GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs

Oana Ursu[1], Nathan Boley[1], Maryna Taranova[1], Y. X. Rachel Wang[2], Galip Gurkan Yardimci[3], William Stafford Noble[3,4], and Anshul Kundaje[1,5]

[1]*Department of Genetics, Stanford University School of Medicine, CA, USA*
[2]*Department of Statistics, Stanford University, CA, USA*
[3]*Department of Genome Sciences, University of Washington, WA, USA*
[4]*Department of Computer Science and Engineering, University of Washington, WA, USA*
[5]*Department of Computer Science, Stanford University, CA, USA*

Abstract

The three-dimensional organization of chromatin plays a critical role in gene regulation and disease. High-throughput chromosome conformation capture experiments such as Hi-C are used to obtain genome-wide maps of 3D chromatin contacts. However, robust estimation of data quality and systematic comparison of these contact maps is challenging due to the multi-scale, hierarchical structure of the data and the resulting idiosyncratic properties of experimental noise.

We introduce a multi-scale concordance measure called GenomeDISCO (DIfferences between Smoothed COntact maps) for assessing the similarity of a pair of contact maps obtained from chromosome capture experiments. We denoise the contact maps using random walks on the contact map graph, and integrate concordance at multiple scales of smoothing. We use simulated datasets to benchmark GenomeDISCO's sensitivity to different types of noise typically affecting chromatin contact maps. When applied to a large collection of Hi-C datasets, GenomeDISCO accurately distinguishes biological replicates from samples obtained from different cell types.

Software implementing GenomeDISCO is available at `http://github.com/kundajelab/genomedisco`.

Contact: akundaje@stanford.edu

1 Introduction

The three-dimensional (3D) conformation of chromatin defines a network of physical interactions among genomic loci, including regulatory elements such as gene promoters, distal enhancers and insulators (Krijger and De Laat, 2016). Thus, 3D chromatin architecture plays a key role in gene regulation and cellular function. Changes in 3D chromatin architecture at multiple scales, ranging from large-scale rearrangement of compartments and topologically-associating domains (TADs) to rewiring of enhancer-promoter interactions, are associated with dynamic cellular processes such as differentiation (Dixon *et al.*, 2015; Fraser *et al.*, 2015) and reprogramming (Krijger *et al.*, 2016; Beagan *et al.*, 2016)). Moreover, aberrant disruption of 3D chromatin architecture has been associated with several diseases (Lupiáñez *et al.*, 2015; Gröschel *et al.*, 2014).

The last decade has witnessed a revolution in high-throughput sequencing-based assays and imaging techniques to map 3D chromatin architecture at multiple scales and resolutions, providing new insights into spatial genome organization (Schmitt *et al.*, 2016). The sequencing-based methods (referred to as 3C-seq experiments) for assaying 3D chromatin architecture such as 3C (Dekker *et al.*, 2002), 4C (Zhao *et al.*, 2006; Simonis *et al.*, 2006), 5C (Dostie *et al.*, 2006), Hi-C (Lieberman-Aiden *et al.*, 2009), Capture Hi-C (CHi-C) (Mifsud *et al.*, 2015), ChIA-PET (Fullwood *et al.*, 2009) and HiChIP (Mumbach *et al.*, 2016) are all variations around the chromosome conformation capture technique. In a Hi-C experiment, genome-wide interactions are mapped by ligating proximal fragments followed by deep sequencing. The result of such an experiment is a

genome-wide contact map, which is a matrix with a sequencing readout of the contact frequency for every pair of genomic loci.

A number of computational methods have been designed to normalize (Yaffe and Tanay, 2011; Hu *et al.*, 2012; Imakaev *et al.*, 2012; Knight and Ruiz, 2013; Servant *et al.*, 2015) and extract statistically significant contacts from the different types of 3D chromatin conformation assays (Ay *et al.*, 2014; Ron *et al.*, 2017; Mifsud *et al.*, 2017; Cairns *et al.*, 2016; Carty *et al.*, 2017). However, principled methods for systematic comparisons of 3D contact maps are equally important and form a core component of two key analyses. First, as an essential quality control tool, it is useful to quantify the concordance of replicate experiments. This is particularly relevant because it is common practice to pool reads across biological replicates of a 3C-seq experiment before downstream analyses. Significant differences between the pooled replicates could result in suboptimal or misleading downstream results. Second, understanding and quantifying similarity between replicates is also an essential step in differential analysis, where the goal is to reliably identify statistically significant differences between contact maps in different biological conditions.

Experimentally derived contact maps exhibit certain properties that are distinct from other types of functional genomic data. First, contact maps explicitly encode the adjacency matrix of a multi-scale, modular network consisting of large-scale compartments, TADs, CTCF/cohesin mediated loops and potentially transient interactions between other types of elements (Schmitt *et al.*, 2016). Second, the contact frequency between a pair of loci is strongly dependent on their linear genomic distance (Dekker *et al.*, 2002; Ay *et al.*, 2014; Duan *et al.*, 2010) and affected by additional biases such as restriction fragment size, GC content and mappability (Yaffe and Tanay, 2011; Imakaev *et al.*, 2012; Cournac *et al.*, 2012; Hu *et al.*, 2012; Schmitt *et al.*, 2016). Third, the resolution of a contact map defined in terms of the size (in nucleotides) of the interacting loci is often a free parameter and heuristically determined based on the depth of sequencing (Rao *et al.*, 2014). Finally, the noise associated with estimates of contact frequencies is also strongly associated with sequencing depth. These properties necessitate the development of new computational methods specifically suited for analysis of Hi-C data.

Statistical measures that have been developed to quantify the reproducibility of 1D functional genomics assays, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq), DNA methylation and RNA sequencing, cannot be trivially applied to 3D contact maps. For instance, simple correlation measures, which are most frequently used as measures of reproducibility (Rao *et al.*, 2014), do not correctly capture the reproducibility of Hi-C data (Yang *et al.*, 2017). This is partly because these simple correlation measures consider each entry in a contact map as an independent measurement, thereby ignoring the rich connectivity and dependence structure in 3D contact maps. More sophisticated reproducibility measures have recently been introduced including comparison of eigenvectors (Yan *et al.*, 2016), and a stratified correlation coefficient (Yang *et al.*, 2017), and these methods alleviate many of the problems with traditional correlation.

In this work, we introduce GenomeDISCO (DIfferences between Smoothed COntact maps), a computational framework for quantifying reproducibility or concordance of contact maps from 3C-seq experiments (Figure 1). We represent a contact map as a network or graph, where nodes are genomic loci and edges are weighted proportional to the appropriately normalized contact frequency between a pair of loci (nodes). We denoise the contact maps using random walks on the graph, and perform comparisons at multiple scales of smoothing. We use systematic simulations to calibrate the method, showing its ability to detect artificially introduced noise, differences in distance dependence curves and differences in structural properties of contact maps. We then apply GenomeDISCO and other related approaches to the largest existing collection of Hi-C experiments (Rao *et al.*, 2014), and benchmark their performance on a comparison of replicate experiments and experiments from different cell types. We provide an efficient implementation of our method as well as comprehensive analysis reports and visualizations in the form of a user-friendly software package at `http://github.com/kundajelab/genomedisco`. GenomeDISCO is also included in the 3D genome analysis suite recommended by the Encyclopedia of DNA Elements (ENCODE) Consortium, at `http://github.com/kundajelab/genomedisco/tree/master/reproducibility_analysis`.

2 Methods

2.1 A graph representation of chromatin contact maps

We represent a chromatin contact map as a graph or network of interactions between genomic loci, with a weighted adjacency matrix A. Each node i in the network is a genomic locus (segment) of a specified resolution or size (in nucleotides). The weight of each edge A_{ij} is a suitably normalized, experimentally-derived contact frequency between a pair of nodes i and j. In practice, we ignore inter-chromosomal interactions and hence represent all chromosomes as independent graphs (contact maps). For simplicity of notation, we refer to each of these chromosomal contact maps as A. We aggregate concordance scores across all chromosomes as described in Section 2.3. For Hi-C contact maps, the nodes in the contact map graph of each chromosome (rows and columns of A) span all consecutive non-overlapping segments of the chromosome at a specified resolution.

2.2 Datasets, preprocessing and normalization

Hi-C Data We use Hi-C datasets from seven cell types from (Rao *et al.*, 2014) as summarized in Supplementary Table 1 (GEO accession numbers included in the table). For each cell type, we downloaded reads mapped to the hg19 human genome reference, filtered for mapping quality (MAPQ $>$ 30). We then computed the number of reads supporting contacts between all pairs of genomic bins of a specified resolution to obtain a contact map. In general, GenomeDISCO expects a user-defined resolution, which can be determined empirically, for instance using the definition provided in (Rao *et al.*, 2014), i.e. choosing the lowest resolution such that at least 80% of the genomic bins have at least 1000 contacts with non-zero counts. In this work, we use a 50 kb resolution for our simulations and 40kb for the real Hi-C datasets.

We normalize the raw contact count matrix, C to its normalized version, A using the square root normalization method (sqrtvc from (Rao *et al.*, 2014)) that corrects for node-specific, factorizable biases.

$$A = D^{-\frac{1}{2}} C D^{-\frac{1}{2}}$$

where D is a diagonal matrix, with each entry D_{ii} corresponding to the degree (row sum) of node n_i. Other normalization frameworks such as ICE (Imakaev *et al.*, 2012) or KR (Knight and Ruiz, 2013) are also compatible with our framework and do not change any presented conclusions. We prefer the sqrtvc normalization since KR and ICE occasionally do not converge for Hi-C datasets with moderate sequencing depth processed at very high resolution, such as 5-10kb. In the case of KR and sqrtvc normalization, the normalized matrix A is a valid or close approximation to a transition probability matrix (i.e. the entries in each row sum to 1).

2.3 The GenomeDISCO score for estimating multi-scale concordance of contact maps

A concordance score that aims to estimate the global similarity between a pair of contact maps must account for the specific properties of experimentally-derived contact maps.

First, contact maps contain structural features that manifest at different scales, such as large-scale compartments, sub-Mb scale TADs and sub-TADs that manifest as densely connected diagonal blocks and CTCF/cohesin mediated loops observed as focal points of enriched contacts. Thus, an ideal concordance score would be able to measure similarity across multiple scales. Previous studies have found that using features defined at multiple scales of the contact map lead to a boost in performance when predicting gene co-expression (Babaei *et al.*, 2015).

Second, genome-wide contact maps such as those from Hi-C experiments measure a very large space of possible contacts and hence require deep sequencing ($>$ billion reads) for reliable estimates of contact frequency. Due to cost and material constraints, typical Hi-C datasets are sequenced at significantly lower coverage (e.g. 100M reads (Lajoie *et al.*, 2015)). These undersampled datasets exhibit a large proportion of contacts with low observed counts with high variance (Carty *et al.*, 2017) including some contacts with 0 observed counts, a phenomenon known as stochastic dropout. To address this issue, we propose a denoising approach to smooth contact maps by leveraging random walks on the contact map graph followed by comparisons at multiple scales of smoothing.

The key steps to comparing a pair of chromatin contact maps $A1$ and $A2$ are as follows (Figure 1).

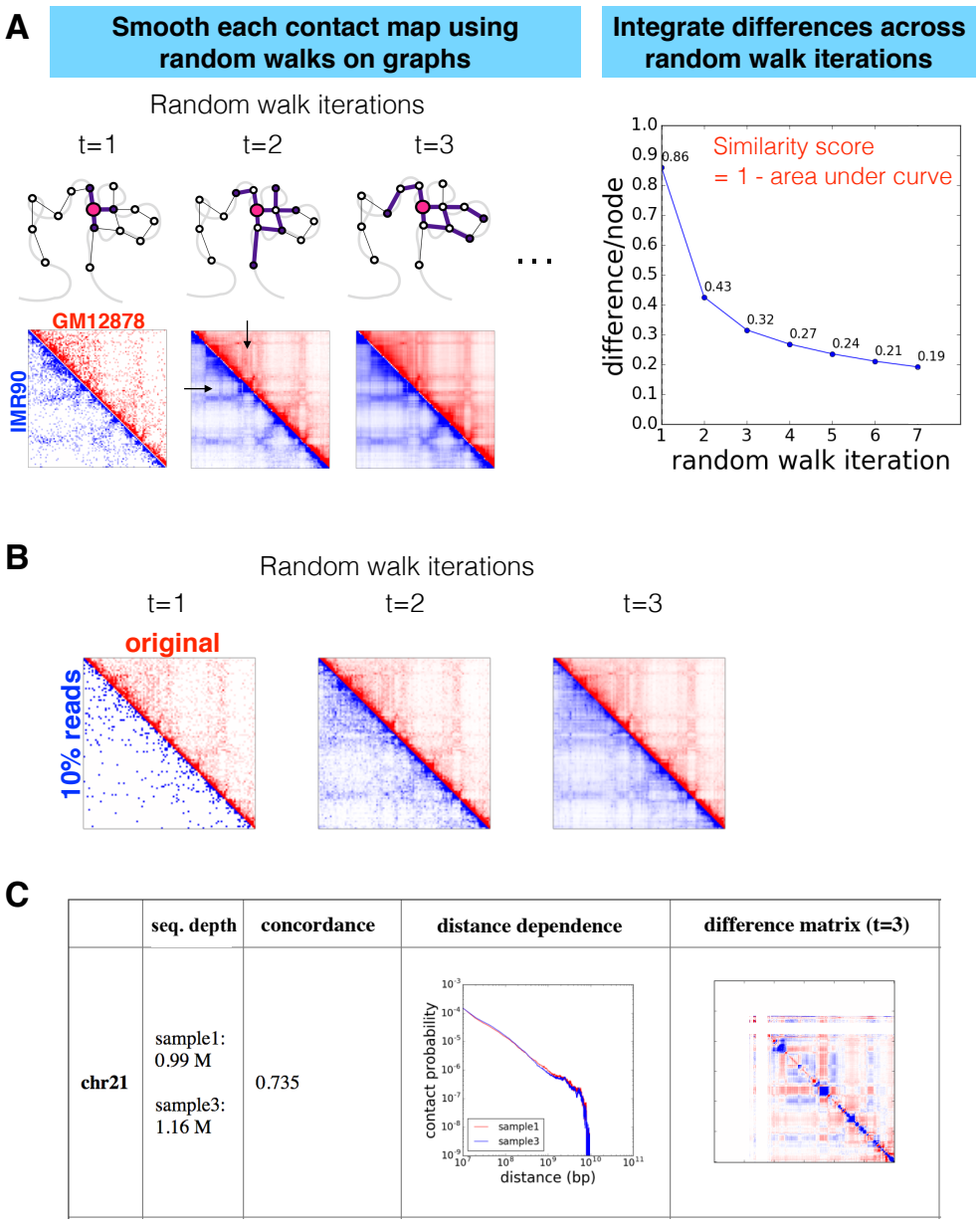

Figure 1: **The GenomeDISCO method for computing concordance using differences in smoothed contact maps**

A) An example comparison between two cell lines, GM12878 (top, red matrix) and IMR90 (bottom, blue matrix). We start from the original matrix (t=1), and perform smoothing of the data using random walks on graphs at multiple steps, t. At each random walk step, we compute the L_1 difference between the contact maps. Finally, we integrate these differences into a concordance score, equal to 1 - normalized area under the curve of the obtained differences.

B) Smoothing process displayed for the IMR90 sample, either at its original sequencing depth (top, red matrix), or when subsampled to 10% of reads (bottom, blue matrix). We find that our smoothing scheme closely recovers the domain and compartment features of the deeply sequenced sample even when starting with a sample with lower sequencing depth.

C) Output from GenomeDISCO, including its diagnostic information including the sequencing depth of the samples being compared, a plot showing the distance dependence curves of the two datasets and the difference matrix between the smoothed matrices (for this example, using t=3).

- **Equalizing sequencing depth** In order to avoid artificial differences due to sequencing depth, we first equalize the sequencing depth of the pair of datasets to be compared by randomly subsampling the count matrix to the minimum depth of the two datasets. This is done by sampling for each entry in the contact map from a binomial distribution with N=count at the entry and p chosen such that p(total sequencing depth) = (desired sequencing depth).

- **Denoising contact maps using random walks** We denoise each contact map separately using random walks on the contact map graph A. In particular, we ask: if we start a random walk at node i, and allow it to take t steps, what is the probability we will reach node j? This probability can be obtained from

$$A^t$$

where A^t_{ij} represents the probability of reaching node j from node i in t steps. We perform random walks of increasing steps from 1 to t_{max}, without any reduction in resolution (bin size) of the nodes. We perform these operations separately for $A1$ and $A2$, obtaining $A1, A1^2, A1^3$ etc., and similarly $A2, A2^2, A2^3$ etc.

- **Computing the difference between smoothed contact maps at each step** For each step t of the random walk, we compute the difference $d_t(A1, A2)$ between $A1^t$ and $A2^t$ as the L_1 distance between the two smoothed contact maps, divided by the average number of non-zero nodes in the two contact maps:

$$d_t(A1, A2) = \frac{\sum_i \sum_j |A1^t_{ij} - A2^t_{ij}|}{|N_{nonzero} = \{n_i | \sum_j n_{ij} > 0\}|}$$

. Since each row of $A1$ and $A2$ sums to ≈ 1, the weighted degree (sum of weights of all edges to/from a node) of each node is ≈ 1. Hence, $d_t(A1, A2)$ scores range from 0 to 2, with small values indicating high similarity.

- **Integrating differences across multiple scales** We compute an integrated multi-scale difference $D(A1, A2)$ between $A1$ and $A2$ as the area under the curve (AUC) obtained by plotting the difference scores $d_t(A1, A2)$ relative to a range of steps $t = [t_{min}, t_{max}]$, divided by $\Delta t = (t_{max} - t_{min})$ i.e. the range of t values used. If $t_{min} = t_{max}$, then $D(A1, A2) = d_{t_{min}}(A1, A2)$. We determined the optimal parameters t_{min} and t_{max} as those that lead to the best separation of biological replicates from datasets corresponding to different cell types or conditions. Based on our benchmarking results on simulated and on real Hi-C datasets from (Rao *et al.*, 2014) (see Section 2.5), we found $t_{min} = t_{max} = 3$ to be optimal for human Hi-C datasets. These parameters may need to be optimized for datasets from other types of assays or other species.

- **Converting the difference to a concordance score** We converted the integrated difference $D(A1, A2)$ into a concordance score $R(A1, A2)$ as follows:

$$R(A1, A2) = 1 - D(A1, A2)$$

$R(A1, A2)$ scores ranges from -1 to 1, with larger values indicating greater similarity.

- **Combining concordance scores across multiple chromosomes** We computed a separate concordance score for each chromosome. To obtain a genome-wide concordance score, we computed the average of the scores across all chromosomes in the genome.

- **Calibrating the concordance score** We calibrated the concordance scores using high quality biological replicates as the gold standard for defining similarity and datasets from different cell types as the gold standard for defining dissimilarity. Specifically, for Hi-C data, we identified concordance score thresholds that best distinguished pairs of simulated replicates from pairs of simulated datasets representing contact maps from different cell types (See Section 2.4. Concordance scores also depend on the baseline resolution of the contact maps (higher resolutions result in an overall decrease in the magnitude of concordance scores). Hence, we provide precomputed calibrated thresholds on concordance scores for a range of resolutions: 10 kb, 40 kb, 500 kb at http://github.com/kundajelab/genomedisco/tree/master/reproducibility_analysis/calibration_tables/GenomeDISCO.calibration_table.txt.

2.4 Simulating different types of noise in contact maps

We first simulated realistic contact maps based on Hi-C datasets from seven cell types from (Rao *et al.*, 2014) (see Supp Table 2) in order to calibrate parameters and evaluate the performance of GenomeDISCO. We pooled reads across all replicates for each cell type. We used a resolution of 50 kb fixed size genomic bins. For the sake of efficiency, we then restricted the contact map to chromosome 21 for our simulations. For each cell type, we rescaled the raw contact count matrix C into a probability matrix P, such that all entries in the upper triangle of P sum to 1 (i.e. a valid probability distribution). Then, to distinguish structural differences from differences in distance dependence curves, we scaled the obtained probabilities to ensure that datasets for all cell types contain identical distance dependence functions. We modified the distance dependence curves to follow one reference distance dependence function. For the Hi-C datasets, we used the GM12878 dataset as the reference since it is the most deeply sequenced cell type. For each genomic distance g (from 0 bp to the length of the chromosome), we divided each entry at genomic distance g by the ratio of the average contact probability at distance g in the target dataset and the average contact probability at distance g in the reference dataset with the desired distance dependence curve. Note that the upper triangle of the resulting scaled matrix is a valid probability distribution, because the upper triangle of the reference matrix is a probability distribution. Finally, we simulated a contact map of a desired read depth N by sampling each entry (i, j) from a binomial distribution, with $p = P_{ij}$. We repeated this process twice per contact map for each simulation configuration (i.e. we sampled twice from the same underlying probability matrix) to generate a pair of "pseudo-replicates", obtaining 14 (7 x 2) simulated contact maps.

We then simulated various types of noise in the contact maps to evaluate the behavior of the GenomeDISCO concordance score.

- **Edge dropout** The edge dropout simulations measured how our concordance score changes as a function of removing edges from contact map graphs. For this simulation, we randomly set x% (for x between 10% and 90%) of the entries in the probability matrix to 0. We then renormalized the upper triangle to a valid probability distribution and then sampled from a binomial distribution as described above. For each level of noise, we computed scores by comparing the original sample (0% dropout) against a sample with x% edge dropout. We estimated the standard deviation of scores for x% edge dropout based on 14 comparisons. Seven of these correspond to concordance scores obtained across the 7 cell types with Hi-C data. Also, for each cell type, we computed 2 scores: one comparison for replicate 1 (0% dropout) vs. replicate 2 (x% dropout), the second comparison for replicate 2 (0% dropout) vs. (replicate 1, x% dropout).

- **Node dropout** The node dropout simulations involve random removal of nodes. As in the edge dropout simulations, we perturbed P. For a given percent of dropout, x, we removed x% of the nodes, which is equivalent to setting all probabilities involving that node to 0. Then we renormalized and sampled reads from the binomial distribution as described above. The comparisons are analogous to the ones described in the edge dropout simulations: we compared the contact maps with 0% dropout against those with x% dropout for a total of 14 comparisons per dropout level.

- **Domain boundary noise** The domain boundary noise simulations were designed to understand how the concordance score changes as a function of uncertainty in the location of domain/TAD boundaries in the contact map. To simulate variation in domain location, we used a reference contact map and shifted it by a specific number of nodes b called the domain boundary noise. The contact frequency for a pair of nodes (i, j) in the shifted contact map is equal to the contact frequency at nodes $(i + b, j + b)$ in the reference contact map. Then, we compared the original matrix with matrices shifted by different values of b (50 kb, 100 kb, 200 kb, 400 kb, 80 kb, 1.6 Mb). We performed this shift using all nodes on chr21, but for scoring, we only used a subset of this chromosome that is contiguous (starting at 20 Mb and ending at 45 Mb). For consistency with the other simulation types described above, this subset was used in all simulations in this paper. As in the previous simulations, we obtained 14 comparisons per shift.

- **Different distance dependence curves** For each contact map, we used two different distance dependence curves, obtained from Hi-C datasets from two different cell lines HMEC and HUVEC, which had the largest difference in distance dependence curves (Rao *et al.*, 2014). To transform a contact map to obey a desired distance dependence function, we used the probability matrix P representation of the contact map (as described above) and rescaled values at each genomic distance such that the average contact probability at that distance corresponds to the average contact frequency of the desired dataset. Finally, given the rescaled probability matrix, we sampled from the binomial distribution to obtain the simulated datasets. For each of the two distance dependence curves of interest, we obtained 2 pseudoreplicates by sampling from the binomial distribution twice. Thus, we obtained 7 cell types x 2 comparisons with shared distance curves: the first comparison involving (replicate 1, distance dependence 1) vs. (replicate 2, distance dependence 1) and the second comparison involving (replicate 1, distance dependence 2) vs. (replicate 2, distance dependence 2). The comparisons with different distance curves were restricted to be in the same cell type i.e. (replicate 1, distance dependence 1) vs (replicate 1, distance dependence 2) and so on.

  To measure the separation between the two groups of comparisons (pairs of samples with the same distance dependence curves and those with different curves), we used the Silhouette score (Rousseeuw, 1987). This entails computing for every comparison i the following quantity: $s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$, where b(i)=lowest difference between the score for comparison i and a score from the opposite group and a(i)=the average difference in score between i and the other comparisons in the group i is a part of. The silhouette score is the average s(i) across all datapoints.

- **Comparing simulated pseudoreplicates within a cell type to simulated maps between different cell types** For this simulation, we use the datasets created in the "Edge noise" simulation, with 0% noise. Since each cell type has 2 simulated pseudoreplicates, we can evaluate concordance of pairs of pseudoreplicates from the same cell type and compare it to concordance of pairs of simulated contact maps from different cell types. We measured the separation between pseudoreplicates and different cell types using the silhouette score as described above.

The simulator code is included in the GenomeDISCO package (`http://github.com/kundajelab/genomedisco/tree/master/genomediscosimulations_from_real_data.py`).

For the results on the simulations, we used $t = 3$, as this value was deemed optimal in our parameter optimization (see Section 2.5).

2.5 Parameter optimization on the real data

We used the following procedure to identify the optimal random walk step parameters $t = [t_{min}, t_{max}]$ on real Hi-C datasets from (Rao *et al.*, 2014), which contains more than 80 experiments across multiple human cell lines. We used half the experiments as a training set and the remaining half as a test set. We then computed GenomeDISCO concordance scores for all pairs of datasets in the training set, for random walks with different t. We considered all combinations of t_{min} and t_{max} within the range of $1 \leq t \leq 5$. We optimized parameters based on their ability to classify pairs of biological replicates from pairs of non-replicates in the training set. We used auPRC (see Figure 4A) to evaluate classification performance. We then used the test set to compare GenomeDISCO with the methods described below.

2.6 Comparison with other methods

We compared our method with two other recently developed concordance scoring methods for Hi-C data; HiCRep (Yang *et al.*, 2017) and HiC-Spector (Yan *et al.*, 2016). For HiCRep we used a maximum distance of contacts equal to 5 Mb and a smoothing parameter h=5, which is what was suggested for 40kb resolution Hi-C data. For HiC-Spector we used 20 eigenvectors. Another commonly used concordance score is correlation (Spearman or Pearson), but (Yang *et al.*, 2017) have already pointed out its deficiencies. Hence, we do not include comparisons to naive correlation measures.

2.7 Analysis of differences in distance dependence curves

To obtain a quantitative measurement of whether two contact maps have different functions that map contact probability as a function of genomic distance, we computed the sum of absolute values of the difference between contact probability at each genomic distance. The values obtained in this manner and used in Figure 4 are based on chr18.

3 Results

3.1 Benchmarking GenomeDISCO on simulated perturbations to 3C-seq datasets

We expect a sound concordance score for 3C-seq datasets to be sensitive to key types of noise and artifacts that typically affect these data (Figure 2B).

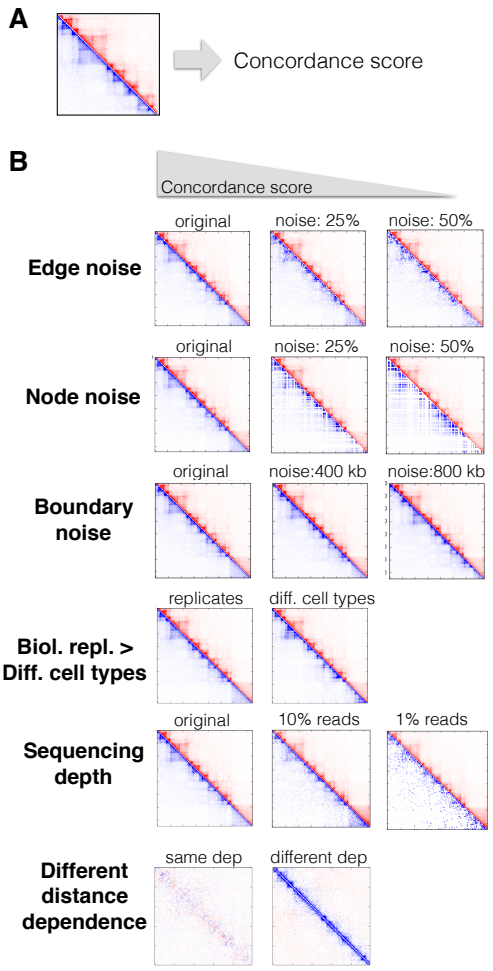

Figure 2: **Definition of concordance for chromosome conformation capture data**
A) A concordance score measures the global similarity between a pair of contact maps. Upper triangle of the matrix (red) is from the first contact map and lower triangle (blue) of the matrix is from the second contact map.
B) Desired features for a concordance score. The score should decrease as we add edge noise, node noise and boundary noise. The score should be higher for replicates from the same cell type than for comparisons between different cell types. The concordance score should drop when sequencing depth is lowered. Scores should be lower for pairs of replicates with different distance dependence curves (the lower panel is showing the difference between the matrices involved in the distance dependence comparisons).

We benchmarked the behavior of GenomeDISCO by using it to compute concordance between a reference Hi-C contact map and a version of the map that is explicitly perturbed with different types and levels of of simulated noise (see Figure 2B) (See Methods). We further compared GenomeDISCO to two other recently developed methods for estimating concordance of HiC data: HiCRep, which measures correlation of contacts stratified by distance (Yang *et al.*, 2017) and HiC-Spector, which computes an eigendecomposition of the Laplacian of the graph, and then compares the L_2 distance between eigenvectors of the 2 contact maps Yan *et al.* (2016).

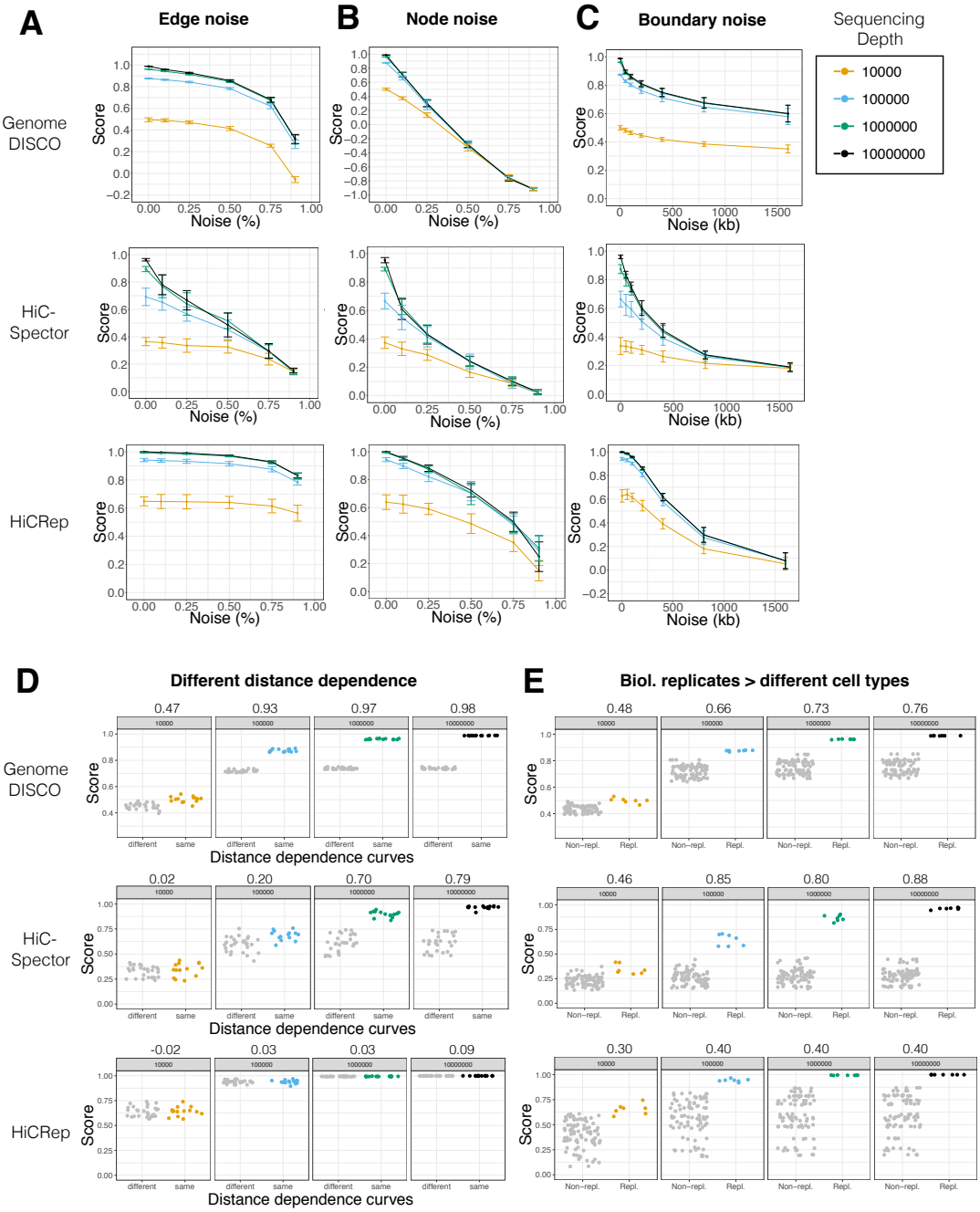

Figure 3: **GenomeDISCO exhibits desired features for a reproducibility score** A-E) Results from simulations, consistent with the expectations in Figure 2B. Error bars represent one standard deviation from the mean score, based on independent simulations across 7 cell types. For D), E), the values above the plots are silhouette scores measuring the separation between the two groups of scores being compared.

We examined the sensitivity of the concordance scores to perturbations that involve random

dropout of edges and nodes as well as misalignment of domain boundaries in the perturbed contact map relative to the reference. Indeed, we found that concordance scores from all three methods decrease with increasing edge drop out (Figure 3A), increasing node drop out (Figure 3B) and increasing domain boundary misalignment (Figure 3C, see Methods).

Next, in order to understand the effect of sequencing depth of the contact maps, we repeated the above three perturbation analyses for reference and perturbed maps subsampled to four depths: 100%, 10%, 1%, 0.1% of 10 million reads restricted to chromosome 21. As expected, we found that the GenomeDISCO score is highest for the deepest samples, and drops consistently with decreasing sequencing depth across all types and levels of perturbations (Figure 3). In addition, the sensitivity to increasing levels of each of the perturbations decreases with lowered sequencing depth. Also, the scores plateau as the sequencing depth is increased from 1 million to 10 million reads, which is expected, since for a 40kb resolution, one would need approx. 1 million reads for chr21.

Contact maps can also differ in their fundamental distance dependence curves that capture the probability of contact as a function of linear genomic distance. Distance dependence functions can differ due to the stage of the cell cycle of cells (Naumova *et al.*, 2013; Nagano *et al.*, 2016) or as a function of perturbation of proteins involved in chromatin 3D architecture, such as RAD21 knockout in yeast (Mizuguchi *et al.*, 2014) or WAPL and SCC4 knockouts in human HAP1 cells (Haarhuis *et al.*, 2017). Replicates from the same condition are often pooled, and if they have different distance dependence curves, the result will be an average that is not representative of either replicate. Hence, being sensitive to differences in distance dependence curves is a useful property of a concordance score. We simulated pairs of contact maps from a common reference contact map by sampling reads according to two different distance dependence curves, obtained from HiC maps from two different cell types (see Methods). We also simulated pairs of contact maps with the same distance distributions. We then compared the concordance of the pairs of simulated maps with different distance curves and to concordance of pairs with the same distance dependence curves at different sequencing depths (as above) using all three methods. GenomeDISCO correctly identified pairs of samples with the same distance dependence curves as more concordant than pairs of samples with different distance dependence curves. As in the other simulations, the margin between the two sets of pairs decreased as we decreased sequencing depth (Figure 3D). Only GenomeDISCO and HiC-Spector are sensitive to differences in distance dependence curves, with GenomeDISCO having better margins of separation at lower sequencing depths as compared to HiC-Spector (GenomeDISCO silhouette scores of 0.47, 0.93, 0.97, 0.98 for 10 million, 1 million, 0.1 million and 10000 reads respectively, and HiC-Spector silhouette scores of 0.02, 0.20, 0.70, 0.79).

Finally, we asked whether pairs of simulated pseudo-replicates sampled from the same reference HiC map are deemed more concordant than pairs of samples from HiC reference maps from different cell types. All three methods were able to successfully discriminate the two sets of pairs with margins decreasing with decreasing sequencing depth (Figure 3E).

3.2 Calibrating and benchmarking GenomeDISCO on real Hi-C datasets

We used $>$ 80 high quality Hi-C datasets from (Rao *et al.*, 2014) spanning multiple human cell-lines (GM12878, HMEC, HUVEC, IMR90, K562, KBM7, NHEK) to benchmark the behavior of our concordance score (Figure 4). Due to the lack of explicit ground truth about the nature of noise in real datasets, we evaluate the validity of the concordance score by expecting higher scores when comparing pairs of biological replicates of Hi-C data with similar distance-dependence characteristics as compared to scores obtained by comparing Hi-C datasets from different cell types. We focused our analysis on a subset of experiments from (Rao *et al.*, 2014) defined as those done with in-situ Hi-C (see Supp Table 1 for a list of the datasets used).

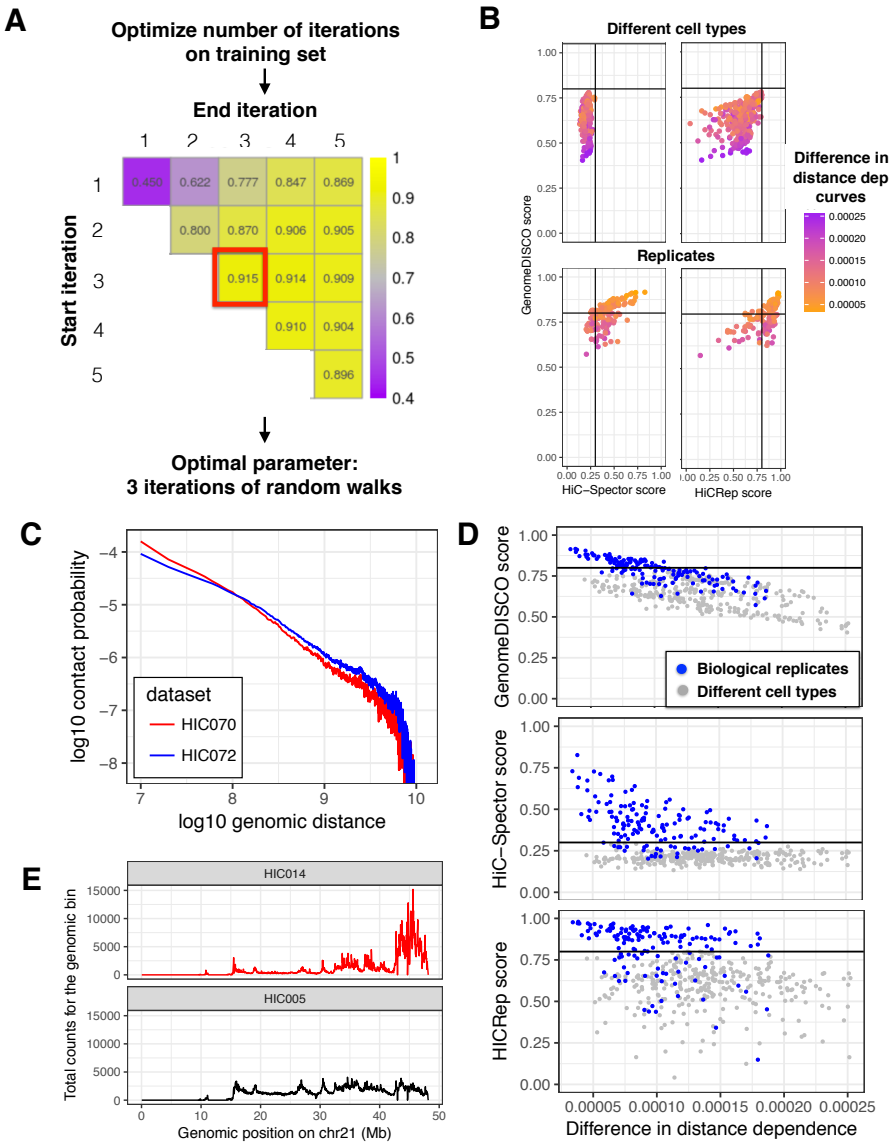

Figure 4: **GenomeDISCO distinguishes biological replicates from nonreplicates, taking distance dependence curves into account**

A) Parameter optimization for GenomeDISCO, by maximizing auPRC for distinguishing biological replicates from non-replicates in a training set of samples. We find that three iterations of random walk works best.

B) Scatterplot of scores obtained with GenomeDISCO vs those obtained with HiCRep and HiC-Spector. For each of the thre methods, we define a threshold that separates low-concordance from high-concordance pairs of samples. The threshold is chosen as the highest score obtained by a comparison between different cell types. GenomeDISCO largely agrees with the other methods. There is a subset of scores that GenomeDISCO selectively ranks as low-concordance and those involve pairs of contact maps with large differences between their distance dependence curves.

C) An example of contact maps with different distance dependence functions that GenomeDISCO deems non-concordant while HiCRep defines as concordant.

D) Concordance scores as a function of difference in distance dependence functions. The difference is measured as the sum of absolute values of differences between the distance curves at each genomic distance.

E) One example dataset whose comparisons consistently receive low concordance scores is experiment HIC014. Upon closer inspection, we found that row sums for each genomic bin for sample HIC014 are non-uniform, compared to e.g. HIC005, at a similar sequencing depth of 300 million reads.

First, we optimized the parameters $t = [t_{min}, t_{max}]$ (minimum and maximum number of random walk steps to integrate over) for GenomeDISCO (See Methods) as follows: we split the datasets from (Rao *et al.*, 2014) into a training and a test set. Then we used the training set to try a set of combinations of $t = [t_{min}, t_{max}]$ and identified the optimal set of as those that maximize auPRC of distinguishing biological replicates from non-replicates on the training set (Figure 4A). We obtained the lowest auPRC when ($t_{min} = t_{max} = 1$) and also obtained low auPRCs whenever t_{min} was set to 1, indicating that smoothing the contact map is critically important to obtain optimal performance. We obtained optimal auPRC for $t_{min} = t_{max} = 3$.

Next, we used GenomeDISCO, HiCRep and HiC-Spector to compute concordance scores for all the pairs of biological replicates and pairs of samples from different cell types. For each method we defined an empirical threshold for classifying sample-pairs into one of two categories labeled high-concordance and low-concordance. The threshold was determined as the highest score across all pairs of samples from different cell types, since we expect concordant biological replicates to be at least as concordant as samples from different cell types. We then analyzed the similarities and differences between the three methods in terms of their classification of the pairs of biological replicates. (Figure 4B).

Out of 149 pairs of biological replicates in the test set, we found that the methods agreed for a majority of cases (104/149 biological replicate pairs were classified consistently between GenomeDISCO and HiCRep, and 93/149 between GenomeDISCO and HiC-Spector). For a small subset of replicate-pairs, HiCRep and/or HiC-Spector classified them as high-concordance, while GenomeDISCO classified them as low concordance. For 36/45 of the discrepancies between GenomeDISCO and HiCRep and 44/56 of the discrepancies with HiC-Spector, the comparisons involved samples with large differences in distance dependence curves (difference in distance dependence curve higher than 0.0001, a value which was found to distinguish pairs of biological replicates in the high-concordance class from those in the low concordance class). For example, samples HIC070 and HIC072 (biological replicates for the K562 cell line) are classified as low-concordance by GenomeDISCO (score 0.641), but classified as high-concordance by HiCRep (score 0.908). These samples have a marked difference in their distance dependence curves (ranked as the 6th largest difference in distance dependence curve among all biological replicate pairs) (Figure 4C). In fact, GenomeDISCO scores in general drop proportional to the difference in distance dependence curves between the pair of samples being compared (Figure 4D). Consistent with our simulation results, HiC-Spector scores also drop with increasing differences in distance dependence curves for pairs of replicates (Figure 4D). But a larger proportion of these replicates fall in the high-concordance class for HiC-Spector as compared to GenomeDISCO.

We also found a subset of replicates that were deemed low-concordance by all three methods (39 for GenomeDISCO vs HiCRep and 26 for GenomeDISCO vs HiC-Spector). For example, three replicate pairs classified as low-concordance by all three methods despite being very deeply sequenced (>300 million reads) involved sample HIC014 from the GM12878 cell type (specifically HIC014 vs HIC020, HIC014 vs HIC022 and HIC014 vs HIC026). Upon closer inspection, we found that HIC014 exhibited an unusual pattern of uneven coverage across the genome (Figure 4E), likely explaining the observed results. In fact, 13/39 pairs of non-concordant pairs of samples between GenomeDISCO and HiCRep (and 12/26 against HiC-Spector) involve comparisons against sample HIC014.

4 Discussion

Here, we present GenomeDISCO, a new approach specifically designed for evaluating concordance and reproducibility of chromatin contact maps obtained from chromosome conformation capture experiments. Our benchmarking experiments on simulated contact maps and high quality Hi-C datasets, which include systematic comparisons to two other methods HiCRep and HiC-Spector, indicate that GenomeDISCO displays competitive accuracy in distinguishing biological replicates from different cell types with the desired sensitivity to sequencing depth, node and edge dropout noise, changes in domain boundaries and subtle differences in distance dependence.

GenomeDISCO introduces a novel approach of using random walks on the contact map graph for progressive smoothing and evaluation of concordance at multiple scales. A weighted graph is a natural representation of a chromatin contact map. A random walk on a contact map graph progressively upweights direct edges involving node pairs that have many high-weight indirect paths of progressively increasing lengths that connect the node pairs. In contrast to other smoothing

approaches, the random walk smoothing algorithm has the advantage of adapting seamlessly to different structural properties and constraints encoded in contact maps from different experimental protocols. For instance, the HiCRep method (Yang *et al.*, 2017), which is specifically designed for comparing Hi-C contact maps, uses a fixed size local 2D smoothing window around each entry in the contact map. This approach is well-suited to genome-wide contact maps such as Hi-C where adjacent neighbors of a node in the contact map are contiguous in linear genomic space. However, it is less appropriate for targeted maps such as those obtained from CHi-C, ChIA-PET or HiChIP where the targeted nodes are not contiguous in linear genomic space and neighbors' of a node in the contact map may be distant loci. GenomeDISCO can be applied as-is to asymmetric or symmetric contact maps from targeted assays such as CHi-C, ChIA-PET or HiChIP.

Another advantage of the random walk is that it is ideally suited to progressively highlight structural properties of contact maps at increasing scales. Short random walks naturally reinforce local network structure, such as loop cliques, whereas longer random walks can highlight global structures such TADs and compartments. GenomeDISCO can compare contact maps at multiple scales and integrate the similarities into a unified score.

Further, GenomeDISCO is sensitive to subtle differences in distance dependence curves. Since it is common to pool multiple Hi-C replicates, it is essential to know if samples exhibit differences, so as to not eliminate signal during pooling. For instance, pooling two samples with very different distance dependence curves may lead to a pooled sample that is difficult to interpret. In addition, since in some cases variation in distance dependence curves is biologically meaningful (for instance, as when distance dependence varies as a function of cell cycle (Naumova *et al.*, 2013; Nagano *et al.*, 2016), or when proteins governing genome 3D structure are knocked out (Mizuguchi *et al.*, 2014; Haarhuis *et al.*, 2017)), it is essential to be able to measure these changes in the distance dependence. On the other hand, two datasets can have different distance dependence curves but still be concordant in terms of enrichments of contacts when accounting for the different distance dependence function of each dataset. Thus, if one is interested in evaluating concordance of contact enrichment (e.g. as measured by methods that call significant contacts), then one can normalize the observed contact frequencies by the expected distance-dependent contact frequencies (which would correct for most differences in distance dependence) for the pair of contact maps before feeding them into GenomeDISCO. One can obtain these observed/expected ratios or associated q-values from Fit-Hi-C (Ay *et al.*, 2014). Another approach is to normalize the contact maps such that their distance dependence curves are matched, with methods such as HiCDiff (Stansfield and Dozmorov, 2017). Alternatively, since HiCRep is not as sensitive to differences in distance dependence curves, one can cross-check results against it to learn if the main reason for low reproducibility from GenomeDISCO is a difference in distance dependence. One quick look at GenomeDISCO's diagnostic plots is usually sufficient to understand if this is the case.

Further, GenomeDISCO provides a variety of diagnostic analyses which are useful to dig deeper in the potential reasons for low concordance. The diagnostic analyses include the comparison of distance dependence curves, sequencing depth (since low sequencing depth leads to lower reproducibility) and a difference matrix between smoothed contact maps.

Finally, what determines a good threshold for concordance of biological replicates? Based on our extensive analyses of simulated datasets and extensive collections of Hi-C data, we define an empirical GenomeDISCO score threshold of 0.8 at 40kb resolution. We also provide a set of precomputed standards based on pseudoreplicates for frequently used resolutions, allowing a direct calibration of a given score to an upper bound. We recommend subsampling datasets to equal sequencing depth before computing reproducibility, since sequencing depth directly affects the reproducibility score. GenomeDISCO performs this subsampling by default.

Three-dimensional chromatin architecture is the next frontier in deciphering genome function. Rapid innovations and improvements of experimental protocols based on chromosome conformation capture are providing us a powerful collection of tools to directly interrogate 3D chromatin architecture. Ensuring high quality reproducible experiments is an essential part of this revolution in understanding chromatin architecture. GenomeDISCO is a user-friendly, efficient and accurate diagnostic tool to evaluate the reproducibility of 3D chromatin conformation capture experiments.

Acknowledgements

We thank Michael Snyder, Jonathan Pritchard, Howard Chang, Michael Bassik and the Kundaje lab for helpful discussions. We thank Suhas Rao for clarifications related to the Hi-C datasets

used. We thank Anna Shcherbina and Chris Probert for help with visualization.

Funding

O.U. is supported by a Howard Hughes Medical Institute International Student Research Fellowship and a Gabilan Stanford Graduate Fellowship. A.K. is supported by NIH grants 1DP2GM12348501, 3U41HG007000-04S1, 3R01ES02500902S1 and 1U01HG009431-01.

References

Ay, F., Bailey, T. L., and Noble, W. S. (2014). Statistical confidence estimation for hi-c data reveals regulatory chromatin contacts. *Genome research*, **24**(6), 999–1011.

Babaei, S., Mahfouz, A., Hulsman, M., Lelieveldt, B. P., de Ridder, J., and Reinders, M. (2015). Hi-c chromatin interaction networks predict co-expression in the mouse cortex. *PLoS Comput Biol*, **11**(5), e1004221.

Beagan, J. A., Gilgenast, T. G., Kim, J., Plona, Z., Norton, H. K., Hu, G., Hsu, S. C., Shields, E. J., Lyu, X., Apostolou, E., *et al.* (2016). Local genome topology can exhibit an incompletely rewired 3d-folding state during somatic cell reprogramming. *Cell Stem Cell*, **18**(5), 611–624.

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.-M., Osborne, C., *et al.* (2016). Chicago: robust detection of dna looping interactions in capture hi-c data. *Genome biology*, **17**(1), 127.

Carty, M., Zamparo, L., Sahin, M., González, A., Pelossof, R., Elemento, O., and Leslie, C. S. (2017). An integrated model for detecting significant chromatin interactions from high-resolution hi-c data. *Nature Communications*, **8**.

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., and Mozziconacci, J. (2012). Normalization of a chromosomal contact map. *BMC genomics*, **13**(1), 436.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, **295**(5558), 1306–1311.

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., *et al.* (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**(7539), 331.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., *et al.* (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, **16**(10), 1299–1309.

Duan, Z., Andronescu, M., Schutz, K., Mcllwain, 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**(7296), 363.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., *et al.* (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Molecular systems biology*, **11**(12), 852.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Han, X., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., *et al.* (2009). An oestrogen receptor α-bound human chromatin interactome. *Nature*, **462**(7269), 58.

Gröschel, S., Sanders, M. A., Hoogenboezem, R., de Wit, E., Bouwman, B. A., Erpelinck, C., van der Velden, V. H., Havermans, M., Avellino, R., van Lom, K., *et al.* (2014). A single oncogenic enhancer rearrangement causes concomitant evi1 and gata2 deregulation in leukemia. *Cell*, **157**(2), 369–381.

Haarhuis, J. H., van der Weide, R. H., Blomen, V. A., Yáñez-Cuna, J. O., Amendola, M., van Ruiten, M. S., Krijger, P. H., Teunissen, H., Medema, R. H., van Steensel, B., *et al.* (2017). The cohesin release factor wapl restricts chromatin loop extension. *Cell*, **169**(4), 693–707.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, **28**(23), 3131–3133.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, **9**(10), 999–1003.

Knight, P. A. and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*, **33**(3), 1029–1047.

Krijger, P. H. L. and De Laat, W. (2016). Regulation of disease-associated gene expression in the 3d genome. *Nature reviews molecular cell biology*, **17**(12), 771–782.

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

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The hitchhiker's guide to hi-c analysis: practical guidelines. *Methods*, **72**, 65–75.

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., *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, **326**(5950), 289–293.

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., *et al.* (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**(5), 1012–1025.

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., *et al.* (2015). Mapping long-range promoter contacts in human cells with high-resolution capture hi-c. *Nature genetics*, **47**(6), 598–606.

Mifsud, B., Martincorena, I., Darbo, E., Sugar, R., Schoenfelder, S., Fraser, P., and Luscombe, N. M. (2017). Gothic, a probabilistic model to resolve complex biases and to identify real interactions in hi-c data. *PloS one*, **12**(4), e0174744.

Mizuguchi, T., Fudenberg, G., Mehta, S., Belton, J.-M., Taneja, N., Folco, H. D., FitzGerald, P., Dekker, J., Mirny, L., Barrowman, J., *et al.* (2014). Cohesin-dependent globules and heterochromatin shape 3d genome architecture in s. pombe. *Nature*, **516**(7531), 432.

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., and Chang, H. Y. (2016). Hichip: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods*, **13**(11), 919.

Nagano, T., Lubling, Y., Varnai, C., Dudley, C., Leung, W., Baran, Y., Cohen, N. M., Wingett, S., Fraser, P., and Tanay, A. (2016). Cell cycle dynamics of chromosomal organisation at single-cell resolution. *bioRxiv*, page 094466.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., and Dekker, J. (2013). Organization of the mitotic chromosome. *Science*, **342**(6161), 948–953.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., *et al.* (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**(7), 1665–1680.

Ron, G., Moran, D., and Kaplan, T. (2017). Promoter-enhancer interactions identified from hi-c data using probabilistic models and hierarchical topological domains. *bioRxiv*, page 101220.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65.

Schmitt, A. D., Hu, M., and Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, **17**(12), 743–755.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, **16**(1), 259.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nature genetics*, **38**(11), 1348.

Stansfield, J. and Dozmorov, M. G. (2017). Hicdiff: A method for joint normalization of hi-c datasets and differential chromatin interaction detection. *bioRxiv*, page 147850.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, **43**(11), 1059–1065.

Yan, K.-K., Yardımcı, G. G., Noble, W. S., and Gerstein, M. (2016). Hic-spector: A matrix library for spectral and reproducibility analysis of hi-c contact maps. *bioRxiv*.

Yang, T., Zhang, F., Yardimci, G. G., Hardison, R. C., Noble, W. S., Yue, F., and Li, Q. (2017). Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *bioRxiv*.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., *et al.* (2006). Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics*, **38**(11), 1341.