

ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis

Guipeng Li¹, Yang Chen¹, Michael P. Snyder^{2,*} and Michael Q. Zhang^{1,3,*}

¹MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China, ²Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA and ³Department of Biological Sciences, Center for Systems Biology, University of Texas, Dallas, 800 West Campbell Road, RL11, Richardson, TX 75080-3021, USA

Received June 18, 2016; Revised August 30, 2016; Accepted September 04, 2016

ABSTRACT

ChIA-PET2 is a versatile and flexible pipeline for analyzing different types of ChIA-PET data from raw sequencing reads to chromatin loops. ChIA-PET2 integrates all steps required for ChIA-PET data analysis, including linker trimming, read alignment, duplicate removal, peak calling and chromatin loop calling. It supports different kinds of ChIA-PET data generated from different ChIA-PET protocols and also provides quality controls for different steps of ChIA-PET analysis. In addition, ChIA-PET2 can use phased genotype data to call allele-specific chromatin interactions. We applied ChIA-PET2 to different ChIA-PET datasets, demonstrating its significantly improved performance as well as its ability to easily process ChIA-PET raw data. ChIA-PET2 is available at <https://github.com/GuipengLi/ChIA-PET2>.

INTRODUCTION

A number of high-throughput methods based on nuclear proximity ligation have been developed to detect genome-wide chromatin interactions, including high-throughput chromosome conformation capture (Hi-C) and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) (1,2). While Hi-C was developed to capture all chromatin interactions and is effective for mapping large-scale structures such as chromatin compartments and topologically associated domains (1,3), ChIA-PET is emerging as an important experimental method for detecting specific protein-mediated chromatin loops genome-wide at high resolution. In principle, ChIA-PET requires several main steps such as cross-linking the chromatin, proximity ligating the interacting fragments with linkers and sequencing the paired-ends of DNA fragments to estimate the frequency of chromatin interactions (2,4). The use of ChIA-PET has helped deepen our view of 3D genome organi-

zation and chromatin impact on gene regulation. For instance, ER- α -binding sites are anchored at gene promoters through long-range chromatin interactions to regulate target genes (2). CTCF, cohesin and ZNF143 are the key architectural factors of 3D chromatin structure (5,6). Enhancer-promoter interactions are highly cell-type specific (6). The *CFTR* gene promoter interacts with different distal cell-type specific regulatory elements that are all located within the same TAD (7). Cell identity controlling genes are located within CTCF-CTCF loops in mammalian chromosomes (8). Pivotal genes in the reprogramming process are transcribed within physical proximity to each other in embryonic stem cells (9). Chromatin loop disruption may alter gene expression (10), and more importantly, oncogene activation could be caused by genetic variants that disrupt chromatin loops (11). Allele-specific chromatin loops were observed at some imprinted loci, such as the *H19/IGF2* locus (5). All these findings require the accurate detection of chromatin loops, in which robust ChIA-PET assays and efficient data analysis play an important role.

However, similar to other genome-wide high-throughput sequencing experiments, ChIA-PET usually requires hundreds of millions of paired-end sequencing reads. The unique protocol of ChIA-PET requires a specific bioinformatics workflow to process and analyze data, which makes the study of ChIA-PET data challenging. A robust, versatile and easy-to-use analysis pipeline for ChIA-PET data is in great need, especially for experimental biologists. A workflow for processing ChIA-PET raw data often requires these common steps, linker trimming, read alignment, paired-end tag (PET) filtering, PCR duplicate removal, peak calling and chromatin interaction calling. There exist several published tools, e.g., ChIA-PET Tool (CPT) (12), ChiaSig (13), MICC (14), Mango (15) and 3CPET (16), to process and analyze ChIA-PET data. As summarized in Table 1 for the comparison of these tools with our ChIA-PET2, CPT fails to correct the major sources of bias (14,15) while ChiaSig and MICC focus on significant loop calling which is only a part of the data analysis workflow. Though Mango

*To whom correspondence should be addressed. Tel: +1 972 883 2528; Fax: +1 972 883 4551; Email: michael.zhang@utdallas.edu
Correspondence may also be addressed to Michael P. Snyder. Tel: +1 650 736 8099; Email: mpsnyder@stanford.edu