Measuring significant changes in chromatin conformation with ACCOST

Kate B. Cook[1], Borislav H. Hristov[1], Karine G. Le Roch[2], Jean Philippe Vert[3,4] and William Stafford Noble [1,5,*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065, USA, [2]Department of Cell Biology, University of California, Riverside, CA 92521, USA, [3]Google Brain, Paris, 75009, France, [4]Centre for Computational Biology, MINES ParisTech, PSL University, Paris, 75009, France and [5]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA 98195-2355, USA

ABSTRACT

Chromatin conformation assays such as Hi-C cannot directly measure differences in 3D architecture between cell types or cell states. For this purpose, two or more Hi-C experiments must be carried out, but direct comparison of the resulting Hi-C matrices is confounded by several features of Hi-C data. Most notably, the genomic distance effect, whereby contacts between pairs of genomic loci that are proximal along the chromosome exhibit many more Hi-C contacts that distal pairs of loci, dominates every Hi-C matrix. Furthermore, the form that this distance effect takes often varies between different Hi-C experiments, even between replicate experiments. Thus, a statistical confidence measure designed to identify differential Hi-C contacts must accurately account for the genomic distance effect or risk being misled by large-scale but artifactual differences. ACCOST (Altered Chromatin COnformation STatistics) accomplishes this goal by extending the statistical model employed by DEseq, re-purposing the 'size factors,' which were originally developed to account for differences in read depth between samples, to instead model the genomic distance effect. We show via analysis of simulated and real data that ACCOST provides unbiased statistical confidence estimates that compare favorably with competing methods such as diffHiC, FIND and HiCcompare. ACCOST is freely available with an Apache license at https://bitbucket.org/noblelab/accost.

INTRODUCTION

An increasing number of experimental techniques—including ChIA-PET ([1]), Hi-C ([2]), Hi-ChIP ([3]), PLAC-seq ([4]), SPRITE ([5]) and GAM ([6])—allow for the high-throughput characterization of pairwise chromatin contacts. These techniques have helped to elucidate the roles that chromatin 3D architecture play in critical cellular processes such as gene regulation, DNA replication and splicing. Notably, aspects of chromatin structure have been implicated in the etiology of several disease phenotypes, in which genetic modifications induce changes in chromatin structure, which in turn induce changes in gene expression ([7]).

Accordingly, a key statistical challenge for 3D chromatin analyses is to assign statistical confidence measures to observed differences in chromatin structure. Such differences may arise between different developmental stages, cell types, individuals or disease states. Empirically, chromatin structure exhibits differences at multiple scales, including differences in nuclear shape and volume, full chromosomes, compartments, topologically associating domains (TADs), sub-TADs or promoter-enhancer loops. In this work, we focus on calling differential interactions at the finest level, i.e. for a given contact matrix defined with respect to genomic loci ('bins') of size w bp, we ask whether the observed contact count associated with two w-bp loci x and y in experimental condition A is significantly different from the corresponding (x, y) contact count from condition B.

A successful method for assigning confidence estimates to differential interactions should exhibit four key properties. First, the method must control the false discovery rate (FDR) associated with an accepted set of differential interactions. Controlling FDR, as opposed to family-wise error rate, makes sense in the context of most 3D chromatin studies, which are often hypothesis-generating studies aimed to produce a large set of discoveries. Second, a good method must take into account the particular biases associated with the assay (such as Hi-C) that generated the data. As we will see, accomplishing this goal turns out to be very challenging. Third, because of the high cost associated with generating 3D chromatin data—in particular, the cost of sequencing deeply enough to characterize all pairs of loci in

*To whom correspondence should be addressed. Tel: +1 206 221 4394; Fax: +1 206 685 7301; Email: wnoble@uw.edu
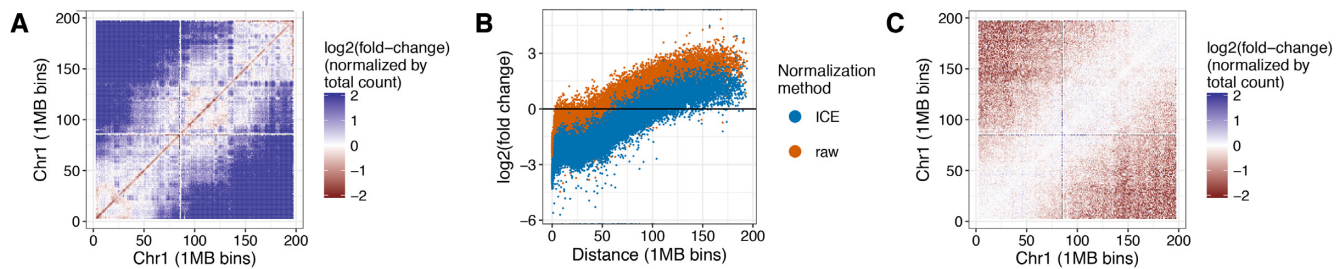
Figure 1. Genomic distance effect leads to artifactual differences in Hi-C contacts. (**A**) The figure plots the log fold-change in Hi-C contact counts between embryonic stem cell (ESC) and cortex Hi-C data (see Methods for details), plotted at 10 kb resolution on human chromosome 1. (**B**) Scatter plot of Hi-C log-fold change as a function of genomic distance, where each point corresponds to a pair of 10 kb loci. The two point colors correspond to raw counts (red) and counts normalized using ICE (blue). (**C**) Similar to panel A, but for two replicate Hi-C experiments performed in mouse cortex.

the genome—an ideal method will be capable of leveraging structure in the observed data to enable calling significant differential interactions even when only a single pair of experiments has been performed. Fourth, as with essentially any statistical testing procedure, an ideal method should maximize statistical power; i.e. the method should identify as many differential interactions as possible, given the data, while still controlling the FDR.

Note that, although multiple genome-wide 3D chromatin assays have been developed, most existing statistical testing frameworks focus on the first and still most widely used assay, Hi-C. Accordingly, we focus here on Hi-C analysis, though many of the methods described here may also apply directly to data generated using assays such as ChIA-PET, SPRITE, Hi-ChIP or PLAC-seq.

Early Hi-C studies generally employed simple fold-change statistics to identify differential contacts (8,9), but these methods have subsequently been supplanted by a variety of more statistically sophisticated techniques.

- HOMER (10) implements a simple test for differential interactions, based on binomially-distributed counts. HOMER can only consider comparisons between two libraries and cannot take into account replicate experiments.
- ChromoR (11) carries out variance stabilization of Hi-C data using the the Haar-Fisz transform (12), which transforms Poisson distributed data into Gaussian distributed coefficients at multiple scales. The resulting coefficients are subjected to wavelet-based denoising. ChromoR then analyzes the resulting transformed data using either change-point analysis for the purpose of detecting chromosomal aberrations or domain boundaries, or Bayes factor analysis for the detection of differential contacts.
- HiBrowse (13) is a Python tool with a web interface that employs the edgeR RNA-seq package (14) to compare pairs of Hi-C experiments.
- The R package diffHic (15) also uses the edgeR statistical framework but dispenses with the web interface. DiffHic offers a range of functionality, including mapping, quality filtering, binning, normalization and estimation of statistical confidences for differential contacts. Notably, the package is able to accommodate complex experimental designs, including paired or blocked designs and experiments involving more than two groups.

- FIND (16) uses a spatial Poisson model to measure differences in chromatin loops, with the aim of relaxing the independence assumption between adjacent loci.
- HiCcompare (17) normalizes the data and then tests for differential interactions on the normalized data directly, using a Z-score comparison. HiCcompare does not leverage replicate experiments.

For all of these methods, a key challenge is accurately accounting for the unique features of Hi-C data. For example, several widely used methods can normalize Hi-C data to account for positional biases that arise due to GC content, mappability and density of restriction-enzyme sites along the genome (18,19). Variants of these normalization routines are integrated into all of the methods outlined above.

A second prominent and potentially problematic feature of Hi-C data is the 'genomic distance effect,' in which pairs of loci that are close together along the genome are very likely to be observed in contact simply due to the random polymer looping behavior of DNA. The genomic distance effect manifests itself most obviously in the form of a marked enrichment of high counts near the diagonal in the Hi-C contact matrix.

This work was primarily motivated by the observation that the genomic distance effect can vary markedly between two Hi-C experiments. The result, in a Hi-C matrix showing log fold-changes in contact counts, is an enrichment of apparently differential contacts far from the diagonal (Figure 1A). Notably, these differences remain even after normalizing the data using a method such as iterative correction and eigenvalue decomposition (ICE) (18) (Figure 1B). The source of this kind of large-scale change in genomic distance effect is not always clear. In some settings, such differences may indicate changes in chromatin compaction (20) or differences in enrichment of various stages of the cell cycle (21). Alternatively, differences in the genomic distance effect may reflect experimental artifacts: we have observed non-trivial differences even between replicate Hi-C runs (Figure 1C). Regardless of the source of these observed differences, we reasoned that any method aiming to identify local changes in chromatin structure should attempt to ignore large-scale differences in the genomic distance effect.

One other method—HiCcompare, which was published while our method was under development—also attempts to correct for differences in genomic distance effect. HiCcompare works by plotting the log fold change in contact

counts as a function of genomic distance, and then fitting a LOESS regression curve to this plot. The regression curve is then used to normalize the observed interaction counts. Differential contacts are then identifying by converting the normalized values to Z-scores and then computing P-values using the standard normal distribution.

In this paper, we introduce ACCOST (Altered Chromatin COnformation STatistics), which assigns statistical confidence estimates to differences in Hi-C contacts, while taking into account differences in the genomic distance effect. Specifically, ACCOST leverages counts at similar genomic distances to estimate the variance of the contact counts, similar to how tools such as DESeq (22) estimate gene expression variance by grouping together genes with similar expression. ACCOST includes distance-dependent normalization factors, thereby allowing for accurate measurement of differential contact counts at short and long ranges. The method accommodates but does not require biological replicates. We demonstrate, via simulation and analysis of real Hi-C data, that ACCOST delivers unbiased statistical confidence estimates while successfully controlling for systematic changes in the genomic distance effect.

MATERIALS AND METHODS

Model

Formally, we consider m contact count matrices $C^k \in \mathbb{N}^{n \times n}$, for $k = 1, \ldots, m$, where $C_{i,j}^k$ is the interaction count between loci i and j in the kth experiment (for $1 \leq i, j \leq n$ and $1 \leq k \leq m$). Each experiment is done in one of two conditions $\mathcal{A} = \{A, B\}$, and we denote by $\rho(k) \in \mathcal{A}$ the condition corresponding to the kth contact count matrix, for $1 \leq k \leq m$. Let $M(A) = \rho^{-1}(A)$ be the set of matrices done in condition A and $M(B) = \rho^{-1}(B)$ those done in condition B. If we denote by $m_A = |M(A)|$ the total number of contact count matrices done in condition A and by $m_B = |M(B)|$ those in B, it follows that $m_A + m_B = m$.

We model the count $C_{i,j}^k$ as a negative binomial (NB) random variable with mean $\mu_{i,j}^k$ and variance $\left(\sigma_{i,j}^k\right)^2$. In order to be able to estimate the mean and the variance from few Hi-C replicates, we make the following assumptions, inspired by the assumptions of the DESeq model for RNA-seq data (22):

- The mean $\mu_{i,j}^k$, that is the expected number of counts between loci i and j in the kth matrix corresponding to condition $\rho(k)$, is equal to a quantity $q_{i,j}^{\rho(k)}$ which represents the 'true' (but unknown) number of interactions between loci i and j in condition $\rho(k)$, multiplied by an experiment- and loci-specific factor as follows:

$$\mu_{i,j}^k = \beta_i^k \beta_j^k s_{|i-j|}^k q_{i,j}^{\rho(k)}, \qquad (1)$$

where $\beta^k \in \mathbb{R}^n$ is an experiment-specific vector of loci-specific biases to account for various biological and technical biases (18) and $s^k \in \mathbb{R}^n$ is a vector of distance-specific size factors to account for the experiment-specific 'genomic distance effect'. This generalizes the notion of size factors used to correct for library size in RNA-seq

(22), but considering each distance and pair of loci separately.

- As in the DESeq model, the variance is the sum of a shot noise term and of a raw variance term:

$$\left(\sigma_{i,j}^k\right)^2 = \mu_{i,j}^k + \left(\beta_i^k \beta_j^k s_{|i-j|}^k\right)^2 v^{\rho(k)}\left(q_{i,j}^{\rho(k)}\right), \qquad (2)$$

where v^A is a smooth non-negative function, for $A \in \mathcal{A}$.

We now explain how we estimate the parameters β, s, q and v of the model from a set of count matrices $C^1, \ldots, C^m$:

- For the vectors of locus-specific biases β^k (for $k = 1, \ldots, m$), we borrow a standard ICE normalization estimator $\hat{\beta}^k$ performed independently on each count matrix (18).
- For the distance-specific size factors s^k (for $k = 1, \ldots, m$), we enforce that the median normalized count for pairs of bins at each given distance in each matrix is the same, by taking for $d \in [0, n-1]$ and $k \in [1, m]$:

$$\hat{s}_d^k = \underset{|i-j|=d}{\text{median}} \frac{C_{i,j}^k}{\hat{\beta}_i^k \hat{\beta}_j^k}.$$

- For the normalized count matrix q^A, for $A \in \mathcal{A}$, we first define the normalized count matrices of each experiment:

$$\forall k \in [1, m], \forall i, j \in [1, n], \quad \hat{q}_{i,j}^k = \frac{C_{i,j}^k}{\hat{\beta}_i^k \hat{\beta}_j^k \hat{s}_{|i-j|}^k},$$

and average these estimates over replicates for each condition:

$$\forall A \in \mathcal{A}, \forall i, j \in [1, n], \quad \hat{q}_{i,j}^A = \frac{1}{m_A} \sum_{k \in M(A)} \hat{q}_{i,j}^k.$$

- To estimate the normalized raw variance functions v^A (for $A \in \mathcal{A}$), we start by estimating the variance of $\hat{q}_{i,j}^A$ for any $i, j \in [1, n]$. Since we are in situations where the number of replicates m_A can be small (down to $m_A = 1$), we propose to leverage information within each matrix to estimate this variance. For each bin (i, j), we consider a set of other bins $N(i, j) \subset [1, n]^2$ where we believe $q_{i,j}^A$ should be almost constant. In the ACCOST implementation, we choose $N(i, j)$ to be the set of all other bins (u, v) at the same distance, i.e., $|u - v| = |j - i|$. This choice is motivated by the overwhelming impact of the distance effect on the contact count. Given such a neighborhood, we form the estimates:

$$\hat{t}_{i,j}^A = \frac{1}{m_A |N(i, j)|} \sum_{k \in M(A)} \sum_{(u,v) \in N(i,j)} \hat{q}_{u,v}^k,$$

$$\hat{w}_{i,j}^A = \frac{1}{m_A |N(i, j)| - 1} \sum_{k \in M(A)} \sum_{(u,v) \in N(i,j)} \left(\hat{q}_{u,v}^k - \hat{t}_{i,j}^A\right)^2.$$

Assuming that $\hat{\beta}^k$ and $\hat{s}^k$ are exact and nonrandom estimators of β^k and s^k, on the one hand, and that $q_{u,v}^A = q_{i,j}^A$ for all $(u, v) \in N(i, j)$, on the other hand, we deduce from our model (1)-(2) that the expectation and variance of $\hat{q}_{u,v}^k$

for $(u, v) \in N(i, j)$ and $k \in M(A)$ are respectively given by:

$$E\hat{q}_{u,v}^k = q_{i,j}^A,$$

$$Var\left(\hat{q}_{u,v}^k\right) = \frac{q_{i,j}^A}{\hat{\beta}_u^k \hat{\beta}_v^k \hat{s}_{|u-v|}^k} + v^A\left(q_{i,j}^A\right).$$

A simple computation shows that if $X_1, \ldots, X_N$ are independent random variables with the same mean and different variances, then:

$$E \frac{1}{N-1} \sum_{i=1}^N \left(X_i - \frac{1}{N} \sum_{j=1}^N X_j \right)^2 = \frac{1}{N} \sum_{i=1}^N Var(X_i).$$

Using this result with the random variables $\{\hat{q}_{u,v}^k : (u, v) \in N(i, j), k \in M(A)\}$ leads to:

$$E\hat{w}_{i,j}^A = v^A\left(q_{i,j}^A\right) + \frac{q_{i,j}^A}{m_A|N(i, j)|} \sum_{k \in M(A)} \sum_{(u,v) \in N(i,j)} \frac{1}{\hat{\beta}_u^k \hat{\beta}_v^k \hat{s}_{|u-v|}^k}.$$

Let us define

$$\hat{z}_{i,j}^A = \frac{\hat{q}_{i,j}^A}{m_A|N(i, j)|} \sum_{k \in M(A)} \sum_{(u,v) \in N(i,j)} \frac{1}{\hat{\beta}_u^k \hat{\beta}_v^k \hat{s}_{|u-v|}^k},$$

then $E\left[\hat{w}_{i,j}^A - \hat{z}_{i,j}^A\right] = v^A\left(q_{i,j}^A\right)$, and we can estimate v^A by regressing $\hat{w}_{i,j}^A - \hat{z}_{i,j}^A$ as a polynomial function of $\log(\hat{q}_{i,j}^A)$, encouraging smoothness. Negative values of $w - z$ are rare and are filtered out by the algorithm.

To test for differential count between conditions at a given locus (i, j), we define as test statistics the total counts in different conditions:

$$\forall A \in \mathcal{A}, \quad C_{i,j}^A = \sum_{k \in M(A)} C_{i,j}^k.$$

$C_{i,j}^A$ is a sum of m_A independent NB-distributed random variables. We approximate its distribution by an NB distribution with mean and variances:

$$E(C_{i,j}^A) = \sum_{k \in M(A)} EC_{i,j}^k, \quad Var(C_{i,j}^A) = \sum_{k \in M(A)} Var(C_{i,j}^k).$$

Under the null hypothesis where $q_{i,j}^A = q_{i,j}^B$, we can get an estimate of the shared normalized count:

$$\hat{q}_{ij}^0 = \frac{1}{m} \sum_{i=1}^m \hat{q}_{ij}^k,$$

from which we deduce an estimate of the mean and variance of $C_{i,j}^A$ for $A \in \mathcal{A}$:

$$\hat{\mu}_{i,j}^A = \sum_{k \in M(A)} \hat{\beta}_i^k \hat{\beta}_j^k \hat{s}_{|i-j|}^k \hat{q}_{ij}^0,$$

$$(\hat{\sigma}_{i,j}^A)^2 = \hat{\mu}_{i,j}^A + \sum_{k \in M(A)} \left(\hat{\beta}_i^k \hat{\beta}_j^k \hat{s}_{|i-j|}^k \right)^2 \hat{v}^A(\hat{q}_{ij}^0).$$

Given the two conditions $\mathcal{A} = \{A, B\}$, we then follow the technique of DESeq to obtain a P-value by conditioning on

the total count $C_{i,j} = C_{i,j}^A + C_{i,j}^B$:

$$p_{i,j} = \frac{\sum_{a+b=C_{i,j}, p(a,b) \leq p(C_{i,j}^A, C_{i,j}^B)} p(a, b)}{\sum_{a+b=C_{i,j}} p(a, b)},$$

where

$$p(a, b) = P\left(C_{i,j}^A = a | \hat{\mu}_{i,j}^A, (\hat{\sigma}_{i,j}^A)^2\right) P\left(C_{i,j}^B = b | \hat{\mu}_{i,j}^B, (\hat{\sigma}_{i,j}^B)^2\right)$$

$$= NB\left(a; \hat{\mu}_{i,j}^A, (\hat{\sigma}_{i,j}^A)^2\right) NB\left(b; \hat{\mu}_{i,j}^B, (\hat{\sigma}_{i,j}^B)^2\right),$$

since $P(C_{i,j}^A = a)$ and $P(C_{i,j}^B = b)$ are independent under the null hypothesis.

Data sets

For validation and comparison of methods, we used Hi-C data from human, mouse and the malaria parasite *Plasmodium falciparum*.

- The human Hi-C data is derived from two cell lines, GM12878 and IMR90 (23). The data were downloaded in text form using the `juicer dump` command of Juicer tools (24). Replicates are numbered as in Supplementary Table S1 of (23), heading 'biological replicate number.'
- Two mouse Hi-C data sets were analyzed. The two data sets were generated in the same lab, but were produced for two different studies. The first data set is derived from mouse embryonic stem cells (mESCs) (25), and the second is from mouse cortex (26).
- The *Plasmodium* Hi-C data represents parasites at two distinct stages in the *P. falciparum* life cycle, trophozoite and sporozoite (27).

All three data sets were analyzed at 10 kb resolution.
All datasets were ICE normalized using `iced` (https://github.com/hiclib/iced).

Competing methods

For analysis by diffHiC, contact count matrices were loaded and transformed into ContactMatrix objects. Following the diffHiC documentation, we filtered by average count and applied LOESS normalization before using diffHiC to estimate *P*-values. As prescribed in the diffHiC manual, we applied LOESS normalization separately to the near-diagonal bin pairs (defined as pairs with genomic distance below 50 Mb) and to other pairs.

Following the FIND documentation, we loaded contact count matrices and ran FIND with parameters windowSize = 3, method = hardCutoff, and alpha = 0.7.

Following the HiCcompare documentation, we loaded contact count matrices and ran HiCcompare with parameters A.min = 15 and adjust.dist = true.

RESULTS

ACCOST correctly accounts for the genomic distance effect

ACCOST extends the statistical model used by DEseq (22) to explicitly capture observed differences in the genomic distance effect between different Hi-C experiments. The key
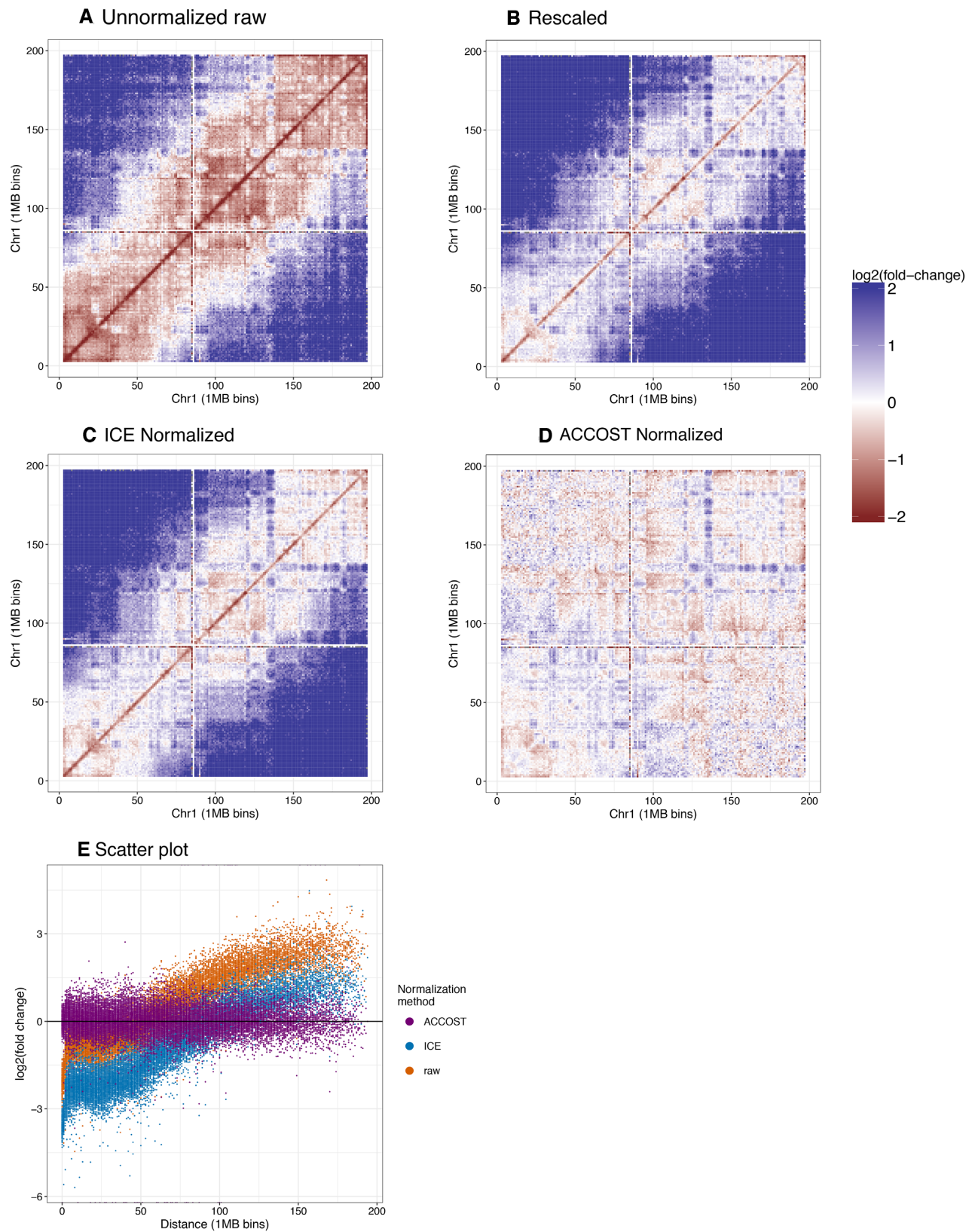
Figure 2. Genomic distance normalization highlights differences in contact counts. Panels A–D plot the log fold change of Hi-C contacts between ESC and cortex on chromosome 1. (**A**) Raw data. (**B**) Data after rescaling the two matrices to have the same total count, to account for differences in sequencing depth. (**C**) Data after ICE normalization. (**D**) Data after ACCOST normalization, which additionally accounts for the genomic distance effect. (**E**) Similar to Figure 1B, but including ACCOST.
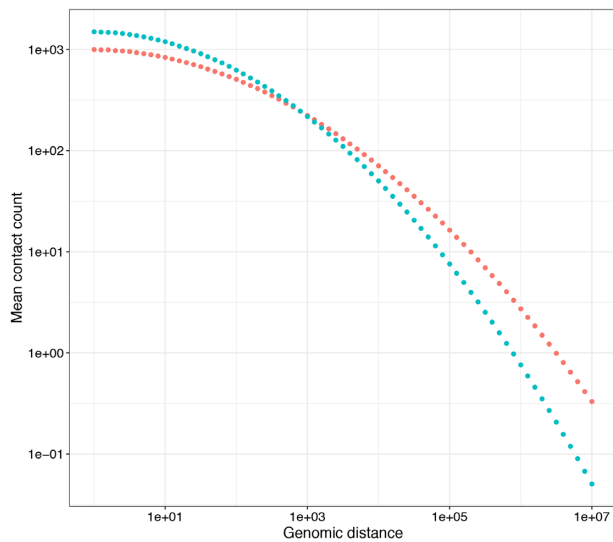
Figure 3. Simulated distance decay curves. The two decay curves were constructed such that the mean contact count between neighboring genomic bins was 1000 and 1500 for the two distributions. No position-specific biases (as are corrected for using ICE) were added.
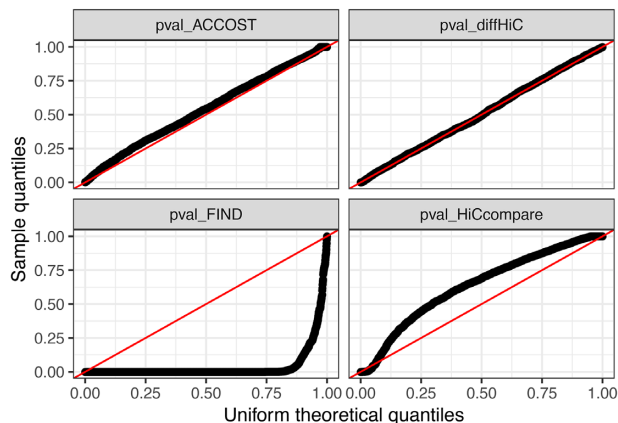
Figure 4. Uniformity of *P*-values generated by different methods. Each panel is a quantile–quantile (Q–Q) plot of the empirical (y-axis) versus uniform (x-axis) distribution of *P*-values generated by ACCOST, diffHiC, FIND and HiCcompare on simulated null data. The *P*-values for FIND and HiCcompare deviate strongly from the expected uniform distribution.

idea is to re-purpose DEseq's 'size factors,' which were originally developed to account for differences in read depth between samples, to instead model the genomic distance effect and the loci-specific biases. The extension is non-trivial because a separate set of factors must be created for each genomic distance (see Materials and Methods). To estimate the variance in contact counts in cases where the number of biological replicates is small, we assume that pairs of loci at the same genomic distance act similarly, and calculate, for each genomic distance, sample means and variances for all pairs of loci across all replicates. We then estimate a smooth function of the variance as a function of the mean (Methods) to come to our final estimate of the variance. This approach allows ACCOST to calculate a *P*-value relative to the null hypothesis that the true number of interactions between loci i and j is equal in the two conditions being tested.

Visualization of log-fold change values between two different cell types, mouse ESC and cortex, illustrates the effect of ACCOST normalization. A raw log fold-change comparison (Figure 2A) highlights differences primarily due to the difference in sequencing depth of the two libraries. Correcting for this effect by simple linear rescaling leaves a clear pattern associated with the genomic distance effect, with large fold-changes farther from the diagonal (Figure 2B). This pattern is not eliminated by ICE normalization (Figure 2C). Because the ICE procedure is based on multiplying each pairwise contact count by two locus-specific bias terms, and because these bias terms are similar in the two experiments, the log fold change values are quite similar before and after ICE normalization. In contrast, the pattern associated with the genomic distance effect is successfully removed by ACCOST's normalization (Figure 2D–E).

To quantitatively confirm that ACCOST correctly accounts for differences in the genomic distance effect, we simulated two Hi-C matrices with different distance decay curves. We generated counts from Poisson distributions with means that decay as illustrated in Figure 3. Note that we did not employ a negative binomial distribution because we did not want the simulations to favor ACCOST. The significance of differences between simulated datasets was then assessed using ACCOST and three existing methods (diffHiC, FIND and HiCcompare). Because the only differences in the simulated data sets arise due to changes in the genomic distance effect, all methods should produce uniform (null) *P*-values. Empirically, we observe nearly uniform *P*-values for both ACCOST and diffHiC (Figure 4, top row). In contrast, both FIND and HiCcompare are misled by the genomic distance effect, with FIND assigning many very small *P*-values, and HiCCompare deviating both above and below the expected uniform distribution (Figure 4, bottom row).

We further compared ACCOST to diffHic by assessing the significance of changes in contact counts between GM12878 and IMR90 cells. These cells derive from two markedly different cell types (B-lymphocyte and fibroblast, respectively) and have been observed to have different short- and long-range contact patterns (23), consistent with differing genome organization at multiple scales. At short range (<100 Mb, Figures 5A and C), differential contact *P*-values are evenly distributed between positive and negative log fold-changes. However, at long range (>100 Mb, Figures 5B and D) diffHic *P*-values are skewed consistently toward more frequent contacts in IMR90 (negative log fold-change). Further investigation of this phenomenon shows a striking increase in the number of significant contacts called by diffHiC at genomic distance of 50 Mb (Figures 5E), which corresponds to the threshold between the two LOESS normalizations performed by diffHiC.

Analysis of *Plasmodium* Hi-C data

As an illustration of how ACCOST might be applied in practice, we ran the software on Hi-C data from two stages in the *P. falciparum* life cycle: trophozoite, the most transcriptionally active stage in the erythrocytic cell cycle of the parasite, and sporozoite, the stage responsible for the transmission of the disease from mosquito to human. Malaria
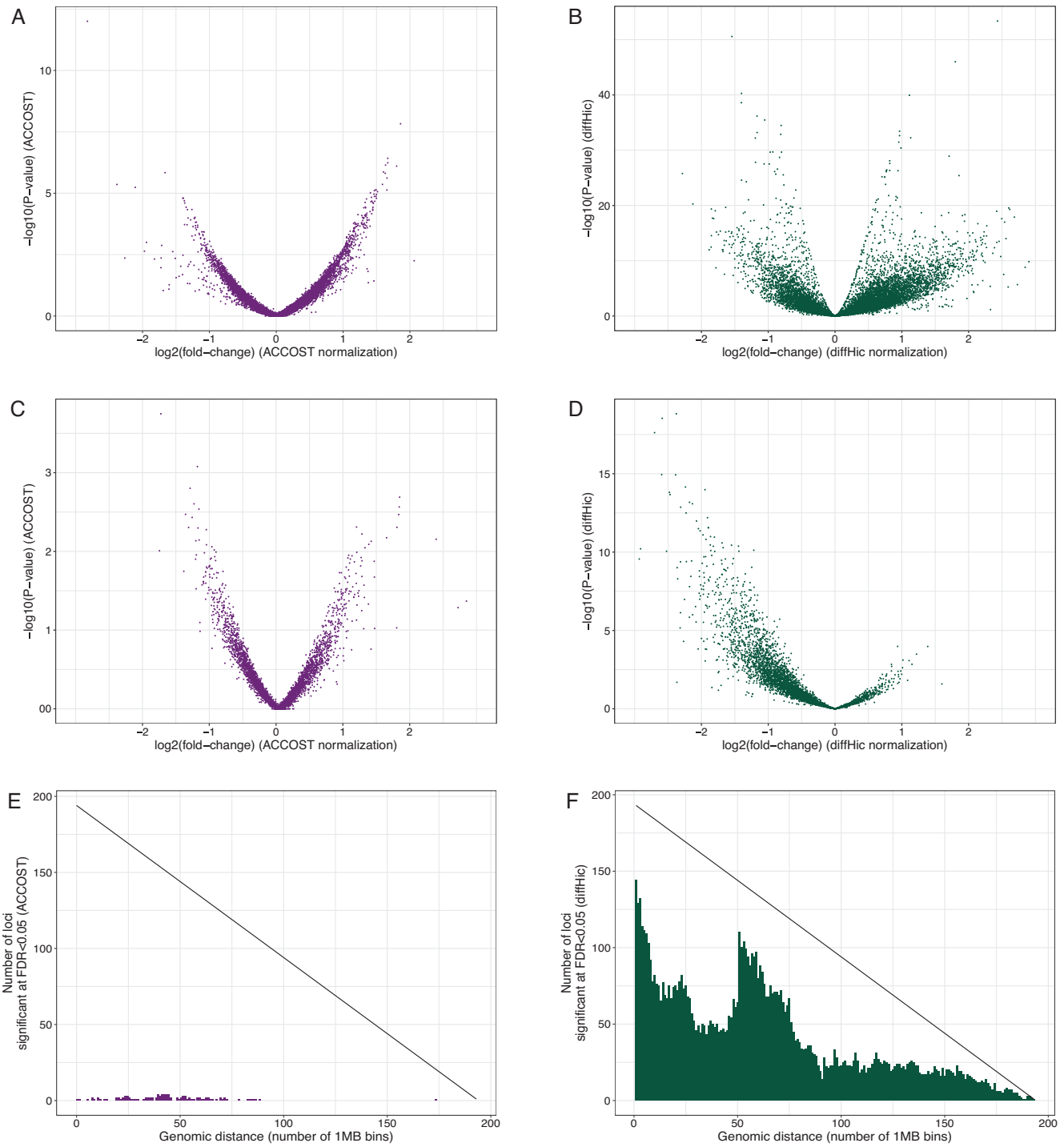
Figure 5. ACCOST *P*-values are not skewed at long distances. At short range (<100 Mb), ACCOST and diffHic yield symmetric distributions of *P*-values (**A** and **B**). At long range (>100 Mb), diffHic fold-changes are skewed in one direction, while ACCOST fold-changes are more evenly distributed (**D** and **C**). Histograms of bin pairs for which each method indicates a significant change in contact counts with FDR <0.05. The black line indicates the total number of bin pairs measured (**E** and **F**).
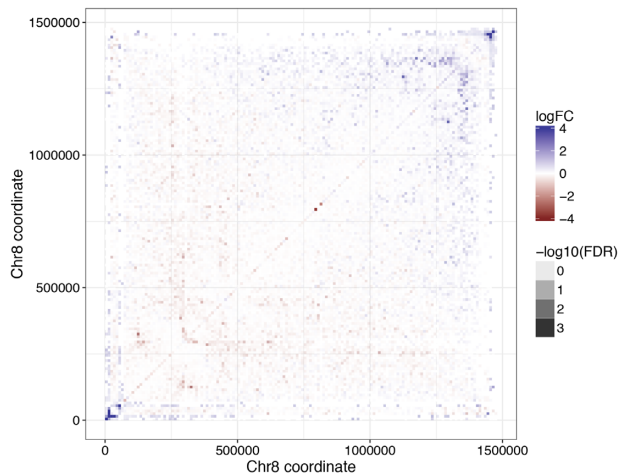
Figure 6. Comparison of life cycle stages in *Plasmodium falciparum*. AC-COST reveals several significant changes in the Hi-C interaction counts between two distinct stages in the life cycle of the parasite highlighting the role of genes expressed in each stage.

transmission specifically occurs when sporozoites are injected from the mosquito's salivary gland into the human skin during blood feeding. ACCOST highlights several changes in interactions between these two stages, one of them particularly interesting in the right arm of chromosome 8 (Figure 6). In this chromosome, two clusters of genes interact significantly more at the sporozoite stage. One of these cluster involves genes specifically expressed at the sporozoite stage, including the sporozoite invasion-associated protein 2 and the sporozoite and liver stage tryptophan-rich protein. The generation of a loop at position ~130k may be critical to enhancing gene expression of these genes at this particular stage. The other cluster of genes occurring at positions between 133k and 145k encode proteins that are exported to the surface of the infected red blood cells and implicated in mechanisms driving parasite immune evasion. These genes are critical to the parasite survival during the erythrocytic cycle and are expressed at a high level at the trophozoite stage. However, because these genes are not required in the mosquito stages, they are repressed in the sporozoite stage; accordingly, we observe them interacting significantly with the parasite heterochromatin cluster near the telomere ends at this stage of the parasite life cycle.

DISCUSSION

We have demonstrated that ACCOST provides unbiased confidence estimates for differential Hi-C contacts by correctly controlling for the genomic distance effect caused by the polymer behavior of DNA. ACCOST thus enables biologists to better interpret Hi-C data by focusing on the subset of differential contacts that exhibit the most statistically significant differences. As is apparent from Figure 6, many of the observed differences exhibit complex structures related to known features of Hi-C data, such as loops, stripes, domains, and compartments. A clear direction for future work is to use ACCOST as a building block in a scheme to identify large, potentially irregularly shaped regions of dif-

ferential contact. A second useful direction for future work would be extensions to handle differentiation or cell cycle time series data.

In the experiment reported in Section 3.1, the simulated counts follow a Poisson distribution, while much of AC-COST complexity is due to the fact that it assumes a negative binomial distribution. This mismatch is intentional, since we did not want our simulation setup to favor AC-COST, but the mismatch may also explain why the AC-COST P-values in Figure 4 are not perfectly uniform.

The failure of FIND to account for the genomic distance effect is not surprising, since the method makes no attempt to model this effect. However, one surprising outcome of our simulation analysis is the conclusion that HiC-compare fails to account for the genomic distance effect, whereas diffHiC does a good job of handling the effect. We note that diffHic models the mean-variance relationship and then normalizes what it calls 'trended biases' between libraries. Although the method never explicitly models the genomic distance, this trended bias seems empirically to capture the genomic distance effect. On the other hand, we do not have a good hypothesis for why HiCcompare does not work well. Similarly, the tendency of diffHiC to produce a greater number of significant P-values in one direction relative to the other remains unexplained (Figure 5).

As currently implemented, ACCOST uses one genomic distance bin per Hi-C bin. However, this is done merely for convenience. In principle, one could decouple these two discretization schemes. This would allow ACCOST, for example, to increase the size of genomic distance bins at long distances.

REFERENCES

1. Fullwood,M.J., Liu,M.H., Pan,Y.F., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H., Chew,E.G. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
2. 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. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
3. Mumbach,M.R., Rubin,A.J., Flynn,R.A., Dai,C., Khavari,P.A., Greenleaf,W.J. and Chang,H.Y. (2016) HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919–922.
4. Fang,R., Yu,M., Li,G., Chee,S., Liu,T., Schmitt,A.D. and Ren,B. (2016) Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.*, **26**, 1345–1348.
5. Quinodoz,S.A., Ollikainen,N., Tabak,B., Palla,A., Schmidt,J.M., Detmar,E., Lai,M.M., Shishkin,A.A., Bhat,P., Takei,Y. *et al.* (2018) Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, **174**, 744–757.
6. Beagrie,R.A., Scialdone,A., Schueler,M., Kraemer,D.C., Chotalia,M., Xie,S.Q., Barbieri,M., de Santiago,I., Lavitas,L.M., Branco,M.R. *et al.* (2017) Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, **543**, 519–524.

7. Krumm,T. and Duan,Z. (2019) Understanding the 3D genome: emerging impacts on human disease. *Semin. Cell Dev. Biol.*, **90**, 62–77.

8. Dixon,J.R., Jung,I., Selvaraj,S., Shen,Y., Antosiewicz-Bourget,J.E., Lee,A.Y., Ye,Z., Kim,A., Rajagopal,N., Xie,W. *et al.* (2015) Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331.

9. Wang,J., Lan,X., Hsu,P.Y., Hsu,H.K., Huang,K., Parvin,J., Huang,T.H. and Jin,V.X. (2013) Genome-wide analysis uncovers high frequency, strong differential chromosomal interactions and their associated epigenetic patterns in E2-mediated gene regulation. *BMC Genomics*, **14**, 70.

10. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

11. Shaviti,Y. and Lio,P. (2014) Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.*, **10**, 1576–1585.

12. Fryzlewicz,P. and Nason,G.P. (2004) A Haar-Fisz algorithm for Poisson intensity estimation. *J. Comput. Graph. Stat.*, **13**, 621–638.

13. Paulsen,J., Sandve,G.K., Gundersen,S., Lien,T.G., Trengereid,K. and Hovig,E. (2014) HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. *Bioinformatics*, **30**, 1620–1622.

14. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

15. Lun,A.T.L. and Smyth,G.K. (2015) diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, **16**, 258.

16. Djekidel,M.N., Chen,Y. and Zhang,M.Q. (2018) FIND: difFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Res.*, **28**, 412.

17. Stansfield,J.C., Cresswell,K.G., Vladimirov,V.I. and Dozmorov,M.G. (2018) HiCcompare: an R-package for joint normalization and comparison of Hi-C datasets. *BMC Bioinformatics*, **19**, 279.

18. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

19. Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.

20. Wang,S., Su,J.H., Beliveau,B.J., Bintu,B., Moffitt,J.R., Wu,C.T. and Zhuang,X. (2016) Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, **353**, 598–602.

21. Naumova,N., Imakaev,M., Fudenberg,G., Zhan,Y., Lajoie,B.R., Mirny,L.A. and Dekker,J. (2013) Organization of the mitotic chromosome. *Science*, **342**, 948–953.

22. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

23. Rao,S.S.P., Huntley,M.H., Durand,N., Neva,C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **59**, 1665–1680.

24. Durand,N.C., Shamim,M.S., Machol,I., Rao,S.S., Huntley,M.H., Lander,E.S. and Aiden,E. L. (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.

25. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

26. Shen,Y., Yue,F., McCleary,D.F., Ye,Z., Edsall,L., Kuan,S., Wagner,U., Dixon,J., Lee,L., Lobanenkov,V.V. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

27. Ay,F., Bunnik,E.M., Varoquaux,N., Bol,S.M., Prudhomme,J., Vert,J.-P., Noble,W.S. and Le Roch,K.G. (2014) Three-dimensional modeling of the *P.falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Res.*, **24**, 974–988.