HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient

Tao Yang[1], Feipeng Zhang[2], Galip Gurkan Yardimci[5] , Ross C. Hardison[1, 4] , William Stafford Noble[5, 6], Feng Yue[1, 3*], Qunhua Li[1, 2*]

1. Bioinformatics and Genomics Program, Pennsylvania State University, University Park, PA 16802

2. Department of Statistics, Pennsylvania State University, University Park, PA 16802

3. Department of Biochemistry and Molecular Biology, Pennsylvania State University College of Medicine, Hershey, PA 17033

4. Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802

5. Department of Genome Sciences, University of Washington, Seattle, WA 98105

6. Department of Computer Science and Engineering, University of Washington, Seattle, WA 98105

*Corresponding author

Email addresses:

TY: txy146@psu.edu; FZ: fxz13@psu.edu; GY: gurkan@uw.edu; RH: ross@bx.psu.edu; WN: william-noble@uw.edu; FY: fyue@hmc.psu.edu; QL: qunhua.li@psu.edu.

Abstract

Hi-C is a powerful technology for studying genome-wide chromatin interactions. However, current methods for assessing Hi-C data reproducibility ignore spatial features in Hi-C data, such as domain structure and distance dependence. We present the stratum-adjusted correlation coefficient (SCC), a reproducibility measure that accounts for these features. SCC can assess pairwise differences between Hi-C matrices under a wide range of settings and can be used to determine optimal sequencing depth. The measure consistently shows higher accuracy than existing approaches in distinguishing subtle differences in reproducibility and depicting interrelationships of cell lineages. The R package HiCRep implements our approach.

Keywords: Hi-C, reproducibility, quality control, stratification, chromatin interaction, 3D genome organization

Background

The three-dimensional (3D) genome organization across a wide range of length scales is important for proper cellular functions [1–3]. At large distances, non-random hierarchical territories of chromosomes inside the cell nucleus are tightly linked with gene regulation [4]. At a finer resolution, the interactions between distal regulatory elements and their target genes are essential for orchestrating correct gene expression across time and space (e.g. different tissues). A progression of high-throughput methods based on chromatin conformation capture (3C) [5] has emerged, including 4C [6], 5C [7], Hi-C [8], ChIA-PET [9], Capture Hi-C [10], and HiChIP [11]. These methods offer an unprecedented opportunity to study higher-order chromatin structure at various scales. Among them, the Hi-C technology and its variants are of particular interest due to their relatively unbiased genome-wide coverage and ability to measure chromatin interaction intensities between any two given genomic loci.

However, the analysis and interpretation of Hi-C data are still in their early stages. In particular, no sound statistical metric to evaluate the quality of Hi-C data has been developed. When biological replicates are not available, investigators often rely on either visual inspection of the Hi-C interaction heatmap or examination of the ratio of long-range interaction read pairs over the total sequenced reads [12–14], but neither of these approaches is supported by robust statistics. When two or more biological replicates are available, it is a common practice to use either Pearson or Spearman correlation coefficients between the two Hi-C data matrices as a metric for data quality [13,15–20]. However, such correlation approaches may lead to incorrect conclusions because they do not take into consideration the unique characteristics of Hi-C data, such as domain structures and distance dependence, which refers to the pattern that the chromatin interaction frequencies between two genomic loci, on average, decrease substantially as their genomic distance increases. As we will demonstrate here, two unrelated biological samples can have a high Pearson correlation coefficient, while two visually similar replicates can have a low Spearman correlation coefficient. It is also not uncommon to observe higher Pearson and Spearman correlations between unrelated samples than those between real biological replicates.

In this work, we develop a novel framework for assessing the reproducibility of Hi-C data that takes into account the unique spatial features of the data. Our method first minimizes the effect of noise and biases by smoothing the Hi-C matrix, and then it addresses the distance-dependence effect by stratifying Hi-C data according to their genomic distance. We further develop a stratum-adjusted correlation coefficient (SCC) as a measurement of Hi-C data reproducibility. The value of SCC, which ranges from -1 to 1, can be used to compare the degrees of differences in reproducibility. Our framework can also infer confidence intervals for SCC, and further it can estimate the statistical significance of the difference in reproducibility measurements for different data sets. We applied this framework to three different groups of publicly available Hi-C data sets. The SCC was able to distinguish biological replicates from non-replicates, whereas Pearson and Spearman correlations failed to do so consistently. We also show that the SCC metric can be used as a distance measure to compare Hi-C data matrices from different cell types. When comparing Hi-C data from human embryonic stem cells and lineage-specific, differentiated cells derived from them, we found that only SCC correctly resolved all the interrelationships between different cell lineages, demonstrating the power of the proposed framework. Our algorithm was implemented as an R package called HiCRep, which is freely available on GitHub.

Results

Spatial patterns in Hi-C data and their influence on reproducibility assessment

Unlike many other genomic data types, Hi-C data exhibits unique spatial patterns. One prominent pattern is the strong decay of interaction frequency with the increase of genomic distance between interaction loci, i.e. the so-called distance dependence. This pattern is generally thought to result from non-specific interactions, which are more likely to occur between loci at closer genomic distance than those at a greater distance [21,22]. This distance dependence is found consistently in every Hi-C matrix and is one of the most dominant patterns in the matrix of interaction frequencies measured by Hi-C [21]. This dominance of the high interaction frequencies at short distances generates strong but spurious association between Hi-C matrices even when the samples are unrelated, as revealed by the high Pearson correlation between any two Hi-C matrices. As an example, we computed the Pearson correlations of Hi-C contact matrices between two biological replicates and between two unrelated cell lines, hESC and IMR90 [23]. Strikingly, the Pearson correlation between a hESC sample and an IMR90 sample is even higher than the correlation between two biological replicates in hESC (ρ =0.92 vs ρ = 0.91), despite the high similarity between the biological replicates (Figure 1A). Further investigation confirmed that this is because the Hi-C data in hESC and IMR90 share a highly similar pattern of distance-dependent interactions (Figure 1B). Therefore, the Pearson correlation coefficient cannot distinguish real biological replicates from unrelated samples.

Another important pattern of Hi-C data is the domain structure in their contact maps. These structures represent contiguous regions in which loci tend to interact more frequently with each other than with outside regions. While the interactions within the structures can be highly variable between difference cell types, the domain structures, such as topologically associating domains (TADs), are stable across cell types [23–25]. Therefore, we expect a higher reproducibility at the domain level than at the individual contact level. This difference should be reflected in the reproducibility assessment. However, both Pearson and Spearman correlation coefficients only consider point interactions, and do not take domain structures into account. A consequence of this is that Spearman correlation can be driven to low values by the stochastic variation in the point interactions and overlook the similarity in domain structures. As a result, two biological replicates that have highly similar domain structures may have a low Spearman correlation coefficient; conversely, a sample may have a higher Spearman correlation with an unrelated sample than with its biological replicates when the stochastic variation is high. For instance, despite the high similarity between the biological replicates in IMR90 and hESC, their Spearman correlations are only 0.47 and 0.37, respectively. However, the Spearman correlation between an IMR90 sample and a hESC sample is higher (0.44) than the correlation between the two hESC replicates, even though there are many differences in the domain structures of the two cell lines. Therefore, we need a more sophisticated evaluation metric to incorporate both structural aspects of variation for a better assessment of the reproducibility of Hi-C data.

Overview of the HiCRep method

We develop a novel two-stage approach to evaluate the reproducibility of Hi-C data (Figure 2). The first stage is smoothing the raw contact matrix in order to reduce local noise in the contact map and to make domain structures more visible. The smoothing is accomplished by applying a 2D mean filter, which replaces the read count of each contact in the contact map with the average counts of all contacts in its neighborhood. In the second stage, we apply a stratification approach to account for the pronounced distance dependence in the Hi-C data. This stage proceeds in two steps. First we stratify the smoothed chromatin interactions according to their genomic distance, and then we apply a novel stratum-adjusted correlation coefficient statistic (SCC) to assess the reproducibility of the Hi-C matrices. The SCC statistic is calculated by computing a Pearson correlation coefficient for each stratum (Figure 2) and then aggregating the stratum-specific correlation coefficients using a weighted average, with the weights derived from the generalized Cochran-Mantel-Haenszel (CMH) statistic [26,27]. The value of SCC ranges from -1 to 1 and can be interpreted in a way similar to the standard correlation. A great advantage of our approach is that we can derive the asymptotic variance of SCC and use it to assess statistical significance when comparing reproducibility from different samples. More detailed descriptions of the HiCRep method and the SCC statistic are presented in the Methods section.

Distinguishing pseudo, real and non-replicates

We first evaluated the performance of our method on samples whose expected levels of reproducibility are known: pseudo-replicates (PR), biological replicates (BR) and non-replicates (NR). Biological replicates refer to two independent Hi-C experiments performed on the same cell types. Non-replicates refer to Hi-C experiments performed on different cell types. Pseudo replicates are generated by pooling reads from biological replicates together and randomly partitioning them into two equal portions. The difference between two pseudo-replicates only reflects sampling variation, without biological or technical variation. Therefore, we expect that the reproducibility of pseudo-replicates is the highest, followed by biological replicates and then non-replicates.

For testing, we first generated PR, BR and NR using Hi-C data in the hESC and IMR90 cell lines [23] (details in Methods). We compared the performance of SCC with Pearson correlation and Spearman correlation and investigated whether these metrics can distinguish PR, BR and NR (Figure 3A and Table S1). For the hESC dataset, SCC correctly ranks the reproducibility of the three types of replicate pairs (PR>BR>NR), whereas Pearson and Spearman correlations both incorrectly rank BR lower than one or more of the NRs. For the IMR90 dataset, although all three methods infer the correct order of reproducibility, SCC separates BR from NR by a much larger margin than the other metrics. For example, the largest difference between BR and NR reported by SCC is 0.24, compared to only 0.08 by Pearson and 0.13 by Spearman correlations.

The sequencing depths differ substantially for the hESC (replicate 1: 60M; replicate 2: 271M) and IMR90 (replicate 1: 201M; replicate 2: 153M) datasets. To ensure that these differences were not confounding our evaluations, we subsampled all the replicates to 60 million reads and repeated the same analysis. As shown in Figure 3A (blue dots), even with the same number of reads, Pearson and Spearman correlations still fail to distinguish real replicates from all non-replicates. On the contrary, our method consistently ordered the reproducibility of replicates correctly, indicating that SCC can capture the intrinsic differences between the samples, even those that differ in sequencing depth.

We expanded this analysis to a larger Hi-C dataset recently released by the ENCODE consortium. This dataset consists of Hi-C data from eleven cancer cell lines, with two biological replicates for each cell type (details are in Methods). For each cell type, we formed twenty non-replicate pairs with the remaining ten cell types and computed SCC, Pearson and Spearman correlations for BR and all NRs. As shown in Figure 3B and Table S2, SCC clearly distinguishes BRs from NRs (a p-value = 1.665×10^{-15}, one-sided Kolmogorov-Smirnov test), while the other two methods fail to do so (Pearson: p-value = 0.084; Spearman: p-value = 0.254, K-S test). Because the sequencing depth of the Hi-C data varies across cell types, we also examined the separation between BRs and NRs for each cell type. As shown in Figure 3C, SCC separates the BRs and NRs for all the cell types by a margin of at least 0.1, whereas the other two methods fail to separate them in more than half of the cell types (additional file 1: Figure S1). Furthermore, SCC illustrates a desirable correspondence to the sequencing depth. When the average sequencing depth between the biological replicates is relatively low (<30M), SCC monotonically increases with the sequencing depth; this behavior likely reflects insufficient coverage at the lower sequencing depths. In contrast, the value for SCC remains high and stable for greater sequencing depths (Figure 3C), reflecting saturation of reproducibility and likely reflecting sufficient coverage. We investigate this property further in a later section.

Evaluating biological relevance by constructing cell lineages

Next we used our method as a similarity measure to infer the interrelationship between cell types on a cell lineage. Because this assessment requires the reproducibility measure to quantify the subtle differences between closely related cells, it serves as a biologically relevant approach to evaluating the accuracy of the reproducibility measure. More importantly, it also evaluates the potential of our method as a measure for quantifying the similarities or differences of Hi-C matrices in different cell or tissue types.

For this analysis, we used the Hi-C data in human embryonic stem (ES) cells and in four cell lineages derived from them [13], namely, mesendoderm (ME), mesenchymal stem cells (MS), neural progenitor cells (NP), and trophoblast-like cells (TB), with two biological replicates for each cell type. Using the A/B compartments in Hi-C data, Dixon et al. [13] inferred the distance to the parental ES cell from the nearest to the farthest as ME, NP, TB and MS (Figure 4A). Importantly, the same relationships were also supported by previous analysis of gene expression data (additional file 1: Figure S2) in the same cell types [28].

We first calculated the pairwise similarities between the ten samples (two replicates in each cell type) using SCC, Pearson and Spearman correlations (Table S3). As shown in Figure

S3 (additional file 1), SCC again provided the best separation between real replicates and non-replicates among all three methods of comparison.

Next, we reconstructed the relationships among the cell lineages by performing hierarchical clustering based on the pairwise similarity scores. As shown in Figure 4B, the dendrogram constructed based on SCC precisely depicts the interrelationships: all the biological replicates are grouped together as terminal clusters, and the relationships between cell lines exactly follow the tree structure in Dixon et al. [13] and Xie et al. [28] (Figure 4A). In contrast, the dendrograms constructed based on Pearson (Figure 4C) and Spearman correlation coefficients (Figure 4D) group several non-replicates together and infer different relationships between some cell lines. For example, when using Pearson correlation, two ME replicates are not clustered together and NP is unexpectedly placed as the least related cell type to ES cells. When using Spearman correlation, an ES replicate is clustered with an ME replicate and again NP is unexpectedly predicted as the least related cell type to ES cells.

We further expanded this analysis using the recently published Hi-C data in fourteen human primary tissues and two cell lines [29] (Table S4). Because biological replicates are not available for all the samples, our analysis focused on quantifying the relationships between tissues or cells. Again, the lineage constructed based on SCC reasonably depicted the tissue and germ layer origins of the samples (Figure 5A): hippocampus and cortex were grouped together; right ventricle and left ventricle were grouped together; endodermal tissues such as pancreas, lung, and small bowel were placed in the same lineage. Neither Pearson nor Spearman correlation performed as well as SCC. For example, right and left ventricles were not grouped together by Spearman correlation (Figure 5C). These results confirm the potential of our method as a measure for quantifying the difference in Hi-C data between cell or tissue types.

SCC is robust to different choices of resolution

Depending on the sequencing depth, Hi-C data analysis may be performed at different resolutions. A good reproducibility measure should perform well despite the choice of resolution. To evaluate the robustness of our method, we repeated the clustering analysis for the human ES and ES-derived cell lineages using data processed at several different resolutions (i.e., 10Kb, 25Kb, 40Kb, 100Kb, 500Kb, 1Mb). Again, as shown in Figure 6 and Table S5, we observed that SCC inferred the expected relationships at all resolutions considered, whereas Pearson and Spearman correlations inferred the expected relationships only at 500Kb and 1Mb. Furthermore, unlike Pearson and Spearman correlations, whose values drastically change at different resolutions, the values of SCC remain in a consistent range across all resolutions. These results confirm the robustness of our method to the choice of resolution.

Detecting differences in reproducibility due to sequencing depth

Sequencing depth is known to affect the signal-to-noise ratio and the reproducibility of Hi-C data [21]. Insufficient coverage can reduce the reproducibility of a Hi-C experiment. As a quality control tool, a reproducibility measure is expected to be able to detect the differences in reproducibility due to sequencing depth. To evaluate the sensitivity of our method to sequencing depth, we subsampled all the samples in the H1 ES cell lineage [13] to create a series of subsamples with different sequencing depths (25%, 50% and 75% of the original sequencing depth). We then computed SCC for all samples. As shown in Figure 7A and Table S6, SCC monotonically decreases with sequencing depth in all data sets. This confirms that our method can reflect the change of reproducibility due to sequencing depth.

Using SCC to guide the selection of the optimal sequencing depth

Having established that SCC can reflect the change of reproducibility due to the change of sequencing depth, we propose to use the saturation of SCC as a criterion to determine the most cost-effective sequencing depth that achieves a reasonable reproducibility. To illustrate how to use our method to determine the optimal sequencing depth, we created subsamples at a series of reduced sequencing depths from the Hi-C data in the H1 ES cell in [13] (original depth=500M) by down-sampling. As shown in Figure 7B and Table S7, SCC initially increases with the increase of sequencing depth when the number of total reads is less than 200 million. SCC increases little after this point (less than 0.01) and eventually reaches a plateau (Figure 7B). To determine the lowest sequencing level that achieves similar reproducibility as the original data, we compared the 90% confidence intervals of SCC at all the reduced sequencing depths with that of the original depth. Starting at 300M (60% of the original depth), the confidence intervals overlap with that of the original depth. This indicates that the reduced samples can achieve a similar level of reproducibility as the original one by using about 60% of the original depth for this dataset. Further increase of sequencing depth beyond this point does not significantly improve reproducibility.

As a comparison, we performed a similar analysis using a dataset with relatively low sequencing depth (30M Hi-C reads from the A549 cell line). We observe that all the reduced samples with less than 90% of the original sequencing depth (27M) have a significantly lower reproducibility than the original sample at the 90% significance level (Figure 7B and Table S7). From 90% to the original depth, there is still an increase of SCC of 0.01, compared with less than 0.001 for the hESC dataset, suggesting that this dataset may not reach saturation in reproducibility at its original sequencing depth. For this dataset, further increase of sequencing depth may improve reproducibility.

Discussion

Although there has been a dramatic increase in the scope and complexity of Hi-C experiments, analytical tools for data quality control have been lacking. Current approaches for assessing Hi-C data reproducibility may lead to incorrect conclusions because they fail to take into consideration the unique spatial characteristics of Hi-C data. In this work, we developed a new method for assessing the reproducibility of Hi-C contact frequency maps. By effectively taking account of the spatial features of Hi-C data, our reproducibility measure overcomes the limitations of Pearson and Spearman correlations and can differentiate the reproducibility of samples at a fine level. The empirical evaluation showed that SCC distinguished subtle differences between closely related cell lines, biological replicates and pseudo replicates, and it produced robust results at different resolutions.

Our statistic has several properties that make it well-suited as a reproducibility measure for providing standardized, interpretable, automatable and scalable quality control. First, this statistic has a fixed scale of [-1, 1], which makes it easy to standardize the quality control process and compare reproducibility across samples. Second, our statistic is intuitive and easy to interpret. It can be interpreted as a weighted average correlation coefficient over different interaction distances. This straightforward interpretation makes it accessible to experimentalists. Third, our statistic is fast to compute and is directly applicable to the raw contact matrix. It is easily scalable for monitoring data quality for a large number of experiments. Furthermore, we also provide an estimator for the variance of this statistics, such that the statistical significance of the difference in reproducibility can be inferred. Using this estimator, we establish a procedure to determine the sufficiency of sequencing depth.

In summary, we develop a novel method to accurately evaluate the reproducibility of Hi-C experiments. The presented method is a first step toward ensuring high reproducibility of Hi-C data. We also show that this method can be used as a similarity measure for quantifying the differences in Hi-C data between different cell and tissue types. Thus, HiCRep is a valuable tool for the study of 3D genome organization. It is freely available as an R package at https://github.com/MonkeyLB/hicrep.

Methods

Data Preprocessing

We generated the Hi-C contact matrix using the pipeline from [13]. Briefly, the paired-end reads were first aligned to the hg19 reference genome assembly using BWA [30]. The unmapped reads were filtered, and potential PCR duplicates were removed using Picardtools (https://broadinstitute.github.io/picard/). For most analysis, we used 40kb bins. To obtain contact maps at this resolution, we divided the genome into 40kb bins as in [13] and obtained the interaction frequency by counting the number of reads falling into each pair of bins. Here we chose to apply our method directly to raw data without bias correction, so that the reproducibility assessment is free of assumptions made in the bias correction procedures [15,16,31], and faithfully reflects the nature of the raw data. Only the intra-chromosomal interactions were used for our analysis. Given that the interactions over 5Mb in distance are rare, only the contacts within the range of 0~5Mb were used in the reproducibility assessment. All the datasets were preprocessed using the same preprocessing procedure.

2D mean filter smoothing

Because the space of interactions surveyed by Hi-C experiments is very large, achieving sufficient coverage is still challenging. When samples are not sufficiently sequenced, the local variation introduced by under-sampling can make it difficult to capture large domain structures. To handle this issue, we first smooth the contact map before assessing reproducibility. Although smoothing will reduce the individual spatial resolution, it can improve the contiguity of the regions with elevated interaction, consequently enhancing the domain structures. It has been found effective in commonly-used Hi-C normalization methods [15,32].

We use a 2D mean filter to smooth the contact map. The filter replaces the read count of each contact in the contact map with the mean counts of all contacts in its genomic neighborhood. This filter is fast to compute and is effective for smoothing rectangular shapes [33] like domain structures in Hi-C data. Specifically, let $C_{n\times n}$ denote a $n \times n$ contact map and c_{ij} denote the counts of the interaction between loci i and j. Given a span size h>0, the smoothed contact map after passing a h^{th} 2D mean filter is defined as follows:

$$x_{ij}(h) = \frac{\sum_{m=\max(1,i-h)}^{\min(i+h,n)} \sum_{l=\max(1,j-h)}^{\min(j+h,n)} C_{ml}}{(1+2h)^2}$$

A visualization of the smoothing effect with different window sizes is shown in Figure S4 (additional file 1).

Selection of smoothing parameter

The span size h is a tuning parameter controlling the smoothing level. A very small h might not reduce enough local variation to enhance the boundaries of domain structures, while a large h will make the boundaries of domain structures blurry and limit the spatial resolution. Therefore, the optimal h should be adaptively chosen from the data.

To select *h* objectively, we developed a heuristic procedure to search for the optimal choice. Our procedure is designed based on the observation that the correlation between contact maps of replicate samples first increases with the level of smoothness and then plateaus when sufficient smoothness is reached. To proceed, we used a pair of reasonably deeply sequenced interaction maps as the training data. We randomly sampled 10% of the data ten times. For each subsample, we computed the stratum-adjusted correlation coefficient (SCC, described in a later section) at a series of *h*'s in the ascending order and recorded the smallest *h* at which the increment of SCC was less than 0.01. The mode of *h* among the ten subsamples was selected as the final span size. The detailed steps are shown in Algorithm 1 in additional file 1.

Because the level of local variation in a contact map depends on the resolution used to process the data, the span size required to achieve sufficient smoothness varies according to resolution. Hence, a proper *h* for each resolution needs to be trained separately. However, at a given resolution, it is desirable to use the same *h* for all datasets, so that the downstream reproducibility assessment can be compared on the same basis. To reduce the chance of over-smoothing due to sparseness caused by insufficient coverage when training *h*, we used a deeply sequenced data set as training data.

Here we obtained *h* in our analysis from the Human H1 ESC dataset [13]. This dataset was deeply sequenced (330M and 740M reads for its two replicates) and had a reasonable quality [13], making it suitable as training data. We processed the data using a series of resolutions (10Kb, 25Kb, 40Kb, 100Kb, 500Kb and 1Mb), and then selected *h* for each resolution using the procedure described above. We obtained *h*=20, 11, 5, 3, 1, and 0 for the resolution of 10Kb, 25Kb, 40Kb, 100Kb, 500Kb and 1Mb, respectively. These values were used throughout our study for all datasets at the corresponding resolutions. The robustness of our procedure was assessed using the Human H1 ESC dataset and four derived cell lines (details are in the Results section).

Stratification by distance

To take proper account of the distance effect in reproducibility assessment, we stratify the contacts by the genomic distance between their interaction loci. Specifically, let $X_{n\times n}$ be an $n\times n$ smoothed contact map at a resolution of bin size b. We compute the interaction distance for each contact x_{ij} as $d_{ij} = |j - i|\times b$ and then stratify the contacts by d_{ij} into K strata, $X_k = \{x_{ij}: (k\text{-}1)b < d_{ij} \leq kb\}, k = 1, \ldots K$. Here we consider the interaction distance of 0 ~ 5Mb. This leads to K = 125 for the bin size b = 40kb. If x_{ij} is 0 in both samples, then it is excluded from the reproducibility assessment.

Stratum-adjusted correlation coefficient (SCC)

Our reproducibility statistic is motivated from the generalized Cochran-Mantel-Haenszel (CMH) statistic M^2. The CMH statistic is a stratum-adjusted summary statistic for testing if two variables are associated while being stratified by the third variable [27], for example, the association between treatment and response stratified by age. Though originally developed for categorical data, it is also applicable to continuous data [26] and can detect consistent linear association across strata. However, the magnitude of M^2 depends on the sample size; therefore, it cannot be used directly as a measure of the strength of the association. When there is no stratification, the CMH statistic is related to the Pearson correlation coefficient ρ as $\mathrm{M}^2 = \rho^2(N\text{-}1)$, where N is the number of observations [27]. This relationship allows the strength of association summarized by M^2 to be represented using a measure that has a fixed scale and is comparable across different samples. However, ρ does not involve stratification. This motivates us to derive a stratum-adjusted correlation coefficient (SCC) to summarize the strength of association from the CMH statistic when there is stratification.

Derivation of stratum-adjusted correlation coefficient (SCC)

Let *(X, Y)* denote a pair of samples with *N* observations. The observations are stratified into *K* strata, and each stratum has N_k observations such that $\sum_{k=1}^{K} N_k = N$. Denote the observations in stratum *k* as (x_{1_k}, y_{1_k}), …, (x_{N_k}, y_{N_k}) and the corresponding random variables as (X_k, Y_k), respectively. In our context, (x_{i_k}, y_{i_k}) are the smoothed counts of the *i*th contact on the *k*th stratum in the two contact maps *X* and *Y*. Let $T_k = \sum_{i=1}^{N_k} x_{i_k} y_{i_k}$, the CMH statistics is defined as

$$M^2 = \frac{\left[\sum_k [T_k - E(T_k)]\right]^2}{\sum_k Var(T_k)}, \qquad (1)$$

where $E(T_k)$ and $Var(T_k)$ are the mean and variance of T_k under the hypothesis that X_k and Y_k are conditionally independent given the stratum,

$$E(T_k) = \frac{\sum_{i=1}^{N_k} x_{i_k} \sum_{j=1}^{N_k} y_{j_k}}{N_k}, \qquad (2)$$

and

$$Var(T_k) = \frac{1}{N_k - 1}\left[\sum_{i=1}^{N_k} x_{i_k}^2 - \frac{\left(\sum_{i=1}^{N_k} x_{i_k}\right)^2}{N_k}\right]\left[\sum_{j=1}^{N_k} y_{j_k}^2 - \frac{\left(\sum_{j=1}^{N_k} y_{j_k}\right)^2}{N_k}\right]. \qquad (3)$$

To derive the stratum-adjusted correlation coefficient from the CMH statistic, write the Pearson correlation coefficient ρ_k for the kth stratum as $\rho_k = \frac{r_{1k}}{r_{2k}}$, where

$$r_{1k} = E(X_k Y_k) - E(X_k)E(Y_k) = \frac{\sum_{i=1}^{N_k} x_{i_k} y_{i_k}}{N_k} - \frac{\sum_{i=1}^{N_k} x_{i_k} \sum_{j=1}^{N_k} y_{j_k}}{N_k{}^2} \qquad (4)$$

$$r_{2k} = \sqrt{\mathrm{var}(X_k)\,\mathrm{var}(Y_k)} = \sqrt{\left[\frac{\sum_{i=1}^{N_k} x_{i_k}^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} x_{i_k}}{N_k}\right)^2\right]\left[\frac{\sum_{i=1}^{N_k} y_{i_k}^2}{N_k} - \left(\frac{\sum_{i=1}^{N_k} y_{i_k}}{N_k}\right)^2\right]} \qquad (5)$$

It is easy to see that $r_{1k} = \frac{T_k - ET_k}{N_k}$ and $r_{2k} = \frac{\sqrt{(N_k - 1)Var(T_k)}}{N_k}$. Then we can represent M^2 using ρ_k,

$$M^2 = \frac{\left(\sum_{k=1}^{K} N_k r_{2k} \rho_k\right)^2}{\sum_{k=1}^{K} (N_k r_{2k})^2 / (N_k - 1)}. \qquad (6)$$

Define

$$\rho_s = \frac{\sum_{k=1}^{K} N_k r_{2k} \rho_k}{\sum_{k=1}^{K} N_k r_{2k}} = \sum_{k=1}^{K} \left(\frac{N_k r_{2k}}{\sum_{k=1}^{K} N_k r_{2k}} \right) \rho_k \;, \qquad (7)$$

then

$$M^2 = \frac{\rho_s{}^2}{\frac{1}{\left(\sum_{k=1}^{K} N_k r_{2k}\right)^2} \sum_{k=1}^{K} (N_k r_{2k})^2 / (N_k - 1)}. \; (8)$$

From (8), it can be seen that $\rho_s{}^2$ reflects the strength of association in M^2. This strength relates to M^2 in a similar way as the Pearson correlation to M^2 in the case without stratification. As shown in (7), ρ_s is a weighted average of the stratum-specific correlation coefficients, with weights $w_k = \frac{N_k r_{2k}}{\sum_{k=1}^{K} N_k r_{2k}}$ assigned according to the variance and sample size of a stratum. We call ρ_s the stratum-adjusted correlation coefficient (SCC). Similar to standard correlation coefficients, it satisfies $-1 \le \rho_s \le 1$. A value of $\rho_s = 1$ corresponds to a perfect positive correlation, a value of $\rho_s = 1$ corresponds to a perfect negative correlation, and a value of $\rho_s = 0$ corresponds to no correlation.

The variance of ρ_s can be computed as

$$\mathrm{var}(\rho_s) = \frac{\sum_{k=1}^{K} (N_k r_{2k})^2 \, \mathrm{var}(\rho_k)}{\left(\sum_{k=1}^{K} N_k r_{2k}\right)^2}, \qquad (9)$$

where $\mathrm{var}(\rho_k) = \frac{\sum_{k=1}^{K} \left(\rho_k - \sum_{k=1}^{K} \rho_k \big/ K\right)^2}{K}$ is the asymptotic standard error for the Pearson correlation coefficient in a single stratum [34].

The idea of obtaining an average correlation coefficient based on the CMH statistic has been explored in [35] in the context of contingency tables with ordered categories. However, its derivation has several errors, which lead to a different statistic that ignores the sample size differences in different strata.

Variance stabilized weights

The downside for Equation (7) is that it is based on the implicit assumption in the CMH statistic that the dynamic ranges of X and Y are constant across strata. However, in Hi-C data, the read counts for contacts with short interaction distances have a much larger dynamic range than those with long interaction distances. As a result, the weights for the strata with large dynamic ranges will dominate (7), due to the large values of their r_{2k}.To normalize the dynamic range, we rank the contact counts in each stratum separately and then normalize the ranks by the total number of observations N_k in each stratum, such that all strata share a similar dynamic range. We then compute r_{2k} in the weights in (7) and (9) using the normalized ranks, instead of the actual counts, i.e.

$$r_{2k} = \sqrt{Var\left[\frac{Rank(X_k)}{N_k}\right]Var\left[\frac{Rank(Y_k)}{N_k}\right]} \; (10).$$

The stratum-specific correlation ρ_k is still computed using actual values rather than ranks, as actual values have better sensitivity than ranks when there are a large number of low counts.

Implementation of our pipeline

We have implemented our method as an R package. It is publicly available as the HiCRep package on GitHub https://github.com/MonkeyLB/hicrep.

Declarations

Funding

This work was supported by NIH training grant T32 GM102057 (CBIOS training program to The Pennsylvania State University), a Huck Graduate Research Innovation Grant to TY, and by NIH grants R01GM109453 (to QL), U01CA200060 (to FY), R24DK106766 (to RCH and FY), U54HG006998 (to RCH), and U41HG007000 (to WSN).

Availability of data and materials

HiCRep has been implemented in R and deposited at Github with URL https://github.com/MonkeyLB/hicrep. The datasets analyzed in this study were obtained from the public domain, as described below. Hi-C data sets used in this project can be visualized in the 3D genome browser (http://3dgenome.org).

We obtained the Hi-C data of human embryonic stem cells (hESCs) and human IMR90 fibroblasts from Dixon, et al. (2012) [23] (GEO accession number: GSE35156). Each cell type has two biological replicates.

We obtained the Hi-C data of human embryonic stem (ES) cells and four human ES-cell-derived lineages, mesendoderm (ME), mesenchymal stem (MS) cells, neural progenitor (NP) cells and trophoblast-like (TB) cells from Dixon, et al. (2015) [13] (GEO accession number: GSE52457). Each cell type has two biological replicates.

We obtained the Hi-C data of eleven human cancer cell lines from the ENCODE consortium (https://www.encodeproject.org). This dataset includes cell lines of G401, A549, CAKi2, PANC1, RPMI7951, T47D, NCIH460, SKMEL5, LNCaP, SKNMC and SKNDZ. Each cell line has two biological replicates. The sequencing depths of the datasets can be found in Table S8.

We obtained the Hi-C data of fourteen human primary tissues from Schmitt, et al. (2016) [29] and Leung, et al. (2015) [36]. The tissues include adrenal gland (GSM2322539), bladder (GSM2322540, GSM2322541), dorsolateral prefrontal cortex (GSM2322542), hippocampus (GSM2322543), lung (GSM2322544), ovary (GSM2322546), pancreas (GSM2322547), psoas muscle (GSM2322551), right ventricle (GSM2322554), small bowel (GSM2322555), spleen (GSM2322556), liver (GSM1419084), left ventricle (GSM1419085), and aorta (GSM1419086). The tissues were collected from four donors, each of which provides a subset of tissues. To minimize variation due to individual difference, we used the samples from the two donors with the largest number of tissues. If one tissue sample consists of multiple replicates from a single donor, the replicates were merged into a single dataset. We obtained the GM12878 cell data from Selvaraj et al. (2013) (GSM1181867, GSM1181867) [37] and the IMR90 cell data from Dixon, et al. (2012) (GSM862724, GSM892307) [23].

Authors' contributions

QL and FY designed the study. QL and TY formulated the problem and developed the computational method and the SCC statistic. FZ provided assistance in the formulation of SCC statistic. QL, FY and TY designed the data analysis and performance evaluation. GY and WN designed the data analysis for ENCODE cancer cell data. TY implemented the method, developed the software package and analyzed the data. QL, TY, and FY interpreted the results. RH provided assistance on the biological interpretation of the results. TY, RH, FY and QL wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Ethics approval is not applicable for this study.

References

1. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. 2013;14:390–403.

2. Sexton T, Cavalli G. The role of chromosome domains in shaping the functional genome. Cell. 2015. p. 1049–59.

3. Bickmore WA. The Spatial Organization of the Human Genome. Annu. Rev. Genomics Hum. Genet. Annual Reviews; 2013;14:67–84.

4. Misteli T. Higher-order genome organization in human disease. Cold Spring Harb. Perspect.

Biol. 2010.

5. Dekker J. Capturing Chromosome Conformation. Science. 2002;295:1306–11.

6. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006;38:1348–54.

7. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16:1299–309.

8. Lieberman-aiden E, Berkum NL Van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326:289–93.

9. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, et al. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. 2010;11:R22.

10. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;47:598–606.

11. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Meth. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.;

2016;13:919–22.

12. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012/04/13. 2012;485:376–80.

13. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015;518:331–6.

14. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;503:290–4.

15. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. - Supplement. Nat. Methods. 2012;9:999–1003.

16. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: Removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28:3131–3.

17. Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell. 2014. p. 771–5.

18. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.

19. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. Genome Biol. 2015;16:183.

20. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.

21. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods. 2015;72:65–75.

22. Fudenberg G, Mirny LA. Higher-order chromatin structure: Bridging physics and biology. Curr. Opin. Genet. Dev. 2012. p. 115–24.

23. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.

24. Rapkin LM, Anchel DR, Li R, Bazett-Jones DP. A view of the chromatin landscape. Micron. 2012;43:150–8.

25. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502:59–64.

26. Mantel N. Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure. J. Am. Stat. Assoc. 1963;58:690–700.

27. Agresti A. Categorical Data Analysis. 3rd ed. New York: Wiley. 2012.

28. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013;153:1134–48.

29. Schmitt AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. 2016;17:2042–59.

30. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.

Bioinformatics. 2009;25:1754–60.

31. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC Genomics. 2012;13:436.

32. 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.

33. Davies R. Computer and Machine Vision: Theory, Algorithms, Practicalities. 4th ed. Oxford: Academic Press. 2012.

34. Brown MB, Benedetti JK. Sampling behavior of tests for correlation in two-way contingency tables. J. Amer. Stat. Assoc. 1977;72:309–15.

35. Rubenstein LM, Davis CS. Estimation of the average correlation coefficient for stratified bivariate data. Stat. Med. 1999;18:567–80.

36. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature. 2015;518:350–4.

37. Selvaraj S, R Dixon J, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol. 2013;31:1111–8.

Figure 1. An illustration example. (A) Hi-C contact maps of the biological replicates of hESC and IMR90. (B) Relationship between genomic distance and the average contact frequency for a hESC sample and an IMR90 sample. Data is from chromosome 22: 32000000 – 40000000.

.

Figure 2. A schematic representation of our method.

Figure 3. Discrimination of pseudo replicates (PR), biological replicates (BR) and non-replicates (NR). (A) Reproducibility scores for the illustration example (hESC and IMR90 cell lines) in Figure 1. Red dots are the results in the original samples, and blue dots are the results after equalizing the sequencing depth in all samples. (B-C) Reproducibility scores for the BR and NR in the ENCODE 11 cancer cell lines. The triangle represents the score for a BR and the boxplot represents the distribution of the scores for NRs. (B) Reproducibility scores for BRs and NRs in all cell types. (C) SCC for BRs and the corresponding NRs in each cell type. From left to right, the cell lines are ordered according to the average sequencing depths of the biological replicates.

Figure 4. Estimating interrelationship between the ten samples in the human H1 ESC lineage. (A) The lineage relationship between the ES cell and its five derived cells based on previous analysis from gene expression data and A/B compartments in Hi-C data in [17, 18]. (B-D) Estimated interrelationship based on the pairwise similarity score calculated using (B) SCC (C) Pearson correlation and (D) Spearman correlation. Heatmaps show the similarity scores. Dendrograms are resulted from a hierarchical clustering analysis based on the similarity scores. For easy visualization, the cell lines in the heatmaps are ordered according to their known distances to ES cells in (A). A decreasing trend of scores is expected from left to right (from bottom to top, respectively) if the estimated interrelationship agrees with the known lineage.

Figure 5. Estimated interrelationship for fourteen human primary tissues and two cell lines in [29]. The dendrograms are resulted from a hierarchical clustering analysis based on the pairwise similarity calculated using (A) SCC, (B) Pearson correlation and (C) Spearman correlation.

Figure 6. Estimated similarity between the human H1 ES cell and its derived cells at different resolutions. (A) SCC, (B) Pearson correlation coefficient, and (C) Spearman correlation coefficient.

Figure 7. Detecting the change of reproducibility due to sequencing depth using SCC. (A) SCC of downsampled biological replicates (25%, 50%, 75%, 100% of the original sequencing depth) for the five cell lines on the H1 ES cell lineage. (B) Saturation curves of SCC for datasets with different coverages. Plotted is the SCC at different subsamples (10%-90%) of the original samples with 90% confidence intervals. The blue dots represent H1 human ESC data (original sequencing depth=500M). The red dots represent the A549 data (original sequencing depth=30M).

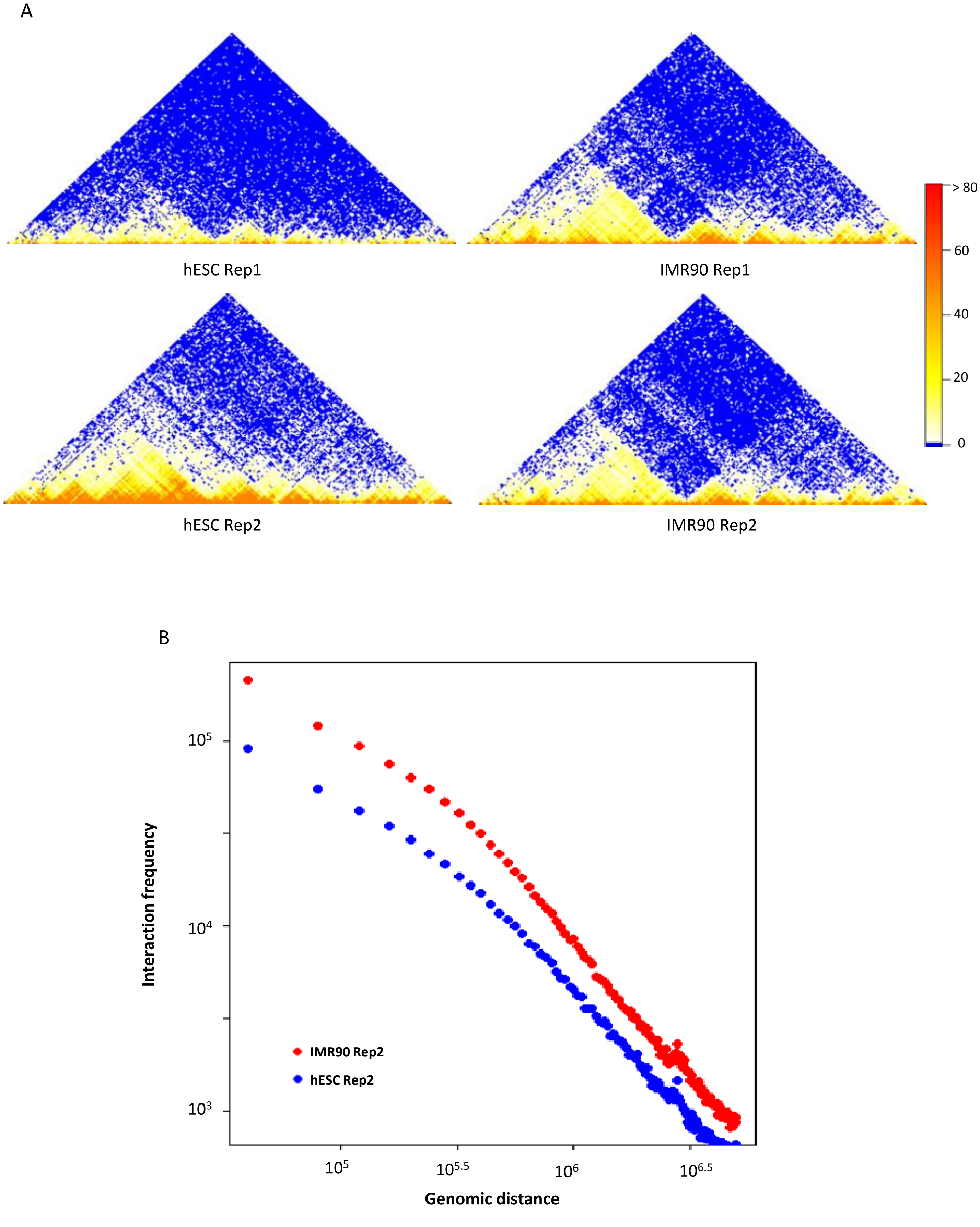
A
hESC Rep1
IMR90 Rep1
hESC Rep2
IMR90 Rep2
> 80
60
40
20
0
B
Interaction frequency
10^5
10^4
10^3
IMR90 Rep2
hESC Rep2
10^5
$10^{5.5}$
10^6
$10^{6.5}$
Genomic distance

Figure 1

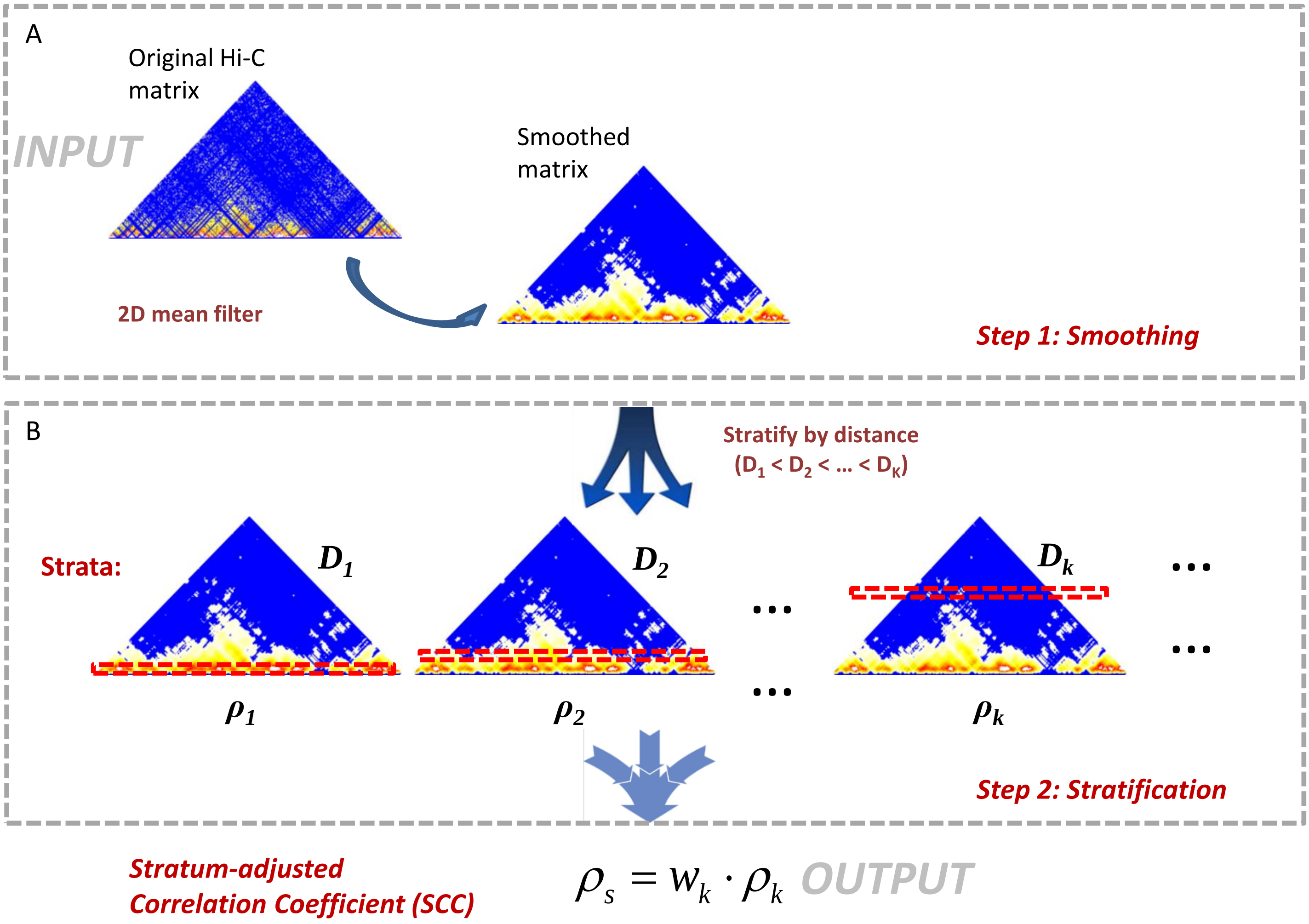
A
Original Hi-C matrix
INPUT
Smoothed matrix
2D mean filter
Step 1: Smoothing
B
Stratify by distance
($D_1 < D_2 < … < D_K$)
Strata:
D_1
D_2
D_k
…
ρ_1
ρ_2
ρ_k
Step 2: Stratification
Stratum-adjusted Correlation Coefficient (SCC)
$\rho_s = w_k \cdot \rho_k$
OUTPUT

Figure 2

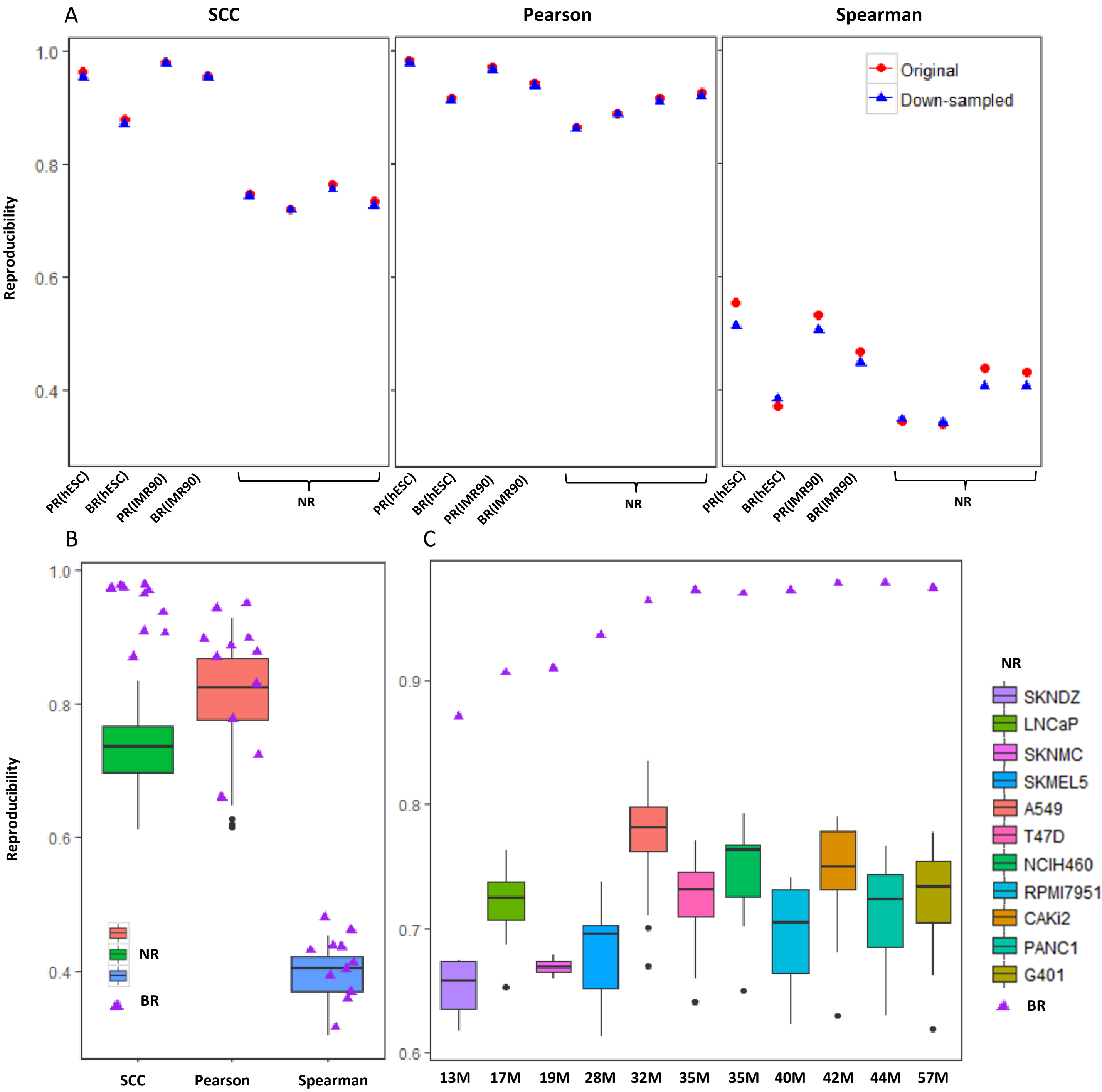
A
SCC
Pearson
Spearman
Reproducibility
1.0
0.8
0.6
0.4
Original
Down-sampled
PR(hESC)
BR(hESC)
PR(IMR90)
BR(IMR90)
NR
B
C
0.4
0.6
0.8
1.0
NR
BR
SCC
Pearson
Spearman
0.6
0.7
0.8
0.9
13M
17M
19M
28M
32M
35M
35M
40M
42M
44M
57M
NR
SKNDZ
LNCaP
SKNMC
SKMEL5
A549
T47D
NCIH460
RPMI7951
CAKi2
PANC1
G401
BR

Figure 3

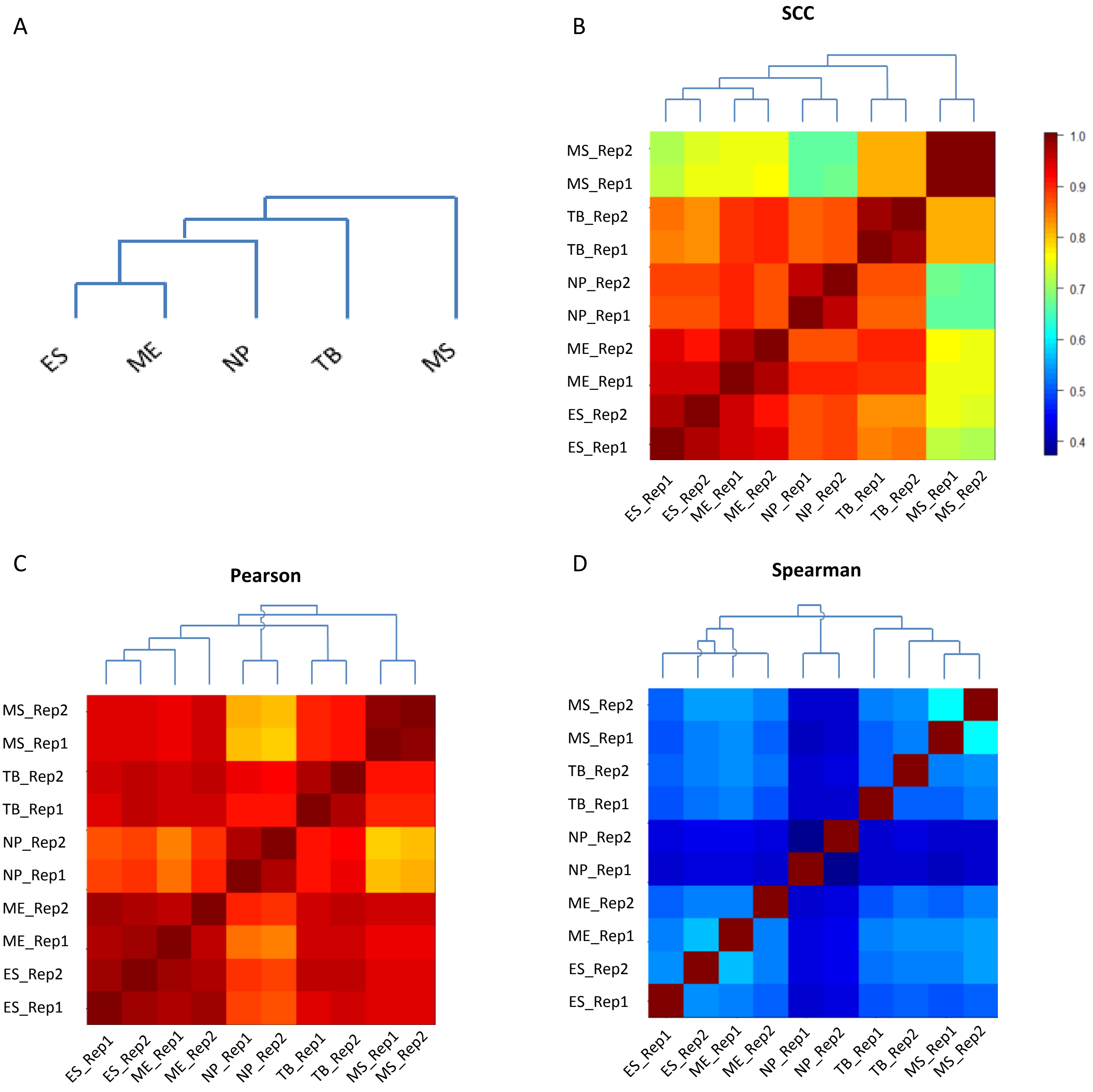
A
ES
ME
NP
TB
MS
B
SCC
MS_Rep2
MS_Rep1
TB_Rep2
TB_Rep1
NP_Rep2
NP_Rep1
ME_Rep2
ME_Rep1
ES_Rep2
ES_Rep1
ES_Rep1
ES_Rep2
ME_Rep1
ME_Rep2
NP_Rep1
NP_Rep2
TB_Rep1
TB_Rep2
MS_Rep1
MS_Rep2
1.0
0.9
0.8
0.7
0.6
0.5
0.4
C
Pearson
MS_Rep2
MS_Rep1
TB_Rep2
TB_Rep1
NP_Rep2
NP_Rep1
ME_Rep2
ME_Rep1
ES_Rep2
ES_Rep1
ES_Rep1
ES_Rep2
ME_Rep1
ME_Rep2
NP_Rep1
NP_Rep2
TB_Rep1
TB_Rep2
MS_Rep1
MS_Rep2
D
Spearman
MS_Rep2
MS_Rep1
TB_Rep2
TB_Rep1
NP_Rep2
NP_Rep1
ME_Rep2
ME_Rep1
ES_Rep2
ES_Rep1
ES_Rep1
ES_Rep2
ME_Rep1
ME_Rep2
NP_Rep1
NP_Rep2
TB_Rep1
TB_Rep2
MS_Rep1
MS_Rep2

Figure 4

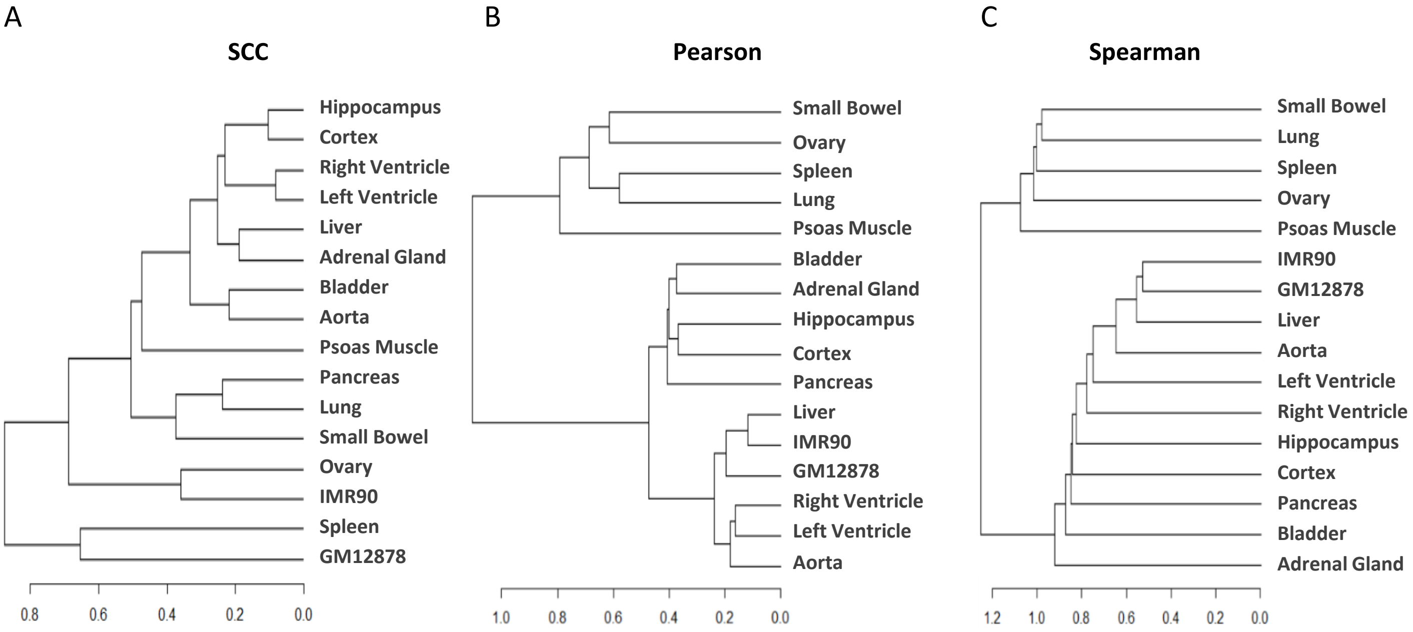
A
SCC
Hippocampus
Cortex
Right Ventricle
Left Ventricle
Liver
Adrenal Gland
Bladder
Aorta
Psoas Muscle
Pancreas
Lung
Small Bowel
Ovary
IMR90
Spleen
GM12878
0.8
0.6
0.4
0.2
0.0
B
Pearson
Small Bowel
Ovary
Spleen
Lung
Psoas Muscle
Bladder
Adrenal Gland
Hippocampus
Cortex
Pancreas
Liver
IMR90
GM12878
Right Ventricle
Left Ventricle
Aorta
1.0
0.8
0.6
0.4
0.2
0.0
C
Spearman
Small Bowel
Lung
Spleen
Ovary
Psoas Muscle
IMR90
GM12878
Liver
Aorta
Left Ventricle
Right Ventricle
Hippocampus
Cortex
Pancreas
Bladder
Adrenal Gland
1.2
1.0
0.8
0.6
0.4
0.2
0.0

Figure 5

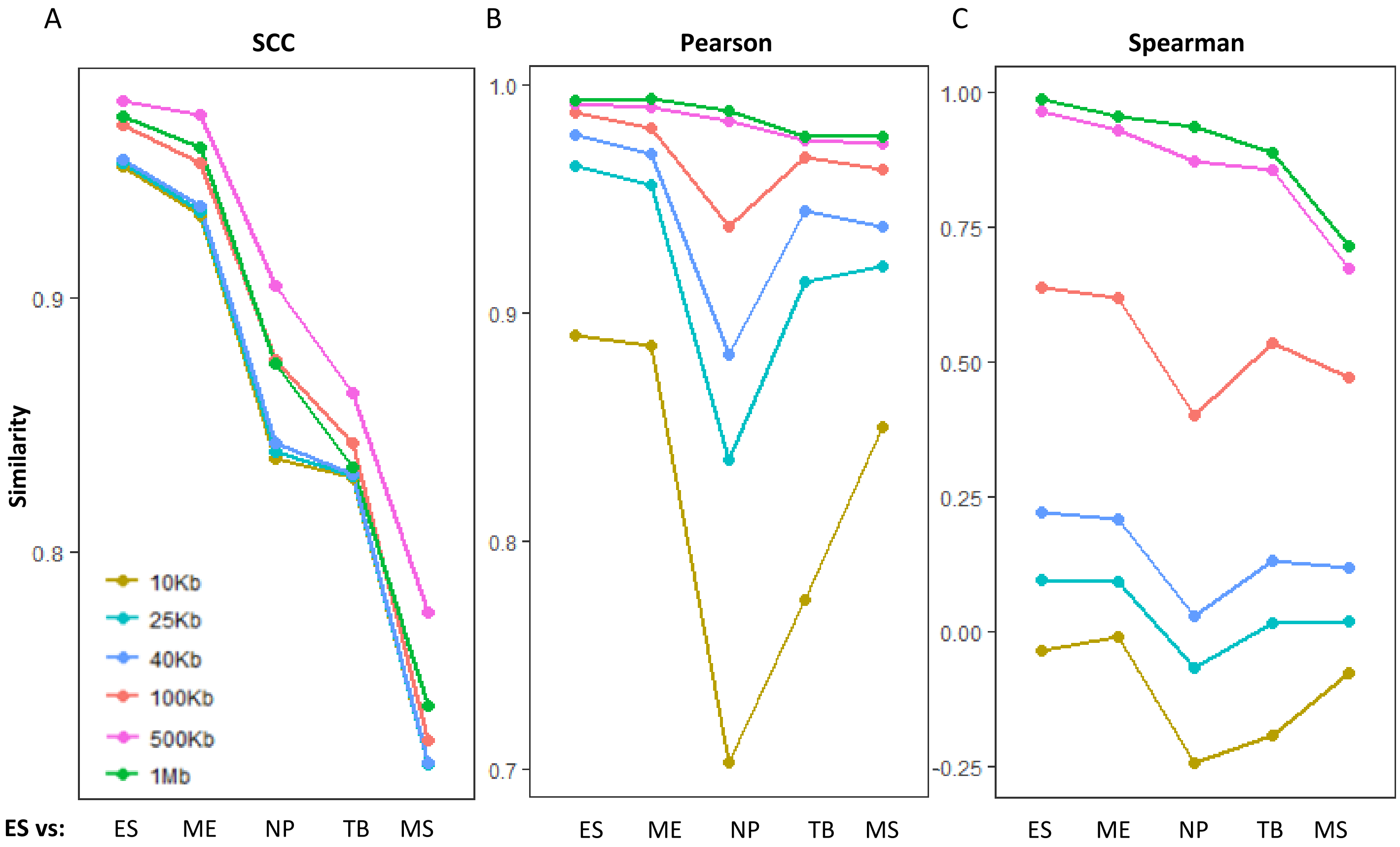
A
SCC
B
Pearson
C
Spearman
Similarity
0.9
0.8
10Kb
25Kb
40Kb
100Kb
500Kb
1Mb
ES vs:
ES
ME
NP
TB
MS
1.0
0.9
0.8
0.7
ES
ME
NP
TB
MS
1.00
0.75
0.50
0.25
0.00
-0.25
ES
ME
NP
TB
MS

Figure 6

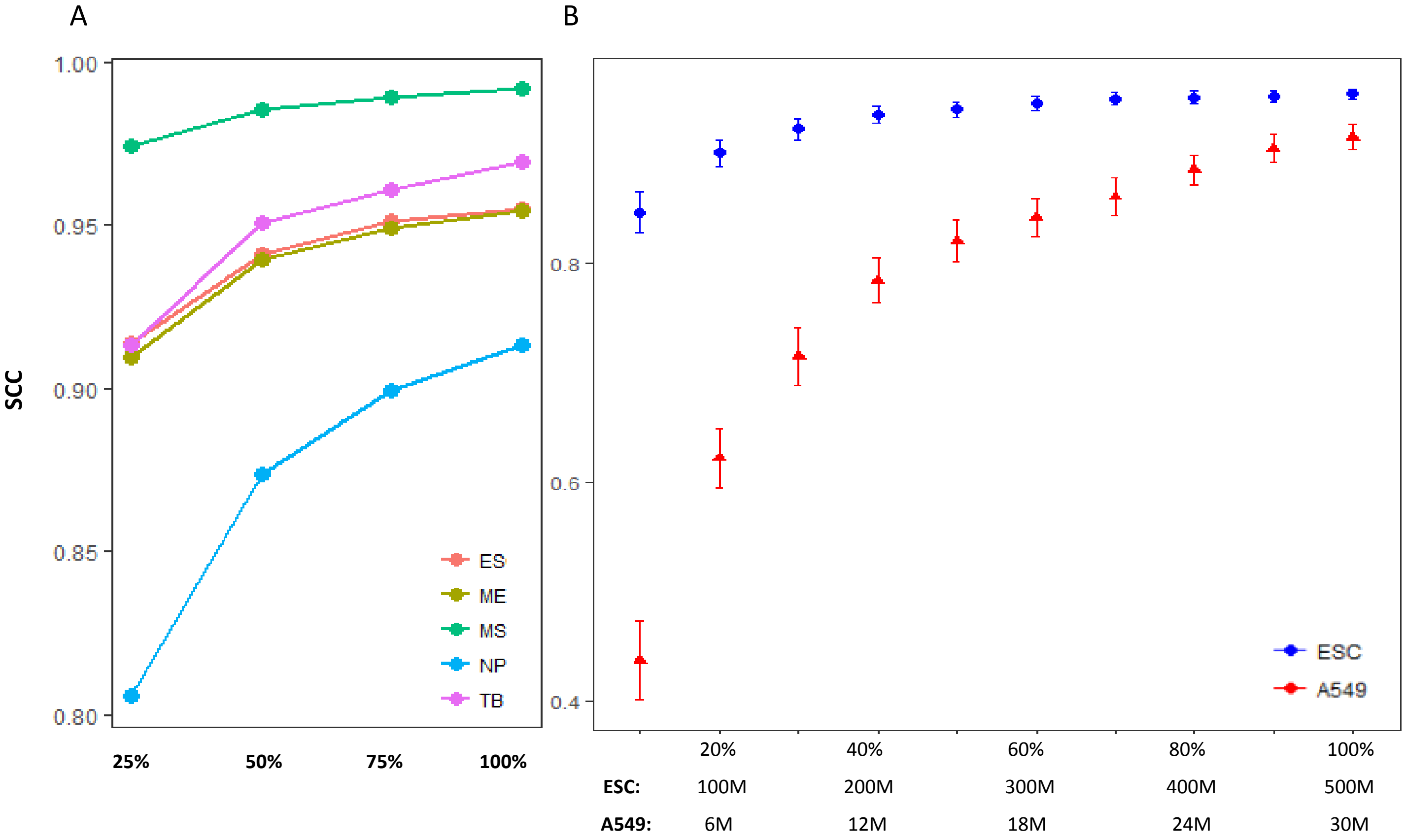
A
B
SCC
1.00
0.95
0.90
0.85
0.80
25%
50%
75%
100%
ES
ME
MS
NP
TB
0.8
0.6
0.4
20%
40%
60%
80%
100%
ESC:
100M
200M
300M
400M
500M
A549:
6M
12M
18M
24M
30M
ESC
A549

Figure 7