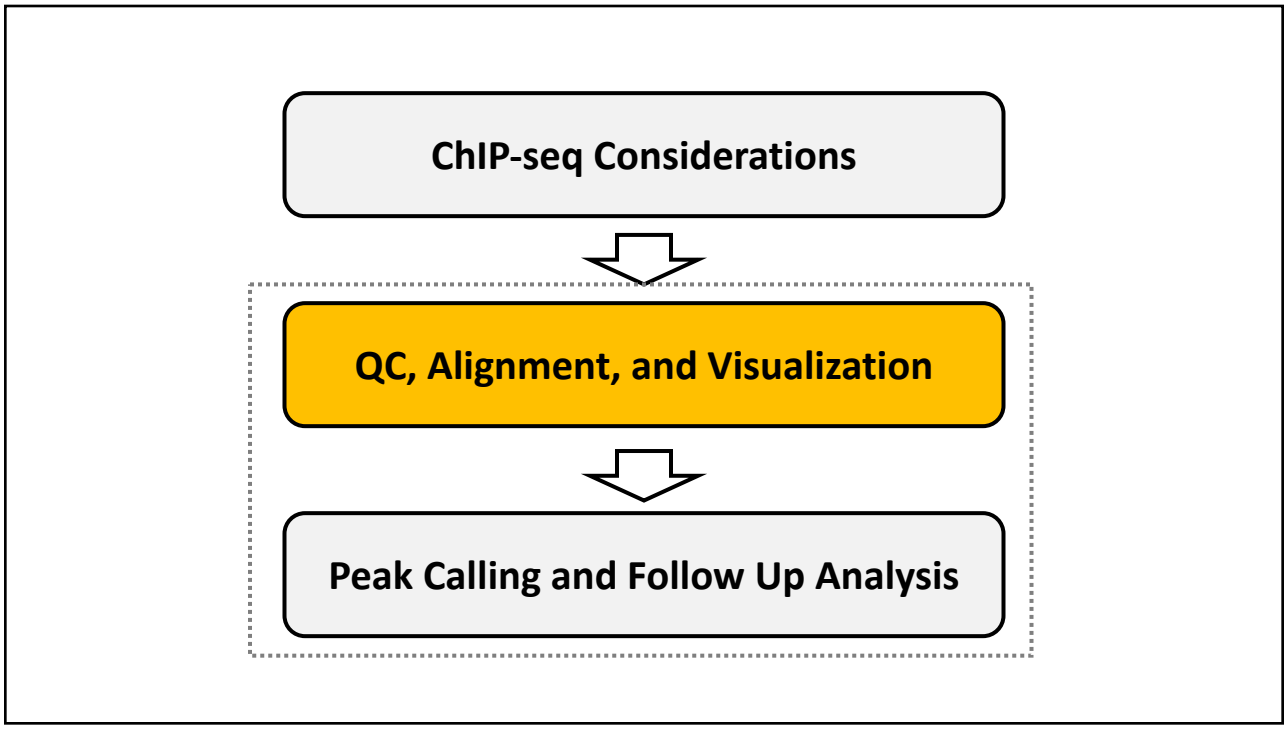
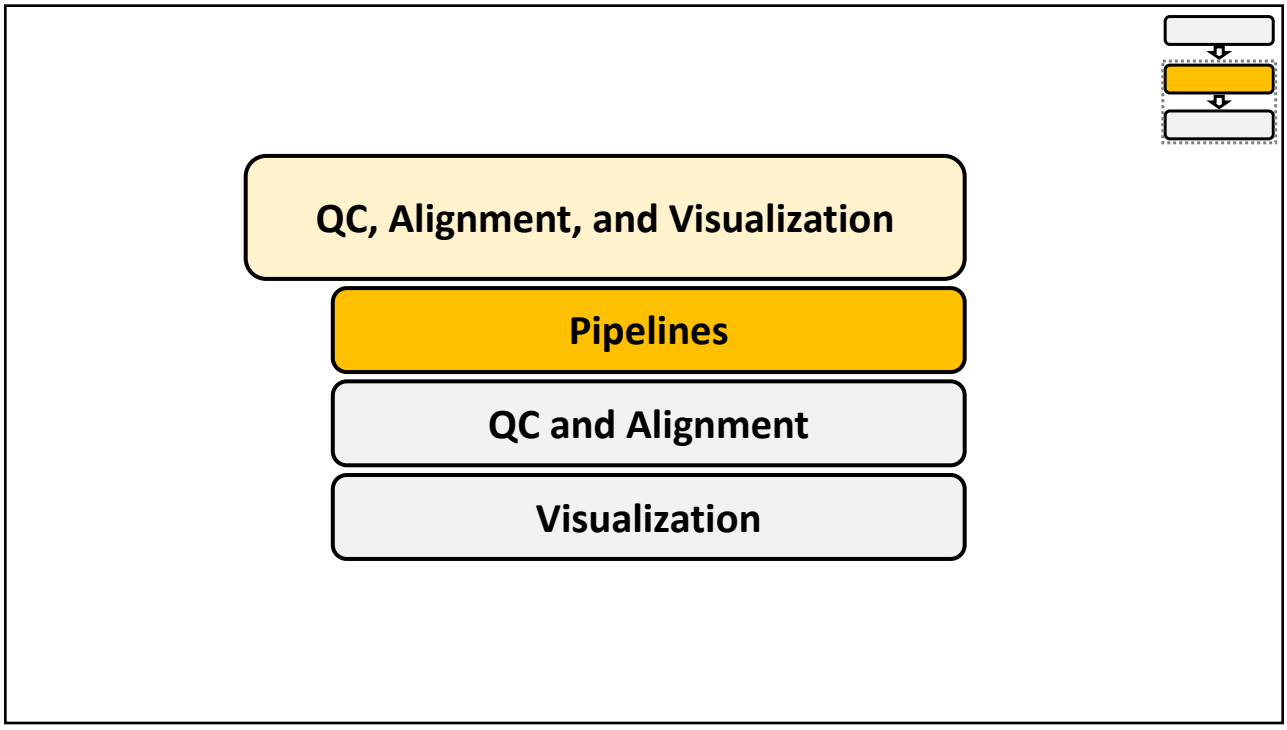
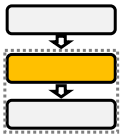
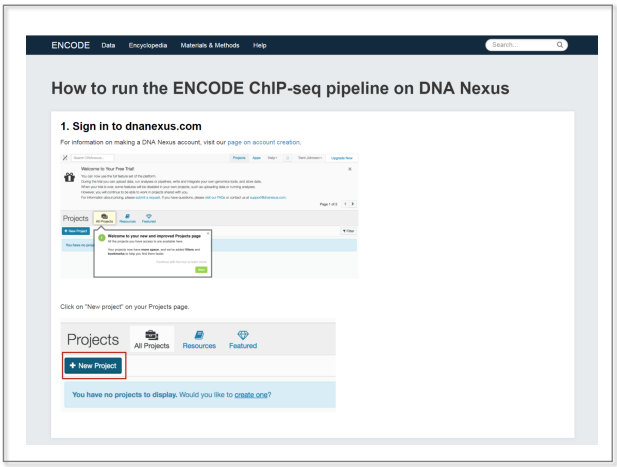

ChIP-seq Considerations
QC, Alignment, and Visualization
Peak Calling and Follow Up Analysis


QC, Alignment, and Visualization
Pipelines
QC and Alignment
Visualization

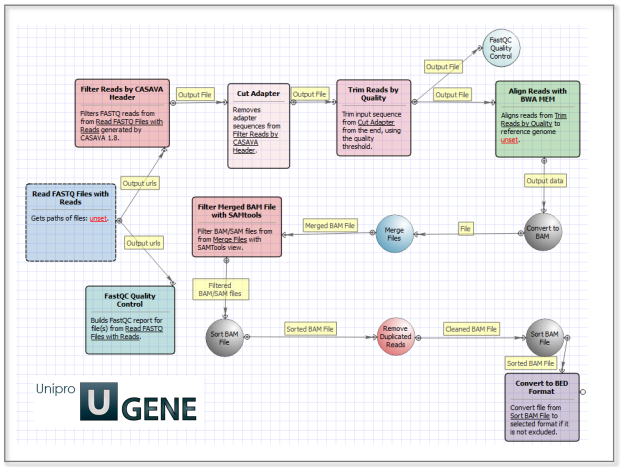
ChIP-seq Pipelines



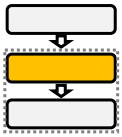
Cloud-based



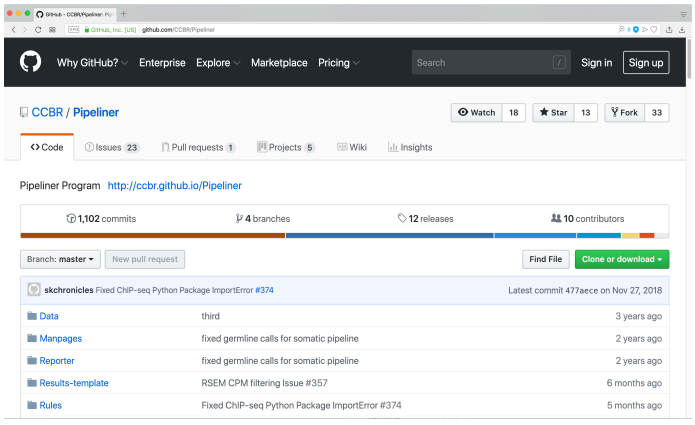
Local http://ugene.net



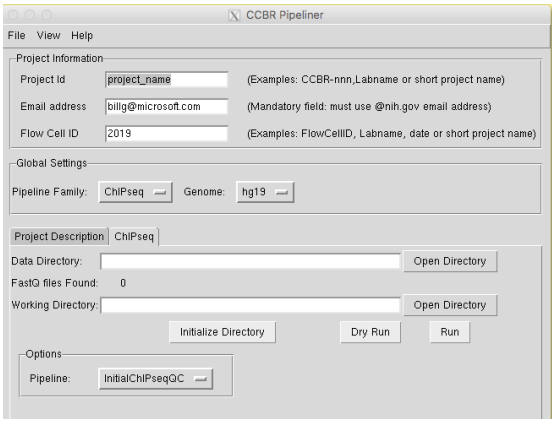
ChIP-seq Pipeline for Biowulf



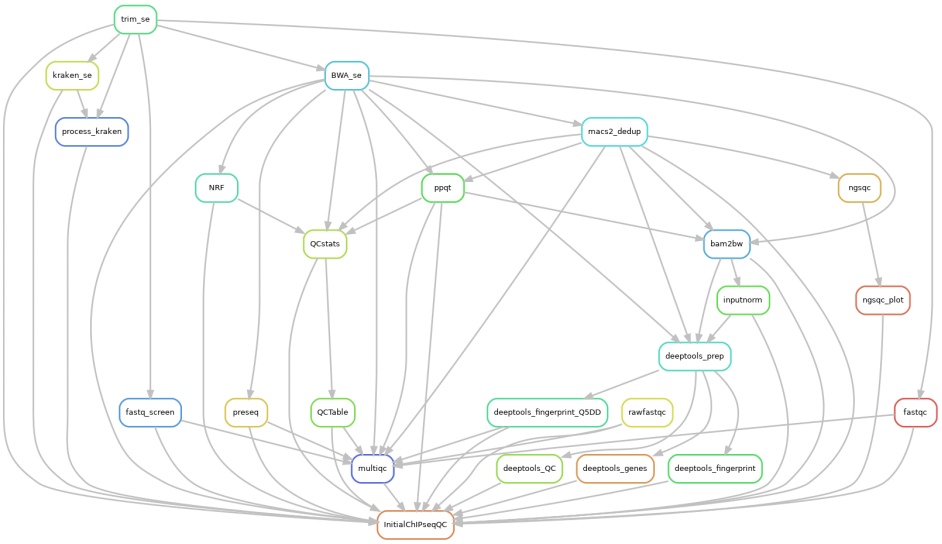
github.com/CCBR/Pipeliner



GUI



CCBR/NCBR ChIP-seq Pipeline



QC, Alignment, and Visualization

Pipelines

QC and Alignment

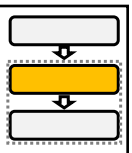
Pre-alignment QC

Alignment

Post-alignment QC

Visualization

File Formats



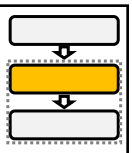
FASTA

```
>Universal Adapter
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
```

FASTQ

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=72
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACCAAGTTACCCTTAACAACTTAAGGGTTTTCAAATAGA
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=72
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9ICIIIIIIIIIIIIIIIIIIIDIIIIIII>IIIIII/
@SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=72
GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGAAGCAGAAGTCGATGATAATACGCGTCGTTTTATCAT
+SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=72
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBIIIIIIIIIIIIIIIIIIIIIGII>IIIII-I)8I
```

Trimming



Williams *et al. BMC Bioinformatics* (2016) 17:103
DOI 10.1186/s12859-016-0956-2

BMC Bioinformatics

RESEARCH ARTICLE Open Access

CrossMark

Trimming of sequence reads alters RNA-Seq gene expression estimates

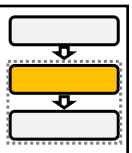
Claire R. Williams[1], Alyssa Baccarella[2], Jay Z. Parrish[1*] and Charles C. Kim[2,3*]

Abstract

Background: High-throughput RNA-Sequencing (RNA-Seq) has become the preferred technique for studying gene expression differences between biological samples and for discovering novel isoforms, though the techniques to analyze the resulting data are still immature. One pre-processing step that is widely but heterogeneously applied is trimming, in which low quality bases, identified by the probability that they are called incorrectly, are removed. However, the impact of trimming on subsequent alignment to a genome could influence downstream analyses including gene expression estimation; we hypothesized that this might occur in an inconsistent manner across different genes, resulting in differential bias.

Results: To assess the effects of trimming on gene expression, we generated RNA-Seq data sets from four samples

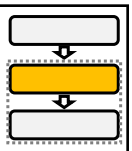
Many tools are available

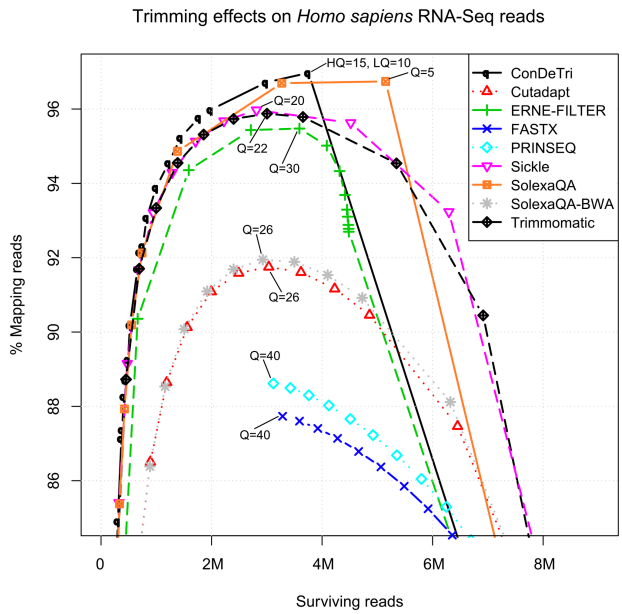


• BBDUK	• FASTX-Toolkit	• Sickle
• Biopieces	• Goby	• Trimgalore
• **Cutadapt**	• ngs_backbone	• **Trimmomatic**

More than 30 published adapter trimming tools…

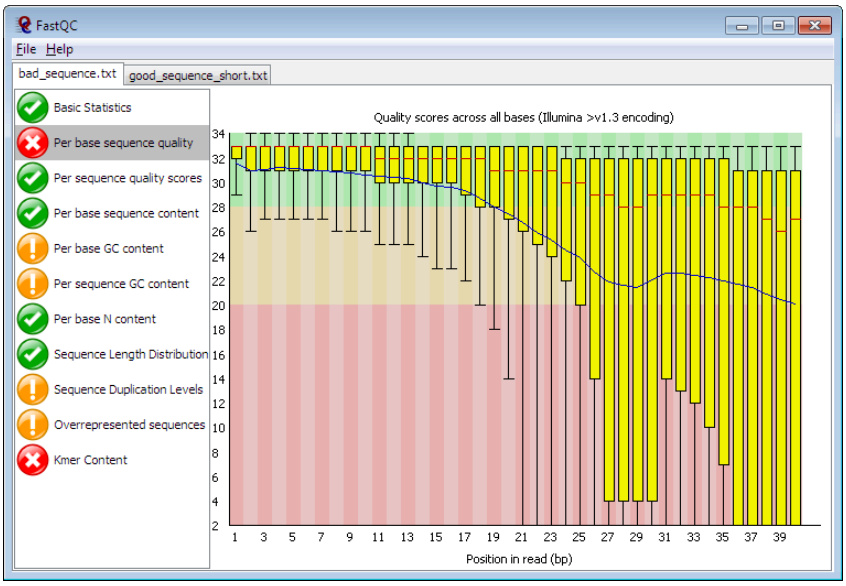
Trimming programs





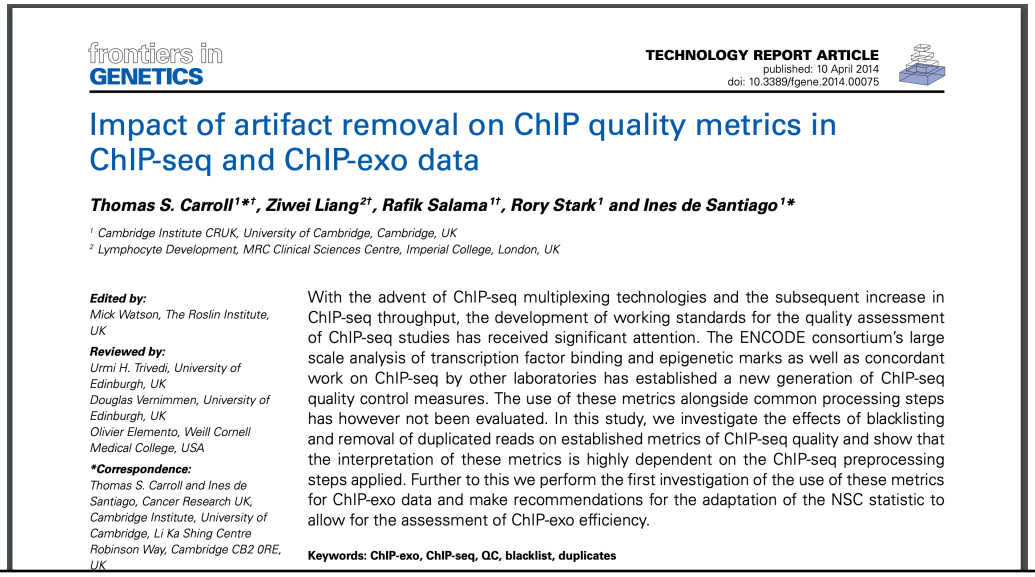
Del Fabbro et al 2013. PLOS One

FastQC



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

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*

Edited by:
Mick Watson, The Roslin Institute, UK

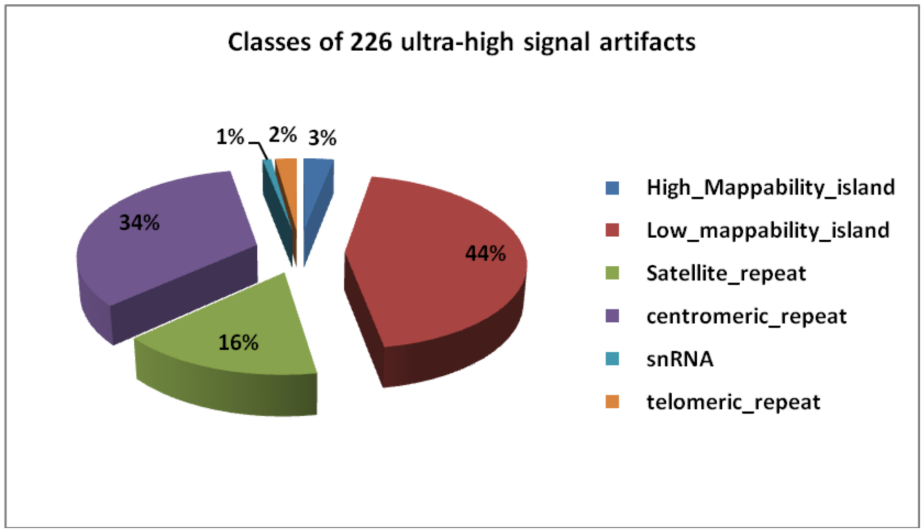
Reviewed by:
Urmi H. Trivedi, University of Edinburgh, UK
Douglas Vernimmen, University of Edinburgh, UK
Olivier Elemento, Weill Cornell Medical College, USA

****Correspondence:***
Thomas S. Carroll and Ines de Santiago, Cancer Research UK, Cambridge Institute, University of Cambridge, Li Ka Shing Centre Robinson Way, Cambridge CB2 0RE, UK

With the advent of ChIP-seq multiplexing technologies and the subsequent increase in ChIP-seq throughput, the development of working standards for the quality assessment of ChIP-seq studies has received significant attention. The ENCODE consortium's large scale analysis of transcription factor binding and epigenetic marks as well as concordant work on ChIP-seq by other laboratories has established a new generation of ChIP-seq quality control measures. The use of these metrics alongside common processing steps has however not been evaluated. In this study, we investigate the effects of blacklisting and removal of duplicated reads on established metrics of ChIP-seq quality and show that the interpretation of these metrics is highly dependent on the ChIP-seq preprocessing steps applied. Further to this we perform the first investigation of the use of these metrics for ChIP-exo data and make recommendations for the adaptation of the NSC statistic to allow for the assessment of ChIP-exo efficiency.

Keywords: ChIP-exo, ChIP-seq, QC, blacklist, duplicates

Blacklists



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

QC, Alignment, and Visualization

Pipelines

QC and Alignment

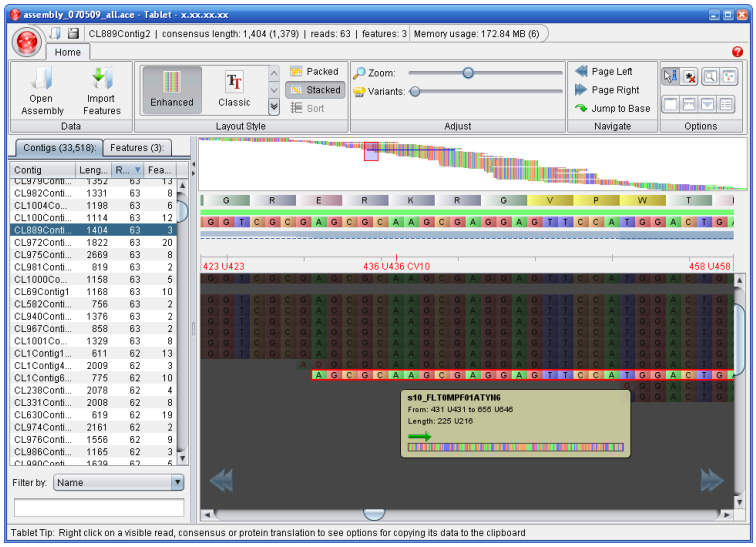
Pre-alignment QC

Alignment

Post-alignment QC

Visualization

Mapping



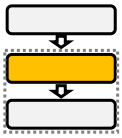
- Bowtie
- BWA
- ELAND
- HISAT
- MAQ
- NovoAlign
- SOAP
- STAR
- …and others

SAM format

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Medium.com

Best aligner?



"The evaluation of Bowtie, Bowtie2, BWA, mrsFAST, and Novoalign show their ability to correctly map the reads. Moreover, **Novoalign mapped the largest percentage** of reads, similar to GSNAP, specially for highly repeated genomes. However, it maintained the lowest throughput among the genome indexing tools in most of the experiments"

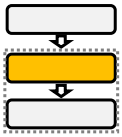
"In general, **there is no *the-best* tool** among all of the tools; each tool was *the-best* in certain conditions. The short sequence mapping problem is still an active problem and new tools are needed to be developed"

Hatem et al, *BMC Bioinformatics* 2013

"So Bowtie is definitely faster and we are able to reproduce the sensitivity gain, however if you account for false-positives, **BWA clearly wins out**"

seqanswers.com

The Bowtie sequence aligner was originally developed by Ben Langmead *et al.* at the **University of Maryland** in 2009. It is Bowtie is open-source software and is currently maintained by **John Hopkins University**.



QC, Alignment, and Visualization

Pipelines

QC and Alignment

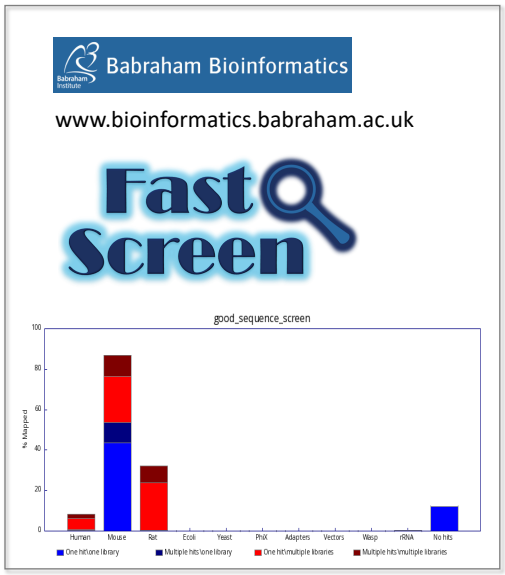
Pre-alignment QC

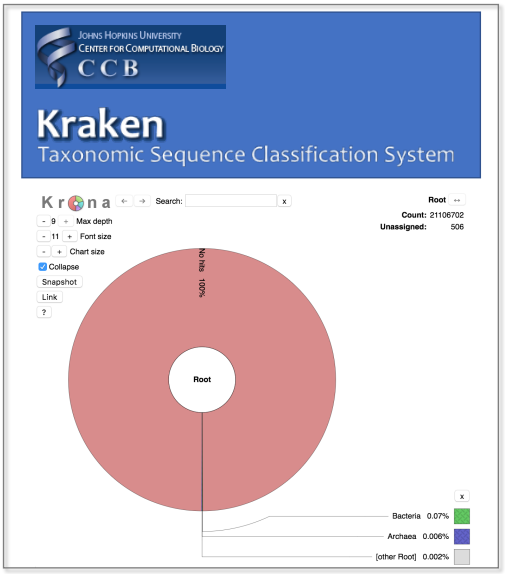
Alignment

Post-alignment QC

Visualization

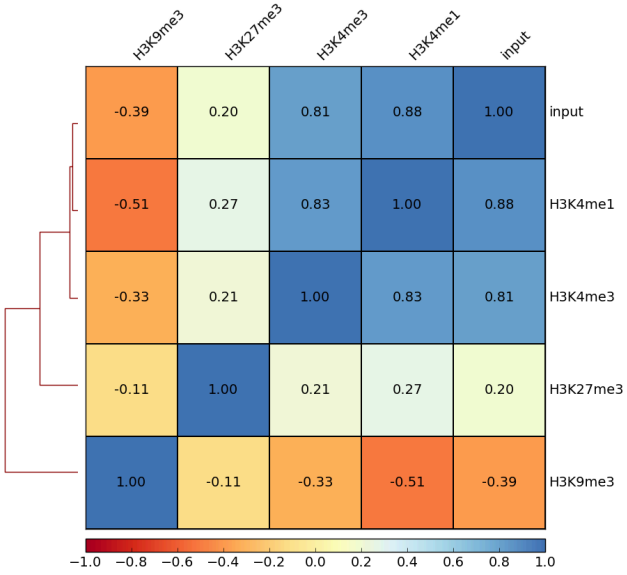
Contamination





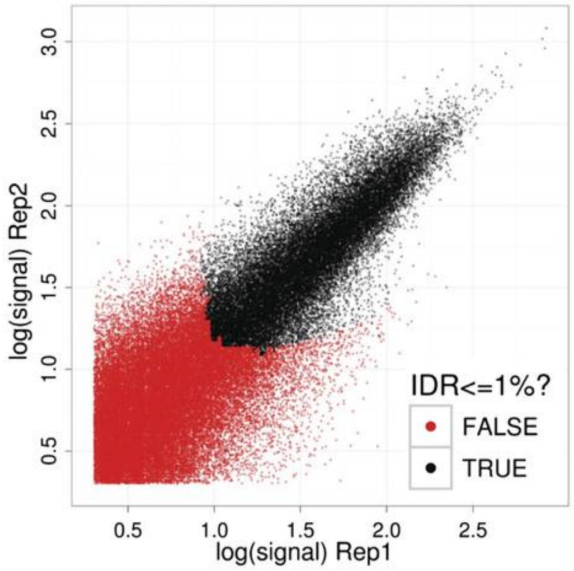
Wingett. 2018. F1000 Res Wood. 2014. Genome Biol

Correlation



deeptools.readthedocs.io

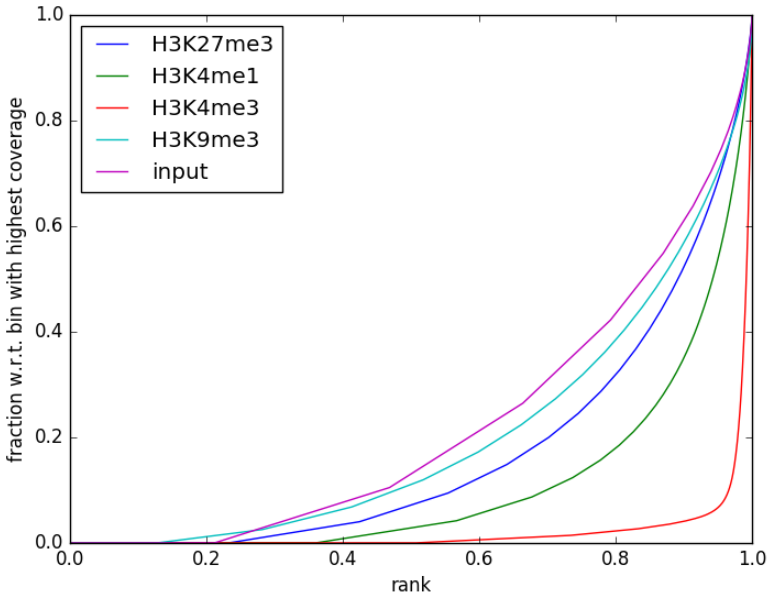
The irreproducible discovery rate (IDR)



NB: ENCODE developers do NOT recommend using as it is for broad chromatin marks ChIP-seq

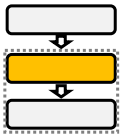
Landt et al 2012. Genome Res

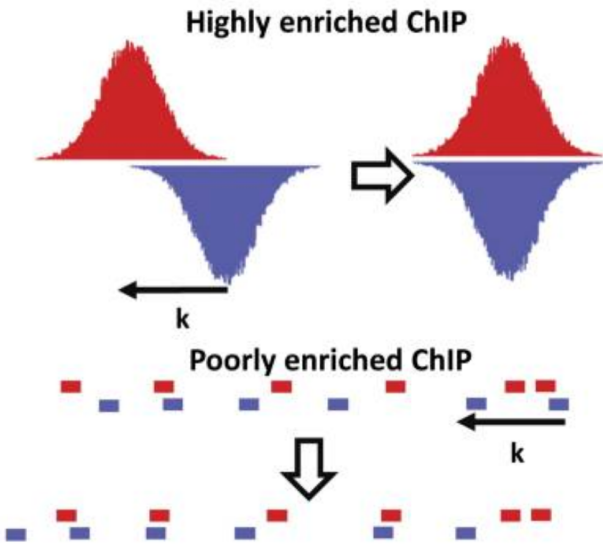
Fingerprint plot



deeptools.readthedocs.io

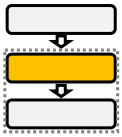
Generation of cross-correlation plot

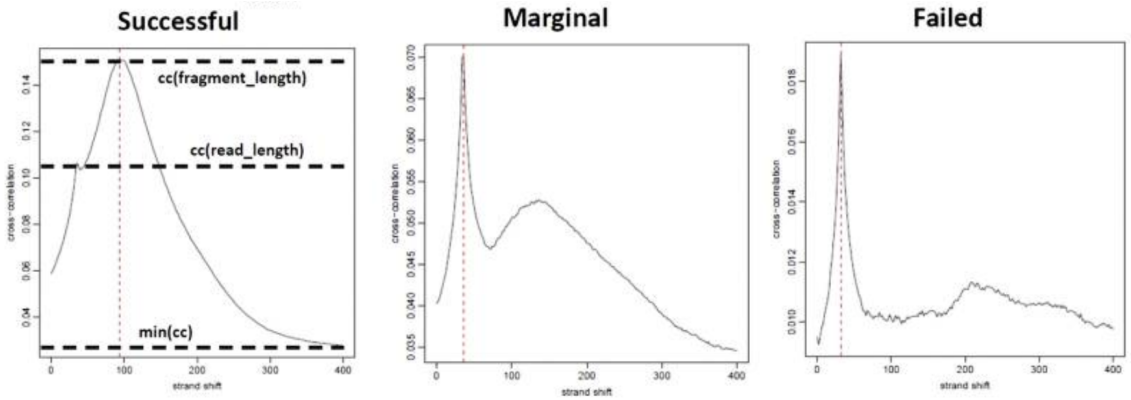




Landt et al 2012. Genome Res

Phantom peaks





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

Landt et al 2012. Genome Res

Library Complexity

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

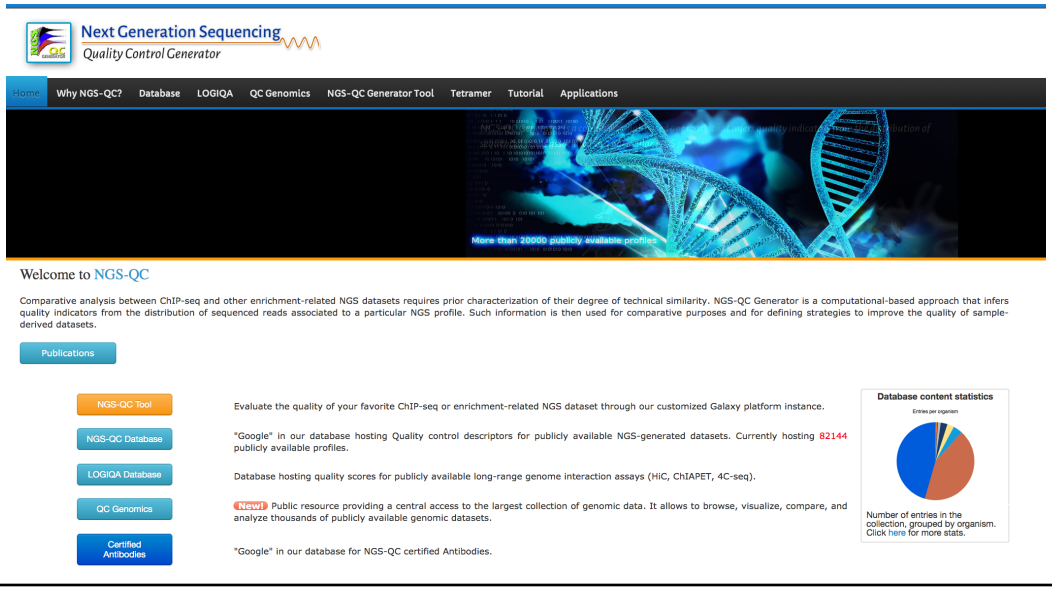
PCR Bottleneck Coefficient (PBC) shows how skewed the distribution of read counts per location is towards 1 read per location.
PBC = N1/Nd
(N1= number of genomic locations to which EXACTLY one unique mapping read maps;
Nd = the number of genomic locations to which AT LEAST one unique mapping read maps, i.e. the number of non-redundant, unique mapping reads)

Non-Redundant Fraction (NRF) - Unique Reads/Total Mapped Reads

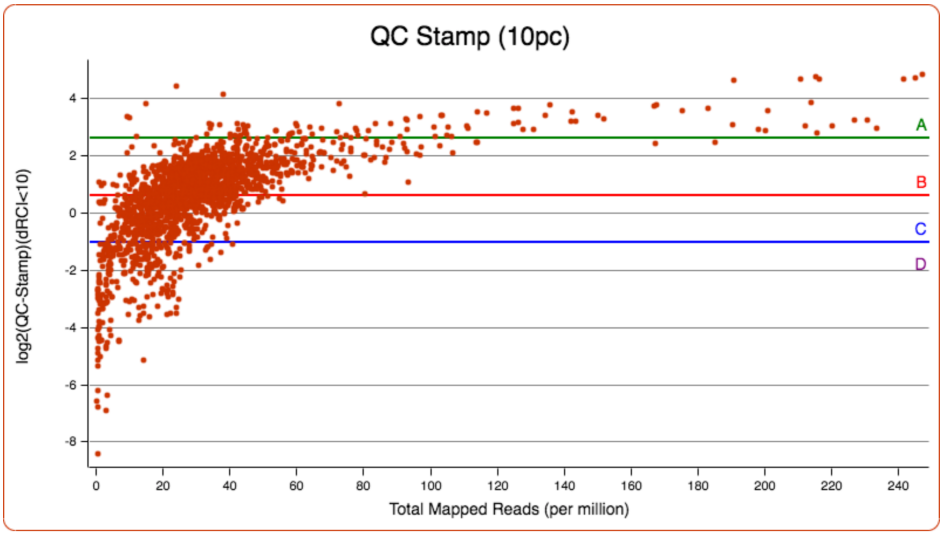
Qtag is a thresholded version of RSC (-2:veryLow, -1:Low, 0:Medium, 1:High, 2:veryHigh)

	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

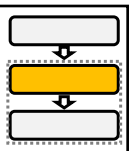
www.ngs-qc.org

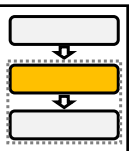


H3K27ac in humans



Clear zoom | Update Results for the selected region





QC, Alignment, and Visualization

Pipelines

QC and Alignment

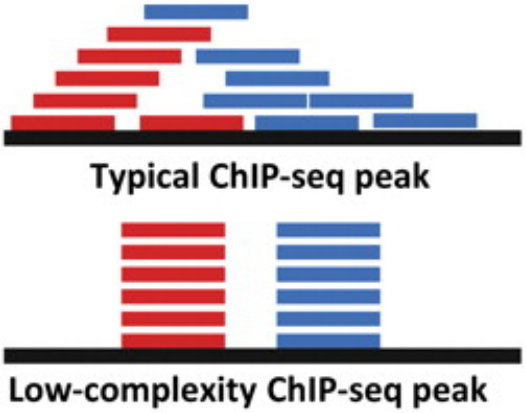
Visualization

Duplication

BigWigs

Normalization

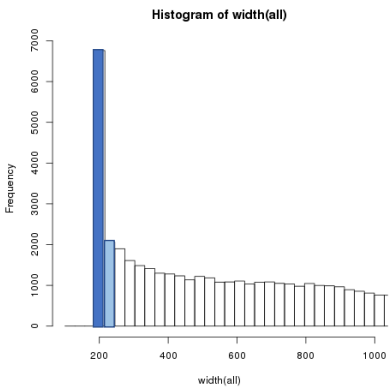
Duplication



Landt et. al. Genome Res. 2012

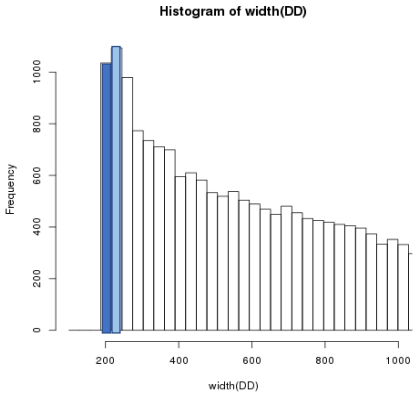
Do you need to remove duplicates?

All reads



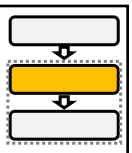
# peaks	61,314
# bases covered	84,157,874

No Duplicates

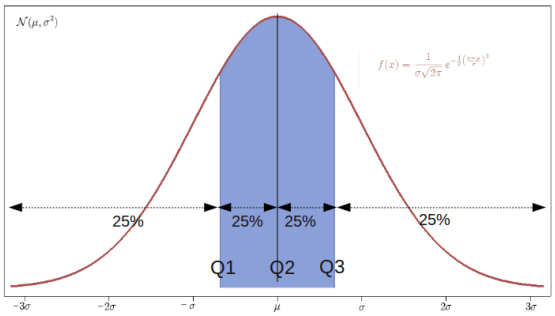


	25,175
	36,168,022

Two ways to remove duplicates

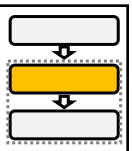


- Partial duplicate removal
  - Uses a binomial distribution of read numbers across the entire genome and removes the upper quantile.
- Remove all duplicates
  - If reads map to the same start and end position, remove all but one of the reads.

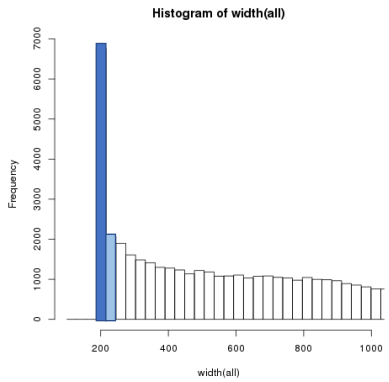


Wikipedia. 2019.

Effect of partial/total duplicate removal

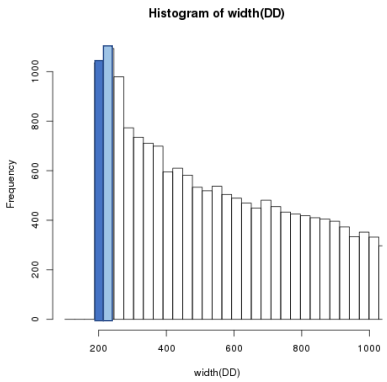


All reads



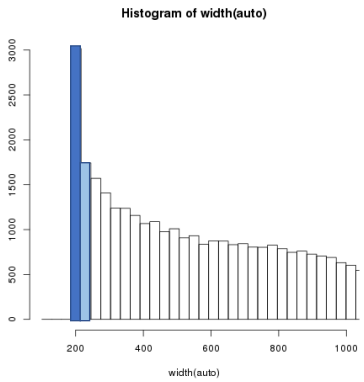
# peaks	61,314
# bases covered	84,157,874

No Duplicates



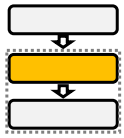
# peaks	25,175
# bases covered	36,168,022

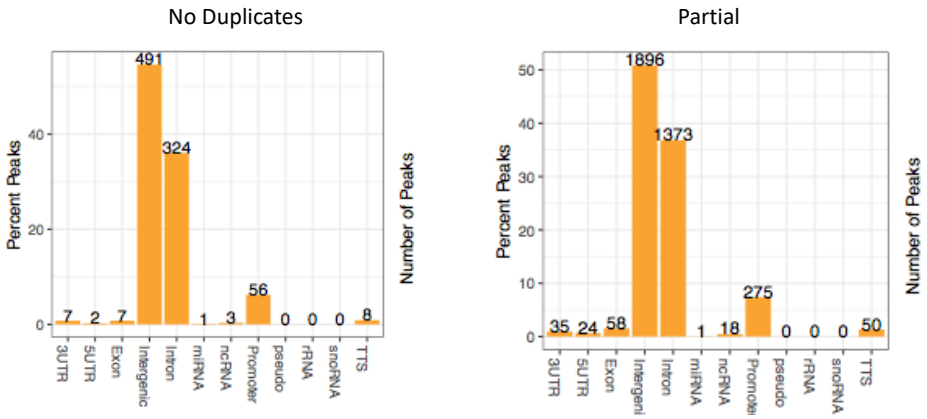
Partial

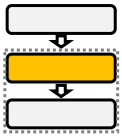


# peaks	47,479
# bases covered	69,159,165

Effect of partial/total duplicate removal







QC, Alignment, and Visualization

Pipelines

QC and Alignment

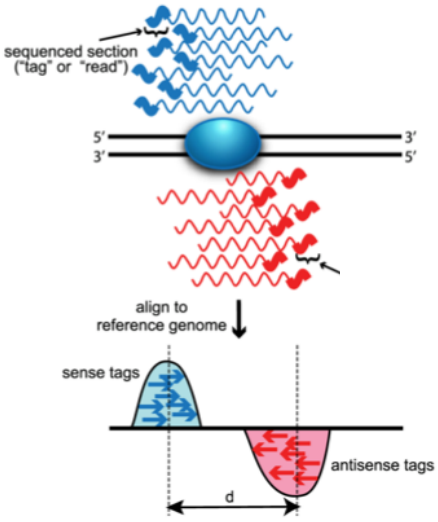
Visualization

Duplication

BigWigs

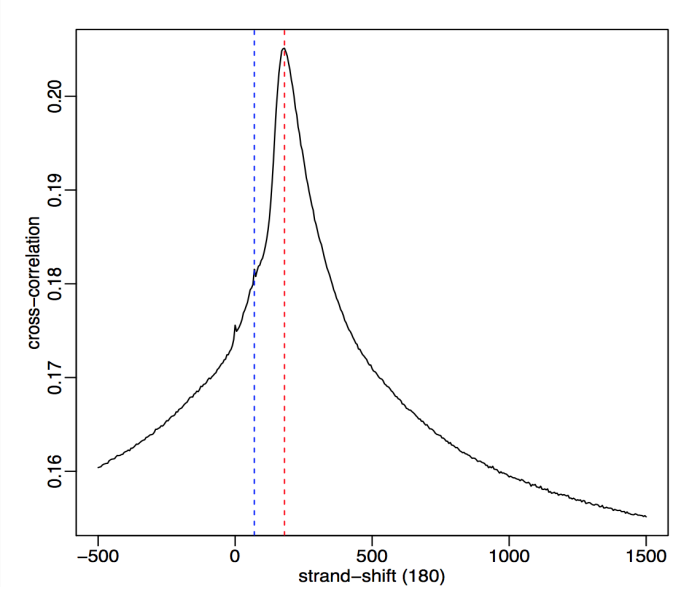
Normalization

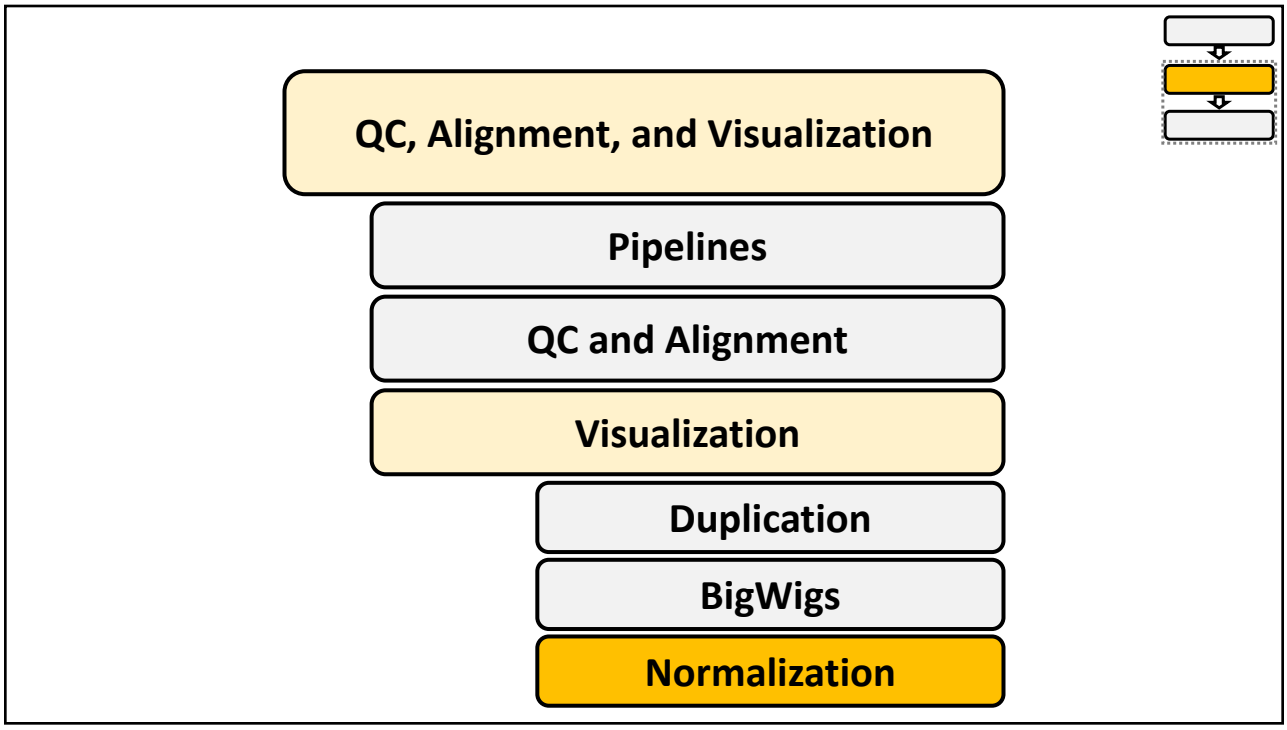
BigWig generation:
Read extension for single end sequencing data



Wilbanks et. al. PLOS ONE. 2010.

Calculating the read extension

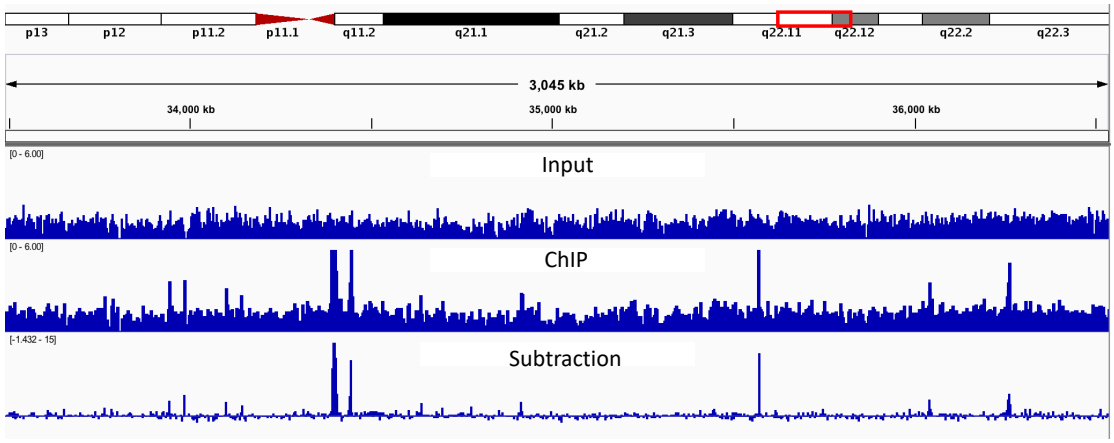




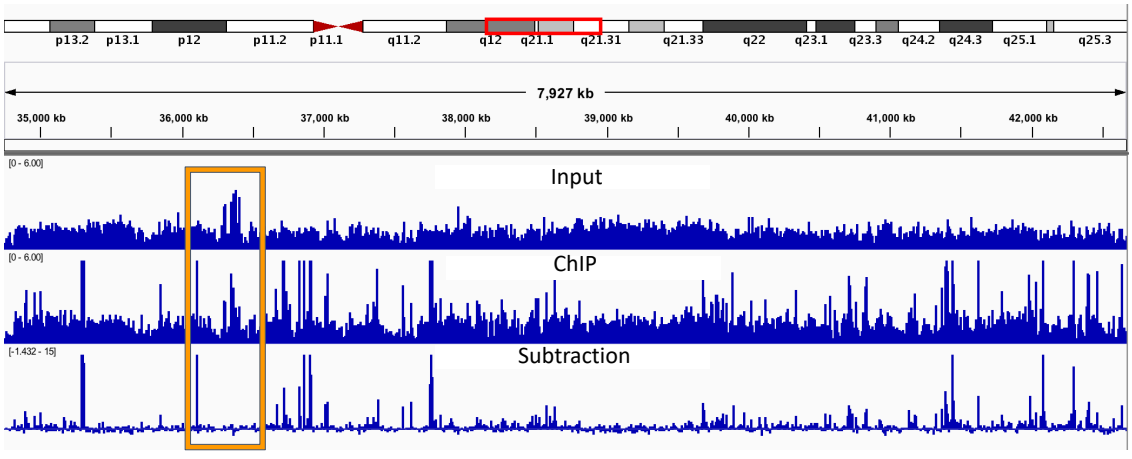
Normalization for library size

- RPKM:
  - reads per kilobase per million reads
  - defined as:
    - RPKM (per bin) = # of reads per bin / (# of mapped reads (in millions) * bin length (kp))
- 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

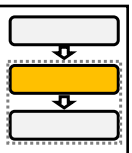
Normalization: Subtracting the input from the library normalized reads



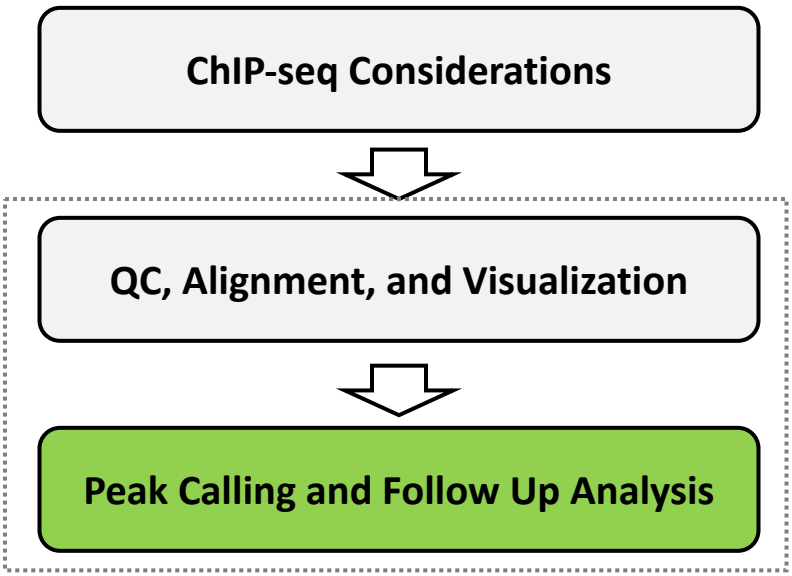
Input Subtracted Normalization

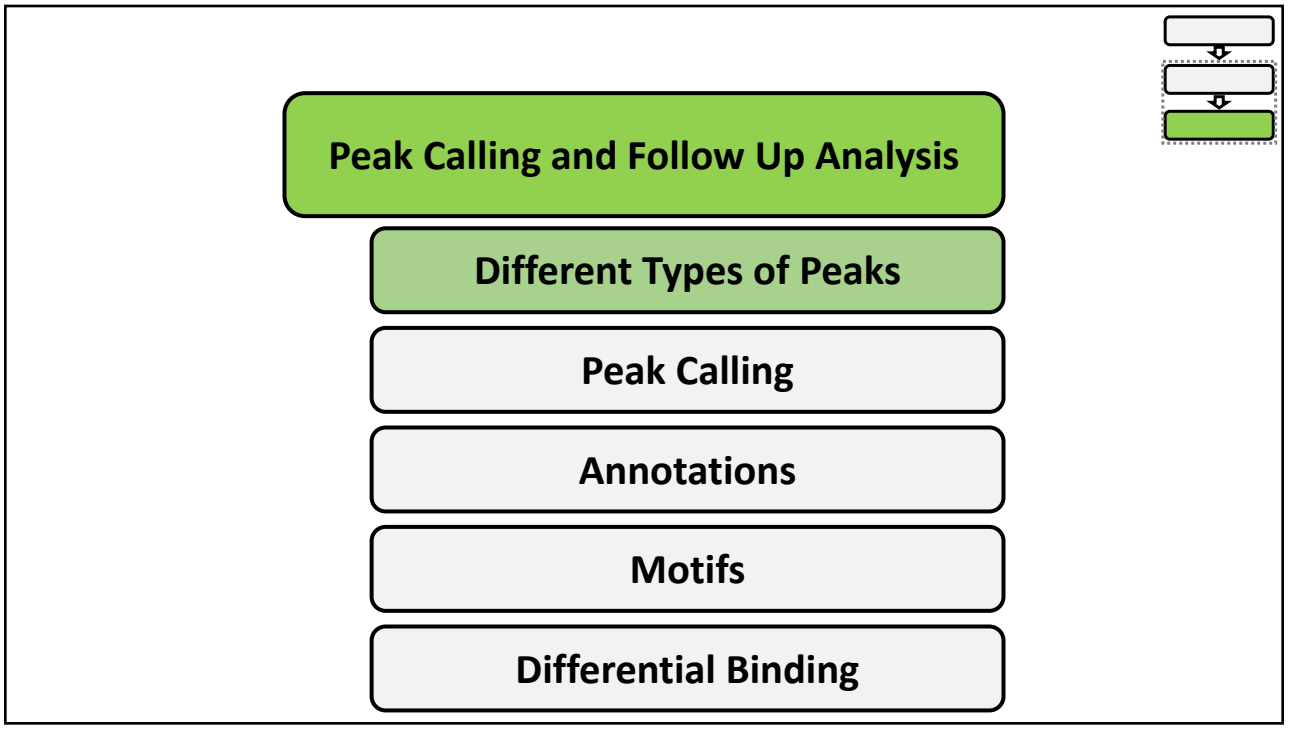


Tools



- Duplicate Removal:
  - MACS: https://github.com/taoliu/MACS
  - Picard; https://broadinstitute.github.io/picard/
- Visualization:
  - deeptools: https://deeptools.readthedocs.io/en/develop/
- Viewers:
  - IGV: https://software.broadinstitute.org/software/igv/
  - UCSC genome browser: https://genome.ucsc.edu/

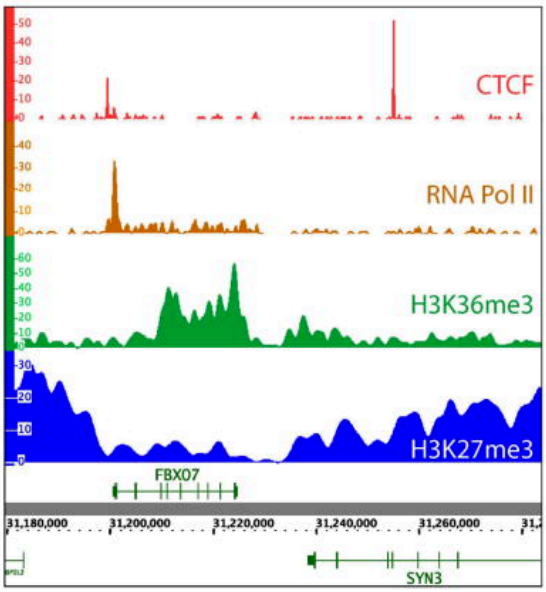




Proteins bind in different ways

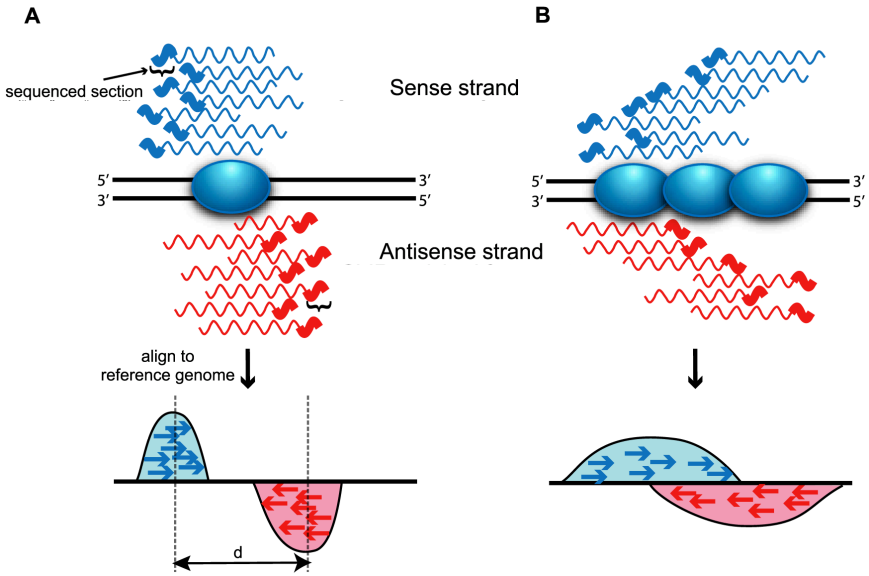
- Transcription factor
  - Tight, high peaks
- RNA Pol II
  - Enriched at TSS but bound throughout the gene body
- Histones
  - Some are sharper and located near TSS
  - Some are broader and spread out across the length of active or inactive genes

Proteins bind in different ways

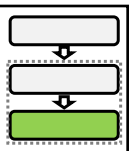


Park et al 2009. Nat Rev Genet

What causes these different shapes?

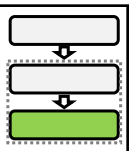


Wilbanks et al 2010. PLOS ONE

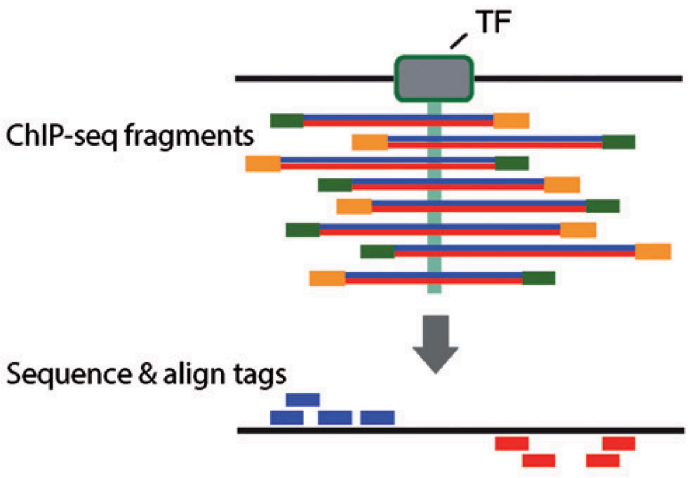


Peak Calling and Follow Up Analysis

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

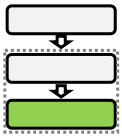


How are peaks called?

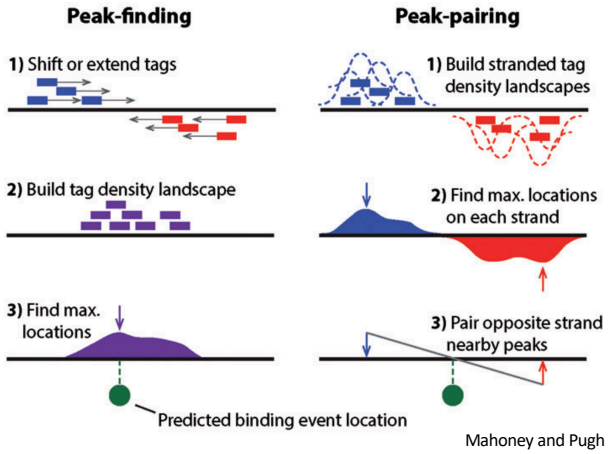


Mahoney and Pugh et al 2015. Criti Rev Biochemi and MolBio

General concept of most peak callers

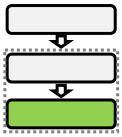


Count the number of reads within a window and determine whether this number is above background



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

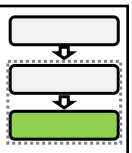
There are many peak callers out there…



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
TM	SISSRs	PolyPeak	spp-mtc	etc…

Thomas et al 2017. Briefings in Bioinformatics

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			

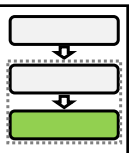
X* = Windows-only GUI or cross-platform command line interface

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

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

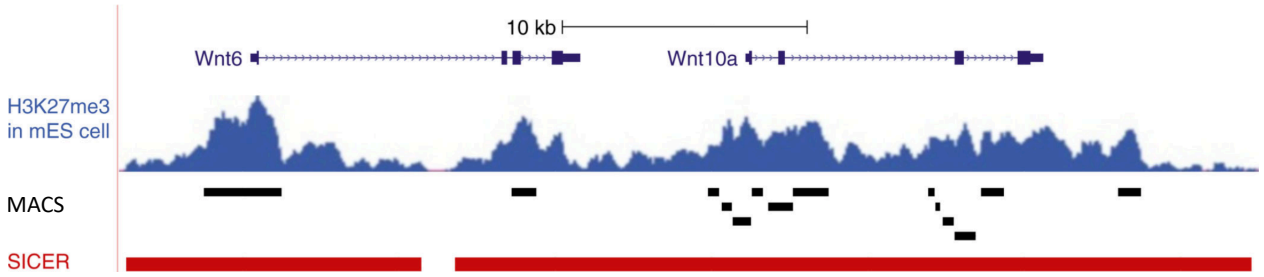
Wilbanks et al 2010. PLOS ONE

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

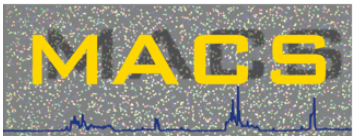
MACS works well for narrow peaks while SICER is designed for broad peaks



Xu et al 2014. Methods Mol Biol

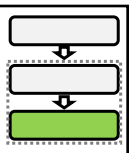
Model-based Analysis of ChIP-Seq (MACS)

- 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

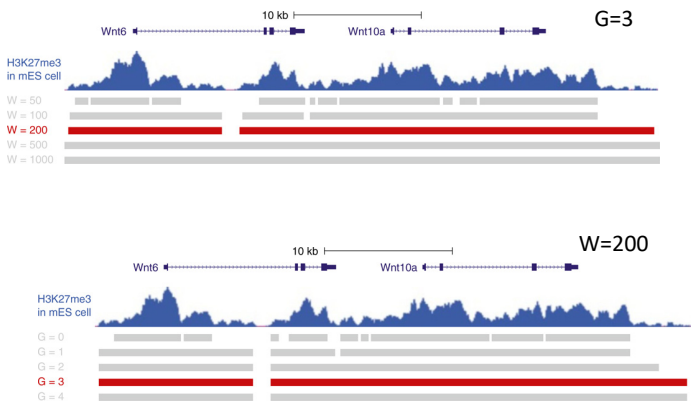


Feng et al 2012. Nature Protocols

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

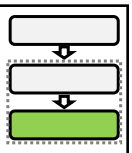


- Uses windows and gaps to identify ”islands” of enrichment
- Gaps allow for short regions lacking binding within an island, more pattern variability across island
- Compares to a randomized background and control background to calculate FDR



Xu et al 2014. Methods Mol Biol

Output file formats



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

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

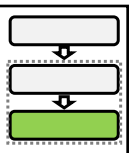
This format is used to provide called peaks of signal enrichment based on pooled, normalized (interpreted) data. It is a BED6+4 forr

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.

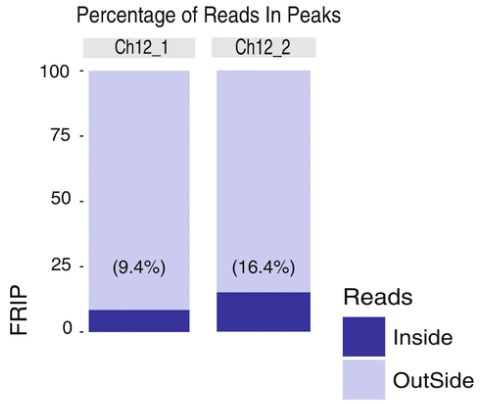
Here is an example of narrowPeak format:

```
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
```

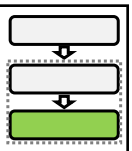
FRiP (Fraction of Reads in Peaks)



- 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

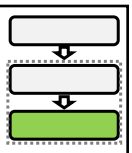
Peak Calling

Annotations

Motifs

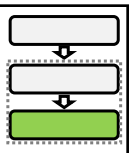
Differential Binding

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

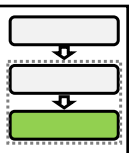
UROPA

- 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

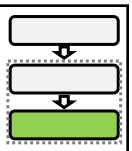
Annotation tools: example HOMER output table

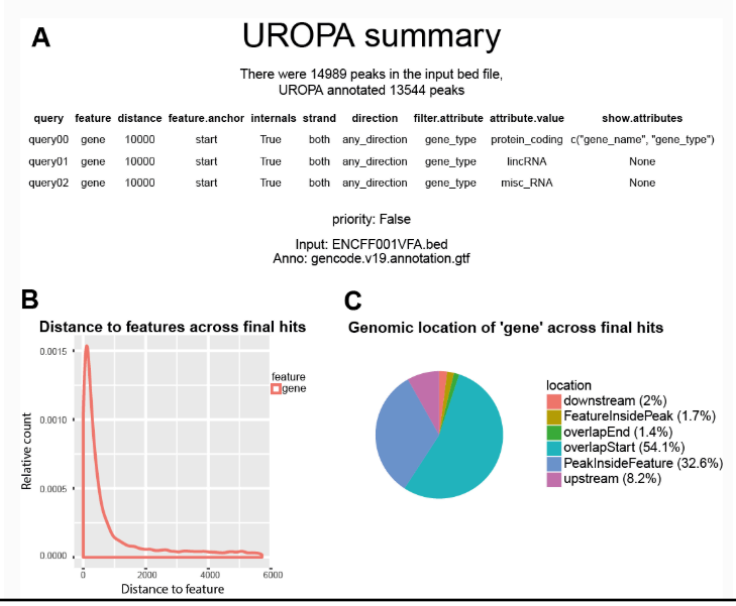


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	PeakID	Chr	Start	End	Strand	Peak Sco	Focus Ra	Annotation	Detailed Anno	Distance to T	Nearest Pror	PromoterID	Nearest Unig	Nearest Refs	Nearest Ense	Gene Name	Gene Alias	Gene Descrip
2	chr18-1	chr18	69007968	69008268	+	593	0.939	intron (NR_03	intron (NR_03	74595	NR_034133	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	transmembr
12	chr14-3	chr14	88712188	88712488	+	443.1	0.844	intron (NM_0(	intron (NM_0(	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

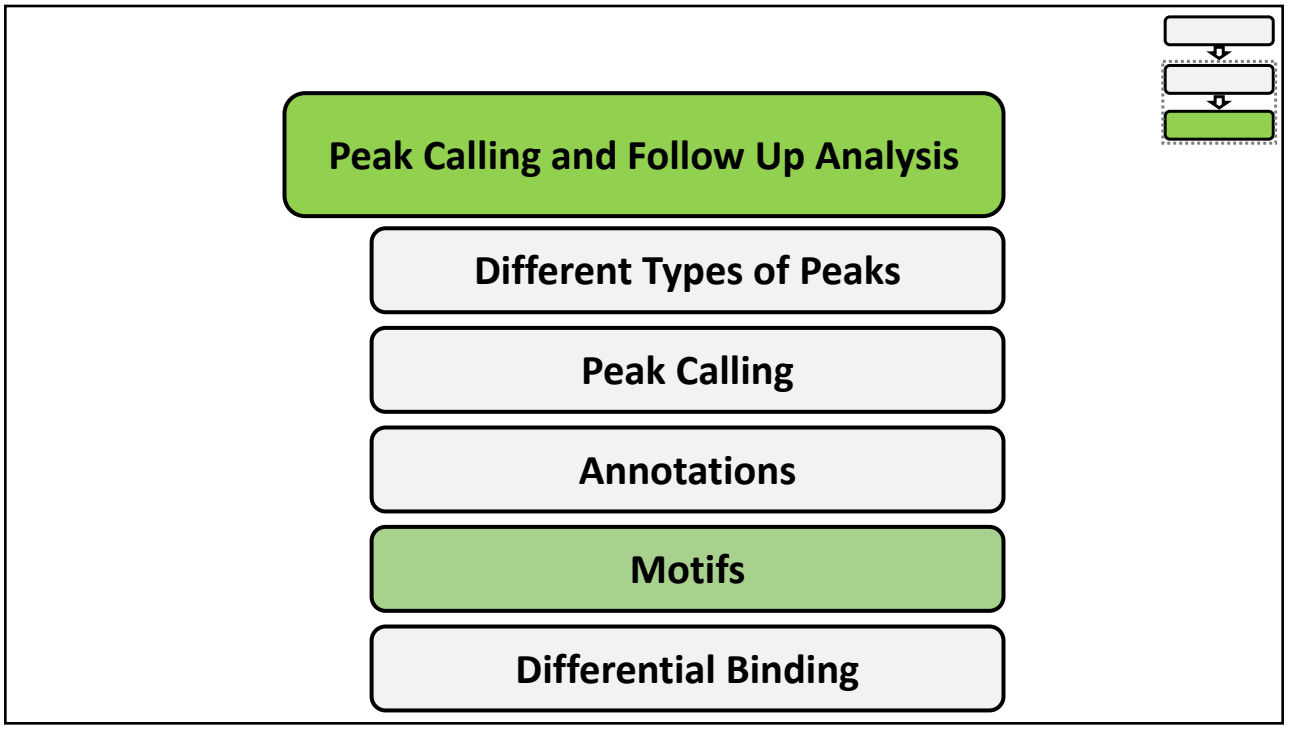
Heinz et al 2010. Mol Cell

UROPA output figures



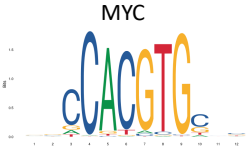


Kondili et al 2017. Scientific Reports

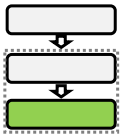


Motifs: things to consider

- 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?



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

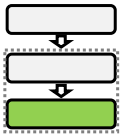
AME
Analysis of Motif Enrichment

HOMER

- Runs for both known and novel motifs simultaneously

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

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.

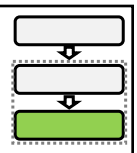
Select the type of control sequences to use
Shuffled input sequences | User-provided control sequences | NONE NEW!

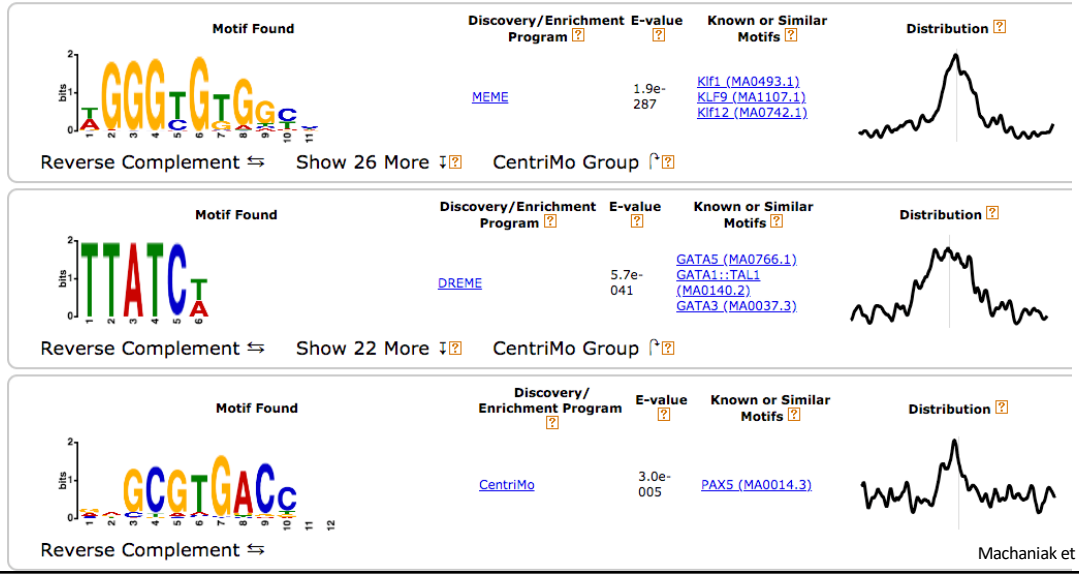
Select the sequence alphabet
Use sequences with a standard alphabet or specify a custom alphabet.
DNA, RNA or Protein | Custom | Choose File | No file chosen

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)

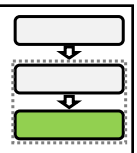
MEME-ChIP output





Machaniak et al 2011. Bioinformatics

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%)

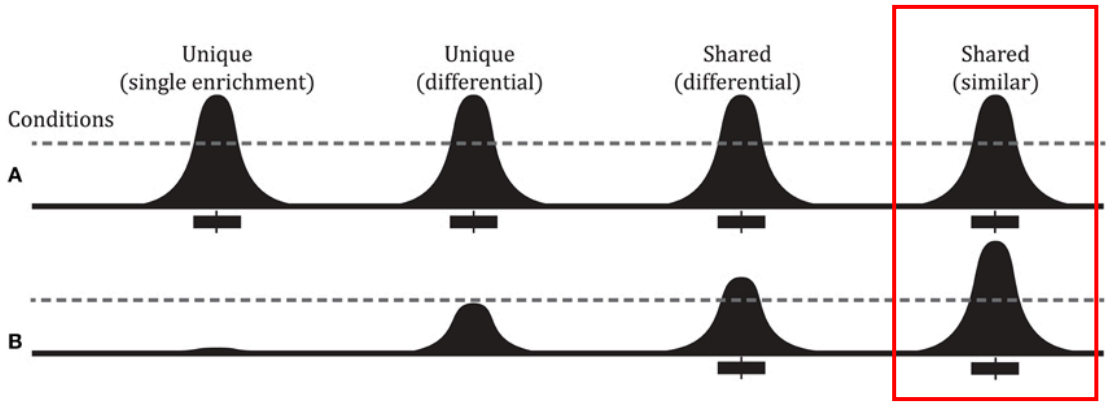
HOMER output

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		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		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)
						41.8bp	MA0102.1_Cebpa	motif

Peak Calling and Follow Up Analysis

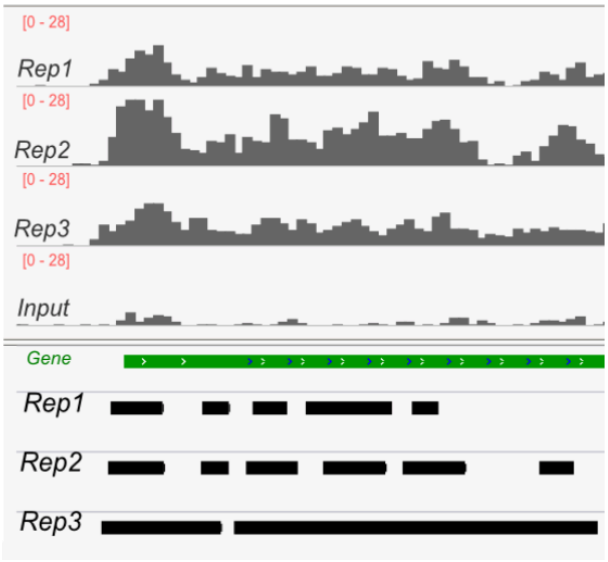
- Different Types of Peaks
- Peak Calling
- Annotations
- Motifs
- **Differential Binding**

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



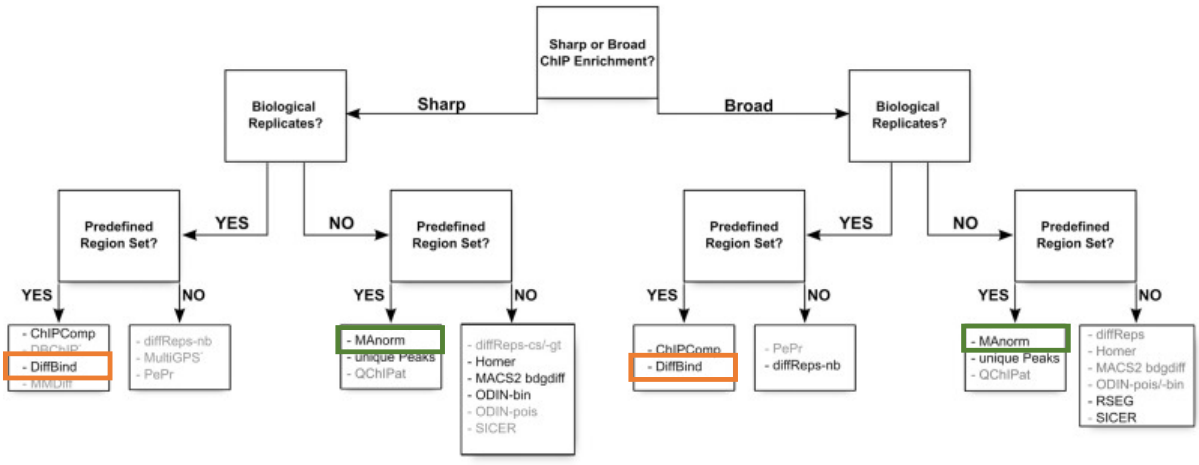
Wu et al 2015. Front Genet

Differential peak calling is dependent on peak calling quality



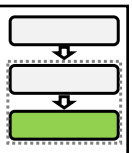
Yang et al 2014. Comput Struct Biotechnol J

Differential peak calling



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

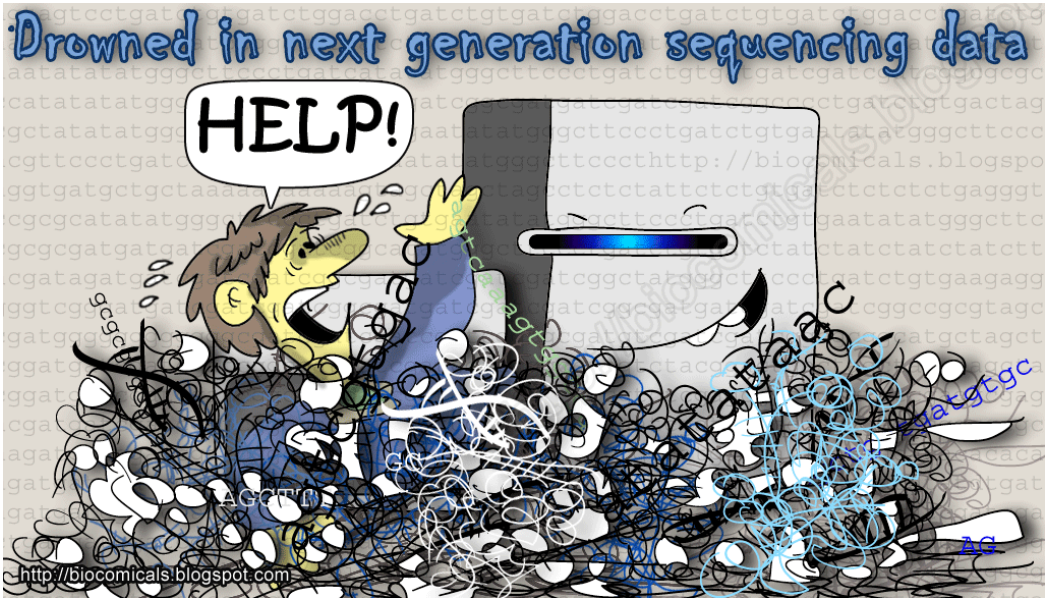
DIFFBIND

- 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

Conclusions

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


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