

# Analysis of Next-Generation Sequencing Data for Epigenetics

Chongzhi Zang

[zang@virginia.edu](mailto:zang@virginia.edu)

[zanglab.org](http://zanglab.org)

NYU: Epigenetics and Environmental Diseases – Spring 2021

April 20, 2021

# Outline

- 1st Half:
  - NGS introduction
  - NGS data analysis strategy
  - ChIP-seq data analysis
- 2nd Half:
  - Other NGS data analysis
  - Downstream analysis and integration
  - Online resources

# Learning Objectives

- Understand how NGS works and key QC measures
- Learn how ChIP-seq data analysis is done
- Understand general strategies of NGS data analysis and online resources

Microhabitats save mammals, but not birds, from warming pp. 553 & 633

Gut microbiota modulate immunotherapy pp. 573, 595, & 602

Physically distanced quantum gates pp. 576 & 604

# Science

\$15  
5 FEBRUARY 2021  
sciencemag.org

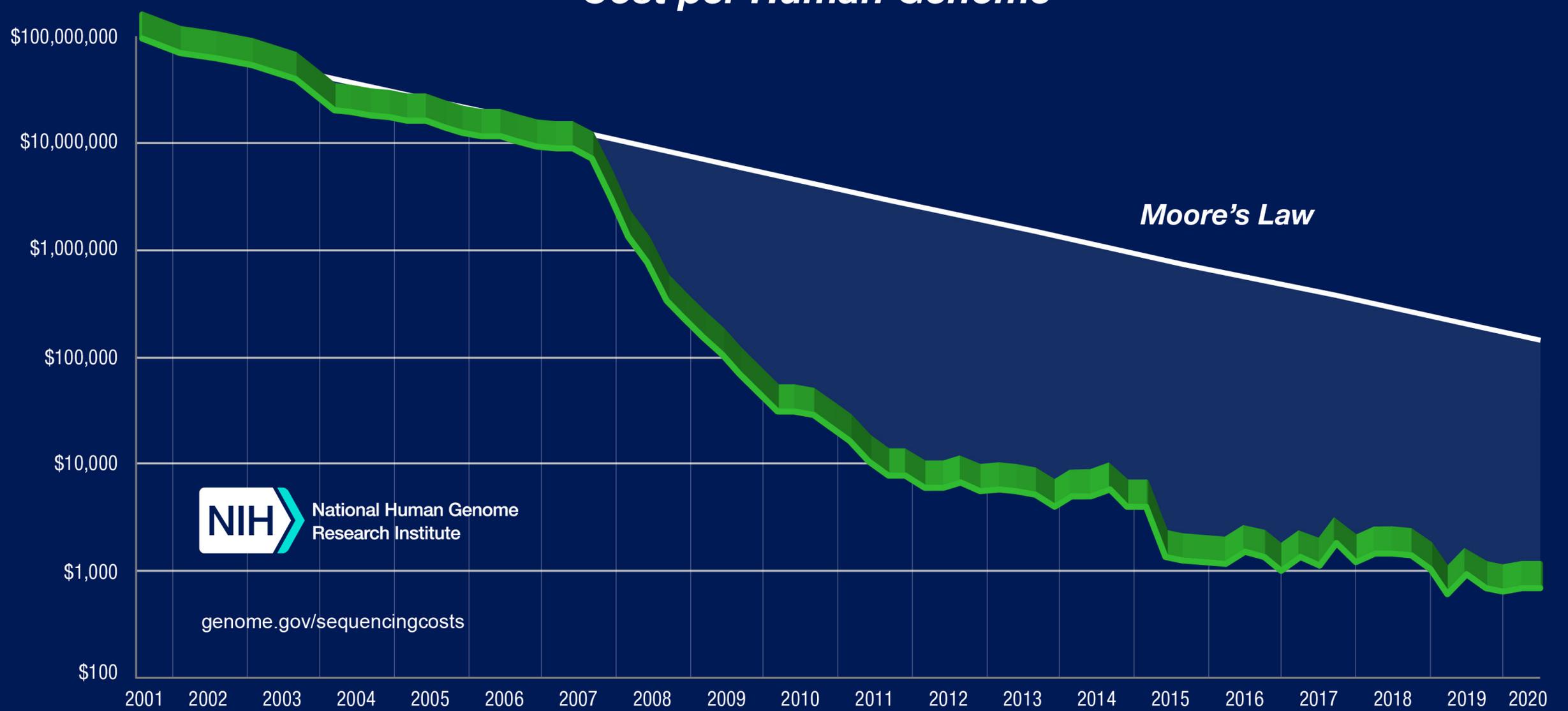
AAAS

SPECIAL ISSUE

## HUMAN GENOME AT



# Cost per Human Genome



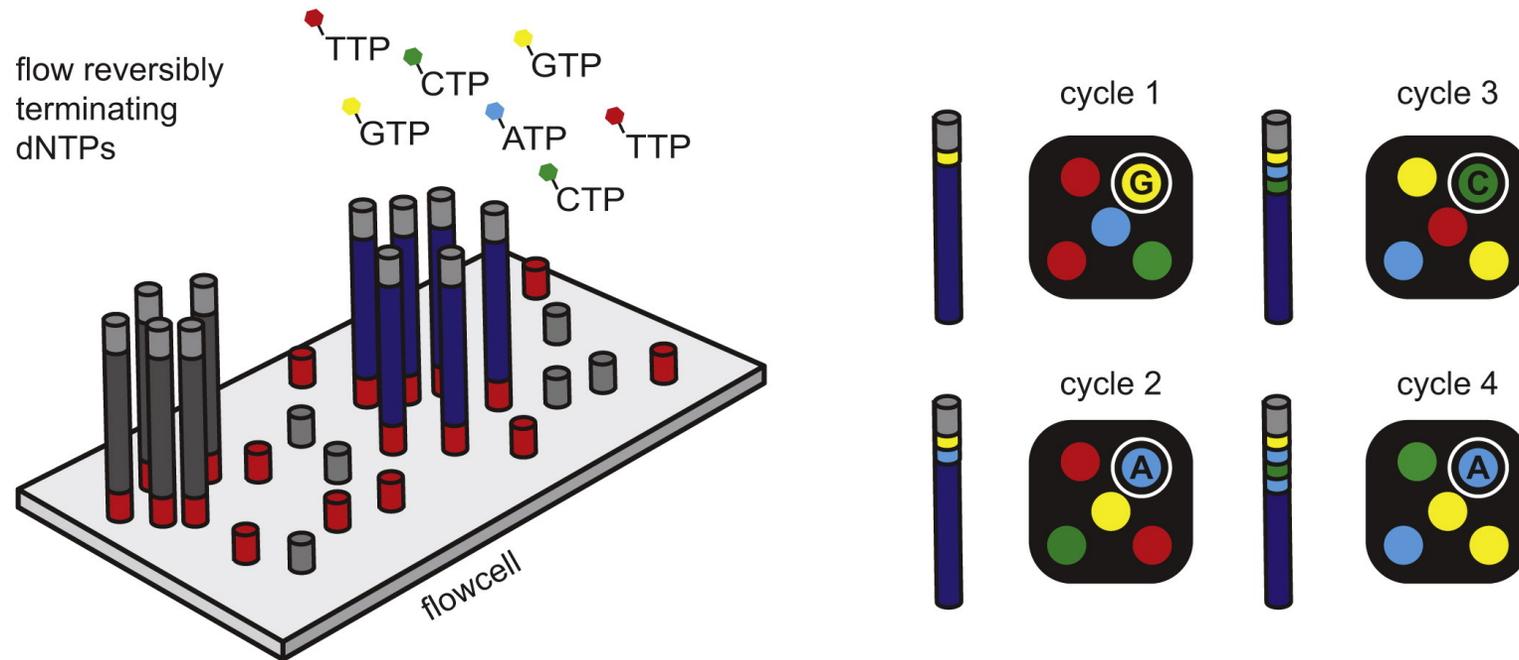
**NIH** National Human Genome Research Institute

[genome.gov/sequencingcosts](http://genome.gov/sequencingcosts)

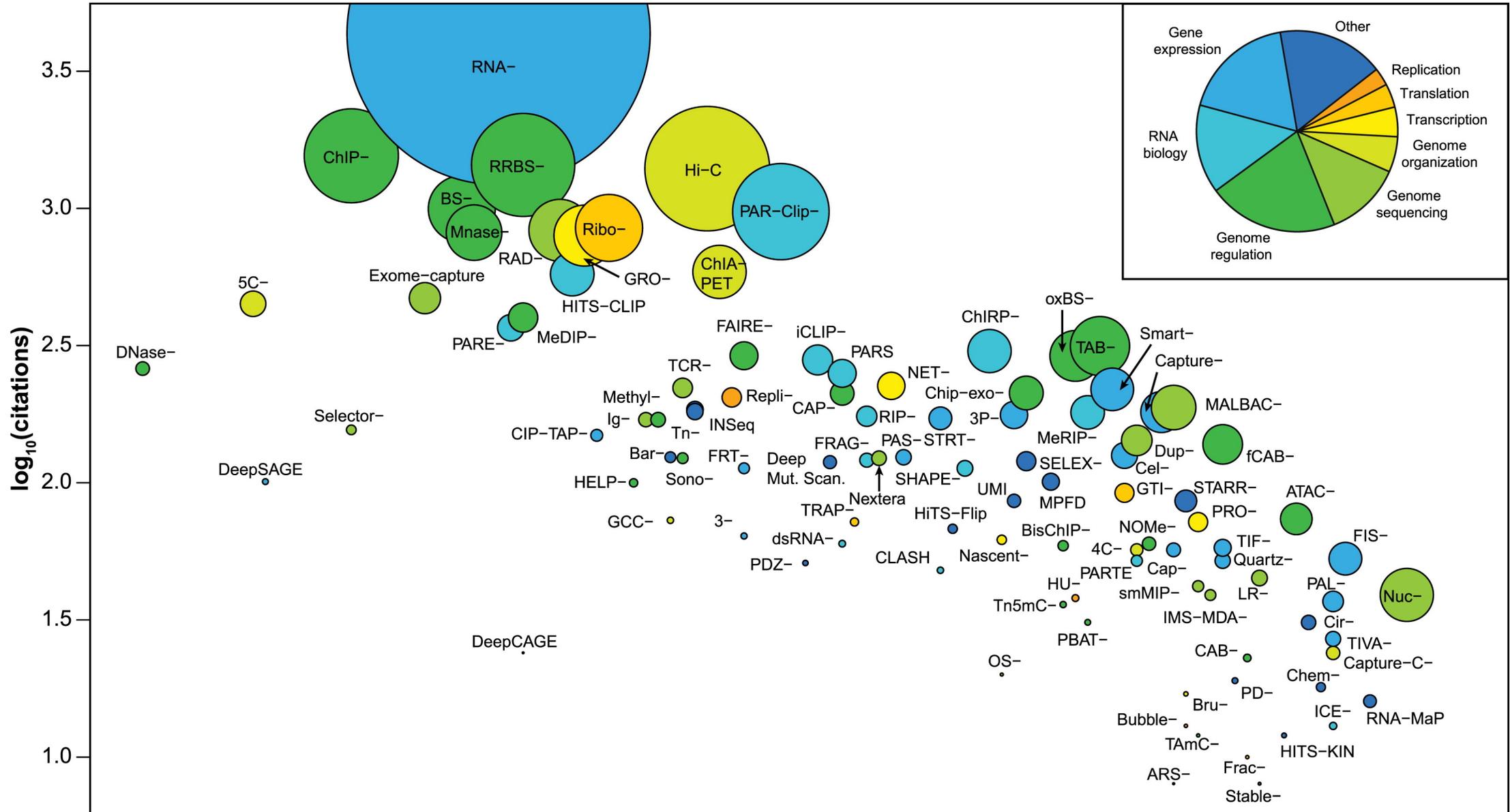
*Moore's Law*



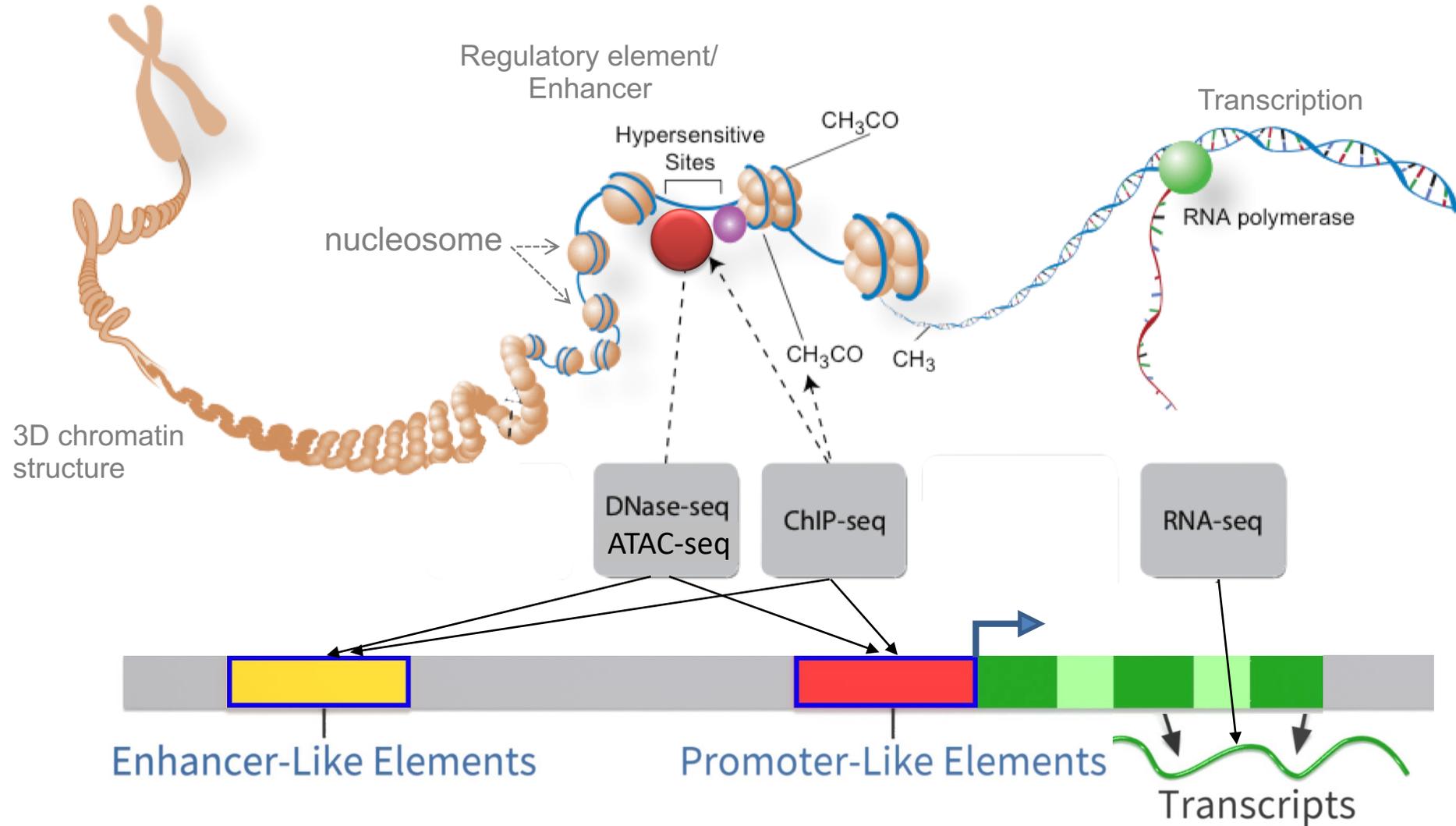
# Illumina/Solexa sequencing technology

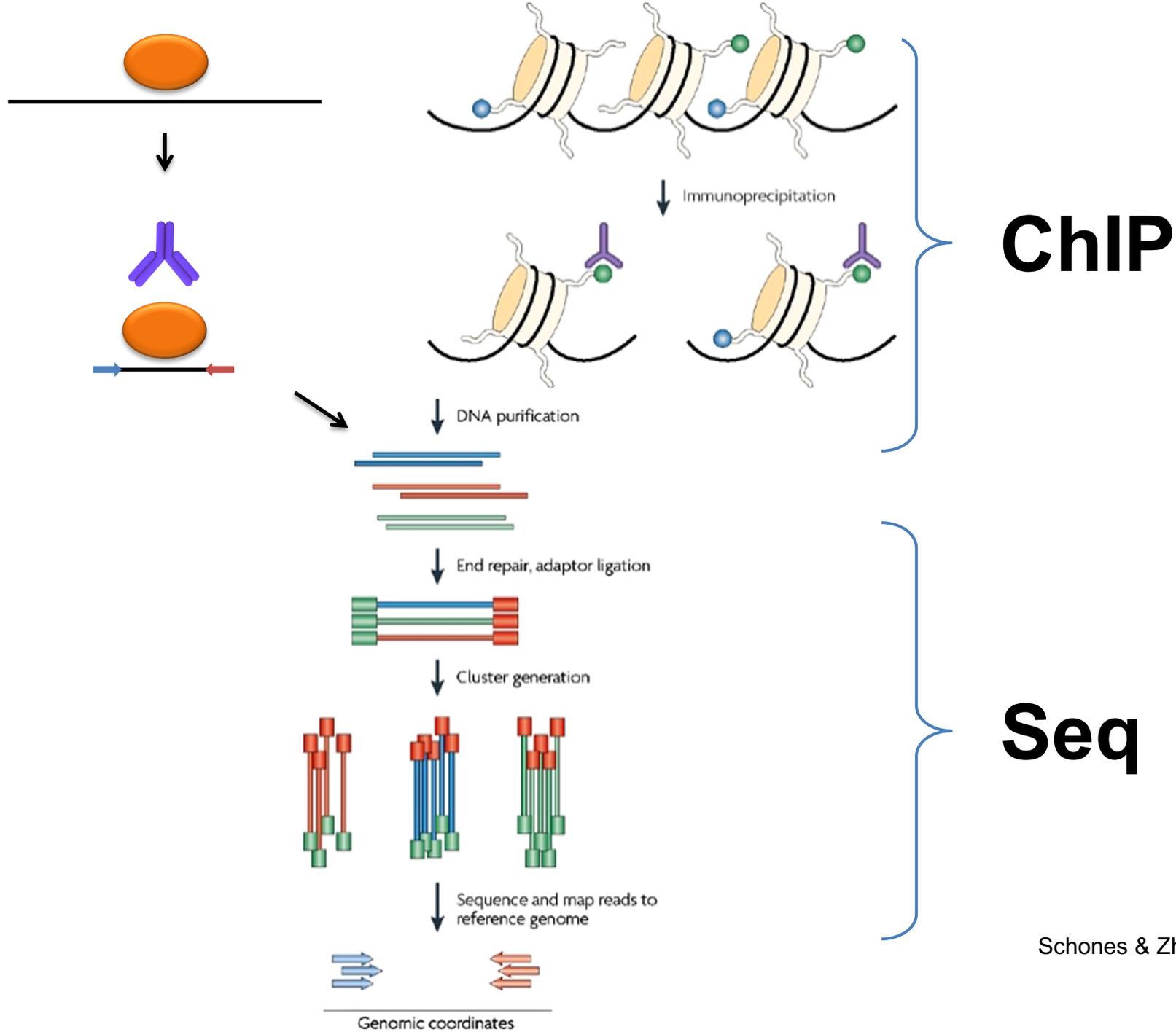


# NGS-based applications (-seq)



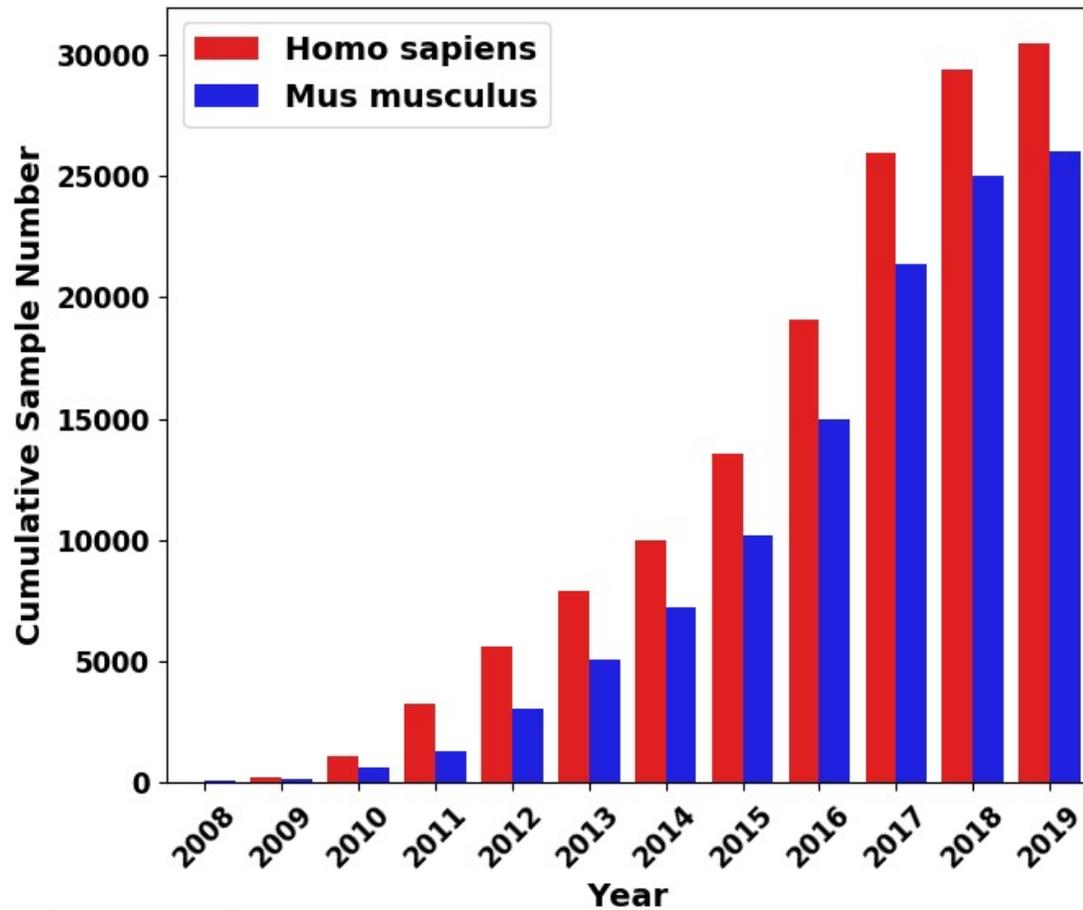
# NGS helps functional studies of the genome and epigenome



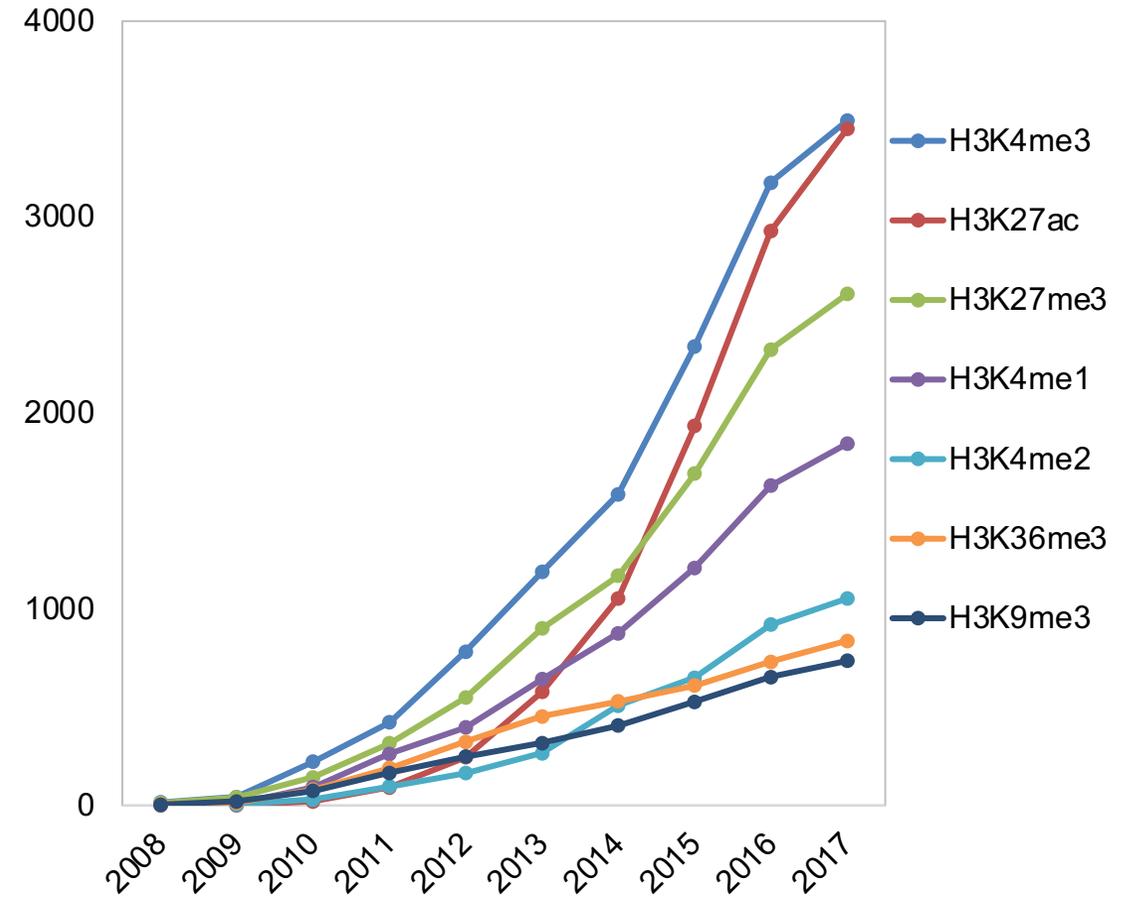


Schones & Zhao. *Nat. Rev. Genet.* 2008

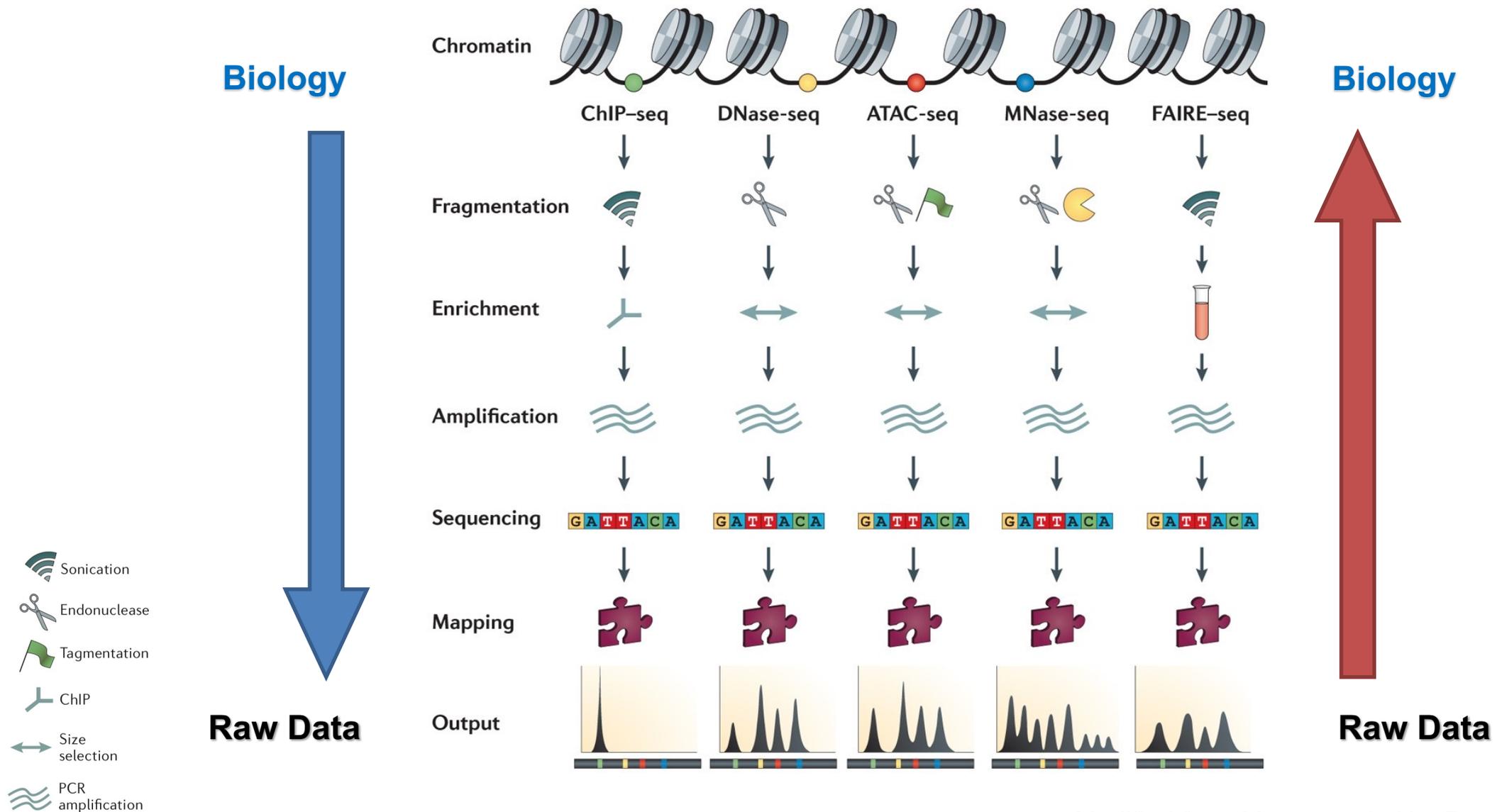
# ChIP-seq has become a dominant method for profiling epigenomes

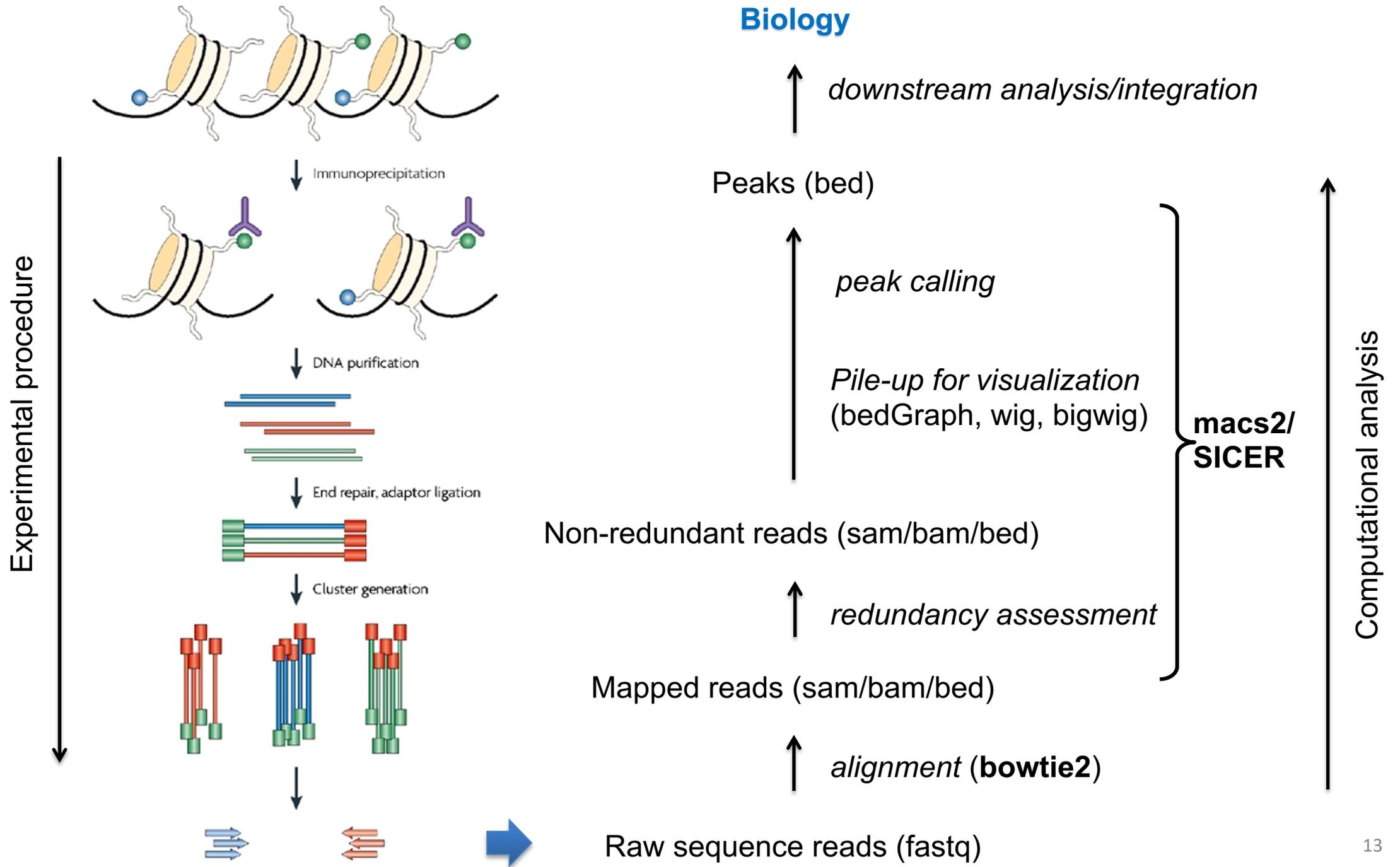


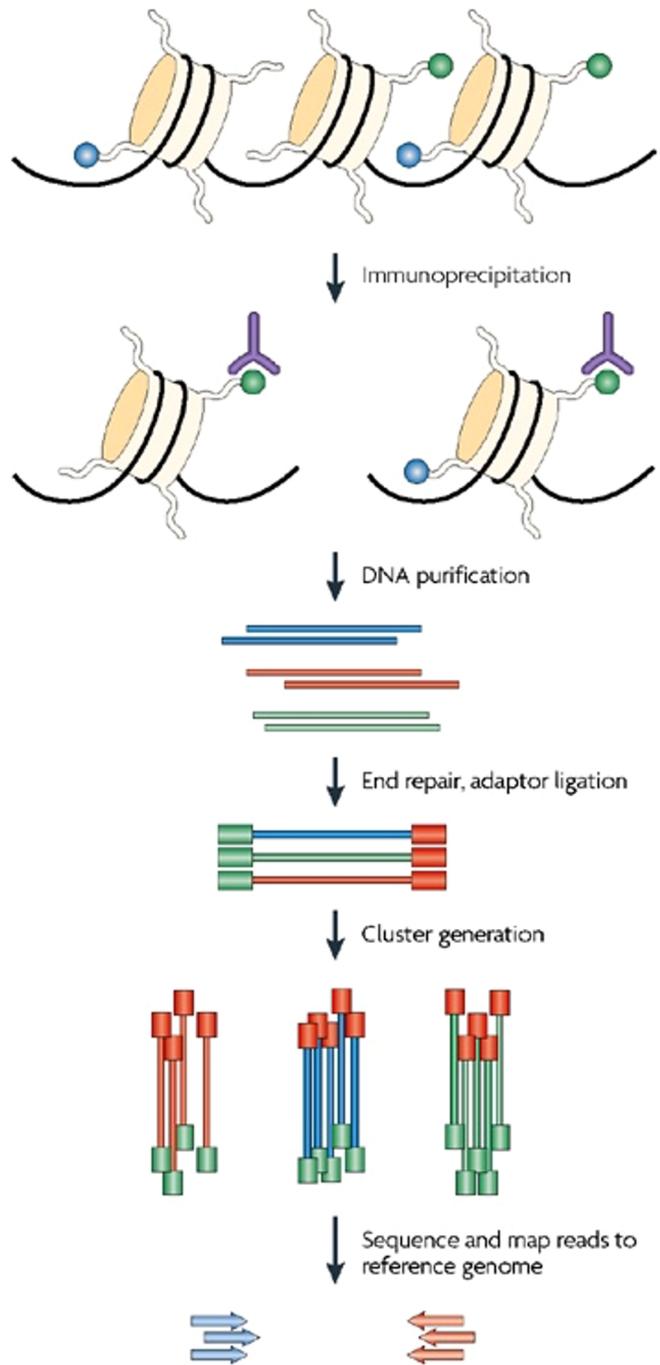
cistrome.org/db



# NGS data analysis strategy: Reconstruction of biology



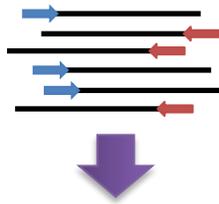




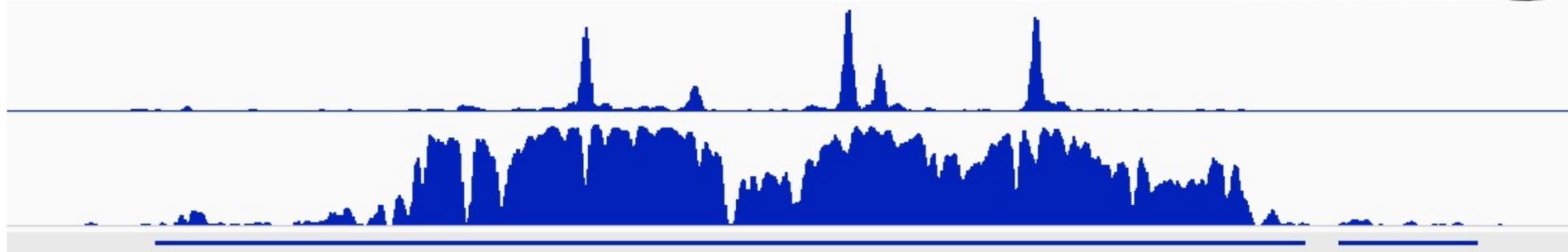
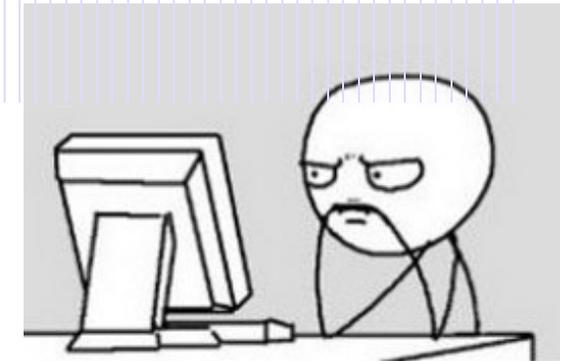
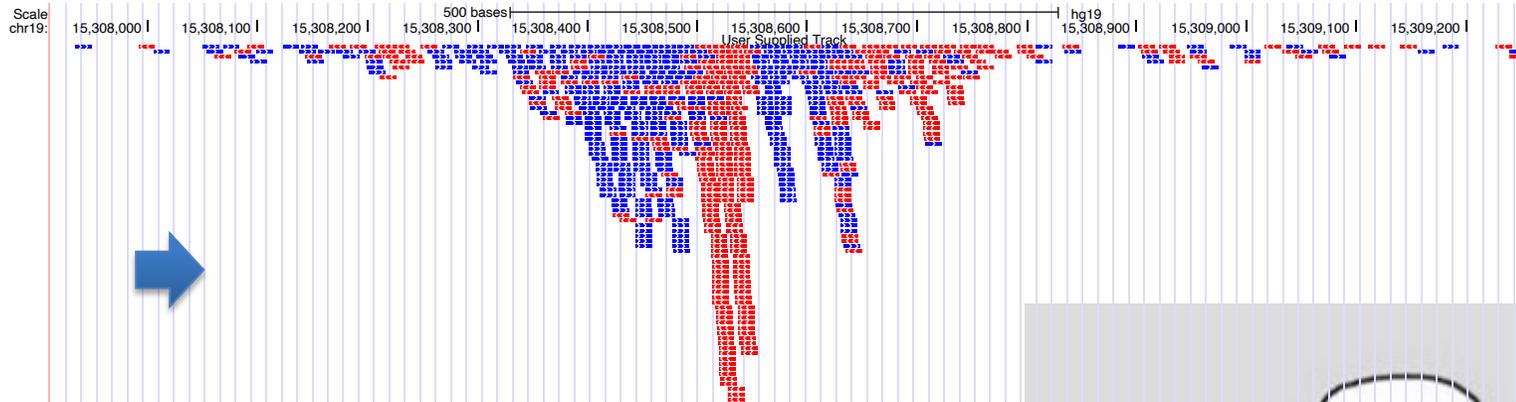
## ChIP-seq: Study design

- **Background Control: Input or IgG**
  - Input chromatin: sonicated/digested chromatin without immunoprecipitation
  - IgG: “unspecific” immunoprecipitation
- **Study Control:**
  - Control exp sample: ChIP + input
  - Treated exp sample: ChIP + input

# ChIP-seq data analysis overview

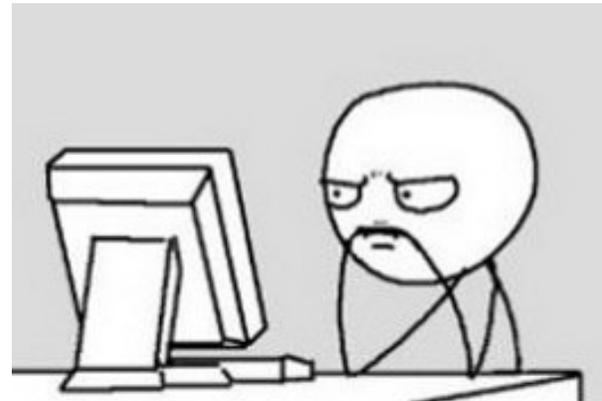


```
@ILLUMINA-8879DC:231:KK:3:1:1070:945 1:Y:0:  
NNAATACAGTCAGAAACATATCATATTGGAGAATA  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1153:945 1:Y:0:  
NNAAGCACACAGAAGATAACTAAACAATCAAGTAG  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1222:945 1:Y:0:  
NNAAGGCTCTTGAGAAGAAATCATTCTGGATGGCA  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1304:939 1:Y:0:  
NNCCAGGCTCCCGCATTCTCCTGCCTCAGCTTCT  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1354:945 1:Y:0:  
NNCTCTCCTTAGCTAACTTTCAACTAAGCCAAA  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1411:932 1:Y:0:  
NNGTAGGACCATGGCGTTGCGACACAAAAAATT  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1496:937 1:Y:0:  
NNNTTCATCGGGTTGAGAGTCCCTTGTTCATGCA  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1533:939 1:Y:0:  
NNNATTTCCCGTTCCAGGTCGCAATTTCCGCCGTT  
#####  
@ILLUMINA-8879DC:231:KK:3:1:1573:940 1:Y:0:  
NNGGGTGCGCCTTTAGTCCAGCTACTCAGGAAC
```



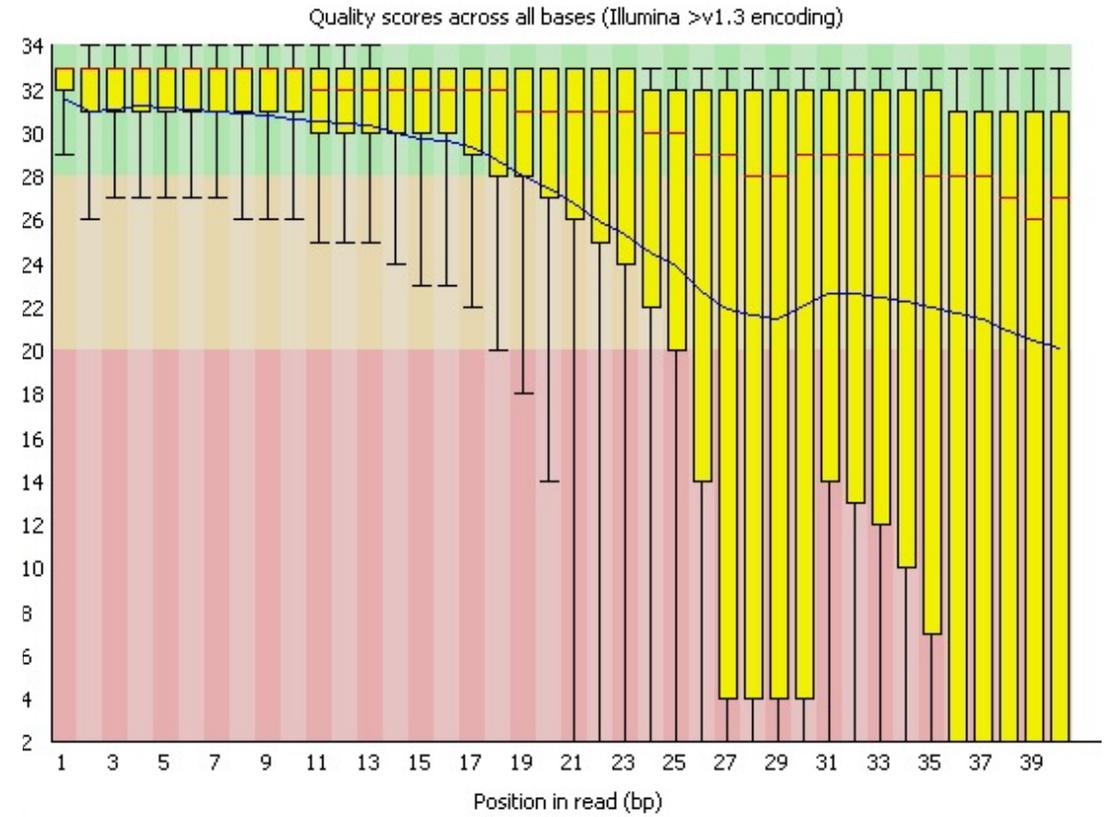
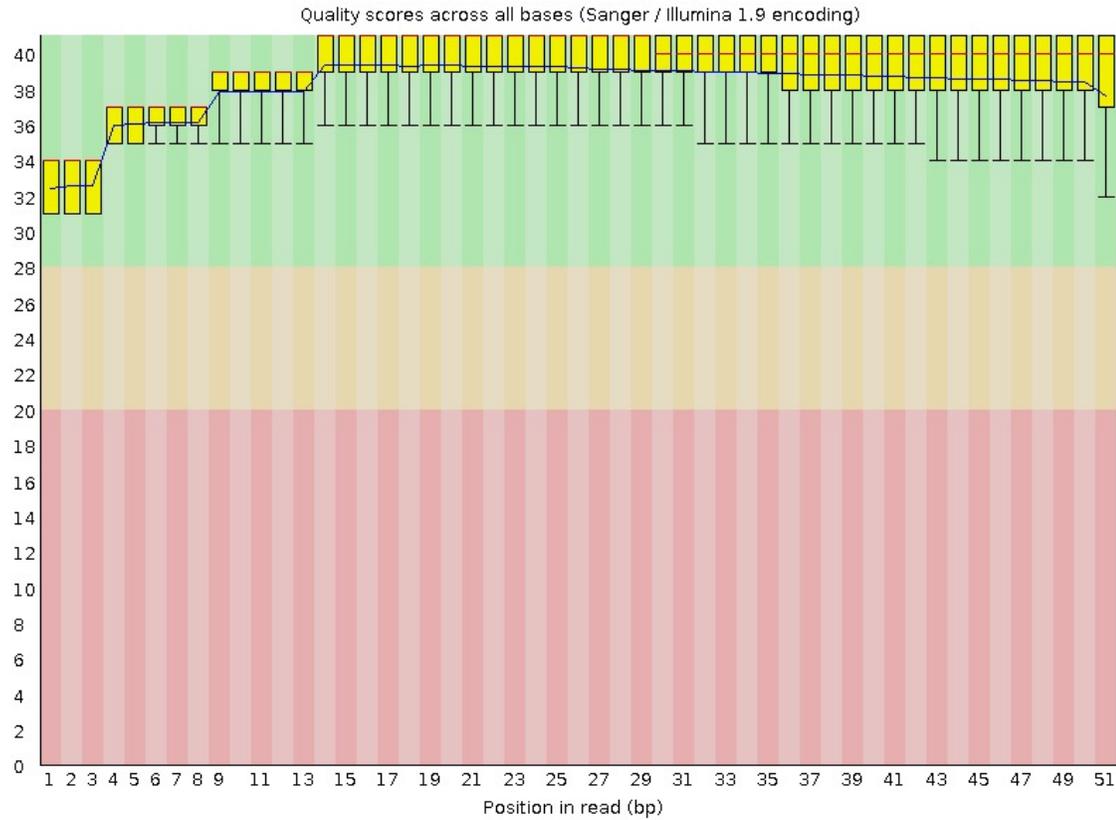
# ChIP-seq data analysis overview

- Where in the genome do these sequence reads come from? - Sequence alignment and quality control
- What does the enrichment of sequences mean? - Peak calling
- What can we learn from these data? – Downstream analysis and integration

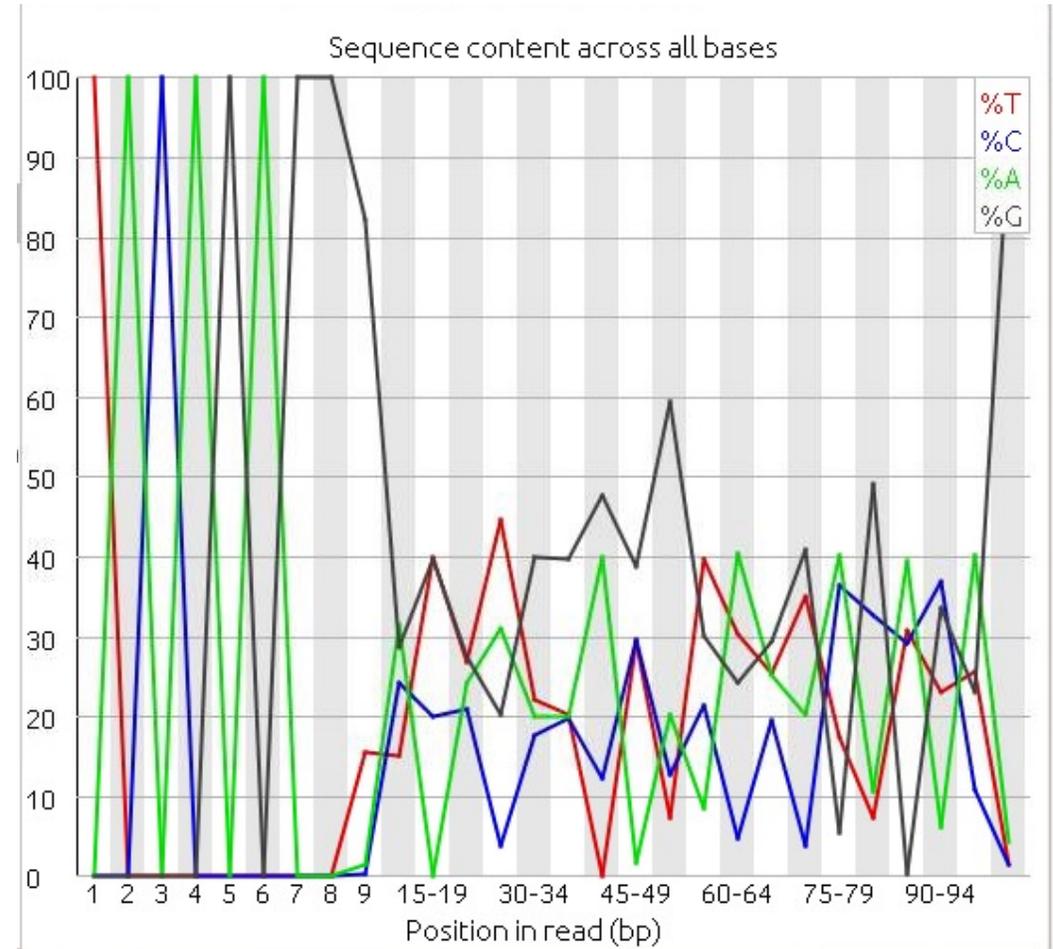
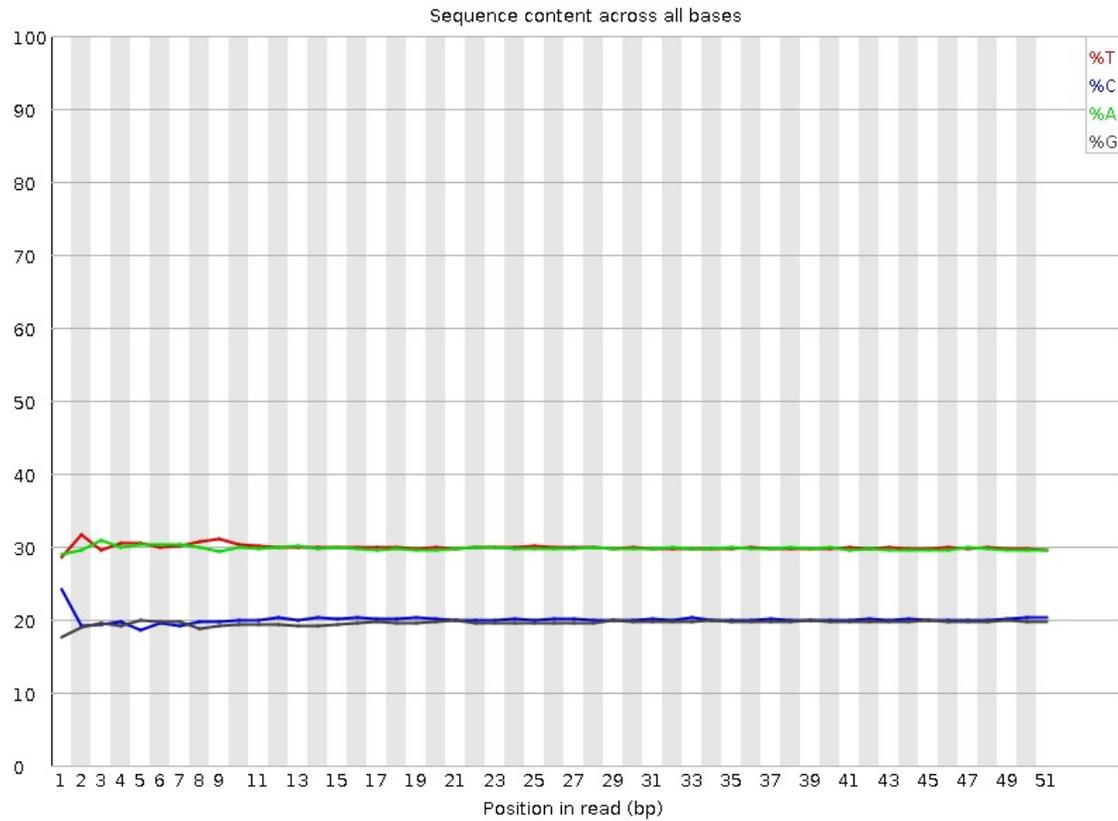




# Sequencing quality assessment: fastqc

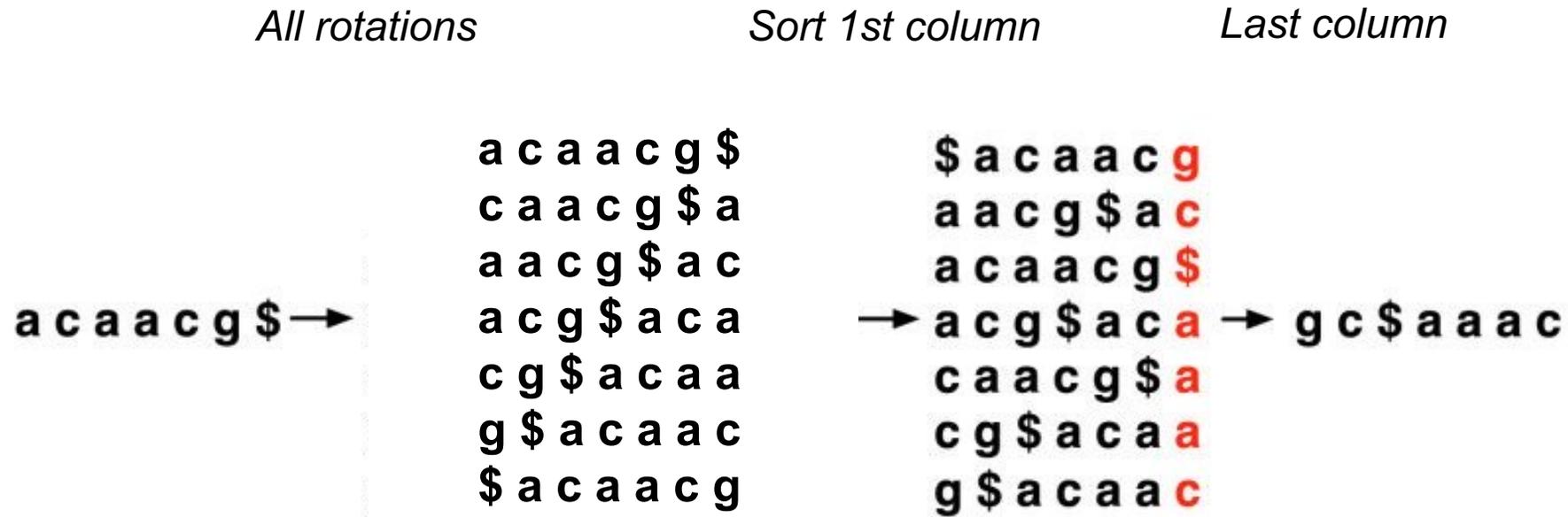


# Sequencing quality assessment: fastqc



# Burrows-Wheeler transform

Reversible permutation of the characters in a string, originally used for data compression



BW matrix

bowtie, BWA

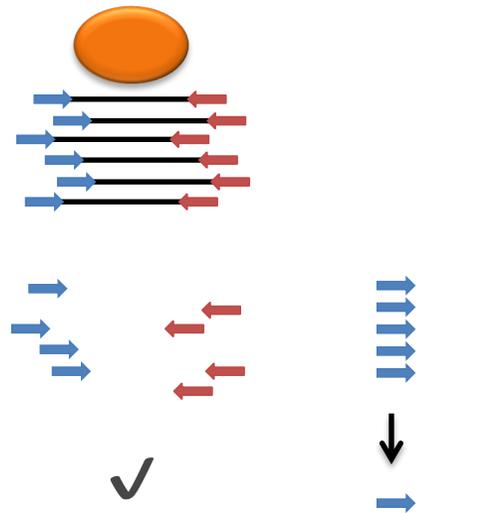
# Sequence read mapping: bowtie2/BWA

- alignment of each sequence read: **bowtie2** or **BWA**

{ cannot map to the reference genome X  
can map to multiple loci in the genome X  
can map to a unique location in the genome ✓

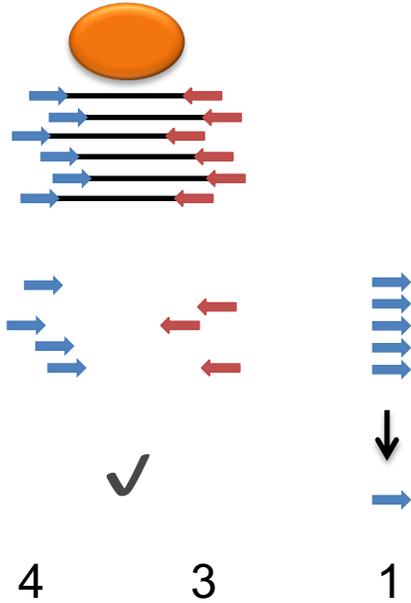
or Quality score

- redundancy assessment:



Langmead et al. 2009,  
Zang et al. 2009

# Redundancy Control



# mapped reads: 12  
 # non-redundant reads: 8  
 # locations w/ reads: 8  
 # locations w/ 1 read: 7

- Non-redundant rate:

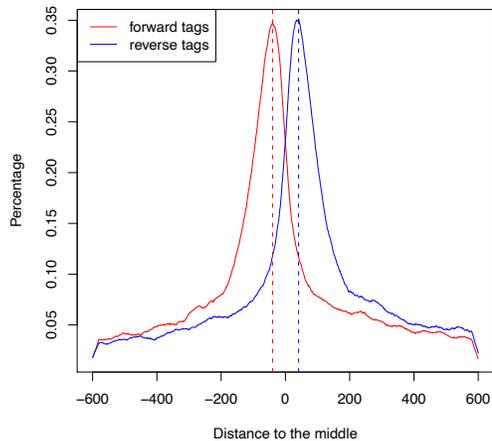
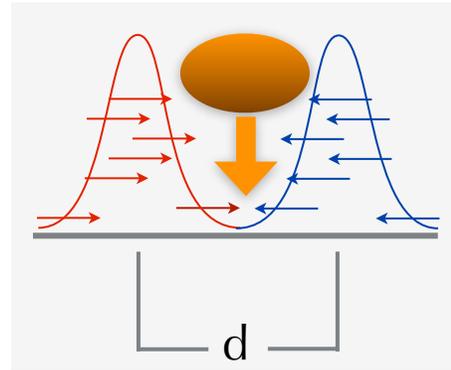
$$\frac{\text{\# non-redundant reads}}{\text{\# mapped reads}} = 8/12 = 66.7\%$$

- PBC (PCR Bottleneck Coefficient):

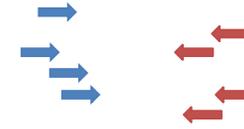
$$\frac{\text{\# locations w/ 1 read}}{\text{\# locations w/ reads}} = 7/8 = 87.5\%$$

# DNA fragment size estimation

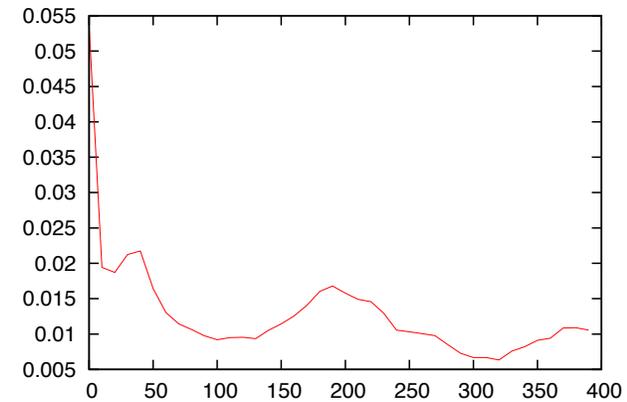
peak model



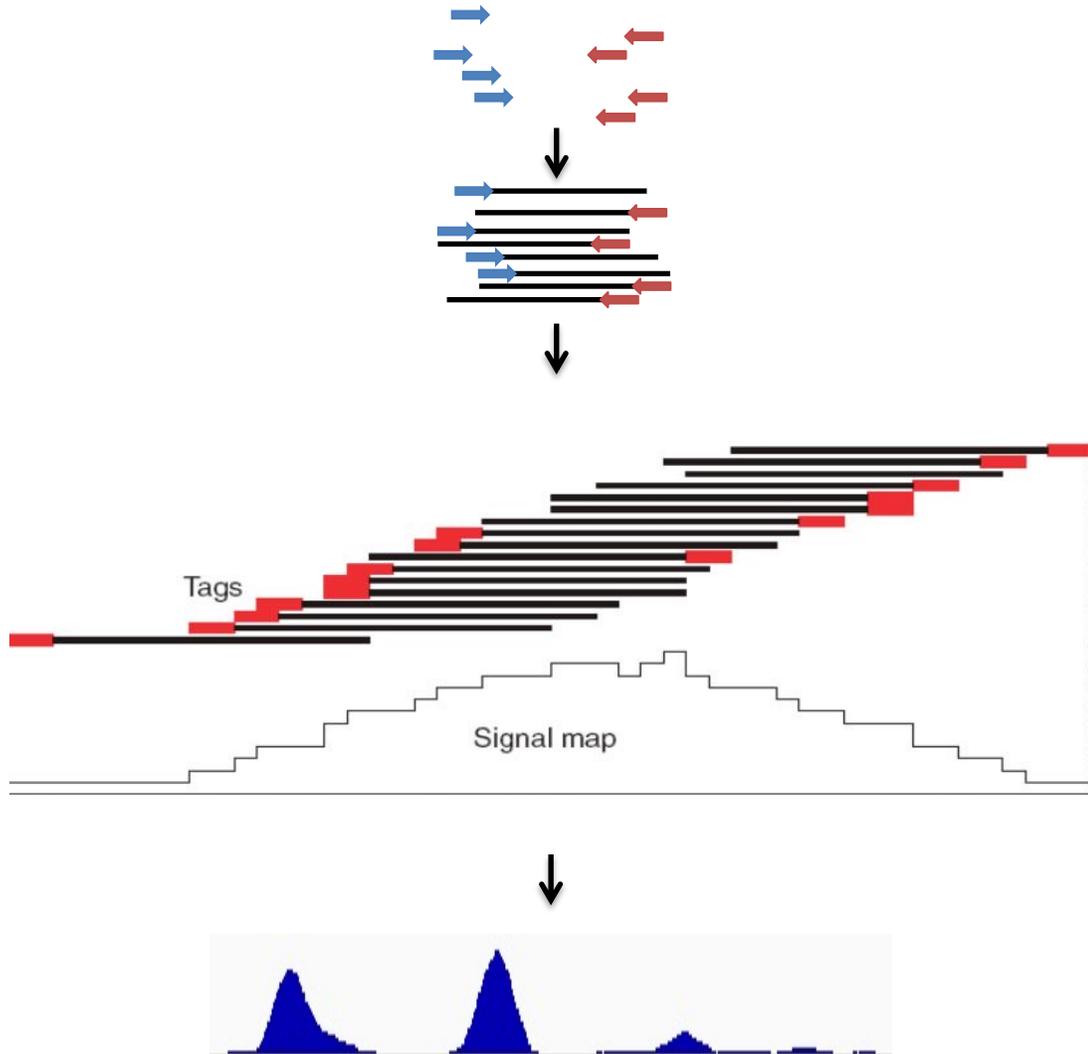
cross-correlation



$$C(r) = \frac{1}{X} \int_x (T_+(x) - \overline{T_+}) (T_-(x+r) - \overline{T_-})$$



# Pile up: visualization



- bedGraph:

chr4	10344200	10344250	5
chr4	10344250	10344300	10
chr4	10344300	10344350	25
chr4	10344350	10344400	15
chr4	10344400	10344450	8

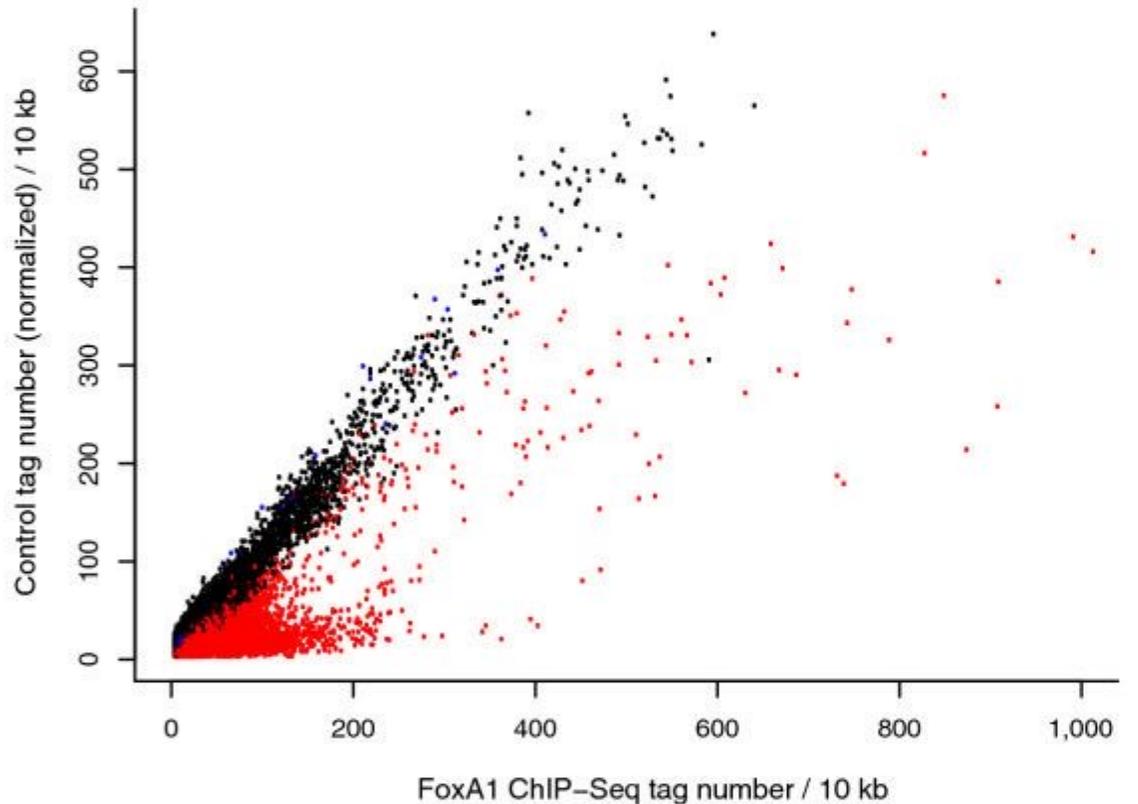
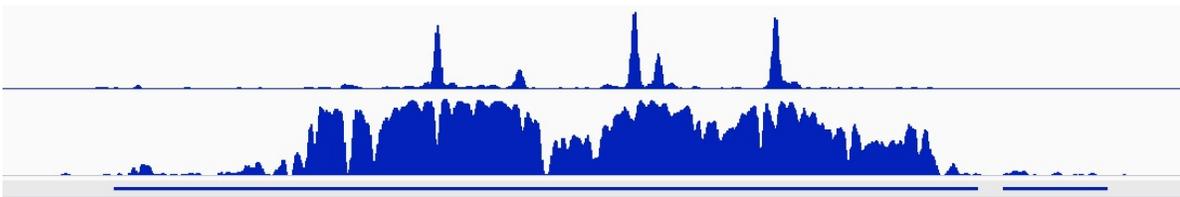
- wiggle:

```
track type=wiggle_0
variableStep chrom=chr4 span=50
10344200 5
10344250 10
10344300 25
10344350 15
10344400 8
```

- bigWig: indexed binary format

# ChIP-seq: Peak calling

- Goal: Identify regions in the genome enriched for sequence reads:
  - Compared to genomic background
  - Compared to input control

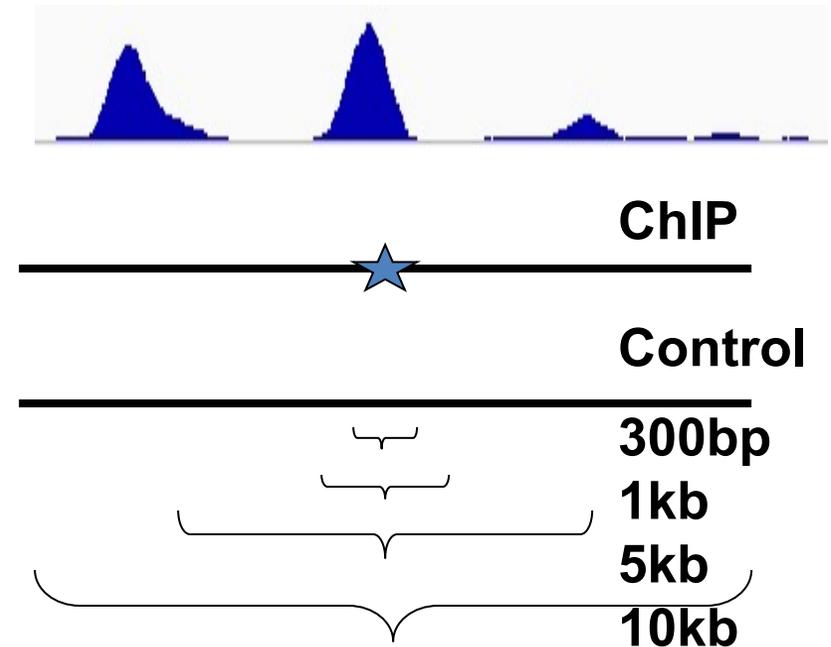


# MACS: model

- **Model-based Analysis for ChIP-Seq**
- Read distribution along the genome  $\sim$  Poisson distribution  
( $\lambda_{BG}$  = total tag / genome size)
- ChIP-seq show local biases in the genome
  - Chromatin and sequencing bias
  - 200-300bp control windows have to few tags
  - But can look further

$$\text{Dynamic } \lambda_{local} = \max(\lambda_{BG}, [\lambda_{ctrl}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}])$$

- B-H adjustment to correct for FDR
  - p-value  $\rightarrow$  q-value



Zhang et al, *Genome Bio*, 2008

# MACS: Critical input parameters

```
macs2 callpeak [-h] -t TFILE [TFILE ...] [-c [CFILE]] [-g GSIZE] [-q QVALUE | -p PVALUE] [--  
outdir OUTDIR] [-n NAME] [-B]
```

**-g GSIZE** Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts: 'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for C. elegans (9e7) and 'dm' for fruitfly (1.2e8), Default:hs

**-q QVALUE** Minimum FDR (q-value) cutoff for peak detection. DEFAULT: 0.05. -q, and -p are mutually exclusive.

**--outdir OUTDIR** If specified all output files will be written to that directory. Default: the current working directory

**-n NAME** Experiment name, which will be used to generate output file names. DEFAULT: "NA"

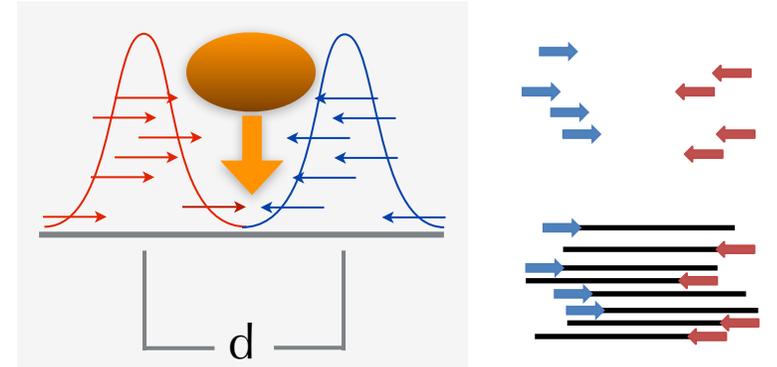
**-B, --bdg** Whether or not to save extended fragment pileup, and local lambda tracks (two files) at every bp into a bedGraph file. DEFAULT: False

# MACS: Output interpretation

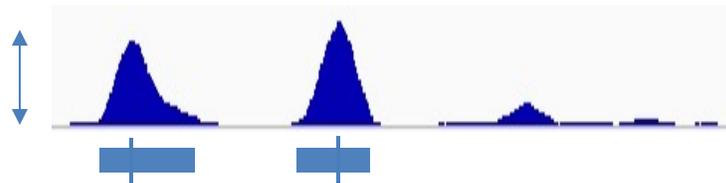
```
# This file is generated by MACS version 2.1.2
# Command line: callpeak -t ../bowtie2/AR.sam -g hs -n AR --bdg
# ARGUMENTS LIST:
# name = AR
# format = AUTO
# ChIP-seq file = ['../bowtie2/AR.sam']
# control file = None
# effective genome size = 2.70e+09
# band width = 300
# model fold = [5, 50]
# qvalue cutoff = 5.00e-02
# The maximum gap between significant sites is assigned as the read length/tag size.
# The minimum length of peaks is assigned as the predicted fragment length "d".
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 10000 bps
# Broad region calling is off
# Paired-End mode is off
```

# MACS: Output interpretation

```
# tag size is determined as 51 bps
# total tags in treatment: 19442622
# tags after filtering in treatment: 17218335
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.11
# d = 141
# alternative fragment length(s) may be 141 bps
```



chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-	
log10(qvalue)	name								
chr1	2603	2989	387	2870	18.00	6.68596	3.52825	3.66748	AR_peak_1
chr1	138179	138371	193	138281	18.00	14.90779	7.93021	11.47829	AR_peak_2
chr1	36515	36714	200	36609	16.00	12.59143	7.05394	9.25447	AR_peak_3
chr1	201091	201231	141	201114	10.00	7.58293	5.23859	4.50002	AR_peak_4
chr1	69373	69558	186	69452	18.00	9.61904	4.93737	6.41821	AR_peak_5



# MACS: Output interpretation

- Excel

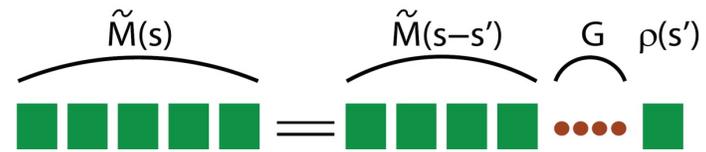
chr	start	end	length	abs_summit	pileup	-log10(pvalue)	fold_enrichment	-	
chr1	log10(qvalue)	name							
chr1	2603	2989	387	2870	18.00	6.68596	3.52825	3.66748	AR_peak_1
chr1	138179	138371	193	138281	18.00	14.90779	7.93021	11.47829	AR_peak_2
chr1	36515	36714	200	36609	16.00	12.59143	7.05394	9.25447	AR_peak_3
chr1	201091	201231	141	201114	10.00	7.58293	5.23859	4.50002	AR_peak_4
chr1	69373	69558	186	69452	18.00	9.61904	4.93737	6.41821	AR_peak_5

- narrowPeak

chr	start	end	name	score	fold	p	q	sm	
chr1	591170	591325	AR_peak_290	82	.	6.63900	11.50806	8.21785	25
chr1	629218	629993	AR_peak_291	295	.	3.42374	33.50185	29.54851	636
chr1	630286	630453	AR_peak_292	106	.	2.39458	14.04047	10.64496	81
chr1	630765	631382	AR_peak_293	239	.	3.14283	27.79379	23.97848	480
chr1	631877	632366	AR_peak_294	224	.	3.06645	26.24850	22.47273	380

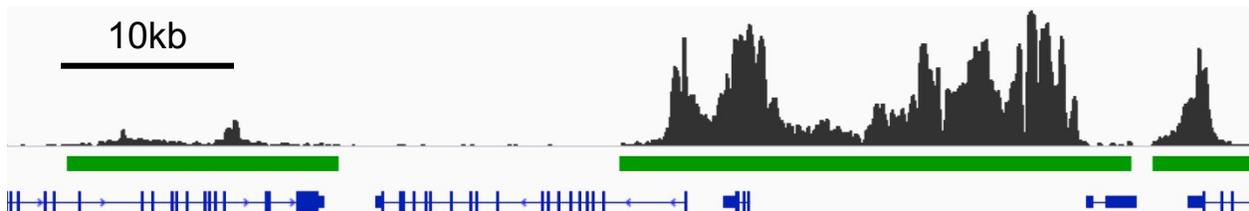
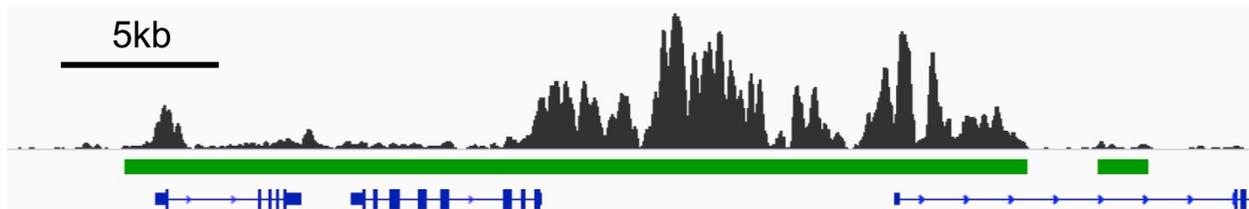
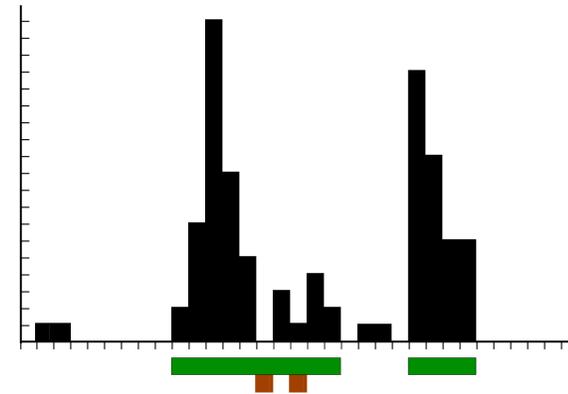
# Call broad peaks: SICER

- **Spatial-clustering Identification of ChIP-Enriched Regions**



$$\tilde{M}(s) = G(\lambda, l_0, g) \int_{s_0}^s ds' \tilde{M}(s-s') \rho(s')$$

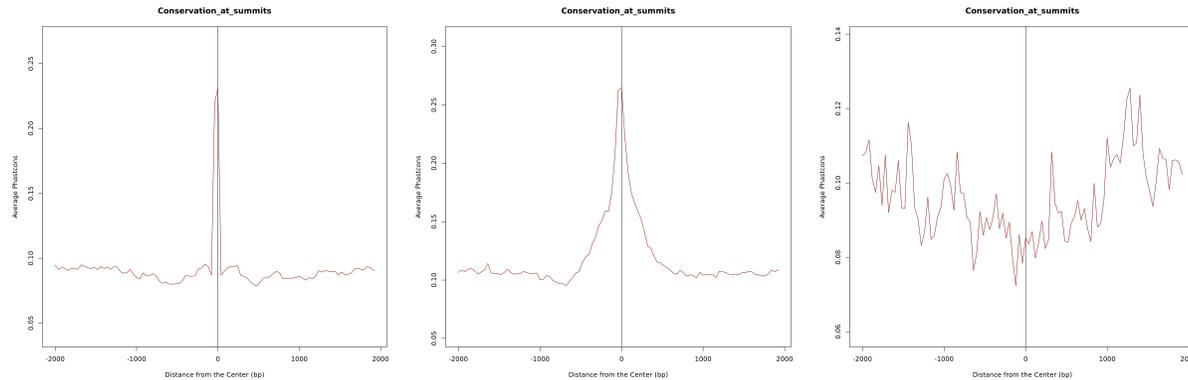
$$M(s) = t^{g+1} \tilde{M}(s) t^{g+1}$$



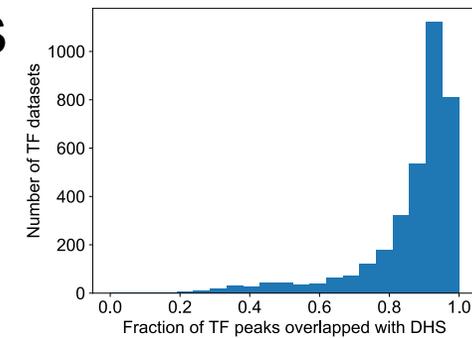
# Quality Control

- FRiP (Fraction of Reads in Peaks) score
  - 1-10% for TF is normal
- Number of peaks
  - Number of peaks with high fold-enrichment, e.g, 5, 10, ...
  - 2000

- Sequence conservation



- Fraction of peaks within regulatory regions
  - 80%

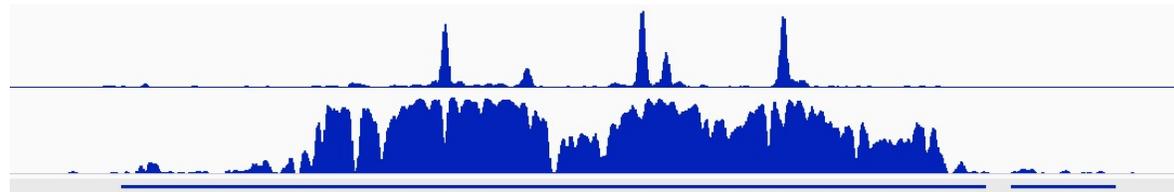


# Data formats

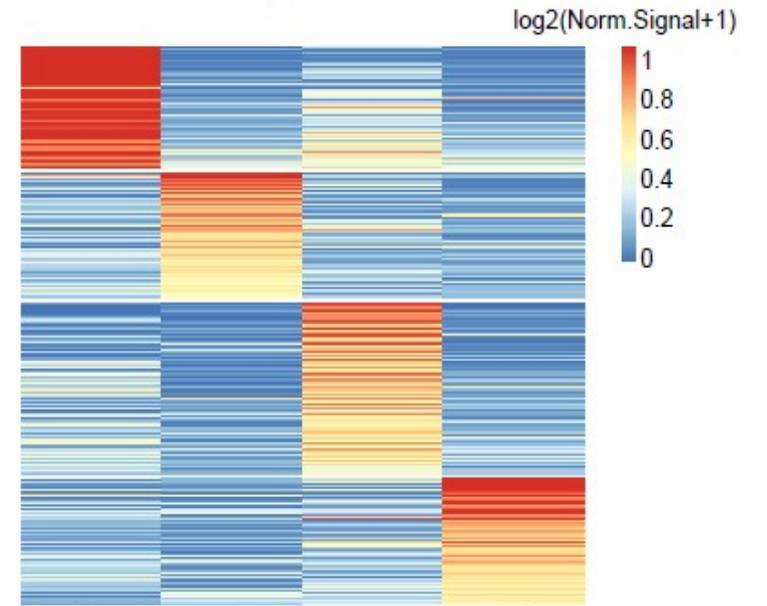
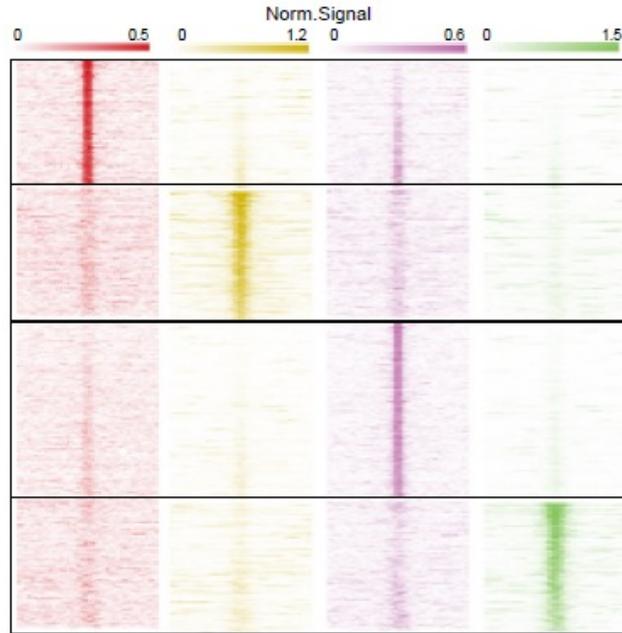
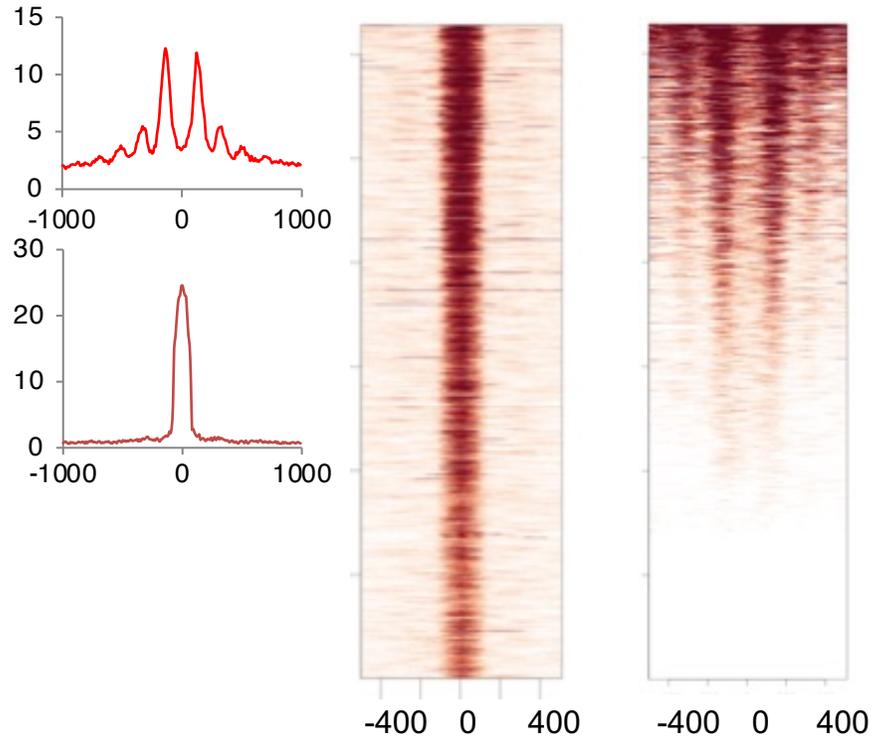
- BED:

chr11	10344210	10344260	255	0	-
chr4	76649430	76649480	255	0	+
chr3	77858754	77858804	255	0	+
chr16	62688333	62688383	255	0	+
chr22	33031123	33031173	255	0	-

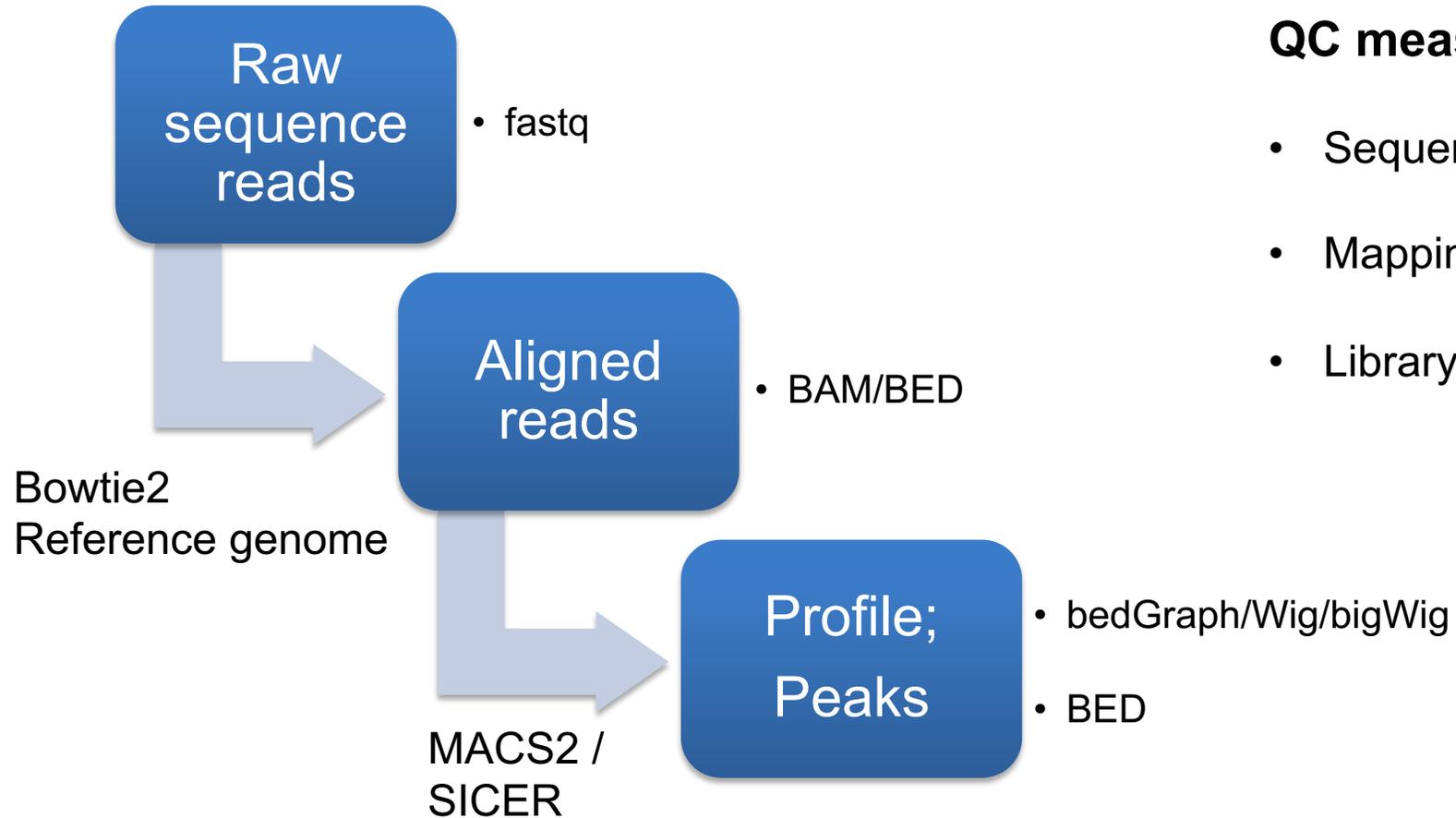
- SAM/BAM: aligned sequencing reads
- bedGraph, Wig, bigWig: pile-up profiles for browser visualization



# Data visualization

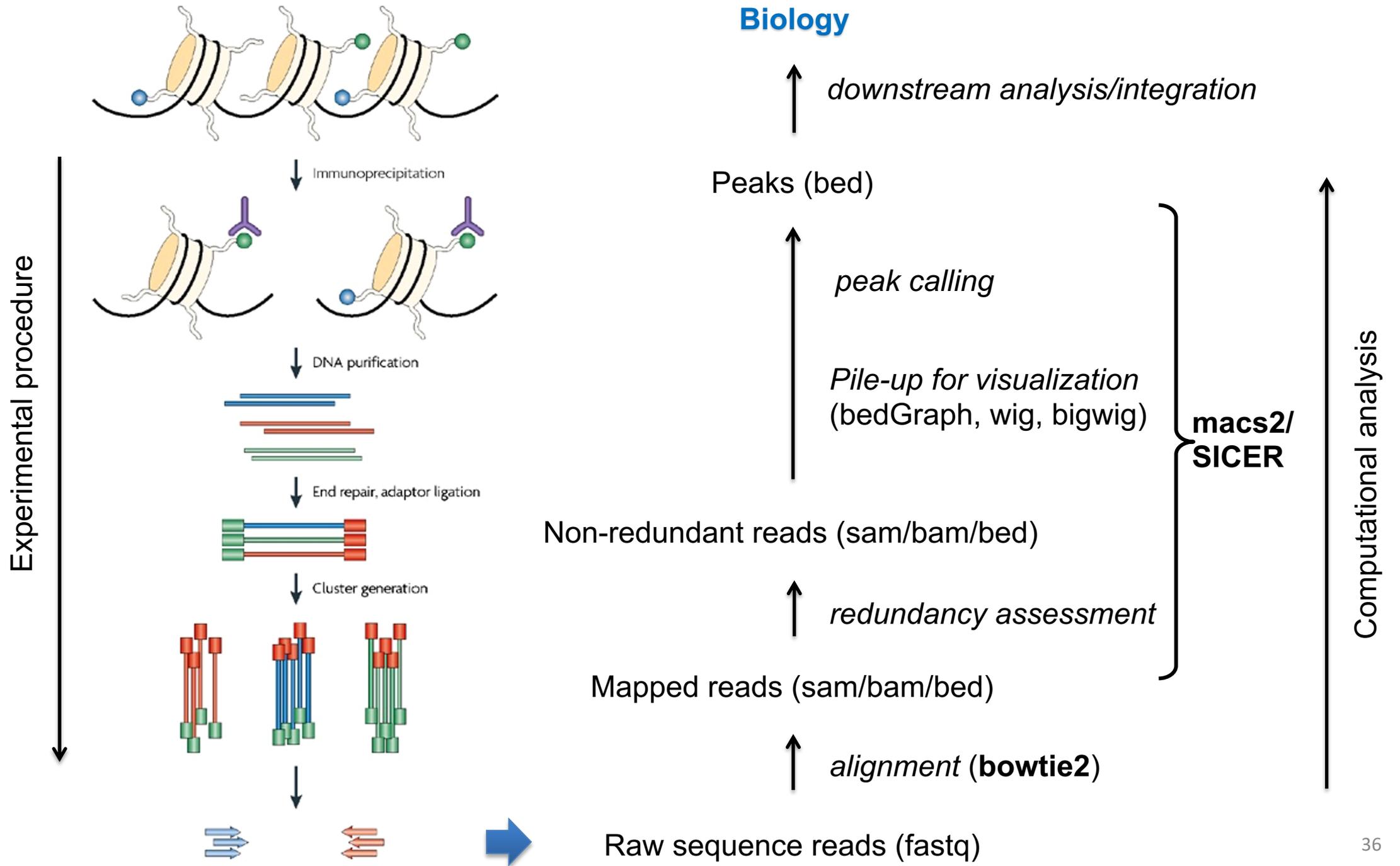


# ChIP-seq Data flow and QC summary



## QC measures

- Sequence quality (fastqc)
- Mapping quality (uniquely mapped ratio)
- Library complexity (PBC)
- Fold enrichment, peaks
- Signal/Noise (FRIP score)
- Regulatory annotation



**Break**

# Outline

- 1st Half:
  - NGS introduction
  - NGS data analysis strategy
  - ChIP-seq data analysis
- **2nd Half:**
  - Other NGS data analysis
  - Downstream analysis and data integration
  - Online resources

# Other epigenetics NGS data

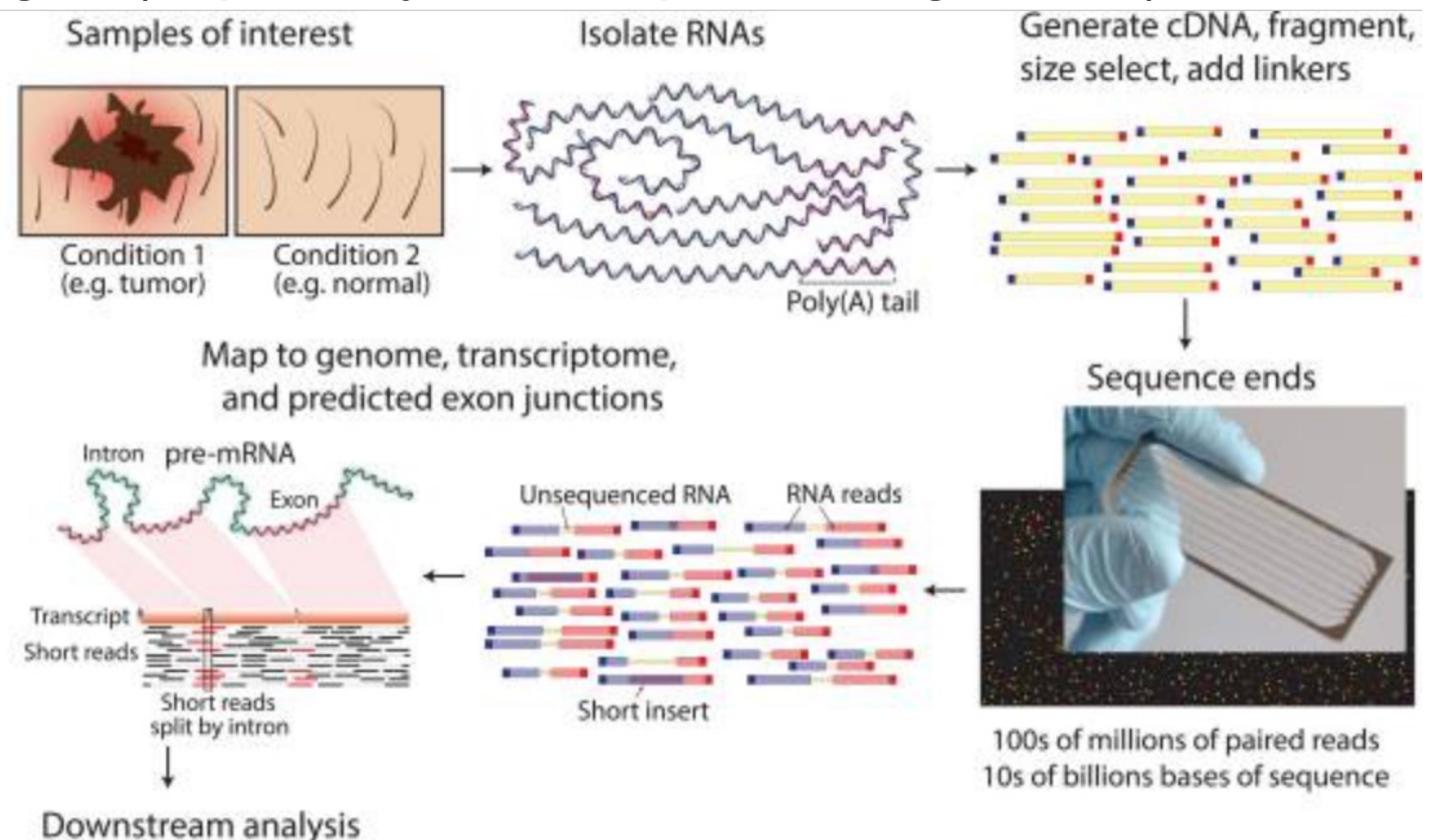
- RNA-seq
- BS-seq
- Hi-C
- Single-cell techniques (scRNA-seq, scATAC-seq, etc.)

# RNA-seq

- Detect gene expression (transcription)
  - Readout of epigenetic changes (especially transcriptional regulation)

- Protocols

- mRNA (polyA)
- Total RNA (rRNA depletion)
- ...



# RNA-seq

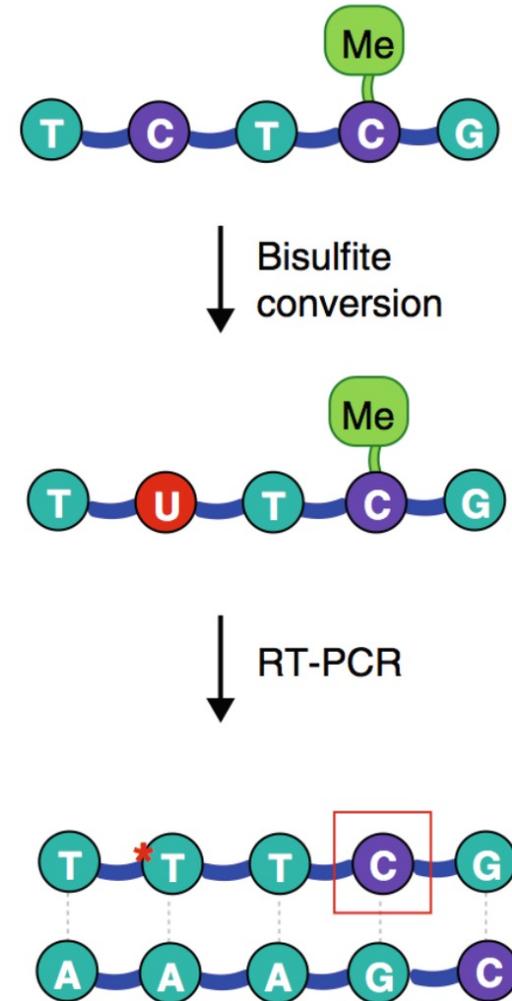
- QC
  - Exon (gene body) coverage
    - %reads located on exon (genebody) regions
  - Splicing junctions
  - Strand specificity
  - Software
    - RseQC: <http://rseqc.sourceforge.net>

# RNA-seq

- Analysis
  - Mapping
    - Hisat2: *Kim et al., Nat Biotechnol. 2019*
  - Gene expression index / reads count
    - Stringtie (novel transcript / splicing events): *Pertea et al., Nat Biotechnol. 2015*
    - HT-seq: *Anders et al., Bioinformatics 2015*
  - Differential expression
    - DESeq2: *Love et al., Genome Biol. 2014*
  - Visualization
    - Pile up the aligned reads (bedtools / ucsc tools)
    - IGV / genome browser view (UCSC genome browser, epigenome browser)

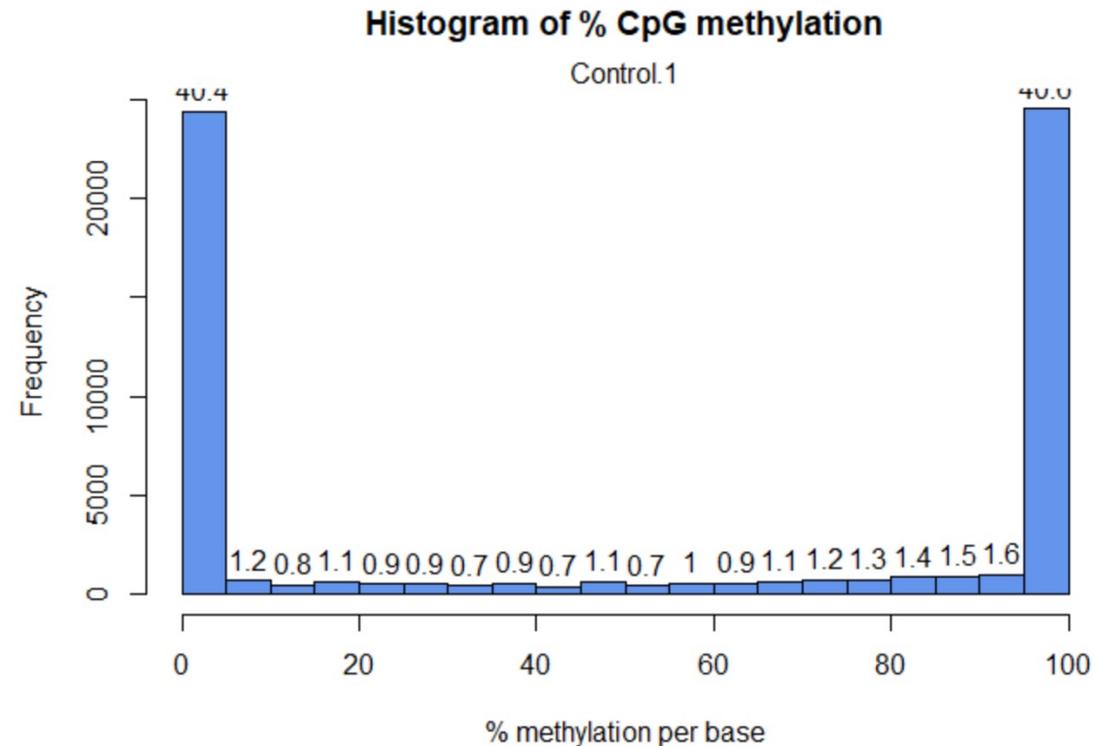
# BS-seq

- Detect genome-wide DNA-methylation profiles
- Quantitatively detect DNA-methylation level at CpG sites
- Require higher reads coverage
- Similar to whole genome sequencing (WGS)
- Many variants developed with different advantages
  - RRBS-seq
  - WGBS-seq



# BS-seq

- QC
  - Methylation level of lambda DNA
    - Naked DNA with no methylation
    - (whether the BS treatment works)
  - Global CpG methylation status distribution
    - Bi-modal
    - Reads coverage
    - Number of CpG sites with enough reads
    - (whether the detection is reliable)
  - Software:
    - BseQC, MOABS

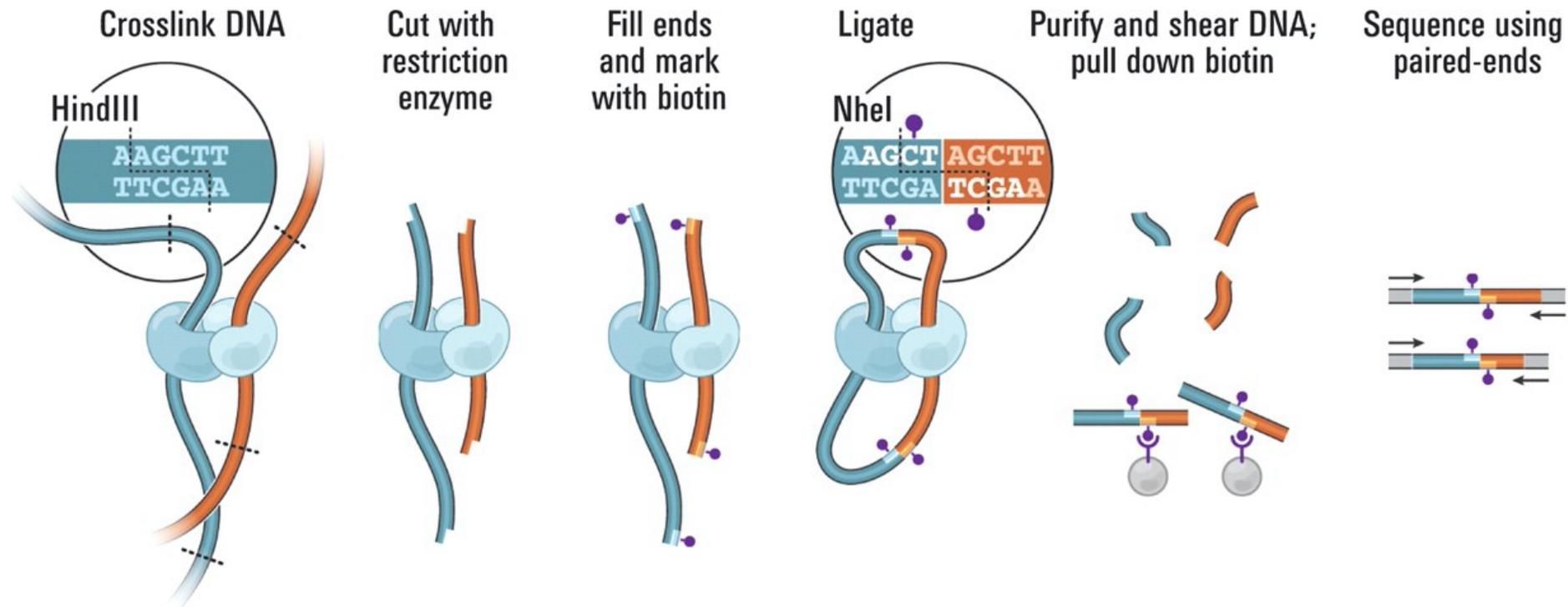


# BS-seq

- Analysis
  - Mapping: bsmmap
  - Methylation status calling: MOABS
    - *Sun et al., Genome Biol. 2014*
  - Differential methylated region (DMR): MOABS
- Output
  - CpG level methylation table
    - Cutoff on reads coverage (5/10)
  - Region level methylation table
    - Cutoff on qualified CpG (5/10)
  - Bigwig track for genome browser visualization
  - Differential methylated region (DMR)

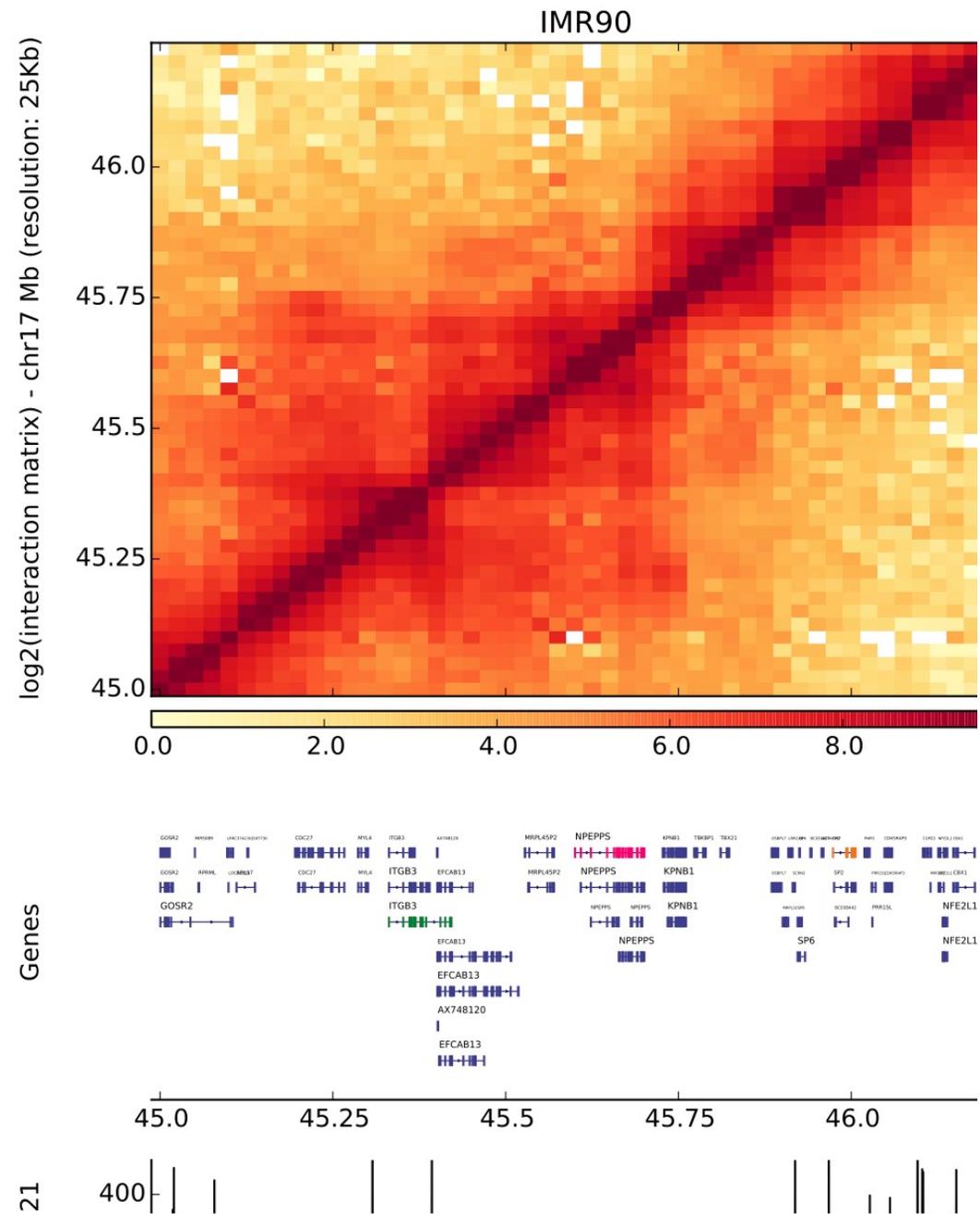
# Hi-C

- Chromatin interaction: 3D genome organization
- ChIA-PET
- HiChIP/PLAC-seq



# Hi-C

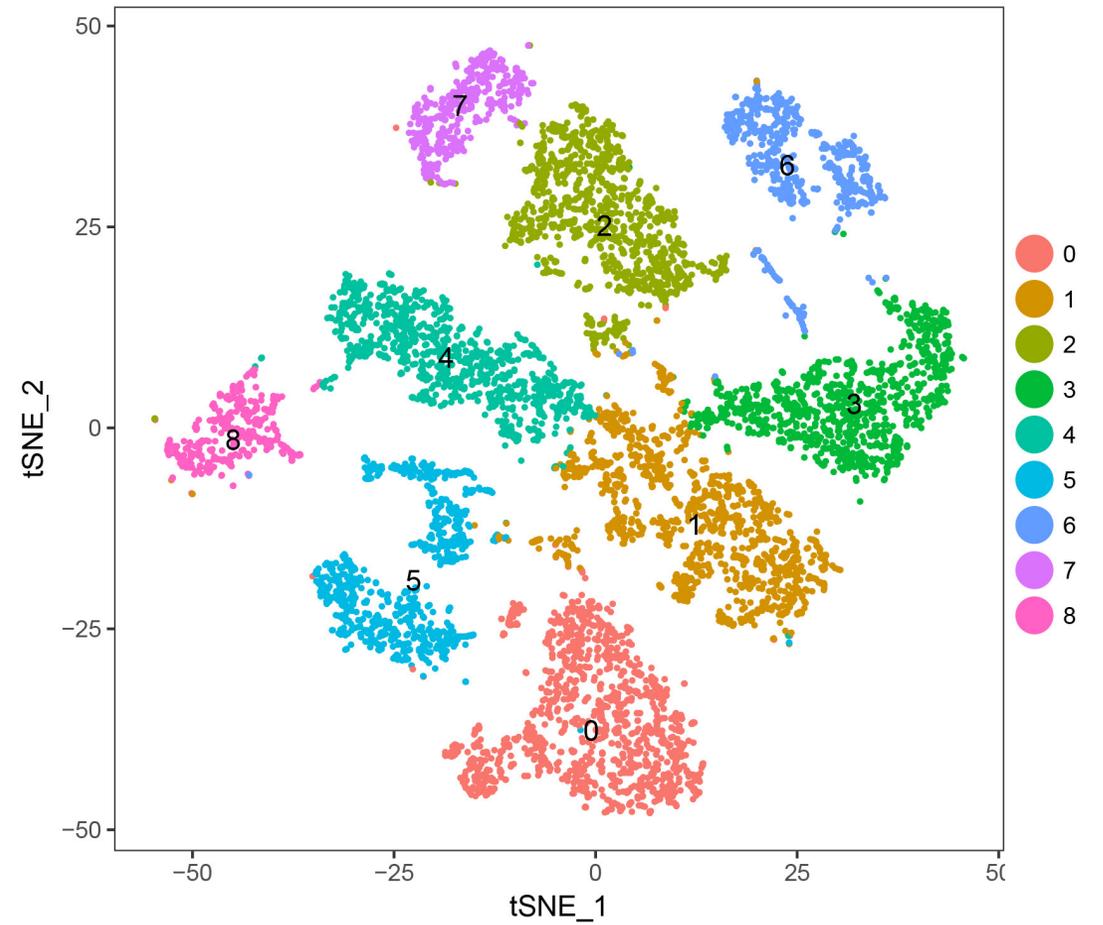
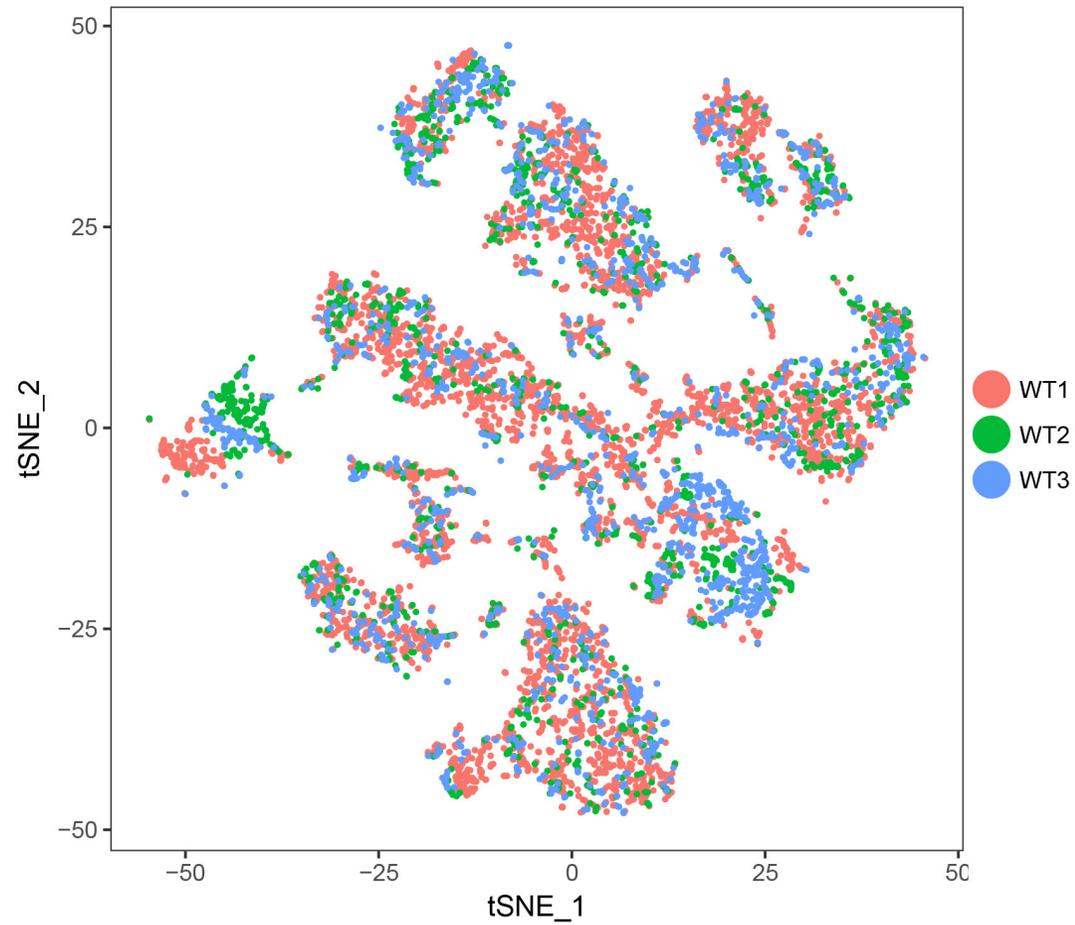
- QC
  - % “Bad” reads pair (invalid ligation product)
    - Self-circle
    - Dangling-end
    - Inter-chromosomal contact (not necessarily bad)
  - % short range / long ranges interaction
- Analysis
  - Interaction matrix
  - Heatmap visualization
- Software:
  - HiC-pro (QC + process)
  - Juicer (visualization)



# Single-cell data analysis (scRNA-seq or scATAC-seq)

- QC:
  - Sequencing quality, sequencing depth, unique UMI, cells with enough read count etc.
- Analysis:
  - Cell – feature count matrix
  - Dimensionality reduction
  - Clustering vs. t-SNE/UMAP visualization
  - Downstream analysis

# Clustering vs. Visualization



# Downstream analysis and integration

1. DNA sequences at the peaks: motif discovery
2. Annotation of the peaks
3. Integration with other omics data/information for functional analyses

# Position weight matrix (PWM) representation of DNA sequence motifs

GAGGTAAAC  
 TCCGTAAGT  
 CAGGTTGGA  
 ACAGTCAGT  
 TAGGTCATT  
 TAGGTACTG  
 ATGGTAACT  
 CAGGTATAC  
 TGTGTGAGT  
 AAGGTAAGT

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$



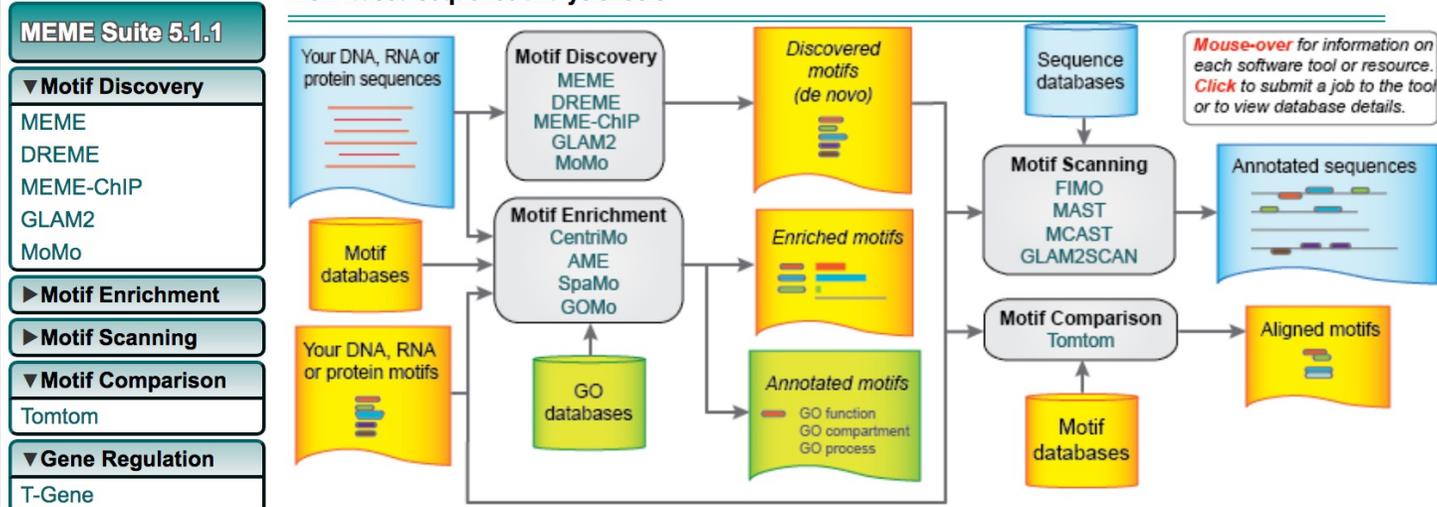
$$R_i = \log_2(4) - H_i$$

$$H_i = - \sum_b f_{b,i} \times \log_2 f_{b,i}$$

# MEME ( meme-suite.org )

## The MEME Suite

Motif-based sequence analysis tools



- MEME Suite 5.1.1**
- ▼ **Motif Discovery**
  - MEME
  - DREME
  - MEME-ChIP
  - GLAM2
  - MoMo
- ▶ **Motif Enrichment**
- ▶ **Motif Scanning**
- ▼ **Motif Comparison**
  - Tomtom
- ▼ **Gene Regulation**
  - T-Gene
- ▶ **Manual**
- ▶ **Guides & Tutorials**
- ▶ **Sample Outputs**
- ▶ **File Format Reference**
- ▶ **Databases**
- ▶ **Download & Install**
- ▶ **Help**
- ▶ **Alternate Servers**
- ▶ **Authors & Citing**
- ▶ **Recent Jobs**
- ◀ **Previous version 5.1.0**

<b>MEME</b> Multiple Em for Motif Elicitation	<b>CentriMo</b> Local Motif Enrichment Analysis	<b>FIMO</b> Find Individual Motif Occurrences
<b>DREME</b> Discriminative Regular Expression Motif Elicitation	<b>AME</b> Analysis of Motif Enrichment	<b>MAST</b> Motif Alignment & Search Tool
<b>MEME-ChIP</b> Motif Analysis of Large Nucleotide Datasets	<b>SpaMo</b> Spaced Motif Analysis Tool	<b>MCAST</b> Motif Cluster Alignment and Search Tool
<b>GLAM2</b> Gapped Local Alignment of Motifs	<b>GOMo</b> Gene Ontology for Motifs	<b>GLAM2Scan</b> Scanning with Gapped Motifs
<b>MoMo</b> Modification Motifs	<b>Tomtom</b> Motif Comparison Tool	<b>GT-Scan</b> Identifying Unique Genomic Targets
<b>T-Gene</b> Predicting Target Genes		

# HOMER ( [homer.ucsd.edu](http://homer.ucsd.edu) )

← → ↻ ⓘ Not Secure | [homer.ucsd.edu/homer/introduction/basics.html](http://homer.ucsd.edu/homer/introduction/basics.html) ☆ C ⋮



## HOMER

Software for motif discovery and ChIP-Seq analysis

### Introduction to HOMER

The best way to learn about HOMER is to go through the tutorial pages. We've tried to spell out what happens in each step and explain the "why". A brief description of the Motif Finding component of HOMER is found below. Explanation of the sequencing analysis components of HOMER are integrated into the tutorials.

### General Introduction to Motif Discovery with HOMER

HOMER is a collection of tools that are commonly needed for the analysis of gene expression profiling (microarray) and genome-wide location analysis experiments (ChIP-Seq or ChIP-Chip). There are also routines for other types of sequencing experiments, such as DNase-Seq or GRO-Seq.

Some of the things HOMER does NOT DO is find differentially expressed genes (although it has some routines to help with this), cluster gene expression profiles, or search for all the instances Transfac motifs in order to make you hopelessly confused!!! The idea was not to completely reinvent the wheel if possible.

Unfortunately, HOMER must be run as a command-line tool, and may be difficult to use if you are new to UNIX. While commands have been distilled to be as simple and user-friendly as possible, basic knowledge of the UNIX environment and file system is critical (but can probably be learned quickly after typing `unix tutorial` into google). I am proud to say that many of the people using HOMER are completely new to UNIX, so it is indeed possible. In addition, a spreadsheet program (i.e. EXCEL) is needed to graph and visualize some of the results produced by HOMER.

Below is a description of how motif analysis is executed with HOMER. Documentation describing the steps of analysis for [Next-Gen Sequencing](#) (or genomic position analysis) or [Microarrays](#) (gene-based analysis) are covered in separate sections.

### *De Novo* Motif Discovery Strategy

# GREAT ( great.stanford.edu )

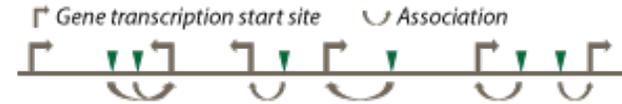
GREAT predicts functions of *cis*-regulatory regions.

- Input:** A set of Genomic Regions (such as transcription factor binding events identified by ChIP-Seq).

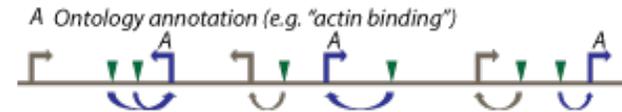
Example:  $\nabla$  SRF ChIP-Seq called peaks



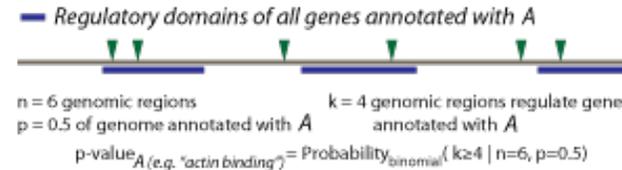
- GREAT associates both proximal and distal input Genomic Regions with their putative target genes.



- GREAT uses gene Annotations from numerous ontologies to associate genomic regions with annotations.



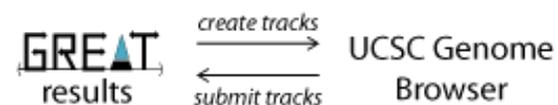
- GREAT calculates statistical Enrichments for associations between Genomic Regions and Annotations.



- Output:** Annotation terms that are significantly associated with the set of input Genomic Regions.

	Ontology term	p-value
SRF peaks regulate genes involved in:	Actin cytoskeleton	$10^{-9}$
	FOS gene family	$10^{-8}$
	TRAIL signaling	$10^{-7}$

- Users can create UCSC custom tracks from term-enriched subsets of Genomic Regions. Any track can be directly submitted to GREAT from the UCSC Table Browser.



# ChIPseeker: an R/Bioconductor package



Search:

[Home](#) [Install](#) [Help](#) [Developers](#) [About](#)

[Home](#) » [Bioconductor 3.10](#) » [Software Packages](#) » ChIPseeker

## ChIPseeker

platforms **all** rank **123 / 1823** posts **2 / 0 / 1 / 0** in Bioc **6 years**  
build **warnings** updated **since release** dependencies **152**

DOI: [10.18129/B9.bioc.ChIPseeker](https://doi.org/10.18129/B9.bioc.ChIPseeker) [f](#) [t](#)

### ChIPseeker for ChIP peak Annotation, Comparison, and Visualization

Bioconductor version: Release (3.10)

This package implements functions to retrieve the nearest genes around the peak, annotate genomic region of the peak, statistical methods for estimate the significance of overlap among ChIP peak data sets, and incorporate GEO database for user to compare the own dataset with those deposited in database. The comparison can be used to infer cooperative regulation and thus can be used to generate hypotheses. Several visualization functions are implemented to summarize the coverage of the peak experiment, average profile and heatmap of peaks binding to TSS regions, genomic annotation, distance to TSS, and overlap of peaks or genes.

Author: Guangchuang Yu [aut, cre] , Yun Yan [ctb], Hervé Pagès [ctb], Michael Kluge [ctb], Thomas Schwarzl [ctb], Zhougeng Xu [ctb]

Maintainer: Guangchuang Yu <guangchuangyu at gmail.com>

Citation (from within R, enter `citation("ChIPseeker")`):

Yu G, Wang L, He Q (2015). "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization." *Bioinformatics*, **31**(14), 2382-2383. doi: [10.1093/bioinformatics/btv145](https://doi.org/10.1093/bioinformatics/btv145).

#### Documentation »

*Bioconductor*

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

#### Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

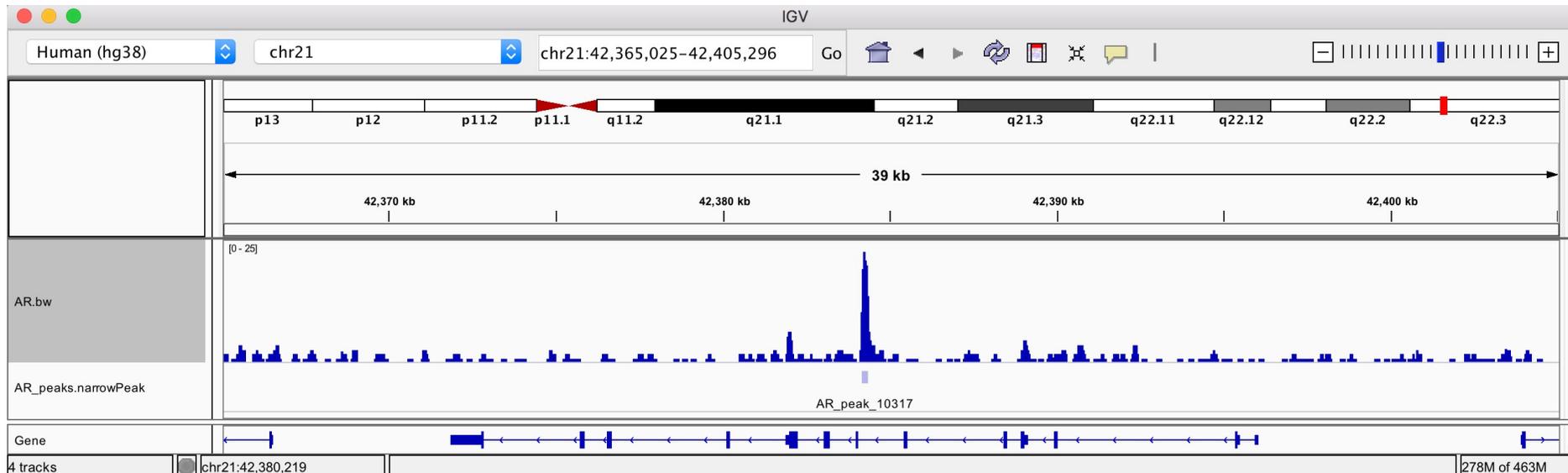
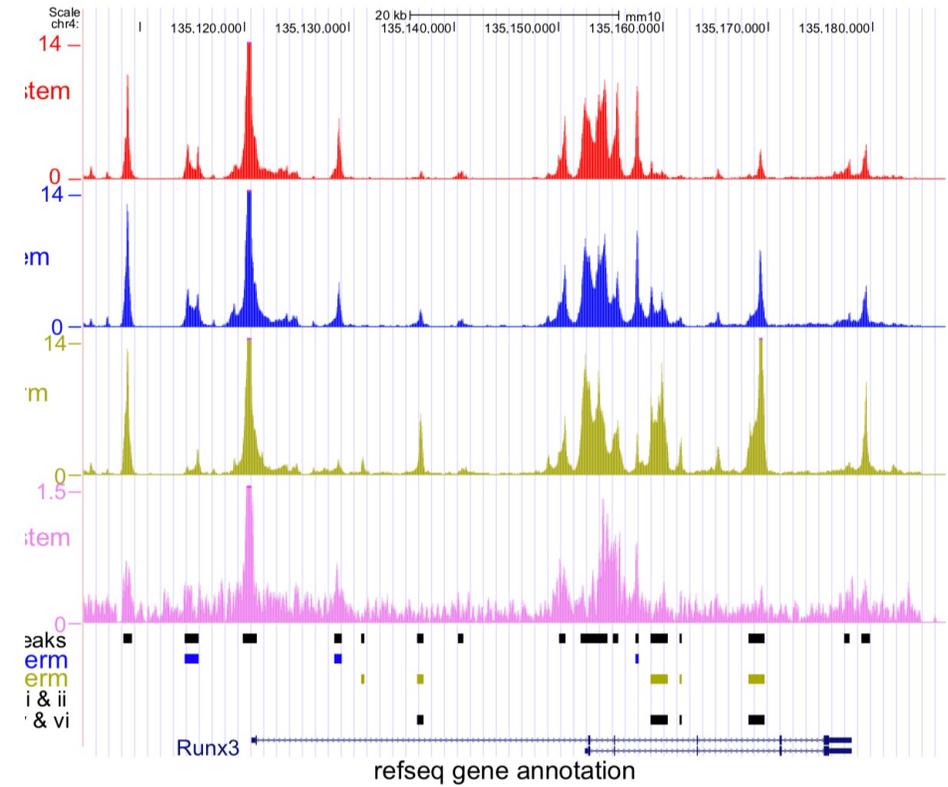
- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package developers

# ChIP-seq: Downstream analyses

- Data visualization
  - UCSC genome browser: <http://genome.ucsc.edu/>
  - WashU epigenome browser: <http://epigenomegateway.wustl.edu/>
  - IGV: <http://software.broadinstitute.org/software/igv/>
- Integration with gene expression
  - BETA: <http://cistrome.org/BETA/>
- Integration with other epigenomic data
  - BART: <http://bartweb.org/>
  - MARGE: <http://cistrome.org/MARGE/>
  - GREAT: <http://great.stanford.edu>
  - ENCODE SCREEN: <http://screen.umassmed.edu/>

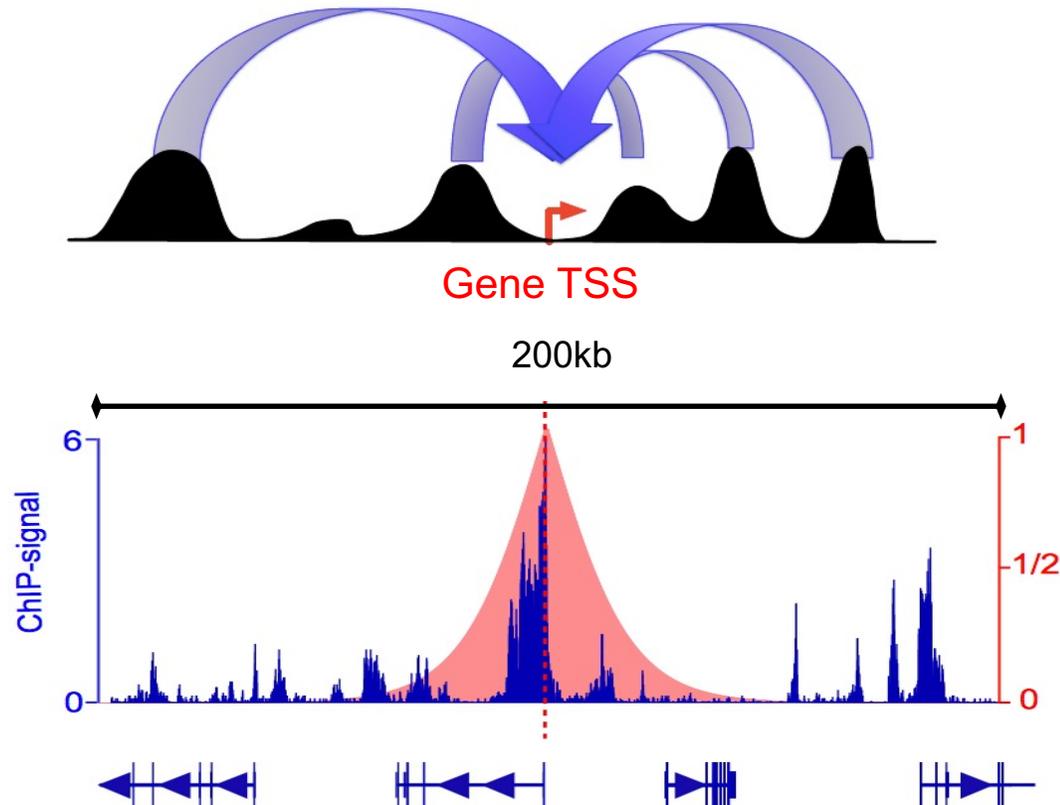
# Data Visualization

- bedGraph to bigWig
- macs2 output data
- IGV, UCSC, etc.



# BETA

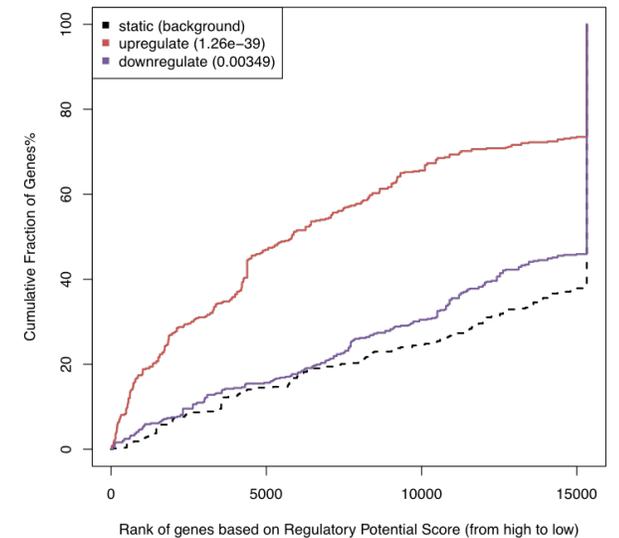
- Binding Expression Target Analysis (*Wang et al. Nat Protoc. 2013*)
- defines Regulatory Potential on each gene



$$P_i = \sum_{|j| < 10^5} W_j Z_j$$

$$W_j = \frac{2 \exp(-\alpha |j|)}{1 + \exp(-\alpha |j|)}$$

$$\alpha = \frac{\log 3}{10^4}$$



Wang, et al. *Nat Protoc* 2013  
 Wang, Zang et al. *Genome Res* 2016  
 Qin et al. *Genome Bio* 2020

# Galaxy: web-interface analysis platform

- <https://usegalaxy.org/>

The screenshot shows the Galaxy web interface. At the top, there is a navigation bar with the Galaxy logo, a search bar, and links for 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', 'Login or Register', and a user profile icon. The main content area features a central announcement for the JXTX James P. Taylor Foundation, which includes a logo with sneakers and text about the foundation's mission and a 'Donate Now' button. To the left is a sidebar with 'Tools' and various tool categories like 'GENERAL TEXT TOOLS', 'GENOMIC FILE MANIPULATION', and 'COMMON GENOMICS TOOLS'. To the right is a 'History' sidebar showing an empty history and a message to load data from an external source. A blue banner at the bottom of the main content area provides information about SARS-CoV-2 data analysis resources.

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

**JXTX**  
James P. Taylor Foundation  
Design by Rebekka Paisner

**James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.**

To ensure implementation of this vision, the Galaxy community has established a foundation—JXTX: The James P. Taylor Foundation for Open Science. The JXTX Foundation's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

To make this happen we are accepting contributions. More details can be found on [the @jxtx page in the Galaxy Hub](#). Please, help us continue what James has started.

[Donate Now](#)

**Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at [covid19.galaxyproject.org](https://covid19.galaxyproject.org)**

**PennState** **JOHNS HOPKINS UNIVERSITY** **OREGON HEALTH & SCIENCE UNIVERSITY** **TACC** **CYVERSE™**

The Galaxy Team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State, the Department of Biology at Johns Hopkins University and the Computational Biology Program at Oregon Health & Science University.

This instance of Galaxy is utilizing infrastructure generously provided by CyVerse at the Texas Advanced Computing Center, with support from the National Science Foundation.

# Run MACS on Cistrome, a Galaxy-based platform

- <http://cistrome.org/ap/>

The screenshot displays the Cistrome Galaxy web interface. The main content area is titled 'Upload File (version 1.1.4)'. It features a 'File Format' dropdown menu set to 'Auto-detect', with a note: 'Which format? If for expression data, choose cel.zip or xys.zip. See help below'. Below this is a 'File (Please avoid Windows format text file):' section with a 'Choose File' button and the text 'No file chosen'. A tip explains that files larger than 2GB will fail and suggests using the URL method or ASPERA. There is also a 'URL/Text:' section with a text input area and a note: 'Here you may specify a list of URLs (one per line) or paste the contents of a file.' A table titled 'Files uploaded via ASPERA:' is currently empty, with a message: 'Your ASPERA upload directory contains no files.' Below the table is a 'Convert spaces to tabs:' section with a 'Yes' checkbox and a note: 'Use this option if you are entering intervals by hand.' The 'Genome:' dropdown is set to 'Human Dec. 2013 (GRCh38/hg38) (hg38)'. An 'Execute' button is at the bottom of the form.

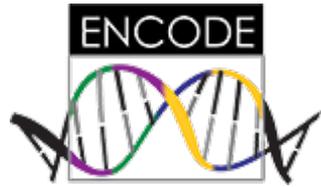
The left sidebar contains tool categories: 'CISTROME TOOLBOX' (Import Data, Upload File, CistromeFinder, CistromeCR, Expression CEL file packager, GenomeSpace import) and 'GALAXY TOOLBOX' (Get Data, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences).

The right sidebar shows a 'History' panel with a list of 10 items, each with a name, an eye icon, an edit icon, and a delete icon. The items are: 68: Heatmap log, 67: Heatmap k-means clustered regions, 66: Heatmap R script, 65: Heatmap image, 64: Heatmap log, 63: Heatmap k-means clustered regions, 62: Heatmap R script, 61: Heatmap image, 60: Heatmap log, 59: Heatmap k-means clustered regions, 58: Heatmap R script, 57: Heatmap image, 56: Heatmap log, and 55: Heatmap k-means clustered regions.

# SICER2

- <https://zanglab.github.io/SICER2/>

The screenshot shows a web browser window with the URL `zanglab.github.io/SICER2/`. The page features a blue header with the text "SICER2 Documentation" and a search bar labeled "Search docs". A left sidebar contains a navigation menu with the following items: "Quick Start", "SICER2", "Introduction", "Installation", "Using SICER2", "Using SICER2 for differential peak calling", "Example Use", "Workflow of SICER2", "Understanding SICER2 Outputs", and "Contact". The main content area displays the breadcrumb "Docs » Quick Start" and a link to "Edit on GitHub". The title "SICER2" is prominently displayed, followed by the subtitle "Redesigned and improved ChIP-seq broad peak calling tool SICER". A status bar indicates "build passing". Below this, there is a link to the "GitHub Repo" and a section titled "Introduction". The introduction text states: "Chromatin immunoprecipitation combined with high-throughput sequencing (ChIP-seq) can be used to map binding sites of a protein of interest in the genome. Histone modifications usually occupy broad chromatin domains and result in diffuse patterns in ChIP-seq data that make it difficult to identify signal enrichment. SICER, a spatial clustering approach for the identification of ChIP-enriched regions, was developed for calling broad peaks from ChIP-seq data." A second paragraph explains: "Usability of the original SICER software has been affected by increased throughputs of ChIP-seq experiments over the years. We now present SICER2 a more user-friendly version of SICER that has been redesigned and streamlined to handle large ChIP-seq data sets. This new Python package supports multiple job submissions on cluster systems and parallel processing on multicore architectures."



# ENCODE

<https://www.encodeproject.org/>

ENCODE Data Encyclopedia Materials & Methods Help

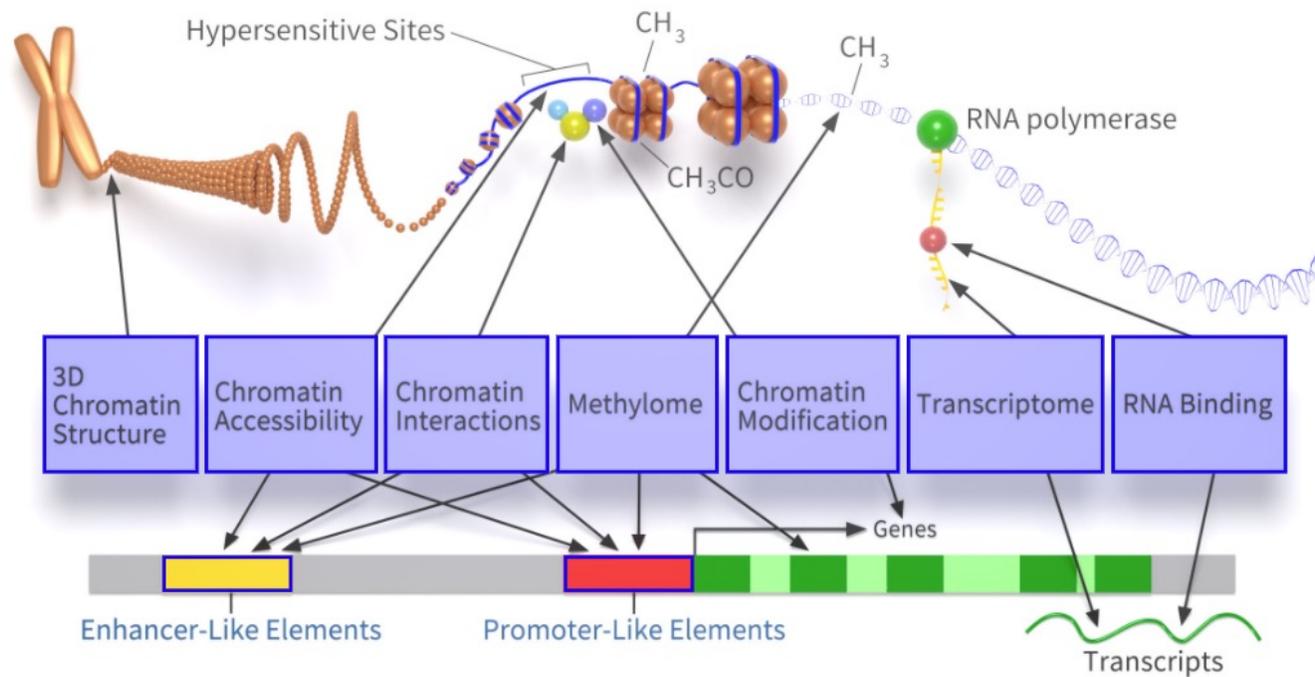
New >>

Search...



Sign in / Create account

## ENCODE: Encyclopedia of DNA Elements



About ENCODE Project

Getting Started

Experiments

Search ENCODE portal ?

ENCODE Q

Functional Characterization Experiments

About ENCODE Encyclopedia

candidate Cis-Regulatory Elements

Search for candidate Cis-Regulatory Elements ?

Hosted by SCREEN

Human GRCh38 Q

Mouse mm10 Q

[Visit hg19 site](#)

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

HUMAN

MOUSE

WORM

FLY

# Cistrome Data Browser

<http://cistrome.org/db/>

The screenshot shows the Cistrome Data Browser website. At the top, there is a navigation bar with links for Home, Documentation, About, Statistics, Batch download, ToolKit, Cistrome-GO, and Liu Lab. Below this is a large blue banner with the Cistrome logo and the text "Cistrome Data Browser".

A "Tips" section contains the following advice:

- Check what factors regulate your gene of interest, what factors bind in your interval or have a significant binding overlap with your peak set. Have a try at [CistromeDB Toolkit](#).
- If you have a Transcription Factor ChIP-seq (and TF perturbed expression) data, [Cistrome-GO](#) help you predict the function of this TF.
- Please help us curate the samples which has incorrect meta-data annotation by clicking the button on the inspector page. Thank you!

The search interface includes a "Containing word(s):" input field, a "Search" button, and an "Options" dropdown menu.

Below the search bar are three filter panels:

- Species:** All (selected), Homo sapiens, Mus musculus
- Biological Sources:** All (selected), 1-cell pronuclei, 1015c, 10326, 1064Sk, 106A
- Factors:** All (selected), AATF, ABCC9, ACSS2, ACTB, ADNP

The "Results" section displays a table with the following columns: Batch, Species, Biological Source, Factor, Publication, and Quality Control.

Batch	Species	Biological Source	Factor	Publication	Quality Control
<input type="checkbox"/>	Homo sapiens	HeLa; Epithelium; Cervix	BTAF1	Johannes F, et al. Bioinformatics 2010	●●●●●●

Mei *et al.* *Nucleic Acids Res.* 2017  
Zheng *et al.* *Nucleic Acids Res.* 2018

# BART web: inferring transcriptional regulators from a variety of inputs

## BART: Binding Analysis for Regulation of Transcription

User input

Gene list

ChIP-seq

Region set

Hi-C maps

>1000 H3K27ac ChIP-seq

Adaptive Lasso regression

Mapping

**Cis-regulatory profile**

Cis-regulatory element repertoire  
(2.7 million in the human genome,  
1.5 million in the mouse genome)

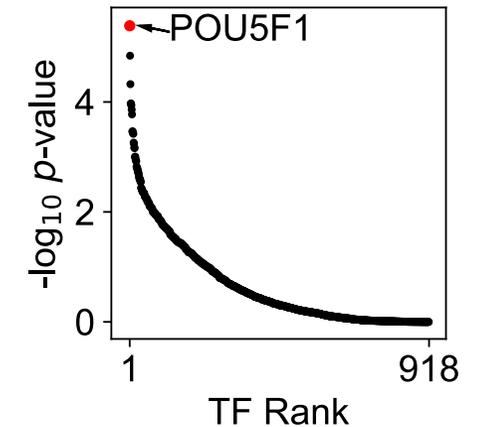
differential  
interaction

>13,000 TF ChIP-seq datasets

ROC associations

Statistical tests,  
Background adjustment,  
Irwin-Hall rank integration

Output prediction



# Limitations of NGS for epigenetics research

- Dependent on assays (e.g., antibody availability and quality for ChIP-seq)
- Semi-quantitative: does not detect global change
- Needs many cells – difficult for clinical samples
- Cellular heterogeneity

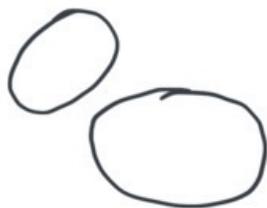
# Take-home messages

- Why am I learning these if I am not a bioinformatician?
  - Help improve experimental design
  - Quality control
  - Better interpret the experimental data
  - Take advantage of existing tools and data resources

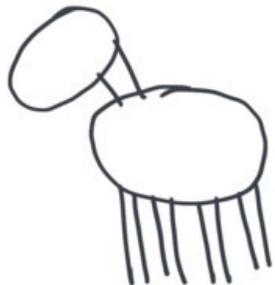
# HOW TO: DRAW A HORSE

BY VAN OKTOP

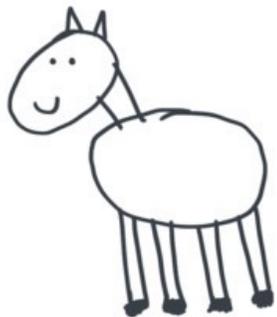
---



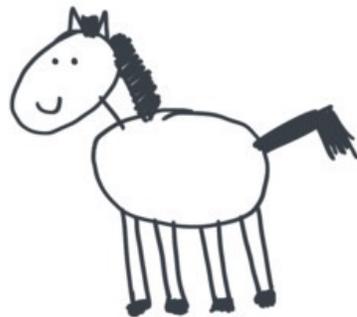
① DRAW 2 CIRCLES



② DRAW THE LEGS

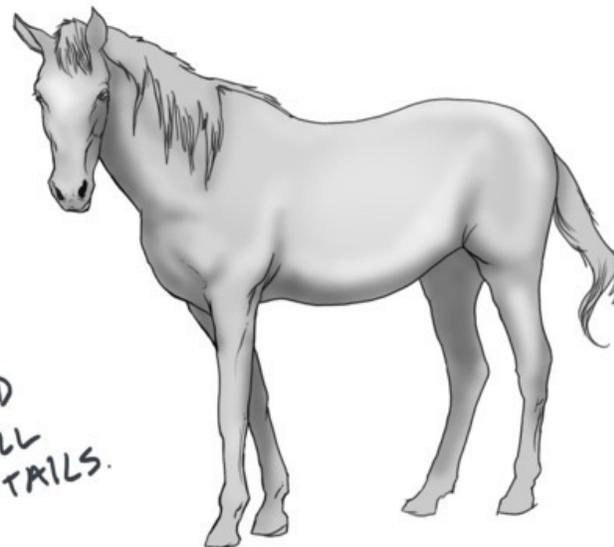


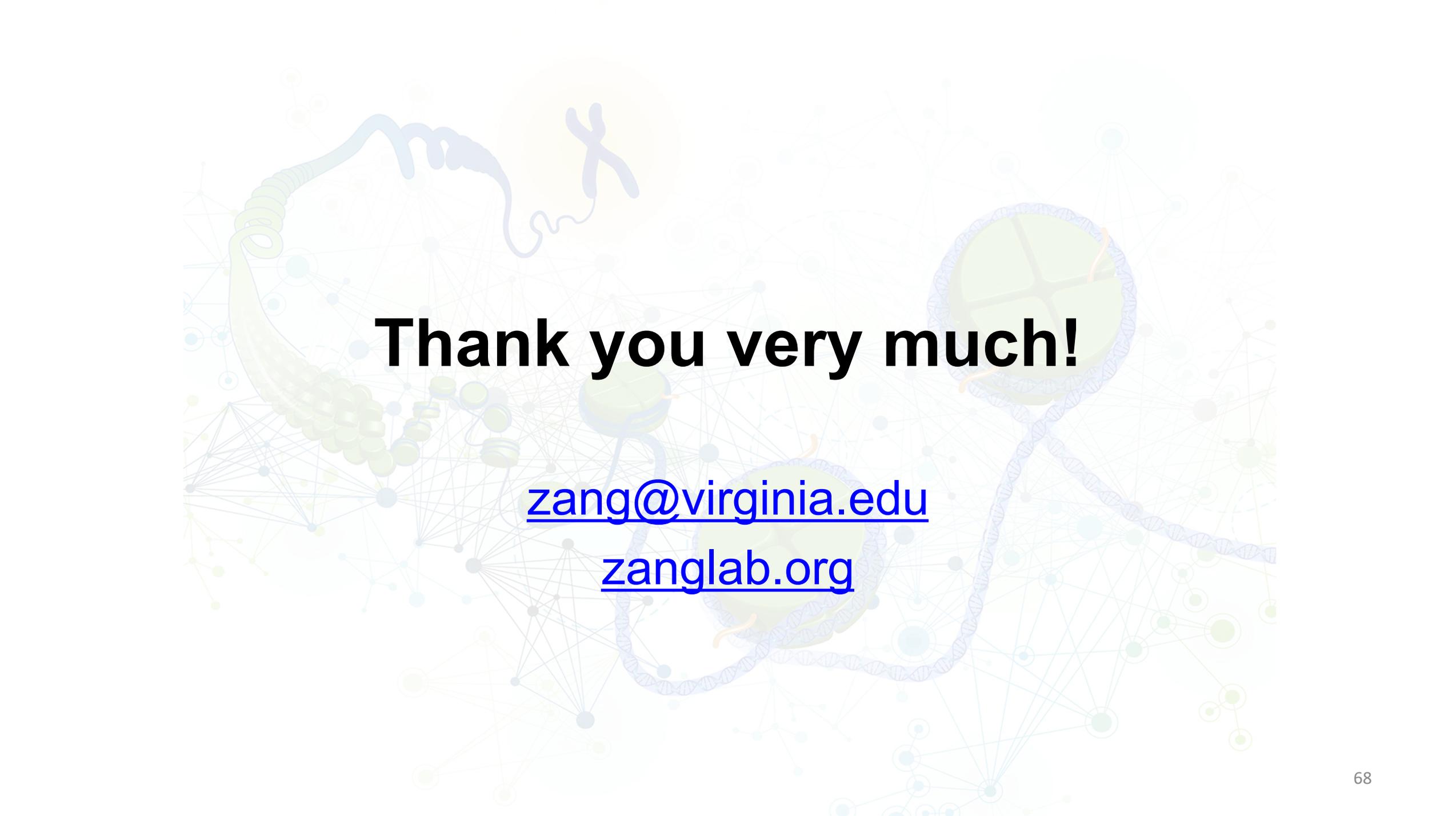
③ DRAW THE FACE



④ DRAW THE HAIR

⑤  
ADD  
SMALL  
DETAILS.





**Thank you very much!**

[zang@virginia.edu](mailto:zang@virginia.edu)

[zanglab.org](http://zanglab.org)