Bioinformatics Advance Access published December 4, 2014

Application Note

HIPPIE: A high-throughput identification pipeline for promoter interacting enhancer elements

Yih-Chii Hwang[1], Chiao-Feng Lin[2,3], Otto Valladares[2,3], John Malamon[2,3], Pavel Kuksa[2,3], Qi Zheng[4], Brian D. Gregory[1,4*], and Li-San Wang[1,2,3*]

[1] Genomics and Computational Biology Graduate Group, University of Pennsylvania Perelman School of Medicine.
[2] Institute for Biomedical Informatics, University of Pennsylvania Perelman School of Medicine.
[3] Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine.
[4] Department of Biology, University of Pennsylvania, Philadelphia, PA

Associate Editor: Prof. Gunnar Ratsch

ABSTRACT

Summary: We implemented HIPPIE (High-throughput Identification Pipeline for Promoter Interacting Enhancer elements) to streamline the workflow from mapping raw Hi-C reads, identifying DNA–DNA interacting fragments with high confidence and quality control, detecting histone modifications and DNase hypersensitive enrichments in putative enhancer elements, to ultimately extracting possible intra- and inter-chromosomal enhancer–target gene relationships.

Availability: This software package is designed to run on high-performance computing clusters with Oracle Grid Engine (OGE). The source code is freely available under the MIT license for academic and nonprofit use. The source code and instructions are available at the Wang lab website (http://wanglab.pcbi.upenn.edu/hippie/). It is also provided as an Amazon Machine Image to be used directly on Amazon Cloud with minimal installation.

Contact: lswang@mail.med.upenn.edu or bdgregor@sas.upenn.edu

1 INTRODUCTION

Genome-wide chromosome conformation capture (Hi-C) has been utilized to reveal three-dimensional connectivity of chromatin regions in eukaryotic nuclei (Lieberman-Aiden *et al.*, 2009). Due to its capability to capture all possible chromatin interactions in a genome, it has been recently employed to observe long-range regulatory elements with their geographically proximal target gene promoters (Hwang *et al.*, 2013). Although there have been workflows successfully expediting the analysis of one-dimensional high-throughput sequencing results such as whole-exome sequencing, ChIP-seq, DNase-seq, and RNA-seq; there are limited tools to untangle two-dimensional DNA–DNA physical interactions using Hi-C datasets. In an effort to reduce the obstacle of processing these large-scale datasets, and to establish an analysis protocol to detect candidate long-range regulatory elements, we implemented an automated workflow that processes Hi-C results starting from read mapping with quality controls, and corrects for biases in interactions based on the linear distance, mappability, GC content, and fragment lengths of each pair of Hi-C reads. This pipeline identifies candidate promoter-interacting enhancer elements by integrating Hi-C results with epigenomics data such as histone modifications and DNase hypersensitivity sites.

2 METHODS

HIPPIE takes Hi-C raw reads as the input and generates a list of enhancers with their interacting target gene(s) as the output. We built HIPPIE with five step-wise phases (Figure 1): (I) read mapping, (II) quality control, (III) identification of significant DNA–DNA interacting regions, (IV) enhancer–target gene predictions, and (V) characterization of these long-range interactions. Although HIPPIE is streamlined and automated, each phase of HIPPIE can be independently called with commonly used file formats generated by different platforms and programs, such as FASTQ, SAM, BAM, or BED. Thus, it can readily be combined with other upstream processing and/or downstream analyses. The implementations of each phase are described below.

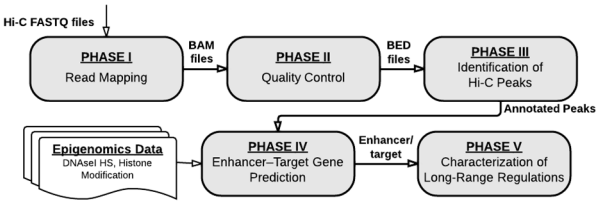

Fig. 1. An overview of HIPPIE.

Read mapping in HIPPIE uses the sequence alignment package BWA (Li and Durbin, 2009). It takes raw Hi-C paired-end sequencing reads in FASTQ format as input, and applies SAMtools (Li *et al.*, 2009) to compress the read alignment SAM files to BAM files and produces mapping quality metrics. The quality control steps discard reads not passing a user-defined mapping quality criterion (default minimum quality score = 30), remove potential PCR duplicates, ignore mitochondrial sequences, and exclude random contigs.

Identification of interacting DNA fragments consists of calling significant Hi-C peaks and annotating their genomic features. Because the resolution of Hi-C is constrained by the length distribution of the fragments produced by the chosen restriction enzyme (the sequence between two consecutive restriction sites along the genomic DNA), we retained the restriction fragments that harbor significantly higher specific than nonspecific read coverage (Supplementary Material) as "Hi-C peaks". Next, we applied BEDtools (Quinlan and Hall, 2010) to annotate these peaks with genetic features downloaded from the UCSC Genome Browser (Karolchik *et al.*, 2014), including annotations for promoters, exons, introns, other functional RNAs, etc.

Enhancer–target gene prediction reveals the interactions of the annotated peaks, and produces a list of candidate enhancer elements

*To whom correspondence should be addressed:
lswang@mail.med.upenn.edu, bdgregor@sas.upenn.edu.

(CEEs) and the gene(s) with which they interact as supported by Hi-C reads. To correct for Hi-C experimental biases in their linear distance between restriction fragments, GC content, mappability and length reported in (Yaffe and Tanay, 2011), we implemented the algorithm introduced by (Jin *et al.*, 2013) and extracted statistically significant DNA–DNA interactions (p-value ≤ 0.1, negative binomial distribution test). For enhancer prediction, our pipeline selects Hi-C peaks that interact with a promoter, reside in a DNase hypersensitive region, as well as harbor high levels of enhancer-associated histone modifications (H3K27ac or H3K4me1) but not promoter-associated marks or repressive marks (H3K4me3 and H3K27me3). An option of using ENCODE genome segmentations (Hoffman *et al.*, 2013) for candidate enhancers is also provided. This step is followed by characterization of enhancer–promoter interactions, which summarizes the overall properties of the interactions such as their linear distance distribution, as well as reports the enrichment of specific histone modifications and GWAS single nucleotide polymorphisms (SNPs) within the CEEs.

Note the phases are not only streamlined with error control, but also modularized for individual calls. For instance, users can map their Hi-C reads with other algorithms, and call peaks with HIPPIE starting at phase III; or one can directly import the interaction regions and utilize HIPPIE for enhancer–target gene identifications (phase IV).

3 USING HIPPIE

HIPPIE was built specifically for long-range enhancer–gene pair interaction detection upon the architecture of our previous DNA sequencing workflow (Lin *et al.*, 2013). For instance, we implemented job dependencies and error checking to automate the entire process. To run HIPPIE, users first prepare a configuration file describing the software and data paths, as well as their Hi-C library information. For each library, HIPPIE generates a corresponding bash script for Oracle Grid Engine job submission commands that can be invoked at the command line. When errors occur, all following jobs will be held for users to troubleshoot and re-execute the stalled phase or step. This modular architecture reduces the potential for unnecessary, repeated jobs. A complete run of HIPPIE produces candidate enhancer elements (CEEs) in BED format that are annotated with their target gene symbol(s), together with Hi-C read count supporting the interaction and interaction p-values.

To run HIPPIE, users can either install the package on their own cluster system, or simply access a pre-created Amazon Machine Image (AMI) from Amazon Web Services (AWS) on an Elastic Compute Cloud (EC2) instance (AMI ID: ami-3b0fb252).

We evaluated HIPPIE on our cluster using publicly available Hi-C datasets (Dixon et al., 2012). These datasets are 36 and 100 base pair (bp) paired-end sequencing with a total of 59.4 giga bases (1.35 billion single reads) from the Illumina GA II platform (GEO accessions GSM862723 and GSM892306). The total CPU time required for HIPPIE to process these datasets is 437.26 core-hours. The break-down of CPU time for each phase is as follows: read mapping: 64.4%, quality control: 5.8%, identification of peaks: 26.8%, enhancer–target gene interaction prediction: 2.8%, and characterization: 0.1%. The maximum memory usage is 4.77G for read mapping. We identified 3,707 candidate enhancer elements with 3,190 targeted RefSeq genes.

4 COMPARISON WITH OTHER TOOLS

While there are publicly available pipelines for processing Hi-C reads, there are no open-source software packages that take raw reads as input and ultimately identify enhancer–target gene pairs along with their interaction characteristics (Table 1). Among them, Hicpipe takes mapped reads and corrects the contact maps based on possible experimental biases (Yaffe and Tanay, 2011). HiC-inspector aligns reads and generates a contact matrix with user-defined read densities but does not have statistical filtering steps for the identified fragments (https://github.com/HiC-inspector). HiCUP maps reads with filtering out artifacts and self-interacting reads without any statistical model (http://www.bioinformatics.babraham.ac.uk/projects/hicup/). None of those identify long-range regulatory elements; nor provide error checking.

Table 1. Comparison among Hi-C processing pipelines

	HIPPIE	HiCUP	HiC-inspector	hicpipe
DNA – DNA Interactions				
Mapping algorithm	BWA	Bowtie	Bowtie	-
PCR artifacts filtering	✓	✓	-	-
Restriction Fragment size	Exact size	-	User-defined max. size	Bias correction
User-defined threshold for peak calling	✓	-	-	✓
GC-content normalization	✓	-	-	✓
Enhancer–target gene prediction				
Epigenomics Annotation	✓	-	-	-
Enhancer–target distance	✓	-	-	-
Enhancer GWAS enrichment	✓	-	-	-
Enhancer histone modification enrichment	✓	-	-	-

ACKNOWLEDGEMENTS

We thank Dr. Weixin Wang and Dr. Yuk Yee Leung for insightful discussions and the Gregory and Wang labs for constructive input. Funding: This work was supported by National Institute on Aging [U24-AG041689]; National Institute of General Medical Sciences [R01-GM099962]

REFERENCES

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 1–5.

Hoffman,M.M. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–41.

Hwang,Y.-C. *et al.* (2013) High-throughput identification of long-range regulatory elements and their target promoters in the human genome. *Nucleic Acids Res.*, **41**, 4835–46.

Jin,F. *et al.* (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–4.

Karolchik,D. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–70.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–60.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–93.

Lin,C. *et al.* (2013) DRAW+SneakPeek: Analysis Workflow and Quality Metric Management for DNA-Seq Experiments. *Bioinformatics*, btt422–.

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

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.