מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

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

Dena Leshkowitz
Bioinformatics Unit
WIS
March 2016

Definition

- ChIP-seq is short for

  chromatin **i**mmuno-**p**recipitation followed by **seq**uencing

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

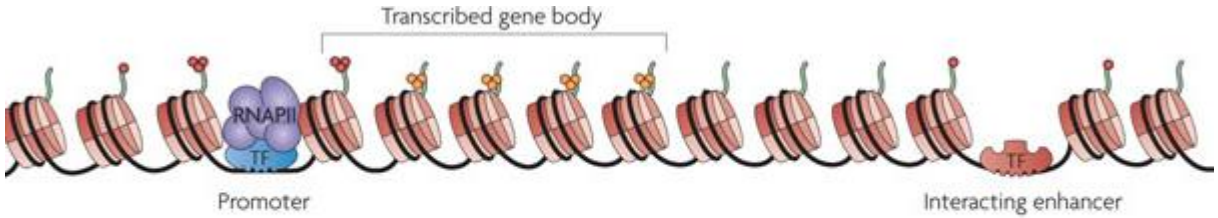
ChIP-Seq – Lecture Outline

- 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 regions biological relevance?

Transcription Regulation

Characterize DNA-protein interactions in vivo, such as:

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



Nature Reviews | Genetics

Epigenomics: Roadmap for regulation

Casey E. Romanoski, Christopher K. Glass, Hendrik G. Stunnenberg, Laurence Wilson & Genevieve Almouzni

Affiliations | **Corresponding authors**

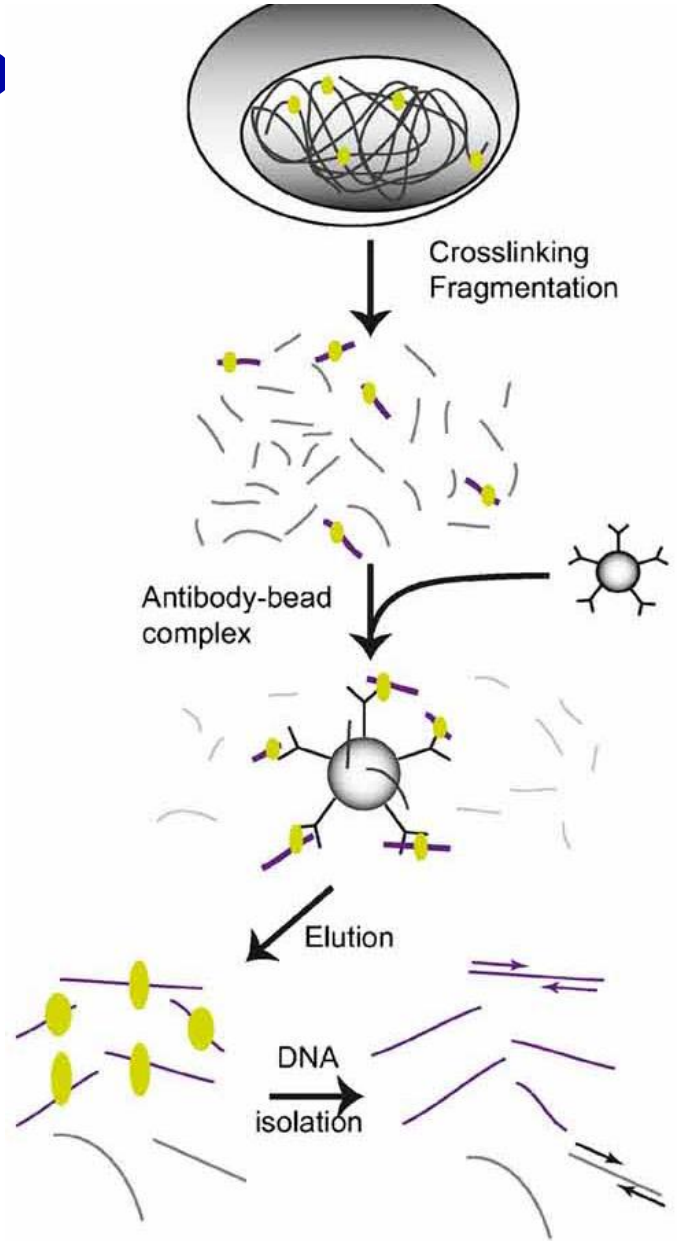
Nature **518**, 314–316 (19 February 2015) | doi:10.1038/518314a
Published online 18 February 2015

www.itsjustabadday.com

ChIP – How is it done?

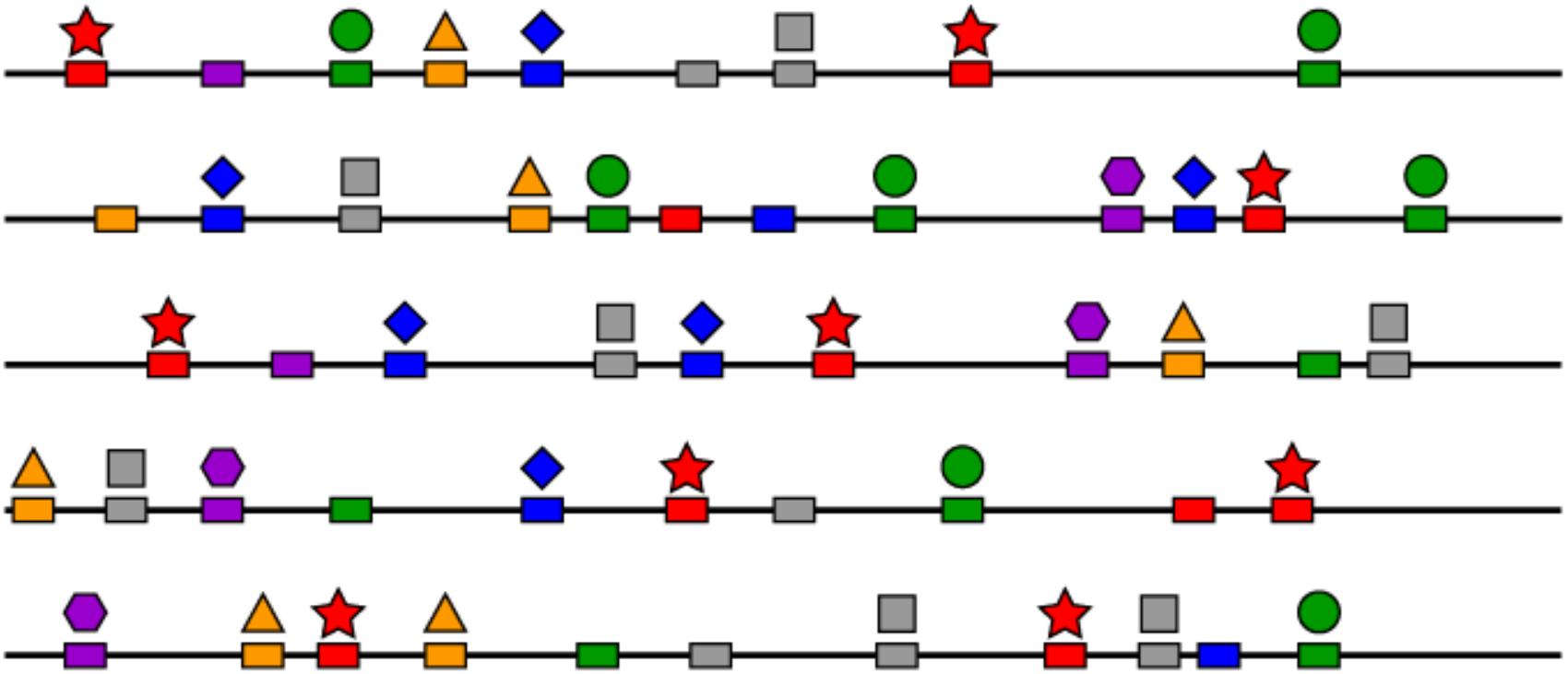
ChIP is a technique that permits to

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

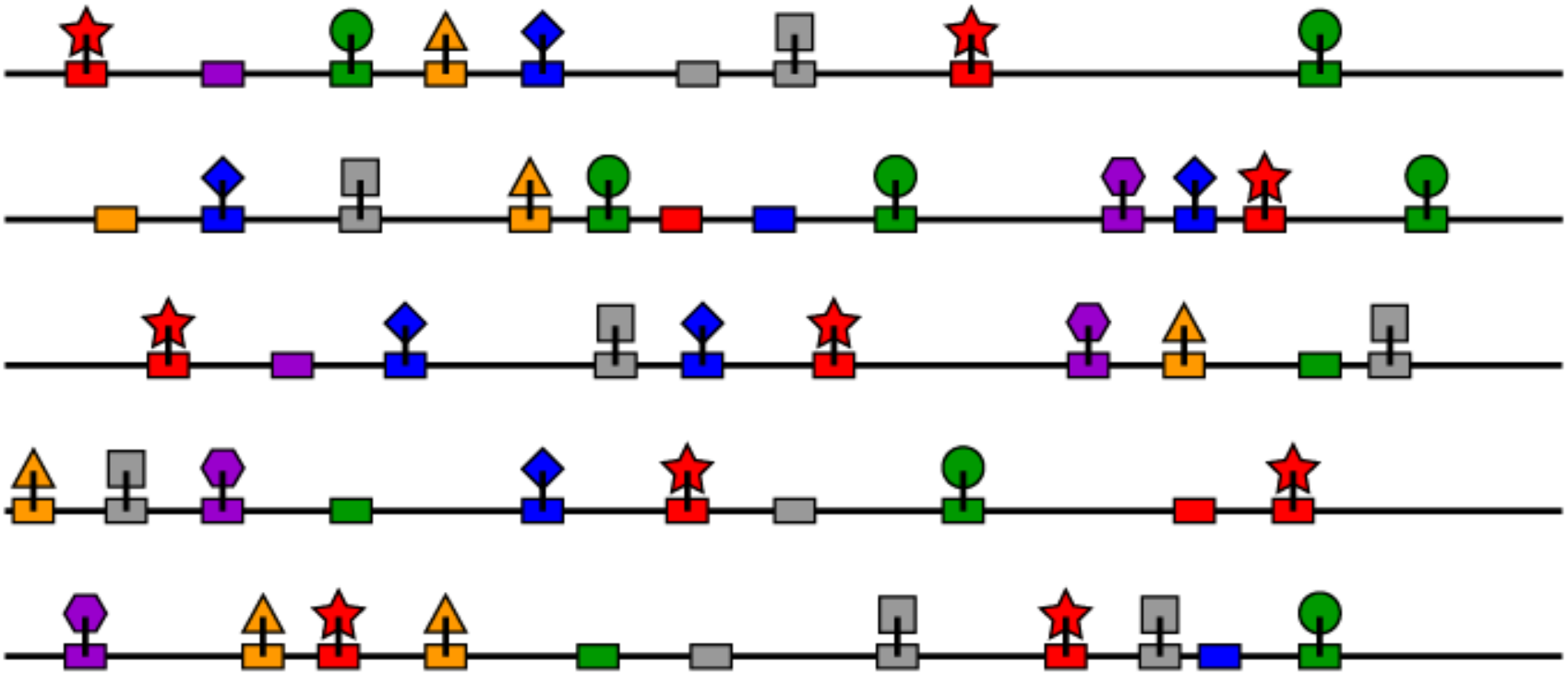


Barski and Zhao J Cell Biochem. 2009 May 1;107(1):11-8(2009)

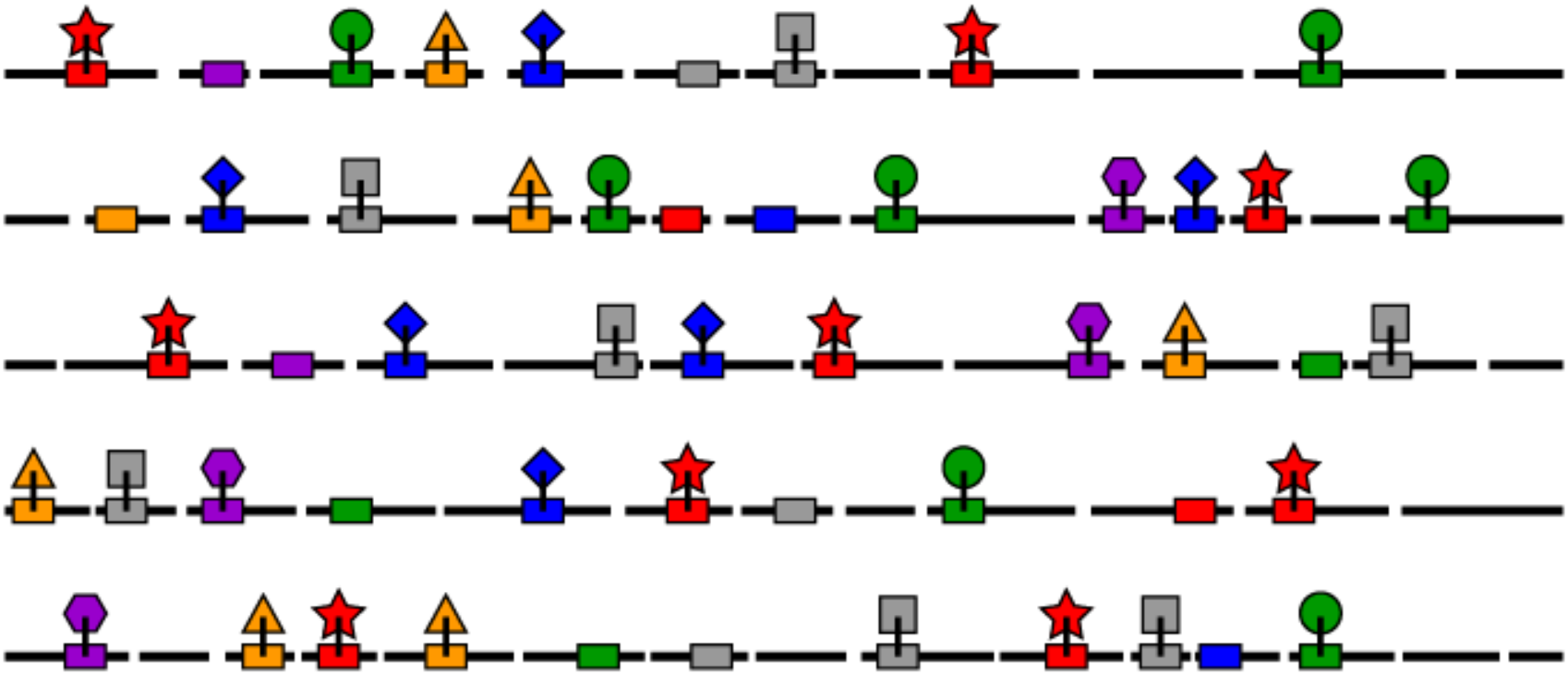
Chromatin ImmunoPrecipitation (ChIP)



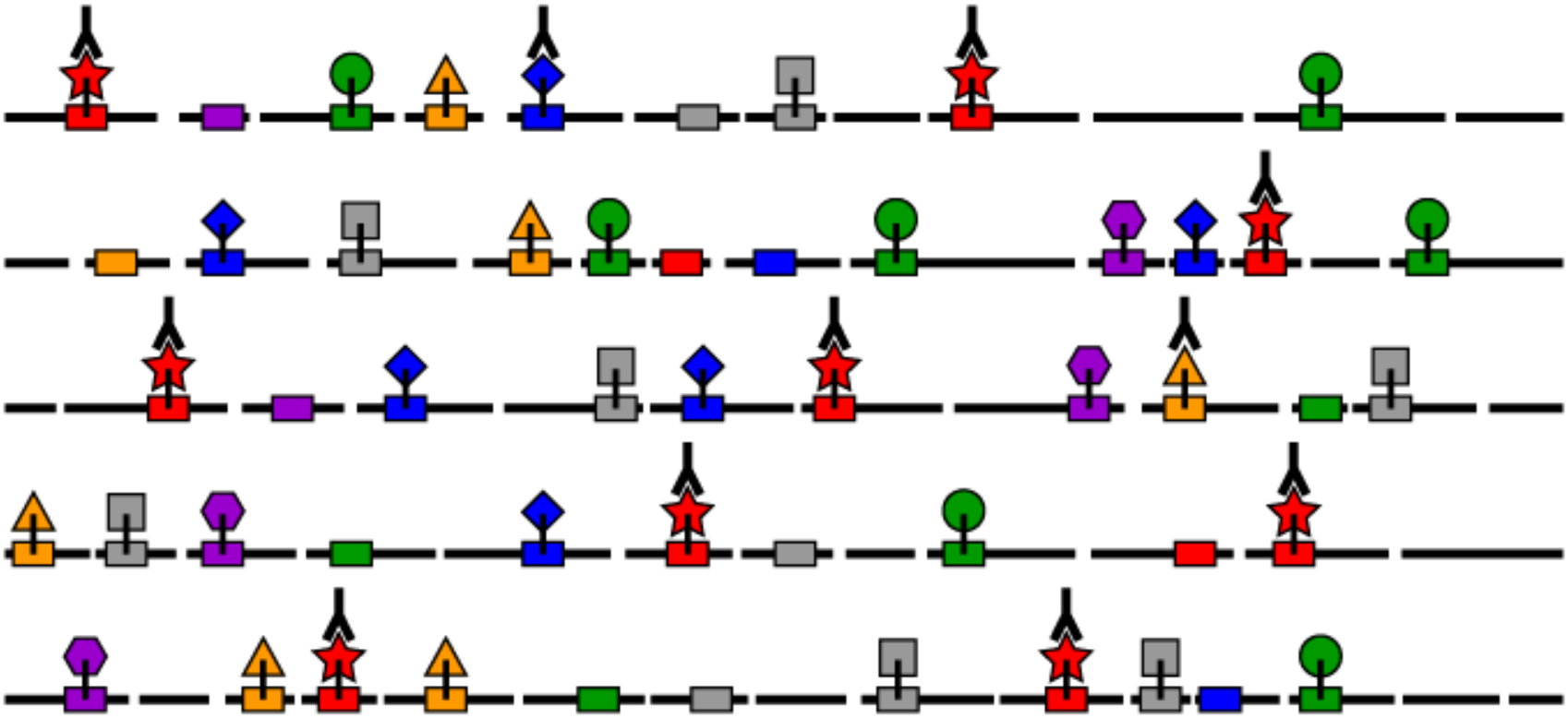
TF/DNA Crosslinking *in vivo*



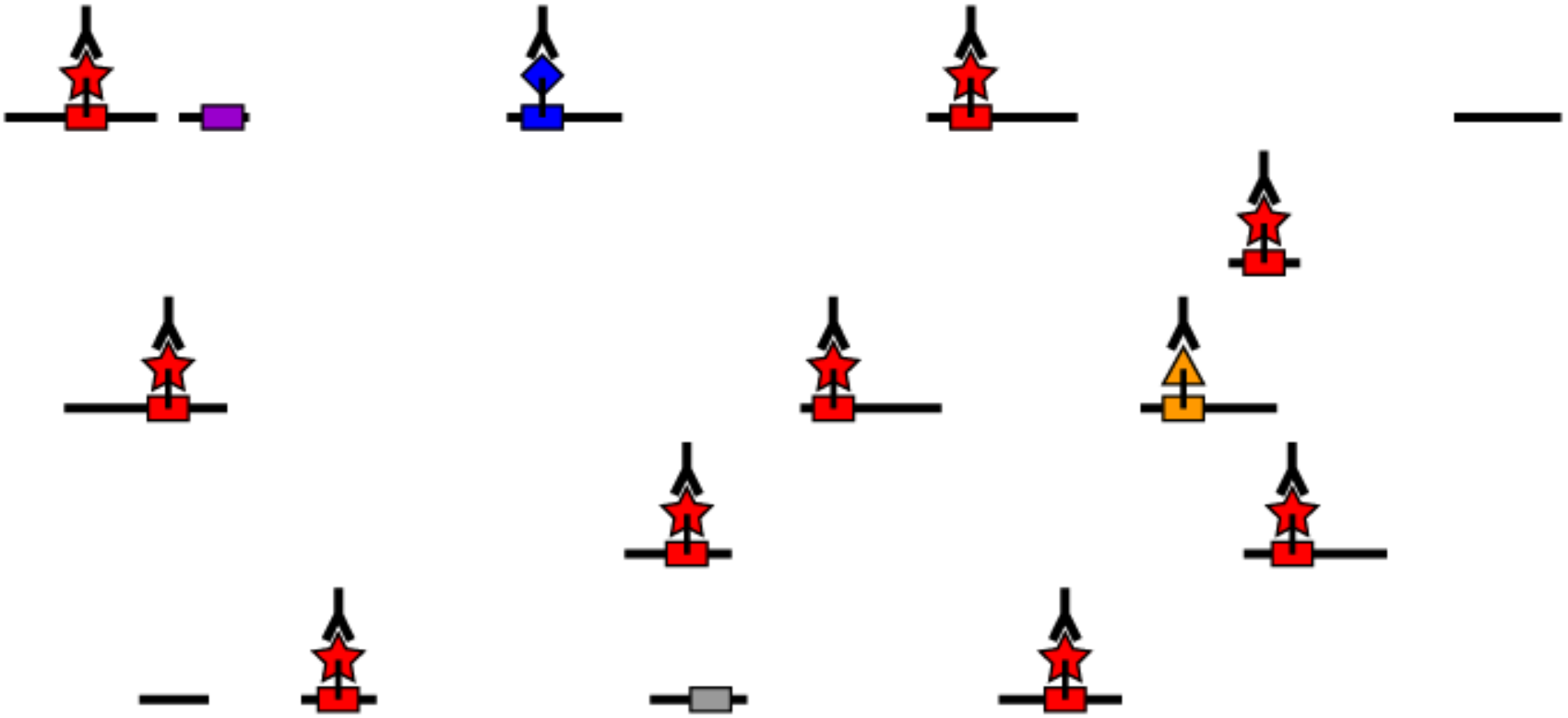
Sonication (~200bp)



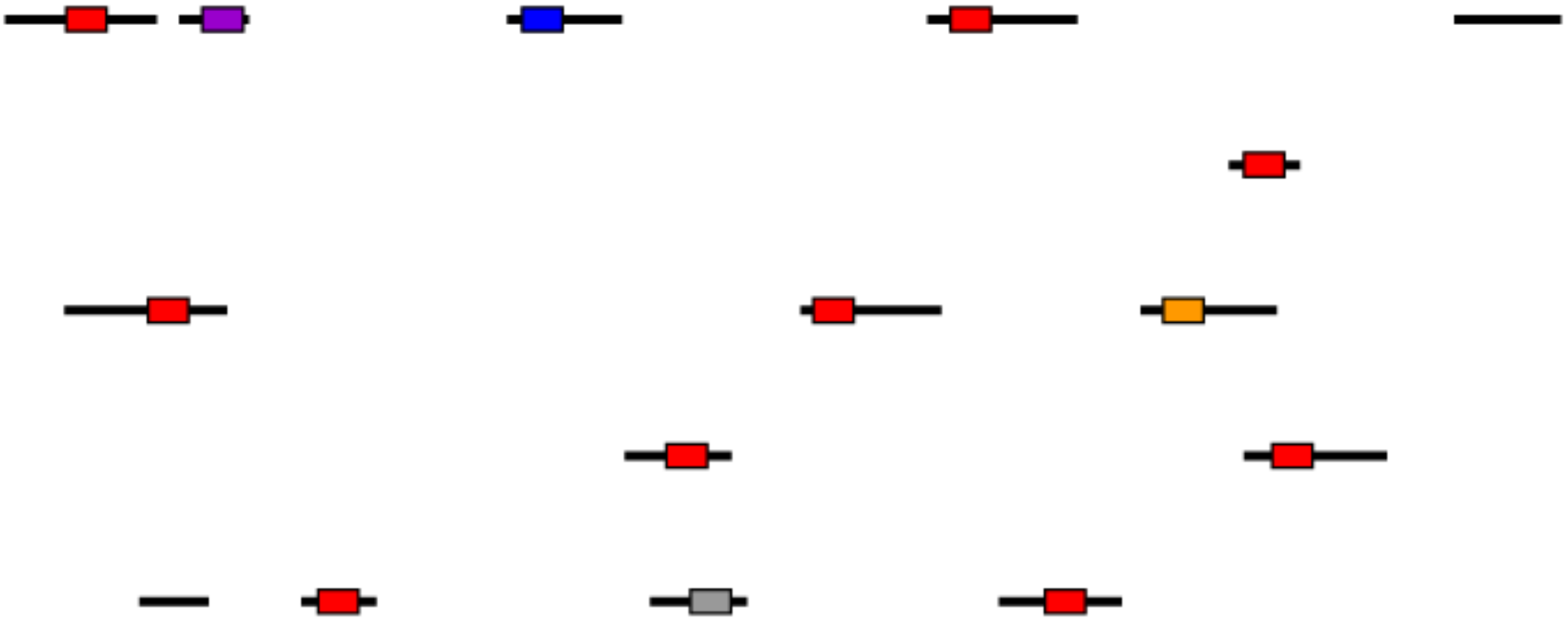
TF-specific Antibody



Immunoprecipitation



Reverse Crosslink and DNA Purification



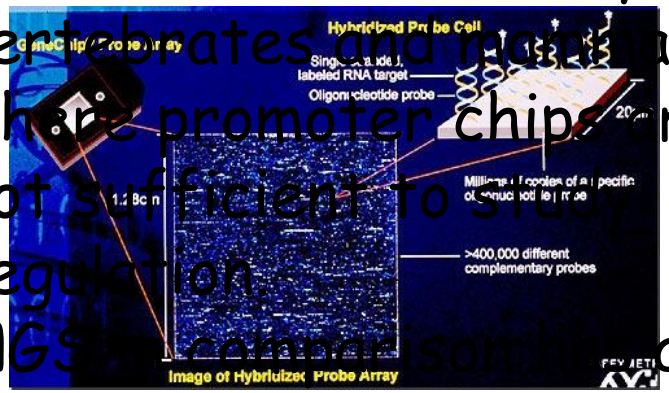
ChIP High-throughput technology

Discover the DNA binding regions in the genomic scale

DNA Microarrays

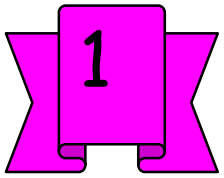
More attractive mostly for vertebrates and mammals where promoter chips are not sufficient to study regulation.

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



ChIP-Chip

Next Generation Sequencing



ChIP-Seq

NGS Technologies

	454 (Roche)	Illumina HiSeq2500	Solid (ABI)
			
Seq method	Pyro-sequencing	Seq-by-Synthesis (SBS)	Seq by oligo ligation

Experimental Design

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

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

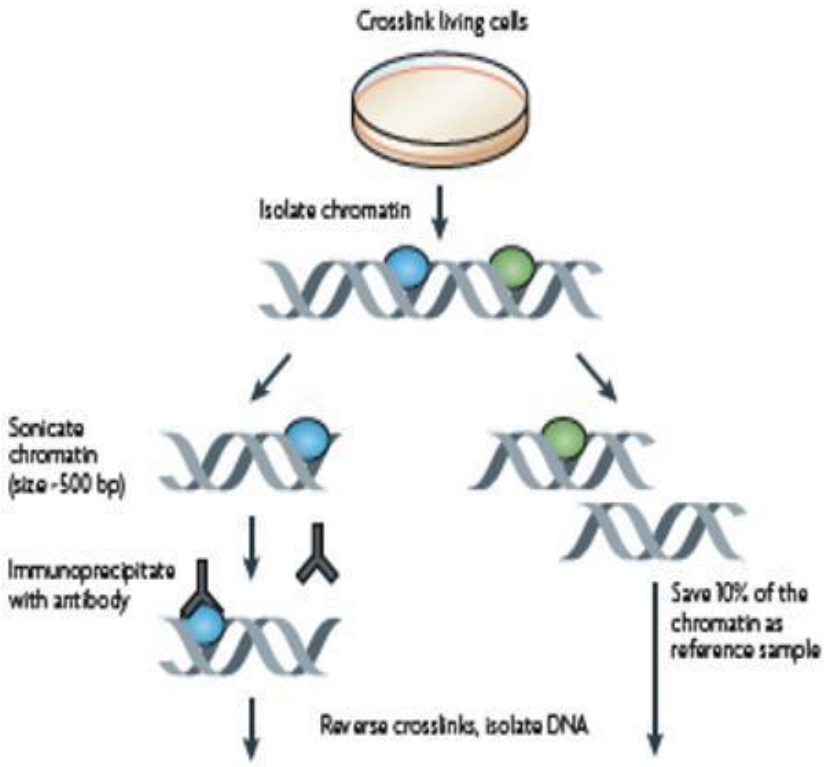
  ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.

- **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



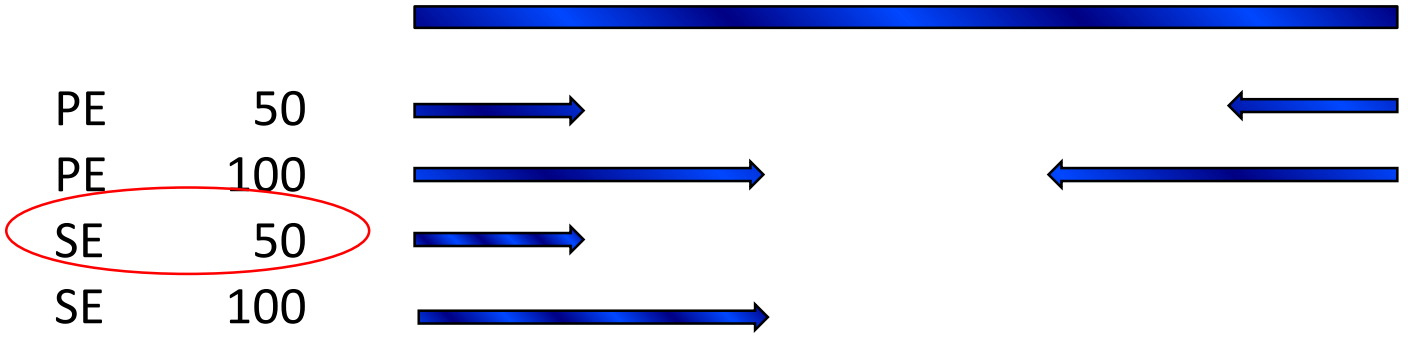
Nat Rev Genet. 2009 10(9): 605-616.

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

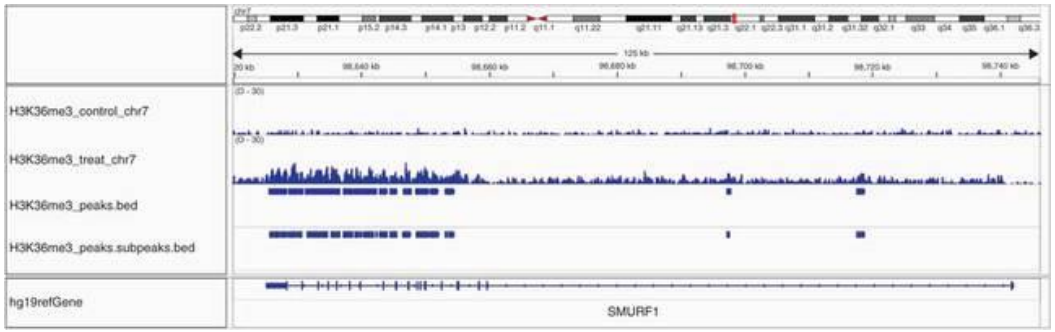
- Use validated antibody maybe even two different antibodies
- Need to have biological replicates
- If we have a good reference genome (mouse, human…) no need to sequence long reads (>50 bases) and no need in paired-end sequencing



- How much reads (sequences) per sample?

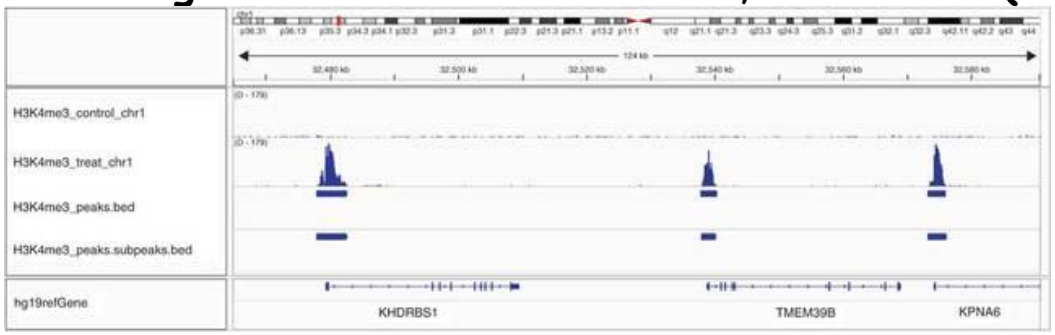
Coverage Required - Types of Peaks

H3K36me3 (exon regions)



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

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

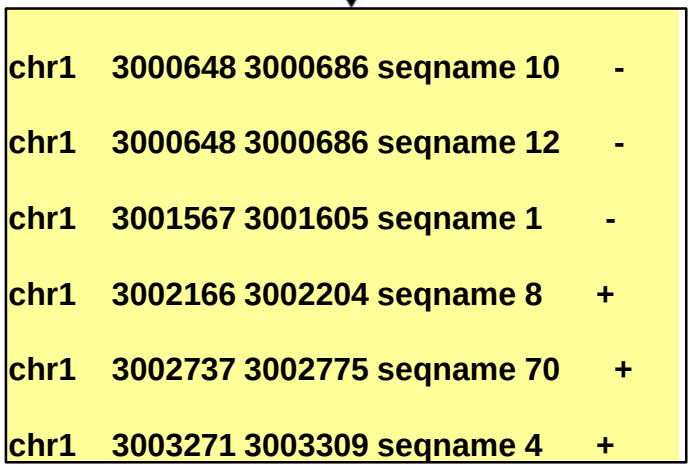
- Experimental issues: how is a ChIP-seq experiment done?
- Analysis of the sequence data: how to detect the binding regions?
- Downstream analysis: how to extract the biological relevance?

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 (bed file)
- Usually exclude sequences that map to more than one location on the genome

Sequencing

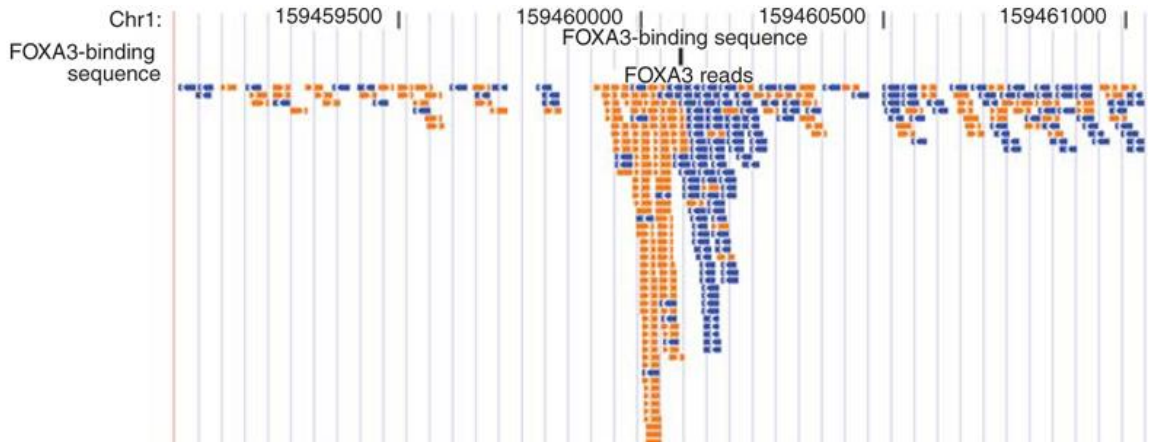


chr1	3000648	3000686	seqname 10	-
chr1	3000648	3000686	seqname 12	-
chr1	3001567	3001605	seqname 1	-
chr1	3002166	3002204	seqname 8	+
chr1	3002737	3002775	seqname 70	+
chr1	3003271	3003309	seqname 4	+

Mapped Reads – bed format

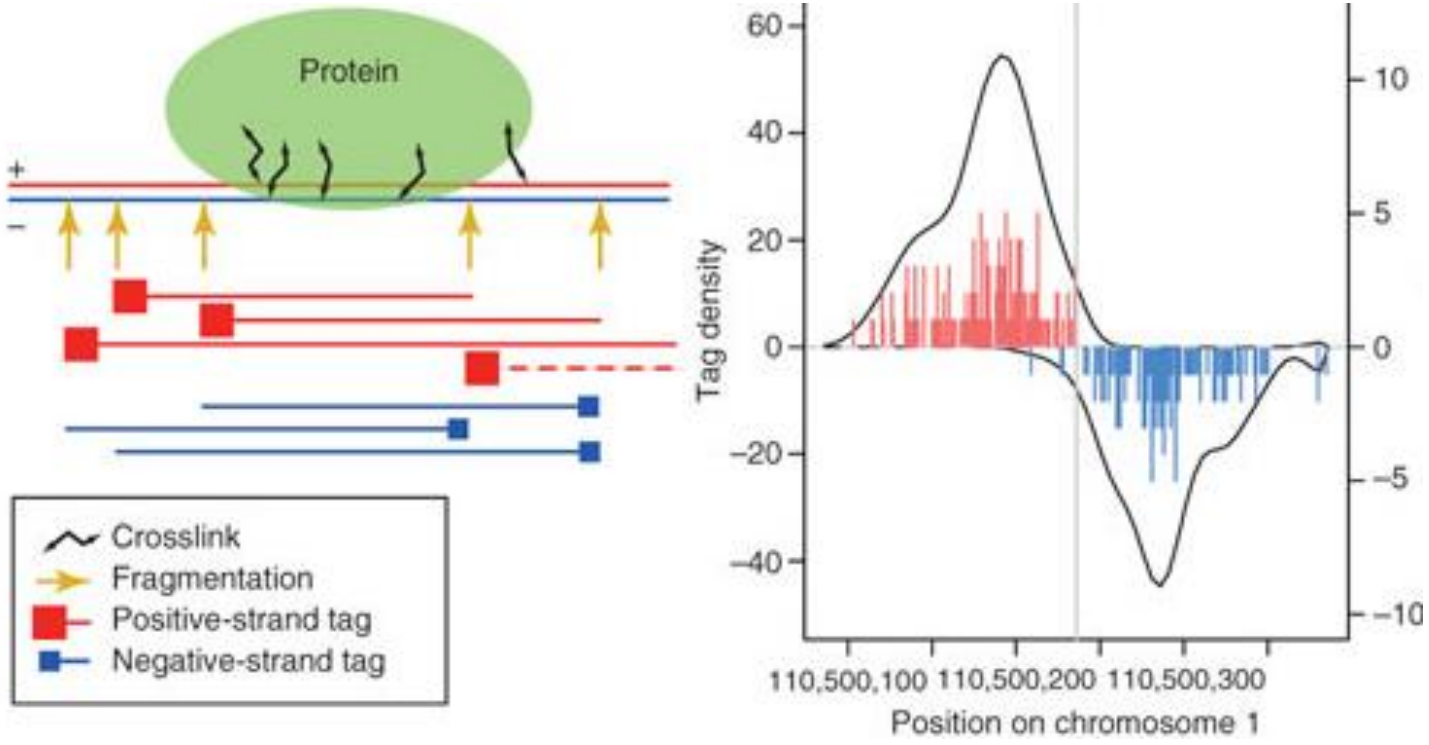
Bioinformatics Challenge– Detecting the Binding Regions

View mapped reads in a genome browser



Application note Nature Methods 6, (2009)

ChIP-Seq
Bimodal distribution



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

Bioinformatics Challenge- Detecting the Binding Regions

Open Access

Method

Model-based Analysis of ChIP-Seq (MACS)

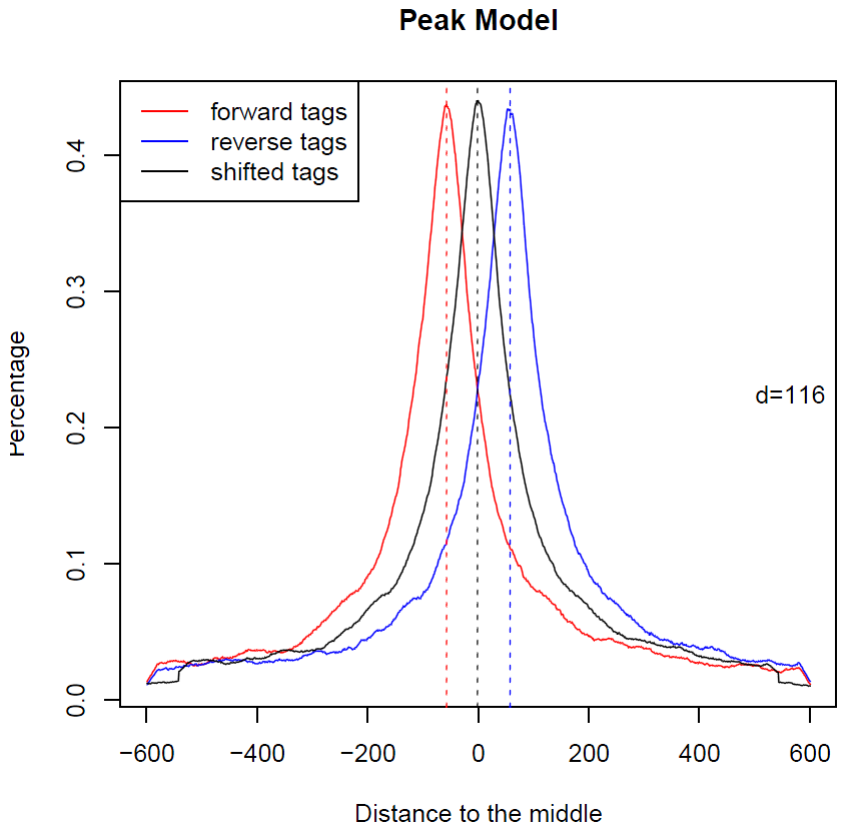
Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nussbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

Addresses: [*]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 44 Binney Street, Boston, MA 02115, USA. [†]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. [‡]Gene Security Network, Inc., 2686 Middlefield Road, Redwood City, CA 94063, USA. [§]Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital and Department of Pathology, Harvard Medical School, 13th Street, Charlestown, MA 02129, USA. [¶]Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA, 02142, USA. [¥]Department of Genetics, Stanford University Medical Center, Stanford, CA 94305, USA. [#]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.

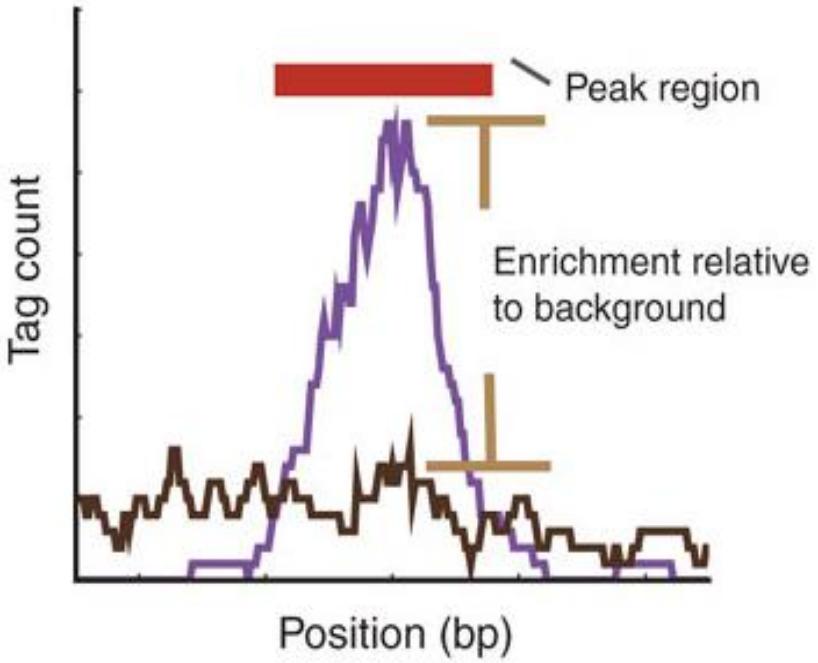
MACS

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



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

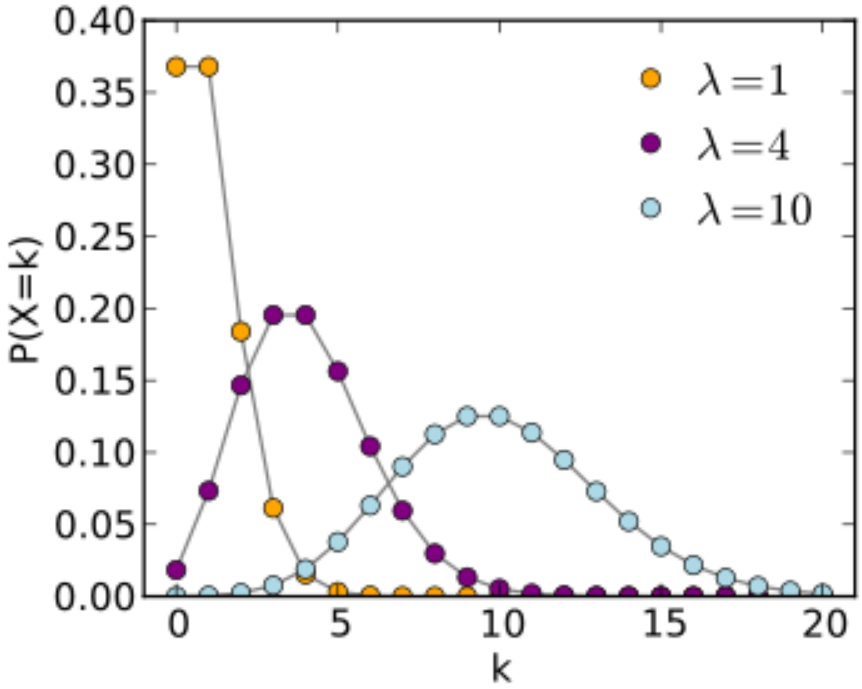
- MACS models the reads distribution along the genome by a Poisson distribution
  - We are counting reads in a fixed region
  - We can compute the expected (mean) number of events, i.e. number of reads in a specified region:

    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

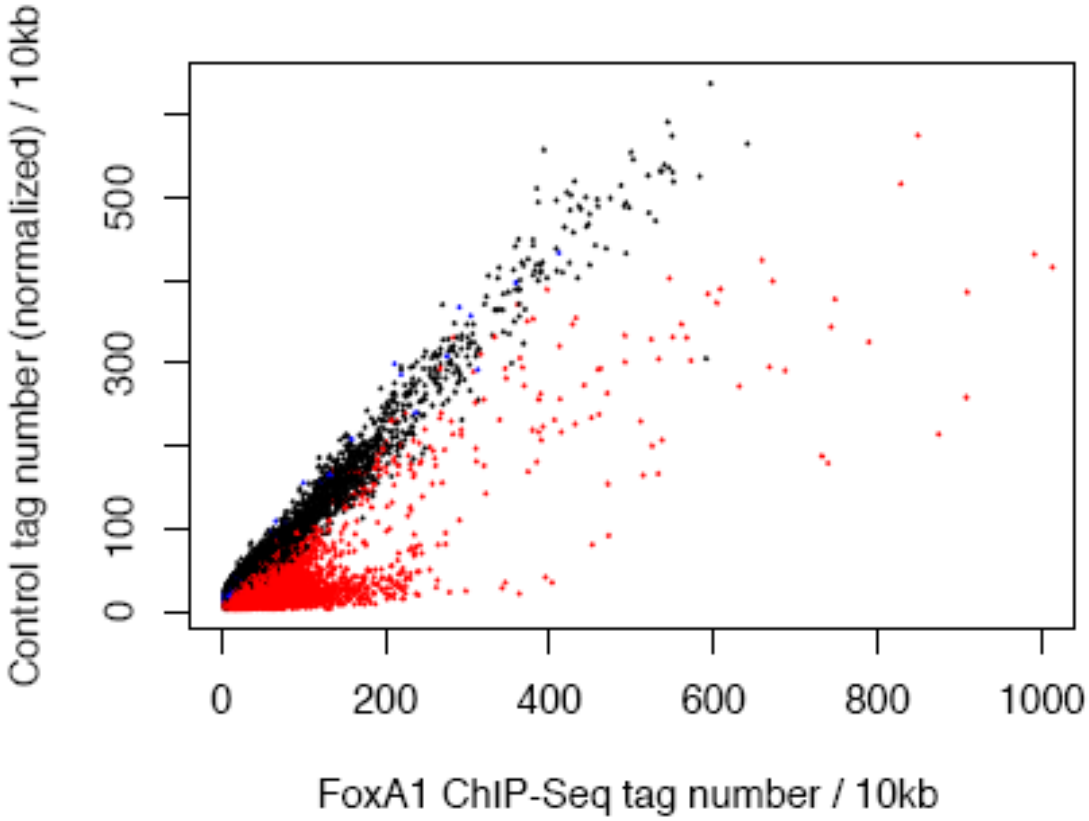


http://en.wikipedia.org/wiki/Poisson_distribution

Peak Detection

ChIP-Seq show local biases in the genome

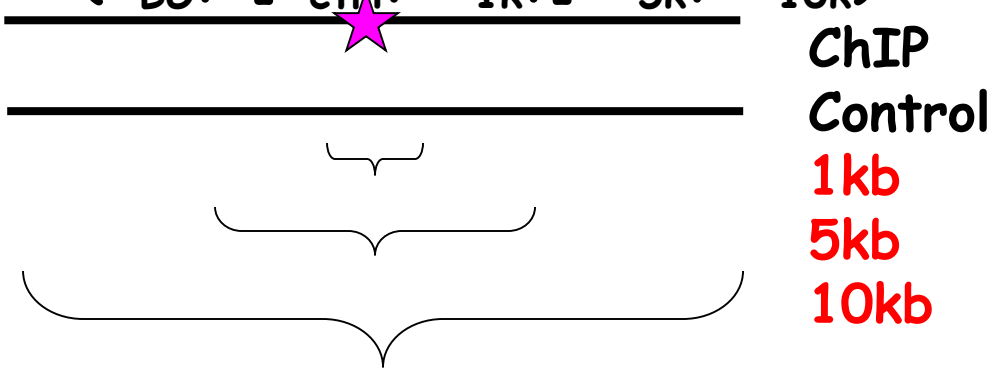
Chromatin and sequencing bias



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

MACS
FDR (False Discovery Rate)

FDR is calculated by:

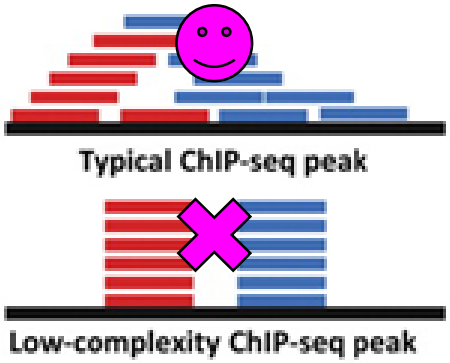
- **Swapping the control and treatment data**
- **Calculating p-values for these 'negative peaks'**
- **Calculating the FDR for a certain p-value**

$$\%FDR = \left(\text{Number of 'negative peaks'} / \text{Number of CHIP peaks}\right) * 100$$

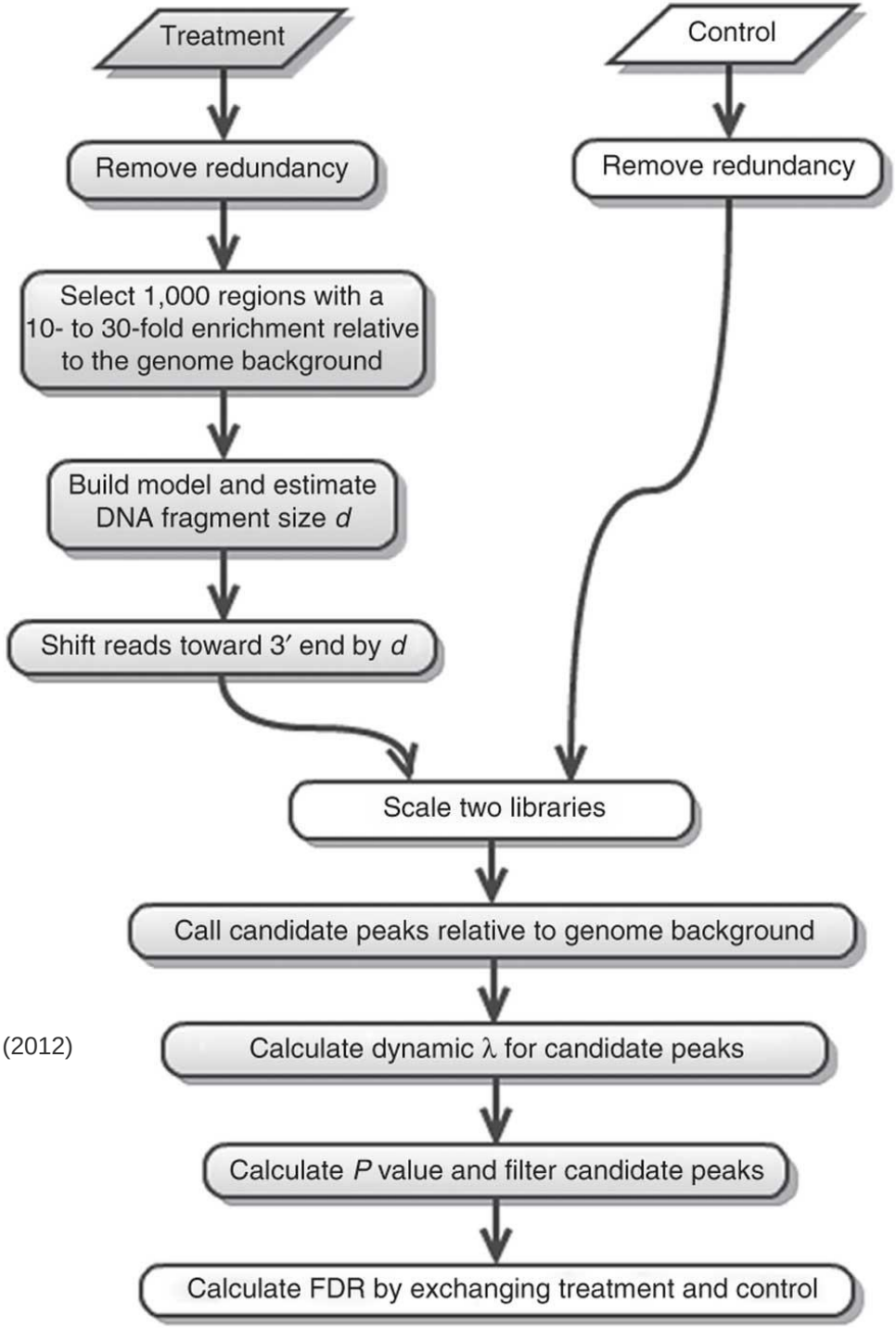
Example: 10 / 100 * 100 = 10% FDR

- Can we apply this method when comparing 2 biological conditions?
- Problem – unbalanced number of reads control-treatment

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- should have non-redundant frequency >=0.8



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

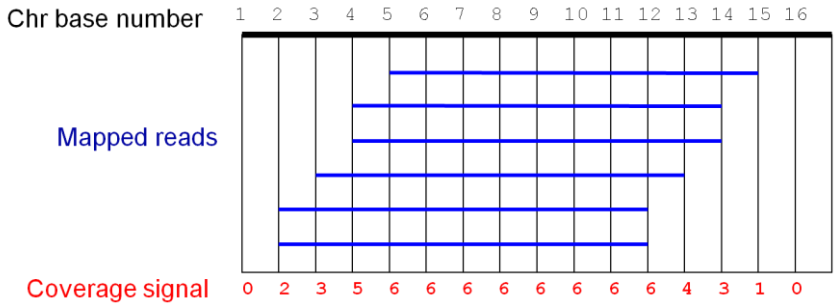
MACS Peak Information (.xls)

- **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



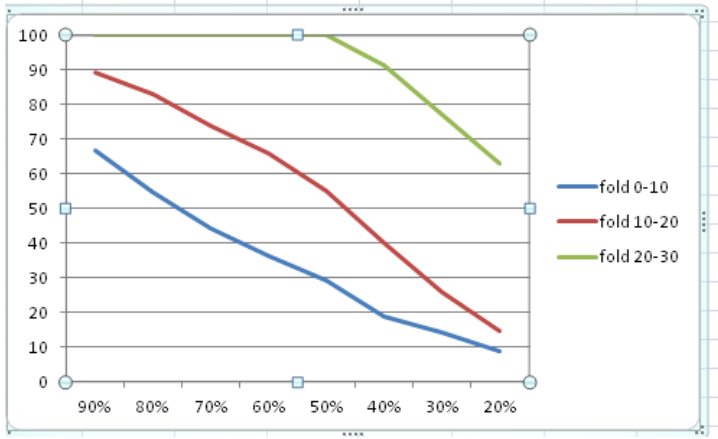
Wiggle format example

```
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
```

MACS Sampling for Sequencing Saturation

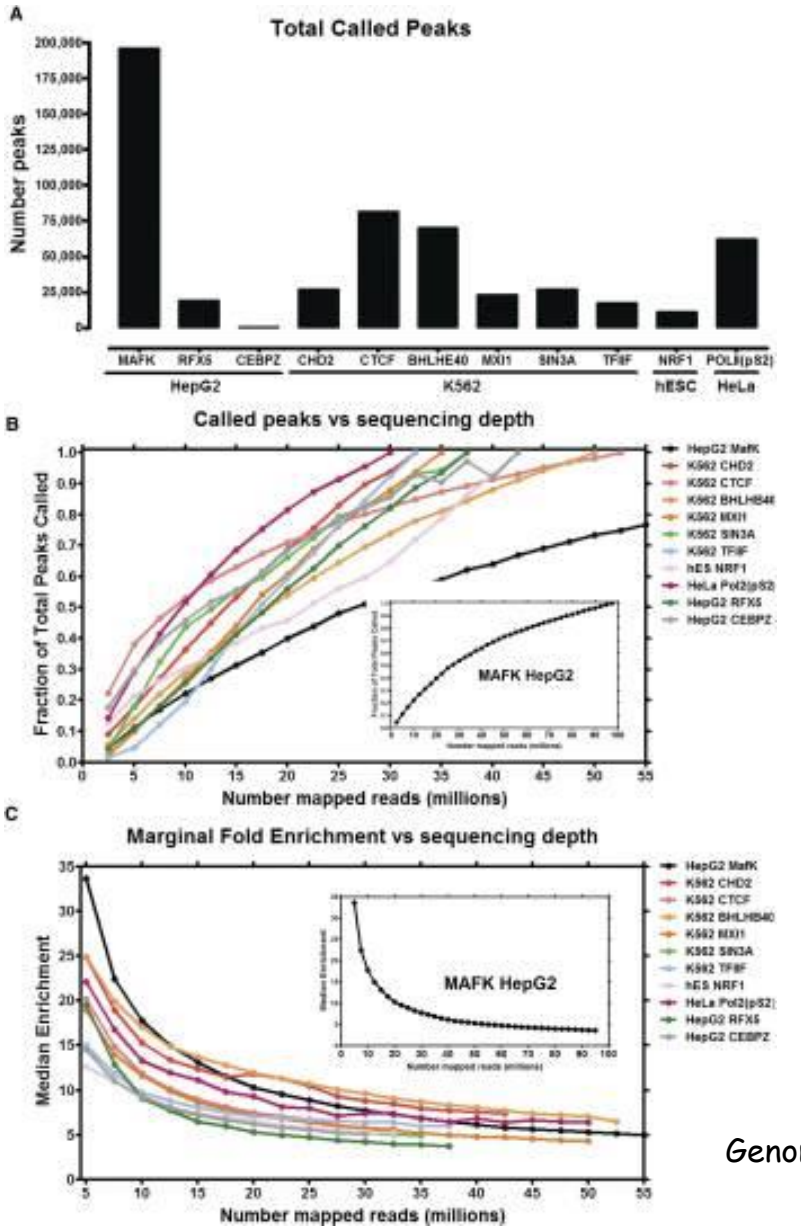
Macs produces a diagnosis report to test whether saturation in peak detection has been reached.

		cover by sampling							
FC range	# of Peaks	90%	80%	70%	60%	50%	40%	30%	20%
0-10	2502	66.75	54.64	44.12	36.33	29.26	18.9	14.19	8.87
10-20	1374	89.23	83.04	73.65	66.01	54.95	40.17	25.91	14.63
20-30	35	100	100	100	100	100	91.43	77.14	62.86
30-40	1	100	100	100	100	100	100	100	100



Standards: one should detect 99% of the peaks that show at least 2-fold enrichment over control with 90% of the data

Peak Counts Depend on Sequencing Depth



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

Biological Replicates ENCODE

- **Biological replicates are required for each dataset**
- **Criteria to decide that the biological replicates are in agreement-**
  - **The number of targets identified for each replicate cannot differ by more than a factor of 2**
  - **Irreproducible discovery rate (IDR)**

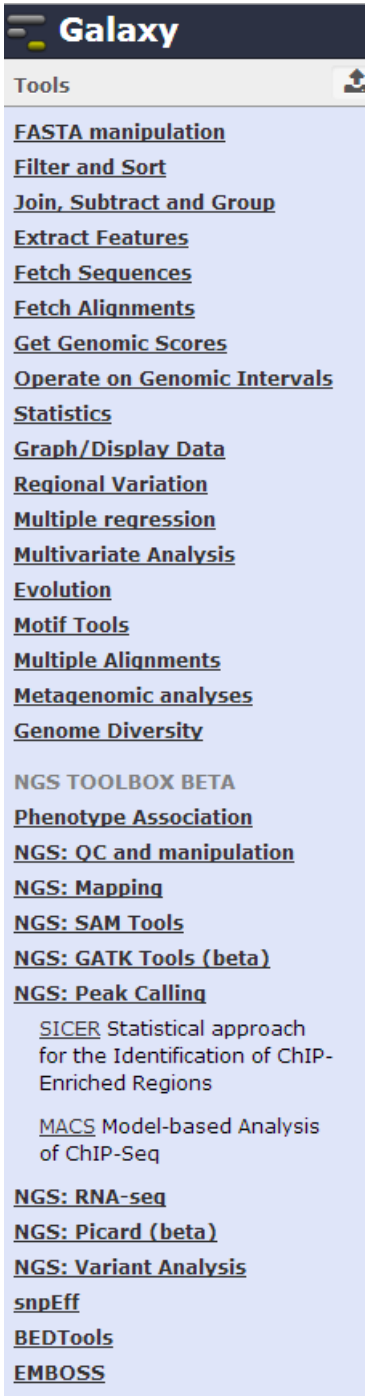
    For submission to ENCODE, we currently require that the number of bound regions identified in an IDR comparison between replicates to be at least 50% of the number of regions identified in an IDR comparison between two "pseudoreplicates" generated by pooling and then randomly partitioning all available reads from all replicates (Np/Nt < 2)
- **Reads from replicates which meet these criteria are usually combined and the data rescored.**
- **If requirement is not meet a third replicate is required.**

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.

  - Chipster



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.

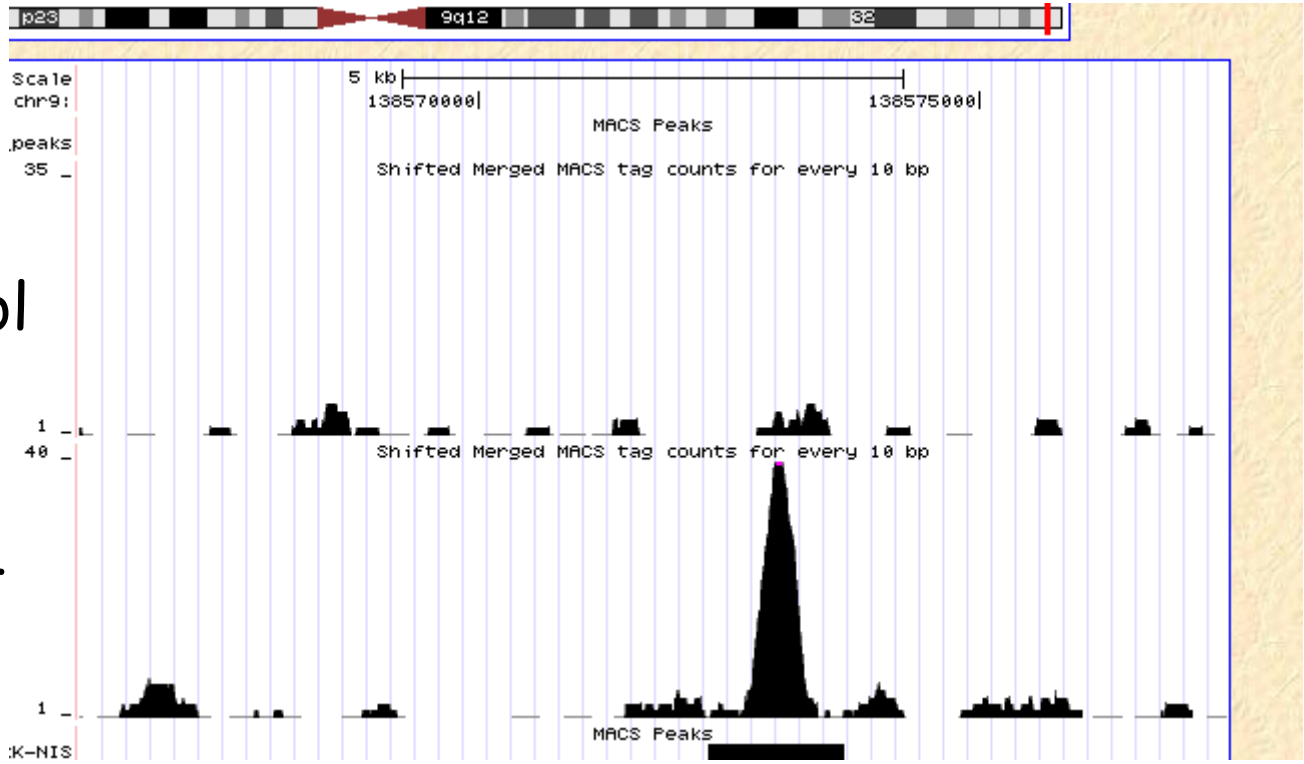
ChIP-Seq – Lecture Outline

- 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

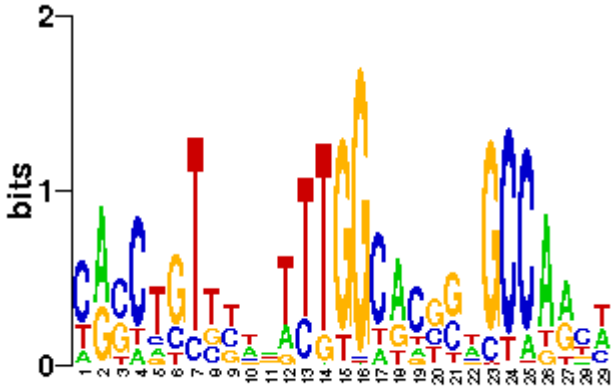
treat



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

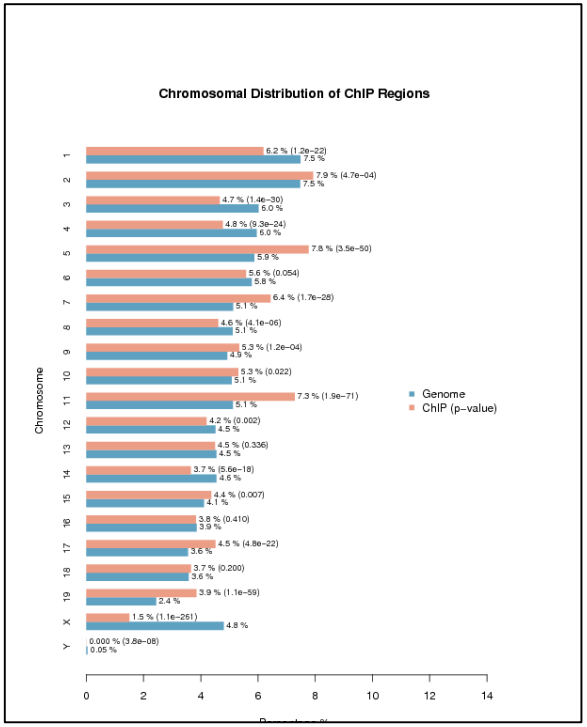


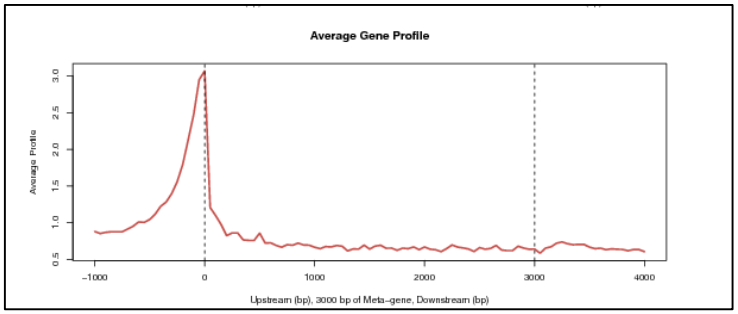
Recommended portals:

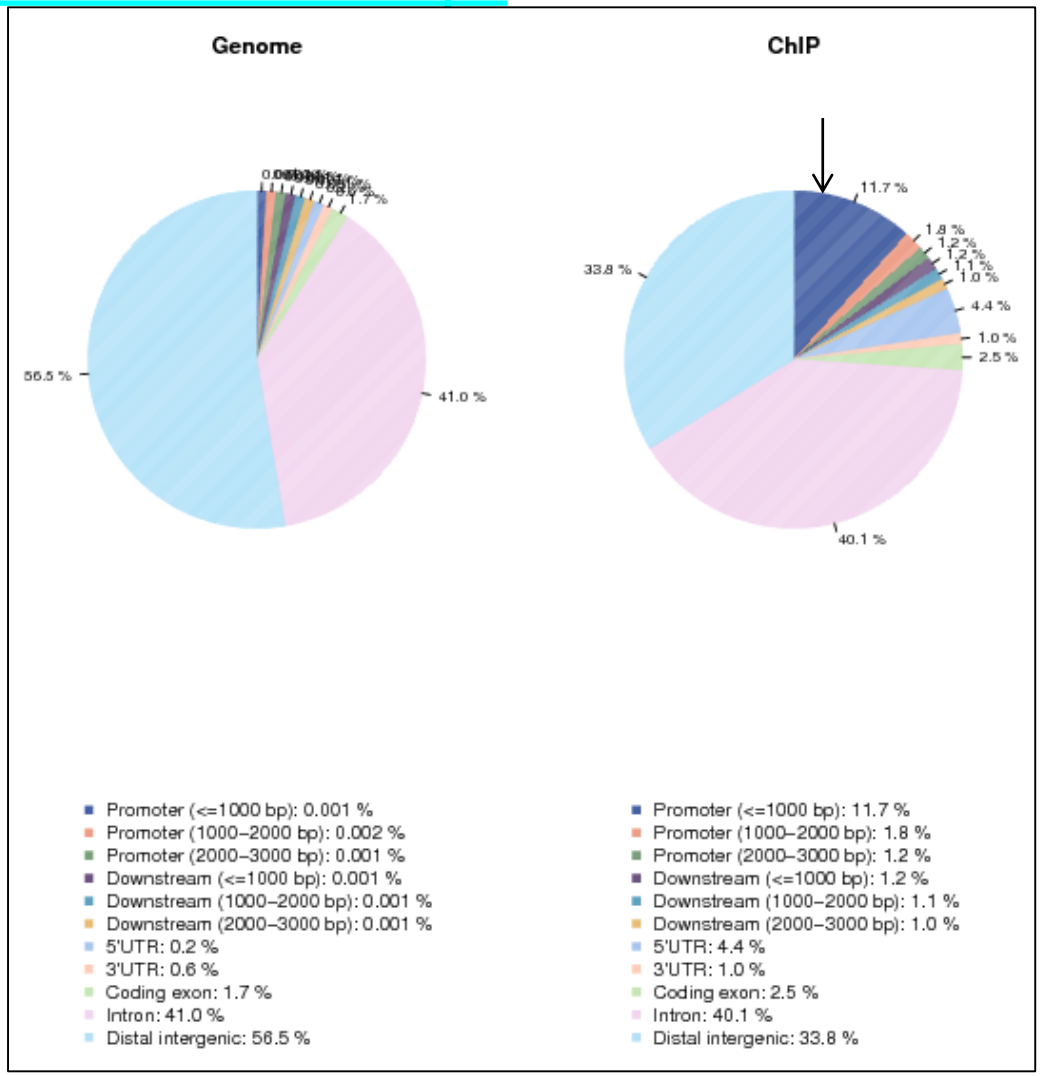
- MEME-ChIP (public)
- Homer (command line)
- Cistrome (public- SeqPos)
- Genomatix
- Chipster

CEAS: Enrichment of Genome Features

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

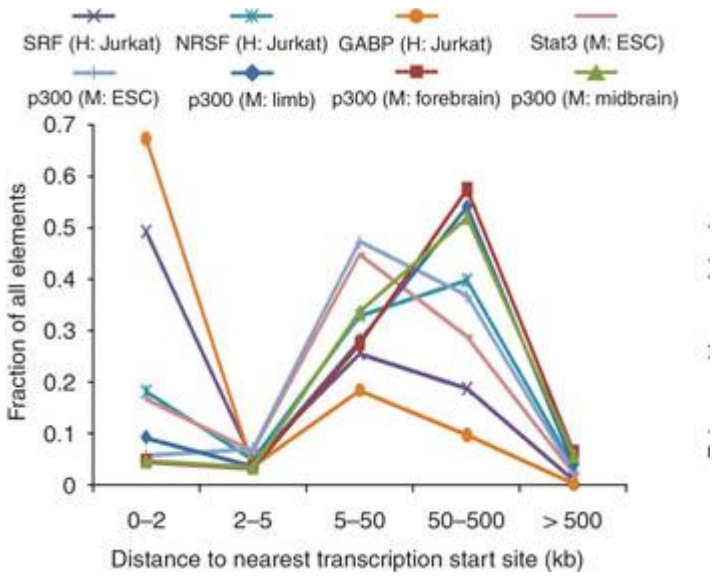






Functional Interpretation

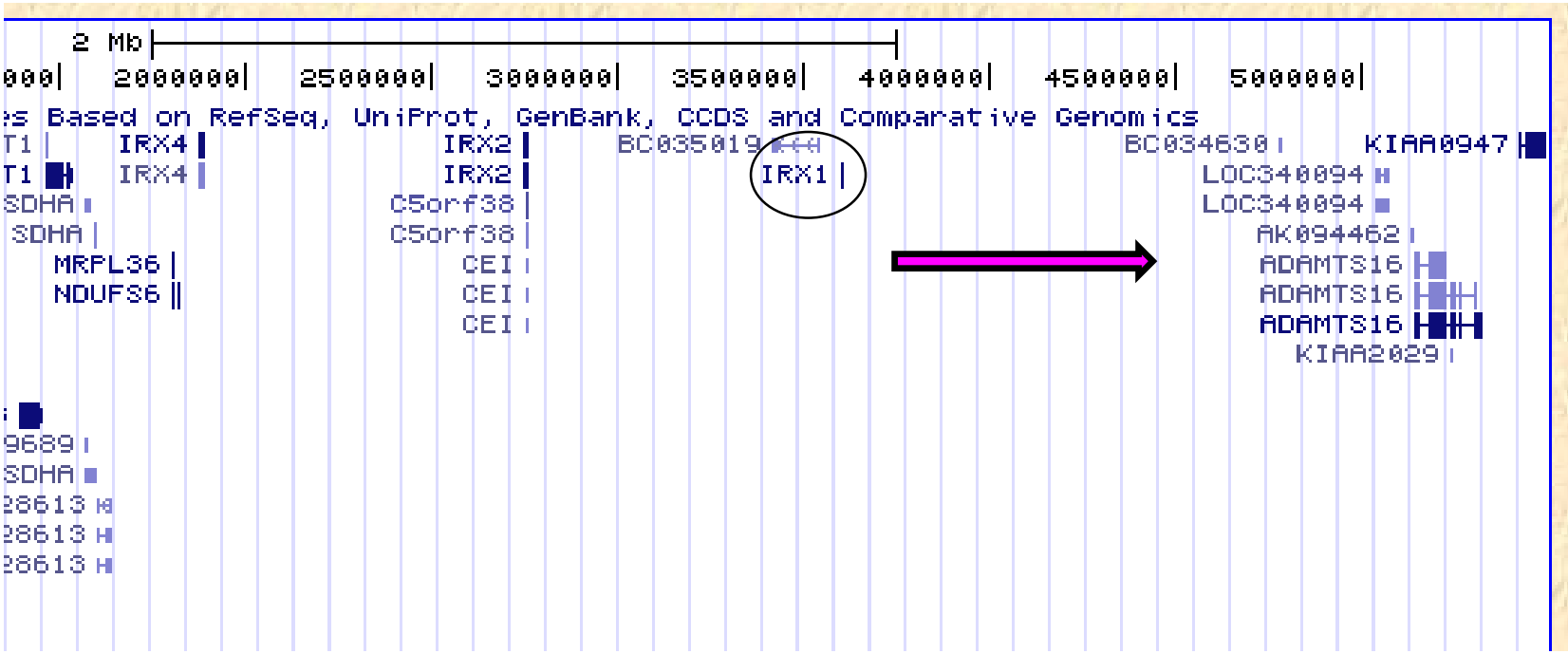
- **Which processes and functions does our TF regulate?**
  - **Associate peaks with genes**
  - **Associating only proximal genomic regions to genes (<5 kb) – for most TF ignores a large fraction of binding data**



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?

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**

GREAT ANALYSIS

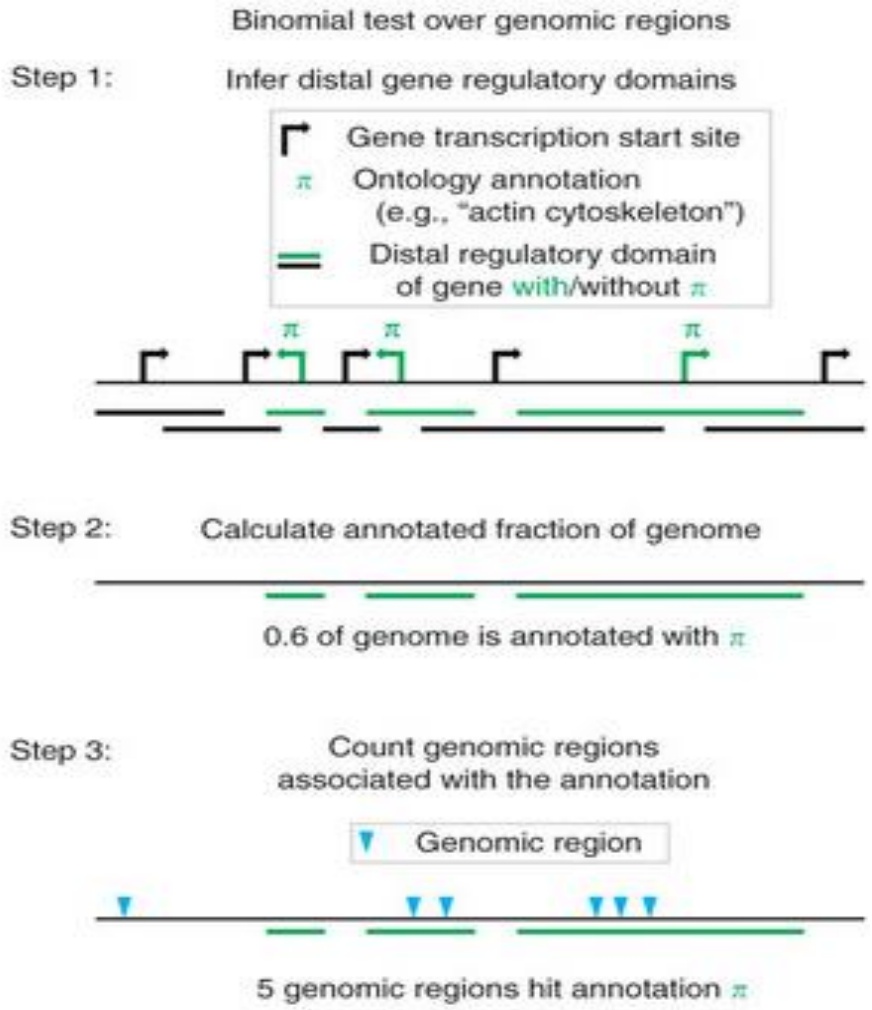
Mouse Phenotype (no terms)

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].

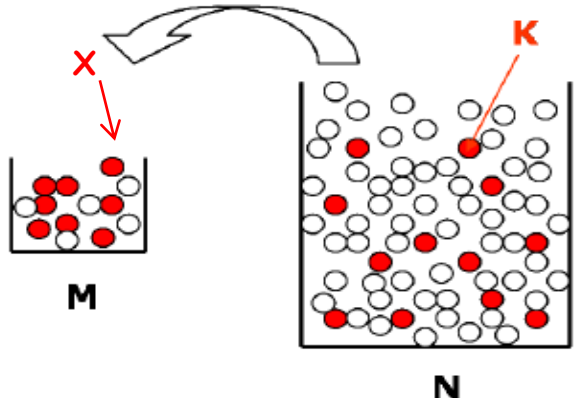


Nature Biotechnology 28 , 495–501 (2010)

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

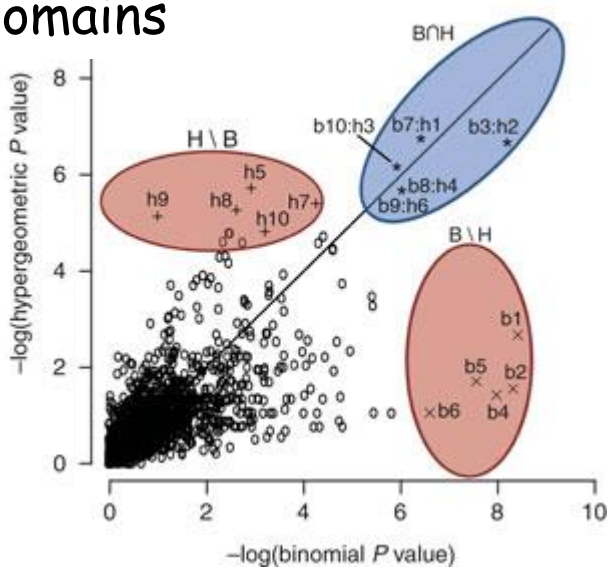
A gene based approach would expect 14% of the genes near peaks to be associated with 'multicellular organismal development'

Why do we have this discrepancy?

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?

Exercise

- Run MACS2 (command line)
- Compare peaks from two IP replicates
- Annotate the peaks using CEAS and GREAT
- View the mapped reads and peaks with a genome browser (IGV)


My Peak
4
2
0
0
2
4
6
8
THANKS