Basics of ChIP-Seq

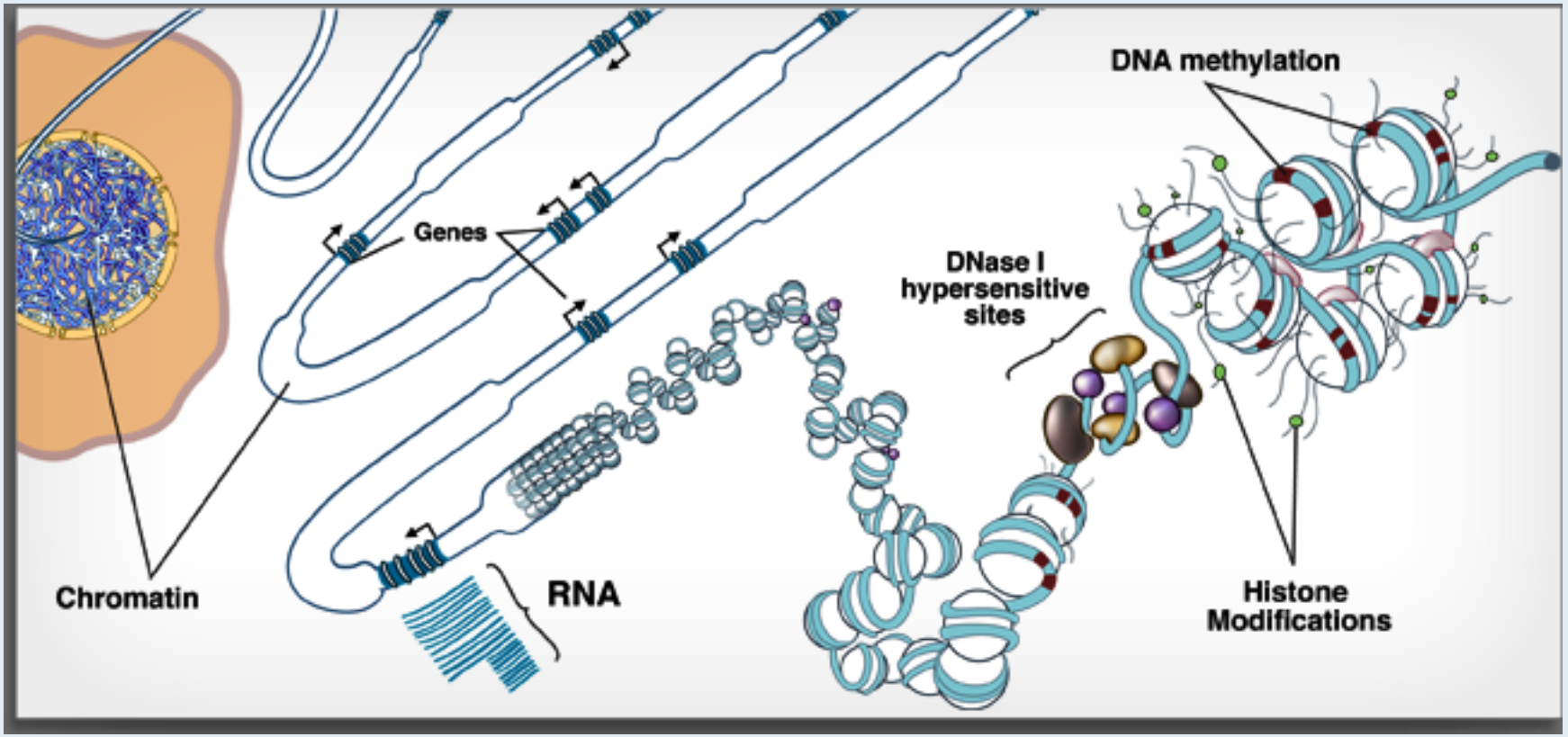
Lauren Mills Ph.D
RISS @ MSI

Slides : https://www.msi.umn.edu/tutorial-materials

Outline

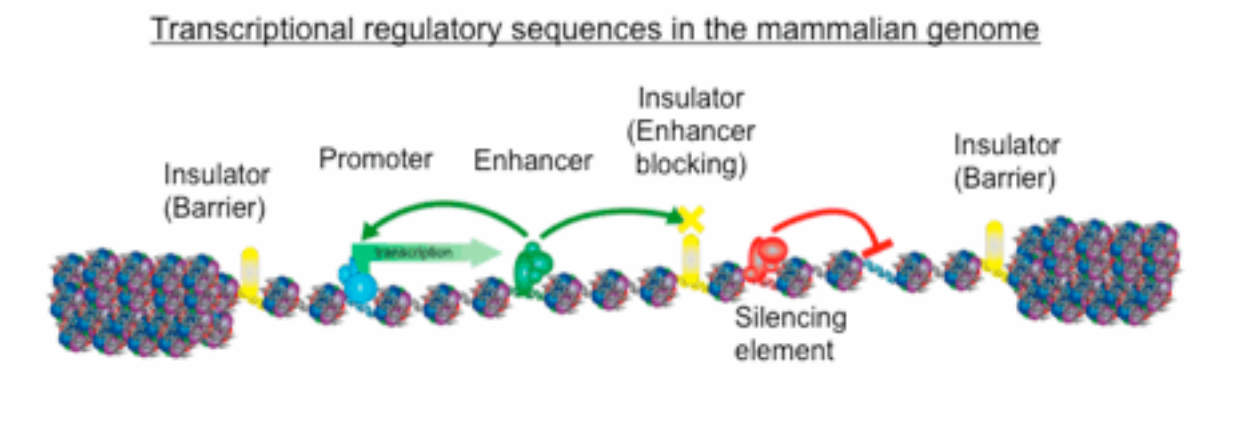
- What can you learn from ChIP?
- Planning an experiment
- Sequencing considerations
- Identifying Peaks
- Looking at your data
- Interacting with data
- Data Sources

ChIP-Seq



ChIP-Seq

- Chomatin immunoprecipitation to isolate fragments of DNA bound by protein of interest.
- Can be used to identify regions of the genome bound by specific Transcription Factors, Histone Modifications or RNA Polymerase.
- Biological function of non-coding parts of the genome i.e., enhancers, silencers, insulators.



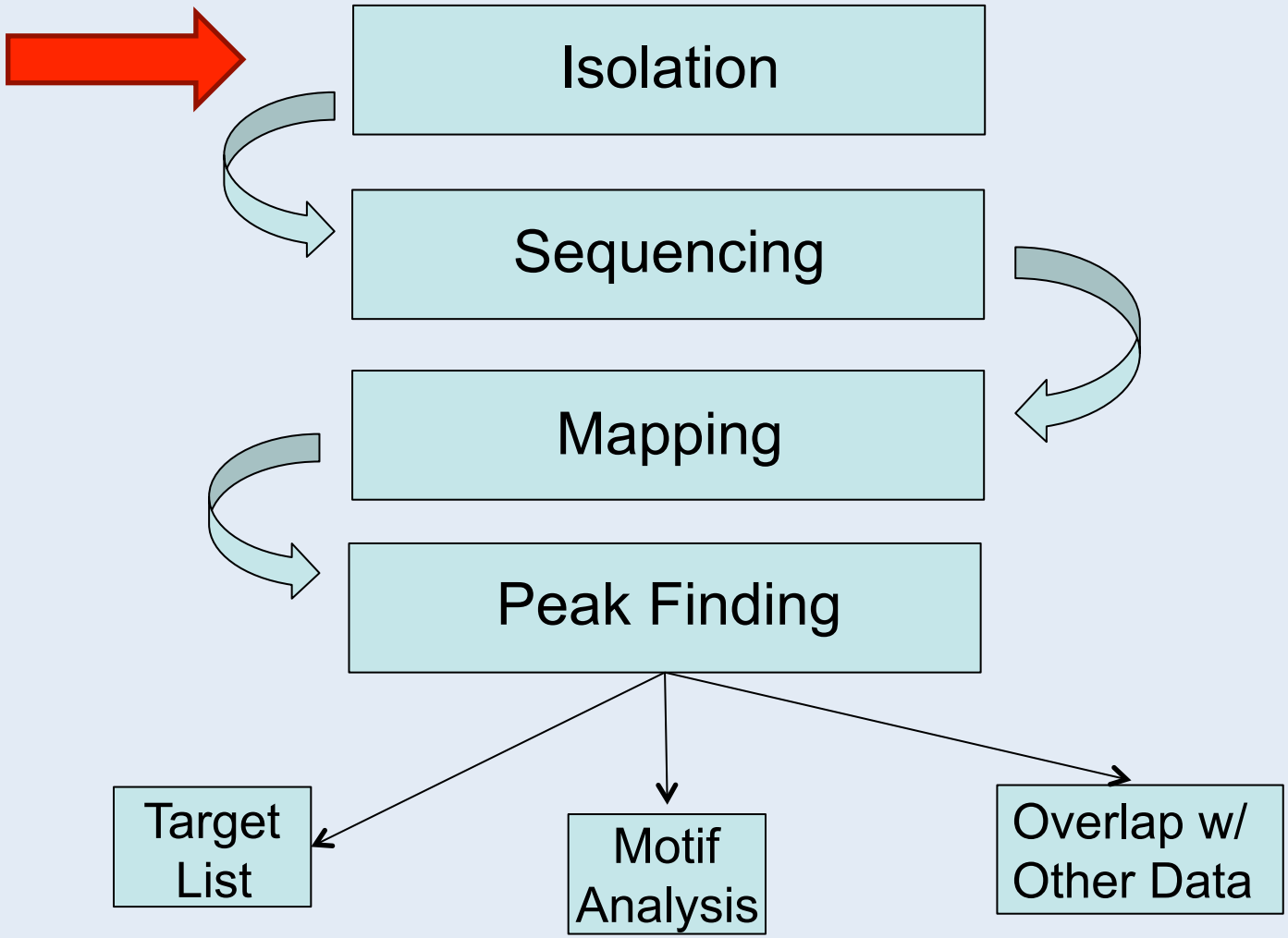
ChIP-Seq

- Transcription Factors
  - Where are they?
  - Sequence preference
  - Correlation with gene expression
- Chromatin Marks
  - Where are they?
  - What combinations do they come in
  - How do they relate to biological status

Public Data Repositories

- Roadmap Epigenomics
  - Human focused
  - Large number of histone marks, DNA methylation, DNase hypersensitivity across many many cell types
- ENCODE
  - Human and Mouse (some)
  - Started with ChIP-chip data has moved into ChIP-seq and other enrichment based sequencing
  - Came before Roadmap
- modENODE
  - ENCODE but for Fruit Fly (7 species) and Worm (4 species)
  - Same goals as ENCODE different organisms

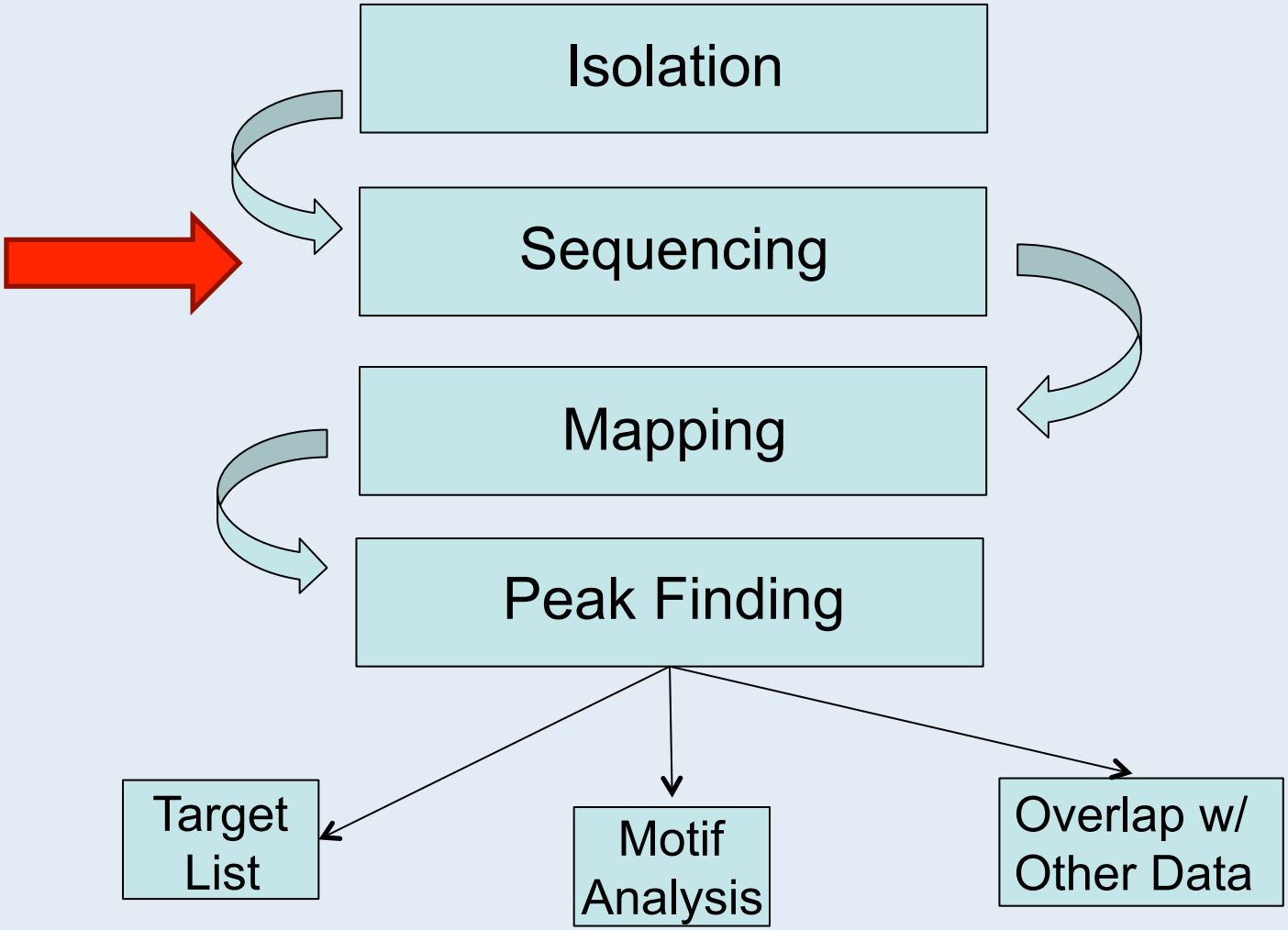
Basic ChIP-Seq Workflow



Isolation

- Results are very dependent on “wet-lab” protocols
  - Use a validated antibody
  - Validate enrichment using PCR before prepping for sequencing
- Input samples are needed for each cell type, antibody and sonication.
- Antibody list from ENCODE: https://www.encodeproject.org/search/?type=antibody_lot

Basic ChIP-Seq Workflow



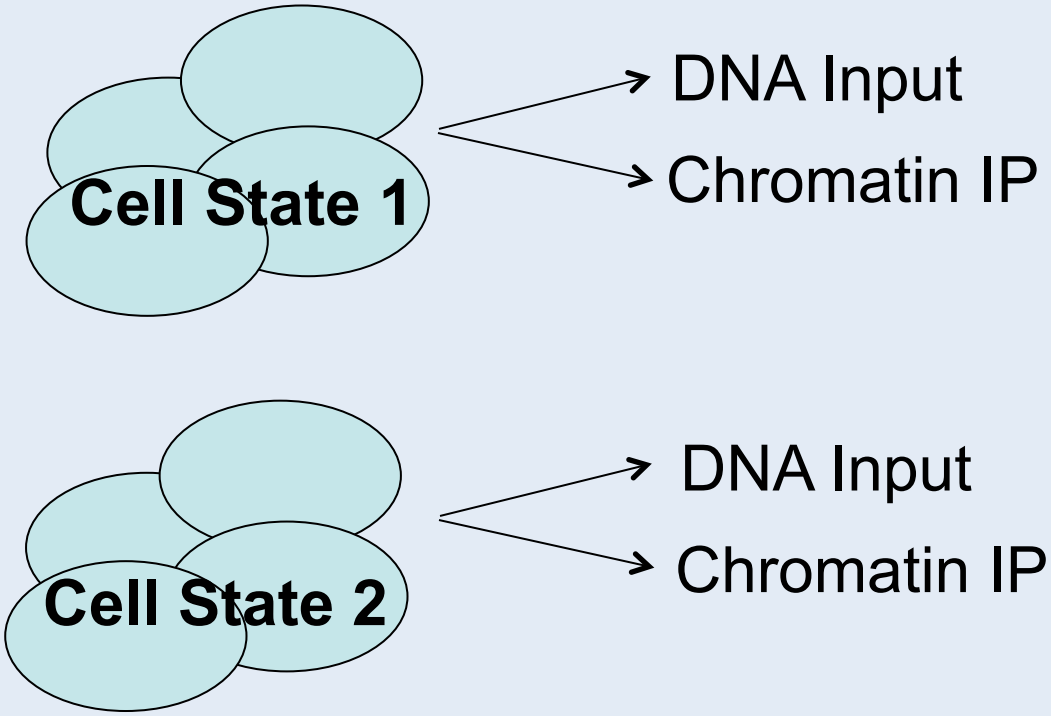
Sequencing Recommendations

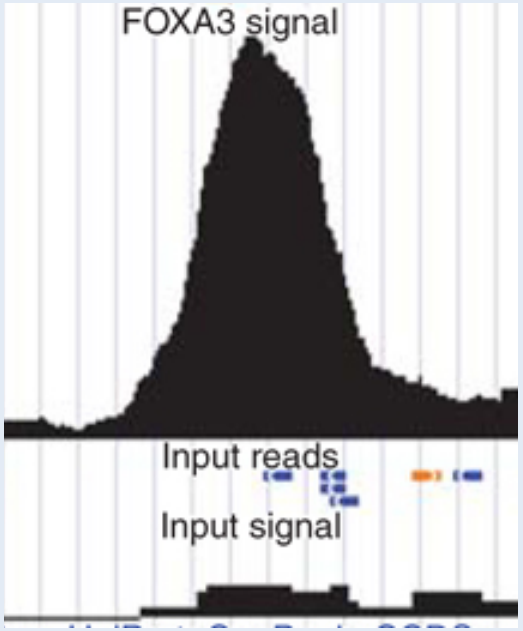
- Large Genomes (human, plant)
  - TF: 10 million uniquely mapped reads
  - Chromatin: 20 million uniquely mapped reads
- Small Genomes (worm, fly, yeast)
  - TF: 2 million uniquely mapped reads
  - Chromatin: 5 million uniquely mapped reads
- At least Two biological replicates are recommended

Sequencing Recommendations

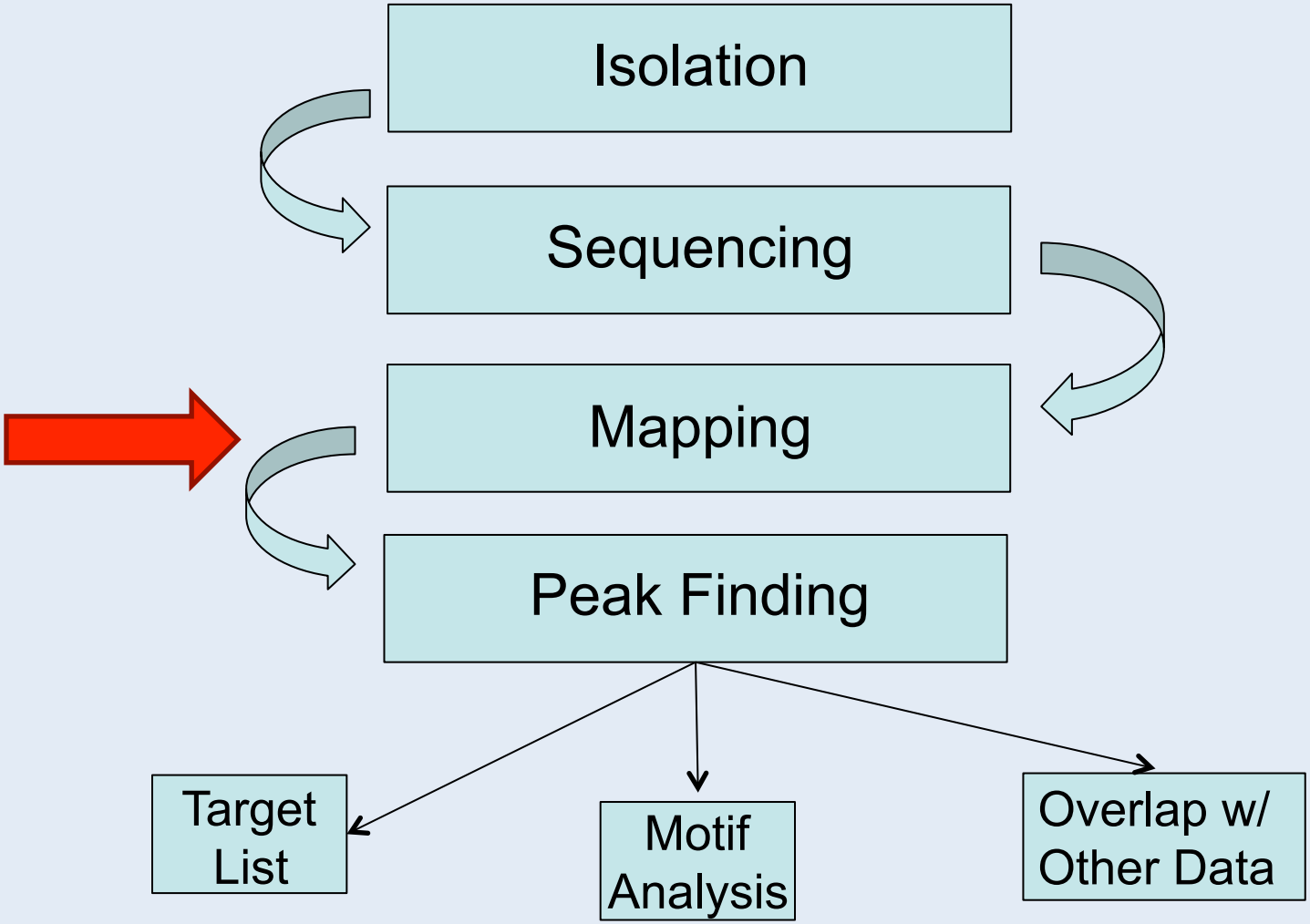
- Enrichment regions will be at least as large as the fragments of DNA sequenced. (Sonication)
- Longer reads do not give you better information. 100bp max is recommended.
- Paired end data is not required and some algorithms can not use paired end data.

Experimental Design





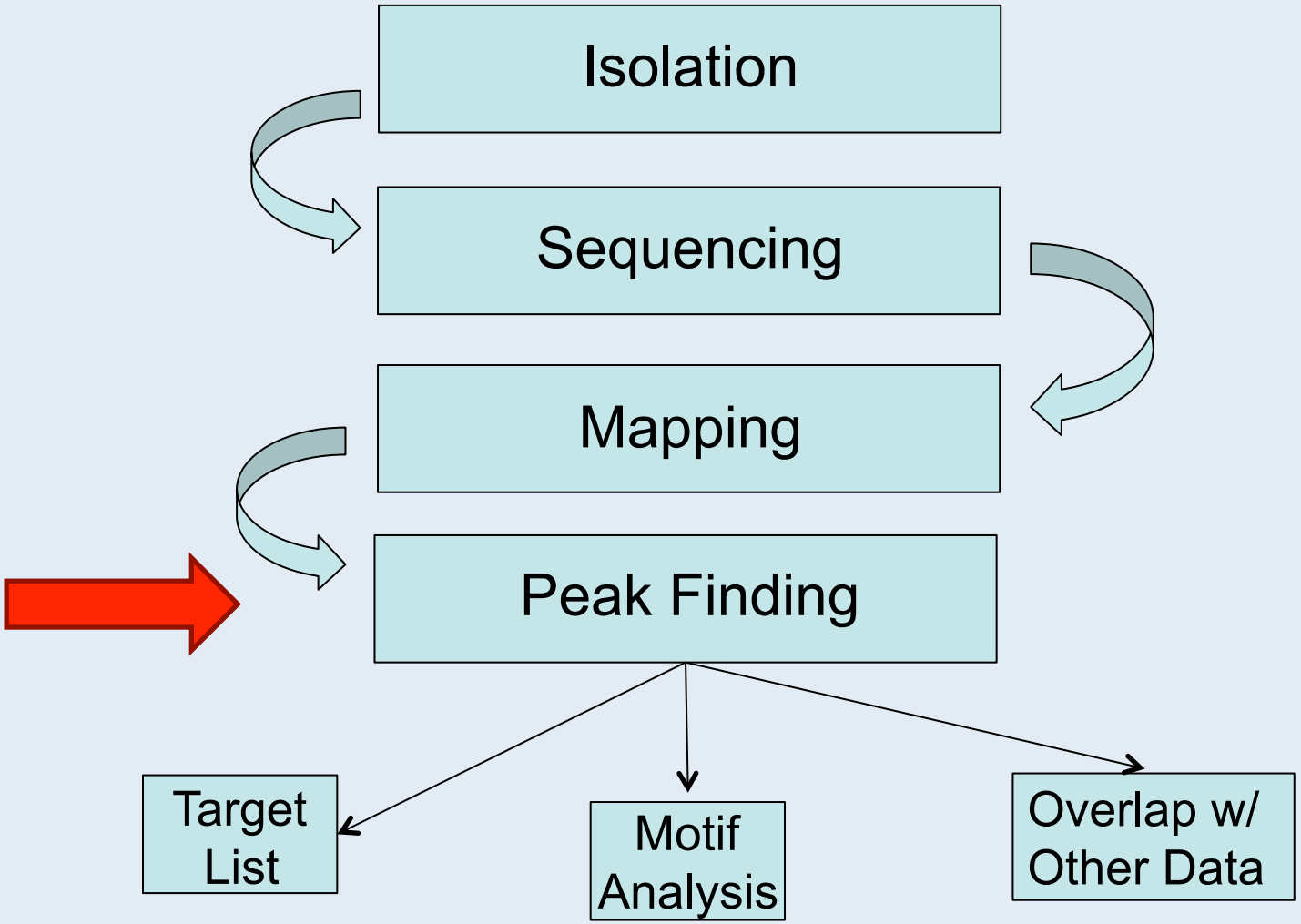
Basic ChIP-Seq Workflow



Mapping Your Reads

- Map biological replicates independently using standard mapping tools i.e., BWA Bowtie.
- Remove PCR duplicates, these are 100% identical reads. SAMtools rmdup
- Depending on the peak finding software used you may need to convert from SAM/BAM to BED format (BedTools)
- Check your mapping stats, if < 80% of your reads map then there may be an issue.

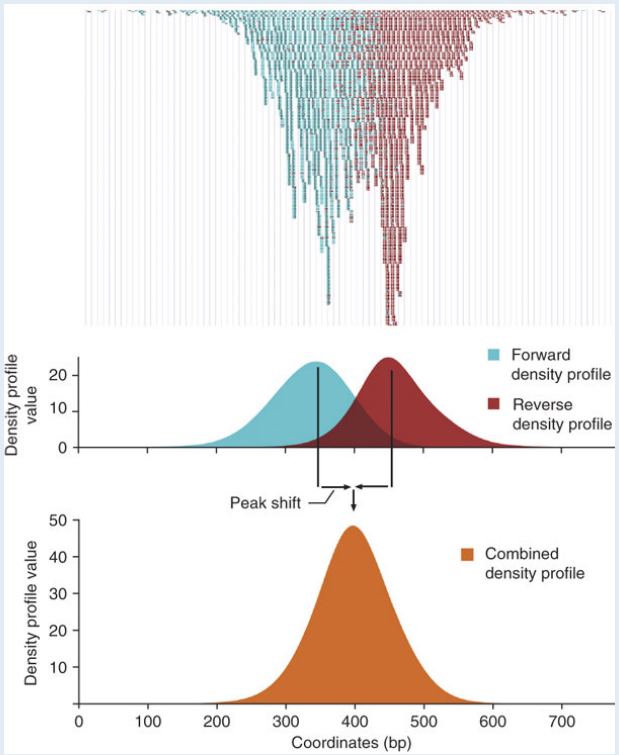
Basic ChIP-Seq Workflow



Peak Calling

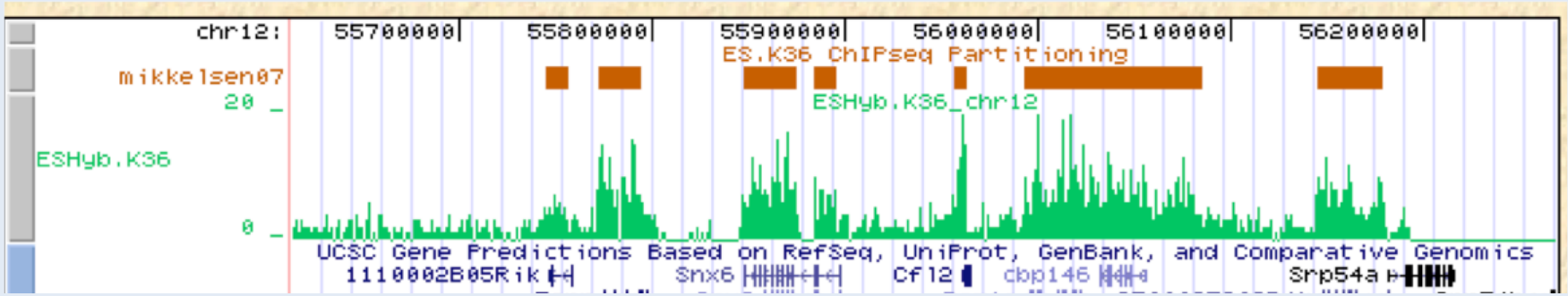
- Peak callers are algorithms that identify regions that have more reads than background.
- Three types: point-source, broad peak callers and those that can do both.
- MACS, MACS2, ZINBA, SICER, CCAT

chr1	3504555	3504771	MACS_peak_1	51.97
chr1	3505222	3505607	MACS_peak_2	92.84
chr1	3506002	3506130	MACS_peak_3	53.32
chr1	4485926	4486384	MACS_peak_4	91.62
chr1	4613288	4613545	MACS_peak_5	73.09
chr1	4758830	4759138	MACS_peak_6	74.85
chr1	4759374	4760160	MACS_peak_7	163.21
chr1	4773712	4774711	MACS_peak_8	283.47



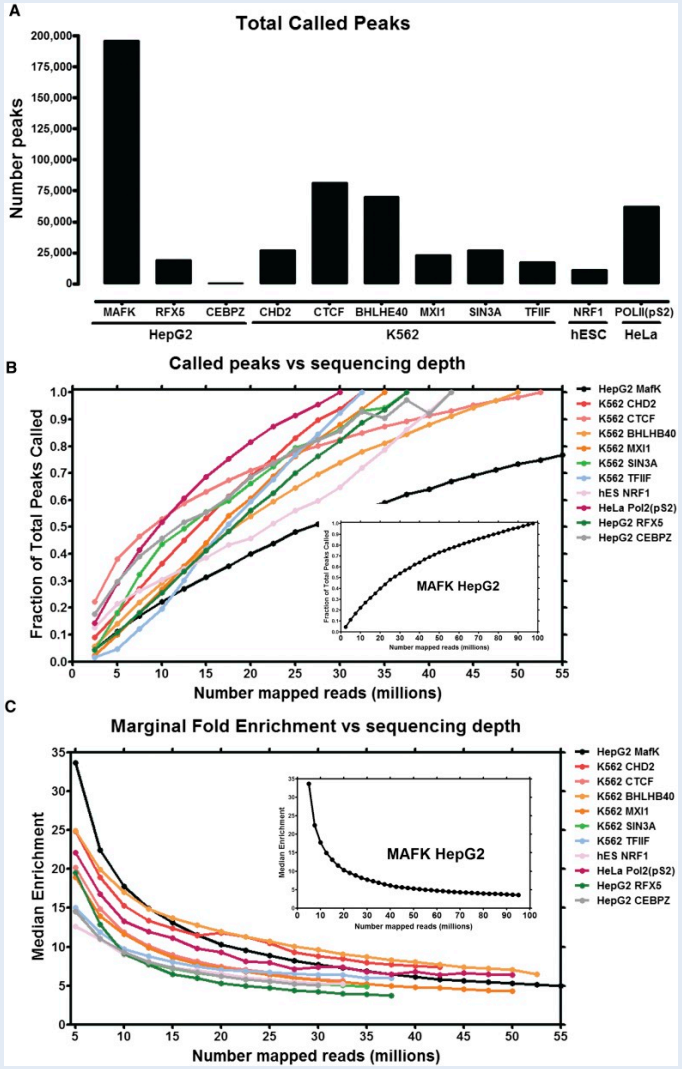
Peak Data

- WIG family- continuous data
  - WIG, BigWIG, bedGraph
- BED family – discrete locations
  - Bed, Bed6, Bed12
- Your data can be visualized in genome browsers such as IGV, UCSC and Ensembl



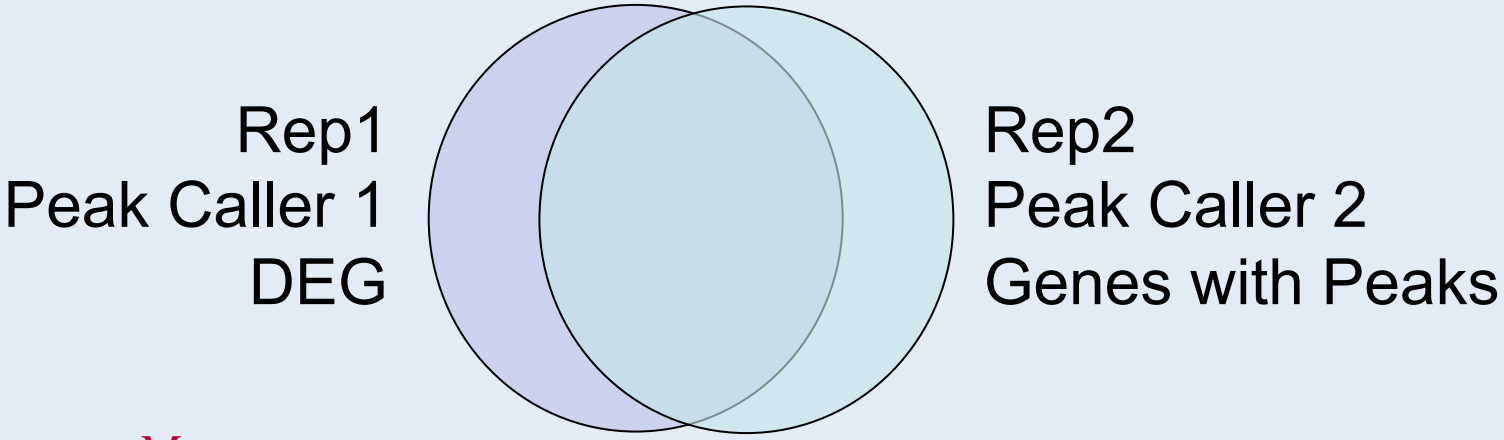
Which Peaks?

- Depending on the software peaks can be evaluated by FDR, reads in the peak, enrichment ratio.
- It is common to get 10s of thousands of peaks in a sample.
- The more reads you have the more peaks you will get.



Which Peaks?

- What do positive control regions look like? P-value, enrichment, score, rank
- Which peaks are found in all of the biological replicates?
- Top 10,20,30% of peaks?

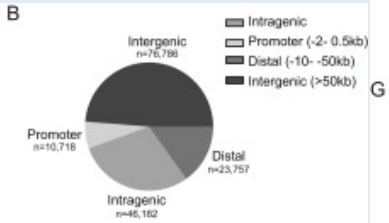


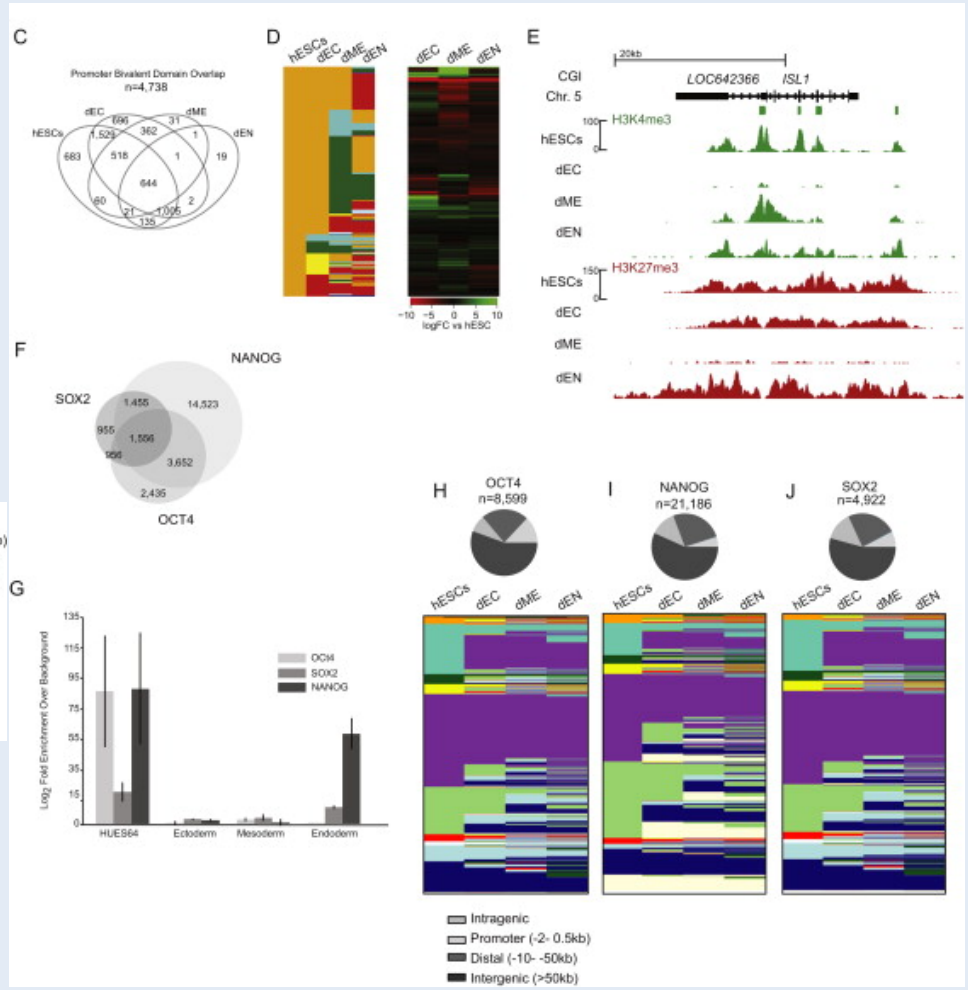
Visualize the Data

- What genomic context do your peaks reside in?
- Do they co-localize with other peaks? Differentially expressed genes? Know TFBS?
- What changes between samples, states, time?

A

State	hESCs	dEC	dME	dEN
H3K4me3 & H3K27me3	11,828	9,174	2,737	4,776
H3K4me3 & H3K27ac	11,583	6,815	9,182	9,959
H3K4me3	15,443	11,262	14,408	36,180
H3K27me3 & H3K4me1	3,064	5,481	674	1,640
H3K27ac	20,345	12,330	29,818	23,186
H3K4me1	40,035	54,836	48,159	29,402
H3K27me3	15,374	16,921	20,405	20,574
H3K9me3	61,426	56,272	60,980	58,431
IMR	10,272	12,667	16,930	13,652
HMR	63,499	63,724	51,111	59,339
None	15,030	18,312	13,186	10,543

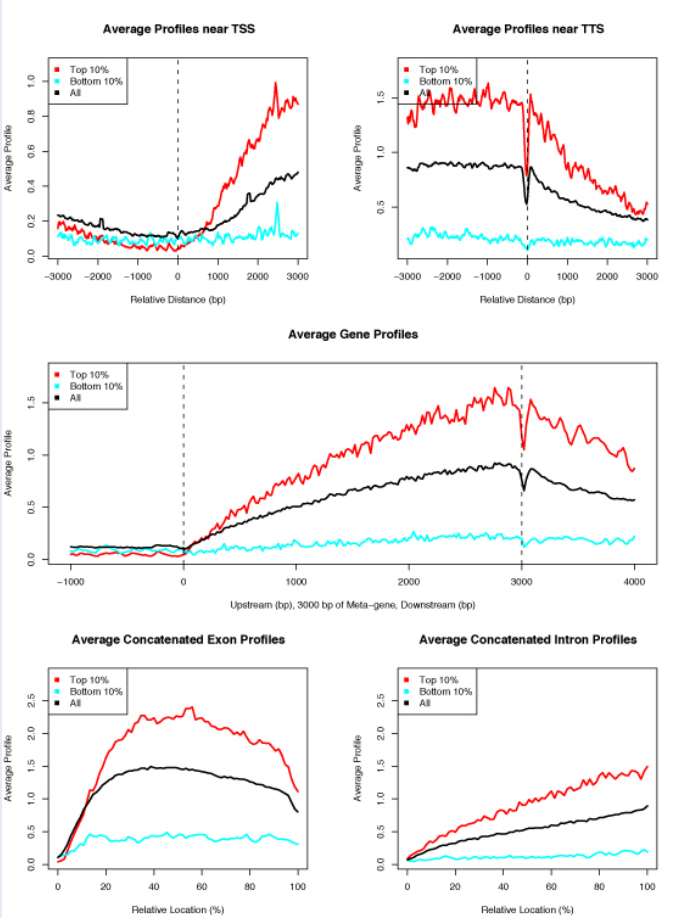


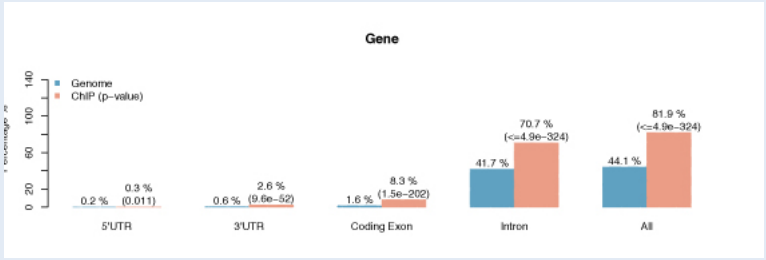


Ren and colleagues
Cell, Vol. 153, Issue 5, p1134–1148

UNIVERSITY OF MINNESOTA

Visualize the Data





- Most visualizations are custom (R, MATLAB)
- CEAS, venn Diagram and heatmap tool in Galaxy
- Command line and R packages for visulizations: http://omictools.com/data-visualization-c479-p1.html

Next Steps

- Motif Identification
  - Use only highest confidence peaks
  - Multiple tools and methods, MEME is very popular
- Differential peak calling
  - Still a new problem but tools do exist
  - Samples need to be treated uniformly
- More Advanced Integrative analysis
  - Machine learning techniques can identify patterns that exist in large data sets
  - ChromHMM

Links for more Info

- MACS: https://github.com/taoliu/MACS/blob/macs_v1/README.rst
- MACS2: https://github.com/taoliu/MACS/
- SCIER: http://home.gwu.edu/~wpeng/Software.htm
- ZINBA: https://code.google.com/p/zinba/
- OmicTools Chip_seq: http://omictools.com/chip-seq-c1215-p1.html
- chromHMM: http://compbio.mit.edu/ChromHMM/
- MEME: http://meme.nbcr.net/meme/doc/meme-chip.html
- ENCODE guidelines paper: http://genome.cshlp.org/content/22/9/1813.full?sid=bf1bf8f4-9ecf-4dd6-9761-78af1e542311
- Roadmap Epigenomics: http://www.roadmapepigenomics.org/
- ENCODE: https://www.encodeproject.org/
- modENCODE: http://www.modencode.org/
- UCSC epigenome browser: http://www.epigenomebrowser.org/