Bioinformatics, 34(4), 2018, 675–677
doi: 10.1093/bioinformatics/btx625
Advance Access Publication Date: 3 October 2017
Applications Note

Genome analysis

HiCapTools: a software suite for probe design and proximity detection for targeted chromosome conformation capture applications

Anandashankar Anil, Rapolas Spalinskas, Örjan Åkerborg and Pelin Sahlén*

KTH – Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Solna 171 65, Sweden

*To whom correspondence should be addressed

Associate Editor: Bonnie Berger

Received on March 13, 2017; revised on September 5, 2017; editorial decision on September 27, 2017; accepted on September 29, 2017

Abstract

Summary: Folding of eukaryotic genomes within nuclear space enables physical and functional contacts between regions that are otherwise kilobases away in sequence space. Targeted chromosome conformation capture methods (T2C, chi-C and HiCap) are capable of informing genomic contacts for a subset of regions targeted by probes. We here present HiCapTools, a software package that can design sequence capture probes for targeted chromosome capture applications and analyse sequencing output to detect proximities involving targeted fragments. Two probes are designed for each feature while avoiding repeat elements and non-unique regions. The data analysis suite processes alignment files to report genomic proximities for each feature at restriction fragment level and is isoform-aware for gene features. Statistical significance of contact frequencies is evaluated using an empirically derived background distribution. Targeted chromosome conformation capture applications are invaluable for locating target genes of disease-associated variants found by genome-wide association studies. Hence, we believe our software suite will prove to be useful for a wider user base within clinical and functional applications.

Availability: https://github.com/sahlenlab/HiCapTools.

Contact: pelinak@kth.se

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

1 Introduction

Promoters play a pivotal role in regulating expression levels of the corresponding genes (Smale and Kadonaga, 2003). Promoters and enhancers contain binding sites for both ubiquitous and tissue-specific transcription factors, with chromatin loops bringing promoters proximal to distal enhancers enabling modulated regulation (Maston *et al.*, 2006; Shen *et al.*, 2012; Spitz and Furlong, 2012; Visel *et al.*, 2009). Chromosome conformation capture adapted for high-throughput sequencing (Hi-C) preserves DNA looping information and provides a list of regions in close proximity (Lieberman-Aiden *et al.*, 2009). This powerful methodology enabled us to understand how genomes are organized in 3D space within the nucleus (Dixon *et al.*, 2012; Ea *et al.*, 2015). However, a linear increase in Hi-C resolution requires a quadratic increase in sequencing depth, making it a costly method to detect interactions occurring between features such as promoters or enhancers. To map promoter-anchored interactions, a targeted Hi-C approach can be used where Hi-C material is hybridized to a set of promoter targeting sequence capture probes (Dixon *et al.*, 2012; Dryden *et al.*, 2014; Jäger *et al.*, 2015; Ma *et al.*, 2015; Sahlén *et al.*, 2015). This allows focusing on proximities of targeted regions and restores the linear relationship between the read depth and sensitivity. Hi-C uses a four or six cutter restriction enzyme for fragmentation of the genome, and this choice dictates the resolution of targeted Hi-C.

CHiCAGO is a software tool which can be used to detect DNA looping in targeted Hi-C data (Cairns *et al.*, 2016). It deploys a convolution of two distributions corresponding to Brownian collisions and technical noise to call feature interactions. Here we present HiCapTools, a software package that can fully support a targeted Hi-C experimental setup with modules to determine probe placement and to detect proximities in the output. In contrast to CHiCAGO, HiCapTools uses a negative control probe set to generate a background contact frequency distribution to calculate the statistical significance of observed proximities.

2 Implementation

HiCapTools has two modules: the first selects capture probes to target sequences of interest. The second module processes a mapped and filtered targeted Hi-C dataset (Wingett *et al.*, 2015) to detect proximities between probes and the rest of the genome. The two modules will be described separately below.

The software is implemented using C++11 and compiled using the gcc compiler (version 4.9) and packaged with CMake tool (version 3.5.1). Detailed instructions on installation and usage are provided in the supplementary text.

Module 1: probe design (PD1)

The module generates a list of regions that will be targeted by sequence capture probes. Since most chromosome conformation capture applications fragment the genome using restriction enzymes, informative products contain a ligation site between two restriction fragments, i.e. a junction. Therefore, probes are placed precisely next to the restriction sites to maximize capture of fragments containing junctions. HiCapTools is currently compatible only with Hi-C performed with restriction enzymes. The software requires two mandatory and two optional input files: coordinates of restriction fragments, a list of transcripts or features, coordinates of the repeat regions (optional) and alignability scores of the genome (optional). The module reads restriction fragment coordinates, repeat regions and alignability scores into memory (Kent, 2014), and stores them using interval trees (Garrison, 2015). It then reads features [such as transcripts and single nucleotide variants (SNVs)] and locates neighboring restriction sites. In the minimal mode, the module reports sequences that target restriction fragments closest to the features, as dictated by the lower and upper thresholds of distance from the feature (user provided). However, if repeat and alignability files are provided, regions with low sequence quality are avoided while placing probes. In this case PD1 successively searches restriction sites within a given distance from the feature and chooses probes satisfying the conditions set by the user (Supplementary Fig. S1 and Supplementary Material). The user can also set the distance between probes to avoid placing probes too close to each other.

Module 2: proximity detector (PD2)

The second module of the suite reports proximities between targeted regions and the rest of the genome. The module takes a sorted binary alignment file (BAM) and requires that invalid junctions and duplicate read pairs are removed beforehand. The program processes only pairs mapped on targeted regions. BAMTools is used to read the alignment files (Barnett *et al.*, 2011). First a probe region is set to determine all alignments on that probe (Supplementary Fig. S2).Then restriction fragments containing the mate of each alignment are located and counted. PD2 then uses fragments and counts to generate two lists of proximities – one that is between targeted and non-targeted regions (feature to distal) and the second between targeted regions themselves (feature to feature).

It is possible to include negative controls, i.e., a set of probes in the design that target regions with no known annotation or regulatory potential, and PD1 can select such regions and selects probes for targeting (Supplementary text). Proximities of such probes can then be used to obtain contact frequencies at different distances occurring due to structural constraints (Supplementary text). This is achieved by binning the observed distances of such probes (default bin size is 1 kb). Mean and standard deviation of each bin is calculated to generate contact frequency versus distance. To avoid over-penalizing for distances observed only a few times (particularly the case for distances over 200 kb), contact frequency distribution was smoothed using the moving average method (Kenney and Keeping, 1962). Statistical significance of observed proximities was assigned by means of a p-value, obtained relative to the background distribution of the corresponding distance bin, under a normality assumption (Gautschi, 1972; Bochkanov). The value of *P*-value based filtering was assessed using an in-house generated targeted Hi-C dataset obtained from the GM12878 cell line (Marco *et al.*, 2017). The dataset was overlapped with selected enhancer associated H3K4me1 peaks from ENCODE (ENCODE Project Consortium, 2004), and peaks obtained with the enhancer track of the tfNet-repository (Diamanti *et al.*, 2016). Overlap enrichment was calculated relative to promoter-distance and fragment-length matched but otherwise random regions. A stricter *P*-value increases enrichment (Supplementary Fig. S3a and b), in particular for regions at larger distance from the corresponding promoter. A control set with enhancer inactive H3K9me3 peaks lack a similar enrichment signal (Supplementary Fig. S3c).

We then processed the same dataset above using the CHiCAGO tool (Cairns *et al.*, 2016) and compared its regulatory element enrichment levels to those obtained with HiCapTools (Supplementary Material and Supplementary Table S1). We found that CHiCAGO performs better at short (<50 kb) distances whereas the opposite is true for distant (>500 kb) proximities (Supplementary Fig. S3b). The two tools show similar results at medium distances.

3 Discussion

Targeted Hi-C applications are gaining significant interest in the functional and disease biology fields, particularly in cases where non-coding variants play important roles. Special care is taken to make the software accessible for users with minimal bioinformatics background such as clinical researchers. Therefore, HiCapTools should be able to attract a wide user base given its relevance and convenience.

Acknowledgements

We thank Phil Ewels, Francesco Vezzi, Ayla De Paepe and Sailendra Pradhananga for their help in testing and naming the software.

Funding

This work has been supported by the Swedish Research Council [Grant Agreement No. 78081].

Conflict of Interest: none declared.

References

Barnett,D.W. *et al.* (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691–1692.

Bochkanov,S. ALGLIB. http://www.alglib.net

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

Diamanti,K. *et al.* (2016) Maps of context-dependent putative regulatory regions and genomic signal interactions. *Nucleic Acids Res.*, **19**, gkw800.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Dryden,N.H. *et al.* (2014) Unbiased analysis of potential targets of breast cancer susceptibility loci by capture Hi-C. *Genome Res*, **24**, 1854–1868.

Ea,V. *et al.* (2015) Contribution of topological domains and loop formation to 3D chromatin organization. *Genes (Basel)*, **6**, 734–750.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (80-.)*, **306**, 636–640.

Garrison,E. (2015) A minimal C++ interval tree implementation. https://github.com/ekg/intervaltree

Gautschi,W. (1972) Error function and fresnel integrals. In: Abramowitz,M., and Stegun,I.A. (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, pp. 297–309.

Jäger,R. *et al.* (2015) Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.*, **6**, 6178.

Kenney,J.F. and Keeping,E.S. (1962) *Mathematics of Statistics*. 3rd edn. Van Nostrand, Princeton, NJ.

Kent,J. (2014) ENCODE-DCC/kentUtils. Jim Kent command line bioinformatic utilities. https://github.com/ENCODE-DCC/kentUtils

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Ma,W. *et al.* (2015) Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat. Methods*, **12**, 71–78.

Marco,C. *et al.* (2017) Allele specific chromatin signals, 3D interactions and refined motif predictions for immune and B cell related diseases. NAR-03066-2017, In preparation.

Maston,G.A. *et al.* (2006) Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.

Sahlén,P. *et al.* (2015) Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol.*, **16**, 156.

Shen,Y. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

Smale,S.T., and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.

Spitz,F., and Furlong,E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.

Visel,A. *et al.* (2009) Genomic views of distant-acting enhancers. *Nature*, **461**, 199–205.

Wingett,S. *et al.* (2015) HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*, **4**, 1310.