ChIPSeq Data Analysis

Peter FitzGerald, CCR, NCI

Talk Outline

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

Course Introduction

Course Outline

Day 1

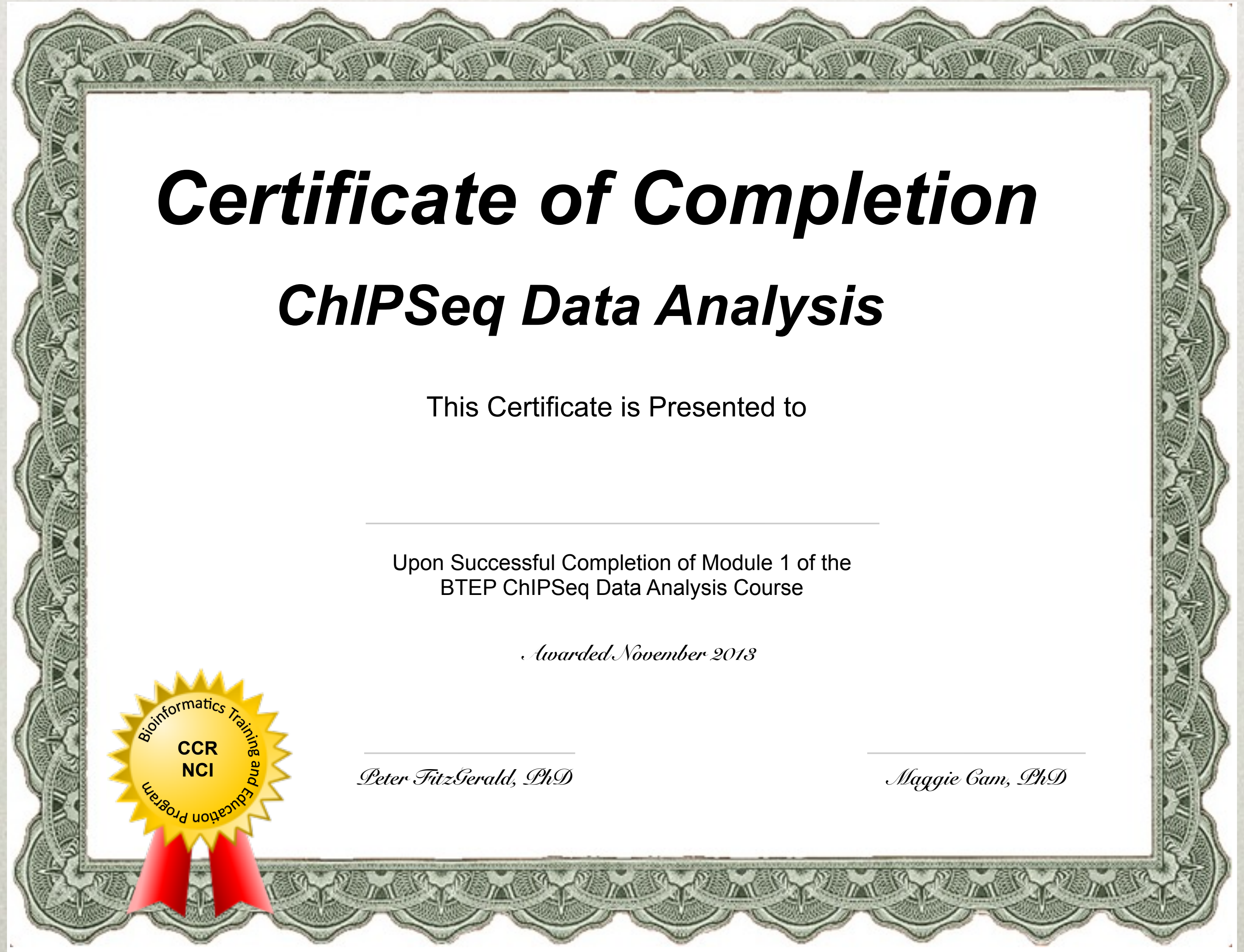
- Design and Analysis Overview (9:30 - 12:30)
- Genomatix (The basics & Data Import and Mapping) - (1:30 - 4:30)

Day 2

- Genomatix (Workflows & Biological Perspective) - (9:30 - 12:30)
- CISTROME (1:30 - 4:30)

Course Reward

Professional Looking Certificate*



Certificate of Completion

ChIPSeq Data Analysis

This Certificate is Presented to

Upon Successful Completion of Module 1 of the
BTEP ChIPSeq Data Analysis Course

Awarded November 2013

Bioinformatics Training and Education Program

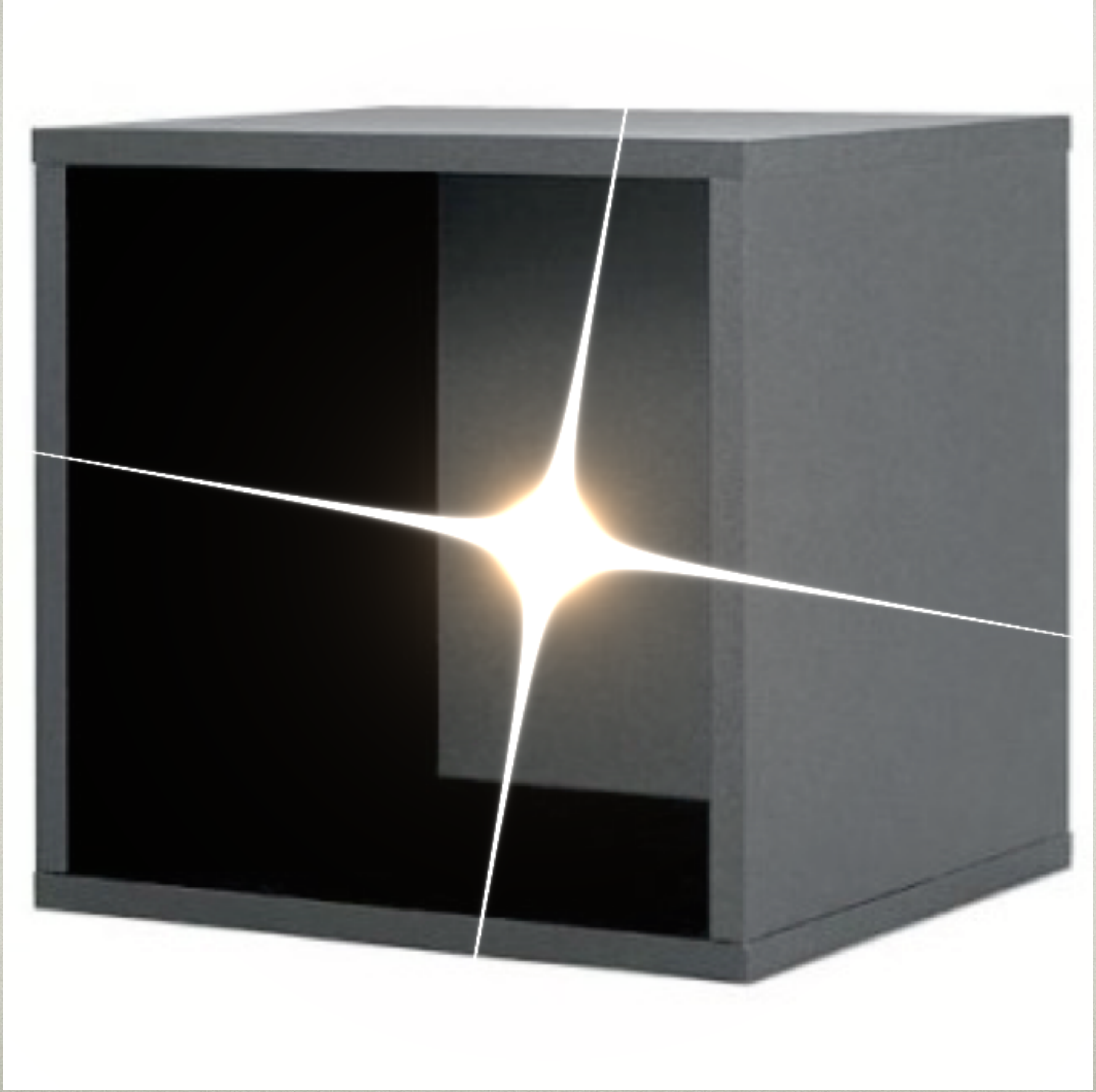
CCR
NCI

Peter FitzGerald, PhD

Maggie Cam, PhD

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 (Cistrome)

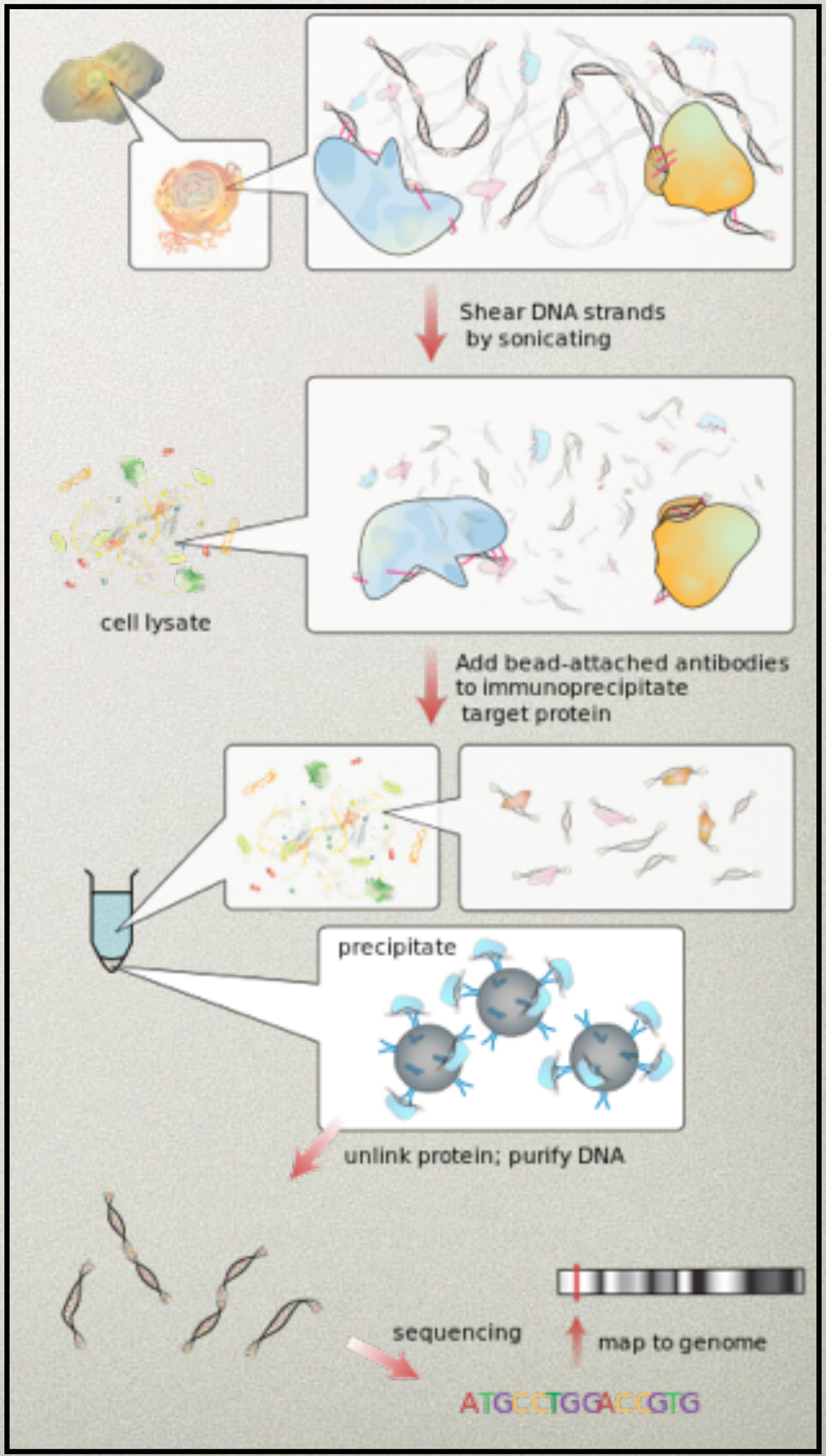


ChIP-SEQ Background

ChIPSEQ

Chromatin ImmunoPrecipitation (ChIP) and massively parallel **seq**uencing (SEQ)

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


Shear DNA strands by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein
precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG

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)

Imported Author Today, 3:18 PM
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

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.

Subset of Techniques

- ChIPSeq
- ChIPExo
- FAIRE-Seq(Formaldehyde-Assisted Isolation of Regulatory Elements)
- DNase Hypersensitivity
- DNase Footprinting

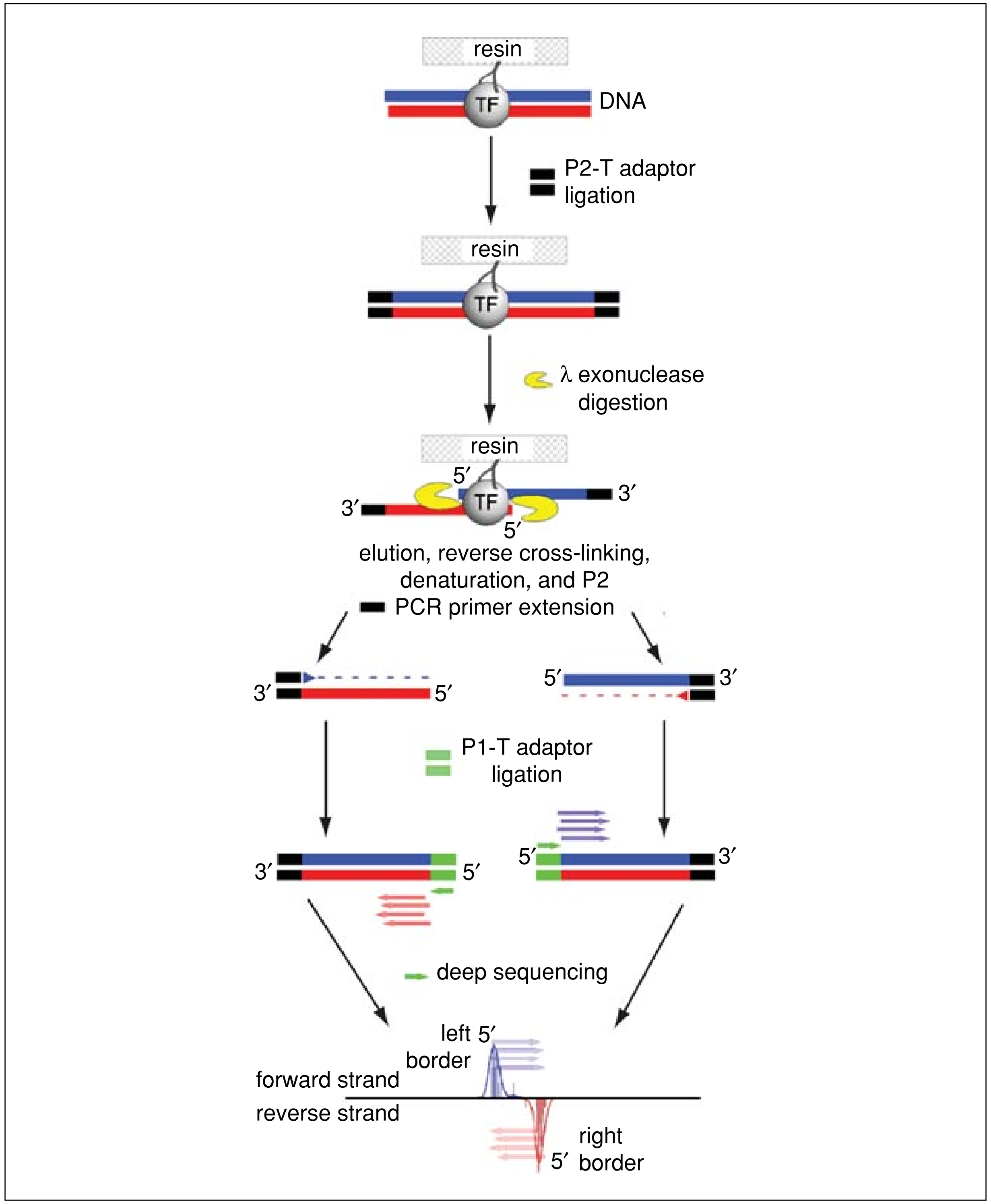
Different Variations

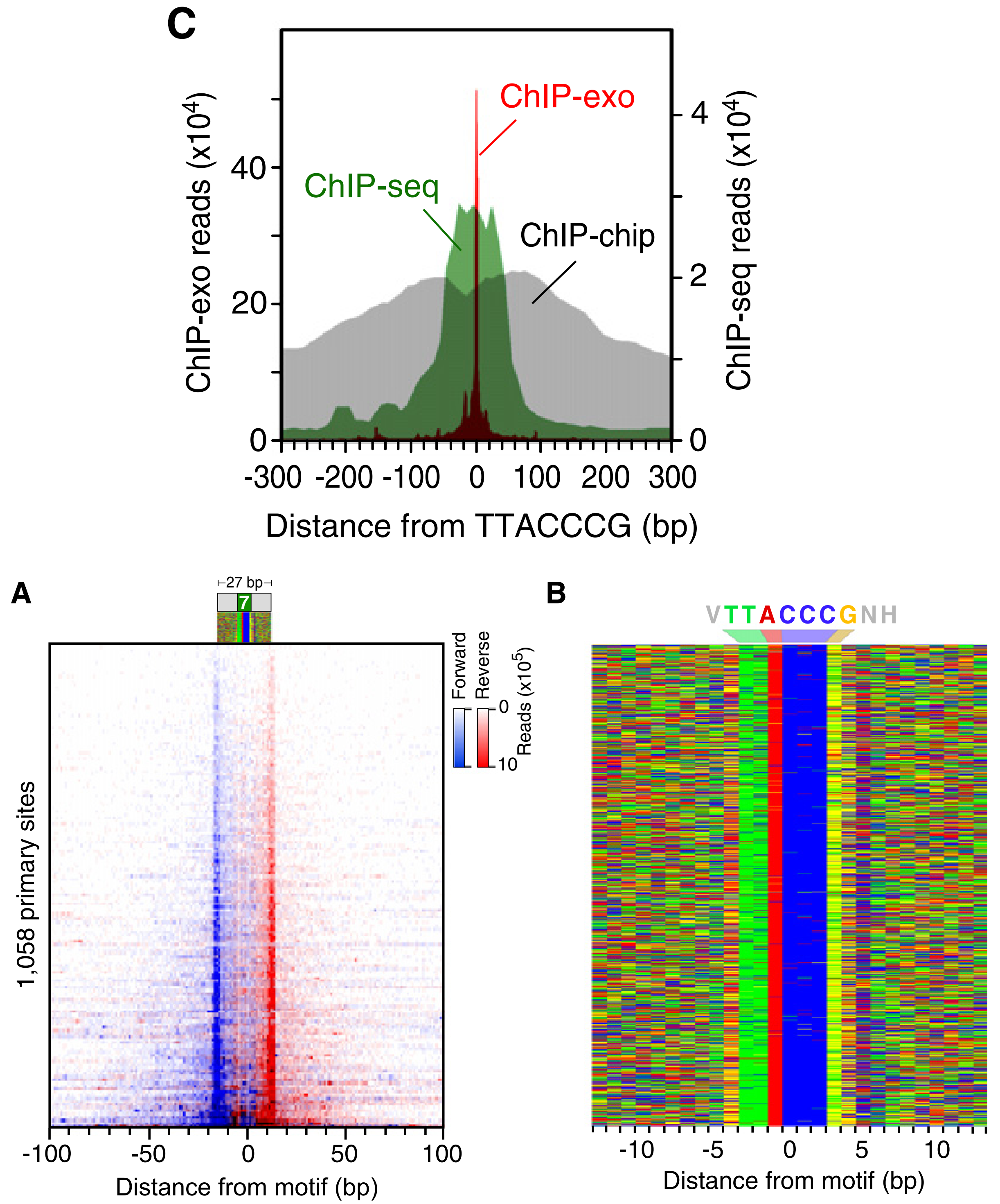
- Native ChIP (N-ChIP)
- Cross link protein and DNA (Formaldehyde) (X-ChIP)
- Protein-Protein cross linking (disuccinimidyl glutarate) and formaldehyde (HDAC- chromatin remodelers)

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

Subset of Techniques

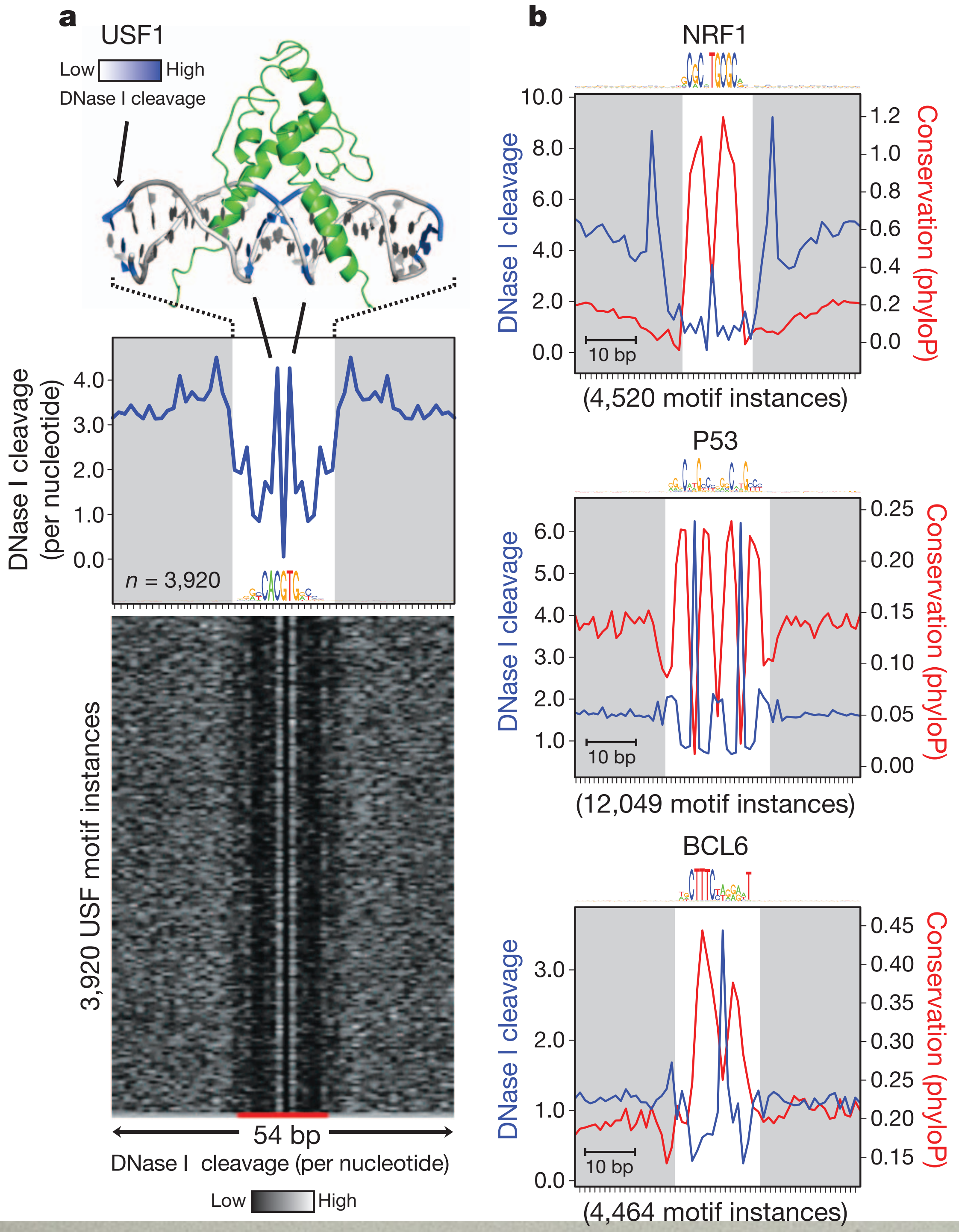
ChIP-exo



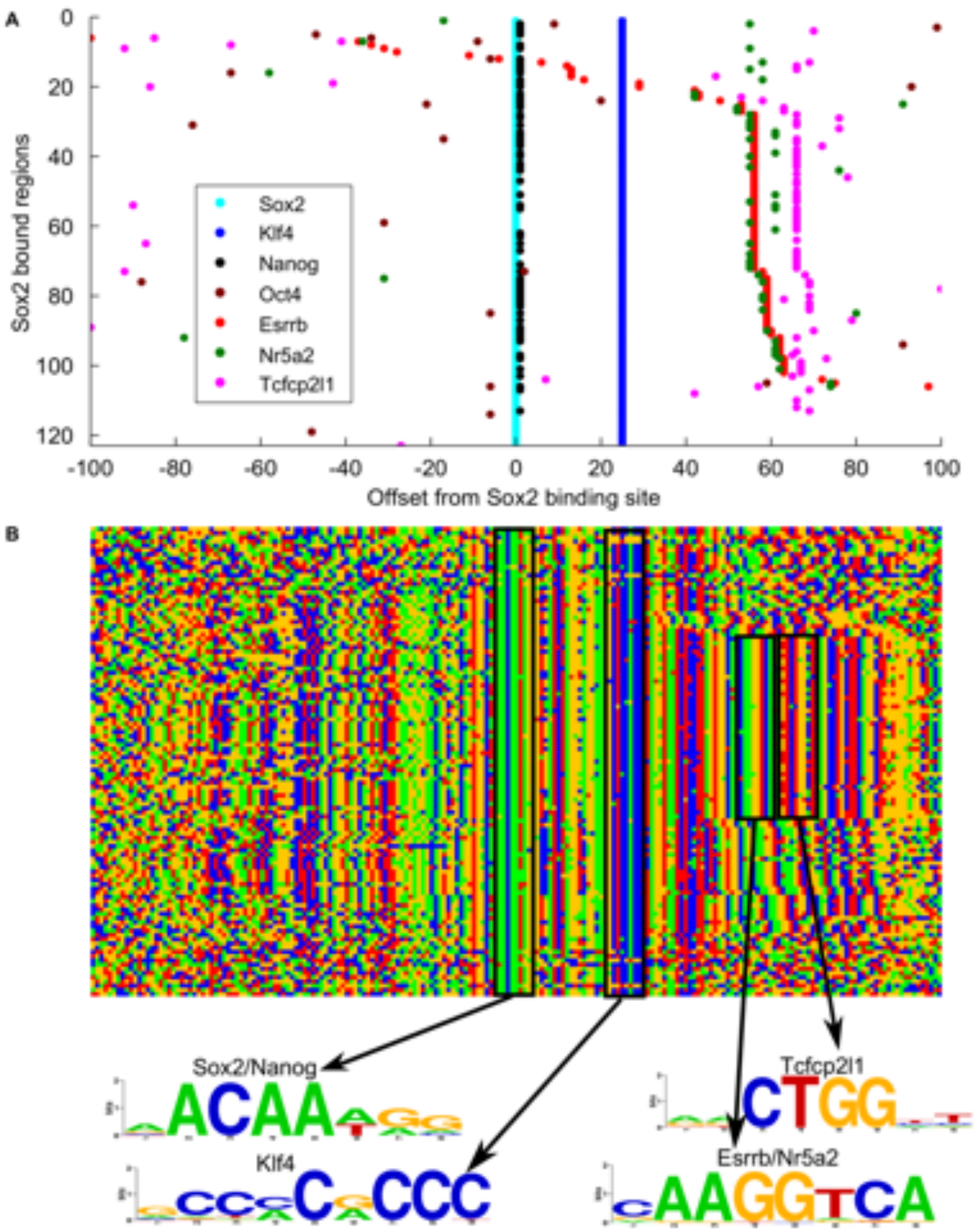


Subset of Techniques

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.

Workshop on Reproducibility of Data Collection and Analysis

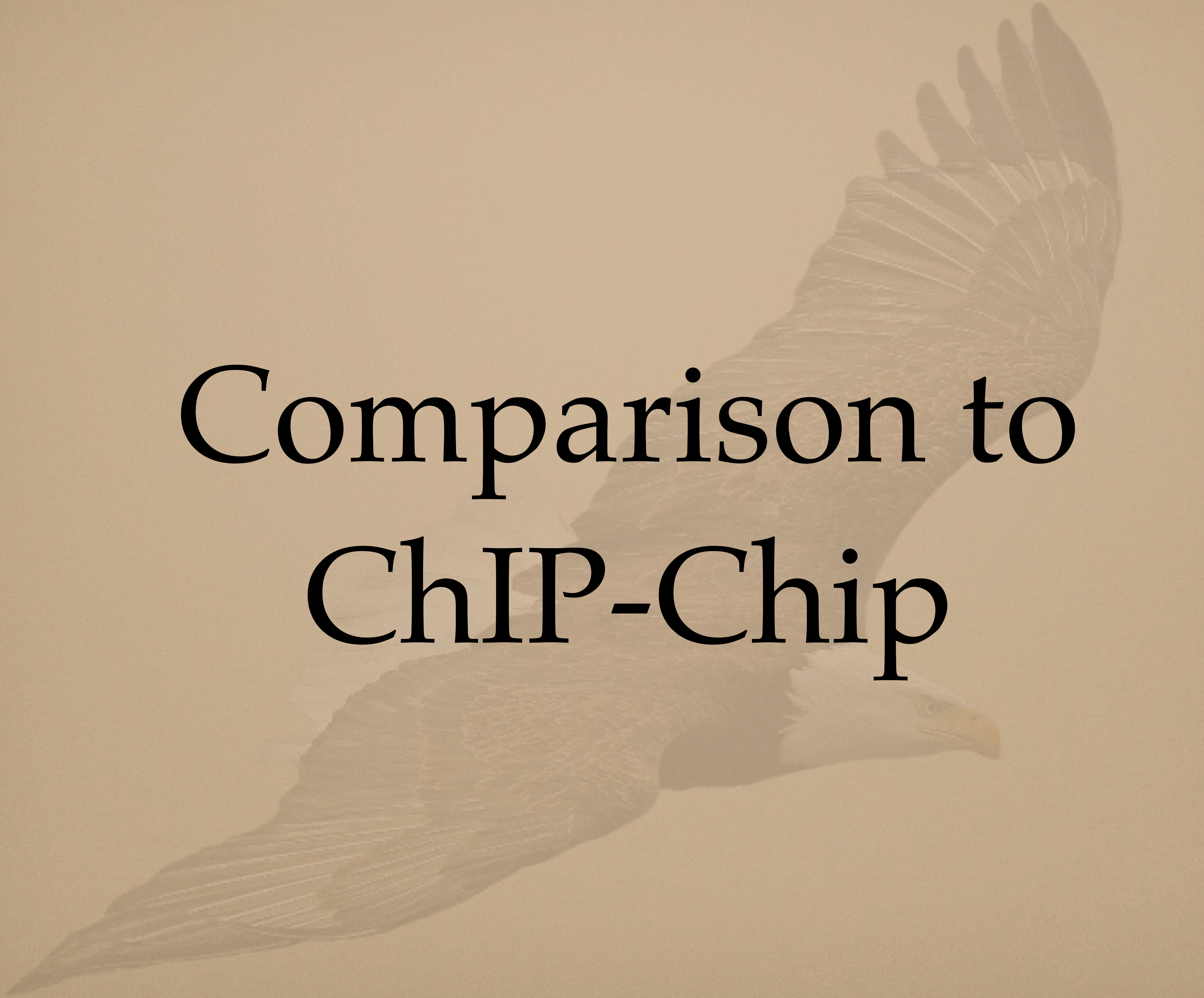
Modern Technologies in Cell Biology: Potentials and Pitfalls

Monday November 24th
8:30 a.m. to 4:30 p.m.
Lipsett Amphitheater, Building 10.

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 !

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.

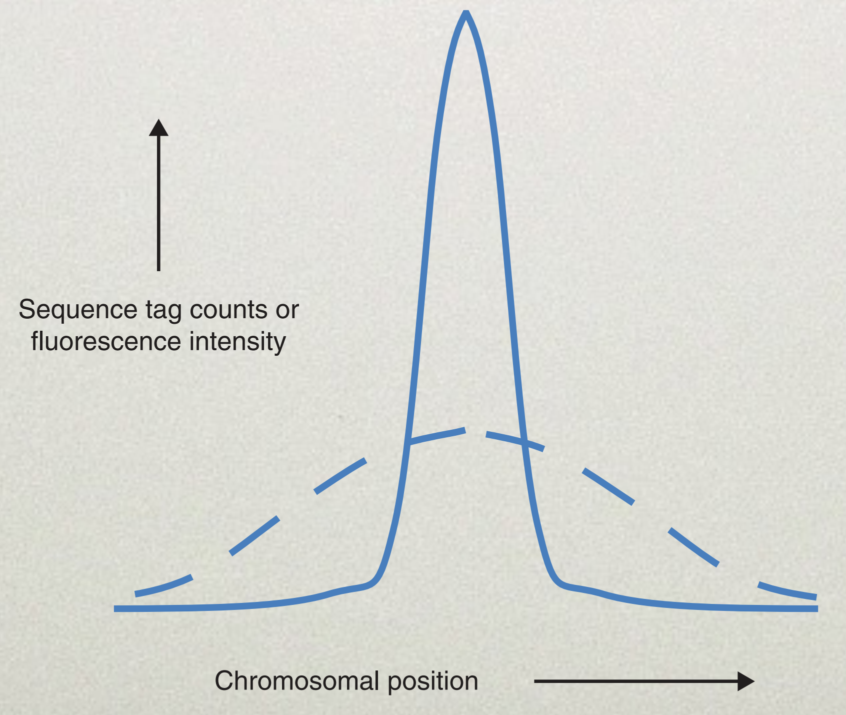


Figure 2. Comparison of ChIP-seq and ChIP-chip. Representative signals from ChIP-seq (solid line) and ChIP-chip (dashed line) show both greater dynamic range and higher resolution with ChIP-seq. Whereas three binding peaks are identified using ChIP-seq, only one broad peak is detected using ChIP-chip.

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

Steps in ChIP-seq

- Wet Lab Experiment
- Generate Sequences Data
- MAP sequences to genome
- Identify "peaks"
- Find motifs
- Correlate peaks/motifs with biology
- Differential studies

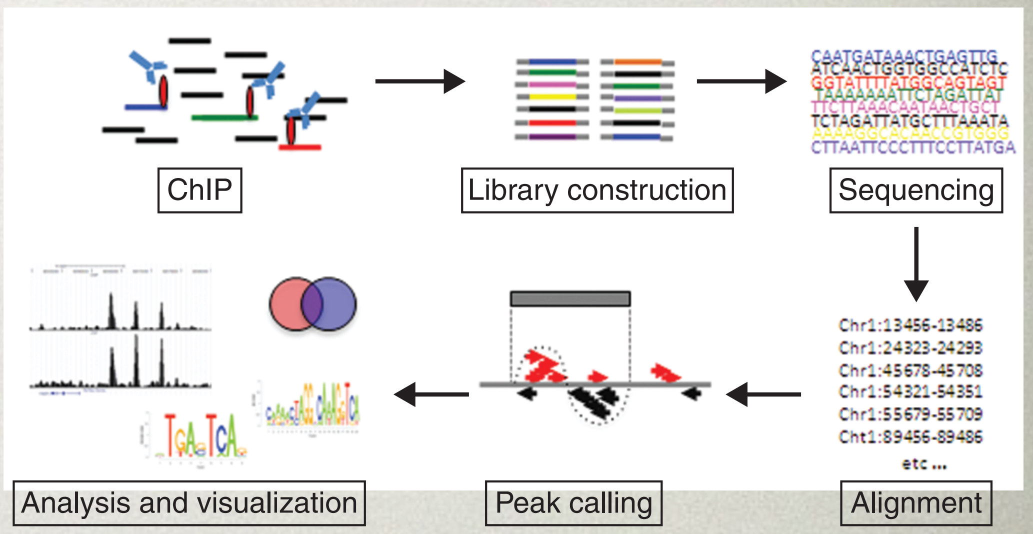


Figure 1. Flow scheme of the central steps in the ChIP-seq procedure.

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:

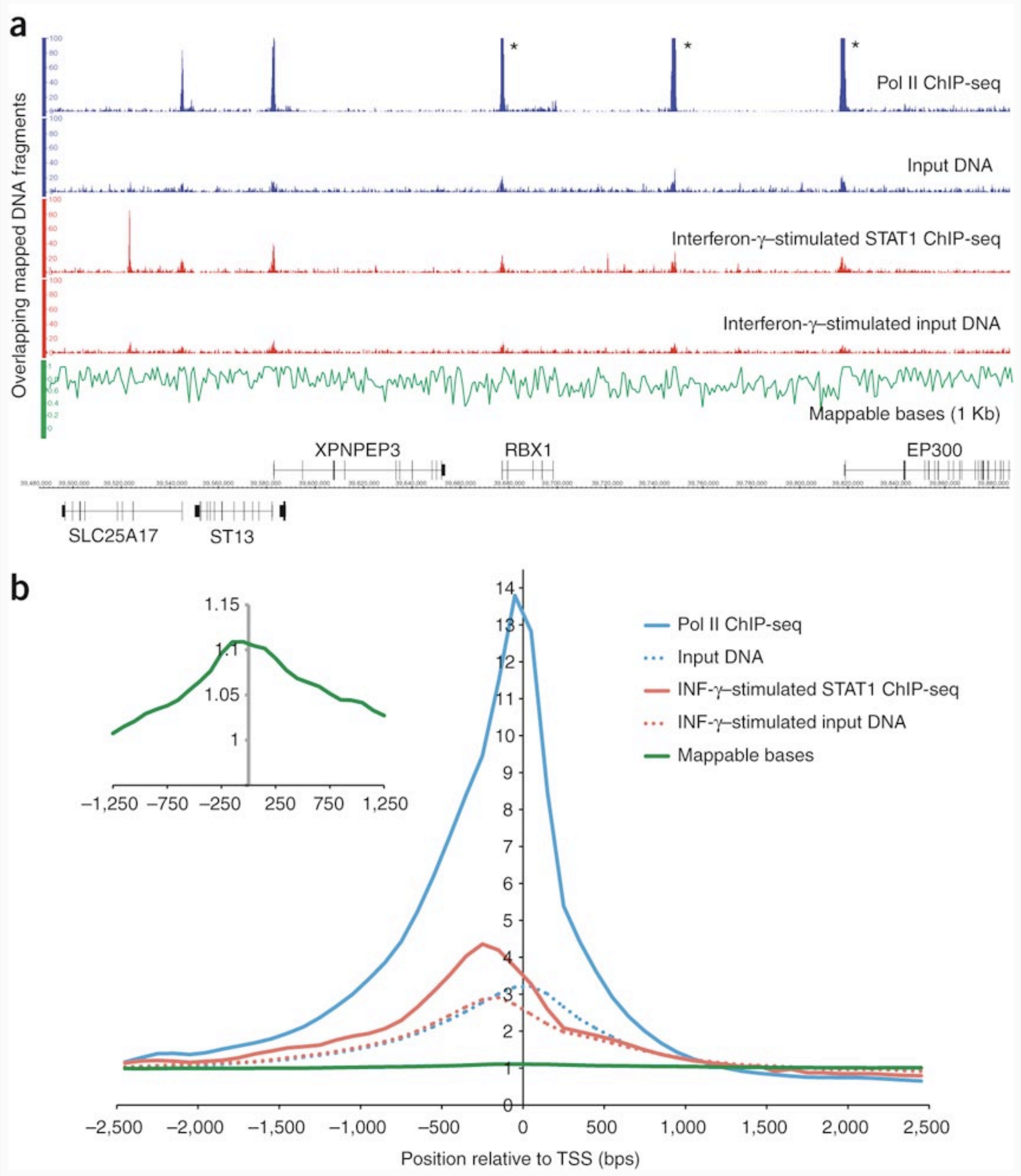
- input DNA (that is, DNA prior to immunoprecipitation, IP) [solubility, shearing, amplification]
- mock IP (treated the same as the IP but without any antibody) [low level of pull down DNA]
- non-specific IP (that is, using an antibody against a protein not known to be involved in DNA binding or chromatin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

Correction or Masking??



Rozowsky, Nature Biotechnology, 2009

Sequencing

There are three commonly used choices for this step:

- HiSeq
- MiSeq
- SOLID

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.

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

ENCODE Recommendations - Part I

- Antibody characterization -
  - Primary: immunoblot (cross-reactivity) and immunostain (location)
  - Secondary (any of the following validation methods)
    - Knockdown or knockout of the target protein
    - IP followed by mass spectrometry
    - IP with multiple antibodies against different parts of the target protein or members of the same complex
    - IP with an epitope-tagged version of the protein
    - Motif enrichment *(For ENCODE data to be submitted, motifs should be enriched at least fourfold compared with all accessible regions (e.g., DNase hypersensitive regions) and present in >10% of analyzed peaks)*

ENCODE Recommendations - Part II

ChIP experimental design guidelines

- **Sequencing and library complexity**
  - ➢ ENCODE's goal is to obtain ≥10 million uniquely mapping reads per replicate experiment
  - ➢ Target NRF (non-redundancy fraction) ≥0.8 for 10 million reads – *NRF is defined as the ratio between the Number of positions in the genome that unique reads map to / Total number of uniquely mappable reads*

- **Control libraries**
  - ➢ ENCODE generates and sequences a control ChIP library for each cell type, tissue, or embryo collection and sequences the library to the appropriate depth
  - ➢ Importantly, a new control is always performed if the culture conditions, treatments, chromatin shearing protocol, or instrumentation is significantly modified

- **Reproducibility**
  - ➢ Experiments are performed at least twice to ensure reproducibility
  - ➢ Concordance is determined from analysis using the IDR methodology (next slide)

ENCODE Recommendations – Part III

ChIP-seq quality assessment guidelines

- *A set of data quality thresholds established for submission of ChIP-seq data sets.*
  - *Balancing data quality with practical attainability*

1. **Cross-correlation analysis**
   - Calculate and report NSC and RSC for each experiment
   - The NSC (Normalized strand cross-correlation) and RSC (relative strand cross-correlation) metrics use cross-correlation of stranded read density profiles to measure enrichment independently of peak calling
   - If NSC values < 1.05 and RSC values < 0.8 → ENCODE recommends additional replicate be attempted or the experiment explained in the data submission

2. **Irreproducible discovery rate (IDR) -** established for mammalian cells - point source features
   - Biological replicates are performed for each ChIP-seq data set and subjected to peak calling
   - IDR analysis is then performed with a 1% threshold

ENCODE Recommendations - Part IV

Data reporting guidelines (similar to GEO)

1. **Metadata - minimal information**
   - Investigator, organism, or cell line, experimental protocol
   - Indication as to whether an experiment is a technical or biological replicate
   - Precise source of the antibody; Catalog and lot number for any antibody used
   - Information used to characterize the antibody

2. **Analysis Details**
   - Peak calling algorithm and parameters used, including threshold and reference genome used to map peaks
   - A summary of the number of reads and number of targets for each replicate and for the merged data set
   - Criteria that were used to validate the quality of the resultant ChIP-seq data (i.e., overlap results or IDR29)
   - Experimental validation results (e.g., qPCR) and link to the control track that was used
   - An explanation if the experiment fails to meet any of the standards.

3. **High-throughput sequencing data**
   - Raw data (FASTQ files) should be submitted to both GEO and SRA
   - Each replicate should be submitted independently
   - Target region and peak calling results

Data Analysis

analysis Pipeline

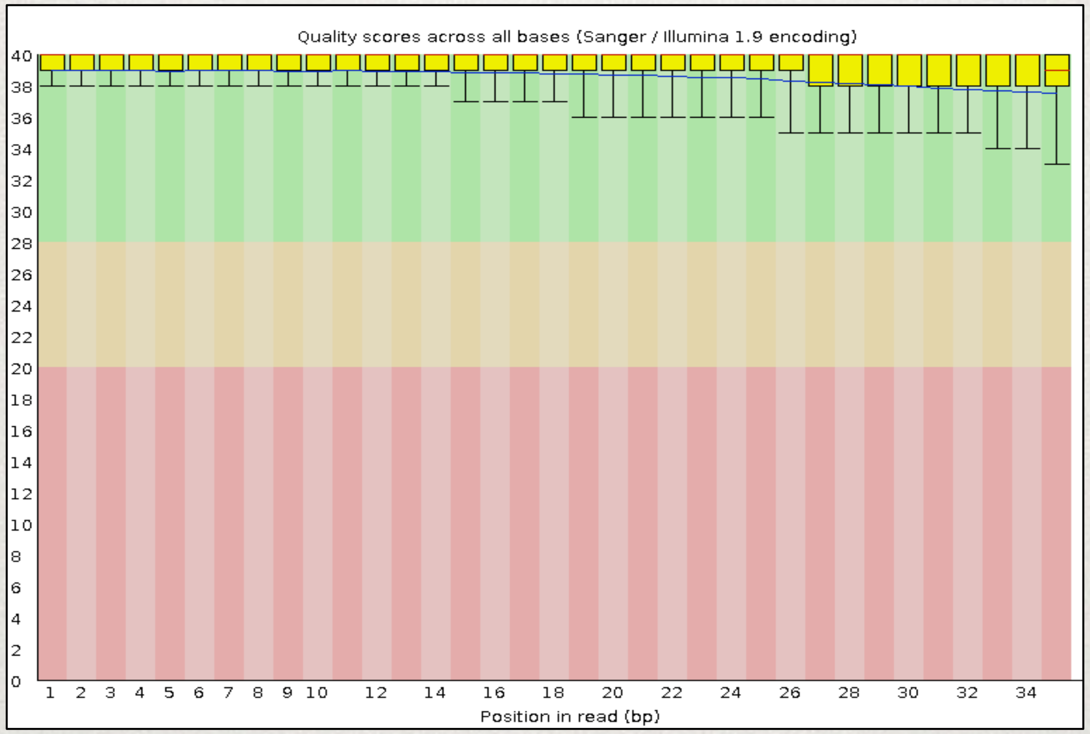
- Aligners
- Peak finders
- Motif finders
- GSEA
- Pathway analysis
- Differential effects
- Visualizers

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

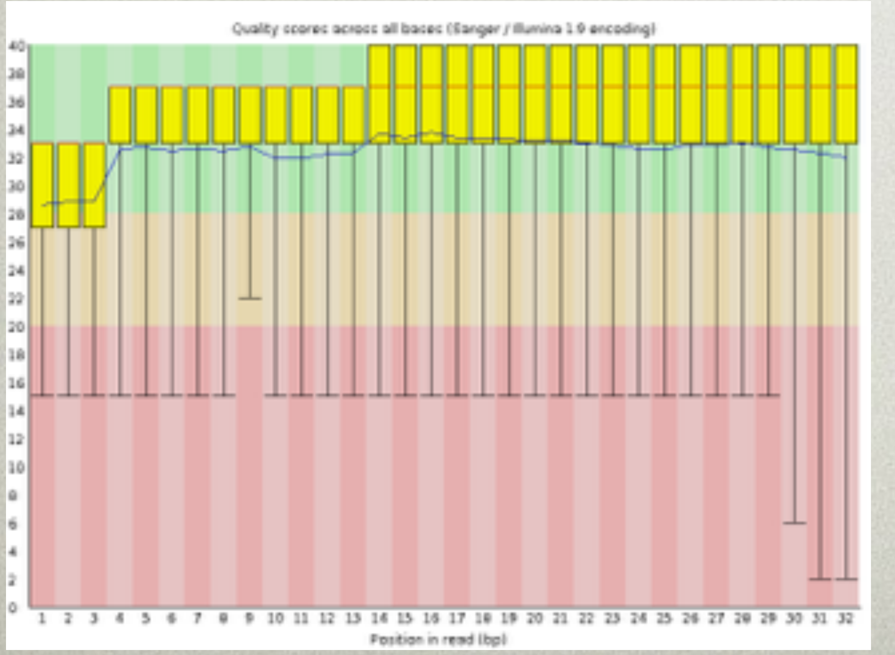
Good data is always more robust to analytical choices than poor data.

Read Quality

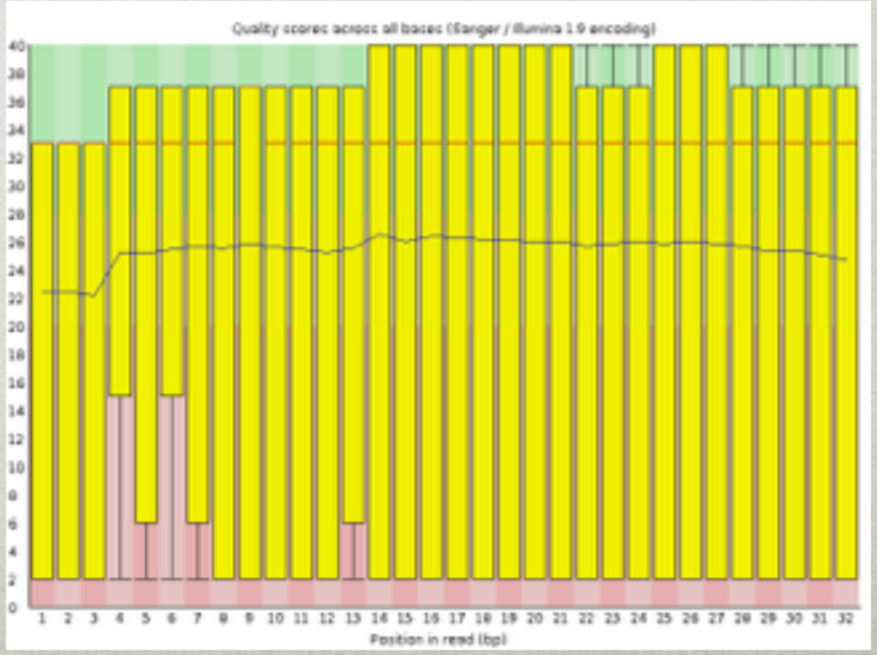
Great!



Okay



Bad!

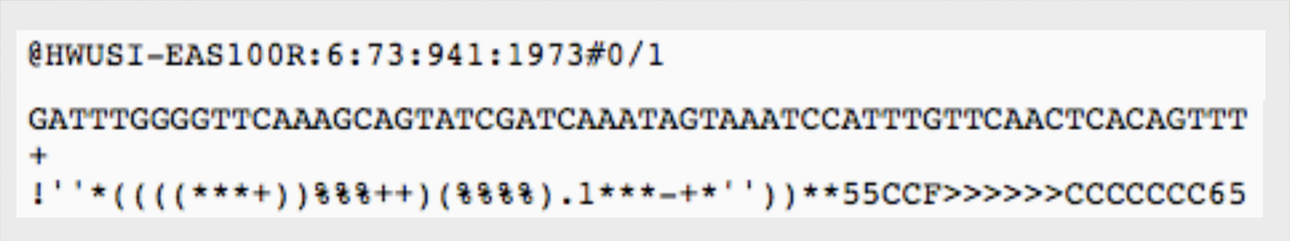


File Formats

File Formats

- Fastq
- SAM/BAM
- BED
- GFF/GTF
- WIG

http://genome.ucsc.edu/FAQ/FAQformat.html



```
@HWUSI-EAS100R:6:73:941:1973#0/1

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

File Formats

FASTA

```
>HWI-ST398_0092:1:1:5372:2486#0/1
TTTTTCGTTCTTTTCATGTACCGCTTTTTGTTCGGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
```

FASTQ

```
@HWI-ST398_0092:6:73:5372:2486#0/1
TTTTTCGTTCTTTTCATGTACCGCTTTTTGTTCGGTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAT
+HWI-ST398_0092:1:1:5372:2486#0/1
ffffeedfcedffffeffdefff_fffffdccfdZdeeadefecZedaecdbRdTY^ZYT``_T`_^bc_Wceaa[
```

6 - Flowcell lane

73 - Tile number

5372:2486 - 'x','y'-coordinates of the cluster within the tile
#0 - index number for a multiplexed sample (0 for no indexing)

/1 - the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

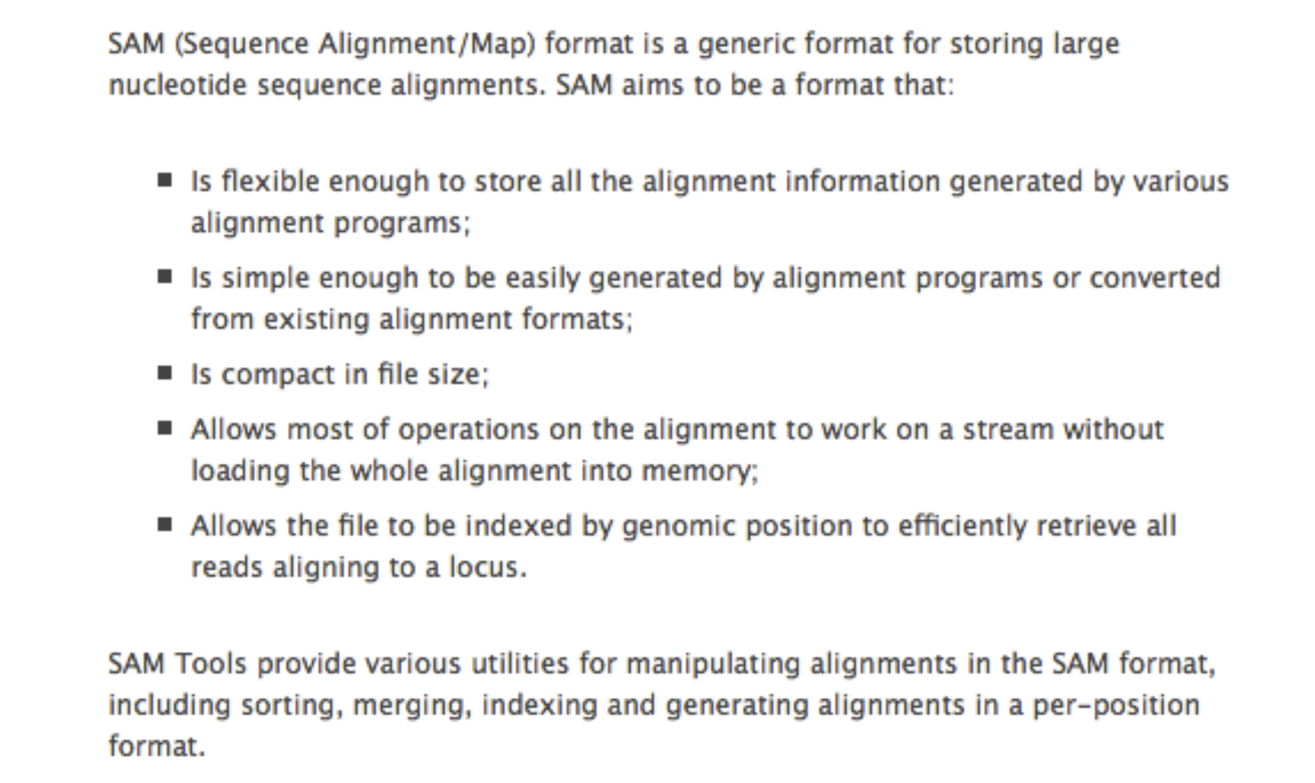
File Formats

FASTQ

Phred Quality Scores

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

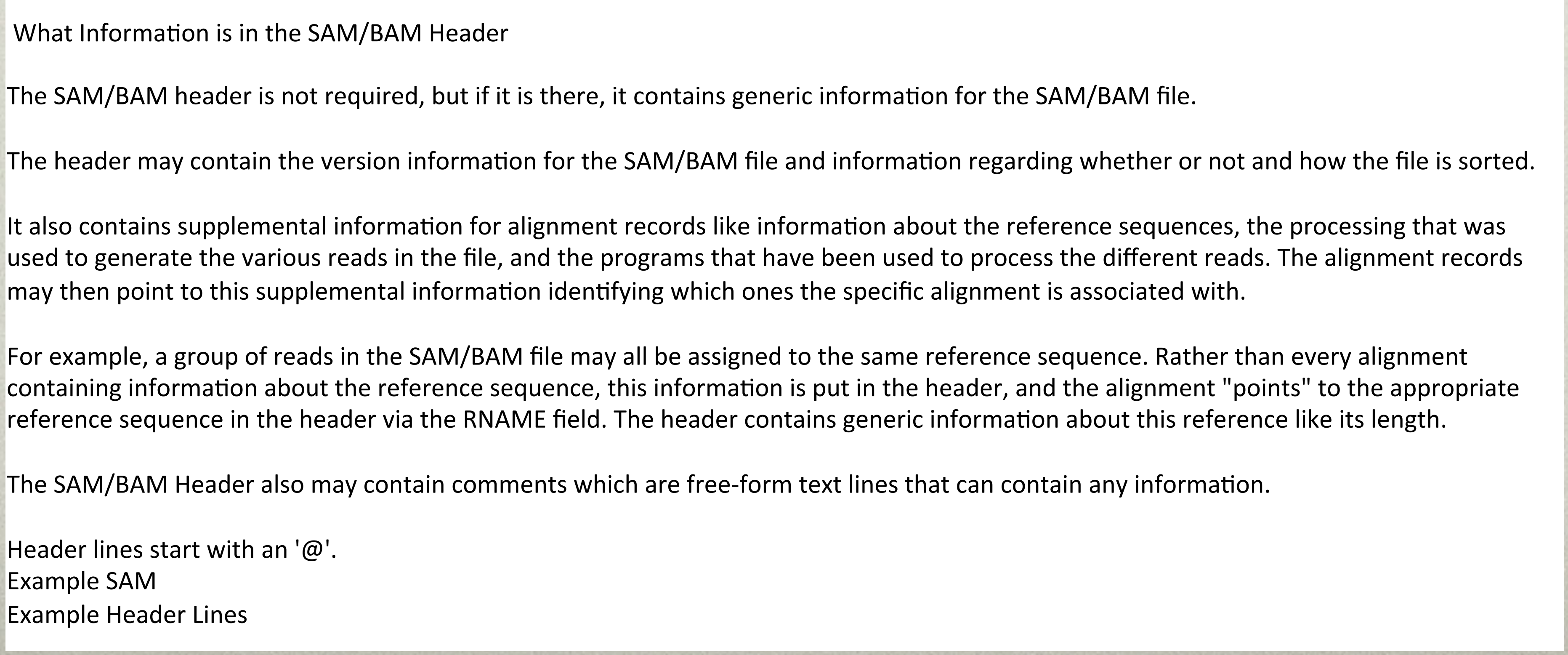
File Formats SAM



SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.



What Information is in the SAM/BAM Header

The SAM/BAM header is not required, but if it is there, it contains generic information for the SAM/BAM file.

The header may contain the version information for the SAM/BAM file and information regarding whether or not and how the file is sorted.

It also contains supplemental information for alignment records like information about the reference sequences, the processing that was used to generate the various reads in the file, and the programs that have been used to process the different reads. The alignment records may then point to this supplemental information identifying which ones the specific alignment is associated with.

For example, a group of reads in the SAM/BAM file may all be assigned to the same reference sequence. Rather than every alignment containing information about the reference sequence, this information is put in the header, and the alignment "points" to the appropriate reference sequence in the header via the RNAME field. The header contains generic information about this reference like its length.

The SAM/BAM Header also may contain comments which are free-form text lines that can contain any information.

Header lines start with an '@'.
Example SAM
Example Header Lines

File Formats- SAM

```
8_100_10000_12419     163    chrVII  271183  255    40M    =      271294  151    TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG      bbbabbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb      XA:i:0  MD:Z:40 NM:i:0
```

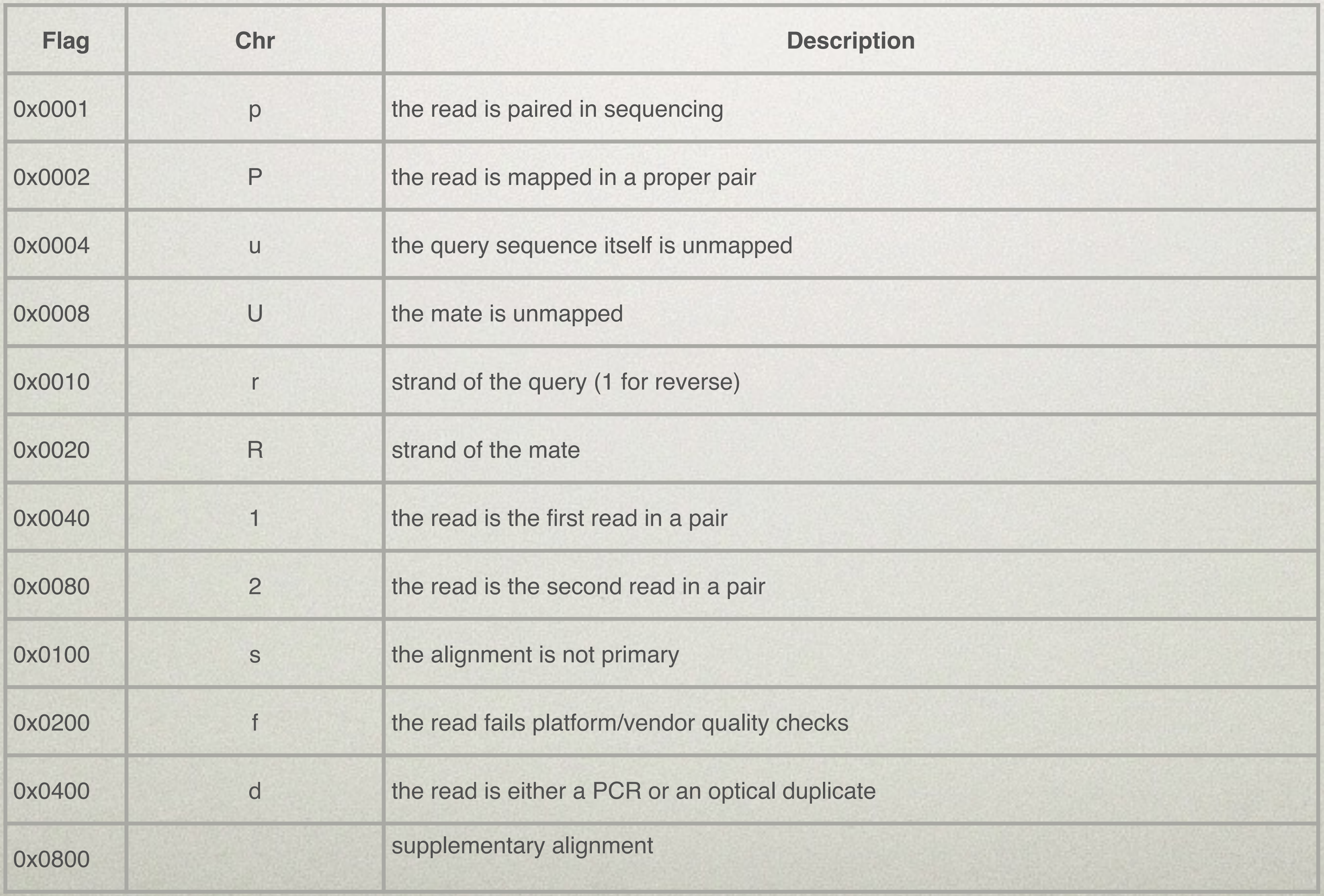
8_100_10000_12419	163	chr7	271183	255	40M	=	271294	151	TGGTGTATTAT ACGCTACCGT	bbbabbbbbbbbb bbbbbbbcbbbbc	XA:i:0 MD:Z:40 NM:i:0
QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	TLEN	SEQ	QUAL	OPT

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe (‘=’ if same as RNAME)
8	MPOS	1-based Mate POSistion
9	TLEN	inferred Template LENgth (insert size)
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

File Formats- SAM

```
8_100_10000_12419      163    chrVII  271183  255     40M     =      271294  151     TGGTGTATTATACGCTACCGTGCGGTGCCGGGGGCAACCG       bbbabbbbbbbbbbbbbbbbcbbbbcbbbbbbbbbbbbbb       XA:i:0  MD:Z:40 NM:i:0
```

http://picard.sourceforge.net/explain-flags.html



Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate
0x0800		supplementary alignment

File Formats BAM

BAM is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments. **BAM** is compressed in the **BGZF** format. BGZF files support random access through the BAM file index.

***BGZF** is block compression implemented on top of the standard gzip file format. The goal of BGZF is to provide good compression while allowing efficient random access to the BAM file for indexed queries. The BGZF format is 'gunzip compatible', in the sense that a compliant gunzip utility can decompress a BGZF compressed file.*

File Formats BED

BED files are tab delimited text files BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track.

The first three required BED fields are: (UCSC-definitions)

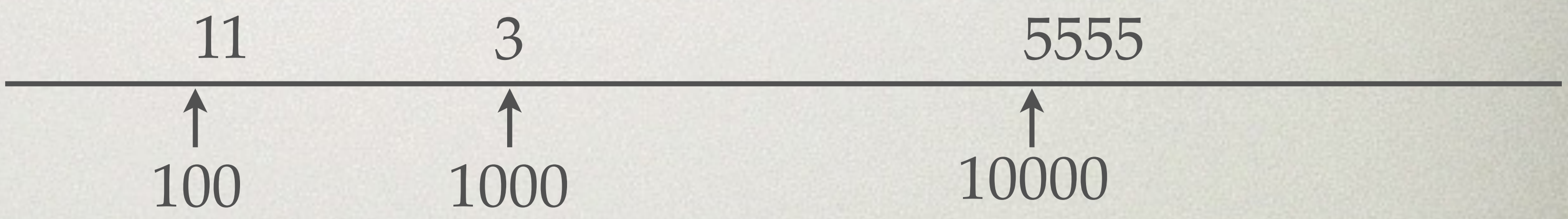
1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold.
   The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.
4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score (between 0 and 1000).
6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RBG value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to *blockCount*.
12. **blockStarts** - A comma-separated list of block starts. All of the *blockStart* positions should be calculated relative to *chromStart*. The number of items in this list should correspond to *blockCount*.

File Formats WIG

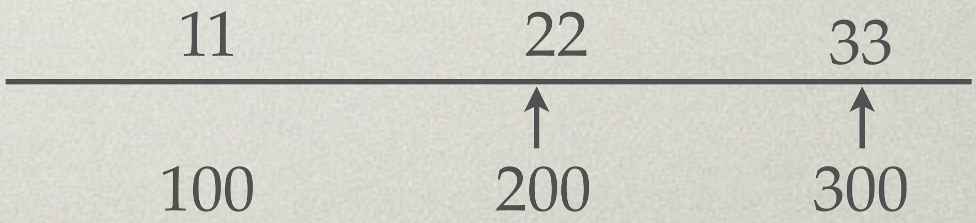
Line oriented text file with two options:

- Variable step
- Fixed step

```
variableStep  chrom=chr1 span=2
100 1
variableStep  chrom=chr1 span=1
1000 3
variableStep  chrom=chr1 span=4
10000 5
```



```
fixedStep  chrom=chr1 start=100 step=100 span=2
1
2
3
```



File Formats GFF/GTF

- GFF (General Feature Format)
- GTF (Gene Transfer Format)

1. **seqname** - The name of the sequence. Must be a chromosome or scaffold.
2. **source** - The program that generated this feature.
3. **feature** - The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
4. **start** - The starting position of the feature in the sequence. The first base is numbered 1.
5. **end** - The ending position of the feature (inclusive).
6. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). If there is no score value, enter ".".
7. **strand** - Valid entries include '+', '-', or '.' (for don't know/don't care).
8. **frame** - If the feature is a coding exon, *frame* should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
9. **group** - All lines with the same group are linked together into a single item.

GTF is a refined form of the GFF with group attributes

- **gene_id *value*** - A globally unique identifier for the genomic source of the sequence.
- **transcript_id *value*** - A globally unique identifier for the predicted transcript.

GFF3 http://www.sequenceontology.org/resources/gff3.html

Mapping

Mapping
Which Genome Version?

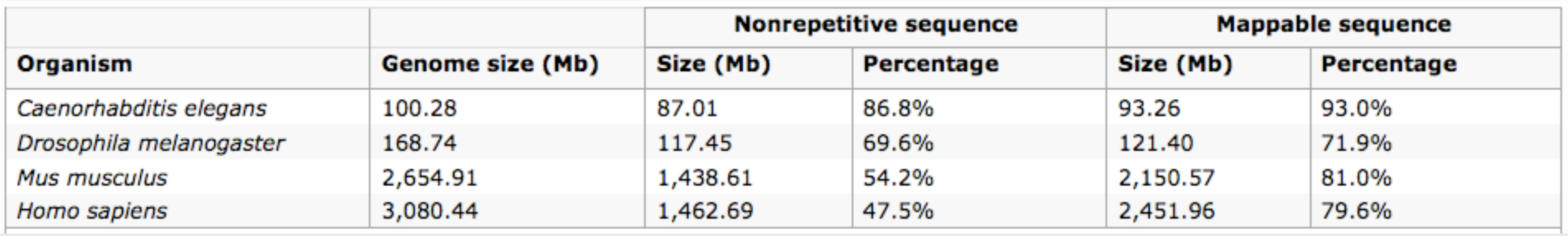
- Which version of the genome do you **want/need** to use. (*Record and report it!!*)

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

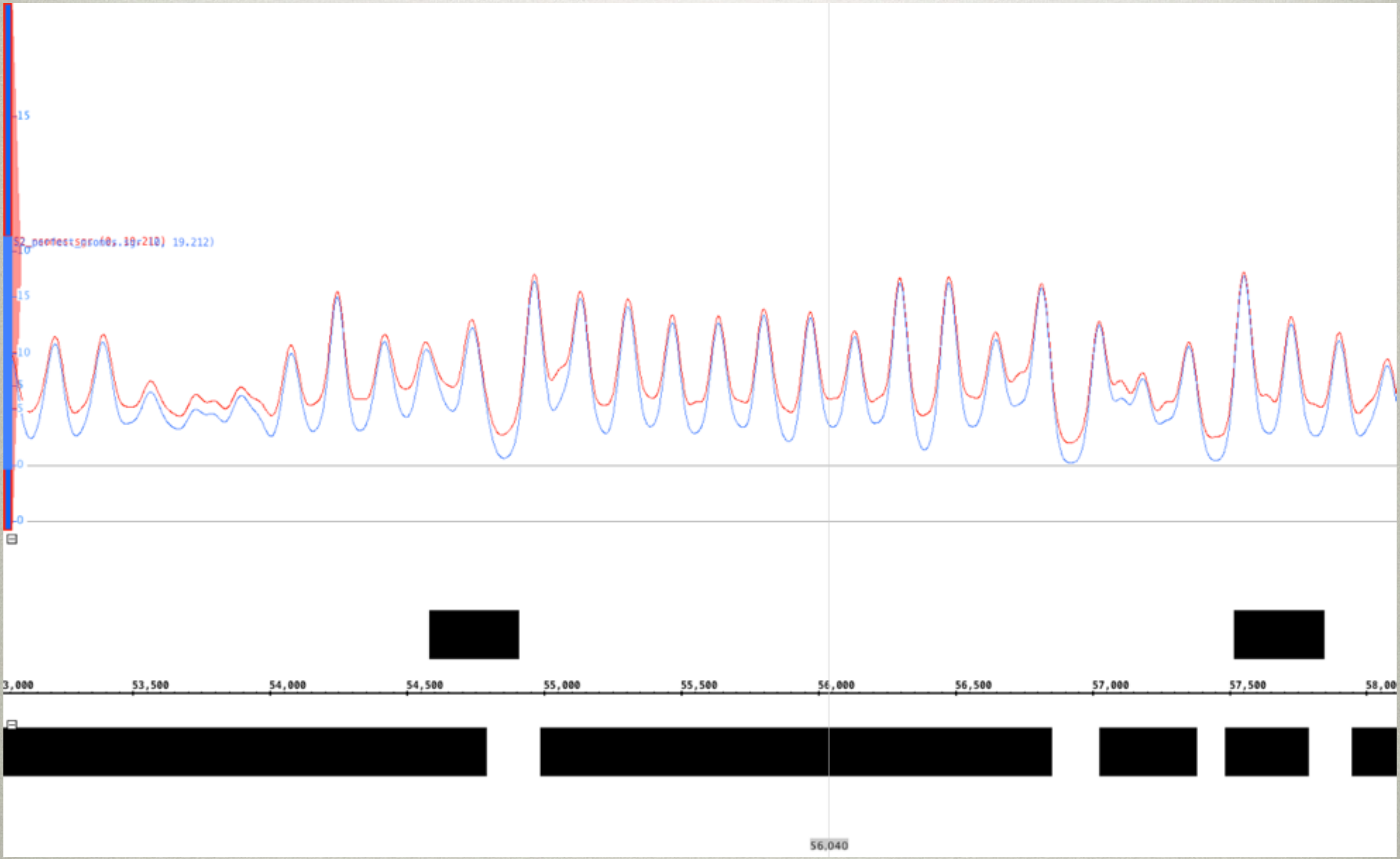
Mapping

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).

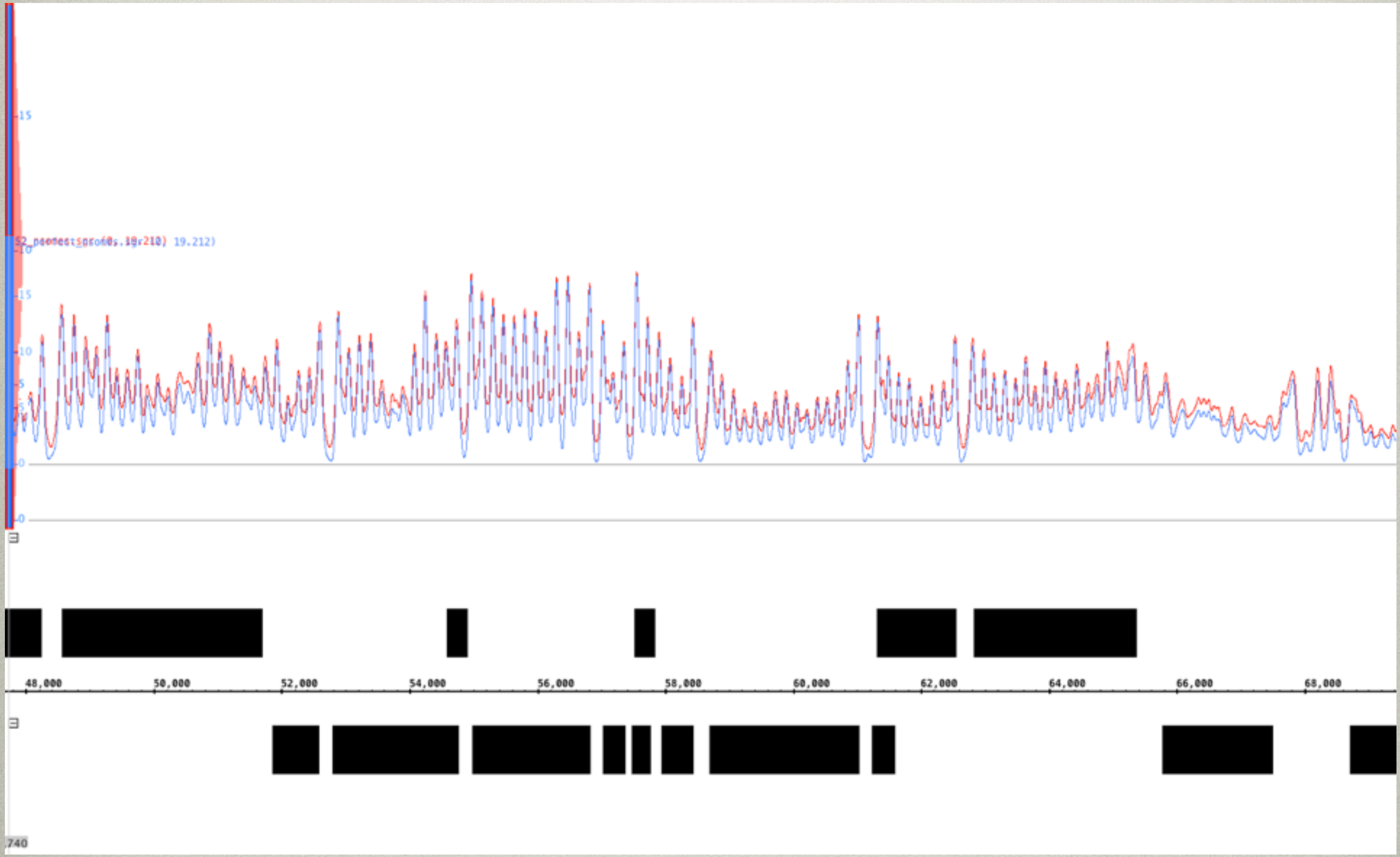
Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome 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)

Mapping Quality



Mapping Quality

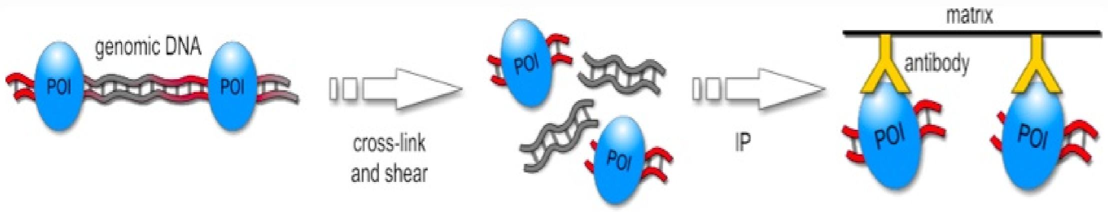


PEAK-Calling

Peak Calling

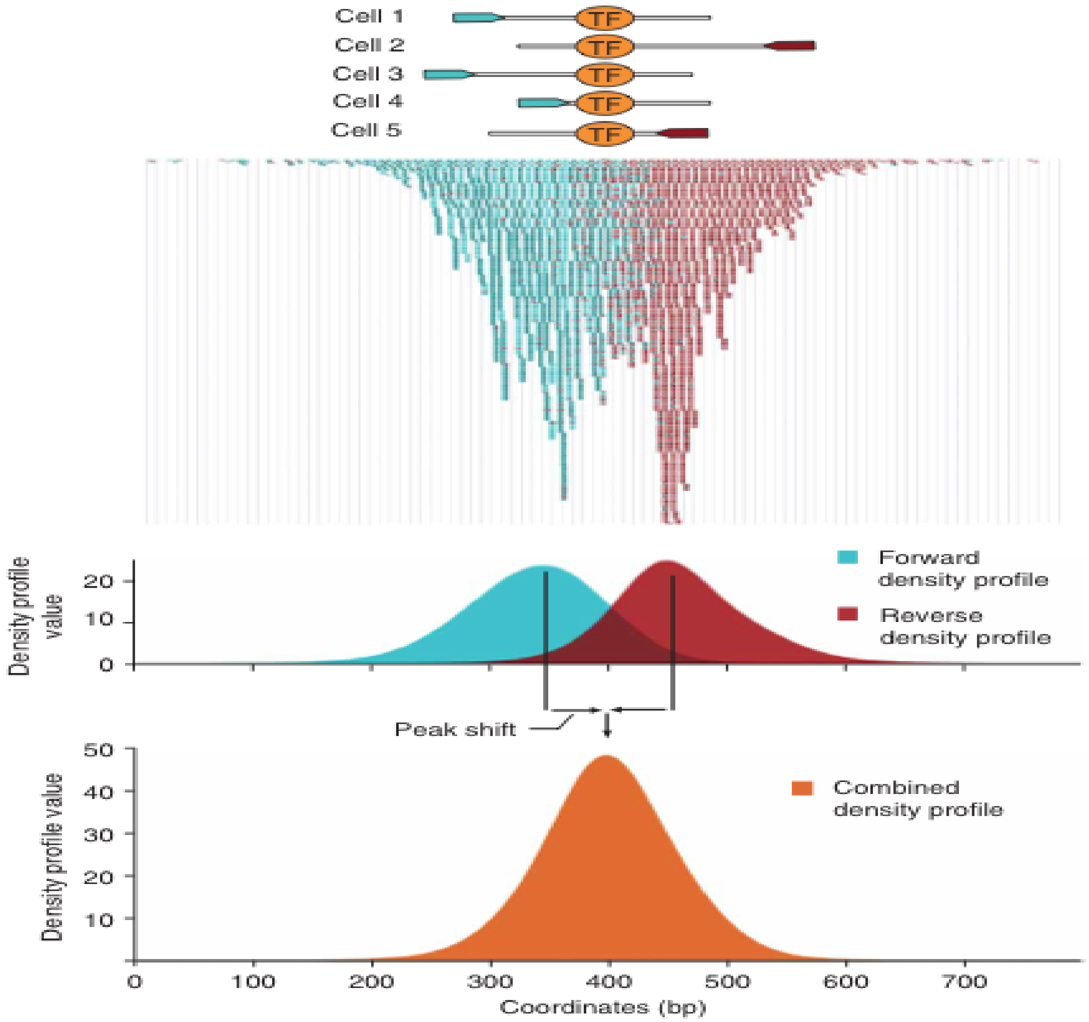
What is the ultimate goal of peak calling?

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
Density profile value
20
10
0
Forward density profile
Reverse density profile
Peak shift
Density profile value
50
40
30
20
10
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

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

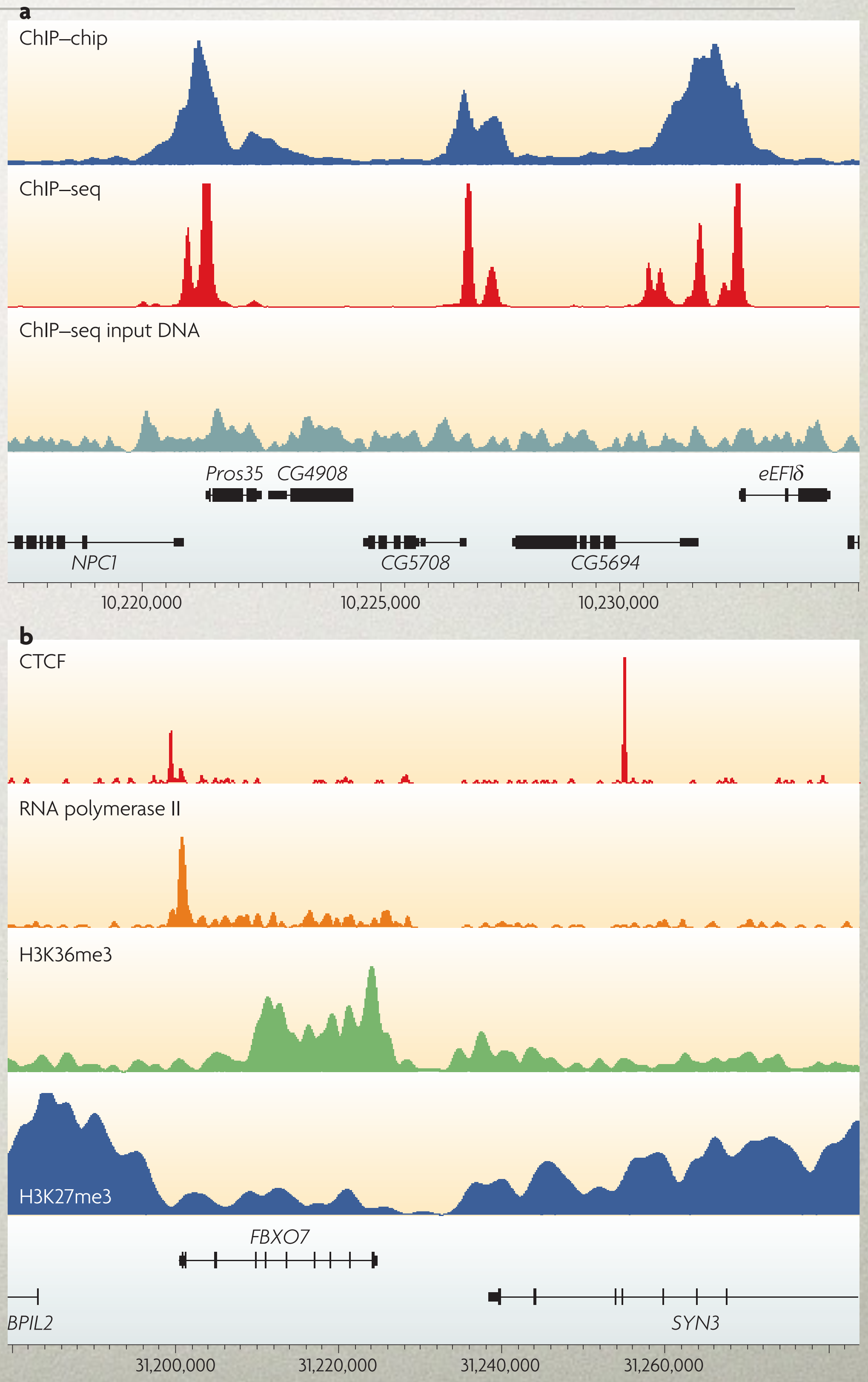
Figure 2 | chiP profiles. a | Examples of the profiles generated byNcahtruorme Raetivnieiwmsml uGneonpetriec-s cipitation followed by sequencing (ChIP–seq) or by microarray (ChIP–chip). Shown is a section of the binding profiles of the chromodomain protein Chromator, as measured by ChIP–chip (unlogged 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.

Sharp

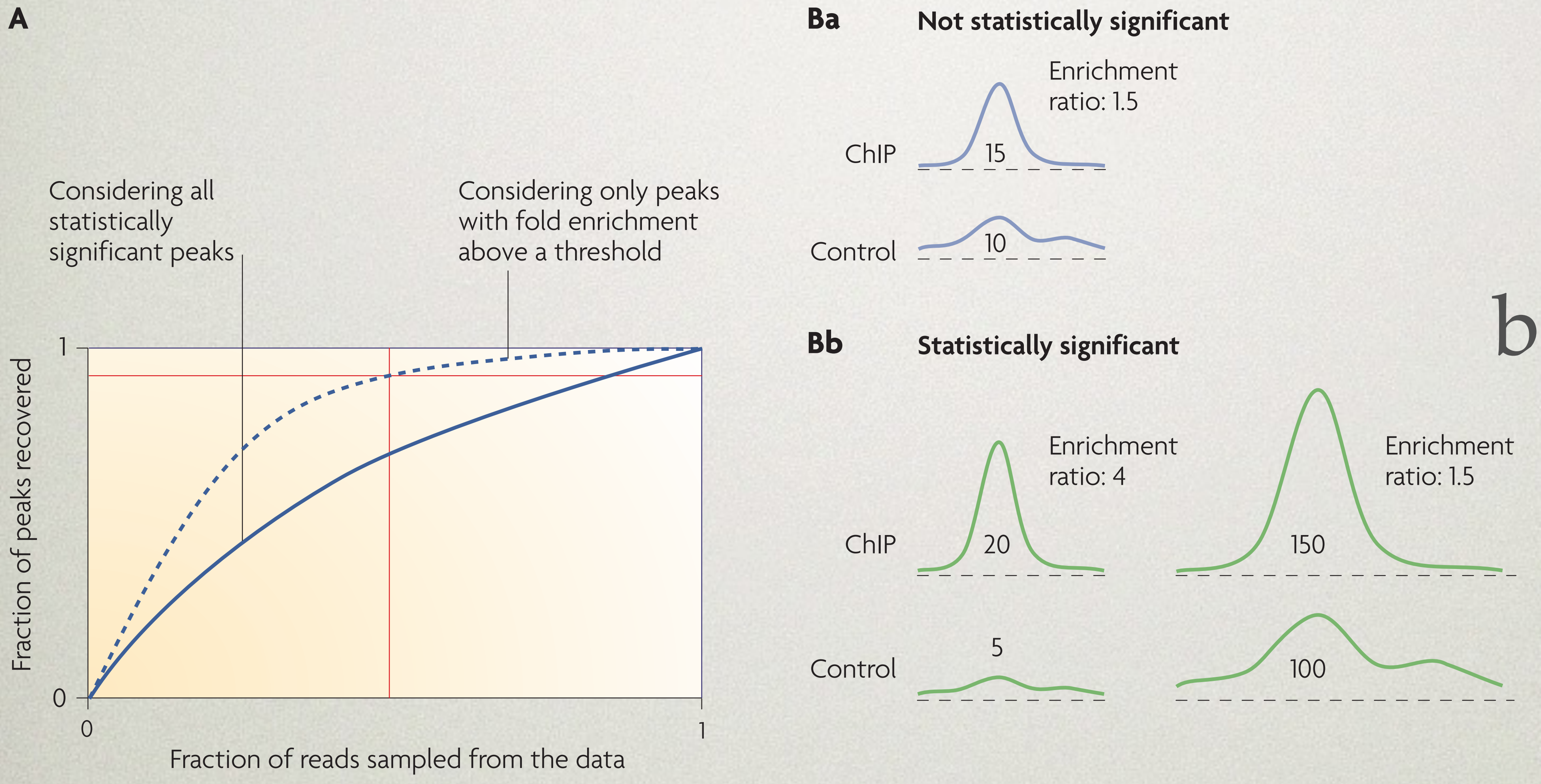
Mixed

Medium

Broad



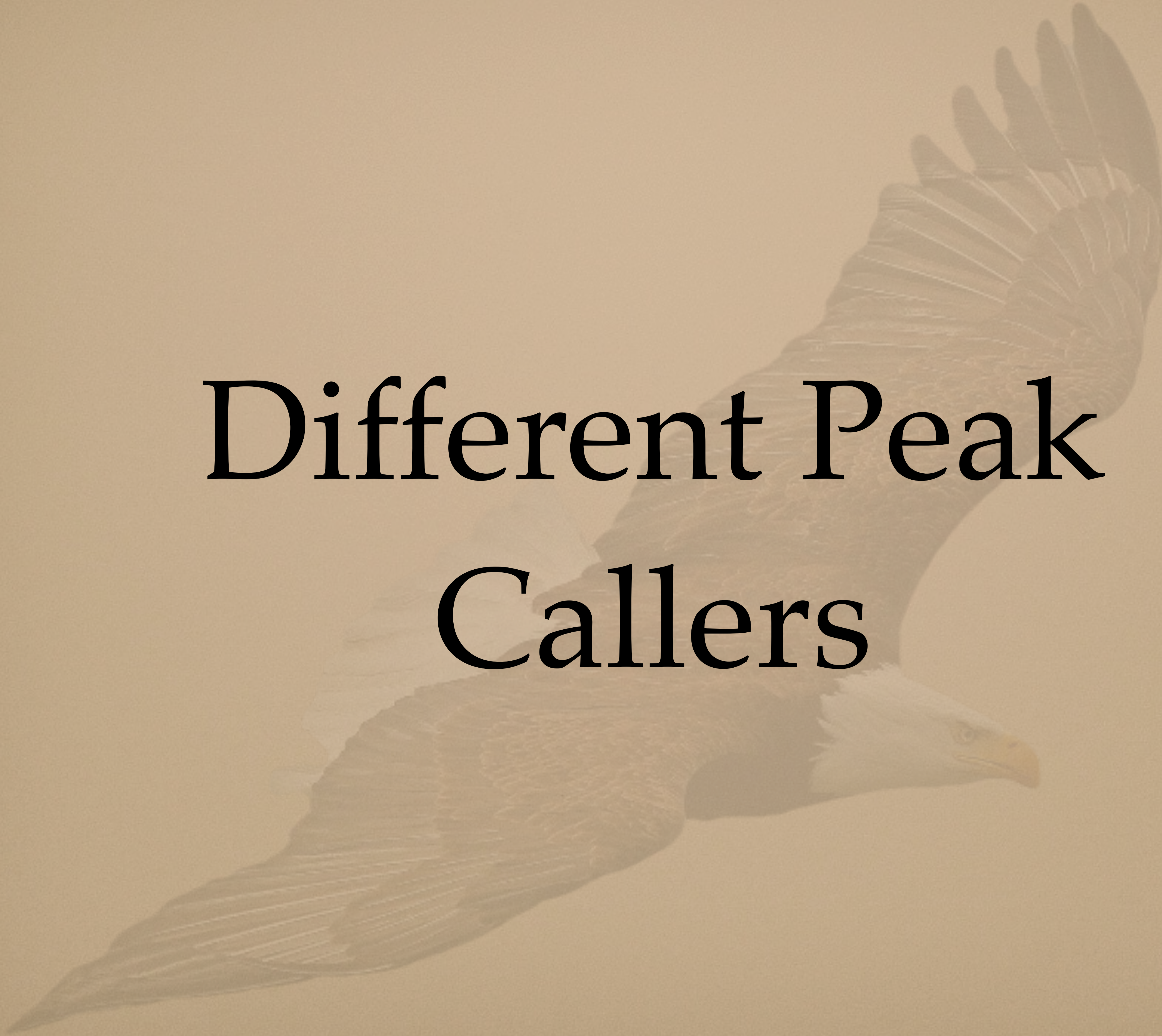
Types of Peaks



But is it biologically relevant

Figure 3 I Depth of sequencing. A I 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
Nature Reviews I Genetics 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 I 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 I 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

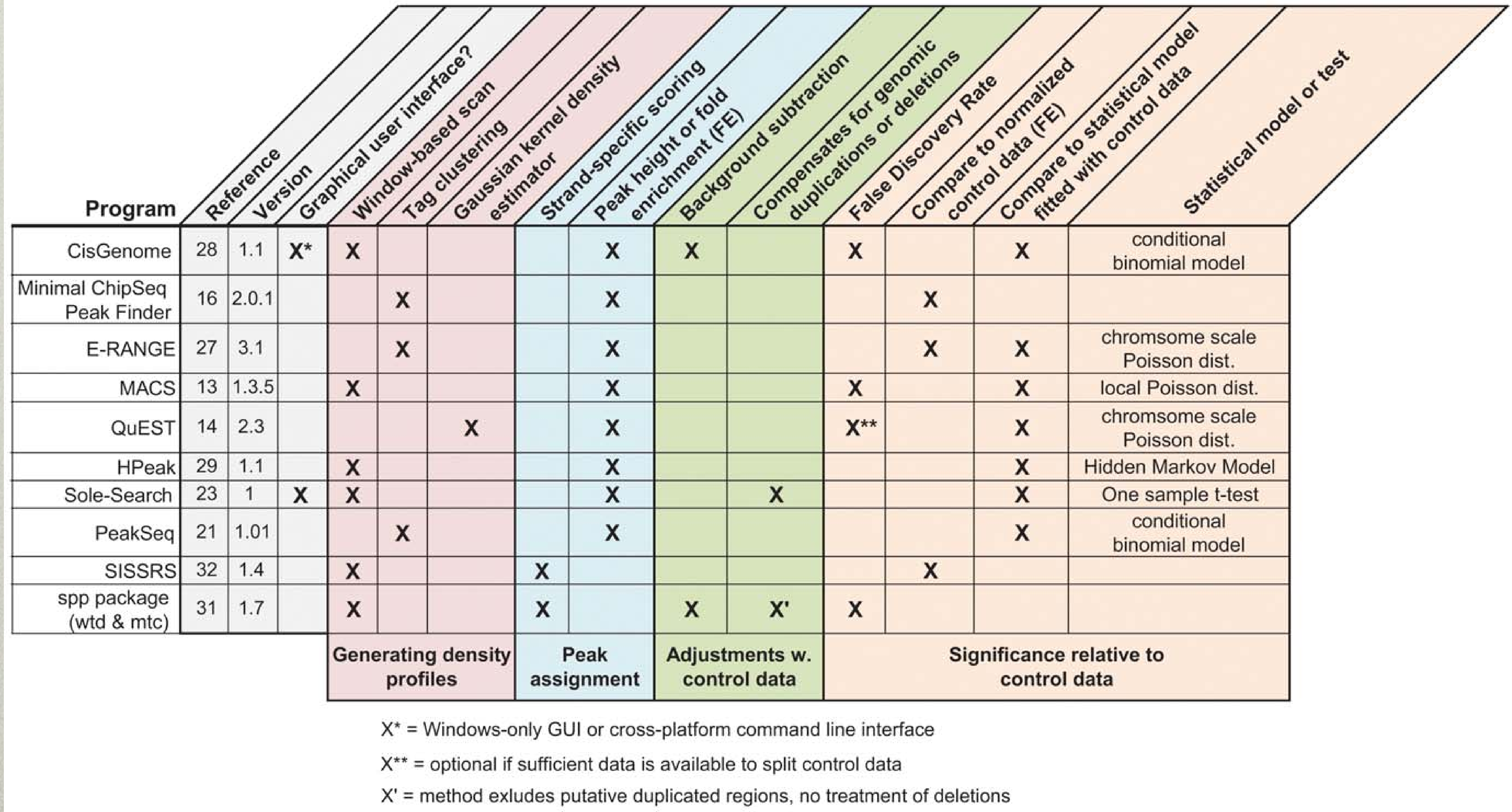
- Potentially the most critical, especially for "poor quality experiments"

- **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

Testing of ChIP-Seq Algorithms



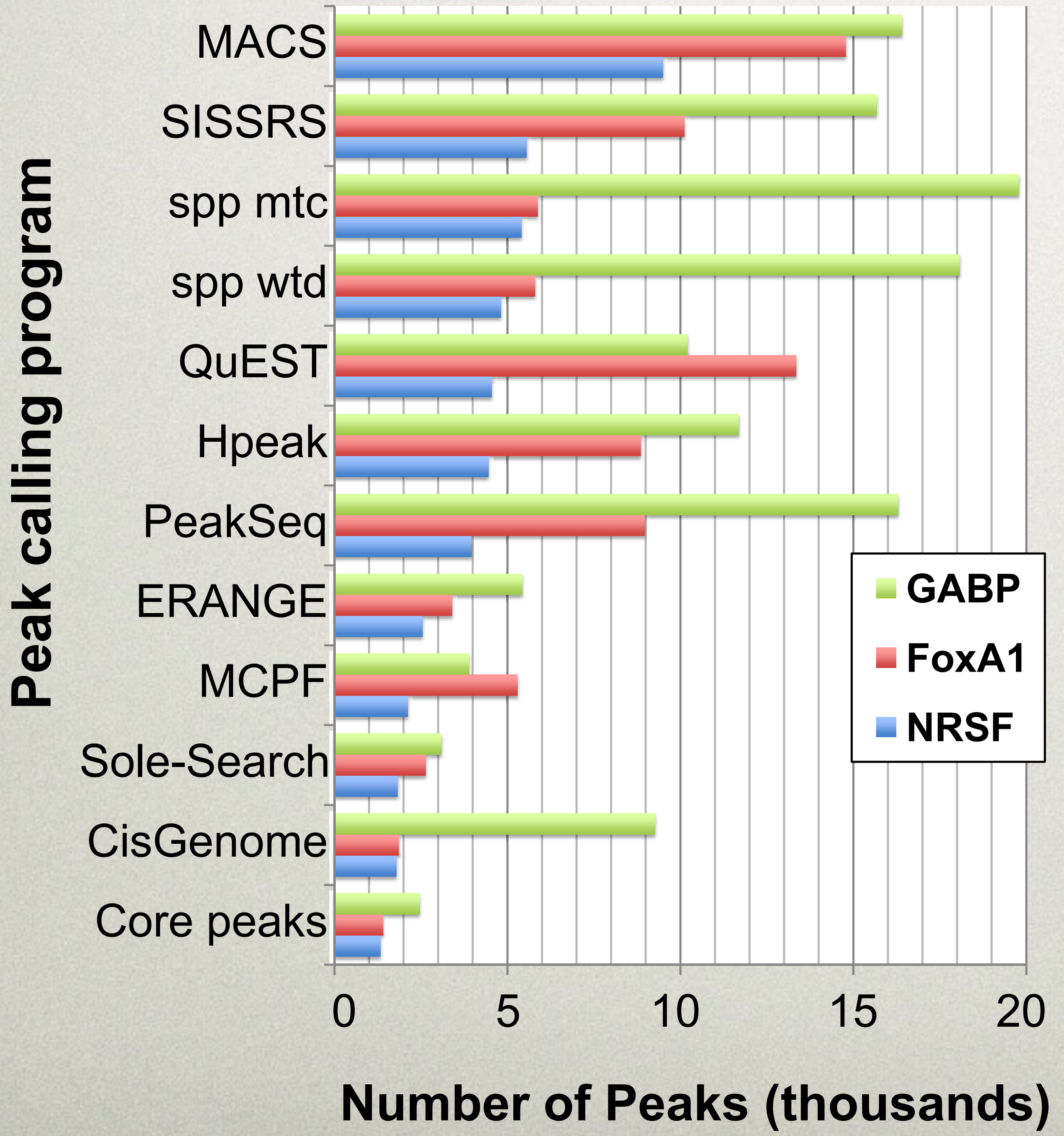
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			
Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background 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

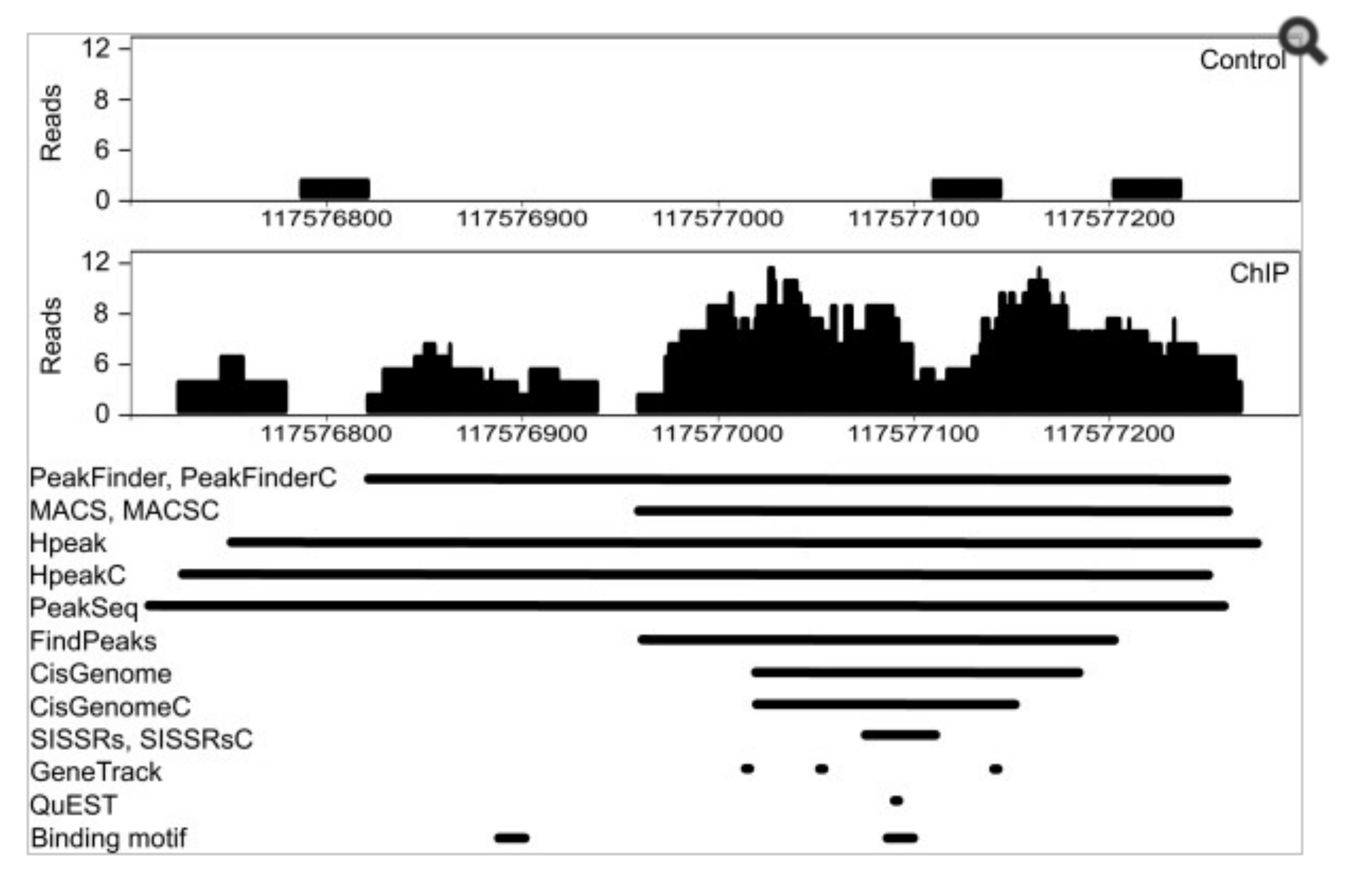
X** = optional if sufficient data is available to split control data

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

Peak Calling Bias



Peak Calling



Imported Author Today, 3:18 PM
While some packages simply aggregate mapped tags without regard to strand, others use strand information to locate the peaks more sensitively. Some peak-calling algorithms require the user to supply a control library whereas others can work without one, but there are several known sources of bias in sequencing reads with ChIP-seq, so that the estimation of confidence in the peaks without a control library is highly unreliable and should be avoided [6]. Confidence in the peaks is quantified using measures such as P-value

Imported Author Today, 3:18 PM
STAT6
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2804666/figure/F5/

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

ChIP-seq Peak Callers

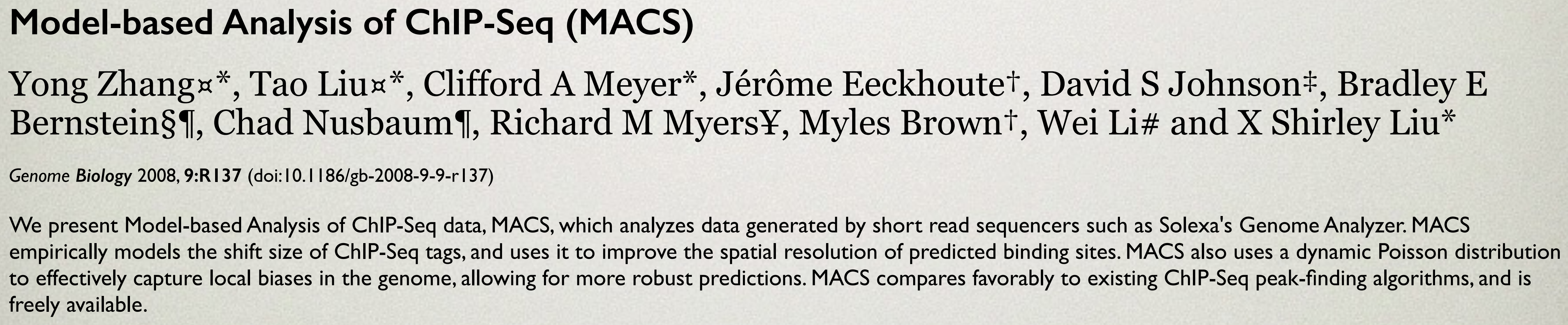
MACS
A widely-used, fast, robust ChIP-seq peak-finding algorithm that accounts for the offset in forward-strand and reverse-strand reads to improve resolution and uses a dynamic **Poisson distribution** to effectively capture local biases in the genome. MACS 1.4 is being used for the current uniform 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 packag**e, 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

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

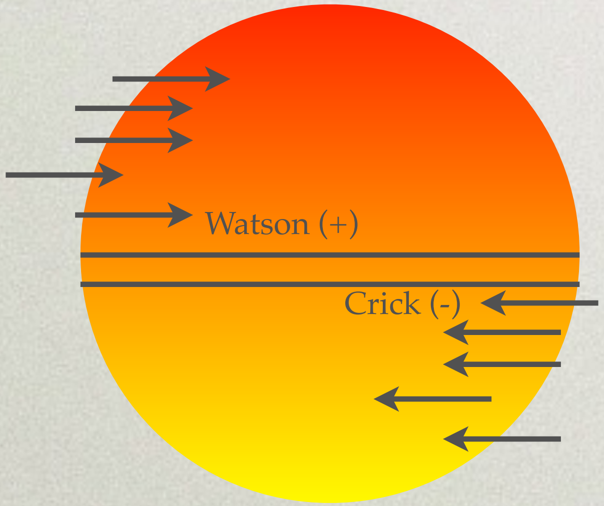
- Most widely used
- Robust, provided your data fits the model
- Ignores PCR artifacts
- Does NOT do much QC for you (*garbage in garbage out*)
- Python based - many dependencies
- Availability:Helix/Biowulf, Genomatix and Galaxy
- Two common versions (1.4.2 and 2.0.10)

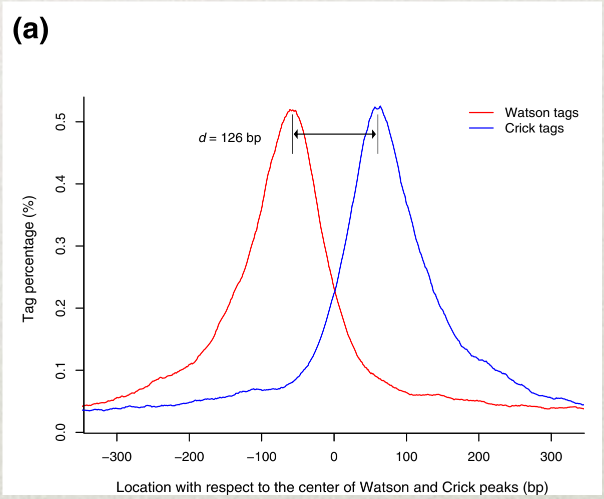
MACS

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:

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

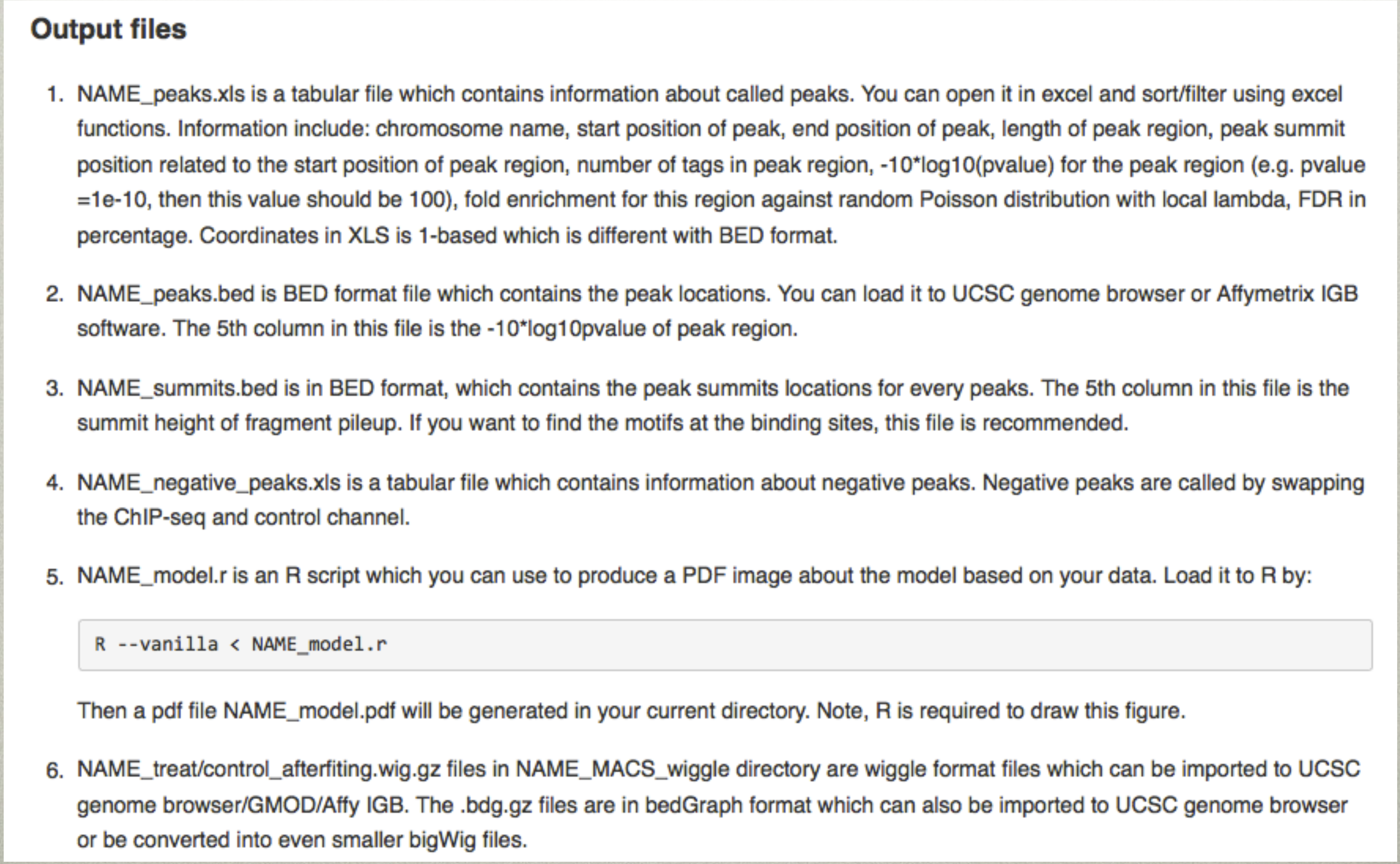
where λ_{1k}, λ_{5k} and λ_{10k} are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample...where no control sample available then λ_{1k} is not used.

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 files

1. NAME_peaks.xls is a tabular file which contains information about called peaks. You can open it in excel and sort/filter using excel functions. Information include: chromosome name, start position of peak, end position of peak, length of peak region, peak summit position related to the start position of peak region, number of tags in peak region, -10*log10(pvalue) for the peak region (e.g. pvalue =1e-10, then this value should be 100), fold enrichment for this region against random Poisson distribution with local lambda, FDR in percentage. Coordinates in XLS is 1-based which is different with BED format.

2. NAME_peaks.bed is BED format file which contains the peak locations. You can load it to UCSC genome browser or Affymetrix IGB software. The 5th column in this file is the -10*log10pvalue of peak region.

3. NAME_summits.bed is in BED format, which contains the peak summits locations for every peaks. The 5th column in this file is the summit height of fragment pileup. If you want to find the motifs at the binding sites, this file is recommended.

4. NAME_negative_peaks.xls is a tabular file which contains information about negative peaks. Negative peaks are called by swapping the ChIP-seq and control channel.

5. NAME_model.r is an R script which you can use to produce a PDF image about the model based on your data. Load it to R by:

   ```
   R --vanilla < NAME_model.r
   ```

   Then a pdf file NAME_model.pdf will be generated in your current directory. Note, R is required to draw this figure.

6. NAME_treat/control_afterfiting.wig.gz files in NAME_MACS_wiggle directory are wiggle format files which can be imported to UCSC genome browser/GMOD/Affy IGB. The .bdg.gz files are in bedGraph format which can also be imported to UCSC genome browser or be converted into even smaller bigWig files.

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)

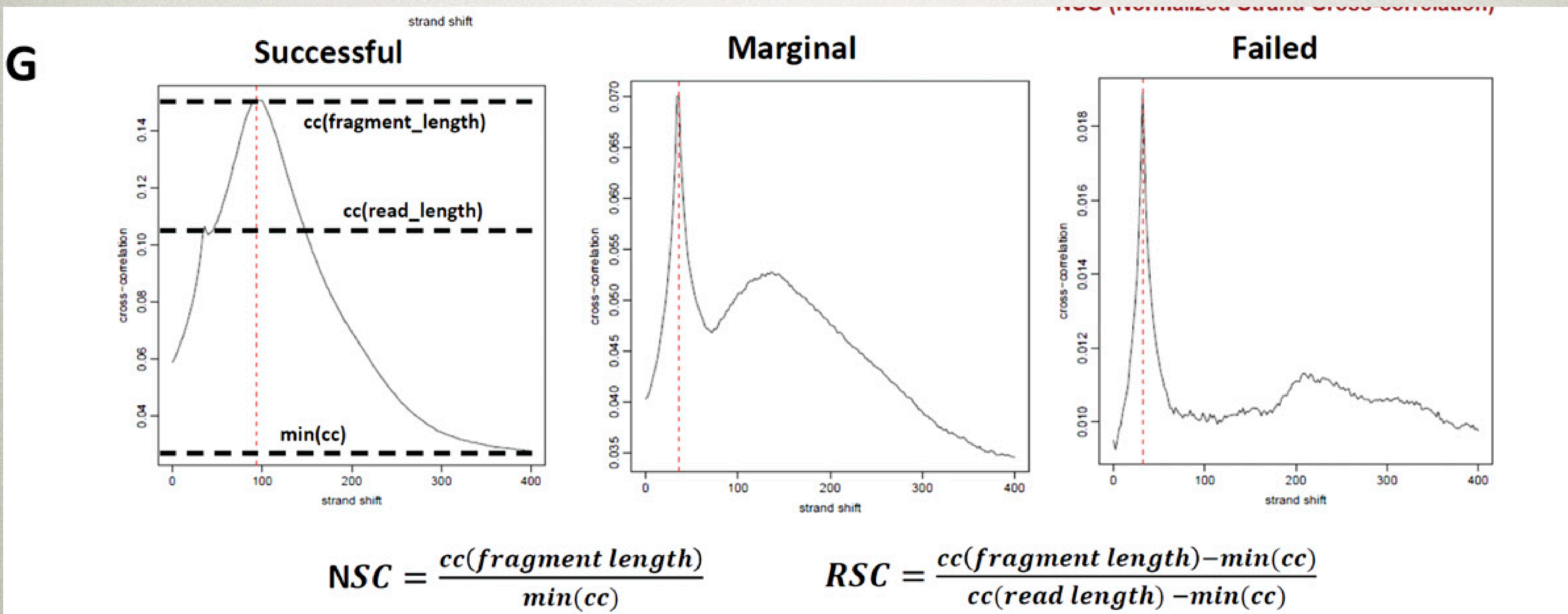
In layman's terms, the IDR method compares a pair of ranked lists of identifications (such as ChIP-seq peaks). These ranked lists should not be pre-thresholded i.e. they should provide identifications across the entire spectrum of high confidence/enrichment (signal) and low confidence/enrichment (noise). 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.

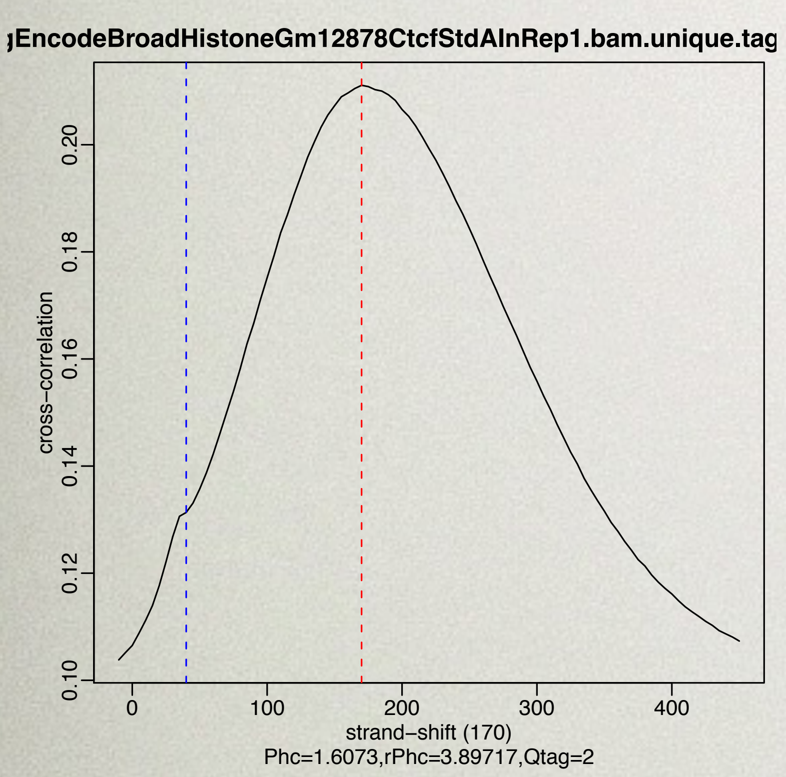
FRiP - Fraction of reads in the Peaks
NSC - Normalized Strand Correlation
RSC - Relative Strand Correlation

QC of Output (encode)

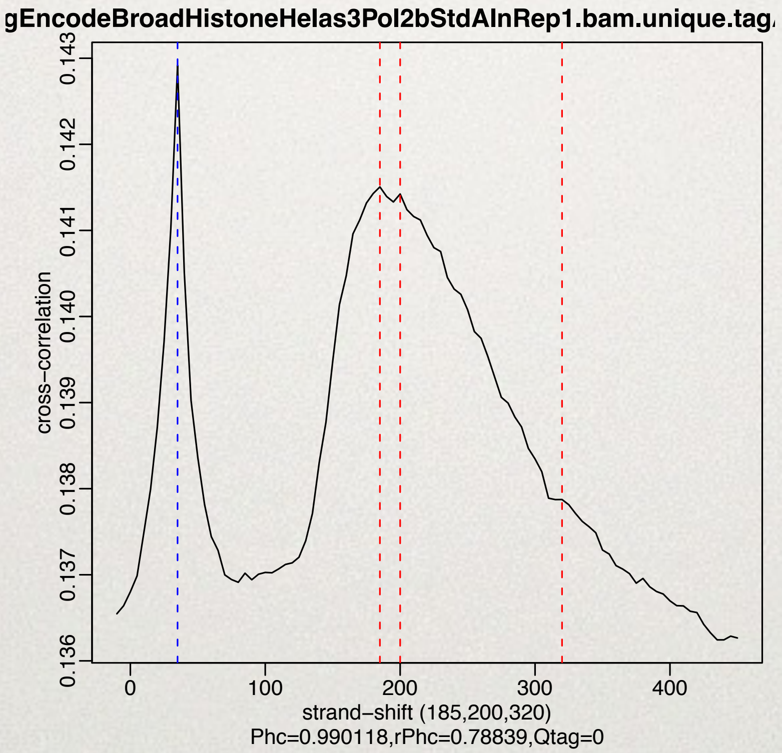


$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

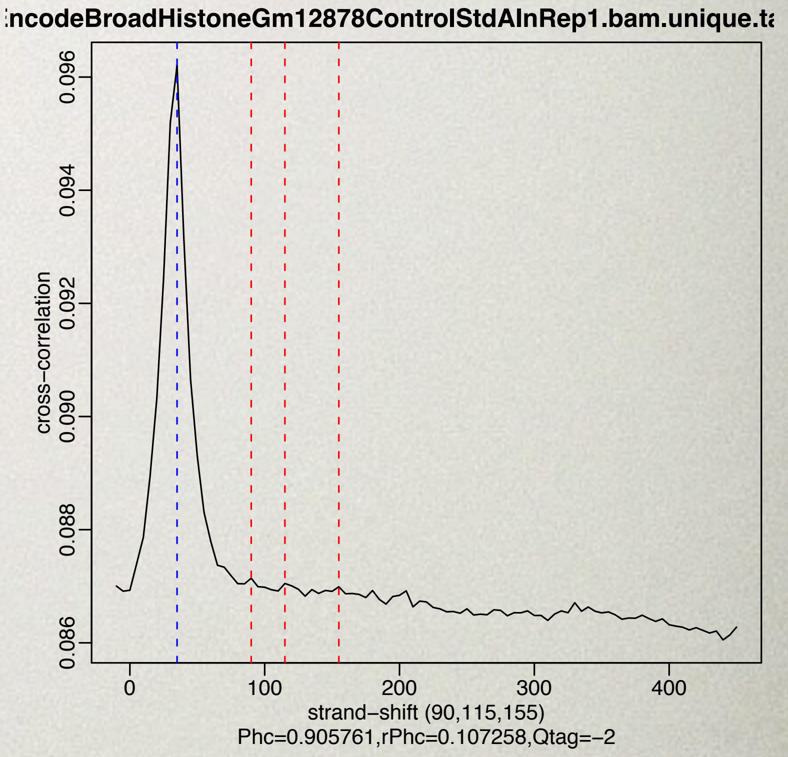
Cross Correlation Plots



Good



Poor

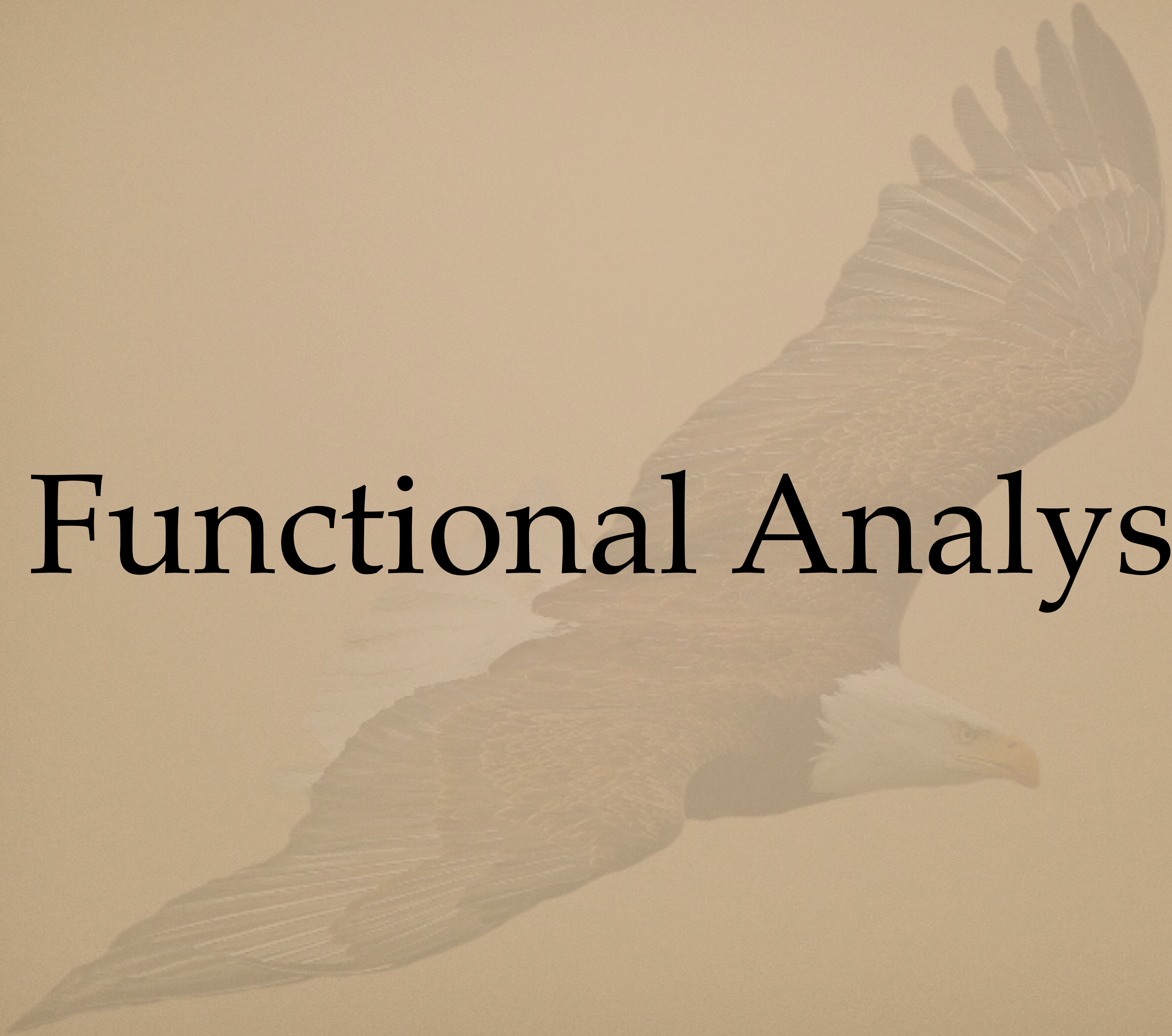


Input

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



Function 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
- Correlation with SNP data to find allele-specific binding

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

Replicates/Controls

Replicates

- Nature of the biological sample
  - Cell line vs Tissue

Controls

- Comparative studies
- Time courses
- Cancer vs Normal

ENCODE ChIP-Seq peaks are screened against a specially curated empirical blacklist of regions in the human genome and 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.

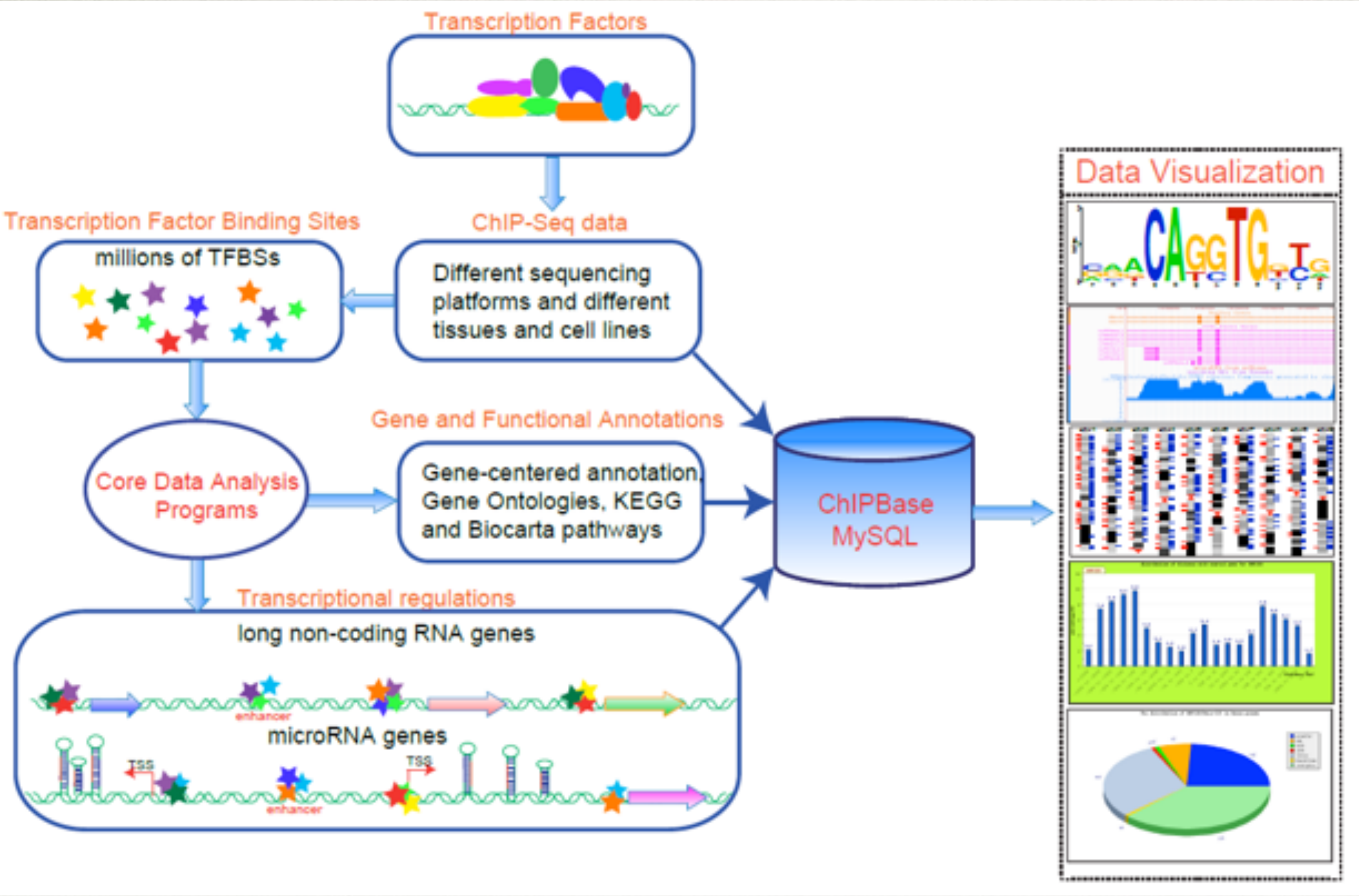
Overlap with pathological repeat regions such as centromeric, telomeric and satellite repeats that often have few unique mappable locations interspersed in repeats.

Where to Find ChIPSeq Data

Types of ChIPSeq Data

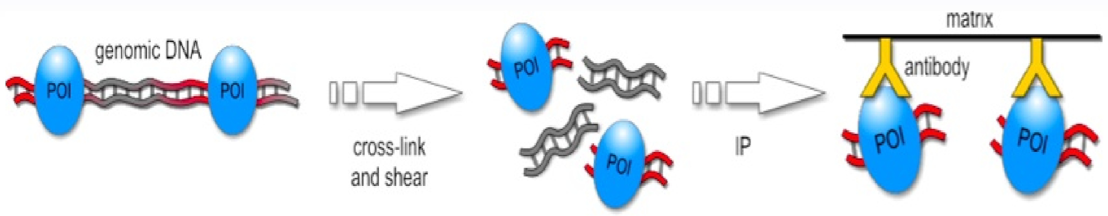
- NCBI (GEO) (SRA -tabular)
- UCSC (various - bam,bed,fastq,other)
- ENCODE (various - bigBed (.bb) and bigWIG (.bw))
- ChIPBase (CSV)
- **Cistrome Browser**

- http://deepbase.sysu.edu.cn/chipbase/ (CHIP-BASE)


Transcription Factors
ChIP-Seq data
Different sequencing platforms and different tissues and cell lines
Transcription Factor Binding Sites
millions of TFBSs
Core Data Analysis Programs
Gene and Functional Annotations
Gene-centered annotation, Gene Ontologies, KEGG and Biocarta pathways
Transcriptional regulations
long non-coding RNA genes
enhancer
microRNA genes
TSS
TSS
enhancer
ChIPBase MySQL
Data Visualization

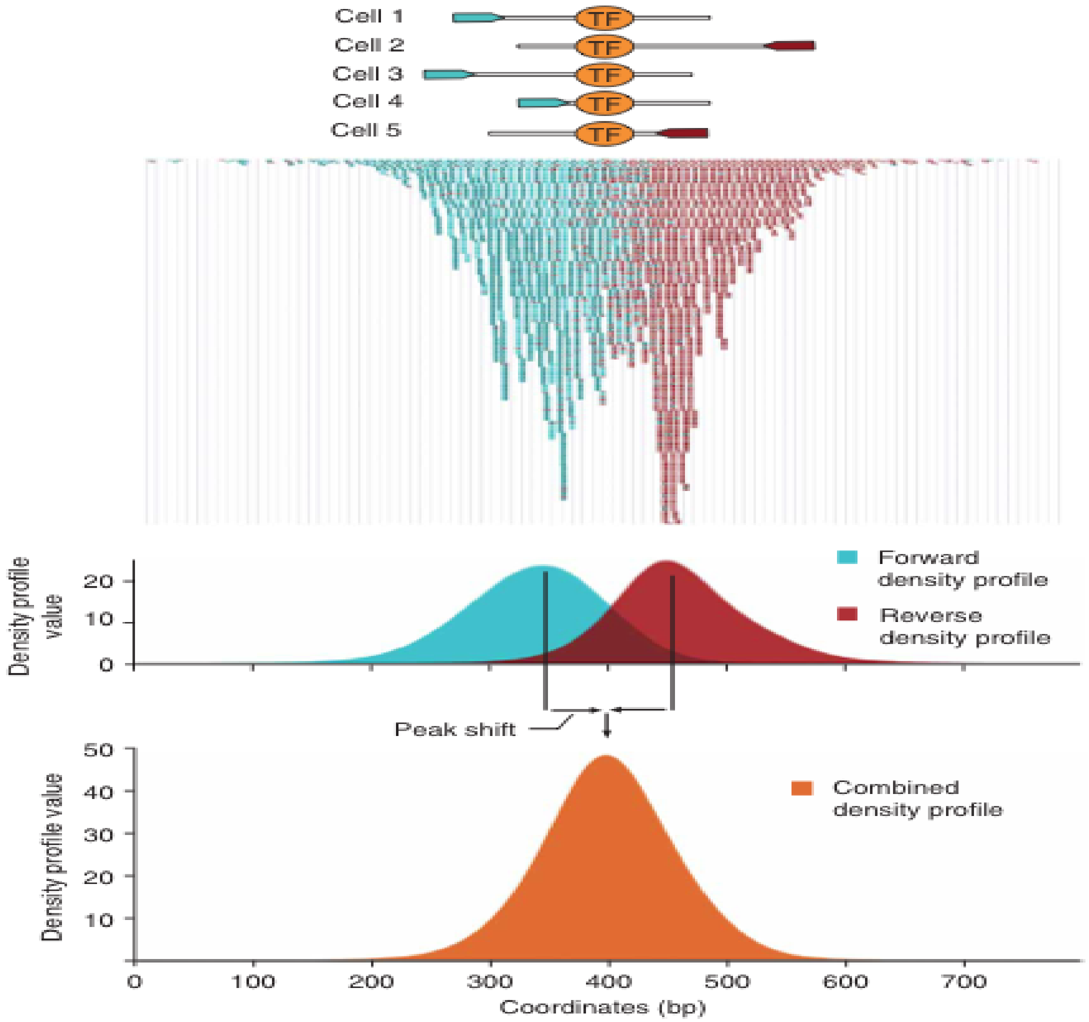
ChIPBase, an integrated resource and platform for decoding **transcription factor binding maps**, **expression profiles** and transcriptional regulation of **long non-coding RNAs (lncRNAs, lincRNAs)**, **microRNAs, other ncRNAs(snoRNAs, tRNAs, snRNAs, etc.) and protein-coding genes** from **ChIP-Seq** data. ChIPBase currently includes **millions of** transcription factor binding sites (TFBSs) among 6 species. ChIPBase provides several web-based tools and browsers to explore TF-lncRNA, TF-miRNA, TF-mRNA, TF-ncRNA and TF-miRNA-mRNA regulatory networks.(**Release 1.1: 1 November 2012**, **Tutorial**)

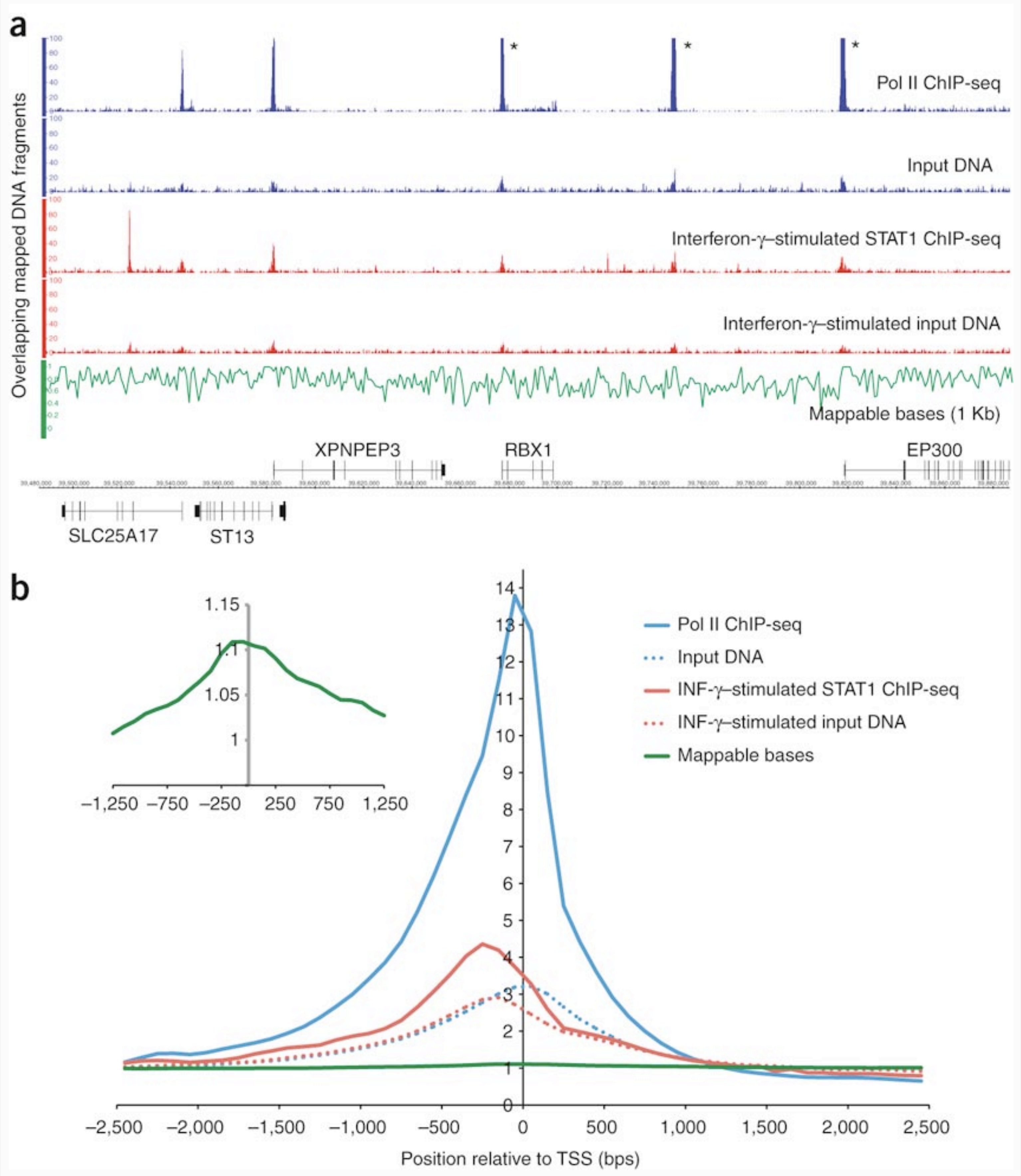
Visualization


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



Rozowsky, Nature Biotechnology, 2009

Types of Peaks

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

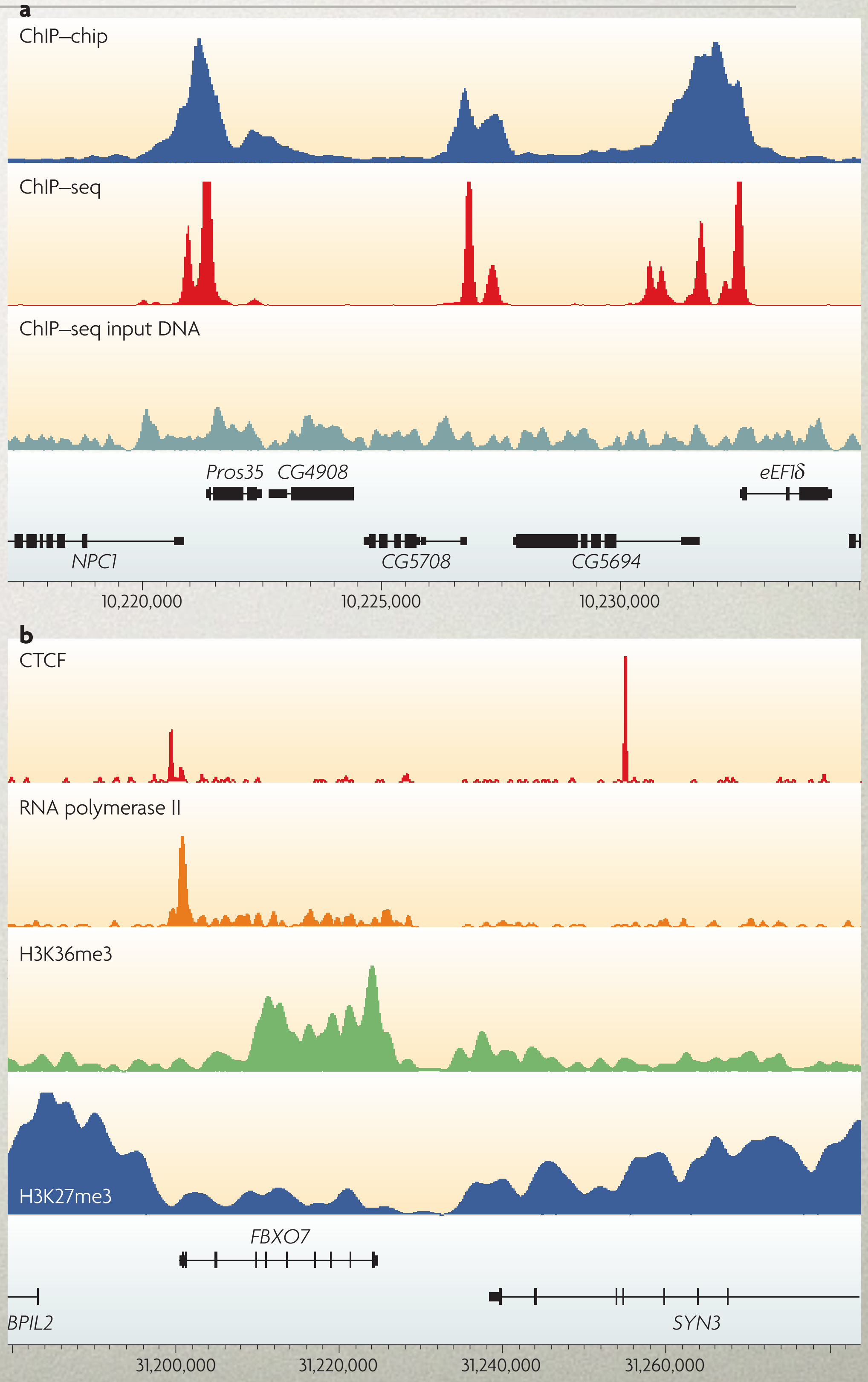
Figure 2 l chiP profiles. a l Examples of the profiles generated byNcahtruorme Raetivnieiwmsml uGneonpetriec-s cipitation followed by sequencing (ChIP–seq) or by microarray (ChIP–chip). Shown is a section of the binding profiles of the chromodomain protein Chromator, as measured by ChIP–chip (unlogged 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 l 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.

Sharp

Mixed

Medium

Broad



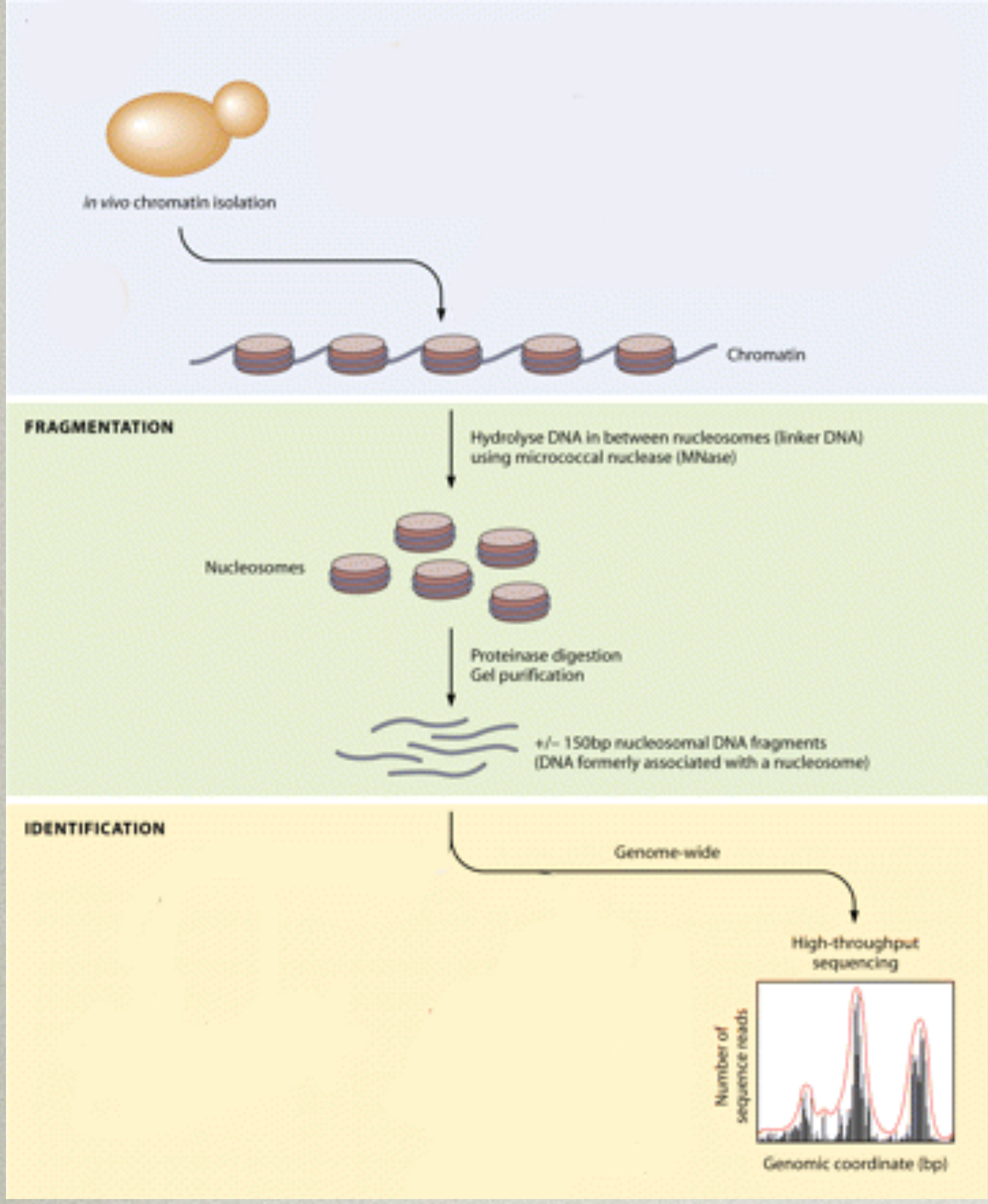
Visualization

Nothing can match the insight obtained by looking at your data

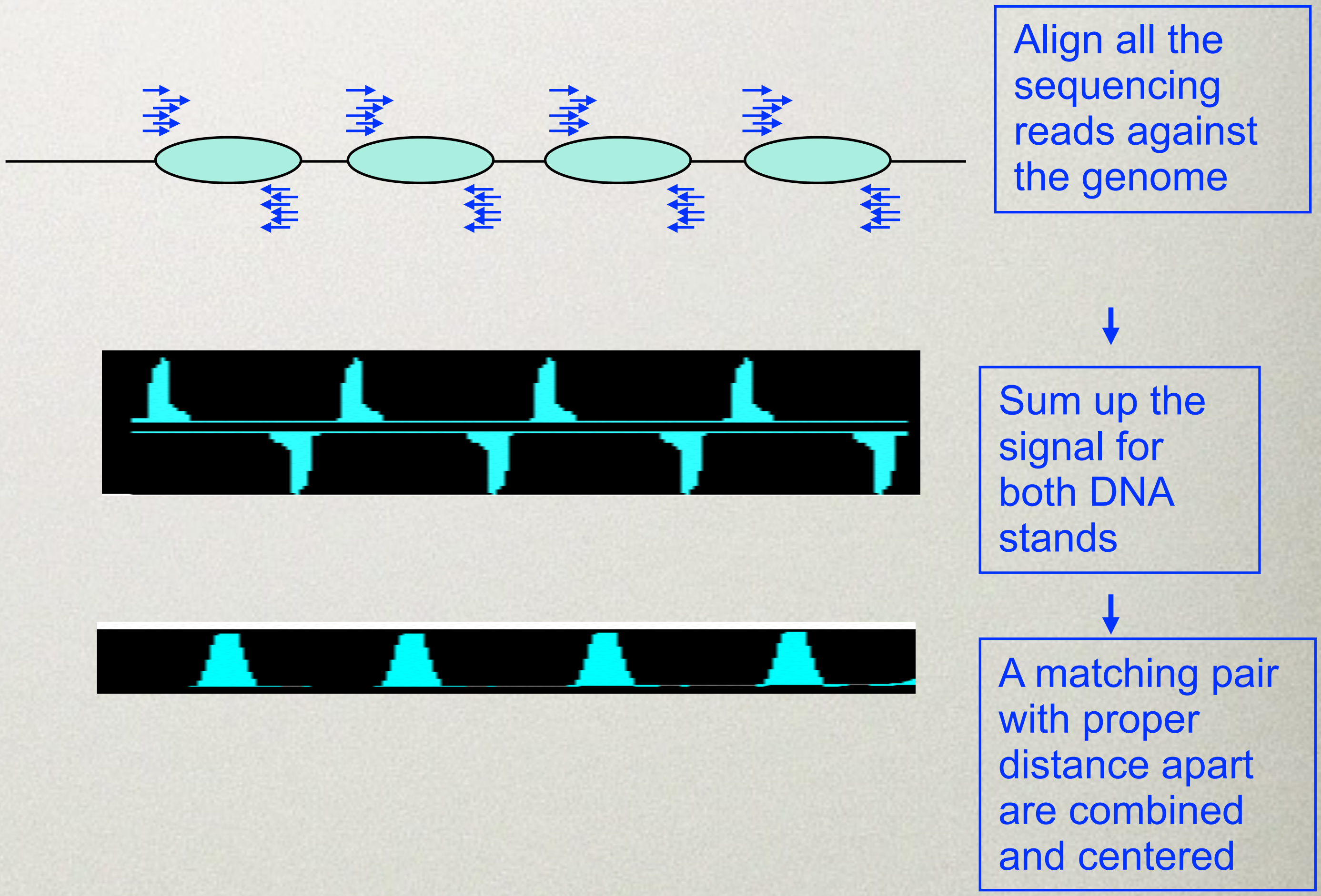
- IGV
- UCSC Genome Browser
- Heatmaps
- NGS-plot

Heat Maps

Yeast as Model Organism


in vivo chromatin isolation
Chromatin
FRAGMENTATION
Hydrolyse DNA in between nucleosomes (linker DNA) using micrococcal nuclease (MNase)
Nucleosomes
Proteinase digestion
Gel purification
+/– 150bp nucleosomal DNA fragments (DNA formerly associated with a nucleosome)
IDENTIFICATION
Genome-wide
High-throughput sequencing
Number of sequence reads
Genomic coordinate (bp)

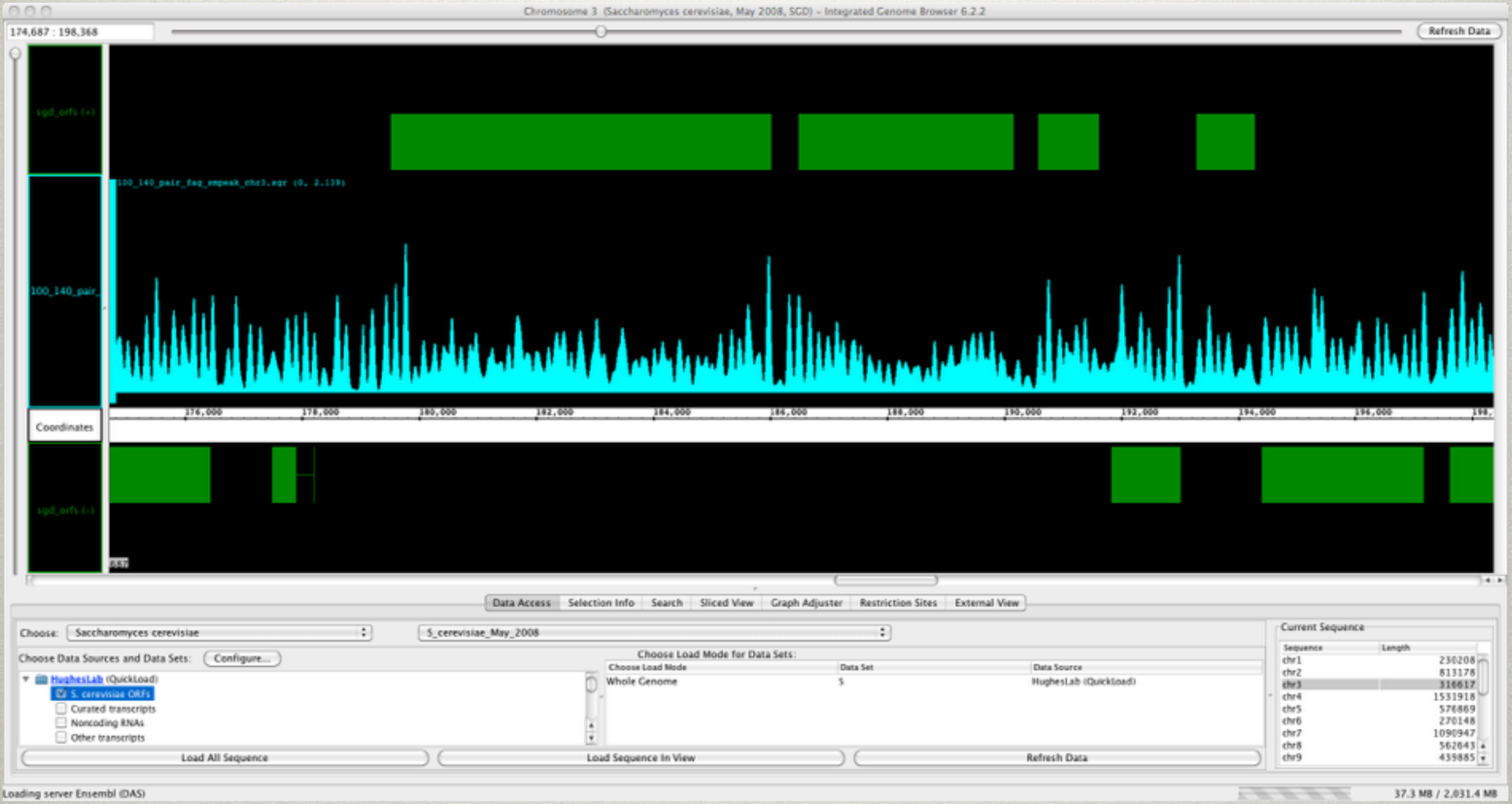
Steps of converting the sequencing reads to nucleosome positions



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



Choose: Saccharomyces cerevisiae

S_cerevisiae_May_2008

Choose Data Sources and Data Sets:

- HughesLab (QuickLoad)
  - ☑ S. cerevisiae ORFs
  - □ Curated transcripts
  - □ Noncoding RNAs
  - □ Other transcripts

Choose Load Mode for Data Sets:

Choose Load Mode	Data Set	Data Source
Whole Genome	S	HughesLab (QuickLoad)

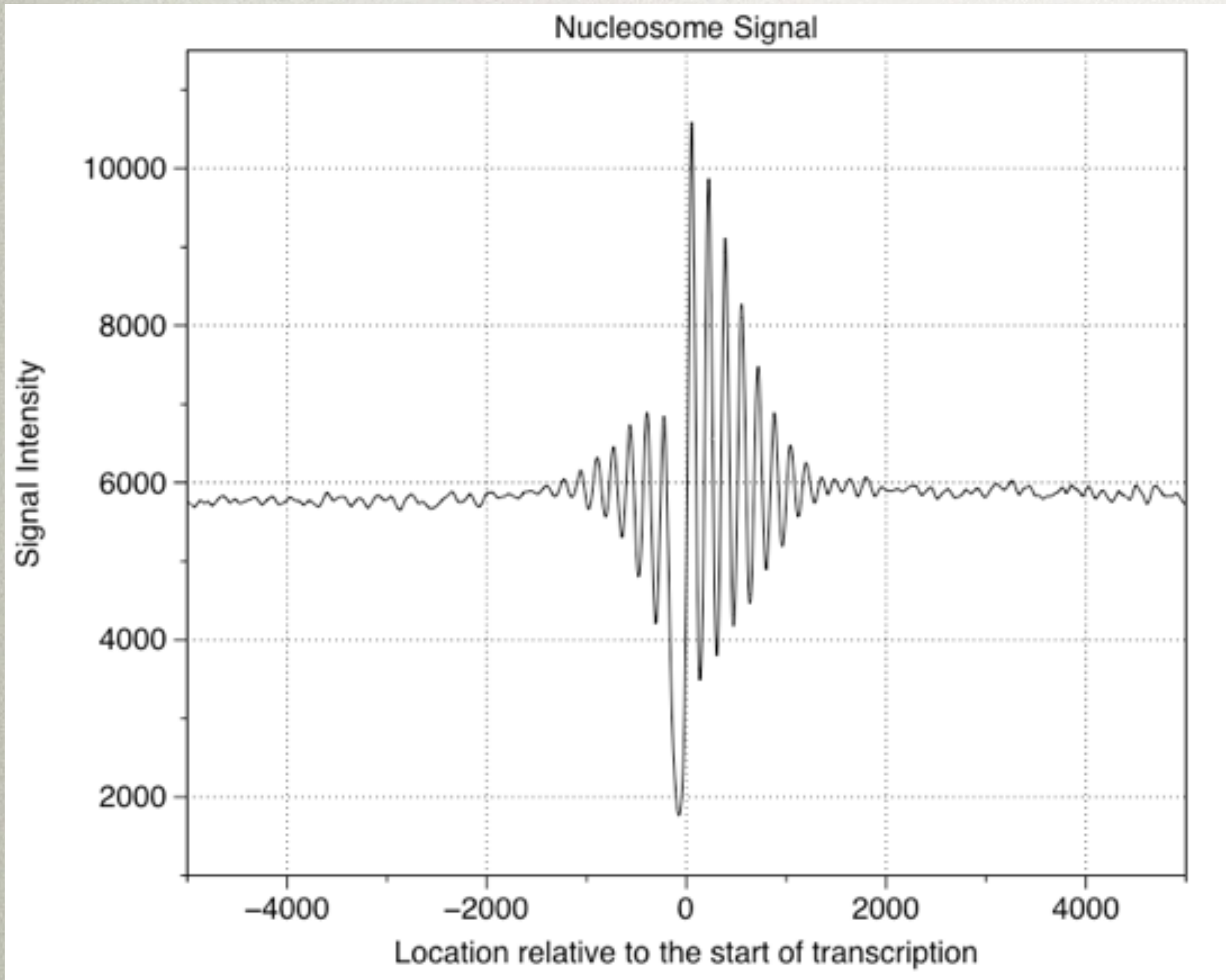
Current Sequence

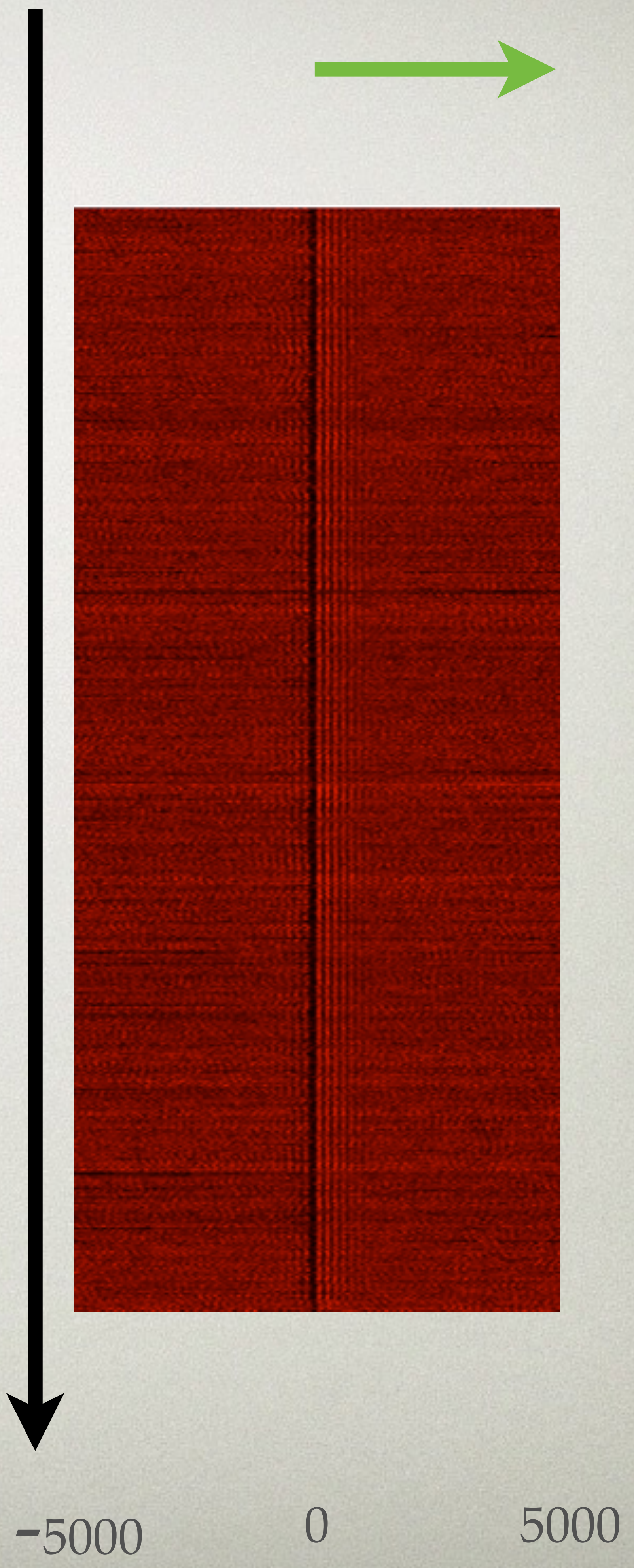
Sequence	Length
chr1	230208
chr2	813178
chr3	316617
chr4	1531918
chr5	576869
chr6	270148
chr7	1090947
chr8	562643
chr9	439885

Loading server Ensembl (DAS)

37.3 MB / 2,031.4 MB

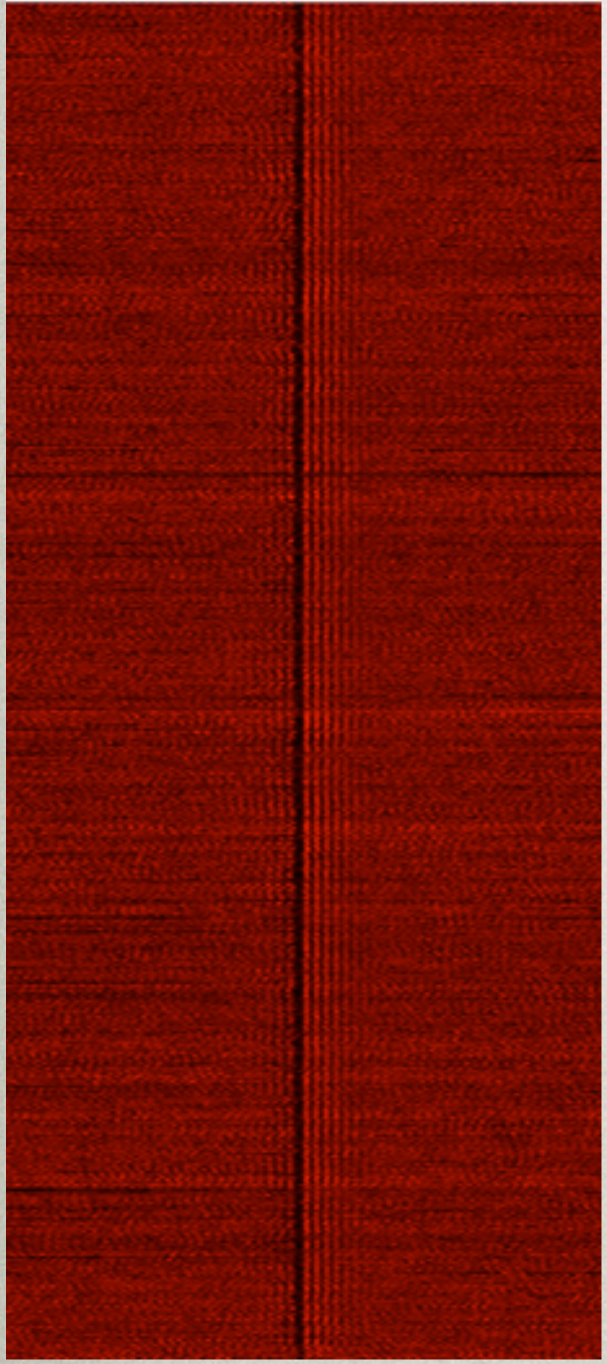
* start site of all genes (~4000)

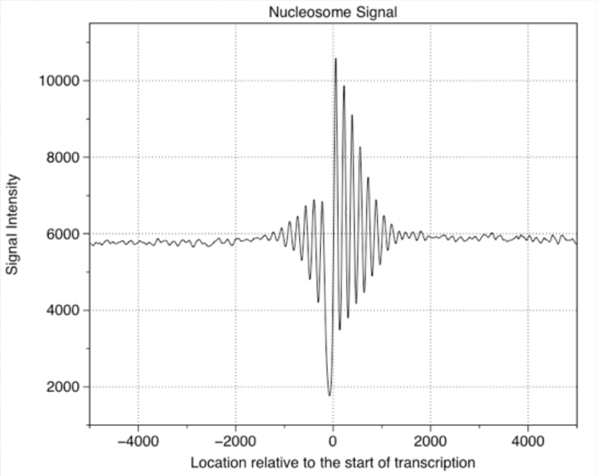


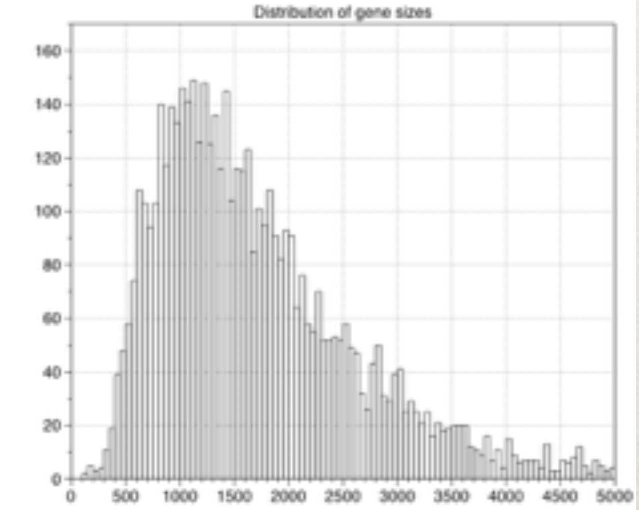


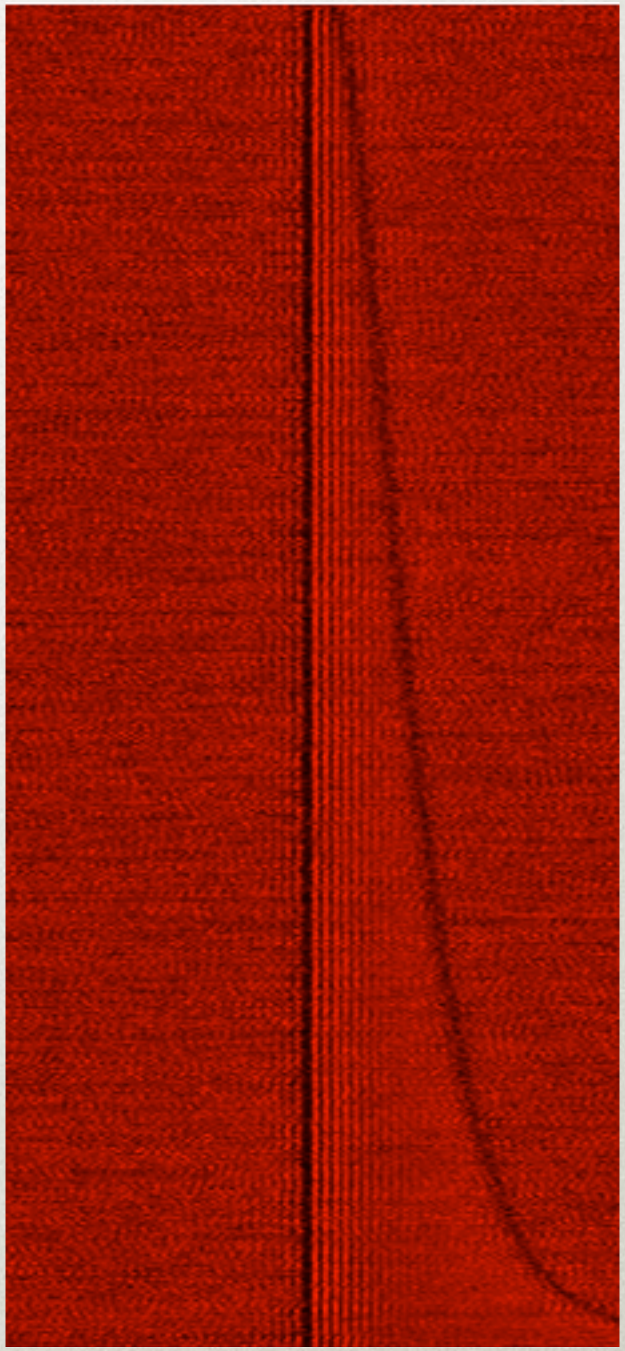
Original

Sorted by Gene Size

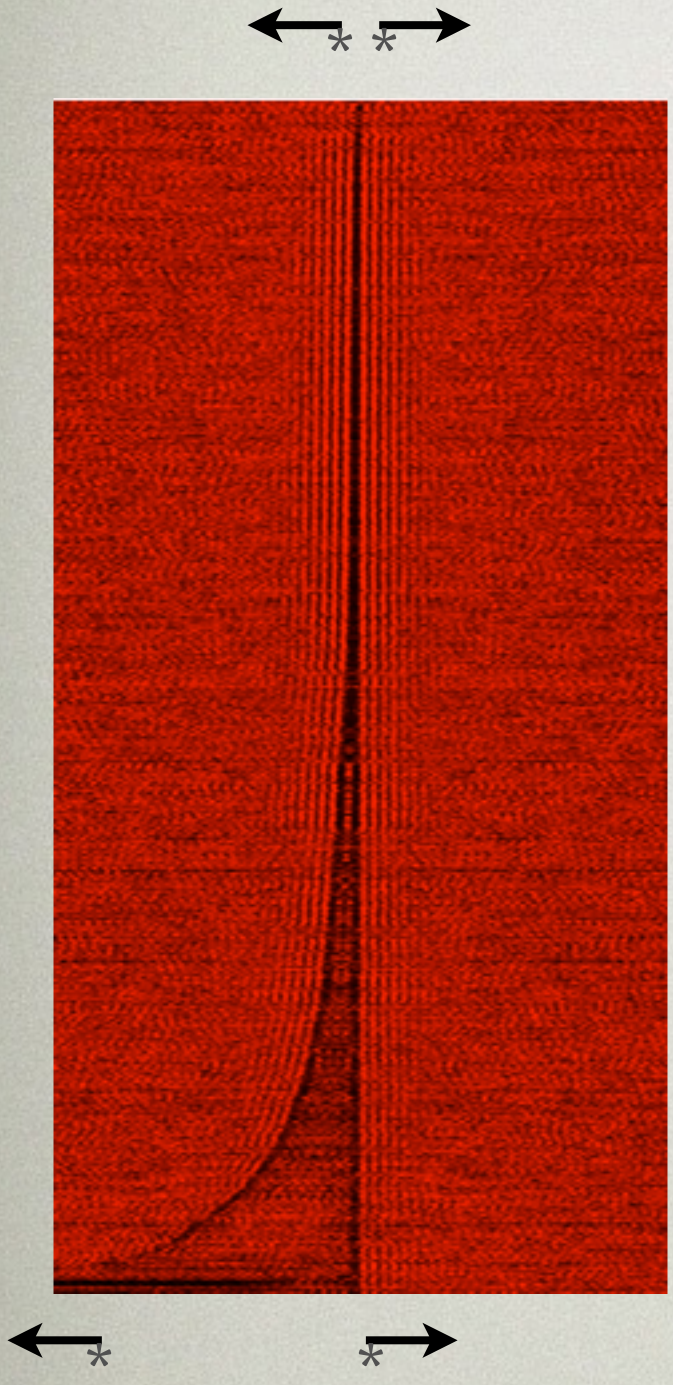


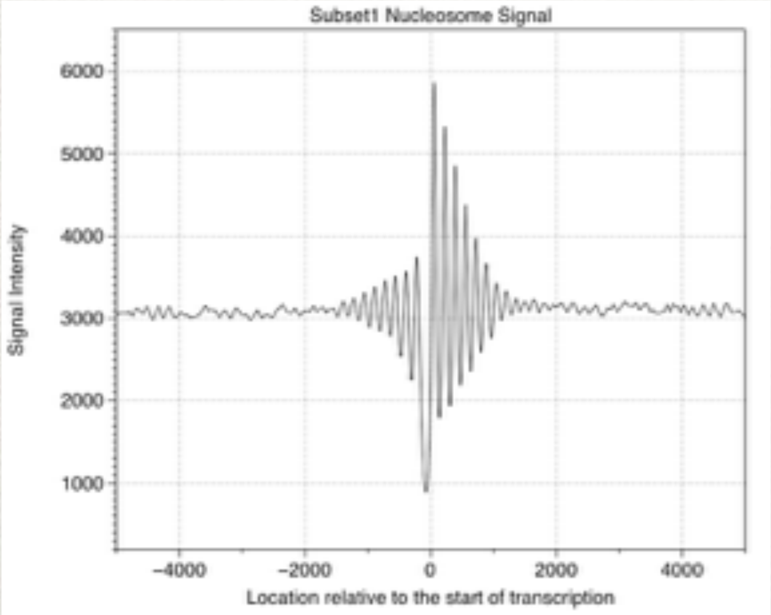


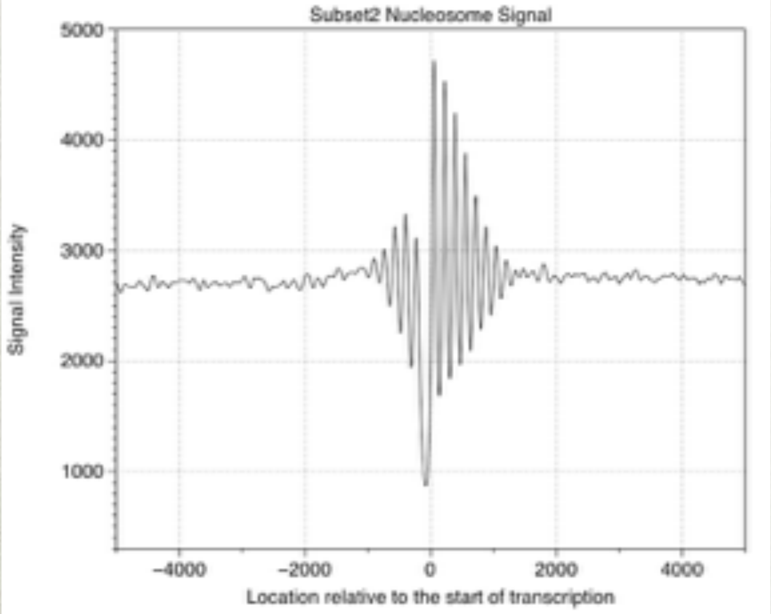


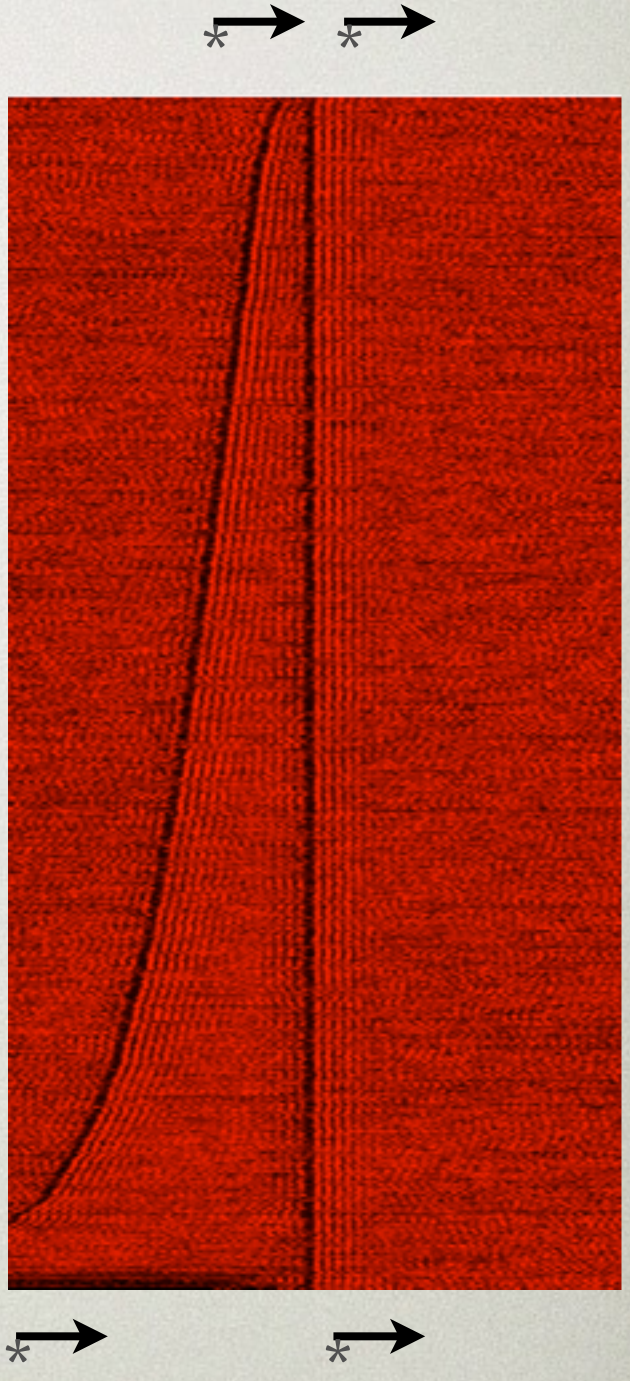


Sorted by nearest neighbour





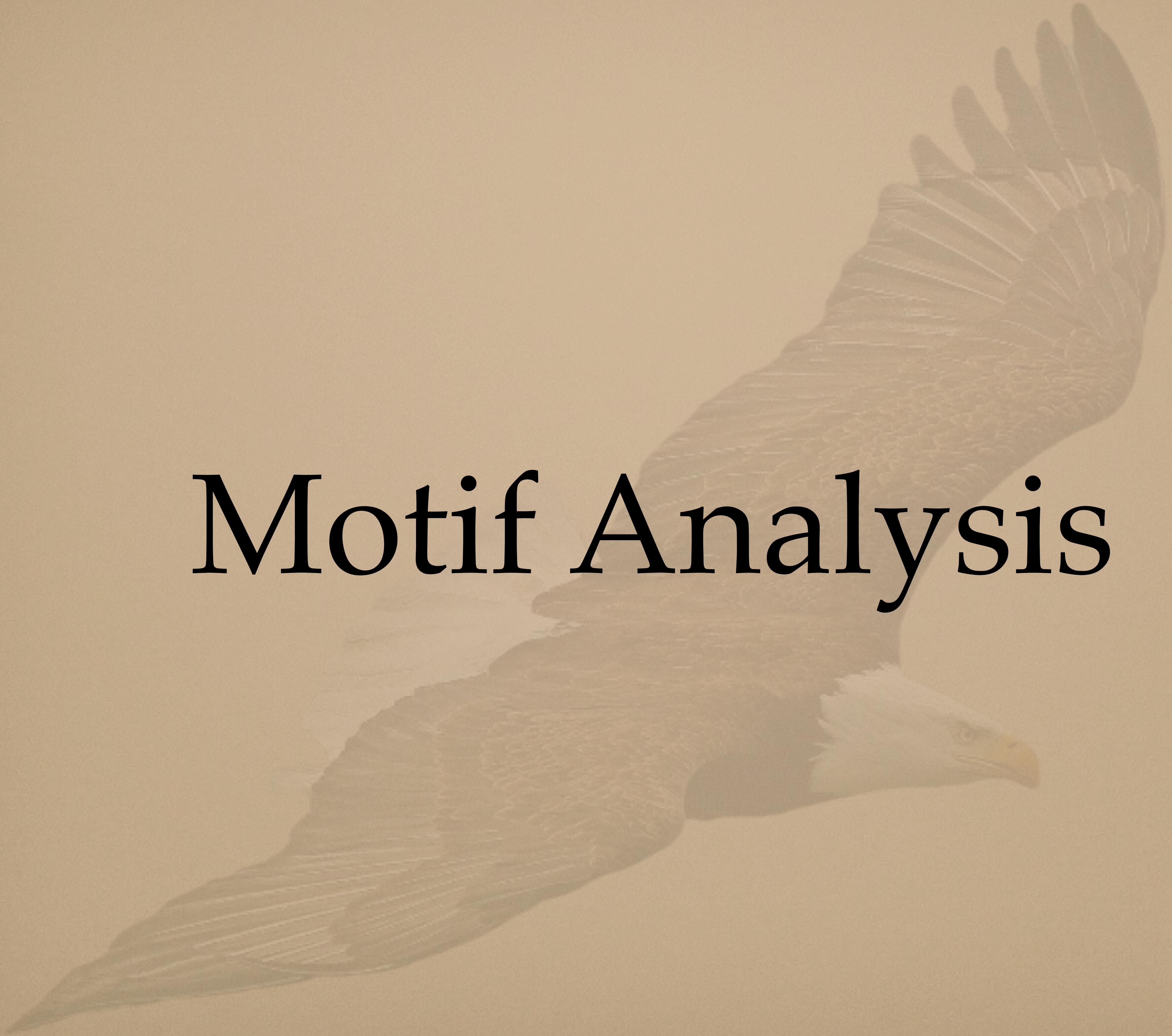




Take home message

- 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.



Motif Analysis

Motif Analysis

- Known Motifs
- Novel Motif finding programs

The MEME Suite

Motif-based sequence analysis tools

http://meme.nbcr.net/meme/

MEME-ChIP uses a combination of motif discovery using MEME (good for wide motifs) and DREME (good for shorter motifs) and comparison of both found motifs and the sequence data against databases of known motifs.

Results-link

Course Outline

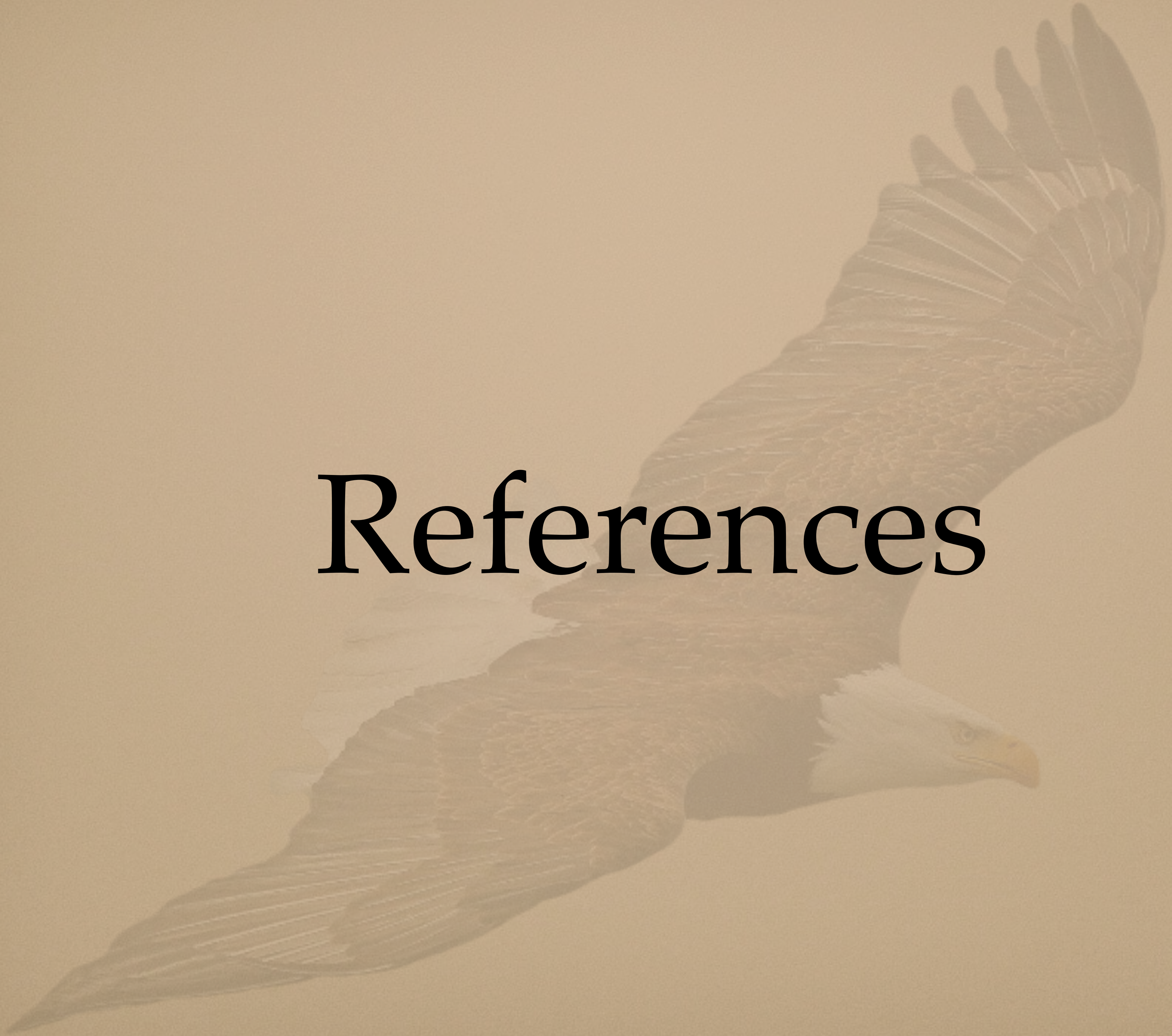
Day 1

- Design and Analysis Overview (9:30 - 12:30)
- Genomatix (The basics & Data Import and Mapping) - (1:30 - 4:30)

Day 2

- Genomatix (Workflows & Biological Perspective) - (9:30 - 12:30)
- CISTROME (1:30 - 4:30)

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

- 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)

- 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)