HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2021 May 20.

Published in final edited form as:

Nat Methods. 2020 January ; 17(1): 37–40. doi:10.1038/s41592-019-0624-3.

FreeHi-C simulates high fidelity Hi-C data for benchmarking and data augmentation

Ye Zheng[1], **Sündüz Keleş**[1,2,3]

[1]Department of Statistics, University of Wisconsin - Madison, Madison, WI 53706, USA

[2]Department of Biostatistics and Medical Informatics, University of Wisconsin - Madison, Madison, WI 53706, USA

Abstract

Ability to simulate high-throughput chromatin conformation (Hi-C) data is foundational for benchmarking Hi-C data analysis methods. Here we present a non-parametric strategy named FreeHi-C to simulate Hi-C data from the interacting genome fragments. Data from FreeHi-C exhibit high fidelity to biological Hi-C data. FreeHi-C boosts the precision and power of differential chromatin interaction detection through data augmentation under preserved false discovery rate control.

Recent maturation of chromosome conformation capture (3C)[1] and Hi-C sequencing technologies[2, 3] revealed transformative insights on long-range gene regulation[4]. A growing number of methods[5–12] emerged for the analysis of Hi-C and other 3C-derived data. Benchmarking and evaluation of Hi-C methods have relied on either conducting distance-stratified permutations of the contact matrices[5] or directly simulating them with the most general spatial structures[6–11]. Additionally, pooling and random partitioning of replicates to generate pseudo-replicates and downsampling are prevalent approaches for studying similarity metrics and sequencing depth effects[5, 8, 11, 13]. A systematic Hi-C simulation method, Sim3C[14], was proposed for the design of Hi-C experiments with respect to the power analysis and the selection of restriction enzyme and sequencing depth. However, the strong focus of Sim3C on microbial genomics and metagenomics negates its utility for common Hi-C experiments[15] (Supplementary Fig. 1, Supplementary Note).

Here we present FreeHi-C (**Fr**agment interactions **e**mpirical **e**stimation for fast simulation of **Hi-C** data), as a robust method that nonparametrically simulates realistic read-level Hi-C data by emulating the standard Hi-C experimental protocol[2, 3] (Fig. 1a). FreeHi-C takes as input Hi-C sequencing data and estimates the frequency of genomic fragment interactions. This is fundamentally different from existing methods that simulate Hi-C contact matrices under a series of assumptions[5–7, 10]. Subsequently, FreeHi-C generates pairs of sequencing

[3]Corresponding author: keles@stat.wisc.edu.

Author Contributions S.K. and Y.Z. conceived the project. Y.Z. and S.K. designed the research and developed the method. Y.Z. developed the simulation framework and performed the experiments. Both authors contributed to the preparation of the manuscript.

Competing Financial Interests The authors declare no competing financial interests.

reads from the interacting fragment pairs with embedded random nucleotide mutations and indels while conserving the proportion of chimeric reads. Thereby, the variability and read-level characteristics of original Hi-C sequencing libraries can be preserved in the simulated sequences.

We illustrated the versatile features of FreeHi-C with Hi-C datasets of two human cell lines, GM12878 and A549 (4 independent cell culture replicates each), and malaria parasite *Plasmodium falciparum* 3D7, as representatives of large and small genomes, respectively. Analysis of GM12878 and A549 are carried out at 40kb resolution and 10kb for *Plasmodium falciparum* 3D7 unless otherwise stated. Replicates are simulated at the same sequencing depth as the seed replicates unless explicitly stated (Supplementary Note, Supplementary Figs. SN1–5). Compared to Sim3C realizations, FreeHi-C simulations capture the detailed chromatin interaction structures, such as chromatin loops and TADs, better; hence contact matrices from FreeHi-C exhibit markedly higher fidelity to the seed biological data (Supplementary Fig. 1). Specifically, comparison of the distribution of contact counts stratified by the genomic distance between the simulated and biological data shows that FreeHi-C yielded contact count distributions are similar to those of the seed biological data (Supplementary Fig. 2) in contrast to Sim3C (Supplementary Note). Similar high fidelity observations hold for paired comparisons of simulated replicates (Supplementary Note). In addition, we show that the simulated data of different replicates preserved the relations existing between replicates of *P. falciparium* (Fig. 1b) and conveyed that chromatin architecture in the ring stage is more similar to that of the schizont stage while the trophozoite stage has striking differences from the other two[16] (Supplementary Fig. 3a). In addition to recapitulating the clustering structure of the biological samples, FreeHi-C simulated replicates also preserve genomic domain structures, such as A/B compartment and TADs. A/B compartments correlation is higher between the seed biological replicate and the FreeHi-C simulated replicates than those among biological samples (Fig. 1c and Supplementary Note). Similarly, the concordance of the TAD structures are markedly higher for FreeHi-C simulated replicates compared to other biological samples (Fig. 1d, Supplementary Note).

Next, we demonstrate how FreeHi-C enables benchmarking a wide range of Hi-C analysis methods and compare it with the downsampling strategy. We first focus on the assessment of the reproducibility of Hi-C contact matrices by HiCRep[13]. HiCRep reproducibility quantification successfully clustered the ring and schizont stages together (Fig.1b), this result can be challenged as being biased due to the significant differences in the sequencing depths of the samples, i.e., the ring and schizont stages have 45–375% more reads (Supplementary Note). Therefore, it is desirable to adjust for the sequencing depth either by simulation or downsampling. By down simulating the ring and schizont stages to the sequencing depth of trophozoite sample, or up simulating the trophozoite and ring stages to the schizont sample, HiCRep reproducibility consistently clusters the ring and schizont stages together (Fig. 1e, Supplementary Fig. 3). Downsampling, however, leads to the schizont stage being misclustered with the trophozoite (Fig. 1f) indicating that downsampling may generate low-quality Hi-C matrices (i.e., reproducibility ranges from 0.32 to 0.91) that potentially lead to biased inference. We further illustrate how FreeHi-C

enables sequencing depth debiasing for HiCRep[13] in evaluating reproducibility and Fit-Hi-C12 in detecting significant interactions in Supplementary Note.

Another pivotal benchmarking utility of FreeHi-C is for comparing the performance of different methods that address the same Hi-C inference problem. We illustrate this in the context of detection of differential chromatin interactions (DCIs) by comparing diffHic[6], multiHiCcompare[17], and Selfish[11] across a series of sequencing depths for false discovery rate (FDR) control and power through both FreeHi-C simulation and downsampling (Supplementary Note). Both diffHic and multiHiCcompare can control the FDR with either the GM12878 or A549 datasets and that diffHic is generally more conservative (Supplementary Fig. 4 and 5). Remarkably, downsampling displays an increasing FDR trend as the sequencing depth increases. However, this trend can only be investigated up until the original depth of the samples since the sequencing depths of downsampled samples cannot exceed their original depths (Fig. 1g). FreeHi-C simulation elucidates the trends when the depths of the samples go beyond their original ones and reveals a conservative FDR control at higher depths. Additionally, downsampling and FreeHi-C simulation have distinct implications for detection power of the methods (Supplementary Note).

A key impediment for differential chromatin interaction inference with Hi-C data is the limited number of biological replicates. We identified three commonly encountered scenarios as a function of the number of replicates available per condition (Fig. 2, Supplementary Fig. 6): one replicate per condition (ORPC), uneven numbers of replicates per condition where one of the conditions have only a single replicate (URPC), and multiple numbers of replicates per condition (MRPC). Under these settings, we applied multiHiCcompare for DCIs detection. In the ORPC setting, the FDR control with the BH procedure[18] ensures that the observed FDR is below the target FDR (Fig. 2a and Supplementary Fig. 7); however, this setting exhibits extremely conservative FDR control. This comes at the cost of low power (Supplementary Fig. 8a). When we augment each of the conditions with FreeHi-C simulated replicates, FDR is still well controlled (Fig. 2a and Supplementary Fig. 7), and power increases by an average of 300 fold across the five levels of FDR thresholds (Supplementary Fig. 8b). It is reasonable to argue that, under ORPC, the key target should be the ranking of the interactions, rather than a thresholded list of DCIs, because the top significant DCIs are typically utilized for downstream analysis. We scrutinized the accuracy of the top significant DCIs identified with and without FreeHi-C augmentation by comparing the ranked lists to the gold standard set of DCIs detected by utilizing all the 4 replicates per condition. FreeHi-C simulated samples yield a significantly higher precision of 100–75% compared to ~10% in the ORPC setting (Fig. 2b, see also Supplementary Note for evaluations under additional gold standard settings). We next assessed the biological relevance of the “ranked up” and “ranked down” DCIs due to FreeHi-C augmentation by external RNA-seq and CTCF ChIP-seq data. Both comparisons supported the new ranking of the top DCIs (Supplementary Figs. 8c–d, Supplementary Note). The URPC setting conveyed similar results (Supplementary Fig. 6).

Finally, we generalized the FreeHi-C augmentation strategy as a meta-analysis approach for multiple replicates per condition (Figs. 2c–d, Methods). FDR control is well preserved as the number of simulated replicates in augmentation increases (Fig. 2c and Supplementary Fig.

9a). Notably, augmentation with simulated replicates not only boosts the number of significant DCIs identified (Supplementary Figs. 9b–c), but also yields a significantly better ranking with higher precision of the top significant DCIs for further quantitative and experimental validation (Fig. 2d, Supplementary Figs. 10–11, Supplementary Note). We successfully validated the biological relevance of the DCIs identified with meta-analysis via FreeHi-C augmentation by leveraging RNA-seq and CTCF ChIP-seq data as in the ORPC setting (Supplementary Figs. 12–14).

Analytical methods for analyzing data from Hi-C and related experiments are growing at a fast pace without uniform benchmarking and evaluation. Since FreeHi-C enables simulating read-level Hi-C data in a data-driven manner, it relies on seed biological data. However, FreeHi-C framework can conceptually accommodate additional user-induced features such as spike-ins while adhering to seed biological data quality and signal-to-noise ratio.

Methods

FreeHi-C simulation framework

Processing and training module—FreeHi-C implements the steps outlined in Fig. 1b. It takes as input raw Hi-C sequencing data in the form of FASTQ files, processes the reads, and learns the parameters for the simulation module. The sequence processing module follows a standard protocol[21, 22] by aligning raw paired-end read files individually and then joining the read ends to form read pairs, followed by interaction validation checking and duplicate removal steps. After obtaining the valid read-pairs (i.e., interactions), it fits an interaction-level mixture model to estimate the genomic fragment interaction frequencies, with the genomic fragments defined by the experimental restriction enzyme cutting sites. This sampling model considers two multinomial distributions with parameters $\overrightarrow{\pi_0} = \left\{\pi_{ij}^0\right\}$ and $\overrightarrow{\pi_1} = \left\{\pi_{ij}^1\right\}$, $i, j = 1, \ldots, L$, where L denotes total number of genomic fragments, for generating interaction events. The distribution indexed by the parameter $\overrightarrow{\pi_0}$ reflects background interactions, driven by experimental artifacts such as genomic distance whereas the second distribution, $\overrightarrow{\pi_1}$, characterizes true biological interaction signals. For each genome fragment “interaction” event, one interaction is drawn from each of the multinomial distributions, and one of them is recorded as the actual interaction event with probability α. Consequently, the distribution specified by this interaction-level sampling model is multinomial with vector-valued parameter $\alpha\overrightarrow{\pi_0} + \left(1 - \alpha\right)\overrightarrow{\pi_1}$. This intuitive sampling model, estimation of which does not require deconvolution of the two distributions, is key for FreeHi-C’s successful capture of the interactions in a given biological sample. Once the genomic fragment interaction is sampled, FreeHi-C generates a read pair from this interaction by taking into account strand configuration of the ends of the read pair, mismatch(es), insertion(s), deletion(s), the proportion of chimeric reads, and base quality scores of the reads. Hence, as part of the training module, FreeHi-C estimates this set of parameters empirically from the collection of valid reads. Specifically, FreeHi-C estimates the frequency distributions of numbers of insertions, deletions, and different types of mismatches across all the valid read pairs. Additionally, FreeHi-C processing and training module records the proportion of chimeric reads, i.e., reads pairs where one or both of the

read ends are sequenced over the ligation junction, that are rescuable with the aim of preserving this proportion for the simulated reads. Finally, FreeHi-C empirically estimates the distribution of base quality scores for each locus of the Hi-C reads and uses these estimates to ensure that the simulated reads have similar base quality scores as the seed biological replicate.

Simulation module—FreeHi-C simulates fragment pairs from the estimated interaction-level mixture model as genomic fragments that form crosslinks in the Hi-C experiment protocol[3]. The ligation procedure in the experiment leads to two fragment junction sites. FreeHi-C randomly generates one of these which is then passed onto the next step to emulate DNA shearing. Two DNA shearing loci are randomly selected within ± 500bp, by default, of the selected ligation site. These two shearing loci also work as the starting points of the sequencing procedure. FreeHi-C extracts the sequences of the given length, for example, 36bp for 36bp paired-end sequencing, from these loci and assigns strand direction accordingly. During this sequencing step, reads closer than the requested read length to the ligation sites can be generated, as an emulation of chimeric reads. The final step is to introduce noise to the read sequences so that the mismatches and indels match to those in the reads of the original biological sample. Utilizing the empirical distribution of the sequence base quality scores across individual locus, FreeHi-C simulates such scores for each read at the nucleotide level. A key strength of FreeHi-C is that it can generate as many reads as specified by the user and outputs these in the FASTQ format. Furthermore, it processes the resulting reads according to the standard analysis protocol of Hi-C reads by the processing module. Through the post-simulation processing, FreeHi-C can directly provide genomics contact counts in a sparse matrix format (BED) compatible with the standard input format of downstream Hi-C analysis. Processing of the raw reads and learning of the parameters can be implemented on individual read pairs followed by a final aggregation; hence this module can be efficiently parallelized. Furthermore, simulations based on the same parameter settings are parallelized at the read-pair generation level.

Data augmentation with FreeHi-C simulated samples

To account for the fact that simulated samples cannot provide additional full degrees of freedom when testing for differential chromatin interactions with two or more replicates per condition, we employed a FreeHi-C simulation augmented meta-analysis strategy. The meta-analysis approach pairs biological and simulated samples in numbers concordant with the original differential testing design and aggregates p-values of candidate differential interactions across comparisons by Fisher's method[23]. More specifically, simulation replicates for each of the original n biological replicates are generated per condition and considers 2^n-1 additional tests to preserve the degrees of freedom of the original test statistic. For example, for a setting with 2 biological replicates per condition, we generated 4 FreeHi-C simulation samples, one per original biological replicate, and evaluated the following comparisons, where c_1 and c_2 refer to the two testing conditions under consideration.

Test 1: ($\mathrm{Rep1}_{c1,\mathrm{bioSample}}$, $\mathrm{Rep2}_{c1,bioSample}$) vs. ($\mathrm{Rep1}_{c2,\mathrm{bioSample}}$, $\mathrm{Rep2}_{c2,bioSample}$)

Test 2: ($\text{Rep1}_{c1,simulation}$, $\text{Rep2}_{c1,bioSample}$) vs. ($\text{Rep1}_{c2,simulation}$, $\text{Rep2}_{c2,bioSample}$)

Test 3: ($\text{Rep1}_{c1,\text{bioSample}}$, $\text{Rep2}_{\text{c1},simulation}$) vs. ($\text{Rep1}_{c2,bioSample}$, $\text{Rep2}_{c2,simulation}$)

Test 4: ($\text{Rep1}_{c1,simulation}$, $\text{Rep2}_{c1,simulation}$) vs. ($\text{Rep1}_{c2,simulation}$, $\text{Rep2}_{c2,simulation}$)

The p-values of these tests are then aggregated by Fisher's method[23] as $-2\sum_{i=1}^{M}\log(p_i) \sim \chi^2_{2M}$, where $M = 4$ in the above example and the degree of freedom for the χ^2 distribution is 8. However, the above 4 tests are not independent, the resulting aggregated p-values are anti-conservative. To dampen this effect, instead of ranking the differential interactions with the Fisher's p-value in the BH procedure[18], we rank them based on median of their adjusted p-values from individual test. Specifically, instead of ordering the hypothesis sequence $H_{(i)}$, $i = 1, 2, \ldots, M$, by the aggregated p-values obtained from Fisher's method, we order them by the new significance rank determined by the median adjusted p-values of across all four tests for each individual contact and denote such ordering as $H_{(r(i))}$, $i = 1, 2, \ldots, M$, where $r(i)$ is the index of the contact that is ranked i^{th} in the new significant ranking list. Let k be the largest i for which $p_{(r(i))} \leq \frac{i}{M}\alpha$, then the BH procedure rejects all $H_{(r(i))}$, $i = 1, 2, \ldots, k$. This is a more conservative procedure than the ordinary BH procedure on the Fisher's p-values because the number of rejections, k, is always smaller or equal to the number of rejections using the BH procedure with Fisher's p-values. The computational experiments support that this reverses the potential anti-conservative effect of aggregating dependent p-values with Fisher's method.

Evaluating the ranking of detected DCIs

The gold standard differential interaction set is approximated by the most significant DCIs detected by multiHiCcompare based on a quasi-likelihood negative binomial generalized log-linear model to test the coefficients with BH procedure adjustment for multiple comparisons (one-sided test; FDR ≤ 0.001, 0.005, 0.01, 0.05, respectively) using 4 biological replicates of GM12878 versus 4 biological replicates of A549. In the sets presented in this paper, we rank the DCIs by their significance order and quantify the fraction of the top N significant differential interactions that appear in the gold standard set where N varies as 500, 1000, etc. This quantity refers to recovery rate or precision. A more conservative gold standard set of DCIs is defined as the intersection of the significant DCIs detected by multiHiCcompare (FDR ≤ 0.01) and those identified by diffHic (FDR ≤ 0.1).

Another type of gold standard DCI list is specially defined for tests of uneven number of replicates per condition (Supplementary Figs. 24–25). In this setting, the gold standard DCI set is defined as the set of most significant interactions in the comparison of rep2 and rep4 of GM12878 with rep1 and rep4 of A549. Accordingly, we measure the precision of the results from the following four comparisons: (i) one replicate out of rep2 and rep4 of GM12878 versus one out of rep1 and rep4 of A459; (ii) one replicate out of rep2 and rep4 of GM12878 with its FreeHi-C augmentation versus one out of rep1 and rep4 of A459 with its FreeHi-C augmentation; (iii) rep2 and rep4 of GM12878 versus one out of rep1 and rep4 of A459 with its FreeHi-C augmentation; (iv) one out of rep2 and rep4 of GM12878 with its FreeHi-C augmentation versus rep1 and rep4 of A459.

Evaluating DCIs detected by FreeHi-C augmentation with RNA-seq and CTCF ChIP-seq

We evaluated the significance of the observed proportion of differentially expressed (DE) genes between GM12878 and A549 that overlap with significant DCIs using a randomization test. DE genes are detected by DESeq2[24] based on a negative binomial model. An empirical null distribution for the observed overlap statistics is constructed by randomly selecting an equal number of interactions as significant ones from all valid bin-pairs and overlapping these with the DE genes for 10,000 times. The significance level of observed overlap is quantified by the percentage of random selection results that is larger than or equal to the observed statistics. A similar strategy is implemented for evaluating co-localization of DCIs with differential CTCF ChIP-seq peaks. Differential CTCF ChIP-seq peaks are defined by peaks that are uniquely enriched in only one cell line.

Reporting Summary

Further information on research design is available in the Life Sciences Reporting Summary linked to this article.

Data availability

To study the operating features of FreeHi-C, we utilized two publicly available human Hi-C datasets as examples of large genomes with independent experiments using four cell cultures, which are referred to as four biological replicates, from GM12878[3] cell line and another four from A549[25]. Raw FASTQ files for GM12878 were downloaded from GEO[26] (https://www.ncbi.nlm.nih.gov/geo) under the accession code GSE63525 and raw sequences for A549 were obtained from the ENCODE portal[27] (https://www.encodeproject.org) with accession code ENCSR662QKG. For evaluation of FreeHi-C performance on small genomes, we leveraged three different stages of malaria parasite *Plasmodium falciparum* red blood cell cycles[16]. Raw sequences for *P. falciparum* are downloaded from GEO[26] (https://www.ncbi.nlm.nih.gov/geo) under the accession code GSE50199. GM12878 and A549 are both processed at 40kb resolution, and *P. falciparum* at 10kb.

For validating the differential interaction detection with a differential expression analysis, we utilized RNA-seq gene expression data from the ENCODE portal (accession ENCSR000AED for GM12878 and ENCSR000CTM for A549). Similarly, the CTCF ChIP-seq peak signal files were also downloaded from ENCODE under accession ENCSR000DZN for GM12878 and ENCSR000DPF for A549. The data used in this paper are summarized in the Supplementary Table 1.

All the simulated data used in the analysis and Juicebox[28] visualization data are available at Zenodo[29] (http://doi.org/10.5281/zenodo.3345896).

Code availability

FreeHi-C pipeline is implemented in Python with C accelerated core calculations and it naturally fits in the high-performance computing environments for parallelization. The source codes and instructions for running FreeHi-C are publicly available at https://github.com/keleslab/FreeHiC.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants HG009744 and HG007019 to S.K.. We thank the peer reviewers and the reviewing editors of this work for their insightful comments.

Reference

1. Dekker J, Rippe K, Dekker M, Kleckner N: Capturing chromosome conformation. Science 295(5558) (2002) 1306–11 [PubMed: 11847345]
2. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al.: Comprehensive map- ping of long-range interactions reveals folding principles of the human genome. Science 326(5950) (2009) 289–293 [PubMed: 19815776]
3. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7) (2014) 1665–1680 [PubMed: 25497547]
4. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, Wilson M, Sridharan R: A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Research 44(4) (2016) 1977 [PubMed: 26546512]
5. Yardımcı GG, Ozadam H, Sauria ME, Ursu O, Yan KK, Yang T, Chakraborty A, Kaul A, Lajoie BR, Song F, et al.: Measuring the reproducibility and quality of hi-c data. Genome Biology 20(1) (2019) 57 [PubMed: 30890172]
6. Lun AT, Smyth GK: diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics 16(1) (2015) 258 [PubMed: 26283514]
7. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S: Comparison of computational methods for Hi-C data analysis. Nature Methods 14(7) (2017) 679 [PubMed: 28604721]
8. Ursu O, Boley N, Taranova M, Wang YR, Yardimci GG, Stafford Noble W, Kundaje A: Genomedisco: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. Bioinformatics 34(16) (2018) 2701–2707 [PubMed: 29554289]
9. Djekidel MN, Chen Y, Zhang MQ: Find: differential chromatin interactions detection using a spatial poisson process. Genome Research 28(3) (2018) 412–422
10. Stansfield JC, Cresswell KG, Vladimirov VI, Dozmorov MG: Hiccompare: an R package for joint normalization and comparison of hi-c datasets. BMC Bioinformatics 19(1) (2018) 279 [PubMed: 30064362]
11. Ardakany AR, Ay F, Lonardi S: Selfish: discovery of differential chromatin interactions via a self-similarity measure. Bioinformatics (2019) i145–i153 [PubMed: 31510653]
12. Ay F, Bailey TL, Noble WS: Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research 24(6) (2014) 999–1011 [PubMed: 24501021]
13. Yang T, Zhang F, Yardimci GG, Song F, Hardison RC, Noble WS, Yue F, Li Q: HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Research (2017) gr–220640
14. DeMaere MZ, Darling AE: Sim3c: simulation of hi-c and meta3c proximity ligation sequencing technologies. GigaScience 7(2) (2017) gix103
15. DeMaere MZ, Darling AE: bin3c: exploiting hi-c sequencing data to accurately resolve metagenome-assembled genomes. Genome Biology 20(1) (2019) 46 [PubMed: 30808380]
16. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG: Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene ex- pression. Genome Research 24(6) (2014) 974–988 [PubMed: 24671853]

17. Stansfield JC, Cresswell KG, Dozmorov MG: multiHiCcompare: joint normalization and comparative analysis of complex hi-c experiments. Bioinformatics (2019)
18. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1) (1995) 289–300
19. Zheng X, Zheng Y: Cscoretool: fast hi-c compartment analysis at high resolution. Bioinformatics 34(9) (2017) 1568–1570
20. Giorgetti L, Lajoie BR, Carter AC, Attia M, Zhan Y, Xu J, Chen CJ, Kaplan N, Chang HY, Heard E, et al.: Structural organization of the inactive x chromosome in the mouse. Nature 535(7613) (2016) 575 [PubMed: 27437574]

Reference

21. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology 16(1) (2015) 259 [PubMed: 26619908]
22. Zheng Y, Ay F, Keles S: Generative modeling of multi-mapping reads with mHi-C advances analysis of Hi-C studies. eLife 8 (2019) e38070 [PubMed: 30702424]
23. Fisher RA: Statistical methods for research workers. Genesis Publishing Pvt Ltd (2006)
24. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. Genome biology 15(12) (2014) 550 [PubMed: 25516281]
25. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Galip Gurkan Yardımcı AC, Bann DV, Wang Y, Clark R, Zhang L, Yang H, Liu T, Iyyanki S, An L, Pool C, Sasaki T, Rivera-Mulia JC, Ö zadam H, Lajoie BR, Kaul R, Buckley M, Lee K, Diegel M, Pezic D, Ernst C, Hadjur S, Odom DT, Stamatoyannopoulos JA, Broach JR, Hardison RC, Ay F, Noble WS, Dekker J, Gilbert DM, Yue F: Integrative detection and analysis of structural variation in cancer genomes. Nature Genetics (9 2018) https://www.nature.com/articles/s41588-018-0195-8.
26. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al.: NCBI GEO: archive for functional genomics data sets update. Nucleic Acids Research 41(D1) (2012) D991–D995 [PubMed: 23193258]
27. Consortium EP, et al.: An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414) (2012)
28. Durand Neva C., et al. "Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom." Cell systems 3.1 (2016): 99–101. [PubMed: 27467250]
29. Zheng Ye, & Keles Sunduz. (2019). Supplementary Data for "FreeHi-C: high fidelity Hi-C data simulation for benchmarking and data augmentation" [Data set]. Zenodo. 10.5281/zenodo.3345896

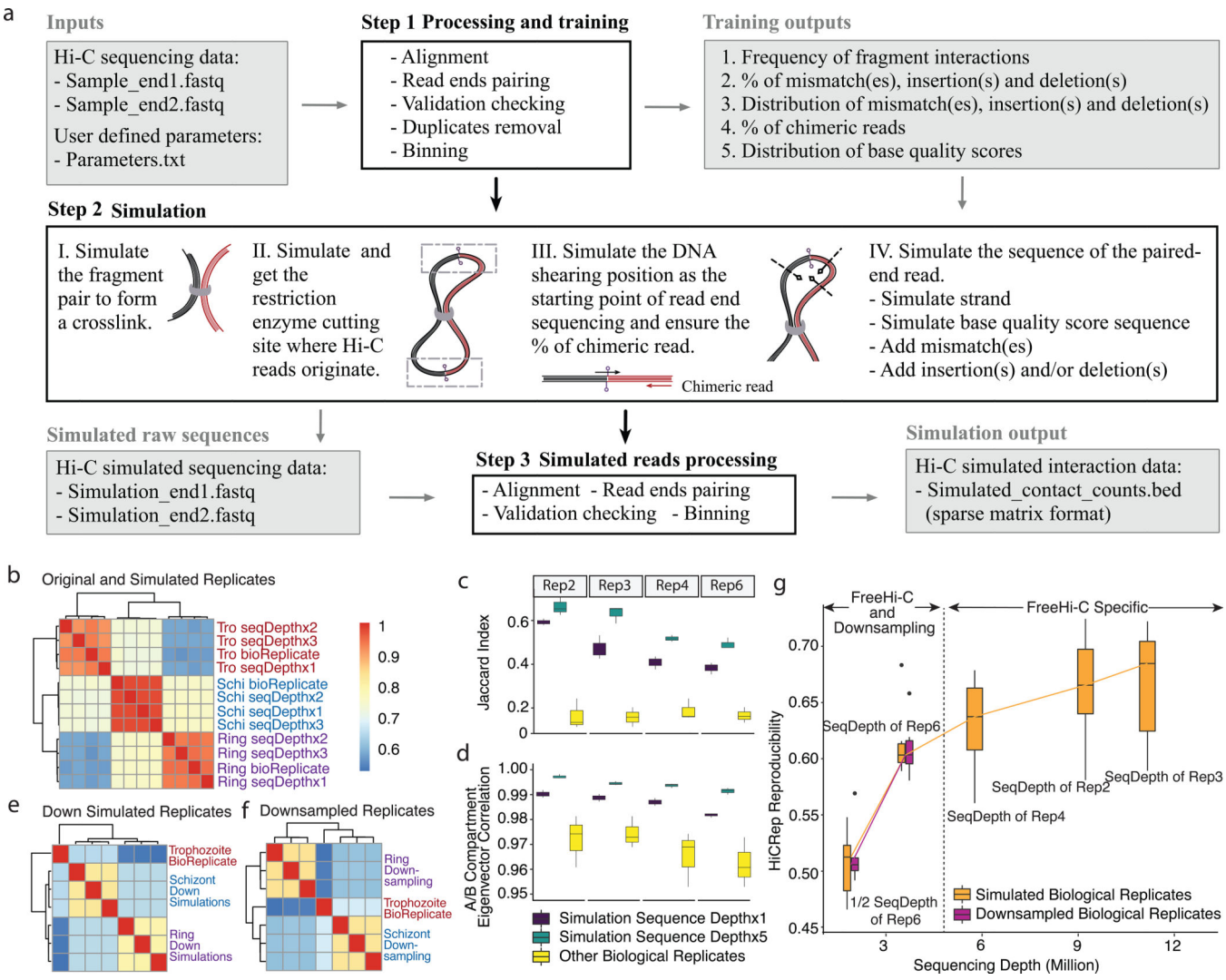

Figure 1. FreeHi-C enables simulating high fidelity Hi-C data.

a. FreeHi-C simulation workflow. Black arrows connect the processing procedures, and grey arrows show the data flow. **b.** Hierarchical clustering of the original Hi-C biological replicates and the FreeHi-C simulated replicates for the ring, trophozoite, schizont stages of *P. falciparum*. Heatmap clustering is obtained with the inherited R function *hcluster* in the *pheatmap* package using the default parameters. Distance is quantified by HiCRep[13]. **c.** Pearson correlation analysis of the A/B compartment eigenvector between the seed biological replicate (delineated at the top of each panel) and FreeHi-C simulation of 1 × sequencing depth, 5 × original sequencing depth, and other biological replicates. A/B compartment eigenvector is calculated by CscoreTool[19] (n = 3). **d.** Jaccard index of the TADs detected using the seed biological replicate (delineated at the top of each panel) and FreeHi-C simulation of the 1 × sequencing depth, 5 × original sequencing depth, and other biological replicates. TAD boundaries are detected using the Insulation Score[20] (n = 3). **e** and f. Hierarchical clustering of the FreeHi-C simulated (**e**) and downsampled (**f**) replicates matching the sequencing depth of the original *P. falciparum* trophozoite stage sample. Distance is calculated by HiCRep[13]. **g.** HiCRep[13] reproducibility of the contact matrices

between pairs of biological replicates of GM12878 simulated by FreeHi-C (orange) or downsampled (purple) to 0.5 × sequencing depth of replicate6, 1 × sequencing depth of replicate6, and sequencing depths of replicate4, replicate2, and then replicate3, respectively (n = 4). In **c**, **d**, and **g**, the center lines indicate medians, box limits indicate the 25th and 75th percentiles. The upper whisker extends from the hinge to the largest value no further than 1.5 × inter-quartile ranges from the hinge. The lower whisker extends from the hinge to the smallest value at most 1.5 × inter-quartile of the hinge. Data beyond the end of the whiskers are outlying points and are plotted individually.

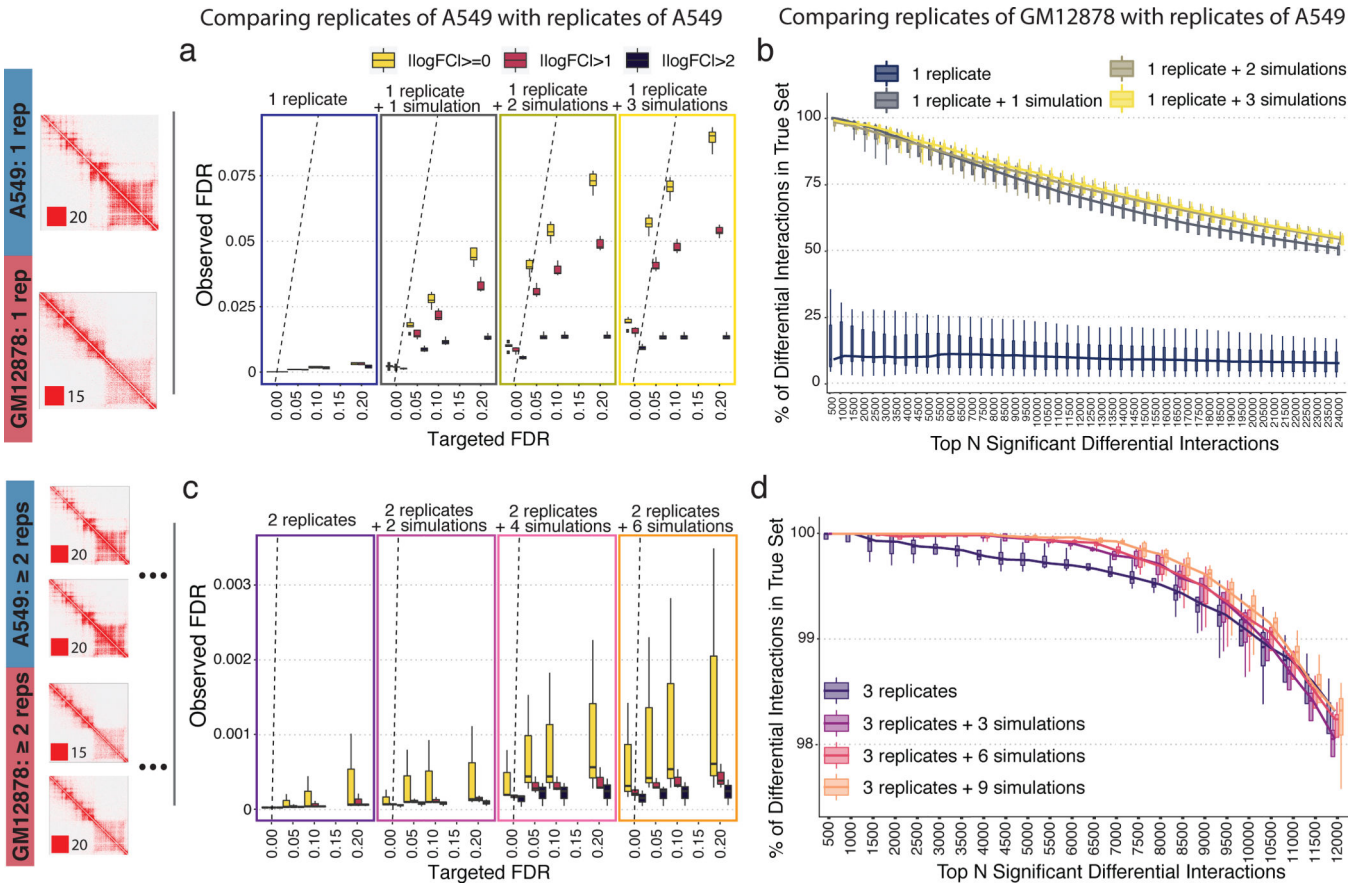

Figure 2. Data augmentation with FreeHi-C simulated replicates improves differential chromatin interactions (DCIs) detection.

a (n = 16) and **b** (n = 16) refer to one replicate per condition (ORPC). **c** (n = 3) and **d** (n = 16) refer to multiple replicates per condition (MRPC) settings. **a** (n = 16) and **c** (n = 3) delineate observed false discovery rates of within-sample comparisons for A549 data (i.e., comparisons of replicate(s) of A549 with other replicate(s) of A549). The dashed lines are y = x. **b** (n = 16) and **d** (n = 16) display precision, computed as the percentage of top significant DCIs of each specific analysis in the gold standard differential chromatin interaction list, as a function of top-ranking DCIs. The gold standard set is defined by comparing the full set of 4 replicates of GM12878 with 4 replicates of A549 filtered by FDR ≤ 0.01. |logFC| refers to the absolute value of natural log transformed fold-change. Differential chromatin interaction detection is performed by HiCcompare[10], by converting the normalized contact counts into Z-scores, and multiHiCcompare[17], using a quasi-likelihood negative binomial generalized log-linear model (one-sided test). The p-values are adjusted by Benjamini-Hochberg procedure[18] for multiple comparisons. For all the boxplots in this figure, the center lines correspond to the medians, box limits correspond to the 25th and 75th percentiles and whiskers comprise all data points within 1.5 × the inter-quartile range.