MaxHiC: robust estimation of chromatin interaction frequency in Hi-C and capture Hi-C experiments

Hamid Alinejad-Rokny[#*] [1, 2], Rassa Ghavami[#] [3], Hamid R. Rabiee[3], Narges Rezaei[4], Kin Tung Tam[1], Alistair R. R. Forrest* [1]

[1]Harry Perkins Institute of Medical Research, QEII Medical Centre and Centre for Medical Research, the University of Western Australia, Nedlands, Perth, WA, 6009, AU.

[2]Systems Biology and Health Data Analytics Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, 2052, AU.

[3]Bioinformatics and Computational Biology Lab, Department of Computer Engineering, Sharif University of Technology, Tehran, 11365, IR.

[4]Center for Complex Biological Systems, University of California Irvine, California, 92697, US.

[#]Joint first author. These authors contributed equally in this paper.

*Correspondence should be addressed to**:** HAR (h.alinejad@unsw.edu.au) and ARRF (alistair.forrest@perkins.uwa.edu.au).

Abstract

Hi-C is a genome-wide chromosome conformation capture technology that detects interactions between pairs of genomic regions, and exploits higher order chromatin structures. Conceptually Hi-C data counts interaction frequencies between every position in the genome and every other position. Biologically functional interactions are expected to occur more frequently than random (background) interactions. To identify biologically relevant interactions, several background models that take biases such as distance, GC content and mappability into account have been proposed. Here we introduce MaxHiC, a background correction tool that deals with these complex biases and robustly identifies statistically significant interactions in both Hi-C and capture Hi-C experiments. MaxHiC uses a negative binomial distribution model and a maximum likelihood technique to correct biases in both Hi-C and capture Hi-C libraries. We systematically benchmark MaxHiC against major Hi-C background correction tools and demonstrate using published Hi-C and capture Hi-C datasets that 1) Interacting regions identified by MaxHiC have significantly greater levels of overlap with known regulatory features (e.g. active chromatin histone marks, CTCF binding sites, DNase

sensitivity) and also disease-associated genome-wide association SNPs than those identified by currently existing models, and 2) the pairs of interacting regions are more likely to be linked by eQTL pairs and more likely to identify known enhancer-promoter pairs than any of the existing methods. We also demonstrate that interactions between different genomic region types have distinct distance distribution only revealed by MaxHiC. MaxHiC is publicly available as a python package for the analysis of Hi-C and capture Hi-C data.

Keywords: Genome architecture, Chromosome conformation capture (Hi-C), Capture Hi-C, Chromosomal contact map, Chromatin interaction, Enhancer-promoter interactions, Gene regulation.

Introduction

Hi-C is a genome-wide chromosome conformation capture technology used to identify interactions between pairs of genomic regions. Hi-C libraries consist of a pair of reads which can be mapped to the genome for identifying pairs of interacting regions. The data is typically summarized as a contact matrix, in which each row and column correspond to a (fixed-bin) fragment of the genome and each cell shows the number of reads supporting an interaction between the two fragments. Hi-C data has multiple biases due both to the multiple steps used in the protocol and the proclivity of closely located genomic regions that are more likely to randomly interact. Not all interactions are biologically relevant and thus the important challenge in the analysis of Hi-C data is to distinguish between random/artefactual interactions and those interactions that are more likely to be functional/regulatory.

There are two main approaches used to identify significant interactions in Hi-C data. The first, as used in Hi-C-Norm (1) and Hi-C-DC (2), attempts to quantify known sources of biases (fragment length, GC content and mappability), and use them as variables in a background expectation function. The second approach, as used in ICE (3), Fit-Hi-C (4), CHiCAGO (5) and GOTHiC (6), assumes that all Hi-C biases are reflected in the read count data of interactions. Therefore, the observed read counts are used to assign a P-value to each individual interaction. In order to account for unknown sources of bias a separate bias parameter is assumed for each bin. The tools

using the first approach are limited to analysis of Hi-C libraries generated with restriction enzymes, whereas those using the second approach can also work with DNase Hi-C experiments in which the fragment sizes are unknown. Rao *et el.* (7) also proposed HiCCUP, which is not a background correction model, but can be used to identify chromatin loops. HiCCUP identifies clusters of Hi-C contact matrixes and reports the centroid of the cluster, in which the frequency of the contact matrix is significantly higher than the local background. Using a local background and reporting only the centroid of a cluster results in identification of a small number of significant Hi-C interactions. Notably, as HiCCUP is implemented using GP-GPUs, the user needs to have specific hardware to run it.

We introduce MaxHiC (Maximum Likelihood estimation for Hi-C), a negative binomial model that uses a maximum likelihood technique to correct the complex combination of known and unknown biases in both Hi-C and capture Hi-C libraries. Benchmarking against current leading background models (GOTHiC, CHiCAGO, Fit-Hi-C, Fit-Hi-C2 (8) and HiCCUP) shows that MaxHiC identifies interactions that contain substantially more evidence of biological relevance. Interacting regions identified by MaxHiC are significantly enriched for H3K27ac, H3K4me1, CTCF binding sites, DNase hypersensitive sites, and GWAS SNPs. They are also more likely to be annotated as promoters, enhancers, and insulator elements. More importantly the interacting pairs are substantially enriched for known enhancer-promoter pairs and non-coding polymorphisms linked by eQTL to target genes.

MaxHiC as an open-source python package is available at https://github.com/forrest-lab/MaxHiC and https://github.com/Rassa-Gvm/MaxHiC.

Results

A negative binomial regression background model for Hi-C data

Here we present MaxHiC, a novel background correction model for identifying statistically significant interacting regions in both general and capture Hi-C libraries. In MaxHiC, the genome is divided into fixed size non-overlapping bins, and the read counts supporting interactions between bins are tested for significance. The read-count

of interactions are assumed to follow a negative binomial distribution which is widely used to model over dispersed count data (9, 10). The mean parameter of the negative binomial distribution for each interaction is calculated as a function of the bias factors of its two ends and their genomic distance.

It is known that the contact frequency between pairs of regions linked by Hi-C reads decays as a function of distance along the chromosome (11). A large fraction of Hi-C reads represent random ligations products of nearby genomic regions. In MaxHiC, distance is modelled by a function that decreases at increasing genomic distances to reach a small but constant non-zero value to account for random ligations. For *trans*-interactions, we use the same constant value as observed for distant *cis* interactions. Bias factors of bins are calculated as a function of their total read-count and mappability of their surrounding bins. All of the parameters of the model are learned by maximizing the logarithm of likelihood of the observed interactions using the ADAM algorithm (12). The model is trained in multiple iterations; in each iteration, putative interactions with the best P-values are identified and set aside. These significant interactions are ignored in the next phase of training so the trained background model won't become biased. The same approach is also used in MaxHiC to analyze Capture-Hi-C data. The only difference is that separate sets of parameters are learned for modelling target-target, target-non-target and non-target-non-target interactions to account for the fact that interactions involving one or two of the targeted regions will be enriched to different levels compared to non-targeted regions. A schematic of the MaxHiC model for analysis of general Hi-C is shown in **Figure 1a** and a detailed explanation of both models is provided in the methods section.

MaxHiC vs other leading Hi-C tools

In order to evaluate the performance of MaxHiC, we performed a comprehensive benchmarking between MaxHiC, GOTHiC, CHiCAGO, Fit-Hi-C, Fit-Hi-C2 and HiCCUP (7) on several published Hi-C datasets (see more details about the samples in **Supplementary table S1**).

We first used MaxHiC to identify significant interactions (at 1kb, 5kb and 10kb bin sizes) in two publicly available Hi-C data from Rao *et al.* (7) (on the GM12878 EBV

transformed B-lymphoblastoid cell line and human mammary epithelial cells (HMEC)) and one set capture Hi-C data from Mifsud *et al*. (13) (also on GM12878). For each library we also used GOTHiC, CHiCAGO, Fit-Hi-C, Fit-Hi-C2, and HiCCUP with their default parameters to identify significantly interacting regions. **Table 1** and **Supplementary table S2** summarise the numbers of significantly interacting pairs found using each method. **Supplementary figure S1** shows Venn diagrams of the interacting pairs identified by each method at 10kb bin sizes. Notably the number of interactions called as significant by each method varied greatly (from 10,485 using HiCCUP, to 8,136,100 using GOTHiC).

The sets of interactions called as significant or insignificant by each method (**Figure 1b** for sample GM12878 and **Supplementary figure S2a** for HMEC) had distinct distance distributions. At 10kb bin size, GOTHiC and HiCCUP had the smallest median distances between interacting regions of 170kb, while that for MaxHiC, Fit-Hi-C, Fit-Hi-C2 and CHiCAGO were 270kb, 7,590kb, 7,426kb and 93,213kb respectively. The low median length observed for GOTHiC is reflected in the observation that it calls almost all putative interactions below 100kb as significant.

Comparing the interactions called as significant by each method revealed interactions identified by MaxHiC had substantially more read support than those identified by all methods except HiCCUP, however HiCCUP only predicted significant interactions for a small range of distances ~50kb to ~5000kb. Both the average number of reads and minimum number of reads supporting a significant interaction were higher for those called by MaxHiC than other models (**Table 1, Supplementary tables S2**). By examining the average read count of significant and insignificant interactions called by each method (**Figure 1c** for GM12878 and **Supplementary figure S2b** for HMEC), it is revealed that at all distances, significant interactions called by MaxHiC and Fit-Hi-C had substantially more reads than those called as insignificant. GOTHiC only demonstrated such trend at distances above 100kb. At all distances, paired regions called as significantly interacting by MaxHiC had higher average read count support than the other five methods. This was also reflected in the minimum number and average number of reads required to call an interaction as significant (**Supplementary figure S3**). We observed similar pattern when considering interactions calculated using different fragment sizes (**Supplementary figure S4** shows the same plots using

1kb, 5kb, 10kb, and 50kb bin sizes). Note: HiCCUP was not able to run on 1kb bin size and found no loops when run on 50kb bin size.

Evidence of regulatory potential for interacting regions identified by MaxHiC

We next investigated whether interacting regions identified by each method as significant were enriched for genomic and epigenetic annotations indicative of regulatory regions. Comparison of the interacting regions with histone mark profiles generated for GM12878 and HMEC by the ENCODE consortium (14-16) revealed the regions identified by MaxHiC were significantly enriched for H3K9ac, H3K27ac, H3K79me2, H3K4me3, H3k4me1, and H4K20me1 marks (**Figure 2**). They were also enriched for DNAse hypersensitive sites and CTCF binding but not for the repressive mark H3K27me3. Importantly there was very little if any enrichment of these features in regions called as significant by GOTHiC, CHiCAGO, Fit-Hi-C, and Fit-Hi-C2. Interestingly HiCCUP had similar levels of enrichment however MaxHiC identified 34 fold more interactions. For the GM12878 analysis we also observed significant enrichment of transcription factor binding sites from ENCODE ChIP-seq data (**Supplementary figure S5 and Supplementary table S3**). Again, MaxHiC and HiCCUP showed similar levels of enrichment for GM12878 specific transcription factor binding sites. Together this suggests that both MaxHiC and HiCCUP identify interactions between regions with regulatory potential however MaxHi-C identifies substantially more (34 fold).

Next, the interacting regions identified by MaxHiC were significantly enriched for ENCODE annotated enhancers, promoters, insulators and transcription states (17, 18) (**Figure 2**). They were also significantly depleted of regions annotated as heterochomatin. Again, HiCCUP showed similar levels of enrichment while Fit-Hi-C and Fit-Hi-C2 showed some lower level of enrichment for some of the chromatin states. Surprisingly regions identified by CHiCAGO were depleted of these states.

Besides the regulatory regions predicted by ENCODE, significant interactions identified by MaxHiC also contained a much higher fraction of promoters and enhancers identified by FANTOM5 as active in GM12878 and HMEC (in comparison to all FANTOM5 promoters and enhancers) (**Figure 2**) indicating that MaxHiC not only

enriches for promoters and enhancers but those active in the cell type from which the Hi-C data was generated.

Lastly, interacting regions identified by MaxHiC and HiCCUP on the GM12878 dataset were significantly more likely to overlap autoimmune SNPs identified in GWAS while those identified in the HMEC were significantly more likely to overlap breast trait associated SNPs (**Figure 2,** GWAS traits considered are provided in **Supplementary table S4**). No such enrichment was observed for SNPs associated with unrelated traits such as kidney, lung, and liver traits.

MaxHiC enriches for interactions between regulatory regions

We next examined the fractions of interacting pairs identified by each method involving regulatory elements at both ends. We specifically focused on promoter-enhancer, promoter-promoter, enhancer-enhancer and insulator-insulator pairs. In the GM12878 dataset approximately 12% of the raw unfiltered pairs fell into at least one of the above categories (**Figure 3a**). Strikingly, the read pairs identified by MaxHiC were almost 4 times (47%) more likely to fall into one of these four categories (**Figure 3a**). Notably ~17% of the pairs corresponded to promoter-enhancer interactions, ~15% enhancer-enhancer, ~8% promoter-promoter and ~7% to insulator-insulator interactions. Similar levels of enrichment were observed for HiCCUP, but as noted above HiCCUP identified 34 fold fewer significant interactions. In contrast, the other methods had distributions similar to the unfiltered pairs with Fit-Hi-C, Fit-Hi-C2 and GOTHiC showing fractions slightly above that of the unfiltered pairs. Extending this to any interaction involving an enhancer, promoter or insulator increased the fraction of annotated pairs to 70%. Similar fractions were observed with the HMEC data. **Figure 3b** shows the relative enrichment for interactions involving pairs of chromatin features and predicted chromatin states. Similar to **Figure 2** we observed enrichment of interactions involving active chromatin states and depletion of heterochromatin-heterochromatin pairs using the interactions called as significant by MaxHiC and HiCCUP and to a much lesser extent with Fit-Hi-C and Fit-Hi-C2.

We next used expression quantitative trait loci (eQTL) pairs, identified in EBV-transformed lymphocytes as part of the Genotype-Tissue Expression (GTEx) Project

(19) to examine enrichment of eQTL SNPs in regions that interact with promoters. Approximately 2.1% of the interactions identified by MaxHiC in the GM12878 cell line overlapped at least one eQTL pair. This corresponded to a 15.7 fold enrichment (**Figure 3b**). In contrast GOTHiC, CHiCAGO and Fit-Hi-C's interactions were lower (8.3, 1.8 and 7.9 fold enriched respectively). Matching analyses on HMEC confirmed similar enrichments for regulatory features and eQTLs from mammary tissue (**Figure 3b**).

Lastly using a mouse cortex Hi-C data set (20), and 822 experimentally validated brain specific enhancers from the VISTA enhancer database (21) we demonstrated that interactions identified by MaxHiC and HiCCUP are significantly enriched for experimentally validated enhancer-promoter pairs. In particular, MaxHiC identified 4 fold more validated enhancer-promoter pairs from VISTA than expected. GOTHiC and Fit-Hi-C also enriched for these pairs but only 1.9 and 2 fold, respectively. There was no enrichment in the significant interactions identified by CHiCAGO (**Figure 3b**).

Genomic spacing of interacting regulatory elements

Using the putative regulatory region annotations from the analysis above, we next examined the spacing preferences for interactions between different regulatory features. We first examined preferred spacing for H3K27ac-H3K27ac and CTCF-CTCF structural loops. For MaxHiC, Fit-Hi-C, and Fit-Hi-C2 we observed the spacing between CTCF pairs to be larger (130kb) than H3K27ac-H3K27ac pairs (80kb) (**Figure 4a**). In both cases there were long tails reaching beyond 1Mb. Performing the same analysis on annotated regulatory regions revealed promoter-promoter, enhancer-promoter, and enhancer-enhancer pairs had almost identical spacing preferences (with median spacings of 70kb, 70kb and 90kb respectively), while insulator-insulator pairs were considerably further apart with a median spacing of 170kb (**Figure 4b**). We did not see this trend for HiCCUP however we note HiCCUP did not identify significant interactions below 50kb and above 5Mb (**Figure 1b, c**).

Discussion

Enhancers have important roles in controlling gene expression. Linking distal regulatory regions such as enhancers to the genes that they regulate is critical to the interpretation of regulatory variants identified in whole genome data. Importantly, genomic variants in enhancers and regulatory elements have been linked to human complex diseases (22, 23). Therefore, pairing enhancer regions to their regulatory targets is crucial for diagnosis of genetic diseases. Compared to simplistic assignment of enhancers to their nearest promoters, computational methods have improved the prediction of enhancer-promoter pairs (24-27). However, physical interaction data is still far more convincing in validating the pairs. Hi-C data offers one approach to identify these links, however Hi-C sequencing libraries are extremely noisy because of self-ligations, random ligations artifacts and other complex sources of biases such as GC content and mappability of the sequences. Given this, accurate identification of significantly interacting regions is dramatically affected by the tool used. Here, we have presented MaxHiC, a new tool for identifying significantly interacting regions from Hi-C data (and capture Hi-C data see **Supplementary figure S6** and **Supplementary table S5**). Through a dramatically improved discrimination of biologically relevant interactions from background, we showed that MaxHiC enriches for interacting regions with substantially more evidence of regulatory potential (**Figures 1 and 2**) than those identified by existing Hi-C analysis tools (4-6, 8). This is a critical advance as enhancer-promoter interaction maps are needed to interpret the impact of non-coding variants identified in intergenic regions.

From our analyses, only MaxHiC and HiCCUP (and to a much lesser extent Fit-Hi-C/2) enrich for interactions between regulatory regions. Specifically, interactions identified by MaxHiC were enriched for regulatory features, disease-associated genome-wide association SNPs, eQTL pairs and experimentally validated enhancer-promoter interactions. Although interacting regions identified by HiCCUP had similar levels of enrichment it found 34 fold less significantly interacting regions. Additionally, HiCCUP appears to have limitations on the distances at which interactions can be found and the bin sizes possible. It also required specialised GPU architecture to run. From this observation we conclude that MaxHiC is a substantially better tool than existing Hi-C analysis packages.

We also implemented a capture version of MaxHiC to analyse capture Hi-C experiments; in the capture Hi-C technology, target regions (also called baits) are sequenced with higher depth than the others, resulting in three different types of interactions target-target, target-other and other-other with different properties. Capture Hi-C baits, but not other ends, have a further source of bias associated with uneven capture efficiency, thus for capture Hi-C experiments we model each class separately. As **Supplementary figure S6** shows the capture version of MaxHiC also identifies target-interacting regions that are enriched for regulatory features and disease-associated GWAS SNPs. The levels of enrichment are much higher than the regions identified by GOTHiC, Fit-Hi-C, and Fit-Hi-C2. Same as the non-capture Hi-C analysis, MaxHiC and HiCCUP show similar levels of enrichment in capture data, however, MaxHiC identified 36 times more interactions than HiCCUP (36 times more interactions at 10kb fragment; 7.3 times more interactions at 5kb fragment). Surprisingly, the average read count of significant interacting regions identified by MaxHiC were much higher than HiCCUP's interactions (1.4 fold and 2.4 fold in 10kb and 5kb fragments, respectively). Importantly, MaxHiC's interacting regions had higher level of enrichment than the interacting regions identified by CHiCAGO, a specific model designed for the analysis of capture Hi-C data.

In conclusion, MaxHiC is a new open source tool for identifying significant interacting regions from Hi-C and capture Hi-C data. We have demonstrated that it significantly outperforms existing tools in terms of enrichment of interactions between known regulatory regions. We will continue to develop MaxHiC related tools for identifying topologically associating domains (TADs) and for identifying interactions that are significantly different between conditions. As single-cell Hi-C data becomes more broadly available we will also adapt MaxHiC to work with these datasets. Lastly, we believe that with minor modifications the same principles used here to analyse Hi-C data may be more broadly extended to identify significant interactions from RNA-chromatin interaction datasets (24, 25).

Online methods

Tool availability

The MaxHiC source code and python package, a sample dataset, and instructions on how to run MaxHiC are provided at https://github.com/forrest-lab/MaxHiC and https://github.com/Rassa-Gvm/MaxHiC. For more details about each parameter, please visit the github page. A tutorial on using the MaxHiC package is also provided in **supplementary file 1** (instructions for MaxHiC). MaxHiC is memory and speed optimized and will work with both single core and multicore machines. It can work with very large Hi-C data sets (as bam file or contact maps generated by other tools) to identify significant interactions at a broad range of resolutions (1kb-500kb).

Publicly available Hi-C data and data access

Samples GM12878 B-lymphoblastoid cell line (SRR1658572) and Human Mammary Epithelial Cell (SRR1658680) from Rao *et al*. (7), capture Hi-C sample GM12878 B-Lymphoblastoid (ERR4360330) from Mifsud *et al*. (13) and mouse cortex Hi-C sample (SRX128219) from Dixon *et al*. (20) were used to benchmark MaxHiC against other models. All samples are publically available. More details about the samples are provided in **Supplementary table S1**.

Mapping, filtering and Interaction calling

MaxHiC takes as input either bam files or contact map files generated by packages such as HiCUP (26) and HiC-Pro (27). For the analyses presented here, FASTQ paired-end reads were aligned to the human hg19 genome and filtered through the HiC-Pro pipeline developed in our lab (statistics are presented in **Supplementary table S1**). To remove duplicate reads, filter for valid interactions, and generate Hi-C interaction matrices (5kb and 10kb bin-sizes), we set the following parameters: MIN_FRAG_SIZE = 230; MAX_FRAG_SIZE = 1100; MIN_INSERT_SIZE = 120; and MAX_INSERT_SIZE = 990. All experimental artifacts, such as circularized reads and re-ligations, singletons have been filtered out using HiC-Pro. We also removed all self-

interactions from the downstream analyses. In-house scripts have then used to statistics evaluation of Hi-C libraries (see **Supplementary table S1**).

The MaxHiC algorithm

In the following, we explain MaxHiC algorithm and mathematics behind the model for the analysis of both Hi-C and capture Hi-C experiments.

A negative binomial regression background model for Hi-C data

In MaxHiC, a negative binomial distribution has been used to model the read count of interactions based on the bias factors of the two ends and their genomic distance. The negative binomial distribution is the analytic form of a Poisson distribution with its mean parameter following a gamma distribution to account for over-dispersion in data. As shown in equation 1, the observed read count of each interaction, x_{ij}, is modelled using its expected read count, μ_{ij}, and a shared dispersion factor, r, between all interactions.

$$x_{ij} \sim NB(\mu_{ij}, r) \quad (1)$$

The parameters of the model are learned using a maximum likelihood approach. Gradient descent has been used to find the values of the parameters maximizing the logarithm of the likelihood function. In order to have higher accuracy, by having a higher difference between read count of meaningful and meaningless interactions, and also to have higher speed, the proposed model uses the interactions between DNA bins, not DNA fragments. DNA is divided into pieces and all reads recorded for each fragment are assigned to the bin containing their middle points. All of the recorded interactions between pairs of bins with at least one read are the input samples used for training the model. After the model has been trained, the P-value of each interaction can be calculated based on the value of reliability, *1 - CDF*, for its observed read count.

Calculating the expected value for the read count. The expected read count for the interaction between bins i and j is calculated according to equation 2. $f(d_{ij})$ indicates the expected value of read count to be recorded in the experiment for any two loci with distance d_{ij}, v_i and v_j are bias factors of the two interacting bins. As mentioned above, bins have different characteristics and different probabilities of showing up in the

experiment that affects the expected value of their interactions. It's assumed that the effects of two bins are independent from each other, so the total bias factor has been calculated as a product of the bias factors of the two ends.

$$\mu_{ij} = v_i \times v_j \times f(d_{ij}) \quad (2)$$

Calculating bias factors for bins. In order to be applicable to all types of Hi-C protocols, accounting for the unknown sources of bias and also the relation between the known sources, the second approach, mentioned in the introduction, has been adopted in this study. A uniform bias factor has been assumed for each bin, which becomes more valid as the resolution of the analysis increases and the length of bins become shorter. We expect the bins included in more interactions with higher reads to have higher tendency to show up in the experiment. In addition, the total number of interactions that a bin has been included in, would have no effect on its bias. To deal with these problems, the bias factor for bins have been calculated using equation 3 in which v_{base}, b and b_m are the parameters of the function and s_i is the total number of reads observed for bin i. Here it is assumed that the total number of reads observed for each bin is a measure of its ability to show up in the final results of the test, and bias factor is assumed to be a linear function of s_i in log-log space. v_{base} is also considered as the base possible value for the bias factor. This prevents the bins with very few recorded reads to affect the parameters in the training process and it also prevents their related interactions with very few reads from showing up as significant because of their abnormally low bias. Many tools do this by putting a threshold and filtering bins with less bias, but in MaxHiC it is done automatically without any need for adjusting the threshold. As illustrated in equation 3, v_i is proportional to s_i raised to a power *b.* The power can be learned and if it becomes less than one it can express a saturation phenomena. As duplicated read pairs are discarded in post-processing, having more read count for two interacting bins as the read count of their interaction increases, becomes lower than their initial potential.

$$v_i = maximum\left(e^{b \times \ln(s_i) + b_m}, v_{base}\right) \Rightarrow v_i = maximum(s_i^b \times e^{b_m}, v_{base}) \quad (3)$$

Calculating expected read count as a function of distance. The expected read count as a function of distance for cis interactions is illustrated in equation 4. The function is the maximum of two values. The first one, $f^v(d_{ij})$, is a degree 3 polynomial

function of genomic distance in log-log space which specifies the expected read count of interactions based on the linear genomic distance between the two interacting loci. The reason for using an exponential function is the fact that Brownian motion and diffusion can be modelled using an exponential function. This assumption is also consistent with the power law function observed for the read count as a function of distance which has been introduced by Lieberman *et al*. for the first time (28). The degree of 3 has been used so the function can have varying curvature in different distances. In this function, the logarithm of distance has been used instead of the distance itself, as DNA is not a straight string and has multiple folding with different scales that causes proximity to have a closer relation with the logarithm of the length than the length itself. A similar concept has been used in Fit-Hi-C (4), CHiCAGO (5) and HiC-DC (2) methods. The biggest difference is the fact that all of these methods consider a strictly decreasing function for the expected read count which approaches zero as the genomic distance increases and causes all interactions between far genomic loci to be considered significant while they may have only one read. This cannot be completely logical as farther than some specific distance, random collisions occur due to three dimensional closeness not the genomic closeness. This three dimensional closeness can be due to closeness to the fixed three dimensional structures such as loops or random dynamic crossing of DNA string while two different parts of it are having contact. The expected read count for these kinds of random interactions has been modeled using a constant parameter, f^c. Therefore, in close genomic distances $f^{v(d_{ij})}$ is dominating while the value of $f^{v(d_{ij})}$ becomes very small, the value of f^c dominates the expected read count. For trans interactions, similar to the cis interactions for far loci, a constant parameter has been used to denote the expected value of read count for any random trans interactions.

$$f(d_{ij}) = \max\left(f^{v(d_{ij})}, f^c\right) \tag{4}$$

$$f^{v(d_{ij})} = e^{(a_3.\ln(d)^3 + a_2.\ln(d)^2 + a_1.\ln(d) + a_0)}$$

Maximum function is not a smooth function to use in optimization. Therefore, instead of maximum, the soft-maximum function has been used. This soft-maximum resembles the maximum function. When an argument, for example α, is larger than the other, it's exponential becomes far more larger and so it's related factor, γ_α^τ, approaches 1 while

the other factor, γ_{β}^{τ}, approaches 0. Therefore, the final value will be approximately equal to α itself.

$$\tau = softmax(\alpha, \beta) \Longrightarrow \tau = \gamma_{\alpha}^{\tau} \times \alpha + \gamma_{\beta}^{\tau} \times \beta \quad (5)$$

$$\gamma_{\alpha}^{\tau} = \frac{e^{k \times \alpha}}{e^{k \times \alpha} + e^{k \times \beta}}, \gamma_{\beta}^{\tau} = \frac{e^{k \times \beta}}{e^{k \times \alpha} + e^{k \times \beta}}$$

Training the model in multiple rounds. As mentioned above, the method models read count of interactions occurring based on Brownian motion, not the meaningful chromatin contacts, which are called real interactions from now on. But real interactions are also among the data and can affect the parameters of the model and pull them up. For example, in distances where there are more read interactions, the expectation may become higher than other distances. To deal with this bias, the model is trained in multiple rounds, which can be specified by user (the default value for iterations is 4). In each round, first the parameters of the model are learned using a stochastic gradient descent approach. The training is done in iterations until all the parameters of the model have changed by less than 0.0001 compared to the previous iteration. Training also will stop when reaching 1000 iterations to handle oscillating situations around the optimum values. This number is high enough to give the model enough time to be trained well and has been chosen by the developers based on running the model on multiple Hi-C datasets (25 samples with 2-4 different resolutions). Next, a P-value is calculated for each interaction using the trained model and the real interactions are identified based on a fixed user-defined P-value as significance limit. The read count of these real interactions is substituted with their expected read count for calculating the total number of reads for each bin, as if they were not read what would their read count be, to remove the bias of having real interactions from bias factors of the bins. They are also completely ignored in the training of the model's parameters. This is another main difference between MaxHiC and other models.

An automatic way of removing the effect of noisy data. Typically ,a high percentage of interacting fragments in Hi-C have only 1 read between; for example, the typical percentage can be above 85% for a resolution of 5Kb. End bins of many of these interactions have very low coverage which results in a very small expected value for the interaction. This can pull the value of parameters down for the benefit of

themselves in the model. In order to cope with such noisy data, the expectation for read count is calculated based on equation 6 instead of equation 2. In this way, the maximum of expected value and a base expected read count, $base_e$, is used as the final expected value. This will prevent the mentioned noisy interactions to affect greatly the parameters of the model.

$$\mu_{ij} = softmax(v_i \times v_j \times f(d_{ij}), base_e) \quad (6)$$

A negative binomial regression background model for capture Hi-C data

Capture-Hi-C is a version of Hi-C in which some regions in DNA, which are called *Baits* or *Target Regions*, are sequenced with higher depth than the others. This will result in three different types of interactions in Capture-Hi-C with different properties, Target-Target, Target-Other and Other-Other. The background model developed for Capture-Hi-C follows the same statistical model as the one developed for General-Hi-C, but in order to account for the differences between these types of interactions, three different sets of parameters are used, each for one interaction type. Each set of parameters is learned separately from the others based on its own interactions, in the process of learning the parameters of the model. In calculating the bias factors for bins, the sum of read counts cannot be used directly as different interactions correspond to different types. To remove this barrier, all the read counts of interactions are converted to their equivalent read-counts in Target-Target interaction type:

$$x_{ij}^{t2} = x_{ij}^{t1} \times \frac{\mu_{ij}^{t2}}{\mu_{ij}^{t1}} \quad (7)$$

In equation 7, t_1 and t_2 stand for the original interaction type and the one we want to convert the read-count to, which is assumed to be Target-Target, respectively. The expected value for each interaction type is easily calculated by using the set of parameters related to that interaction type and the calculated properties of the interaction, sum of reads for bins and their genomic distance, in the expected read-count formula regardless of the original type of the interaction.

Assessment of feature enrichment (ENCODE regulatory features, FANTOM5 promoters/enhancers)

Annotation datasets. We downloaded tissue-specific ChIP-Seq data in the form of processed peak calls for histone modifications (e.g. H3K27ac, H3K4me1, etc), CTCF binding sites, DNase hypersensitive sites, and ChromHMM predicted promoters and enhancers from the ENCODE project (17). We used FANTOM5 active promoters and transcribed enhancers from the FANTOM5 consortium (29, 30). FANTOM5 tissue-specific enhancers were obtained from (30).

Region-based analysis. To compute the enrichment of interactions overlapping an epigenetic feature, we first calculated the fraction of significant interactions with at least one end overlapping a feature. We then calculated fraction of all interactions overlapping with the feature as the genome-wide expectation. The enrichment value was then calculated by dividing the fraction of significant interactions overlapping a feature by the genome-wide expectation.

$$\textbf{Enrichment value} = \frac{\text{fraction of significant interactions overlapping a feature}}{\text{fraction of all interactions overlapping a feature}} \quad (8)$$

Pair-based analysis. To compute the enrichment of interactions that both sides of an interaction overlap with an epigenetic feature, we repeated the above analyses in which both sides of interaction overlap with the feature.

Capture Hi-C analysis. In the capture Hi-C analysis, to ensure our results were not affected by the strong peak signal over target regions, only target-interacting regions ("bait-to-any interactions") are considered and signals that mapped as bait-bait interactions and those signals that mapped outside of the target-interacting regions are excluded from the analysis. To calculate the fraction of significant target-interacting regions overlapping an epigenetic feature in the capture Hi-C analysis, we first calculated the fraction of significantly target-interacting regions overlapping a feature. We then randomly selected fragments that had no interaction with the target regions (we selected exactly the same number of fragments as the number of significant target-interacting fragments). The enrichment value then calculated by dividing the fraction of significant interacting regions overlapping a feature on the average fraction of 1000

permutations. We performed the above analysis for each Hi-C background model, separately.

$$\textbf{Enrichment value in capture Hi-C data} = \frac{\text{fraction of significant interactions overlapping a feature}}{\text{average fraction of random interacting regions overlapping a feature in 1000 permutations}} \quad (9)$$

GWAS data analysis

GWAS SNPs for autoimmune disease, neurological/behavioural traits and kidney/liver/lung disorders were downloaded from Maurano *et al*. (31). To compute the enrichment of interactions overlapping GWAS SNPs, we first calculated the fraction of significant interactions with at least one end containing GWAS SNPs. We then calculated fraction of all interactions overlapping GWAS SNPs as the genome-wide expectation. The enrichment value was then calculated by dividing the fraction of significant interactions overlapping GWAS SNPs by the genome-wide expectation. In the analysis of capture Hi-C library, we only considered overlapping of GWAS SNPs with target-interacting regions in both the significant target-interacting regions and the permutated target-interacting regions.

VISTA enhancer analysis

VISTA gene-enhancer pairs were downloaded from VISTA enhancer database (https://enhancer.lbl.gov; (21), mouse mm9 assembly (download date: 20/03/2019). In total, 822 mouse brain-related vista enhancer were found in the vista enhancer database. To calculate the fraction of vista gene-enhancer pairs overlapping with significant interactions identified by the four models, we considered those Hi-C interactions where one side of the interaction overlapped the enhancer and the other overlapped the promoter of the target gene identified by VISTA. To calculate the enrichment we divided these by the genome-wide expectation.

Overlapping Hi-C interactions with eQTL pairs

We used the set of GTEx v7 eQTLs identified as significant in EBV-transformed lymphocytes from the Genotype-Tissue Expression (GTEx) Project (19). To calculate the fraction of eQTL pairs overlapping with significant interactions, we considered those Hi-C interactions where one side of the interaction overlapped an eQTL SNP and the other side overlapped the promoter of the eQTL target gene identified by GTEx. To calculate enrichment, we divided these by the genome-wide expectation.

Author contributions

HAR, ARRF, and RG designed the study. HAR, ARRF, and RG wrote the manuscript; HAR, ARRF, RG, and HRR edited the manuscript; RG implemented the tool with help from HAR. RG and HRR designed the statistical background model. HAR and RG carried out all the analyses including the statistical analyses, significant region identification, benchmarking, annotations, and text mining. HRR and NR helped with the statistical analysis. HAR generated figures 2, 3, 4, S1, S5, S6 and all tables. RG generated figures 1, S2, S3, S4 with supervision from HAR. ARRF provided his advices in generating figures. All authors have read and approved of the final version of the paper.

Conflict of interest

The authors declare no competing financial and non-financial interests.

Funding

This work was carried out with the support of an Australian Research Council Discovery Project grant (DP160101960) to ARRF and a UNSW Scientia Fellowship to HAR.

Acknowledgments

This work was supported by an Australian Research Council discovery project grant to ARRF (DP160101960) and a WA Department of Health Near-Miss Merit Award to HAR. ARRF is currently supported by an Australian National Health and Medical

Research Council Fellowship APP1154524. ARRF was also supported by funds raised by the MACA Ride to Conquer Cancer and a Senior Cancer Research Fellowship from the Cancer Research Trust. HRR and RG were supported by IRN National Science Foundation (INSF), Grant No. 96006077. HAR is supported by a UNSW Scientia Fellowship. HAR was previously a research associate in Harry Perkins Institute of Medical Research under supervision of ARRF. Analysis was made possible with computational resources provided by the Telethon Kids Bioinformatics Server with funding from the Australian Government and the Government of Western Australia.

References

1. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28(23):3131-3.
2. Carty M, Zamparo L, Sahin M, Gonzalez A, Pelossof R, Elemento O, et al. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun. 2017;8:15454.
3. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999-1003.
4. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. 2014;24(6):999-1011.
5. Cairns J, Freire-Pritchett P, Wingett SW, Varnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 2016;17(1):127.
6. Mifsud B, Martincorena I, Darbo E, Sugar R, Schoenfelder S, Fraser P, et al. GOTHiC, a probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. PLoS One. 2017;12(4):e0174744.
7. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80.
8. Kaul A, Bhattacharyya S, Ay F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc. 2020;15(3):991-1012.
9. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
10. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
11. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148(3):458-72.
12. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint. 2014;arXiv:1412.6980.

13. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47(6):598-606.
14. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
15. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010;28(8):817-25.
16. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43-9.
17. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
18. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41(2):827-41.
19. Consortium G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60.
20. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376-80.
21. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35(Database issue):D88-92.
22. Forrester WC, Epner E, Driscoll MC, Enver T, Brice M, Papayannopoulou T, et al. A deletion of the human beta-globin locus activation region causes a major alteration in chromatin structure and replication across the entire beta-globin locus. Genes Dev. 1990;4(10):1637-49.
23. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet. 2003;12(14):1725-35.
24. Bonetti A, Agostini F, Suzuki AM, Hashimoto K, Pascarella G, Gimenez J, et al. RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. Nat Commun. 2020;11(1):1018.
25. Li X, Zhou B, Chen L, Gou LT, Li H, Fu XD. GRID-seq reveals the global RNA-chromatin interactome. Nat Biotechnol. 2017;35(10):940-50.
26. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Res. 2015;4:1310.
27. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;16:259.
28. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93.
29. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462-70.
30. Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJ, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5' ends. Nature. 2017;543(7644):199-204.
31. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190-5.

Tables

Table1. Statistical summary of significant interactions identified by MaxHiC, GOTHiC, CHiCAGO, Fit-Hi-C, Fit-HiC2, and HiCCUP in the GM12878 Hi-C library from Rao *et al.* (7). at bin-sizes 5kb and 10kb. Statistical summaries for HMEC library are provided in **Supplementary tables S2**, respectively.

1kb	MaxHiC	GOTHiC	CHiCAGO	Fit-Hi-C	Fit-Hi-C2	HiCCUP*	All
percentage of significant interactions	0.03	9.4	-	16.6	17.1	-	100
average read-count of interactions	18.1	2.9	-	1.1	1.5	-	1.2
min read-count of cis interactions	4	2	-	1	1	-	1
median distance of cis interactions (kb)	3	14	-	30832	29426	-	544
5kb							
percentage of significant interactions	0.17	9.25	0.17	0.8	0.9	9.40E-05	100
average read-count of interactions	21.3	7.85	1.29	5.6	5.9	32.39	1.77
min read-count of cis interactions	3	2	1	1	1	11	1
median distance of cis interactions (kb)	130	85	101350	5235	5014	110	15173
10kb							
percentage of significant interactions	0.32	7.32	0.02	1.06	1.1	9.25e-05	100
average read-count of interactions	29.4	15.17	1.83	12.3	13.5	58.62	2.38
min read-count of interactions	3	2	1	2	2	11	1
median distance of cis interactions (kb)	270	170	93213	7590	7426	150	6990

* HiCCUP calls interactions with FDR < 0.05 as significant. For all other methods we used a threshold of P-value < 0.001 to identify significant interactions.

Figure legends

Figure 1. Background modelling as a function of distance in MaxHiC and other HiC tools. a) MaxHiC mathematical model. **b)** Average read-count of significant (green), insignificant (red), and all interactions (blue) in different genomic distances identified by different HiC tools at 10kb bin size on the Rao *et al.* GM12878 data. **c)** Number of significant interactions identified by the six models in different genomic distances on Rao *et al.* GM12878 sample at fragment size 10kb. HICCUP doesn't have any significant interactions in in less than 50kb or more than 5Mb.

Figure 2. Enrichment of regulatory potential in regions identified by MaxHiC and other tools. a) Enrichment of GM12878 specific histone marks, putative regulatory regions and GWAS signal in regions identified by MaxHiC in the Rao *et al.* GM12878 dataset. **b)** As in (a), but enrichment of signals in the HMEC dataset is shown. ENCODE chromatin states are from GM12878 and HMEC data respectively. For FANTOM5, enrichment is shown for all FANTOM5 promoters and enhancers and for those only active in GM12878 and HMEC. GWAS SNPs for autoimmune disease, neurological/behavioural traits and kidney/liver/lung disorders (**Supplementary table 4**) were downloaded from Maurano *et al*. (31). Note: HiCCUP calls interactions with FDR < 0.05 as significant. For all other methods we used a threshold of P-value < 0.001 to identify significant interactions.

Figure 3. MaxHiC enriches for interactions between regulatory regions, eQTL-target pairs and known enhancer-promoter pairs. a) Fraction of regulatory elements in the significant interactions identified by the four models. **b)** Enrichment of cell line specific chromatin pairs, chromatin states pairs, and eQTL pairs that overlapped with both ends of the significant interactions identified by the six models at fragment size 10kb. The enrichment is calculated by dividing of the fraction of the features overlapping with both ends of the significant interactions identified by each model on fraction of the features overlapping with both ends of all interactions (*see methods*). **c)** Enrichment of previously validated promoter-enhancer pairs active in brain (from VISTA (21)) in the significant interactions identified in a mouse cortex Hi-C dataset from Dixon *et al.* (20)].

Figure 4. Distance distribution for interactions between different classes of regulatory elements. a) Comparison of spacing observed for H3k27ac-H3k27ac and CTCF-CTCF structural loops in the interactions identified by MaxHiC in the Rao *et al.* GM12878 dataset. **b)** Comparison of spacing observed for promoter-promoter, enhancer-promoter, enhancer-enhancer and insulator-insulator pairs in the interactions identified by MaxHiC in the Rao *et al.* GM12878 dataset. Plots use fragment bin-size 5kb and window size 5kb.

Supplementary figures legends

Figure S1. Venn diagram of the significant interactions identified by each model in the Rao *et al.* GM12878 and HMEC datasets. **a)** Venn diagram of significant interactions identified by each model in the GM12878 data at fragment size 10kb. **b)** Venn diagram of significant interactions identified by each model in the HMEC data at fragment size 10kb. We did not include Fit-Hi-C2 in this figure as it is almost the same as Fit-Hi-C.

Figure S2. Comparisons of the number of significant interactions and average read-counts of significant interactions identified by four Hi-C background models applied to the HMEC Hi-C data. a) Average read-count of significant (green), insignificant (red), and all interactions (blue) in different genomic distances on the HMEC sample at fragment size 10kb. **b)** Number of significant interactions identified by the four models in different genomic distances on the HMEC sample at fragment size 10kb.

Figure S3. Distance distribution of read-count support of significant interactions identified by the four models in GM12878 and HMEC. a) Min read-count of significant interactions in the Rao *et al.* GM12878 data, **b)** as in a but for HMEC, **c)** Average read-count of significant interactions in the Rao *et al.* GM12878 data and **d)** as in c but for the HMEC data. Fragment size 10kb.

Figure S4. Comparisons of the number of significant interactions and average read-counts of significant interactions identified at fragment sizes 1kb, 5kb, 10kb, 50kb. Plots as in **Figure1.** Note: HiCCUP did not work on fragment size 1kb and did not identify any significant chromatin loops using fragment size 50kb.

Figure S5. Enrichment of transcription factor binding sites from ENCODE ChIP-seq data on Rao *et al*. GM12878 data. As the plot shows the enrichments in MaxHiC's interactions are similar to the enrichments in the chromatin loops identified by HiCCUP (which identified 34 fold fewer interactions).

Figure S6. Comparison of the four models in the analysis of capture Hi-C experiment in capture Hi-C sample GM12878. We also developed a separate version of MaxHiC to handle capture Hi-C libraries. In the capture Hi-C library, we are particularly interested to identify significant interactions that at least one end of interaction is baited (refer to capture regions). **a)** Enrichment of GM12878 specific histone marks, putative regulatory regions and GWAS signal in regions identified by CHiCAGO, GOTHiC, Fit-Hi-C, Fit-Hi-C2, HiCCUP and capture version of MaxHiC, on the capture Hi-C library GM12878 from Mifsud *et al.* (13). This showed that significant target-interacting pairs identified by MaxHiC and HiCCUP have much higher enrichment for regulatory pairs and Autoimmune related GWAS SNPs. Importantly, MaxHiC identified 36 fold more interactions than HiCCUP but at the same level of enrichments for regulatory features.

Supplementary tables legends

Supplementary table S1. Details of publicly available Hi-C samples used in this study.

Supplementary Table S2. The statistical summaries of significant interactions identified by MaxHiC, GOTHiC, CHiCAGO, Fit-Hi-C, Fit-Hi-C2, and HiCCUP on the Rao HMEC Hi-C library at fragment sizes 1kb, 5kb and 10kb.

Supplementary Table S3. Enrichment of transcription factor binding sites in the significant interactions identified by MaxHiC, GOTHiC, CHiCAGO, Fit-Hi-C, Fit-Hi-C2, and HiCCUP on the Rao GM Hi-C library at fragment sizes 10kb.

Supplementary Table S4. Details of autoimmune and breast cancer traits used in this study for GWAS analysis.

Supplementary Table S5. The statistical summaries of significant interactions identified by MaxHiC, GOTHiC, CHiCAGO, Fit-Hi-C, Fit-Hi-C2, and HiCCUP on the Mifsud GM12878 capture Hi-C library at fragments 5kb and 10kb. As the table shows, the average read count of significant interactions identified by MaxHiC are much higher than those identified by CHiCAGO, GOTHiC, Fit-Hi-C, Fit-HiC2 and HiCCUP.

Figure 1- MaxHiC versus other models on sample Rao (GM12878 Lymphoblastoid)

a) MaxHiC model

negative binomial regression model:

$$x_{ij} \sim NB(\mu_{ij}, r)$$

expected value as a function of observed read count and distance:

$$\mu_{ij} = softmax(v_i \times v_j \times f(d_{ij}), base_e)$$

bias factors for bins:

$$v_i = \exp(b \times ln(s_i) + b_m) => v_i = s_i^b \times \exp(b_m)$$

function of distance:

$$f(d_{ij}) = \max(f^v(d_{ij}), f_c^c)$$

$$f^v(d_{ij}) = \exp\,(a_3.\ln(d)^3 + a_2.\ln(d)^2 + a_1.\ln(d) + a_0)$$

multiple iterations for training the model using Maximum Likelihood estimation

b) number of interactions vs distance

c) average read count vs distance

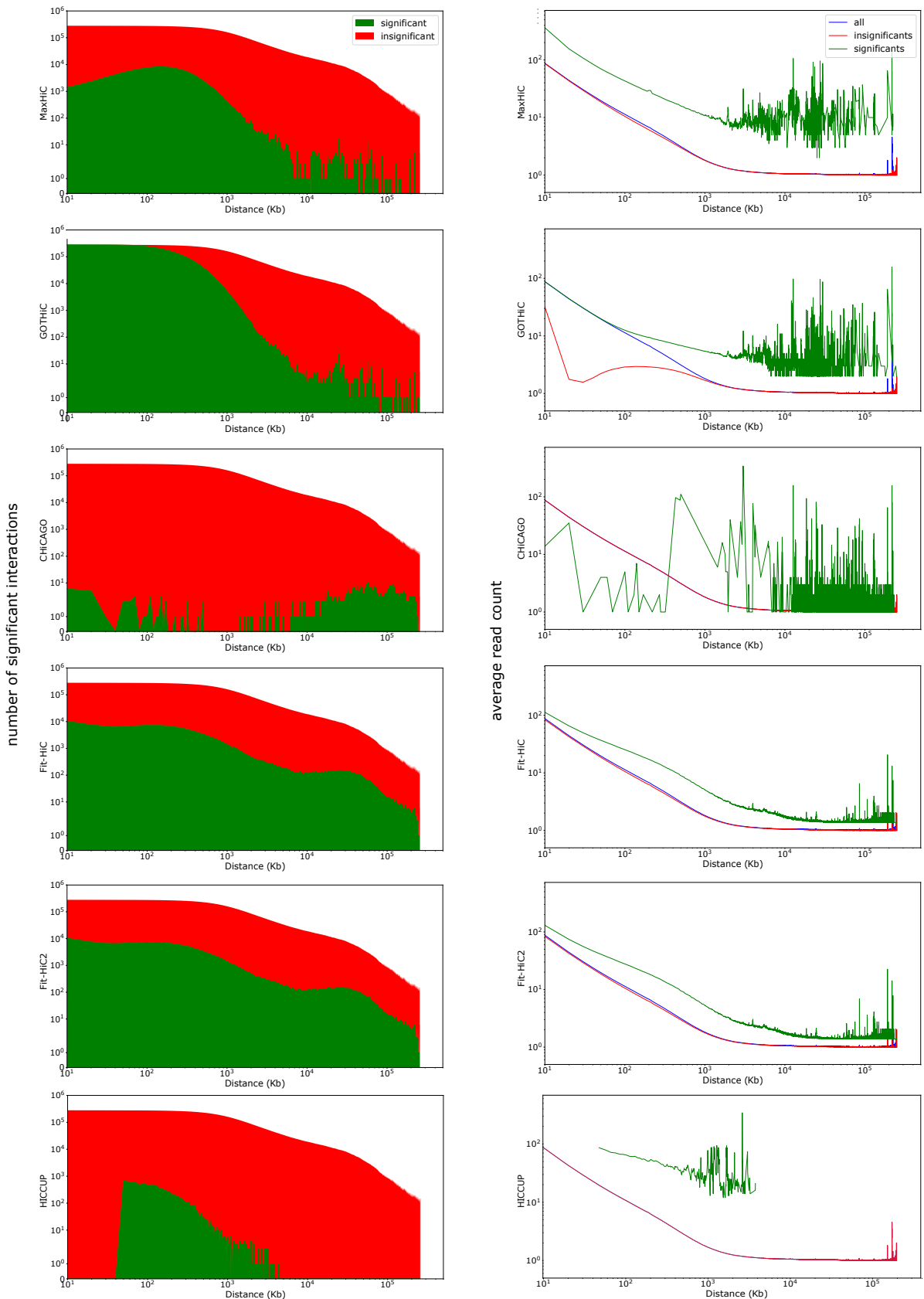


Figure 2 – enrichment in region-based analysis

a) GM12878

		MaxHiC	GOTHiC	CHiCAGO	Fit-HiC	Fit-HiC2	HiCCUP
ENCODE chromatin marks	H3k9ac	2.26	-1.07	-1.43	1.15	1.25	2.34
	H3k27ac	1.98	-1.07	-1.47	1.07	1.27	1.99
	H3k79me2	1.94	-1.10	-1.45	1.09	1.22	1.97
	CTCF	1.91	-1.01	-1.14	1.04	1.26	1.97
	H3k4me3	1.76	-1.04	-1.36	-1.07	-1.00	1.75
	DNase	1.74	-1.02	-1.23	-1.00	1.11	1.72
	H3k4me1	1.63	-1.06	-1.41	1.05	1.17	1.67
	H4k20me1	1.37	-1.15	-1.15	1.03	1.13	1.34
	H3k27me3	-1.09	-1.14	-1.21	-1.06	-1.04	-1.04
ENCDOE chromatin states	Active_Promoter	2.72	-1.03	-1.30	1.27	1.43	2.75
	Strong_Enhancer	2.65	-1.07	-1.48	1.15	1.36	2.66
	Weak_Promoter	2.38	-1.02	-1.23	1.21	1.31	2.32
	Txn_Transition	2.37	-1.05	-1.18	1.22	1.29	2.34
	Insulator	2.19	1.03	-1.06	1.19	1.29	2.15
	Weak_Enhancer	1.88	-1.04	-1.35	-1.03	1.18	1.82
	Txn_Elongation	1.80	-1.06	-1.19	1.10	1.21	1.82
	Weak_Txn	1.73	-1.07	-1.28	1.06	1.16	1.72
	Poised_Promoter	1.67	1.05	1.10	1.16	1.25	1.69
	Repetitive/CNV	1.15	-1.01	1.19	1.00	1.03	1.17
	Repressed	1.05	-1.03	-1.17	1.03	1.08	1.07
	Heterochrom/lo	-1.46	1.02	1.00	1.01	-1.02	-1.37
FANTOM5	F5 enhancer (all)	1.53	-1.02	-1.19	1.05	1.15	1.54
	F5 promoter (all)	1.49	-1.02	-1.10	1.07	1.18	1.55
	F5 promoter (GM)	2.31	-1.00	-1.21	1.22	1.35	2.34
	F5 enhancer (GM)	2.41	-1.10	-1.20	1.23	1.37	2.45
GWAS	Autoimmune	3.80	2.01	1.38	2.30	2.53	3.72
	Breast	1.07	1.30	-1.01	-1.02	1.04	1.05
	Kidney & liver & lung	-1.09	-1.18	-1.04	-1.18	-1.08	-1.06

b) HMEC

		MaxHiC	GOTHiC	CHiCAGO	Fit-HiC	Fit-HiC2	HiCCUP
ENCODE chromatin marks	H3k9ac	1.97	-1.12	-1.99	1.11	1.17	1.99
	H3k27ac	1.73	-1.11	-2.19	1.06	1.15	1.78
	H3k79me2	1.57	-1.18	-1.93	-1.01	1.06	1.53
	CTCF	1.94	-1.01	-1.32	1.18	1.26	1.91
	H3k4me3	1.51	-1.18	-1.93	-1.12	-1.00	1.55
	DNase	1.32	-1.03	-1.12	1.04	1.10	1.41
	H3k4me1	1.51	-1.07	-1.66	1.03	1.11	1.56
	H4k20me1	1.15	-1.15	-1.23	1.01	1.04	1.17
	H3k27me3	-1.06	-1.09	-1.04	-1.00	-1.01	-1.03
ENCDOE chromatin states	Active_Promoter	2.55	-1.09	-2.16	1.13	1.19	2.61
	Strong_Enhancer	2.16	-1.10	-2.28	1.09	1.17	2.22
	Weak_Promoter	2.32	-1.05	-1.59	1.15	1.24	2.34
	Txn_Transition	2.17	-1.12	-1.78	1.15	1.27	2.19
	Insulator	2.48	1.08	-1.25	1.28	1.38	2.35
	Weak_Enhancer	1.59	-1.06	-1.60	1.03	1.13	1.52
	Txn_Elongation	1.54	-1.13	-1.64	-1.01	1.09	1.57
	Weak_Txn	1.49	-1.10	-1.64	1.02	1.07	1.43
	Poised_Promoter	1.82	1.05	1.17	1.17	1.24	1.75
	Repetitive/CNV	1.09	-1.10	1.11	-1.15	-1.07	1.12
	Repressed	1.18	-1.00	-1.03	1.04	1.14	1.22
	Heterochrom/lo	-1.34	-1.00	1.03	-1.02	-1.09	-1.27
FANTOM5	F5 enhancer (all)	1.62	-1.04	-1.22	1.07	1.24	1.57
	F5 promoter (all)	1.54	-1.05	-1.18	1.06	1.21	1.58
	F5 promoter (GM)	2.44	-1.09	-1.31	1.42	1.53	2.39
	F5 enhancer (GM)	2.60	-1.10	-1.29	1.43	1.55	2.57
GWAS	Autoimmune	1.38	2.30	1.23	2.14	1.92	1.46
	Breast	4.92	2.14	1.41	2.36	2.55	4.84
	Kidney & liver & lung	1.33	1.36	-1.01	1.29	1.35	1.39

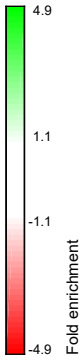
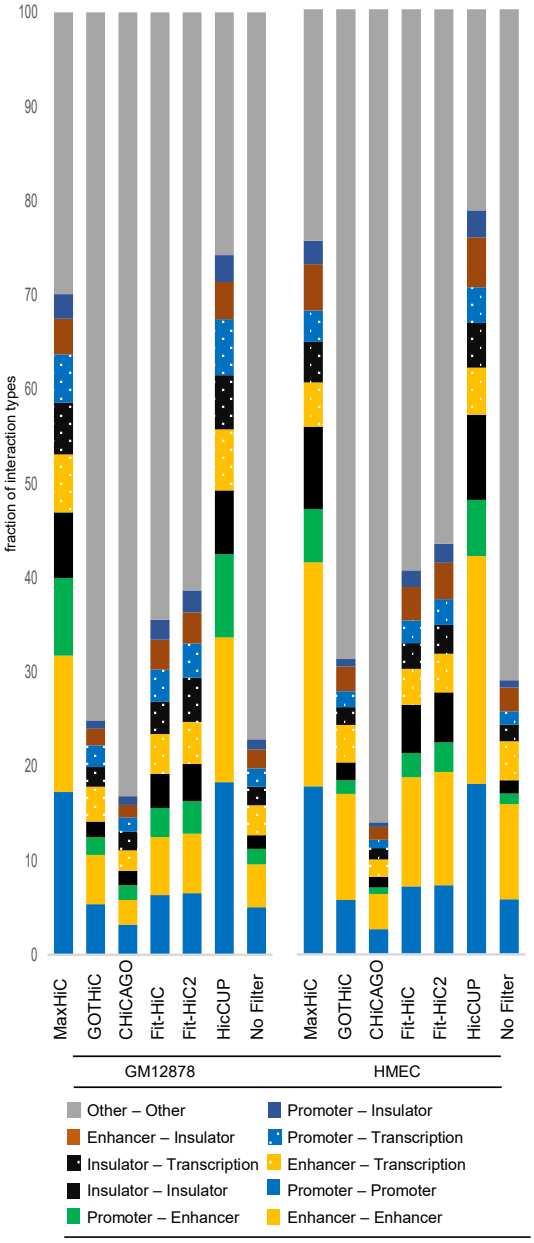


Figure 3 – comparison between MaxHiC and current leading models

a) fraction of regulatory pairs



b) enrichment of genomic pairs

	GM12878						HMEC					
	MaxHiC	GOTHiC	CHiCAGO	Fit-HiC	Fit-HiC2	HicCUP	MaxHiC	GOTHiC	CHiCAGO	Fit-HiC	Fit-HiC2	HiCCUP
chromatin pairs												
H3k27ac-H3k27ac	4.96	1.26	-2.15	1.97	2.11	5.02	4.13	1.13	-4.05	1.53	1.59	4.18
CTCF-CTCF	4.24	1.3	-1.51	1.71	1.86	5.18	4.5	1.4	-1.87	1.74	1.82	4.62
H3k27ac-CTCF	3.95	1.15	-1.77	1.64	1.81	4.24	3.68	1.08	-2.79	1.48	1.55	3.78
DNase-DNase	3.65	1.19	-1.7	1.59	1.78	3.59	2.28	1.09	-1.45	1.22	1.28	2.12
H3k4me1-H3k4me1	3.3	1.21	-2.14	1.51	1.69	3.73	2.79	1.1	-2.95	1.32	1.41	2.91
chromatin state pairs												
Active promoter-Active promoter	7.43	1.33	-1.02	1.47	1.77	7.94	7.5	1.26	-2.34	1.79	1.87	7.68
Strong enhancer-Strong enhancer	7.07	1.32	-1.51	2.18	2.48	7.69	5.11	1.15	-3.66	1.72	1.85	5.24
Active promoter-Strong enhancer	5.9	1.19	-1.36	1.93	1.99	6.48	5.09	1.02	-3.4	1.56	1.72	5.18
Txn_Transition-Txn_Transition	5.57	1.58	-1.32	2.38	2.54	5.45	5.34	1.75	-1.66	2.16	2.28	5.21
Promoter-Promoter	5.27	1.28	-1.07	1.68	1.83	5.83	5.68	1.27	-1.61	1.69	1.82	5.79
Insulator-Insulator	5.16	1.31	1.06	1.48	1.75	5.07	6.61	1.37	-1.25	2.09	2.31	6.33
Promoter-Txn_Transition	4.71	1.26	-1.25	1.91	1.98	5.08	4.33	1.21	-2	1.56	1.69	4.47
Promoter-Insulator	4.18	1.18	-1.04	1.4	1.73	4.76	4.99	1.17	-1.6	1.59	1.73	5.15
Enhancer-Txn_Transition	4.01	1.19	-1.41	1.73	1.86	4.59	3.36	1.05	-2.34	1.31	1.46	3.47
Promoter-Enhancer	3.88	1.17	-1.41	1.62	1.74	4.23	3.49	1.03	-2.07	1.33	1.49	3.54
Txn_Elongation-Txn_Transition	3.84	1.48	-1.07	1.87	1.93	4.19	3	1.46	-1.86	1.57	1.68	3.14
Enhancer-Enhancer	3.77	1.21	-1.55	1.58	1.73	4.05	2.96	1.1	-2.47	1.32	1.45	3.08
Txn_Transition-Weak_Txn	3.71	1.24	-1.19	1.8	1.85	3.99	3.21	1.13	-2.26	1.47	1.59	3.37
Insulator-Txn_Transition	3.65	1.09	-1.19	1.54	1.55	3.91	4	1.05	-1.87	1.6	1.79	4.15
Promoter-Txn_Elongation	3.59	1.28	-1.19	1.71	1.75	3.87	3.07	1.24	-1.89	1.2	1.34	3.21
Promoter-Weak_Txn	3.56	1.21	-1.3	1.6	1.68	3.82	3.21	1.1	-2.02	1.29	1.39	3.34
Enhancer-Insulator	3.44	1.15	-1.27	1.4	1.56	3.85	3.47	1.13	-1.9	1.45	1.57	3.59
Insulator-Weak_Txn	3.24	1.14	-1.17	1.37	1.43	3.65	3.22	1.11	-1.86	1.38	1.53	3.37
Txn_Elongation-Txn_Elongation	3.2	1.6	1.02	1.83	1.86	3.6	2.22	1.62	-1.89	1.46	1.58	3.36
Weak_Txn-Weak_Txn	3.2	1.33	-1.34	1.61	1.68	3.15	2.57	1.18	-2.33	1.32	1.48	2.69
Enhancer-Weak_Txn	3.14	1.21	-1.45	1.48	1.49	3.58	2.47	1.08	-2.31	1.24	1.38	2.59
Insulator-Txn_Elongation	3.11	1.11	-1.11	1.46	1.54	3.55	2.99	1.07	-1.73	1.43	1.58	3.14
Enhancer-Txn_Elongation	3.04	1.2	-1.39	1.47	1.53	3.37	2.19	1.04	-2.23	1.1	1.18	2.32
Txn_Elongation-Weak_Txn	2.97	1.3	-1.22	1.61	1.69	3.31	2.19	1.16	-2.22	1.25	1.38	2.33
Repetitive/CNV-Repetitive/CNV	2.94	1.51	3.4	1.1	1.24	3.3	7.24	2.76	1.38	-1.07	1.13	7.62
Txn_Transition-Repetitive/CNV	2.55	1.05	1.18	1.29	1.38	2.78	3.61	-1.02	-1.09	1.09	1.24	3.74
Promoter-Repetitive/CNV	2.53	1.08	1.24	1.3	1.42	2.65	2.44	-1.01	-1.07	1.06	1.23	2.57
Insulator-Repressed	2.4	1.29	-1.19	1.49	1.58	2.51	2.56	1.33	-1.17	1.67	1.83	2.65
Insulator-Repetitive/CNV	2.27	1.1	1.34	1.18	1.28	2.34	2.29	1.04	-1.04	1.19	1.31	2.37
Promoter-NA	2.17	-1.02	-1.22	1.18	1.29	2.23	2.2	-1.05	-1.49	1.11	1.25	2.41
Repetitive/CNV-Weak_Txn	2.03	1.07	1.09	1.16	1.27	2.12	1.71	-1.09	-1.41	-1.08	1.06	1.83
Enhancer-Repetitive/CNV	2.03	1.02	-1.01	-1.09	1.13	2.13	1.76	-1.13	-1.37	-1.03	1.09	1.89
Insulator-NA	1.95	1.06	-1.06	1.19	1.27	2.05	2.18	1.08	-1.25	1.28	1.38	2.38
Txn_Elongation-Repetitive/CNV	1.93	1.07	1.03	1.19	1.23	1.99	1.77	-1.12	-1.28	-1.08	1.12	1.89
Repressed-Repressed	1.87	1.75	-1.63	1.45	1.53	1.98	2.19	2.09	-1.02	1.64	1.83	2.38
Enhancer-NA	1.7	-1.05	-1.35	1.05	1.12	1.82	1.48	-1.06	-1.61	1.01	1.08	1.68
Enhancer-Repressed	1.68	1.05	-1.62	1.05	1.09	1.72	1.45	-1.02	-1.58	1.08	1.19	1.58
Txn_Transition-Heterochrom/lo	1.61	-1.3	-1.22	-1.09	-1.01	1.68	1.71	-1.38	-1.7	-1.02	1.2	1.84
Promoter-Heterochrom/lo	1.6	-1.12	-1.21	-1.04	1.06	1.66	1.85	-1.12	-1.37	1.03	1.18	1.94
Insulator-Heterochrom/lo	1.59	1.05	-1.04	1.13	1.19	1.63	1.92	1.06	-1.15	1.25	1.34	2.04
Promoter-Repressed	1.56	1.13	-1.37	1.16	1.21	1.62	1.98	1.08	-1.26	1.21	1.37	2.08
Heterochrom/lo-Weak_Txn	1.39	-1.17	-1.3	-1.06	1.07	1.46	1.34	-1.17	-1.59	-1.03	1.05	1.46
Enhancer-Heterochrom/lo	1.36	-1.13	-1.36	-1.08	1.05	1.45	1.34	-1.1	-1.53	-1.02	1.08	1.45
Txn_Elongation-Heterochrom/lo	1.33	-1.25	-1.21	-1.11	1.02	1.42	1.19	-1.33	-1.54	-1.12	1.04	1.28
Repetitive/CNV-Heterochrom/lo	-1.13	-1.03	1.17	-1.05	-1.02	-1.09	-1.15	-1.11	1.2	-1.16	1.05	-1.08
Heterochrom/lo-Heterochrom/lo	-1.93	1.05	1.08	-1.04	-1.09	-1.56	-1.6	1.07	1.19	1	1.03	-1.28
eQTL - targets												
EBV transformed cell	15.74	8.73	1.86	7.97	8.98	15.97	6.4	8.7	1.4	5.8	6.12	6.86
Mammary tissue	2.96	2.71	-1.05	3.25	3.12	9.24	9.5	6.1	1.3	5.7	5.97	9.93
Ovary	9.24	9.76	2.7	8.88	8.63	8.78	5.3	7.8	1.1	4.9	5.21	5.72
Prostate	5.36	7.4	1.7	6.75	6.32	5.63	4.3	7.8	1.4	6.4	6.58	4.61

Fold enrichment: 15.7, 1.5, -1.5, -15.7

c) VISTA known enhancers

	MaxHiC	GOTHiC	CHiCAGO	Fit-HiC	Fit-HiC2	HicCUP
VISTA known brain enhancer	3.98	1.92	-1.07	1.98	2.23	4.05

Figure 4 – distance distribution for different interacting regions in sample Rao GM12878 Lymphoblastoid

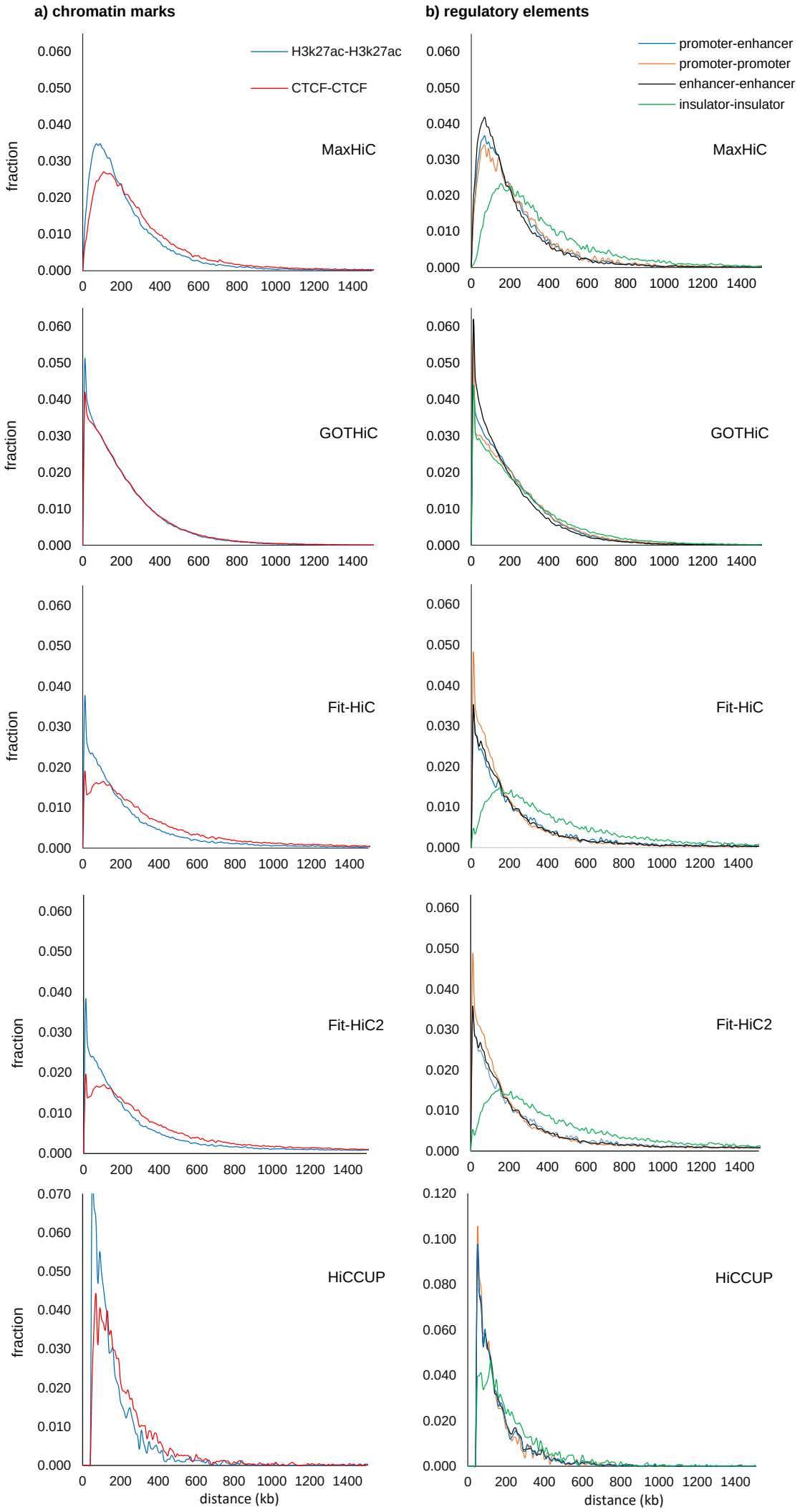
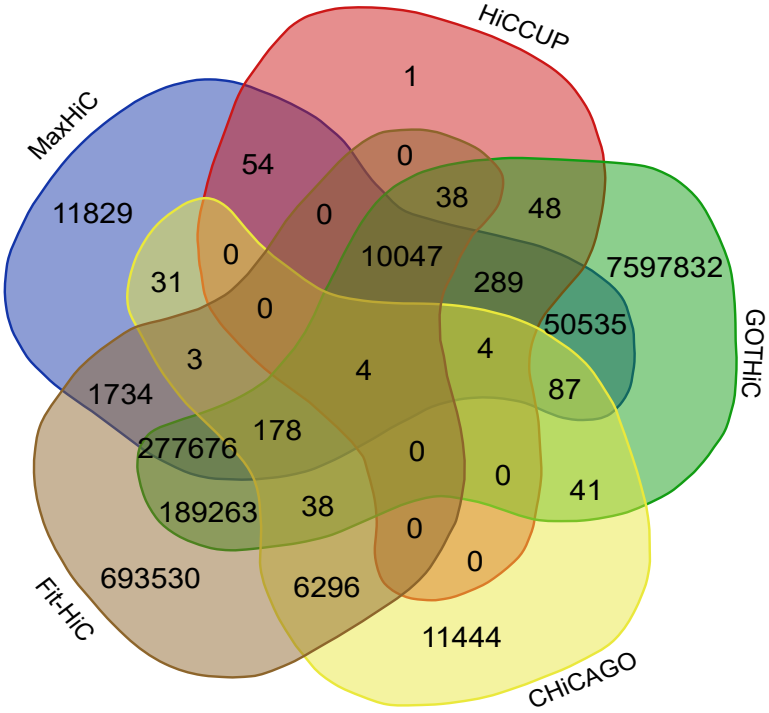


Figure S1 - Intersection between significant interactions identified by the four models

a) GM12878 cell line - 10kb



b) HMEC cell line - 10kb

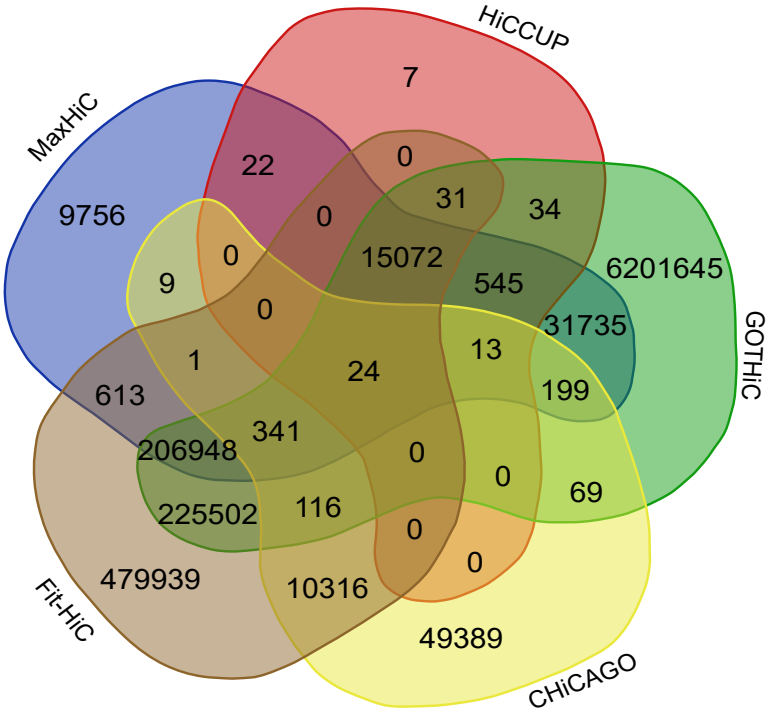


Figure S2 - MaxHiC versus other models on sample Rao Human Mammary Epithelial Cell

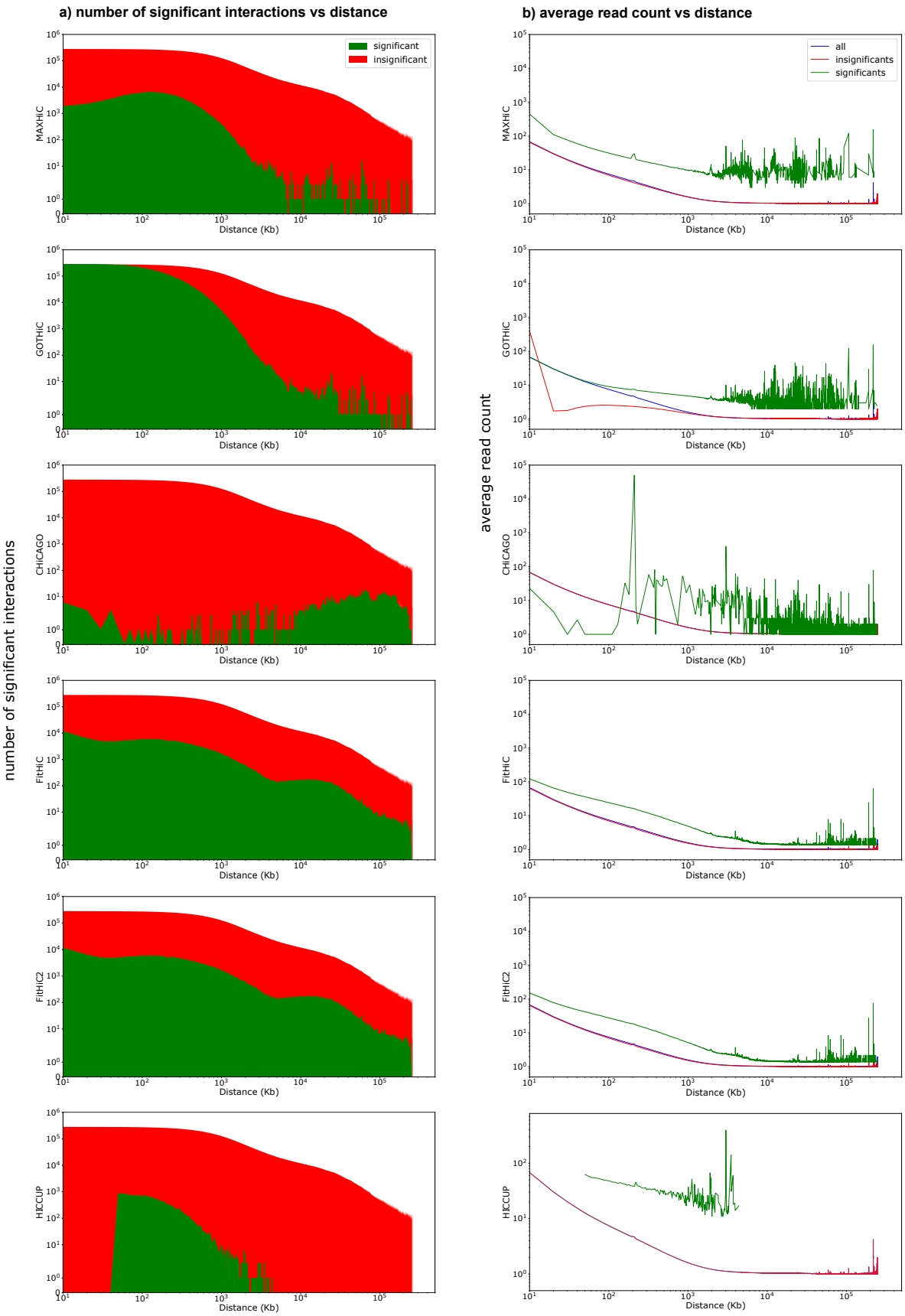
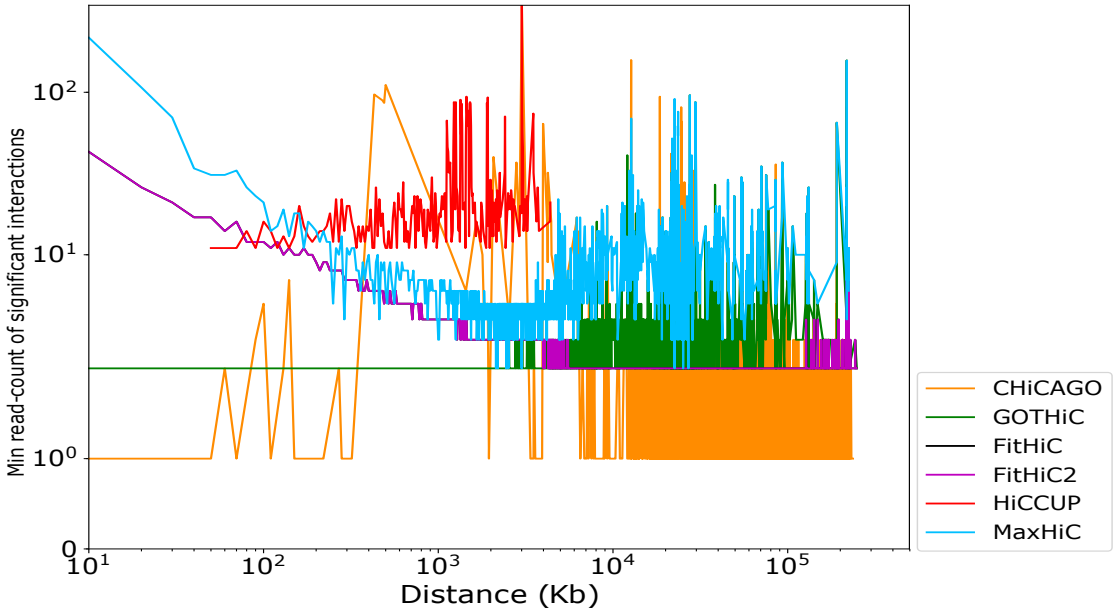


Figure S3 - min read-count and average read count of significant interaction vs distance on sample Rao GM12878 B-lymphoblastoid

a) min read-count vs distance



b) average read-count vs distance

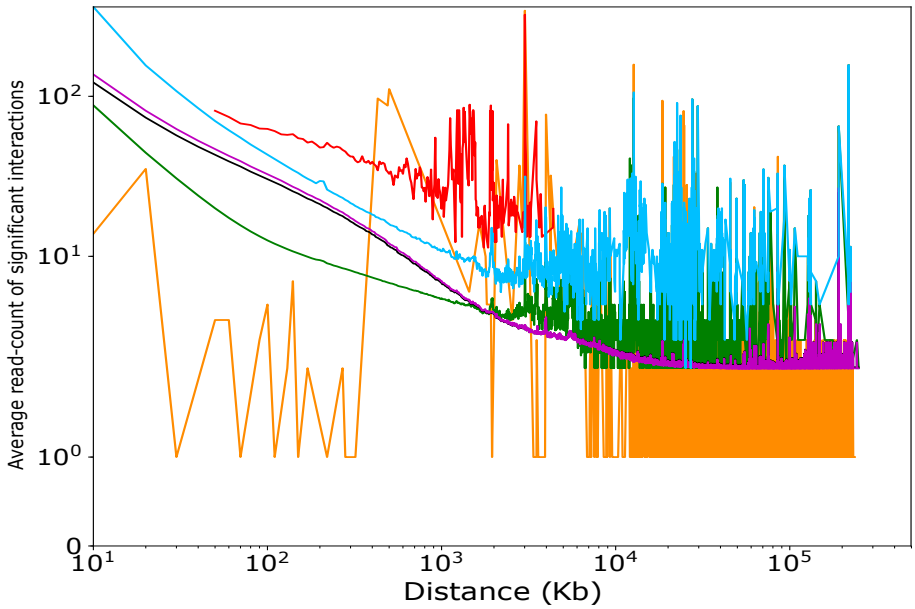


Figure S4 - average read-count of significant interactions vs distance on sample Rao GM12878 Lymphoblastoid

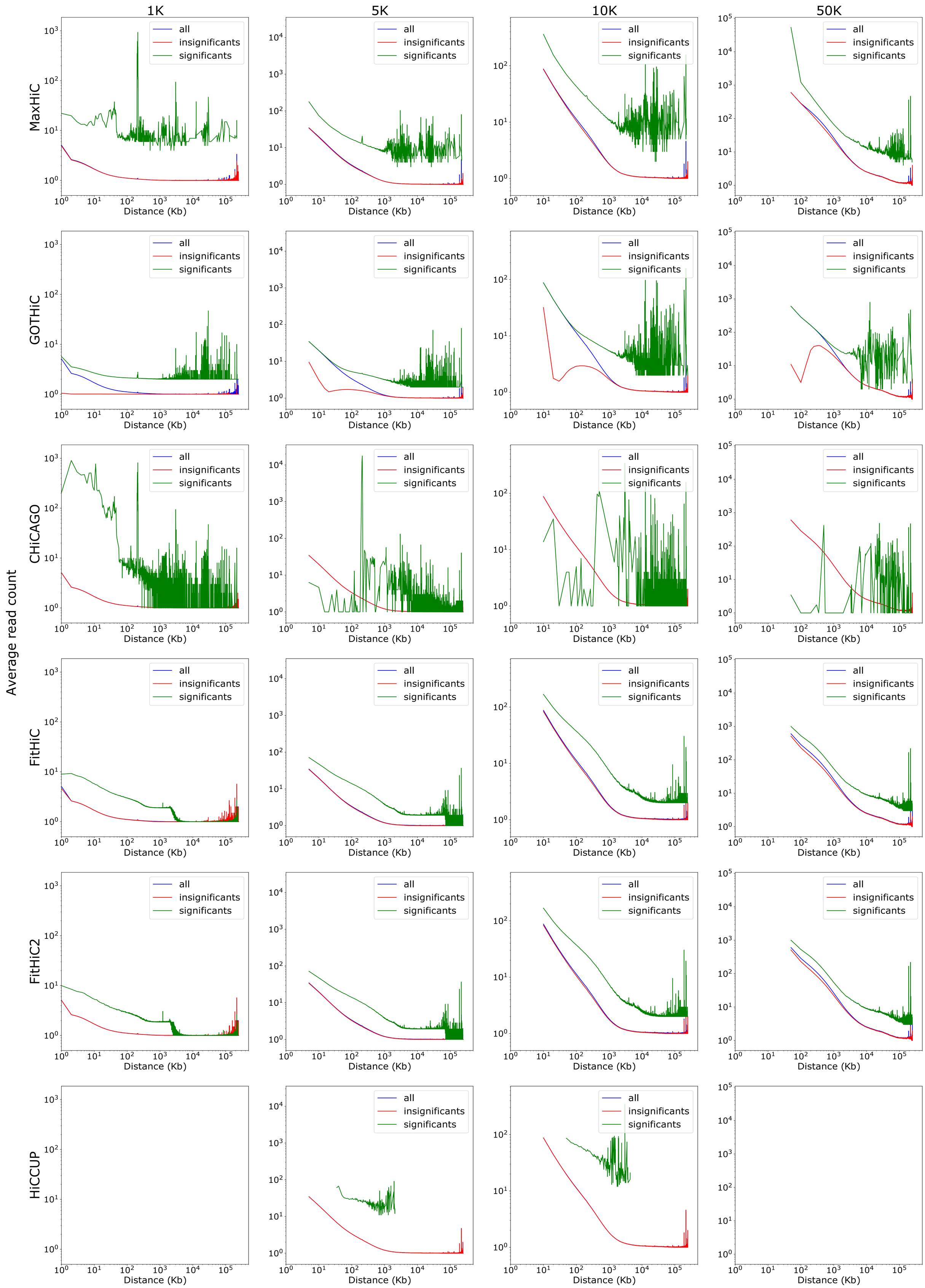


Figure S5 - enrichemnt of TFBS in the Rao GM12878

	MaxHiC	GOTHiC	CHiCAGO	Fit-HiC	Fit-HiC2	HiCCUP
POLR3G	4	1	-2.7	1.5	1.7	4.2
RXRA	3.9	1	-1.4	1.5	1.7	4.1
BCLAF1	3.4	-1	-1.6	1.4	1.4	3.5
WRNIP1	3.2	-1	-1.5	1.3	1.4	3.4
CHD1	3.2	-1	-1.5	1.3	1.4	3.5
RCOR1	3.1	-1.1	-1.8	1.3	1.4	3.3
STAT3	3.1	-1.1	-2	1.3	1.4	3.3
CEBPB	3.1	-1.1	-2.2	1.2	1.3	3.4
ZEB1	3	1	-1.2	1.3	1.4	3.1
STAT5A	3	-1	-1.5	1.2	1.3	3.2
ELK1	3	-1	-1.4	1.3	1.4	3.2
STAT1	2.9	-1	-1.2	1.3	1.4	3.1
ZNF143	2.9	1.1	-1.1	1.3	1.4	3.1
NFYA	2.9	-1	-1.1	1.3	1.4	3.1
MYC	2.9	-1	-1.2	1.3	1.4	3
MAX	2.9	-1	-1.3	1.3	1.4	3.1
MAZ	2.9	-1	-1.2	1.3	1.4	3.1
ETS1	2.9	-1	-1.4	1.3	1.4	3
FOXM1	2.8	-1	-1.7	1.2	1.3	3
TBL1XR1	2.8	-1	-1.7	1.2	1.3	3
ZBTB33	2.8	-1	-1.5	1.3	1.4	3
MTA3	2.8	-1.1	-1.8	1.2	1.3	3
SIN3A	2.8	-1	-1.4	1.3	1.4	3
TBP	2.8	-1.1	-1.7	1.2	1.3	3
RFX5	2.8	-1	-1.3	1.2	1.3	3
YY1	2.8	1	-1.3	1.3	1.4	3
CHD2	2.8	-1	-1.5	1.2	1.3	3
IKZF1	2.8	-1.1	-2.1	1.2	1.3	3
E2F4	2.8	-1	-1.5	1.3	1.4	3
MXI1	2.7	-1	-1.4	1.2	1.3	2.8
USF2	2.7	-1	-1.4	1.2	1.3	2.9
MEF2C	2.7	-1	-1.9	1.2	1.3	2.9
FOS	2.7	-1	-1.1	1.3	1.4	2.9
BHLHE40	2.7	-1	-1.5	1.2	1.3	2.9
PML	2.7	-1	-1.6	1.2	1.3	2.9
TCF3	2.7	-1	-1.3	1.2	1.3	2.8
TAF1	2.7	-1	-1.5	1.2	1.3	2.9
POU2F2	2.7	-1	-1.5	1.2	1.3	2.8
ATF3	2.7	1	1.1	1.3	1.4	2.9
SRF	2.7	-1	-1.4	1.2	1.3	2.9
NFE2	2.7	-1	-1.2	1.2	1.3	2.9
ELF1	2.7	-1	-1.2	1.2	1.3	2.9
NFATC1	2.7	-1	-1.4	1.2	1.3	2.8
SMC3	2.7	1.1	-1.3	1.3	1.4	2.9
SP1	2.7	-1	-1.4	1.2	1.3	2.9
EGR1	2.6	-1	-1.3	1.2	1.3	2.8
USF1	2.6	-1	-1.2	1.2	1.3	2.8
TCF12	2.6	-1	-1.4	1.2	1.3	2.8
PBX3	2.6	-1	-1.2	1.2	1.3	2.8
JUND	2.6	-1.1	-2.3	1.2	1.3	2.8
GABPA	2.6	-1	-1.2	1.2	1.3	2.8
POLR2A	2.6	-1	-1.2	1.2	1.3	2.8
RELA	2.6	-1.1	-1.9	1.1	1.2	2.8
NRF1	2.5	-1	-1.4	1.2	1.1	2.6
IRF4	2.5	-1	-1.8	1.2	1.1	2.7
MEF2A	2.5	-1	-2	1.1	1.2	2.7
BRCA1	2.5	-1	-2	1.3	1.4	2.7
SIX5	2.5	-1	-1.2	1.2	1.3	2.7
BCL3	2.5	-1	-1.3	1.2	1.3	2.7
NR2C2	2.4	1	1.2	1.2	1.3	2.6
PAX5	2.4	-1	-1.4	1.1	1.2	2.7
EP300	2.4	-1.1	-1.9	1.1	1.2	2.6
ATF2	2.4	-1	-2	1.1	1.2	2.6
EZH2	2.3	1	-1.2	1.2	1.3	2.5
RAD21	2.3	1	-1.3	1.2	1.3	2.5
BCL11A	2.3	-1.1	-2	1.1	1.2	2.5
NFIC	2.2	-1.1	-2	1.1	1.2	2.4
NFYB	2.2	-1	-1.3	1.1	1.2	2.3
ZZZ3	2.1	-1	-1.4	1	1.1	2.4
EBF1	2	-1	-1.4	1.1	1.2	2.3
RUNX3	2	-1	-1.6	1.1	1.2	2.3
REST	2	1	-1.2	1.1	1.2	2.2
BATF	1.9	-1	-1.8	1	1	2
CTCF	1.9	1	-1.2	1.1	1.2	2.1
SPI1	1.9	-1	-1.5	1	1.1	2
ZNF274	1.6	1	-1.5	1.7	1.8	1.8

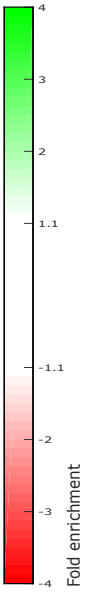


Figure S6 - enrichment for target interacting regions

a) GM12878

		MaxHiC	GOTHiC	CHiCAGO	Fit-HiC	Fit-HiC2	HiCCUP
ENCODE chromatin marks	H3k9ac	1.99	-1.02	1.55	1.10	1.16	1.97
	H3k27ac	1.84	-1.08	1.61	1.08	1.14	1.85
	H3k79me2	1.83	-1.03	1.36	1.12	1.17	1.86
	CTCF	1.72	-1.09	1.65	1.15	1.19	1.68
	H3k4me3	1.67	-1.09	1.31	1.05	1.11	1.63
	DNase	1.42	-1.09	1.26	1.02	1.08	1.45
	H3k4me1	1.35	-1.09	1.16	1.10	1.13	1.31
	H4k20me1	1.24	-1.22	1.13	1.01	1.05	1.25
	H3k27me3	-1.04	-1.27	-1.12	-1.09	-1.06	-1.03
ENCDOE chromatin states	Active_Promoter	2.75	-1.11	2.13	1.18	1.23	2.81
	Strong_Enhancer	2.46	-1.10	2.02	1.21	1.34	2.42
	Weak_Promoter	2.12	-1.09	1.68	1.18	1.29	2.05
	Txn_Transition	2.10	1.05	1.68	1.23	1.31	2.13
	Insulator	1.98	-1.10	1.42	1.20	1.28	1.95
	Weak_Enhancer	1.63	-1.09	1.15	1.12	1.18	1.60
	Txn_Elongation	1.56	-1.01	1.14	1.05	1.12	1.66
	Weak_Txn	1.52	-1.09	1.12	1.15	1.19	1.54
	Poised_Promoter	1.46	1.06	1.08	1.06	1.12	1.39
	Repetitive/CNV	1.06	-1.03	-1.10	1.00	1.02	1.10
	Repressed	-1.01	-1.08	-1.14	1.04	1.03	-1.02
	Heterochrom/lo	-1.28	1.04	-1.11	1.03	1.01	-1.32
FANTOM5	F5 promoter (GM)	2.19	-1.06	1.43	1.17	1.23	2.23
	F5 enhancer (GM)	2.07	1.03	1.54	1.16	1.25	2.01