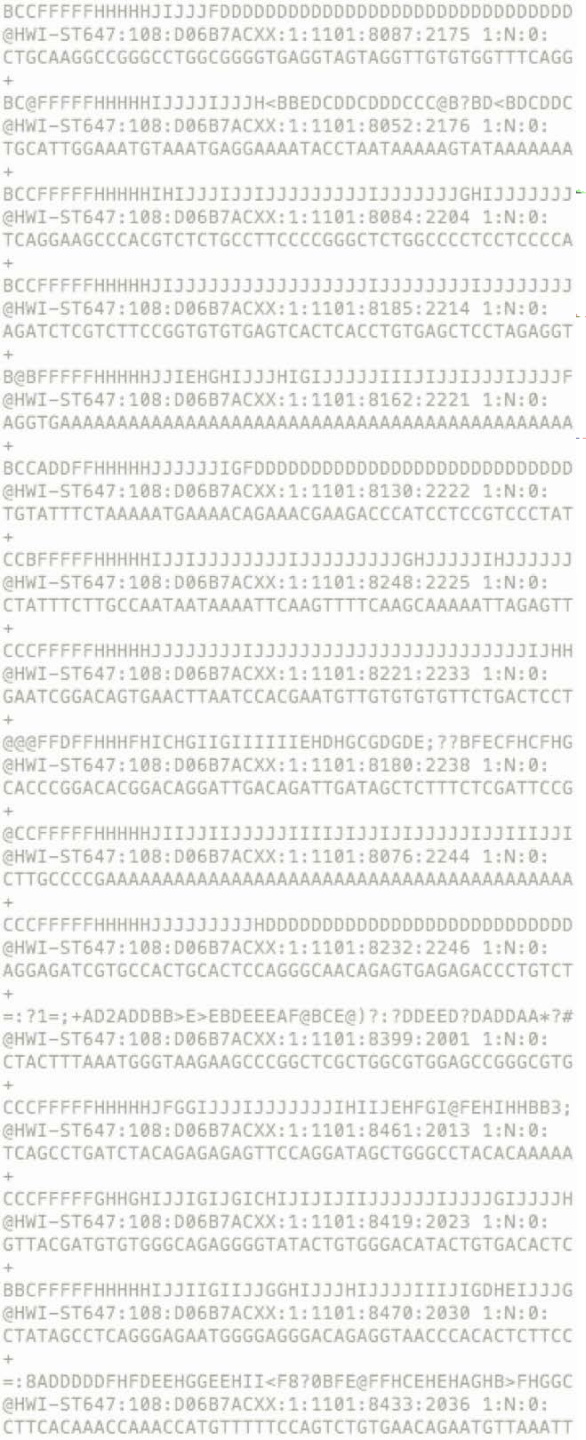
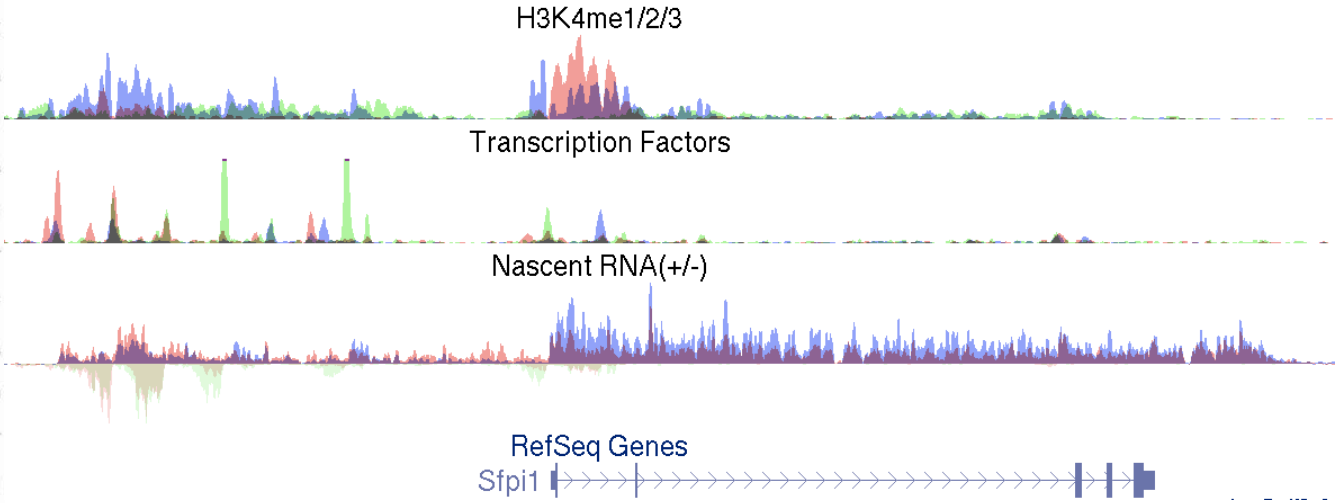


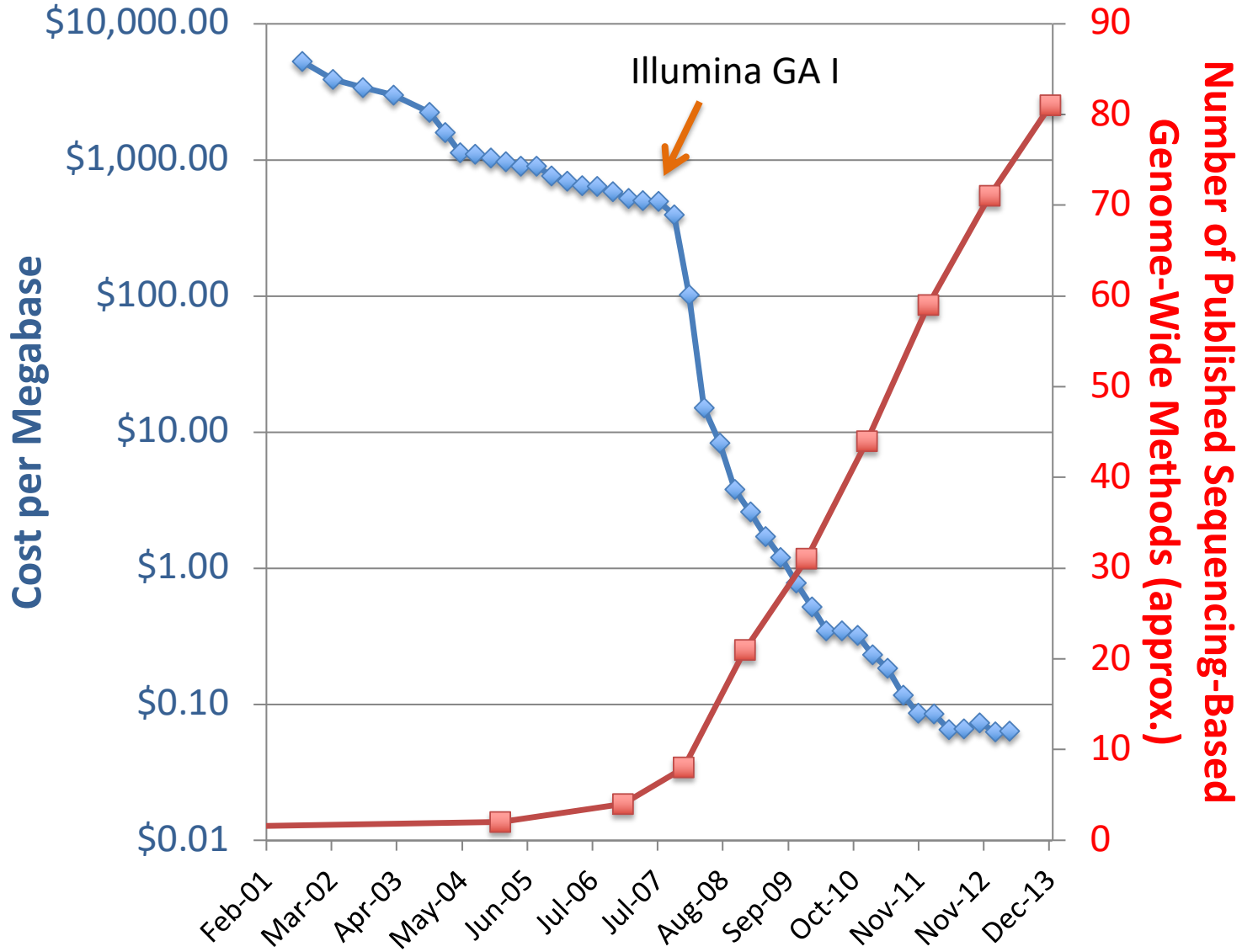


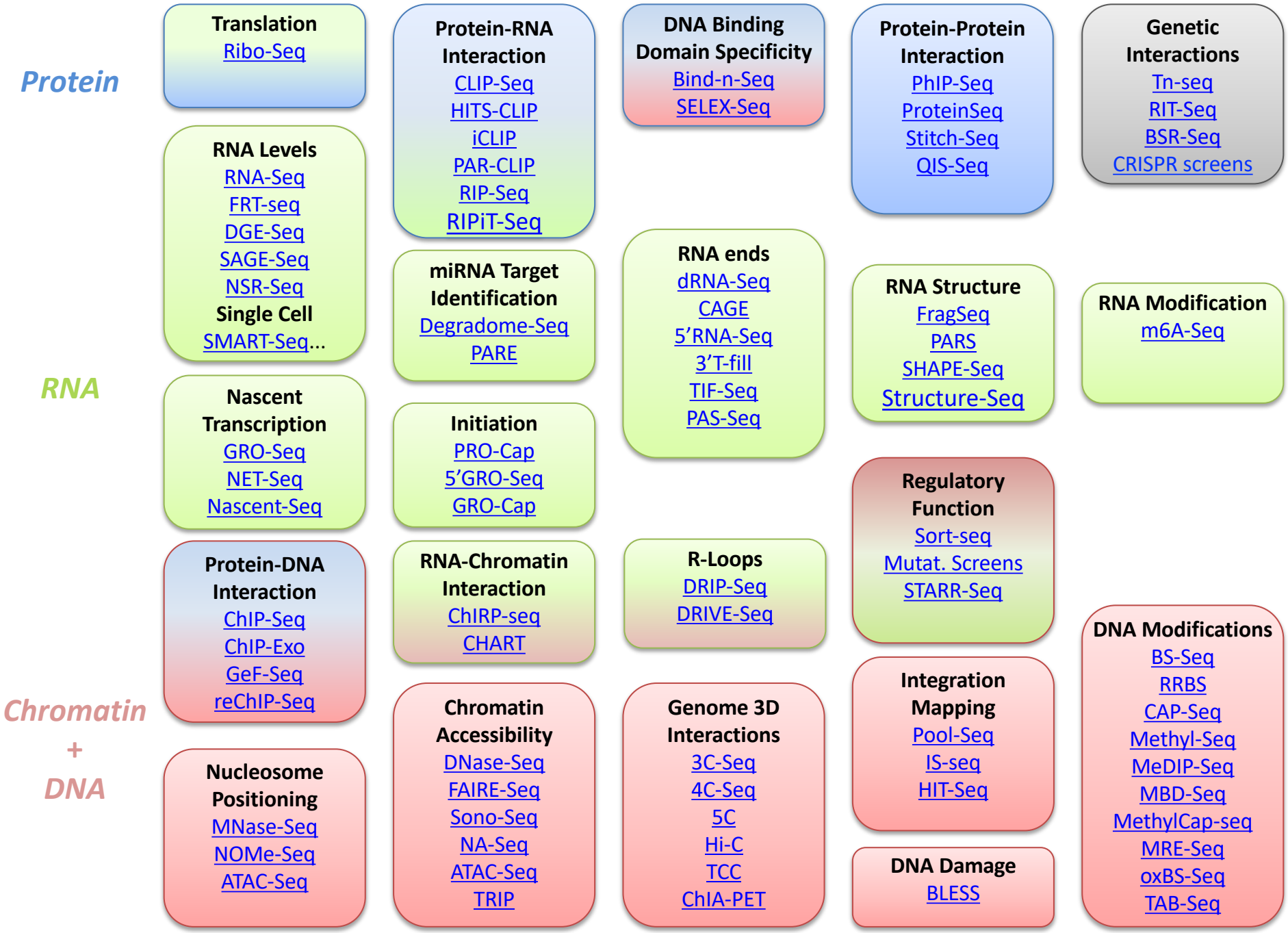
ChIP-Seq

Experimental design and analysis strategies

Sven Heinz, Ph.D.
Department of Medicine
Division of Endocrinology
UC San Diego

Genomics - The Sequencing Revolution




Protein
RNA
Chromatin
+
DNA
Translation
Ribo-Seq
RNA Levels
RNA-Seq
FRT-seq
DGE-Seq
SAGE-Seq
NSR-Seq
Single Cell
SMART-Seq...
Nascent Transcription
GRO-Seq
NET-Seq
Nascent-Seq
Protein-DNA Interaction
ChIP-Seq
ChIP-Exo
GeF-Seq
reChIP-Seq
Nucleosome Positioning
MNase-Seq
NOMe-Seq
ATAC-Seq
Protein-RNA Interaction
CLIP-Seq
HITS-CLIP
iCLIP
PAR-CLIP
RIP-Seq
RIPiT-Seq
miRNA Target Identification
Degradome-Seq
PARE
Initiation
PRO-Cap
5'GRO-Seq
GRO-Cap
RNA-Chromatin Interaction
ChIRP-seq
CHART
Chromatin Accessibility
DNase-Seq
FAIRE-Seq
Sono-Seq
NA-Seq
ATAC-Seq
TRIP
DNA Binding Domain Specificity
Bind-n-Seq
SELEX-Seq
RNA ends
dRNA-Seq
CAGE
5'RNA-Seq
3'T-fill
TIF-Seq
PAS-Seq
R-Loops
DRIP-Seq
DRIVE-Seq
Genome 3D Interactions
3C-Seq
4C-Seq
5C
Hi-C
TCC
ChIA-PET
Protein-Protein Interaction
PhIP-Seq
ProteinSeq
Stitch-Seq
QIS-Seq
RNA Structure
FragSeq
PARS
SHAPE-Seq
Structure-Seq
Regulatory Function
Sort-seq
Mutat. Screens
STARR-Seq
Integration Mapping
Pool-Seq
IS-seq
HIT-Seq
DNA Damage
BLESS
Genetic Interactions
Tn-seq
RIT-Seq
BSR-Seq
CRISPR screens
RNA Modification
m6A-Seq
DNA Modifications
BS-Seq
RRBS
CAP-Seq
Methyl-Seq
MeDIP-Seq
MBD-Seq
MethylCap-seq
MRE-Seq
oxBS-Seq
TAB-Seq

Sequencing-based methods

https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf

ChIP-Seq: **Ch**romatin **I**mmuno**p**recipitation coupled to High-throughput **Seq**uencing

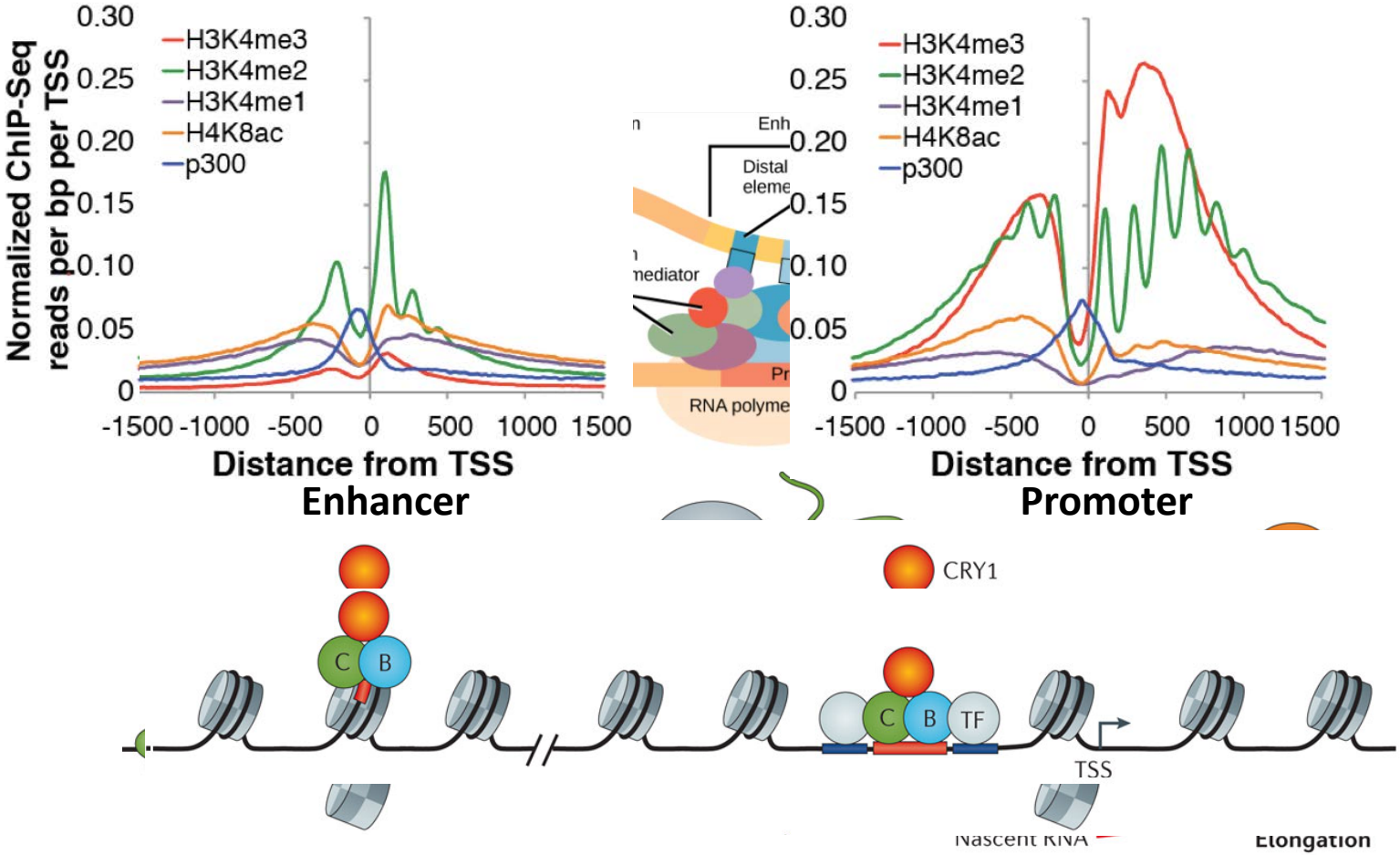
- Map the genomic location where specific proteins make contact with DNA *in vivo*.
  - Specific protein and its associated DNA are purified using an antibody.

- Next to RNA-Seq currently the most popular sequencing-based method.

Why ChIP-Seq?

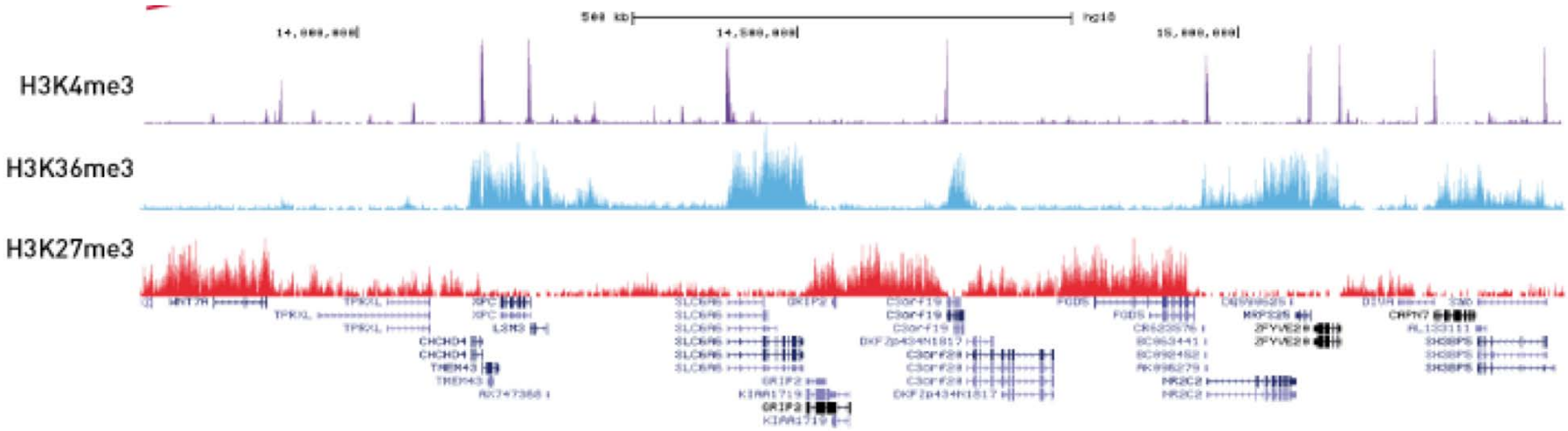
- Define active genomic regions in euchromatin: promoters, enhancers, gene bodies
- Define inactive regions: heterochromatin, repressed chromatin
- Track genome activity by chromatin changes
- Identify TF- and coregulator-bound sites
- Locate domain boundary factors

Why ChIP-Seq?



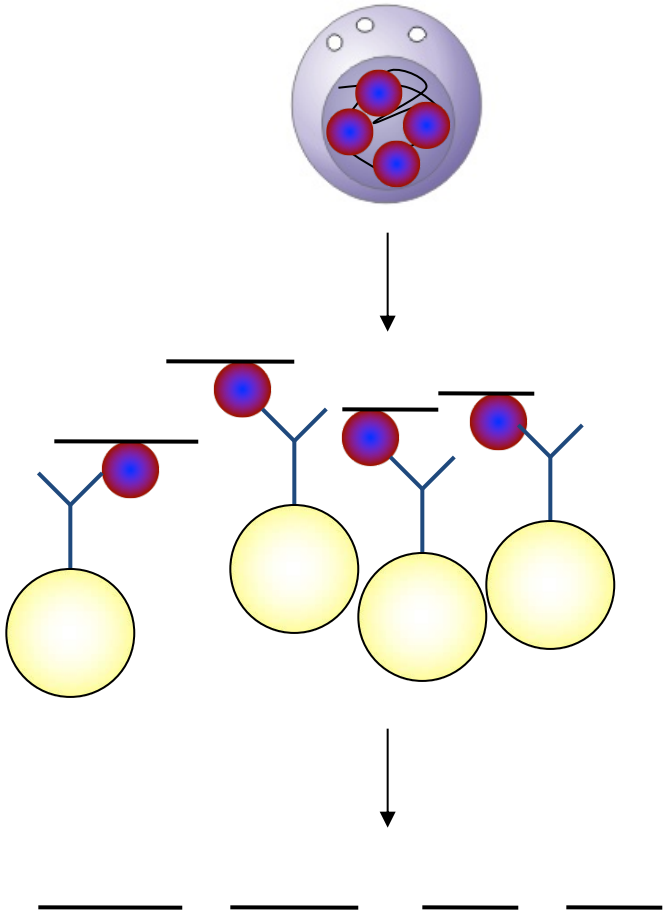
Takahashi Nat Rev Genet 2016

Histone modifications mark genome features



	Euchromatin					Hetero chromatin
	H3K4me1/2	H3K4me3	H3K27ac	H3K36me3	H3K27me3	H3K9me3
Promoter	+	+	active		repressed	inactive
Enhancer	+	-	active		repressed	inactive
Coding	-	-	-	+	repressed	inactive

Chromatin Immunoprecipitation (ChIP)

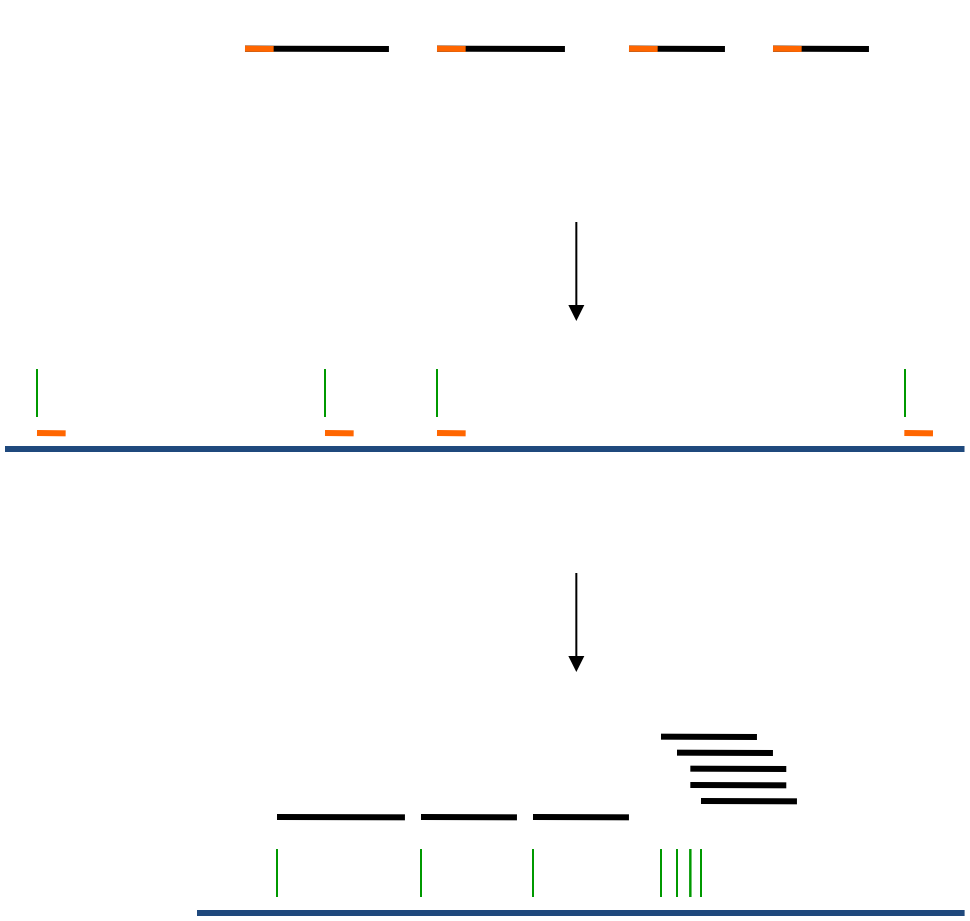


Formaldehyde-fix proteins ( ) to DNA, fragment DNA (sonication or MNase digest)

Immunoprecipitate

Isolate DNA

ChIP-seq



Sequence ChIPped DNA fragment ends (50-mer tags, up to 2 Billion tags/run)

Map tags to genome to obtain genomic positions

Visualize fragment positions

Downstream analysis

- Peak finding
- Motif finding
- …

Example Study: ChIP-Seq for the transcription factor PU.1 in Macrophages and B cells

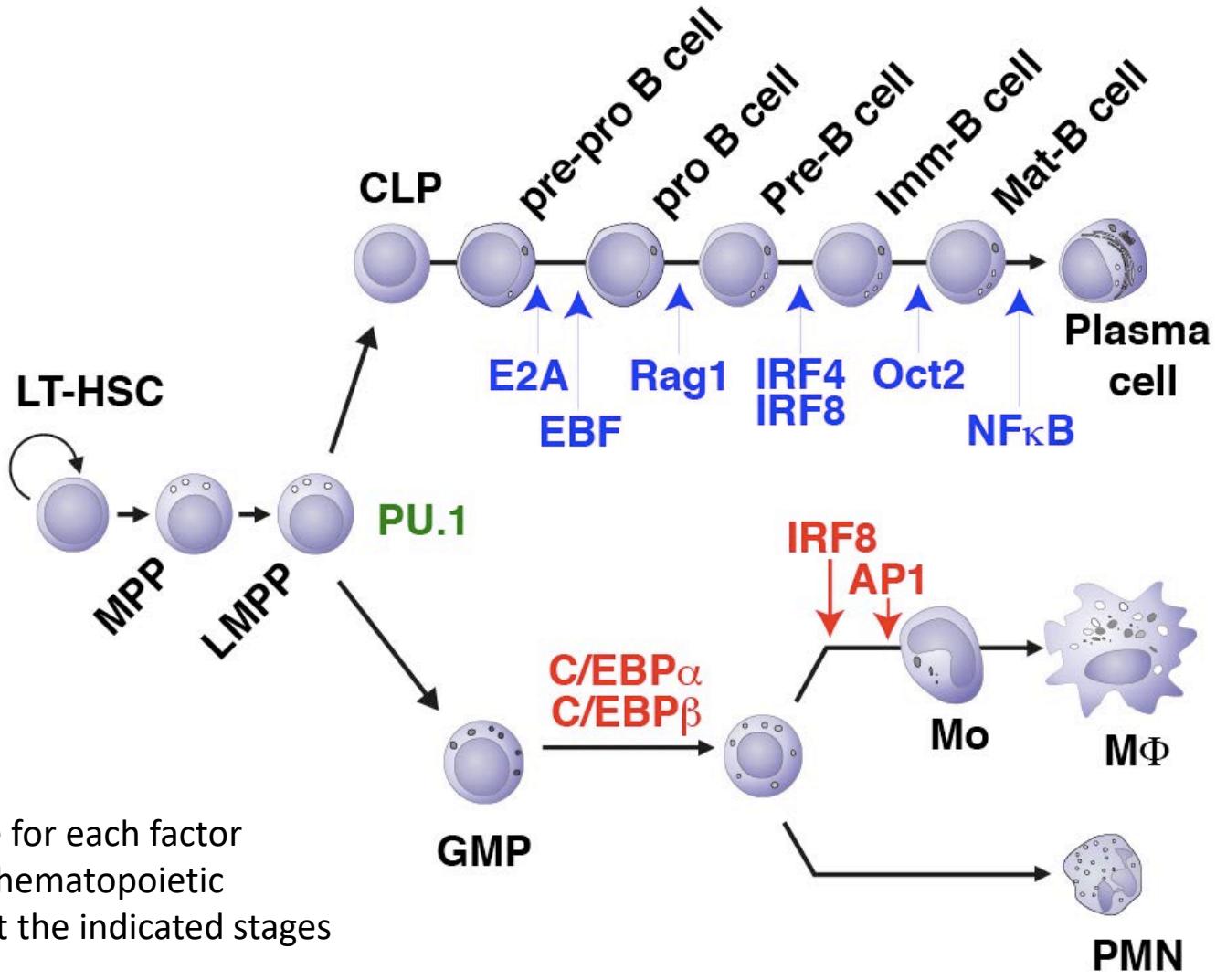
Chris Benner

Nathan Spann

How do transcription factors find their targets?

- Most transcription factors bind ~6-10 bp sequences *in vitro*
- There are 100k – 5 million predicted binding sites for different transcription factors in mammalian genomes
- Usually less than 5% of these sites are actually bound
- Factors bind to different sites in different cell types - why?

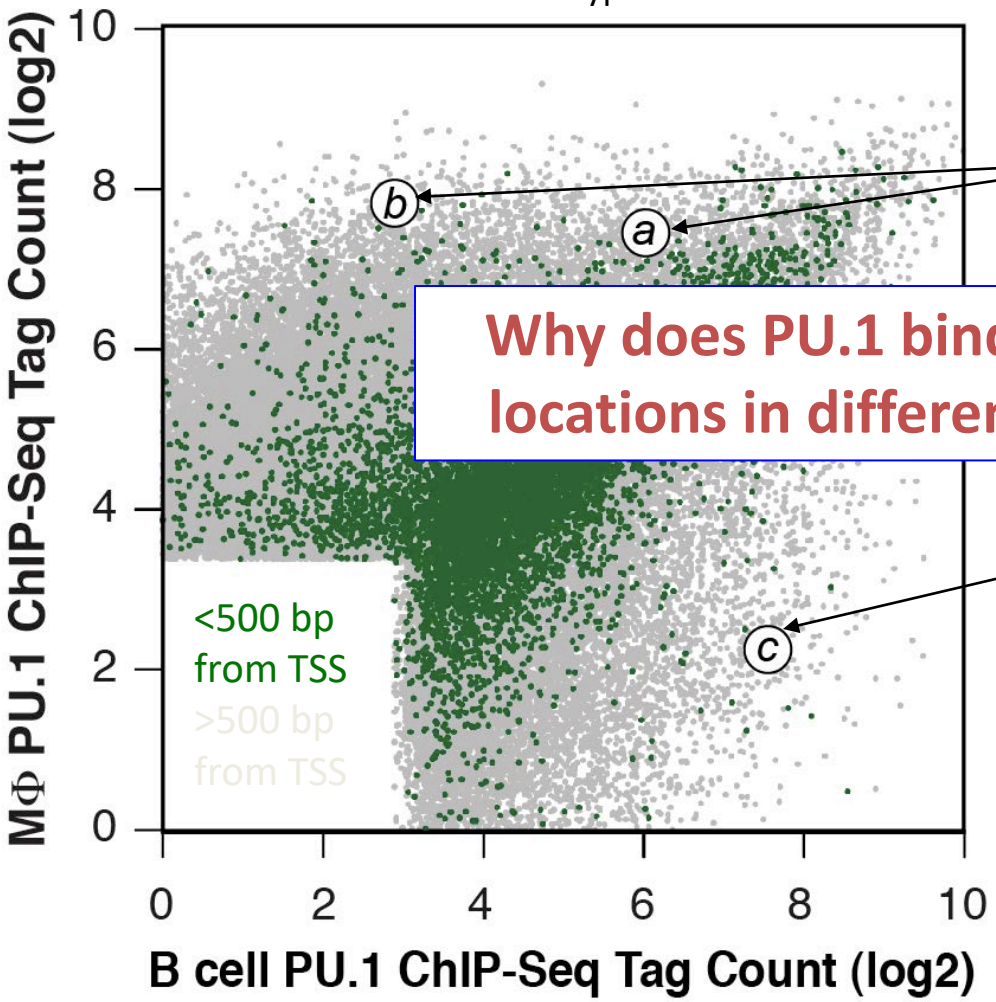
Genetic Evidence of Transcription Factors in Macrophage and B cell Development

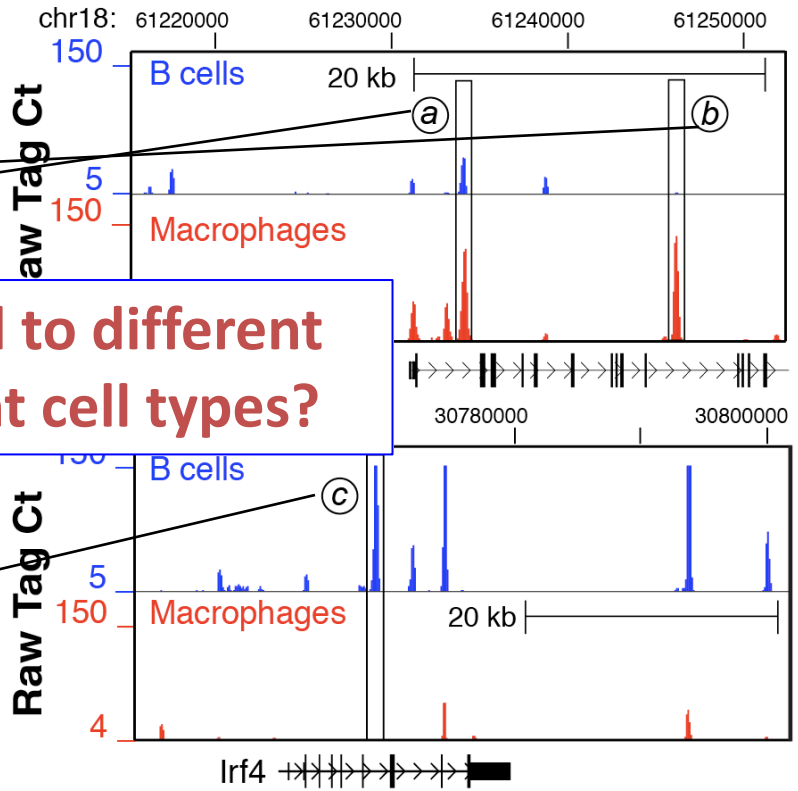


Knock-out mice for each factor arrests/affects hematopoietic development at the indicated stages

PU.1 cistromes differ greatly between macrophages and B cells

Each data point represents a PU.1 peak found in at least one of the two cell types





Why does PU.1 bind to different locations in different cell types?

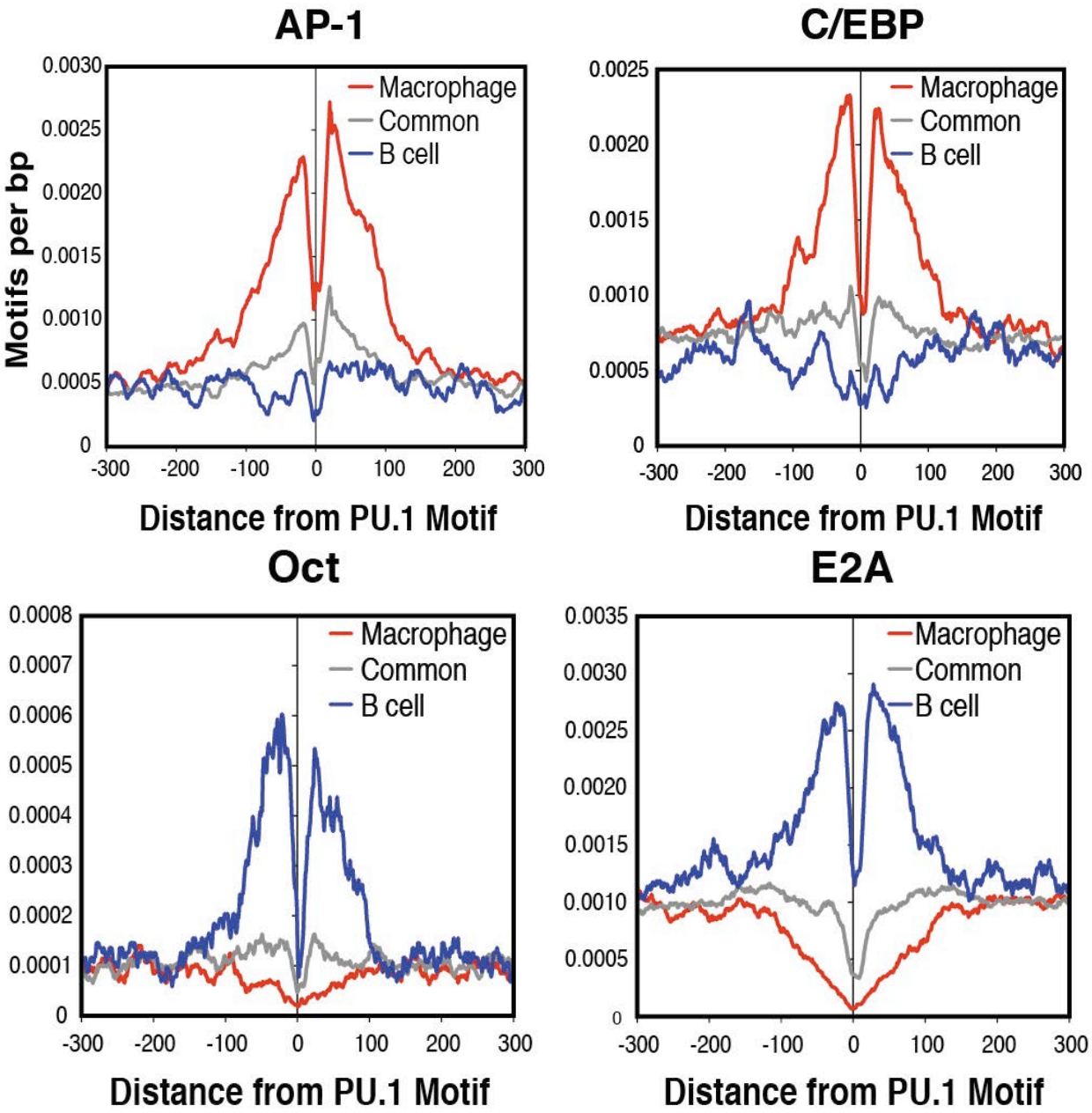
Cell-specific PU.1 binding sites are co-enriched for cell type-specific TF motifs

Macrophage-specific PU.1 binding sites

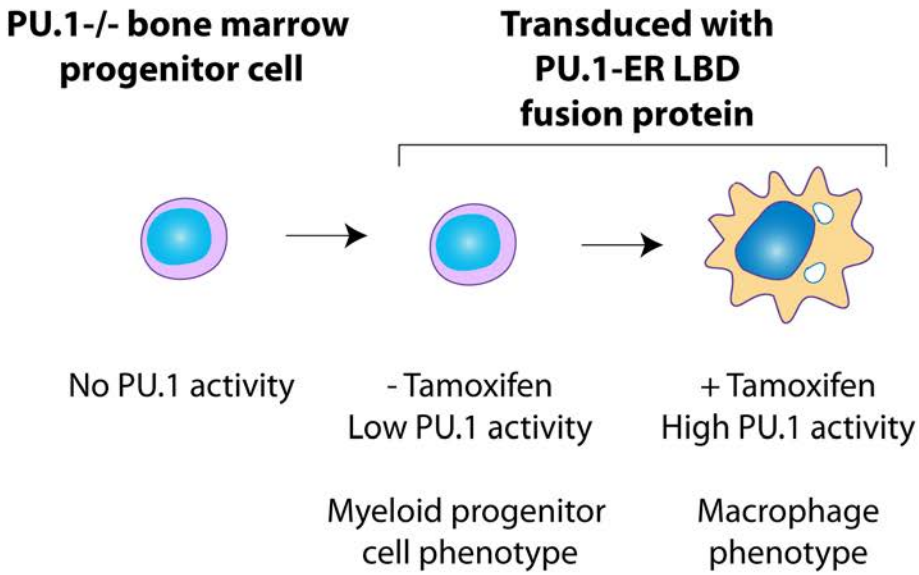
B-cell-specific PU.1 binding sites

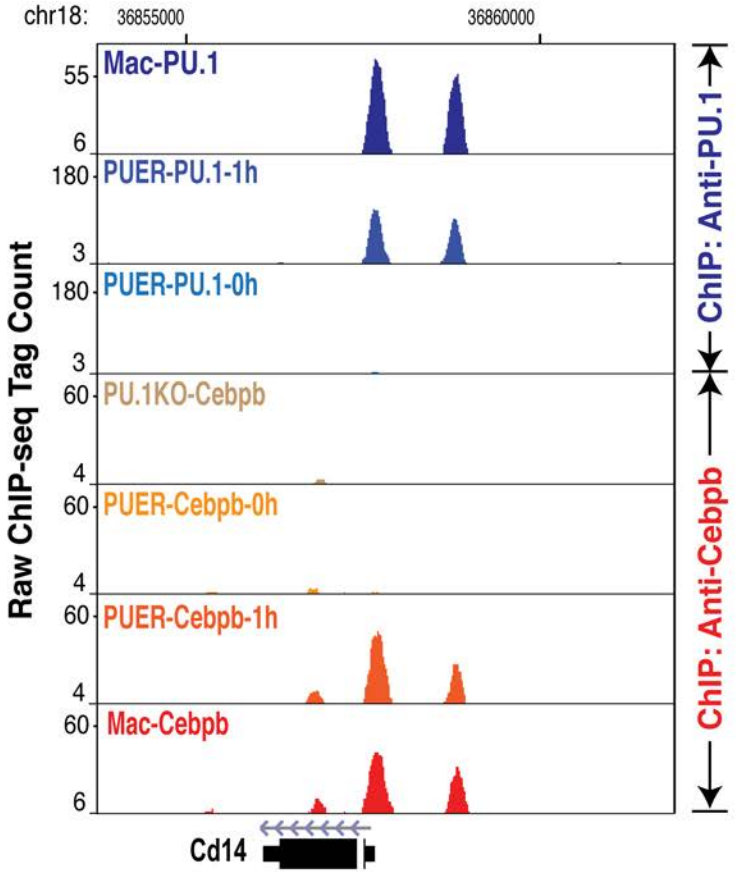
Chris Benner

Motif co-enrichment is found within ~ 100 bp of PU.1

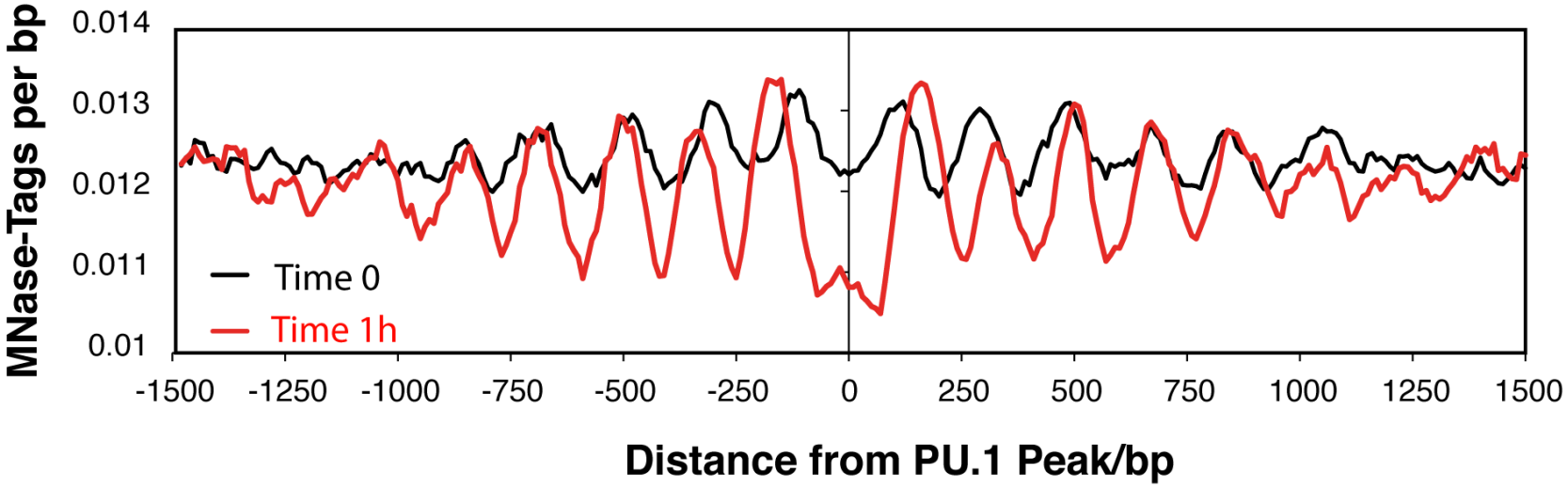


C/EBPβ binding is dependent upon PU.1 at many sites throughout the genome

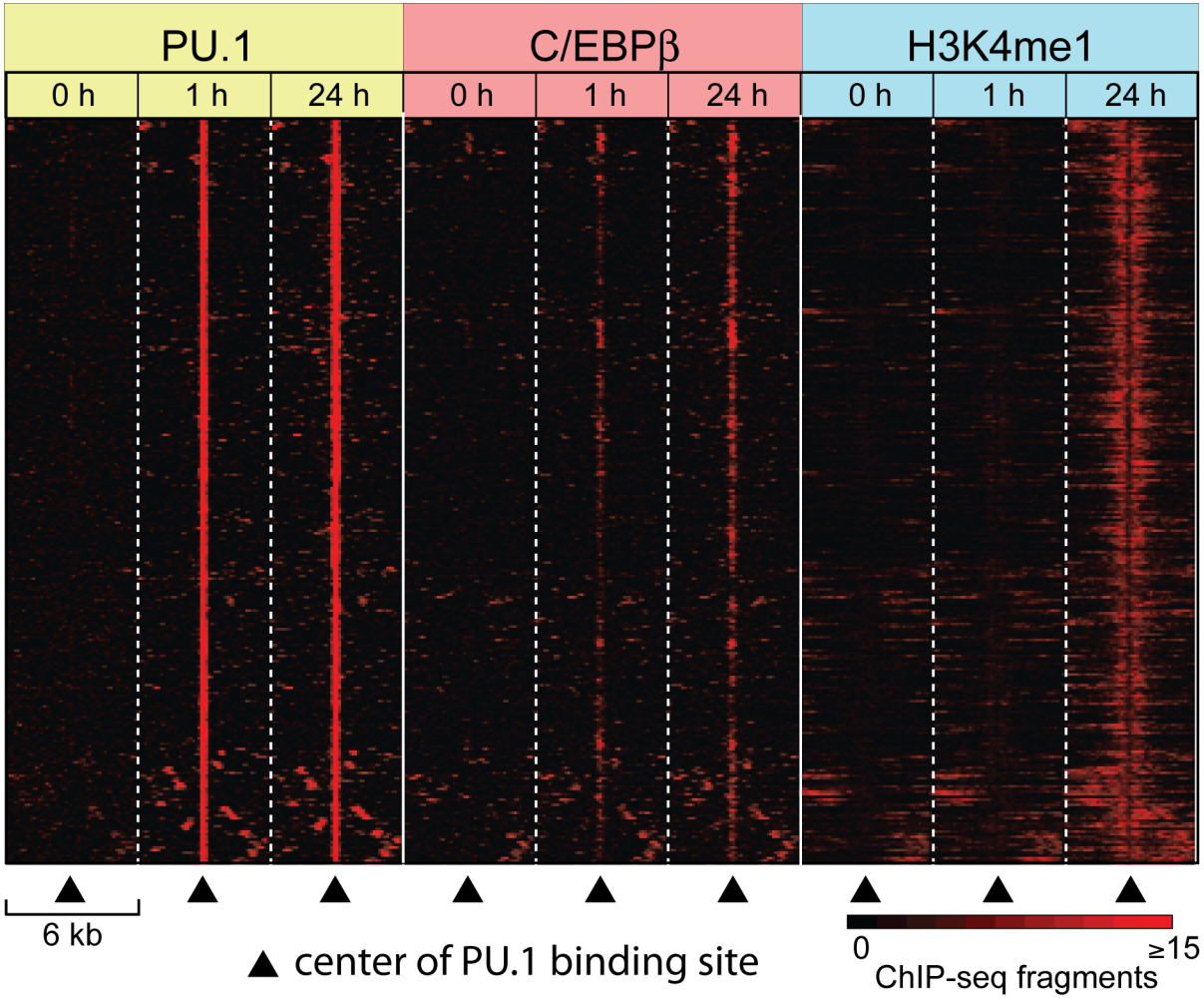




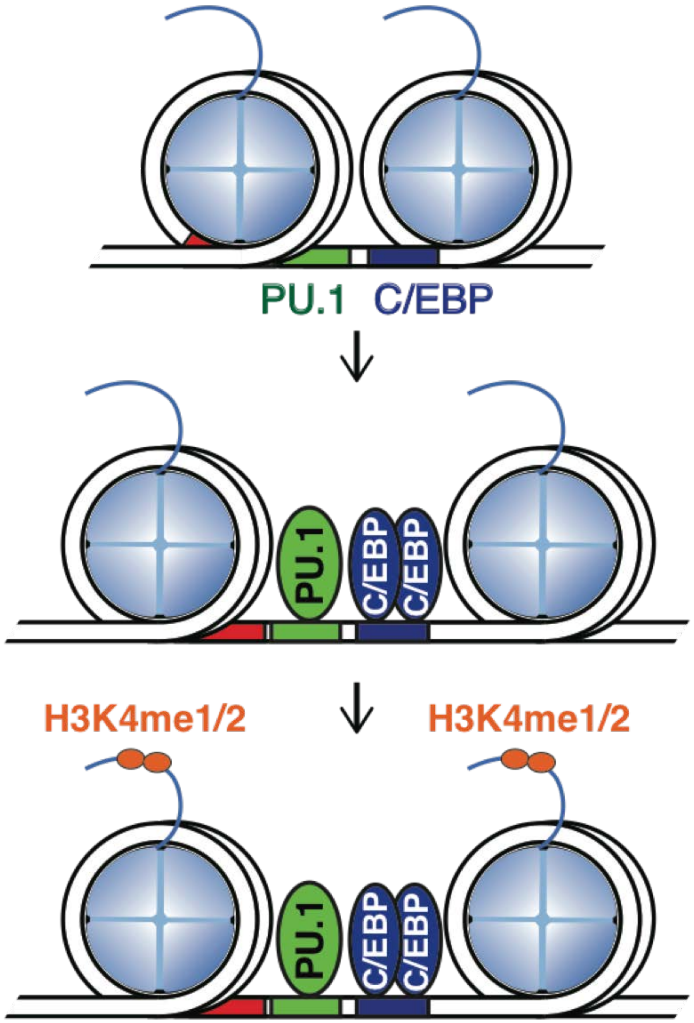
Induced PU.1 binding results in rapid nucleosome remodeling in PUER cells



PU.1 binding directs the ‘writing’ of the H3K4me1 enhancer mark

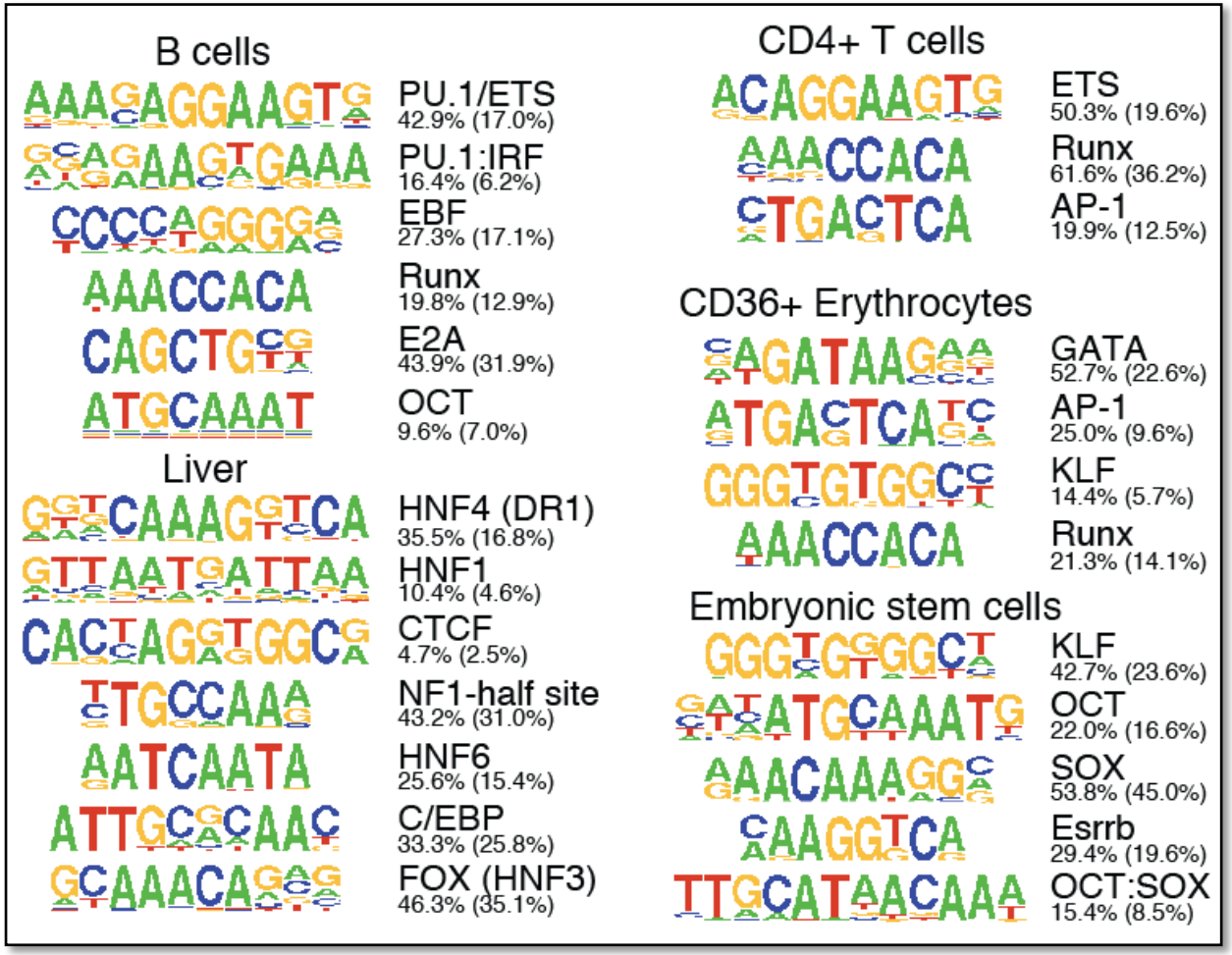


A collaborative model of lineage-determining transcription factor action



Heinz et al. Mol Cell 2010

Motifs enriched at H3K4me1-marked sites point to lineage-determining TFs



Similar motifs found with:

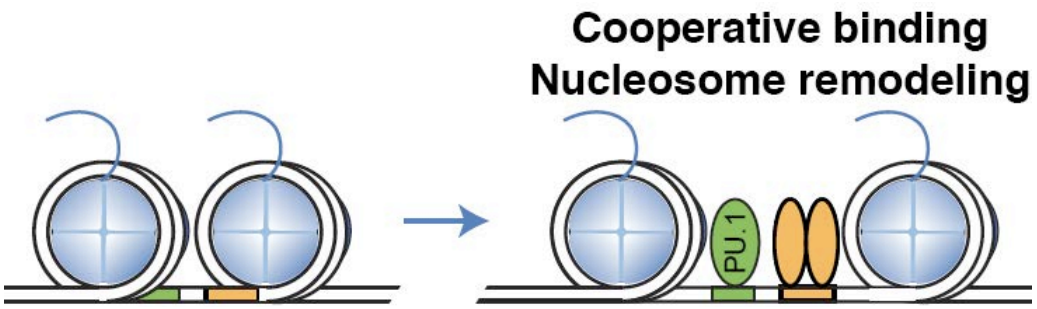
- H3K4me1/2
- H3K27ac, H3ac
- H4K8ac, H4ac
- DNase, FAIRE
- p300, co-activators
- NCoR, co-repressors
- RNA Pol II
- Any Activating Transcription Factor

What won't find these motifs:

- CTCF, NRSF
- H3K27me3, H3K9me3

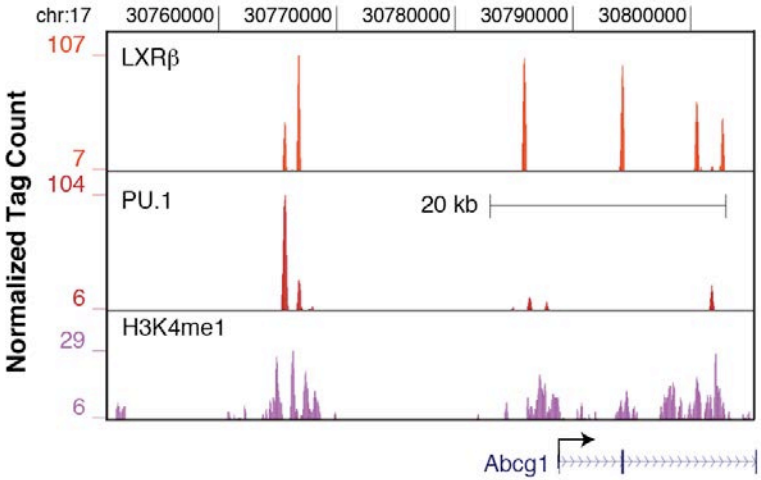
Lineage-defining transcription factors establish the cell enhancer repertoire

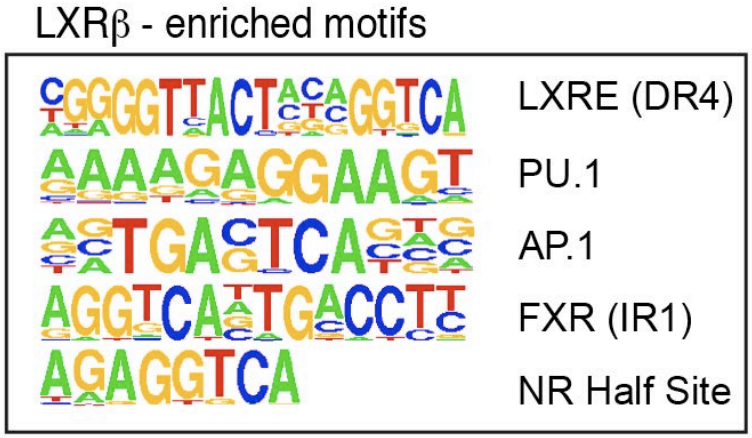
- Factors with motifs enriched at enhancers are:
  - Critical for the existence of that cell type
  - Typically the most highly expressed factors in that cell type
  - Can usually be used to reprogram that cell type
    - Oct4/Sox2/Klf4/Myc: Fibroblasts -> ES cells
    - PU.1+C/EBP: Fibroblasts -> Macrophages
    - C/EBP: Bcells -> Macrophages



Other Factors

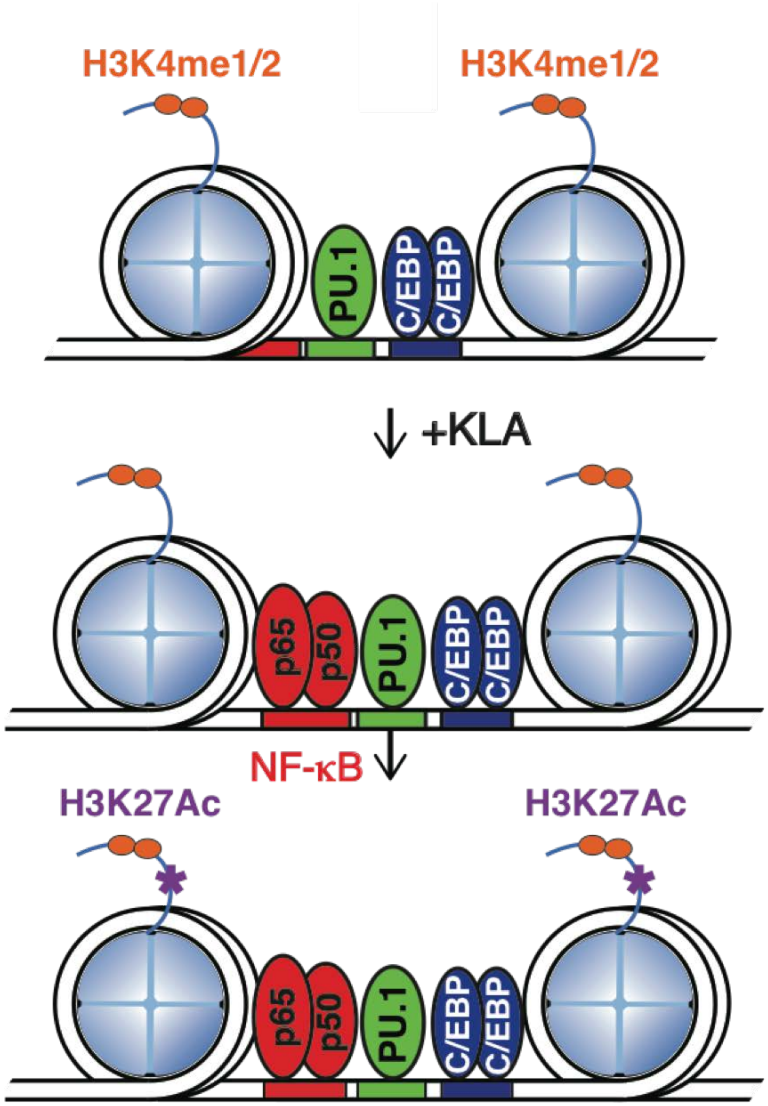
- Pretty much every activating transcription factor in macrophages is enriched for PU.1, AP-1, etc.
  - LXR, SRF, NF-κB (p65), Bcl6, PPARγ, etc.





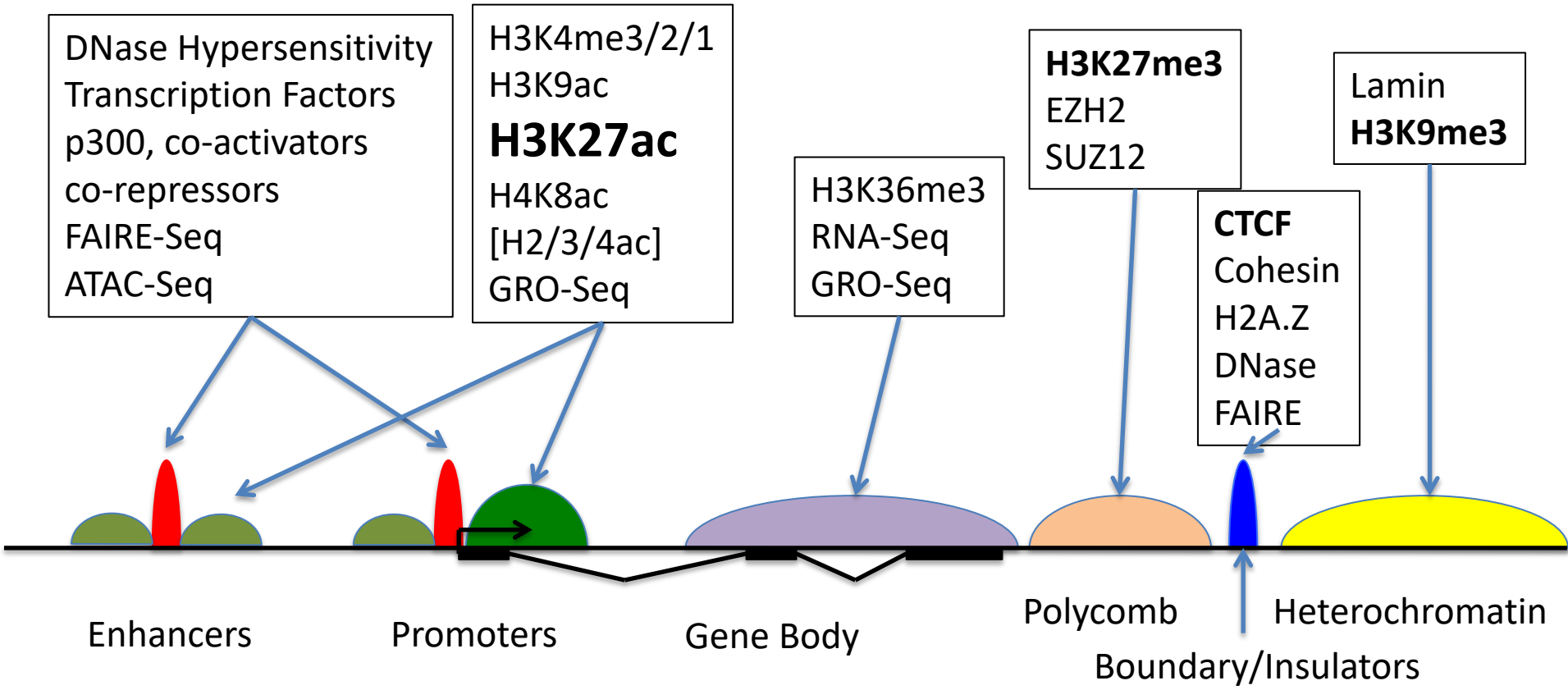
- Similar results for other cell types:
  - How many papers doing ChIP-Seq in ES cells did NOT show motif enrichment for the OCT4 and/or SOX2 motif?

A hierarchy of TF action



Heinz et al. Nature 2013

Planning an experiment: Types of Regions to Interrogate



Planning ChIP-Seq Experiments

- Purpose of the Study
  - Identify transcription factor targets
  - Identify enhancers (p300, H3K4me1/2, H3/H4ac, DNase, ATAC-seq)
  - Identify lineage-determining transcription factors
  - Check Pol II status in transcription
  - Find regions of repressive chromatin
  - Study genome-wide effects of a knock-out/siRNA knock-down

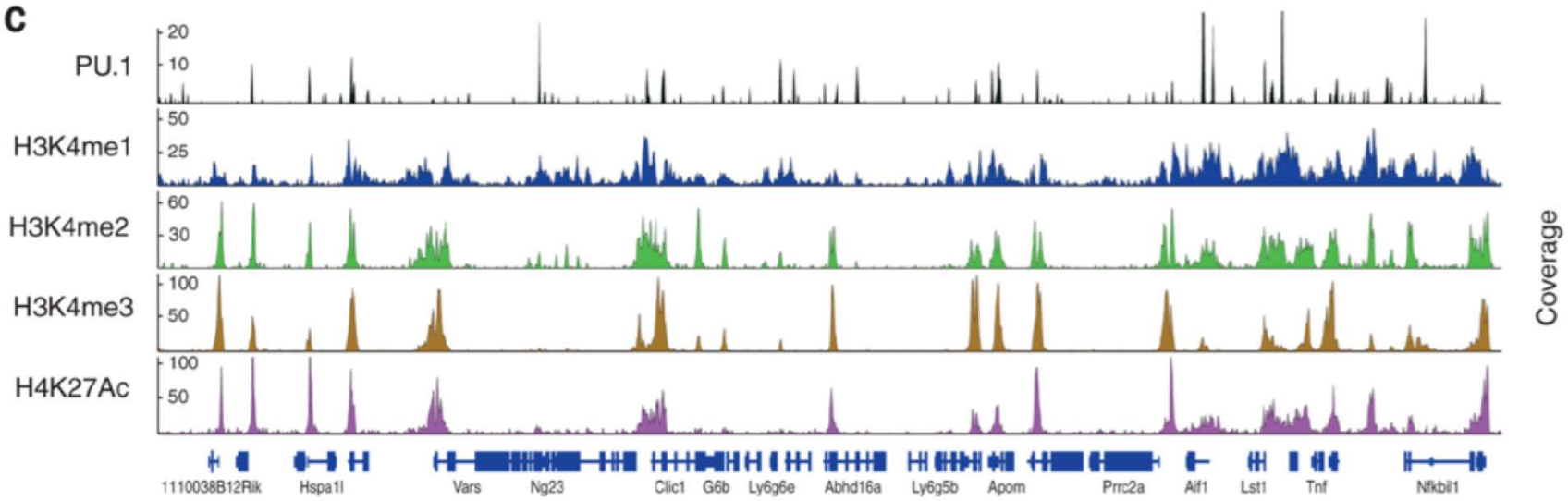
- Choosing the right thing to ChIP
  - Transcription factor? Is it expressed, is it present in the nucleus
  - Histone modifications? p300?
  - Total Pol II vs CTD phosphorylation

What do you need for ChIP-Seq?

- **~0.1 - 150 million cells*** (depending on protocol and antibody quality)
  - Works with just about any organism and cell type, some solid tissues can be hard to work with
- **Most important: ChIP-grade antibody**
  - Must be able to recognize its epitope under formaldehyde-crosslinked conditions. (Antibodies that work in Western blot may not be adequate for ChIP; IHC-P is a good bet)

What do you need for ChIP-Seq?

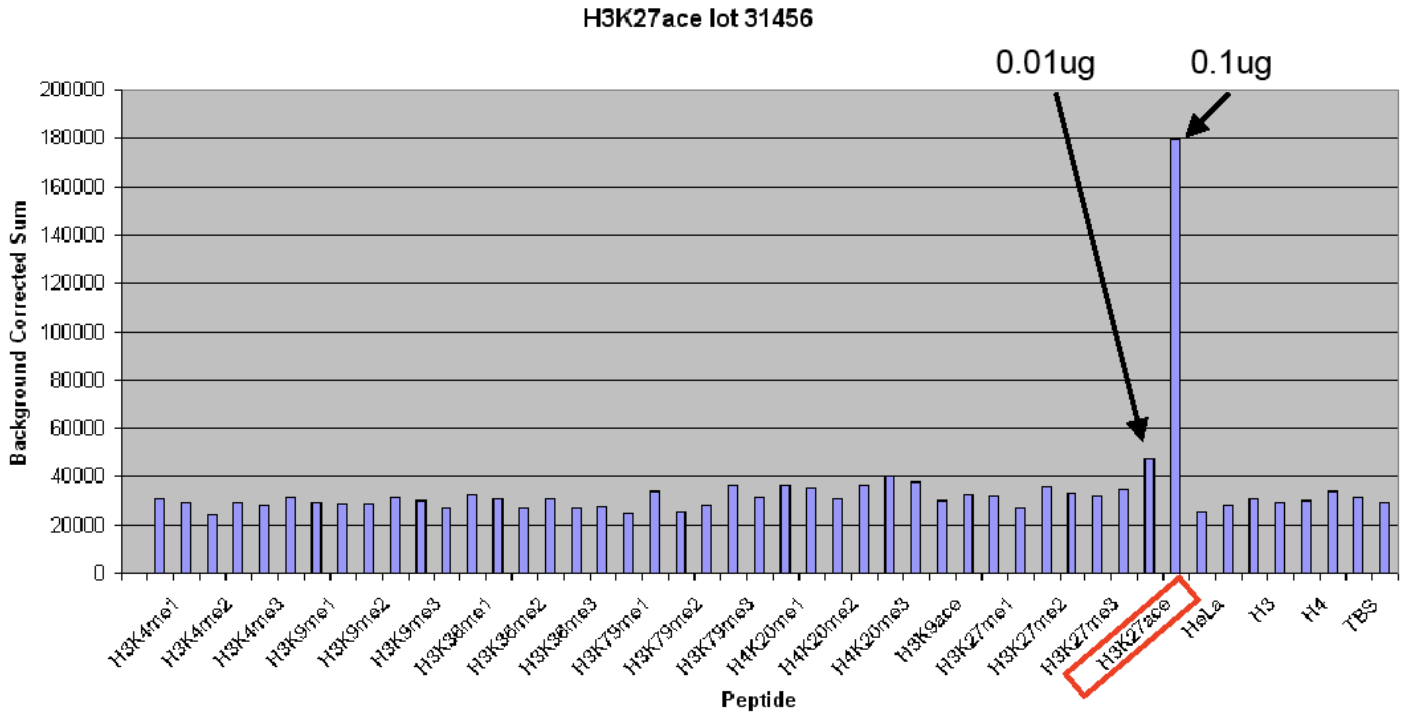
- **<u>ChIP-grade antibody</u>**
  - Easy test: compare IP and ChIP (fixed) in Western blot
  - Most histone modifications work well, but most are highly correlated…
    - Unless you study histone marks, stick to well-working and well-accepted marks: H3K4me1/2/3, H3K27ac/me3, H3K36me3; H3K9me3



Lara-Astiaso, Science 2014

Choosing the right antibody

- Good idea to validate specificity
  - Knock-out or knock-down with siRNA, then ChIP
  - Peptide arrays for histone modification
- Better: Vendor did it already



From Encode

Crosslinking

- **Standard:** 1% formaldehyde crosslinks proteins to DNA, but due to its short linker length: might not stabilize protein-protein interactions well

- **Alterative: Double-crosslinking**, first with protein-protein crosslinker, then with formaldehyde improves ChIP for some targets (e.g. p65/NF-κB; Nowak et al. Biotechniques 2005), best with cell suspension.

Protocol considerations

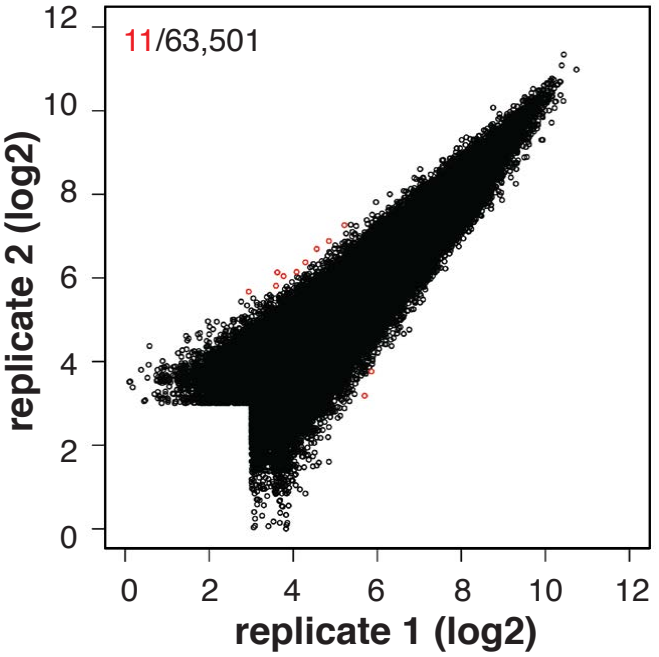
- Many protocols work, and often equally well (if the antibody is “good”)
- Biggest difference: detergent in the lysis buffer
  - Ionic detergents need to be diluted an/or sequestered by non-ionic detergents for the antibody to work. Non-ionic detergents don’t strip proteins off the chromatin, require more sonication
  - RIPA (1% Triton X-100, 0.1% SDS, 0.1% deoxycholate) works better for “weak” antibodies, and specific cell types, but less convenient for sonication and potentially higher background
- Too little sonication (especially in RIPA) leaves little product in the size range (100-400 bp) required for sequencing -> increased duplication rate!
- Too stringent washes (> 300-500 mM NaCl) decrease signal to noise ratio
- LiCl is less disruptive than NaCl (for magnetic beads, 250 mM LiCl often sufficient)
- Column or bead cleanup has less open chromatin bias than phenol/chloroform (see FAIRE!)
- ChIP-Seq protocol suggestions:

  http://www.ncbi.nlm.nih.gov/pubmed/19275939 Schmidt Methods 2009
  http://www.ncbi.nlm.nih.gov/pubmed/20513432 Heinz Mol Cell 2010
  http://www.ncbi.nlm.nih.gov/pubmed/22940246 Garber Mol Cell 2012
  http://www.ncbi.nlm.nih.gov/pubmed/23171294 Gilfillan BMC Genomics 2012

Planning ChIP-Seq Experiments

- Controls, Input need to be performed
  - Min number inputs = cell*organism*protocol*library-prep
- Replicates
  - Technical replicates probably not needed
  - Biological replicates – YES!
    - The more quantitative the study, the greater the need for replication.

NF-κB ChIP-Seq, biological replicates, mouse macrophages



Sequencing ChIP-Seq Experiments

Study Type	Optimal Sequencing Strategy
Normal ChIP-Seq	Single ended, short reads
Allele-specific ChIP-Seq	Paired end, **long** reads

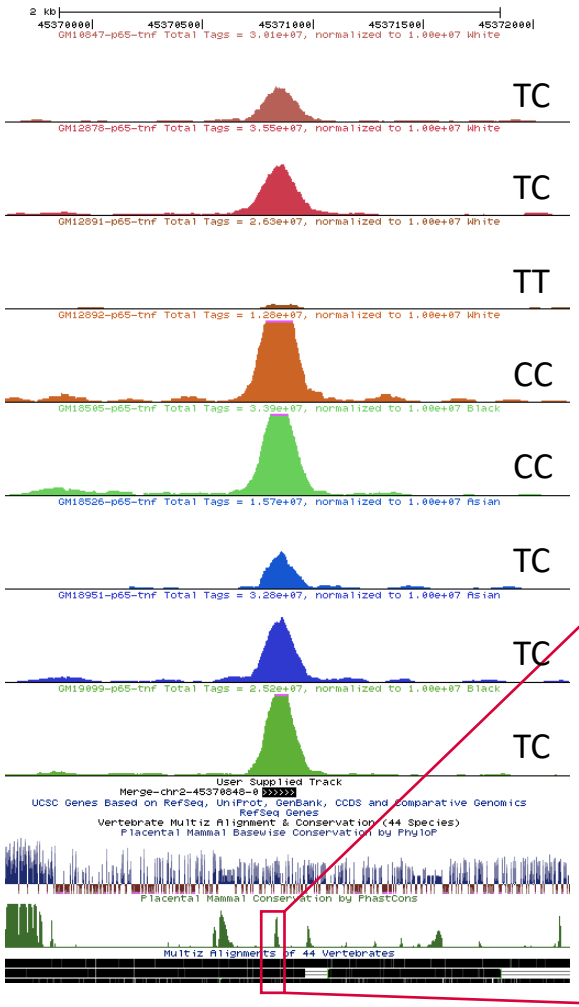
Short reads – more bang for the buck

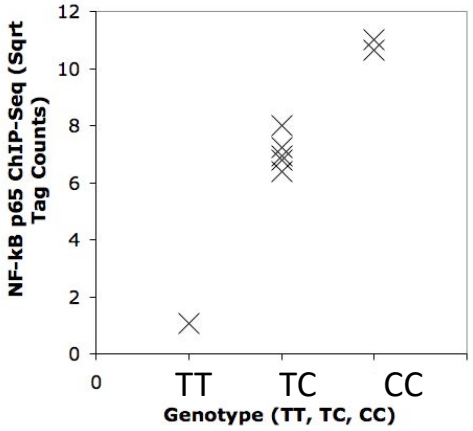
- More quantification per bp - once a read can be placed unambiguously in the genome, additional nucleotides are redundant information

Long/paired-end reads - more likely to cover sequence variants (i.e. SNPs)

- Key to identifying allele-specific binding or for genetic studies - if a read doesn't contain a sequence variant, it can't be used to distinguish alleles.

Characterize non-coding SNPs





Reference Base: C

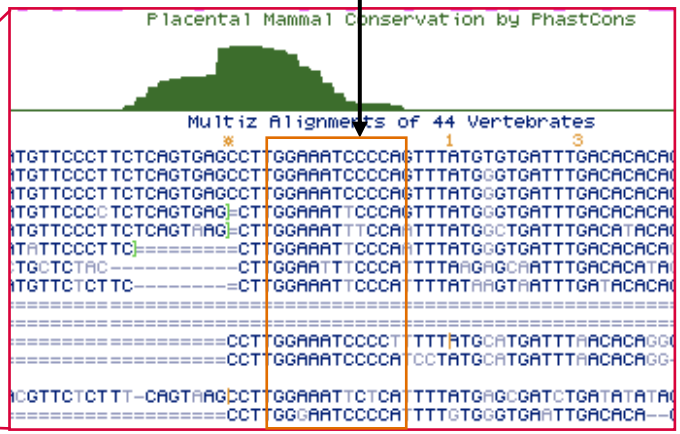
SNP: T

Correlation: -0.97

p-value: 6.18e-5

C: GGAAATC**C**CC

T: GGAAATC**T**CC



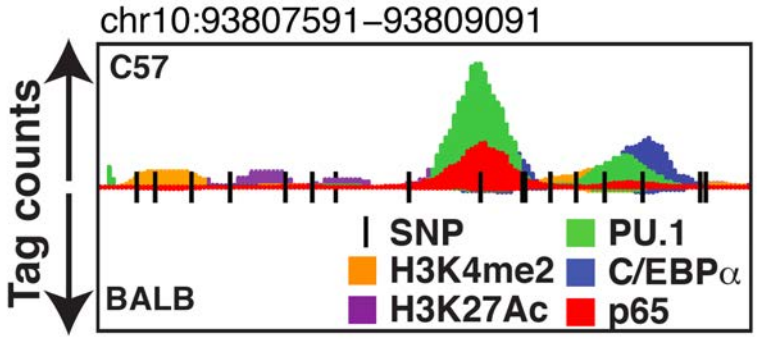
Kasowski 2010

Prioritize non-coding SNPs

- QTL (expression quantitative trait loci) are genomic regions containing sequence variants with influence on expression levels of one or more genes
  - *cis*-QTL: affects gene on the same chromosome
  - *trans*-QTL: affects genes on other chromosome

  - Expression: eQTL
  - TF binding: bQTL
  - Histone mark: hQTL

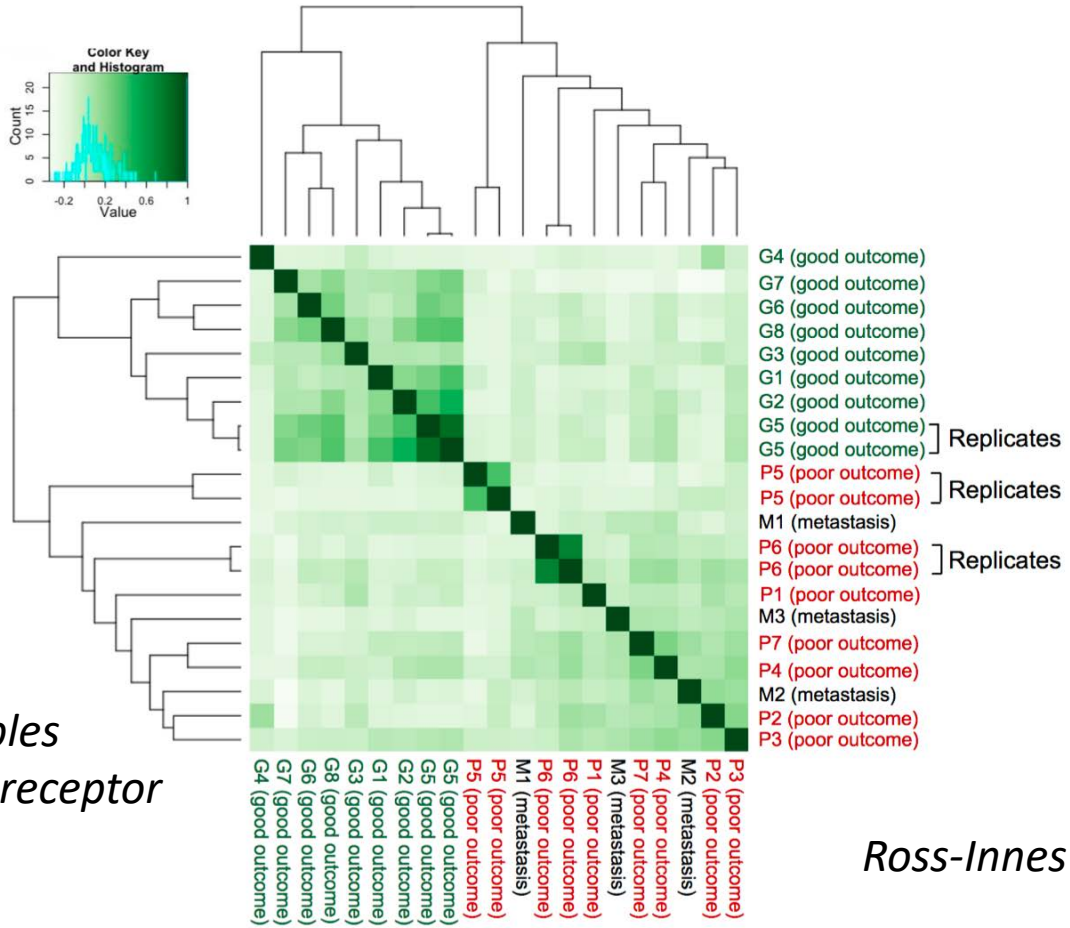
Pinpointing causal SNPs by ChIP-Seq combined with motif analysis



Heinz and Romanoski et al. Nature 2013

Potential clinical applications of NGS-based epigenome profiling

ChIP-Seq distinguishes patient outcomes



Breast cancer samples
ChIP-seq: estrogen receptor
Pearson r matrix

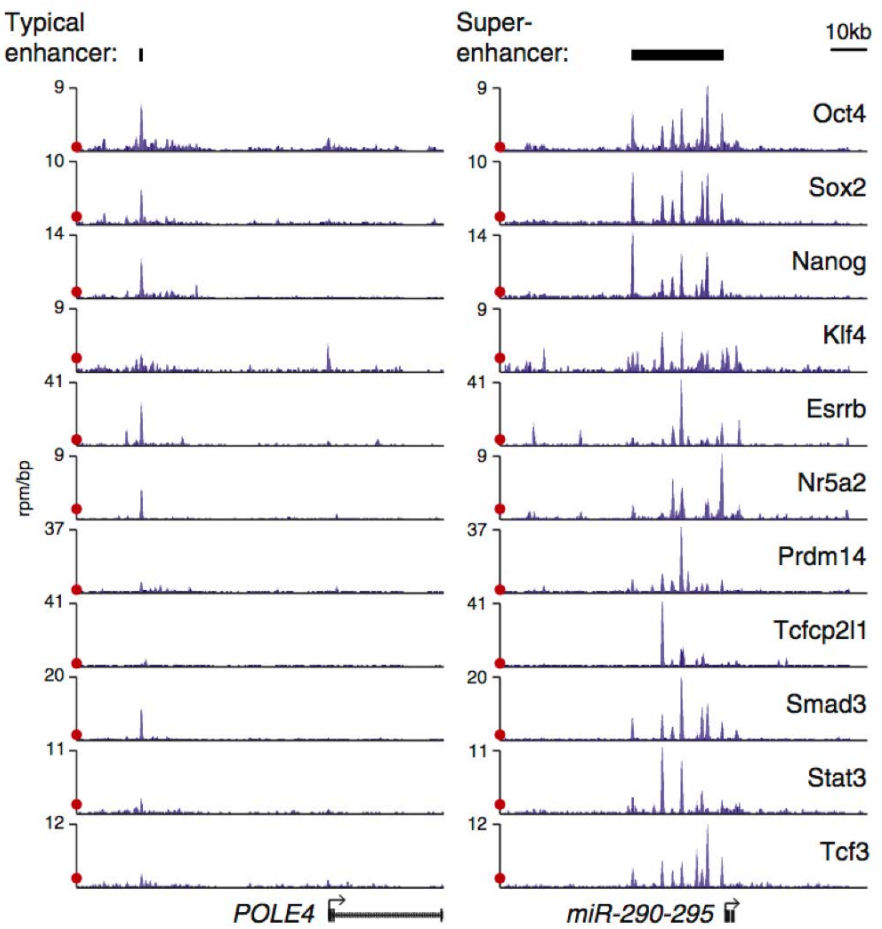
Ross-Innes Nature 2012

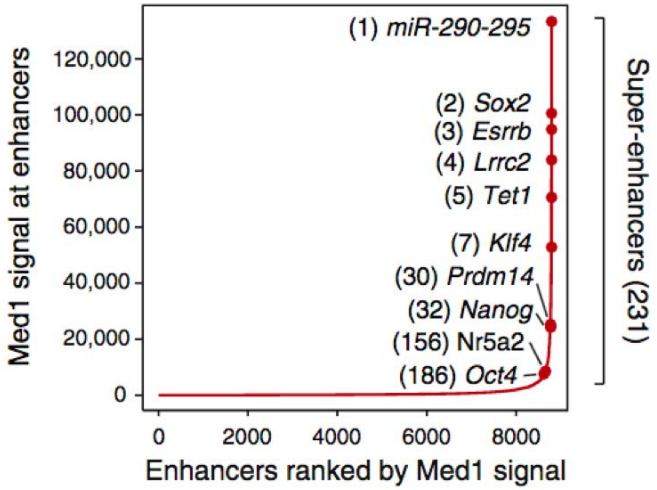
Potential clinical applications of NGS-based epigenome profiling

Super-enhancers mark gene loci that are important for cell identity

- ChIP-Seq of genome activity markers (Med1, H3K27ac)
- Identify enhancer-rich regions (super-enhancers)

Super-enhancers



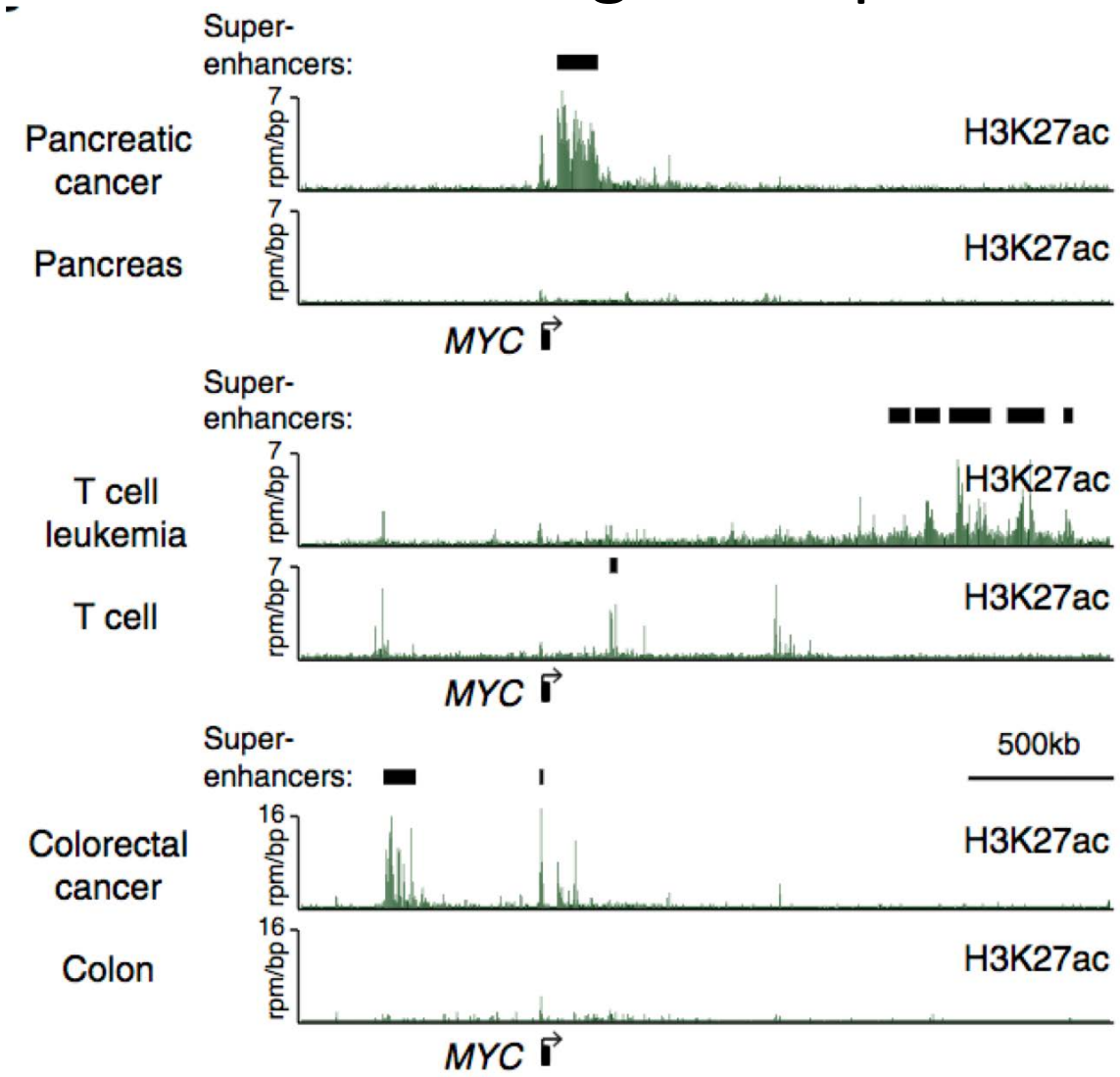


Transcription factor	Motif	P-value
Oct4		$9.19*10^{-64}$
Sox2		$3.01*10^{-67}$
Nanog		$9.46*10^{-17}$
Klf4		$4.33*10^{-6}$
Esrrb		$2.55*10^{-84}$
Nr5a2	n.a.	n.a.

Transcription factor	Motif	P-value
Prdm14	n.a.	n.a.
Tcfcp2l1		$6.83*10^{-11}$
Smad3		$9.31*10^{-11}$
Stat3		$2.90*10^{-10}$
Tcf3		$5.46*10^{-27}$

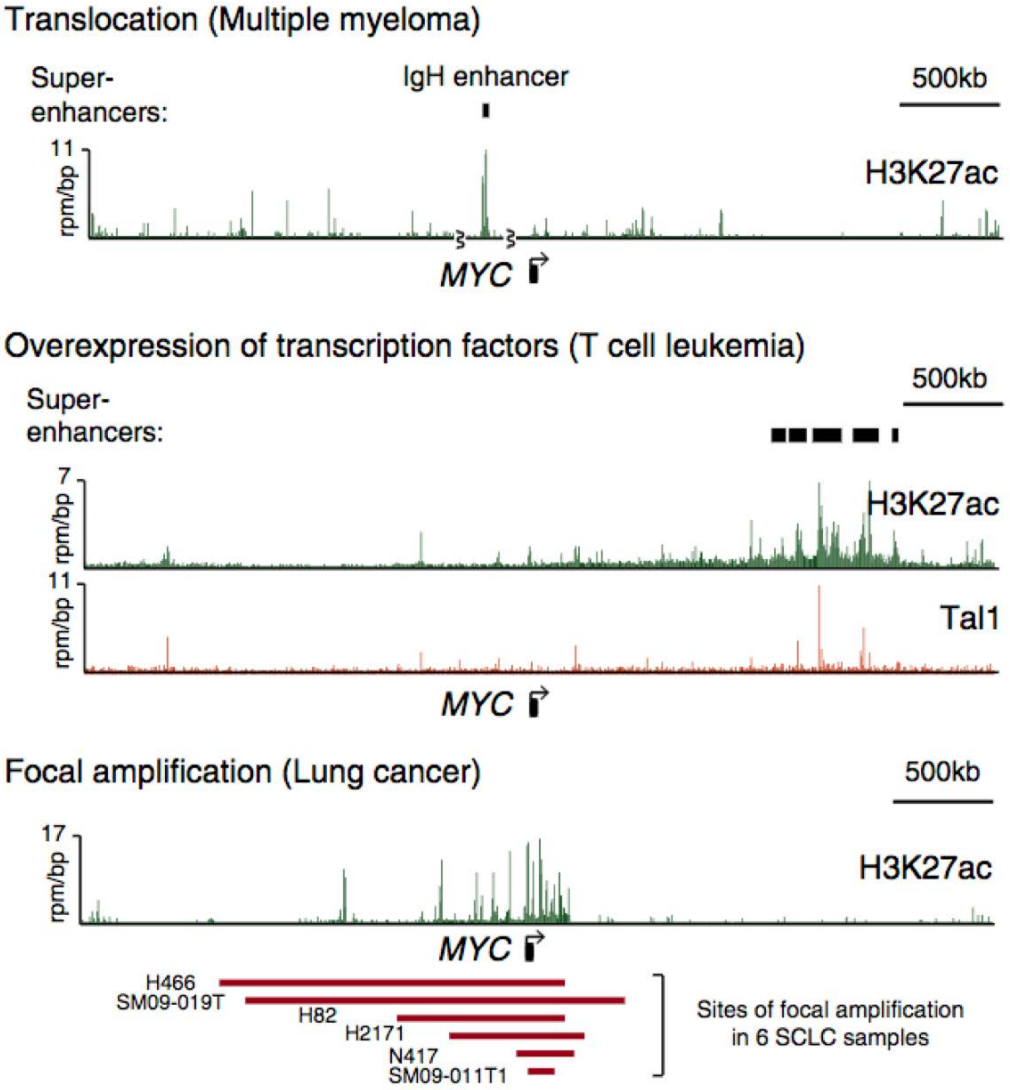
Hnisz et al. Cell 2013

Cancer cells activate or acquire super-enhancers that drive oncogene expression



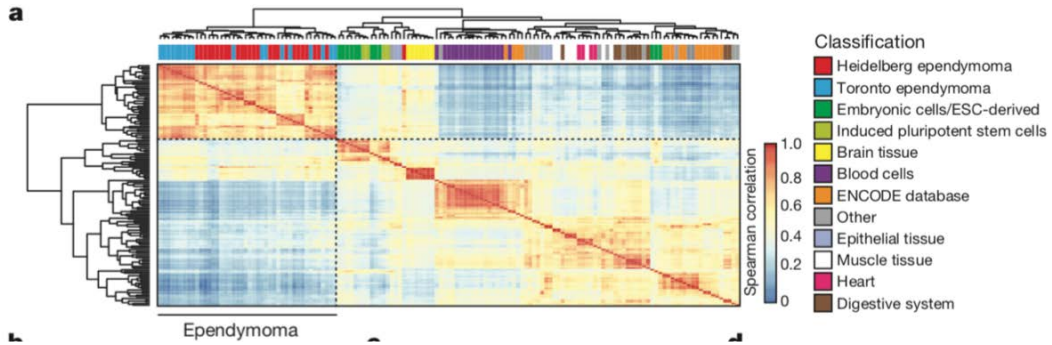
Hnisz et al. Cell 2013

Cancer cells activate or acquire super-enhancers that drive oncogene expression



Hnisz et al. Cell 2013

H3K27ac pattern identifies ependymoma subtypes and possible targets

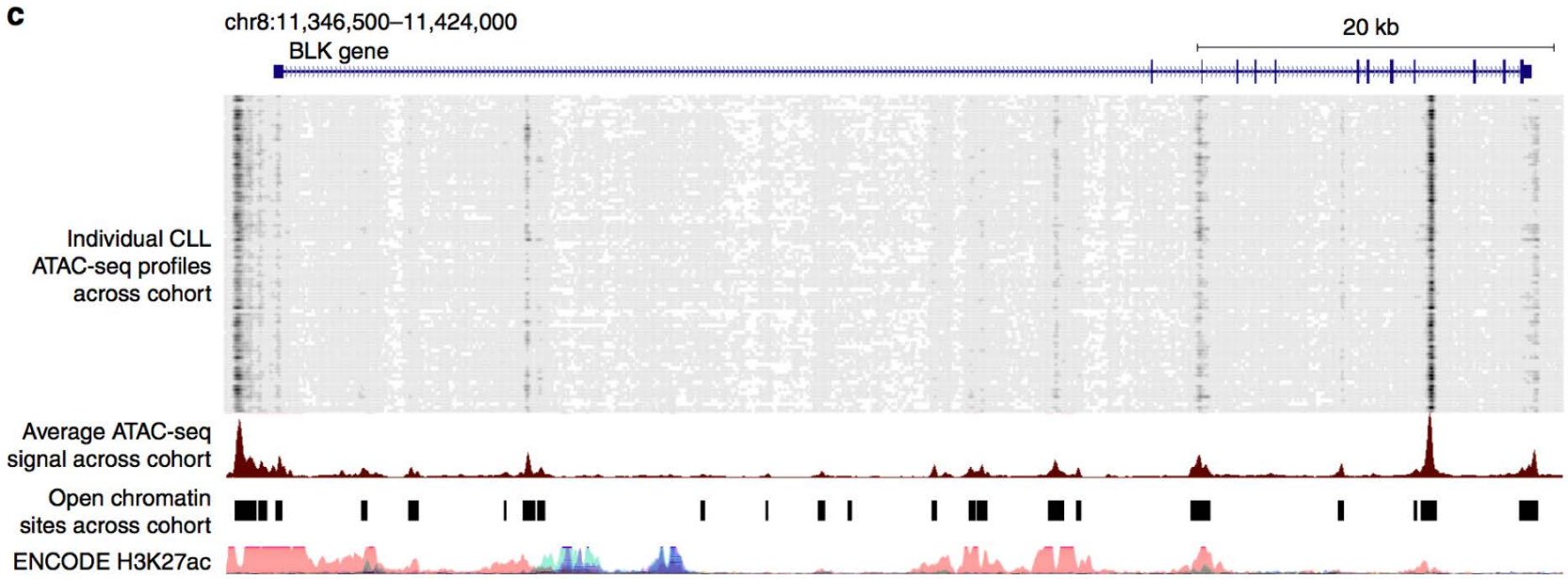


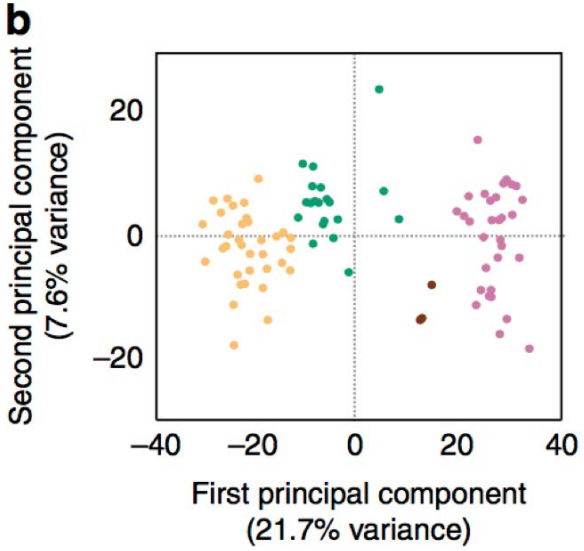
Mack et al. Nature 2018

Potential clinical applications of NGS-based epigenome profiling

- Open chromatin (here: by ATAC-seq, similar to DNase HS) can be used to stratify disease risk

ATAC-seq stratifies patient CLL samples



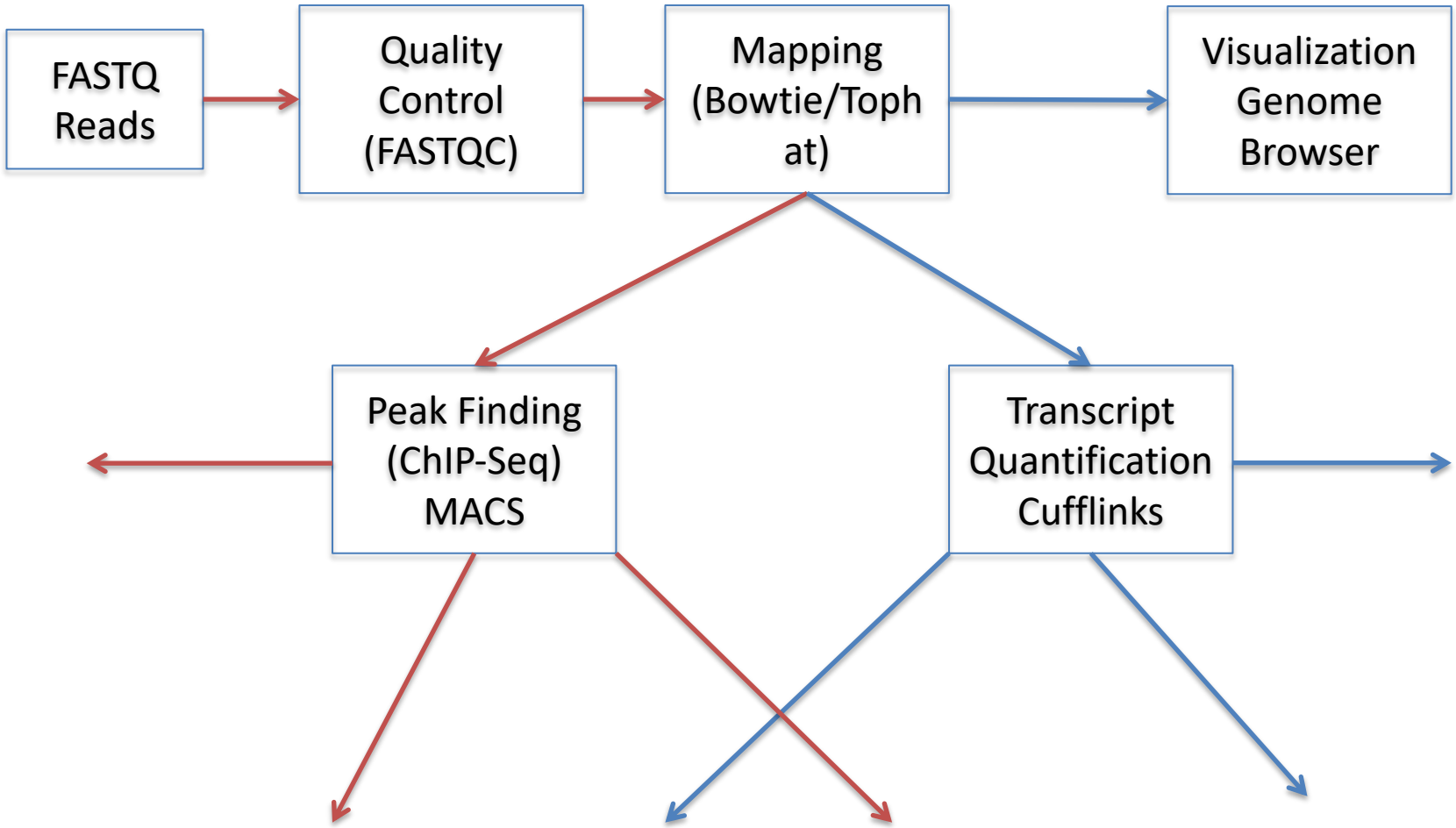


CLL

- Most common leukemia
- Heterogeneous (indolent vs. rapid progression)
- Low mutation burden
- *IGHV* status: mutated vs. germline (aggressive)
- Other biomarkers less predictive

Rendeiro et al. Nat. Comm. 2016

Workflow


FASTQ Reads
Quality Control (FASTQC)
Mapping (Bowtie/Toph at)
Visualization Genome Browser
Peak Finding (ChIP-Seq) MACS
Transcript Quantification Cufflinks

Examples of ChIP-Seq Software

Read Mapping

BWA - handles short gaps, good for 50-100+ bp

Bowtie2 – fast, good for 20-75 bp

BLAT - Very long reads 454

Others: Maq, Mosaik, Novoalign, SOAP2, ZOOM…

(RNA-Seq: Tophat, STAR – spliced alignment)

Peak Finding

HOMER – transcription factors (TF) and histone modification regions (Histone)

MACS2 – industry standard, TF (new version to handle Histone)

Others: QuEST, GLITR, CisGenome, PeakSeq, Sole-Search, findPeaks, SISSRS, E-RANGE, GenomeStudio, various R/Bioconductor packages

Additional Analysis

Motif Finding:

HOMER

CisFinder

MEME/DREME

ChIPMunk

Gene Ontology:

GREAT – location corrected GO analysis for peaks

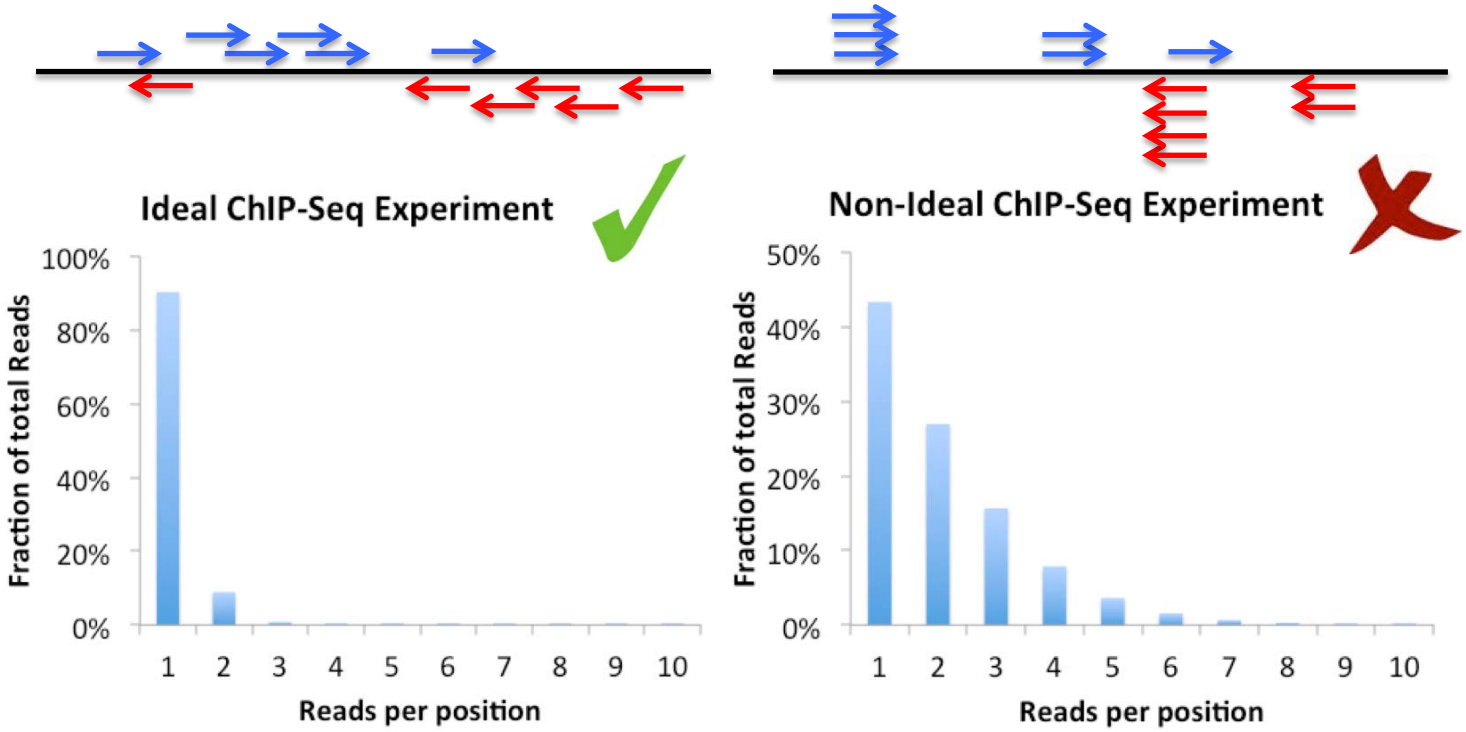
Lots of stuff in R/Bioconductor

Read Alignment: Coordinate Systems

- Version of the Genome is VERY important
  - Make sure you keep track of it
- Canonical Chromosomes vs. All
  - chr5_random – may contain segments of other chromosomes, making it tougher to map reads to a unique position
  - However, chr5_random also contains unique genomic sequence that couldn't be placed in the genome.
- Localizing reads to the genome
  - Most studies only keep reads that map to a unique location in the genome (means that you are "blind" to repeat regions in the genome)
  - How do you tell if your transcription factor binds to a specific class of repetitive element?

Quality Control:
Sequencing Library Complexity

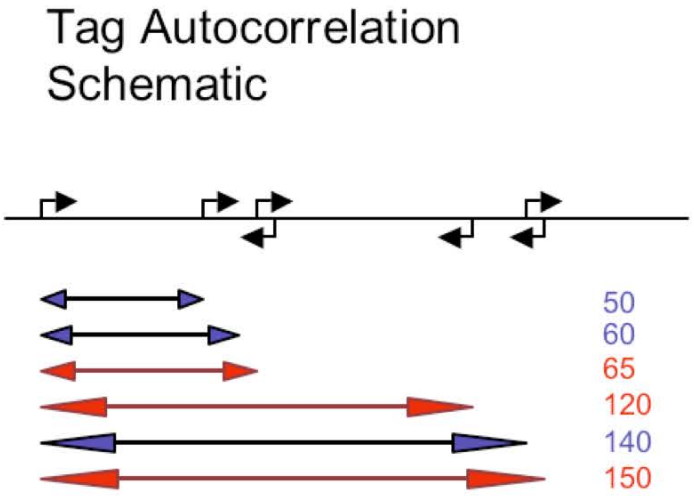
- Reads from most types of sequencing experiments should not be identical
- If an experiment is “clonal”, there is a good chance that not enough starting material was used during library construction/over amplified

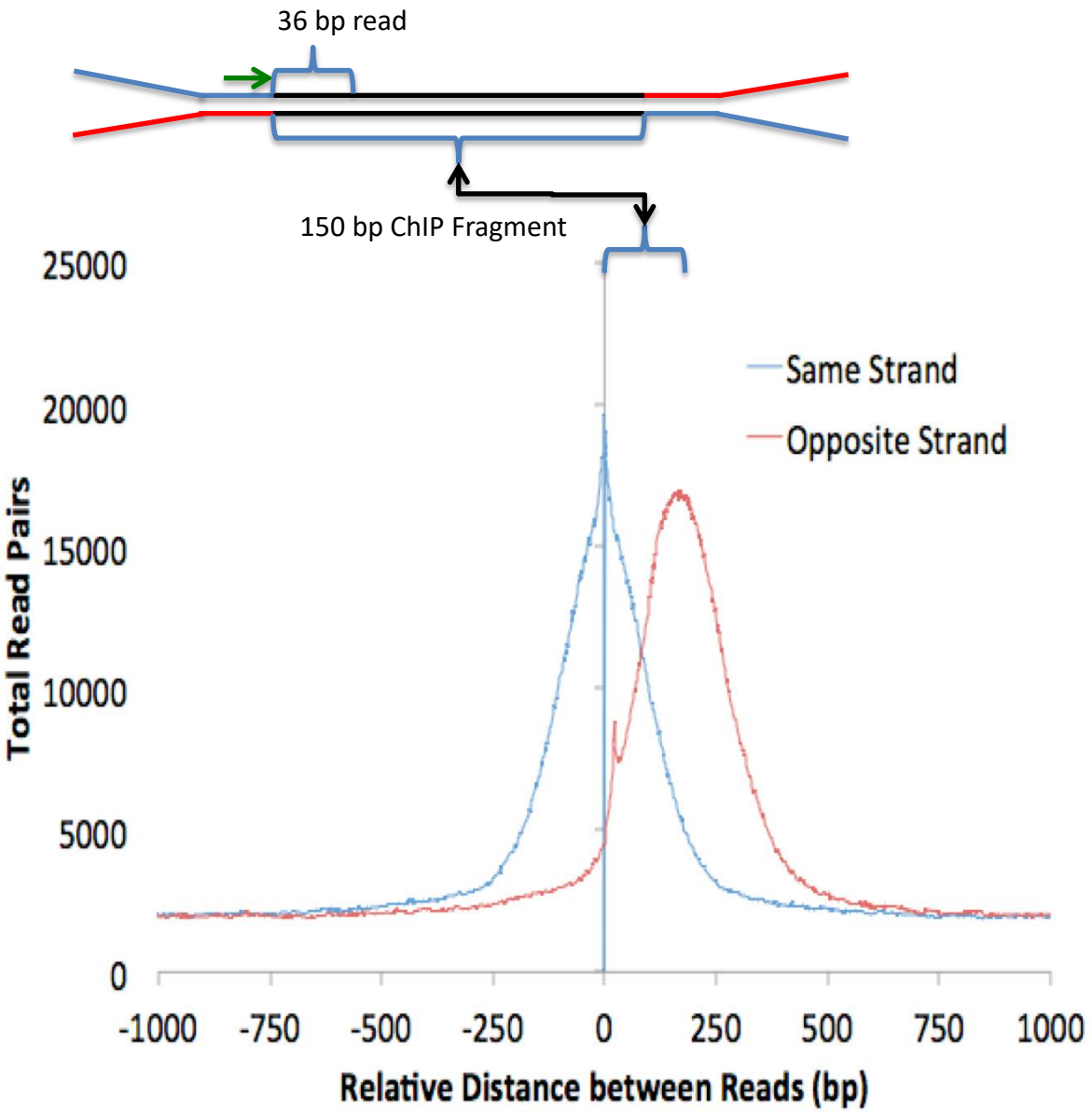


Quality Control:

Autocorrelation

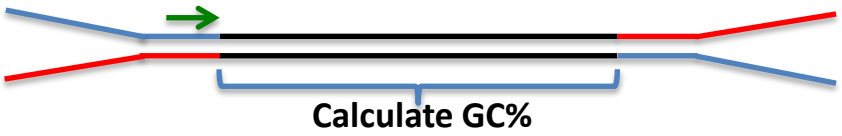
Use the distribution between reads to estimate the length of fragments cut-out for sequencing



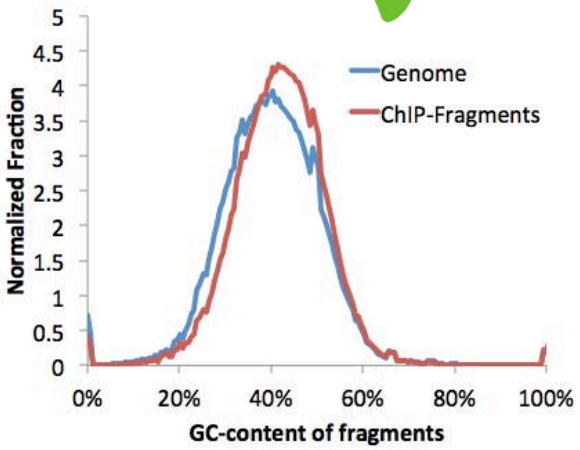


Quality Control:

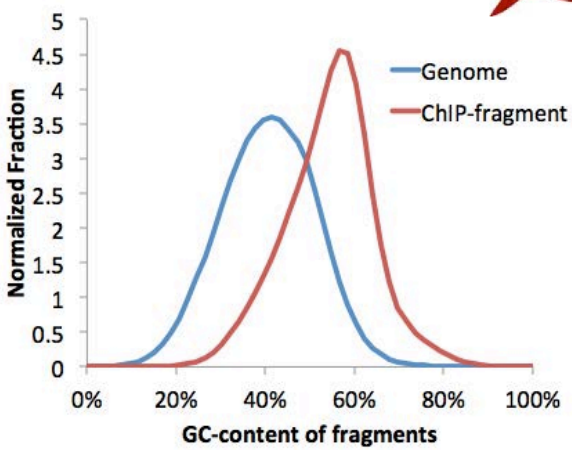
GC% Fragment Bias

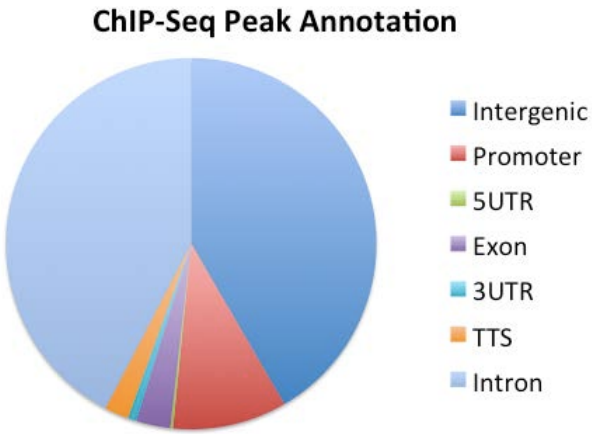


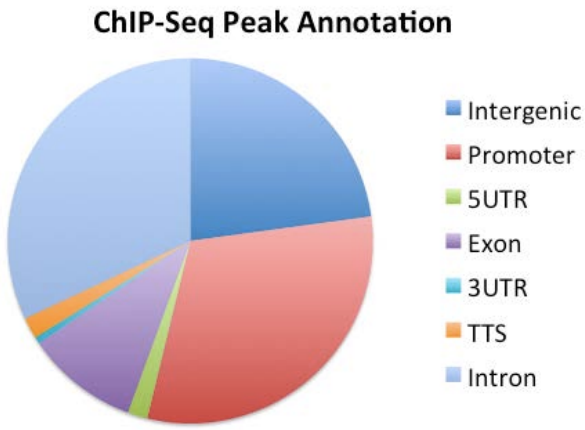
Ideal Experiment



Non-Ideal Experiment

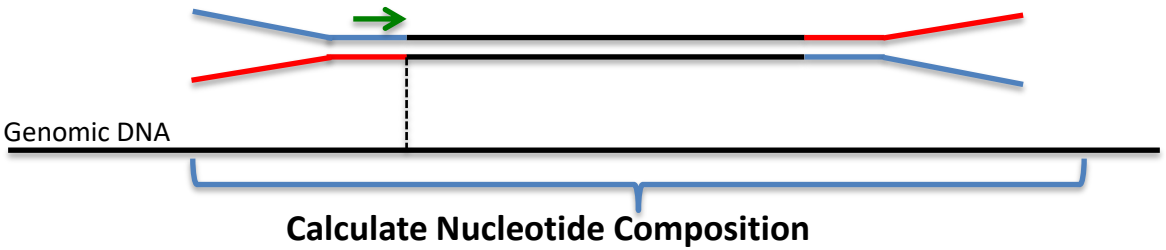


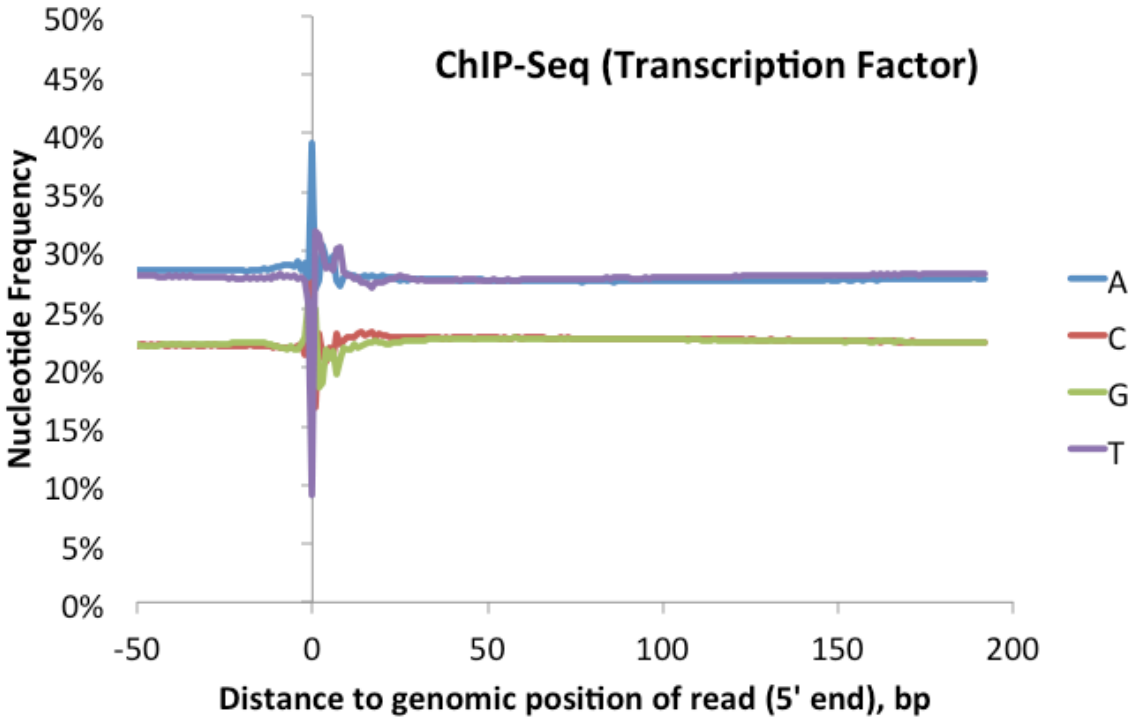




Quality Control:

Position-specific Nucleotide Composition Bias

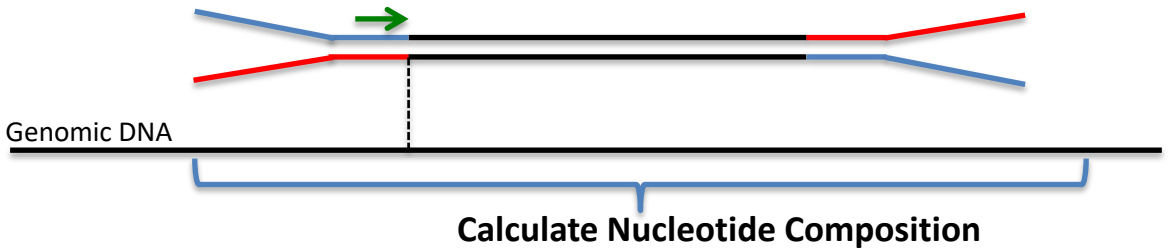


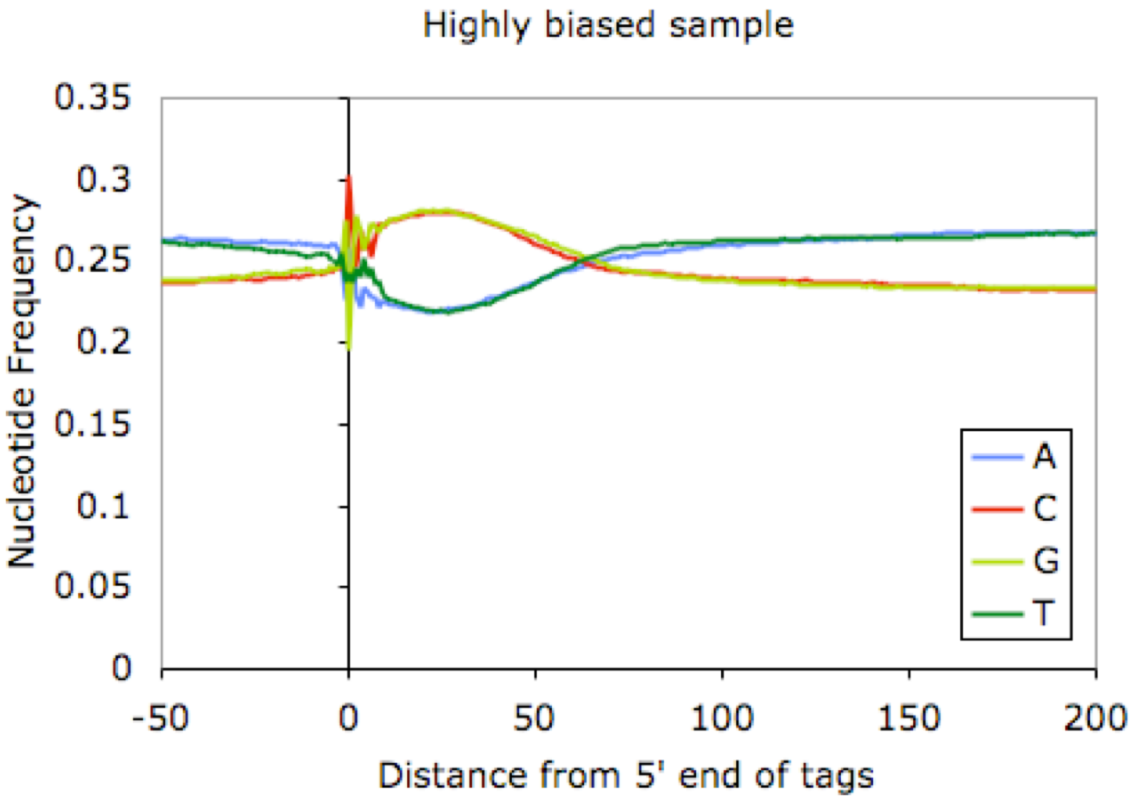


Reveals bias in enzymatic manipulation of sample

Quality Control:

Position-specific Nucleotide Composition Bias

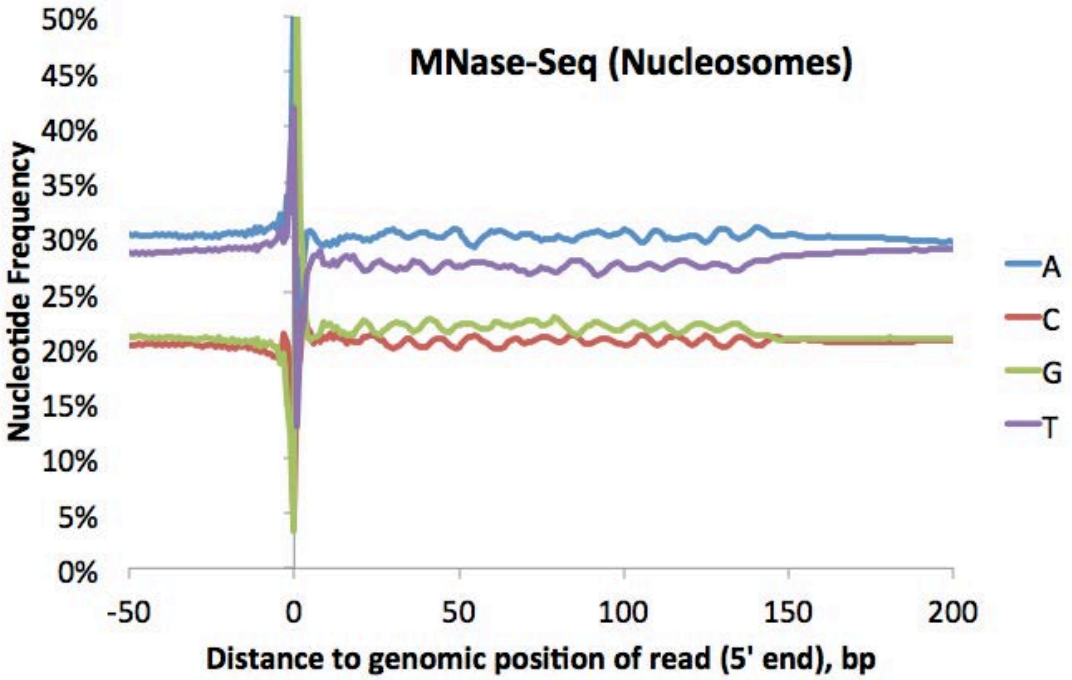


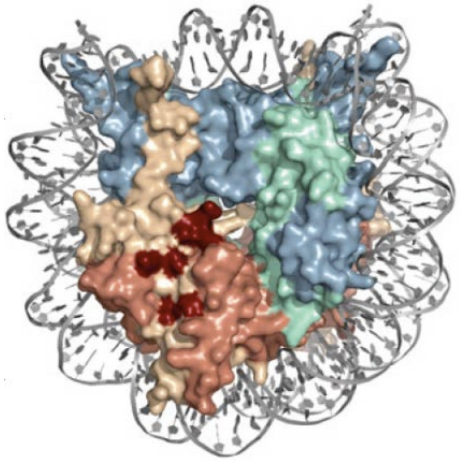


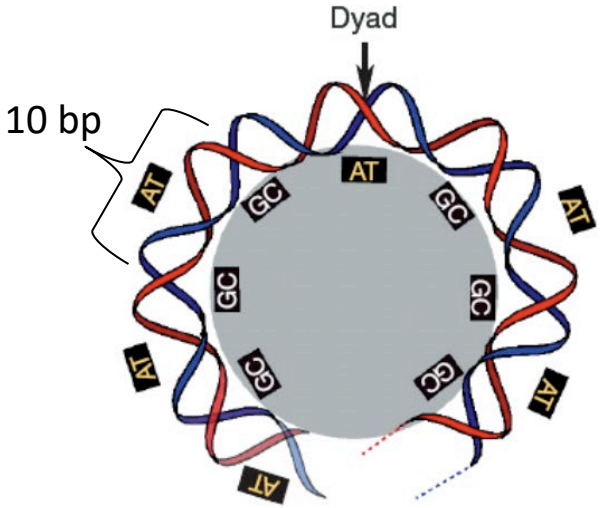
Reveals bias in enzymatic manipulation of sample

Quality Control:

Position-specific Nucleotide Composition Bias







Traditional Nucleosome Positioning Sequences

(From Albert et al. Nature Nature 446, 572-576 (29 March 2007)

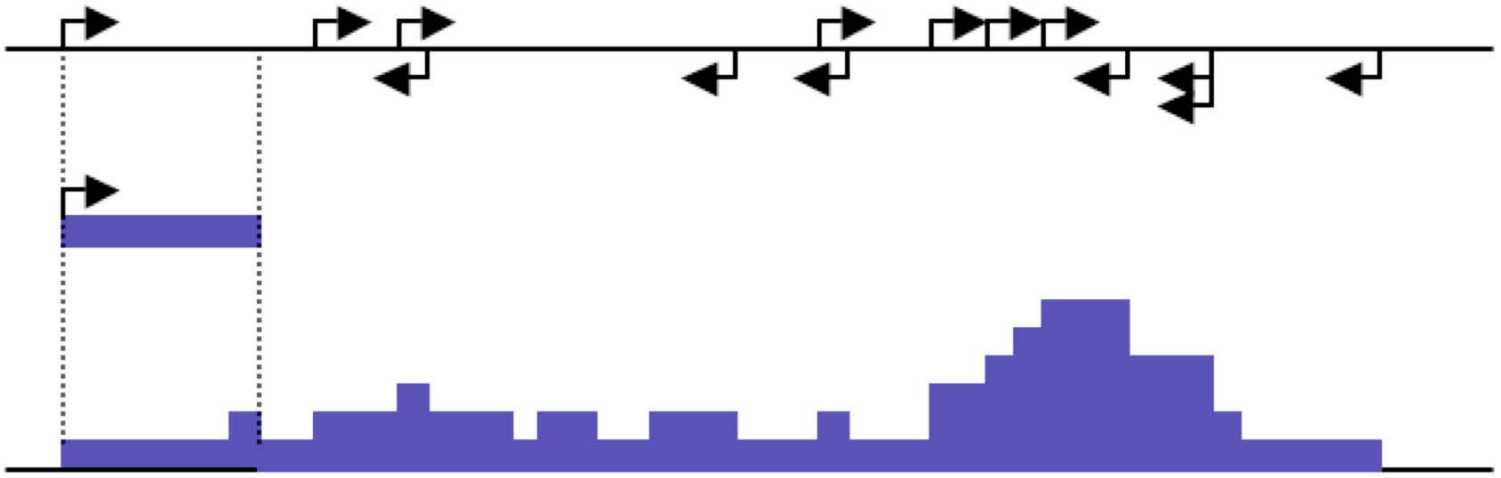
Data Visualization

UCSC Genome Browser – Give these guys a Nobel prize for their impact on the genomics/bioinformatics community!

- UCSC GB integrates a very large number of diverse data sets and allows you to visualize them across the genome.
- Upload and visualize your own sequencing experiments to the genome browser!
- Other Genome Browsers / Data Viewers are out there as well (e.g. **WashU Epigenome Browser** (*nice* interface), **IGV** (local)).
- Some are faster and make it easier to zoom in and out etc., but none of them really match the depth of data available at UCSC

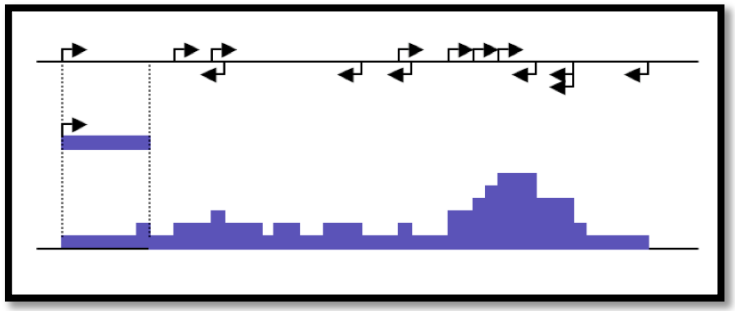
Creating Tag Pileups

- Extend each read to the estimated ChIP-fragment length (i.e. ~150 bp)
- Add up the coverage of each fragment to build a pileup



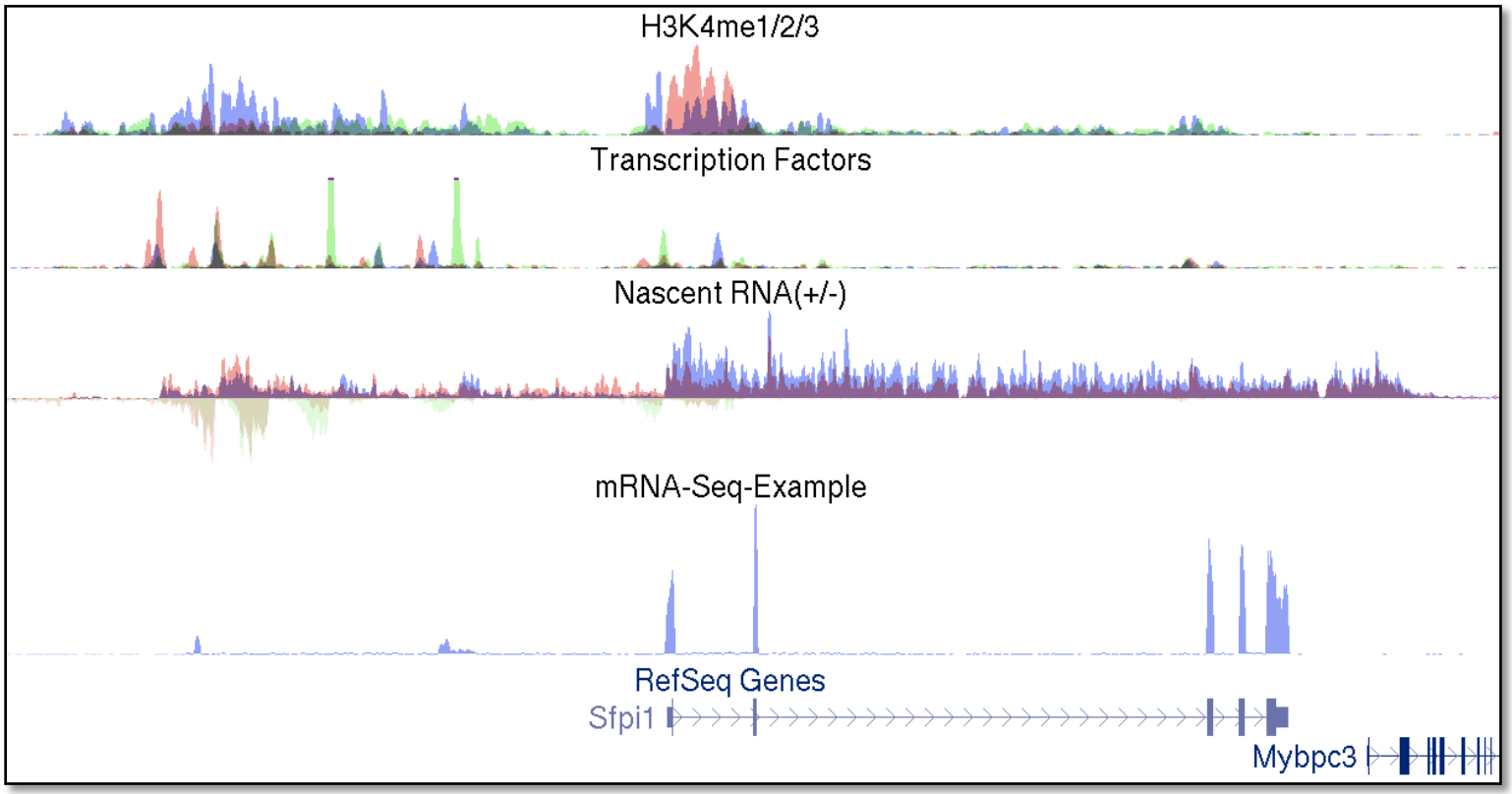
Quality Control:

Read Density Visualization



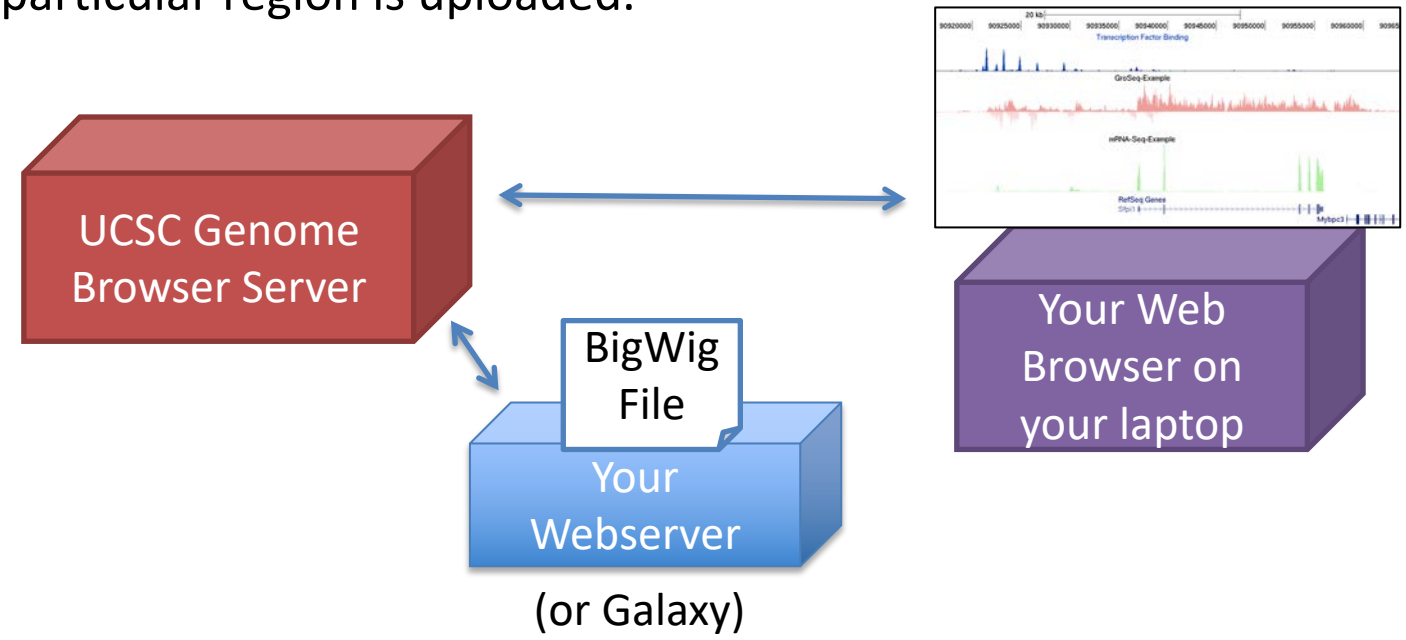
Raw read fragments are piled up to produce read densities and visualized using **UCSC Genome Browser**

HOMER supports creation of “bigWig” and Track Hubs (translucent combination tracks)



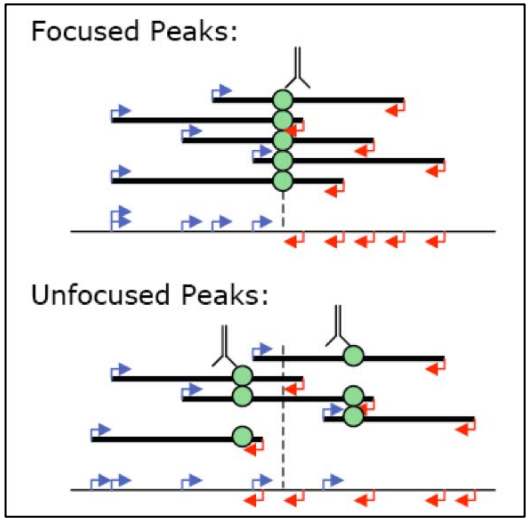
BigBed/BigWig Files

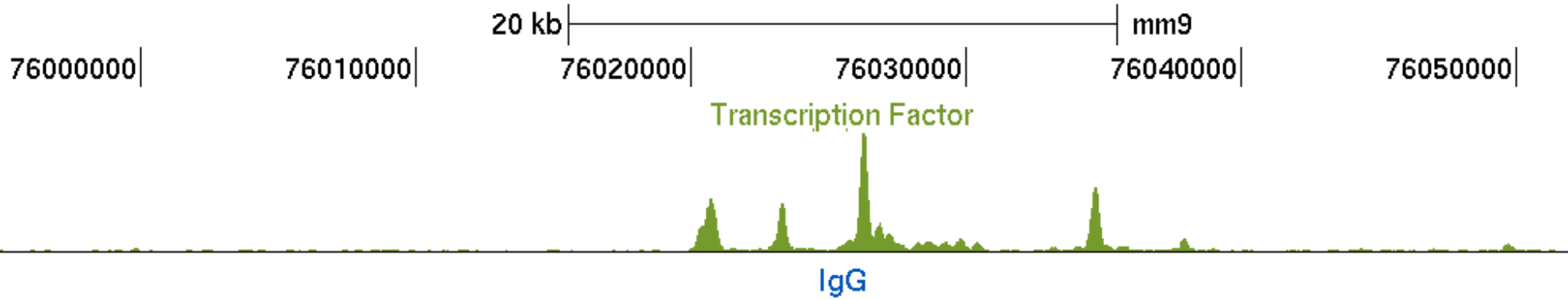
- Quick NOTE: bigWig/bigBed Files
  - There is a practical limit to how much custom data can be loaded onto UCSC at a time.
  - In order to upload full experiments, UCSC has implemented tools that allow you to host indexed data files on your own webserver. These files are then accessed by UCSC, and only the data needed for a particular region is uploaded.



Peak Finding

- **<u>Unbiased detection of regions with high read density</u>**
- If reads were selected randomly from the genome, they would follow a Poisson distribution (more or less)

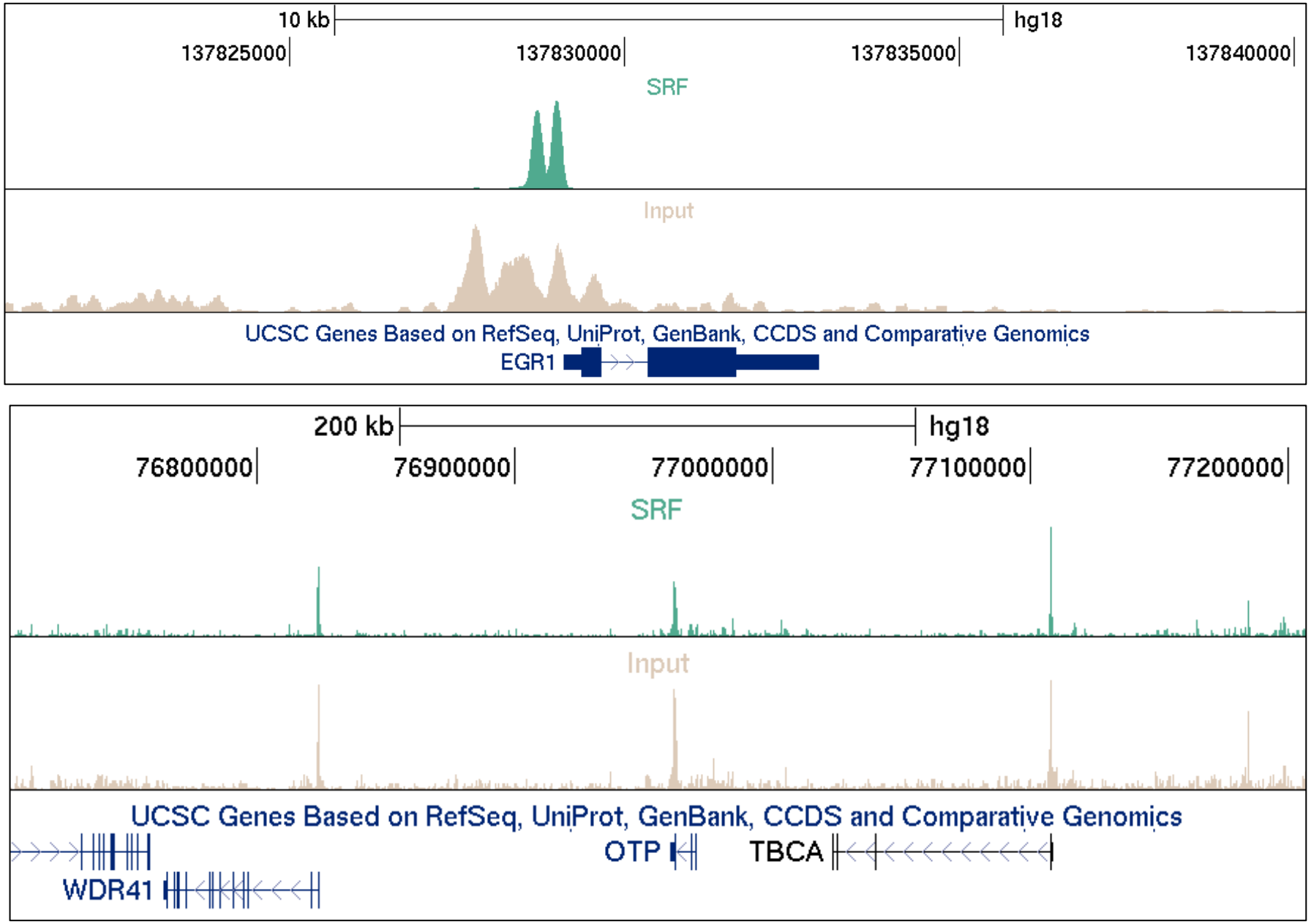




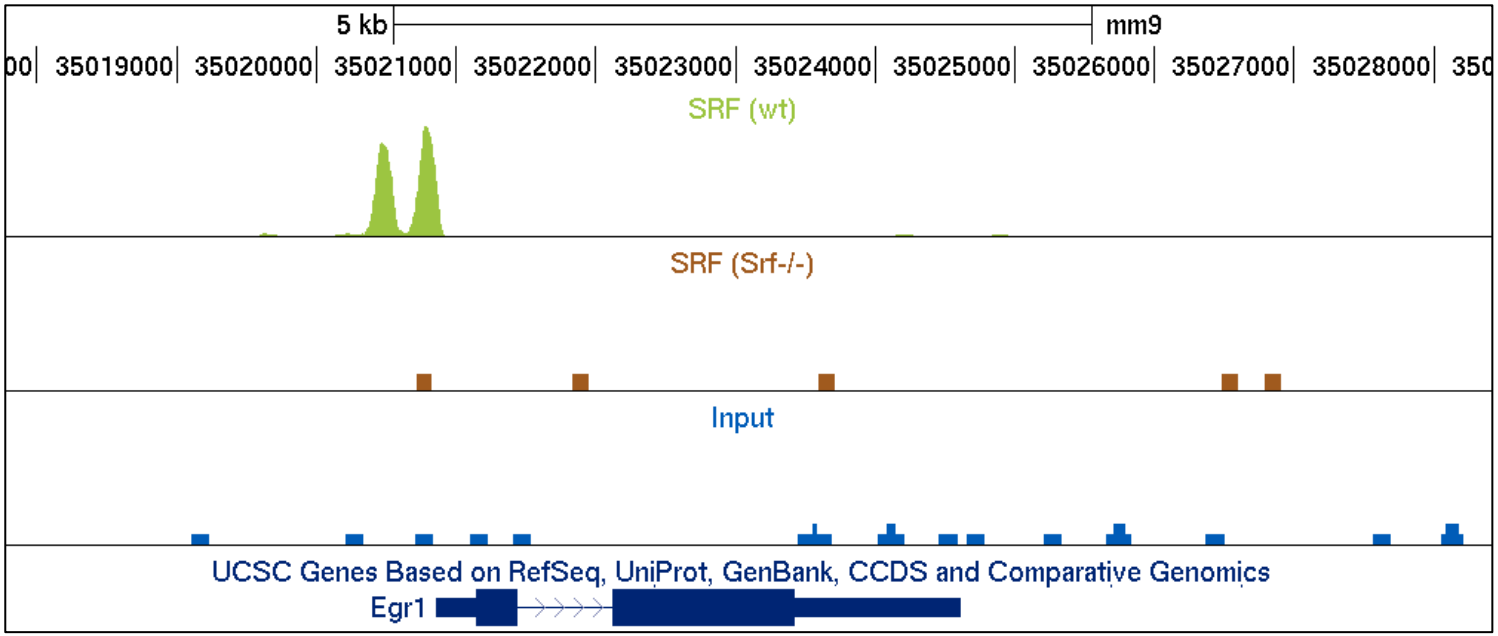
Problems with a simple approach to peak finding

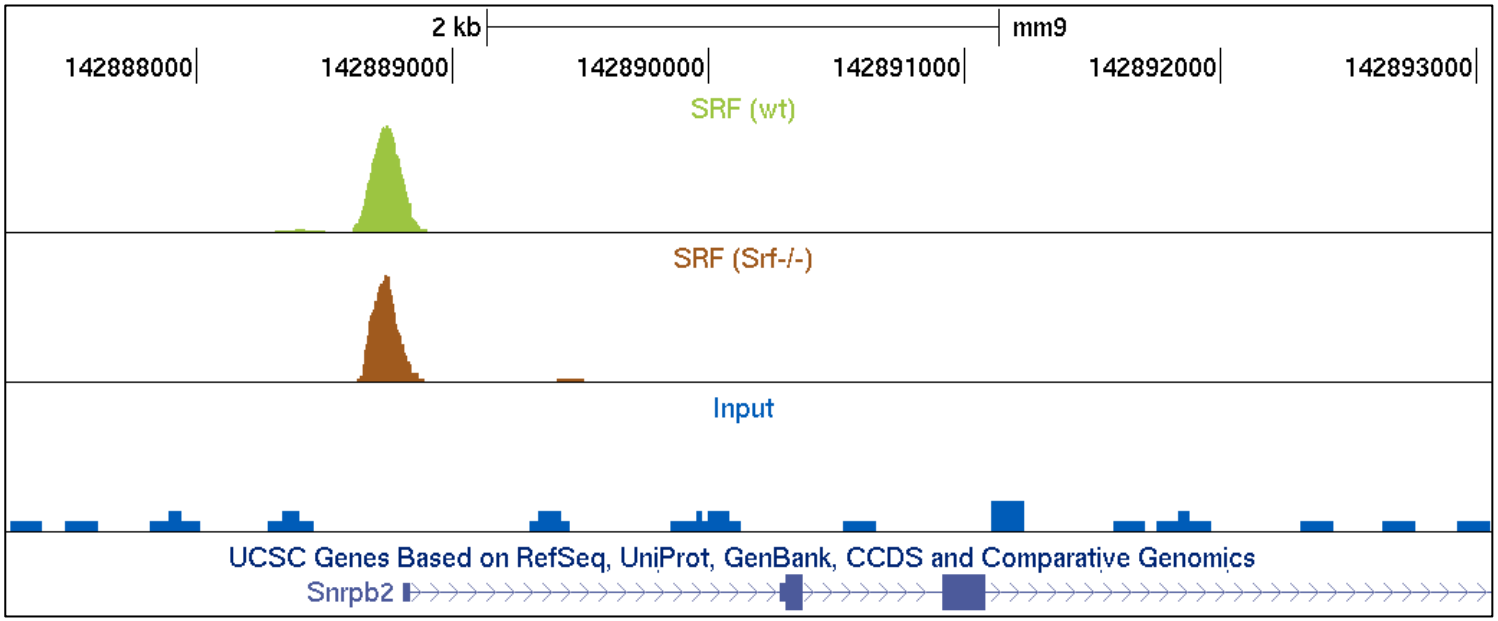
- Random distribution of reads is not a realistic control!
  - Read alignment bias (repeats are hard to map reads to)
  - Sample genome (i.e. cell line) is different from the reference genome (copy number variants, etc.)
  - ChIP protocol introduces bias
    - enzymatic preference for specific sequences/adapter ligation
    - GC-bias introduced introduced by washes/PCR
    - DNA extraction protocol may bias certain regions
  - Antibody – May not be very specific
  - Contaminants – plasmids, satellite DNA, etc.
- Important to run a control
  - Input (ChIP protocol without antibody selection)
  - IgG (ChIP protocol using a negative control antibody) [Input is preferred over IgG]
  - ChIP using Knockout cells (best control, if possible)

Dirty Input


10 kb
hg18
137825000
137830000
137835000
137840000
SRF
Input
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
EGR1
200 kb
hg18
76800000
76900000
77000000
77100000
77200000
SRF
Input
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
OTP
TBCA
WDR41

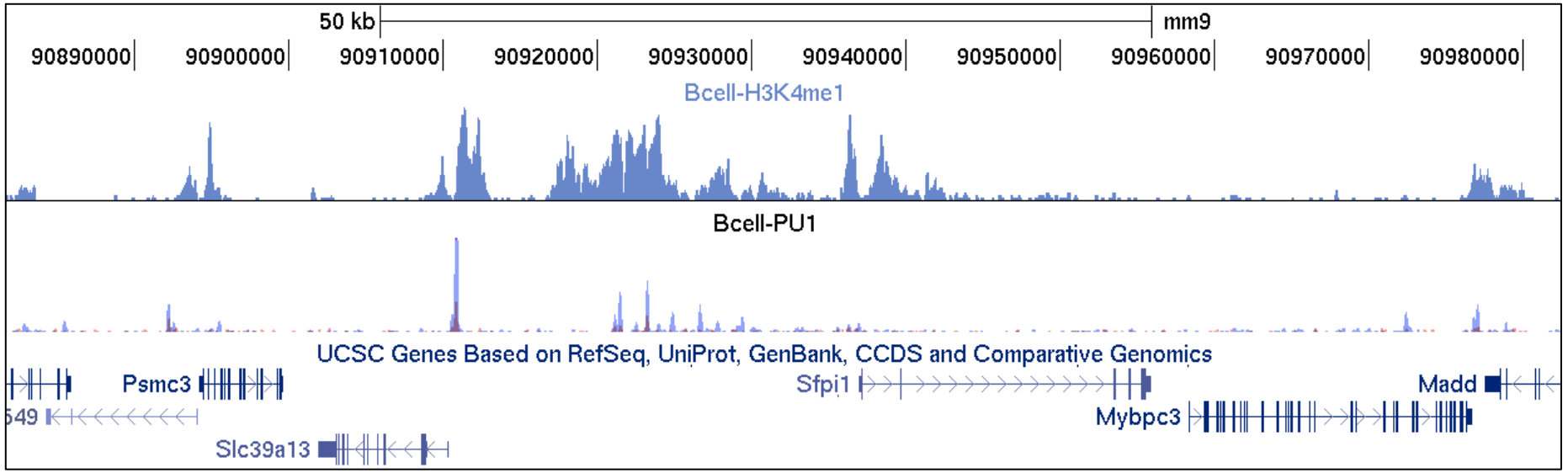
ChIP-Seq for SRF in SRF-/- Mice



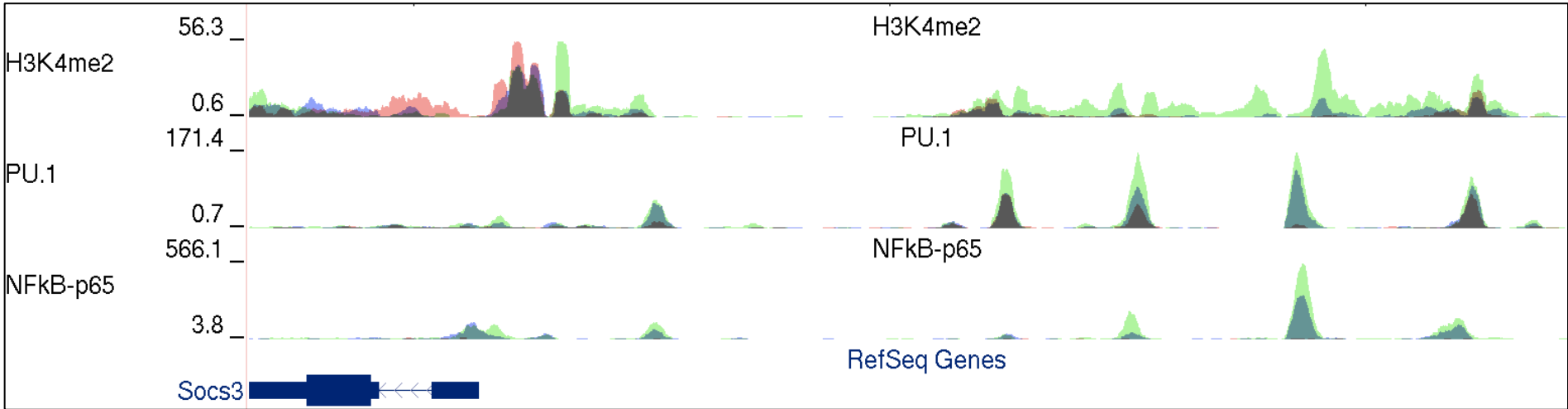


Histone Modifications

Generally want to define broad domains with continual signal enrichment

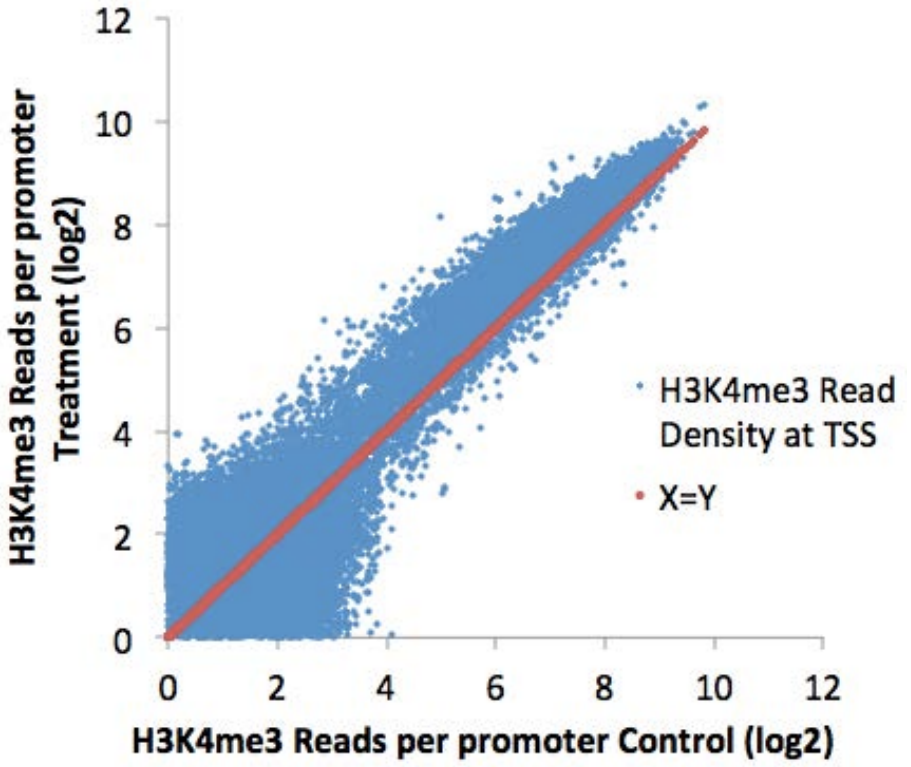


If MNase is used for histone modifications, it can be possible to identify individual nucleosomes and nucleosome-free regions:

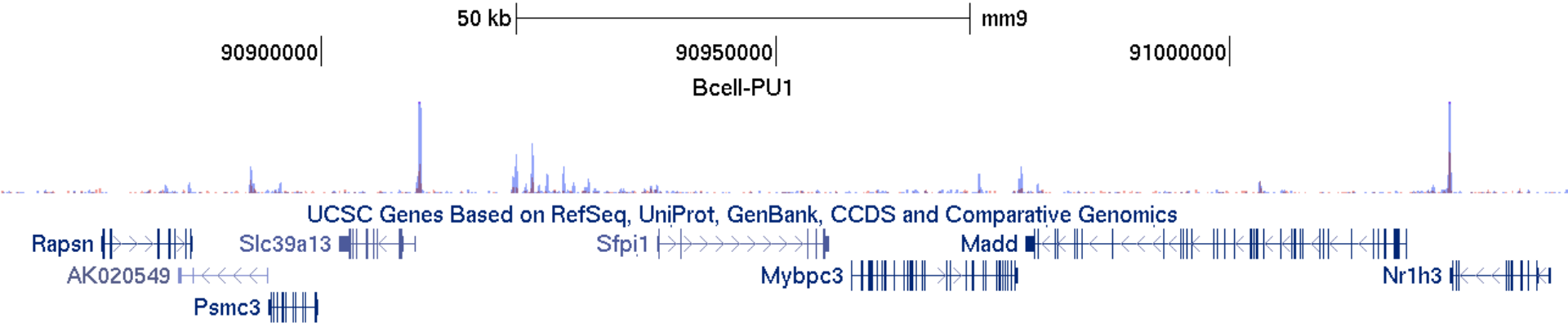


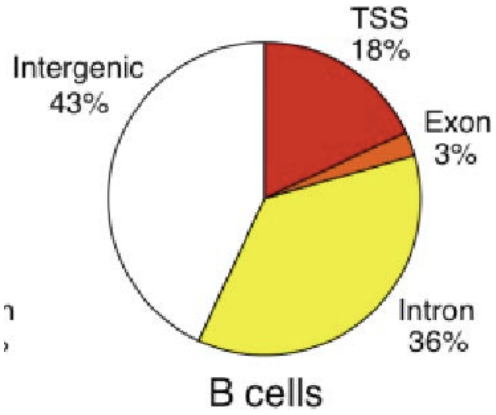
Comparing Experiments, Normalization

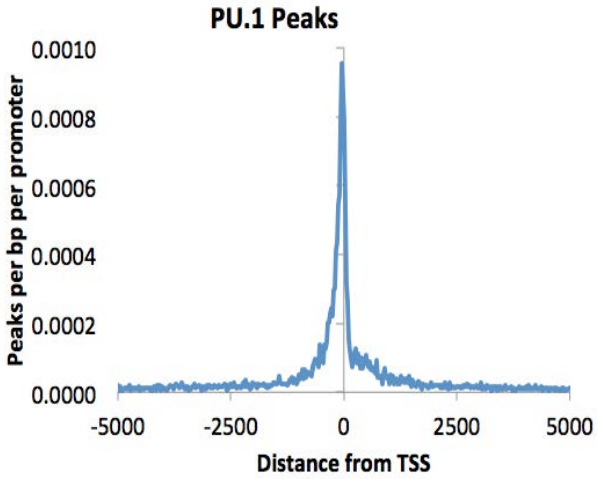
- Need to control for sequencing depth:
  - Normalize to total mapped tags
  - Normalize for IP efficiency – difficult! Requires assumptions about regions that should have equal coverage
- Differential Peaks:
  - Consider the number of reads at a given genomic locus in each separate experiment
  - Exactly like RNA-Seq, and a lot like microarray differential expression (except raw data is integer counts)
  - Data commonly modeled using Poisson or Negative Binomial distributions
  - Do not overlook fold change – unlike other metrics, it is sequencing depth-independent



ChIP-Seq Peak Annotation







Gene Ontology with ChIP-Seq Data: GREAT

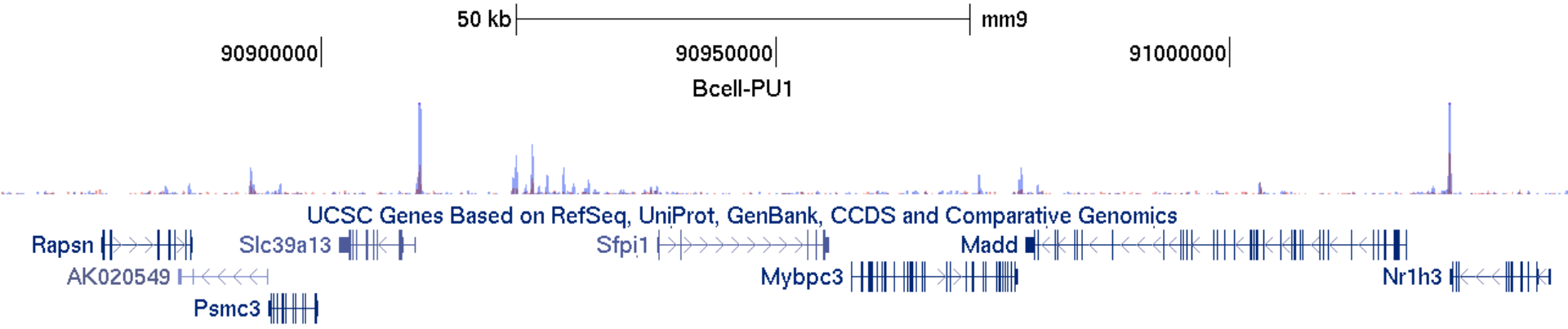
GO Molecular Function

Table controls: Export — Shown top rows

Term Name	Binom Rank	Binom Raw P-Value	Binom FDR Q-Val
transcription repressor activity	6	5.1863e-23	2.5733e-20
DNA regulatory region binding	9	1.3450e-13	4.4491e-11
promoter binding	10	2.1783e-13	6.4849e-11
transcription corepressor activity	13	7.9016e-9	1.8095e-6
SMAD binding	18	4.8776e-7	8.0670e-5
transcription repressor binding	25	5.1618e-6	6.1467e-4
specific transcriptional repressor activity	43	2.0405e-4	1.4127e-2
extracellular matrix structural constituent	60	7.8257e-4	3.8828e-2

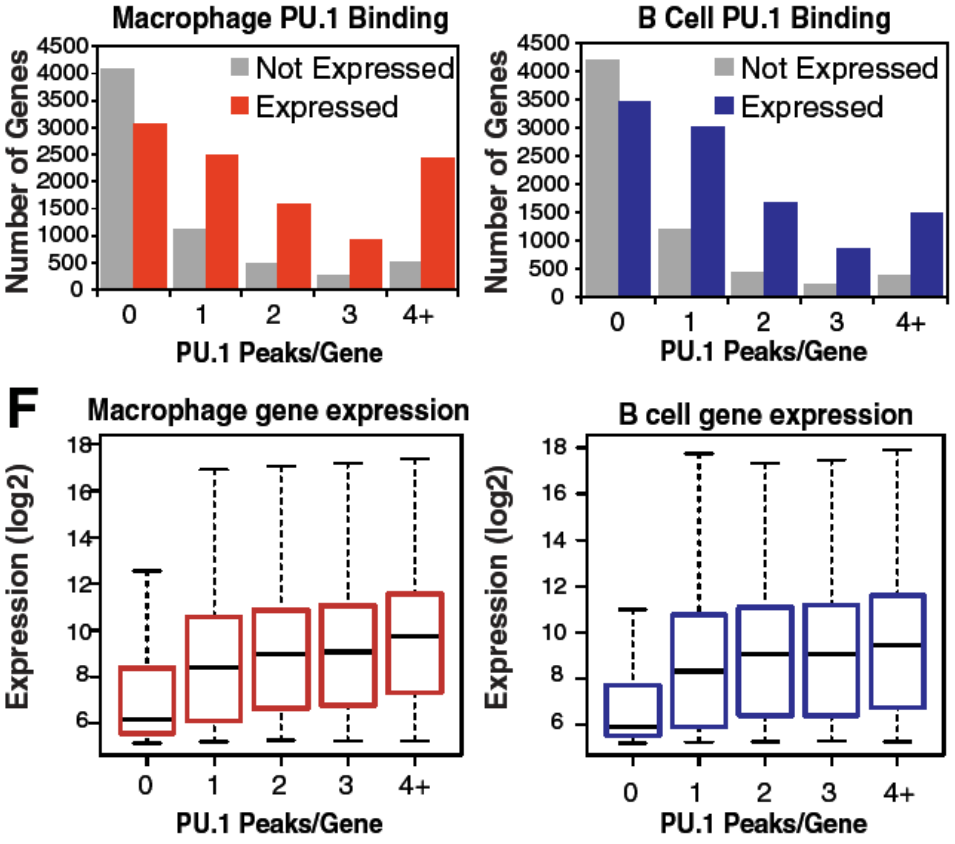
Assigning ChIP-Seq Peaks to Target Genes

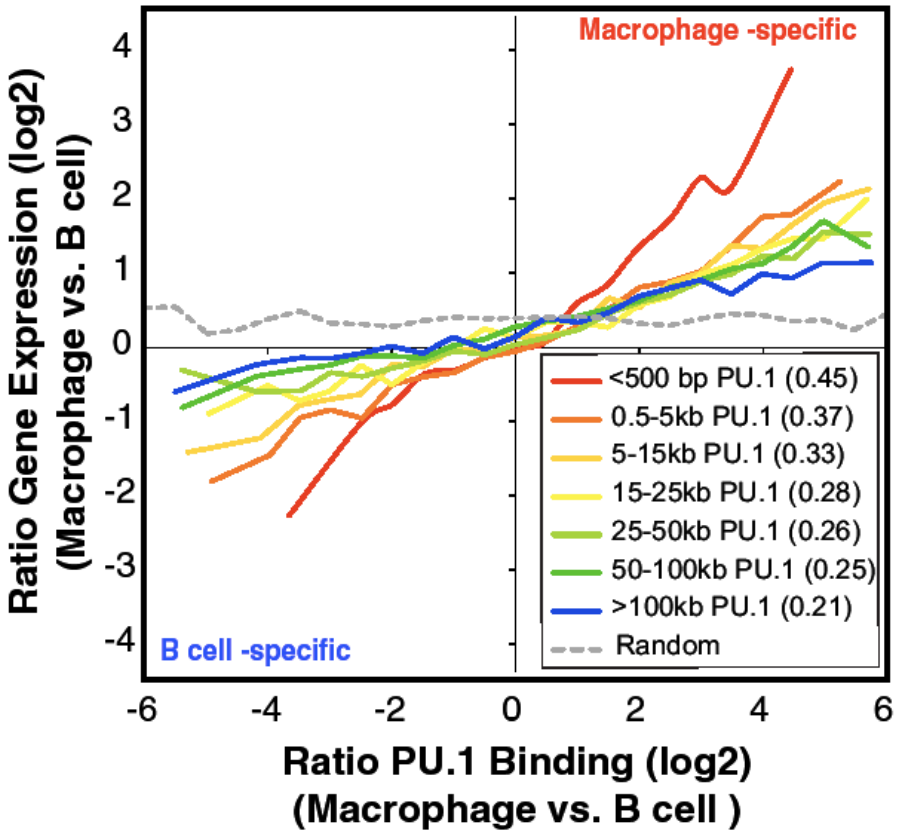
- Assign Peak to nearest TSS
- Assign all peaks within a certain distance to TSS
- Only assign conserved peaks/within same syntenic blocks
- Advanced association (additional information required)
  - eQTL, bQTL, …QTL – used genetics to help correlate gene activity with a given genomic region
  - 3D genomic interactions (looping chromatin) 3C, ChIA-Pet, Hi-C



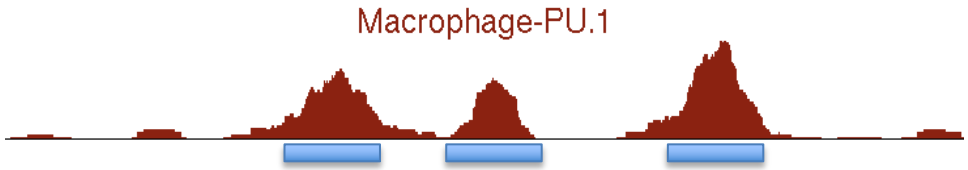
Correlating Gene Expression with ChIP-Seq

- Correlation is usually poor at best…



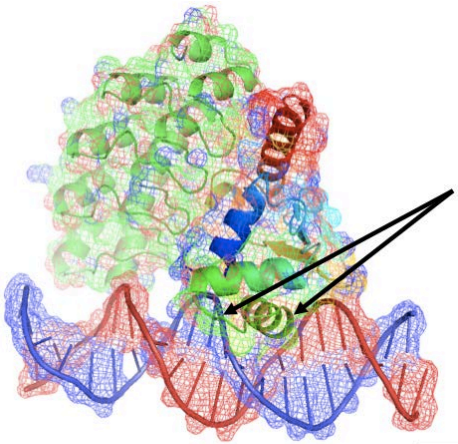


Motif Discovery



- Sequence specific DNA-binding Transcription factors *should* bind to sites that contain regulatory elements bound by the factor
- Very Important aspect of ChIP-Seq
  - Quality control – if you cannot find the expected motif, maybe the ChIP wasn't as good or as specific as originally thought.
  - Provides in vivo binding specificity
  - Important hypothesis generating tool. Identify co-factors, etc.
  - Can help identify artifacts
- Possible to perform Motif discovery on different types of ChIP-Seq experiments including histone modifications and Transcription factors.
- Key things to be aware of:
  - CpG Islands: Mammalian genomes contain GC rich regions that can through off many motif finding algorithms
  - TSS/Proximal Promoter Motifs: Most TFs are biased to the TSS, where certain elements are very common (i.e. SP1, CAAT box, E-Box…)
  - Transcription Factor Hot-Spots

De novo Motif Discovery in HOMER



Contacts between protein and DNA form the basis for DNA sequence specificity

GGGAATTCCC
GGGAAATCCC
GGGAAACCCC
GGGAATCCCC
GGGAGATCCC
GGGAACTCCC
GGGAAGTCCC
GGGAAATTCC
GGGAATCCCC
GGGAGTCCCC

Position	1	2	3	4	5	6	7	8	9	10
pA	0	0	0	1	0.8	0.4	0	0	0	0
pC	0	0	0	0	0	0.1	0.4	0.9	1	1
pG	1	1	1	0	0.2	0.1	0	0	0	0
pT	0	0	0	0	0	0.4	0.6	0.1	0	0

Probability Matrix

Crystal Structure from Batchelor et al. Science. 1998 Feb 13;279(5353)

Differential Motif Discovery: Finds sequences that are specifically enriched in the target set

Group	Sequence
Target	TTCTGAACCACACTCCAAGACCAGGAAGTGGCCCCTATGGCCAGAATCCT…
Target	CTCAGTCCCCGAGGAAGTAGAAAAGACAGAACCACATAGATTAGGGTGCT…
Target	AACCACAGTCATAAATGTAATAGGTTAACTCTTTGAGGAAGTAACCACACTC…
Target	AAAGAGCCAACCACATTGTGGAGGTTAGAGATTTTAGGAGCTAGCGGCGAC…
Target	TGATTTCCGACATAACCACAGCTCACTTCCGAGGAAGTCAACAAAGCAATTT…
Background	CCGCCCCGGGACGTGCCACCCGACGCGCGCAACCACACCATCGTGGGCA…
Background	TTGAGAGCCGAGATTTTATATAACCACAGGGCGGGTTGGGAAAAAAAGCCG…
Background	AAACACCAACAGGAAGTTTCGCGTAGAGAAAATTACCCAGTATAAAAATTGT…
Background	CCCAGATATATGAGTTTGTGGAACCACAAAACCCGGGTTTGTGAAGAGTAT…
Background	CAAGTGGCAAAGACTCTGTAGTTGTTAACACCACCTGACCCTATGGCAGAC…

Motif significance calculated using zero or one occurrence per sequence (ZOOPS) coupled with cumulative hypergeometric distribution to calculate enrichment

Pre-processing Phase:

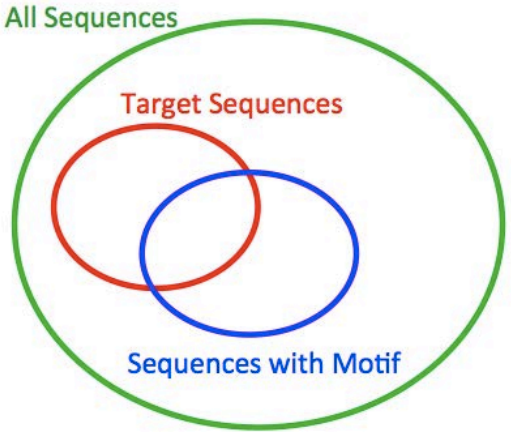
- Remove redundant sequences
- Normalize GC-content

Exhaustive Search Phase:

- Screen all possible oligos for enrichment

Local Optimization Phase:

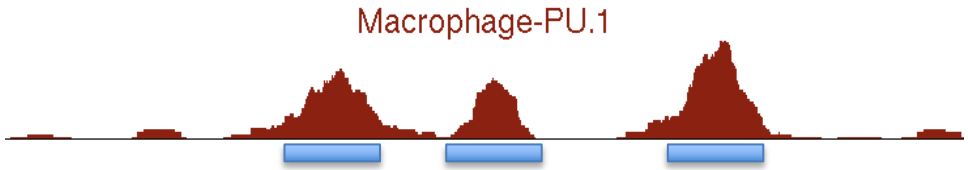
- Expand promising oligos into probability matrices
- Iteratively improve matrices by considering individual contributions from different oligos



Objective: Find the motifsthat are highly ***enriched*** in target sequences

(Maximize the overlap)

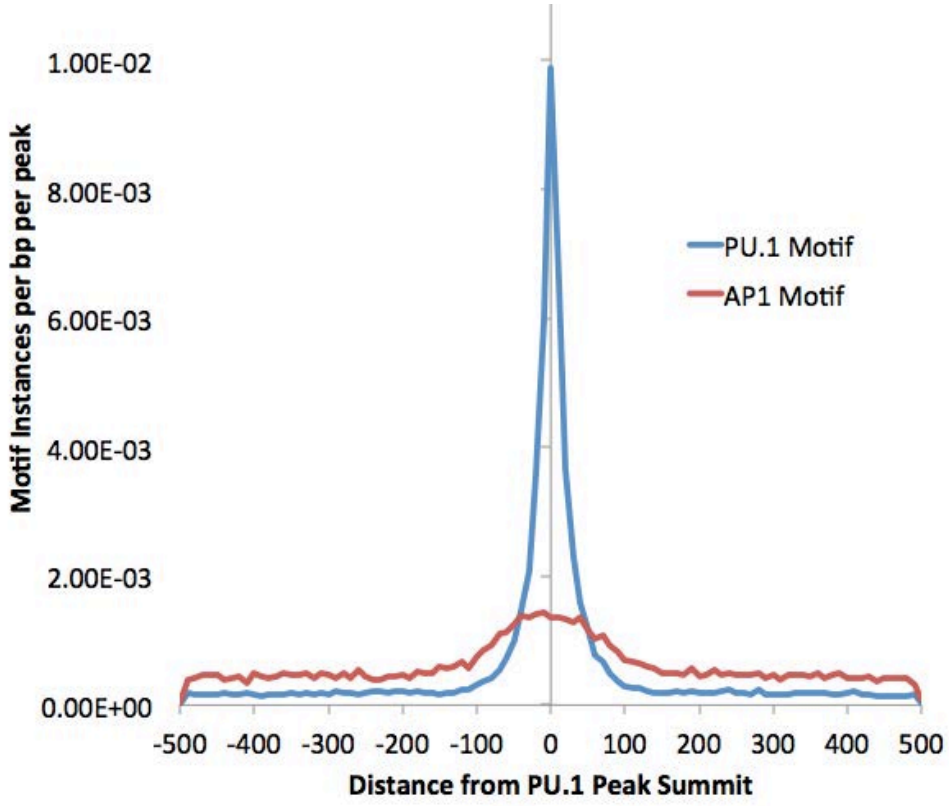
ChIP-Seq Motif Finding



Extract Sequences at peaks, compare to random genomic sequence (matched for GC%)

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details
1		1e-15178	-3.495e+04	62.95%	8.94%	32.7bp (63.7bp)	PB0058.1_Sfpi1_1 More Information | Similar Motifs Found
2		1e-800	-1.844e+03	24.47%	13.08%	52.7bp (59.7bp)	MF0006.1_bZIP_cEBP-like_subclass More Information | Similar Motifs Found
3		1e-762	-1.755e+03	12.92%	5.12%	51.4bp (60.4bp)	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq/Homer More Information | Similar Motifs Found
4		1e-370	-8.533e+02	2.35%	0.41%	48.3bp (56.4bp)	MA0139.1_CTCF More Information | Similar Motifs Found
5		1e-286	-6.597e+02	7.00%	3.26%	52.5bp (68.3bp)	Sp1(Zf)/Promoter/Homer More Information | Similar Motifs Found
6		1e-273	-6.306e+02	4.04%	1.44%	55.7bp (66.6bp)	PB0037.1_Isgf3g_1 More Information | Similar Motifs Found
7		1e-268	-6.175e+02	11.44%	6.61%	54.2bp (57.0bp)	RUNX(Runt)/HPC7-Runx1-ChIP-Seq/Homer More Information | Similar Motifs Found
8		1e-227	-5.239e+02	7.15%	3.69%	54.2bp (53.6bp)	PL0005.1_hlh-30 More Information | Similar Motifs Found

Distribution of Motifs at ChIP-Seq Peaks



Links

- UCSC Genome Browser

  http://genome.ucsc.edu
- Galaxy

  https://usegalaxy.org
- HOMER

  http://homer.ucsd.edu/homer
- GREAT

  http://bejerano.stanford.edu/great/public/html/

Questions?