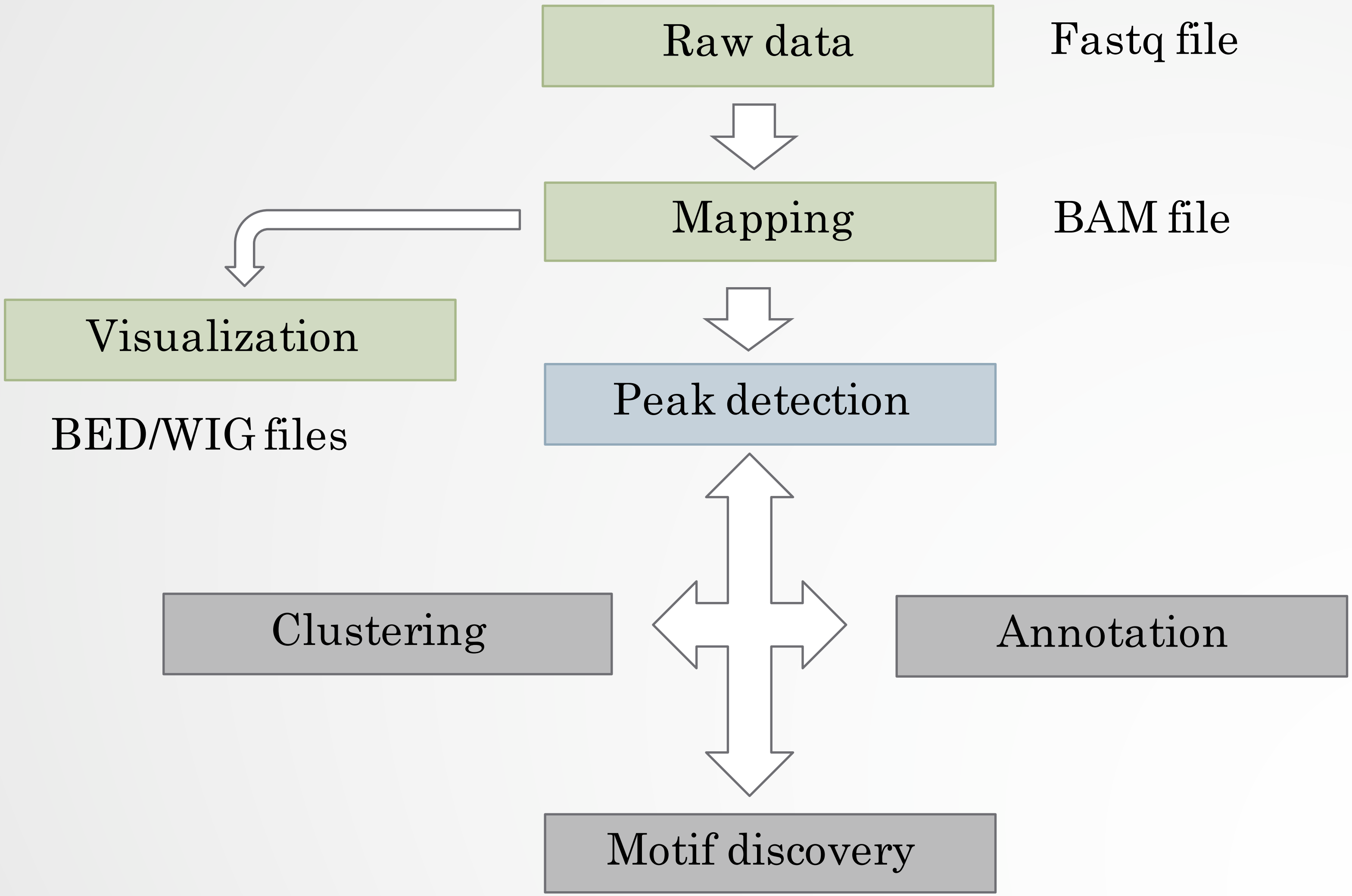
ChIP-seq: Peak Calling

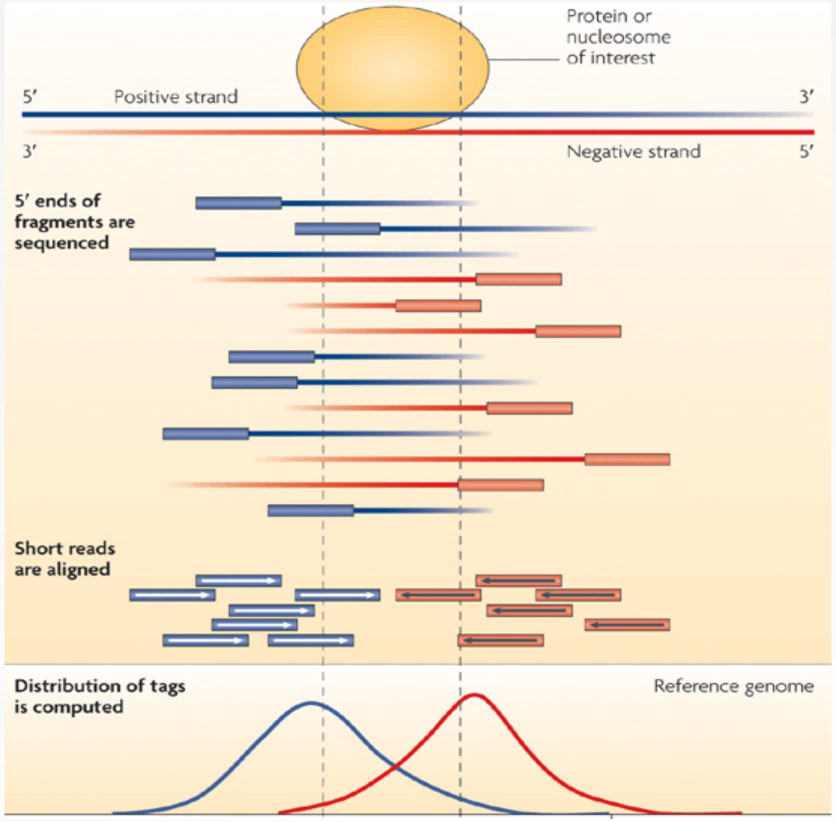
Stéphanie Le Gras
(slegras@igbmc.fr)

Guidelines


Raw data
Fastq file
Mapping
BAM file
Visualization
BED/WIG files
Peak detection
Clustering
Annotation
Motif discovery

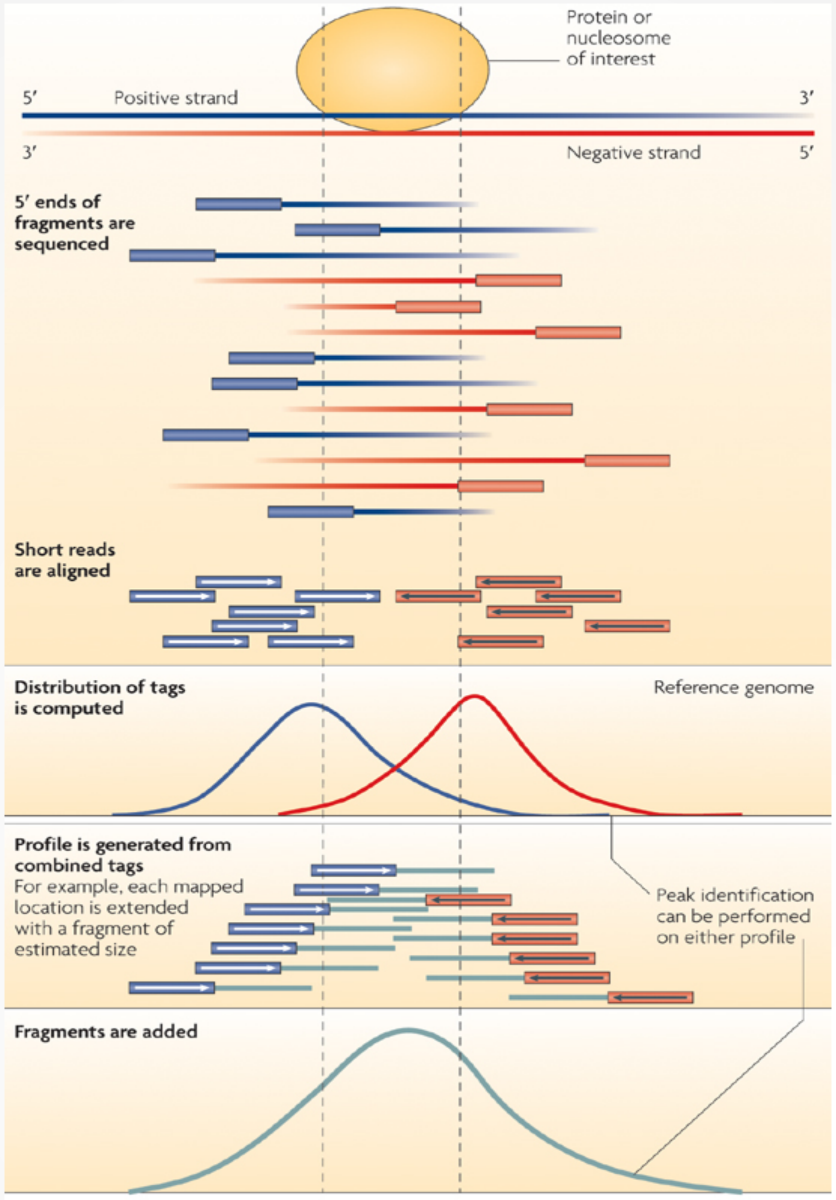
From reads to peaks

- Chip-seq peaks are a mixture of two signals:
  - + strand reads (Watson)
  - - strand reads (Cricks)
- The sequence tag density accumulates on forward and reverse strands centered around the binding site



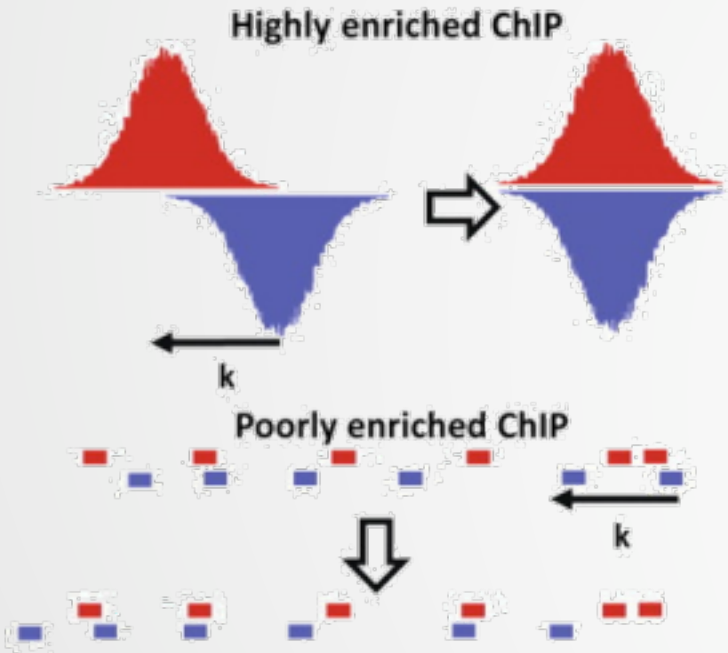
From reads to peaks

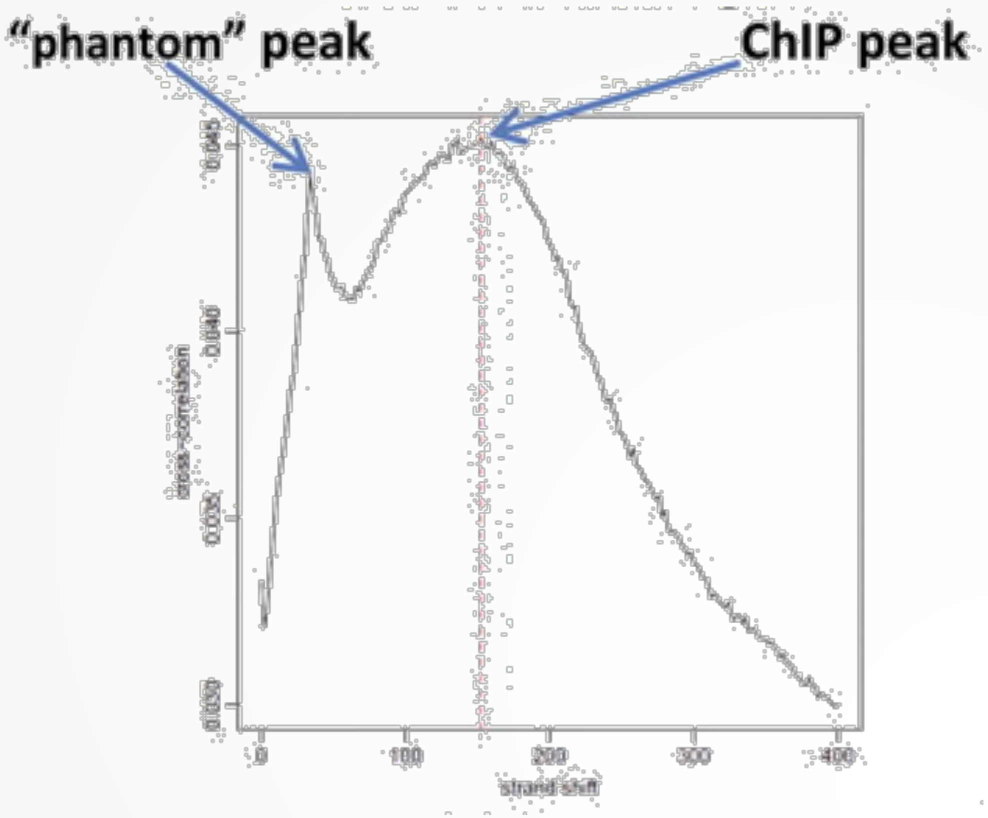
- Get the signal at the right position
  - Read shift
  - Extension
- Estimate the fragment size
- Do paired-end



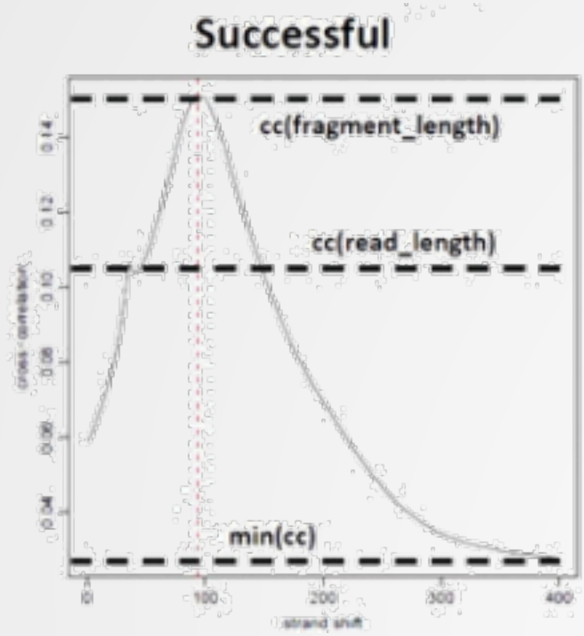
QC: cross correlation analysis

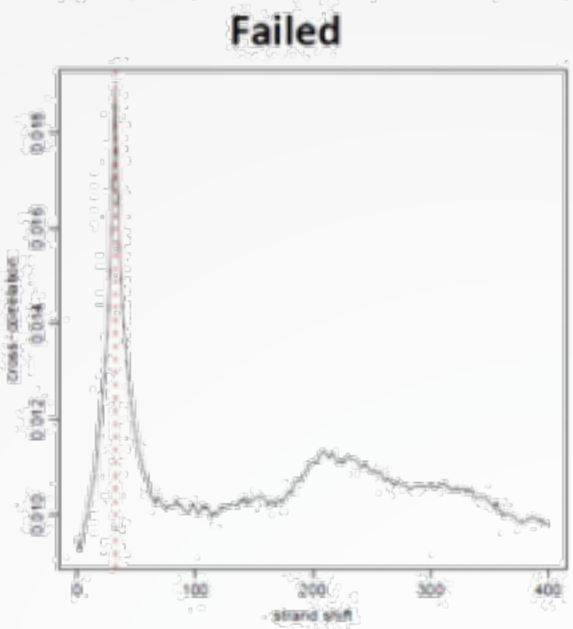
- The cross-correlation metric is computed as the Pearson's linear correlation between the Crick strand and the Watson strand, after shifting Watson by k base pairs.





QC: cross correlation analysis





NSC: normalized strand coefficient

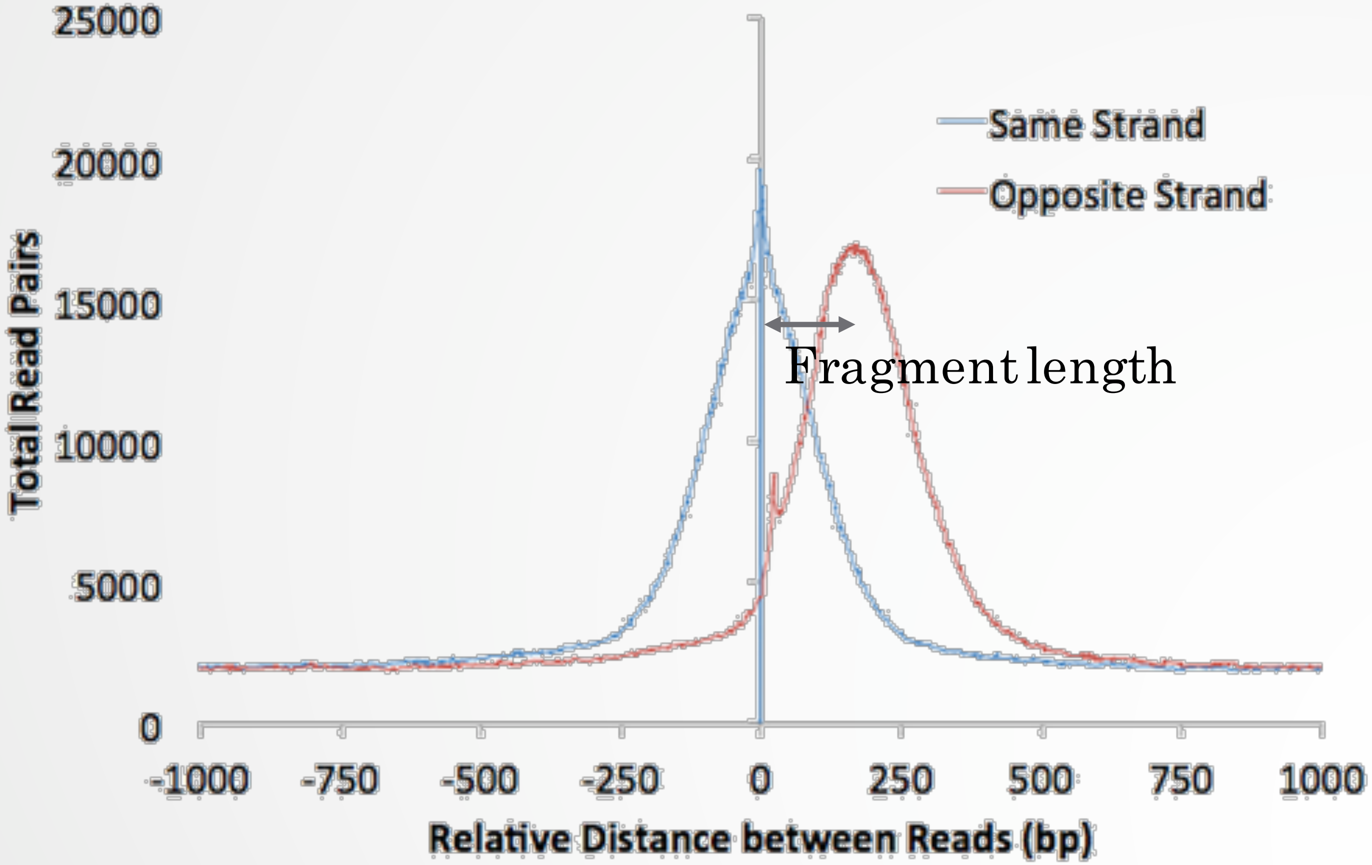
$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

Relative strand correlation (RSC)

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

Estimating the fragment size

- Homer (Heinz et al, 2010): Compute distribution of distances between adjacent reads in the genome



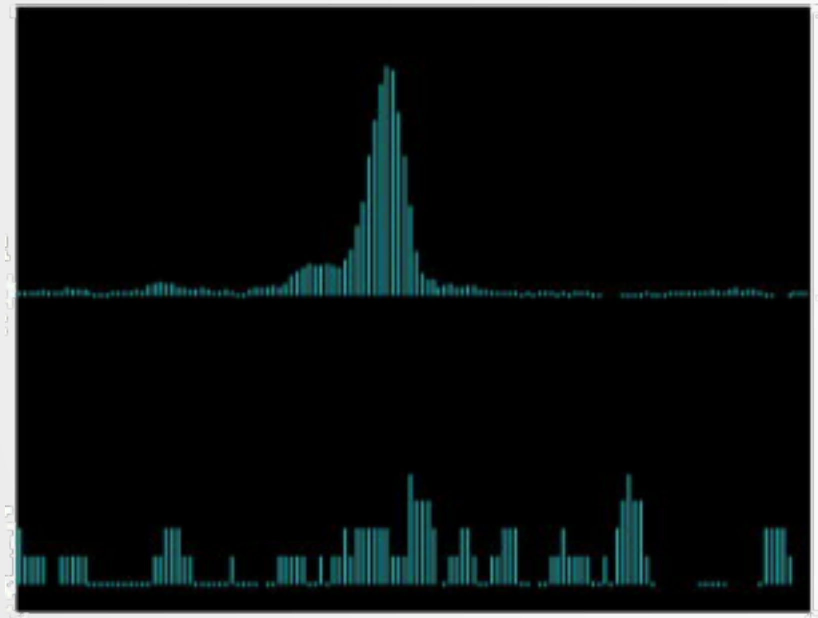
Sequencing depth normalization

- Normalizing by total read numbers: scaling the local read density by a multiplicative ratio of total sequencing depths.
- Normalizing to 1x or 10x coverage: norm.binCount / 1x coverage = real bin count / real coverage
- RPKM: number of reads per bin / (number of mapped reads (in millions) * bin length (kbp))

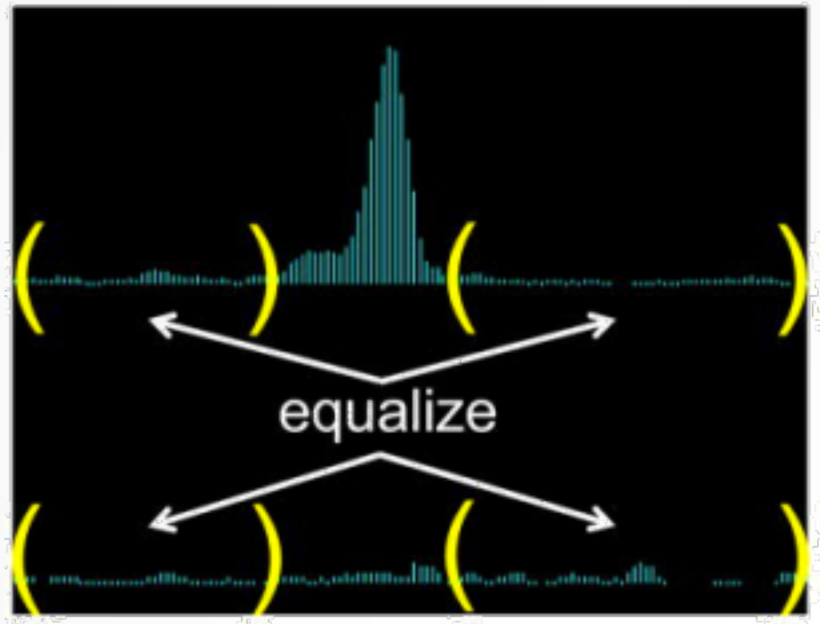
Input normalization

- Naïve subtraction: treatment – input. To be done, first scale the two libraries by the total number of reads (library size)

- Normalizing by total read count



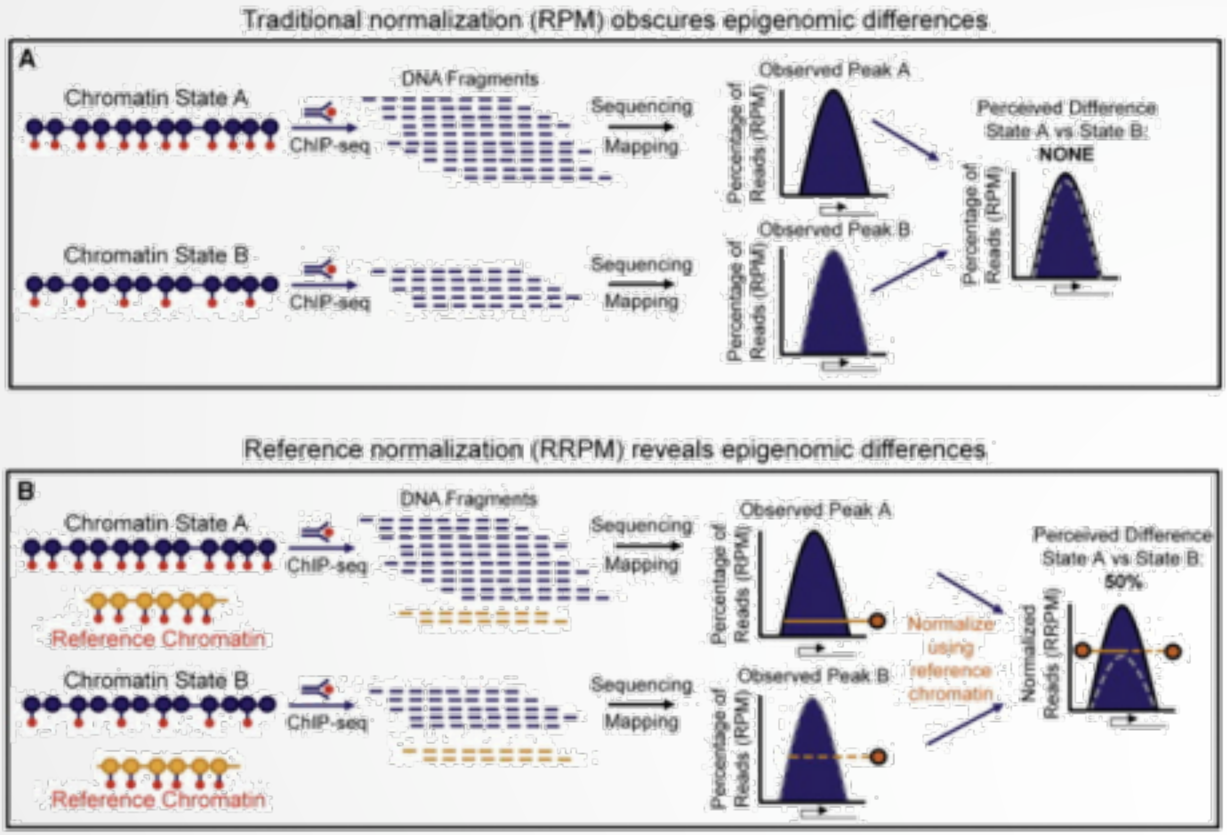
- Normalizing the background



Diaz et al, 2012

Spike-in

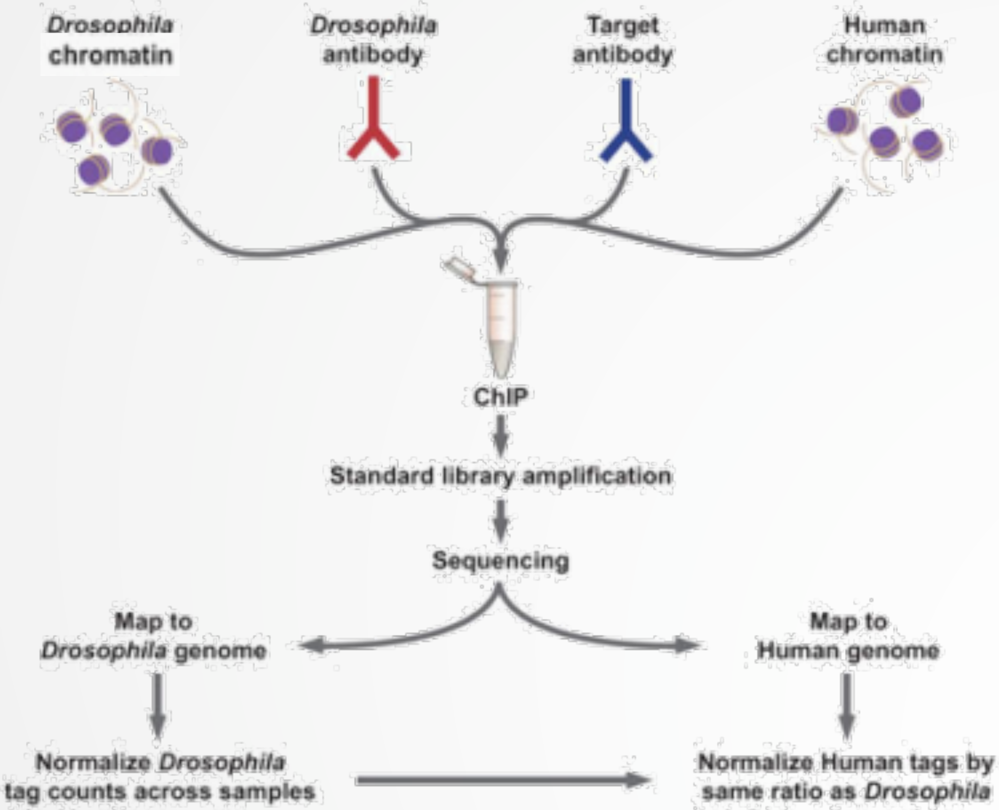
- Current normalization methods fail to detect global changes as they make the assumption that globally nothing change but a small portion of the genome
- Insert external chromatin used as reference chromatin



Orlando et al, 2014

Spike-in

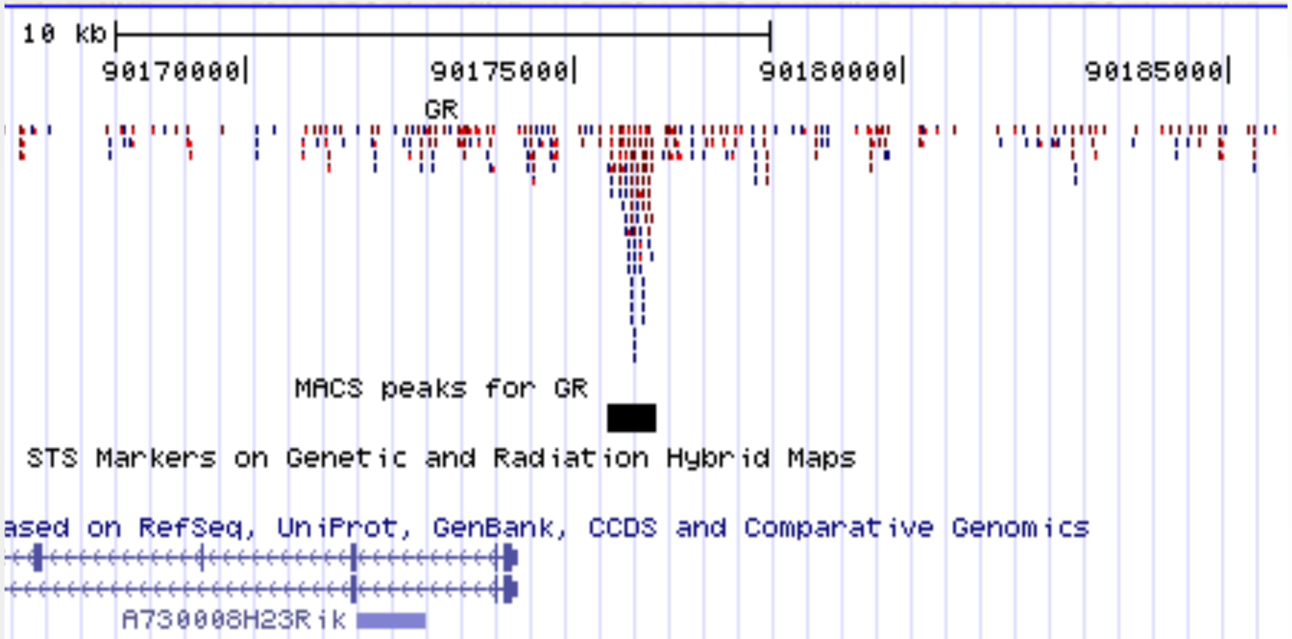
- Spike-in normalization can be applied to ChIP-Seq data to reduce the effects of technical variation and sample processing bias



http://www.activemotif.com/catalog/1091/chip-normalization

Peak detection

- Discover interaction sites from aligned reads
- Idea: loci with a lot of reads/fragments = signal site



Peak detection

- Loci with lots of reads could also be due to
  - Sequencing biases
  - Chromatin biases (e.g CNVs)
  - PCR biases/artefacts
  - Biases/artefacts of unknown origin
  - So need to separate signal from noise
- Need to use a control to correct for the biases (Expect that the biaises are similar in input and in IP)

Peak finders

Basic components of peak callers:

- A signal profile definition along each chromosome
- A background model
- Peak call criteria
- Post-call filtering of artifactual peaks
- Significance ranking of called peaks

Pepke et al, 2009

Peak finders

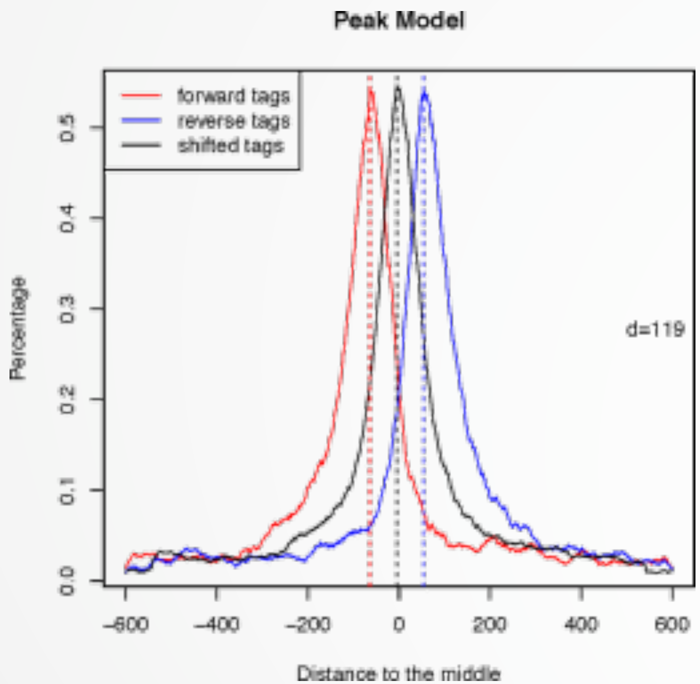
Pepke et al, 2009

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff High quality peak estimate, per-region estimate, or input	High quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally *P* values	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	NA	Number of reads under peak	1: Monte Carlo simulation 2: NA	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation (KDE)	*s* s.d. above KDE for 1: random background, 2: control	Input or estimated	KDE for local background	Peak height	1: None 2: None	Threshold s.d. value, KDE bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrichment	2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross-correlation	KDE for enrichment and empirical FDR estimation	*q* value	1: NA 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$ as a function of profile threshold	KDE bandwidth, peak height, subpeak valley depth, ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	*P* value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and *P* values	*q* value	1: None 2: From Poisson *P* values	Window length, gap size, FDR (with control) or *E*-value (no control)	No / Yes	15
SISSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ +$ N_- threshold in	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11

MACS [Zhang et al, 2008]

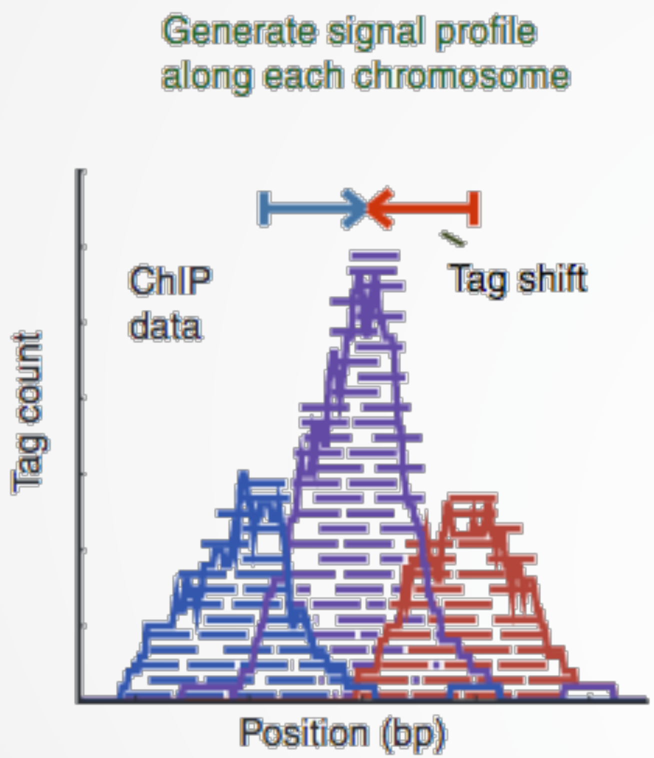
1. Modeling the shift size of ChIP-Seq tags

- slides *2bandwidth* windows across the genome to find regions with tags more than *mfold* enriched relative to a random tag genome distribution
- randomly samples 1,000 of these highly enriched peaks
- separates their Watson and Crick tags, and aligns them by the midpoint between their Watson and Crick tag centers
- define *d* as the distance in bp between the summit of the two distributions



MACS [Zhang et al, 2008]

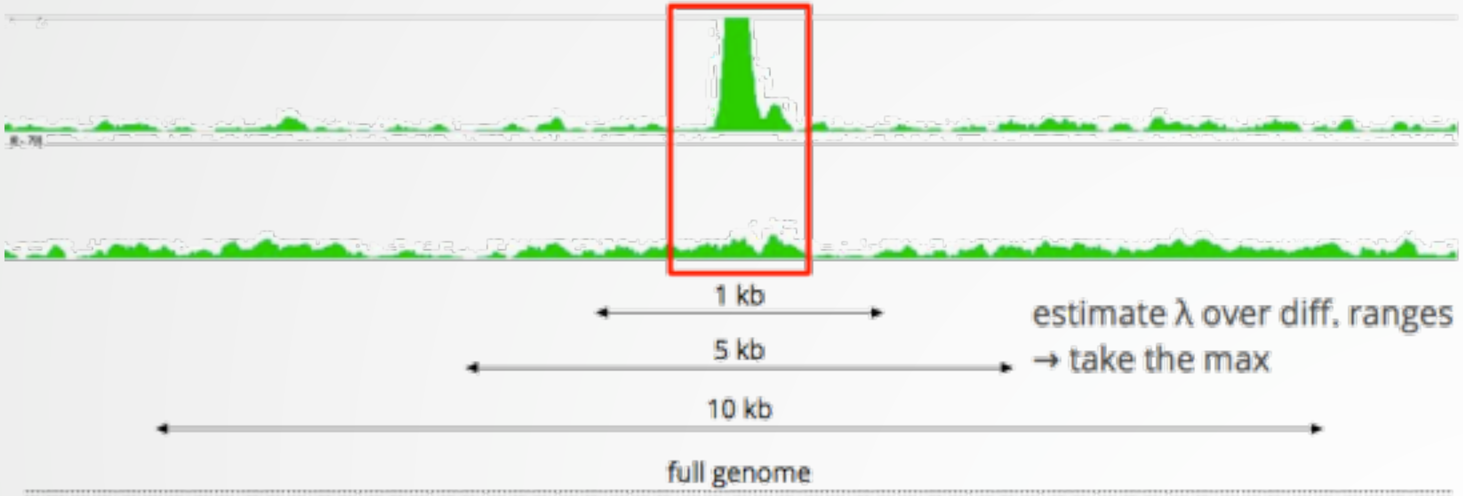
- **2. Peak detection**
  - Normalization: linearly scales the total control read count to be the same as the total ChIP read count
  - Duplicate read removal
  - Tags are shifted by *d*/2



Pepke et al, 2009

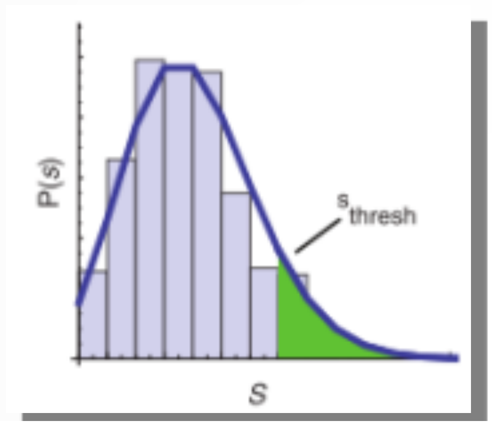
MACS [Zhang et al, 2008]

- Slides 2d windows across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution *p*-value based on λ_{BG}, default 10^{-5})
- Estimate parameter λ_{local} of Poisson distribution



Source:
C. Herrmann

- Keep peaks significant under λ_{BG} and λ_{local} and with p-value < threshold

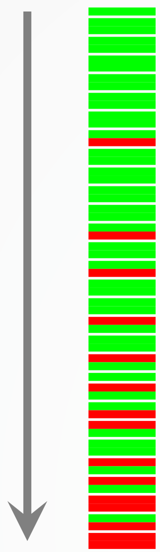


MACS [Zhang et al, 2008]

3. **Multiple testing correction (FDR)**

- Swap treatment and input and call negative peaks
- Take all the peaks (neg + pos) and sort them by increasing p-values

$$\mathrm{FDR}(p) = \frac{\#\ \textbf{Negative}\ \text{peaks with p-value} < p}{\#\ \textbf{Selected peaks}}$$



FDR = 2/27 = 0.074

Exercise: peak calling

We now want to call MITF peaks.

- 1. Use **Macs2 callpeak** to perform the peak calling on the data. Use default parameters except for
  - ChIP-Seq Treatment File: mitf.bam
  - ChIP-Seq Control File: ctrl.bam
  - Effective genome size: Human
  - Outputs: Peaks as tabular file, summits, Summary page (html), Plot in PDF
- 2. Look at the resulting datasets. How many peaks are found?
- 3. What is the fragment size estimated by Macs2? What do you think of the value?
- 4. Rerun **Macs2** using the same parameters as before but changing the shift size:
  - Build Model: Do not build the shifting model (--nomodel)
  - The arbitrary extension size in bp: 100
- 5. How many peaks are now found?

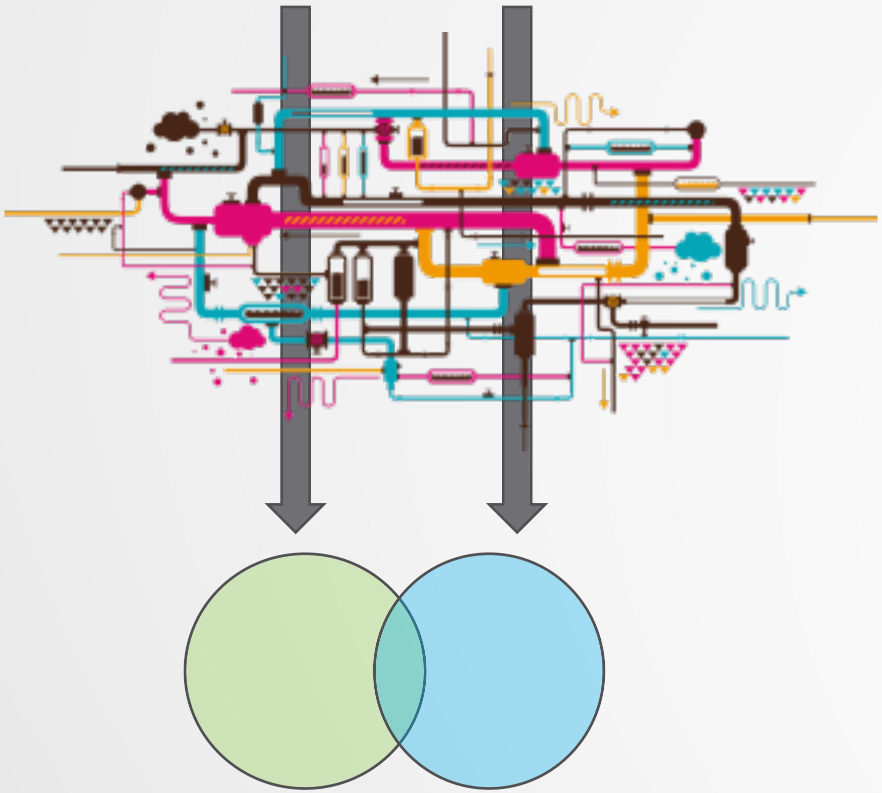
How to deal with replicates

Analyze samples separately and takes union or intersection of resulting peaks

Merge samples prior to the peak calling (e.g recommended by MACS)

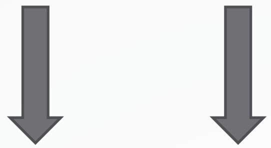
Sample 1.a

Sample 1.b

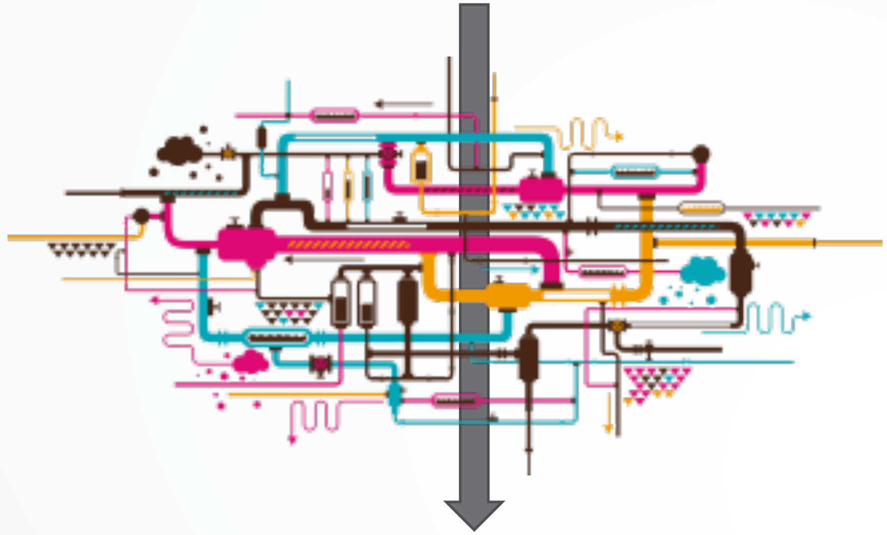


Sample 1.a

Sample 1.b



Sample 1

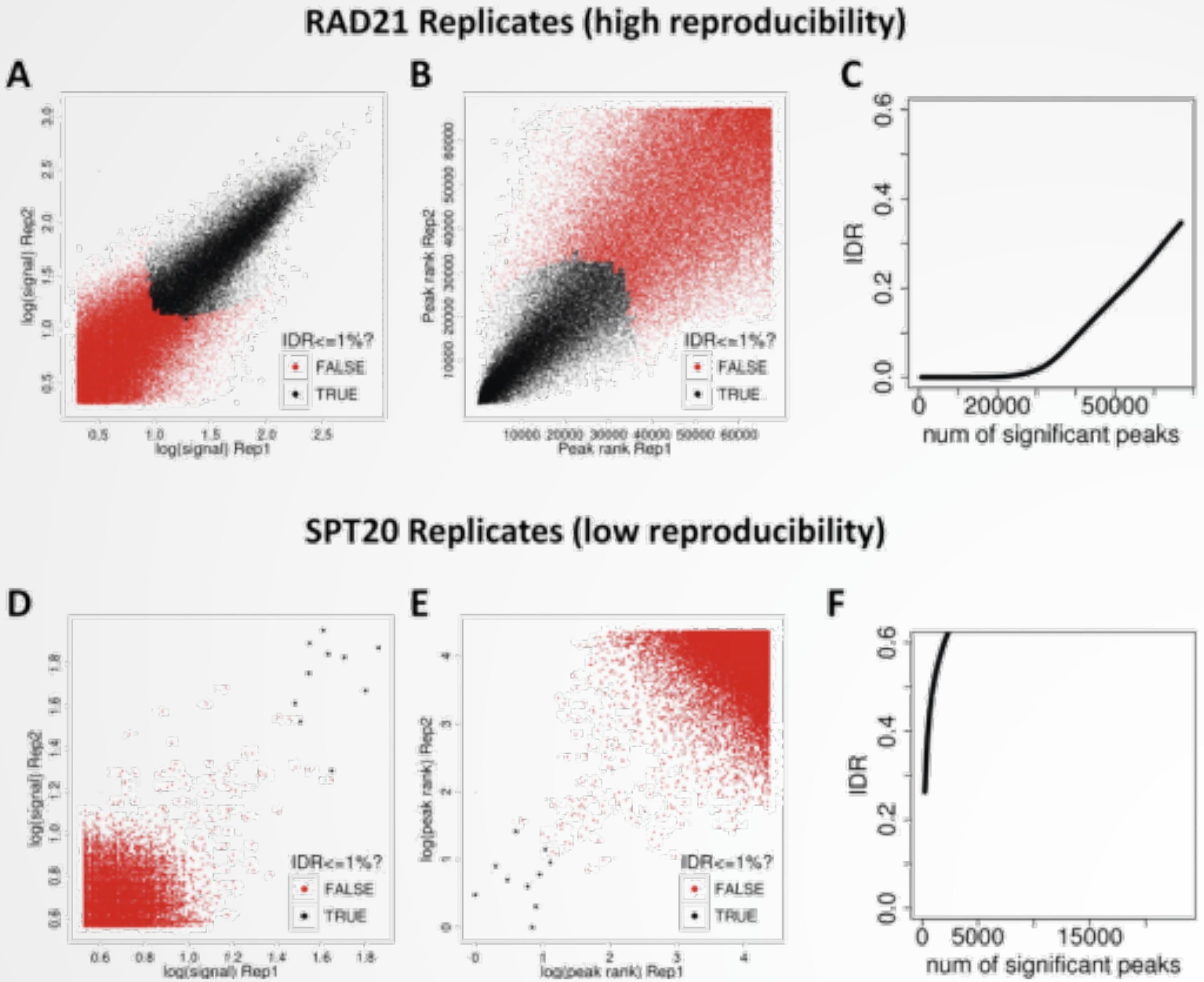


IDR

- Measures consistency between replicates
- Uses reproducibility in score rankings between peaks in each replicate to determine an optimal cutoff for significance.
- Idea:
  - The most significant peaks are expected to have high consistency between replicates
  - The peaks with low significance are expected to have low consistency

https://sites.google.com/site/anshulkundaje/projects/idr

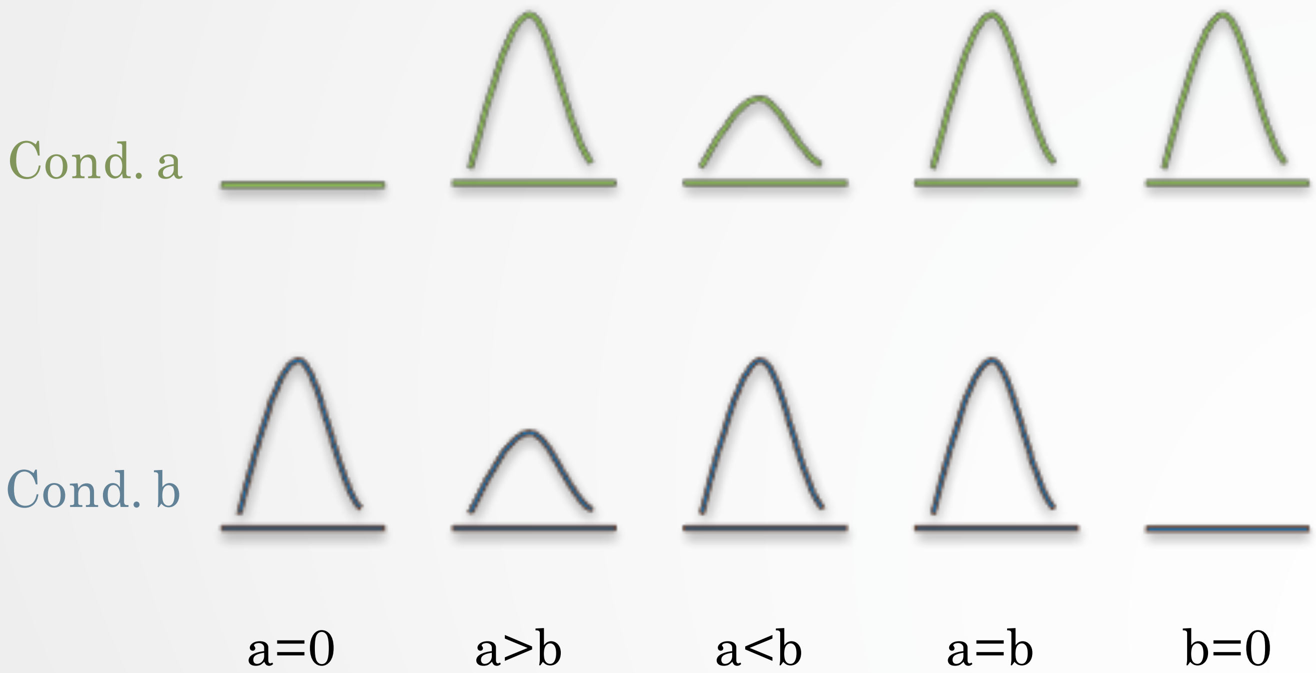
IDR



(!) IDR doesn't work on broad source data!

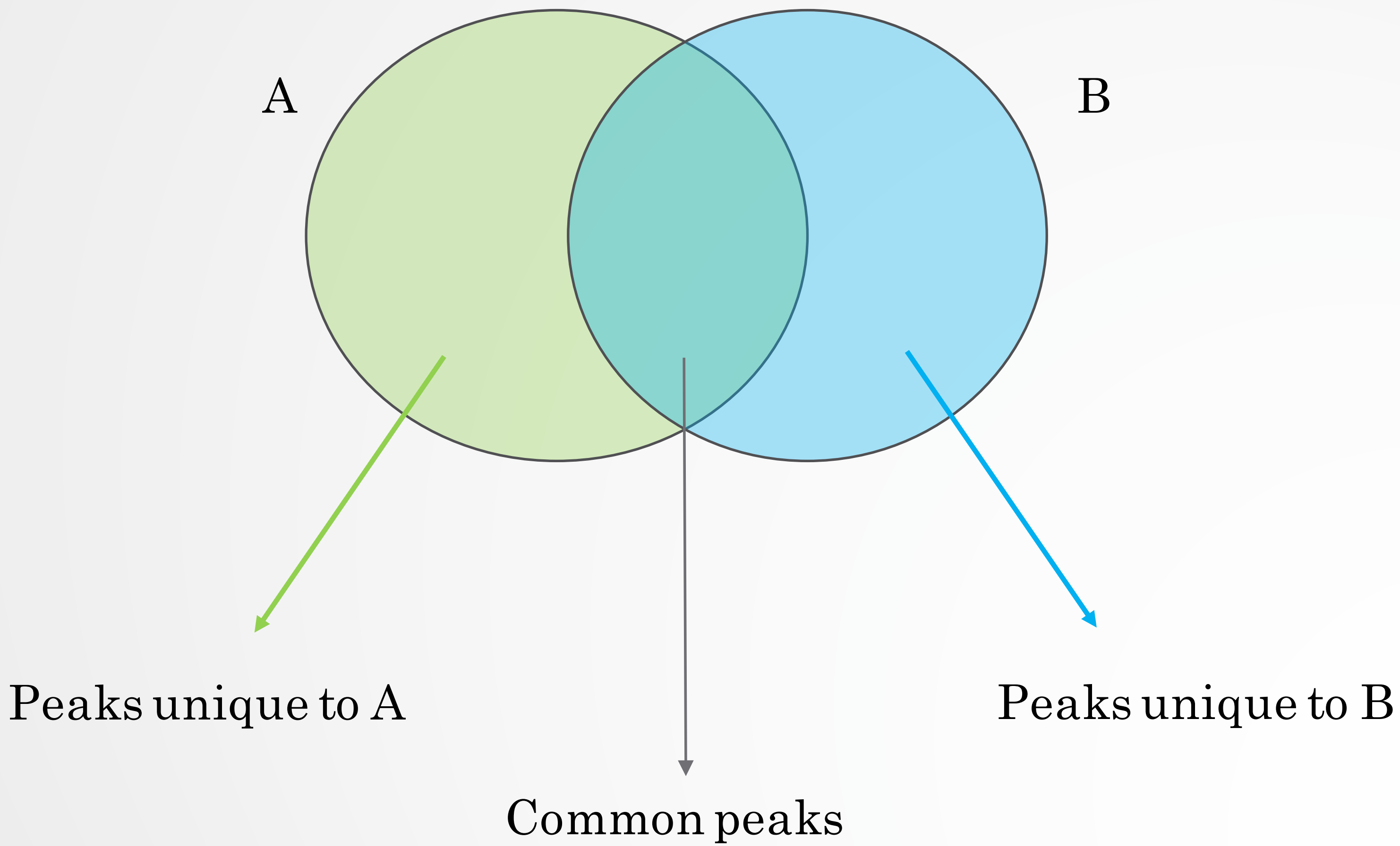
Differential binding analysis

- Find differential binding events by comparing different conditions
  - qualitative analysis: binding vs no binding
  - quantitative analysis: weak binding vs strong binding



Differential binding analysis

Qualitative approach



Differential binding analysis

Quantitative approach

- Do the peak calling on all data
- Take union of all peaks
- Do quantitative analysis of differential binding events based on read counts

- Statistical models
  - No replicates: assume simple Poisson model
  - With replicates: perform differential test using DE tools from RNA-seq (EdgeR, DESeq,…) based on read counts