Response letter for resubmission

Reviewer 1

-- Ref 1.1 –source code--

Reviewer Comment	The tool did not supply the detailed documents and manual. The source code is listed in github, but I only found a very short readme file. The source code is not finally cleaned.
Author Response	Thanks for the advice. We have cleaned up the source code and expanded the readme file. We have provided links to a set of test data and added a tutorial showing how to use the code with the test data.

EXCEPT

-- Ref 1.2 –Details of ENCODE data--

Reviewer Comment	In the manuscript, authors used ENCODE data to demonstrate the reproducibility of replicates. What's the meaning of box and red cross? Why there are 33 pairs of pseudo replicates? Author should include one example (using ENCODE data) to illustrate how to use the package.
Author Response	We have included a set of ENCODE data and a tutorial for illustrating our package. Figure 1 is a standard boxplot in which a box represents the first and third quartiles, and the red crosses are outliers. We have modified the caption for clarification. The generation of three kinds of replicates was explained in detail in the Supplementary information.
Excerpt From Revised Manuscript	Figure 1. ….The boxplot shows the distribution of Q in 23 chromosomes, with red crosses as the outliers.

-- Ref 1.3 –Benchmark--

Reviewer Comment	The author tested the running time of Hi-Spector, but the whole section (Benchmark) is not clear and locks of details.
Author Response	We have put additional analysis and details in the Supplementary Information.

Reviewer 2

-- Ref 2.1 –Intuition of the Laplacian--

Reviewer Comment	The idea of considering Hi-C contact matrix as a graph and using a method in graph theory to solve the problem is interesting and potentially useful. However, the authors should give a more detailed explanation for the biological meaning of the Laplacian matrix or the advantage of using the eigenvectors of the Laplacian matrix. The authors simply state that the normalized Laplacian is related to random walk process in the graph and that the eigenvalues “capture the large-scale structure” of the contact map, but why this is appropriate for the problem at hand is not clearly explained
Author Response	We have reorganized the text and added a paragraph on the mathematical intuition behind our method. More details were given in the Supplement. Essentially, the eigenvectors offer a canonical way to decompose a contact map. We have further shown that our method can better separate biological replicate and non-replicates compared to the correlation coefficient suggested by the reviewer, which is essentially a different way to decompose a matrix such that all elements are independent.
Excerpt From Revised Manuscript	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 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 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 pseudo replicates, biological replicates, and non-replicates compared to the simple-minded correlation coefficient (see Figure S3 and the Supplement).

BUT γ IS CALLED L ?

-- Ref 2.2 –Reproducibility for other features--

Reviewer Comment	Reproducibility of Hi-C matrices could be performed for various features. (A/B compartments, TADs, loops, distance dependence etc.) This metric is probably most sensitive to A/B compartment type long-range structures. This is fine, but it should be stated.

Author Response	We agree with the reviewer that two contact maps can be compared on many different levels. What we refer to as “reproducibility” here focuses on the direct comparison of the contact maps based directly on the matrix elements. A comprehensive comparison of features like TADs and loops depends strongly on the choices of methods and parameters, which is beyond the scope of this manuscript.

SAY ONE SENT. ?

-- Ref 2.3 –Comparison with other methods--

Reviewer Comment	How does this compare to other methods used for measuring Hi-C data reproducibility? For example, how much improvement is there by using HiC-Spector compared with the simple Spearman correlation? An exhaustive comparison may be out of scope, but there is currently no comparison
Author Response	Thanks for the suggestion. We have performed a comparison with a simple correlation metric. We used the Pearson correlation for the logarithmic values of the matrix elements (plus a pseudo count), which is close to Spearman correlation. We found that such a simple-minded method cannot separate biological replicates and non-replicates (Figure S3). This means that treating all matrix elements independently is not a good way to define reproducibility. We have included this analysis in the Supplement.

EXCEPT

-- Ref 2.4 –Noise along the diagonal--

Reviewer Comment	How sensitive is the metric to noise along the diagonal? Some tools specifically tune out regions <20kb or so, for some analyses.
Author Response	Thank you very much for this question. We recalculated the reproducibility scores by removing the diagonal entries of the contact maps. We found that the two set of reproducibility scores agree pretty well. For instance, based on 253 (11*23) pairs of matrices from the 11 pairs of biological replicates, the correlation coefficient between the 2 sets is 0.82 (see Figure S4). We have put this analysis in the Supplement.

-- Ref 2.5 –Documentation--

Reviewer	The documentation should be more detailed to allow for

Comment	easier testing of the software, e.g., the format of input and output. The current version does not seem to support popular formats such as .hic.
Author Response	We have included a set of test data and a tutorial for software testing. The Julia version allows for a standard HiC-Pro format. The new python script further allows both HiC-Pro and .hic format in file input.

-- Ref 2.6 –Raw versus normalized matrices--

Reviewer Comment	Could this tool be used both on raw and normalized matrices?
Author Response	We have focused on raw matrices because they are the direct results of the experiments. We did calculate the reproducibility scores for pairs of normalized matrices; however, the results were not satisfactory. This is because in a normalized matrix, for each row or column, the sum is 1. The vector [1,1,...1] will be by definition an eigenvector of the matrix. Essentially, the normalization procedure transforms the spectrum such that many of the leading eigenvectors are close to [1,1,...1]. Consequently, the leading eigenvectors of two normalized matrices appear to be very similar. Unless more eigenvectors are included, the metric cannot capture the distance between two matrices. However, if the normalization is performed at the whole genome level but reproducibility is quantified on the intra-chromosomal level, then the metric we defined should work. In this submission, we have included these technical issues in the Supplement.

-- Ref 2.7 –Issue with Julia--

Reviewer Comment	The source code for HiC-Spector was written in Julia, but from my experience of installing and debugging Julia and the software package on a cluster environment, I am a little concerned about its user-friendliness for the general community. It took me many hours to install and run the software.
Author Response	Julia is a rather new technical language. We understand the reviewer's concern. Therefore, we have included a python script for calculating the reproducibility score. The script can be easily run in command line mode.

-- Ref 2.8 –W matrix--

Reviewer Comment	8. In Line 38 on the right column: 'The larger the value of W_ij, the closer is the distance between loci i and j'. True in the graph theoretic sense, but not in the biological sense?
Author Response	Using interaction frequency as a measure (inverse measure) of physical distance is a basic assumption behind Hi-C experiments. Recent experiments based on single molecular imaging provided strong evidence in support of this assumption (Wang et al. Science 2016). Under this proxy, the mathematical statement is true in the biological sense. On the other hand, there is no graph theoretical restriction on defining a large value of W as a close distance. We have modified the text for clarification.
Excerpt From Revised Manuscript	The matrix elements represent the frequencies of contact between genomic loci and therefore serve as a proxy of spatial distance. In principle, the larger the value of W_{ij, biologically the closer is the distance between loci i and j.

ADD

-- Ref 2.9 –Knight-Ruiz--

Reviewer Comment	9. In Line 55 on the second page, 'For instance, we have a function ... contact maps (Imakaev et al., 2012)'. The algorithm used for contact maps normalization in Imakaev et al., 2012 is not the Knight-Ruiz algorithm.
Author Response	We knew that. We have modified the text for clarification..
Excerpt From Revised Manuscript	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.

-- Ref 2.10 –Equations--

Reviewer Comment	10. Equation (1) does not seem to match what is in the source code. Please double-check. 11. In Equation (2), Q(A,B) does not range from 0 to 1 as stated and is also inconsistent with what is described in the source code. Please double-check.
Author Response	Thank you very much for pointing these out. There are typos in the equations. There should not be a power of 2 in Equation (1) and the

	factor 1/r should appear inside the bracket for Equation (2). However, we should mention that the source code is right. Indeed, the source code does slightly more than what is written in the equations. First, some chromosomal bins have zero coverage and thus the entries of some rows or columns are all zeros. Those columns or rows are excluded for calculating the eigenvectors. The eigenvectors of matrices A and B are then matched to arrive at v^A and v^B with the same dimension. Second, while the leading eigenvectors tend to capture the large-scale structures of the graph, in some rare cases, there are eigenvectors which are extremely localized. We have therefore filtered such eigenvectors. Moreover, the actual distance between v^A and v^B is not the simple Euclidean distance displayed in Equation (1). This is because the sign of an eigenvector is free to change (i.e. if v_A is an eigenvector, -v_A is an eigenvector). Suppose d(v^A, v^B) is the standard Euclidean distance. The distance we employed is in fact min(d(v^A, v^B),d(v^A, -v^B)). All these technical details are hard to summarize in a single mathematical equation; we therefore use Equation (1) to illustrate the essential idea. Nevertheless, in this revision, we have added the details in the Supplement. For Equation 2, for the sake of simplicity, we have moved it and the details to the Supplement.

Reviewer 3

-- Ref 3.1 -Details of ENCODE data--

Reviewer Comment	list the detailed information on the datasets they used for figure 1. List the cell types, # of reads, link for download
Author Response	We have provided the information in Supplementary Table 1.

-- Ref 3.2 –Code Documentation--

Reviewer Comment	Write a tutorial/README for their code. Use real example, starting from two matrix files and how they computed the results.
Author Response	We have provided some of the contact maps used in this study and a tutorial script for illustrating how to use our code (see also our response to Ref. 1.1 and Ref.2.5).