Bioinformatics, 36(12), 2020, 3625–3631
doi: 10.1093/bioinformatics/btaa202
Advance Access Publication Date: 24 March 2020
Original Paper

OXFORD

Genome analysis

HiChIP-Peaks: a HiChIP peak calling algorithm

Chenfu Shi [1,*], **Magnus Rattray**[2,3] **and Gisela Orozco** [1,3]

[1]Division of Musculoskeletal and Dermatological Sciences, School of Biological Sciences, Centre for Genetics and Genomics Versus Arthritis, [2]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester M13 9PT, UK and [3]NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9PT, UK

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on June 25, 2019; revised on February 18, 2020; editorial decision on March 15, 2020; accepted on March 19, 2020

Abstract

Motivation: HiChIP is a powerful tool to interrogate 3D chromatin organization. Current tools to analyse chromatin looping mechanisms using HiChIP data require the identification of loop anchors to work properly. However, current approaches to discover these anchors from HiChIP data are not satisfactory, having either a very high false discovery rate or strong dependence on sequencing depth. Moreover, these tools do not allow quantitative comparison of peaks across different samples, failing to fully exploit the information available from HiChIP datasets.

Results: We develop a new tool based on a representation of HiChIP data centred on the re-ligation sites to identify peaks from HiChIP datasets, which can subsequently be used in other tools for loop discovery. This increases the reliability of these tools and improves recall rate as sequencing depth is reduced. We also provide a method to count reads mapping to peaks across samples, which can be used for differential peak analysis using HiChIP data.

Availability and implementation: HiChIP-Peaks is freely available at https://github.com/ChenfuShi/HiChIP_peaks.

Contact: chenfu.shi@postgrad.manchester.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The 3D conformation of the chromatin is fundamental in the regulation of gene expression; regulatory elements such as enhancers have been shown to act by physically interacting with their target promoters (Bulger and Groudine, 2011; Nolis *et al.*, 2009; Shlyueva *et al.*, 2014; Yao *et al.*, 2015). These regulatory elements are highly regulated and context specific (Alasoo *et al.*, 2018; Kundaje *et al.*, 2015; Simeonov *et al.*, 2017). However, the requirement of large number of cells (tens of millions) to obtain chromatin interactions maps at sufficient resolution and the high cost associated with widely used chromatin conformation techniques, such as Hi-C, have hindered the study of chromatin interactions in primary and patient-derived cells (Rao *et al.*, 2014).

HiChIP is a recently developed technique to analyse chromatin conformation which consists of an *in situ* Hi-C library preparation followed by a chromatin immunoprecipitation (ChIP) step, usually targeting the histone modification H3K27ac or cohesin. It has many advantages compared with traditional methods, such as Hi-C, Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) and Capture Hi-C such as lower cost, higher sensitivity, lower input requirements and reduced sequencing required (Mumbach *et al.*, 2016, 2017). Unfortunately, few tools exist to specifically analyse HiChIP data, with most publications relying on tools originally developed for Hi-C. HiChIP provides a new set of computational challenges because it combines biases introduced by two independent techniques: ChIP and *in situ* Hi-C library preparation. This phenomenon is particularly evidenced by libraries enriched for H3K27ac because this histone modification has a significantly more specific enrichment compared with cohesin.

It is theoretically possible to extract two types of information from HiChIP data: the position of enriched regions for the ChIP and the long-range interactions involving these regions. The enriched regions, also called anchors or peaks, are usually identified prior to the identification of long-range interactions. Previous tools used either MACS2 on close range read pairs (FitHiChIP; Bhattacharyya *et al.*, 2019) or an adaptation of it (Hichipper; Lareau and Aryee, 2018). Vanilla MACS2 (Zhang *et al.*, 2008) and its implementation in FitHiChIP has been shown to be strongly biased due to HiChIP-specific biases, primarily the biotin pulldown (Lareau and Aryee, 2018; Supplementary Fig. S1). Hichipper tries to solve this problem by modelling a corrected background as a function of proximity to restriction sites and using that background for MACS2 peak calling. This results in many small peaks which then need to be merged to match the restriction fragments, which causes them to lose statistical metrics, such as *P*-values or scores rendering comparisons between samples infeasible. Our tests show also that using all the reads results in poor specificity while using only self-circle and dangling end reads results in very few reads being retained and correspondingly reduced sensitivity.

For this reason, many recent publications used independent ChIP-Seq as input to define anchors (Pelikan *et al.*, 2018). However,

that too can be a problem because the peaks definition can strongly influence the expected signal from a region and can be extremely variable if not done from exactly the same sample.

Here, we propose a method to extract the location of ChIP-Seq peaks from HiChIP data that improves significantly on previous attempts. We analysed the HiChIP protocol and library preparation and developed an algorithm and a data representation that takes in consideration how the libraries are generated and Hi-C and HiChIP-specific biases, such as the biotin pulldown (Fig. 1). We opted for a re-ligation (restriction) site centred representation and we use short range interactions to identify the signal from the chromatin immune-precipitation. We then model the background signal as a negative binomial distribution to model over-dispersion and identify regions of enriched signal. We show that our approach is highly reproducible when compared with reference ChIP-Seq datasets and we show how this can improve the performance of downstream tools to call chromatin loops from HiChIP data. We also provide a method to count reads mapping to peaks across samples, which can be used to analyse differentially bound regions from HiChIP data and show that this can identify biologically significant differences.

The software is available as a Python 3 package on GitHub and PyPi along with code to reproduce the results presented here.

2 Materials and methods

2.1 A novel representation for HiChIP data

Hi-C maps have typically been analysed using a fixed size bin matrix format. This can introduce significant biases because the expected number of reads depends on the number of restriction sites included within each bin. Moreover, reads are not uniformly distributed in the genome but are strongly biased around restriction sites because of the on-bead library preparation (Fig. 2 and Supplementary Fig. S1). This causes sparsity and non-uniformity in the data, which can bias methods based on genomic position alone. An alternative approach is to analyse maps at a restriction fragment resolution. Analysing the raw data from Mumbach *et al.* (2017), we find that a significant number of reads contain uncut restriction sites (Supplementary Table S1). This suggests that the cutting frequency is low and that especially for libraries prepared using frequently cutting enzymes (four cutters), such as MboI, the read assignment to the restriction fragment can be misleading. Moreover, traditional pair classifications, such as dangling end and self-circle, can be misleading since they can be wrongly classified as valid pairs (Fig. 1B) with significant biases due to fragment size (Fig. 1C). Importantly, the low frequency of cutting implies that the detectable signal is directly correlated to the number of restriction sites and only indirectly to the effective genomic size.

Approaching the problem in a novel way, we develop a data structure that focuses on the re-ligation site. This is the location from which reads are generated during the library preparation (Fig. 1A) and this data structure maximizes our detection power while at the same time minimizing biases introduced by the Hi-C library preparation. Reads are assigned based on the direction of the read to the nearest restriction site to which they point. This method significantly reduces the previously mentioned biases and maximizes information at the highest meaningful resolution. Miss-assignment of the reads due to small fragment size would be automatically corrected in a logical way (Fig. 1C). We implement this data structure as a sparse matrix in which the diagonal contains all the re-ligation pairs and the diagonal +1 contains pairs traditionally classified as

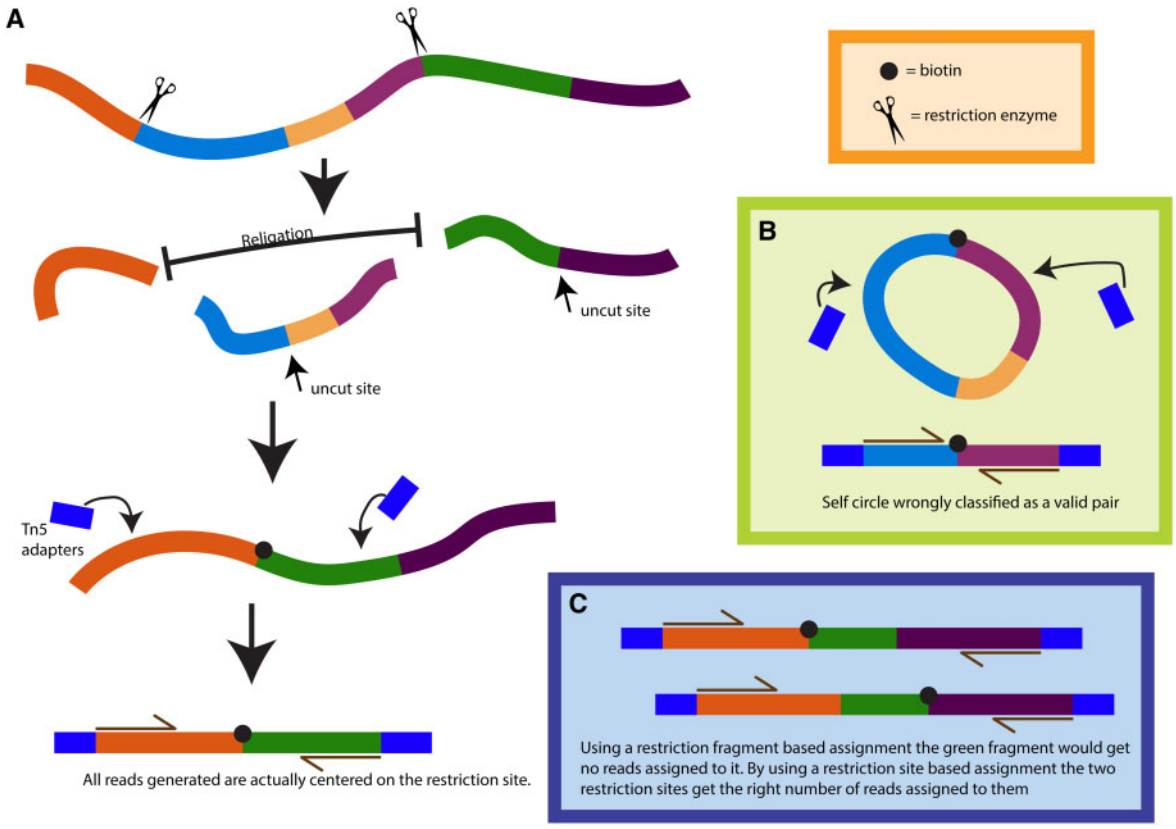

Fig. 1. Justification for re-ligation site based data structure. (**A**) The Hi-C protocol creates reads that are centred on the re-ligation site. Mapping reads at a higher resolution is not biologically relevant and only creates sparsity. (**B**) Example of how traditional self-circle classifications are not reliable with libraries generated using frequently cutting enzymes. (**C**) Example of how a data representation based on the restriction fragment can heavily bias the read counts depending on the size of the fragment while basing the data representation on the re-ligation site can reduce the bias by compensating each read assigned incorrectly with another one

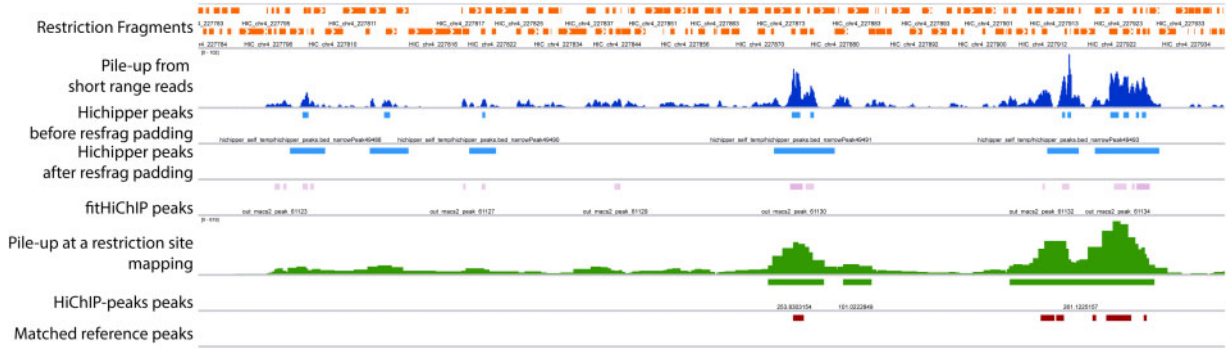

Fig. 2. Visualization of the different methods used for peak calling from HiChIP data. Our novel restriction site centred data structure allows us to correct for the bias introduced by the Hi-C library preparation. This is contrast with in contrast with methods that make use of MACS2 applied to the pile-up from short range (in this case from Hichipper). Our method inherently simplifies the peak calling approach and allows us to have higher sensitivity and lower false positive rate compared with the methods implemented in Hichipper and FitHiChIP. Data shown are from GM12878 cells

self-circle and dangling ends. Starting from this data structure, we develop our peak calling algorithm.

2.2 Peak calling

To limit the bias in the definition of self-circles and dangling ends and increase sensitivity, we decided to also include reads that map to close range interactions. By default, we include interactions that are within two sites of the re-ligation pairs. We perform a moving integration as a smoothing function over three restriction sites to reduce noise. This allows us to use more significant reads in the successive steps and to regulate these settings depending on the input data. For example, libraries generated using the commercially available Arima-HiC kit, which uses two restriction enzymes, generate next to no reported self-circle and dangling ends reads (Supplementary Fig. S2) but our method can be easily tuned by changing the previous parameters.

Visual inspection of the distribution of the background signal suggests that it closely matches a negative binomial distribution (Supplementary Fig. S3), similar to what has been found with ChIP-seq data (Diaz *et al.*, 2012). In order to model the distribution of the background, it is important to note that a large majority of reads will locate in peaks (up to 80%) and inclusion of these reads would highly bias parameter estimation. We, therefore, first remove the most significant peaks using a Poisson-based model (similarly to MACS2; Zhang *et al.*, 2008) with a very stringent setting (P-value $<1 \times 10^{-8}$) and a genomic background. We then estimate the negative binomial mean and over-dispersion parameters using the residual background reads.

Most fragment-size bias is removed thanks to our novel data structure but we still find a small amount of bias (Supplementary Fig. S4). We correct for this by using a LOWESS fit with the residual background and then correct the expected background level within each region using the learned regression function.

$$
\begin{aligned}
&bg \sim NB(\lambda, \sigma)\\
&\lambda = \lambda_g + l(s)\\
&\lambda_g = genomic\ \ background\ \ signal\\
&\sigma = variance\ \ of\ \ background\ \ signal\\
&s = size\ \ of\ \ fragments\\
&l(s) = size\ \ function\ \ estimated\ \ with\ \ LOWESS\ \ fit
\end{aligned}
$$

The fitted negative binomial model represents the data well after fragment-size bias correction with a P-value distribution that is close to uniform away from zero, as expected, with a spike close to zero corresponding to data inferred to be within peaks (Supplementary Fig. S5). We then use the Benjamini–Hochberg false discovery rate (FDR) correction and combined contiguously significant re-ligation sites into peaks.

2.3 Differential peak analysis

We take advantage of the data structure and the expected background model to develop an addition to our main software. To call differentially bound regions, we first combine the peaks from all the samples to create a list of consensus peaks. Similarly to DiffBind (Ross-Innes *et al.*, 2012; Stark and Brown, 2011), we then count how many reads were assigned to those regions from each sample, correct the values by removing the expected background based on the negative binomial model and fragment size and then analyse the results using DESeq2 (Love *et al.*, 2014) to normalize the read counts across samples and perform differential expression analysis.

$$
\begin{aligned}
&signal = x - \lambda\\
&x = counts\ \ mapped\ \ to\ \ peak
\end{aligned}
$$

The model assumptions of DESeq2 are satisfied as evidenced by the P-value distributions shown in Supplementary Figure S20.

Unsupervised hierarchical clustering was done using Euclidean distance on the signal from all the peaks with rlog normalization. Motif enrichment analysis was done using differentially bound peaks between Tregs and naïve T cells and Th17 and naïve T cells. These regions were submitted to HOMER v 4.8.3 (Heinz *et al.*, 2010) with the findMotifsGenome.pl command and '–size given' parameter.

2.4 Data pre-processing

HiChIP data from Mumbach *et al.* (2017) were downloaded from SRA for naïve T cells, Th17, Tregs and K562 (SRP no. SRP112520). Reads were filtered and the adapters were removed using fastp v0.19.4 (Chen *et al.*, 2018). The reads were then mapped to the GRCh38 genome with HiC-Pro v2.11.0 (Servant *et al.*, 2015), using default settings. Replicates were merged together as described in the Section 3.2.

2.5 Hichipper peak calling

We called anchors using the Hichipper v 0.7.5 pipeline (Lareau and Aryee, 2018) on the HiC-Pro results with default settings making sure to include the modified background correction and restriction fragment aware padding. We called peaks with the setting EACH, SELF (for self-circle and dangling ends only) or EACH, ALL (for all reads).

2.6 FitHiChIP (MACS2 short range) peak calling

We used the supplied tool with FitHiChIP (Bhattacharyya *et al.*, 2019) to call peaks from HiC-Pro results with default settings. This tool uses all reads from dangling ends, re-ligation and self-circle pairs and also all reads within 1 kb from the valid pairs and supplies all the reads to MACS2 2.1.1 (Zhang *et al.*, 2008) for peak calling.

2.7 Peak calling comparison

We downloaded reference H3K27ac tracks for GM12878 cells from the Encode website (accession no. ENCSR000AKC), replicated peak set (accession no. ENCFF367KIF).

For the naïve T cells, we used the processed peaks from the roadmap project (Sample E038; Kundaje *et al.*, 2015). We used the tool LiftOver to convert the genomic coordinates from hg19 to hg38.

All comparisons were done using bedtools v2.27.1 (Quinlan and Hall, 2010) annotate function and then analysed in python. No extension of the peaks was done. Peaks on X and Y chromosomes were excluded from the comparison. For HiChIP-Peaks plots are presented as lines that result in cumulative sum of the results with the peaks sorted by *P*-value.

2.8 Subsampling analysis

We created subsampled datasets from the GM12878 HiChIP data (Mumbach *et al.*, 2017) by subsampling the raw reads creating datasets with 500, 250, 125 and 62.5 million reads. For the naïve T cells dataset, we used the combined data, the two biological replicates with the two technical combined or the four technical replicates as individual samples.

We compared how many of the peaks called using the full dataset could be recovered from the subsampled datasets for Hichipper and HiChIP-Peaks. We calculated precision and recall rates using bedtools v2.27 annotate.

For loop calling, we used Hichipper with default settings. We either used the default peak calling algorithm from Hichipper or we supplied the peaks called using HiChIP-Peaks from the respective dataset. In the former case, we used the skip-resfrag-pad setting to avoid Hichipper expanding the peaks. Overlaps were calculated using bedtools pair-to-pair.

To compare the loops called, we first filtered the loops by FDR < 0.10 as reported by Mango (Phanstiel *et al.*, 2015). We then checked if the loops called in the full dataset could be found in the subsampled datasets and calculated the recall rate. A loop was considered recalled if both ends overlapped both ends of a loop in the subsampled dataset.

2.9 Loops comparison with reference datasets

We compared the results from the loops called from Hichipper using default settings or using the peaks generated from HiChIP-Peaks with a matched reference. We sourced promoter capture Hi-C data for the GM12878 from Javierre *et al.* (2016). For the naïve T cells, we used data generated from Mifsud *et al.* (2015) but we downloaded the CHiCAGO loop calls from Bhattacharyya *et al.* (2019). We also downloaded H3K27ac ChIA-PET data from Heidari *et al.* (2014).

We filtered the loops reported by Hichipper by FDR <0.01 and overlaps were then calculated using bedtools pair-to-pair. The results were then analysed and processed in Python.

Loops were also called with FitHiChIP (Bhattacharyya *et al.*, 2019) using the following settings: coverage normalization, stringent background with merging enabled and 5 kb bin size, with either HiChIP-Peaks peaks or peaks generated using the included tool as described in Section 3.

3 Results

3.1 HiChIP-Peaks improves reference peak recovery

To evaluate the performance of our peak calling algorithm, we chose two of the cell lines reported by Mumbach *et al.* (2017) for which a reference ChIP-seq track was available either from ENCODE or Roadmap project. We combined all the reads from different replicates from naïve T cells and from GM12878 cells, respectively. Using different metrics, we show that our method is superior to previous attempts at calling peaks and allows for scoring of the peaks identified.

Specifically, our method is able to recover more peaks from the reference with significantly lower FDR (Fig. 3A and B) and calling fewer peaks (Supplementary Fig. S6) than Hichipper or FitHiChIP (note real FDR cannot be zero because the reference ChIP-Seq does not come from the same sample as the HiChIP). In particular, we note that both Hichipper with all reads and FitHiChIP present significant FDR problems with >70% of peaks called not observed in the reference. The reason for these differences can be explained by looking at the results of the various methods (Fig. 2 and Supplementary Fig. S1B and C). We see how the bias introduced by the library preparation can bias other methods and how our method significantly reduces this effect. In particular, we notice how other methods based on MACS2 tend to call many small peaks around restriction sites and have also false positive problems created by the non-uniform background that the library preparation method introduces. Although our method identifies on average larger peaks than Hichipper, our method is still superior when comparing the total amount of genome covered with the recalled peaks at an FDR of 0.01 or 0.001 (Supplementary Fig. S7). FitHiChIP performs well on this metric but the comparison cannot be considered comparable because the peaks called from FitHiChIP are small but dispersed along the genome (Fig. 2 and Supplementary Fig. S1B and C).

Moreover, we note that with Hichipper it is not possible to change the sensitivity: changing the q-value setting does not produce any difference in number of peaks called or genome covered.

3.2 HiChIP-Peaks is more stable than Hichipper when read depth is reduced

Using the best settings for Hichipper (SELF reads) we compared the stability of the results when the number of reads in the dataset is reduced. We analysed the individual technical replicates of the naïve T cells that contain about 100 million reads per sample. We show that our method is consistently able to maintain accuracy and

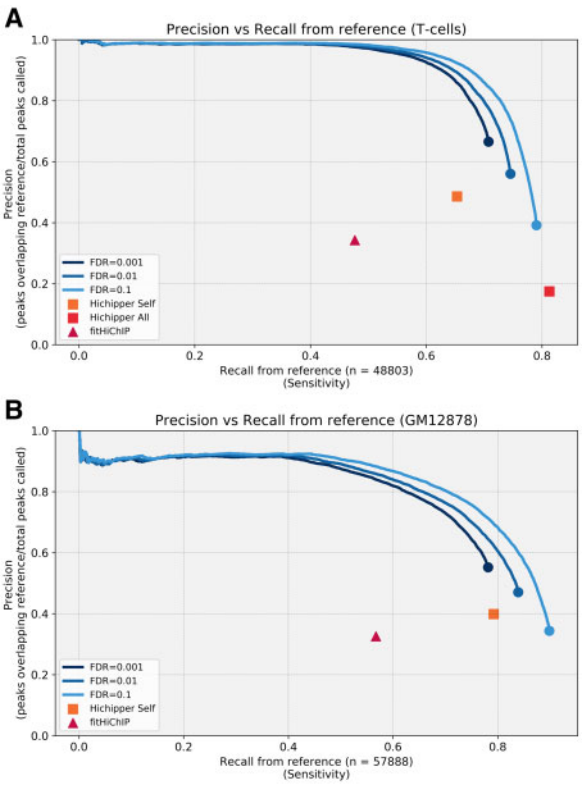

Fig. 3. Precision versus recall from reference in (**A**) T cells and (**B**) GM12878 cells. For our software, we sort peaks in ascending *P*-value order and show the true positive rate as the number of peaks recovered increases. We provide three different FDR settings as the FDR setting changes the size of the peaks themselves and the lines do not overlap perfectly. We show results for Hichipper and FitHiChIP (default settings) for comparison. Hichipper in (ALL) mode fails to run with the GM12878 dataset

sensitivity, while Hichipper suffers greatly when the number of reads is less than optimal (Fig. 4). We then tested how a reduced dataset would affect peak calling using the peaks identified from the full dataset as the reference. We used progressively subsampled datasets for the GM12878 dataset and we tested technical replicates and biological replicates from the naïve T-cell dataset. HiChIP-Peaks demonstrates a much higher recall rate at a higher precision compared with Hichipper in both of these cell types when the read count is reduced (Fig. 5A, C and Supplementary Fig. S8).

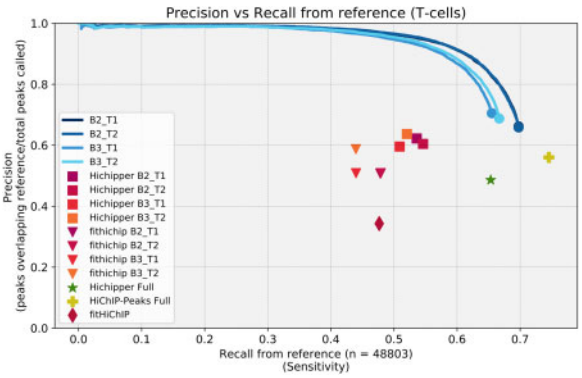

Fig. 4. Effect of reduced read depth on peak calling performance. Precision versus recall from reference (naïve T cells dataset). We show that our software maintains high consistency while Hichipper's sensitivity goes down rapidly when read count goes down. Our software is set at a FDR of 0.01

3.3 Increasing the stability of loop calling in Hichipper

We decided to test whether our improved peak calling would affect loop calling results from Hichipper. Because the biggest differences in peak calling were found when lowering the number of reads, we decided to test Hichipper's stability using progressively subsampled datasets from the GM12878 dataset and by combining fewer technical replicates from the naïve T cells datasets. We note that, using peaks from our algorithm, Hichipper is able to recall loops identified using the full dataset at a higher recall rate at the same level of precision compared with using its own peaks (Fig. 5B, D and Supplementary Figs S9 and S10). This shows that stability and accuracy of the peaks called significantly impacts the loop calling results and our algorithm can greatly improve the stability of the results, especially when number of reads available is limited.

Additionally, because of the higher accuracy of the peaks, we note that the anchors of the loops identified using our peaks overlap more the reference ChIP-seq. The percentage of loops overlapping a peak in at least 1 anchor goes from 84.2% to 92.3%, and the loops overlapping a peak at both anchors goes from 40.9% to 58.6% in GM12878 cells. In T cells, the values go from 96.8% to 99.2% and 62.3% to 76.8%.

Next, we wanted to test how the loops identified from these methods overlapped with loops identified with other techniques. To do this, we sourced matched promoter capture Hi-C and ChIA-PET data from publicly available sources. Supplying our peaks to

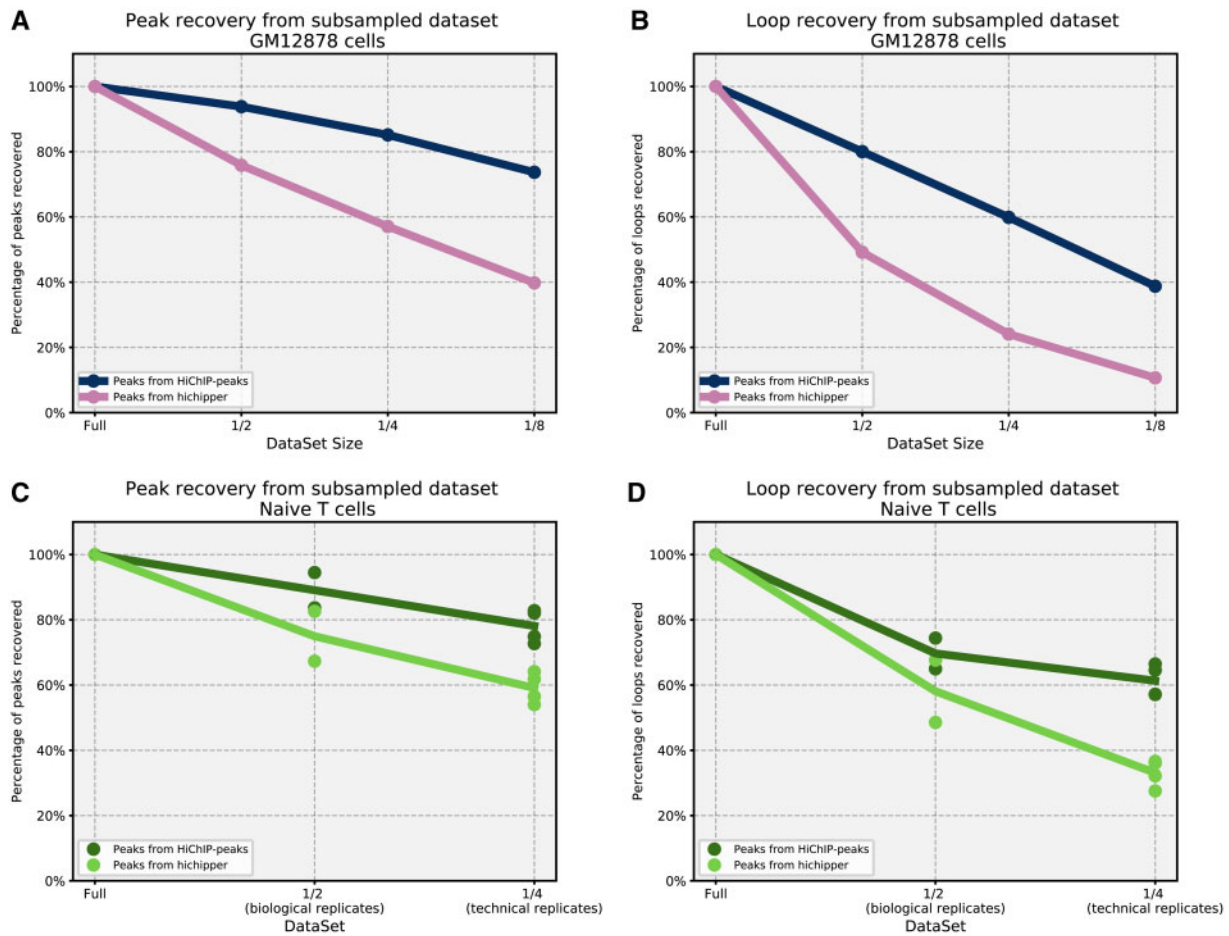

Fig. 5. Effect of subsampled datasets on peak and loop calling. (**A**) Recall rate of peaks from full dataset using subsampled datasets from GM12878 dataset. (**B**) Recall rate of loops called (chr22) from full dataset using GM12878 datasets. (**C**) Same as (A) but with naïve T cells dataset. We merged the technical replicates for each biological replicates for half and we used each technical replicate individually for one-fourth. (**D**) Same as (B) but with naïve T cells datasets

Hichipper seems to allow the recovery of a higher fraction of reference loops for the same number of loops called (Supplementary Figs S11–S13). Because the size of the loop anchors could form a bias in the datasets, we wanted to test how much of the effect was due to covering more base pairs of the genome. To do this, we compared the recall rate with the genomic coverage (in base pairs) of the loop anchors used. The difference between the two techniques is greatly reduced, but we still see a small improvement using peaks generated by our method (Supplementary Figs S14 and S15).

We noted however in our comparisons that the overlaps between the different techniques are very low, ranging from 5% to 15% recall and precision rates in all tested conditions. This can be partly explained by looking at the overlaps of the loops with reference H3K27ac ChIP-seq peaks which was very poor for all datasets including the K562 ChIA-PET reference. For promoter capture Hi-C, only 18.6% (GM12878) and 23.9% (CD4 naïve T cells) of the loops overlapped a reference peak at both anchors, which is particularly low considering that the captured regions are promoters which are highly overlapping with this histone modification. For the K562 ChIA-PET dataset, the overlaps were even lower with only 7.7% of the loops overlapping a reference peak at both loop anchors. This seems to indicate that these techniques are identifying different classes of loops than the HiChIP methods considered here.

We also tested the effects of the peak calling in FitHiChIP (Bhattacharyya *et al.*, 2019), but the loop calling was not significantly affected compared with the effects seen in Hichipper (Supplementary Figs S16–S19). About 80% of the loops are replicated between the two settings, and the number of loops overlapping reference loops is also unaffected. This is likely due to how FitHiChIP bins the data and in the way it removes ChIP bias before calling loops.

3.4 Novel data representation allows accurate differential peak calling from HiChIP data alone

Using the novel data representation, we provide an interface to analyse differentially bound regions in HiChIP datasets, fully exploiting the information contained in them.

We carried out a proof-of-concept study by analysing the four technical replicates of the naïve T cells individually. Our results show that the sensitivity and reproducibility of our software is sufficiently good that we can easily differentiate between technical and biological replicates of the same cell type (Fig. 6A). We find almost 3000 peaks (more than 10% of all peaks) differentially bound (FDR < 0.10, log2FoldChange > 0.5) between biological replicates of the same cell type further affirming the importance of peak calling on individual HiChIP datasets instead of using combined or external ChIP-seq datasets.

We then analysed data from the two other T-cell types, Th17 and Tregs. We merged the technical replicates into biological replicates. Although the read depth is very different between the different samples (37–60 m reads used in the peak calling) our software performs remarkably consistently, producing similar number of peaks with high overlap. In Tregs, e.g. biological replicate 3 (R3) contains 1.54 times the number of reads in biological replicate 2 (R2). Our software identified 25 771 peaks in R2 versus 27 105 peaks in R3. Moreover, 88.4% of the peaks in R2 were also called in R3, and 81.2% of the peaks in R3 were also called in R2. Biological differences vastly outweigh technical differences and samples cluster by cell type (Fig. 6B and C). We identify thousands of peaks that are significantly differentially bound between the different cell types. As expected the differences between Th17 and Tregs are smaller than between Th17 and naïve T cells. To test whether the differentially bound peaks have biological significance, we ran motif enrichment analysis with HOMER on the peaks from Th17 versus naïve T cells and Tregs versus naïve T cells contrasts. The results clearly indicate enrichment in binding sites for transcription factors involved in the interferon pathway, ETS-RUNX and others (Supplementary Table S2), consistent with models of T-cell activation (Christie and Zhu, 2014) and confirming the accuracy of our differential peak calling method.

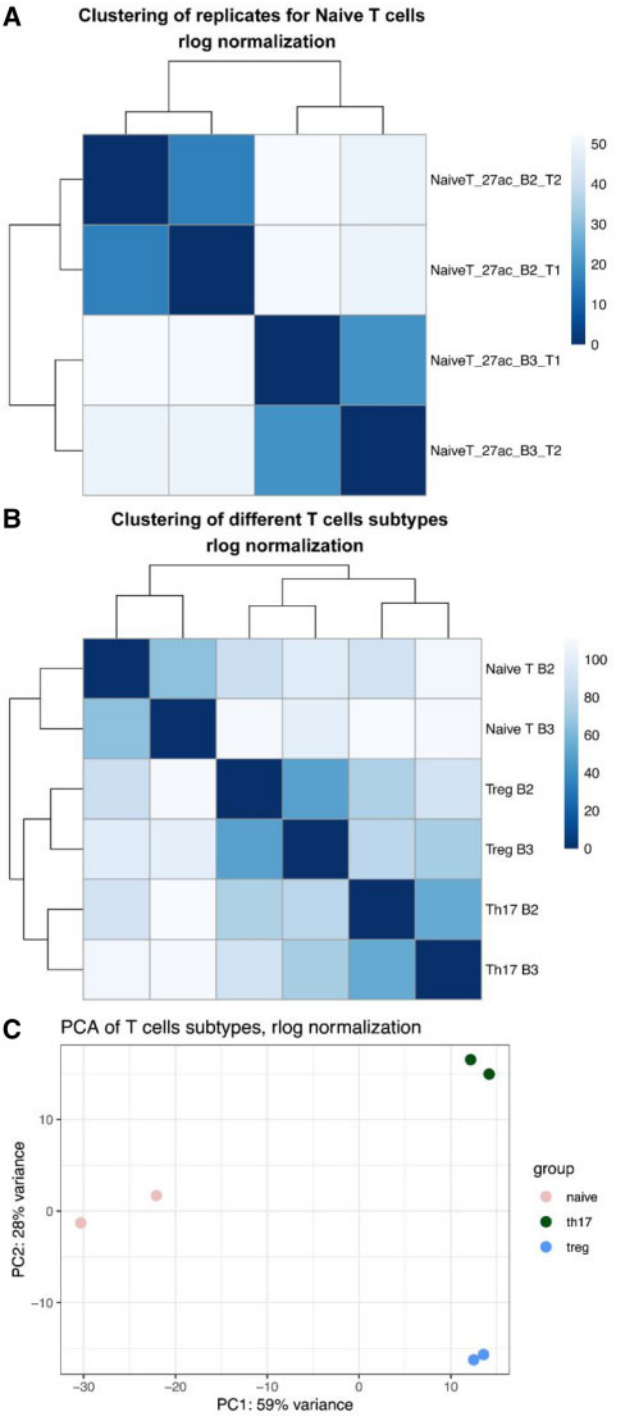

Fig. 6. Differential peak analysis. (**A**) Unsupervised hierarchical clustering using Euclidean distance of scores from the four technical replicates of naïve T cells. (**B**) Unsupervised hierarchical clustering of the three different T-cell related cell lines. (C) Principal component analysis (PCA) of the three different cell lines

4 Discussion

HiChIP is quickly gaining importance, especially in studies involving primary cells of various tissue types thanks to the lower input and sequencing requirements. Previously ChIP-seq tracks were used to identify peak regions as the quality of peak calling from HiChIP was deemed insufficient. This either added a significant cost and sample requirement to the experiment design or often researchers relied on data not generated from the same sample.

We show that our software can reliably and efficiently identify enriched regions using only HiChIP datasets, even when read depth is relatively low, and that using HiChIP-Peaks significantly improves the reliability of Hichipper's loop calling. This shows also how good peak calling is of fundamental importance for Hichipper's functionality. Our results also demonstrate that accurate peak calling from each sample is important because each biological replicate can have different peaks, which can affect the identified loops, especially when studying more transient and regulated regions.

As the popularity of chromatin conformation methods increase, commercial kits, such as the Arima HiChIP kit, are starting to be developed. The kit is highly efficient thanks to its dual restriction enzyme protocol, but this results in the absence of reported dangling ends and self-circles (Supplementary Fig. S2). This impacts the performance of Hichipper using the SELF setting, which, according to our analysis, is the best of the currently available methods. Therefore, our method, HiChIP-Peaks, has the potential to be the only method of choice when using commercially available kits such as the Arima HiChIP to generate HiChIP libraries.

Our results show that our alternative data structure for representing Hi-C reads limits biases due to how reads are generated in this protocol and maximizes resolution within the constraints of the technology. This data structure can also be used for other kinds of analysis with simple generalizations.

Acknowledgements

The authors would like to acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester.

Funding

This work was supported by the Wellcome Trust [award references 207491/Z/17/Z and 215207/Z/19/Z], the Versus Arthritis [award reference 21754], the National Institute for Health Research Manchester Biomedical Research Centre and the Medical Research Council [award reference MR/N00017X/1].

Conflict of Interest: none declared.

References

Alasoo,K., *et al.*; HIPSCI Consortium. (2018) Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.*, **50**, 424–431.

Bhattacharyya,S. *et al.* (2019) Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.*, **10**, 4221.

Bulger,M. and Groudine,M. (2011) Functional and mechanistic diversity of distal transcription enhancers. *Cell*, **144**, 327–339.

Chen,S. *et al.* (2018) Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Christie,D. and Zhu,J. (2014) Transcriptional regulatory networks for CD4 T cell differentiation. *Curr. Top. Microbiol. Immunol.*, **381**, 125–172.

Diaz,A. *et al.* (2012) Normalization, bias correction, and peak calling for ChIP-seq. *Stat. Appl. Genet. Mol. Biol.*, **11**, Article 9.

Heidari,N. *et al.* (2014) Genome-wide map of regulatory interactions in the human genome. *Genome Res.*, **24**, 1905–1917.

Heinz,S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

Javierre,B.M. *et al.* (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.e19.

Kundaje,A. *et al.*; Roadmap Epigenomics Consortium. (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Lareau,C.A. and Aryee,M.J. (2018) Hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods*, **15**, 155–156.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Mifsud,B. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.

Mumbach,M.R. *et al.* (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919–922.

Mumbach,M.R. *et al.* (2017) Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.*, **49**, 1602–1612.

Nolis,I.K. *et al.* (2009) Transcription factors mediate long-range enhancer--promoter interactions. *Proc. Natl. Acad. Sci. USA*, **106**, 20222–20227.

Pelikan,R.C. *et al.* (2018) Enhancer histone-QTLs are enriched on autoimmune risk haplotypes and influence gene expression within chromatin networks. *Nat. Commun.*, **9**, 2905.

Phanstiel,D.H. *et al.* (2015) Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, **31**, 3092–3098.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rao,S.S.P. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Ross-Innes,C.S. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

Servant,N. *et al.* (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, **16**, 259.

Shlyueva,D. *et al.* (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.

Simeonov,D.R. *et al.* (2017) Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature*, **549**, 111–115.

Stark,R. and Brown,G. (2011) DiffBind: differential binding analysis of ChIP-Seq peak data. *Bioconductor*. Available online at: http://bioconductor.org/packages/release/bioc/html/DiffBind.html.

Yao,L. *et al.* (2015) Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 550–573.

Zhang,Y. *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.