מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

Hands – on ATAC-seq analysis

Ester Feldmesser

Bareket Dassa

18-7-18

Nat Methods. 2013 Dec;10(12):1213-8. doi: 10.1038/nmeth.2688. Epub 2013 Oct 6.

Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.

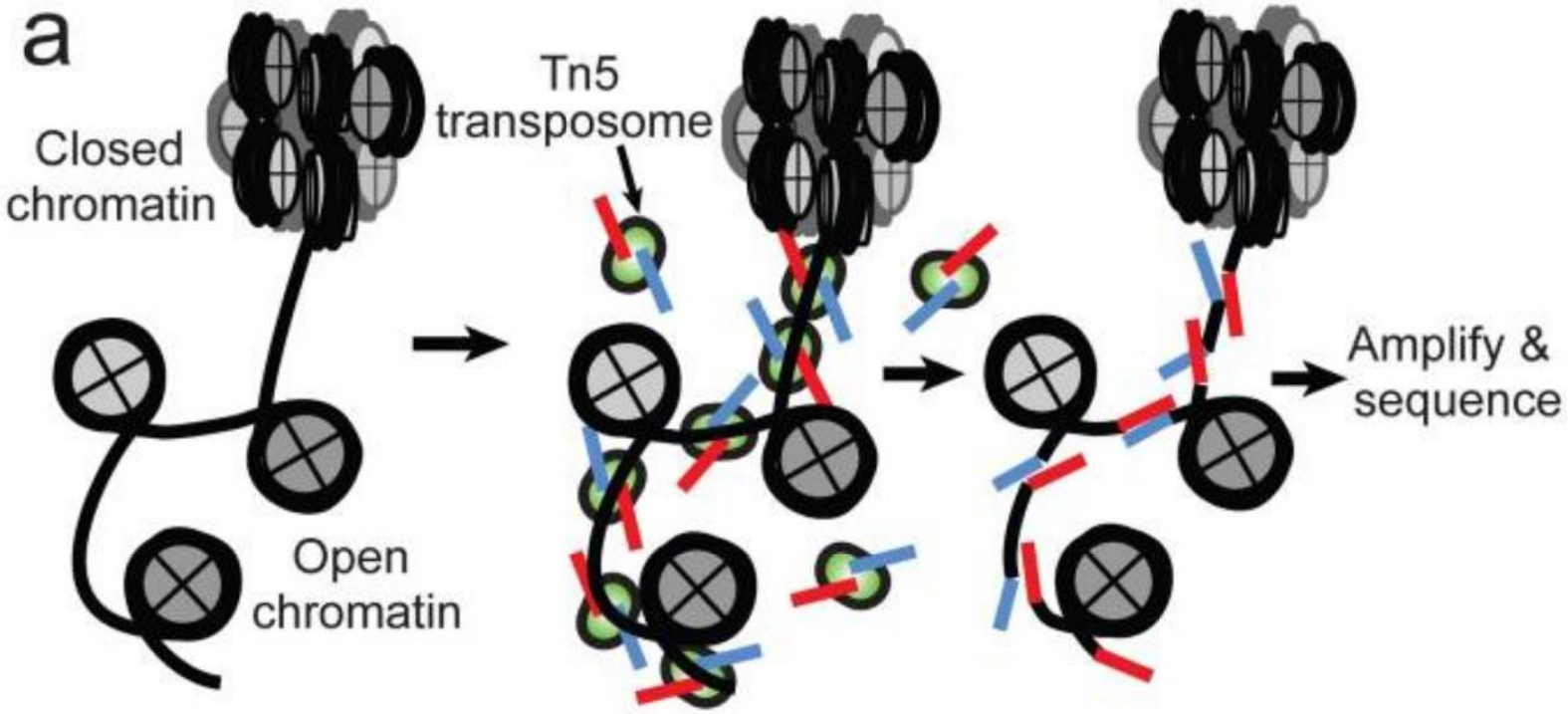
Buenrostro JD[1], Giresi PG, Zaba LC, Chang HY, Greenleaf WJ.

Author information

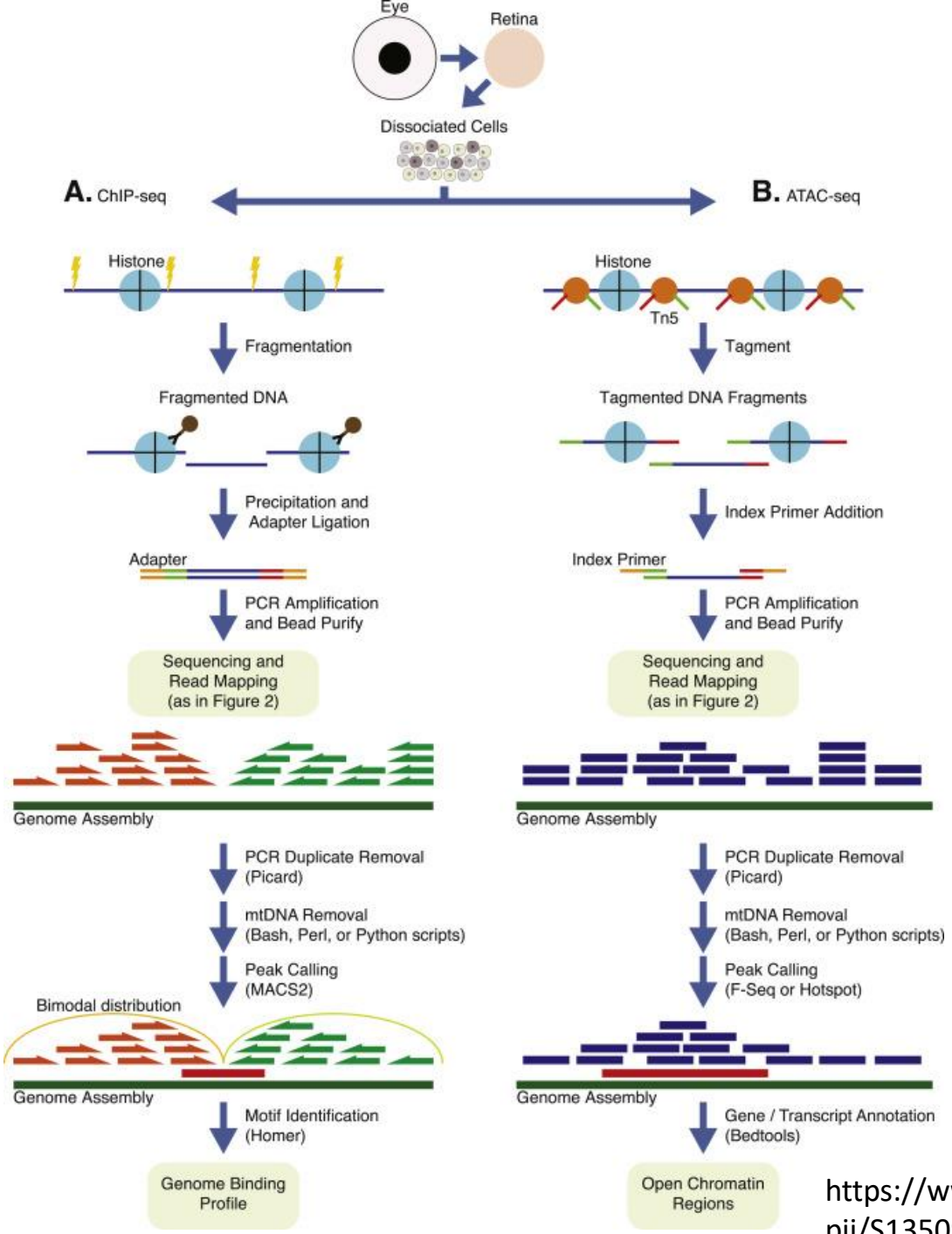
Abstract

We describe an assay for transposase-accessible chromatin using sequencing (ATAC-seq), based on direct in vitro transposition of sequencing adaptors into native chromatin, as a rapid and sensitive method for integrative epigenomic analysis. ATAC-seq captures open chromatin sites using a simple two-step protocol with 500-50,000 cells and reveals the interplay between genomic locations of open chromatin, DNA-binding proteins, individual nucleosomes and chromatin compaction at nucleotide resolution. We discovered classes of DNA-binding factors that strictly avoided, could tolerate or tended to overlap with nucleosomes. Using ATAC-seq maps of human CD4(+) T cells from a proband obtained on consecutive days, we demonstrated the feasibility of analyzing an individual's epigenome on a timescale compatible with clinical decision-making.

ATAC-seq is a sensitive, accurate probe of open chromatin state



(a) ATAC-seq reaction schematic. Transposase (green), loaded with sequencing adapters (red and blue), inserts only in regions of open chromatin (nucleosomes in grey) and generates sequencing library fragments that can be PCR amplified.



https://www.sciencedirect.com/science/article/pii/S1350946216300301

Experimental design

Control: Naked DNA, option to use a black list

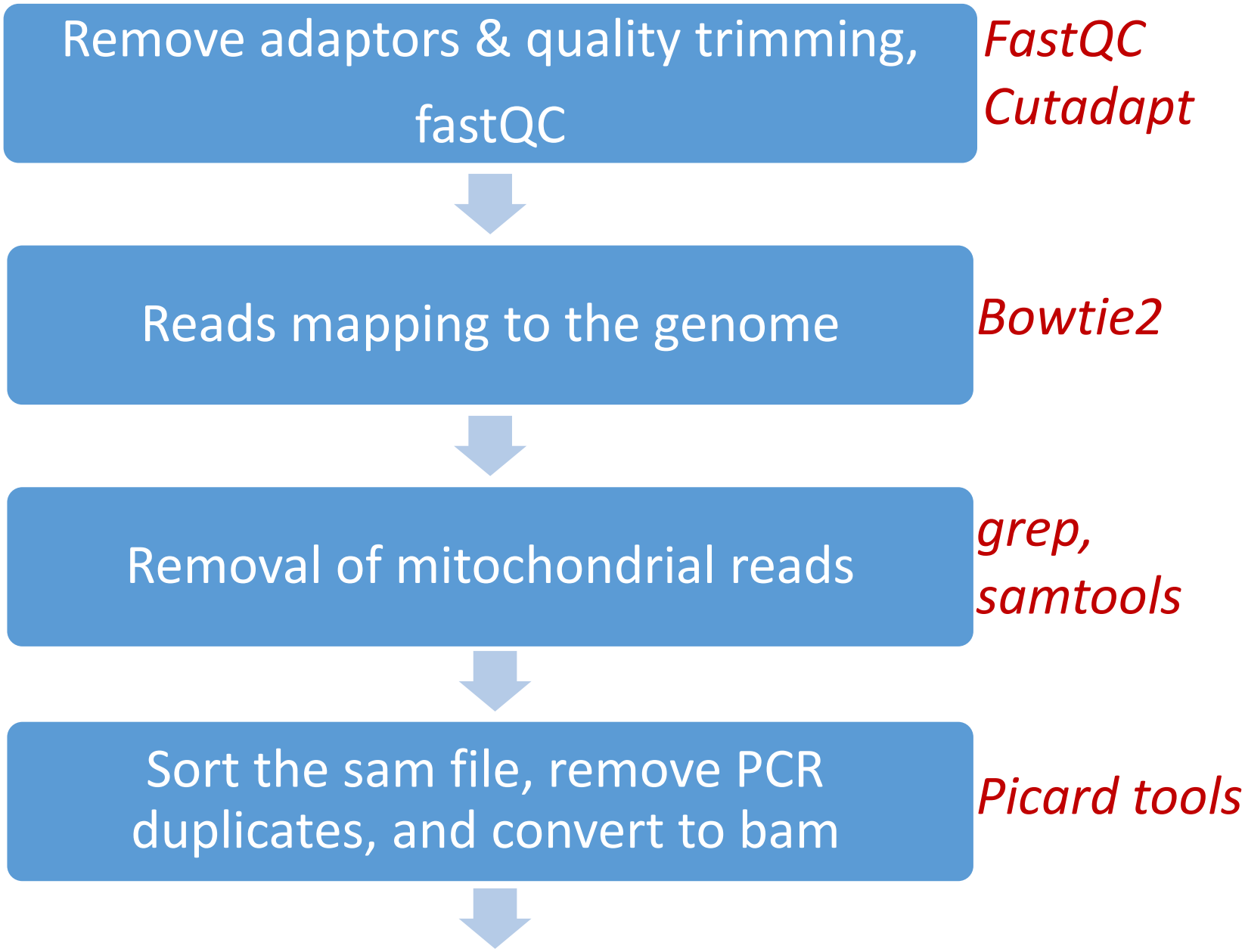
Replicates: At least 2 biological replicates

Library type: Paired-end

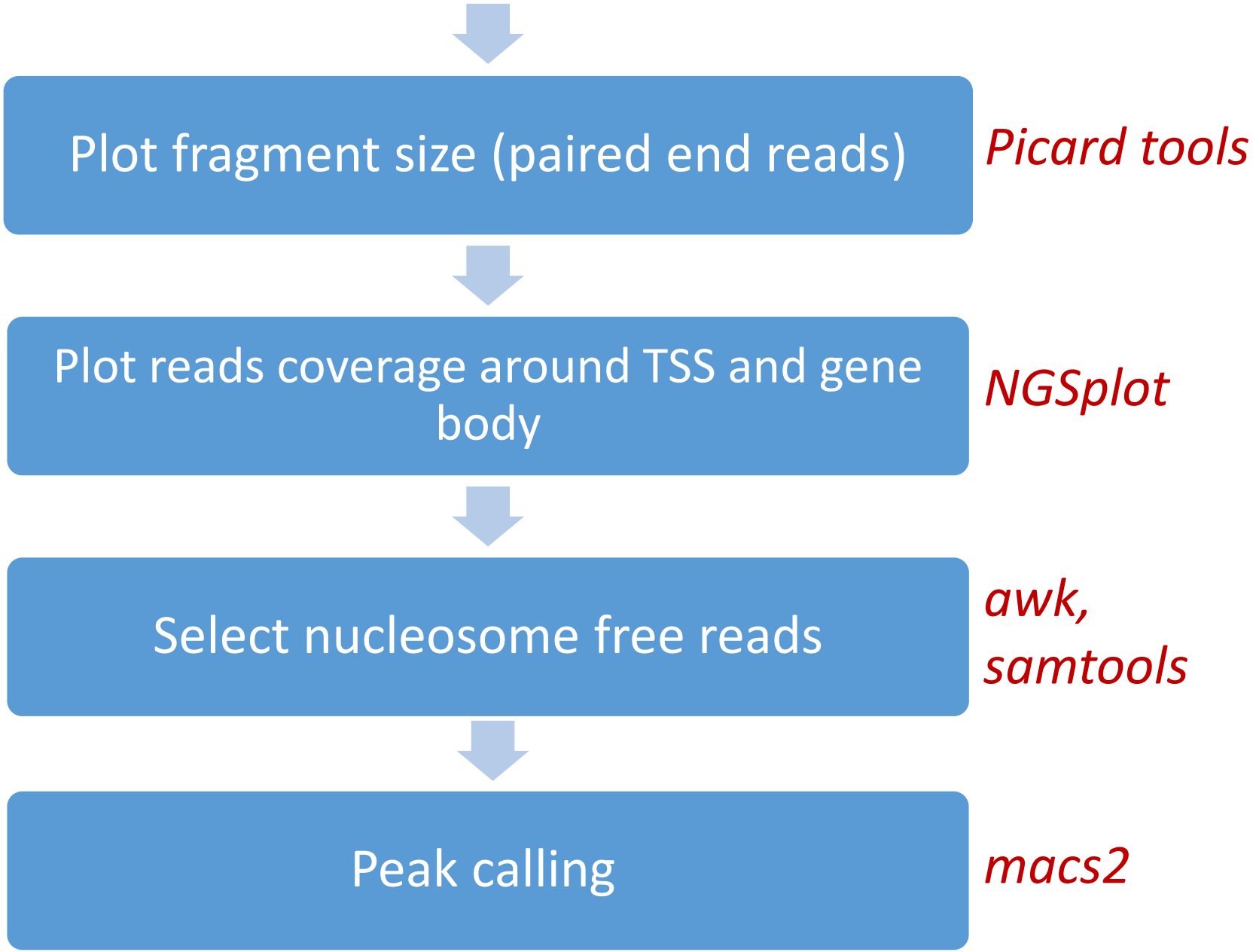
Sequencing depth: Depends on the genome size. Each replicate should have 50 million for paired-ended, non-duplicate, non-mitochondrial aligned reads (i.e. 25 million fragments).

Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Clifford A. Meyer & X. Shirley Liu
http://www.nature.com/nrg/journal/v15/n11/full/nrg3788.html#t1
Encode guidelines:
https://www.encodeproject.org/atac-seq/
The ENCODE blacklist :
https://sites.google.com/site/anshulkundaje/projects/blacklists

ATAC QC and analysis workflow



ATAC QC and analysis workflow (2)



ATAC-seq analysis pipeline by Tsviya Olender

- Perl scripts (for PE, no naked use in MACS2)
- Uses the WEXAC cluster
- Submission to bsub
- Generates log files

Commands run by the pipeline script

cutadapt -q 25 -a $adaptor1 -A $adaptor2 --minimum-length 30 -o $fastq1_t -p $fastq2_t $fastq1 $fastq2

fastqc -o $fastqcD -f fastq $fastqF

bowtie2 -X2000 --local -p4 --mm -x /shareDB/iGenomes/Mus_musculus/UCSC/mm10/Sequence/Bowtie2Index/genome -1 sample_name.R1.fastq.gz -2 sample_name.R2.fastq.gz -S sample_name.sam >& sample_name.log

grep -v 'chrM' sample_name.sam | **samtools view** -b -h -F 4 -f 0x2 - > sample_name_mapped.bam

java -jar /apps/RH7U2/general/picard/2.8.3/**picard.jar SortSam** SO=coordinate I=sample_name_mapped.bam O=sample_name_mapped_sorted.bam

java -jar /apps/RH7U2/general/picard/2.8.3/**picard.jar MarkDuplicates** INPUT=sample_name_mapped_sorted.bam OUTPUT=sample_name_mapped_sorted_rem.bam M=sample_name_metrics.txt REMOVE_DUPLICATES=true

samtools index sample_name_mapped_sorted_rem.bam

samtools flagstat sample_name_mapped_sorted_rem.bam > flagstat_ sample_name_mapped_sorted_rem

ngs.plot.r -G mm10 -R tss -C sample_name_mapped_sorted_rem.bam -O KO_2_high.tss -D refseq -T KO_2_high

ngs.plot.r -G mm10 -R genebody -C sample_name_mapped_sorted_rem.bam -O KO_2_high.genebody -D refseq -T KO_2_high

java -jar /apps/RH7U2/general/picard/2.8.3/**picard.jar CollectInsertSizeMetrics** I= sample_name_mapped_sorted_rem.bam MINIMUM_PCT=0.5 O= sample_name_ins_sz_mtrcs.log H= sample_name.pdf W=1000

samtools view sample_name_rem.bam | **awk** -F "\t" '{if (($9>- 120) && ($9< 120)) print $_}' >> sample_name_mapped_sorted_rem_temp.sam

samtools view -h -b sample_name_mapped_sorted_rem_temp.sam > sample_name_mapped_sorted_rem_free.bam

samtools index sample_name_mapped_sorted_rem_free.bam

igvtools count -w 5 sample_name_mapped_sorted_rem_free.bam sample_name_mapped_sorted_rem.tdf mm10

bedtools bamtobed -i sample_name_mapped_sorted_rem_free.bam > sample_name_mapped_sorted_rem.bed

samtools flagstat sample_name_mapped_sorted_rem_free.bam > sample_name_mapped_sorted_rem_free.flagstat

macs2 callpeak -t $openRegionsbam --bw 120 -B -f BAMPE --SPMR -B -g mm --nomodel --shift -50 --extsize 100 --broad -n $sample --keep-dup all --outdir MACS_2

Links to background material

fastq format: https://en.wikipedia.org/wiki/FASTQ_format
sam format: https://samtools.github.io/hts-specs/SAMv1.pdf
bam format: binary of sam
bam.bai format: index of bam
tdf: https://software.broadinstitute.org/software/igv/TDF

Fastqc: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml
Samtools: http://www.htslib.org/doc/samtools.html
Picard tools: http://broadinstitute.github.io/picard/
ngsplot: https://github.com/shenlab-sinai/ngsplot
https://github.com/shenlab-sinai/ngsplot/wiki/ProgramArguments101
MACS2: https://github.com/taoliu/MACS

IGV: https://software.broadinstitute.org/software/igv/download
GREAT: http://bejerano.stanford.edu/great/public/html/
Cistrome: http://cistrome.org/ap/root

How to run the pipeline

1. You need to copy the following folder to your home directory (folder name can be edited)
   cp -R /shareDB/ATAC-seq_pipeline/ .
   Instructions are in file "**how_to_run.txt**"
   This folder includes all the script files to run the pipeline, parameters and samples files

   atacpipeline_V2.pl
   collect_qual_params_v2.pl
   how_to_run.txt
   run_ATAC_Ts_V2_params.txt
   run_ATAC_Ts_V2.pl
   samples.txt
   lib/

2. Edit files and do not change their names:
   Prepare a file with names of all samples, i.e **samples.txt**
   Change hard coded flags in **run_ATAC_Ts_V2_params.txt**

```
[params]
genome = /shareDB/BioServices/bowtie2_db/mm10/mm10
adaptor1 = CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
adaptor2 = CTGTCTCTTATACACATCTGACGCTGCCGACGAGTGTAGATCTCGGTGGTCGCCGTATCATT
TSS_file = /home/labs/olenderlab/lvzvia/MyPipeLines/ATAC_V2/lib/TSS_+2500_-
2500_uniqueProm.bed
crude_reads_location_for_merging
=/home/labs/olenderlab/lvzvia/reinerlab/Aditya_AtacSeq/170517_NB501540_0020_AHNTWLB
GX2/fastq

[setup_run]
combine_fastq = 0      collects fastq files which are divided into several parts
fastqc = 0             run fastqc
trim_adapter = 1       runs cutadapt
make_body = 1          runs bowtie alignment
make_plots = 1         generates ngsplot plots
nucleosome_free = 1    filters bam file, to contain only paired-end reads with < 130bp insert
call_peaks = 1         uses MACS2 to call peaks
countTSS_reads = 1     counts the reads on the TSS regions, performs FPKM normalization
```

3. Change hard coded path to “$progD” in **atacpipeline_V2.pl**

4. Prepare input folder called “1_fastq” in the same script folder.

The expected name convention should be:

SRR5121093_R1_001.fastq.gz

SRR5121093_R2_001.fastq.gz

Fastq files should be compressed.

(note: the script also handles a format of multiple fastq files per sample, which we do no discuss here)

Run the pipeline

```
module load perl
perl PATH_TO_prog/atacpipeline_V2.pl samples.txt run_ATAC_Ts_V2_params.txt  new-short 8000
```

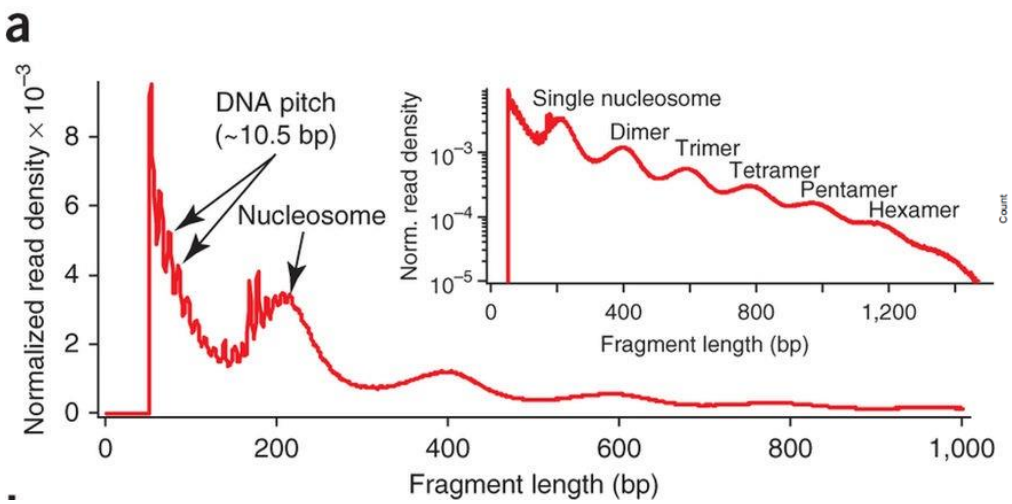
Output files

1_fastq
2_fastqc
3_align
4_plots
5_nucleosome_free
6_MACS_2
7_TSS
8_reports

igv.log
sample_err.txt
sample.log

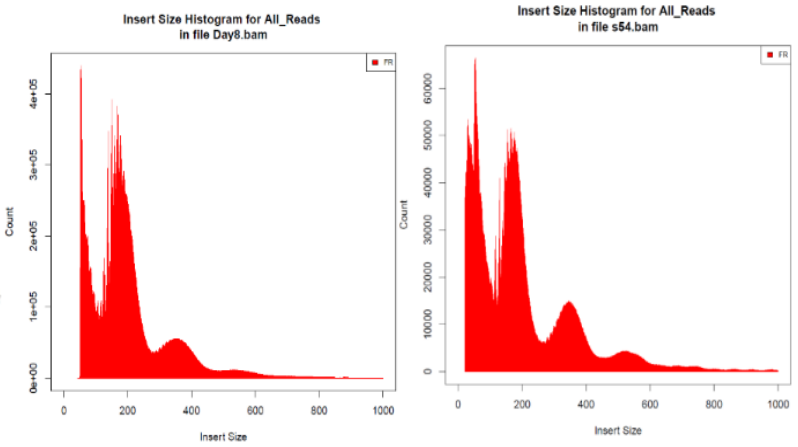
ATAC-seq quality control

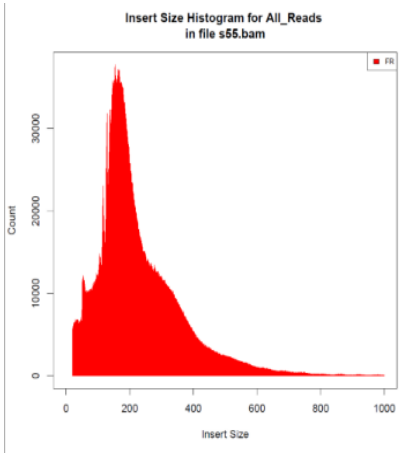
An example for an expected insert-size histogram:



https://www.nature.com/articles/nmeth.2688

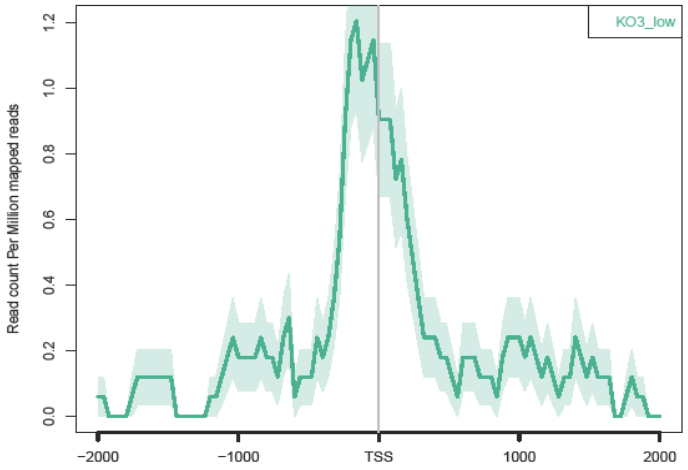
Example for a good sample



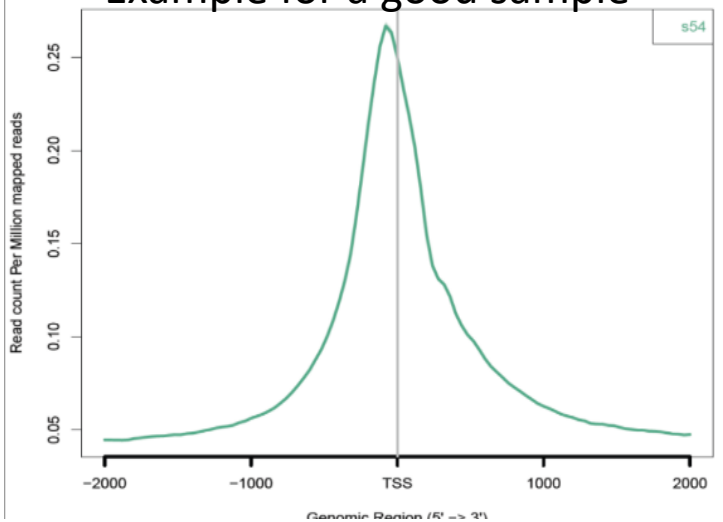


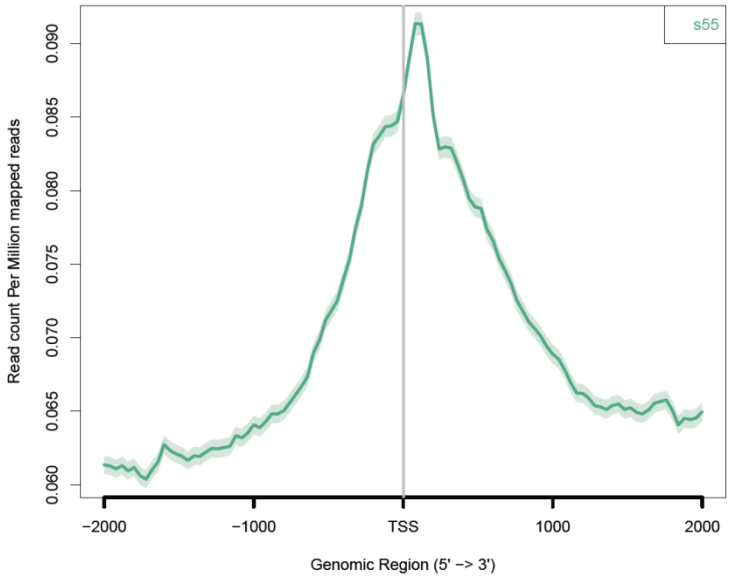
Example for a bad sample

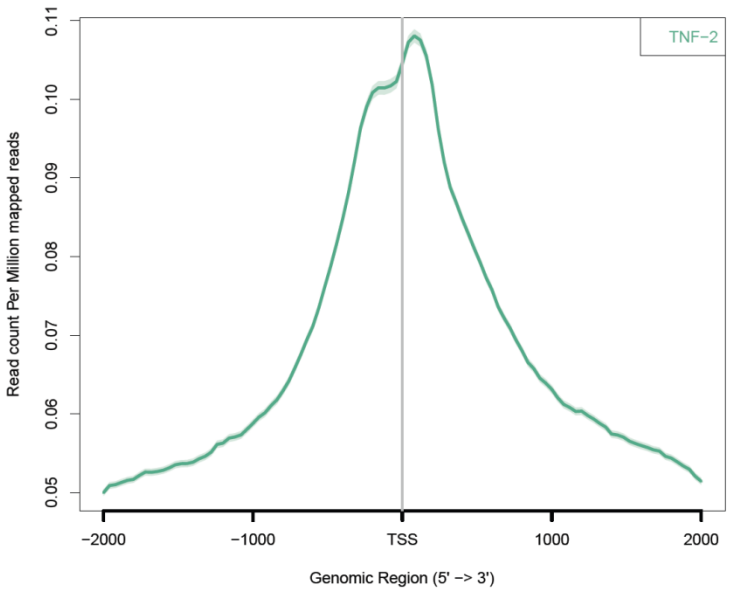
Average profiles of read coverage across TSS regions:



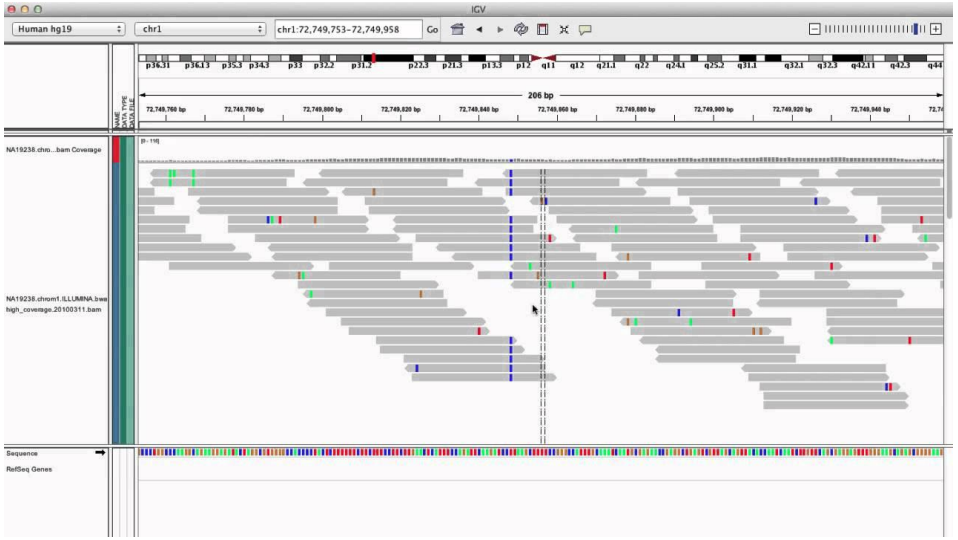
Example for a good sample





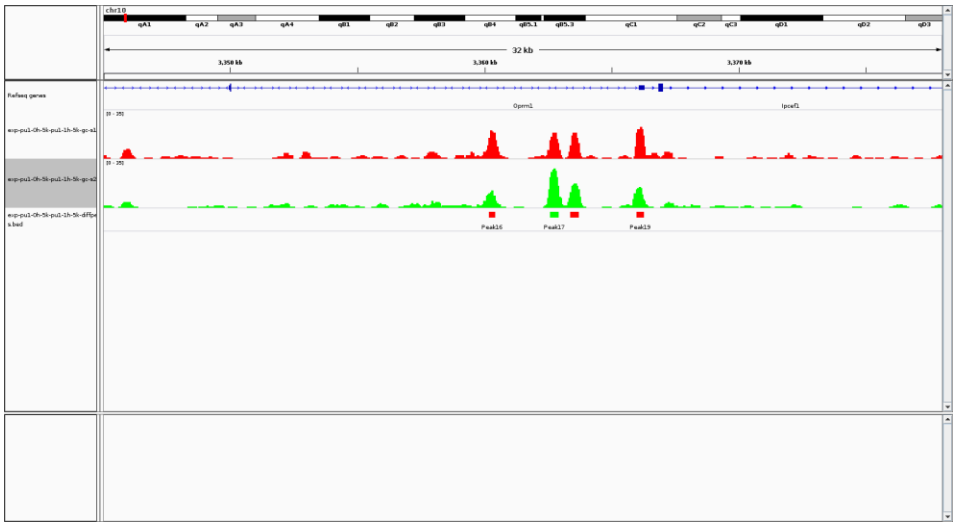


Visualization of read alignment: IGV



Input files (folder 5):

- **bam and bai**
- **tdf**



Summary of statistics

To generate a summary of statistics, and to activate the "collect_bedtools_count.pl" script:
Change the hard coded param: $pipelinePATH

perl ./collect_qual_params_V2.pl samples.txt > filename.txt

1. Example of output table:

sample	crudeReads	passedCutAdapt	Mapped	NucleosomeFree
SRR5121093	158501420	148741972	113280450	68814330
SRR5121094	130146804	122962938	68304344	34175582
SRR5121095	129649322	122002352	101210360	45549744
SRR5121096	143291944	133959192	116776692	48089750

* Read counts. For fragment, divide by 2

2. Example of “TSS_counts_table.txt”

				cKO_126_S5	cKO_130_S8	
chr18_37331527_37336527	Pcdhb6	+	2	0	0.203425358988774	0
chr18_77062943_77067943	Pias2	+	22	42	2.23767894887651	4.27193
chr18_50566199_50571199	Pudp	-	12	2	1.22055215393264	0.203425354

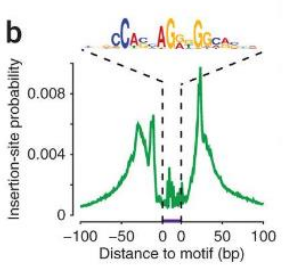
Downstream analysis

GREAT

NGSplot

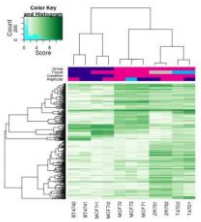
ChIPseeker

Homer



CEAS (cistrome)

Genomatix



Differential peak analysis

Diffbind