mHi-C: robust leveraging of multi-mapping reads in Hi-C analysis

Ye Zheng[1], **Ferhat Ay**[2,3] & **Sündüz Keleş**[1,4]

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

[2]La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA

[3]School of Medicine, University of California San Diego, La Jolla, CA 92093, USA

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

Current Hi-C analysis approaches are unable to account for reads that align to multiple locations, and hence underestimate biological signal from repetitive regions of genomes. We developed *mHi-C*, a *m*ulti-read mapping strategy to probabilistically allocate Hi-C multi-reads. mHi-C exhibited superior performance over utilizing only uni-reads and heuristic approaches aimed at rescuing multi-reads on benchmarks. Specifically, mHi-C increased the sequencing depth by an average of 20% leading to higher reproducibility of contact matrices and larger number of significant contacts across biological replicates. mHi-C also revealed biologically supported *bona fide* promoter-enhancer interactions and topologically associating domains involving repetitive genomic regions, thereby unlocking a previously masked portion of the genome for conformation capture studies.

DNA is highly compressed in nucleus and organized into a complex three-dimensional structure. This compressed form brings distal functional elements into close spatial proximity of each other[1,2] and has far-reaching influence on gene regulation. Changes in DNA folding and chromatin structure remodeling may result in cell malfunction with devastating consequences[3-6]. Hi-C technique[7,8] emerged as a high throughput technology for interrogating the three-dimensional configuration of the genome and identifying regions that are in close spatial proximity in a genome-wide fashion. Thus, Hi-C data is powerful for discovering key information on the roles of the chromatin structure in the mechanisms of gene regulation.

There are a growing number of published and well-documented Hi-C analysis tools and pipelines[9-14], and their operating characteristics were recently studied[15-17] in detail. However, a key and common step in these approaches is the exclusive use of uniquely mapping reads. Limiting the usable reads to only uniquely mapping reads underestimates signal originating from repetitive regions of the genome which are shown to be critical for tissue specificity[18]. Such reads from repetitive regions can be aligned to multiple positions (Fig. 1a - left) and are referred to as multi-mapping reads or multi-reads for short. The critical drawbacks of discarding multi-reads have been recognized in other classes of genomic studies such as transcriptome sequencing (RNA-seq)[19], chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq)[20,21], as well as genome-wide mapping of protein-RNA binding sites (CLIP-seq or RIP-seq)[22]. In this work, we developed mHi-C, a hierarchical model that probabilistically allocates Hi-C multi-reads to their most likely genomic origins by utilizing specific characteristics of the paired-end reads of the Hi-C assay. mHi-C is implemented as a full analysis

pipeline (https://github.com/keleslab/mHiC) that starts from unaligned read files and produces a set of statistically significant contacts at a given resolution.

For developing mHi-C and studying its operating characteristics, we utilized samples from: **(i)** a primary human fibroblast cell line (IMR90)[23] and **(ii)** three different asexual stages of the malaria parasite *Plasmodium falciparum* 3D7[24]. These samples represent different study designs and span a wide range of sequencing depths from both a large ($\sim$3B bp) and a small ($\sim$21M bp) genome. Before applying mHi-C model to these datasets and investigating biological implications, we established the substantial contribution of multi-reads to the sequencing depth. At read-end level (Supplementary Table 1 for terminology), after the initial step of aligning to the reference genome (Supplementary Fig. 1), multi-reads constitute approximately 10% of all the mappable read ends (Supplementary Tables 2 and 3). Moreover, the contribution of multi-reads to the set of aligned reads further increases by an additional 1 to 2% when chimeric reads (Supplementary Table 1) are also taken into account (Supplementary Tables 4 and 5). Most notably, Supplementary Fig. 3 for IMR90 and Supplementary Figs. 4a,b for the *P. falciparum* demonstrate that multi-reads make up a larger percentage of usable reads compared to uniquely mapping chimeric reads that are routinely rescued in Hi-C analysis pipelines[12,14]. At the read pair level, after joining of both ends, multi-reads increase the sequencing depth by 18% to 23% and thereby provide a substantial increase to the depth before the read pairs are further processed into bin pairs (Fig. 1a - right, Supplementary Table 6 for IMR90 and Supplementary Table 7 for the *P. falciparum*).

As part of the post alignment pre-processing steps, Hi-C reads go through a series of validation checking to ensure that the reads that represent biologically meaningful fragments are retained and used for downstream analysis (Supplementary Figs. 1 and 2, Supplementary Table 1). mHi-C pipeline tracks multi-reads through these processing steps. Remarkably, in the application of mHi-C to both the IMR90 and *P. falciparum* datasets, a subset of the high quality multi-reads are reduced to uni-reads either in the validation step when only one candidate contact passes the validation screening, or because all the alignments of a multi-read reside within the same bin (Fig. 1b - left, Supplementary Table 1 and see Methods). Collectively, mHi-C rescues 5% more valid read pairs (Fig. 1b - right) that originate from multi-reads and are mapped unambiguously without carrying out any multi-reads specific procedure. Such 5% corresponds to millions of reads for deeper sequenced datasets (Supplementary Tables 8 and 9). For the remaining multi-reads (Fig. 1b - right, colored in red), mHi-C implements a novel multi-mapping model and probabilistically allocates them.

mHi-C generative model (Fig. 1c and see Methods) is constructed at the bin-pair level to accommodate the typical signal sparsity of genomic interactions. The bins are either fixed-size non-overlapping genome intervals, 40kb for IMR90 and 10kb for *P. falciparum* in this paper, or fixed number of restriction fragments derived from the Hi-C protocol. We denote the observed alignment indicator vector for a given paired-end read i by vector $\boldsymbol{Y}_i$ and use unobserved hidden variable vector $\boldsymbol{Z}_i$ to indicate its true genomic origin. Contacts captured by Hi-C assay can arise as random contacts of nearby genomic positions or true biological interactions. mHi-C

generative model acknowledges this feature by utilizing data-driven priors, $\pi_{(j,k)}$ for bin pair j and k, as a function of contact distance between the two bins. mHi-C updates these prior probabilities for each candidate bin pair that a multi-read can be allocated to by leveraging local contact counts. As a result, for each multi-read i, it estimates posterior probabilities of genomic origin variable $\boldsymbol{Z_i}$. Specifically, $Pr(Z_{i,(j,k)} = 1 \mid \boldsymbol{Y_i}, \boldsymbol{\pi})$ denotes the posterior probability, i.e., allocation probability, that the two read ends of multi-read i originate from bins j and k. These posterior probabilities, which can also be viewed as fractional contacts of multi-read i, are then utilized to assign each multi-read to most likely genomic origin. Our results in this paper only utilized reads with allocation probability greater than 0.5. This ensured the output of mHi-C to be compatible with the standard input of the downstream normalization and statistical significance estimation methods[11,25].

Before quantifying mHi-C model performance, we first provide direct visual comparison of the contact matrices between Uni-setting and Uni&Multi-setting. Fig. 1d and Supplementary Figs. 5-7 clearly illustrate how multi-mapping reads fill in the low mappable regions and lead to more complete matrices, corroborating that repetitive genomic regions are under-represented without multi-reads. In addition to increasing the sequencing depth in extremely low contact bins, higher bin-level coverage after leveraging multi-mapping read pairs appears as a global pattern across the genome (Fig. 1e, Supplementary Fig. 8 for IMR90 and Supplementary Fig. 4c for *P. falciparum*). Overall, for the combined replicates of IMR90, 94.06% of the 40kb bins are covered by at least 50 contacts under the Uni&Multi-setting, compared to 93.54% under Uni-setting, thereby allowing us to study 15.80Mb more of the genome. Those major improvements in coverage provide direct evidence that mHi-C is rescuing multi-reads that originate from biologically valid fragments.

In addition to direct comparison of the raw contact matrices, we also identified the set of significant contacts by Fit-Hi-C[11] and performed a focused comparison of the identified interactions. At 5% false discovery rate (FDR), mHi-C detects 20% - 35% more novel significant contacts for relatively highly sequenced replicates, e.g., replicates 1 to 4 for IMR90 (red lines in Fig. 1f and Supplementary Fig. 9). The gains are markedly larger for smaller sequencing depth replicates, i.e., replicates 5 and 6 for IMR90 and the *P. falciparum* datasets (red lines in Supplementary Fig. 4d). Overall gains in the number of novel contacts persist as the cutoff for mHi-C posterior probabilities of multi-read assignments varies. Furthermore, improvement in the number of significant contacts is more evident for the top contacts identified at the smaller FDR thresholds. At fixed FDR, significant contacts identified by the Uni&Multi-setting also include the majority of significant contacts inferred from the Uni-setting, indicating that incorporating multi-reads is extending the significant contact list (low level of blue lines in Fig. 1f, Supplementary Fig. 9 for IMR90 and Supplementary Fig. 4d for *P. falciparum*). In general, harnessing the multi-reads with posterior probability larger than 0.5 and controlling the FDR at 1%, Fit-Hi-C identifies 24.53% more significant contacts compared to Uni-setting (comparing dark green to dark purple bars in Supplementary Figs. 10 and 11) and 32.70% of these are unique to Uni&Multi-setting (light green bars in Supplementary Figs. 10 and 11) collectively for all the

six replicates of IMR90. We next asked whether these novel contacts could have been identified by the Uni-setting by employing a more liberal FDR threshold. We observed that 56.30% of the newly detected significant contacts at 1% FDR (i.e., 18.41% compared to the all the significant contacts under Uni-setting) cannot be recovered even by a more liberal significant contact list under Uni-setting at 10% FDR (yellow bars in Supplementary Figs. 10 and 11). Conversely, Uni&Multi-setting is unable to recover only 1.13% of the Uni-setting contacts once the FDR is controlled at 10% for the Uni&Multi-setting (light blue bars in Supplementary Figs. 10 and 11), highlighting again that Uni&Multi-setting predominantly adds on novel significant contacts while retaining contacts that are identifiable under the Uni-setting. Similar analysis for *P. falciparum* stages is summarized in Supplementary Fig. 4e. We further confirmed this consistent power gain by a Receiver Operating Characteristic (ROC) and a Precision-Recall (PR) analysis (Supplementary Fig. 12). Supplementary Fig. 12b illustrates that at the same false discovery rate (1-precision), mHi-C achieves consistently higher power (recall) than the Uni-setting in addition to better AUROC performance (Supplementary Fig. 12a).

Before further investigating biological implications of mHi-C rescued multi-reads, we considered intuitive heuristic strategies for rescuing multi-reads at different stages of the Hi-C analysis pipeline as alternatives to mHi-C (Fig. 1g and see Methods). Specifically, AlignerSelect and DistanceSelect rescue multi-reads by simply choosing one of the alignments of a multi-read pair by default aligner strategy and based on distance, respectively. In addition to these, we designed a direct alternative to mHi-C, named SimpleSelect, as a model-free approach that imposes genomic distance priors in contrast to leveraging of the local interaction signals of the bins by mHi-C (e.g., local contact counts due to other read pairs in candidate bin pairs). Reanalysis of the Hi-C datasets with these strategies suggests that AlignerSelect and DistanceSelect allocate more valid reads compared to SimpleSelect or mHi-C (Supplementary Fig. 13). However, the numbers of significant contacts detected by SimpleSelect and mHi-C exceed those identified by AlignerSelect and DistanceSelect consistently for all the replicates and across different FDR thresholds (Fig. 1h). This indicates that AlignerSelect and DistanceSelect tend to select alignments of multi-reads that are due to random interactions; as a result, their gains in depth do not translate into more significant interactions. SimpleSelect, by-design, over-emphasizes the genomic distance in determining the origins of multi-reads and assigns multi-reads to short-range contact positions compared to mHi-C (Supplementary Fig. 14). The distance bias of simpleSelect is especially apparent in lower depth samples (Fig. 1i). We evaluated the direct biological consequence of heavily biasing read assignment by genomic distance, as employed by SimpleSelect, by comparing the significant contacts among three life stages of *P. falciparum*. We reasoned that the better multi-reads allocation strategy would reveal a differential analysis pattern more consistent with the Uni-setting, whereas a genomic distance biased strategy - SimpleSelect - will underestimate differences since multi-reads will be more likely to be allocated to candidate bin pairs with shortest genomic distance regardless of other local contact signals. Supplementary Fig. 15 corroborates this drawback of SimpleSelect and demonstrates that mHi-C differential patterns agree better with that of the Uni-setting. More-

over, Supplementary Fig. 15 suggests that rings stage is more similar to schizonts, an observation consistent with existing findings on *P. falciparum* life stages[24,26]. Further comparisons between SimpleSelect and mHi-C are discussed in Methods and illustrate consistently better performance of mHi-C over SimpleSelect. We focus the remainder of our investigations on the downstream impact of mHi-C rescued multi-reads.

We assessed the impact of multi-reads rescued by mHi-C on the reproducibility from the point of both raw contact counts and significant interactions detected. First, we used the stratum-adjusted correlation coefficient[27], a recent state-of-the-art approach for evaluating the reproducibility of Hi-C contact matrices. Fig. 2a demonstrates that integrating multi-reads leads to increased reproducibility and reduced variability of stratum-adjusted correlation coefficients among biological replicates. Next, we considered the degree of overlap among the significant contacts identified at 5% and 10% FDR across replicates (Supplementary Fig. 16a). We observed that significant contacts specific to Uni&Multi-setting have consistently higher reproducibility than those specific to Uni-setting across all pairwise comparisons. Since random contacts tend to arise due to short genomic distances between loci, we stratified the significant contacts based on distance and reassessed the reproducibility as a function of genomic distance between the contacts (Supplementary Fig. 16b). Notably, significant contacts identified only by the Uni-setting and those common to both settings have a stronger gradual descending trend as a function of the genomic distance, indicating decaying reproducibility for long range interactions. In contrast, Uni&Multi-setting maintains a relatively higher and stable reproducibility for longer genomic distances.

To further establish the biological implications of mHi-C rescued multi-reads, we investigated genomic features of novel contacts. Annotation of the significant contacts with ChromHMM segmentations from the Roadmap Epigenomics project[28] highlights marked enrichment of significant contacts in annotations involving repetitive DNA (Fig. 2b, Supplementary Fig. 17a). Most notably, ZNF genes & repeats and Heterochromatin states exhibit the largest discrepancy of the average significant contact counts between the Uni- and Uni&Multi-settings. To complement the evaluation with ChromHMM annotations, we evaluated the Uni-setting and Uni&Multi-setting significant contact enrichment of genomic regions harboring histone marks and other biochemical signals[23,29] (See Methods) by comparing their average contact counts to those without such signal (Fig. 2c). Notably, while we observe that multi-reads boost average number of significant contacts for biochemically active regions of the genome, they contribute more to regions that harbor H3K27me3 peaks (Fig. 2c, Supplementary Fig. 17b). Such regions are associated with downregulation of nearby genes through forming heterochromatin structure[30]. Supplementary Figs. 18-20 further provide specific examples of how increased marginal contact counts due to multi-reads are supported by signals of histone modifications, CTCF binding sites, and gene expression. Such regions harbor many genes of biological significance. For example, NBPF1 (Supplementary Fig. 18) is implicated in many neurogenetic diseases and its family consists of dozens of recently duplicated genes primarily located in segmental duplications[31]. In addition, RABL2A within the highlighted region of

Supplementary Fig. 19 is member of RAS oncogene family.

We find that a significant impact of multi-reads is for detection of promoter-enhancer interactions. Overall, mHi-C identifies 14.89% more significant promoter-enhancer contacts at 5% FDR (Supplementary Table 10). Among them, 13,313 enhancer-promoter contacts are reproducible among six replicates for IMR90 under Uni&Multi-setting (Supplementary Table 11) and 62,971 reproducible for at least two replicates (Supplementary Table 12) leading to 15.84% more novel promoter-enhancer interactions specific to Uni&Multi-setting. Fig. 2d provides WashU epigenome browser[32] display of such novel highly reproducible enhancer-promoter interactions. Supplementary Fig. 21 depicts these interactions in more details across the 6 replicates with more examples in Supplementary Fig. 22. We next validated the novel promoter-enhancer interactions by investigating the expression levels of the genes contributing promoters to these interactions. Fig. 2e supports that genes with significant interactions in their promoters generally exhibit higher expression levels (comparing bars 1-5 to bars 8-9 in Fig. 2e). Furthermore, if these contacts involve an enhancer, the average gene expression can be 38.17% higher than that of the overall promoters with significant interactions (comparing bars 10-11 to bars 1-2 in Fig. 2e). Most remarkably, newly detected significant promoter-enhancer contacts (bar 13 in Fig. 2e) exhibit a stably higher gene expression level, highlighting that, without multi-reads, biologically supported promoter-enhancer interactions are underestimated. In addition, an overall evaluation of significant interactions (5% FDR) that considers interactions from promoters with low expression (TPM $\leq$ 1) as false positives illustrates that mHi-C specific significant promoter interactions have false positive rates comparable to or smaller than that of significant promoter interactions common to Uni- and Uni&Multi-settings (Supplementary Fig. 23). In contrast Uni-setting specific contacts have elevated false positive rates.

We also investigated the impact of mHi-C rescued multi-reads on the topologically associating domains (TADs)[33]. We used the DomainCaller[34] to infer TADs of IMR90 under both settings. Uni&Multi-setting identifies similar number of TADs compared to Uni-setting TADs (Supplementary Fig. 24). We utilized TAD detection across individual replicates to evaluate power and false positive rates of both settings. Notably, even though TADs are stable structures even across species[34-36], Uni&Multi-setting identifies 2.01% more reproducible TADs across replicates with 2.36% lower non-reproducible TADs (Fig. 2f). Several studies have revealed the role of CTCF in establishing the boundaries of genome architecture[37-39]. We observed that a larger proportion of the detected TADs have CTCF peaks with convergent CTCF motif pairs at the boundaries once multi-reads are utilized. (Supplementary Figs. 25a-c). Fig. 2g provides an explicit example of how the gap in the contact matrix due to depletion of multi-reads biases the inferred TAD structure. In addition to discovery of novel TADs (Supplementary Figs. 26) by filling in the gaps in the contact count matrix and boosting the domain signals, mHi-C also refines TAD boundaries (Supplementary Fig. 27 and 28) and eliminates potential false positive TADs that are split by the white gaps in Uni-setting (Supplementary Figs. 29-31). The novel, adjusted, and eliminated TADs are largely supported by CTCF ChIP-seq signal[21] as well as convergent CTCF motifs (Supplementary Fig. 25d), providing evidence for mHi-C driven

modifications to these TADs and revealing a lower false discovery rate for mHi-C compared to Uni-setting (Fig. 2h and Supplementary Fig. 25e).

Hi-C data are powerful for identifying long-range interacting loci, chromatin loops, topologically associating domains (TADs), and A/B compartments[40]. Our Hi-C multi-mapping strategy, mHi-C, yields remarkable increase in sequencing depth. This translates into significant and consistent gains in reproducibility of the raw contact counts, identified significant contacts and promoter-enhancer interactions, and domain structures. In addition to discovering novel promoter-enhancer interactions, mHi-C provides a less biased assessment of Hi-C signal originating from highly repetitive regions.

Acknowledgments This work was supported by NIH HG009744 and NIH HG007019 (S.K.). F.A. is partially supported by Institute Leadership Funds from La Jolla Institute for Allergy and Immunology.

Author Contributions F.A. and S.K. conceived the project. Y.Z., F.A., and S.K. designed the research. Y.Z. and S.K. developed the method. Y.Z. constructed the pipeline and performed experiments. All authors contributed to the preparation of the manuscript.

Competing Interests The authors declare no competing financial interests.

Correspondence keles@stat.wisc.edu.

1. Dekker, J., Rippe, K., Dekker, M., Kleckner, N.: Capturing chromosome conformation. Science **295**(5558) (2002) 1306–11

2. de Laat, W., Duboule, D.: Topology of mammalian developmental enhancers and their regulatory landscapes. Nature **502**(7472) (2013) 499–506

3. Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., et al.: Two independent modes of chromatin organization revealed by cohesin removal. Nature **551**(7678) (2017) 51

4. Rosa-Garrido, M., Chapski, D.J., Schmitt, A.D., Kimball, T.H., Karbassi, E., Monte, E., Balderas, E., Pellegrini, M., Shih, T.T., Soehalim, E., et al.: High resolution mapping of chromatin conformation in cardiac myocytes reveals structural remodeling of the epigenome in heart failure. Circulation (2017) Circulation–117

5. Lee, D., Tan, W., Anene, G., Li, P., Danh, T., Tiang, Z., Ng, S.L., Efthymios, M., Autio, M., Jiang, J., et al.: Gene neighbourhood integrity disrupted by CTCF loss in vivo. bioRxiv (2017) 187393

6. Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., Mirny, L.: Emerging Evidence of Chromosome Folding by Loop Extrusion. bioRxiv (2018) 264648

7. 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.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science **326**(5950) (2009) 289–293

8. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., Aiden, E.L.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**(7) (2014) 1665–1680

9. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K.: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular Cell **38**(4) (2010) 576–589

10. Hwang, Y.C., Lin, C.F., Valladares, O., Malamon, J., Kuksa, P.P., Zheng, Q., Gregory, B.D., Wang, L.S.: HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics **31**(8) (2014) 1290–1292

11. Ay, F., Bailey, T.L., Noble, W.S.: Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research **24**(6) (2014) 999–1011

12. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., Barillot, E.: HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biology **16**(1) (2015) 259

13. Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al.: Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics **47**(6) (2015) 598–606

14. Lun, A.T., Smyth, G.K.: diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics **16**(1) (2015) 258

15. Ay, F., Noble, W.S.: Analysis methods for studying the 3D architecture of the genome. Genome Biology **16**(1) (2015) 183

16. Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., Bicciato, S.: Comparison of computational methods for Hi-C data analysis. Nature Methods **14**(7) (2017) 679

17. Yardimci, G., Ozadam, H., Sauria, M.E., Ursu, O., Yan, K.K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B.R., Song, F., et al.: Measuring the reproducibility and quality of Hi-C data. bioRxiv (2017) 188755

18. Xie, M., Hong, C., Zhang, B., Lowdon, R.F., Xing, X., Li, D., Zhou, X., Lee, H.J., Maire, C.L., Ligon, K.L., Gascard, P., Sigaroudinia, M., Tlsty, T.D., Kadlecek, T., Weiss, A., O'Geen, H., Farnham, P.J., Madden, P.A.F., Mungall, A.J., Tam, A., Kamoh, B., Cho, S., Moore, R., Hirst, M., Marra, M.A., Costello, J.F., Wang, T.: DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nature Genetics **45**(7) (2013) 836–41

19. Li, B., Dewey, C.N.: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **12**(1) (2011) 323

20. Chung, D., Kuan, P.F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E.H., Dewey, C., Keleş, S.: Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Computational Biology **7**(7) (2011) e1002111

21. Zeng, X., Li, B., Welch, R., Rojo, C., Zheng, Y., Dewey, C.N., Keleş, S.: Perm-seq: mapping protein-DNA interactions in segmental duplication and highly repetitive regions of genomes with prior-enhanced read mapping. PLoS Computational Biology **11**(10) (2015) e1004491

22. Zhang, Z., Xing, Y.: CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Research **45**(16) (2017) 9260–9271

23. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C., Ren, B.: A high-resolution map of three-dimensional chromatin interactome in human cells. Nature **503**(7475) (2013) 290

24. Ay, F., Bunnik, E.M., Varoquaux, N., Bol, S.M., Prudhomme, J., Vert, J.P., Noble, W.S., Le Roch, K.G.: Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Research **24**(6) (2014) 974–988

25. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., Mirny, L.A.: Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods **9**(10) (2012) 999–1003

26. Bunnik, E.M., Cook, K.B., Varoquaux, N., Batugedara, G., Prudhomme, J., Shi, L., Andolina, C., Ross, L.S., Brady, D., Fidock, D.A., et al.: Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages. bioRxiv (2018) 242123

27. Yang, T., Zhang, F., Yardimci, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., Li, Q.: HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Research (2017) gr–220640

28. Roadmap Epigenomics Consortium: Integrative analysis of 111 reference human epigenomes. Nature **518**(7539) (2015) 317–330

29. The ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature **489** (09 2012) 57–74

30. Ferrari, K.J., Scelfo, A., Jammula, S., Cuomo, A., Barozzi, I., Stützer, A., Fischle, W., Bonaldi, T., Pasini, D.: Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity. Molecular Cell **53**(1) (2014) 49–62

31. Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., et al.: GeneCards Version 3: the human gene integrator. Database **2010** (2010)

32. Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E.A., Koebbe, B.C., Nielsen, C., Hirst, M., Farnham, P., et al.: The human epigenome browser at Washington University. Nature Methods **8**(12) (2011) 989–990

33. Pombo, A., Dillon, N.: Three-dimensional genome architecture: players and mechanisms. Nature Reviews. Molecular Cell Biology **16**(4) (2015) 245–257

34. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature **485**(7398) (2012) 376–380

35. 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.: Chromatin architecture reorganization during stem cell differentiation. Nature **518**(7539) (2015) 331

36. Dekker, J., Heard, E.: Structural and functional diversity of topologically associating domains. FEBS Letters **589**(20PartA) (2015) 2877–2884

37. Ong, C.T., Corces, V.G.: CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics **15**(4) (2014) 234

38. Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S.Z., Penrad-Mobayed, M., Sachs, L.M., Ruan, X., Wei, C.L., Liu, E.T., Wilczynski, G.M., Plewczynski, D., Li, G., Ruan, Y.: CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell **163**(7) (2015) 1611–1627

39. Hsu, S.C., Gilgenast, T.G., Bartman, C.R., Edwards, C.R., Stonestrom, A.J., Huang, P., Emerson, D.J., Evans, P., Werner, M.T., Keller, C.A., et al.: The BET protein BRD2 cooperates with CTCF to enforce transcriptional and architectural boundaries. Molecular Cell **66**(1) (2017) 102–116

40. Yu, M., Ren, B.: The Three-Dimensional Organization of Mammalian Genomes. Annual Review of Cell and Developmental Biology (0) (2017)

41. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al.: GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research **22**(9) (2012) 1760–1774

42. Consortium, E.P., et al.: The ENCODE (ENCyclopedia of DNA elements) project. Science **306**(5696) (2004) 636–640

43. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al.: NCBI GEO: archive for functional genomics data setsupdate. Nucleic Acids Research **41**(D1) (2012) D991–D995

44. Li, H., Durbin, R.: Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics **26**(5) (2010) 589–595

45. Martin, M.: Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal **17**(1) (2011) pp–10

46. Belaghzal, H., Dekker, J., Gibcus, J.H.: Hi-C 2.0: an optimized hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods **123** (2017) 56–65

47. Grant, C.E., Bailey, T.L., Nobel, W.S.: FIMO: Scanning for occurrences of a given motif. Bioinformatics **7** (2011) 1017

48. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al.: JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research **46**(D1) (2017) D260–D266

49. Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J.M., Ouellette, S.B., Ahzir, A., Kumar, N., et al.: HiGlass: Web-based visual comparison and exploration of genome interaction maps. bioRxiv (2017) 121889

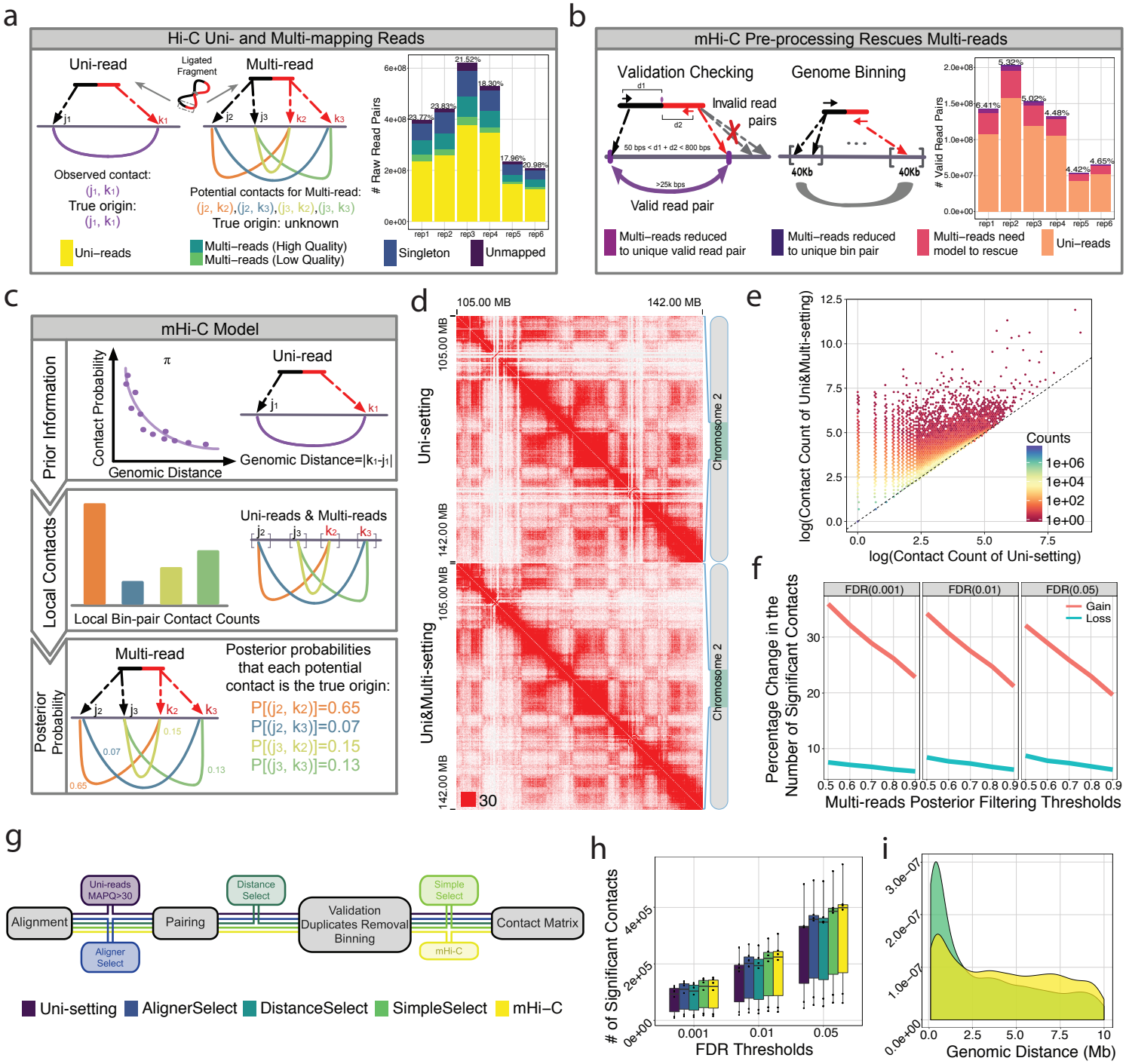
a
Hi-C Uni- and Multi-mapping Reads
Uni-read
Ligated Fragment
Multi-read
Observed contact:
True origin:
Potential contacts for Multi-read:
True origin: unknown
Uni–reads
Multi–reads (High Quality)
Multi–reads (Low Quality)
Raw Read Pairs
Singleton
Unmapped
b
mHi-C Pre-processing Rescues Multi-reads
Validation Checking
Genome Binning
Invalid read pairs
Valid read pair
Valid Read Pairs
Multi–reads reduced to unique valid read pair
Multi–reads reduced to unique bin pair
Multi–reads need model to rescue
Uni–reads
c
mHi-C Model
Prior Information
Contact Probability
Genomic Distance
Uni-read
Genomic Distance=|k1-j1|
Local Contacts
Uni-reads & Multi-reads
Local Bin-pair Contact Counts
Posterior Probability
Multi-read
Posterior probabilities that each potential contact is the true origin:
P[(j2, k2)]=0.65
P[(j2, k3)]=0.07
P[(j3, k2)]=0.15
P[(j3, k3)]=0.13
d
105.00 MB
142.00 MB
Uni-setting
Uni&Multi-setting
Chromosome 2
30
e
log(Contact Count of Uni&Multi-setting)
log(Contact Count of Uni-setting)
Counts
f
FDR(0.001)
FDR(0.01)
FDR(0.05)
Gain
Loss
Percentage Change in the Number of Significant Contacts
Multi-reads Posterior Filtering Thresholds
g
Alignment
Uni-reads MAPQ>30
Aligner Select
Pairing
Distance Select
Validation Duplicates Removal Binning
Simple Select
mHi-C
Contact Matrix
Uni-setting
AlignerSelect
DistanceSelect
SimpleSelect
mHi-C
h
# of Significant Contacts
FDR Thresholds
i
Genomic Distance (Mb)

Figure 1 Overview of mHi-C model and global impact of multi-reads in Hi-C data analysis. a. Standard Hi-C pipelines utilize uni-reads while discarding multi-mapping reads which give rise to multiple potential contacts. Right panel summarizes the total number of reads in different categories as a result of alignment to reference genome across the six IMR90 replicates. Percentages of high-quality multi-reads compared to uni-reads are depicted on top of each bar. **b.** Multi-mapping reads can be reduced to uni-reads within validation checking and genome binning pre-processing steps. Right panel classifies the aligned reads after validation checking and binning. Percentage improvements in sequencing depths due to multi-reads becoming uni-reads are depicted on top of each bar. **c.** mHi-C modeling starts from the prior built by only uni-reads to quantify the relationship between random contact probabilities and the genomic distance between the contacts. This prior is updated by leveraging local bin pair contacts including both uni- and multi-reads and results in posterior probabilities that quantify the evidence for each potential contact to be the true genomic origin. **d.** Contact matrices of IMR90 combined replicates (rep1-6) are compared under Uni-setting and Uni&Multi-setting for chr2:105,000,000-142,000,000. White gaps of Uni-reads contact matrix, due to lack of reads from repetitive regions, are filled in by multi-reads, hence resulting in a more complete contact matrix. **e.** Bin-level contact counts under Uni&Multi-setting and Uni-setting (IMR90 rep 1). The dashed line is y = x. mHi-C rescues multi-reads from valid ligation fragments, resulting in significant increase in contact counts. **f.** Percentage change in the number of significant contacts: gained (red) and lost (blue) by the Uni&Multi-setting compared to the Uni-setting (IMR90 rep 1). **g.** Intuitive heuristic strategies (AlignerSelect, DistanceSelect, SimpleSelect) for model-free assignment of multi-reads at various stages the of Hi-C analysis pipeline. SimpleSelect is designed as a direct model-free alternative to mHi-C. **h.** Numbers of significant contacts detected by different multi-read allocation strategies at different FDR thresholds. **i.** Comparison of genomic distance distributions of significant contacts between SimpleSelect and mHi-C for IMR90 rep5 (FDR $<$ 0.001).

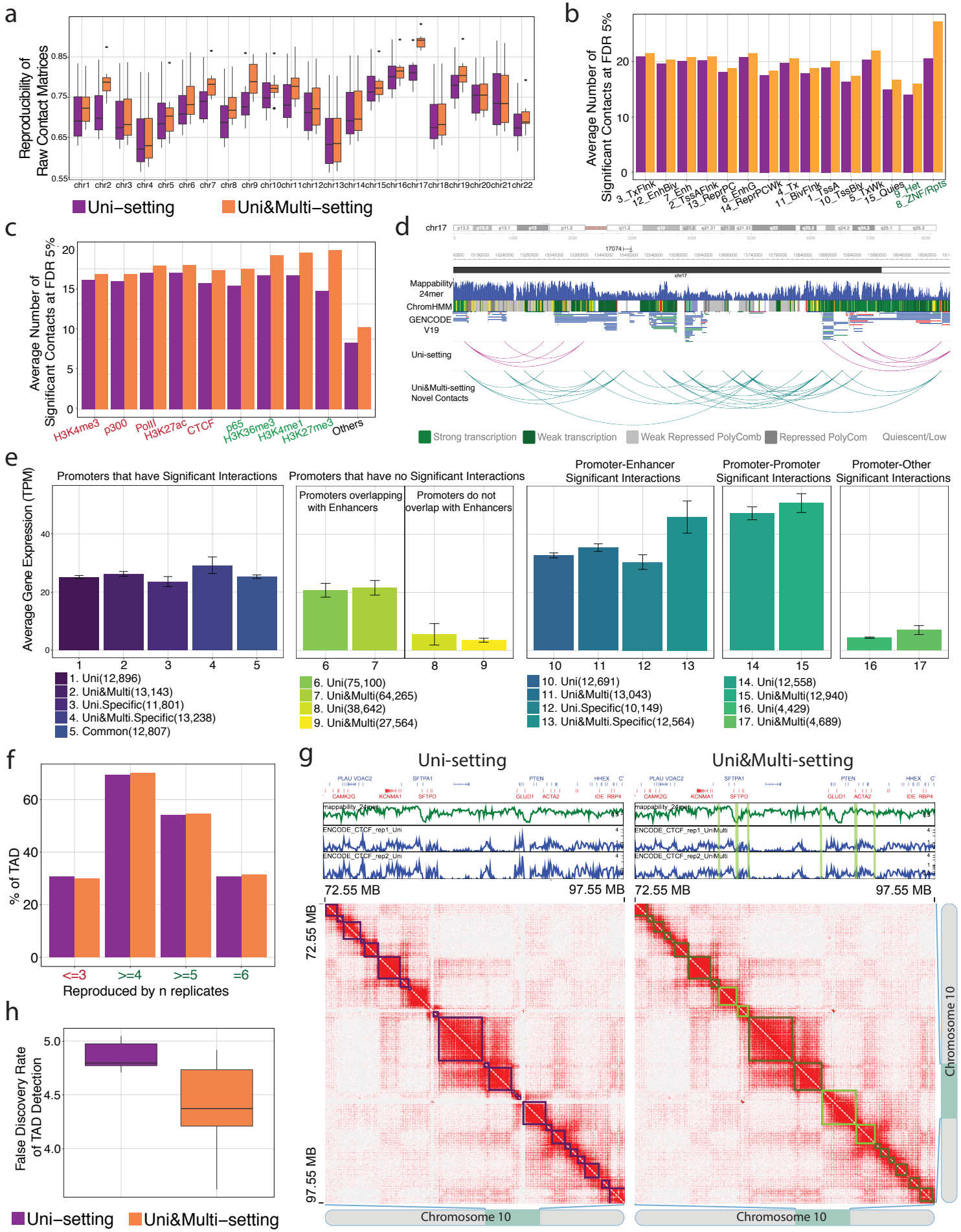
a
Reproducibility of Raw Contact Matrices
Uni−setting
Uni&Multi−setting
b
Average Number of Significant Contacts at FDR 5%
3_TxFlnk
12_EnhBiv
7_Enh
2_TssAFlnk
13_ReprPC
6_EnhG
14_ReprPCWK
4_Tx
11_BivFlnk
1_TssA
10_TssBiv
5_TxWk
15_Quies
9_Het
8_ZNF/Rpts
c
Average Number of Significant Contacts at FDR 5%
H3K4me3
p300
PolII
H3K27ac
CTCF
p65
H3K36me3
H3K4me1
H3K27me3
Others
d
chr17
Mappability 24mer
ChromHMM
GENCODE V19
Uni-setting
Uni&Multi-setting Novel Contacts
Strong transcription
Weak transcription
Weak Repressed PolyComb
Repressed PolyCom
Quiescent/Low
e
Average Gene Expression (TPM)
Promoters that have Significant Interactions
Promoters that have no Significant Interactions
Promoters overlapping with Enhancers
Promoters do not overlap with Enhancers
Promoter-Enhancer Significant Interactions
Promoter-Promoter Significant Interactions
Promoter-Other Significant Interactions
1. Uni(12,896)
2. Uni&Multi(13,143)
3. Uni.Specific(11,801)
4. Uni&Multi.Specific(13,238)
5. Common(12,807)
6. Uni(75,100)
7. Uni&Multi(64,265)
8. Uni(38,642)
9. Uni&Multi(27,564)
10. Uni(12,691)
11. Uni&Multi(13,043)
12. Uni.Specific(10,149)
13. Uni&Multi.Specific(12,564)
14. Uni(12,558)
15. Uni&Multi(12,940)
16. Uni(4,429)
17. Uni&Multi(4,689)
f
% of TAD
<=3
>=4
>=5
=6
Reproduced by n replicates
g
Uni-setting
Uni&Multi-setting
72.55 MB
97.55 MB
Chromosome 10
h
False Discovery Rate of TAD Detection
Uni−setting
Uni&Multi−setting

Figure 2 Downstream analysis and biological impacts of mHi-C rescued multi-reads. a. HiCRep[27] reproducibility evaluation of raw contact matrices among IMR90 six replicates. Box plots depict all pairwise comparisons between the replicates. **b.** Average number of significant contacts (5% FDR) identified by Uni- and Uni&Multi-setting for each chromHMM annotation category. **c.** Average number of significant contacts (5% FDR) across ChIP-seq peaks associated with different genomic functions. Red/Green labels denote smaller/larger differences between the two settings compared to the differences observed in the "Others" category that depict non-peak regions. **d.** mHi-C identifies novel significant enhancer-promoter contacts (green) that are reproducible among at least two replicates in addition to those reproducible under the Uni-setting (red). **e.** Average gene expression with standard errors for five different scenarios of interactions that group promoters into six different categories. In the first panel, significant interactions involving promoters are classified into five settings and the average gene expressions across genes with the corresponding promoters are depicted. Second panel involves two alignment settings and genes without any promoter interactions at 5% FDR. This panel is further separated into two categories: promoters that overlap with enhancer annotated regions and those that do not. The latter one serves as the baseline for average expression. Genes contributing to the third and fourth panel have promoter-enhancer, promoter-promoter interactions at 5% FDR. The fifth panel considers genes whose promoters have significant interactions with non-enhancer and non-promoter regions. Numbers in the parenthesis correpond to the number of transcripts in each category. **f.** Percentage of topologically associating domains (TADs) that are reproducibly detected under Uni-setting and Uni&Multi-setting. TADs that are not detected in at least 4 of the 6 replicates are considered as non-reproducible. **g.** Comparison of the contact matrices with superimposed TADs between Uni- and Uni&Multi-setting for chr10:72,550,000-97,550,000. TADs affected by white gaps involving repetitive regions are highlighed in light green. Gene tracks from GENCODE[41] followed by 24mer mappability tracks are displayed above each contact matrix. The third and fourth tracks are CTCF peaks from ENCODE[42]. Light green shaded areas correspond to new TAD boundaries. **h.** False discovery rate of TADs detected under two settings. TADs that are not reproducible and lack CTCF peaks at the TAD boundaries are labeled as false positives.

METHODS

Datasets Datasets utilized for evaluating mHi-C are a primary human fibroblast cell line (IMR90)[23] and three different asexual stages of the malaria parasite *Plasmodium falciparum* 3D7[24] with accession numbers GSE43070 (IMR90) and GSE50199 (*Plasmodium falciparum*) obtained from NCBI Gene Expression Omnibus[43]. IMR90 datasets have 36 bps paired-end reads while the read length for the *Plasmodium falciparum* datasets is 40 bps. These samples represent different study designs and span a wide range of sequencing depths from both a large and a small genome. Restriction enzymes, HindIII, which recognizes sequence AAGCTT, and MboI, which recognizes GATC, were used in the IMR90 and *Plasmodium falciparum* studies, respectively. Taking sequencing depths, genome sizes, and the distribution of restriction enzyme recognition sites on the genomes into consideration, we studied the interactions at the resolution of 40kb for IMR90 and 10kb for the *P. falciparum*.

mHi-C workflow We developed a complete pipeline customized for incorporating high quality multi-mapping reads into the Hi-C data analysis work flow. The overall pipeline, illustrated in Supplementary Figs. 1 and 2, incorporates the essential steps of the Hi-C analysis pipelines. In what follows, we outline the major steps of the analysis to explicitly track multi-reads and describe how mHi-C utilizes them.

Read end alignment: uni- and multi-reads and chimeric reads. The first step in the mHi-C pipeline is alignment of each read end separately to the reference genome. The default aligner in the mHi-C software is BWA[44]; however mHi-C can work with any aligner that outputs multi-reads. The default alignment parameters are (i) edit distance maximum of 2 including mismatches and gap extension; and (ii) maximum number of gap open of 1. mHi-C sets the maximum number of alternative hits saved in the XA tag to be 99 to keep track of multi-reads. If the number of alternative alignments exceed the threshold of 99 in the default setting, these alignments are not recorded in XA tag. We regarded these alignments as low quality multi-mapping reads compared to those multi-mapping reads that have relatively smaller number of alternative alignments. In summary, low quality multi-mapping reads are discarded together with unmapped reads, only leaving uniquely mapping reads and high quality multi-mapping reads for downstream analysis. mHi-C pipeline further restricts the maximum number of mismatches (maximum to be 2 compared to 3 in BWA default setting) to ensure that the alignment quality of multi-reads is comparable to that of standard Hi-C pipeline.

Chimeric reads, that span ligation junction of the Hi-C fragments (Supplementary Fig. 1) are also a key component of Hi-C analysis pipelines. The ligation junction sequence can be derived from the restriction enzyme recognition sites and used to rescue chimeric reads. mHi-C adapts the pre-splitting strategy of diffHiC[14], which is modified from the existing Cutadapt[45] software. Specifically, the read ends are trimmed to the center of the junction sequence. If the trimmed left $5'$ ends are not too short, e.g., ≥ 25 bps, these chimeric reads are remapped to the reference genome. As the lengths of the chimeric reads become shorter, these reads tend to become multi-reads.

Valid fragment filtering. While each individual read end is aligned to reference genome separately, inferring interacting loci relies on alignment information of paired-ends. Therefore, read ends are paired after unmapped and singleton read pairs as well as low quality multi-mapping ends (Supplementary Fig. 1 and Supplementary Table 1) are discarded. After pairing, read end alignments are further evaluated for their representation of valid ligation fragments that originate from biologically meaningful long-range interactions (Supplementary Fig. 1). First, reads that do not originate from around restriction enzyme digestion sites are eliminated since they primarily arise due to random breakage by sonication[46]. This is typically achieved by filtering the reads based on the total distance of two read end alignments to the restriction site. We required the total distance to be within 50-800 bps for IMR90 and 50-500 bps for *P. falciparum*. The lower bound of 50 for this parameter is motivated by the chimeric reads with as short as 25 bps on both ends. Second, a single Hi-C interaction ought to involve two restriction fragments. Therefore, read ends falling within the same fragment, either due to dangling end or self-circle ligation, are filtered. Third, because the nature of chromatin folding leads to abundance of random short-range interactions, interactions of two regions that are too close in genomic distance are highly likely to be random interaction without regulatory implications. As a result, reads with ends aligning too close to each other are also filtered. mHi-C uses the twice the resolution rule, namely 80kb for IMR90 and 20kb for *P. falciparum*. Notably, as a result of this valid fragment filtering, some multi-mapping reads can be counted as uniquely mapping reads (Supplementary Table 1 - 2b). This is because, although a read pair has multiple potential genomic origins dictated by its multiple alignments, only one of them ends up passing the validation screening. Once the multi-mapping uncertainty is eliminated, such read pairs are passed to the downstream analysis as uni-reads. We remark here that standard Hi-C analysis pipelines do not rescue these multi-reads.

Duplicate removal. To remove PCR duplicates, mHi-C considers the following two issues. First, due to allowing maximum number of 2 mismatches in alignment, some multi-reads may have the exact same alignment position and strand direction with uni-reads. If such duplicates arise, uni-reads are granted higher priority and the overlapping multi-reads together with all their potential alignment positions are discarded completely. This ensures that the uni-reads that arise in standard Hi-C analysis pipelines will not be discarded as PCR duplicates in the mHi-C pipeline. Second, if a multi-mapping read alignment is duplicated with another multi-read, the read with smaller alphabetical read query name will be preserved. More often than not, if multi-read A overlaps multi-read B at a position, then it is highly likely that they will overlap at other positions as well. This convention ensures that it is always the read pair A alignments that are being retained (Supplementary Fig. 2).

Genome binning. Due to the typical limited sequencing depths of Hi-C experiments, reference genome is divided into small non-overlapping intervals, i.e., bins, to secure enough number of contact counts across units. The unit can be fix-sized genomic intervals or a fixed number of consecutive restriction fragments. mHi-C can switch between the two unit options with ease. After binning, the interaction unit reduces from alignment position pairs to bin pairs.

Remarkably, multi-mapping reads, ends of which are located within the same bin pair, reduce to uni-reads as their potential multi-mapping alignment position pairs support the same bin pair contact. Therefore, there is no need to distinguish the candidate alignments within the same bin (Supplementary Fig. 2 and Supplementary Table 1 - 3b).

mHi-C generative model and parameter estimation. mHi-C infers genomic origins of multi-reads at the bin pair level (Supplementary Table 1). We denoted the whole alignment vector for a given paired-end read i by vector $\mathbf{Y}_i$. If the two read ends of read i align to only bin j and bin k, respectively, we set the respective components of the alignment vector as: $Y_{i,(j,k)}$ = 1 and $Y_{i,(j',k')} = 0, \forall \quad j' \neq j, k' \neq k$. Index of read, i, ranges from 1 to N, where N is total number of valid Hi-C reads, including both uni-reads and multi-reads that pass the essential processing in Supplementary Figs. 1 and 2. Overall, the reference genome is divided into M bins and j represents the bin index of the end, alignment position of which is upstream compared to the other read end position indicated by k. Namely, j takes on a value from 1 to $M-1$ and k runs from $j+1$ to the maximum value M. For uniquely mapping reads, only one alignment is observed, i.e., $\sum_{(j,k)}^{(M-1,M)} Y_{i,(j,k)} = 1$. However, for multi-mapping reads, we have $\sum_{(j,k)}^{(M-1,M)} Y_{i,(j,k)} > 1$.

We next defined a hidden variable $Z_{i,(j,k)}$ to denote the true genomic origin of read i. If read i originates from position bin pairs j and k, we have $Z_{i,(j,k)} = 1$. In addition, a read can only originate from one alignment position pair on the genome; thus, $\sum_{(j,k)}^{(M-1,M)} Z_{i,(j,k)} = 1$ for both uni- and multi-reads. We define O_i = $\{(j, k)$: $Z_{i,(j,k)}$ = 1$\}$ to represent true genomic origin of read i and S_{O_i} as the set of location pairs that read pair i can align to. Hence, $Y_{i,(j,k)}$ = 1, if $(j, k) \in S_{O_i}$. Under the assumption that the true alignment can only originate from those observed alignable positions, O_i must be one of the location pairs in S_{O_i}. We further assume that the indicators of true origin for read i, $\boldsymbol{Z_i}$ = $(Z_{i,(1,2)}, Z_{i,(1,3)}, ..., Z_{i,(M-1,M)})$ are random draws from a Dirichlet - Multinomial distribution. Specifically,

$$\mathbf{Z_i} \overset{i.i.d.}{\sim} \text{Multinomial}(\pi_{(1,2)}, \pi_{(1,3)}, \cdots, \pi_{(j,k)}, \cdots, \pi_{(M-1,M)}), \quad i = 1, \cdots, N, \tag{1}$$

where $\pi_{(j,k)}$ can be interpreted as contact probability between bin j and k ($j < k$). We further assume that

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\gamma_{(1,2)}, \gamma_{(1,3)}, \cdots, \gamma_{(j,k)}, \cdots, \gamma_{(M-1,M)}), \tag{2}$$

where $\boldsymbol{\pi} = (\pi_{(1,2)}, \pi_{(1,3)}, \cdots, \pi_{(M-1,M)})$ and $\gamma_{(j,k)}$ is a function of genomic distance and quantifies random contact probability. Specifically, we adapt the univariate spline fitting approach from Fit-Hi-C[11] for estimating random contact probabilities with respect to genomic distance and set $\gamma_{(j,k)} = \text{Spline}(j,k) \times N + 1$. Here, N is the total number of valid reads as defined above and $\text{Spline}(j,k)$ denotes the spline estimate of the random contact probability between bins j and k. Therefore, $\text{Spline}(j,k) \times N$ is the average random contact counts (i.e., pseudo-counts)

between bin j and k. As a result, the probability density function of π can be written as:

$$P(\boldsymbol{\pi}|\boldsymbol{\gamma}) = \frac{\Gamma(\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}\gamma_{(j,k)})}{\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\Gamma(\gamma_{(j,k)})}\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\pi_{(j,k)}^{(\gamma_{(j,k)}-1)}$$

$$= \frac{\Gamma(\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}(Spline(j,k)\times N+1))}{\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\Gamma(Spline(j,k)\times N+1)}\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\pi_{(j,k)}^{Spline(j,k)\times N}.$$

We next derive the full data joint distribution function.

Lemma 1. Given the true genomic origin under the mHi-C setting, the set of location pairs that a read pair can align to will have observed alignments with probability 1.

Proof.

$$P(\boldsymbol{Y_i}|Z_{i,(j,k)}=1) = P(\boldsymbol{Y_i}|O_i)$$

$$= \prod_{j=1}^{M-1}\prod_{k=j+1}^{M} P(Y_{i,(j,k)}|O_i)$$

$$= \prod_{j=1}^{M-1}\prod_{k=j+1}^{M} [1(Y_{i,(j,k)}=1,(j,k)\in S_{O_i}) + 1(Y_{i,(j,k)}\neq 1,(j,k)\notin S_{O_i})]$$

$$= 1. \qquad \square$$

Based on Lemma 1, we can get the joint distribution $P(\mathbf{Y},\mathbf{Z}|\pi)$ as

$$P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{i}^{N} P_\pi(\boldsymbol{Y_i},\boldsymbol{Z_i})$$

$$= \prod_{i}^{N}\prod_{j=1}^{M-1}\prod_{k=j+1}^{M} P_\pi(\boldsymbol{Y_i}, Z_{i,(j,k)}=1)^{Z_{i,(j,k)}}$$

$$= \prod_{i}^{N}\prod_{j=1}^{M-1}\prod_{k=j+1}^{M} [P_\pi(\boldsymbol{Y_i}|Z_{i,(j,k)}=1)\pi_{(j,k)}]^{Z_{i,(j,k)}}$$

$$= \prod_{i}^{N}\prod_{j=1}^{M-1}\prod_{k=j+1}^{M} \pi_{(j,k)}^{Z_{i,(j,k)}}.$$

Using the Dirichlet-Multinomial conjugacy, we derive the posterior distribution of π as

$$P(\boldsymbol{\pi}|\boldsymbol{Z}) \propto P(\boldsymbol{\pi},\boldsymbol{Z}) = \prod_{i=1}^{N} P(\boldsymbol{Z_i}|\boldsymbol{\pi})P(\boldsymbol{\pi})$$

$$\propto \prod_{j=1}^{M-1}\prod_{k=j+1}^{M} \pi_{(j,k)}^{(\sum_{i=1}^{N} Z_{i,(j,k)}+\gamma_{(j,k)}-1)}$$

$$= \prod_{j=1}^{M-1}\prod_{k=j+1}^{M} \pi_{(j,k)}^{(\sum_{i=1}^{N} Z_{i,(j,k)}+Spline(j,k)\times N)}.$$

We next derive an Expectation-Maximization algorithm for fitting this model.

E-step.

$$Z^{(t)}_{i,(j,k)} = E(Z_{i,(j,k)}|\boldsymbol{Y_i}, \boldsymbol{\pi}) = \frac{\pi^{(t)}_{(j,k)}}{\sum_{(j',k')\in S_{O_i}} \pi^{(t)}_{(j',k')}} 1[(j,k) \in S_{O_i}].$$

M-step.

$$\pi^{(t+1)}_{(j,k)} = \frac{\sum_{i=1}^{N} Z^{(t)}_{i,(j,k)} + Spline(j,k) \times N}{N + \sum_{j'=1}^{M-1} \sum_{k'=j+1}^{M} Spline(j',k') \times N}.$$

Estimate of the contact probability $\pi_{(j,k)}$ in the M-step can be viewed as an integration of local interaction signal, encoded in $\sum_{i=1}^{N} Z^{(t)}_{i,(j,k)}$, and random contact signal due to prior, i.e., Spline$(j,k) \times N$.

The by-products of the EM algorithm are posterior probabilities, P($Z_{i,(j,k)}$ = 1 | $\boldsymbol{Y_i}$, $\boldsymbol{\pi}$), which are utilized for assigning each multi-read to the most likely genomic origin. To keep mHi-C output compatible with the input required for the widely used significant contact detection methods, we filtered multi-reads with maximum allocation posterior probability less than or equal to 0.5 and assigned the remaining multi-reads to their most likely bin pairs. This ensured the use of at most one bin pair for each multi-read pair. We repeated our computational evaluations by varying this threshold on the posterior probabilities to ensure robustness of the overall conclusions to this threshold.

Software availability mHi-C pipeline is implemented in Python and accelerated by C. The source codes and instructions for running mHi-C are publicly available at `https://github.com/keleslab/mHiC`. Each step is organized into independent script with flexible user-defined parameters and implementation options. Therefore, analysis can be carried out from any step of the work-flow and easily fits in high performance computing environments for parallel computations.

Assessing false positive rates for significant contacts and TADs identification under the Uni- and Uni&Multi-settings To quantify false positive rates of the Uni- and Uni&Multi-settings, at the significant contact level, we defined true positives and true negatives by leveraging deeply sequenced replicates of IMR90 dataset (replicates 1-4). Significant contacts reproducibly identified across all four replicates at 0.1% FDR by both the Uni- and Uni&Multi-settings were labeled as true positives (i.e., true interactions). True negatives were defined as all the contacts that were not deemed significant at 25% FDR in any of the four replicates. We then evaluated significant contacts identified by smaller depth replicates 5 & 6 with ROC and PR curves (Supplementary Figs. 12) by using these sets of true positives and negatives as the gold standard. To quantify false positive rates at the topologically associating domains (TADs) level (Fig. 2h), we utilized TADs that are reproducible in more than three replicates of IMR90 dataset and/or harbor CTCF peaks at the boundaries as true positives. The rest of the TADs are supported neither by multiple replicates nor by CTCF, hence are regarded as false positives.

Model-free multi-reads allocation strategies The simplified and intuitive strategies depicted in Fig. 1g correspond to rescuing multi-reads at different essential stages of the Hi-C analysis pipeline. AlignerSelect relies on the base aligner, e.g., BWA, to determine the primary align-

ment of each individual end of a multi-read pair. DistanceSelect enables the distance prior to dominate. It selects the read end alignments closest in the genomic distance as the origins of the multi-read pair and defaults to the primary alignment selected by base aligner for inter-chromosomal multi-reads. Finally, SimpleSelect follows the overall mHi-C pipeline closely by making use of the standard Hi-C validation checking and binning procedures. For the reads that align to multiple bins, it selects the bin pair closest in the genomic distance as the allocation of the multi-read pair. Bin-pair allocations for inter-chromosomal multi-reads are set randomly in this strategy.

We designed SimpleSelect as a model-free alternative approach to mHi-C. When we evaluated SimpleSelect assignments, we observed that a large fraction of these were assigned low posterior probabilities by the mHi-C model that leverages both the distance prior and the local contact counts (Supplementary Fig. 32). A direct comparison of the mHi-C allocation probability of each multi-read assignment from the two approaches exhibits agreement along the diagonal, indicating that mHi-C is able to capture contact distance impact similar to SimpleSelect for 31 to 41% of the multi-reads and, in addition, highlights leveraging of local signals by mHi-C to designate the rest of the multi-reads to more probable regions in the off-diagonal triangle (Supplementary Fig. 33).

Evaluating reproducibility Reproducibility in contact matrices was evaluated using HiCRep in the default settings. We further assessed the reproducibility in terms of identified contacts by grouping the significant contacts into three categories: contacts only detected under Uni-setting, contacts unique to Uni&Multi-setting, and contacts that are detected under both settings. The reproducibility is calculated by overlapping significant contacts between every two replicates and recording the percentage of interactions that are also deemed significant in another replicate (Supplementary Fig. 16).

Chromatin states of novel significant contacts We annotated the novel significant contacts with the 15 states ChromHMM segmentations for IMR90 epigenome (ID E017) from the Roadmap Epigenomics project[28]. All six replicates of IMR90 are merged together in calculating the average enrichment of significant contacts among the 15 states (Fig. 2b).

ChIP-seq analysis ChIP-seq peak sets for IMR90 cells were obtained from ENCODE portal (`https://www.encodeproject.org/`) and GEO[43]. Specifically, we utilized H3K4me1 (ENCSR831JSP), H3K4me3 (ENCSR087PFU), H3K36me3 (ENCSR437ORF), H3K27ac (ENCSR002YRE), H3K27me3 (ENCSR431UUY) and CTCF (ENCSR000EFI) from the ENCODE project and p65 (GSM1055810), p300 (GSM1055812) and PolII (GSM1055822) from GEO[43]. In addition, raw data files in fastq format were processed by Permseq[21] utilizing DNase-seq of IMR90 (ENCODE accession ENCSR477RTP) to incorporate multi-reads and, subsequently, peaks were identified using ENCODE uniform ChIP-seq data processing pipeline (`https://www.encodeproject.org/pages/pipelines/#DNA-binding`). CTCF motif quantification for topologically associating domains was carried out with FIMO[47] under the default settings using CTCF motif frequency matrix from JASPAR[48]

Promoters with significant contacts Significant contacts across six replicates of the IMR90 study were annotated with GENCODE V19[41] gene annotations and enhancer regions from ChromHMM. Gene expression calculations utilized RNA-seq quantification results from the ENCODE project with accession number ENCSR424FAZ.

Marginal Hi-C tracks in Supplementary Figs. 18-20 Uni-setting and Uni&Multi-setting Hi-C tracks displayed on the UCSC genome browser figures (Supplementary Figs. 18-20) are obtained by aggregating contact counts of six replicates of IMR90 for each genomic coordinate along the genome.

Visualization of contact matrices and interactions We utilized Juicebox[34], HiGlass[49] and WashU epigenome browser[32] for depicting contact matrices and interactions, respectively, throughout the paper.