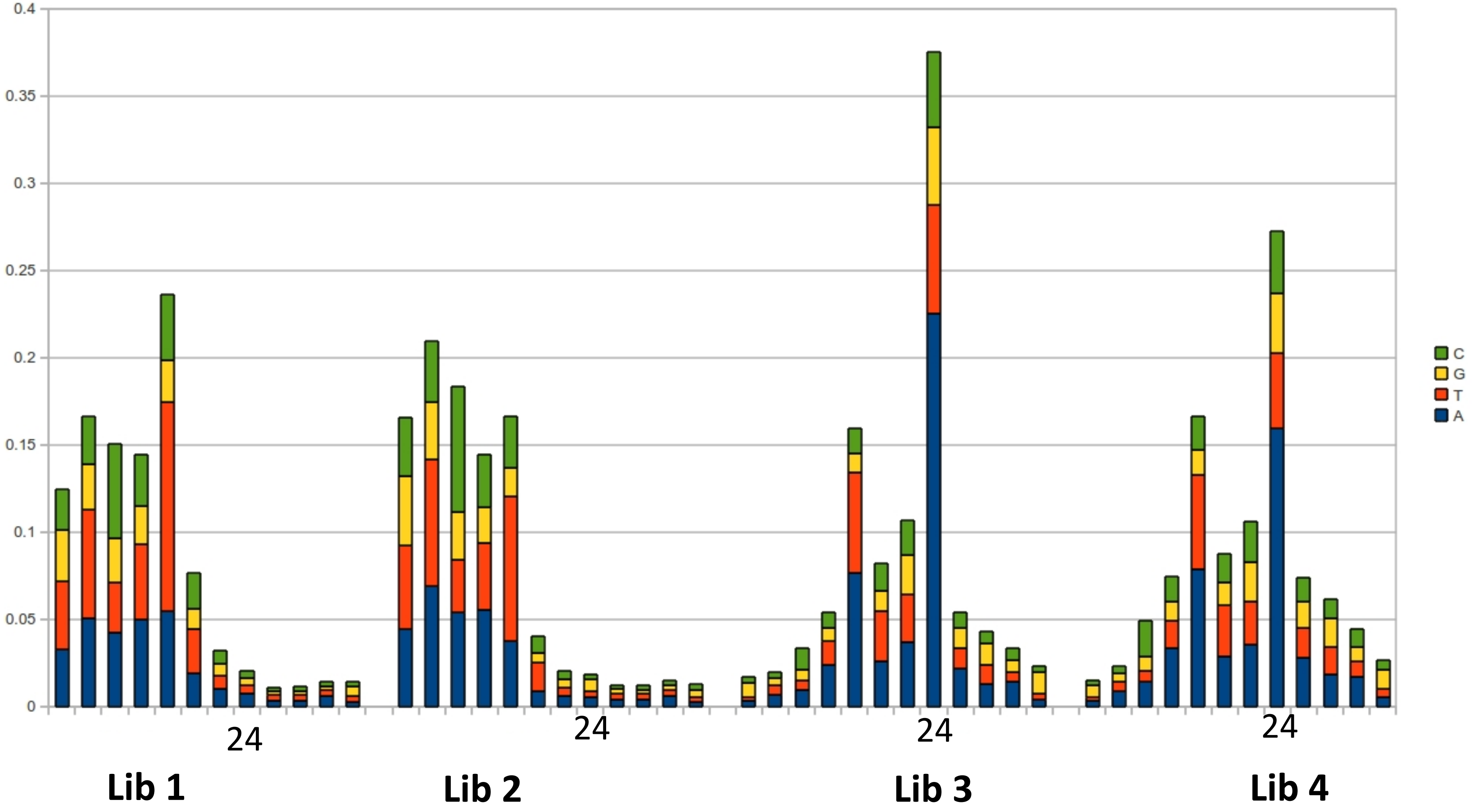
# Next-generation Sequencing

Lecture 11

# Applications of RNA-seq

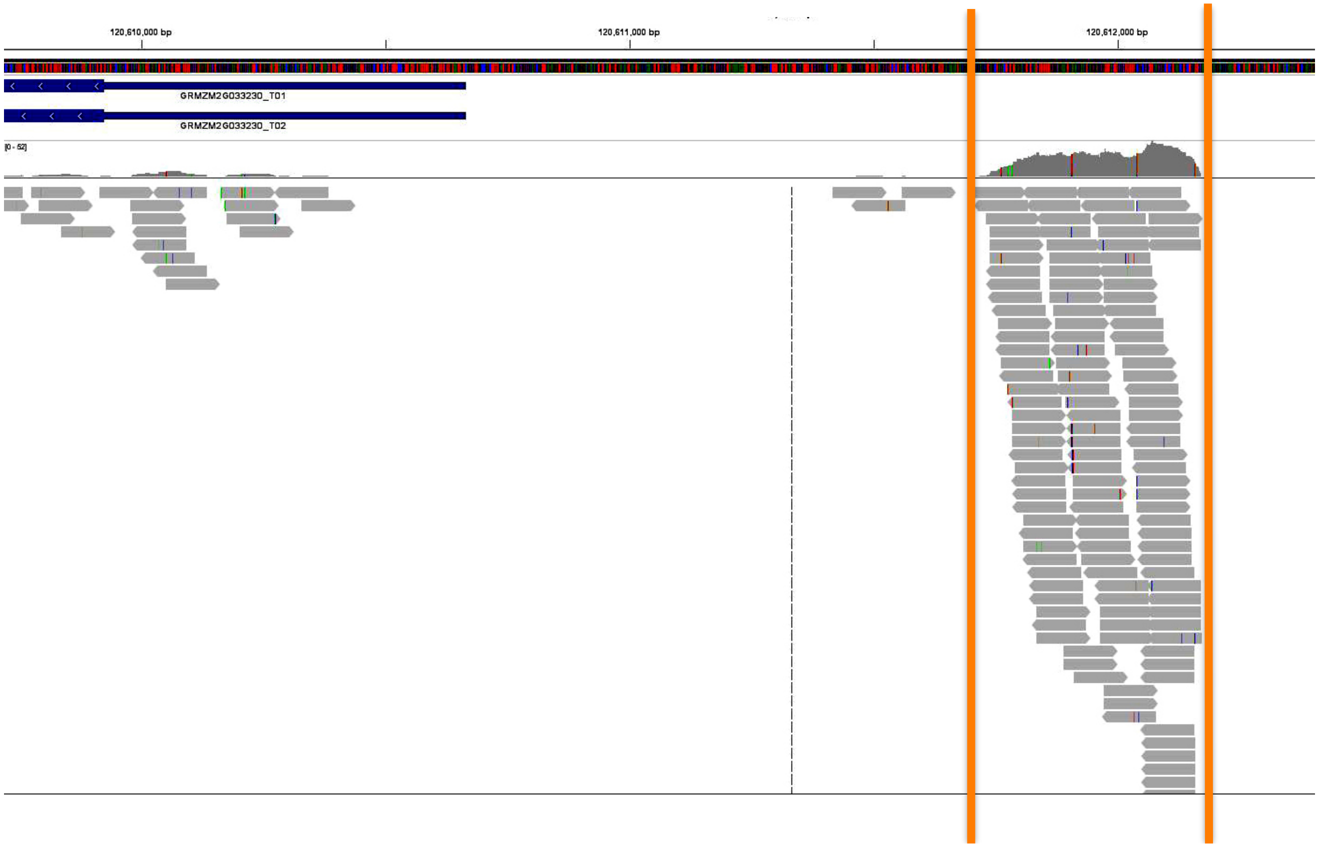
- Gene expression
  - Expression of individual genes/loci
  - Quantitatively discriminate isoforms using junction reads and coverage of individual exons, introns, etc.
- Annotation
  - New features of the transcriptome: genes, exons, splicing, ncRNAs
- SNP
- Fusion gene detection

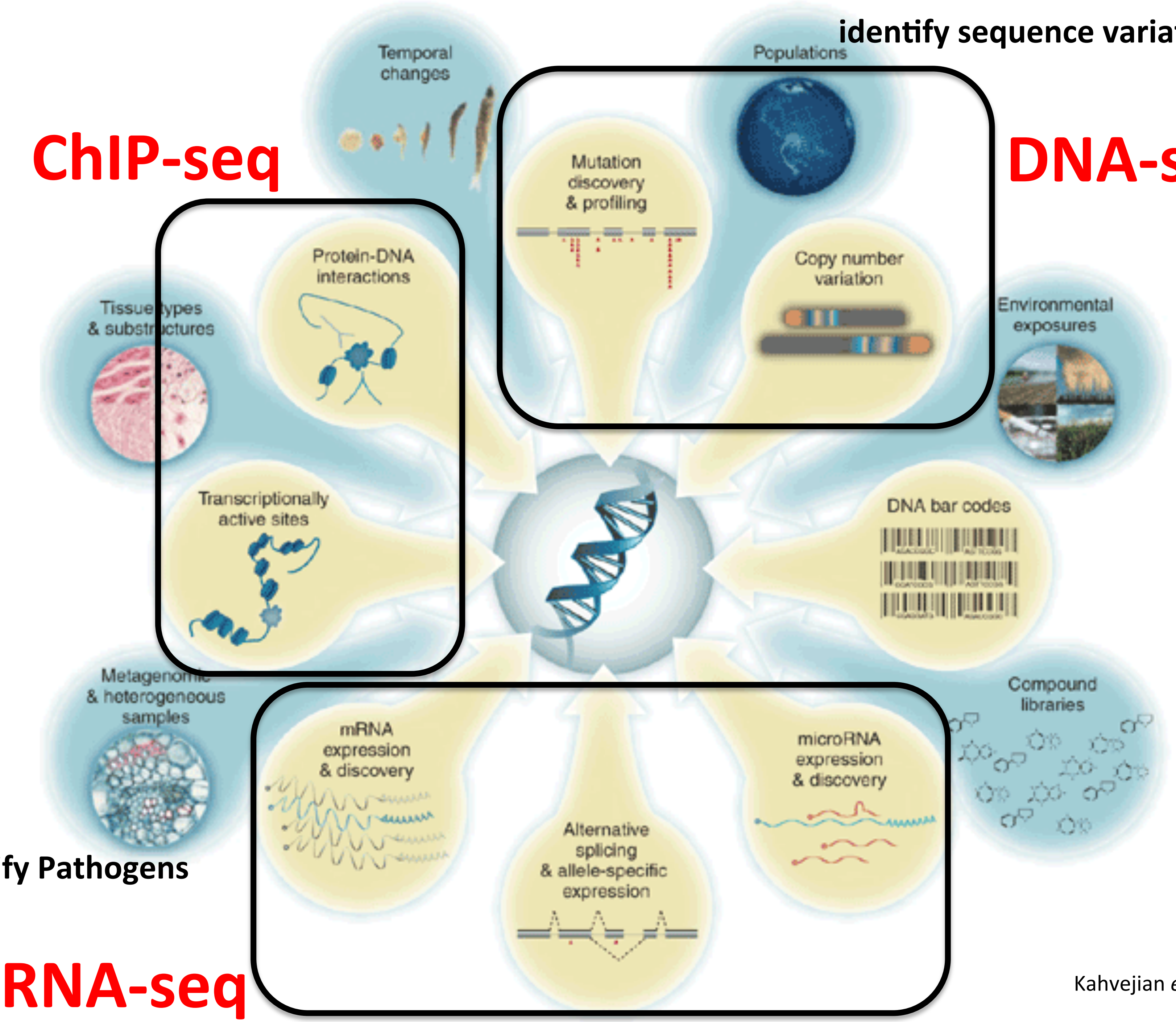
# Which small RNA libraries are good?



# lncRNA Candidate 1

- Chr 8
- Near GRMZM2G033230
- Length: 473
- Reads number: 114




Temporal changes
Populations
Mutation discovery & profiling
Copy number variation
Protein-DNA interactions
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
Compound libraries
mRNA expression & discovery
microRNA expression & discovery
Alternative splicing & allele-specific expression

identify sequence variations

**ChIP-seq**

**DNA-seq**

Identify Pathogens

**RNA-seq**

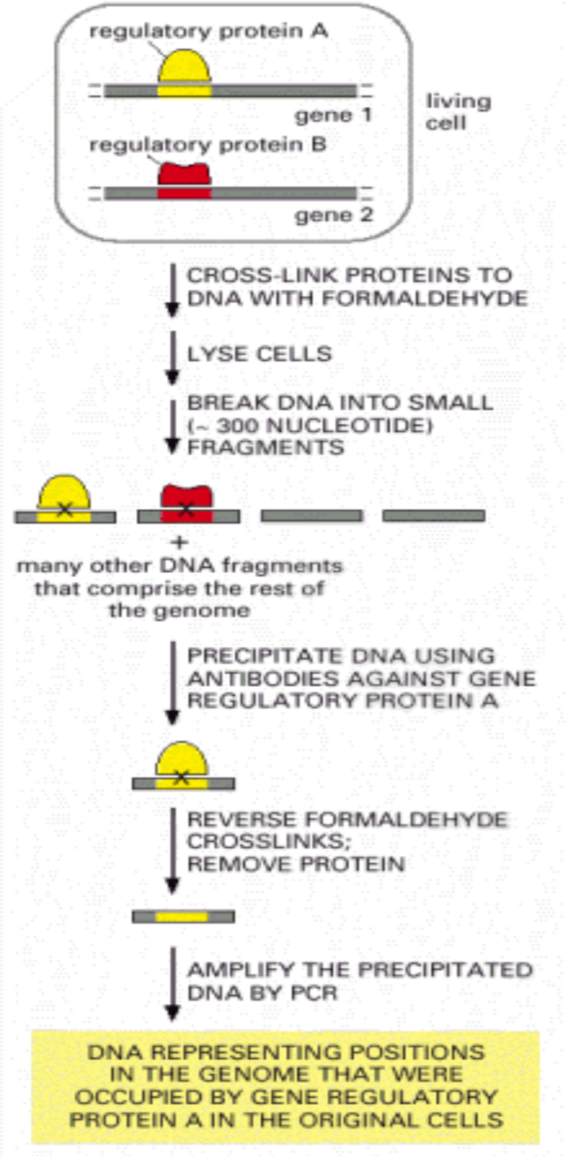
Kahvejian *et al*, 2008

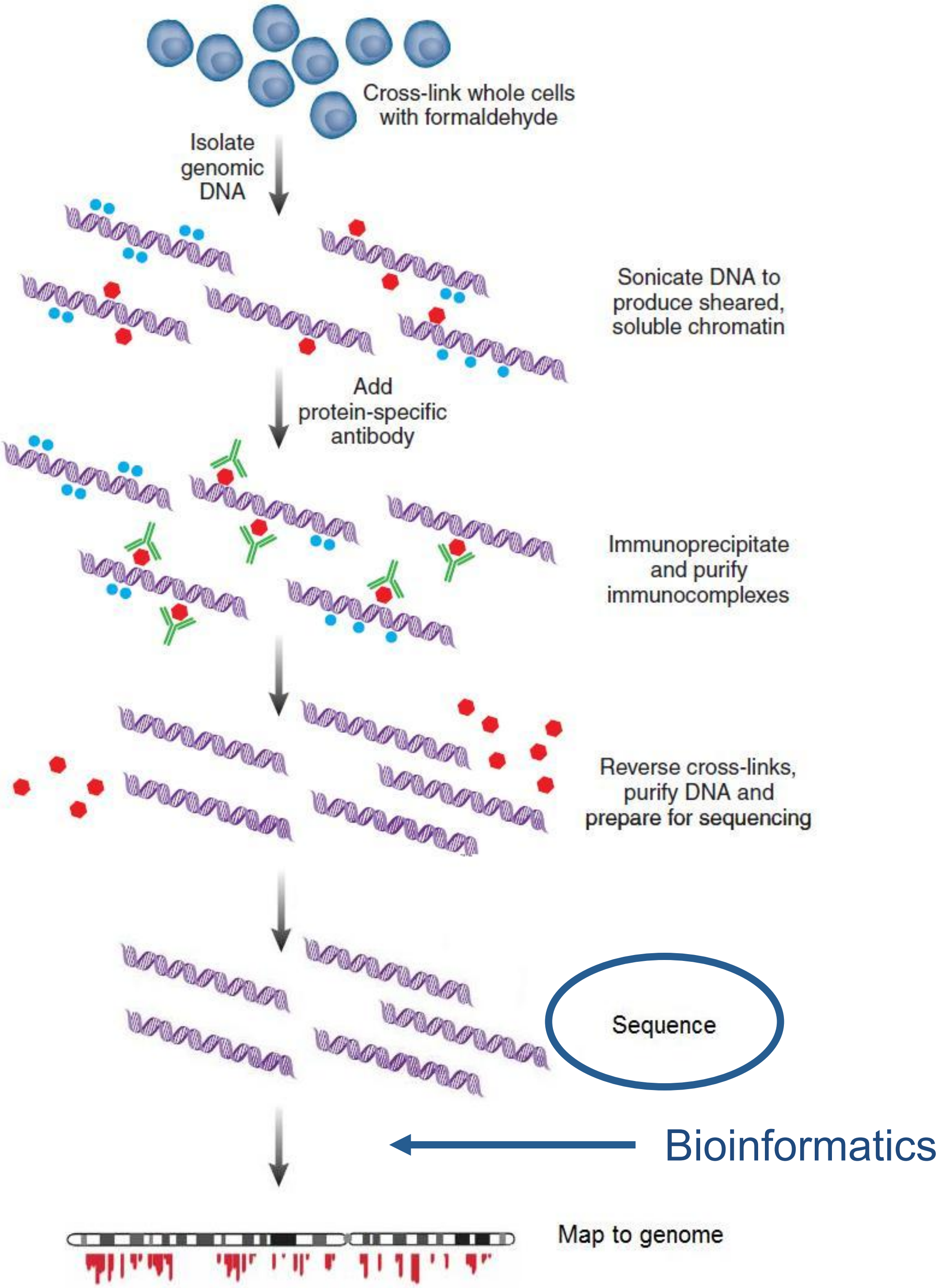
# Protein-DNA interaction

- DNA is the information carrier of almost all living organisms.
- Protein is the major building block of life.
- Interaction between DNA and protein play vital roles in the development and normal function of living organisms, and disease if something goes wrong.
- An important mechanism of protein-DNA interaction is via direct binding, i.e., a protein binds to a particular fragment of the DNA.

# Chromatin Immunoprecipitation (ChIP)

- ChIP is a method to investigate protein-DNA interaction in vivo.
- In ChIP, antibodies are used to select specific proteins or nucleosomes, which enrich for DNA fragments that are bound to these proteins or nucleosomes.
- The output of ChIP is enriched fragments of DNA that were bound by a particular protein.
- The identity of DNA fragments need to be further determined by a second method.


regulatory protein A
gene 1
regulatory protein B
gene 2
living cell
CROSS-LINK PROTEINS TO DNA WITH FORMALDEHYDE
LYSE CELLS
BREAK DNA INTO SMALL (~ 300 NUCLEOTIDE) FRAGMENTS
+
many other DNA fragments that comprise the rest of the genome
PRECIPITATE DNA USING ANTIBODIES AGAINST GENE REGULATORY PROTEIN A
REVERSE FORMALDEHYDE CROSSLINKS; REMOVE PROTEIN
AMPLIFY THE PRECIPITATED DNA BY PCR
DNA REPRESENTING POSITIONS IN THE GENOME THAT WERE OCCUPIED BY GENE REGULATORY PROTEIN A IN THE ORIGINAL CELLS



Cross-link whole cells with formaldehyde

Isolate genomic DNA

Sonicate DNA to produce sheared, soluble chromatin

Add protein-specific antibody

Immunoprecipitate and purify immunocomplexes

Reverse cross-links, purify DNA and prepare for sequencing

Sequence

Bioinformatics

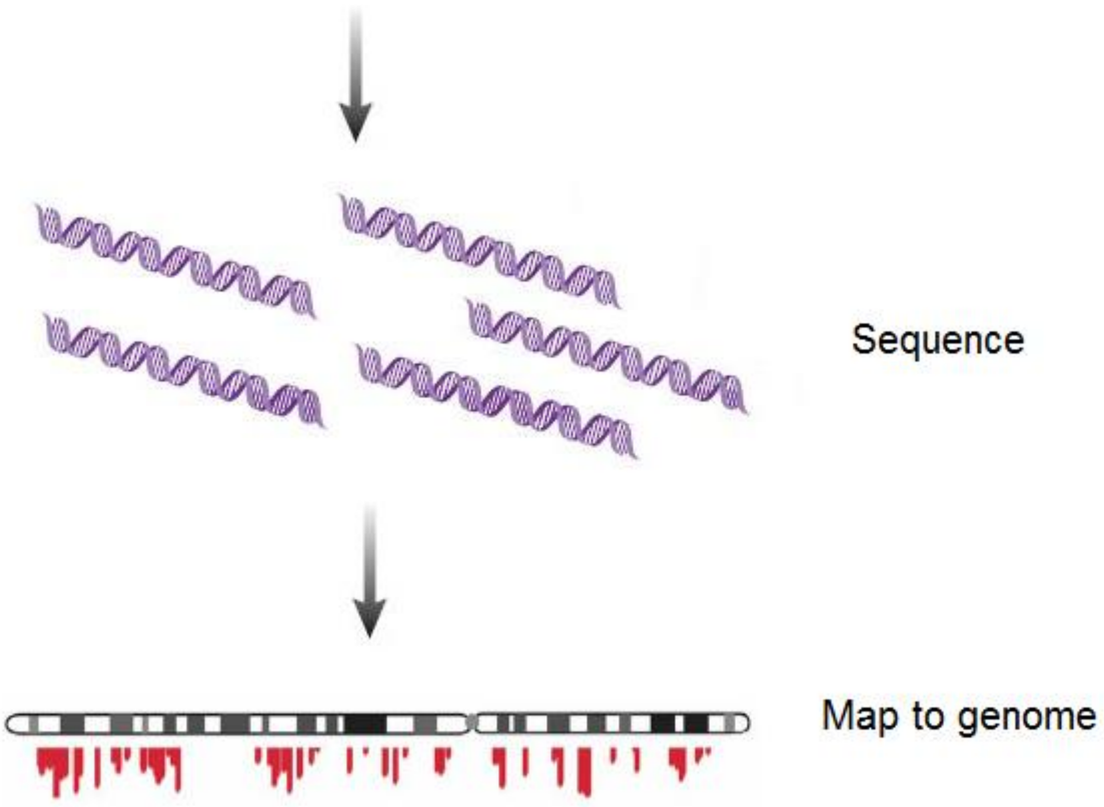
Map to genome

# ChIP-seq

Although the short reads (~35bp) generated by NGS platforms pose serious difficulties for certain applications - for example, de novo genome assembly - they are aceptable for ChIP-seq.

The more precise mapping of protein-binding sites provided by ChIP-seq allows for a more accurate list of targets for transcription factors and enhancers, in addition to better identification of sequence motifs.

Nat Rev Genet. 2009 Oct;10(10):669-80.



- The idea is that if a segment of DNA contains a protein binding site, this sequence will appear more often in the precipitated fraction.

# ChIP-seq v.s. ChIP-chip

- ChIP-seq has higher resolution, fewer artifacts, greater coverage and a larger dynamic range than ChIP-chip, and therefore provides substantially improved data.
- In ChIP-seq, the DNA fragments of interest are sequenced directly instead of being hybridized on an array.
- The main disadvantage with ChIP-seq is its current cost and availability. The overall cost of ChIP-seq, which includes machine depreciation and reagent cost, will have to be lowered further for it to be comparable with the cost of ChIP-chip.
- For high-resolution profiling of an entire large genome, ChIP-seq is already less expensive than ChIP-chip.
- However, as the cost of sequencing continues to decline and institutional support for sequencing platforms continues to grow, ChIP-seq is likely to become the method of choice for nearly all ChIP experiments in the near future.

# Drawbacks of ChIP-seq

- All profiling technologies produce unwanted artifacts, and ChIP-seq is no exception. Although sequencing errors have been reduced substantially as the technology has improved, they are still present, especially towards the end of each read.
- There is also bias towards GC-rich content in fragment selection, both in library preparation and in amplification before and during sequencing, although notable improvements have been made recently.
- In addition, when an insufficient number of reads is generated, there is a loss of sensitivity or specificity in detection of enriched regions.
- There are also technical issues in performing the experiment, such as loading the correct amount of sample: too little sample will result in too few tags.

Nat Rev Genet. 2009 Oct;10(10):669-80.

# What does ChIP-seq can do?

- Chromatin-immunoprecipitation followed by sequencing is a powerful tool
- Epigenetics:
  - histone modifications
  - DNA methylation (different from bisulfite-seq)
- Locating transcription factor (TF) DNA interactions
- Detecting what nucleic acid sequences any protein is interacting with
  - ribosomal profiling

# Work flow of ChIP-seq

- Experimental design and sample preparation
- Sequencing
- Data analysis
  - Data preprocessing
  - Short reads mapping
  - Peak analysis
  - Post-processing: annotation

# Sample preparation

(1)The DNA-binding protein is crosslinked to DNA *in vivo* by treating cells with formaldehyde.

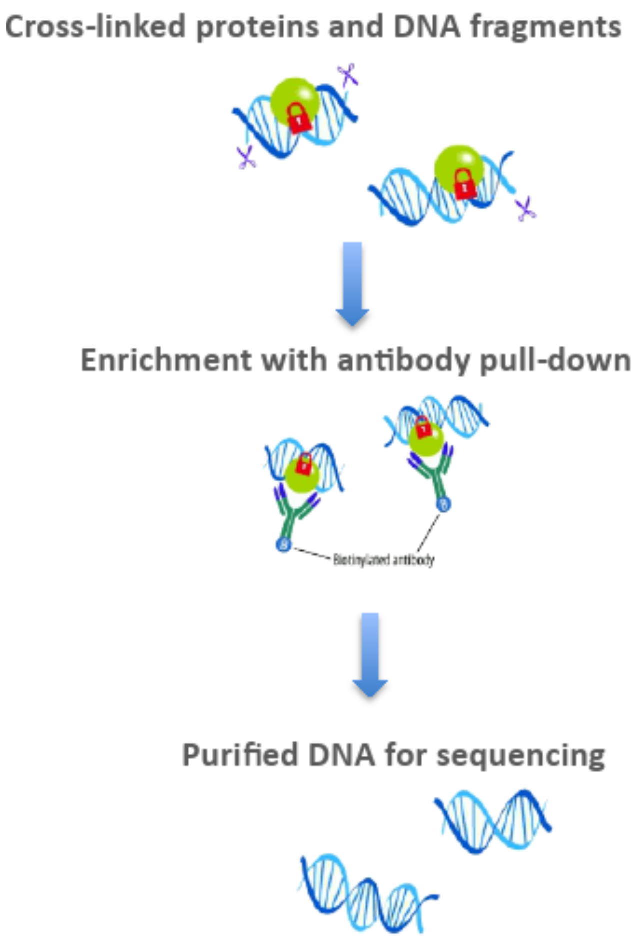
(2) the chromatin is sheared by sonication into small fragments.

(3) Introduce tagged antibody that targets the protein of interest， which is used to immunoprecipitate the DNA-protein complex.

(4) The crosslinks are reversed.

(5) Purification of DNA.

During the construction of a sequencing library, the immunoprecipitated DNA is subjected to size selection (typically in the ~150-300bp range, although there seems to be a bias towards shorter fragments in sequencing).


Cross-linked proteins and DNA fragments
Enrichment with antibody pull-down
Biotinylated antibody
Purified DNA for sequencing

# Issues for library construction

- Libraries may be constructed from ChIP DNA by standard protocols specific to the sequencing platform. Typically, library construction includes end repair, the addition of single adenosine residues, adaptor ligation, size selection and gel purification, followed by PCR with primers specific to the sequencing platform.
- During the size-selection step, it is important that the agarose gel be melted at room temperature (~22 ºC) rather than at 50 ºC, as the latter temperature might result in a bias for guanosine and cytidine because of loss of sequences rich in adenosine and thymidine.
- During the PCR amplification step, it is important that adaptor-ligated DNA products are not over-amplified, which may result in a loss of specific signal, bias or redundancy in the number of sequence tags.
- Over-amplification can typically be avoided by decreasing the number of PCR cycles or decreasing the amount of template DNA used for PCR.

Nat Immunol. 2011 Sep 20;12(10):918-22

# Antibody issues

- There are often multiple antibodies for a particular protein
    - For P53, there are two widely used ones
- The antibody might not be specific.
- Might detect direct and indirect interactions with DNA
- Cross-linking may occur for spatially proximal proteins that are bound to DNA very far apart in the sequence.

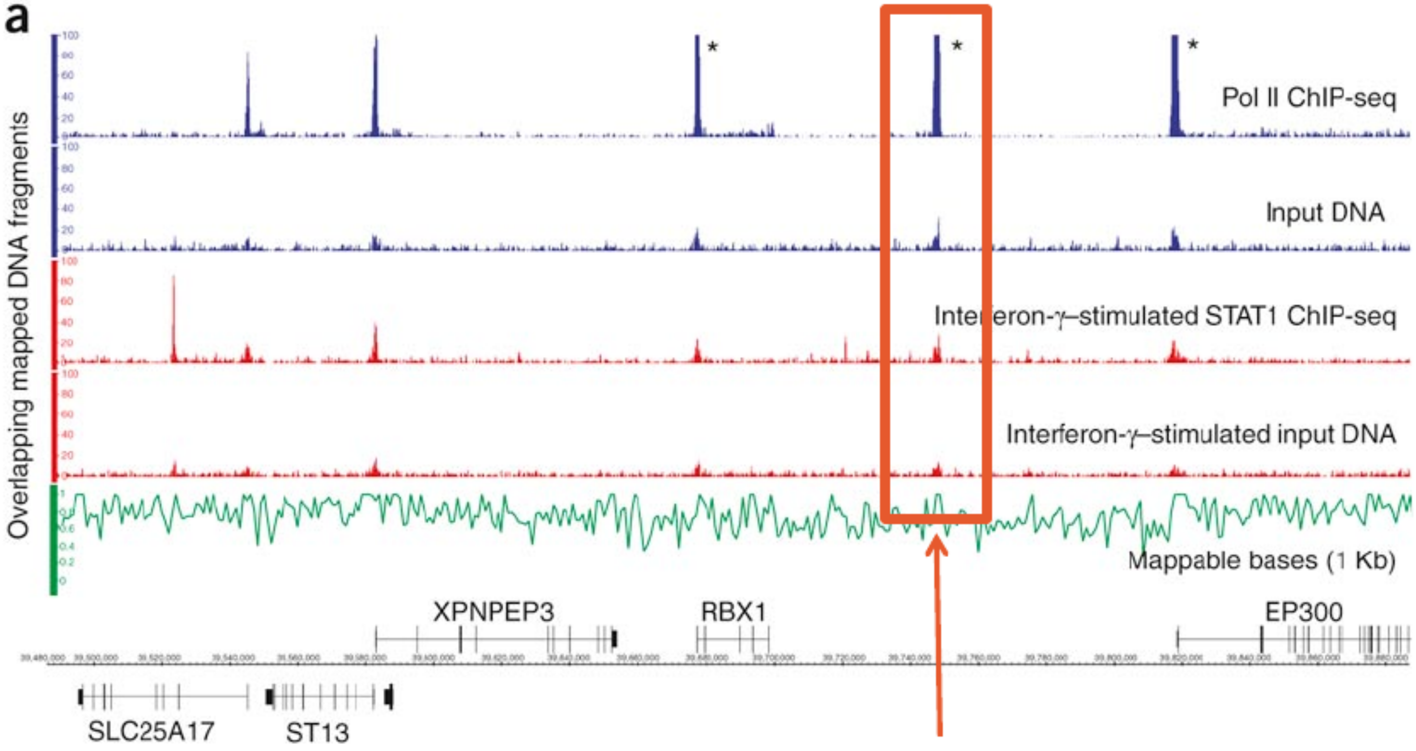
# Controls

- It is important that relevant controls are used
- It is, however, not so clear what those should be, and at what level they are useful.
- Commonly used controls:

  Input DNA (randomly sheared DNA)

  - Unspecific antibodies (IgG, antibody to some other proteins, antibody from other species, etc)
  - Some other proteins (GFP, etc)
- Used to identify anomalies in the genome or artifacts that might be due to reagents, not biology.

# The need for controls



ChIP

Input

ChIP

Input

Peaks line up

Rozowsky et al., 2009

# Replicates

- Many factors, including cell-culture conditions, ChIP and library construction, may contribute to variability between data sets.
- To ensure reliability of the data, biological replicate experiments are necessary.
- Although there is no consensus on the correct number of replicates needed, at least duplicate biological experiments should be done.
- Although only one ChIP-grade antibody is available for the analysis of most histone modifications and transcription factors, it is recommended that ChIP-seq data be confirmed through the use of a different antibody wherever possible, to control for a potential antibody cross-reactivity.

Nat Immunol. 2011 Sep 20;12(10):918-22

# Work flow of ChIP-seq

- Experimental design and sample preparation
- Sequencing
- Data analysis
  - Data preprocessing
  - Short reads mapping
  - Peak analysis

# Bioinformatics Challenges

- Rapid mapping of these short sequence reads to the reference genome
- Visualize mapping results
  - Thousand of enriched regions
- Peak analysis
  - Peak detection
  - Finding exact binding sites
- Compare results of different experiments
  - Normalization
  - Statistical tests

# Analysis

- Quality controls
- Map to the genome
    - Does it like repetitive DNA?
- Determine fragment length
- Determine signal/background
- Deal with controls (if present)
- Decide if we are looking for peaks or sausages?
- If transcription factor, do we know the binding motif?

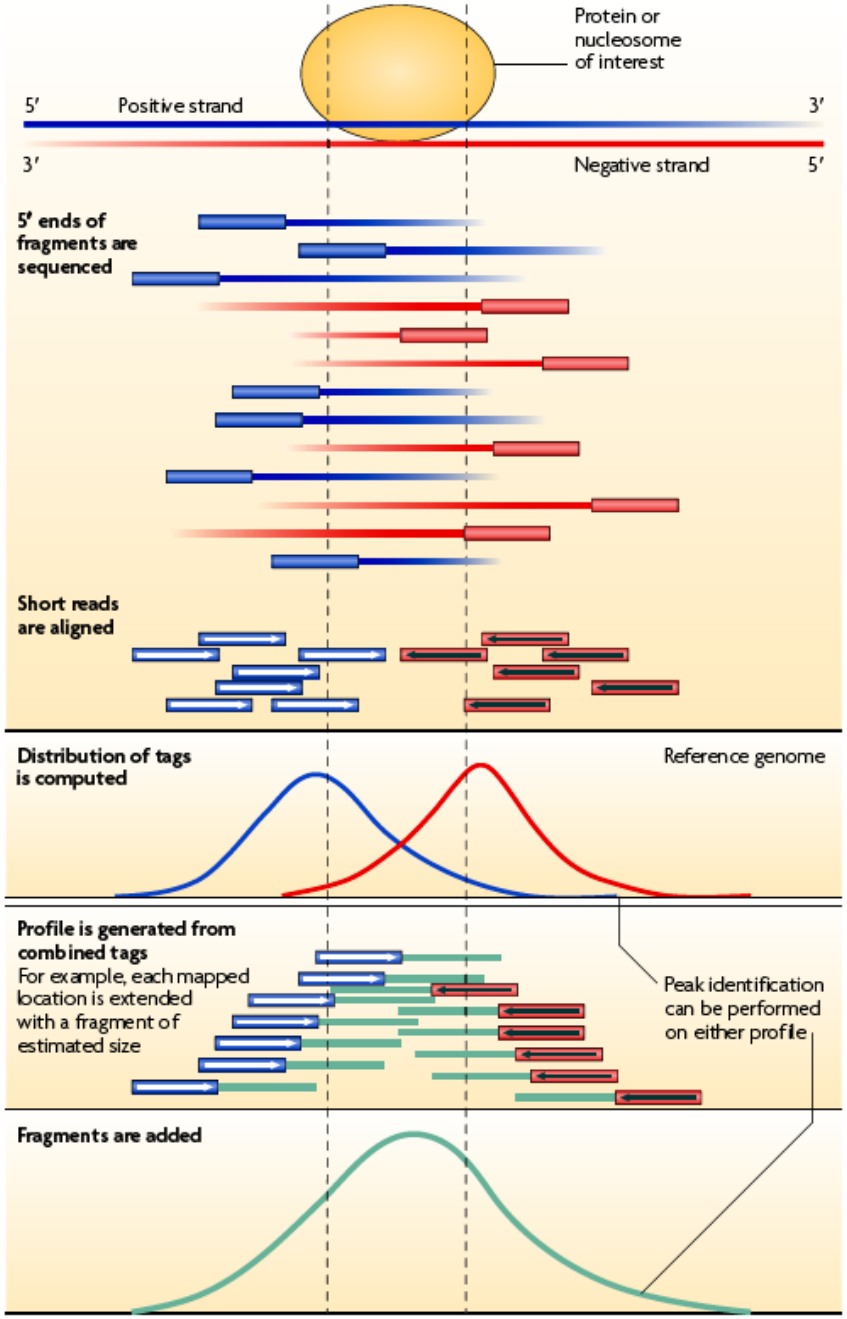
# Sequence Mapping & Filtering

- Alignment for ChIP-seq should allow for a small number of mismatches due to sequencing errors, SNPs an indels or the difference between the genome of interest and the reference genome.
- Only sequence reads mapped to a unique position on the reference genome are kept (about 50%). Reads mapped to multiple sites ('multi-reads') are usually discarded during 'normal' analysis. Consequently, peaks in highly repetitive regions are overlooked.
- However, repetitive regions have been linked to important biological functions such as disease susceptibility, immunity and defense. Note: A new method has been proposed to incorporate multi-reads into peak detection through the use of a weighted alignment scheme.
- A minimum five fold enrichment over the control sampled is required.

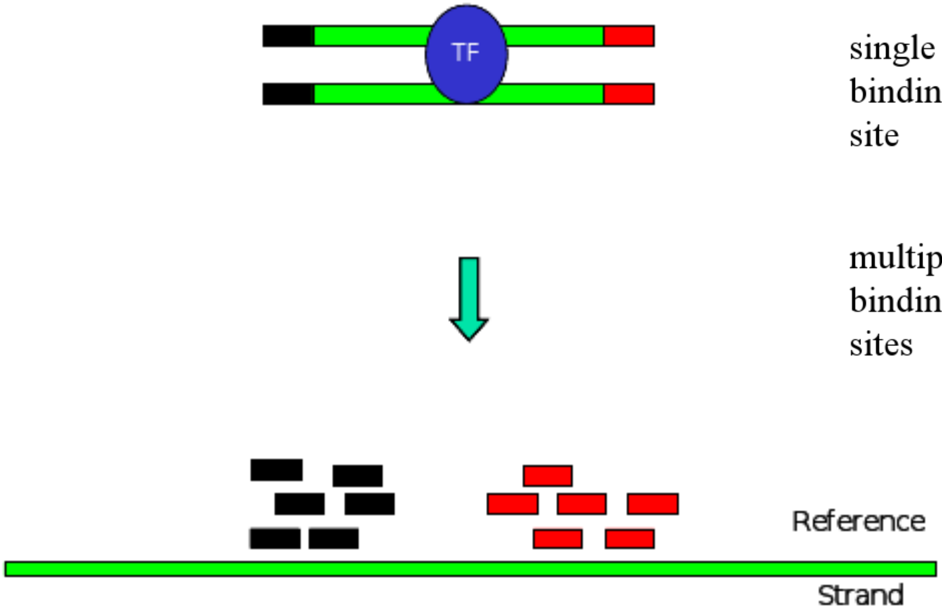
Nat Immunol. 2011 Sep 20;12(10):918-22

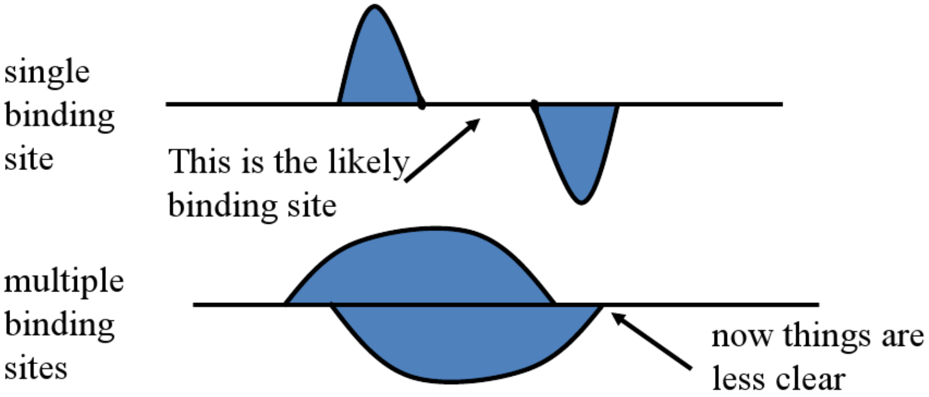
# Strand-specific profiles at enriched sites

the fragments are sequenced at the 5' end, and the locations of mapped reads should form two distributions, one on the positive strand and the other on the negative strand, with a consistent distance between the peaks of the distributions.


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

# One binding site has Potentially two peaks in read counts





single binding site

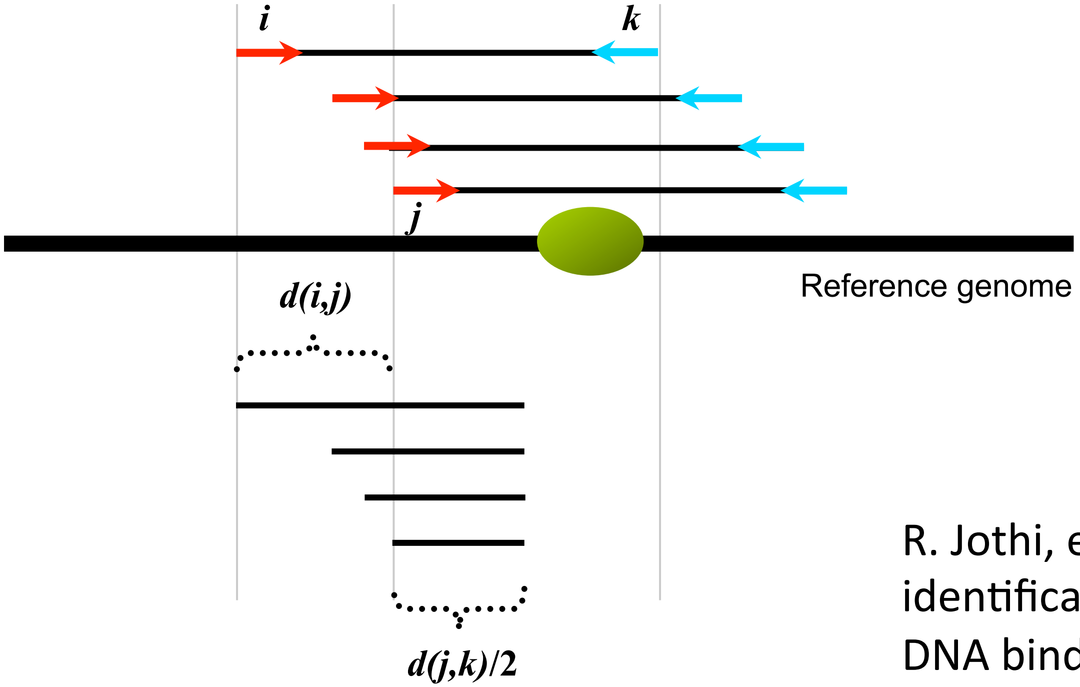
This is the likely binding site

multiple binding sites

now things are less clear

# Estimating fragment length

$$length = \frac{1}{n}\sum_{i=1}^{n}\left\{2d(i,j)+d(j,k)\right\}$$



R. Jothi, et al. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Research, 36:5221-31, 2008

Pictorial illustration of the DNA fragment length estimation.

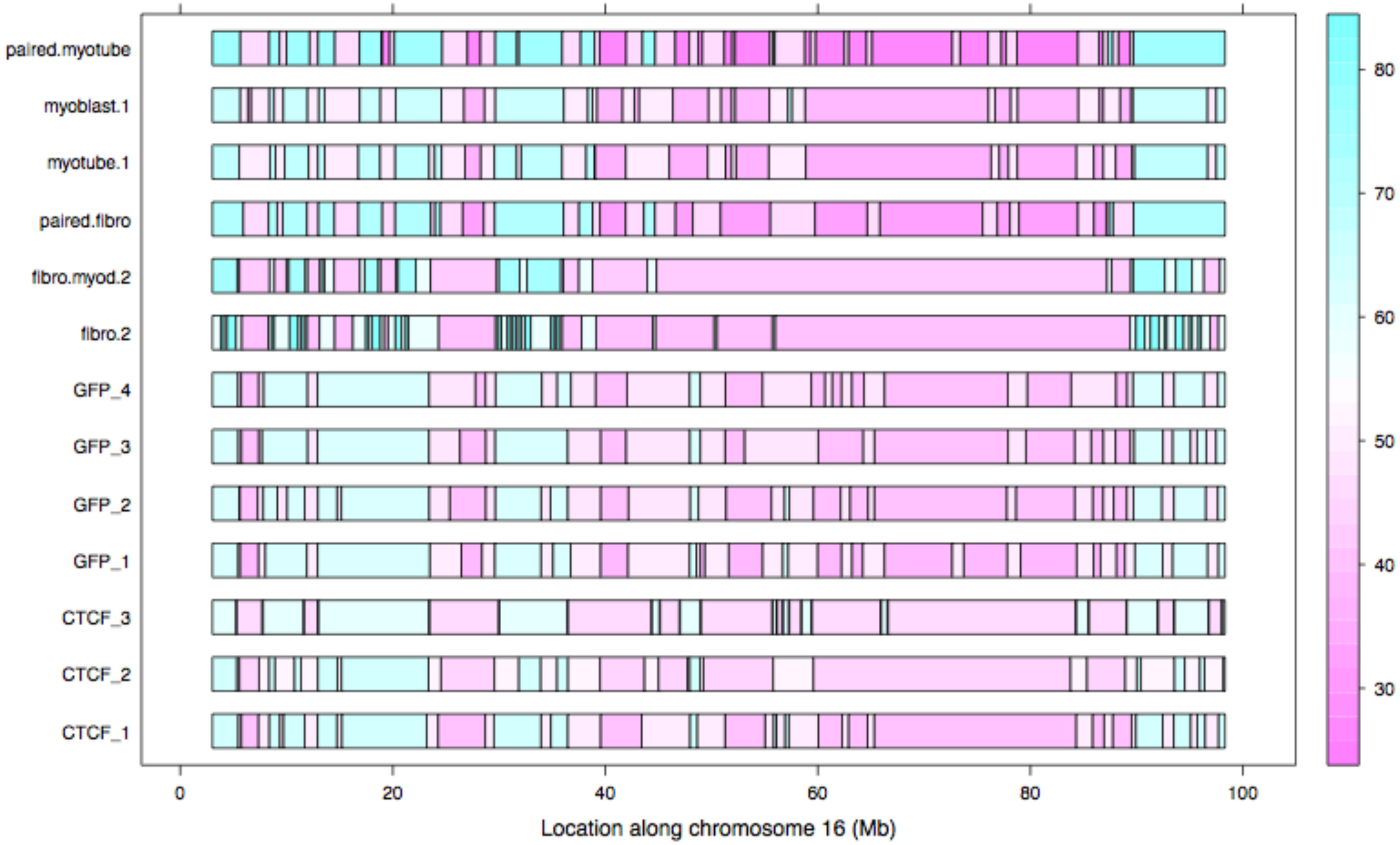
# Using fragment length

- With the estimated fragment length, one can shift the read position by half of the fragment length (more useful for transcriptional factor binding).
- Or one can extend the read to the fragment length to cover a larger section of the genome.

# Signal vs. Background

- We observe both reads that correspond to
  - Signal: binding we are interested in
  - Background: low density reads from throughout the genome.
- We want to separate these two types of signal
  - The background varies within a genome and between individuals.

# Background variation



# More issues about signal and background

- Another important issue in data analysis is comparison of the amount of histone modification or binding of transcription factors in two different cell types or under different conditions.

- Because of variations in ChIP conditions, the amount of noise may vary substantially between different samples even with the same antibody.

- Because scaling the data to sequenced depth does not eliminate systematic errors, normalization algorithms are needed for comparisons across samples.

Nat Immunol. 2011 Sep 20;12(10):918-22

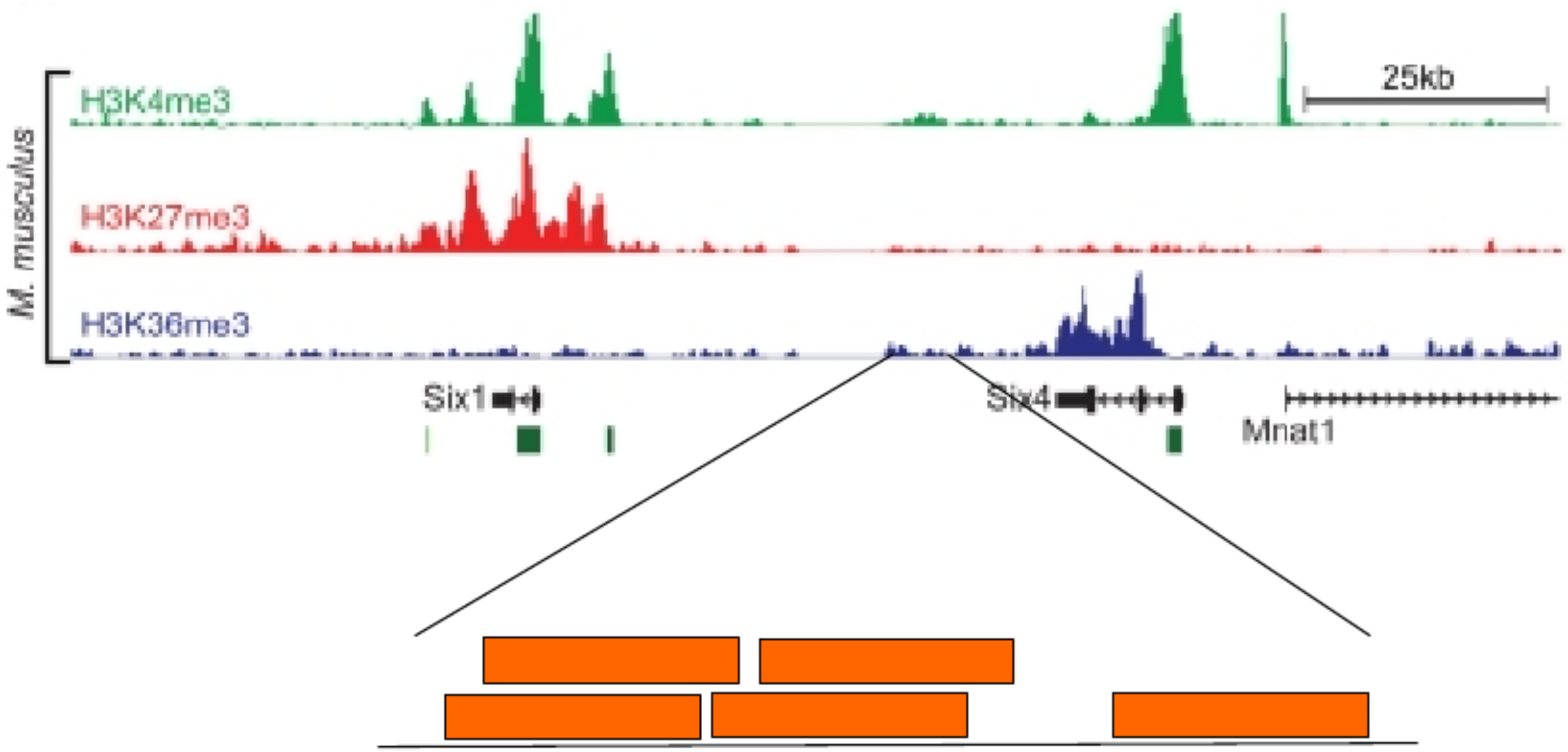
# Identification of enriched regions: Peak analysis

- After sequenced reads are aligned to the genome, the next step is to identify regions that are enriched in the ChIP sample relative to the control with statistical significance.
- Peak discovery: Determining the exact binding sites from short reads generated from ChIP-Seq experiments.
- Several 'peak callers' that scan along the genome to identify the enriched regions are currently available .

# Quantifying binding-peak finding

- Good algorithm should
  - Identify real peaks
  - Estimate confidence (e.g., calculate p-value and q-value)

# Peak finding



- Basic idea: count the number of reads in windows and determine whether this number is above background, and if so, define the region boundary.

# Peak finding

- Calling a region as bound with a protein can be done in different ways:
  - Hard thresholds (number of reads above some number k)
  - Kernel density estimators.
  - Hidden Markov models (HMM)
  - Compare bin counts to a background distribution determined from the input sample (or assuming a Poisson or negative Binomial distribution for example). This used by the chipseq package of BioConductor.

# Use peak height to test for the significance of the peak.

Assuming spatial Poisson process, let *X* be the height of a peak.

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \qquad x = 0,1,2.........$$

where λ is the mean of Poisson process (average read count at each position).

$$P(X \geq t) = \sum_{x=t}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!}$$

This gives a p-value for peak height of t.

# Use the mass (total read count) of a peak to determine its significance

- The total mass or tread count can be modeled with a geometric distribution (each read has to reach another read before it ends to keep the peak going). Suppose *X* is the mass of a peak

$$P(X = x) = p(1 - p)^{x-1}, x = 1,2......$$

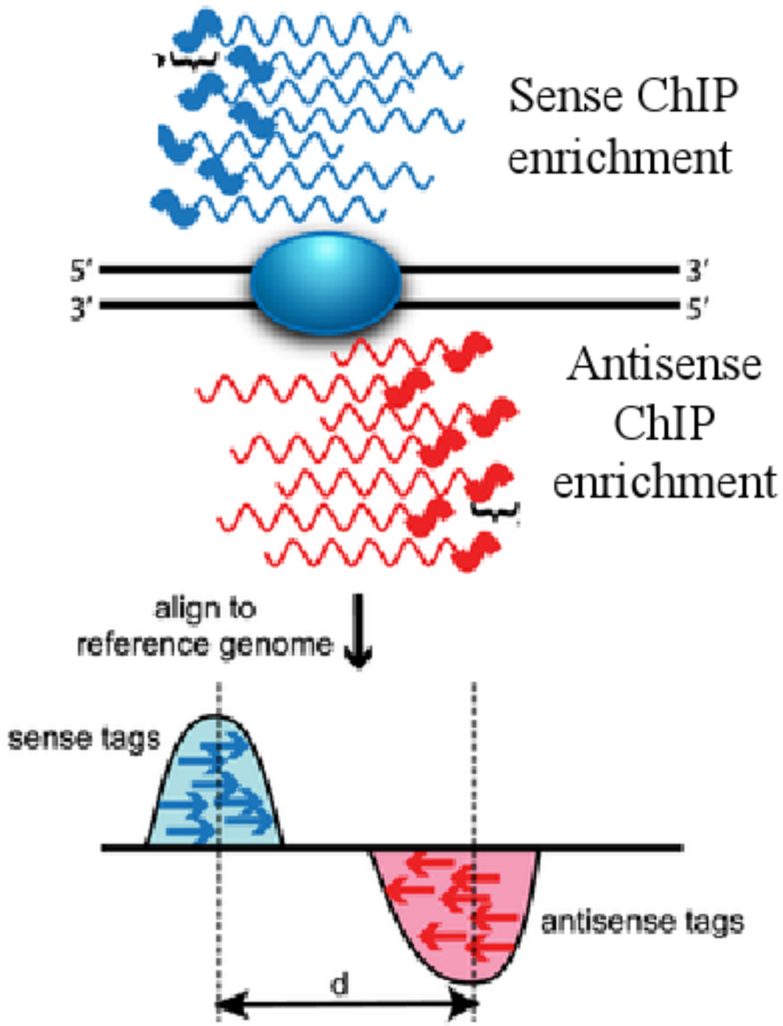
*p* is the probability that a position has no read.

$$P(X \geq t) = \sum_{x=t}^{\infty} p(1 - p)^{x-1}$$

This gives a p-value for getting a peak with mass *t* or bigger.

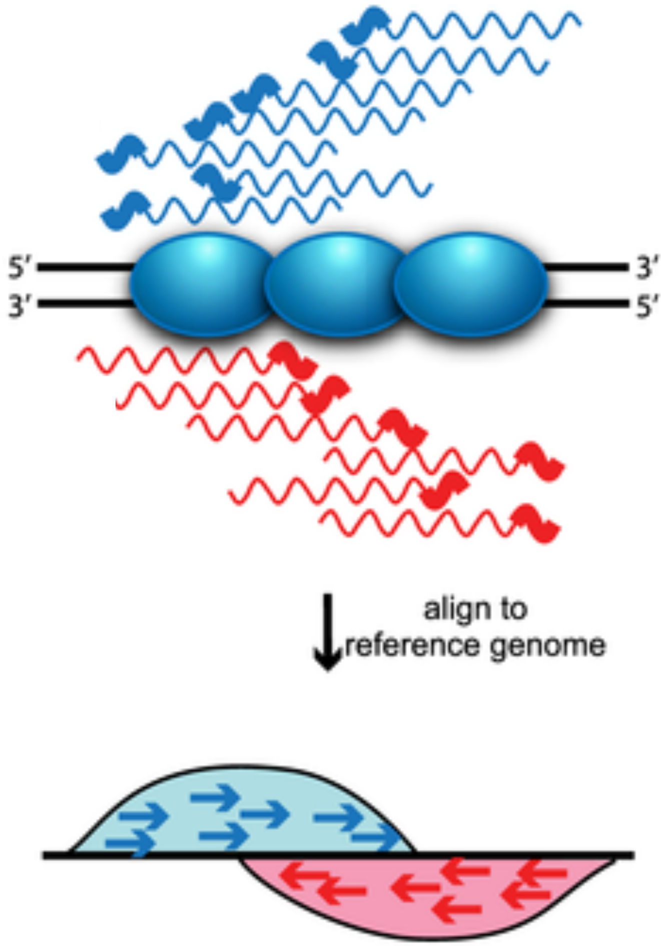
# Strand Specific methods

- Another feature that some methods consider is that reads can be from the plus or minus strands.
- In this case, for a given TF two peaks will be observed, separated by a constant distance, d
- This can be modeled either post-hoc, or by using strand specific calls



# Peak finding

- However, this is only useful where the protein being assayed has a sharp, well defined binding site.
- For histone modifications with broad and sometimes shallow peak, this information is less useful.



# MACS (Model-based Analysis for ChIP-seq)

MACS performs a peak-calling from ChIP-seq mapped reads through two main steps:

1. Modeling the shift size of ChIP-seq tags
2. Peak detection

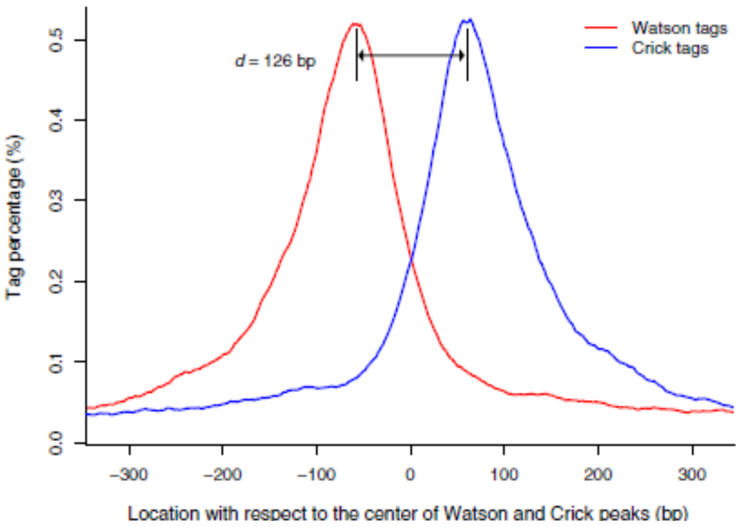
Basic idea:

- MACS takes advantage of the bimodal pattern of sense and antisense tags  to empirically model the shifting size to better locate the real binding sites.

Genome Biol (2008) vol. 9 (9) pp. R137

# MACS: shift size

- Given a sonication size (bandwidth) and a high-confidence fold-enrichment (mfold), MACS slides 2 bandwidth windows across the genome to find regions with tags more than mfold enriched relative to a random tag genome distribution.
- After that, MACS randomly samples 1000 of these high-quality peaks, separates their Watson and Crick tags, and aligns them between their Watson and Crick tag centers.
- The distance between the modes of the Watson and Crick peaks in the alignment is defined as 'd' and MACS shifts all the tags by d/2 toward the 3' ends to the most likely protein-DNA interaction sites.
- Modeling of the shift size is a way for MACS to guide peak detection.



# Some Issues

- ChIP-seq users are often curious as to whether they have sequenced enough to saturate all the binding sites. In principle, sequencing saturation should be dependent on the fold-enrichment, since higher-fold peaks are saturated earlier than lower-fold ones.

- MACS produces a saturation table to report, at different fold-enrichments, the proportion of sites that could still be detected when using 90% to 20% of the tags.

- while peaks with over 60-fold enrichment have been saturated, deeper sequencing could still recover more sites less than 40-fold enriched relative to the chromatin input DNA.

# Some Issues

When read counts from ChIP and controls are not balanced, the sample with more reads often gives more peaks even though MACS and other peak finders normalize the total read counts between the two samples.

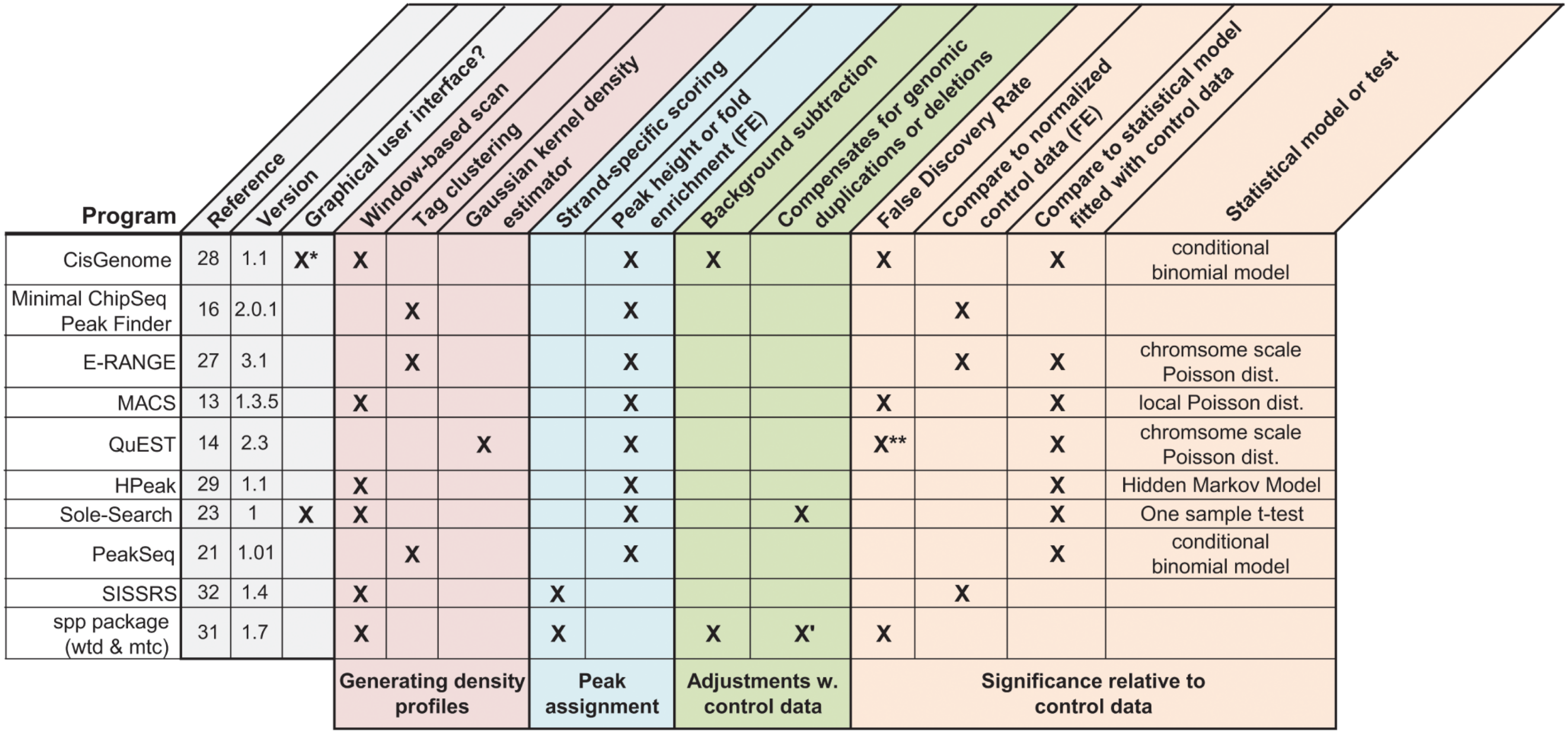
ChIP-seq users are suggested that if they sequence more ChIP tags than controls, the significance test of their ChIP peaks might be overly optimistic.

If a user has replicated files for ChIP or/and control, it is recommended to concatenate all replicates into one input file: pool of replicates.

| Program | Website | Language |
|---|---|---|
| MACS | http://liulab.dfci.harvard.edu/MACS/ | Python |
| QuEST | http://mendel.stanford.edu/SidowLab/downloads/quest/ | Perl |
| XSET | Not publicly released | |
| FindPeaks | http://vancouvershortr.sourceforge.net/ | java |
| TIROE | Not publicly released | |
| PeakSeq | http://www.gersteinlab.org/proj/PeakSeq/ | Perl / C |
| E-RANGE | http://woldlab.caltech.edu/rnaseq/ | Python |
| CisGenome | http://www.biostat.jhsph.edu/~hji/cisgenome/ | C/C++ |
| BayesPeak | http://www.compbio.group.cam.ac.uk/Resources/BayesPeak/csbayespeak.html | Perl / C |
| spp (R package) | http://compbio.med.harvard.edu/Supplements/ChIP-seq/ | R (not a formal package) |
| SISSRS | http://sissrs.rajajothi.com/ | Perl |
| CSDeconv | http://www.unisa.edu.au/maths/phenomics/csdeconv/ | MATLAB R2009a |
| SWEMBL | http://www.ebi.ac.uk/~swilder/SWEMBL/ | C |
| GeneTrack | http://code.google.com/p/genetrack/ | |
| HPeak | http://www.sph.umich.edu/csg/qin/HPeak/ | Perl |
| PICS | http://www.bioconductor.org/packages/release/bioc/html/PICS.html | R, Bayesian method |
| Bioconductor ChIPseq | http://www.bioconductor.org/packages/release/bioc/html/chipseq.html | R |

Wilbanks et al, 2010, PLoS One.

# Summary of some peak finders



| | | | | Generating density profiles | | | Peak assignment | | Adjustments w. control data | | Significance relative to control data | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Program | Reference | Version | Graphical user interface? | Window-based scan | Tag clustering | Gaussian kernel density estimator | Strand-specific scoring | Peak height or fold enrichment (FE) | Background subtraction | Compensates for genomic duplications or deletions | False Discovery Rate | Compare to normalized control data (FE) | Compare to statistical model fitted with control data | Statistical model or test |
| CisGenome | 28 | 1.1 | X* | X | | | | X | X | | X | | X | conditional binomial model |
| Minimal ChipSeq Peak Finder | 16 | 2.0.1 | | | X | | | X | | | | X | | |
| E-RANGE | 27 | 3.1 | | | X | | | X | | | | X | X | chromsome scale Poisson dist. |
| MACS | 13 | 1.3.5 | | X | | | | X | | | X | | X | local Poisson dist. |
| QuEST | 14 | 2.3 | | | | X | | X | | | X** | | X | chromsome scale Poisson dist. |
| HPeak | 29 | 1.1 | | X | | | | X | | | | | X | Hidden Markov Model |
| Sole-Search | 23 | 1 | X | X | | | | X | | X | | | X | One sample t-test |
| PeakSeq | 21 | 1.01 | | | X | | | X | | | | | X | conditional binomial model |
| SISSRS | 32 | 1.4 | | X | | | X | | | | | X | | |
| spp package (wtd & mtc) | 31 | 1.7 | | X | | | X | | X | X' | X | | | |

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

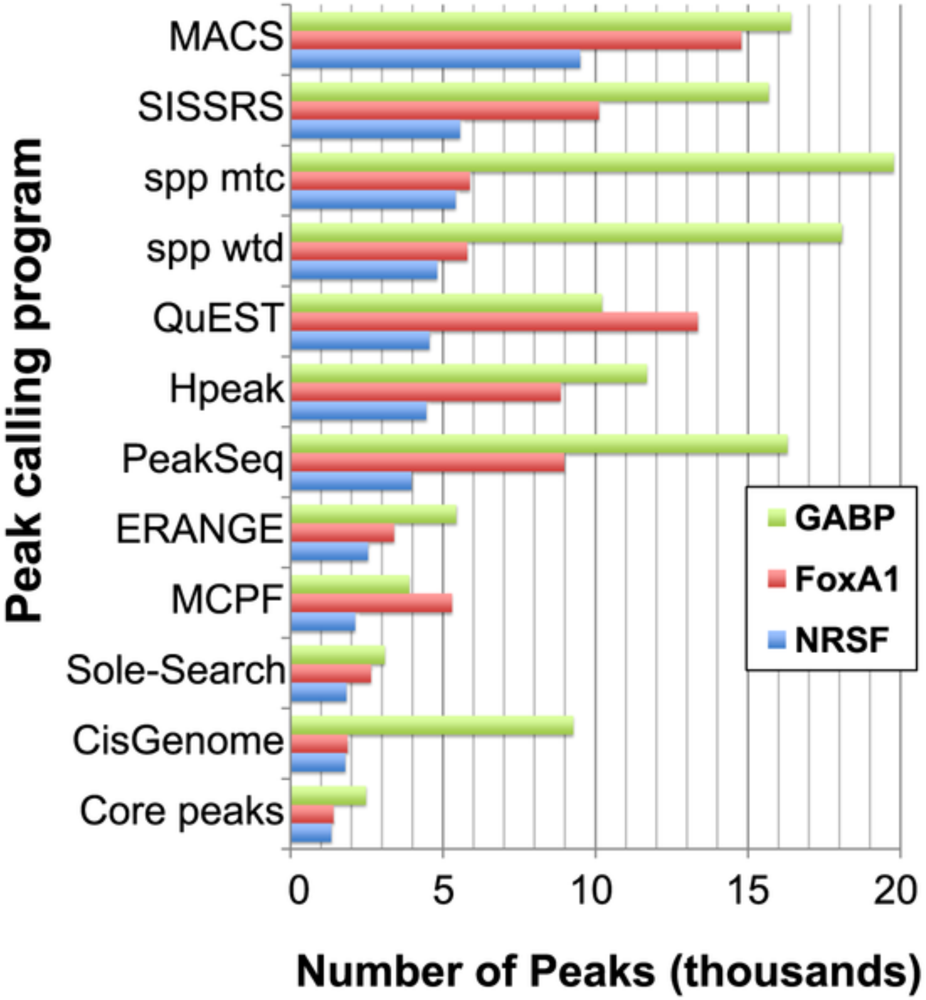
Wilbanks et al, 2010, PLoS One.

# Performance comparisons

- It is difficult to compare performance among different tools, because all methods rely on particular parameter values and need to be tuned accordingly to work best.

- However, some groups have applied multiple methods to the same dataset using their default parameters and compared results.

# Performance of 11 methods for calling binding sites for 3 TFs.



- The performance varies for different TFs.
- The performance of two bioconductor packages, PICS and BayesPeak, is similar to that of MACS
- Of course, more is not necessarily better.

Wilbanks et al. 2010, PLoS ONE.

# Agreement between different methods (NRSF).

| NRSF | CisGenome | Sole-Search | WOLD | ERANGE | PeakSeq | Hpeak | QuEST | wtd | mtc | SISSRS | MACS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CisGenome | X | 80 | 76 | 64 | 44 | 40 | 36 | 37 | 33 | 31 | 19 |
| Sole-Search | 82 | X | 81 | 68 | 45 | 40 | 36 | 38 | 34 | 37 | 19 |
| MCPF | 91 | 95 | X | 81 | 53 | 48 | 42 | 47 | 41 | 48 | 22 |
| ERANGE | 91 | 93 | 94 | X | 61 | 54 | 47 | 52 | 46 | 49 | 26 |
| PeakSeq | 98 | 99 | 100 | 100 | X | 85 | 66 | 78 | 69 | 78 | 43 |
| Hpeak | 98 | 99 | 100 | 100 | 91 | X | 69 | 83 | 74 | 80 | 43 |
| QuEST | 91 | 92 | 91 | 89 | 76 | 74 | X | 74 | 68 | 76 | 44 |
| spp wtd | 98 | 99 | 99 | 97 | 87 | 85 | 72 | X | 84 | 76 | 45 |
| spp mtc | 98 | 98 | 99 | 96 | 87 | 86 | 75 | 94 | X | 77 | 47 |
| SISSRS | 97 | 98 | 100 | 99 | 89 | 86 | 75 | 88 | 79 | X | 46 |
| MACS | 100 | 99 | 100 | 100 | 97 | 94 | 87 | 93 | 88 | 93 | X |

Percentage of total number of peaks called by one method (column) that are also discovered by another method (row).

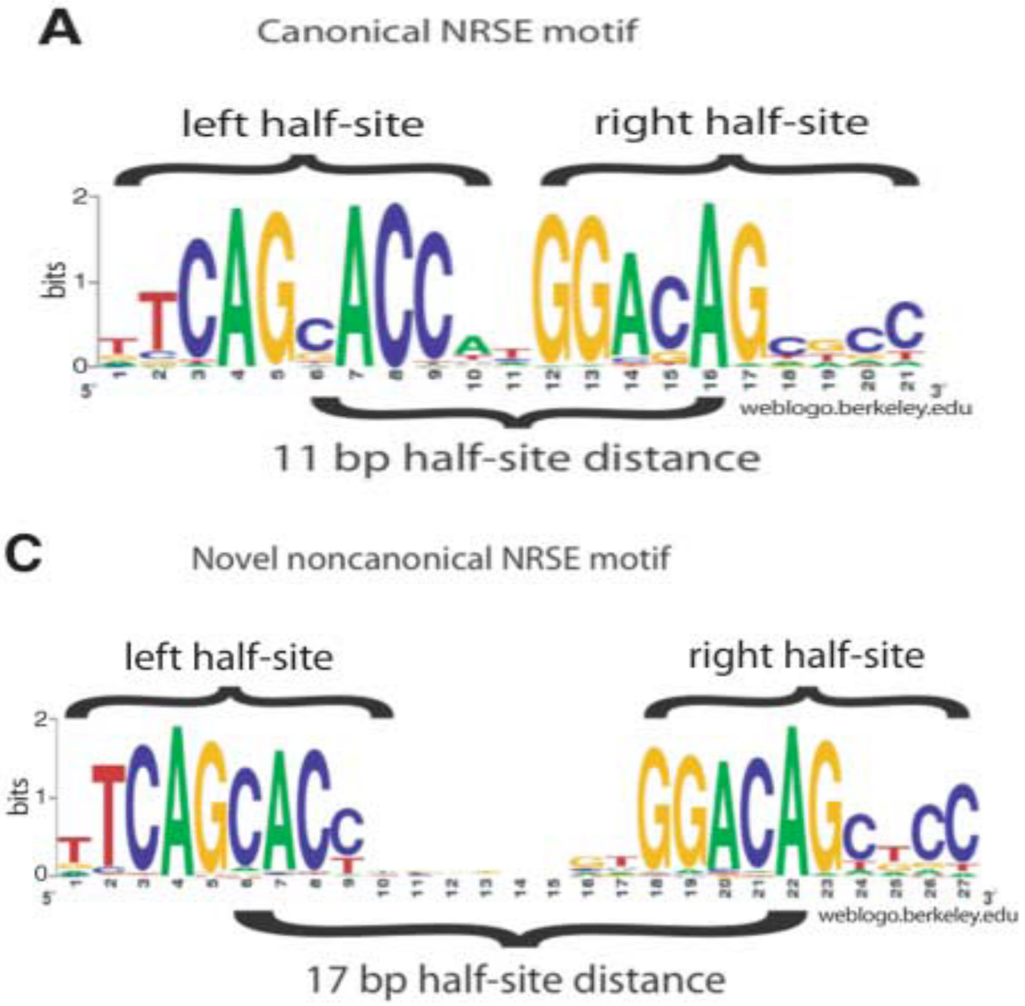
# What can we do?

- Try several methods and take the intersection of calls?
- If biological replicates exist, only consider peaks called in multiple samples?
- Use confidence measures associated with each peak in downstream analysis?
- Employing some combination of the first and second points are pretty common in practice.
- In general, methods have been developed for identifying regions where transcription factors bind. Methods for identifying regions where histone modifications occur are less mature, although some approaches (e.g., those based on HMMs) may be useful

# Post-processing

- Once we have decided what regions are peaks we need to try and interpret them.
- Typically that involves putting them in some form of genomic context.
- Bioconductor package Iranges/rtraklayer and various annotation packages can help.
- Identify protein binding motifs on DNA.

# Motif


A
Canonical NRSE motif
left half-site
right half-site
bits
2
1
0
5'
3'
weblogo.berkeley.edu
11 bp half-site distance
C
Novel noncanonical NRSE motif
left half-site
right half-site
bits
2
1
0
5'
3'
weblogo.berkeley.edu
17 bp half-site distance

# Differentially Enriched Peaks

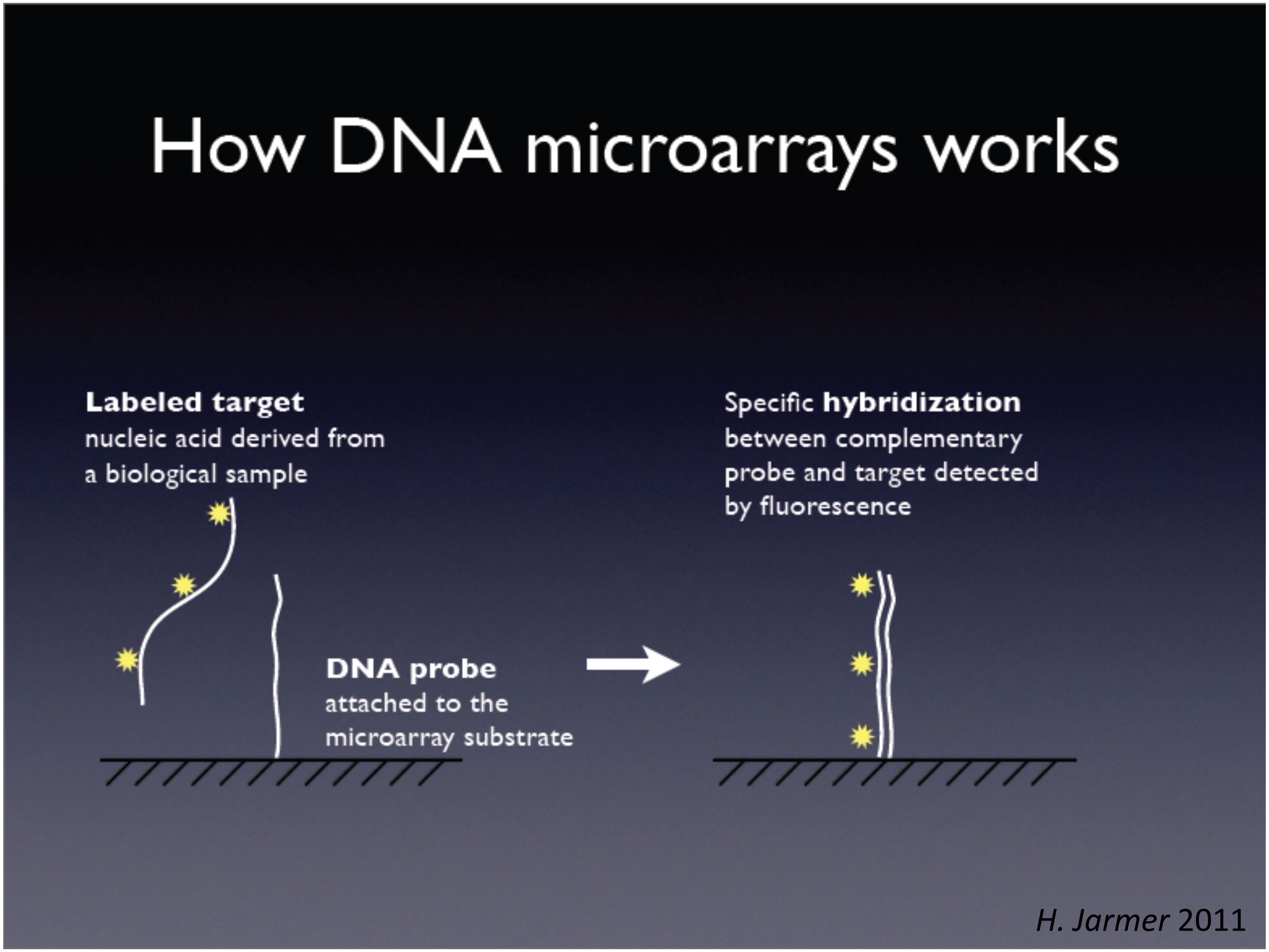
- Suppose we have two treatment conditions.
- We want to know which peaks are differentially enriched (low in one condition and high in the other).
- One could use some cut-off in one condition, and then look for peaks in the other.
- Instead, we combine the data into one collection, choose a fairly relaxed cut-off to define intervals of interest.
- We can then find DE peaks by a number of methods.
- A regression type approach with DESeq or edgeR seems to work
- Normalization is an important problem

------how to deal with different numbers of reads in the different samples.

# Post-post-processing

Validation of a number of peaks is always recommended in a ChIP-seq analysis !!!

Nat Rev Genet. 2009 Oct;10(10):669-80.



# How DNA microarrays works

**Labeled target**
nucleic acid derived from
a biological sample

**DNA probe**
attached to the
microarray substrate

Specific **hybridization**
between complementary
probe and target detected
by fluorescence

*H. Jarmer* 2011