Standards and Guidelines for Chromatin Immunoprecipitation Based Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)
January 2017

I. Introduction

ChIA-PET has enabled identification of genome-wide chromatin-chromatin interactions mediated by a protein of interest. In brief, chromatin is subjected to cross-linking with formaldehyde after which an immunoprecipitation step is performed for a protein of interest. Strands of enriched chromatin undergo proximity ligation and are then subjected to paired-end sequencing. Each paired-end tag (PET) represents two regions found to interact. This document aims to outline standards in experimental methodology, sample and experimental recording, and data analysis that will guide the production of high quality ChIA-PET datasets. Due to rapid methodological and technological advances, this document should be revised annually.

II. Information to be supplied with each sample used for a ChIA-PET experiment.

1. For cell lines the following information should be recorded and provided -
    a) Cell line source and lot number.
    b) Growth time/passage number.
    c) Cell density.
    d) Protocol used to culture cell line.
2. Method of sonication used and number of cycles of sonication performed (if applicable).
3. Protocol used to generate ChIA-PET libraries. Generally, speaking there are two broad types of protocols that may be used, which differ in how paired-end tags are generated for Illumina-based sequencing. Briefly, interacting chromatin strands, which have been cross-linked, are ligated into circular molecules. These circular entities have to be "cut" and subjected to paired-end sequencing. This "cutting" can be done using either tagmentation or non-tagmentation based approaches. In non-tagmentation approaches (e.g. see Heidari et al., Genome Research 2014), restriction enzymes are used to cut the circularized fragments. Such approaches such as these tend to produce much shorter PETs (typically 15-30 bp) and involve more experimental steps as sequencing adaptors must be added. In contast, newer approaches typically use the Tn5 transposase which both cleaves the

circular fragment and attaches sequencing adaptors. These produce much larger mappable PETs (~100-150 bp).

4. Quantitation of gDNA produced after performing immunoprecipitation step as well as after library amplification via PCR. Quantitation should be ideally done using the Qiagen High-Sensitivity assay.
5. Linker sequence, if used.

III. Performance of ChIA-PET Sequence Experiment: Number of Replicates and Sequencing Depth

1. Replication: In order to ensure that the data are reproducible, experiments should be performed with two biological replicates. A biological replicate is defined as an independent growth of cells/tissue and subsequent analysis. Technical replicates of the same library are not required.
2. Sequencing depth: A full ChIA-PET dataset should contain >150 million reads for tagmentation-based protocols. This number was derived based on subsampling a large ChIA-PET dataset generated in the GM12878 lymphoblastoid cell line (LCL) and then examining the reproducibility in frequency of interactions across the genome. Very little gain in reproducibility was observed past 150 million raw reads. It is important to note, however, that the total number of called interactions (using the mango pipeline) is likely to increase with higher number of reads. The number 150 million was chosen based on observing no large gains in reproducibility and producing ~10,000 significant interactions (based on mango pipeline).

IV. Information supplied concerning pre- and post-sequencing mapping, read statistics and quality scores.

1. Pre-mapping data filtering/handling details. Details must be provided of data analysis steps undertaken prior to read mapping. These can include – trimming of low quality bases from reads, removal of adaptor sequences, and removal of linker sequences.
2. Mapping of sequence data. Information regarding which mapping program was used should be documented.
3. Mapping algorithm thresholds and settings.
    a. Number of allowable mismatches, minimal score, maximum allowed sum quality scores at mismatches, etc. Essentially, a list

of all parameters used by the mapping program should be recorded.

   b. Were reads only allowed to match uniquely or were multiple genomic mapping positions allowed?

4. Post-mapping data filtering/handling and results.
   a. Duplicate paired-end tags should be removed after mapping, i.e., PETs sharing both 5' read alignment positions.
   b. Any other post-mapping filtering steps should be detailed (e.g., removal of PETs which are either closer or further apart then some pre-set distance).
   c. The total number of "usable" PETs should be recorded. Here "usable" is defined as a uniquely mapping PET which has passed any post-mapping filtering steps.
   d. The percent of usable PETs - # of usable PETs / total # of PETs *100.
5. Report Spearman Rank correlation of the # of PETs connecting all peak-pairs with >2 PETs linking them between two replicates.
6. Report the percent of interactions found to overlap between the two replicates. Here an interactions is defined as FDR<10% as determined by the mango pipeline. Overlap is defined as the two anchor regions (i.e., peaks) being within 1 kbp of each other.