ChIP-Seq Analysis

Presenter:
Paula Moolhuijzen| Centre for Crop Disease Management (CCDM), Curtin University

Xi Li | CSIRO, AU
Remco Loos | EMBL-EBI, UK
Myrto Kostadima | EMBL-EBI, UK

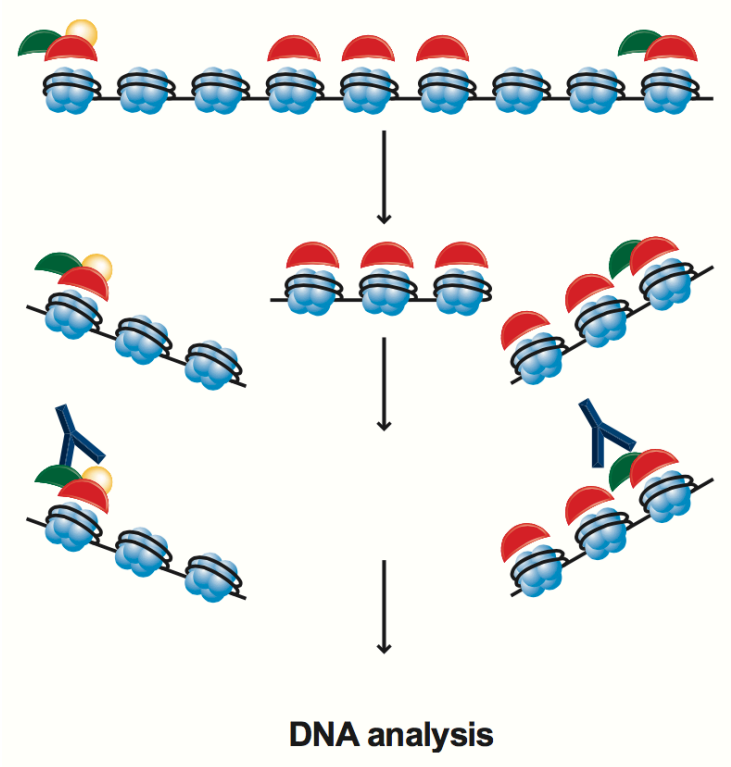
DATE: 12-07-2016

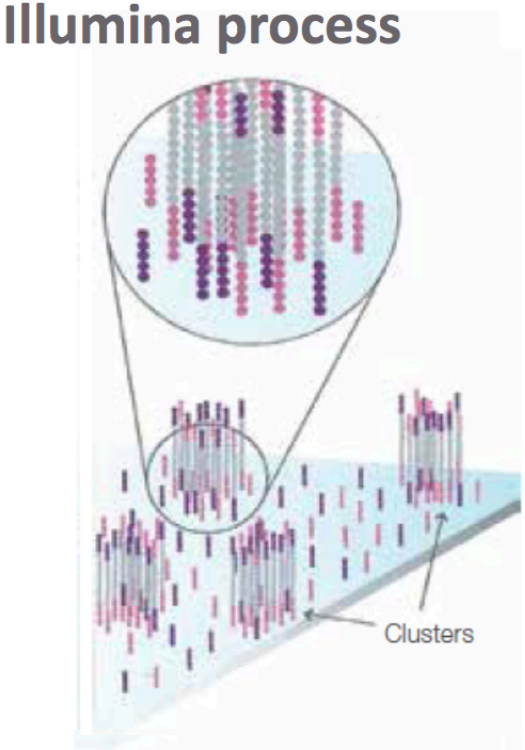
ChIP-Seq Overview

- **Introduction to ChIP-Seq**
  - **Background**
- **Experimental Design**
- **Overview of Analysis**
  - **How to do**
- **Introduction to hands-on workshop**
  - **Let's do**

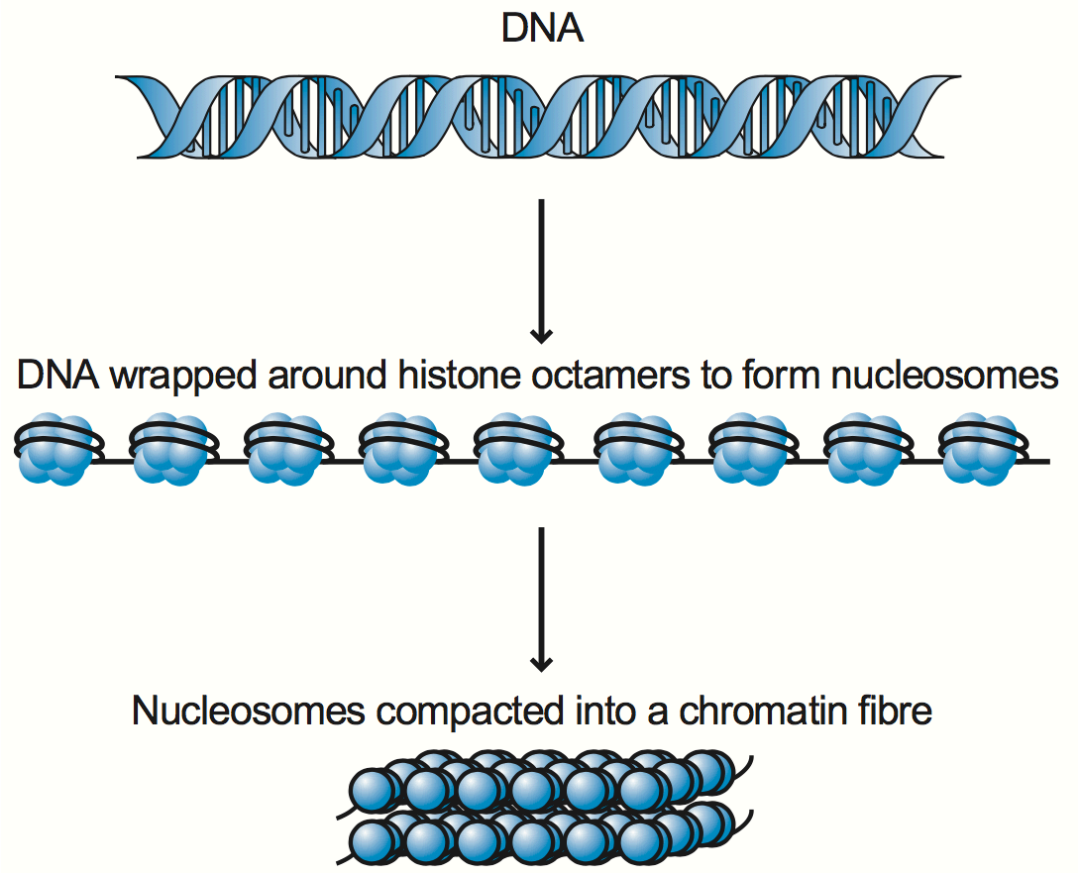
ChIP-Seq - Introduction

Chromatin ImmunoPrecipitation + Sequencing





Structure and Function of chromatin



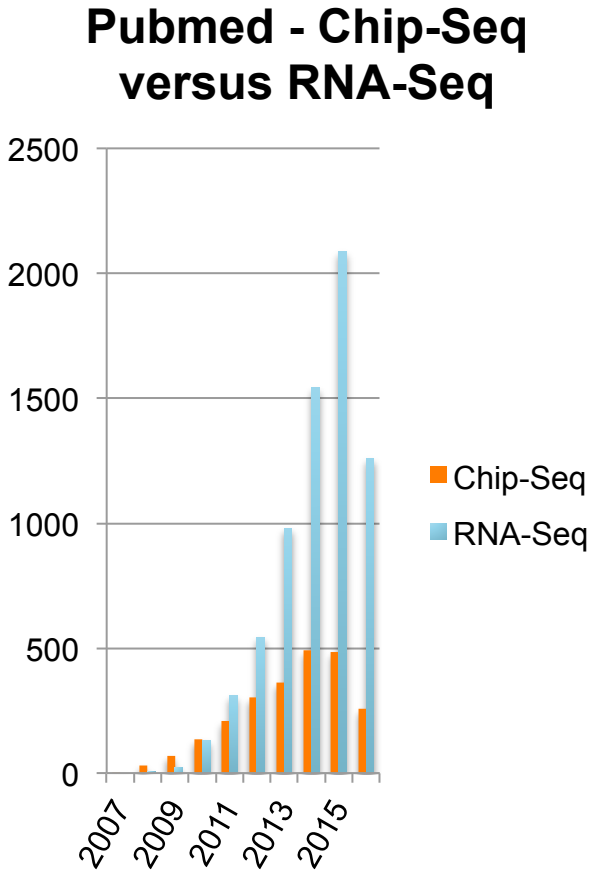
- Chromatin packages DNA to enable it to fit in the cell
- Chromatin serves as a mechanism to control gene expression

ChIP-Seq – Why Study?

- Study of gene regulation:
  - Protein-DNA interaction: Transcription factor binding locations, core transcriptional machinery
  - Histone modifications, Nucleosome positioning, DNA methylation

ChIP-Seq – First Application

- One of the early applications of NGS
- First studies published in 2007
  - Johnson et al (Science) – NRSF, Genome-wide mapping of in vivo protein-DNA interactions
  - Barski et al (Cell) - High-resolution profiling of histone methylations in the human genome
  - Robertson et al (Nature Methods) . Genome-wide profiles of STAT1 DNA association
  - Mikkelsen et al (Nature) - Natural variation of histone modification and its impact on gene expression in the rat genome

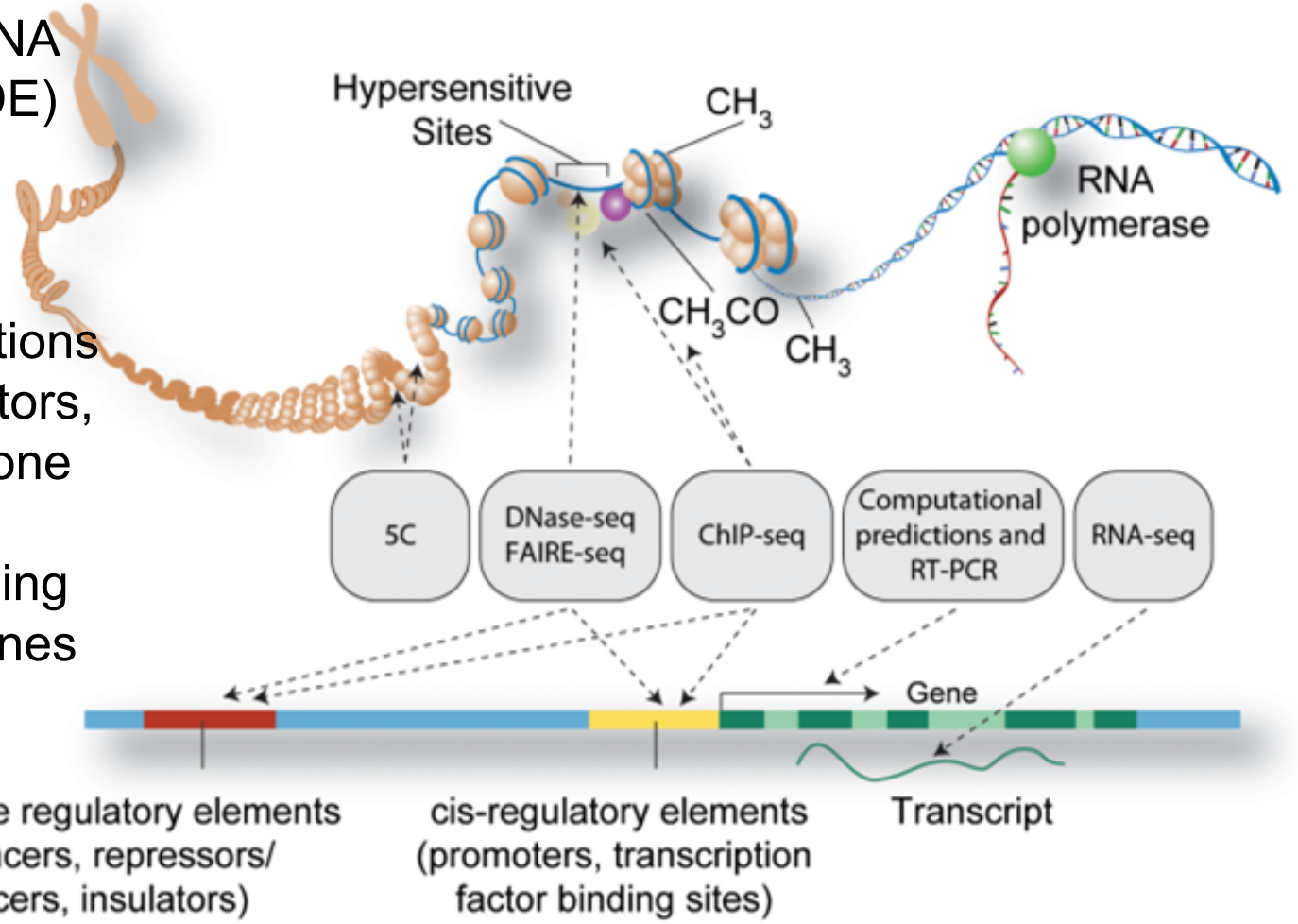


Total Publications (PubMed)
ChIP-Seq + 2,349
RNA-Seq +6,883

ChIP-Seq – ENCODE Key Task

Encyclopedia of DNA Elements (ENCODE) Project: Key Tasks

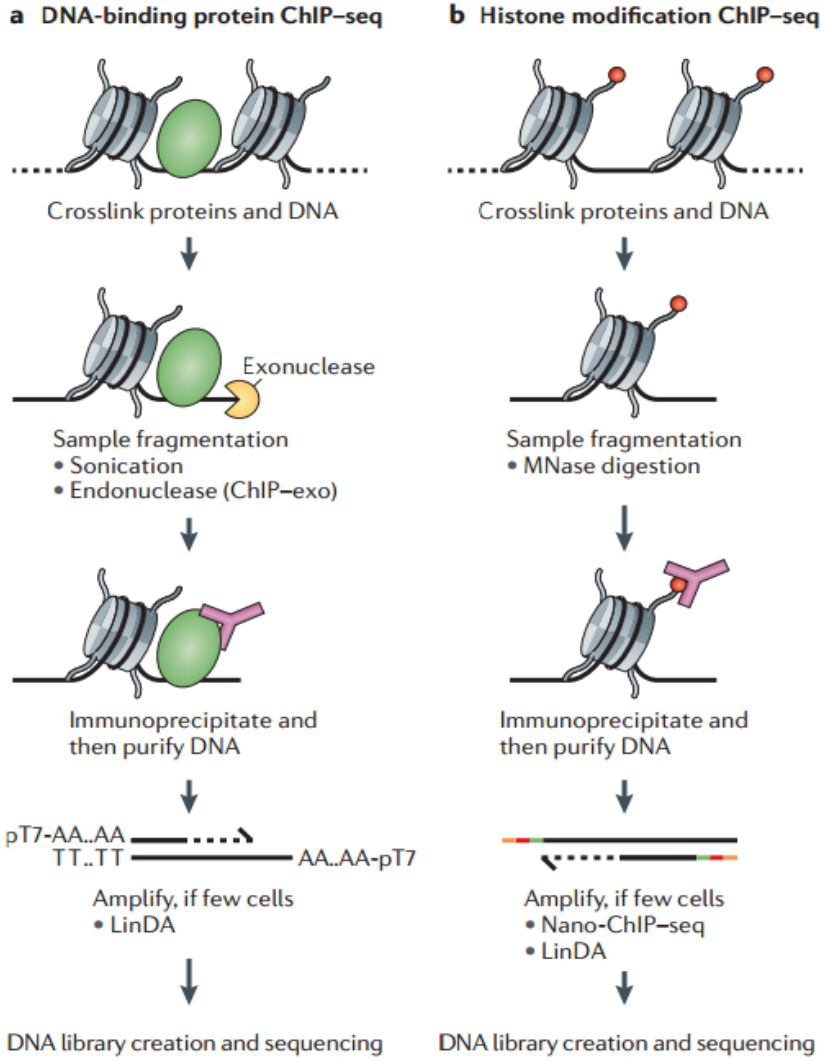
Mapping of the chromosomal locations of transcription factors, nucleosomes, histone modifications, chromatin remodeling enzymes, chaperones and polymerases



A User's Guide to the Encyclopedia of DNA Elements (ENCODE), 2011 PLOS | BIOLOGY

ChIP-Seq - lab procedures

1. **Cross-linking**: proteins bound to chromatin
2. **Shearing:** fragments the chromatin
3. **Immunoprecipitation:** captures the DNA fragments bound to one protein using an antibody specific to it.
4. **Sequencing:** and sequences the ends of the captured fragments using next-generation sequencing (NGS).



Furey, TS, 2012 Nature Reviews | Genetics

ChIP-chip vs ChIP-seq

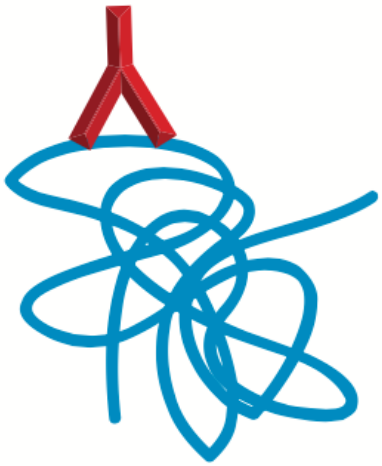
	ChIP-chip	**ChIP-seq**
Resolution	Array-specific	**High - single nucleotide**
Coverage	Limited by sequences on the array	Limited by “alignability” of reads to the genome, increases with read length
Repeat elements	Masked out	Many can be covered (47% of human genome is non-repetitive but ~80% is uniquely mapped)
Cost	$400-800 per array (1-6M probes), multiple arrays needed for human genome	Around $1000 per lane; 20-30M reads
Source of noise	Cross hybridization	Sequencing bias, GC bias, sequencing error
Amount of ChIP DNA required	High, few micrograms	Low 10-50ng
Dynamic range	Lower detection limit and saturation at high signal	Not limited
Multiplexing	Not possible	Possible

Chip-Seq Overview

- Introduction to Chip-Seq
- **Experimental Design**
- Overview of Analysis
- Introduction to hands-on workshop
  - Let's Do

Experimental Design considerations

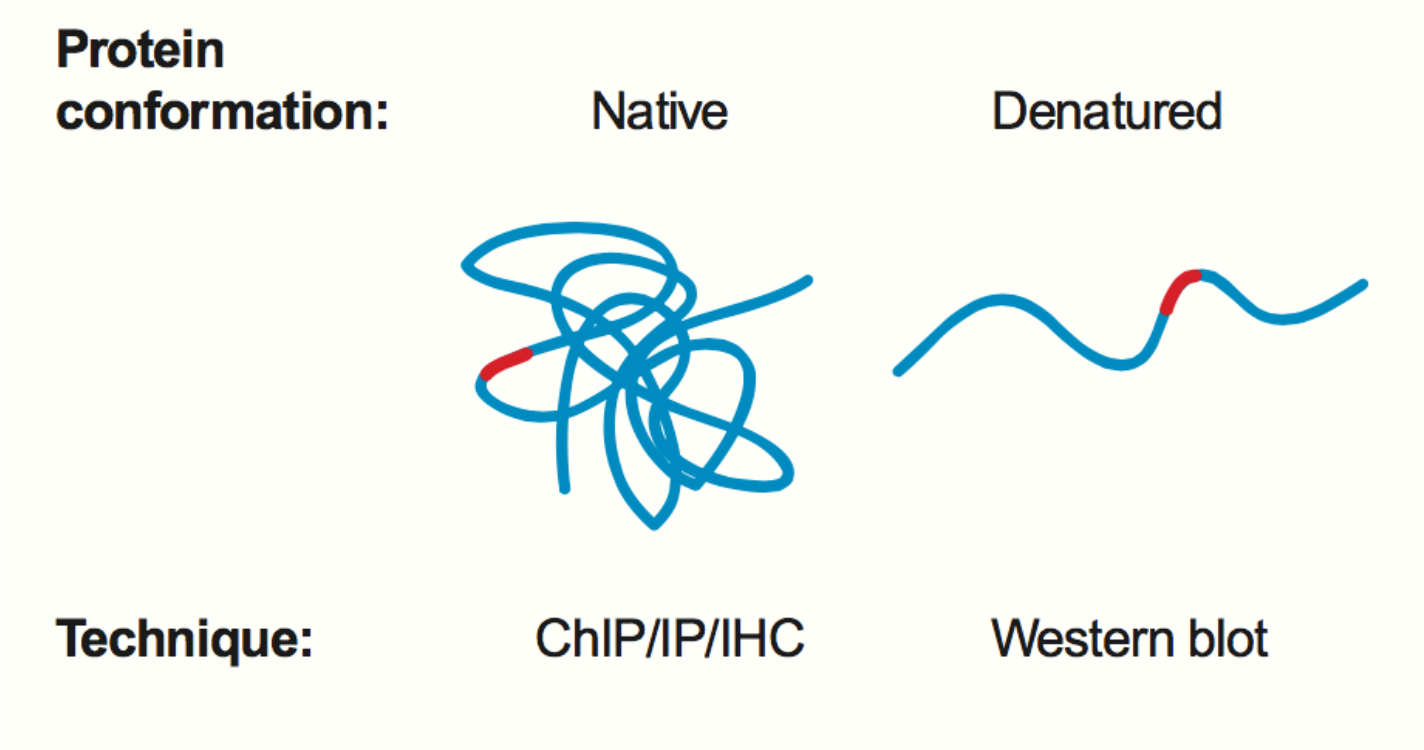
- **Antibody quality**
- Control experiment
- Depth of sequencing
- Multiplexing
- Paired-end reads



Antibody quality

- Antibody quality - a sensitive and specific antibody will give a high level of enrichment
  - Limited efficiency of antibody is the main reason for failed ChIP-seq experiments
  - Check your antibody ahead if possible.
    - Immunoprecipitation / immunohistochemistry / immunocytochemistry are good indicators of success in ChIP
    - Western blotting to check the reactivity of the antibody with unmodified and non-histone proteins.

Antibody quality



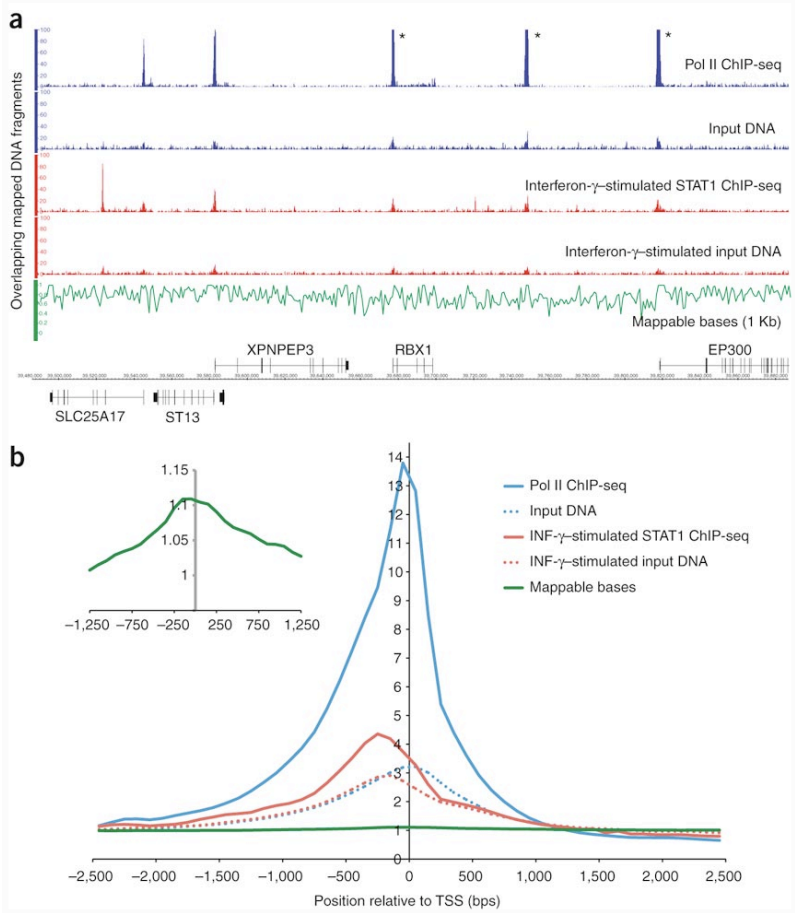
- If an antibody works in IP, IHC or ICC, there is a good chance that the epitope will also be recognized in ChIP

Experimental Design

- Antibody quality
- **Control experiment**
- Depth of sequencing
- Multiplexing
- Paired-end reads

Why we need a control sample

- Open chromatin regions are fragmented more easily than closed regions.
- Repetitive sequences might seem to be enriched (inaccurate repeats copy number in the assembled genome).
- Uneven distribution of sequence tags across the genome
- A ChIP-seq peak should be compared with the same region in a matched control

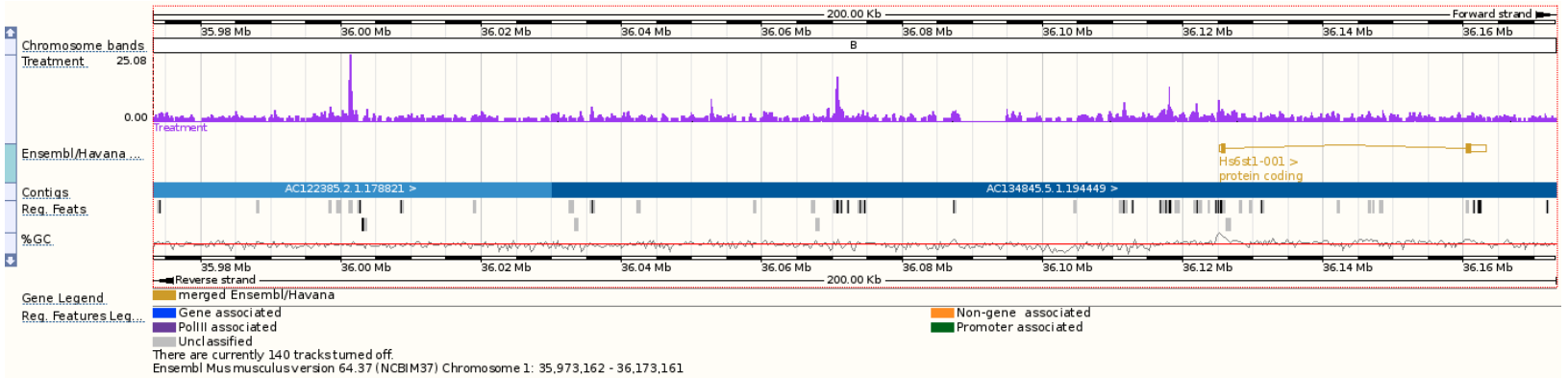


Rozowsky, Nature Biotechnology, 2009

Control type

- Input DNA
- Mock IP - DNA obtained from IP without antibody
  - Very little material can be pulled down leading to inconsistent results of multiple mock IPs.
- Nonspecific IP - using an antibody against a protein that is not known to be involved in DNA binding
- There is no consensus on which is the most appropriate
- Sequencing a control can be avoided when looking at:
  - time points
  - differential binding pattern between conditions

Experimental Design



- **Depth of sequencing**
- Multiplexing
- Paired-end reads

Depth of sequencing

- More prominent peaks are identified with fewer reads, whereas weaker peaks require greater depth
- Number of putative target regions continues to increase significantly as a function of sequencing depth
- GA1 generated 4-6M reads, GA2 12-15M reads, GA2X 18-30M, **HiSeq2500 up to 250 M reads per lane**
- With current sequencing technologies, one lane is usually sufficient

Consider size of the genome and number and size of DNA binding sites

Sequence Saturation: MACS "diag" table

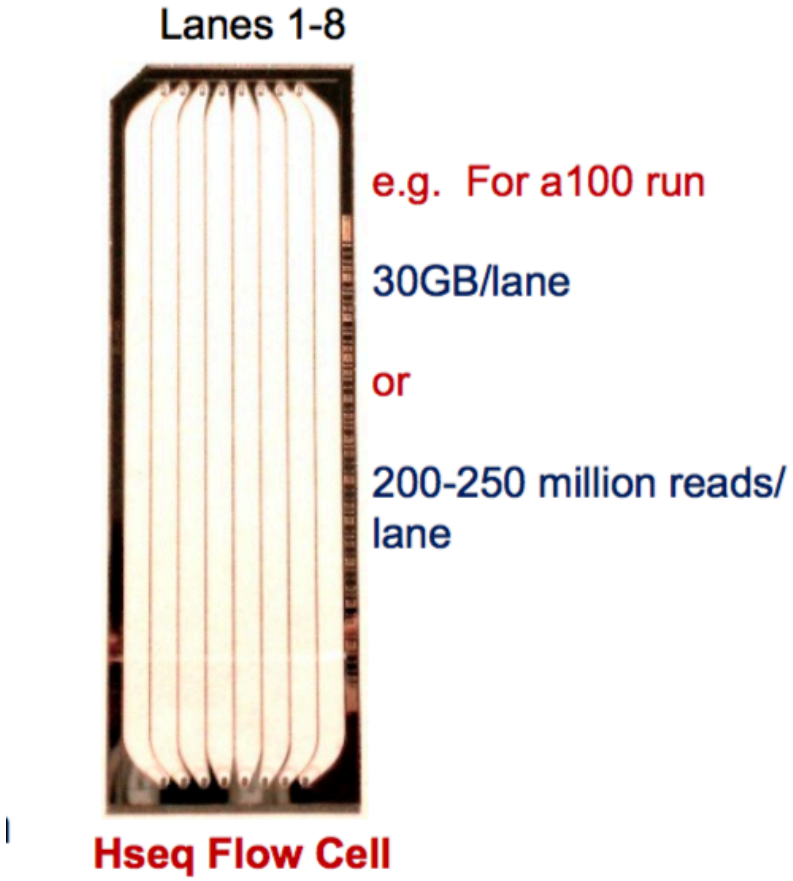
		% of peaks covered after sampling 90% ... 20% of the total tags							
FC	# peaks	90%	80%	70%	60%	50%	40%	30%	20%
0-20	31530	75.01	55.98	39.58	26.01	15.35	7.43	2.64	0.51
20-40	5481	99.62	97.7	92.52	80.46	61.34	36.75	14.61	2.81
40-60	235	100	100	100	100	99.57	90.21	68.51	28.09
60-80	40	100	100	100	100	100	100	95	62.5
80-100	7	100	100	100	100	100	100	100	85.71
100-120	2	100	100	100	100	100	100	100	100
120-140	5	100	100	100	100	100	100	100	100
160-180	1	100	100	100	100	100	100	100	100

Experimental Design

- Antibody quality
- Control experiment
- Depth of sequencing
- **Multiplexing**
- Paired-end reads

Multiplexing

- Number of reads per run continue to increase
- The ability to sequence multiple samples at the same time becomes important, especially for small genomes
- Different barcode adaptors are ligated to different samples
- Useful in experimental design to control technical variation

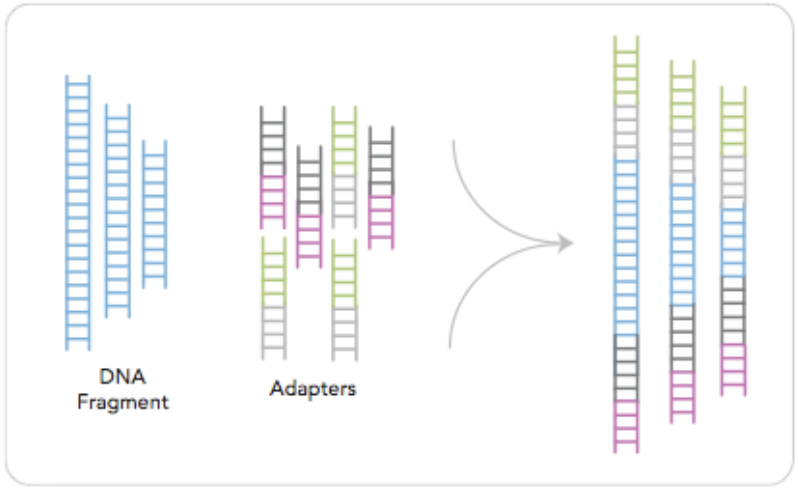


Experimental Design

- Antibody quality
- Control experiment
- Depth of sequencing
- Multiplexing
- **Paired-end reads**

Paired-end sequencing

- Reads are sequenced from both ends
- Increase “mappability” - especially in repetitive regions
- Costs ~twice as much as single end reads

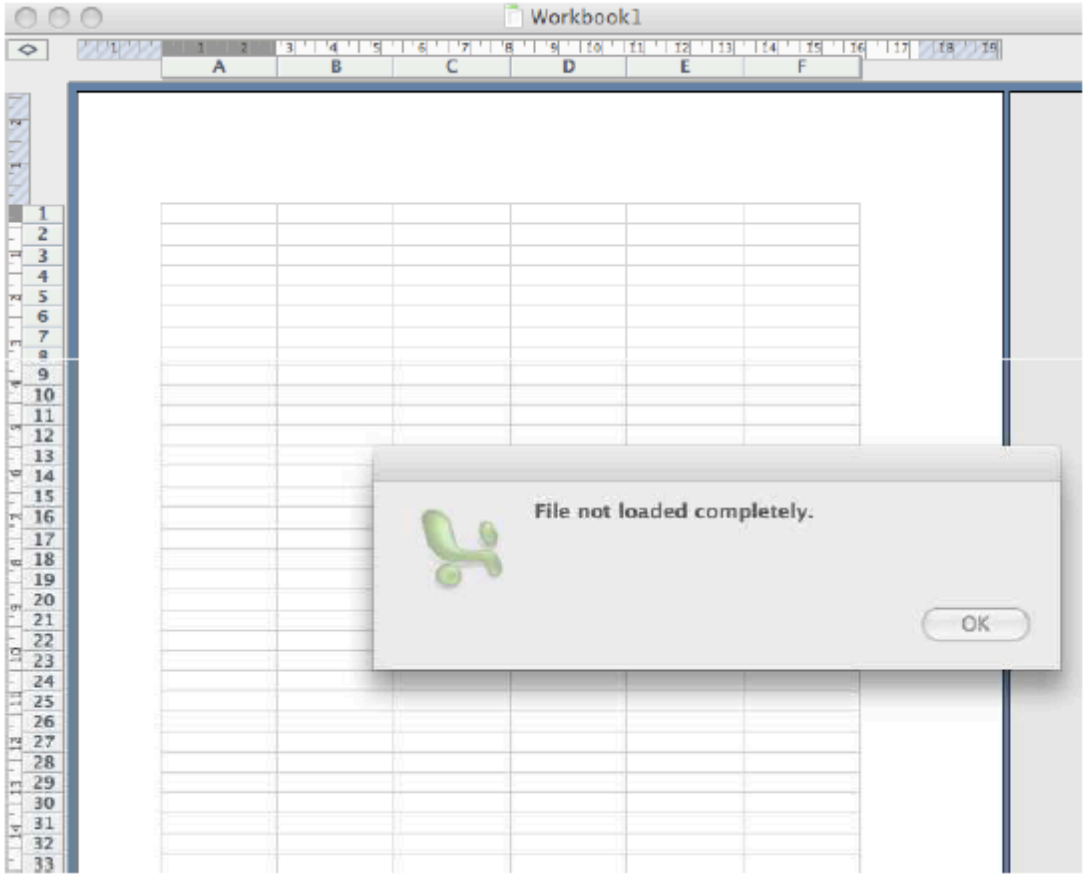


- For ChIP-Seq, usually not worth the extra cost, unless you have a specific interest in repeat regions
- Can assist in identifying duplicated regions

Chip-Seq Overview

- Introduction to Chip-Seq
- Experimental Design
- Overview of Analysis
- Introduction to hands-on workshop
  - Let's Do

A Challenge - Bioinformatics



Mapping *in-vivo* interactions of Protein-DNA poses multiple computational challenges

Chip-Seq – Key analysis steps

1. Sequence alignment
   - Align sample(s) and control to a reference genome
2. Peak Calling
   - Read alignment depth of coverage
3. Enrichment Analysis
   - Peak annotation
4. Motif Analysis
   - Identify specific sequence motifs for binding sites
5. Differential binding analysis

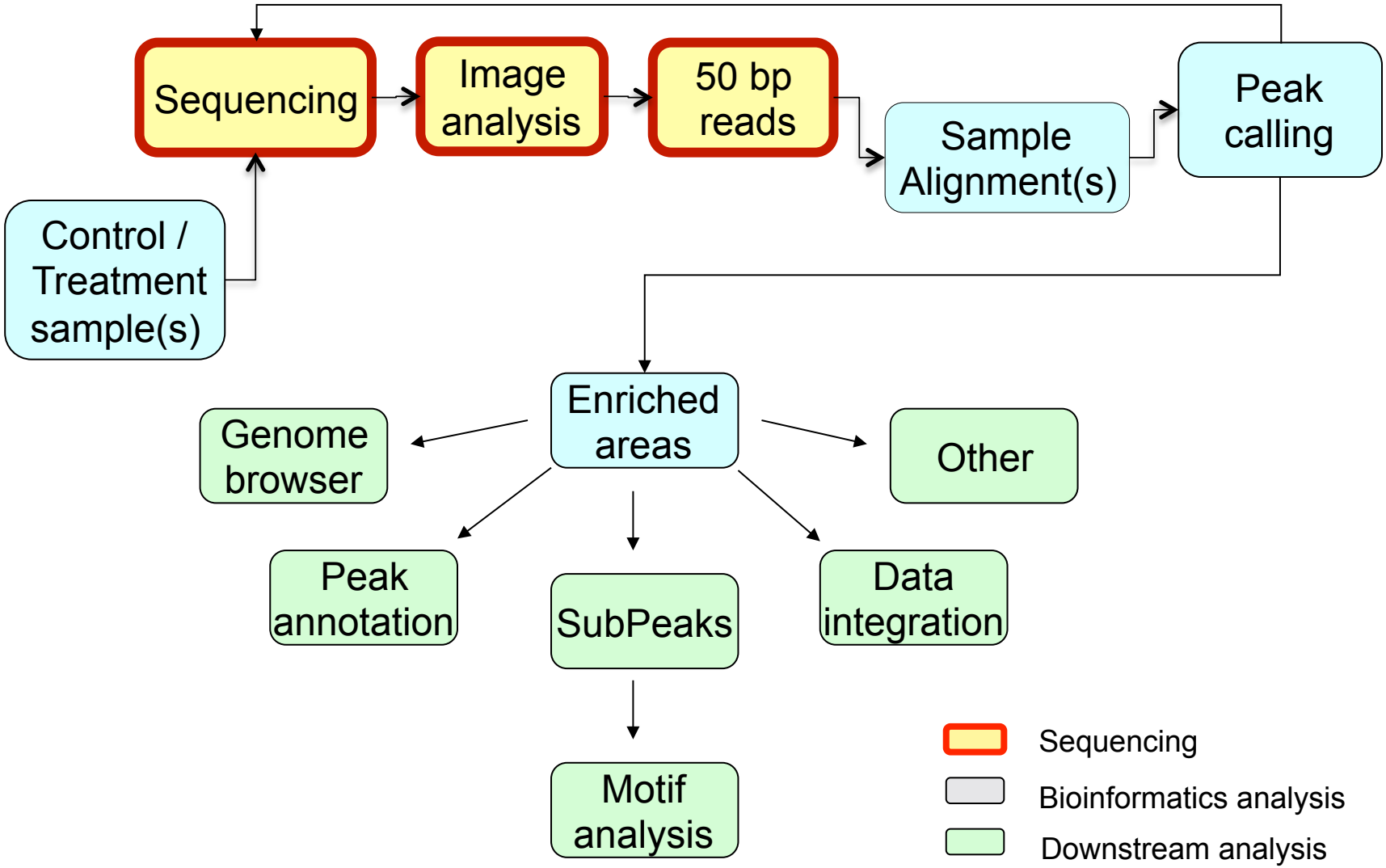
Analysis – Overview tools

Short-read aligners		
BWA	http://bio-bwa.sourceforge.net	Fast and efficient; based on the Burrows–Wheeler transform
Bowtie	http://bowtie-bio.sourceforge.net	Similar to BWA, part of suite of tools that includes TopHat and CuffLinks for RNA-seq processing
GSNAP	http://research-pub.gene.com/gmap	Considers a set of variant allele inputs to better align to heterozygous sites
Wikipedia list of aligners	http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment	A comprehensive list of available short-read aligners, with descriptions and links to download the software
Peak callers		
MACS	http://liulab.dfci.harvard.edu/MACS	Fits data to a dynamic Poisson distribution; works with and without control data
PeakSeq	http://info.gersteinlab.org/PeakSeq	Takes into account differences in mappability of genomic regions; enrichment based on FDR calculation
ZINBA	http://code.google.com/p/zinba	Can incorporate multiple genomic factors, such as mappability and GC content; can work with point-source and broad-source peak data
Differential peak calling		
edgeR	http://www.bioconductor.org/packages/2.9/bioc/html/edgeR.html	Uses negative binomial distribution to model differences in tag counts; uses replicates to better estimate significant differences
DESeq	http://www-huber.embl.de/users/anders/DESeq	Also uses negative binomial distribution modelling, but differs in the calculation of the mean and variance of the distribution
baySeq	http://www.bioconductor.org/packages/release/bioc/html/baySeq.html	Uses empirical Bayes approach to identify significant differences; assumes negative binomial distribution of data
SAMSeq	http://www.stanford.edu/~junli07/research.html#SAM	Based on the popular SAM software; a non-parametric method that uses resampling to normalize for differences in sequencing depth

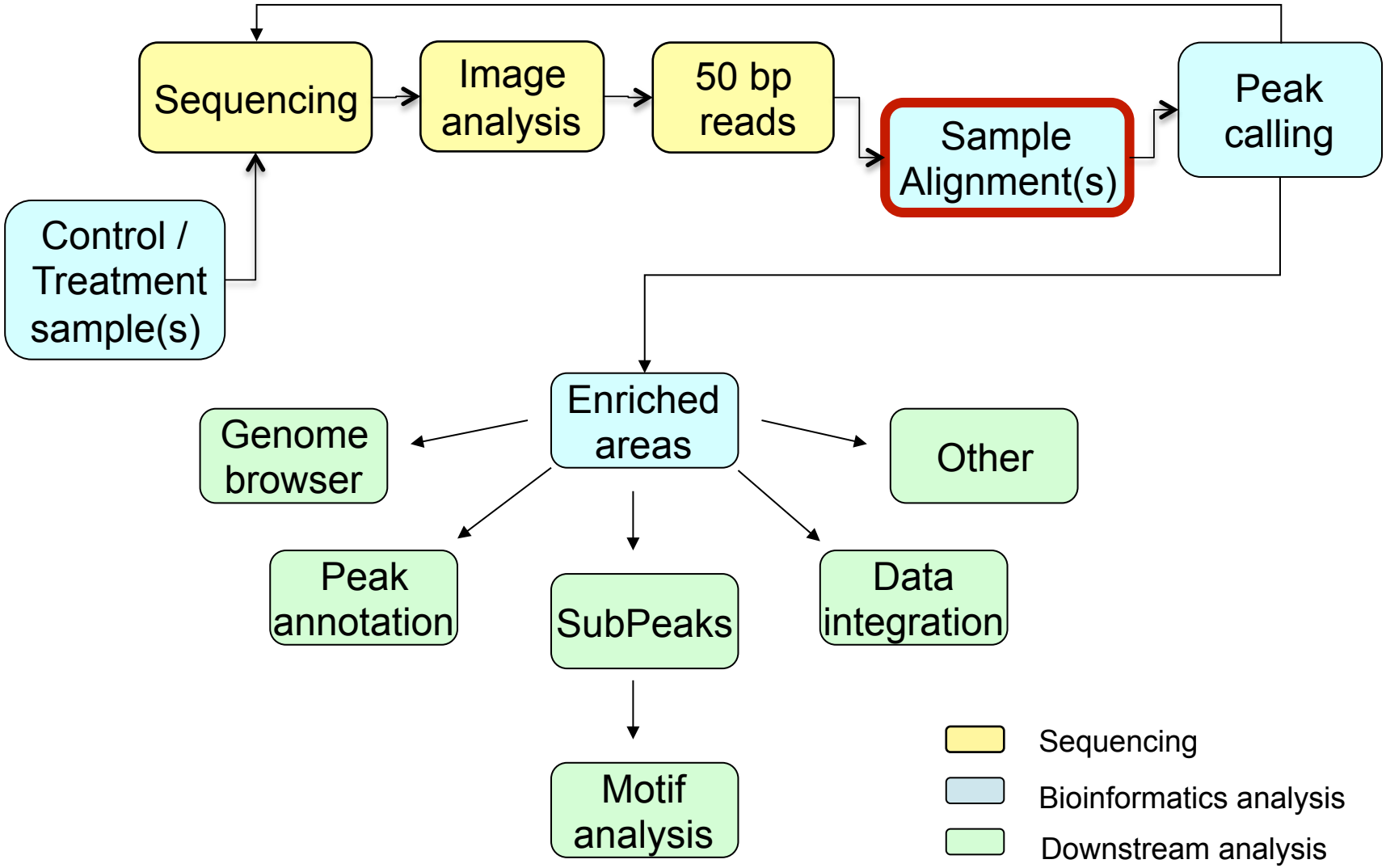
Furey, TS, 2012

Nature Reviews | Genetics

Analysis - Overview



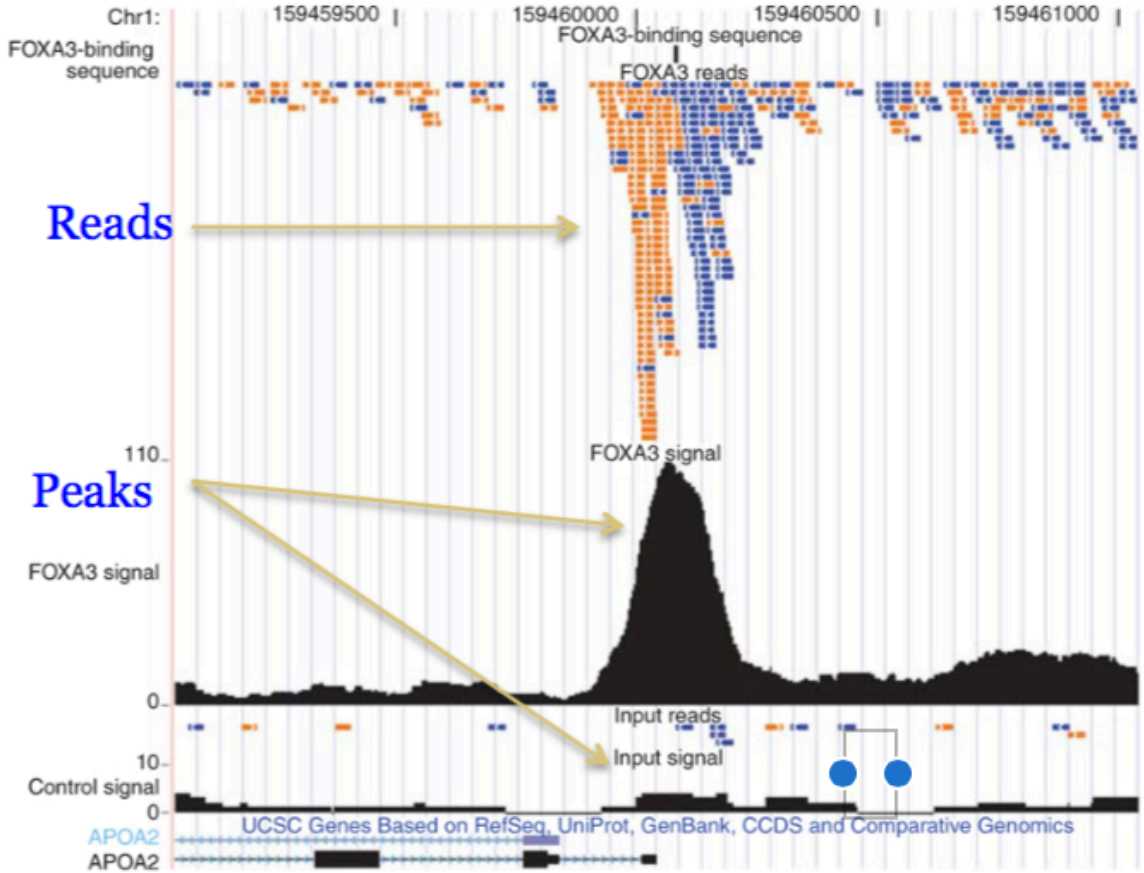
Analysis - Overview



Alignment

Computational mapping of the sequenced DNA identifies the genomic locations of bound

- DNA-binding enzymes,
- modified histones,
- chaperones,
- nucleosomes, and
- transcription factors



Alignment- Genome Mappability

- Not all of the genome is 'available' for mapping
- Align your reads to the unmasked genome

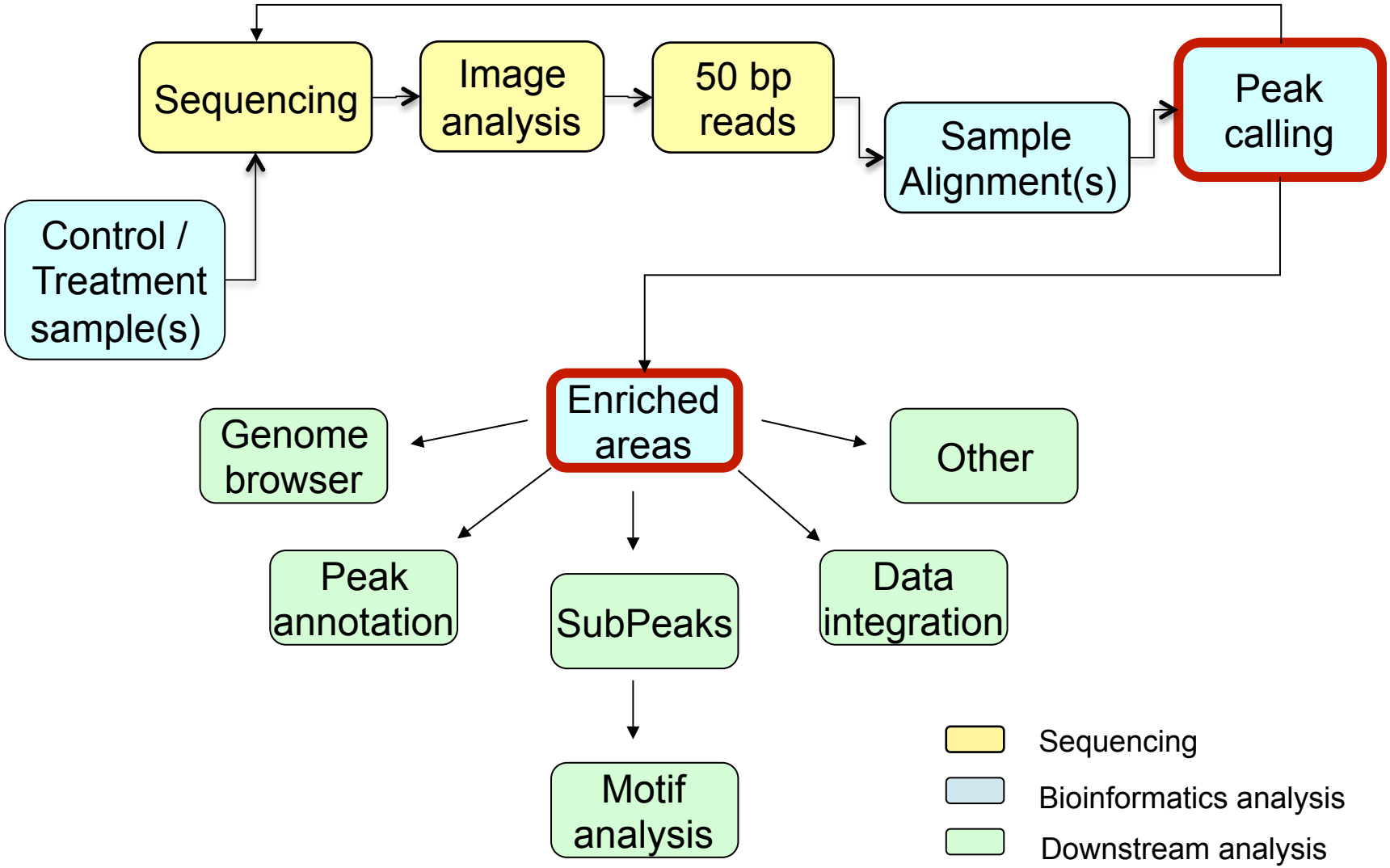
		Nonrepetitive sequence		Mappable sequence	
Organism	**Genome size (Mb)**	**Size (Mb)**	**Percentage**	**Size (Mb)**	**Percentage**
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens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

*Calculated based on 30nt sequence tags

Rozowsky, 2009

- For ChIP-seq, usually short reads are used (50bp)
- Limited gain in using longer reads (again, unless you have a specific interest in repeat regions)

Analysis - Overview



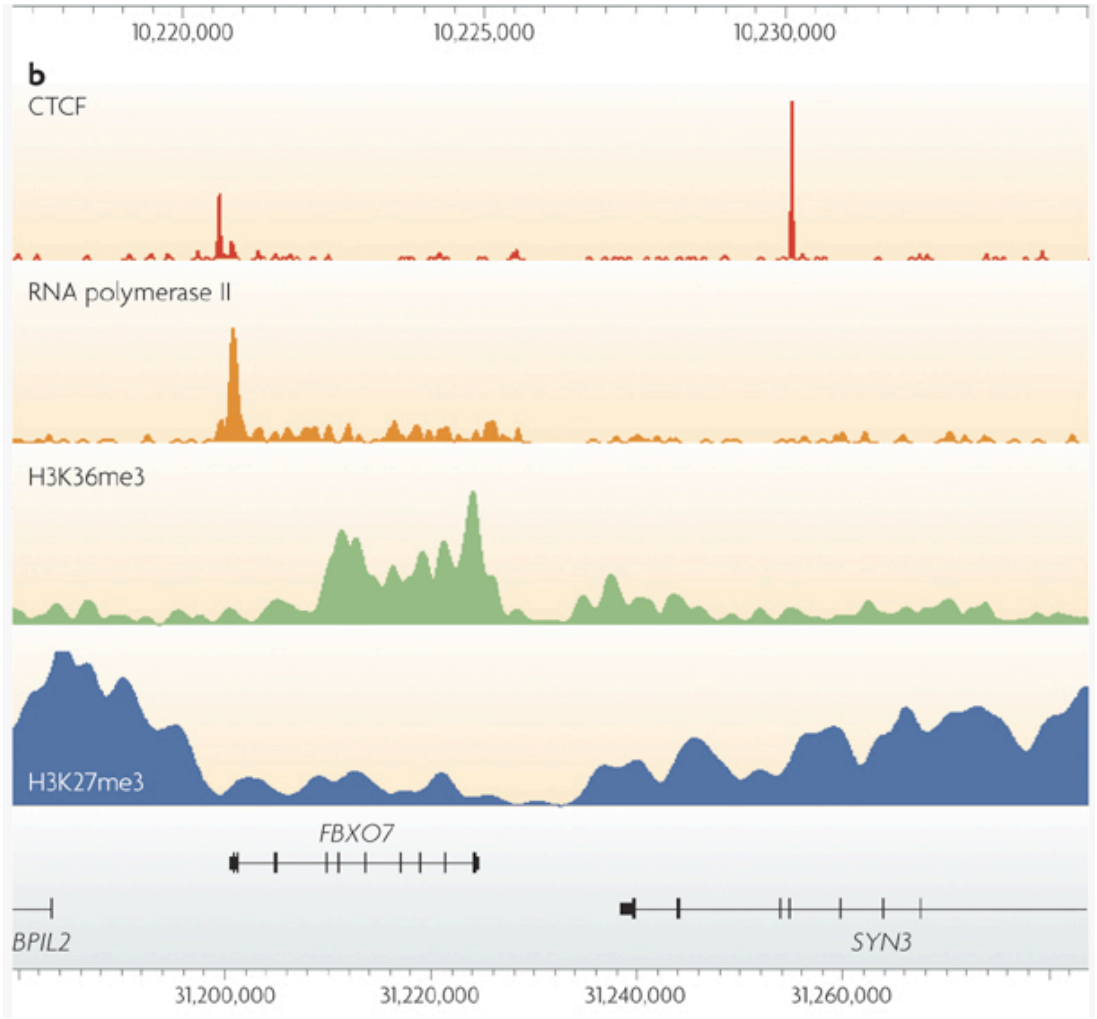
Peak Calling

- Basic - regions are scored by the number of tags in a window of a given size. Then assess by enrichment over control and minimum tag density.
- Advanced - take advantage of the directionality of the reads.
- Advanced methods make more assumptions, making them less appropriate in certain cases

Peaks => regions with "significant" number of mapped reads

Peak Calling - Challenges

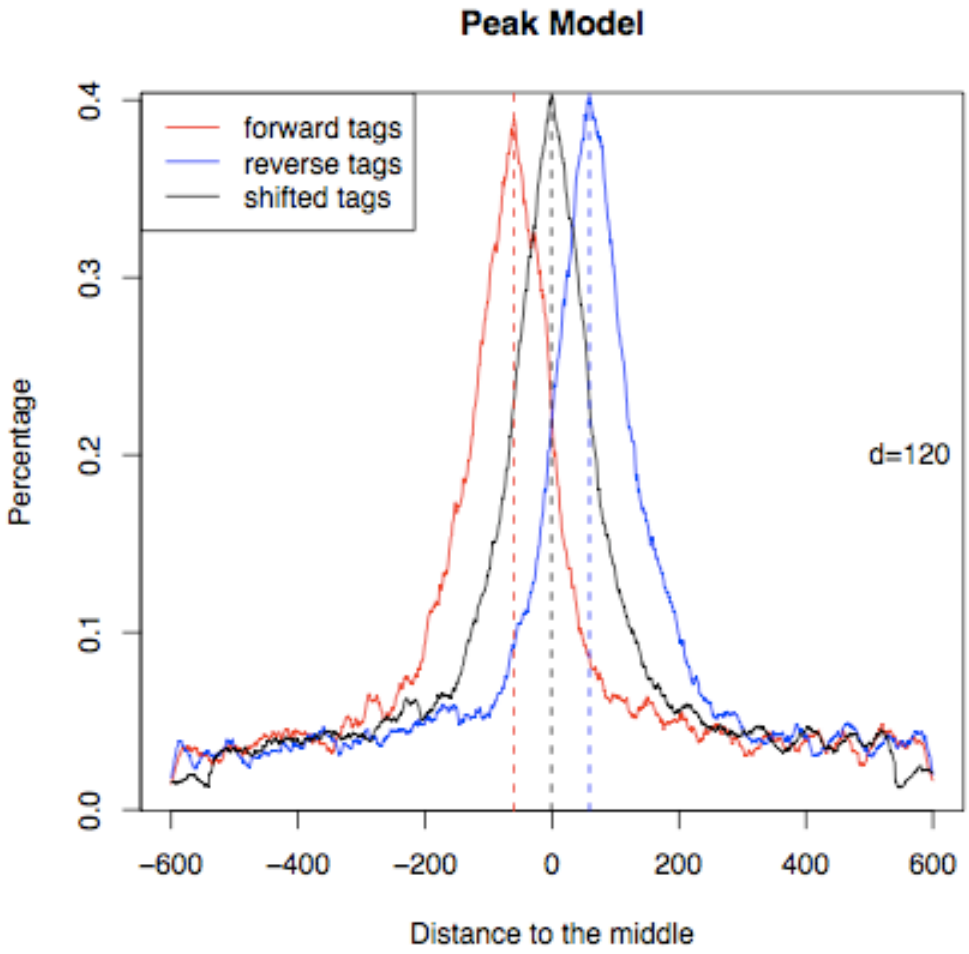
- Adjust for sequence alignments - regions that contain **repetitive elements** have different expected tag count
- Different ChIP-seq applications produce **different type of peaks**. Most current tools have been designed to detect sharp peaks (TF binding, histone modifications at regulatory elements)
- Alternative tools exist **for broader peaks** (histone modifications that mark domains - transcribed or repressed), e.g. SICER



Park J, Nature Reviews Genetics, 2009

MACS tool

- Model the shift size between +/- strand tags
  - Scan the genome to find regions with tags more than m-fold enriched relative to random tag distribution
  - Randomly sample 1000 of these (high quality peaks) and calculate the distance between the modes of their +/- peaks
  - Shift all the tags by d/2 toward the 3' end.



MACS - Peak detection

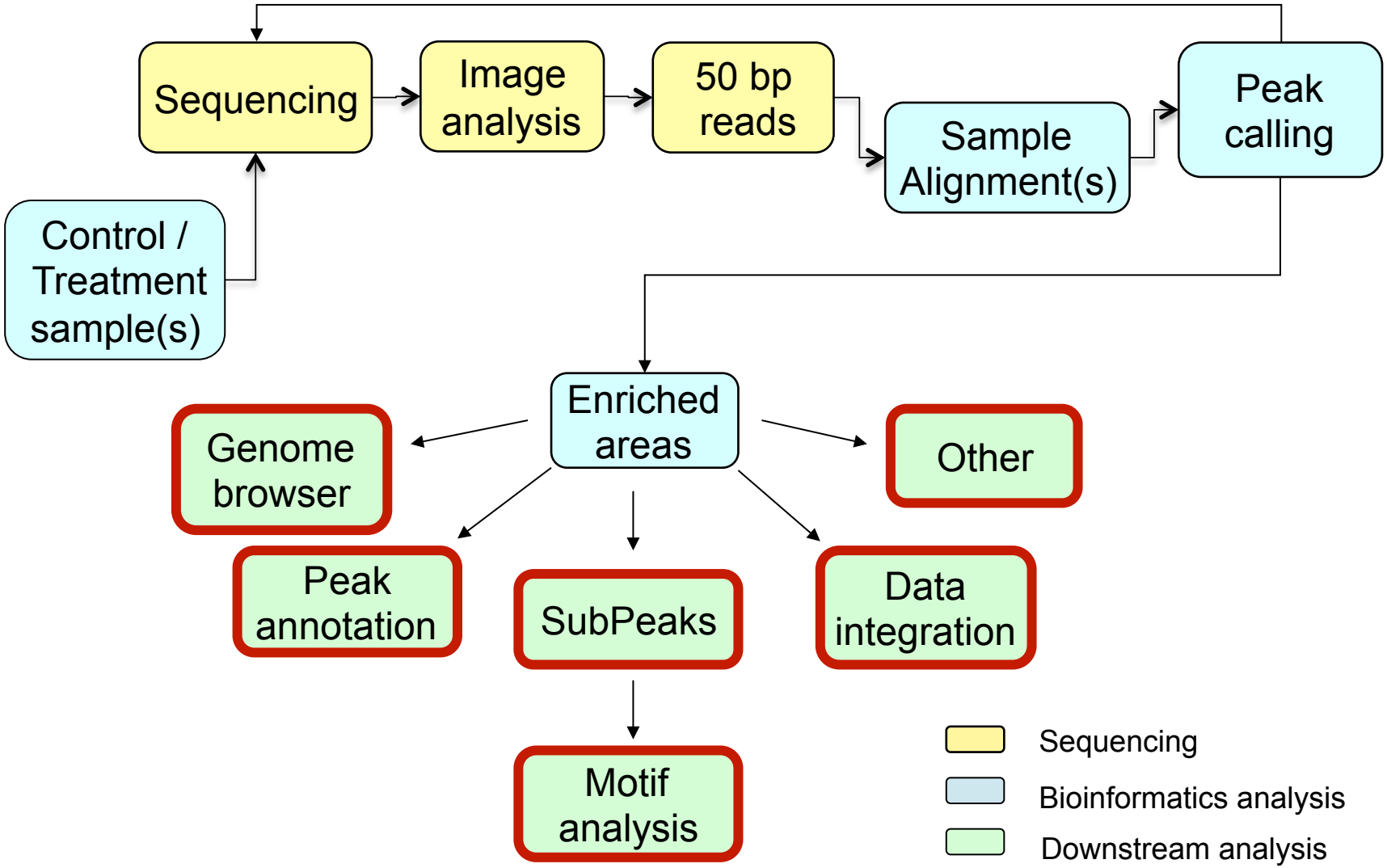
1. Duplicate tags are removed (in excess of what can be expected by chance)

2. Candidate peaks with significant tag enrichment are found in a sliding window across the genome (Poisson distribution, global background, p-value 10e-5)

3. Overlapping peaks are merged, and each tag base distance extended from its center

4. Peaks are eliminated that are not significant with respect to local background levels. The control sample is used to eliminates peaks that are also significantly represented in the location.

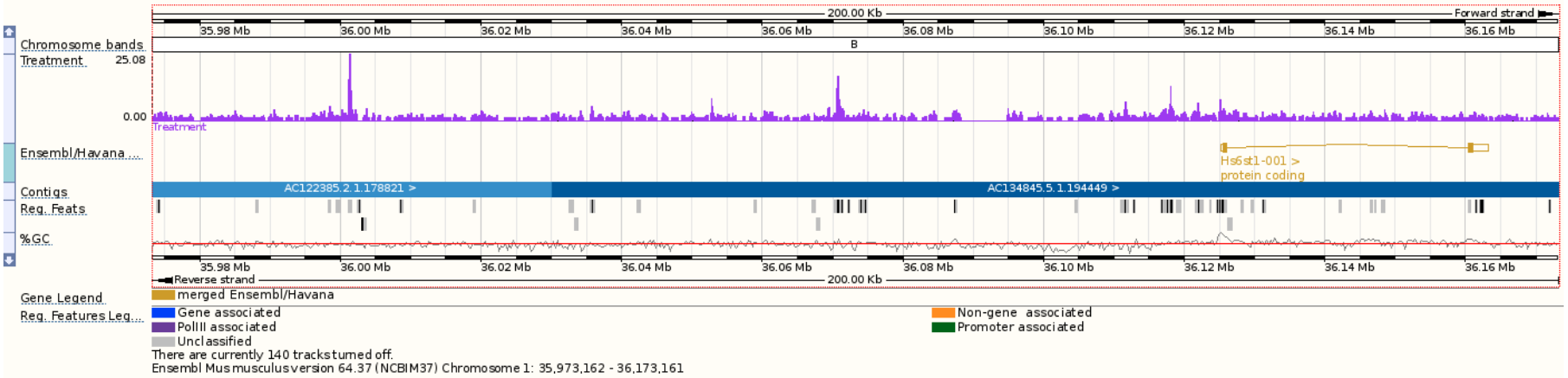
Analysis - Overview



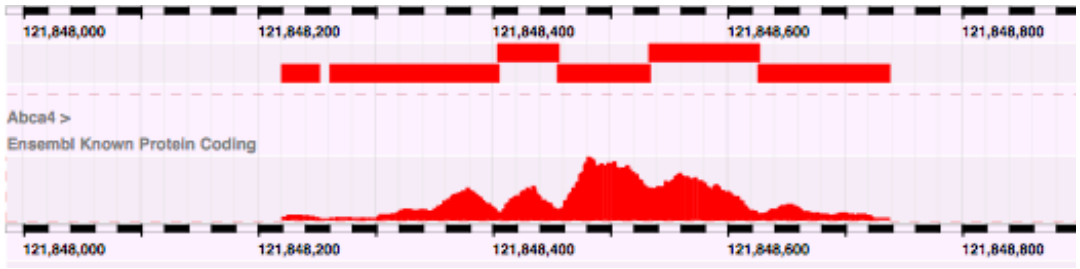
Analysis downstream to peak calling

- **Peak Annotation:** finding interesting **features surrounding peak** regions: PeakAnalyzer
- **Visualization:** genome browser: Ensembl, UCSC, IGV
- **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 expression data
- Correlation with SNP data to find allele-specific binding

Visualization in a genome browser



Motif Analysis

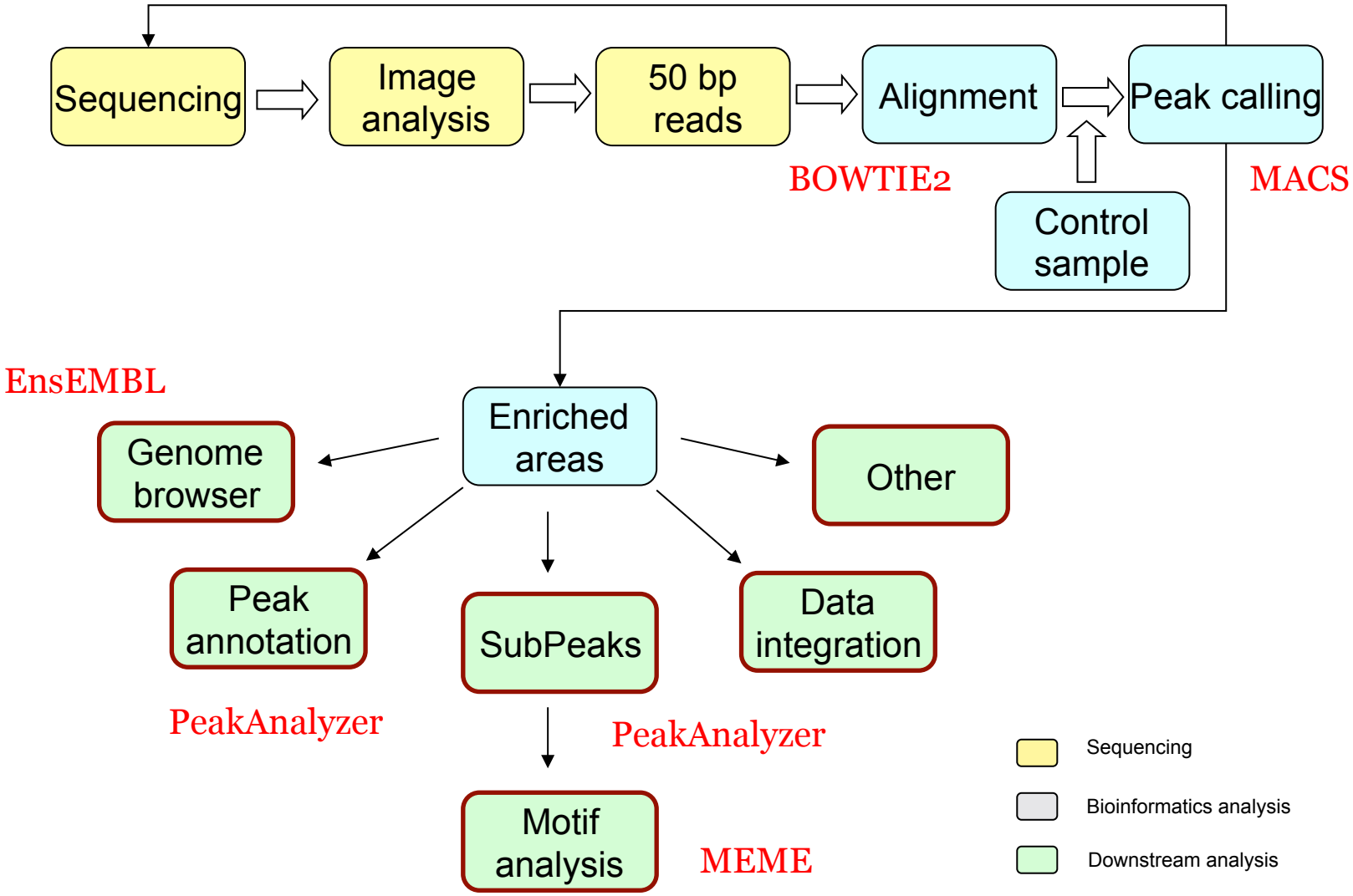


GAATCCCACA TTTGCATAACAAAAG ACTCCTGGTG
CAGCTGCTCT TCTGCATAACAAAGG GTGGCCCTGC
CCGGTTTTTC TTTGCATAACAATAA GATCTGGCTA
TTATTCTCAC TTTGCATAGGAATGG GGCAGTTAGA
CACAGCCACA TTTGCATAACAGAAG CCGAGCCCGC
CTTGGGTGAA TTTGCAAGACAAAGG ACAATGATCA

Discovery of binding sequence motifs

1. Split peaks
2. Fetch summit sequences
3. Run motif prediction tool

Analysis - Overview



Chip-Seq Overview

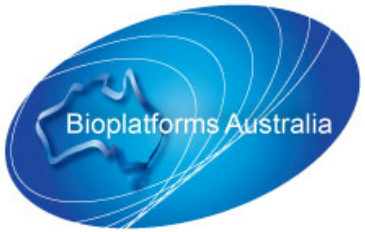
- Introduction to Chip-Seq
- Experimental Design
- Overview of Analysis
- Introduction to hands-on workshop
  - Let's Do

Hands-on

- The data we will use today was reported in Chen, X et al. (2008) Integration of external signaling pathways with the **core transcriptional network in embryonic stem cells.** Cell. Jun 13;133(6):1106-17 (http://www.cell.com/cell/pdf/S0092-8674(08)00617-X.pdf)

- You have already performed the first step, alignment of the reads to the genome, in the previous session**. We start from the aligned reads**. Go to the Chip-Seq module in your electronic handout.

- **In a terminal shell go to the /home/trainee/chipseq** directory where we will perform simple ChIP-Seq analysis
  - Detect immuno-enriched areas using the peak caller program MACS
  - Visualize peak regions in the Ensembl genome browser
  - Perform functional annotation (PeakAnalyzer) and detect potential binding sites (motifs) in the predicted binding regions using motif discovery tool, MEME.



Thank you

EMBL-EBI