

# ChIP-Seq Data Analysis 1

Lauren Vanderlinden

BIOS 6660

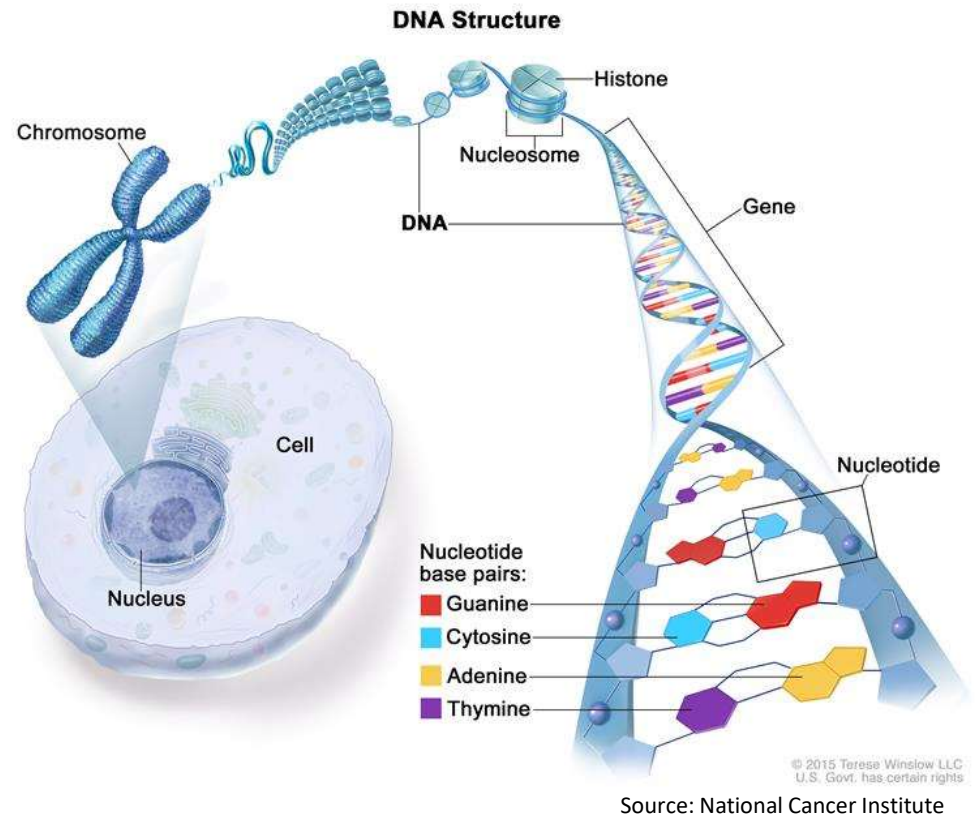
Spring 2019

# Overview From Last Time

- Walked through RNA-seq pre-processing on yampa
- Many independent scripts
- Good to have a main document which shows which scripts you used
-

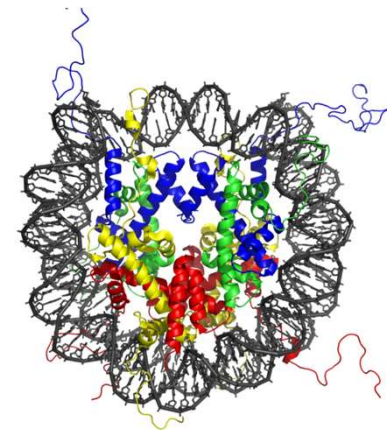
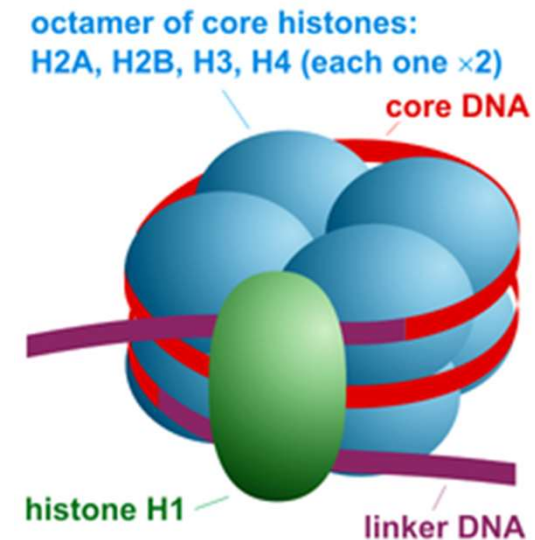
# Chromosomes

- How DNA is stored in the cell
- >3 billion bases, need to fit in a small space
- DNA is wrapped around proteins called histones which coil into chromosomes
- Chromatin is the DNA and proteins making up the chromosomes



# Nucleosome

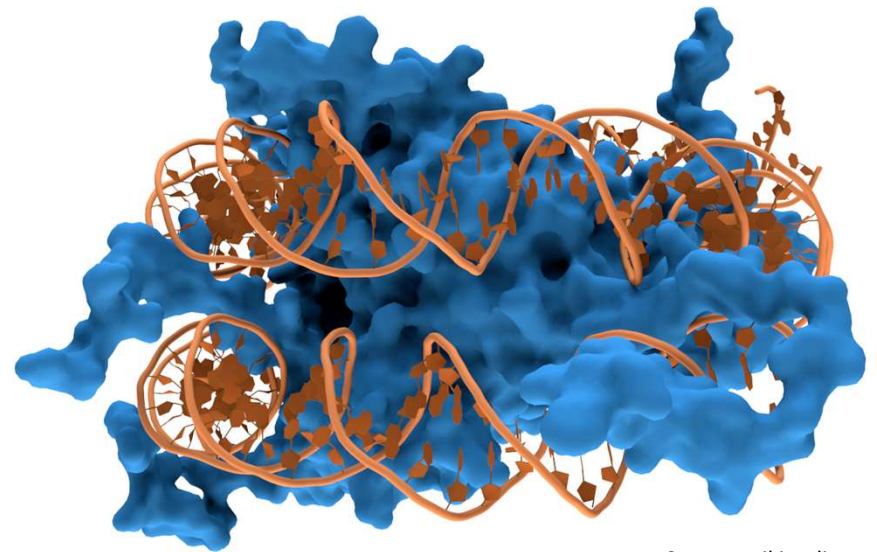
- Basic unit of packaging
- Consists of DNA wrapped around a core of 8 histone proteins
- Compared to a spool
- Nucleosome positions in the genome are not random
- Determines the accessibility of the DNA to regulatory proteins



Source: Wikipedia

# DNA-Protein Physical Interaction

- Many types of DNA-binding proteins
  - Transcription Factors
    - Regulate gene expression
  - Polymerases
    - Enzymes synthesizes long chains of polymers or nucleic acids
  - Nucleases
    - Enzyme which can cleave DNA
  - Histones
    - Chromosome packaging
- For ChIP-Seq we will be interested in transcription factors and histone binding



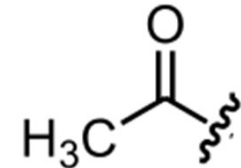
Source: wikipedia

A histone is shown in blue and DNA in gold  
3D structure is very complex

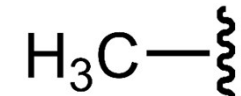
# Chromatin Remodeling

- Dynamic modification of chromatin architecture to allow access of condensed genomic DNA
- Allows regulatory transcription machinery proteins to access DNA
  - regulates gene expression
- 2 main methods:
  1. Covalent histone modifications by specific enzymes
    - Histone acetyltransferases (HATs)
    - Deacetylases
    - Methyltransferases
    - Kinases
  2. ATP-dependent chromatin remodeling complexes which either move, eject or restructure nucleosomes

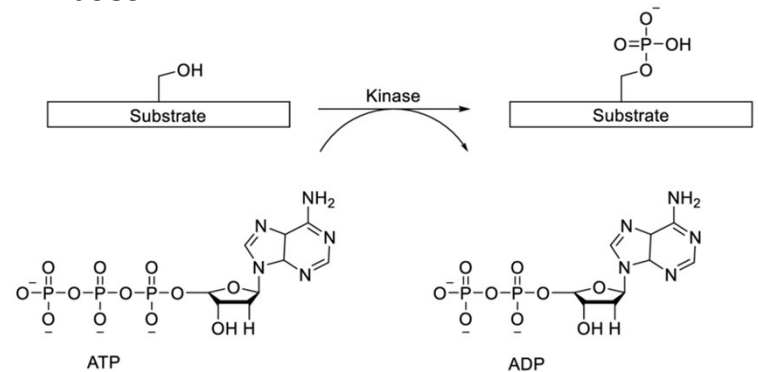
Acetyl group:



Methyl group:



Kinases:



Source: Wikipedia

# Histone Modifications

- mono-, di-, or trimethylation (M) at arginine or lysine residues
- acetylation (A) at lysine residues
- phosphorylation (P) at serine residues.
- Some amino acids, such as lysine 9 of histone H3, can be subject to either acetylation or methylation, but not both
- H3 trimethyl K4 is popular to look at

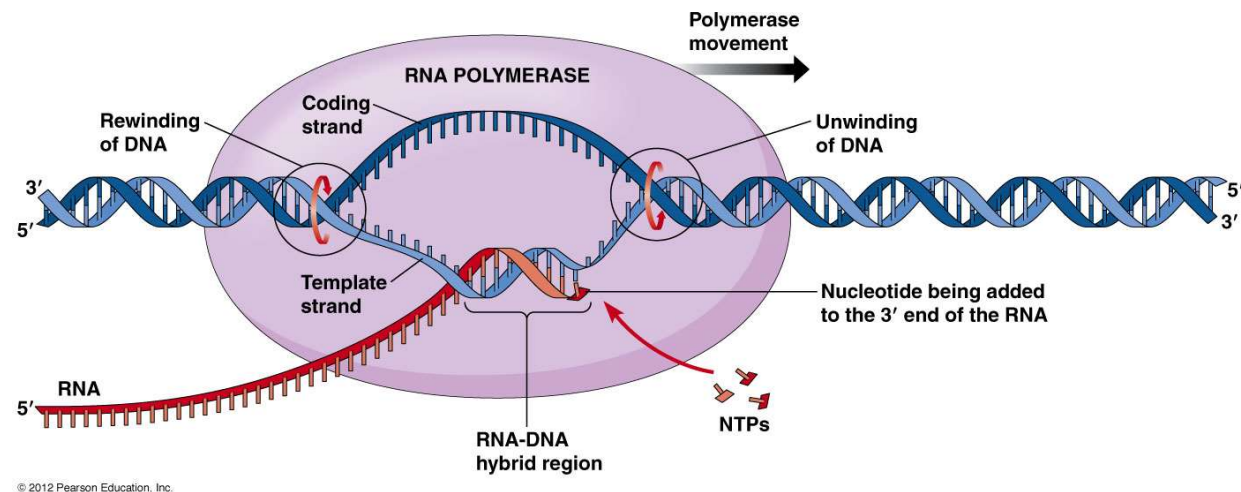


Source: Dressler et al, 2008

# Transcription



- RNA Polymerase needs to attach to DNA
- But how does it know where to go and bind to?
- Promoter regions
  - $\approx 100\text{--}1,000$  bp long right before the transcription start site (TSS)
  - TATA box
- Enhancer Regions
  - $\approx 50\text{--}1,500$  bp long
  - Can be located up to 1 Mb away
  - always cis-acting



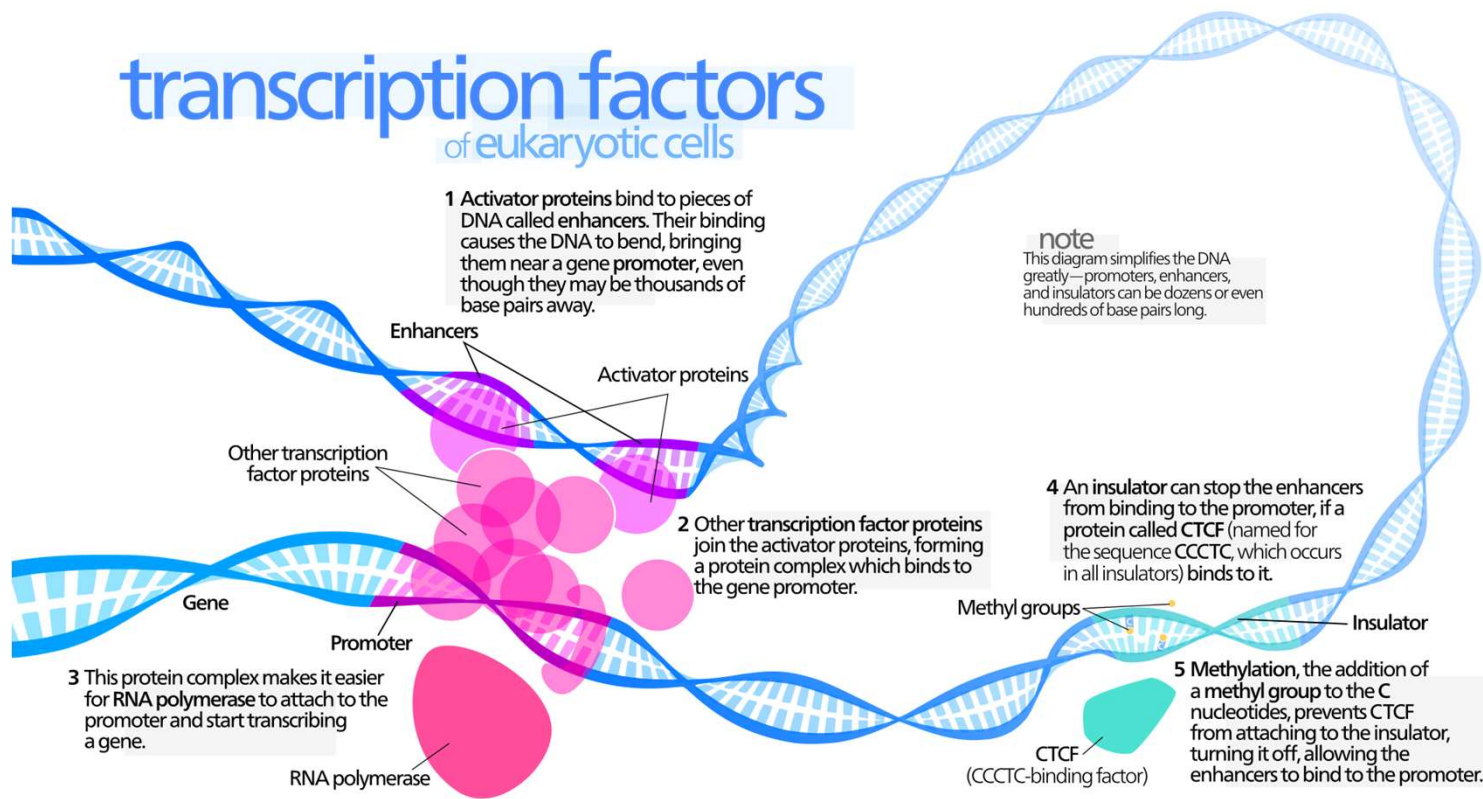
Source: Pearson Education



# Transcription Factors

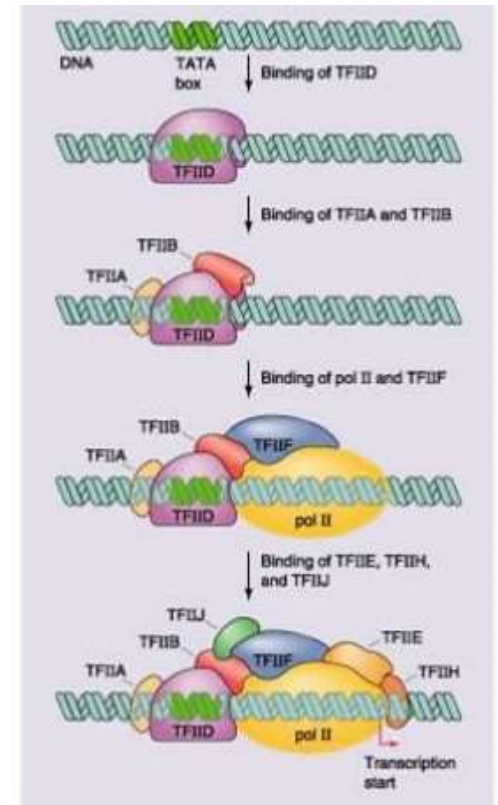
- Basal, or general, transcription factors are necessary for RNA polymerase to function at a site of transcription in eukaryotes
  - Considered the most basic set of proteins needed to activate gene transcription
- Bind to either enhancer or promoter regions of DNA adjacent to the genes that they regulate
- Can cause gene to be either **up-** or **down-** regulated
- Many mechanisms:
  - **Stabilize** or **block** the binding of RNA polymerase to DNA
  - Catalyze **acetylation** or **deacetylation** of histones
  - Recruit **coactivator** or **corepressor** proteins to the transcription factor DNA complex

# transcription factors of eukaryotic cells



Source: wikipedia

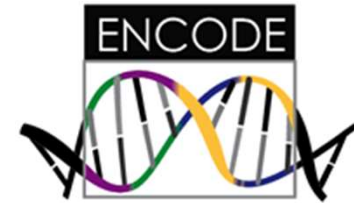
## Example of Basal TF



Source: slideServe

# ENCODE

- <https://www.encodeproject.org/>



ENCODE Data Encyclopedia Materials & Methods Help Search...

## ENCODE: Encyclopedia of DNA Elements

Diagram illustrating the ENCODE project's approach to identifying DNA elements. The diagram shows a DNA strand with various elements: Long-range regulatory elements (enhancers, repressors/silencers, insulators), Promoters, Genes, and Transcripts. Above the DNA, various experimental techniques are mapped to specific elements: 5C, ChIA-PET, Hi-C for long-range elements; DNase-seq, FAIRE-seq, ATAC-seq for promoters; ChIP-seq for genes; WGBS, RRBS, methyl array for transcripts; and Computational predictions, RNA-seq, CLIP-seq, RIP-seq for other elements. The diagram also shows Hypersensitive Sites, CH<sub>3</sub>, CH<sub>3</sub>CO, and RNA polymerase.

Buttons: About ENCODE Project, Getting Started, Experiments

Search ENCODE portal ⓘ

ENCODE Q

Buttons: About ENCODE Encyclopedia, Candidate Regulatory Elements

Search for Candidate Regulatory Elements ⓘ  
Hosted by SCREEN

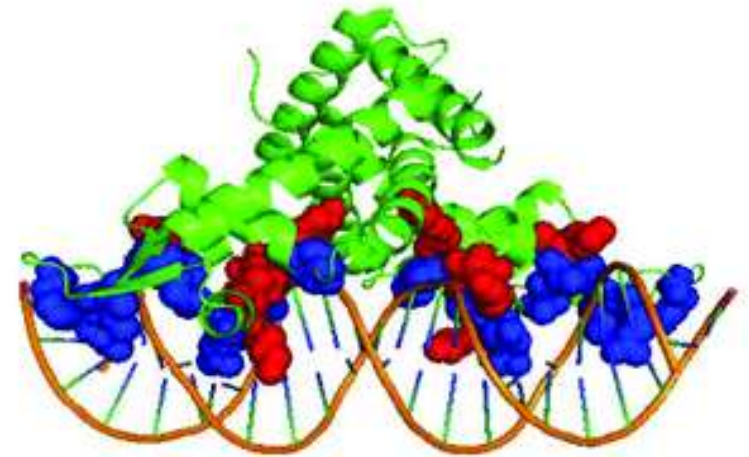
Buttons: Human hg19 Q, Mouse mm10 Q

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

# ChIP-Seq

1. Chromatin immunoprecipitation (**ChIP**)
2. DNA sequencing (**Seq**)

- **Goal:** identify the binding sites of DNA-associated proteins.



Source: Shen et al, 2017

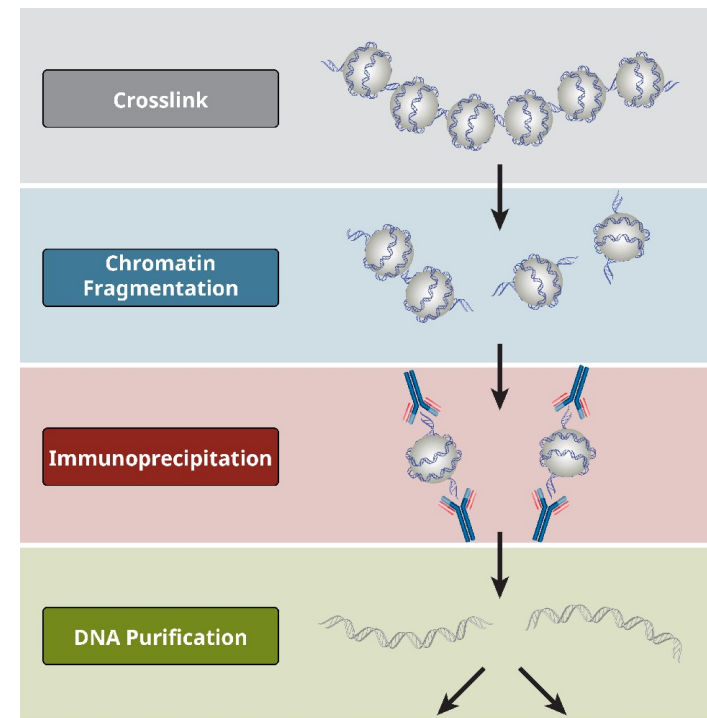
# Chromatin Immunoprecipitation (Steps 1&2)

## 1. Crosslink DNA to protein so stays stable OR Native Approach

- Crosslink using formaldehyde or UV light
- Use crosslinking when looking at TF
- Use native state when looking at histone modifications

## 2. Fragment

- Sonication will shear chromatin

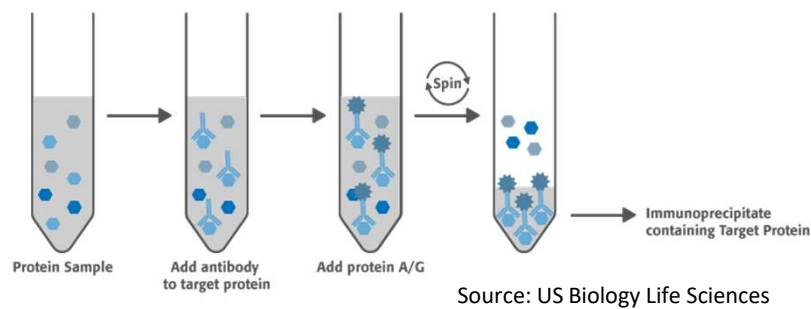


Source: Cell Signaling Technologies

# Chromatin Immunoprecipitation (Steps 3&4)

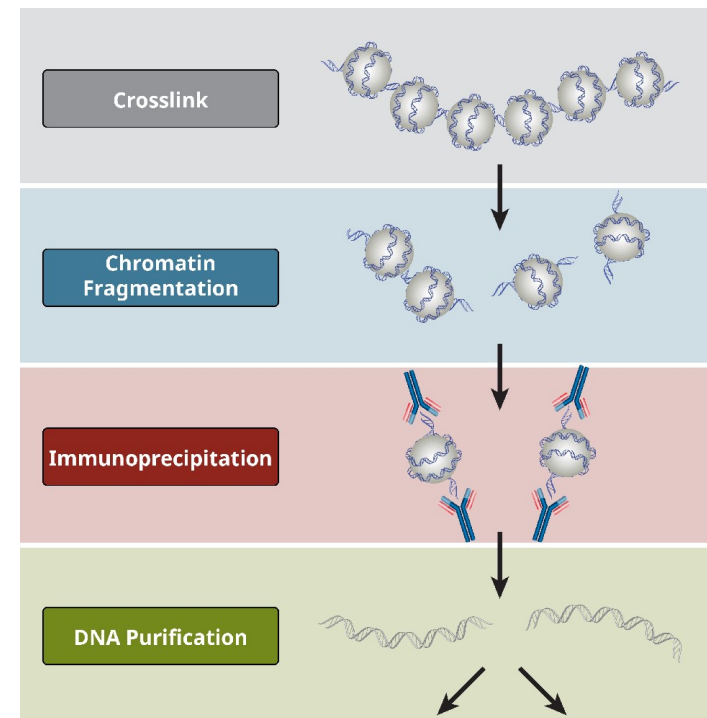
## 3. Immunoprecipitation

- Use an antibody to select protein of interest



## 4. Purify DNA

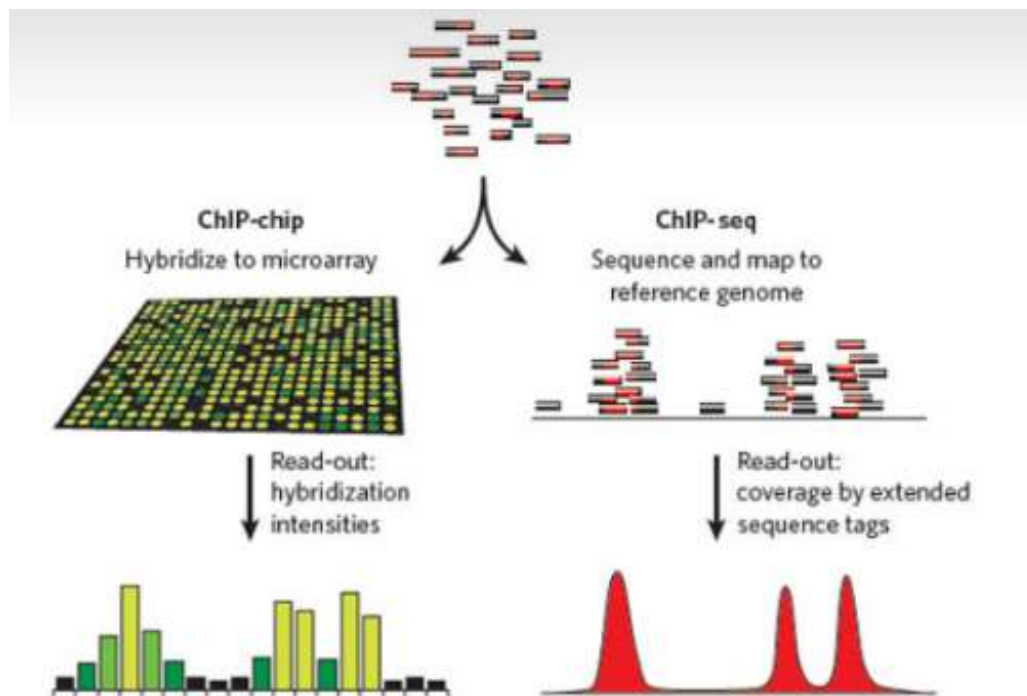
- Cross-link reversal
- Purify



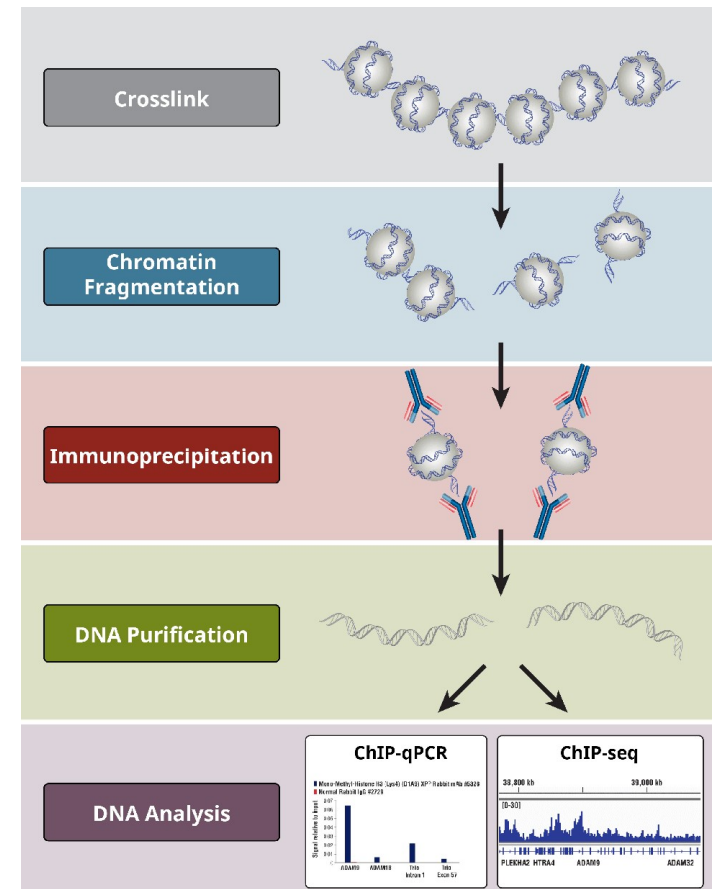
Source: Cell Signaling Technologies



# ChIP-Paired With Some DNA Technology



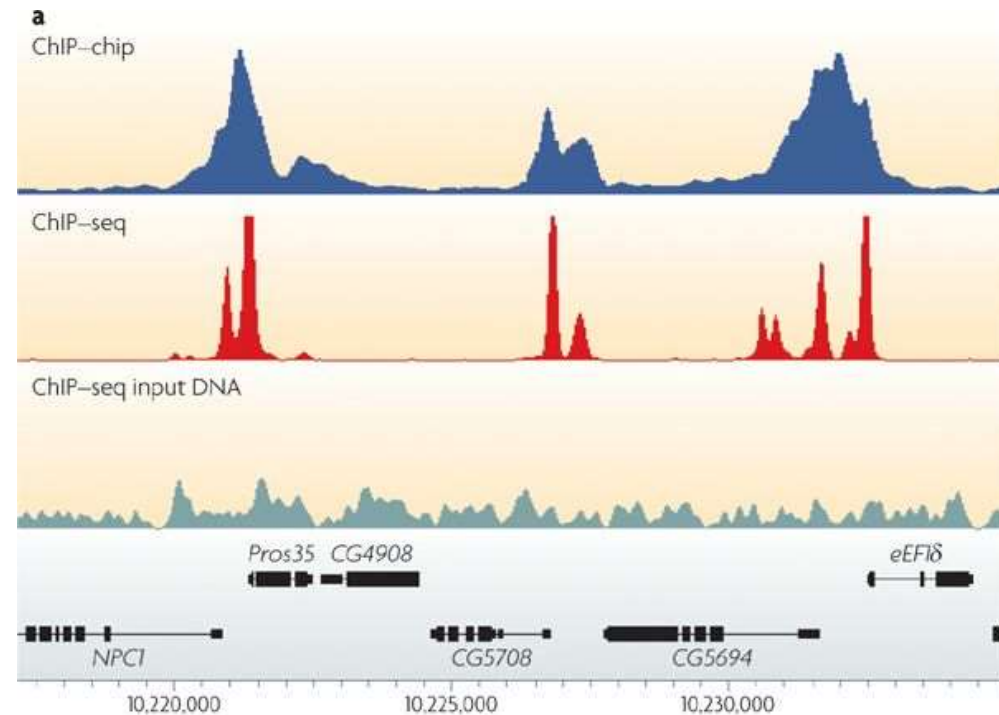
Source: Katerina Kechris



Source: Cell Signaling Technologies

# Negative Controls

- Sensitivity and specificity of antibody
- Found large artifacts
  - Find peaks in control
- Biases:
  - Shearing of DNA
  - GC content
  - Regions of open chromatin
- Compare your true ChIP-Seq to a background
- Control sample “input” or “IgG”
  - Sonicated chromatin without immunoprecipitation
  - IgG: “unspecific” IP

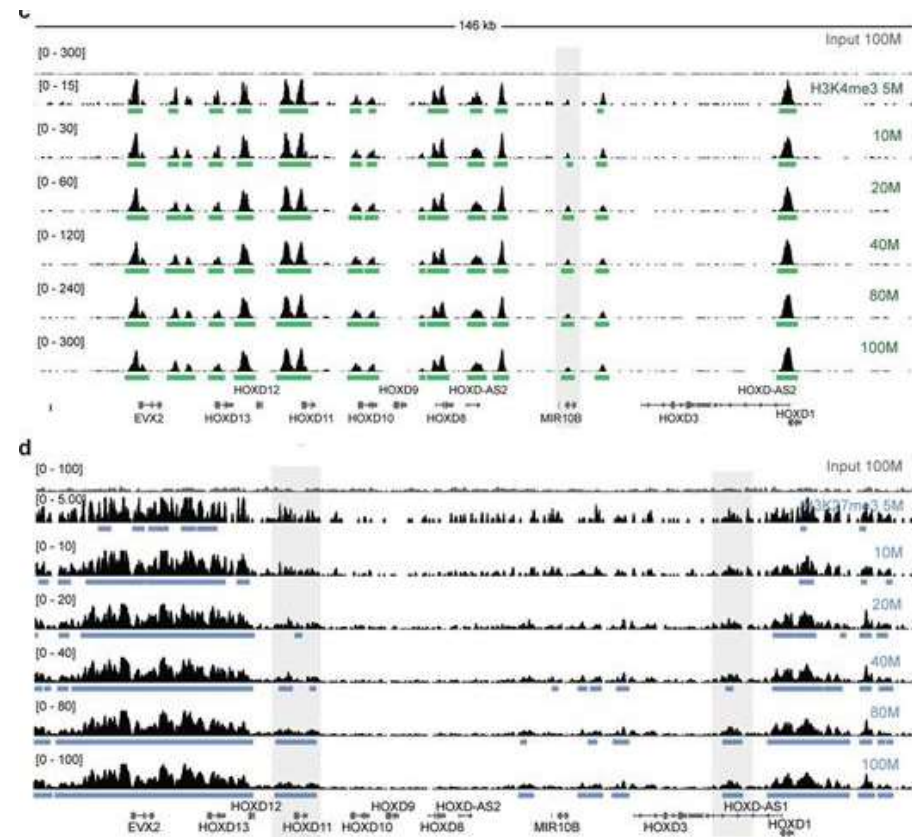


Source: Parker et al, 2009



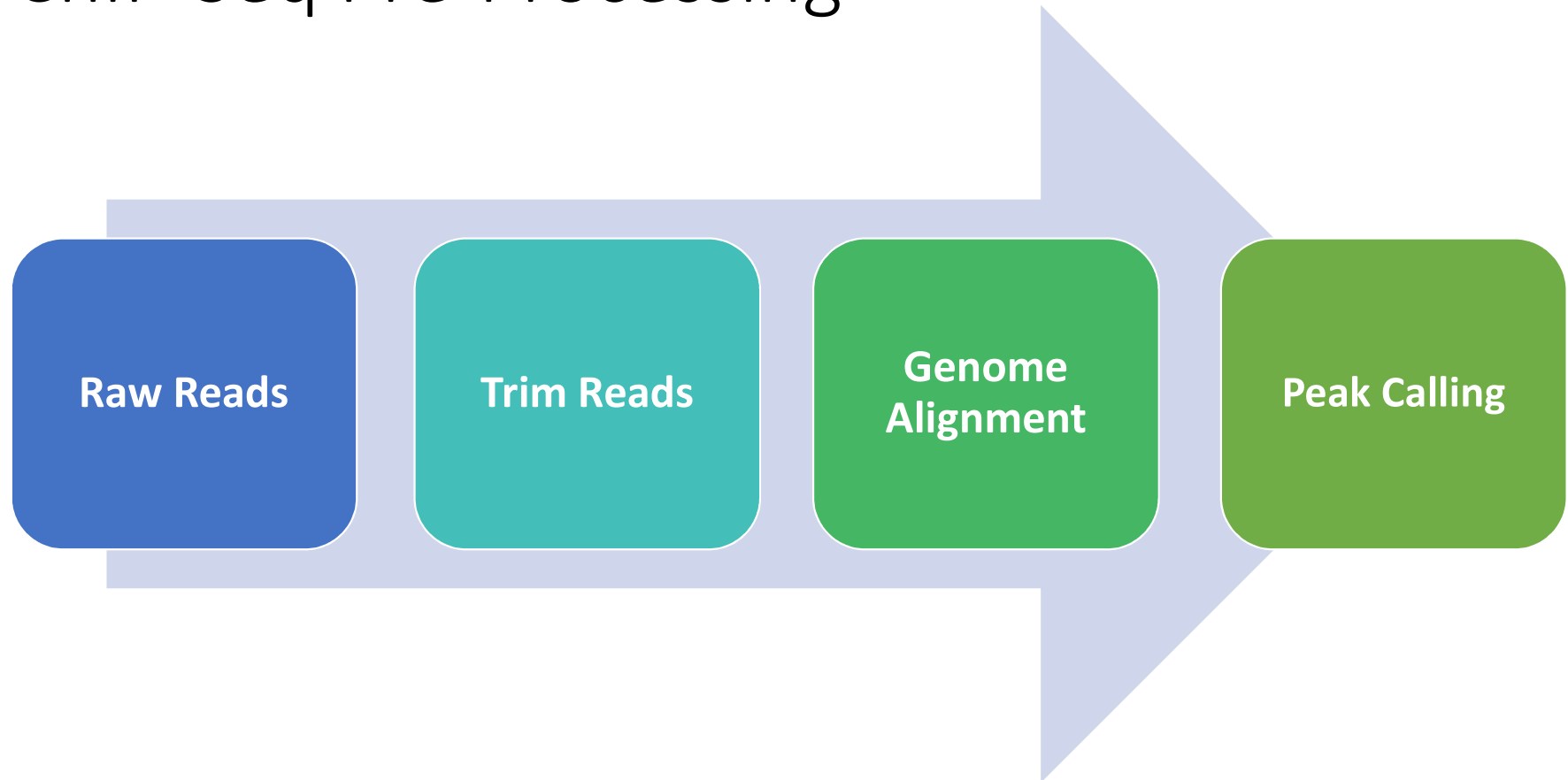
# Sequencing & Experimental Design

- Platform
  - Illumina
- Fragment Length
  - 25-35 bp common
- Depth
  - 40-50 Million
- R/CSSP
  - Power analysis for sequencing depth in ChIP-Seq analysis
- Use biological replicates
  - N=3 for cell lines
  - Usually use same control

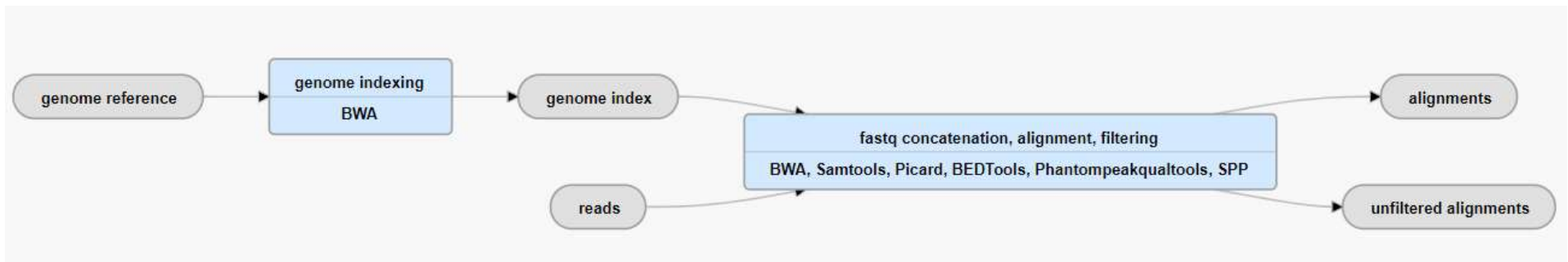


Source: Jung et al, 2014

# ChIP-Seq Pre-Processing



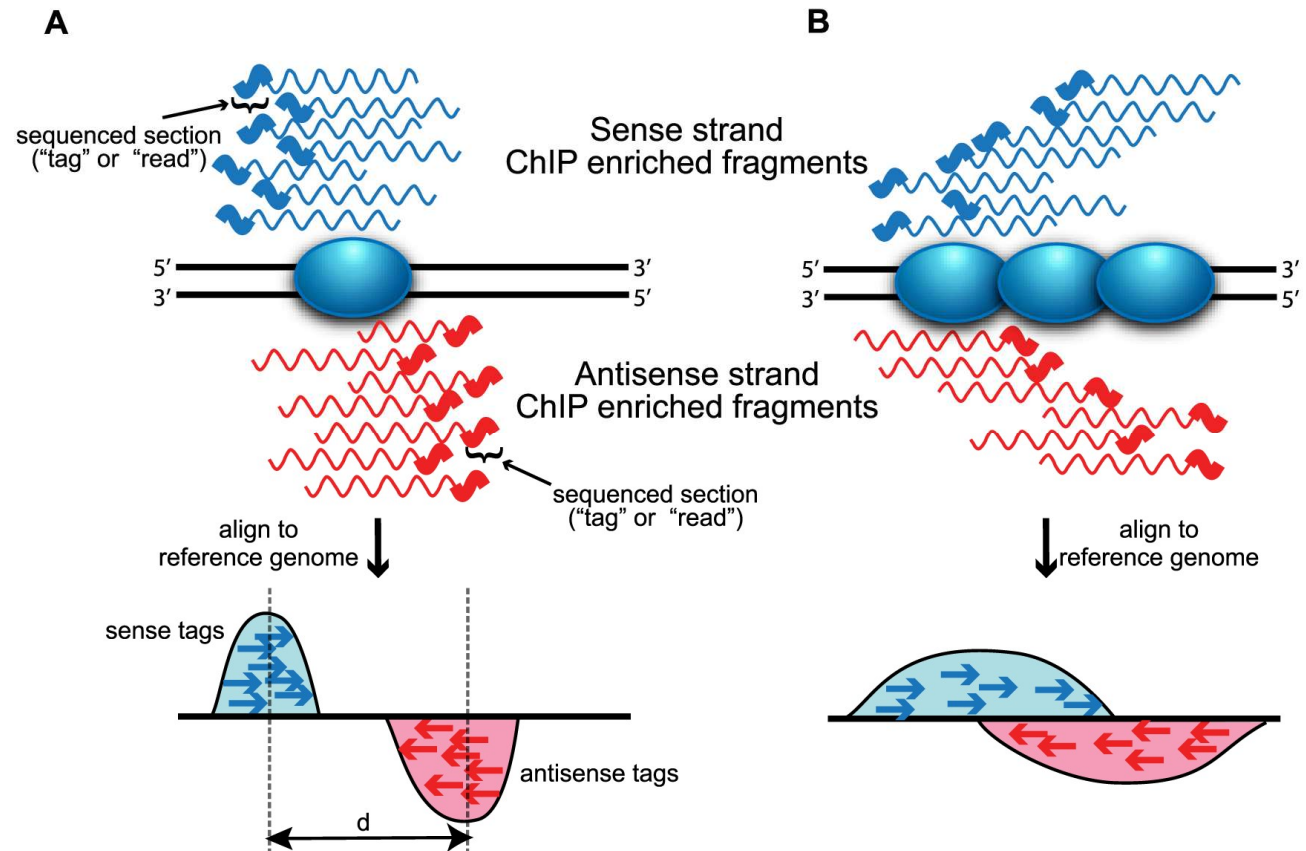
# ENCODE Pipeline



- Raw reads in FASTQ format
  - ENCODE recommends using BWA
    - Unspliced aligner (no gaps allowed)
    - Okay because we are dealing with DNA
  - Take the aligned bam file to a peak calling algorithm
- QC
    - Uniquely mapped reads, at most 2 mismatches
    - >50% total reads uniquely mappable
    - >50% reads non-redundant
      - Non-Redundant Fraction (NRF)
      - Need NRF > 0.8 for 10M reads
    - Wide distribution across genome, low GC bias, etc.

# Peak Calling

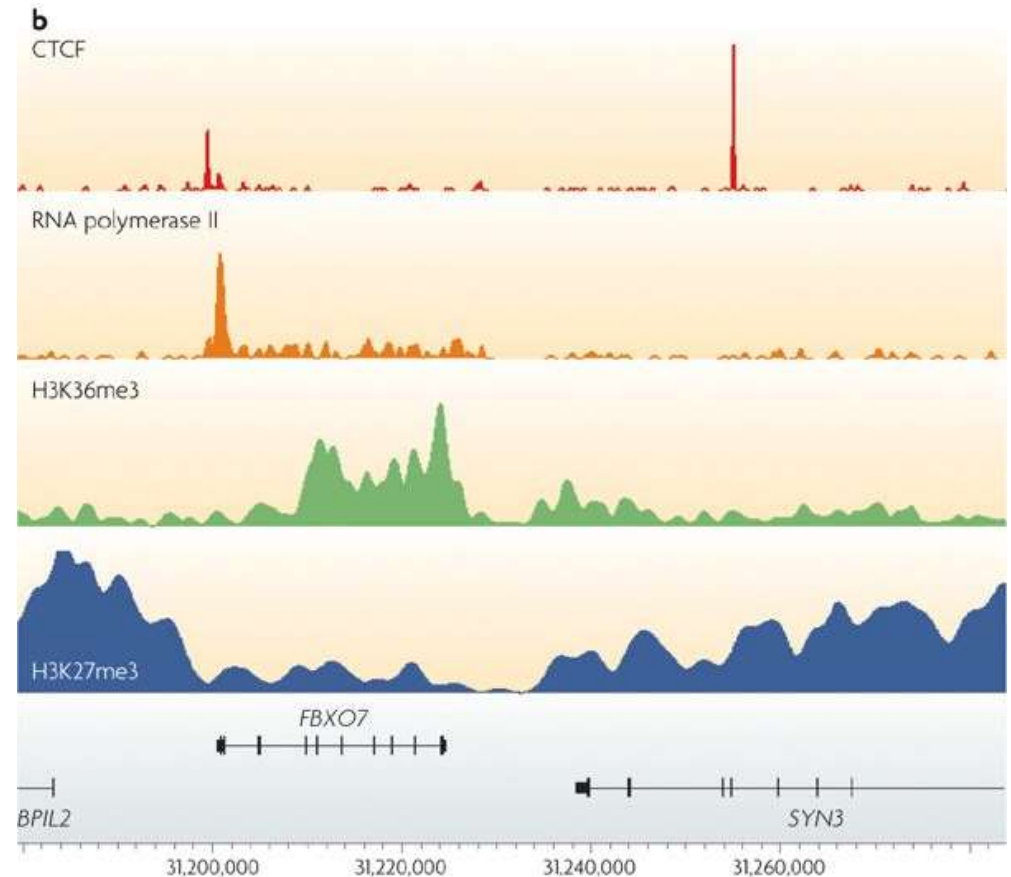
- What area of the genome do we say is interacting/binding with protein?
- Since we have DNA, we have 2 strands
  - Sense
  - Antisense
- A) TF B) Histone



Source: Wilbanks et al, 2010

# Histone vs TF Peaks

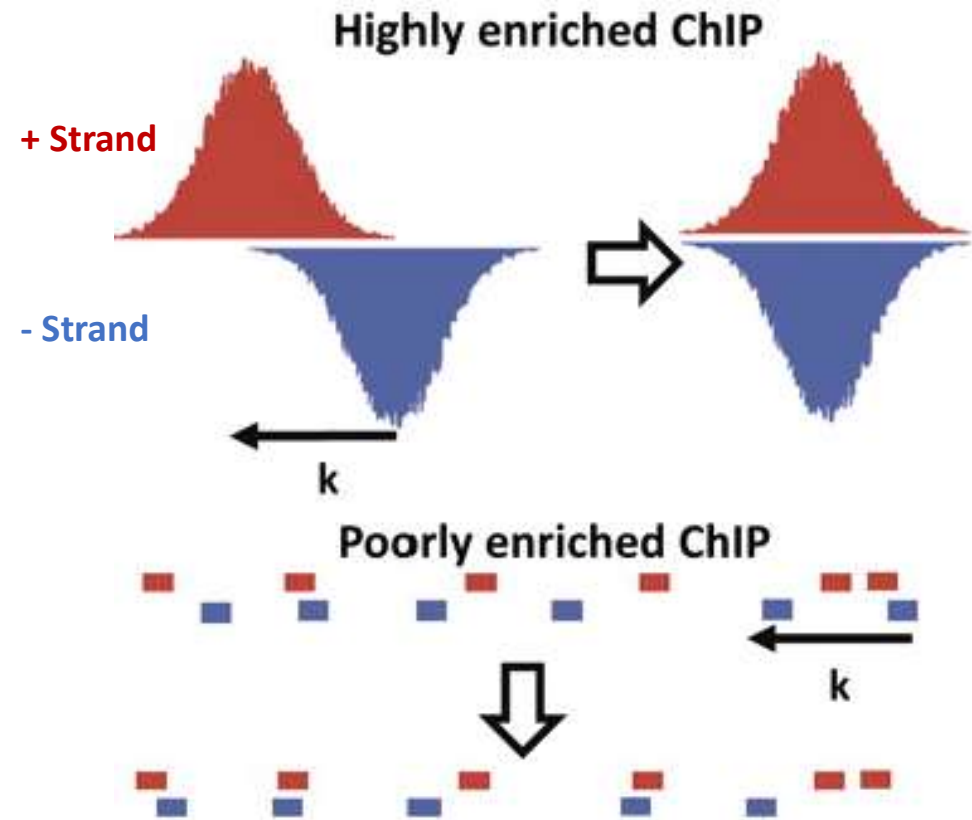
- Transcription factor peaks are sharper
- Transcription factor DNA binding sites are typically 10 nt long
- Histone peaks typically ~4x times length as TF peaks
- Use different peak calling algorithms based on type of ChIP you used



Source: Parker et al, 2009

# Read Shifting

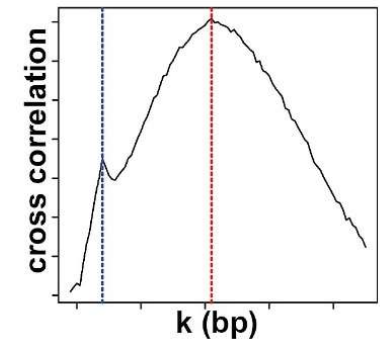
- Offset in forward or reverse strand reads
- Majority ChIP-Seq single-end sequencing
- Reads align to either the sense/antisense strands and the 3' or 5' extremes of the DNA fragments pulled down.
- The reads are shifted and the data from both strands combined to determine the most likely bases involved in protein binding.
- How big a "shift" is determined by the fragment size generated in the ChIP-seq library preparation
  - Estimate from sequence data
  - Determine empirically



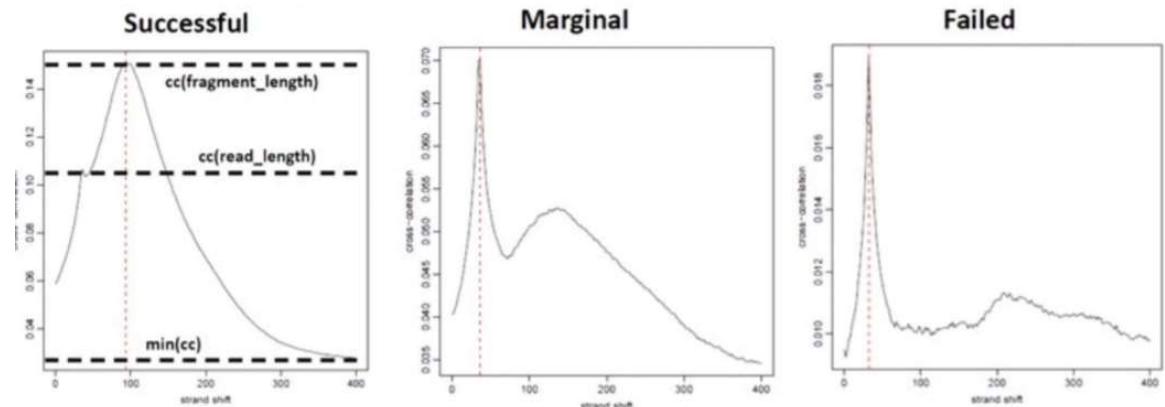
Source: Indiana University

# Cross-Correlation

- Pearson correlation between positive and negative strand profiles at different strand shift distances,  $k$ .
- Cross-correlation peaks at 2 places:
  - Read length (blue)
  - Average fragment length (red)



Source: Bailey et al, 2013



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

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

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

Source: Cornell University

# Normalization

- Want to compare ChIP to control/input
- Total read counts between ChIP and control/input not the same
- Unfortunately the control/input does not have a uniform distribution of reads across genome
  - GC content, CNV, mappability
- Issue is ChIP really consists of 2 distributions (IP fragments + background) where the control/input just consists of 1 distribution (background)
- Estimate background reads ratio in ChIP reads
  - Scale by ChIP/control normalization factor
  - $N_1$  and  $N_2$  total reads for ChIP and control
  - $\Pi_0 N_1$  background and  $(1 - \Pi_0) N_1$  enriched signal reads
  - $r$  is the normalization factor

$$r = \frac{\Pi_0 * N_1}{N_2}$$



# Peak Calling Programs

## Within R:

BayesPeak, PICS,  
MOSAIcs, iSeq,  
ChIPseqR, CSAR, ChIP-  
Seq, SPP, NarrowPeaks

## Command Line Tools:

CisGenome, **E-RANGE**,  
FindPeaks, F-Seq, GLITR,  
**MACS**, PeakSeq, QuEST,  
SICER, SiSSRs, spp, Useq,  
MUSIC, BCP

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific density	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		chromosome scale Poisson dist.
MACS	13	1.3.5		X			X			X		X		local Poisson dist.
QuEST	14	2.3			X		X			X**		X		chromosome 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 excludes putative duplicated regions, no treatment of deletions

Source: Parker et al, 2009

The options are endless!

# New Recommendations

- Poisson test to rank their candidate peaks are more powerful than those that use a Binomial test
- Best for TF:
  - BCP and MACS2
- Best for Histones:
  - BCP and MUSIC



Briefings in Bioinformatics, 18(3), 2017, 441–450

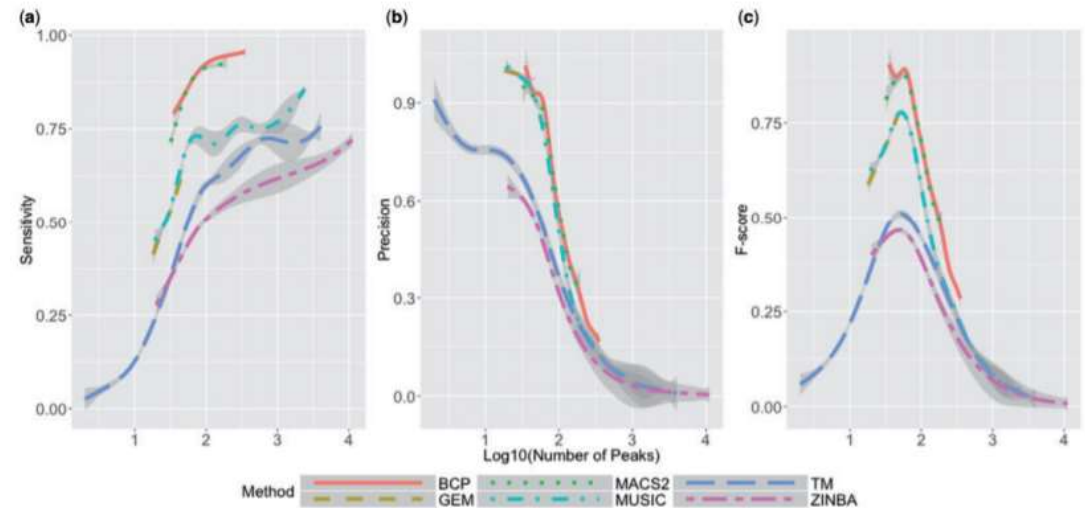
doi: 10.1093/bib/bbw035

Advance Access Publication Date: 11 May 2016

Paper

## Features that define the best ChIP-seq peak calling algorithms

Reuben Thomas, Sean Thomas, Alisha K. Holloway and Katherine S. Pollard



# MACS/MACS2

1. Removes redundant reads
  - Detects read length automatically
  - Filters duplicates
  - Calculates max # duplicate reads in single positions warranted by sequencing depth, removes excess of this number
2. Accounts for read-shifting for the offset in forward or reverse strand reads
  - Models distance between paired forward and reverse strand peaks
  - Slides window across genome to find enriched regions ( $M\text{-fold} > \text{background}$ )
  - Size bandwidth is 2x bandwidth parameter (can opt for a broad option for histones)
  - Expected background # reads  $\times$  length / mappable genome size
3. Peak Detection Phase
  - Extends reads in 3' direction to the fragment length from modeling
  - Scales samples linearly to same read number of control/input
  - Scans genome again window size 2x fragment length
  - P-value from dynamic Poisson distribution capture local biases and read background levels
  - Benjamini-Hochberg FDR correction

# MACS2 Code

Filter duplicates for each bam file

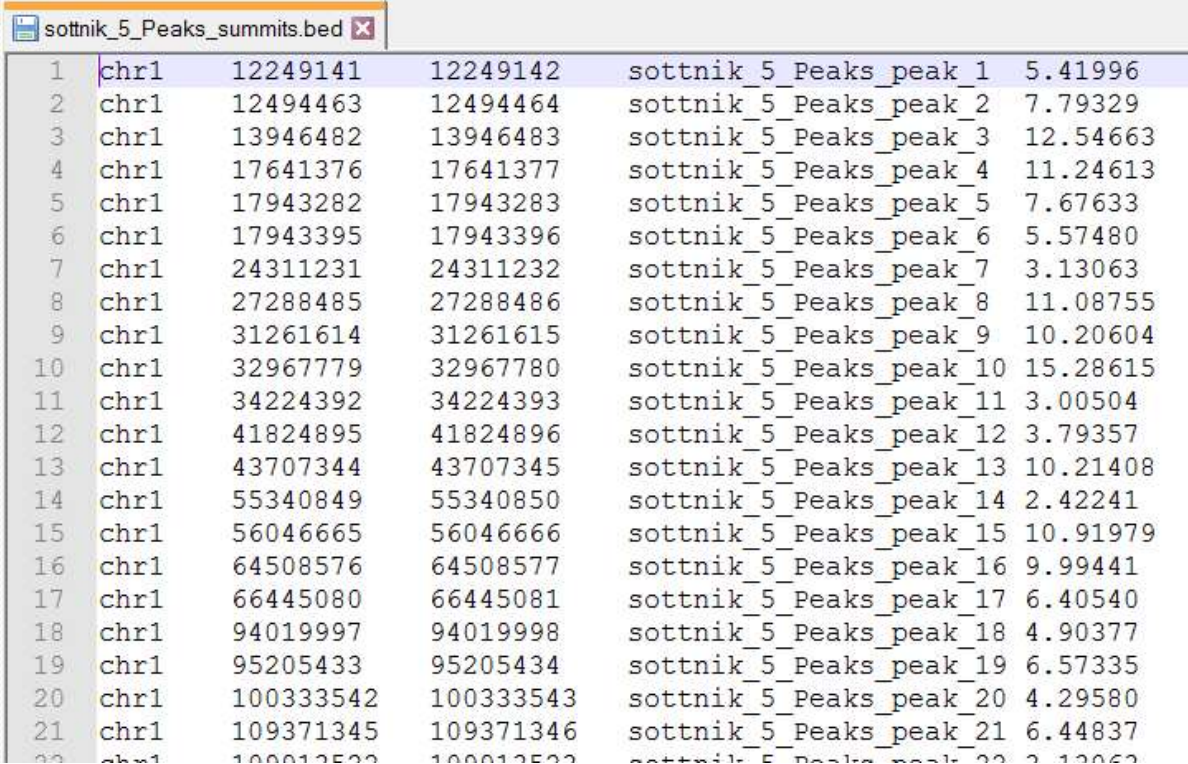
-i        input file  
-f        format of input file  
-g        genome you aligned to  
--keep-dup        number of duplicates to keep  
-o        output file

```
# Filter duplicates  
# Write down the count of reads for each sample after duplicate removal  
  
macs2 filterdup \  
-i "~/Course_Materials/ChIPSeq/Preprocessed/Alignment_BWA/tp53_r2.fastq_trimmed.fastq_sorted.bam" \  
-f BAM -g hs --keep-dup=1 --verbose=3 -o "tp53_r2.fastq_trimmed.fastq_sorted_filterdup.bed"  
  
macs2 filterdup \  
-i "~/Course_Materials/ChIPSeq/Preprocessed/Alignment_BWA/TAp73beta_r2.fastq_trimmed.fastq_sorted.bam" \  
-f BAM -g hs --keep-dup=1 --verbose=3 -o "TAp73beta_r2.fastq_trimmed.fastq_sorted_filterdup.bed"  
  
macs2 filterdup \  
-i "~/Course_Materials/ChIPSeq/Preprocessed/Alignment_BWA/input.fastq_trimmed.fastq_sorted.bam" \  
-f BAM -g hs --keep-dup=1 --verbose=3 -o "input.fastq_trimmed.fastq_sorted_filterdup.bed"
```

# BED File Format

- Tab delimited file
- No Column Headers
- First 3 columns required

Column 1	Chromosome
Column 2	Start (bp)
Column 3	End (bp)
Column 4	Name
Column 5	Score
Column 6	Strand



1	chr1	12249141	12249142	sottnik_5_Peaks_peak_1	5.41996
2	chr1	12494463	12494464	sottnik_5_Peaks_peak_2	7.79329
3	chr1	13946482	13946483	sottnik_5_Peaks_peak_3	12.54663
4	chr1	17641376	17641377	sottnik_5_Peaks_peak_4	11.24613
5	chr1	17943282	17943283	sottnik_5_Peaks_peak_5	7.67633
6	chr1	17943395	17943396	sottnik_5_Peaks_peak_6	5.57480
7	chr1	24311231	24311232	sottnik_5_Peaks_peak_7	3.13063
8	chr1	27288485	27288486	sottnik_5_Peaks_peak_8	11.08755
9	chr1	31261614	31261615	sottnik_5_Peaks_peak_9	10.20604
10	chr1	32967779	32967780	sottnik_5_Peaks_peak_10	15.28615
11	chr1	34224392	34224393	sottnik_5_Peaks_peak_11	3.00504
12	chr1	41824895	41824896	sottnik_5_Peaks_peak_12	3.79357
13	chr1	43707344	43707345	sottnik_5_Peaks_peak_13	10.21408
14	chr1	55340849	55340850	sottnik_5_Peaks_peak_14	2.42241
15	chr1	56046665	56046666	sottnik_5_Peaks_peak_15	10.91979
16	chr1	64508576	64508577	sottnik_5_Peaks_peak_16	9.99441
17	chr1	66445080	66445081	sottnik_5_Peaks_peak_17	6.40540
18	chr1	94019997	94019998	sottnik_5_Peaks_peak_18	4.90377
19	chr1	95205433	95205434	sottnik_5_Peaks_peak_19	6.57335
20	chr1	100333542	100333543	sottnik_5_Peaks_peak_20	4.29580
21	chr1	109371345	109371346	sottnik_5_Peaks_peak_21	6.44837



# MACS2 Code

- Predict Fragment Length

-i	input file
-g	genome you aligned to
-m	M-FOLD Range to look at (default is 5 – 50)
-bw	bandwidth of region used to compute fragment size (default 300)

```
# predict fragment length  
# write down the fragment length for each sample  
  
macs2 predictd -i tp53_r2.fastq_trimmed.fastq_sorted_filterdup.bed -g hs -m 5 20  
macs2 predictd -i TAp73beta_r2.fastq_trimmed.fastq_sorted_filterdup.bed -g hs -m 5 20  
macs2 predictd -i input.fastq_trimmed.fastq_sorted_filterdup.bed -g hs -m 5 20
```

# MACS2 Code

- t       ChIP file
- c       control/input file
- n       names to generate output
- broad   look for broad peaks (histones)
- bdg     generate bedgraph files

```
# MACS2 callpeak options
```

```
macs2 callpeak -h
```

```
# -t sample -c control -g effective genome size needs to be empirically computed using  
# a hg38.fa genome file for  
# hg38 but for this practical use 'hs' which is = 2.6e9, the value for hg19  
# -f filetype --bdg generate bedgraph
```

```
macs2 callpeak \
```

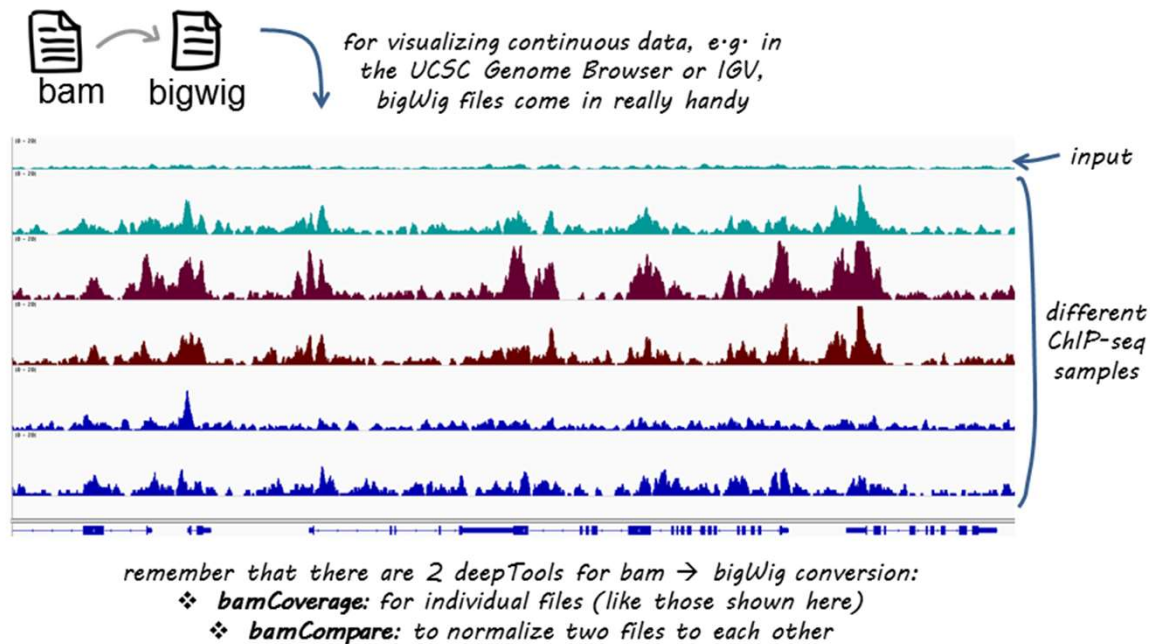
```
-t "-/Course_Materials/ChIPSeq/Preprocessed/Alignment_BWA/tp53_r2.fastq_trimmed.fastq_sorted.bam" \  
-c "-/Course_Materials/ChIPSeq/Preprocessed/Alignment_BWA/input.fastq_trimmed.fastq_sorted.bam" \  
-g hs -n tp53_r2.fastq_trimmed.fastq_sorted_standard -f BAM --keep-dup auto --bdg
```

```
macs2 callpeak \
```

```
-t "-/Course_Materials/ChIPSeq/Preprocessed/Alignment_BWA/TAp73beta_r2.fastq_trimmed.fastq_sorted.bam" \  
-c "-/Course_Materials/ChIPSeq/Preprocessed/Alignment_BWA/input.fastq_trimmed.fastq_sorted.bam" \  
-g hs -n TAp73beta_r2.fastq_trimmed.fastq_sorted_standard -f BAM --keep-dup auto --bdg
```

# Bedgraph & BigWig Files

- For visualization specifically in the genome browser



deeptools  
samtools

Provide good  
options for  
going back and  
forth between  
file formats



# MAC2 Output

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	#	This	file	is	generated	by	MACS	version	2.1.0.20150731											
2	#	Command	line:	callpeak	-t	sottnik_5.markdup.Sample.bed	-c	controlSample.bed	-f	BED	-g	hs	-n	sottnik_5_Peaks	--outdir	/data/home/vanderll/Sottnik/data_processed/v1.alreadyAligned/sottnik_5_peaks	-B	--SPMR		
3	#	ARGUMENTS	LIST:																	
4	#	name	=	sottnik_5_Peaks																
5	#	format	=	BED																
6	#	ChIP-seq	file	=	['sottnik_5.markdup.Sample.bed']															
7	#	control	file	=	['controlSample.bed']															
8	#	effective	genome	size	=	2.70e+09														
9	#	band	width	=	300															
10	#	model	fold	=	[5, 50]															
11	#	qvalue	cutoff	=	5.00e-02															
12	#	Larger	dataset	will	be	scaled	towards	smaller	dataset.											
13	#	Range	for	calculating	regional	lambda	is:	1000	bps	and	10000	bps								
14	#	Broad	region	calling	is	off														
15	#	MACS	will	save	fragment	pileup	signal	per	million	reads										
16																				
17	#	tag	size	is	determined	as	49	bps												
18	#	total	tags	in	treatment:	24339985														
19	#	tags	after	filtering	in	treatment:	24339985													
20	#	maximum	duplicate	tags	at	the	same	position	in	treatment	=	1								
21	#	Redundant	rate	in	treatment:	0.00														
22	#	total	tags	in	control:	24339980														
23	#	tags	after	filtering	in	control:	24339980													
24	#	maximum	duplicate	tags	at	the	same	position	in	control	=	1								
25	#	Redundant	rate	in	control:	0.00														
26	#	d	=	49																
27	#	alternative	fragment	length(s)	may	be	49	bps												

# MACS2 Output

Absolute summit

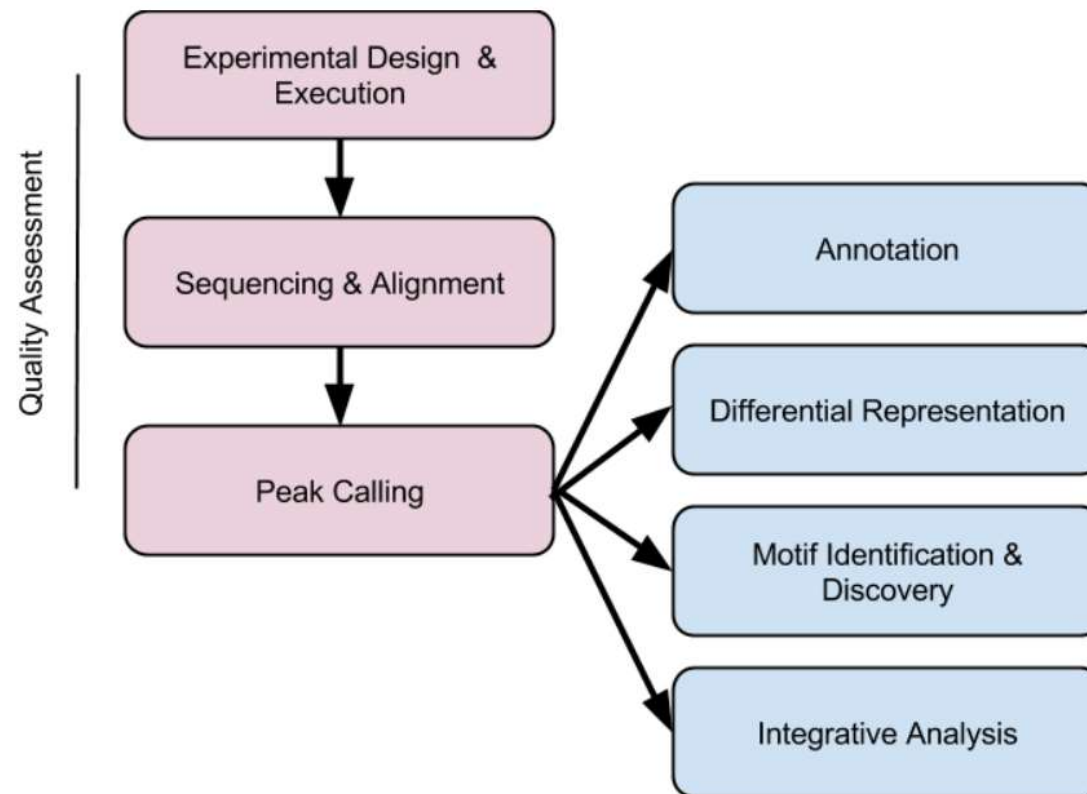
$-\log_{10}(\text{p-value})$

$-\log_{10}(\text{q-value})$

Fold Enrichment

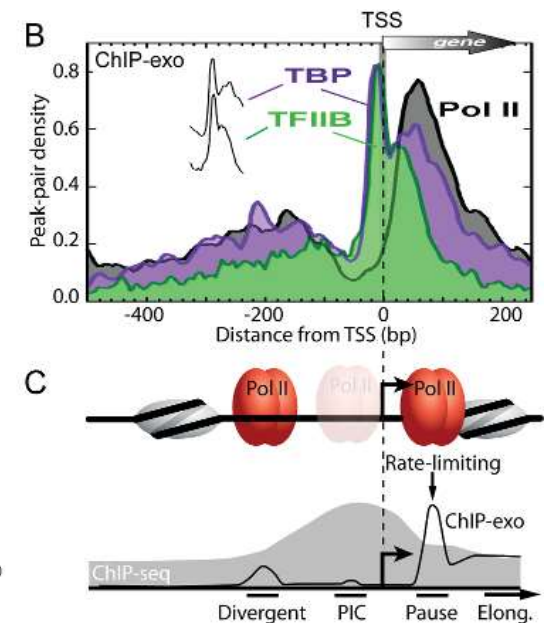
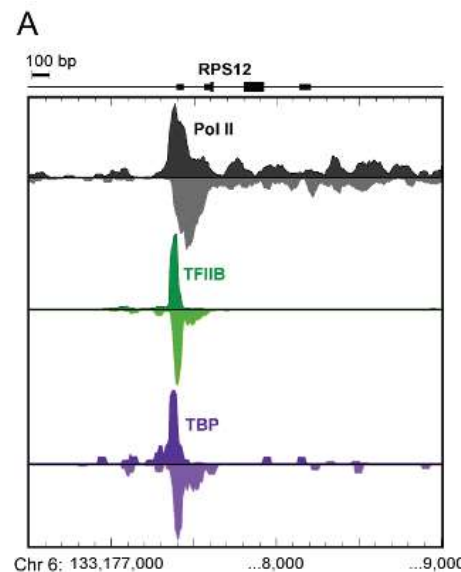
27	# alternative fragment length(s) may be 49 bps											
28	chr	start	end	length	abs_summit	pileup	$-\text{LOG}_{10}(\text{p-value})$	fold_enrichment	$-\text{LOG}_{10}(\text{q-value})$	name		
29	chr1	12249114	12249173	60	12249142	10	10.10805	7.53115	5.41996	sottnik_5_Peaks_peak_1		
30	chr1	12494400	12494509	110	12494464	15	12.64637	8.08081	7.79329	sottnik_5_Peaks_peak_2		
31	chr1	13946446	13946680	235	13946483	16	17.83532	11.04613	12.54663	sottnik_5_Peaks_peak_3		
32	chr1	17641249	17641549	301	17641377	27	16.40858	7	11.24613	sottnik_5_Peaks_peak_4		
33	chr1	17943202	17943304	103	17943283	15	12.52288	8	7.67633	sottnik_5_Peaks_peak_5		
34	chr1	17943361	17943415	55	17943396	13	10.30259	7.07071	5.5748	sottnik_5_Peaks_peak_6		
35	chr1	24311219	24311307	89	24311232	10	7.68112	6.00109	3.13063	sottnik_5_Peaks_peak_7		
36	chr1	27288454	27288541	88	27288486	17	16.24134	9.81997	11.08755	sottnik_5_Peaks_peak_8		
37	chr1	31261495	31261660	166	31261615	22	15.25735	7.66667	10.20604	sottnik_5_Peaks_peak_9		
38	chr1	32967664	32967826	163	32967780	18	20.85935	12.34568	15.28615	sottnik_5_Peaks_peak_10		
39	chr1	34224371	34224600	230	34224393	14	7.53305	5	3.00504	sottnik_5_Peaks_peak_11		
40	chr1	41824872	41824930	59	41824896	13	8.38157	5.78273	3.79357	sottnik_5_Peaks_peak_12		
41	chr1	43707163	43707459	297	43707345	15	15.26552	9.74481	10.21408	sottnik_5_Peaks_peak_13		

# Analyses with Peaks



# Peak Annotation

- Is my peak in an gene?
- What about promoter region?
- Enhancer region?
- Could just look at the nearest gene and calculate distance from gene
- R/ChIPseeker helps annotate peaks



Source: Perreault et al, 2016

# ChIPseeker Code

```
## loading packages
library(ChIPseeker)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
library(clusterProfiler)

##read in peak calls (bed file)
peak <- readPeakFile("/path/macs2_peaks.bed")
```

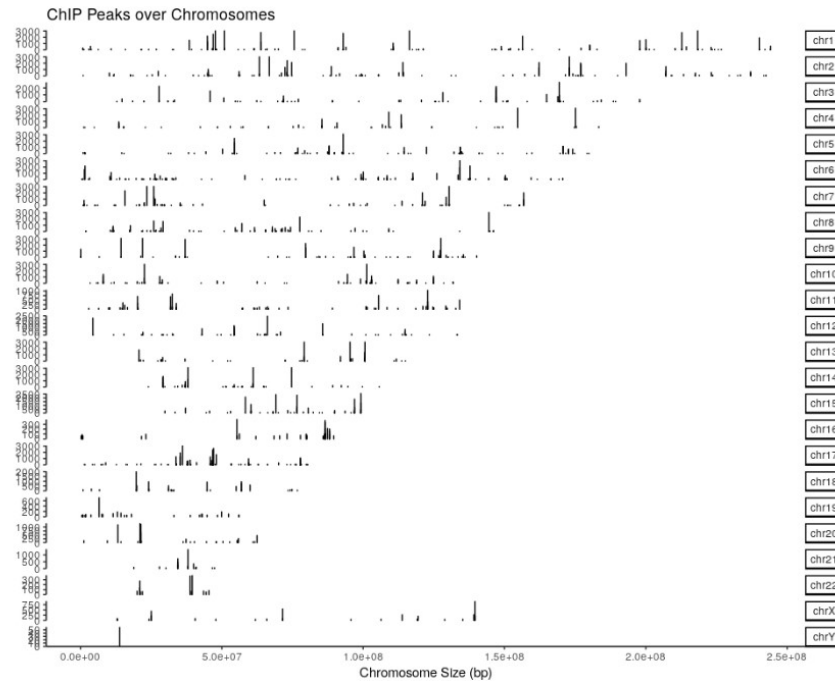
peak is a GRanges object,  
which is a type of S4 object

```
## GRanges object with 1331 ranges and 2 metadata columns:
##      seqnames      ranges strand |      V4      V5
##      <Rle>        <IRanges> <Rle> | <factor> <numeric>
## [1] chr1      815093-817883      * | MACS_peak_1  295.76
## [2] chr1     1243288-1244338      * | MACS_peak_2   63.19
## [3] chr1     2979977-2981228      * | MACS_peak_3  100.16
## [4] chr1     3566182-3567876      * | MACS_peak_4  558.89
## [5] chr1     3816546-3818111      * | MACS_peak_5   57.57
## ...      ...      ...      ... | ...      ...
## [1327] chrX 135244783-135245821      * | MACS_peak_1327  55.54
## [1328] chrX 139171964-139173506      * | MACS_peak_1328  270.19
## [1329] chrX 139583954-139586126      * | MACS_peak_1329  918.73
## [1330] chrX 139592002-139593238      * | MACS_peak_1330  210.88
## [1331] chrY  13845134-13845777      * | MACS_peak_1331   58.39
## -----
## seqinfo: 24 sequences from an unspecified genome; no seqlengths
```



# ChIPseeker Code – coverage plot

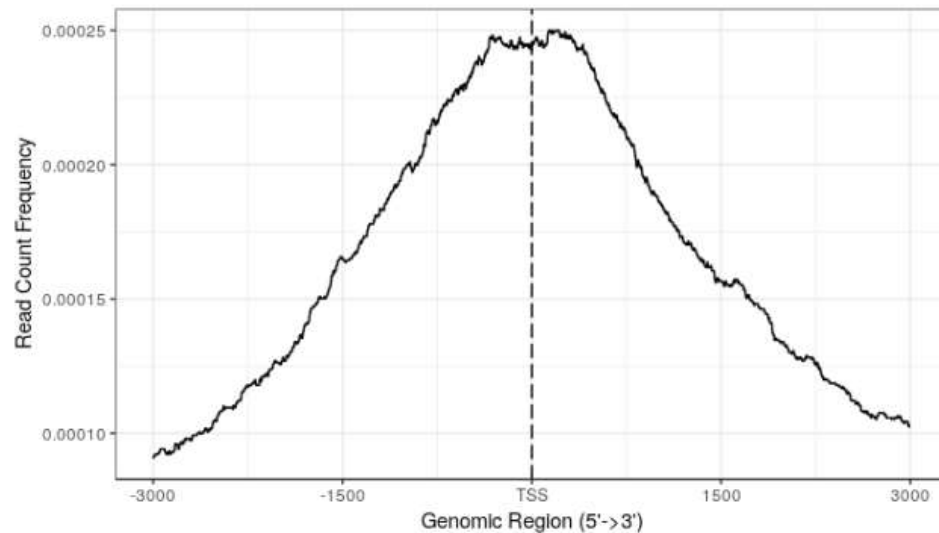
```
covplot(peak, weightCol="V5")
```



# ChIPseeker Code – TSS Distance Code

```
#find what you consider promoter regions
promoter <- getPromoters(Txdb=txdb, upstream=3000, downstream=3000)
##find distance peaks are to defined promoter regions
tagMatrix <- getTagMatrix(peak, windows=promoter)
```

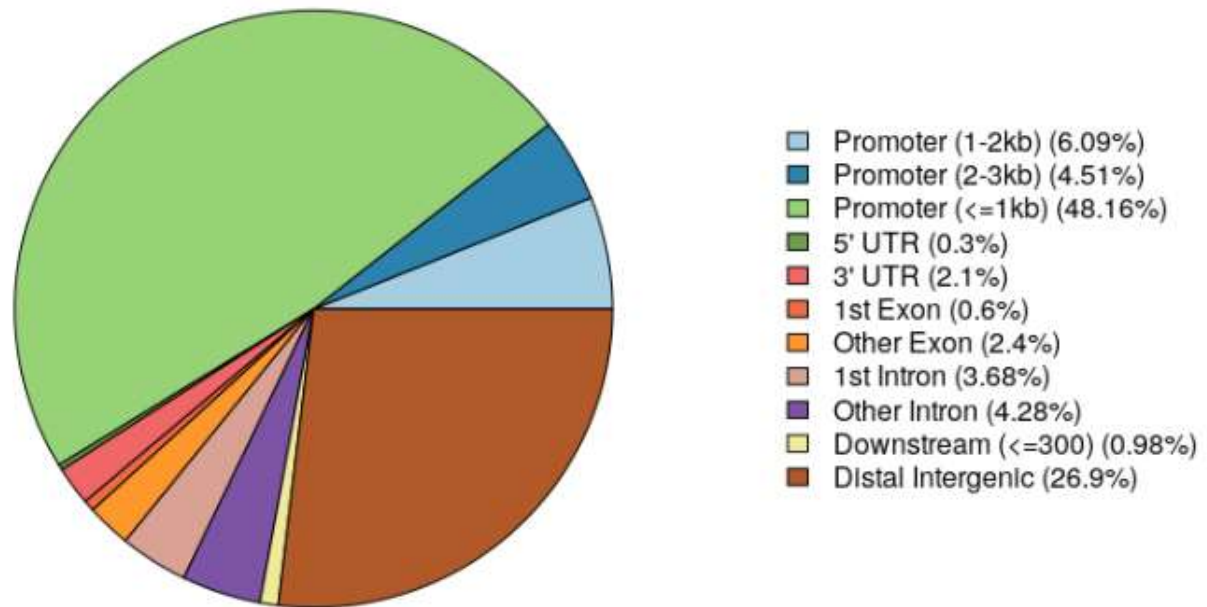
```
plotAvgProf(tagMatrix, xlim=c(-3000, 3000),
            xlab="Genomic Region (5'→3')", ylab = "Read Count Frequency")
```



# ChIPseeker Code – Peak Annotation

```
##get peak annotation  
peakAnno <- annotatePeak("/path/macs2_peaks.bed", tssRegion=c(-3000, 3000),  
                        TxDb=txdb, annoDb="org.Hs.eg.db")
```

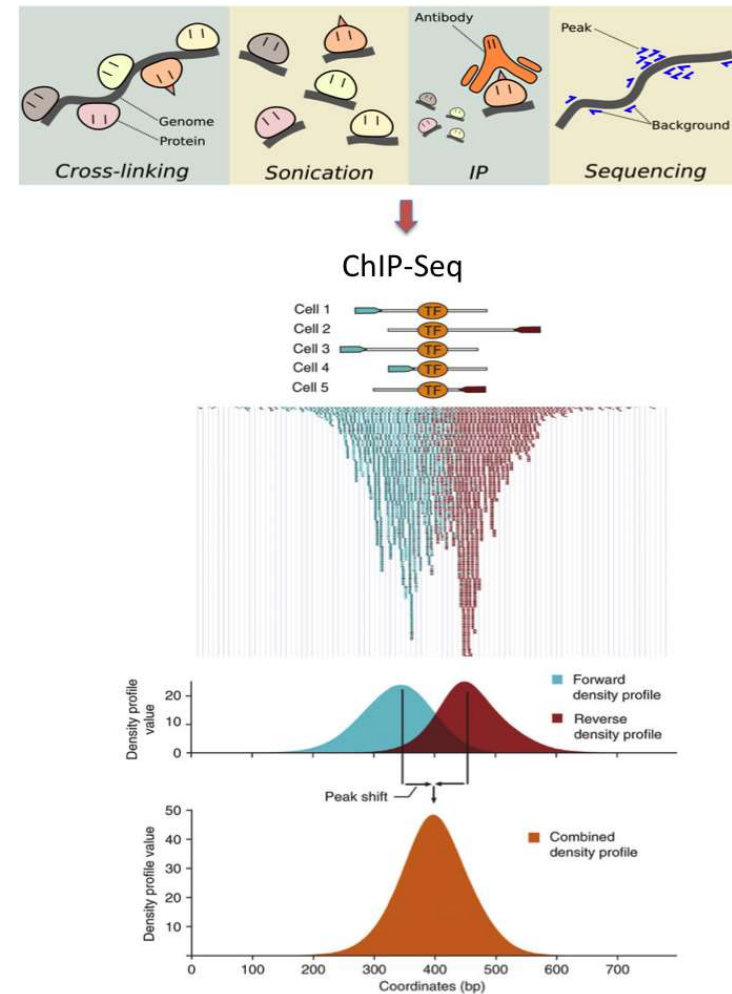
```
plotAnnoPie(peakAnno)
```





# Summary

- Use chromatin immunoprecipitation to extract DNA fragments which bind to protein of interest
- Make peak calls
- Annotate peak calls



# References

- Dressler GR, *Epigenetics, Development, and the Kidney*. JASN November **2008**, 19 (11) 2060-2067
- Shen C, Ding Y, Tang J, Song J, Guo F. *Identification of DNA-protein Binding Sites through Multi-Scale Local Average Blocks on Sequence Information*. Molecules. **2017** Nov 28;22(12)
- Wilbanks EG, Facciotti MT. *Evaluation of algorithm performance in ChIP-seq peak detection*. PLoS One **2010**, vol. 5 pg. e11471
- Park PJ. *ChIP-seq: advantages and challenges of a maturing technology*. Nature Reviews Genetics **2009** volume 10, pages 669–680.
- Jung, Y. L. *et al.* Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res.* **42**, e74 (**2014**).
- T. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, J. Zhang. *Practical guidelines for the comprehensive analysis of ChIP-seq data*. PLoS Comput. Biol., 9 (**2013**), p. e1003326

# References Continued

- Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei C, Lin F, Sung W: A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* **2010**, 26(9):1199–1204.
- Perreault AA, Venters BJ. *The ChIP-exo Method: Identifying Protein-DNA Interactions with Near Base Pair Precision*. J Vis Exp. **2016** Dec 23;(118). doi: 10.3791/55016.
- <https://www.youtube.com/watch?v=nkWGmaYRues>
- [https://biohpc.cornell.edu/lab/doc/Chip-seq\\_workshop\\_lecture1.pdf](https://biohpc.cornell.edu/lab/doc/Chip-seq_workshop_lecture1.pdf)
- [https://bioinformatics-core-shared-training.github.io/cruk-autumn-school-2017/ChIP/Materials/Practicals/Prctical4\\_PeakCalling\\_SS.pdf](https://bioinformatics-core-shared-training.github.io/cruk-autumn-school-2017/ChIP/Materials/Practicals/Prctical4_PeakCalling_SS.pdf)