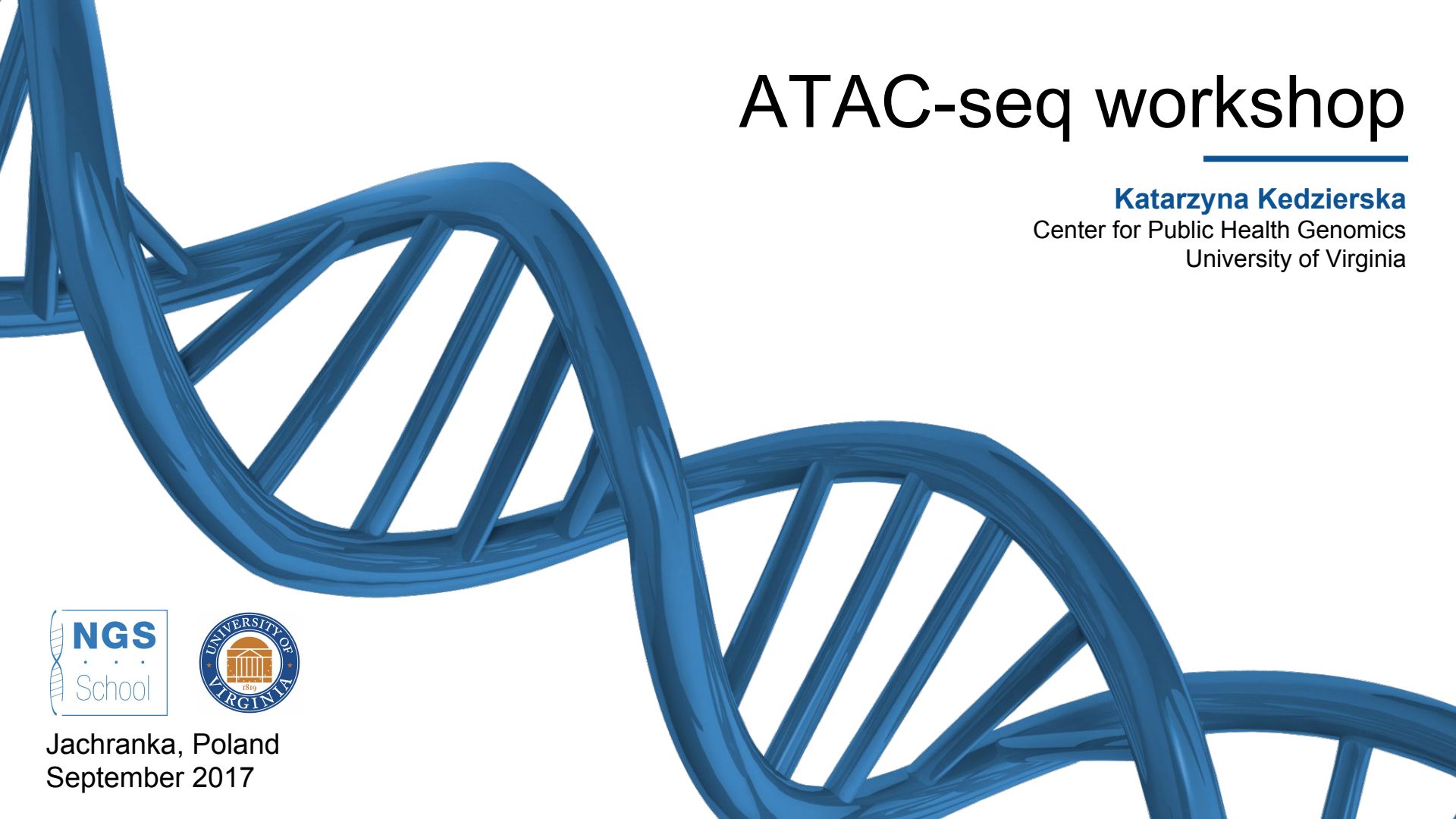
# ATAC-seq workshop

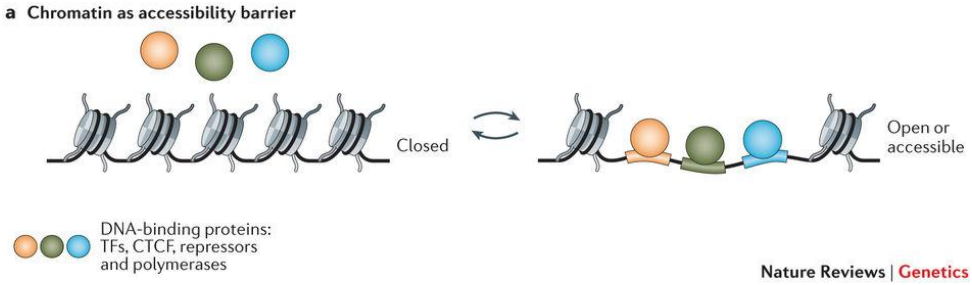
**Katarzyna Kedzierska**
Center for Public Health Genomics
University of Virginia



NGS School

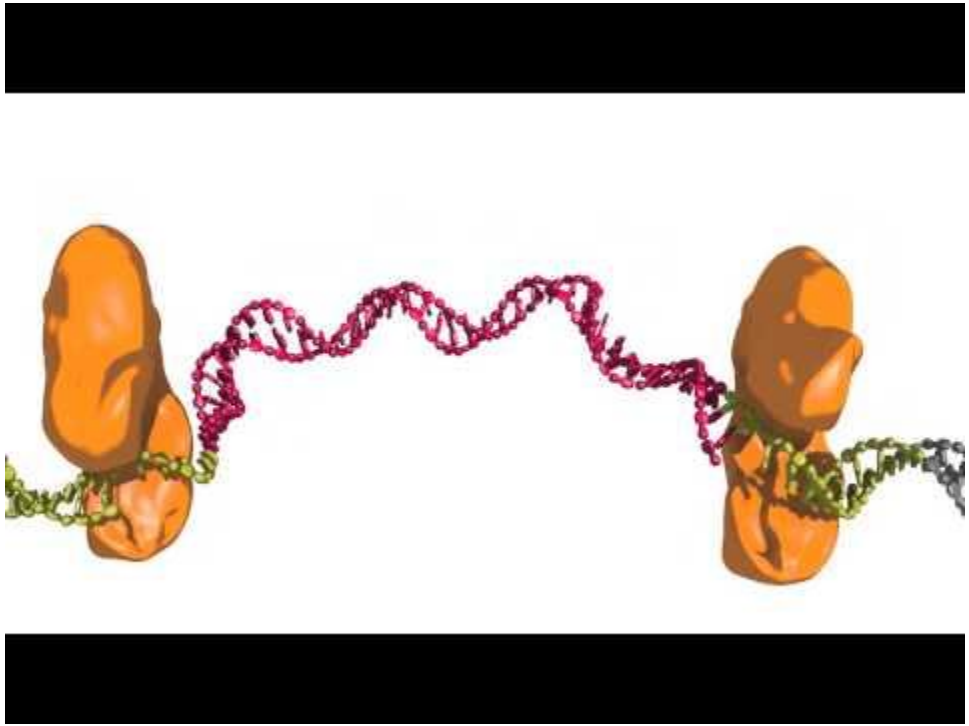
Jachranka, Poland
September 2017

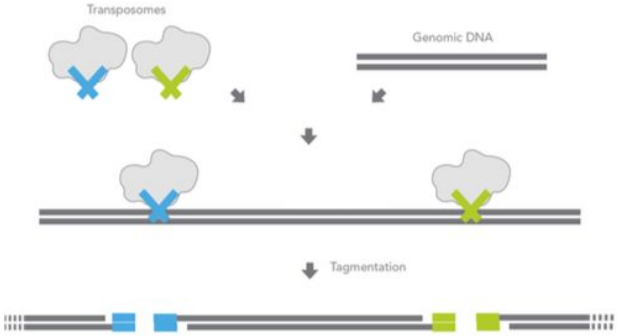
# Accessible chromatin

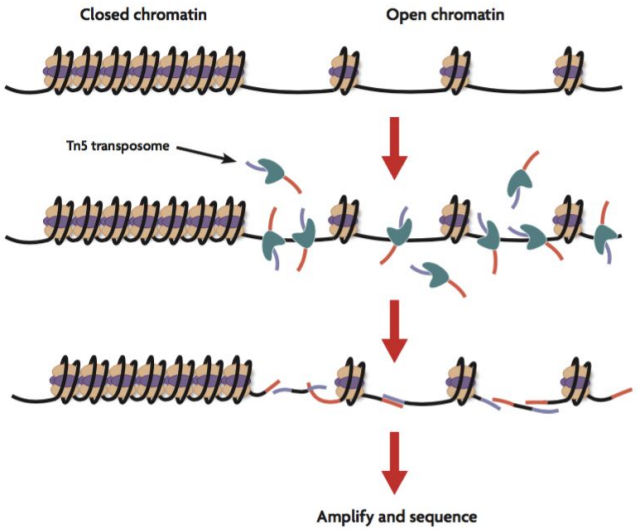



If chromatin is open there's place for DNA-binding proteins, like TFs or polymerases to bind.

# Transposition





# Research questions

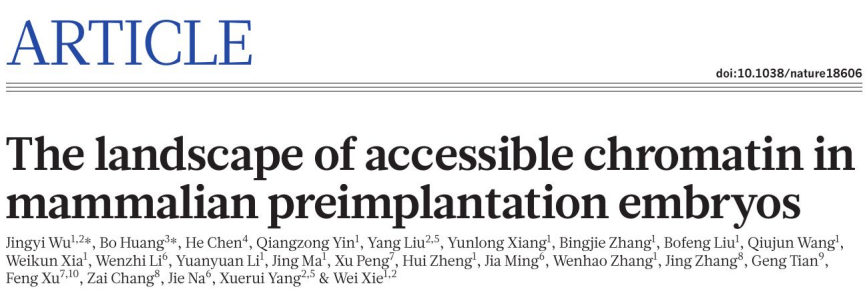
generate epigenomic profiles

map accessible chromatin across tissues or conditions

retrieve nucleosome positions

identify important transcription factors

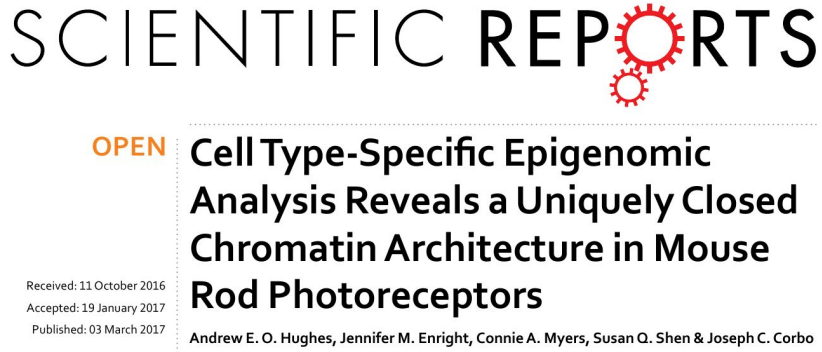
generate occupancy profiles of TFs (footprinting)


ARTICLE

doi:10.1038/nature18606

The landscape of accessible chromatin in mammalian preimplantation embryos

Jingyi Wu[1,2*], Bo Huang[3*], He Chen[4], Qiangzong Yin[1], Yang Liu[2,5], Yunlong Xiang[1], Bingjie Zhang[1], Bofeng Liu[1], Qiujun Wang[1], Weikun Xia[1], Wenzhi Li[6], Yuanyuan Li[1], Jing Ma[1], Xu Peng[7], Hui Zheng[1], Jia Ming[6], Wenhao Zhang[1], Jing Zhang[8], Geng Tian[9], Feng Xu[7,10], Zai Chang[8], Jie Na[6], Xuerui Yang[2,5] & Wei Xie[1,2]


SCIENTIFIC REPORTS

OPEN

Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors

Received: 11 October 2016
Accepted: 19 January 2017
Published: 03 March 2017

Andrew E. O. Hughes, Jennifer M. Enright, Connie A. Myers, Susan Q. Shen & Joseph C. Corbo

# Advantages of ATAC-seq

doesn’t require sonication or phenol-chloroform extraction (FAIRE-seq)

no antibodies needed (ChIP-seq)

no sensitive enzymatic digestion (MNase-seq or DNA-seq)

and significant reduction of the required input material and time needed to process the samples

# Variation of the method

**scATAC-seq** - single cell ATAC-seq

LETTER

doi:10.1038/nature14590

**Single-cell chromatin accessibility reveals principles of regulatory vari**

Jason D. Buenrostro[1,2], Beijing Wu[1]*, Ulrik
Howard Y. Chang[2] & William J. Greenleaf[1]

ARTICLES

nature genetics

Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution

M Ryan Corces[1–3,11], Jason D Buenrostro[3–5,11,12], Beijing Wu[4], Peyton G Greenside[4,6], Steven M Chan[7], Julie L Koenig[1,2], Michael P Snyder[3,4], Jonathan K Pritchard[4,8,9], Anshul Kundaje[4,10], William J Greenleaf[3,4], Ravindra Majeti[1,2,12] & Howard Y Chang[3,12]

**FastATAC** - one-step membrane permeabilization and transposition, requires 5k cells; optimized for primary blood cells

**OmniATAC** - modified version of ATAC protocol, published on Aug 28th 2017.

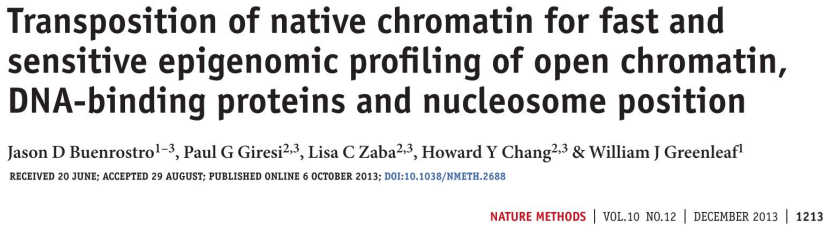
An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues

M Ryan Corces, Alexandro E Trevino, Emily G Hamilton, Peyton G Greenside, Nicholas A Sinnott-Armstrong, Sam Vesuna, Ansuman T Satpathy, Adam J Rubin, Kathleen S Montine, Beijing Wu, Arwa Kathiria, Seung Woo Cho, Maxwell R Mumbach, Ava C Carter, Maya Kasowski, Lisa A Orloff, Viviana I Risca, Anshul Kundaje, Paul A Khavari, Thomas J Montine, William J Greenleaf & Howard Y Chang

Affiliations | Contributions | Corresponding authors

*Nature Methods* (2017) | doi:10.1038/nmeth.4396
Received 04 April 2017 | Accepted 11 July 2017 | Published online 28 August 2017

# Sources of information

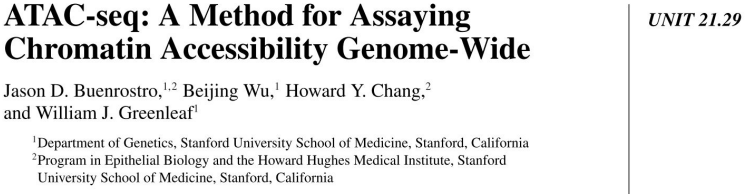


**Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position**

Jason D Buenrostro[1–3], Paul G Giresi[2,3], Lisa C Zaba[2,3], Howard Y Chang[2,3] & William J Greenleaf[1]

RECEIVED 20 JUNE; ACCEPTED 29 AUGUST; PUBLISHED ONLINE 6 OCTOBER 2013; DOI:10.1038/NMETH.2688





**ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide**

Jason D. Buenrostro,[1,2] Beijing Wu,[1] Howard Y. Chang,[2] and William J. Greenleaf[1]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California
[2]Program in Epithelial Biology and the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California



ATAC-seq forum:
https://sites.google.com/site/atacseqpublic/home

Encode guidelines:
https://www.encodeproject.org/atac-seq/ - official pipeline currently in beta tests

Bioconductor support, Biostars forum.

LETTER

doi:10.1038/nature14590

**Single-cell chromatin accessibility reveals principles of regulatory variation**

Jason D. Buenrostro[1,2], Beijing Wu[1]*, Ulrike M. Litzenburger[2]*, Dave Ruff[3], Michael L. Gonzales[3], Michael P. Snyder[1], Howard Y. Chang[2] & William J. Greenleaf[1,4]

# ATAC-seq experimental procedure

**Input:** crude nuclei or fixed tissue

ATAC-see reveals the a
transposase-mediated

Xingqi Chen[1], Ying Shen[1], Will Draper[2], Jasc
Ansuman T Satpathy[1], Ava C Carter[1], Rajarsl
William J Greenleaf[1,3,8], Jan T Liphardt[2] & H

RECEIVED 22 AUGUST; ACCEPTED 19 SEPTEMBER; PUBLISHED ONLINE 17

*Published online 28 November 2016*

*Nucleic Acids Research, 2017, Vol. 45, No. 6* ***e41***
*doi: 10.1093/nar/gkw1179*

## Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes

**Zefu Lu[1], Brigitte T. Hofmeister[2], Christopher Vollmers[3], Rebecca M. DuBois[3] and Robert J. Schmitz[1,*]**

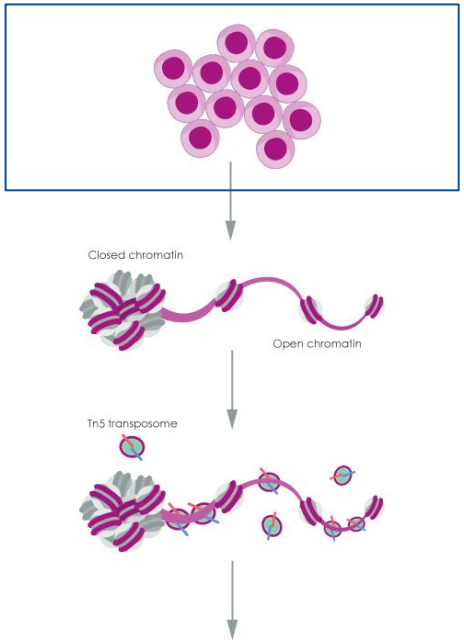
[1]Department of Genetics, University of Georgia, Athens, GA 30602, USA, [2]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA and [3]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Received August 05, 2016; Revised November 03, 2016; Editorial Decision November 11, 2016; Accepted November 15, 2016

| | |
|---|---|
| human | $3.3 * 10^9$ |
| mouse | $2.7 * 10^9$ |
| zebrafish | $1.5 * 10^9$ |
| fruit fly | $1.2 * 10^8$ |
| *A. thaliana* | $1.4 * 10^8$ |

Amount of nuclei depends on **genome** size.

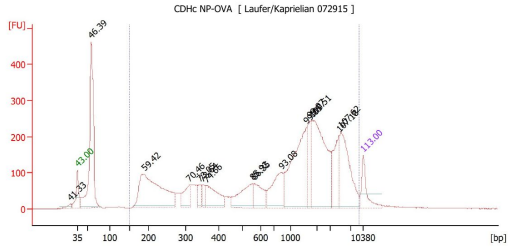
(500, 50k or more) // human and mouse



# ATAC-seq experimental procedure

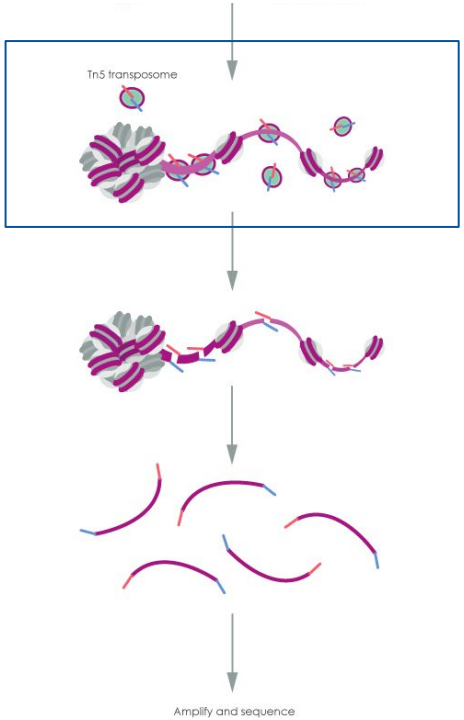
**Transposition reaction**

default: 30 min in 37℃




3. **Question: How do I know how many cells to add to the transposition reaction?**

**Answer:** Assuming cells are happy, the biggest source of failure comes from variations in cell number. We see biggest differences in the requirement of the number of cells between species; however, variation exists between cell types as well. If desired, a good way to troubleshoot or improve signal-to-noise for your particular application is to do a titration of cells, and if your cheap like I am, I would scale the reaction down 10x and titrate using 5,000 cells and 5uL transposition reactions. When you find a sample that best matches the gel above, then simply scale up to the 50uL reaction.




Over-transposition:

- Increase number of nuclei
- Decrease enzyme volume (non linear)

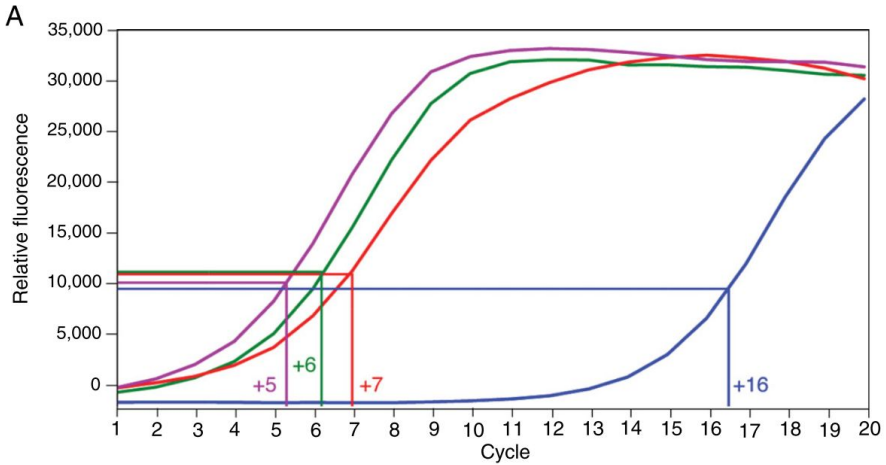
Under-transposition:

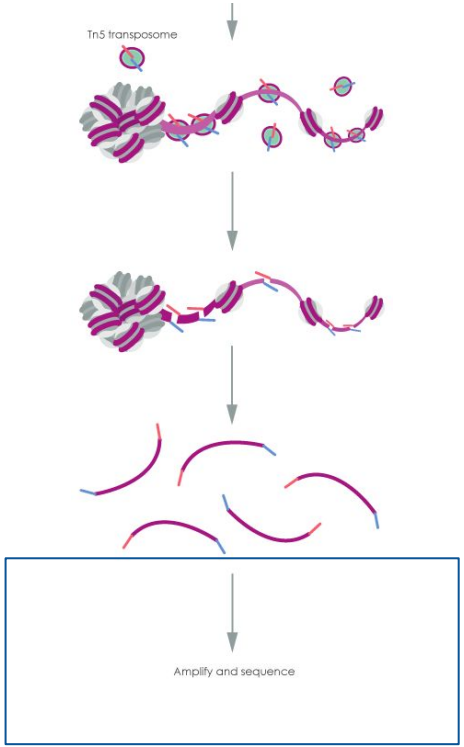
- Decrease number of nuclei
- Increase time of reaction

# ATAC-seq experimental procedure

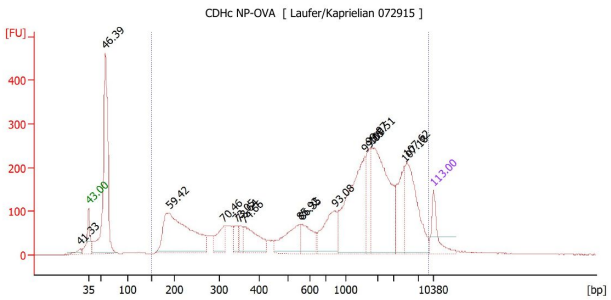
Perform PCR - 5 cycles -> take aliquot and do qPCR to calculate how many additional cycles need to be run.

# of additional cycles: ⅓ of the max fluorescence intensity



# ATAC-seq experimental procedure



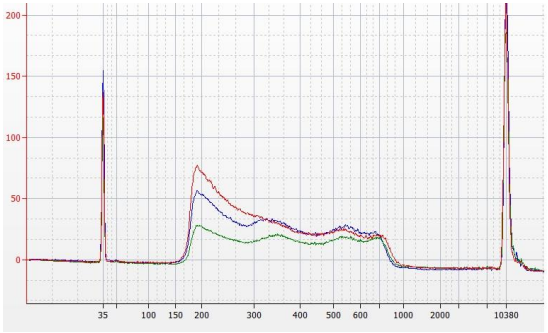
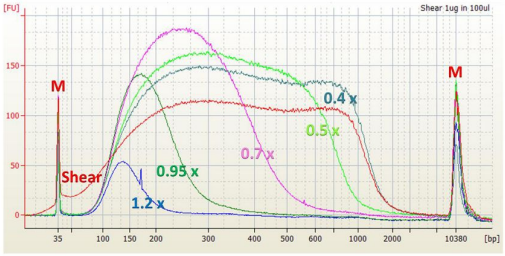




**Figure 3** Agilent High Sensitivity DNA chip Electropherogram.



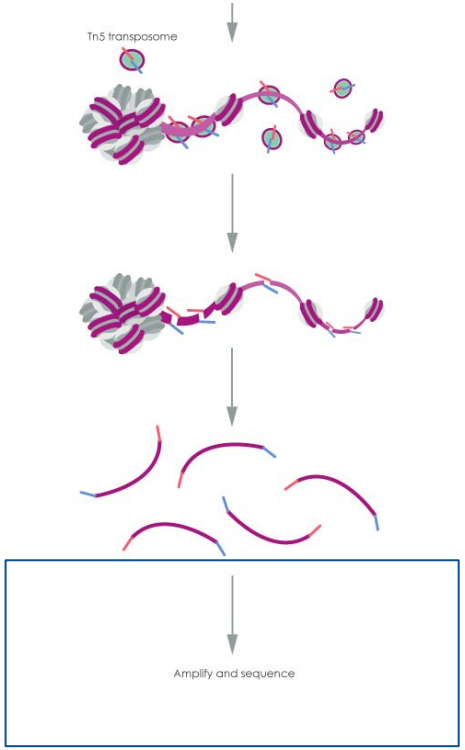

M = upper and lower markers for High Sensitivity DNA chip.

Shear = 1 µL of 20 ng/µL input control sample in water.

1.2x to 0.4x = 1 µL of shear, size selected with given ratio of **SPRIselect** volume to sample volume.

*SPRIselect User Guide, Beckman Coulter*

**size selection:** double sided / right side beads selection



# Experimental design

Control:

experiment dependent, no need for “input”

Replicates:

experiment dependent, at least **two** biological replicates - if there is high variability among samples I would recommend more

Library type:

paired-end, with single-end only some analysis can be performed

Sequencing depth:

depends on the genome size, assuming 70% mappability ratio to satisfy encode standard 70 milion reads would be needed

# Analysis workflow

| | |
|---|---|
| **Reads processing**<br>quality assessment<br>filtering and trimming if necessary | fastqc with multiqc<br>trim galore |
| **Alignment** | bowtie / bwa mem |
| **Alignment processing**<br>quality filtering<br>filtering out blacklisted regions<br>shifting alignments | samtools<br>bedtools<br>R ATACseqQC |
| **Peak calling** | macs2 |

# Before you start

*There are two types of people: those who backup, and … those who will backup.*

Keep raw, unprocessed data until your experiment is safely deposed in the database (ENA, GenBank or DDBJ).

# Reads processing - fastq file format

@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<????#=

sequence identifier (always starts with @)
sequence [ATCG]
quality score identifier
quality score

Sequence identifier always starts with @ but how it is constructed depends on the source of files (sequencing platform or database)

| | |
|---|---|
| <is filtered> | N - means that read passed the filtering<br>Y - means that read did not pass Illumina filtering |

# Example of reads filtering

```
zcat sample.raw.fastq.gz
grep -A 3 '^@.*[^:]*:N:[0-9]*:'
grep -v '^\-\-$'
gzip
> sample.filtered.fastq.gz
```

- ➔ Open and print to stdout reads in gzipped fastq files
- ➔ Find all lines having the desired pattern (:N:) and print that line and 3 following
- ➔ Don't print lines with only -
- ➔ Gzip output
- ➔ Save output to this file

```
zcat sample.raw.fastq.gz  | grep -A 3 '^@.*[^:]*:N:[0-9]*:' | grep -v '^\-\-$' |
gzip > sample.filtered.fastq.gz
```

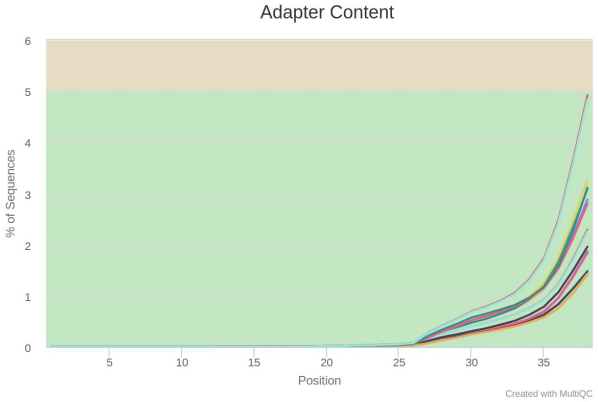
# Reads processing

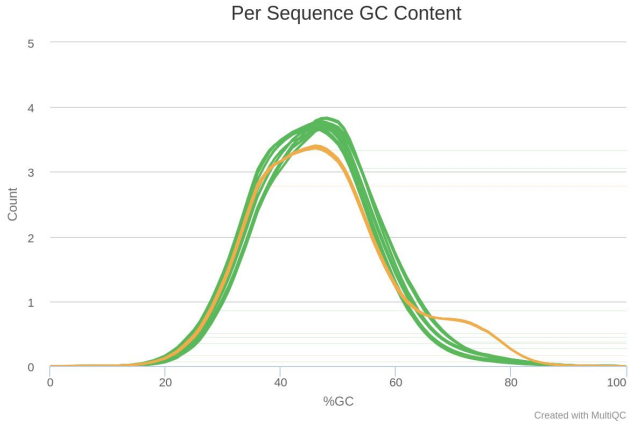
MultiQC

Aggregate results from bioinformatics analyses across many samples into a single report

MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools.

http://multiqc.info/




***Trim Galore!***
wrapper, cutadapt + fastQC

trims poor quality 3' bases, trims adapters (either specified or found in first 1 milion sequences) and discards too short reads

https://github.com/FelixKrueger/TrimGalore

# Alignment

## Aligning short reads to reference genome

## Bowtie2

```
bowtie2 \
        -x ${reference_genome} \
        -1 <(zcat ./cleaned/${sample}_R1.fastq.gz) \
        -2 <(zcat ./cleaned/${sample}_R2.fastq.gz)  \
        -p ${n_threads} \
        --very-sensitive \
        -X 2000
```

## bwa mem (or bwa aln)

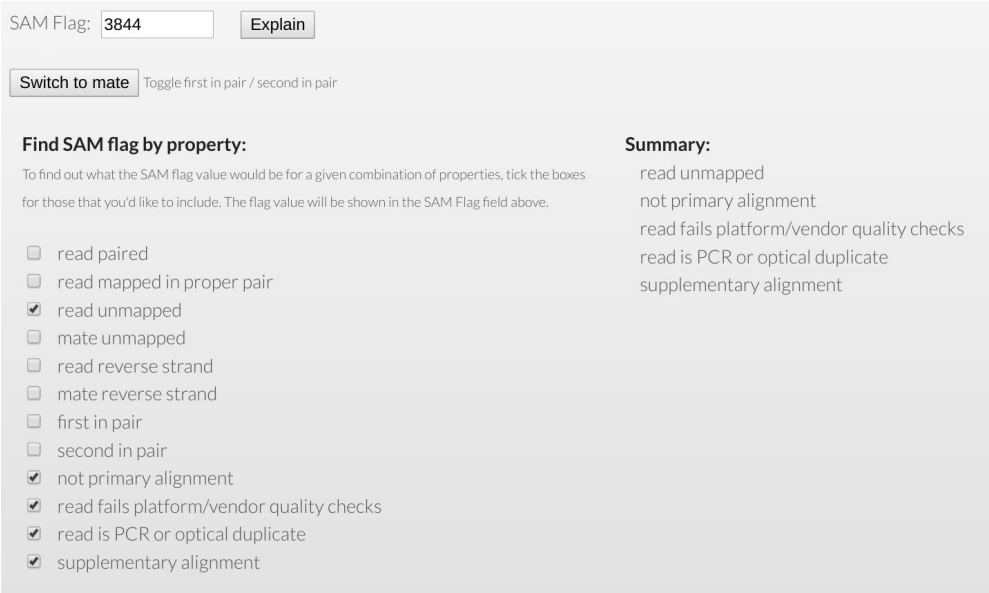
```
bwa mem \
        -v 3 \
        -t ${n_threads} ${reference_genome} ./cleaned/${sample}_R1.fq.gz ./cleaned/${sample}_R2.fq.gz 2> ./bwa/${sample].log |
               samtools view -b -@ ${compression_threads} -o ./bwa/${sample}_raw.bam
```

# Alignment - SAM

```
@HD     VN:1.3  SO:coordinate
@SQ     SN:1      LN:195471971
@SQ     SN:10   LN:130694993
@SQ     SN:11   LN:122082543
@SQ     SN:12   LN:120129022
@RG     ID:id   SM:sample LB:lib
@PG     ID:bwa  PN:bwa  VN:0.7.15-r1142-dirty   CL:bwa mem -v 3 -t 16 -R @RG\tID:id\tSM:sample\tLB:lib ./ref/Mdna.toplevel.fa ./sample_R1_trimmed.fq.gz
./sample_R2_trimmed.fq.gz
HWI-ST1309F:284:C8KYNANXX:2:2110:2778:8477   99          1          3000081 40          35M       =          3000141 79
CCCATCTGGTCCTGGGCTTTTTTTTTTTTTTTTT              BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFF        NM:i:0  MD:Z:35 AS:i:35 XS:i:35
```

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | `[!-?A-~]{1,254}` | Query template NAME |
| 2 | FLAG | Int | $[0, 2^{16}-1]$ | bitwise FLAG |
| 3 | RNAME | String | `\*\|[!-()+-<>-~][!-~]*` | Reference sequence NAME |
| 4 | POS | Int | $[0, 2^{31}-1]$ | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | $[0, 2^{8}-1]$ | MAPping Quality |
| 6 | CIGAR | String | `\*\|([0-9]+[MIDNSHPX=])+` | CIGAR string |
| 7 | RNEXT | String | `\*\|=\|[!-()+-<>-~][!-~]*` | Ref. name of the mate/next read |
| 8 | PNEXT | Int | $[0, 2^{31}-1]$ | Position of the mate/next read |
| 9 | TLEN | Int | $[-2^{31}+1, 2^{31}-1]$ | observed Template LENgth |
| 10 | SEQ | String | `\*\|[A-Za-z=.]+` | segment SEQuence |
| 11 | QUAL | String | `[!-~]+` | ASCII of Phred-scaled base QUALity+33 |

# Alignment - Flags




https://broadinstitute.github.io/picard/explain-flags.html

**samtools view**

-q INT Skip alignments with MAPQ smaller than INT [0].

-f INT Only output alignments with **all bits** set in INT present in the FLAG field.

-F INT **Do not** output alignments with **any bits** set in INT present in the FLAG field.

-G INT **Do not** output alignments with **all bits** set in INT present in the FLAG field. This is the opposite of -f such that -f12 -G12 is the same as no filtering at all.

# Alignment processing

Blacklisted regions - regions having high signal / read counts independent of cell line or experiment type.
Available for human, mouse, worm and fruit fly.
https://sites.google.com/site/anshulkundaje/projects/blacklists

```
#n - number of chromosomes in a given organism
chromosomes=$(echo $(for i in {1..n}; do echo "chr"$i; done | xargs) "chrX" "chrY");

bedtools intersect -v -abam ./bwa/${sample}.bam -b ${blacklisted_regions} |
     samtools view -h -b -F 3844 -f 2 -q 5 ${sample}_filtered.bam ${chromosomes} |
          samtools sort -n -T ${sample}_tmp -o ${sample}_sorted.bam -@ ${n_threads};
```

# Alignment processing - quality check

## Collect statistics

```
samtools flagstat ${sample}.bam > ${sample}.txt
```

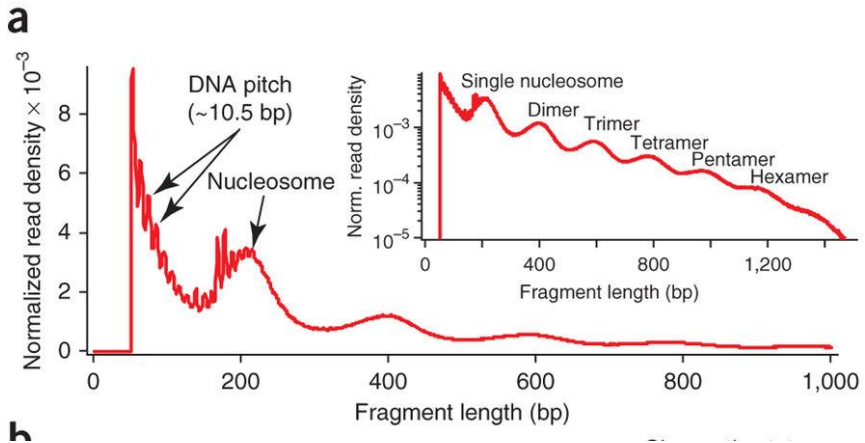
## ENCODE standards:

at least 25 milion non-duplicate, non-mitochondrial reads

more than 80%, preferable more than 95% mapped reads

Non-Redundant Fraction (i.e. # of non-duplicate reads / total # of reads) > 0.9

# Alignment processing - quality check

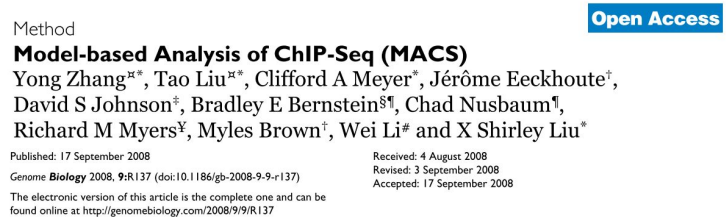



Check fragment size distribution.

Experimental design (size selection)

# Peak calling - MACS2

```
macs2 callpeak \
        --verbose 3 \
        --treatment ${sample}_sorted.bam \
        -g hg \
        -B \
        -q 0.05 \
        --extsize 200 \
        --nomodel \
        --shift -100 \
        --nolambda \
        --keep-dup all \
        -f BAM \
        --outdir ./peaks/${sample}\
        --call-summits
```



Method

Open Access

**Model-based Analysis of ChIP-Seq (MACS)**

Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nusbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

Published: 17 September 2008

*Genome Biology* 2008, **9**:R137 (doi:10.1186/gb-2008-9-9-r137)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/9/R137

Received: 4 August 2008
Revised: 3 September 2008
Accepted: 17 September 2008

**From the MACS2 manual**

Here are some examples for combining --shift and --extsize:

1. To find enriched cutting sites such as some DNAse-Seq datasets. In this case, all 5' ends of sequenced reads should be extended in both direction to smooth the pileup signals. If the wanted smoothing window is 200bps, then use '--nomodel --shift -100 --extsize 200'.

2. For certain nucleosome-seq data, we need to pileup the centers of nucleosomes using a half-nucleosome size for wavelet analysis (e.g. NPS algorithm). Since the DNA wrapped on nucleosome is about 147bps, this option can be used: '--nomodel --shift 37 --extsize 73'.

README for MACS: https://github.com/taoliu/MACS

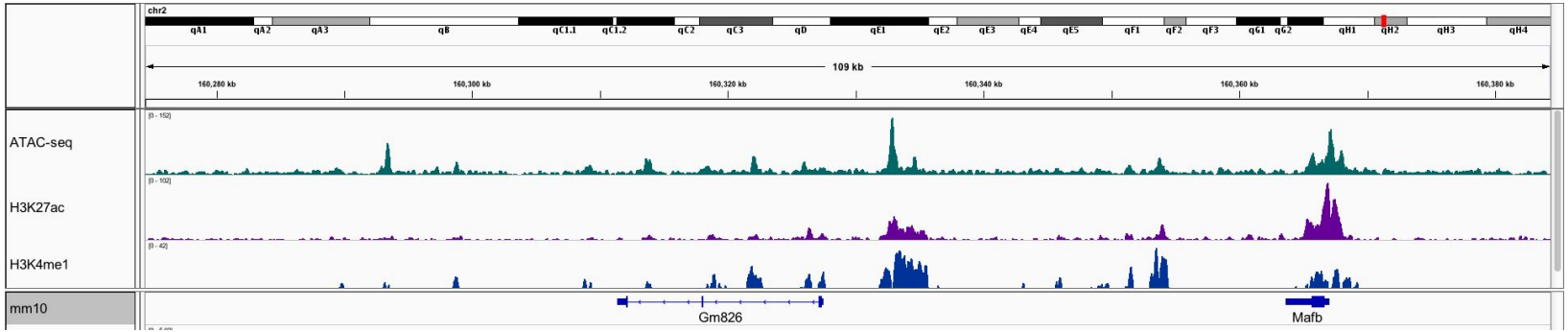
# Peak calling - file formats

**BED (Browser Extensible Data)** - 3 columns (chrom, chromStart, chromEnd) required
**BED6, BED6+4** - BED files with additional columns
**bedGraph** - version of BED file used for visualisation

MACS2 outputs:

1. **NAME_peaks.xls** - header with run description; chrom, chromStart, chromEnd, length, summitPosition (absolute), pileup (at summit), -log10(pvalue), fold_enrichment, -log10(qvalue), name
2. **NAME_peaks.narrowPeak** (BED6+4) - chrom, chromStart, chromEnd, name, score, strand, integer score,  fold-change, -log10pvalue, -log10qvalue, summitPosition (from peak start)
3. **NAME_summits.bed** (BED) - location of summits
4. **NAME_treat_pileup.bdg** (bedGraph) - chrom, chromStart, chromEnd, signal

# Peak calling - quality control

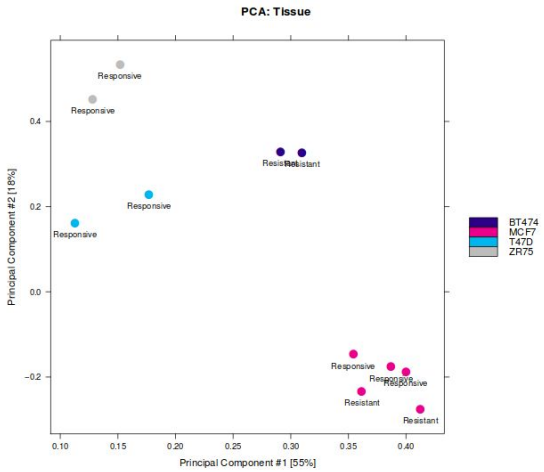



**Fraction of reads in peaks (FRiP)** - Fraction of all mapped reads that fall into the called peak regions, i.e. usable reads in significantly enriched peaks divided by all usable reads.

FRiP should be >0.3, though values greater than 0.2 are acceptable

# Peak calling - quality control

**Visualize for example by PCA plot**




http://setosa.io/ev/principal-component-analysis/

**Transcription Start Site (TSS) Enrichment Score**

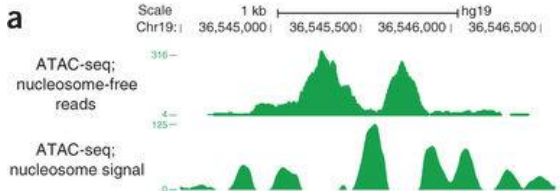
| Annotation used | Value | Resulting Data Status |
|---|---|---|
| **hg19 Refseq TSS annotation** | <6 | Concerning |
| | 6-10 | Acceptable |
| | >10 | Ideal |
| **GRCh38 Refseq TSS annotation** | <5 | Concerning |
| | 5-7 | Acceptable |
| | >7 | Ideal |
| **mm9 GENCODE TSS annotation** | <5 | Concerning |
| | 5-7 | Acceptable |
| | >7 | Ideal |
| **mm10 Refseq TSS annotation** | <10 | Concerning |
| | 10-15 | Acceptable |
| | >15 | Ideal |

# Accessible chromatin ≠ open chromatin

**Open chromatin** - can be defined as nucleosome free region

**Accessible chromatin** - regions of chromatin that are accessible for transposase

It all depends on what you need and how you call peaks!

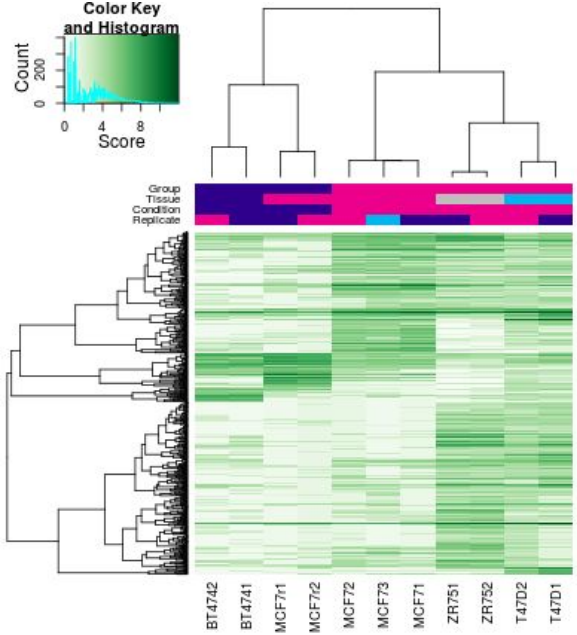


# Differentially accessible regions

Task: identify the sites that are accessible in one, but not the other sample

Diffbind package in R based on DESeq2

1. Scan peaksets and merge them creating consensus.
2. Create matrix of counts, peaks x samples.
3. Calculate the library size, normalize.
4. Apply statistical tests to assess which sites are differentially open.



# Gene set enrichment

**Enrichr**

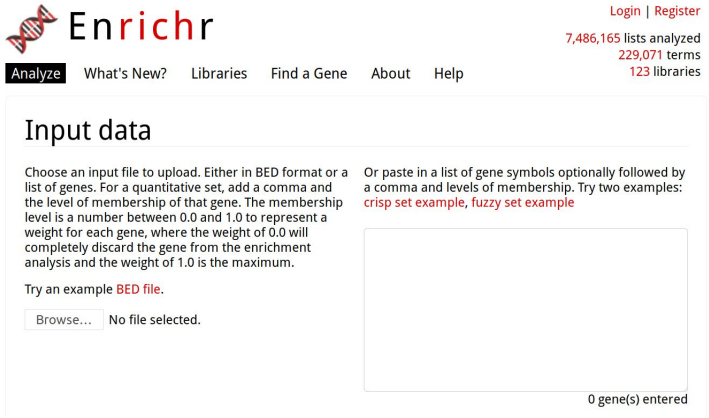
can use both gene symbols or BED file

http://amp.pharm.mssm.edu/Enrichr/

**GREAT**

“GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes.”

http://bejerano.stanford.edu/great/public/html/





GREAT improves functional interpretation of *cis*-regulatory regions

Cory Y McLean[1], Dave Bristor[1,2], Michael Hiller[2], Shoa L Clarke[3], Bruce T Schaar[2], Craig B Lowe[4], Aaron M Wenger[1] & Gill Bejerano[1,2]

NATURE BIOTECHNOLOGY VOLUME 28 NUMBER 5 MAY 2010 495

# Motif search

Identify transcription factors bound to the chromatin

**HOMER (v4.9, 2-20-2017)**

Software for motif discovery and next generation sequencing analysis

**The MEME Suite**

**Motif-based sequence analysis tools**

# Footprinting

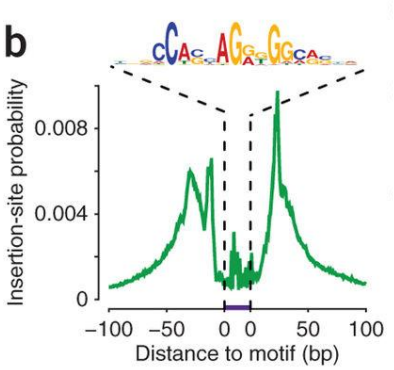
CENTIPEDE

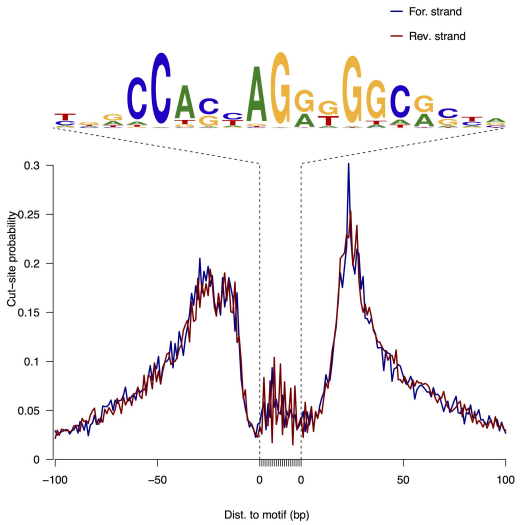
integrates histone modifications or DNase I cleavage patterns with genomic information such as gene annotation and evolutionary conservation to generate genome-wide map of transcription factor binding sites

*Pique-Regi, R., Degner, J., & Pai, A. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research, 3, 447–455. https://doi.org/10.1101/gr.112623.110.Freely*

ATACseqQC

doesn't use the conservation (PhyloP)



# Workshop outline

Workshop will cover:

1. Shifting and splitting the reads with R package ATACseqQC
2. Calling peaks with MACS2
3. Motif search with HOMER
4. Identifying differentially bound sites with R package Diffbind
5. Basic enrichment analysis with GREAT and Enrichr
6. Transcription Factor footprinting with R package ATACseqQC

# Software

| Reads processing | Alignment | Alignment processing | Peak calling |
|---|---|---|---|
| cutadapt<br>Trim Galore!<br>Trimmomatic | bwa<br>bowtie2 | Samtools<br>bedtools<br>Picard<br>Sambamba | MACS2<br>Hotspot<br>Homer<br>ZINBA |

MEME suite

CENTIPEDE - tutorial  msCENTIPEDE  R ATACseqQC  pyDNase

NucleoATAC  chromVAR

R DiffBind  R ChIPseeker  R ChIPQC

grep, awk

# Summary

1. Optimize the procedure and analysis for a given experiment.
2. Design the experiment.
   a. Choose proper controls;
   b. Consider tissue and sample type;
   c. Set your goals.
3. Keep unprocessed data until you depose it in a database.
4. Carefully read software documentation before using it.
5. Do quality checks and follow guidelines.
   a. Check raw and processed (filtering, trimming) reads;
   b. Filter alignment (blacklisted, uncanonical, low quality);
   c. Check fragment size distribution, mappability ratio, NRF;
   d. Calculate FRiP and TSS enrichment.

# Acknowledgments

Aakrosh Ratan

Center for Public Health Genomics

University of Virginia