

# ChIP-Seq: Using High-Throughput Sequencing to Discover Protein-DNA Interactions

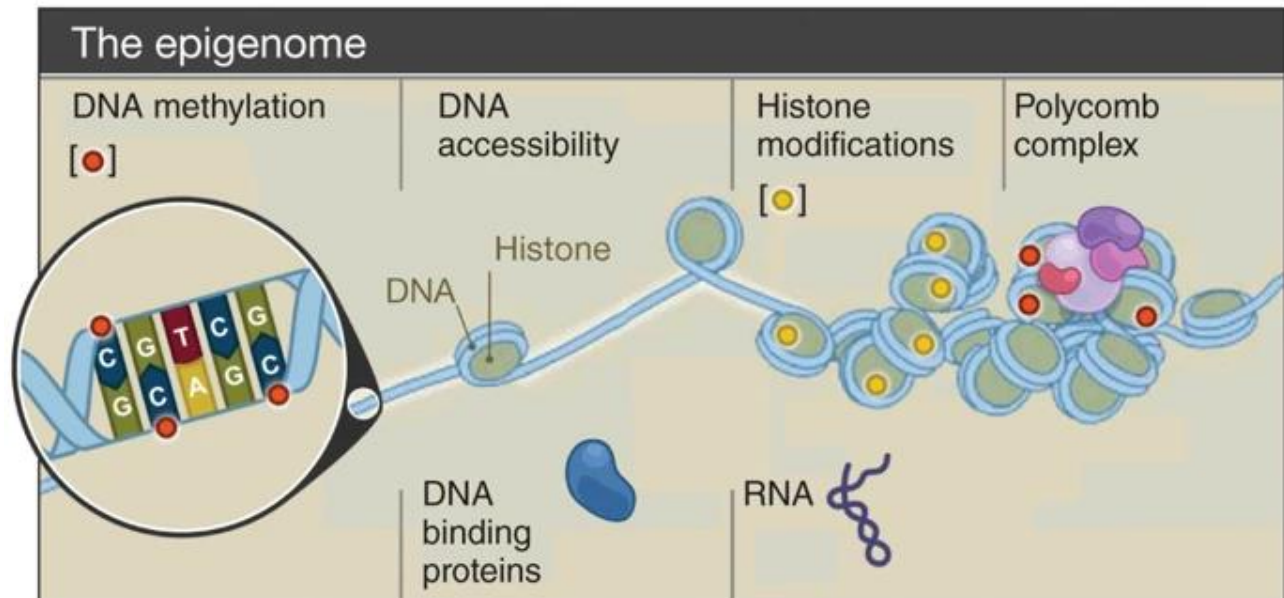
Dena Leshkowitz

Introduction to Deep Sequencing Analysis  
for Biologists

Course 2019-2020

# Epigenome

- ✓ How does the same genome sequence give rise to over 200 different cell types through remarkably consistent differentiation programs?
- ✓ Gene regulation and genome function are intimately related to the physical organization of genomic DNA
- ✓ Epigenetic modifications are **reversible** modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence.





# Definition

- ChIP-seq is short for chromatin immuno-precipitation followed by sequencing
- It provides quantitative, genome-wide view of DNA- protein binding events

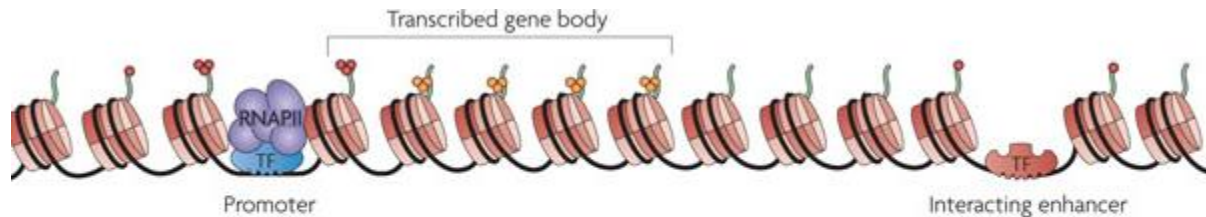
# ChIP-Seq - Lecture Outline

- Experimental issues: how is a ChIP-Seq experiment done?
- Analysis of the sequence data: how to detect the binding regions?
- Downstream analysis: how to extract the biological relevance?

# Aim: Transcription Regulation

Characterize genome wide DNA-protein interactions in vivo, such as:

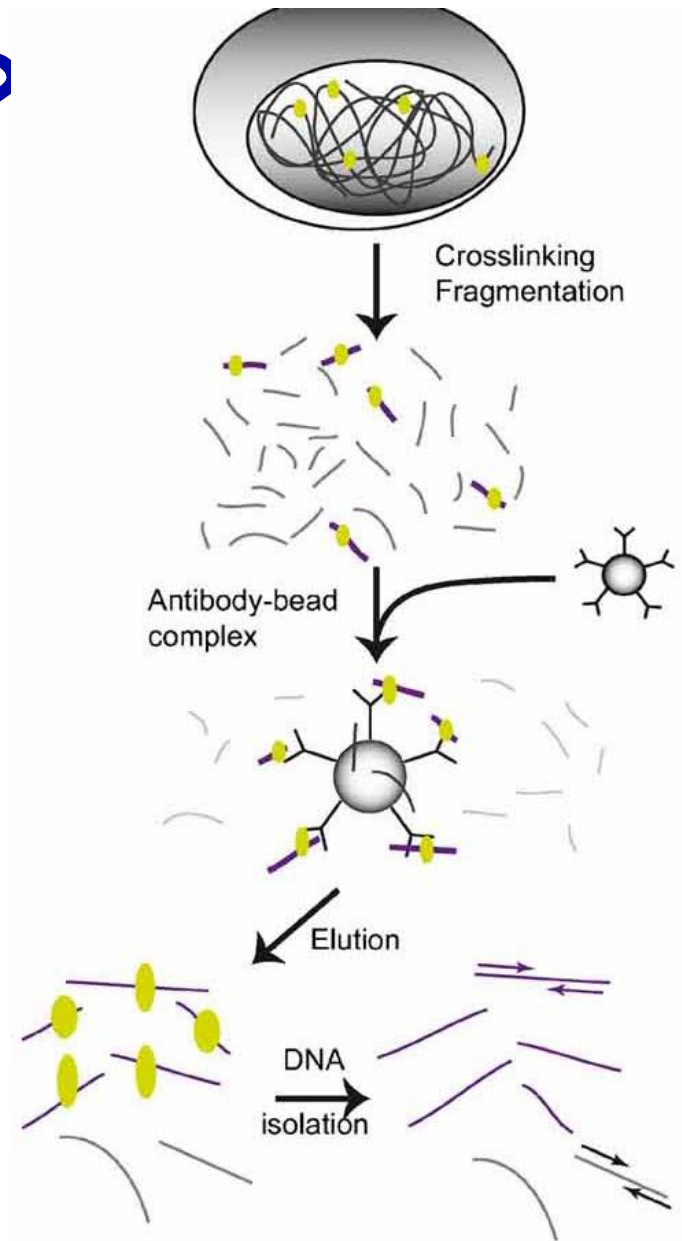
- TF to promoter or enhancer
- RNA polymerase II
- Modified histones



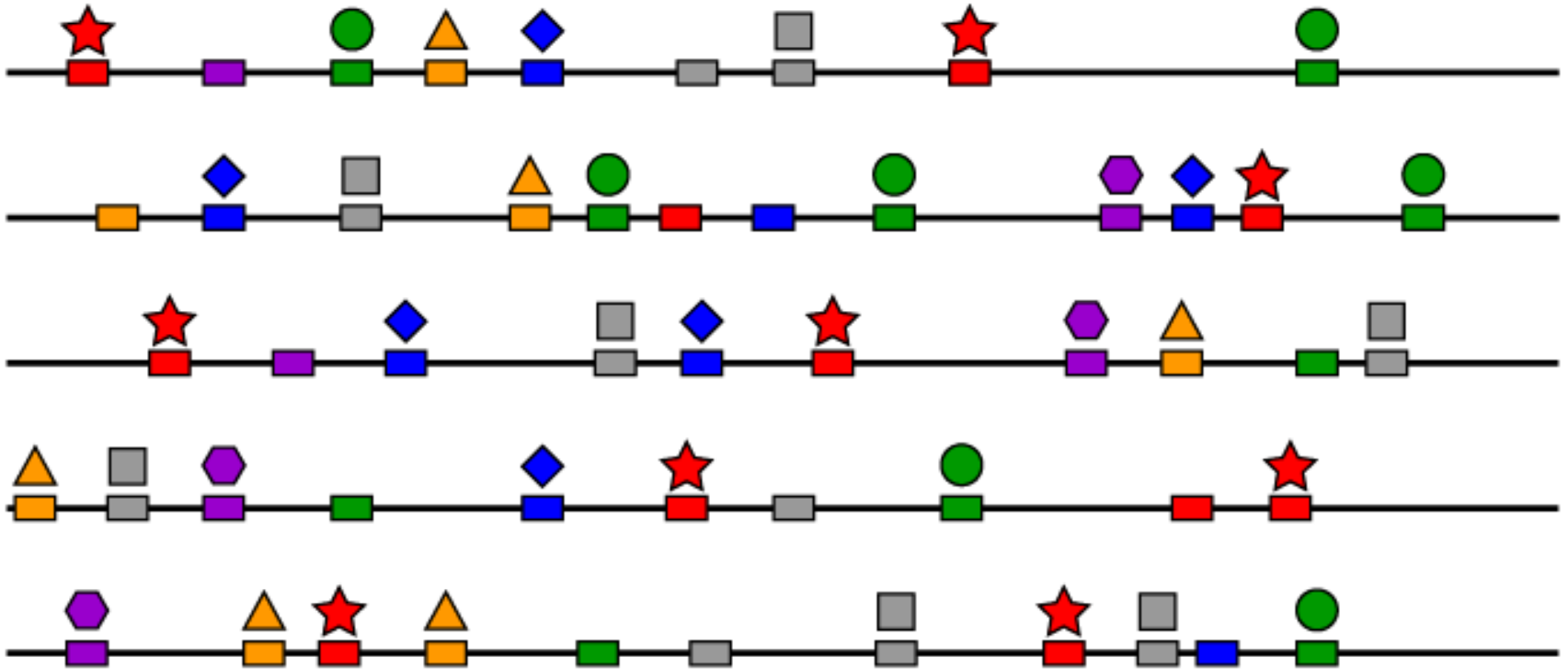
# ChIP - How is it done?

ChIP is a technique that permits to

- “Freeze” the protein-DNA binding events inside the cell nucleus
- Use antibodies to extract the DNA bound by a specific protein

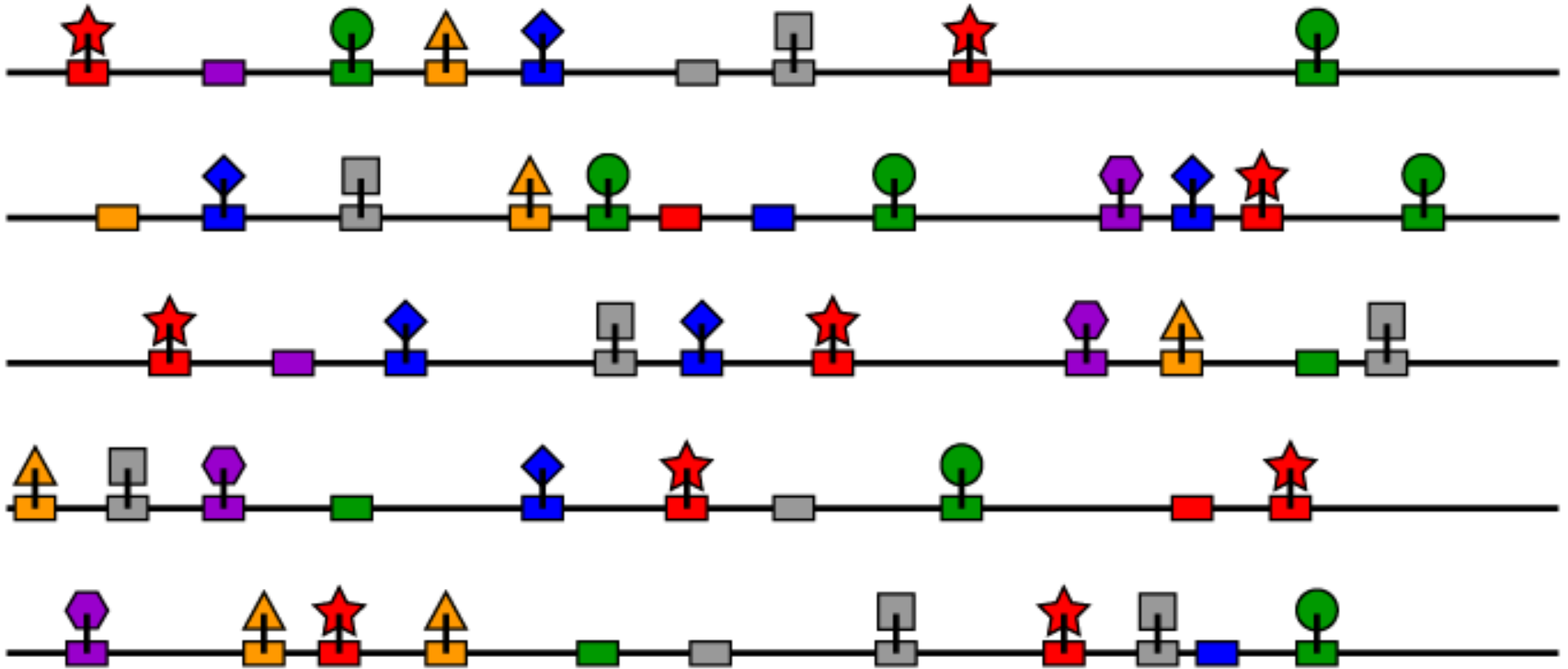


# Chromatin Immunoprecipitation (ChIP)

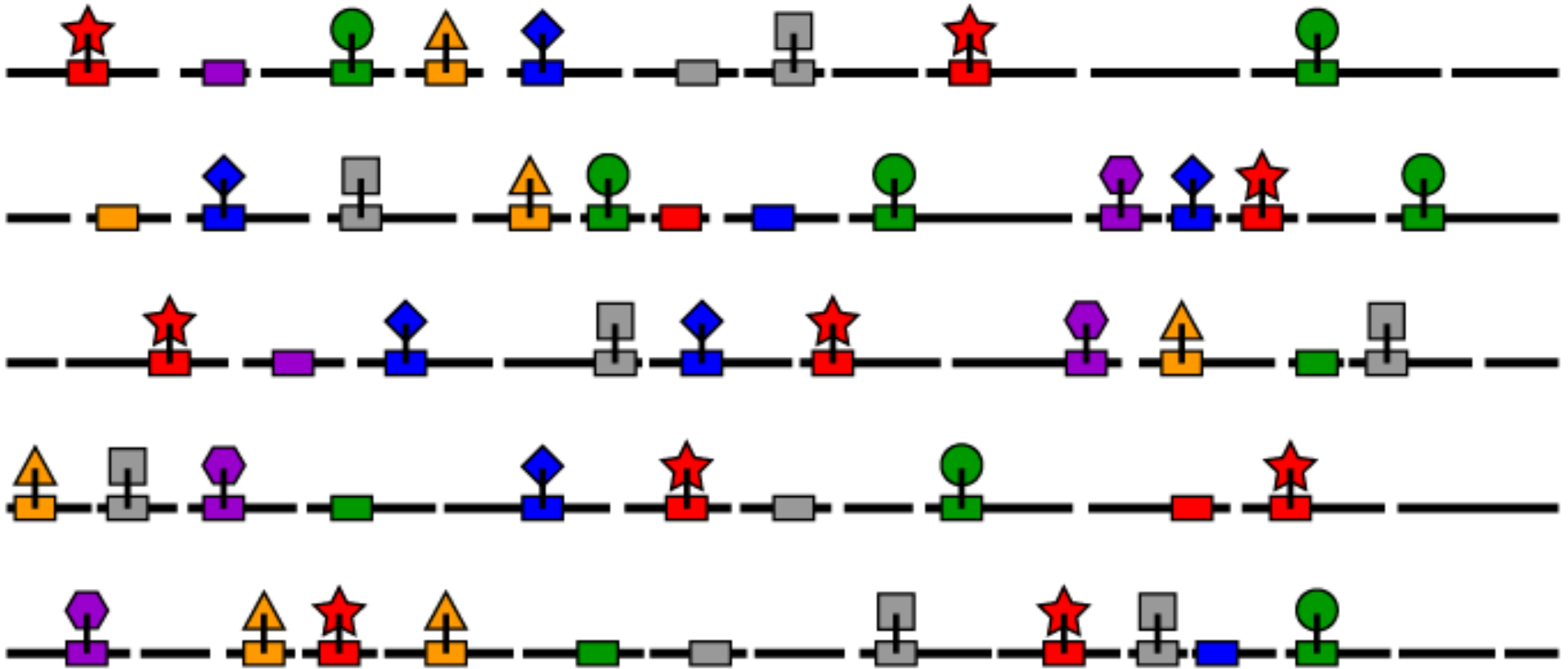




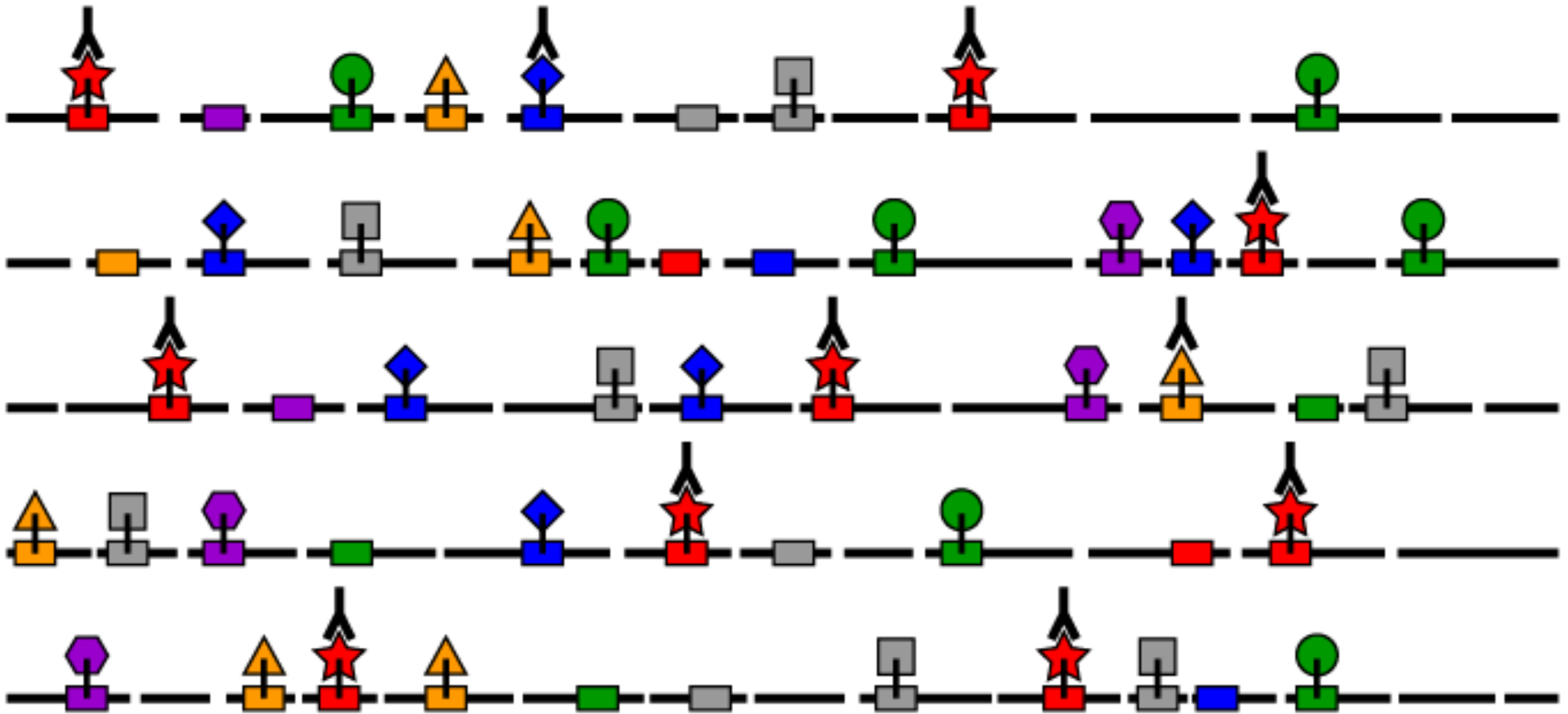
# TF/DNA Crosslinking *in vivo*



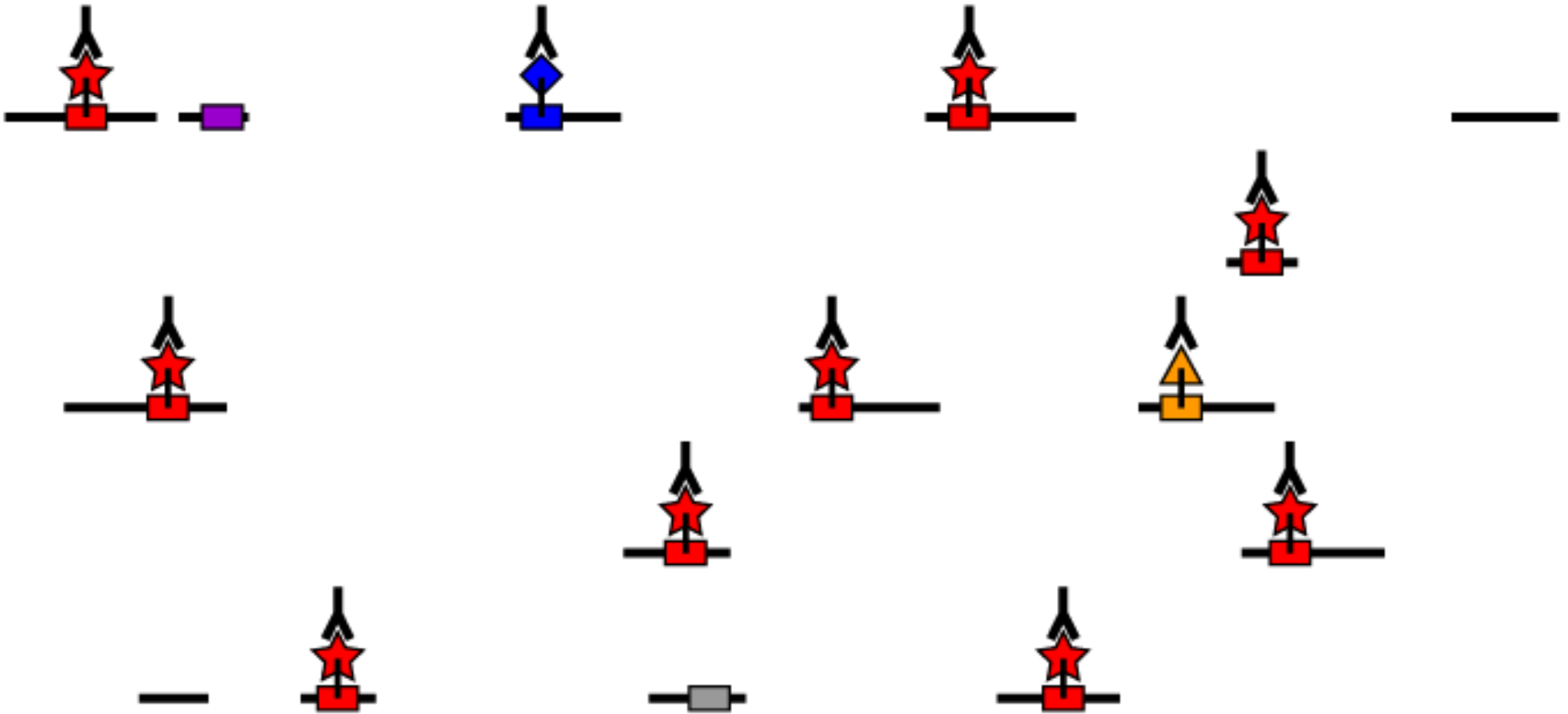
# Sonication (~200bp)



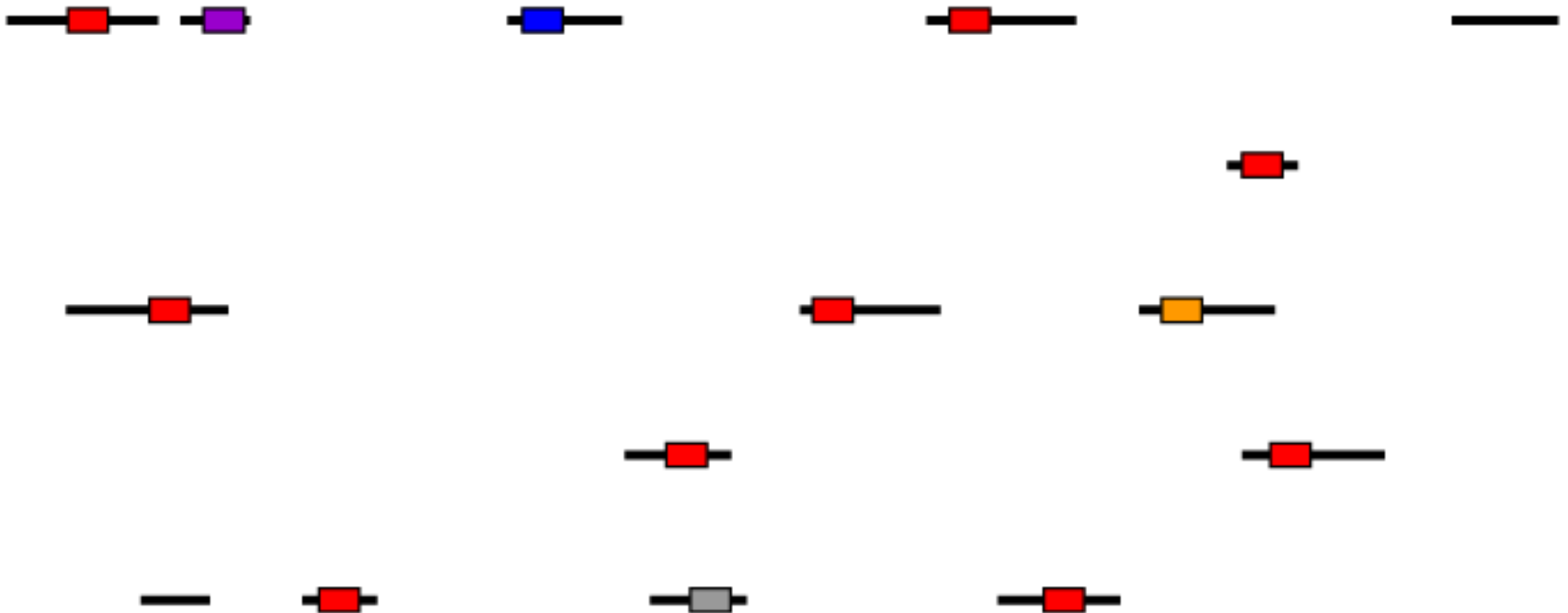
# TF-specific Antibody



# Immunoprecipitation



# Reverse Crosslink and DNA Purification

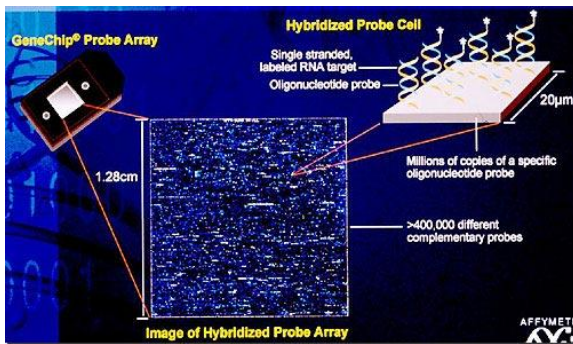


# ChIP High-throughput technology

Discover the DNA binding regions in the genomic scale

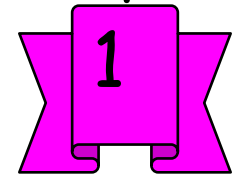
DNA Microarrays

Next Generation Sequencing



More attractive for where promoter chips are not sufficient to study regulation.

NGS in comparison has a wider dynamic range and better base-resolution.



ChIP-Chip

ChIP-Seq

# ENCODE Guidelines

The screenshot shows a web browser window with the URL <https://www.encodeproject.org/about/experiment-guidelines/>. The browser's address bar and tabs are visible at the top. Below the browser window, the ENCODE website header is shown with navigation links: ENCODE, Data, Encyclopedia, Materials & Methods, Help, and a search bar. The main content area features the heading "Current ENCODE Experiment Guidelines". Under this heading, there are two main sections: "Antibody Characterization Guidelines" and "Experiment Guidelines".

## Antibody Characterization Guidelines

- [Antibody Characterization Standards for RNA binding proteins](#) (November 2016)
- [Antibody Characterization Standards for antibodies used in fRIP assays](#) (November 2016)
- [Antibody Characterization Standards for histone modifications and chromatin-associated proteins](#) (October 2016)
- [Antibody Characterization Standards for transcription factors](#) (May 2016)

## Experiment Guidelines

- [ENCODE Experimental Guidelines for ENCODE3 RNA-seq](#) (Includes long and small RNA-seq, CAGE, RAMPAGE) (January 2017)
- [ENCODE Experimental Guidelines for ENCODE3 DNase-seq](#) (January 2017)
- [ENCODE Experimental Guidelines for ENCODE3 eCLIP](#) (January 2017)
- [ENCODE Experimental Guidelines for ENCODE3 ChIP-seq](#) (January 2017)
- [ENCODE Characterization Guidelines for ChIP-seq using Epitope-tagged Transcription Factors](#) (January 2017)
- [ENCODE Experimental Guidelines for ENCODE4 ChIA-PET v2.0](#) (July 2017)
- [ENCODE Standards and Guidelines for Whole Genome Shotgun Bisulfite Sequencing \(WGBS\)](#) (July 2015)
- [ENCODE Experimental Guidelines for ENCODE4 HiC v1.0](#) (January 2018)

For previous versions of experiment and antibody guidelines, click [here](#).

# Experimental Design

- **ENCODE consortium's Standards, Guidelines and Best Practices**

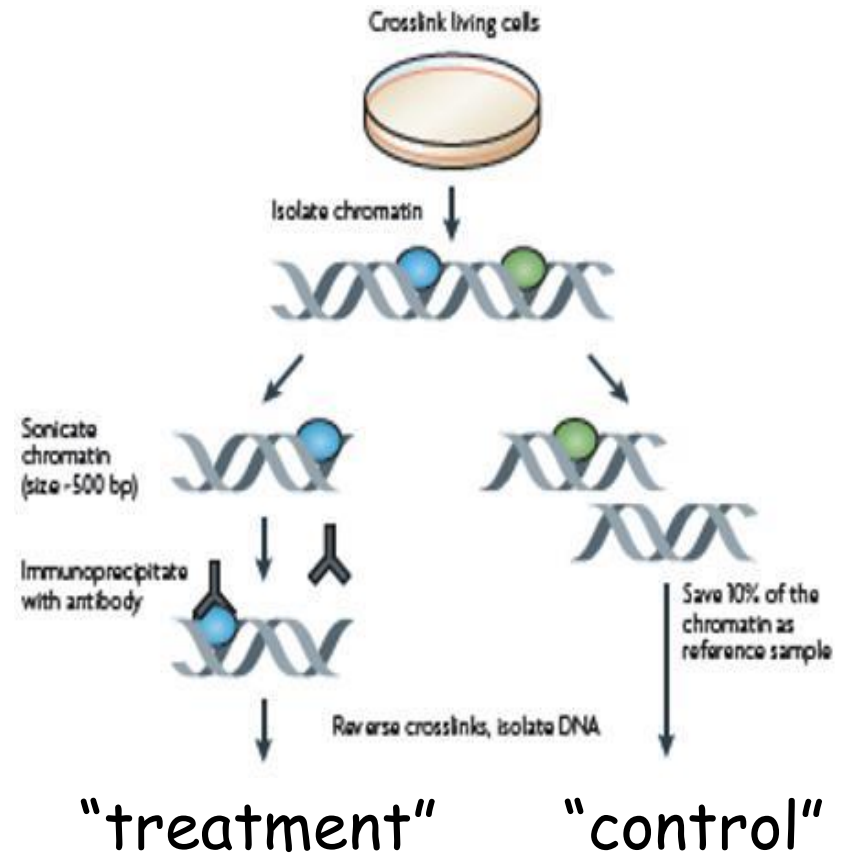
Genome Res. 2012 Sep;22(9):1813-31. doi:  
10.1101/gr.136184.111.

- **Consult with the person who will analyse the data before performing the experiment**
- **Kick-off meeting**



# Types of ChIP Controls

- ✓ "Input" DNA before IP
- ✓ "Mock" IP with no antibody
- ✓ IP with Pre-Immune Serum
- ✓ IP with a non-relevant antibody
- ✓ IP with knock out (cells without the relevant protein)
- ✓ Control should be same cell and condition as the IP in order to account for genetic and epigenetic features

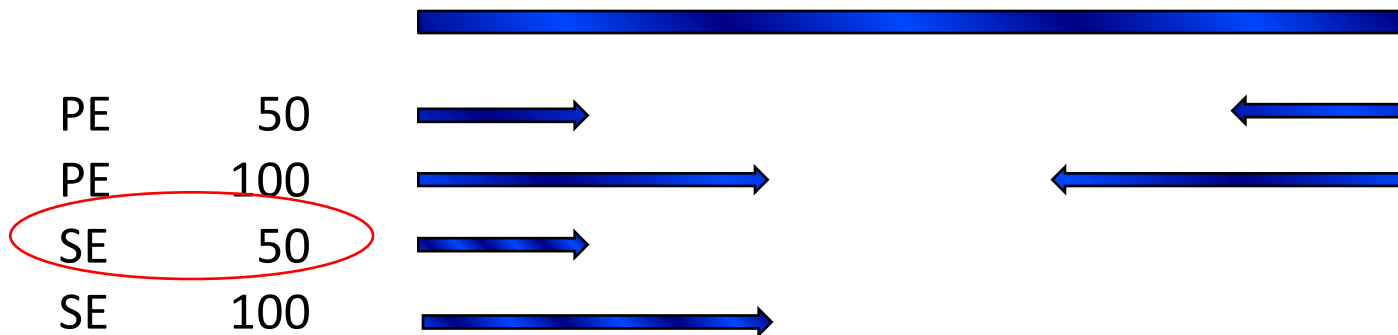


# Why Do I Need a Control Sample?

- Majority of DNA sequenced in a ChIP reaction is background, the IP is an enrichment method
- We will have a lot of **biased** DNA fragments that did not bind our protein of interest
  - DNA that is more prone to breakage (open-chromatin regions)
  - Specific amplified DNA in the genome we sequence (but not in the reference genome)
  - Artificially high signal in some types of repeat regions such as satellite, telomeric and centromeric repeats
  - Other technical or sequence bias
- Need to have a control!

# Other Experimental Design Issues

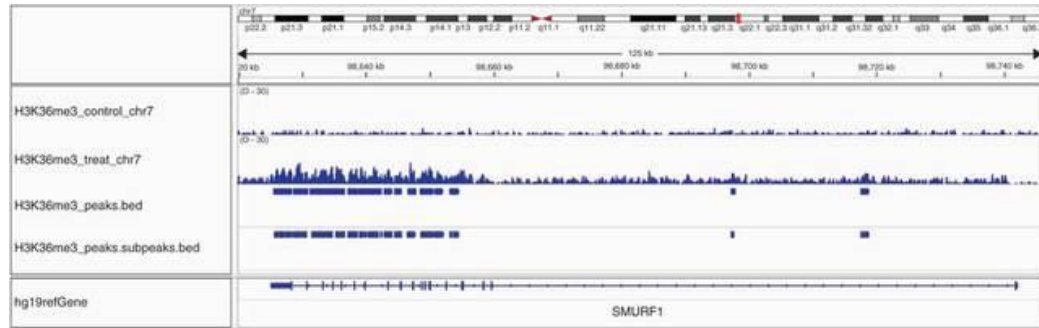
- Use validated antibody maybe even two different antibodies
- **Need to have biological replicates**
- If we have a good reference genome (mouse, human...) no need to sequence long reads (>50 bases) and no need in paired-end sequencing



- How much reads (sequences) per sample? <sup>-19-</sup>

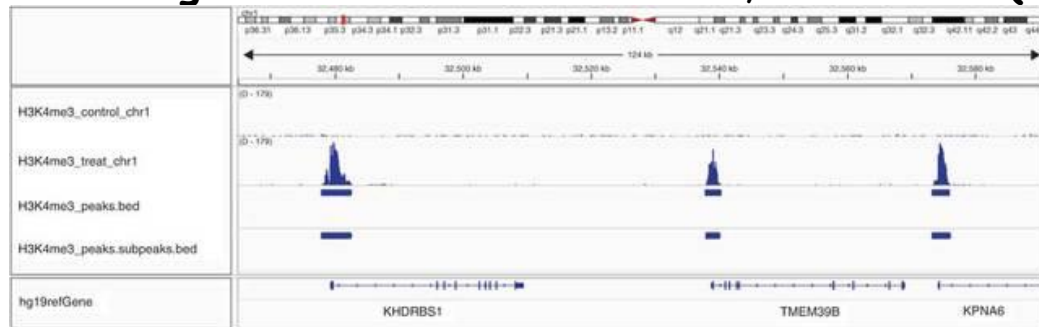
# Coverage Required - Types of Peaks

Broad peaks-  
H3K36me3  
(exon regions)



Feng et al. Nature Protocols 7, 1728-1740 (2012)

Sharp peaks-  
H3K4me3  
(promoter regions)



ENCODE recommendations:

uniquely aligned read	Sharp Peaks	Broad peaks
mammalian cells	10M	15-20M
flies and worms	2M	5-10M

# ChIP-Seq - Lecture Outline

- Experimental issues: how is a ChIP-seq experiment done?
- Analysis of the sequence data: how to detect the binding regions?
- Downstream analysis: how to extract the biological relevance?

# From Sequences to Mapped Reads

- Mapping million of sequences-reads:  
find their best alignment to the genome
- Instead of sequence we have coordinates and orientation
- Usually exclude sequences that map to more than one location on the genome

```
CAATGATAAACTGAGTTG
ATCAACTGGTGGCCATCTC
GGTATTTTATGGCAGTAGT
TAAAAAAATTCAGATTAT
TCTTAAACAATAACTGCT
TCTAGATTATGCTTTAAATA
AAATGGGACACCCGTGGG
CTTAATTCCCTTTCCTATGA
```

Sequencing



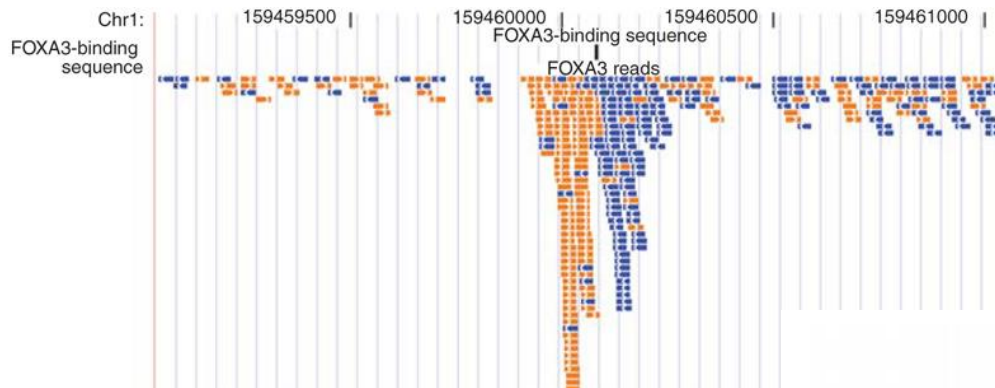
```
@HD VN:1.0 SO:coordinate
@SQ SN:chr20 LN:64444167
@PG ID:TopHat VN:2.0.14 CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-re
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6 18 GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C181DACXX:7:1102:4284:73714 16 chr20 190930 3 100M * 0 0
CCGTGTTAAGGTGGATCGGTCACCTCCAGCTAGGCTTAGGGATTCTTAGTGGCTAGGAAATCCAGTAGTCTCTCAGTCCCCCTCT
C BBDCCDDCCDDDDDDDDCCDDCCBDC?DDDDDDDDDDDDDDCCDDDDDDDDDDDDCCDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
AS:i:-15 XM:i:3 XO:i:0 XG:i:0 MD:Z:55C20C13A9 NM:i:3 NH:i:2 CC:Z:- CP:i:5532714 HI:i:0
HWI-ST1145:74:C181DACXX:7:1114:2759:41961 16 chr20 193953 50 100M * 0 0
TCTGGATCATCTGGTGTAGTGGCTCTGACTCAGAGACCTTCGCTCCCTGGGGCAGTGGACCTCCAGTATCCCTGACATAAGGGCATGGACGA
G DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
AS:i:-16 XM:i:3 XO:i:0 XG:i:0 MD:Z:60G16T18T3 NM:i:13 NH:i:1
HWI-ST1145:74:C181DACXX:7:1294:14760:4030 16 chr20 270877 50 100M * 0 0
GGCTTTATTGGTAAAAAAGGAATAGCAGATTAAATCAGAAATTCACCTGGCCAGCAGACCAACCAAGAAAGGAAGGAAGCAAGAAAAACCA
C DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD
AS:i:-11 XM:i:2 XO:i:0 XG:i:0 MD:Z:8A85G13 NM:i:2 NH:i:1
HWI-ST1145:74:C181DACXX:7:1218:11167:8699 0 chr20 271218 50 50M470N50M * 0 0
```

Alignment

Mapped Reads - bam

# Bioinformatics Challenge- Detecting the Binding Regions

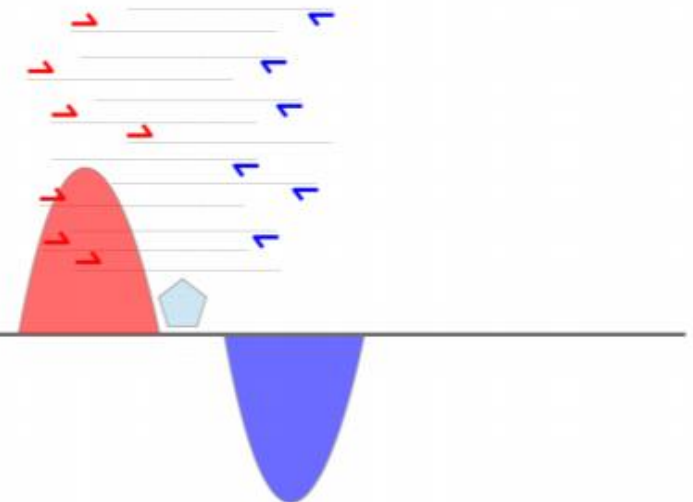
View mapped reads in a genome browser



Application note Nature Methods 6, (2009)

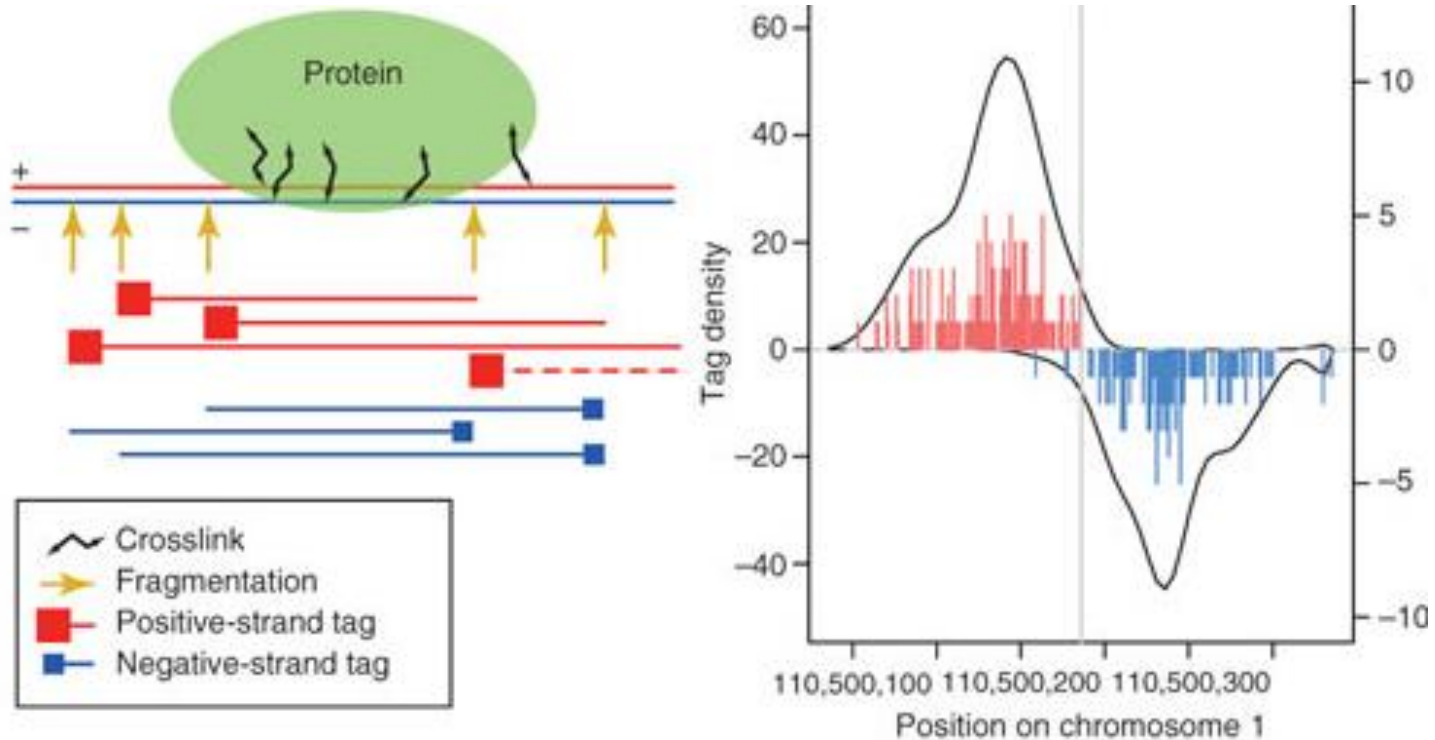
ChIP seq on DNA  
binding TF

read densities  
on +/- strand



# ChIP-Seq

## Bimodal distribution



Kharchenko et al. Nature Biotechnology 26, 1351 - 1359 (2008)



# Bioinformatics Challenge- Detecting the Binding Regions

Open Access

Method

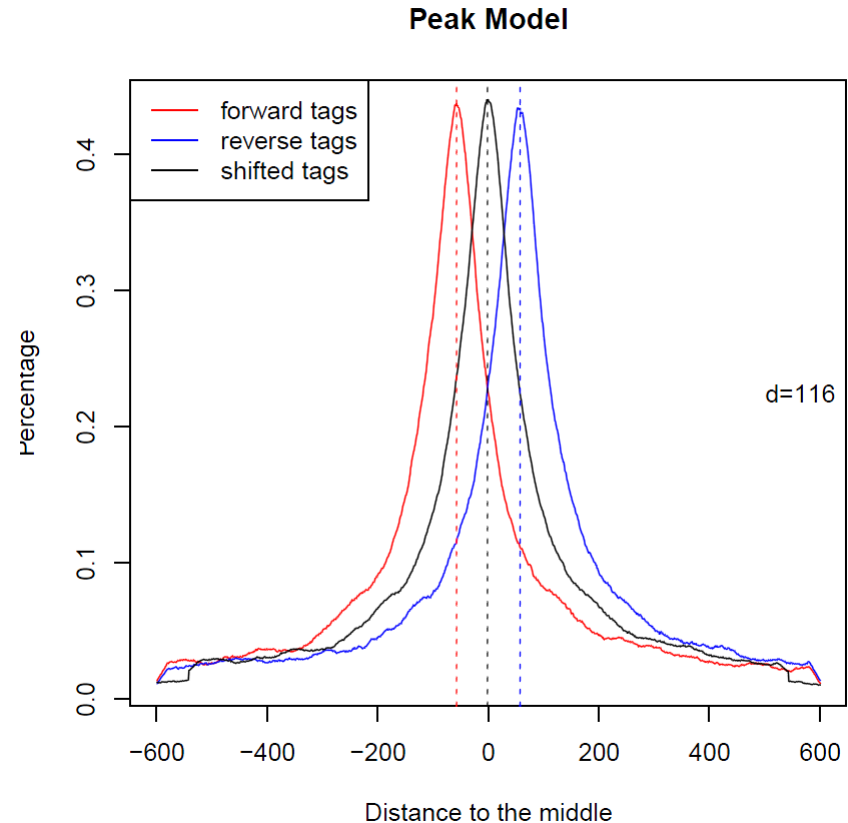
## Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang<sup>\*,\*</sup>, Tao Liu<sup>\*,\*</sup>, Clifford A Meyer<sup>\*</sup>, Jérôme Eeckhoute<sup>†</sup>,  
David S Johnson<sup>‡</sup>, Bradley E Bernstein<sup>§¶</sup>, Chad Nussbaum<sup>¶</sup>,  
Richard M Myers<sup>¥</sup>, Myles Brown<sup>†</sup>, Wei Li<sup>#</sup> and X Shirley Liu<sup>\*</sup>

Addresses: <sup>\*</sup>Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 44 Binney Street, Boston, MA 02115, USA. <sup>†</sup>Division of Molecular and Cellular Oncology, Department of Medical Oncology, Dana-Farber Cancer Institute and Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA. <sup>‡</sup>Gene Security Network, Inc., 2686 Middlefield Road, Redwood City, CA 94063, USA. <sup>§¶</sup>Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital and Department of Pathology, Harvard Medical School, 13th Street, Charlestown, MA 02129, USA. <sup>¶</sup>Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142, USA. <sup>¥</sup>Department of Genetics, Stanford University Medical Center, Stanford, CA 94305, USA. <sup>#</sup>Division of Biostatistics, Dan L Duncan Cancer Center, Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.

# MACS

- Building peak model
  - Uses 'high-quality' peaks to estimate fragment width  $d$
  - Searches for highly significant enriched regions (above a certain fold)
  - Separates +/- reads and detects the distance between the +/- read distribution
  - Shifts all reads  $d/2$  towards the 3' end

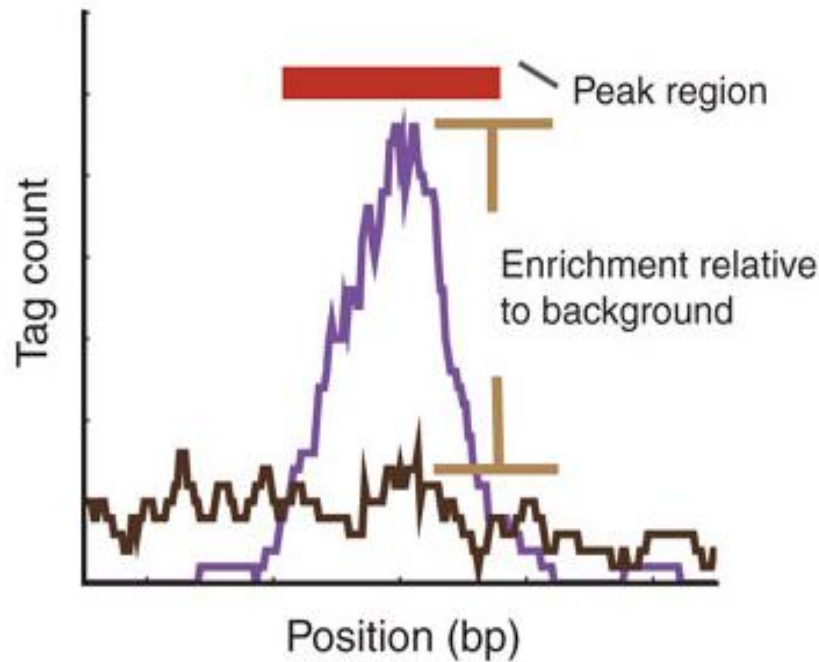


# MACS2- Paired end reads

- If the sequencing was paired end, MACS2 will use the actual insert sizes of pairs of reads (properly paired) to build fragment

```
end-1 ← HWI-EAS-249_38:7:1:7:1166/1
        Chr5  40  1  15902374  15902413  C  40 100 40
end-2 ← HWI-EAS-249_38:7:1:7:1166/2
        Chr5  1  40  15902154  15902193  F  40 100 40
```

# How to Assess If a Peak is Significant?



# Poisson distribution

- MACS models the reads distribution along the genome by a Poisson distribution
- The Poisson distribution is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed area, if these events occur with a known average rate and independently. (wikipedia)

# Poisson Distribution

- The Poisson distribution is sometimes called the law of small numbers because it is the probability distribution of the number of occurrences of an event that happens rarely but has very many opportunities to happen.

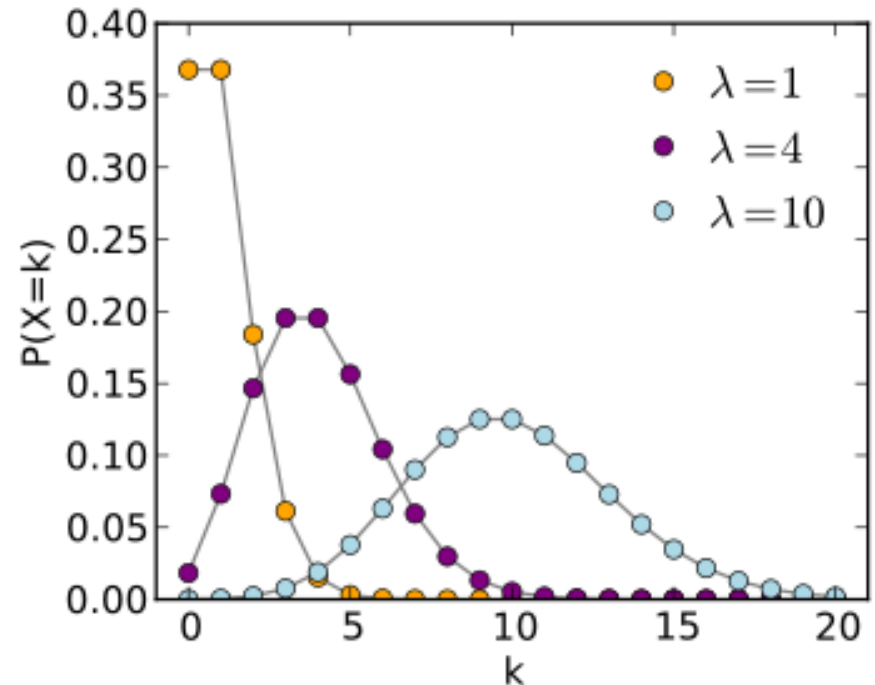
# Peak Detection

- MACS models the reads distribution along the genome by a Poisson distribution
  - We are counting reads in a fixed region
  - We can compute the expected (mean) number of events, i.e. number of reads in a specified region:  
 $\lambda_{BG} = \text{total read counts} / \text{effective genome size}$
  - The expectation is a small number
  - We have many reads
- In this model  $\lambda_{BG}$  is also the variance of the distribution

# Poisson Distribution Probability

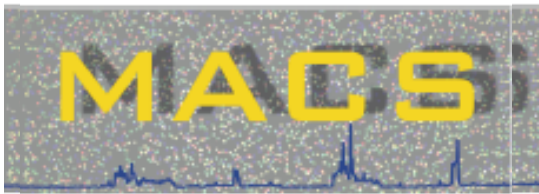
$$P(k; \lambda) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

Is the probability of observing  $k$ , for which  $\lambda$  is the expectation



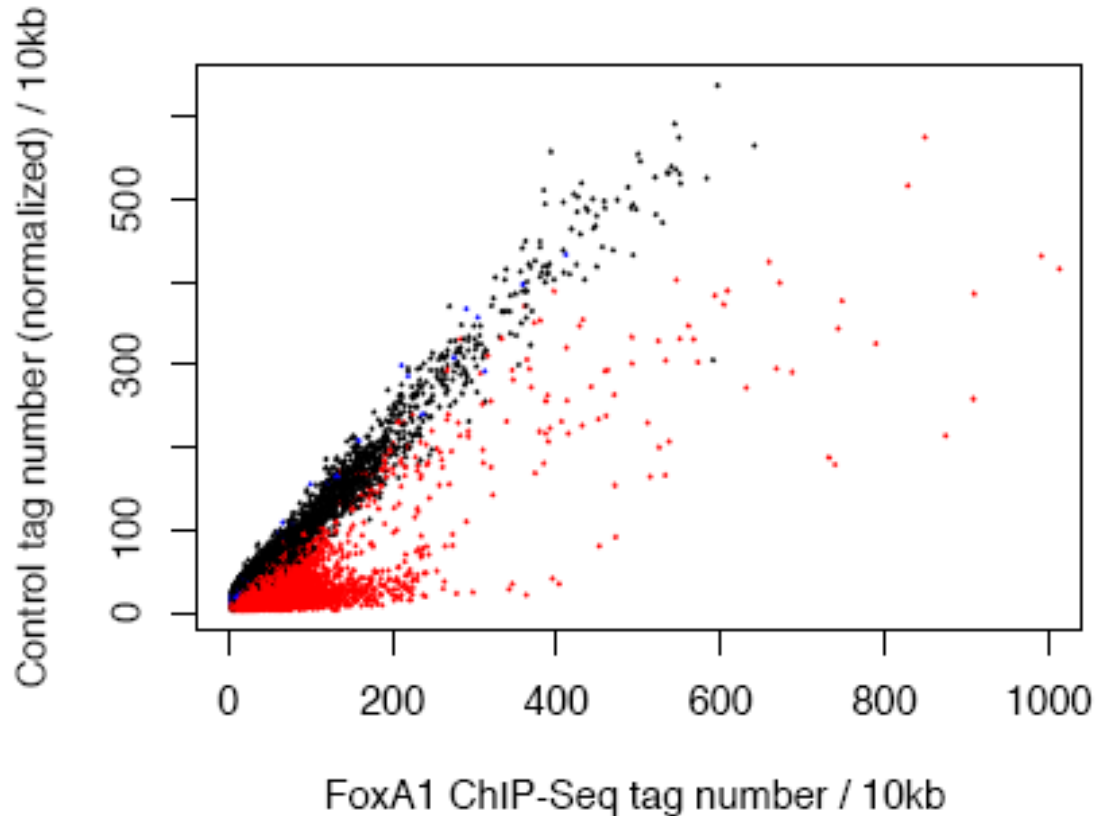
[http://en.wikipedia.org/wiki/Poisson\\_distribution](http://en.wikipedia.org/wiki/Poisson_distribution)

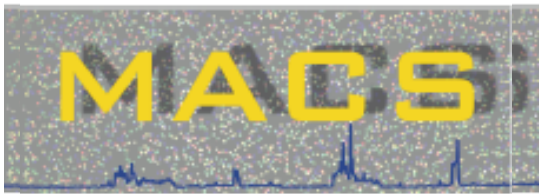




# Peak Detection

ChIP-Seq show local biases in the genome  
Chromatin and sequencing bias





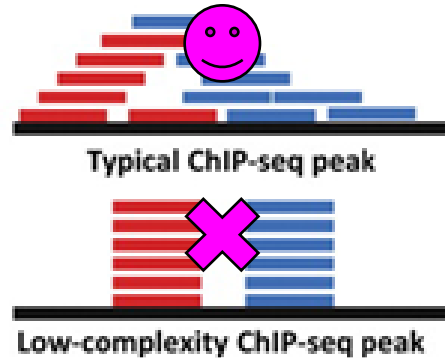
# Peak Calls

- The expected number of reads is
  - $\lambda_{BG}$  = total read counts / effective genome size
- Since ChIP-Seq data show local biases in the genome, a local expected value is calculated for each peak!

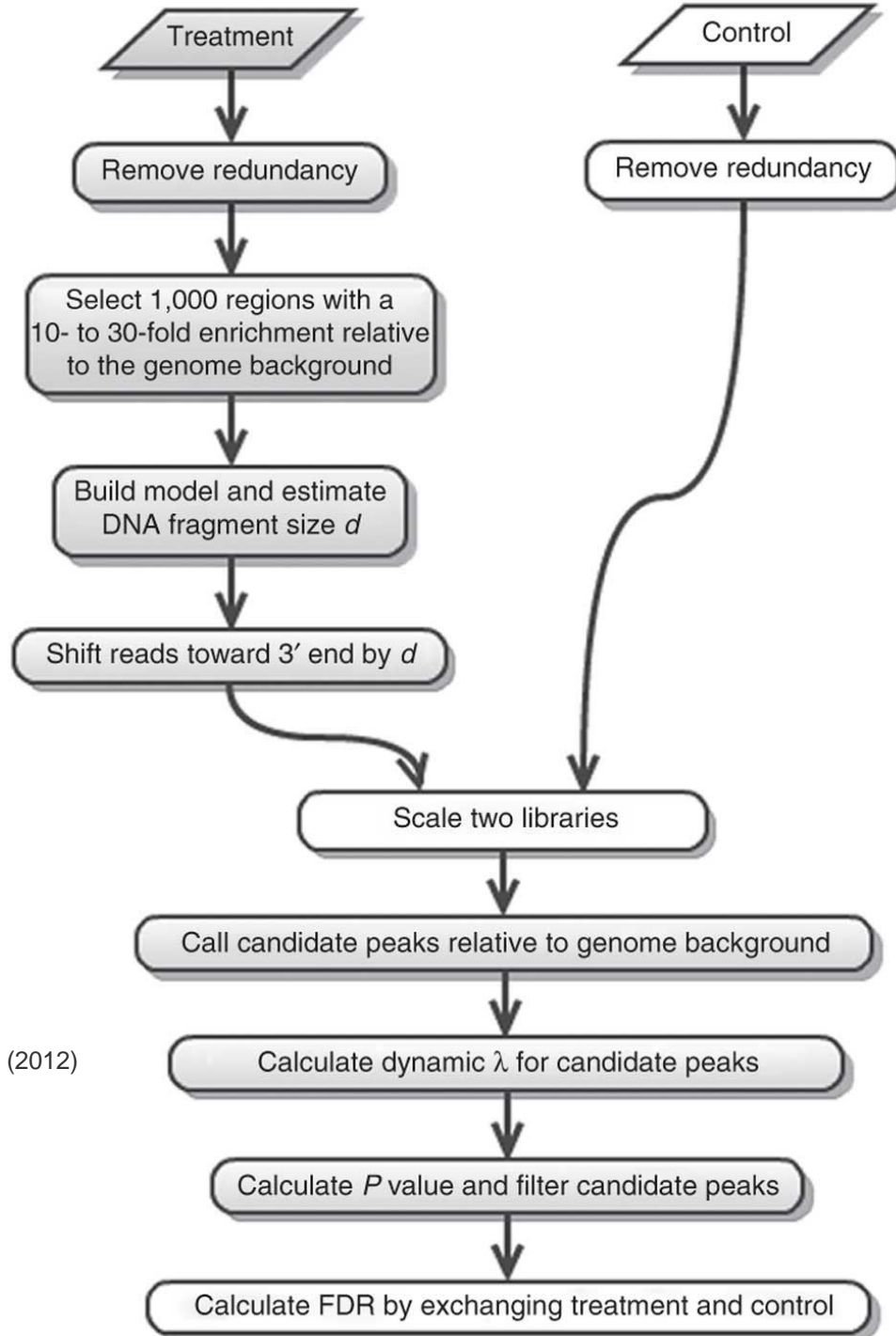
$$\text{Dynamic } \lambda_{\text{local}} = \max(\lambda_{BG}, [\lambda_{\text{ctrl}}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}])$$

ChIP Control  
1kb  
5kb  
10kb

# Redundant Reads - Tags



- Over amplification of ChIP-DNA by PCR may cause the same original DNA fragment to be sequenced repeatedly
- MACS removes the redundant reads i.e. reads at the exact same genome location and the same strand if their number exceeds the expected redundancy.
- Expected is based on the genome size and the number of reads.
- ENCODE guidelines- 10M uniquely mapped reads should have non-redundant frequency  $\geq 0.8$



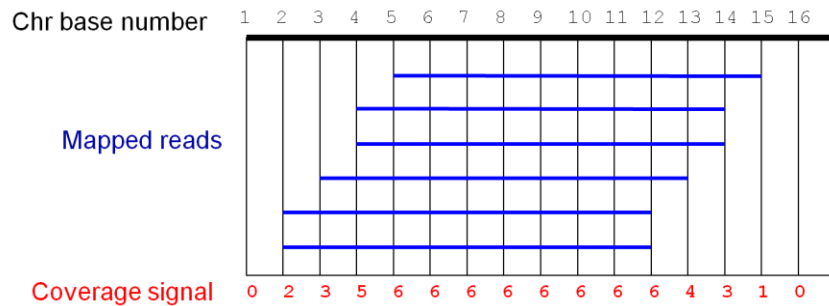
# MACS Peak Information (.xls)

- **Summit**  
peak summit position related to the start position of peak region,
- **Tags**  
number of tags-reads in peak region
- **$-10 \cdot \log_{10}(\text{pvalue})$**   
a PHRED like quality score for the peak region e.g. this value would be 100 for a p-value of  $1e-10$
- **Fold enrichment**
  - for this region against random Poisson distribution with local  $\lambda$

chr	start	end	length	summit	tags	$-10 \text{LOG}_{10}$ *(pvalue)	Fold enrich ment	FDR (%)
chr1	4838075	4838758	684	278	68	459.98	42.53	0.84

# MACS: Shifted Wiggle Files

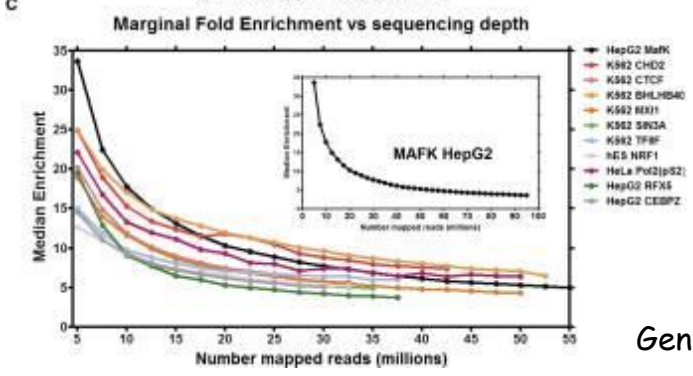
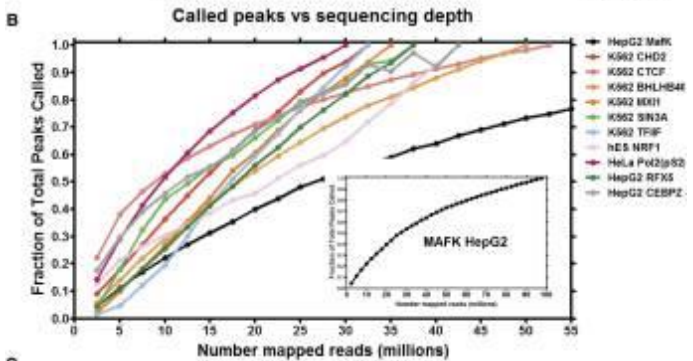
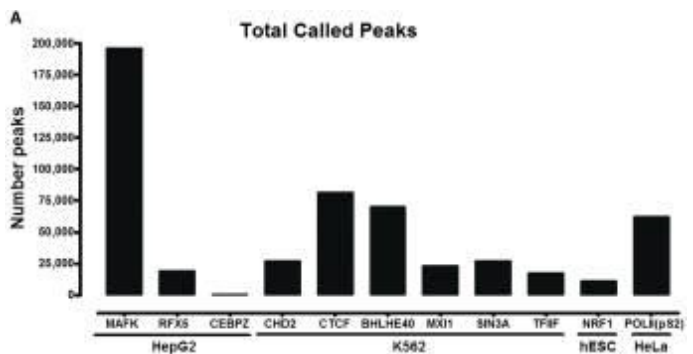
Shifted reads displayed as a coverage signal



Wiggle format example

variableStep	chrom=chr1
1	0
2	2
3	3
4	5
5	6
6	6
7	6
8	6
9	6
10	6
11	6
12	6
13	4
14	3
15	1

# Peak Counts Depend on Sequencing Depth



# Biological Replicates ENCODE

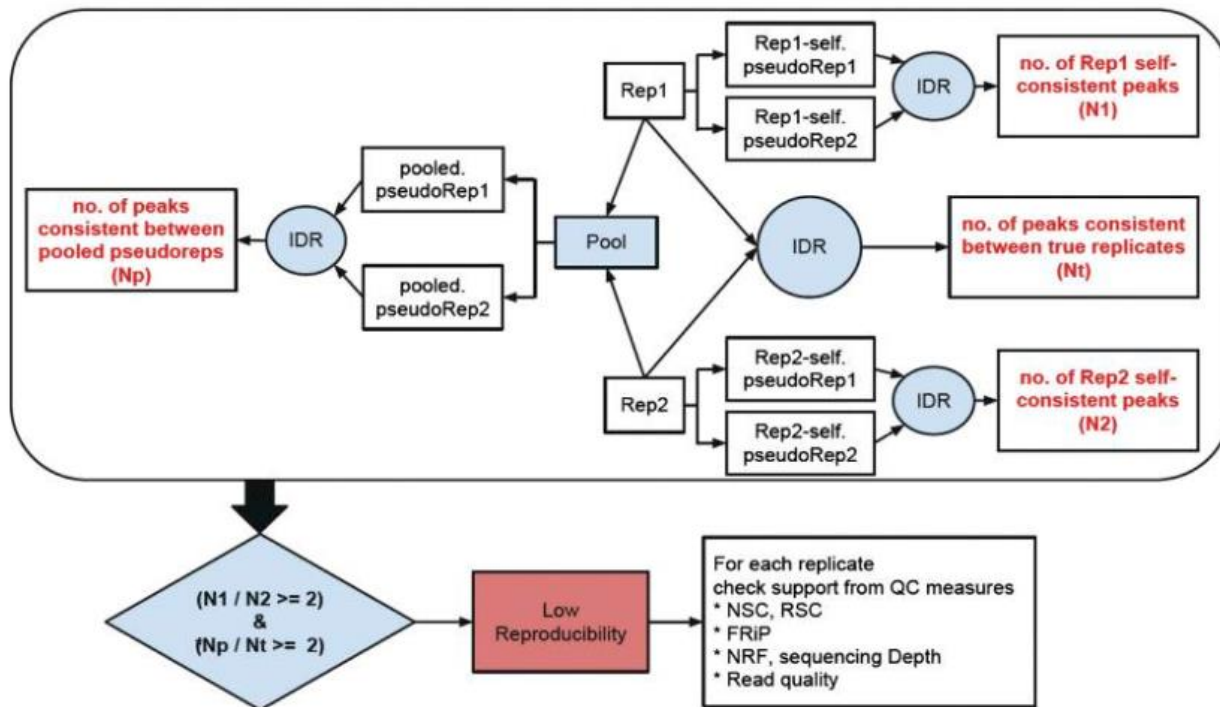
- Biological replicates are required for each dataset
- Criteria to decide that the biological replicates are in agreement- Irreproducible discovery rate (IDR)
- Reads from replicates which meet these criteria are usually combined and the data rescored.
- If requirement is not met a third replicate is required.



# Biological Replicates Evaluation

[https://hbctraining.github.io/Intro-to-ChIPseq/lessons/09\\_handling-replicates-idr.html](https://hbctraining.github.io/Intro-to-ChIPseq/lessons/09_handling-replicates-idr.html)

1. Evaluate peak consistency between **true replicates**
2. Evaluate peak consistency between **pooled pseudo-replicates**
3. Evaluate **self-consistency** for each individual replicate



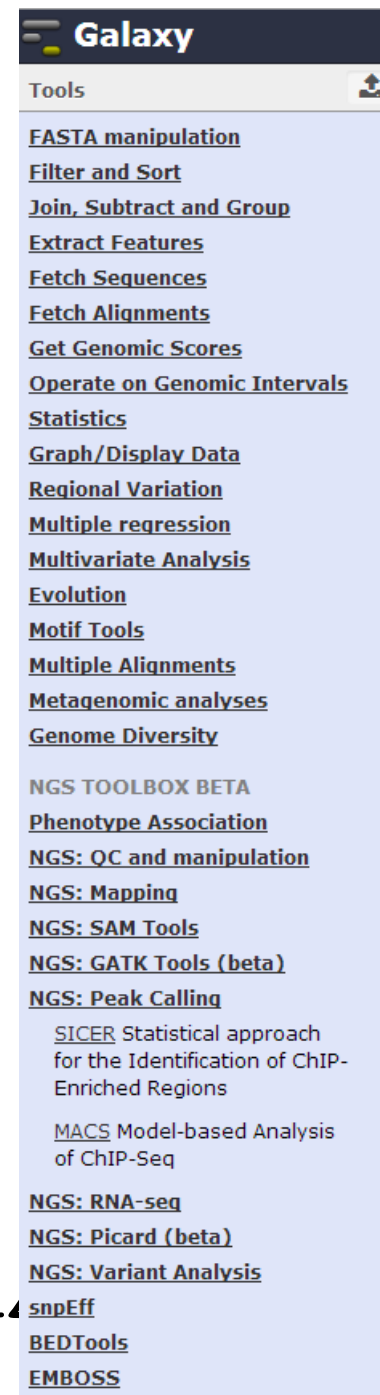
This figure is taken from the ENCODE ChIP-seq Guidelines.

# How to Run MACS

- As a command line program (on a Linux server - exercise)
- Web portals such as:
  - GALAXY (public)
  - CISTROME (public)

Disadvantage - They require you to load the mapped reads takes a **LONG TIME**; Usually do not run the latest version of MACS.

- UTAP



The image shows a screenshot of the Galaxy web interface. At the top, there is a dark header with the word "Galaxy" in white. Below the header, there is a light blue sidebar with a "Tools" section. The "Tools" section is expanded, showing a list of tool categories and individual tools. The tools listed are: FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, Genome Diversity, NGS TOOLBOX BETA, Phenotype Association, NGS: QC and manipulation, NGS: Mapping, NGS: SAM Tools, NGS: GATK Tools (beta), NGS: Peak Calling, SICER Statistical approach for the Identification of ChIP-Enriched Regions, MACS Model-based Analysis of ChIP-Seq, NGS: RNA-seq, NGS: Picard (beta), NGS: Variant Analysis, snpEff, BEDTools, and EMBOSS.

# Public Tools for ChIP-Seq Analysis

**Table 1** | Publicly available ChIP-seq software packages discussed in this review

	Profile	Peak criteria <sup>a</sup>	Tag shift	Control data <sup>b</sup>	Rank by	FDR <sup>c</sup>	User input parameters <sup>d</sup>	Artifact filtering: strand-based/duplicate <sup>e</sup>	Refs.
CsGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally <i>P</i> values	<i>P</i> value	1: None 2: $\frac{\# \text{ control}}{\# \text{ ChIP}}$	Optional peak height, ratio to background	Yes / No	4, 18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	<i>s</i> s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLTR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\# \text{ control}}{\# \text{ ChIP}}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson <i>P</i> value	Estimate from high quality peak pairs	Used for Poisson fit when available	<i>P</i> value	1: None 2: $\frac{\# \text{ control}}{\# \text{ ChIP}}$	<i>P</i> -value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial <i>P</i> value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	<i>q</i> value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	<i>q</i> value	1: NA 2: $\frac{\# \text{ control}}{\# \text{ ChIP}}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	<i>P</i> value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and <i>P</i> values	<i>q</i> value	1: None 2: From Poisson <i>P</i> values	Window length, gap size, FDR (with control) or <i>E</i> -value (no control)	No / Yes	15
SICSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region <sup>f</sup>	Average nearest paired tag distance	Used to compute fold-enrichment distribution	<i>P</i> value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson <i>P</i> value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	<i>P</i> value	1: Monte Carlo simulation 2: $\frac{\# \text{ control}}{\# \text{ ChIP}}$	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial <i>P</i> value	Estimated or user specified	Subtracted before peak calling	<i>q</i> value	1, 2: binomial 2: $\frac{\# \text{ control}}{\# \text{ ChIP}}$	Target FDR	No / Yes	20

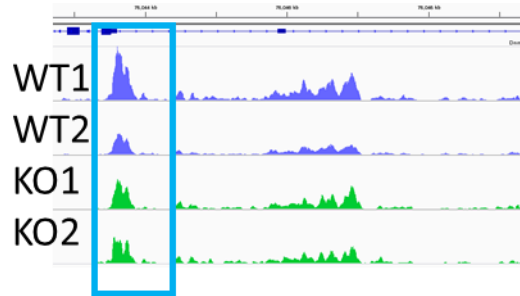
<sup>a</sup>The labels 1: and 2: refer to one-sample and two-sample experiments, respectively. <sup>b</sup>These descriptions are intended to give a rough idea of how control data is used by the software. 'NA' means that control data are not handled. <sup>c</sup>Description of how FDR is or optionally may be computed. 'None' indicates an FDR is not computed, but the experimental data may still be analyzed; 'NA' indicates the experimental setup (1 sample or 2) is not yet handled by the software.  $\frac{\# \text{ control}}{\# \text{ ChIP}}$ , number of peaks called with control (or same portion thereof) and sample reversed. <sup>d</sup>The lists of 'user input parameters' for each program are not exhaustive but rather comprise a subset of greatest interest to new users. <sup>e</sup>'Strand-based' artifact filtering rejects peaks if the strand-specific distributions of reads do not conform to expectation, for example by exhibiting extreme bias of tag populations for one strand or the other in a region. 'Duplicate' filtering refers to either removal of reads that occur in excess of expectation at a location or filtering of called peaks to eliminate those due to low complexity read pairs that may be associated with, for example, microsatellite DNA.  $N_+$  and  $N_-$  are the numbers of positive and negative strand reads, respectively.

© 2009 Nature America, Inc. All rights reserved.

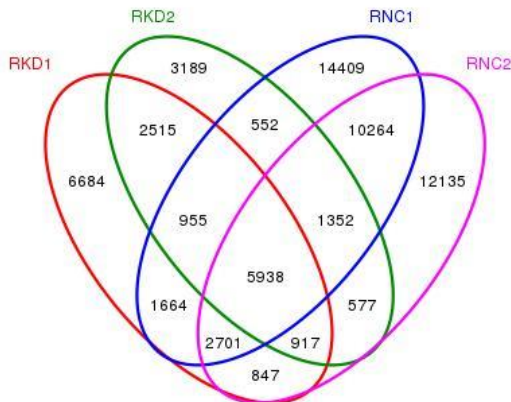


# Comparing Peaks

- Is the binding to WT higher than to KO?



Venn Diagram using peaks regions



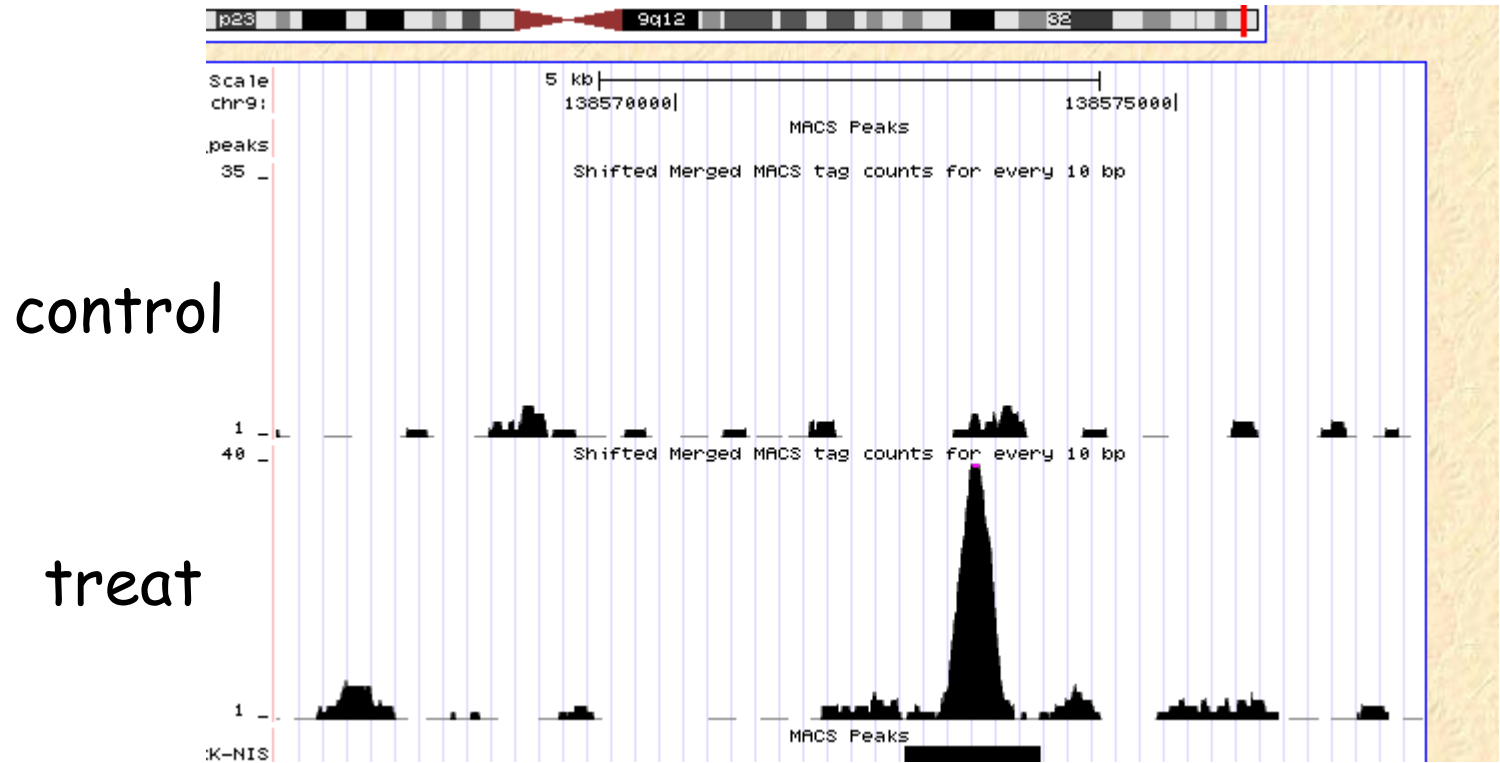
DiffBind: Differential binding analysis of ChIPSeq peak data  
Extract fragment counts for all peaks in all samples and perform a statistical test, such as DESeq2

-44-

# ChIP-Seq - Lecture Outline

- Experimental issues: how is a ChIP-seq experiment done?
- Analysis of the sequence data: how to detect the binding regions?
- **Downstream analysis: how to extract the biological relevance?**

# Loading to a Genome Browser



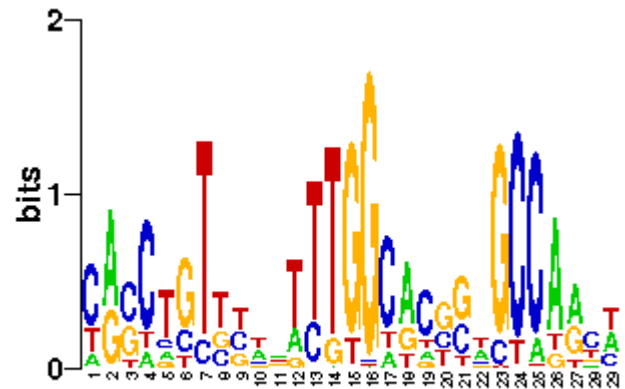
# TF Motifs Tools

Suppose I have the list of enriched genomic regions, what next?

- Find the TF binding motifs enriched in comparison to genomic background
- Predict the TF motif

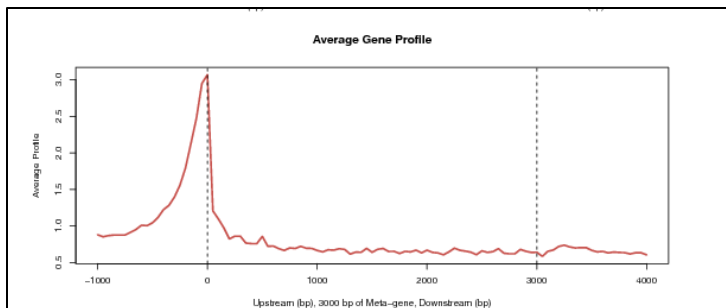
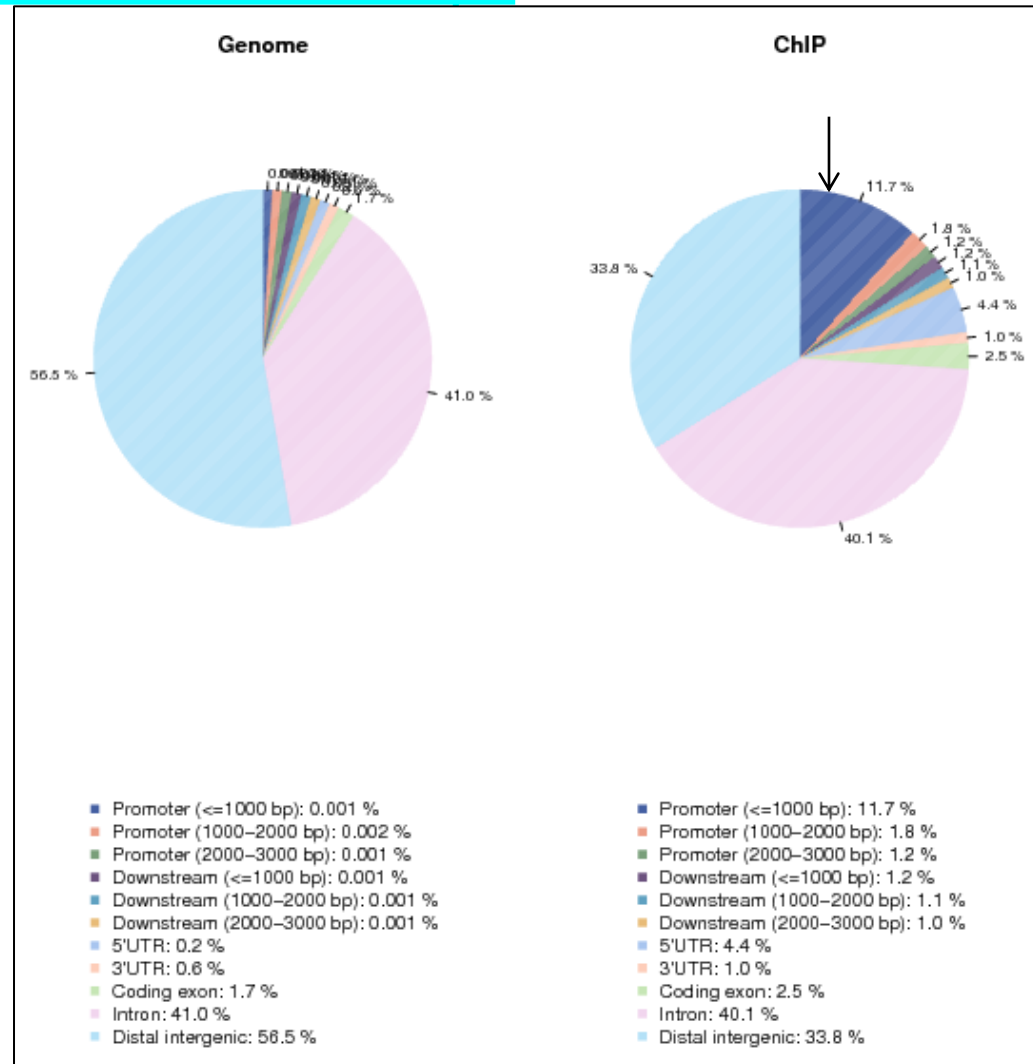
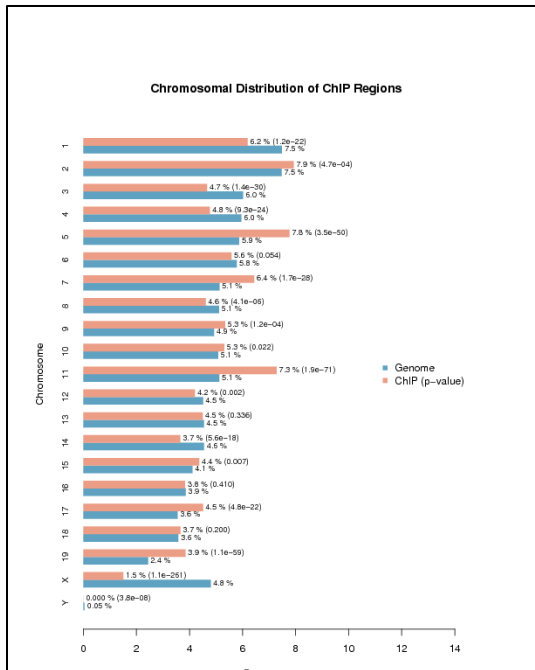
Recommended portals:

- MEME-ChIP (public)
- Homer - command line



# CEAS: Enrichment of Genome Features

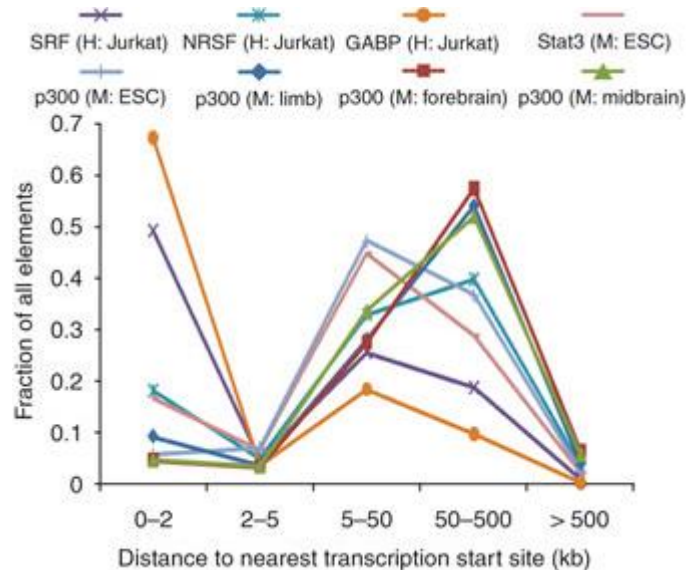
<http://cistrome.dfci.harvard.edu/ap/>





# Functional Interpretation

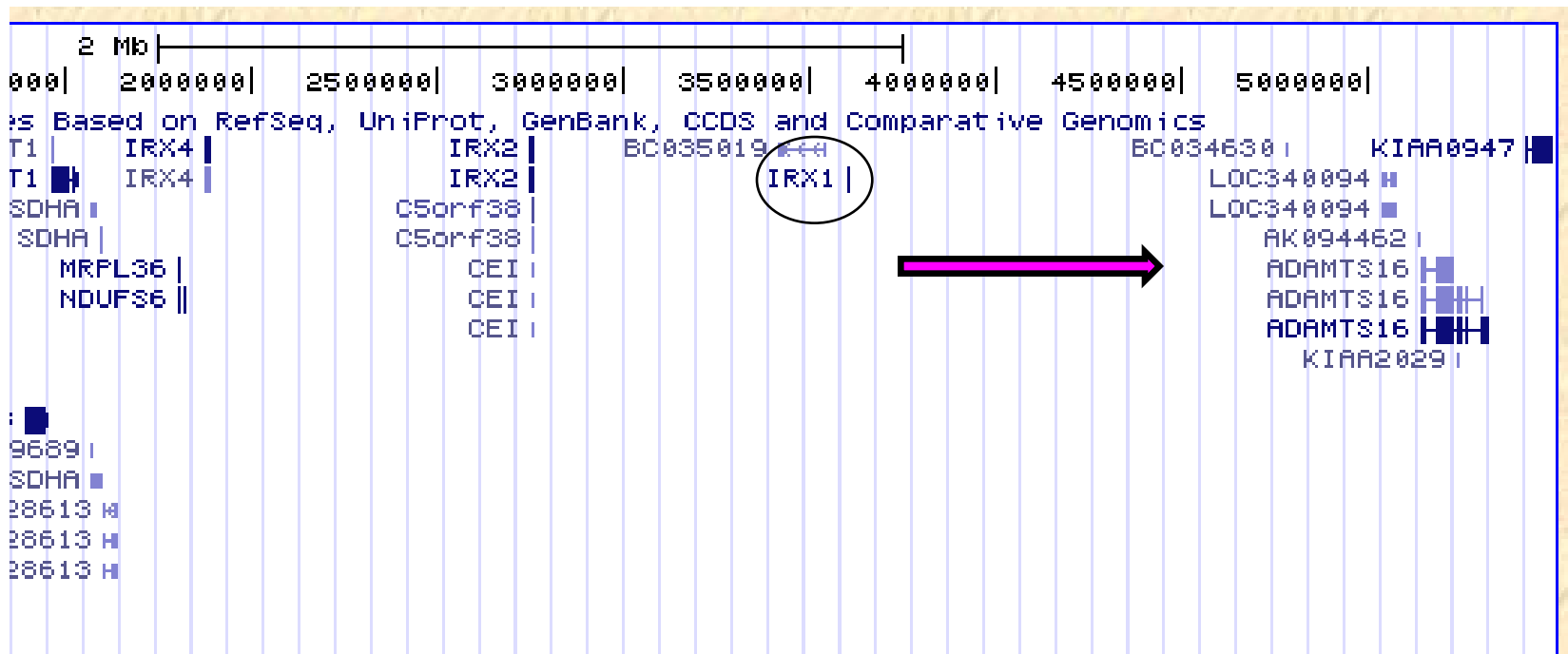
- Which processes and functions does our TF regulate?
  - Associate peaks with genes
  - Associating only proximal genomic regions to genes (<5 kb) – for most TF ignores a large fraction of binding data



Nature Biotechnology 28 ,  
495-501 (2010)

# Functional Interpretation

- Some genes are found in "gene deserts" and therefore the regulatory genomic region we can assign to them is large (<1Mb)



# In Which Processes and Functions is Our TF Involved?

## Associating Peaks-Genes-Ontology

- Ontology term 1: gene1, gene2, gene3 ...
- Ontology term 2: gene2, gene5, gene8 ...
- ...
- Are my peaks located near genes enriched for certain ontology terms?
- Which statistical test should we apply?

## GREAT improves functional interpretation of *cis*-regulatory regions

Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger & Gill Bejerano

**Affiliations** | **Contributions** | **Corresponding author**


*Nature Biotechnology* **28**, 495–501 (2010) | doi:10.1038/nbt.1630

Published online 02 May 2010

- **GREAT's genomic region-based statistical test**
- **The probability of hitting a term is calculated as the fraction of the genome that is associated with that term**

# GREAT ANALYSIS

**Mouse Phenotype (no terms)** Global controls

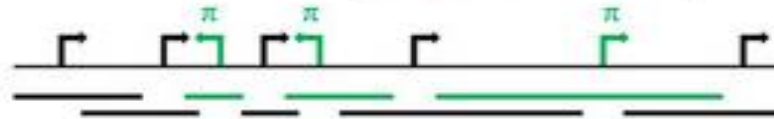
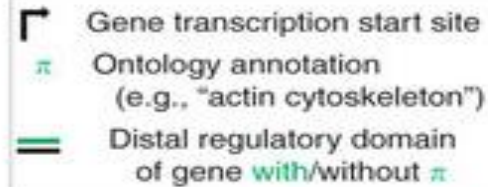
Table controls: Export Shown top rows in this table: 20 Set Term annotation count: Min: 1 Max: Inf Set Visualize this table:  [select one]

Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
decreased heart rate	19	6.4691e-14	2.4889e-11	9.9387	19	19.39%	2	2.1817e-2	10.1569	7	104	5.22%
increased sensitivity to xenobiotic induced morbidity/mortality	76	3.8303e-10	3.6841e-8	10.5656	13	13.27%	1	2.9367e-2	10.7788	7	98	5.22%
abnormal xenobiotic induced morbidity/mortality	104	3.3973e-9	2.3879e-7	8.7957	13	13.27%	4	3.8428e-2	8.3835	7	126	5.22%
complete preweaning lethality	107	4.1608e-9	2.8426e-7	8.6458	13	13.27%	3	4.6184e-2	8.5187	7	124	5.22%
decreased circulating adrenaline level	406	9.3188e-4	1.6778e-2	15.9885	3	3.06%	5	3.3941e-2	50.3010	3	9	2.24%

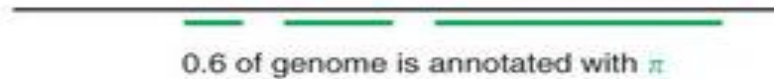
The test set of 98 genomic regions picked 134 (1%) of all 20,221 genes.  
*Mouse Phenotype* has 7,310 terms covering 6,642 (33%) of all 20,221 genes, and 456,354 term - gene associations.  
 7,310 ontology terms (100%) were tested using an annotation count range of [1, Inf].

## Binomial test over genomic regions

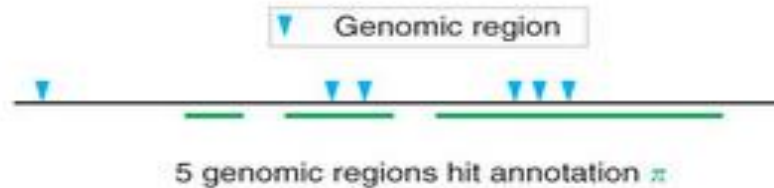
Step 1: Infer distal gene regulatory domains



Step 2: Calculate annotated fraction of genome



Step 3: Count genomic regions associated with the annotation



Step 4: Perform binomial test over genomic regions

$n = 6$  total genomic regions

$p_{\pi} = 0.6$  fraction of genome annotated with  $\pi$

$k_{\pi} = 5$  genomic regions hit annotation  $\pi$

$$P = \Pr_{\text{binom}}(k \geq 5 \mid n = 6, p = 0.6)$$

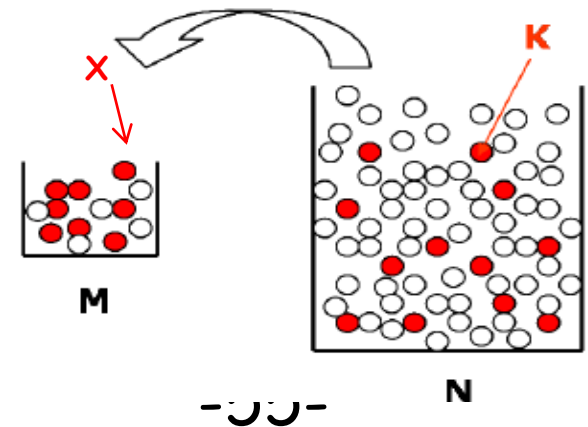
Nature  
Biotechnology 28,  
495-501 (2010)

# Gene List Enrichment Test

The hypergeometric test is the standard gene enrichment test for gene lists (such as differential gene lists from microarray expression studies).

The hypergeometric p-value equals the probability of choosing  $x$  from  $K$  (red balls- genes with a certain ontology) when randomly drawing  $M$  genes from the genome with  $N$  genes.

$$P(X = x | N, M, K) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}$$



# Binomial vs Hypergeometric



statistical test  
calculated for a set of  
genomic regions

GREAT expects 33% of  
all input peaks to be  
associated with  
'multicellular organismal  
development'

statistical test  
calculated for a set of  
genes

A gene based approach  
would expect 14% of the  
genes near peaks to be  
associated with  
'multicellular organismal  
development'

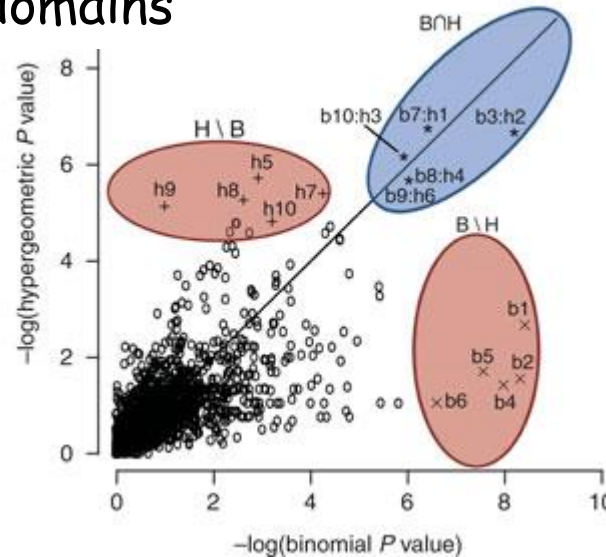
Why do we have this discrepancy? -56-



# GREAT Uses Both the Hypergeometric and the Binomial Test

Significant by hypergeometric :  
general terms arising from genes with large regulatory domains

Significant in both tests: specific and accurate -supported by multiple genes and binding events



Significant by binomial test:  
many peaks near few genes

Nature Biotechnology  
28 , 495-501 (2010)

# Ontologies Included in GREAT

Currently, GREAT includes

- Gene Ontology (GO)
- Ontologies covering phenotypes and human disease
- Pathways
- Gene expression
- Regulatory motifs
- Gene families

# ChIP-Seq - Lecture Outline

- Experimental issues: how is a ChIP-seq experiment done?
- Analysis of the sequence data: how to detect the binding regions?
- Downstream analysis: how to extract the biological relevance?

# My Peak

