Computational Pipeline for ChIP-seq Data Analysis

Minghui Wang, Qi Sun
Bioinformatics Facility
Institute of Biotechnology

Purpose of workshop

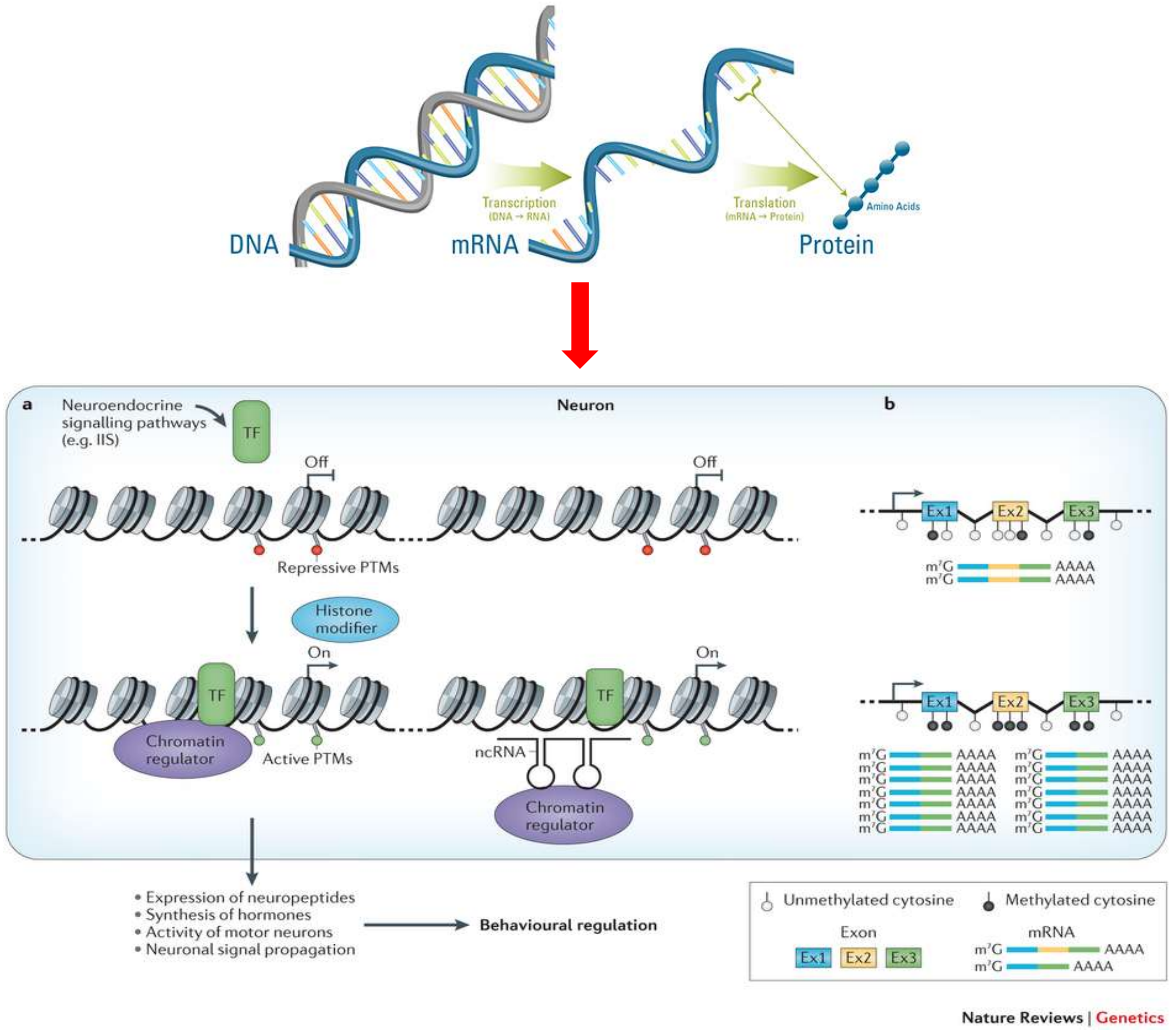
Know about basic consideration before performing ChIP-seq experiment

Introduce basic steps of analyzing ChIP-seq and widely used software

- mapping reads to genome
- identifying binding sites
- visualization of enriched regions
- discovering binding site motifs

…

The levels of regulators



-Epigenetic level

-Transcriptional factor

Hua Yan et al (2014). Nature reviews genetics

ChIP-seq

ChIP-seq Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing.

Goal Identify genome-wide binding sites of proteins of interesting

-Transcriptional factor/Histone marks/…

Research object of ChIP-seq

Transcription factors

Histone marks

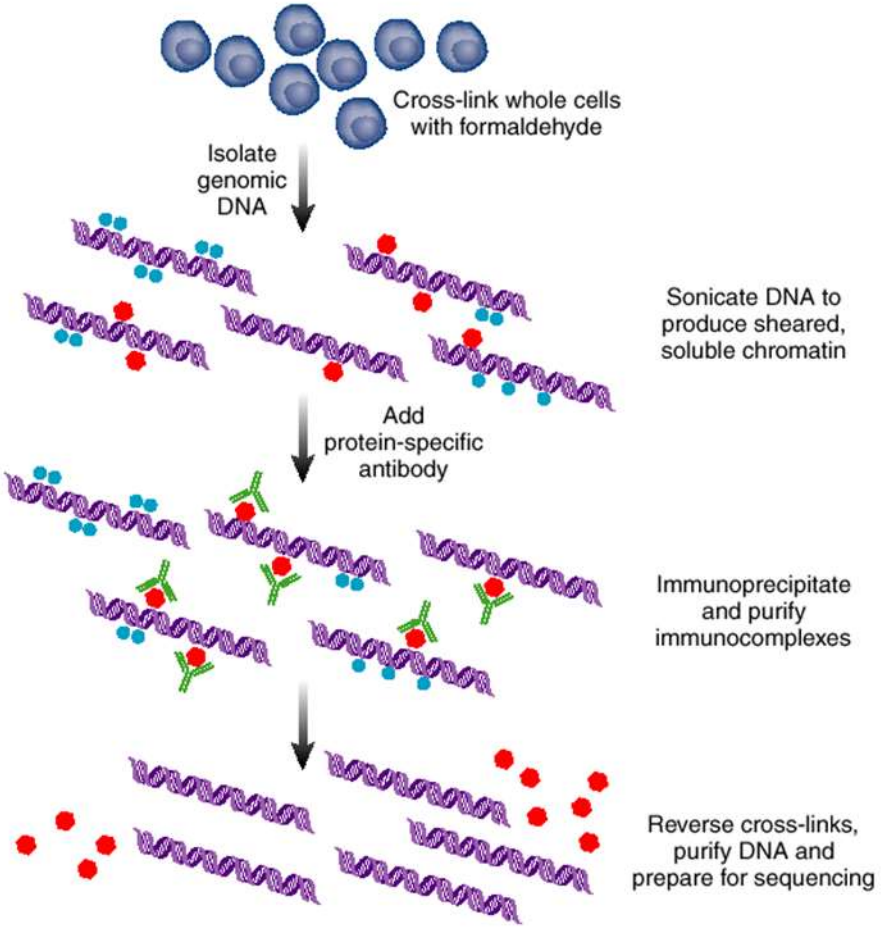
- H3K4me3
- H3K27me3

Nucleosomes

Rna polymerase

- RNA pol II

Major steps of ChIP-seq

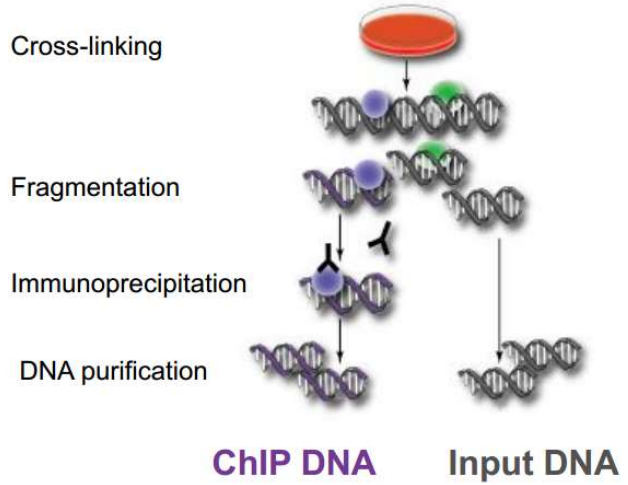


Elaine R Mardis ER (2007) Nature Methods 4:613-614

ChIP-Seq experimental design

Most experimental protocols involve a control sample that is processed the same way as the test sample except that no immunoprecipitation step (input) or no specific antibody (IgG)

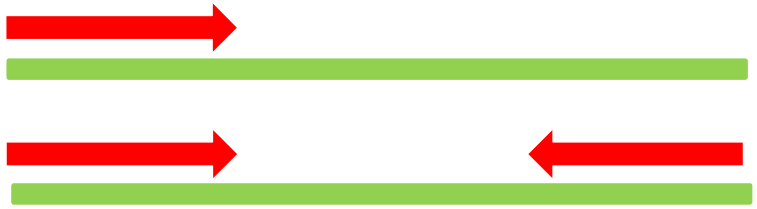
- GC sequencing bias
- amplification bias.
- mapping artifacts
- non-specific pull-down



ChIP-Seq experimental design

Single-end

Paired-end



Sequence length 36~100 bp

- increase “mappability” of reads specially in repetitive regions.
- double sequencing cost.

ChIP-Seq experimental design

Mammalian cells

- sharp peaks (TFs) 10 million uniquely mapped reads
- broad peaks 12-20 million uniquely mapped reads

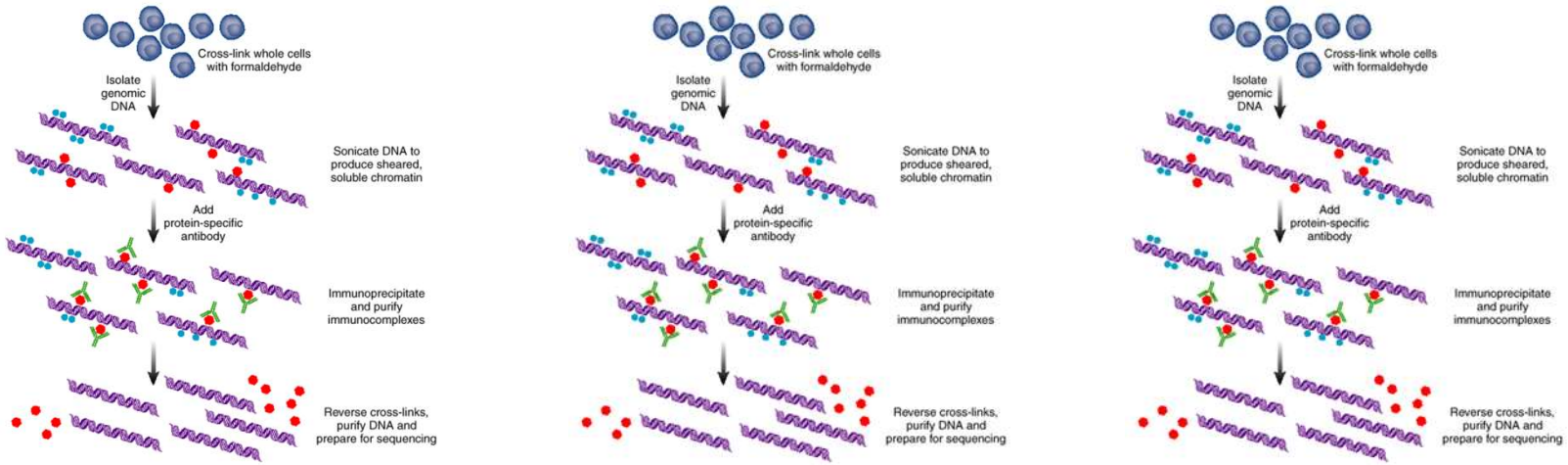
Flies/Worms

- sharp peaks (TFs) 2 million
- broad peaks 5-10 million

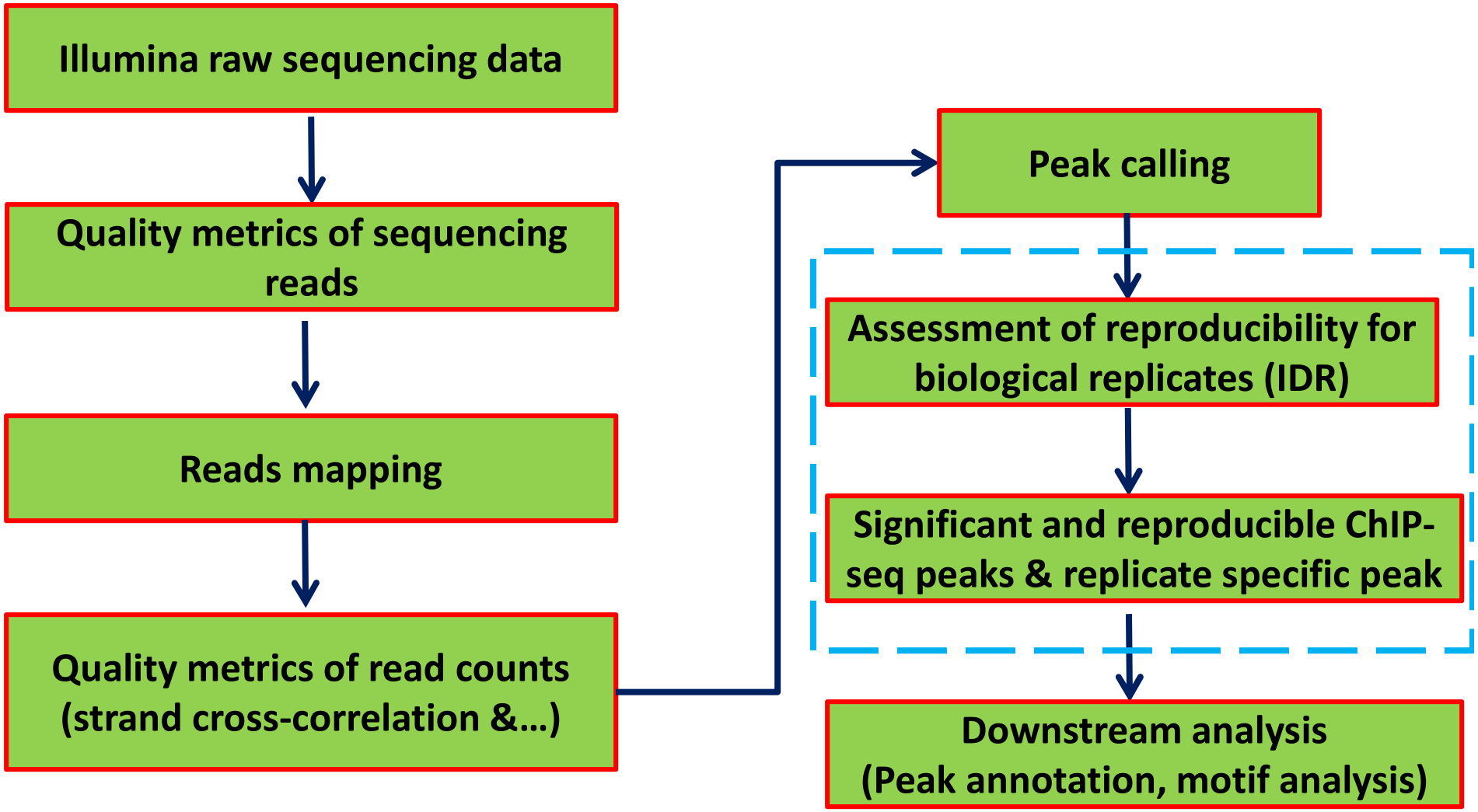
ChIP-Seq experimental design

Single sample

Replicates

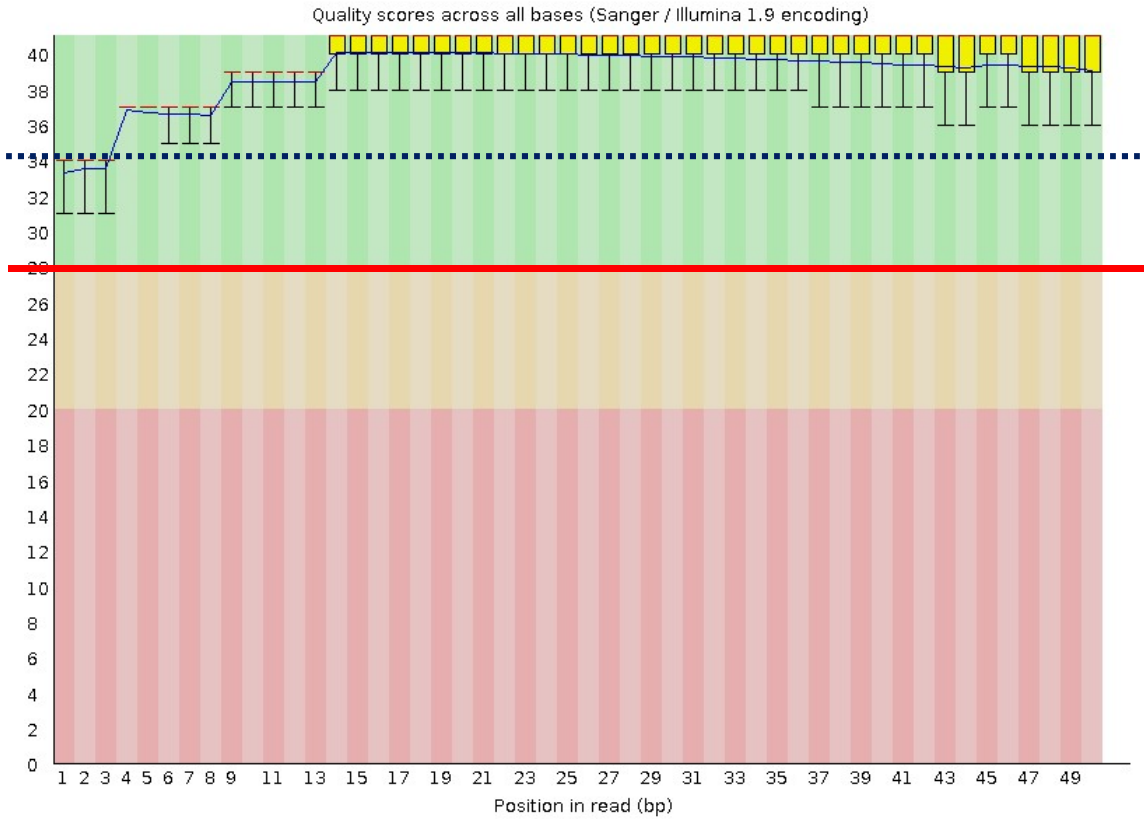


Data analysis protocol



Quality metrics of sequencing reads

- FastQC can be used for an overview of the data quality
- Phred quality scores used for trimming low quality bases
- *P* = 10^(-*Q*/10); *Q*=30 base is called incorrectly 1 in 1000



fastqc input.fastq

fastx_trimmer [-f N] [-l N] [-m MINLEN] [-i INFILE] [-o OUTFILE]

Reads mapping

Most popular software: Bowtie, BWA, MAQ etc

```
java -jar /programs/trimmomatic/trimmomatic-0.36.jar SE -phred33 input.fq.gz
output.fq.gz ILLUMINACLIP:adapter.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

Reference genome Output base name

```
bowtie2-build   <input>          <output name>
bowtie2  -x <output name> {-1 <m1> -2 <m2> | -U <r>} –p  8 -S [<hit>]
```

Pair-end or single end Output sam file

CPU cores

- Multiple mapping hits were discarded

Reference genome; FASTA format: 2 lines for each read (“>name”, sequence)

```
>I
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
```

Illumina raw data; FASTQ format: 4 lines per read (“@name”, sequence, “+”, quality string)

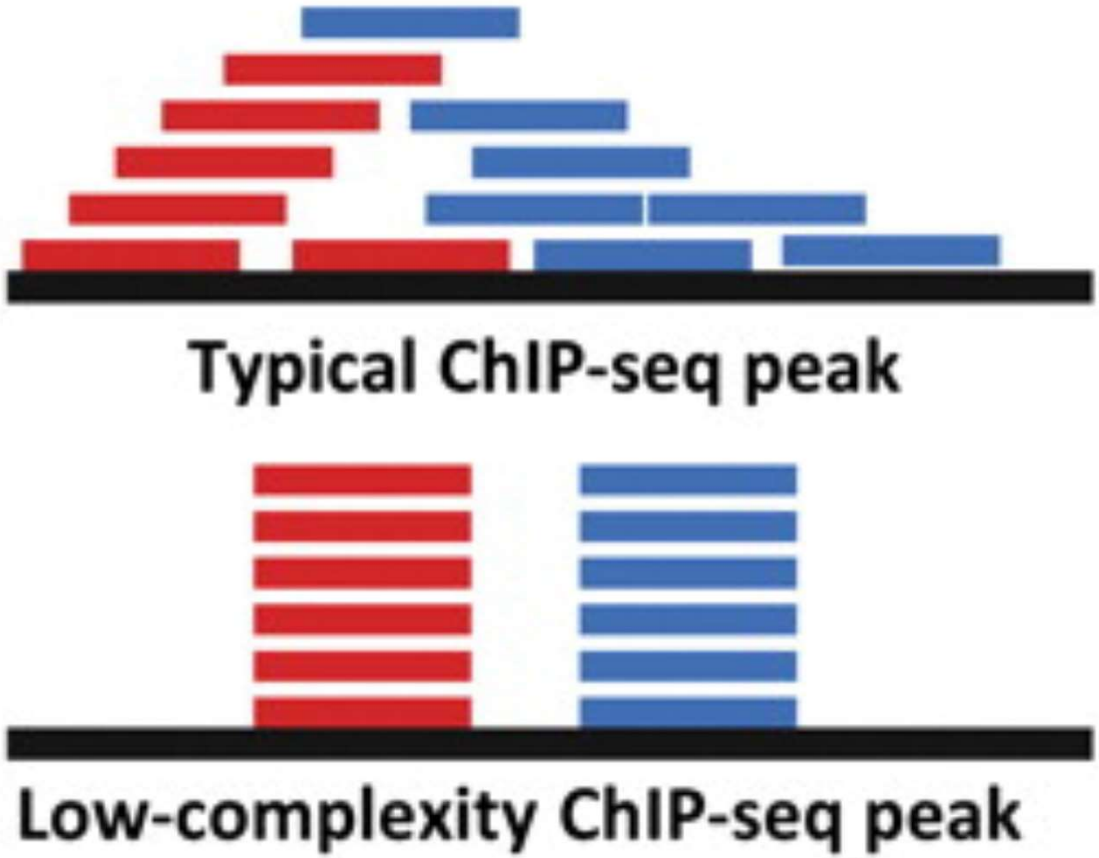
```
@ILLUMINA:405:C269YACXX:1:1101:3833:1996 1:N:0:NAAA
CAATGGAAGAACAGACACTACATATATTGAGCACATTATCATGTTA
+
FFFFHHHFHJJJIHGIGGIGGGIJGIJJJIGEHIGIIJIIHIIHII
```

SAM output

```
ILLUMINA:405:C269YACXX:1:2115:4141:76972        0       I       350017  42      46M     *       0
0       AAATAGATAATTCTTTGAAAATGAATAATTTTTTGGAAAAATCTTT  FFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJ
JJJJJJ  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:46 YT:Z:UU
ILLUMINA:405:C269YACXX:1:2203:10724:54489       0       I       350046  42      46M     *       0
0       TTTTTGGAAAAATCTTTTTTTCTTATTTTTCTAGACGTACACCTTT  FFFFHHHHJJJJJJJJJJJJJJIJJJJJJJJJJJJJJHIJ
JJJJJJ  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:46 YT:Z:UU
```

- Mapping reads proportion (%)

Quality Control

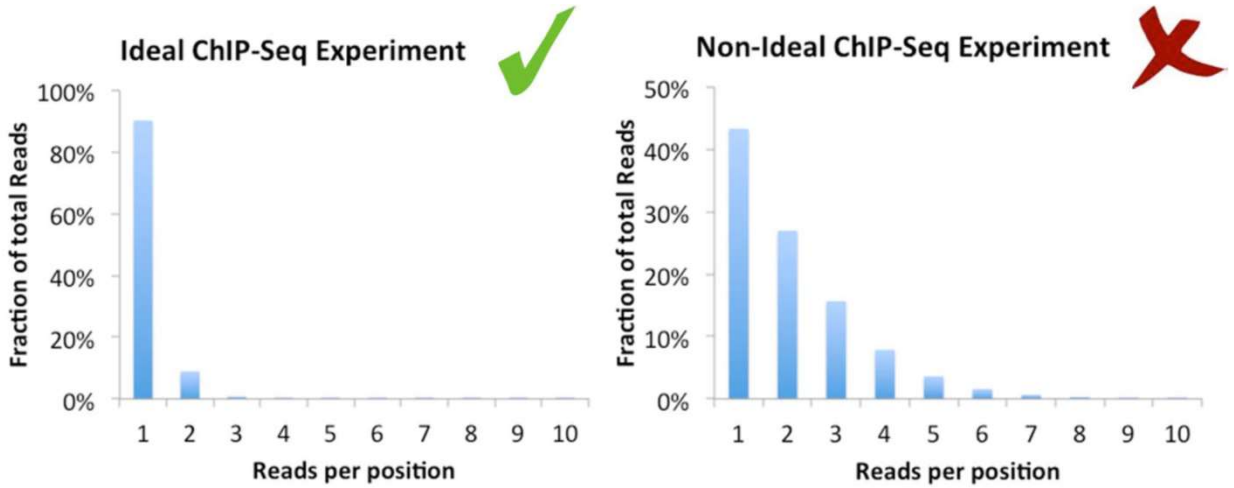


Quality Control

Nonredundant fraction (NRF)

$$\mathrm{NRF} = \frac{\#\text{unique start positions of uniquely mappable reads}}{\#\text{uniquely mappable reads}}$$

ENCODE recommends target of NRF >0.8 for 10 million uniquely mapped reads

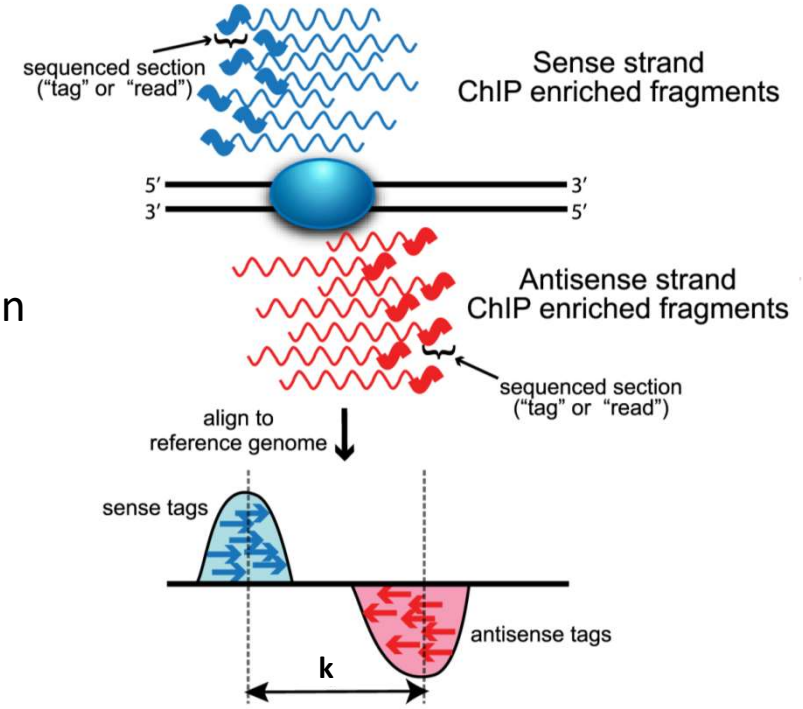


samtools rmdup & picard MarkDuplicates

Signal-to-noise ratio

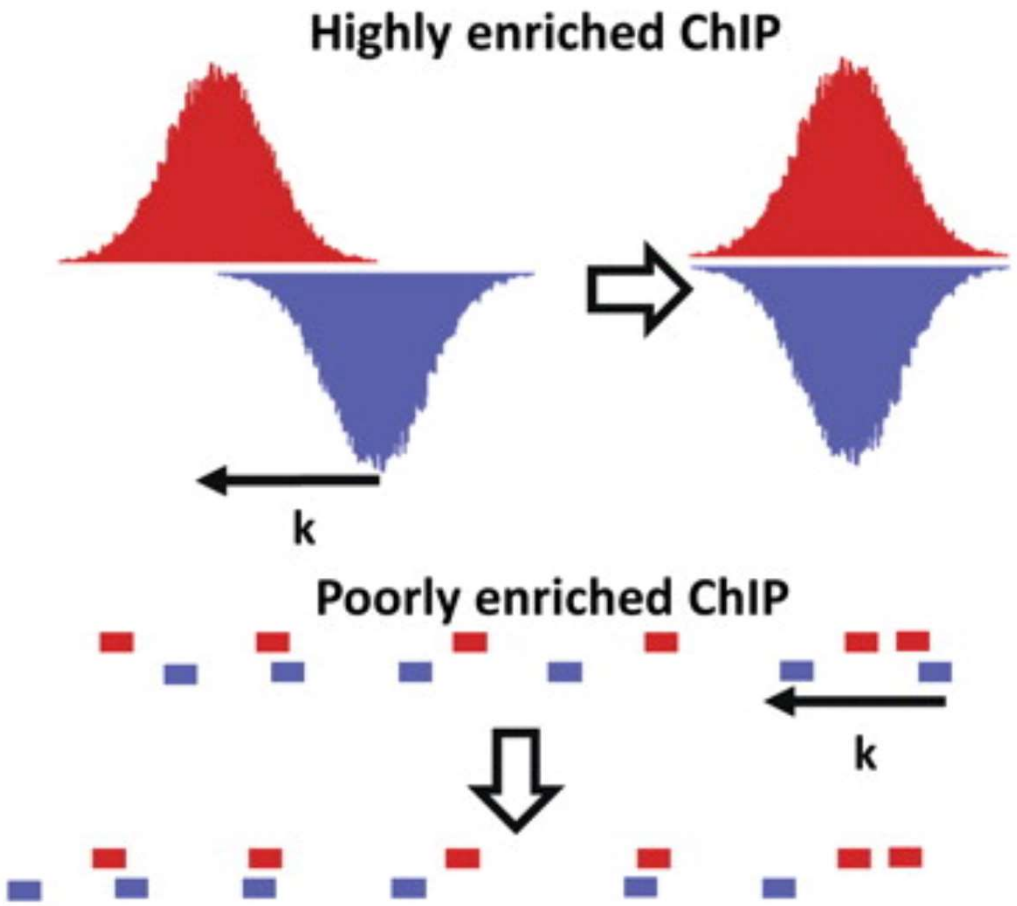
DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5' end.

- With ChIP-seq, the alignment of the reads to the genome results in two peaks (one on each strand) that located on flanking sides of the protein or nucleosome of interest.
- The distance between strands specific peaks (k) represents the average sequenced fragment.

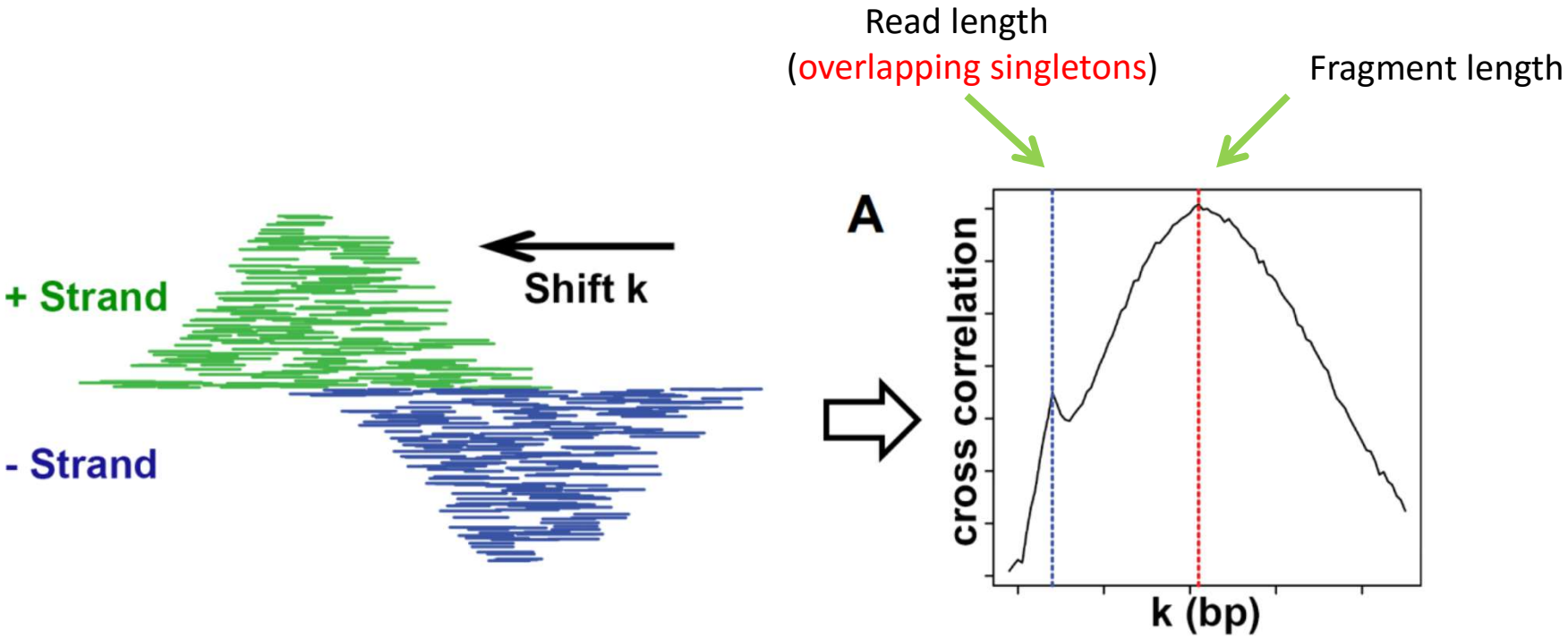


Wilbanks et al (2010) PLoS ONE 5:e11471

Cross-correlation



Cross-correlation

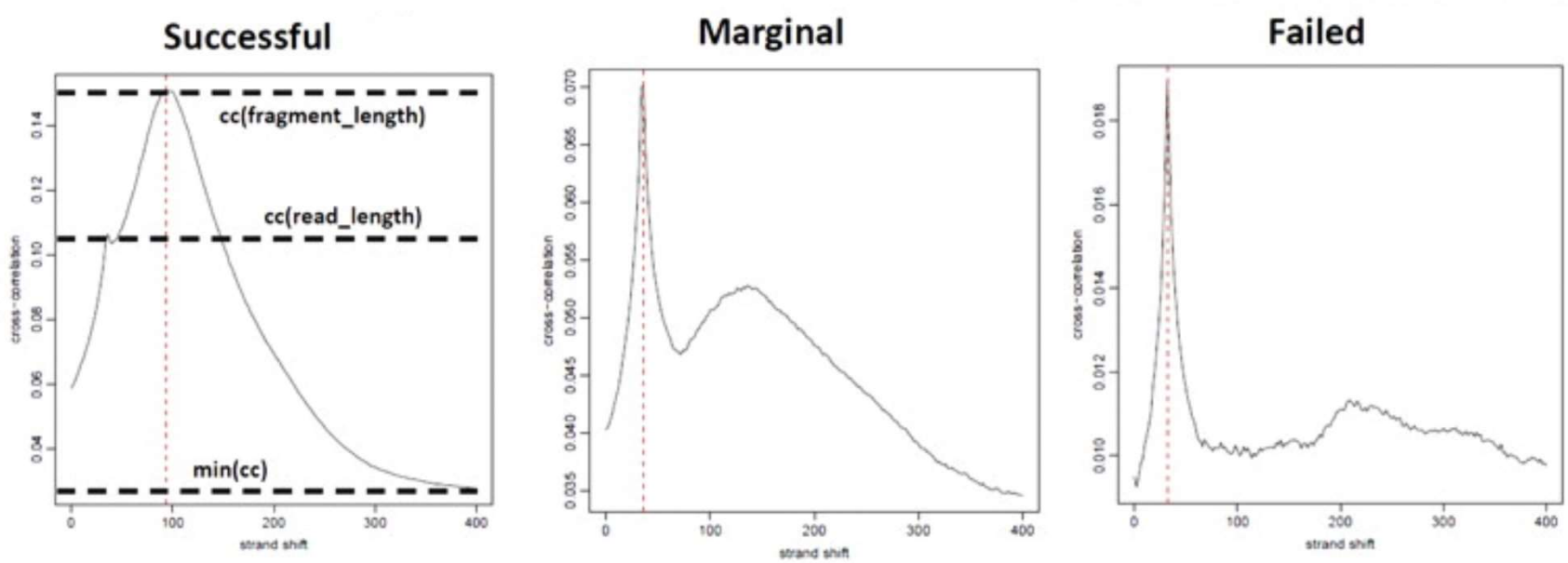


Strand cross-correlation is computed as the Pearson correlation between the positive and the negative strand profiles at different strand shift distances, k

Rscript run_spp.R –c=test.bam –savp –out=output_spp.out

Bailey et al (2013). PLOS Computational Biology

Cross-correlation



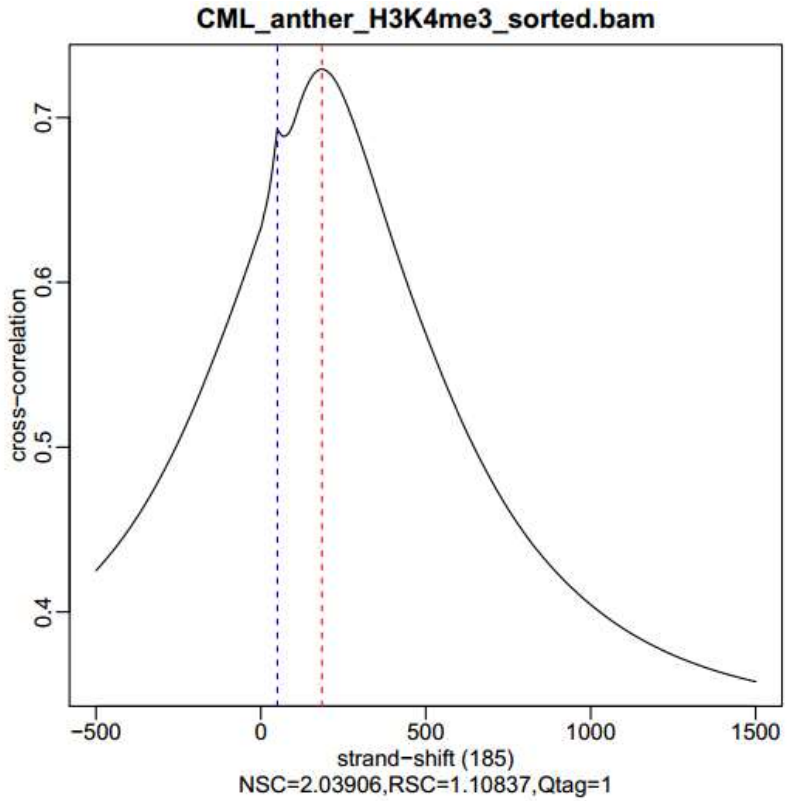
$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Bad data with NSC values < 1.05 and RSC values < 0.8

Cross-correlation

Rscript run_spp.R -c=../../CML_anther_H3K4me3_sorted.bam -savp -out=aa.spp

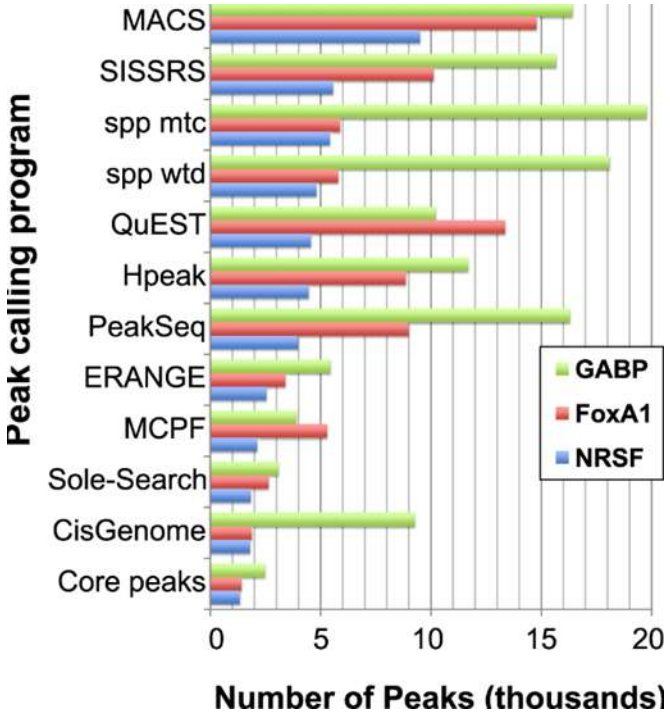
Loading required package: caTools
Reading ChIP tagAlign/BAM file ../../CML_anther_H3K4me3_sorted.bam
opened /tmp/RtmpKAzmGq/CML_anther_H3K4me3_sorted.tagAlign3c9c7d90dab4
done. read 123823458 fragments
ChIP data read length 50
Calculating peak characteristics
Minimum cross-correlation value 0.3577009
Minimum cross-correlation shift 1500
Top 3 cross-correlation values 0.729372332432529
Top 3 estimates for fragment length 185
Window half size 690
Phantom peak location 50
Phantom peak Correlation 0.6930339
Normalized Strand cross-correlation coefficient (NSC) 2.039056
Relative Strand cross-correlation Coefficient (RSC) 1.108365



Peak-calling program

- MACS → Yong Zhang et al
- cisGenome → Hongkai Ji et al
- spp →Peter Park et al
- rbrads → Julie Ahringer et al
- BayesPeak → Simon Tavaré et al
- …

R environment



Elizabeth et al (2010) PLoS ONE 5(7): e11471

Peak caller MACS2

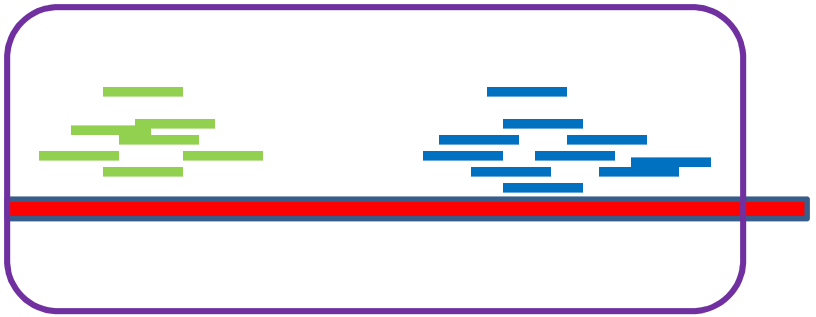
Model-based Analysis of ChIP-seq data (MACS), which has been one of the most commonly used peak callers. MACS introduced a more sophisticated way of modeling the fragment size.

http://liulab.dfci.harvard.edu/MACS/index.html
https://github.com/taoliu/MACS

Zhang et al (2008) Genome Biology 9:R137

Parameters and concepts of MACS2

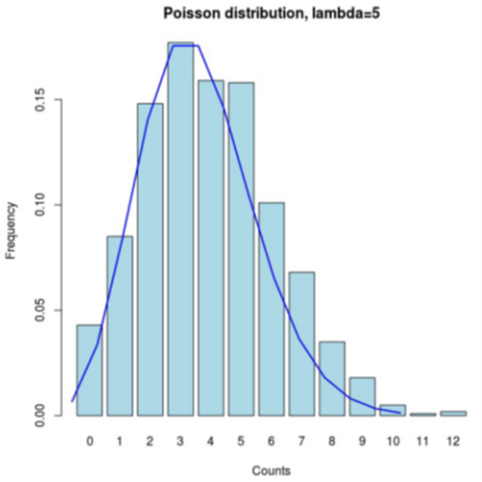
-DNA treatment & input sample
-DNA fragment length
-Band width
-Effect genome size
-Non-redundant reads
-call summits
-mfold
-qvalue



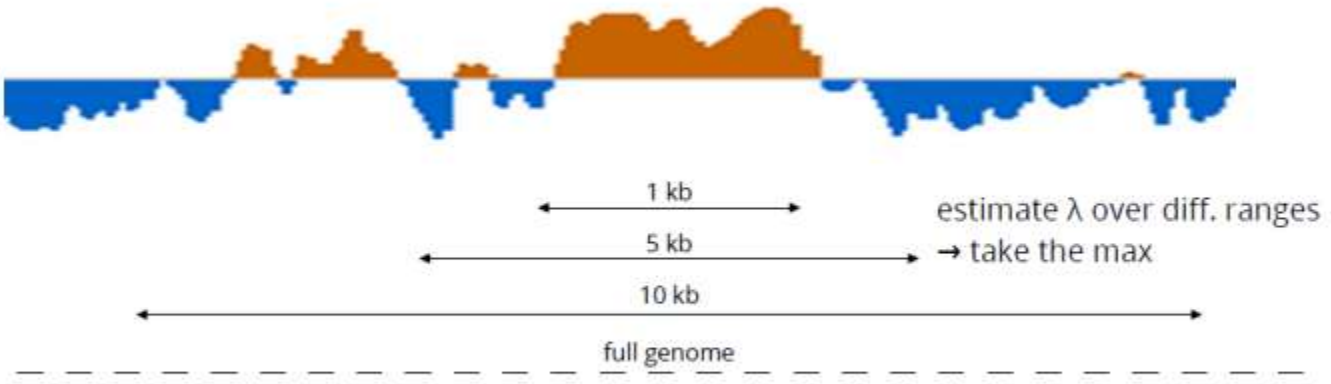
$$\lambda = \frac{\ell \times N}{G^*}$$

.

$$P(H \geq h) = \sum_{k=h}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = 1 - \sum_{k=0}^{h-1} \frac{e^{-\lambda}\lambda^k}{k!}$$



Dynamic local lambda



Usage of MACS2

macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01

- callpeak: Main MACS2 Function to Call peaks from alignment results.
- bdgpeakcall: Call peaks from bedGraph output.
- bdgbroadcall: Call broad peaks from bedGraph output.
- bdgcmp: Deduct noise by comparing two signal tracks in bedGraph.
- bdgdiff: Differential peak detection based on paired four bedgraph files.
- filterdup: Remove duplicate reads at the same position, then convert acceptable format to BED format.
- predictd: Predict d or fragment size from alignment results.
- pileup: Pileup aligned reads with a given extension size (fragment size or d in MACS language). Note there will be no step for duplicate reads filtering or sequencing depth scaling, so you may need to do certain post- processing.
- randsample: Randomly sample number/percentage of total reads.
- refinepeak: (Experimental) Take raw reads alignment, refine peak summits and give scores measuring balance of forward- backward tags. Inspired by SPP

Callpeak - options

Various options to indicate/control input, output, peak modelling and peak calling

macs2 callpeak

```
usage: macs2 callpeak [-h] -t TFILE [TFILE ...] [-c [CFILE [CFILE ...]]]
        [-f {AUTO,BAM,SAM,BED,ELAND,ELANDMULTI,ELANDEXPORT,BOWTIE,
             BAMPE}]
        [-g GSIZE] [--keep-dup KEEPDUPLICATES]
        [--buffer-size BUFFER_SIZE] [--outdir OUTDIR] [-n NAME]
        [-B] [--verbose VERBOSE] [--trackline] [--SPMR]
        [-s TSIZE] [--bw BW] [-m MFOLD MFOLD] [--fix-bimodal]
        [--nomodel] [--shift SHIFT] [--extsize EXTSIZE]
        [-q QVALUE] [-p PVALUE] [--to-large] [--ratio RATIO]
        [--down-sample] [--seed SEED] [--nolambda]
        [--slocal SMALLLOCAL] [--llocal LARGELOCAL] [--broad]
        [--broad-cutoff BROADCUTOFF] [--call-summits]
```

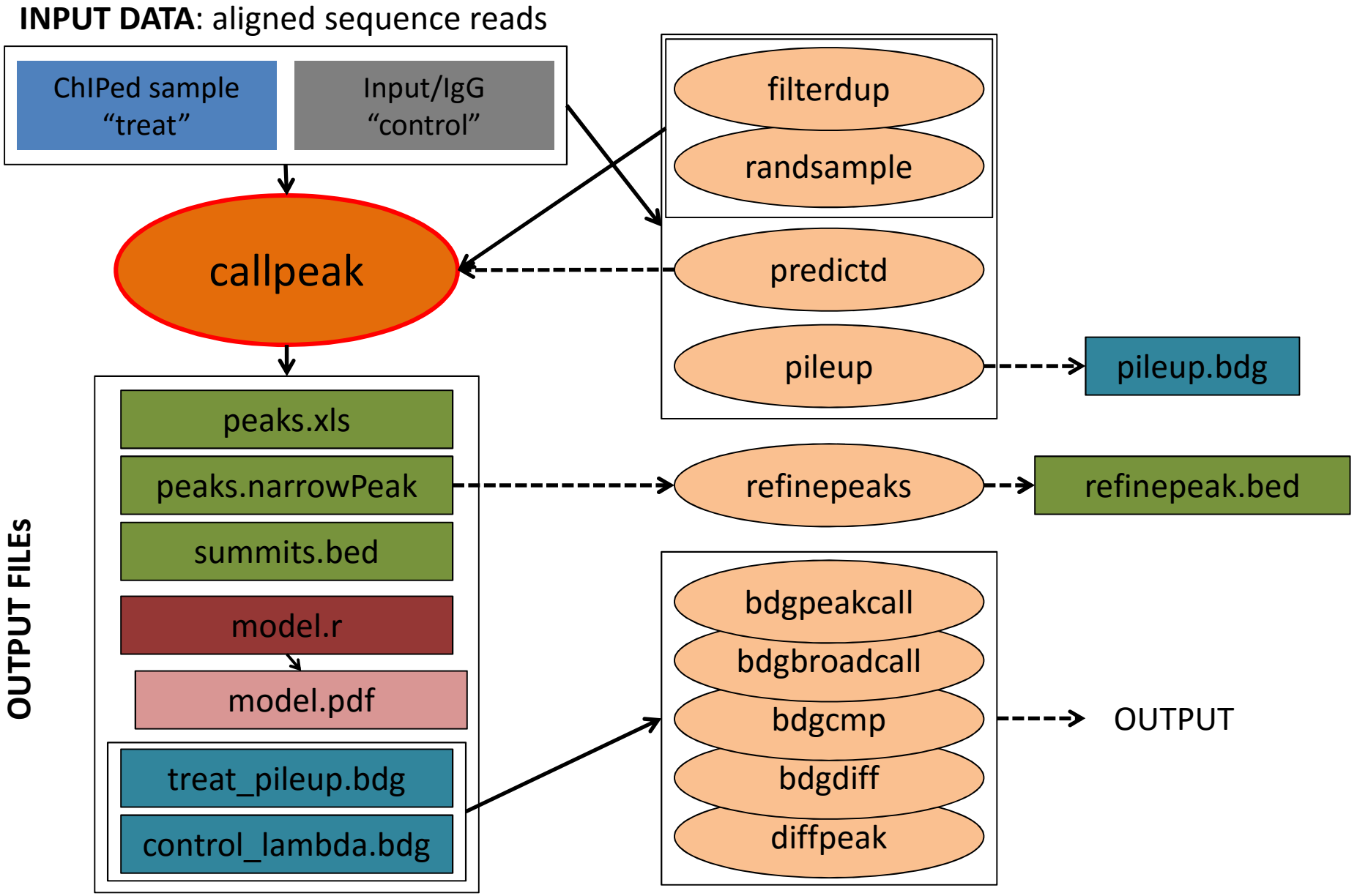
Input

output

Modelling

Peak Calling

MACS2 – program(s)



Examples of MACS setting

- Default setting

```
macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test
```

- Adjust mfold limits and bandwidth

```
macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01 –m 10 30
bw 300
```

- Stop shifting model setting

```
macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01 --nomodel --extsize
200  --shift 0
```

- Post-processing

```
macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01 –nomodel --extsize
200  --shift 0 --call-summits
```

Output of MACS2

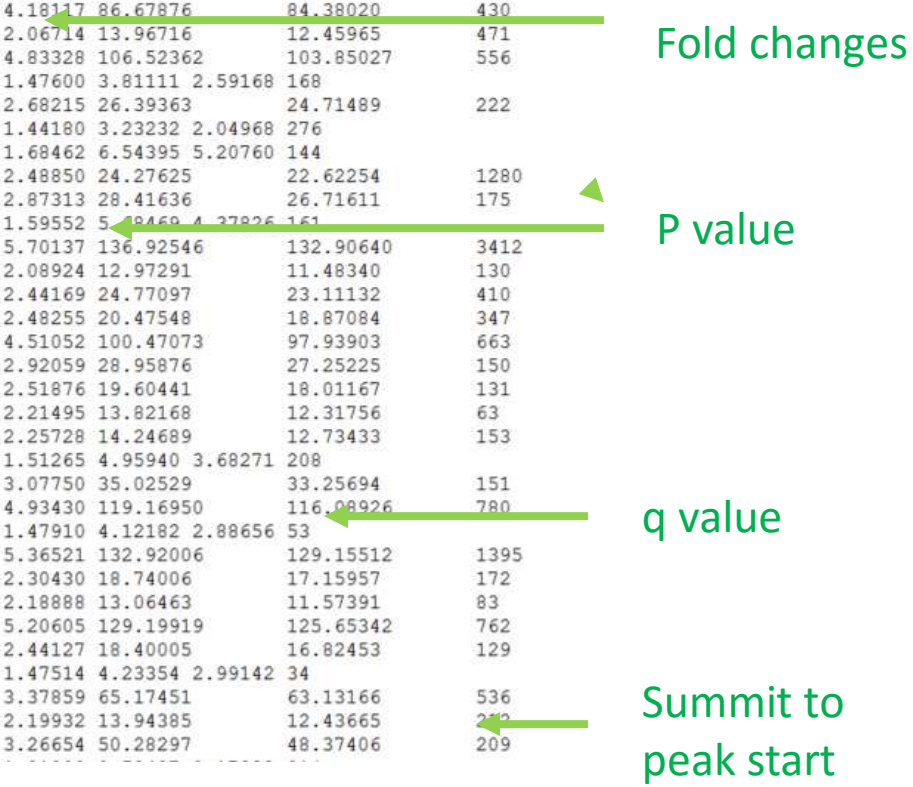
```
# This file is generated by MACS version 2.1.0.20150731
# Command line: callpeak -t H3K4me3_wt_combine.bam -c H3_wt_combine.bam -n H3K4me3_wt_h3_narrow --nomodel --shift 0
# ARGUMENTS LIST:
# name = H3K4me3_wt_h3_narrow
# format = AUTO
# ChIP-seq file = ['H3K4me3_wt_combine.bam']
# control file = ['H3_wt_combine.bam']
# effective genome size = 9.00e+07
# band width = 300
# model fold = [2, 10]
# qvalue cutoff = 1.00e-02
# Larger dataset will be scaled towards smaller dataset.
# Range for calculating regional lambda is: 1000 bps and 10000 bps
# Broad region calling is off

# tag size is determined as 101 bps
# total tags in treatment: 30801118
# tags after filtering in treatment: 24581775
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.20
# total tags in control: 50559028
# tags after filtering in control: 41283295
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.18
# d = 178
chr     start   end     length  abs_summit      pileup  -log10(pvalue)  fold_enrichment -log10(qvalue)  name
I       3771    4563    793     4075    246.00  82.34708        4.57456 80.10690        H3K4me3_wt_h3_narrow_peak_1
I       16402   17106   705     16903   201.00  58.86108        4.01713 56.87683        H3K4me3_wt_h3_narrow_peak_2
I       24137   24786   650     24375   187.00  33.28239        2.77739 31.53087        H3K4me3_wt_h3_narrow_peak_3
I       26239   27330   1092    26669   287.00  86.67876        4.18117 84.38020        H3K4me3_wt_h3_narrow_peak_4
I       39624   40498   875     40095   134.00  13.96716        2.06714 12.45965        H3K4me3_wt_h3_narrow_peak_5
I       46700   47876   1177    47256   303.00  106.52362       4.83328 103.85027       H3K4me3_wt_h3_narrow_peak_6
```

Summit

FC

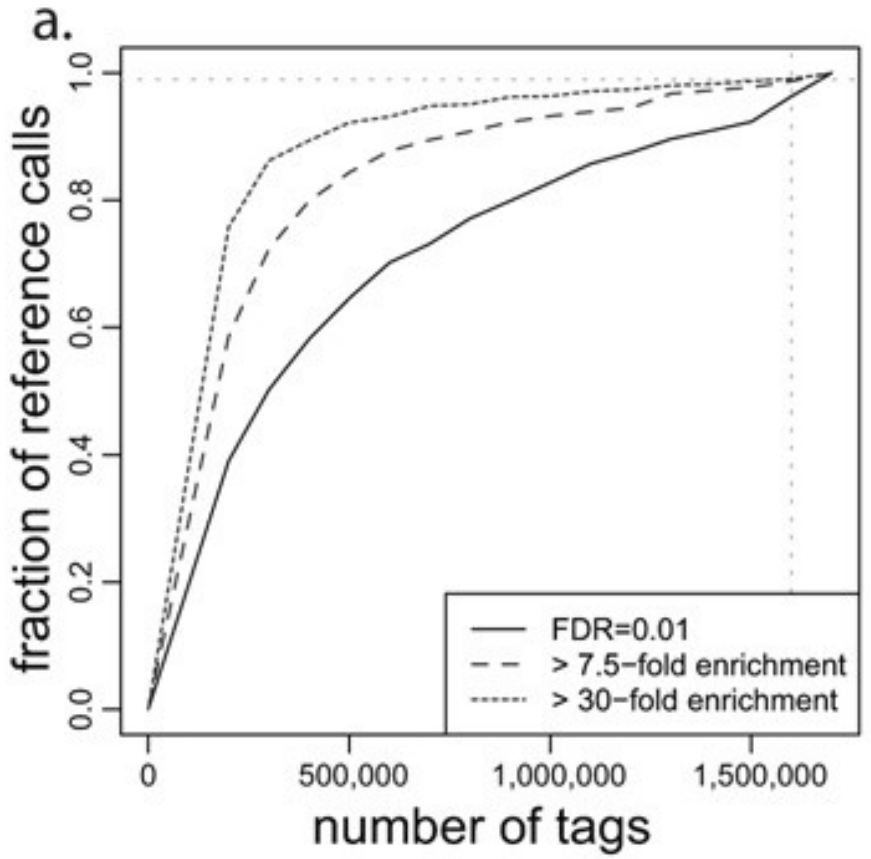
Output of MACS2

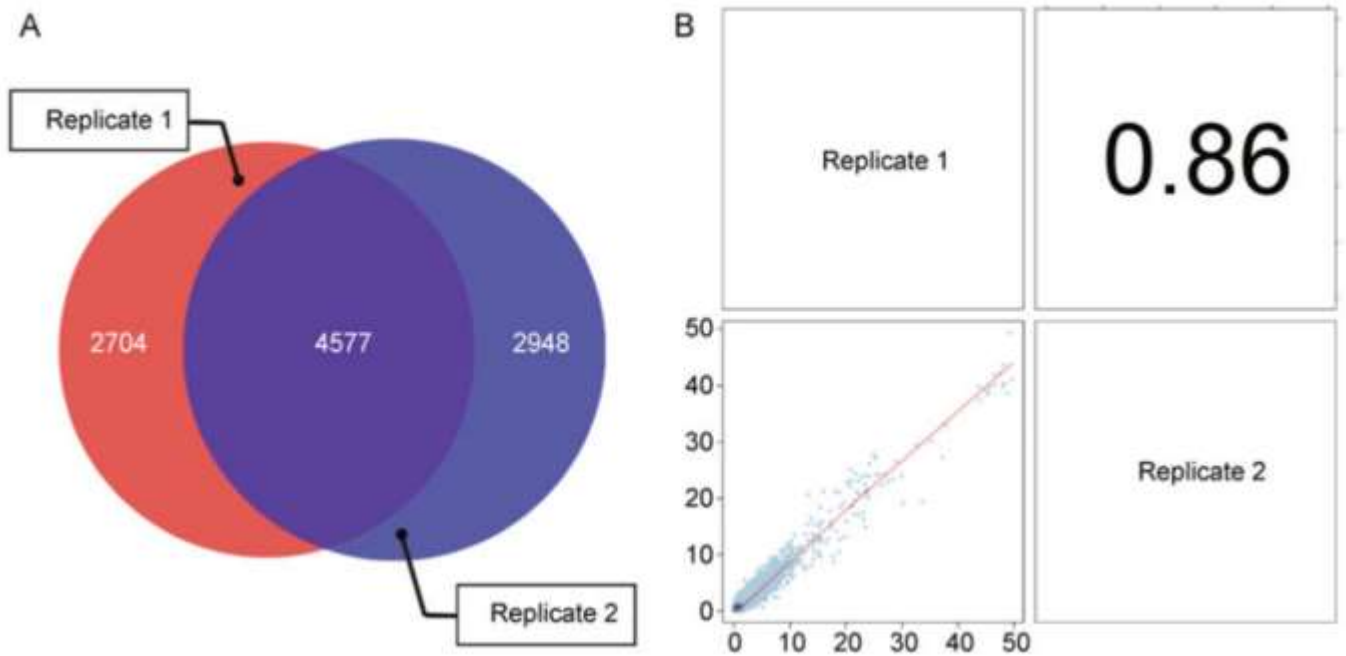
```
[mingh@cbsum1c2b007 H3K4me3]$ more H3K4me3_wt_h3_narrow_peaks.narrowPeak
I       3770    4563    H3K4me3_wt_h3_narrow_peak_1     801     .       4.57456 82.34708        80.10690        304
I       16401   17106   H3K4me3_wt_h3_narrow_peak_2     568     .       4.01713 58.86108        56.87683        501
I       24136   24786   H3K4me3_wt_h3_narrow_peak_3     315     .       2.77739 33.28239        31.53087        238
I       26238   27330   H3K4me3_wt_h3_narrow_peak_4     843     .       4.18117 86.67876        84.38020        430
I       39623   40498   H3K4me3_wt_h3_narrow_peak_5     124     .       2.06714 13.96716        12.45965        471
I       46699   47876   H3K4me3_wt_h3_narrow_peak_6     1038    .       4.83328 106.52362       103.85027       556
I       63262   63440   H3K4me3_wt_h3_narrow_peak_7     25      .       1.47600 3.81111 2.59168 168
I       70036   70674   H3K4me3_wt_h3_narrow_peak_8     247     .       2.68215 26.39363        24.71489        222
I       71021   71303   H3K4me3_wt_h3_narrow_peak_9     20      .       1.44180 3.23232 2.04968 276
I       72183   72675   H3K4me3_wt_h3_narrow_peak_10    52      .       1.68462 6.54395 5.20760 144
I       92180   94658   H3K4me3_wt_h3_narrow_peak_11    226     .       2.48850 24.27625        22.62254        1280
I       96293   96756   H3K4me3_wt_h3_narrow_peak_12    267     .       2.87313 28.41636        26.71611        175
I       106983  107430  H3K4me3_wt_h3_narrow_peak_13    43      .       1.59552 5.18469 4.37826 161
I       107835  112973  H3K4me3_wt_h3_narrow_peak_14    1329    .       5.70137 136.92546       132.90640       3412
I       113096  113434  H3K4me3_wt_h3_narrow_peak_15    114     .       2.08924 12.97291        11.48340        130
I       128242  129073  H3K4me3_wt_h3_narrow_peak_16    231     .       2.44169 24.77097        23.11132        410
I       180049  180740  H3K4me3_wt_h3_narrow_peak_17    188     .       2.48255 20.47548        18.87084        347
I       182198  183518  H3K4me3_wt_h3_narrow_peak_18    979     .       4.51052 100.47073       97.93903        663
I       215397  215715  H3K4me3_wt_h3_narrow_peak_19    272     .       2.92059 28.95876        27.25225        150
I       215831  216244  H3K4me3_wt_h3_narrow_peak_20    180     .       2.51876 19.60441        18.01167        131
I       216422  217849  H3K4me3_wt_h3_narrow_peak_21    123     .       2.21495 13.82168        12.31756        63
I       237416  237664  H3K4me3_wt_h3_narrow_peak_22    127     .       2.25728 14.24689        12.73433        153
I       250818  251170  H3K4me3_wt_h3_narrow_peak_23    36      .       1.51265 4.95940 3.68271 208
I       251488  252492  H3K4me3_wt_h3_narrow_peak_24    332     .       3.07750 35.02529        33.25694        151
I       288149  289602  H3K4me3_wt_h3_narrow_peak_25    1160    .       4.93430 119.16950       116.08926       780
I       313206  313515  H3K4me3_wt_h3_narrow_peak_26    28      .       1.47910 4.12182 2.88656 53
I       313676  315746  H3K4me3_wt_h3_narrow_peak_27    1291    .       5.36521 132.92006       129.15512       1395
I       315973  316279  H3K4me3_wt_h3_narrow_peak_28    171     .       2.30430 18.74006        17.15957        172
I       316456  316664  H3K4me3_wt_h3_narrow_peak_29    115     .       2.18888 13.06463        11.57391        83
I       322624  324424  H3K4me3_wt_h3_narrow_peak_30    1256    .       5.20605 129.19919       125.65342       762
I       342969  343228  H3K4me3_wt_h3_narrow_peak_31    168     .       2.44127 18.40005        16.82453        129
I       346655  346968  H3K4me3_wt_h3_narrow_peak_32    29      .       1.47514 4.23354 2.99142 34
I       348076  348967  H3K4me3_wt_h3_narrow_peak_33    631     .       3.37859 65.17451        63.13166        536
I       363772  364103  H3K4me3_wt_h3_narrow_peak_34    124     .       2.19932 13.94385        12.43665        248
I       364266  365366  H3K4me3_wt_h3_narrow_peak_35    483     .       3.26654 50.28297        48.37406        209
```

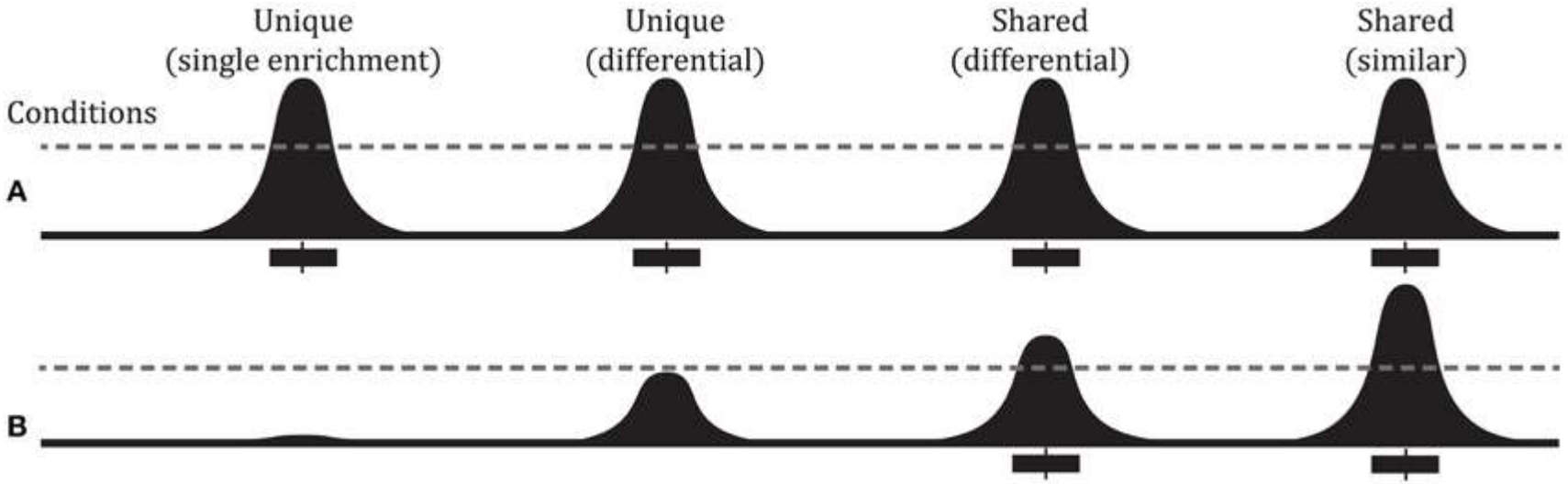


Peak quality checking

- Assessing coverage saturation







How to make the best use of the variability between replicates ???

Consistency of replicates: IDR

- IDR the irreproducible discovery rate
- Each list of peaks is ranked according to p-value or signal score
- The IDR method adopted the bivariate rank distributions over the replicates in order to separate signal from noise based on consistency and reproducibility of identifications

Old version

```
Rscript batch-consistency-analysis.r [peakfile1] [peakfile2] -1 [outfile.prefix] 0 F p.value
Rscript batch-consistency-plot.r [npairs] [output.prefix] [input.file.prefix1] [input.file.prefix2] [input.file.prefix3]
```

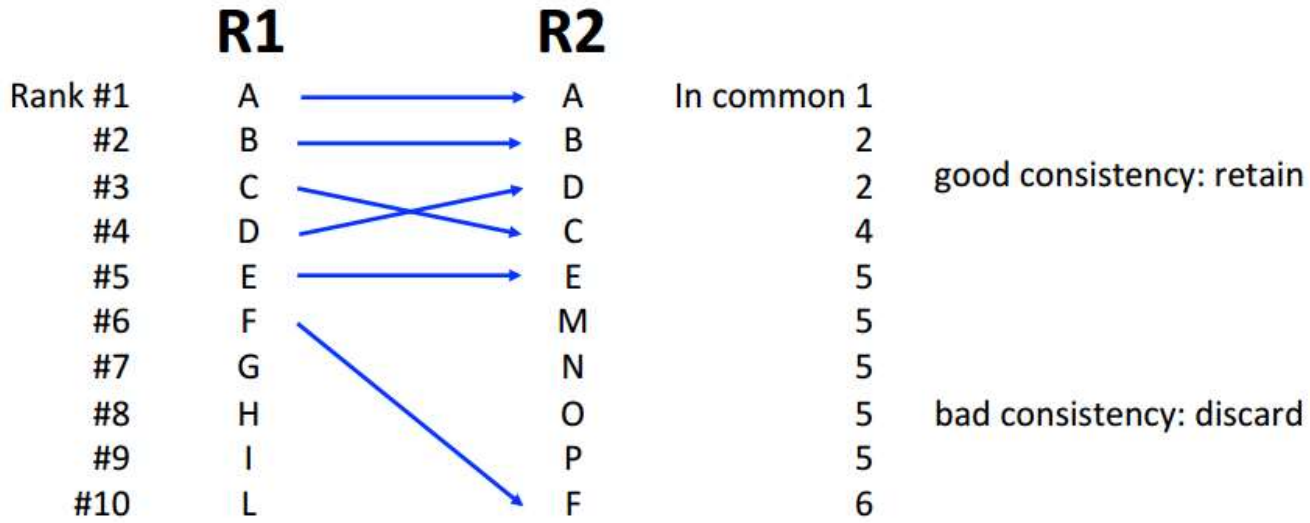
New version

```
idr --samples ../idr/test/data/peak1 ../idr/test/data/peak2
```

https://github.com/nboley/idr

IDR example

cat rep1a.narrowPeak|sort -k8,8nr|head –n 100000 >rep1a_sorted.narrowPeak
cat rep1b.narrowPeak|sort -k8,8nr|head –n 100000 >rep1b_sorted.narrowPeak

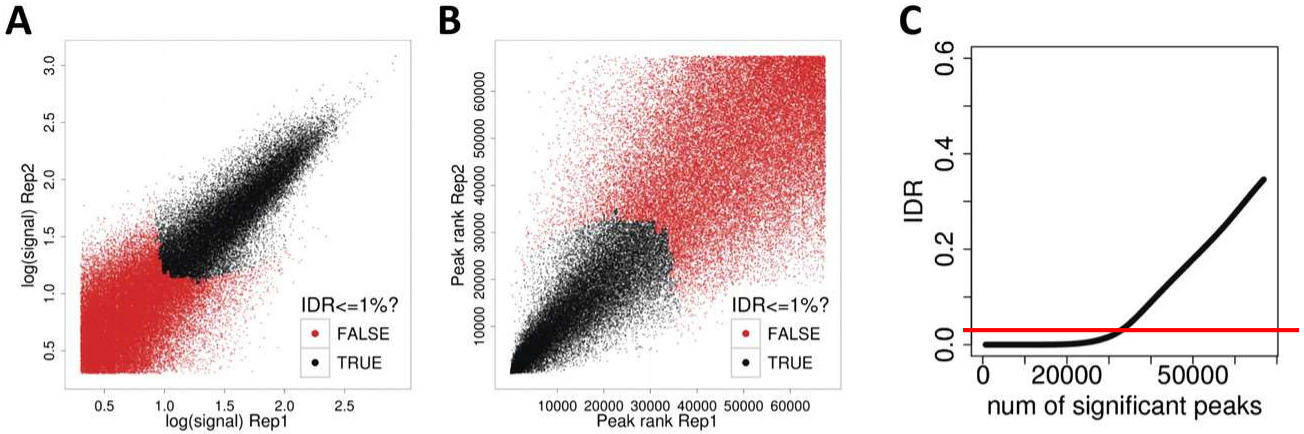



```
Rscript batch-consistency-analysis.r rep1a_sorted.narrowPeak  rep1b_sorted.narrowPeak \
-1 rep1a_vs_1b 0 F p.value
```

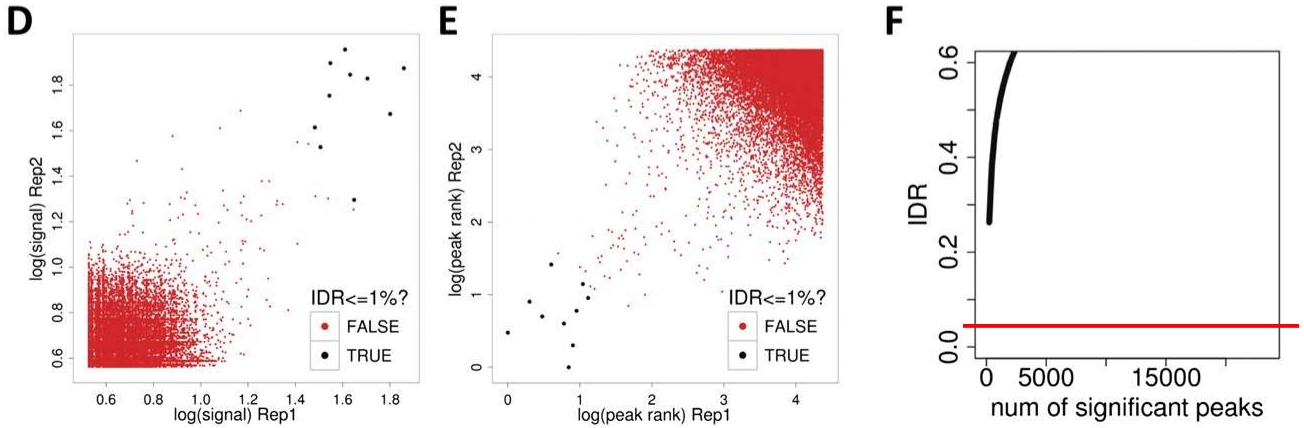
F=narrowPeak
T=broadPeak

IDR

RAD21 Replicates (high reproducibility)



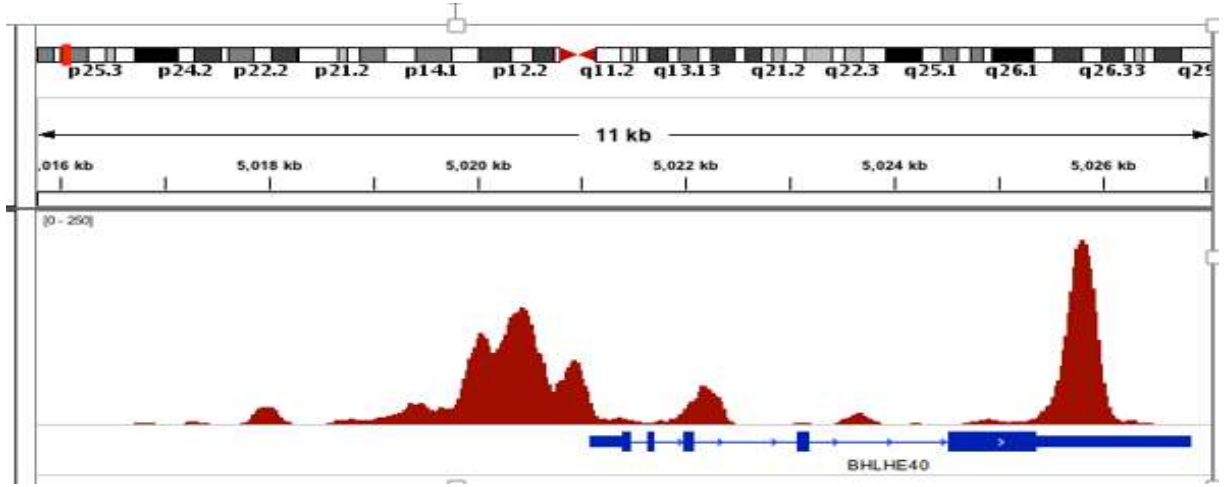
SPT20 Replicates (low reproducibility)



Combine treatment bam

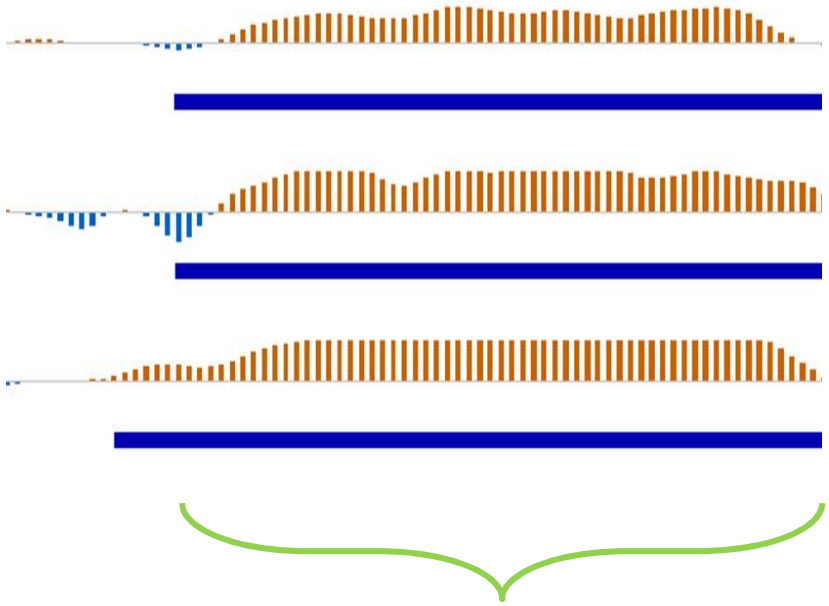
Combine input bam

MACS2

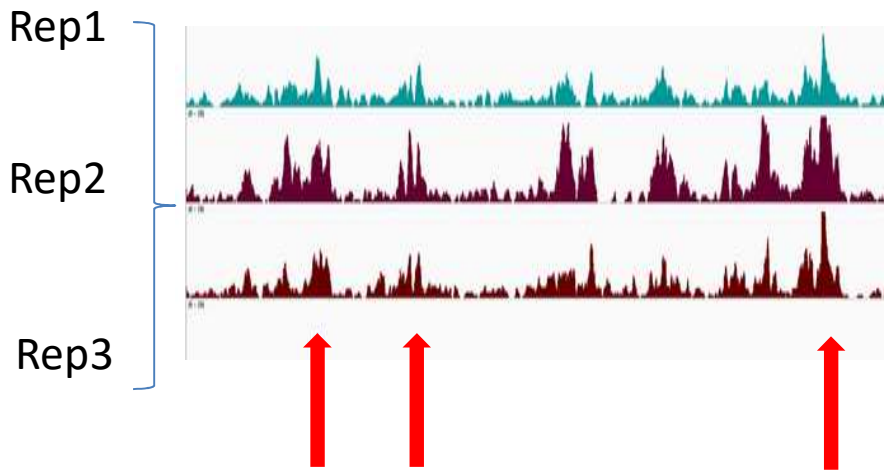


bedtools multicov [OPTIONS] -bams aln.1.bam aln.2.bam ... aln.n.bam -bed <bed/gff/vcf>

Peak region merging and statistics



Final peak list



bedtools multicov [OPTIONS] -bams aln.1.bam aln.2.bam ... aln.n.bam -bed <bed/gff/vcf>

Multiple replicates

$$g(N_{ij}) = \mu + x_i \beta_i + z_j u_j + \varepsilon_{ij}$$

N_{ij} : observed reads count for i^{th} sample and j^{th} biological replicate

β_i : i^{th} sample effect (fixed)

u_j : random effect due to j^{th} biological replicate

ε_{ij} : error

Link function: log-link for Poisson family

More complex comparison

Parameter	Contrast 1	Contrast 2	Contrast 3
β_{Young_ChIP}	1	0	0.5
$\beta_{Young\_control}$	-1	0	-0.5
β_{Old_ChIP}	0	1	-0.5
$\beta_{Old\_control}$	0	-1	0.5

Yong IP Vs control; Old IP Vs control and Yong Vs Old under control

Tools for differential peak calling

- edgeR
- DESeq2
- DiffBind
- MMDiff
- MAnorm