Introduction to ChIP-seq

Myrto Kostadima

Ensembl Regulation Project Leader

myrtullina

EMBL-EBI

Ensembl

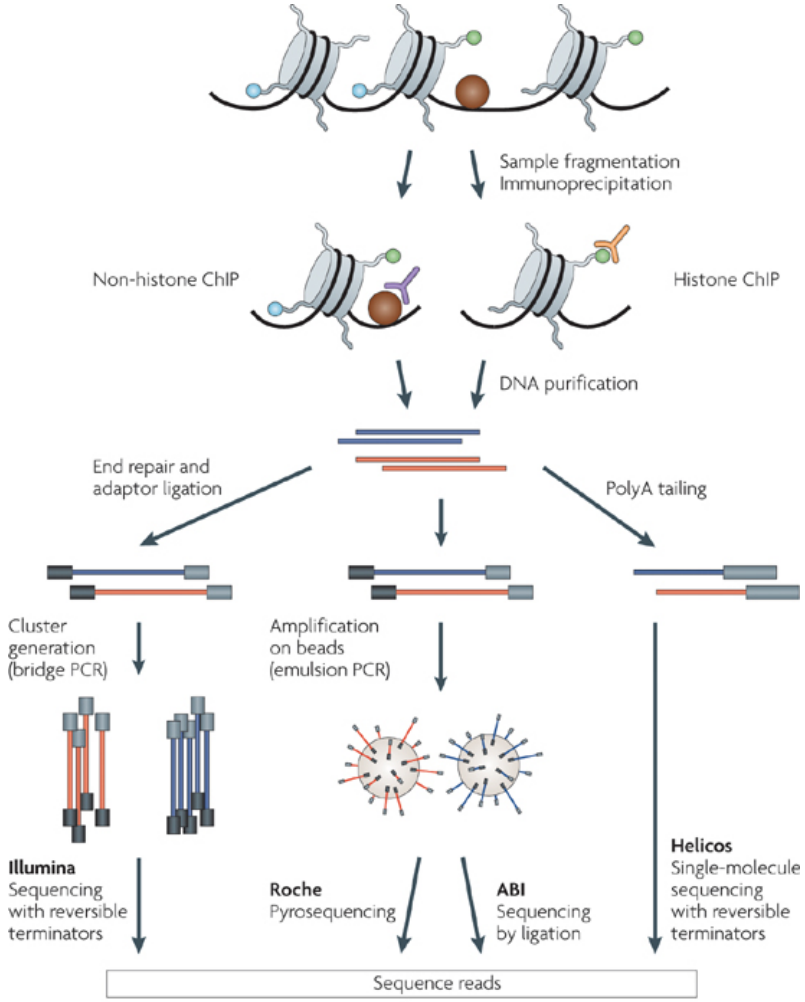
CHROMATIN IMMUNOPRECIPITATION FOLLOWED BY SEQUENCING

- One of the early applications of NGS
- First studies published in 2007
  - Johnson et al (Science) - NRSF
  - Barski et al (Cell) - histone methylation
  - Robertson et al (Nature Methods) - STAT1
  - Mikkelsen et al (Nature) - histone modification
- ~2,700 publications currently in PubMed

APPLICATIONS

- Protein-DNA interaction
  - Identification of transcription factor binding, core transcriptional machinery
- Histone modifications

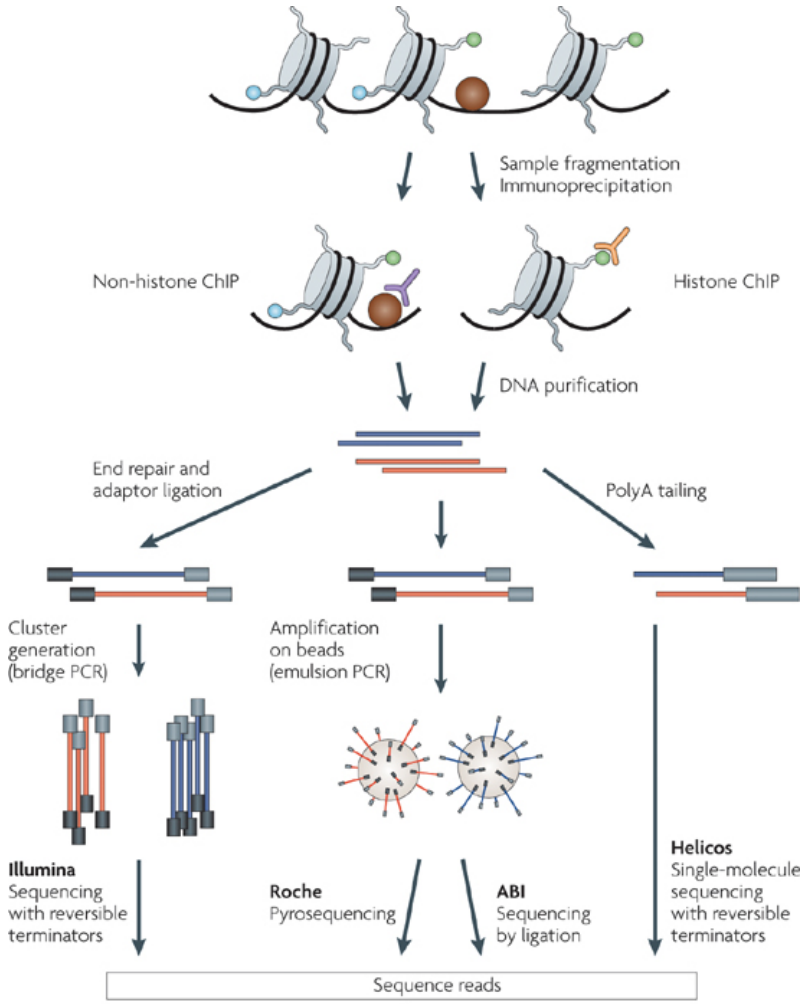
EXPERIMENT OVERVIEW



Peter J. Park (2009)

Nature Reviews | Genetics

EXPERIMENT OVERVIEW



Peter J. Park (2009)

Resolution	High - single nucleotide
Coverage	Limited by "alignability" of reads to the genome, increases with read length
Repeat elements	Many can be covered (only 80% of the human genomes uniquely mappable)
Cost	Around 1000$ per lane; 20-30M reads
Source of noise	Sequencing bias, GC bias, sequencing error
Amount of ChIP DNA required	Low 10-50ng
Dynamic range	Not limited
Multiplexing	Possible

SAMPLE PREPARATION

	Transcription factor binding	Histone modifications and Nucleosome positioning
Crosslinking	Formaldehyde	Usually not
Fragmentation	Sonication (200-600bp)	Mnase treatment
Immunoprecipitation	Antibody specific to protein	Antibody specific to histone modification or histone

Library construction

- Size selection ~150-300bp
- Adapter ligation
- Cluster generation (amplification)
- Sequence by synthesis

EXPERIMENTAL DESIGN

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

ANTIBODY QUALITY

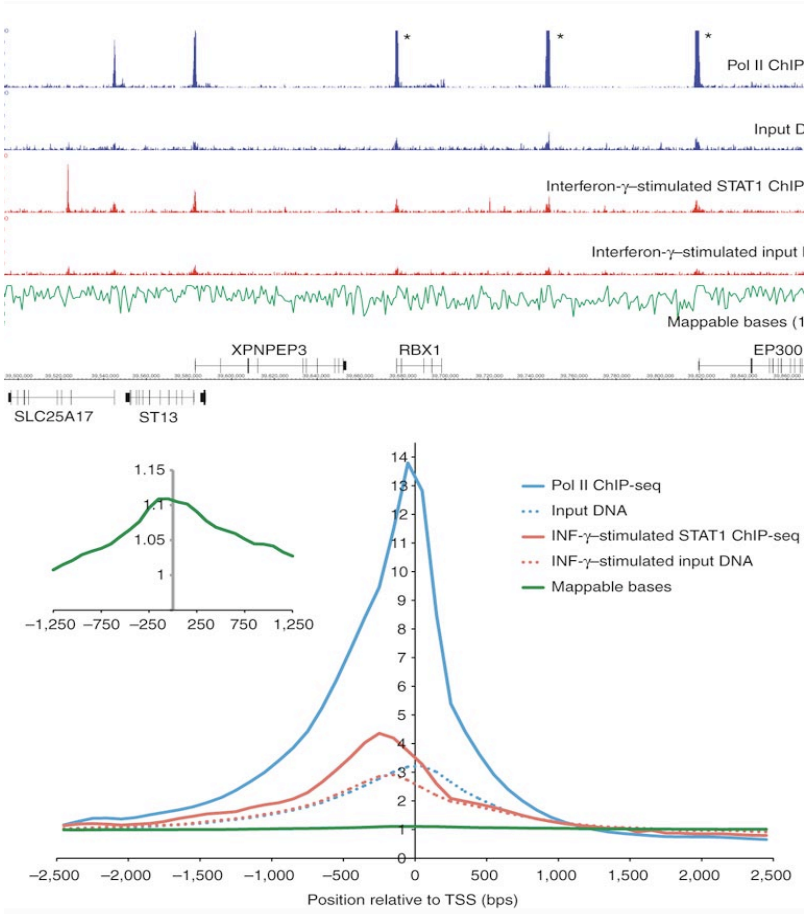
- Antibody quality - a sensitive and specific antibody will give a high level of enrichment
  - Limited efficiency of antibody is the main reason for failed ChIP-seq experiments
- Check your antibody ahead if possible.
  - Western blotting to check the reactivity of the antibody with unmodified and non-histone proteins.
- Optimize ChIP protocol
  - If known positives and negatives are available, perform qPCR to demonstrate enrichment for these regions

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, (2009) Nature Biotechnology

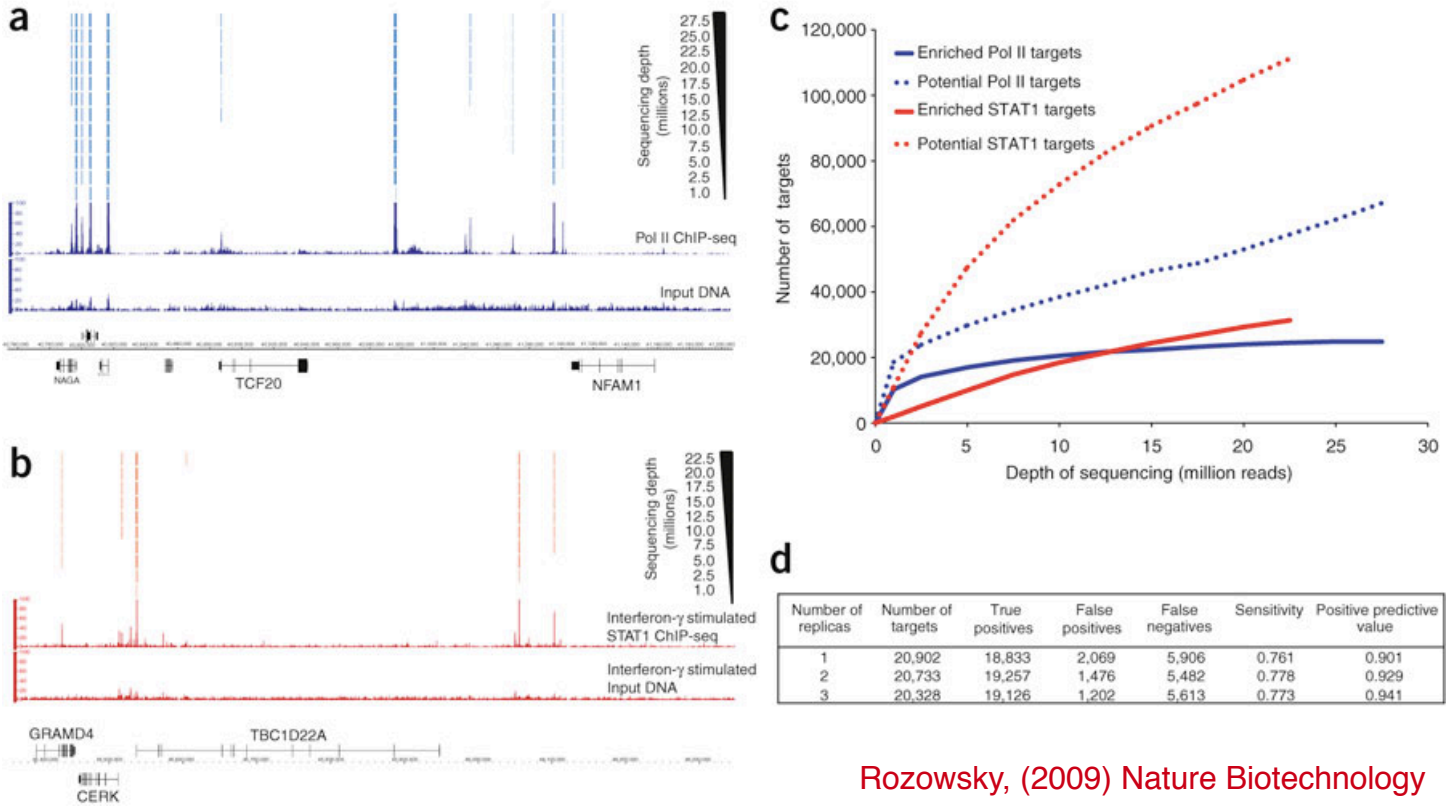
CONTROL TYPE

- 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
- Sequencing a control can be avoided when looking at:
  - time points
  - differential binding pattern between conditions

EXPERIMENTAL DESIGN

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

SEQUENCING DEPTH



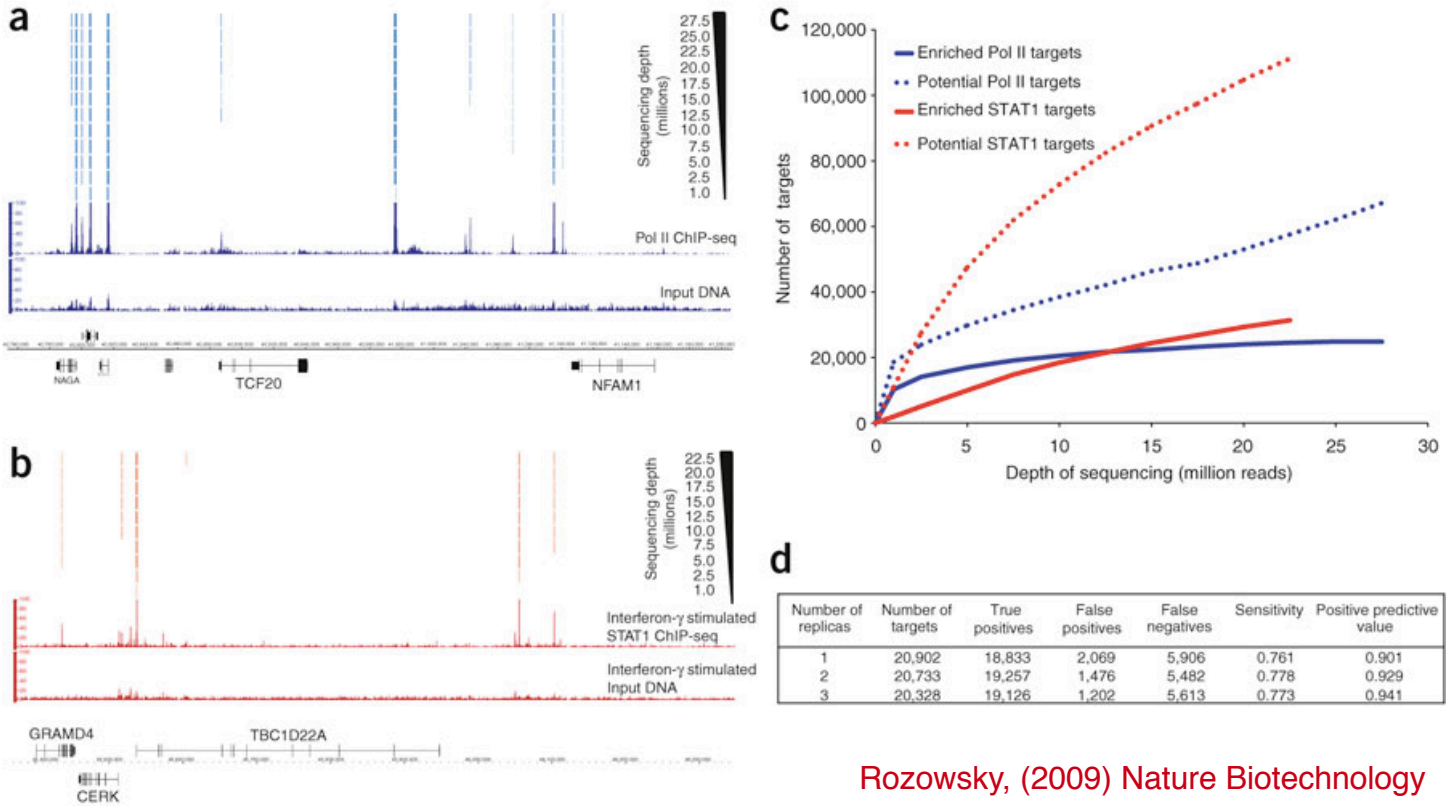
d

Number of replicas	Number of targets	True positives	False positives	False negatives	Sensitivity	Positive predictive value
1	20,902	18,833	2,069	5,906	0.761	0.901
2	20,733	19,257	1,476	5,482	0.778	0.929
3	20,328	19,126	1,202	5,613	0.773	0.941

Rozowsky, (2009) Nature Biotechnology

- More prominent peaks are identified with fewer reads, versus weaker peaks that require greater depth
- Number of putative target regions continues to increase significantly as a function of sequencing depth

SEQUENCING DEPTH



d

Number of replicas	Number of targets	True positives	False positives	False negatives	Sensitivity	Positive predictive value
1	20,902	18,833	2,069	5,906	0.761	0.901
2	20,733	19,257	1,476	5,482	0.778	0.929
3	20,328	19,126	1,202	5,613	0.773	0.941

Rozowsky, (2009) Nature Biotechnology

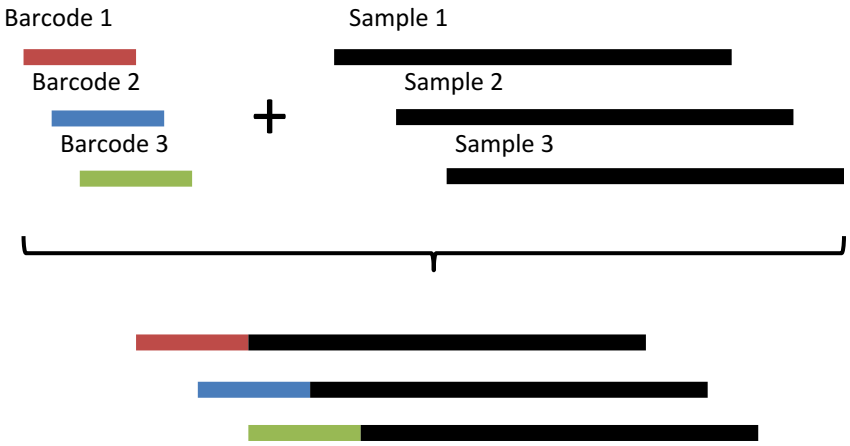
- GA2X 18-30M, MiSeq 25M, HiSeq up to 350M
- With current sequencing technologies for human/mouse >20M uniquely mapped duplicate free reads is usually sufficient.

EXPERIMENTAL DESIGN

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

Sample barcoding and de-multiplexing

Barcoding



De-multiplexing

- # Read 1
  - ATTAGACCTAAGCA
- # Read 2
  - GAGCAACGACTAC
- # Read 3
  - ATTAGGCCATACAT
- # Read 4
  - CCATAGGCTGACTA
- ...

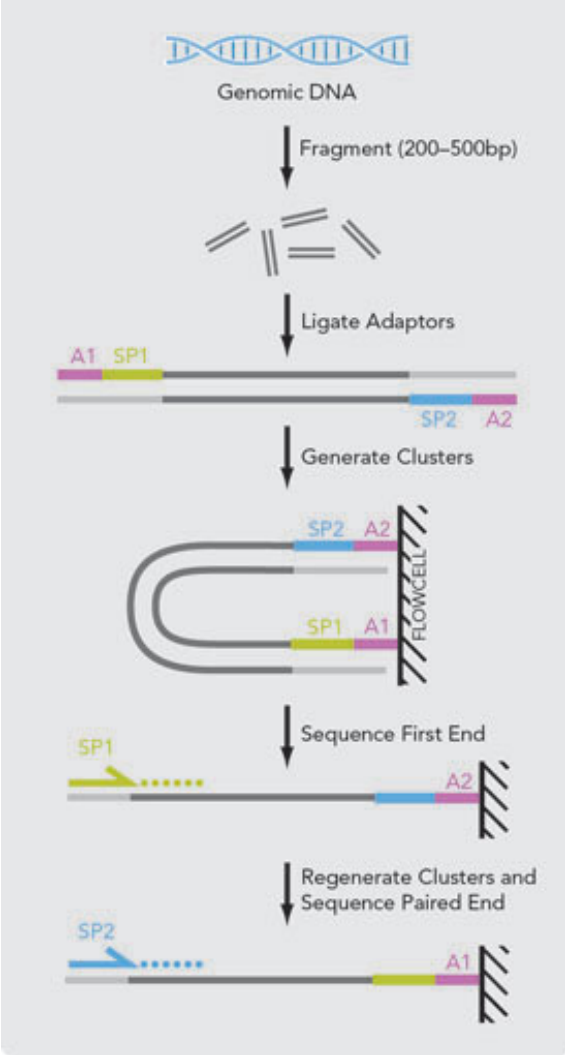
EXPERIMENTAL DESIGN

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

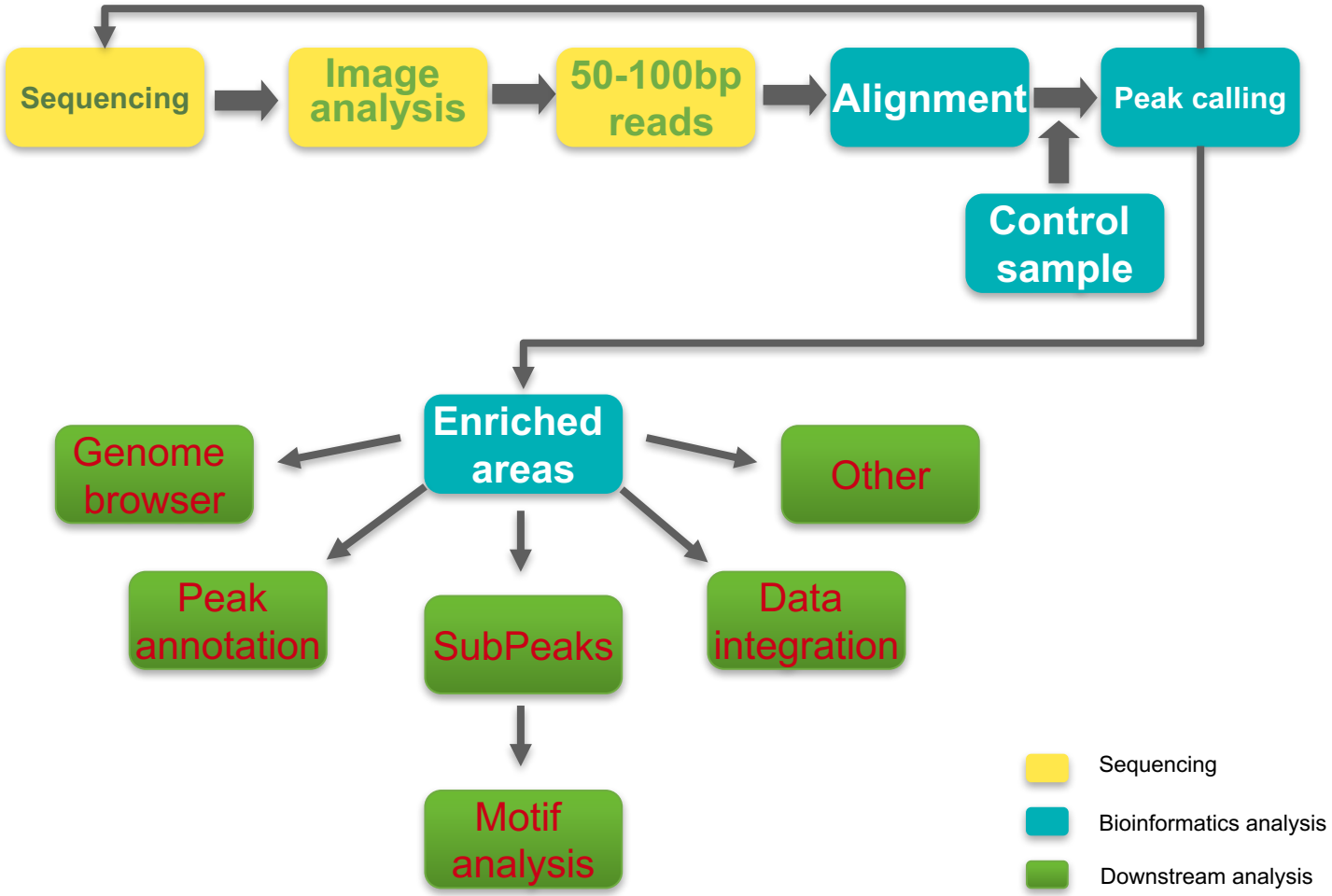
PAIRED-END SEQUENCING

- 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



ANALYSIS OVERVIEW



(Short) Read Alignment

GOAL: Given a reference sequence and a set of short reads, align each read to the reference sequence

MAPPABILITY

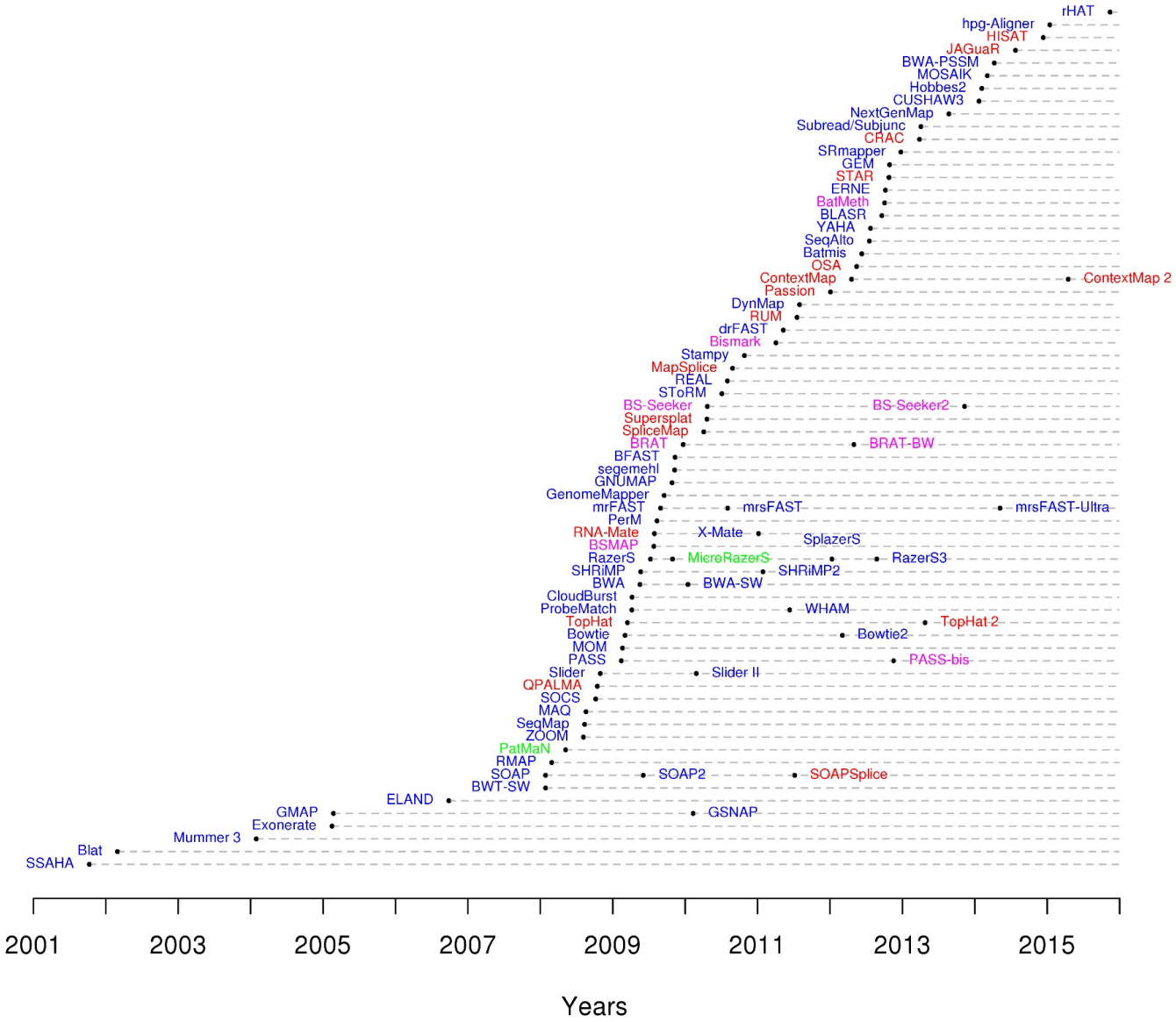
- Not all of the genome is ‘available’ for mapping
- Align your reads to the unmasked genome

		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 (50/100bp)
- Limited gain in using longer reads (again, unless you have a specific interest in repeat regions)



DNA

RNA

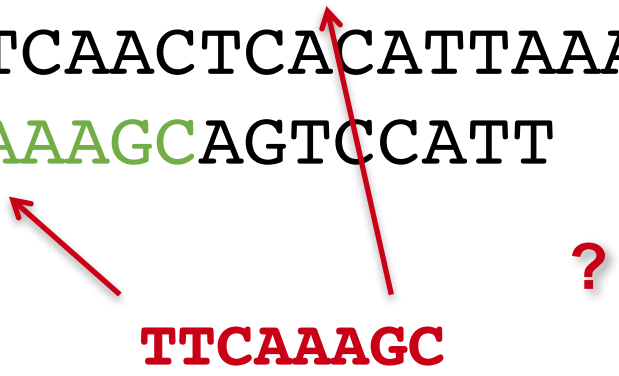
Methylation

microRNA

http://www.ebi.ac.uk/~nf/hts_mappers/

Reads can map in multiple locations

```
>CHROMOSOME_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAA
TCCATTTGTTCAACTCACATTAAATAGTCGATCAAAT
AGTTGGTTCAAAGCAGTCCATT
```



- Some parts of the genome will not be unique:
  - Common, repeated motifs (proteins domains)
  - Repeat regions

This can not always be resolved

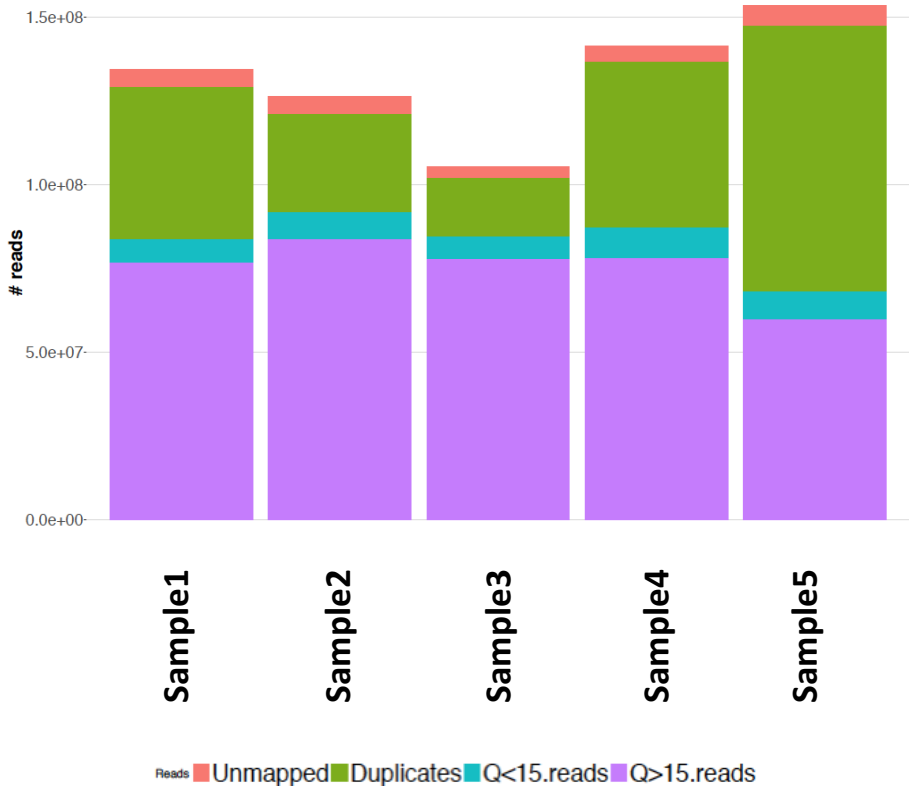
And what do I do with duplicates?

- Reads that align in exactly the same place (same start + same CIGAR string)
- Duplicates can occur from:
  - Artefacts from sequencing (PCR artefacts)
  - Real biological signal
  - We cannot tell apart which one, unless we use barcodes.

Parekh S et al., Sci Rep. 2016 May 9;6:25533.

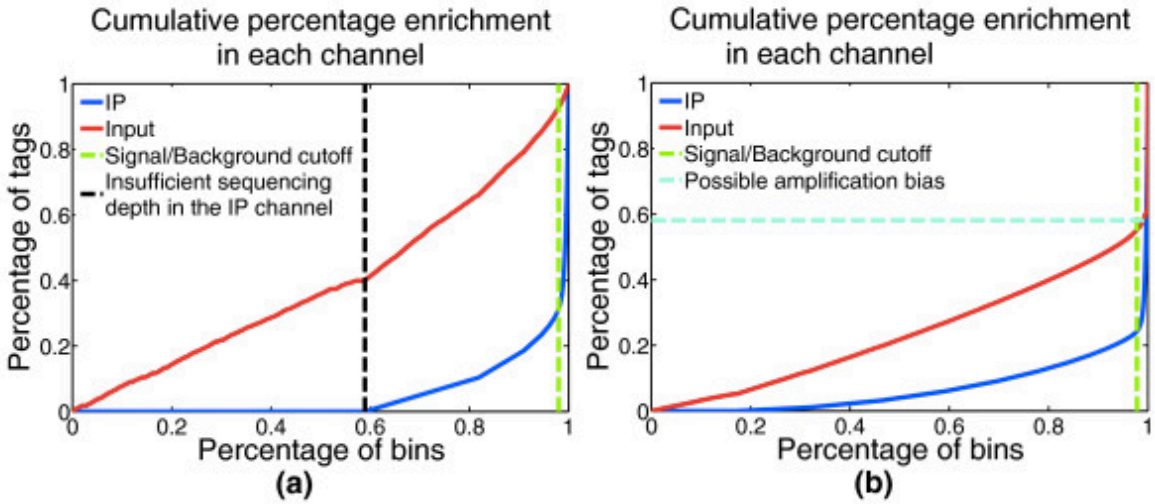
LET'S TALK ABOUT QUALITY CONTROL

Read alignment statistics



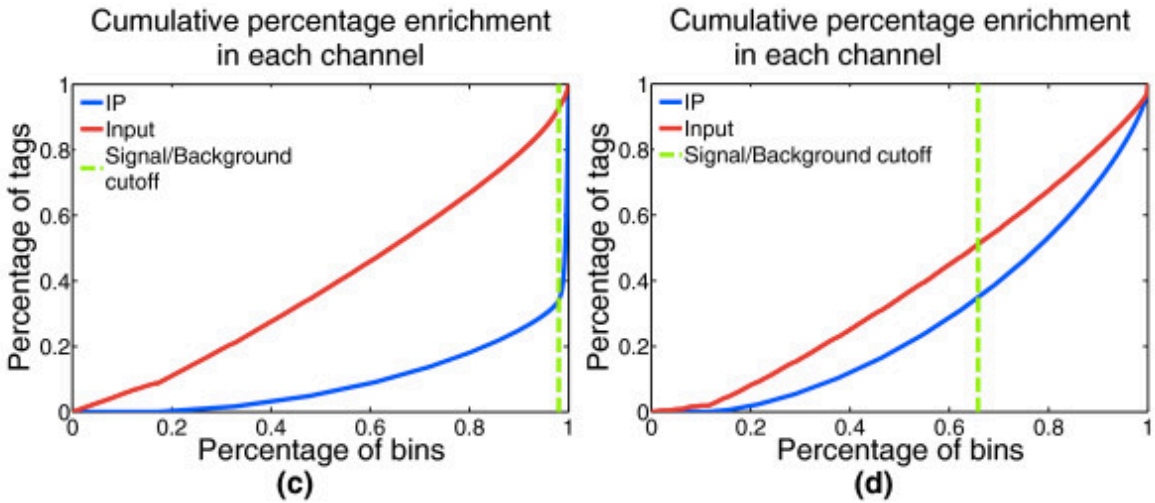
Did my ChIP experiment work?

60% of the genome didn't have sufficient coverage in the IP



over 60% of the reads map to a small percentage of the genome, indicating amplification bias

IP worked

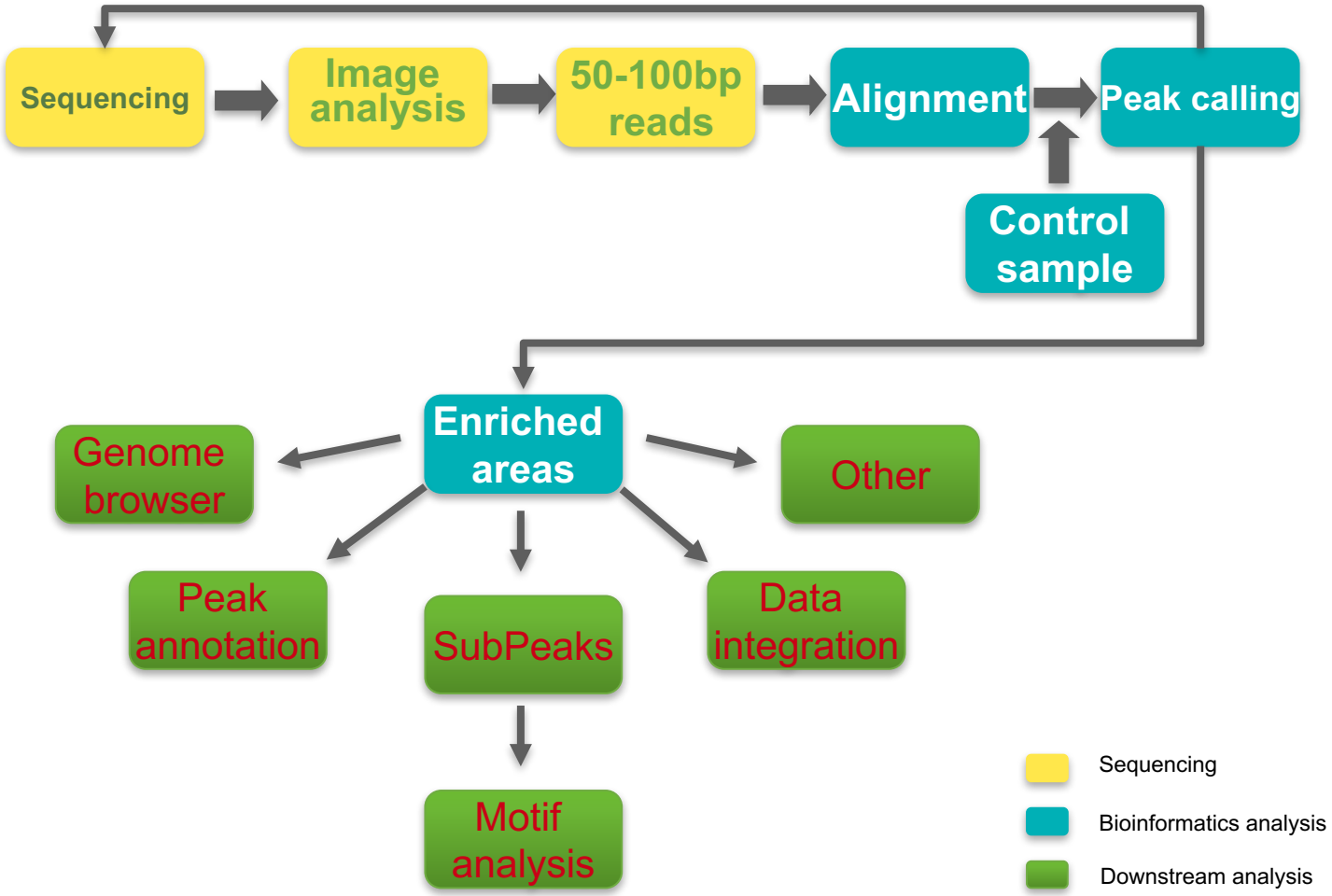


weak IP: IP and Input curves are not well separated

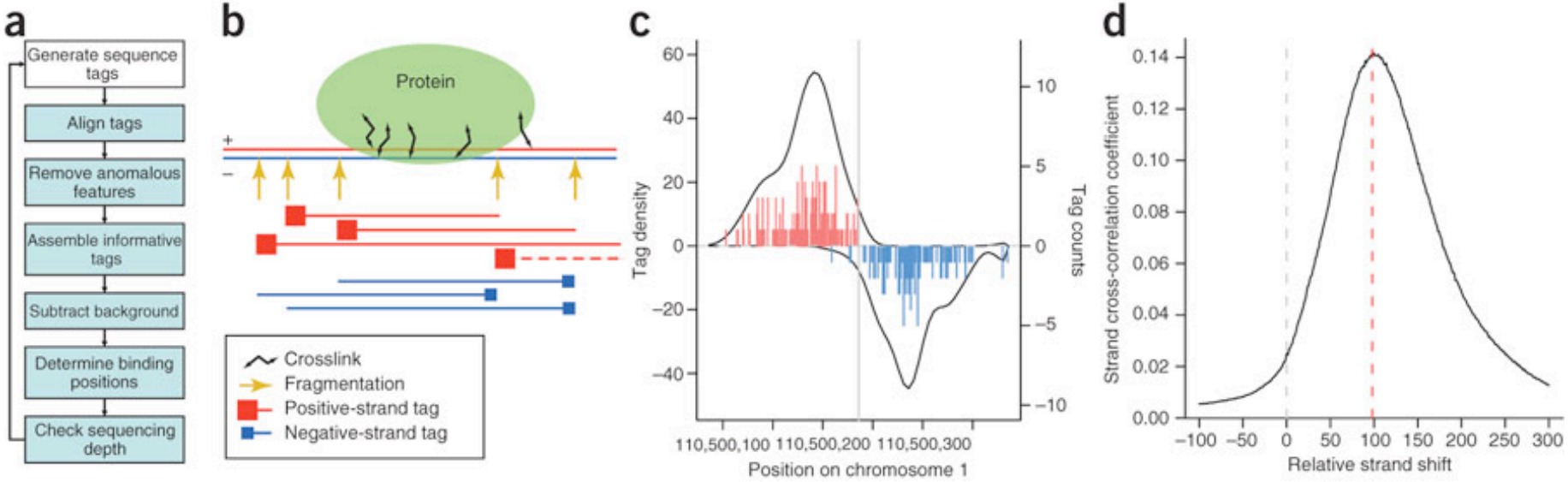
Diaz et al. (2012)
https://github.com/songlab/chance

https://deeptools.readthedocs.io/en/latest/content/tools/plotFingerprint.html

ANALYSIS OVERVIEW



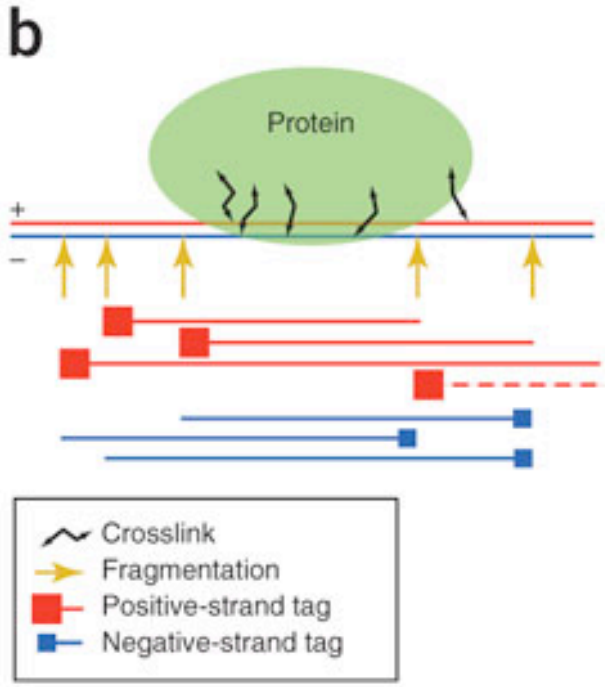
STRAND SPECIFIC PROFILE



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.



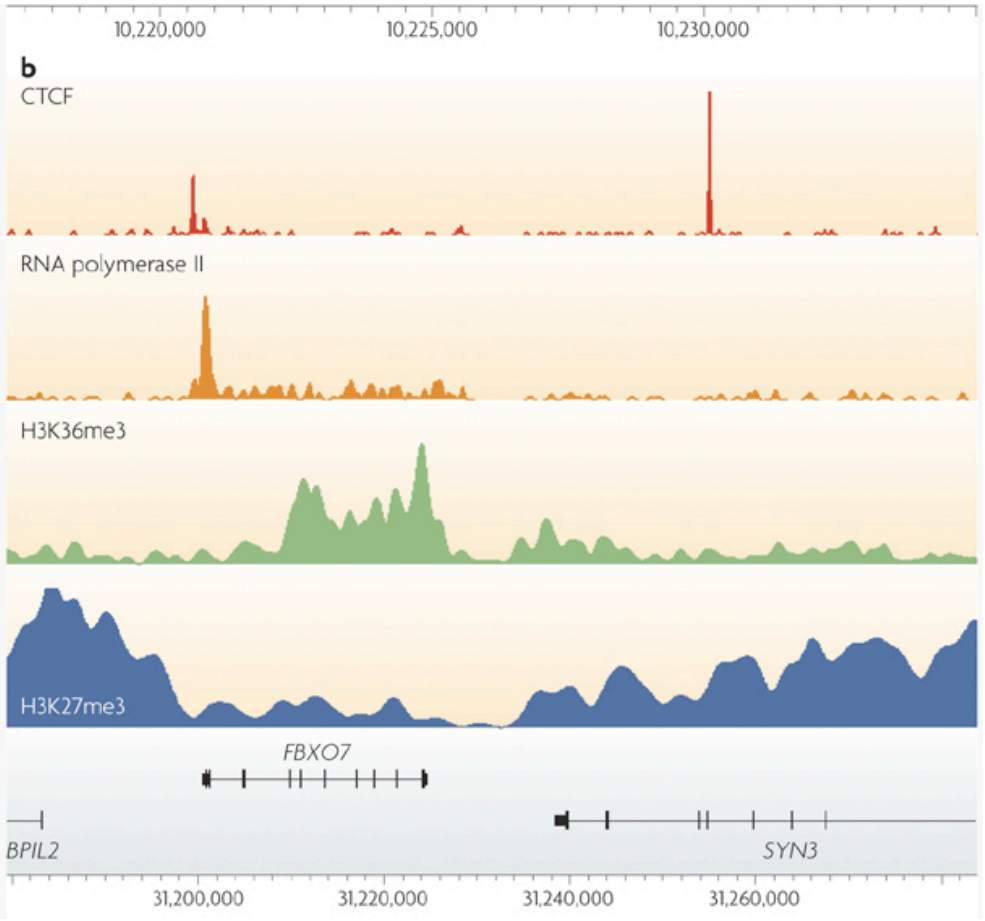
Kharchenko (2008) Nature Biotechnology

PEAK CALLING-CHALLENGES

- Adjust for sequence alignability - 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)

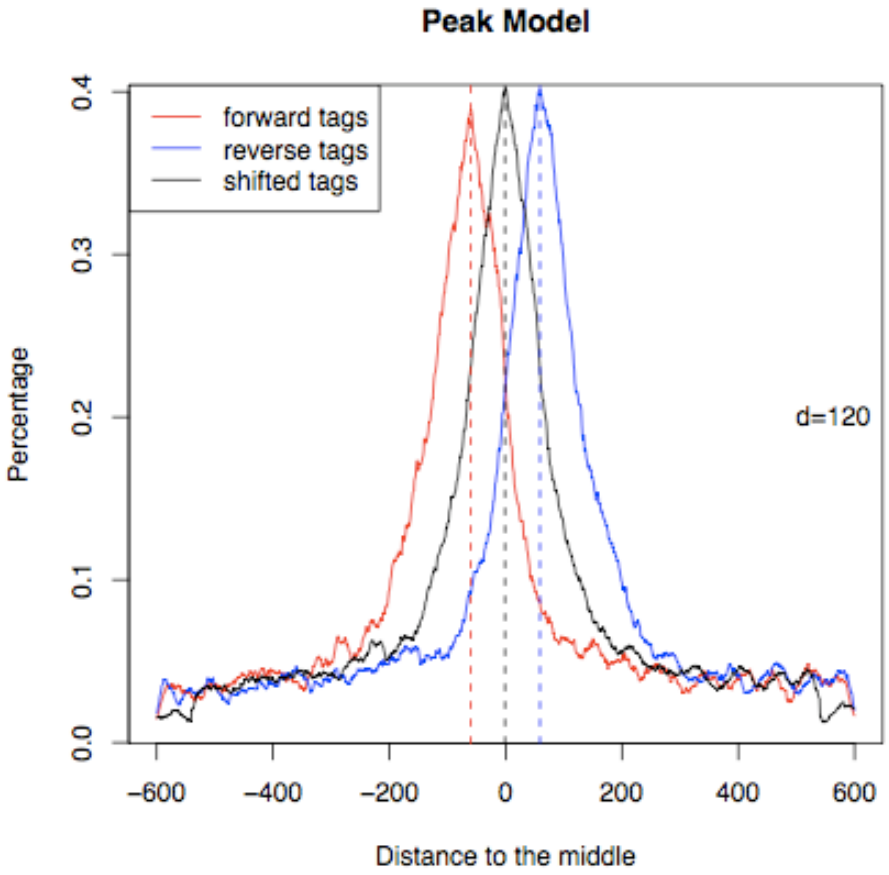
- Alternative tools exist for broader peaks (histone modifications that mark domains - transcribed or repressed), e.g. SICER



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 mfold (between 10-30) 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.

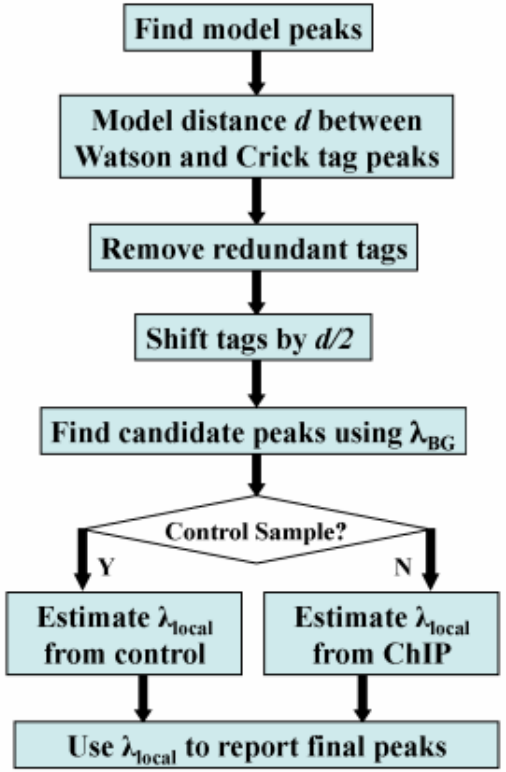


Is a peak greater than expected by chance?

- x = observed read count
- λ= expected read number
- P = Probability to find a peak higher than x

$$\lambda = \frac{(read\_size) * (mapped\_reads)}{Alignable\_genome\_size}$$

$$P = 1 - \sum_{k=0}^{x-1} \frac{e^{-\lambda} \lambda^{k}}{k!}$$



- Tag distribution along the genome can be modelled by a Poisson distribution

Slide adapted from http://www.slideshare.net/lucacozzuto/macs-course

e.g. non –significant peak

tag count = 2
total reads = 30,000,000
read length = 36
mappable human genome = 2,700,000,000

$$\lambda_{BG} = \frac{(36) * (30,000,000)}{2,700,000,000} = 4$$

$$P = 1 - \sum_{k=0}^{1} \frac{e^{-4} * 4^{k}}{k!} = 0.9$$

Slide adapted from http://www.slideshare.net/lucacozzuto/macs-course

e.g. significant peak

tag count = 10
total reads = 30,000,000
read length = 36
mappable human genome = 2, 700,000,000

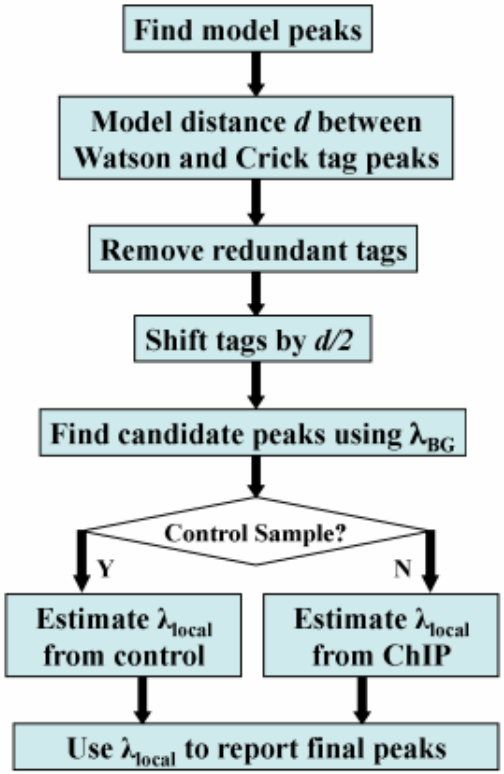
$$\lambda_{BG} = \frac{(36) * (30{,}000{,}000)}{2{,}700{,}000{,}000} = 4$$

$$P = 1 - \sum_{k=0}^{9} \frac{e^{-4} * 4^{k}}{k!} = 0.008$$

Slide adapted from http://www.slideshare.net/lucacozzuto/macs-course

The benefit of having a control ChIP experiment

- Candidate peaks are evaluated via comparing them against a “local” distribution.
- Fold enrichment = Enrichment over the λ_{local}
- False Discovery Rate (FDR) is calculated (#peaks in control) / (#peaks in IP) Control peaks are calculated by swapping control and sample.



Slide adapted from http://www.slideshare.net/lucacozzuto/macs-course

MACS PEAK DETECTION

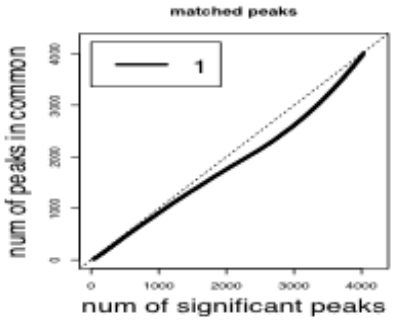
- Remove duplicate tags (in excess of what can be expected by chance)
- Slide window across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution, global background, p-value 10e-5)
- Also looks at local background levels and eliminates peaks that are not significant with respect to local background
- Uses the control sample to eliminate peaks that are also present there

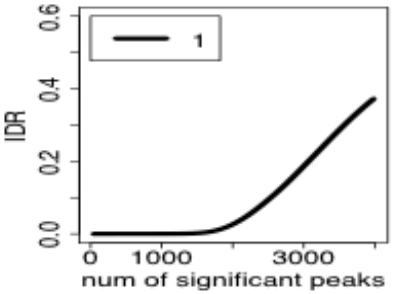
Do we need replicates?

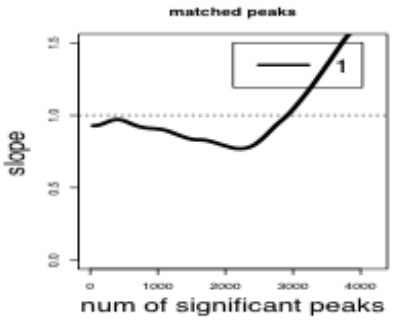
- To ensure that experiments are reproducible at least 2 replicates must be performed.
- To access replicate agreement, Irreproducible Discovery Rate analysis can be used:
  - If replicates measure the same biology high scores (e.g. low p-values) represent strong evidence of being genuine signals and are ranked high and also more consistently on the replicates than those of spurious signals.
- Broad histone modifications are hard to quantify for reproducibility

IDR ANALYSIS

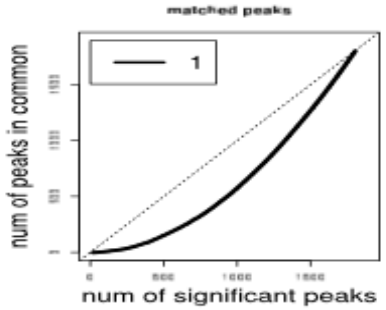
- Plotting the consistency between a pair of rank lists (that contains both significant and insignificant peaks) will indicate how many peaks have been reliably detected as the point showing change in consistency
- IDR score for each signal: reflects the probability for the signal to belong to the irreproducible group.

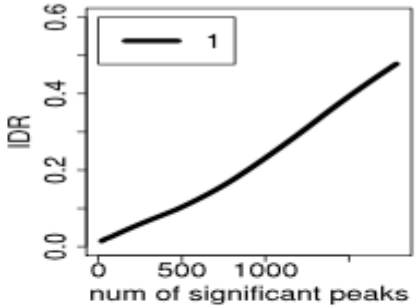


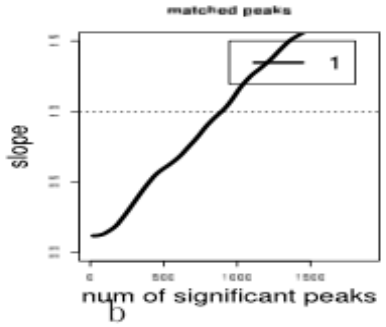




reasonable consistency

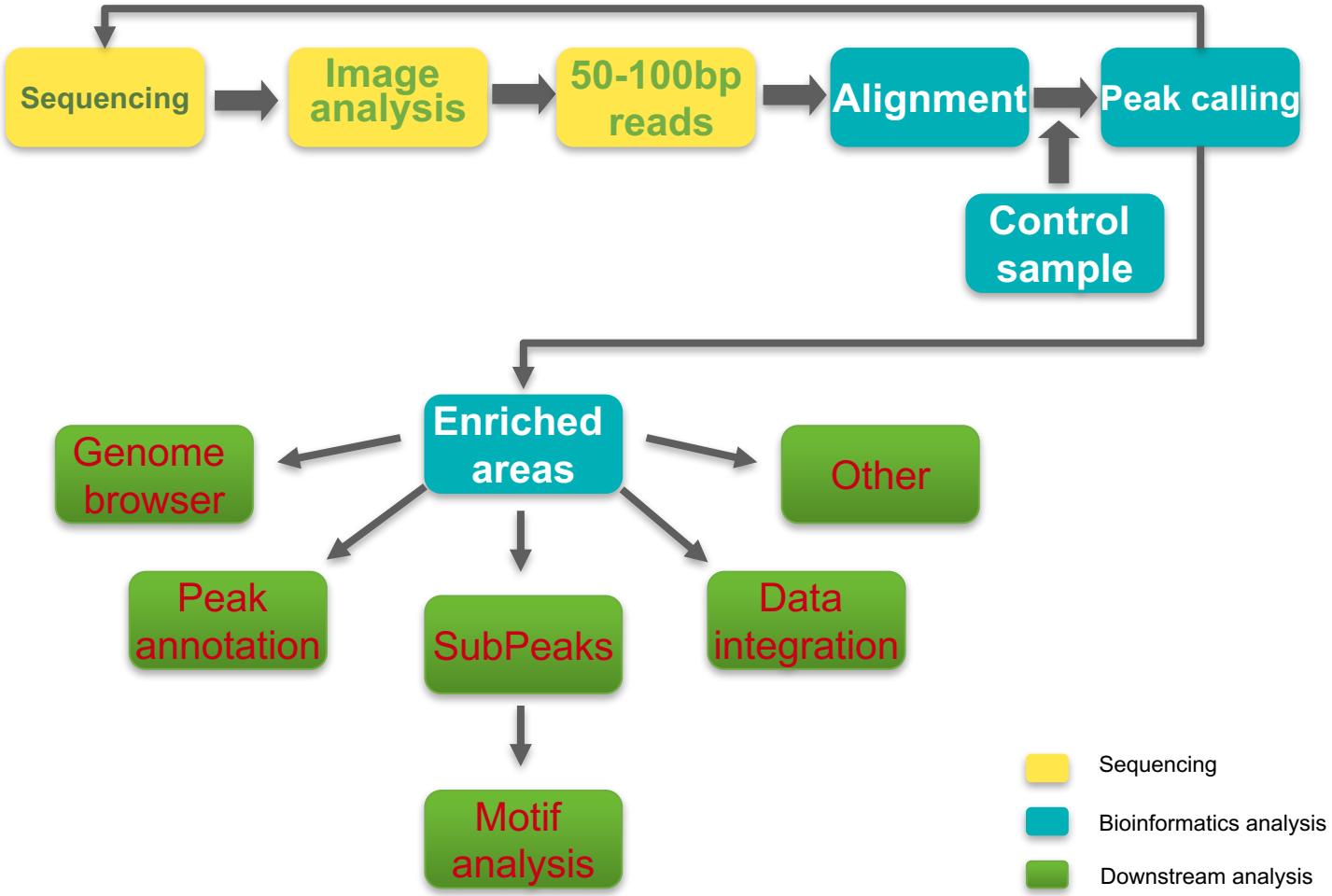






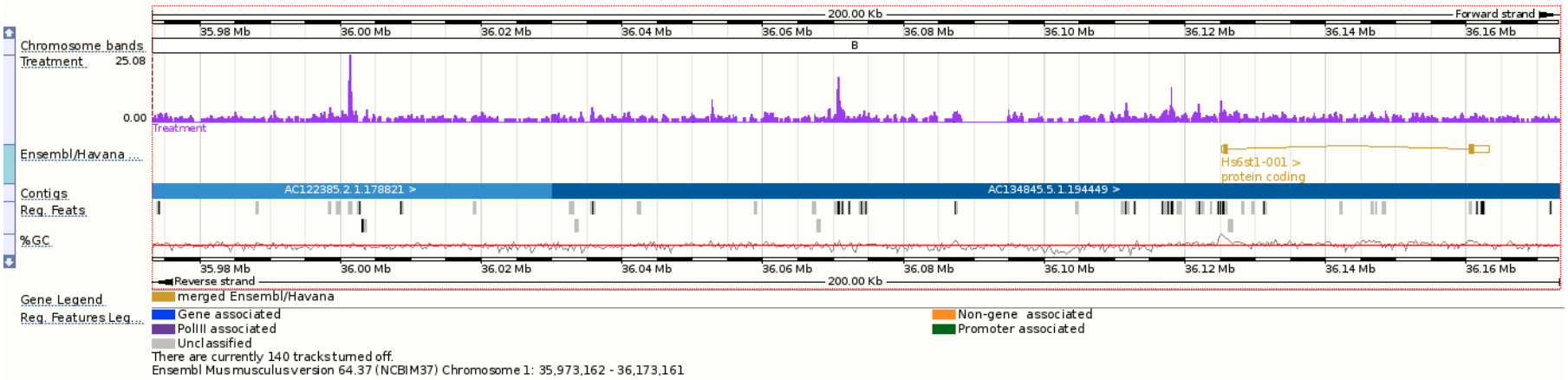
poor consistency

ANALYSIS OVERVIEW



ANALYSIS DOWNSTREAM TO PEAK CALLING

- **Visualisation** - genome browser: Ensembl, UCSC, IGV



ANALYSIS DOWNSTREAM TO PEAK CALLING

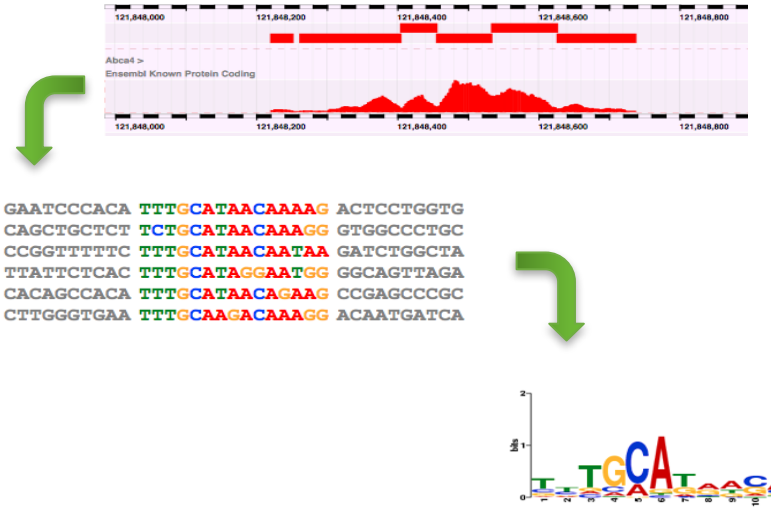
- **Visualization** - genome browser: Ensembl, UCSC, IGV
- **Peak Annotation** - finding interesting features surrounding peak regions:
  - PeakAnalyzer
  - ChIPpeakAnno (R package)
  - GREAT
  - bedtools
  - PAVIS

ANALYSIS DOWNSTREAM TO PEAK CALLING

- **Visualization** - genome browser: Ensembl, UCSC, IGV
- **Peak Annotation** - finding interesting features surrounding peak regions
- Correlation with **expression data**

ANALYSIS DOWNSTREAM TO PEAK CALLING

- **Visualization** - genome browser: Ensembl, UCSC, IGV
- **Peak Annotation** - finding interesting features surrounding peak regions:
- Correlation with expression data
- Discovery of **binding sequence motifs**



ANALYSIS DOWNSTREAM TO PEAK CALLING

- **Visualization** - genome browser: Ensembl, UCSC, IGV
- **Peak Annotation** - finding interesting features surrounding peak regions:
- Correlation with expression data
- **Discovery of binding sequence motifs**
- **Gene Ontology analysis** on genes that bind the same factor or have the same modification

ANALYSIS DOWNSTREAM TO PEAK CALLING

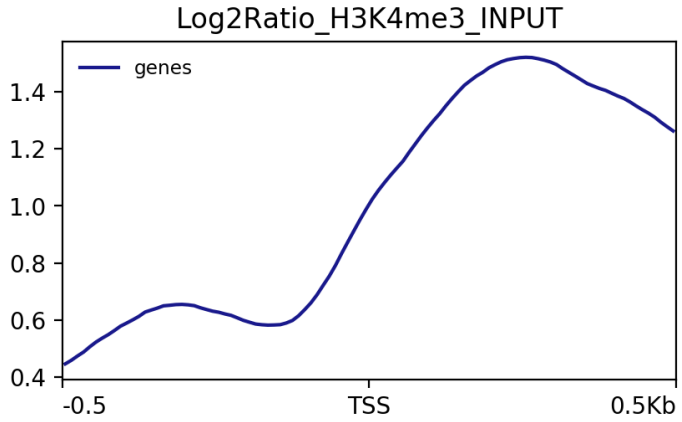
- **Visualization** - genome browser: Ensembl, UCSC, IGV
- **Peak Annotation** - finding interesting features surrounding peak regions:
- Correlation with expression data
- **Discovery of binding sequence motifs**
- Gene Ontology analysis on genes that bind the same factor or have the same modification
- Correlation with SNP data to find **allele-specific binding**

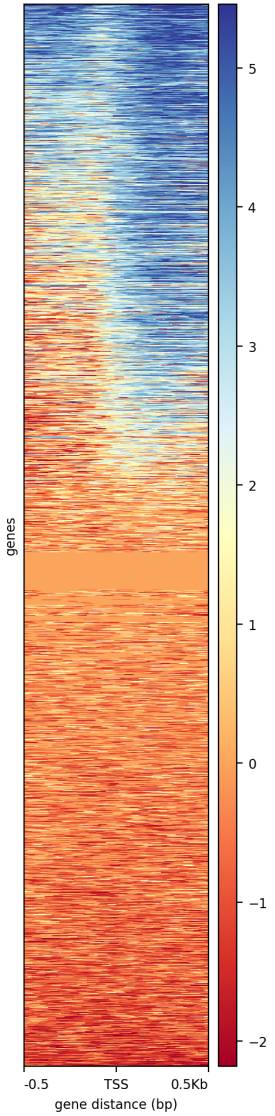
NON-PEAK BASED ANALYSIS

- Can be more appropriate for non-specific binding sites
- Does not involve calling significant peaks, and discarding the rest of the signal as noise
- Instead, the whole signal is used for analysis
- Global analysis, e.g. looking for enrichment at TSS
- Tools include CEAS and EpiChip and deepTools

NON-PEAK BASED ANALYSIS

- Example: Profile and Heatmap for H3K4me3 around the TSS





"It's an absolute myth that you can send an algorithm over raw data and have insights pop up"

-Jeffrey Heer (University of Washington)