ChIP-Seq Data Analysis: Probing DNA-Protein Interactions

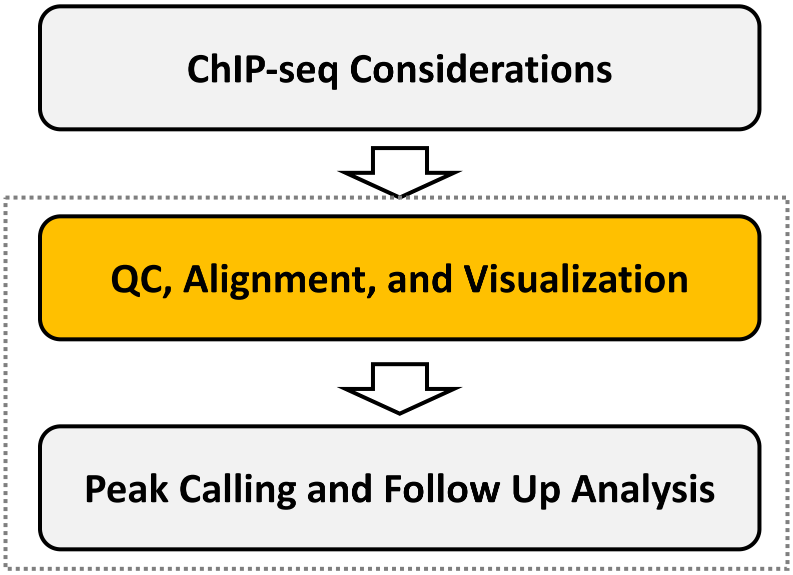
Paul Schaughency[1,2] , Tovah Markowitz[1], Vishal Koparde[3]

Schedule

9:30 - 10:15	Introduction to ChIP-Seq
10:15 - 10:30	Q&A
10:30 - 11:20	QC, Alignment, and Visualization
11:20 - 12:00	Peak Calling and Follow Up Analysis
12:00 - 12:30	Q&A

[1]NIAID Collaborative Bioinformatics Resource (NCBR), [2]Center for Cancer Research Sequencing Facility (CCR-SF) Bioinformatics, [3]Center for Cancer Research Collaborative Bioinformatics Resource (CCBR)

1



ChIP-seq Considerations

QC, Alignment, and Visualization

Peak Calling and Follow Up Analysis

2

QC, Alignment, and Visualization

Pipelines

QC and Alignment

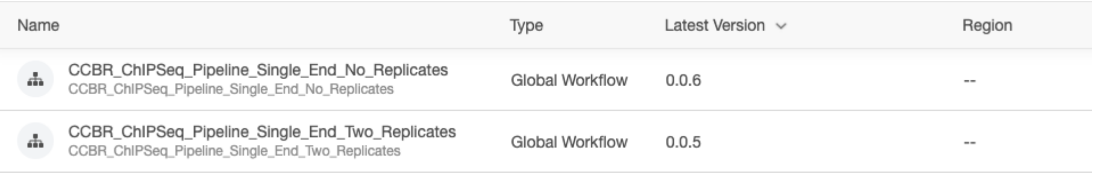
Visualization

3

ChIP-seq Pipelines

Cloud-based

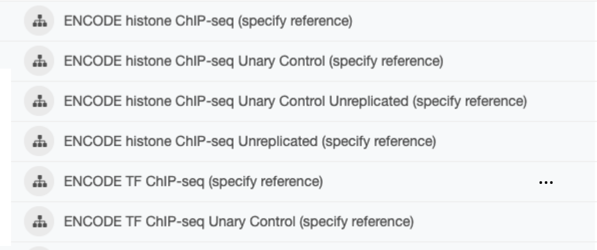
CCBR/NCI



Name	Type	Latest Version	Region
CCBR_ChIPSeq_Pipeline_Single_End_No_Replicates CCBR_ChIPSeq_Pipeline_Single_End_No_Replicates	Global Workflow	0.0.6	--
CCBR_ChIPSeq_Pipeline_Single_End_Two_Replicates CCBR_ChIPSeq_Pipeline_Single_End_Two_Replicates	Global Workflow	0.0.5	--

- Cost (Ingress/Egress/Compute/Storage)
- Secure (FISMA-Moderate)
- Share-able outside NIH

National Human Genome Research Institute ENCODE



ENCODE histone ChIP-seq (specify reference)

ENCODE histone ChIP-seq Unary Control (specify reference)

ENCODE histone ChIP-seq Unary Control Unreplicated (specify reference)

ENCODE histone ChIP-seq Unreplicated (specify reference)

ENCODE TF ChIP-seq (specify reference)

ENCODE TF ChIP-seq Unary Control (specify reference)

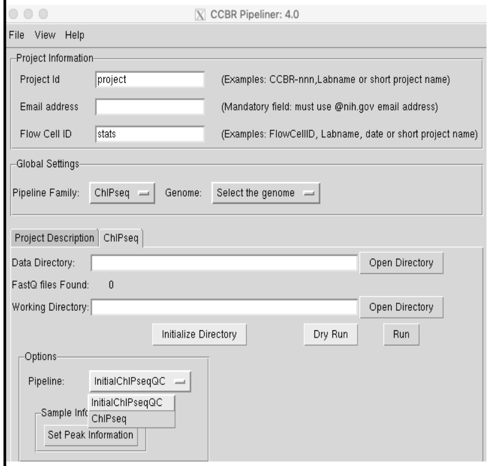
4

ChIP-seq Pipelines

Biowulf-based

CCBR/NCBR

BIOWULF
HIGH PERFORMANCE COMPUTING AT THE NIH


CCBR Pipeliner: 4.0

File View Help

Project Information

Project Id: project (Examples: CCBR-nnn,Labname or short project name)

Email address: (Mandatory field: must use @nih.gov email address)

Flow Cell ID: stats (Examples: FlowCellID, Labname, date or short project name)

Global Settings

Pipeline Family: ChIPseq Genome: Select the genome

Project Description | ChIPseq

Data Directory: Open Directory

FastQ files Found: 0

Working Directory: Open Directory

Initialize Directory | Dry Run | Run

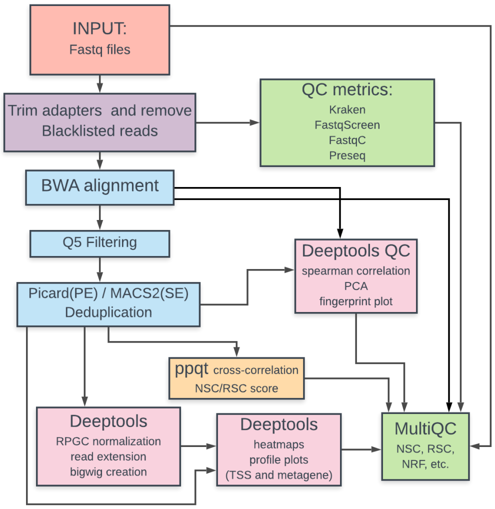
Options

Pipeline: InitialChIPseqQC

InitialChIPseqQC
ChIPseq

Sample Info

Set Peak Information


INPUT: Fastq files

Trim adapters and remove Blacklisted reads

QC metrics: Kraken, FastqScreen, FastqC, Preseq

BWA alignment

Q5 Filtering

Deeptools QC: spearman correlation, PCA, fingerprint plot

Picard(PE) / MACS2(SE) Deduplication

ppqt cross-correlation NSC/RSC score

Deeptools: RPGC normalization, read extension, bigwig creation

Deeptools: heatmaps, profile plots (TSS and metagene)

MultiQC: NSC, RSC, NRF, etc.

Genomes supported

- hg19
- hg38
- mm9
- mm10

5

QC, Alignment, and Visualization

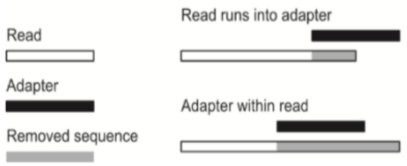
Pipelines

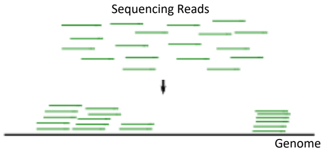
QC and Alignment

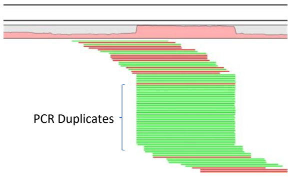
Visualization

6

Pipeline: Basic bioinformatics concepts







Base quality trimming
+
Adapter removal

Alignment or “Mapping”

Deduplication or
Duplicate removal

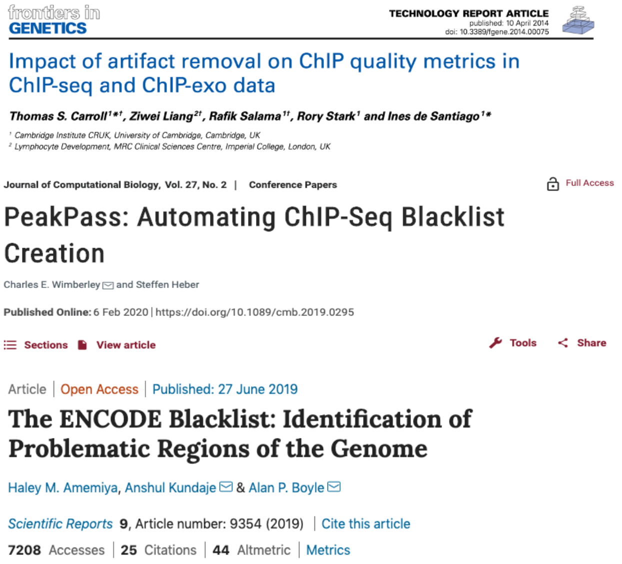
CutAdapt

BWA

Picardtools or MACS2

7

Blacklists



frontiers in GENETICS

TECHNOLOGY REPORT ARTICLE
published: 10 April 2014
doi: 10.3389/fgene.2014.00075

Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data

Thomas S. Carroll[1*†], Ziwei Liang[2†], Rafik Salama[1†], Rory Stark[1] and Ines de Santiago[1*]

[1] Cambridge Institute CRUK, University of Cambridge, Cambridge, UK
[2] Lymphocyte Development, MRC Clinical Sciences Centre, Imperial College, London, UK

Journal of Computational Biology, Vol. 27, No. 2 | Conference Papers

Full Access

PeakPass: Automating ChIP-Seq Blacklist Creation

Charles E. Wimberley and Steffen Heber

Published Online: 6 Feb 2020 | https://doi.org/10.1089/cmb.2019.0295

Article | Open Access | Published: 27 June 2019

The ENCODE Blacklist: Identification of Problematic Regions of the Genome

Haley M. Amemiya, Anshul Kundaje & Alan P. Boyle

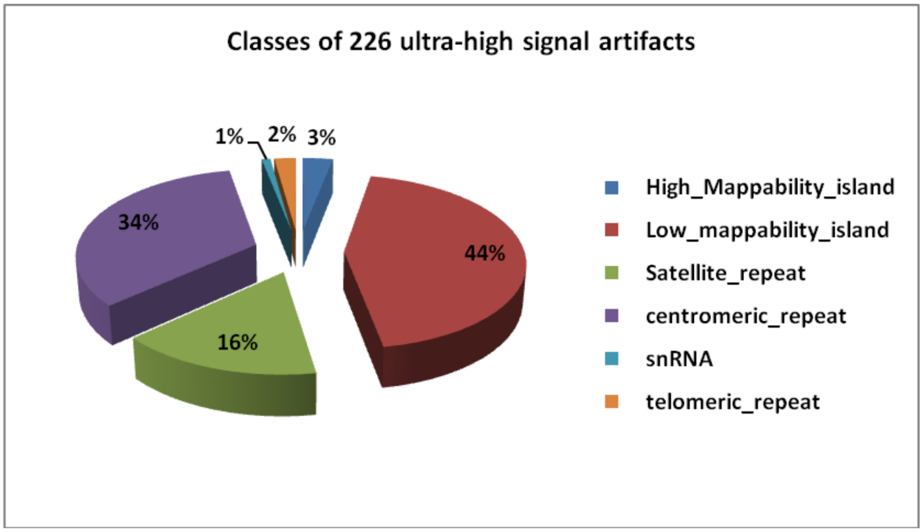
Scientific Reports 9, Article number: 9354 (2019) | Cite this article

7208 Accesses | 25 Citations | 44 Altmetric | Metrics

- ChIP-Seq blacklists contain genomic regions that frequently produce artifacts and noise in ChIP-Seq experiments.
- Remove reads to these regions to improve signal-to-noise ratio
- Reference genome specific lists are calculated in a manually curated + automated manner

8

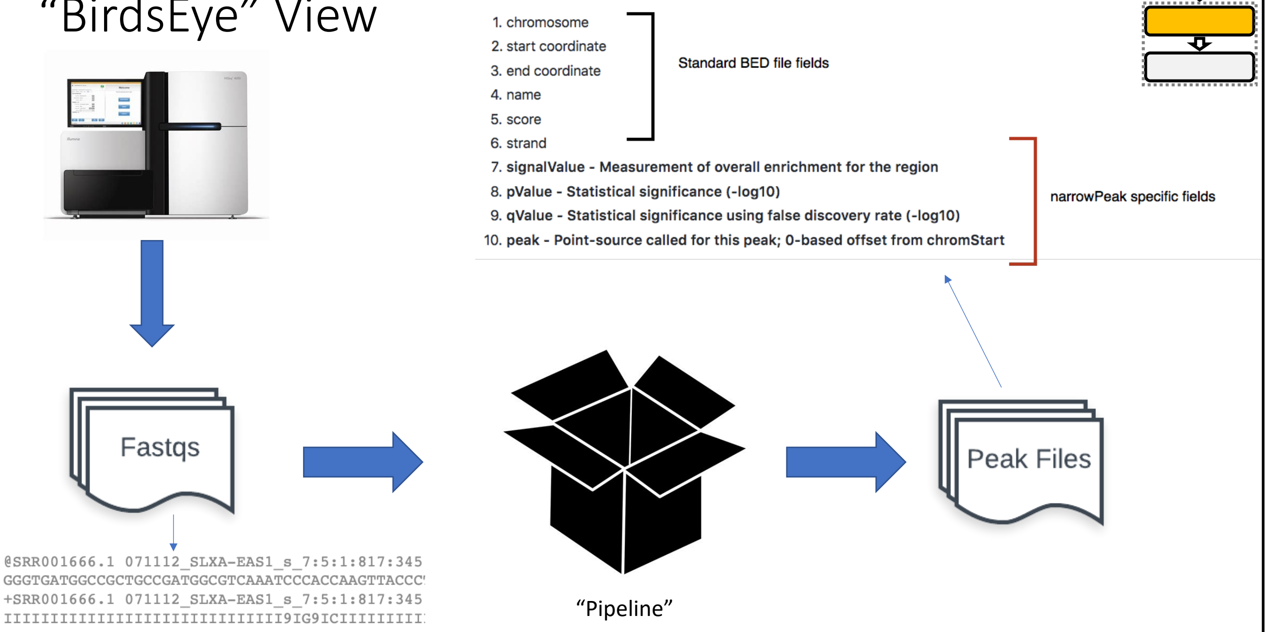
Blacklists



"A comprehensive collection of signal artifact blacklist regions in the human genome", by Anshul Kundaje

9

"BirdsEye" View



1. chromosome
2. start coordinate
3. end coordinate
4. name
5. score
6. strand
7. **signalValue - Measurement of overall enrichment for the region**
8. **pValue - Statistical significance (-log10)**
9. **qValue - Statistical significance using false discovery rate (-log10)**
10. **peak - Point-source called for this peak; 0-based offset from chromStart**

Standard BED file fields

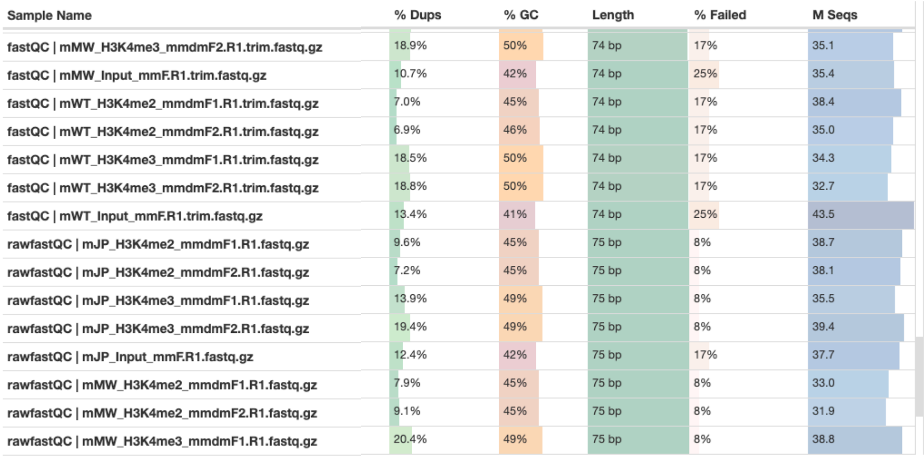
narrowPeak specific fields

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACCAAGTTACCC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9ICIIIIIIIIII
```

10

MultiQC report

General Stats



Sample Name	% Dups	% GC	Length	% Failed	M Seqs
fastQC | mMW_H3K4me3_mmdmF2.R1.trim.fastq.gz	18.9%	50%	74 bp	17%	35.1
fastQC | mMW_Input_mmF.R1.trim.fastq.gz	10.7%	42%	74 bp	25%	35.4
fastQC | mWT_H3K4me2_mmdmF1.R1.trim.fastq.gz	7.0%	45%	74 bp	17%	38.4
fastQC | mWT_H3K4me2_mmdmF2.R1.trim.fastq.gz	6.9%	46%	74 bp	17%	35.0
fastQC | mWT_H3K4me3_mmdmF1.R1.trim.fastq.gz	18.5%	50%	74 bp	17%	34.3
fastQC | mWT_H3K4me3_mmdmF2.R1.trim.fastq.gz	18.8%	50%	74 bp	17%	32.7
fastQC | mWT_Input_mmF.R1.trim.fastq.gz	13.4%	41%	74 bp	25%	43.5
rawfastQC | mJP_H3K4me2_mmdmF1.R1.fastq.gz	9.6%	45%	75 bp	8%	38.7
rawfastQC | mJP_H3K4me2_mmdmF2.R1.fastq.gz	7.2%	45%	75 bp	8%	38.1
rawfastQC | mJP_H3K4me3_mmdmF1.R1.fastq.gz	13.9%	49%	75 bp	8%	35.5
rawfastQC | mJP_H3K4me3_mmdmF2.R1.fastq.gz	19.4%	49%	75 bp	8%	39.4
rawfastQC | mJP_Input_mmF.R1.fastq.gz	12.4%	42%	75 bp	17%	37.7
rawfastQC | mMW_H3K4me2_mmdmF1.R1.fastq.gz	7.9%	45%	75 bp	8%	33.0
rawfastQC | mMW_H3K4me2_mmdmF2.R1.fastq.gz	9.1%	45%	75 bp	8%	31.9
rawfastQC | mMW_H3K4me3_mmdmF1.R1.fastq.gz	20.4%	49%	75 bp	8%	38.8

Key parameters:

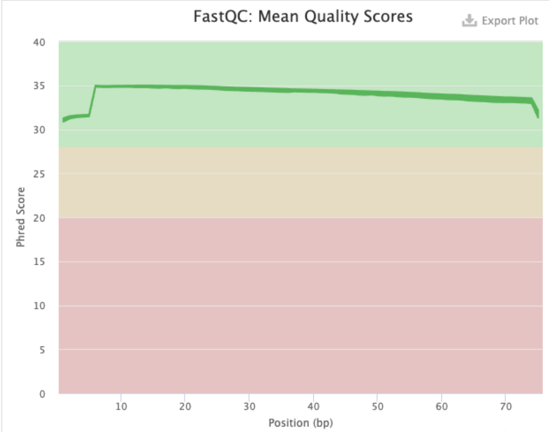
- Number of reads
- GC
- Mapping percentage

11

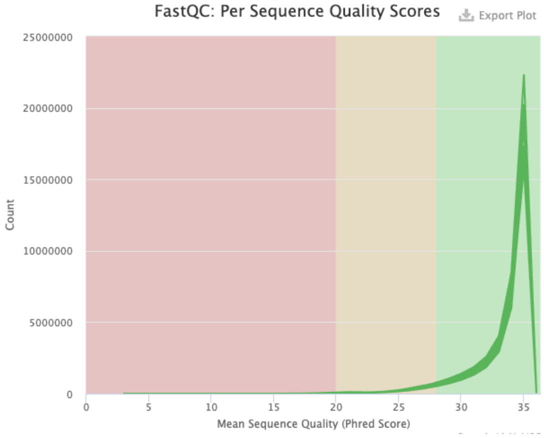
MultiQC report

Sequence-related metrics → FASTQC

Sequence Quality



Per Sequence Quality Scores

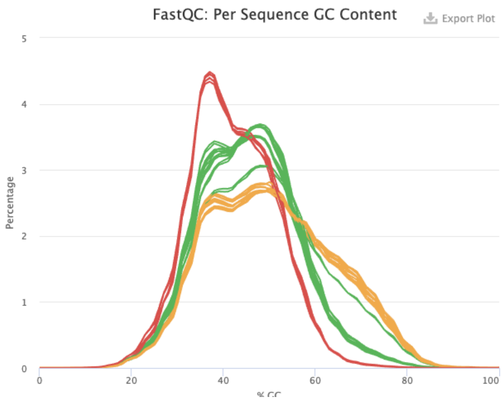


12

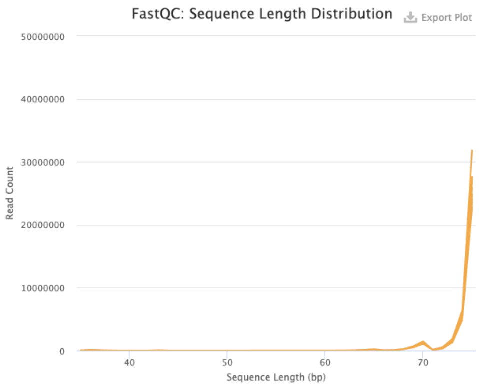
MultiQC report

Sequence-related metrics → FASTQC

GC Content



Length Distribution

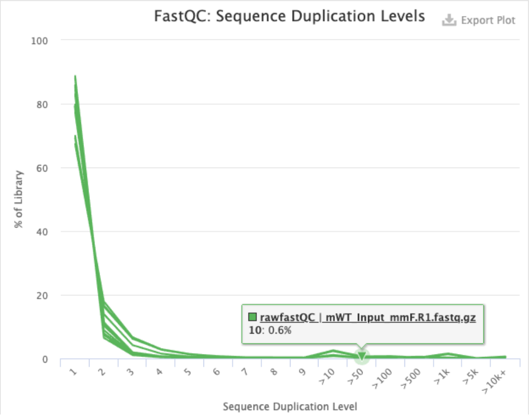


13

MultiQC report

Sequence-related metrics → FASTQC

Duplication Levels



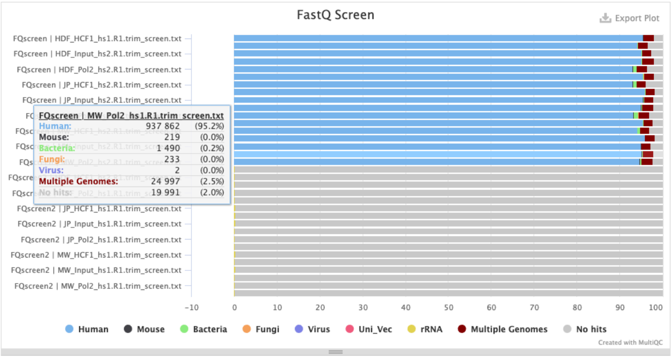
Other metrics:

- Over-represented sequences
- Adapter Content

14

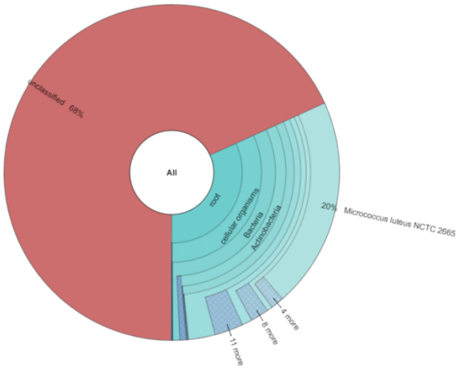
MultiQC report: Contaminants

FastQ screen


FastQ Screen
Export Plot
FQscreen | HDF_HCF1_hs1.R1.trim_screen.txt
FQscreen | HDF_Input_hs2.R1.trim_screen.txt
FQscreen | HDF_Pol2_hs2.R1.trim_screen.txt
FQscreen | JP_HCF1_hs2.R1.trim_screen.txt
FQscreen | JP_Input_hs2.R1.trim_screen.txt
FQscreen2 | JP_HCF1_hs1.R1.trim_screen.txt
FQscreen2 | JP_Input_hs1.R1.trim_screen.txt
FQscreen2 | JP_Pol2_hs1.R1.trim_screen.txt
FQscreen2 | MW_HCF1_hs1.R1.trim_screen.txt
FQscreen2 | MW_Input_hs1.R1.trim_screen.txt
FQscreen2 | MW_Pol2_hs1.R1.trim_screen.txt
-10
0
10
20
30
40
50
60
70
80
90
100
● Human
● Mouse
● Bacteria
● Fungi
● Virus
● Uni_Vec
● rRNA
● Multiple Genomes
● No hits
Created with MultiQC

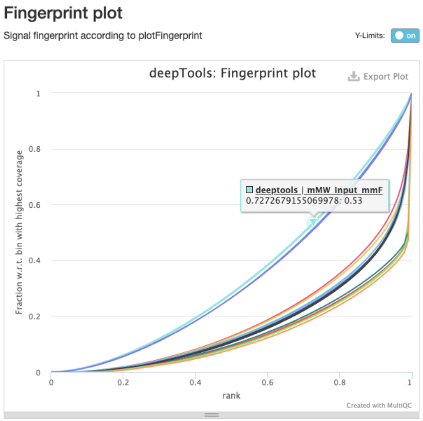
```
FQscreen | MW_Pol2_hs1.R1.trim_screen.txt
Human:              937 862    (95.2%)
Mouse:                  219     (0.0%)
Bacteria:             1 490     (0.2%)
Fungi:                  233     (0.0%)
Virus:                    2     (0.0%)
Multiple Genomes:    24 997     (2.5%)
No hits:             19 991     (2.0%)
```

Kraken + Krona



MultiQC report

Fingerprint plot



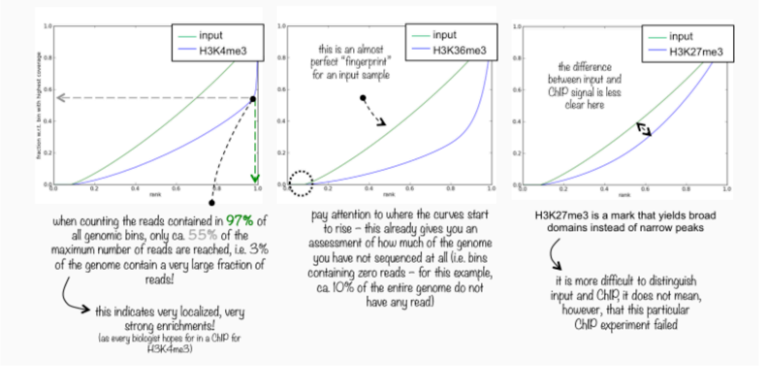
Fingerprint plot

Signal fingerprint according to plotFingerprint

Y-Limits: on

deepTools: Fingerprint plot
Export Plot
Fraction w.r.t. bin with highest coverage
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
rank
Created with MultiQC

```
deeptools | mMW_Input_mmF
0.7272679155069978: 0.53
```



this is an almost perfect "fingerprint" for an input sample

the difference between input and ChIP signal is less clear here

when counting the reads contained in 97% of all genomic bins, only ca. 55% of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

this indicates very localized, very strong enrichments! (as every biologist hopes for in a ChIP for H3K4me3)

pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads – for this example, ca. 10% of the entire genome do not have any read)

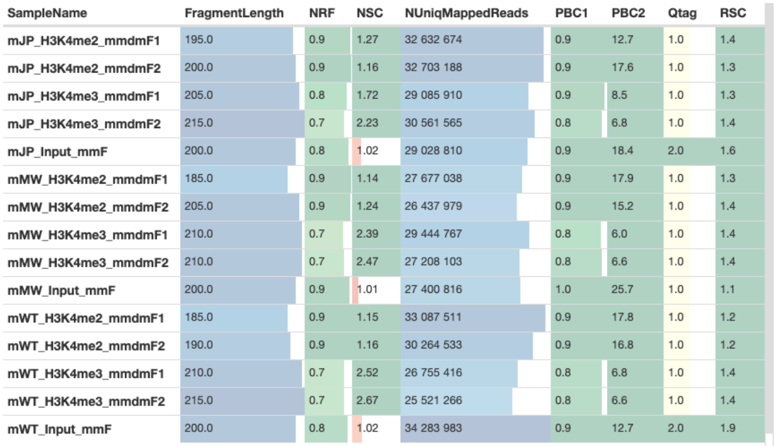
H3K27me3 is a mark that yields broad domains instead of narrow peaks

it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed

- Answers the question "Did my ChIP work?"
- Input close to 45⁰ as possible
- Input above IP
- Broad histones → farther away from 45⁰

MultiQC report

ChIPSeq specific metrics



SampleName	FragmentLength	NRF	NSC	NUniqMappedReads	PBC1	PBC2	Qtag	RSC
mJP_H3K4me2_mmdmF1	195.0	0.9	1.27	32 632 674	0.9	12.7	1.0	1.4
mJP_H3K4me2_mmdmF2	200.0	0.9	1.16	32 703 188	0.9	17.6	1.0	1.3
mJP_H3K4me3_mmdmF1	205.0	0.8	1.72	29 085 910	0.9	8.5	1.0	1.3
mJP_H3K4me3_mmdmF2	215.0	0.7	2.23	30 561 565	0.8	6.8	1.0	1.4
mJP_Input_mmF	200.0	0.8	1.02	29 028 810	0.9	18.4	2.0	1.6
mMW_H3K4me2_mmdmF1	185.0	0.9	1.14	27 677 038	0.9	17.9	1.0	1.3
mMW_H3K4me2_mmdmF2	205.0	0.9	1.24	26 437 979	0.9	15.2	1.0	1.4
mMW_H3K4me3_mmdmF1	210.0	0.7	2.39	29 444 767	0.8	6.0	1.0	1.4
mMW_H3K4me3_mmdmF2	210.0	0.7	2.47	27 208 103	0.8	6.6	1.0	1.4
mMW_Input_mmF	200.0	0.9	1.01	27 400 816	1.0	25.7	1.0	1.1
mWT_H3K4me2_mmdmF1	185.0	0.9	1.15	33 087 511	0.9	17.8	1.0	1.2
mWT_H3K4me2_mmdmF2	190.0	0.9	1.16	30 264 533	0.9	16.8	1.0	1.2
mWT_H3K4me3_mmdmF1	210.0	0.7	2.52	26 755 416	0.8	6.8	1.0	1.4
mWT_H3K4me3_mmdmF2	215.0	0.7	2.67	25 521 266	0.8	6.6	1.0	1.4
mWT_Input_mmF	200.0	0.8	1.02	34 283 983	0.9	12.7	2.0	1.9

Quantifying library complexity

- **NRF:** Number of distinct mapping reads after removing duplicates/total number of reads
- **PBC1:** Number of genomic locations where exactly one read maps uniquely/number of distinct genomic locations to which one read maps uniquely
- **PBC2**: Number of genomic locations where only one read maps uniquely/number of genomic locations where two reads map uniquely

Quantifying CrossCorrelation

- **NSC:** cross-correlation value/minimum cross-correlation
- **RSC**: (cross-correlation value - minimum cross-correlation) / (correlation at phantom peak - minimum cross-correlation)
- **Qtag**: Overall Quality score

17

Library Complexity

National Human Genome Research Institute

ENCODE

ENCODE guidelines

PBC1	PBC2	Bottlenecking level	NRF	Complexity	Flag colors
< 0.5	< 1	Severe	< 0.5	Concerning	Orange
0.5 ≤ PBC1 < 0.8	1 ≤ PBC2 < 3	Moderate	0.5 ≤ NRF < 0.8	Acceptable	Yellow
0.8 ≤ PBC1 < 0.9	3 ≤ PBC2 < 10	Mild	0.8 ≤ NRF < 0.9	Compliant	None
≥ 0.9	≥ 10	None	> 0.9	Ideal	None

PCR Bottlenecking Coefficient 1 (PBC1)

- PBC1=$M_1/M_{DISTINCT}$ where
  - M_1: number of genomic locations where exactly one read maps uniquely
  - $M_{DISTINCT}$: number of distinct genomic locations to which some read maps uniquely

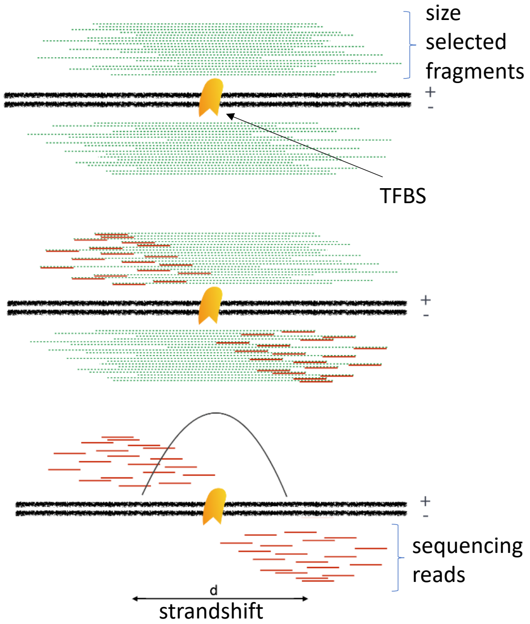
PCR Bottlenecking Coefficient 2 (PBC2)

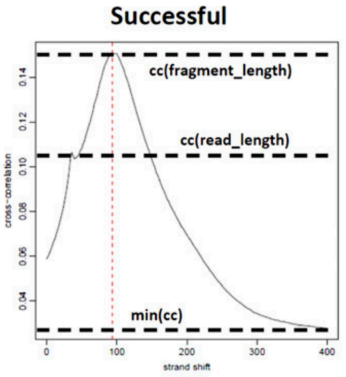
- PBC2=M_1/M_2 where
  - M_1: number of genomic locations where only one read maps uniquely
  - M_2: number of genomic locations where two reads map uniquely

Non-Redundant Fraction (NRF) – Number of distinct uniquely mapping reads (i.e. after removing duplicates) / Total number of reads.

18

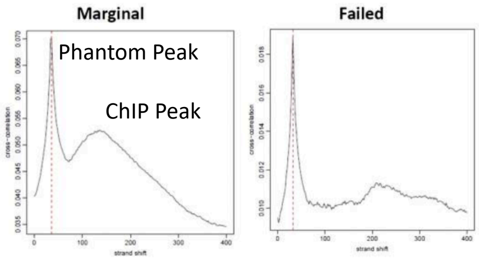
Cross correlation





$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

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



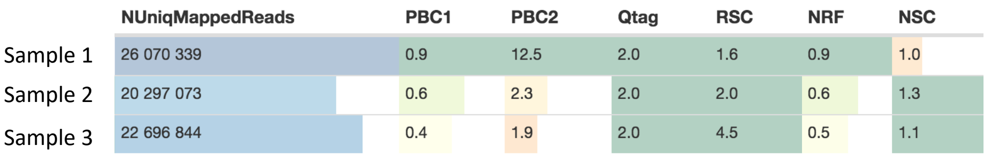
Landt et al 2012. Genome Res

Cross correlation

NSC	Comments
> 1.1	Optimal
1.05 - 1.1	Acceptable
< 1.05	low signal to noise or very few peaks (biological or technical)

RSC	Comments
> 1	Optimal
0.8 - 1	Acceptable
< 0.8	poor quality ChIP, low read sequence quality, shallow sequencing depth, or few peaks

NSC → Normalized Strand Co-efficient
RSC → Relative Strand Co-efficient
Qtag → is a thresholded version of RSC (-2:veryLow, -1:Low, 0:Medium, 1:High, 2:veryHigh)

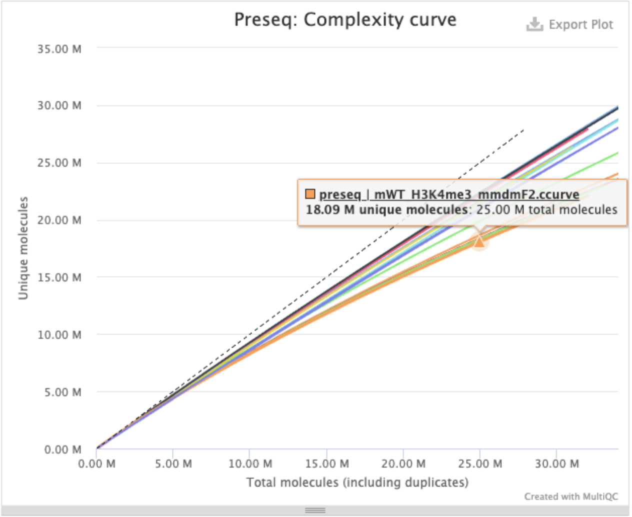


	NUniqMappedReads	PBC1	PBC2	Qtag	RSC	NRF	NSC
Sample 1	26 070 339	0.9	12.5	2.0	1.6	0.9	1.0
Sample 2	20 297 073	0.6	2.3	2.0	2.0	0.6	1.3
Sample 3	22 696 844	0.4	1.9	2.0	4.5	0.5	1.1

MultiQC report:
More library complexity

Preseq

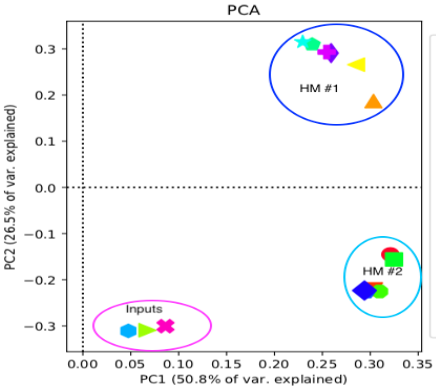
- Answers the question: Do "more" sequences mean "new" sequences?
- Inputs are expected to be closer to the dotted line than corresponding IP-ed sample



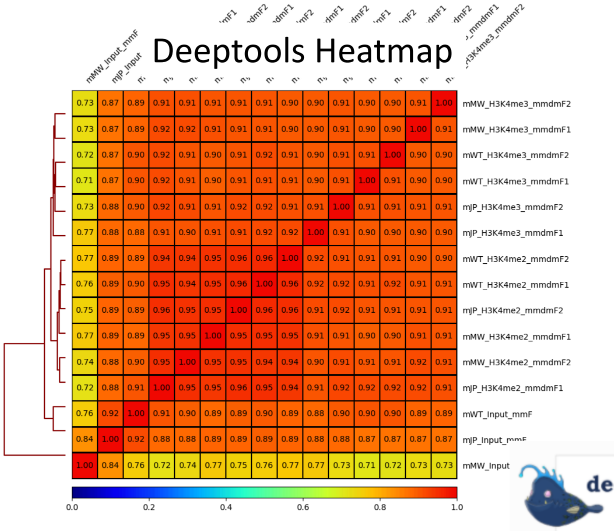
MultiQC report:
inter-sample comparison

Deeptools PCA

- "Inputs" are generally together
- Verify replicate concordance

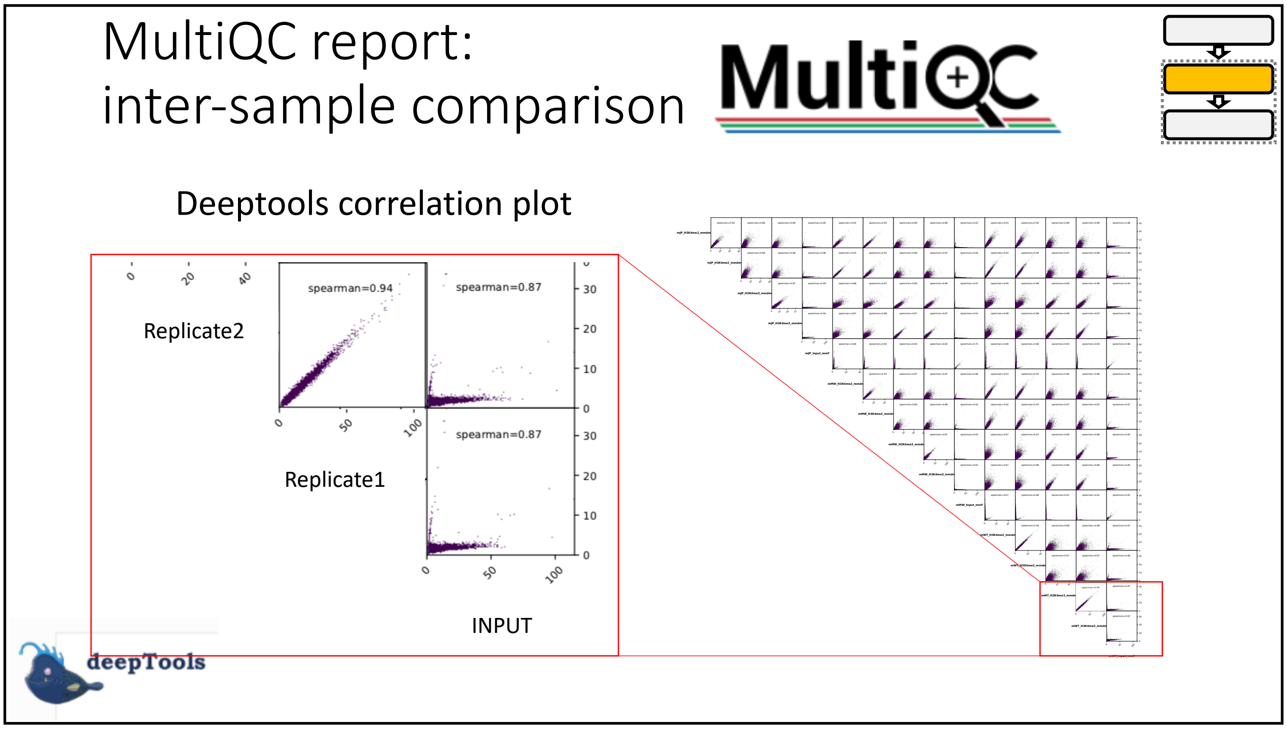


Deeptools Heatmap



0.73	0.87	0.89	0.91	0.91	0.91	0.91	0.91	0.90	0.90	0.91	0.90	0.90	0.91	1.00	mMW_H3K4me3_mmdmF2
0.73	0.87	0.89	0.92	0.92	0.91	0.91	0.91	0.90	0.90	0.91	0.90	0.90	1.00	0.91	mMW_H3K4me3_mmdmF1
0.72	0.87	0.90	0.92	0.91	0.90	0.91	0.92	0.91	0.90	0.91	0.91	1.00	0.90	0.90	mWT_H3K4me3_mmdmF2
0.71	0.87	0.90	0.92	0.91	0.90	0.91	0.91	0.91	0.90	0.91	1.00	0.91	0.90	0.90	mWT_H3K4me3_mmdmF1
0.73	0.88	0.90	0.92	0.91	0.91	0.92	0.92	0.91	0.91	1.00	0.91	0.91	0.91	0.91	mJP_H3K4me3_mmdmF2
0.77	0.88	0.88	0.91	0.90	0.91	0.91	0.92	0.92	1.00	0.91	0.90	0.90	0.90	0.90	mJP_H3K4me3_mmdmF1
0.77	0.89	0.89	0.94	0.94	0.95	0.96	0.96	1.00	0.92	0.91	0.91	0.91	0.90	0.90	mWT_H3K4me2_mmdmF2
0.76	0.89	0.90	0.95	0.94	0.95	0.96	1.00	0.96	0.92	0.92	0.91	0.92	0.91	0.91	mWT_H3K4me2_mmdmF1
0.75	0.89	0.89	0.96	0.95	0.95	1.00	0.96	0.96	0.91	0.92	0.91	0.91	0.91	0.91	mJP_H3K4me2_mmdmF2
0.77	0.89	0.89	0.95	0.95	1.00	0.95	0.95	0.95	0.91	0.91	0.90	0.90	0.91	0.91	mMW_H3K4me2_mmdmF1
0.74	0.88	0.90	0.95	1.00	0.95	0.95	0.94	0.94	0.90	0.91	0.91	0.91	0.92	0.91	mMW_H3K4me2_mmdmF2
0.72	0.88	0.91	1.00	0.95	0.95	0.96	0.95	0.94	0.91	0.92	0.92	0.92	0.92	0.91	mJP_H3K4me2_mmdmF1
0.76	0.92	1.00	0.91	0.90	0.89	0.89	0.90	0.89	0.88	0.90	0.90	0.90	0.89	0.89	mWT_Input_mmF
0.84	1.00	0.92	0.88	0.88	0.89	0.89	0.89	0.89	0.88	0.88	0.87	0.87	0.87	0.87	mJP_Input_mmF
1.00	0.84	0.76	0.72	0.74	0.77	0.75	0.76	0.77	0.77	0.73	0.71	0.72	0.73	0.73	mMW_Input

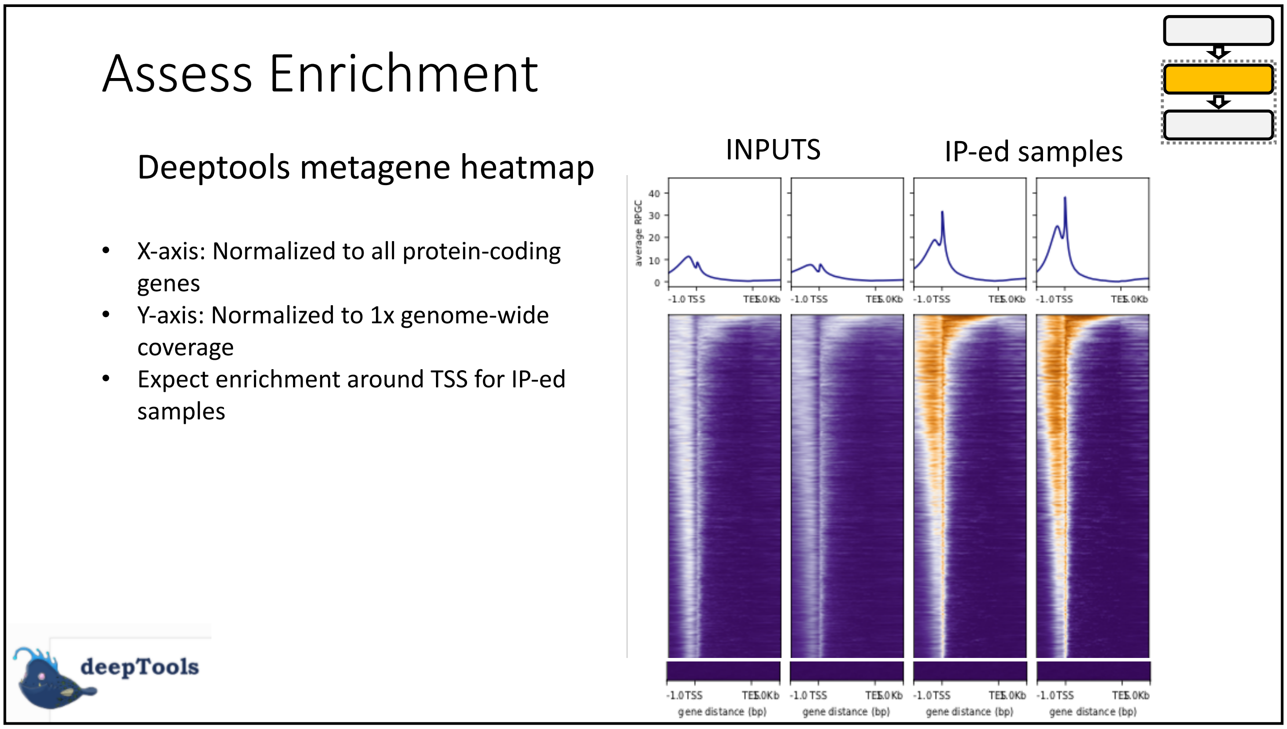
deepTools



MultiQC report: inter-sample comparison

Deeptools correlation plot

23



Assess Enrichment

Deeptools metagene heatmap

- X-axis: Normalized to all protein-coding genes
- Y-axis: Normalized to 1x genome-wide coverage
- Expect enrichment around TSS for IP-ed samples

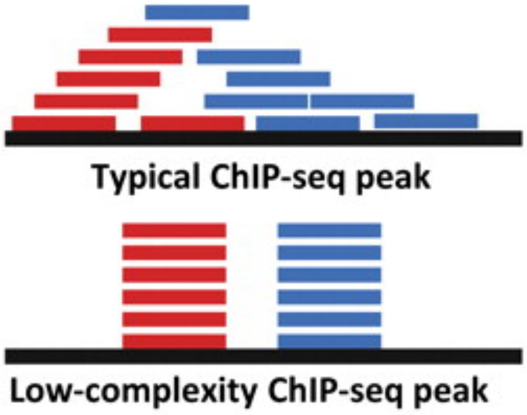
24

QC, Alignment, and Visualization

- Pipelines
- QC and Alignment
- Visualization
  - **Duplication**
  - BigWigs
  - Normalization

25

Duplication

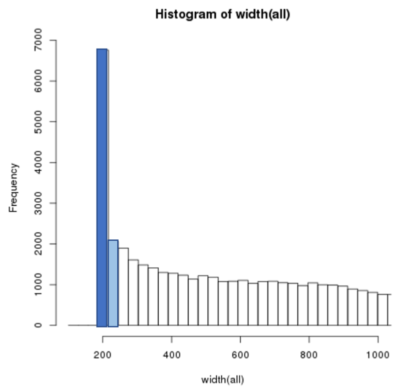


Landt et. al. Genome Res. 2012

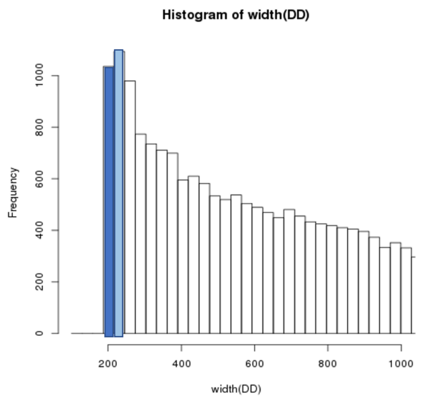
26

Do you need to remove duplicates?

All reads



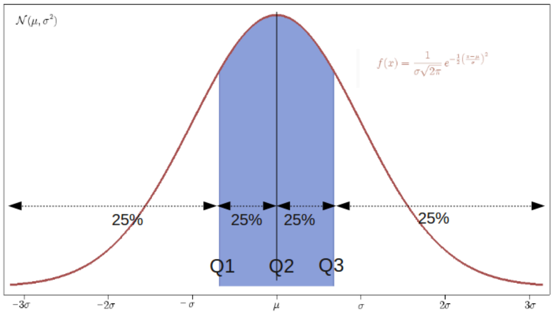
No Duplicates



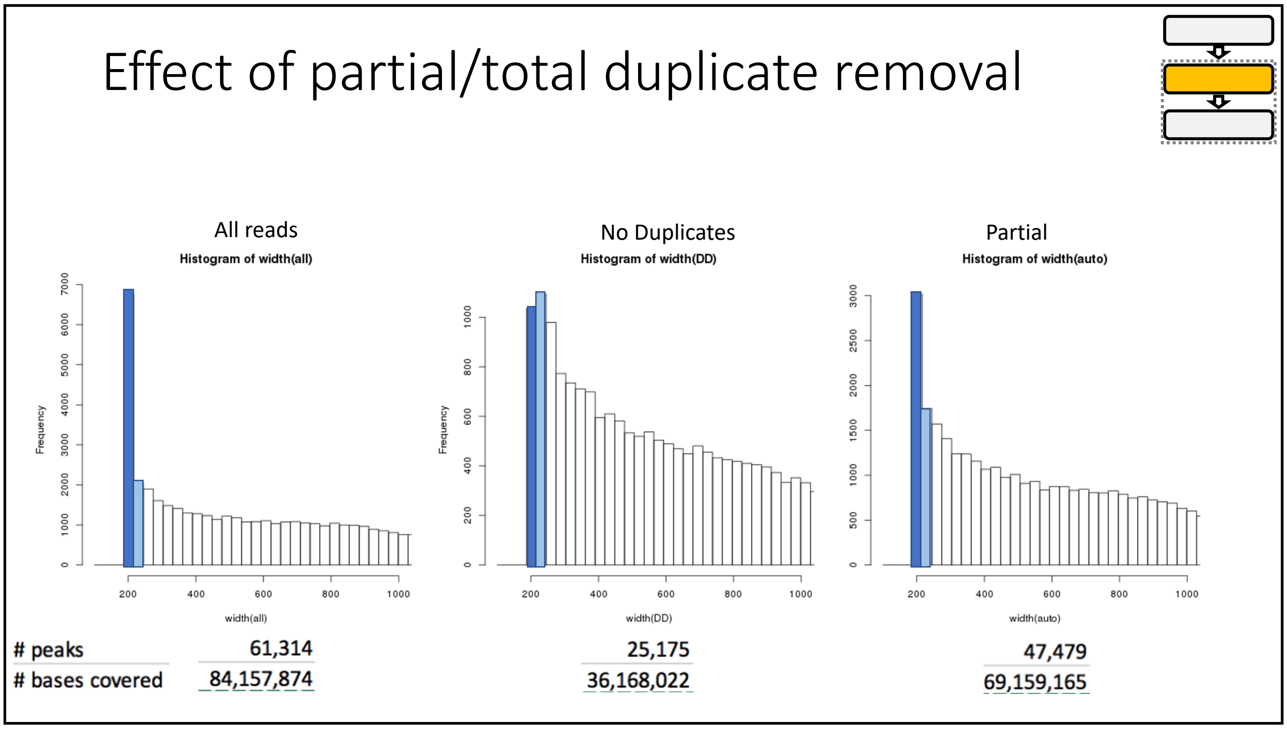
	All reads	No Duplicates
# peaks	61,314	25,175
# bases covered	84,157,874	36,168,022

Two ways to remove duplicates

- Partial duplicate removal
  - Uses a binomial distribution of read numbers across the entire genome and removes the upper quantile.
- Remove all duplicates
  - If reads map to the same start and end position, remove all but one of the reads.



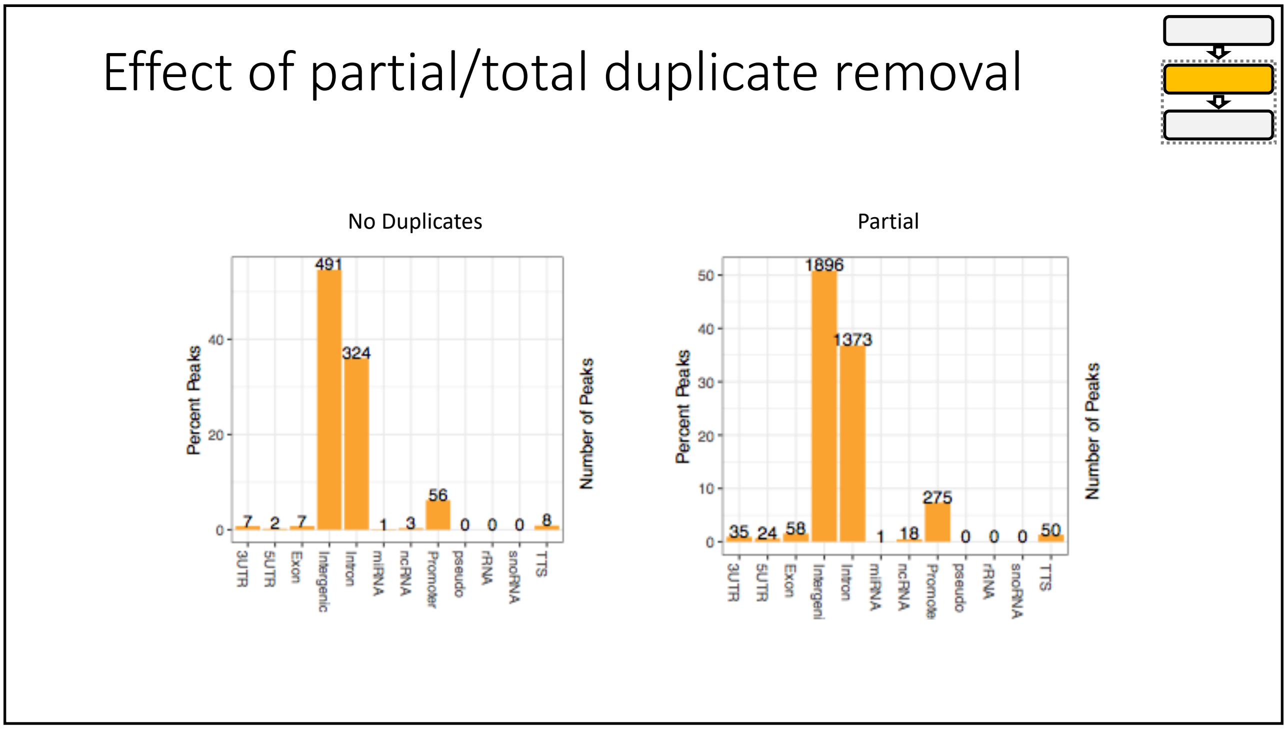
Wikipedia. 2020.



Effect of partial/total duplicate removal

	All reads	No Duplicates	Partial
# peaks	61,314	25,175	47,479
# bases covered	84,157,874	36,168,022	69,159,165

29



Effect of partial/total duplicate removal

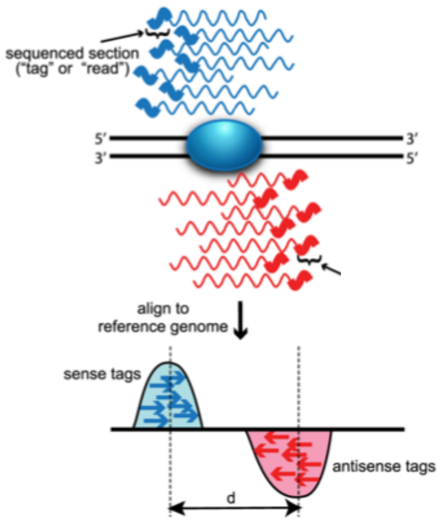
30

QC, Alignment, and Visualization

- **Pipelines**
- **QC and Alignment**
- **Visualization**
  - **Duplication**
  - **BigWigs**
  - **Normalization**

31

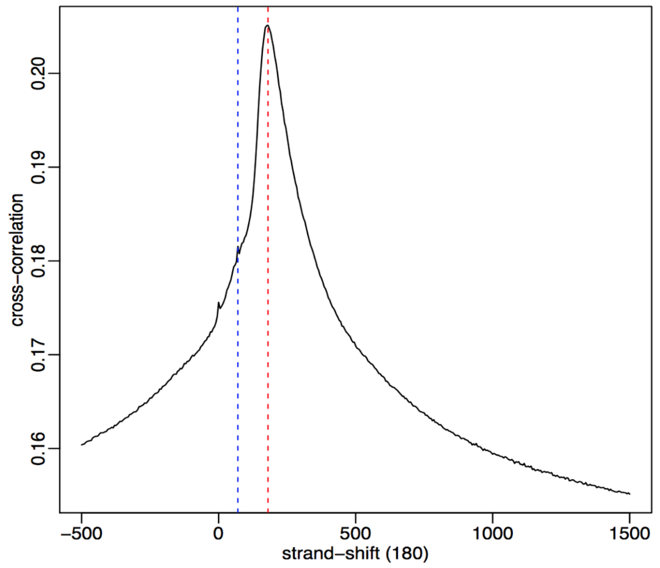
BigWig generation:
Read extension for single end sequencing data



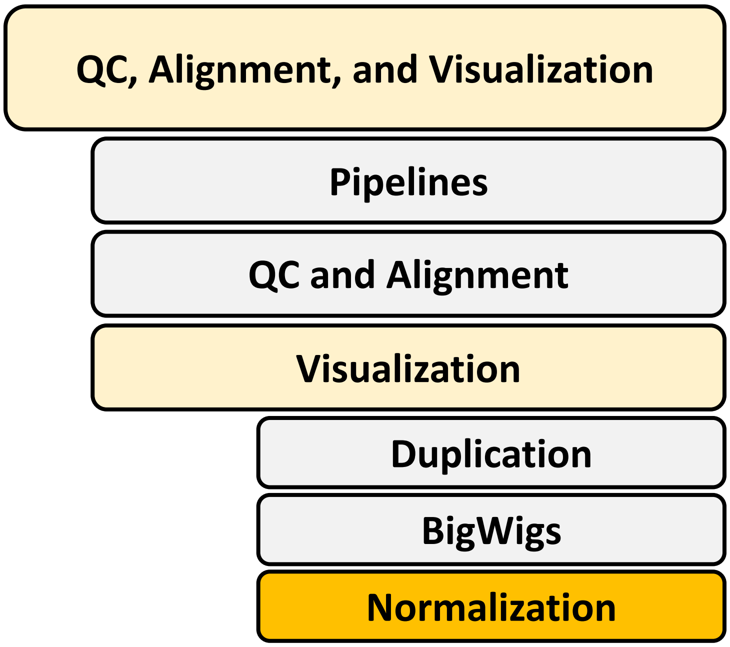
Wilbanks et. al. PLOS ONE. 2010.

32

Calculating the read extension



33



QC, Alignment, and Visualization

Pipelines

QC and Alignment

Visualization

Duplication

BigWigs

Normalization

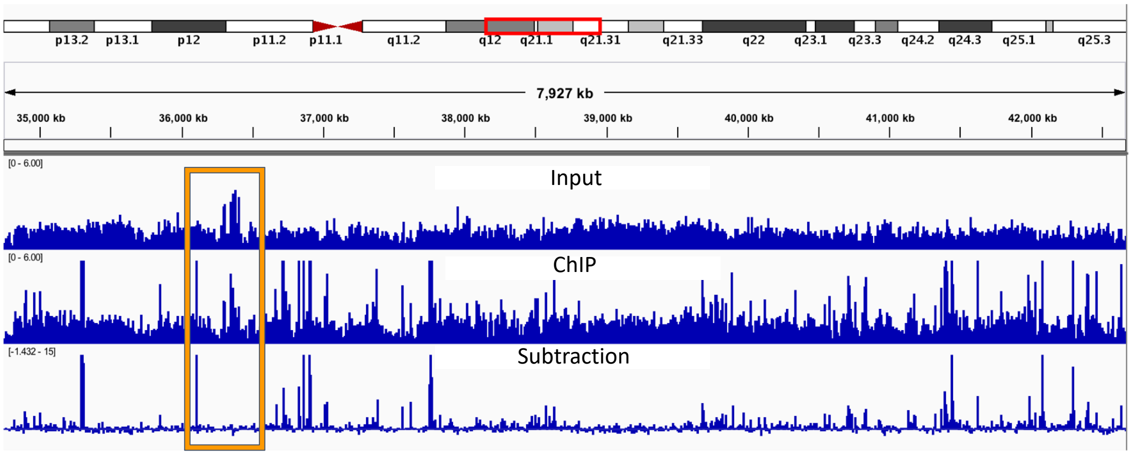
34

Normalization for library size

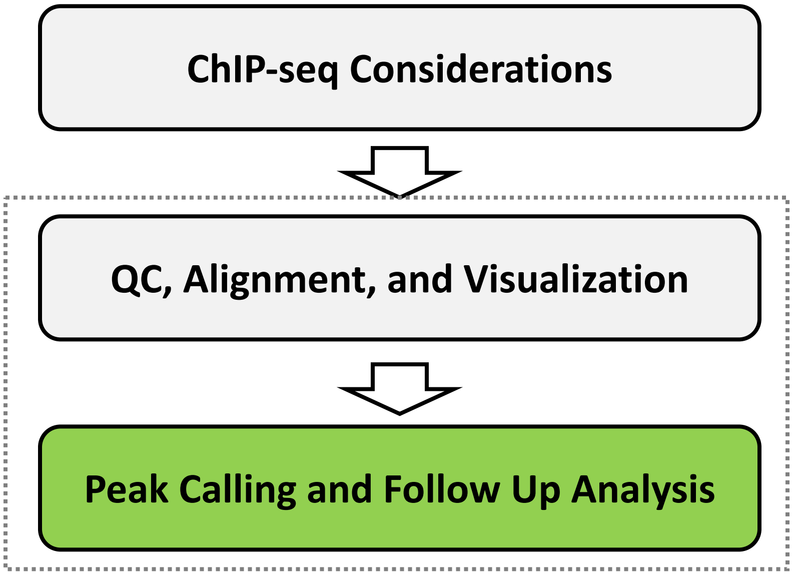
- RPKM:
  - reads per kilobase per million reads
  - defined as:
    - RPKM (per bin) = # of reads per bin / (# of mapped reads (in millions) * bin length (kb))
- RPGC:
  - reads per genomic content
  - used to normalize reads to 1x depth of coverage
  - defined as:
    - RPGC = (total # of mapped reads * fragment length) / effective genome size

35

Input Subtracted Visualization



36



ChIP-seq Considerations

QC, Alignment, and Visualization

Peak Calling and Follow Up Analysis

37

Peak Calling and Follow Up Analysis

- Different Types of Peaks
- Peak Calling
- Annotations
- Motifs
- Differential Binding

38

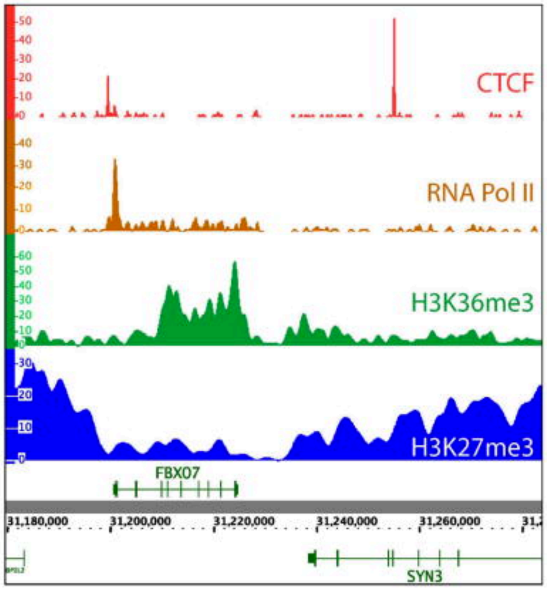
Proteins bind in different ways

- Transcription factor
  - Tight, high peaks
- RNA Pol II
  - Enriched at TSS but bound throughout the gene body
- Histones
  - Some are sharper and located near TSS
  - Some are broader and spread out across the length of active or inactive genes

BROAD PEAKS	NARROW PEAKS
H3F3A	H2AFZ
H3K27me3	H3ac
H3K36me3	H3K27ac
H3K4me1	H3K4me2
H3K79me2	H3K4me3
H3K79me3	H3K9ac
H3K9me1	
H3K9me2	
H4K20me1	

39

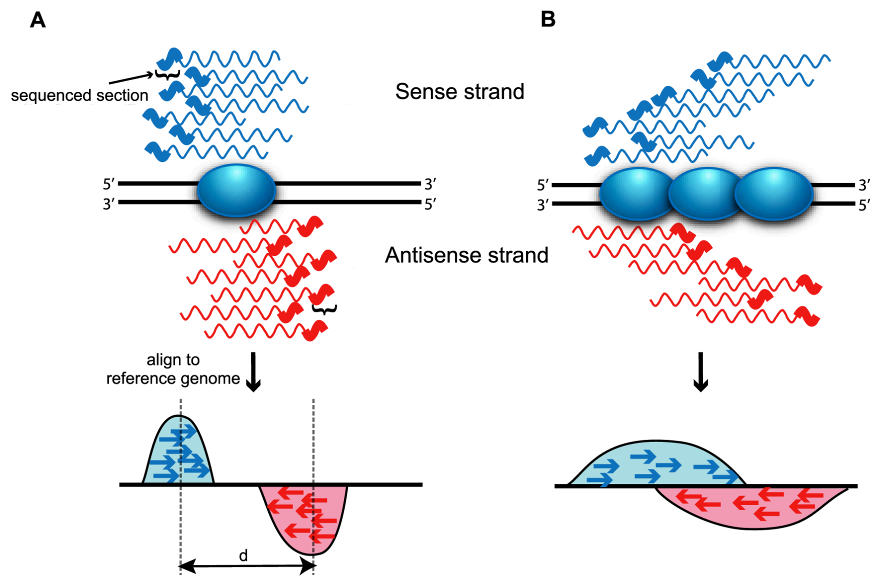
Proteins bind in different ways



Park et al 2009. Nat Rev Genet

40

What causes these different shapes?



Wilbanks et al 2010. PLOS ONE

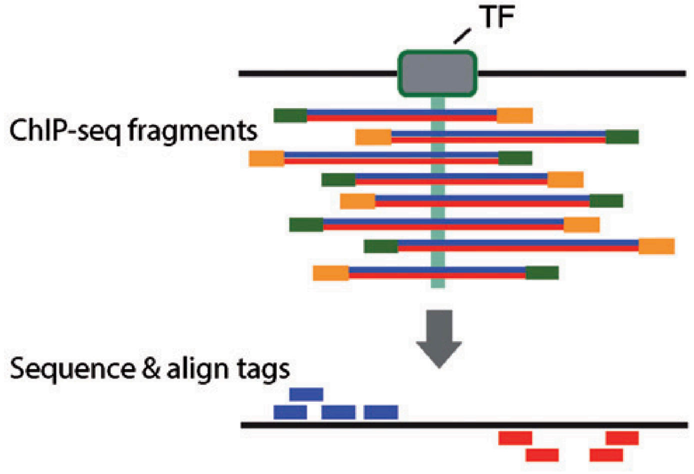
41

Peak Calling and Follow Up Analysis

- Different Types of Peaks
- Peak Calling
- Annotations
- Motifs
- Differential Binding

42

How are peaks called?

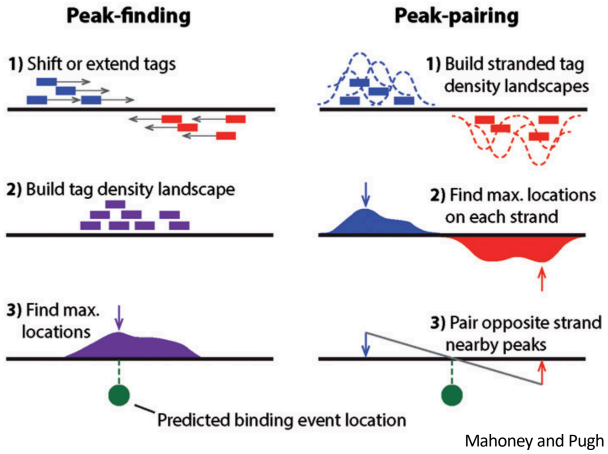


Mahoney and Pugh et al 2015. Criti Rev Biochemi and MolBio

43

General concept of most peak callers

Count the number of reads within a window and determine whether this number is above background


Peak-finding
1) Shift or extend tags
2) Build tag density landscape
3) Find max. locations
Peak-pairing
1) Build stranded tag density landscapes
2) Find max. locations on each strand
3) Pair opposite strand nearby peaks
Predicted binding event location

Mahoney and Pugh et al 2015. Criti Rev Biochemi and MolBio

44

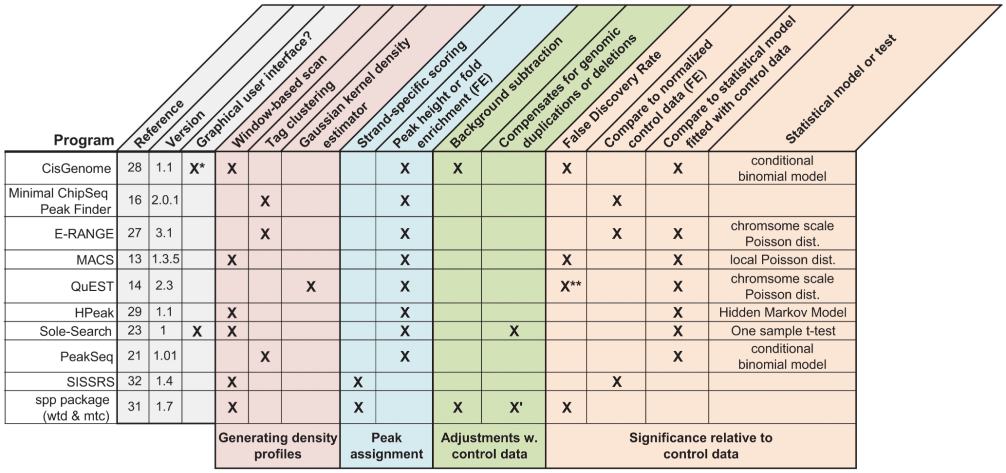
There are many peak callers out there…

GEM	CCAT	Fseq	Hotspot	spp-msp
BCP	ChIPDiff	QuEST	Qeseq	Sole-Search
MUSIC	ERANGE	RSEG	Hpeak	CisGenome
MACS2	PeakSeq	TPIC	BayesPeak	Gene Track
ZINBA	**SICER**	W-ChIPPekas	spp-wtd	FindPeaks
Genrich	SISSRs	PolyPeak	spp-mtc	etc…

Thomas et al 2017. Briefings in Bioinformatics

45

Each peak caller has different methods and benefits



Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

Wilbanks et al 2010. PLOS ONE

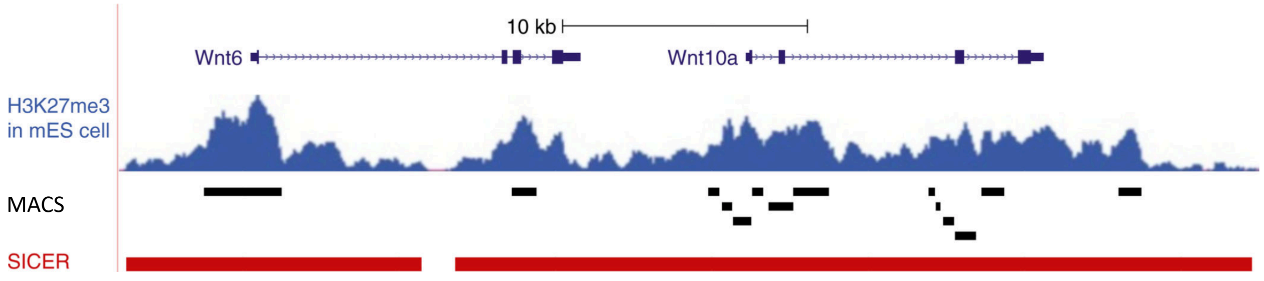
46

Peak calling: things to keep in mind

- Peak callers are designed to deal with different types of peaks
  - Pay attention to what they're designed to handle
- Peak callers are optimized for a specific type of peak/dataset
  - Tuning the parameters is often important
  - Including the p-value, q-value, and/or FDR
- Peaks will not completely overlap across replicates or tools

47

MACS works well for narrow peaks while SICER is designed for broad peaks



Xu et al 2014. Methods Mol Biol

48

Model-based Analysis of ChIP-Seq (MACS)

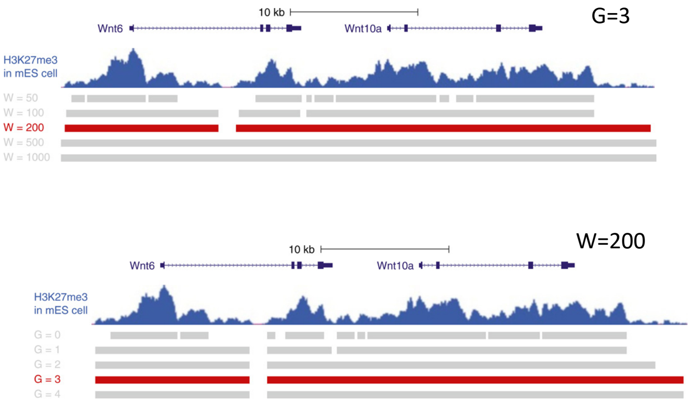
- Extend reads and scale to library size
- Call candidate peaks relative to:
  - control sample
  - genome background
  - large local region
  - small local region
- Calculate FDR by calling peaks in the control relative to the ChIP

Feng et al 2012. Nature Protocols

49

Spatial Clustering for Identification of ChIP-Enriched Regions (SICER)

- Uses windows and gaps to identify ”islands” of enrichment
- Gaps allow for short regions lacking binding within an island, more pattern variability across island
- Compares to a randomized background and control background to calculate FDR



G=3

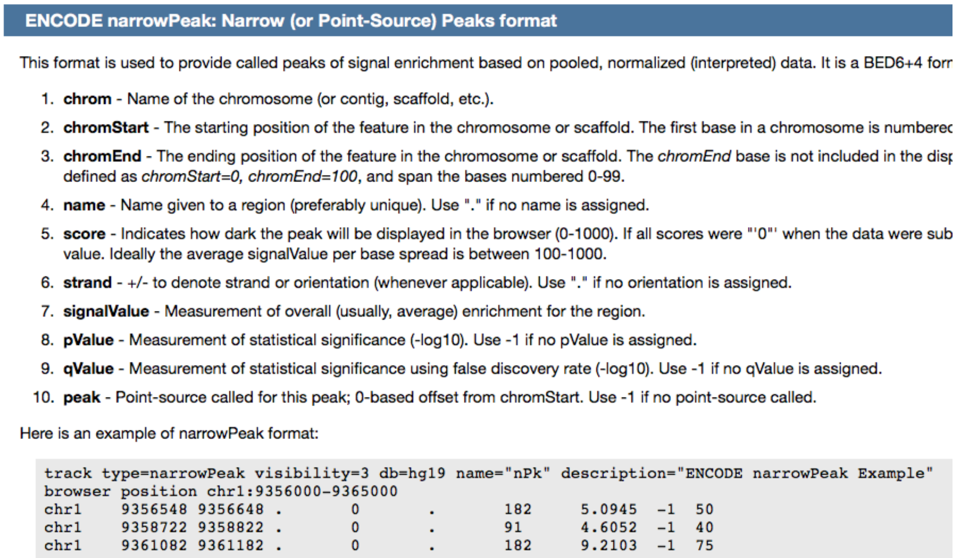
W=200

Xu et al 2014. Methods Mol Biol

50

Output file formats

- https://genome.ucsc.edu/FAQ/FAQformat.html



ENCODE narrowPeak: Narrow (or Point-Source) Peaks format

This format is used to provide called peaks of signal enrichment based on pooled, normalized (interpreted) data. It is a BED6+4 forr

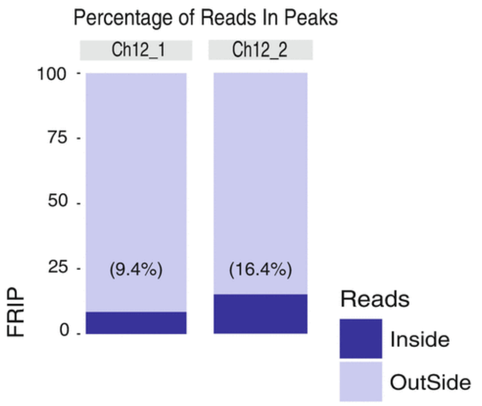
1. **chrom** - Name of the chromosome (or contig, scaffold, etc.).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numberec
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the disp defined as *chromStart=0, chromEnd=100*, and span the bases numbered 0-99.
4. **name** - Name given to a region (preferably unique). Use "." if no name is assigned.
5. **score** - Indicates how dark the peak will be displayed in the browser (0-1000). If all scores were "'0"' when the data were sub value. Ideally the average signalValue per base spread is between 100-1000.
6. **strand** - +/- to denote strand or orientation (whenever applicable). Use "." if no orientation is assigned.
7. **signalValue** - Measurement of overall (usually, average) enrichment for the region.
8. **pValue** - Measurement of statistical significance (-log10). Use -1 if no pValue is assigned.
9. **qValue** - Measurement of statistical significance using false discovery rate (-log10). Use -1 if no qValue is assigned.
10. **peak** - Point-source called for this peak; 0-based offset from chromStart. Use -1 if no point-source called.

Here is an example of narrowPeak format:

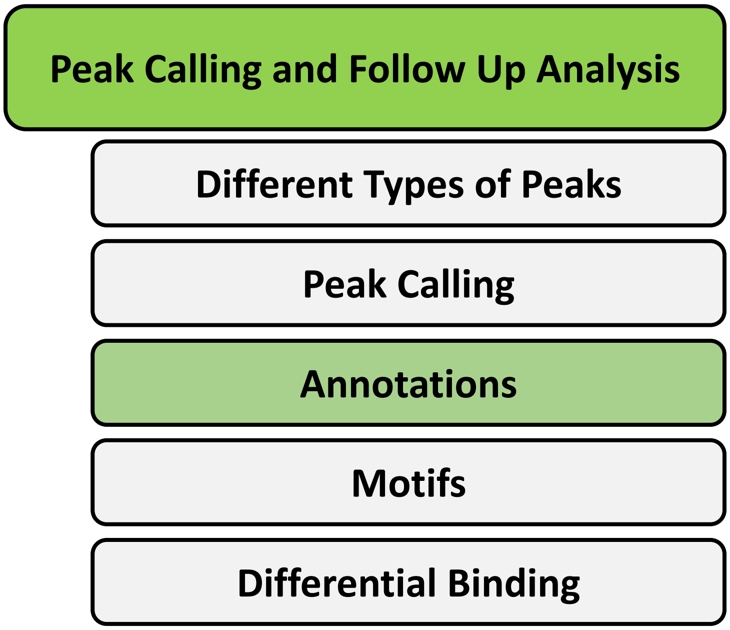
```
track type=narrowPeak visibility=3 db=hg19 name="nPk" description="ENCODE narrowPeak Example"
browser position chr1:9356000-9365000
chr1    9356548 9356648 .       0       .       182     5.0945  -1  50
chr1    9358722 9358822 .       0       .       91      4.6052  -1  40
chr1    9361082 9361182 .       0       .       182     9.2103  -1  75
```

FRiP (Fraction of Reads in Peaks)

- Measures global ChIP enrichment
- Quick understanding of quality of the IP and peak calling algorithm
- Good quality FRiP for a transcription factor: > 5%



de Santiago, Carroll 2017. Chromatin Immunoprecipitation



Peak Calling and Follow Up Analysis

Different Types of Peaks

Peak Calling

Annotations

Motifs

Differential Binding

Annotations: questions to ask

- Is this protein enriched around promoters?
  - Many tools are biased towards promoters/TSS sites
- What is a gene?
  - Do you have a reason to include pseudogenes, lincRNAs, etc?
- Do you care about introns/alternative transcripts?
- What happens if a peak overlaps multiple genes?

Annotation tools

HOMER

- Straight-forward to use
- Only protein coding genes
- Focused on nearest TSS
- One annotation per peak

UROPA

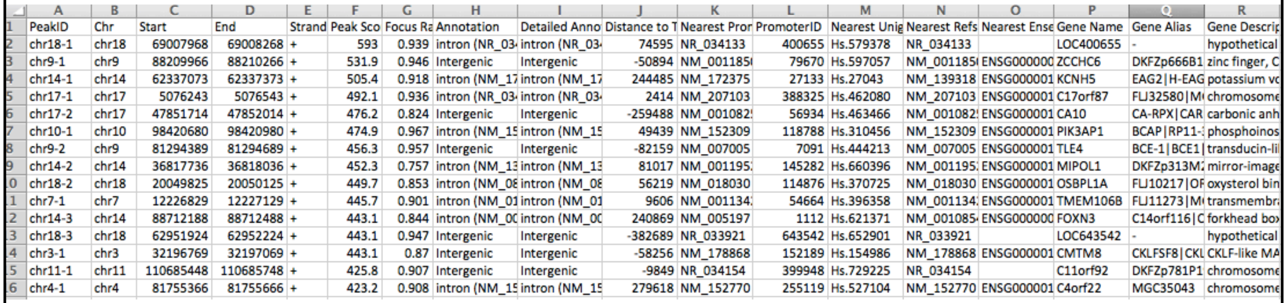
- More complicated to set up
- Takes any gene list input
- Focuses where the user decides
- Creates two tables: one of top annotation per peak, and one of all possible annotations given the input conditions

Heinz et al 2010. Mol Cell

Kondili et al 2017. Scientific Reports

55

Annotation tools: example HOMER output table

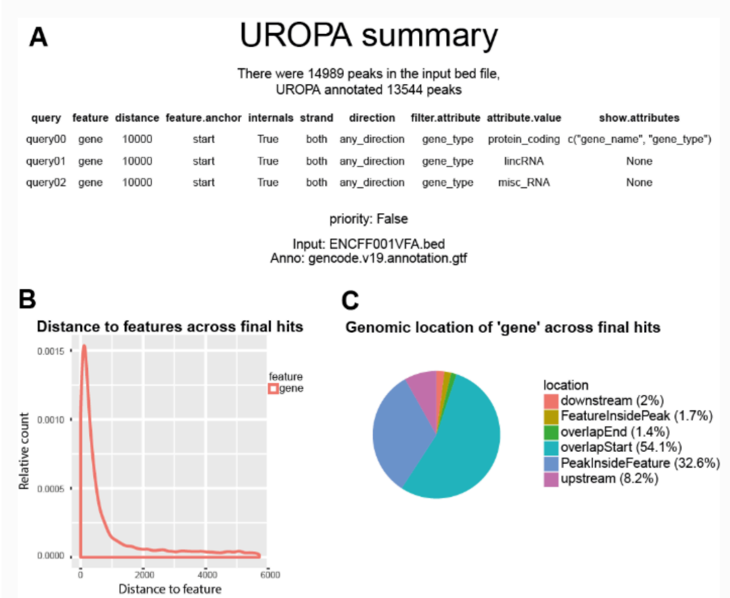


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	400655	Hs.579378	NR_034133		LOC400655	-	hypothetical
3	chr9-1	chr9	88209966	88210266	+	531.9	0.946	Intergenic	Intergenic	-50894	NM_001185	79670	Hs.597057	NM_001185	ENSG000000	ZCCHC6	DKFZp666B1	zinc finger, C
4	chr14-1	chr14	62337073	62337373	+	505.4	0.918	intron (NM_17	intron (NM_17	244485	NM_172375	27133	Hs.27043	NM_139318	ENSG000001	KCNH5	EAG2|H-EAG	potassium vo
5	chr17-1	chr17	5076243	5076543	+	492.1	0.936	intron (NR_03	intron (NR_03	2414	NM_207103	388325	Hs.462080	NM_207103	ENSG000001	C17orf87	FLJ32580|M	chromosome
6	chr17-2	chr17	47851714	47852014	+	476.2	0.824	Intergenic	Intergenic	-259488	NM_001082	56934	Hs.463466	NM_001082	ENSG000001	CA10	CA-RPX|CAR	carbonic anh
7	chr10-1	chr10	98420680	98420980	+	474.9	0.967	intron (NM_15	intron (NM_15	49439	NM_152309	118788	Hs.310456	NM_152309	ENSG000001	PIK3AP1	BCAP|RP11-	phosphoinos
8	chr9-2	chr9	81294389	81294689	+	456.3	0.957	Intergenic	Intergenic	-82159	NM_007005	7091	Hs.444213	NM_007005	ENSG000001	TLE4	BCE-1|BCE1|	transducin-li
9	chr14-2	chr14	36817736	36818036	+	452.3	0.757	intron (NM_13	intron (NM_13	81017	NM_001195	145282	Hs.660396	NM_001195	ENSG000001	MIPOL1	DKFZp313M	mirror-image
10	chr18-2	chr18	20049825	20050125	+	449.7	0.853	intron (NM_08	intron (NM_08	56219	NM_018030	114876	Hs.370725	NM_018030	ENSG000001	OSBPL1A	FLJ10217|OF	oxysterol bin
11	chr7-1	chr7	12226829	12227129	+	445.7	0.901	intron (NM_01	intron (NM_01	9606	NM_001134	54664	Hs.396358	NM_001134	ENSG000001	TMEM106B	FLJ11273|M	transmembra
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_00	intron (NM_00	240869	NM_005197	1112	Hs.621371	NM_001085	ENSG000000	FOXN3	C14orf116|C	forkhead box
13	chr18-3	chr18	62951924	62952224	+	443.1	0.947	Intergenic	Intergenic	-382689	NR_033921	643542	Hs.652901	NR_033921		LOC643542	-	hypothetical
14	chr3-1	chr3	32196769	32197069	+	443.1	0.87	Intergenic	Intergenic	-58256	NM_178868	152189	Hs.154986	NM_178868	ENSG000001	CMTM8	CKLFSF8|CKL	CKLF-like MA
15	chr11-1	chr11	110685448	110685748	+	425.8	0.907	Intergenic	Intergenic	-9849	NR_034154	399948	Hs.729225	NR_034154		C11orf92	DKFZp781P1	chromosome
16	chr4-1	chr4	81755366	81755666	+	423.2	0.908	intron (NM_15	intron (NM_15	279618	NM_152770	255119	Hs.527104	NM_152770	ENSG000001	C4orf22	MGC35043	chromosome

Heinz et al 2010. Mol Cell

56

UROPA output figures



A

UROPA summary

There were 14989 peaks in the input bed file,
UROPA annotated 13544 peaks

query	feature	distance	feature.anchor	internals	strand	direction	filter.attribute	attribute.value	show.attributes
query00	gene	10000	start	True	both	any_direction	gene_type	protein_coding	c("gene_name", "gene_type")
query01	gene	10000	start	True	both	any_direction	gene_type	lincRNA	None
query02	gene	10000	start	True	both	any_direction	gene_type	misc_RNA	None

priority: False

Input: ENCFF001VFA.bed
Anno: gencode.v19.annotation.gtf

Kondili et al 2017. Scientific Reports

57

Annotation tools

PAVIS

- Online tool
- Annotates based on nearest TSS
- Has an “intuitive” interface

manticore.niehs.nih.gov/pavis2

PAVIS

GREAT

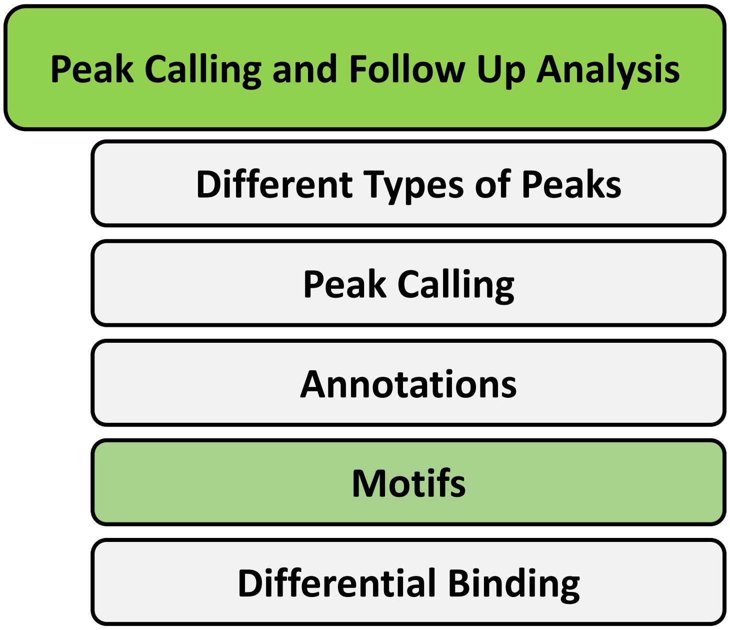
GREAT

- Online tool
- Annotates based on nearest TSS
- Each peak can be associated with up to two genes (one in each direction)
- Only works with four reference genomes (human and mouse)
- Also includes functional enrichment analyses

http://great.stanford.edu/

Huang W et al 2013. Bioinformatics
McLean CY et al 2010. Comp Biol

58



Peak Calling and Follow Up Analysis

Different Types of Peaks

Peak Calling

Annotations

Motifs

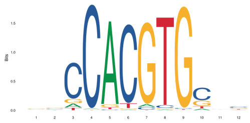
Differential Binding

59

Motifs: things to consider

- Transcription factor motifs:
  - Tends to be small and robust; often centrally located in peaks
- Other proteins:
  - More varied, degenerated motifs, if any at all
  - Rarely centrally located
- Motifs are identified as enriched in peaks relative to some background: should it be the entire genome, just promoters, or something else?
- Search for known motifs or novel motifs?

MYC



60

Motif Calling Tools

MEME Suite

- MEME-ChIP: novel motifs
  - MEME
  - DREME: small, robust motifs
  - Centrimo: centrally enriched motifs
- AME: known motifs

MEME-ChIP
Motif Analysis of Large Nucleotide Datasets

AME
Analysis of Motif Enrichment

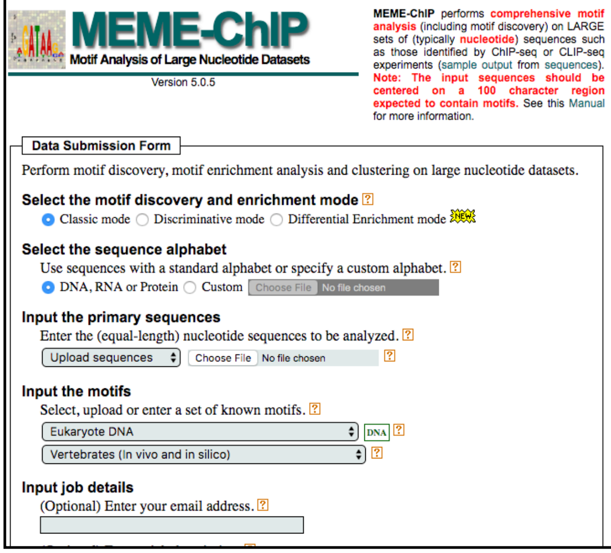
HOMER

- Runs for both known and novel motifs simultaneously

Heinz et al 2010. Mol Cell
Bailey et al 2009. Nucleic Acids Research

61

MEME: meme-suite.org



MEME-ChIP
Motif Analysis of Large Nucleotide Datasets
Version 5.0.5

MEME-ChIP performs **comprehensive motif analysis** (including motif discovery) on LARGE sets of (typically **nucleotide**) sequences such as those identified by ChIP-seq or CLIP-seq experiments (sample output from sequences). **Note: The input sequences should be centered on a 100 character region expected to contain motifs.** See this Manual for more information.

Data Submission Form

Perform motif discovery, motif enrichment analysis and clustering on large nucleotide datasets.

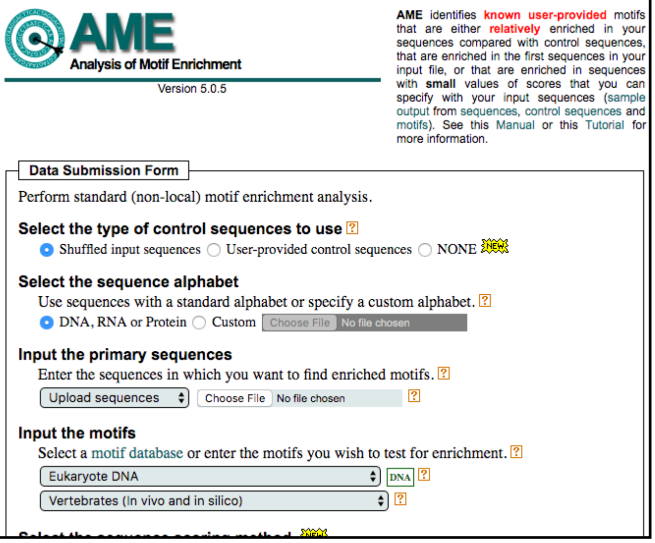
Select the motif discovery and enrichment mode ?
● Classic mode ○ Discriminative mode ○ Differential Enrichment mode NEW

Select the sequence alphabet
Use sequences with a standard alphabet or specify a custom alphabet. ?
● DNA, RNA or Protein ○ Custom Choose File No file chosen

Input the primary sequences
Enter the (equal-length) nucleotide sequences to be analyzed. ?
Upload sequences | Choose File No file chosen ?

Input the motifs
Select, upload or enter a set of known motifs. ?
Eukaryote DNA | DNA ?
Vertebrates (In vivo and in silico) ?

Input job details
(Optional) Enter your email address. ?



AME
Analysis of Motif Enrichment
Version 5.0.5

AME identifies **known user-provided** motifs that are either **relatively** enriched in your sequences compared with control sequences, that are enriched in the first sequences in your input file, or that are enriched in sequences with **small** values of scores that you can specify with your input sequences (sample output from sequences, control sequences and motifs). See this Manual or this Tutorial for more information.

Data Submission Form

Perform standard (non-local) motif enrichment analysis.

Select the type of control sequences to use ?
● Shuffled input sequences ○ User-provided control sequences ○ NONE NEW

Select the sequence alphabet
Use sequences with a standard alphabet or specify a custom alphabet. ?
● DNA, RNA or Protein ○ Custom Choose File No file chosen

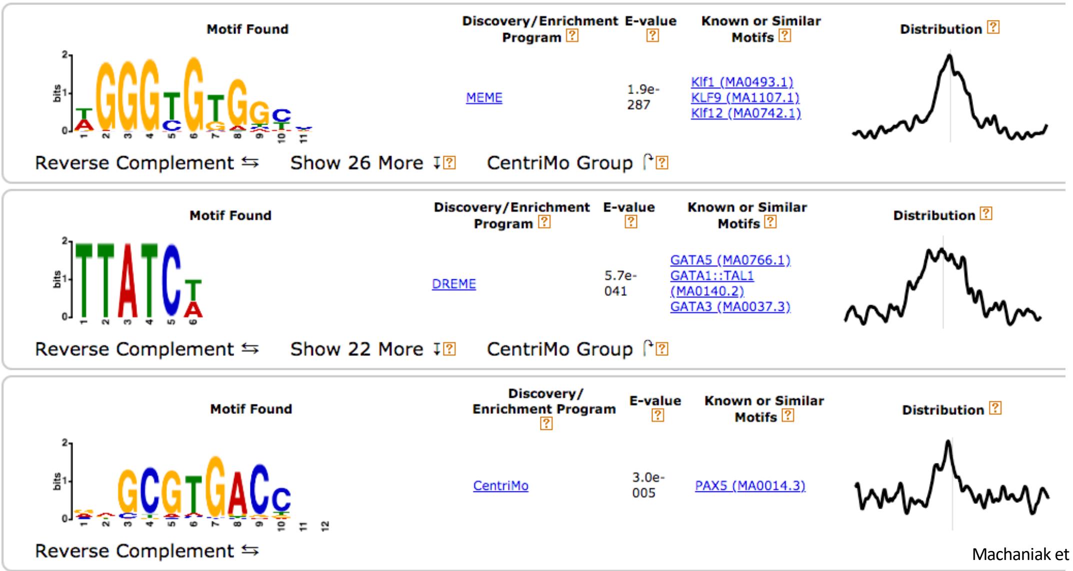
Input the primary sequences
Enter the sequences in which you want to find enriched motifs. ?
Upload sequences | Choose File No file chosen ?

Input the motifs
Select a motif database or enter the motifs you wish to test for enrichment. ?
Eukaryote DNA | DNA ?
Vertebrates (In vivo and in silico) ?

Select the sequence scoring method NEW

62

MEME-ChIP output



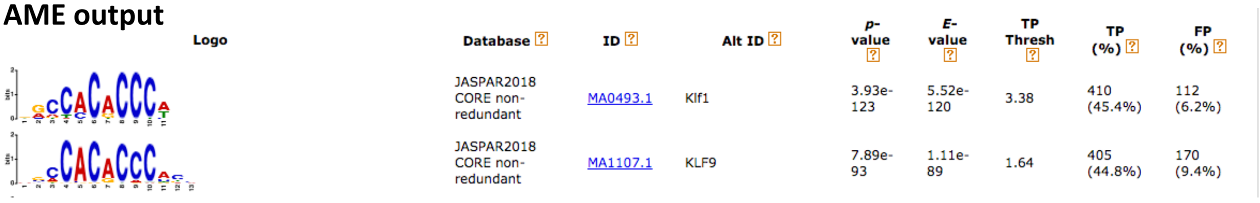
Motif Found	Discovery/Enrichment Program ?	E-value ?	Known or Similar Motifs ?	Distribution ?
	MEME	1.9e-287	Klf1 (MA0493.1) KLF9 (MA1107.1) Klf12 (MA0742.1)	
Reverse Complement ⇆	Show 26 More ↧?	CentriMo Group ↱?		
	DREME	5.7e-041	GATA5 (MA0766.1) GATA1::TAL1 (MA0140.2) GATA3 (MA0037.3)	
Reverse Complement ⇆	Show 22 More ↧?	CentriMo Group ↱?		
	CentriMo	3.0e-005	PAX5 (MA0014.3)	
Reverse Complement ⇆				

Machaniak et al 2011. Bioinformatics

63

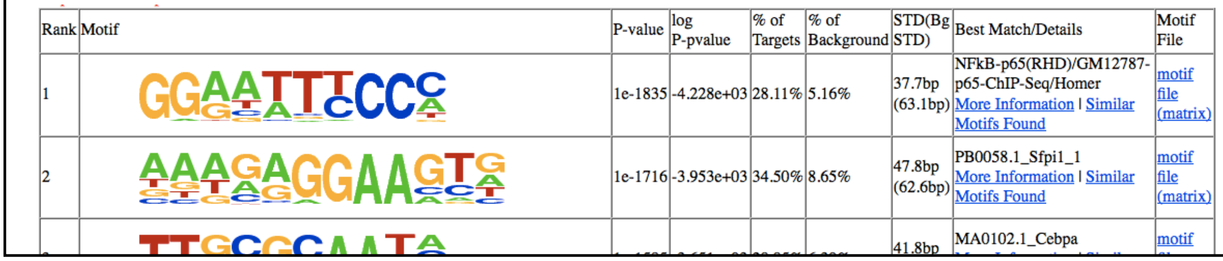
Motif seach: tabular outputs

AME output



Logo	Database ?	ID ?	Alt ID ?	p-value ?	E-value ?	TP Thresh ?	TP (%) ?	FP (%) ?
	JASPAR2018 CORE non-redundant	MA0493.1	Klf1	3.93e-123	5.52e-120	3.38	410 (45.4%)	112 (6.2%)
	JASPAR2018 CORE non-redundant	MA1107.1	KLF9	7.89e-93	1.11e-89	1.64	405 (44.8%)	170 (9.4%)

HOMER output



Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1	GGAATTTCCC	1e-1835	-4.228e+03	28.11%	5.16%	37.7bp (63.1bp)	NFkB-p65(RHD)/GM12787-p65-ChIP-Seq/Homer More Information | Similar Motifs Found	motif file (matrix)
2	AAAGAGGAAGTG	1e-1716	-3.953e+03	34.50%	8.65%	47.8bp (62.6bp)	PB0058.1_Sfpi1_1 More Information | Similar Motifs Found	motif file (matrix)
	TTGCGCAATA					41.8bp	MA0102.1_Cebpa	motif

64

Peak Calling and Follow Up Analysis

Different Types of Peaks

Peak Calling

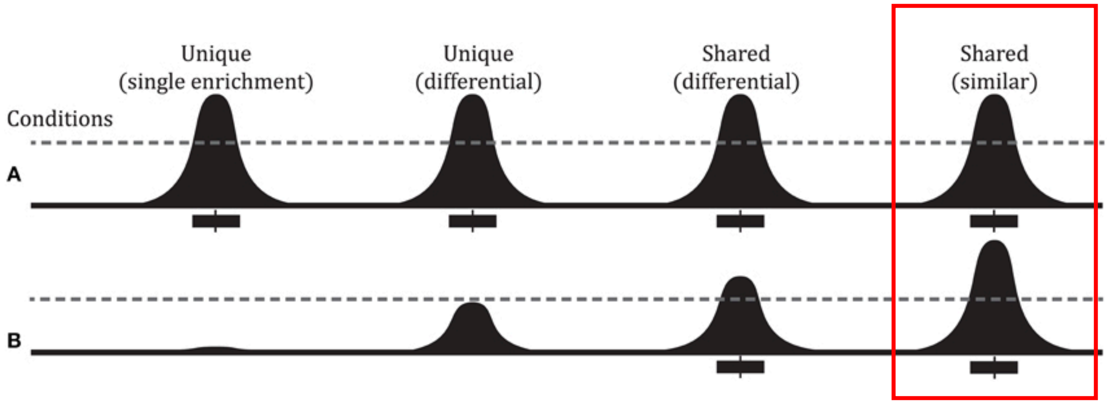
Annotations

Motifs

Differential Binding

65

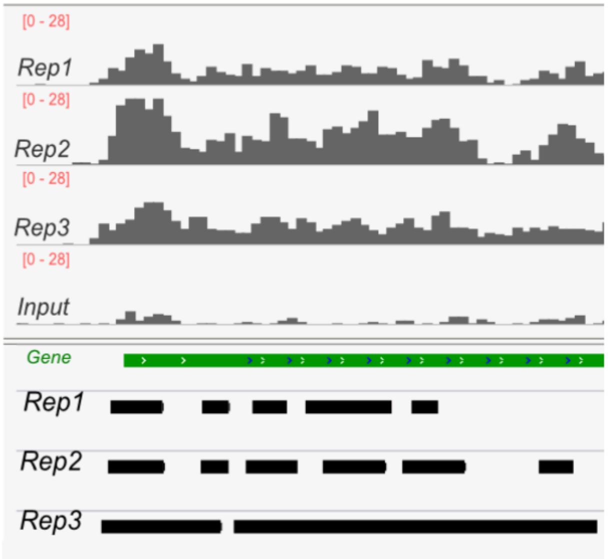
Key assumption of differential peak calling: most peaks are similar across conditions



Wu et al 2015. Front Genet

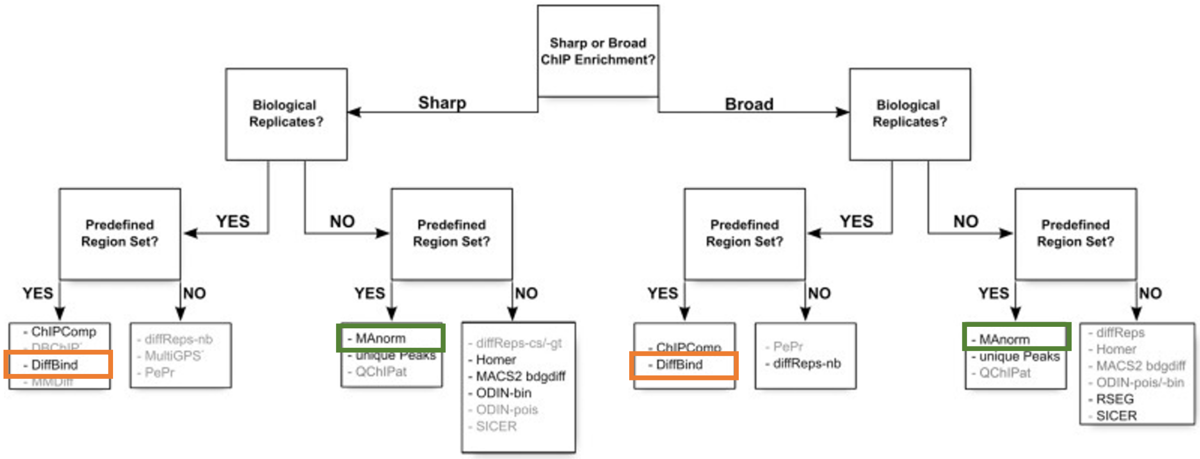
66

Differential peak calling is dependent on peak calling quality



Yang et al 2014. Comput Struct Biotechnol J

Differential peak calling



Sharp or Broad ChIP Enrichment?

Sharp → Biological Replicates?

- YES → Predefined Region Set?
  - YES → - ChIPComp - DBChIP - DiffBind - MMDiff
  - NO → - diffReps-nb - MultiGPS - PePr
- NO → Predefined Region Set?
  - YES → - MAnorm - unique Peaks - QChIPat
  - NO → - diffReps-cs/-gt - Homer - MACS2 bdgdiff - ODIN-bin - ODIN-pois - SICER

Broad → Biological Replicates?

- YES → Predefined Region Set?
  - YES → - ChIPComp - DiffBind
  - NO → - PePr - diffReps-nb
- NO → Predefined Region Set?
  - YES → - MAnorm - unique Peaks - QChIPat
  - NO → - diffReps - Homer - MACS2 bdgdiff - ODIN-pois/-bin - RSEG - SICER

Steinhauser et al 2016. Brief Bioinformatics

Differential peak calling tools

MANORM

- Cannot handle replicates
- Lacks statistical power
- Needs peaks to be defined from an outside source
- Works for both narrow and broad peaks

DIFFBIND

- Requires replicates of all conditions
- Has a statistical framework
- Needs peaks to be defined from an outside source
- Works for both narrow and broad peaks

Ross-Innes et al 2012. Nature
Shao et al 2012. Genome Biology

69

Comparing your data to other ChIP-seq data

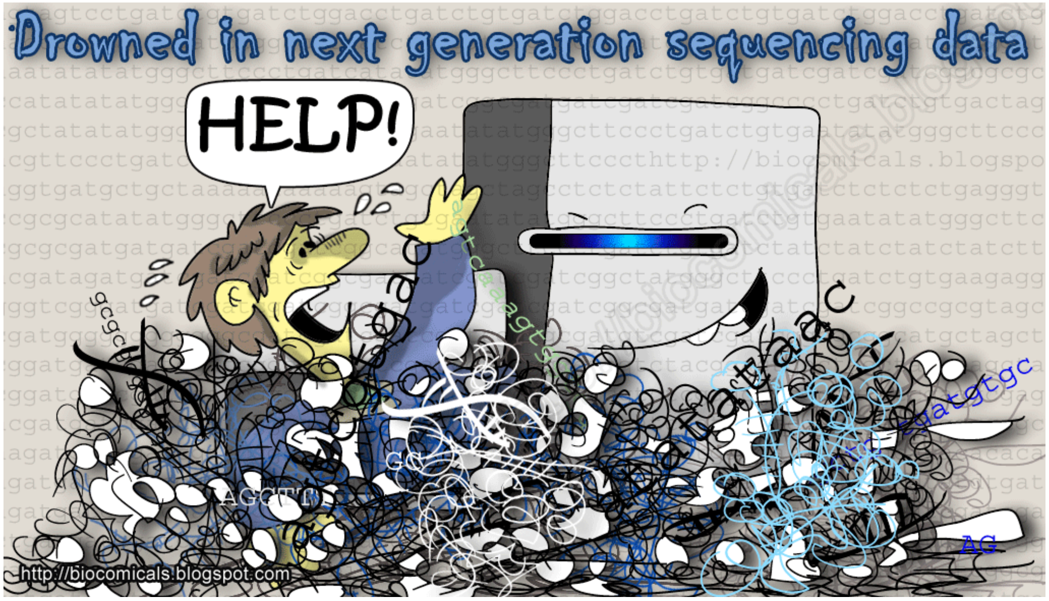
- ENCODE (Encyclopedia of DNA Elements)
  - www.encodeproject.org
  - Visualizations and peak analyses of mouse, human, *Drosophila*, and *C. elegans* data in healthy control conditions. Data types include ChIP-seq, DNase-seq, ATAC-seq, HiC, and more.
- Cistrome
  - cistrome.org
  - Cistrome Analysis Pipeline, Cistrome Data Browser, Cistrome Cancer, Cistrome-GO, CistromeDB Toolkit, Landscape *In Silico* deletion Analysis
  - Visualizations and peak analyses of many public mouse and human ChIP-seq, DNase-seq, and ATAC-seq datasets reanalyzed using their pipeline
- GTRD (Gene Transcription Regulation Database)
  - gtrd.biouml.org
  - Used DNase-seq, ChIP-seq, and motif databases to identify transcription factor binding sites for human and mouse genomes

70

Conclusions

- ChIP-seq is not trivial.
- Every experiment is unique.
- Experimental design is critical for ChIP-seq.

71


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

72