Bioinformatics, 35(22), 2019, 4764–4766
doi: 10.1093/bioinformatics/btz450
Advance Access Publication Date: 14 June 2019
Applications Note

Genome analysis

Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data

Jonathan Cairns[1,2,†], William R. Orchard[1,3,4,5,†], Valeriya Malysheva[1,3,4,†] and Mikhail Spivakov [1,3,4,*]

[1]Regulatory Genomics Group, Nuclear Dynamics Programme, Babraham Institute, Cambridge CB22 3AT, UK, [2]Data Sciences and Quantitative Biology, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge CB4 0WG, UK, [3]Functional Gene Control Group, Epigenetics Section, MRC London Institute of Medical Sciences, London W12 0NN, UK, [4]Institute of Clinical Sciences, Faculty of Medicine, Imperial College, London W12 0NN, UK and [5]Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, UK

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Inanc Birol

Received on January 22, 2019; revised on May 3, 2019; editorial decision on May 23, 2019; accepted on June 5, 2019

Abstract

Summary: Capture Hi-C is a powerful approach for detecting chromosomal interactions involving, at least on one end, DNA regions of interest, such as gene promoters. We present Chicdiff, an R package for robust detection of differential interactions in Capture Hi-C data. Chicdiff enhances a state-of-the-art differential testing approach for count data with bespoke normalization and multiple testing procedures that account for specific statistical properties of Capture Hi-C. We validate Chicdiff on published Promoter Capture Hi-C data in human Monocytes and CD4$^+$ T cells, identifying multitudes of cell type-specific interactions, and confirming the overall positive association between promoter interactions and gene expression.

Availability and implementation: Chicdiff is implemented as an R package that is publicly available at https://github.com/RegulatoryGenomicsGroup/chicdiff.

Contact: mikhail.spivakov@lms.mrc.ac.uk

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

1 Introduction

Differential signal detection in sequencing data is one of the most common tasks in genomic analyses. Multiple tools have been developed for this purpose, many of which, including DESeq and EdgeR, are based on the negative binomial models for count data (Anders and Huber, 2010; Robinson *et al.*, 2010). Such tools are theoretically suitable for the analysis of most sequencing data types, including chromatin immunoprecipitation and Hi-C, leading to the development of wrapper packages around DESeq and EdgeR that facilitate differential analyses for such data (Lareau and Aryee, 2018; Ross-Innes *et al.*, 2012). However, both of these algorithms have been developed with standard RNA sequencing data in mind and therefore not account for or benefit from the specific properties of data resulting from other assays, prompting the development of assay-specific differential analysis tools (Chen *et al.*, 2015; Liu and Ruan, 2017; Stansfield *et al.*, 2018; Xu *et al.*, 2008).

Capture Hi-C (CHi-C) is a powerful experimental technique for detecting chromosomal interactions globally and at high resolution (Schoenfelder *et al.*, 2015). In CHi-C, the genome-wide pulldown of pairs of interacting genomic fragments by Hi-C is followed by sequence capture to selectively enrich Hi-C material for interactions involving (at least on one end) fragments of interest, termed 'baits'. Differential analyses of CHi-C data are challenging due to sample normalization issues, sparsity and uneven signal-to-noise ratios across interaction distances and different capture baits, which are not accounted for by standard differential analysis algorithms.

We have previously reported CHiCAGO, a statistical pipeline for robust detection of significant interactions in CHi-C data from a single condition (Cairns *et al.*, 2016). Here, we present Chicdiff, an R package for differential CHi-C data analysis. Chicdiff combines moderated differential testing for count data implemented in DESeq2 (Love *et al.*, 2014) with CHi-C-specific procedures for signal normalization informed by CHiCAGO and *P*-value weighting. Jointly, procedures implemented in Chicdiff enable a robust and sensitive detection of differential interactions in CHi-C data.

2 Approach

A schematic of the overall analysis approach is presented in Supplementary Figure S1. The following sections and Supplementary Note describe specific steps in more detail.

2.1 Feature selection

CHi-C data are often sparse, particularly at large interaction distances, limiting the power of differential signal detection. In part, this problem can be mitigated based on the fact CHi-C signals commonly spread to adjacent fragments (Eijsbouts *et al.*, 2019), most likely owing to the tethering of these fragments into the vicinity of the baits by nearby specific interactions. Therefore, to increase power, Chicdiff pools read across several fragments (by default, five in each direction) surrounding each interacting fragment of interest for each bait. A functionality is provided to prioritize fragment-level interactions within each detected differentially interacting region *post-hoc* (see Supplementary Note).

2.2 Data normalization and significance testing

Typically, in differential count analyses, a single normalization (scaling) factor is estimated per sample to account for differences in library size. However, we found that in CHi-C data, normalization can be further improved by taking into account the differences in the background levels for specific pairs of fragments between samples. In CHi-C, unlike in many other data types, such as RNA-seq, it is possible to obtain such background estimates from the data, and procedures for this are implemented in the Chicago package (Cairns *et al.*, 2016). Chicdiff combines scaling factors based on these background estimates with sample-level scaling factors in a manner that minimizes the total dispersion of read counts across replicates and conditions at each interaction.

The count and scaling matrices generated as described above are provided as input for the DESeq2 package, which tests each interaction for differences between conditions using a negative binomial model with moderated dispersion estimation.

2.3 Weighted multiple testing treatment

As with other Hi-C-derived data types, signal-to-noise ratios and effect sizes in CHi-C data vary highly with interaction distance. This makes a strong case for non-uniform multiple testing correction, such that *P*-values for differential tests on longer-distance interactions are corrected more stringently compared with those on short-distance interactions. To do this, Chicdiff uses the Independent Hypothesis Weighting (IHW) method (Ignatiadis *et al.*, 2016) to learn *P*-value weights based on interaction distance in a manner that maximizes the number of rejected null hypotheses. However, training IHW weights on the test regions is not appropriate, since their *P*-values are often not uniform under the null due to selection bias, which violates IHW's core assumption. Therefore, instead we learn weights on a separate 'weight training set' of fragment pairs randomly drawn from the full interaction count data for each sample (i.e. not limited to CHiCAGO-detected significant interactions), thus avoiding selection bias. The distance-dependent weights learned this way are applied to the *P*-values in the test set, and the resulting weighted *P*-values are reported to the user.

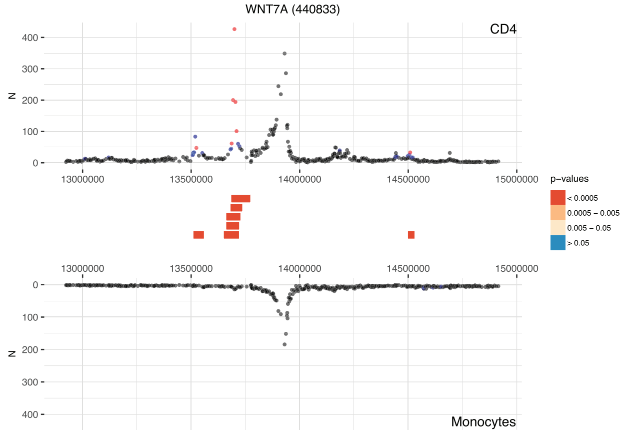

Fig. 1. Example of differential interactions detected by Chicdiff. Profiles of Promoter CHi-C interaction counts detected for *WNT7A* promoter in naive CD4$^+$ T cells (top) and monocytes (bottom) generated by Chicdiff (data from Javierre *et al.*, 2016). Significant interactions detected for each condition separately by CHiCAGO are colour-coded (blue: 3<score≤5; red: score>5). Significant differentially interacting regions detected by Chicdiff are depicted as red blocks. Interactions beyond 1 Mb each way cropped out

3 Use example

We applied Chicdiff to detect interactions specific to naive CD4+ T cells versus monocytes based on promoter CHi-C data from Javierre *et al.* (2016). This resulted in 208 232 detected differential interacting regions (weighted adjusted *P*-value <0.05; see Supplementary Table S1 for further summary statistics). An example of differential interactions is shown in Figure 1, and a heatmap of a subset of differential and non-differential interactions is shown in Supplementary Figure S2. As expected, differential promoter-interacting regions were enriched for differential enhancer activity between the two cell types (Supplementary Fig. S3). In addition, many genes whose promoters engaged in differential interactions showed consistent differences in expression (Supplementary Fig. S4). Supplementary Figures S5–S9 validate the Chicdiff approach by comparing the differential interaction calls obtained with and without pooling across multiple fragments, with Chicdiff versus standard DESeq2 normalization, and with and without *P*-value weighting, with respect to the expression of associated genes and other parameters.

Acknowledgements

We thank Chris Wallace for helpful discussions and Michiel Thiecke for testing the developmental version of Chicdiff.

Funding

This work was supported by core funding from UK Research and Innovation and a Babraham Institute Translational Advisory Group Award.

Conflict of Interest: J.C. is currently an employee of AstraZeneca.

References

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Cairns,J. *et al*. (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol*., **17**, 127.

Chen,L. *et al*. (2015) A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics*, **31**, 1889–1896.

Eijsbouts,C. *et al*. (2019) Fine mapping chromatin contacts in capture Hi-C data. *BMC Genomics*, **20**, 77.

Ignatiadis,N. *et al*. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods*, **13**, 577–580.

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.

Lareau,C.A. and Aryee,M.J. (2018) diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics*, **34**, 672–674.

Liu,L. and Ruan,J. (2017) Utilizing networks for differential analysis of chromatin interactions. *J. Bioinform. Comput. Biol*., **15**, 1740008.

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

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

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

Schoenfelder,S. *et al*. (2015) The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res*., **25**, 582–597.

Stansfield,J.C. *et al*. (2018) HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics*, **19**, 279.

Xu,H. *et al*. (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–2349.