Measuring the reproducibility and quality of Hi-C data

Galip Gürkan Yardımcı, Hakan Ozadam, Michael E.G. Sauria, Oana Ursu, Koon-Kiu Yan, Tao Yang, Abhijit Chakraborty, Arya Kaul, Bryan R. Lajoie, Fan Song, Ye Zhang, Ferhat Ay, Mark Gerstein, Anshul Kundaje, Qunhua Li, James Taylor, Feng Yue, Job Dekker, William S. Noble

Abstract

Hi-C is currently the most widely used assay to investigate the 3D organization of the genome and to study its role in gene regulation, DNA replication, and disease. However, Hi-C experiments are costly to perform and involve multiple complex experimental steps; thus, accurate methods for measuring the quality and reproducibility of Hi-C data are essential to determine whether the output should be used further in a study. Using real and simulated data, we profile the performance of several recently proposed methods for assessing reproducibility of population Hi-C data, including HiCRep, GenomeDISCO, HiC-Spector and QuASAR-Rep. By explicitly controlling noise and sparsity through simulations, we demonstrate the deficiencies of performing simple correlation analysis on pairs of matrices, and we show that methods developed specifically for Hi-C data produce better measures of reproducibility. We also show how to use established (e.g., ratio of intra to interchromosomal interactions) and novel (e.g., QuASAR-QC) measures to identify low quality experiments. In this work, we assess reproducibility and quality measures by varying sequencing depth, resolution and noise

levels in Hi-C data from 13 cell lines, with two biological replicates each, as well as 176 simulated matrices. Through this extensive validation and benchmarking of Hi-C data, we describe best practices for reproducibility and quality assessment of Hi-C experiments. We make all software publicly available at http://github.com/kundaje/3DChromatin_ReplicateQC to facilitate adoption in the community.

Introduction

The Hi-C assay couples chromosome conformation capture (3C) with next-generation sequencing, making it possible to profile the three-dimensional structure of chromatin in a genome-wide fashion [1]. Recently, application of the Hi-C assay has allowed researchers to profile the 3D genome during important biological processes such as cellular differentiation [2, 3], X inactivation [4-6] and cell division [7]; and to identify hallmarks of 3D organization of chromatin, such as compartments [1], topologically associating domains (TADs) [8, 9] , and DNA loops [10]. Because the Hi-C assay measures the 3D conformation of a genome in the form of pairs of mapped reads ("interactions") connecting different loci, many such pairs are required to adequately characterize all pairwise interactions across a complete genome [10-12]. Consequently, the Hi-C assay can be costly to run. It is thus essential to have accurate and robust methods to evaluate the quality and reproducibility of Hi-C experiments, both to ensure the validity of scientific conclusions drawn from the data and to indicate when an experiment should be repeated or sequenced more deeply. Reproducibility measures

are also important for deciding whether two replicates can be pooled, a strategy that is frequently used to obtain a large number of Hi-C interactions [10].

A rich collection of literature for assessing the quality and reproducibility of a large collection of next generation sequencing based genomics assays, such as ChIP-seq [13] and DNase-seq [14], has been complied over the past decade [15-17]. For these assays, enrichment of signal ("peaks") at loci of interest [18] and assay-specific properties of sequencing fragments have been used as indicators of the quality of an experiment [15]. Correlation coefficient [19-21] and statistical methods such as the irreproducible discovery rate (IDR) [16] have been used to measure the reproducibility of such assays. However, all of these methods are designed to operate on data that is laid out in one dimension along the genome. Furthermore, unlike other functional genomics assays, Hi-C data must be analyzed at an effective resolution determined by the user [12, 22, 23]. For these reasons, existing methods for assessing genomic data quality and reproducibility are not directly applicable to Hi-C data.

A variety of methods have been used previously to measure the quality and reproducibility of Hi-C experiments Ad hoc measures include using, for reproducibility, the Pearson or Spearman correlation coefficient [2, 24-26] and, for data quality, statistics that describe the properties of Hi-C fragment pairs [1, 27]. The drawbacks of using correlation as a reproducibility measure for genomics experiments, both because of its susceptibility to outliers and because it implicitly treats all elements of the Hi-C matrix as independent measurements, has been documented [15, 28]. In practice, because most of the Hi-C signal arises from interactions between loci less than 1 Mb apart [22, 23], the correlation coefficient will be dominated by these short range

interactions. To alleviate such problems, distance based stratification [29] and dimensionality reduction of Hi-C signal [30], prior to measuring the correlation, have been proposed. Conversely, simple mapping statistics may be used to indicate a high or low percent of invalid or artefactual Hi-C fragments [23, 31], but such statistics reflect only the mapping stage of the analysis and cannot be immediately combined into a robust quality score.

To overcome these problems, members of the ENCODE Consortium have recently developed methods for assessing both the quality and the reproducibility of the Hi-C assay [32-34]. In this study, we used large sets of real and simulated Hi-C data to assess and compare the performance of methods for measuring the reproducibility of Hi-C data and evaluating Hi-C data quality. We generated multiple benchmarks for testing the performance of reproducibility measures and established that all of these methods can accurately measure reproducibility of Hi-C data, whereas correlation coefficient cannot. Similarly, we have used real and simulated datasets to profile the performance of quality control methods and compared these methods to established statistics that have been used as indicators of high quality Hi-C experiments. Here, we offer a thorough assessment of quality control and reproducibility methods and describe best practices for analyzing the quality and reproducibility of Hi-C data.

Results

Experimental and simulated Hi-C datasets for performance evaluation

We performed two replicate Hi-C experiments on cells from 11 immortalized human cancer cell lines from a variety of tissues and lineages (Supp. Table. 1). After aligning and filtering of paired end sequencing reads, we obtain 10 to 61 million paired reads per experiment. These Hi-C interactions serve as a readout of three dimensional proximity of the corresponding genomic loci. The interactions are binned into fixed-sized bins, and a count of the number of Hi-C interactions that connect each pair of bins is stored in a Hi-C contact matrix. Unless otherwise noted, we used 40 kilobase (kb) bins because this value achieves reasonable sparsity of the Hi-C contact matrices, based on the depth of sequencing of the data sets used in our study. Also, this resolution has been adopted in multiple previous studies [7, 8]. We use the resulting Hi-C matrices as input to every reproducibility and quality control analysis in this study, except where indicated.

For use in assessing reproducibility and quality measures for Hi-C data, we designed a model for simulating noisy Hi-C experiments (Fig. 1A). Our noise model aims to simulate a contact matrix from a Hi-C experiment performed on chromatin that lacks any high order structure, such as loops and topologically associating domains. For this purpose, our simulation models two main phenomena: the “genomic distance effect,” i.e., the higher prevalence of crosslinks between genomic loci that are close together along the genome [1], and random ligations generated by the Hi-C protocol [23]. For the

first phenomenon, we use real Hi-C data, and we sample from the empirical marginal distribution of counts as a function of genomic distance. The second phenomenon, random ligation noise, is modeled by generating Hi-C interactions between random bin pairs (see Methods for details). Counts generated by these two "noise" components of the model can be mixed with different proportions to produce simulated "pure noise" Hi-C matrices. We then mix the simulated contacts with experimental contact matrices in varying proportions to obtain noise injected matrices.

In addition to noise, we tested the effects of sparsity and the resolution of Hi-C matrices on the performance of each method. We profiled the effects of sparsity explicitly by downsampling real Hi-C matrices to contain a set of fixed total number of intra-chromosomal Hi-C interactions. Binning resolution further controls the sparsity of a Hi-C matrix, at the same time dictating the scale of chromatin organization that can be observed in a Hi-C matrix. By binning deeply sequenced Hi-C datasets containing at least 400 million intrachromosomal Hi-C interactions from two cell types, we generated Hi-C matrices binned at high, mid and low resolutions (10 kb, 40 kb, 500 kb) and used these to investigate the effect of resolution on each method as well (Supp. Table. 1). A schematic of the full range of datasets used in this study to validate each method is shown in Fig.1B.

Measures for quality and reproducibility of Hi-C data

Four recently developed methods for measuring the quality of and reproducibility of Hi-C experiments were assessed in this study (Fig. 1C). HiCRep [33], GenomeDISCO [34], HiC-Spector [32] and QuASAR-Rep (in submission) measure reproducibility, and

QuASAR-QC measures quality of Hi-C data. The four reproducibility methods we evaluate employ a variety of transformations of the Hi-C contact matrix. HiCRep stratifies a smoothed Hi-C contact matrix according to genomic distance and then measures the weighted similarity of two Hi-C contact matrices at each stratum. In this way, HiCRep explicitly corrects for the genomic distance effect and addresses the sparsity of contact matrices through stratification and smoothing, respectively. GenomeDISCO uses random walks on the network defined by the Hi-C contact map to perform data smoothing before computing similarity. The resulting score is sensitive to both differences in 3D DNA structure and differences in the genomic distance effect [34], and makes it thus more challenging for two contact maps to be reproducible, as they have to satisfy both criteria to be deemed similar. HiC-Spector transforms the Hi-C contact map to a Laplacian matrix and then summarizes the Laplacian by matrix decomposition. QuASAR calculates the interaction correlation matrix, weighted by interaction enrichment. The two variants of QuASAR, QuASAR-QC and QuASAR-Rep, both assume that spatially close regions of the genome will establish similar contacts across the genome, and they measure quality and reproducibility, respectively, by testing the validity of this assumption for a single and pair of replicates.

Reproducibility measures correctly rank noise-injected datasets

To assess the performance of the reproducibility measures, we simulated pairs of Hi-C matrices with varying noise levels. Intuitively, a good reproducibility measure should declare the least noisy replicate pair as most reproducible and the noisiest replicate pair as least reproducible. We paired a real Hi-C contact matrix with a noisier version of the same matrix using a wide range of simulated noise levels (5%, 10%, 15%, 20%, 30%, 40% and 50%). This procedure yielded seven pairs of replicates for each of 11 different cell types. We performed this approach using two different sets of randomly generated noise matrices, using one-third genomic distance noise and two-thirds random ligation noise or vice versa. Each replicate pair was assigned a reproducibility measure by HiCRep, GenomeDISCO, HiC-Spector, QuASAR-Rep and Spearman correlation.

Our analysis showed that all reproducibility measures were able to correctly rank the simulated datasets. Averaged over 11 different cell types, we observed a monotonic trend for all of these measures (Fig. 2A). Indeed, for every cell type and every measure, increasing the noise level always led to a decrease in estimated reproducibility (Supp. Fig. 1).

Comparing the two noise models, we saw less consistent trends. HiC-Spector assigned higher reproducibility scores to matrices with 66% genomic distance noise and 33% random ligation noise. Spearman correlation and GenomeDISCO showed the opposite behavior whereas QuASAR-Rep and HiCRep gave similar scores regardless of the underlying noise proportions. This variability suggests that the various reproducibility measures exhibit different sensitivities to different sources of noise, thus potentially yielding complementary assessments of reproducibility.

Assessment using real data sets reveals differences among reproducibility measures

Inevitably, any simulation approach is only as good as its underlying assumptions; thus, we also analyzed the performance of the four reproducibility measures using real data. Specifically, we asked whether the reproducibility measures can discriminate between pairs of independent Hi-C experiments repeated on the same cell type versus pairs of experiments from different cell types. In this setup, we used three types of replicate pairs: matrices from the same cell type (which we call “biological replicates,” although each pair represents the same cells being prepped twice, rather than two different sets of cells), matrices from different cell types (non-replicates) and matrices sampled from combined biological replicates (pseudo-replicates).

Because pseudo-replicates are generated from pooled biological replicates, their variation solely stems from statistical sampling, with no biological (including distance effect) or technical variance. Therefore, we expect pseudo-replicates to exhibit the highest reproducibility. Conversely, non-replicate pairs are expected to have the lowest

degree of reproducibility, because they contain all the experimental variation observed in biological replicates, as well as cell type specific differences in 3D chromatin organization.

In contrast to the simulation analysis, the analysis using real datasets showed distinct differences among the five methods. For each of the eleven cell types and each reproducibility measure, we assigned reproducibility scores to a single biological replicate pair, 20 non-replicate pairs, and three pseudo-replicate pairs (Fig. 2B). The reproducibility score of a replicate pair is the score obtained by averaging reproducibility scores assigned to each chromosome. Strikingly, the Spearman correlation failed to separate biological replicates from non-replicates, whereas all four other measures succeeded (Supp. Fig. 2A). These differences are statistically significant according to a one-sided Kolmogorov-Smirnov test ($P < 0.01$). Intuitively, we prefer a measure that separates non-replicates from biological replicates with a clear margin. By this measure, the HiC-Spector measure yields the largest separation, followed by HiCRep, QuASAR-rep and GenomeDISCO (Fig. 2B). Among them, HiC-Spector and HiCRep correctly rank all replicates types for all eleven comparisons, with a clear separation between biological replicates and non-replicates. GenomeDISCO ranks a biological replicate lower than a non-replicate for a single case out of eleven. The pair of biological replicates that GenomeDISCO ranks lower than non-replicates shows a marked difference in genomic distance effect (Supp Fig 3), to which this method is sensitive [34]. QuASAR-rep is able to correctly rank most replicate types. The one exception is SKNDZ, where the biological replicate and pseudo-replicate pairs receive very similar scores. This phenomenon likely occurs because of the extremely low sequencing depth

in this sample pair (approximately 15 and 9 million of intra-chromosomal Hi-C interactions). Indeed, we also observe that all of the non-replicate pairs that receive very low scores (<0.2) involve SKNDZ. As expected, the Spearman correlation performs worse than the Hi-C-specific measures, ranking non-replicates higher than biological replicates in eight cases.

Pseudo-replicate reproducibility scores provide an upper bound for each reproducibility measure. In general, these scores show similar trends to those described above. For example, the Spearman correlation scores assigned to pseudo-replicates show a wide separation from the rest of the scores, even though non-replicates and biological replicates are intermingled. On the other hand, GenomeDISCO, HiC-Spector, HiCRep, and QuASAR-rep show the desired behaviour: a high degree of separation between non-replicates and biological replicates, and a relatively small separation between biological replicates and pseudo-replicates.

Reproducibility can be determined over a range of experimental coverage

To directly investigate the effects of the coverage of a Hi-C experiment on the reproducibility measures, we downsampled real Hi-C matrices to contain fewer interactions and examined the effects on the resulting reproducibility scores. We limited this analysis to real data from six cell types with higher coverage, and we subsampled each replicate multiple times to contain 5 to 30 million total Hi-C interactions (see Methods for details). These datasets were used for testing the ability of each method to

distinguish among different replicate types at lower coverage levels, and for explicitly profiling the dependence of reproducibility scores on coverage levels.

All four Hi-C reproducibility measures retained their ability to distinguish between replicate types, even at extremely low coverage levels. Visualization of the reproducibility scores revealed that all four measures successfully separate non-replicates from biological replicates even with only five million Hi-C interactions, a feat that Spearman correlation cannot achieve at even the highest coverage level (Fig. 2C). As before, pseudo-replicate pairs continue to serve as an upper bound for reproducibility measures. However, the separation between pseudo-replicates and biological replicates is reduced at lower coverage levels, and so is the separation between biological replicates and non-replicates. Furthermore, this analysis suggests we can infer empirical thresholds for these reproducibility measures that can effectively separate all biological replicates from non-replicates at a given coverage levels, as explained in methods section. These empirical thresholds, selected as the midpoint between the most reproducible non-replicate pair and the least reproducible replicate pair, are shown as dashed lines in Fig. 2C and can be found in Supp. Table 2.

Consistent with the trends observed in the analysis of real datasets, the reproducibility of downsampled replicate pairs exhibits a dependence on sequencing depth. We observe that reproducibility scores associated with biological replicates become significantly smaller as coverage decreases, according to a one-sided Wilcoxon signed rank test ($P < 0.05$, Supp. Fig. 4). The HiCRep, GenomeDISCO, QuASAR-rep and Spearman correlation scores exhibit a statistically significant drop for every level of coverage. In contrast, reproducibility scores from HiC-Spector only start to significantly

decay below 20 $\times 10^6$ interactions, exhibiting a lesser degree of dependence on the coverage level. This may be because the leading eigenvectors used by HiC-Spector tend to capture local or mesoscopic structures, which are less likely to be affected by coverage. Despite varying levels of dependence on coverage, downsampling analysis convincingly shows that all measures exhibit a dependence on coverage. Thus, coverage of different replicate pairs must be factored into reproducibility analyses, especially for comparative purposes.

Reproducibility measures are robust to changes in resolution

The resolution of a Hi-C matrix effectively dictates the scale of 3D organization observable from the data: a low resolution matrix can only reveal compartments and TADs[1, 8], whereas high resolution matrices reveal additional finer scale structures like chromatin loops [10]. To investigate the effect of resolution on reproducibility, we used deeply sequenced Hi-C replicates with at least 400 million intra-chromosomal interactions generated from the HepG2 and HeLa cell lines. From these data, we generated real and simulated replicate pairs at 10kb, 40kb and 500kb resolution, and we measured the reproducibility of each replicate pair.

HiCRep, GenomeDISCO, HiC-Spector and QuASAR-Rep accurately measure reproducibility at both high and low resolutions, whereas Spearman correlation performance is dependent on resolution. The four Hi-C-specific methods can correctly rank pseudo, biological and non-replicate pairs at 10kb, 40kb and 500kb resolutions (Fig. 3A) with a clear margin between biological replicate and non-replicate pairs; however, Spearman correlation fails at this task for data binned at 10kb resolution for

biological replicate data sets obtained from HepG2. Notably, the reproducibility scores from the four methods are largely independent of resolution. While GenomeDISCO and QuASAR-rep exhibit some dependence of resolution, assigning lower reproducibility scores to replicates with lower coverage, they maintain a clear boundary between biological and non-replicates at all resolutions. However, the Spearman correlation exhibits the highest degree of dependence to resolution and fails to maintain such boundaries. Simulated datasets further validate that reproducibility scores from each method decrease with increasing levels of noise at 10kb, 40kb and 500kb resolution (Fig. 3B)

Next, we used deeply sequenced datasets to further investigate the effect of coverage on reproducibility scores of biological replicates at three resolution levels using a wider range of coverage values (30, 60, 120, 240, and 400 million intra-chromosomal interactions). For HiCRep, QuASAR-rep and GenomeDISCO, we observed that reproducibility scores tend to plateau at 240 million interactions at 10kb and 40kb resolutions, whereas reproducibility scores of 500kb resolution matrices benefit little from higher coverage (Fig. 3C). Consistent with our previous observations, HiC-Spector exhibits a lower degree of dependence on coverage, with scores reaching maxima at 120kb. Spearman correlation exhibits different trends at different resolutions, underscoring its unsuitability to the task. Overall, the four Hi-C reproducibility measures exhibit robustness to coverage and resolution differences, as measured by their ability to distinguish between replicate and non-replicate pairs.

Noise reduces the consistency and the prevalence of higher order structures in Hi-C matrices

Having investigated four different methods for evaluating the reproducibility of a given pair of Hi-C matrices, we now focus on methods for evaluating the quality of a single Hi-C matrix. As before, we perform this evaluation by injecting noise into real Hi-C data, producing a collection of 88 matrices corresponding to 11 cell types and 8 different noise profiles (see Methods). Among our four Hi-C reproducibility measures, only one (QuASAR-QC) provides a variant to assess the quality of a single matrix. The procedure yields a single, bounded summary statistic indicative of homogeneity of the underlying sample population and the signal-to-noise ratio of the interaction map. In addition to QuASAR-QC analysis, we profiled two well-known features of 3D organization: statistically significant long range contacts [35, 36], which include DNA loops, and topologically associating domains (TADs). Intuitively, we expect that significant contacts and TADs should be harder to detect in noisy matrices, and that such matrices should have a lower degree of consistency.

Our analysis suggests that QuASAR-QC is indeed sensitive to the noise and the coverage of a Hi-C matrix. For each simulated Hi-C matrix from 11 cell types, QuASAR-QC detects a perfectly monotonic relationship between the noise level and the consistency of the matrix (Fig. 4A). The same trend is observed in deeply sequenced HepG2 and HeLa cell types at 10kb, 40kb and 500kb resolutions (Supp. Fig 5). Although the majority of noise-free combined replicates are assigned a QuASAR-QC score ranging from 0.23 to 0.30, three cell types have strikingly lower QuASAR-QC

scores ranging from 0.12 and 0.6. The Hi-C matrices from these three cell types (LNCaP, SKNDZ, SKNMC) contain fewer Hi-C interactions. Thus, the lower consistency scores are likely due to the sparsity that results from low experimental coverage (Supp. Table 1). Furthermore, investigation of contact probabilities at given genomic distances for each cell type revealed that the three cell types with lower QuASAR-QC scores have significantly higher contact probabilities at genomic distances larger than 50 megabases (Supp. Fig. 6). Because such long range contacts are unlikely to occur due to the organization of chromatin, it is likely that such long range contacts represent random ligation of uncrosslinked DNA fragments, which is a known source of noise in a Hi-C experiment [23]. Thus, the QuASAR-QC measure is potentially sensitive to both the level of simulated noise and the differences in level of inherent noise that each combined replicate contains.

Statistically significant mid-range (50kb-10Mb) interactions are depleted in noisy Hi-C matrices. We identified statistically significant Hi-C contacts using Fit-Hi-C [35] for each of the Hi-C matrices that make up our simulated dataset. Because robust identification of such contacts requires deeply sequenced datasets that contain large numbers of Hi-C interactions, we chose to use a somewhat liberal false discovery rate threshold of 0.05 to facilitate discovery of statistically significant contacts. For 11 cell types, we observed that eight out of eleven cell types exhibit a perfect or near perfect anti-correlation between the injected noise percentage and the total number of significant interactions (Fig. 4B). For the other three cell lines (LNCaP, SKNDZ, SKNMC), Fit-Hi-C identifies almost no significant contacts with or without any noise injection, further supporting the conclusion that these Hi-C data sets have low quality. These three cell

lines are also the cell lines that have the lowest QuASAR-QC scores, corroborating the results between these two independent analyses. For the deeply sequenced two data sets (HepG2 and HeLa), we observed a similar trend at both 10kb and 40kb resolutions, with a higher number of significant mid-range contacts due to the higher coverage, as expected (Supp. Fig. 7).

Surprisingly, we found that topologically associating domain detection is highly robust to noise. We identified TADs using the insulation score [5, 37] method for the 88 simulated matrices, and we characterized the changes in total number of TADs and TAD size distribution and the changes to TAD boundaries with respect to noise level. The total number of identified TADs and their size distribution are only altered at the highest level of noise injection (Supp Figure 8 and 9). In addition, TAD boundaries between the original replicate and noise injected levels exhibit the same degree of variation between two biological replicates, further supporting the idea that TAD boundaries identified with the insulation score approach are highly robust to noise (Fig. 4C, Supp Fig. 10).

Quality control measures require different levels of experimental coverage

Continuing our assessment of Hi-C quality measures, we used downsampled Hi-C matrices to investigate the relationship between experimental coverage and each QC measure using a similar setup as before (see Methods).

Quality control metrics exhibit a predictable dependence on the coverage of Hi-C matrices. For each of the six cell types we downsampled, we observed that QuASAR-QC scores are lower for Hi-C matrices with fewer interactions (Fig. 4D). We observe the

same trend for deeply sequenced matrices at 10kb and 40kb resolutions; however, QuASAR-QC scores at 500kb tend to benefit less from deeper coverage, likely because coarse resolutions do not require large numbers of Hi-C interactions. (Supp. Fig. 11). Similarly, the number of statistically significant long range interactions also decreases as we reduce the number of total Hi-C interactions. However, the number of significant interactions decrease at a much higher rate: even at 15 million interactions most cell lines lose the majority of significant interactions (Fig. 4E). Larger numbers of significant interactions are detected in deeply sequenced datasets, due to added statistical power, but a similar relationship between coverage and number of significant contacts is observed at both 10kb and 40kb resolutions (Supp. Fig. 12). Conversely, we found that TADs detected by insulation score are robust to low coverage levels. Using the same approach for noise-injected datasets, we found that total number of TADs and their size distribution are not altered by lower coverage (Supp Figures 13 and 14). Indeed, the distances between TAD boundaries identified at lower coverage and original replicates only differ from the baseline distribution at 10 million or fewer interactions (Fig. 4F, Supp Fig15).

Quality control measures are consistent with mapping statistics

To further validate the performance of the quality controls measures at our disposal, we investigated the relationship between the QuASAR-QC scores assigned to real Hi-C matrices and various read-mapping statistics that have been used previously to evaluate Hi-C data quality [23]. The three statistics we compared against are percentage of fragment pairs that can be mapped uniquely to the genome (aligned

pairs), percentage of fragment pairs from the same restriction fragments (invalid pairs), and the percentage of intrachromosomal interactions (intra-chromosomal percentage).

Overall, we observe varying degrees of correlation between the quality control measures and the mapping statistics for biological replicates. The percentage of aligned pairs is correlated with higher quality experiments, consistent with what one would intuitively expect from high quality sequencing libraries (Fig. 5A). The percentage of invalid pairs is also weakly anti-correlated with QuASAR-QC scores, consistent with the fact that invalid pairs represent uninformative Hi-C interactions (Fig. 5B). However, we observed the highest degree of correlation between QuASAR-QC scores and intra-chromosomal percentage (Fig. 5C). In a typical Hi-C experiment, a portion of inter-chromosomal interactions result from random ligation of non-crosslinked fragments; thus, a significant enrichment of inter-chromosomal interactions, which results in a depletion of intra chromosomal interactions, indicates a low quality Hi-C experiment. In particular, six biological replicates with lower than 30% intra-chromosomal interactions have the lowest QuASAR-QC scores; these replicates are from the LNCaP, SKNDZ, and SKNMC cell types. These replicates were also identified to have lower quality in our simulation studies (Fig. 5A) and are depleted for significant mid-range interactions, establishing the consistency of quality control measures overall. We note that this finding is consistent with the previously suggested range of 40-60% intra-chromosomal interactions for high quality experiments [23]. These trends can be reproduced when we include deeply sequenced Hi-C data, except for aligned pair percentage (Fig. 4D-F). The deeply sequenced datasets are generated by a four cutter enzyme (DpnII), which

presumably results in a different range of successful alignment percentage compared to HindIII.

Discussion

We evaluated recently proposed methods for measuring the quality and reproducibility of Hi-C experiments. Using a rich set of Hi-C experiments from a variety of human cell types, we tested whether these methods can identify reproducible and high quality experiments. Furthermore, we generated Hi-C contact matrices with controlled levels of noise by designing a simulated noise injection process. Our analysis shows that these measures perform well and improve upon the shortcomings of using generic or qualitative approaches.

The Hi-C reproducibility measures that we evaluated measure reproducibility more accurately than the Spearman correlation for real and simulated datasets. In particular, measures specifically designed for Hi-C data can better distinguish subtle differences in the 3D organization of different cell types, because these methods directly account for the specific noise properties of this data type [32, 33]. Although all four methods perform satisfactorily, we found that they exhibit different sensitivities to different sources of noise and different spatial patterns, thus potentially yielding complementary assessments of reproducibility. HiC-Spector, QuASAR-rep HiCRep exhibit similar performance across all tests. For example, previously reported analysis suggests that GenomeDISCO is particularly sensitive to differences in the genomic distance effect.[34]. Given that the reproducibility methods are successful at distinguishing

biological replicates from non-replicates, these methods can also be used to check for sample swaps during large-scale experiments.

The QuASAR-QC measure provides an interpretable score that can accurately rank simulated datasets according to noise levels and distinguish low quality real Hi-C experiments from high quality ones (in submission). This measure correlates with previously established statistics that indicate high quality in a Hi-C experiment and have been used as qualitative indicators of quality. Each of these statistics captures different sources of error in a Hi-C assay. In contrast, QuASAR-QC offers a single score that allows direct ranking of multiple experiments.

Significant mid-range interactions, such as DNA loops, are also depleted in low quality Hi-C experiments in both simulated and real datasets. Surprisingly, we found that TAD detection is fairly robust to all but high levels of noise, presumably because TAD detection only requires that a dataset contains a sufficient proportion of valid short range Hi-C interactions and ignores mid- and long-range interactions. Unfortunately, it is challenging to convert the enrichment of such features into a quality control measure, due to other quality-independent biological processes which can cause variation of these features. However, a near total depletion of these features, mid-range interactions in particular, may certainly indicate lower quality overall.

In practice, different Hi-C experiments are performed with very different levels of coverage, reflecting library complexity and the amount of sequencing performed. Hi-C data can be binned at high or low resolutions which, combined with resolution, determines the sparsity of the resulting Hi-C contact matrix [12, 22]. Only at high resolutions are fine-scale structures such as DNA loops detectable [10]. Our

experiments confirmed that sparser datasets yield lower reproducibility and quality scores. Interestingly, we observed that the four reproducibility measures can distinguish between replicates from the same cell type versus replicates from different cell types at high levels of sparsity and at both high and low resolutions. The robustness of reproducibility methods to differences in coverage and resolution shows that they can be applied to Hi-C datasets of smaller genomes where higher resolutions are more common and can assess reproducibility reliably for experiments that have not been deeply sequenced. By using the suggested empirical thresholds, experimenters can assess with relatively low sequencing depth whether a pair of Hi-C libraries yields data that is reproducible enough to warrant sequencing more deeply.

We release a software package that incorporates the four reproducibility measures and the QuASAR-QC measure (https://github.com/kundajelab/3DChromatin_ReplicateQC). Until recently, proven measures have been lacking and currently there is no standard for measuring for quality and reproducibility of Hi-C data. This tool will both greatly simplify the task of measuring both the quality and reproducibility of Hi-C datasets robustly by using the methods we show to be accurate in this study. We also propose a set of empirical quality and reproducibility thresholds for use at various coverage levels, which are built into the software package to make it easy to determine whether samples pass quality and reproducibility standards (Supp. Table 2).

While the methods we compared are tailored for Hi-C data, similar chromosome conformation capture assays such capture Hi-C [38, 39]and ChIA-PET [40] are used to study three dimensional interactions in the genome. These assays differ from Hi-C due to their targeted nature; however, they share many properties of Hi-C assay, such as

the genomic distance effect, and can be represented as a contact matrix similar to Hi-C [41]. Reproducibility and quality measures of these assays are lacking in general, raising the possibility of adaptation of the methods we evaluate here to these assays.

In summary, we show that recently proposed Hi-C quality and reproducibility measures accurately measure these qualities on a large collection of real and simulated data. By profiling various parameters of Hi-C contact matrices, we describe best practices for applying and interpreting these measures. We also make available a convenient software tool that simplifies the application of these measures to Hi-C datasets. We hope that adoption of this standard toolkit will help to improve the quality and reproducibility of Hi-C data generated in the future.

Methods

Measures of reproducibility

HiCRep. This method assesses reproducibility by taking into account two dominant spatial features of Hi-C data: distance dependence and domain structure. The method first smooths the given Hi-C matrices to help capture domain structures and reduce stochastic noise due to insufficient sampling. It then addresses the distance-dependence effect by stratifying Hi-C data according to genomic distance. Specifically, the method consists of two stages.

In the first stage, HiCRep smooths the Hi-C raw contact map using a 2D mean filter, which replaces the read count of each contact with the average counts of all contacts in its neighborhood. The neighborhood size is obtained from a deeply sequenced benchmark dataset using a training procedure. In this analysis, neighborhood size parameter of 20, 5, and 1 are

used for the resolutions of 10 kb, 40 kb, and 500 kb, respectively. Smoothing improves the contiguity of regions with elevated interaction, consequently enhancing the domain structures.

In the second stage, HiCRep takes into account the distance dependence effect by a stratification and aggregation strategy. This stage consists of two steps. The algorithm first stratifies the contacts according to the genomic distances of the contacting loci and computes the correlation coefficients within each stratum. HiCRep then assesses the reproducibility of the Hi-C matrix by applying a novel stratum-adjusted correlation coefficient statistic (SCC) to aggregate the stratum-specific correlation coefficients using a weighted average, with the weights derived from the Cochran-Mantel-Haenszel (CMH) statistic. The SCC has a range of [-1, 1] and is interpreted in a way similar to the standard correlation coefficient.

GenomeDISCO. This method focuses on two key aspects of contact maps: the need for smoothing, and the multiscale nature of these maps. The need for smoothing arises because contact maps are insufficiently sampled, especially at low sequencing depths. This means that a pair of genomic regions can exhibit a low count either from a lack of contact or from insufficient sampling. This problem is addressed by smoothing the data, essentially assuming that two contact maps are reproducible as long as they capture similar higher order structures, even if they differ in terms of individual contacts. GenomeDISCO investigates contact maps at multiple scales by comparing them at different levels of smoothing and computing a reproducibility score that takes all these comparisons into account.

The smoothing approach is based on random walks on networks. Each contact map is treated as a network, where each node is a genomic region and each edge is weighted by the Hi-C count matrix, following normalization. In this work, square root was used normalization, but similar results were obtained by using alternative normalization methods, including simple row- and column-based normalization or Knight-Ruiz normalization [42] (Data not shown). Random walk are performed on networks to smooth the data, asking for each pair of nodes what is the

probability of reaching node i from node j, if t steps are allowed in a random walk biased by the edge weights. The smoothed data can be computed by raising the adjacency matrix of our weighted network to the power t. Lower values of t perform local smoothing of the data, revealing structures such as domains, while larger values of t emphasize compartments. This graph-based smoothing scheme aims to preserve sharp domain boundaries that 2D methods may dilute.

To obtain the GenomeDISCO reproducibility score, each contact map is separately smoothed across a range of t values. For each value of t, the L1 distance (i.e., the sum of the absolute values in the difference matrix) between the two smoothed contact maps is computed and normalized by the average number of nodes with nonzero total counts across the two original contact maps compared. Afterwards, a combined distance between the two contact maps is obtained by computing the area under the curve of the L1 difference as a function of t. This allows us to consider multiple levels of smoothing and thus multiple scales when computing our scores. Finally, this distance is converted into a reproducibility score as follows:

$$\text{Reproducibility} = 1 - (\text{combined distance})$$

This score is in the range [-1,1], with higher scores representing higher reproducibility. This is because, for each node, the maximum L1 difference is 2, corresponding to the case when the node has mutually exclusive contacts in the two contact maps being compared. Thus, the combined distance lies in the range [0, 2], making the reproducibility score fall in the range [-1, 1].

Parameter optimization on an orthogonal dataset revealed the optimal t={3} [34], which was used in this study.

In all pairwise comparisons in this paper, the sample with higher coverage was downsampled to match the coverage of the other sample.

HiC-Spector. The starting point of spectral analysis is the Laplacian matrix L, which is defined as L=D-W, where W is a symmetric and non-negative matrix representing a chromosomal contact map, and D is a diagonal matrix in which $D_{ii} = \sum_j W_{ij}$. The matrix L is further normalized by the transformation $D^{-1/2}LD^{-1/2}$, and its leading eigenvectors are found. As in other commonly used dimensionality reduction procedures, the first few eigenvalues are of particular importance because they capture the basic structure of the matrix, whereas the latter eigenvalues are essentially noise. Given two contact maps W^A and W^B, their corresponding Laplacian matrices L^A and L^B and corresponding eigenvectors are calculated. 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 L^A and L^B. A distance metric is defined as:

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

Here || || represents the Euclidean distance between the two vectors. The parameter r is the number of leading eigenvectors used. In general, S_d provides a metric to gauge the similarity between two contact maps. The distance is then linearly rescaled to a reproducibility score ranging from 0 to 1.

QuASAR. The Quality Assessment of Spatial Arrangement Reproducibility (QuASAR) measure uses the concept that within a distance matrix, as the distance between two features approaches zero, the correlation between the rows corresponding to those two features will approach one. This relationship is highlighted by calculating the interaction correlation matrix, weighted by interaction enrichment. To determine reproducibility across replicates, the correlation of weighted correlation matrices was calculated, as follows. For the reproducibility analyses of 11 cell types, contact matrices were re-binned from 40kb to 120kb resolution prior to analysis by QuASAR-rep. In the other reproducibility analyses using deeply sequenced data

sets, the contact matrices were not re-binned. In every case, matrices were filtered by removing intra-chromosomal interaction matrix rows and columns such that all remaining rows and columns contained at least five non-zero entries. The background signal-distance relationship was estimated as the mean number of reads for each inter-bin distance. The interaction correlation matrix was calculated across all pairwise sets of rows and columns from the log-transformed enrichment matrix excluding bins falling on the diagonal in either set from the correlation calculation. Row/column pairs with zero reads occurriThe Quality Assessment of Spatial Arrangement Reproducibility (QuASAR) measure uses the concept that within a distance matrix, as the distance between two features approaches zero, the correlation between the rows corresponding to those two features will approach one. This relationship is highlighted by calculating the interaction correlation matrix, weighted by interaction enrichment. To determine reproducibility across replicates, the correlation of weighted correlation matrices was calculated, as follows. For the reproducibility analyses of 11 cell types, contact matrices were re-binned from 40kb to 120kb resolution prior to analysis by QuASAR-rep. In the other reproducibility analyses using deeply sequenced data sets, the contact matrices were not re-binned. In every case, matrices were filtered by removing intra-chromosomal interaction matrix rows and columns such that all remaining rows and columns contained at least five non-zero entries. The background signal-distance relationship was estimated as the mean number of reads for each inter-bin distance. The interaction correlation matrix was calculated across all pairwise sets of rows and columns from the log-transformed enrichment matrix excluding bins falling on the diagonal in either set from the correlation calculation. Row/column pairs with zero reads occurring at their intersection were excluded from the correlation matrix. Note that, to distinguish the use of QuASAR for assessing reproducibility versus data quality (described below), we refer in the main text to “QuASAR-Rep” and “QuASAR-QC”.

Processing of reproducibility scores. All the reproducibility measures we use in this study assign a reproducibility score to a pair of Hi-C contact matrices. Due to the sparsity and noise nature of inter-chromosomal matrices, reproducibility scores are only calculated for intra-chromosomal matrices. The final reproducibility score assigned to a pair of Hi-C experiments in this study is the mean of the reproducibility scores assigned to pairs of Hi-C contact matrices of each chromosome.

Empirical reproducibility score thresholds. To infer empirical thresholds for distinguishing non-replicates for biological replicate pairs for each method, we used the distribution of reproducibility scores assigned to non-replicate pairs and biological replicate pair at a given coverage level. Similar to the concept of a maximal margin hyperplane, the empirical threshold we inferred is midpoint of the reproducibility score of the highest scoring non-replicate pair and the reproducibility score of the lowest scoring biological-replicate pair. For each coverage level from 30 million Hi-C interactions to 5 million interactions, we inferred a single empirical threshold for each reproducibility metric. These thresholds are available in Supp. Table 2.

Measures of quality

QuASAR. The sample quality measure from QuASAR ("QuASAR-QC") uses the same transformation as described above for reproducibility except contact matrices were re-binned at 1Mb resolution. However, instead of looking at weighted correlation matrices between samples, the quality score is found by taking the weighted mean across all chromosomes and subtracting the unweighted mean of correlation matrices across all chromosomes.

TAD boundary calling and analysis TAD boundaries were identified using the insulation score [37]. This score captures the density of signal in the Hi-C contact matrix around the diagonal, as a function of genomic position. Because the signal is weaker at the boundary of two TADs, minima in the insulation score profile correspond to TAD boundaries. We used the TAD calling

software described in Giorgetti et al. [5], employing the previously used parameters (--ss 80000 --im iqrMean --is 480000 --ids 320000) for calculation of the insulation score and identification of minima.

To characterize the effects of noise and coverage on TAD boundary identification, we used noise injected and downsampled datasets as explained before and used insulation score method as desribed in the previous section. For noise injected datasets, we found that number of identified TADs across the genome are only altered at the highest noise levels: the number of total TADs increased only by 5% with 50% noise injection (Supp. Fig 8). Consistent with the changes in the total number of TADs, the distribution of TAD sizes is only altered at high noise levels. For 7 out of 11 cell types, we detect a statistically significant reduction in the TAD size distribution ($P < 0.01$, Kolmogorov-Smirnov test) only at either 40% or 50% noise (Supp. Fig 9). Furthermore, positions of TAD boundaries are not altered with increasing noise levels. For 11 cell types, we calculated the distances between the TAD boundaries of the combined noise-free biological replicate and the TAD boundaries from noise-injected replicates. These distances were compared against the TAD boundary distances from biological replicate pairs, which serves as a baseline for how much the TAD boundaries fluctuate between different replicates from the same cell type. Again, we found that the boundary distances are significantly larger than the baseline distribution (one sided KS test, $P < 0.05$) only at the 50% noise level for four cell types and never larger for the remaining four cell types (Fig. 4C, Supp Fig. 10).

We adopted the same approach for investigating the effect on coverage on insulation score identified TADs. For each of the six downsampled cell lines, we identified TADs using insulation score method and compared total number of TADs, the size distribution of TADs and the differences between TAD boundaries between the original replicate and downsampled replicates. We observe that the total number of TADs detected and TAD size distributions are similar at all coverage levels (Supp. Figures 13 and 14). We calculated the distances between

TAD boundaries identified from downsampled replicates against the TAD boundaries from original biological replicates, and we compared this distribution against the distances between biological replicates as a baseline. For five of the six cell types, downsampling causes the TAD boundaries to shift away from the original boundaries significantly (Kolmogorov-Smirnov test, $P < 0.05$) only 10 million and lower number of interactions, further supporting the idea that TAD boundary by insulation score detection is mostly robust to low coverage (Fig. 4F, Supp Fig. 15).

Number of significant contacts For a given normalized Hi-C contact map, we computed the number of contacts that are deemed statistically significant using Fit-Hi-C [35]. Hi-C contact maps were binned at 40kb resolution and normalized using the Knight-Ruiz matrix balancing algorithm [42]. Deeply sequenced Hi-C data from 2 cell types were binned at 10kb and 40kb resolutions for Fit-Hi-C analysis (Supp. Table 2). Fit-Hi-C assigns a statistical significance to each contact between two bins by assigning a p-value and a q-value. For each experiment, we counted the number of contacts that are above a given q-value threshold for every intra-chromosomal interactions and aggregated them over all chromosomes and used this sum as the total number of significant contacts for a given experiment.

Mapping statistics

We have used three statistics to summarize alignment quality, valid Hi-C fragment pairs, and the ratio of intrachromosomal and interchromosomal Hi-C interactions. A thorough description of these statistics and their application is reviewed in Lajoie et al [23]. First statistic we use is percentage of aligned pairs, which corresponds to the percentage of Hi-C fragment pairs that uniquely map to the genome on both sides. Typically, single sided and non-unique alignments are discarded in Hi-C pipelines [22, 23]. The second statistic is invalid pairs, which the percentage of aligned pairs that map against the same restriction fragment. These fragment pairs are non-informative since they do not correspond to a fragment between two different

regions [23]. The third statistic is the percentage of intra-chromosomal valid pairs. Random ligations are much more likely to result in interchromosomal fragments; thus a high ratio of non-informative random ligation events result in an enrichment of inter-chromosomal interactions and a depletion of intra-chromosomal interactions [23].

Simulation of noisy Hi-C matrices

To generate noise for Hi-C data in a realistic manner, we simulated two Hi-C contact matrices that would result from two processes that are not dictated by the 3D organization of chromatin. These “pure noise” matrices are mixed with the real Hi-C contact matrix to generate the final, noisy Hi-C matrix. The first noise matrix models the genomic distance effect, namely the higher probability of observing a Hi-C interaction between two regions that are close along the one dimensional length of a chromosome. Because such regions are constrained to be close to each other, they are more likely to interact compared to more distal regions, in the absence of any higher order structure. This effect has been documented early on and is generally corrected in Hi-C contact matrices to better visualize medium and long range interactions [1]. The second noise matrix models the ligation of non-crosslinked DNA fragments during the ligation step of the Hi-C protocol. Fragments pairs that results from random ligation are uninformative since they can link two regions independently of 3D organization.

Additionally, the Hi-C assay is subject to the same biases that other next generation sequencing assays suffer from. These biases result include a bias in favor of GC rich regions and a bias against regions of low mappability. During the generation of both types of noise matrices, we factored in such biases by using the sum of each row as a proxy for the the overall bias of a bin. Coverage normalization of Hi-C matrices [1] similarly uses marginals to counter such biases.

To generate the genomic distance noise matrix G, we sampled from empirical distributions derived from of real Hi-C matrix. In this setting, the genomic distance D is defined as the

number of bins that lie between a pair of bins i and k, i.e. $|i - k| = D$. For every value of D, we build a vector S by collecting the set of real Hi-C matrix entries M_{ik} for which $|i - k| = D$. We then randomly select values from S for insertion into G, again considering only entries G_{ik} for which $|i - k| = D$. This sampling strategy effectively shuffles the matrix entries in M at a fixed distance, thus preserving the original genomic distance effect while disrupting other higher order structures. However, instead of uniformly sampling from S, we adopted a stratified sampling strategy to better model GC and mappability biases. Specifically, S was broken into multiple strata before sampling. The strata are determined by products of marginals, i.e. M_{ik} is assigned to a certain stratum based on the product of the marginals of bin i and bin k. For a given value of D, we chose stratum size in such a way that each each stratum contains 100 elements. When sampling the G_{ik}, we sampled a value from from the stratum that M_{ik} belongs to. By repeating the stratified sampling for every value of D, the final matrix G is obtained.

To generate the random ligation noise matrix R, we generated random Hi-C interactions and aggregated them to build a Hi-C contact matrix. We generated these interactions by randomly choosing two bins *i* and *k*, and adding one to the matrix entry R_{ik} in the random noise contact matrix. Instead of sampling the bins uniformly, the probability of sampling a bin was set proportional to marginal of that bin, thus modeling the GC and mappability bias of each bin. The sampling process was repeated N times, where N is the total number of interactions in the original Hi-C contact matrix M, to generate a random ligation noise matrix.

After both noise matrices are generated from the original Hi-C matrix, these matrices were mixed in varying proportions to generate a series of noisy Hi-C matrices. Each such matrix is a mixture of three matrices: a real matrix, a genomic distance noise matrix, and a random ligation noise matrix. To generate a simulated matrix with *c* total counts from, we sampled counts uniformly at random from one real and two simulated matrices at a given target ratios. In practice, we varied the total proportion of noise from 0% to X%, and for each total noise level we

consider two settings for the relative proportions of genomic distance noise random ligation noise: we either used one third of matrix G and two thirds of matrix R, or vice versa. We note that most analyses in this study were robust to either scenario.

The software for injecting noise into Hi-C contact matrices is available at https://github.com/gurkanyardimci/hic-noise-simulator.

Downsampling

Downsampled data sets were generated by converting an input Hi-C matrix into a set of pairwise individual intra-chromosomal interactions and uniformly sampling a given number of interactions from this set. Following downsampling, we re-binned the set of chosen interactions into a Hi-C matrix.

For analysis of reproducibility measures, we limited the analysis to real data from six cell types with replicates of at least 30 million interactions, and we downsampled each individual replicate to have a wide range of total interactions (30×10^6, 25×10^6, 20×10^6, 15×10^6, 10×10^6, 5×10^6). Using a single pseudo-replicate and a single biological replicate pair for each cell type and 15 non-replicates at each coverage level, we generated a total of 162 replicate pairs. These datasets were used for testing the ability of each method to distinguish among different replicate types at lower coverage levels, and for explicitly profiling the dependence of reproducibility scores on coverage levels.

For analysis of QC measures, we generated downsampled biological replicates from the same six cell types to have fewer interactions (30×10^6, 25×10^6, 20×10^6, 15×10^6, 10×10^6, 5×10^6, 10^6), resulting in a set of 84 matrices. In addition, we applied the same setup to deeply sequenced datasets from two cell types at a wider range of coverage values (30×10^6, 60×10^6,

120 $\times 10^6$, 240 $\times 10^6$, 400 $\times 10^6$), at multiple resolutions, resulting in 30 matrices. For each downsampled matrix, we calculated QuASAR scores and identified statistically significant long range contacts and TAD boundaries.

Generation of pseudo-replicates

Given two biological replicate experiments, we generated pseudo-replicates by aggregating the two replicates and downsampling from the combined matrix. Combination of two biological replicates is performed by summing the two Hi-C contact matrices of these replicates. Following combination, the resulting combined Hi-C matrix is downsampled as described above to generate pseudo-replicates. We forced the pseudo-replicates to have the average of total number of interactions of two seed biological replicates.

Funding

G.G.Y and W.S.N are supported by awards NIH U41HG007000, U24HG009446. H.O. is supported by DK107980. B.R.J is supported by awards HG004592, HG003143 and J.D. is supported by awards HG004592, HG003143, DK107980. M.E.G.S. and J.T. is supported by awards NIH R24 DK106766, U41 HG006620. O.U is supported by Howard Hughes Medical Institute International Student Research Fellowship and a Gabilan Stanford Graduate Fellowship award and A.K. is supported by awards NIH DP2OD022870, U24HG009397, R01ES025009-02S1. T.Y. is supported by NIH T32 GM102057 (CBIOS training program to The Pennsylvania State University), a Huck Graduate Research Innovation Grant and Q.L. is supported by awards NIH R24 DK106766, U41 HG006620.

Authors' Contributions

GGY, JD, WSN designed the experiments. Y.Z. and B.R.J. generated and processed the data. G.G.Y., H.O., M.E.G.S., O.U., K.Y., T.Y, A.C., A.K., F.A. ran the experiments. G.G.Y and W.S.N. analyzed the results. G.G.Y, O.U. and W.S.N made the figures. All authors contributed to the preparation of the manuscript.

Acknowledgements

We thank Giancarlo Bonora and Kate Cook for useful discussions.

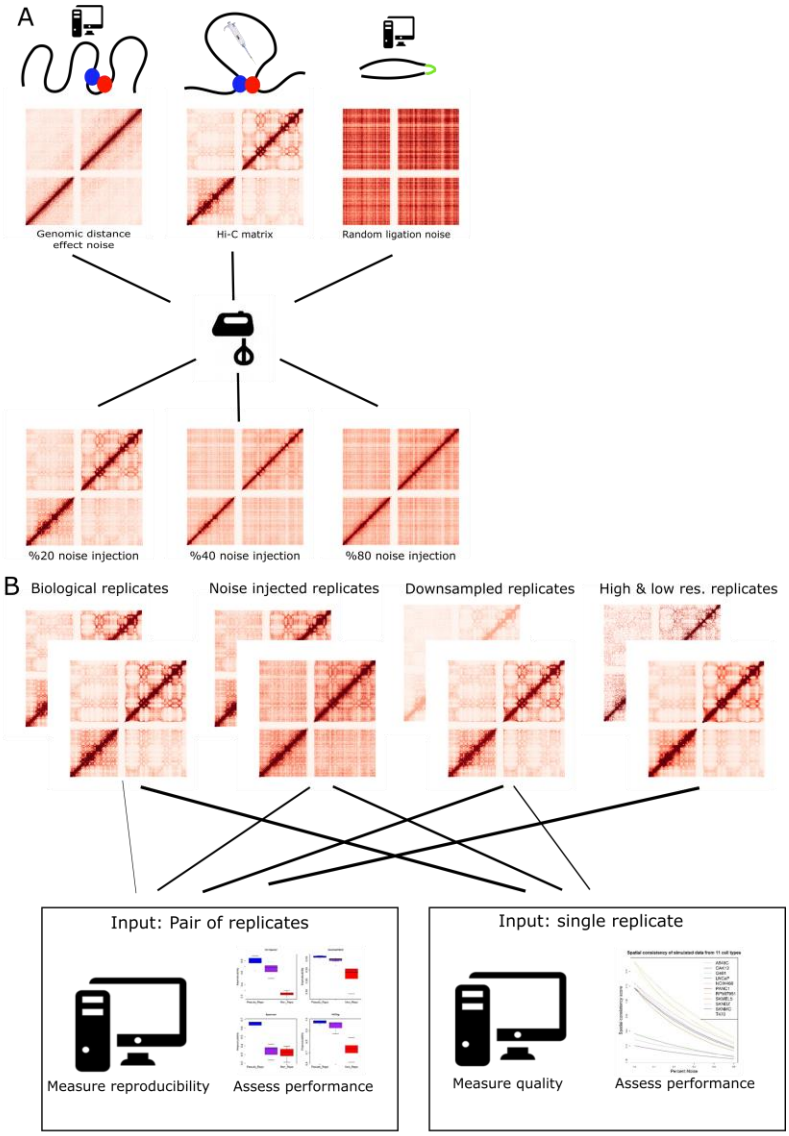


Fig 1. **Overview of the study.** (A) Schematic showing the approach for generating noise-injected Hi-C matrices. In the upper panel, we generate two types of noise from real Hi-C data (center): random ligation noise (right) and genomic distance effect noise (left). The three matrices are then mixed to generate noisy datasets (lower panel). By changing the mixing proportions, we can create datasets with varying percentages of noise. (B) To benchmark the performance of various quality control and reproducibility measures, we compiled a large number of Hi-C replicates from 13 cell types, and simulated noise-injected datasets from the original data. Real and simulated datasets binned at different resolutions and downsampled to different coverage levels are the inputs to reproducibility and quality control measures where each replicate pair and single replicate are assigned a score. Performance of each measure is evaluated on their ability to correctly rank real and simulated datasets. (C) Summary of the basic principles of the four reproducibility methods evaluated in this study.

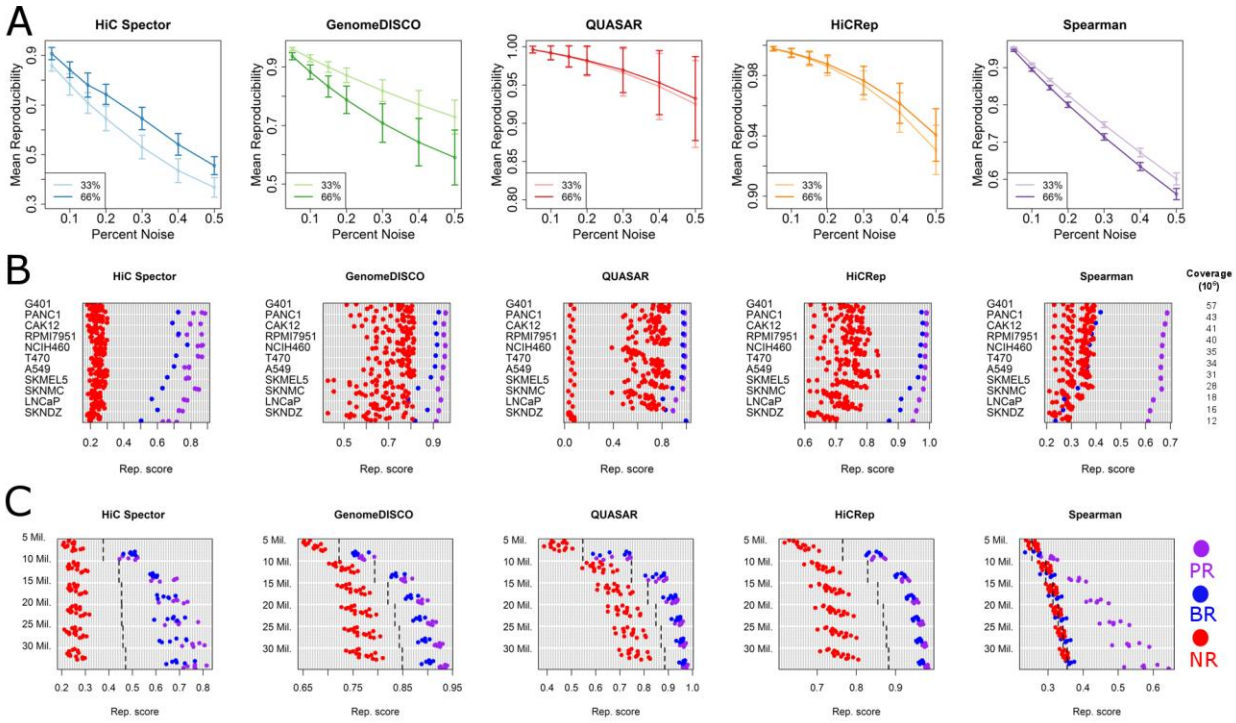


Figure 2. **Comparison of reproducibility measures.** (A) Curves showing the mean reproducibility score assigned to 11 cell types at each noise injection level for 33% and 66% random ligation noise configurations. Vertical bars represent one standard deviation away from the mean. The monotonic decay of each curve shows that each measure is able to capture the expected decay in reproducibility levels with respect to noise level. (B) Reproducibility scores assigned to biological replicate (blue), non-replicate (red) and pseudo-replicate (purple) pairs for each cell type. Note that the point corresponding to the pseudo-replicate pair for SKNDZ for QuASAR is occluded by one of the biological replicate points. (C) Reproducibility scores assigned to biological replicate (blue), non-replicate (red), and pseudo-replicate (purple) pairs from six cell types at seven different coverage levels. Dashed lines indicate empirical threshold for distinguishing biological replicate pairs from non-replicate pairs.

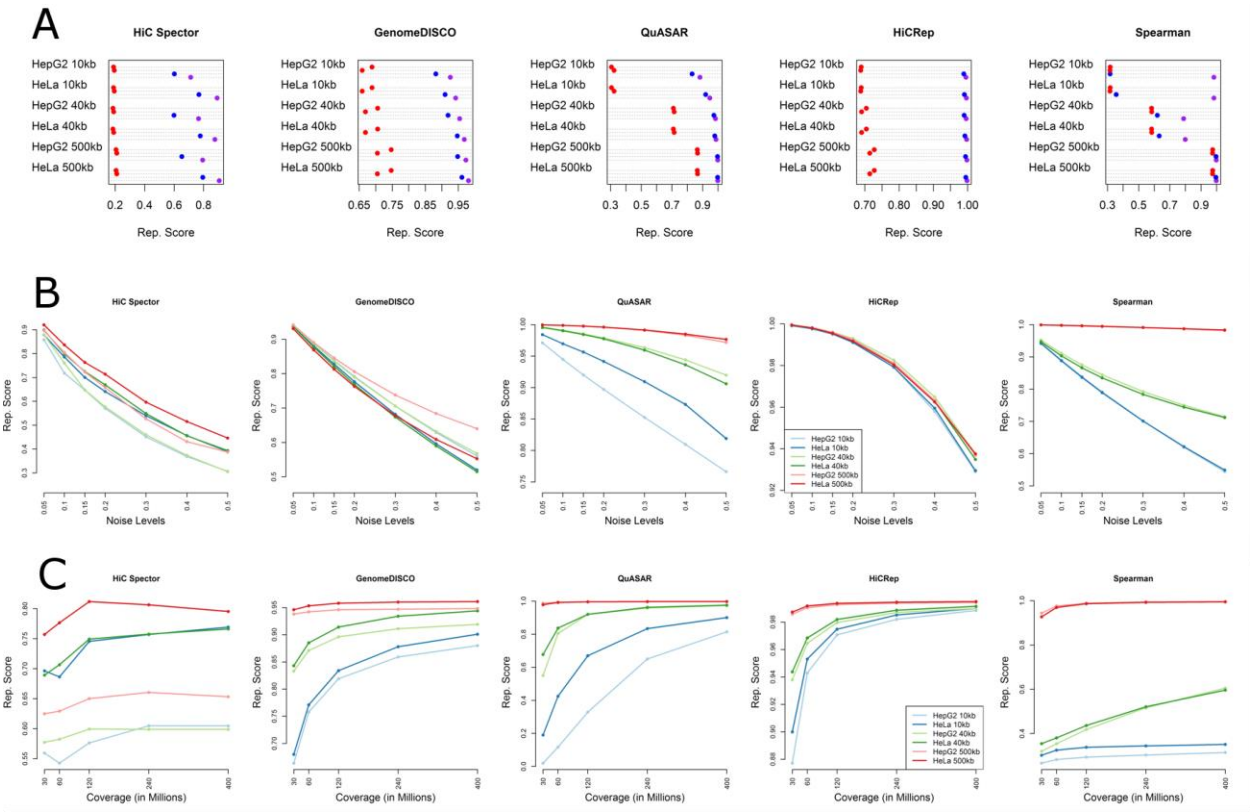


Fig 3. **Effects of resolution on reproducibility measures**. (A) Reproducibility scores assigned to biological replicate (blue), non-replicate (red), and pseudo-replicate (purple) pairs from HepG2 and HeLa Hi-C datasets at 10kb, 40kb and 500kb resolutions. (B) Reproducibility scores of noise injected replicate pairs decay with increasing levels of noise at low and high resolutions. (C) Reproducibility scores assigned to downsampled biological replicate pairs at different resolutions. Reproducibility scores from each measure plateau around 120 or 240 million interactions for all resolutions.

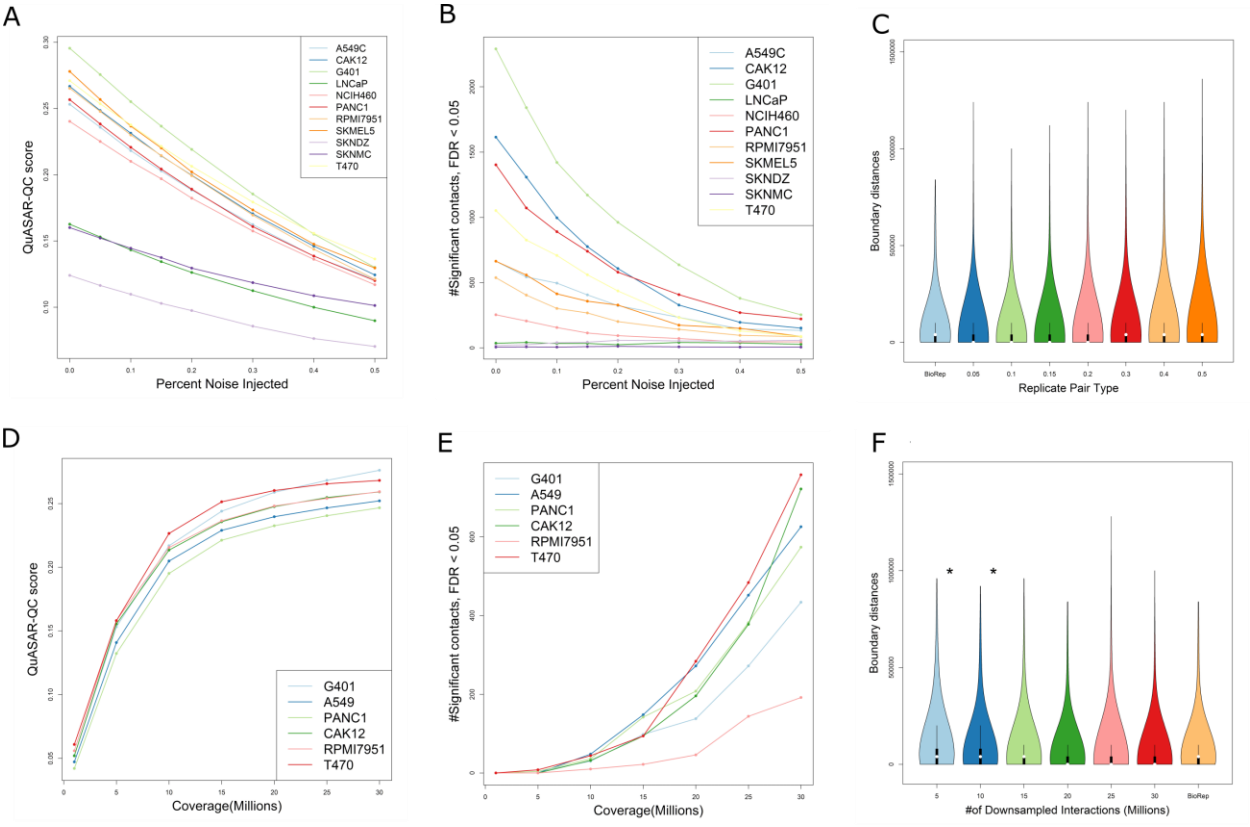


Figure 4. **Quality measures.** (A) QuASAR-QC scores assigned to noise injected matrices from 11 cell types (B). Total number of significant contacts above a 5% FDR threshold from noise injected matrices from 11 cell types. (C) Violin plots showing the distribution of TAD boundary distances between biological replicates and noise-injected replicates for T470 cells. There is no significant change in the distribution of TAD boundary distances at any given noise level. (D) QuASAR-QC scores assigned to downsampled replicates from six different cell types. (E) Total number of significant contacts above a 5% FDR threshold from downsampled replicates from six different cell types. (F) Violin plots showing the distribution of distances between domain boundaries in biological replicates and noise-injected replicates for T470 cells. In panels C and F, asterisks indicate that the distribution of boundary distances is significantly larger than the null distribution, which is obtained by comparing biological replicates.

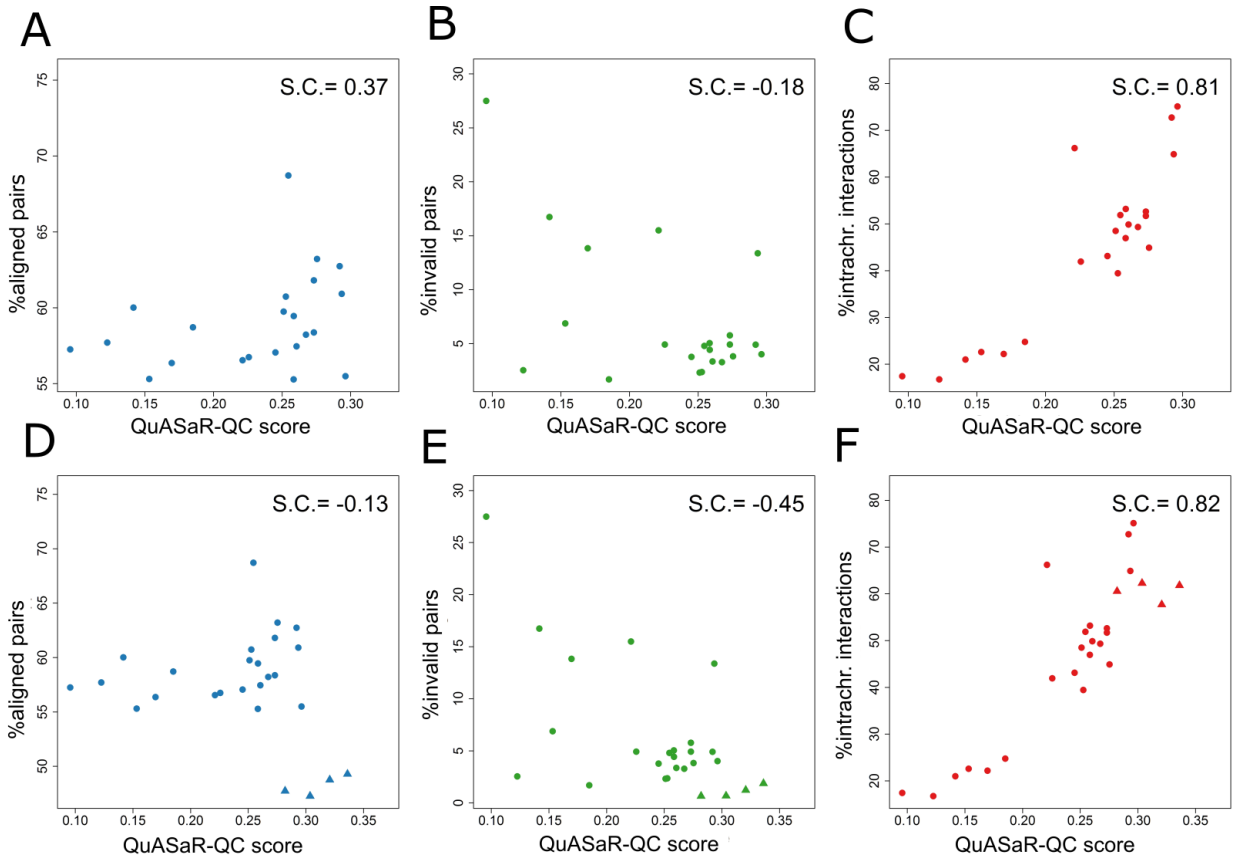
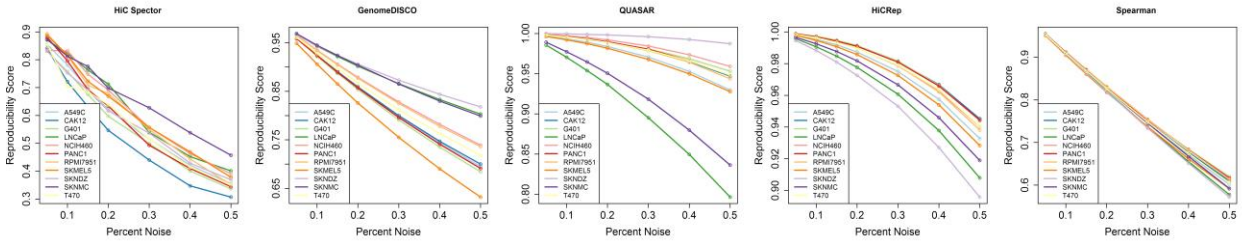
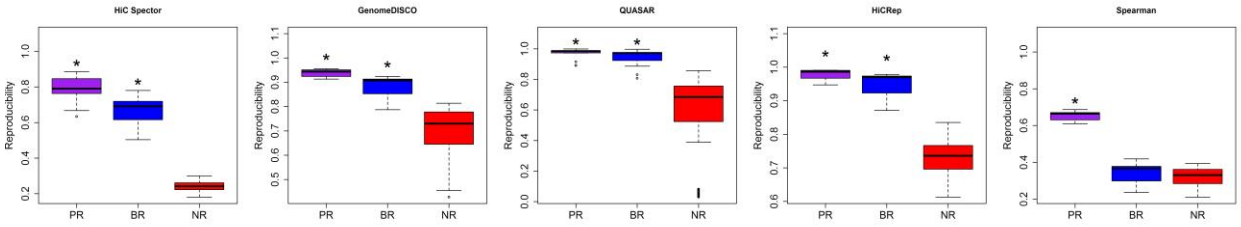


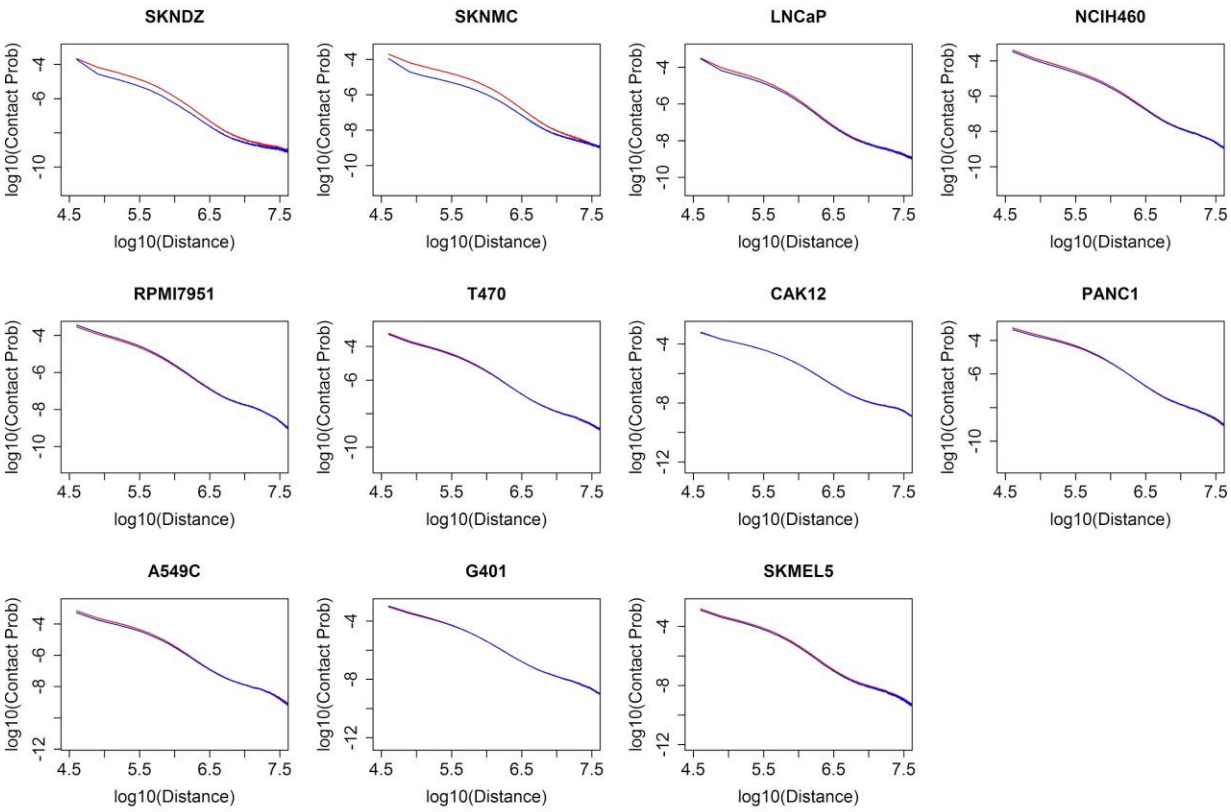
Figure 5. **Comparison of QuASAR-QC to previously described statistics.** (A-C) Scatter plots of QuASAR-QC scores of biological replicates from 11 cell types plotted against quality statistics that describe (A) alignment mapping percentage, (B) percentage of artefactual Hi-C fragments and (C) percentage of intra-chromosomal interactions. (D-F) Same as (A-C), but deeply sequenced Hi-C replicates are included in the scatter plots as triangles, regular replicates are shown as circles. In each plot, the Spearman correlation between the QuASAR-QC scores and the statistic are indicated in the upper right corner.



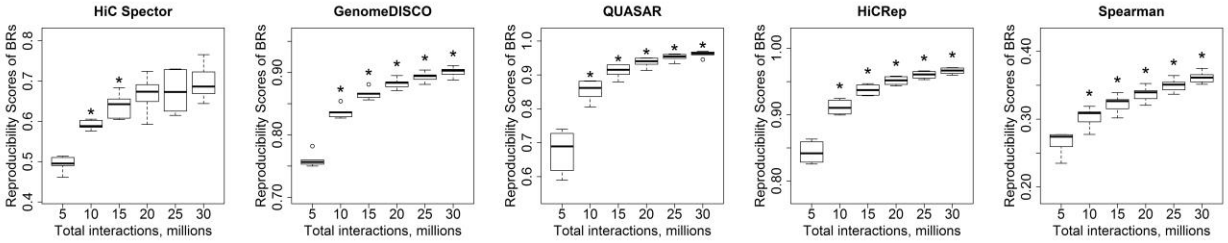
Supp. Fig. 1. Reproducibility scores assigned to each noise injected replicate pair for each cell type for 33% random ligation configuration. For every cell type and every measure, we see a monotonic trend of decreasing scores with higher levels of noise. The same trends for each measure are observed in the 66% random ligation noise configuration (not shown).



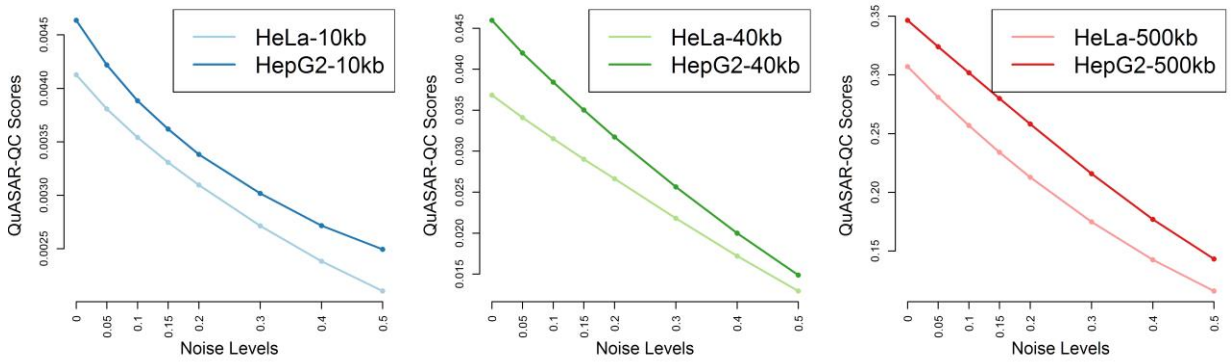
Supp. Fig. 2. Boxplots showing the distribution of reproducibility scores assigned to each replicate pair category by each measure: pseudo replicates (PR), biological replicates (BR) and non-replicates (NR). Asterisks indicate that the marked distribution is significantly larger than the preceding one according to a one-sided Kolmogorov-Smirnov test ($P < 0.01$).



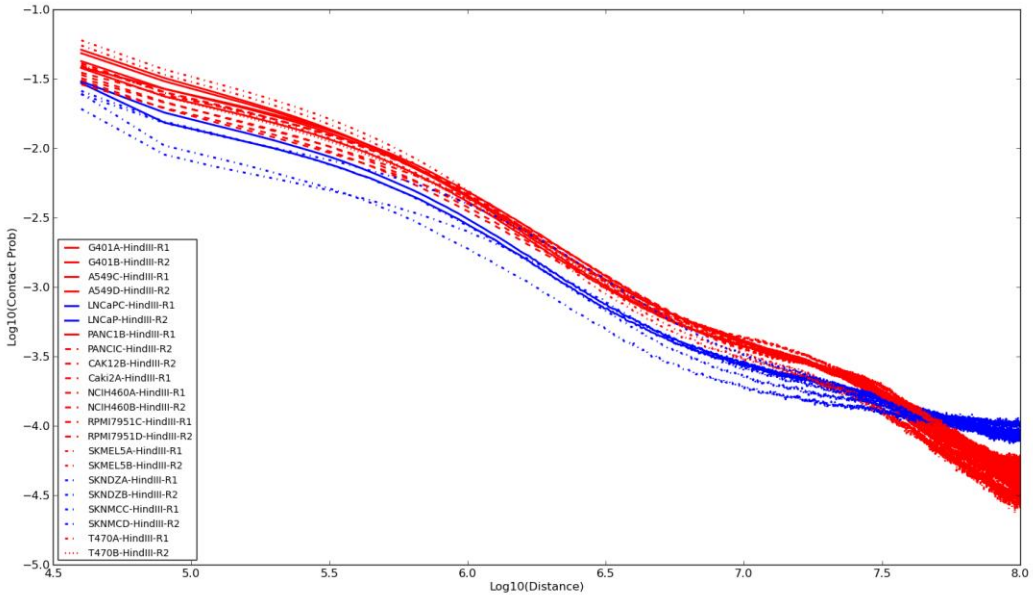
Supp Fig 3. Contact probability curves for biological replicates pairs from each replicate pair. The curve for first replicate is shown in red and the second replicate is in blue.



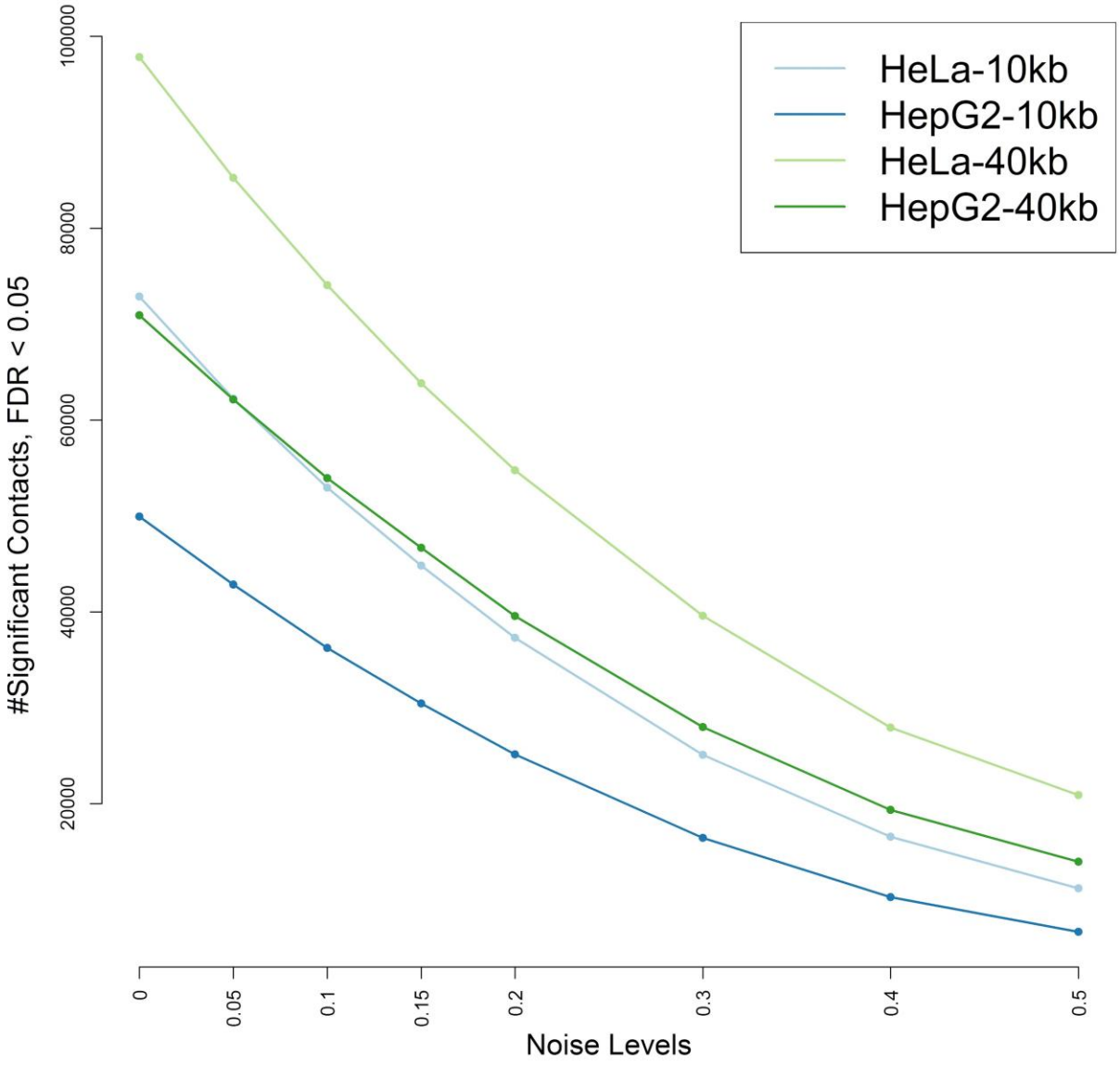
Supp. Fig. 4. Boxplots showing the distribution of reproducibility scores assigned to six downsampled biological replicates at each coverage level. Asterisks above each distribution indicate that the distribution of reproducibility scores assigned to that distribution is significantly larger than the previous distribution according to a one-sided Kolmogorov-Smirnov test ($P < 0.05$)



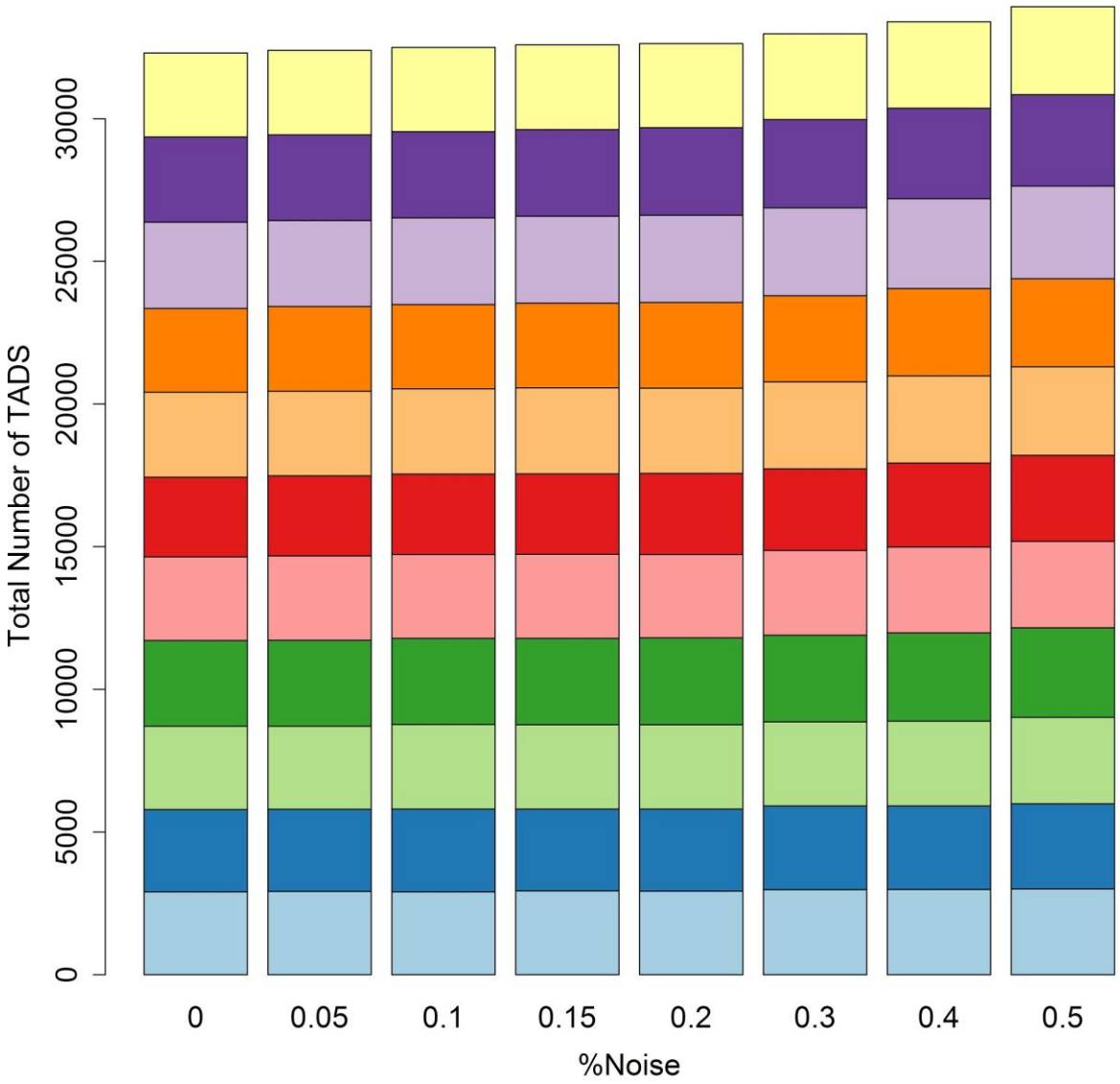
Supp. Figure 5. QuASAR-QC scores assigned to noise injected simulated datasets from deeply sequenced replicates. QuASAR-QC scores decrease with increasing levels of noise at all resolutions.



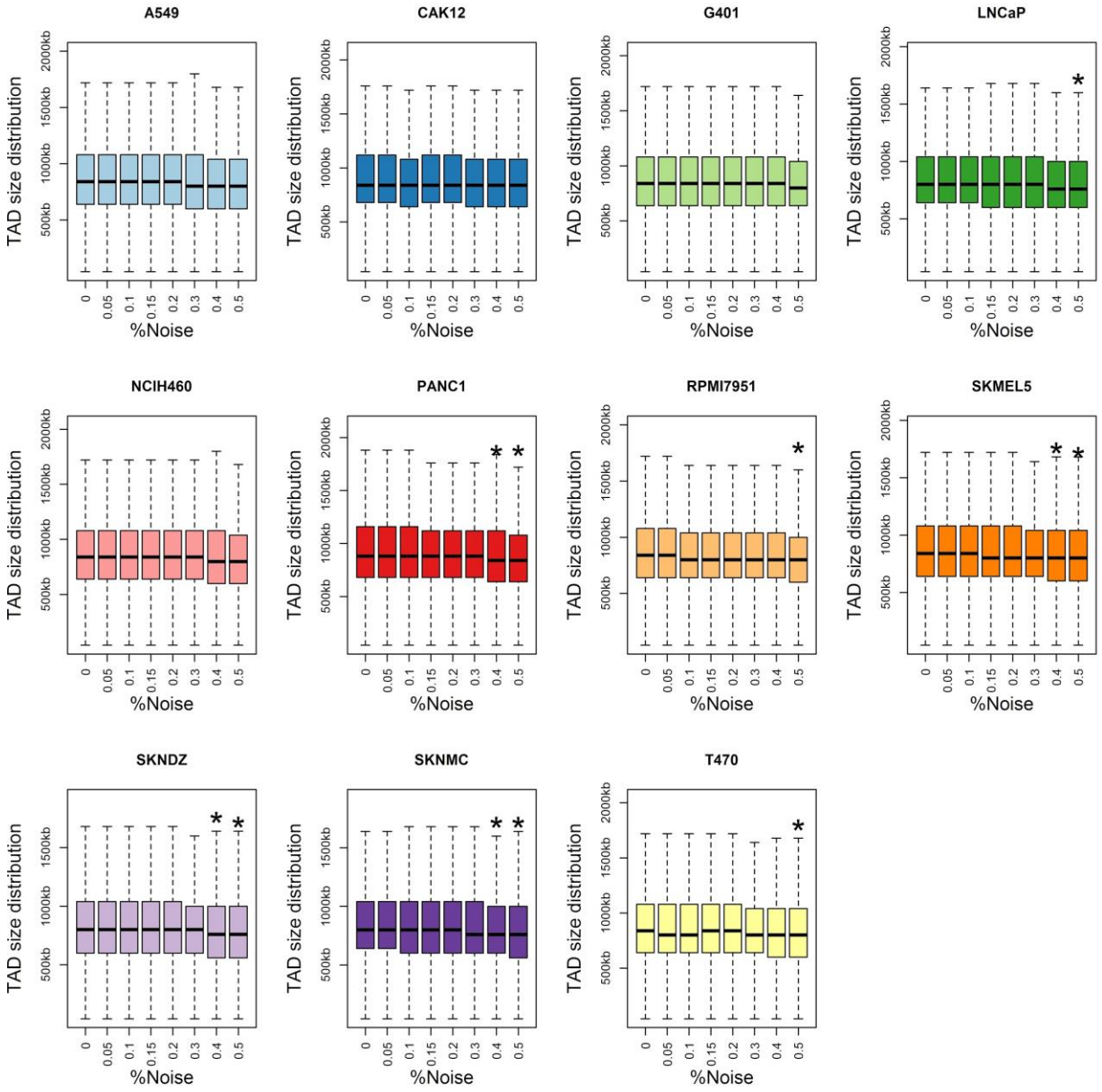
Supp Figure 6. Curves showing the probability of contact between two loci against genomic distance. The blue curves are generated using Hi-C experiments done on SKNDZ, SKNMC and LNCaP cell lines.



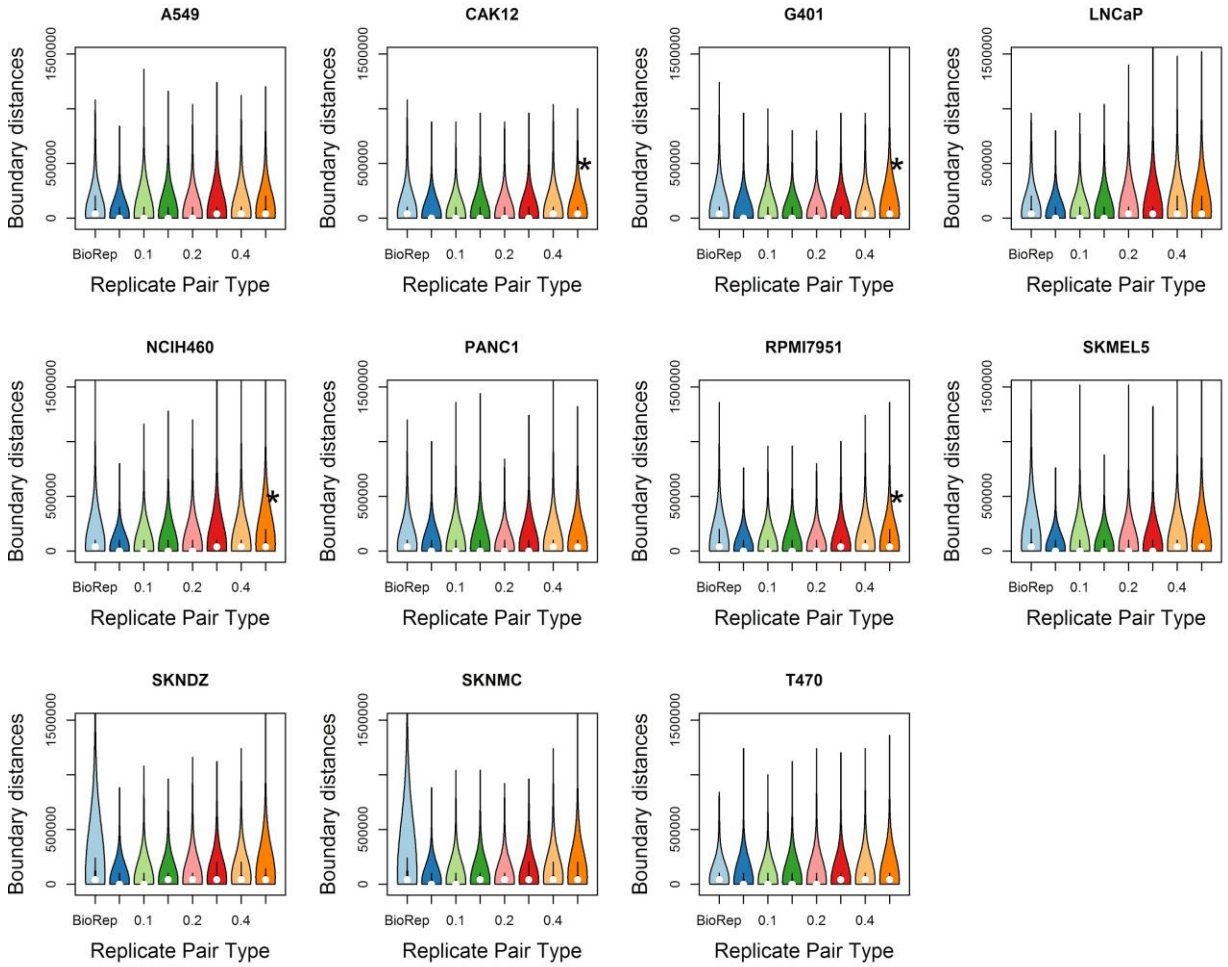
Supp. Figure 7. Total number of significant mid-range contacts identified by Fit-Hi-C with an FDR threshold of 0.05 from simulated datasets at 10kb and 40kb resolutions. The total number of contacts for each cell type and resolution decreases monotonically with increasing levels of noise.



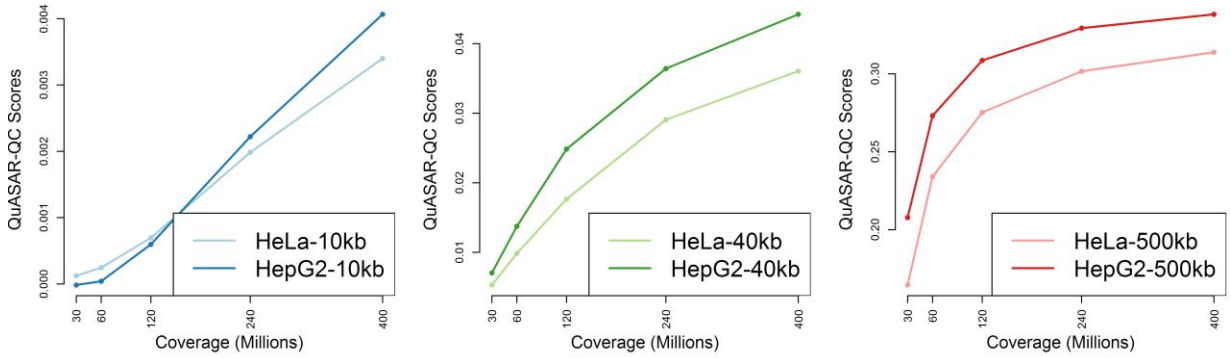
Supp. Fig 8. Bar plots showing the number of TADs for each cell line (coded by color) at each noise injection level.



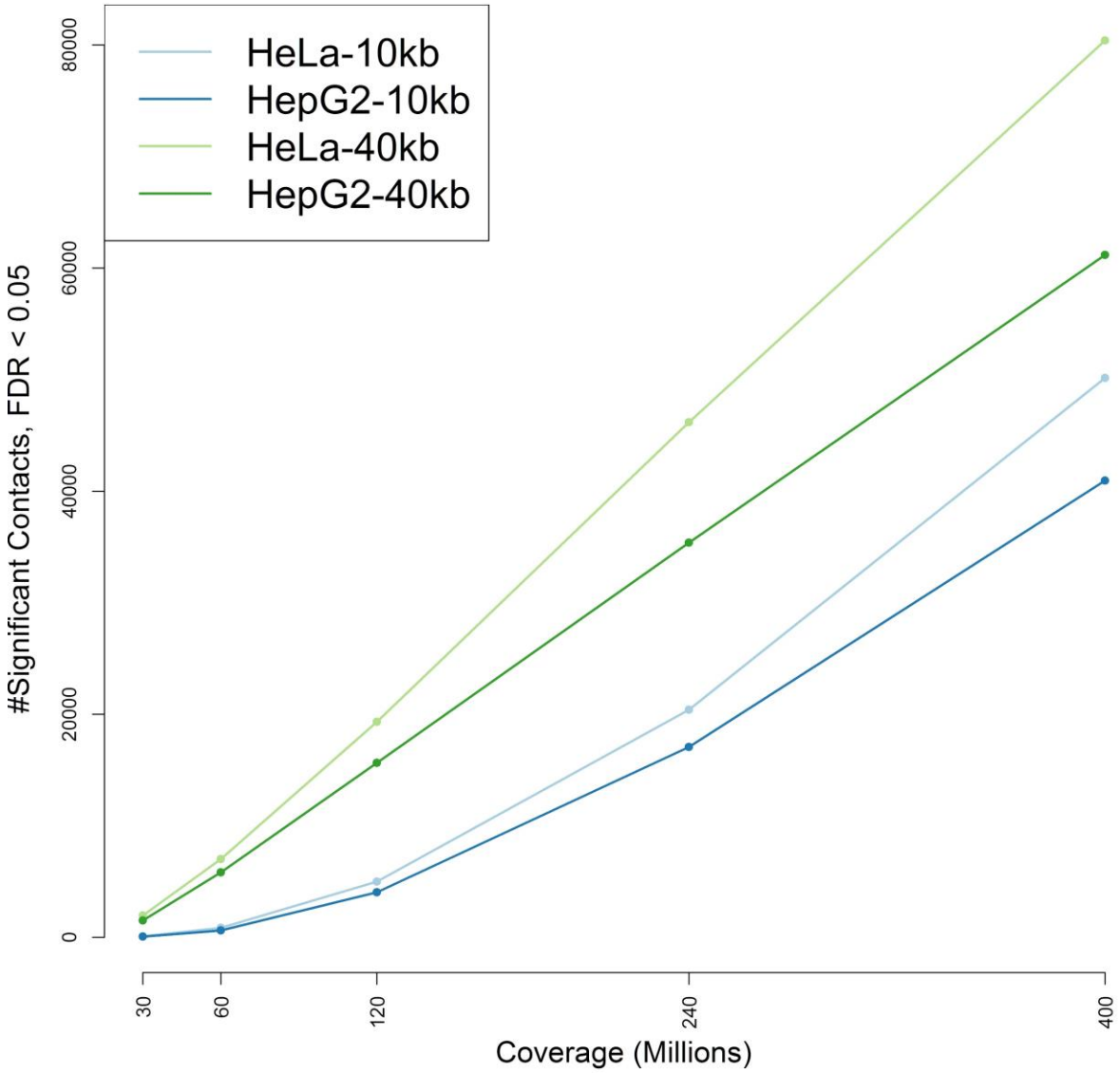
Supp. Fig 9. Boxplots showing the distribution of TAD sizes at each noise injection level. Each plot corresponds to a simulated dataset from an individual cell type. Distributions marked with an asterisk are significantly different from the original distribution TAD sizes detected from the noise-free replicate (KS test, $P < 0.01$).



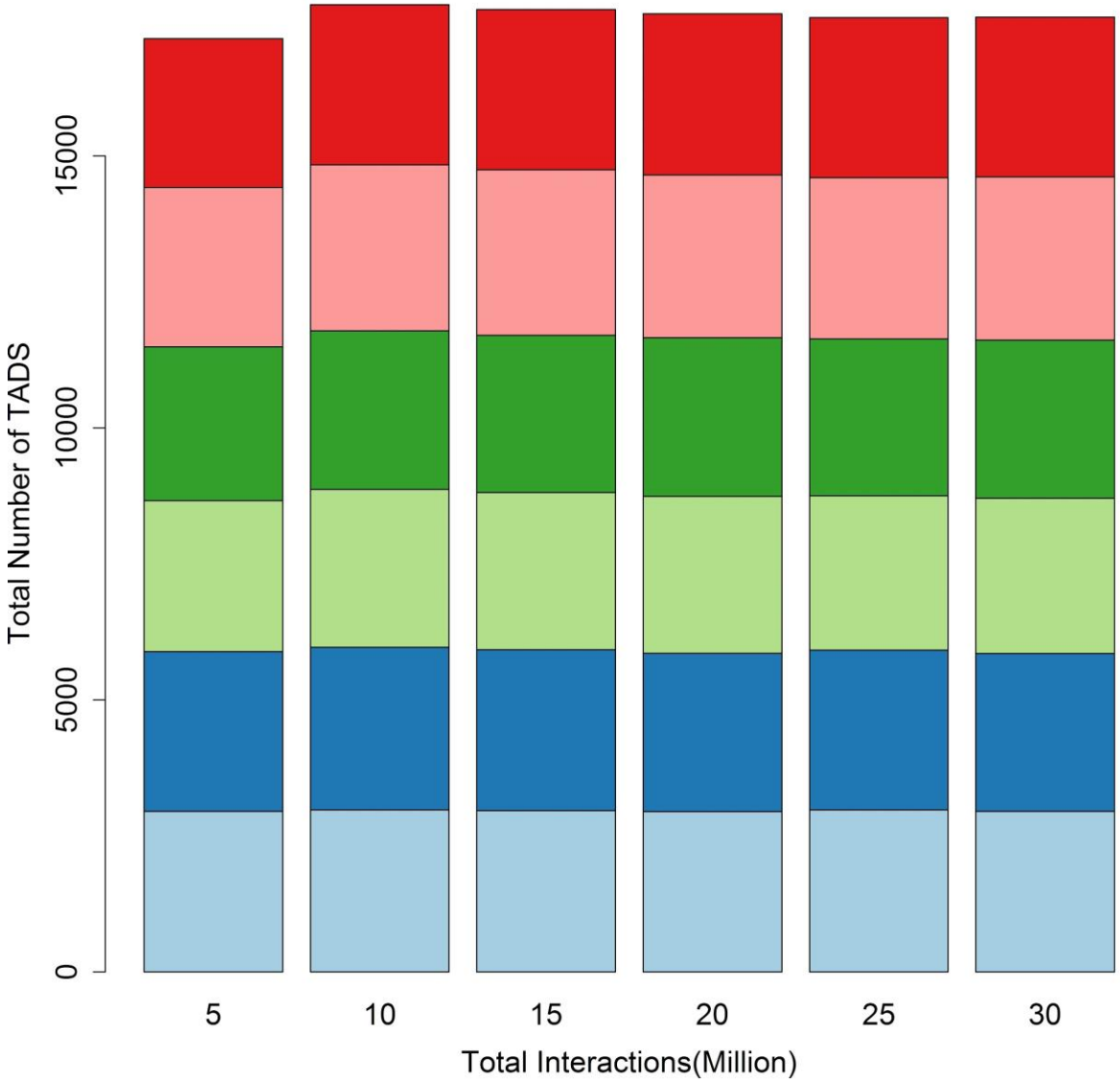
Supp. Fig 10. Violin plots showing the distribution of distances between domain boundaries between biological replicates and simulated replicates. Each panel corresponds to a single cell type.



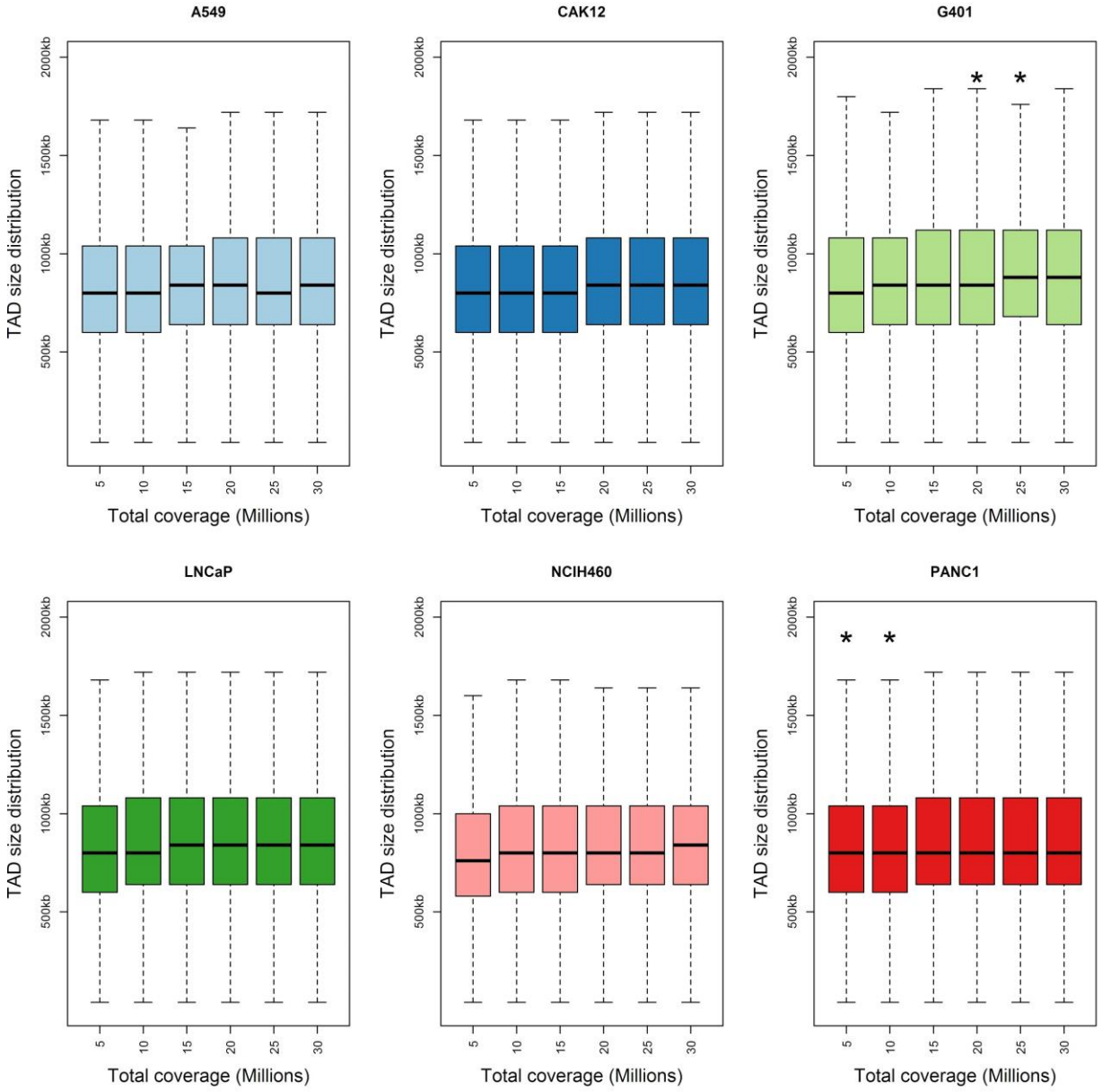
Supp. Figure 11. Curves showing the QuASAR scores assigned to deeply sequenced cell types downsampled to 30, 60, 120, 240 and 400 million interactions at 10kb, 40kb and 500kb resolutions.



Supp. Figure 12. Curves showing the total number significant mid-range interactions detected by FIt-Hi-C from deeply sequenced cell types downsampled to 30, 60, 120, 240 and 400 million interactions at 10kb,40kb.

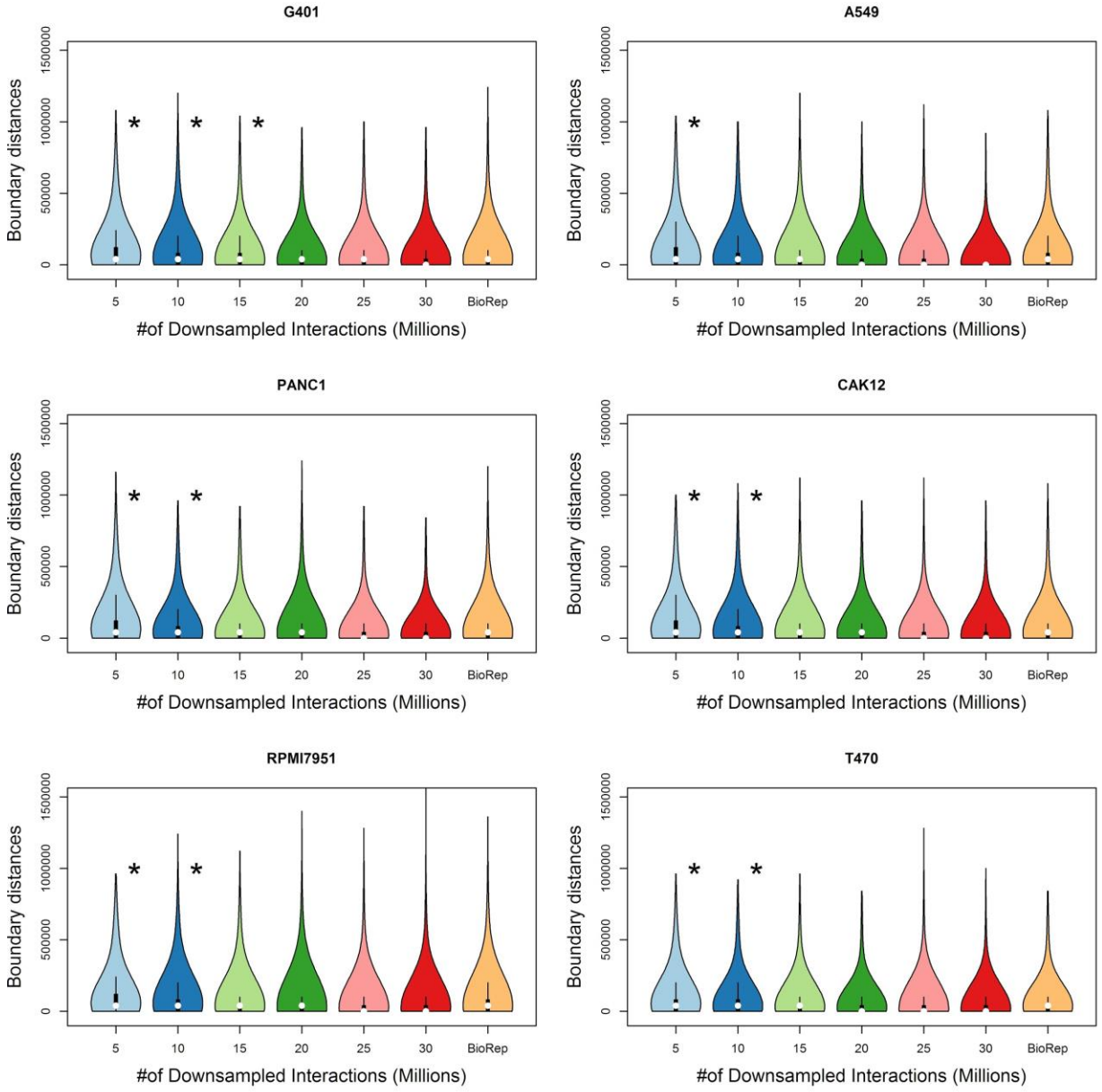


Supp. Fig 13. Barplots showing the number of TADs for each downsampled cell line (coded by color) at each coverage level.



Supp. Fig 14. Boxplots showing the distribution of TAD sizes at downsampling level.

Each plot corresponds to a simulated dataset from an individual cell type.



Supp. Fig 15. Violin plots showing the distribution of distances between domain boundaries between biological replicates and downsampled replicates. Each panel corresponds to a single cell type.

Biosample	Tissue/Morphology	1st replicate coverage	2nd replicate coverage	ENCODE sample IDs	Resolutions
A549	Lung/Epithelial	33,028,385	30,167,658	ENCSR444WCZ	40kb
CAKI2	Kidney/Epithelial	36,297,274	47,032,721	ENCSR401TBQ	40kb
G401	Kidney/Epithelial	61,278,507	52,988,386	ENCSR079VIJ	40kb
LNCaP	Prostate/Epithelial	17,976,198	15,357,134	ENCSR346DCU	40kb
NCIH460	Lung/Epithelial	41,579,896	28,892,164	ENCSR489OCU	40kb
PANC1	Pancreas/Epithelial	37,454,217	50,535,714	ENCSR440CTR	40kb
RPMI7951	Skin/Epithelial	31,953,729	48,764,886	ENCSR862OG	40kb
SKMEL5	Skin/Stellate	45,742,471	10,651,488	ENCSR312KHQ	40kb
SKNDZ	Brain/Epithelial	15,631,291	9,813,185	ENCSR105KFX	40kb
SKNMC	Brain/Epithelial	24,914,561	12,578,436	ENCSR834DXR	40kb
T47D	Mammary Gland/Epithelial	33,902,719	35,957,065	ENCSR549MGQ	40kb
HepG2	Liver/Epithelial	412,741,167	456,705,426	In procession	10kb,40kb,500kb
HeLa	Cervix/Epithelial	515,837,715	494,774,796	In procession	10kb,40kb,500kb

Supp. Table 1. Thirteen human cancer cell types that Hi-C experiments were performed on, together with the tissue type and lineage the cells were immortalized from. Two replicate experiments were performed in each cell type. The coverage columns list the total number of intra-chromosomal interactions for the 1st and the 2nd replicate for each cell type. ENCODE sample ID of each experiment is provided in the corresponding column. The first 11 cell types with lower coverage vaues are binned at only 40kb resolution, whereas the last 2 cell types with large number of Hi-C interactions are binned at three different resolutions.

	30×10^6	25×10^6	20×10^6	15×10^6	10×10^6	5×10^6
HiC-Spector	0.471	0.46	0.455	0.454	0.443	0.376
GenomeDISCO	0.849	0.843	0.834	0.82	0.794	0.722
QuASAR-Rep	0.885	0.87	0.849	0.816	0.747	0.547
HiCRep	0.882	0.877	0.868	0.855	0.829	0.765

Supp. Table 2. Empirical thresholds for distinguishing non-replicates from biological replicates for each measure at a given coverage level. Each column corresponds to empirical threshold inferred by using biological replicates and non-replicates that have been downsampled to the value in the column header (see Methods).

1. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.
2. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation.* Nature, 2015. **518**(7539): p. 331-6.
3. Krijger, P.H., et al., *Cell-of-Origin-Specific 3D Genome Structure Acquired during Somatic Cell Reprogramming.* Cell Stem Cell, 2016. **18**(5): p. 597-610.
4. Deng, X., et al., *Bipartite structure of the inactive mouse X chromosome.* Genome Biol, 2015. **16**: p. 152.
5. Giorgetti, L., et al., *Structural organization of the inactive X chromosome in the mouse.* Nature, 2016. **535**(7613): p. 575-9.
6. Darrow, E.M., et al., *Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture.* Proc Natl Acad Sci U S A, 2016. **113**(31): p. E4504-12.
7. Naumova, N., et al., *Organization of the mitotic chromosome.* Science, 2013. **342**(6161): p. 948-53.
8. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
9. Nora, E.P., et al., *Spatial partitioning of the regulatory landscape of the X-inactivation centre.* Nature, 2012. **485**(7398): p. 381-5.
10. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
11. Jin, F., et al., *A high-resolution map of the three-dimensional chromatin interactome in human cells.* Nature, 2013. **503**(7475): p. 290-4.
12. Schmitt, A.D., M. Hu, and B. Ren, *Genome-wide mapping and analysis of chromosome architecture.* Nat Rev Mol Cell Biol, 2016. **17**(12): p. 743-755.
13. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.
14. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.
15. Landt, S.G., et al., *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.* Genome Res, 2012. **22**(9): p. 1813-31.
16. Li, Q.H., et al., *Measuring Reproducibility of High-Throughput Experiments.* Annals of Applied Statistics, 2011. **5**(3): p. 1752-1779.
17. Qin, Q., et al., *ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline.* BMC Bioinformatics, 2016. **17**(1): p. 404.
18. Ji, H., et al., *An integrated software system for analyzing ChIP-chip and ChIP-seq data.* Nat Biotechnol, 2008. **26**(11): p. 1293-300.
19. Frank, C.L., et al., *Regulation of chromatin accessibility and Zic binding at enhancers in the developing cerebellum.* Nat Neurosci, 2015. **18**(5): p. 647-56.
20. Bardet, A.F., et al., *A computational pipeline for comparative ChIP-seq analyses.* Nat Protoc, 2011. **7**(1): p. 45-61.
21. Ho, J.W., et al., *ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis.* BMC Genomics, 2011. **12**: p. 134.
22. Ay, F. and W.S. Noble, *Analysis methods for studying the 3D architecture of the genome.* Genome Biol, 2015. **16**: p. 183.
23. Lajoie, B.R., J. Dekker, and N. Kaplan, *The Hitchhiker's guide to Hi-C analysis: practical guidelines.* Methods, 2015. **72**: p. 65-75.
24. Tjong, H., et al., *Physical tethering and volume exclusion determine higher-order genome organization in budding yeast.* Genome Res, 2012. **22**(7): p. 1295-305.
25. Hu, M., et al., *HiCNorm: removing biases in Hi-C data via Poisson regression.* Bioinformatics, 2012. **28**(23): p. 3131-3.

26. Gorkin, D.U., D. Leung, and B. Ren, *The 3D genome in transcriptional regulation and pluripotency.* Cell Stem Cell, 2014. **14**(6): p. 762-75.
27. van Berkum, N.L., et al., *Hi-C: a method to study the three-dimensional architecture of genomes.* Journal of visualized experiments : JoVE, 2010(39).
28. Teng, M., et al., *A benchmark for RNA-seq quantification pipelines.* Genome Biol, 2016. **17**: p. 74.
29. Imakaev, M., et al., *Iterative correction of Hi-C data reveals hallmarks of chromosome organization.* Nat Methods, 2012. **9**(10): p. 999-1003.
30. Serra, F., et al., *Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors.* PLoS Comput Biol, 2017. **13**(7): p. e1005665.
31. Nagano, T., et al., *Comparison of Hi-C results using in-solution versus in-nucleus ligation.* Genome Biol, 2015. **16**: p. 175.
32. Yan, K.K., et al., *HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps.* Bioinformatics, 2017.
33. Yang, T., et al., *HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient.* Genome Res, 2017.
34. Ursu, O., et al., *GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs.* BioRxiv, 2017.
35. Ay, F., T.L. Bailey, and W.S. Noble, *Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts.* Genome Res, 2014. **24**(6): p. 999-1011.
36. Carty, M., et al., *An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data.* Nat Commun, 2017. **8**: p. 15454.
37. Crane, E., et al., *Condensin-driven remodelling of X chromosome topology during dosage compensation.* Nature, 2015. **523**(7559): p. 240-4.
38. Ma, W., et al., *Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes.* Nat Methods, 2015. **12**(1): p. 71-8.
39. Sanborn, A.L., et al., *Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes.* Proc Natl Acad Sci U S A, 2015. **112**(47): p. E6456-65.
40. Fullwood, M.J. and Y. Ruan, *ChIP-based methods for the identification of long-range chromatin interactions.* J Cell Biochem, 2009. **107**(1): p. 30-9.
41. Cairns, J., et al., *CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data.* Genome Biol, 2016. **17**(1): p. 127.
42. Knight, P.A. and D. Ruiz, *A fast algorithm for matrix balancing.* Ima Journal of Numerical Analysis, 2013. **33**(3): p. 1029-1047.