ChIPSeq Data Analysis

Introduction

Peter FitzGerald, CCR, NCI
Head Genome Analysis Unit, CCR, NCI

Course Outline

Day 1

- Introduction - Design and Analysis Overview (9:30 - 11:30 am)
- An Overview on Experimental Subtypes and Variations of ChIP-Seq (11:30 - 12:30 pm)
  Alexei Lobanov
- Analysis of ChIP-Seq data: Raw Data to Results (1:00 pm – 4:00 pm)
  Bong-Hyun Kim

Day 2

- Hands-on Tutorial for Analysis of ChIP-Seq data with the Genomatix Genome Analyzer (GGA)(9:30 - 12:30 pm)

- Mining ChIP-Seq data from Public Databases (1:00 - 4:0 pm)
  Bong-Hyun Kim

Talk Outline

- Introduction/Background
- Comparison to ChIP-chip
- Experimental Design
- Data analysis
- Analysis in Detail
- Functional Analysis
- Visualization
- File Formats

Course Goals

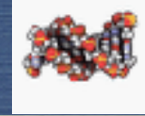
- Provide some basic knowledge on how to generate and interpret ChIPSeq data.
- Equip you with the fundamental knowledge required to understand what the data analysis entails.
- Impart enough understanding of the analytic process to enable you to establish strategic partnership with bioinformatician collaborators.
- Provide hands-on experience with both a commercial (Genomatix) and an Open Source Tool (MACS, SICER, MEME)

ChIP-Seq Background

ChIPSEQ

Chromatin
Immuno**P**recipitation (ChIP)
and massively parallel
sequencing (SEQ)

First reported by several groups in **2007**... now the most widely used technique for analyzing DNA:Protein interactions

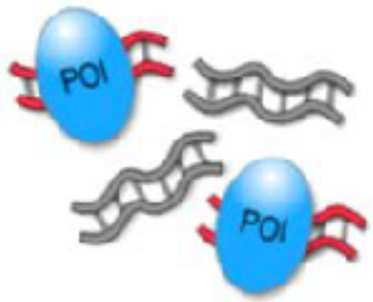


ChIP-seq overview

DNA + bound protein

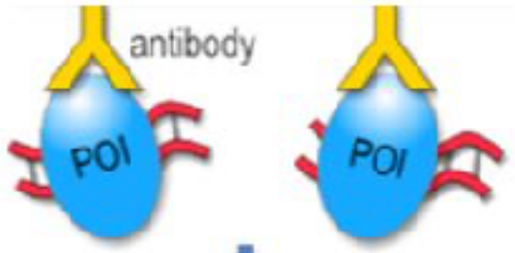
cross-link and shear

Fragment DNA



IP

Immunoprecipitate

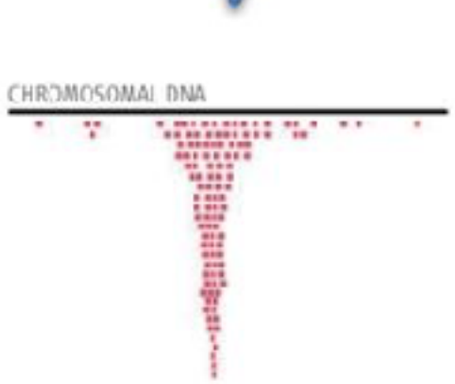


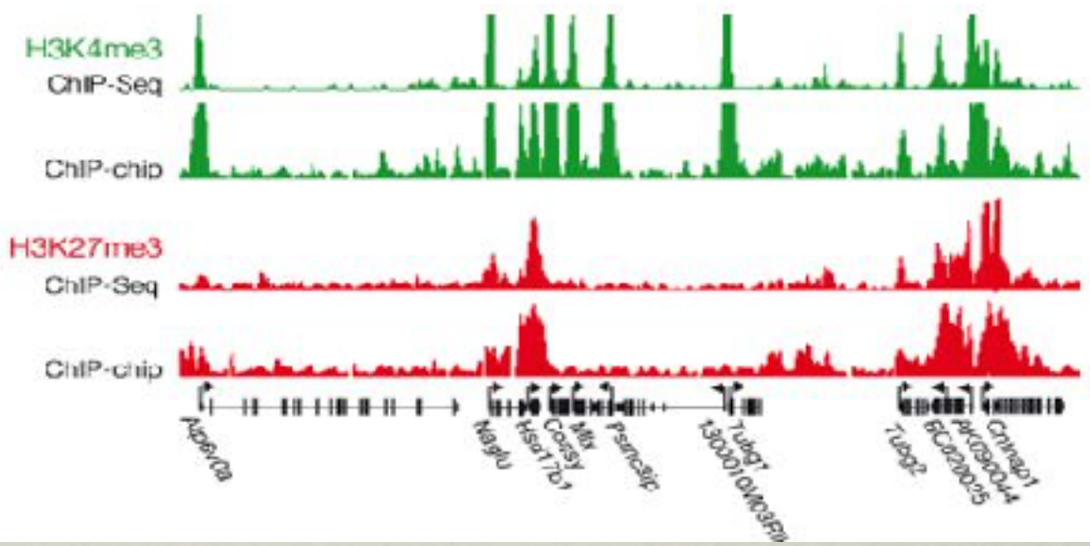
Release DNA

Prepare sequencing library

Sequence

Map sequence tags to genome & identify peaks





Adapted from slide set by: Stuart M. Brown, Ph.D.,
Center for Health Informatics & Bioinformatics, NYU School of Medicine

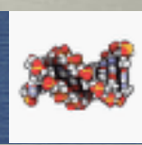
What can be done with this Technique

Can be use to interrogate ANY DNA-binding protein physically associated with a DNA segment on a genome wide basis.

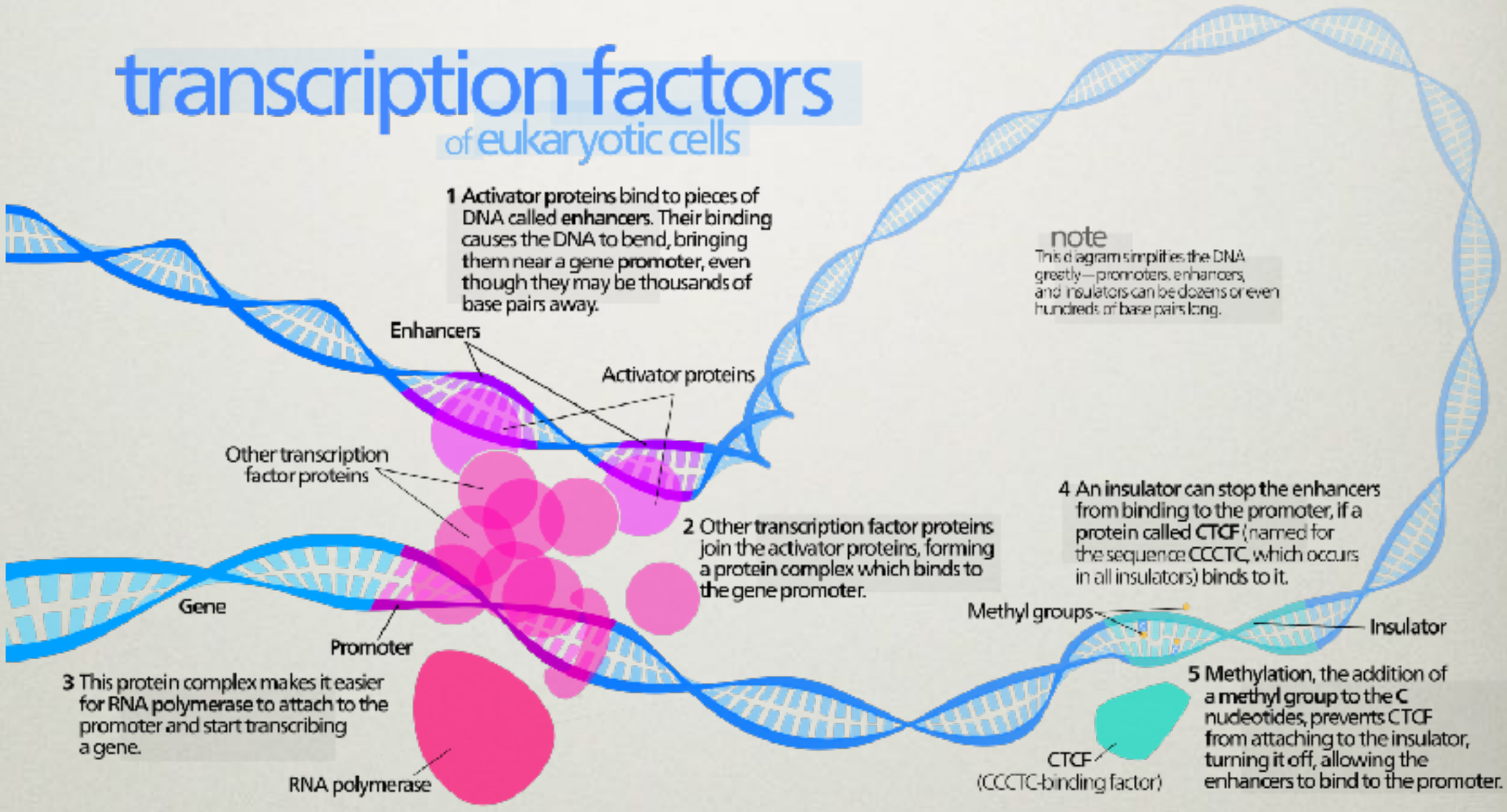
- Transcription factors (p53, STAT1)
- Basal transcription machinery (Pol II)
- Histones and modified histones (H3_ml4)
- Chromatin modifying enzymes (histone acetylase)

Transcription Factors

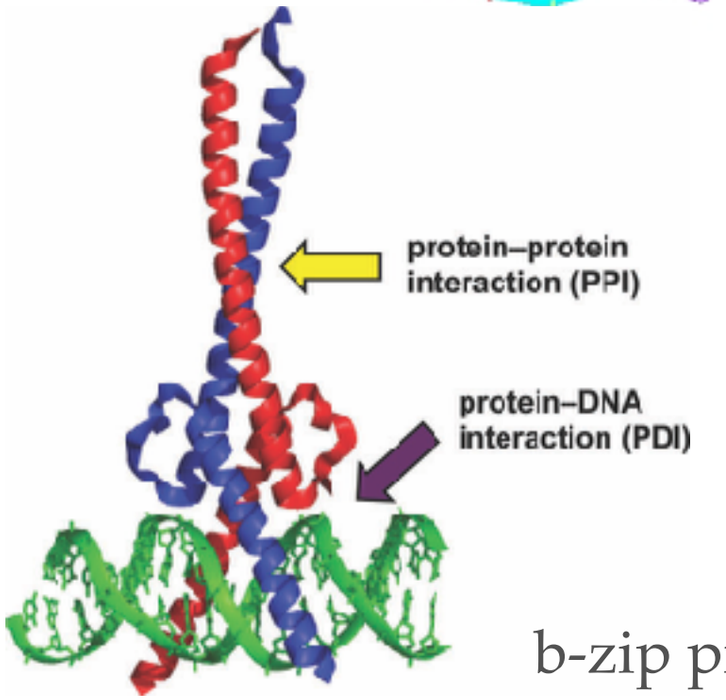
The first action of a transcription factor is to find and to bind DNA segments and ChIP-seq allows the **binding sites** of transcription factors to be identified across **entire genomes**. The **DNA sequence motif** that is recognized by the binding protein can be computed; the **precise regulatory sites in the genome** for any transcription factor can be identified; the direct **downstream targets** of any transcription factor can be determined; and the **clustering of transcription-regulatory** proteins at specific DNA sites can be assessed.



Transcription Factors

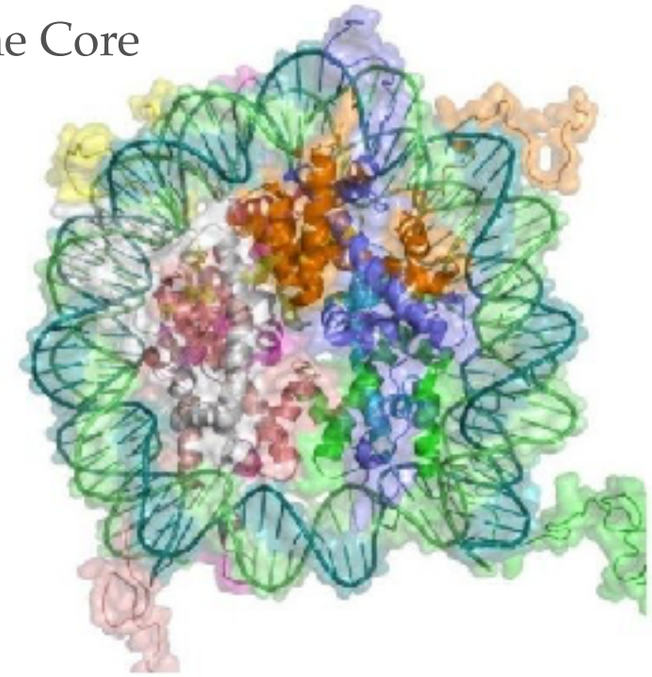


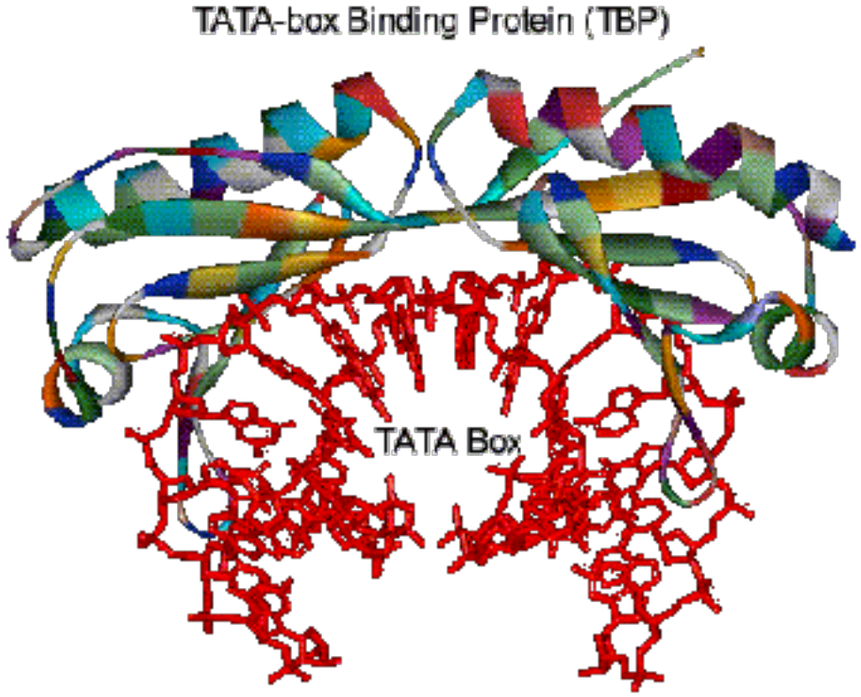
Transcription Factors



b-zip protein

Histone Core





Subset of Techniques

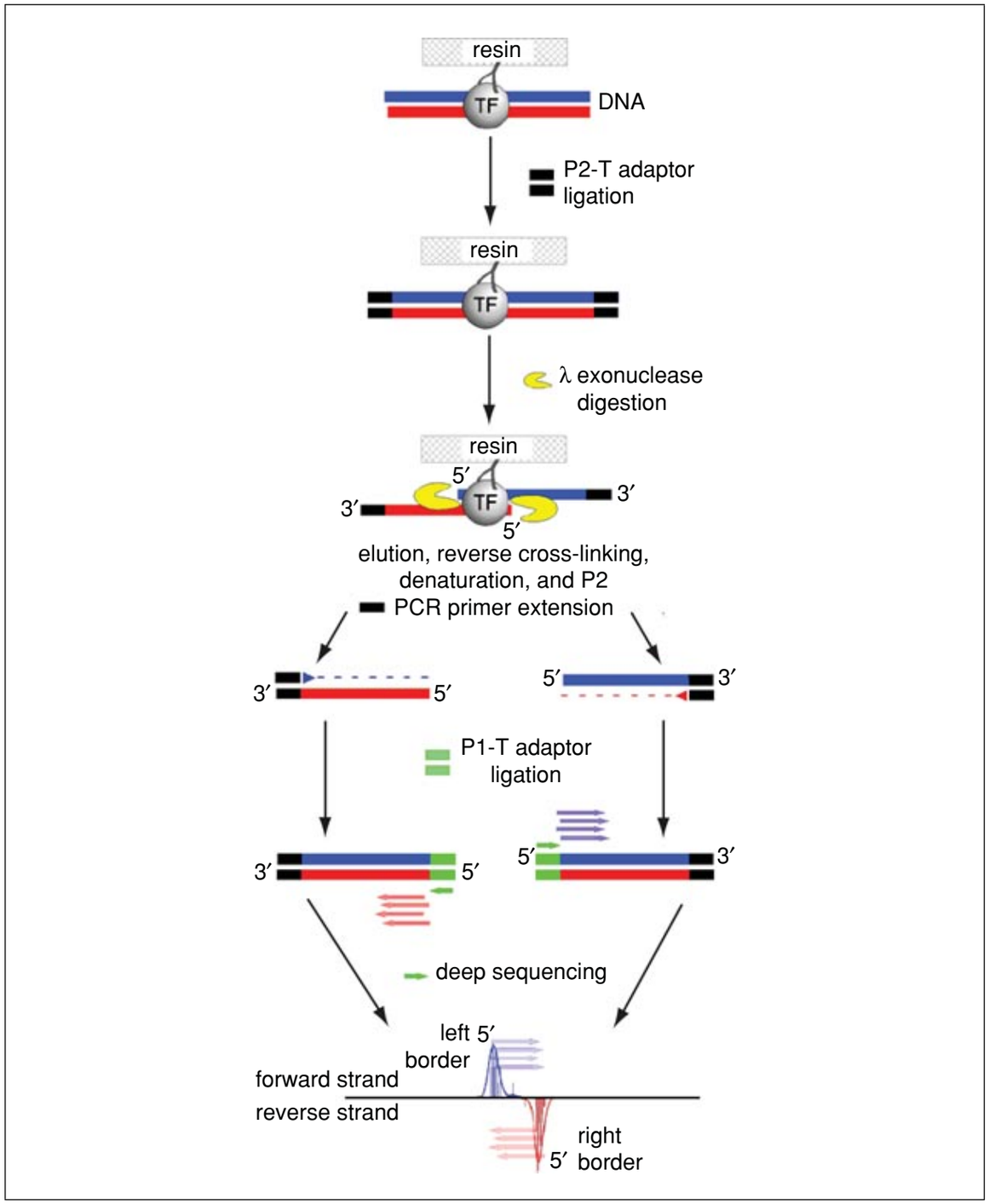
- ChIPSeq
- ChIPExo
- DNase Hypersensitivity
- DNase Footprinting
- ATAC-Seq
- MNase-Seq

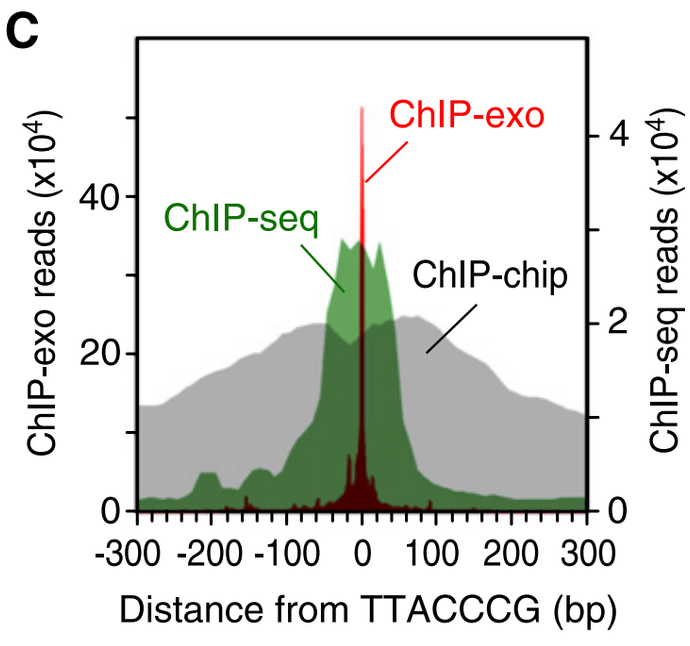
Different Variations

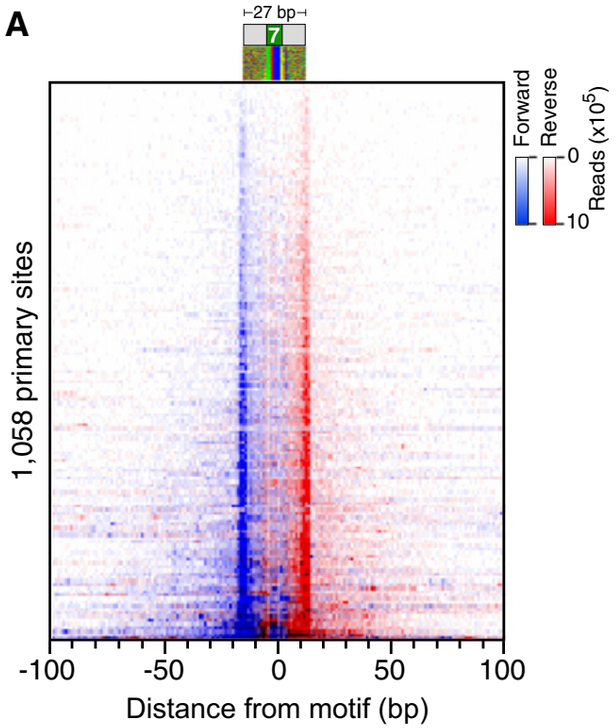
- Native ChIP (N-ChIP)
- Cross link protein and DNA (Formaldehyde) (X-ChIP)
- Protein-Protein cross linking (disuccinimidyl glutarate) and formaldehyde (HDAC- chromatin remodelers)

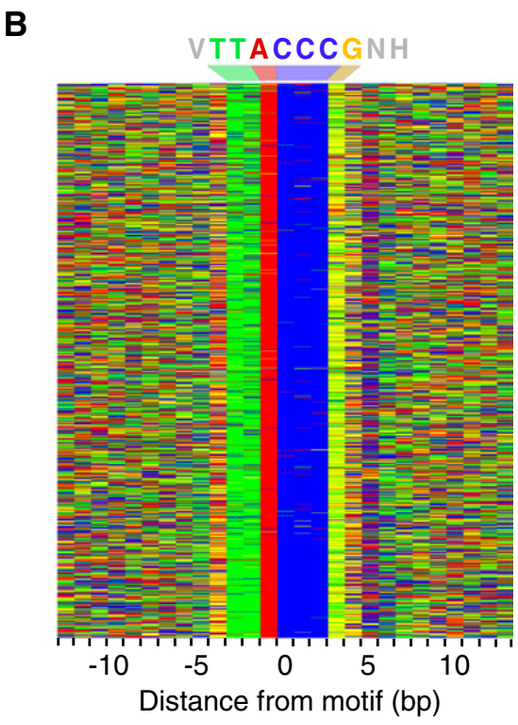
- Sonication (Fragmentation ...200-300bp)
- Enzymatic digestion (Micrococcal nuclease)
- Enzymatic digestion (DNAase)
- Enzymatic digestion (Exonuclease)

ChIP-exo


resin
TF
DNA
P2-T adaptor ligation
resin
TF
λ exonuclease digestion
resin
5′
TF
3′
3′
5′
elution, reverse cross-linking, denaturation, and P2 PCR primer extension
3′
5′
5′
3′
P1-T adaptor ligation
3′
5′
5′
3′
deep sequencing
left 5′ border
forward strand
reverse strand
right 5′ border

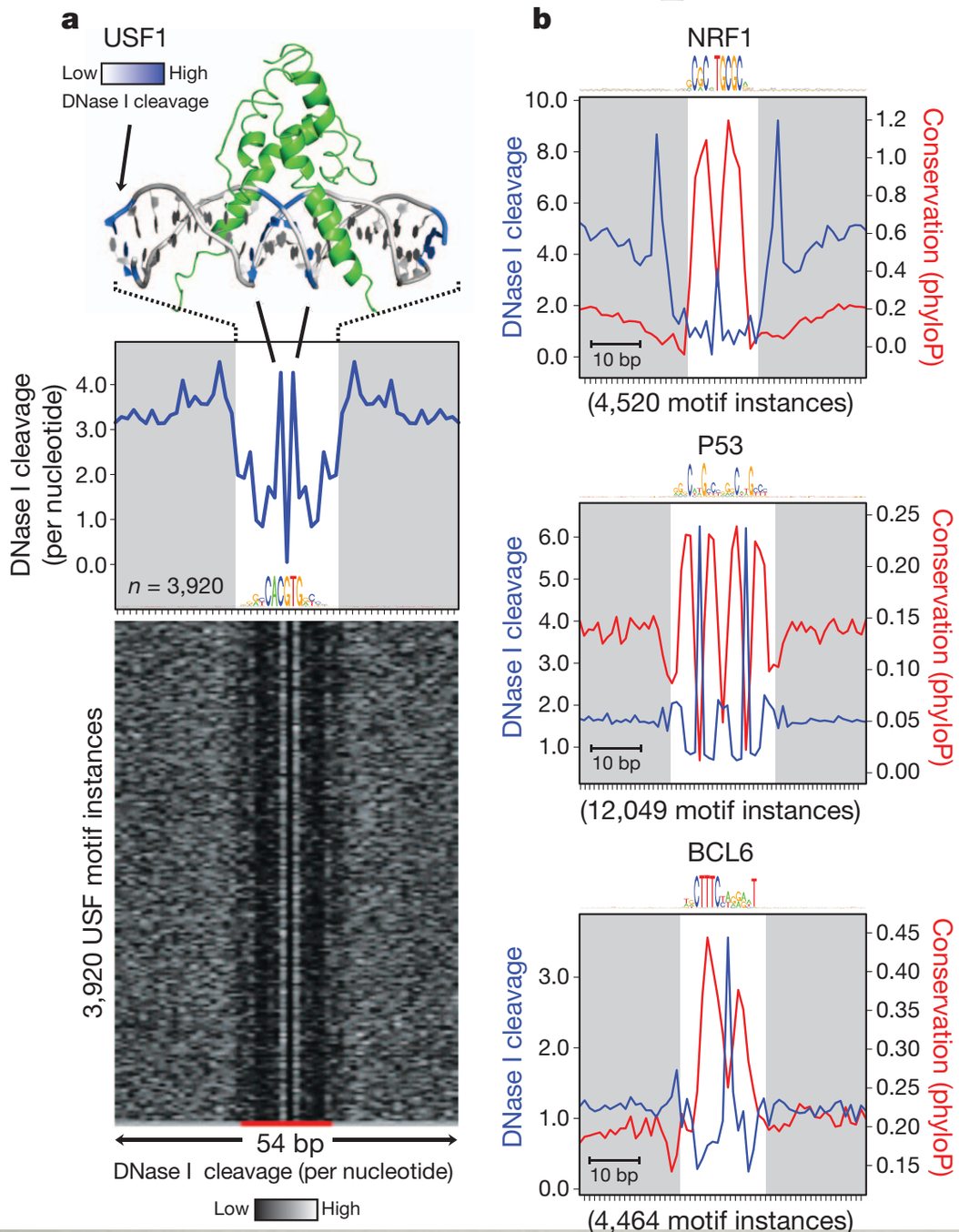

C
ChIP-exo
ChIP-seq
ChIP-chip
ChIP-exo reads (x10^4)
ChIP-seq reads (x10^4)
40
20
0
4
2
0
-300 -200 -100 0 100 200 300
Distance from TTACCCG (bp)


A
27 bp
7
Forward
Reverse
0
10
Reads (x10^5)
1,058 primary sites
-100 -50 0 50 100
Distance from motif (bp)

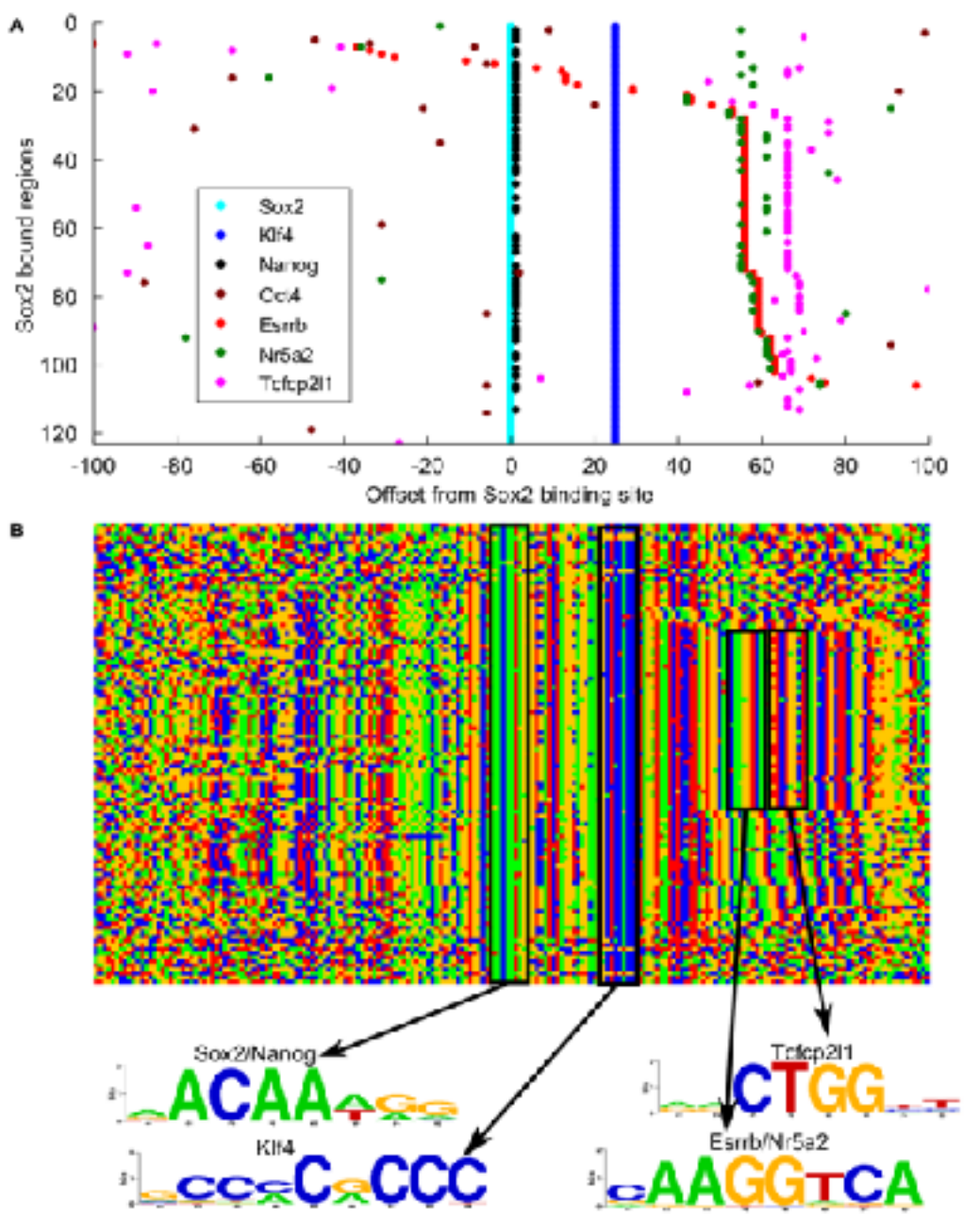

B
VTTACCCGNH
-10 -5 0 5 10
Distance from motif (bp)

DNase Footprint

DNase Footprint



Data Integration



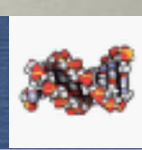
Whoops!

DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence

Myong-Hee Sung,[1,2] Michael J. Guertin,[1,2] Songjoon Baek,[1,2] and Gordon L. Hager[1,*]
[1]Laboratory of Receptor Biology and Gene Expression, National Cancer Institute, NIH, Building 41, 41 Library Drive, Bethesda, MD 20892, USA

Molecular Cell *56*, 275–285, October 23, 2014

Genomic footprinting has emerged as an unbiased discovery method for transcription factor (TF) occupancy at cognate DNA in vivo. A basic premise of footprinting is that sequence-specific TF-DNA interactions are associated with localized resistance to nucleases, leaving observable signatures of cleavage within accessible chromatin. This phenomenon is interpreted to imply protection of the critical nucleotides by the stably bound protein factor. However, this model conflicts with previous reports of many TFs exchanging with specific binding sites in living cells on a timescale of seconds. We show that TFs with short DNA residence times have no footprints at bound motif elements. Moreover, the nuclease cleavage profile within a footprint originates from the DNA sequence in the factor-binding site, rather than from the protein occupying specific nucleotides. These findings suggest a revised understanding of TF footprinting and reveal limitations in comprehensive reconstruction of the TF regulatory network using this approach.



STEPS

- Crosslinking
- Fragmentation
- ChIP (Immunoprecipitation)
- Library Construction
- Sequencing
- Alignment
- Peak Calling - Motif discovery - Visualization
- Correlating peaks/motifs with biology

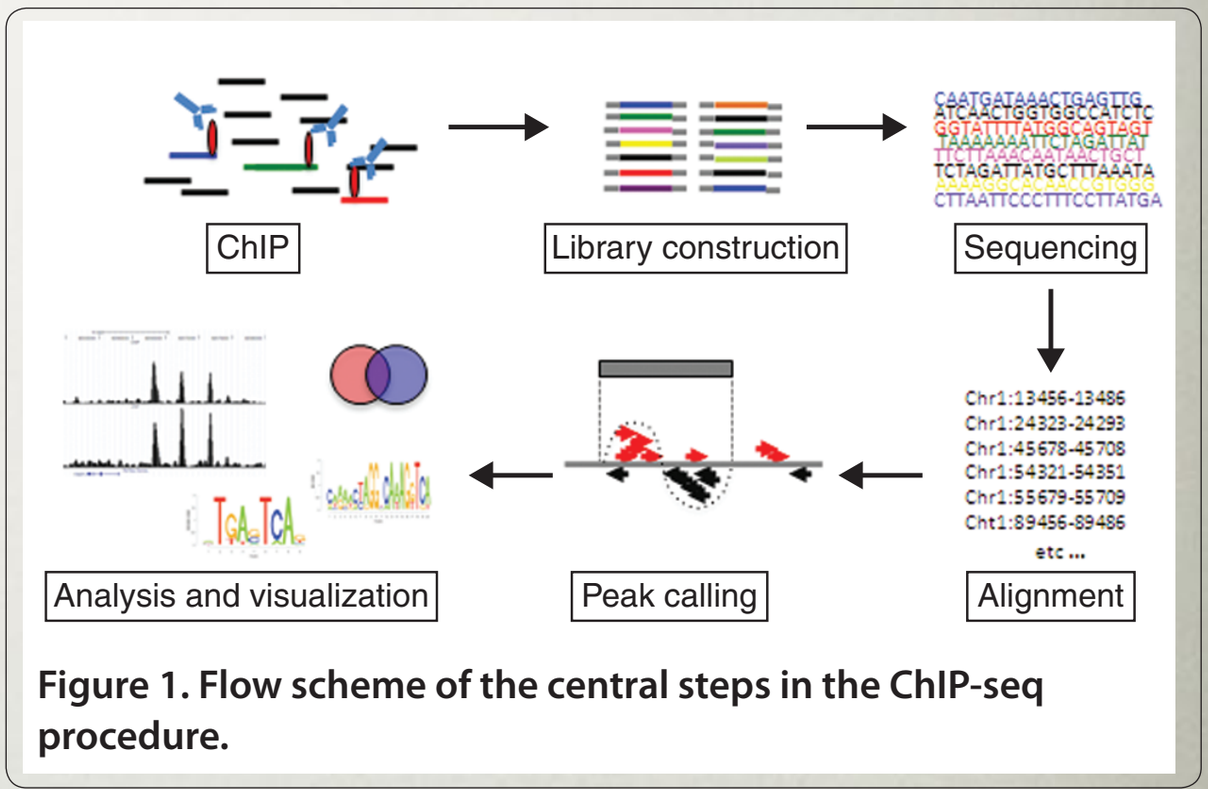


Figure 1. Flow scheme of the central steps in the ChIP-seq procedure.

On the surface ChIP-SEQ is a very simple straightforward technique with lots of potential...

Unfortunately, a number of technical and biological issues often make it a very challenging endeavor !

Confounding Factors

- Antibody "ChIP" efficiency
- Chromatin "fragile" sites
- Protein residence time
- Footprint of bound protein
- DNA repeat regions
- What is a real peak (signal to noise)

Comparison to ChIP-Chip

Comparison to ChIP-ChIP

- Nucleic acid hybridization is complex and is dependent on many factors including the GC-content, length, concentration, and secondary structure of both the target and probe sequences.

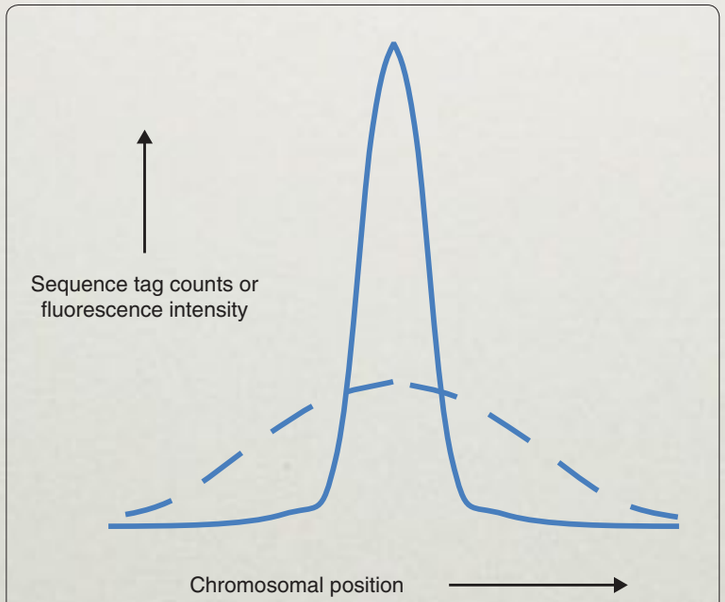


Figure 2. Comparison of ChIP-seq and ChIP-chip. Representative signals from ChIP-seq (solid line) and ChIP-chip (dashed line) show both greater dynamic range and higher resolution with ChIP-seq. Whereas three binding peaks are identified using ChIP-seq, only one broad peak is detected using ChIP-chip.

Comparison of ChIP-chip and ChIP-SEQ

	ChIP-chip	ChIP-Seq
Resolution	Array-specific, generally 30–100bp	Single nucleotide
Coverage	Limited by sequences on the array; repetitive regions usually masked out	Limited only by alignability of reads to the genome; increases with read length; many repetitive regions can be covered
Cost	$400–$800 per array (1–6 million probes); multiple arrays may be needed for large genomes	$1000–$2000 per Illumina lane (6–15 million reads prior to alignment)
Source of platform noise	Cross-hybridization between probes and non-specific targets	Some GC-bias may be present
Experimental design	Single- or double-channel, depending on platform	Single channel
Cost-effective cases	Large fraction enriched (broad binding), profiling of selected regions	Small fraction enriched (sharp binding), large genomes
Required amount of ChIP DNA	High (few μg)	Low (10–50 ng)
Dynamic range	Lower detection limit, saturation at high signal	Not limited
Amplification	More required	Less required; single molecule sequencing without amplification is available
Multiplexing	Not possible	Possible

Experimental Design

ChIP-SEQ
Before you Start

- Do you need really need to do the experiment ?
  - Is there existing data ?
  - Is there similar data...same factor different conditions/cell type/organism
  - Is there similar data...different but similar factor
- Do you have a plan on how to analyze the data.

ChIP-SEQ Design Issues

- Antibody Selection
  - *Probably the most critical experiment decision*
- DNA Control
- Depth of Sequencing (How many reads)
- Replicates
- Experimental Goals (Positive control)
- Algorithm choices - mapping and peak-calling

Its all about the antibody

- Must have specificity for target molecule
- Must immunoprecipitate the target
  (Must ChIP well!)
- Do you have Quality control metric to access the quality of your antibody (don't rely on vendor)
  (Western blots, Chip PCR)

Its all about the antibody

“Having a third party validate every batch would be a fabulous thing,” says Peter Park, a computational biologist at Harvard Medical School.

He notes that the consortium behind ENCODE — a project aimed at identifying all the functional elements in the human genome — tested more than 200 antibodies targeting modifications to proteins called histones and found that more than 25% failed to target the advertised modification.

BIOMEDICINE: NATURE NEWS I 1 AUGUST 2013 I VOL 500 I NATURE I 15

Control

Its alway best to have one!

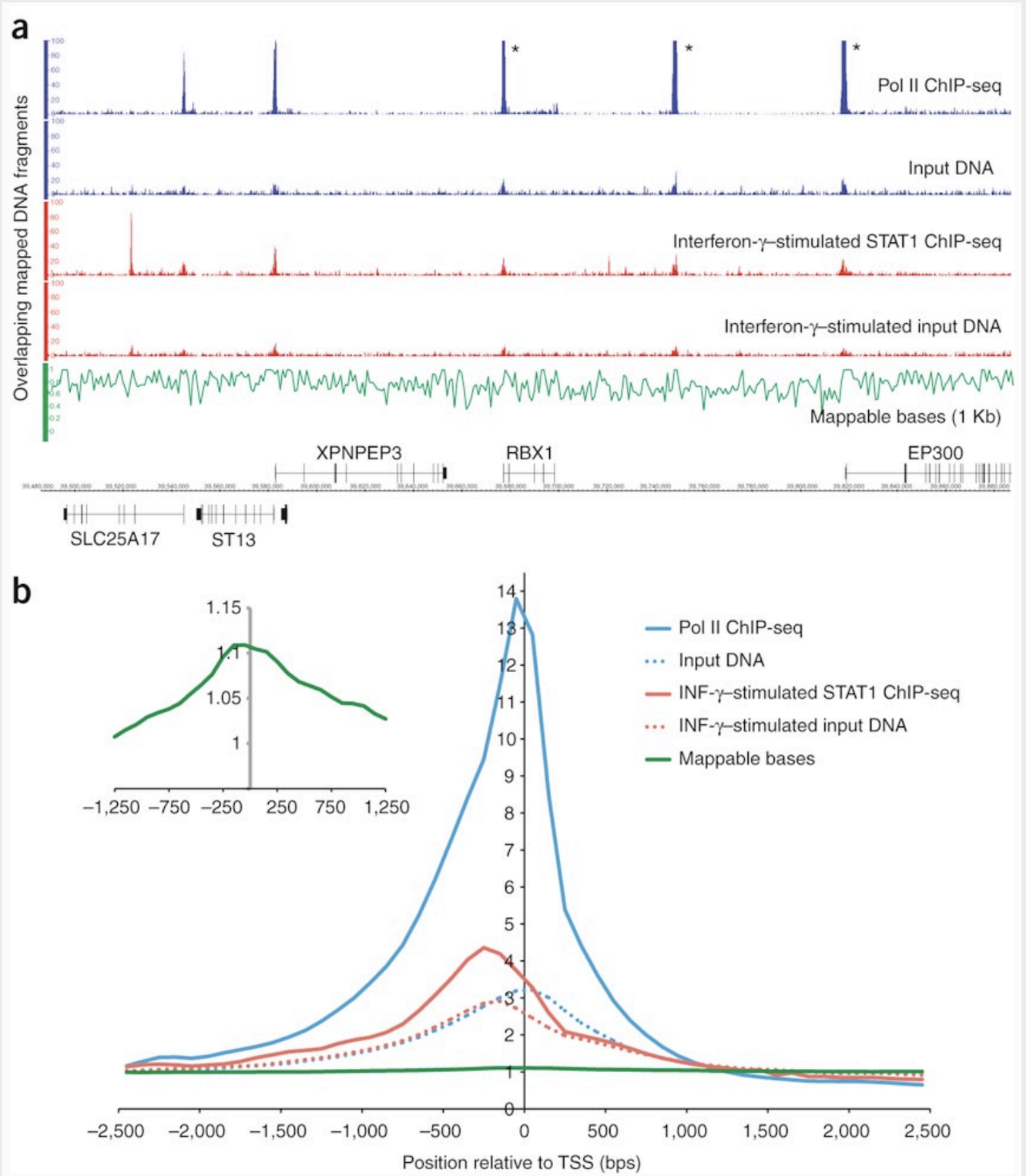
There are three commonly used choices for this control:

- input DNA (that is, DNA prior to immunoprecipitation, IP) [solubility, shearing, amplification]
- mock IP (treated the same as the IP but without any antibody) [low level of pull down DNA]
- non-specific IP (that is, using an antibody against a protein not known to be involved in DNA binding or chromatin modification, such as IgG).
  [low level of pull down DNA]

No consensus although most use input DNA... control not necessarily needed for differential binding experiments

Why you need a control

- Preferential sequencing of G+C rich regions
- Repeat regions
- Genomic Amplifications
- Genomic Landmarks (TSS) higher than normal in control
- Chromatin structure - shearing is different: euchromatin vs heterochromatin, active vs silenced genes
- PCR biased amplification (remove identical reads)



Rozowsky, Nature Biotechnology, 2009

Sequencing

These days it is almost always Illumina Sequencing

- HiSeq
- MiSeq
- NextSeq

Paired-end vs single end reads

- Increased mappability - especially in repeat region
- Double the costs

  Usually not worth the extra cost, except for special circumstances

Sequencing

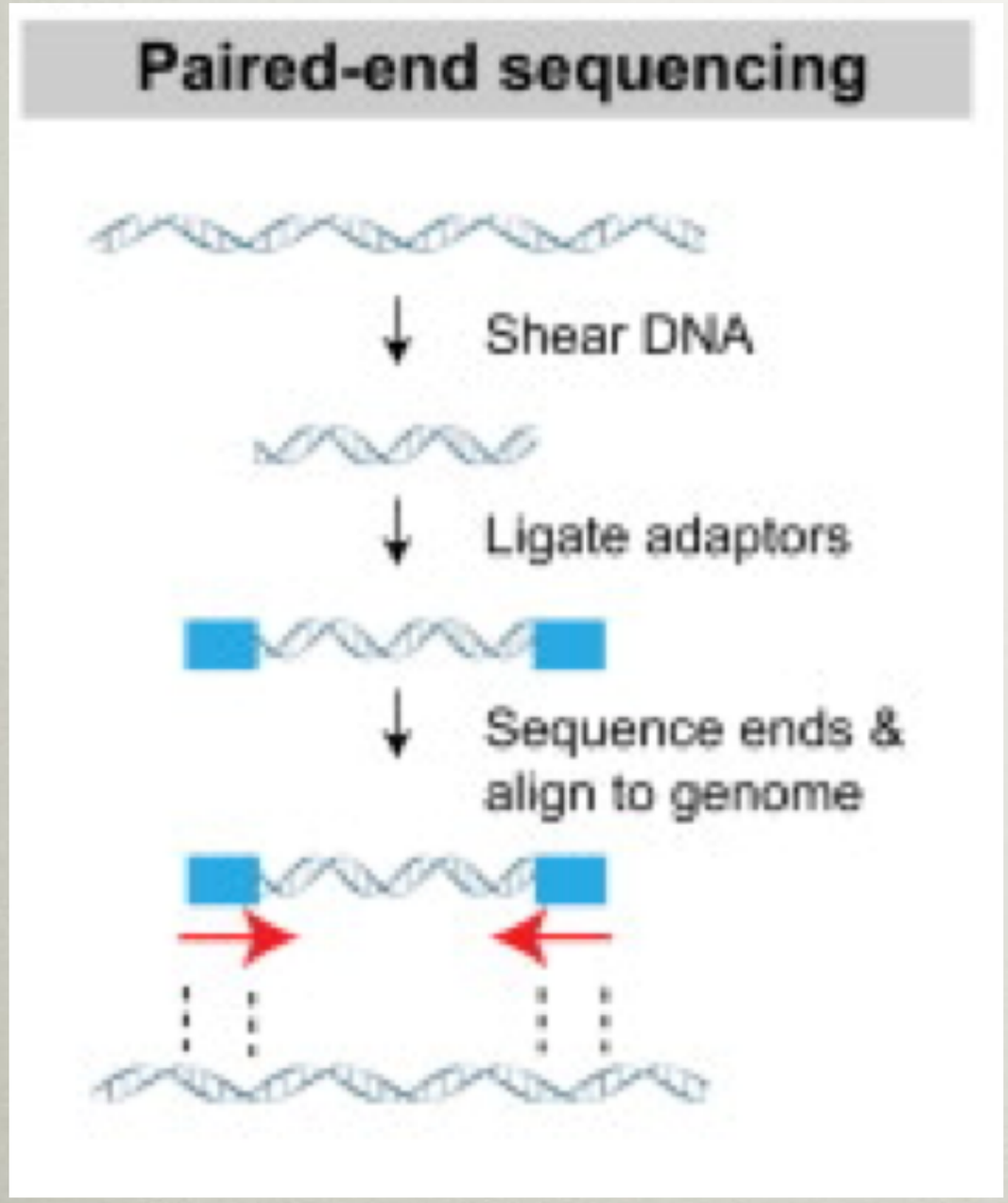
How many reads and how long ?

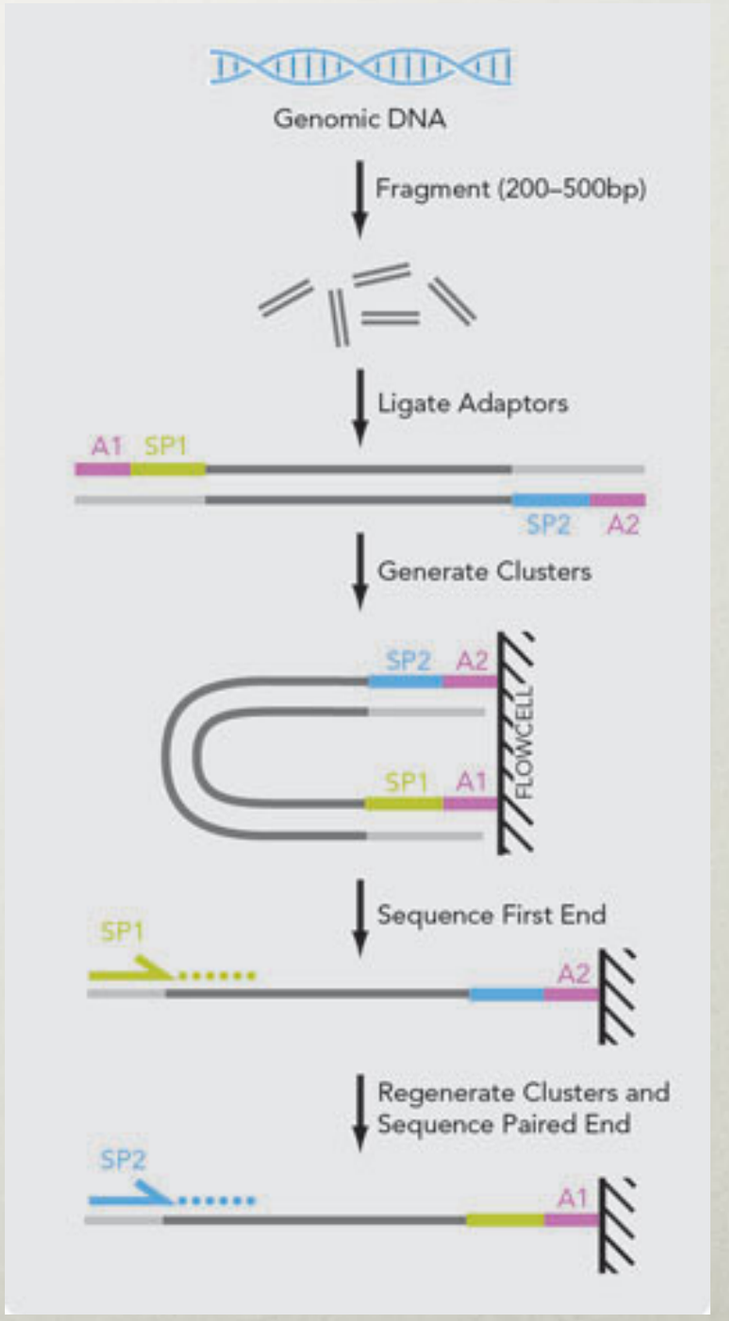
Normally short reads (36bp) are sufficient

Human - Sharp peak≅20M - Broad peak≅40M
high frequency elements (nucleosomes) need more.

- Prominent peaks are identified with fewer reads, while weaker peaks require more reads.
- The number of putative target regions (peaks) increases as a function of read depth...may not plateau.

Sequencing


Paired-end sequencing
Shear DNA
Ligate adaptors
Sequence ends &
align to genome


Genomic DNA
Fragment (200–500bp)
Ligate Adaptors
A1
SP1
SP2
A2
Generate Clusters
SP2
A2
FLOWCELL
SP1
A1
Sequence First End
SP1
A2
Regenerate Clusters and
Sequence Paired End
SP2
A1

ChIP-SEQ Technical Problem Points

- How many reads are needed
  - How much binding
- Saturation
- Hi vs low affinity sites
- Minimal enrichment saturation ratio, MSER
- **Can one library be compared quantitatively with another on a site-by-site basis?**

Sequencing Biases

- Linker ligation
- PCR amplification
- Hydridization

→ Platform Biases

Can be controlled for (to some degree) by using input DNA to identify and correct for "sequencing biases".

Replicates

Having replicates is **ALWAYS** good, and many times its essential.

In general Biological replicates are more useful than technical replicates.

The need for replicates and the appropriate number is largely dependent on experimental goals (general or specific) and the quality of the data (which may have its basis in biology rather than technique).

Experimental Goals

- Make sure your experimental design is appropriate to meet your desired goals.

- Talk to the people who are going to analyze the data **BEFORE** you do the experiment.

Snapshot of ENCODE Recommendations

Really good antibody to start with!

EXPERIMENTAL DESIGN GUIDELINES

- At least 2 replicates
- Input Control for each condition
- Reproducibility
- Library complexity
- Adequate Sequencing depth to capture events across genome

DATA QUALITY ASSESSMENT

- Metrics at every stage possible to assess quality of experiment
- Cross-correlation for stranded reads
- Irreproducible Discovery Rate (IDR) for peak concordance in replicates

DATA REPORTING GUIDELINES

- Minimal Information for Chip-seq Experiment (MICE)
- Analysis Details
- High-throughput sequencing data

Data Analysis

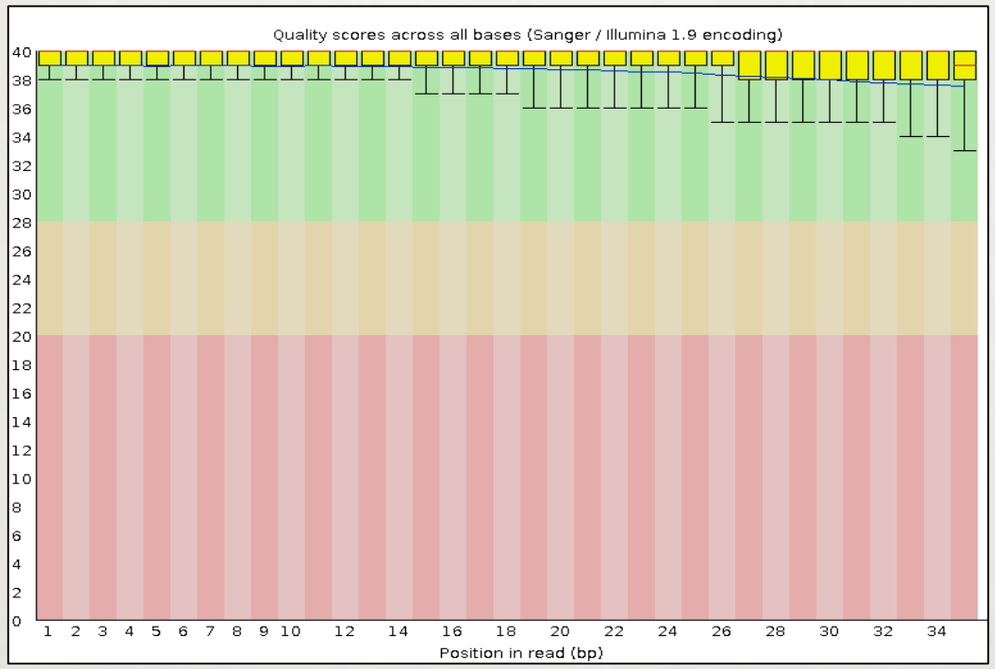
Analysis Pipeline

- Sequencing Quality Control
- Mapping
- Peak finders
- Visualizers
- Motif finders
- GSEA
- Pathway analysis
- Differential effects

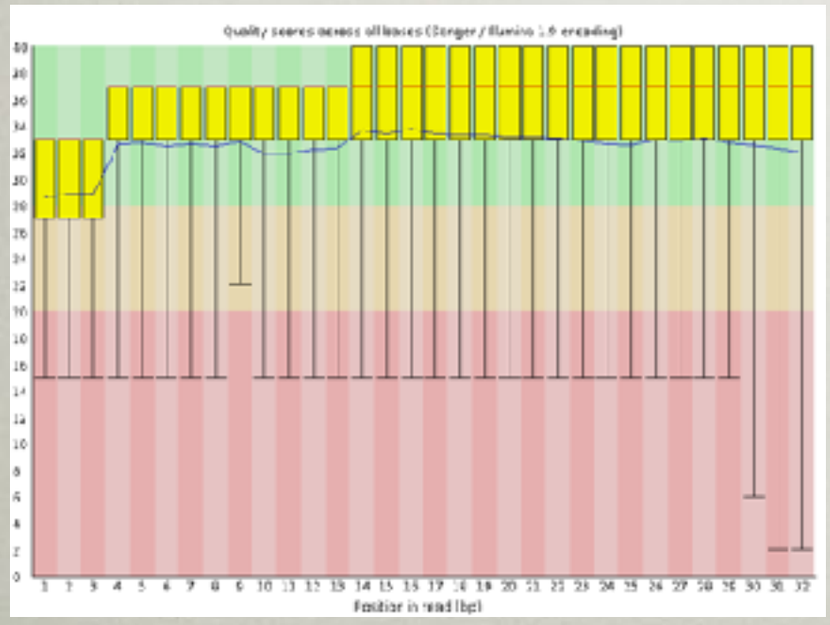
Which program / method you use at each step will be influenced by many factors

Read Quality (FastQC)

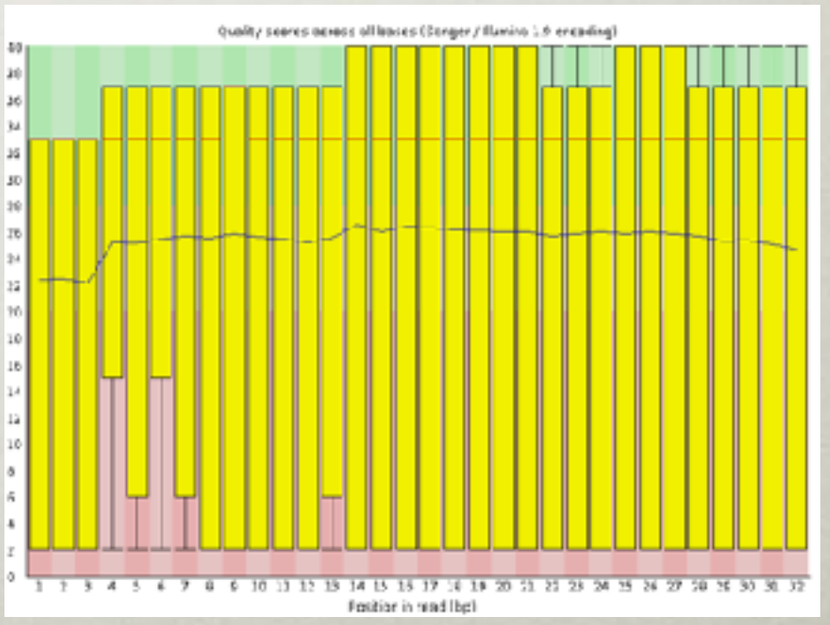
Great!



Okay



Bad!



Good quality **sequence** data does not mean a successful ChIPSeq experiment.

Mapping

Mapping
Which Genome Version?

- Which version of the genome do you **want/need** to use. (*Record and report it!!*)

Considerations

- Genome annotation
- Parallel experiments
- Experiments you want to compare it too.
- Available browsers

Mapping Bias

Not all the genome is "*available*" for mapping

		Nonrepetitive sequence		Mappable sequence	
Organism	**Genome size (Mb)**	**Size (Mb)**	**Percentage**	**Size (Mb)**	**Percentage**
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens	3,080.44	1,462.69	47.5%	2,451.96	79.6%

*Calculated based on 30nt sequence tags

Rozowsky, 2009

Mapping Bias

- Effects of repetitive DNA
- Length of reads
- Many choices of mappers
- How important is the mapper you use ?

- Bowtie
- BWA
- BFAST
- Novoalign
- ELAND
- STAR

Mapping

Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

Aligner less critical than some for other NGS applications... most important is how they handle repeat regions and PCR amplification products and mismatches (indels)

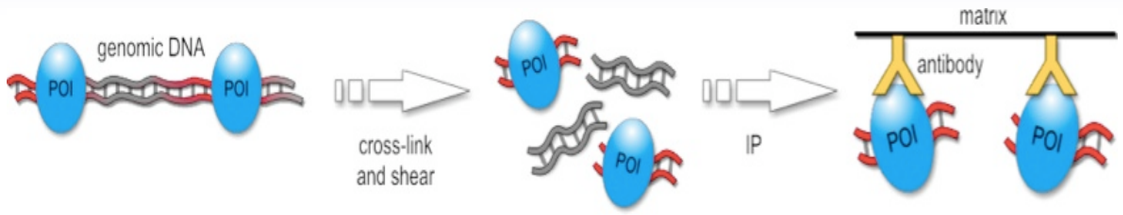
Peak-Calling

Good data is always more robust to analytical choices than poor data.

Peak Calling

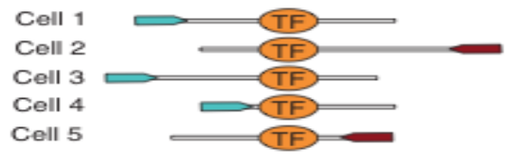
What is the ultimate goal of peak calling?

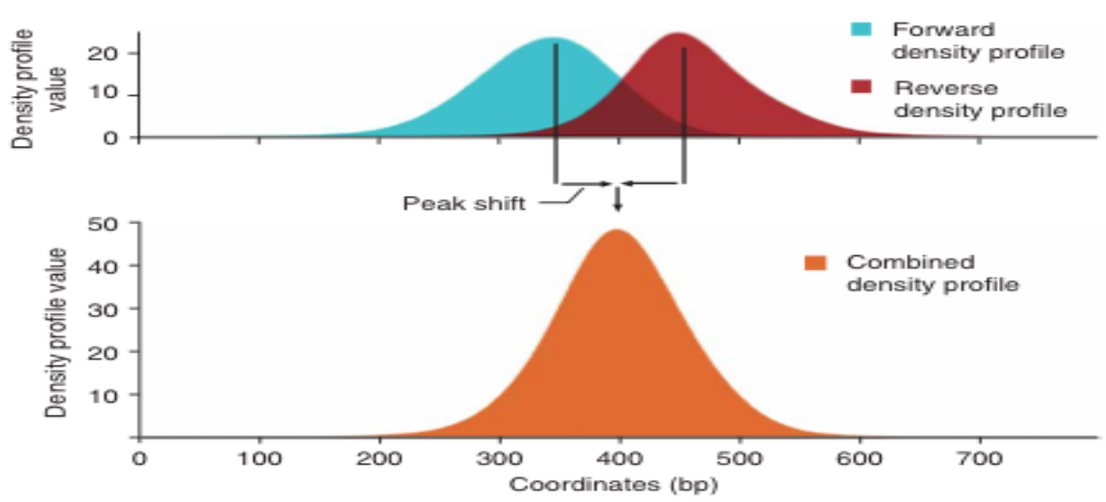
It is to determine if and where there is enrichment compare to a control


genomic DNA
POI
POI
cross-link
and shear
POI
POI
IP
matrix
antibody
POI
POI

⇓

ChIP-Seq


Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
TF
TF
TF
TF
TF


Density profile value
20
10
0
Forward density profile
Reverse density profile
Peak shift
50
40
30
20
10
Density profile value
Combined density profile
0
100
200
300
400
500
600
700
Coordinates (bp)

Peak Calling

- Read Shifting
- Background estimation (uses control)
- Artifact removal
- Significance cutoff (FDR)
- *Multiple Programs with differing ability*
- *No consensus*
- *Often effected by parameter selection*

Types of Peaks

Each has its own challenges

- Narrow (Sharp)
- Broad
- Nucleosome like

Types of Peaks

Peaks have different shapes (characteristic of the protein?) and each presents its own challenges

Sharp

Mixed

Medium

Broad

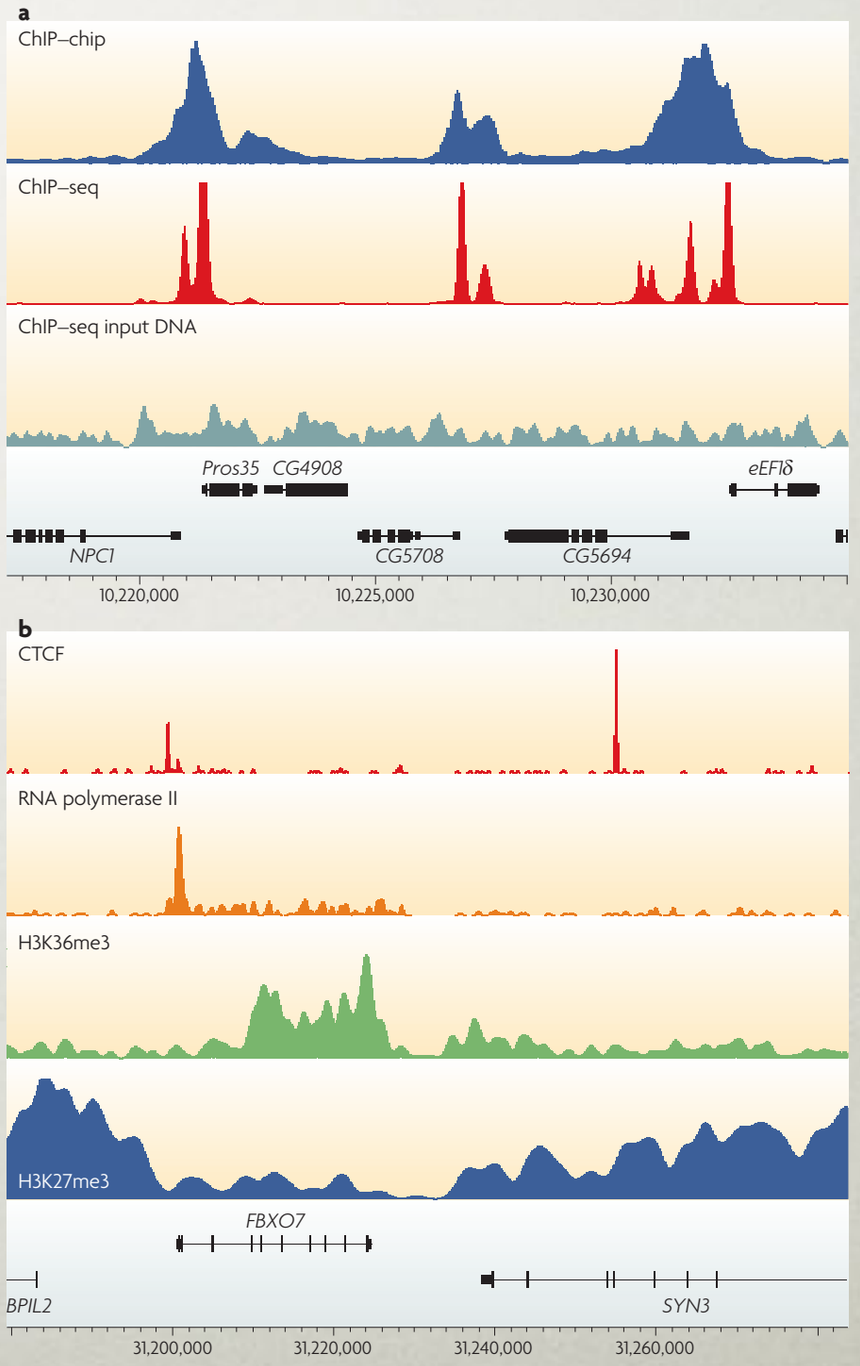
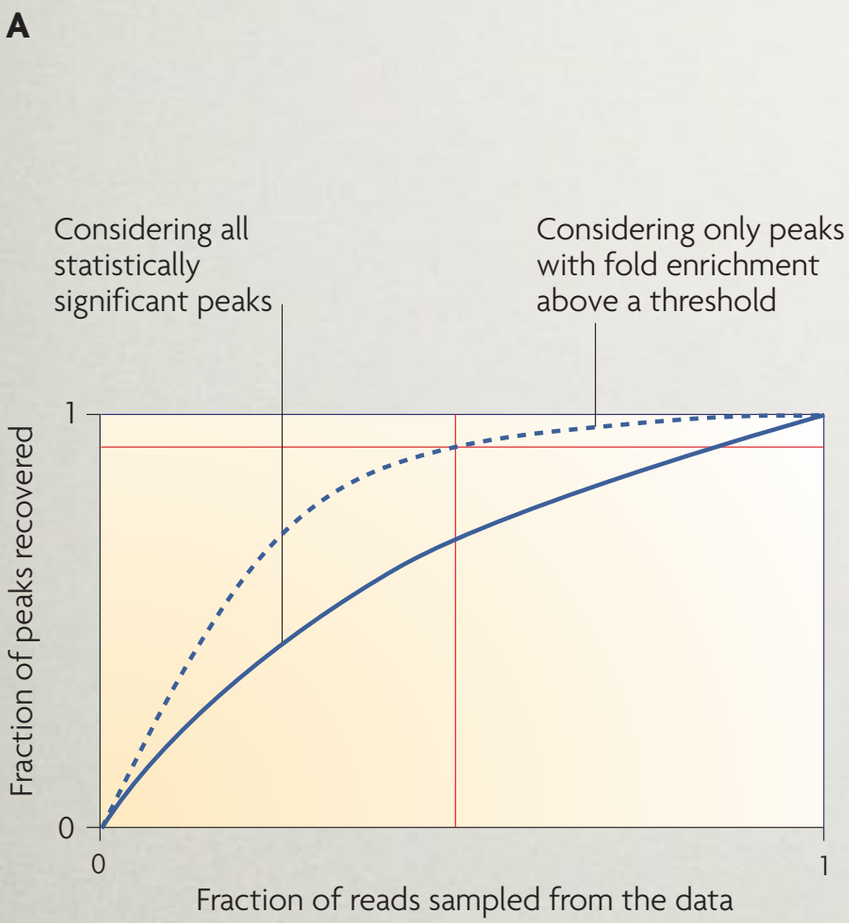
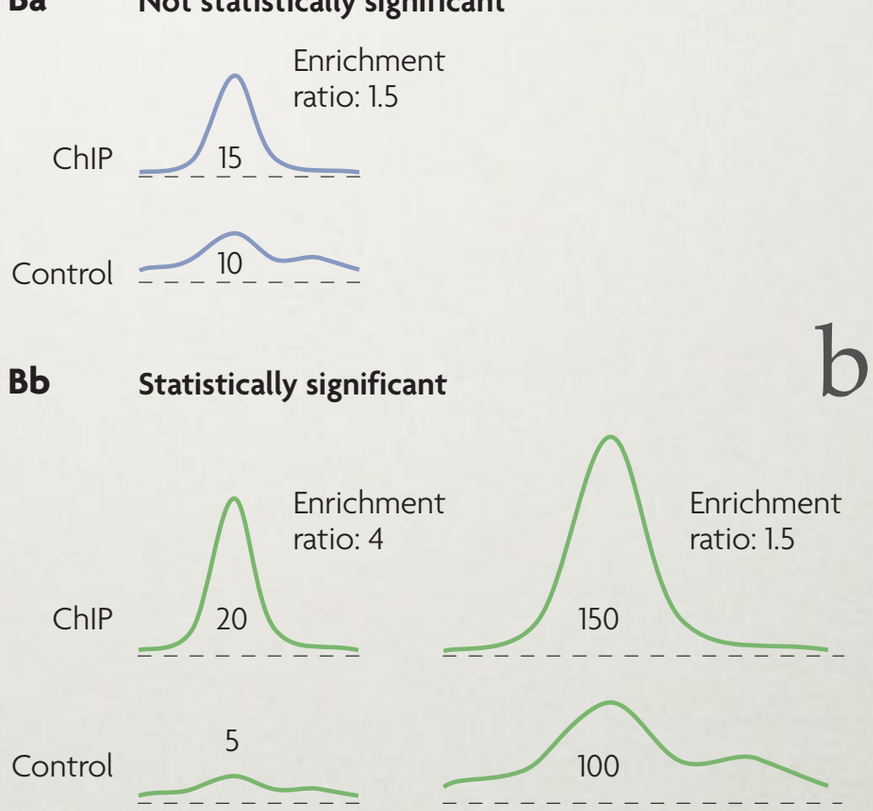


Figure 2 | chiP profiles. a | Examples of the profiles generated byNcahtruorme Raetivnieiwmsml uGneonpetriec-s cipitation followed by sequencing (ChIP–seq) or by microarray (ChIP–chip). Shown is a section of the binding profiles of the chromodomain protein Chromator, as measured by ChIP–chip (unlogged intensity ratio; blue) and ChIP–seq (tag density; red) in the Drosophila melanogaster S2 cell line. The tag density profile obtained by ChIP–seq reveals specific positions of Chromator binding with higher spatial resolution and sensitivity. The ChIP–seq input DNA (control experiment) tag density is shown in grey for comparison. b | Examples of different types of ChIP–seq tag density profiles in human T cells. Profiles for different types of proteins and histone marks can have different types of features, such as: sharp binding sites, as shown for the insulator binding protein CTCF (CCCTC-binding factor; red); a mixture of shapes, as shown for RNA polymerase II (orange), which has a sharp peak followed by a broad region of enrichment; medium size broad peaks, as shown for histone H3 trimethylated at lysine 36 (H3K36me3; green), which is associated with transcription elongation over the gene; or large domains, as shown for histone H3 trimethylated at lysine 27 (H3K27me3; blue), which is a repressive mark that is indicative of Polycomb-mediated silencing. BPIL2, bactericidal/permeability-increasing protein-like 2; FBXO7, F box only 7; NPC1, Niemann-Pick disease, type C1; Pros35, proteasome 35 kDa subunit; SYN3, synapsin III. Data for part b are from Ref. 25.

Types of Peaks





But is it biologically relevant

Figure 3 | Depth of sequencing. A | To determine whether enough tags have been sequenced, a simulation can be carried out to characterize the fraction of the peaks that would be recovered if a smaller number of tags had been sequenced. In many cases, new statistically significant peaks are discovered at a steady rate with an increasing number of tags (solid curve) — that is, there is no saturation of binding sites. However, when a minimum threshold is imposed for the enrichment ratio between chromatin immunoprecipitation (ChIP) and input DNA peaks, the rate at which new peaks are discovered slows down (dashed curve) — that is, saturation of detected binding sites can occur when only sufficiently prominent binding positions are considered. For a given data set, multiple curves corresponding to different thresholds can be examined to identify the threshold at which the curve becomes sufficiently flat to meet the desired saturation criteria (defined by the intersection of the orange lines on the graph). We refer to such a threshold as the minimum saturation enrichment ratio (MSER). The MSER can serve as a measure for the depth of sequencing achieved in a data set: a high MSER, for example, might indicate that the data set was undersampled, as only the more prominent peaks were saturated (see Ref. 48 for details). Ba | A peak that is not statistically significant — the enrichment ratio between the ChIP and control experiments is low (1.5) and the number of tag counts (shown under the peaks) is also low. Bb | Two ways in which a peak can be statistically significant. On the left, although the number of tag counts is low, the enrichment ratio between the ChIP and control experiments is high (4). On the right, the peaks have the same enrichment ratio as those in a but have a larger number of tag counts; this example shows that continued sequencing might lead to less prominent peaks becoming statistically significant and that there might not necessarily be a saturation point after which no further binding sites are discovered.

Different Peak Callers

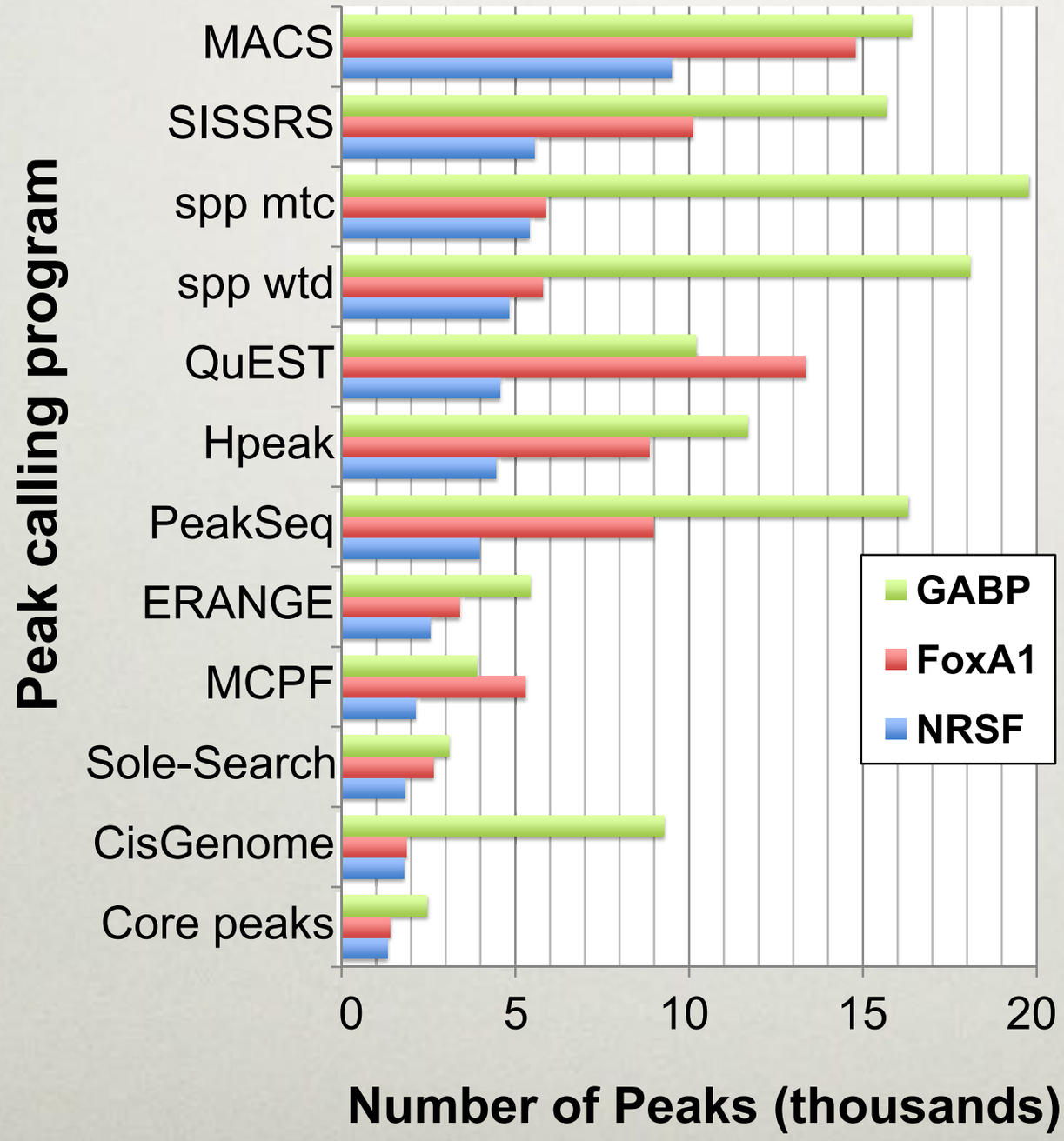
Peak Calling Bias

Potentially the most critical, especially for "poor quality experiments"

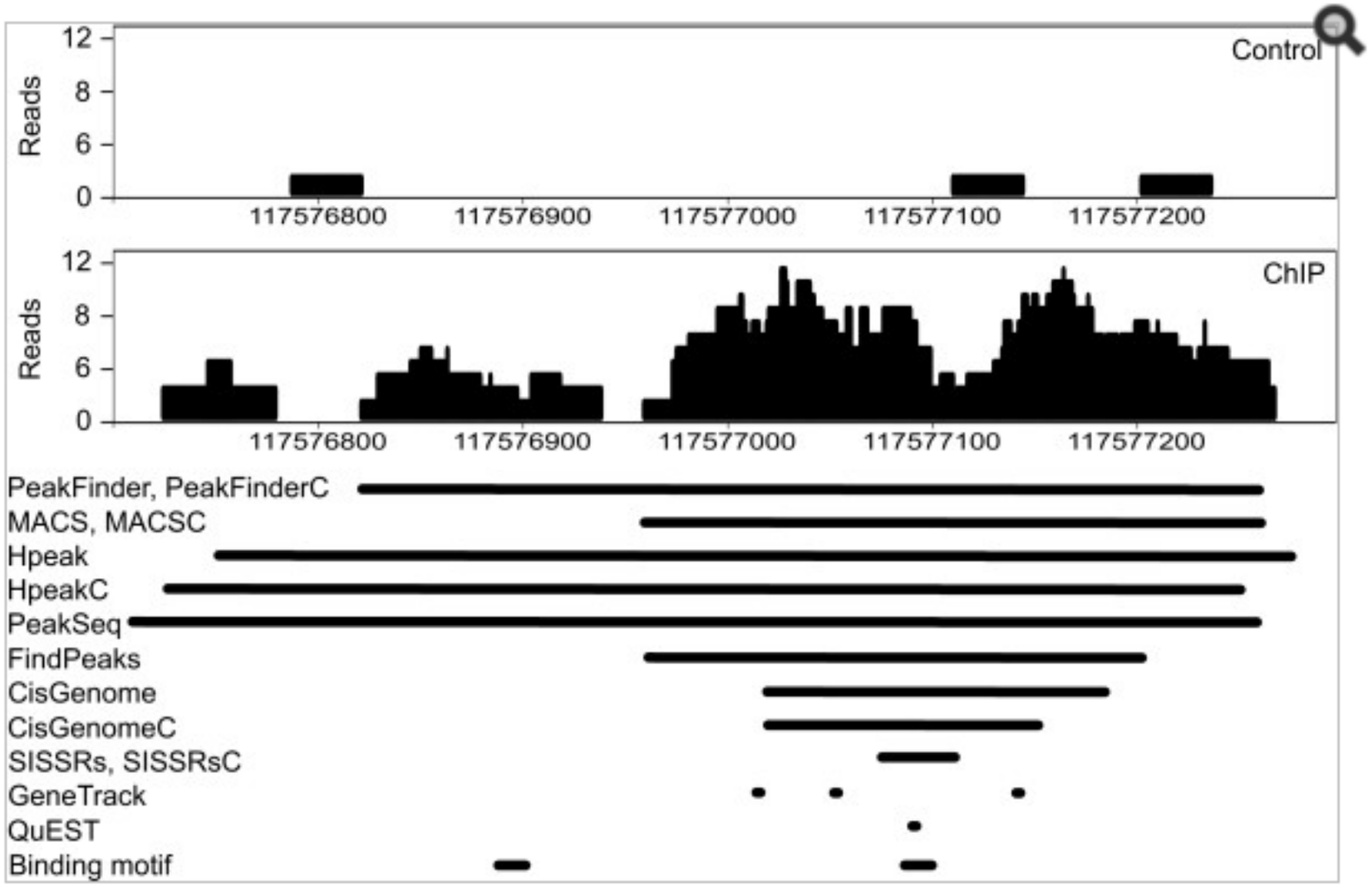
- **MACS**
- SISSRs
- **PeakSeq**
- **SICER**
- Useq
- CisGenome
- CCAT
- **SPP**
- NGSA

Different models, call different numbers of peaks, different sized peaks, optimized for different shaped peaks

Peak Calling Bias



Peak Calling



http://encodeproject.org/ENCODE/encodeTools.html

ChIP-seq Peak Callers

MACS
A widely-used, fast, robust ChIP-seq peak-finding algorithm that accounts for the offset in forward-strand and reverse-strand reads to improve resolution and uses a dynamic **Poisson distribution** to effectively capture local biases in the genome. MACS 1.4 is being used for the current uniform peak calling pipeline.
Feng J, Liu T, Zhang Y. Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinformatics. 2011 Jun;Chapter 2:Unit 2.14.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

PeakSeq
Identifies enriched regions in ChIP-seq type experiments and explicitly compares signal experiments to control experiments.
Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009 Jan;27(1):66-75.

SPP
A ChIP-seq peak calling algorithm, implemented as an **R package**, that accounts for the offset in forward-strand and reverse-strand reads to improve resolution, compares enrichment in signal to background or control experiments, and can also estimate whether the available number of reads is sufficient to achieve saturation, meaning that additional reads would not allow identification of additional peaks.
Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008 Dec; 26(12):1351-9.

MACS

(Peak Calling)

Model-based Analysis of ChIP-Seq MACS

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# and X Shirley Liu*

Genome ***Biology*** 2008, **9:R137** (doi:10.1186/gb-2008-9-9-r137)

We present Model-based Analysis of ChIP-Seq data, MACS, which analyzes data generated by short read sequencers such as Solexa's Genome Analyzer. MACS empirically models the shift size of ChIP-Seq tags, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome, allowing for more robust predictions. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, and is freely available.

Peak Callers - MACS

MACS is (for Transcription Factor binding) one of the most popular peak callers, it is also one of the oldest and this probably contributes to its success. It is a good method, good enough for many experimental conditions and requires very little justification if cited as the tool used in a publication. MACS performs removal of redundant reads, read-shifting to account for the offset in forward or reverse strand reads. It uses control samples and local statistics to minimize bias and calculates an empirical FDR.

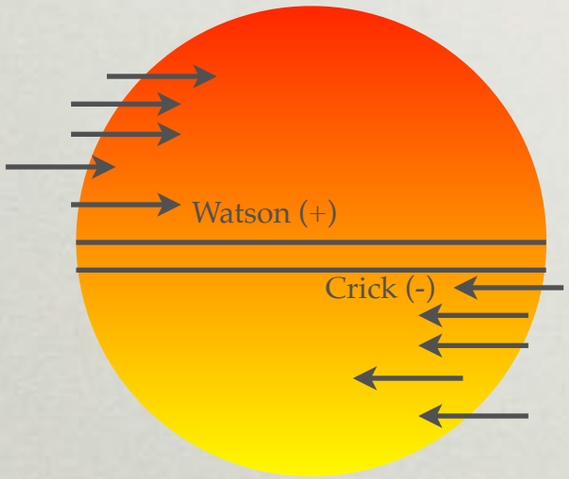
Model-based Analysis of ChIP-Seq MACS

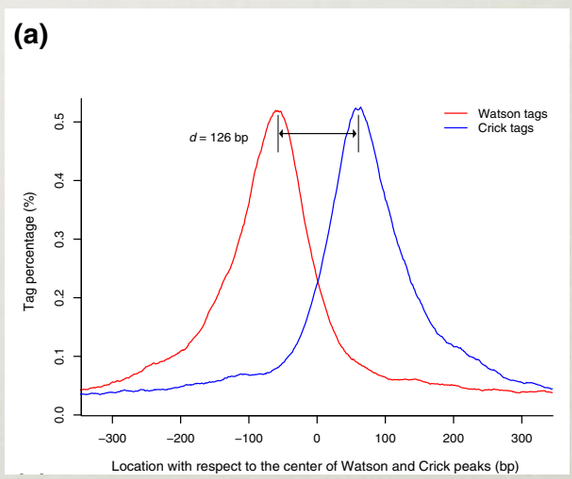
- Most widely used
- Robust, provided your data fits the model
- Ignores PCR artifacts
- Does NOT do much QC for you
  (*garbage in garbage out*)
- Python based - many dependencies
- Availability:Helix/Biowulf, Genomatix and Galaxy
- Two common versions (1.4.2 and 2.0.10)

MACS
READ SHIFTING

• MACS takes advantage of the expected bimodal distribution pattern to empirically model the shifting size to better locate the precise binding sites.

• 1000 high quality peaks where > mfold-enrichment relative to random tag distribution





• Define distance d, and shifts all tags d/2 distance towards the 3’ end

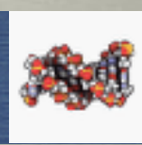
MACS
Peak Detection

- Linearly scales the total control tag count to the same and the ChIP tag count
- Removes duplicate tags in excess of what is expected by the sequencing depth (binomial distribution p-value $<10^{-5}$)
- Tag distribution is modeled by a Poisson distribution, and using a 2d window to find peaks with a significant tag enrichment (Poisson distribution p-value based on λ_{BG}, default 10^{-5}).
- Overlapping enriched tags are merges and each tag position is extended d bases from its center.
- The location (summit) of the highest fragment pileup is predicted to be the precise binding location

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

λ captures both the mean and the variance of the distribution.

e is a constant (natural log)=2.71828



MACS
Peak Detection Extras

Background

Instead of using a uniform background (λ_{BG}) from the whole genome they use a dynamic parameter, λ_{local} for each candidate peak where:

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

where λ_{1k}, λ_{5k} and λ_{10k} are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample...where no control sample available then λ_{1k} is not used.

MACS
Peak Detection Extras

Background

λ_{local} captures the influence of local biases, and is robust against occasional low tag counts at small local regions. MACS uses λ_{local} to calculate the p-value of each candidate peak and removes potential false positives due to local biases (that is, peaks significantly under λ_{BG}, but not under λ_{local}). Candidate peaks with p-values below a user-defined threshold p-value (default 10^{-5}) are called, and the ratio between the ChIP-Seq tag count and λ_{local} is reported as the fold_enrichment.

MACS
Practical Use

OUTPUT

- NAME_peaks.xls
- NAME_peaks.bed
- NAME_summits.bed
- NAME_negative_peaks.xls
- NAME_model.r
- NAME_treat/control_aftershifting.wig.gz

Peak Calling

When do you know a ChIP-seq is not working?

If there is a control library, a ChIP-seq that is not working should result in few called peaks, and side-by-side inspection of selected genomic loci in the ChIP and control libraries should show poor enrichment. However, even when two identical libraries are sequenced, there will be several areas that may show significant count differences (as part of an FDR). The ultimate test would be the quantitative PCR validation of selected ChIP-seq peaks. For some transcription factors with well characterized motifs it can make sense to check for the occurrence of the motif in a significant fraction of the called peaks.

MACS
Practical Use

Macs come in two version

- Differences poorly documented
- Different syntax
- 1.4 used pvalues 2.0 uses qvalues (FDR)

Using macs for peak calling in unix:

- macs14 –t test.bam –c control.bam –f BAM –n name –g hs –w -bdg

- macs2 callpeak -t test.bam -c control.bam -f BAM -g hs –n name -B -q 0.01

Quality Control on the called peaks

QC of Output (ENCODE)

- Visual Inspection
  (known positive control - similar dataset)
- Measure global ChIP enrichment (FRIP) >1%
- Cross Correlation analysis (two peaks)
- Consistency for replicates (Analysis using IDR)

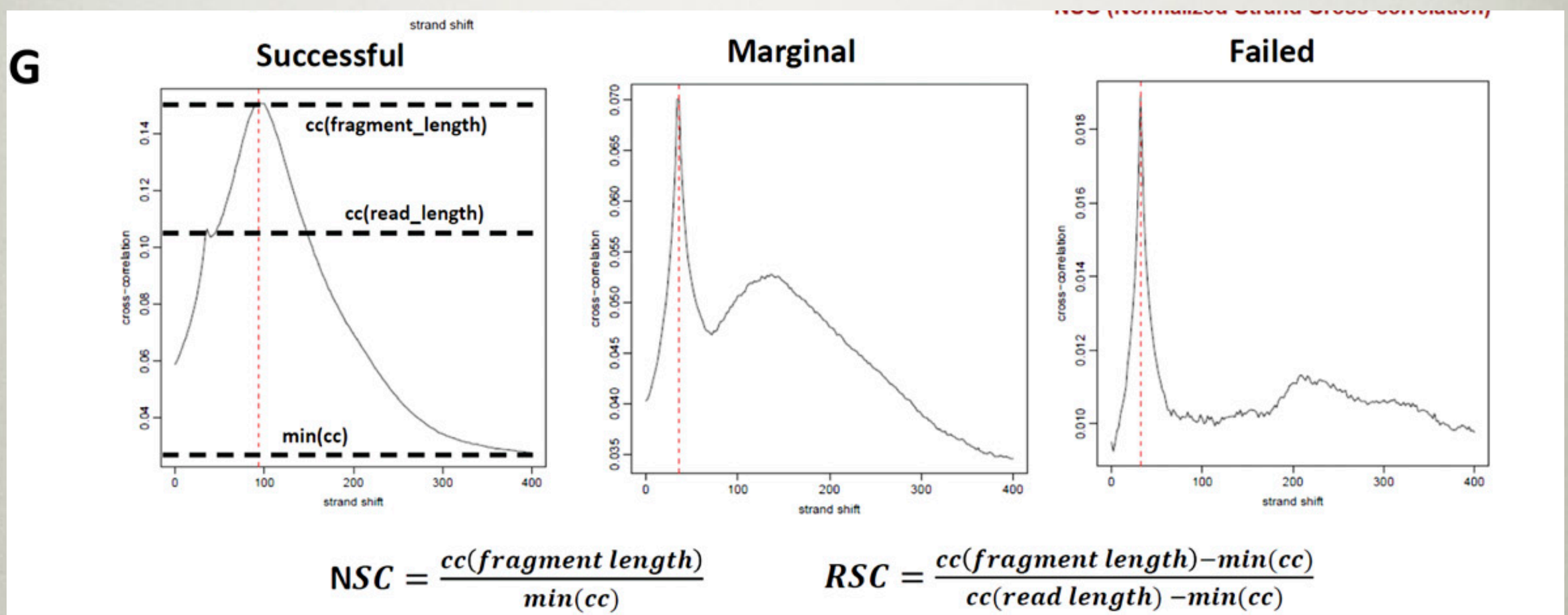
In layman's terms, the IDR method compares a pair of ranked lists of identifications (such as ChIP-seq peaks). These ranked lists should not be pre-thresholded i.e. they should provide identifications across the entire spectrum of high confidence/enrichment (signal) and low confidence/enrichment (noise). The IDR method then fits the bivariate rank distributions over the replicates in order to separate signal from noise based on a defined confidence of rank consistency and reproducibility of identifications i.e the IDR threshold.

QC of Output (ENCODE)

Thus far, the most successful point-source factor experiments for ENCODE have FRiP values of 0.2–0.5 (factors such as REST, GABP, and CTCF) and NSC/RSC values of 5–12. Although these quality scores and characteristics were routinely obtained for the best-performing factor/antibody combinations, they are not the rule; for most transcription factors, the ChIP quality metrics were substantially lower and more variable.

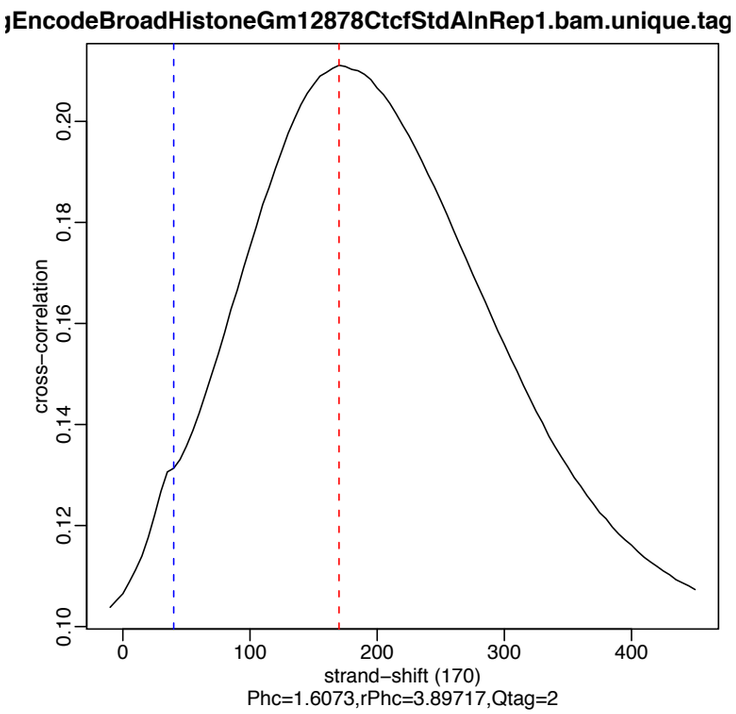
FRiP - Fraction of reads in the Peaks
NSC - Normalized Strand Correlation
RSC - Relative Strand Correlation

QC of Output (ENCODE)

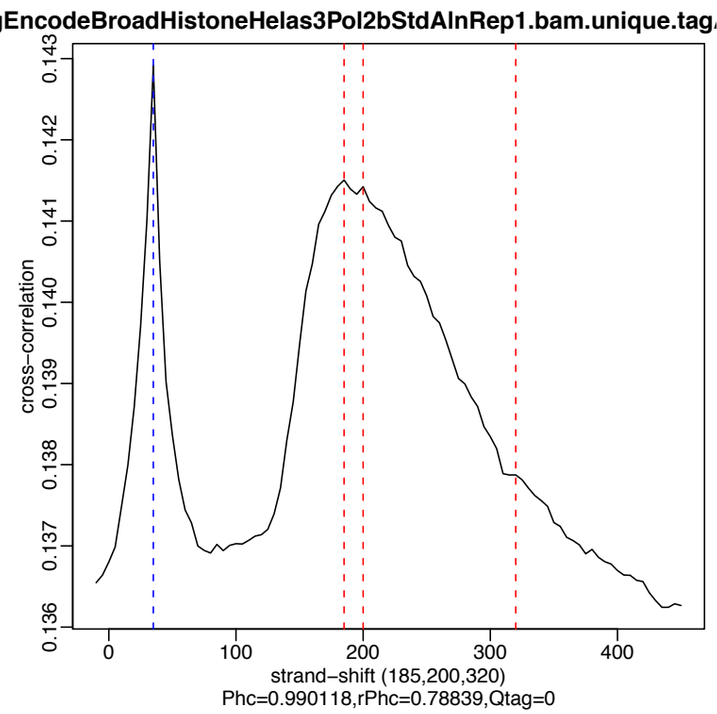


$$\mathbf{NSC} = \frac{cc(fragment\ length)}{min(cc)} \qquad \boldsymbol{RSC} = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

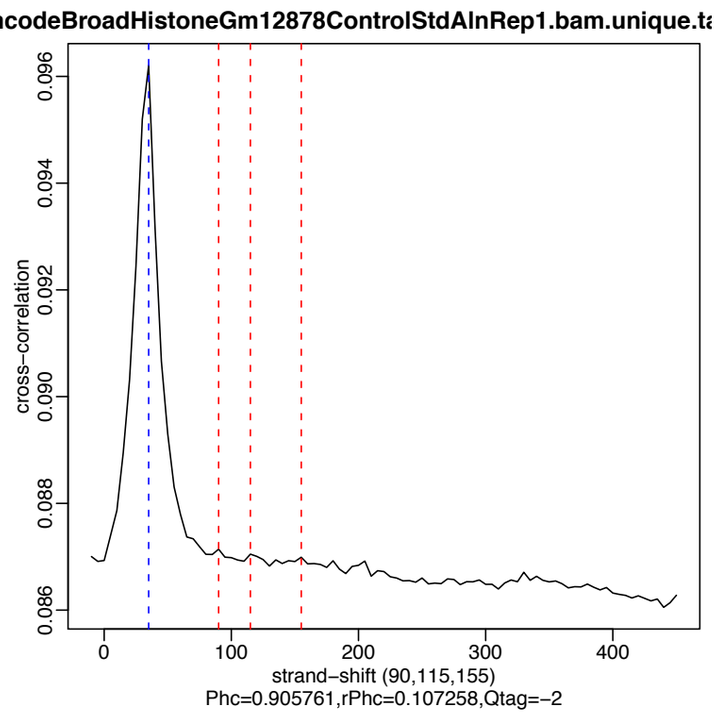
Cross Correlation Plots



Good



Poor

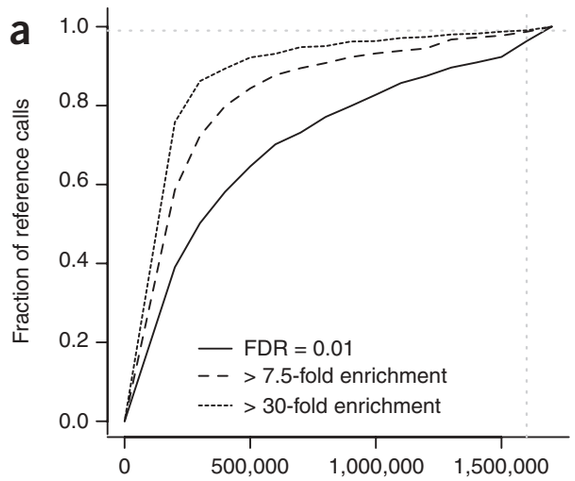


Input

Quality Control

- Clonal Tag Counts
- Sequencing Fragment Length Estimation (tag autocorrelation)
- Checking for Sequence Bias

As more pronounced binding positions are identified using smaller sequencing depth, an experiment of given depth may saturate detection of the binding positions that exceed a certain tag enrichment ratio relative to the background. We refer to this enrichment ratio as the minimal saturated enrichment ratio (MSER).



- http://biowhat.ucsd.edu/homer/chipseq/qc.html#Sequencing_Fragment_Length_Estimation
- https://sites.google.com/site/anshulkundaje/projects/idr
- CHANCE

What quality is need for for further analysis

- Motif Analysis (**low**)
- Discovering regions to test for biological function such as transcriptional enhancement, silencing, or insulation (**Medium - High**)
- Deducing and mapping combinatoric occupancy (**High**)
- Integrative analysis (**High**)

Functional Analysis

Analysis downstream to peak calling

- Visualization - genome browser: Ensembl, UCSC, IGB
- Peak Annotation - finding interesting features surrounding peak regions:
- Correlation with expression data
- Discovery of binding sequence motifs
- Split peaks
- Fetch summit sequences
- Run motif prediction tool
- Gene Ontology analysis on genes that bind the same factor or have the same modification

Function analysis

- **Visualization**
  - IGV & IGB
  - UCSC Genome
  - Heatmaps
- **C**is-regulatory **E**lement **A**nnotations **S**ystem (CEAS)
- Homer
- MEME
- **GREAT** predicts functions of cis-regulatory regions

ENCODE ChIP-Seq peaks are screened against a specially curated empirical **blacklist** of regions in the human genome. Peaks overlapping the blacklisted regions were discarded.

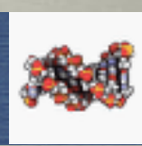
(http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz)

These artifact regions typically show the following characteristics:

Unstructured and extreme artifactual high signal in sequenced input-DNA and control datasets as well as open chromatin datasets irrespective of cell type identity.

An extreme ratio of multi-mapping to unique mapping reads from sequencing experiments.

Overlap with pathological repeat regions such as centromeric, telomeric and satellite repeats that often have few unique mappable locations interspersed in repeats.

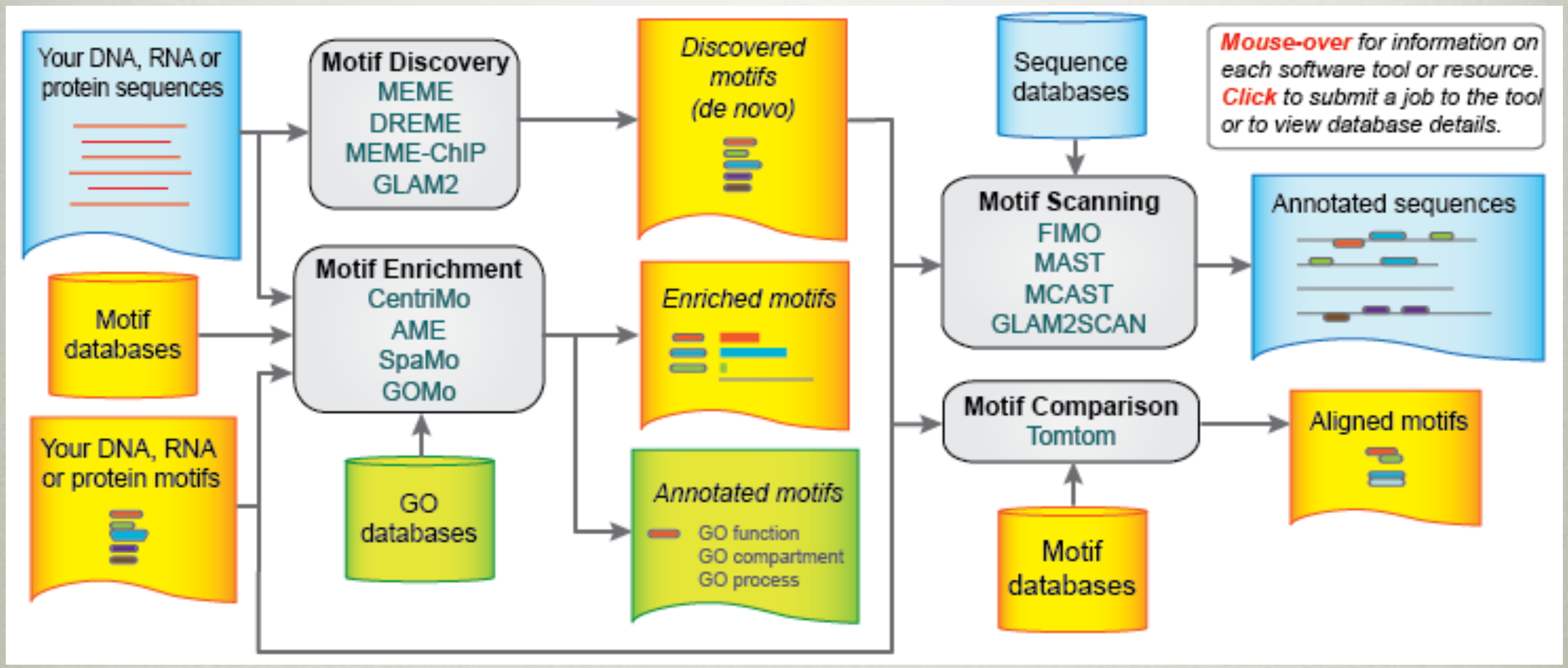


Motif

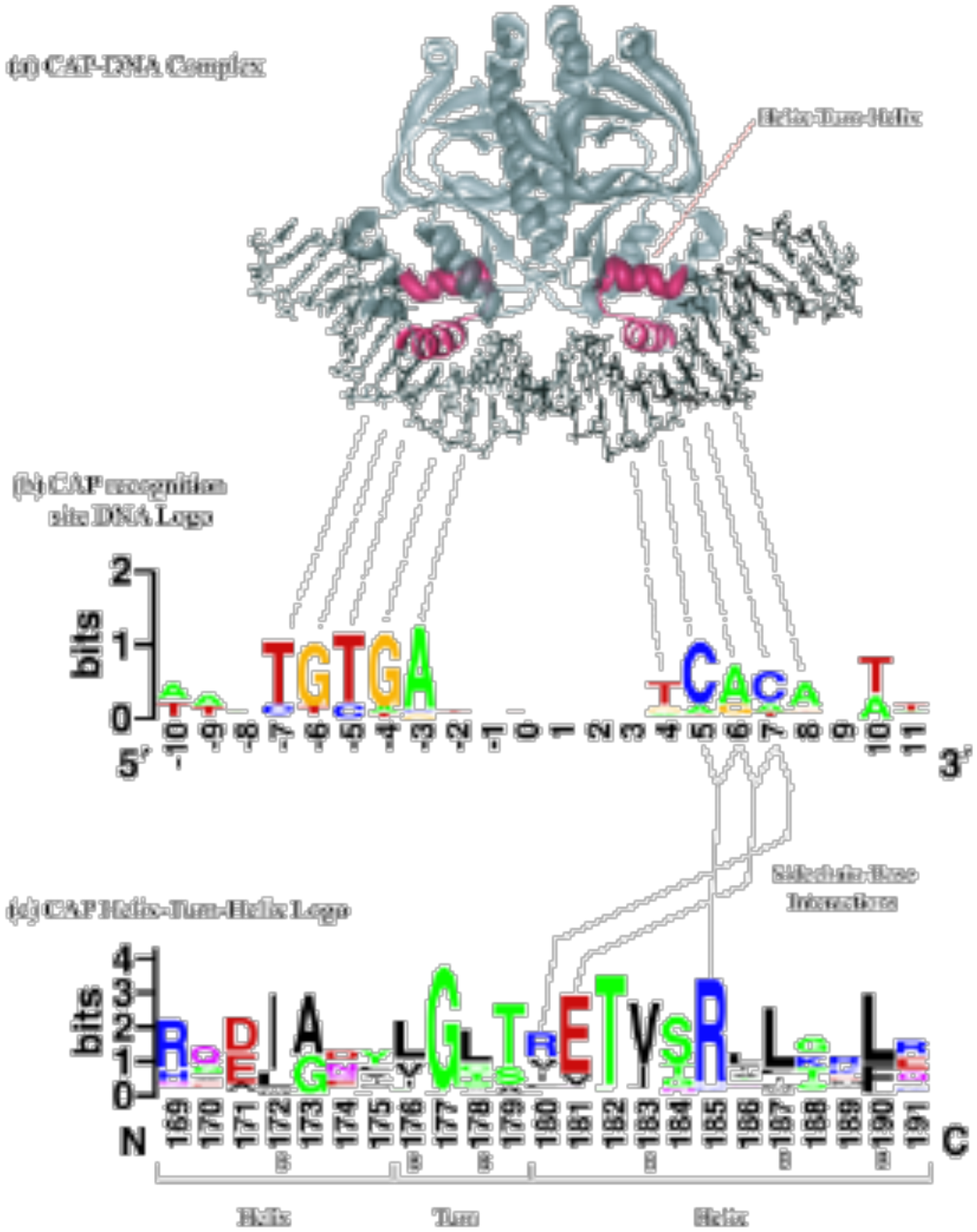
ANALYSIS

MEME

Motif-based sequence



http://meme-suite.org



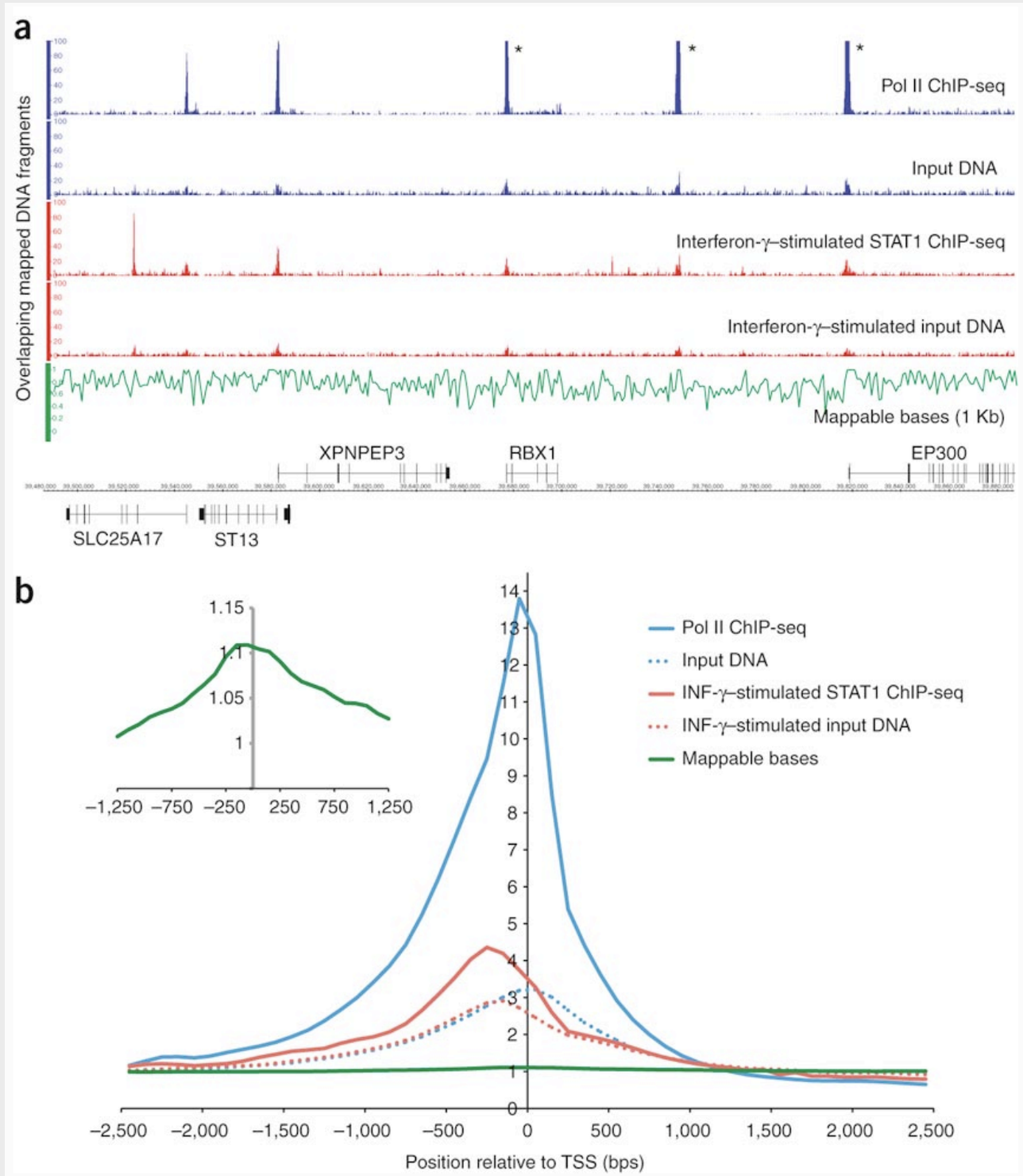
http://weblogo.threeplusone.com

Where to Find ChIPSeq Data

Types of ChIPSeq Data

- NCBI (GEO) (SRA -tabular)
- UCSC (various - bam,bed,fastq,other)
- ENCODE (various - bigBed (.bb) and bigWIG (.bw))
- ChIPBase (CSV)
- Cistrome Browser

Visualization



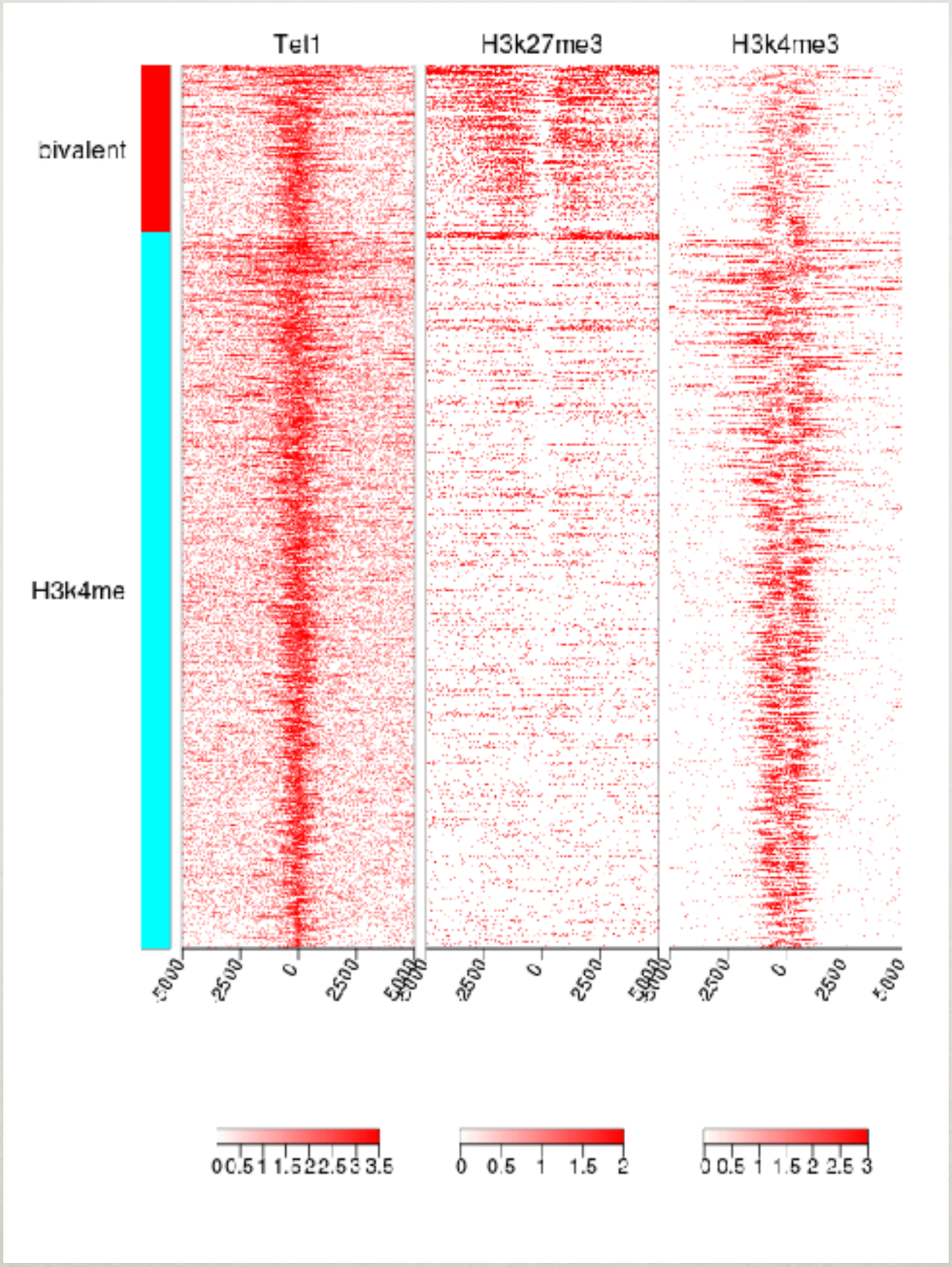
Rozowsky, Nature Biotechnology, 2009

Visualization

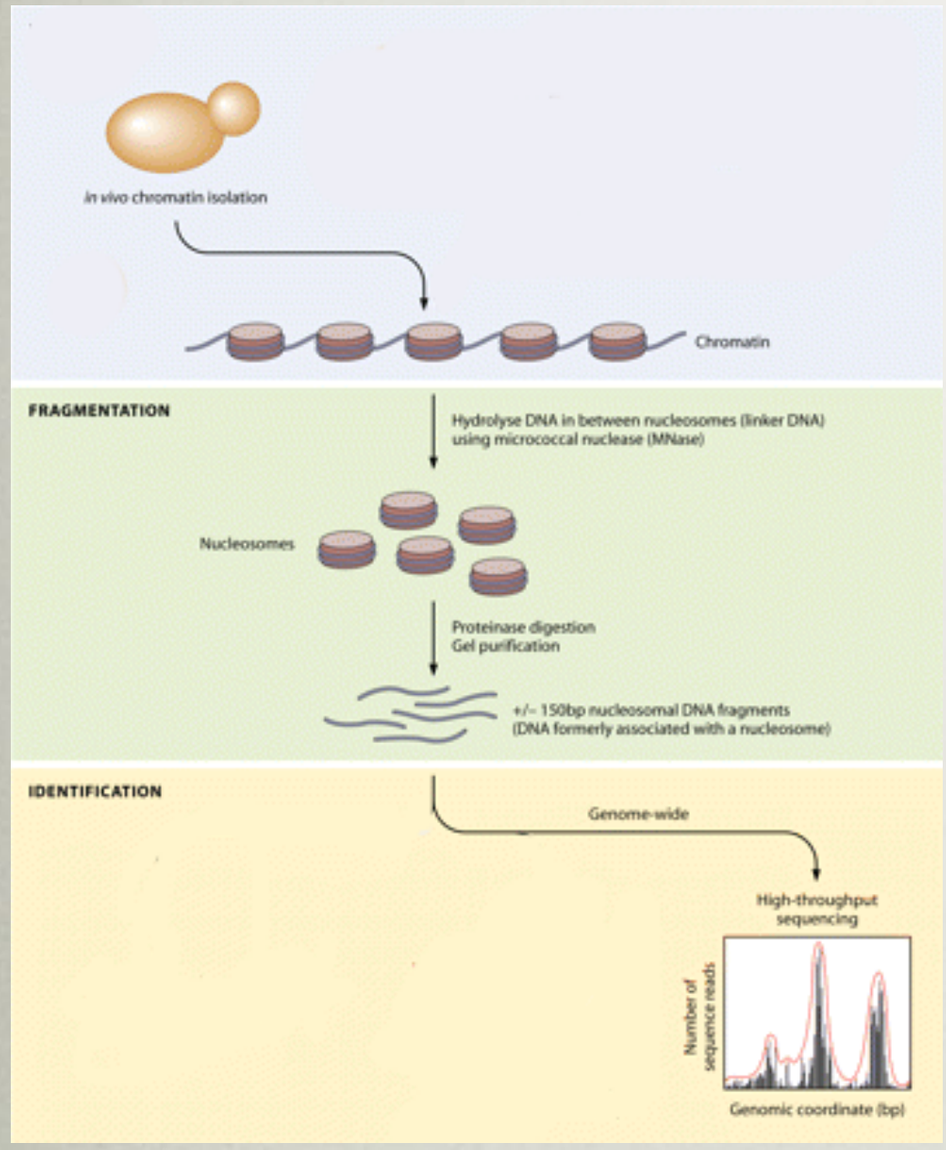
Nothing can match the insight obtained by looking at your data

- IGV
- UCSC Genome Browser
- Heatmaps
- NGS-plot

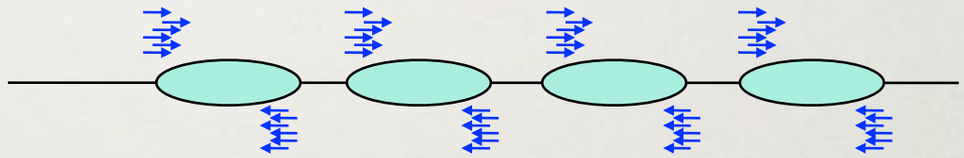
Heat Maps


Tel1
H3k27me3
H3k4me3
bivalent
H3k4me
-5000
-2500
0
2500
5000
0 0.5 1 1.5 2 2.5 3 3.5
0 0.5 1 1.5 2
0 0.5 1 1.5 2 2.5 3

Yeast as Model Organism



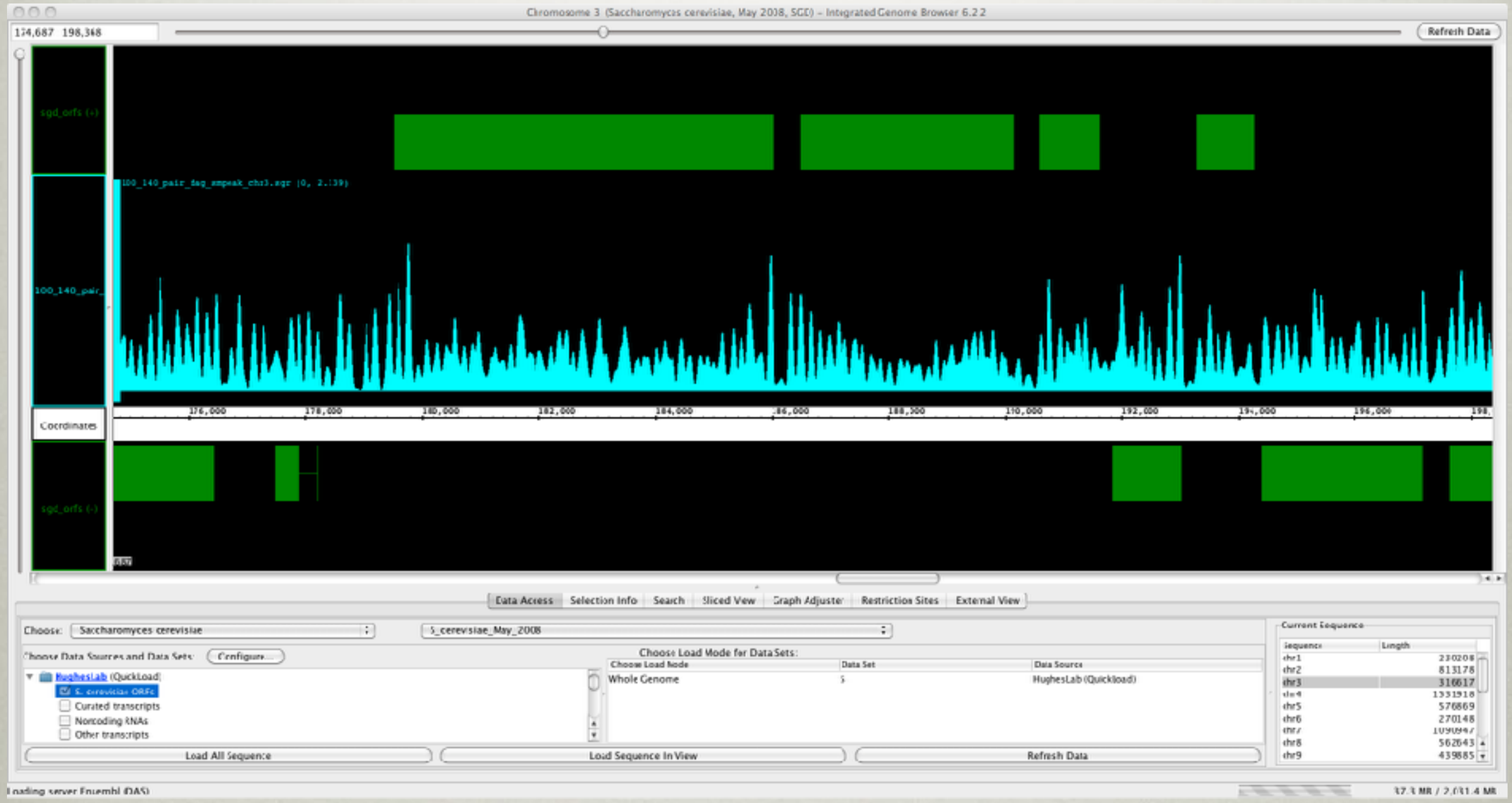
Steps of converting the sequencing reads to nucleosome positions



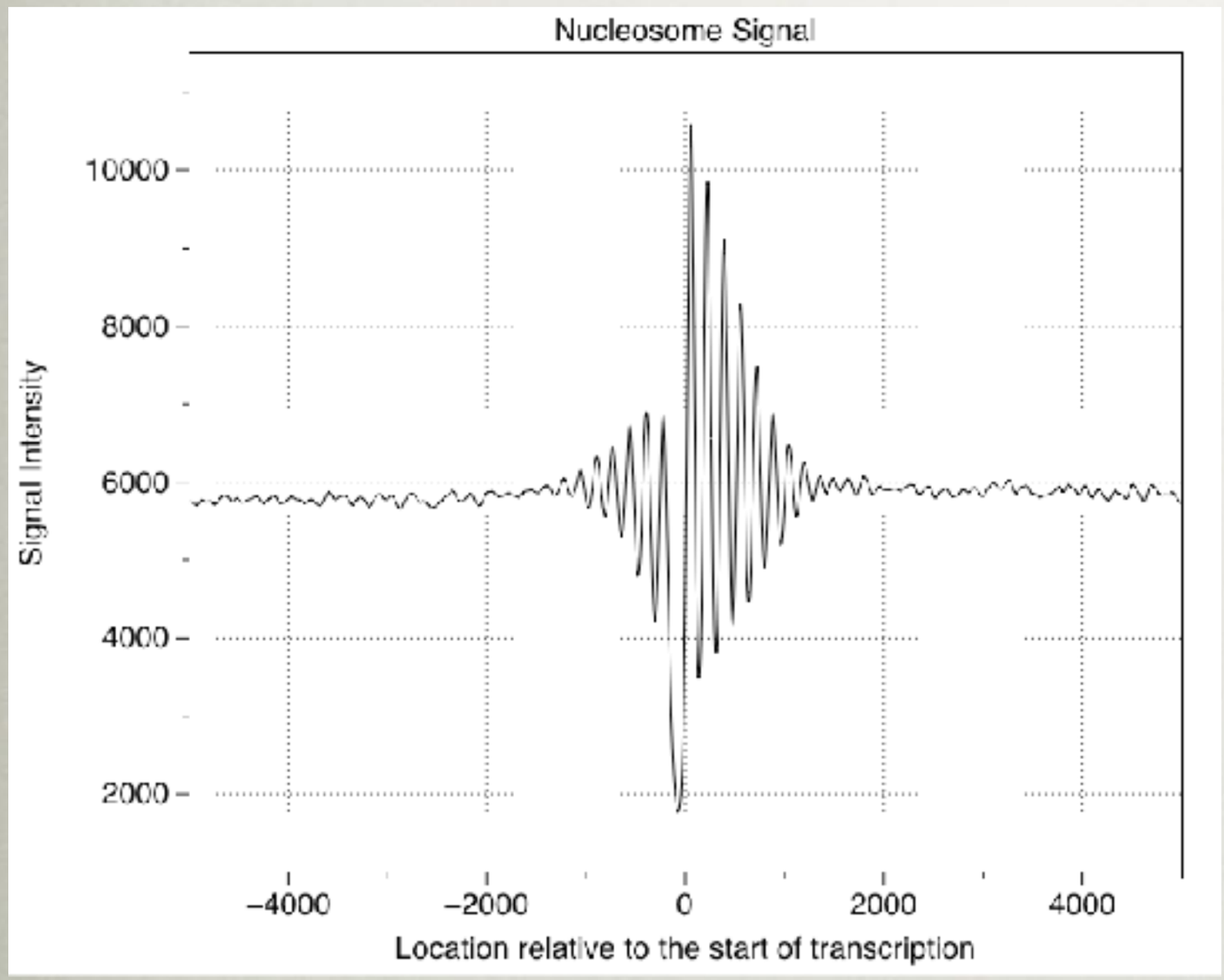
Align all the sequencing reads against the genome

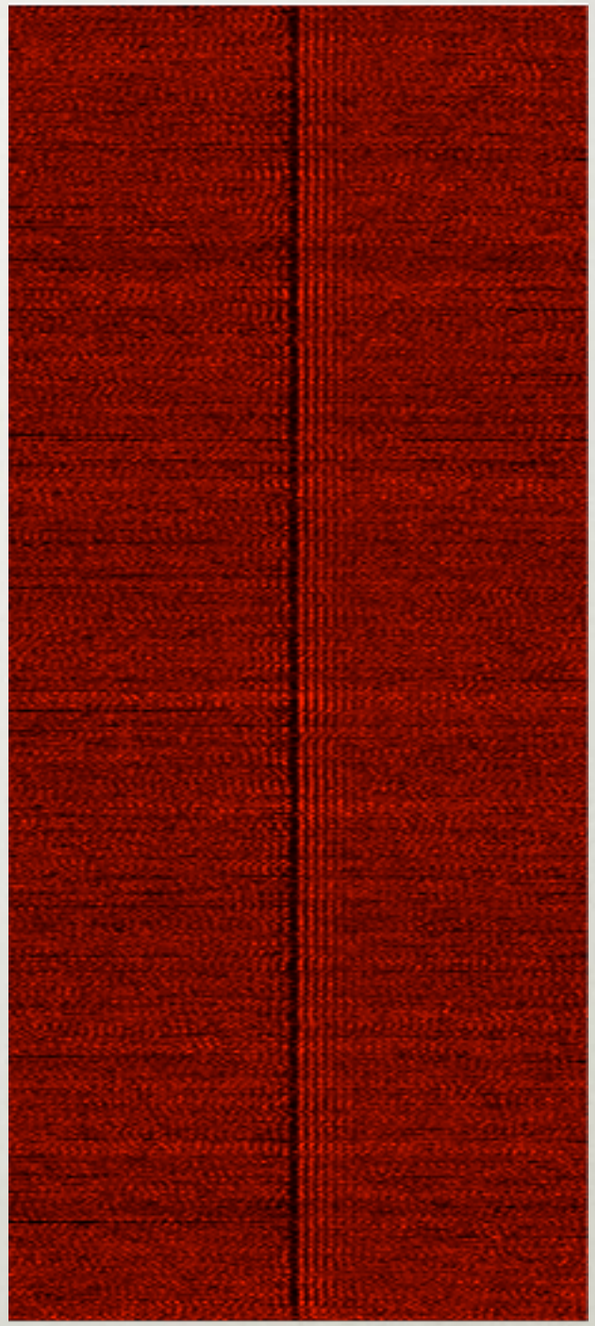
Sum up the signal for both DNA stands

A matching pair with proper distance apart are combined and centered


Chromosome 3 (Saccharomyces cerevisiae, May 2008, SGD) – Integrated Genome Browser 6.2.2
174,687 198,368
Refresh Data
sgd_orfs (+)
100_140_pair_
100_140_pair_dap_smpeak_chr3.sgr [0, 2.139]
Coordinates
sgd_orfs (-)
Data Access
Selection Info
Search
Sliced View
Graph Adjuster
Restriction Sites
External View
Choose:
Saccharomyces cerevisiae
S_cerevsiae_May_2008
Choose Data Sources and Data Sets:
Configure...
HughesLab (QuickLoad)
S. cerevisiae ORFs
Curated transcripts
Noncoding RNAs
Other transcripts
Choose Load Mode for DataSets:
Choose Load Mode
Data Set
Data Source
Whole Genome
S
HughesLab (Quickload)
Load All Sequence
Load Sequence In View
Refresh Data
Current Sequence
Sequence
Length
chr1 230208
chr2 813178
chr3 316617
chr5 576869
chr6 270148
chr8 562643
chr9 439885

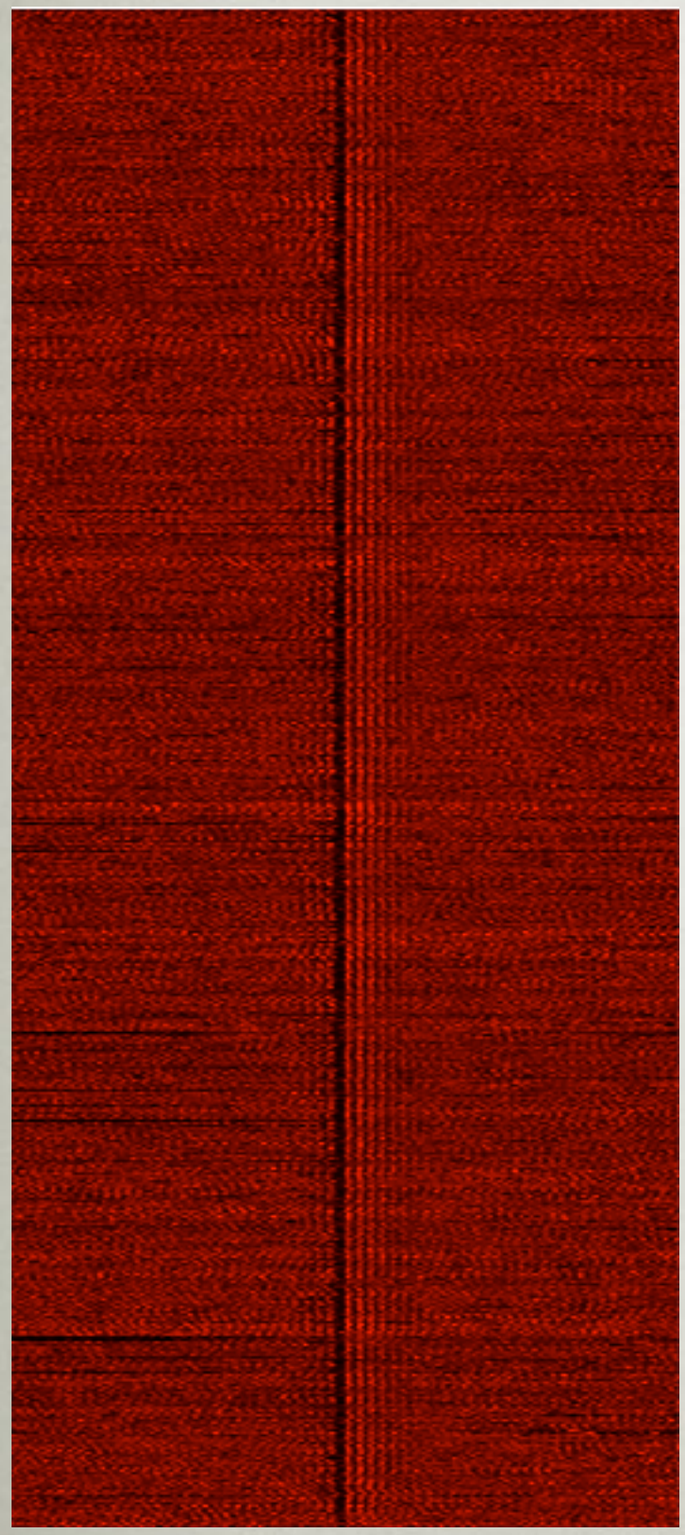
* start site of all genes (~4000)

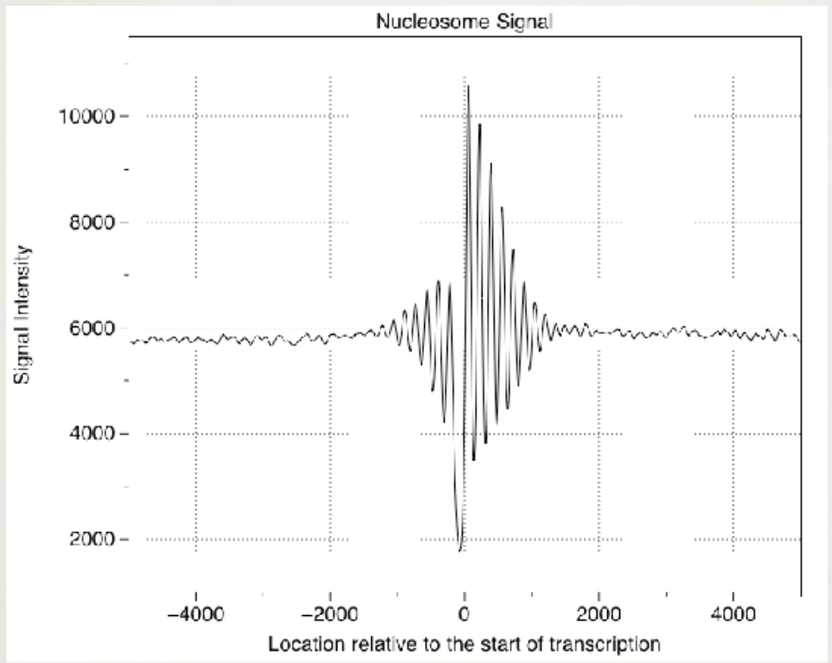


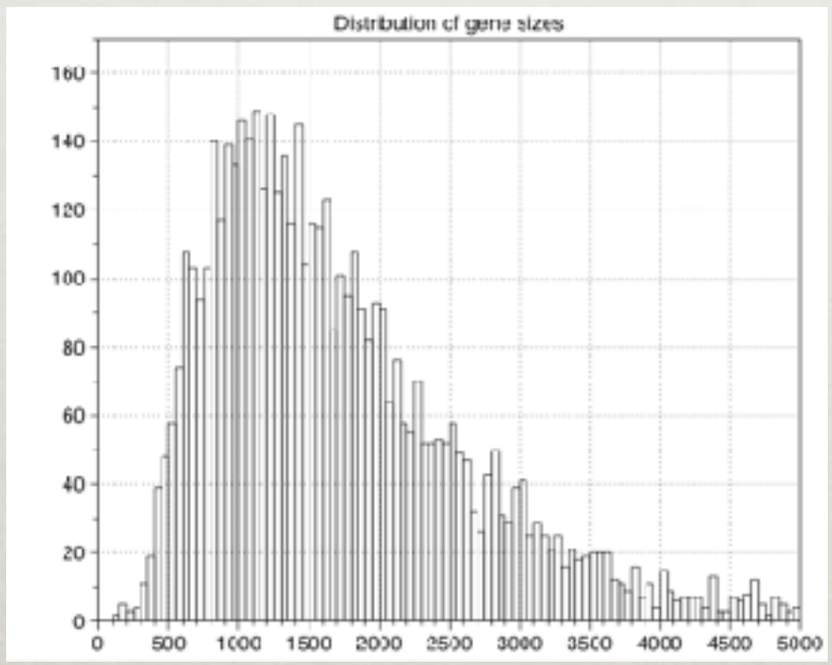


−5000 0 5000

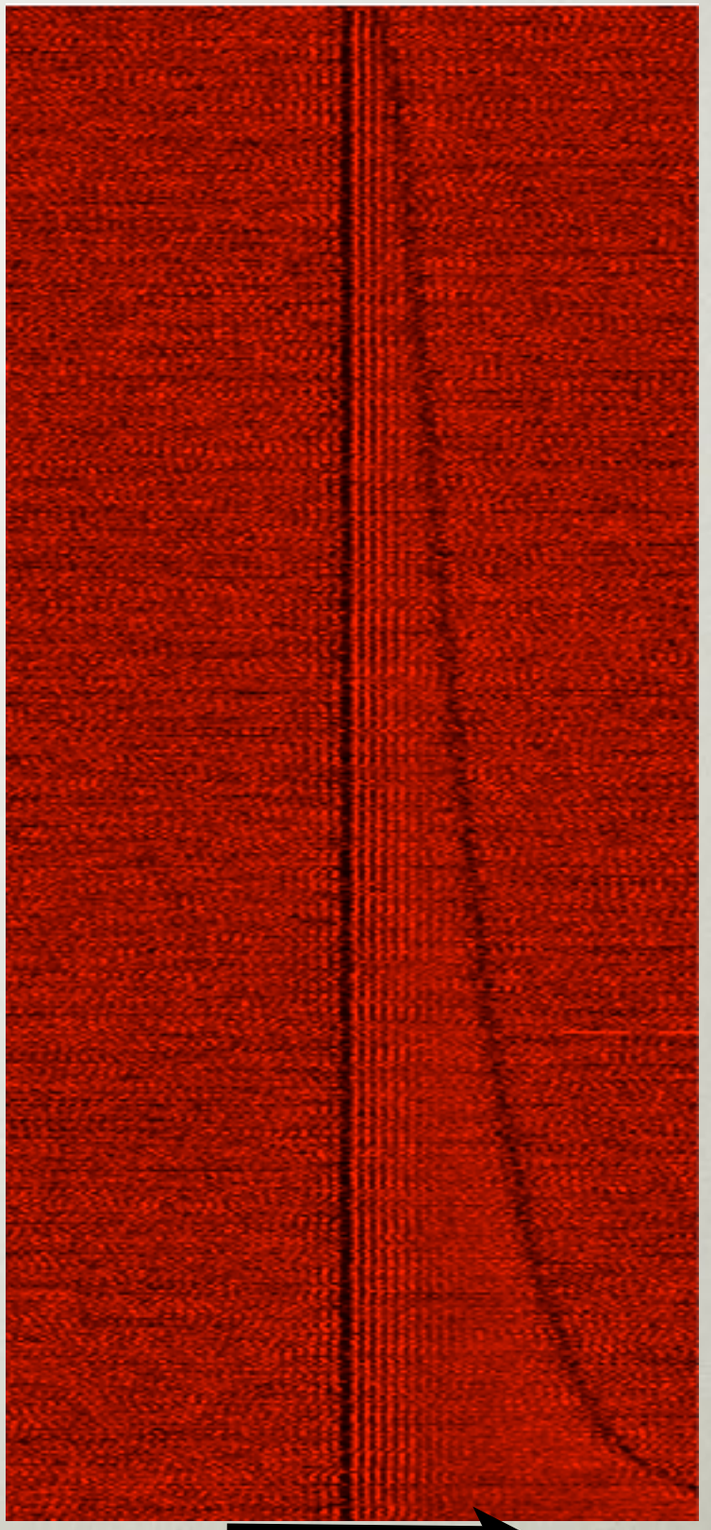
Original



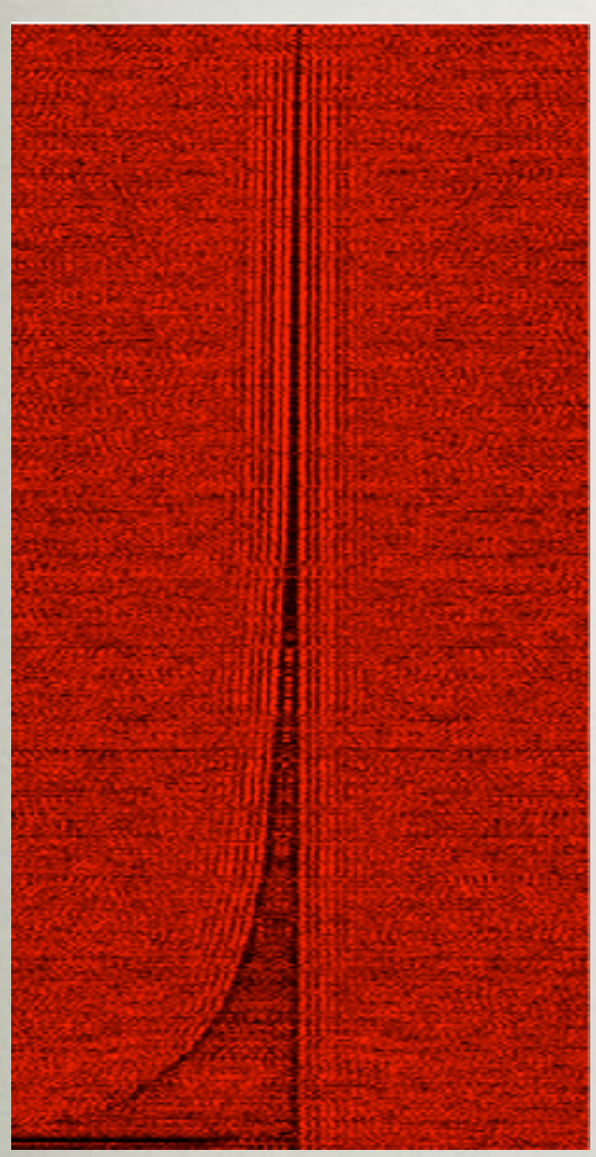


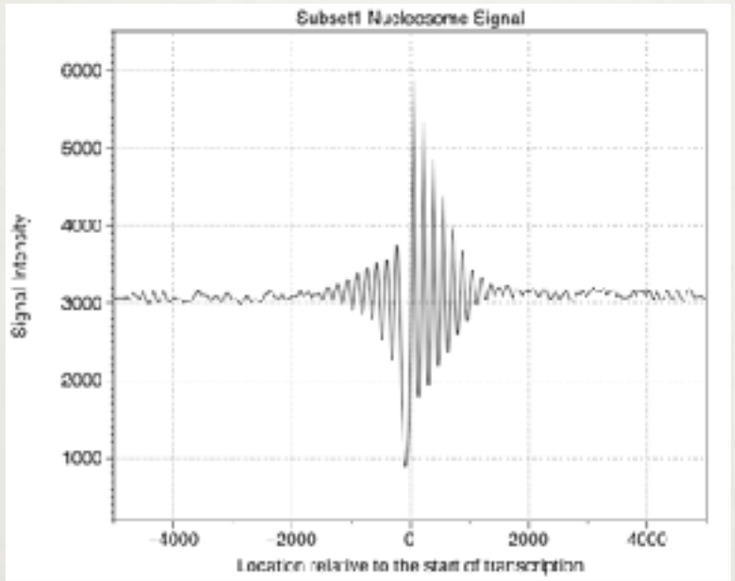


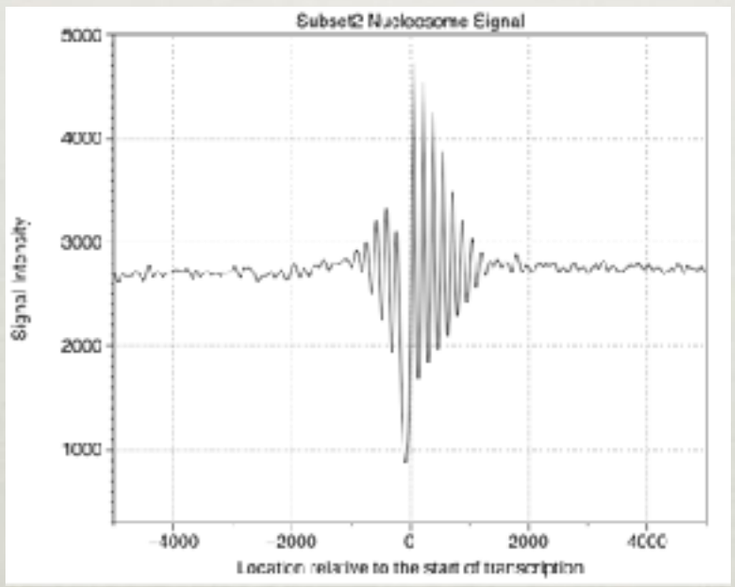
Sorted by Gene Size

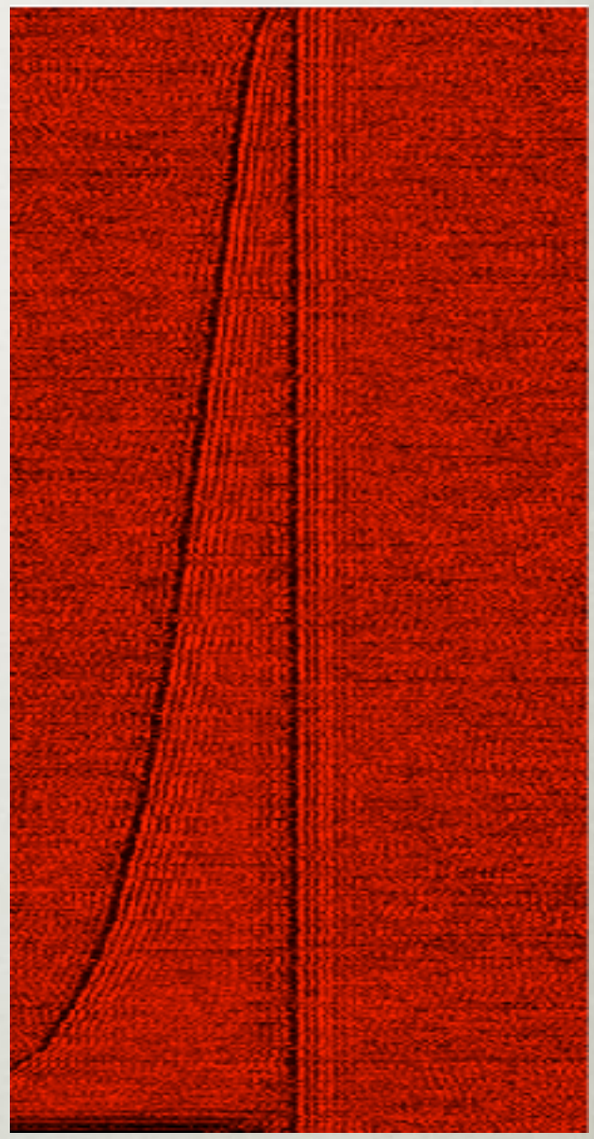


Sorted by nearest neighbour









Take home message

- Think about what the data may be telling you and explore different ways of looking at the same data.
- Be wary of summation plots/statistics... they may be "correct" but they can lead you astray or hide the better story.

References

Early ChIPSeq References

- Johnson DS, et al. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316(5830):1497–502. [PubMed: 17540862]
- Barski A, et al. High-resolution profiling of **histone** methylations in the human genome. Cell. 2007; 129(4):823–37. [PubMed: 17512414]
- Robertson G, et al. Genome-wide profiles of **STAT1** DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007; 4(8): 651–7. [PubMed: 17558387]
- Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007; 448(7153):553–60. [PubMed: 17603471]

Review References

- Park, P. J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet., 10, 669–680.
- Hyunjin Shin, Tao Liu, Xikun Duan, Yong Zhang and X. Shirley Liu, Computational methodology for ChIP-seq analysis Quantitative Biology 2013, 1(1): 54–70 DOI 10.1007/s40484-013-0006-2

References

- http://www.slideshare.net/COST-events/chipseq-data-analysis (SLIDES)

- http://bbcf.epfl.ch/bbcflib/tutorial_chipseq.html

- http://www.biocodershub.net/community/get-the-most-of-your-chip-seq-experiments/

- http://collaboratory.lifesci.ucla.edu/node/35 (Course)

- https://github.com/songlab/chance (QC suite...interesting)

References

- http://ccg.vital-it.ch/chipseq/ AND http://chip-seq.sourceforge.net
- http://www.youtube.com/watch?v=4oFdS9EN9Pk
- http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/chip-seq-analysis/chip-seq-practical
- http://medias01-web.embl.de/Mediasite/Play/94ec103b215c4b45a397400fde4029421d (VIDEO)
- http://liulab.dfci.harvard.edu/MACS/
- http://gettinggeneticsdone.blogspot.com/2013/06/encode-chip-seq-significance-tool-which.html
- https://usegalaxy.org/u/james/p/exercise-chip-seq
- http://sissrs.rajajothi.com
- http://meme.nbcr.net/meme/doc/meme-chip.html (MEME_CHIP)
- https://sites.google.com/a/brown.edu/genomics-club/guidance/peak-callers (list of sites)