Peak-calling for ChIP-seq

Izzy Newsham
MRC Cancer Unit
University of Cambridge

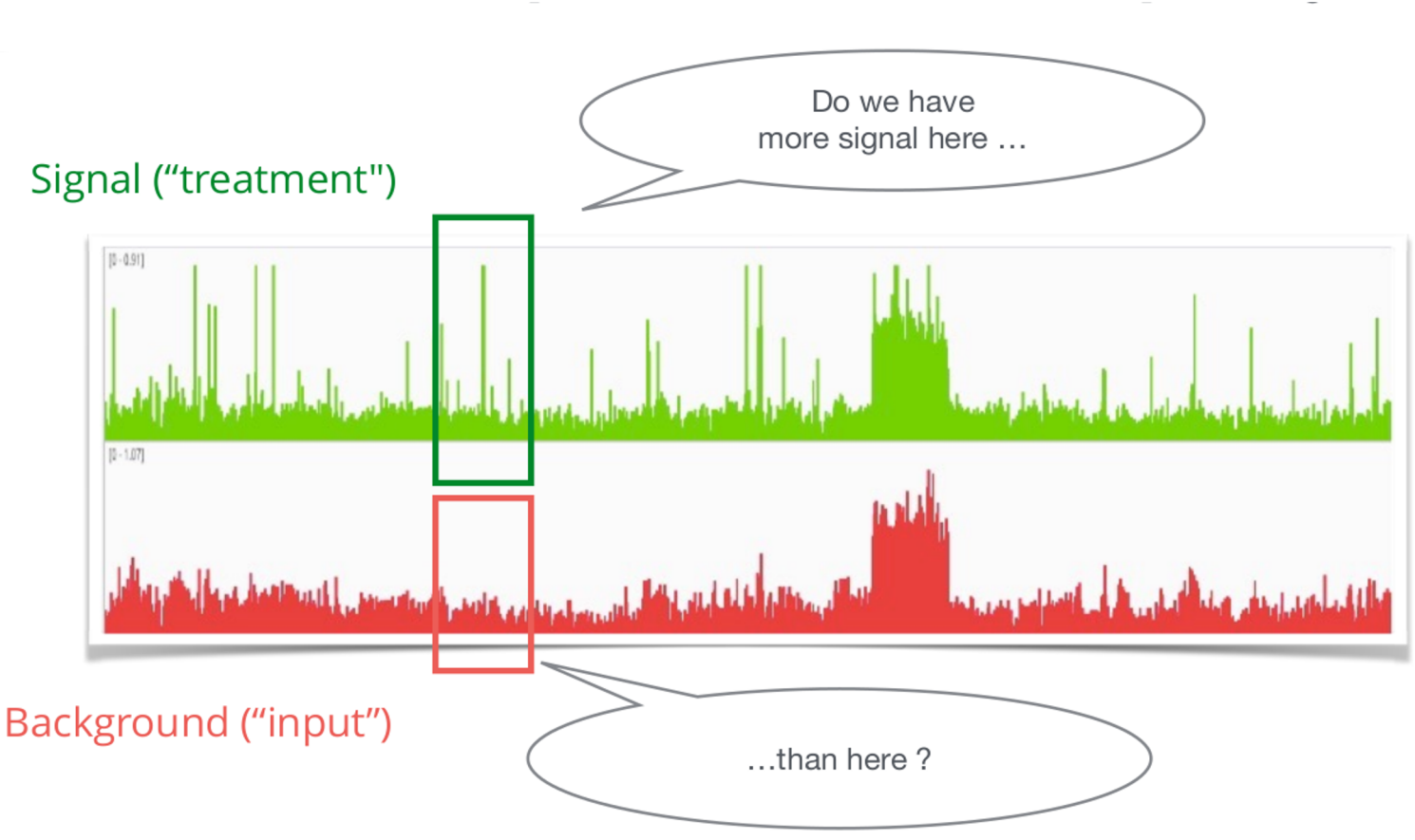
CRUK Bioinformatics Summer School 2021
27th July 2021

Slides adapted from Shamith Samarajiwa

Overview

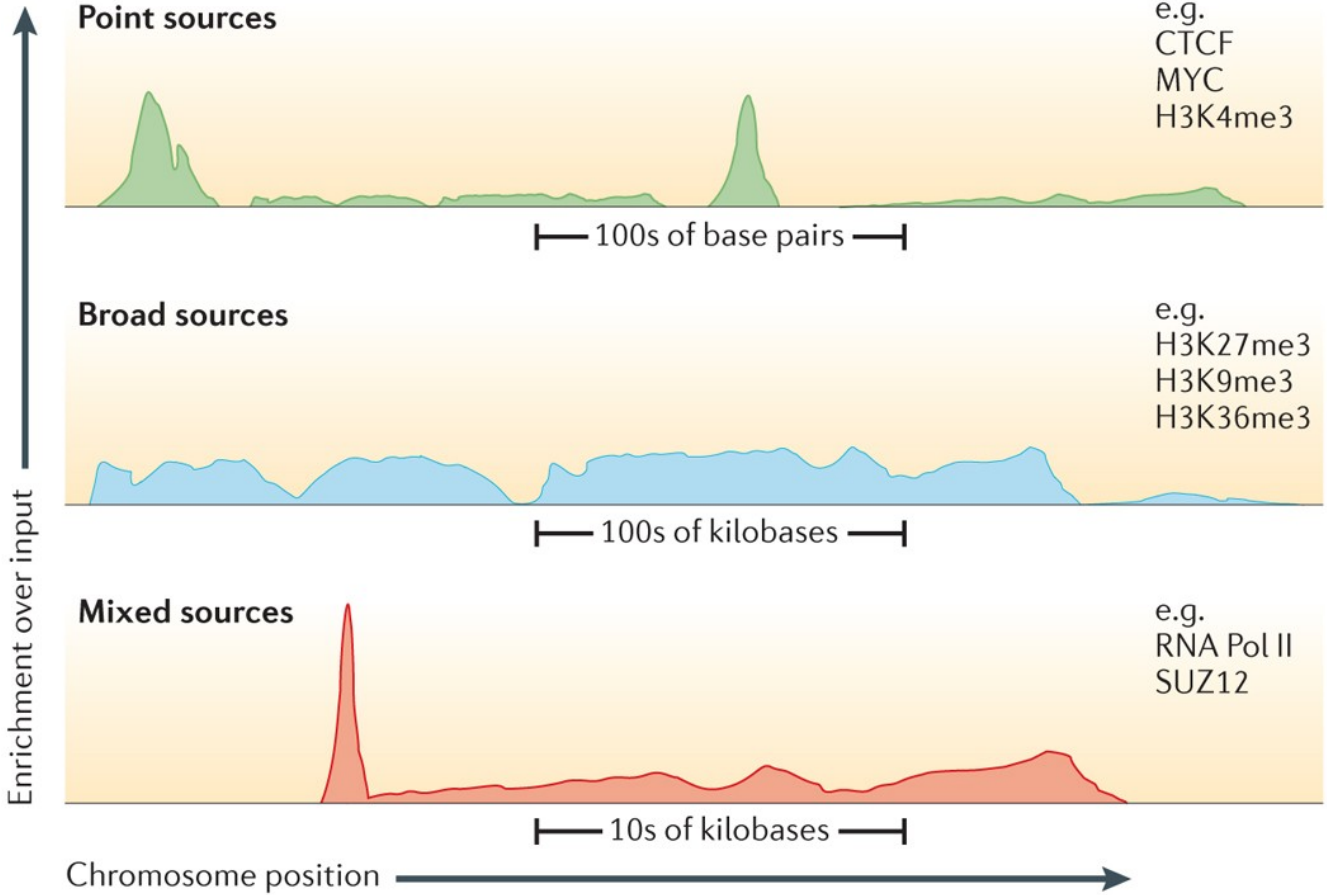
- Peak types
- ENCODE Project
- Software packages
- Important concepts for peak calling
  - Duplicates
  - Identifying the peak locations
- MACS2
  - Steps of MACS2 peak calling

Peaks: Signal to Noise



modified from Carl Herrmann

Peak Shapes



Sims et al., 2014 Nat Rev Genet.

Nature Reviews | Genetics

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

Narrow, Broad and Mixed Peaks

Narrow:

- Most TF peaks are narrow
- Particularly sharp peaks from ChIP-exo data
- Some histone marks, such as H3K4me3

ChIP-seq peaks from epigenomic data can be narrow or broad

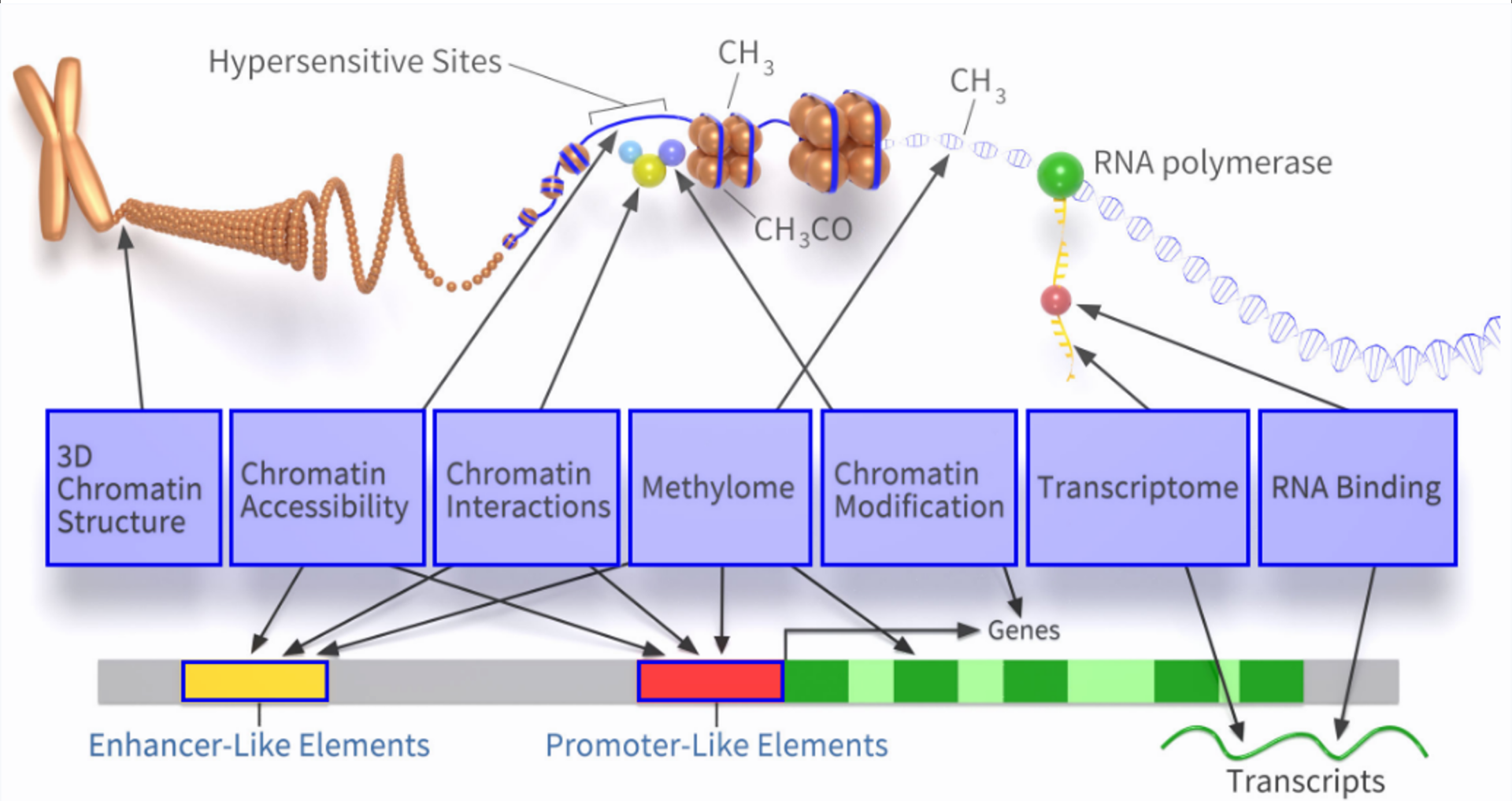
Broad:

- Histone marks such as H3K9me3 or H3K27me3
- DNA binding proteins such as HP1 , Lamins (Lamin A or B), HMGA

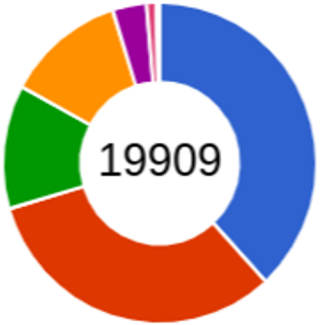
Mixed:

- RNA polymerase II peaks - depending on whether its detecting transcription initiation at the TSS or propagation along the gene body

ENCODE: Encyclopedia of DNA Elements

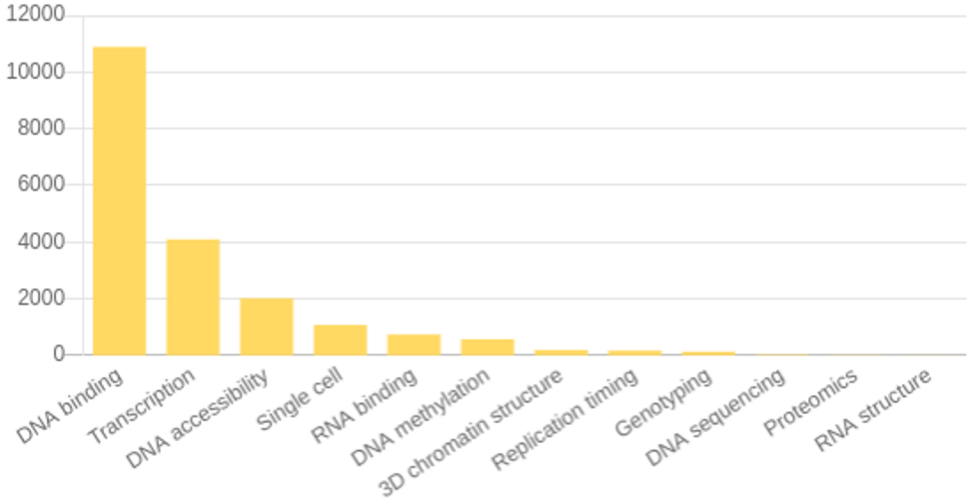


Biosample Type



- cell line
- tissue
- whole organisms
- primary cell
- in vitro differentiated cells
- cell-free sample
- organoid
- technical sample

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

Peak Calling Software

MACS2 *(MACS3 soon)*	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.

Broad peak and Domain callers

- ***MACS2*** default setting calls narrow peaks

  For broad peaks: *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.

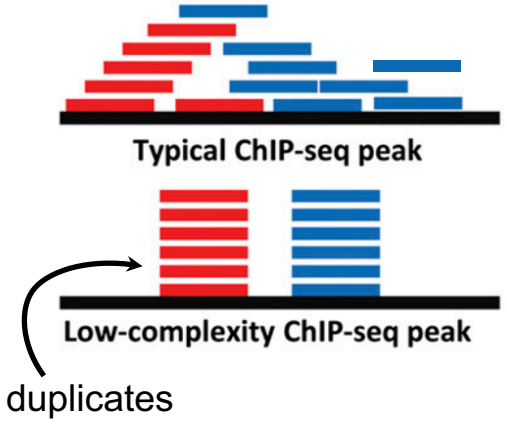
- Others: ***Enriched Domain Detector (EDD), RSEG, BroadPeak****,* ***PeakRanger (CCAT)***

Important concepts

- Duplicates in ChIP-seq
- Identifying the peak locations

Duplicate Removal

- Duplicates are reads or pairs of reads that have **identical or near-identical sequences** (due to sequencing errors) and map to the **same genomic position** and strand



Modified from: Landt *et al.*, ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.

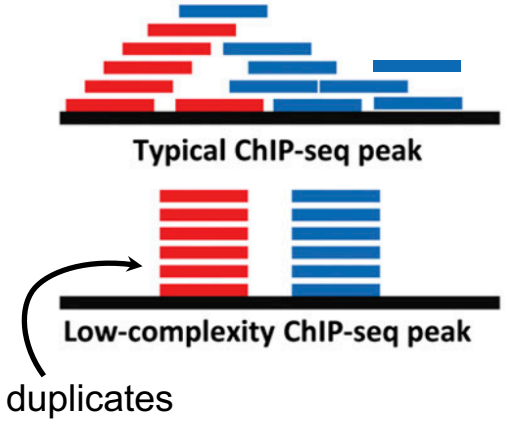
Duplicate Removal

- Duplicates are reads or pairs of reads that have **identical or near-identical sequences** (due to sequencing errors) and map to the **same genomic position** and strand

Two duplicate types:

- During library preparation, ChIP DNA undergoes a PCR amplification step
- **Increased sequencing depth**, **low immunoprecipitation efficiency** or **insufficient amounts of starting material**, can contribute to PCR duplicates formation
- These types of duplicates **need to be filtered out**

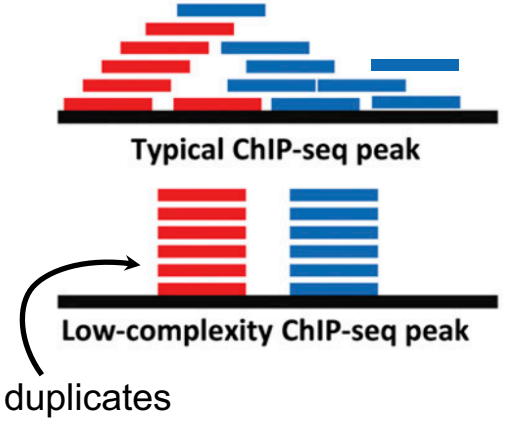
- However **natural duplicates** arise from sequencing of independent DNA fragments derived from the same genomic locations
- These **should not be removed** as they are part of the true signal



Modified from: Landt *et al.*, ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.

Duplicate Removal

- Examination of read alignment (BAM files) in a genome browser can help identify PCR duplicates
- Tian et al. suggest most duplicates in (narrow) peaks are natural duplicates, and **removing duplicates results in loss of true signal**



Modified from: Landt *et al.*, ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.

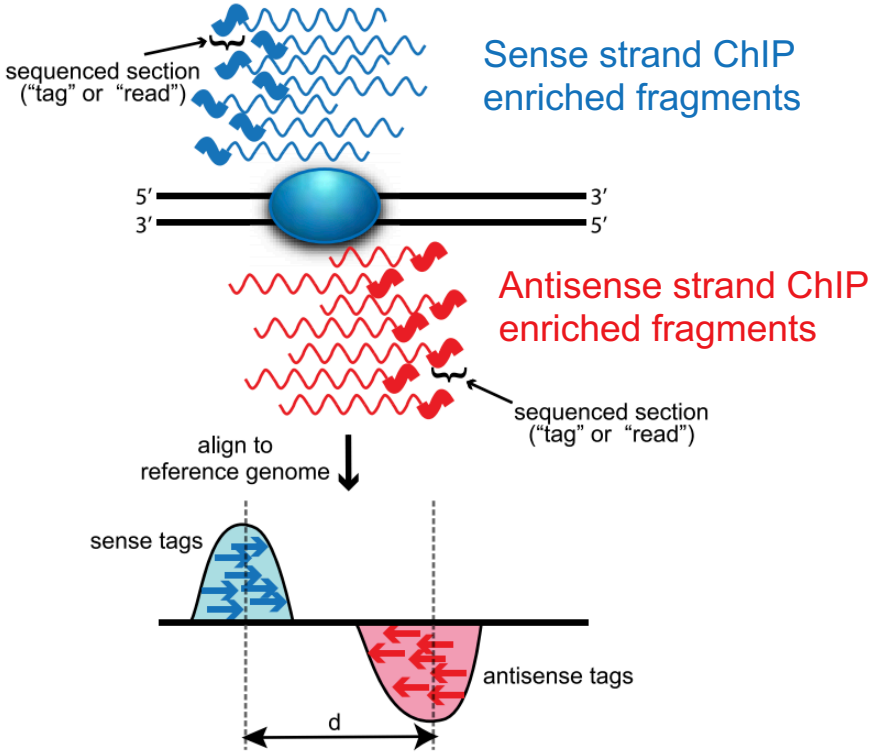
> PLoS One. 2019 Apr 3;14(4):e0214723. doi: 10.1371/journal.pone.0214723. eCollection 2019.

Identification of factors associated with duplicate rate in ChIP-seq data

Shulan Tian [1], Shuxia Peng [1], Michael Kalmbach [2], Krutika S Gaonkar [1], Aditya Bhagwate [1], Wei Ding [3], Jeanette Eckel-Passow [1], Huihuang Yan [1], Susan L Slager [1]

Identifying true peak locations

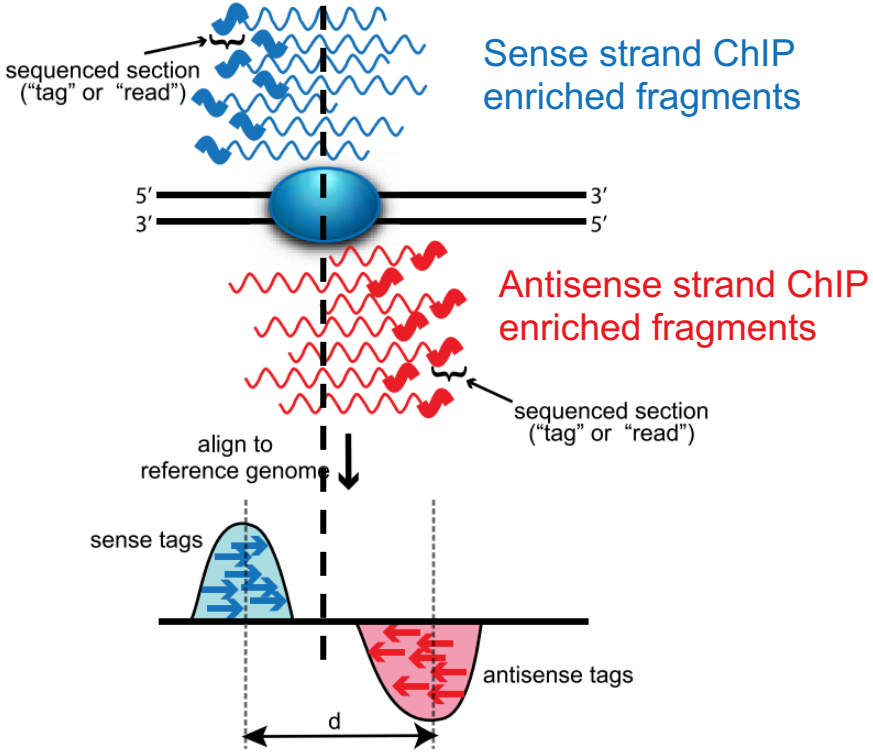
Reads display strand-dependent bimodality:



Modifed from: Wilbanks and Facciotti, Evaluation of Algorithm Performance in ChIP-Seq Peak Detection

Identifying true peak locations

Reads display strand-dependent bimodality:

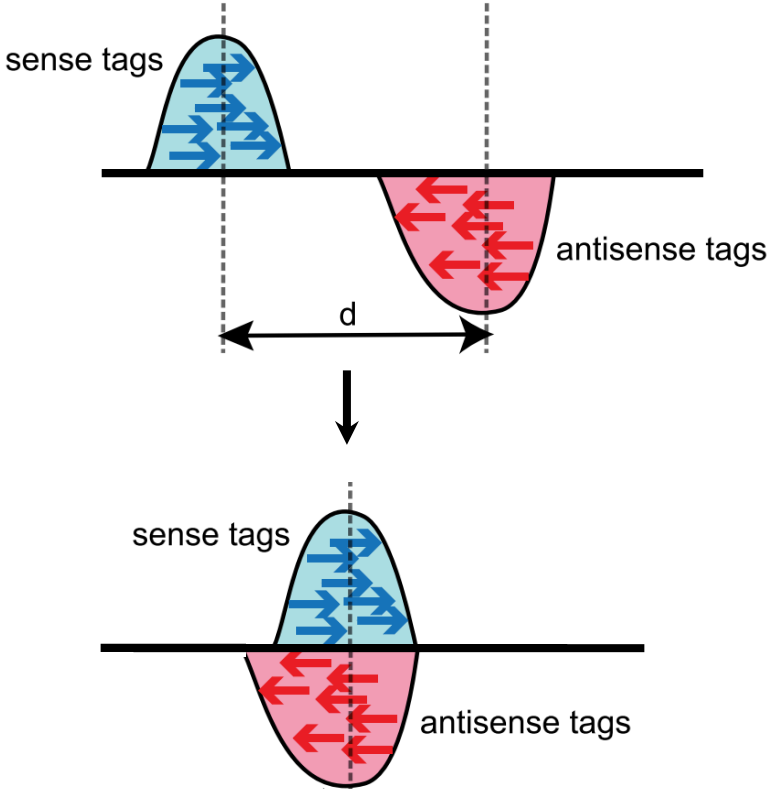


Modifed from: Wilbanks and Facciotti, Evaluation of Algorithm Performance in ChIP-Seq Peak Detection

Identifying true peak locations

- So we need to **shift the reads** so they all align at the true binding site

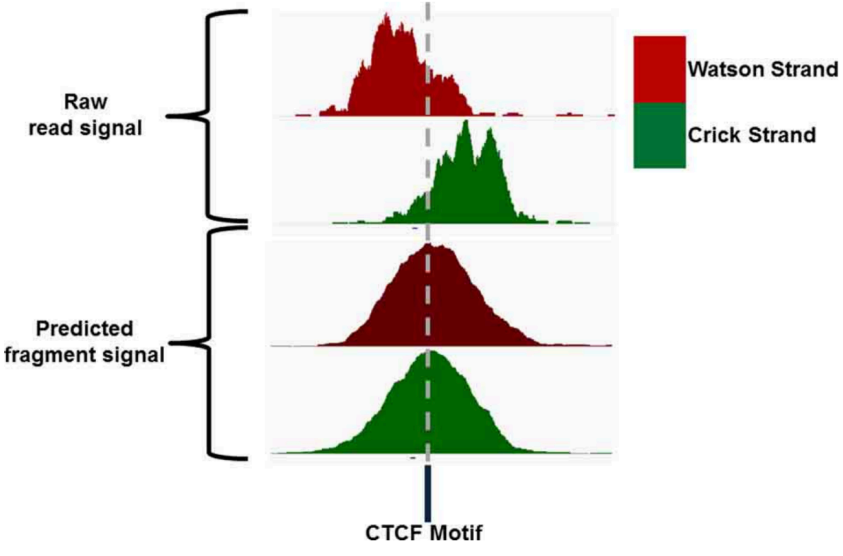
- In order to do this, we need to find the **fragment length, d**
- d can be detected **experimentally** or **estimated from the strand asymmetry** in the data
- The optimal size range of chromatin for ChIP-Seq analysis should be between 150 and 300 bp



Identifying true peak locations

- So we need to **shift the reads** so they all align at the true binding site

- In order to do this, we need to find the **fragment length, d**
- d can be detected **experimentally** or **estimated from the strand asymmetry** in the data
- The optimal size range of chromatin for ChIP-Seq analysis should be between 150 and 300 bp



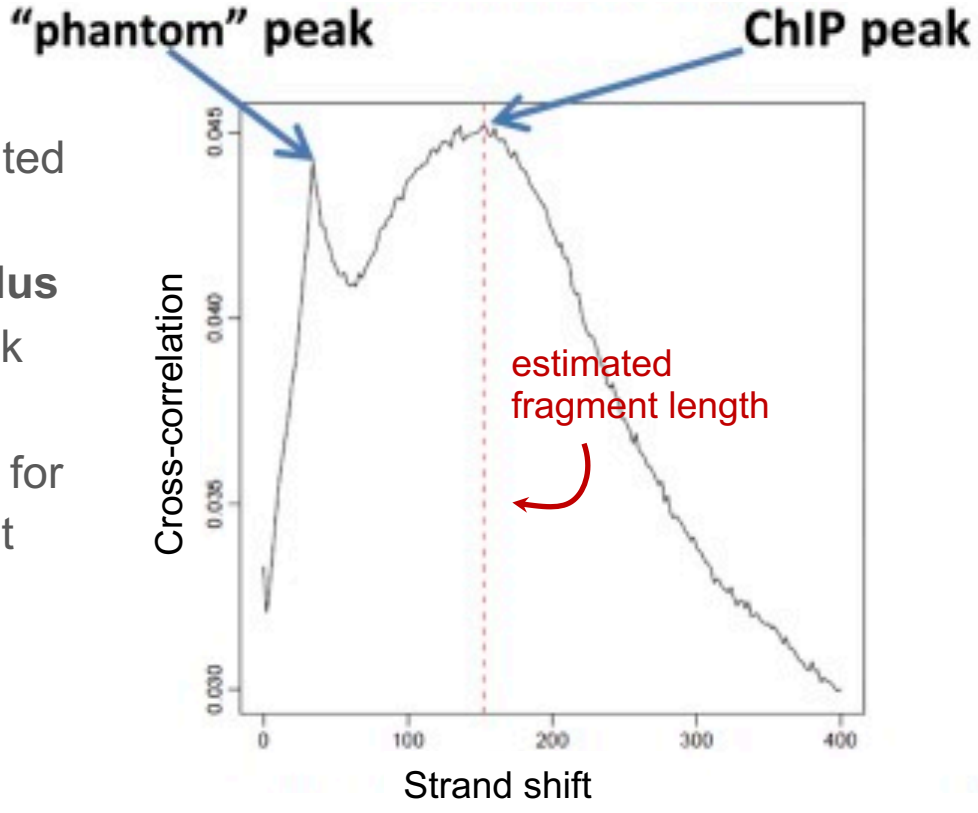
Carroll, Liang, Salama, Stark and Santiago. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data

Identifying true peak locations

The cross-correlation plot

- The strand cross-correlation is computed as the **Pearson's linear correlation** between the **minus strand** and the **plus strand**, after shifting minus strand by k base pairs
- The result is a cross-correlation value for each shift value, that is plotted against each other to generate the cross-correlation plot

- It is an important **quality control plot**

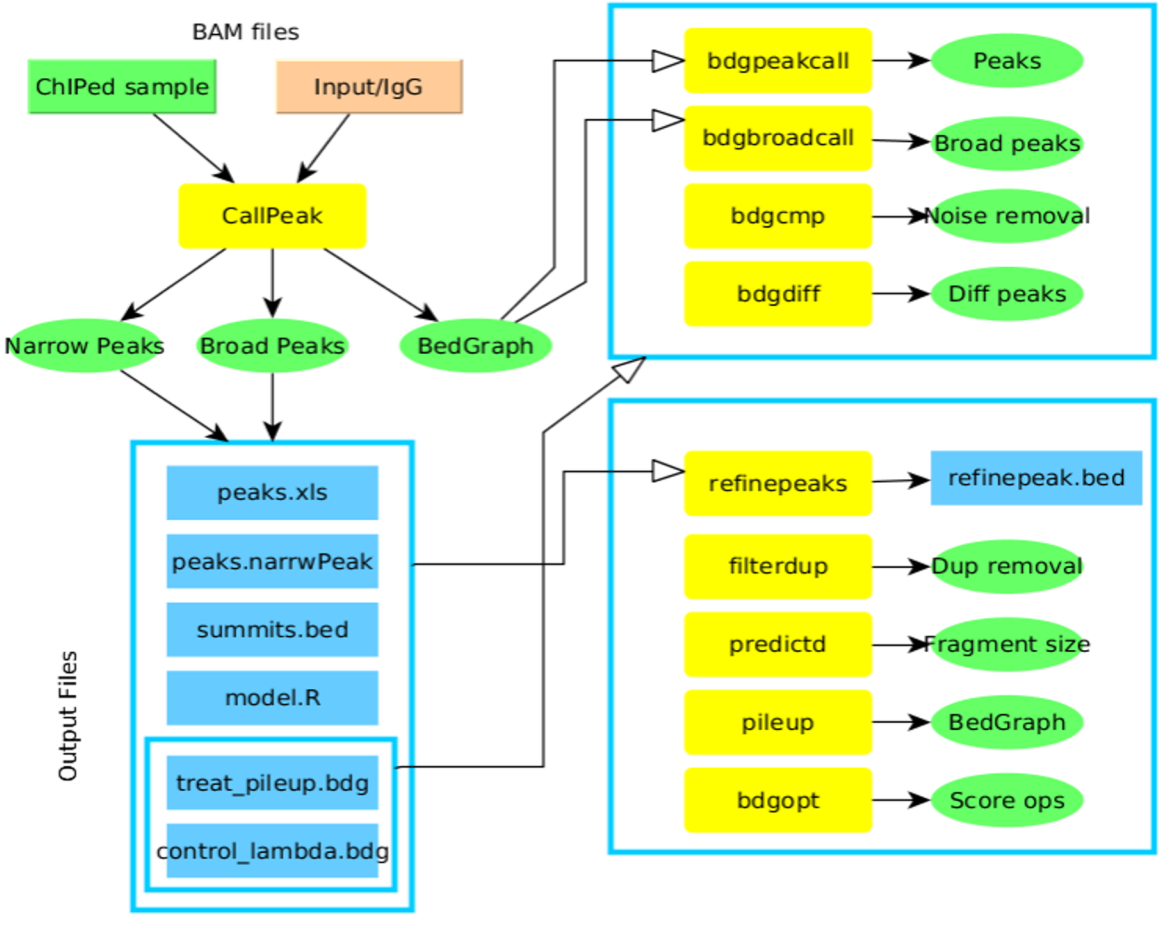


Modifed from: Landt *et al.*, ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.

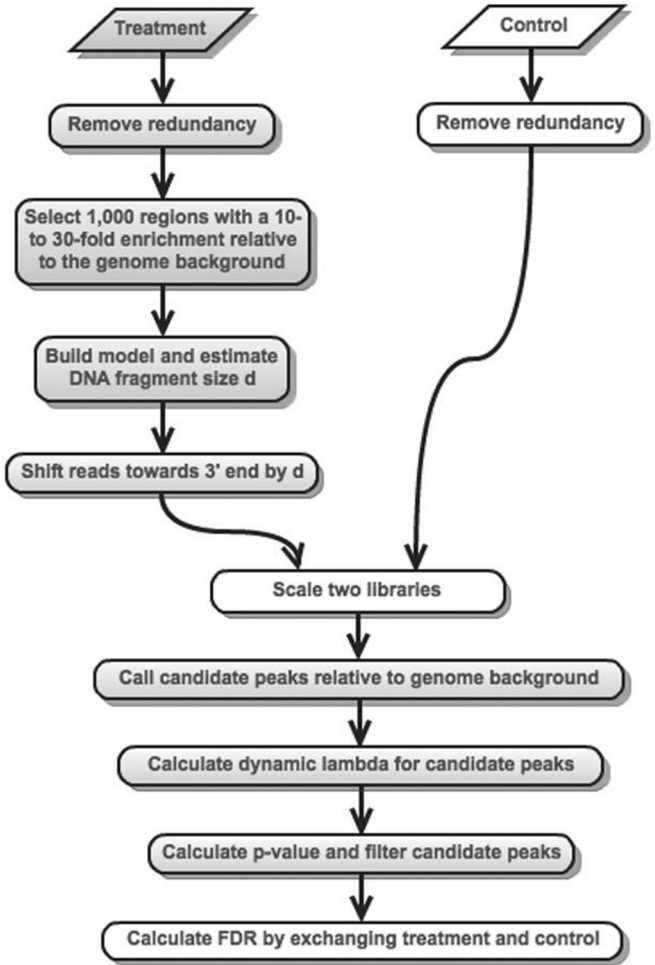
MACS2

- **Most widely used** peak caller (not the best!)
- Identifies genome-wide locations of TF binding, histone modification or NFRs from ChIP-seq or ATAC-seq data
- Can be used without a **control** but a control sample results in more accurate peaks
- **Controls bias** due to GC content, mappability, DNA repeats or CNVs
- Can call **narrow or broad** peaks
- Many settings for optimizing results
- MACS3 (alpha version is currently available)

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

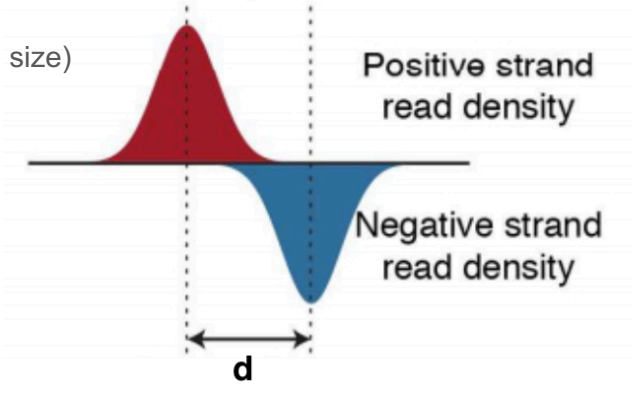

Treatment
Control
Remove redundancy
Remove redundancy
Select 1,000 regions with a 10- to 30-fold enrichment relative to the genome background
Build model and estimate DNA fragment size d
Shift reads towards 3' end by d
Scale two libraries
Call candidate peaks relative to genome background
Calculate dynamic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

(bandwidth = the sonication size)

- **Slide a window** of length 2*bandwidth across the genome
- For each window, calculate the fold-enrichment and **retain enriched windows** with **enrichment > MFOLD**
- **Sample 1000** of these windows
- Compute the **read-densities for both strands**. The distance between the peaks from each strand is d
- **Shift all reads** towards the 3' end by **d/2**



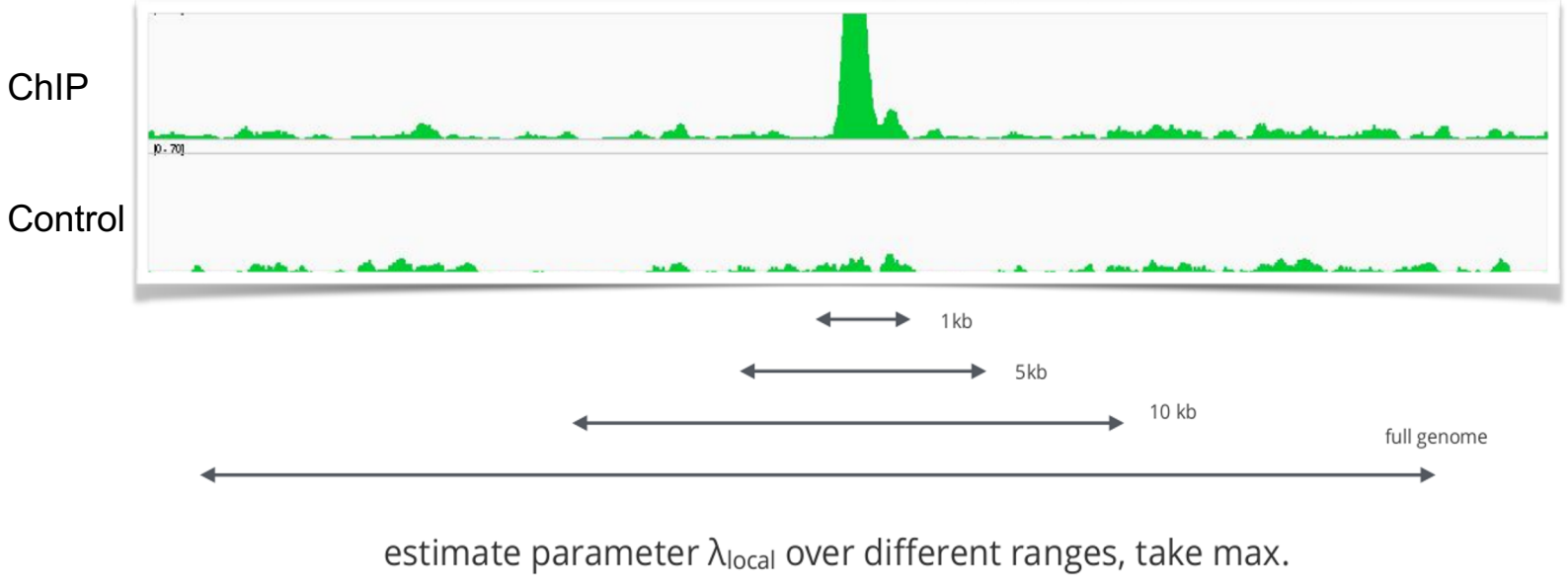
https://github.com/hbctraining/Intro-to-ChIPseq/blob/master/lessons/05_peak_calling_macs.md

Step 2: Identify local noise

- **Slide a window** of length 2*d across the genome
- For each window, model the read counts in the control sample as a **Poisson distribution**
  - Estimate the λ_{local} parameter of Poisson distribution:
  - $\lambda_{local} = \max(\lambda BG, \lambda 1k, \lambda 5k, \lambda 10k)$

MACS2

Step 2: Identify local noise

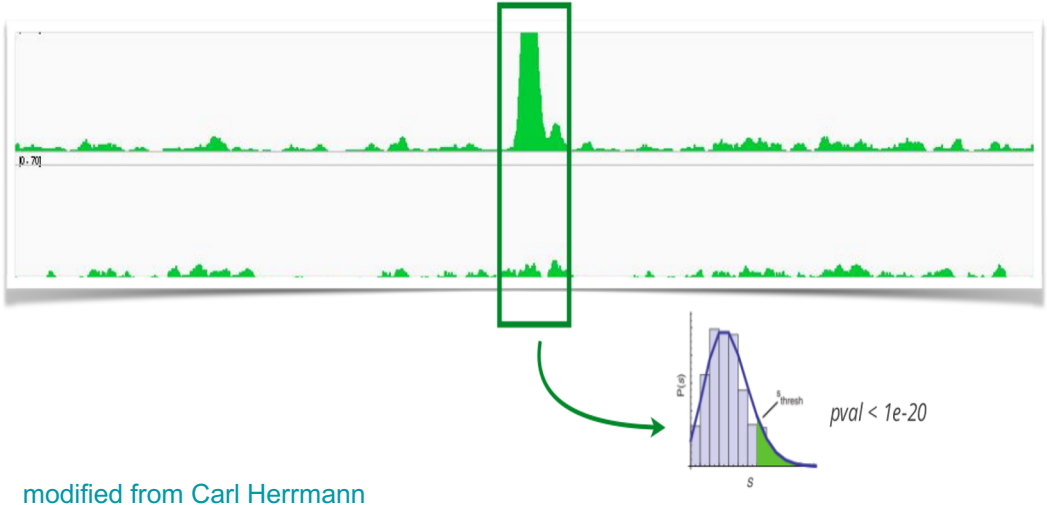


estimate parameter λ_{local} over different ranges, take max.

modified from Carl Herrmann

Step 3: Identify enriched (peak) regions

- Calculate a **p-value** to determine if the read counts in the ChIP sample follow this control distribution (with mean λ_{local}) or not
- Determine regions with p-value < PVALUE



modified from Carl Herrmann

Step 3: Identify enriched (peak) regions

- Calculate a **p-value** to determine if the read counts in the ChIP sample follow this control distribution (with mean λ_{local}) or not
- Determine regions with p-value < PVALUE
- **Merge overlapping** enriched regions
- Determine **summit position** - where the enriched region has the most fragments piled up
- Calculate the **fold-enrichment**
  - Ratio between the number of ChIP reads and λ_{local}

MACS2

Step 4: Estimate FDR

As each called peak is independent, we need to perform multiple testing correction

- Calculate p-values for negative peaks, by peak calling after swapping treatment and control

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

FDR = 2/25 = 0.08

Step 4: Estimate FDR

- Calculate p-values for negative peaks, by peak calling after swapping treatment and control

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

FDR = 2/25 = 0.08

In MACS2, this has been replaced by the **Benjamini-Hochberg** correction method

Quality control

There are various **quality metrics** and plots to check your ChIP-seq and peak calling has worked

An important metric: Irreproducible Discovery Rate (IDR)

- We expect to have **high consistency between replicates** for the most significant called peaks.
- IDR **measures consistency between replicates** in high-throughput experiments
- software: https://github.com/nboley/idr

More on quality metrics in the next lecture

References

- Sims et al., Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014
- Landt *et al.,* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012, 22:1813-1831. PMID: 22955991
- Tian et al. Identification of factors associated with duplicate rate in ChIP-seq data. PLOS One. 2019
- Wilbanks *et al.,* Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One. 2010, Jul 8;5(7):e11471. PMID: 20628599
- Carroll, Liang, Salama, Stark and Santiago. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Front Genet. 2014
- Zhang et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008
- https://github.com/hbctraining/Intro-to-ChIPseq/blob/master/lessons/05_peak_calling_macs.md
- Carl_Herrmann_ChIP-seq_slides