HiFive: A normalization approach for higher-resolution HiC and 5C chromosome conformation data analysis

Michael EG Sauria[1], Jennifer E Phillips-Cremins[1,3], Victor G Corces[1], James Taylor[1,2§]

[1]Department of Biology, Emory University, Atlanta, GA 30322, USA

[2]Departments of Biology and Computer Science, Johns Hopkins University, Baltimore, MD 21211, USA

[3]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19103, USA

[§]Corresponding author

Email addresses:

MEGS: mgehrin@emory.edu

JEPC: jcremins@seas.upenn.edu

VGC: vcorces@emory.edu

JT: james@taylorlab.org

Abstract

While sequencing-based proximity assays have established many megabase-scale features of chromosome organization, experimental noise and biases have left finer-scale organization poorly understood. We developed HiFive, an empirically-driven probabilistic modeling approach for normalizing and analyzing HiC and 5C data. HiFive allows reconstruction of chromatin structural features at higher resolution than previously described and meets or outperforms other approaches. With novel feature detection, denoising, and efficient algorithms, HiFive enables utilization of these data with a fraction of the computational power and time without sacrificing resolution. Using HiFive, we demonstrate transcriptionally-associated differential spatial organization of the mouse genome.

Keywords

HiC, 5C, chromatin conformation, normalization, software, transcriptional activity, spatial organization

Rationale

Although the vast majority of the human genome was sequenced more than a decade ago, it is clear that sequence alone is insufficient to explain the complex gene and RNA regulatory patterns seen across time and cell type in eukaryotes. The context surrounding sequences—whether from combinations of DNA binding transcription factors (TFs) [1-3], methylation of the DNA itself [4, 5], or local histone modifications [5, 6]—is integral to how the cell utilizes each sequence element. Although we have known about the potential roles that sequentially distant but spatially proximal sequences and their binding and epigenetic contexts play in regulating expression and function, it has only been over the past decade that new sequencing-based techniques have enabled high-throughput analysis of higher-order structures of chromatin and investigation into how these structures interact amongst themselves and with other genomic elements, to influence cellular function.

Several different methods for assessing chromatin interactions have been devised, all based on the sequencing of hybrid fragments of DNA created preferentially between spatially close sequences. These approaches include ChIA-Pet [7], tethered chromosome capture [8], and the chromatin conformation capture technologies of 3C, 4C, 5C, and HiC [9-12] (Supplemental Figure 1; see Supplemental Materials: Proximity-mediated ligation assays). While these assays have allowed a rapid expansion of our understanding of the nature of genome structure, they also have presented some formidable challenges.

In both HiC and 5C, systematic biases resulting from the nature of the assays have been observed [13, 14], resulting in differential representation of sequences in the resulting datasets. While analyses at a larger scale are not dramatically affected by these biases due to the large number of data points being averaged over, higher-resolution approaches must first address these challenges. Several analysis methods have been described in the literature and applied to correcting biases in HiC [13, 15-18] and 5C data [19-23]. There is still, however, room for improving our ability to remove this systematic noise from the data and resolve finer-scale features.

A second challenge posed by data from these types of assays is one of resources. Unlike other next-generation sequencing assays where even single-base resolution is limited to a few billion data points, these assays assess pairwise combinations, potentially increasing the size of the dataset by several orders of magnitude. For a three billion base-pair genome cut with a six-base restriction enzyme (RE), the number of potential interaction pairs is more than half a trillion. Even allowing that the vast majority of those interactions will be absent from the sequencing data, the amount of information that needs to be handled and the complexity of normalizing these data still pose a major computational hurdle, especially for investigators without access to substantial computational resources.

Here we present HiFive, an analysis method developed for handling both HiC and 5C data using a combination of empirically determined and probabilistic signal modeling. We demonstrate that HiFive allows memory- and computationally-efficient HiC and 5C data handling and normalization while retaining high-resolution data for downstream analyses of interaction signals, making fine-scale chromatin structural analysis accessible to a wider range of investigators.

Implementation

Organization of HiFive

HiFive is based on a hierarchy of data modules stored in HDF5-formatted files (a management structure for handling complex and large sets of data) using the package h5py for easy, compact, and fast access that may be shared across experiments and analyses as shown in Figure 1. All aspects of the software are written in Python2 and make use of the Cython and NumPy packages to accelerate computationally intensive operations. This gives HiFive speed similar to C-based code with the human readability and ease of use of Python. In addition, certain scripts and functions support parallelization using the package mpi4py to utilize the Message Passing Interface (MPI), greatly increasing the scalability of analysis. This allows computationally intensive processes to be automatically split across multiple processors on a cluster or multicore computer. The overall goal of this implementation is to reduce storage, memory, and processing power requirements without sacrificing analytical power.

For additional details of HiFive's organization, see *Supplemental Materials: HiFive's organization*.

Data filtering

In both experimental approaches, data are filtered in a two-stage process. After mapping is completed and read pairs have been assigned to fragments or fends for 5C and HiC respectively, a set of filters is applied to the data when creating the final dataset. Once a HiC or 5C object is created and linked to a data object, a second filter is applied (Supplemental Figures 2 and 3).

The first round of filtering is designed to remove non-valid pairings of reads ends. In 5C, this is limited to read ends that originated from primers with the same orientation. HiC is more complicated due to the random shearing and non-sequence-specific assay approach. There are two primary requirements to be considered a value HiC read. First, the distance between the ends of the sequenced fragment cannot be too long, as inferred by the sum of distances from mapping coordinates to the next downstream restriction site. Second, the pair of read ends cannot have originated as the result of a failed restriction cut, circularization of a single fragment, or sequencing of a PCR duplicate (Supplemental Figure 4). A more detailed discussion about read filtering can be found in the *Supplemental Material: Read filtering*.

Once data objects are created and linked to a HiC or 5C object, a filter based on read coverage can be applied to either. Removal of fragments is accomplished using an iterative filtering process. For each fragment or fend, the number of non-filtered interacting sequences for which the pair had a non-zero read count is calculated within a specified maximum interaction distance range. Sequences with fewer than the specified minimum-count of valid interaction pairs are discarded and the process is repeated until all remaining sequences have at least the minimum number of valid interaction pairs.

Estimation of distance-dependent signal

In both HiC and 5C, the largest influence driving interaction signal intensity is the genomic distance between interaction sequences. Given an experiment with dense-enough coverage at shorter ranges of inter-fragment distances, we observe a roughly power-law function such that the log of the interaction counts varies as a linear function of the log of the inter-fragment distance. We find that this relationship holds best when the distance is measured from midpoint to midpoint of the fragment pair and when considering interactions covering ranges from 1 Kb to 1Mb. Because of the uneven distribution of RE sites throughout the genome, we believe it is imperative that any normalization scheme must take fragment spacing into account in order to generate unbiased estimates of relative interaction frequency. To this end, HiFive finds a distance-dependent signal approximation function prior to normalization and incorporates this function into the calculation of expected counts.

In order to approximate the nearly linear relationship between the log of non-zero counts and the log of inter-fragment distances seen in 5C data, HiFive uses a simple regression approach (Supplemental Figure 5a). Because the log of the counts is used, zeros are necessarily excluded from the calculations. We do not, however, feel that this has any negative impact on parameter estimation. To the contrary, we see increased variation at lower read counts suggesting that unobserved fragment pairs are the least accurate data in the experiment. For valid non-zero interaction counts *s* and distances *d* between fragments *i* and *j*, the distance function *D* is estimated with slope *b* and intercept *a as*:

$$D(d_{i,j}) = e^{\log(d_{i,j})b+a}$$

Counts, *s*, are valid if not filtered in the creation of the data object and originate from non-filtered fragments. These counts are denoted by subset *A*. The slope is defined as:

$$b = \frac{\sum_{i,j}^{A}\left[\log(d_{i,j}) - \overline{\log\,(d_A)}\right]\left[\log(\mathrm{s}_{i,j}) - \overline{\log(s_A)}\right]}{\sum_{i,j}^{A}\left[\log(d_{i,j}) - \overline{\log(d_A)}\right]^2}$$

and the intercept is defined as:

$$a = s_A - log(d_A)b$$

Once calculated, the intercept is held constant. The slope may be updated throughout the normalization procedure.

HiC data have a similar general relationship between signal and inter-fend distance, though the relationship is not linear over the larger range of distances included. In addition, the sparseness of non-zero interaction counts makes direct assessment of the distance-dependent relationship difficult. HiFive overcomes these challenges by using a piecewise approximation approach (Supplemental Figure 5b). Although other non-parametric options are available that may yield more precise estimates, the size of datasets involved in HiC analysis make them impractical from a computational and time standpoint. To find the piecewise approximation, HiFive begins by splitting interactions into *N* bins of equal log-distance sizes covering the complete range of distances for intra-chromosomal interactions. The mean of each bin is calculated from all valid counts, *s*, spanning the bin's range, where valid interactions exclude those filtered out when creating the HiC data object and those originating from removed fends and are denoted as subset *A*. Thus, for the bin *n* with an upper bound of *c*, the mean *μ* is calculated as follows:

$$\mu_n = \frac{\sum_{i,j}^{c_{n-1}<d_{i,j}\leq c_n} s_{i,j}}{\left|\{s_{i,j}: c_{n-1} < d_{i,j} \leq c_n\}\right|}$$

To account for noise in the data, HiFive can apply a triangular smoothing function to the bin means using smoothing parameter *k*, giving the smoothed mean value *μ*':

$$\mu_n' = e^{\frac{\log(\mu_n)k+\sum_{i=1}^{k-1}\log\,(\mu_{n-1}\mu_{n+1})(k-i)}{k^2}}$$

The HiC distance function is defined as a piecewise linear approximation of these smoothed means, where a bin's mean corresponds to a specific distance *g* falling within its upper and lower boundaries denoted as c_u and c_l, respectively:

$$g = c_u e^{(c_l - c_u)(1 - 0.5^{0.5})}$$

For an interaction between fends *i* and *j* with a distance falling between two defined points, such as between the distances denoted by *g* associated with bins *n-1* and *n*, the corresponding distance-mean is estimated:

$$D(d_{i,j}) = \frac{(g_n - d_{i,j})\mu'_{n-1} + (d_{i,j} - g_{n-1})\mu'_n}{g_n - g_{n-1}}$$

Data normalization

HiFive's normalization approach uses a combination of empirical distance-dependence estimation and probabilistic modeling to estimate expected signal and enrichment. HiFive works on two key assumptions. First, the majority of interaction reads are derived from a combination of signal that is dependent on inter-fragment distance and fragment-specific bias. Second, the effects of the fragment biases can be described as the product of the individual biases associated with the interaction fragment pair. Thus the expected signal *E* for the interaction between fragments or fends *i* and *j* is the product of the bias correction *f* for each end of the interaction and the distance function *D* as follows:

$$E_{i,j} = D(d_{i,j}) f_i f_j$$

This approach allows adjustment for factors known to contribute to differences in fragment observation rates, such as guanine and cytosine (GC) content, fragment length, and mappability [13, 15] without explicitly limiting the correction to a specific relationship of factors. Although the general framework is the same for both HiC and 5C, differences in the experimental procedures necessitate technique-specific variants of the distance-dependence function and underlying probability distribution.

The 5C observed interactions are modeled with a log-normal distribution around the predicted values with standard deviation σ such that:

$$s_A \sim LogN(E_A, \sigma^2)$$

The standard deviation is estimated at the same time as the distance-dependence parameters prior to learning fragment corrections such that:

$$\sigma = \sqrt{\frac{\sum\left[\log(s_{i,j}) - \log(E_{i,j}) - \overline{[\log(s_A) - \log(E_A)]}\right]^2}{|A| - 1}}$$

Because estimated counts change as the corrections are learned, HiFive includes the ability to periodically update the parameters of the distance-dependence function. If specified by the user, the slope *b* in the distance function and σ are updated, although the term $s_{i,j}$ is replaced with the bias-corrected count $s'_{i,j}$ for calculating these parameters such that:

$$s'_{i,j} = \frac{s_{i,j}}{f_i f_j}$$

HiC interaction counts are modeled as a series of Poisson processes with λ being defined as the predicted value for a given interaction such that:

$$s_A \sim Pois(E_A)$$

Unlike 5C, only a fraction of interactions are used to learn fend bias corrections. Specifically, counts (including zeros) are included in the model if they belong to the set of valid interactions, they are intra-chromosomal, and the distance between their fends is less than or equal to a user-specified maximum interaction distance range. This is done for two reasons; 1) the vast majority of observed interactions occur over short interaction distances and 2) including all possible interactions or simply all possible *cis* interactions is computationally unfeasible for this kind of model.

Like with 5C data, the HiC distance function can be updated during the learning of fend correction values. This is done by substituting the raw count term $s_{i,j}$ with the corrected interaction count term $s'_{i,j}$ as described above for finding the distance bin means.

HiFive learns bias parameters for both HiC and 5C data using a two-stage gradient descent approach that maximizes the likelihood of the observed data under the probability distributions described above. In the burn-in phase, fend or fragment bias parameters are updated using a constant learning rate. This is followed by an annealing phase in which the learning rate decreases in a linear fashion to zero.

HiFive-Express, an iterative approximation method for bias correction

Due to the memory and computational requirements associated with rigorous HiC and 5C normalization, HiFive also includes a fast and computationally inexpensive iterative approximation normalization alternative for both data types referred to as HiFive-Express. HiFive-Express makes use of the same framework and underlying predicted value scheme as HiFive, changing only the approach to bias correction value calculations described above. While the results are not as robust as the more computationally expensive HiFive normalization, they are still sufficient for many applications, allowing resolution down to the individual fragment or fend level depending on assay type. In addition, HiFive-Express can be performed on a single processor in minutes with a comparatively small memory footprint and can easily make use of all data, up to and including trans interactions for HiC data.

Unlike HiFive's probability-based maximization, HiFive-Express attempts to minimize the distance between one and the fraction defined by interaction counts over predicted values. Predicted counts can either be found using bias values alone or with bias and distance function values. The advantage of this approach is two-fold: 1) the calculations are simpler (and therefore faster) than calculating the gradients and 2) because the cost is calculated in terms of a fraction with observed counts as the numerator, zero counts always have a fractional value of zero, reducing the needed calculations down to only observed interactions (a small fraction of the total possible interactions) and a single sum of the number of unobserved interactions, although this second point only applies to HiC data. In addition, because HiFive-Express uses an iterative update, no learning rate parameter is necessary.

In 5C data, the normalization using HiFive-Express is still limited to non-zero interactions as the correction approximation is performed on the log of the observed counts. For each round, the fragment correction value f_i is updated as follows from the non-zero subset of observed interactions involving fragment i, A_i:

$$f_i' = f_i + \frac{\sum_j^{A_i}\left[\log\left(s_{i,j}\right) - D\left(d_{i,j}\right) - f_i - f_j\right]}{2|A_i|}$$

If distance-dependence is not taken into account, $D(d_{i,j})$ is set to zero.

HiFive-Express uses a similar approach for HiC data although because counts rather than log counts are used, all valid possible interactions are considered including unobserved interactions. For each round of HiC correction approximation, the fend correction factor f_i is updated as follows from the valid set of interactions involving fend i, A_i:

$$f_i' = f_i\sqrt{\frac{1}{|A_i|}\sum_j^{A_i}\frac{s_{i,j}}{D(d_{i,j})f_i f_j}}$$

If distance-dependent signal is ignored, $D(d_{i,j})$ is set to one.

Dynamic binning

To avoid the trade-off of resolution versus coverage, HiFive employs a dynamic binning approach that takes an initial partitioning of the data space and adjusts each bin size to meet a user-defined minimum number of interactions (Figure 2a-c). The search space can be limited, possibly resulting in invalid bins for data-sparse regions, or allowed to run until the criterion has been met for each bin. Initial partitioning can be based on fend or fragment boundaries, uniform-sized bins, or arbitrary user-defined bins. Data used when expanding bins can also be at fend- or fragment-level resolution, uniformly sized, or user-defined. In the case of fend- and fragment-resolution data arrays, bins would expand by a single interaction at a time given the non-uniform spacing, whereas a uniformly spaced data array would expand bins in all directions for a bin-size increase of (n + 1) * 4 per round, where n is the current number of steps the new boundary is from the bin. This process results in a more informative overview of the data, both visually and with respect to removing the effects of stochasticity from data-sparse regions.

See Supplemental Materials: Need for dynamic binning for further discussion.

Boundary index

One approach that has been useful in marking structurally significant features in chromatin conformation data has been identifying shift-points where interactions move from one set of high-interacting partners to another. Dixon, Selvaraj [24] described a statistic called the directionality index (DI) that measured the difference in overall interaction strength for upstream versus downstream interactions with a set of fragments. This yields a positive or negative value, depending on the bias. Boundaries are then called using an HMM to determine transition points from upstream to down stream bias in the data. While this has proved useful for identifying what they labeled as topological domains, we find that this approach has limitations in identifying smaller structures and boundary features nested within larger ones. In order to investigate these features, we have devised a variant of this statistic called the boundary index (BI), a statistic that captures shifts in interaction partner preference (Figure 2d and e). Because the BI detects interaction preference changes rather than overall bias, it is able to identify small domains and subdomain structures that are missed by the DI described by Dixon, Selvaraj [24].

The boundary index is found at each RE site by taking fends up and downstream of the site falling in equal sized intervals (“width”), calculating the log enrichment of interactions between fends in these intervals with groups of fends up and downstream of the site grouped into a specified size intervals (“height”), up to some maximum distance from the site (“window”). The BI for that site is the mean absolute difference between enrichments for interactions with fends in the upstream width versus the interactions with fends in the downstream width. Thus for boundary point P and width W, we define fends in sets I and J:

$$I = \{i : P - W \leq i < P\}$$

$$J = \{j : P \leq j < P + W\}$$

These fends interact with fends within a distance defined by the window upstream and downstream of P and divided into equal-sized intervals defined by the height, H, such that they make up N sets (the number of height-sized bins on one side of P) denoted by K and M for upstream and downstream sets, respectively:

$$K_n = \{k : P - Hn \leq k < P - H(n-1)\}$$

$$M_n = \{m : P + H(n-1) \leq m < P + Hn\}$$

Thus, the BI for point P is:

$$BI_p = \frac{1}{2|N|}\left[\left|log\left(\frac{\sum_k^{K_n}\sum_i^{I|i>k} s_{k,i}}{\sum_k^{K_n}\sum_i^{I|i>k} E_{k,i}}\right) - log\left(\frac{\sum_k^{K_n}\sum_j^{J} s_{k,j}}{\sum_k^{K_n}\sum_j^{J} E_{k,j}}\right)\right| \right.$$

$$\left. + \left|log\left(\frac{\sum_m^{M_n}\sum_i^{I} s_{i,m}}{\sum_m^{M_n}\sum_i^{I} E_{i,m}}\right) - \log\,(\left(\frac{\sum_m^{M_n}\sum_j^{J|j<m} s_{j,m}}{\sum_m^{M_n}\sum_j^{J|j<m} E_{j,m}}\right)\right|\right]$$

A user-defined minimum number of bins must be included in the calculation or *BI* is not found for point *P*. A set of BI values can then be smoothed to reduce noise and enable better peak calling using a Gaussian smoother such that for all BI positions within a distance of 2.5 R of *P*, defined as set *T*, the smoothed value BI'_p is defined as:

$$BI'_p = \frac{\sum_t^T BI_t e^{\frac{-(p-t)^2}{2R^2}}}{\sum_t^T e^{\frac{-(p-t)^2}{2R^2}}}$$

Three dimensional modeling

By using one of HiFive's various binning options and filling in missing values (and increasing the reliability of non-missing values) using the dynamic binning approach, HiFive provides an easy, fast, and intuitive way to approximate the consensus 3-dimensional conformation of chromatin from the two-dimensional array of interaction data generated using either the HiC or 5C assay. By disabling search-space limitations, HiFive guarantees the production of a completely filled matrix of values. By taking these values as *n* examples with *n* features, we are able to make use of simple dimensionality-reduction approaches such as principal component analysis (PCA) to find sets of three coordinates to best describe the bin-similarities. To do this efficiently, we make use of the fast PCA function within the Python machine learning package mlpy to compute only the first three components, enabling us to apply this to higher-resolution data for more detailed models without prohibitive computing time.

Results and discussion

HiC Unit of Interaction

One of the ways that HiFive achieves high resolution of HiC data lies in its treatment of DNA fragments that result from RE digestion of the genome. Both HiC and 5C experiments rely on fractionation of the genome by REs. Unlike 5C, however, HiC data is composed of reads that, theoretically, can map anywhere along the restriction fragments. It has been shown that fragment length is inversely related to interaction signal intensity [13]. In addition, the HiC assay maps reads with an orientation indicating which end of restriction fragments was ligated. With these two facts in mind, we assessed the similarity between interactions within 1 Mb of a set of fends compared to the adjacent fend on either side of them in the raw data and data corrected for distance-dependence, fend bias, and both. We found that fends originating from the same restriction fragment did not show any more similarity in their non-zero log-interactions with other fragments than adjacent fends originating from neighboring restriction fragments (Supplemental Figure 6). We concluded that the nature of the assay coupled with the filtering of reads results in fends that originate from the same fragment acting as independent units of interaction. As such, for all normalization of downstream analysis, they were treated independently.

The practical result of assessing fends independently is that the number of possible interacting sites is doubled. Further, this increases the number of interactions and therefore the interaction map resolution by a factor of four compared to assessing whole restriction fragments.

5C Performance

In order to assess HiFive's performance in processing 5C data, we applied HiFive's normalization to mouse 5C data produced by Nora, Lajoie [19] (Supplemental Table 1) and examined the model fit across a variety of factors. These included assessing the effects of normalization on the distance-dependence relationship, the differences between fragments across their signal strength and variance, the relationship between genomic characteristics and signal strength, and the reproducibility of bias correction values across replicates and methods.

After normalization with HiFive, 5C interactions showed an improved fit to the distance-dependence line (Supplemental Figure 5a). This improvement was seen in both the standard HiFive and HiFive-Express algorithms. Reduction in variance amongst reads of similar inter-fragment distances was particularly strong at low counts. The algorithm used (HiFive vs. HiFive-Express) made little difference in the improvement of fit.

To determine whether systemic biases associated with fragment characteristics were removed, two previously cited sources of systematic noise, fragment length and GC content [14] were assessed across individual reads as well as for fragment averages. In both cases, the magnitude of the effects (slope of the regression line) was reduced, regardless of the correction algorithm (Supplemental Figure 7). Additionally, there was a marked reduction in signal variance for fragment means for both approaches (Supplemental Figure 8). While fragments showed a wide range of mean log interaction counts prior to bias correction, normalized mean log counts showed little difference and maintained a consistent but slightly lower variance between fragments for both correction algorithms.

We also evaluated HiFive's 5C normalization performance by examining consistency between 5C data and HiC data for the same corresponding region and cell type. At the same time, we compared HiFive against two other methods, the approach described in Nora, Lajoie [19], a variation of the approach using HiCLib as applied to 5C data described in Naumova, Imakaev [20], and raw 5C data. The resulting normalized values for each method, along with raw 5C data, were compared to HiC data covering the same region normalized using either HiFive or HiCPipe, both using normal and dynamic binning to account for the lower coverage in the HiC dataset. For the comparisons to the Nora *et al* approach, the distance-dependent signal was removed from both HiFive and HiCPipe-normalized HiC data prior to comparison as the Nora et al approach necessarily removes this portion of the signal. HiFive and HiCLib results were tested both with and without the distance-dependent signal portion present.

Normalizations performed using HiFive showed improved correlation between HiC and 5C data compared to alternative methods and raw 5C signal across both replicates when data was dynamically binned, regardless of HiC normalization method (Figure 3). HiFive-Express also performs well when comparing 5C data against dynamically binned HiC data. Further, in the absence of the distance-dependent portion of the signal, HiFive-Express still performs well, regardless of the HiC correction method. It is unclear, how valid comparisons to the normally binned HiC are, given the extreme differences in coverage between the two types of assays. The dynamically binned HiC, though, clearly recapitulates all of the same structural features seen in the 5C data, suggesting it is a better standard for validation against. Visually, the biases associated with individual fragments are clearly apparent as striations in the raw data. After correction with HiFive and HiFive-Express, these striations are almost entirely absent. Other methods still show marked striping indicative of incomplete fragment-bias removal.

HiC Performance

To assess HiFive's performance in normalizing HiC data, we applied HiFive's normalization approaches to data from Dixon, Selvaraj [24] (Supplemental Table 1) and examined the effects of correction across a number of factors. These included the following: correlations of GC content, sequence mappability, and fend length to signal; between-fend signal variation; and correction value differences between data replicates and algorithms. There were hardly any normalization effects on the distance-dependence relationship. For further details, see *Supplemental Materials: HiC normalization and distance-dependence*.

Examining the effects of fend length, GC content, and mappability, we find that only the former and latter show a strong influence on the mean signal of its associated fend, although average GC content does show some influence on signal after removing the distance-dependence portion of the signal (Supplemental Figure 10). Correction using HiFive and HiFive-Express both greatly reduced effects of length, mappability, and GC content to the extent that it was present. Only HiFive reduced the variance among fend means for counts with and without the distance-dependent signal present while HiFive-Express appeared to overcorrect if distance-dependence was still included in the signal when stats were calculated. When the distance-dependent signal was removed prior to finding fend means, the variance for interactions corrected with HiFive-Express was reduced to nearly zero, reflecting fundamental differences in the underlying transformation of counts between these two approaches for HiC data. These correction trends held for correction of both HindIII- and NcoI-produced data. Comparing these same effects on bias relationships in data corrected by HiCPipe, we found much higher variance and stronger bias relationships after normalization (Supplemental Figure 11). The large variance in counts is especially interesting given the mappability cutoff of that HiCPipe recommends of 50%. HiCNorm could not be included due to the unreasonable resources needed to produce fend-level corrections ad HiCLib was not included because its performance was comparable to HiFive's (see below).

Next, we examined the range of fend means and variance across all valid fends before and after correction. Raw fend means showed a large range of means and variances for counts within one Mb interaction range for both cis and trans interactions (Supplemental Figures 12 and 13). After correction with either HiFive or HiFive-Express, fend means and variances were highly consistent, although variances were slightly higher when HiFive-Express was used in both cis and trans interactions. This was true regardless of RE used. Conversely we see the trend of increasing means persist to a much larger percent in HiCPipe-corrected data for both RE-produced datasets (Supplemental Figures 14 and 15). We also found that fend variance for cis interactions was much more varied across bins.

To assess the effectiveness of HiFive's HiC normalization compared to other methodologies, we examined correlations of corrected data across datasets produced using different REs. To do this, normalized data was generated for each dataset using HiFive, HiFive-Express, HiCPipe [13], HiCLib [16], and HiCNorm [15]. Data were binned at four resolutions, 10 Kb, 25 Kb, 100 Kb and 1 Mb, and inter-dataset correlations were calculated across a series of maximum interaction distance ranges for cis interactions. Trans interaction correlations were also determined at the 1 Mb resolution.

At the 1 Mb and 100 Kb resolutions, HiFive showed superior performance to all other methods across all interaction (Figure 4a). HiCPipe performed nearly as well at these lower resolutions, followed by HiCNorm. HiFive-Express performed equal or better to HiCPipe using 100 Kb bins and just slightly worse than HiCPipe using 1 Mb bins across all interaction distance ranges. Neither HiCNorm nor HiCLib matched the performance of HiFive, HiFive-Express, or HiCPipe at the 1 Mb bin size. At the 100 Kb bin size, HiCLib was below the other methods and for all but the shortest interaction range cutoffs, HiCNorm performed nearly identically to HiCPipe. At 25 Kb, HiFive and HiFive-Express outperformed other methods, though HiFive-Express actually showed better inter-dataset correlations for short range interactions than the standard HiFive normalization. At the 10 Kb resolution, HiFive showed a slight performance dip relative to HiCPipe and HiFive-Express when comparing interactions under 200 Kb apart, though HiFive-Express still outperformed HiCPipe across all ranges and both showed better performance than

HiCNorm and HiCLib. Including interactions with ranges greater than 200 Kb, HiFive showed better agreement between datasets than HiCPipe.

We believe one of HiFive's strengths is its ability to predict expected values for unobserved interactions. To assess this in the context dataset correlations, we dynamically binned data at the 25 Kb, 100 Kb, and 1 Mb bin sizes and reran the correlation analysis between RE datasets (Figure 4b). HiCLib was excluded because dynamic binning requires predicted counts to calculate while HiCLib returns only corrected values. Across all size ranges and bin sizes, HiFive showed superior performance. HiFive-Express also performed the second best for all but 1 Mb binned data.

Although they were designed with short-range interactions and high-resolution analysis as their target purpose, HiFive and HiFive-Express still performed relatively well in normalizing trans interactions (Figure 4c). Interestingly, HiFive-Express performed better on this measure than HiFive, showing performance just slightly below that of HiCPipe. We also examined trans-interaction correlations after dynamically binning the trans interaction heatmaps. While all correlations increased, the performance of HiFive, HiFive-Express and HiCPipe more than double when using dynamically binned data. Further, HiFive showed superior performance to HiFive-Express, though neither quite reaches the level of correlations produced by HiCPipe for trans interactions. For further discussion on the two algorithmic approaches and their differing performances, see *Supplemental Materials: HiFive's HiC algorithm performance differences*.

The boundary index captures more significant features than the directionality index

To verify the performance of our boundary index statistic, we began by using a cutoff yielding 3,078 peaks and comparing these to the 3,051 TAD boundaries generated by Dixon, Selvaraj [24] across chromosomes 1-19. We found that there was a high amount of overlap between the two sets of boundary calls with more than 60% of the DI-based boundaries in regions with BI coverage falling within 40 Kb of a BI-based peak (Figure 5). We partitioned boundaries into overlapping and unique and then found occupancy profiles for them based on CTCF, Smc1, H3K4me3, and PolII occupancy sites as well as TSS locations (Supplemental Table 2) [25-27]. Occupancy data was found up and downstream of each and binned at 10 Kb intervals. In sites that were considered equivalent between the two methods, we found nearly identical profiles, although signal was higher for the transcriptionally associated features H3K4me3, PolII, and TSSs centered on BI peaks. We attribute this to finer-scale positioning of the boundary. Across sites unique to each method, the general profile shape was similar. Like the overlapping sites, though, we saw a stronger signal for boundary sites generated using the DI method, particularly at the boundary site itself. Interestingly, we saw an increased background signal in unique BI-called sites compared to unique DI boundaries. This may be indicative of BI-called sites having a higher sensitivity for finding structural features occurring within domains as a higher background suggests additional structure features are occurring in relatively close proximity to these sites.

Three-dimensional modeling

HiFive's modeling strategy was assessed by determining the ability of its structural modeling approach to reproduce the fend-corrected interaction signal based on coordinate-distances alone for individual chromosomes as well as across the whole genome simultaneously. Two segmentations were used to create the distance matrices used for coordinate estimation, a partitioning based on BI peaks and a fixed-width binning producing a comparable number of breaks. After dynamically binning the data to create complete distance matrices, we used the mlpy fast-PCA algorithm to calculate the first three components as coordinates.

The PCA-based modeling approach, when coupled with dynamic binning for complete distance matrices shows a high degree of robustness. Correlations between models and data ranged from 82.5-88.9% (86.1% mean) and 82.7-89.2% (86.5% mean) for BI-peak and fixed width partitioning, respectively (Supplemental Figure 17). The whole genome models showed much lower correlations between model predictions and data at 44.5% and 41.7% for BI-peak and fixed width approaches, respectively. Comparisons between distance matrices for the two alternate binning approaches showed extremely high correlation with individual chromosome model correlations ranging from 98.27-99.97% (99.67% mean) and 99.39% correlation between whole genome models. This consistence in resulting models is also apparent when directly comparing the shaped and structures produced in 3D renderings of the chromosomes. Further results can be found in *Supplemental Materials: Three-dimensional modeling*.

Examining the physical shape of the modeled chromosomes, two distinct features are evident both in the individual and whole genome models (Figure 6a). The chromatin appears to exist predominantly as a combination of two different states: tight, dense coils occupying a small volume of space, and de-condensed stretches looped or folded back and forth creating accessible but still fairly low-volume compartments of chromatin. In both individual and whole genome models, there appears to be a polarization of states with the tightly coiled condensed chromatin orienting in the same general direction suggesting either a common characteristic or external organizer restricting the spatial arrangement of these condensed domains. While we feel confident in the organization of small structures suggested by this modeling approach, we do acknowledge that it is based on data produced from a heterogeneous population and a mix of interactions from pairs of homologous chromosomes. That being said, the smaller structures are likely to be stable and reproducible, whereas the whole chromosome shape or relative positioning amongst chromosomes is an amalgam of configurations representing general organizational trends that likely do hold for large portions of the assayed cell population.

Spatial organization of transcriptional activity

In order to assess the interplay between physical conformation of the chromatin and transcription, we used expression data from Shen, Yue [26], placing gene TSSs in their approximate locations in space according to our physical models. The physical placement of gene TSSs within the modeled chromatin, both on an individual and whole genome basis, showed a polarization of gene clusters corresponding to expression levels (Figure 6b and e). Genes with few or no detectable transcripts are seen clustered together and away from transcriptionally active genes, which are located on less condensed stretches of chromatin and tend towards the opposite region of chromosome-occupied space. This trend is also visible in the whole genome model, with the majority of low expression genes presenting in a sheet across the top of the model and active genes existing on parallel strands extending away from this sheet and converging in space. Interestingly, the highly compact coils appear mainly devoid of genes and are flanked by groups of inactive genes. To quantify this spatial organization, we examined the physical spacing of TSSs with respect to each other partitioned by expression level.

Using this approach we find that expression level is predictive of gene spacing such that low expression gene TSSs are spaced further apart from each other and from more highly expressed genes, given the intervening sequence between them, compared with pairs of more highly expressed genes. This holds true for both intra- and inter-chromosomal gene spacing, suggesting that not only are transcriptionally active genes occupying a separate space from less or non-transcribed genes but also that active genes are being brought together in a way that silent genes are not. It is unclear whether this is an active shepherding process or a consequence of the genes' transcriptional activity but these observations hold with other experimental data about the co-localization of co-expressed genes [28-30]. In addition, although clustering of inactive genes is not seen in individual chromosomes, we do see shorter distances amongst inactive or very low expression genes.

To examine additional organizing features, we painted the interaction models using a series of genomic annotations (Figure 6c, Supplemental Table 3). To mark regions associated with active PolII transcription we used H3K4me1 and H3K4me3 [6]. Repressive regions were marked with H3K9me3 and H3K27me3. Lamin-associated domains (LADs) were marked using Lamin B [31]. We also marked regions of RNA polymerase III (PolIII) activity with transcription factor for polymerase III C (TFIIIC) associated with PolIII [32]. Because these are regions associated with transcribing tRNAs and 5S RNAs which have both been shown to localize to the nucleolar periphery, this mark served as our surrogate for nucleolar-associated domains (NADs) [33]. There is an extreme polarization between LADs and TFIIIC regions. We also noted that TFIIIC marked regions showed an increased compaction compared to other actively transcribed regions. Active marks for PolII transcription occurred in chromatin regions surrounding TFIIIC marks and far from Lamin B. Repressive marks flanked Lamin B-associated regions and tended to show little association with TFIIIC. Based on these consistent findings across all chromosomes (Supplemental Figures 18-36), we propose that LADs and NADs serve as scaffolds in nuclear organization, allowing locally controlled formation of domains in stretches between LAD and

NAD connections but globally arranging positions of these domains to some extent along the radius of the nucleus (Supplemental Figure 37). This is a particularly intriguing possibility as Lamin A/C and BAF competitively compete with Lamin B chromatin binding regions and Lamin A/C has been shown to co-localize to the nucleolar periphery while Lamin B shows no such localization [34]. Coupled with the observation that some LADs overlap with identified NADs, this may serve as a mechanism for regulating pluripotency and lineage specificity [33, 35].

Conclusions

HiFive provides an easy-to-use, fast, and efficient framework for working with a variety of chromatin conformation data types. Because of the modular storage scheme, re-analysis and downstream analysis is made easier without additional storage requirements. We have attempted to make all aspects of the normalization procedures adjustable, allowing tuning of these analysis approaches to a user's specific needs. HiFive's approach relies solely on interaction data to find bias corrections, meaning that it is limited to higher sequencing depth experiments. However, as sequencing costs become cheaper we do not foresee coverage in experiments decreasing but rather the opposite. Although other methods like HiCPipe and HiCNorm may have comparable results in some scenarios, the output is more difficult to work and limited to a simple heatmap and model parameters. We also found that HiFive performs better in removing known sources of variance than HiCPipe at resolutions not practical with HiCNorm. HiFive provides support to any number of manipulations of the data on the fly by providing library of tools. HiFive also provides the only ready-to-use normalization approach for basic 5C data analysis. In addition to rigorous normalization of data, HiFive also provides users with alternative options for fast analysis with minimal computational requirements at only a slight accuracy cost, opening high-resolution HiC and 5C analysis to a much larger portion of the scientific community. HiFive can be downloaded at https://bitbucket.org/bxlab/hifive.

List of abbreviations used

TF: transcription factor; MPI: message parsing interface; Fend: fragment end; RE: restriction enzyme; DI: directionality index; BI: boundary index; ESC: embryonic stem cell; bp: base pair; Kb: kilobase; Mb: megabase; FPKM: fragments per kilobase of exon per million fragments mapped; TSS: transcription start site; GEO: Gene Expression Omnibus. 3C: chromosome conformation capture; 5C: chromosome conformation capture carbon copy; TFIIIC: transcription factor for polymerase III C.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MEGS, JEPC, VGC, and JT conceived the project and developed feature requirements. JEPC made significant contributions to the design of the 5C tools. MEGS developed all algorithms, designed and wrote all software, and wrote the manuscript. JT contributed to the manuscript and supported the project. All authors read and approved the final manuscript.

Additional data files

The following additional data files are available with the online version of the paper. Additional data file 1 contains a detailed methods section, three tables listing the sources of datasets used, and supplemental figures. Additional data file 2 contains HiFive documentation. Additional data file 3 contains a tar archive containing the HiFive library, including a set of useful scripts for using the library. Additional data file 4 contains a tar archive containing all scripts using to generate the data, analyses, and figures presented in this paper.

Acknowledgements

Research reported in this publication was supported by the National Institutes of Health under awards R01GM035463 to VC and R01DK065806 to JT and by American Recovery and Reinvestment Act (ARRA) funds through grant number RC2HG005542 to JT. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Arnone, M.I. and E.H. Davidson, *The hardwiring of development: organization and function of genomic regulatory systems.* Development, 1997. **124**(10): p. 1851-64.

2. Zinzen, R.P., et al., Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature, 2009. **462**(7269): p. 65-70.

3. He, A., et al., Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc Natl Acad Sci U S A, 2011. **108**(14): p. 5632-7.

4. Varriale, A., DNA Methylation, Epigenetics, and Evolution in Vertebrates: Facts and Challenges. Int J Evol Biol, 2014. **2014**: p. 475981.

5. Cantone, I. and A.G. Fisher, *Epigenetic programming and reprogramming during development.* Nat Struct Mol Biol, 2013. **20**(3): p. 282-9.

6. Kimura, H., *Histone modifications for human epigenome analysis.* J Hum Genet, 2013. **58**(7): p. 439-45.

7. Fullwood, M.J., et al., *Chromatin interaction analysis using paired-end tag sequencing.* Curr Protoc Mol Biol, 2010. **Chapter 21**: p. Unit 21 15 1-25.

8. Kalhor, R., et al., Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol, 2012. **30**(1): p. 90-8.

9. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.

10. Zhao, Z., et al., Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet, 2006.

38(11): p. 1341-7.

11. Dostie, J., et al., Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res, 2006. **16**(10): p. 1299-309.

12. 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.

13. Yaffe, E. and A. Tanay, Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet, 2011. **43**(11): p. 1059-65.

14. van Berkum, N.L. and J. Dekker, Determining spatial chromatin organization of large genomic regions using 5C technology. Methods Mol Biol, 2009. **567**: p. 189-213.

15. Hu, M., et al., *HiCNorm: removing biases in Hi-C data via Poisson regression.* Bioinformatics, 2012. **28**(23): p. 3131-3.

16. Imakaev, M., et al., Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods, 2012. **9**(10): p. 999-1003.

17. Jin, F., et al., A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature, 2013. **503**(7475): p. 290-4.

18. Hu, M., et al., *Bayesian inference of spatial organizations of chromosomes.* PLoS Comput Biol, 2013. **9**(1): p. e1002893.

19. Nora, E.P., et al., Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature, 2012. **485**(7398): p. 381-5.

20. Naumova, N., et al., *Organization of the mitotic chromosome.* Science, 2013. **342**(6161): p. 948-53.

21. Sanyal, A., et al., *The long-range interaction landscape of gene promoters.* Nature, 2012. **489**(7414): p. 109-13.

22. Phillips-Cremins, J.E., et al., Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell, 2013. **153**(6): p. 1281-95.

23. Rousseau, M., et al., Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinformatics, 2011. **12**: p. 414.

24. Dixon, J.R., et al., Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 2012. **485**(7398): p. 376-80.

25. Ren, B. *Mouse Encode Project at Ren Lab*. Available from: http://chromosome.sdsc.edu/mouse/download.html.

26. Shen, Y., et al., *A map of the cis-regulatory sequences in the mouse genome.* Nature, 2012. **488**(7409): p. 116-20.

27. Kagey, M.H., et al., Mediator and cohesin connect gene expression and chromatin architecture. Nature, 2010. **467**(7314): p. 430-5.

28. Zhao, Z.W., et al., Spatial organization of RNA polymerase II inside a mammalian cell nucleus revealed by reflected light-sheet superresolution microscopy. Proc Natl Acad Sci U S A, 2014. **111**(2): p. 681-6.

29. Zhang, Y., et al., Chromatin connectivity maps reveal dynamic promoter-enhancer long-

range associations. Nature, 2013. **504**(7479): p. 306-10.

30. Rieder, D., et al., Co-expressed genes prepositioned in spatial neighborhoods stochastically associate with SC35 speckles and RNA polymerase II factories. Cell Mol Life Sci, 2014. **71**(9): p. 1741-59.

31. Shimi, T., et al., The A- and B-type nuclear lamin networks: microdomains involved in chromatin organization and transcription. Genes Dev, 2008. **22**(24): p. 3409-21.

32. Carriere, L., et al., Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells. Nucleic Acids Res, 2012. **40**(1): p. 270-83.

33. Nemeth, A., et al., *Initial genomics of the human nucleolus.* PLoS Genet, 2010. **6**(3): p. e1000889.

34. Kind, J. and B. van Steensel, Stochastic genome-nuclear lamina interactions: Modulating roles of Lamin A and BAF. Nucleus, 2014. **5**(2): p. 124-130.

35. van Koningsbruggen, S., et al., High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Mol Biol Cell, 2010. **21**(21): p. 3735-48.

Figures

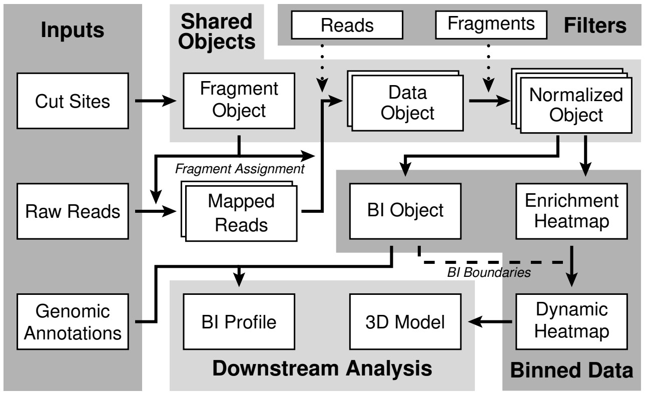

Figure 1 - The software architecture of HiFive. Self-contained units of information are denoted by boxes, with solid arrows denoting dependencies for object creation. The split line marked ‘fragment assignment’ depends on the type of data being handled and acts upstream or downstream of the mapped reads objects for 5C and HiC, respectively. Dotted lines denote filters limiting information passed from one object to the next. The dashed line denotes an optional input.

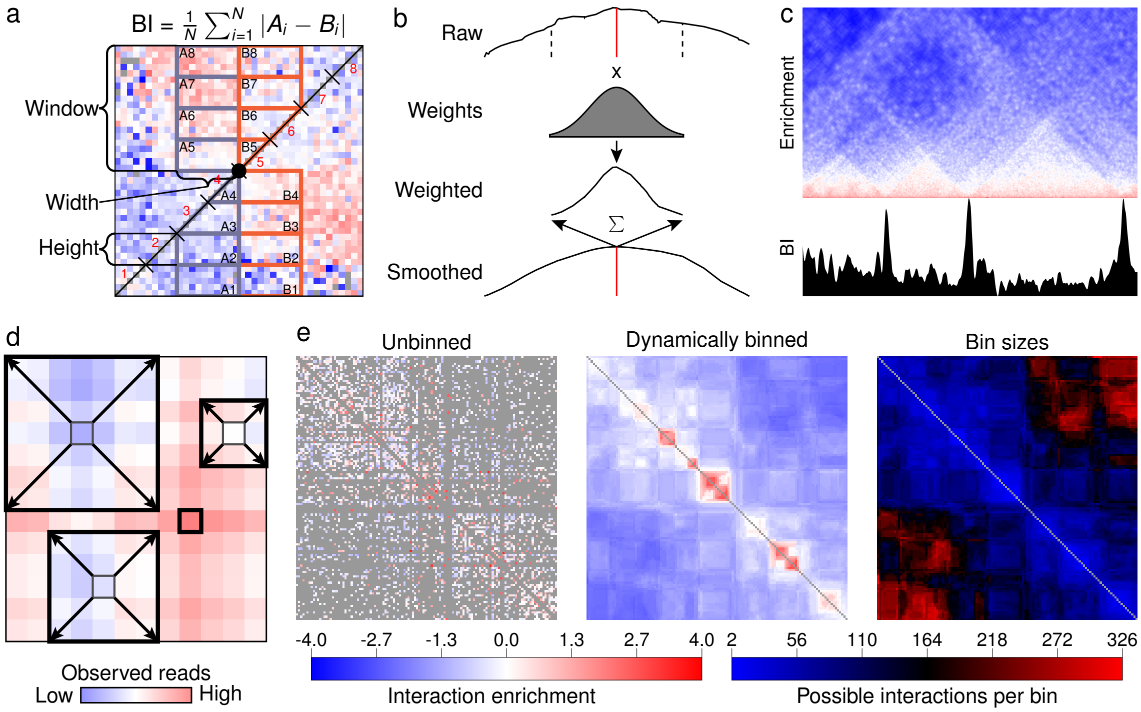

Figure 2 - Expanding data interpretation using dynamic binning and the boundary index statistic. a) A schematic detailing the calculation of the boundary index. The genomic coordinates (along the diagonal) are portioned into regions (red numbers). For each region, all fend interactions are found with fends within some width up (A) and downstream (B) of the boundary point (black circle). b) BI smoothing is accomplished by applying Gaussian-distributed weights to a set of surrounding points and finding the weighted mean. c) BI peaks correspond to changes in enrichment patterns. d) Dynamic binning considers each bin individually, expanding its borders in all directions until a user-defined minimum number of observations have been incorporated or a maximum bin size has been reached. This results in different sized bins containing roughly equivalent amounts of data. e) Unbinned data is shown with unobserved interactions in gray. Dynamically binned data is in the same color scale but has no empty bins as they have been expanded as needed to incorporate sufficient reads. The size of each bin is indicated by the number of potential interactions (both observed and unobserved) that are now included in it.

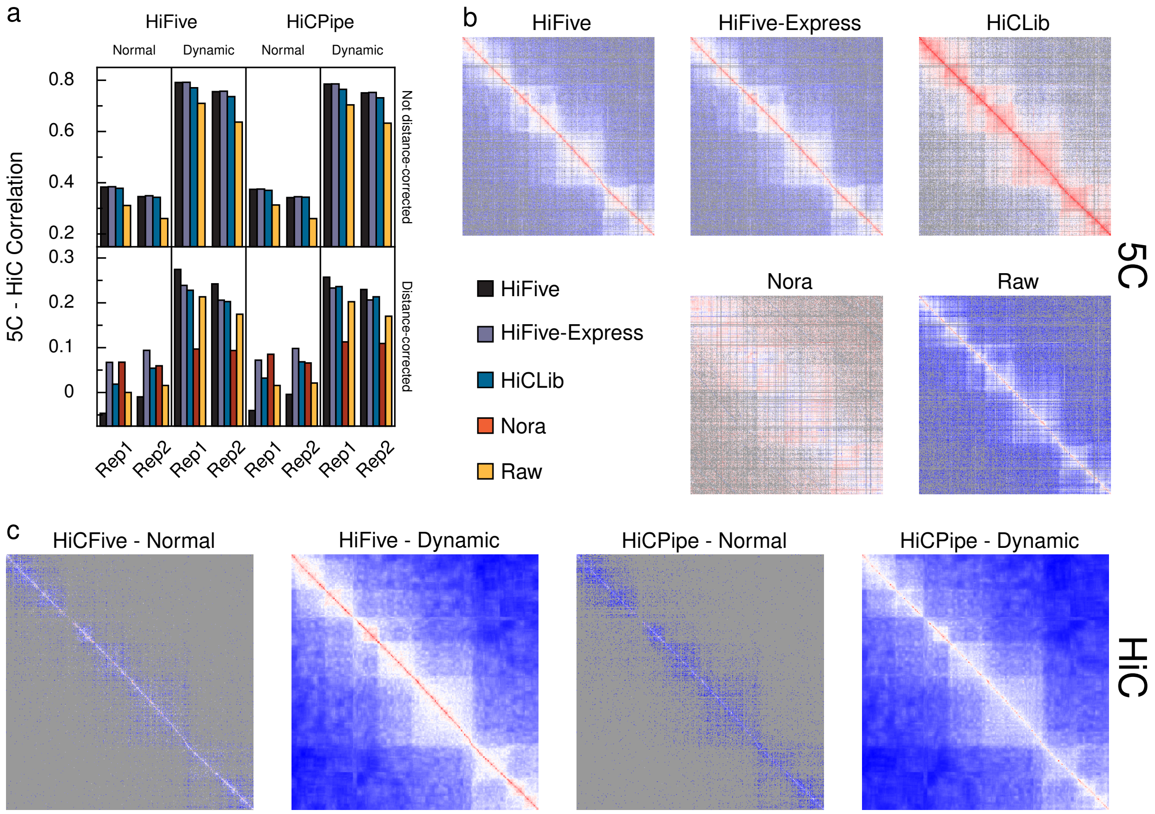

Figure 3 - 5C data normalization and correlation with corresponding HiC data. a) Methods in the upper panel were compared without removal of the distance-dependent portion of the signal, whereas the lower panel shows the correlations in which the distance-dependent signal was removed. b) Visualization of normalized and logged 5C counts. Colors are scaled to maximize the dynamic range, with blue corresponding to the lowest counts, red to the highest, and white to the midpoint between the two. Gray denotes interactions where no reads were observed. Rows and columns correspond to forward and reverse probes, respectively. c) HiC data corrected using HiFive or HiCPipe and binned using the same boundaries as the 5C data. Coloring is as described for b. Dynamically binned HiC data show bins sized to include 20 interactions per bin.

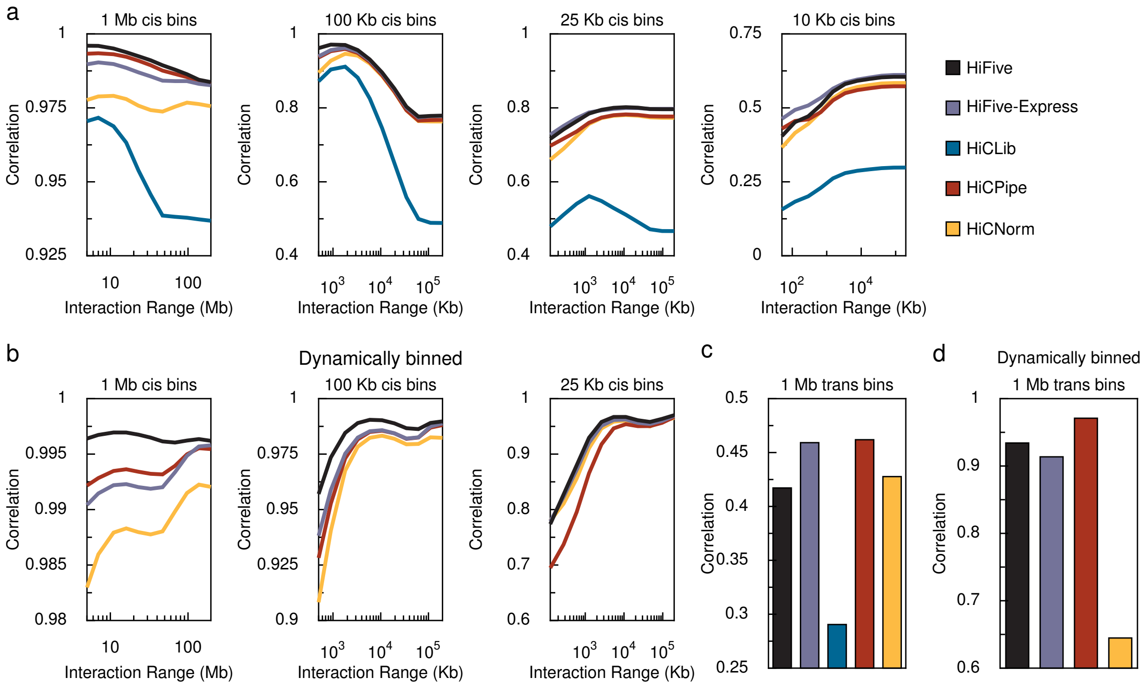

Figure 4 - HiC normalization and inter-dataset correlation. a) HiC data from mouse ESCs produced using HindIII and NcoI were normalized using a number of different analysis methods and intra-chromosomal interaction correlations were compared across a range of bin sizes and maximum interaction distances between the two datasets. Interaction range indicates the largest interaction distance included in for the correlation calculation. b) Heatmap data were dynamically binned prior to calculating inter-dataset correlations. Interaction rangers are as described for panel a. c) Correlation between 1Mb-binned inter-chromosomal interactions across datasets after normalization using a variety of methods. d) Inter-chromosome heatmaps were dynamically binned and inter-dataset correlations were calculated for each of the shown methods.

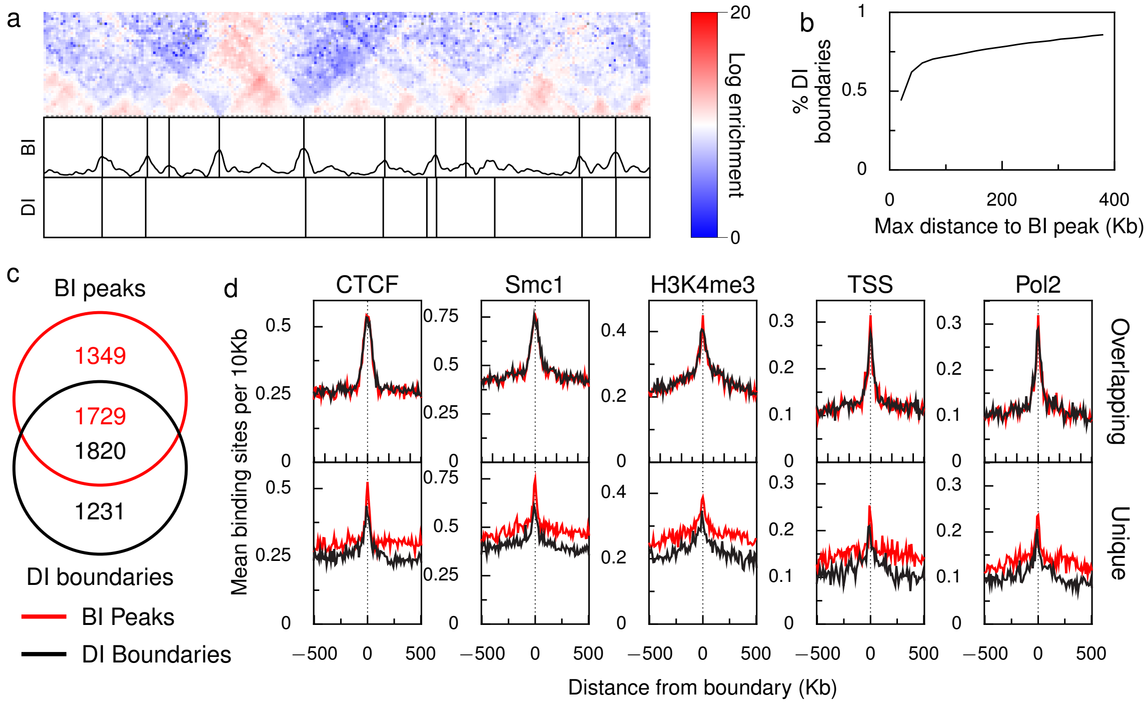

Figure 5 - Boundaries identified using boundary index scoring and associated signals. Boundaries identified from peaks in pooled HindIII BI signal compared to topological domain boundaries found using DI. a) Interaction enrichment signal for a 5 Mb stretch of chromosome 19 and its associated BI signal, BI peaks, and DI domain boundaries. b) The percentage of DI boundaries that have a BI peaks within a given window size. c) Overlap of BI peaks and DI boundary sets using a 40 Kb cutoff for defining overlap. d) Frequency of annotation data peaks across a 1 Mb window centered on each boundary or peak and binned every 10 Kb.

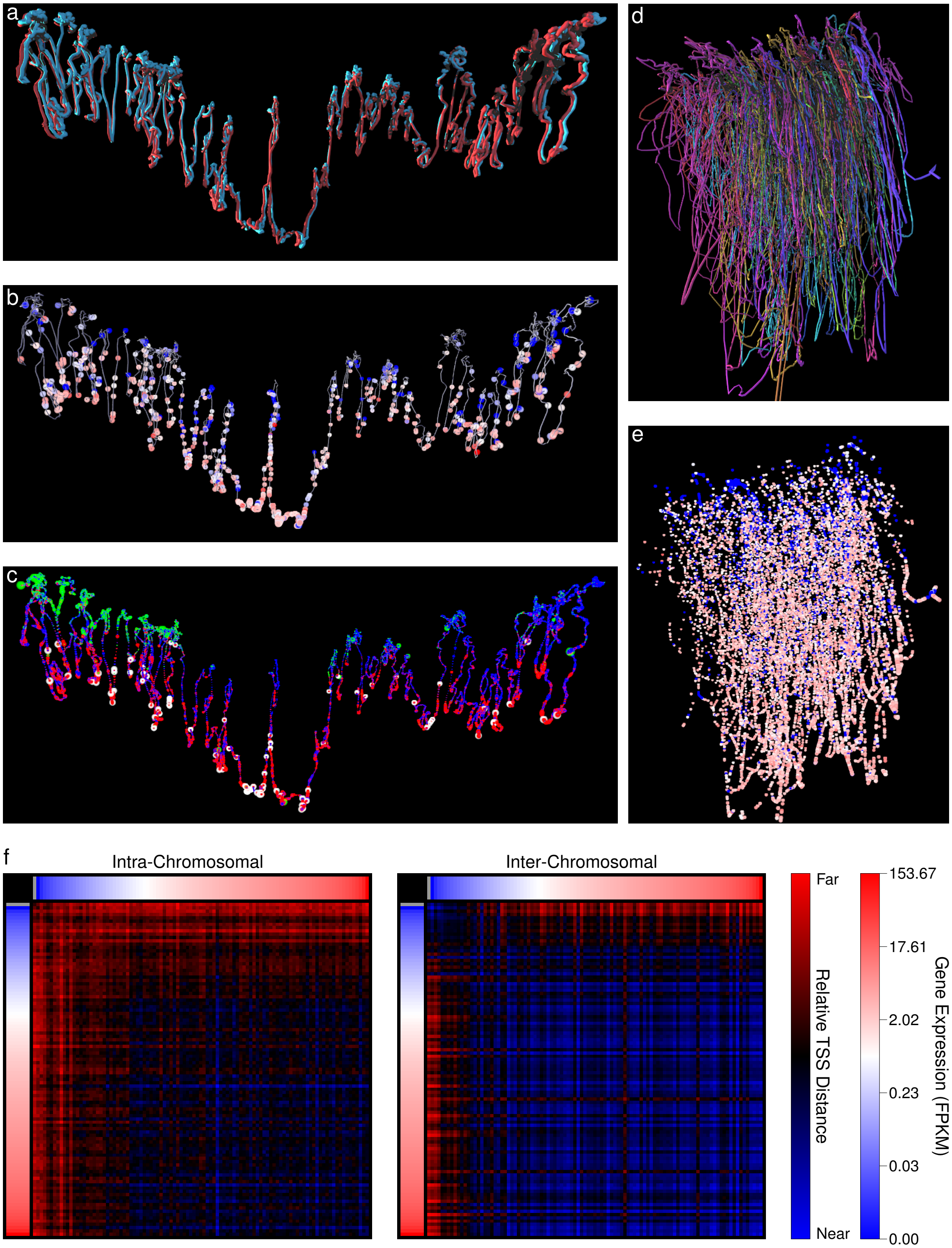

Figure 6 - Three-dimensional interpretations of structural data reveal differential spatial organization of genes by activity. a) A rendering of the chromosome 3 models showing the fixed width bin model in blue and the BI-peak bin model in red. The dotted white line indicates the section blown up in subfigure b. Strand thickness is proportional to the square root of the ratio of

physical spacing of the points versus the sequential distance, such that a thicker strand when a longer sequence length occurs between closer 3D endpoints. b) A skeletal rendering of chromosome 3 with points indicating gene TSSs. Genes are color coded by expression level in FPKMs. c) A model painted with binned annotation data. H3K4me1 and H3K4me3 are shown in red. H3K27me3 and H3K9me3 are shown in blue. LaminB is shown in green. Transcription factor for polymerase IIIC is shown in white. Sizes are proportional to signal strength and signals are binned at 10 Kb intervals. d) A whole genome model rending chromosomes rendered in different colors and proportional strand thickness. e) All mouse genes in the same physical configuration as in panel b and colored according to expression levels. f) Mean relative distances between gene TSSs, binned by expression levels (top and left of plots). All non-observed genes are in a single bin and colored gray in the mean expression scale. Intra-chromosomal distances were determined by calculating the log-ratio of the mean physical distance over the mean sequential distance. Inter-chromosomal distances are simply the mean log-physical distances between TSSs.