

# ChIP-seq module

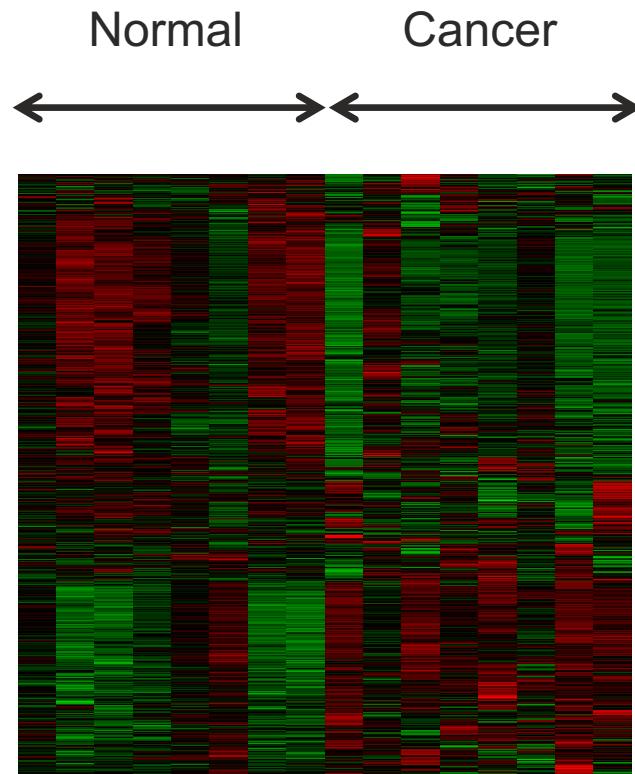
BD2K Course 2018

Y. Ada Zhan



# Introduction

# Learn from RNA-Seq



Differential Gene Expression

*What may cause the differences?*

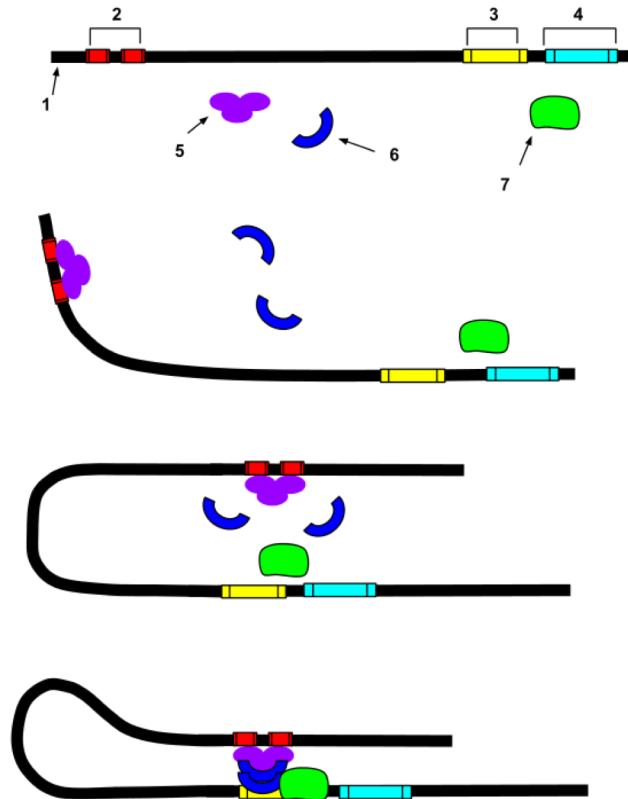


Steve Munger, 2017

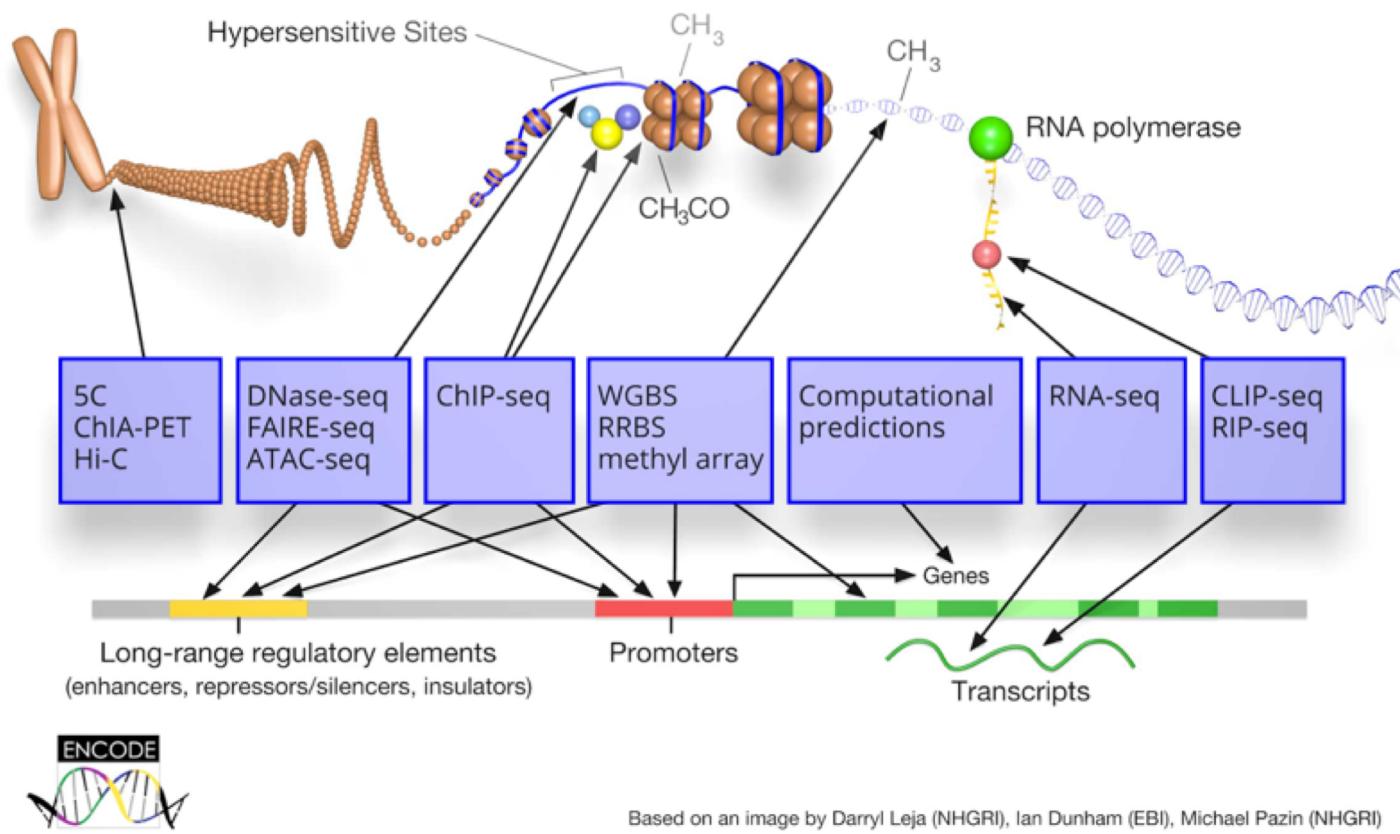
THE JACKSON LABORATORY

# What can go wrong or be different?

1. DNA
2. Enhancer
3. Promoter
4. Gene
5. Transcription Activator Protein
6. Mediator Protein
7. RNA Polymerase

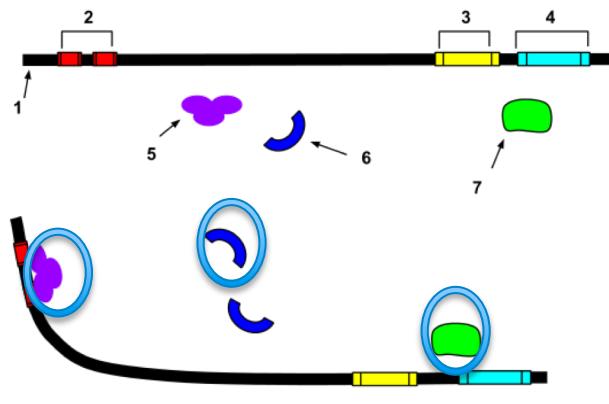


- Structure variation (mutation, deletion, translocation etc. in DNA sequences)
- Chemical modifications on DNA (methylation, acetylation, etc.)
- Chemical modifications on histone



# ChIP-seq

## Chromatin ImmunoPrecipitation followed by Sequencing



- Work on proteins that interact with DNA
- Get the DNA sequences that are bound by the protein of interest
- Explore how genes are regulated

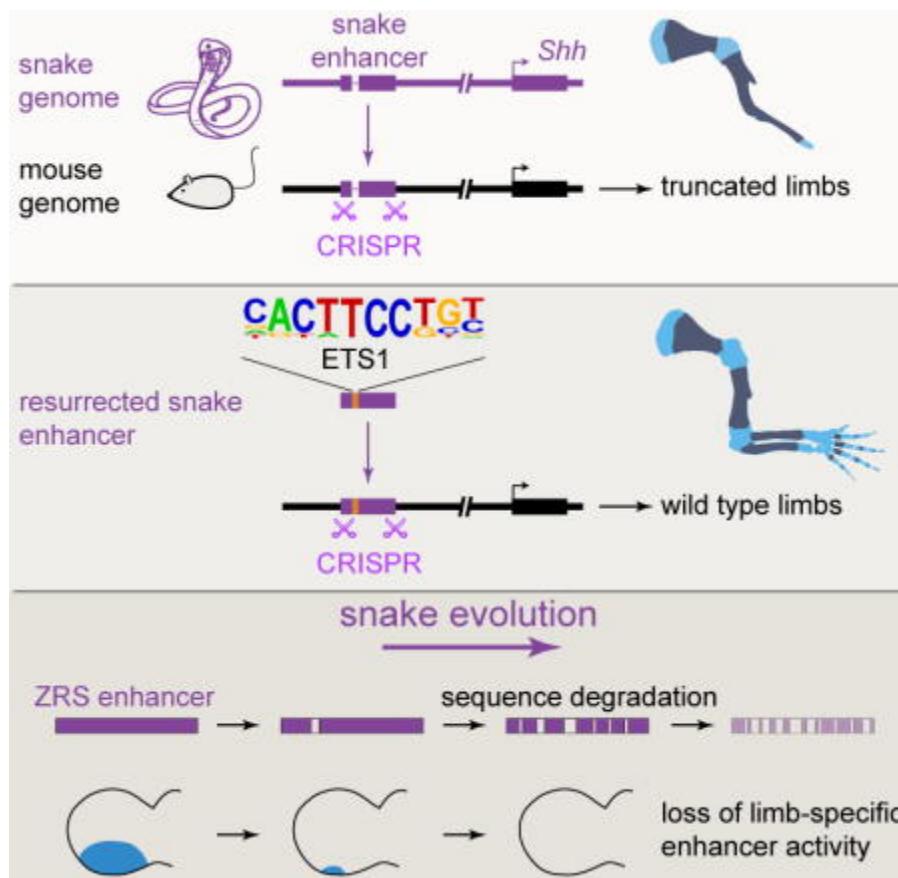
### Applications:

To detect interesting regions in genome at base-pair resolution for

- DNA-binding proteins, e.g. transcription factor, RNAPII, etc
- Histone modifications → status of enhancers



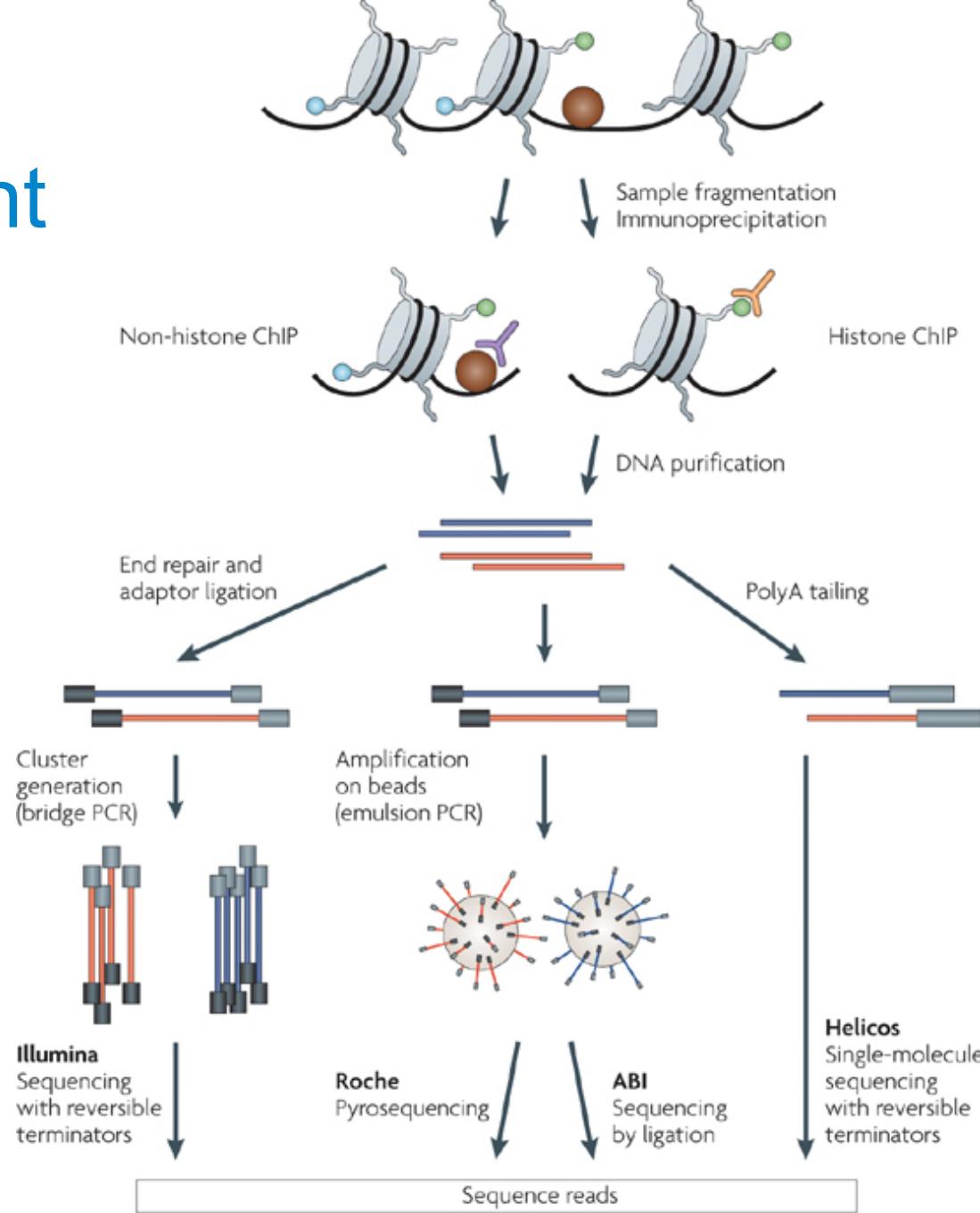
# Example on enhancer



- ChIP-Seq can help find enhancers
- Other techniques, like eQTL, Hi-C or ChIA-PET, are required to identify the interactions between particular enhancers and genes.



# ChIP-seq Experiment

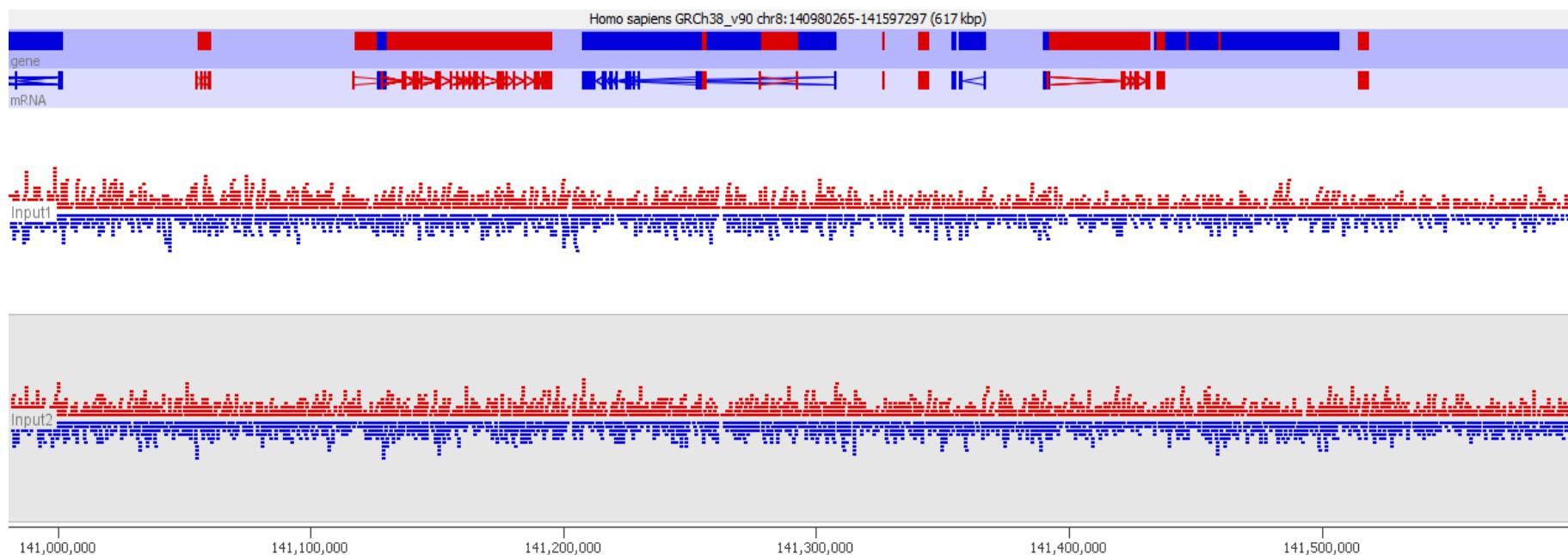


# Experimental Considerations

- Antibodies: High sensitivity and specificity
- Cell number: More cells, better signal-to-noise (typically  $10^6$  to  $10^7$  cells)
- Controls: To remove background noise
  - IgG (Reads are only informative if the ChIP hasn't worked.)
  - Input Chromatin (sonicated / Mnase etc)
    - Genomic library - everywhere equally
    - Technical issues can cause variation
- Replicates:

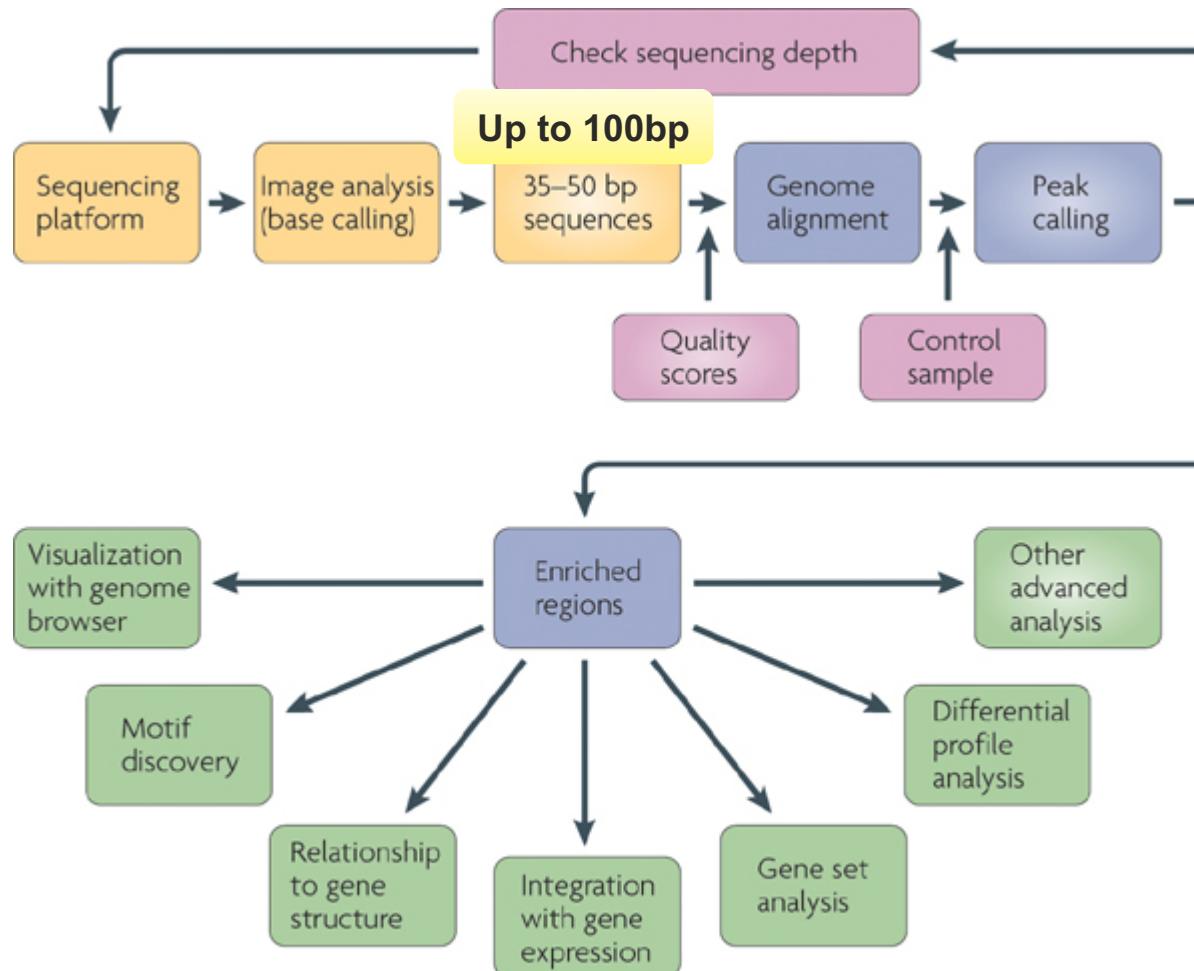


# Examine controls

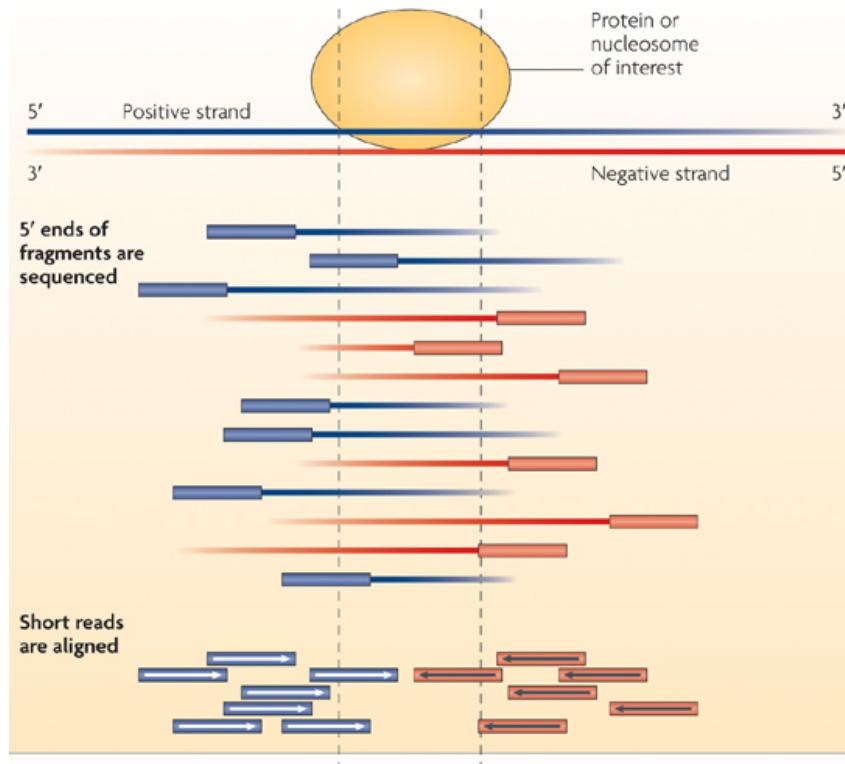


- Does the coverage look even
- If there are multiple inputs to do they look similar

# Overview of ChIP-seq analysis



# Sequence Alignment



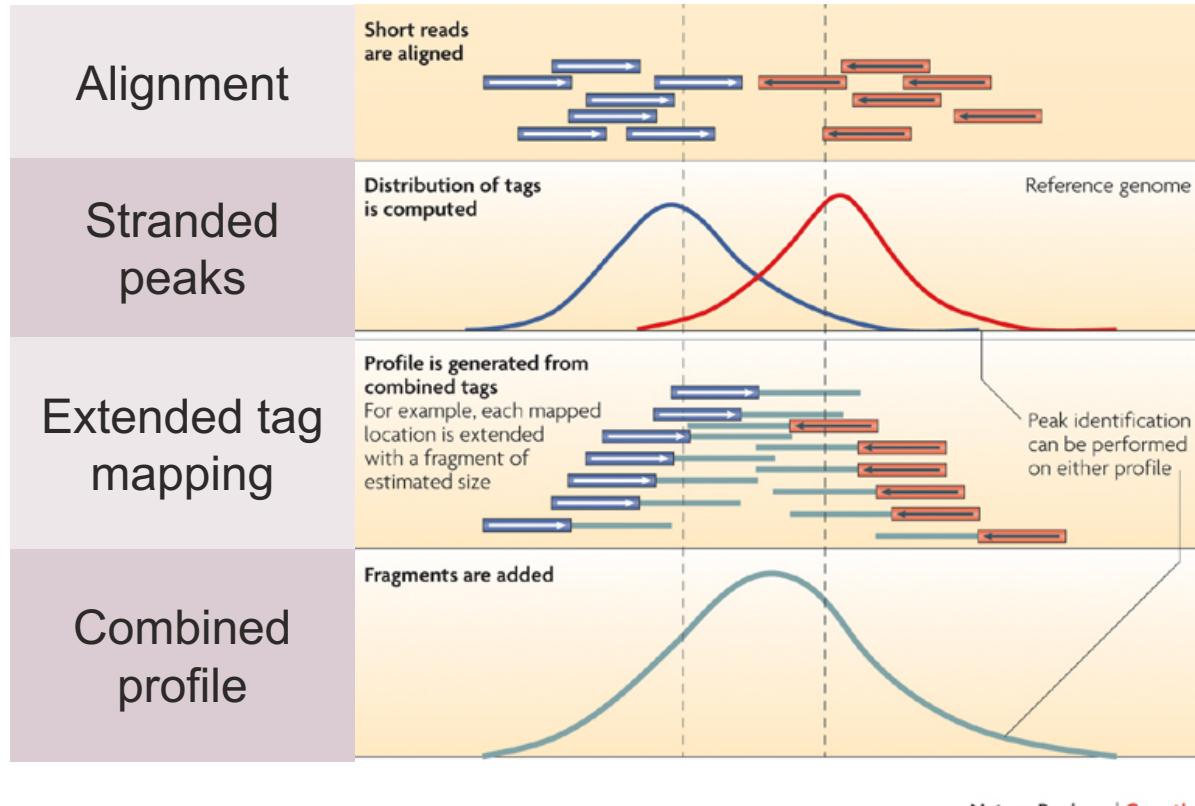
# Sequence Alignment

- Bowtie (35 – 50 bp)
- Bowtie2 (50 – 100 bp, no upper limit actually)
- BWA (different algorithms for shorter and longer sequences than 70bp)
- Others: MAQ, ELAND, ...



# Peak calling

Identify enriched regions above controls



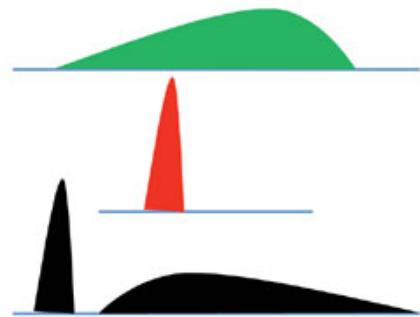
Nature Reviews | Genetics



THE JACKSON LABORATORY

# Peak calling

## Software



Peak calling (choose the right tool)

Type of peak	Example	Representative tools
Broad	H3K27me3	CCAT, SICER
Sharp	CTCF	MACS
Sharp & broad	Pol II	ZINBA



# Peak calling

## Software used in Encode pipeline

- **SPP**

A ChIP-seq peak calling algorithm, implemented as an R package, that accounts for the offset in forward-strand and reverse-strand reads to improve resolution, compares enrichment in signal to background or control experiments, and can also estimate whether the available number of reads is sufficient to achieve saturation, meaning that additional reads would not allow identification of additional peaks. SPP will be used in the ENCODE 3 uniform peak calling pipeline.

Kharchenko PV, Tolstorukov MY, Park PJ. [Design and analysis of ChIP-seq experiments for DNA-binding proteins](#). *Nat Biotechnol*. 2008 Dec;26(12):1351-9. PMID: [19029915](#); PMC: [PMC2597701](#)

- **GEM**

GEM is a Java software package for analyzing genome wide ChIP-seq/ChIP-exo data. GEM can decompose single observed peaks into multiple binding events, determine binding event location at high spatial resolution, and discover explanatory DNA sequence motifs with an integrated model of ChIP reads and proximal DNA sequences. GEM is able to process single-end or paired-end data and can be run in single-condition mode or multi-condition mode. GEM will be used in the ENCODE 3 uniform peak calling pipeline.

Guo Y, Mahony S, Gifford DK. [High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints](#). *PLoS Comput Biol*. 2012;8(8):e1002638. PMID: [22912568](#); PMC: [PMC3415389](#)

- **PeakSeq**

Identifies enriched regions in ChIP-seq type experiments and explicitly compares signal experiments to control experiments. PeakSeq will be used in the ENCODE 3 uniform peak calling pipeline.

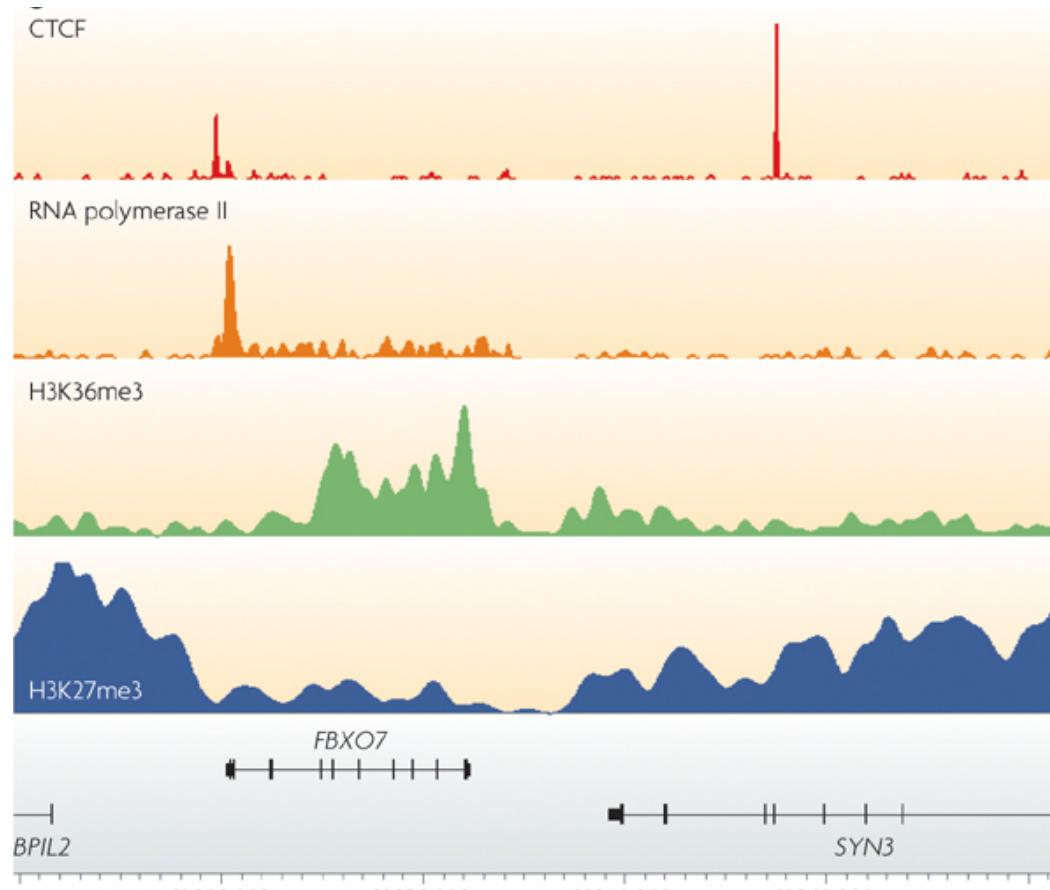
Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. [PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls](#). *Nat Biotechnol*. 2009 Jan;27(1):66-75. PMID: [19122651](#); PMC: [PMC2924752](#)

- **MACS**

A widely-used, fast, robust ChIP-seq peak-finding algorithm that accounts for the offset in forward-strand and reverse-strand reads to improve resolution and uses a dynamic Poisson distribution to effectively capture local biases in the genome. MACS 1.4 was used in



# ChIP-seq profile



Nature Reviews | Genetics



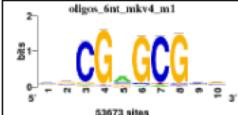
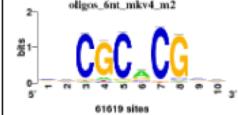
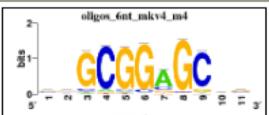
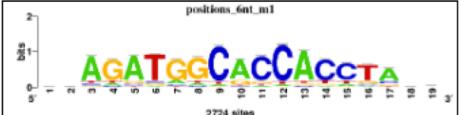
THE JACKSON LABORATORY

# After peak calling

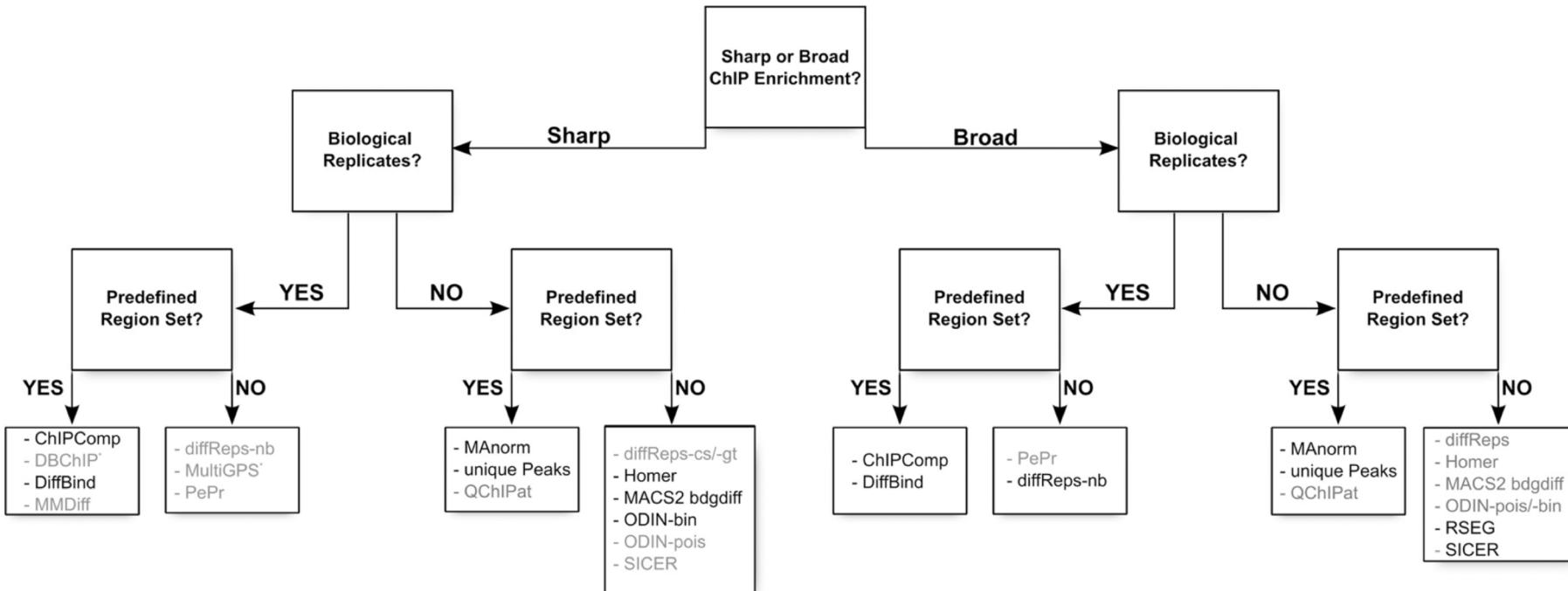
- Motif analysis (MEME, HOMER, ...)
- Integrative analysis with expression data
- Differential binding
- Peak annotation
- ...



# Motif analysis

Motif	Logo	3 Top hits in databases
oligos_6nt_mkv4_m1		<u>versus Homer:</u> homer_326.ZNF519, <u>versus jaspar_core_nonredundant_vertebra</u> no match
oligos_6nt_mkv4_m2		<u>versus Homer:</u> homer_123.RTACGTGC, homer_124.HIF2a, homer_38.NCCACGTG, <u>versus jaspar_core_nonredundant_vertebra</u> ARNT::HIF1A, Ahr::Arnt, Hes1,
oligos_6nt_mkv4_m3		<u>versus Homer:</u> homer_25.BORIS, homer_266.Sp1, homer_145.KLF14, <u>versus jaspar_core_nonredundant_vertebra</u> EGR1, SP2, KLF16,
oligos_6nt_mkv4_m4		<u>versus Homer:</u> homer_266.Sp1, <u>versus jaspar_core_nonredundant_vertebra</u> SP1,
oligos_6nt_mkv4_m5		<u>versus Homer:</u> homer_266.Sp1, <u>versus jaspar_core_nonredundant_vertebra</u> no match
positions_6nt_m1		<u>versus Homer:</u> no match <u>versus jaspar_core_nonredundant_vertebra</u> no match

# Differential binding

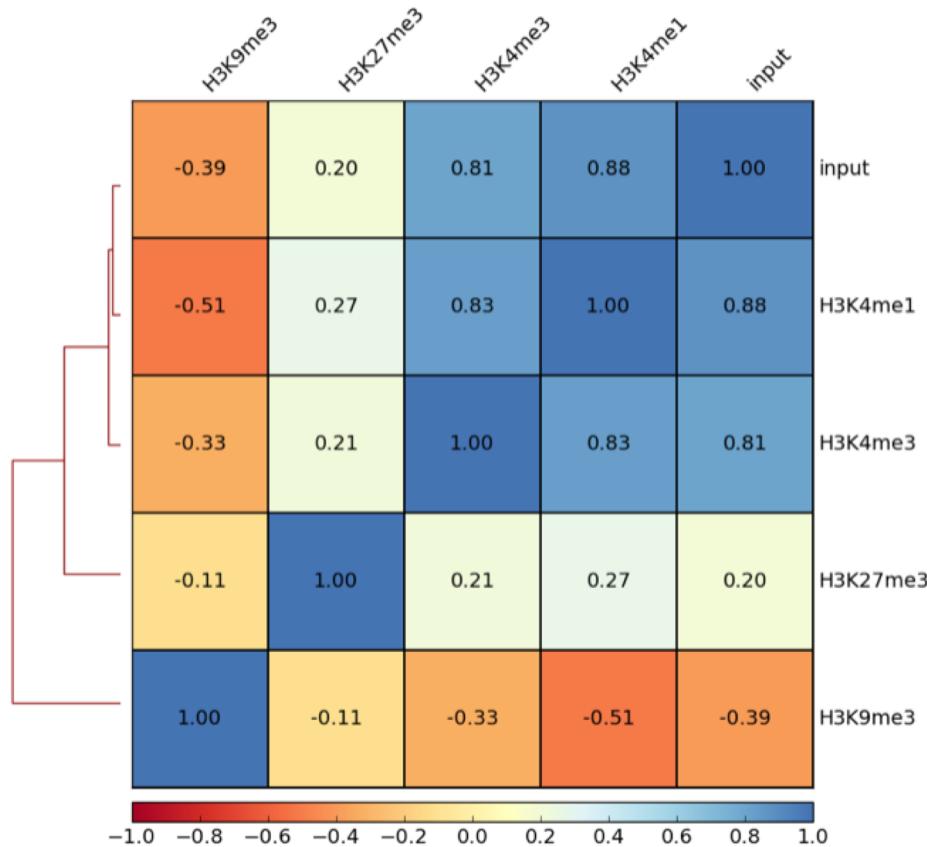


From: A comprehensive comparison of tools for differential ChIP-seq analysis

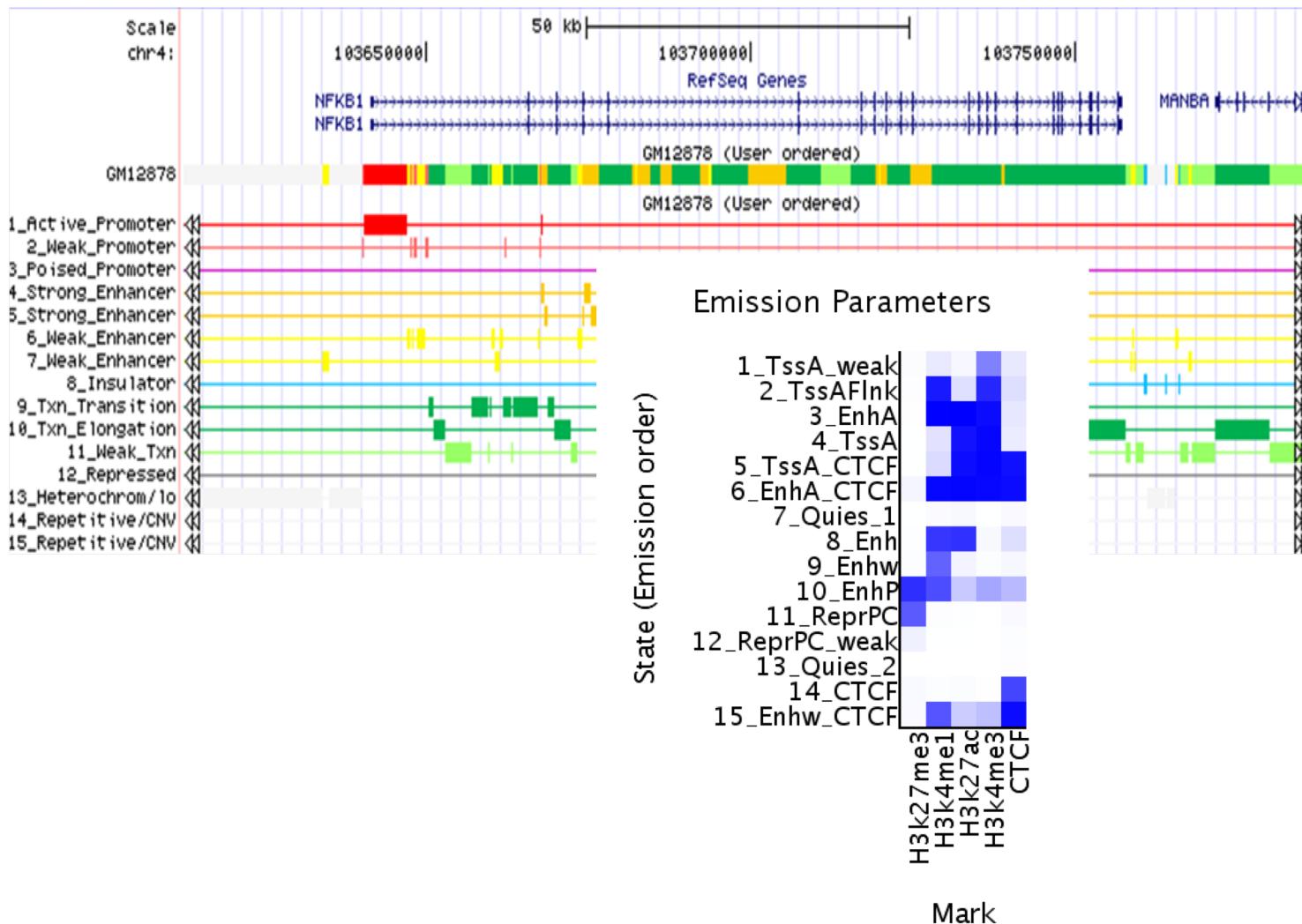
Brief Bioinform. 2016;17(6):953-966. doi:10.1093/bib/bbv110

Brief Bioinform | © The Author 2016. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

# Correlation between different binding factors



# Chromatin state



# ChIP-Seq on-line analysis

  ChIP-Seq  
On-line Analysis Tools

To get all news messages from us [Follow @EPD\\_SIB](#)

Computational Cancer Genomics | ExPASy | EPFL | Home Page

**ChIP-Seq Tools**

- ChIP-Cor
- ChIP-Extract
- ChIP-Peak
- ChIP-Part
- ChIP-Center
- ChIP-Track
- ChIP-Convert

**MGA Database**

- MGA-Search
- MGA Data Overview
- MGA FTP-Site
- Genome Assembly Table

**Other Resources**

- EPD
- SSA
- PWMScan

**Documentation**

- Tutorials
- General Documentation

**References**

**Frequently Asked Questions**

News: 2018-10-08 -- New Software release (1.5.4) [more](#)

The ChIP-Seq Web Server provides access to a set of useful tools performing common ChIP-Seq data analysis tasks, including positional correlation analysis, peak detection, and genome partitioning into signal-rich and signal-poor regions. Users can analyse their own data by uploading mapped sequence tags in various formats, including BED and BAM. The server also provides access to hundreds of publicly available data sets such as ChIP-seq data, RNA-seq data (i.e. CAGE), DNA-methylation data, sequence annotations (promoters, polyA-sites, etc.), and sequence-derived features (CpG, phastCons scores).

The source code is available on [sourceforge](#)

**The ChIP-Seq Tools**

- [ChIP-Cor](#): Generation of an aggregation plot (feature correlation plot) for specific genomic features.
- [ChIP-Extract](#): Extraction of specific genome annotation features around reference genomic anchor points. The output is a table with rows representing each reference anchor point and columns the feature tag occurrence at specific distances. This table can be used to generate heatmaps.
- [ChIP-Center](#): Read tag shifting to estimated center-positions of DNA fragments.
- [ChIP-Peak](#): Narrow peak caller that uses a fixed width peak size.
- [ChIP-Part](#): Broad peak caller algorithm used for broad regions of enrichment found in ChIP-seq experiments targeted at histone marks.
- [ChIP-Track](#): Generation of UCSC Genome Browser annotation tracks for data visualization.
- [ChIP-Convert](#): Format conversion tool for BED, BAM, SGA, and other data formats. Data export from our local data repository, the Mass Genome Annotation (MGA) database, is also available.

<http://rsat.ulb.ac.be/dpeaks/>



THE JACKSON LABORATORY

# For command line user – Anaconda and bioconda

- Anaconda (package management platform) <https://anaconda.org>
- Bioconda (package repository for bioinformatics) <https://bioconda.github.io>
- Deeptools <http://deeptools.readthedocs.io/en/develop/>



# Data from database

- Encode (<https://www.encodeproject.org>)
- modEncode  
(<http://www.modencode.org/publications/about/index.html>)
- Roadmap (<http://www.roadmapepigenomics.org>)
- ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>)

-----

- GENCODE ([human and mouse genome](#))



# Hands on Work



# What we will do?

Objective: Perform a basic ChIP-seq pipeline

- In terminal:
  - Quality check (FASTQC: check the quality of raw sequencing data in fastq format)
  - Mapping (bowtie: align sample sequences to the reference genome)
  - Peak calling (MACS2: locate immuno-enriched regions)
- On MEME online server
  - Motif analysis (MEME-ChIP: identify the binding motifs)



# Tools

Software	Version	Usage
fastqc	0.11.3	Quality checking
Bowtie	0.12.8	mapping
Samtools	1.3.1	SAM, BAM file manipulation
MACS2	2.0.10.20120913	Peak calling
bedtools	2.26.0	Bed file manipulation

Note:

- You may install the software through bioconda
- MACS2 requires python2.7.X
- Check software dependencies on their installation page on-line



# Datasets

- Assay: ChIP-seq
- Target: CTCF
- Biosample: Homo Sapiens GM12878 (A lymphoblastoid cell line produced from the blood of a female donor with northern and western European ancestry by EBV transformation)
- Sample type: DNA
- Chromosome: Chr1
- Lab: Bradley Bernstein, Broad
- Source: encodeproject.org
- Platform: Illumina Genome Analyzer
- ENCODE accession id:
  - ENCFF000ARP (sample) -> GM12878\_CTCF\_chr1.fastq
  - ENCFF000ARK (control) -> GM12878\_control\_chr1.fastq



# Getting help

- fastqc --help
- bowtie -h
- samtools or samtools <command>
- macs2 -h



- Check where you are:
  - pwd
- Find out what are in the directory:
  - ls
- Go to ‘ChIPseq’ folder:
  - cd ChIPseq
  - Check what's in the box: ls
  - Examine files: less filename or cat filename
  - Take a look at two files: readme.txt and workflow.sh
    - readme.txt : describes the data structure in this folder
    - workflow.sh : contains all the commands for this practice



# Check the sequencing quality

```
# Go the data folder if you are not there:  
cd ChIPseq
```

```
# Perform fastqc quality check:  
fastqc GM12878_control_chr1.fastq  
fastqc GM12878_CTCF_chr1.fastq
```

```
# Check output files  
ls  
(You will find .zip and .html files for each .fastq.  
Download .html files and open them in your  
browser)
```



# Index the genome

```
# To make the alignment more memory efficient, bowtie  
will index the genome first.  
# The reference genome is the chromosome 1 of human  
genome build GRCh38
```

**bowtie-build hg38/GRCh38.chr1.fa hg38/GRCh38.chr1**



# Sequence alignment

```
# learn bowtie options  
bowtie -h
```

```
# map to reference genome  
bowtie -m 1 -S ./hg38/GRCh38.chr1 GM12878_control_chr1.fastq \  
> GM12878_control_chr1.sam
```

```
bowtie -m 1 -S ./hg38/GRCh38.chr1 GM12878_CTCF_chr1.fastq \  
> GM12878_CTCF_chr1.sam
```

```
# Take a look at the first 5 lines of the output  
head -n 5 GM12878_CTCF_chr1.sam
```



# Compress SAM

```
# Learn samtools view options  
samtools view
```

```
# Compress SAM to BAM file to save storage  
samtools view -bSo GM12878_control_chr1.bam \  
GM12878_control_chr1.sam  
samtools view -bSo GM12878_CTCF_chr1.bam \  
GM12878_CTCF_chr1.sam
```



# Index BAM

```
# Index the BAM files for IGV visualization in two steps

## sort the alignment according to chromosome position
samtools sort GM12878_control_chr1.bam \
GM12878_control_chr1.sorted
samtools sort GM12878_CTCF_chr1.bam \
GM12878_CTCF_chr1.sorted

## index the sorted files
samtools index GM12878_control_chr1.sorted.bam
samtools index GM12878_CTCF_chr1.sorted.bam
```

Note: make sure the sorted.bam and .bai index files in the same directory for IGV loading



# Peak calling

```
# Learn macs2  
macs2 -h
```

```
# Identify peak enrichment  
macs2 callpeak -t GM12878_CTCF_chr1.sorted.bam \  
-c GM12878_control_chr1.sorted.bam \  
-f BAM -g 175000000 -n GM12878_CTCF_chr1 -B -q 0.01
```

- The output files description can be found on MACS2 page :  
<https://github.com/taoliu/MACS>
- The R script is one output of macs2.
- The output .bed file can be loaded to IGV directly.

# Motif analysis

```
# summit files include the estimation of peak summit at single base  
level  
# extend summits 100bp on both directions  
bedtools slop -i GM12878_CTCF_chr1_summits.bed \  
-g hg38/GRCh38.chr1.size -b 100 \  
> GM12878_CTCF_chr1_summits_ext.bed  
  
# get sequences for the extended summit regions (i.e. fasta)  
bedtools getfasta -fi hg38/GRCh38.chr1.fa \  
-bed GM12878_CTCF_chr1_summits_ext.bed \  
-fo GM12878_CTCF_chr1_summits_ext.fa
```

# Motif analysis using MEME

Go to site: <http://meme-suite.org/tools/meme-chip>

The screenshot shows the MEME-ChIP web interface. On the left is a sidebar with a navigation menu:

- MEME Suite 4.11.4**
  - ▼ Motif Discovery**
    - MEME
    - DREME
    - MEME-ChIP
    - GLAM2
  - ▼ Motif Enrichment**
    - CentriMo
    - AME
    - SpaMo
    - GOMo
  - Motif Scanning**
  - Motif Comparison**
  - Manual**
  - Guides & Tutorials**
  - Sample Outputs**
  - File Format Reference**
  - Databases**

The main content area is titled "Data Submission Form". It contains the following sections:

- Select the motif discovery and enrichment mode**: A radio button group with "Normal mode" selected.
- Select the sequence alphabet**: A radio button group with "DNA, RNA or Protein" selected. A blue arrow points to the "Choose File" button next to the input field "no file selected".
- Input the primary sequences**: A section for entering nucleotide sequences. It includes a "Upload sequences" dropdown, a "Choose File" button (with "no file selected"), and a help icon.
- Input the motifs**: A section for selecting motifs. It includes dropdown menus for "Eukaryote DNA" (set to "DNA") and "Vertebrates (In vivo and in silico)".

