Updated on 01/14/2022

Assay for Transposase Accessible Chromatin with high-throughput Sequencing (ATAC-seq)

Preprocessing, aligning, and quality checking

Sequences are quality-checked using *FastQC*, and processed with *Trimmomatic* [1] to trim adapters and low quality bases. Next, reads are aligned to the reference genome with Burrows-Wheeler Aligner (*BWA-MEM*). Duplicate reads are then flagged by *Picard MarkDuplicates*. *SAMtools* [2] is used to remove duplicate, unmapped, non-uniquely mapped, and mitochondrial DNA reads, followed by removing sequences which were mapped to the ENCODE blacklist regions [3] via *BEDtools* [4]. Finally, reads aligned to the "+" strand are shifted +4 bp, and reads aligned to the "-" strand are shifted -5 bp, to adjust for a 9 bp target sequence duplication generated by Tn5 transposase [5]. A post-alignment quality assessment report is then created by ATACseqQC [6].

Peak calling and annotation

Peaks can be called using *MACS2* [7] or *HMMRATAC* [8]. *MACS2* employs a local Poisson model and represents a more widely used option. HMMRATAC is based on a semi-supervised machine learning approach and designed specifically for ATAC-seq data. Overall, *HMMRATAC* is considered to be more accurate than *MACS2* and generates additional nucleosome position information, but at the same time requires higher computational resources. Peak differential binding analysis can be done with *csaw* [9] or *DiffBind* [10]. *HOMER* [11] is used for peak annotation.

Required:

1. Raw data files (fastq)
2. Metadata spreadsheet with sample information

Deliverables:

1. Data quality report
2. Peak differential binding results
3. Peak annotation results

References:

1. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.
2. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

3. Amemiya, H.M., A. Kundaje, and A.P. Boyle, *The ENCODE blacklist: identification of problematic regions of the genome.* Scientific reports, 2019. **9**(1): p. 1-5.
4. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
5. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nature methods, 2013. **10**(12): p. 1213-1218.
6. Ou, J., et al., *ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data.* BMC genomics, 2018. **19**(1): p. 1-13.
7. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome biology, 2008. **9**(9): p. 1-9.
8. Tarbell, E.D. and T. Liu, *HMMRATAC: a Hidden Markov ModeleR for ATAC-seq.* Nucleic acids research, 2019. **47**(16): p. e91-e91.
9. Lun, A.T. and G.K. Smyth, *csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows.* Nucleic acids research, 2016. **44**(5): p. e45-e45.
10. Stark, R. and G. Brown, *DiffBind: differential binding analysis of ChIP-Seq peak data.* R package version, 2011. **100**(4.3).
11. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Molecular cell, 2010. **38**(4): p. 576-589.

Note: For sequencing data acquisition please contact Emory Integrated Genomics Core (EIGC@emory.edu).

Questions? Comments?

Please email us at EICC@emory.edu