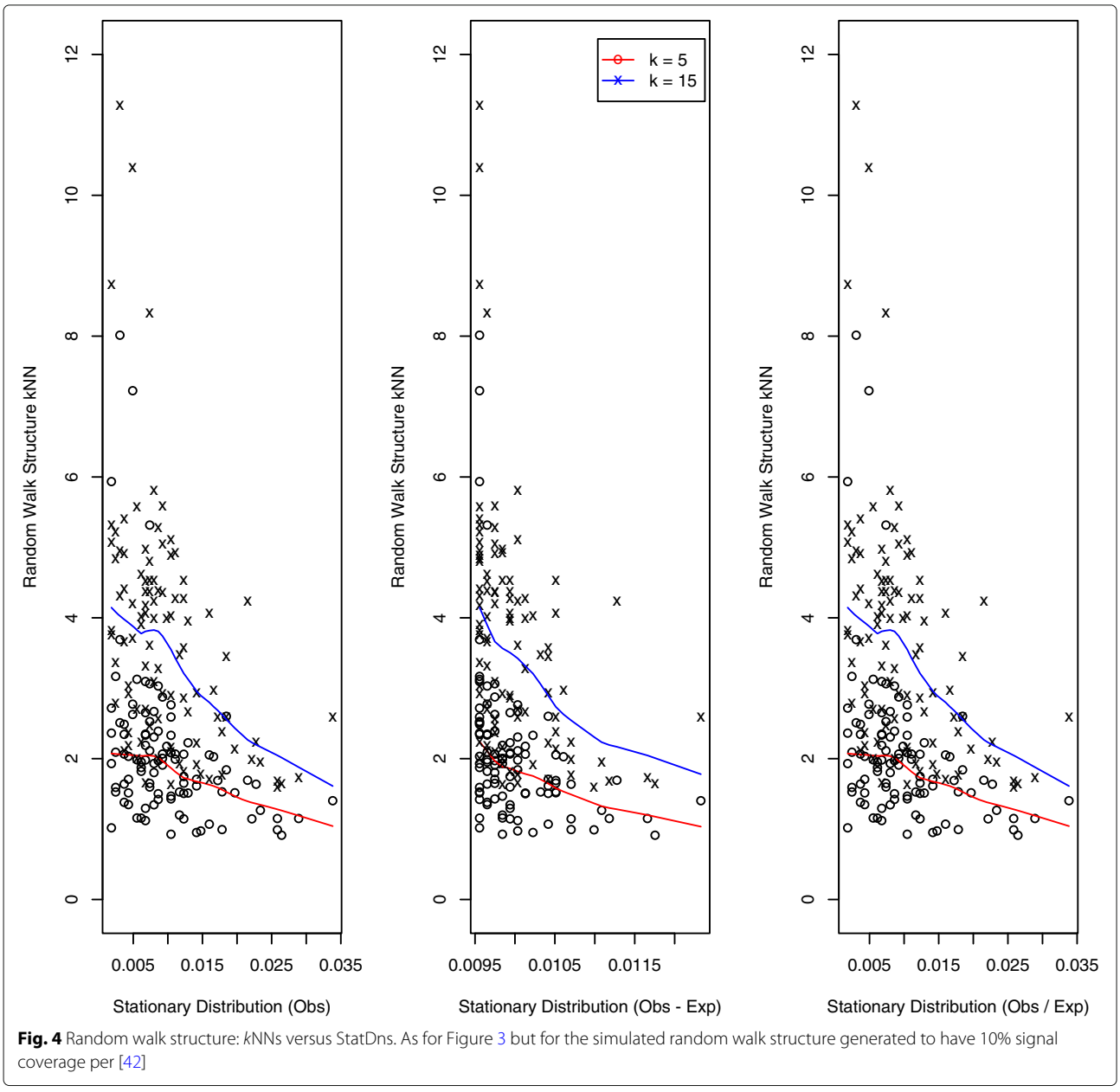

Fig. 4 Random walk structure: *k*NNs versus StatDns. As for Figure 3 but for the simulated random walk structure generated to have 10% signal coverage per [42]

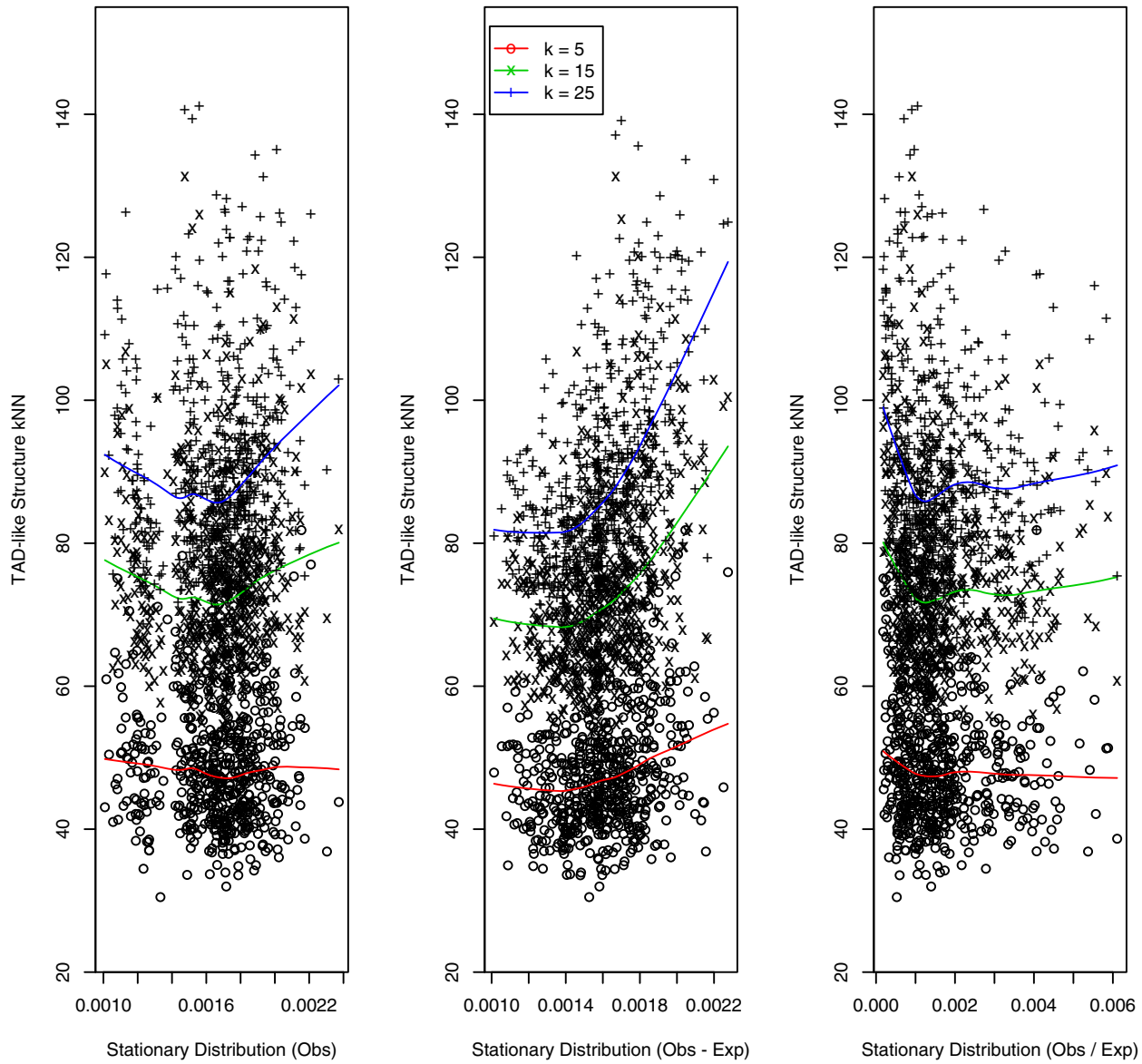

Fig. 5 TAD-like structure: *k*NNs versus StatDns. As for Fig. 3 but for the simulated TAD-like structure generated to have mid-level noise and structural variability ($\alpha = 100$ and $\Delta t = 10^3$) per [34] and with *k*NNs: $k = 5$ (o, red lowess smooth), $k = 15$ (x, green lowess smooth) and $k = 25$ (+, blue lowess smooth)

right panels corresponding to O and O/E normalization display decreasing relationships with *k*NN distances these are driven by elevated *k*NN values for small StatDn probabilities. Results for $O - E$ normalization are effectively constant. Analogous findings were obtained from other (weighted, non-metric) MDS reconstruction approaches, as well as for HSA-based reconstruction.

Similarly, results from StatDn evaluation of a reconstruction for IMR90 chromosome 21 by HSA are depicted in Fig. 8. Here the left and middle panels corresponding to O and $O - E$ normalization display decreasing relationships with *k*NN for the bulk of the data but exhibit increasing trends in the upper tail: the region containing the HIR. These same trends were evident in reconstructions obtained using MDS.

Discussion

Many potential difficulties surrounding use of StatDns were delineated in Methods under Normalization and

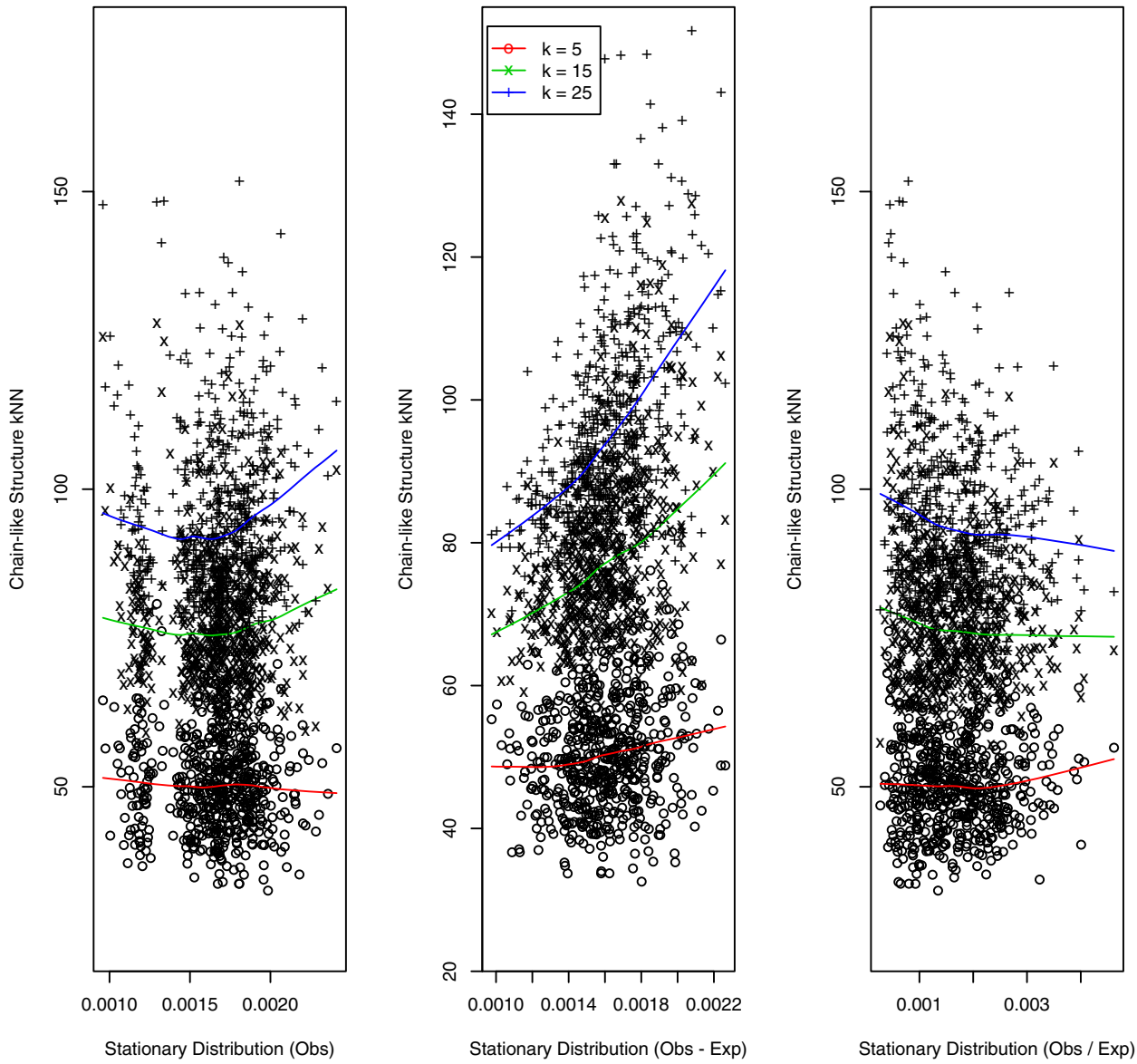

Fig. 6 Chain-like structure: *k*NNs versus StatDns. As for Fig. 5 but for the simulated chain-like structure

Interpretation Issues and these concerns have been borne out by the empirical results. It is important to note that these problems cannot be ascribed to deficiencies of the reconstruction algorithms since they are also exhibited with simulated structures that bypass the reconstruction step. Moreover, for some of the explorations based on chromatin configuration reconstruction, we have deliberately opted to utilize a minimalist MDS approach, thereby limiting the influence of assumptions and parameter tuning. These findings, wherein StatDns do not recapitulate inferred 3D MDS reconstructions, also pertain to an alternate state-of-the-art reconstruction algorithm, HSA, and hold across all cell lines and chromosomes examined. Thus, the overall weight of evidence, both theoretic and empirical, is such that StatDns, especially those based on the prescribed $O - E$ normalization, cannot be recommended as a means for evaluating 3D genome reconstruction. Indeed, these problematic underpinnings of StatDns, including the logic surrounding their definition, call into question their usage for any purpose, not just reconstruction assessment as examined here.

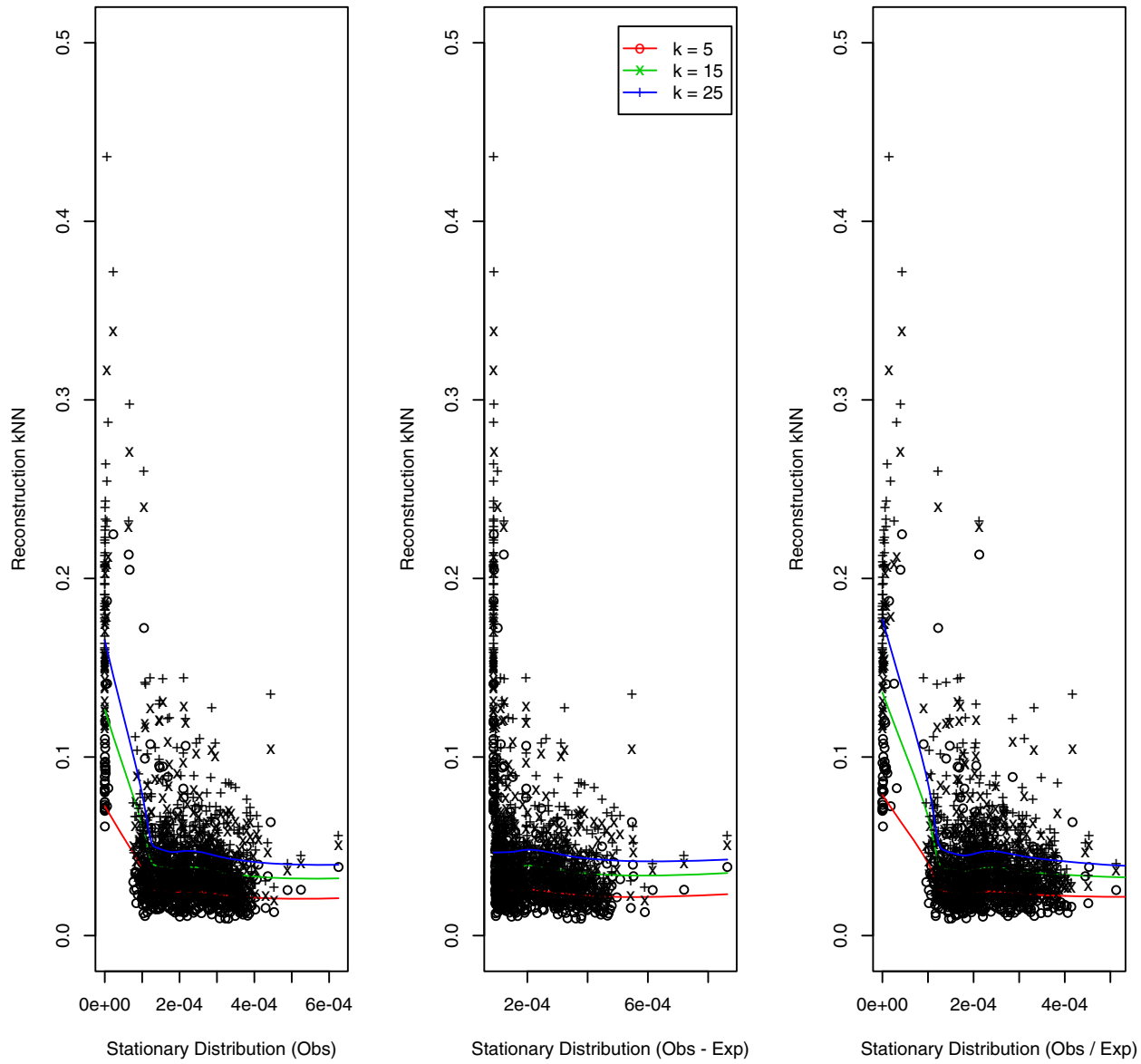

Fig. 7 GM12878 Chromosome 9: *k*NNs versus StatDns. As for Fig. 5 but for reconstructed GM12878 Chromosome 9 where the reconstruction utilized unweighted metric MDS. While plotted points correspond to 500 randomly sampled loci (≈ 10% of the total), the depicted lowess smooths are based on the entire sample

This conclusion begs the question as to whether alternate, established structural units derived from Hi-C contact matrices, such as TADs [9] and contact domains [23], might serve as components for (non-orthogonal) reconstruction assessment. However, these constructs are by definition local and so do not provide a basis for effecting large-scale structure interrogation. It was the purported ability of StatDns to capture frequent, long-range interactions that motivated this evaluation of their validation potential. Conversely, TADs [24] and FISH distances [29] have been used to improve the reconstruction process itself. Again, given their uncertain foundation, we see no analogous role for StatDns.

Conclusion

Our analyses demonstrate that, as constructed, StatDns do *not* provide a suitable measure for assessing the accuracy of 3D genome reconstructions. Whether this is attributable to specific choices surrounding their formulation or to the logic underlying their very definition remains to be determined.

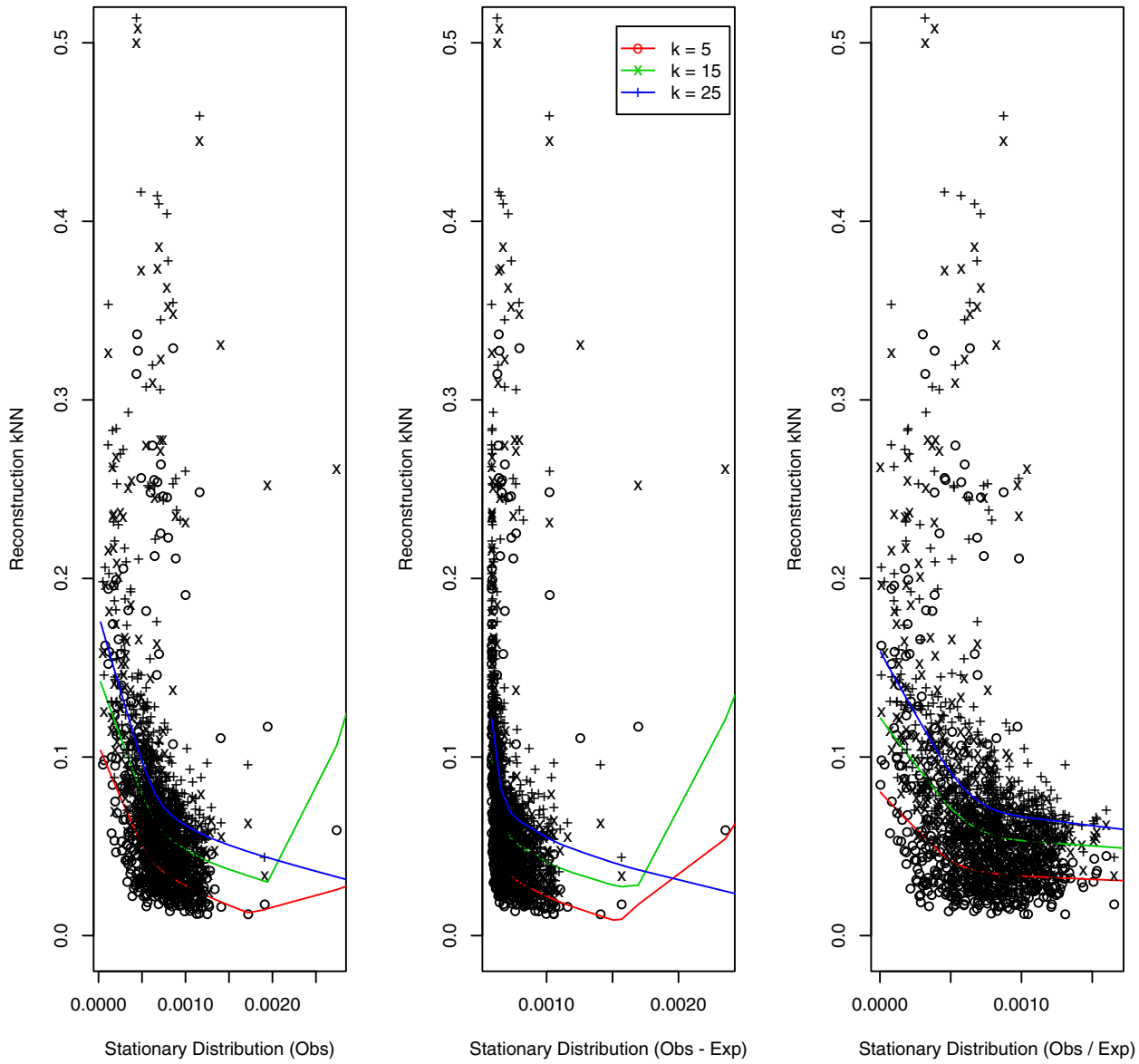

Fig. 8 IMR90 Chromosome 21: *k*NNs versus StatDns. As for Fig. 7 but for reconstructed IMR90 Chromosome 21 where the reconstruction utilized HSA. While plotted points correspond to 500 randomly sampled loci (≈ 35% of the total), the depicted lowess smooths are based on the entire sample

Abbreviations

3D: Three dimensional; FISH: Fluorescence *in situ* hybridization; GEO: Gene expression Omnibus; HIRs: Highly interactive regions; HSA: Hamiltonian simulated annealing; *k*NNs: *k* Nearest neighbors; MDS: Multi-dimensional scaling; SCC: Stratified correlation coefficient; SKLLS: Sobhy, Kumar, Lewerentz, Lizana, Stenberg; StatDn: Stationary distribution; TAD: Topologically associated domain; TPM: Transition probability matrix

Acknowledgements

The authors thank Trevor Hastie for helpful comments.

Authors' contributions

MRS conceived the study, performed analyses and wrote the manuscript. KF-B performed analyses. Both authors have read and approved the final manuscript.

Funding

Support was provided by NIH grant R01GM109457. The funding body played no role in the design of the study, the collection, analysis, and interpretation of data, or in the writing of the manuscript.

Availability of data and materials

Hi-C data for GM12878 cells is available from GEO with accession GSE63525: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. Hi-C data for IMR90 cells is available from GEO with accession GSE35156: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35156.
Contact maps and associated structures corresponding to chain-like and TAD-like models [34] were obtained from http://sgt.cnag.cat/3dg/datasets/. The noised-up helical (regular) and random walk structures and attendant contact matrices utilized in [42] are available from https://people.umass.edu/ouyanglab/hsa/downloads.html#Data.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
MRS is an Associate Editor for *BMC Bioinformatics*.

Author details
[1]Division of Bioinformatics, Department of Epidemiology and Biostatistics, UCSF, 550 16th Street, 94158 San Francisco, CA, USA. [2]Computational Biology, 23andMe, Inc., 899 West Evelyn Avenue, Mountain View, CA 94041, USA.

Received: 3 November 2019 Accepted: 17 February 2020
Published online: 24 February 2020

References

1. Agresti A. Categorical data analysis. 3rd Ed. New York: Wiley; 2012.
2. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG. Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. 2014;24(6):974–88.
3. Beagrie RA, Scialdone A, Schueler M, Kraemer DC, Chotalia M, Xie SQ, Barbieri M, de Santiago I, Lavitas LM, Branco MR, Fraser J, Dostie J, Game L, Dillon N, Edwards PA, Nicodemi M, Pombo A. Complex multi-enhancer contacts captured by genome architecture mapping. Nature. 2017;543(7646):519–24.
4. Ben-Elazar S, Yakhini Z, Yanai I. Spatial localization of co-regulated genes exceeds genomic gene clustering in the *Saccharomyces cerevisiae* genome. Nucleic Acids Res. 2013;41(4):2191–201.
5. Beygelzimer A, Kakadet S, Langford J, Arya S, Mount D, Li S. FNN: Fast nearest neighbor search algorithms and applications. R package version 1.1.3. 2019. https://CRAN.R-project.org/package=FNN. Accessed 2019.
6. Capurso D, Segal MR. Distance-based assessment of the localization of functional annotations in 3D genome reconstructions. BMC Genomics. 2014;15:992.
7. Capurso D, Bengtsson H, Segal MR. Discovering hotspots in functional genomic data superposed on 3D chromatin configuration reconstructions. Nucleic Acids Res. 2016;44(5):2028–35.
8. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC Genomics. 2012;13:436.
9. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80.
10. Duan Z, Andronescu M, Schutz K, Mcllwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. A three-dimensional model of the yeast genome. Nature. 2010;465(7296):363–7.
11. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28(23):3131–3.
12. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Meth. 2011;9:999–1003.
13. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotech. 2011;30:90–8.
14. Lee CS, Wang RW, Chang HH, Capurso D, Segal MR, Haber JE. Chromosome position determines the success of double-strand break repair. Proc Natl Acad Sci. 2016;113(2):E146–54.
15. de Leeuw J, Mair P. Multidimensional scaling using majorization: SMACOF in R. J Stat Softw. 2009;31(3):1–30.
16. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J. 3D genome reconstruction from chromosomal contacts. Nat Meth. 2014;11(11):1141–3.
17. Li W, Gong K, Li Q, Alber F, Zhou XJ. HiCorrector: A fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. Bioinformatics. 2015;31(6):960–2.
18. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range contacts reveals folding principles of the human genome. Science. 2009;326:289–93.
19. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502(7469):59–64.
20. Park J, Lin S. A random effect model for reconstruction of spatial chromatin structure. Biometrics. 2017;73(1):52–62.
21. Qiu Y, Mei J. RSpectra: Solvers for large-scale eigenvalue and SVD problems. R package version 0.15-0. 2019. https://CRAN.R-project.org/package=RSpectra. Accessed 2019.
22. Ramani V, Deng X, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, Shendure J. Massively multiplex single-cell Hi-C. Nat Meth. 2017;14(3):263–6.
23. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
24. Rieber L, Mahony S. miniMDS: 3D structural inference from high-resolution Hi-C data. Bioinformatics. 2017;33:i261–6.
25. Segal MR. Representative curves for longitudinal data via regression trees. J Comp Graph Stat. 1994;3:214–33.
26. Segal MR, Xiong H, Capurso D, Vazquez M, Arsuaga J. Reproducibility of 3D chromatin configuration reconstructions. Biostatistics. 2014;15(3):442–56.
27. Segal MR, Bengtsson HL. Reconstruction of 3D genome architecture via a two-stage algorithm. BMC Bioinformatics. 2015;16:373.
28. Segal MR, Bengtsson HL. Improved accuracy assessment for 3D genome reconstructions. BMC Bioinformatics. 2018;19(1):196.
29. Shavit Y, Hamey FK, Lio P. FisHiCal: an R package for iterative FISH-based calibration of Hi-C data. Bioinformatics. 2014;30(21):3120–2.
30. Sobhy H, Kumar R, Lewerentz J, Lizana L, Stenberg P. Highly interacting regions of the human genome are enriched with enhancers and bound by DNA repair proteins. Sci Rep. 2019;9(1):4577.
31. Stevens TJ, Lando D, Basu S, Atkinson L, Cao Y, Lee S, Leeb M, Wohlfahrt KJ, Boucher W, O'Shaughnessy-Kirwan A, Cramard J, Faure AJ, Ralser M, Blanco E, Morey L, Sansó M, Palayret MGS, Lehner B, Di Croce L, Wutz A, Hendrich B, Klenerman D, Laue ED. 3D structure of individual mammalian genomes studied by single cell Hi-C. Nature. 2017;544(7648):59–64.
32. Szalaj P, Tang Z, Michalski P, Pietal MJ, Luo OJ, Sadowski M, Li X, Radew K, Ruan Y, Plewczynski D. An integrated 3-Dimensional Genome Modeling Engine for data-driven simulation of spatial genome organization. Genome Res. 2016;26(12):1697–709.
33. Tjong H, Gong K, Chen L, Alber F. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome Res. 2012;22:1295–1305.
34. Trussart M, Serra F, Baú D, Junier I, Serrano L, Marti-Renom MA. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. Nucleic Acids Res. 2015;43(7):3465–77.
35. Varoquaux N, Ay F, Noble WS, Vert JP. A statistical approach for inferring the 3D structure of the genome. Bioinformatics. 2014;30(12):i26–33.
36. Wang S, Su J-H, Beliveau BJ, Bintu B, Moffitt JR, Wu C-T, Zhuang X. Spatial organization of chromatin domains and compartments in single chromosomes. Science. 2016;353(6299):598–602.
37. Witten DM, Noble WS. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. Nucleic Acids Res. 2012;40(9):3849–55.
38. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet. 2011;43:1059–65.
39. Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, Yue F, Li Q. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017;27(11):1939–49.
40. Yang T. hicrep: Measuring the reproducibility of Hi-C data. R package version 1.6.0. 2019. https://www.bioconductor.org/packages/release/bioc/html/hicrep.html.

41. Zhang Z, Li G, Toh K-C, Sung W-K. 3D chromosome modeling with semi-definite programming and Hi-C data. J Comp Biol. 2013;20(11): 831–46.
42. Zou C, Zhang Y, Ouyang Z. HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. Genome Biol. 2016;17:40.

Publisher's Note