# Statistical Methods for Profiling 3-Dimensional Chromatin Interactions from Repetitive Regions of Genomes

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

***For correspondence:**
keles@stat.wisc.edu (SK)

[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

**Abstract** 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 interactions across biological replicates. The impact of the multi-reads on the identification of novel significant interactions is influenced marginally by relative contribution of multi-reads to the sequencing depth compared to uni-reads, cis-to-trans ratio of contacts, and the broad data quality as reflected by the proportion of mappable reads of datasets. Computational experiments highlighted that in Hi-C studies with short read lengths, mHi-C rescued multi-reads can emulate the effect of longer reads. 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.

## Introduction

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 (***Dekker et al., 2002***; ***de Laat and Duboule, 2013***) and has far-reaching influence on gene regulation. Changes in DNA folding and chromatin structure remodeling may result in cell malfunction with devastating consequences (***Corradin et al., 2016***; ***Won et al., 2016***; ***Javierre et al., 2016***; ***Rosa-Garrido et al., 2017***; ***Spielmann et al., 2018***). Hi-C technique (***Lieberman-Aiden et al., 2009***; ***Rao et al., 2014***) 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 (***Heinz et al., 2010***; ***Hwang et al., 2014***; ***Ay et al., 2014a***; ***Servant et al., 2015***; ***Mifsud et al., 2015***;

***Lun and Smyth, 2015***), and their operating characteristics were recently studied (***Ay and Noble, 2015***; ***Forcato et al., 2017***; ***Yardimci et al., 2017***) 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 (***Xie et al., 2013***). Such reads from repetitive regions can be aligned to multiple positions (Figure 1A) 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) (***Li and Dewey, 2011***), chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) (***Chung et al., 2011***; ***Zeng et al., 2015***), as well as genome-wide mapping of protein-RNA binding sites (CLIP-seq or RIP-seq) (***Zhang and Xing, 2017***). More recently, ***Sun et al.*** (***2018***) and ***Cournac et al.*** (***2015***) argued for a fundamental of repeat elements in the 3D folding of genomes, associating with server repeat expansion disorders. However, the ambiguity of multi-reads alignment renders it a challenge to investigate the repetitive elements co-localization with the true 3D interaction architecture and signals. In this work, we developed mHi-C (Figure 1–Figure supplements 1 and 2), 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 interactions at a given resolution. We evaluated mHi-C both with leveraging replicate structure of a series of Hi-C datasets of different species and tissues, and also with computational trimming and data-driven simulation experiments.

## Results

### Multi-reads significant increase the sequencing depths of Hi-C data

For developing mHi-C and studying its operating characteristics, we utilized six published studies, resulting in seven datasets with multiple replicates, as summarized in Table 1 with more details in Figure 1-source data 1: Table 1. These datasets represent a variety of study designs from different organisms, i.e., human and mouse datasets as examples of large genomes and a *Plasmodium Falciparum* dataset as a representative of small genomes. Specifically, they span a wide range of sequencing depths (Figure 1B), coverages and cis-to-trans ratios (Figure 1–Figure supplement 3), and have different proportions of mappable and valid reads (Figure 1–Figure supplement 4). Before applying mHi-C model to these datasets and investigating biological implications, we first established the substantial contribution of multi-reads to the sequencing depth across these datasets with diverse characteristics. At read-end level (Appendix 1 Table 2 for terminology), after the initial step of aligning to the reference genome (Figure 1–Figure supplements 1 and 2), multi-reads constitute approximately 10% of all the mappable read ends (Figure 1–Figure supplement 5A). Moreover, the contribution of multi-reads to the set of aligned reads further increases by an additional 8% when chimeric reads (Appendix 1 Table 2) are also taken into account with the actual number in Figure 1–source data 1: Table 2. Most notably, Figure 1–Figure supplement 6A demonstrates that, in datasets with shorter read lengths, multi-reads constitute a larger percentage of usable reads compared to uniquely mapping chimeric reads that are routinely rescued in Hi-C analysis pipelines (***Servant et al., 2015***; ***Lun and Smyth, 2015***). Moreover, multi-reads also make up a significant proportion of the rescued chimeric reads (Figure 1–Figure supplement 6B). At the read pair level, after joining of both ends, multi-reads increase the sequencing depth by 18% to 23% for shorter read length datasets and 10% to 15% for longer read lengths, thereby providing a substantial increase to the depth before the read pairs are further processed into bin pairs (Figure. 1C; Figure 1–source data 1: Table 3).

**Table 1.** Hi-C Data Summary

| Cell line | Replicate | Read length (bp) | Restriction Enzyme | HiC Protocol | Source | Resolution (kb) |
|---|---|---|---|---|---|---|
| IMR90 | rep1-6 | 36 | HindIII | dilution | ***Jin et al.*** (***2013***) | 40 |
| GM12878 | rep2-9 | 101 | MboI | in situ | ***Rao et al.*** (***2014***) | 5, 10*, 40* |
| GM12878 | rep32, rep33 | 101 | DpnII | in situ | ***Rao et al.*** (***2014***) | 5 |
| A549 | rep1-4 | 151 | MboI | in situ | ***Consortium et al.*** (***2012***)** | 10, 40 |
| ESC(2012) | rep1, rep2 | 36 | HindIII | dilution | ***Dixon et al.*** (***2012***) | 40 |
| ESC(2017) | rep1-4 | 50 | DpnII | in situ | ***Bonev et al.*** (***2017***) | 10, 40 |
| Cortex | rep1-4 | 50 | DpnII | in situ | ***Bonev et al.*** (***2017***) | 10, 40 |
| P.falciparum | 3 stages | 40 | MboI | dilution | ***Ay et al.*** (***2014b***) | 10, 40 |

* Replicates 2, 3, 4 and 6 of the GM12878 cell line datasets were process at 10kb and 40kb resolution.

** Data is downloaded from the ENCODE portal (https://www.encodeproject.org/).

## Multi-reads can be rescued at multiple processing stages of mHi-C pipeline

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 (Figure 1–Figure supplements 1 and 2, Appendix 1 Table 2). mHi-C pipeline tracks multi-reads through these processing steps. Remarkably, in the application of mHi-C to all six studies, 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 (Figure 1C; Appendix 1 Table 2 and see Materials and Methods). Collectively, mHi-C can rescue as high as $6.7\%$ more valid read pairs (Figure 1D) that originate from multi-reads and are mapped unambiguously without carrying out any multi-reads specific procedure for large genomes and $10.4\%$ for *P. Falciparum*. Such improvement corresponds to millions of reads for deeper sequenced datasets (Figure 1–source data 1: Table 4). For the remaining multi-reads (Figure 1D, colored in pink), which take up 18% of all the valid reads (Figure 1–Figure supplement 5B), mHi-C implements a novel multi-mapping model and probabilistically allocates them.

mHi-C generative model (Figure 1E and see Materials and 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, or fixed number of restriction fragments derived from the Hi-C protocol. The resolutions at which seven cell lines are processed are summarized in Table 1. In the mHi-C generative model, 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 pairs $j$ 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 bin pairs $j$ 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 (***Imakaev et al., 2012***; ***Knight and Ruiz, 2013***; ***Ay et al., 2014a***).

## Probabilistic assignment of multi-reads leads to more complete contact matrices and improves reproducibility across replicates

Before quantifying mHi-C model performance, we first provide direct visual comparison of the contact matrices between Uni-setting and Uni&Multi-setting using raw contact counts and normalized contact counts. We utilize Knight-Ruiz Matrix Balancing normalization (***Knight and Ruiz, 2013***) method directly provided by Juicebox (***Durand et al., 2016***). Figure 2A and Figure 2–Figure supplements 1-6 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 (Figure 2–Figure supplement 7). For example, for the combined replicates of GM12878, 99.61% of the 5kb bins are covered by at least 100 contacts under the Uni&Multi-setting, compared to 98.72% under Uni-setting, thereby allowing us to study 25.55Mb more of the genome. These major improvements in coverage provide direct evidence that mHi-C is rescuing multi-reads that originate from biologically valid fragments.

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 (***Yang et al., 2017***), a recent state-of-the-art approach for evaluating the reproducibility of Hi-C contact matrices. Figure 2B and Figure 2–Figure supplement 8, 9 illustrate that integrating multi-reads leads to increased reproducibility and reduced variability of stratum-adjusted correlation coefficients among biological replicates across all the study datasets. Furthermore, we observe that, for some chromosomes, e.g., chr17 of IMR90 and chr16 of GM12878, the improvement in reproducibility stands out, without a systematic behavior across datasets. A close examination of improvement in reproducibility as a function of ratio of rescued multi-reads to uni-reads across chromosomes highlights the larger proportion of multi-reads rescued for these chromosomes (Figure 2C).

In addition to direct comparison of the raw contact matrices and their reproducibility, we identified the set of significant interactions by Fit-Hi-C (***Ay et al., 2014a***) and assessed the reproducibility of the identified contacts. Specifically, we considered the degree of overlap among the significant interactions identified at 5% and 10% FDR across replicates (Figure 2–Figure supplement 10A). We observed that significant interactions 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 interactions based on distance and reassessed the reproducibility as a function of genomic distance between the contacts (Figure 2–Figure supplement 10B). Notably, significant interactions 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.

## Multi-reads lead to detection of novel significant interactions

At 5% false discovery rate (FDR), mHi-C detects 20% - 50% more novel significant interactions for relatively highly sequenced study datasets (Figure 3A–Figure supplement 1). The gains are markedly larger for datasets with smaller sequencing depths (e.g., ESC-2012) or extremely high coverage (e.g., *P. falciparum*). Overall gains in the number of novel contacts persist as the cutoff for mHi-C posterior probabilities of multi-read assignments varies (Figure 3–Figure supplement 2). At fixed FDR, significant interactions identified by the Uni&Multi-setting also include the majority of significant interactions inferred from the Uni-setting, indicating that incorporating multi-reads is extending the significant interaction list (low level of purple lines in Figure 3–Figure supplement 2).

Figure 3–Figure supplement 3 summarizes the marginal correlations of percentage change in the number of identified significant interactions (at 40Kb resolution and FDR of 0.05) with the

data characteristics commonly used to indicate the quality of Hi-C datasets (excluding the high coverage *P. Falciparum* dataset). These marginal associations highlight the significant impact of relative contribution of multi-reads to the sequencing depth compared to uni-reads and cis-to-trans ratio of contacts on the effect of mHi-C rescued multi-reads on detection of significant interactions. In addition to these marginal associations, Figure 3B and Figure 3–Figure supplement 4 display the percentage increase in the number of identified significant interactions as a function of the percentage increase in the read depth due to multi-reads and the overall sequencing depth, including both the uni- and multi-reads, across all the study datasets. A consistent pattern highlights that short read datasets with large proportion of mHi-C rescued multi-reads compared to uni-reads enjoy a larger increase in the number of identified significant interactions regardless of the FDR threhold. Figure 3–Figure supplement 5 displays the increase in the number of novel significant interactions for the GM12878 datasets in more detail across a set of FDR thresholds at different resolutions and with two types of restriction enzymes. Specifically, Figure 3–Figure supplement 5C illustrates a clear negative association between the sequencing depth and the percent improvement in the number of identified significant interactions at 5kb resolution due to the larger impact of multi-reads on the replicates of smaller size. Additionally, we note that *P. falciparum* datasets tend to exhibit significantly higher gains in the number of identified contacts epecially under stringent FDR thresholds (Figure 3–Figure supplement 5). This is largely due to the ultra high coverage of these datasets (Figure 3–Figure supplement 6). Despite the impact of coverage, cis-to-trans ratio, commonly utilized as an indicator of raw data quality, tends to exhibit a positive trend with the increase in the number of significant interactions (Figure 3–Figure supplement 7).

In general, harnessing the multi-reads with posterior probability larger than 0.5 and controlling the FDR at 1%, Fit-Hi-C identifies 32.49% more significant interactions compared to Uni-setting (comparing dark green to dark purple bar in Figure 3C) and 56.82% of these are unique to Uni&Multi-setting (light green bar in Figure 3C) collectively for all the four replicates of GM12878 at 40kb resolution. We next asked whether these novel contacts could have been identified under the Uni-setting by employing a more liberal FDR threshold. We observed that 34.89% of the newly detected significant interactions at 1% FDR (i.e., 12.71% compared to the all the significant interactions under Uni-setting) cannot be recovered even by a more liberal significant interaction list under Uni-setting at 10% FDR (yellow bar over light green bar in Figure 3C). Conversely, Uni&Multi-setting is unable to recover only 4.60% of the Uni-setting contacts once the FDR is controlled at 10% for the Uni&Multi-setting (light blue over dark purple bar in Figure 3), highlighting again that Uni&Multi-setting predominantly adds on novel significant interactions while retaining contacts that are identifiable under the Uni-setting. Similar analysis for individual replicates of IMR90 are provided in Figure 3–Figure supplement 8 as well as those of GM12878 at individual replicate level or collective analysis at 5kb, 10kb, and 40kb resolutions in Figure 3–Figure supplement 9-12 .

We further confirmed this consistent power gain by a Receiver Operating Characteristic (ROC) and a Precision-Recall (PR) analysis. Figure 3–Figure supplement 13B 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 (Figure 3–Figure supplement 13A).

## Chromatin features of novel significant interactions

To further establish the biological implications of mHi-C rescued multi-reads, we investigated genomic features of novel contacts. Annotation of the significant interactions with ChromHMM segmentations from the Roadmap Epigenomics project (***Roadmap Epigenomics Consortium, 2015***) highlights marked enrichment of significant interactions in annotations involving repetitive DNA (Figure 3D, Figure 3–Figure supplement 14A). Most notably, ZNF genes & repeats and Heterochromatin states exhibit the largest discrepancy of the average significant interaction 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 interaction enrichment of genomic regions harboring histone marks and other biochemical signals (***The ENCODE Project Consortium,***

***2012***; ***Jin et al., 2013***) (See Materials and Methods) by comparing their average contact counts to those without such signal (Figure 3E and Figure 3 – source data 1). Notably, while we observe that multi-reads boost average number of significant interactions for biochemically active regions of the genome, they contribute more to regions that harbor H3K27me3 peaks (Figure 3E, Figure 3–Figure supplement 14B). Such regions are associated with downregulation of nearby genes through forming heterochromatin structure (***Ferrari et al., 2014***). Figure 3–Figure supplements 15-17 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. Many genes of biological significance reside in these regions. For example, NBPF1 (Figure 3–Figure supplement 15) is implicated in many neurogenetic diseases and its family consists of dozens of recently duplicated genes primarily located in segmental duplications (***Safran et al., 2010***). In addition, RABL2A within the highlighted region of Figure 3–Figure supplement 16 is member of RAS oncogene family.

## Multi-reads lead to novel promoter-enhancer interactions

We found that a significant impact of multi-reads is on the detection of promoter-enhancer interactions. Overall, mHi-C identifies 14.89% more significant promoter-enhancer contacts at 5% FDR collectively for six replicates for IMR90 (Figure 4–source data 1: Table 1 and Figure 4 – source data 2). Among them, 13,313 promoter-enhancer contacts are reproducible among all six replicates under Uni&Multi-setting (Figure 4–source data 1: Table 2) and 62,971 aremreproducible for at least two replicates (Figure 4–source data 1: Table 3) leading to 15.84% more novel promoter-enhancer interactions specific to Uni&Multi-setting. Figure 4A provides WashU epigenome browser (***Zhou et al., 2011***) display of such novel highly reproducible promoter-enhancer interactions on chromosome 1. Figure 4–Figure supplement 1-3 provides more examples of reproducible promoter-enhancer significant interactions on other chromosomes and Figure 4–Figure supplement 4 depicts the reproducibility of these interactions in more details across the 6 replicates.

We next validated the novel promoter-enhancer interactions by investigating the expression levels of the genes contributing promoters to these interactions. Figure 4B supports that genes with significant interactions in their promoters generally exhibit higher expression levels (comparing bars 1-5 to bars 8-9 in Figure 4B). 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 Figure 4B). Most remarkably, newly detected significant promoter-enhancer contacts (bar 13 in Figure 4B) 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 (Figure 4–Figure supplement 5). In contrast Uni-setting specific contacts have elevated false positive rates.

## Multi-reads refine the boundary detection of topologically associating domains

We also investigated the impact of mHi-C rescued multi-reads on the topologically associating domains (TADs) (***Pombo and Dillon, 2015***). We used the DomainCaller (***Dixon et al., 2012***, ***2015***) to infer TADs of IMR90 at 40kb and Arrowhead (***Rao et al., 2014***) for that of GM12878 at 5kb under Uni&Multi-settings (Figure 5 – source data 1, 2). The detected TADs are compared to that under Uni-setting provided by the original publication. By doing that we found Uni&Multi-setting identifies similar number of TADs as Uni-setting (Figure 5–Figure supplement 1). 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 (***Dixon et al., 2012***, ***2015***; ***Dekker and Heard, 2015***), Uni&Multi-setting identifies 2.01% more reproducible TADs across replicates with 2.36% lower non-reproducible TADs (Figure 5A). Several studies have revealed the role of CTCF in establishing the boundaries of genome architecture (***Ong and Corces, 2014***; ***Tang et al., 2015***;

***Hsu et al., 2017***). 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. (Figure 5–Figure supplements 2A-C). Figure 5B 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 (Figure 5–Figure supplements 3) by filling in the gaps in the contact count matrix and boosting the domain signals, mHi-C also refines TAD boundaries (Figure 5–Figure supplements 4 and 5) and eliminates potential false positive TADs that are split by the white gaps in Uni-setting (Figure 5–Figure supplements 6-8). The novel, adjusted, and eliminated TADs are largely supported by CTCF signal identified using both uni- and multi-reads ChIP-seq datasets (***Zeng et al., 2015***) as well as convergent CTCF motifs (Figure 5–Figure supplement 2D), providing evidence for mHi-C driven modifications to these TADs and revealing a lower false discovery rate for mHi-C compared to Uni-setting (Figure 5C and Figure 5–Figure supplement 2E and Figure supplement 9).

Next, we assessed the abundance of different classes of repetitive elements at the reproducible TAD boundaries, which are obtained from the RepeatMasker (***Open, 2015***) and UCSC genome brower (***Casper et al., 2017***) using hg19 assembly. The type of repetitive elements under study are segmental duplications (DUP), short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), long terminal repeat elements (LTR), DNA transposon (DNA) and satellites (SATE). We consider $\pm 1$ bin on either side of the edge coordinate of the topolocial domain as one TAD boundary. At a lower resolution, i.e., 40kb for IMR90, each boundary will be a 120kb region and the percentage of TAD boundaries co-localizing with each type of repetitive elements illustrate negligible difference between Uni-setting and Uni&Multi-setting (Figure 5–Figure supplement 10A). Similarly, due to the large bin sizes at the TAD boundaries, majority of TAD boundaries harbor SINE, LINE, LTR and DNA transposon elements. However, when at the higher resolution of 5kb for the GM12878 dataset, SINE is the leading category of elements that co-localizes with more than 99% of TAD boundaries followed by LINEs (Figure 5–Figure supplement 10B). The observation is consistent with fact that SINE and LINE elements are relatively short and cover a larger range of the human genome compared to other subfamilies (15% for SINE and 21% for LINE) (***Treangen and Salzberg, 2012***). Additionally, we evaluated the repetitive elements enrichment by the average number of repetitive elements co-localized at each the TAD boundary. Such enrichment is compared with that within TAD to investigate the contribution of repetitive elements in forming the chromatin domain. The overal average number of repetitive elements expected at the same length genome region is also included as enrichment baseline. Figure 5–Figure supplement 11 demonstrates that SINE, satellites and segmental duplications are significantly more enriched at the TAD boundary. Furthermore, at high resolution, i.e., 5kb for GM12878, the SINE category also has the highest average enrichment and mHi-C enhanced the average number of SINE elements at each TAD boundary (Figure 5–Figure supplement 11B). However, when the resolution is low, there is also no significant change in the average number of repetitive elements across all the overlapped boundaries (Figure 5–Figure supplement 11A). In summary, under uni&Multi-setting, the detected TAD boundaries tend to harbor more SINE elements. This observation is supported by existing findings that human genome folding is remarkably associated with the SINE family ***Cournac et al.*** (***2015***).

## Large-scale evaluation of mHi-C with computational experiments

Before further investigating the accuracy of mHi-C rescued multi-reads with computational experiments, we considered intuitive heuristic strategies for rescuing multi-reads at different stages of the Hi-C analysis pipeline as alternatives to mHi-C (Figure 6A and see Materials and 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).

## Evaluation with trimming experiments establishes high accuracy of mHi-C

We implemented a trimming strategy on the long read dataset A549 with 151bp read length to evaluate the allocation accuracy of mHi-C and compare it with both a random allocation strategy as baseline and the additional heurustic approaches we developed (Figure 6A). We trimmed the uni-reads of replicate 2 of A549, which has the median sequencing depth across all available four replicates, into read lengths of 36bp, 50bp, 75bp, 100bp, and 125bp, respectively. As a result of this process, a subset of uni-reads at the full read length of 151bp with known alignment positions became multi-reads, generating a gold-standard multi-read set with known true origins. The resulting numbers of valid multi-reads are summarized in Figure 6–Figure supplement 1 in comparison with the numbers of valid multi-reads in the original A549 datasets. Figure 6–Figure supplement 1 exhibits a dramatic decay compared to the number of original uni-reads. This discrepancy is attributed to skipping the chimeric read rescue in these trimming experiments hence a large proportion of uni-reads are rendered unmapped or singleton reads before they can be trimmed into alignable multi-reads (Figure 6–Figure supplement 2). For evaluation with the mHi-C allocation, we considered two approaches for generating evaluation datasets where we combined the trimmed multi-reads from replicate 2 with (i) uni-reads from other replicates, i.e., replicates 1, 3 and 4, in the A549 dataset individually, and (ii) with the trimmed reads of replicate 2 that remain uniquely aligned to the reference genome at the same trimmed read length. In the first setting where the uni-read sets are of the original sequencing depth, we focus on the multi-read assignment accuracy at different resolutions, i.e., 10kb and 40kb. The latter setting enables performance assessment of the trimmed multi-reads of small sequencing depth for recovering the full length larger depth dataset, i.e., uni and multi-reads at trimmed read length compared to uni-reads at full read length.

For setting (i), we investigated the multi-reads allocation accuracy with respect to trimmed read length, sequencing depth, and mappability. Figure 6B exhibits superior performance of mHi-C over both the model-free methods and the random baseline in correctly allocating multi-reads of different lengths to their original positions across intra- and inter-chromosome interactions. Accuracy of mHi-C can reach to 90% for intra-chromosomal contacts (Figure 6–Figure supplement 3A). When the allocated multi-reads are stratified as a function of the mappability (Figure 6C), multi-reads with the lowest mappability range of 0-0.05 from the 36bp trimming setting have accuracy levels of $55.92\%$ and $60.28\%$ at resolutions 10kb and 40kb, respectively. The accuracy quickly reaches a minimum level of 80% for reads with mappability of at least $0.5$. Figure 6–Figure supplement 4 displays results for the other trimming settings. A noticeable observation is that as the total sequencing depth decreases from trim length of 36bp to trim length of 125bp, there is a slight drop in the accuracy, albeit not uniform across all ranges of the mappability at the higher resolution of 10kb. This reemphasizes the dependence of multi-read assignment on the total sequencing depth. On average, as expected and illustrated by Figure 6B, accuracy of multi-read assignment has an increasing trend with the read length. Next, we assessed the allocation accuracy among different classes of repetitive elements (Figure 6D, Figure 6–Figure supplement 5). Allocations involving segmental duplication regions exhibit a systematically poor performance across all five trimming settings at both of the 10kb and 40kb resolutions. This observation is consistent with the general characterization of of segmental duplication regions which can range from 1k to 100kb in length (***Treangen and Salzberg, 2012***). Notably, even for these regions, accuracy of mHi-C is markedly higher than the model-free approaches as well as the baseline random selection as shown in Figure 6B.

Next, we utilized trimmed uni-reads and trimmed multi-reads to recover the original uni-reads set of A549 that we used to generate trimmed datasets (setting (ii) above). mHi-C analysis of trimmed multi-reads with the resulting uni-reads of the trimmed dataset can successfully recover the original contact matrices, topological domain structures, and significant interactions. Specifically, to asses the recovery of the original contact matrix, we compared both the trimmed Uni- and

Uni&Multi-setting with the gold standard Uni-setting at the original read length with HiCRep (***Yang et al., 2017***). Figure 7 and Figure 7A-Figure supplement 1 illustrate that mHi-C achieves significant improvement in the reproducibility across all chromosomes under all trimming settings compared to Uni-setting. In addition, reproducibility of different read lengths in Figure 7A is consistent with the expectation that the longer trimmed reads yield contact matrices that are more similar to the original one. Topological associating domains (TAD) identification with these trimmed sets highlight the susceptibility of TAD boundary detection to the sequencing depth. Figure 7B displays an example where multi-reads leadsto better recovery of TAD structure of the full length dataset. Overall, this performance is attributable to accuracy of mHi-C assignments and the resulting increase in sequencing depth of the trimmed uni-read dataset for this setting (Figure 7-Figure supplement 2). In addition, this analysis reveals that the improvement can be even more dramatic for short read contact matrices (Figure 7–Figure supplement 3-6). Finally, Figure 7–Figure supplement 7 and 8 illustrate that mHi-C multi-reads rescued in trimmed dataset have a direct impact on power. At the bin-to-bin interaction level, more than 75% of the significant interactions detected under Uni&Multi-setting of the trimmed datasets are deemed significance in original full-reads significant interaction list. Conversely, for the trimmed Uni-setting, not only fewer significant are detected but also a small fraction of these appear in the original significant interaction list.

### Data-driven simulations of Hi-C interactions further confirm the allocation accuracy of mHi-C

While the trimming strategy above generated shorter read length datasets from a long read dataset, we developed a simulation strategy (Materials and Methods) to empirically simulate Hi-C reads in paired-end fastq format to better match the sequencing depth ratios of multi-reads over uni-reads observed in our study datasets. We used GM12878 replicate 6 to set the simulation parameters and simulated reads that are 36bp, 50bp, 75bp and 100bp long. We generated numbers of multi-reads comparable to that of replicate 6 (Figure 7–Figure supplement 9A) and utilized uni-reads in replicate 3 to allocate the simulated multi-reads. We next matched the multi-to-uni ratio of original replicate 3 with our simulated multi-reads by two strategies where we either down-sampled replicate 3 uni-reads, respectively for different simulated read lengths, or merged all simulated reads of varying read lengths with the original uni-reads of replicate 3. Both approaches led to similar gains and only differed in the resolution that they can be analyzed at (Figure 7C and Figure 7-Figure supplement 9B). Figure 7C displays accuracy of the allocation strategies where mHi-C exhibits a clear advantage to the model-free approaches. Figure 7D and Figure 7–Figure supplement 10 further examine mHi-C accuracy as a function of mappability, revealing an accuracy around 70% even for the shortest simulated reads (36bp) when mappability is larger than $0.5$. Such accuracy will be maintained at 80% for longer reads. Overall accuracies in the simulation experiments is lower than that of the trimming experiments due to relative lower depth of the simulated datasets.

## Discussion

Hi-C data are powerful for identifying long-range interacting loci, chromatin loops, topologically associating domains (TADs), and A/B compartments (***Lieberman-Aiden et al., 2009***; ***Yu and Ren, 2017***). Multi-mapping reads, however, are absent from the typical Hi-C analysis pipelines, leading to under-representation and under-study of three-dimensional genome organization involving repetitive regions. Consequently, downstream analysis of Hi-C data relies on the incomplete Hi-C contact matrices which have frequent and, sometimes, severe interaction gaps spanning across the whole matrix. While centromeric regions contribute to such gaps, lack of multi-reads in the analysis is a significant contributor. Our Hi-C multi-mapping strategy, mHi-C, probabilistically allocates high-quality multi-reads to their most likely positions (Figure 1) and successfully fills in the missing chunks of contact matrices (Figure 2A and Figure 2–Figure supplement 1-6). As a result, incorporating multi-reads yields remarkable increase in sequencing depth and higher bin coverage is translated into significant and consistent gains in reproducibility of the raw contact counts (Figure 2B, C) and

identifies novel significant interactions (Figure 3) and promoter-enhancer interactions (Figure 4), and domain structures (Figure 5). Our computational experiments with trimmed and simulated Hi-C reads elucidate the significant impact of multi-reads in all facets of the Hi-C data analysis and reveal that even for shortest read length of 36 bp, mHi-C accuracy exceeds 80% (85% for lower depth samples at higher resolution analysis) for regions with underlying mappability of at least 0.5 (Figure 6C). mHiC significantly outperforms a baseline random allocation strategy as well as several other model-free and intuitive multi-read allocation strategies while achieving its worst allocation accuracy of 69% for reads originating from segmental duplications (Figure 6B and D).

The default setting of mHi-C is intentionally conservative. mHi-C in the default setting rescues multi-reads that are high quality and can be allocated to a candidate alignment position with high probability of at least $0.5$. mHi-C allows relaxation of this strict filtering. Instead of keeping reads with allocation probability greater than $0.5$, these posterior allocation probabilities can be utilized as fractional contacts. We chose not pursue this approach in this work as the current downstream analysis pipelines do not accommodate such fractional contacts. Currently, mHi-C model does not take into account potential copy number variations and genome arrangements across the genome. While the model can easily be extended to take into account an estimated copy number map of the underlying sample genomes as we have done for other multi-read problems (***Zhang and Keleş, 2014***), our computational experiments with cancerous human alveolar epithelial cells A549 does not reveal any notable deterioration in mHi-C accuracy for these cells with copy number alternations.

## Materials and Methods

### 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 Figure 1–Figure supplements 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 (***Li and Durbin, 2010***); 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 (Figure 1–Figure supplement 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 (***Lun and Smyth, 2015***), which is modified from the existing Cutadapt (***Martin, 2011***) 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., $\geq 25$ 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 (Figure 1–Figure supplement 1 and Appendix 1 Table 2) 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 (Figure 1–Figure supplement 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 (***Belaghzal et al., 2017***). 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 (Appendix 1 Table 2 - 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 (Figure 1–Figure supplement 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 (Figure 1–Figure supplement 2 and Appendix 1 Table 2 - 3b).

### mHi-C generative model and parameter estimation.

mHi-C infers genomic origins of multi-reads at the bin pair level (Appendix 1 Table 2). 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 Figure 1–Figure supplements 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 (***Ay et al., 2014a***) 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 $\pi$ can be written as:

$$\begin{aligned} P(\boldsymbol{\pi}|\boldsymbol{\gamma}) &= \frac{\Gamma(\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}\gamma_{(j,k)})}{\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\Gamma(\gamma_{(j,k)})}\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\pi_{(j,k)}^{(\gamma_{(j,k)}-1)} \\ &= \frac{\Gamma(\sum_{j=1}^{M-1}\sum_{k=j+1}^{M}(Spline(j,k)\times N+1))}{\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\Gamma(Spline(j,k)\times N+1)}\prod_{j=1}^{M-1}\prod_{k=j+1}^{M}\pi_{(j,k)}^{Spline(j,k)\times N}. \end{aligned}$$

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.**

$$\begin{aligned} &P(\boldsymbol{Y_i}|Z_{i,(j,k)} = 1) = P(\boldsymbol{Y_i}|O_i) \\ &= \prod_{j=1}^{M-1}\prod_{k=j+1}^{M} P(Y_{i,(j,k)}|O_i) \\ &= \prod_{j=1}^{M-1}\prod_{k=j+1}^{M} [1(Y_{i,(j,k)} = 1, (j,k) \in S_{O_i}) + 1(Y_{i,(j,k)} \neq 1, (j,k) \notin S_{O_i})] \\ &= 1. \qquad \square \end{aligned}$$

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

$$
\begin{aligned}
P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\pi}) &= \prod_{i}^{N} P_{\pi}(\boldsymbol{Y_i}, \boldsymbol{Z_i}) \\
&= \prod_{i}^{N} \prod_{j=1}^{M-1} \prod_{k=j+1}^{M} P_{\pi}(\boldsymbol{Y_i}, Z_{i,(j,k)} = 1)^{Z_{i,(j,k)}} \\
&= \prod_{i}^{N} \prod_{j=1}^{M-1} \prod_{k=j+1}^{M} [P_{\pi}(\boldsymbol{Y_i}|Z_{i,(j,k)} = 1)\boldsymbol{\pi}_{(j,k)}]^{Z_{i,(j,k)}} \\
&= \prod_{i}^{N} \prod_{j=1}^{M-1} \prod_{k=j+1}^{M} \boldsymbol{\pi}_{(j,k)}^{Z_{i,(j,k)}}.
\end{aligned}
$$

Using the Dirichlet-Multinomial conjugacy, we derive the posterior distribution of $\pi$ as

$$
\begin{aligned}
P(\boldsymbol{\pi}|\boldsymbol{Z}) \propto P(\boldsymbol{\pi}, \boldsymbol{Z}) &= \prod_{i=1}^{N} P(\boldsymbol{Z_i}|\boldsymbol{\pi})P(\boldsymbol{\pi}) \\
&\propto \prod_{j=1}^{M-1} \prod_{k=j+1}^{M} \boldsymbol{\pi}_{(j,k)}^{(\sum_{i=1}^{N} Z_{i,(j,k)} + \gamma_{(j,k)} - 1)} \\
&= \prod_{j=1}^{M-1} \prod_{k=j+1}^{M} \boldsymbol{\pi}_{(j,k)}^{(\sum_{i=1}^{N} Z_{i,(j,k)} + Spline(j,k) \times N)}.
\end{aligned}
$$

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

**E-step.**

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

**M-step.**

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

Estimate of the contact probability $\pi_{(j,k)}$ in the M-step can be viewed as an integration of local interaction signal, encoded in $\sum_{i=1}^{N} Z_{i,(j,k)}^{(t)}$, 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 interactions 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.

## Assessing false positive rates for significant interactions 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 interaction level, we defined true positives and true negatives by leveraging deeply sequenced replicates of IMR90 dataset (replicates 1-4). Significant interactions 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 interactions identified by

smaller depth replicates 5 & 6 with ROC and PR curves (Figure 2–Figure supplements 13) 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 (Figure 6C, Figure 6–Figure supplement2E), 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.

## 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 interactions 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 interactions between every two replicates and recording the percentage of interactions that are also deemed significant in another replicate (Figure 4–Figure supplement 10).

## Chromatin states of novel significant interactions

We annotated the novel significant interactions with the 15 states ChromHMM segmentations for IMR90 epigenome (ID E017) from the Roadmap Epigenomics project (***Roadmap Epigenomics Consortium, 2015***). All six replicates of IMR90 are merged together in calculating the average enrichment of significant interactions among the 15 states (Figure 4B and Figure 4–Figure supplementary 14A).

## ChIP-seq analysis

ChIP-seq peak sets for IMR90 cells were obtained from ENCODE portal (https://www.encodeproject.org/) and GEO (***Barrett et al., 2012***). 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 (***Barrett et al., 2012***). In addition, raw data files in fastq format were processed by Permseq (***Zeng et al., 2015***) 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 (***Grant et al., 2011***) under the default settings using CTCF motif frequency matrix from JASPAR (***Khan et al., 2017***).

## Promoters with significant interactions

Significant interactions across six replicates of the IMR90 study were annotated with GENCODE V19 (***Harrow et al., 2012***) 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 Figure 4–Figure supplements 15-17

Uni-setting and Uni&Multi-setting Hi-C tracks displayed on the UCSC genome browser figures (Figure 4–Figure supplements 15-17) 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 (***Durand et al., 2016***), HiGlass (***Kerpedjiev et al., 2017***) and WashU epigenome browser (***Zhou et al., 2011***) for depicting contact matrices and interactions, respectively, throughout the paper.

## Model-free multi-reads allocation strategies

The simplified and intuitive strategies depicted in Figure 6A 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 alignment 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. 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 (Figure 6–Figure supplement 6). Figure 6–Figure supplement 7 corroborates this drawback of SimpleSelect and demonstrates that mHi-C differential patterns agree better with that of the Uni-setting. Moreover, Figure 6–Figure supplement 7 suggests that rings stage is more similar to schizonts, an observation consistent with existing findings on *P. falciparum* life stages (***Ay et al., 2014b***; ***Bunnik et al., 2018***).

## 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.

## Simulation procedures

We devised a simulation strategy that utilizes parameters learnt from the Hi-C data and results in data with similar signal to noise characteristics as the actual data.

1. *Construction of the interaction prior based on the uni-reads fragment interaction frequency list of GM12878 dataset (replicate 6).* The frequency list from the prior encompasses both the genomic distance effect and local interaction signal strength and forms the basis for simulating restriction fragment interactions.
2. *Generating the restriction enzyme cutting sites for each simulated fragment pair.* After sampling interacting fragments using the frequency list from Step 1, a genomic coordinate within $\pm$ 500bp of the restriction enzyme cutting site and the corresponding strand direction are selected randomly. Reads of different lengths (36bp, 50bp, 75bp, 100bp) are generated starting from these cutting sites.
3. *Mutating the resulting reads.* Mutation and gap rates are empirically estimated based on the aligned uni-reads of replicate 6. The reads from Step 2 are uniformly mutated with these rates allowing up to 2 mutations and 1 gap.
4. *Simulate sequence quality scores of the reads.* We utilize the empirical estimation of the distribution regarding to the sequence base quality scores across individual locations of the read length and simulate for each read its sequence quality scores at the nucleotide level.

5. *Alignment to the reference genome.* The simulated reads are aligned to the reference genome and filtered for validation as we outline in the mHi-C pipeline, resulting in the set of multi-reads that are utilized by mHi-C.

As final step of the simulation studies, we merge the simulated set of multi-reads with rep3&6 uni-reads and ran mHi-C step4 (binning) - step5 (prior already available) - step6 (assign multi-reads posterior probability) independently at resolutions 10kb and 40kb.

## 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. We thank the peer reviewers and the Reviewing and Senior *eLife* Editors of this work for their constructive comments. Ye Zheng would specificially like to thank Peigen Zhou from University of Wisconsin – Madison for his insightful discussions on accelerating the pipeline and to take this precious opportunity to formally ask Peigen: **Will you marry me**?

## 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.

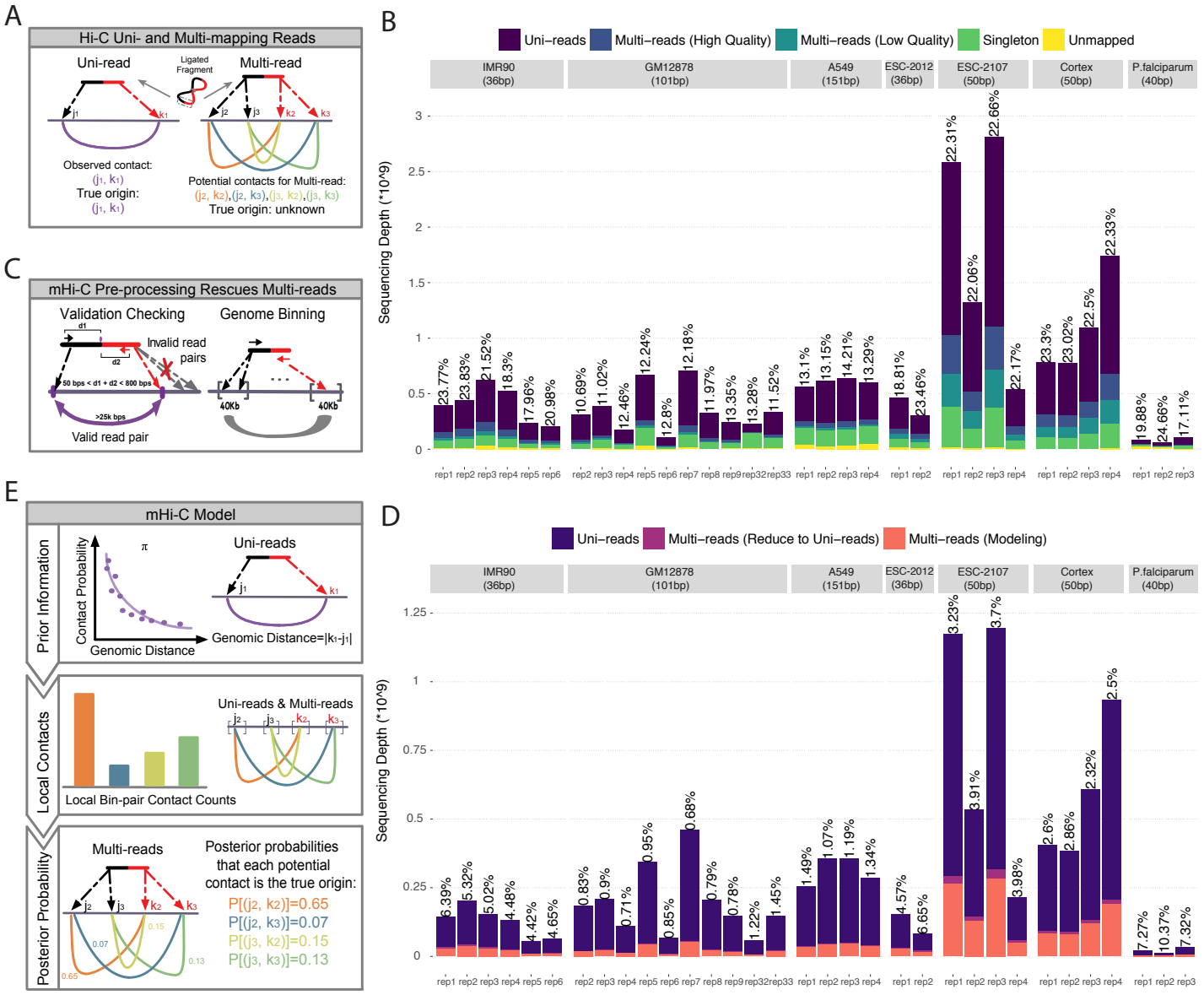



**Figure 1.** Overview of multi-reads and mHi-C pipeline. **(A)** Standard Hi-C pipelines utilize uni-reads while discarding multi-mapping reads which give rise to multiple potential contacts. **(B)** Total number of reads in different categories as a result of alignment to reference genome across the study datasets. Percentages of high-quality multi-reads compared to uni-reads are depicted on top of each bar. **(C)** Multi-mapping reads can be reduced to uni-reads within validation checking and genome binning pre-processing steps. **(D)** 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. **(E)** 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.

**Figure 1–source data 1.** Detailed summary of study datasets.
**Figure 1–Figure supplement 1.** mHi-C pipeline (Alignment - Read end pairing - Valid fragment filtering).
**Figure 1–Figure supplement 2.** mHi-C pipeline (Duplicate removal - Genome binning - mHi-C).
**Figure 1–Figure supplement 3.** Coverage and cis-to-trans ratios across individual replicates of the study datasets as indicators of data quality.
**Figure 1–Figure supplement 4.** Percentages of mappable and valid reads across study datasets as an indicator of data quality.
**Figure 1–Figure supplement 5.** Categorization of reads after alignment across study datasets.
**Figure 1–Figure supplement 6.** Characterization of the abundance of multi-reads due to chimeric reads.

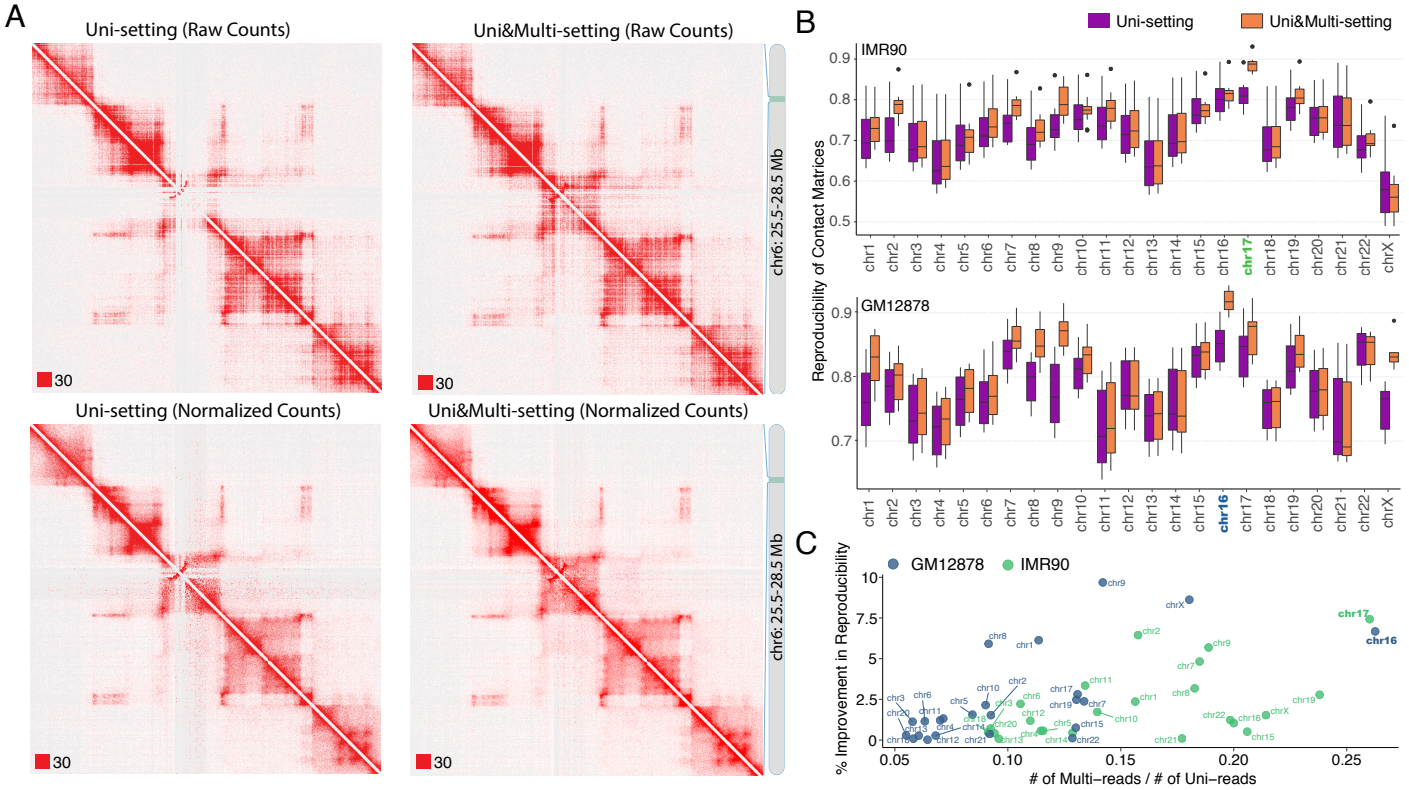



**Figure 2.** Global impact of multi-reads in Hi-C analysis. **(A)** Contact matrices of GM12878 with combined reads from replicates 2-9 are compared under Uni-setting and Uni&Multi-setting using raw and normalized (Knight-Ruiz Matrix Balancing algorithm (***Knight and Ruiz, 2013***)) contact counts for chr6:25.5Mb - 28.5Mb. 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. Such gaps remain in the Uni-setting even after normalization. Red squares at the left bottom of the matrices indicate the color scale. **(B)** Reproducibility of Hi-C contact matrices by HiCRep across all pairwise comparisons between replicates under the Uni- and Uni&Multi-settings (IMR90 and GM12878 are displayed). **(C)** Percent improvement in reproducibility due to the Uni&Multi-setting versus the proportion of number of valid multi-reads compared to numer of the uni-reads in the datasets.

**Figure 2–Figure supplement 1.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting fon chromosome 1.
**Figure 2–Figure supplement 2.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 2.
**Figure 2–Figure supplement 3.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 3.
**Figure 2–Figure supplement 4.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 4.
**Figure 2–Figure supplement 5.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 5.
**Figure 2–Figure supplement 6.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 10.
**Figure 2–Figure supplement 7.** Bin coverage improvement of Uni&Multi-setting compared to Uni-setting for IMR90 at the individual replicate level for two different allocation probability thresholds.
**Figure 2–Figure supplement 8.** Reproducibility at the contact matrix level under the Uni- and Uni&Multi-settings across study datasets.
**Figure 2–Figure supplement 9.** Reproducibility at the contact matrix level at resolutions 40kb (low) and 10kb (high) across study datasets.
**Figure 2–Figure supplement 10.** Reproducibility of significant interactions for IMR90.

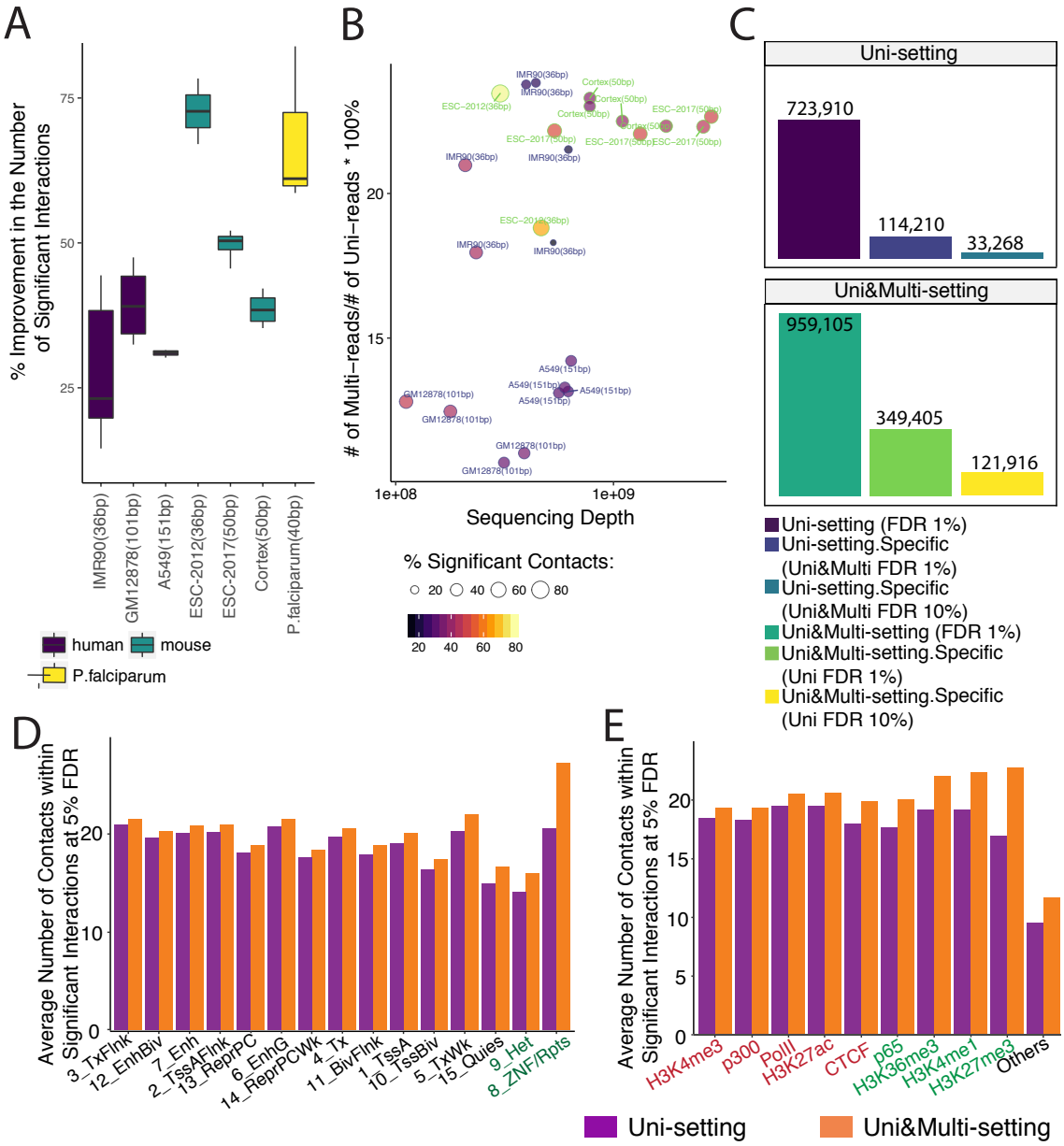

A
% Improvement in the Number of Significant Interactions
IMR90(36bp)
GM12878(101bp)
A549(151bp)
ESC-2012(36bp)
ESC-2017(50bp)
Cortex(50bp)
P.falciparum(40bp)
human
mouse
P.falciparum
B
# of Multi-reads/# of Uni-reads * 100%
Sequencing Depth
% Significant Contacts:
C
Uni-setting
723,910
114,210
33,268
Uni&Multi-setting
959,105
349,405
121,916
Uni-setting (FDR 1%)
Uni-setting.Specific (Uni&Multi FDR 1%)
Uni-setting.Specific (Uni&Multi FDR 10%)
Uni&Multi-setting (FDR 1%)
Uni&Multi-setting.Specific (Uni FDR 1%)
Uni&Multi-setting.Specific (Uni FDR 10%)
D
Average Number of Contacts within Significant Interactions at 5% FDR
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
E
Average Number of Contacts within Significant Interactions at 5% FDR
H3K4me3
p300
PolII
H3K27ac
CTCF
p65
H3K36me3
H3K4me1
H3K27me3
Others
Uni-setting
Uni&Multi-setting

**Figure 3.** Gain in the numbers of novel significant interactions by mHi-C and their characterization by chromatin marks. **(A)** Percentage increase in detected significant interactions (FDR 5%) by comparing contacts identified in Uni&Multi-setting with those of Uni-setting across study datasets at 40kb resolution. **(B)** Percentage change in the numbers of significant interactions (FDR 5%) as a function of overall sequencing depth and the percentage of mHi-C rescued multi-reads in comparison to uni-read sets of individual datasets at 40kb resolution. **(C)** Recovery of significant interactions identified at 1% FDR by analysis at 10% FDR, aggregated over the replicates of GM12878 at 40kb resolution. Detailed descriptions of the groups are provided in Figure 3–Figure supplement 8. **(D)** Average number of contacts falling within the significant interactions (5% FDR) that overlapped with each chromHMM annotation category across six replicates of IMR90 identified by Uni- and Uni&Multi-settings . **(E)** Average number of contacts (5% FDR) that overlapped with significant interactions and different types of ChIP-seq peaks associated with different genomic functions (IMR90 six replicates). 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.

**Figure 3–source data 1.** ChIP-seq peaks detected following the standard ChIP-seq data processing pipeline of ENCODE (***The ENCODE Project Consortium, 2012***) using both uni-reads and multi-reads aligned by Permseq (***Zeng et al., 2015***) for CTCF, PolII, p65, p300, H3K27ac, H3K27me, H3K36me3, H3K4me1 and H3K4me3.
**Figure 3–Figure supplement 1.** Percentage change in the numbers of significant interactions under the Uni&Multi-setting compared to Uni-setting at different FDR thresholds and resolutions.
**Figure 3–Figure supplement 2.** Comparison of significant interactions as a function of posterior probabilities of multi-read assignment (IMR90).
**Figure 3–Figure supplement 3.** Heatmap for marginal correlations of percentage increase in the number of identified significant interactions (FDR 5%) with indicators of data quality.
**Figure 3–Figure supplement 4.** Percentage change in the numbers of significant interactions as a function of overall sequencing depth and the percentage of mHi-C rescued multi-reads in comparison to uni-read sets of individual datasets.
**Figure 3–Figure supplement 5.** Percentage change in the numbers of significant interactions of GM12878 datasets at different resolutions.
**Figure 3–Figure supplement 6.** Percentage change in the numbers of significant interactions with respect to coverage.
**Figure 3–Figure supplement 7.** Percentage change in the numbers of significant interactions with respect to cis-to-trans ratio.
**Figure 3–Figure supplement 8.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of six replicates of IMR90.
**Figure 3–Figure supplement 9.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of four replicates of GM12878 at 40kb resolution.
**Figure 3–Figure supplement 10.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of four replicates of GM12878 at 10kb resolution.
**Figure 3–Figure supplement 11.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of ten replicates of GM12878 at 5kb resolution.
**Figure 3–Figure supplement 12.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for GM12878 summed across replicates at 5kb, 10kb, and 40kb resolutions.
**Figure 3–Figure supplement 13.** ROC and PR curves for replicates 5 and 6 of IMR90.
**Figure 3–Figure supplement 14.** Quantification of significant interactions for chromHMM states and ChIP-seq peak regions (IMR90).
**Figure 3–Figure supplement 15.** Marginalized Hi-C signal, ChIP-seq coverage and peaks and gene expression for chr1 (IMR90).
**Figure 3–Figure supplement 16.** Marginalized Hi-C signal, ChIP-seq coverage and peaks and gene expression for chr2 (IMR90).
**Figure 3–Figure supplement 17.** Marginalized Hi-C signal, ChIP-seq coverage and peaks and gene expression for chr9 (IMR90).

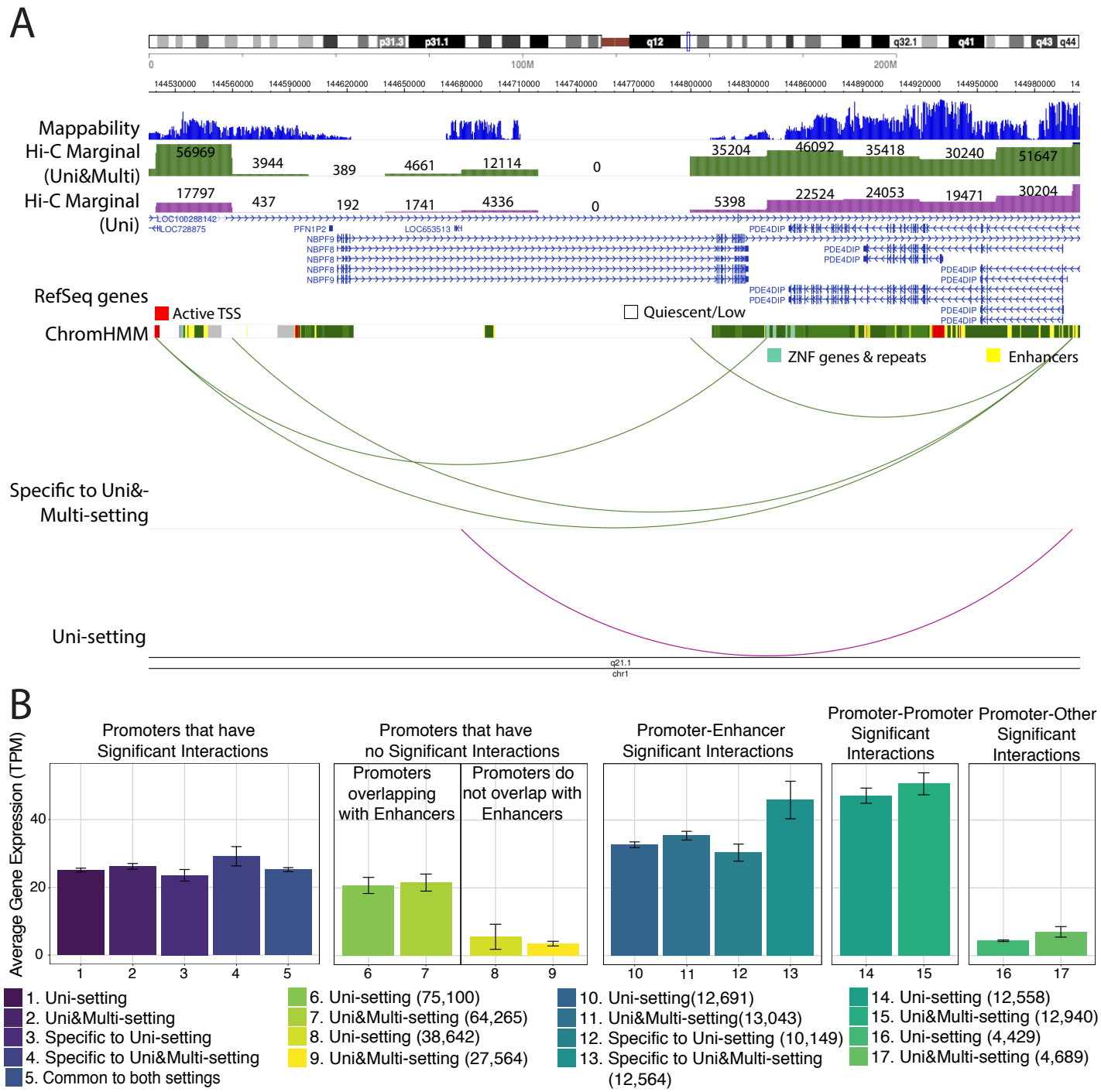

A
Mappability
Hi-C Marginal (Uni&Multi)
Hi-C Marginal (Uni)
RefSeq genes
ChromHMM
Active TSS
Quiescent/Low
ZNF genes & repeats
Enhancers
Specific to Uni&-Multi-setting
Uni-setting
B
Average Gene Expression (TPM)
Promoters that have Significant Interactions
Promoters that have no Significant Interactions
Promoters overlapping with Enhancers
Promoters do not overlap with Enhancers
Promoter-Enhancer Significant Interactions
Promoter-Promoter Significant Interactions
Promoter-Other Significant Interactions
1. Uni-setting
2. Uni&Multi-setting
3. Specific to Uni-setting
4. Specific to Uni&Multi-setting
5. Common to both settings
6. Uni-setting (75,100)
7. Uni&Multi-setting (64,265)
8. Uni-setting (38,642)
9. Uni&Multi-setting (27,564)
10. Uni-setting(12,691)
11. Uni&Multi-setting(13,043)
12. Specific to Uni-setting (10,149)
13. Specific to Uni&Multi-setting (12,564)
14. Uni-setting (12,558)
15. Uni&Multi-setting (12,940)
16. Uni-setting (4,429)
17. Uni&Multi-setting (4,689)

**Figure 4.** Novel promoter-enhancer significant Hi-C contacts are reproducible and associated with active expressed genes. **(A)** mHi-C identifies novel significant promoter-enhancer interactions (green) that are reproducible among at least two replicates in addition to those reproducible under the Uni-setting (red). Marginal contact counts are depicted in the corresponding tracks of Uni- and Uni&Multi-settings. The complete color labels are consistent with ChromHMM 15-state model at http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html. **(B)** 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 correspond to the number of transcripts in each category.

**Figure 4–source data 1.** The number of significant promoter-enhancer Hi-C interactions at FDR 5% under Uni-setting and Uni&Multi-setting, respectively, for six replicates of IMR90.

**Figure 4–source data 2.** Significant promoter-enhancer Hi-C interactions at FDR 5% under Uni-setting and Uni&Multi-setting for six replicates of IMR90 with the number of contact.

**Figure 4–Figure supplement 1.** Examples of significant promoter-enhancer interactions reproducible among 6 replicates under Uni- and Uni&Multi-settings (IMR90) on chromosome 2.

**Figure 4–Figure supplement 2.** Examples of significant promoter-enhancer interactions reproducible among 6 replicates under Uni- and Uni&Multi-settings (IMR90) on chromosome 7.

**Figure 4–Figure supplement 3.** Examples of significant promoter-enhancer interactions reproducible among 6 replicates under Uni- and Uni&Multi-settings (IMR90) on chromosome 17.

**Figure 4–Figure supplement 4.** Significant promoter-enhancer interactions under Uni- and Uni&Multi-settings across 6 IMR90 replicates (Chromosome 17).

**Figure 4–Figure supplement 5.** Expression distribution of genes promoters of which have significant promoter interactions (IMR90).

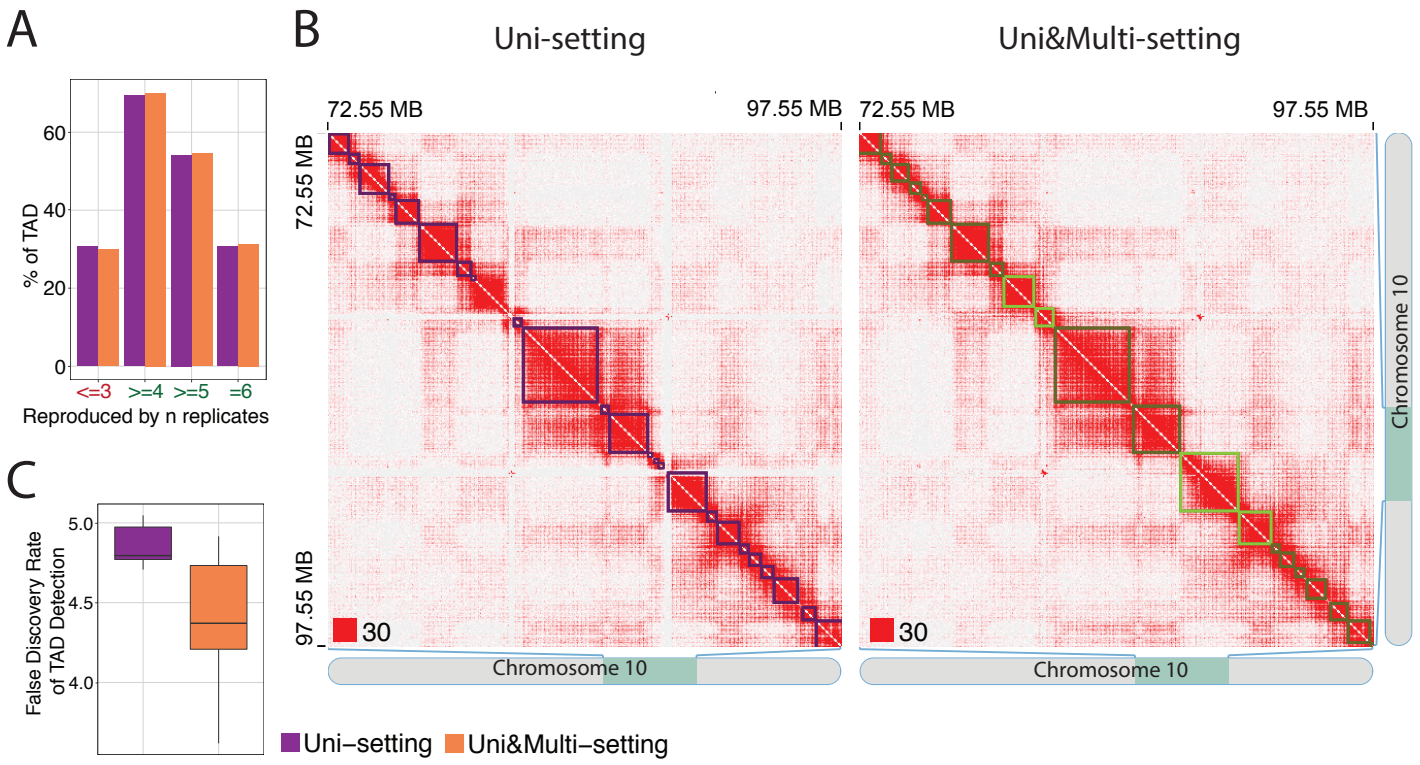



**Figure 5.** mHi-C rescued multi-reads refine detected topologically associating domains. **(A)** Percentage of topologically associating domains (TADs) that are reproducibly detected under Uni-setting and Uni&Multi-setting. TADs that are not detected in at least 4 of the 6 replicates are considered as non-reproducible. **(B)** Comparison of the contact matrices with superimposed TADs between Uni- and Uni&Multi-setting for chr10:72,550,000-97,550,000. Red squares at the left bottom of the matrices indicate the color scale. TADs affected by white gaps involving repetitive regions are highlighted in light green. Light green outlined areas correspond to new TAD boundaries. **(C)** 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.

**Figure 5–source data 1.** Topologically associating domains detected by DomainCaller (***Dixon et al., 2012***) under Uni&Multi-setting for six replicates of IMR90.

**Figure 5–source data 2.** Topologically associating domains detected by Arrowhead (***Rao et al., 2014***) under Uni&Multi-setting for ten replicates of GM12878.

**Figure 5–Figure supplement 1.** The number of topologically associating domains (TADs) detected in each chromosome under Uni-setting and Uni&Multi-setting (IMR90).

**Figure 5–Figure supplement 2.** Comparison of CTCF peaks at the boundaries of topologically associating domains (TADs) under Uni-setting and Uni&Multi-setting across six replicates of IMR90.

**Figure 5–Figure supplement 3.** Novel topologically associating domains (TADs) with CTCF peaks at TAD boundaries (IMR90).

**Figure 5–Figure supplement 4.** Existing topologically associating domains (TADs) with adjusted boundaries supported by CTCF peaks at the new TAD boundaries on chr1 and chr5 (IMR90).

**Figure 5–Figure supplement 5.** Existing topologically associating domains (TADs) with adjusted boundaries supported by CTCF peaks at the new TAD boundaries on chr12 and chr13 (IMR90).

**Figure 5–Figure supplement 6.** False positive topologically associating domains (TADs) detected by the Uni-setting due to the missing reads in low mappability regions on chr2 and chr3 (IMR90).

**Figure 5–Figure supplement 7.** False positive topologically associating domains (TADs) detected by the Uni-setting due to the missing reads in low mappability regions on chr4 and chr16 (IMR90).

**Figure 5–Figure supplement 8.** False positive topologically associating domains (TADs) detected by the Uni-setting due to the missing reads in low mappability regions on chr21 and chrX (IMR90).

**Figure 5–Figure supplement 9.** False discovery rate of TADs detected under two settings. TADs that are not reproducible are labeled as false positives without considering the CTCF peaks at the TAD boundaries.

**Figure 5–Figure supplement 10.** Percentage of TAD boundaries co-localized with different types of repetitive elements under Uni-setting and Uni&Multi-setting for IMR90 at 40kb and GM12878 at 5kb.

**Figure 5–Figure supplement 11.** Repetitive elements enrichment at the reproducible TAD boundaries compared to that within TAD and the average level across the genome for IMR90 at 40kb and GM12878 at 5kb.

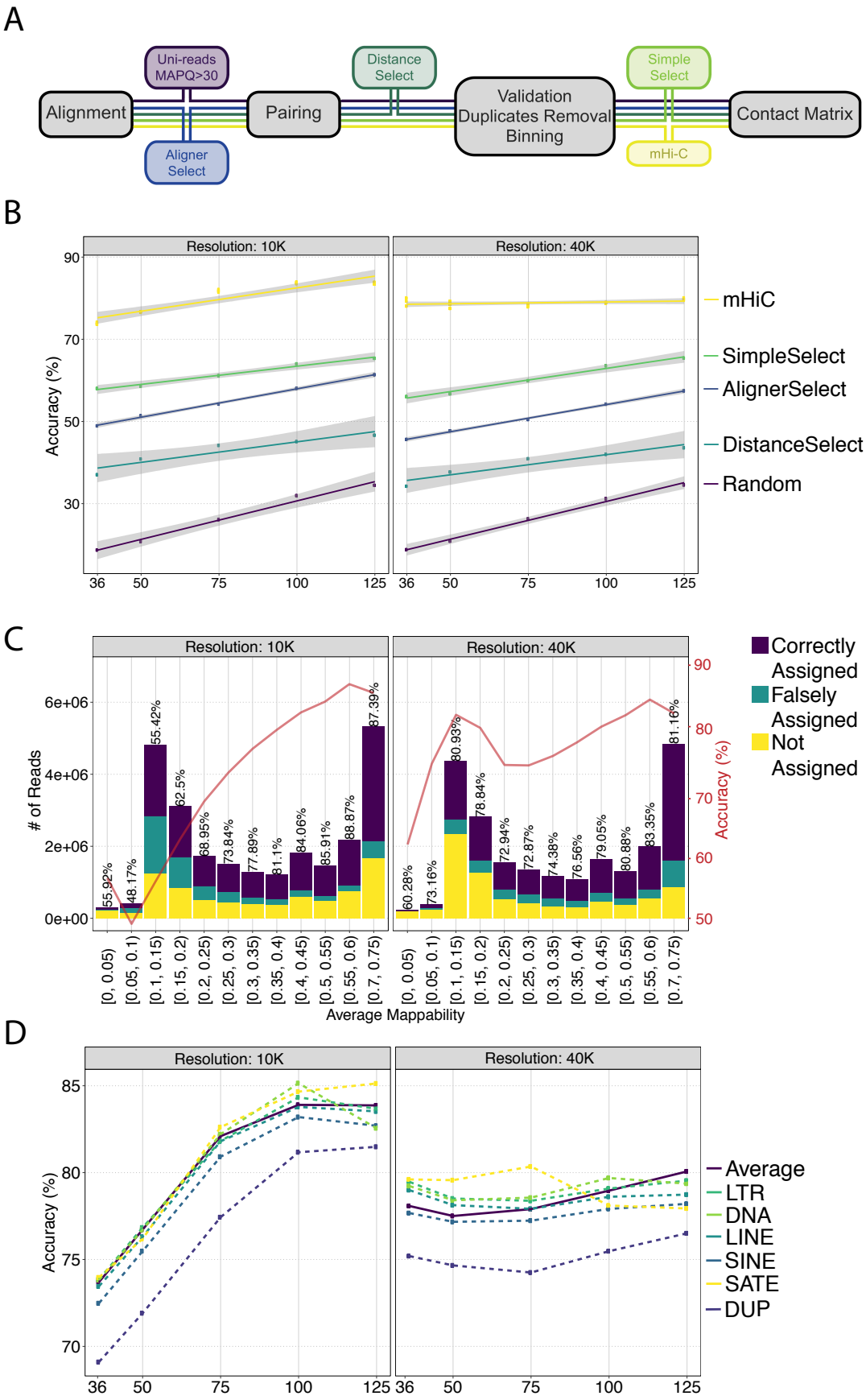

A
Uni-reads MAPQ>30
Alignment
Aligner Select
Pairing
Distance Select
Validation Duplicates Removal Binning
Simple Select
mHi-C
Contact Matrix
B
Resolution: 10K
Resolution: 40K
Accuracy (%)
mHiC
SimpleSelect
AlignerSelect
DistanceSelect
Random
C
Resolution: 10K
Resolution: 40K
# of Reads
Accuracy (%)
Average Mappability
Correctly Assigned
Falsely Assigned
Not Assigned
D
Resolution: 10K
Resolution: 40K
Accuracy (%)
Average
LTR
DNA
LINE
SINE
SATE
DUP

**Figure 6.** Assessing accuracy of mHi-C allocation with trimming experiments utilizing the uni-reads of long read length in A549 datasets. **(A)** Intuitive heuristic strategies (AlignerSelect, DistanceSelect, SimpleSelect) for model-free assignment of multi-reads at various stages the of Hi-C analysis pipeline. **(B)** Accuracy of mHi-C in allocating trimmed multi-reads with respect to trimmed read length, compared with model-free approaches as well as random selection as a baseline. **(C)** Allocation accuracy with respect to mappability. Red solid lines depict the overall accuracy trend. Not assigned category refers to multi-reads with maximum posterior probability of assignment $\leq 0.5$. **(D)** mHi-C accuracy among different types of repetitive element classes.

**Figure 6–Figure supplement 1.** The numbers of uni- and multi-reads after trimming compared with original sequencing depth (A549).
**Figure 6–Figure supplement 2.** Impact of chimeric reads in A549 datasets.
**Figure 6–Figure supplement 3.** Intra-chromosome allocation accuracy with respect to trimmed read length.
**Figure 6–Figure supplement 4.** Allocation accuracy across different mappability regions for trimmed reads of 50bp, 75bp, 100bp, and 125bp respectively.
**Figure 6–Figure supplement 5.** Allocation accuracy across different classes of repetitive elements at 10kb and 40kb resolutions.
**Figure 6–Figure supplement 6.** mHi-C and model-free approaches of utilizing multi-reads.
**Figure 6–Figure supplement 7.** Comparison of significant interactions among the three life stages of *P. falciparum*.

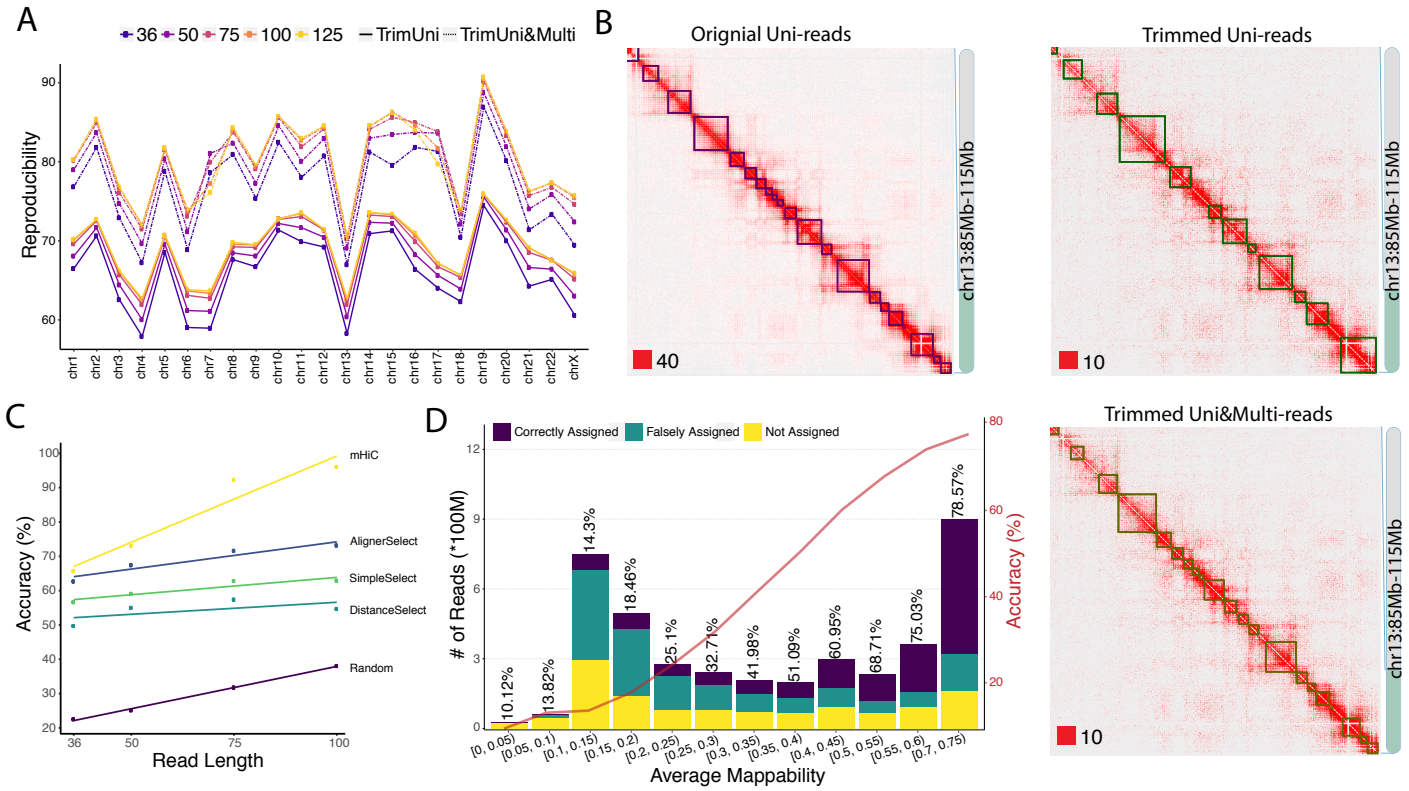



**Figure 7.** Trimmed uni- and multi-reads to recover the original contact matrix of the longer read dataset A549. **(A)** mHi-C rescued multi-reads of the trimmed dataset along with trimmed uni-reads lead to contact matrices that are significantly more similar to original contact matrices compared to only using trimmed uni-reads. **(B)** TAD detection on chromosome 13 with the original longer read uni-reads contact matrix, trimmed uni-reads (125bp long) contact matrix, and trimmed uni- and multi-reads (125bp long) contact matrix. **(C)** Simulation accuracy of mHi-C with respect to different simulated read lengths in comparison to other model-free approaches as well as random selection as baseline. **(D)** mHi-C allocation accuracy across different mappability regions of simulated 36bp reads. The red curve denotes the overal accuracy trend.

**Figure 7–Figure supplement 1.** Reproducibility comparison after adding mHi-C rescued trimmed multi-reads to trimmed uni-reads across different read lengths at 40kb.

**Figure 7–Figure supplement 2.** Allocation accuracy of trimmed reads of different lengths among different mappability regions at 40kb.

**Figure 7–Figure supplement 3.** TAD detection on chromosome 1 of original replicate uni-reads contact matrix, trimmed uni-reads (36bp long) contact matrix and trimmed uni- and multi-reads (36bp long) contact matrix.

**Figure 7–Figure supplement 4.** TAD detection on chromosome 4 of original replicate uni-reads contact matrix, trimmed uni-reads (50bp long) contact matrix and trimmed uni- and multi-reads (50bp long) contact matrix.

**Figure 7–Figure supplement 5.** TAD detection on chromosome 7 of original replicate uni-reads contact matrix, trimmed uni-reads (75bp long) contact matrix and trimmed uni- and multi-reads (75bp long) contact matrix.

**Figure 7–Figure supplement 6.** TAD detection on chromosome 10 of original replicate uni-reads contact matrix, trimmed uni-reads (100bp long) contact matrix and trimmed uni- and multi-reads (100bp long) contact matrix.

**Figure 7–Figure supplement 7.** Numbers of significant interactions identified with trimmed reads under Uni- and Uni&Multi-settings at FDR 0.1%, 1%, 5%, 10% that are also identified (at FDR 10%) with the original contact matrix.

**Figure 7–Figure supplement 8.** Proportion of significant interactions of orignal full reads being recovered in trimmed reads sets.

**Figure 7–Figure supplement 9.** Number of significant interactions identified by simulated Uni&Multi-setting and the allocation accuracy using merged mHi-C allocation process.

**Figure 7–Figure supplement 10.** mHi-C allocation accuracy with respect to mappabilities of simulated reads (50bp, 75bp, 100bp).

## References


**Ay F**, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research. 2014; 24(6):999–1011.

**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 expression. Genome Research. 2014; 24(6):974–988.

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

**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. 2012; 41(D1):D991–D995.

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

**Bonev B**, Cohen NM, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, et al. Multiscale 3D genome rewiring during mouse neural development. Cell. 2017; 171(3):557–572.

**Bunnik EM**, Cook KB, Varoquaux N, Batugedara G, Prudhomme J, Shi L, Andolina C, Ross LS, Brady D, Fidock DA, et al. Changes in genome organization of parasite-specific gene families during the Plasmodium transmission stages. bioRxiv. 2018; p. 242123.

**Casper J**, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D, et al. The UCSC genome browser database: 2018 update. Nucleic acids research. 2017; 46(D1):D762–D769.

**Chung D**, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, 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. 2011; 7(7):e1002111.

**Consortium EP**, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57.

**Corradin O**, Cohen AJ, Luppino JM, Bayles IM, Schumacher FR, Scacheri PC. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. Nature genetics. 2016; 48(11):1313.

**Cournac A**, Koszul R, Mozziconacci J. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. Nucleic Acids Research. 2015; 44(1):245–255.

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

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

**Dixon JR**, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015; 518(7539):331.

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

**Durand NC**, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Systems. 2016; 3(1):99–101.

**Ferrari KJ**, 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. 2014; 53(1):49–62.

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

**Grant CE**, Bailey TL, Nobel WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics. 2011; 7:1017. http://www.ncbi.nlm.nih.gov/pubmed/21330290, doi: 10.1093/bioinformatics/btr064.

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

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

**Hsu SC**, Gilgenast TG, Bartman CR, Edwards CR, Stonestrom AJ, Huang P, Emerson DJ, Evans P, Werner MT, Keller CA, et al. The BET protein BRD2 cooperates with CTCF to enforce transcriptional and architectural boundaries. Molecular Cell. 2017; 66(1):102–116.

**Hwang YC**, Lin CF, Valladares O, Malamon J, Kuksa PP, Zheng Q, Gregory BD, Wang LS. HIPPIE: a high-throughput identification pipeline for promoter interacting enhancer elements. Bioinformatics. 2014; 31(8):1290–1292.

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

**Javierre BM**, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Várnai C, Thiecke MJ, et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell. 2016; 167(5):1369–1384.

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

**Kerpedjiev P**, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Ahzir A, Kumar N, et al. HiGlass: Web-based visual comparison and exploration of genome interaction maps. bioRxiv. 2017; p. 121889.

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

**Knight PA**, Ruiz D. A fast algorithm for matrix balancing. IMA Journal of Numerical Analysis. 2013; 33(3):1029–1047.

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

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

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

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

**Lun AT**, Smyth GK. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics. 2015; 16(1):258.

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

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

**Ong CT**, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics. 2014; 15(4):234.

**Open R**, 4.0 [http://www. repeatmasker. org]. Accessed; 2015.

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

**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. 2014; 159(7):1665–1680.

**Roadmap Epigenomics Consortium**. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518(7539):317–330. http://view.ncbi.nlm.nih.gov/pubmed/25693563.

**Rosa-Garrido M**, Chapski DJ, Schmitt AD, Kimball TH, Karbassi E, Monte E, Balderas E, Pellegrini M, Shih TT, Soehalim E, et al. High resolution mapping of chromatin conformation in cardiac myocytes reveals structural remodeling of the epigenome in heart failure. Circulation. 2017; p. Circulation–117.

**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.

**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. 2015; 16(1):259.

**Spielmann M**, Lupiáñez DG, Mundlos S. Structural variation in the 3D genome. Nature Reviews Genetics. 2018; p. 1.

**Sun JH**, Zhou L, Emerson DJ, Phyo SA, Titus KR, Gong W, Gilgenast TG, Beagan JA, Davidson BL, Tassone F, et al. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. Cell. 2018; .

**Tang Z**, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, Michalski P, Piecuch E, Wang P, Wang D, Tian SZ, Penrad-Mobayed M, Sachs LM, Ruan X, Wei CL, Liu ET, et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell. 2015; 163(7):1611–1627.

**The ENCODE Project Consortium**. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 09; 489:57–74. https://www.encodeproject.org.

**Treangen TJ**, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2012; 13(1):36.

**Won H**, de La Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature. 2016; 538(7626):523.

**Xie M**, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, Gascard P, Sigaroudinia M, Tlsty TD, Kadlecek T, Weiss A, O'Geen H, Farnham PJ, Madden PAF, Mungall AJ, Tam A, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nature Genetics. 2013; 45(7):836–41.

**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; p. gr–220640.

**Yardimci G**, 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. bioRxiv. 2017; p. 188755.

**Yu M**, Ren B. The Three-Dimensional Organization of Mammalian Genomes. Annual review of cell and developmental biology. 2017; 33:265–289.

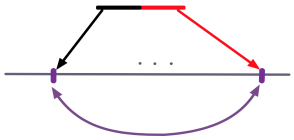
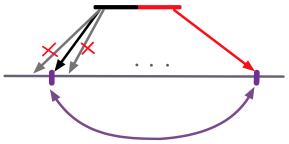
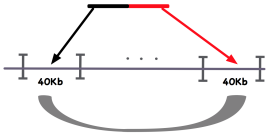
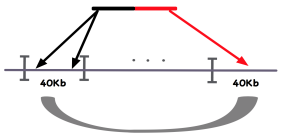
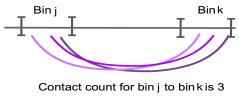
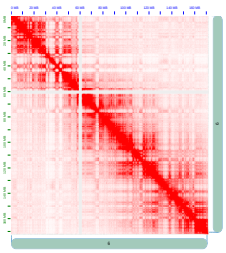
**Zeng X**, Li B, Welch R, Rojo C, Zheng Y, Dewey CN, Keleş S. Perm-seq: mapping protein-DNA interactions in segmental duplication and highly repetitive regions of genomes with prior-enhanced read mapping. PLoS Computational Biology. 2015; 11(10):e1004491.

**Zhang Q**, Keleş S. CNV-guided multi-read allocation for ChIP-seq. Bioinformatics. 2014; 30(20):2860–2867.

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

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

**Appendix 1 Table 2.** Hi-C and mHi-C terminology.

| Terminology | Description | Diagram | Standard Pipeline | mHi-C |
|---|---|---|---|---|
| 1a. Read ends alignment | Each end of a paired-end read is aligned independently. | | Uni-reads | Uni&multi-reads |
| 1b. Chimeric read rescue | Read end spans ligation site. They are rescued by trimming and re-alignment. | | Uni-chimeric ends | Uni&multi-chimeric ends |
| 2a. Read end pairing | Aligned read ends form an alignment position pair. |  | Uni-read pairs | Uni&multi-read pairs |
| 2b. Multi-reads reduced to uni-reads due to validation checking | Only one of the alignment position pairs passes the validation checking. |  | Valid uni-reads | Valid uni- and multi-reduced uni-reads |
| 3a. Bin pair | Bin the genome by a fixed size window. Read end alignments fall within bins and form bin pairs indicating contact from one bin to another. |  | Uni-reads binning | Uni&multi-reads binning |
| 3b. Multi-reads reduced to uni-reads due to binning | Each read end has multiple alignment positions falling within the same bin, thus supporting the same bin pair. |  | Uni-reads bin pairs | Uni- and multi-reduced uni-bin pairs |
| 4. Contact count | Each read pair alignment represents one contact. Contact count for bin pair $j$ and $k$ is the total number of read pairs with ends in $bin_j$ and $bin_k$. |  | Uni-mapping contacts | Uni&multi-mapping contacts |
| 5. Contact matrix | The entry at row j and column k indicates the contact count of $bin_j$ interacting with $bin_k$. |  | Uni-setting contact matrix | Uni&Multi-setting contact matrix |

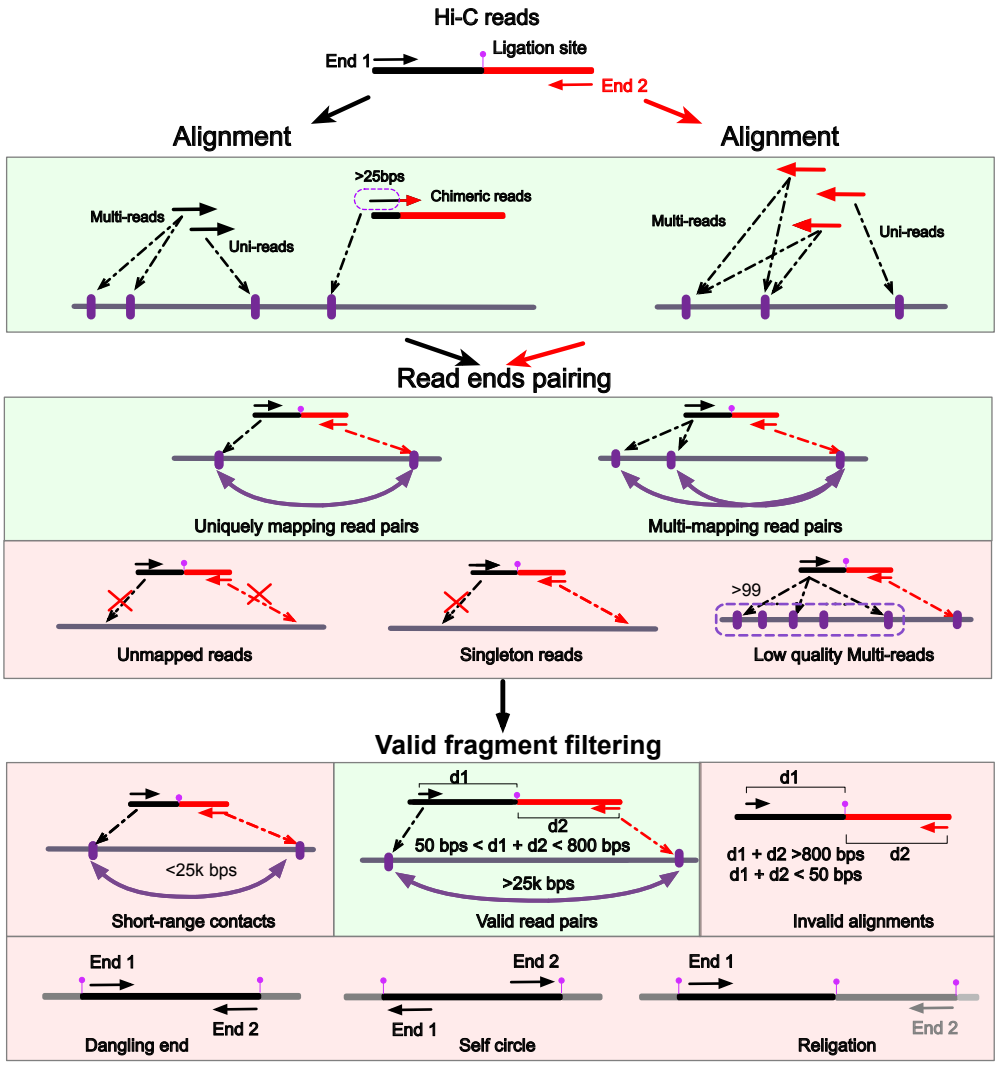



**Figure 1–Figure supplement 1.** mHi-C pipeline (Alignment - Read end pairing - Valid fragment filtering). **1.** Read ends are aligned to reference genome separately allowing multi-reads and chimeric reads to be rescued. **2.** Read ends are paired by their read query names. Multi-reads form more than one read pair with the same read query name. Read ends that fail to align form either unmapped reads or singleton reads and are discarded. Multi-reads with ends aligning to more than 99 positions are regarded as low quality multi-reads and are excluded from the downstream analysis. **3.** Validation checking to filter short-range contacts and alignments far away from restriction enzyme recognition sites. Contacts residing within the same restriction fragment, i.e., dangling end or self circle, as well as adjacent fragments (religation) are discarded. The above three processing steps are applied to each read independently enabling parallel implementation.

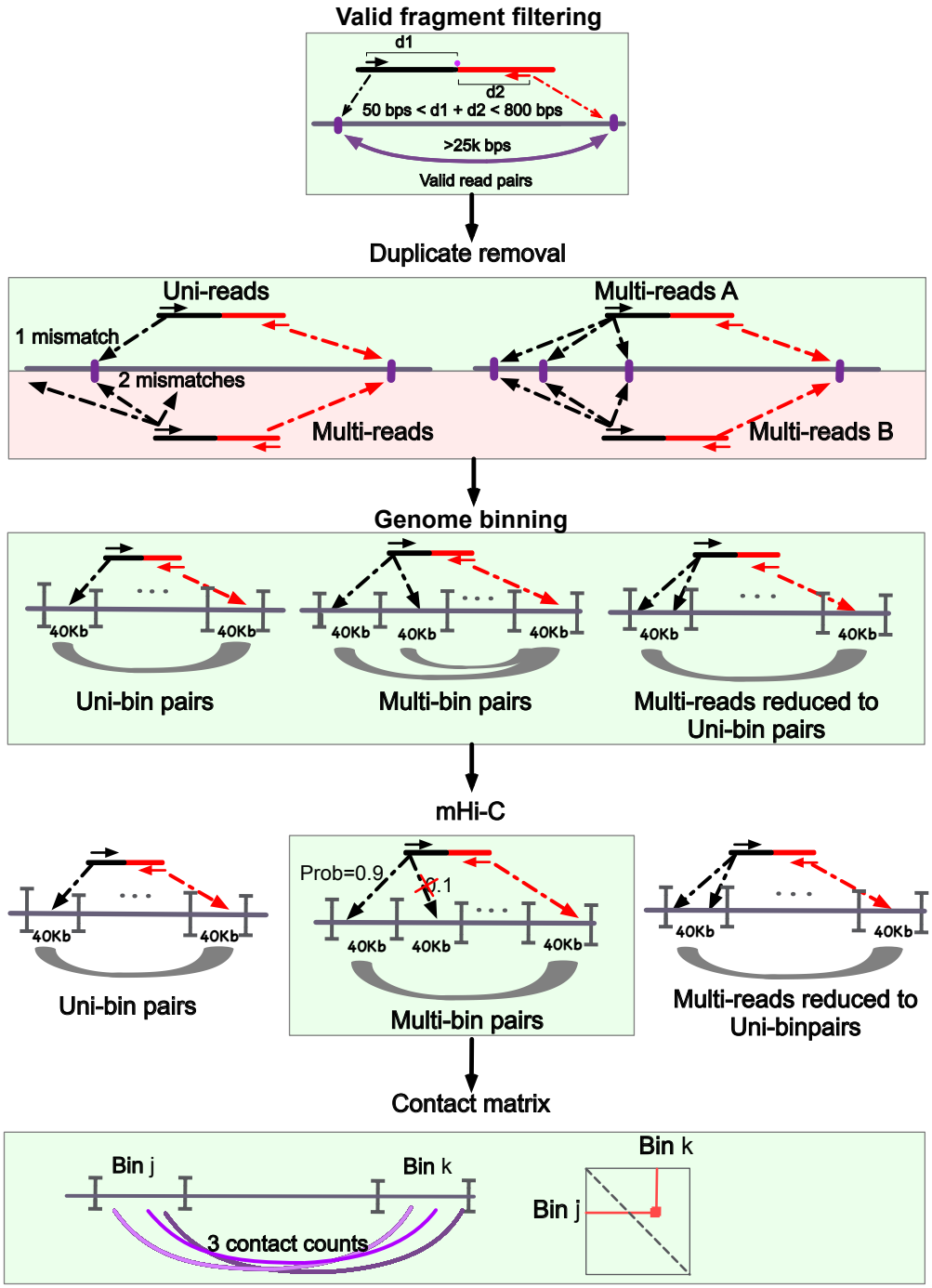



**Figure 1–Figure supplement 2.** mHi-C pipeline (Duplicate removal - Genome binning - mHi-C). **4.** PCR duplicates are removed to ensure that when a uni-read and a multi-read have the same alignment position and strand direction, the uni-read is kept. In the case of multi-reads that overlap with other multi-reads, the ones with alphabetically larger IDs are removed. **5.** Genome is split into fix-sized non-overlapping intervals, i.e., bins or fixed number of restriction fragments and, as a result, read alignment position pairs are reduced to bin pairs. Multi-reads, candidate alignment positions of which fall into the same bin, are reduced to uni-bin pairs. **6.** mHi-C model estimates an allocation probability for each potential contact and enables filtering of contacts by thresholding this allocation probability.

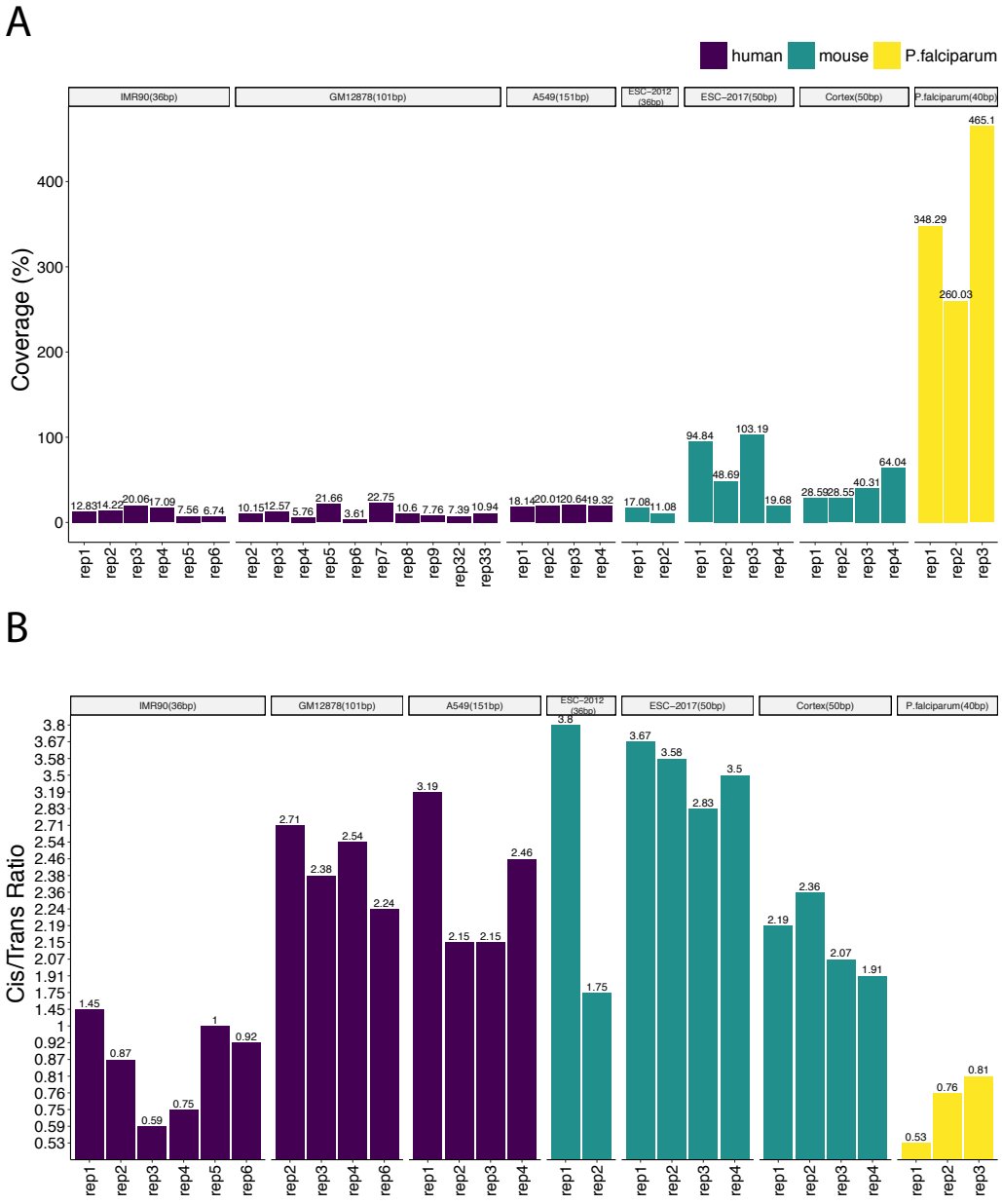



**Figure 1–Figure supplement 3.** Coverage and cis-to-trans ratios across individual replicates of the study datasets as indicators of data quality. **(A)** Coverage is approximated as the ratio of the sequencing depth to the genome size. **(B)** Cis-to-trans ratio is defined as the the number of valid intra-chromosomal contacts divided by the number of valid inter-chromosomal contacts.

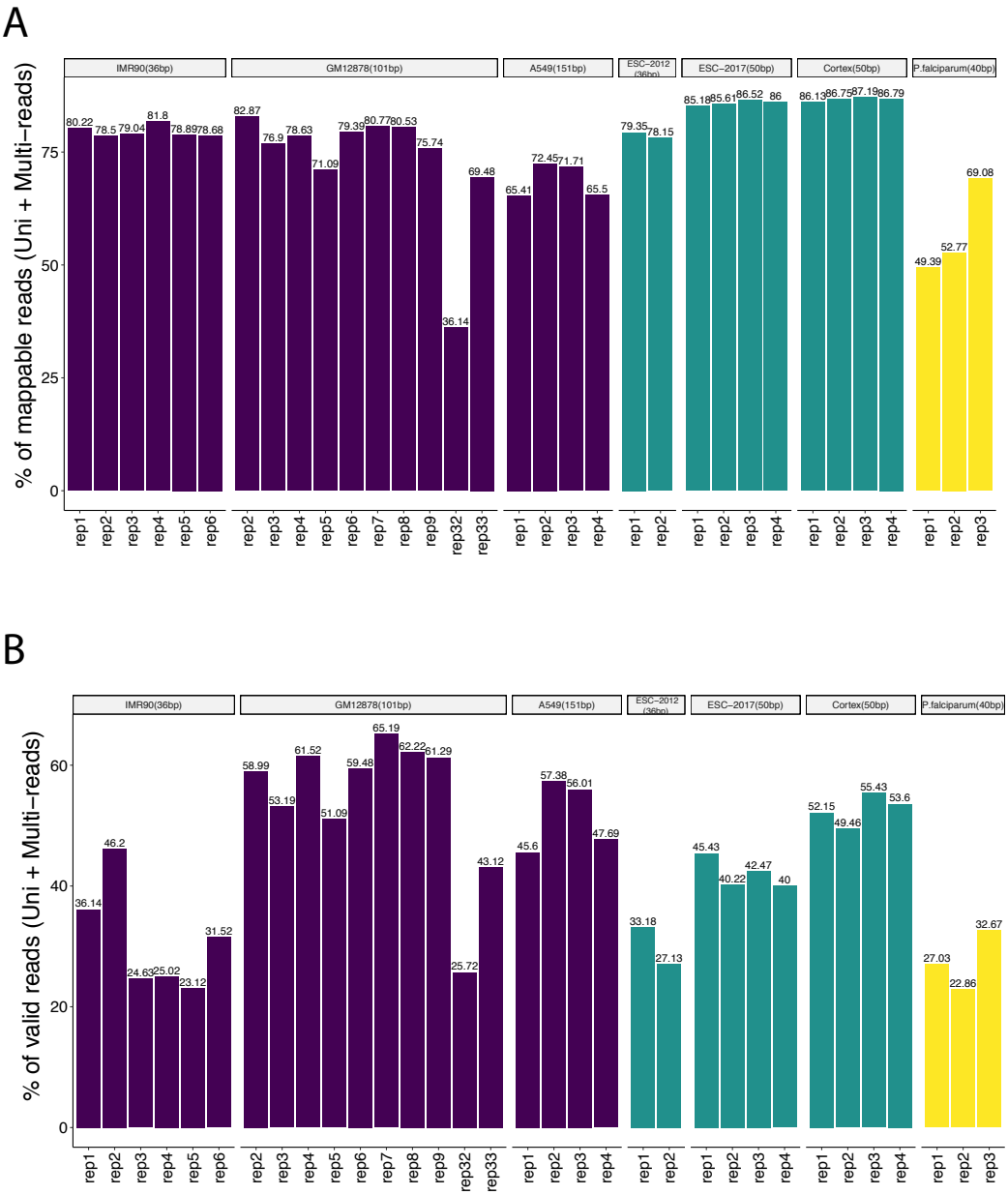



**Figure 1–Figure supplement 4.** Percentages of **(A)** mappable and **(B)** valid reads over the set of all reads for individual replicates of the study datasets as an indicator of data quality. Both the aligned uni-reads and multi-reads are taken into consideration.

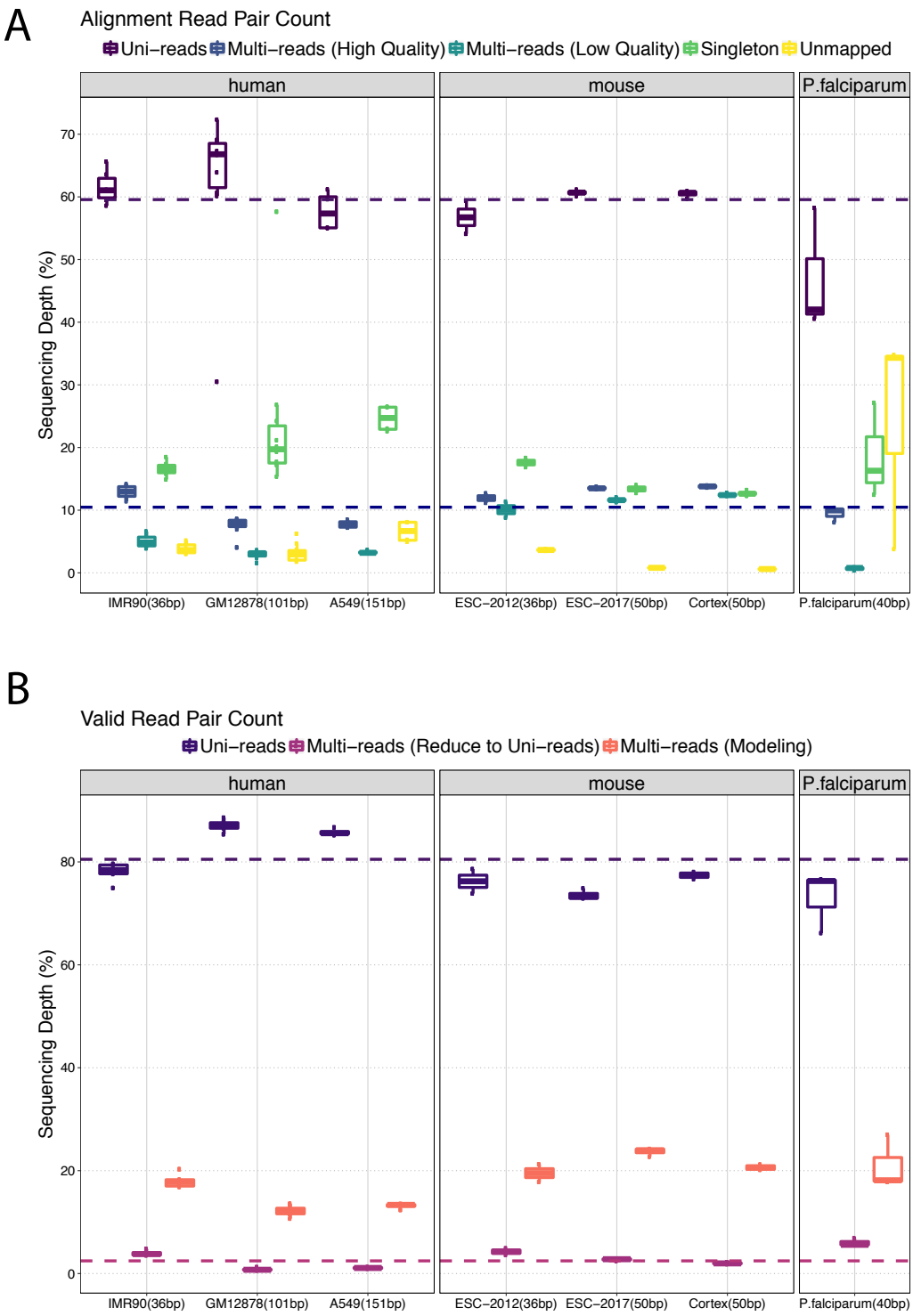



**Figure 1–Figure supplement 5.** Categorization of reads after alignment across study datasets. **(A)** Percentages of mapped reads in different alignment categories. **(B)** Percentages of valid reads with both ends uniquely aligned to reference genome (Uni-reads), at least one end aligning to multiple position and resulting in only one valid alignment after validation and binning (Multi-reads (Reduce to Uni-reads)), and reads with multiple potential valid alignment positions (Multi-reads (Modeling)).

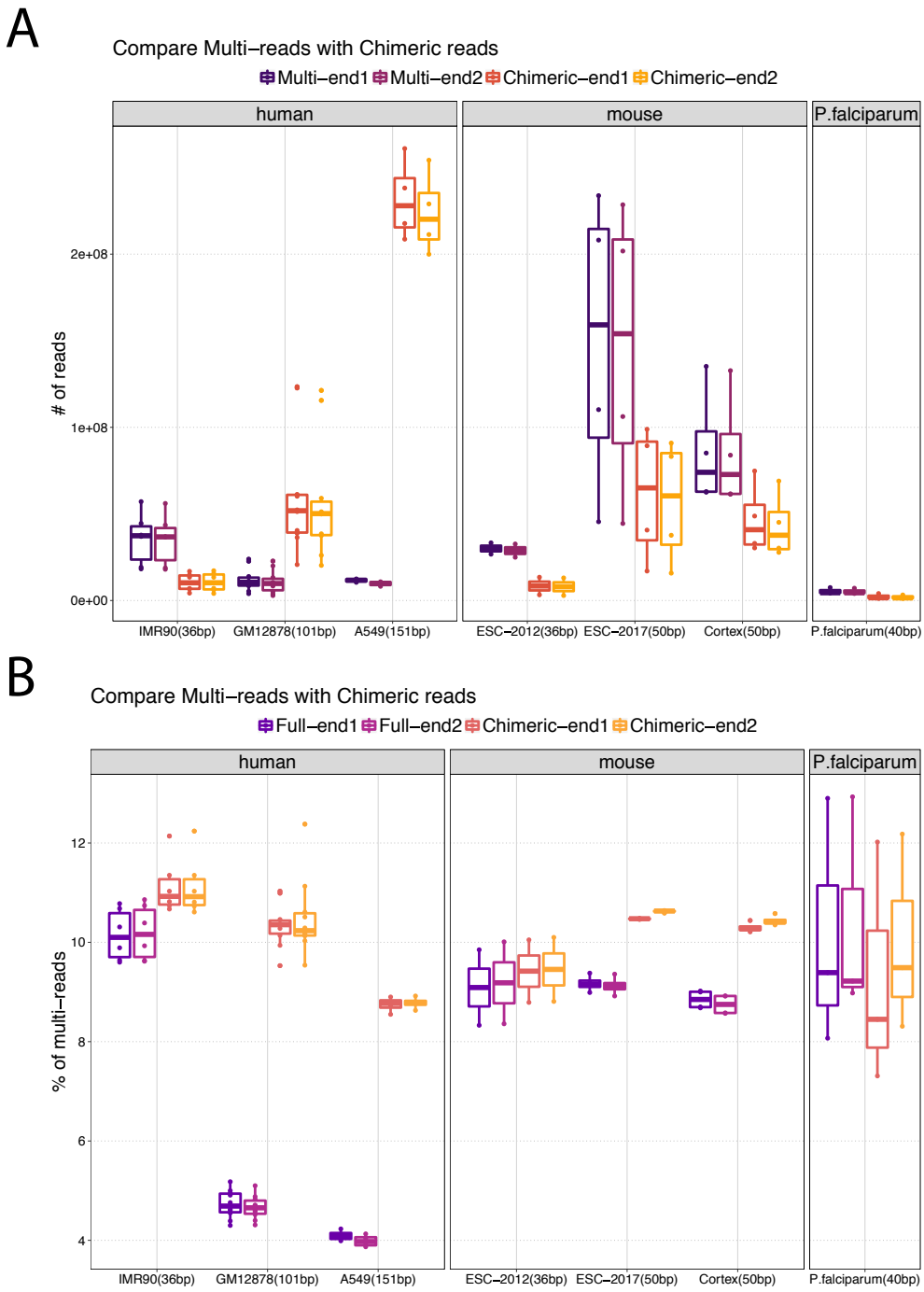



**Figure 1–Figure supplement 6.** Both chimeric reads and multi-reads require extra processing to rescue. **(A)** Numbers of multi-reads compared to the numbers of chimeric reads at each read end level. For Hi-C datasets with have shorter read lengths, multi-reads constitute a larger percentage of the usable reads compared to chimeric reads. **(B)** Proportion of multi-reads among full-length alignable reads compared with that among chimeric reads. As expected, chimeric reads lead to larger percentages of multi-reads.

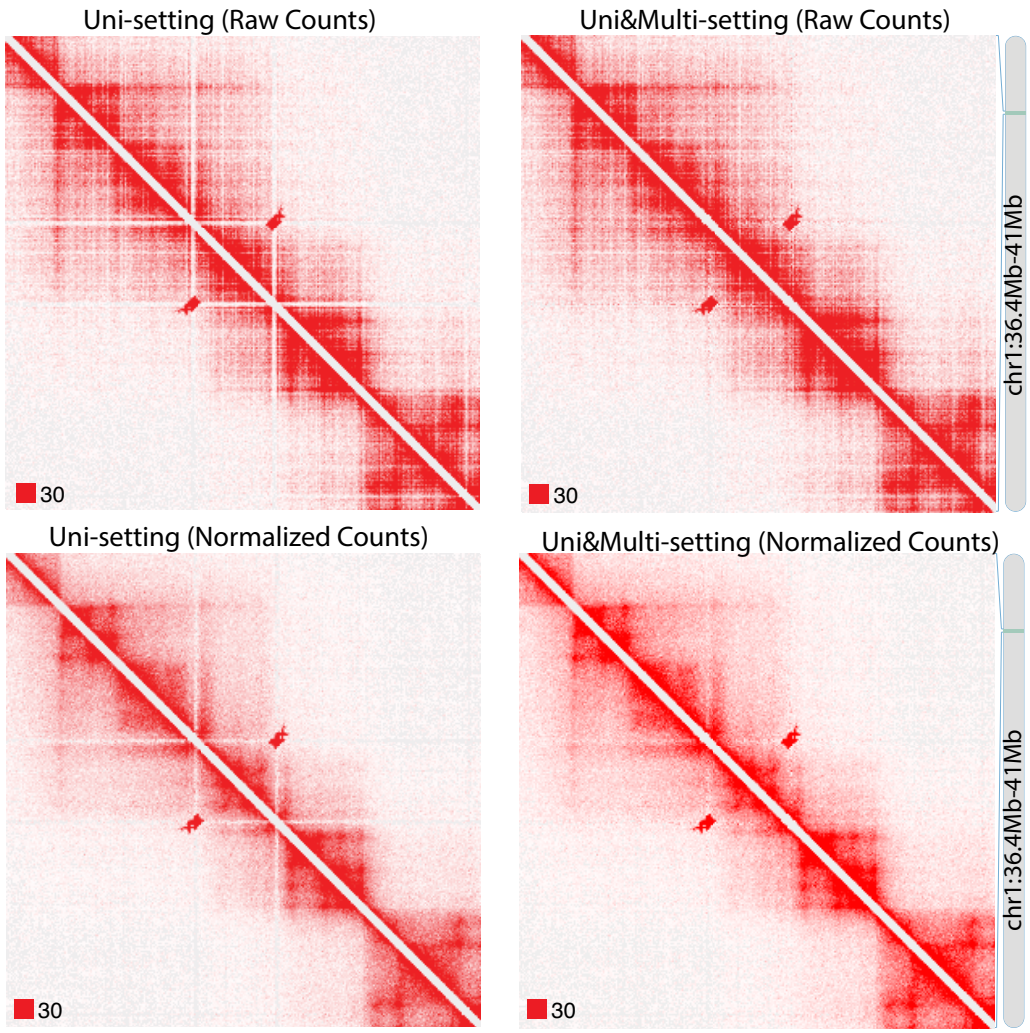



**Figure 2–Figure supplement 1.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 1.

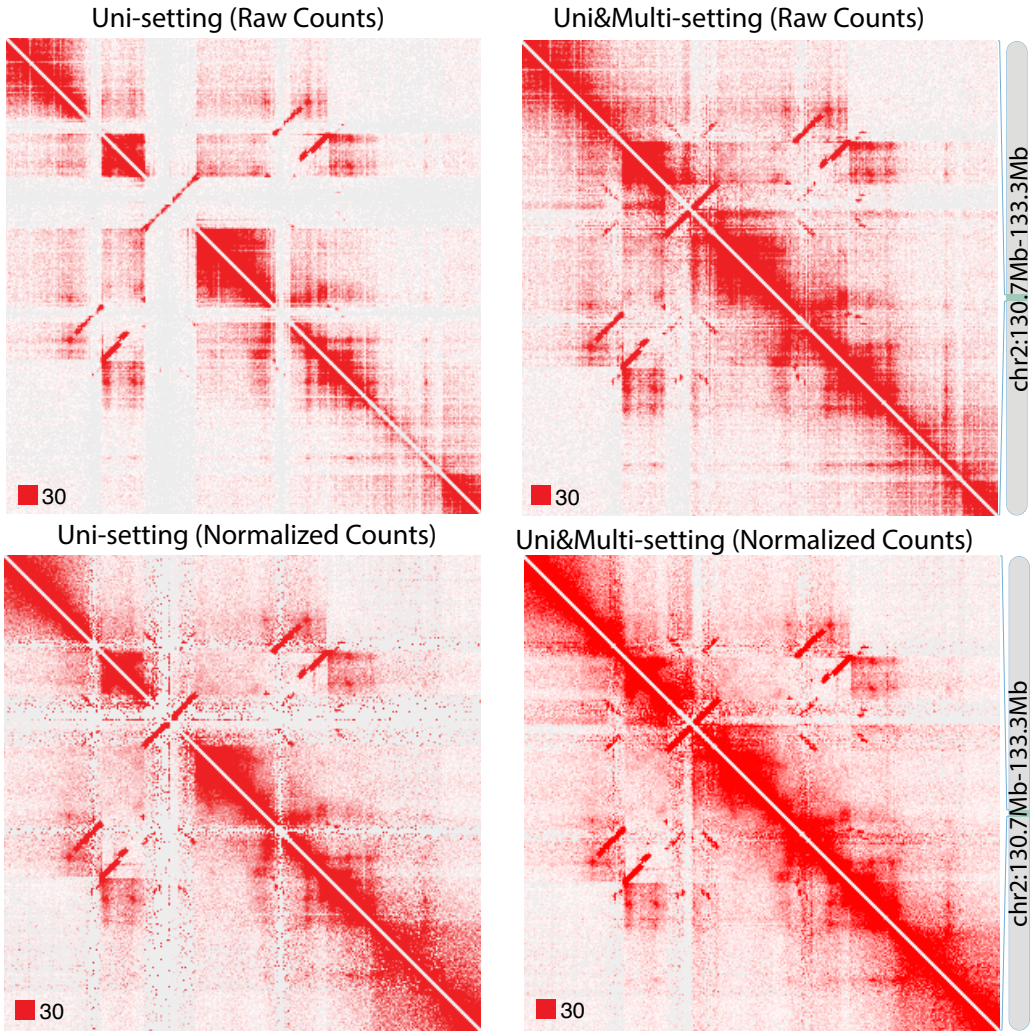



**Figure 2–Figure supplement 2.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 2.

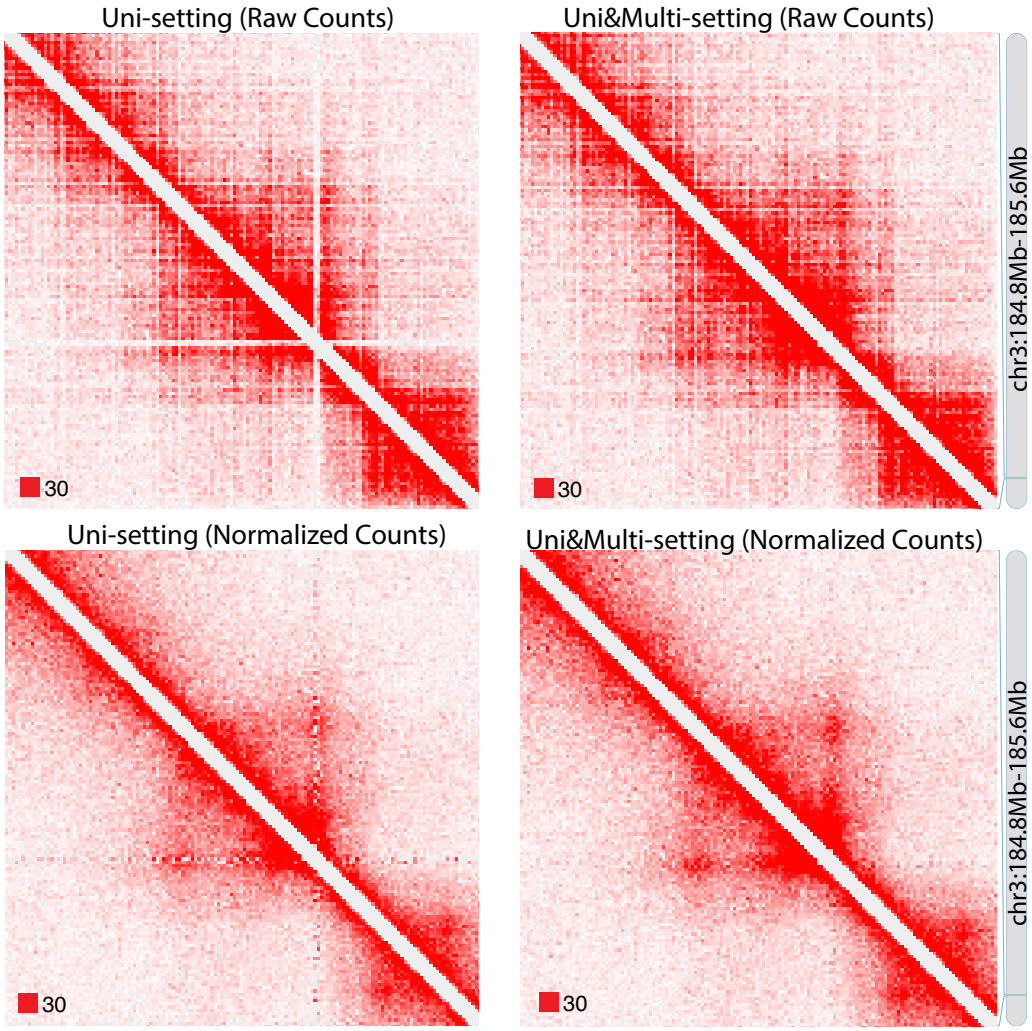



**Figure 2–Figure supplement 3.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 3.

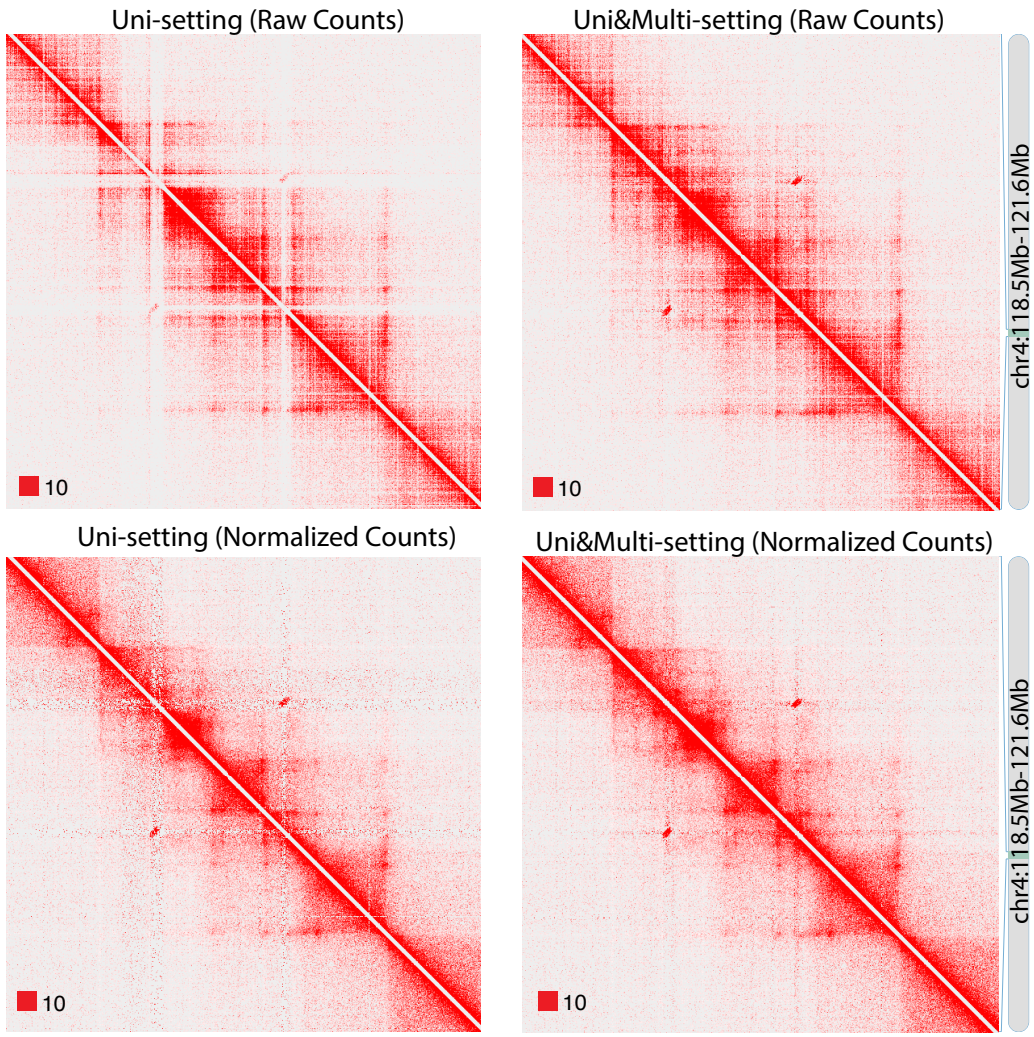



**Figure 2–Figure supplement 4.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 4.

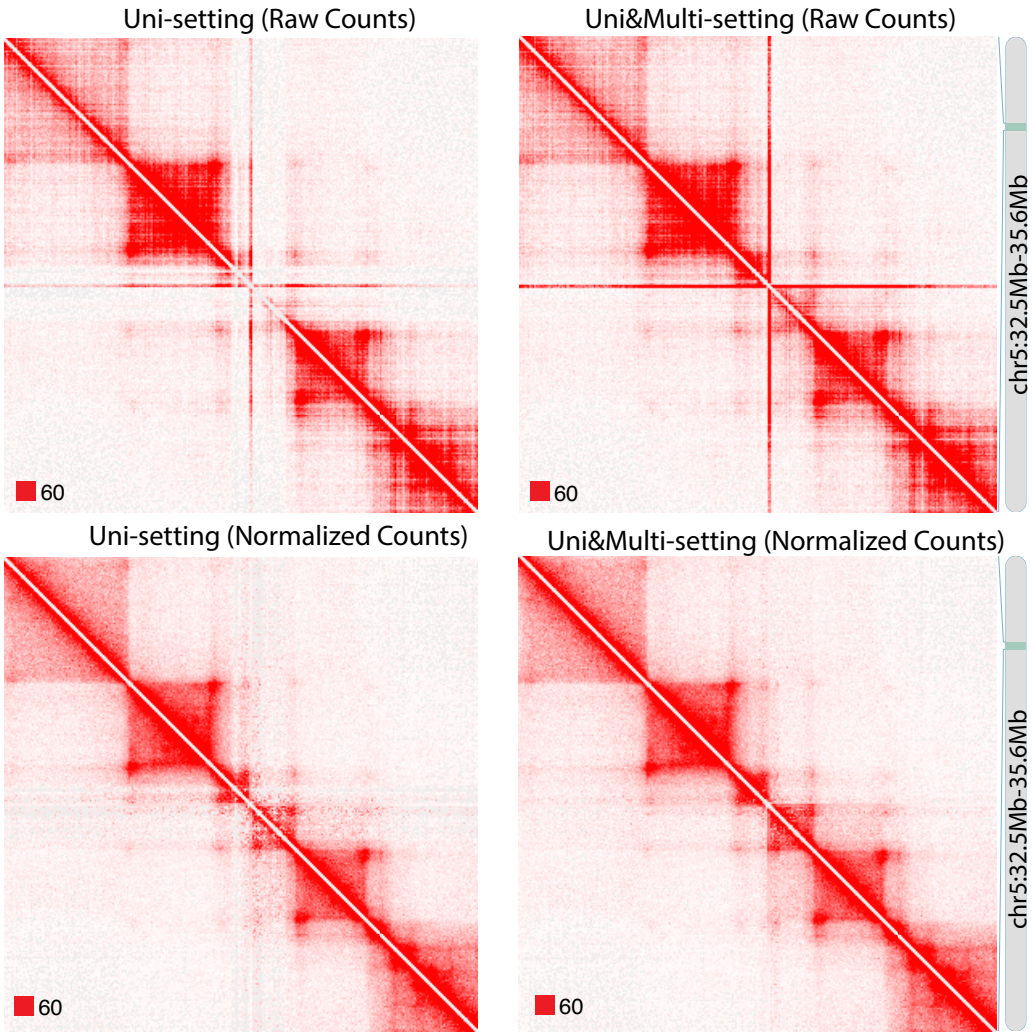



**Figure 2–Figure supplement 5.** Raw and normalized contact matrices of GM12878 under Uni-setting and Uni&Multi-setting on chromosome 5.

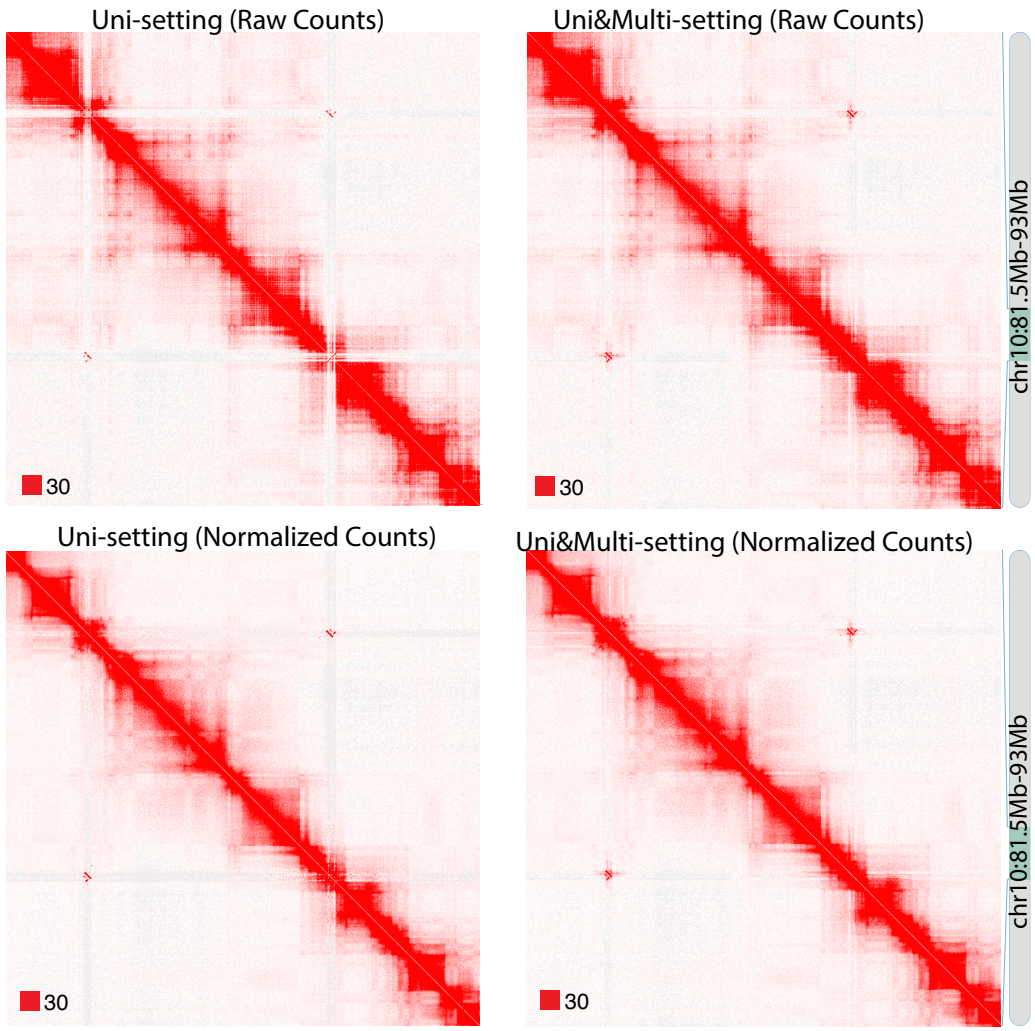



**Figure 2–Figure supplement 6.** Loop signal remains weak even after the normalization of the contact matrix under Uni-setting. Multi-reads recovered the loop signal.

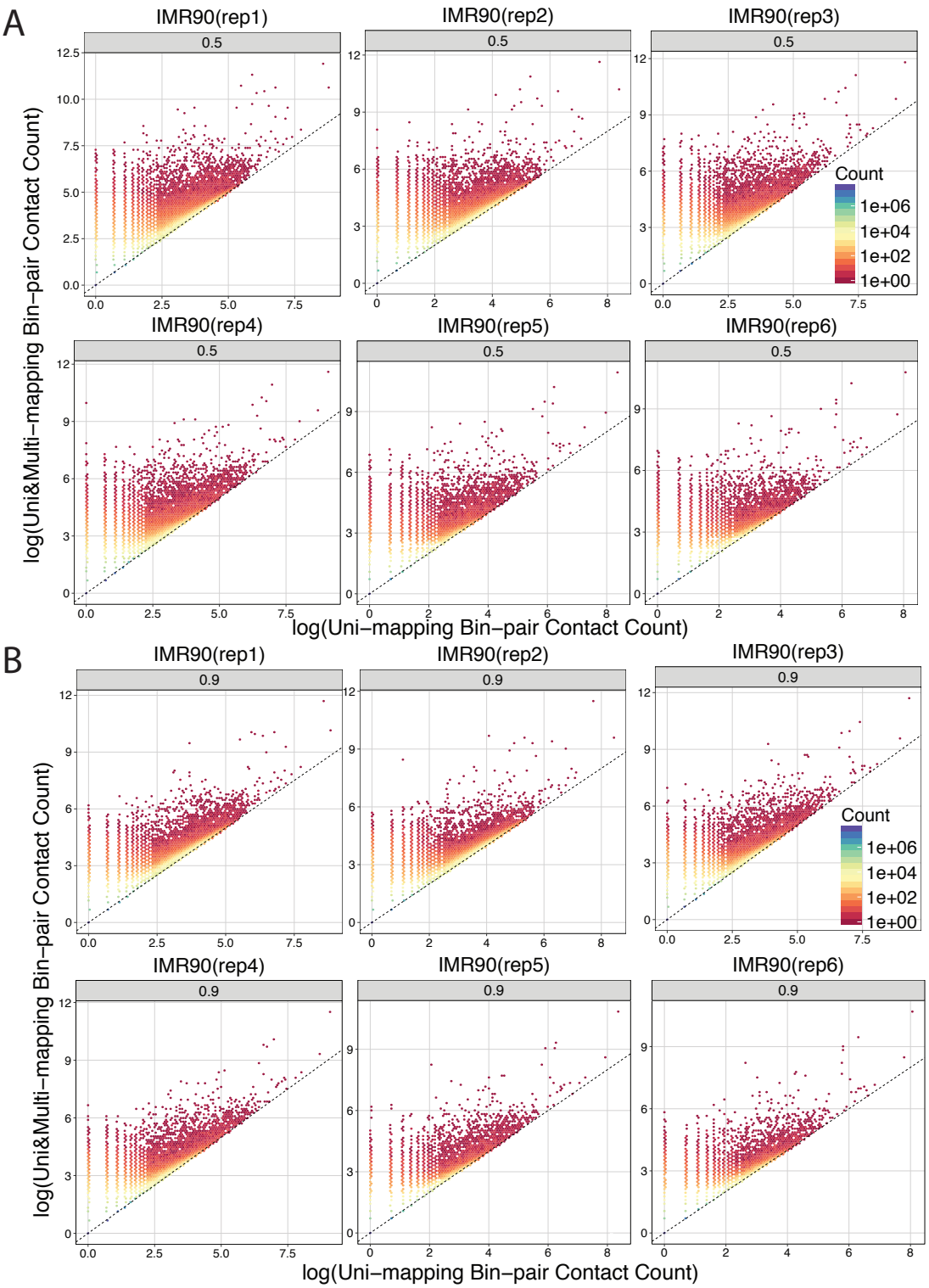



**Figure 2–Figure supplement 7.** Bin coverage improvement of Uni&Multi-setting compared to Uni-setting for IMR90 at the individual replicate level for two different allocation probability thresholds. Uni&Multi-setting for **(A)** includes multi-reads with posterior probability $> 0.5$, whereas **(B)** depicts more strict filtering with an allocation probability greater than $0.9$. The dashed line is y = x. mHi-C rescues multi-reads from valid ligation fragments, resulting in significant increase in contact counts.

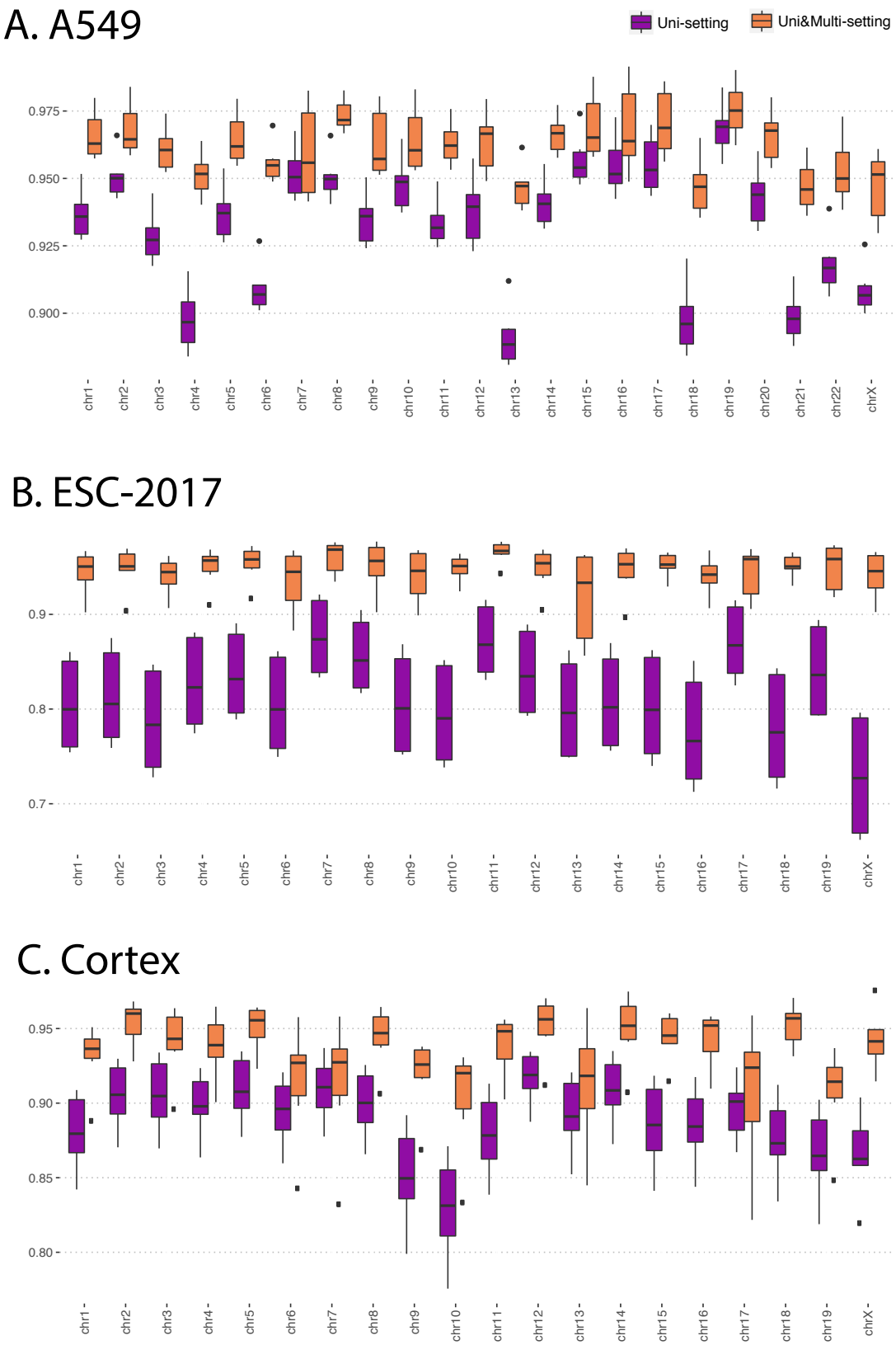



**Figure 2–Figure supplement 8.** Reproducibility at the contact matrix level under the Uni- and Uni&Multi-settings in A549, ESC-2017 and Cortex cell lines. Note that the box plots for the ESC-2012 cell line which only has two replicates are not displayed.

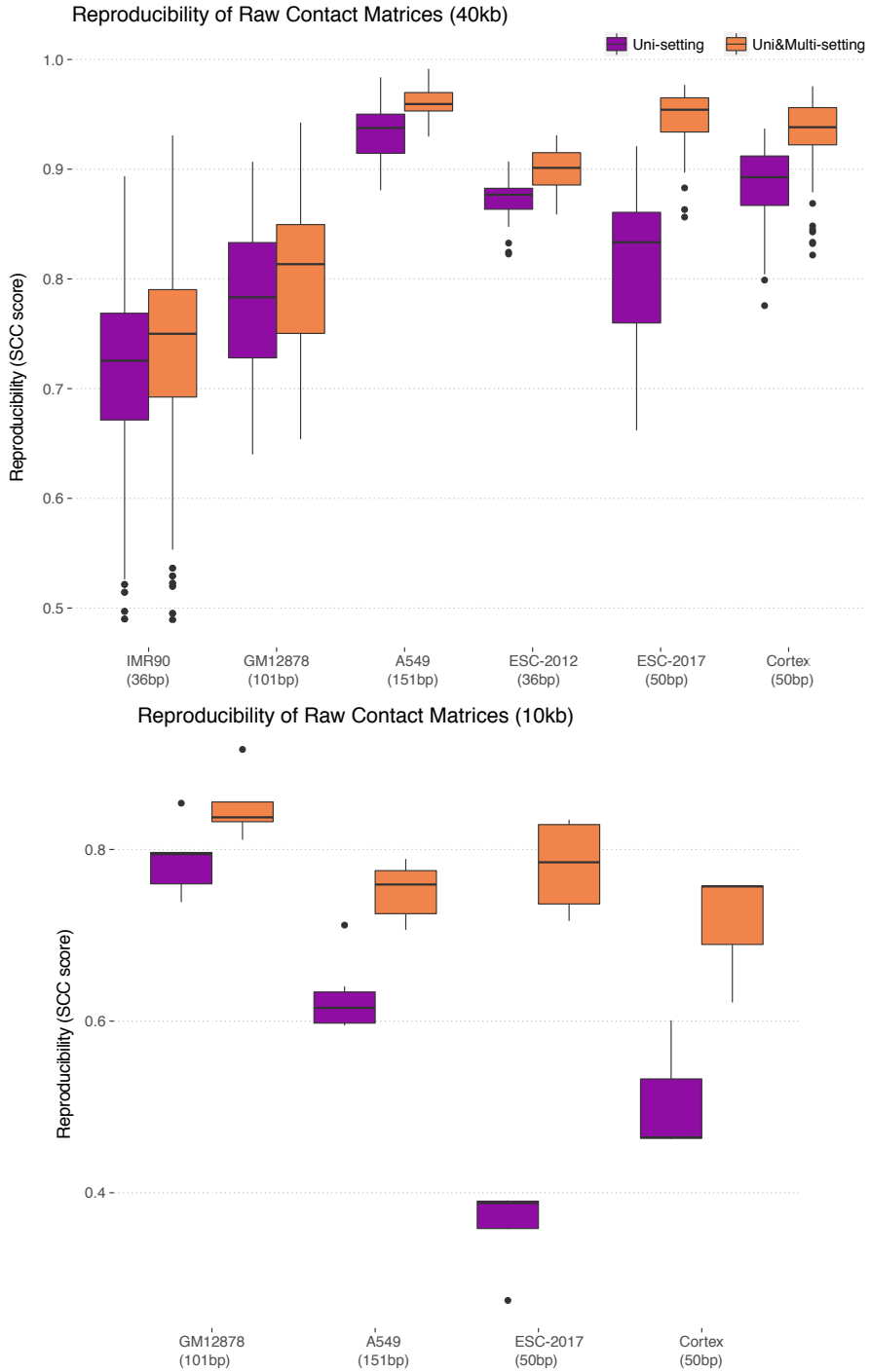



**Figure 2–Figure supplement 9.** Reproducibility at the contact matrix level at resolutions 40kb (low) and 10kb (high) across study datasets.

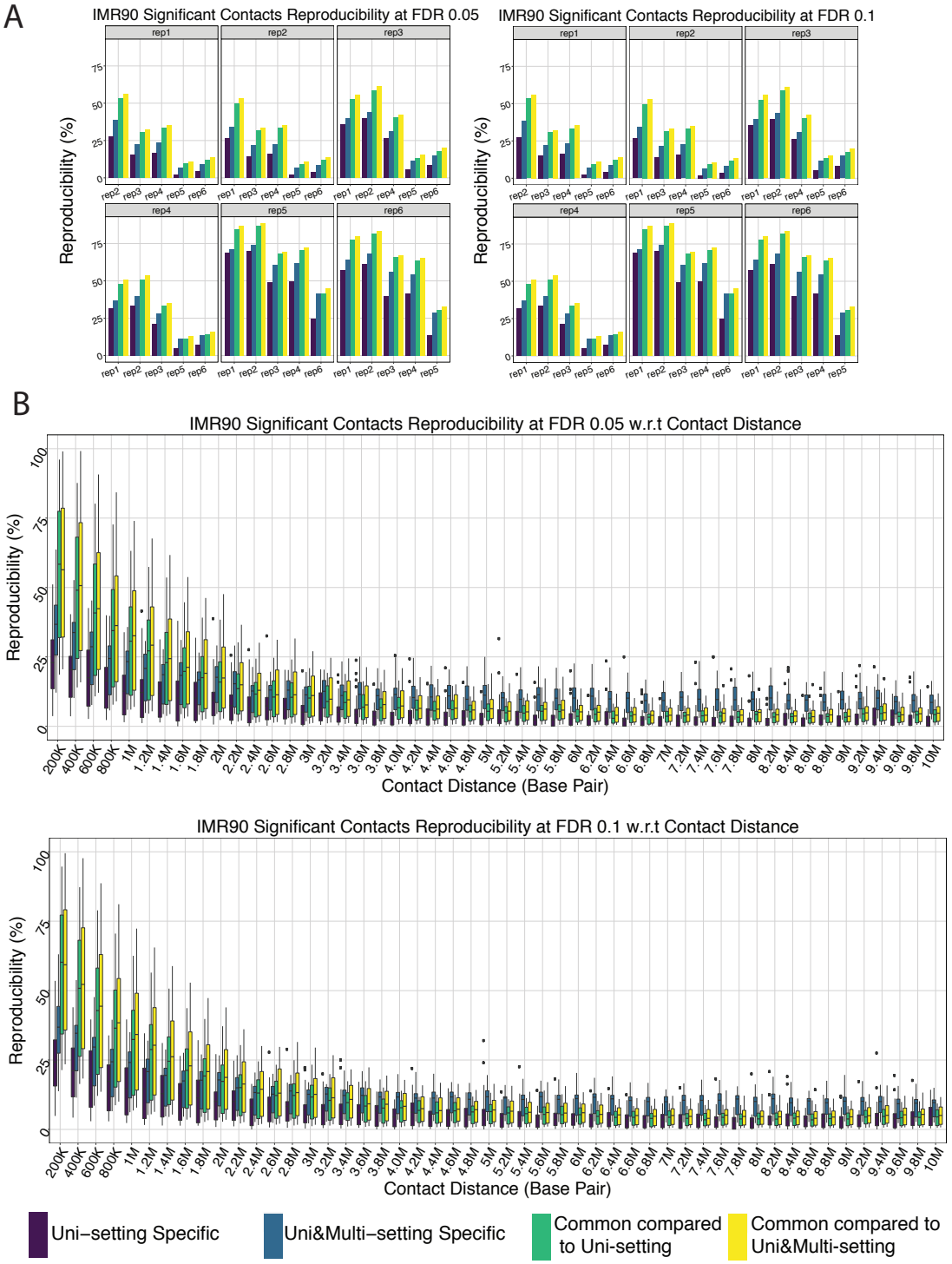



**Figure 2–Figure supplement 10.** Reproducibility of significant interactions for IMR90. **(A)** Significant interactions are classified into three categories: Uni-setting specific or Uni&Multi-setting specific or common to both. Reproducibility is evaluated by the percentage of significant interactions reproduced in another replicate within the same category. **(B)** Reproducibility of significant interactions stratified by genomic distance. Reproducibility is evaluated by the percentage of significant interactions reproduced in another replicate within the same category and genomic distance range.

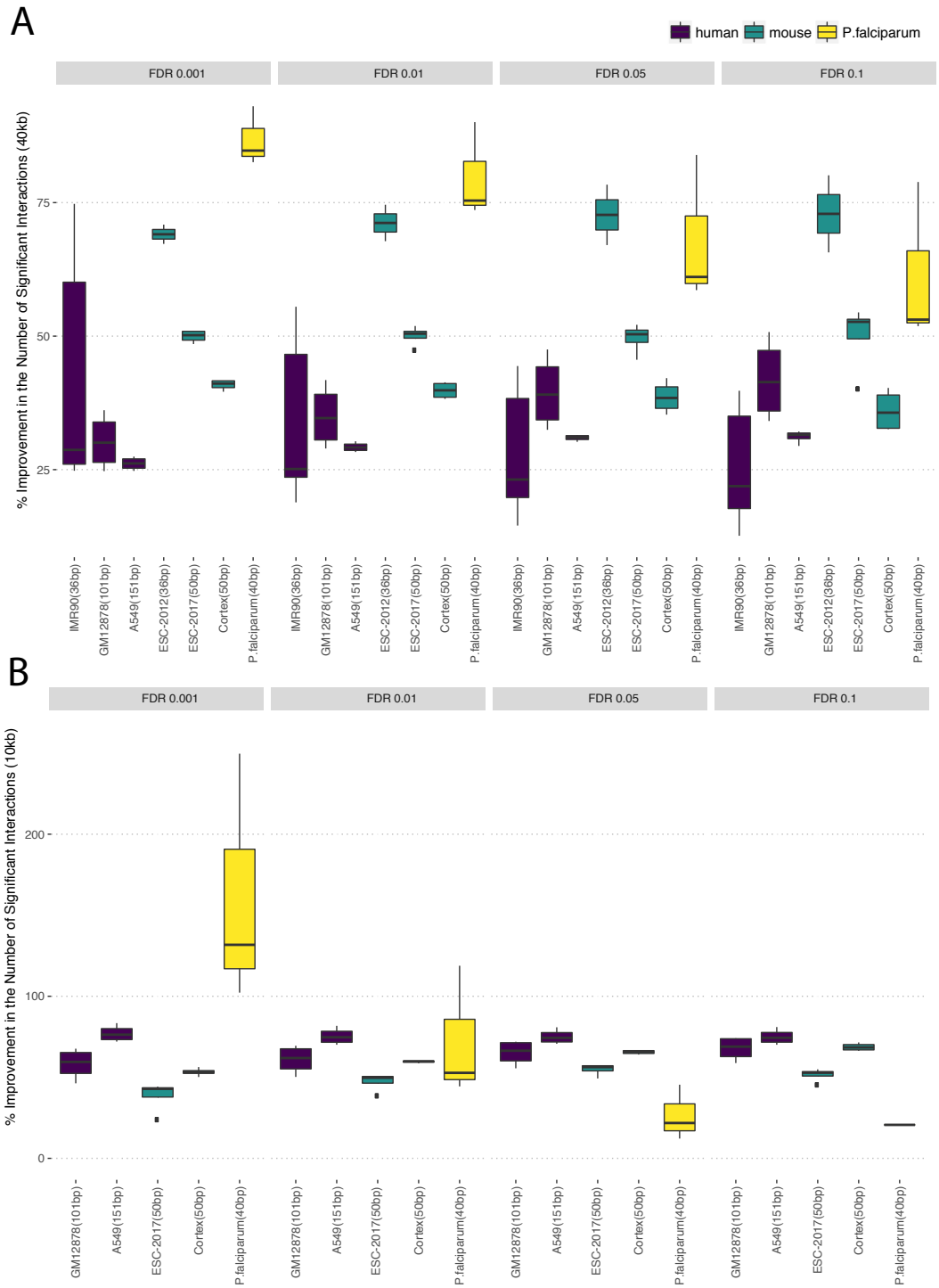



**Figure 3–Figure supplement 1.** Percentage change in the numbers of significant interactions under the Uni&Multi-setting compared to Uni-setting at 0.1%, 1%, 5% and 10% FDR thresholds and resolutions **(A)** 40kb and **(B)** 10Kb.

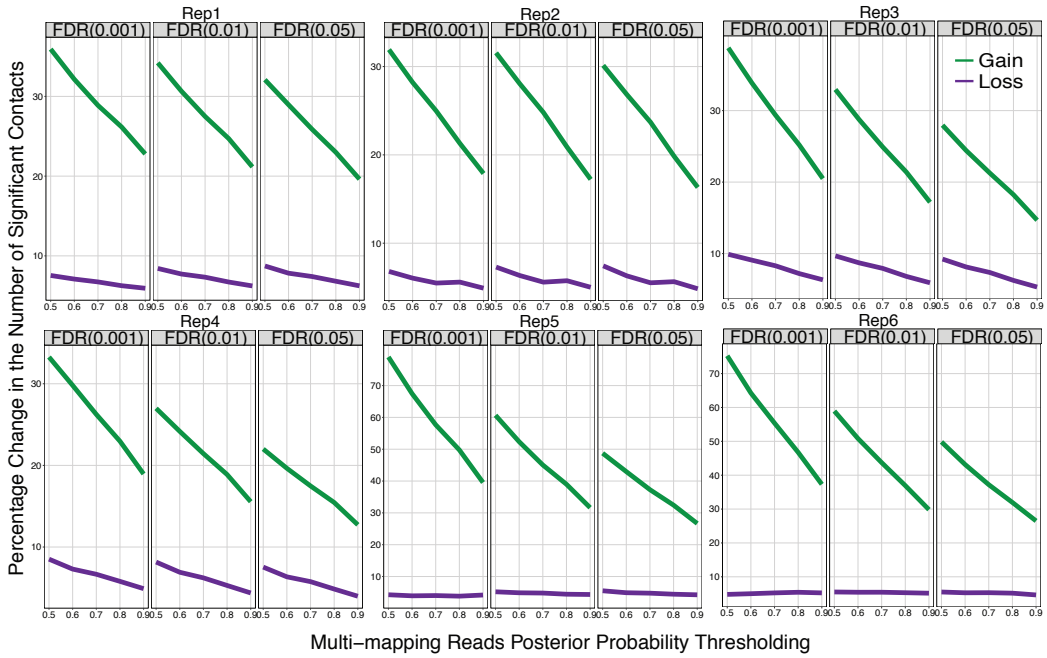



**Figure 3–Figure supplement 2.** Comparison of significant interactions as a function of posterior probabilities of multi-read assignment (IMR90 40kb). Percentage change in the numbers of significant interactions gained (Green) and lost (Purple) by the Uni&Multi-setting compared to the Uni-setting across individual IMR90 replicates for varying FDR and allocation probability thresholds.

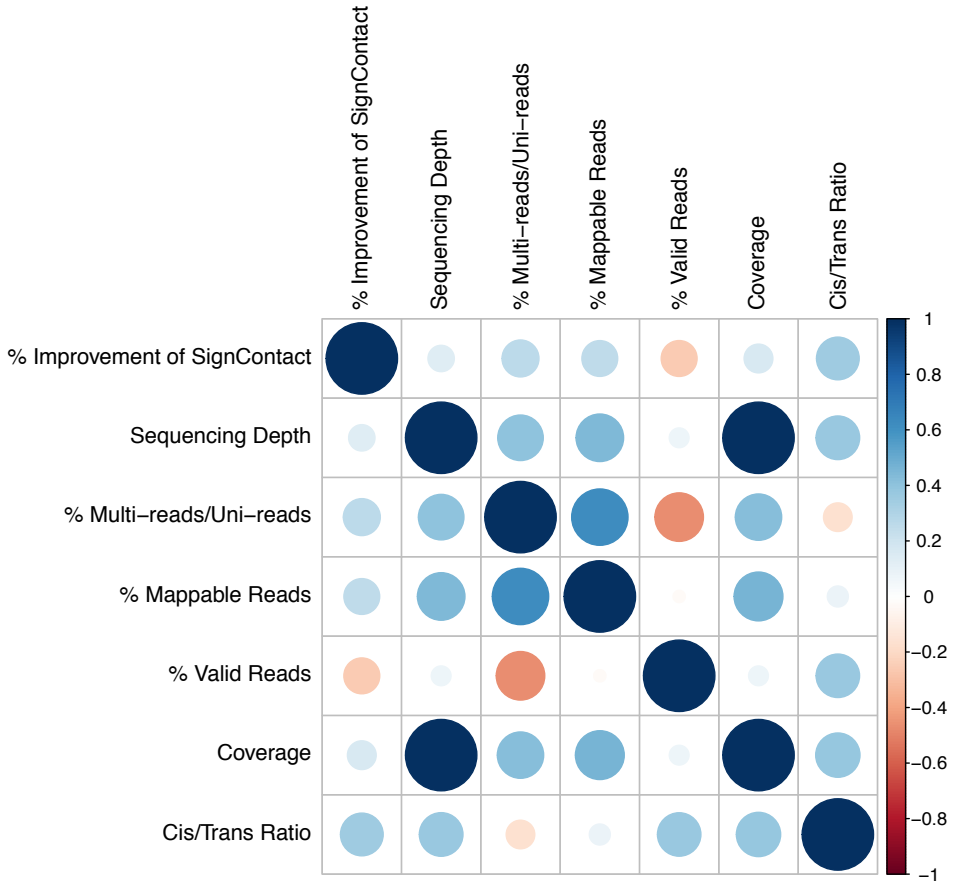



**Figure 3–Figure supplement 3.** Heatmap for marginal correlations of percentage increase in the number of identified significant interactions (FDR 5%) with indicators of data quality across study datasets excluding *P. falciparum* at 40kb. The relative percentage of multi-reads added to uni-reads as well as cis-to-trans ratio are leading impact factors (with p-values $\leq 0.005$) followed by the percentage of mappable reads and valid reads.

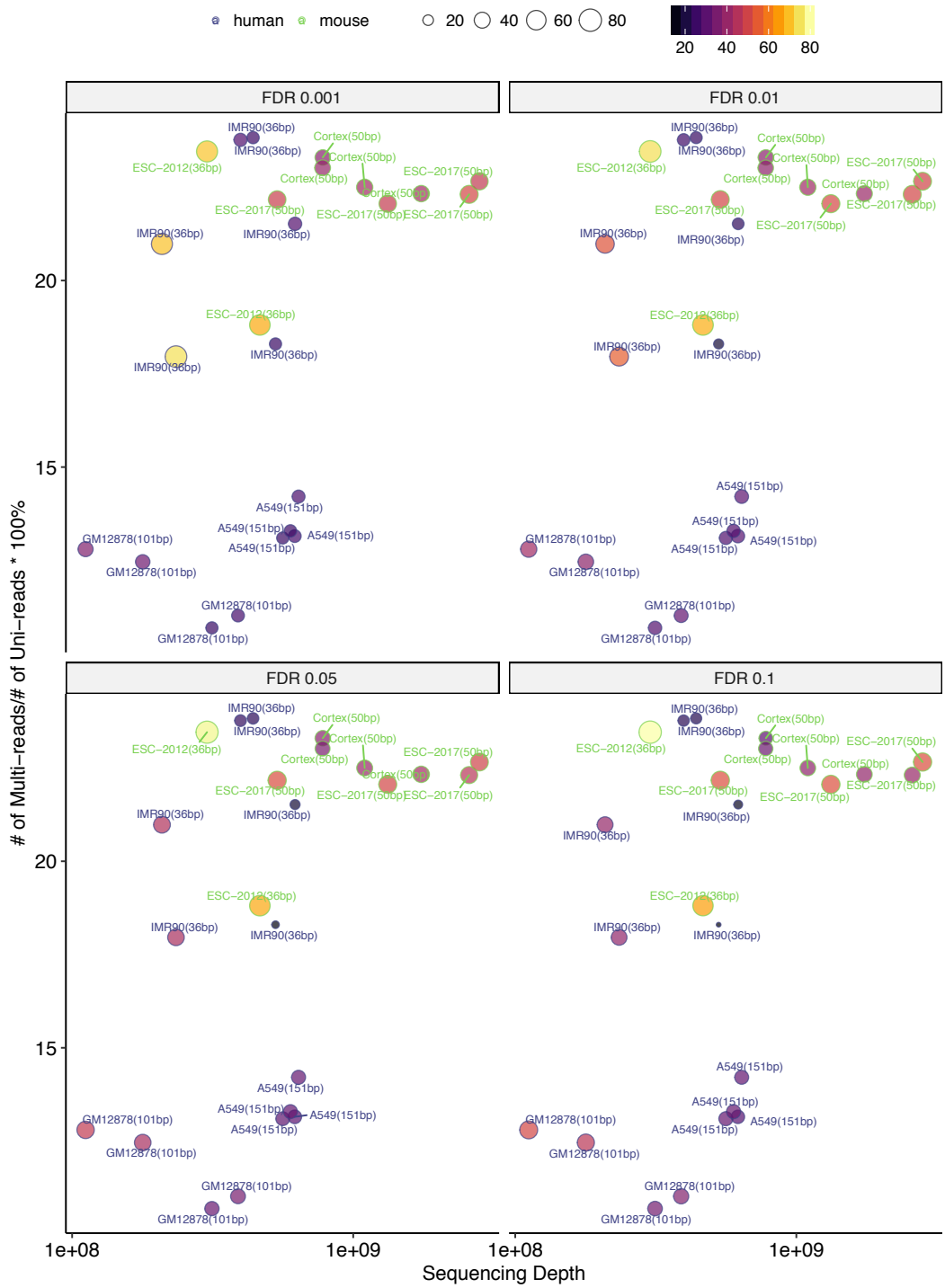



**Figure 3–Figure supplement 4.** Percentage change in the numbers of significant interactions as a function of overall sequencing depth and the percentage of mHi-C rescued multi-reads in comparison to uni-read sets of individual datasets at 40kb.

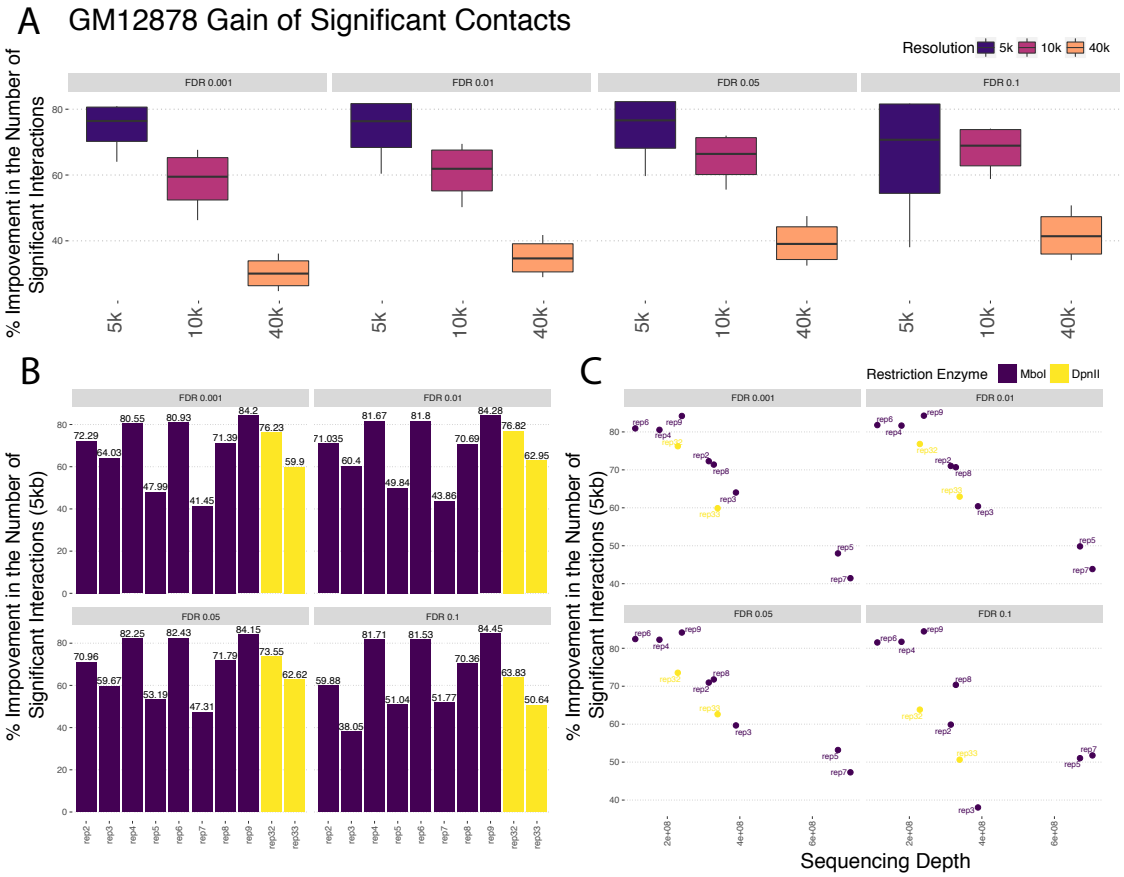



**Figure 3–Figure supplement 5. (A)** Percentage change in the numbers of significant interactions of GM12878 datasets at 5kb, 10kb, 40kb resolutions, respectively, across varying FDR thresholds. **(B)** Percentage change in the numbers of significant interactions of 8 replicates that are based on MboI as restriction enzyme and 2 replicates with DpnII as restriction enzyme at 5kb resolution across different FDR thresholds. **(C)** Percentage change in the numbers of significant interactions with respect to the replicate sequencing depth of all 10 replicates of GM12878 at 5kb resolution across different FDR thresholds.

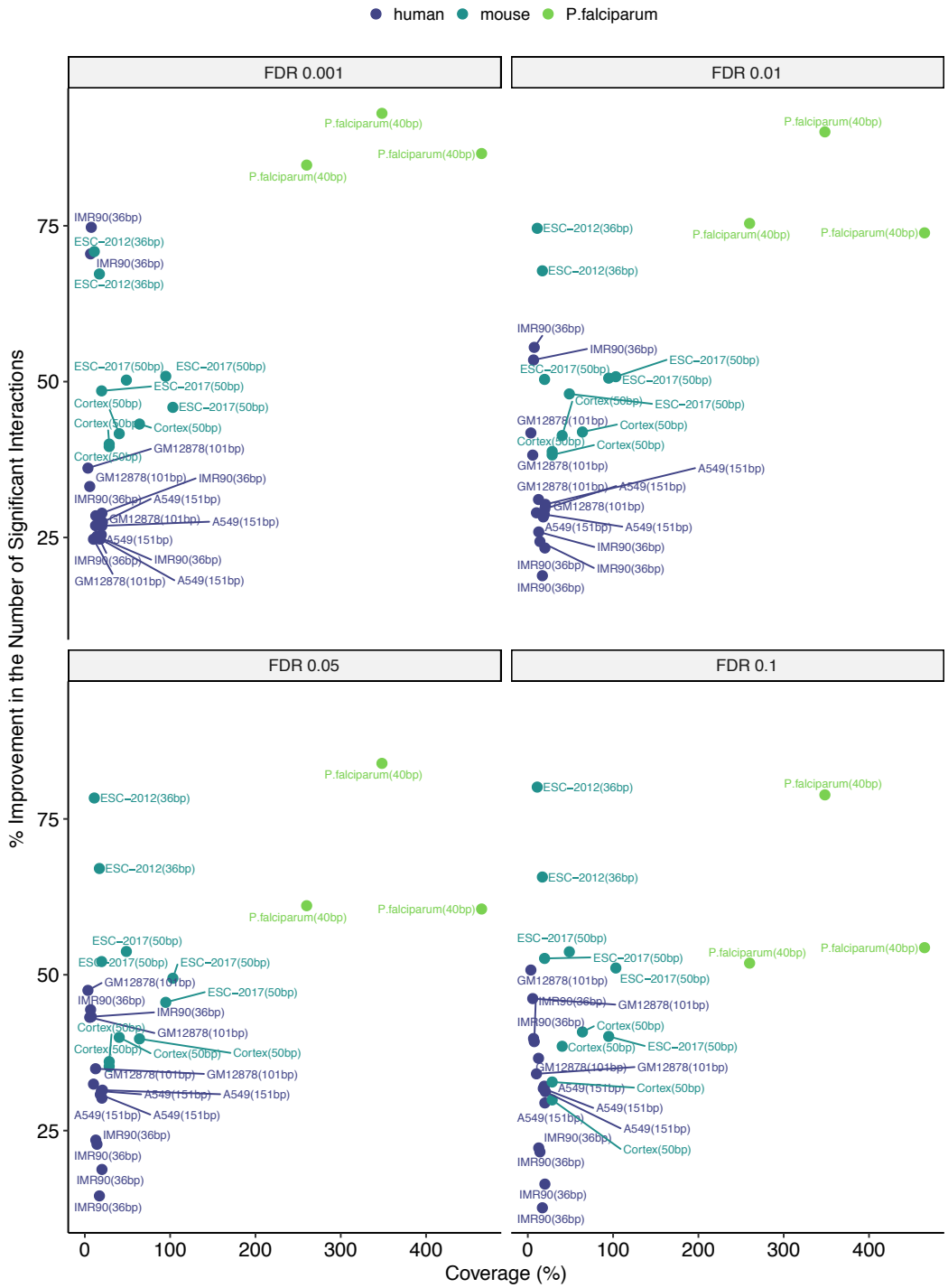



**Figure 3–Figure supplement 6.** Percentage change in the numbers of significant interactions with respect to coverage at 40kb. Coverage is approximated as the ratio of the sequencing depth to the genome size.

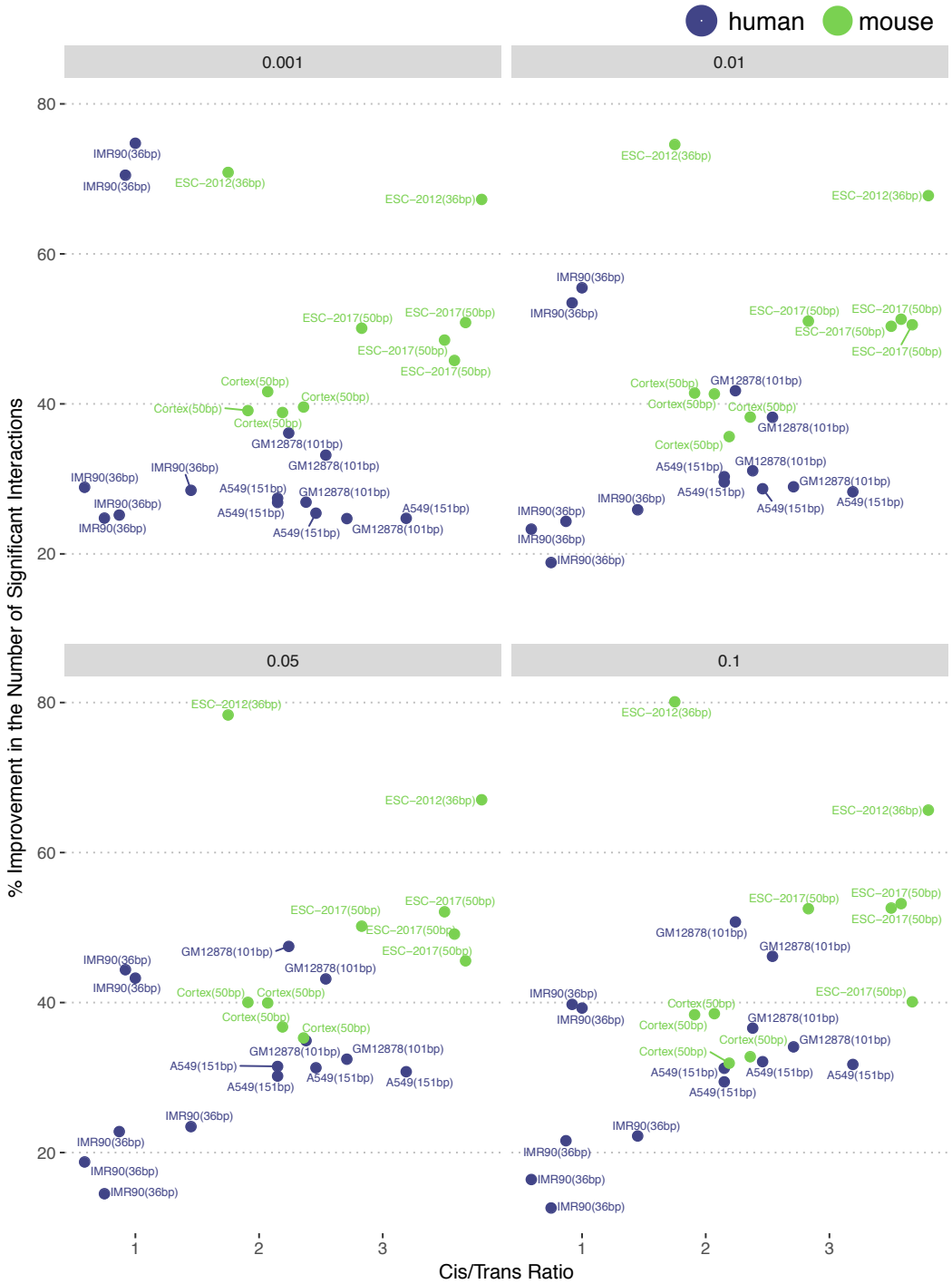



**Figure 3–Figure supplement 7.** Percentage change in the numbers of significant interactions with respect to cis-to-trans ratio excluding *P. falciparum* at 40kb. Cis-to-trans ratio is defined as the the number of valid intra-chromosomal contact counts divided by the number of valid inter-chromosomal contact counts.

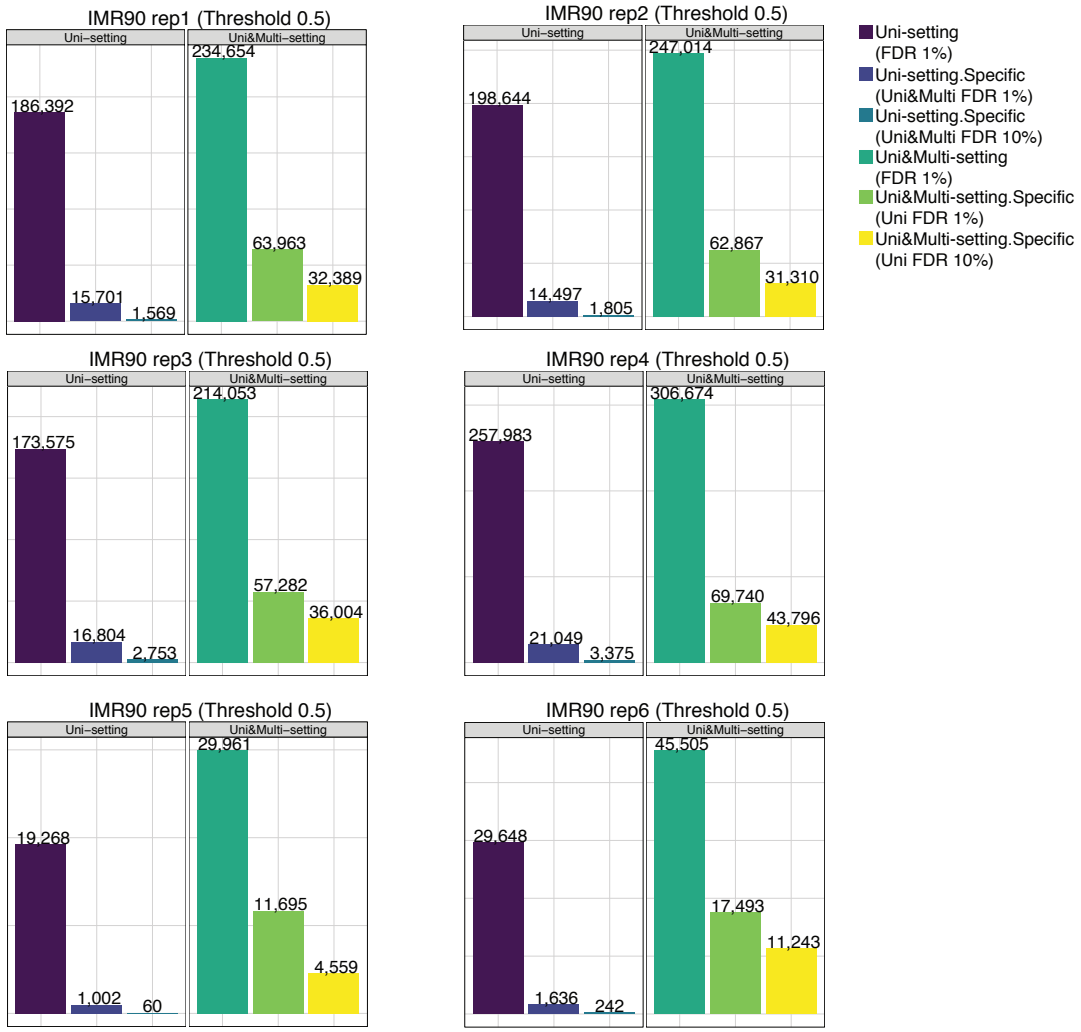



**Figure 3–Figure supplement 8.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of six replicates of IMR90 at 40kb. Uni&Multi-setting. Specific (Uni FDR 10%) is the set of significant interactions identified at 1% FDR by the Uni&Multi-setting but are still unrecoverable by the Uni-setting even with a liberal FDR of 10%. The detailed descriptions of the groups are as follows: Uni-setting (FDR 1%): # of significant interactions identified by the Uni-setting at 1% FDR. Uni&Multi-setting (FDR 1%): # of significant interactions identified by the Uni&Multi-setting at 1% FDR. Uni-setting.Specific (Uni&Multi FDR 1%): # of significant interactions identified by Uni-setting (FDR 1%) but not by Uni&Multi-setting at 1% FDR . Uni-setting.Specific (Uni&Multi FDR 10%): # of significant interactions identified by Uni-setting (FDR 1%) but not by Uni&Multi-setting at 10% FDR. Uni&Multi-setting.Specific (Uni FDR 1%): # of significant interactions identified by Uni&Multi-setting (FDR 1%) but not by Uni-setting at 1% FDR . Uni&Multi-setting.Specific (Uni FDR 10%): # of significant interactions identified by Uni&Multi-setting (FDR 1%) but not by Uni-setting at 10% FDR.

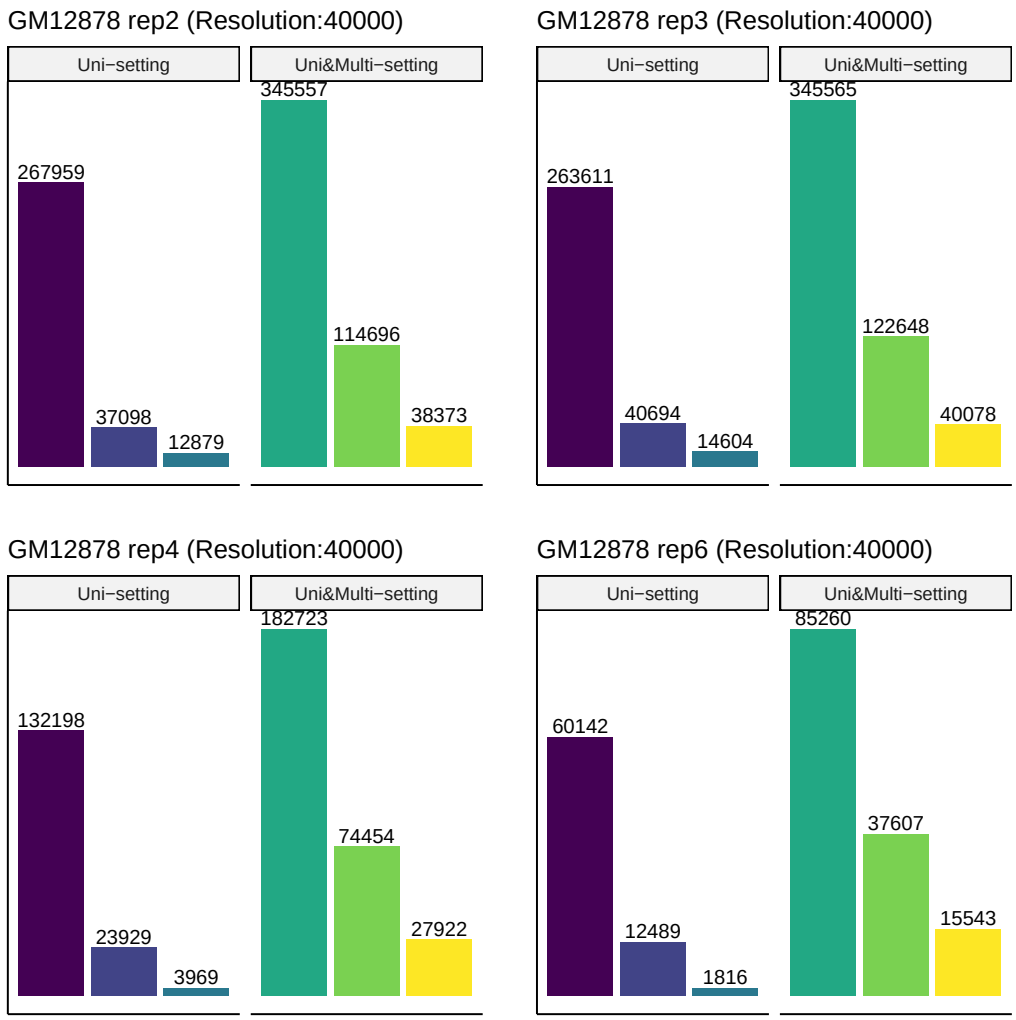



**Figure 3–Figure supplement 9.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of four replicates of GM12878 at 40kb resolution. Color labels are the same as Figure 8

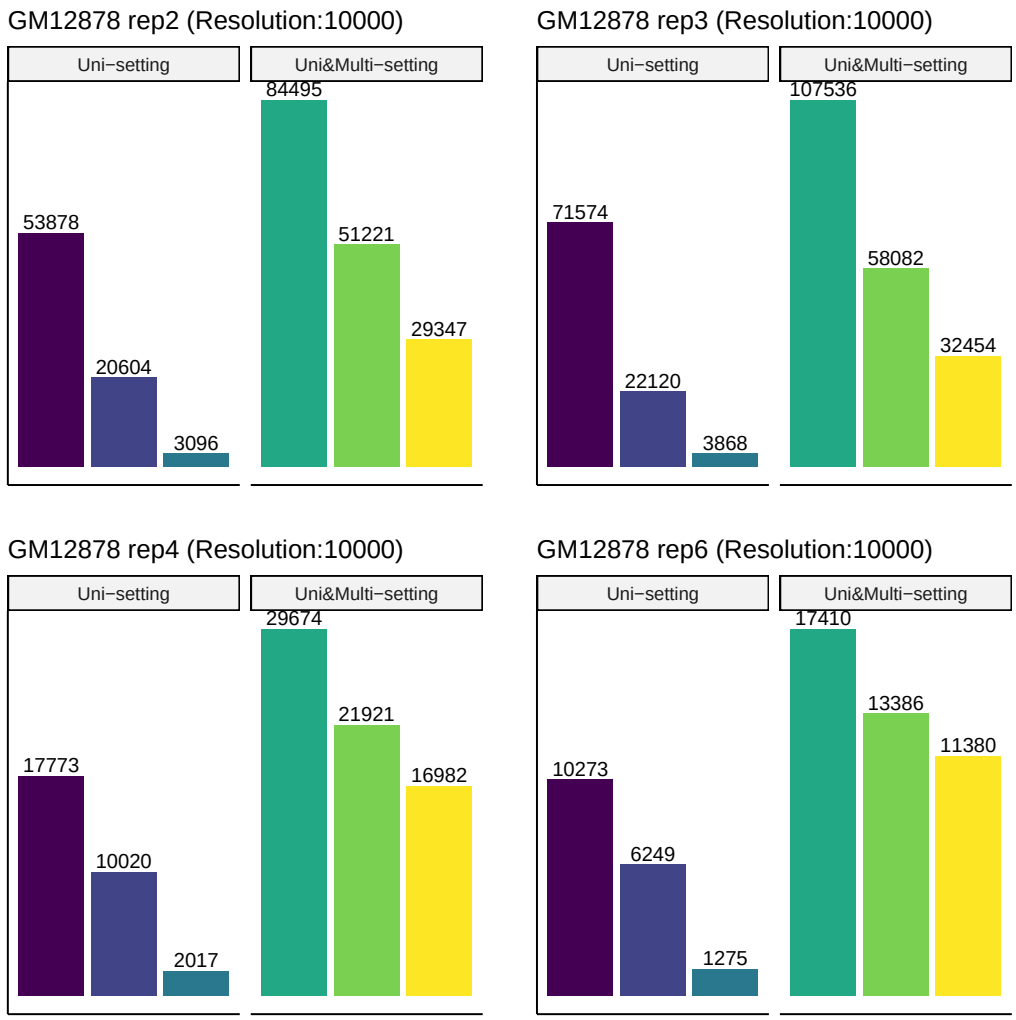



**Figure 3–Figure supplement 10.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of four replicates of GM12878 at 10kb resolution. Color labels are the same as Figure 8

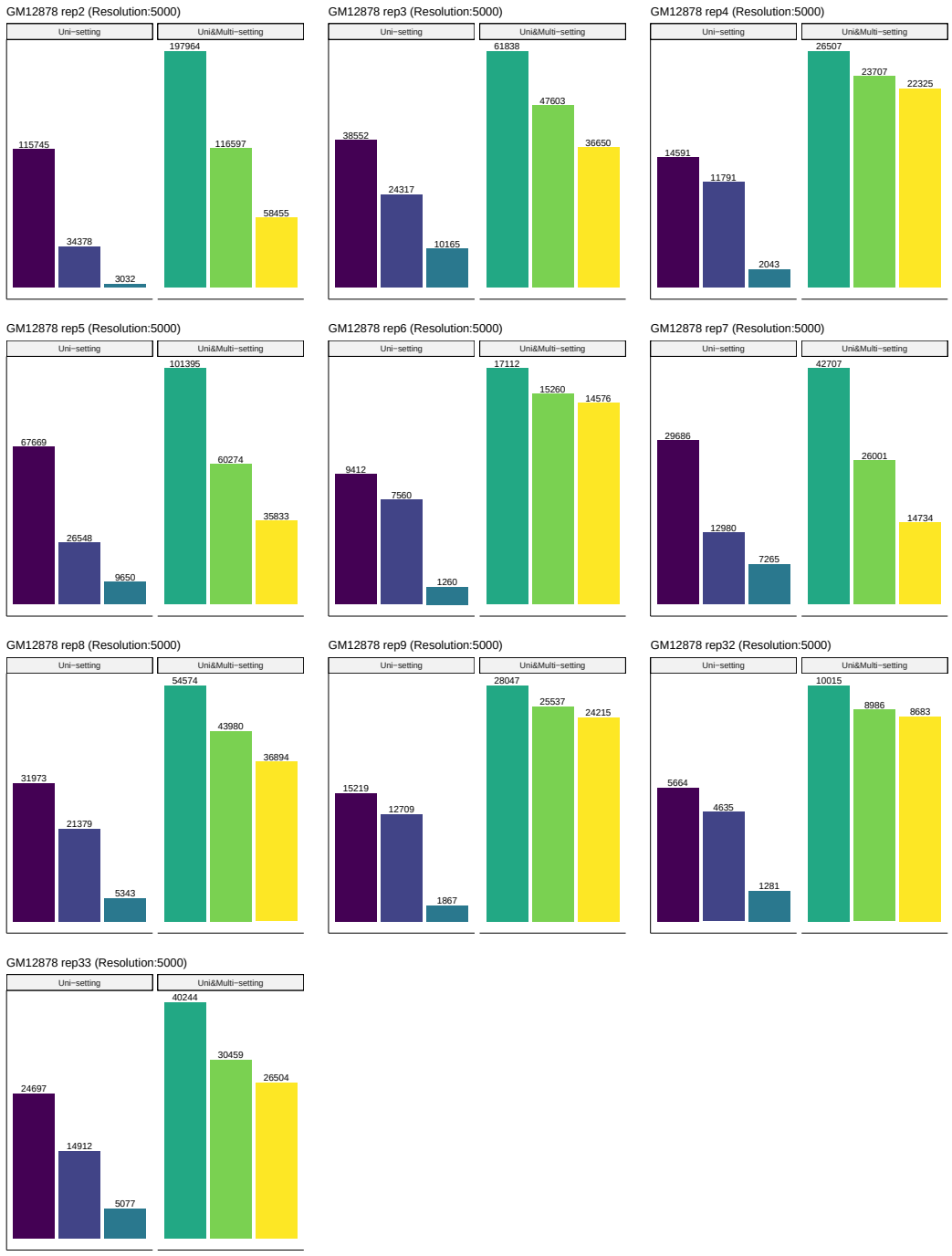



**Figure 3–Figure supplement 11.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for each of ten replicates of GM12878 at 5kb resolution. Color label is the same as Figure 8

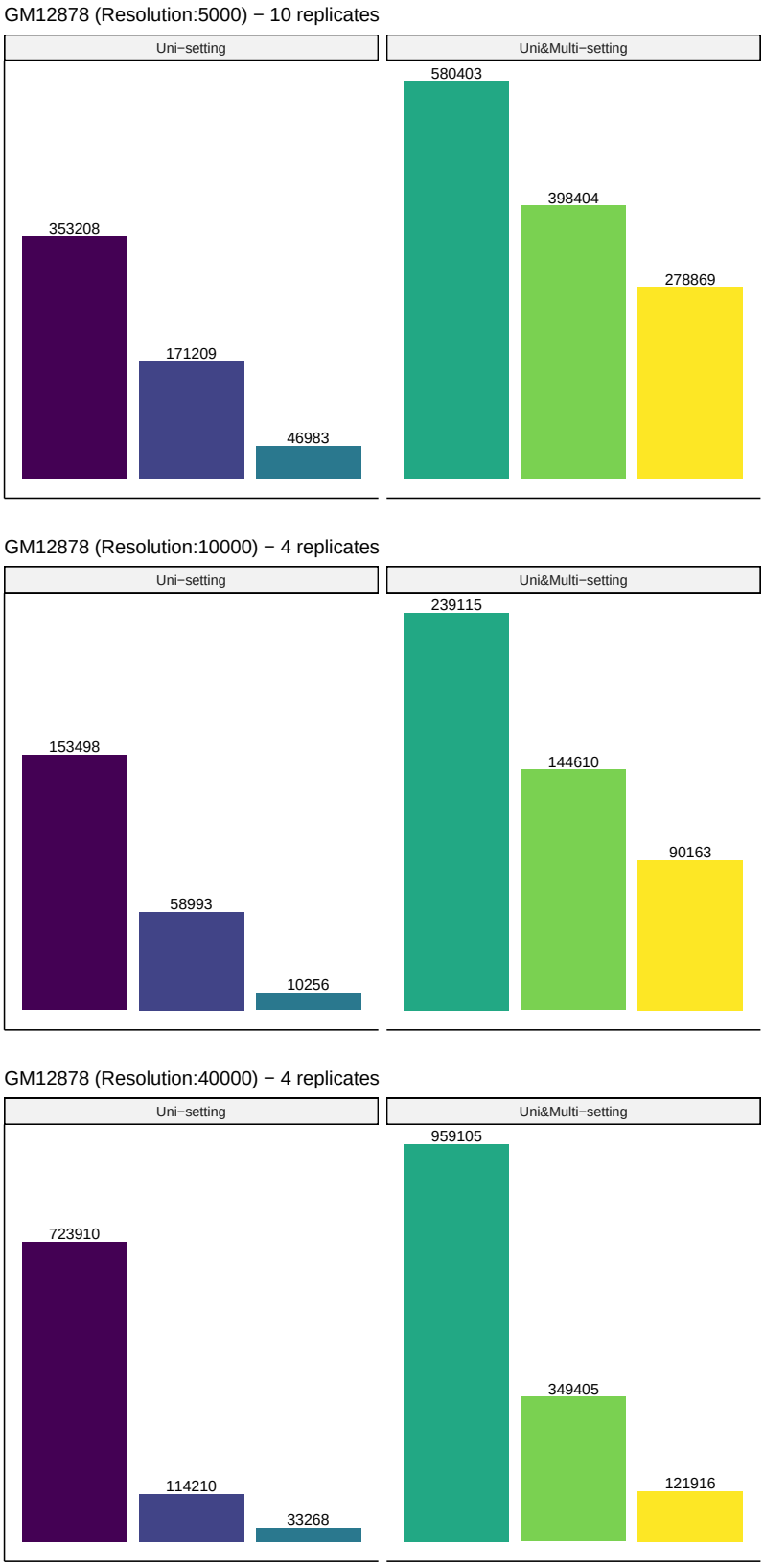



**Figure 3–Figure supplement 12.** Recovery of significant interactions identified at FDR 1% by analysis at FDR 10% for GM12878 summed across replicates at 5kb, 10kb, and 40kb resolutions. Color label is the same as Figure 8

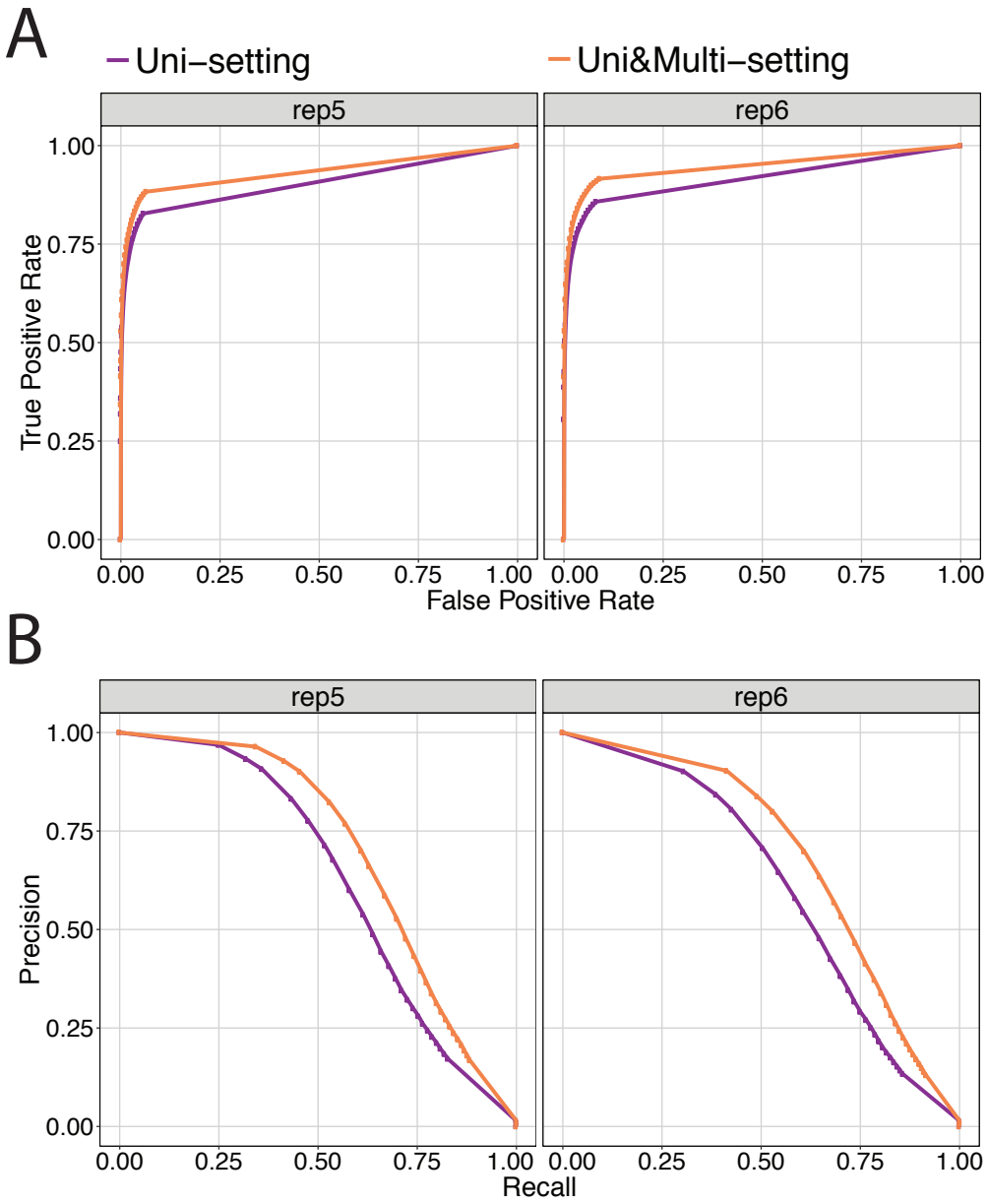



**Figure 3–Figure supplement 13.** ROC and PR curves for replicates 5 and 6 of IMR90. Sets of "True" interactions and "True" non-interactions are defined by reproducible significant/insignificant interactions across replicate 1-4 of both Uni-setting and Uni&Multi-setting (See Materials and Methods). significant interactions of replicates 5 and 6 are utilized to compare ROC **(A)** and PR **(B)** curves among the Uni- and Uni&Multi-settings.

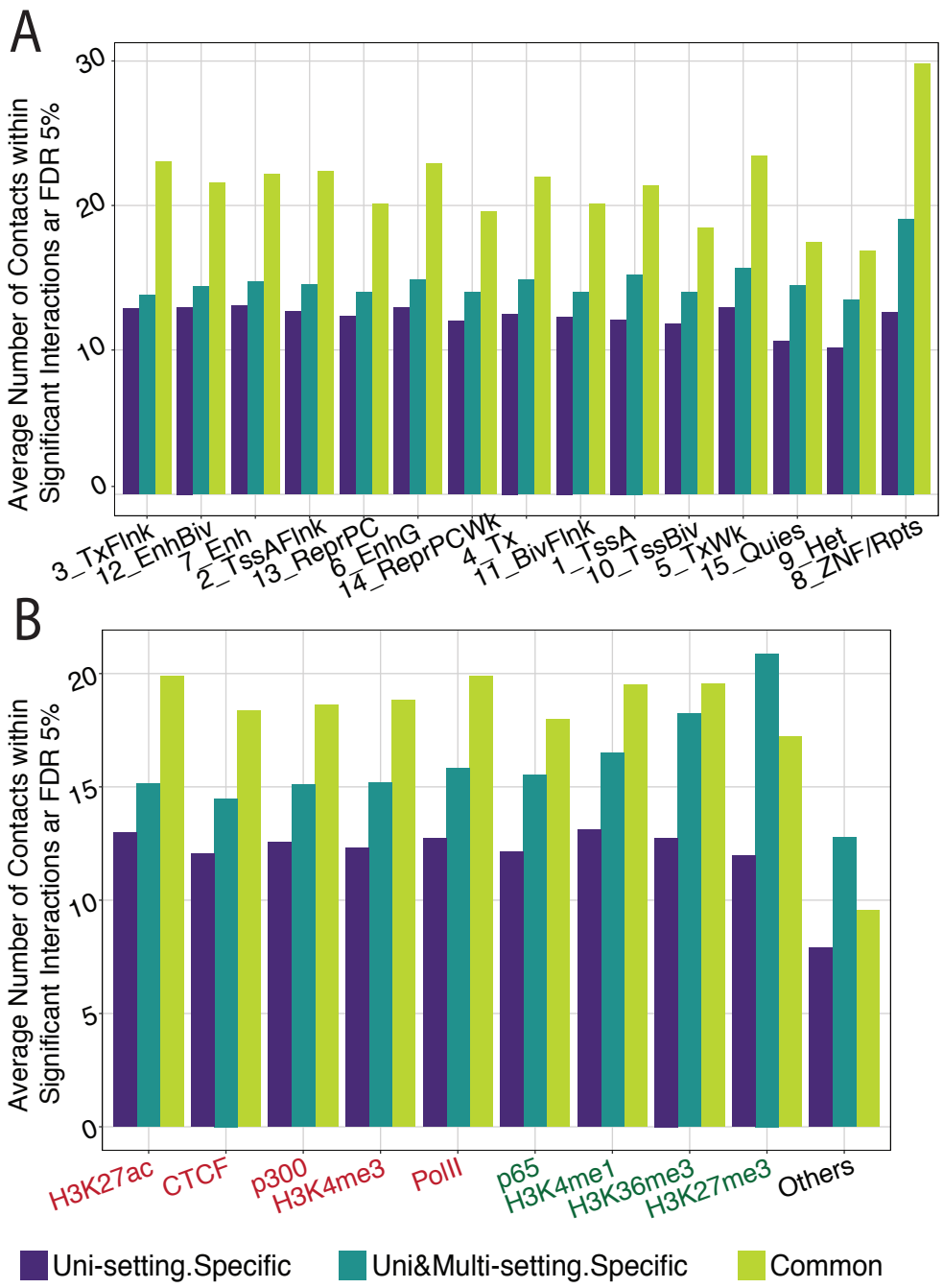



**Figure 3–Figure supplement 14.** Quantification of significant interactions for chromHMM states and ChIP-seq peak regions (IMR90). **(A)** Grouping the significant interactions into three groups (Uni-setting specific, Uni&Multi-setting specific, Common to both settings) reveals the largest enrichment differences in chromHMM annotation categories related to repetitive regions, such as Zinc Finger Genes & Repeats as well as Heterochromatin. **(B)** Average number of significant interactions across regions with a variety of ChIP-seq signals. Red/green labels denote smaller/larger differences between Uni-setting.Specific and Uni&Multi-setting.specific compared to the differences observed in the "Others" category that depicts non-peak regions.

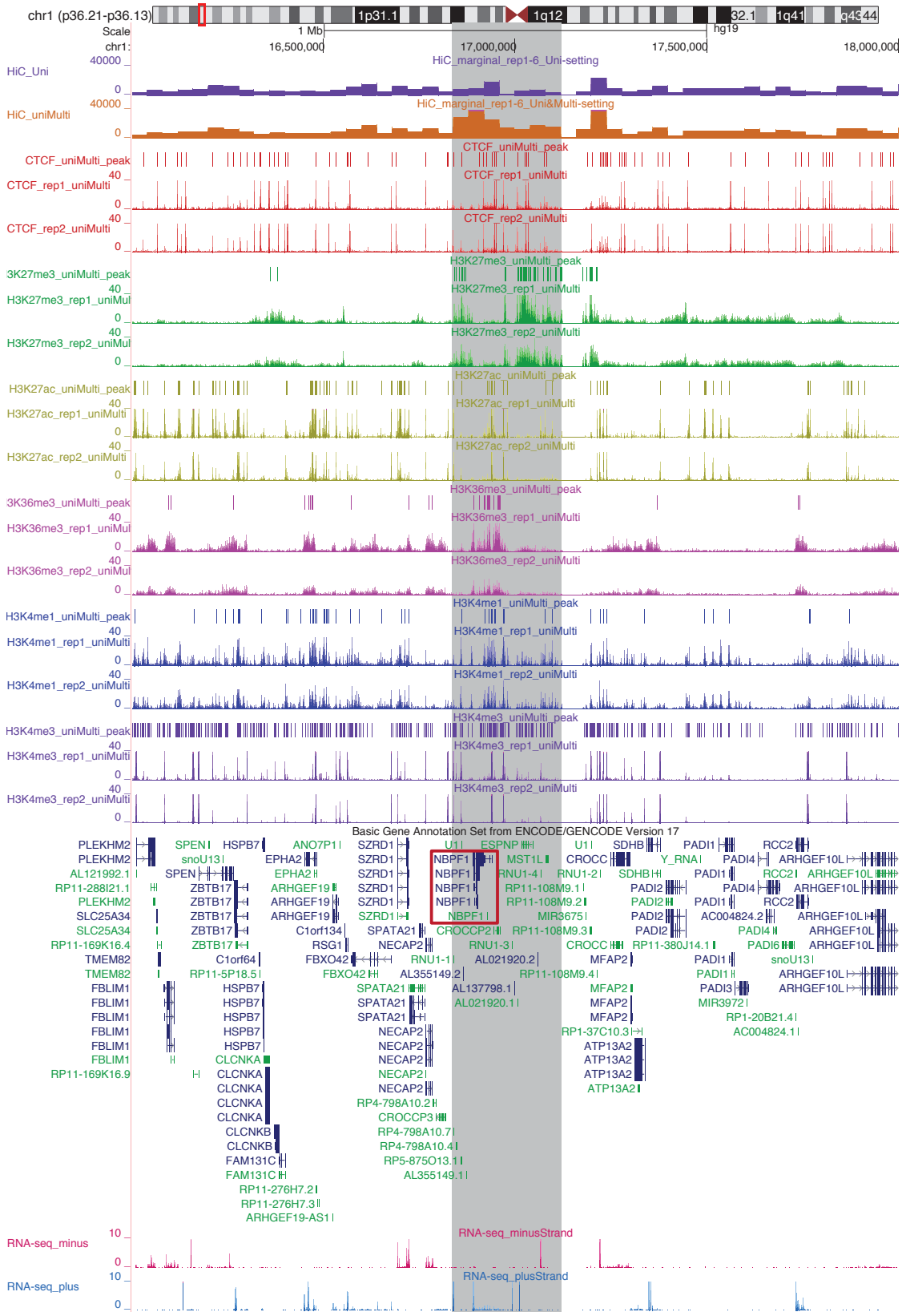



**Figure 3–Figure supplement 15.** Marginalized Hi-C signal (contact counts aggregated across the genomic coordinates for six replicates of IMR90), ChIP-seq coverage and peaks and gene expression for chr1:16,000,000-18,000,000. Highlighted in grey is a region with significantly different marginal Hi-C signal between Uni-setting and Uni&Multi-setting.

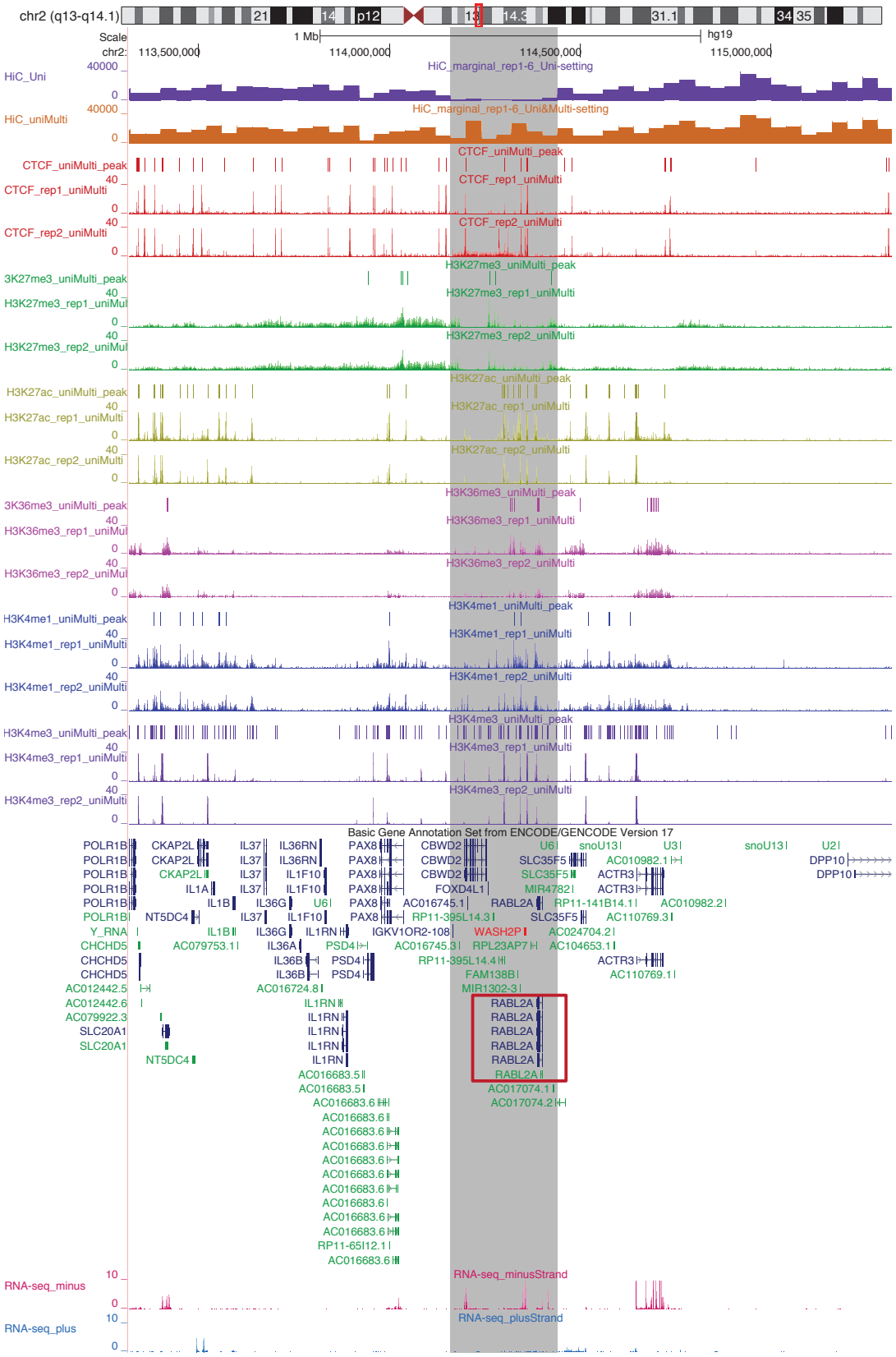



**Figure 3–Figure supplement 16.** Marginalized Hi-C signal (contact counts aggregated across the genomic coordinates for six replicates of IMR90), ChIP-seq coverage and peaks and gene expression for chr2:113460,000-116,000,000. Highlighted in grey is a region with significantly different marginal Hi-C signal between Uni-setting and Uni&Multi-setting.

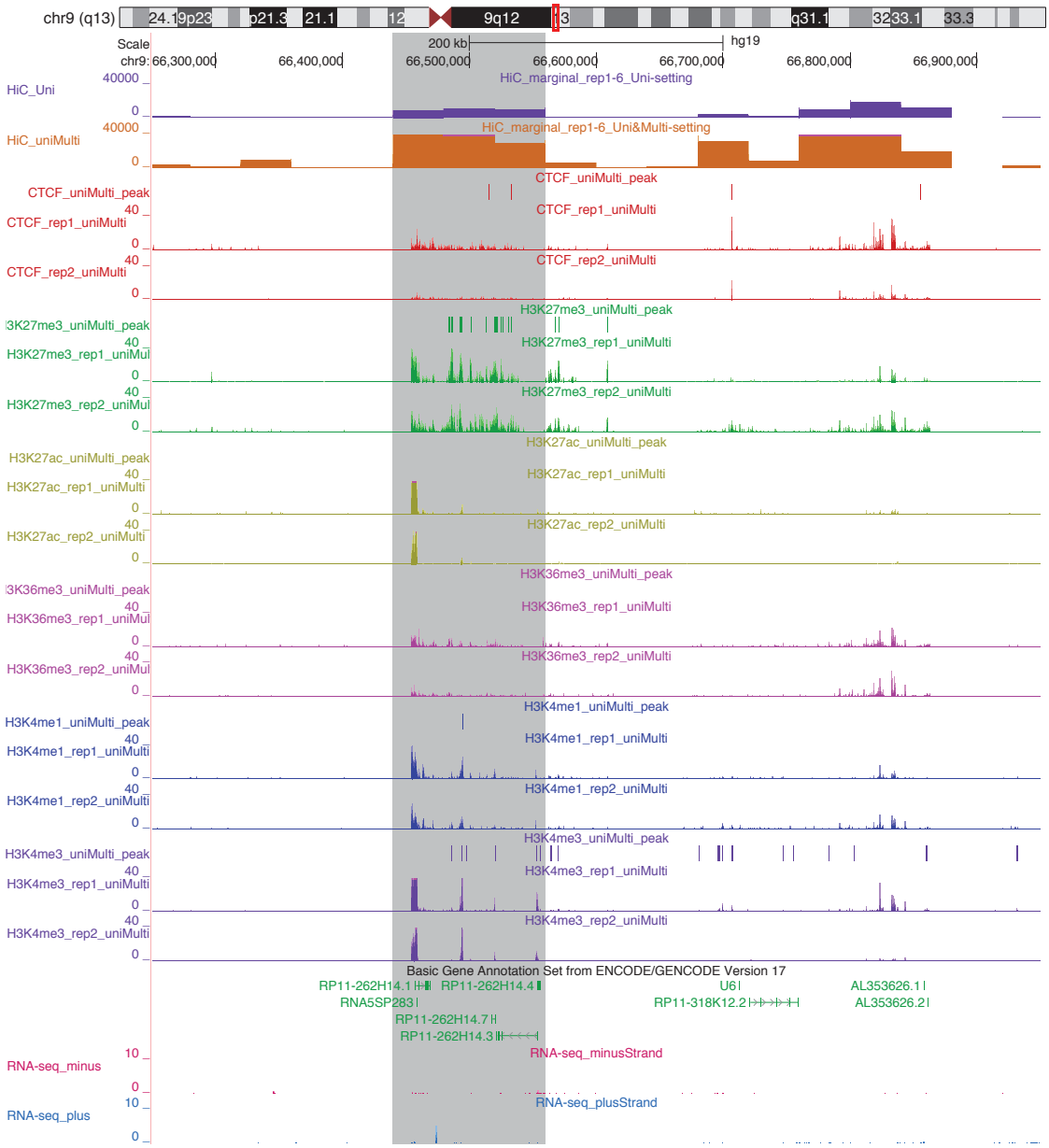



**Figure 3–Figure supplement 17.** Marginalized Hi-C signal (contact counts aggregated across the genomic coordinates for six replicates of IMR90), ChIP-seq coverage and peaks and gene expression for chr9:66,250,000-66,950,000. Highlighted in grey is a region with significantly different marginal Hi-C signal between Uni-setting and Uni&Multi-setting.

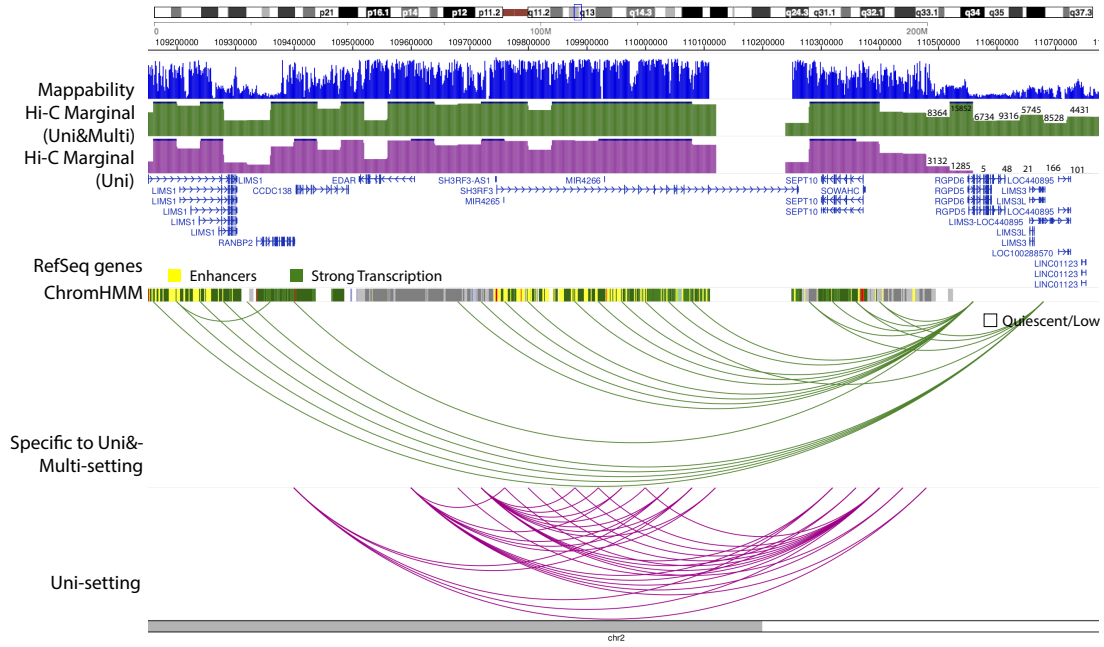



**Figure 4–Figure supplement 1.** Examples of significant promoter-enhancer interactions reproducible among 6 replicates under Uni- and Uni&Multi-settings (IMR90) on chromosome 2.

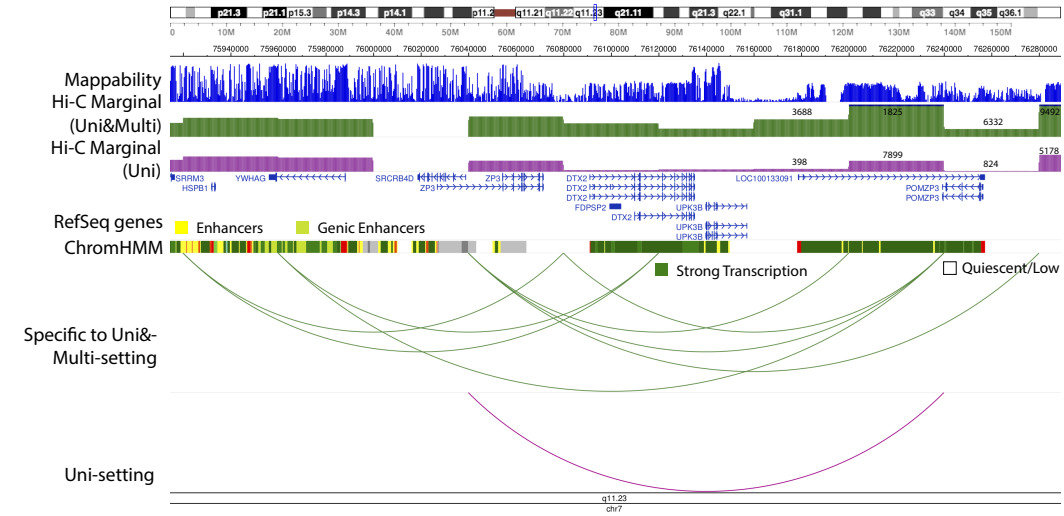



**Figure 4–Figure supplement 2.** Examples of significant promoter-enhancer interactions reproducible among 6 replicates under Uni- and Uni&Multi-settings (IMR90) on chromosome 7.

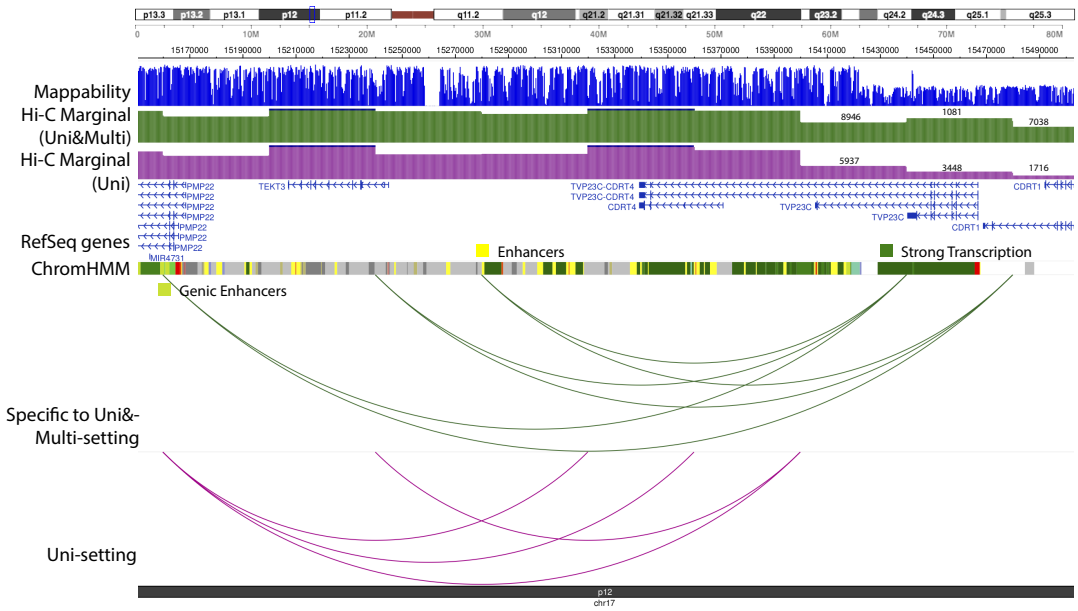



**Figure 4–Figure supplement 3.** Examples of significant promoter-enhancer interactions reproducible among 6 replicates under Uni- and Uni&Multi-settings (IMR90 ) on chromosome 17.

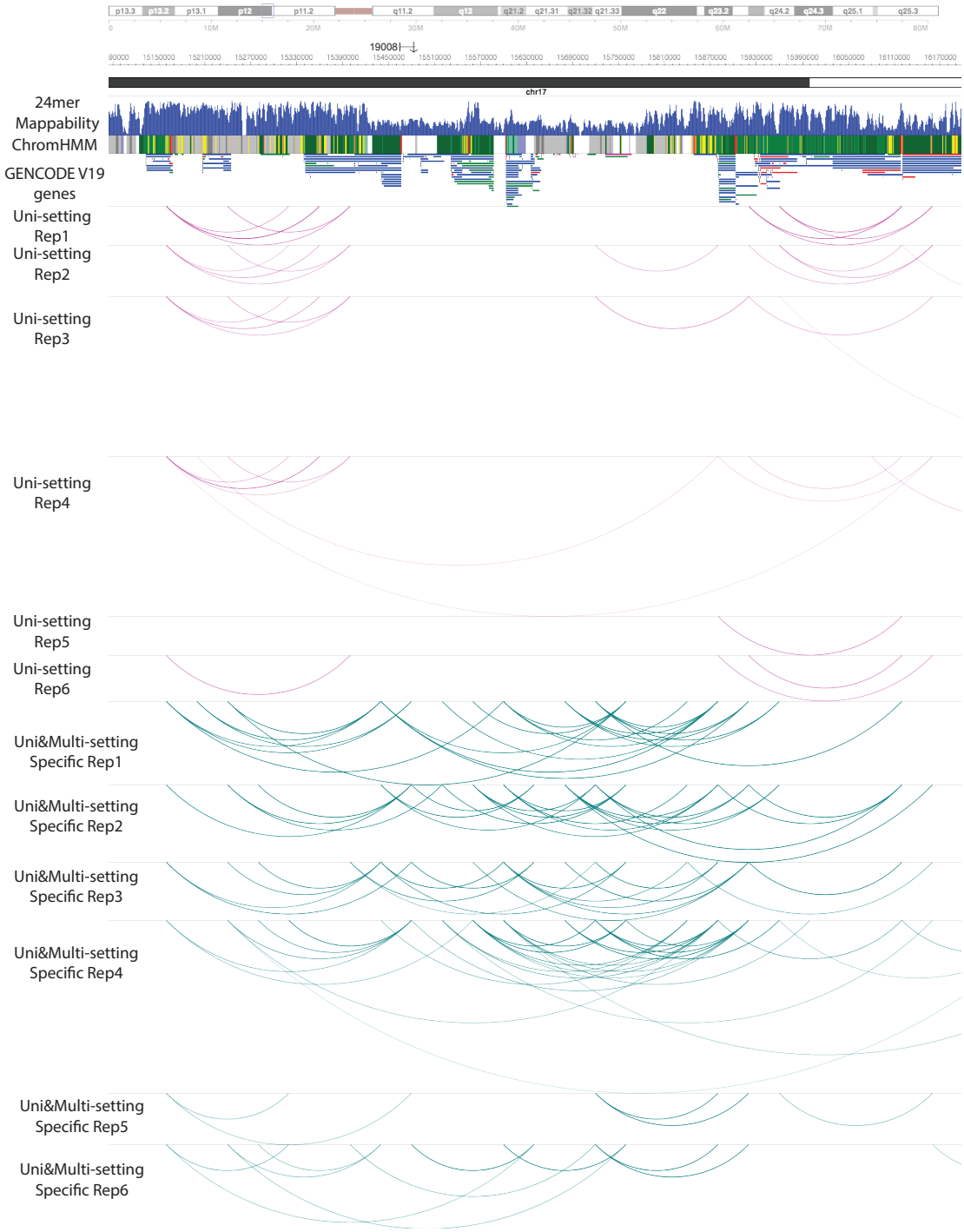



**Figure 4–Figure supplement 4.** Significant promoter-emhancer interactions under Uni- and Uni&Multi-settings across 6 IMR90 replicates (Chromosome 17). This is the individual replicate level data for Figure 4–Figure supplement 4 in a large genome region.

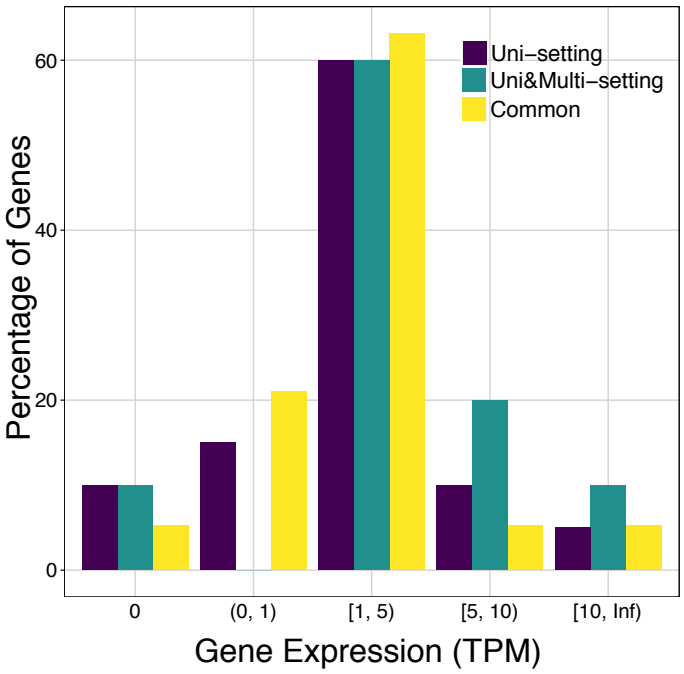



**Figure 4–Figure supplement 5.** Expression distribution of genes promoters of which have significant promoter interactions (IMR90). Genes harboring significant interactions (at 5% FDR) within their promoters are grouped into different gene expression categories.

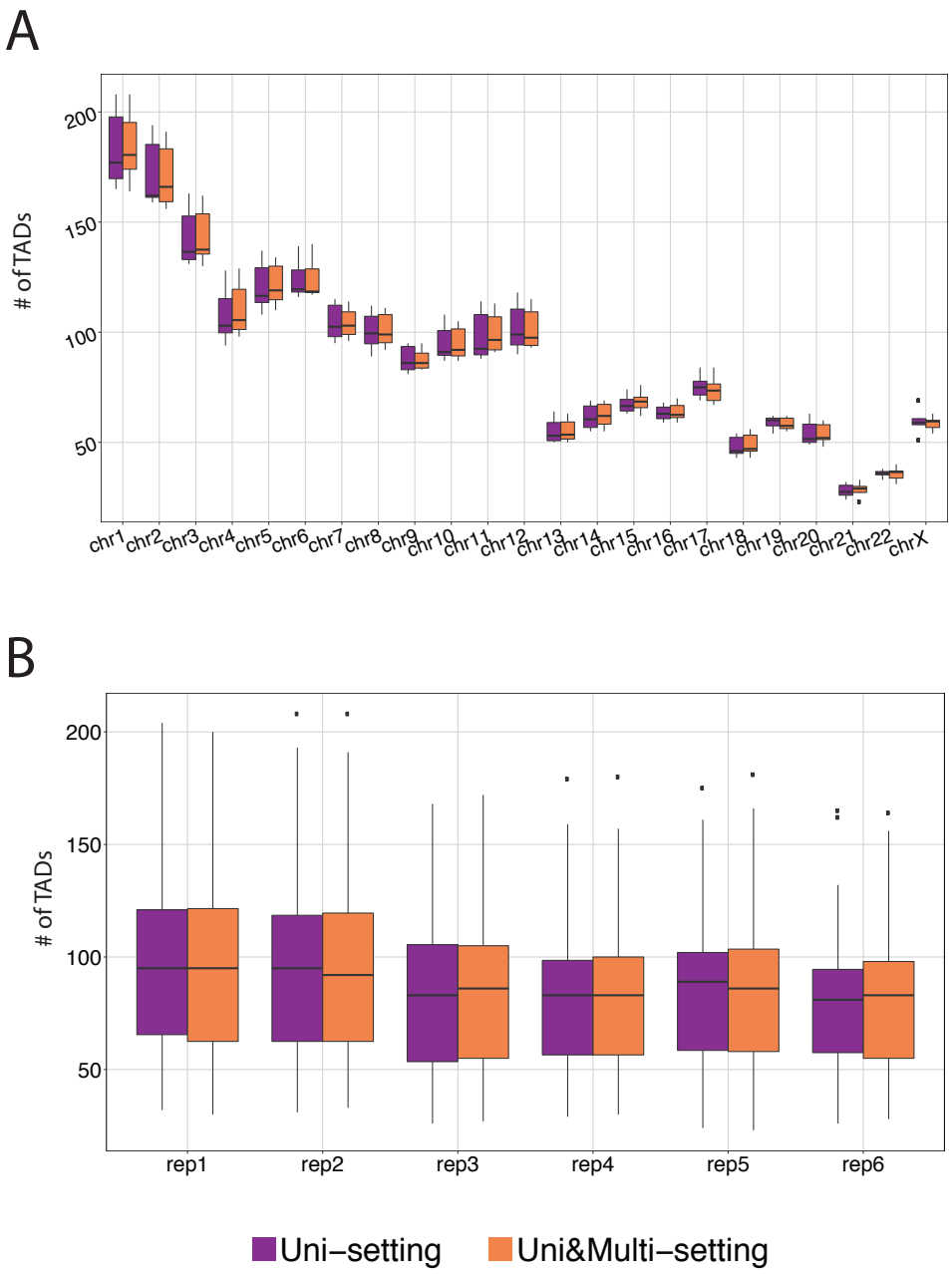



**Figure 5–Figure supplement 1.** The number of topologically associating domains (TADs) detected in each chromosome under Uni-setting and Uni&Multi-setting (IMR90). **(A)** Total number of TADs identified across six replicates for each chromosome. **(B)** Total number of TADs identified across 23 chromosomes for each replicate.

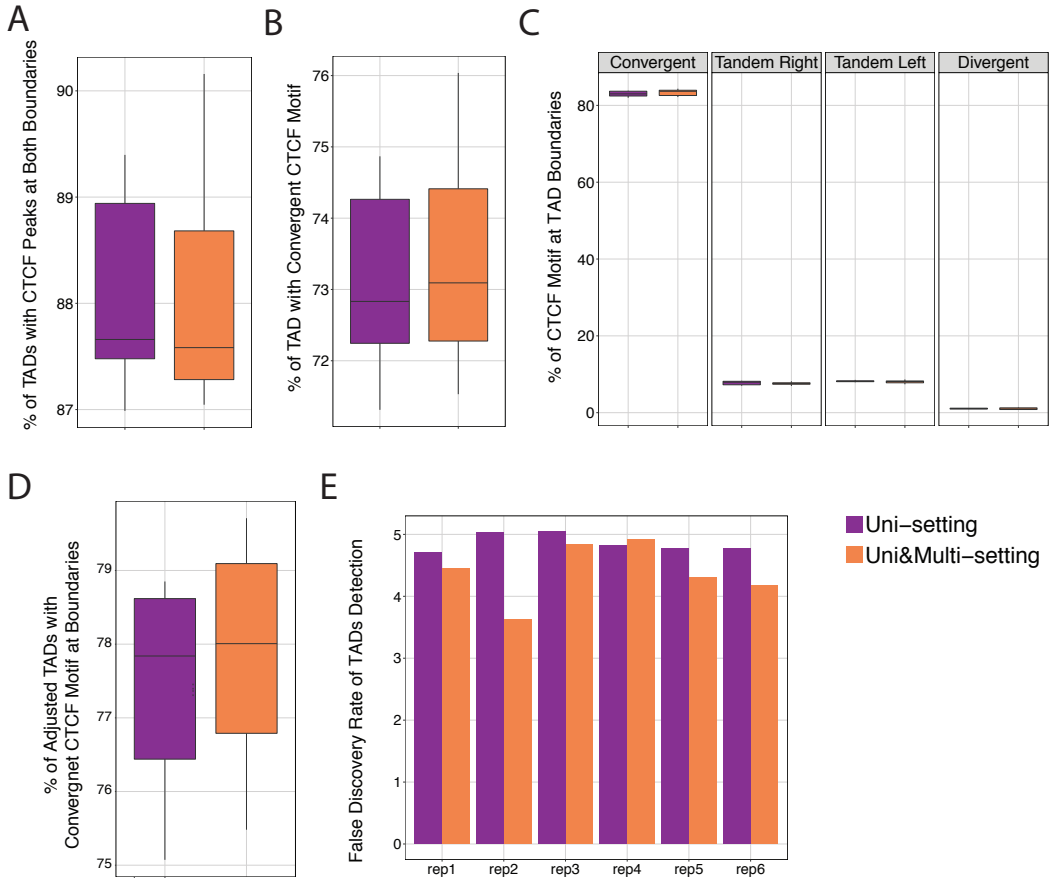



**Figure 5–Figure supplement 2.** Comparison of CTCF peaks at the boundaries of topologically associating domains (TADs) under Uni-setting and Uni&Multi-setting across six replicates of IMR90. **(A)** Percentages of TADs that have CTCF peaks at boundaries. **(B)** Percentages of TADs that have both CTCF peaks and convergent CTCF motifs at the boundaries. **(C)** Percentages of four types of CTCF motifs orientations at TAD boundaries. Convergent motif pairs are those with a forward strand motif upstream of TAD boundaries and a reverse strand motif downstream of TAD boundaries. Tandem Right, similarly, represents forward-forward CTCF motif pairs. Tandem Left refers to reverse-reverse motif pairs. Divergent is reverse-forward motif pairs **(D)** Some TAD boundaries are adjusted under Uni&Multi-setting. Box plots depict the percentage of adjusted TADs that have convergent CTCF motifs at the boundaries. **(E)**. False discovery rate of TADs detected under the two settings. TADs that are not reproducible and lack CTCF convergent motifs at the TAD boundaries are considered as false positives.

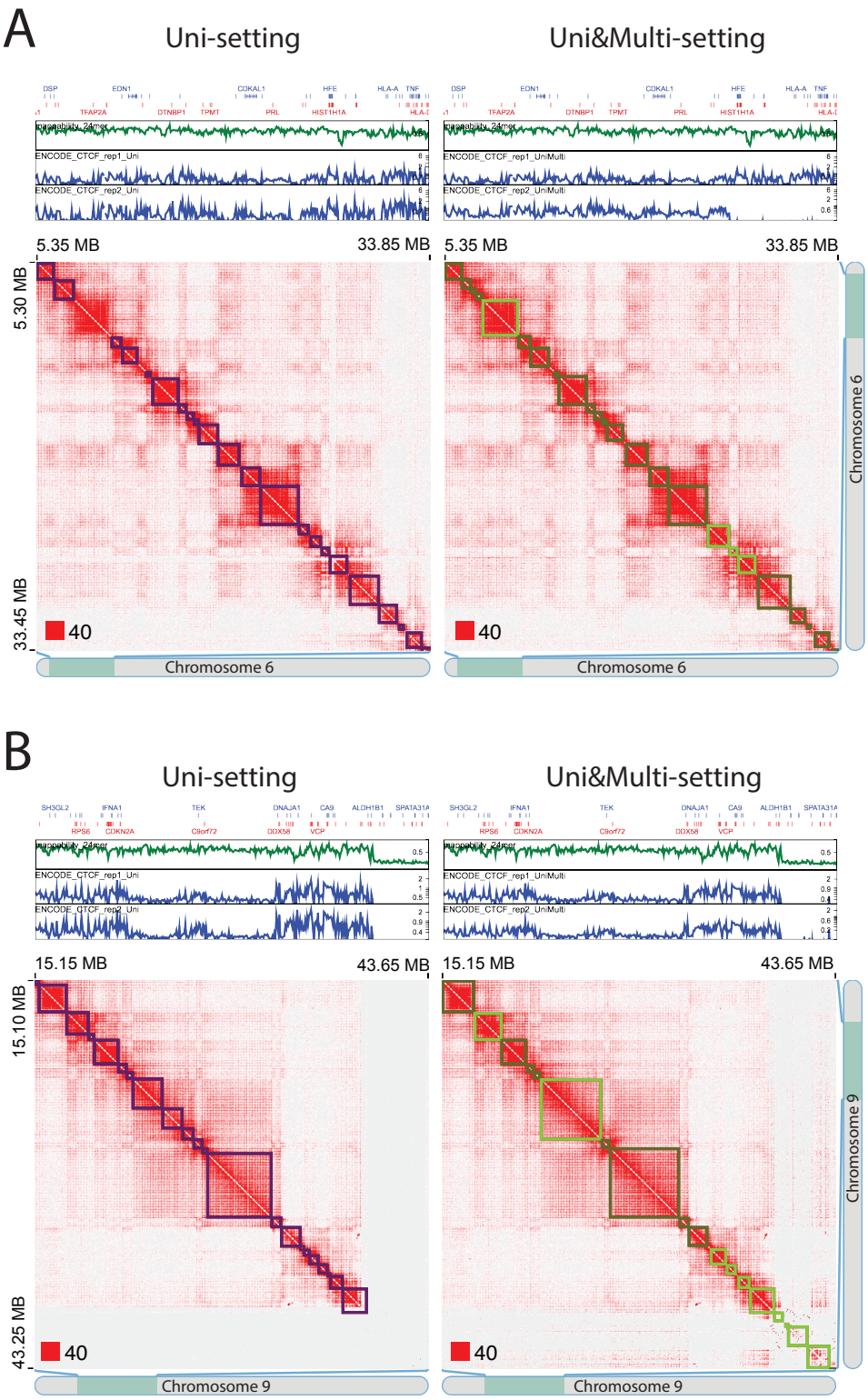



**Figure 5–Figure supplement 3.** Novel topologically associating domains (TADs) with CTCF peaks at TAD boundaries (IMR90). Gene tracks, 24mer mappability tracks as well as CTCF peaks are displayed above the contact matrices. **(A)** Example at chr6:5,350,000-33,850,000. Even in the lack of obviously low mappable contact gaps, multi-reads can enhance the existing interaction signal and reveal detectable TAD structures supported by CTCF peaks. **(B)** Example on chr9:15,150,000-43,650,000. TAD structure, supported by CTCF peaks at the TAD boundaries, becomes detectable as multi-reads fill in the gap in the contact matrix. Red squares at the left bottom of the matrices indicate the color scale.

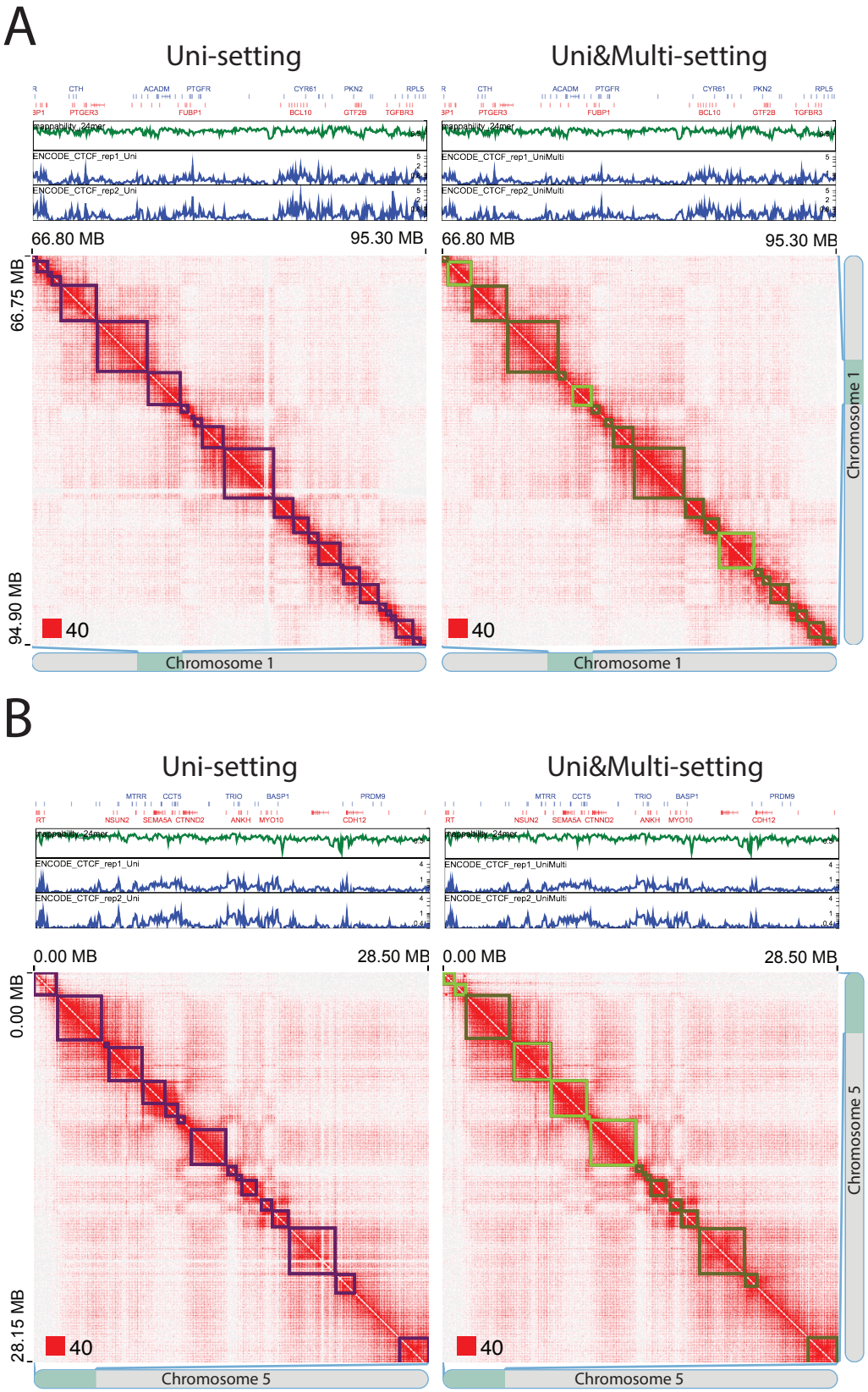



**Figure 5–Figure supplement 4.** Existing topologically associating domains (TADs) with adjusted boundaries supported by CTCF peaks at the new TAD boundaries (IMR90). **(A)** An example from chr1:66,800,000-95,300,000. **(B)** An example from chr5:0-28,500,000. Red squares at the left bottom of the matrices indicate the color scale.

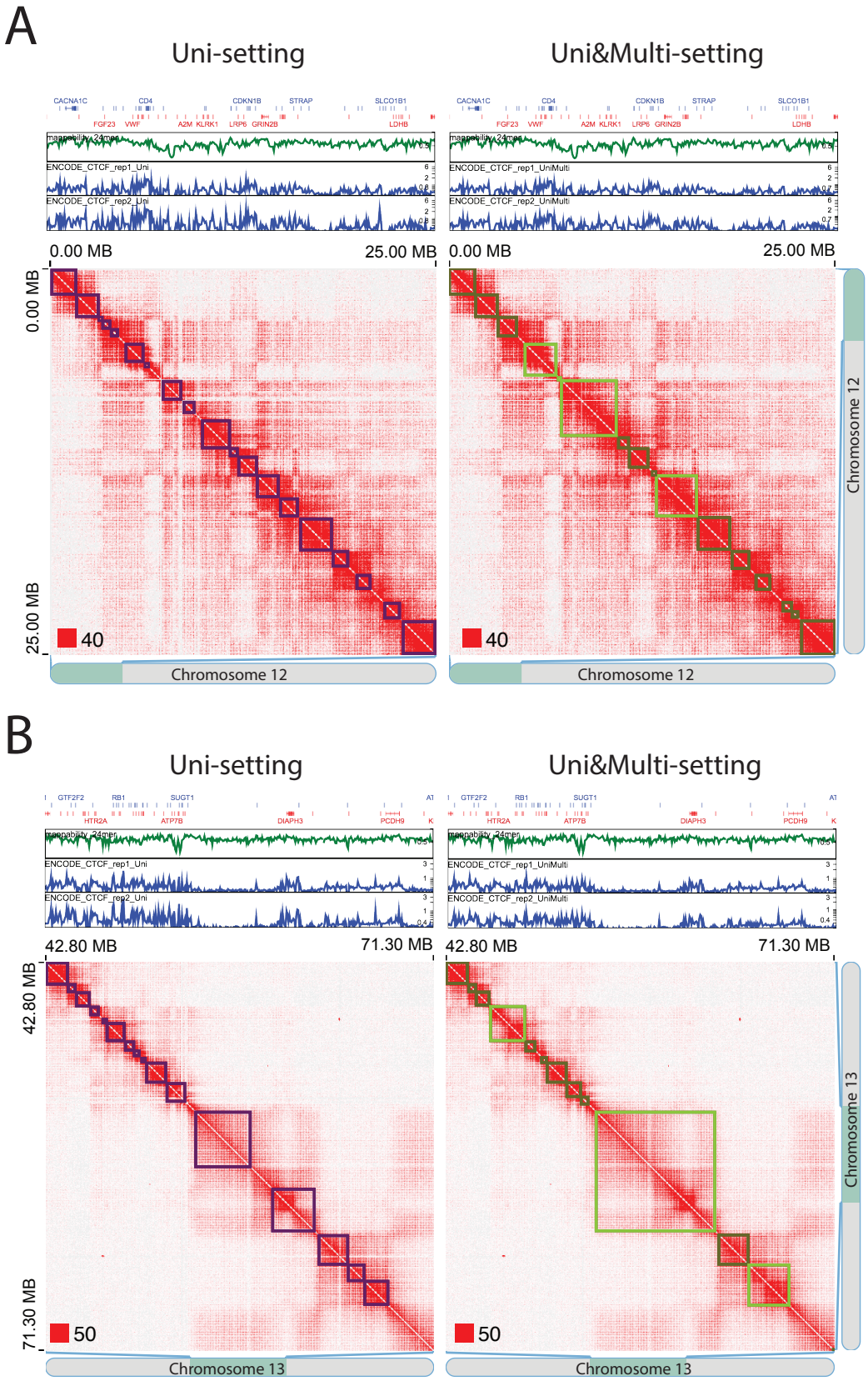



**Figure 5–Figure supplement 5.** Existing topologically associating domains (TADs) with adjusted boundaries supported by CTCF peaks at the new TAD boundaries (IMR90). **(A)** An example from chr12:0-25,000,000. **(B)** An example from chr13:42,800,000-71,300,000. Red squares at the left bottom of the matrices indicate the color scale.

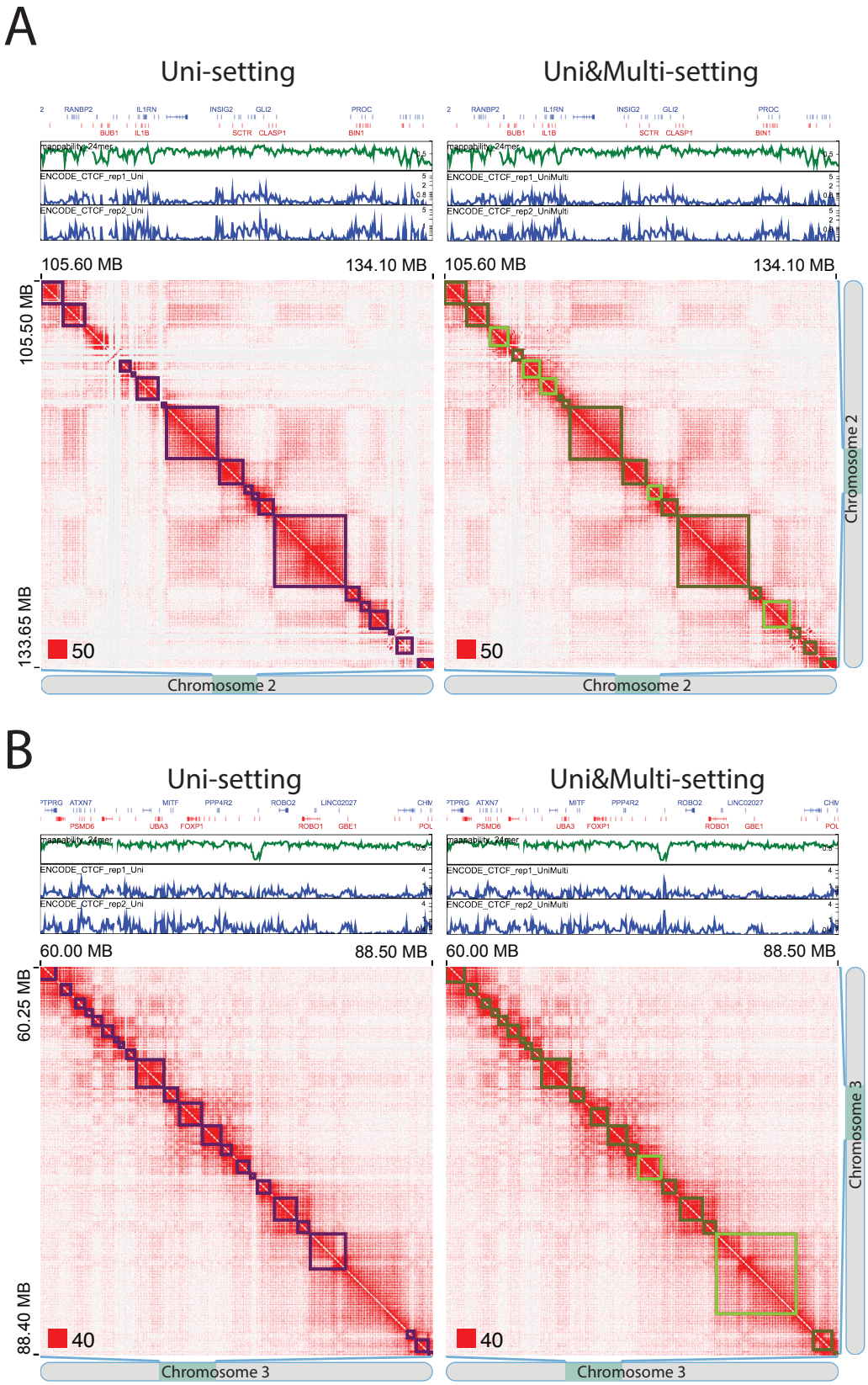



**Figure 5–Figure supplement 6.** False positive topologically associating domains (TADs) detected by the Uni-setting due to the missing reads in low mappability regions (IMR90). TADs that are splitted by white gaps are no longer detected once multi-reads are incorporated, indicating that they are highly likely false positives under the Uni-setting. **(A)** Example on chr2:105,600,000-134,100,000. **(B)** Example on chr3:60,000,000-88,500,000. Red squares at the left bottom of the matrices indicate the color scale.

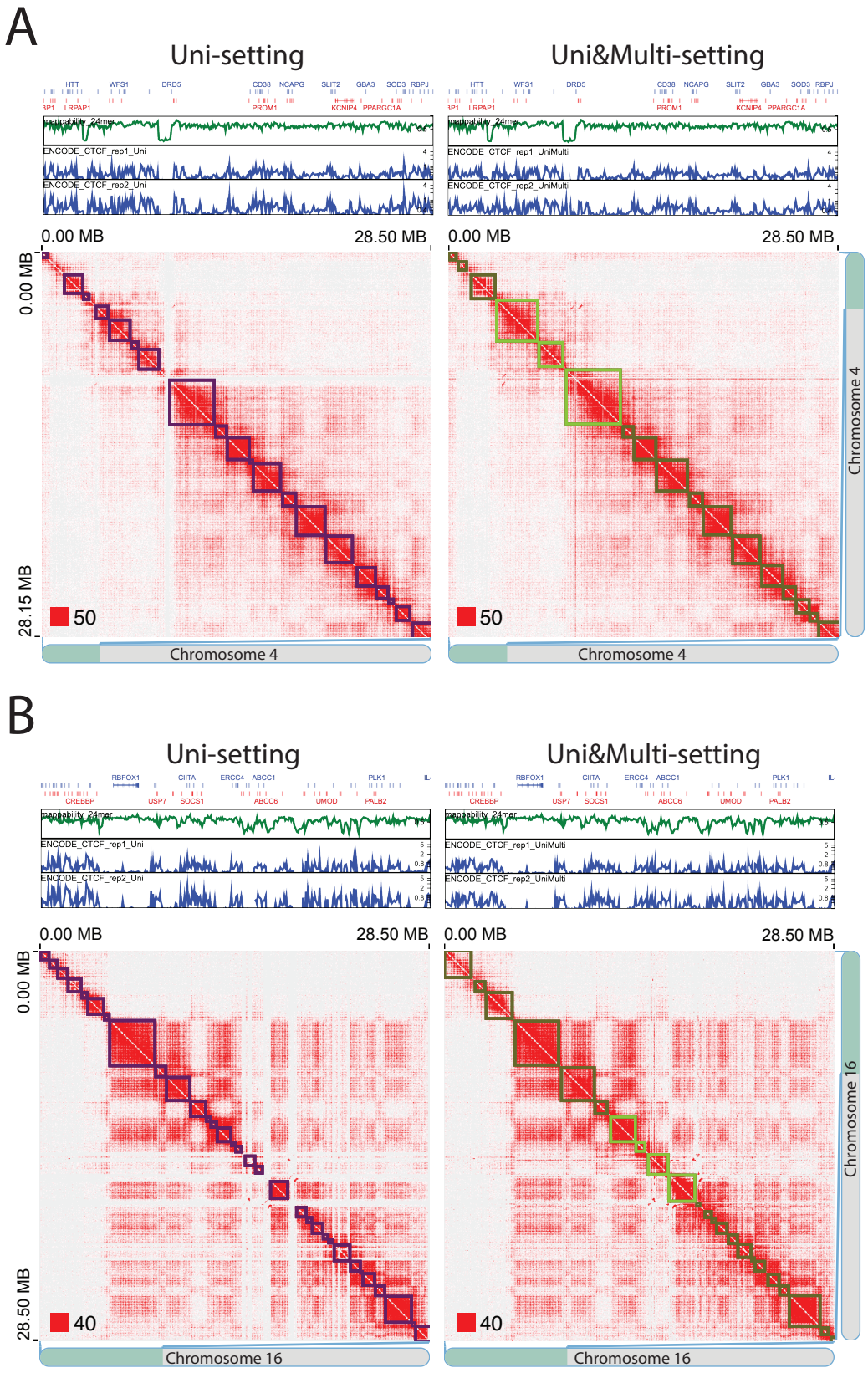



**Figure 5–Figure supplement 7.** False positive topologically associating domains (TADs) detected by the Uni-setting due to the missing reads in low mappability regions (IMR90). TADs that are splitted by white gaps are no longer detected once multi-reads are incorporated, indicating that they are highly likely false positives under the Uni-setting. **(A)** Example on chr4:0-28,500,000. **(B)** Example on chr16:0-28,500,000. Red squares at the left bottom of the matrices indicate the color scale.

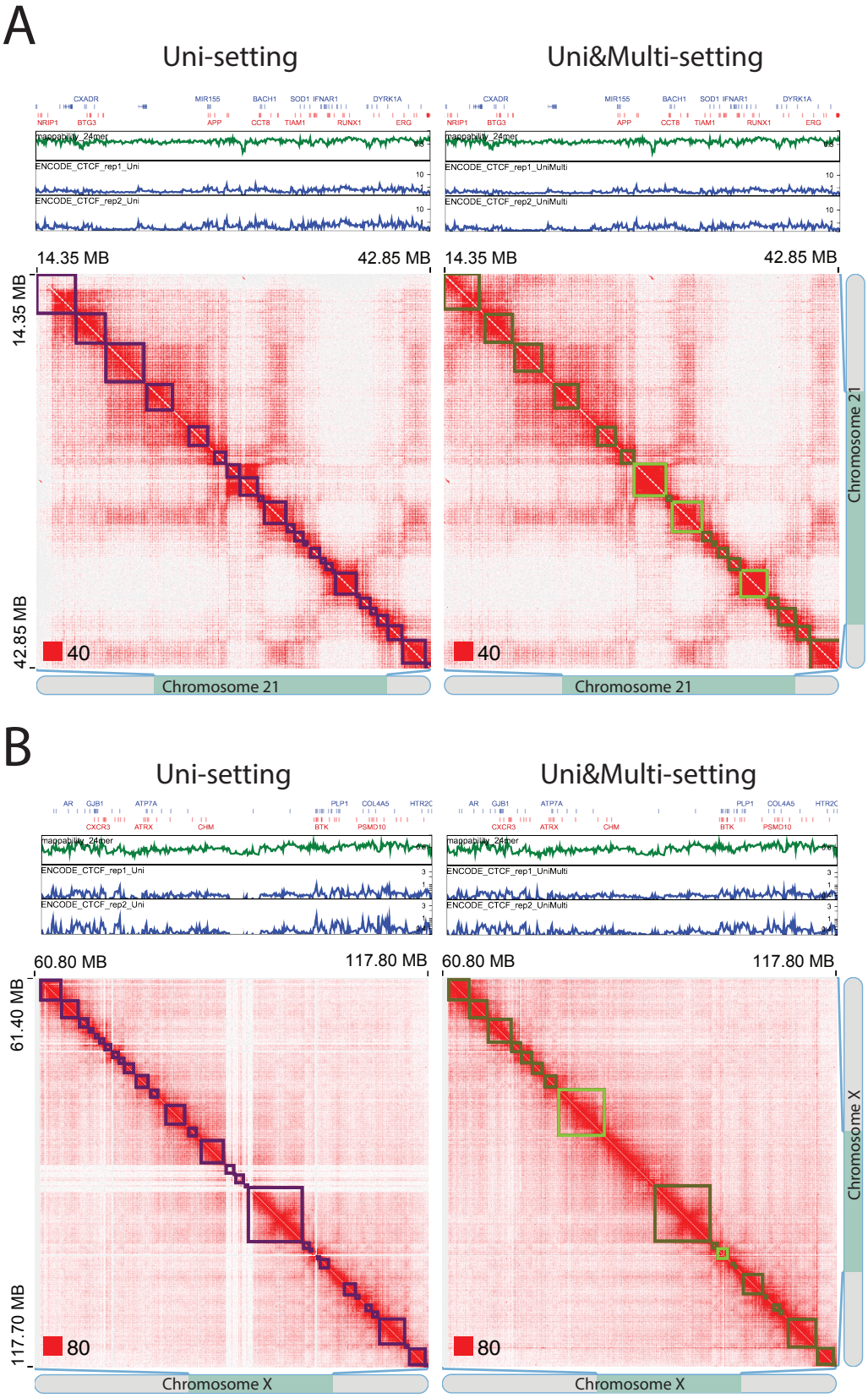



**Figure 5–Figure supplement 8.** False positive topologically associating domains (TADs) detected by the Uni-setting due to the missing reads in low mappability regions (IMR90). TADs that are splitted by white gaps are no longer detected once multi-reads are incorporated, indicating that they are highly likely false positives under the Uni-setting. **(A)** Example on chr21:14,350,000-42,850,000. **(B)** Example on chrX:60,800,000-117,800,000. Red squares at the left bottom of the matrices indicate the color scale.

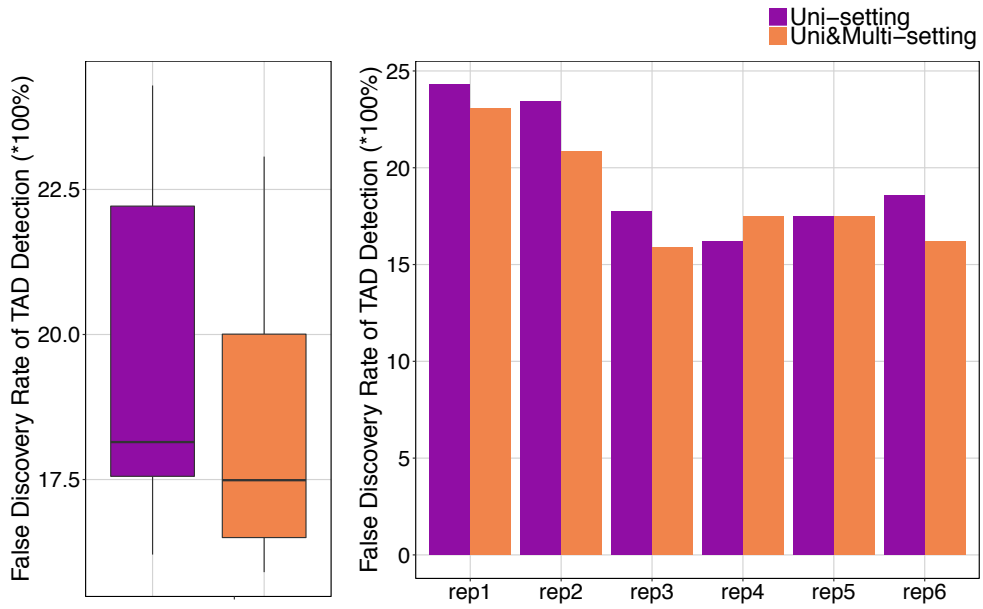



**Figure 5–Figure supplement 9.** False discovery rate of TADs detected under Uni-setting and Uni&Multi-setting (IMR90). TADs that are not reproducible are labeled as false positives without considering the CTCF peaks at the TAD boundaries.

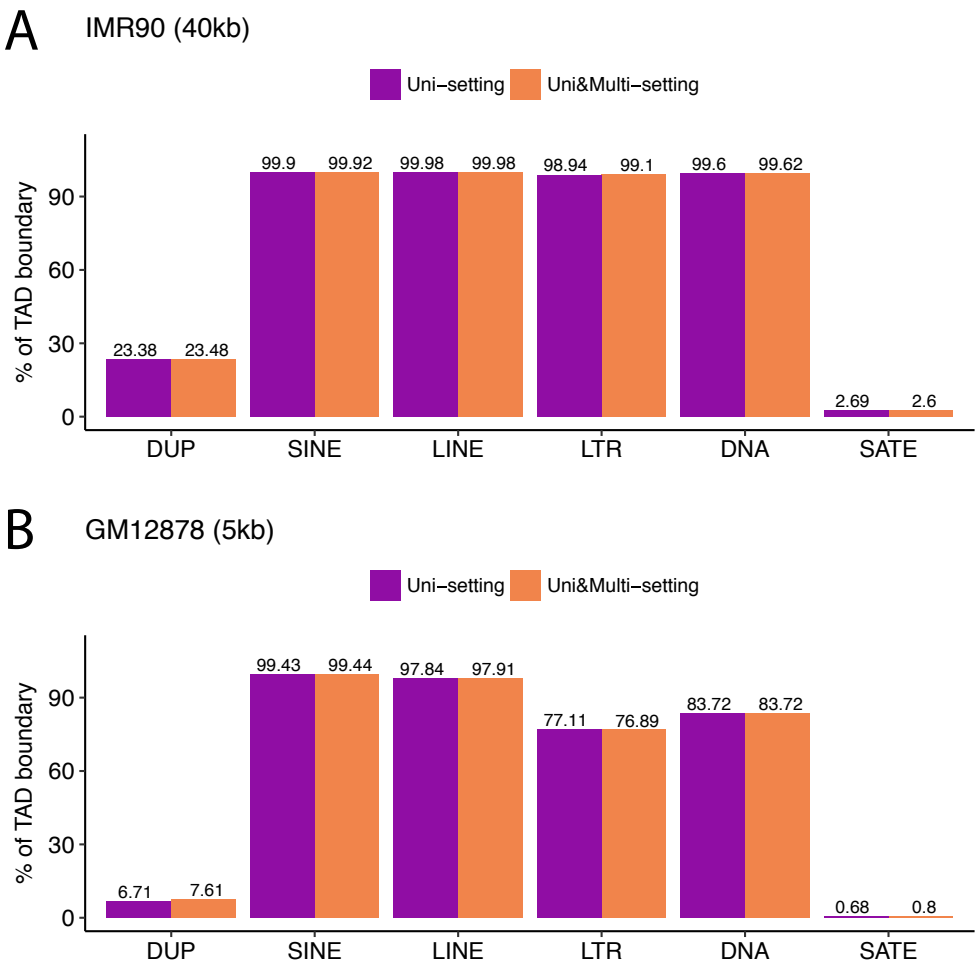



**Figure 5–Figure supplement 10.** Percentage of TAD boundaries co-localized with different types of repetitive elements under Uni-setting and Uni&Multi-setting for IMR90 at 40kb and GM12878 at 5kb.

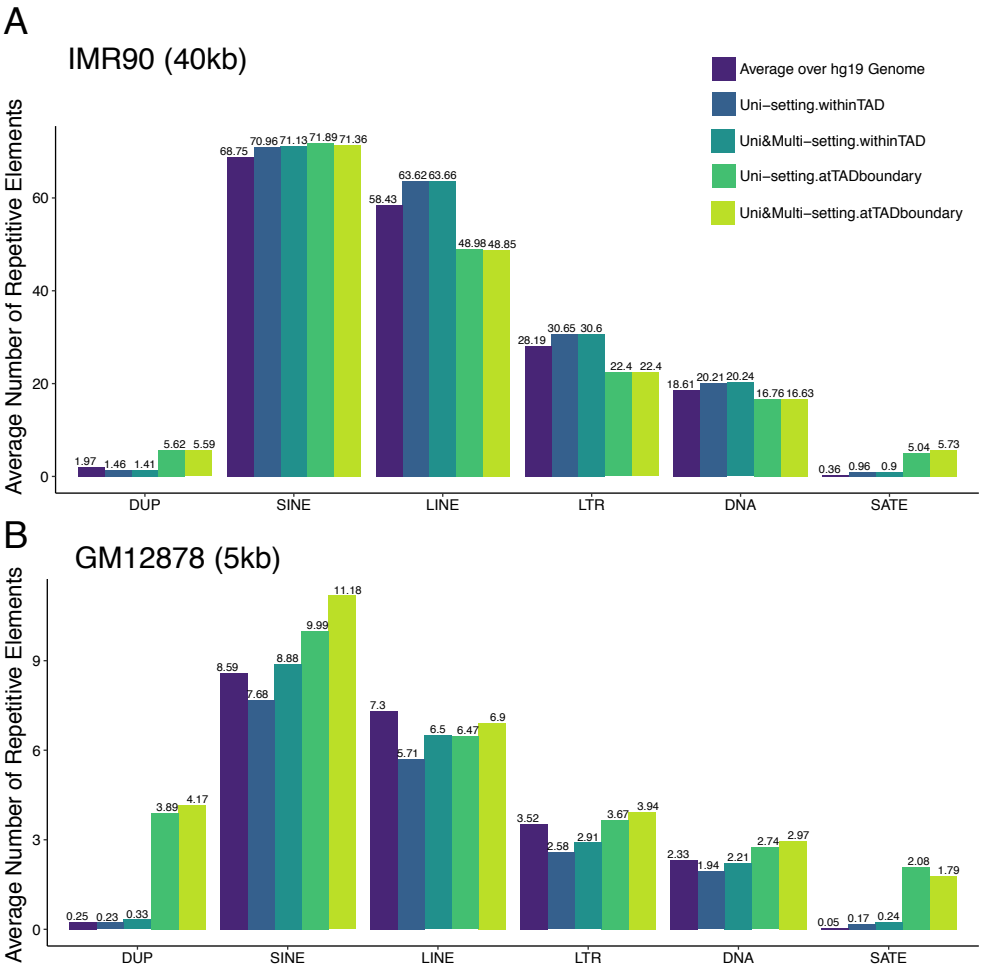



**Figure 5–Figure supplement 11.** The average number of repetitive elements at the reproducible topologically associating domains (TADs) detected under Uni-setting and Uni&Multi-setting for IMR90 at 40kb and GM12878 at 5kb. Such enrichment is compared to that within TAD and across the whole human genome.

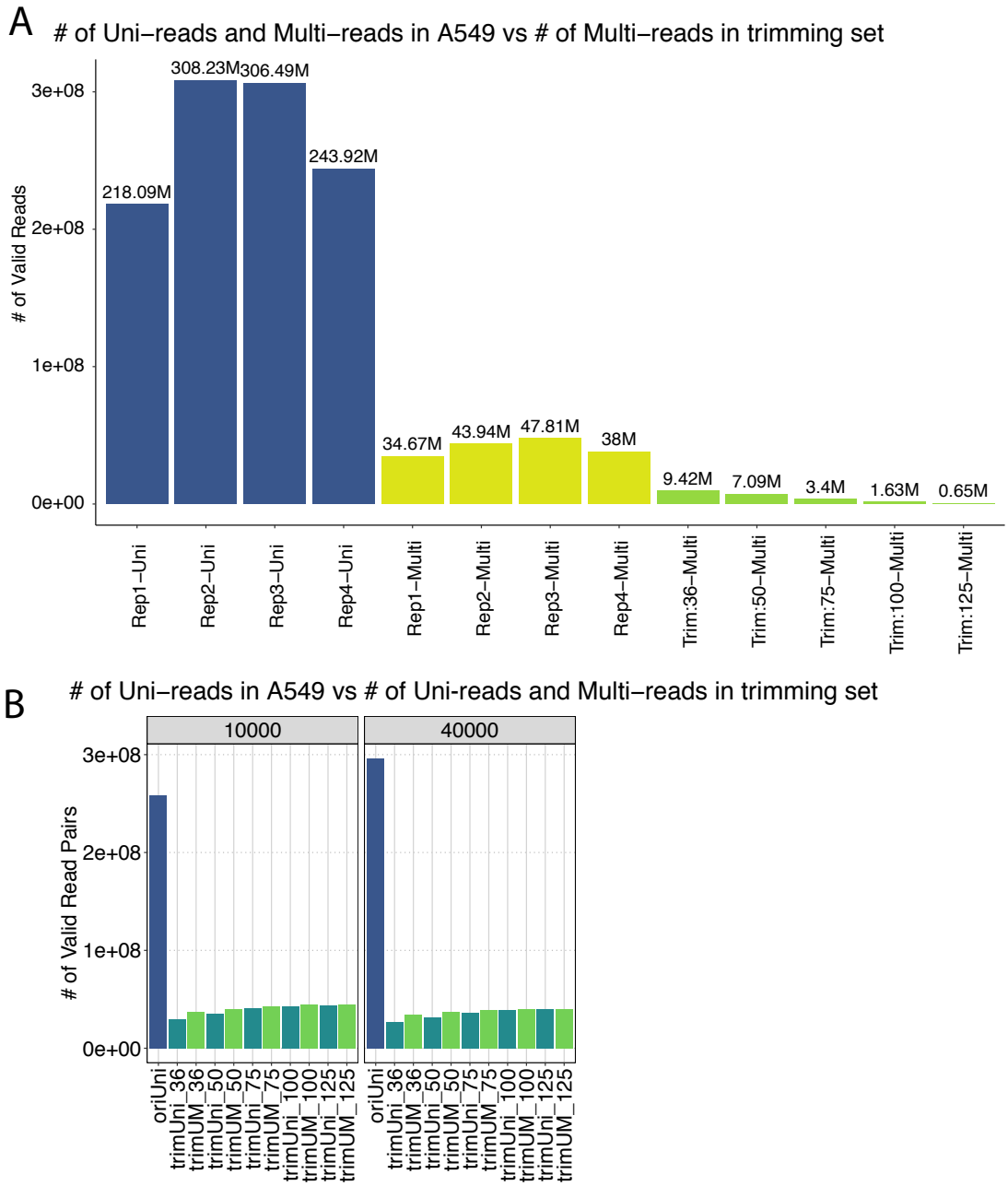



**Figure 6–Figure supplement 1.** The numbers of uni- and multi-reads after trimming compared with the original sequencing depth. **(A)** The numbers of trimmed multi-reads of different trimming read lengths compared to original depths of the four replicates of A549 dataset. **(B)** The numbers of trimmed uni-reads and multi-reads of different trimming read lengths compared to original A549 replicate 2 uni-reads depth.

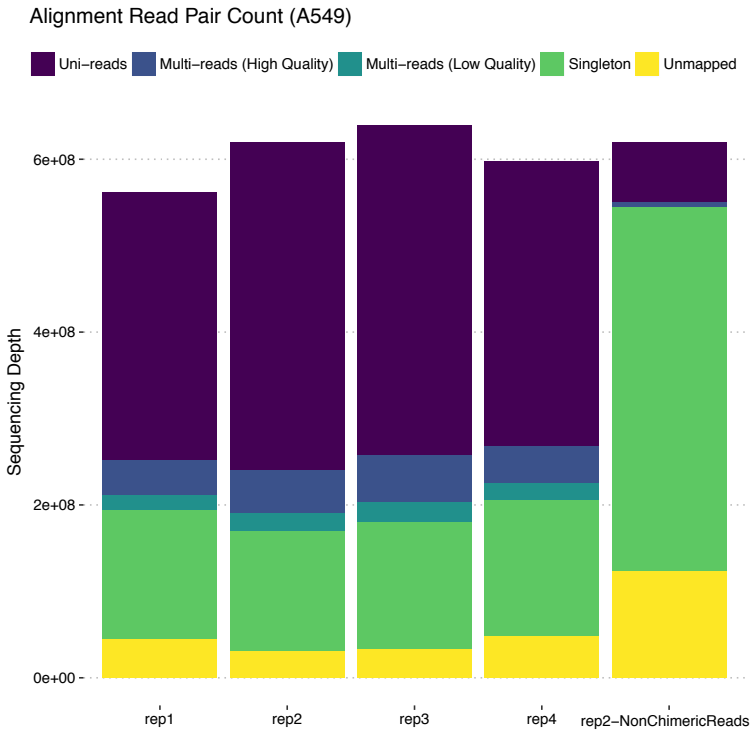



**Figure 6–Figure supplement 2.** Impact of chimeric reads on A549. By comparing rep2 with rep2-NonChimericReads bars, if chimeric reads are not rescued, a large number of uni-reads and multi-reads will be reduced to unmapped reads and singleton reads, leading to much fewer number of usable reads for downstream analysis.

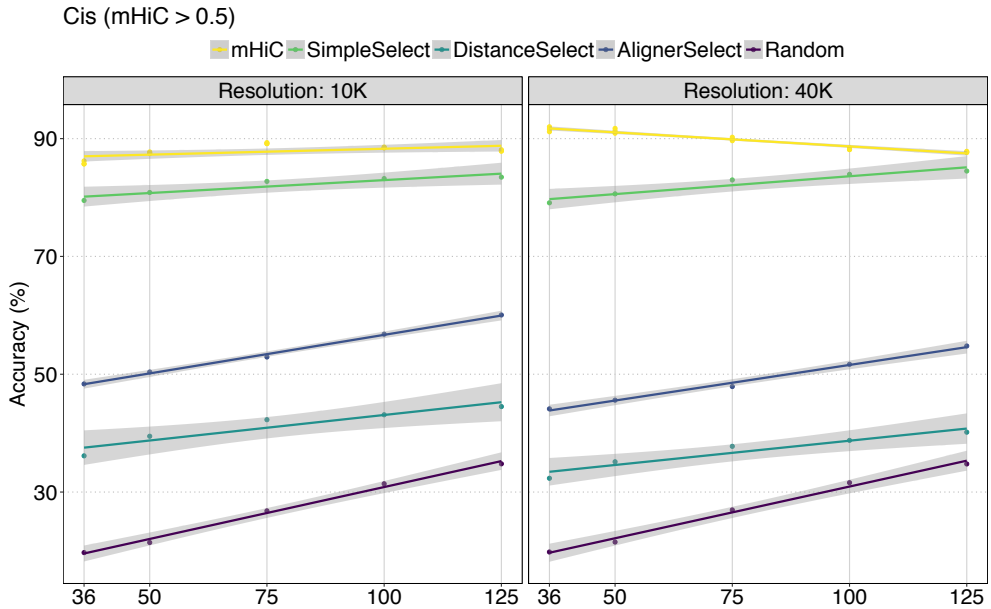



**Figure 6–Figure supplement 3.** Intra-chromosome allocation accuracy with respect to trimmed read length.

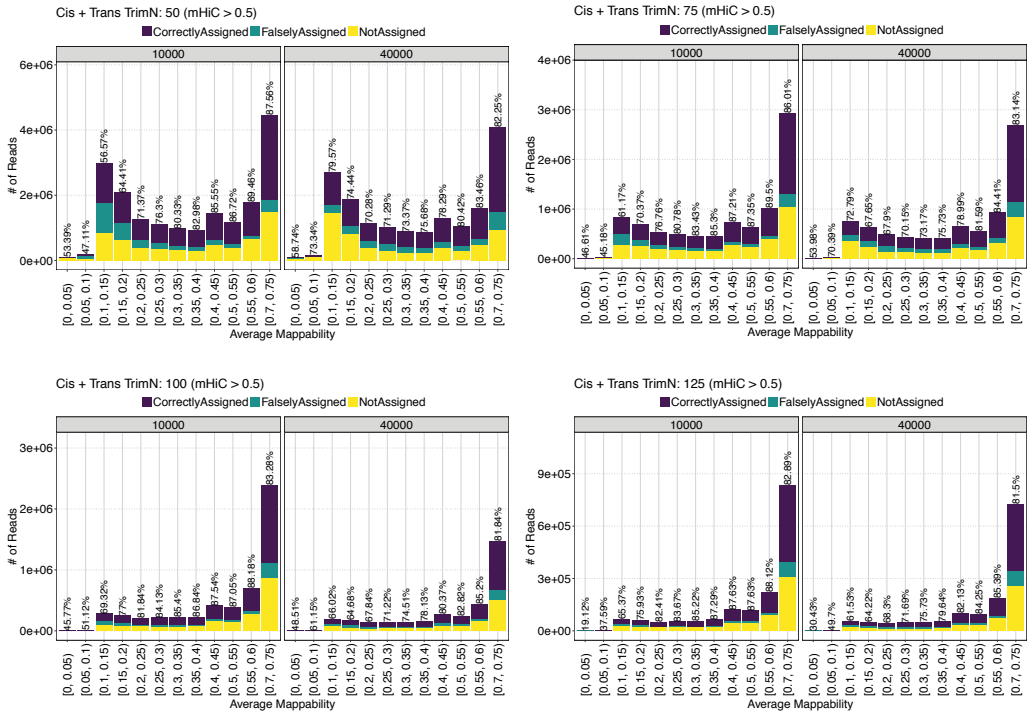



**Figure 6–Figure supplement 4.** Allocation accuracy across different mappability regions for trimmed reads of 50bp, 75bp, 100bp, and 125bp, respectively.

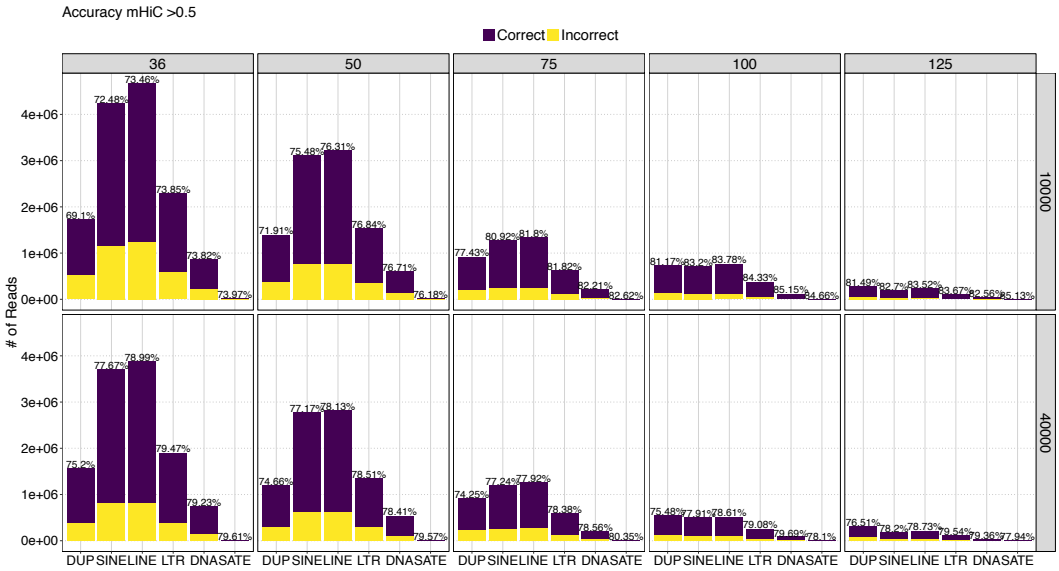



**Figure 6–Figure supplement 5.** Allocation accuracy across different classes of repetitive elements at 10kb and 40kb resolutions.

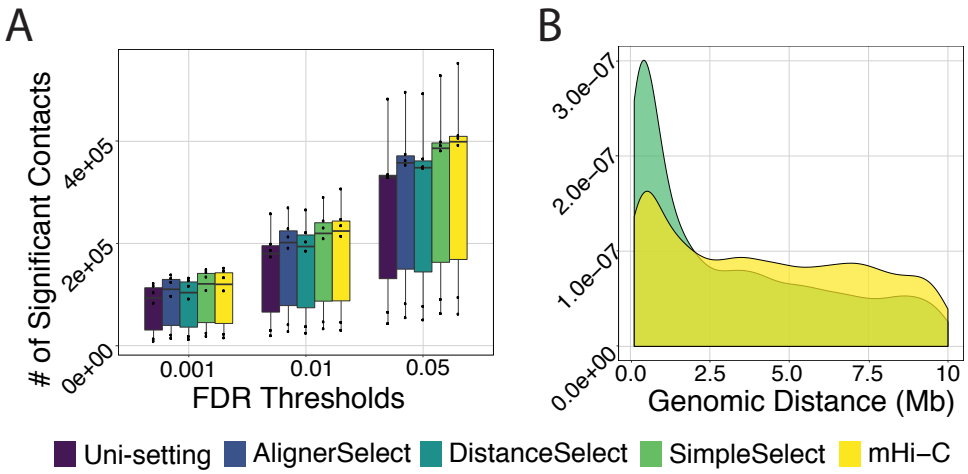



**Figure 6–Figure supplement 6.** **(A)** Numbers of significant interactions detected by different multi-read allocation strategies at different FDR thresholds. **(B)** Comparison of genomic distance distributions of significant interactions between SimpleSelect and mHi-C for IMR90 rep5 (FDR < 0.001)]

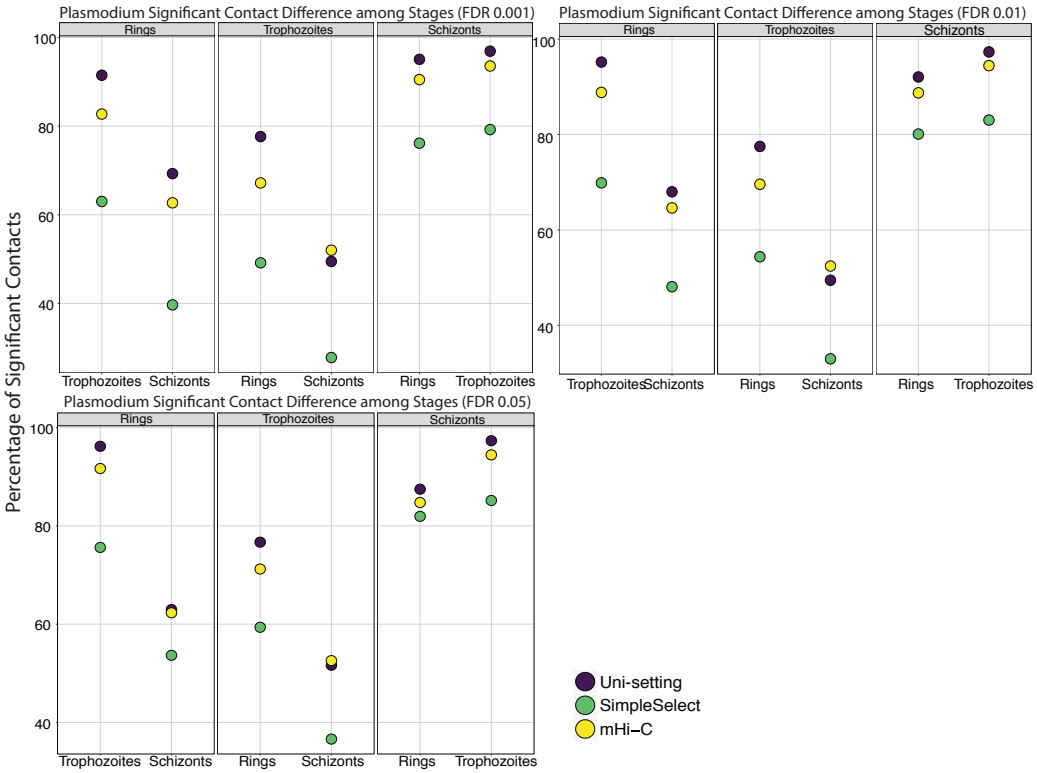



**Figure 6–Figure supplement 7.** Comparison of significant interactions among the three life stages of *P. falciparum*. Y-axis in each panel, namely rings, trophozoites and schizonts, depicts the percentage of contacts that are significant only in the panel condition compared to the other two conditions. Under varying FDR thresholds, mHi-C and Uni-setting tend to have similar percentages of differential interactions among ring - trophozoites - schizonts plasmodium life stages. In contrast, SimpleSelect tends to underestimate differential interactions due to over-emphasizing contact distance prior.

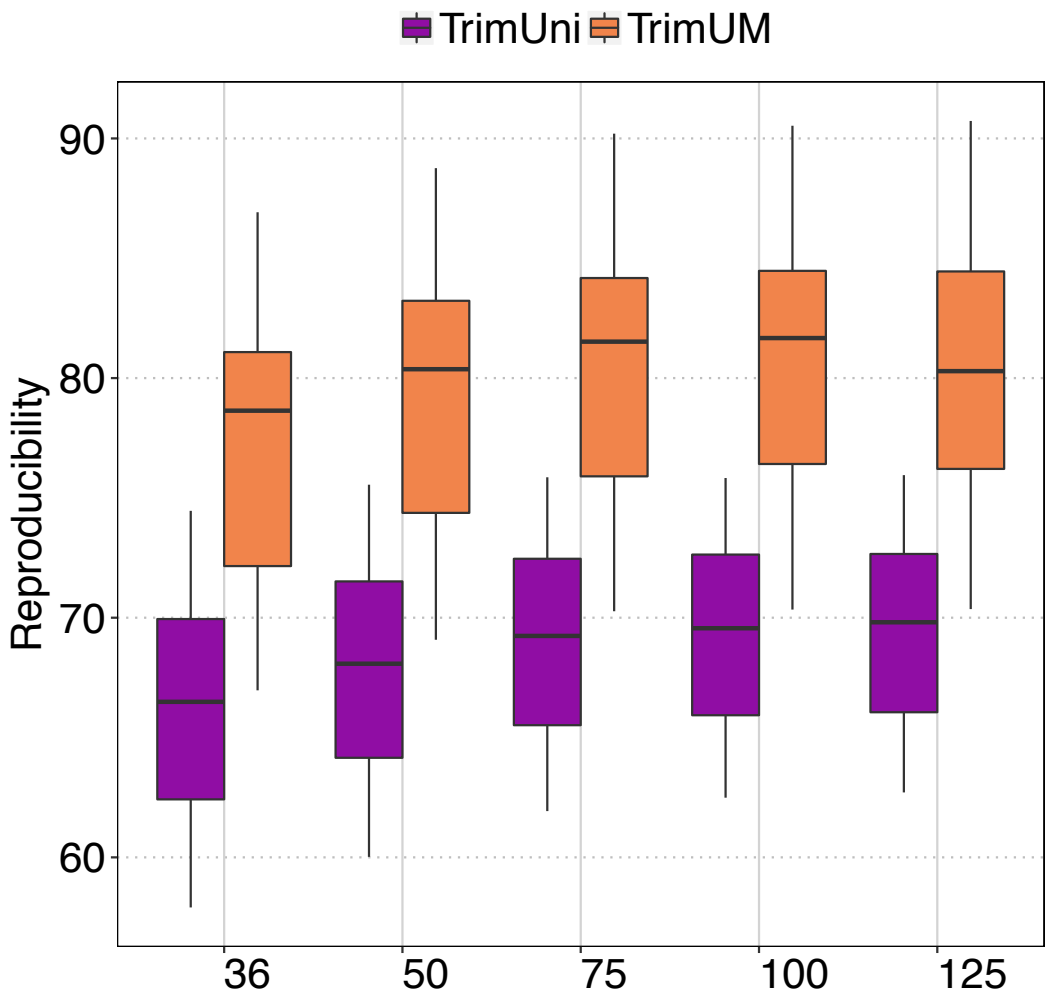



**Figure 7–Figure supplement 1.** Reproducibility comparison after adding mHi-C rescued trimmed multi-reads to trimmed uni-reads across different read lengths at 40kb.

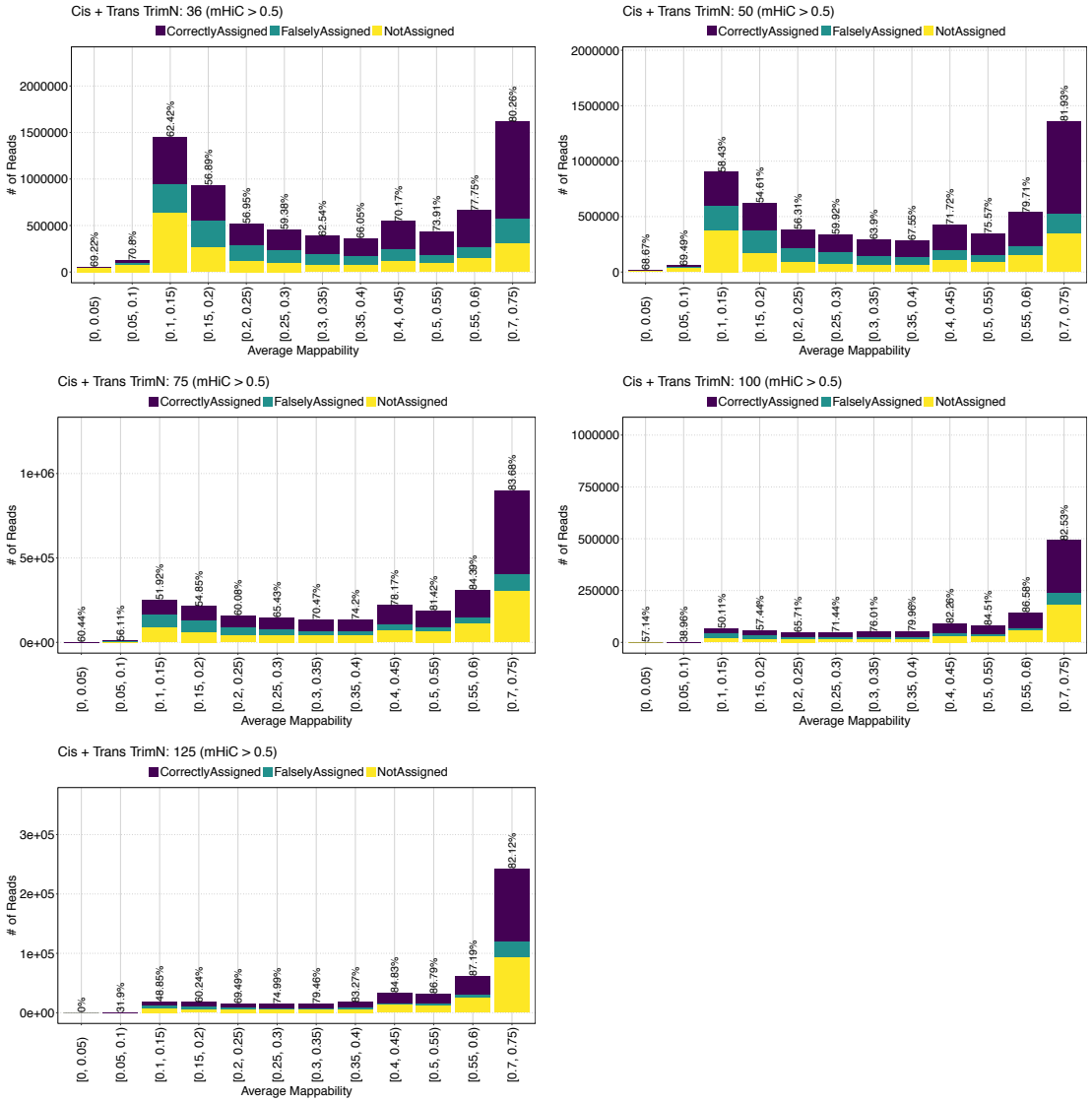



**Figure 7–Figure supplement 2.** Allocation accuracy of trimmed reads of different lengths among different mappability regions.

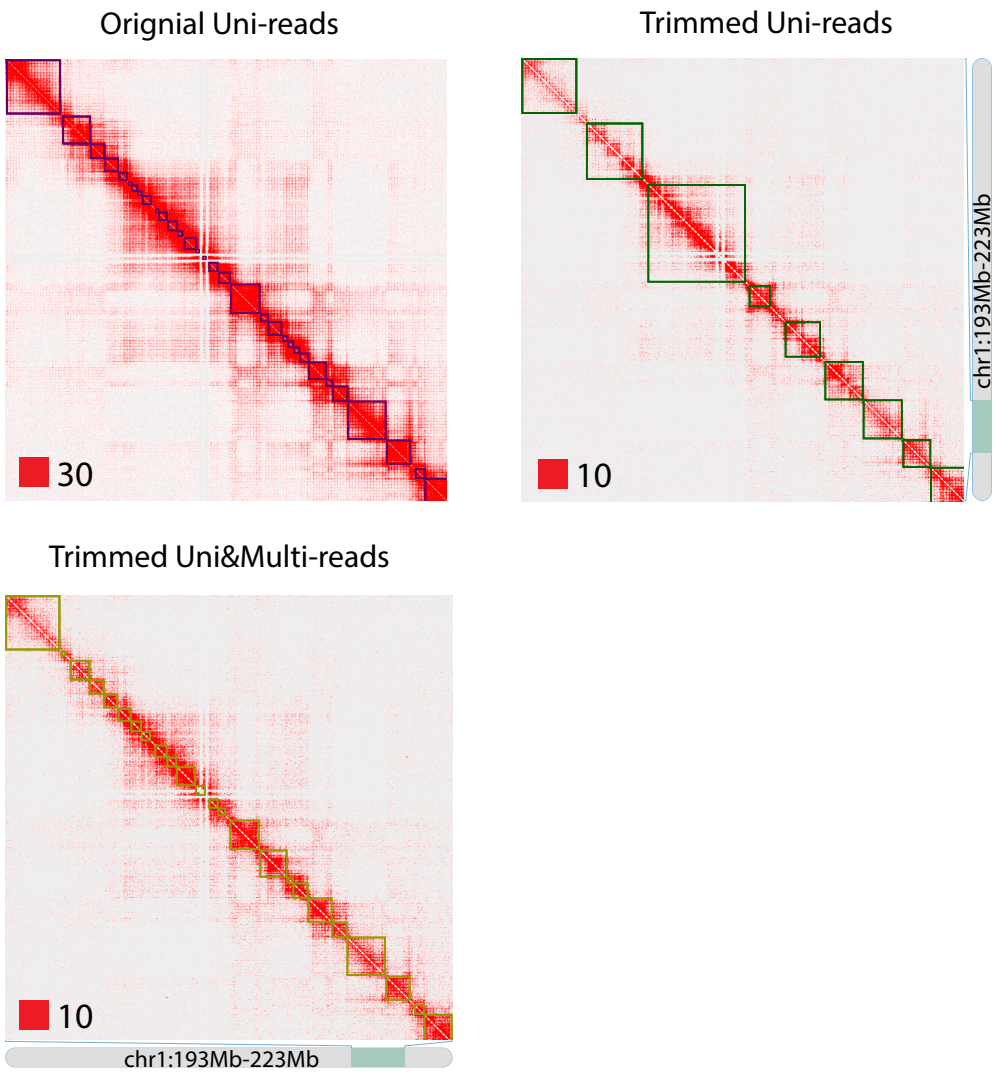



**Figure 7–Figure supplement 3.** TAD detection on chromosome 1 of original replicate uni-reads contact matrix, trimmed uni-reads (36bp long) contact matrix and trimmed uni- and multi-reads (36bp long) contact matrix.

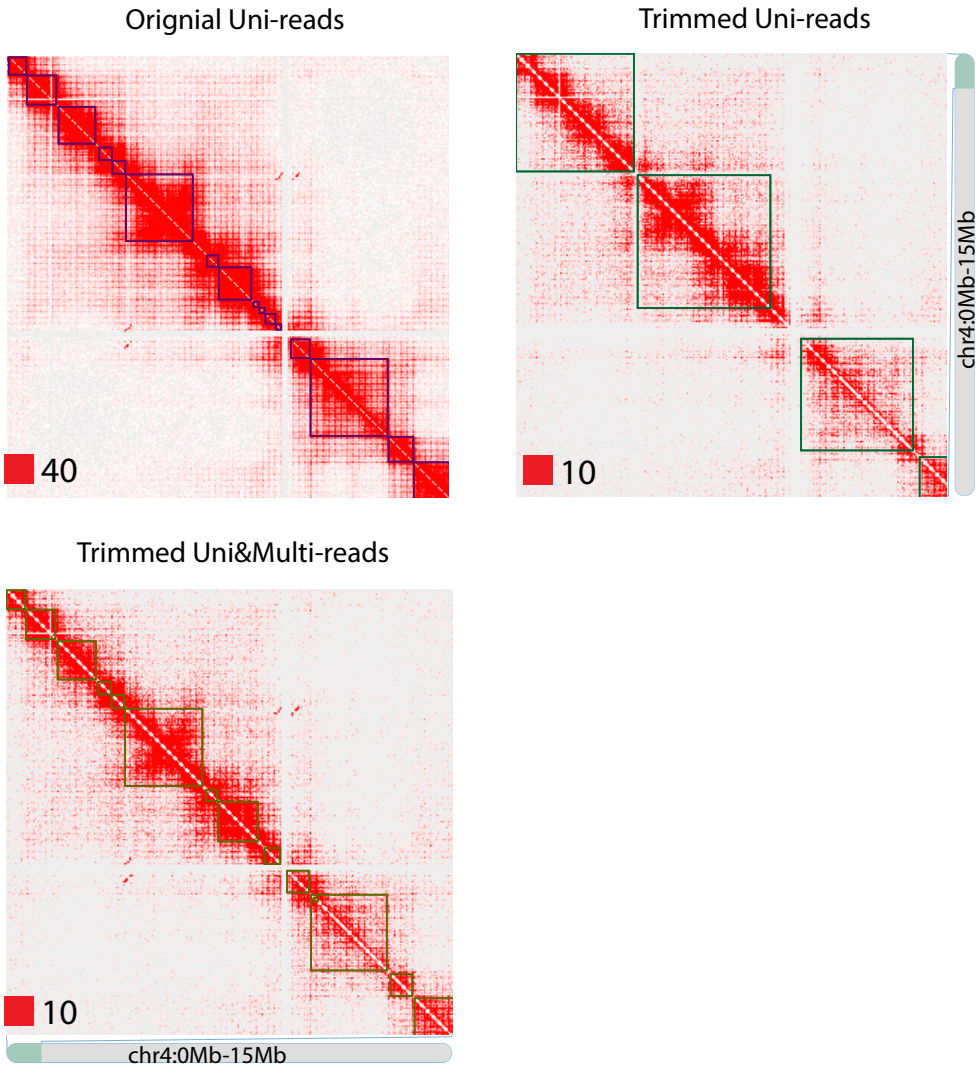



**Figure 7–Figure supplement 4.** TAD detection on chromosome 4 of original replicate uni-reads contact matrix, trimmed uni-reads (50bp long) contact matrix and trimmed uni- and multi-reads (50bp long) contact matrix.

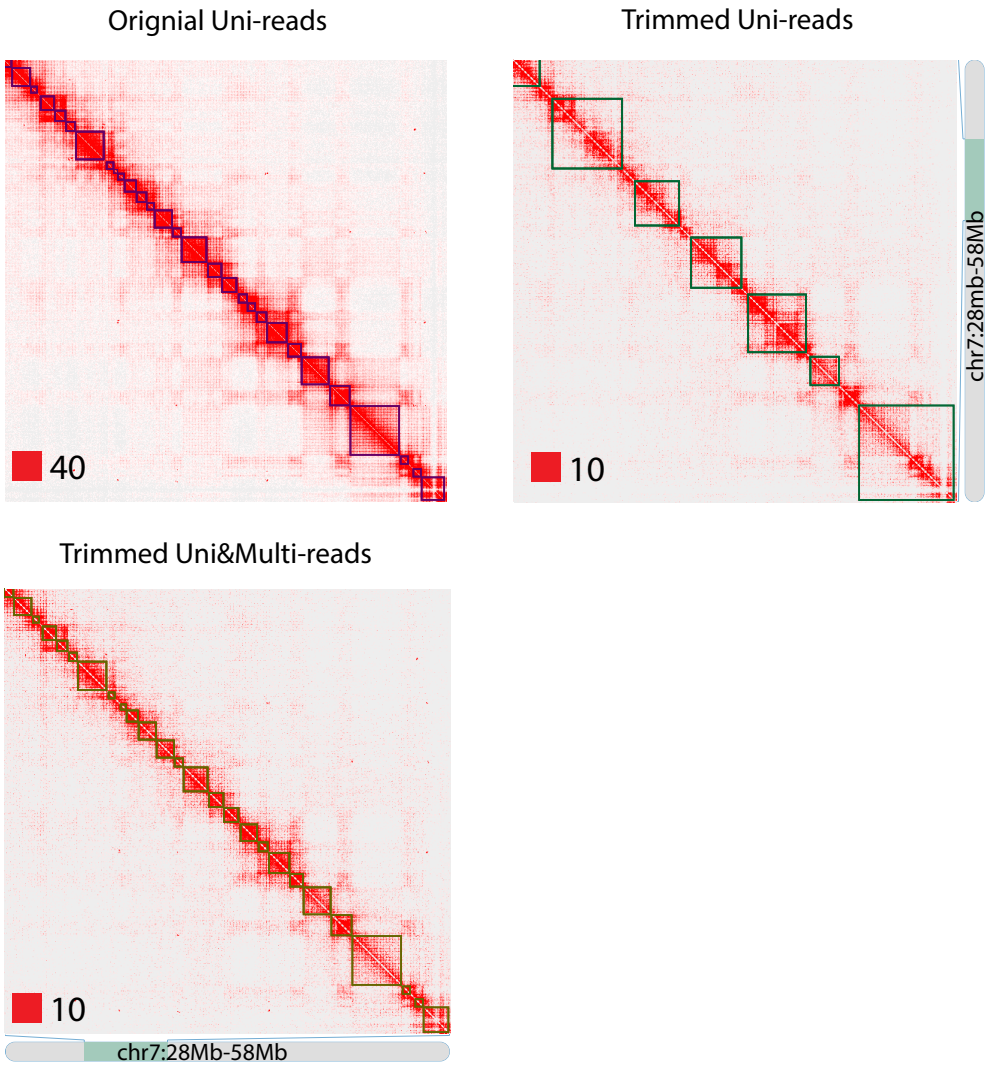



**Figure 7–Figure supplement 5.** TAD detection on chromosome 7 of original replicate uni-reads contact matrix, trimmed uni-reads (75bp long) contact matrix and trimmed uni- and multi-reads (75bp long) contact matrix.

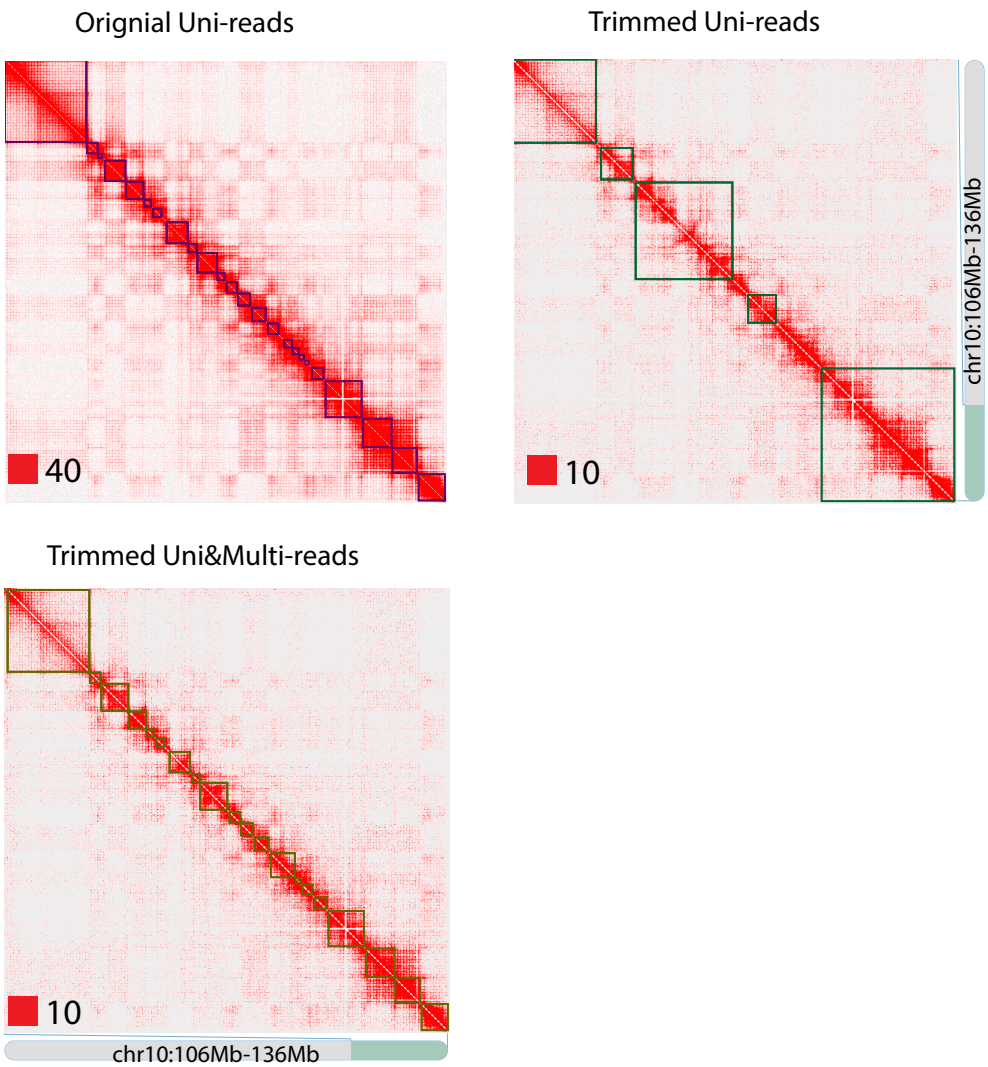



**Figure 7–Figure supplement 6.** TAD detection on chromosome 10 of original replicate uni-reads contact matrix, trimmed uni-reads (100bp long) contact matrix and trimmed uni- and multi-reads (100bp long) contact matrix.

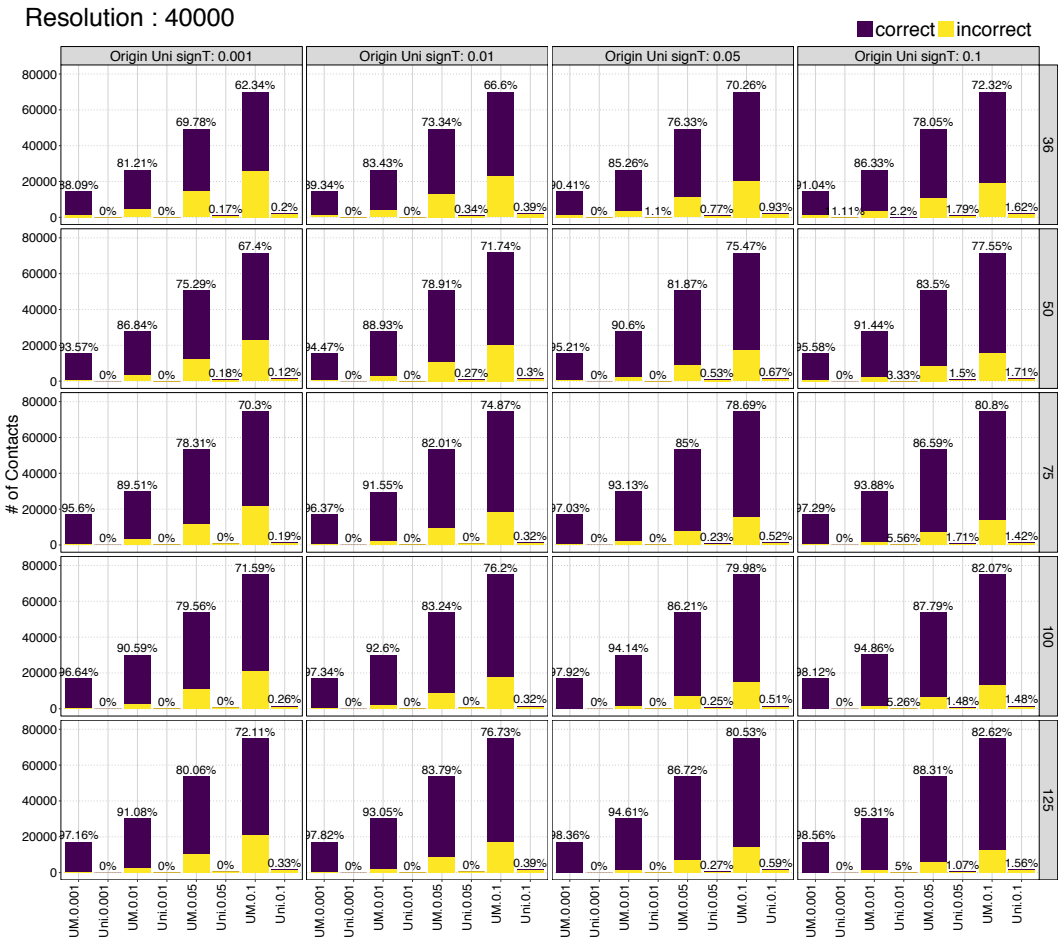



**Figure 7–Figure supplement 7.** Number of significant interactions identified by the mHi-C from trimmed reads that are also identified by the original long read length dataset at FDR 0.1%, 1%, 5%, and 10%, which are labeled at the top of each panel. "UM.0.001" on the x-axis indicates the category of significnat contacts called under trimmed Uni&Multi-setting at FDR 0.1%.

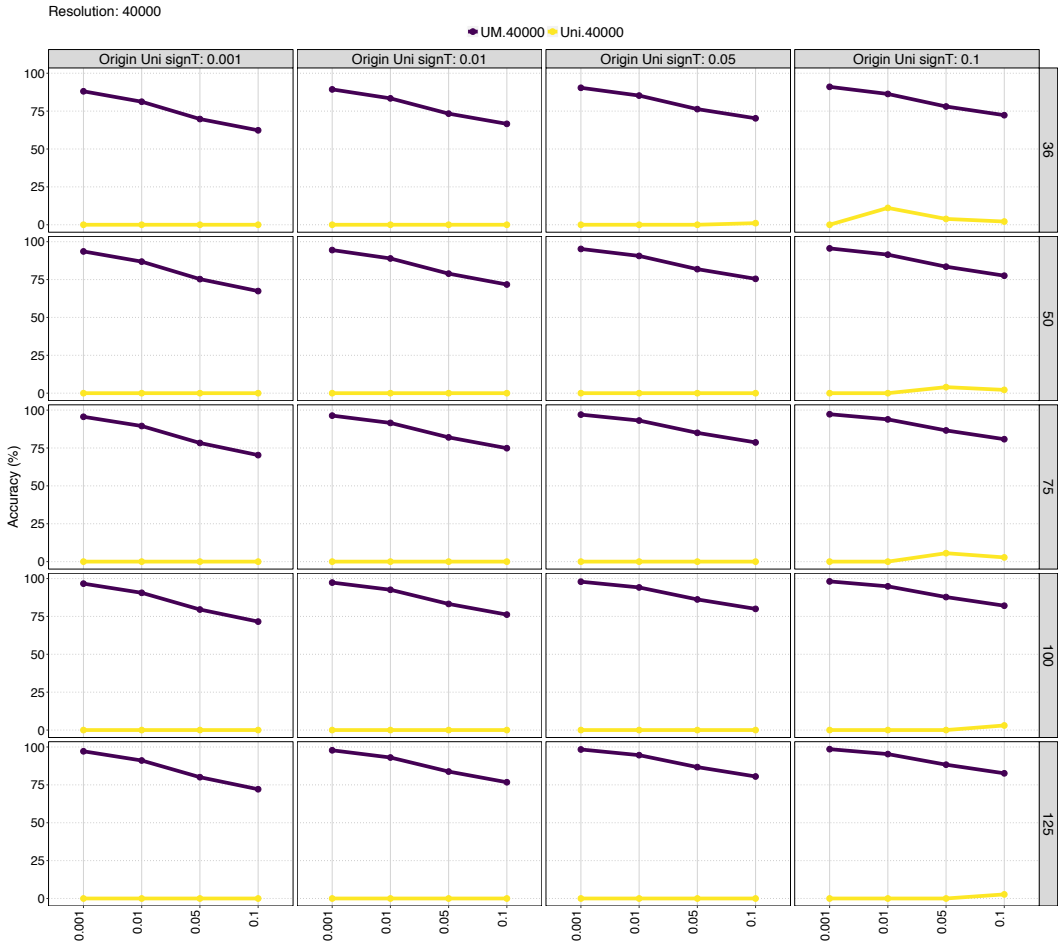



**Figure 7–Figure supplement 8.** Proportion of significant interactions of orignal full reads being recovered in trimmed reads sets under two settings at 40kb resolution.

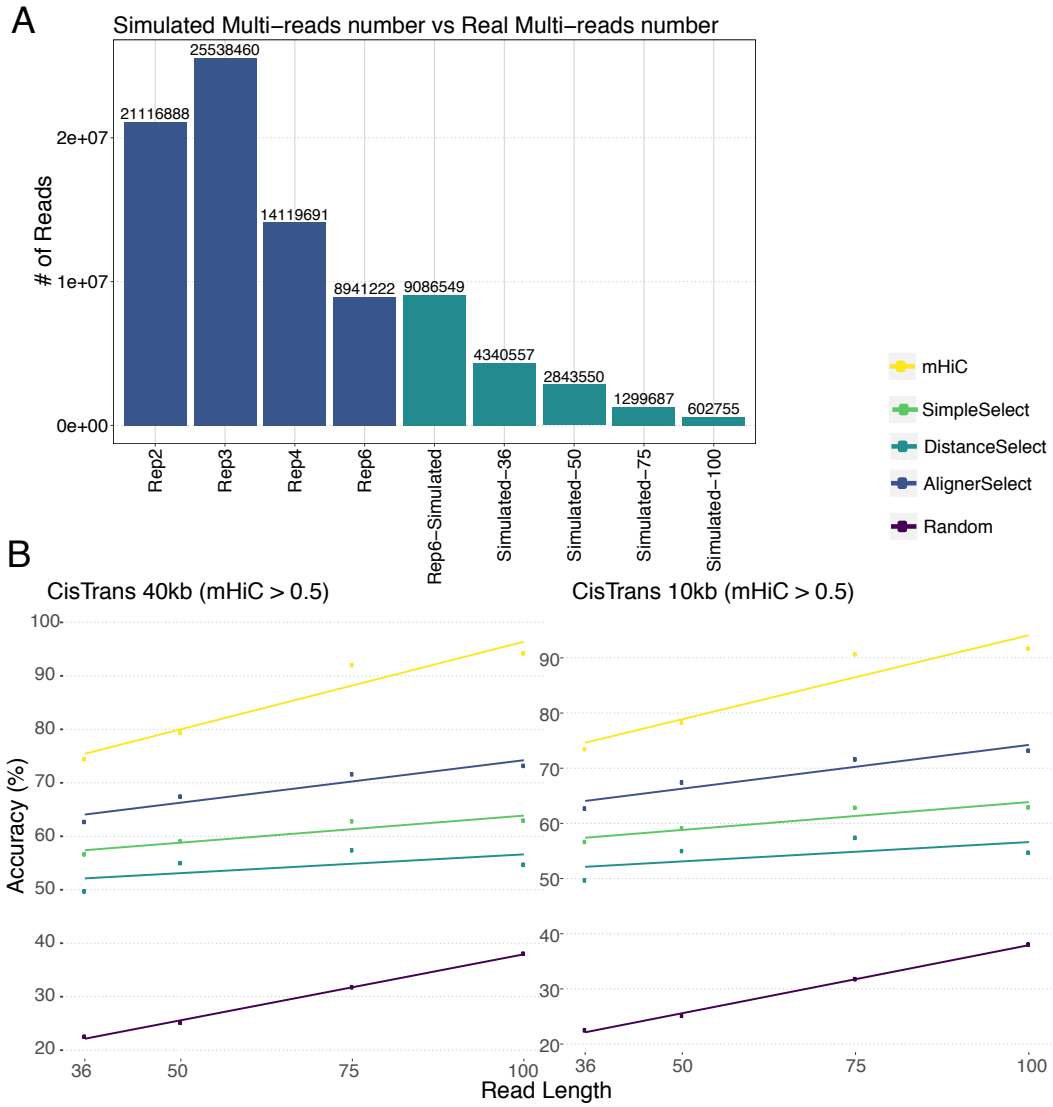



**Figure 7–Figure supplement 9.** Number of significant interactions identified by simulated Uni&Multi-setting and the allocation accuracy using merged mHi-C allocation process. **(A)** Number of simulated multi-reads compared to the multi-reads in real data. **(B)** Allocation accuracy using merged mHi-C assignment process at 40kb and 10kb resolutions.

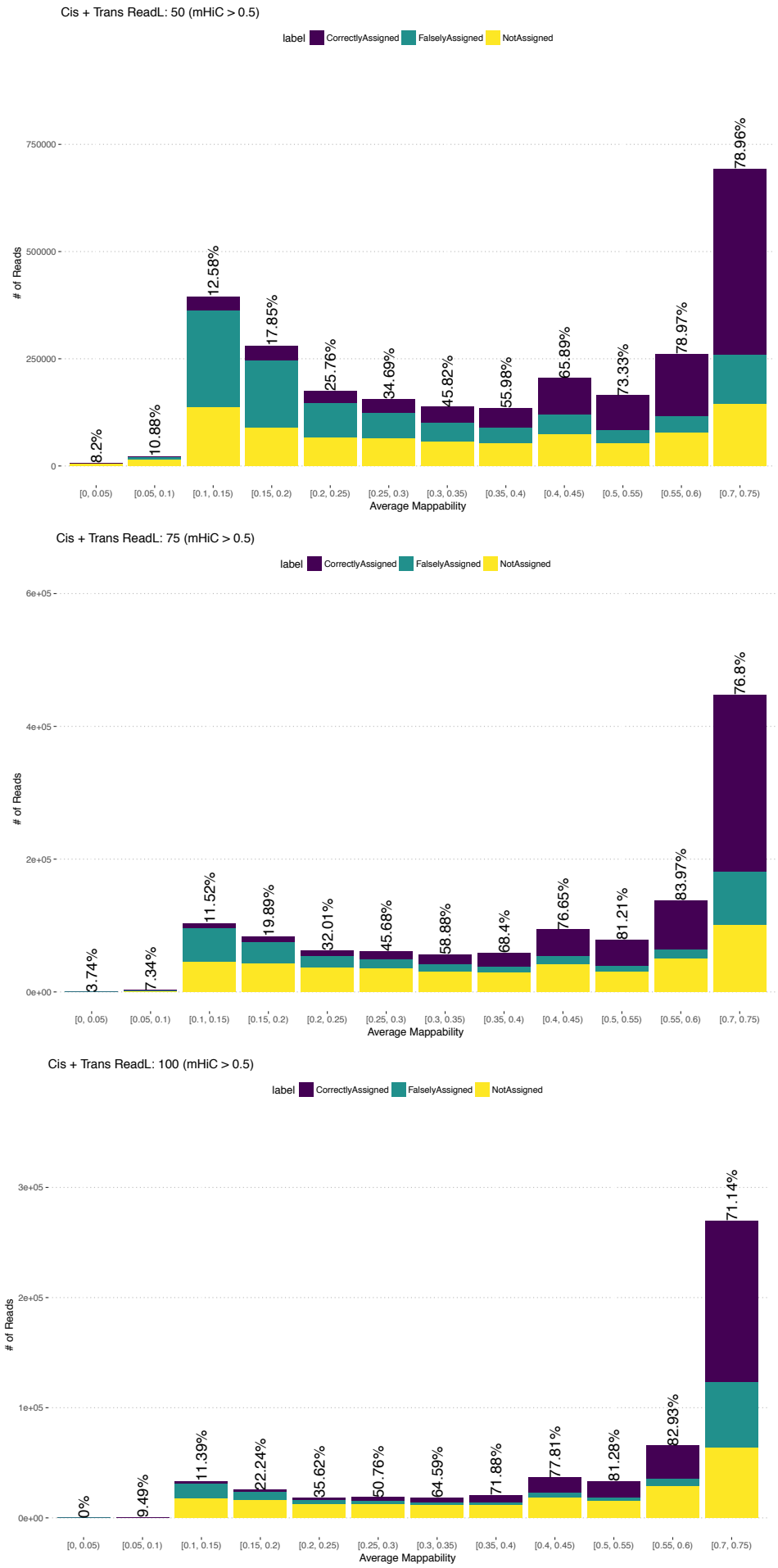



**Figure 7–Figure supplement 10.** mHi-C allocation accuracy with respect to mappabilities of simu-