APPLICATIONS NOTE

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx

Genomics

HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps

Koon-Kiu Yan[1,2*], Galip Gürkan Yardımcı[4], Chengfei Yan[1,2], William S. Noble[4,5] and Mark Gerstein[1,2,3*]

[1]Program in Computational Biology and Bioinformatics, [2]Department of Molecular Biophysics and Biochemistry, [3]Department of Computer Science, Yale University, [4]Department of Genome Sciences, [5]Department of Computer Science and Engineering, University of Washington, Seattle.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Genome-wide proximity ligation based assays like Hi-C have opened a window to the 3D organization of the genome. In so doing, they present data structures that are different from conventional 1D signal tracks. To exploit the 2D nature of Hi-C contact maps, matrix techniques like spectral analysis are particularly useful. Here, we present HiC-spector, a collection of matrix-related functions for analyzing Hi-C contact maps. In particular, we introduce a novel reproducibility metric for quantifying the similarity between contact maps based on spectral decomposition. The metric successfully separates contact maps mapped from Hi-C data coming from biological replicates, pseudo-replicates and different cell types.

Availability: Source code in Julia and the documentation of HiC-spector can be freely obtained at https://github.com/gersteinlab/HiC_spector

Contact: pi@gersteinlab.org

1 Introduction

Genome-wide proximity ligation assays such as Hi-C have emerged as powerful techniques to understand the 3D organization of the genome (Lieberman-Aiden et al., 2009; Kalhor et al., 2011). While these techniques offer new biological insights, they demand different data structures and present new computational questions (Dekker et al., 2013; Ay and Noble, 2015). For instance, a fundamental question of particular practical importance is, how can we quantify the similarity between two Hi-C data sets? In particular, given two experimental replicates, how can we determine if the experiments are reproducible?

Data from Hi-C experiments are usually summarized by so-called chromosomal contact maps. By binning the genome into equally sized bins, a contact map is a matrix whose elements store the population-averaged co-location frequencies between pairs of loci. Therefore, mathematical tools like spectral analysis can be extremely useful in understanding these chromosomal contact maps. The aim of this project is to provide a set of basic analysis tools for handling Hi-C contact maps. In particular, we introduce a simple but novel metric to quantify the reproducibility of the maps using spectral decomposition.

2 Algorithms

We represent a chromosomal contact map by a symmetric and non-negative adjacency matrix W. The matrix elements represent the frequencies of contact between genomic loci. Recent single-cell imaging experiment suggests that the frequency serves as a good proxy of spatial distance (Wang et al., 2016). In principle, the larger the value of W_{ij}, the closer is the distance between loci i and j. The starting point of spectral analysis is the Laplacian matrix L, which is defined as $L = D - W$. Here D is a diagonal matrix in which $D_{ii} = \sum_j W_{ij}$ (the coverage of bin i in the context of Hi-C). As in many other applications, the Laplacian matrix further takes a normalized form $\mathcal{L} = D^{-1/2}LD^{-1/2}$ (Chung, 1997). It can be verified that 0 is an eigenvalue of $\mathcal{L}$, and the set of eigenvalues of $\mathcal{L}$ ($0 \leq \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$) is referred to as the spectrum of $\mathcal{L}$.

Given two contact maps W^A and W^B, we propose to quantify their similarity by decomposing their corresponding Laplacian matrices $\mathcal{L}^A$ and $\mathcal{L}^B$ respectively and then comparing their eigenvectors. Let $\{\lambda_0^A, \lambda_1^A, \ldots, \lambda_{n-1}^A\}$ and $\{\lambda_0^B, \lambda_1^B, \ldots, \lambda_{n-1}^B\}$ be the spectra of $\mathcal{L}^A$ and $\mathcal{L}^B$, and $\{v_0^A, v_1^A, \ldots, v_{n-1}^A\}$ and $\{v_0^B, v_1^B, \ldots, v_{n-1}^B\}$ be their sets of normalized eigenvectors. A distance metric S_d is defined as

$$S_d(A,B) = \sum_{i=0}^{r-1} \left\| v_i^A - v_i^B \right\| \quad . \qquad (1)$$

Here $\|.\|$ represents the Euclidean distance between the two vectors. The parameter r is the number of leading eigenvectors

Deleted: and therefore serve as a

Deleted: biologically

picked from $\mathcal{L}^A$ and $\mathcal{L}^B$. In general, S_d provides a metric to gauge the similarity between two contact maps. v_i^A and v_i^B are more correlated if A and B are two biological replicates as compared to the case when they are two different cell lines (see Figure S1).

For the choice of r, like any principal component analysis, we expect the leading eigenvectors to be more important than the lower ranked eigenvectors. In fact, we observe that the Euclidean distance between a pair of high-order eigenvectors is the same as the distance between a pair of unit vectors whose components are randomly sampled from a standard normal distribution (see Figure S2). In other words, the high-order eigenvectors are essentially noise terms, whereas the signal is stored in the leading vectors. As a rule of thumb, we found the choice $r = 20$ is good enough for practical purposes. Furthermore, as the distance between a pair of randomly sampled unit vectors presents a reference, we linearly rescale the distance metric into a reproducibility score Q ranges from 0 to 1 (see the Supplement).

We used HiC-spector to calculate the reproducibility scores for more than a hundred pairs of Hi-C contact maps. As shown in Figure 1, the reproducibility scores between pseudo-replicates are greater than the scores for real biological replicates, which are greater than the scores between maps from different cell lines (see the Supplement for details). It is worthwhile to point out that two contact maps can be compared in terms of structures like TADs and loops. What we refer to as "reproducibility" is a direct comparison of the contact maps. The comparison of the high order structures usually strongly on the choices of methods and parameters.

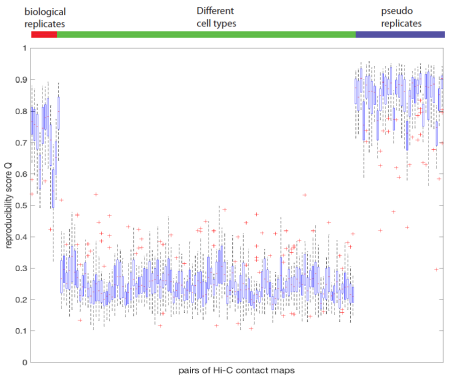

Figure 1 Reproducibility scores for 3 sets of Hi-C contact maps pairs. Contact maps came from Hi-C experiments performed in 11 cancer cell lines. Biological replicates refer to a pair of replicates of the same experiment. Pseudo replicates are obtained by pooling the reads from two replicates together and performing down sampling. There are 11 biological replicates, 33 pairs of pseudo replicates, and 110 pairs of maps between different cell types. The boxplot shows the distribution of Q in 23 chromosomes, with red crosses as the outliers.

Mathematically there are different ways to compare two matrices. For instance, one could assume all matrix elements are independent and define a distance metric using Spearman correlation. The intuition behind S_d is essentially a better way to decompose a contact map. The normalized Laplacian matrix is closely related to a random-walk-process taking place in the underlying graph of W. The leading eigenvector refers to the steady state distribution; the next few eigenvectors correspond to the slower decay modes of the random walk process and capture the densely interacting domains that are highly significant in contact maps. Like typical dimensionality reduction, keeping the first few eigenvectors separates signal from noise. In fact, HiC-spector can better separate biological replicates and non-replicates compared to the correlation coefficient (see Figure S3 and the Supplement).

Apart from the reproducibility score, HiC-spector provides a number of matrix algorithms useful for analyzing contact maps. For instance, to perform a widely used normalization procedure for contact maps (Imakaev et al., 2012), we include the Knight-Ruiz algorithm (Knight and Ruiz, 2012), which is a newer and faster algorithm for matrix balancing. Also, we have included the functions for estimating the average contact frequency with respect to the genomic distance, as well as identifying the so-called A/B compartments (Lieberman-Aiden et al., 2009) using the corresponding correlation matrix.

3 Implementation and Benchmark

HiC-spector is a library written in Julia, a high-performance language for technical computing. A python script for the reproducibility score is also provided for command line calculation. The bottleneck for evaluating Q is matrix diagonalization. The runtime is very efficient but depends on the size of contact maps (see Figure S5 for details).

4 Materials and methods

Hi-C data are generated by the ENCODE consortium (see Supplementary Information). Contact maps in this study were generated using the tool cworld (https://github.com/dekkerlab/cworld-dekker).

Acknowledgements

We want to thank the 3D Nucleome subgroup in the ENCODE consortium for processing the Hi-C data and discussion.

Funding: This work has been supported by NIH award U41 HG007000.
Conflict of Interest: The authors declare no conflict of interest.

References

Ay, F., and Noble, W.S. (2015). Analysis methods for studying the 3D architecture of the genome. Genome Biol. *16*, 183.

Chung, F. (1997). Spectral graph theory (American Mathematical Society).

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

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. Nat. Methods *9*, 999–1003.

Deleted: simple-minded
Deleted: are
Formatted: Font:9 pt
Formatted: Font:9 pt

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Knight, P.A., and Ruiz, D. (2012). A fast algorithm for matrix balancing. IMA J. Numer. Anal. drs019.

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*, 289–293.

Wang, S., Su, J.-H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C., and Zhuang, X. (2016). Spatial organization of chromatin domains and compartments in single chromosomes. Science *353*, 598–602.

HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps

Supplementary Methods

Data and the contact maps

Hi-C data are generated by the ENCODE consortium. Data used in this study came from 11 cancer cell lines (A549, Caki2, G401, LNCaP, NCI-H460, Panc1, RPMI-7951, SK-MEL-5, SK-N-DZ, SK-N-MC, T47D). Raw reads can be downloaded from the ENCODE portal (https://www.encodeproject.org). See Supplementary Table 1 for details. Contact maps were generated by the tool cworld developed in the Dekker lab (https://githib.com/dekkerlab/cworld-dekker). The bin size was set to be 40kb. The analysis reported in Figure 1 was performed on a chromosome-by-chromosome basis. The intra-chromosomal contact maps were not normalized (balanced).

Biological replicates, pseudo-replicates and non-replicates

Pairs of experiments were divided into three classes: biological replicates, pseudo-replicates, and non-replicates. Biological replicates refer to two experimental replicates of the same cell line. The 11 cancer cell lines resulted in 11 pairs of biological replicates. For pseudo-replicates, reads from a pair of biological replicates are pooled together and down sampled into two groups. The procedures were repeated three times for each cell line, arriving at 33 replicates. For non-replicates, the 11 cell types result in 55 pairings. Because there are two replicates for each cell type, there will be 4 possible pairings for each pair of non-replicate cell type pairing. Two of them were used to generate 2*55=110 non-replicate pairs. For each pair of the experiment, 23 reproducibility scores were obtained corresponding to chromosome 1 to 22 and chromosome X.

More details on Laplacian and the definition of S_d

The starting point of spectral analysis is the Laplacian matrix L, which is defined as $L = D - W$. Here D is a diagonal matrix in which $D_{ii} = \sum_j W_{ij}$. The Laplacian matrix further takes a normalized form $\mathcal{L} = D^{-1/2}LD^{-1/2}$. The normalized Laplacian matrix is closely related to a random-walk-process taking place in the underlying graph. The transition matrix for the random walk is $P = WD^{-1}$. The transition matrix and the normalized Laplacian matrix differ by only a similarity transform in which $\mathcal{L} = D^{-1/2}PD^{1/2}$ and therefore they share the same set of eigenvalues. The leading eigenvector captures the steady state distribution; the next few eigenvectors correspond to the slower decay modes of the random walk process and capture the domains that are highly important in contact maps.

The spectral decomposition theorem provides a natural way to separate a matrix into components,

$$\mathcal{L} = \sum_{i=1}^{N} \lambda_i \, v_i^A v_i^{A\prime}$$

Like common dimensionality reduction, keeping the first few eigenvectors separate signal from noise. Spectral theorem offers a natural way to separate a matrix. A simple-minded approach is to treat all matrix elements independently and define a metric by correlation coefficient. Nevertheless, such a simple metric cannot separate contact maps between a pair of biological replicates from maps generated from two different cell lines (Figure S3).

Though the leading eigenvectors tend to capture the large-scale structures of the graph, there are cases such that the eigenvectors are very localized. The localization can be captured by the so-called inverse participation ratio (IPR). Given a unit eigenvector v_i^A, the inverse participation ratio is defined as

$$IPR = \frac{1}{\sum_{j=1}^{N} \left(v_{ij}^A\right)^4}$$

If the eigenvector concentrated on a single node (for instance, v=[1,0,0,…,0]), then $IPR = 1$. If the eigenvector is uniformly spread over the whole graph, i.e. $v = [\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}, \ldots \frac{1}{\sqrt{N}}]$, then $IPR = N$. To find the distance between two contact maps, the localized leading vectors (IPR less than or equal to 5) were filtered.

The distance metric we employed between two matrices is slightly more complicated than the Euclidean distance displayed in Equation (1). Suppose $d(v^A, v^B)$ is the standard Euclidean distance between v^A and v^B. The distance we employed is in fact $\min(d(v^A, v^B), d(v^A, -v^B))$. This is because the sign of an eigenvector is arbitrary (i.e. if v^A is an eigenvector, $-v^B$ is an eigenvector).

Raw versus normalized matrices

As a metric to define the reproducibility of experiments, we focus on raw maps. If contact maps are normalized (ICED) in a whole-genome-to-whole-genome fashion, intra-chromosomal reproducibility scores are well defined. Nevertheless, if A and B are both normalized, the metric $S_d(A,B)$ cannot capture the distance properly. It is because the normalization procedure has drastically transformed the eigenvalues spectrum of a matrix. The leading eigenvectors of two normalized matrices appear to be very similar. Unless more eigenvectors are included, the metric $S_d(A,B)$ cannot capture the distance between two matrices.

The relationship between S_d and the reproducibility score

The distance between a pair of unit vectors whose components are randomly sampled from a standard normal distribution follows a distribution with mean $l = \sqrt{2}$. Figure S2 shows the Euclidean distance between pairs of eigenvectors. As the distance between a

pair of high-order eigenvectors is close to l, in other words, they are essentially noise terms, whereas the signal is stored in the leading vectors. Based on l, the distance S_d between A and B is rescaled to a reproducibility score ranging from 0 to 1 by a linear function:

$$Q(A,B) = \left(1 - \frac{1}{r}\frac{S_d}{l}\right)$$

Contribution of the diagonal elements

It is known that the number of interactions between two loci in the same chromosome that are close together proximally is noisy. To explore this thread, we examined at the contribution of the diagonal elements. Using the 11 pairs of biological replicates, we recalculated the reproducibility scores for 253 (11x23) pairs of intra-chromosomal contact maps by removing the diagonal entries of the contact maps. The new set of reproducibility scores is well correlated (PCC=0.82) with the original set of reproducibility scores based on the full maps (Figure S4).

Implementation and Benchmark

The runtime is efficient but depends on the size of contact maps. Figure S5 shows a benchmark based on contact maps from two replicates of A549 using the Julia implementation. The elapsed time for calculating Q depends on the size of the chromosome. The calculation was performed on a laptop with 2.8GHz Intel Core i7 and 16Gb of RAM.

Supporting Figures

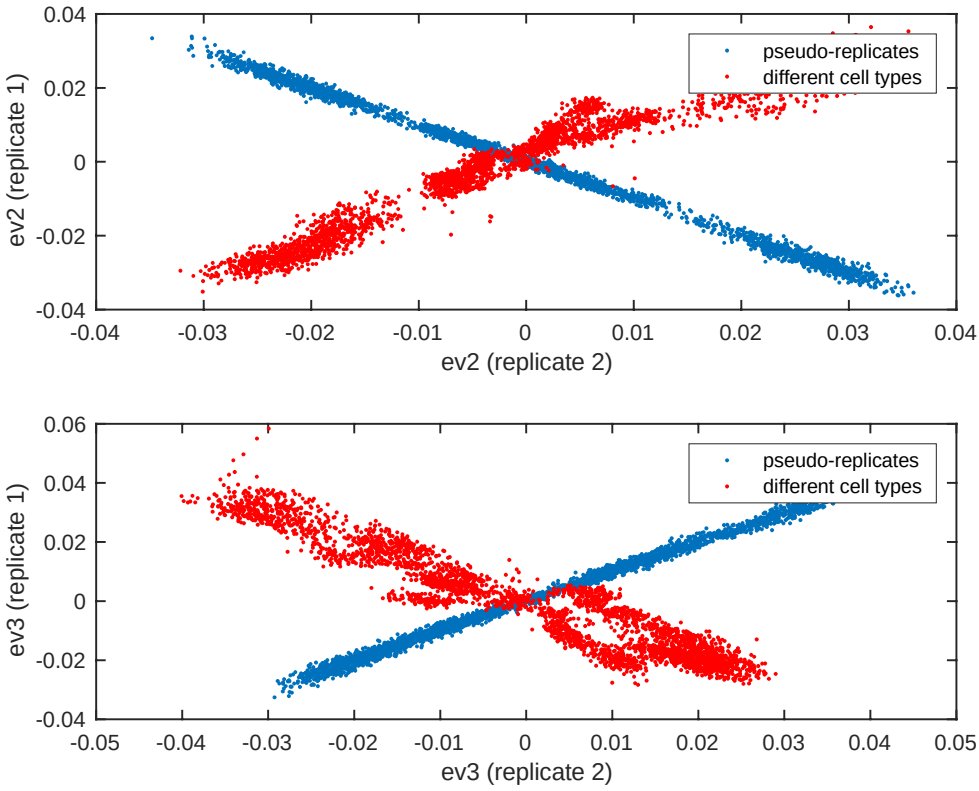

Figure S1: Leading eigenvectors of contact maps. Blue refers to a pair of pseudo-replicates. The corresponding leading eigenvectors are more correlated as compared to red, which refers to a pair of contact maps originating from two different cell lines.

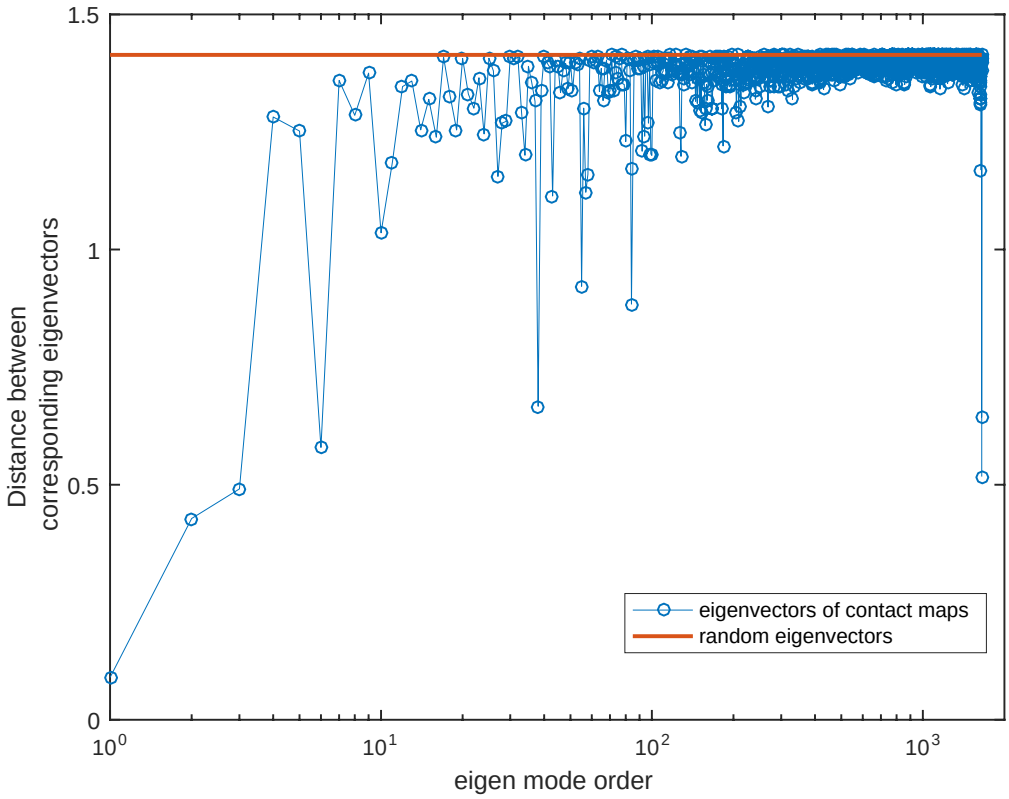

Figure S2: Euclidean distance between corresponding eigenvectors in a pair of Hi-C contact maps. The distance between leading eigenvectors is low. The red line is the distance between two random unit vectors whose components are sampled from a standard normal and then rescaled. The distance between two high-order eigenvectors is very close to the red line, suggesting they are noise instead of actual signal.

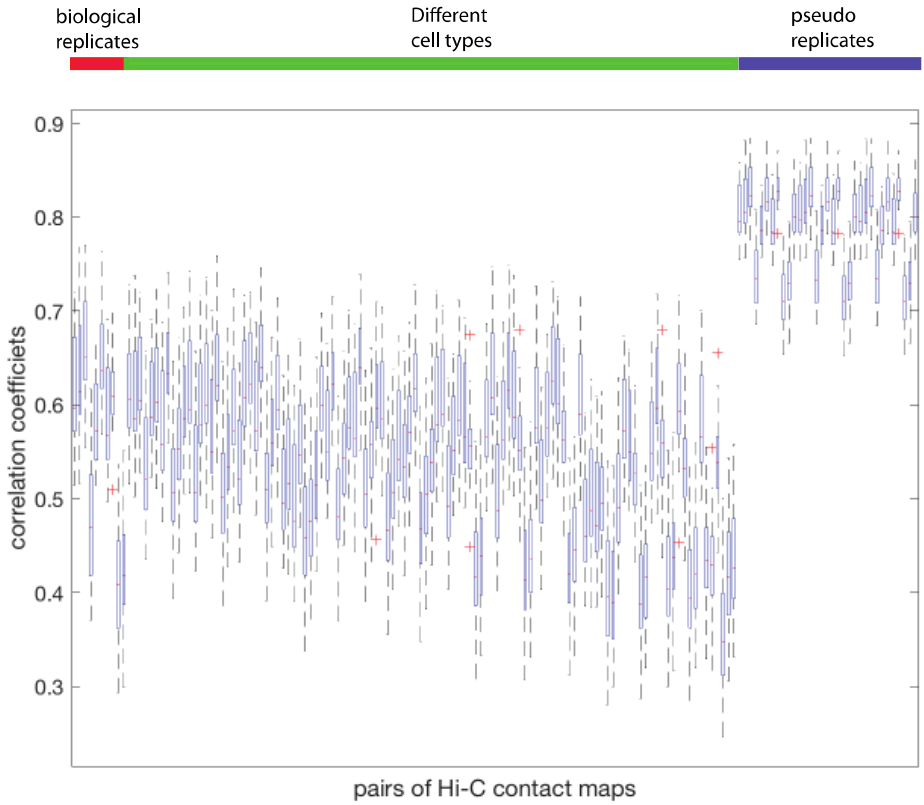

Figure S3: Correlation coefficients between pairs of Hi-C contact maps. Compared to Figure 1 in the main text, though a pair of pseudo-replicates are highly correlated, correlation coefficients cannot separate biological replicates with non-replicates. To find the correlation coefficient for a pair of matrices, all matrix elements are regarded independently. A pseudo-count of value 1 is added to each entry. Logarithm is taken before Pearson correlation coefficient is obtained.

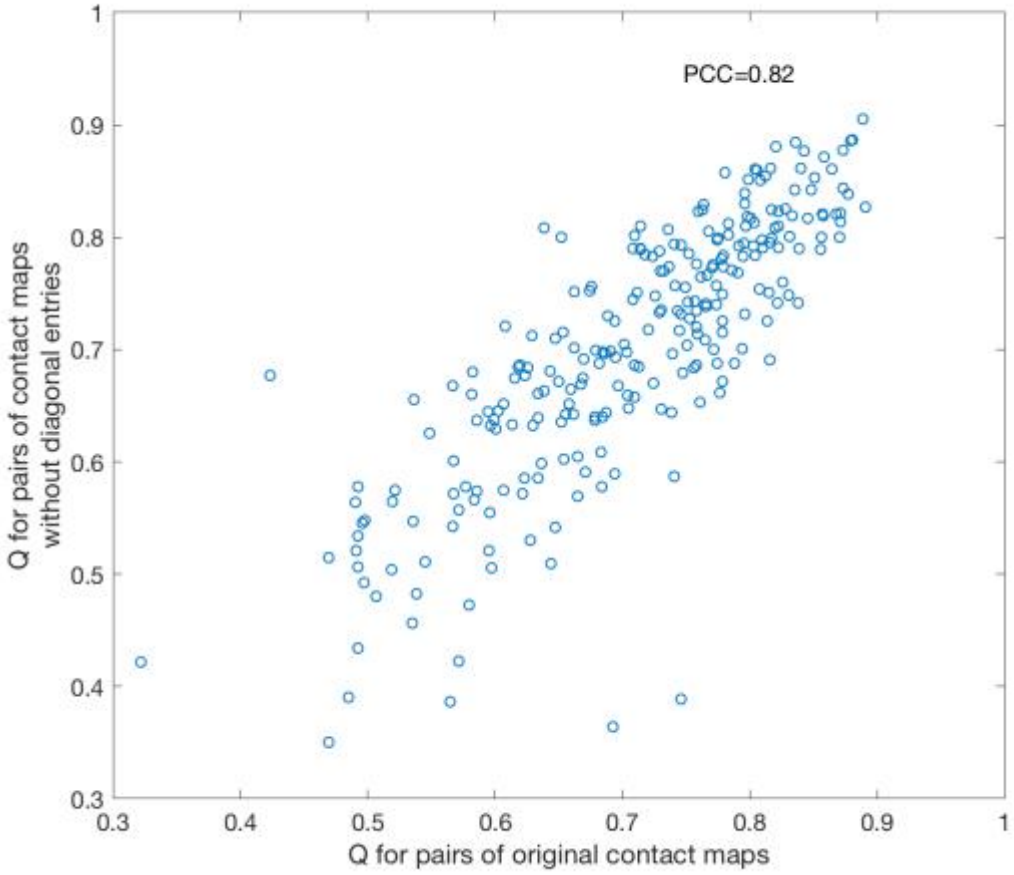

Figure S4: Sensitivity to noise along the diagonal. X-axis: Reproducibility scores for 11x23 contact maps obtained from 11 pairs of biological replicates. Y-axis: Reproducibility scores are recalculated in which the diagonal entries of the maps were removed. The two sets of scores are highly correlated (PCC=0.82).

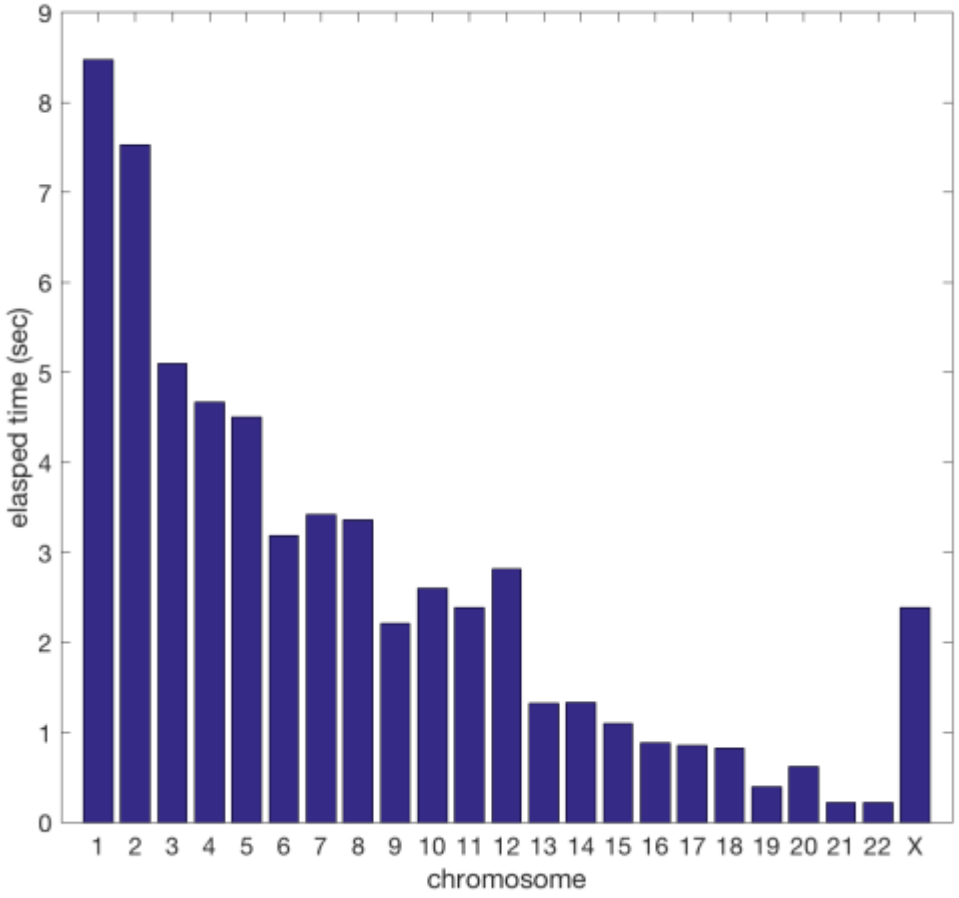

Figure S5: Elapsed time for calculating the reproducibility score for different chromosomes. The bin size used for binning the genome is 40kb.

cell type	# interactions (millions)	Library in ENCODE portal
A549	33	ENCLB571HTP
	30	ENCLB222WYT
Caki2	36	ENCLB555CZE
	47	ENCLB858SVS
G401	61	ENCLB506SDM
	53	ENCLB589RBY
LNCaP	18	ENCLB191OGC
	15	ENCLB473XWD
NCI-H460	42	ENCLB118KAE
	29	ENCLB104ZTM
Panc1	37	ENCLB951HSJ
	51	ENCLB134IVX
RPMI-7951	32	ENCLB210AAY
	49	ENCLB016TGU
SK-MEL-5	46	ENCLB296ZFT
	11	ENCLB462TWE
SK-N-DZ	16	ENCLB524GGK
	10	ENCLB952BSP
SK-N-MC	25	ENCLB215KZO
	13	ENCLB914GYK
T47D	34	ENCLB758KFU
	36	ENCLB183QHG

Table S1: Details of ENCODE Hi-C datasets. Each cell line has two replicates.