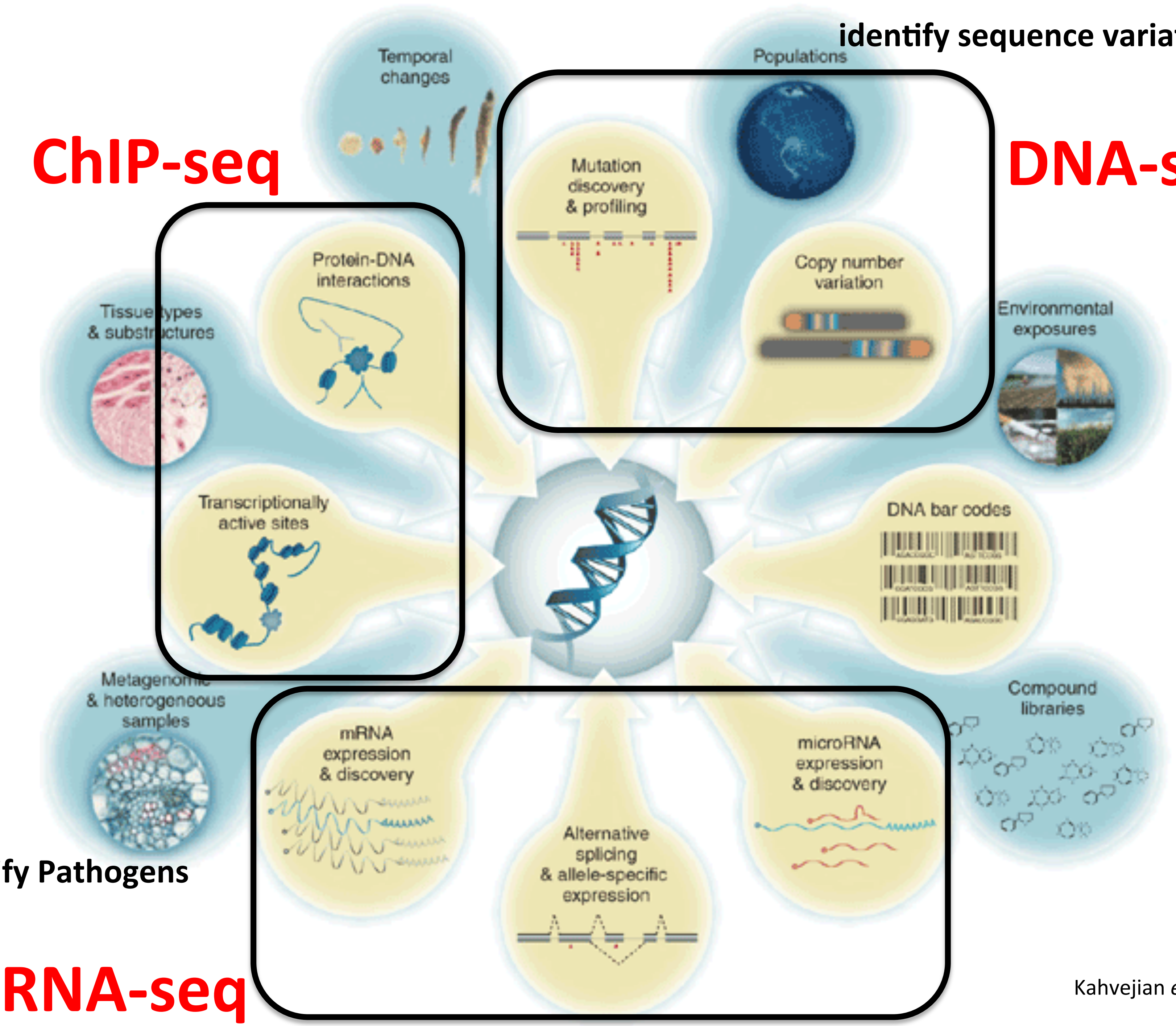
Next-generation Sequencing

Lecture 13


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

identify sequence variations

ChIP-seq

DNA-seq

Identify Pathogens

RNA-seq

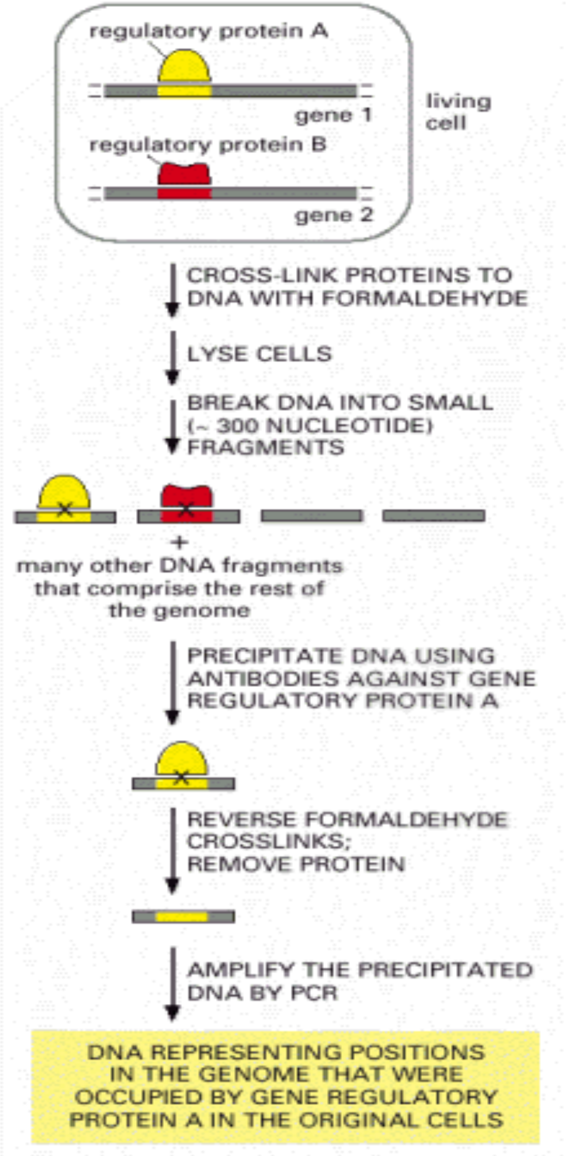
Kahvejian *et al*, 2008

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.
- 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.



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

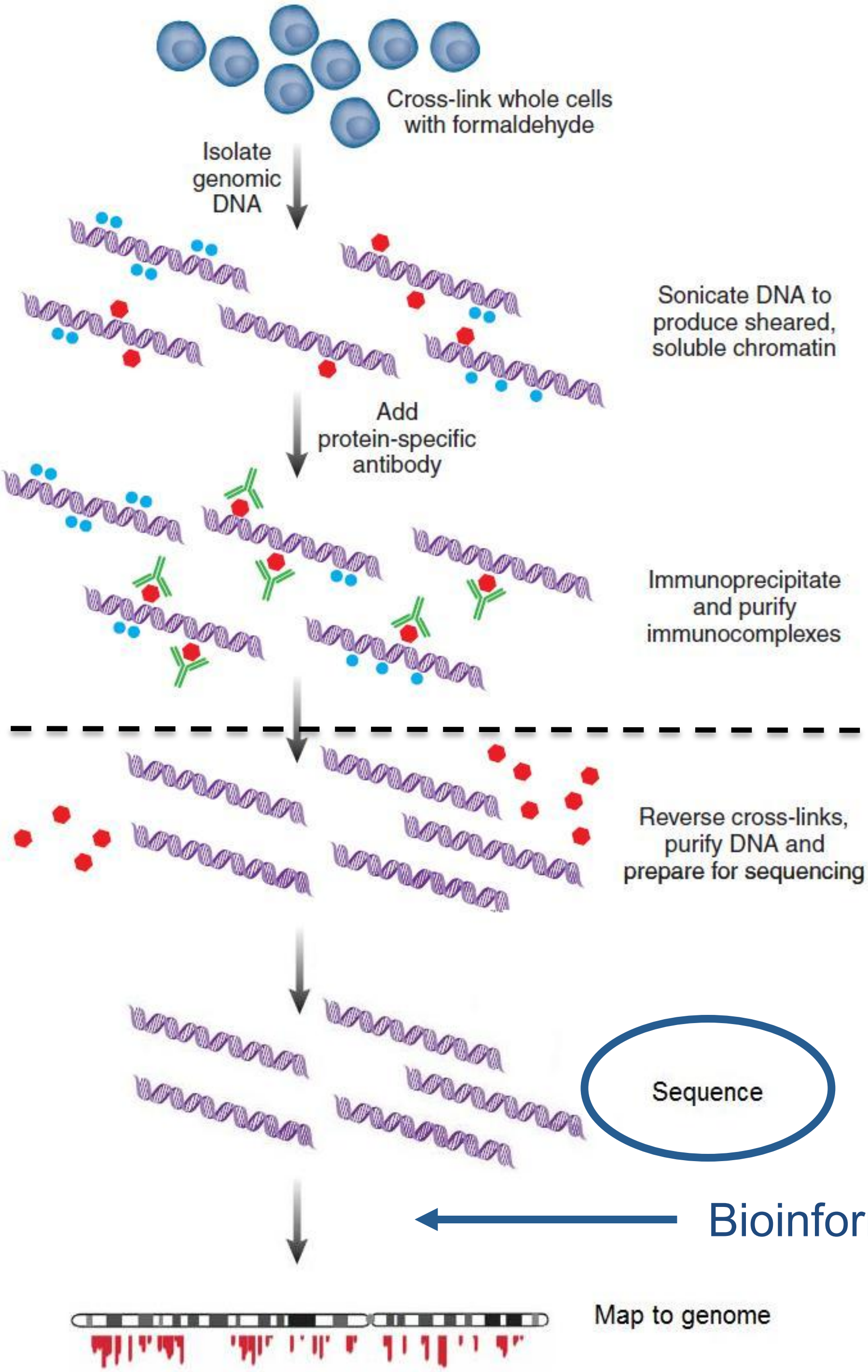
REVERSE FORMALDEHYDE CROSSLINKS; REMOVE PROTEIN

AMPLIFY THE PRECIPITATED DNA BY PCR

DNA REPRESENTING POSITIONS IN THE GENOME THAT WERE OCCUPIED BY GENE REGULATORY PROTEIN A IN THE ORIGINAL CELLS

ChIP

Seq


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
Map to genome

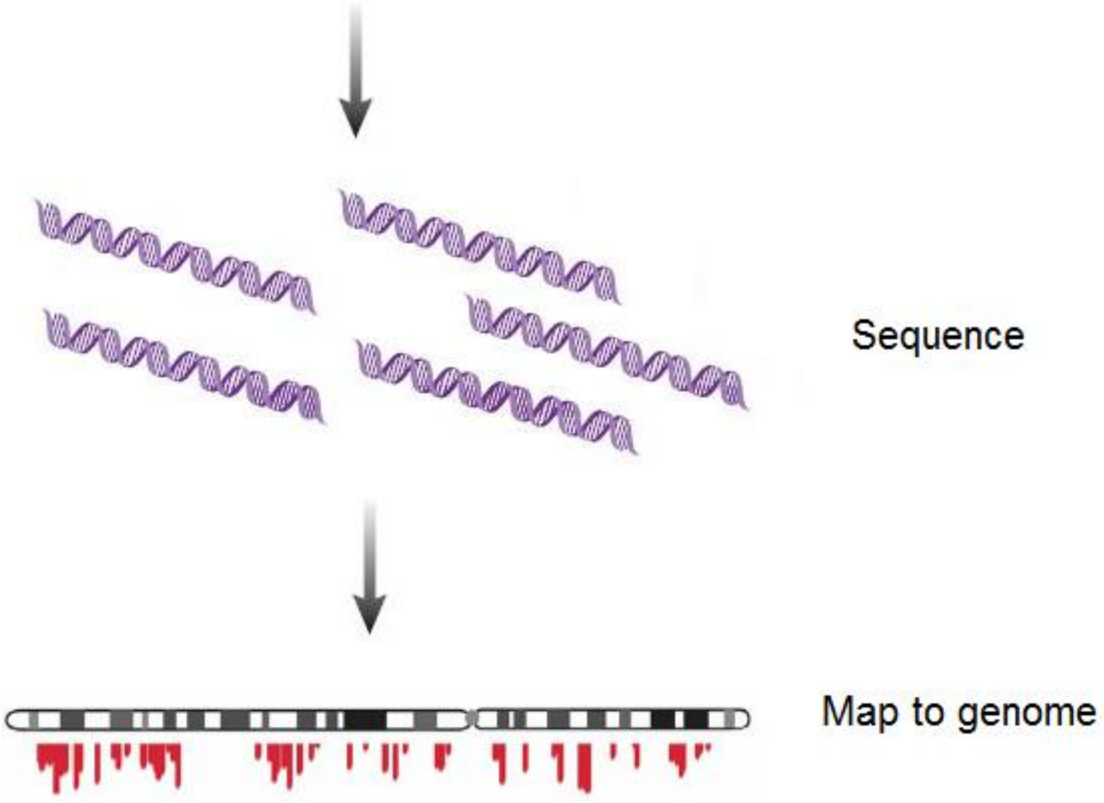
Bioinformatics

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 acceptable 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.

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

ChIP-seq v.s. ChIP-chip

- ChIP-seq has higher resolution, fewer artifacts, greater coverage and a larger dynamic range than ChIP-chip.
- In ChIP-seq, the DNA fragments of interest are sequenced directly instead of being hybridized on an array.
- For high-resolution profiling of an entire large genome, ChIP-seq is already less expensive than ChIP-chip.

Work flow of ChIP-seq

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

Sample preparation

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

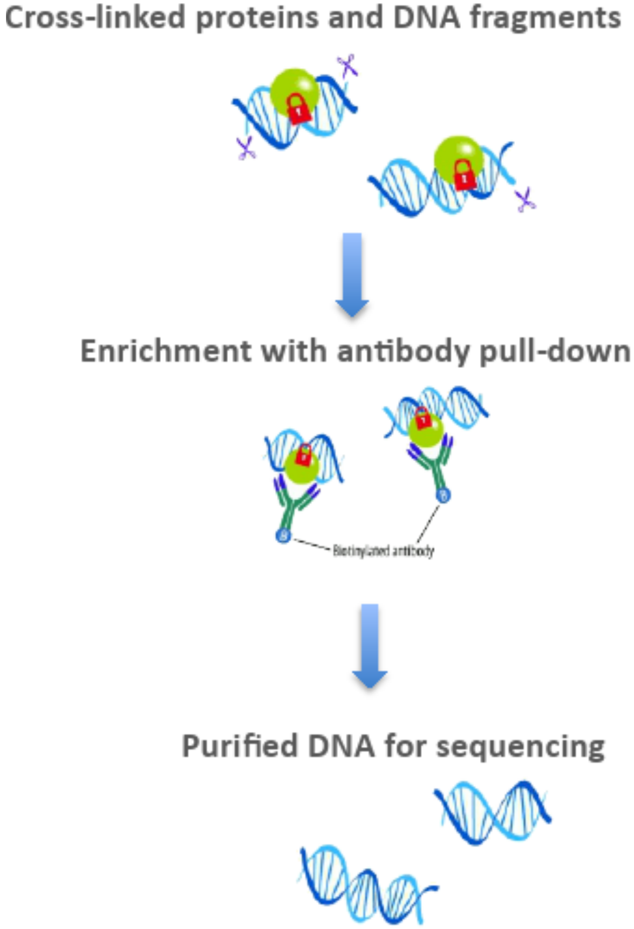
(2) the chromatin is sheared by sonication into small fragments.

(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).



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

Antibody issues

- There are often multiple antibodies for a particular protein
    - For P53, there are two widely used
- 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.

Issues for library preparation

- 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

Work flow of ChIP-seq

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

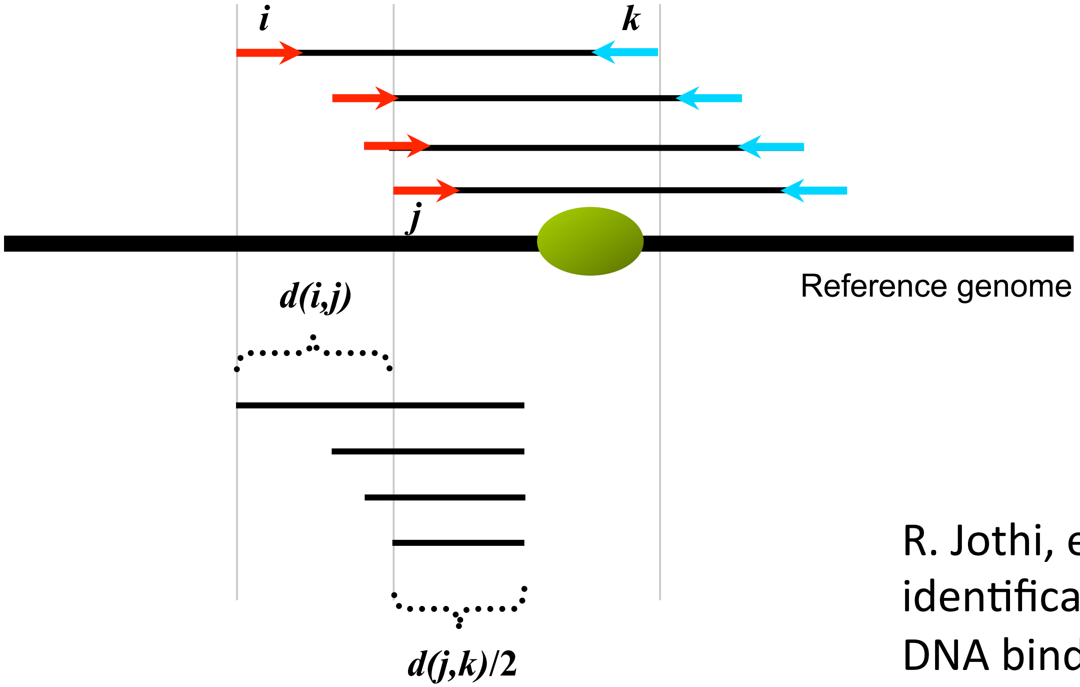
Sequence Mapping & Filtering

- 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.
- However, repetitive regions have been linked to important biological functions such as disease susceptibility, immunity and defense.
- A minimum five fold enrichment over the control sampled is required.

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

Estimating fragment length

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



i,k are paired
j is nearest to k

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

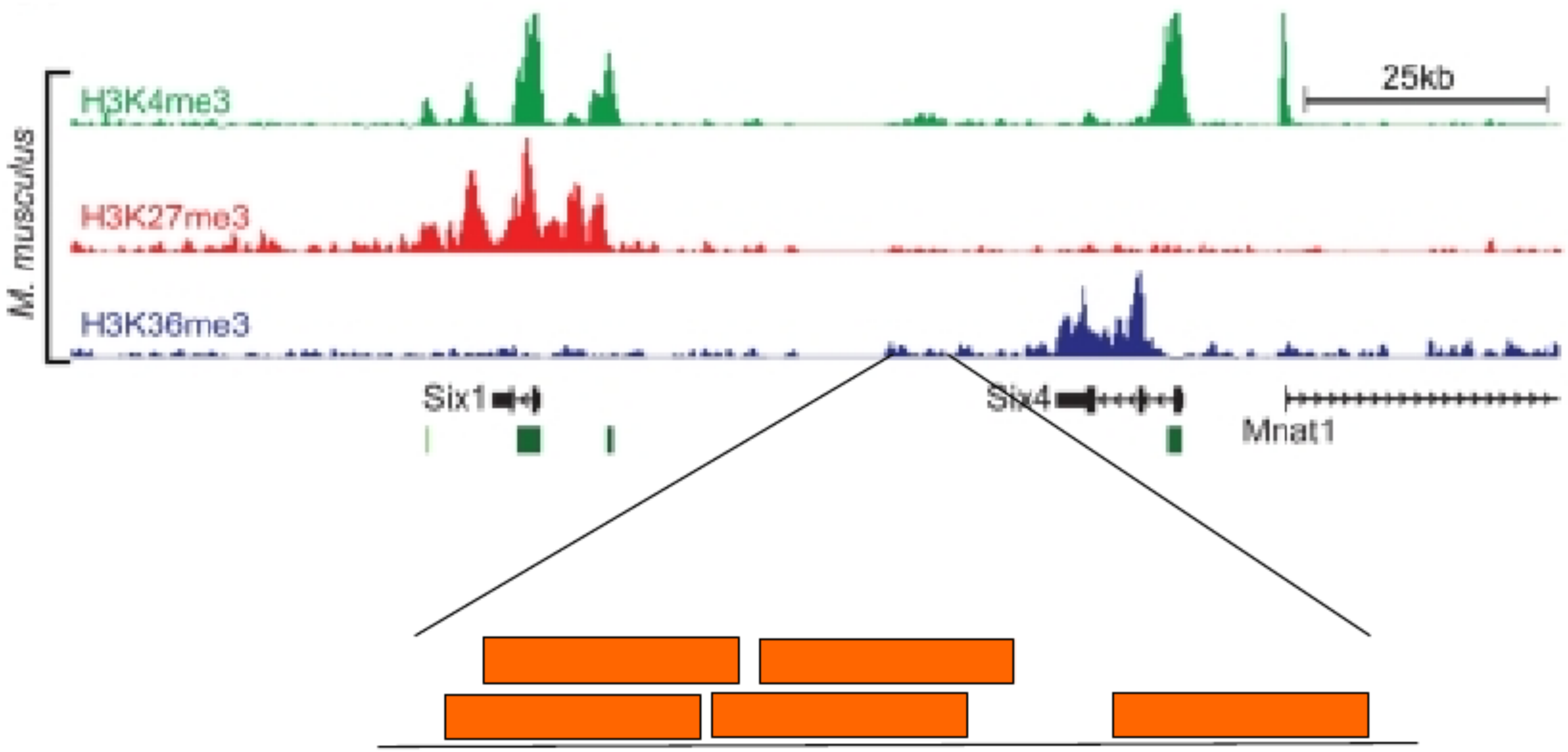
Identification of enriched regions

- 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.

Peak Analysis

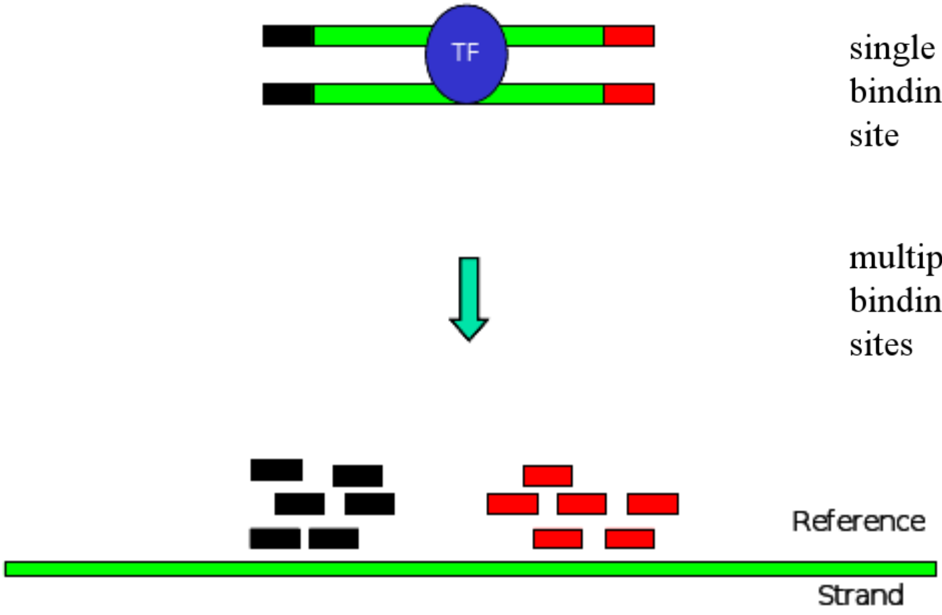
- Identify peaks (peak calling)
- Estimate confidence and find significant peaks (e.g., calculating p-values and removing background noise)

Peak finding

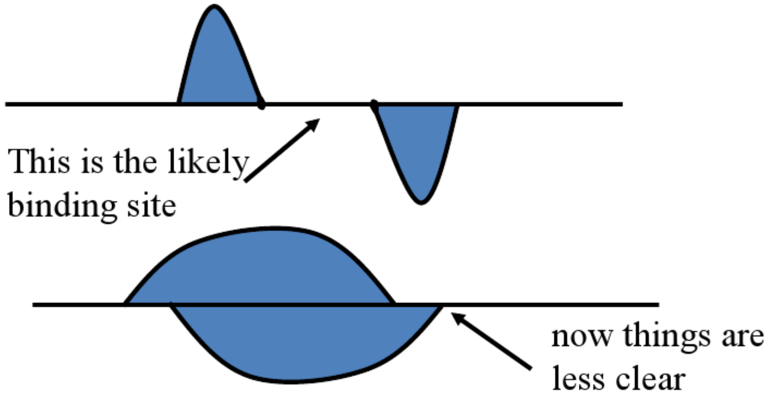


- Basic idea: count the number of reads in windows and determine whether this number is above background, and if so, define the region boundary.

One binding site has Potentially two peaks in read counts



single binding site



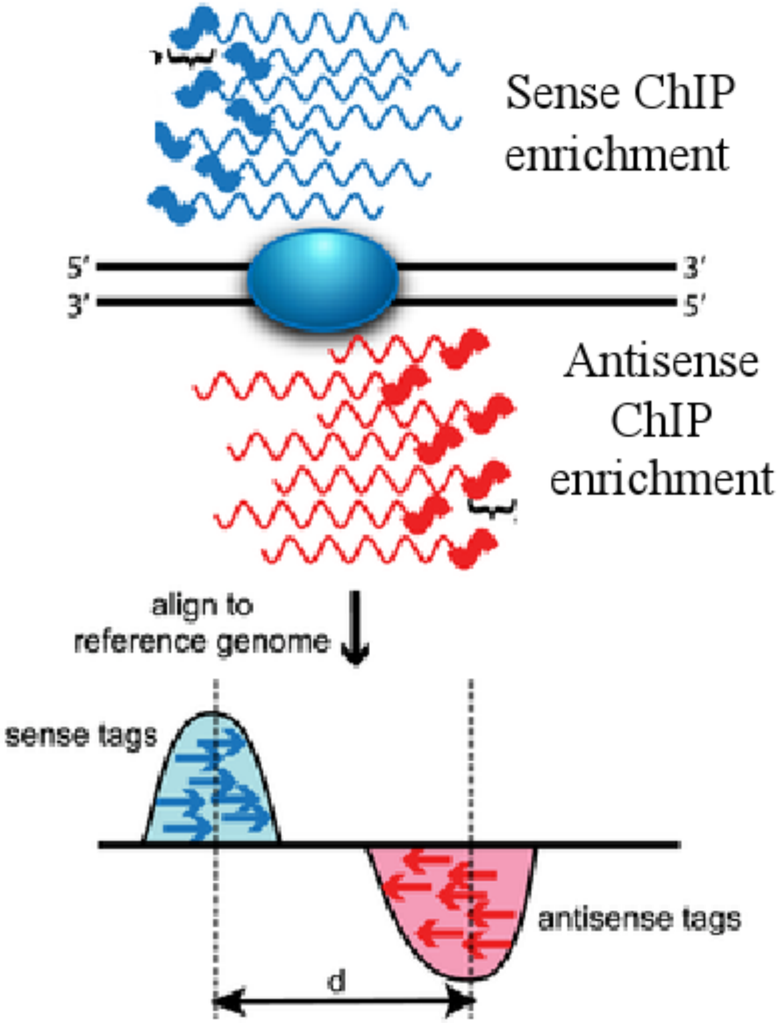
This is the likely binding site

multiple binding sites

now things are less clear

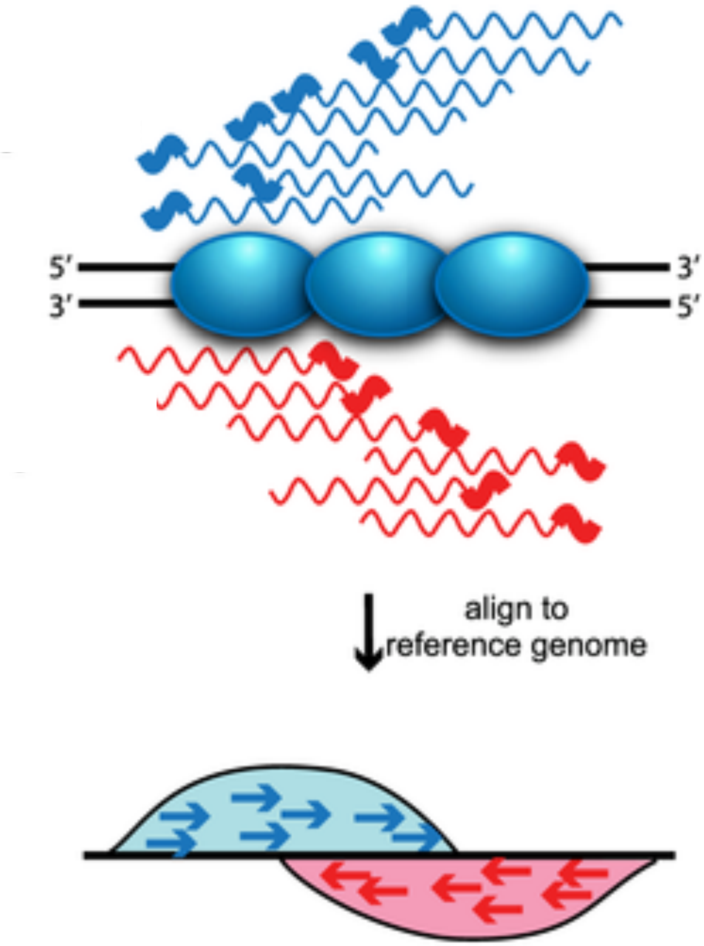
One binding site has two peaks

- 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).



Multiple binding sites

- For example, histone modifications may cause broad and sometimes shallow peak



Peak Analysis

- Identify peaks (peak calling)
- Estimate confidence and find significant peaks (e.g., calculating p-values and removing background noise)

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

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

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

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.

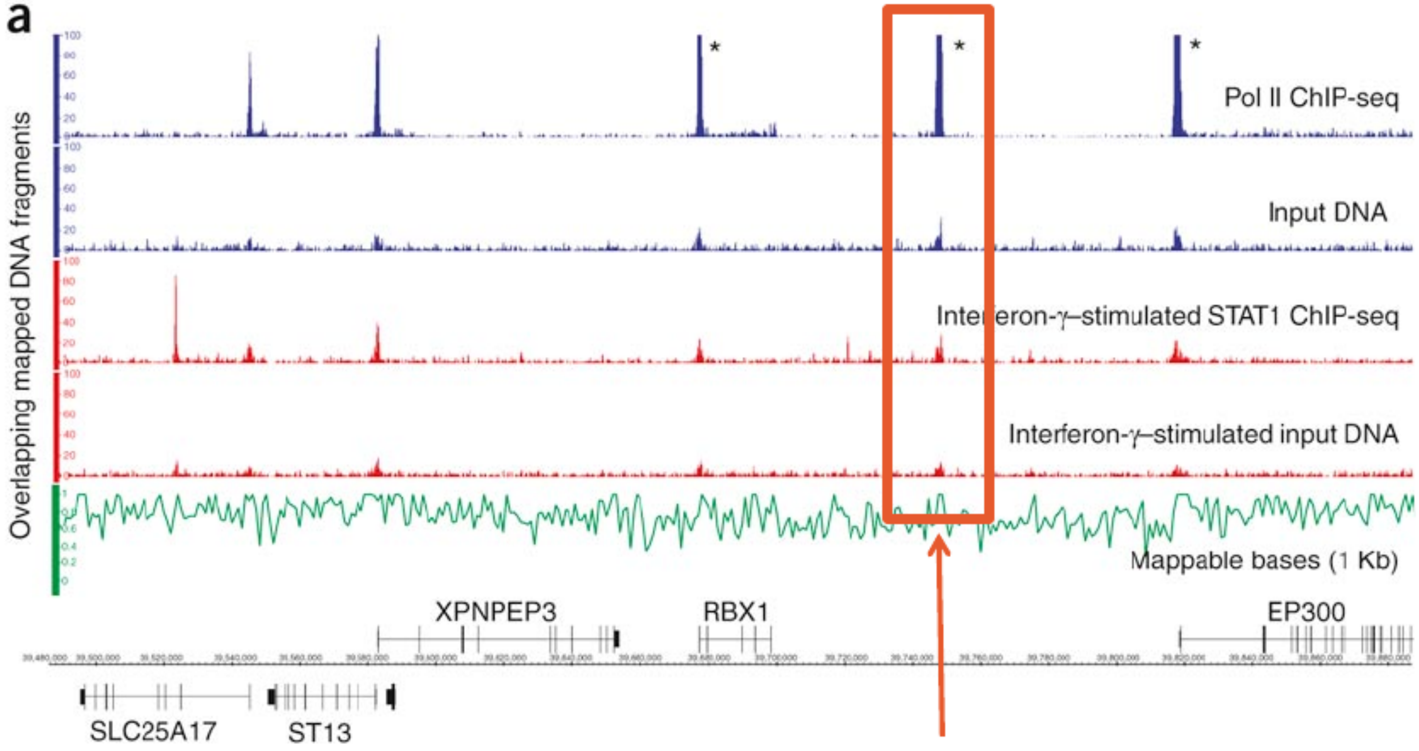
$$P(X \geq t) = \sum_{x=t}^{\infty} p(1-p)^{x-1}$$

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

Controls for background

- 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

Some Issues

- When read counts from ChIP and controls are not balanced, the sample with more reads often gives more peaks even though 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.

- In addition, when an insufficient number of reads is generated, there is a significant loss of sensitivity or specificity in detection of enriched regions.

Replicates

- Many factors may contribute to variability between data sets, to ensure reliability of the data, biological replicate experiments are necessary.
- 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.
- 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.

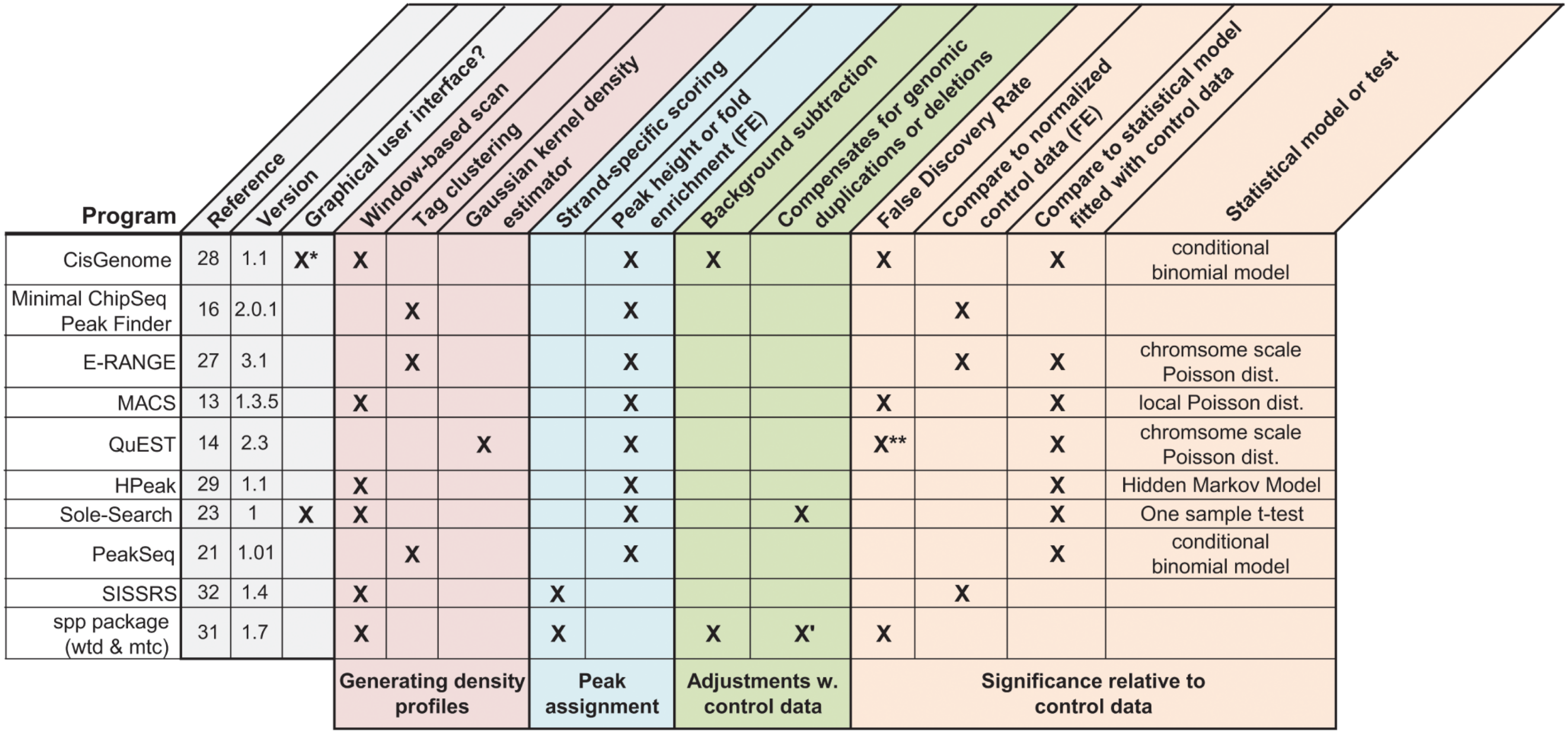
Nat Immunol. 2011 Sep 20;12(10):918-22

Tools for Chip-Seq data analyses

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



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			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

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.

MACS (Model-based Analysis for ChIP-seq)

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

1. Based on pattern of sense and antisense tags, Modeling the shift size of ChIP-seq tags
2. Peak detection using peak height to fit a local Poisson's distribution.

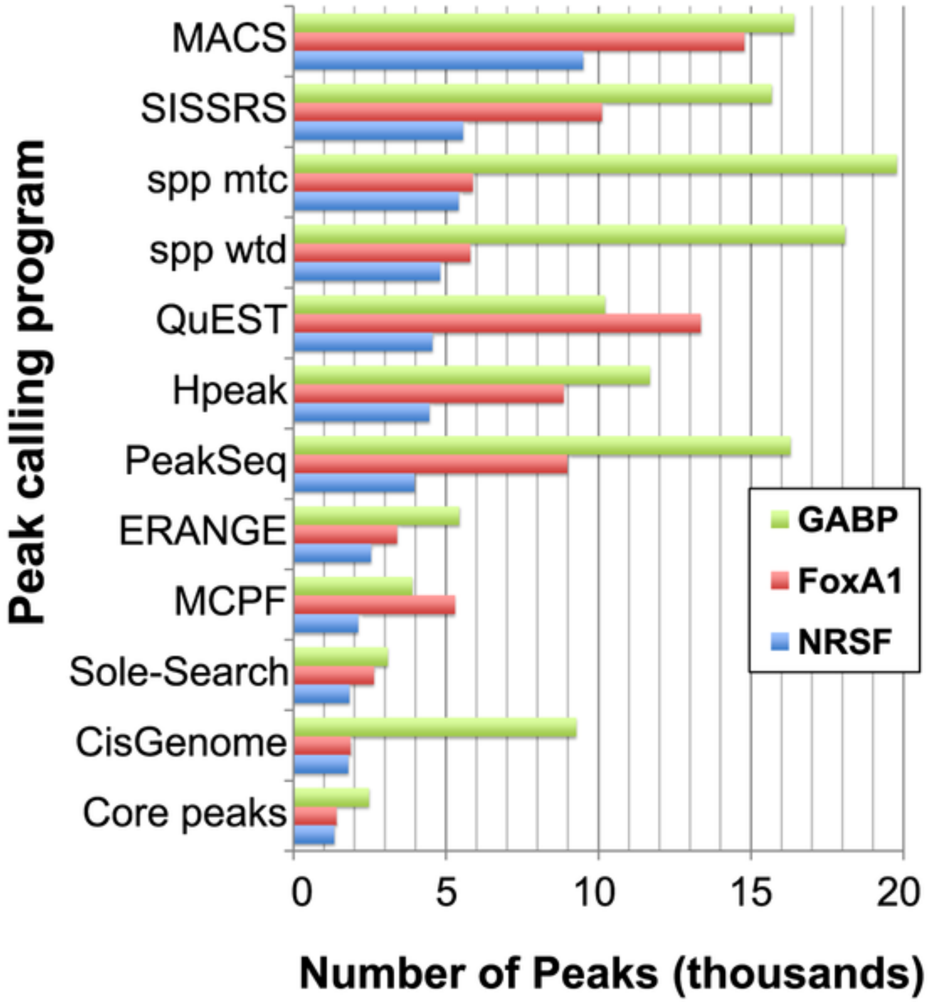
Genome Biol (2008) vol. 9 (9) pp. R137

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.
- More is not necessarily better.

Wilbanks et al. 2010, PLoS ONE.

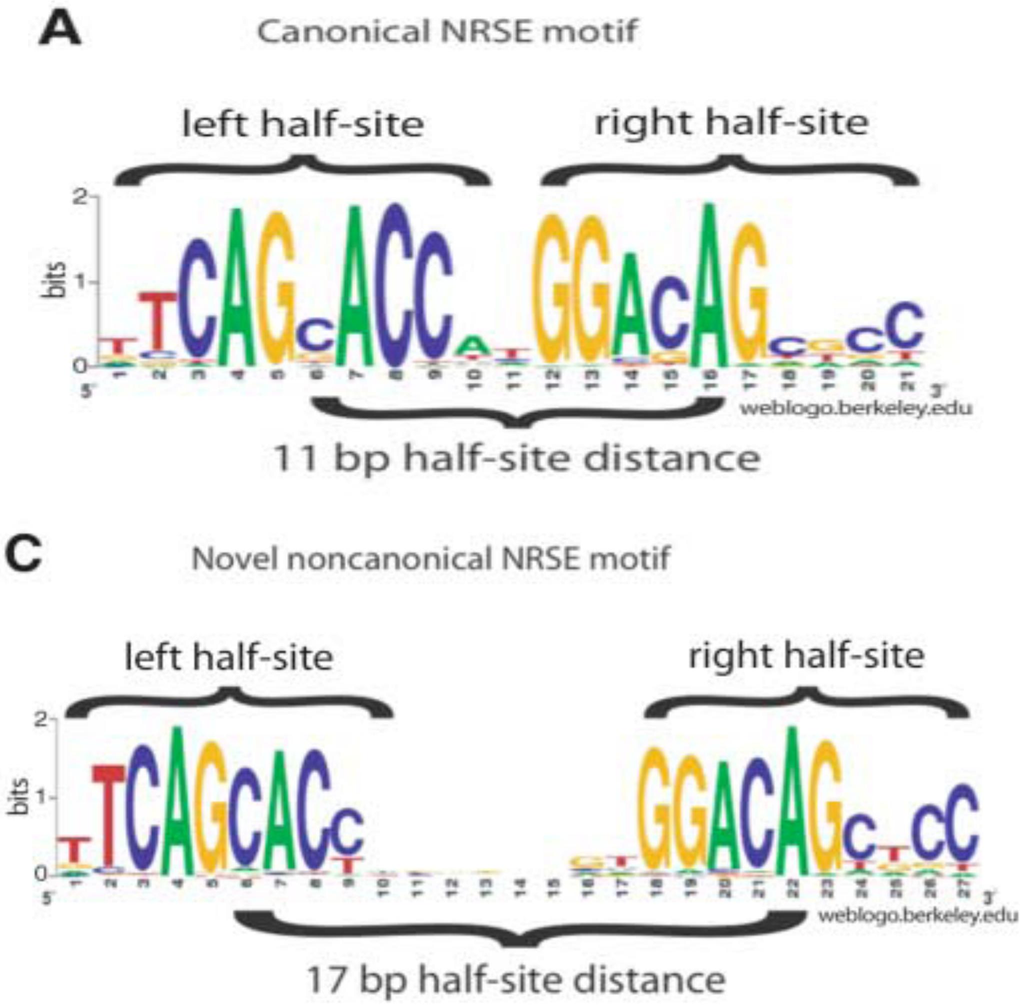
What can we need to know

- Try several methods and take the intersection of calls.
- If biological replicates exist, only consider peaks called in multiple samples.
- 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

- We need to try and interpret discovered peak regions.
- Typically that involves putting them in some forms of genomic context.
- Various annotation packages can help, such as genome browser.
- Identify protein binding motifs on DNA.

Motif


A
Canonical NRSE motif
left half-site
right half-site
bits
2
1
0
5'
3'
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
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'
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
weblogo.berkeley.edu
17 bp half-site distance

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.