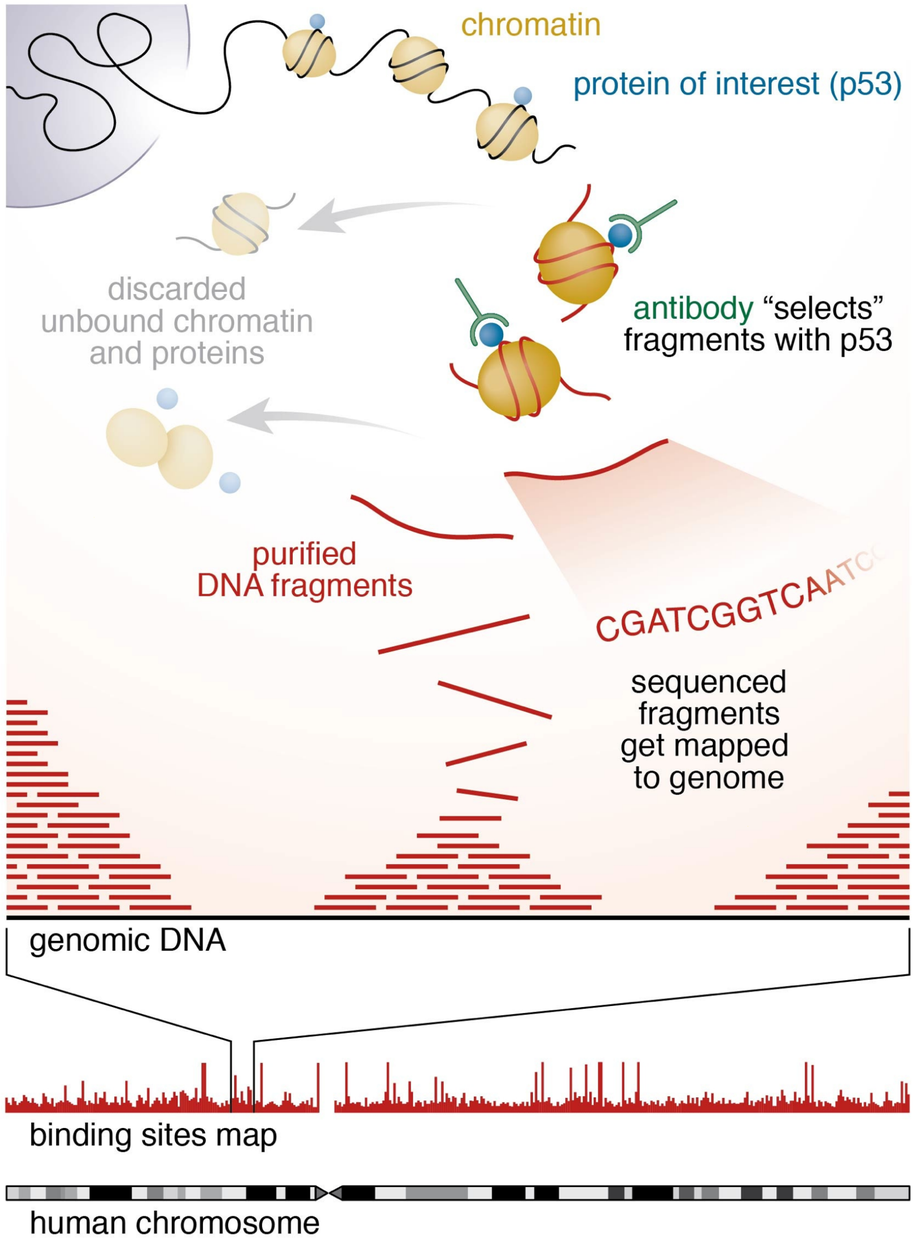
ChIP-seq statistics and tools

HSPH Bioinformatics Core



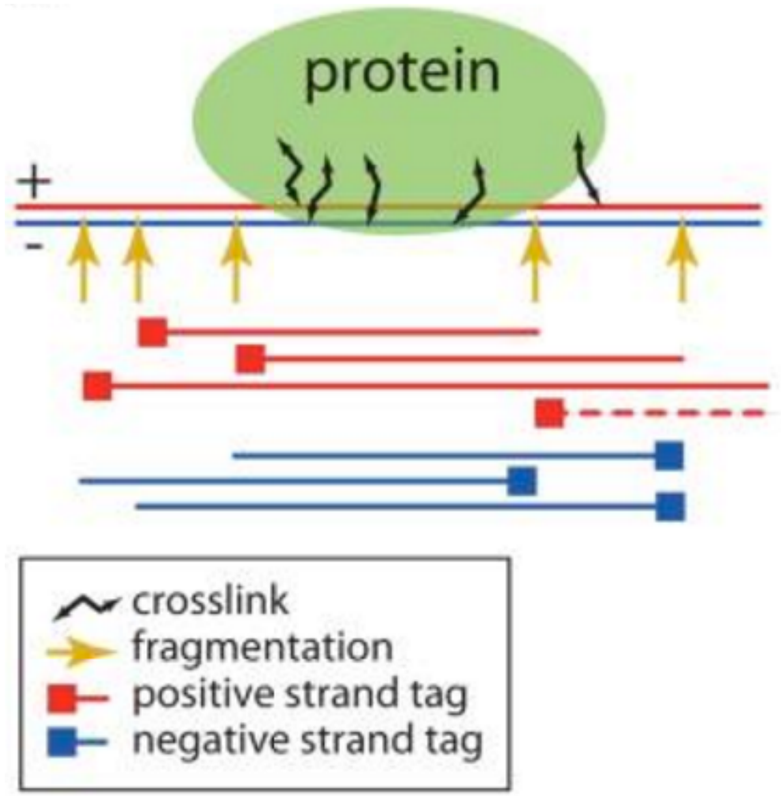
Key aspects of peak calling

- Treating the reads
- Evaluating quality metrics for ChIP-seq data
- Modeling noise levels
- Detecting enriched regions (‘peaks’)
- Dealing with replicates

Treating the reads

Treating the reads

- Positive and negative strand read peaks *do not represent the true location of the binding site*
- Reads can be shifted or elongated to increase resolutions
- The shift distance can be estimated from the data or given as an input parameter



Evaluating quality metrics

Strand cross correlation profile

Strand cross correlation profile

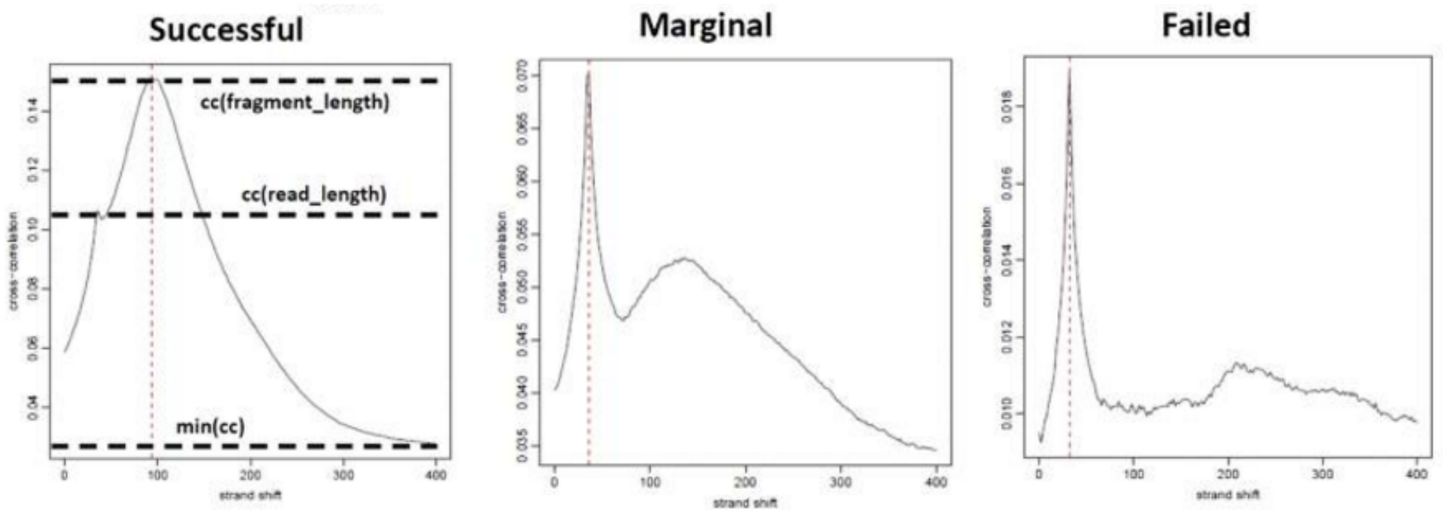
- Compute the number of read starts at each position on the + strand and separately on the - strand for each chromosome

Strand cross correlation profile

- Compute the number of read starts at each position on the + strand and separately on the - strand for each chromosome
- Shift these vectors wrt each other and compute the correlation for each shift.

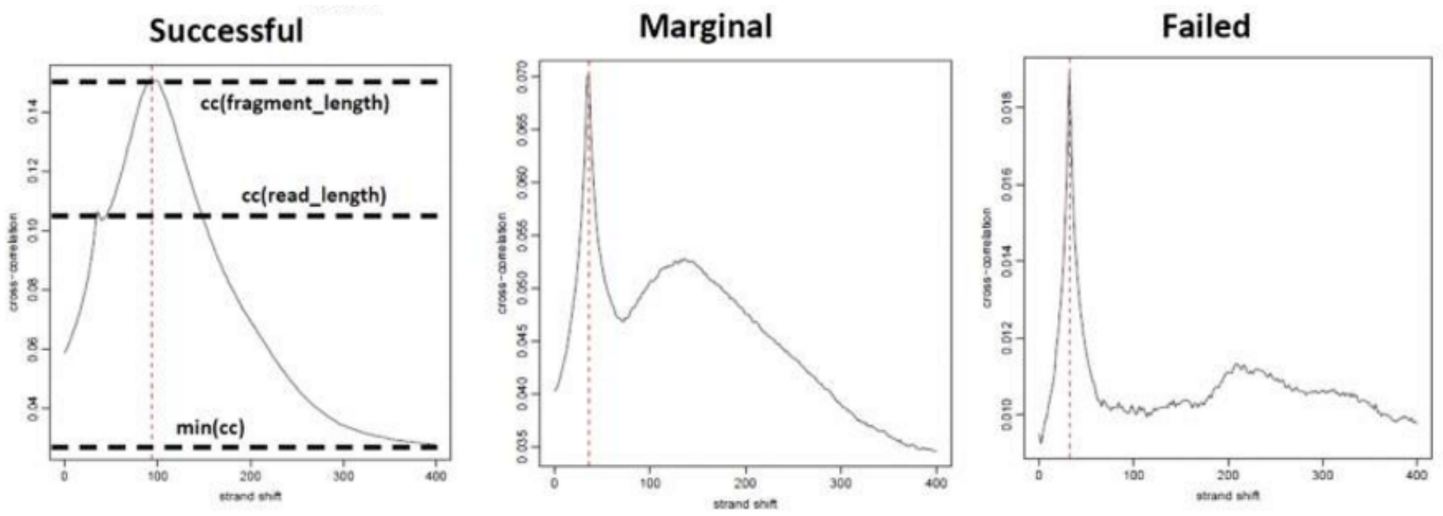
Strand cross correlation profile

- Compute the number of read starts at each position on the + strand and separately on the - strand for each chromosome
- Shift these vectors wrt each other and compute the correlation for each shift.



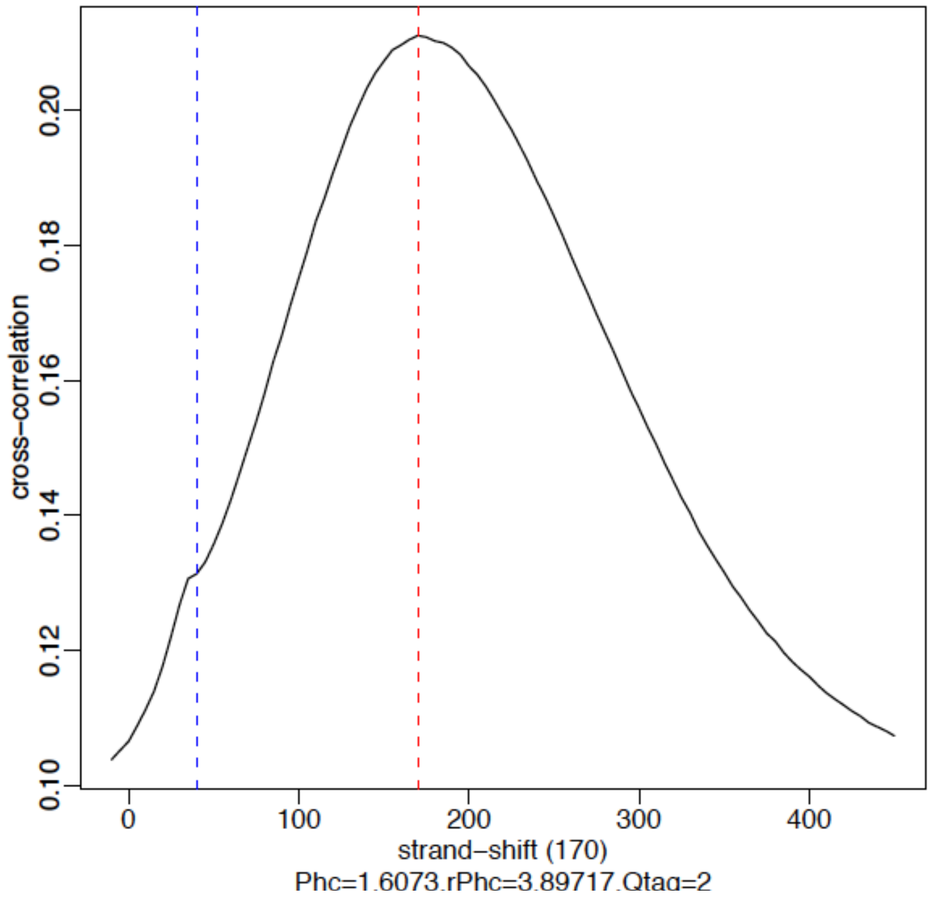
Strand cross correlation profile

- Compute the number of read starts at each position on the + strand and separately on the - strand for each chromosome
- Shift these vectors wrt each other and compute the correlation for each shift.
- Due to the 'shift' phenomenon of reads on the + and - strand around true binding sites, one would get a peak in the cross-correlation profile at the predominant fragment length



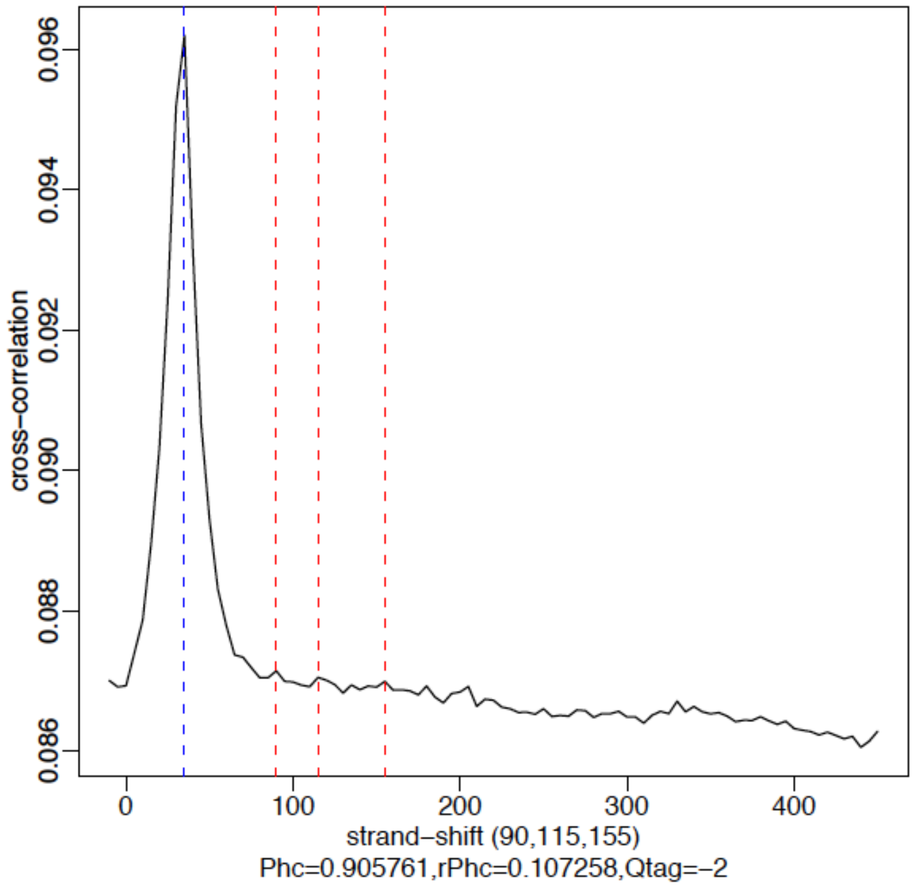
Strong signal: CTCF in human cells

- Great antibody and 45-60K peaks typically
- Red vertical line shows the dominant peak at the true peak shift
- Small bump at the blue vertical line is at read-length



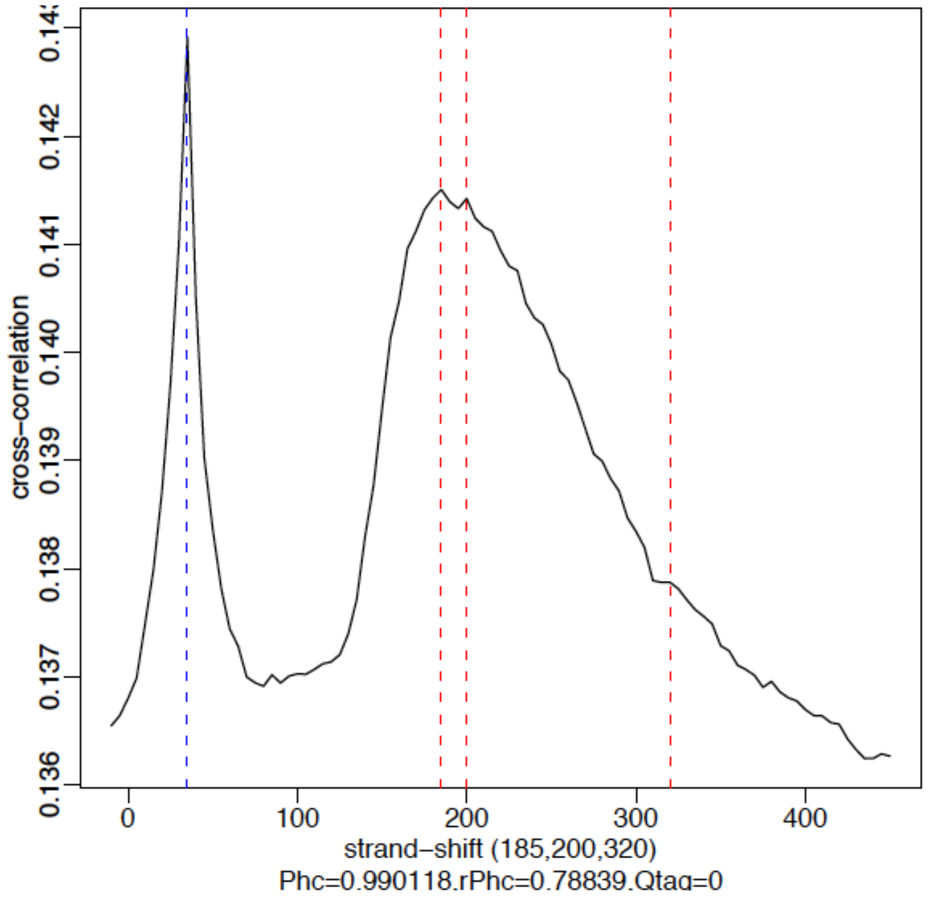
Control dataset (input DNA)

- Note the strongest peak is the blue line (read length) and there is basically almost no other significant peak in the profile
- The absence of a peak is expected since there should be no significant clustering of fragments around specific target sites (except potentially weak biases in open chromatin regions depending on the protocol used)
- The read-length peak occurs due to unique mappability properties of the mapped reads



Weak signal: POL2

- This particular antibody is not very efficient and these are broad scattered peaks
- Has few peaks (~3000 detectable)
- Two peaks in the cross-correlation profile: one at the true peak shift (~185-200 bp) and the other at read length. For such weaker datasets, the read-length peak starts to dominate.



Metrics based on cross-correlation

- **Normalized strand cross-correlation coefficient (NSC):** is the ratio between the fragment-length cross-correlation peak and the background cross-correlation (minimum cross-correlation value over all possible strand shifts). Higher values indicate more enrichment; minimum value is 1.

- **Relative strand cross-correlation coefficient (RSC):** the ratio between the fragment-length peak and the read-length peak, each minus the background value. Minimum value is zero.

Metrics based on cross-correlation

- **Normalized strand cross-correlation coefficient (NSC):** is the ratio between the fragment-length cross-correlation peak and the background cross-correlation (minimum cross-correlation value over all possible strand shifts). Higher values indicate more enrichment; minimum value is 1.

- **Relative strand cross-correlation coefficient (RSC):** the ratio between the fragment-length peak and the read-length peak, each minus the background value. Minimum value is zero.

The current ENCODE practice is for experiments with NSC values below 1.05 and RSC values below 0.8, it is recommended that an additional replicate be attempted or the experiment explained in the data submission as adequate based on additional considerations.

Global ChIP enrichment with FRiP

Global ChIP enrichment with FRiP

- Calculate the fraction of all mapped reads that fall into peak regions; typically, a minority of total reads

Global ChIP enrichment with FRiP

- Calculate the fraction of all mapped reads that fall into peak regions; typically, a minority of total reads
- FRiP values correlate positively and linearly with the number of called regions,

Global ChIP enrichment with FRiP

- Calculate the fraction of all mapped reads that fall into peak regions; typically, a minority of total reads
- FRiP values correlate positively and linearly with the number of called regions,
- ENCODE data sets have a FRiP enrichment of 1% or more when peaks are called using MACS with default parameters. This works well when there are thousands to tens of thousands of called occupancy sites in a large (i.e. mammalian) genome

Global ChIP enrichment with FRiP

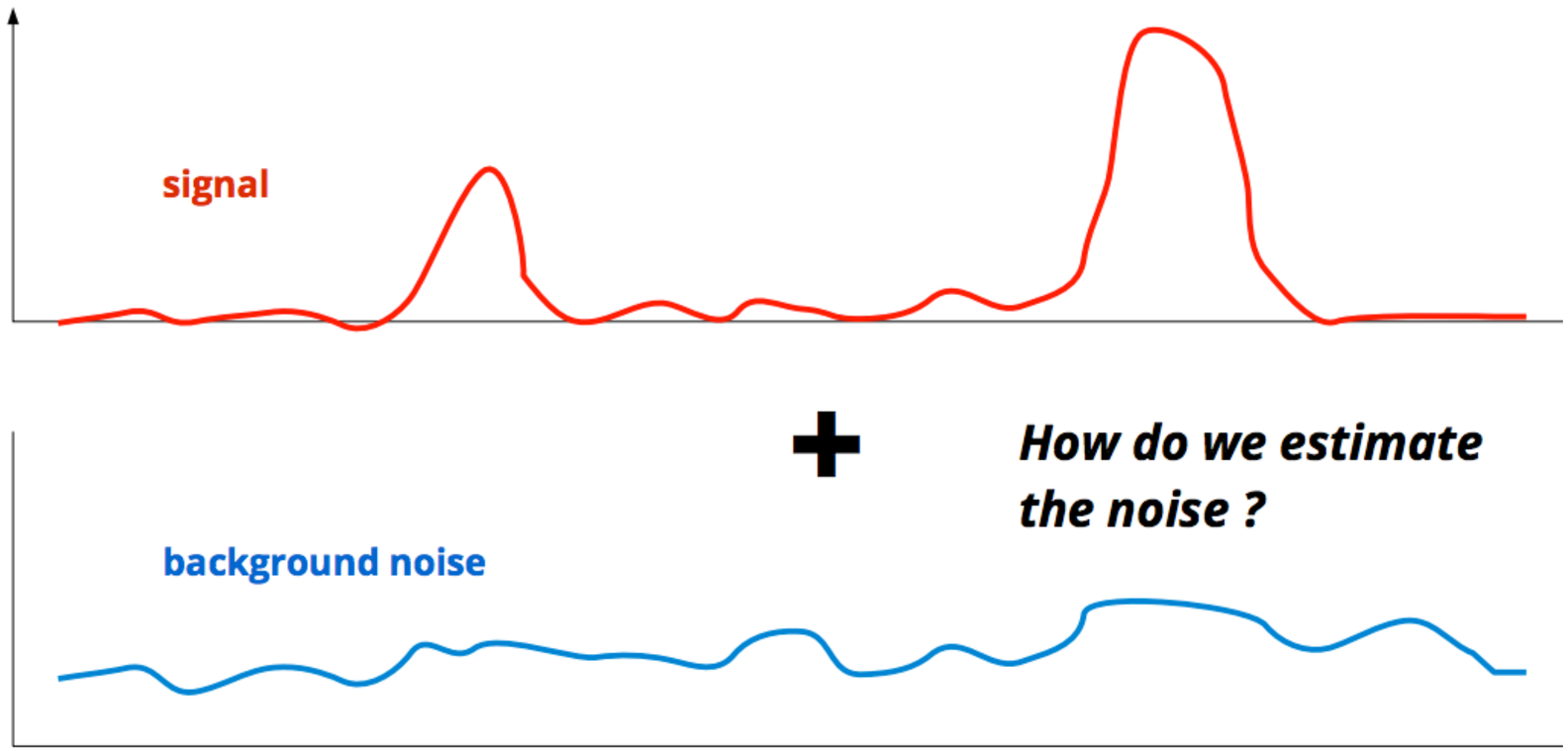
- Calculate the fraction of all mapped reads that fall into peak regions; typically, a minority of total reads
- FRiP values correlate positively and linearly with the number of called regions,
- ENCODE data sets have a FRiP enrichment of 1% or more when peaks are called using MACS with default parameters. This works well when there are thousands to tens of thousands of called occupancy sites in a large (i.e. mammalian) genome

Passing this threshold does not automatically mean that an experiment is successful and a FRiP below the threshold does not automatically mean failure!

FRiP Guidelines

- Useful for comparing results obtained with the same antibody across cell lines or with different antibodies against the same factor.
- FRiP is sensitive to the specifics of peak calling, including the way the algorithm delineates regions of enrichment and the parameters and thresholds used.
- All FRiP values that are compared should be derived from peaks uniformly called by a single algorithm and single parameter set.

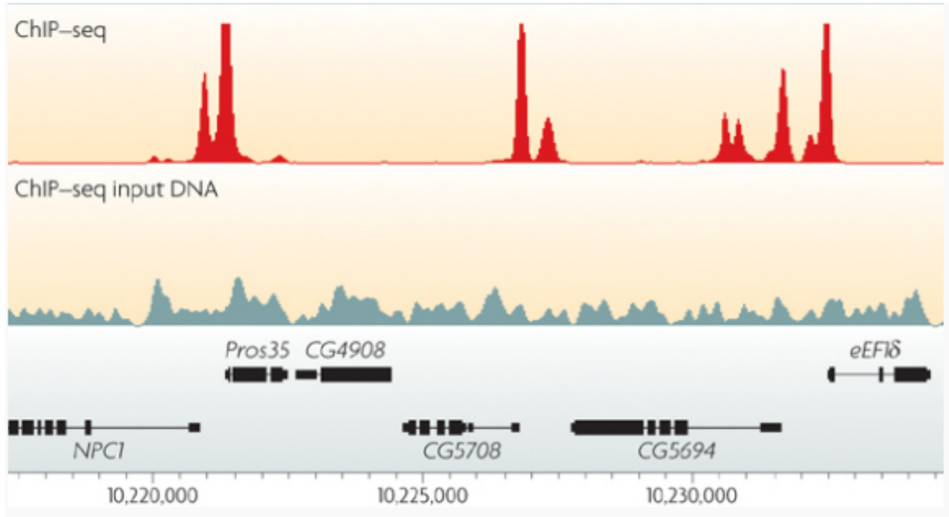
Modeling noise levels



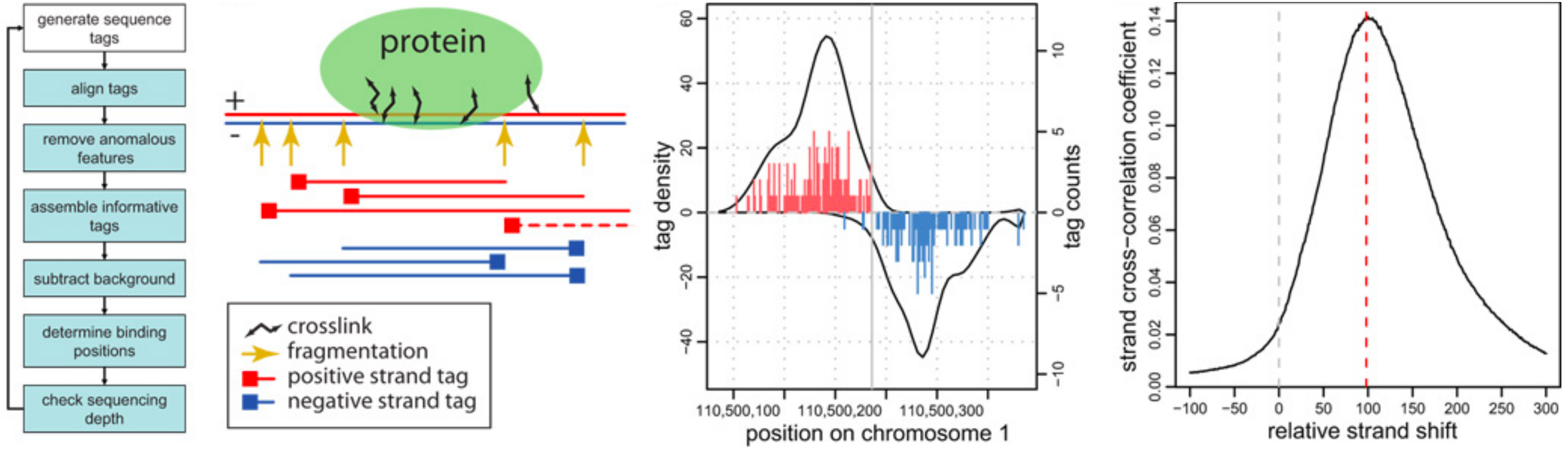
Modeling noise levels

Modeling noise levels

- Noise is not uniform (chromatin conformation, local biases, mappability)
- Input data is mandatory for a reliable estimation of noise (even though some tools don't require it)

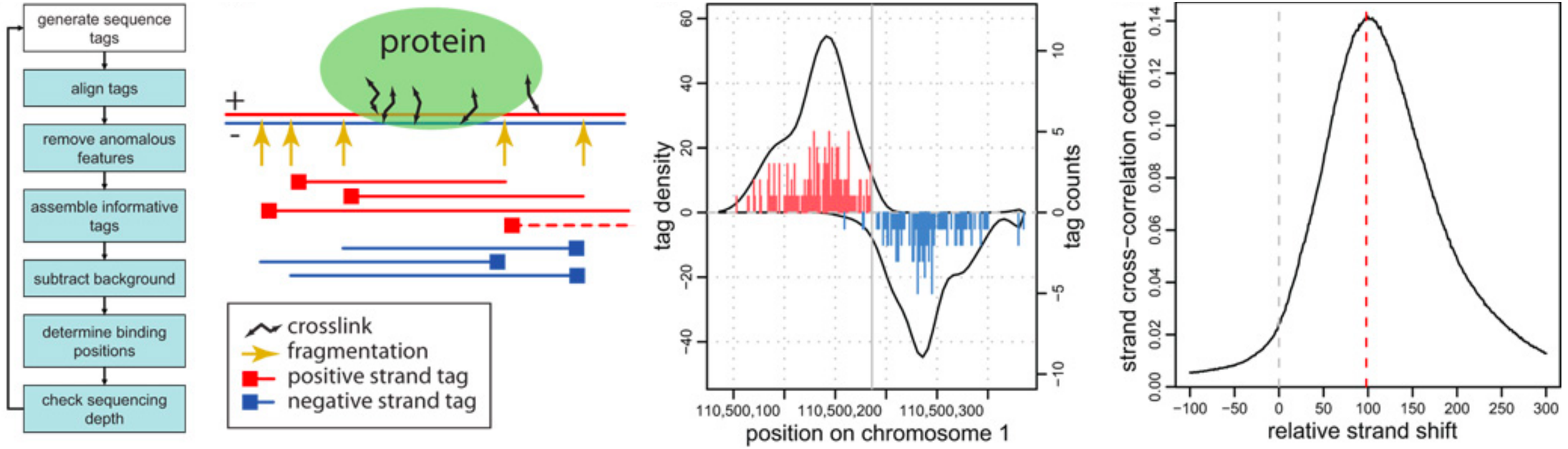


Detecting enriched regions



Kharchenko, Nature Biotech, 2008

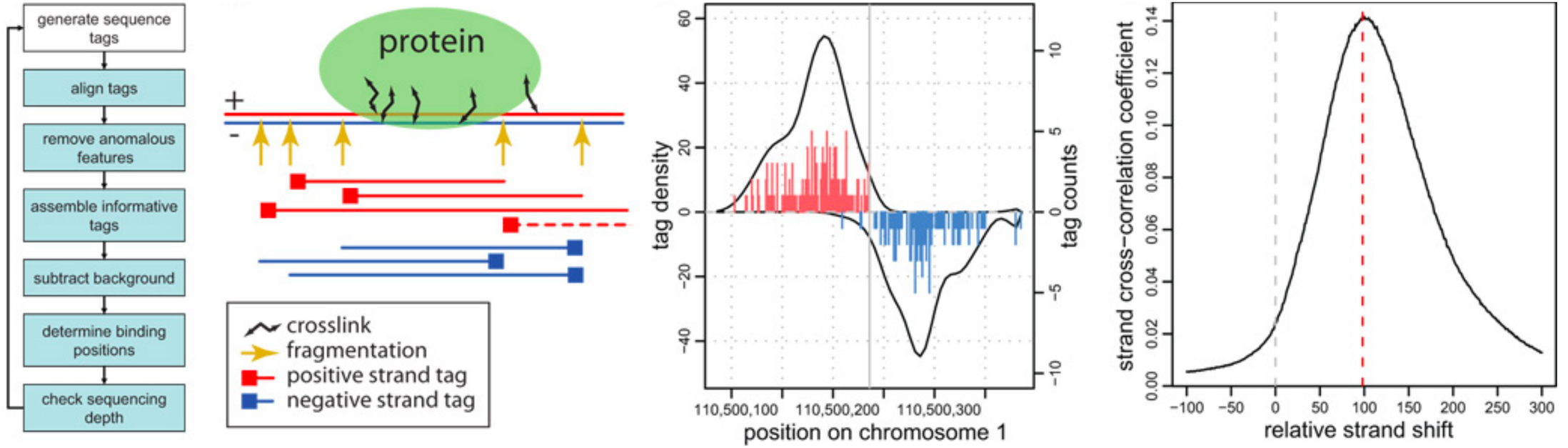
The SPP peak calling pipeline



Kharchenko, Nature Biotech, 2008

- ChIP-seq fragments are sequenced from the 5’ end only

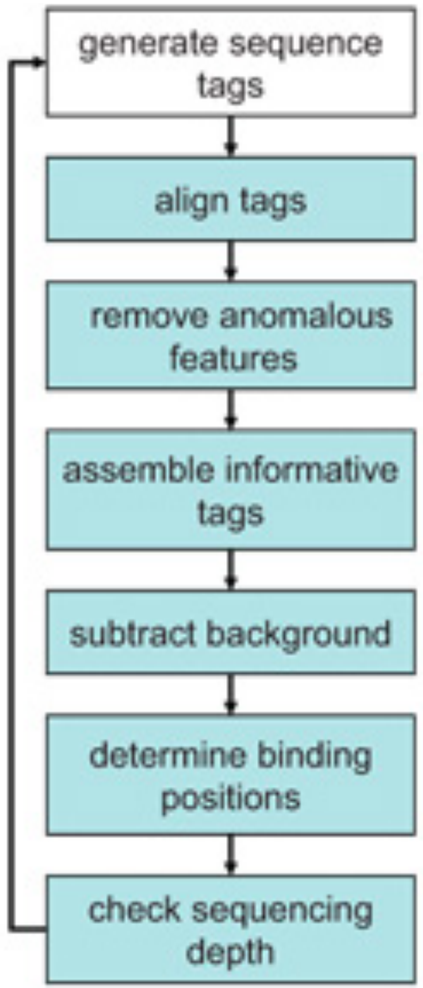
The SPP peak calling pipeline



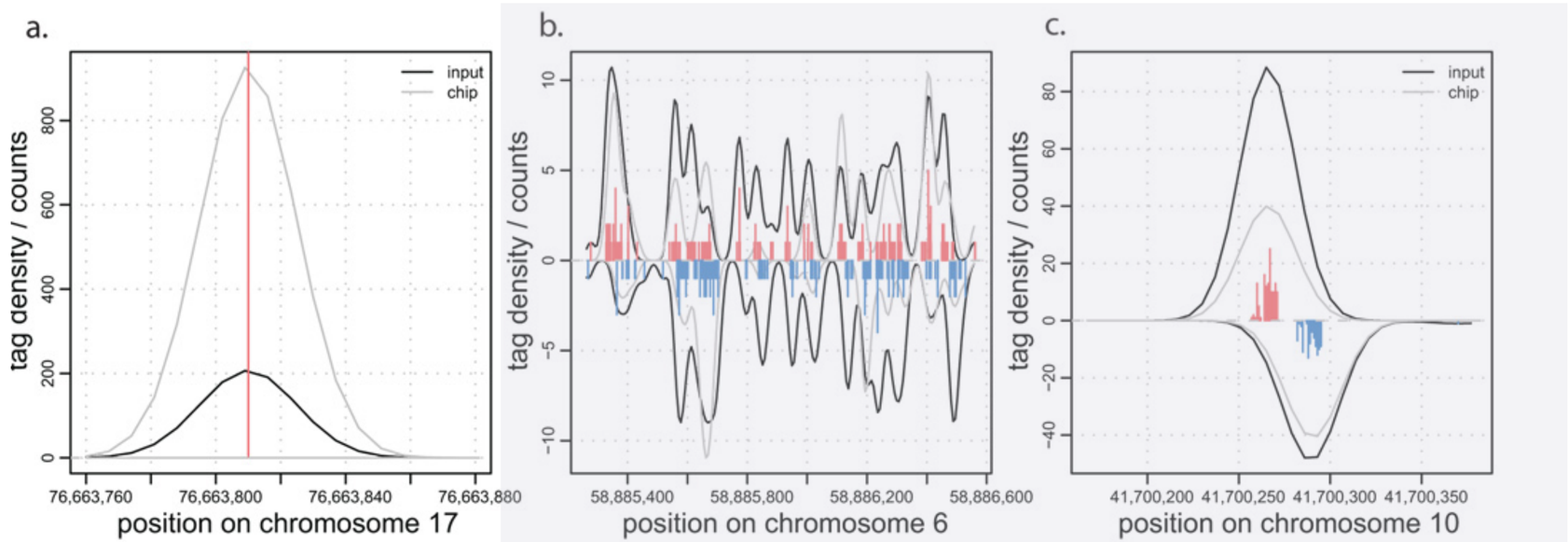
Kharchenko, Nature Biotech, 2008

- ChIP-seq fragments are sequenced from the 5’ end only
- Alignment generates bimodal pattern which is used to estimate the relative strand shift

The SPP peak calling pipeline



Density of tags from ChIP and input samples showing three types of anomalies



Singular positions with extremely high tag count.

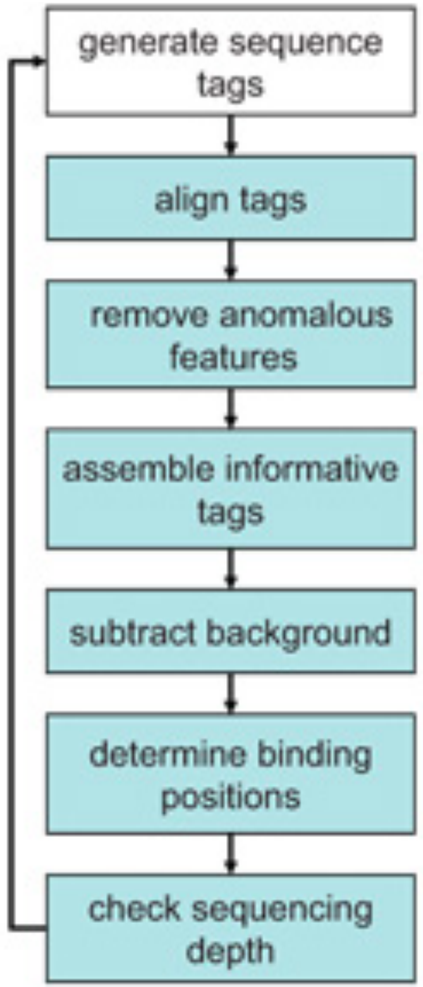
Larger (>1000bp), non-uniform regions of increased background tag density.

Background tag density patterns resembling true protein binding positions.

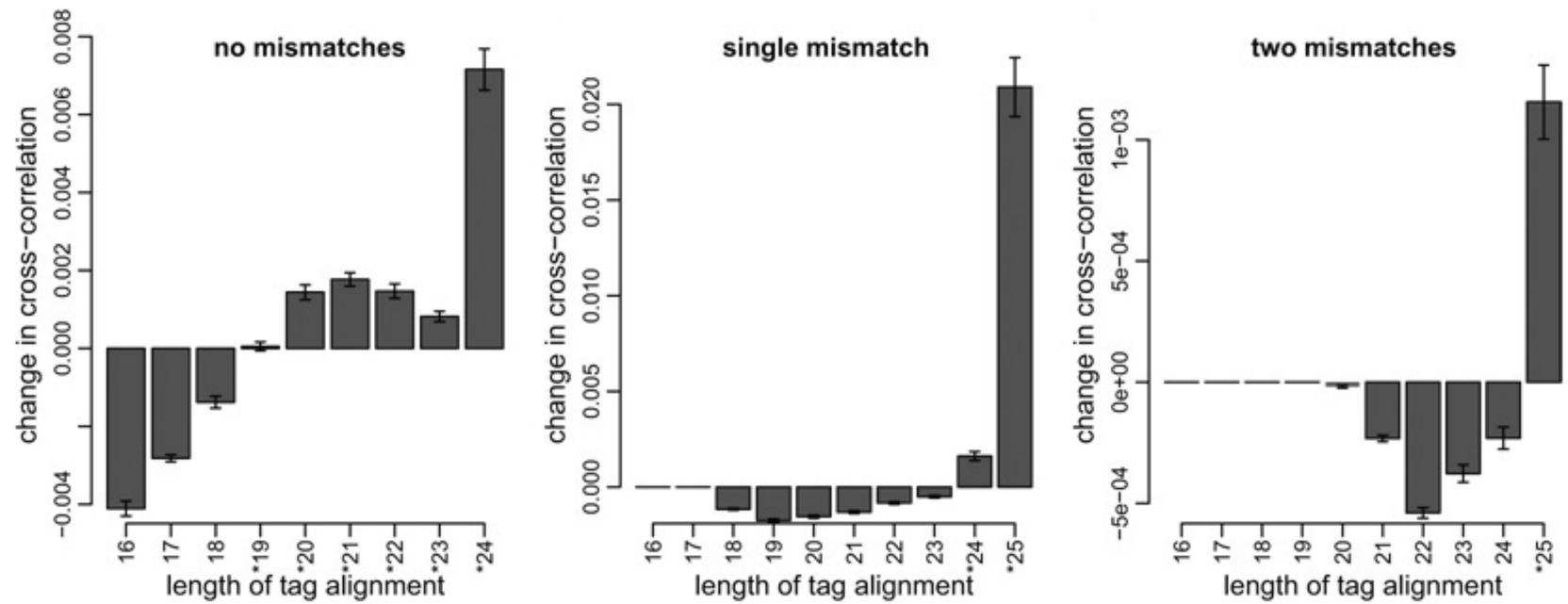
Positions with number of mapped sequence tags (5' ends) with Z-score >10. All tags mapping to such anomalous position (on either strand) are omitted.

Kharchenko, Nature Biotech, 2008

SPP: Remove anomalous features



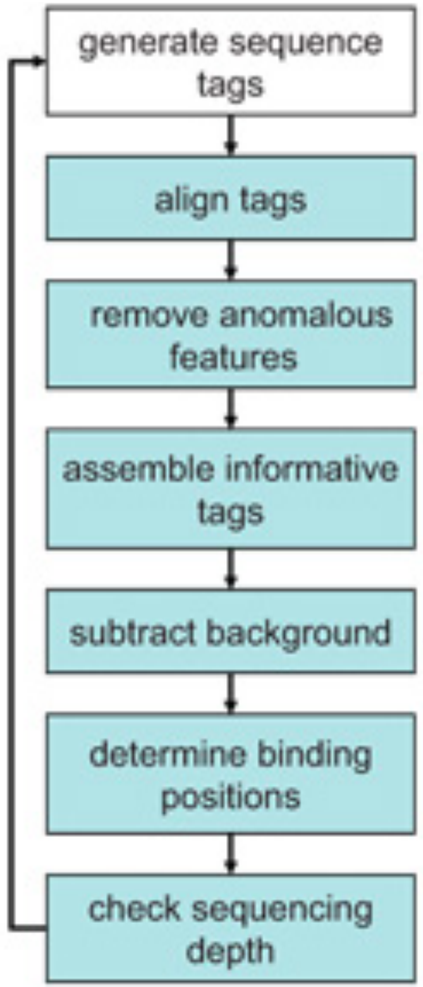
Select informative tag classes based on change in strand cross-correlation magnitude



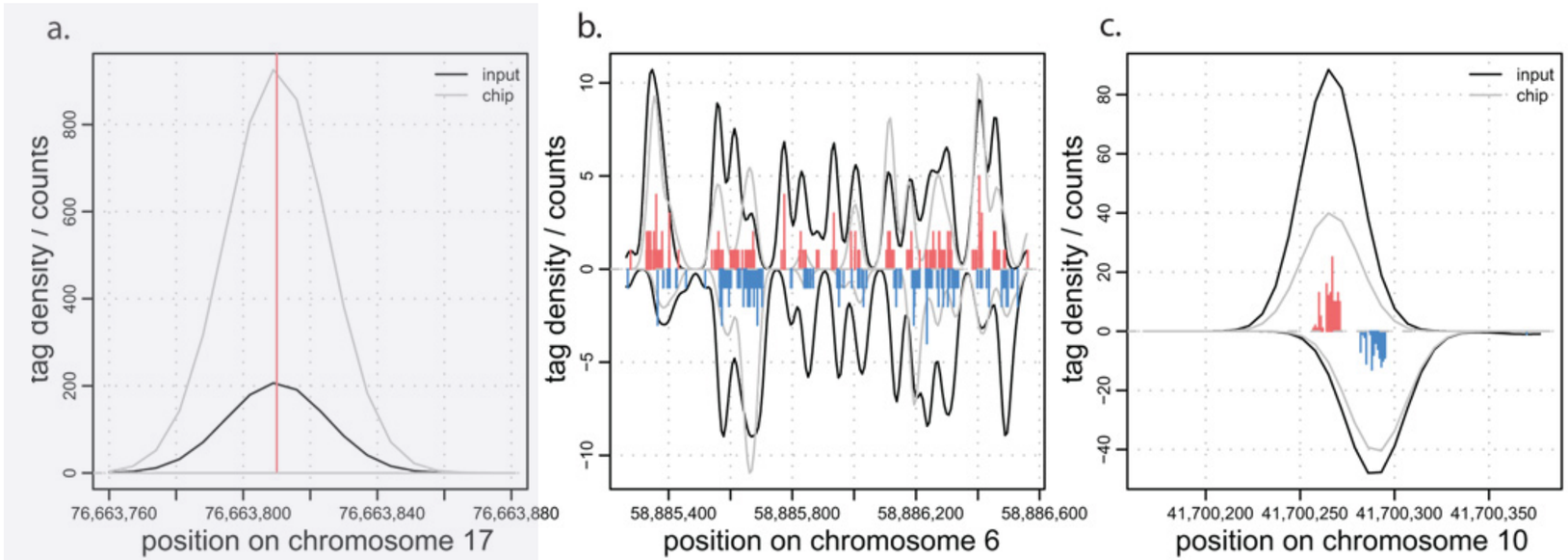
Informative tag classes improve cross-correlation (marked by *), and are incorporated into the final tag set. The y-axis gives the mean change in cross-correlation profile within 40bp around the cross-correlation peak

Kharchenko, Nature Biotech, 2008

SPP: Assemble informative tags



Density of tags from ChIP and input samples showing three types of anomalies



Singular positions with extremely high tag count.

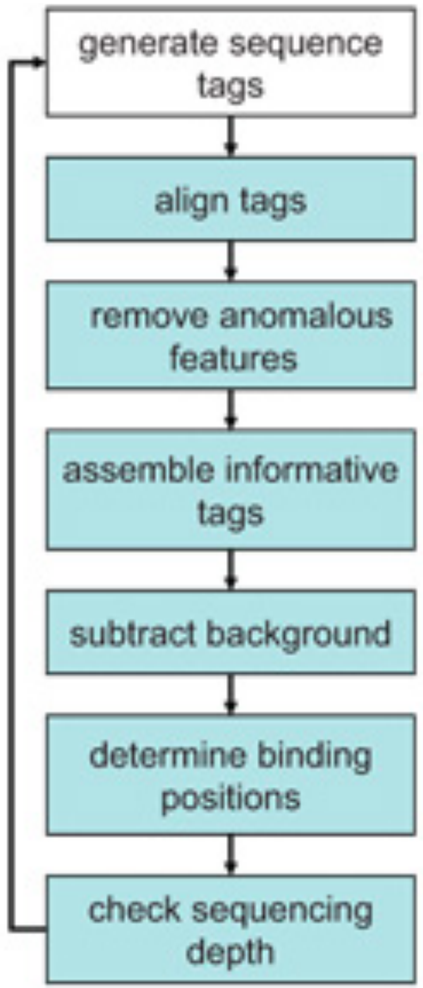
Larger (>1000bp), non-uniform regions of increased background tag density.

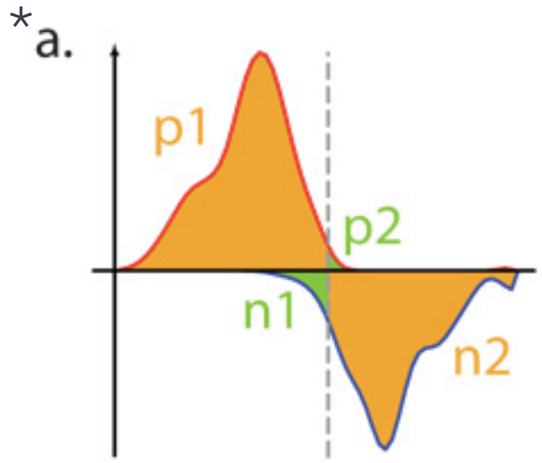
Background tag density patterns resembling true protein binding positions.

WTD and MTC methods were adjusted by subtracting the weighted number of background (input) tags occurring within that window.

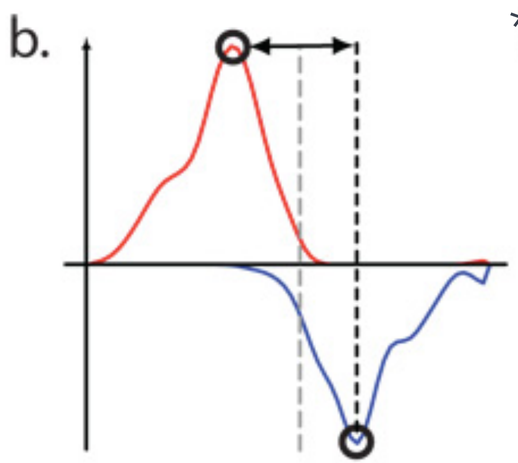
Kharchenko, Nature Biotech, 2008

SPP: Subtract background

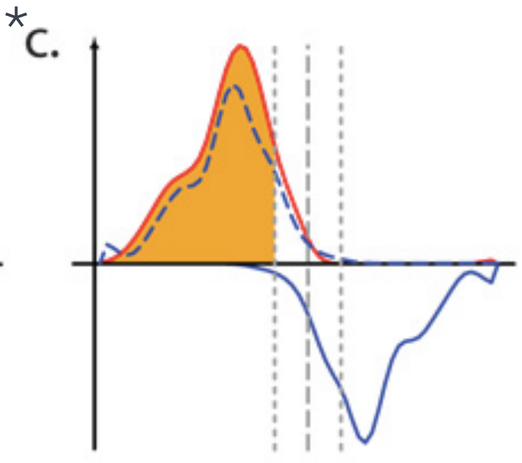




Window Tag Density (WTD)
Calculates the difference between geometric average of the tag counts within the regions marked by orange color (p1 and n2), and the average tag count within the regions marked by green color (n1 and p2). Window size based on average binding tag pattern (estimated from CC plot)



Matching Strand Peaks (MSP)
Identifies local maxima on positive and negative strands and then determines positions where such two peaks are present in the right order, with the expected separation (e.g. 20bp) and comparable magnitude (based on a likelihood ratio test)



Mirror Tag Correlation (MTC)
Similar to WTD. Based on the mirror correlation of the positive and negative strand tag densities. The mirror image of the negative strand density is shown by the blue dashed line. Uses the Pearson linear correlation coefficient.

Kharchenko, Nature Biotech, 2008

SPP: Determine binding positions

Estimation of the False Discovery Rate (FDR)

- Expected proportion of false discoveries among total rejections of the null hypothesis (i.e. how many false positive peaks of the total set of peaks)
- FDR determined empirically by exchanging ChIP and control
- Given by

$$\mathrm{FDR}(s) = \frac{N_r(s) + 0.5}{N_c(s) + 0.5}$$

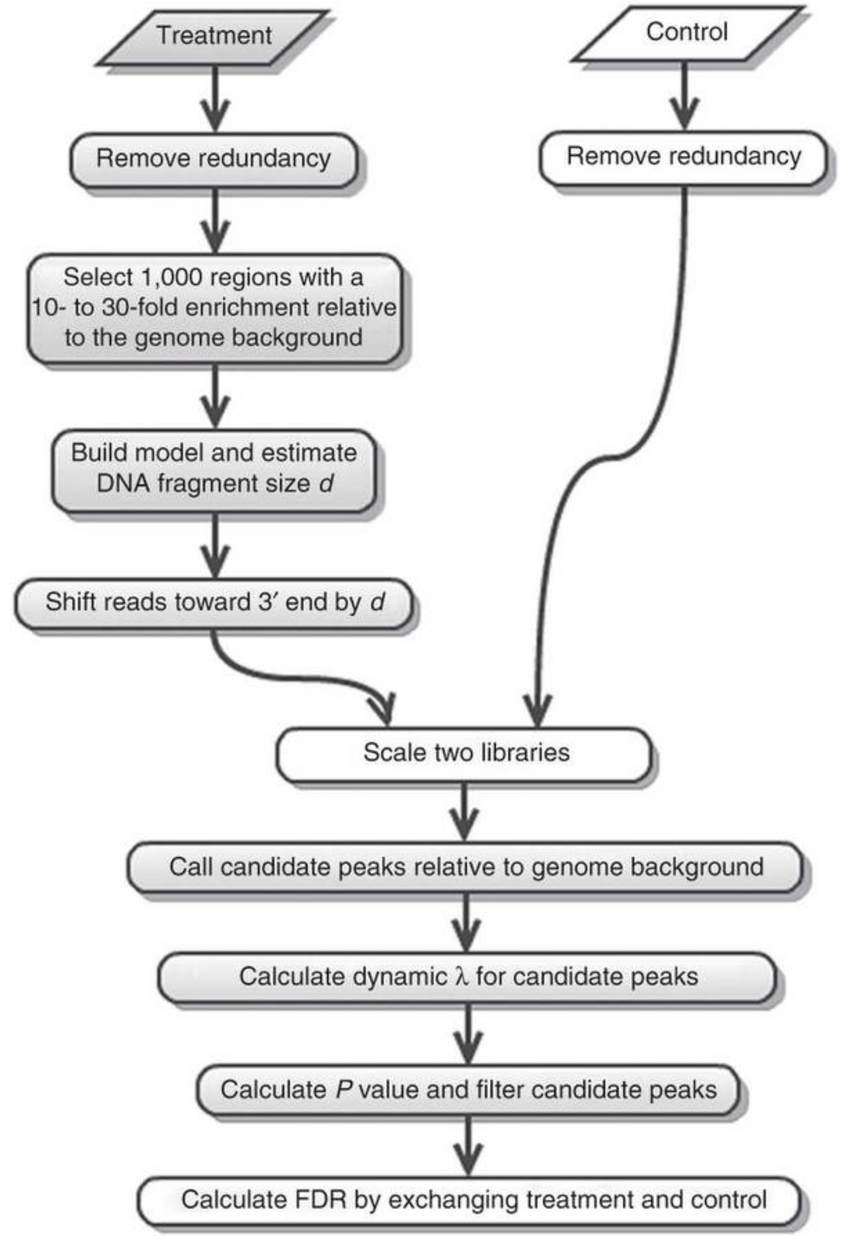
where $N_r(s)$ is the number of binding positions with score s or higher found in the real dataset, and $N_C(s)$ is the number found in a control dataset

MACS peak calling

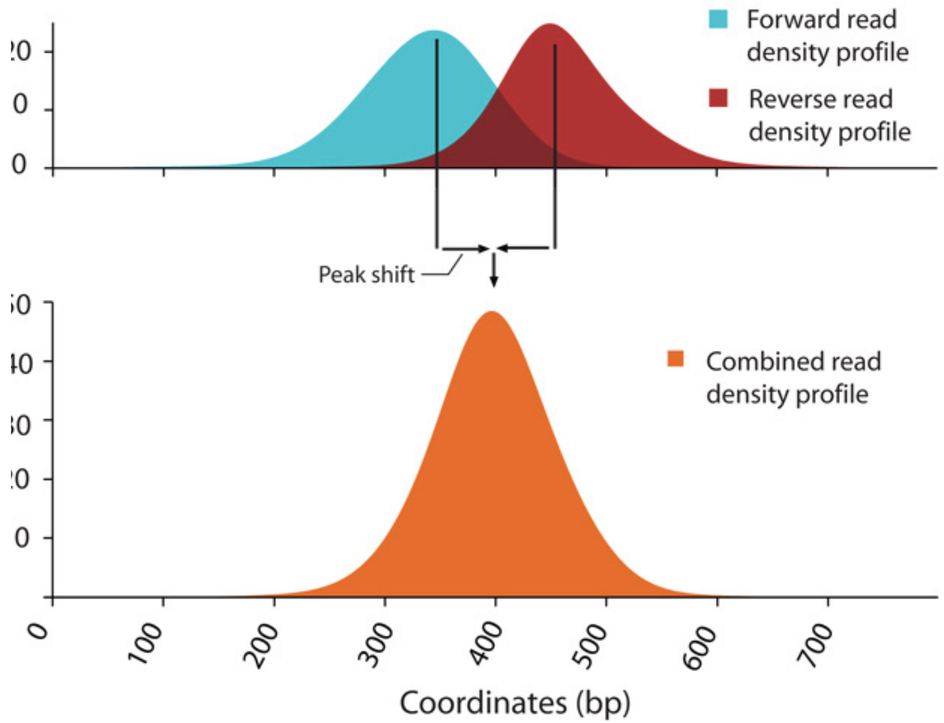
Zhang et al. Model-based analysis of ChIP-Seq, Genome Biol. (2008)

Developed for detection of transcription factor binding sites

Also suited for larger regions

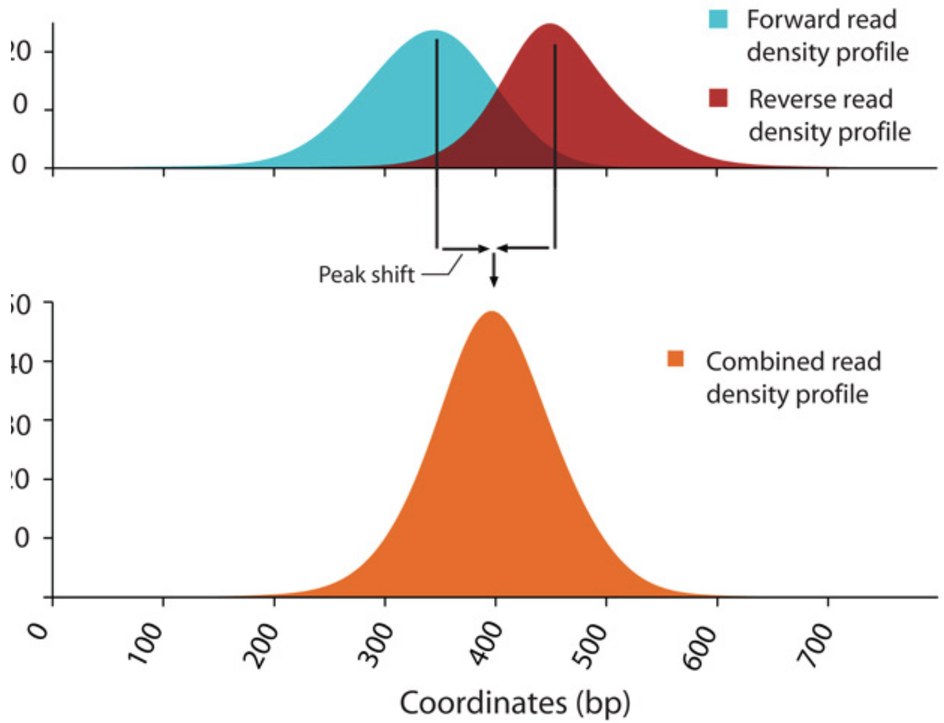


Model read distribution and calculate the shift size, *d/2*



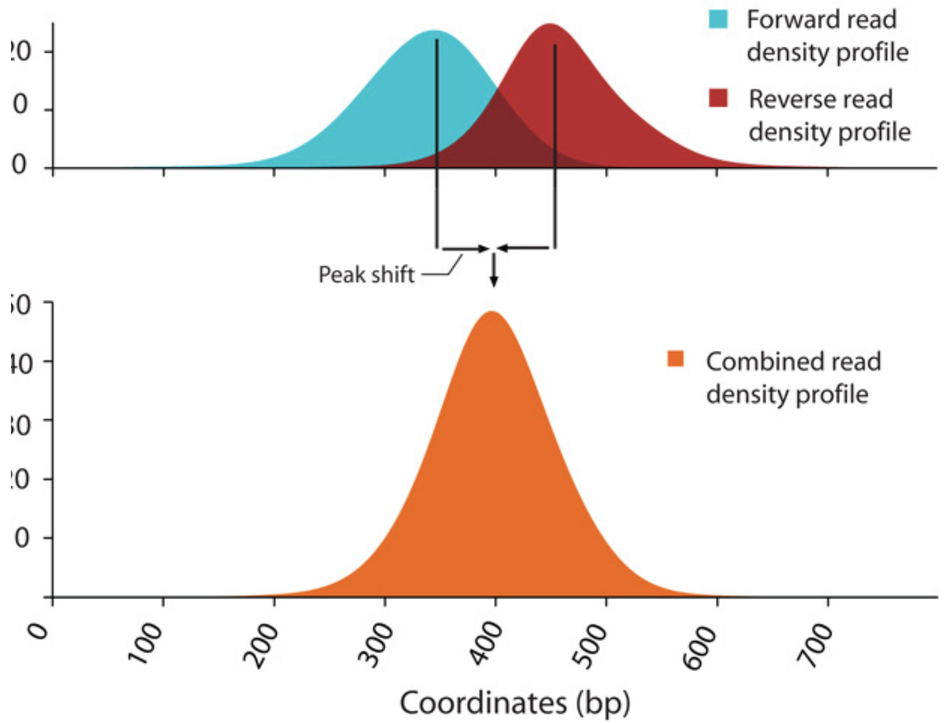
Model read distribution and calculate the shift size, *d/2*

- Find paired peaks using a sliding window 2*size of sequence fragments
  - > Search for regions where reads are enriched more than *MFOLD*



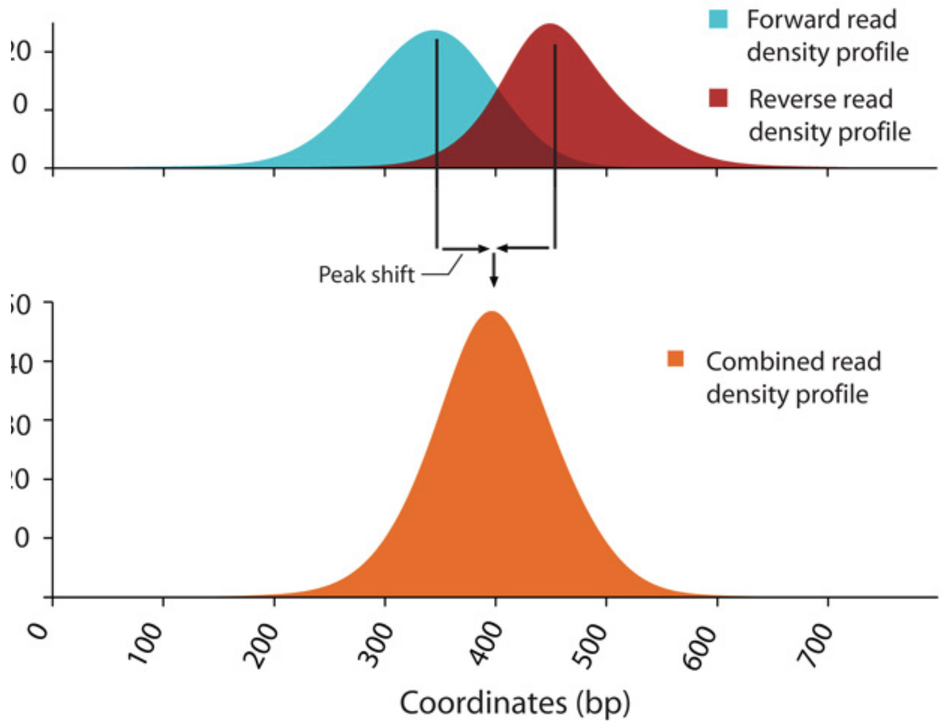
Model read distribution and calculate the shift size, *d/2*

- Find paired peaks using a sliding window 2*size of sequence fragments
  - > Search for regions where reads are enriched more than *MFOLD*
- Estimate the fragment length, '*d*'
  - > Distance between the modes of the positive and negative strand peaks



Model read distribution and calculate the shift size, *d/2*

- Find paired peaks using a sliding window 2*size of sequence fragments
  - > Search for regions where reads are enriched more than *MFOLD*
- Estimate the fragment length, '*d*'
  - > Distance between the modes of the positive and negative strand peaks
- Peak calling
  - > Shift the tags by 1/2 the fragment length, '*d*'
  - > Scan for enriched peaks comparing to background



Peak detection

MACS models the number of reads from a genomic region/window using a Poisson distribution

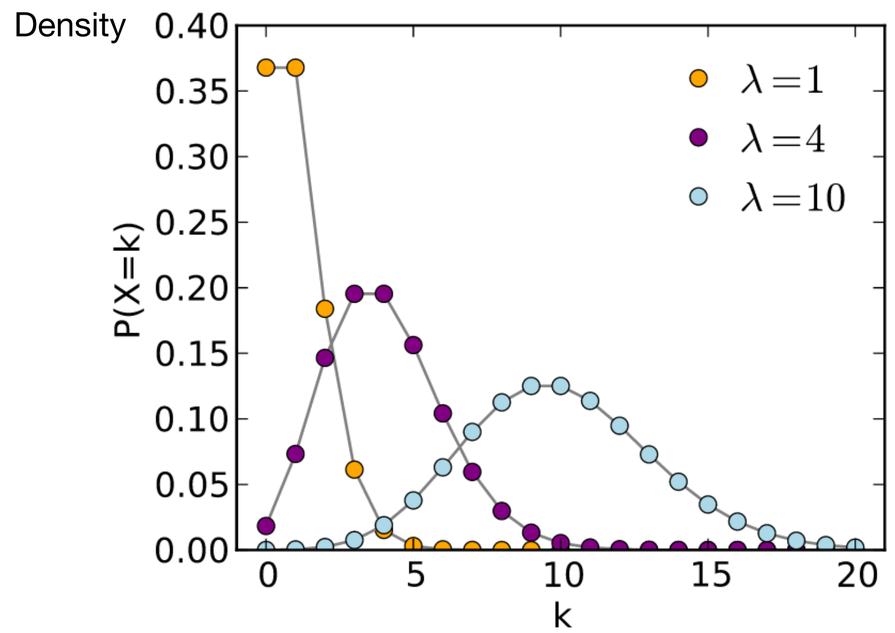
Poisson is a one parameter model, where the parameter λ is the expected number of reads in that window.

$$P_\lambda(X=k) = \frac{\lambda^k}{k! * e^{-\lambda}}$$

λ = mean = expectated value = variance

$$\lambda = \frac{\text{total number of events (k)}}{\text{number of units (n) in the data}}$$

$$= \frac{\text{Read length (nt)} * \text{Total read number}}{\text{Effective genome length (nt)}}$$



http://en.wikipedia.org/wiki/Poisson_distribution

Significance of enrichment

- MACS estimates λ, the expected number of reads, from the control to determine the significance level
- The probability distribution function is given by

$$F_\lambda(n) = P(X \leq n) = \Sigma P_\lambda(k) = e^{-\lambda} \Sigma \lambda^k / k!$$

and the probability of observing more than n reads is $1 - F_\lambda(n)$

Background estimation

- Frequently observe more variance in the data than assumed by the Poisson distribution

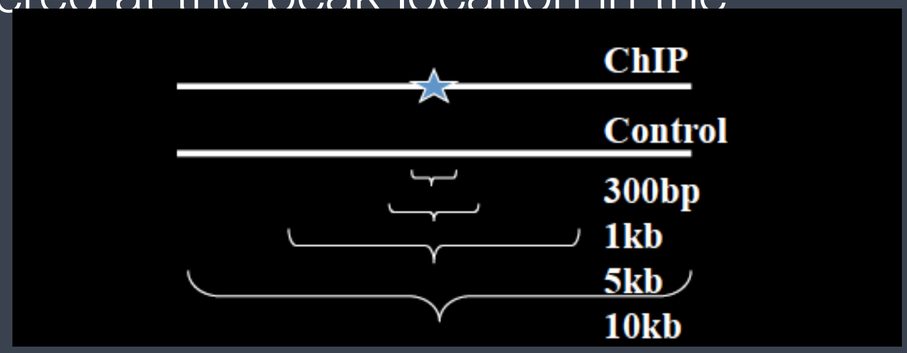
  > Local chromatin structure, PCR, sequencing bias, CNVs leads to false positive peaks

- Need to fit a distribution to smaller regions on the genome using sliding/discrete windows

- MACS uses a dynamic λ_{local}, determined as the maximum value of λ from the background at 1 kb, 5 kb and 10 kb windows centered at the peak location in the control sample

$$\lambda_{BG} = \text{total tags} / \text{genome size}$$

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

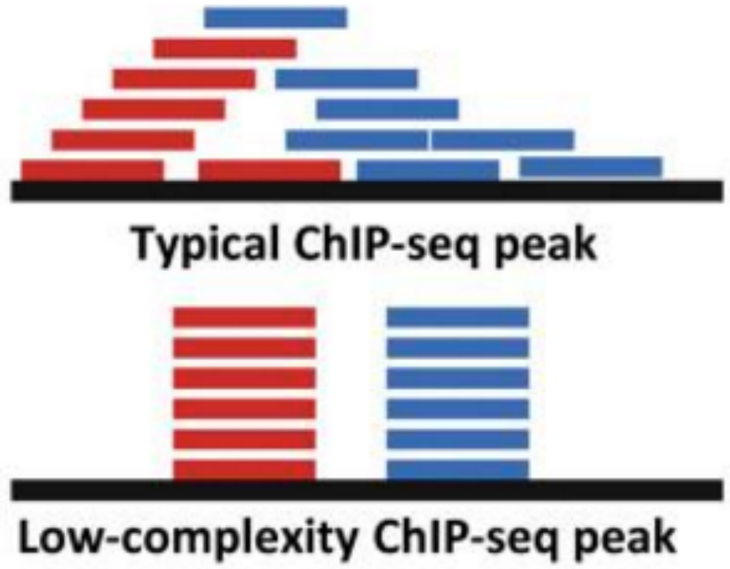


Library scaling

- When ChIP-seq and control samples are sequenced at different depths, MACS either linearly scales down the larger sample (default behavior) or scales up the smaller sample.
- If the total number of reads in the control sample is greater than the number of reads obtained from ChIP-seq by a factor of r (r > 1), then when calculating the P value λlocal will be divided by r by default.

Effective genome length (Mappability)

- Mappability is related to the uniqueness of the k-mers at a particular position the genome
- Repetitive regions have low uniqueness, which means low mappability



Landt et al, Genome Res. 2012

Effective genome length (Mappability)

- Not possible to unambiguously assign reads to all genomic regions
- ‘Mappability’ or uniqueness influences the average mapped depth
- Mappability improves with increased read length

Table 1. Proportions of unique start sites for nucleotide-space short tag alignments

Species	25 (1) (%)	30 (1) (%)	35 (1) (%)	50 (2) (%)	60 (3) (%)	75 (4) (%)	90 (5) (%)
Homo sapiens[a]	66.0	70.9	74.1	76.9	77.5	79.3	80.8
Mus musculus[b]	69.9	74.4	77.1	79.1	79.4	80.7	81.7
Caenorhabditis elegans[c]	85.3	87.7	89.0	89.8	89.9	90.6	91.1
Drosophila melanogaster[d]	67.5	68.4	69.0	69.2	69.2	69.5	69.8

Columns shown are length of tag matched; numbers in parentheses represent the number of mismatches allowed.
[a]Build hg19.
[b]Build mm9.
[c]Build ce6.
[d]Build dm3.

Koehler et al, Bioinformatics (2011)

Estimation of the false discovery rate (FDR)

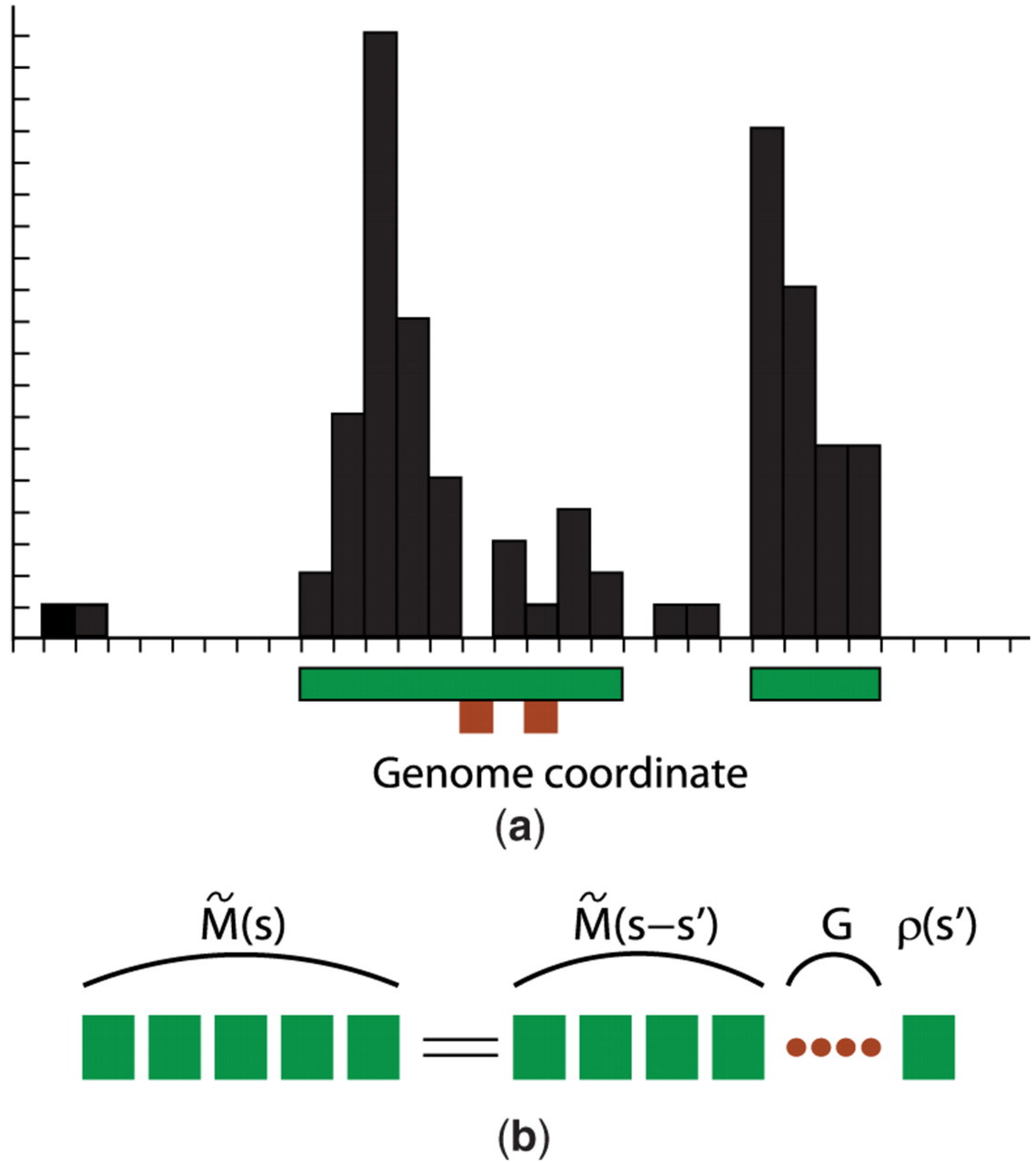
- Expected proportion of false discoveries among total rejections of the null hypothesis (i.e. how many false positive peaks of the total set of peaks)
- In MACSv1.4, FDR determined empirically by exchanging ChIP and control
- In MACSv2, p-values are calculated at every basepair in the genome and then corrected for multiple comparison using the Benjamini-Hochberg correction
- P-values and FDR are affected by sequencing depth with greater sequencing depth leading to lower p-vaues and FDRs

SICER

Hstone modifications tend to cluster to form domains

Identifies spatial clusters of signals unlikely to appear by chance

Pools together enrichment information from neighboring nucleosomes to increase sensitivity and specificity.

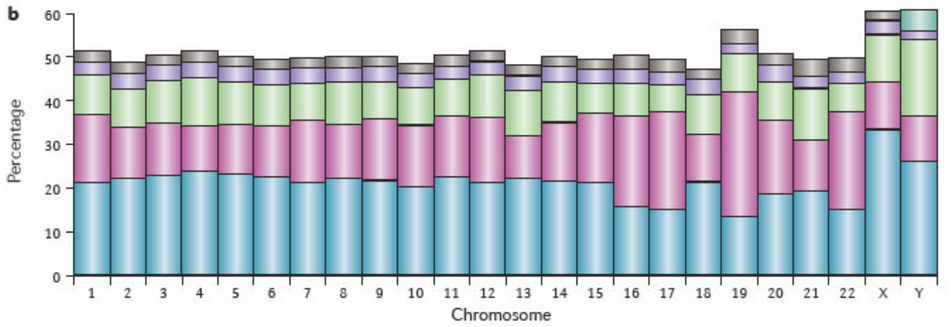


ENCODE Blacklisted Regions

- Regions with anomolous, unstructured, high read counts in NGS experiments
- High ratio of multi-mapping to unique mapping reads
- Overlap repeat elements (satellites, centromeres, telomeres)
- Recommended to use this blacklist to filter regions

a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000



Treangen, T.J. & Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13, 36-46.

Dealing with replicates

Dealing with replicates

Dealing with replicates

- Technical replicates are usually merged before peak calling

Dealing with replicates

- Technical replicates are usually merged before peak calling
- Biological replicates are analyzed separately and compared at the peak call level

Dealing with replicates

- Technical replicates are usually merged before peak calling
- Biological replicates are analyzed separately and compared at the peak call level
- Historical ENCODE guidelines:

> either 80% of the top 40% of the targets identified from one replicate using an acceptable scoring method should overlap the list of targets from the other replicate, OR

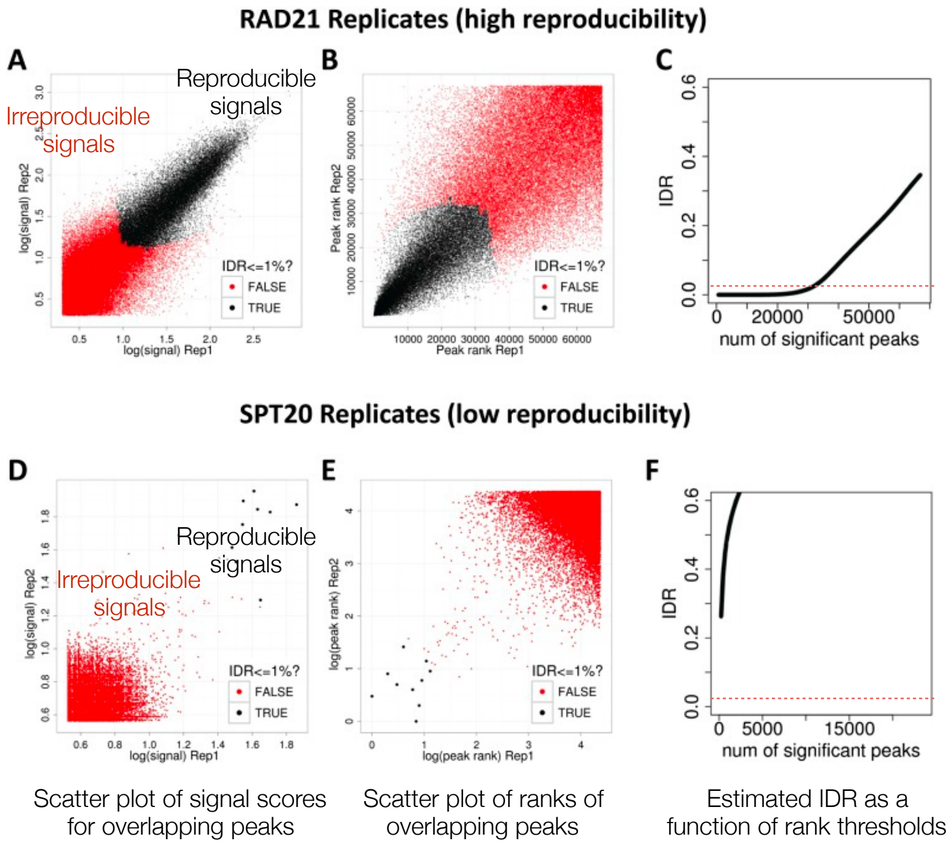
> target lists scored using all available reads from each replicate should share more than 75% of targets in common

Dealing with replicates

- Technical replicates are usually merged before peak calling
- Biological replicates are analyzed separately and compared at the peak call level
- Historical ENCODE guidelines:
  - > either 80% of the top 40% of the targets identified from one replicate using an acceptable scoring method should overlap the list of targets from the other replicate, OR
  - > target lists scored using all available reads from each replicate should share more than 75% of targets in common
- Better approach: IDR

Dealing with replicates: IDR

- Compares a pair of ranked lists of peaks for consistency over the replicates to separate signal from noise.
- The most significant peaks (i.e. genuine signals) are expected to have high consistency between replicates.
- Peaks with low significance (i.e. noise) are expected to have low consistency.



Scatter plot of signal scores for overlapping peaks

Scatter plot of ranks of overlapping peaks

Estimated IDR as a function of rank thresholds

(2012) ENCODE: TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework

Contents

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

Latest: https://github.com/nboley/idr

Differential ChIP-Seq analysis

Detecting differences in ChIP signal between two conditions

Detecting differences in ChIP signal between two conditions

- Challenging compared to RNA-seq and whole genome bisulfite sequencing

Detecting differences in ChIP signal between two conditions

- Challenging compared to RNA-seq and whole genome bisulfite sequencing
- Search space is not limited to a particular region of the genome

Detecting differences in ChIP signal between two conditions

- Challenging compared to RNA-seq and whole genome bisulfite sequencing
- Search space is not limited to a particular region of the genome
- Range of the signal is not constrained to finite interval; requires transformation to apply standard statistical tools

Detecting differences in ChIP signal between two conditions

- Challenging compared to RNA-seq and whole genome bisulfite sequencing
- Search space is not limited to a particular region of the genome
- Range of the signal is not constrained to finite interval; requires transformation to apply standard statistical tools
- Amount of noise in considerable, making variations in the signal challenging to detect, especially when these differences are subtle

Detecting differences in ChIP signal between two conditions

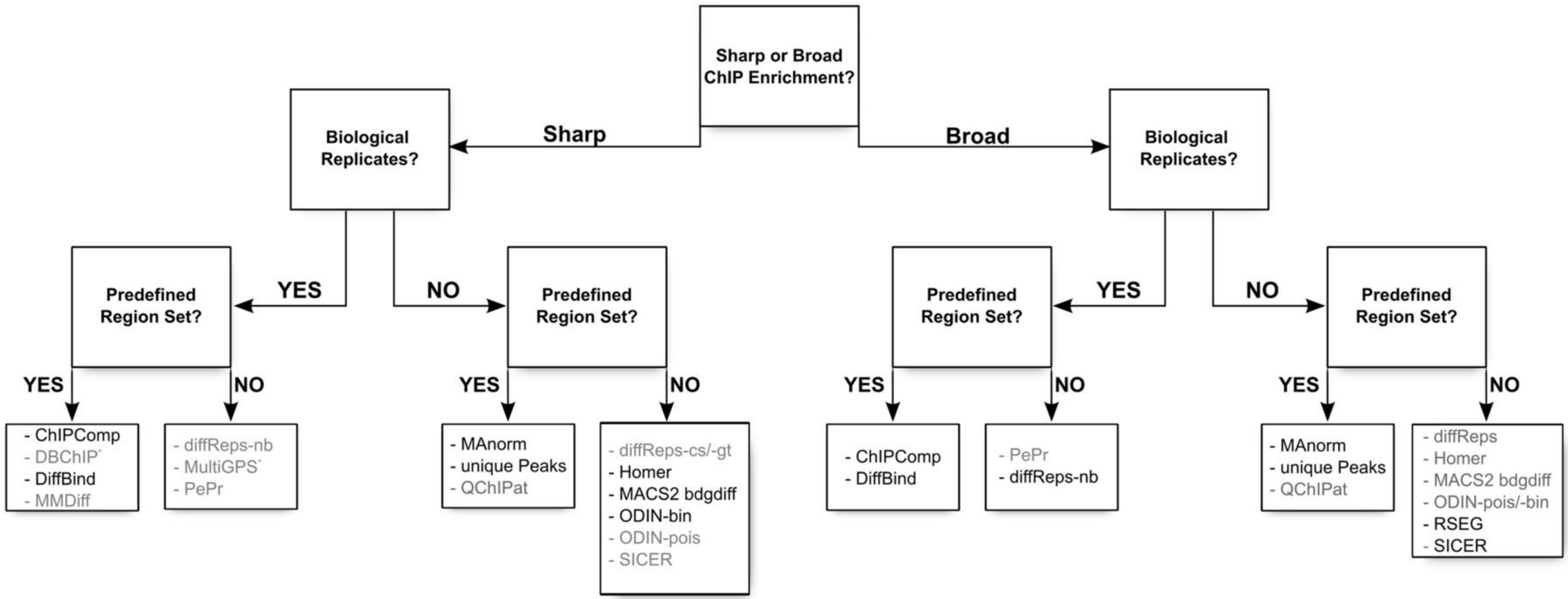
- Challenging compared to RNA-seq and whole genome bisulfite sequencing
- Search space is not limited to a particular region of the genome
- Range of the signal is not constrained to finite interval; requires transformation to apply standard statistical tools
- Amount of noise in considerable, making variations in the signal challenging to detect, especially when these differences are subtle
- The properties of the enriched regions (in particular their length) differ substantially depending on the protein or epigenetic modification targeted by the immunoprecipitation

Software for Differential Binding

Tool	Peak Calling	Normalization	Statistical Test	Sharp Signal	Broad Signal	Biological Replicates	Significance Measure	PMID
SICER	Window based approach, merging of eligible clusters in proximity closer than defined gap size	Library size	Poisson distribution, ChIP norm. counts in possible island against Control norm. counts	n	y	n	FDR	19505939
MACS2	Not required	Library size	Computation of log10 likelihood ratios and setting a pre-defined cutoff for following comparisons: Cond1 > Cond2 and Cond1 > Control1 Cond1 < Cond2 and Cond2 > Control2	y	y	n	log10 likelihood ratio	18798982
ODIN	Not required	- SES normalization (ChIP / input) combined with input subtraction - Library size normalization (Cond1 / Cond2)	Hidden Markov Model (HMM) with a three state topology Emissions are calculated with a Binomial or a mixture of Poisson distribution	y	y	n	p-value	25371479
RSEG	Not required	-	Hidden Markov Model (HMM) with a three state topology NBDiff distribution is used to model read count differences between both conditions	n	y	n	-	21325299
MAnorm	Requires peak calling e.g. with MACS	Genome-wide MA plot combined with LOWESS Regression	Bayesian model approach	y	y	n	p-value	22424423
HOMER	Window based approach; Peak calling done by HOMER	Library size normalization	Fold-change thresholding combined with a Poisson distribution based enrichment analysis	y	y	n	FDR or p-value	20513432
QChIPat	Peak calling possible with BELT, MACS, SISSRs or FindPeaks	1) Nonparametric empirical Bayes correction normalization 2) Quantile normalization 3) Linear normalization	1) Wilcoxon rank sum test 2) Wilcoxon signed rank test	y	y	n	p-value	24564479
diffReps	Sliding window approach	Linear normalization	- Without replicates: G-test or Chi-square test - Replicates: exact negative binomial test	n	y	y/n	p-value	23762400
DBChip	Requires peak calling e.g. with MACS	median ratio strategy (DESeq)	Generalized linear model with negative Binomial distribution	y	n	y/n	FDR	22057161
ChIPComp	Requires peak calling e.g. with MACS	Normalization with a Poisson distribution based model	Wald's test followed by probability calculation Using a Bayesian approach	y	n	y	Posterior probability	25682068
MultiGPS	Expectation maximization learning scheme	edgeR		y	n	y	p-value	24675637
MMDiff	Requires peak calling e.g. with MACS	DESeq	Kernel-based non-parametric test	y	n	y	p-value	24267901
DiffBind	Requires peak calling e.g. with MACS	Differential peak analysis can be performed with: 1) DESeq 2) DESeq2 3) edgeR		y	y	y	p-value or FDR	22217937
PePr	Window based approach	Trimmed Mean of M values (TMM) approach (edgeR)	Binomial distribution	y	y	y	p-value	24894502

Steinhauser et al, Briefings in Bioinformatics (2016)

Decision tree indicating the proper choice of tool depending on the data set



Sebastian Steinhauser et al. Brief Bioinform 2016;bib.bbv110

Briefings in Bioinformatics