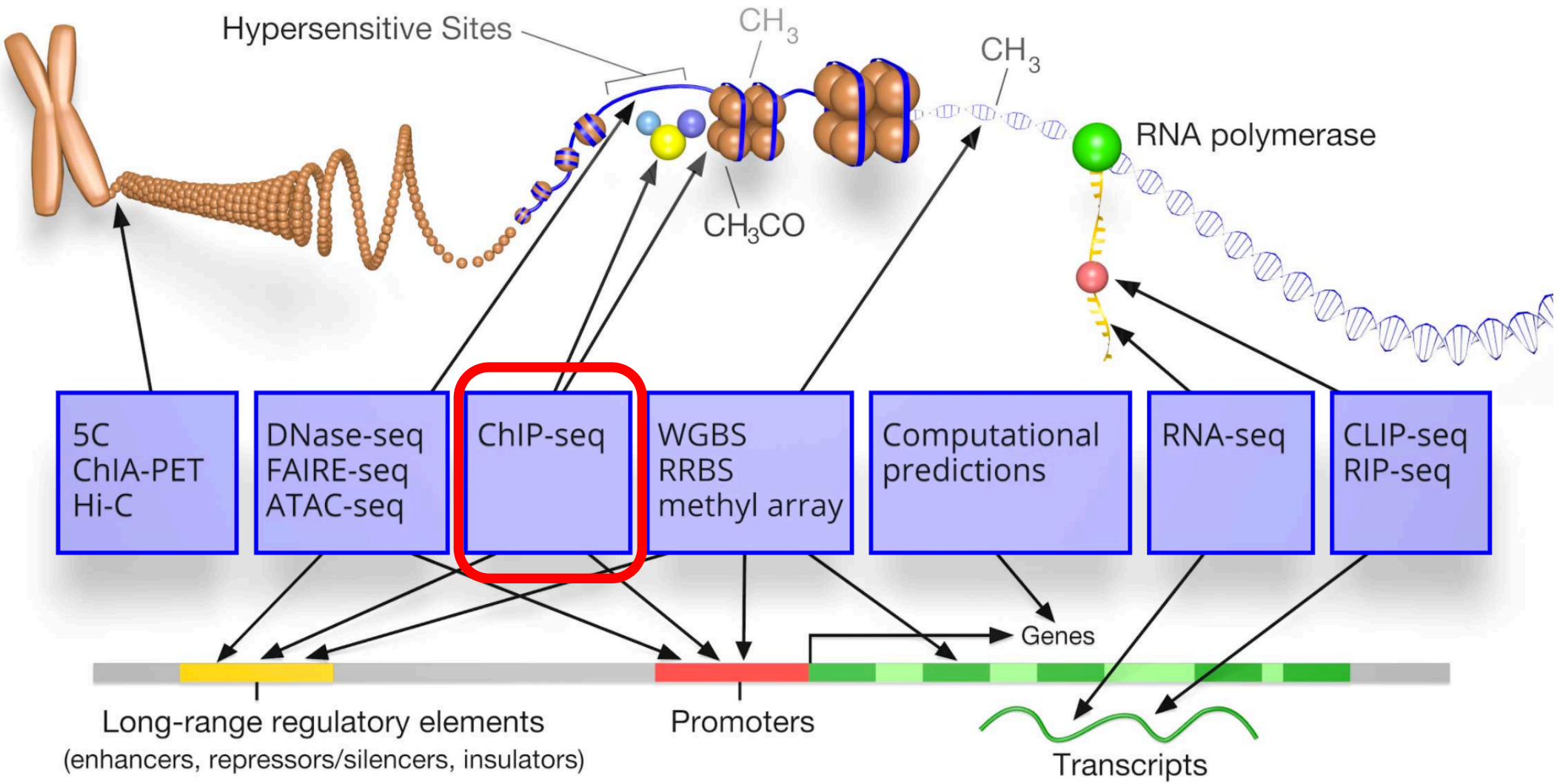
Best Practices in NGS Data Analysis

Introduction to ChIP-seq

(Chromatin Immuno-Precipitation followed by sequencing)

Bong-Hyun Kim
CCR Collaborative Bioinformatics Resource

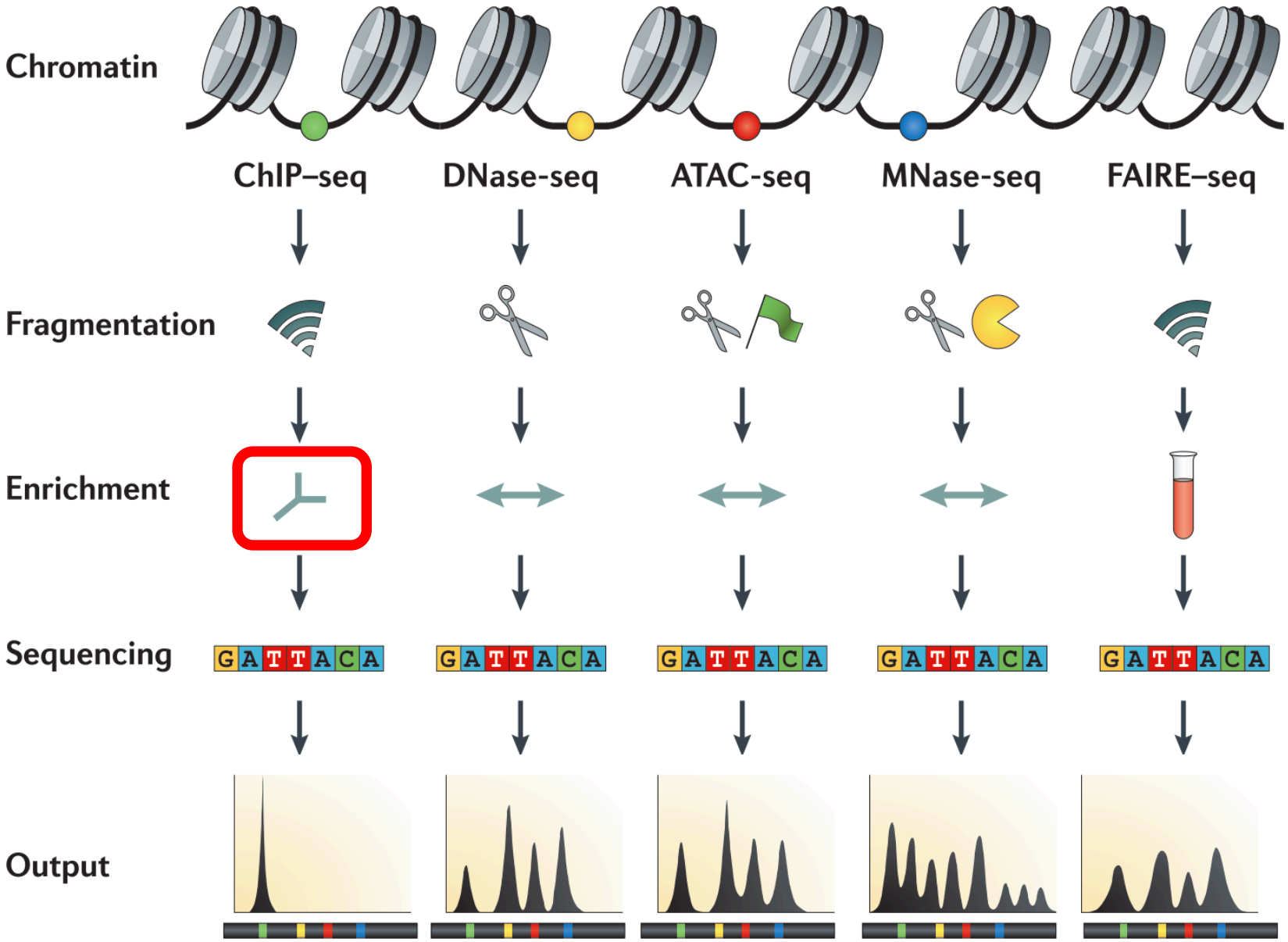
There are many NGS methods to study our (epi)genome.



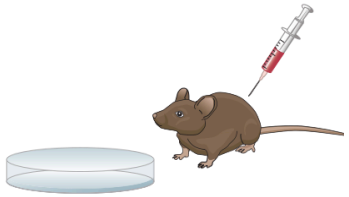
ENCODE

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

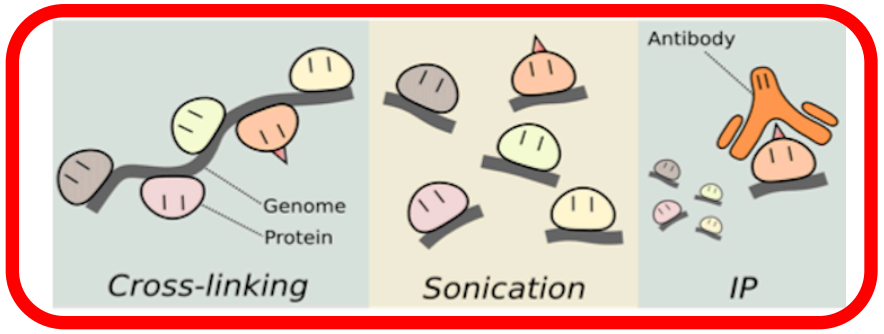
ChIP-seq is different from other chromatin sequencing.



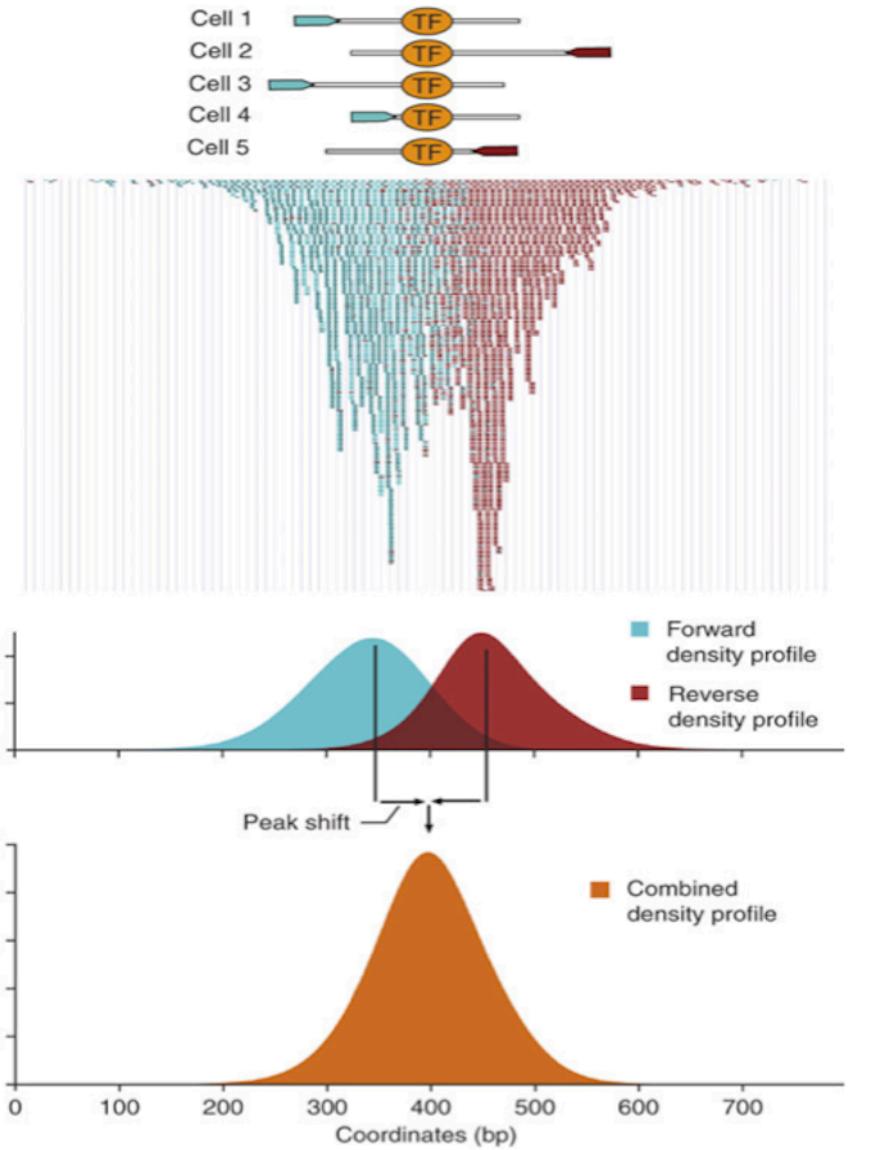
ChIP-seq: Chromatin Immuno-Precipitation followed by sequencing



Sample prep



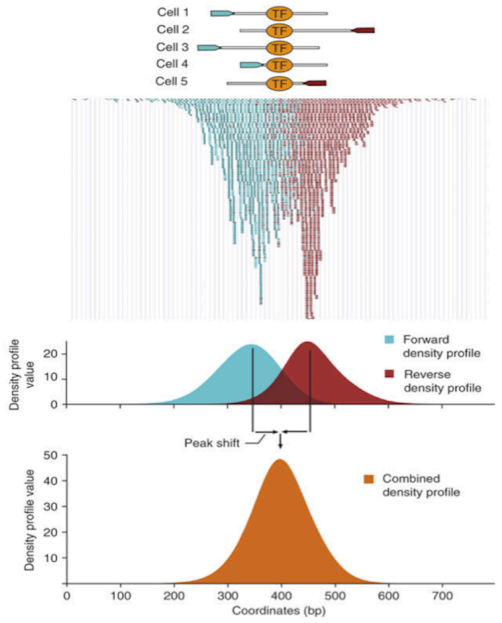
Sequencing

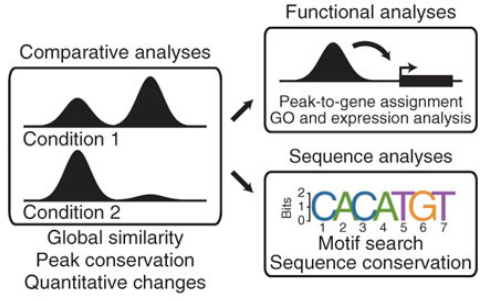


ChIP-seq usually answers the following questions.

- Does binding of a TF or histone distribution changes between conditions?
- Are the modified histones clustered in certain genomic areas?
- What is a binding motif of a transcription factor?
- Binding pattern around genes or TSS

ChIP-seq data analysis overview





Quality control
- Filter on QC score

Trim reads

Map tags
- Filter mapping
- Remove tags mapping to multiple locations
- Filter for unique tags

ChIP–seq example

Convert data to browser-viewable format
- GFF, WIG, BED file formats
- Check data quality by eye

Peak-finding
- Input DNA of similar tag count should be used as background
- MACS can also call SNPs

Known motif analysis
- CisGenome (can also call peaks)
- CEAS website
- The MEME Suite

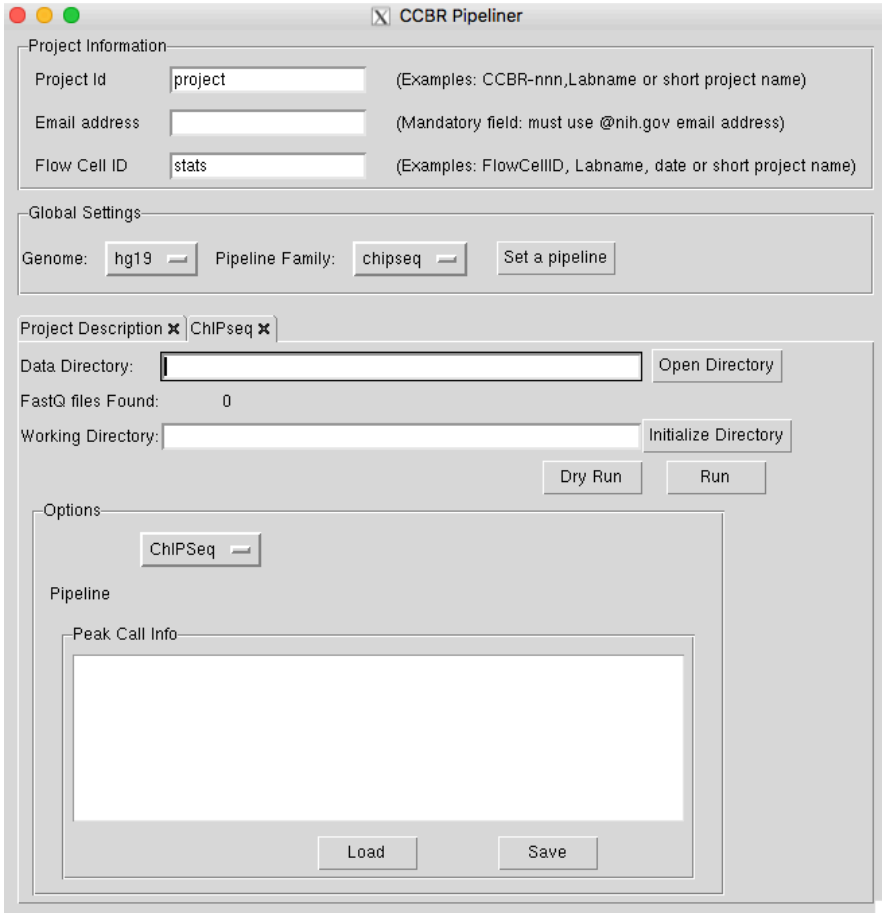
Intersect (UCSC Browser) or join (Galaxy)
- RefSeq TSSs can be used to define promoters (e.g. –2.5 kb/+500 bp)
- Select all promoters or filter based on gene expression
- Cross-reference known SNPs to predict regulatory SNPs

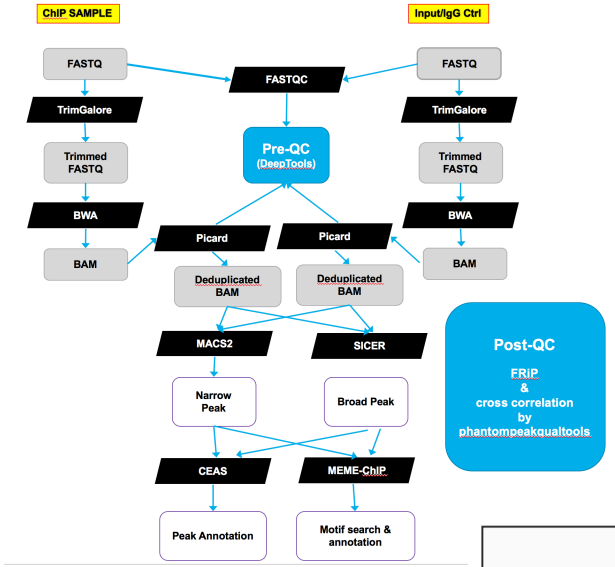
Gene Ontology analysis
- DAVID is an easy web tool

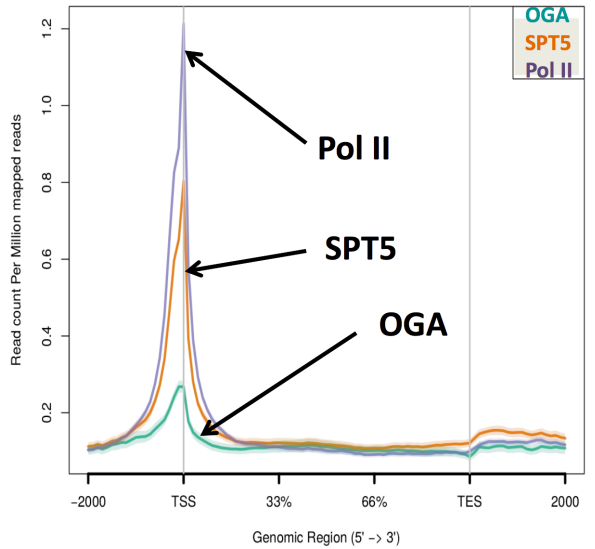
Nature Reviews | Genetics

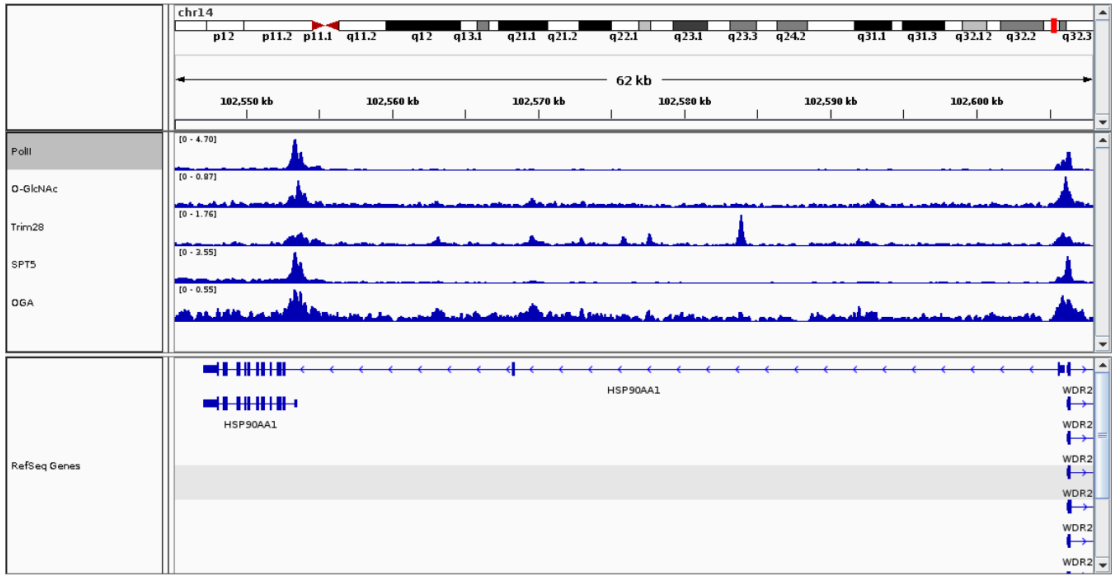
CCBR automated a ChIP-seq analysis pipeline

https://github.com/CCBR/Pipeliner









Stepping into the Regulome: ChIP-Seq/ENCODE Data Analysis (2-day)

BTEP Resources

- Archive
- BTEP Software
- Calendar
- Contacts
- Registration
- Schedule

External Resources

- CCR
- OSTR
- NCI-wide Additional Traning
- NIH Library Bioinformatics
- Other NIH Institutions

Probing DNA-Protein Interactions

ChIP-Seq/ENCODE Data Anaysis (2-day)

This 2-day course, which includes both lecture and hands-on components, will teach the basic concepts and practical aspects of ChIP-Seq data analysis. Learn everything from experimental design to statistical analysis and several downstream motif and pattern discovery methods using both commercial and open source software.

More details to be announced shortly.

NOTE: This is a BYOC (Bring your own laptop Computer) class. Government issued or personal computers are permitted. We will be able to supply a very limited set of computers, so if you want to take the class but cannot bring your own computer please indicate such in the Comment section on the registration form.

You will be able to register for this event on April 1st, 2017 at 12:15am

Date & Time: Monday, 17 April 2017 - 9:00am to Tuesday, 18 April 2017 - 4:00pm

Location: Bldg 10 FAES room 4 (B1C205)

Presenter: Multiple

Affiliation: To Be Announced

Format: Hands-on

Registration Start Date:
Saturday, 1 April 2017 - 12:15am

A successful ChIP-seq leads to quality science!

ARTICLES

nature neuroscience

Genome-wide identification and characterization of functional neuronal activity–dependent enhancers

LETTER

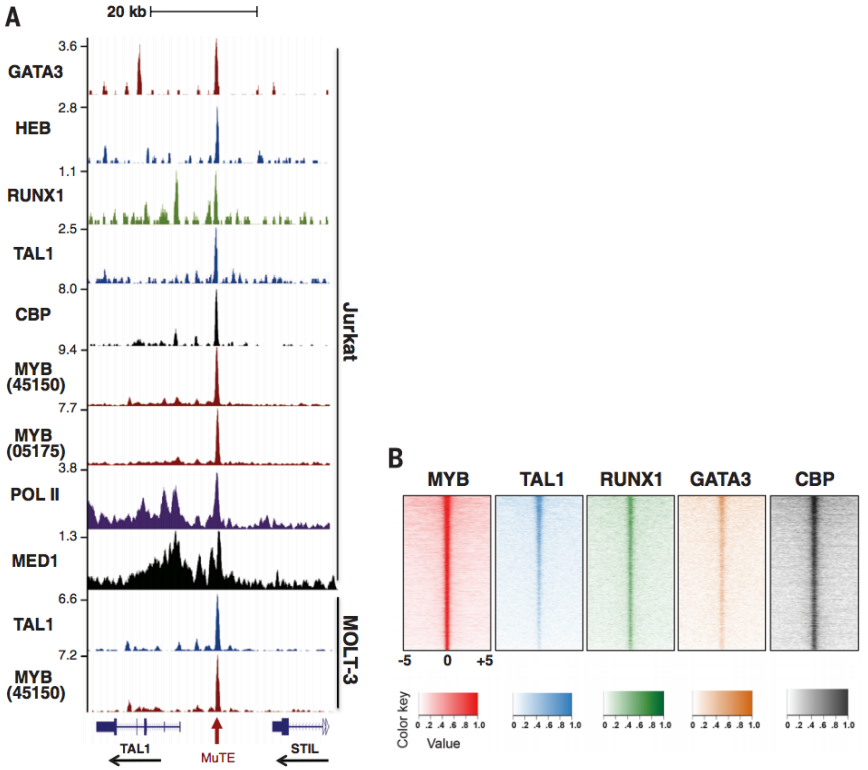
doi:10.1038/nature14289

Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice

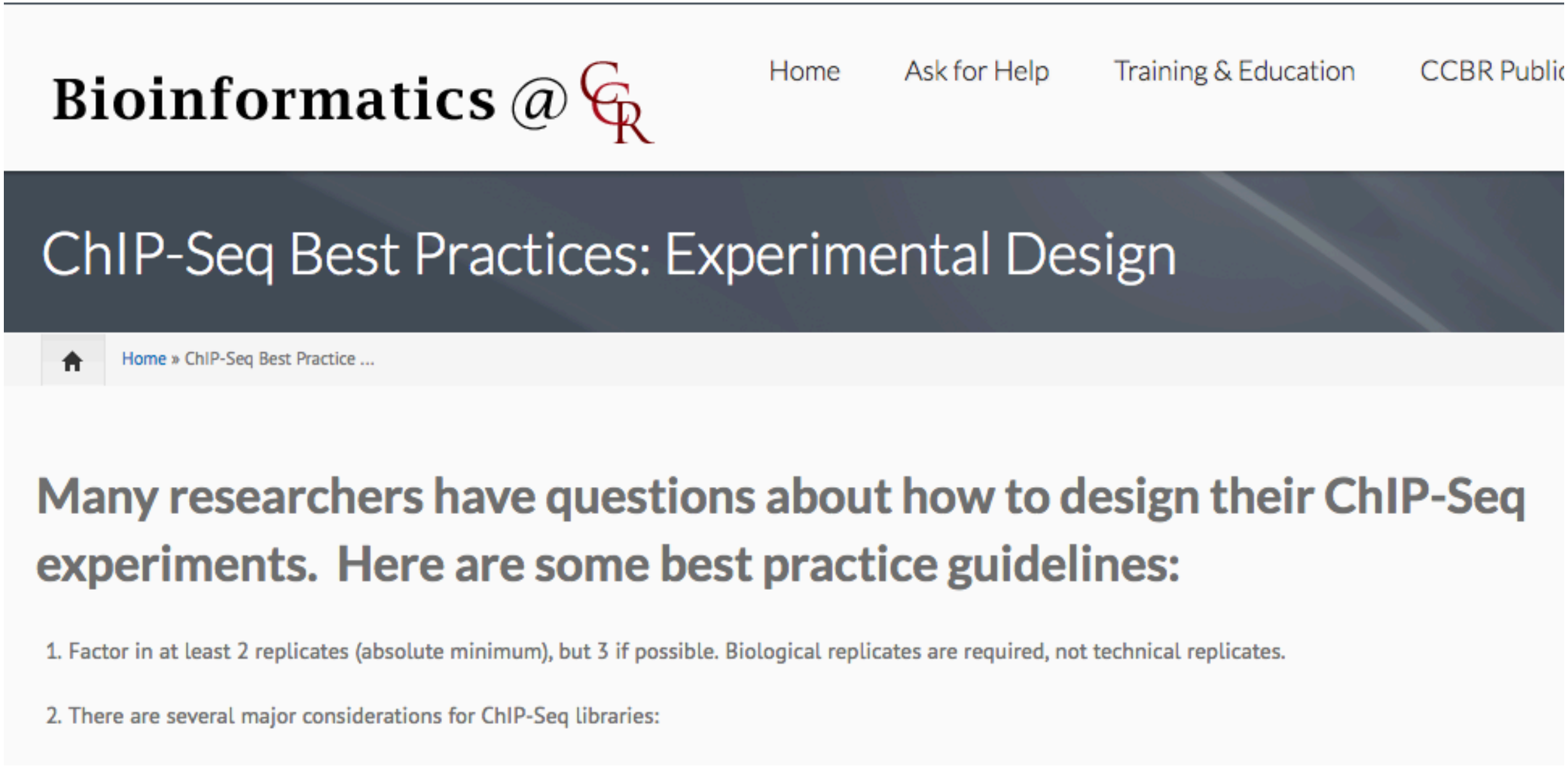
Rene C. Adam[1], Hanseul Yang[1], Shira Rockowitz[2], Samantha B. Larsen[1], Maria Nikolova[1], Daniel S. Oristian[1], Lisa Polak[1], Meelis Kadaja[1], Amma Asare[1], Deyou Zheng[2,3] & Elaine Fuchs[1]

ONCOGENE REGULATION

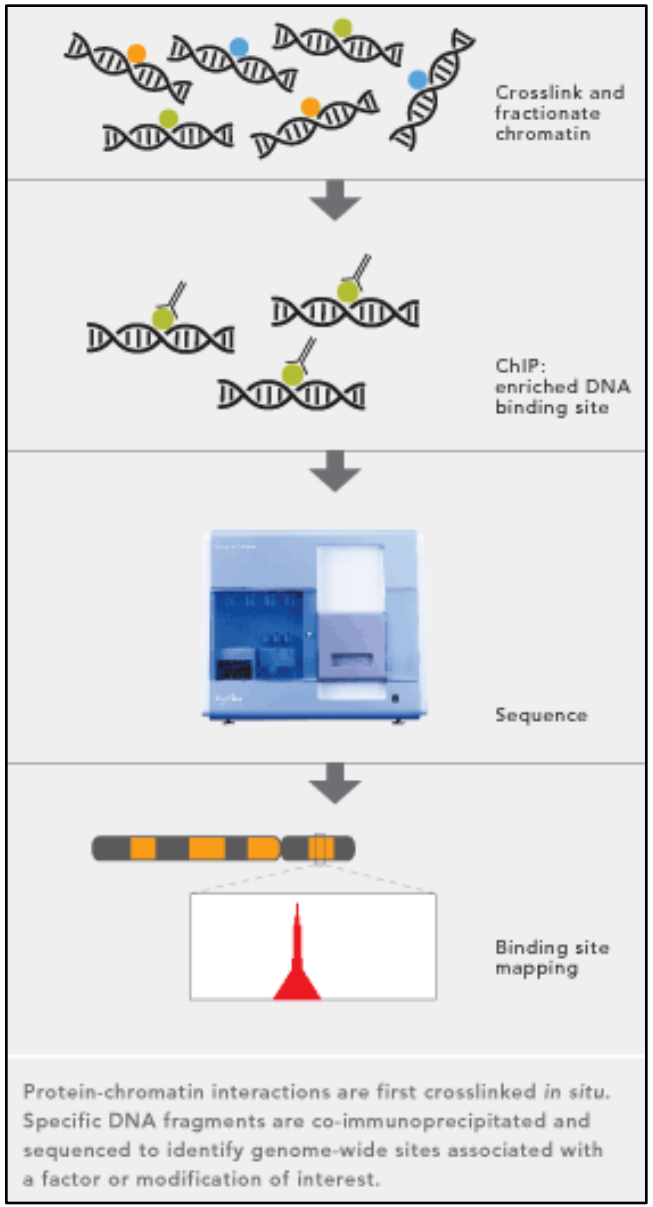
An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element



https://bioinformatics.cancer.gov/content/chip-seq



Considerations for High quality ChIP-seq experiments



Experimental Phase

- Specificity of antibodies
- Low IP efficiency
- Absence of Replicates/Controls

Sequencing Phase

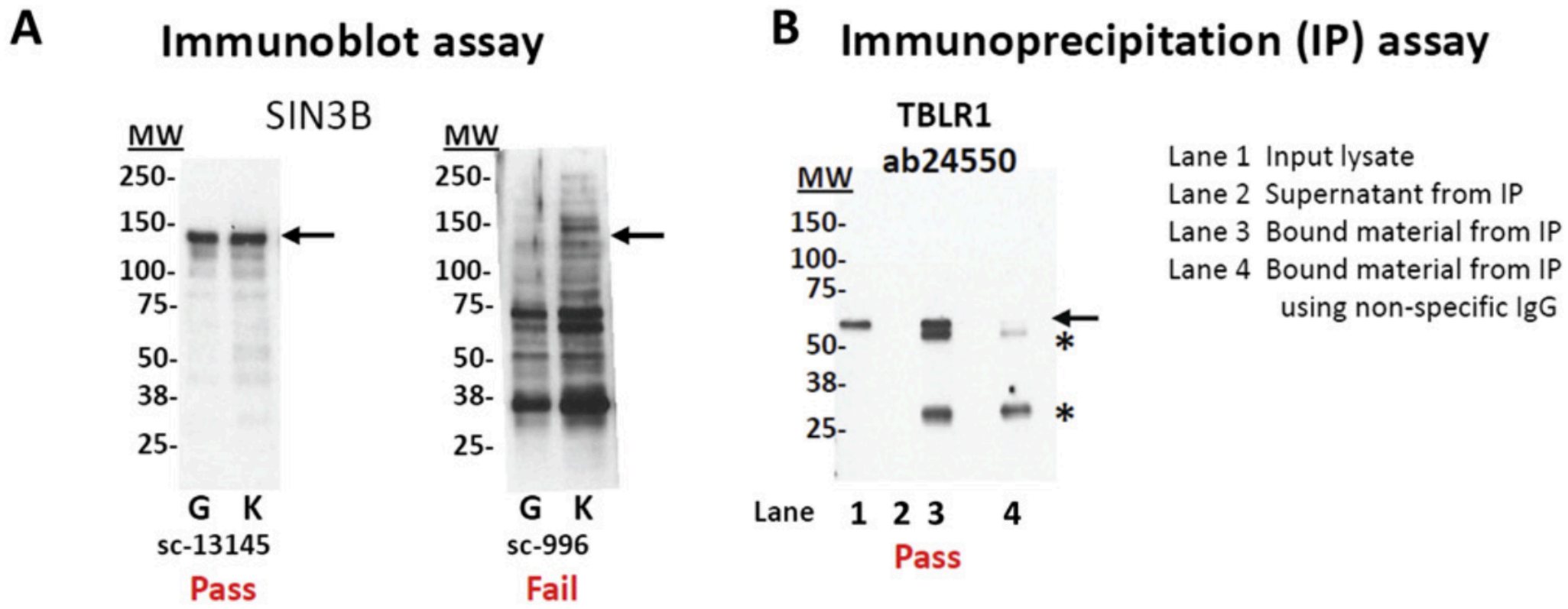
- Library construction
- Sequencing chemistry
- Instrumentation
- Depth of coverage

Analysis Phase

- Different protein classes require different analysis
- Parameter optimization
- Comparisons across experiments

Anand Merchant

A high quality antibody is important!

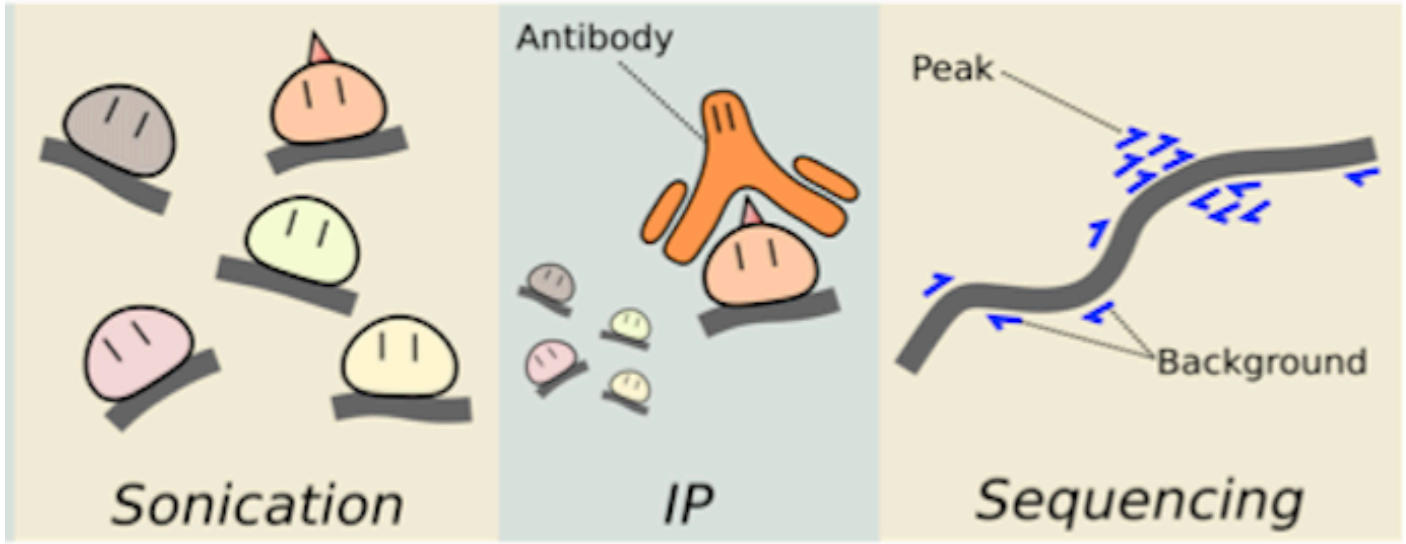


Landt, S. G, et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–31. http://doi.org/10.1101/gr.136184.111

ChIP controls are important!

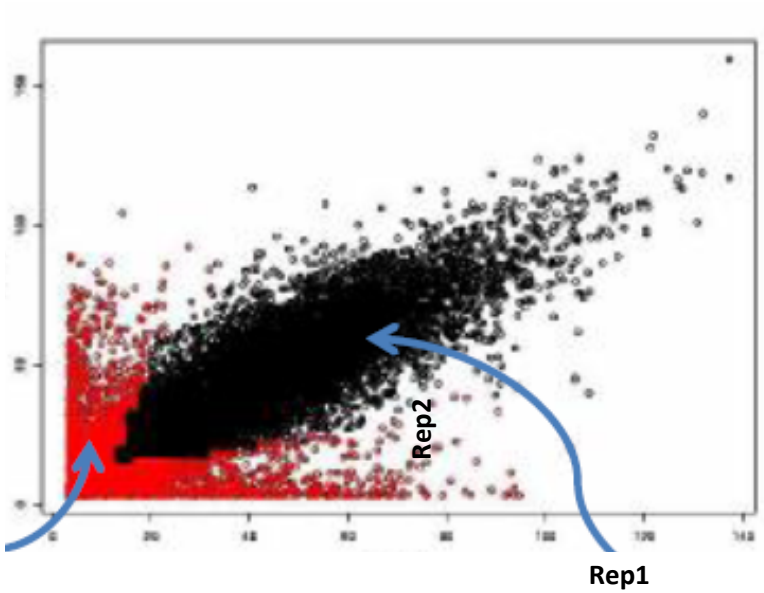
- Input control
  - generally gives more complex backgrounds

- IgG control
  - good in model the effect of antibody (but less complex library)

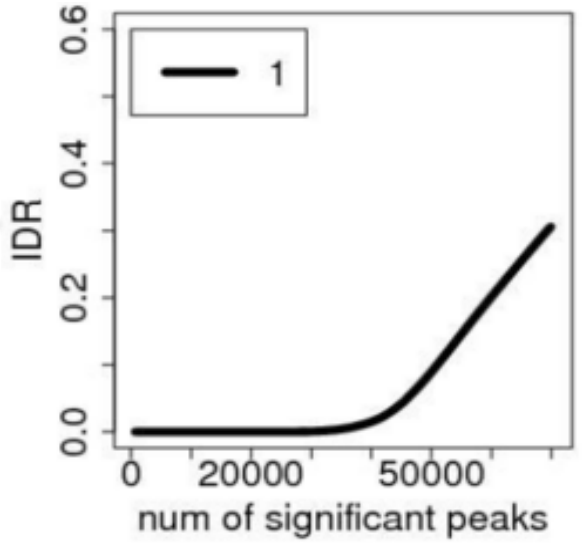


Replicates are important!

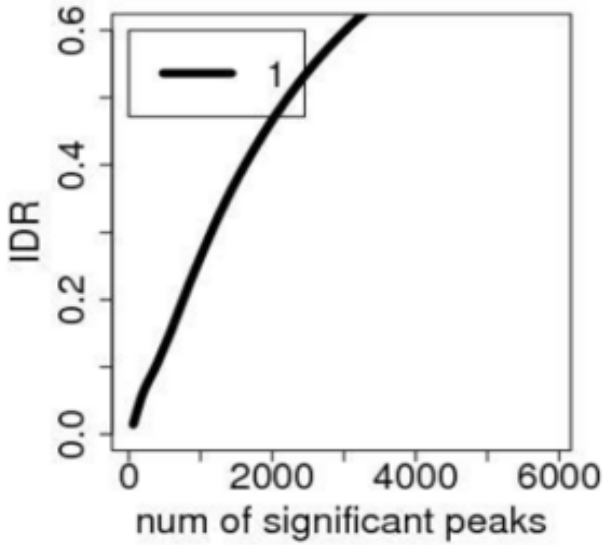
- At least two replicates, three is better.



Good reproducibility

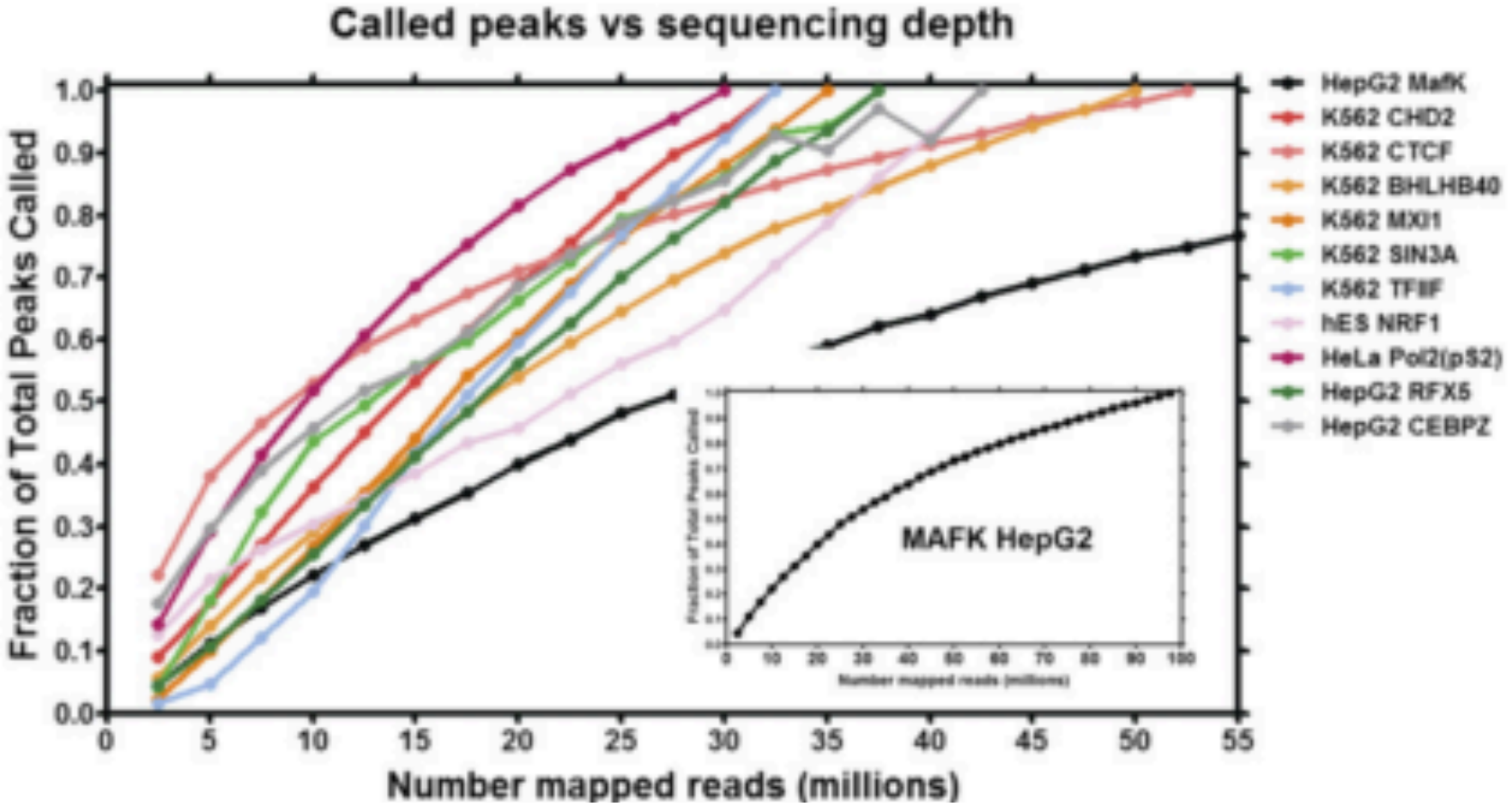


Poor reproducibility

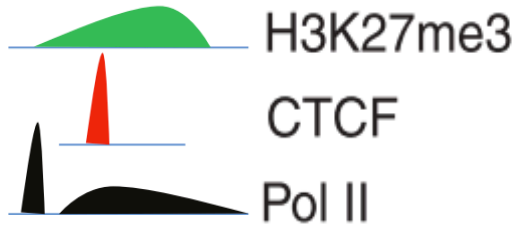


Landt, S. G, et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–31. http://doi.org/10.1101/gr.136184.111

Enough depth of sequencing is important!



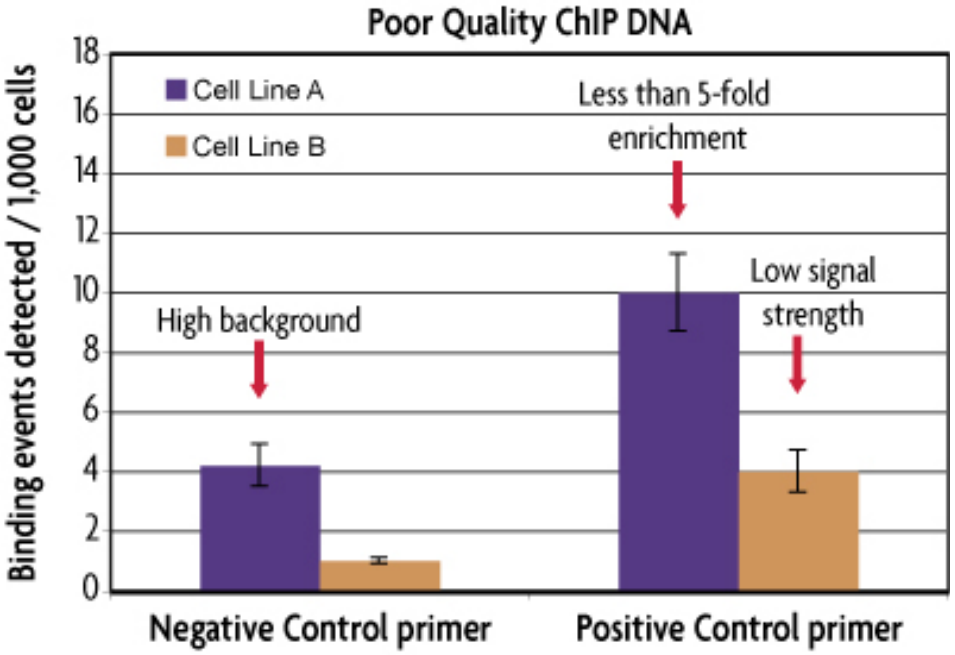
- 30 million reads for TF
- >50 million reads for Histone marks

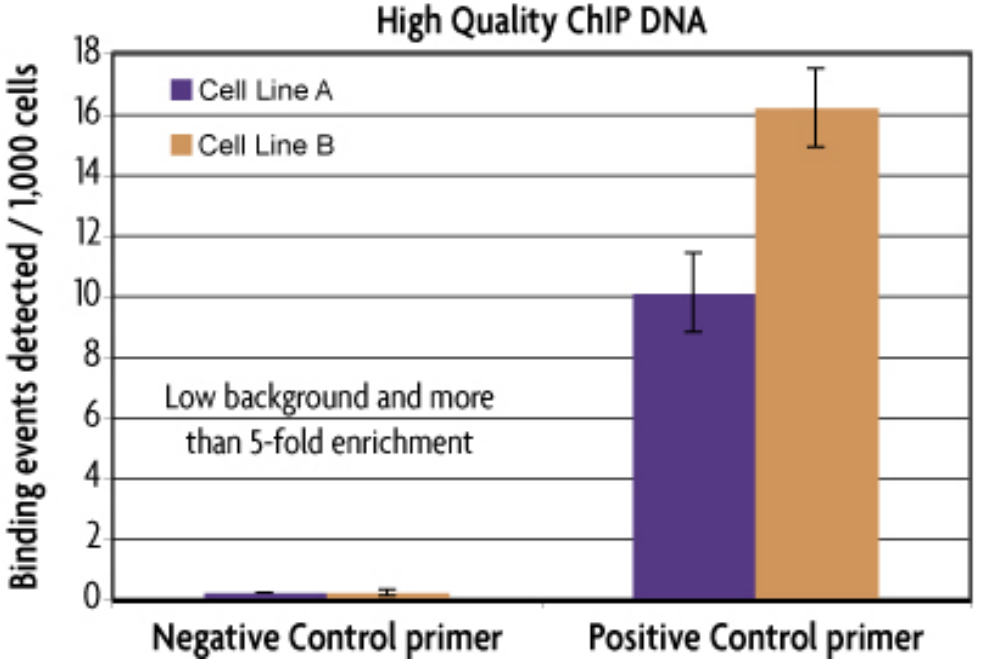


Landt, S. G, et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–31. http://doi.org/10.1101/gr.136184.111

Final check before sequencing!

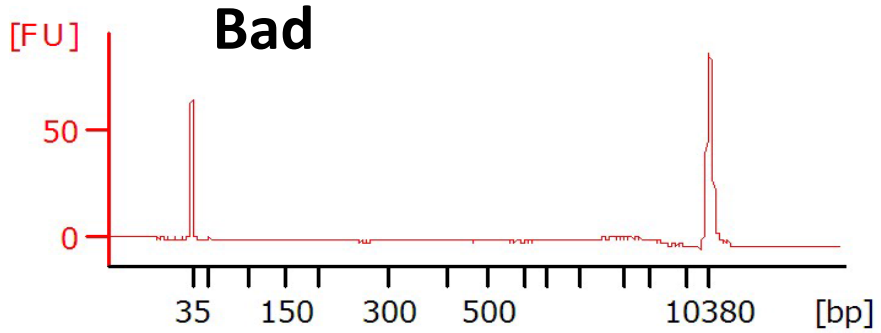
- 10ng is required by SF
- ChIP-qPCR

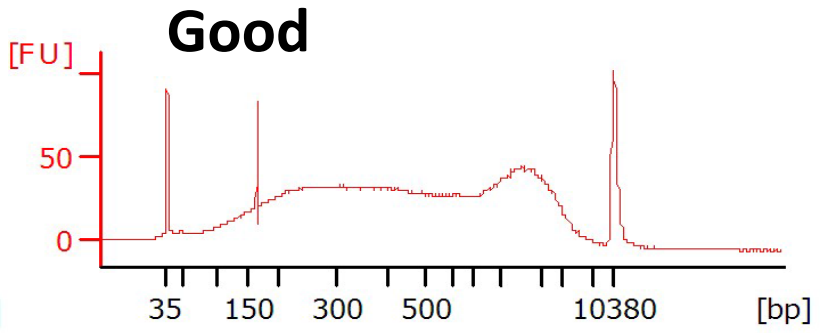




- good amount of DNA at the fragment range 300-500

Bio-analyzer results





https://bioinformatics.cancer.gov/btep/event/201

Center for Cancer Research
BTEP
Bioinformatics Training & Education Program
OSTR,CCR,NCI,NIH

Home | Ask for Help | Training & Education | CCBR Publications

Stepping into the Regulome: ChIP-Seq/ENCODE Data Analysis (2-day)

Home » Stepping into the Regu ...

BTEP Resources

- Archive
- BTEP Software
- Calendar
- Contacts
- Registration
- Schedule

External Resources

- CCR
- OSTR
- NCI-wide Additional Traning
- NIH Library Bioinformatics
- Other NIH Institutions

Probing DNA-Protein Interactions

ChIP-Seq/ENCODE Data Anaysis (2-day)

This 2-day course, which includes both lecture and hands-on components, will teach the basic concepts and practical aspects of ChIP-Seq data analysis. Learn everything from experimental design to statistical analysis and several downstream motif and pattern discovery methods using both commercial and open source software.

More details to be announced shortly.

NOTE: This is a BYOC (Bring your own laptop Computer) class. Government issued or personal computers are permitted. We will be able to supply a very limited set of computers, so if you want to take the class but cannot bring your own computer please indicate such in the Comment section on the registration form.

You will be able to register for this event on April 1st, 2017 at 12:15am

Date & Time: Monday, 17 April 2017 - 9:00am to Tuesday, 18 April 2017 - 4:00pm

Location: Bldg 10 FAES room 4 (B1C205)

Presenter: Multiple

Affiliation: To Be Announced

Format: Hands-on

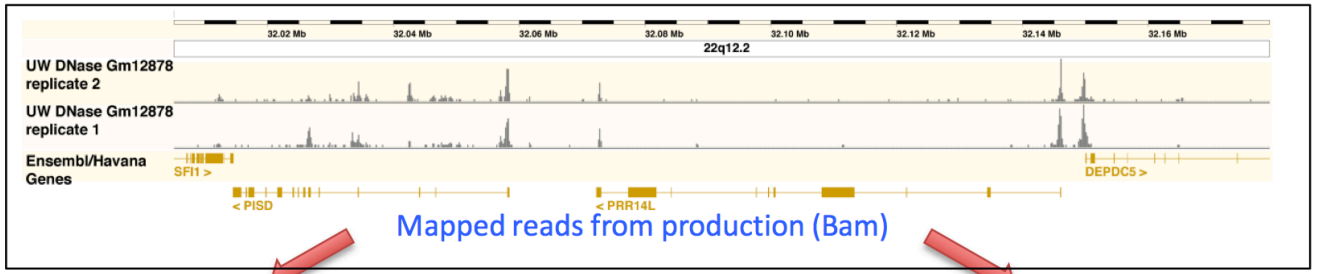
Registration Start Date:
Saturday, 1 April 2017 - 12:15am

Thank you for listening!

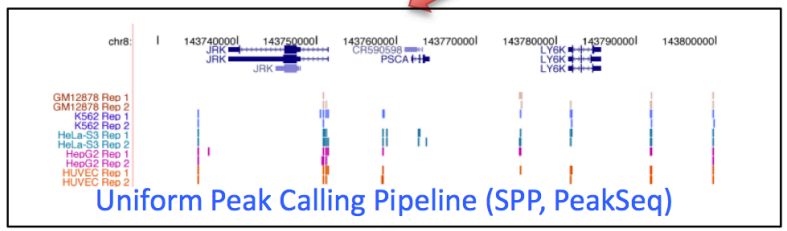
- Any questions?

ENCODE Uniform Analysis Pipeline

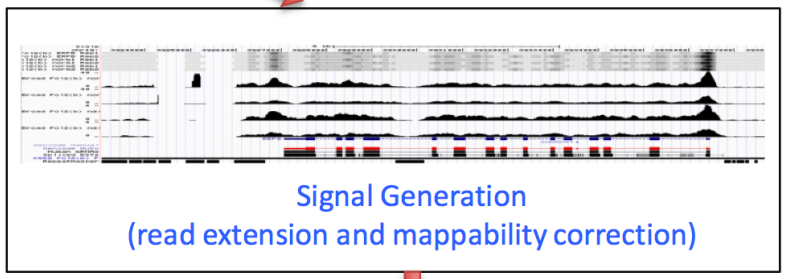
Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour



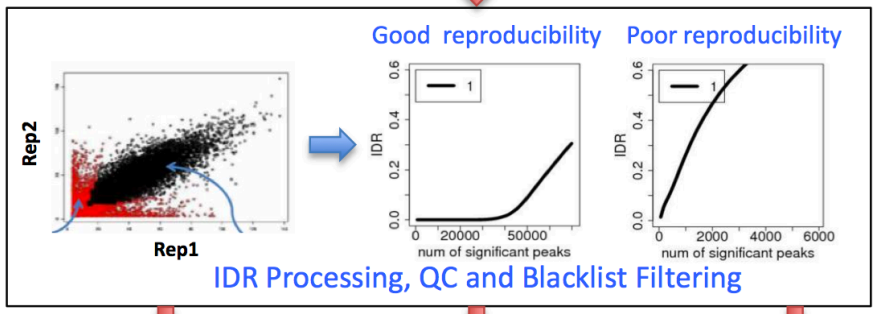
Mapped reads from production (Bam)



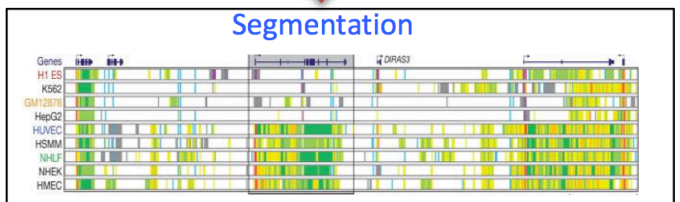
Uniform Peak Calling Pipeline (SPP, PeakSeq)



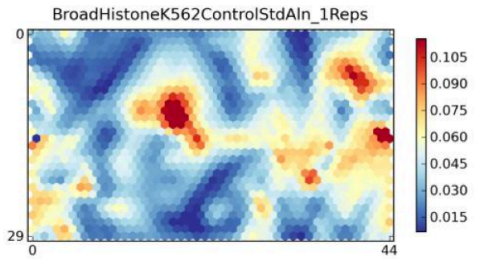
Signal Generation
(read extension and mappability correction)



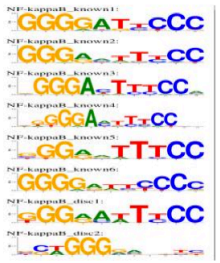
IDR Processing, QC and Blacklist Filtering



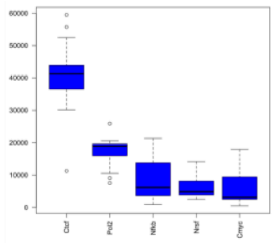
ChromHMM/Segway



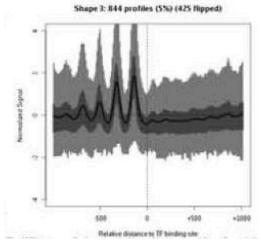
Self Organising Maps



Motif Discovery

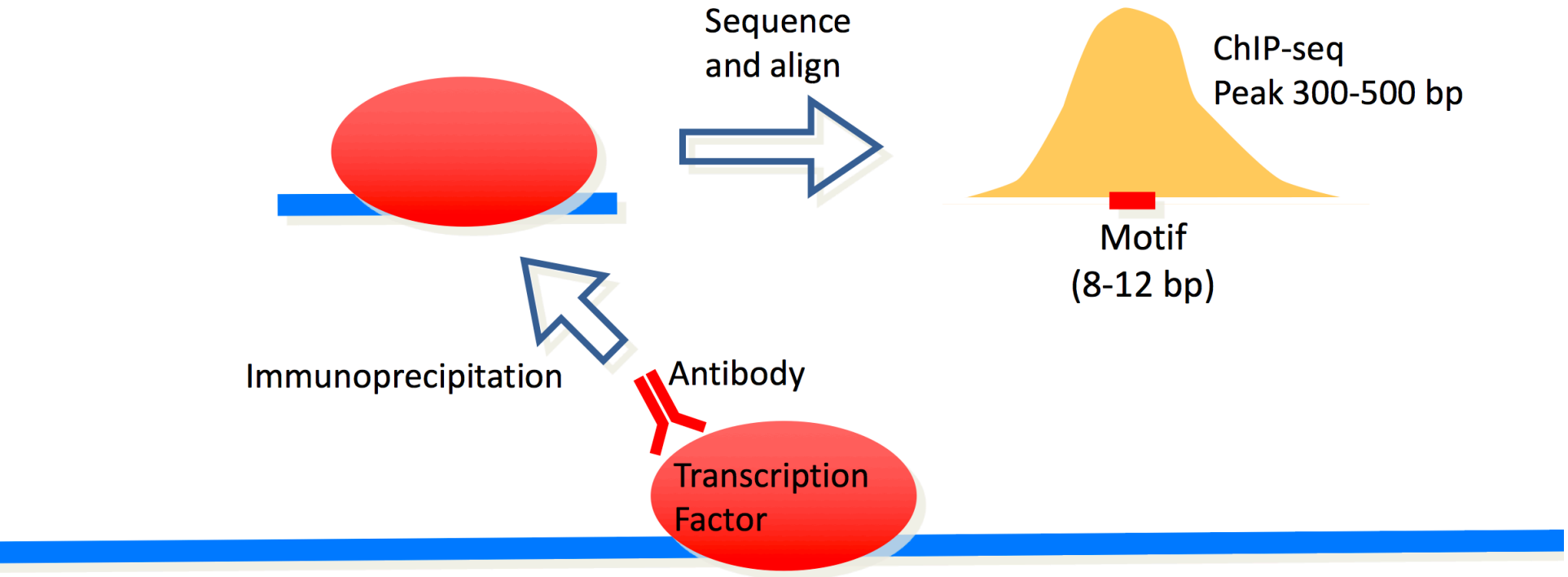


Stats, GSC enrichments, etc.

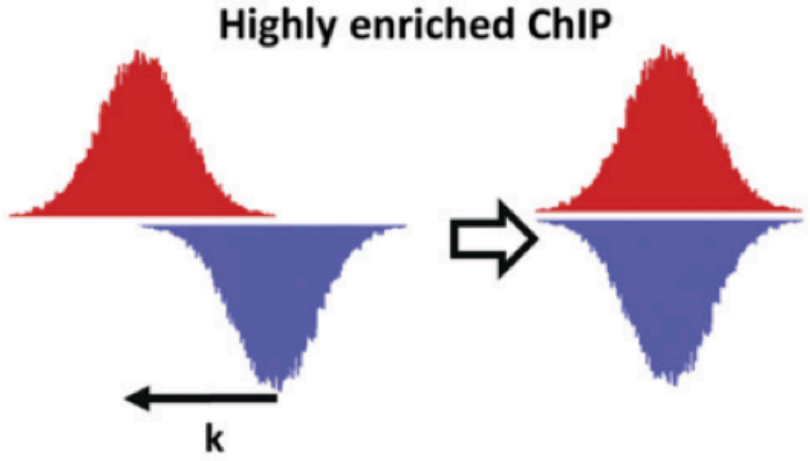


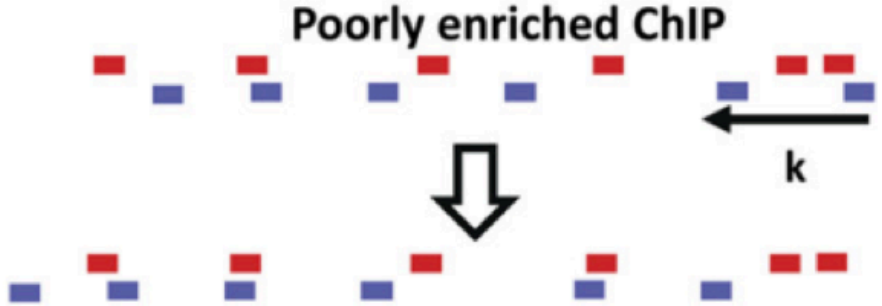
Signal Aggregation over peaks

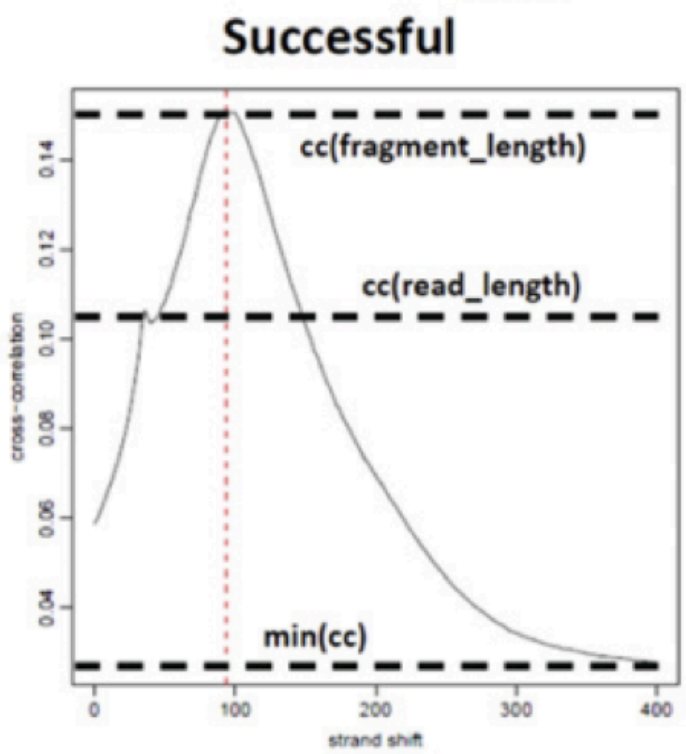
Functional data: ChIP-seq

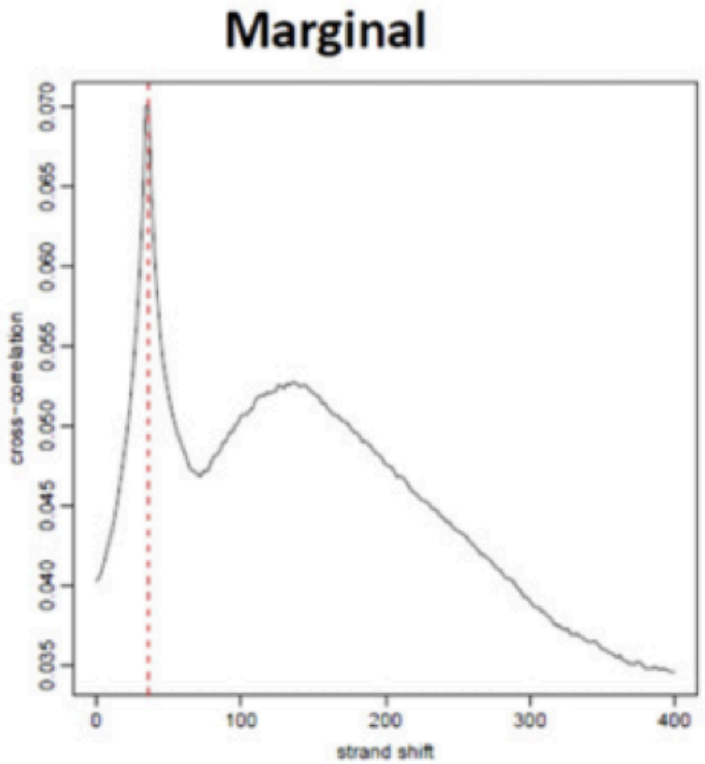


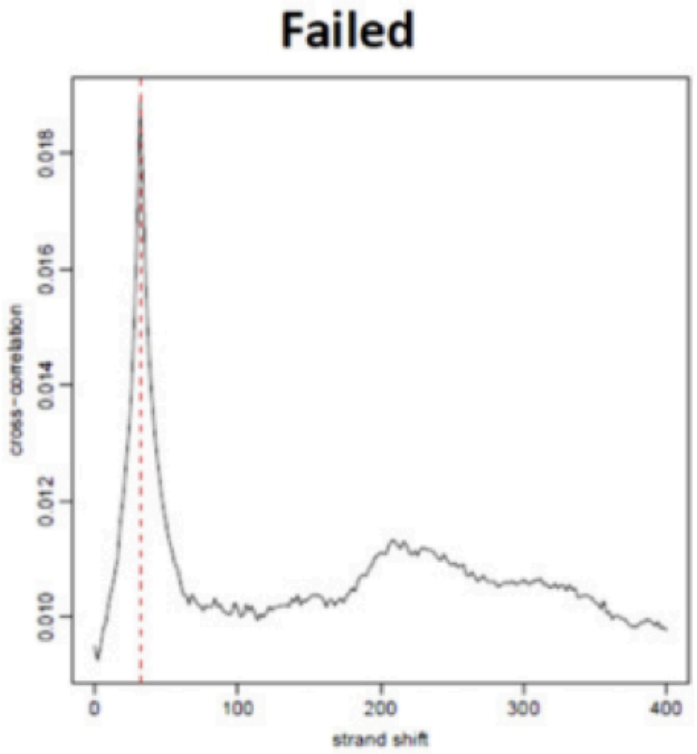
ChIP-exo
Histone Marks





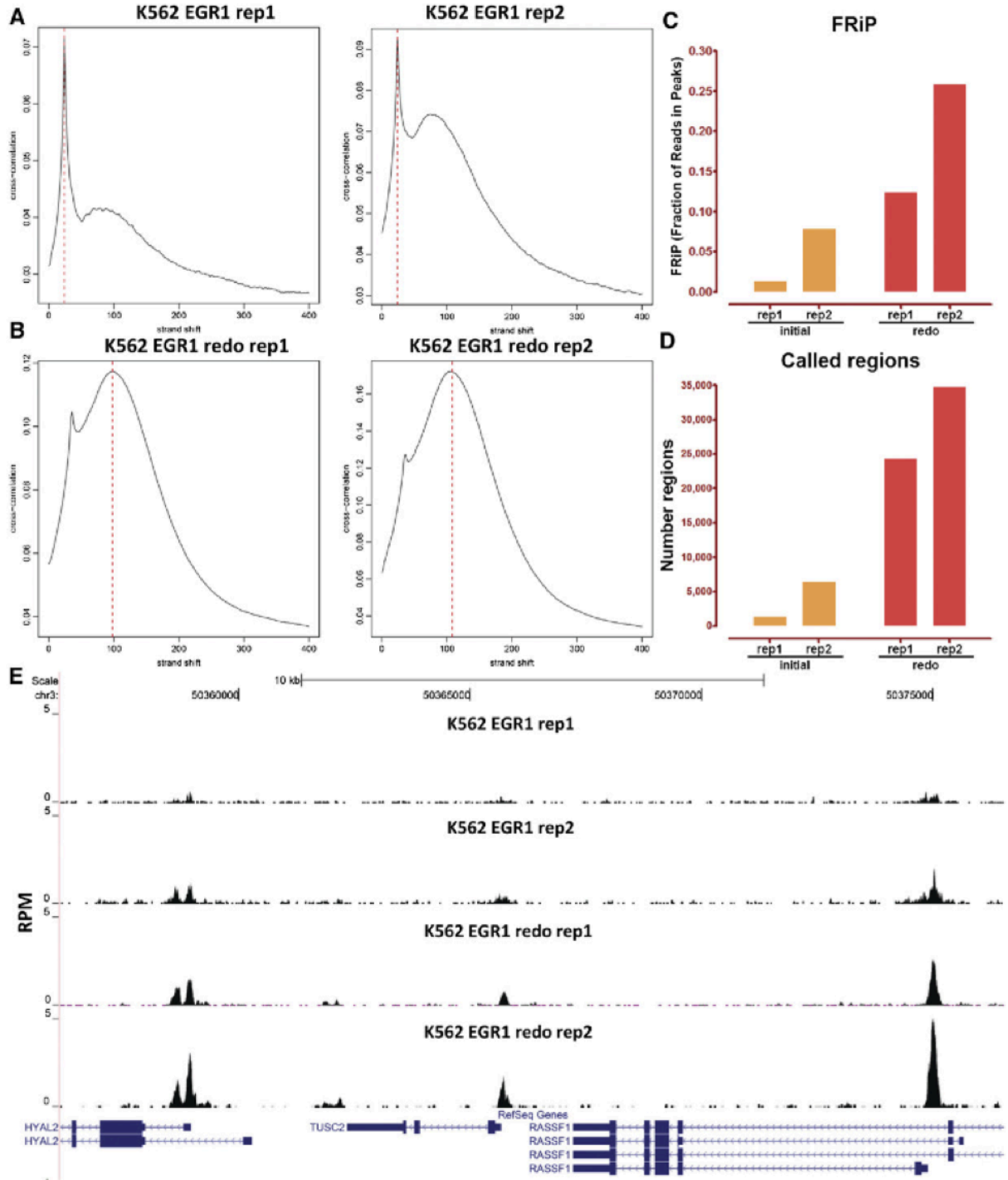




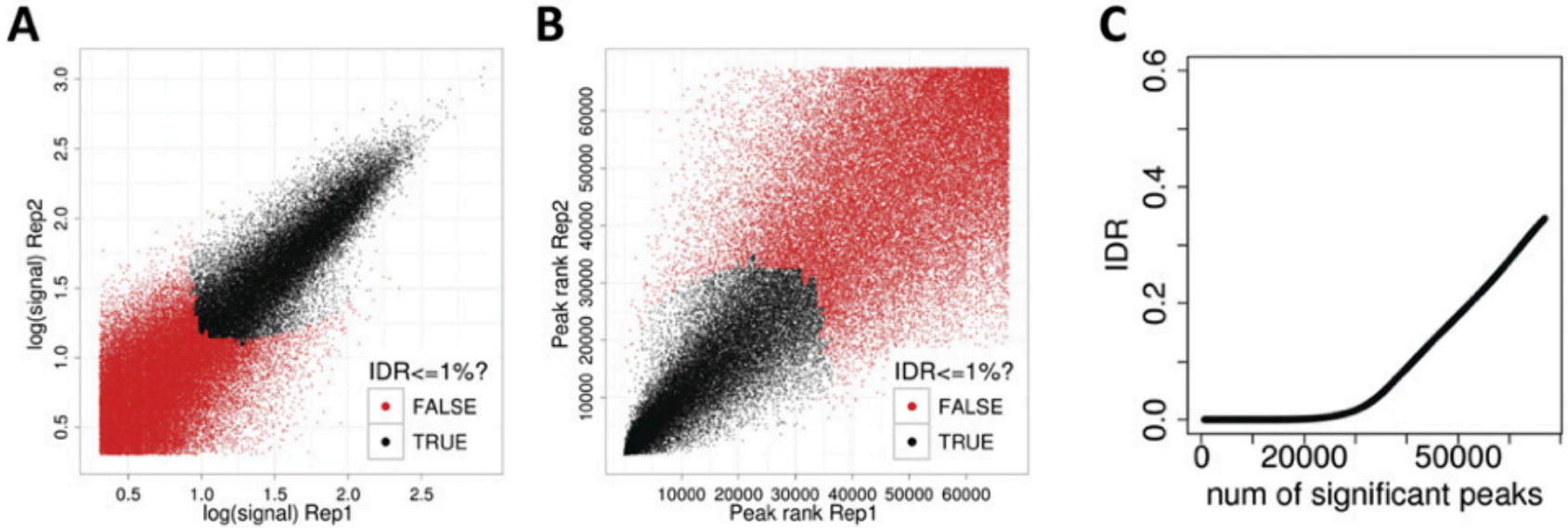


$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

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


A
K562 EGR1 rep1
K562 EGR1 rep2
cross-correlation
strand shift
B
K562 EGR1 redo rep1
K562 EGR1 redo rep2
C
FRiP
FRiP (Fraction of Reads in Peaks)
rep1
rep2
initial
redo
D
Called regions
Number regions
E
Scale
chr3:
10 kb
5036000
5036500
5037000
5037500
K562 EGR1 rep1
K562 EGR1 rep2
RPM
K562 EGR1 redo rep1
K562 EGR1 redo rep2
RefSeq Genes
HYAL2
TUSC2
RASSF1

RAD21 Replicates (high reproducibility)



SPT20 Replicates (low reproducibility)