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

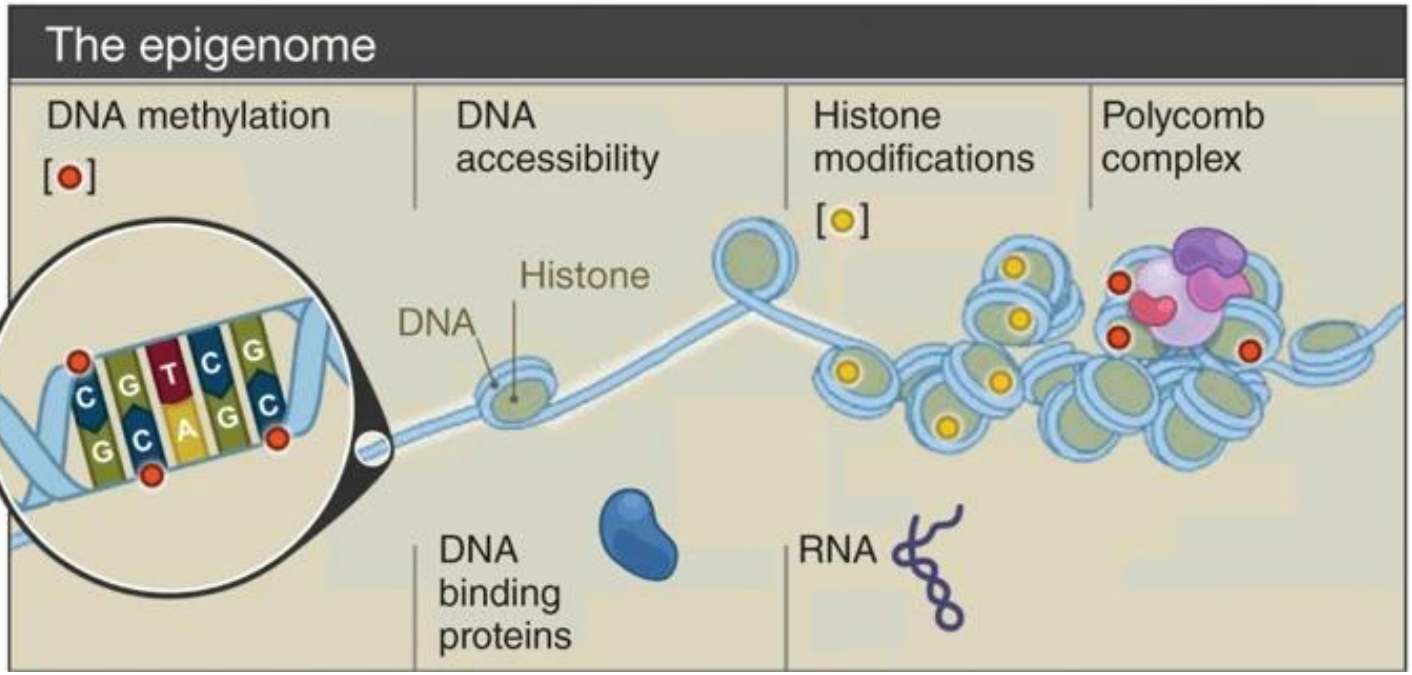
Dena Leshkowitz

Introduction to Deep Sequencing Analysis for Biologists

Course 2021

Epigenome

- How does the same genome sequence give rise to over hundreds of different cell types through remarkably consistent differentiation programs?
- Physical organization and modification of genomic DNA regulates gene expression



www.itsjustabadday.com

Definition

- ChIP-Seq is short for

  Chromatin Immuno-Precipitation followed by Sequencing

- It provides quantitative, genome-wide view of DNA- protein binding events

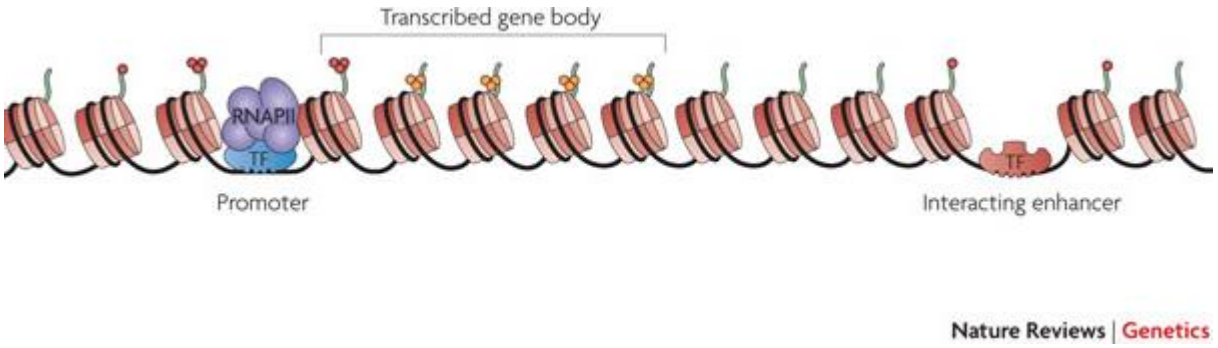
ChIP-Seq – Lecture Outline

- 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:

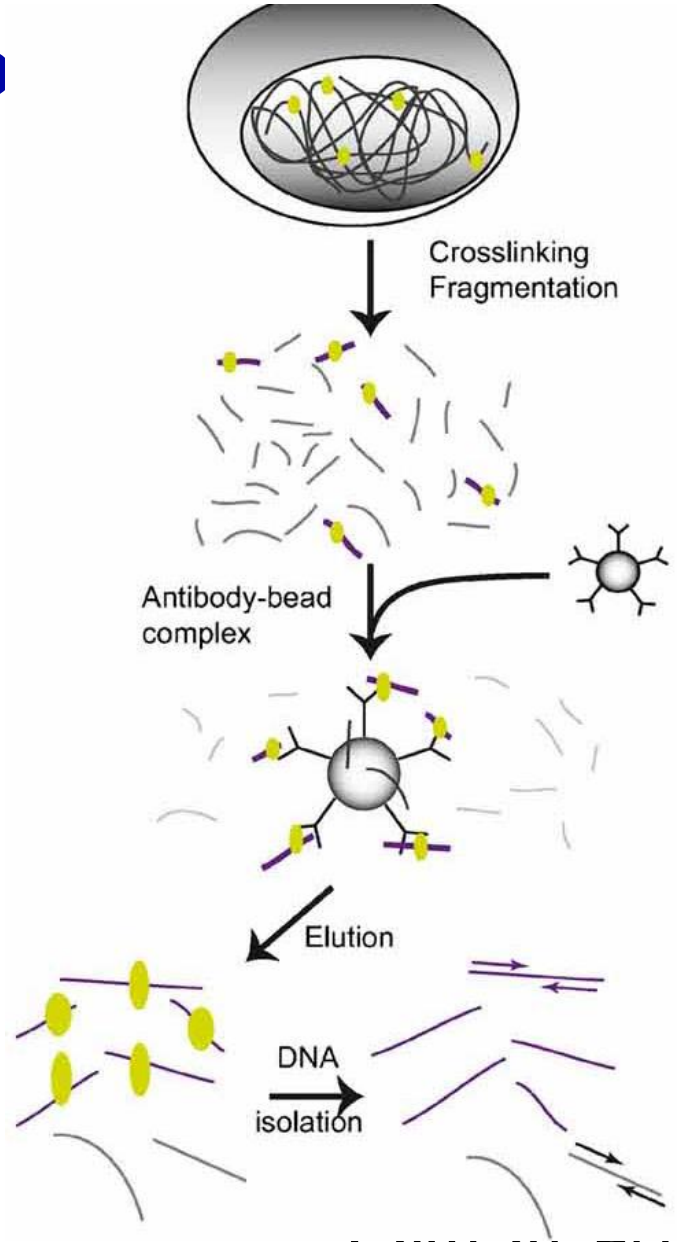
- Binding of Transcription Factors (TF) to promoters or enhancers
- Binding of RNA polymerase II
- Binding of modified histones



ChIP – How is it done?

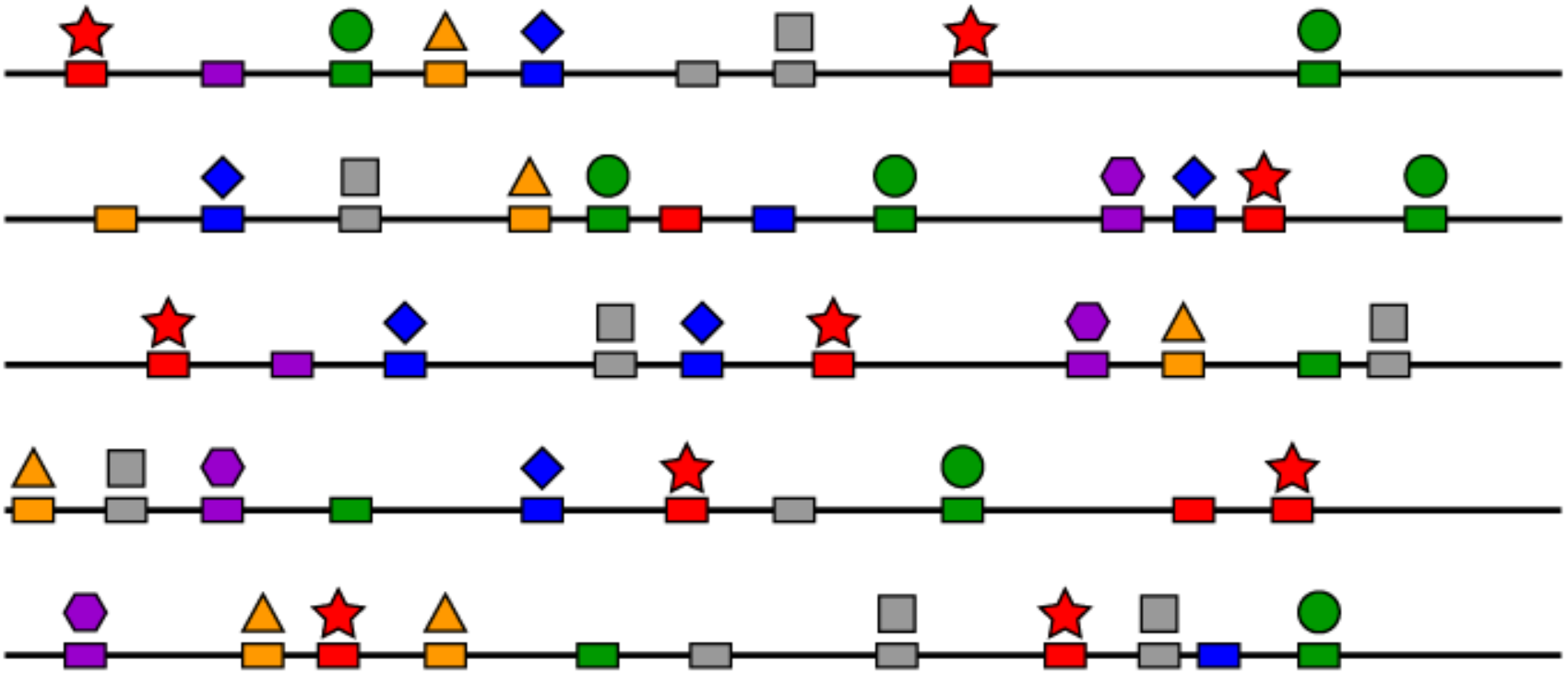
ChIP is a technique that permits to

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



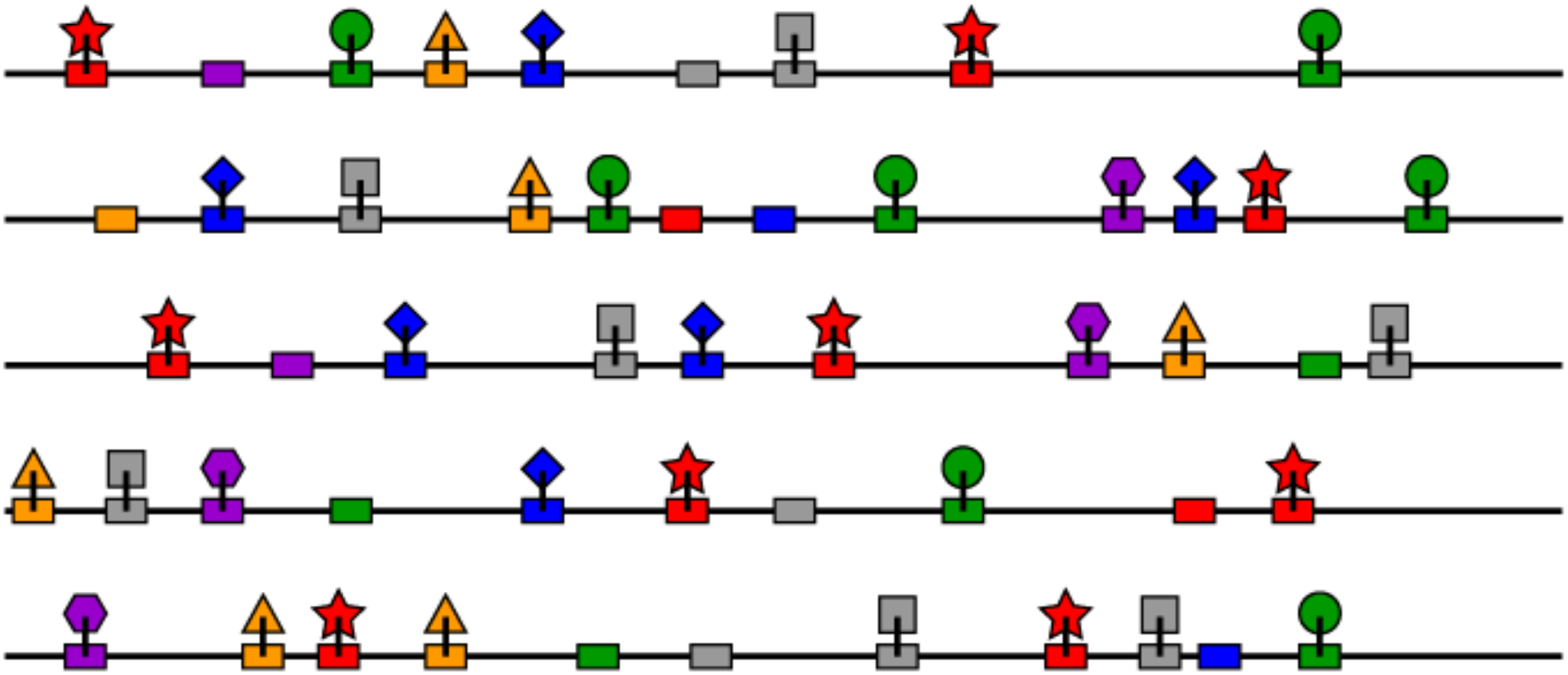
Barski and Zhao J Cell Biochem. 2009 May 1;107(1):11-8(2009)

Chromatin ImmunoPrecipitation (ChIP)



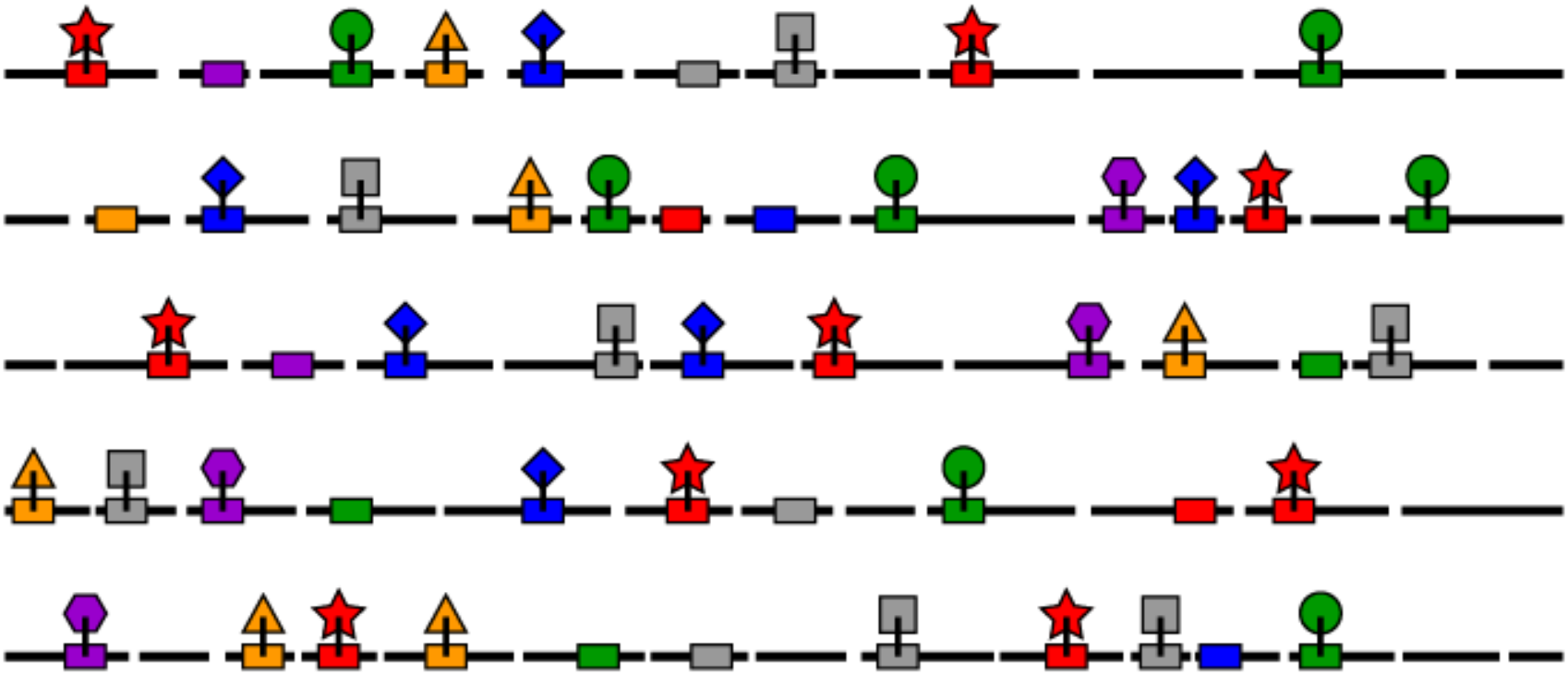
8 out of 60

TF/DNA Crosslinking *in vivo*



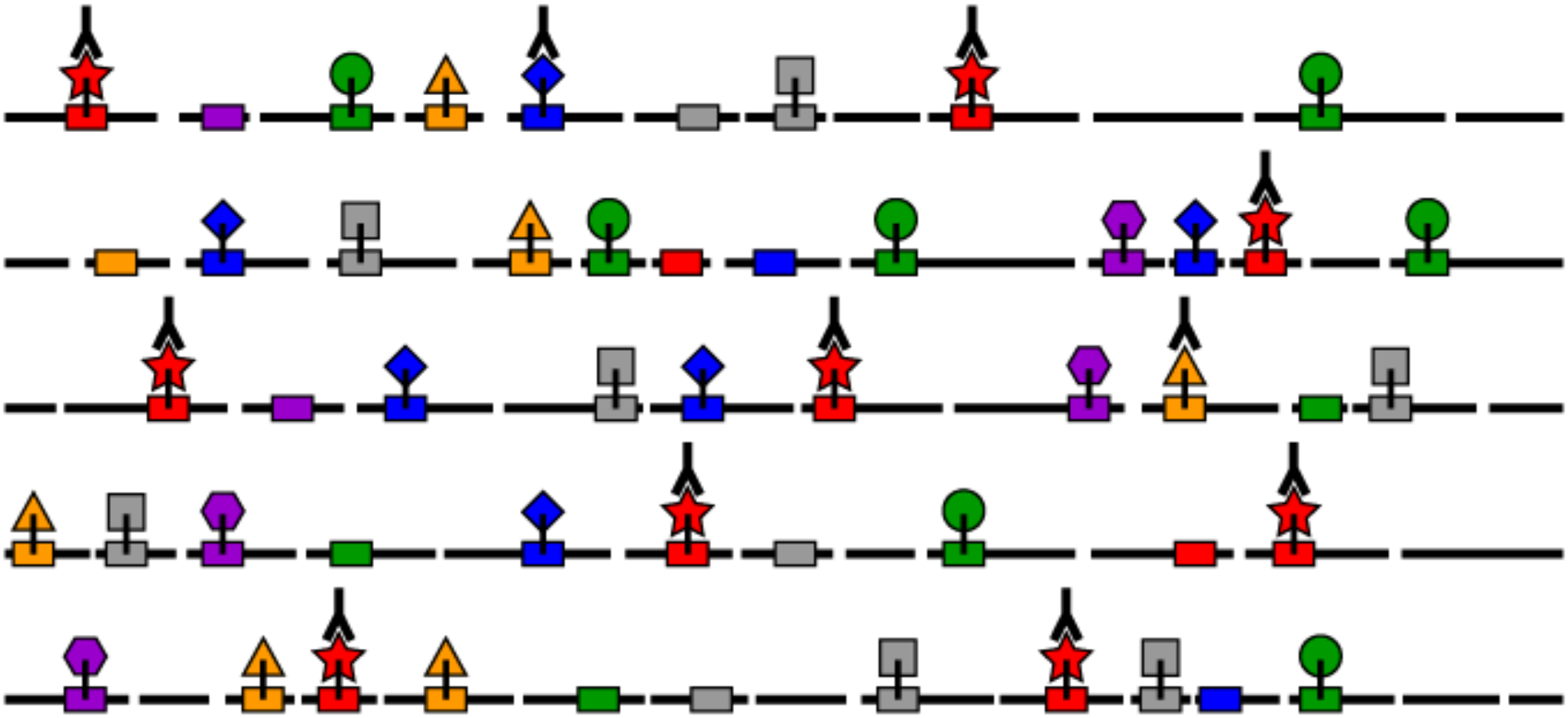
9 out of 60

Sonication (~200bp)



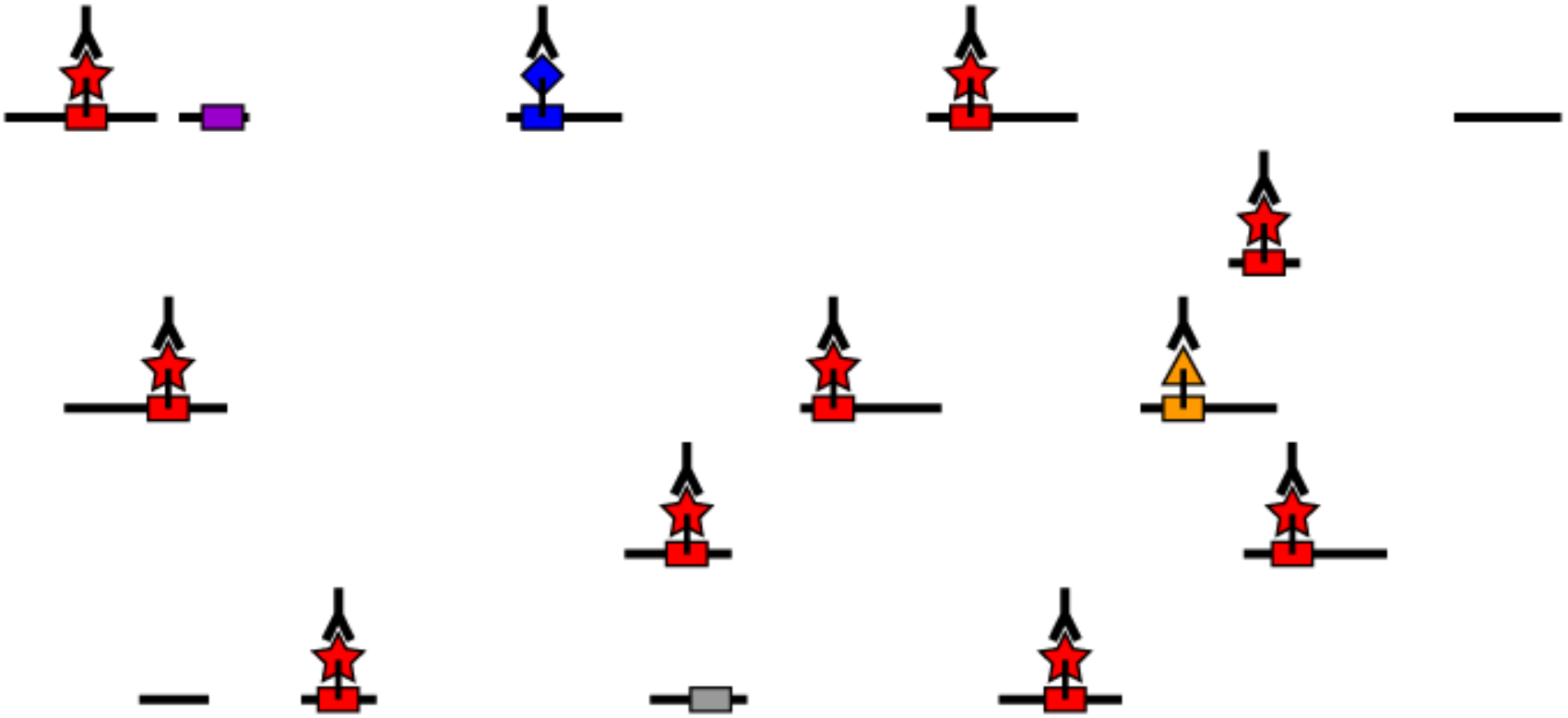
10 out of 60

TF-specific Antibody



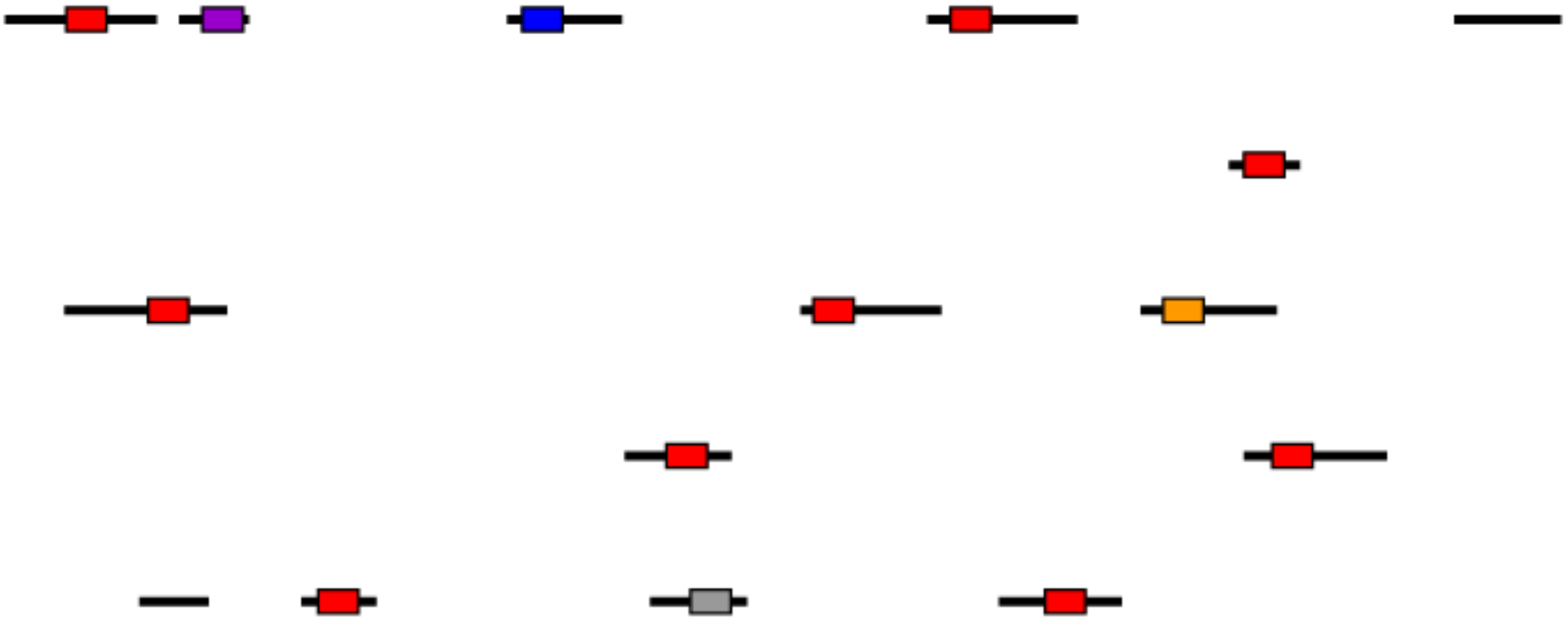
11 out of 60

Immunoprecipitation



12 out of 60

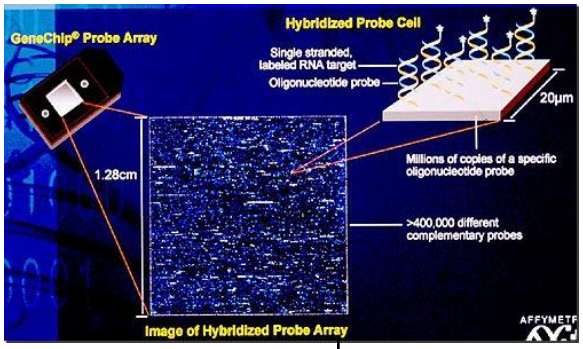
Reverse Crosslink and DNA Purification



ChIP High-throughput technology

Discover the DNA binding regions in the genomic scale

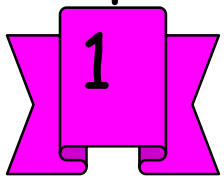
DNA Microarrays



ChIP-Chip

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

Next Generation Sequencing



ChIP-Seq

Experimental Design

- ENCODE consortium's Standards, Guidelines and Best Practices

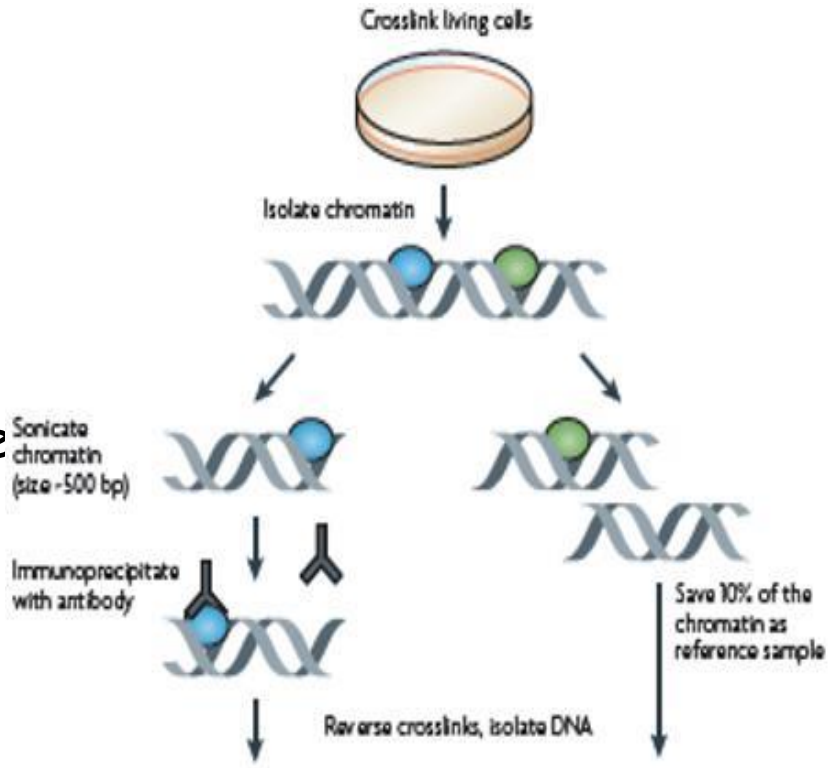
  Genome Res. 2012 Sep;22(9):1813-31. doi: 10.1101/gr.136184.111.

- 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



"treatment" "control"

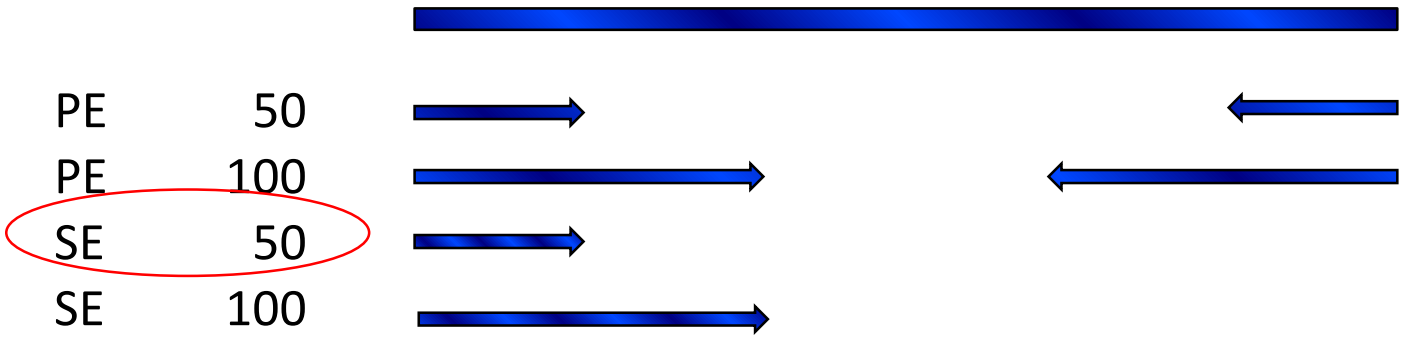
Nat Rev Genet. 2009 10(9): 605–616.

Why Do I Need a Control Sample?

- A lot of the 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

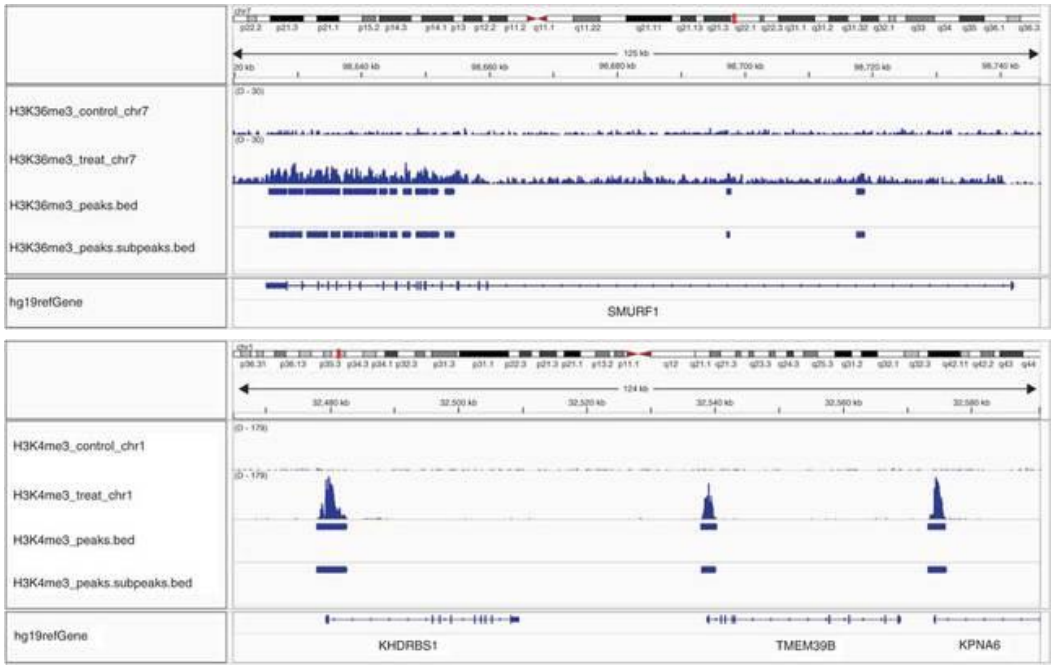


Coverage Required - Types of Peaks

How many reads (sequences) per sample?

Broad peaks- H3K36me3 (exon regions)

Sharp peaks- H3K4me3 (promoter regions)



Feng et al. Nature Protocols 7(2012)

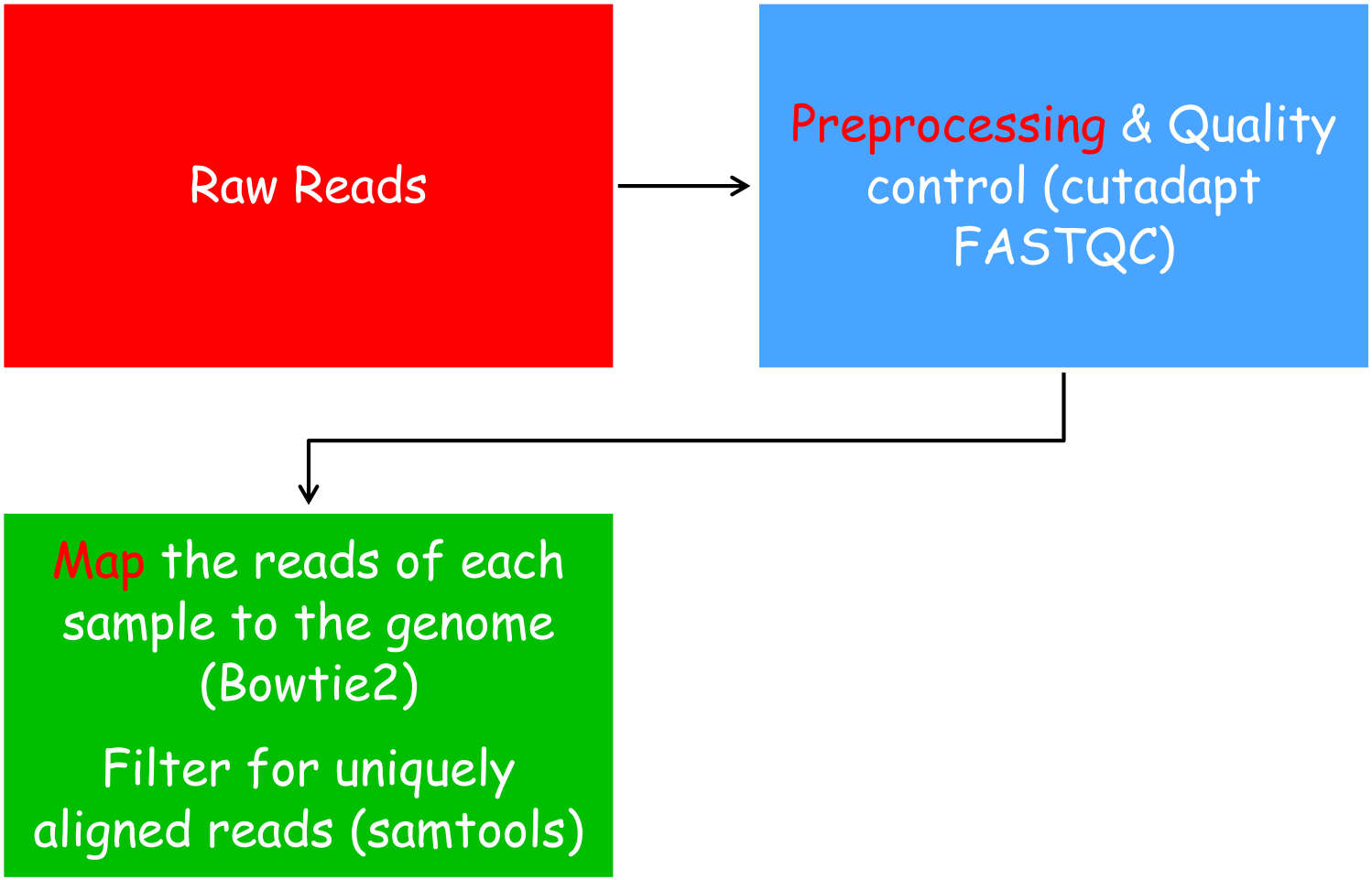
ENCODE recommendations:

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?

ChIP-Seq Workflow



From Sequences to Mapped Reads

Sequences – Fastq

```
@HISEQ:226:C95PJANXX:2:1106:7378:57379 1:N:0:CGCTATGT
CCCGTCTCAAAAGGGAAACAGGTTGATATTCCTGTGCCAATAGTATTATGAGTTTCTTAGA
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEECGGGGDDGGGGGGGGGGGGG
@HISEQ:226:C95PJANXX:2:1106:7467:57387 1:N:0:CGCTATGT
CTCAGGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCCGTAAA
+
CCCCCGGGGGGGGBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:226:C95PJANXX:2:1106:7460:57432 1:N:0:CGCTATGT
CTTGCATTGTTATTTCTTGTCACTACCTCTCTTCTTTAGAATTGGGTAATTTACGCGCCTG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@HISEQ:226:C95PJANXX:2:1106:7269:57438 1:N:0:CGCTATGT
GTGGGGAGTTTGACTGGGGCGGTACATCTGTTAAATATTAACGCAGATGTCCAAAGACAAG
+
CCCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

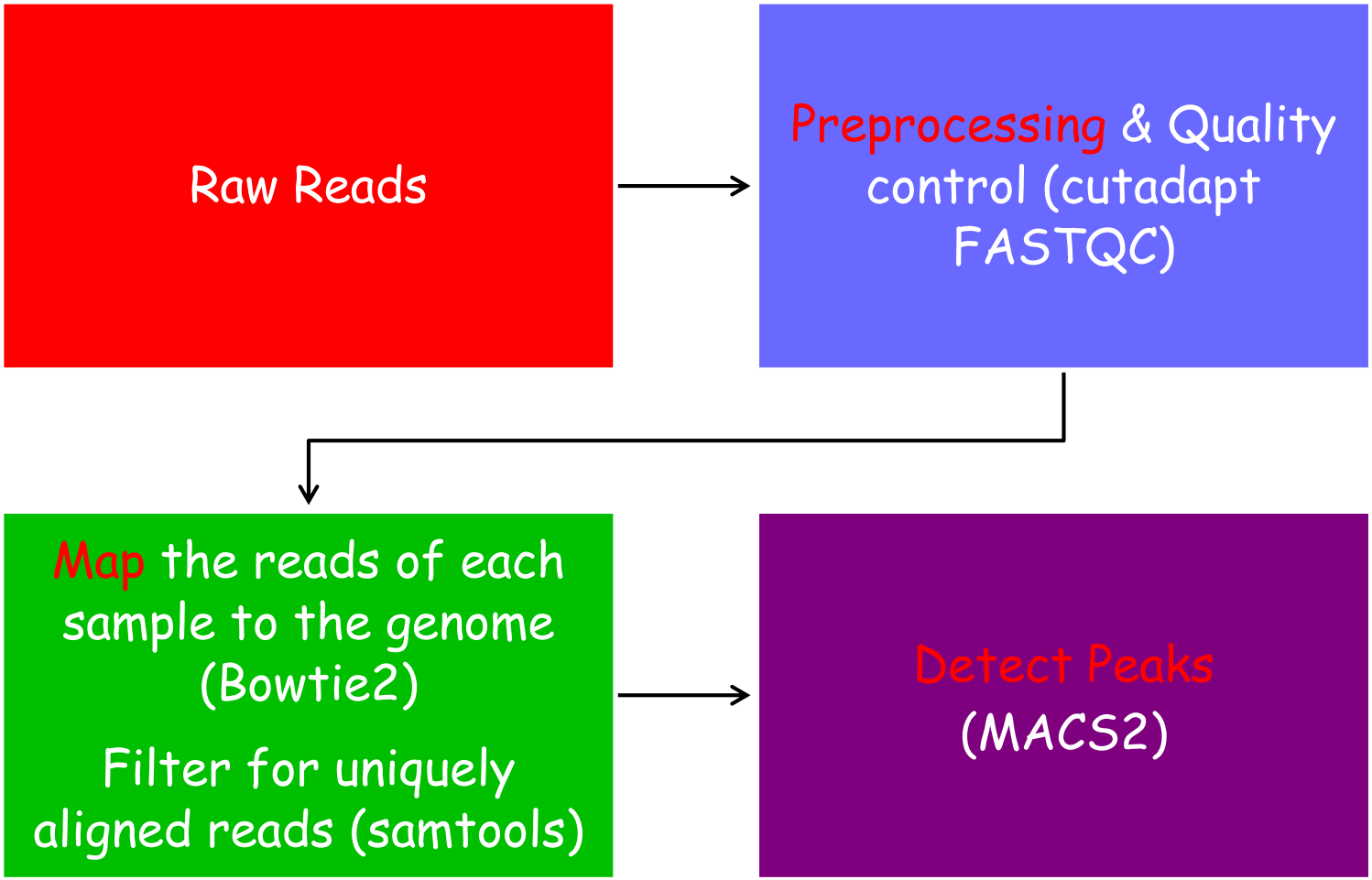
Mapping to genome using Bowtie2

```
HISEQ:153:H8PNPADXX:1:1106:16824:79365  83      chr1    4839877 255     101M    =       4839864 -114
GGTAACTTTCTGTAGTTGGTAAGCCTACCAAGAATTCATGGTTACATCAGGCATCTTATTTTTAAGATATTTCTTTGAC DDEEDDDEDDDDBFFFFHHHHH
IGEJJJJJJJJJJJJJJIIGJIJJJIJJJJJJJJJJJJJJHHHHHFFFFFCCC NH:i:1  HI:i:1  AS:i:200        nM:i:0  MD:Z:1
```

Mapped Reads

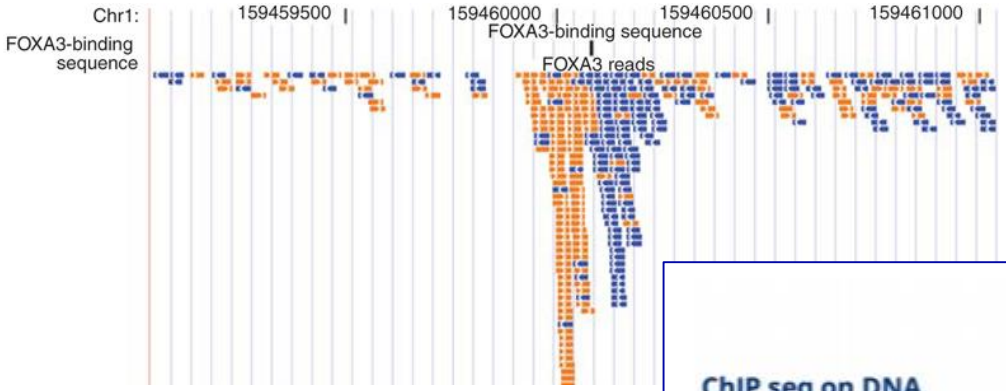
- Instead of sequence we have coordinates and orientation
- Usually exclude sequences that map to more than one location on the genome

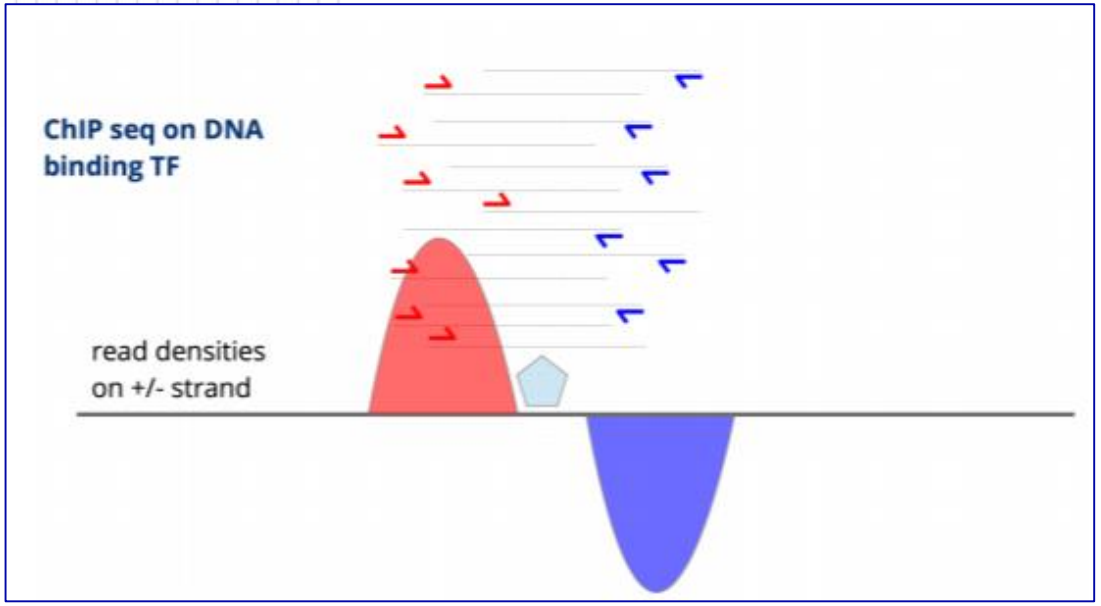
ChIP-Seq Workflow



Detecting the Binding Regions - Peaks

Mapped reads viewed in a genome browser

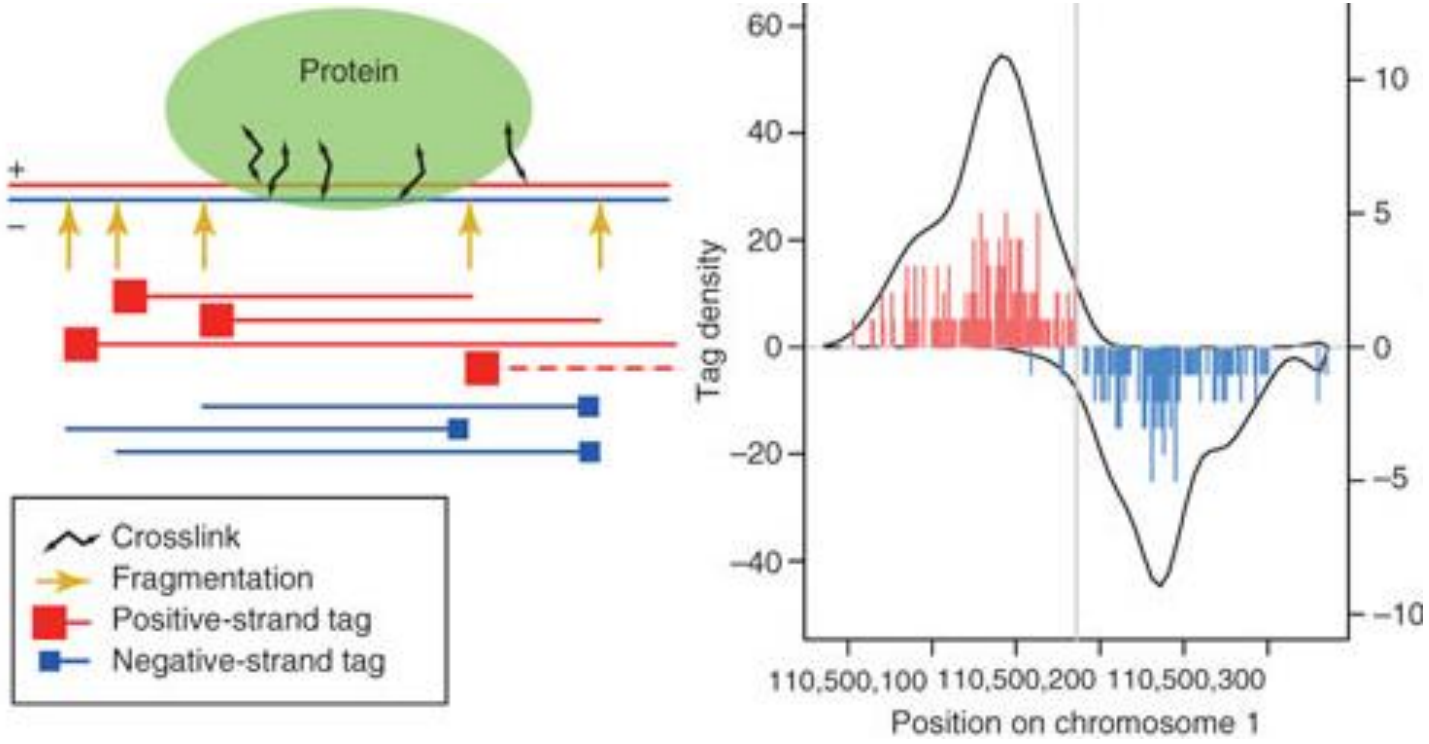




Application note Nature Methods 6, (2009)

http://www.biologie.ens.fr/~mthomas/other/chip-seq-training/booklet/booklet_chip-seq.pdf

ChIP-Seq
Bimodal distribution



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

25 out of 60

Bioinformatics Challenge- Detecting the Binding Regions

Method | Open Access | Published: 17 September 2008

Model-based Analysis of ChIP-Seq (MACS)

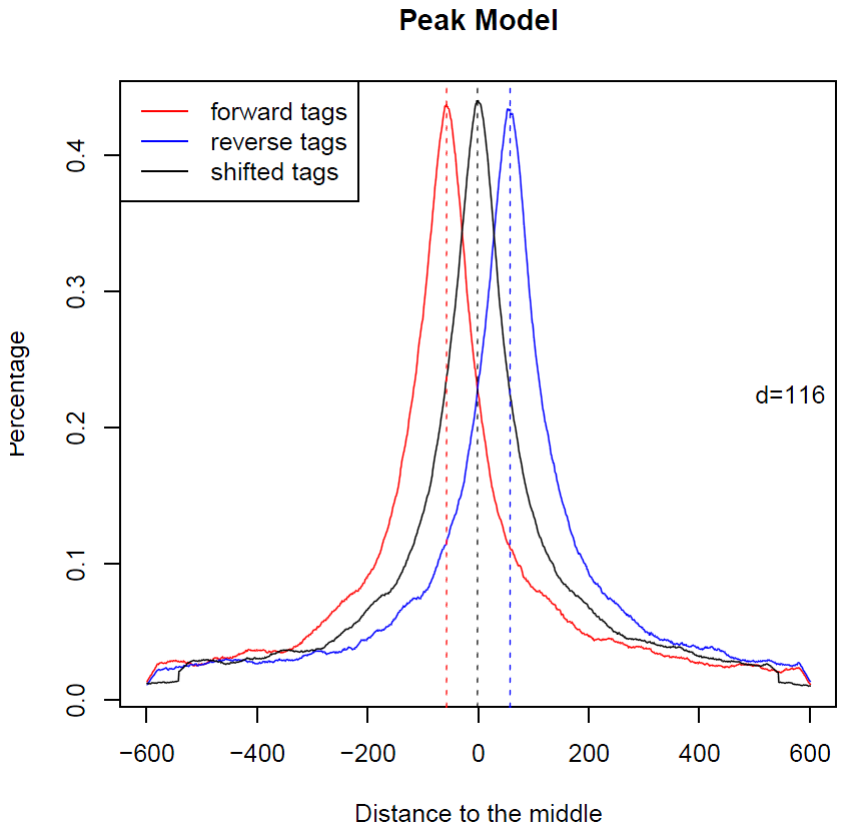
Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li ✉ & X Shirley Liu ✉

Genome Biology **9**, Article number: R137 (2008) | Cite this article

MACS

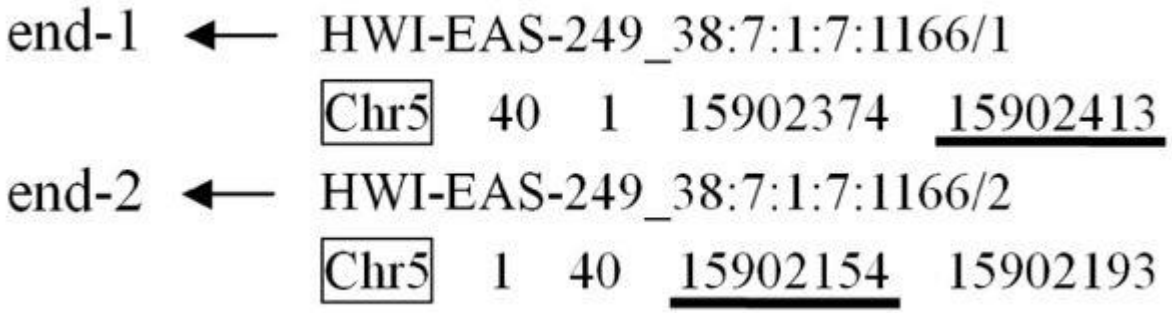
Building peak model

- 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

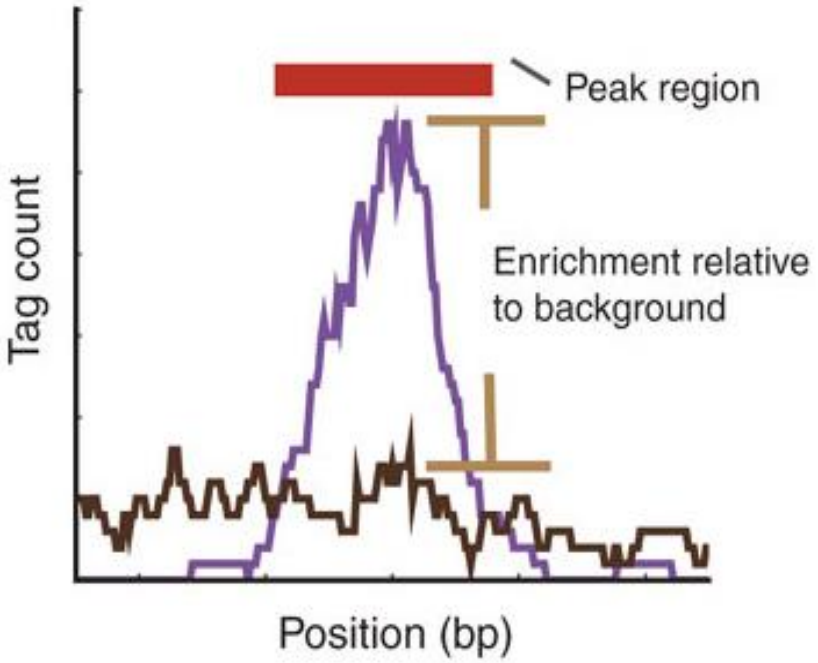


MACS2- Paired end reads

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



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

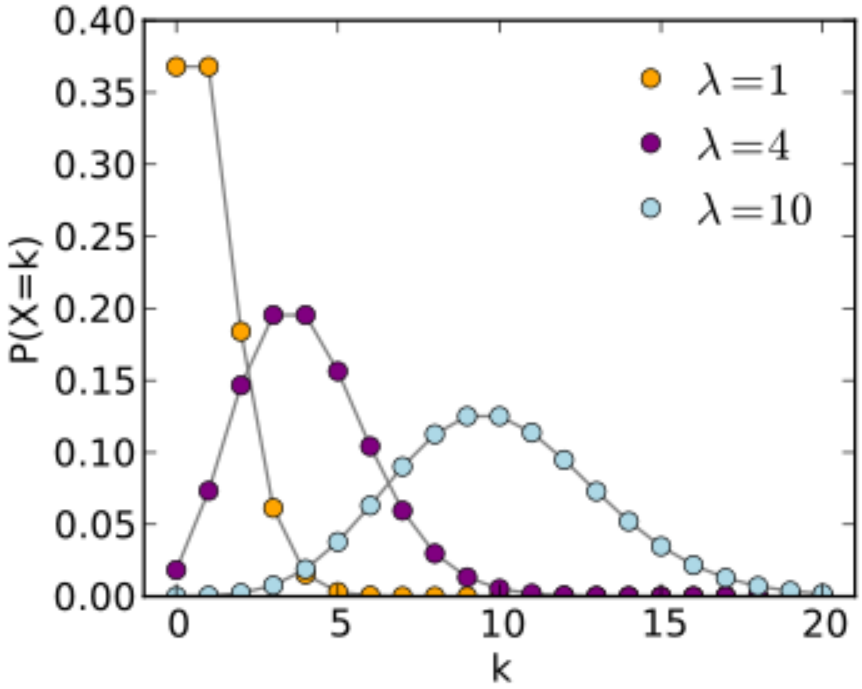
- MACS models the reads distribution along the genome by a Poisson distribution
  - We are counting reads in a fixed region
  - We can compute the expected (mean) number of events, i.e. number of reads in a specified region:

    λ_{BG} = total read counts / effective genome size
  - The expectation is a small number
  - We have many reads
- In this model λ_{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 λ is the expectation

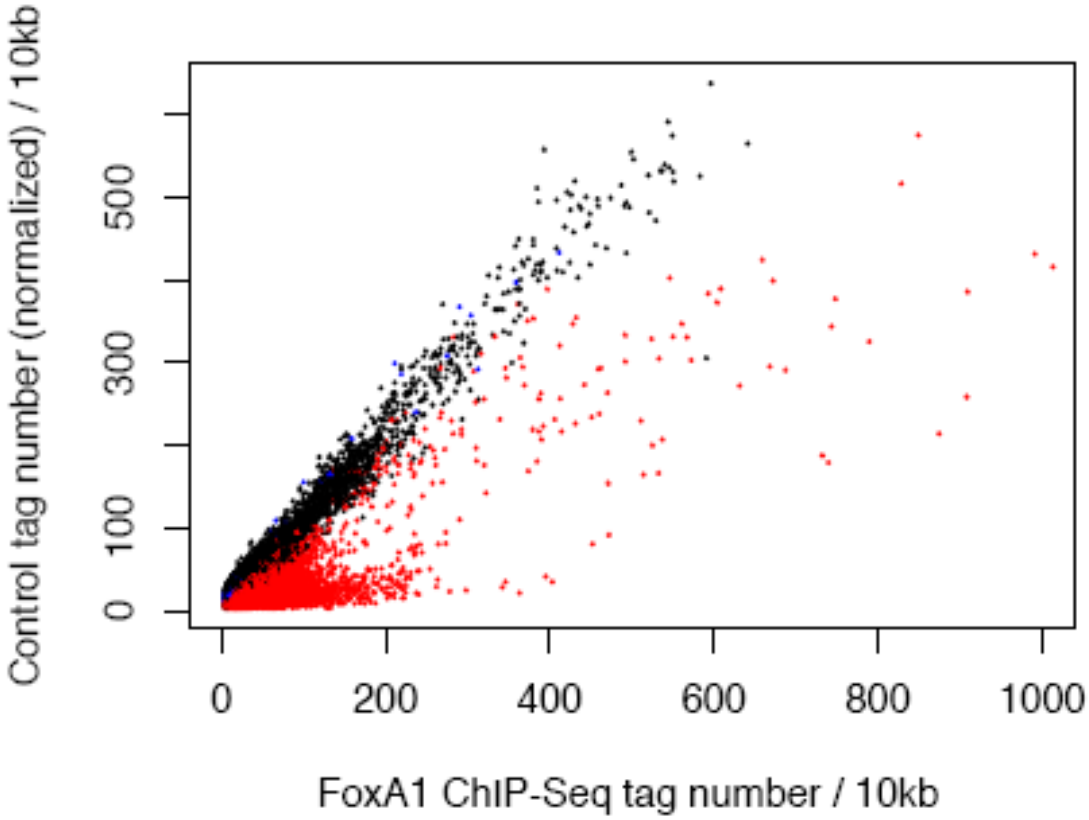


http://en.wikipedia.org/wiki/Poisson_distribution

MACS Peak Detection

ChIP-Seq show local biases in the genome

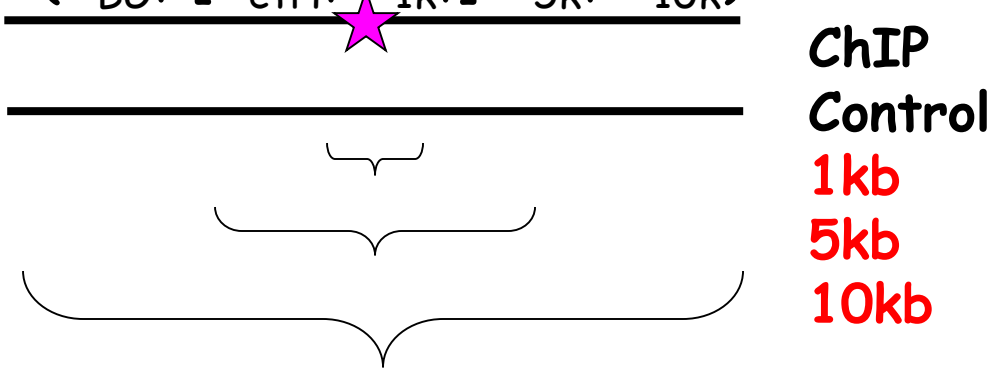
Chromatin and sequencing bias



Peak Calls

- The expected number of reads is
  - λ_{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!

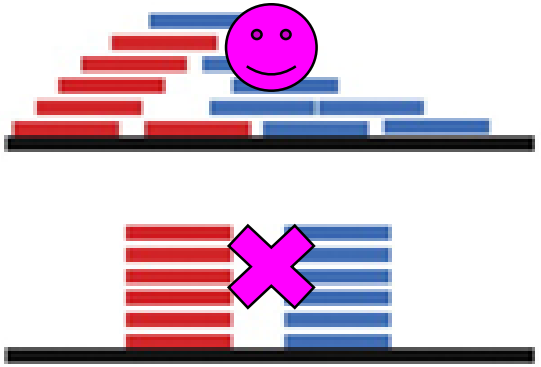
Dynamic λ_{local} = max(λ_{BG}, [λ_{ctrl}, λ_{1k},] λ_{5k}, λ_{10k})

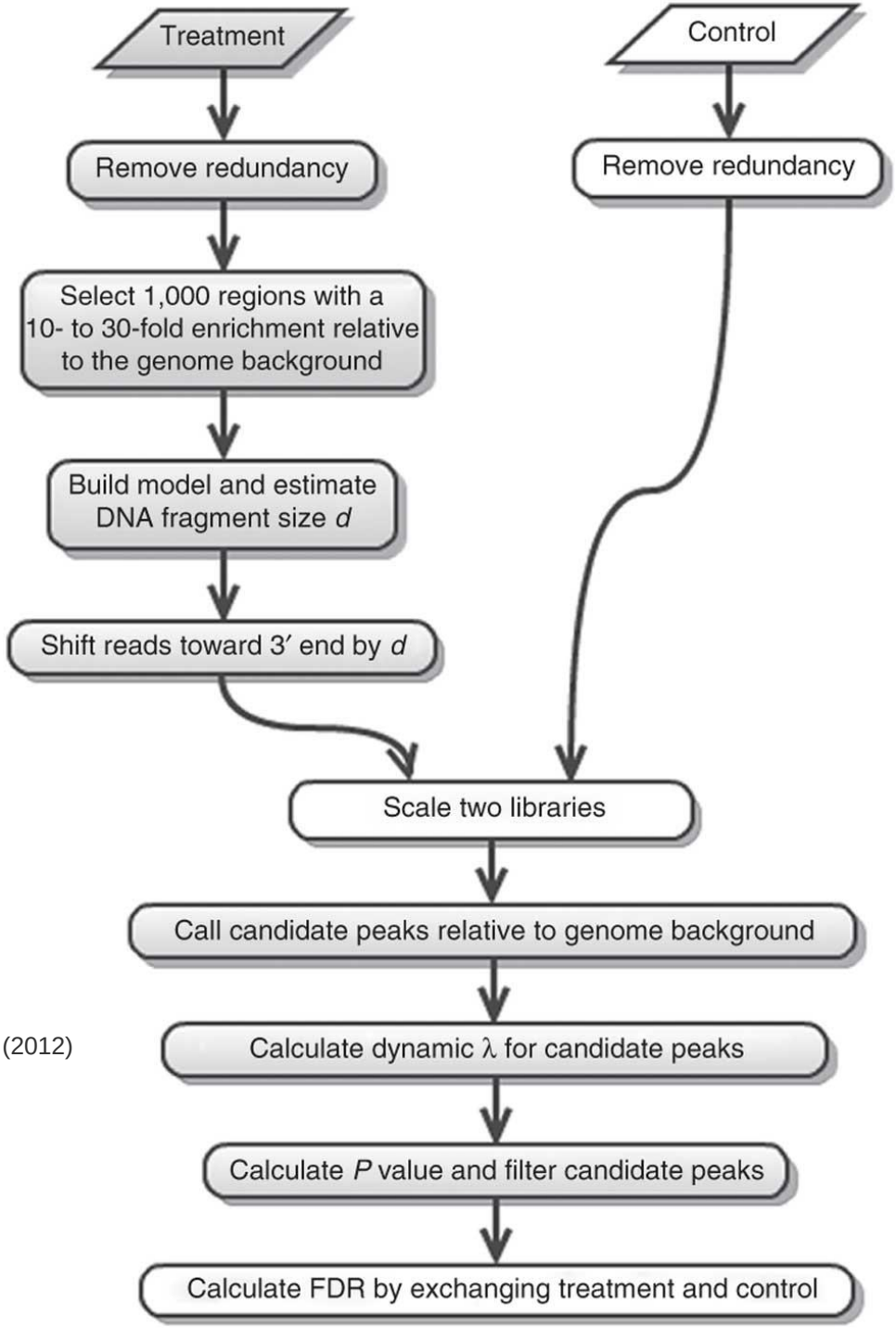


MACS: Zhang et al, Genome Biol 2008

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 predicted redundancy
- Prediction is based on the genome size and the number of reads.
- ENCODE guidelines- 10M uniquely mapped reads should have non-redundant frequency >=0.8





Nature Protocols **7**, 1728–1740 (2012)

37 out of 60

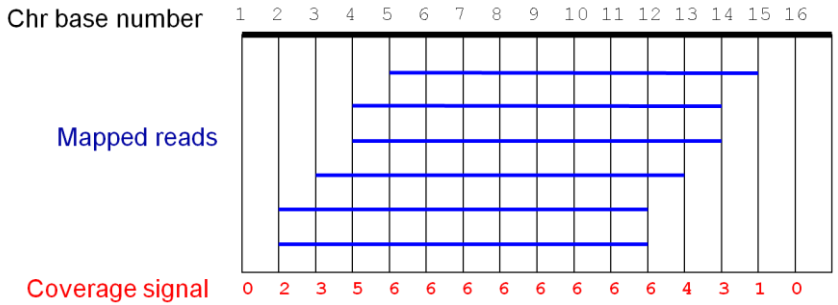
MACS Peak Information (.xls)

- **Summit**
  peak summit position related to the start position of peak region
- **Tags**
  number of tags-reads in peak region
- **-10*$\log_{10}$(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 λ

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

MACS: Shifted Wiggle Files

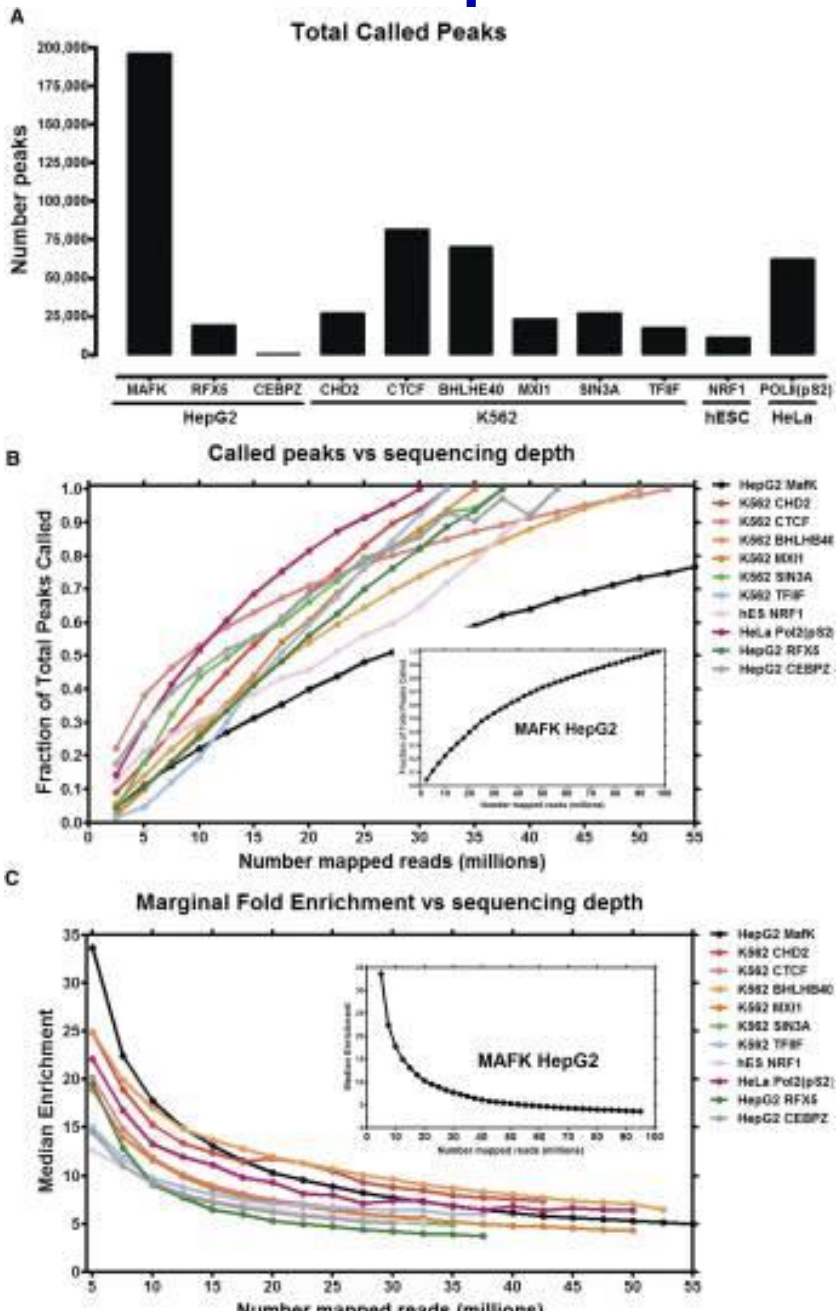
Shifted reads displayed as a coverage signal



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



Genome Res. 2012 Sep; 22(9): 1813–1831.

40 out of 60

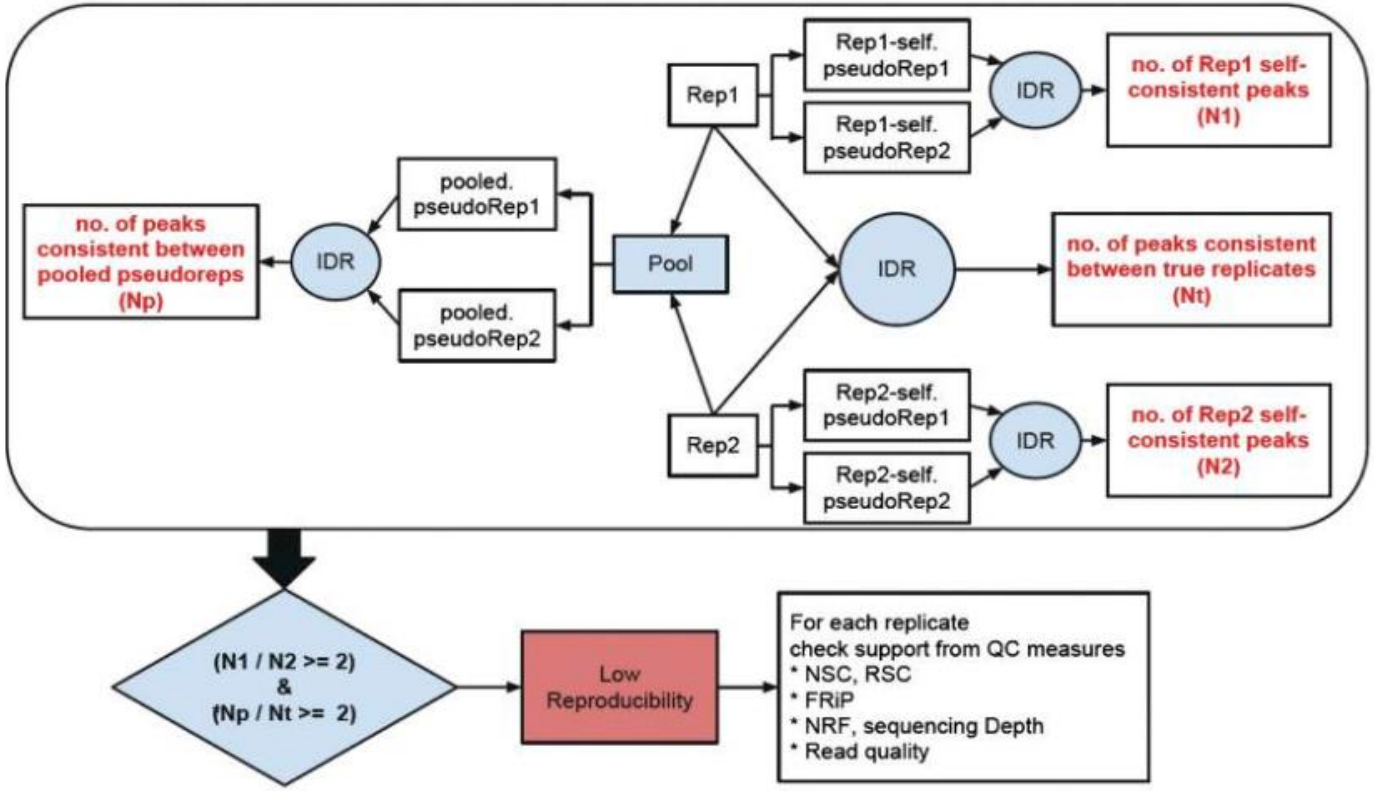
Biological Replicates ENCODE

- 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 meet a third replicate is required.

Biological Replicates Evaluation

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



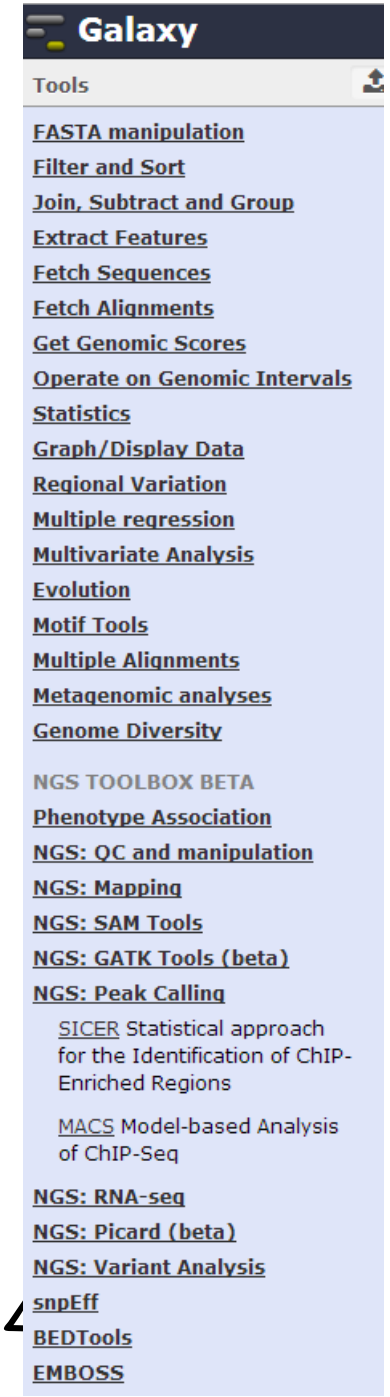
This figure is taken from the ENCODE ChIP-seq Guidelines.

How to Run MACS

- As a command line program (on a Linux server)
- 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



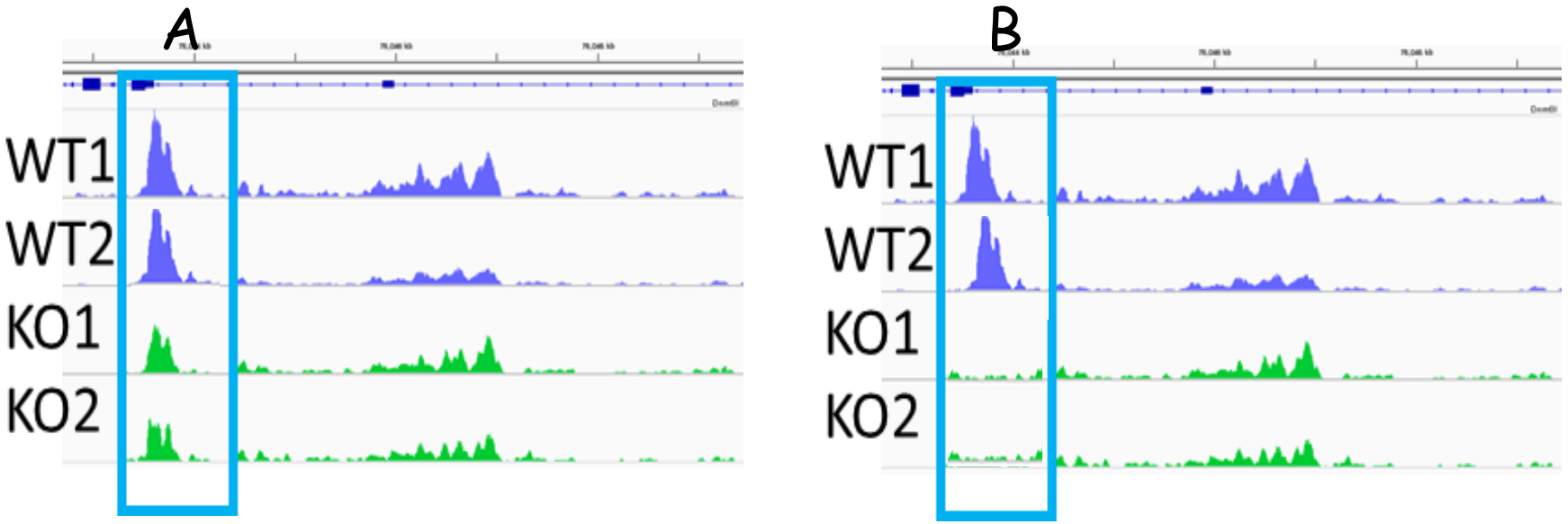
Public Tools for ChIP-Seq Analysis

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

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome 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 *P* values	*P* 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)	*s* 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
GLITR	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 *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* 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	*q* 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	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SISSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson *P* value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	*P* value	1: Monte Carlo simulation 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial *P* value	Estimated or user specified	Subtracted before peak calling	*q* value	1, 2: binomial 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Target FDR	No / Yes	20

[a]The labels 1: and 2: refer to one-sample and two-sample experiments, respectively. [b]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. [c]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. # control / # ChIP, number of peaks called with control (or some portion thereof) and sample reversed. [d]The lists of 'user input parameters' for each program are not exhaustive but rather comprise a subset of greatest interest to new users. [e]'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 pileups that may be associated with, for example, microsatellite DNA. [f]N_+ and N_- are the numbers of positive and negative strand reads, respectively.

Comparing Peaks – Binding Between Samples

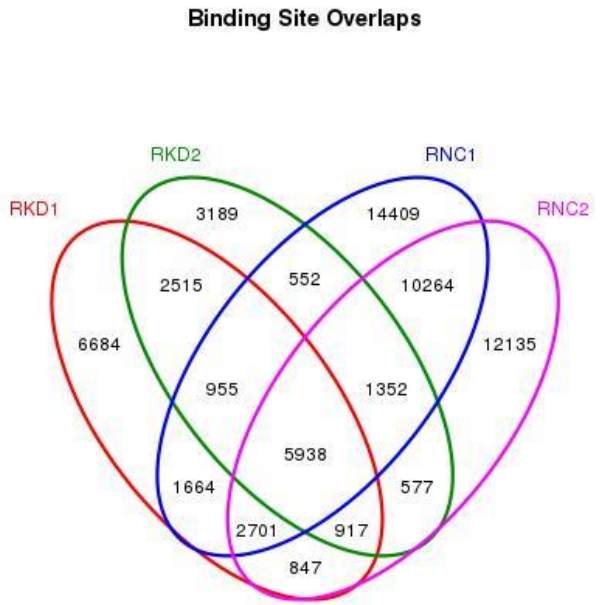


In which of the above is there a difference between WT and KO binding?

45 out of 60

Approaches to Compare Peaks

Venn Diagram
using peaks regions



DiffBind: Identify regions that are **differentially enriched between two or more sample groups**
Extract fragment counts for all peaks in all samples and perform a statistical test, such as DESeq2

What is required to perform a statistical analysis between the groups?

46 out of 60

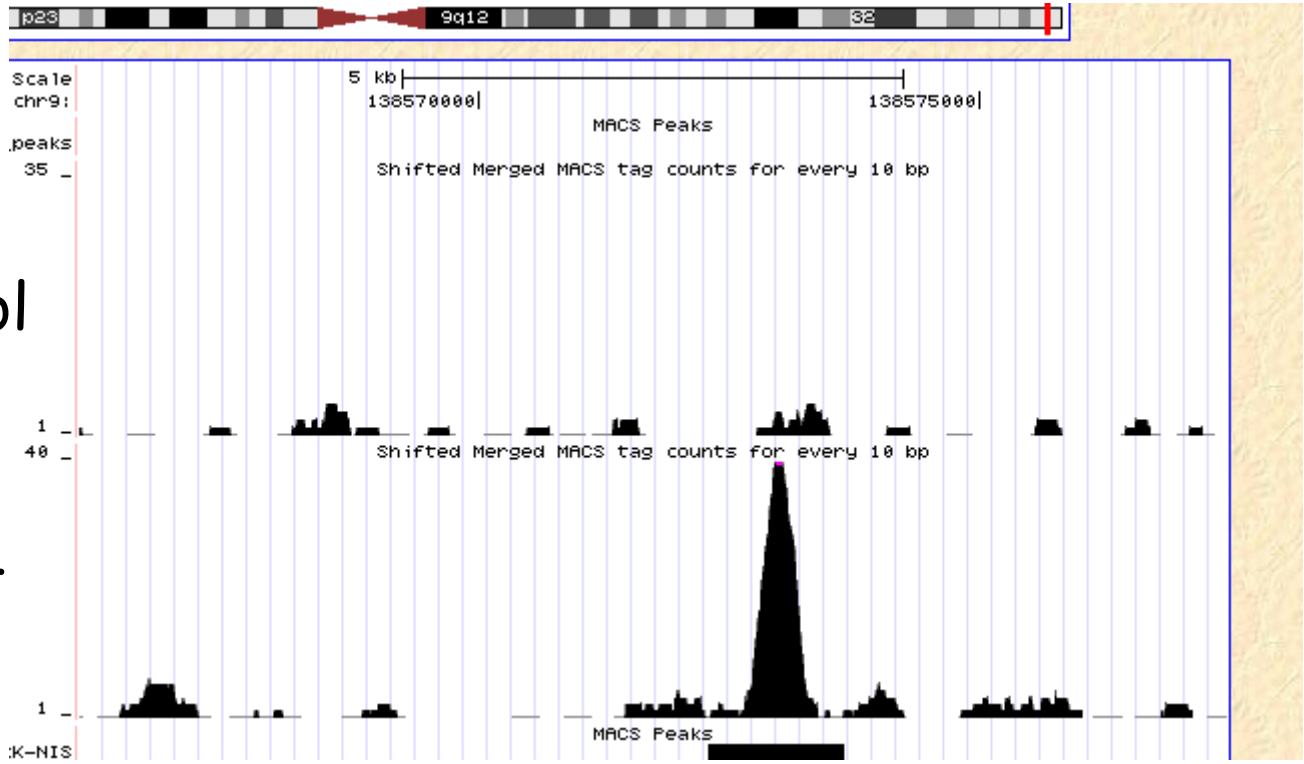
ChIP-Seq – Lecture Outline

- 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

control

treat



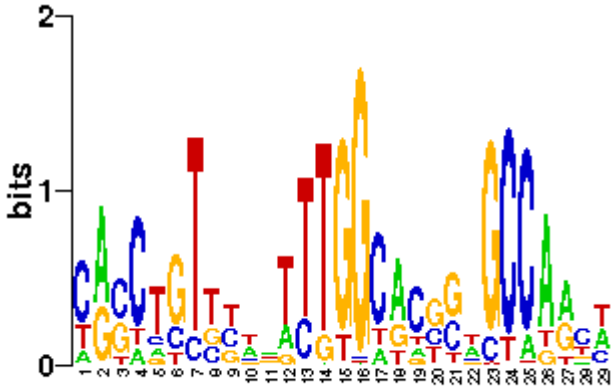
TF Motifs Tools

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

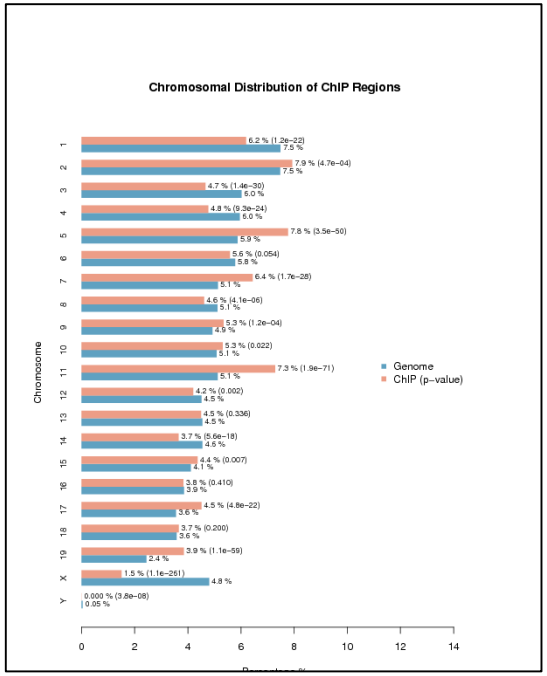
Recommended portals:

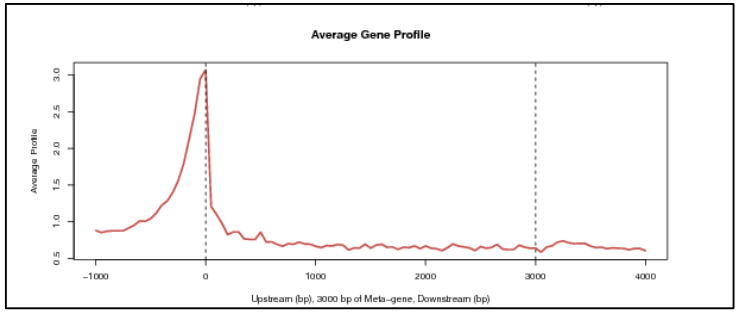
- MEME-ChIP (web)
- Homer – command line

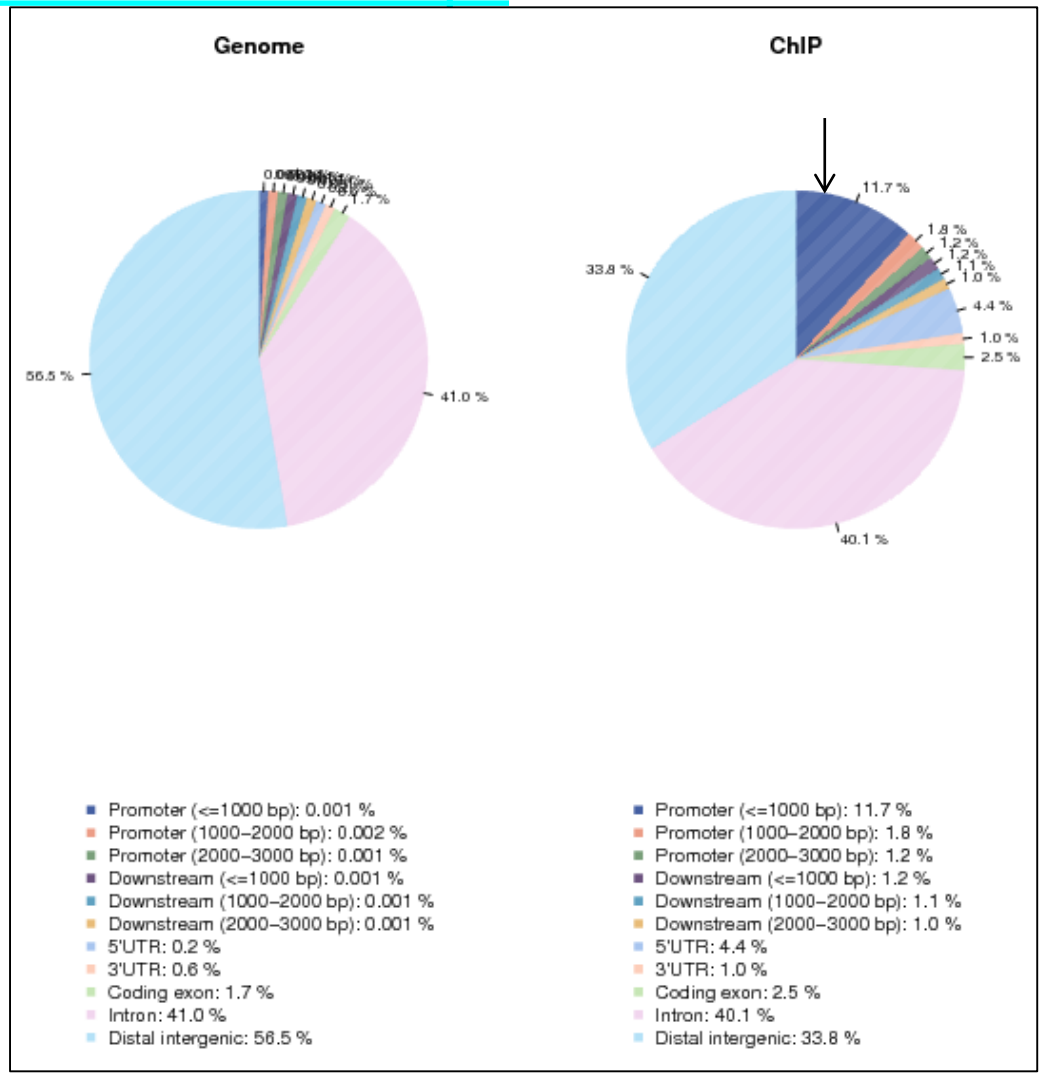


CEAS: Enrichment of Genome Features

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

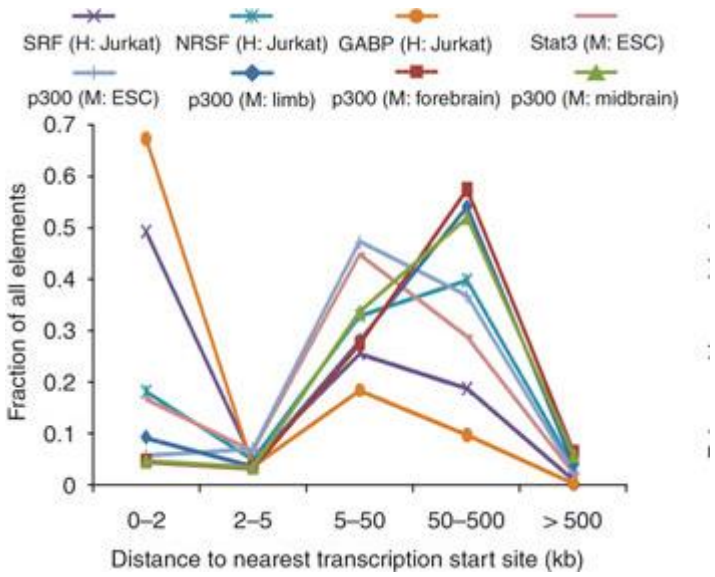






Functional Interpretation

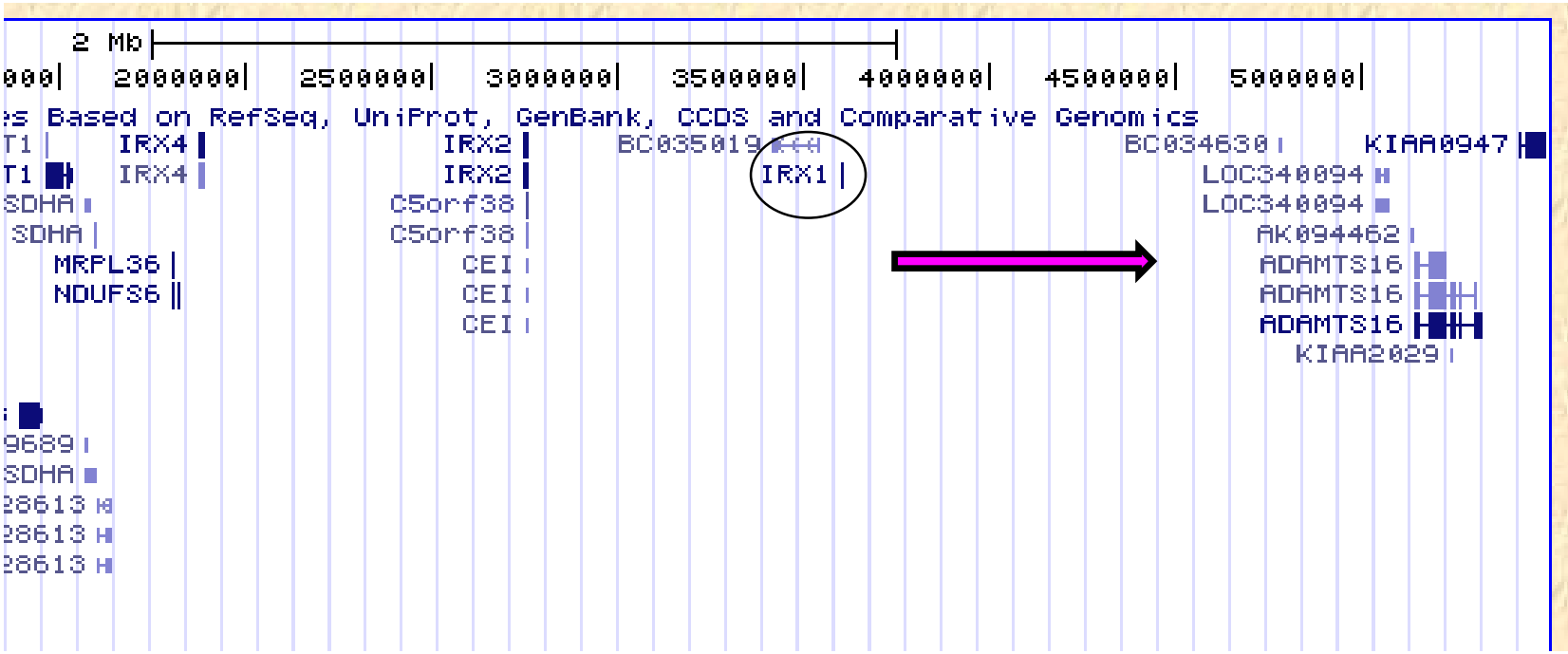
- Which processes and functions does our TF regulate?
  - Associate peaks with genes
  - Associating only proximal genomic regions to genes (<5 kb) – for most TF ignores a large fraction of binding data



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

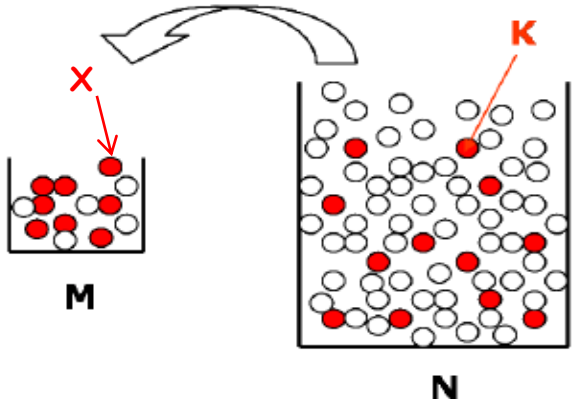
- Ontology term 1: gene1,gene2,gene3 …
- Ontology term 2: gene2,gene5,gene8 …
- …
- Are my peaks located near genes enriched for certain ontology terms?

- Which statistical test should we apply?

Gene List Enrichment Test

- The hypergeometric test is the standard gene enrichment test for gene lists (such as differentially expressed genes)
- 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.

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



NATURE BIOTECHNOLOGY | RESEARCH | ANALYSIS

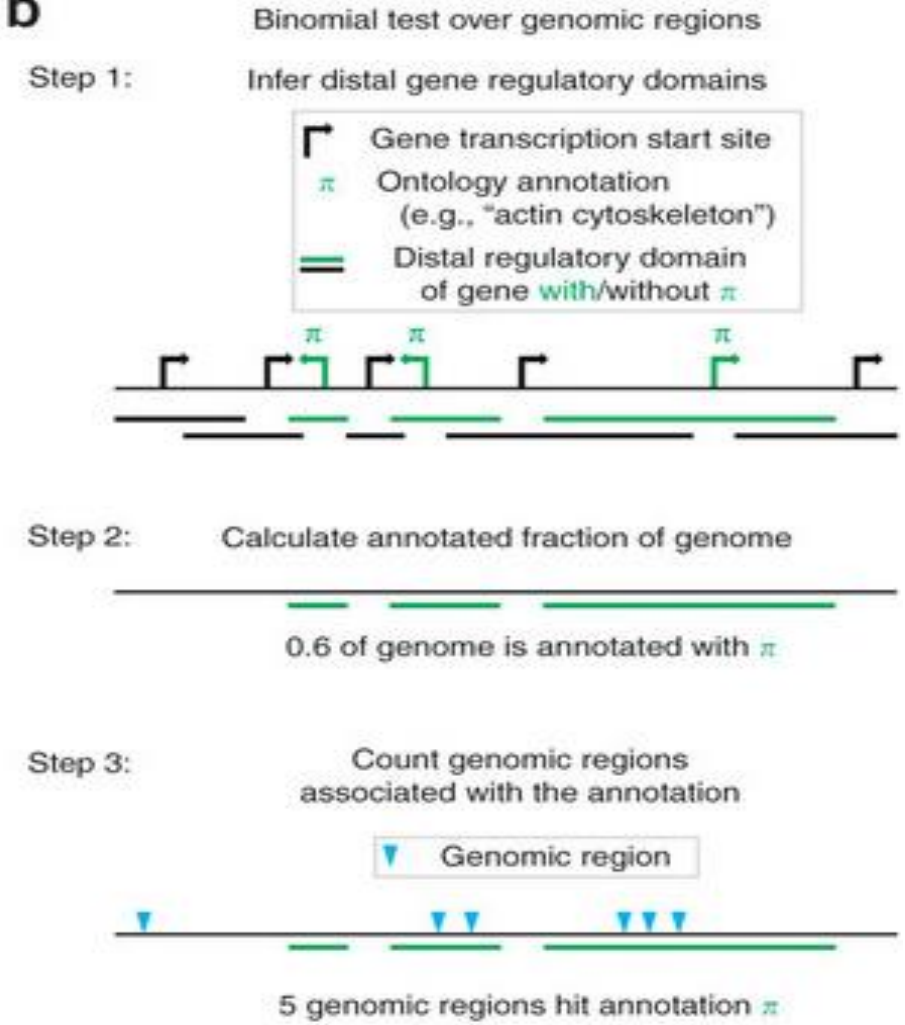
GREAT improves functional interpretation of *cis*-regulatory regions

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



Nature Biotechnology 28, 495–501 (2010)

56 out of 60

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: NEW [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].

57 out of 60

Binomial vs Hypergeometric

Binomial test calculated for a set of genomic regions

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

Hypergeometric test calculated for a set of genes

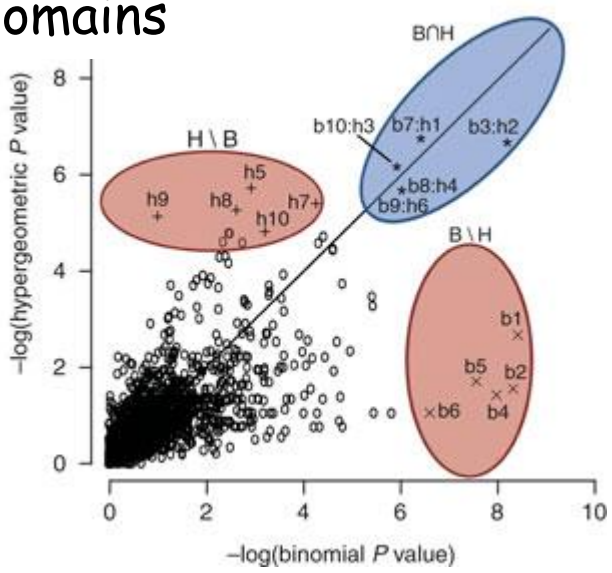
A gene based approach would expect 14% of the genes near peaks to be associated with 'multicellular organismal development'

What is the reason for this discrepancy?

GREAT Uses Both the Hypergeometric and the Binomial Test

Significant by hypergeometric : general terms arising from genes with long 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)

ChIP-Seq – Lecture Outline

- Experimental issues: how is a ChIP-seq experiment done?
- Analysis of the sequence data: how to detect the binding regions-peaks, evaluate replicates and compare between samples
- Downstream analysis: how to extract the biological relevance?
  - Genome Browser
  - DNA motifs enriched
  - Enrichment of genomic features (promoters…)
  - Associate to genes
  - Pathway enrichments

61 out of 60