Updated on 05/03/2019

ATAC-seq Data Analysis

ATAC-seq (Assay for Transposase Accessible Chromatin with high-throughput Sequencing) is a next-generation sequencing approach for the analysis of open chromatin regions to assess the genome-wise chromatin accessibility. ATAC-seq achieves this by simultaneously fragmenting and tagging genomic DNA with sequencing adapters using the hyperactive Tn5 transposase enzyme [REF 1]. Other global chromatin accessibility methods include FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements followed by high throughput sequencing) and DNase-seq. This document aims to provide a workflow for the analysis of ATAC-seq data to identify differential chromatin accessibility.

The following tasks will be performed:

(i) Pre-processing of raw sequencing reads – before mapping the raw reads to the genome, the adapter sequences will be trimmed off. Poor read quality or sequencing errors often lead to low mapping rate.
(ii) Mapping/alignment of sequencing reads to a reference genome – Burrows-Wheeler Aligner (BWA) will be used for mapping of sequencing reads. The output alignment file will be saved as a sequence alignment/map (SAM) format or binary version of SAM called BAM. Duplicate reads will be marked using Picard [REF 2] while reads mapping to mitochondrial DNA and other chromosomes will be excluded from the analysis together with low quality reads (MAPQ<10 and reads in Encode black list regions) using SAMtools [REF 3].
(iii) Filtering and shifting of the mapped reads - the read position will be shifted +4 and -5 bp in the BAM file before peak calling. When the Tn5 transposase cuts open chromatin regions, it introduces two cuts that are separated by 9 bp. Therefore, ATAC-seq reads aligning to the positive and negative strands need to be adjusted by +4 bp and -5 bp respectively to represent the center of the transposase binding site. Picard CollectInsertSizeMetrics will be used to compute the fragment sizes on alignment shifted BAM files.
(iv) Identification and visualization of the ATAC-seq peaks – MACS2 will be used for peak calling with the parameters nomodel or BAMPE [REF 4] and the differentially enriched peaks will be identified using the MACS2 bdgdiff module. Individual peaks separated by <100 bp will be joined together. Peak annotation and functional analysis will be performed using the R package ChIPpeakAnno or HOMER [REF 5 & 6]. First, ATAC-seq peaks will be categorized into different groups based on the nearest RefSeq gene i.e. promoter, untranslated regions (UTRs), intron and exon. Second, peaks that are within 5 kb upstream and 3 kb downstream of the Transcription Start Site (TSS) are associated to the nearest genes. Finally, these genes are then analyzed for over-represented gene ontology (GO) terms and KEGG pathways using ChIPpeakAnno.

All sequencing tracks will be viewed using the Integrated Genomic Viewer (IGV) [REF 7].

Updated on 05/03/2019

REQUIREMENTS:

1. ATAC-seq paired-end FastQ reads.
2. Organism: Human/Mouse
3. Sample sheets. The sample sheets are CSV files that contain sample information.

DELIVERABLES:

1. FastQ data quality report
2. Differential analysis results
3. Functional analysis results
4. Peak visualization plots

REFERENCES:

1. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013; 10:1213–1218.
2. https://broadinstitute.github.io/picard/
3. http://www.htslib.org
4. https://github.com/taoliu/MACS
5. Zhu L, Gazin C, Lawson N, Pagès H, Lin S, Lapointe D, Green M (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics, 11(1), 237. ISSN 1471-2105.
6. http://homer.ucsd.edu/homer/ngs/
7. http://software.broadinstitute.org/software/igv/home

Questions? Comments?

Please email us at EICC@emory.edu