QuASAR: Quality Assessment of Spatial Arrangement Reproducibility in Hi-C Data

Michael EG Sauria[1,†] and James Taylor[1,†]

[1]Departments of Biology and Computer Science, Johns Hopkins University, Baltimore MD 21218
[†]Correspondence to MEGS (mike.sauria@gmail.com) and JT (james@taylorlab.org)

Abstract

Hi-C has revolutionized global interrogation of chromosome conformation, however there are few tools to assess the reliability of individual experiments. Here we present a new approach, QuASAR, for measuring quality within and between Hi-C samples. We show that QuASAR can detect even tiny fractions of noise and estimate both return on additional sequencing and quality upper bounds. We also demonstrate QuASAR's utility in measuring replicate agreement across feature resolutions. Finally, QuASAR can estimate resolution limits based on both internal and replicate quality scores. QuASAR provides an objective means of Hi-C sample comparison while providing context and limits to these measures.

Keywords

chromosome conformation, Hi-C, quality, replicate, reproducibility

Background

Chromosome Conformation Capture assays, particularly Hi-C, are becoming widely used, and have enabled a much better understanding of chromatin spatial interactions and architecture. What has not kept pace is the ability to assess the reliability of these data. One dimensional genomic annotation assays, such as ChIP-seq, have been the focus of quality control efforts for years[1–6]. These methods take a variety of approaches for assessing sample quality including external control samples, orthogonal datasets, replicate comparison, and mapping statistics. This allows direct comparison between similar datasets and a degree of confidence in the biological conclusions associated with the results of analyzing these datasets. The field of chromatin topology would benefit from a similar set of quality assessment resources.

Three areas need to be addressed when considering Hi-C data quality. Two of these, individual sample quality and replicate agreement, are also of great importance in other genomic annotation data types. The final area, limits of data resolution, is a consideration in other types of genomic data but has a special importance in Hi-C data because of the two dimensional nature of the data compared to more traditional one dimensional annotation data. For example, the amount of sequencing required to achieve saturation in a mammalian ChIP-seq experiment is about 40 million reads [7]. While not trivial, this depth of sequencing is often achieved. It is currently unknown what the saturation point of a Hi-C experiment is, but given that most Hi-C samples rarely have more than a couple hundred million reads prior to mapping and filtering, it is unlikely that they approach the saturation point for 2D dataset. Thus it is important to be able to determine reasonable limits for analysis resolution.

A variety of approaches have been use to infer Hi-C sample quality, primarily focused on statistics derived from sequencing quality, read alignment and the position of read ends relative to restriction fragments [8–14]. Sequencing quality and alignment quality give some information about sample integrity and possible contamination. Rao et al. [11] proposed a complexity statistic to determine which Hi-C libraries were worth sequencing. This statistic was based on the percent of unique reads in the sequenced sample. Other groups have pre-screened samples using fragment size profiles as an indication of Hi-C library quality prior to sequencing [12, 13]. While these approaches can provide information about the utility of sample processing, they fail to provide information about how well the Hi-C library captures chromatin conformations reflective of the ground truth. Statistics derived from aligned reads such as PCR duplication rates, self-ligation of restriction fragments, or the percentage reads with inserts of expected size can also be useful in determining the percentage reads in a Hi-C library that have been correctly processed, and it is crucial to perform filtering based on these features. They do not, however, provide information about the actual conformational quality of reads passing these filters, only that the reads conform to a set of expected characteristics based on the Hi-C protocol. For example, a negative control sample lacking the cross-linking step would produce a Hi-C library with the a similar percentage of reads passed to downstream analysis as a normally processed Hi-C sample. The one statistic that has been put forth to directly address sample quality is the ratio of intra- to inter-chromosomal reads [15]. If fragments are randomly ligated, rather than based on proximity, the number of inter-chromosomal interactions would increase, lowering this ratio. The shortcoming of this measure is that it does not necessarily provide information about the quality of intra-chromosomal reads, nor does it account for possible biologically relevant drivers of increased inter-chromosomal interactions such as chromatin decondensation or release from lamina associated domains.

The second important consideration in Hi-C data quality is reproducibility between samples. A primary means of establishing the similarity between two Hi-C samples has been measuring correlation, either Spearman or Pearson, between heatmaps at a chosen resolution [11, 15–19]. The largest drawback to this approach is that the strongest driver of Hi-C signal is genomic distance between loci, regardless of biological interactions [20]. This means that all Hi-C datasets have the same strong underlying distance-based signal decay driving correlation and differences between samples are small by comparison and harder to detect. Two methods have been devised specifically to address this

problem, HiCRep [21] and HiC-Spector [22]. The authors of HiCRep demonstrate that heatmap correlation measures are insufficient at consistently distinguishing unrelated samples from each other compared to biological replicates. This indicates that correlation-based assessments of sample similarity are unreliable. Both of these approaches provide a single summary statistic for the reproducibility between samples, although HiCRep also provides feedback about saturation of the reproducibility measure as a function of sequencing depth.

The final consideration for Hi-C quality assessment is determining what levels of resolution are appropriate for a dataset given its sequencing depth, quality, and reproducibility. Rao et al. [11] proposed a measure of resolution based on minimum number of contacts for some percentage of bins produced by partitioning the genome at a given resolution. This intuitively makes sense from the standpoint of having sufficient data density to resolve features at a given resolution. However, consider a pair of samples with equal numbers of reads, one with a large proportion of random ligation: assuming similar marginal distributions of reads, both samples would have identical resolution limit statistics but different sample qualities at the same resolution.

Here we present a new approach for measuring Hi-C data quality, both within and between samples, using a technique called Quality Assessment of Spatial Arrangement Reproducibility (QuASAR). Using a combination of matrix transformations and sub-sampling, we show that QuASAR not only provides information about a sample's quality and agreement between replicates, but also estimates for return on additional sequencing, absolute quality limits, and an estimate of the maximum reliable resolution that can be used for a Hi-C sample. Together, these results show that QuASAR can facilitate optimized choices in Hi-C data production as well as informed data comparisons and analysis parameter selection.

Results

Spatial consistency concept and matrix transformation strategy

To quantify Hi-C quality, we consider the consistency of inferred spatial arrangement of the Hi-C intra-chromosomal (*cis*) data. Initially, the genome is partitioned into uniform-sized bins at a chosen resolution. For bins that occur close together in space as determined by their read count, there should be a high correlation between their sets of *cis*interactions (Figure 1A). Conversely, bins occurring further apart should show little or no correlation across interactions. Thus, for any given pair of bins, we can identify disagreement between the direct and inferred measure of their interaction. For each sample, we produced a "QuASAR-transformed matrix" by finding the element-wise product of the read count matrix and the local correlation matrix as calculated from a distance-corrected enrichment matrix (Figure 1B). Transformed matrices are calculated across multiple resolutions to examine consistency of different features and scales. In order to target features appropriate to a given resolution, we limit the maximum interaction genomic distance used for analysis as a function of the bin size. This includes in the calculation of correlations, thus the term "local correlation". The resulting transformed matrices can then be used to calculate individual sample quality scores and replication scores for pairs of samples.

The primary drivers of low quality are random ligation products and missing expected interaction fragments. Within the QuASAR transformed matrix, these types of noise appear as individual entries showing deviation from their surroundings (Figure 1C), higher than local background in the case of random ligation and lower for missing reads. A third factor that may impact the quality score is population heterogeneity, which will manifest as a compression of the dynamic range of signal and less differentiation between the correlation and transformed matrices.

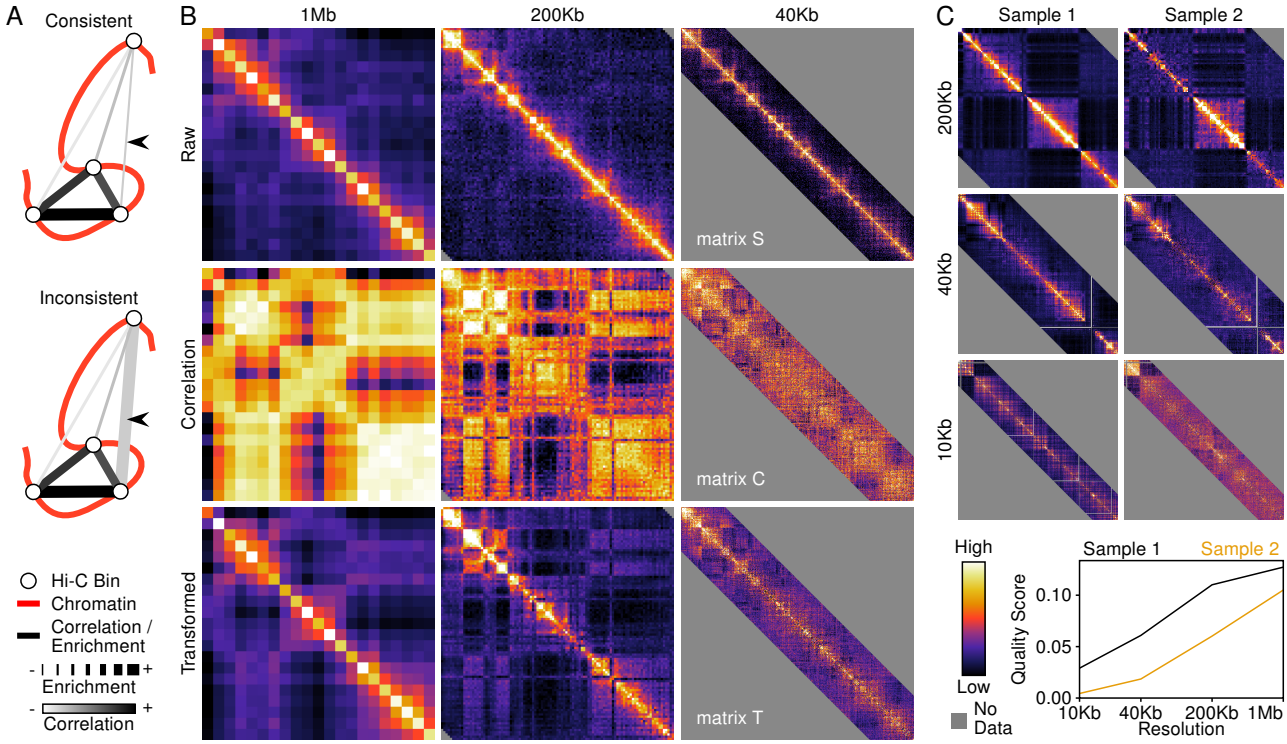

Figure 1: Using spatial consistency between local and regional signals to identify Hi-C sample quality. A) Sample reads are partitioned into bins and pairwise assessment of interaction strength and bin correlation are compared to identify inconsistencies (black arrows). B) Multiple resolutions are considered by QuASAR using transformed matrices derived from local correlation matrices weighted by interaction enrichment. C) Two samples derived from mouse ES cells of differing quality. Lower quality manifests as random points indicating non-specific ligation.

In order to determine an individual sample's quality, we find the mean of the correlation matrix (C) weighted by the read count matrix (R) minus the unweighted mean (Figure 1B):

$$\text{Quality}_k = \frac{\sum_{i=0}^{|C_k|-1} \sum_{j=i+1}^{i+100} t_{kij} I(k,i,j)}{\sum_{i=0}^{|C_k|-1} \sum_{j=i+1}^{i+100} r_{kij} I(k,i,j)} - \frac{\sum_{i=0}^{|C_k|-1} \sum_{j=i+1}^{i+100} c_{kij} I(k,i,j)}{\sum_{i=0}^{|C_k|-1} \sum_{j=i+1}^{i+100} I(k,i,j)}$$

where for bins i and j on chromosome k, c_{kij} is the correlation from matrix C_k, r_{kij} is the raw enrichment value from matrix R_k, and t_{kij} is the transformed value from the QuASAR transformed matrix T_k. I is an indicator function taking on a value of one or zero for valid and invalid correlations, respectively.

Replicate scores were determined by finding the correlation of valid values from the transformed matrices (T_k) for chromosome k between two samples:

$$\text{Replicate}_k = \text{Corr}\left(\left\{T_{Akij} | I_A(k,i,j) = 1, I_B(k,i,j) = 1\right\}, \left\{T_{Bkij} | I_A(k,i,j) = 1, I_B(k,i,j) = 1\right\}\right)$$

where I_A and I_B are the indicator functions for samples A and B, respectively. In order to find sample-wide scores, summations and correlations were calculated across all chromosome matrices simultaneously.

QuASAR evaluation strategy

In order to assess the effectiveness of QuASAR, we used three separate testing approaches. First, we generated noise models for different genome/restriction enzyme combination and created datasets with varying percentages of reads drawn from these models. Second, we used combinations of reads from two separate datasets, either biological replicates in order to generate pseudo-replicates or unrelated samples to create heterogeneous population samples. Finally, we used sub-sampling to investigate the effects of numbers of reads. This was done using *cis*reads that had already passed initial mapping and circularization filters. Because this does not, strictly speaking, qualify as sequencing depth, we refer to the number of valid *cis*reads as "coverage".

To ensure our results were robust, we tested 96 samples across three species, Mouse, Human, and Drosophila melanogaster (Table S1). Samples ranged in coverage from less than 1 million to 185 million reads. All samples were paired biological replicates and were generated from a diverse set of tissues and cell lines, and originated from numerous laboratories.

QuASAR Quality results

QuASAR quality scoring effectiveness was tested using injection of simulated random ligation noise, at levels of 0.1% to 75% relative to read coverage. All samples showed a monotonically decreasing relationship between quality signal and percent noise, with the exception of eight instances (Figure 2A). Exceptions to this relationship all occurred at the lowest level of resolution and the majority occurred at under 0.5% noise, with all of them occurring at 5% noise or less. The highest deviating value as a percentage of the raw sample score, was 100.02273%. Because all of these exceptions occurred at low resolution, which is less responsive to coverage and noise effects, and because increases in quality score were minimal, it is likely that these represent stochastic noise. At all but the lowest resolution, as little as 0.1% noise was detectable in every sample using the QuASAR quality score. These results suggest that this metric is sensitive to even small changes in the amounts of random ligation present in Hi-C samples.

In addition to noise injection, we also examined the effects of heterogeneity on sample scores. Pairs of samples with quality scores differing by less than 1% at the 1 Mb resolution were mixed in varying ratios at 20 million read coverage. At all resolutions, samples showed decreased quality scores when part of a mixture, indicating that QuASAR is sensitive to heterogeneity as well as noise (Figure 2B). Thus, QuASAR quality scores not only detect noise but also the dilution of spatial consistency by superimposition of multiple disparate configurations.

Next, we examined the effects of coverage on quality scoring. As coverage decreased, quality scores decreased in a highly consistent manner for all samples and resolutions (Figure 2C). All quality score vs. log-transformed coverage relationships (for samples with at least 8 million reads for mouse and human samples, 1 million reads for Drosophila) were fit using logistic curves. This suggests that additional sequencing has diminishing returns on quality. Further, there exists some upper limit of quality for each sample. We also find that, consistent with expectations, the amount of coverage necessary to approach this quality asymptote increases with increasing resolution. In other words, a sample requires much less coverage to fully resolve large-scale conformational features in a consistent and high-quality manner. The X-axis offset of different samples also indicates that there should not be an arbitrary guideline for target coverage to resolve a particular resolution as the quality response to coverage is tissue and assay-specific. However, a minimum coverage based on currently tested samples may be appropriate.

Despite different resolutions plateauing at different rates, quality scores showed good agreement in sample ranking across resolutions (Figure S1A). Ranking was more consistent between more similar resolutions, indicating a greater overlap in the features being assessed compared to larger jumps in resolution. Because quality is a function of coverage, we also examined sample rankings at a uniform coverage level to remove any confounding effects. In most cases, the sample rankings still showed good agreement, although the consistency did decrease (Figure S1B).

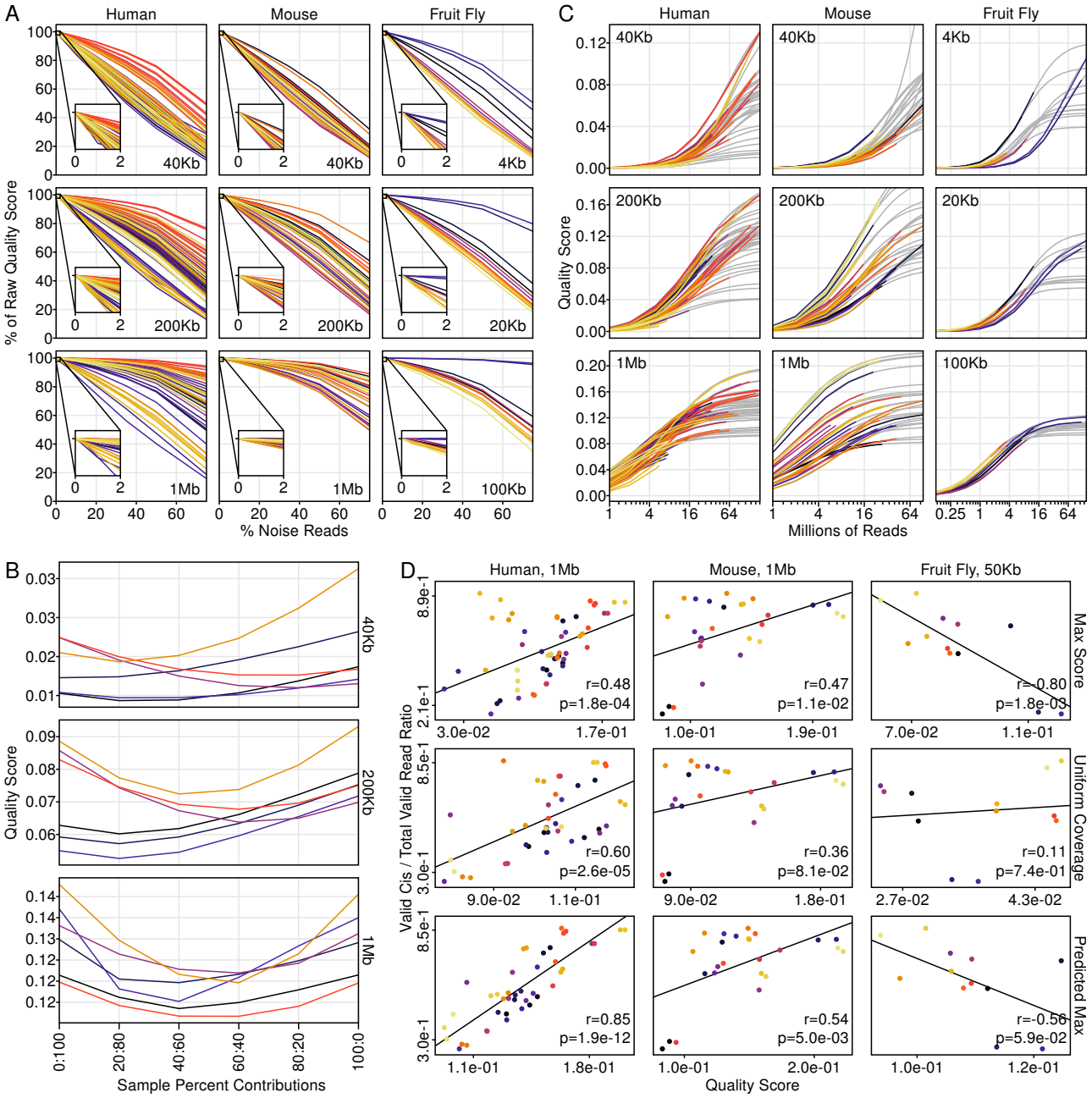

Figure 2: QuASAR quality scoring is sensitive to noise and coverage differences. A) Samples show an almost entirely negative relationship between injected noise and quality score. All scores are reported as a percentage of their unaltered sample score. The insets show a detailed view of values for noise levels between 0 and 2%. All samples within a species are colored consistently across resolutions. B) QuASAR quality scores for mixed samples with varying levels of contribution from each sample pair. All samples have 20 million *cis*reads. C) All quality scores show a monotonically increasing relationship to coverage. Gray lines indicate logistic curves fit to each sample. All samples within a species are colored consistently across resolutions. D) The relationship between quality scores and the percentage intra-chromosomal reads (*cis*) out of all valid reads is depicted. Quality scores were assessed at three different points: the score at maximum sample coverage (top); the score at a uniform coverage across all samples (middle, 10 million reads for human and mouse, 1 million reads for Drosophila); the modeled quality score at infinite coverage (bottom, only samples with at least 4 million reads in human and mouse or 1 million reads in Drosophila). For each plot, the Pearson correlation coefficient and associated p-value are shown.

Interestingly, the mouse-derived samples showed an increase in correlation across resolutions after accounting for coverage.

We also examined how QuASAR quality scores related to various read statistics that have been used as proxies for sample quality. As observed in the coverage analysis, samples do not plateau at the same rate as a function of coverage. This means that a simple comparison across sample qualities at a given resolution may be misleading. In order to address this, we compared read statistics to three different quality reference points: the quality score at each sample's actual coverage; the quality score at a uniform coverage level; and the inferred quality limit as determined by logistic curve fitting. Although we compared scores to four different read statistics, only the percentage of *cis*reads out of all valid reads showed a consistent and strong relationship to QuASAR quality scores (Figure 2D). For human and mouse samples, all three sets of quality scores showed strong correlation to the *cis*read ratio. However, the quality limit scores performed significantly better than the other two sets. The Drosophila samples did not show this same pattern and in fact negatively correlated to the *cis*read ratio. It is unclear why this pattern did not hold for Drosophila samples, although the number of samples is much lower compared to either mouse or human sample sets. The three other read statistics tested, percentage of reads with an insert size too large, percentage of reads from fragment circularization, and percentage of reads from putative missed restriction cuts showed no consistent relationship across any of the quality sets (Figure S2).

Reproducibility results

To determine the performance of QuASAR replicate scoring, we began by finding replicate scores as a function of noise. In order to assess the effects of noise, we calculated a replicate score for each sample across varying resolutions and levels of noise injection into its matched biological replicate. The majority of samples showed a monotonically decreasing relationship between replicate score and noise level (Figure 3A). About 15% of sample-resolution relationships did not strictly hold to this trend. In all of these cases, the scores hovered around the noise-free replicate score before decreasing. Of these, most (35 of 42) occurred at the lowest level of resolution. None of these score fluctuations was over 1% above the noise-free replicate score. These results demonstrate the robustness of QuASAR replicate scores to noise, particularly when comparing macro features (low resolution).

Next, we examined how coverage impacted replicate scoring. For each biological replicate sample pair, samples were down-sampled to various equal numbers of reads and replicate scores were calculated across multiple resolutions. In all sample pairs, replicate scores increased as a function of coverage following a logistic curve and ranging between zero and one (Figure 3B). Every sample pair showed a rightward shift of the curve midpoint as resolution increased, indicating that replicate agreement for macro features, such as compartments, occurred at lower coverage levels than fine-scale features, such as topologically associating domains or loops.

Finally, we calculated replicate scores for all pairwise combinations of samples within each species set across multiple resolutions. For human and mouse samples, pairs were scored at 1 Mb, 200 Kb, 40 Kb, and 10 Kb resolutions while Drosophila samples were scored at 100 Kb, 20 Kb, and 4 Kb (Figure S3). For all sample pairs, we took the highest score across all resolutions. The majority of biological replicate sample pairs scored close to one, the maximum replicate reproducibility score, indicating strong agreement in Hi-C signal between samples (Figure 4A-C). For all replicate pairs we also generated a pseudo-replicate sample composed of half of all reads from each replicate, randomly sampled and combined. Scores between samples and their pseudo-replicates were always higher than true biological replicates. In nearly all cases unrelated sample pair scores showed clear separation from replicate scores. There were two exceptions to this: samples with low biological replicate scores, and human ES and mesendoderm cells. The latter may be because the mesendoderm cells were differentiated from the ES cell line and either retain a strong conformational resemblance or differentiation was incomplete. We also observed elevated reproducibility scores for samples derived from the same tissue or cell line but from unrelated experiments. These samples typically fell in between the range of unrelated and biological replicate pair scores. Two exceptions to this pattern were

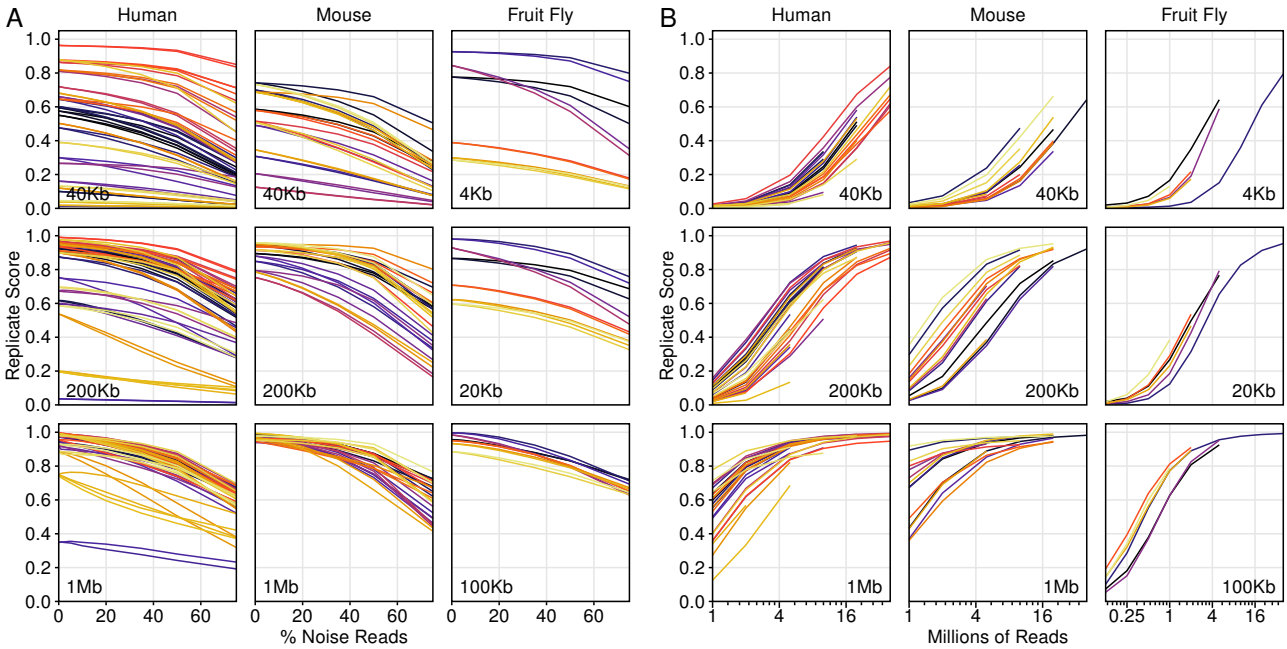

Figure 3: QuASAR replicate scores are sensitive to both noise and coverage. A) Replicate scores were calculated for each pair of sample replicates between the noise-free first replicate and the noise-injected levels of the second replicate and vice versa for each resolution. Sample color coding is consistent within each species across resolutions. B) For each pair of replicates, replicate scores were calculated at each coverage level and resolution. Sample color coding is consistent within each species across resolutions.

human embryonic stem cells and mouse primary fetal liver cells, both of which showed reproducibility scores at or above biological replicate scores. We also observed that there was a relationship between sample quality scores and biological replicate scores such that if at least one of the replicates was of lower quality, the replication score was lower (Figure 4D).

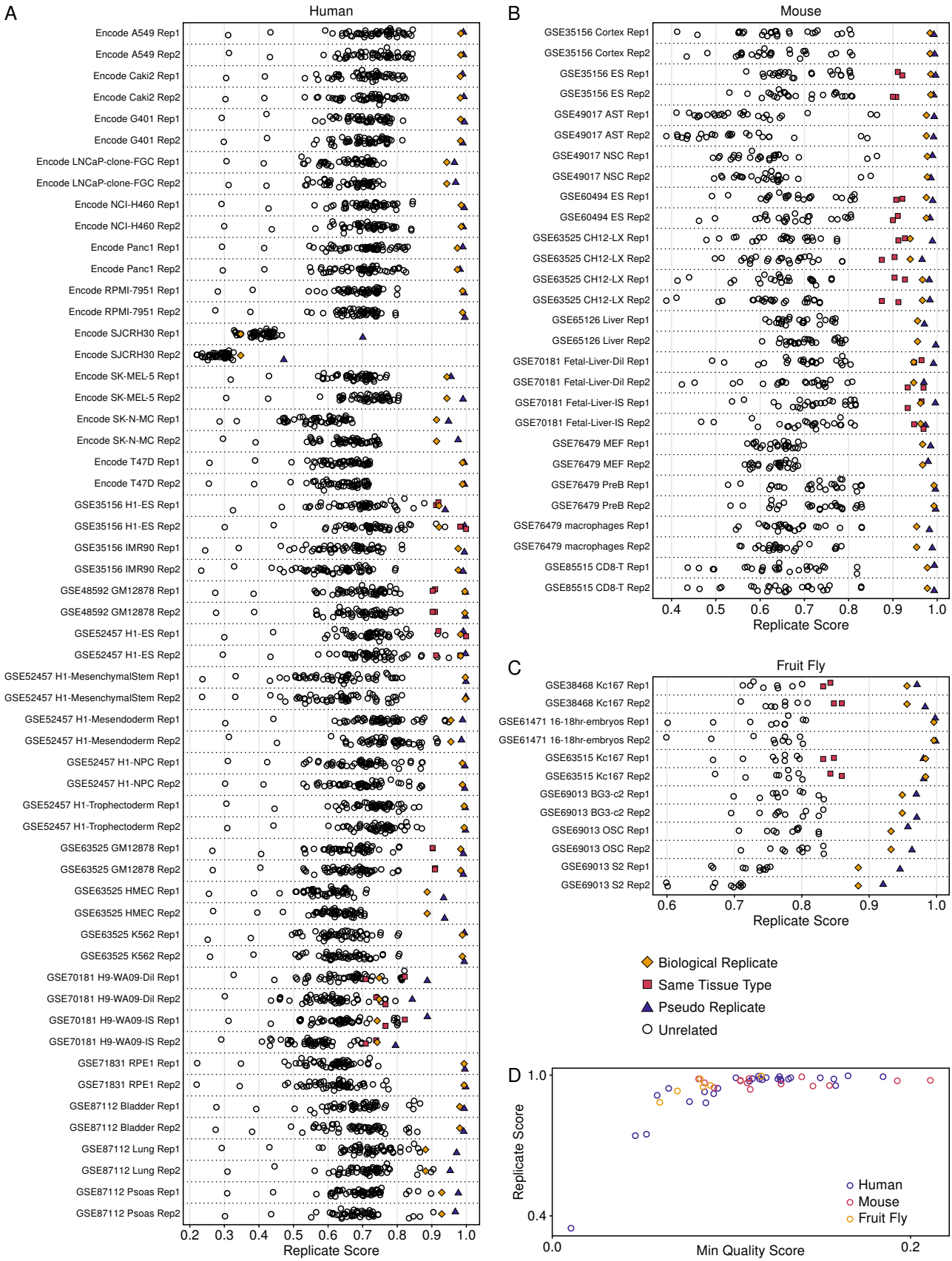
A
Human
Encode A549 Rep1
Encode A549 Rep2
Encode Caki2 Rep1
Encode Caki2 Rep2
Encode G401 Rep1
Encode G401 Rep2
Encode LNCaP-clone-FGC Rep1
Encode LNCaP-clone-FGC Rep2
Encode NCI-H460 Rep1
Encode NCI-H460 Rep2
Encode Panc1 Rep1
Encode Panc1 Rep2
Encode RPMI-7951 Rep1
Encode RPMI-7951 Rep2
Encode SJCRH30 Rep1
Encode SJCRH30 Rep2
Encode SK-MEL-5 Rep1
Encode SK-MEL-5 Rep2
Encode SK-N-MC Rep1
Encode SK-N-MC Rep2
Encode T47D Rep1
Encode T47D Rep2
GSE35156 H1-ES Rep1
GSE35156 H1-ES Rep2
GSE35156 IMR90 Rep1
GSE35156 IMR90 Rep2
GSE48592 GM12878 Rep1
GSE48592 GM12878 Rep2
GSE52457 H1-ES Rep1
GSE52457 H1-ES Rep2
GSE52457 H1-MesenchymalStem Rep1
GSE52457 H1-MesenchymalStem Rep2
GSE52457 H1-Mesendoderm Rep1
GSE52457 H1-Mesendoderm Rep2
GSE52457 H1-NPC Rep1
GSE52457 H1-NPC Rep2
GSE52457 H1-Trophectoderm Rep1
GSE52457 H1-Trophectoderm Rep2
GSE63525 GM12878 Rep1
GSE63525 GM12878 Rep2
GSE63525 HMEC Rep1
GSE63525 HMEC Rep2
GSE63525 K562 Rep1
GSE63525 K562 Rep2
GSE70181 H9-WA09-Dil Rep1
GSE70181 H9-WA09-Dil Rep2
GSE70181 H9-WA09-IS Rep1
GSE70181 H9-WA09-IS Rep2
GSE71831 RPE1 Rep1
GSE71831 RPE1 Rep2
GSE87112 Bladder Rep1
GSE87112 Bladder Rep2
GSE87112 Lung Rep1
GSE87112 Lung Rep2
GSE87112 Psoas Rep1
GSE87112 Psoas Rep2
0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
Replicate Score
B
Mouse
GSE35156 Cortex Rep1
GSE35156 Cortex Rep2
GSE35156 ES Rep1
GSE35156 ES Rep2
GSE49017 AST Rep1
GSE49017 AST Rep2
GSE49017 NSC Rep1
GSE49017 NSC Rep2
GSE60494 ES Rep1
GSE60494 ES Rep2
GSE63525 CH12-LX Rep1
GSE63525 CH12-LX Rep2
GSE63525 CH12-LX Rep1
GSE63525 CH12-LX Rep2
GSE65126 Liver Rep1
GSE65126 Liver Rep2
GSE70181 Fetal-Liver-Dil Rep1
GSE70181 Fetal-Liver-Dil Rep2
GSE70181 Fetal-Liver-IS Rep1
GSE70181 Fetal-Liver-IS Rep2
GSE76479 MEF Rep1
GSE76479 MEF Rep2
GSE76479 PreB Rep1
GSE76479 PreB Rep2
GSE76479 macrophages Rep1
GSE76479 macrophages Rep2
GSE85515 CD8-T Rep1
GSE85515 CD8-T Rep2
0.4 0.5 0.6 0.7 0.8 0.9 1.0
Replicate Score
C
Fruit Fly
GSE38468 Kc167 Rep1
GSE38468 Kc167 Rep2
GSE61471 16-18hr-embryos Rep1
GSE61471 16-18hr-embryos Rep2
GSE63515 Kc167 Rep1
GSE63515 Kc167 Rep2
GSE69013 BG3-c2 Rep1
GSE69013 BG3-c2 Rep2
GSE69013 OSC Rep1
GSE69013 OSC Rep2
GSE69013 S2 Rep1
GSE69013 S2 Rep2
0.6 0.7 0.8 0.9 1.0
Replicate Score
Biological Replicate
Same Tissue Type
Pseudo Replicate
Unrelated
D
1.0
0.4
Replicate Score
0.0
0.2
Min Quality Score
Human
Mouse
Fruit Fly

Figure 4: QuASAR replicate scores reflect both sample quality and consistency between sample origins. A-C) Replicate scores are denoted as points for each sample with an all to all pairwise comparison scheme within each set of species samples. Pairs include true replicates (gold diamonds), pseudo-replicates (purple triangles), same tissue of origin but non-replicates (fuchsia squares), and unrelated samples black circles). D) For each replicate pair, the lower of the two quality scores is plotted against the pair's replicate score.

Determining maximum resolution

One of the key choices to be made in using Hi-C data is determining an appropriate resolution for analysis. To answer this rigorously we propose a combination of quality and replicate scores. We tested replicate score cutoffs ranging from 0.75 to 0.99, using these to classify samples as passing or failing for each resolution (10 Kb, 40 Kb, 200 Kb, and 1 Mb for mouse and human; 4 Kb, 20 Kb, and 100 Kb for Drosophila; Figure 5A). We then used these labels and the associated quality scores to determine the best quality score cutoff as determined by the Gini impurity scores. These scores were then used to classify replicate pairs, with mixed labels (one passing and one failing replicate) being classified as failing. The same strategy was used to find a new optimal replicate cutoff score. The process was repeated until stable cutoffs were found for both replicate and quality scores. We found that across this range of initial replicate cutoffs, there were four sets of stable cutoff values (Figure S4A). We selected the pair of cutoffs with the lowest combined Gini impurity scores as our stringent cutoffs and the second lowest scoring pair as our loose cutoffs. For both cutoff sets, samples were partitioned into two groups with distinct distributions (Figures 5B & S4B) We then estimated resolution limits for each sample based on quality and replicate scores. For each score type, the resolution limit was defined as the point at which the log-transformed resolution vs. score curve equaled the score cutoff (Figure 5A). To further validate our cutoff values, we compared estimated resolution limits determined from quality and replicate analyses. Estimates from the two measures showed significant agreement for both cutoff sets (Figures 5C & S4C). The resolution limits also matched with a visual assessment of the data such that features were resolvable by eye at and above the resolution limits but were sparse or absent below (Figure 5D).

Discussion

We have demonstrated the variability of Hi-C dataset quality across a variety of measures including resolution, coverage, and heterogeneity. Accurate quantification of this variability is necessary to make informed dataset selection and analysis choices. One of the primary strengths of QuASAR is its ability to provide resolution-specific information, which will benefit both producers and consumers of Hi-C data. It is currently difficult, if not impossible, to gauge the sequencing depth required to achieve a target level of resolution for analysis. QuASAR can provide crucial information for estimating the number of reads for a specific library necessary to produce high-confidence results at that target resolution by determining both the underlying library quality and the quality return on sequencing. For end users of Hi-C data, the redundancy of Hi-C datasets for specific cell lines or tissue types presents a challenge in selecting the most suited data for their analysis goals. An objective quality measure allows comparison of Hi-C protocols to evaluate strengths and weaknesses with respect to data quality. In addition to individual sample quality, the ability to find sample similarity between replicates or unrelated samples is of interest to the Hi-C community. Two recent publications describe tools to do just that [21, 22]. While both appear to function well, their outputs, while quantitative, give a binary-feeling classification of same/different. As we demonstrate in this study, there may be more nuance to sample comparisons (Figure 4). It is important not only to show the similarity of samples but to provide context such as at which resolutions the samples appear similar and are they similar enough to provide reliability to analyses. This is especially important given the prevalence of cell line-derived Hi-C data and the heterogeneity and instability of cell line genomes [23, 24]. In this study, we observed that Hi-C sample similarity was lower between cell line-derived datasets produced in different labs than between replicates, even at the lowest resolution of analysis. In

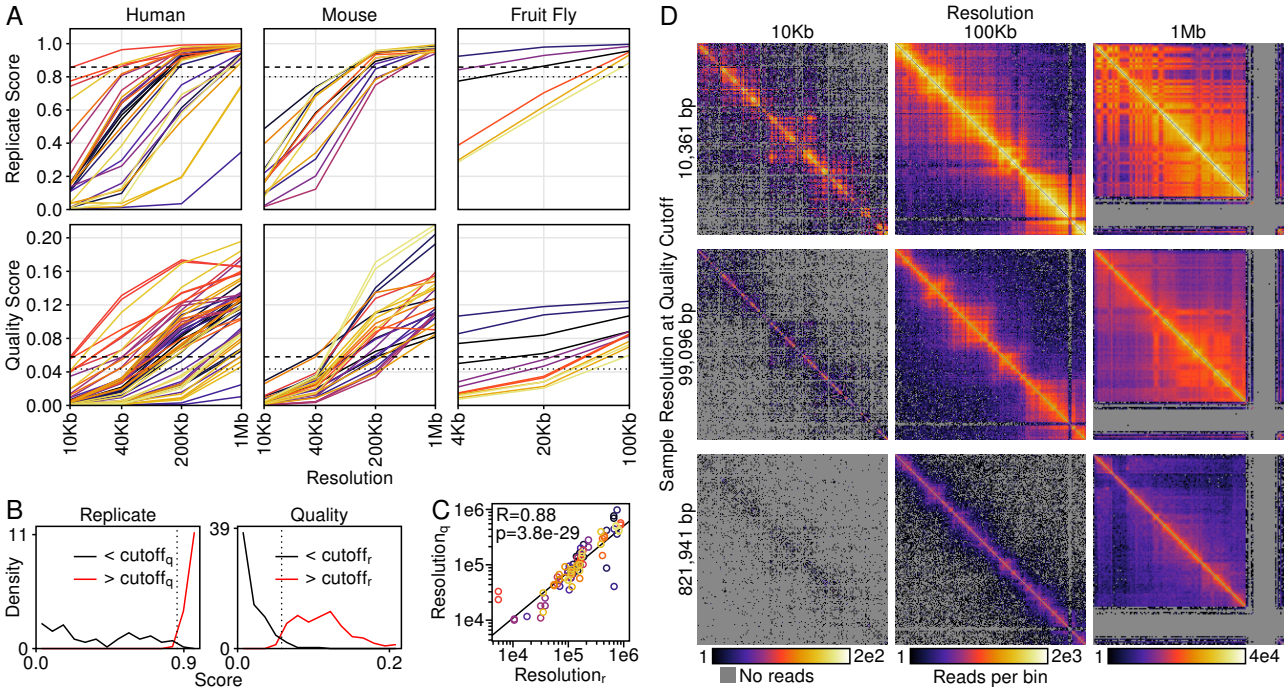

Figure 5: Calculating maximum usable resolutions from QuASAR scoring. A) QuASAR replicate (top) and quality (bottom) scores plotted as functions of log-transformed resolutions. Sample color-coding is consistent within each species plot pairs. Stringent and loose cutoff values are denoted by dashed line and dotted lines, respectively. B) Replicate (left) and quality (right) score distributions, partitioned by stringent quality and replicate cutoffs, respectively. Dotted lines donate the best separation point between partitions based on Gini impurity scoring. C) Resolution limits for each sample determined by stringent replicate and quality cutoffs. The Pearson R value and associated p-value are shown. D) Three samples (rows), selected for resolution limits close to target resolutions 10 Kb, 100 Kb, and 1 Mb, are shown binned at each target resolution (columns).

many cases these scores approached the level of similarity seen for unrelated samples. This may serve as a cautionary tale about mixing dataset origins without verifying their similarity. Thus, to get the best return from time and financial investments in Hi-C data, it is important to evaluate the data critically prior to drawing conclusions. To this end, QuASAR provides an objective means of comparison of both individual samples and replicate agreement while providing context and limits to these measures.

Methods

Hi-C data processing and normalization

Hi-C raw read data were obtained from the Sequence Read Archive (SRA; `https://www.ncbi.nlm.nih.gov/sra`) or, in the case of ENCODE data, the ENCODE Data Portal (`https://www.encodeproject.org`; Table S1). Read ends were aligned using BWA mem version 0.7.12-r1039 and default settings [25] to the appropriate genome build (Table S1). Reads were considered valid and retained if both ends uniquely mapped to single locations, or at least one end spanned a ligation junction and mapped uniquely to two restriction fragments. In cases where both ends mapped to multiple fragments, reads were only kept of the upstream and/or downstream fragment locations matched across read ends. Reads were processed and normalized using HiFive version 1.3.2 [26]. Distance-dependent signal curves were estimated using the settings shown in Table S1. A maximum insert size of 650 bp was used to filter reads. Fends (fragment ends) were filtered to have a minimum of one valid interaction greater than 500 kb. Fend interaction filtering was applied only for the normalization step. All quality analyses were performed on unfiltered reads.

Data were normalized using the "binning" algorithm correcting for GC content, fragment length, and mappability. GC content was calculated from the 200 bp upstream of restriction sites or the length of the fragment, whichever was shorter. Mappability was determined using the GEM mappability function, version 1.315 [27]. Mappability of 36-mers was calculated every 10 bp with an approximation threshold of six, a maximum mismatch of 1 bp, and a minimum match of 28 bp. For each fend, the mean mappability score for the 200 bp upstream of the restriction site, or total fragment size if smaller, was used. For normalization, only intra-chromosomal reads with an interaction distance of at least 500 kb were used. GC content and fragment length were partitioned into 20 bins each and mappability was partitioned into 10 bins. All parameter partitions were done such that together they spanned the full range of values and contained equal numbers of fends in each bin. All bins were seeded from raw count means and GC and length parameters were optimized for up to 100 iterations or until the change in log-probability was less than one, whichever was achieved first. Normalization was used only for noise model construction.

QuASAR matrix transformation and scoring

All intra-chromosomal raw reads were binned at a predefined resolution (1 Mb, 200 Kb, 40 Kb, or 10 Kb for mouse and human datasets, 100 Kb, 20 Kb, or 4 Kb for Drosophila datasets), depending on the analysis. Only numbered chromosomes and the X chromosome were used for analysis in human and mouse datasets while chromosomes 2L, 2R, 3L, 3R, 4, and X were used for fly datasets. For each chromosome, only rows and columns that had at least one read for an interaction occurring over a span of 100 bins or fewer were included. All other rows and columns were marked as invalid. The resulting matrix is defined as R. A scaled matrix, S, was calculated as the square-root of the sum of matrix R plus one. A distance-normalized matrix, N, was calculated by dividing each diagonal of R by the sum of the diagonal divided by the sum of valid rows/columns in that diagonal. A local correlation matrix, C was then calculated from R such that for each pairwise set of rows no more than 100 rows apart, the correlation was calculated between valid column entries not more than 100 columns from either row number, not including self-interactions or the interaction between the pair of bins being correlated:

$$c_{ij} = \mathrm{corr}(n_{ij,local}, n_{ji,local})$$
$$n_{ij,local} = \{n_{ik} \mid j - 100 \leq k < i + 100, k \neq i, k \neq j, I(k) = 1\}$$

where $I(k)$ is an indicator function taking on the value of one for valid rows/columns, and zero otherwise. Entries in matrix C were considered valid if there were at least three points in the correlation and the standard deviation of both $n_{i,local}$ and $n_{j,local}$ were non-zero. The QuASAR transformed matrix, T was the element-wise product of S and C.

QuASAR quality scores were calculated as the sum of valid elements of the transformed matrix T divided by the sum of the corresponding elements of S minus the mean of the corresponding elements of C. Thus, with the indicator function $I(i,j)$ taking a value of one for valid element c_{ij} and zero for an invalid element, the Quasar quality score is defined as follows:

$$\mathrm{Quality}_{chrom} = \frac{\sum_{i=0}^{|C|-1} \sum_{j=i+1}^{i+100} t_{ij} I(i,j)}{\sum_{i=0}^{|C|-1} \sum_{j=i+1}^{i+100} s_{ij} I(i,j)} - \frac{\sum_{i=0}^{|C|-1} \sum_{j=i+1}^{i+100} c_{ij} I(i,j)}{\sum_{i=0}^{|C|-1} \sum_{j=i+1}^{i+100} I(i,j)}$$

This score was calculated on a per chromosome basis. A global quality score was calculated by finding each of these sums across across all chromosomes prior to division.

QuASAR replicate scores were calculated by finding the correlation of the two sample transformation matrices T for all elements that are valid in both matrices. These scores were calculated on a chromosome by chromosome basis as well as a global score across all valid matrix elements for all chromosomes.

Hi-C noise and low-coverage modeling

The noise model employed was based on all noise coming from random ligation. For each genome and restriction enzyme combination, the median bin correction values across all relevant datasets were used to calculate expected values for all bins at the lowest resolution used for analysis (10 Kb for human and mouse data, 4 Kb for Drosophila data). Bins that had zero observed reads in any of the used datasets were set to zero. Expected values were converted into probabilities by dividing values by the sum of all expected values. For noise-injected sample, a random fraction of reads corresponding to the target noise percentage were randomly selected and discarded. The same number of reads were then drawn from the above described noise distribution and combined with the remaining sample reads. This was done prior to filtering or normalization.

Low coverage samples were generated by random selection and removal of reads prior to any filtering or normalization.

Pseudo-replicate and heterogeneous sample generation

Pseudo-replicates were generated by random selection of reads from each replicate across all chromosomes prior to filtering or normalization. For each replicate, half of the reads were selected and combined, meaning that pseudo-replicates had a number of reads equal to the mean of the two replicates. Heterogeneous samples were produced the same way, although the percentage of reads drawn from each sample was varied from 0 to 100% at 20% steps.

Coverage curve estimation

QuASAR quality curves as a function of coverage were calculated using the curve_fit function from the python package SciPy [28]. All lines were estimated using the logistic function as follows:

$$y = \frac{A}{1 + k\frac{x_0}{x}}$$

where A is the quality score upper bound, x_0 is the inflection point coverage level, and k is a scale factor. Initial values for each sample were set to twice its maximum quality score, its maximum coverage, and 0.5 for A, x_0, and k, respectively. A was bounded between -2 and 2, while the other parameters had no limits.

Resolution cutoff calculation

Resolution cutoffs were calculated in an iterative fashion. Initial resolution cutoff values ranged from 0.750 to 0.990 at steps of 0.005. Each sample/resolution combination was labeled passing if its associated replicate score was greater than the initial resolution cutoff. Quality scores were then ordered and potential quality cutoff points were defined as the midpoints between adjacent quality scores. For each potential quality cutoff, the sum of the Gini impurity scores for the two partitions of samples (above and below quality cutoff) was calculated based on replicate cutoff labels, weighted by the number of samples in each partition. This quality cutoff was then used to partition associated replicate/resolution pairs. For all replicates, the lower quality score was used for label determination. The same procedure was followed for finding the new replicate cutoff value as described for the quality cutoff value. This process was repeated until both replicate and quality cutoff values remained constant. Maximum resolution limits were calculated based on quality or replicate curves as a function of the log-transformed resolution. For each sample or replicate pair, the resolution associated with the point at which the quality or replicate cutoff value, respectively, fell on the curve was used as the maximum resolution limit.

Availability of data and materials

The datasets analyzed during the current study are available at `https://bx.bio.jhu.edu/data/quasar`. Analysis code and results are available at `https://github.com/bxlab/Quasar_PaperAnalysis`. QuASAR is packaged as part of the HiFive suite of tools (`https://github.com/bxlab/hifive`).

Acknowledgements

Thanks to all members of the Taylor lab for useful discussion and feedback. Thanks to all of the members of the ENCODE 3D Nucleome subgroup for discussions on reproducibility and quality analyses. MEGS and JT were funded in part by NIH/NIDDK grant R24 DK106766 and NIH/NHGRI grant U41 HG006620.

References

1. Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J. & Snyder, M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22,** 1813–31. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (Sept. 2012). doi:`10.1101/gr.136184.111`.
2. Planet, E., Attolini, C. S., Reina, O., Flores, O. & Rossell, D. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28,** 589–90. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking) (Feb. 2012). doi:`10.1093/bioinformatics/btr700`.
3. Diaz, A., Nellore, A. & Song, J. S. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biol* **13,** R98. ISSN: 1474-760X (Electronic) 1474-7596 (Linking) (Oct. 2012). doi:`10.1186/gb-2012-13-10-r98`.
4. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front Genet* **5,** 75. ISSN: 1664-8021 (Print) 1664-8021 (Linking) (2014). doi:`10.3389/fgene.2014.00075`.
5. Marinov, G. K., Kundaje, A., Park, P. J. & Wold, B. J. Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)* **4,** 209–23. ISSN: 2160-1836 (Electronic) 2160-1836 (Linking) (Feb. 2014). doi:`10.1534/g3.113.008680`.
6. Qin, Q., Mei, S., Wu, Q., Sun, H., Li, L., Taing, L., Chen, S., Li, F., Liu, T., Zang, C., Xu, H., Chen, Y., Meyer, C. A., Zhang, Y., Brown, M., Long, H. W. & Liu, X. S. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. *BMC Bioinformatics* **17,** 404. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking) (Oct. 2016). doi:`10.1186/s12859-016-1274-4`.
7. Jung, Y. L., Luquette, L. J., Ho, J. W. K., Ferrari, F., Tolstorukov, M., Minoda, A., Issner, R., Epstein, C. B., Karpen, G. H., Kuroda, M. I. & Park, P. J. Impact of sequencing depth in ChIP-seq experiments. eng. *Nucleic Acids Research* **42,** e74. ISSN: 1362-4962 (May 2014). doi:`10.1093/nar/gku178`.
8. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. & Cavalli, G. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–72. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (Feb. 2012). doi:`10.1016/j.cell.2012.01.010`.

9. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–80. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (May 2012). doi:10.1038/nature11082.

10. Le Dily, F., Bau, D., Pohl, A., Vicent, G. P., Serra, F., Soronellas, D., Castellano, G., Wright, R. H., Ballare, C., Filion, G., Marti-Renom, M. A. & Beato, M. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev* **28,** 2151–62. ISSN: 1549-5477 (Electronic) 0890-9369 (Linking) (Oct. 2014). doi:10.1101/gad.241422.114.

11. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–80. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (Dec. 2014). doi:10.1016/j.cell.2014.11.021.

12. Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., Leeb, M., Wohlfahrt, K. J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A. J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M. G. S., Lehner, B., Di Croce, L., Wutz, A., Hendrich, B., Klenerman, D. & Laue, E. D. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **advance online publication.** ISSN: 1476-4687 (2017). doi:10.1038/nature21429.

13. Phanstiel, D. H., Van Bortle, K., Spacek, D. V., Hess, G. T., Saad Shamim, M., Machol, I., Love, M. I., Lieberman Aiden, E., Bassik, M. C. & Snyder, M. P. Static And Dynamic DNA Loops Form AP-1 Bound Activation Hubs During Macrophage Development. *bioRxiv* (2017). doi:10.1101/142026.

14. Nora, E. P., Goloborodko, A., Valton, A. L., Gibcus, J. H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L. A. & Bruneau, B. G. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169,** 930–944 e22. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (May 2017). doi:10.1016/j.cell.2017.05.004.

15. Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B.-M., Wingett, S. W. & Fraser, P. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biology* **16,** 175. ISSN: 1474-7596 (2015). doi:10.1186/s13059-015-0753-7.

16. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* **48,** 471–84. ISSN: 1097-4164 (Electronic) 1097-2765 (Linking) (Nov. 2012). doi:10.1016/j.molcel.2012.08.031.

17. Battulin, N., Fishman, V. S., Mazur, A. M., Pomaznoy, M., Khabarova, A. A., Afonnikov, D. A., Prokhortchouk, E. B. & Serov, O. L. Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol* **16,** 77. ISSN: 1474-760X (Electronic) 1474-7596 (Linking) (2015). doi:10.1186/s13059-015-0642-0.

18. Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., Uzawa, S., Dekker, J. & Meyer, B. J. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523,** 240–4. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (July 2015). doi:10.1038/nature14450.

19. Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H. Q., Ong, C. T., Cubenas-Potts, C., Hu, M., Lei, E. P., Bosco, G., Qin, Z. S. & Corces, V. G. Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Mol Cell* **58,** 216–31. ISSN: 1097-4164 (Electronic) 1097-2765 (Linking) (Apr. 2015). doi:10.1016/j.molcel.2015.02.023.

20. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–93. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (Oct. 2009). doi:10.1126/science.1181369.

21. Yang, T., Zhang, F., Yardimci, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F. & Li, Q. HiCRep: assessing the reproducibility of Hi-C data using a stratum- adjusted correlation coefficient. *Genome Res.* ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (Aug. 2017). doi:10.1101/gr.220640.117.

22. Yan, K. K., Gurkan Yardimci, G., Yan, C., Noble, W. S. & Gerstein, M. HiC-Spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics.* ISSN: 1367-4811 (Electronic) 1367-4803 (Linking) (Mar. 2017). doi:10.1093/bioinformatics/btx152.

23. Khan, G. N., Kim, E. J., Shin, T. S. & Lee, S. H. Heterogeneous Cell Types in Single-cell-derived Clones of MCF7 and MDA-MB-231 Cells. eng. *Anticancer Research* **37,** 2343–2354. ISSN: 1791-7530 (2017). doi:10.21873/anticanres.11572.

24. Jemaà, M., Abdallah, S., Lledo, G., Perrot, G., Lesluyes, T., Teyssier, C., Roux, P., van Dijk, J., Chibon, F., Abrieu, A. & Morin, N. Heterogeneity in sarcoma cell lines reveals enhanced motility of tetraploid versus diploid cells. eng. *Oncotarget* **8,** 16669–16689. ISSN: 1949-2553 (Mar. 2017). doi:10.18632/oncotarget.14291.

25. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. eng. *Bioinformatics (Oxford, England)* **25,** 2078–2079. ISSN: 1367-4811 (Aug. 2009). doi:10.1093/bioinformatics/btp352.

26. Sauria, M. E. G., Phillips-Cremins, J. E., Corces, V. G. & Taylor, J. HiFive: a tool suite for easy and efficient HiC and 5C data analysis. eng. *Genome Biology* **16,** 237. ISSN: 1474-760X (Oct. 2015). doi:10.1186/s13059-015-0806-y.

27. Koehler, R., Issac, H., Cloonan, N. & Grimmond, S. M. The uniqueome: a mappability resource for short-tag sequencing. eng. *Bioinformatics (Oxford, England)* **27,** 272–274. ISSN: 1367-4811 (Jan. 2011). doi:10.1093/bioinformatics/btq640.

28. Walt, S. v. d., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* **13,** 22–30. ISSN: 1521-9615 (Mar. 2011). doi:10.1109/MCSE.2011.37.

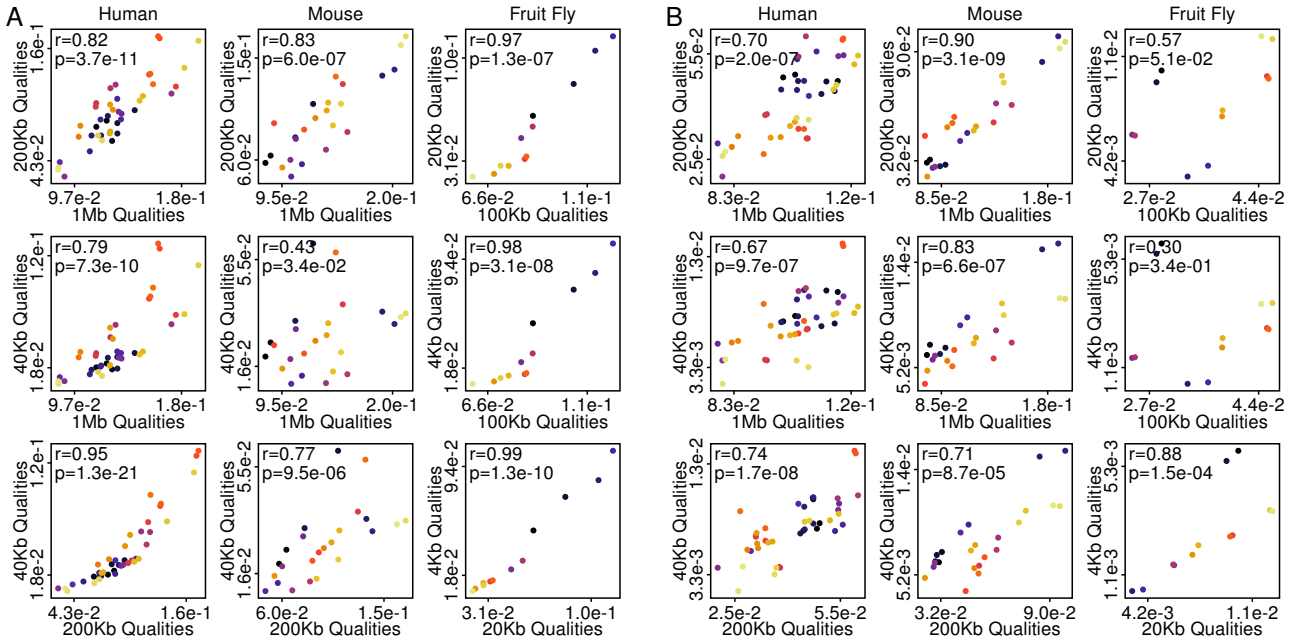

Figure S1: QuASAR quality scores show decreased consistency between larger changes in resolution. A) QuASAR quality scores from unaltered datasets are plotted between different pairs of analysis resolution. For each plot, the Spearman rank-order coefficient of correlation and associated p-value are shown. Sample color coding is consistent within a species across plots. B) QuASAR quality scores from samples down-sampled to a uniform coverage level are shown for different pairs of analysis resolution. For human and mouse data, all samples contain 10 million *cis*reads while for Drosophila samples each contain 1 million *cis*reads. For each plot, the Spearman rank-order coefficient of correlation and associated p-value are shown. Sample color coding is consistent within a species across plots.

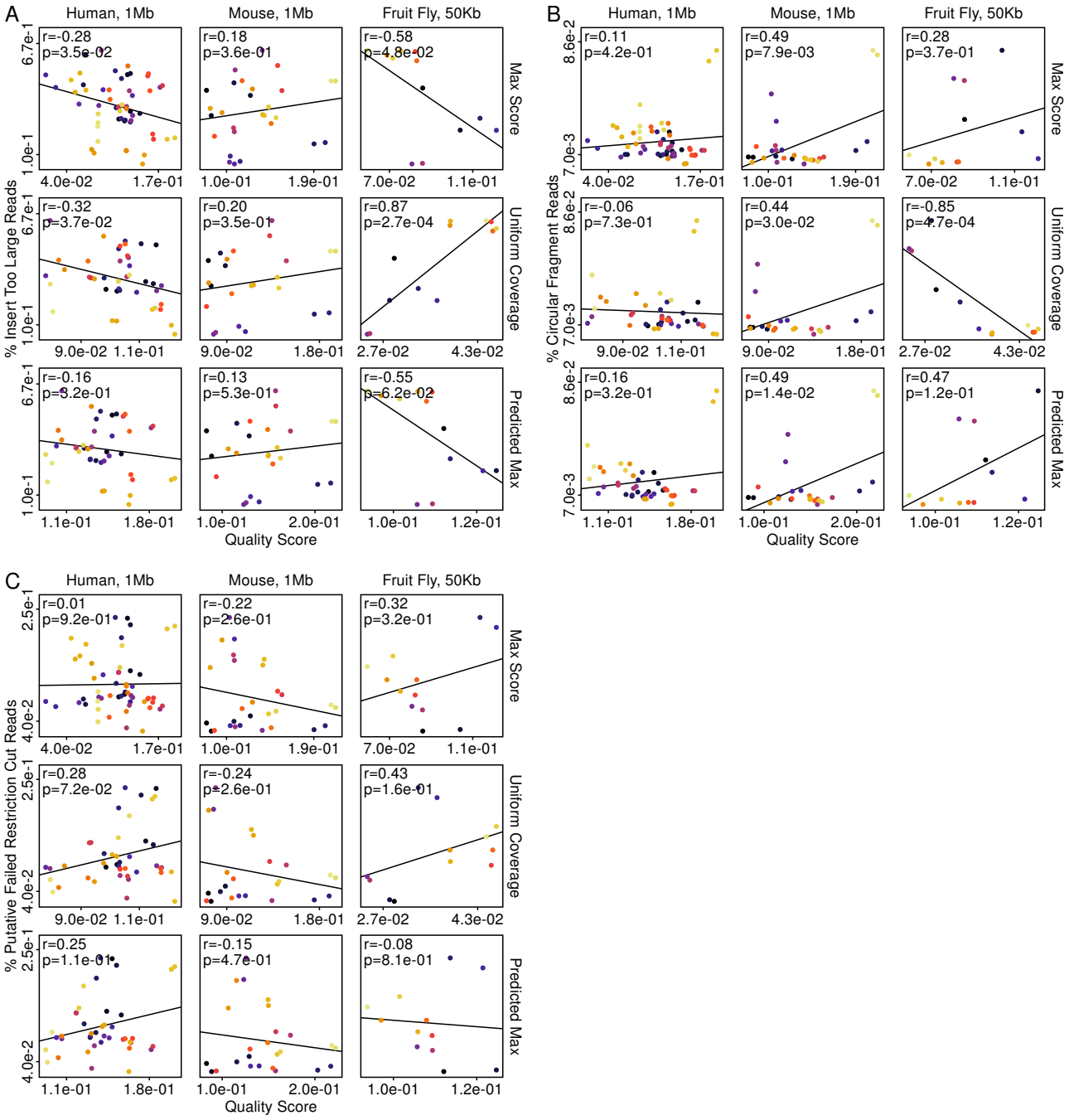

Figure S2: QuASAR quality scores show no consistent relationship to many common descriptive Hi-C statistics. QuASAR quality scores versus common Hi-C mapping statistics. For each panel, quality scores were derived from unaltered samples (top, 1 Mb resolution for human and mouse, 100 Kb resolution for Drosophila), samples down-sampled to equal coverage (middle, 10 million *cis*reads for human and mouse, 1 million *cis*reads for Drosophila), and modeled scores under infinite coverage (bottom, samples with at least 4 million *cis*reads for human and mouse or 1 million reads for Drosophila). For each plot, the Pearson correlation coefficient and associated p-value are shown. A) Quality scores are plotted versus the percentage of reads in each sample with an estimated insert size larger than 650 bp. B) Quality scores are plotted versus the percentage of reads circularized, or self-ligated restriction fragments for each sample. C) Quality scores are plotted versus the percentage of reads for each sample on adjacent restriction fragments that whose ends were in opposing orientations, allowing for the possibility of a failed restriction cut and circularization.

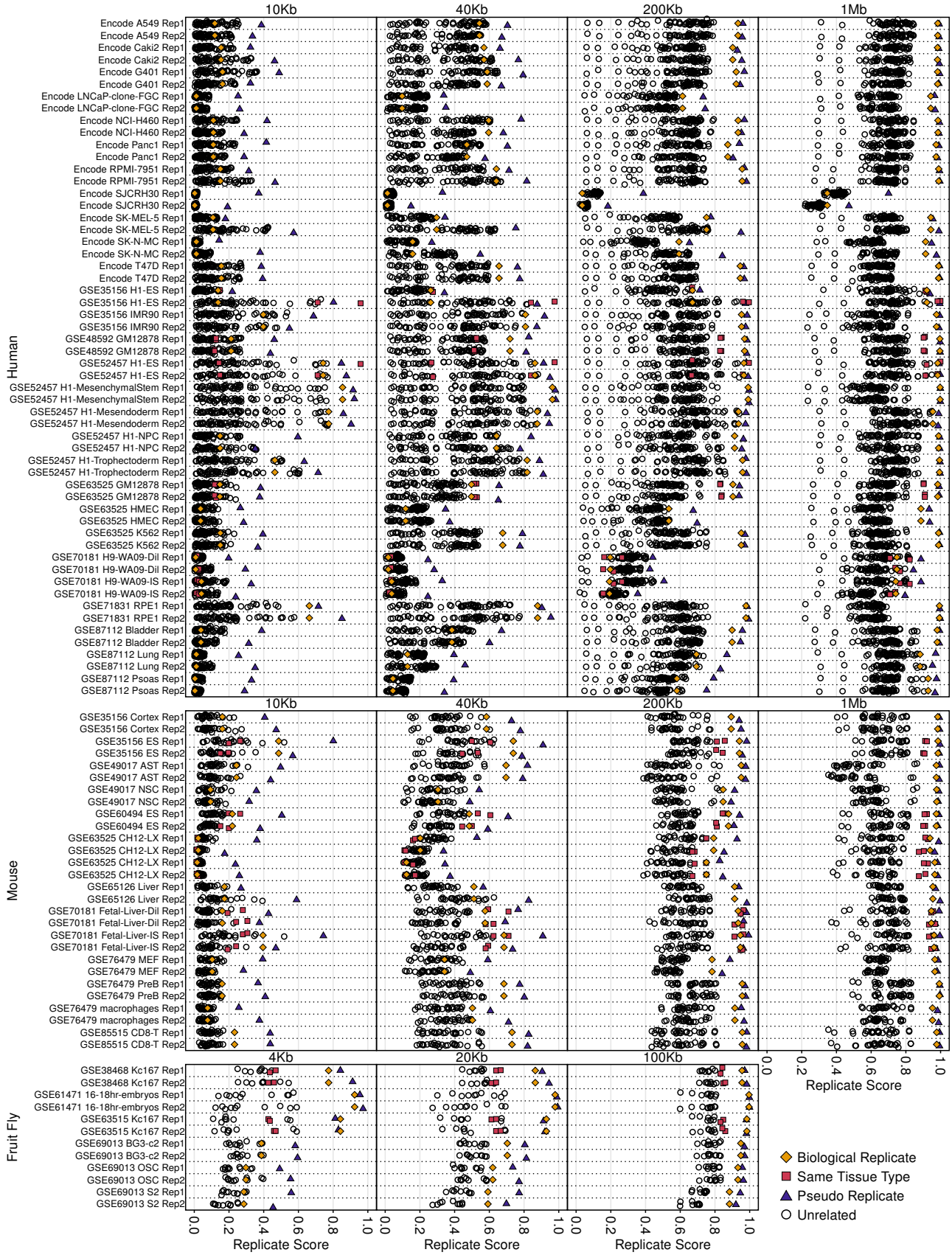
10Kb
40Kb
200Kb
1Mb
Human
Encode A549 Rep1
Encode A549 Rep2
Encode Caki2 Rep1
Encode Caki2 Rep2
Encode G401 Rep1
Encode G401 Rep2
Encode LNCaP-clone-FGC Rep1
Encode LNCaP-clone-FGC Rep2
Encode NCI-H460 Rep1
Encode NCI-H460 Rep2
Encode Panc1 Rep1
Encode Panc1 Rep2
Encode RPMI-7951 Rep1
Encode RPMI-7951 Rep2
Encode SJCRH30 Rep1
Encode SJCRH30 Rep2
Encode SK-MEL-5 Rep1
Encode SK-MEL-5 Rep2
Encode SK-N-MC Rep1
Encode SK-N-MC Rep2
Encode T47D Rep1
Encode T47D Rep2
GSE35156 H1-ES Rep1
GSE35156 H1-ES Rep2
GSE35156 IMR90 Rep1
GSE35156 IMR90 Rep2
GSE48592 GM12878 Rep1
GSE48592 GM12878 Rep2
GSE52457 H1-ES Rep1
GSE52457 H1-ES Rep2
GSE52457 H1-MesenchymalStem Rep1
GSE52457 H1-MesenchymalStem Rep2
GSE52457 H1-Mesendoderm Rep1
GSE52457 H1-Mesendoderm Rep2
GSE52457 H1-NPC Rep1
GSE52457 H1-NPC Rep2
GSE52457 H1-Trophectoderm Rep1
GSE52457 H1-Trophectoderm Rep2
GSE63525 GM12878 Rep1
GSE63525 GM12878 Rep2
GSE63525 HMEC Rep1
GSE63525 HMEC Rep2
GSE63525 K562 Rep1
GSE63525 K562 Rep2
GSE70181 H9-WA09-Dil Rep1
GSE70181 H9-WA09-Dil Rep2
GSE70181 H9-WA09-IS Rep1
GSE70181 H9-WA09-IS Rep2
GSE71831 RPE1 Rep1
GSE71831 RPE1 Rep2
GSE87112 Bladder Rep1
GSE87112 Bladder Rep2
GSE87112 Lung Rep1
GSE87112 Lung Rep2
GSE87112 Psoas Rep1
GSE87112 Psoas Rep2
Mouse
GSE35156 Cortex Rep1
GSE35156 Cortex Rep2
GSE35156 ES Rep1
GSE35156 ES Rep2
GSE49017 AST Rep1
GSE49017 AST Rep2
GSE49017 NSC Rep1
GSE49017 NSC Rep2
GSE60494 ES Rep1
GSE60494 ES Rep2
GSE63525 CH12-LX Rep1
GSE63525 CH12-LX Rep2
GSE63525 CH12-LX Rep1
GSE63525 CH12-LX Rep2
GSE65126 Liver Rep1
GSE65126 Liver Rep2
GSE70181 Fetal-Liver-Dil Rep1
GSE70181 Fetal-Liver-Dil Rep2
GSE70181 Fetal-Liver-IS Rep1
GSE70181 Fetal-Liver-IS Rep2
GSE76479 MEF Rep1
GSE76479 MEF Rep2
GSE76479 PreB Rep1
GSE76479 PreB Rep2
GSE76479 macrophages Rep1
GSE76479 macrophages Rep2
GSE85515 CD8-T Rep1
GSE85515 CD8-T Rep2
Fruit Fly
4Kb
20Kb
100Kb
GSE38468 Kc167 Rep1
GSE38468 Kc167 Rep2
GSE61471 16-18hr-embryos Rep1
GSE61471 16-18hr-embryos Rep2
GSE63515 Kc167 Rep1
GSE63515 Kc167 Rep2
GSE69013 BG3-c2 Rep1
GSE69013 BG3-c2 Rep2
GSE69013 OSC Rep1
GSE69013 OSC Rep2
GSE69013 S2 Rep1
GSE69013 S2 Rep2
0.0
0.2
0.4
0.6
0.8
1.0
Replicate Score
Biological Replicate
Same Tissue Type
Pseudo Replicate
Unrelated

Figure S3: QuASAR replicate scores across all tested resolutions. A-C) Replicate scores are denoted as points for each sample with an all to all pairwise comparison scheme within each set of species samples. Pairs include true replicates (gold diamonds), pseudo-replicates (purple triangles), same tissue of origin but non-replicates (fuchsia squares), and unrelated samples black circles). For each species, the associated resolution is displayed at the top of each block of samples.

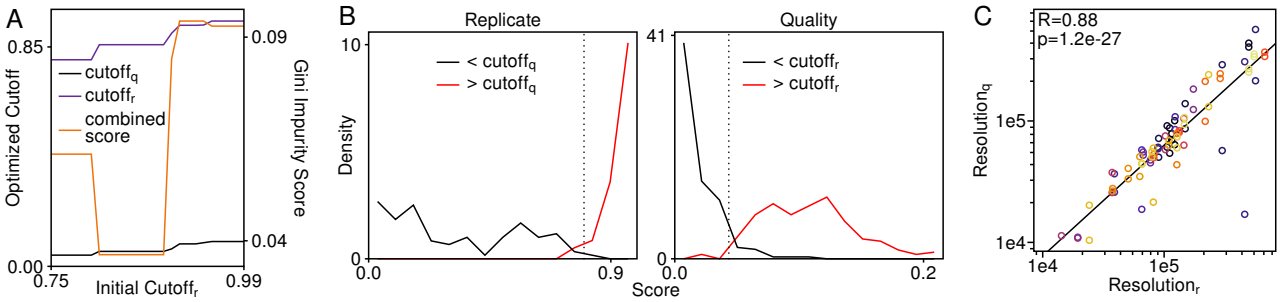

Figure S4: Determining and validating QuASAR score cutoffs for finding maximum usable resolutions. A) Optimized cutoff values and associated Gini impurity score sums as a function of starting replicate cutoff. B) Replicate (left) and quality (right) score distributions, partitioned by loose quality and replicate cutoffs, respectively. Dotted lines donate the best separation point between partitions based on Gini impurity scoring. C) Resolution limits for each sample determined by loose replicate and quality cutoffs. The Pearson R value and associated p-value are shown.

Table S1: Datasets and associated read statistics used for analysis.

Sample Name	Source	Genome	Restriction Enzyme	Total Reads	Cis Reads	Trans Reads	Insert Too Large	Circular Fragments	Failed RE Cut
GSE38468 Kc167 Rep1	GEO	dm6	HindIII	27,317,496	7,153,329	1,728,135	14,800,469	1,434,825	2,190,451
GSE38468 Kc167 Rep2	GEO	dm6	HindIII	31,866,712	12,541,844	2,432,091	10,748,540	3,536,955	2,597,335
GSE61471 16-18hr-embryos Rep1	GEO	dm6	DpnII	417,535,739	119,812,224	42,624,920	164,963,795	17,551,101	72,572,666
GSE61471 16-18hr-embryos Rep2	GEO	dm6	DpnII	426,049,573	149,950,110	53,923,579	143,370,660	8,237,509	70,557,812
GSE63515 Kc167 Rep1	GEO	dm6	DpnII	13,296,137	7,180,064	1,303,959	2,300,660	1,159,378	1,337,490
GSE63515 Kc167 Rep2	GEO	dm6	DpnII	16,574,302	8,899,786	1,701,718	2,917,703	1,415,731	1,618,281
GSE69013 BG3-c2 Rep1	GEO	dm6	HindIII	26,400,803	3,294,546	765,730	18,986,828	435,228	2,901,226
GSE69013 BG3-c2 Rep2	GEO	dm6	HindIII	23,297,036	3,493,941	841,381	15,708,067	386,559	2,853,631
GSE69013 OSC Rep1	GEO	dm6	HindIII	21,380,995	2,808,504	628,555	14,995,492	316,543	2,618,351
GSE69013 OSC Rep2	GEO	dm6	HindIII	27,825,128	3,399,903	719,024	20,097,331	445,878	3,143,223
GSE69013 S2 Rep1	GEO	dm6	HindIII	16,985,283	2,265,853	323,120	11,652,882	326,404	2,411,260
GSE69013 S2 Rep2	GEO	dm6	HindIII	15,659,680	1,612,695	247,311	11,352,367	349,835	2,088,731
Encode A549 Rep1	Encode	hg38	HindIII	72,612,071	22,858,964	22,845,515	19,816,503	531,128	6,559,691
Encode A549 Rep2	Encode	hg38	HindIII	76,607,276	20,620,077	27,333,003	21,615,741	484,748	6,553,431
Encode Caki2 Rep1	Encode	hg38	HindIII	124,611,983	15,055,104	14,092,439	63,585,924	2,844,671	29,033,309
Encode Caki2 Rep2	Encode	hg38	HindIII	93,196,968	23,626,839	26,169,276	29,991,725	878,961	12,529,949
Encode G401 Rep1	Encode	hg38	HindIII	84,180,566	35,343,065	11,022,653	26,205,757	875,727	10,732,964
Encode G401 Rep2	Encode	hg38	HindIII	113,149,805	20,646,983	7,428,005	58,384,950	1,728,126	24,961,144
Encode LNCaP-clone-FGC Rep1	Encode	hg38	HindIII	98,746,284	7,283,328	23,201,271	60,492,800	699,496	7,068,557
Encode LNCaP-clone-FGC Rep2	Encode	hg38	HindIII	77,023,278	7,728,768	27,195,548	35,269,686	551,162	6,277,727
Encode NCI-H460 Rep1	Encode	hg38	HindIII	92,808,397	28,822,618	28,659,574	25,700,402	847,299	8,778,177
Encode NCI-H460 Rep2	Encode	hg38	HindIII	79,834,246	18,929,104	24,283,233	27,177,532	592,543	8,851,522
Encode Panc1 Rep1	Encode	hg38	HindIII	120,072,122	18,626,287	16,865,350	58,867,110	2,119,754	23,592,859
Encode Panc1 Rep2	Encode	hg38	HindIII	99,293,901	11,489,787	9,614,681	52,631,353	2,282,382	23,275,062
Encode RPMI-7951 Rep1	Encode	hg38	HindIII	86,921,033	21,604,431	32,397,270	24,748,314	610,146	7,560,556
Encode RPMI-7951 Rep2	Encode	hg38	HindIII	115,106,366	34,148,233	40,746,072	30,732,770	689,245	8,789,560
Encode SJCRH30 Rep1	Encode	hg38	HindIII	5,935,796	1,193,885	1,635,778	2,638,747	67,606	399,746
Encode SJCRH30 Rep2	Encode	hg38	HindIII	4,563,098	566,739	1,246,769	2,321,581	76,756	351,201
Encode SK-MEL-5 Rep1	Encode	hg38	HindIII	22,536,343	6,479,488	3,202,548	9,753,484	240,286	2,860,413
Encode SK-MEL-5 Rep2	Encode	hg38	HindIII	87,123,588	28,037,119	14,125,496	35,118,308	685,888	9,156,287
Encode SK-N-MC Rep1	Encode	hg38	HindIII	81,487,185	7,164,206	38,532,406	29,268,260	430,500	6,091,634
Encode SK-N-MC Rep2	Encode	hg38	HindIII	107,114,521	15,152,249	44,287,180	37,685,017	870,651	9,119,138
Encode T47D Rep1	Encode	hg38	HindIII	71,523,888	23,196,887	20,325,331	21,523,807	431,158	6,046,328

Table S1: continued on following page

Table S1: continued from previous page

Sample Name	Source	Genome	Restriction Enzyme	Total Reads	Cis Reads	Trans Reads	Insert Too Large	Circular Fragments	Failed RE Cut
Encode T47D Rep2	Encode	hg38	HindIII	79,824,528	24,965,075	25,287,422	21,929,437	609,170	7,032,989
GSE35156 H1-ES Rep1	GEO	hg38	HindIII	99,225,927	15,728,019	11,013,655	62,995,123	1,262,238	8,226,237
GSE35156 H1-ES Rep2	GEO	hg38	HindIII	253,250,501	95,596,874	26,508,473	112,747,912	2,544,935	15,851,285
GSE35156 IMR90 Rep1	GEO	hg38	HindIII	197,933,659	61,886,805	35,023,655	84,610,225	2,589,483	13,822,591
GSE35156 IMR90 Rep2	GEO	hg38	HindIII	128,792,931	40,120,871	38,964,530	44,662,066	1,512,986	3,531,782
GSE48592 GM12878 Rep1	GEO	hg38	HindIII	328,028,071	57,950,306	110,045,885	111,768,542	5,076,394	43,184,730
GSE48592 GM12878 Rep2	GEO	hg38	HindIII	304,035,507	53,229,613	101,861,532	104,811,348	4,253,306	39,877,692
GSE52457 H1-ES Rep1	GEO	hg38	HindIII	351,157,437	128,489,104	35,665,087	159,410,008	3,503,431	24,088,362
GSE52457 H1-ES Rep2	GEO	hg38	HindIII	777,847,319	178,802,683	134,554,773	391,997,800	8,936,942	63,550,893
GSE52457 H1-MesenchymalStem Rep1	GEO	hg38	HindIII	277,934,243	171,430,338	31,690,827	49,700,845	1,890,971	23,220,087
GSE52457 H1-MesenchymalStem Rep2	GEO	hg38	HindIII	311,178,604	185,105,943	36,792,481	63,818,146	1,989,774	23,470,939
GSE52457 H1-Mesendoderm Rep1	GEO	hg38	HindIII	498,214,164	135,013,524	60,198,636	256,421,585	5,217,179	41,361,086
GSE52457 H1-Mesendoderm Rep2	GEO	hg38	HindIII	372,886,094	160,461,301	26,272,049	159,350,579	2,467,841	24,332,902
GSE52457 H1-NPC Rep1	GEO	hg38	HindIII	465,940,053	64,003,384	166,245,873	201,264,454	12,321,072	22,102,878
GSE52457 H1-NPC Rep2	GEO	hg38	HindIII	335,841,327	46,344,708	123,413,974	128,678,675	7,862,163	29,540,171
GSE52457 H1-Trophectoderm Rep1	GEO	hg38	HindIII	414,429,662	73,609,663	63,851,283	229,901,258	8,549,479	38,515,926
GSE52457 H1-Trophectoderm Rep2	GEO	hg38	HindIII	296,493,743	86,565,334	84,448,280	99,629,689	8,558,430	17,290,771
GSE63525 GM12878 Rep1	GEO	hg38	MboI	43,775,044	25,501,372	8,216,991	5,233,259	191,622	4,629,274
GSE63525 GM12878 Rep2	GEO	hg38	MboI	43,450,927	25,334,027	7,933,669	5,268,787	182,547	4,729,473
GSE63525 HMEC Rep1	GEO	hg38	MboI	14,777,441	7,695,744	2,674,969	1,868,373	327,462	2,191,793
GSE63525 HMEC Rep2	GEO	hg38	MboI	13,053,060	7,711,262	2,160,367	1,410,567	189,872	1,572,903
GSE63525 K562 Rep1	GEO	hg38	MboI	60,146,326	32,470,863	17,107,447	5,896,409	45,008	4,607,175
GSE63525 K562 Rep2	GEO	hg38	MboI	47,824,657	28,329,651	15,953,626	2,477,543	22,090	1,024,336
GSE70181 H9-WA09-Dil Rep1	GEO	hg38	HindIII	20,126,169	3,436,591	4,574,042	8,447,808	519,041	3,148,600
GSE70181 H9-WA09-Dil Rep2	GEO	hg38	HindIII	17,739,992	4,008,970	1,311,554	9,195,616	344,916	2,878,826
GSE70181 H9-WA09-IS Rep1	GEO	hg38	HindIII	23,048,639	7,441,353	1,142,568	9,931,964	287,044	4,245,623
GSE70181 H9-WA09-IS Rep2	GEO	hg38	HindIII	21,080,066	5,625,889	578,942	10,177,517	576,933	4,120,676
GSE71831 RPE1 Rep1	GEO	hg38	MboI	145,305,043	66,291,497	11,590,065	25,922,123	10,534,199	30,960,810
GSE71831 RPE1 Rep2	GEO	hg38	MboI	302,117,374	132,699,637	23,177,903	56,145,371	24,179,203	65,904,467
GSE87112 Bladder Rep1	GEO	hg38	HindIII	119,226,332	31,474,525	28,941,017	38,564,740	3,472,838	16,772,414
GSE87112 Bladder Rep2	GEO	hg38	HindIII	118,535,904	27,454,462	24,512,068	42,143,974	2,886,078	21,538,520
GSE87112 Lung Rep1	GEO	hg38	HindIII	73,446,757	15,602,202	27,538,752	22,661,780	3,078,895	4,564,529
GSE87112 Lung Rep2	GEO	hg38	HindIII	58,327,055	13,779,993	31,450,140	9,769,326	1,065,990	2,261,261
GSE87112 Psoas Rep1	GEO	hg38	HindIII	37,635,363	9,889,874	13,236,010	9,908,577	1,012,717	3,587,964

Table S1: continued on following page

Table S1: continued from previous page

Sample Name	Source	Genome	Restriction Enzyme	Total Reads	Cis Reads	Trans Reads	Insert Too Large	Circular Fragments	Failed RE Cut
GSE87112 Psoas Rep2	GEO	hg38	HindIII	26,748,941	7,527,523	9,975,163	6,554,948	615,005	2,076,130
GSE34587 Cortex Rep1	GEO	mm10	HindIII	212,614,457	27,556,396	58,771,240	105,882,359	2,900,093	14,147,018
GSE34587 Cortex Rep2	GEO	mm10	HindIII	163,357,944	31,257,972	57,065,969	63,079,946	2,165,699	8,580,341
GSE35156 Cortex Rep1	GEO	mm10	HindIII	212,614,457	27,556,396	58,771,240	105,882,359	2,900,093	14,147,018
GSE35156 Cortex Rep2	GEO	mm10	HindIII	163,357,944	31,257,972	57,065,969	63,079,946	2,165,699	8,580,341
GSE35156 ES Rep1	GEO	mm10	HindIII	265,710,217	94,313,379	21,047,025	126,339,528	2,798,212	20,927,249
GSE35156 ES Rep2	GEO	mm10	HindIII	164,478,412	42,231,845	20,764,228	87,368,411	2,439,453	11,402,036
GSE49017 AST Rep1	GEO	mm10	HindIII	38,561,251	21,058,621	4,391,813	10,063,301	924,013	2,121,440
GSE49017 AST Rep2	GEO	mm10	HindIII	34,234,548	17,996,920	3,677,052	9,132,562	1,311,286	2,114,621
GSE49017 NSC Rep1	GEO	mm10	HindIII	20,257,064	12,490,185	2,284,336	3,719,539	501,700	1,260,830
GSE49017 NSC Rep2	GEO	mm10	HindIII	17,756,656	10,763,970	1,802,998	3,674,406	415,532	1,099,247
GSE60494 ES Rep1	GEO	mm10	MboI	101,850,576	37,397,967	18,128,121	17,531,817	7,217,983	21,565,649
GSE60494 ES Rep2	GEO	mm10	MboI	73,926,376	23,705,623	10,008,600	13,313,058	8,486,354	18,406,698
GSE63525 CH12-LX Rep1	GEO	mm10	HindIII	71,288,879	11,370,618	7,772,879	48,116,005	169,518	3,839,514
GSE63525 CH12-LX Rep1	GEO	mm10	MboI	21,214,642	6,809,170	3,496,650	6,682,211	514,707	3,705,464
GSE63525 CH12-LX Rep2	GEO	mm10	HindIII	36,593,801	5,810,424	6,863,380	21,517,260	186,512	2,195,847
GSE63525 CH12-LX Rep2	GEO	mm10	MboI	27,609,391	7,938,405	4,545,941	9,157,081	879,386	5,082,898
GSE65126 Liver Rep1	GEO	mm10	HindIII	53,135,939	15,514,435	4,367,439	26,404,465	716,353	6,130,595
GSE65126 Liver Rep2	GEO	mm10	HindIII	254,523,004	48,693,691	17,370,076	154,709,653	2,826,790	30,909,152
GSE70181 Fetal-Liver-Dil Rep1	GEO	mm10	HindIII	196,672,664	42,659,146	79,505,086	57,984,060	5,977,620	10,532,404
GSE70181 Fetal-Liver-Dil Rep2	GEO	mm10	HindIII	99,608,593	25,491,323	9,986,406	53,839,178	1,167,911	9,111,579
GSE70181 Fetal-Liver-IS Rep1	GEO	mm10	HindIII	190,841,201	95,168,973	15,858,814	67,031,529	1,637,766	11,136,586
GSE70181 Fetal-Liver-IS Rep2	GEO	mm10	HindIII	103,085,039	34,296,309	4,580,298	51,951,582	1,683,682	10,565,827
GSE76479 MEF Rep1	GEO	mm10	DpnII	70,549,803	22,951,554	3,746,804	28,181,934	835,505	14,816,840
GSE76479 MEF Rep2	GEO	mm10	DpnII	30,607,750	9,808,604	2,872,556	12,799,295	148,793	4,971,753
GSE76479 PreB Rep1	GEO	mm10	DpnII	76,791,457	27,466,613	5,616,514	29,981,046	925,960	12,798,430
GSE76479 PreB Rep2	GEO	mm10	DpnII	92,594,100	32,866,073	6,587,365	35,823,863	950,826	16,362,255
GSE76479 macrophages Rep1	GEO	mm10	DpnII	38,112,969	12,393,933	6,068,926	16,167,954	341,586	3,135,971
GSE76479 macrophages Rep2	GEO	mm10	DpnII	52,530,406	18,282,498	9,537,324	19,793,224	347,049	4,562,078
GSE85515 CD8-T Rep1	GEO	mm10	MboI	176,761,858	25,970,292	7,609,752	95,364,492	31,336,926	16,478,238
GSE85515 CD8-T Rep2	GEO	mm10	MboI	179,376,530	25,162,394	6,687,851	96,823,586	33,131,244	17,569,215