Computational Pipeline for ChIP-Seq Data Analysis

Minghui Wang, Qi Sun
Bioinformatics Facility
Institute of Biotechnology

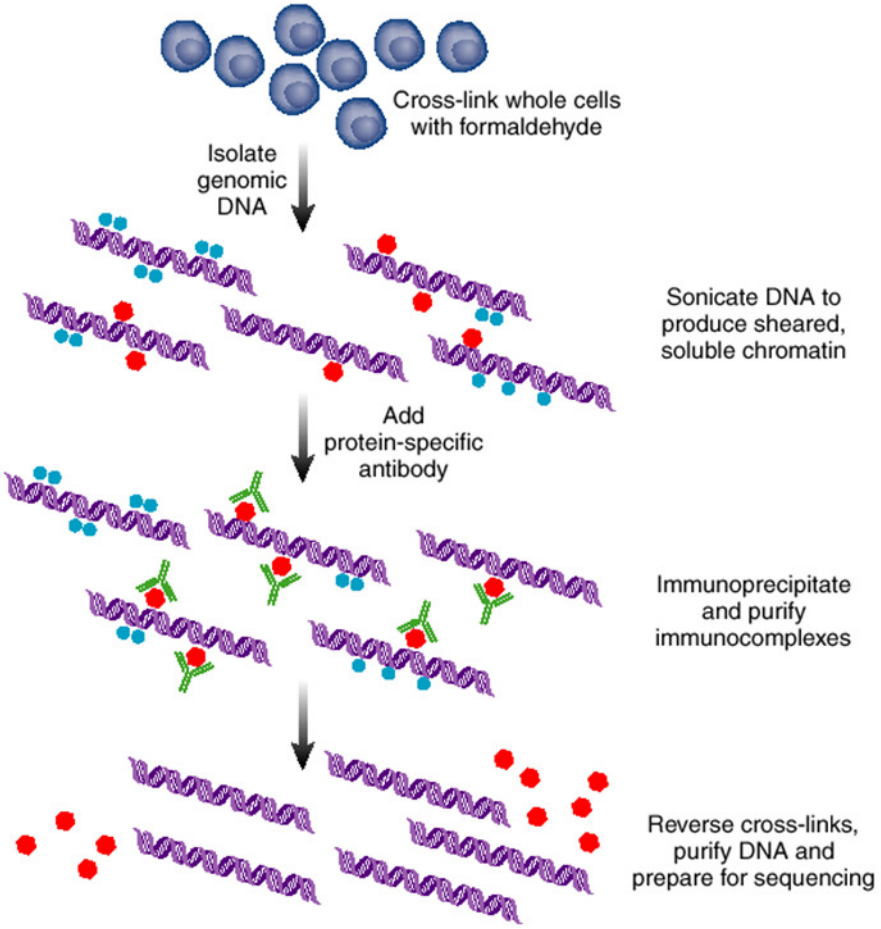
Outline

ChIP-Seq experimental design

Data analysis
- Sequencing data evaluation
- Peak calling & evaluation
- GLM model for multiple replicates

Downstream analysis
- Peak annotation
- Function enrichment

ChIP-seq workflow



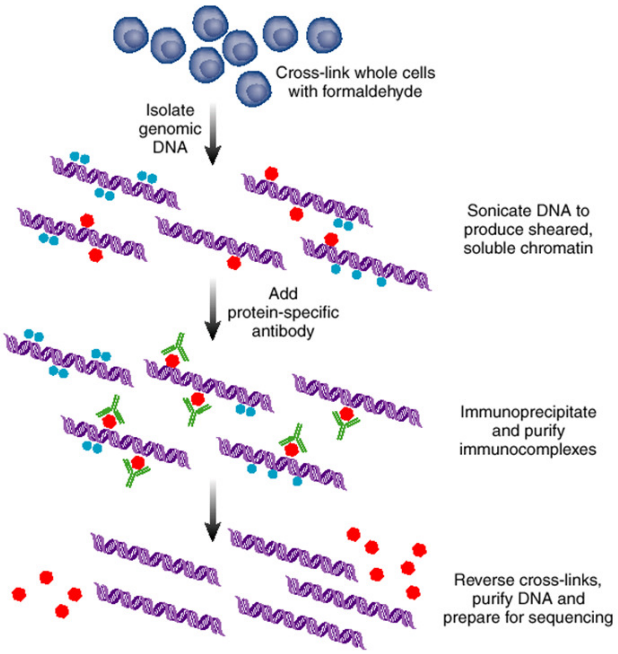
Elaine R Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nature Methods 4:613-614

Controls for ChIP-seq

Most experimental protocols involve a control sample that is processed the same way as the test sample except that no immunoprecipitaion step or no specific antibody

Input DNA & IgG

- Input DNA does not demonstrate “flat” or random (Poisson) distribution.
- Open chromatin regions tend to be fragmented more easily during shearing.
- Amplification bias.
- Mapping artifacts-increased coverage of more “mappable” regions (which also tend to be promotor regions) and repetitive regions due inaccuracies in number of copies in assembled genome.



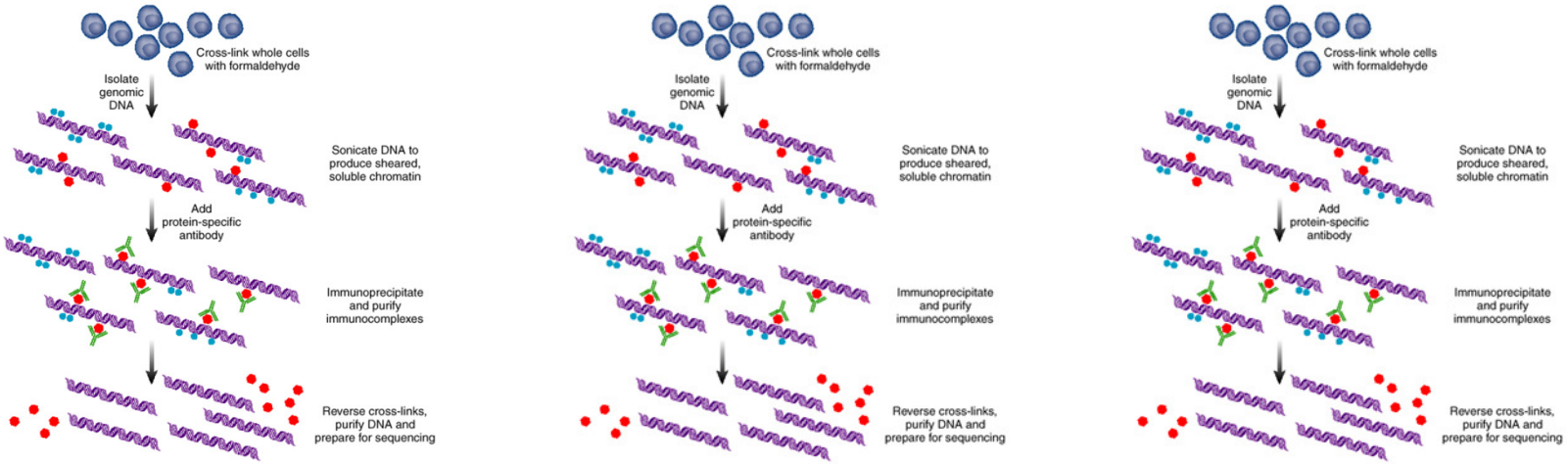
ChIP-Seq experimental design

Single sample

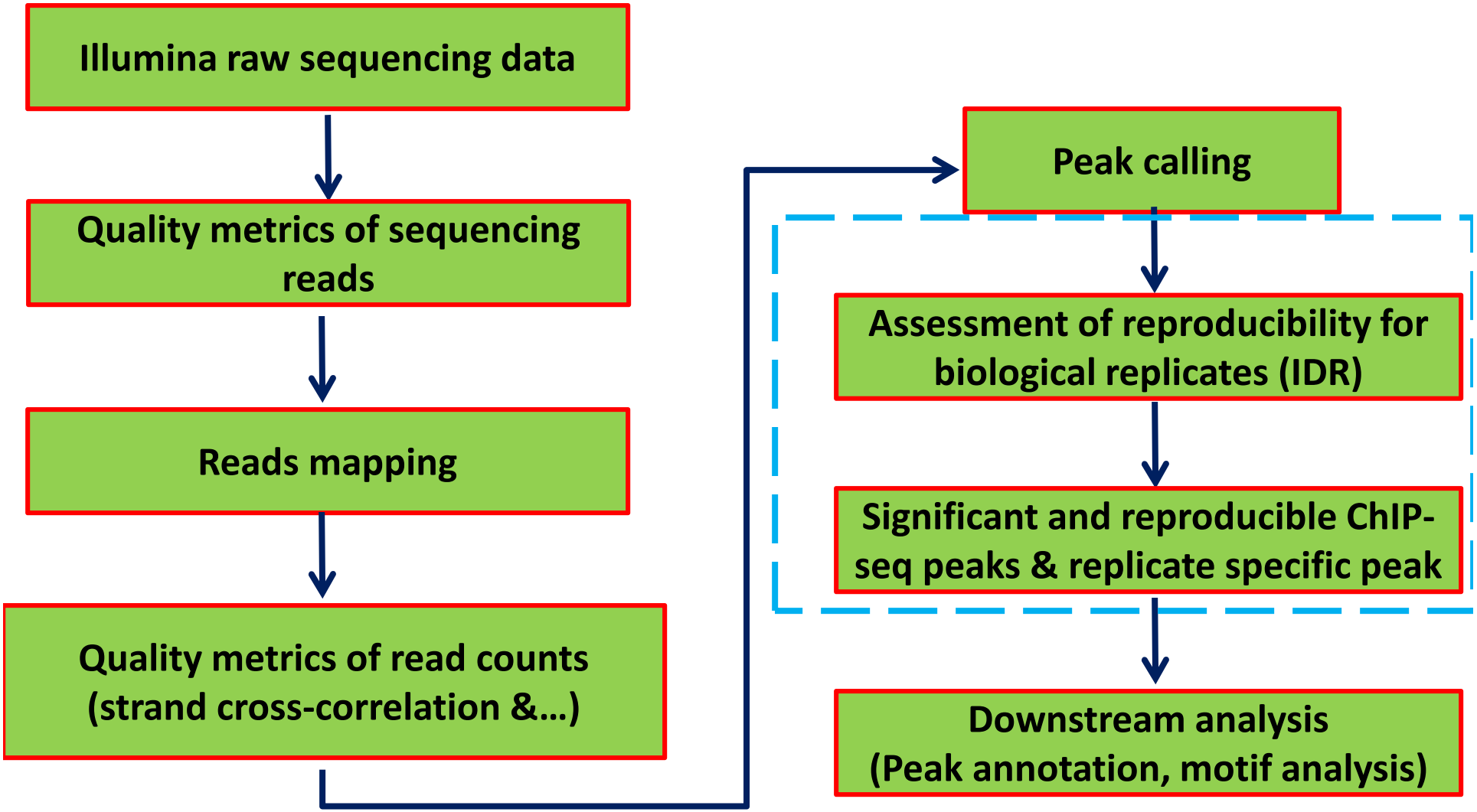
Replicates

Biological replicate

Technical replicate

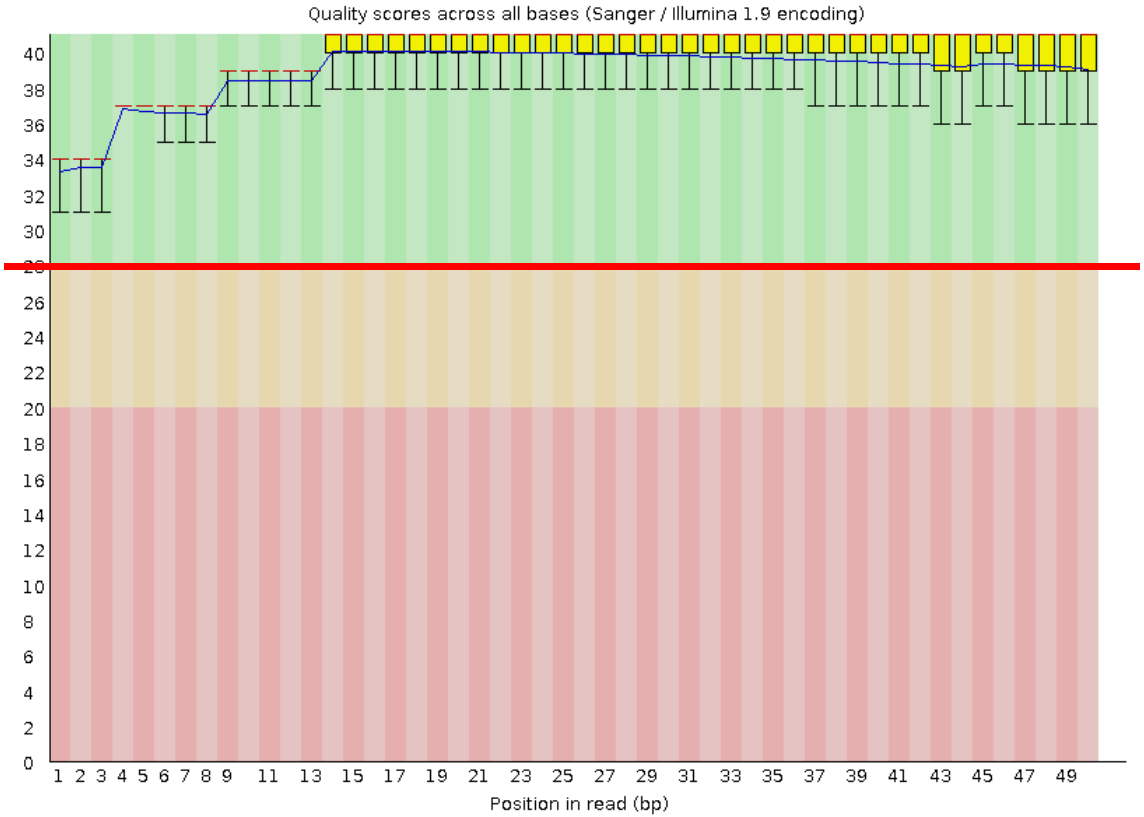


Data analysis protocol



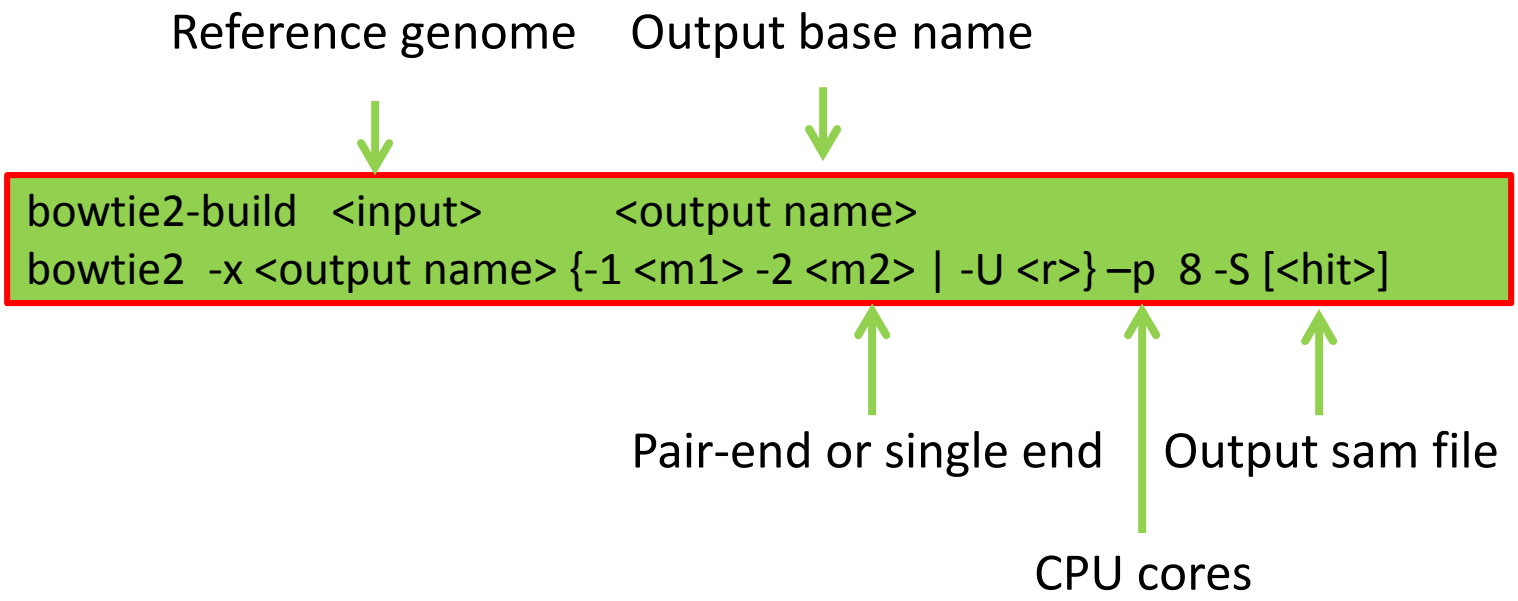
Quality metrics of sequencing reads

- 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



Reads mapping

Most popular software: Bowtie, BWA, MAQ etc



- 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

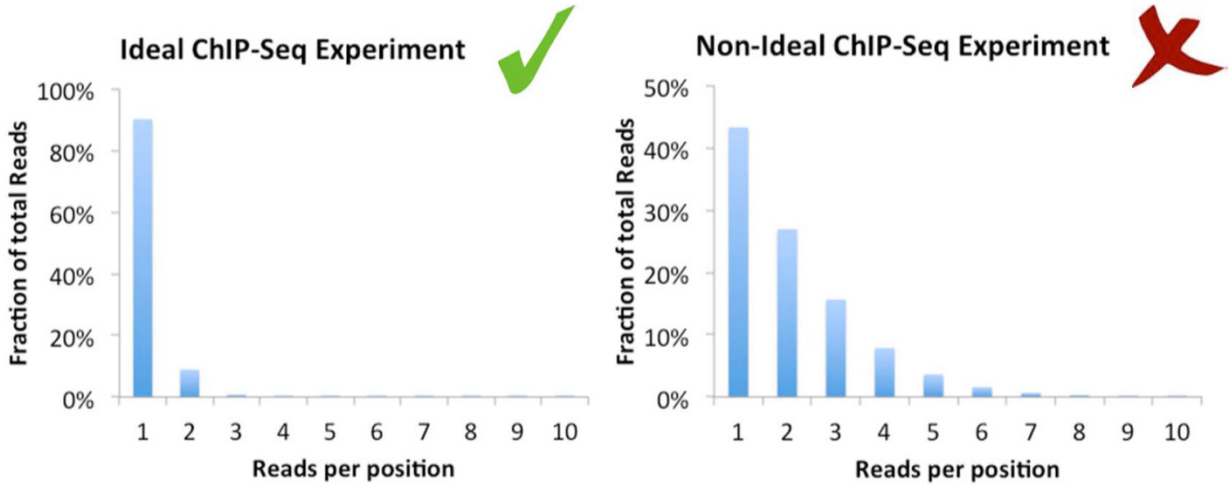
```
ILLUMINA:405:C269YACXX:1:2115:4141:76972	0	I	350017	42	46M	*	0
0	AAATAGATAATTCTTTGAAAATGAATAATTTTTTGGAAAAATCTTT	FFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJ
JJJJJJ	AS:i:0	XN:i:0	XM:i:0	XO:i:0	XG:i:0	NM:i:0	MD:Z:46	YT:Z:UU
ILLUMINA:405:C269YACXX:1:2203:10724:54489	0	I	350046	42	46M	*	0
0	TTTTTGGAAAAATCTTTTTTTCTTATTTTTCTAGACGTACACCTTT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

Quality Control

- Nonredundant fraction (NRF)

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

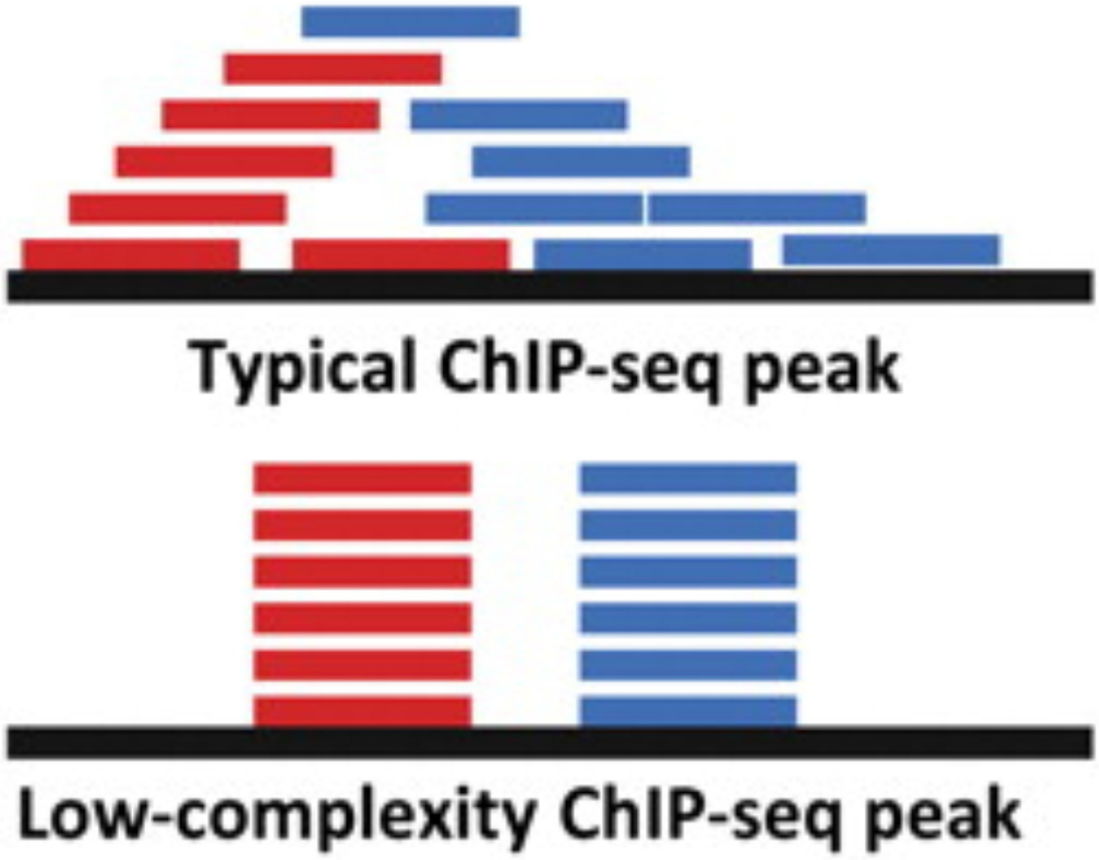
ENCODE recommends target of NRF 0:8 for 10 million uniquely mapped reads



sh run_bam2bed.sh
perl CalbedNRF.pl rep5_D12K4.txt_trim_uniq_sorted_bamtobed.bed

https://github.com/mel-astar/mel-ngs/tree/master/mel-chipseq/chipseq-metrics

Quality Control

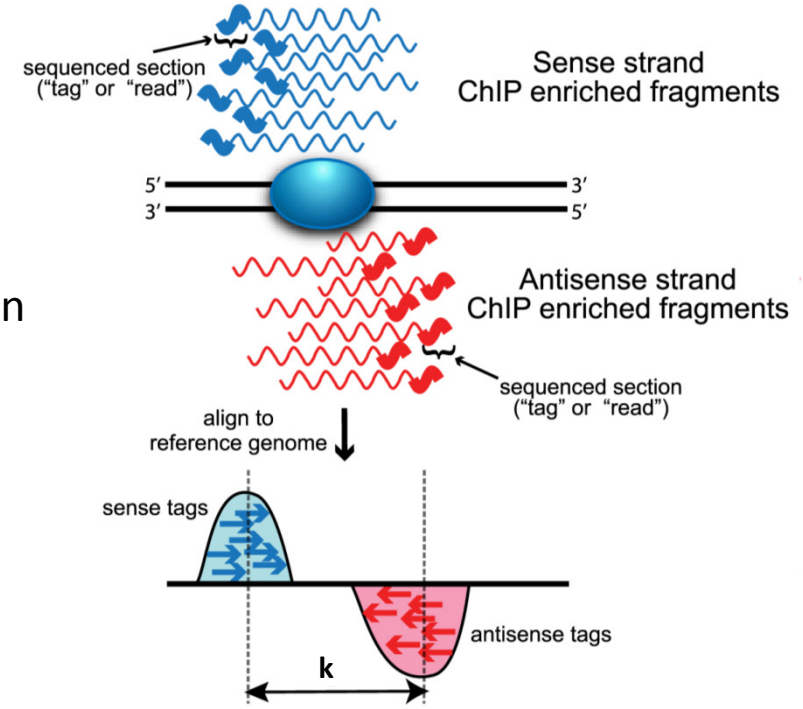


Stephen G. Landt (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res 22: 1813-1831

Cross-correlation

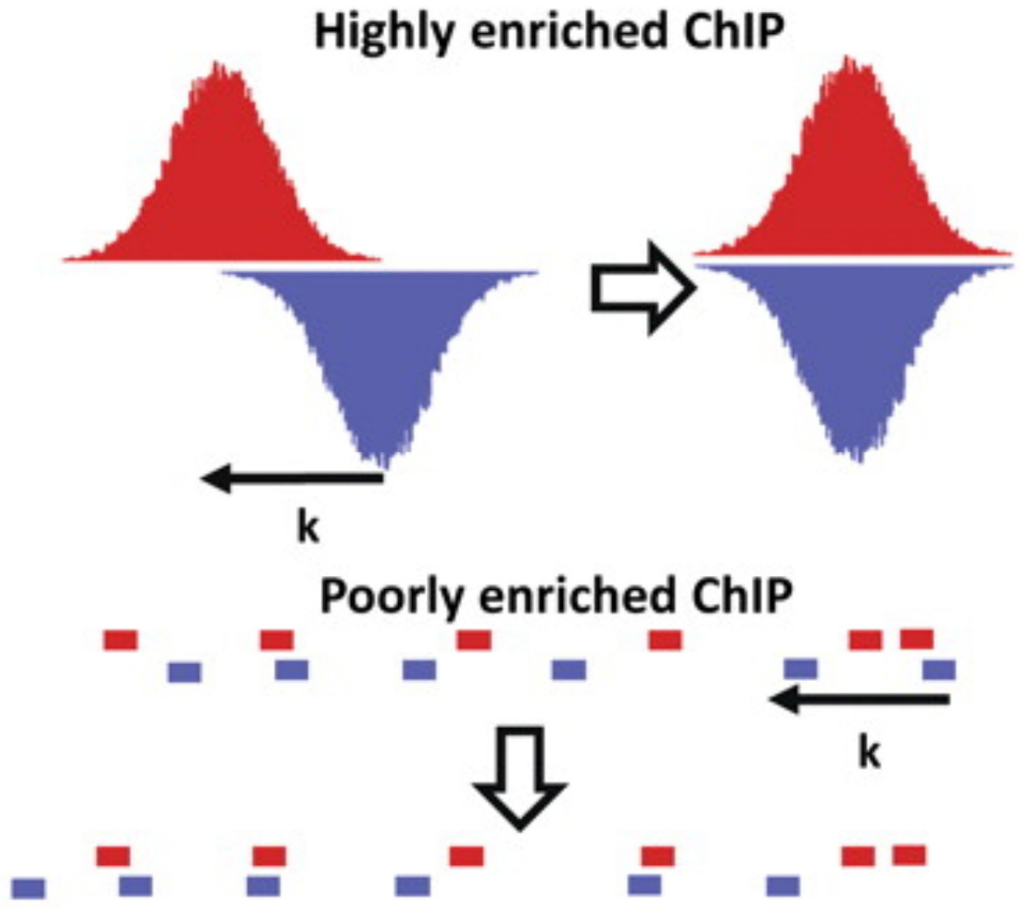
DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5' end.

- 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.



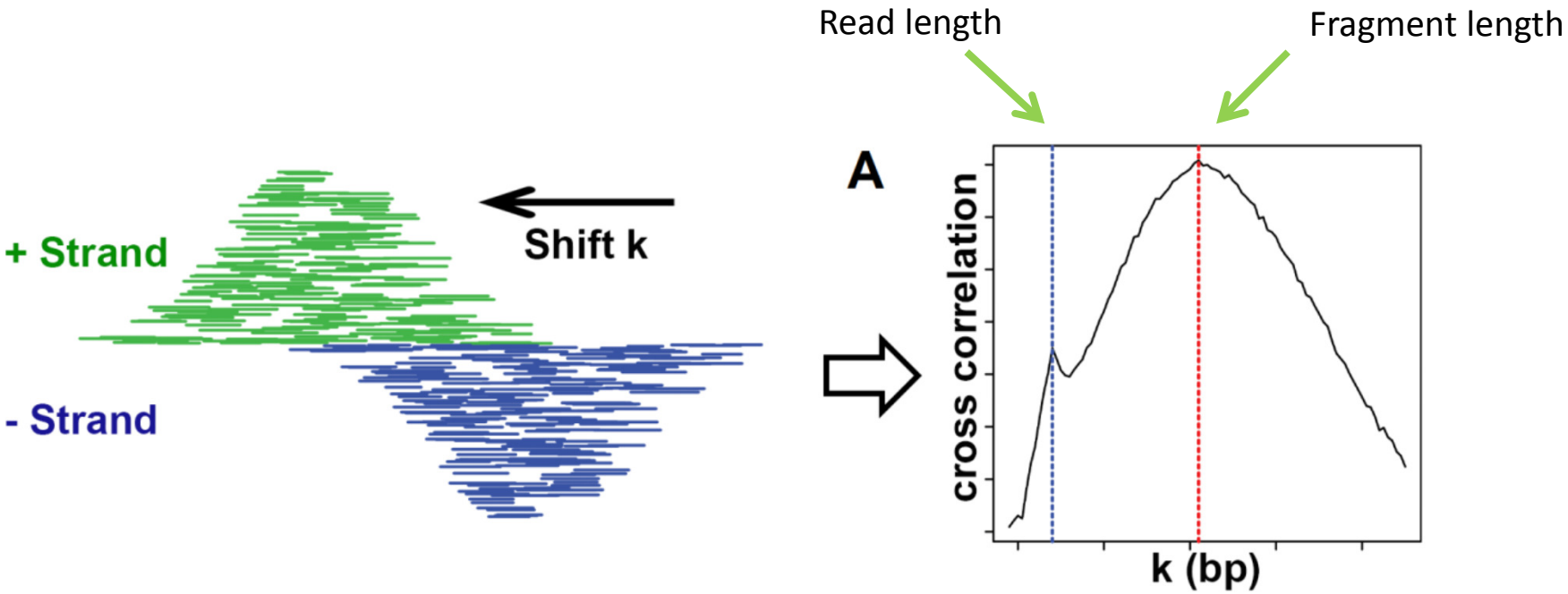
Wilbanks EG (2010) Evaluation of Algorithm Performance in ChIP-Seq Peak Detection PLoS ONE 5:e11471

Cross-correlation



Stephen G. Landt (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res 22: 1813-1831

Cross-correlation

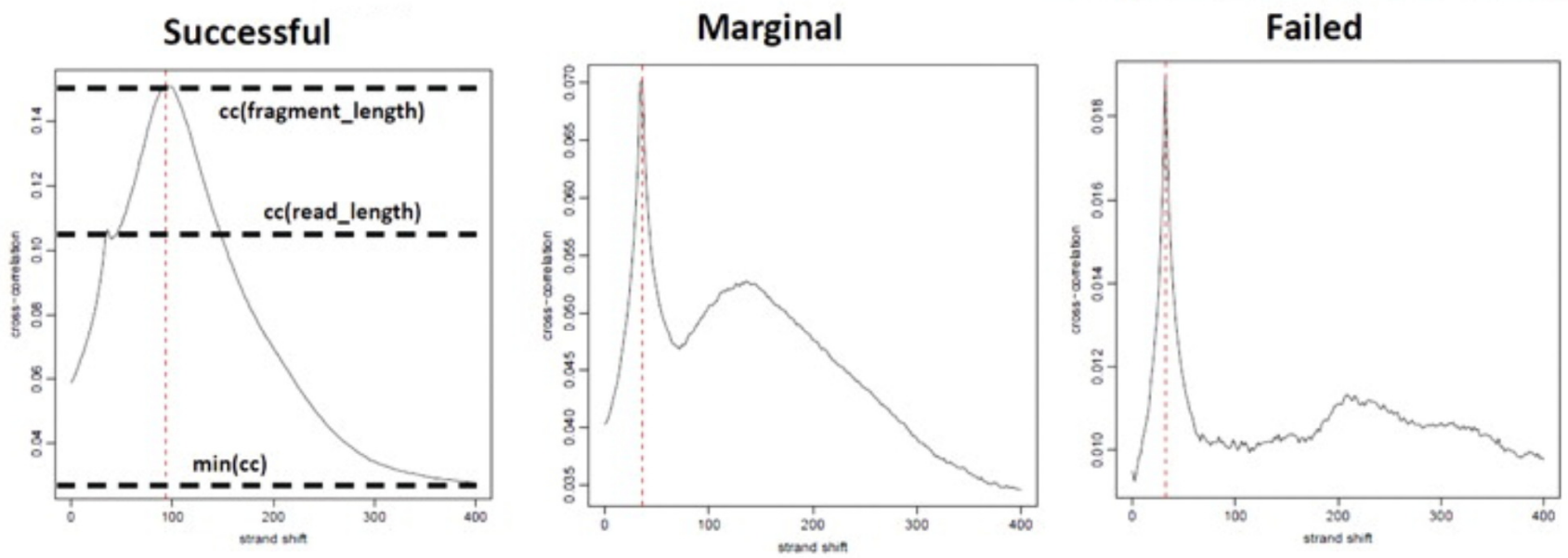


Strand cross-correlation is computed as the Pearson correlation between the positive and the negative strand profiles at different strand shift distances, k

https://sites.google.com/a/brown.edu/bioinformatics-in-biomed/spp-r-from-chip-seq

Bailey, et al (2013). Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data, PLOS Computational Biology

Cross-correlation



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

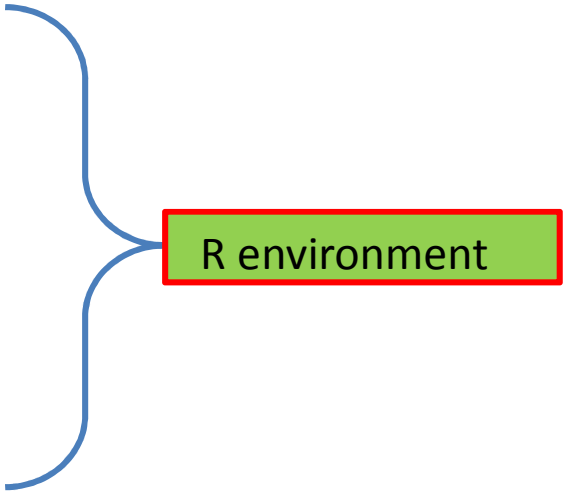
NSC values < 1.05 and RSC values < 0.8

http://code.google.com/p/phantompeakqualtools/

Stephen G. Landt (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res 22: 1813-1831

Peak calling software

- 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
- …



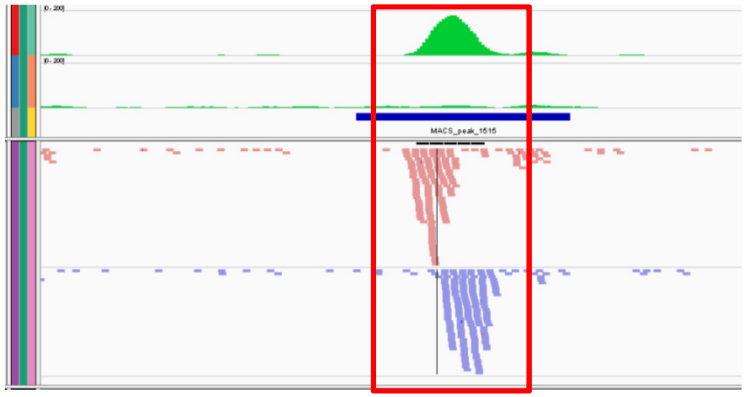
Step1 of MACS2

Estimating fragment length d

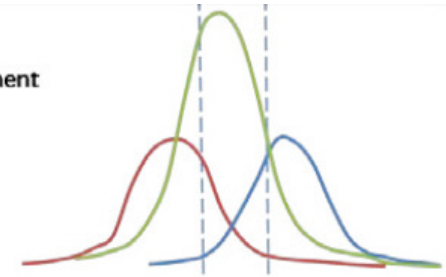
Slide a window of size 2 x *BANDWIDTH*, this value based on sonication size first

Keep top regions with MFOLD enrichment of treatment vs. control

Plot average +/- strand read densities → estimate d



Reads are shifted by d/2 toward the 3' ends, fragment are then added



Algorithm 1 Estimate Fragment Size

1: `Slide a window of` $2 \times \text{bandwidth}$ `across genome`
2: `Identify regions of moderate enrichment (mfold: 10-30 fold)`
3: **for each** peak i of 1000 randomly chosen enriched regions **do**
4: `  separate reads into + and - strand`
5: `  Calculate mode of + and - summit`
6: $d_i \leftarrow |\text{mode}_+ - \text{mode}_-|$
7: **end for**
8: $d \leftarrow \text{average}_i(d_i)$

Zhang Y (2008) Model-based Analysis of ChIP-Seq (MACS) Genome Biology 9:R137

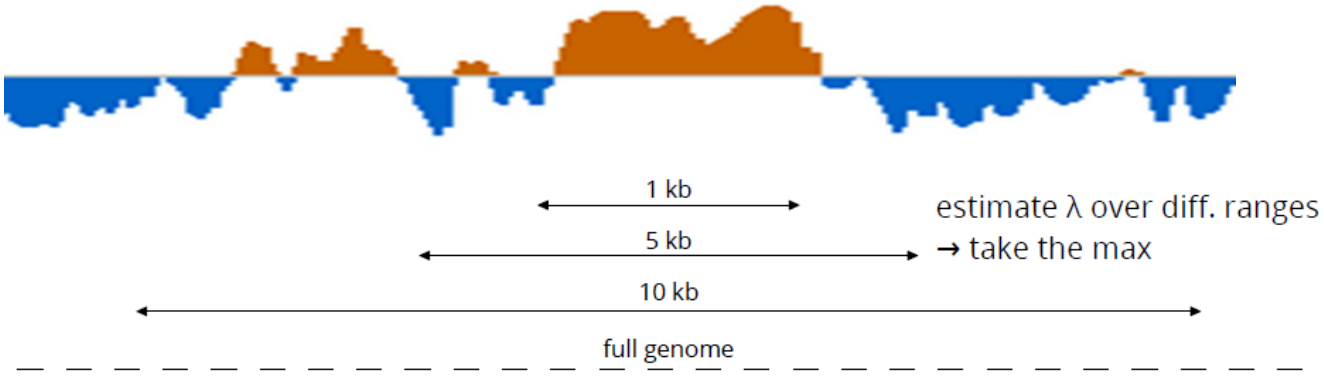
Step2 of MACS2

- Identification of local noise parameter

  shifting all reads by d/2

  slide a window of size 2*d across treatment and input

  estimate parameter λlocal of Poisson distribution



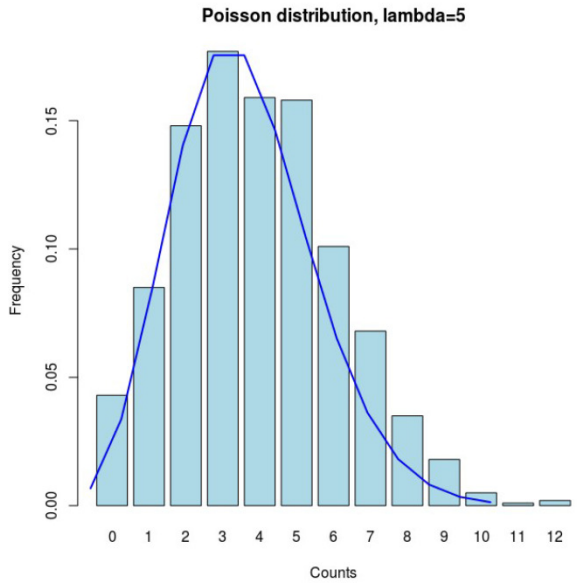
Step3 of MACS2

- Peaks identification

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

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

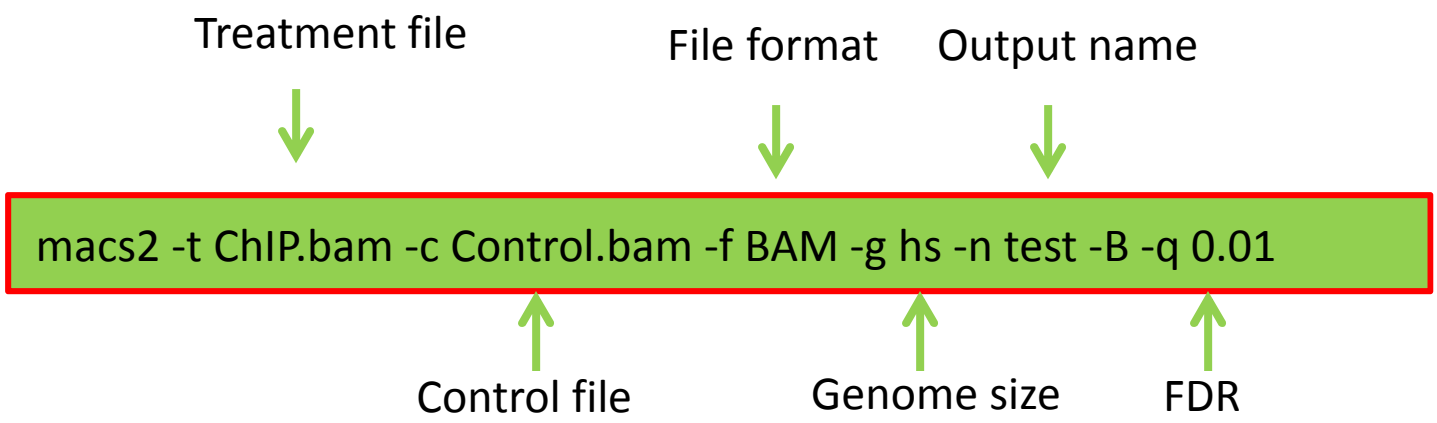
$$p(j) \leq \delta \frac{j}{m}$$



http://liulab.dfci.harvard.edu/MACS/index.html

Zhang Y (2008) Model-based Analysis of ChIP-Seq (MACS) Genome Biology 9:R137

Command of MACS2



Option:

-s TSIZE, tsize=TSIZE

-m MFOLD, --mfold=MFOLD

--bw=BW

--nomodel

--shiftsize=SHIFTSIZE

Output of MACS2

```
# This file is generated by MACS version 2.0.9 20111102 (tag:alpha)
# ARGUMENTS LIST:
# name = rep3_D2K4_H3
# format = AUTO
# ChIP-seq file = ../rep3_D2K4.txt_trim_uniq_sorted.bam
# control file = ../combinerep3_D2H3_sorted.bam
# effective genome size = 9.00e+07
# band width = 150
# 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 46 bps
# total tags in treatment: 8568994
# tags after filtering in treatment: 7814916
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.09
# total tags in control: 28632645
# tags after filtering in control: 21760444
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.24
# d = 150
chr     start   end     length  abs_summit      pileup  -log10(pvalue)  fold_enrichment -log10(qvalue)
I       4060    4291    232     4148    51.00   15.47   3.92    13.65
I       16621   16867   247     16728   45.00   11.98   3.45    10.24
I       24154   24398   245     24267   50.00   14.87   3.84    13.06
I       24563   24868   306     24703   80.00   34.67   5.86    32.50
I       26425   27627   1203    26700   97.00   48.60   7.11    46.21
I       28284   28442   159     28355   44.00   7.72    2.55    6.13
I       30982   31131   150     31068   37.00   7.84    2.84    6.24
I       31802   32130   329     31899   46.00   10.08   2.98    8.41
I       33713   33899   187     33757   44.00   7.72    2.55    6.13
I       34606   35205   600     35057   52.00   9.08    2.59    7.44
```

Input files and parameters setting

Peaks information

Output of MACS2

```
[mingh@cbsumm11 H3K4]$ more rep3_D2K4_H3_peaks.encodePeak
track type=narrowPeak nextItemButton=on
I       4059    4291    MACS_peak_1     136     .       3.92    15.47   13.65   88
I       16620   16867   MACS_peak_2     102     .       3.45    11.98   10.24   107
I       24153   24398   MACS_peak_3     130     .       3.84    14.87   13.06   113
I       24562   24868   MACS_peak_4     325     .       5.86    34.67   32.50   140
I       26424   27627   MACS_peak_5     462     .       7.11    48.60   46.21   275
I       28283   28442   MACS_peak_6     61      .       2.55    7.72    6.13    71
I       30981   31131   MACS_peak_7     62      .       2.84    7.84    6.24    86
I       31801   32130   MACS_peak_8     84      .       2.98    10.08   8.41    97
I       33712   33899   MACS_peak_9     61      .       2.55    7.72    6.13    44
I       34605   35205   MACS_peak_10    74      .       2.59    9.08    7.44    451
I       35353   35741   MACS_peak_11    97      .       3.38    11.43   9.71    78
I       36168   36391   MACS_peak_12    68      .       2.78    8.48    6.86    143
I       39389   39878   MACS_peak_13    148     .       4.07    16.71   14.86   345
I       40039   40344   MACS_peak_14    71      .       2.99    8.81    7.18    214
I       40930   41090   MACS_peak_15    53      .       2.69    6.91    5.35    69
I       46949   47213   MACS_peak_16    180     .       4.45    19.92   18.00   135
I       47288   47607   MACS_peak_17    124     .       3.76    14.27   12.47   203
I       70140   70613   MACS_peak_18    354     .       6.30    37.65   35.43   135
I       93000   93232   MACS_peak_19    62      .       2.84    7.84    6.24    100
I       97597   98073   MACS_peak_20    305     .       5.15    32.72   30.59   292
I       98224   98465   MACS_peak_21    180     .       4.45    19.92   18.00   123
I       107919  108073  MACS_peak_22    60      .       2.78    7.62    6.04    65
I       108184  109005  MACS_peak_23    202     .       4.34    22.16   20.20   604
I       109091  111927  MACS_peak_24    820     .       9.26    85.23   82.07   2142
I       171398  171571  MACS_peak_25    45      .       2.53    6.03    4.51    79
I       182407  182922  MACS_peak_26    307     .       5.83    32.90   30.76   378
I       237822  237986  MACS_peak_27    81      .       3.15    9.83    8.16    65
I       288519  289415  MACS_peak_28    496     .       7.60    52.09   49.64   299
I       310449  310912  MACS_peak_29    148     .       4.07    16.71   14.86   149
I       310963  311271  MACS_peak_30    90      .       3.12    10.75   9.06    145
I       314136  315758  MACS_peak_31    659     .       8.98    68.73   65.96   782
I       315988  316213  MACS_peak_32    81      .       3.15    9.83    8.16    90
```

Enrichment score (fold-change)

-log10pvalue

-log10qvalue

Summit position to peak start

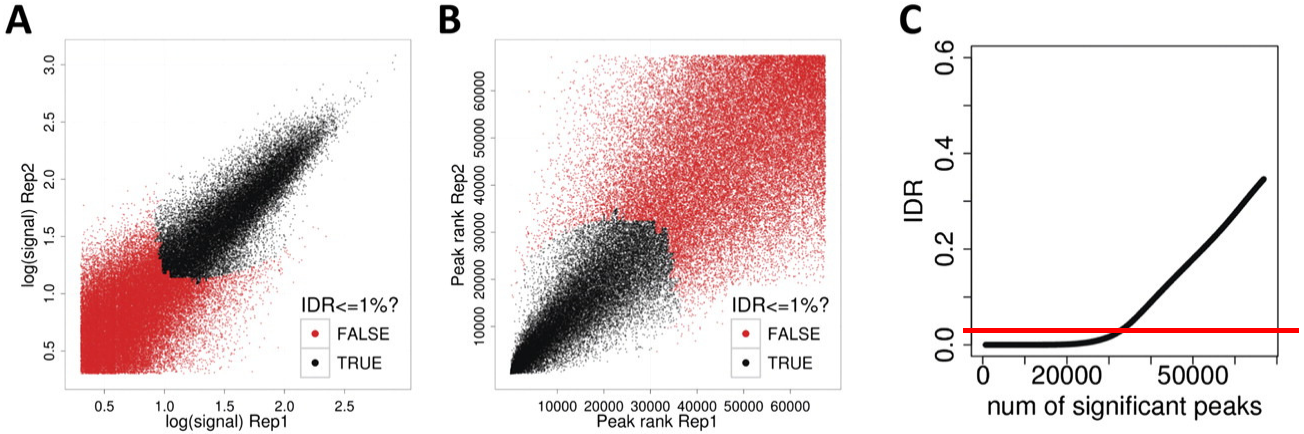
Consistency of replicates: IDR

- 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

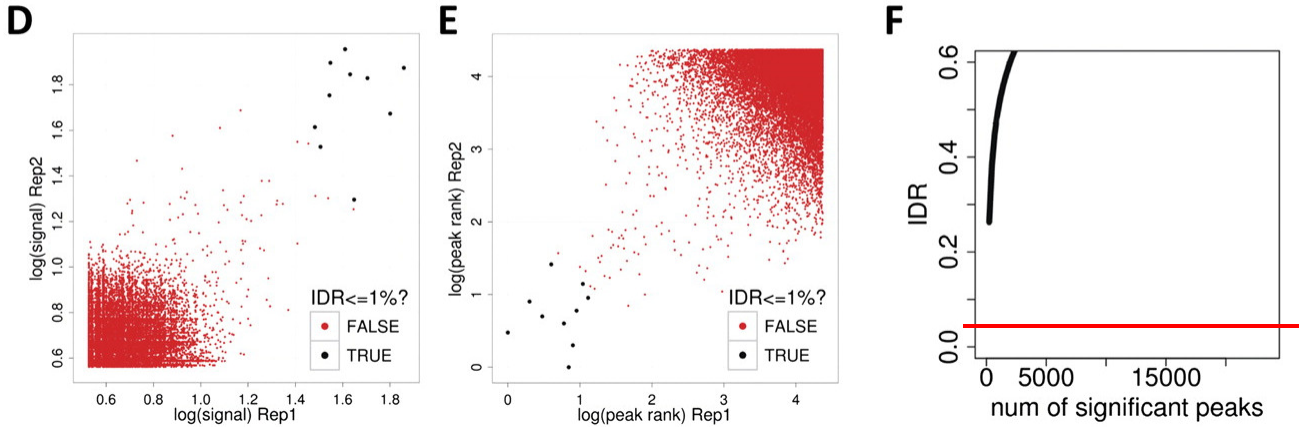
Rscript batch-consistency-analysis.r [peakfile1] [peakfile2] -1 [outfile.prefix] 0 F p.value
Rscript batch-consistency-plot.r [npairs] [output.prefix] [input.file.prefix1] [input.file.prefix2] [input.file.prefix3]

IDR

RAD21 Replicates (high reproducibility)

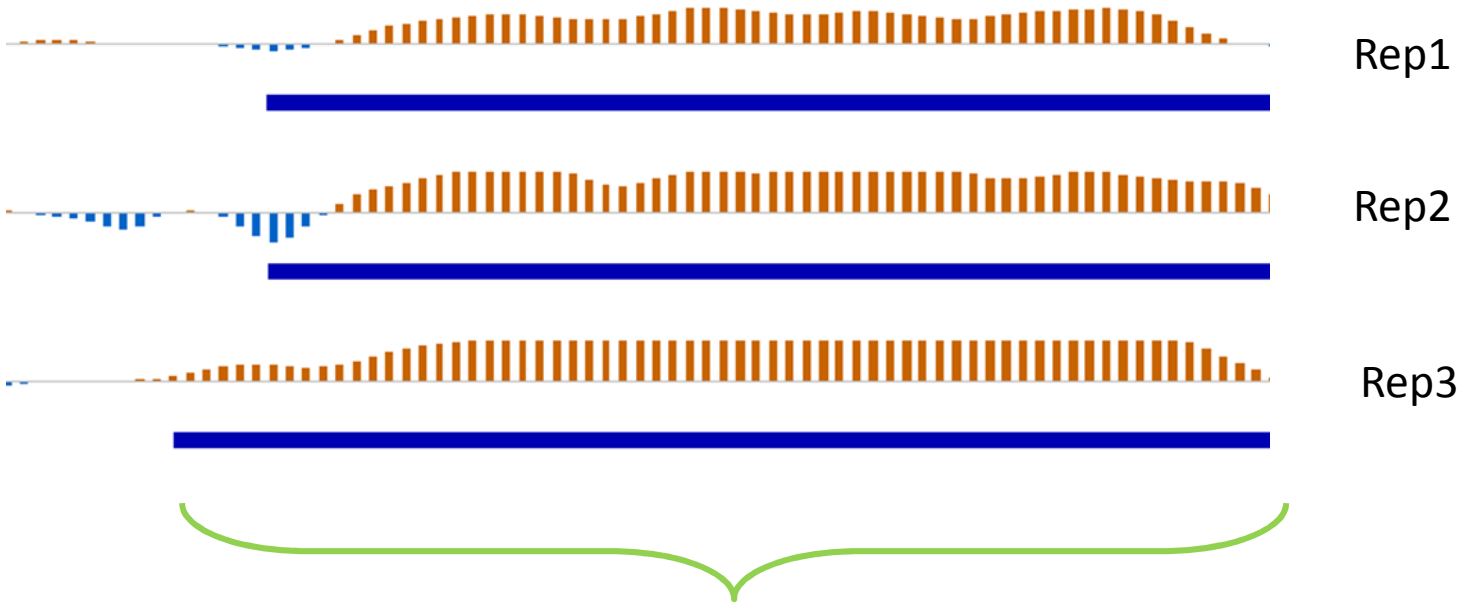


SPT20 Replicates (low reproducibility)



Stephen G. Landt (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res 22: 1813-1831

Peak region merging



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

R scripts for replicates

```
>  time=factor(c(rep(12,2),rep(2,2),rep(12,2),rep(2,2)))
>  trt=factor(rep(c("K4","H3"),4,each=1))
>  design<- model.matrix(~time*trt)
>  design
  (Intercept) time12 trtK4 time12:trtK4
1           1      1     1            1
2           1      1     0            0
3           1      0     1            0
4           1      0     0            0
5           1      1     1            1
6           1      1     0            0
7           1      0     1            0
8           1      0     0            0
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
attr(,"contrasts")$time
[1] "contr.treatment"

attr(,"contrasts")$trt
[1] "contr.treatment"
```

Comparing pairs

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