

ATAC-seq analysis

Bingbing Yuan

BaRC Hot Topics –April 11th 2023
Bioinformatics and Research Computing
Whitehead Institute

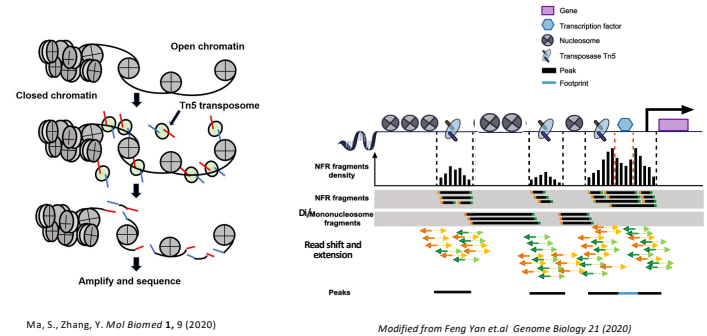


http://barc.wi.mit.edu/hot_topics/



1

Profile generation in ATAC-seq



- Open chromatin
- Motif enrichment
- TF footprinting
- Nucleosome positions

2

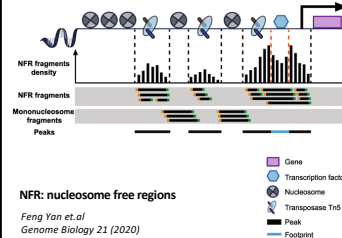
ATAC-seq pipeline goal

- Using short-read sequencing, identify genome-wide regions of open chromatin



3

ATAC-seq analysis workflow



Steps in ATAC-seq data analysis

Goal: Find the open chromatin regions
Peak = open chromatin region

1. **Quality control**
Remove adapters if necessary
2. **Mapping**
Tailored to paired end reads
3. **Post-alignment filtering**
4. **Peak calling and differential analysis**
 - i) Read shift and extension and signal profile generation.
 - ii) Peak assignment
5. **Peak interpretation**
 - i) Find genes next the open chromatin regions
 - ii) Find motifs within peaks



4

ENCODE ATAC-seq pipeline

- Human and mouse with biological replicates
- QC report ([a sample report](#))
- Includes all the steps in a single run

[Instructions to run pipeline using Whitehead server](#)

1. Pre-alignment quality control
2. Aligning reads to genome
3. Post-alignment filtering
4. Post-alignment quality control
5. Peak (accessible regions) calls
6. Assessing Peak Calls with FRiP score (same as ChIP-seq)
7. Blacklist filtering for peaks (same as ChIP-seq)



5



What sequencing works best for ATAC-seq?

- Read depth recommended by ENCODE:
25 million non-duplicate, non-mitochondrial aligned-reads for single-end sequencing
50 million for paired-ended sequencing (25 million fragments)
- No input control sample
- Shorter read lengths (50x50 or 75x75) better than longer reads (100x100 or longer)
- Pair-end reads are recommended over single reads



6



Illumina data format

• Fastq format:

http://en.wikipedia.org/wiki/FASTQ_format

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhghhhhhhhhhedhhhhfhhhhhh
```

/1 or /2 paired-end

→ @seq identifier
→ seq
→ +any description
→ seq quality values

Input qualities	Illumina versions
--solexa-quals	<= 1.2
--phred64	1.3-1.7
--phred33	>= 1.8



7



Check read quality with FastQC

(<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)

1. Run FastQC to check read quality

`fastqc sample.fastq`

2. Open output file:
"fastqc_report.html"

FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ⚠ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ⚠ Kmer Content

Basic Statistics

Measure	Value
Filename	Hept2H3kme3_subset.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1160004
Filtered Sequences	0
Sequence length	36
%GC	45



8



Pre-alignment quality control

1. Check reads quality with fastqc

fastqc read1.fq read2.fq

2. Remove adapter when necessary:

trim_galore --fastqc -nextera --paired --length 30 read1.fq read2.fq

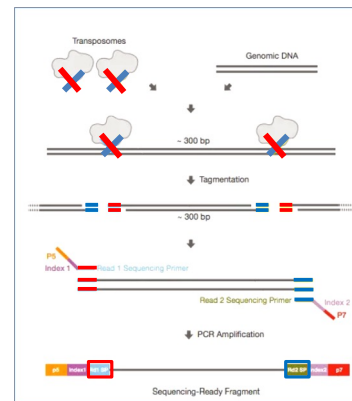
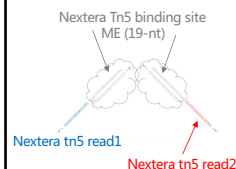
- -nextera: ATAC-seq experiments use the Nextera DNA Library Prep Kit
- --paired: both reads need to pass or they are both removed.
- --length: discard trimmed reads shorter than this length.



9



Nextera adapter in ATAC-seq



Modified from
<https://www.abpbio.com/product/tn5-transposase/>

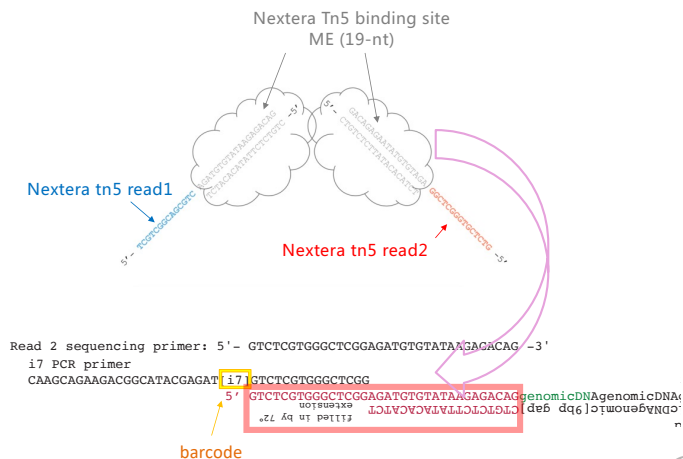


10

10



Nextera adapter in ATAC-seq



Modified from
<https://www.plob.org/article/11443.html>



11



Local genomic files needed for mapping

tak: /nfs/genomes/

- Human, mouse, zebrafish, *C.elegans*, fly, yeast, etc.
- Different genome builds
 - mm10: mouse_mm10_dec_11
 - mm39: mouse_mm39_jun20
- human_hg38_dec13 vs human_hg38_dec13_no_random
 - human_hg38_dec13 includes *_random.fa, *_hap*.fa, etc.
- Sub directories:
 - bowtie
 - Bowtie1: *.ebwt
 - Bowtie2: *.bt2
 - fasta: one file per chromosome
 - fasta_whole_genome: all sequences in one file
 - gtf: gene models from Refseq, Ensembl, etc.



12

12



Map reads to reference genome

Map reads with a non-spliced mapping tools: bowtie2 or BWA

```
bowtie2 --very-sensitive --no-discordant -p 2 -X 2000 -x hg38 -1 read1.fq -2 read2.fq | samtools view -ub - | samtools sort -> | bowtie_out.bam
```

- --no-discordant:
 - Suppress discordant alignments for paired reads
- -X 2000:
 - Increase maximum fragment length to 2k to include nucleosome distribution
 - Used for plotting fragment size distribution



13



Post alignment filtering

- Remove reads with low quality score: MAPQ < 30
`alignmentSieve -b file.bam --minMappingQuality 30 --samFlagInclude 2 -o MAPQ30.bam"`
- Remove duplicates with Picard's 'MarkDuplicates'
`java -jar picard.jar MarkDuplicates I=foo.bam O=noDups.bam M=foo.marked_dup_metrics.txt REMOVE_DUPLICATES=true`
- Remove reads mapped to mitochondria
`samtools view -h file.bam | grep -v chrM | samtools view -b -h -f 0x2 - | samtools sort -> file.sorted.bam`



`-f 0x2`
`samFlagInclude 2` } keep only properly paired reads

14



Post-alignment quality control

- Fragment size distribution
- TSS enrichment score

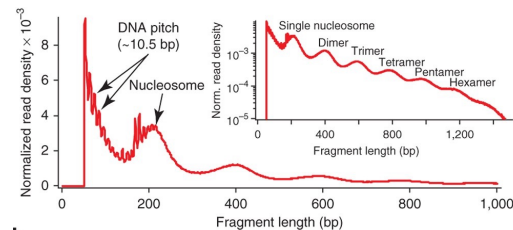


15



Fragment size distribution

```
calculate_ATACseq_fragment_size_distribution.R  
library("ATACseqQC")  
pdf("sample.fragment_sizes.pdf", w=11, h=8.5)  
fragSizeDist("sample.bam", "sample")  
dev.off()
```



Jason D Buenrostro, et.al. Nature Methods 10 (2013)



16

16



Profile generation in ATAC-seq

Ma, S., Zhang, Y. *Mol Biomed* 1, 9 (2020)

Modified from Feng Yan *et al* *Genome Biology* 21 (2020)

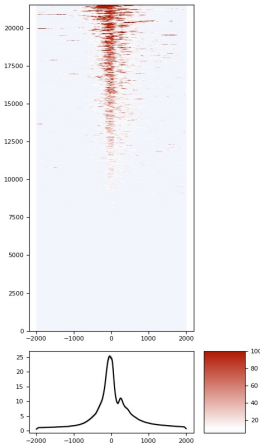


(Transcription Start Site) TSS enrichment score

1. Collect reads around 2000 bp around TSSs
2. Calculate average of read depth of the 100 bp bin at both ends
3. Get a fold change at each position over that average read depth
4. Score is the value at the center of the distribution

- TSS enrichment values depends on TSS annotation.
- Score with TSS per transcript is smaller than with one TSS per gene
- ENCODE standard with one TSS per gene:
<https://www.encodeproject.org/atac-seq/#standards>

```
calculate_TSS_enrichment_score.py
--outdir    ./tss
--outprefix sample_tss
--fastq1    sample_1.fq.gz
--tss       TSS.bed
--chromsizes chromInfo.txt
--bam       sample.bam >
sample_TSS_enrichment_score.txt
```



18

[illegible]

Convert bam to bed

- Consider both reads of a pair:
bedtools bamtobed -i foo.bam > foo.bed
- Convert an input paired-end alignment:

```
K00168:88:HFF7YBBXX:1:1203:28371:20304 163 chr1 3199165 42 51M = 3199703 589  
ACTAAAAACAGACAATGCTCAACATTTACATGAAATGAAGAATAAATAT AA-FFFJA-FAJJ<-<-FA-  
7AJ7-AJJF-<-77AJ-A<F<FJAFJFJJJA MD:Z:51 PG:Z:MarkDuplicates XG:i:0 NM:i:0 XM:i:0  
XN:i:0 XO:i:0 AS:i:0 YS:i:0 YT:Z:CP  
  
K00168:88:HFF7YBBXX:1:1203:28371:20304 83 chr1 3199703 42 51M = 3199165 -589  
AC TTGTGAAAAAATGAGTAA CAGACTTC TGT TAAAATGACCACAGTG TACT  
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ FFFAA MD:Z:51 PG:Z:MarkDuplicates XG:i:0 NM:i:0  
XM:i:0 XN:i:0 XO:i:0 AS:i:0 YS:i:0 YT:Z:CP
```
- To bed output file:

```
chr1 3199164 3199215 K00168:88:HFF7YBBXX:1:1203:28371:20304/2 42 +  
chr1 3199702 3199753 K00168:88:HFF7YBBXX:1:1203:28371:20304/1 42 -
```



Shift reads

- Reads should be shifted + 4 bp and - 5 bp for positive and negative strand respectively, to account for the 9-bp duplication created by DNA repair of the nick by Tn5 transposase

```
cat foo.bed | awk -F '\t' 'BEGIN {OFS = FS}{ if ($6 == "+") {$2 = $2 + 4} else if ($6 == "-") {$3 = $3 - 5} print $0}' >|
```

foo_tn5.bed

Bed format:

```
Chr start end read_name score strand
chr1 3199164 3199215 K00168:88:HFF7YBBXX:1:1203:28371:20304/2 42 +
chr1 3199702 3199753 K00168:88:HFF7YBBXX:1:1203:28371:20304/1 42 -
```

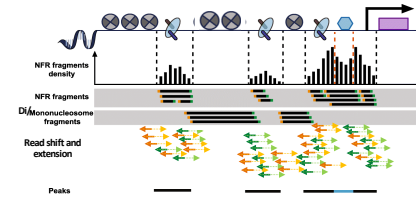


21



Identify potential open chromatin regions (peaks)

regions with reads piled up to a greater extent than the background read coverage



Modified from Feng Yan et.al Genome Biology 21 (2020)

Create a signal profile centered around the cutting sites



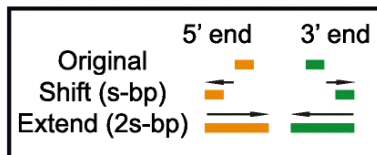
22



Call peaks

```
macs2 callpeak -t foo_tn5.bed -n foo -f BED -g mm -q 0.01 --nomodel --shift -75 --extsize 150 --call-summits --keep-dup all
```

Shift the reads to create a signal profile centered around the cutting sites



Feng Yan et.al Genome Biology 21 (2020)



23



MACS output

Output files:

- Excel peaks file (“_peaks.xls”) contains the following columns
Chr, start, end, length, abs_summit, pileup, -LOG10(pvalue), fold_enrichment, -LOG10(qvalue), name
- “_summits.bed”: contains the peak summits locations for every peaks. The 5th column in this file is -log10qvalue
- “_peaks.narrowPeak” is BED6+4 format file. Contains the peak locations together with peak summit, fold-change, pvalue and qvalue.



24



FRiP (Fraction of reads in peaks) score

- Fraction of all mapped reads that fall into the called peak regions
 - The higher the score, the better
- According to [ENCODE](#), score is preferably over 0.3, values greater than 0.2 are acceptable.

```
calculate_FRiP_score.py Sample.bam Sample_peaks.narrowPeak
```



25



Blacklist filtering for peaks

- Anomalous, unstructured, or high signal in next-generation sequencing experiments independent of cell line or experiment
- Often in repeats (centromeres, telomeres, satellite repeats)
0.5% of genome, but could account for >10% total signal
- BaRC_datasets -> ENCODE_blacklists
ENCODE: human, mouse, fly, or C. elegans

```
bedtools intersect -v -a foo_peaks.narrowPeak -b  
blacklist.bed > bfiltered_peaks.narrowPeak  
# -v option: only report those "A" peaks with no overlaps in "B"
```



26

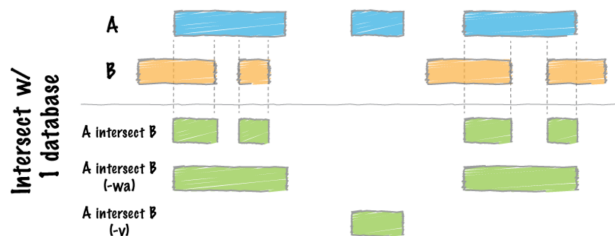


26

bedtools

- intersectBed

intersect



<https://bedtools.readthedocs.io/>



27



Visualize peaks in IGV



28



28

Compare open chromatin between different experimental conditions

- Without replicates, you can use bedtools to compare two samples:
intersectBed (finds the subset of peaks **common** in 2 samples **or unique** to one them)
- If you have replicates, you can use:
 - bedtools merge to merge the peaks
 - bedtools coverage to count the number of reads in peaks
 - DESeq2 on reads assigned the peaks to get differentially open peaks

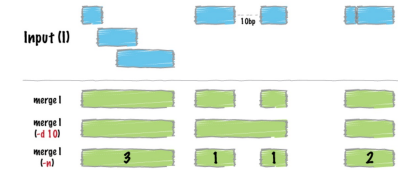


29



bedtools

merge



coverage

Below are the number of features in A (N=...) overlapping B and fraction of bases in B with coverage.

Chromosome	=====
BED File B	=====
BED File A	=====
Result	[N=3, 10/15] [N=1, 2/16] [N=1, 6/6] [N=5, 11/12]



30



Linking peaks to genes: Bedtools

slopBed: extend a feature by a user-defined number of bases

```
$ cat A.bed
chr1 0 100
chr1 500 580
$ cat my.genome
chr1 1000
$ slopBed -i A.bed -g my.genome -b 5
chr1 0 105
chr1 795 585
$ slopBed -i A.bed -g my.genome -l 2 -r 3
chr1 0 103
chr1 798 583
```

groupBy

It groups rows based on the value of a given column/s and it summarizes the other columns

closestBed

Chromosome	=====
BED File A	=====
BED File B	=====
Result	=====



31



Linking peaks to nearby genes:

Get all the genes at a certain distance (*i.e.* 3Kb) of the peak.
The distance we use depends on the area where we want to find regulatory interactions.

- Take all genes and add 3Kb up and down with slopBed
 - slopBed -b 3000 -i GRCh37.p13.HumanENSEMBLgenes.bed -g /nfs/genomes/human_gp_feb_09_no_random/anno/chromInfo.txt > HumanGenesPlusMinus3kb.bed
- Intersect the slopped genes with peaks and get the list of unique genes overlapping
intersectBed -wa -a HumanGenesPlusMinus3kb.bed -b peaks.bed | awk '{print \$4}' | sort -u > Genesat3KbforlessfromPeaks.txt

intersectBed -wa -a HumanGenesPlusMinus3kb.bed -b peaks.bed | head -3

```
chr1 45956538 45968751 ENSG00000236624.CCDC163P
chr1 45956538 45968751 ENSG00000236624.CCDC163P
chr1 51522509 51528577 ENSG00000265538.MIR4421
```



32



Link peaks to closest gene

For each region find the closest gene and filter based on the distance to the gene

The commands below are an example where we are looking for interactions at 3Kb or less.

`closestBed -d -a peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | head`

```
chr1 20870 21204 H3k4me3_chrl_peak_1 5,77592 chr1 14363 29806 ENSG00000227232_WASH7P 0
chr1 28482 30214 H3k4me3_chrl_peak_2 374,48264 chr1 29554 31109 ENSG00000243485_MIR1302-10 0
chr1 28482 30214 H3k4me3_chrl_peak_2 374,48264 chr1 14363 29806 ENSG00000227232_WASH7P 0
```

#the next two steps can also be done on excel

`closestBed -d -a peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10 -c 6,7,8, -o distinct,distinct,distinct | head -3`

```
ENSG00000227232_WASH7P 0 chr1 14363 29806
ENSG00000243485_MIR1302-10 0 chr1 29554 31109
ENSG00000227232_WASH7P 0 chr1 14363 29806
```

`closestBed -d -a peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10 -c 6,7,8, -o distinct,distinct,distinct | awk 'BEGIN {OFS="\t"} {if ($2<3000) {print $3,$4,$5,$1,$2}}' | head -5`

```
chr1 14363 29806 ENSG00000227232_WASH7P 0
chr1 29554 31109 ENSG00000243485_MIR1302-10 0
chr1 14363 29806 ENSG00000227232_WASH7P 0
chr1 134901 139379 ENSG00000237683_ALG27309.1 0
chr1 135141 135895 ENSG00000268903_RP11-34P13.15 0
```



33



Link peaks to closest gene (1 command)

For each region find the closest gene and filter based on the distance to the gene

The command below is an example where we are looking for interactions at 3Kb or less.

`closestBed -d -a peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10 -c 6,7,8, -o distinct,distinct,distinct | awk 'BEGIN {OFS="\t"} {if ($2<3000) {print $3,$4,$5,$1,$2}}' > closestGeneAt3KborLess.bed`

closestBed

-d print the distance to the feature in -b

groupBy

-g columns to group on

-c columns to summarize

-o operation to use to summarize



34



Identify footprints

- DNA sequences directly bound by TFs
- High depth of coverage:
At least ~200 million of reads
- Shift reads to account for the 9-bp duplication

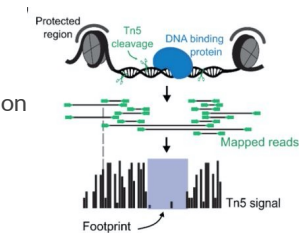
Tools:

HINT: (HmM-based IdeNtification footprints)

Genome Biol **20**, 45 (2019)

TOBIAS:

Nat Commun **11**, 4267 (2020)



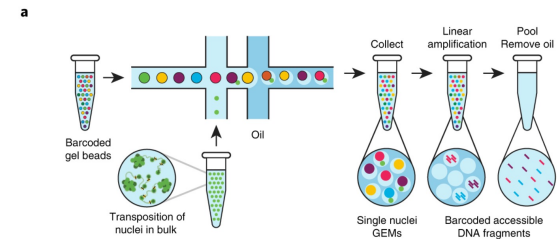
Bentsen, M., Goymann, P., Schultheis, H. et al. *Nat Commun* **11** (2020)



35



Single cell ATAC-seq (scATAC-seq)



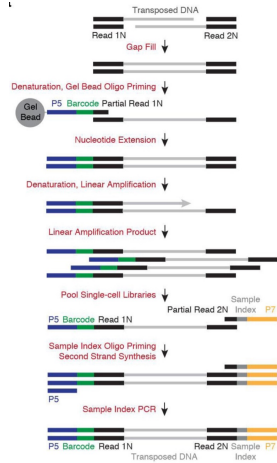
Nat Biotechnol **37**, 925–936 (2019).



36



scATAC-seq



Nat Biotechnol 37, 925–936 (2019).

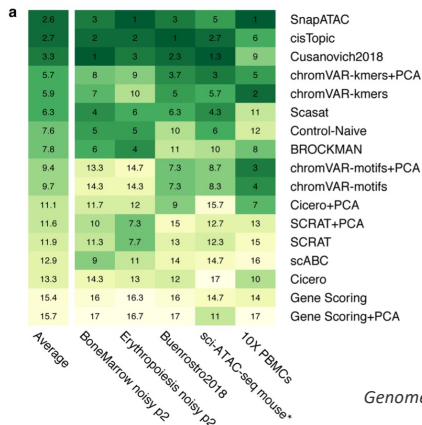
37

scATAC-seq

- scATAC-seq is good for heterogeneous samples with multiple cell types or cell states, where bulk ATAC-seq considers average cell signals and could miss signals from a subset of cells
- Signal is sparse in scATAC-seq
- Remove cells with low quality or doublets
- Make cell-by-feature matrix
- Generate and annotate cell clusters

38

scATAC-seq analysis tools



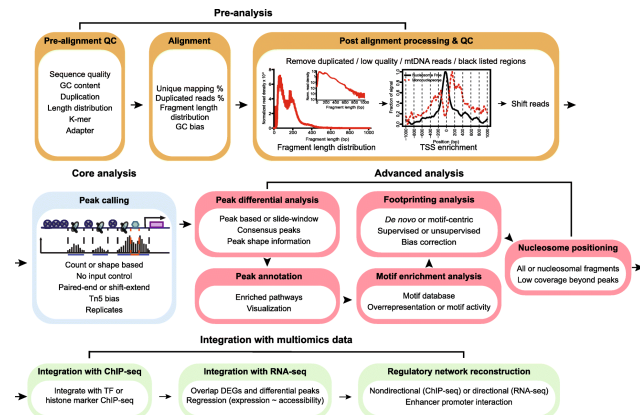
Computational and Structural Biotechnology Journal 18 (2020)

[Seurat scATAC-seq analysis pipeline:](#)

scRNA integration implements packages (MACS2, monocle, Cicero, etc)

40

ATAC-seq analysis



Feng Yan et al
Genome Biology 21 (2020)

41

Exercises:

- [Mouse lung tissue postnatal \(0 days\) from ENCODE](#)
Randomly chosen 100,000 pair-end reads from the 1st replicate
- \\wi-files1\BaRC_Public\Hot_Topics\ ATACseq_2023\ ATAC-seq_2023_commands.txt

42

References

- From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 21, 2020
- Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 2013
- ENCODE ATAC-seq guideline: <https://github.com/ENCODE-DCC/atac-seq-pipeline>
- MACS:
- Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, 2008 <https://liulab-dfci.github.io/software/>
- Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*. 2011
- Picard Tools: <https://broadinstitute.github.io/picard/>
- Bedtools: <https://code.google.com/p/bedtools/>
- BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics* 15, 2010
- Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* 37, 925–936 (2019)
- Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation. *Computational and Structural Biotechnology Journal*, 18, 2020
- Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol* 20, 241 (2019).
- Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* 20, 45 (2019)
- ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* 11, 4267 (2020)

43

Other resources

- Previous Hot Topics
Quality Control
http://barc.wi.mit.edu/education/hot_topics/NGS_QC_2017/slides4perPage.pdf
- Best Practices:
http://barcwiki.wi.mit.edu/wiki/SOPs/atac_Seq

44

Upcoming Hot Topics sessions

- Practical RNA-seq analysis - February 16
- Protein structure prediction with AlphaFold2 - March 23
- Intro to scRNA-seq analysis - March 28
- ATAC-seq analysis - April 11
- Enrichment analysis - April 27
- CRISPR pooled screen analysis – TBA
- Protein structure visualization (?) - let us know if you're interested!



45

