UNIVERSITY OF CAMBRIDGE

MRC Cancer Unit

Peak-calling for ChIP-seq and ATAC-seq

Shamith Samarajiwa

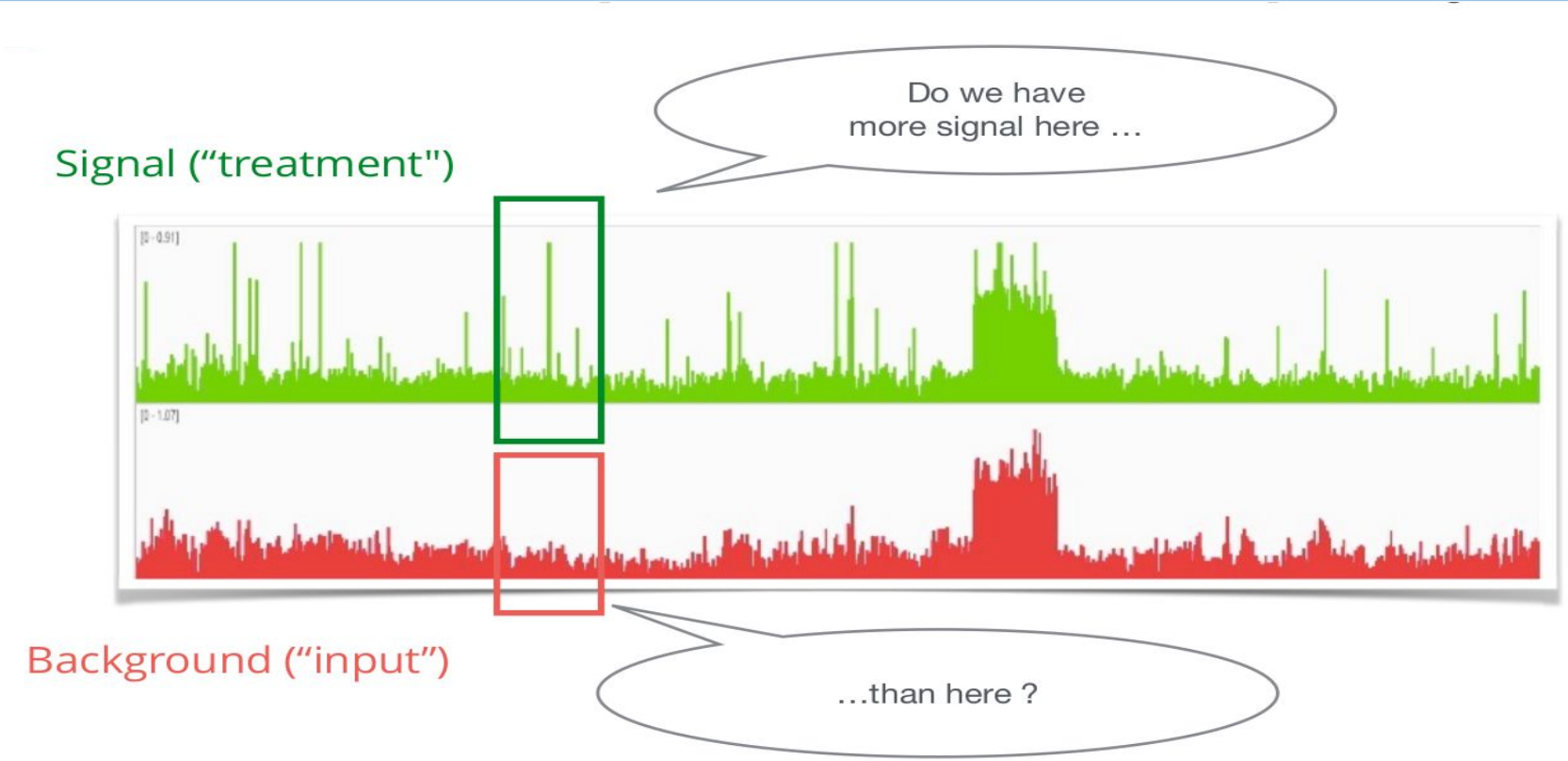
CRUK Autumn School in Bioinformatics 2017

University of Cambridge

Overview

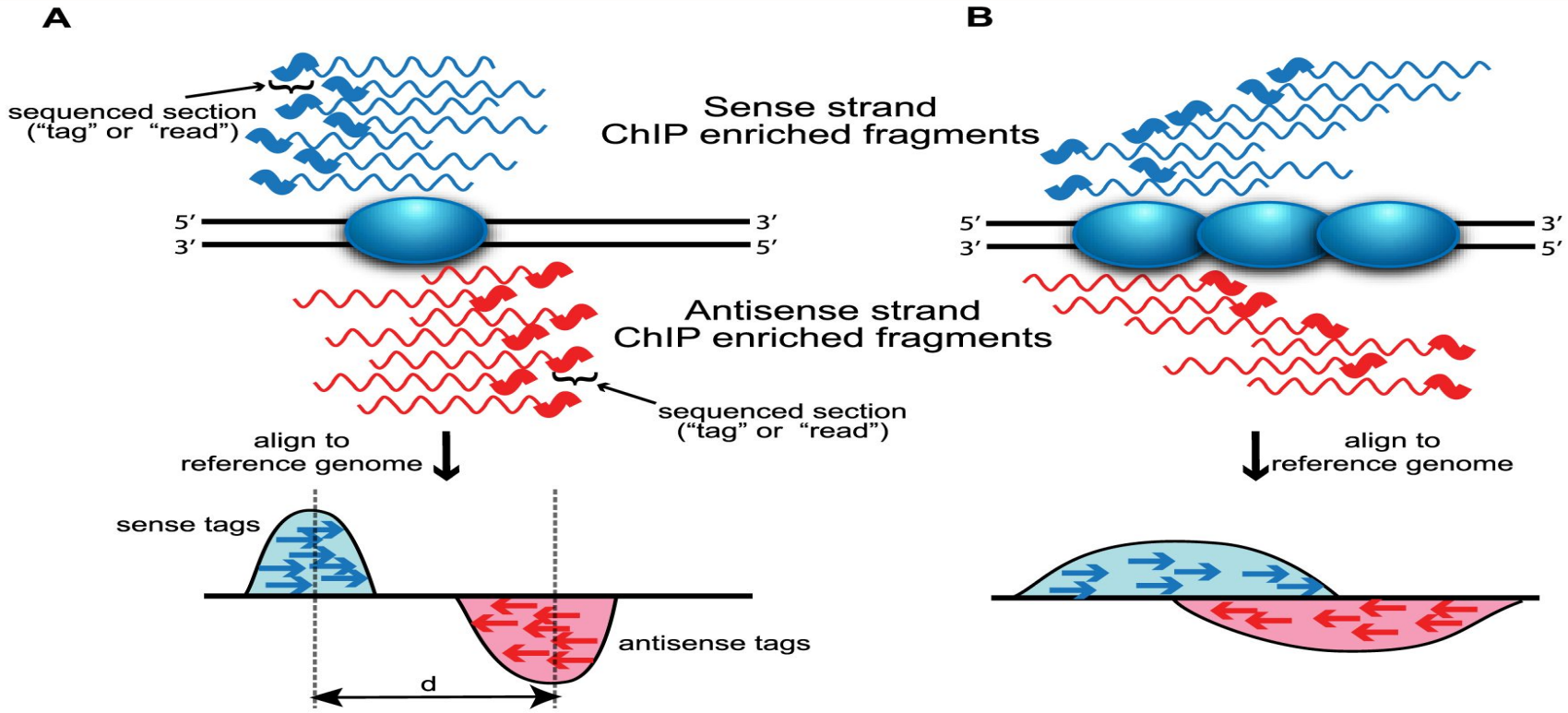
- ★ Peak-calling: identify enriched (signal) regions in ChIP-seq or ATAC-seq data
  - Software packages
  - Practical and Statistical aspects (Normalization, IDR, QC measures)
  - MACS peak calling
  - Overview of transcription factor, DNA binding protein, histone mark and nucleosome free region peaks
  - Narrow and Broad peaks
  - A brief look at the MACS2 settings and methodology
  - ATAC-seq signal detection

Signal to Noise



modified from Carl Herrmann

Strand dependent bimodality



Wilbanks et al. 2010 PLOS One

Peak Calling Software

★ Comprehensive list is at: https://omictools.com/peak-calling-category

MACS2 *(MACS1.4)*	Most widely used peak caller. Can detect narrow and broad peaks.
Epic *(SICER)*	Specialised for broad peaks
BayesPeak	R/Bioconductor
Jmosaics	Detects enriched regions jointly from replicates
T-PIC	Shape based
EDD	Detects megabase domain enrichment
GEM	Peak calling and motif discovery for ChIP-seq and ChIP-exo
SPP	Fragment length computation and saturation analysis to determine if read depth is adequate.

Quality Measures

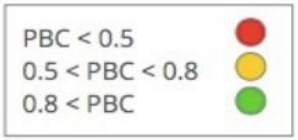
- Fraction of reads in peaks (FRiP) is dependant on data type.

$$FRiP = \frac{reads \in peaks}{total\ reads}$$

FRiP can be calculated with deepTools2

- PCR Bottleneck Coefficient (PBC) is a measure of library complexity

$$PBC = \frac{N_1}{N_2}$$



N1= Non redundant, uniquely mapping reads
N2= Uniquely mapping reads

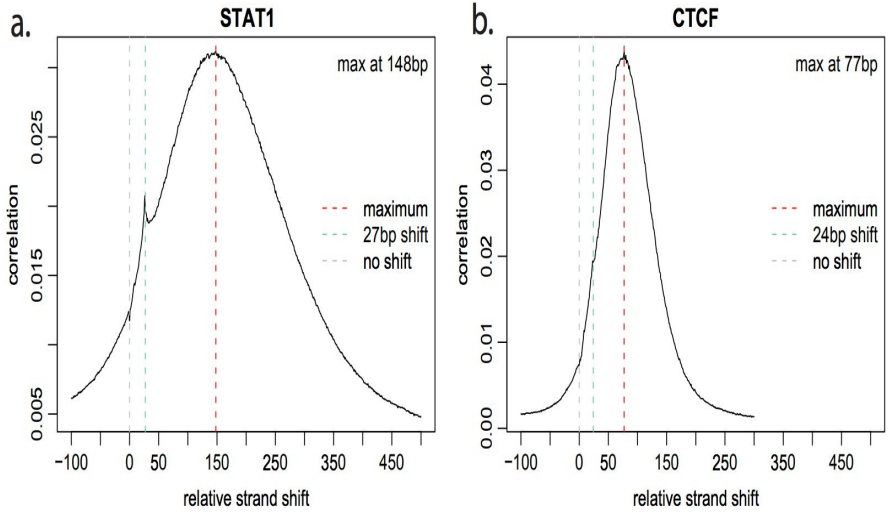
Preseq and preseqR for determining library complexity
Daley et al., 2013, Nat. Methods

Quality Measures

- Relative strand cross-correlation

The RSC is the ratio of the fragment-length cross-correlation value minus the background cross-correlation value, divided by the phantom-peak cross-correlation value minus the background cross-correlation value.

The minimum possible value is 0 (no signal) highly enriched experiments have values greater than 1, and values much less than 1 may indicate low quality.



Supplementary Figure 1. Strand cross-correlation profiles are shown for STAT1 (**a**) and CTCF (**b**). The dashed red line marks the position of the maximum cross-correlation. The light-blue lines mark strand shift corresponding to the length of the Solexa tag reads. A jump of cross-correlation at such shift is present in some datasets, in particular STAT1. The gray line marks 0bp strand shift.

Quality Measures

- Normalised strand cross-correlation NSC

The ratio of the maximal cross-correlation value (which occurs at strand shift equal to fragment length) divided by the background cross-correlation (minimum cross-correlation value over all possible strand shifts).

Higher values indicate more enrichment, values > 1.1 are relatively low NSC scores, and the minimum possible value is 1 (no enrichment). This score is sensitive to technical effects; for example, high-quality antibodies such as H3K4me3 and CTCF score well for all cell types. This score is also sensitive to biological effects; narrow marks score higher than broad marks (H3K4me3 vs H3K36me3, H3K27me3) for all cell types and ENCODE production groups, and features present in some individual cells, but not others, in a population are expected to have lower scores.

A measure of enrichment derived without dependence on prior determination of enriched regions.

IDR: Irreproducible Discovery Rate

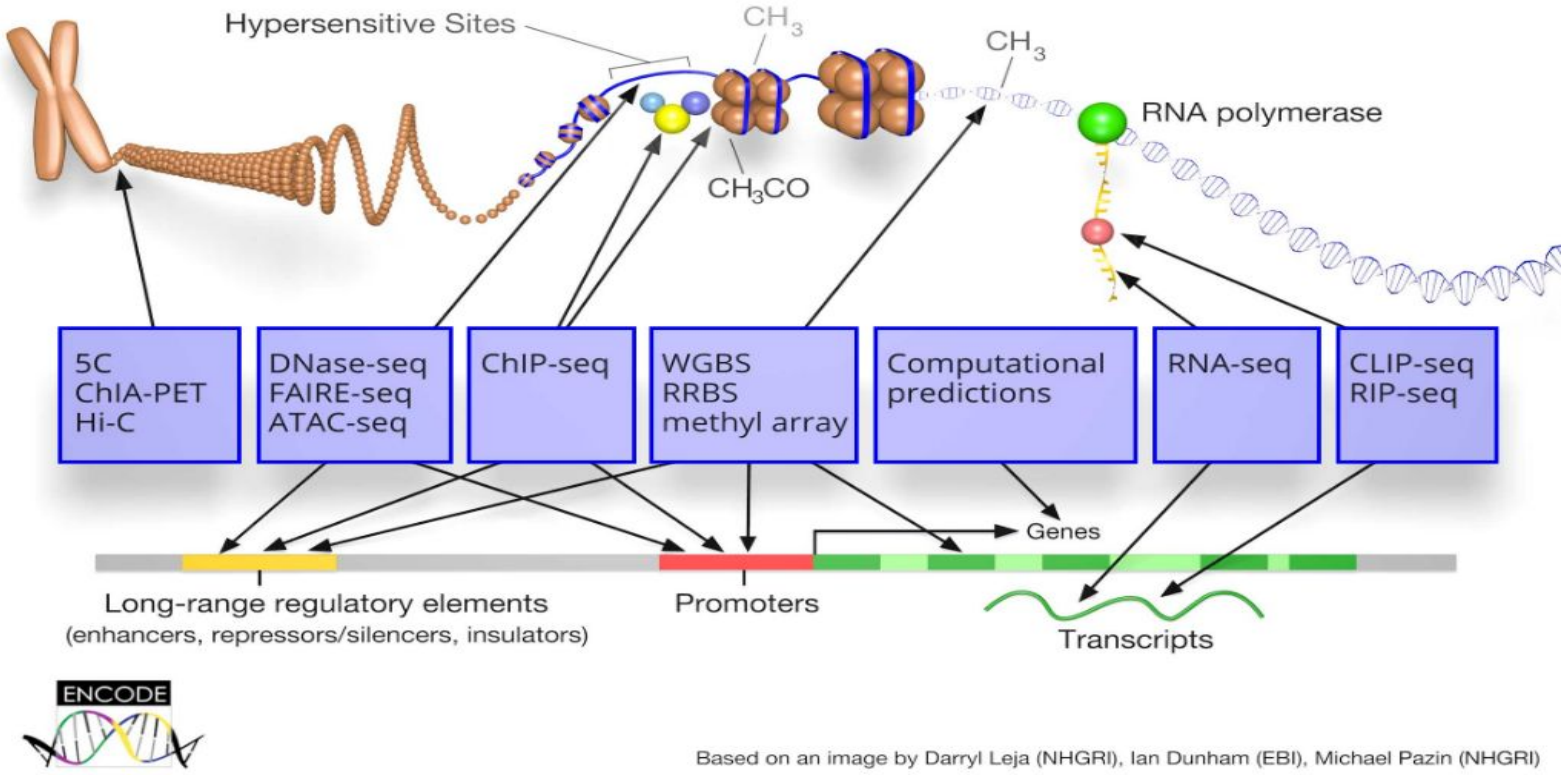
- If two replicates measure the same underlying biology, the most **significant peaks** which are likely to be genuine signals, are expected to have **high consistency between replicates**. Peaks with low significance, which are more likely to be noise, are expected to have low consistency.

- IDR **measures consistency between replicates** in high-throughput experiments. The IDR method compare 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).

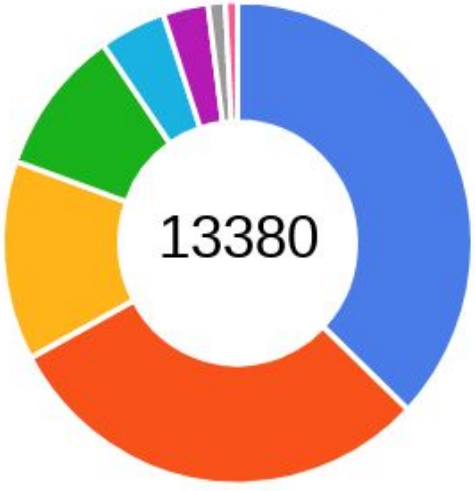
- The method uses reproducibility in score rankings between peaks in each replicate to determine an optimal cutoff for significance. The IDR method then fits the bivariate rank distributions over the replicates in order to separate signal from noise based on a defined confidence of rank consistency and reproducibility of identifications.

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

ENCODE

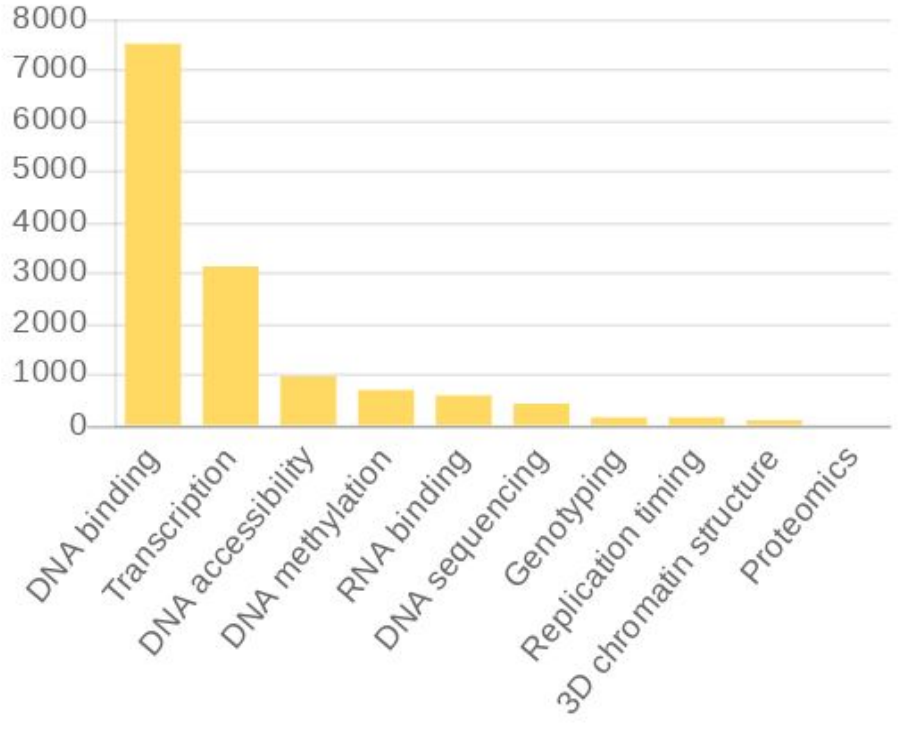

Hypersensitive Sites
CH3
CH3
CH3CO
RNA polymerase
5C
ChIA-PET
Hi-C
DNase-seq
FAIRE-seq
ATAC-seq
ChIP-seq
WGBS
RRBS
methyl array
Computational
predictions
RNA-seq
CLIP-seq
RIP-seq
Genes
Long-range regulatory elements
(enhancers, repressors/silencers, insulators)
Promoters
Transcripts
ENCODE
Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

Biosample Type



- immortalized cell line
- tissue
- primary cell
- whole organisms
- in vitro differentiated cells
- stem cell
- in vitro sample
- induced pluripotent stem cell line

Assay Categories



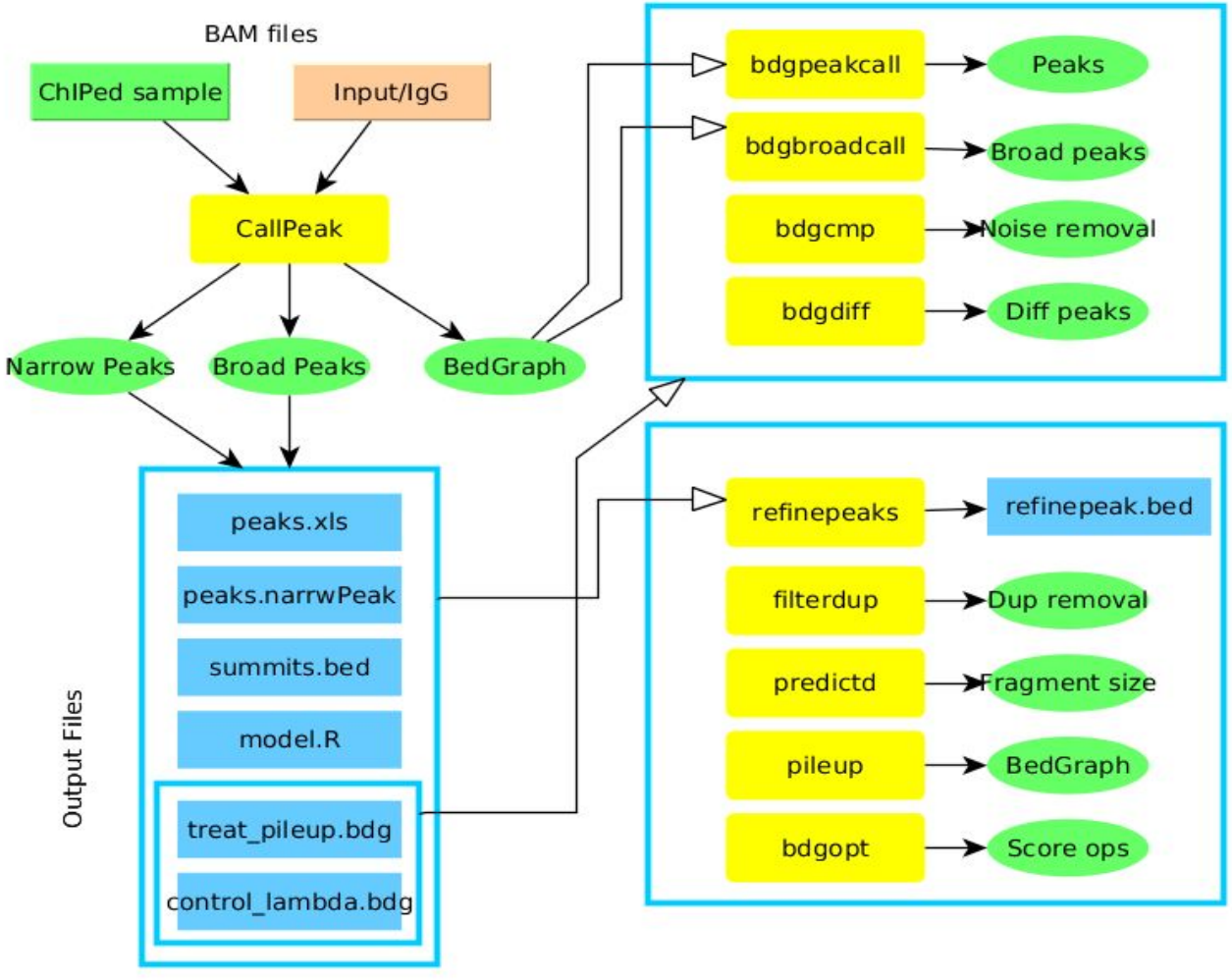
Encode Quality Metrics

Assay	Cell	Target	Treatment	Identifier	N_uniq map reads	MACS FDR 0.01	Self Cons IDR 0.02	Rep Cons IDR 0.01	SPOT	PBC	NSC	RSC	Under seq	Diff rep	Manual low S/N	Auto low S/N
TF-ChIP-seq	A549	CTCF	DEX_100nM	wgEncodeHaibTfbsA549CtcfPcr1xDexaAlnRep1	24,281,189	38,537	45,841	30,324	0.2361	0.71	2.79	2.19	0	0	0	0
TF-ChIP-seq	A549	CTCF	DEX_100nM	wgEncodeHaibTfbsA549CtcfPcr1xDexaAlnRep2	15,453,361	96,884	39,091	30,324	0.1249	0.41	1.84	2.31	0	1	0	0
TF-ChIP-seq	A549	GR	DEX_100nM	wgEncodeHaibTfbsA549GrPcr2xDexaAlnRep1	16,608,102	9,921	12,613	8,283	0.0754	0.91	1.38	1.21	0	1	0	0
TF-ChIP-seq	A549	GR	DEX_100nM	wgEncodeHaibTfbsA549GrPcr2xDexaAlnRep2	28,467,922	8,683	12,880	8,283	0.0723	0.44	1.42	1	0	0	0	0
TF-ChIP-seq	A549	POL2	DEX_100nM	wgEncodeHaibTfbsA549Pol2Pcr2xDexaAlnRep1	19,005,470	12,689	24,395	21,463	0.6166	0.86	2.99	1.32	0	0	0	0
TF-ChIP-seq	A549	POL2	DEX_100nM	wgEncodeHaibTfbsA549Pol2Pcr2xDexaAlnRep2	23,115,884	14,816	28,503	21,463	0.5388	0.86	2.81	1.12	0	0	0	0
TF-ChIP-seq	A549	USF1	DEX_100nM	wgEncodeHaibTfbsA549Usf1Pcr1xDexaAlnRep1	22,289,881	2,631	16,330	8,917	0.0791	0.87	1.28	1.86	0	0	0	0
TF-ChIP-seq	A549	USF1	DEX_100nM	wgEncodeHaibTfbsA549Usf1Pcr1xDexaAlnRep2	12,364,820	3,028	7,659	8,917	0.0517	0.82	1.44	1.9	0	0	0	0
TF-ChIP-seq	A549	GR	DEX_500pM	wgEncodeHaibTfbsA549GrPcr1xDexdAlnRep1	19,646,503	25,233	1,312	1,226	0.0105	0.96	1.05	0.56	0	0	1	1
TF-ChIP-seq	A549	GR	DEX_500pM	wgEncodeHaibTfbsA549GrPcr1xDexdAlnRep2	15,095,316	123,828	1,218	1,226	0.0109	0.94	1.06	0.5	0	0	1	1
TF-ChIP-seq	A549	GR	DEX_50nM	wgEncodeHaibTfbsA549GrPcr1xDexbAlnRep1	19,291,260	57,488	23,821	25,096	0.1289	0.96	1.55	1.42	0	0	0	0
TF-ChIP-seq	A549	GR	DEX_50nM	wgEncodeHaibTfbsA549GrPcr1xDexbAlnRep2	16,754,796	71,917	22,601	25,096	0.1426	0.95	1.64	1.61	0	0	0	0
TF-ChIP-seq	A549	GR	DEX_5nM	wgEncodeHaibTfbsA549GrPcr1xDexcAlnRep1	20,120,740	19,331	8,573	10,348	0.0343	0.98	1.10	0.89	0	1	1	0
TF-ChIP-seq	A549	GR	DEX_5nM	wgEncodeHaibTfbsA549GrPcr1xDexcAlnRep2	20,559,786	31,539	13,796	10,348	0.0641	0.96	1.23	1.17	0	0	0	0
TF-ChIP-seq	A549	CTCF	EtOH_0.02p	wgEncodeHaibTfbsA549CtcfPcr1xEtoh02AlnRep1	22,672,467	31,983	37,735	33,511	0.1601	0.75	1.78	2.67	0	0	0	0
TF-ChIP-seq	A549	CTCF	EtOH_0.02p	wgEncodeHaibTfbsA549CtcfPcr1xEtoh02AlnRep2	14,351,615	236,390	49,814	33,511	0.2040	0.42	3.21	2.55	0	0	0	0
TF-ChIP-seq	A549	POL2	EtOH_0.02p	wgEncodeHaibTfbsA549Pol2Pcr2xEtoh02AlnRep1	17,136,347	17,929	29,121	28,130	0.5602	0.9	2.89	1.19	0	0	0	0
TF-ChIP-seq	A549	POL2	EtOH_0.02p	wgEncodeHaibTfbsA549Pol2Pcr2xEtoh02AlnRep2	19,201,309	16,879	34,156	28,130	0.5687	0.82	3.09	1.12	0	0	0	0
TF-ChIP-seq	A549	USF1	EtOH_0.02p	wgEncodeHaibTfbsA549Usf1Pcr1xEtoh02AlnRep1	16,241,779	7,936	11,349	10,368	0.0648	0.95	1.38	2.02	0	0	0	0
TF-ChIP-seq	A549	USF1	EtOH_0.02p	wgEncodeHaibTfbsA549Usf1Pcr1xEtoh02AlnRep2	13,242,129	11,812	11,204	10,368	0.0793	0.85	1.72	1.99	0	0	0	0
TF-ChIP-seq	AG04449	CTCF	None	wgEncodeUwTfbsAg04449CtcfStdAlnRep1	9,952,444	97,323	62,334	44,965	0.5513	0.85	11.97	2.11	0	0	0	0
TF-ChIP-seq	AG04449	CTCF	None	wgEncodeUwTfbsAg04449CtcfStdAlnRep2	23,572,200	42,477	42,096	44,965	0.2187	0.94	2.68	1.61	0	0	0	0
TF-ChIP-seq	AG04450	CTCF	None	wgEncodeUwTfbsAg04450CtcfStdAlnRep1	21,170,101	44,837	43,626		0.2450	0.9	2.62	1.73	0	0	0	0
TF-ChIP-seq	AG09309	CTCF	None	wgEncodeUwTfbsAg09309CtcfStdAlnRep1	14,311,099	37,977	35,062	35,451	0.3278	0.89	3.93	1.8	0	0	0	0
TF-ChIP-seq	AG09309	CTCF	None	wgEncodeUwTfbsAg09309CtcfStdAlnRep2	10,263,622	34,845	31,992	35,451	0.1768	0.95	2.31	1.52	0	0	0	0
TF-ChIP-seq	AG09319	CTCF	None	wgEncodeUwTfbsAg09319CtcfStdAlnRep1	22,451,182	53,232	42,690	34,945	0.3807	0.8	4.32	1.67	0	0	0	0
TF-ChIP-seq	AG09319	CTCF	None	wgEncodeUwTfbsAg09319CtcfStdAlnRep2	25,700,109	45,377	38,947	34,945	0.2775	0.87	2.97	1.73	0	0	0	0
TF-ChIP-seq	AG10803	CTCF	None	wgEncodeUwTfbsAg10803CtcfStdAlnRep1	26,964,677	39,701	38,287	39,892	0.2254	0.88	2.36	1.63	0	0	0	0

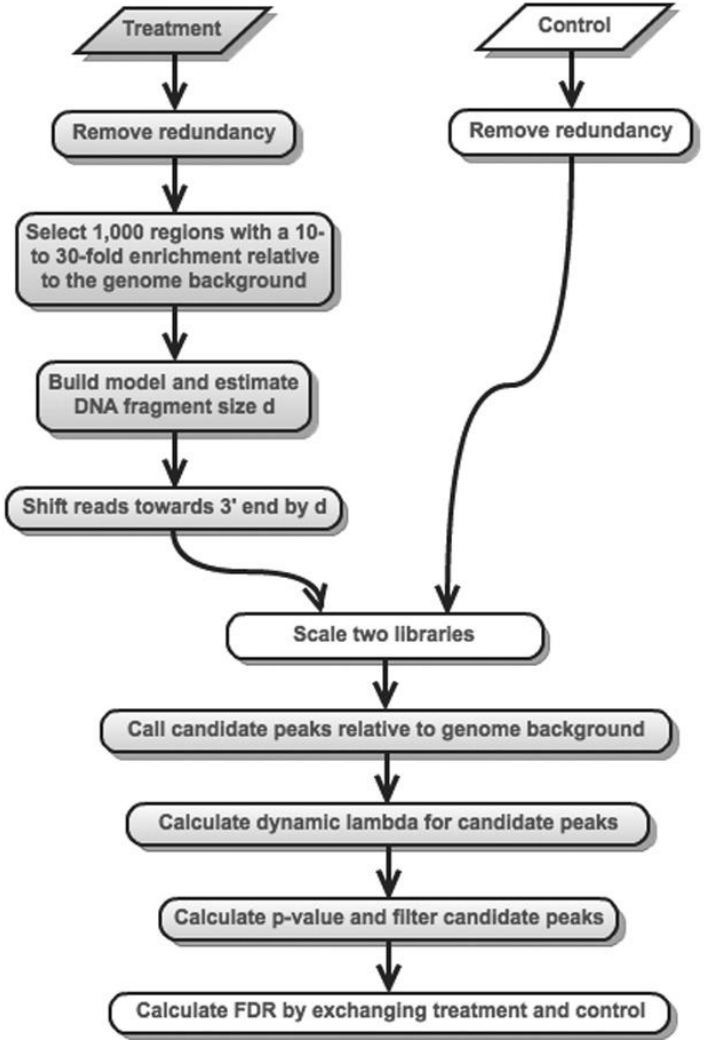
MACS2

- Most widely used peak caller.
- identifies genome-wide locations of TF binding, histone modification or NFRs from ChIP-seq or ATAC-seq data.
- Can be used without a **control** (Input -samples of sonicated chromatin OR IgG - nonspecific antibody) - Not Recommended for ChIP-seq!
- Controls eliminate **bias** due to GC content, mappability, DNA repeats or CNVs.
- Can call narrow and broad peaks.
- Many settings for optimizing results.

MACS2


BAM files
ChIPed sample
Input/IgG
CallPeak
Narrow Peaks
Broad Peaks
BedGraph
bdgpeakcall
Peaks
bdgbroadcall
Broad peaks
bdgcmp
Noise removal
bdgdiff
Diff peaks
Output Files
peaks.xls
peaks.narrwPeak
summits.bed
model.R
treat_pileup.bdg
control_lambda.bdg
refinepeaks
refinepeak.bed
filterdup
Dup removal
predictd
Fragment size
pileup
BedGraph
bdgopt
Score ops

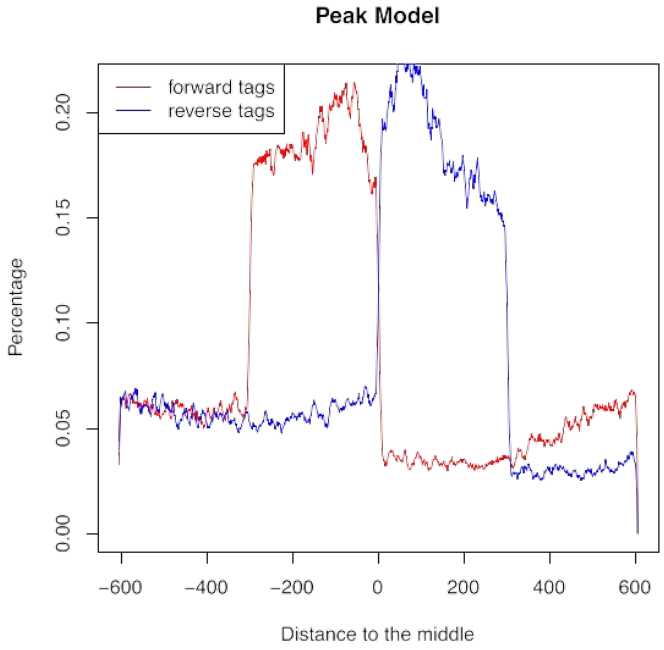
MACS1.4


Treatment
Control
Remove redundancy
Remove redundancy
Select 1,000 regions with a 10- to 30-fold enrichment relative to the genome background
Build model and estimate DNA fragment size d
Shift reads towards 3' end by d
Scale two libraries
Call candidate peaks relative to genome background
Calculate dynamic lambda for candidate peaks
Calculate p-value and filter candidate peaks
Calculate FDR by exchanging treatment and control

Peak calling with MACS2

★ **Step 1: Estimate fragment length *d* and adjust read position**

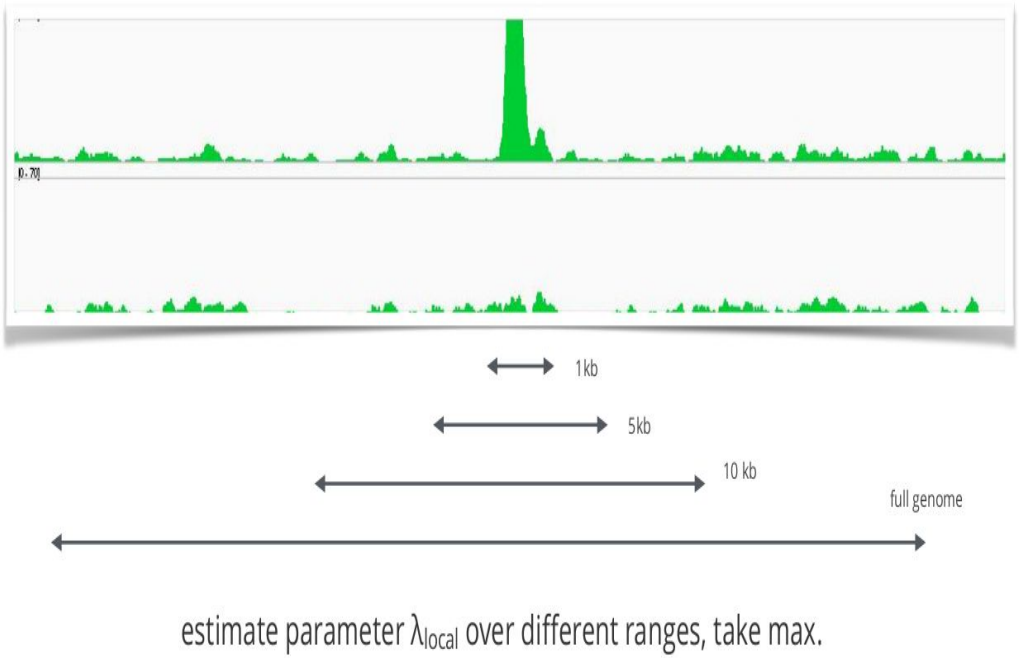
- Slide a window of length 2 x **bw** bandwidth (half of estimated sonication size) across genome.
- Retain windows with > **MFOLD** (*fold-enrichment of treatment / background*)
- Compute the average +/- strand specific read-densities for these bins.



MACS2

★ **Step 2: Identify local noise**

- slide a window of size 2*d across treatment and input
- estimate λ_{local} parameter of Poisson distribution

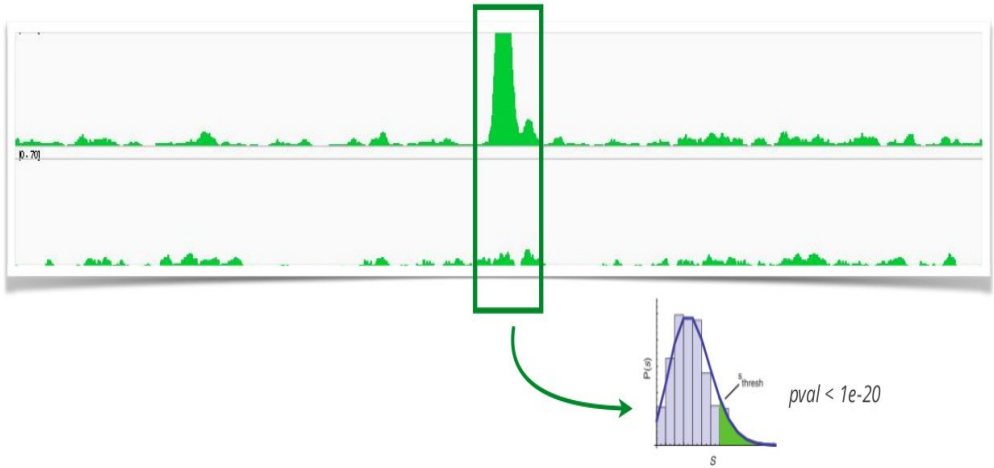


modified from Carl Herrmann

MACS2

★ **Step 3: Identify enriched (peak) regions**

- determine regions with p-value < PVALUE
- determine summit position within enriched regions as max density



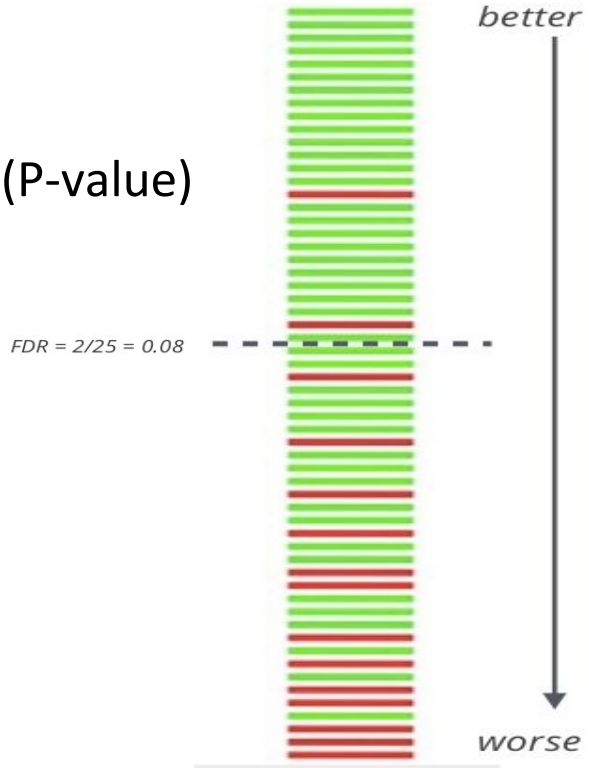
estimate parameter λ_{local} over different ranges, take max.

modified from Carl Herrmann

MACS2

- ★ Step 4: Estimate FDR
  - positive peaks (P-values)
  - swap treatment and input, call negative peaks (P-value)

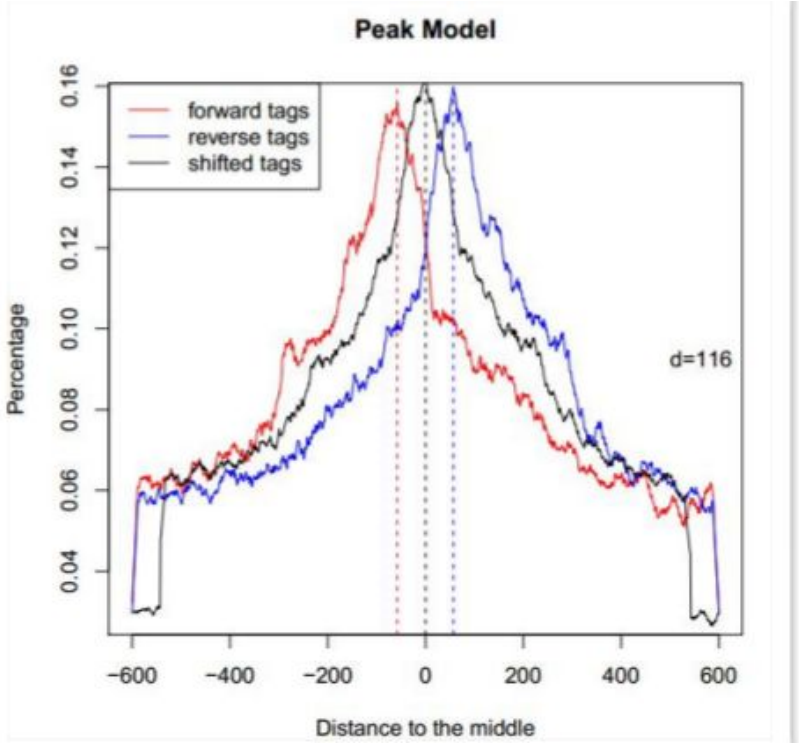
$$FDR = \frac{\#\text{ negative peaks with pval} < p}{\#\text{ positive peaks with pval} < p}$$



modified from Carl Herrmann

Reads to Peaks

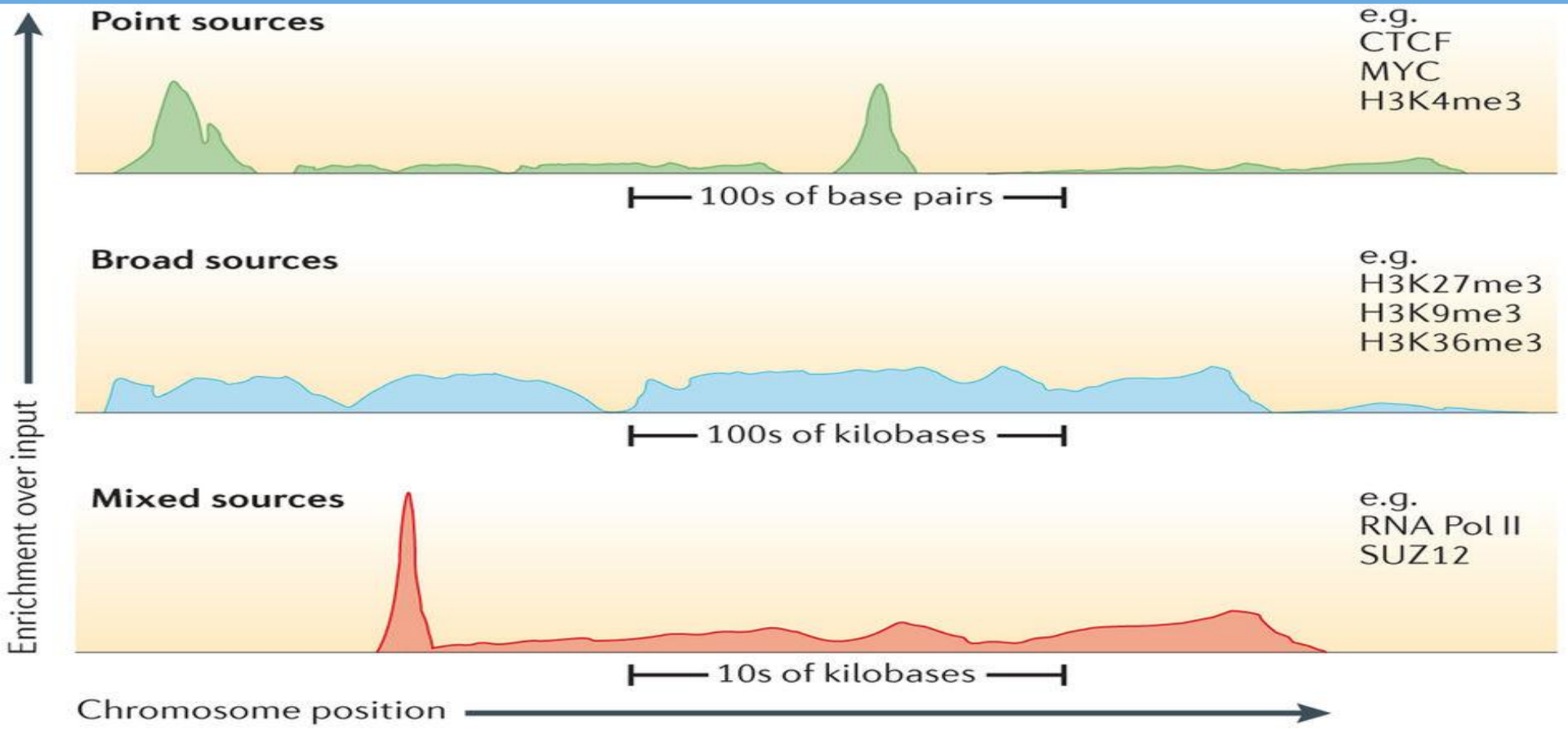
- +ive and -ive strand reads do not represent true binding sites
- Fragment length *d* can be detected experimentally or estimated from strand asymmetry in data
- Reads from both strands can be **extended** to the length of *d* **OR**
- Reads can be **shifted** towards 3' by *d/2*



Narrow, Broad and Mixed Peaks

- Different data types have different peak shapes. Use appropriate peak callers or domain detectors. Same TF may have different peak shapes reflecting differences in biological conditions. Replicates should have similar binding patterns.
- Most TF peaks are narrow, with particularly sharp peaks from ChIP-exo data.
- ChIP-seq peaks from epigenomic data can be narrow, broad or gapped. Histone marks such as H3K9me^3 or H3K27me^3 are broad while others such as H3K4me^3 and proteins such as CTCF are narrow. Other DNA binding proteins such as HP1 , Lamins (Lamin A or B), HMGA etc. form broad peaks or domains.
- PolII peaks can be narrow or broad depending on whether its detecting transcription initiation at the TSS or propagation along the gene body.
- ATAC-seq data representing nucleosome free regions (NFRs) can be narrow or broad depending on the properties of regulatory regions underlying them.

Peak Shapes



Sims et al., 2014 Nat Rev Genet.

Nature Reviews | Genetics

Broad peak and Domain callers

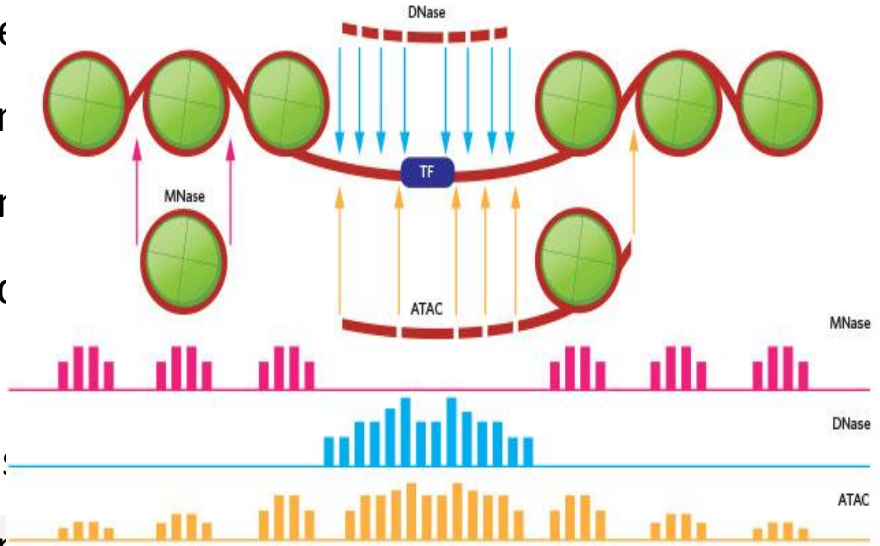
- ***MACS2*:** *macs2 callpeak --broad*

- ***Epic*:** Useful for finding medium or diffusely enriched domains in chip-seq data. Epic is an improvement over the original SICER, by being faster, more memory efficient, multi core, and significantly easier to install and use.

- Others: ***Enriched Domain Detector (EDD), RSEG, BroadPeak, PeakRanger (CCAT)***

ATAC-seq settings

- If using paired end reads use "--format BAMPE" to let MACS2 pileup the whole fragments in general. If you want to focus or looking for where the 'cutting sites' are, ther "--nomodel --shift -100 --extsize 200" should work.
- Since the DNA wrapped on a nucleosome is about 147bp, for single nucleosome detection use " --nomodel --shift -37 --extsize 73".



BASED ON *EPIGENETICS CHROMATIN*, 7:33, 2014.
Scientist, Volume 30 Issue 1 | January 2016

References

- Landt *et al.,* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012, 22:1813-1831. PMID: 22955991.
- Wilbanks *et al.,* Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One. 2010, Jul 8;5(7):e11471. PMID: 20628599
- Nakato *et al.,* Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinform. 2017 Mar 1;18(2):279-290. PMID: 26979602
- Buenrostro *et al.,* ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015 Jan 5;109:21.29.1-9. PMID: 25559105
- Sims