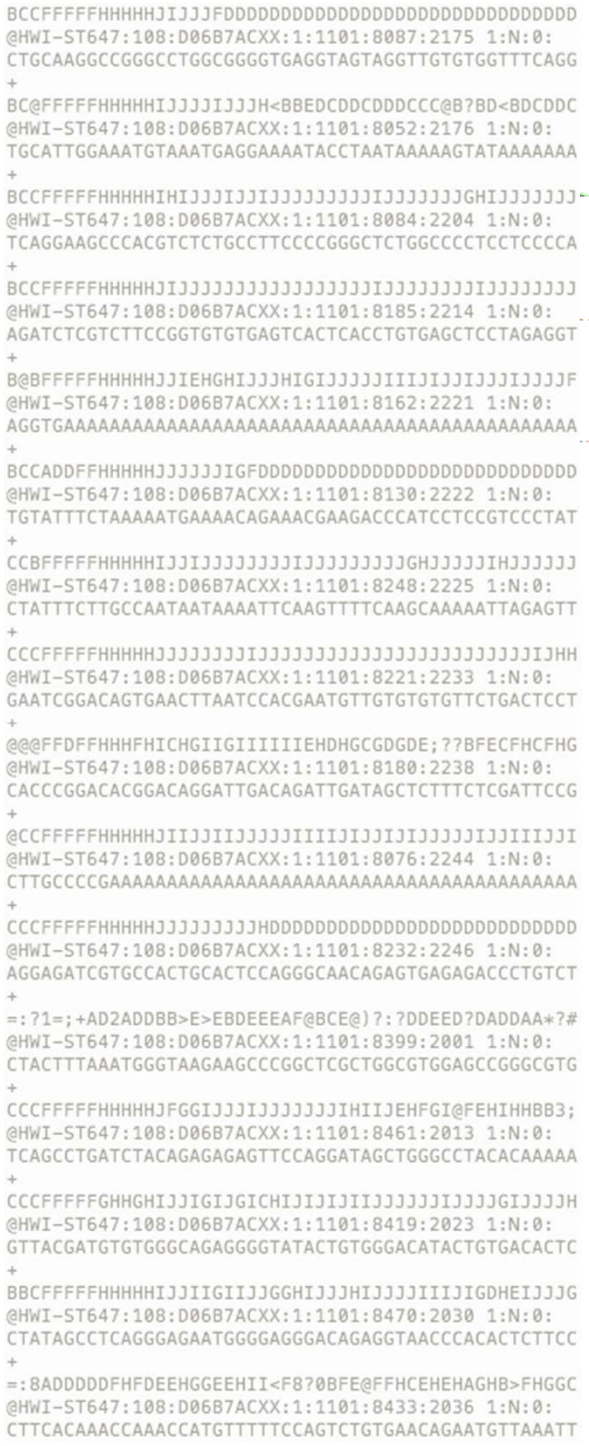
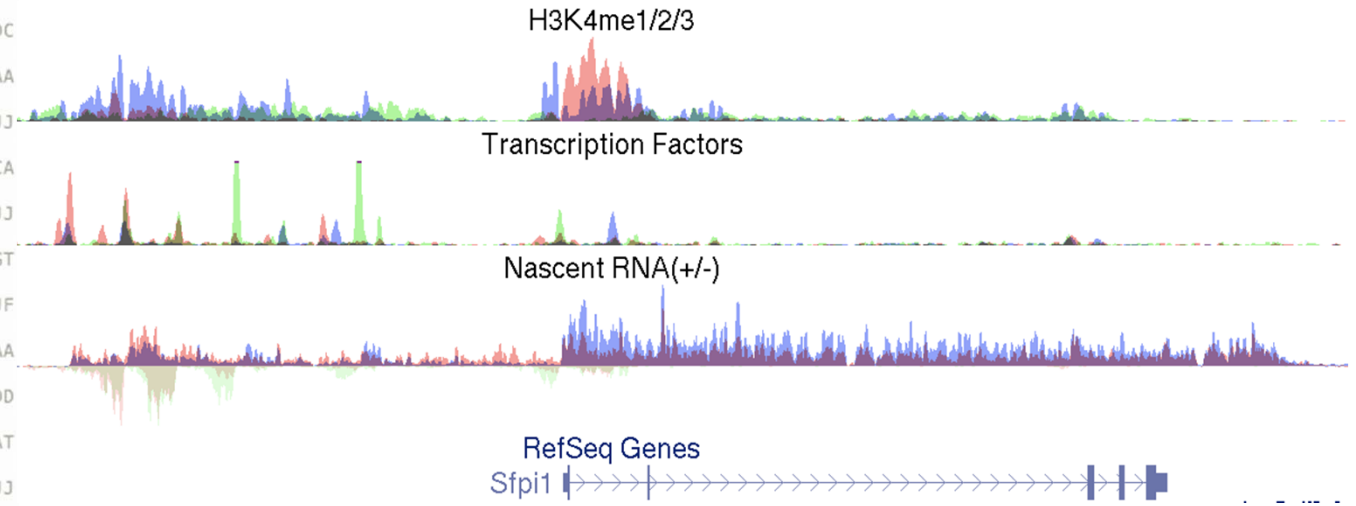




ChIP-Seq

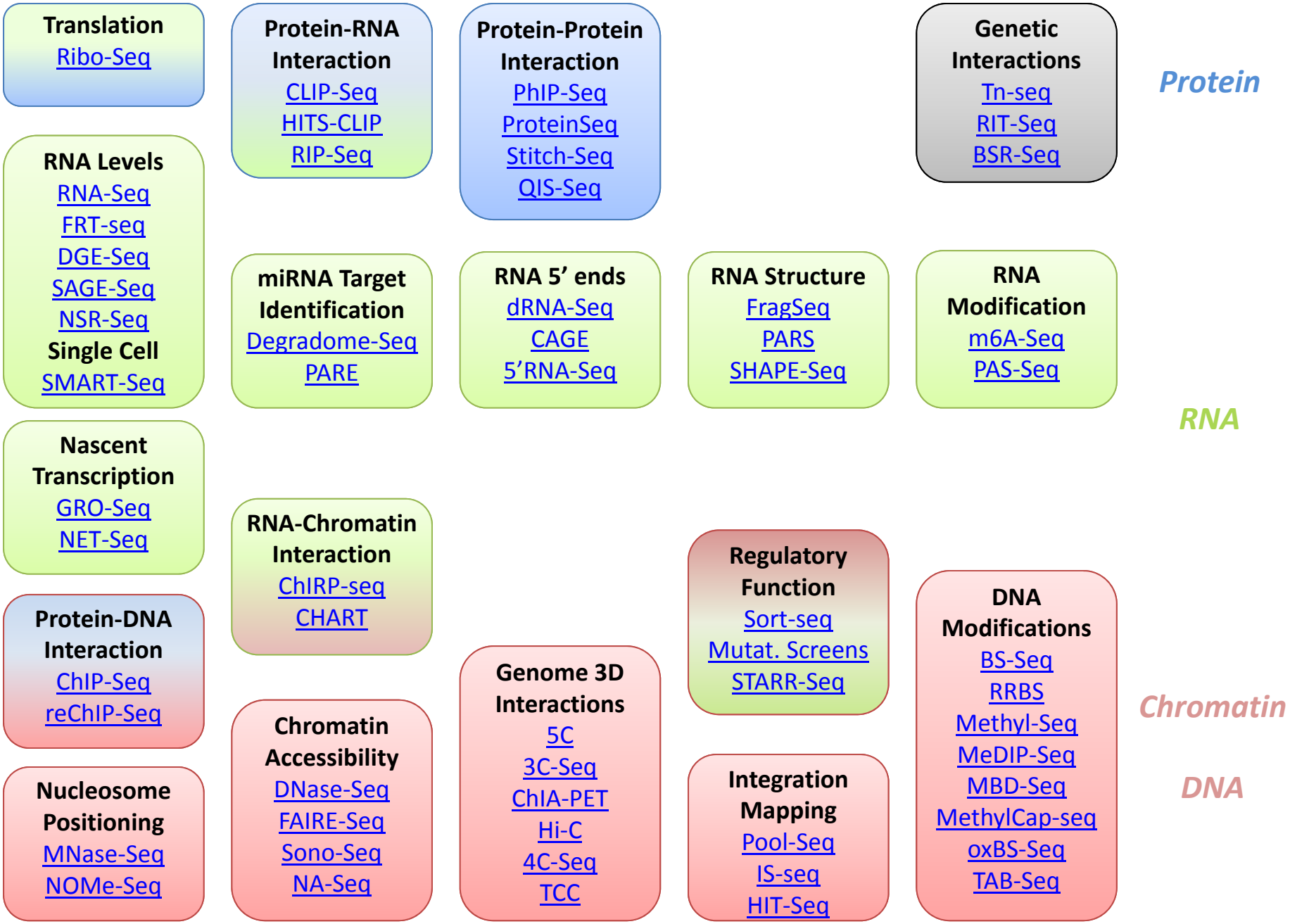
Experimental design and analysis strategies

Sven Heinz

UCSD

High-throughput sequencing - application characteristics

- **Genome sequencing - DNA-Seq** (assembly)
  - Coverage, long reads increase significance
- **Resequencing** (Reference sequence (genome/transcriptome) known, use data for mapping and counting/quantification)
  - High read counts increase significance
- **Tag sequencing** (use synthetic DNA barcodes for counting)
  - High read counts increase significance


Translation
Ribo-Seq
Protein-RNA Interaction
CLIP-Seq
HITS-CLIP
RIP-Seq
Protein-Protein Interaction
PhIP-Seq
ProteinSeq
Stitch-Seq
QIS-Seq
Genetic Interactions
Tn-seq
RIT-Seq
BSR-Seq
Protein
RNA Levels
RNA-Seq
FRT-seq
DGE-Seq
SAGE-Seq
NSR-Seq
Single Cell
SMART-Seq
miRNA Target Identification
Degradome-Seq
PARE
RNA 5' ends
dRNA-Seq
CAGE
5'RNA-Seq
RNA Structure
FragSeq
PARS
SHAPE-Seq
RNA Modification
m6A-Seq
PAS-Seq
RNA
Nascent Transcription
GRO-Seq
NET-Seq
RNA-Chromatin Interaction
ChIRP-seq
CHART
Regulatory Function
Sort-seq
Mutat. Screens
STARR-Seq
DNA Modifications
BS-Seq
RRBS
Methyl-Seq
MeDIP-Seq
MBD-Seq
MethylCap-seq
oxBS-Seq
TAB-Seq
Chromatin
Protein-DNA Interaction
ChIP-Seq
reChIP-Seq
Genome 3D Interactions
5C
3C-Seq
ChIA-PET
Hi-C
4C-Seq
TCC
Integration Mapping
Pool-Seq
IS-seq
HIT-Seq
DNA
Nucleosome Positioning
MNase-Seq
NOMe-Seq
Chromatin Accessibility
DNase-Seq
FAIRE-Seq
Sono-Seq
NA-Seq

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.

Overview

- Getting Started
  - Introduction to the ChIP-Seq protocol
  - Handling sequence data, quality control
  - Basic analysis pipeline:
    - Mapping
    - Visualization
    - Peak finding
    - Control experiments
    - Peak annotation
    - Motif Finding
    - GO enrichment
- Using ChIP-Seq
  - Example of a ChIP-Seq study: The PU.1 transcription factor
  - Planning experiments
  - Advanced methods/ modifications to ChIP-Seq

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, although some solid tissues can be hard to work with

What do you need for ChIP-Seq?

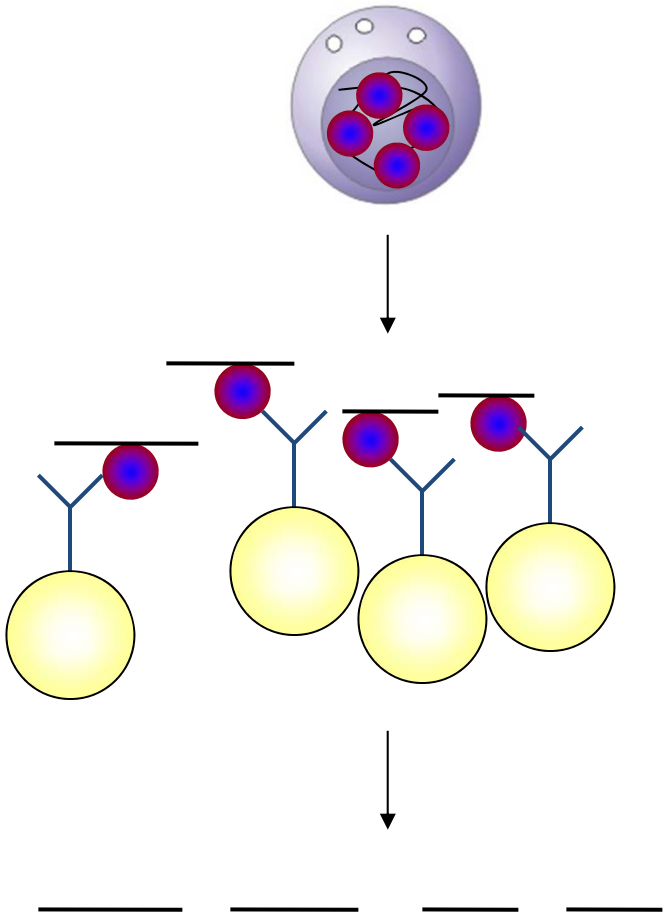
- **ChIP-grade antibody.**
  - Must be able to recognize its epitope under cross-linked conditions. (Antibodies that work in Western blot may not be adequate for ChIP)
  - Most histone modifications work well, but many are highly correlated…
    - Unless you study histone marks themselves, stick to well-working and well-accepted marks: H3K4me1/2/3, H3K27ac/me3, H3K36me3

Table 2 | Summary of ENCODE histone modifications and variants

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5′ end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3′ regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5′ end of genes
H4K20me1	Region	Preference for 5′ end of genes

ENCODE Project Consortium

Chromatin Immunoprecipitation (ChIP)



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

Immunoprecipitate

Isolate DNA

Quantifying DNA pulled down by ChIP

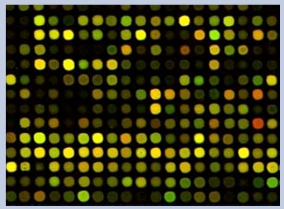
q-PCR

Use specific primers to quantify a region(s) of DNA using a PCR reaction

Controls: non-bound region in same ChIP, input-normalize, IgG

Pros: Fast (3hrs), high signal to noise

Cons: Very Biased, only check a limited number of regions (*very very big problem* – lacks internal controls other methods have)

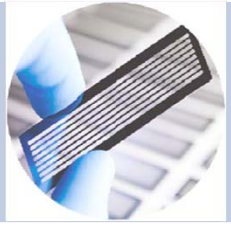


ChIP-Chip

Use microarrays printed with DNA from various regions, measure ChIP sample by labeling and hybridizing to the slide

Pros: Measure several regions, medium signal to noise

Cons: measurement by Hybridization is sloppy, somewhat biased (need to select regions a priori), analysis can be tricky



ChIP-Seq

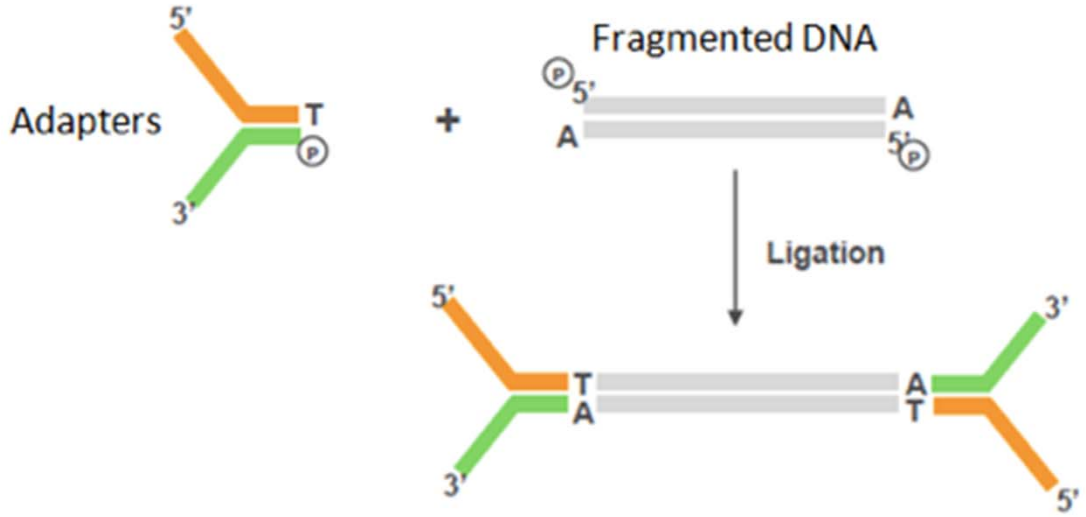
Directly measure ChIP sample using next-gen sequencing

Pros: Unbiased, genome-wide

Cons: Might need to sequence a lot to see all binding events.

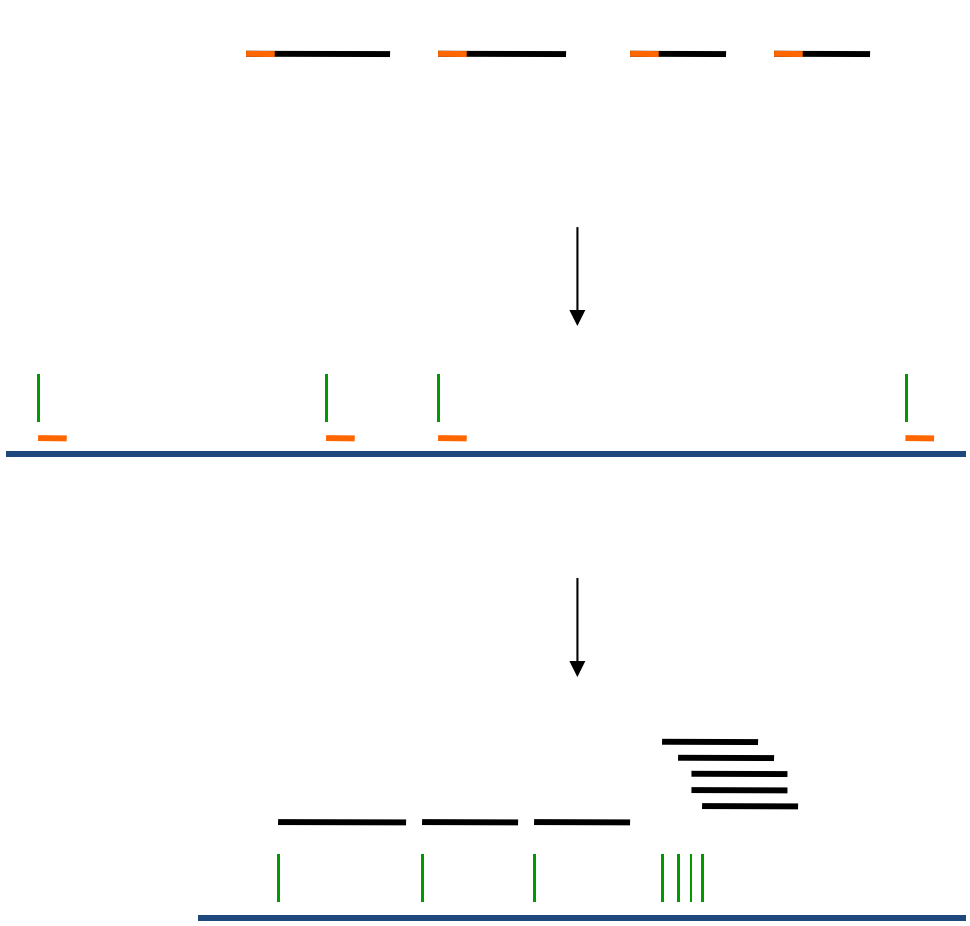
Creating a Sequencing Library

- Ligate general adapters to the ends of DNA fragments



- Amplify

ChIP-seq



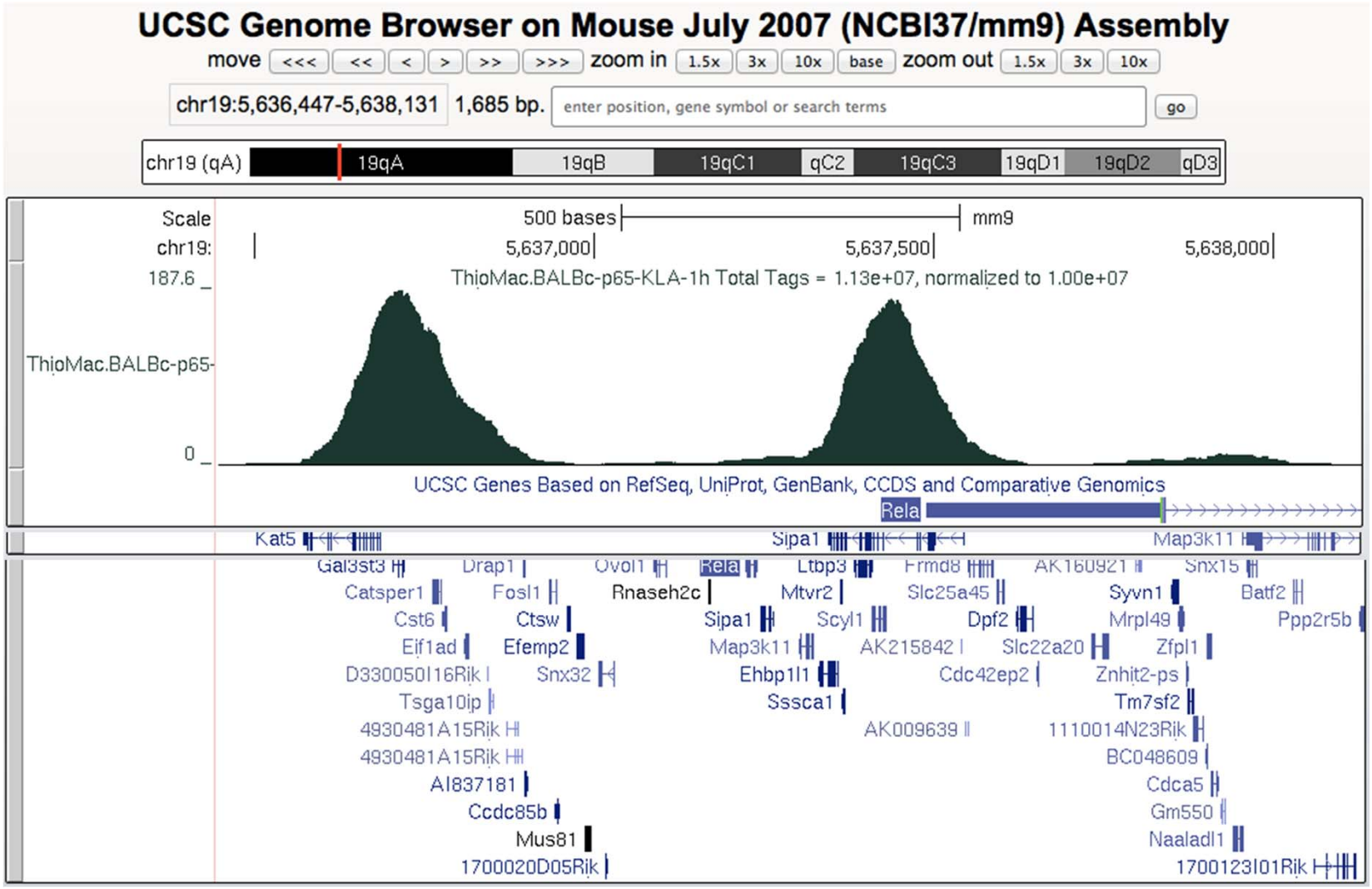
Sequence ChIPped DNA fragment ends (50-mer tags, up to 2 Billion tags/run)

Map tags to genome to obtain genomic positions

Visualize fragment positions

Downstream analysis

Genome Browser



http://genome.ucsc.edu

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

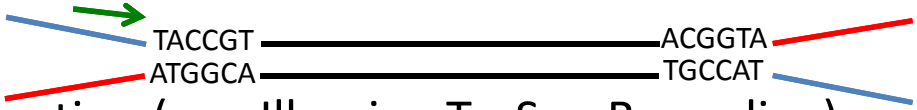
Bottom line:

Shorter reads provide more quantification per bp. Once a read can be placed unambiguously in the genome, additional nucleotides redundant information

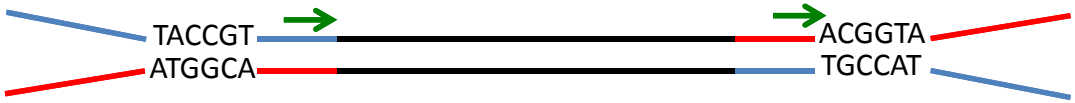
Long/Paired-end Reads increase the likelihood that you will cover sequence variants (i.e. SNPs), which is key when performing genetic studies or trying to identify allele specific binding. If a read doesn't contain a sequence variant, it can't be used to distinguish alleles.

Sequencing Depth and Barcoding

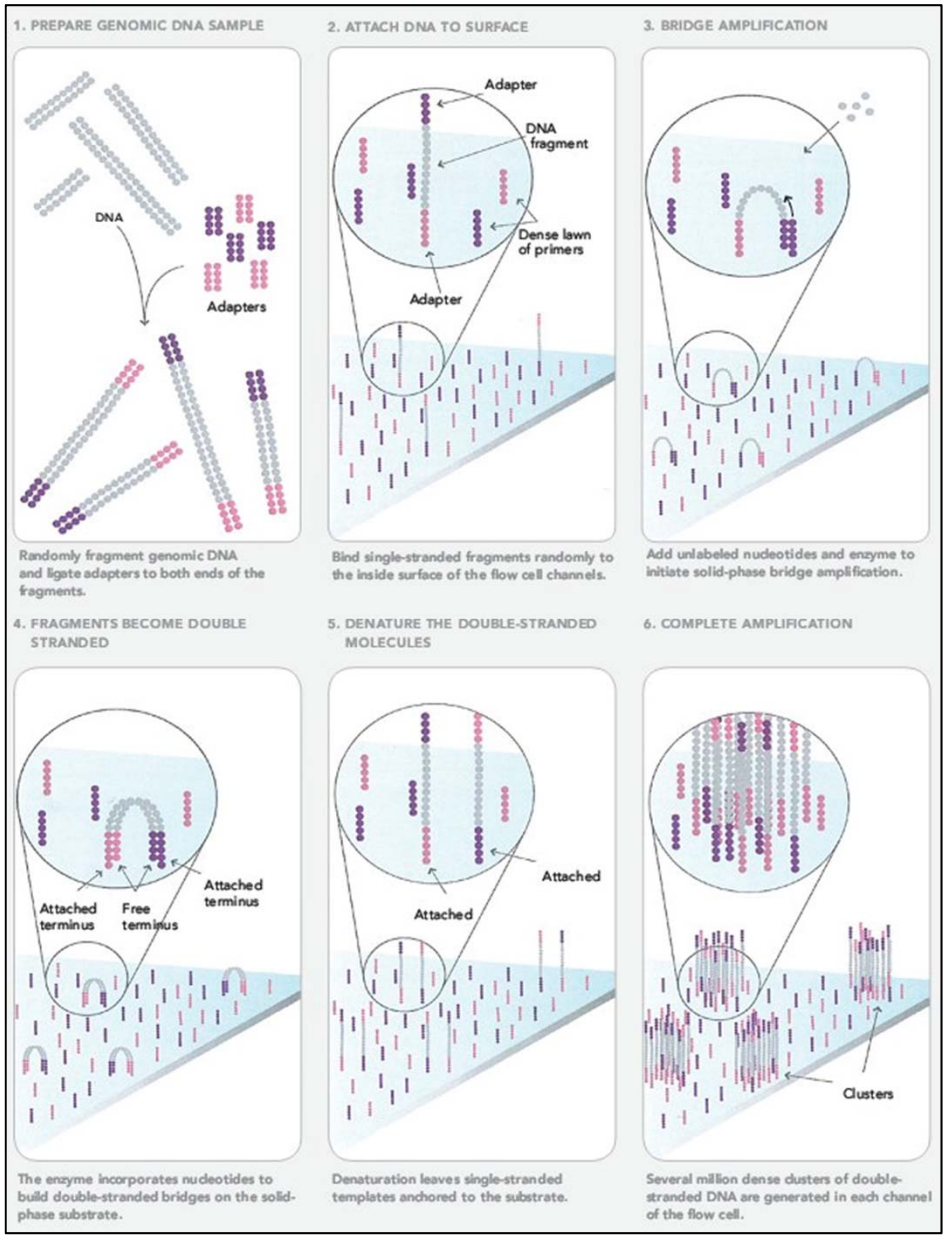
- 10-20 million reads can reveal can provide reasonable for a ChIP-Seq sample
  - May require more reads to identify rare binding events
- Modern Illumina Sequencers offer 200-300+ million reads per lane
- Solution: Multiplex Samples (Barcoding):
  - Simple solutions, such as adding nucleotide code to adapter - barcode is part of output sequence (i.e. first several nucleotides)
    - Problem: Barcode sequence can introduce sequence preferences for ligation and other bias

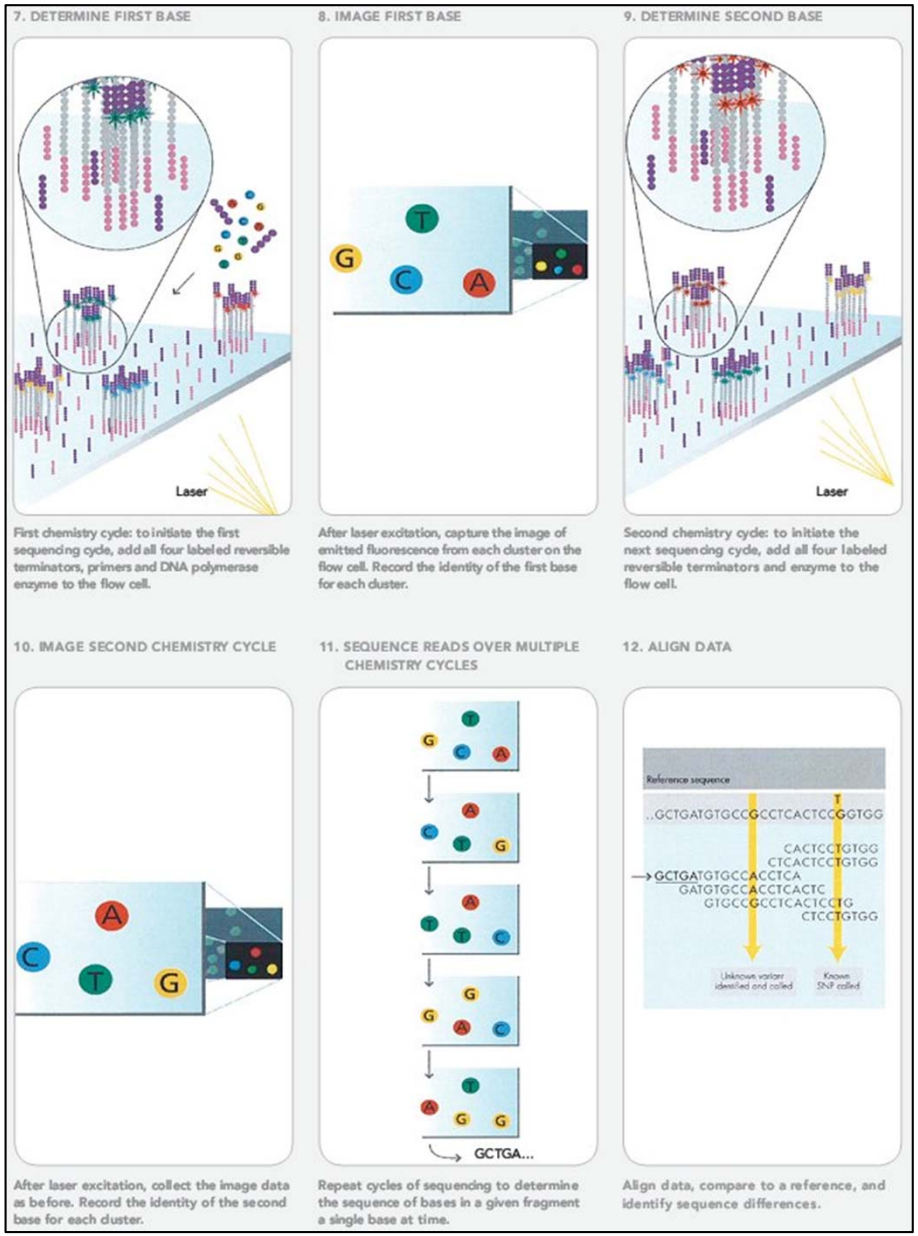


  - 2nd sequencing reaction (e.g. Illumina TruSeq Barcoding)
    - Barcode is Isolated inside the sequencing adapters avoiding bias due to the barcode sequence – a 2nd sequencing reaction is used to sequence the barcode and then match to the correct sample.



On the Sequencer…





Old Stuff from the Illumina website…

So you sent your sample for sequencing…
This is what you get back…

```
@HWI-ST647:108:D06B7ACXX:1:1101:7815:1954 1:N:0:
NCATCATCGAATCTTTGAACGCACATTGCGCCCTCTGGTATTCCAGAGGG
+
#4?DDFFFHHHHHJJJIJJJJJFGHIIIJJJJJJJJJJDHIJJJJJIJJI
@HWI-ST647:108:D06B7ACXX:1:1101:7860:1962 1:N:0:
NGAGAGGGTAAGGGACTAGGATGATAACAGGTGAGCCATTTGAGTCCCTA
+
#4=DDDFFFHHHHJIJJJJJJJJJJJIIJJJFHIJIJJJJJJJJIJJJJJ
@HWI-ST647:108:D06B7ACXX:1:1101:7914:1964 1:N:0:
NGATTGGCAGAAAAGTGACAATGGATTTTATTTCACTCATATTCAGATTT
+
#1=DFFFFHHHHHJJFHIJJIJJJJJJJJJJJJJJJJIIJJJJIJIJJJ
@HWI-ST647:108:D06B7ACXX:1:1101:7971:1987 1:N:0:
GAACCTGAATCCCGGACAGGTGCAGAGCCTGCCCCTGCCAACCCCACCCA
+
@@@FFFFFHHHHHJJBHIJJHIJGIGIJJJJJJJJIGJJJJJJJJIJJIA
@HWI-ST647:108:D06B7ACXX:1:1101:7820:1988 1:N:0:
AATTTGGCTGCTCATTACTTACGCAAATTTCTGCAGCTGGTTTTTCTCTT
+
BBCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJIJJJJJJDGIJJJJJJJ
```

About 1 billion lines of this stuff…

per lane

FASTQ file format

Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
32	20	040		Space	64	40	100	@	@	96	60	140	`	`
33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

```
@HWI-ST647:108:D06B7ACXX:1:1101:7815:1954 1:N:0:
NCATCATCGAATCTTTGAACGCACATTGCGCCCTCTGGTATTCCAGAGGG
+
#4?DDFFFHHHHHJJJIJJJJJFGHIIIJJJJJJJJJJDHIJJJJJIJJI
@HWI-ST647:108:D06B7ACXX:1:1101:7860:1962 1:N:0:
NGAGAGGGTAAGGGACTAGGATGATAACAGGTGAGCCATTTGAGTCCCTA
+
#4=DDDFFFHHHHJIJJJJJJJJJJJIIJJJFHIJIJJJJJJJJIJJJJJ
@HWI-ST647:108:D06B7ACXX:1:1101:7914:1964 1:N:0:
NGATTGGCAGAAAAGTGACAATGGATTTTATTTCACTCATATTCAGATTT
+
#1=DFFFFHHHHHJJFHIJJIJJJJJJJJJJJJJJJJJJIIJJJJIJIJJJ
@HWI-ST647:108:D06B7ACXX:1:1101:7971:1987 1:N:0:
GAACCTGAATCCCGGACAGGTGCAGAGCCTGCCCCTGCCAACCCCACCCA
+
@@@FFFFFHHHHHJJBHIJJHIJGIGIJJJJJJJJJIGJJJJJJJJIJJIA
@HWI-ST647:108:D06B7ACXX:1:1101:7820:1988 1:N:0:
AATTTGGCTGCTCATTACTTACGCAAATTTCTGCAGCTGGTTTTTCTCTT
+
BBCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJIJJJJJJDGIJJJJJJJ
```

- NOTE: quality scores can be encoded differently by different vendors/software versions

- Usually have extension *.fq, *.fastq, or *_sequence.txt

$Q_{Sanger} = -10 \log_{10} p$; Q = 10 -> p = 0.1
Q = 20 -> p = 0.01
Q = 30 -> p = 0.001

p = probability of wrong base call

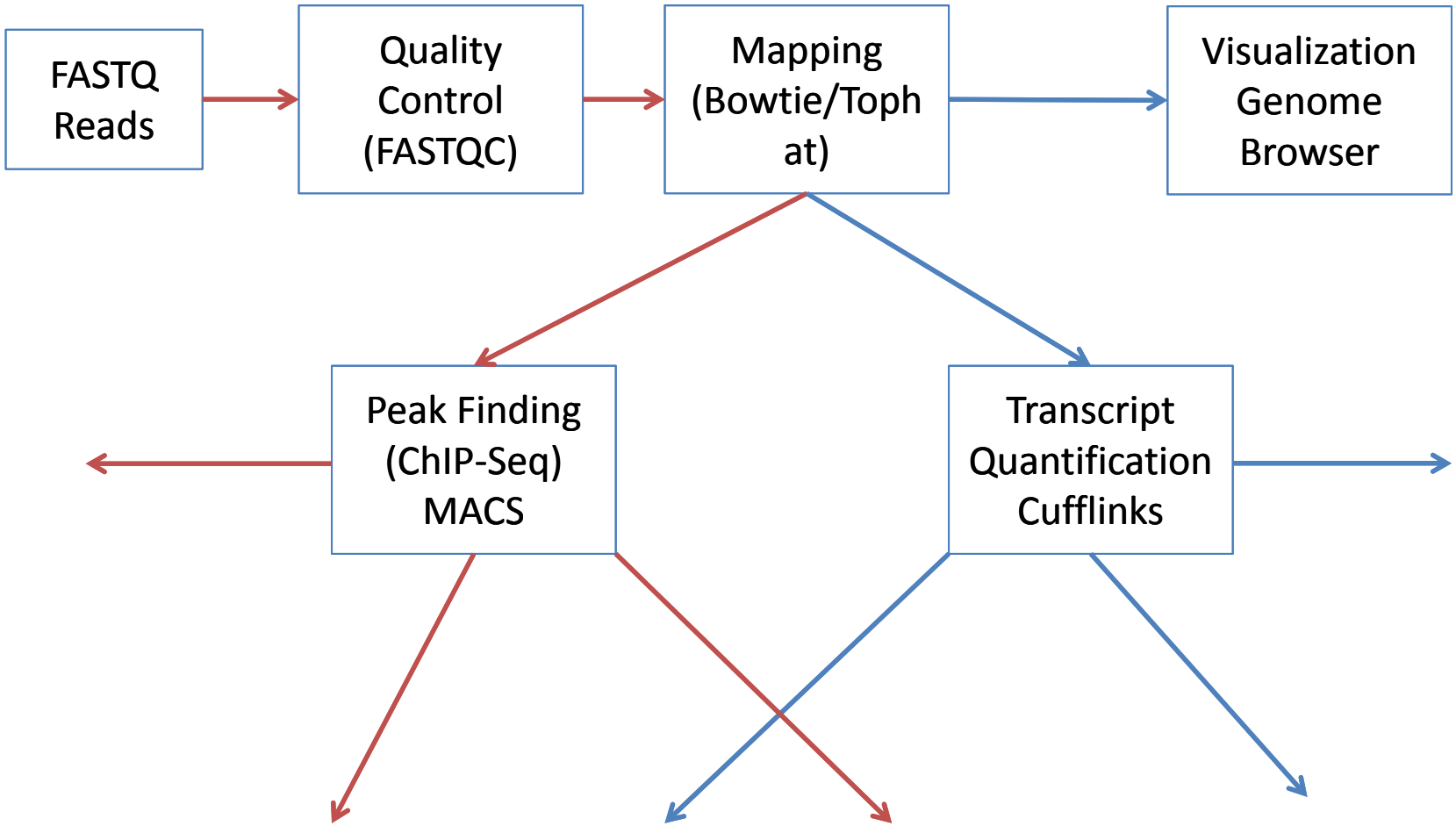
Raw data and backups

- The **FASTQ** sequence files constitute the Raw data from a sequencing experiment

  SAVE these and ensure they are backed-up!!!

- When it comes time to publish your study, you may be required to provide them.
  - Everything you do from this point on will manipulate the data one way or another
  - As processing techniques improve, genome versions update, etc., it's nice to able to revisit old experiments and reprocess them to compare them to new experiments in a safe and fair manner
- Most Journals want you to submit your data to NCBI GEO and/or the NCBI SRA (short read archive)
  - They generally want the **FASTQ** files
- NOTE: The raw data alone from a sequencing run can take up close to 50-100 Gb, zipped!
- If you have human patient samples, take note of precautions you might need to follow to protect identity

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

Bowtie – fast, good for 20-75 bp

Eland – Old Illumina alignment program

BLAT - Very long reads 454

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

(RNA-Seq: Tophat – spliced alignment)

Peak Finding

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

MACS – 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

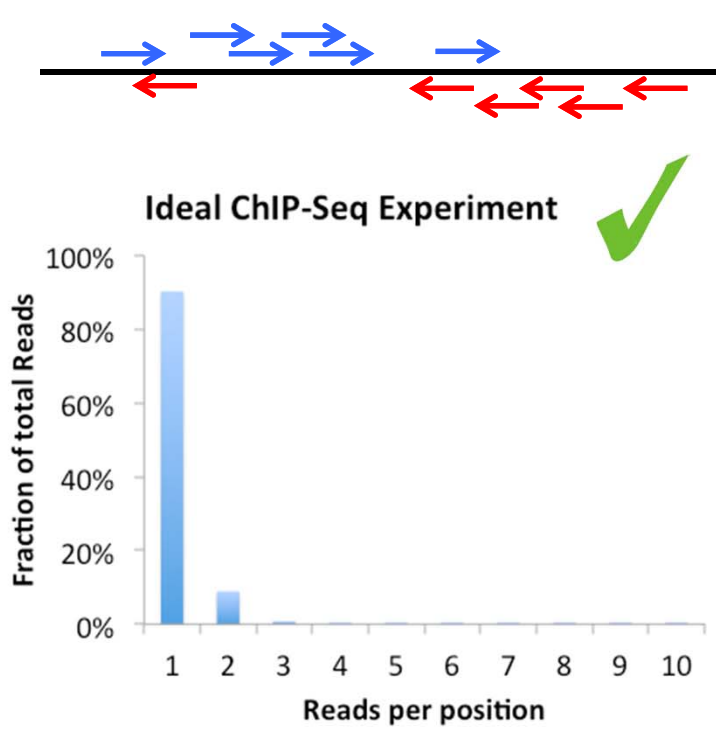
Lots of stuff in R/Bioconductor

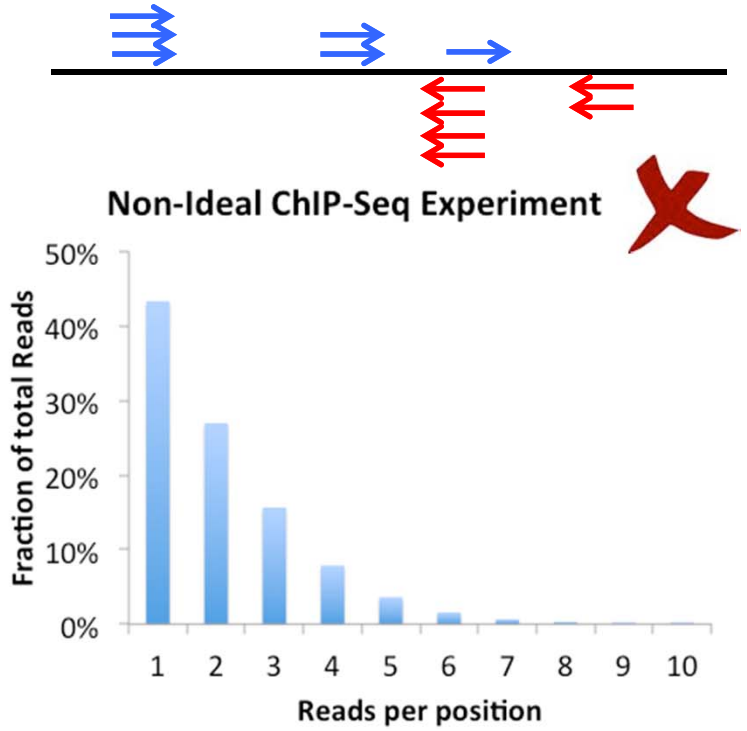
Read Alignment: Coordinate Systems

- 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

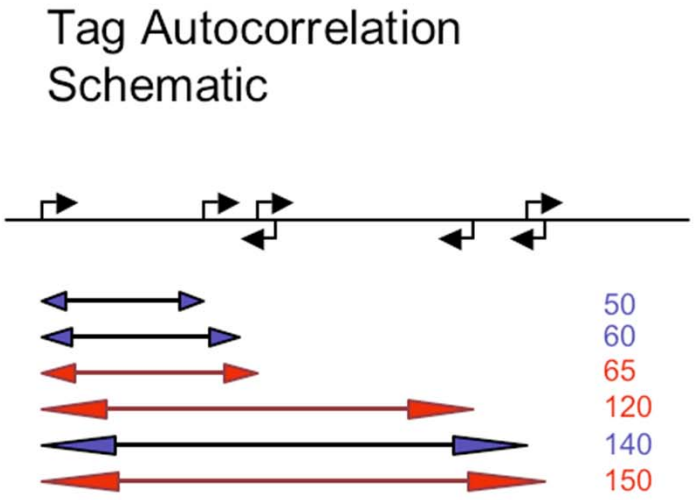


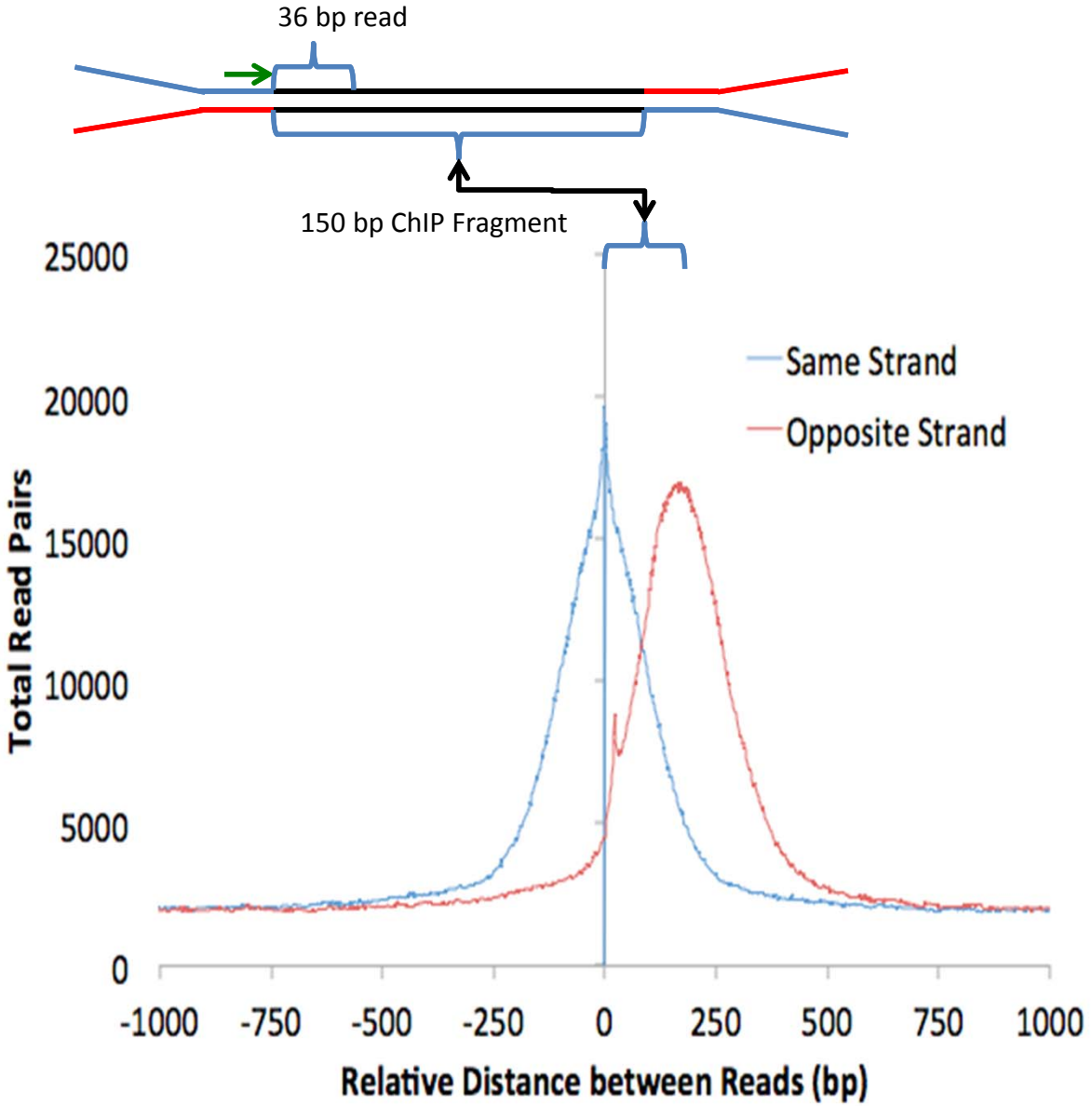


Quality Control:

Autocorrelation

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

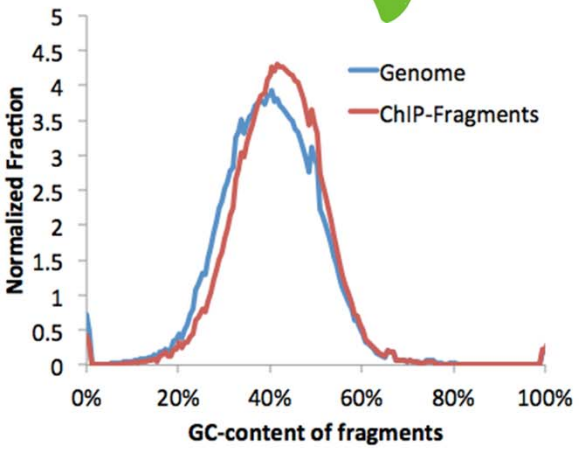




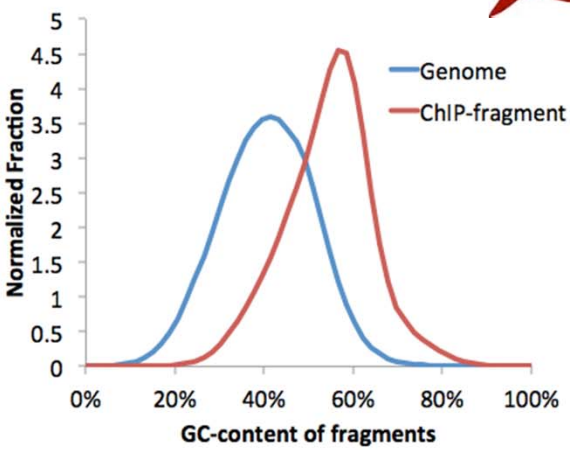
Quality Control:

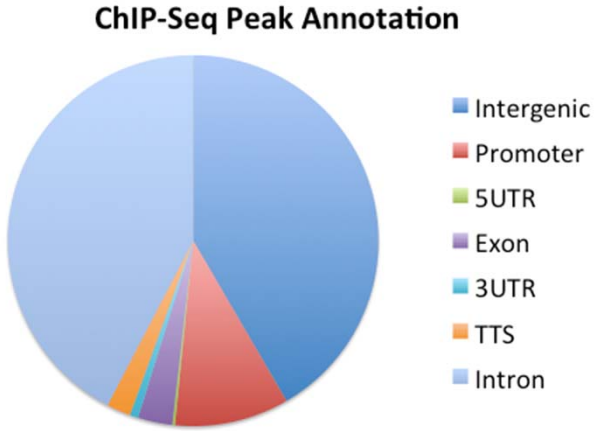
GC% Fragment Bias

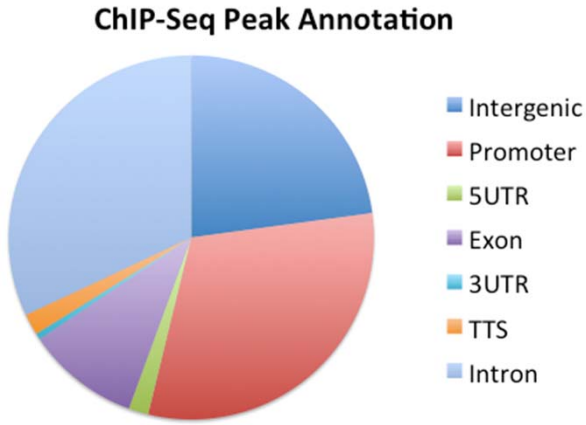
Ideal Experiment



Non-Ideal Experiment

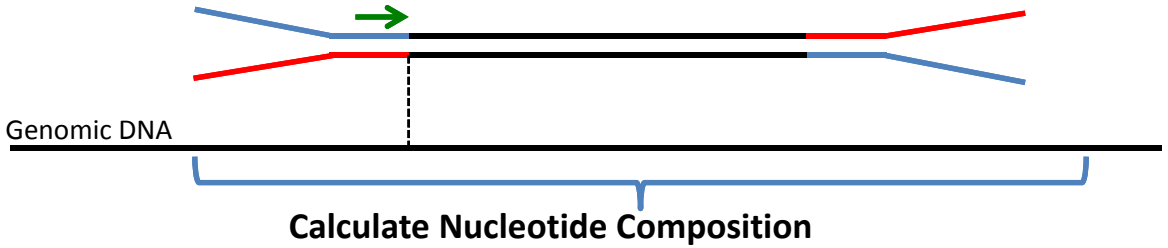


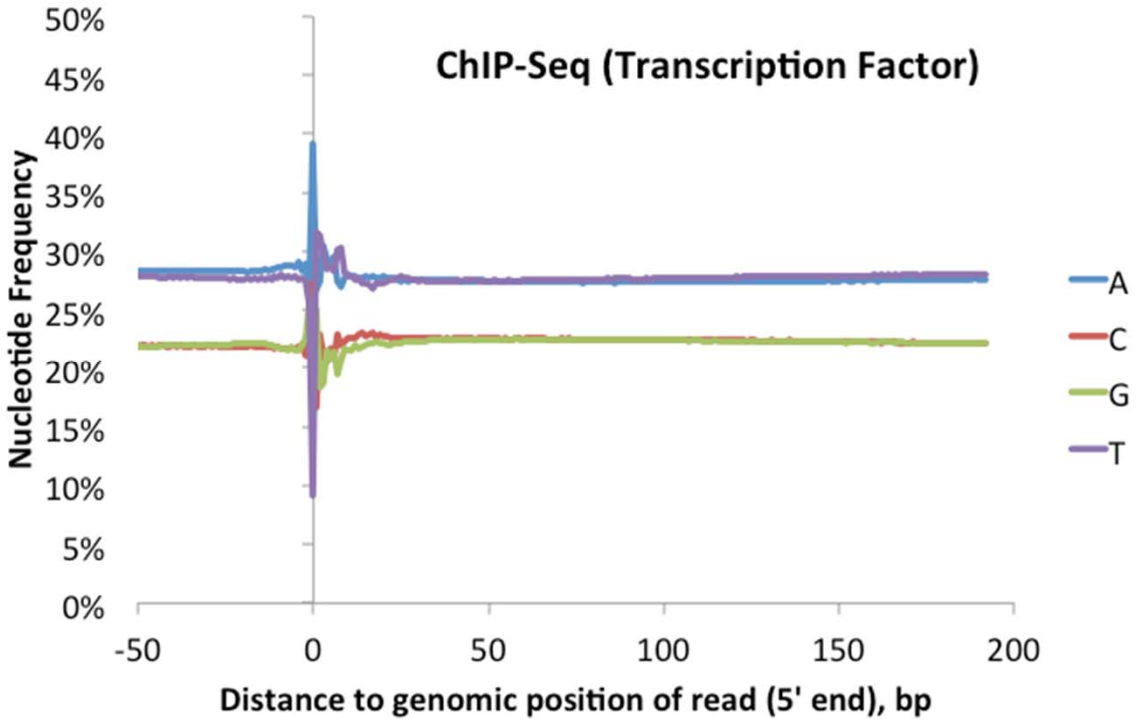




Quality Control:

Position-specific Nucleotide Composition Bias

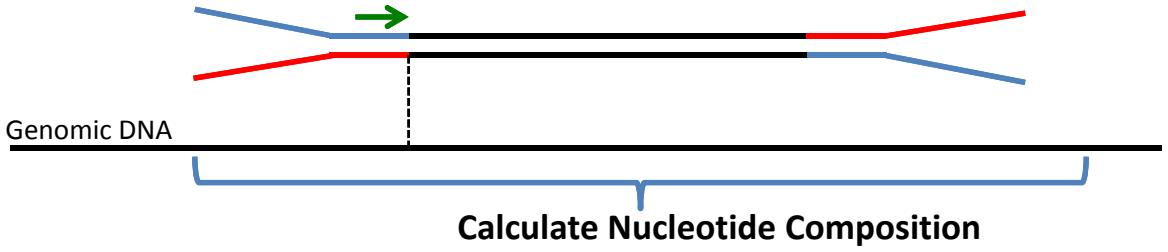


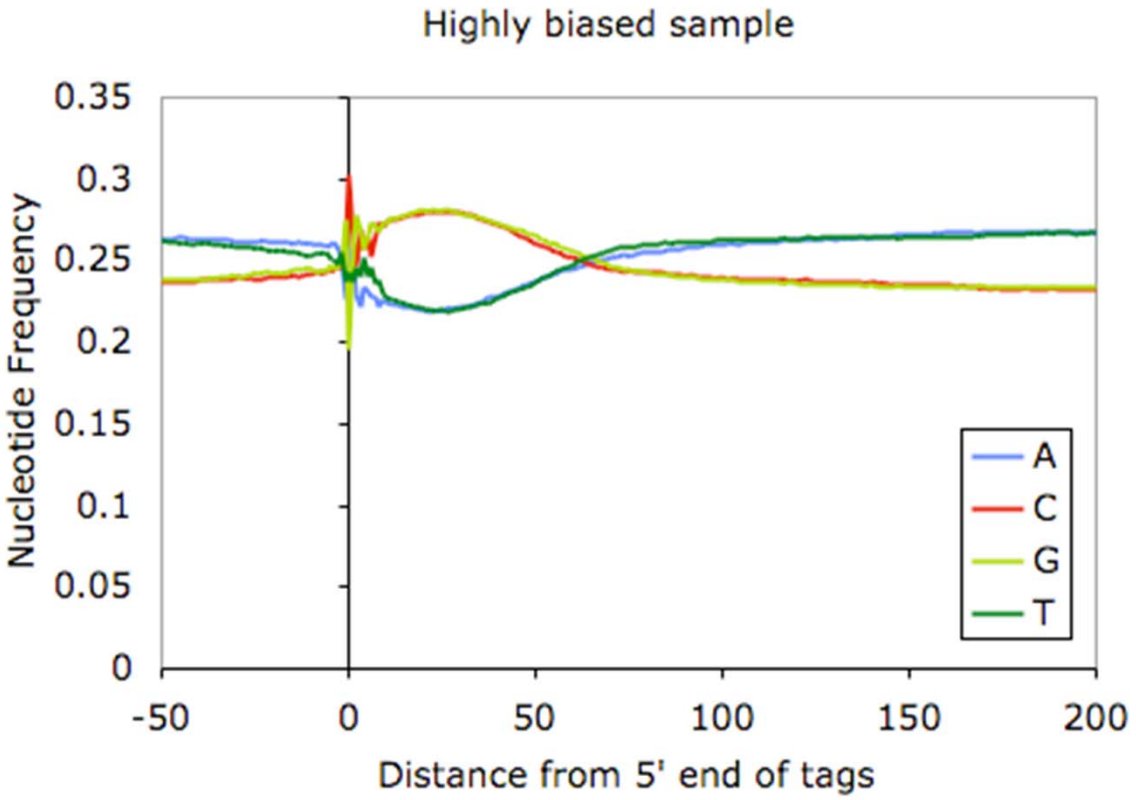


Reveals bias in enzymatic manipulation of sample

Quality Control:

Position-specific Nucleotide Composition Bias

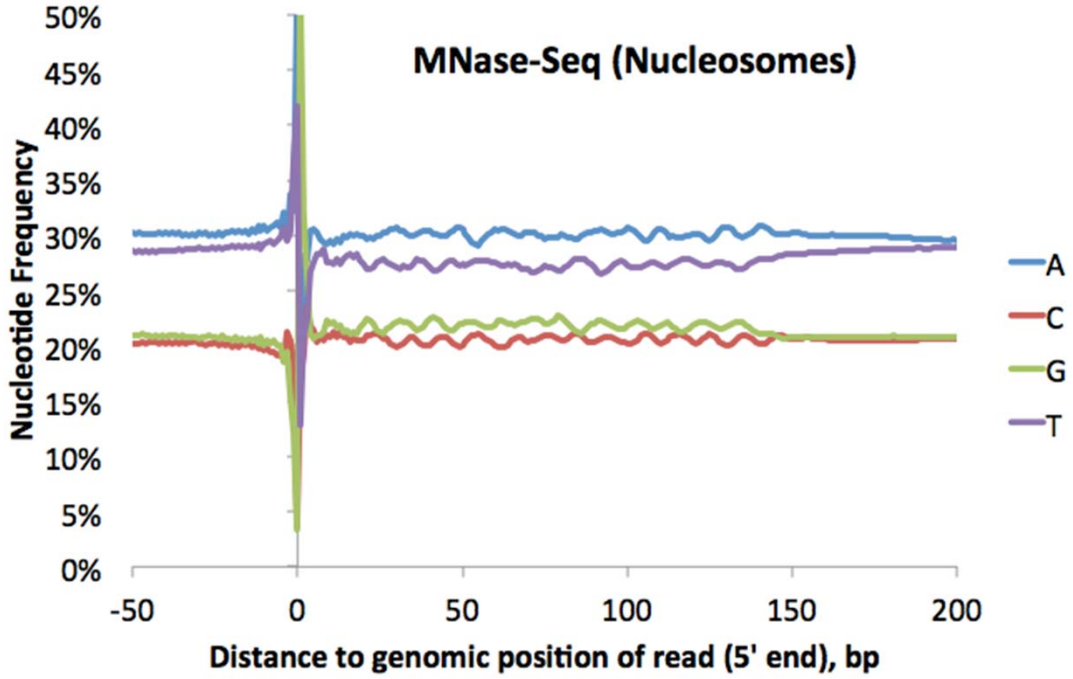


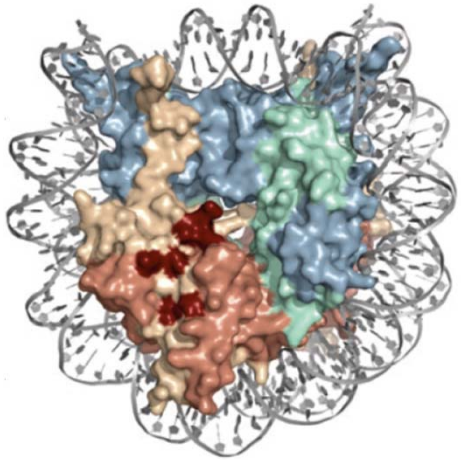


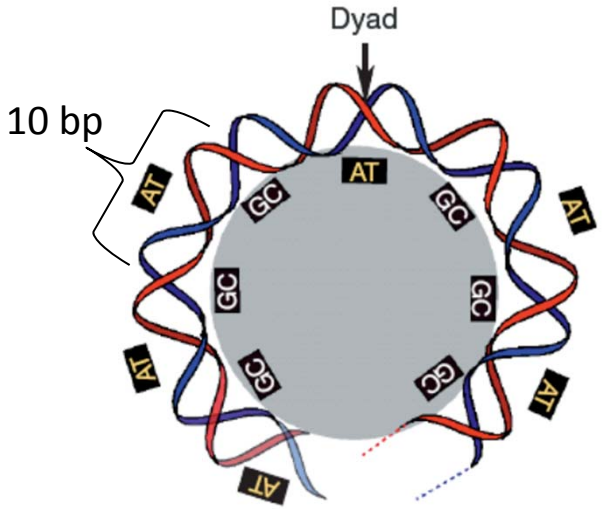
Reveals bias in enzymatic manipulation of sample

Quality Control:

Position-specific Nucleotide Composition Bias







Traditional Nucleosome Positioning Sequences

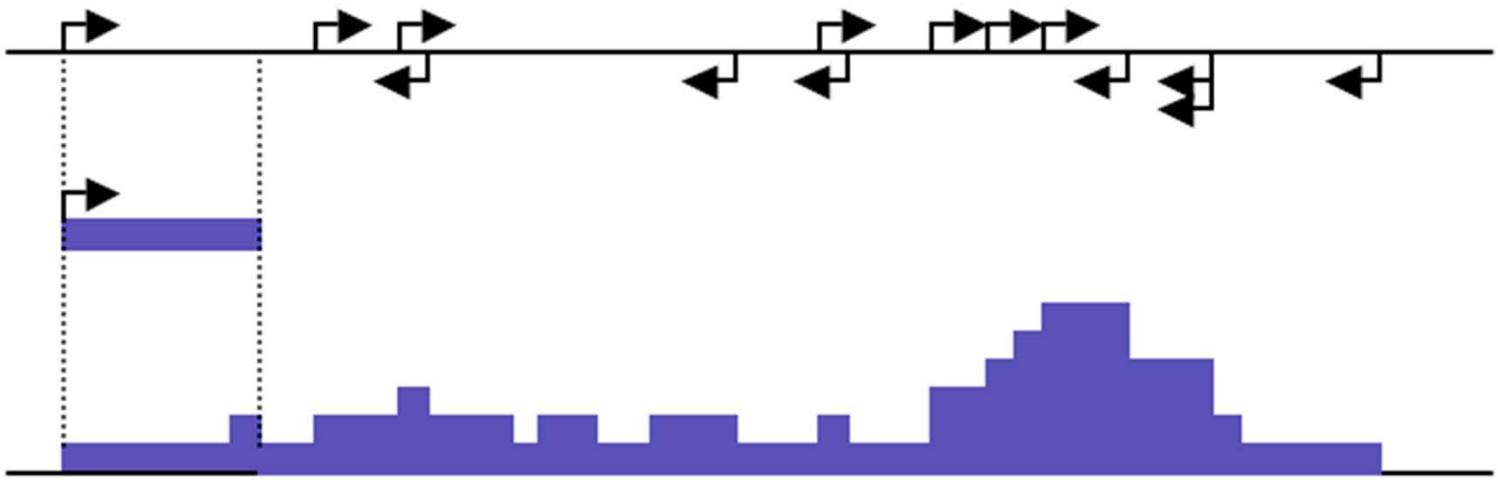
(From Albert et al. Nature Nature 446, 572-576 (29 March 2007)

Visualization of Data

- UCSC Genome Browser
  - UCSC GB integrates a very large number of diverse data sets and allows you to visualize them across the genome
  - You can upload your own sequencing experiments to the genome browser
- Other Genome Browsers / Data Viewers are out there as well (e.g. IGV).
  - 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

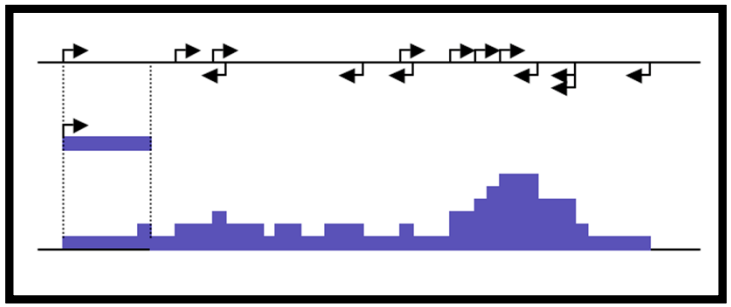
Creating Tag Pileups

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



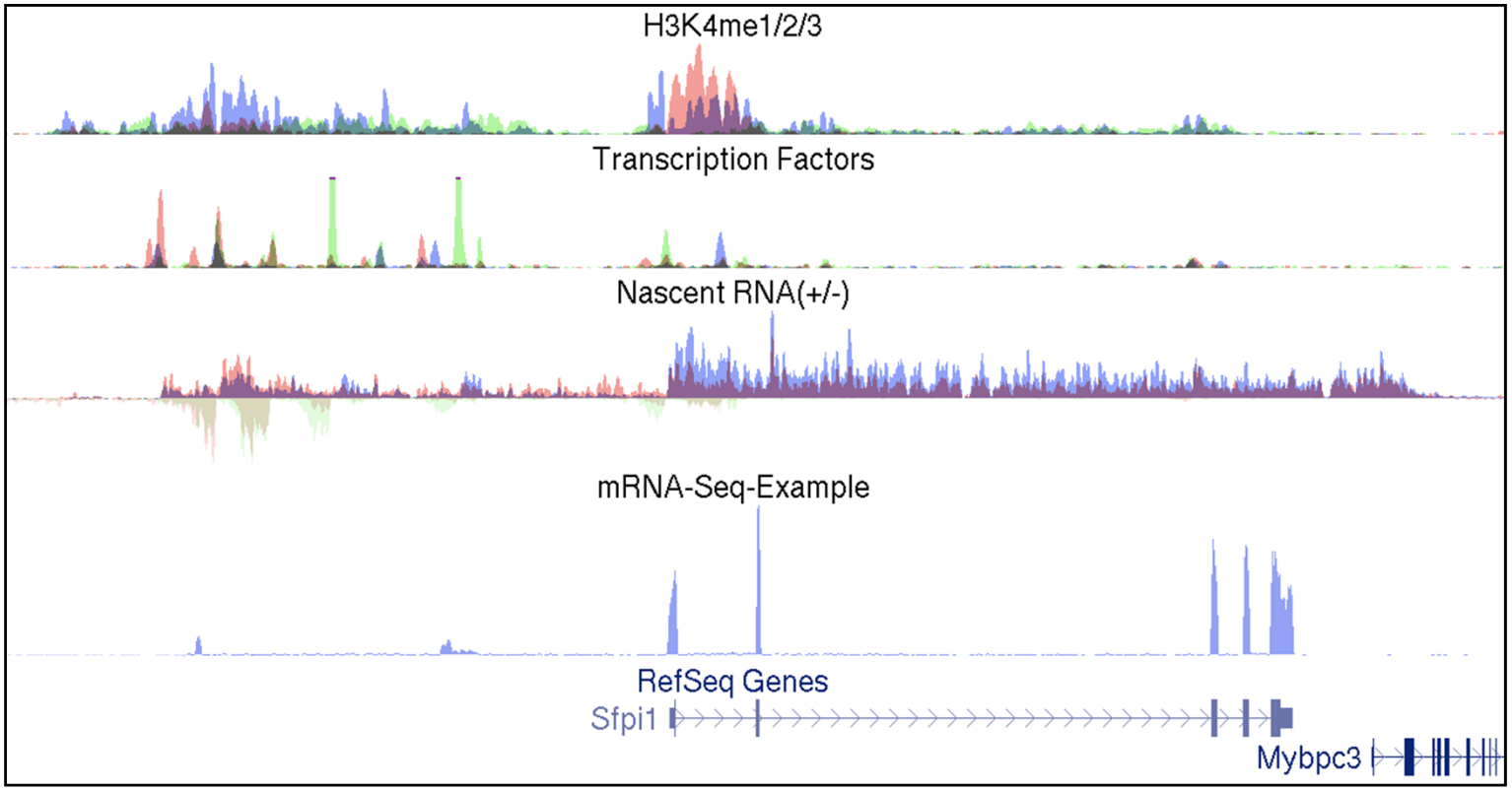
Quality Control:

Read Density Visualization



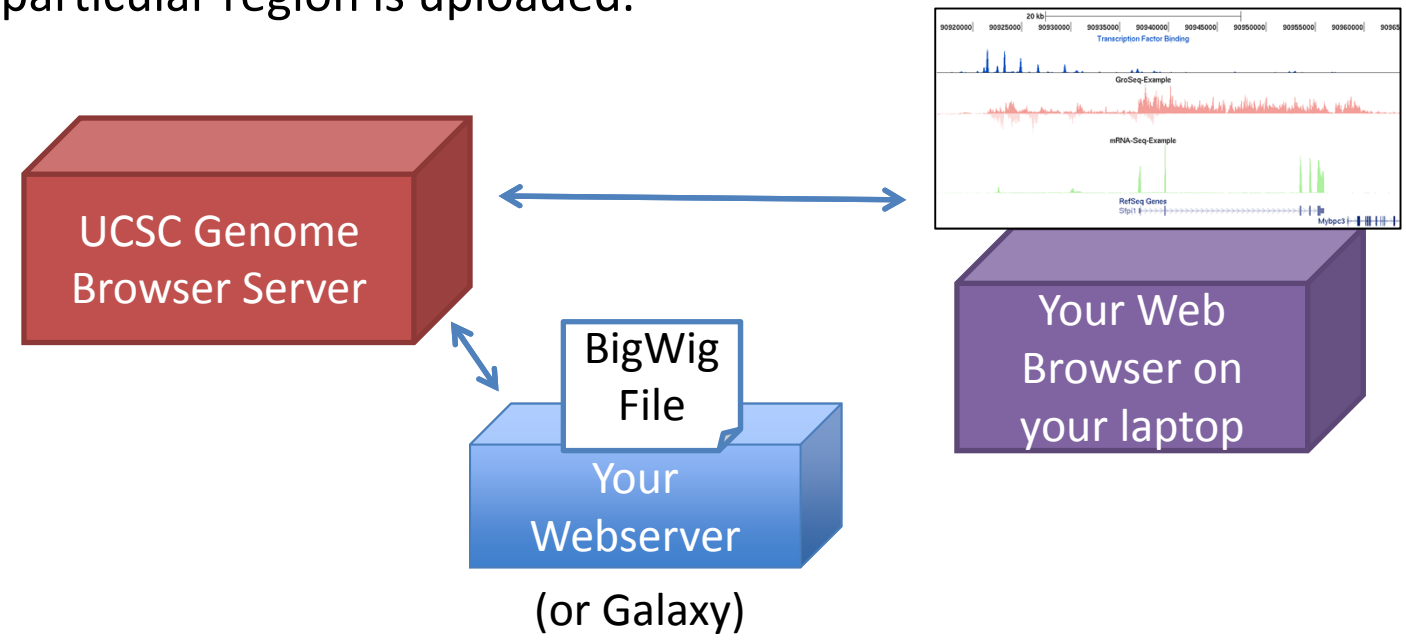
Raw read fragments are piled up to produce read densities and visualized using **UCSC Genome Browser**

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



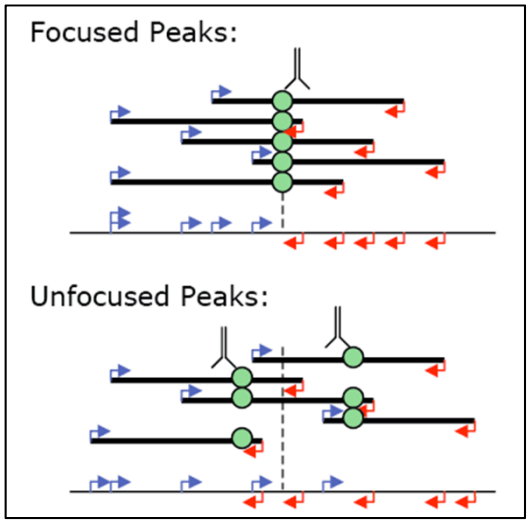
BigBed/BigWig Files

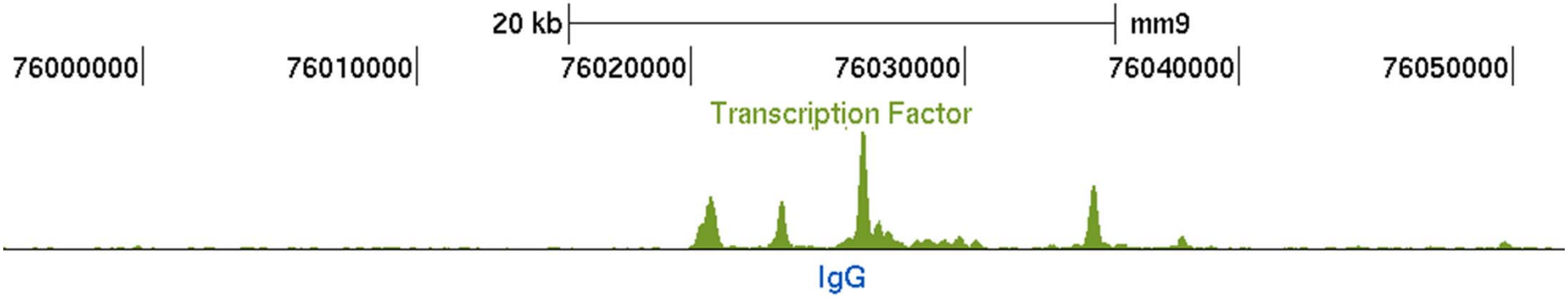
- Quick NOTE: bigWig/bigBed Files
  - There is a practical limit to how much custom data can be loaded onto UCSC at a time.
  - In order to upload full experiments, UCSC has implemented tools that allow you to host indexed data files on your own webserver. These files are then accessed by UCSC, and only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

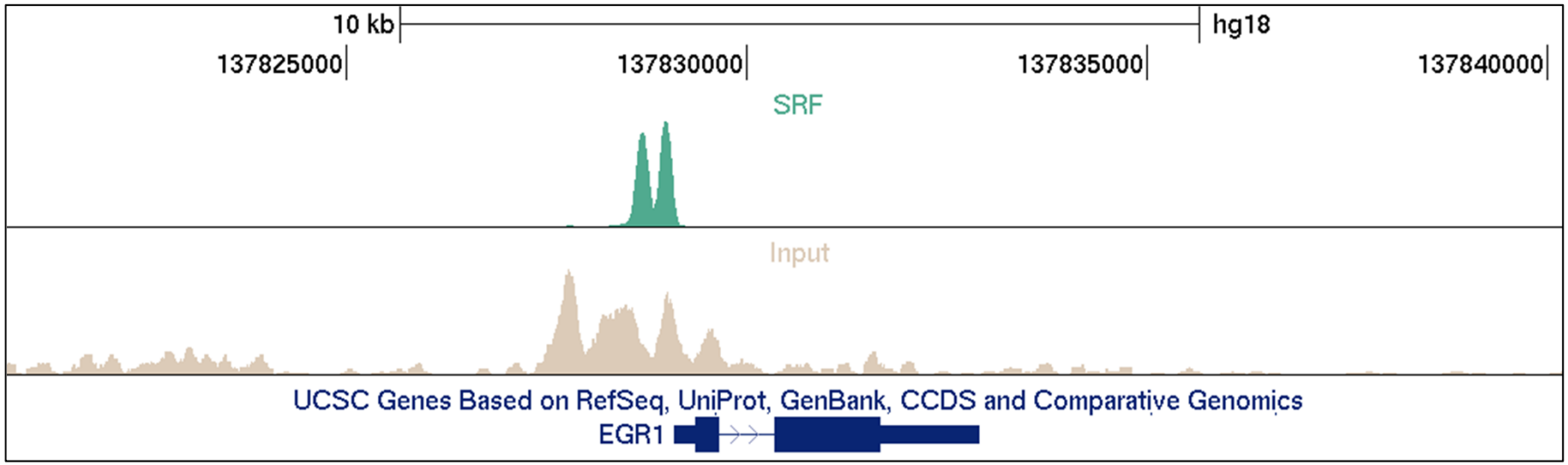


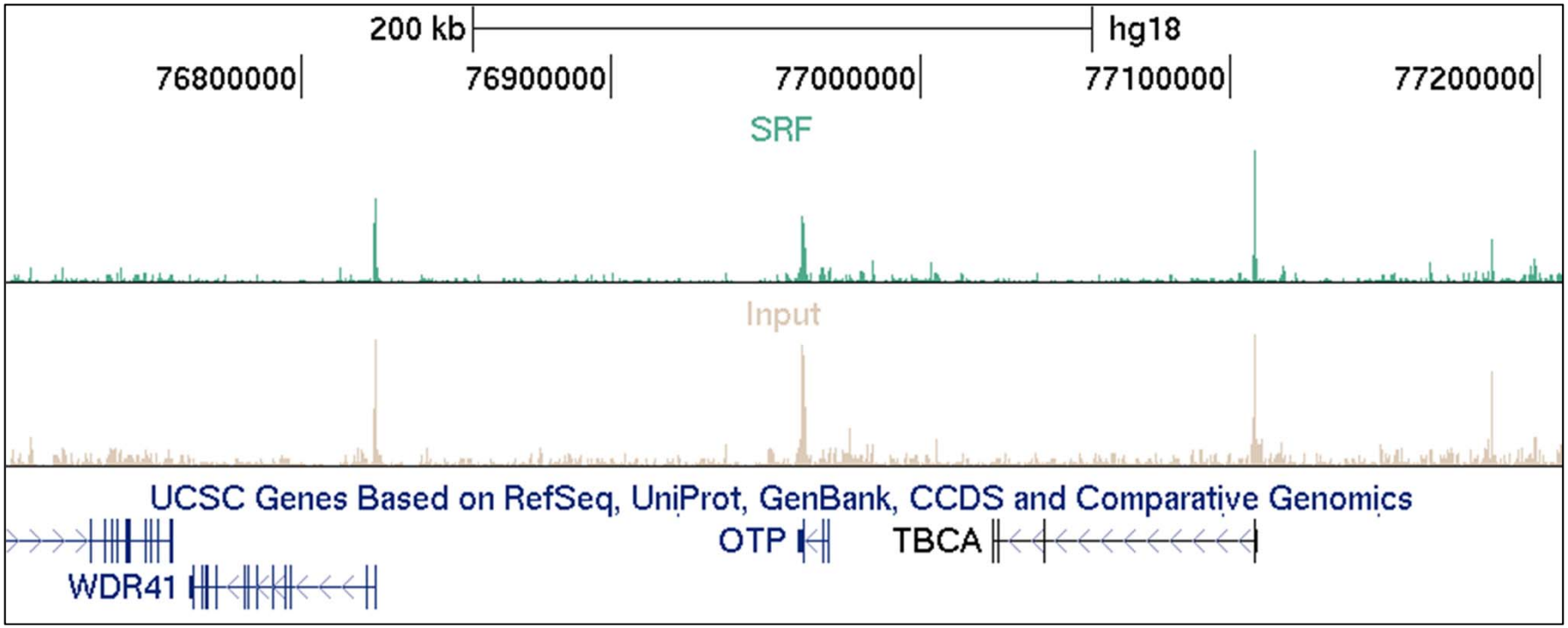


Problems with a simple approach to Peak finding

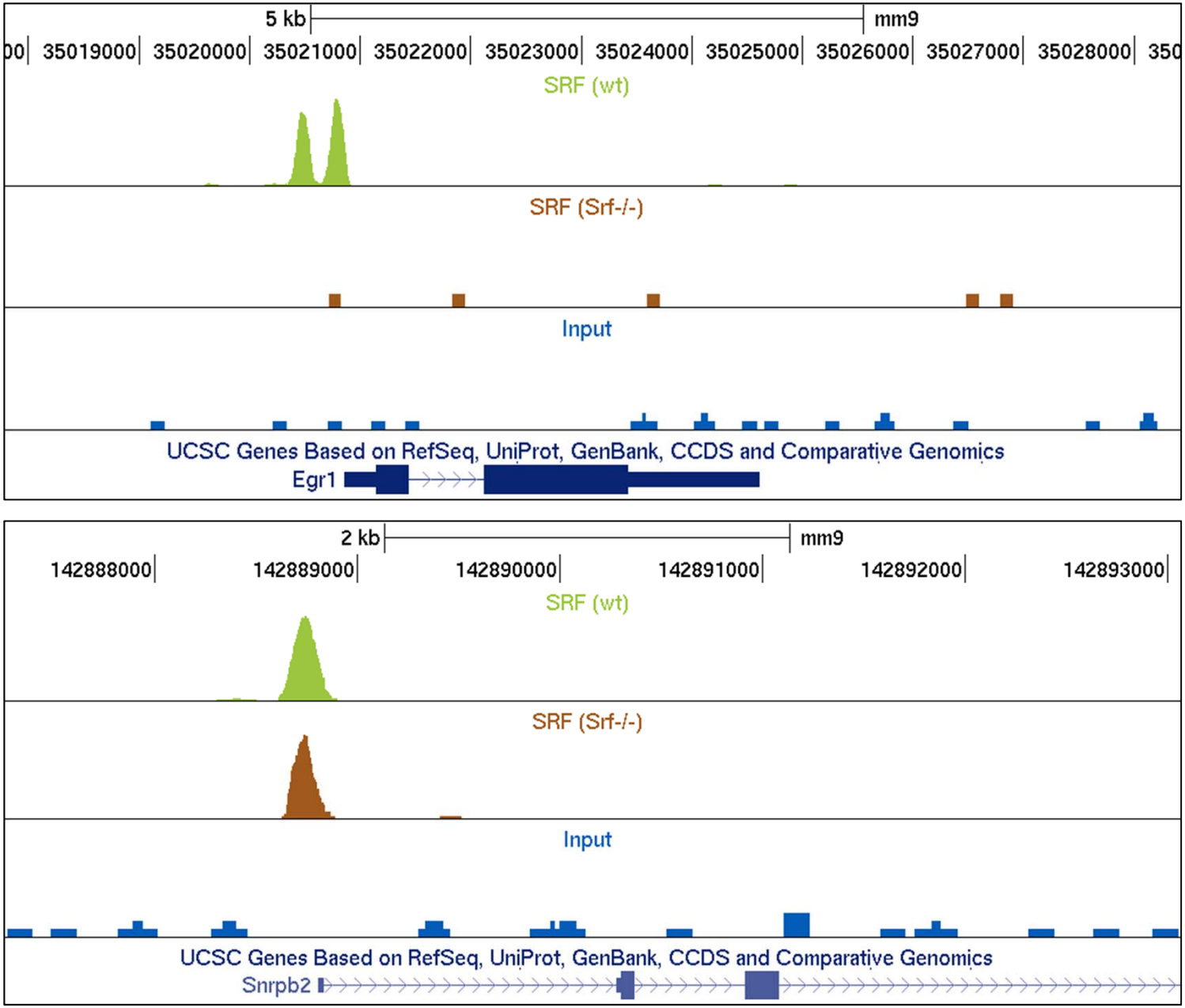
- Random distribution of reads is not a realistic control!
  - Read alignment bias (repeats are hard to map reads to)
  - Sample genome (i.e. cell line) is different from the reference genome (copy number variants, etc.)
  - ChIP protocol introduces bias
    - enzymatic preference for specific sequences/adapter 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

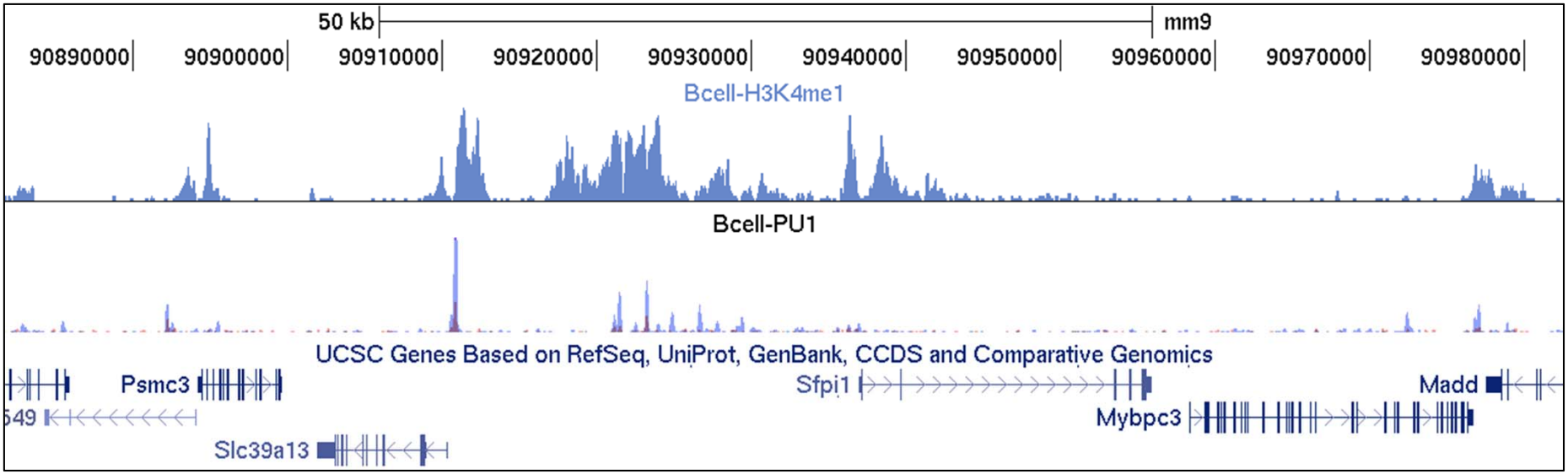

200 kb
hg18
76800000
76900000
77000000
77100000
77200000
SRF
Input
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
OTP
TBCA
WDR41

ChIP-Seq for SRF in SRF-/- Mice

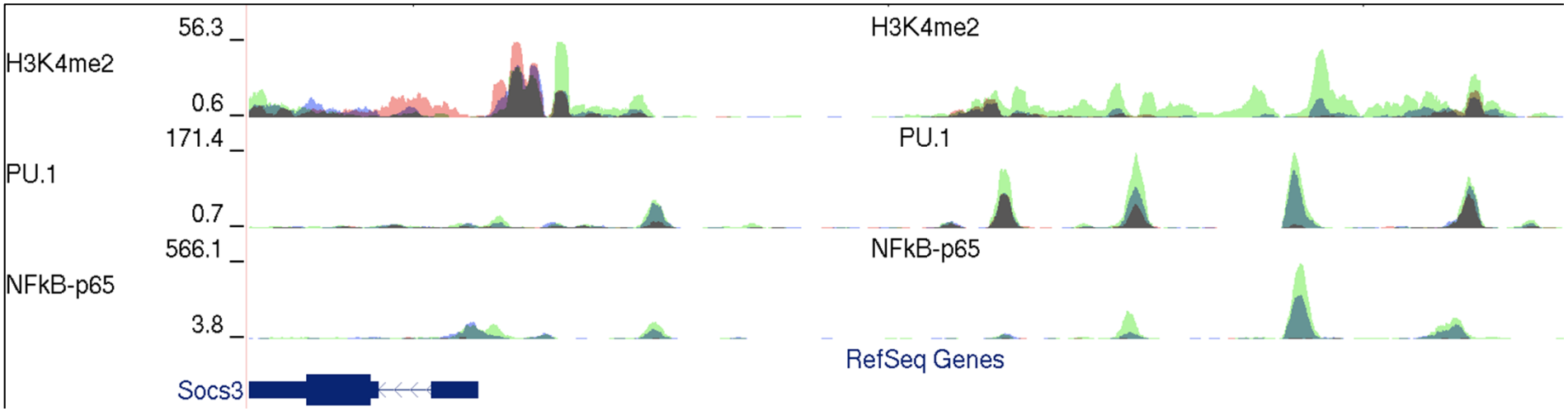


Histone Modifications

Generally want to define broad domains with continual signal enrichment

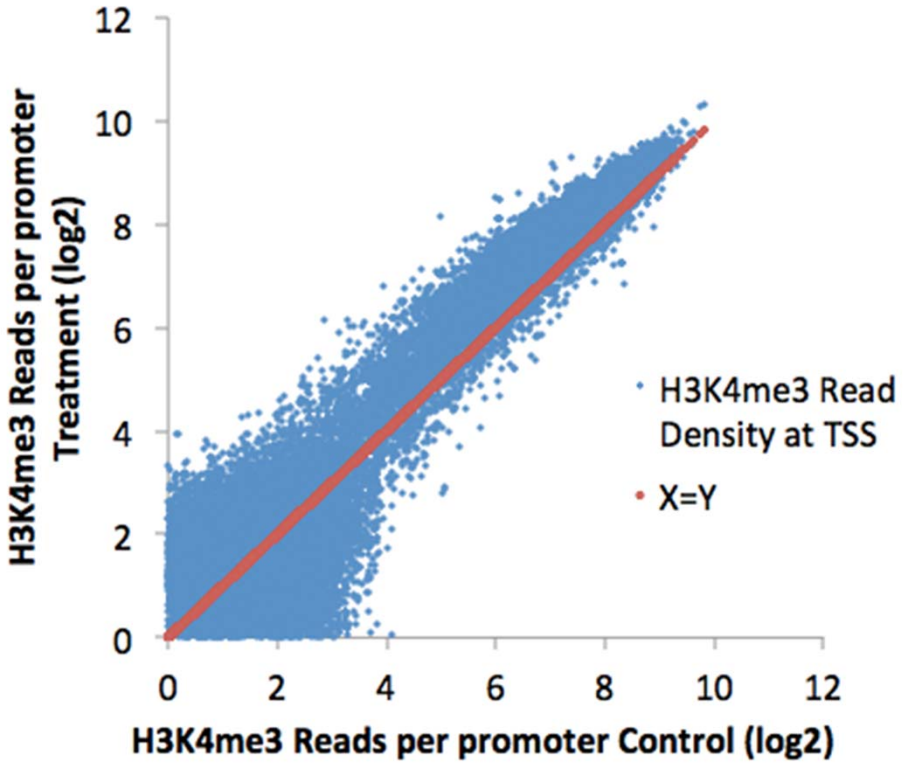


If Mnase is used for histone modifications, it can be possible to identify individual nucleosomes and nucleosome free regions:

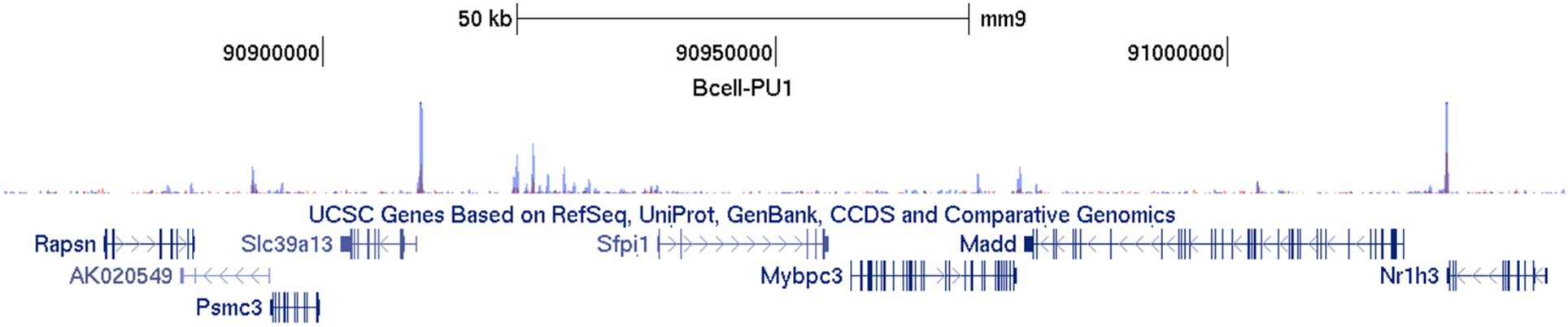


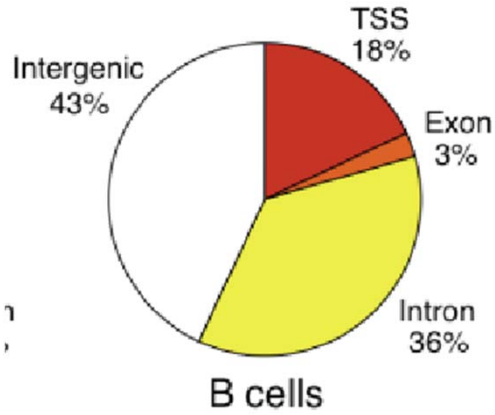
Comparing Experiments, Normalization

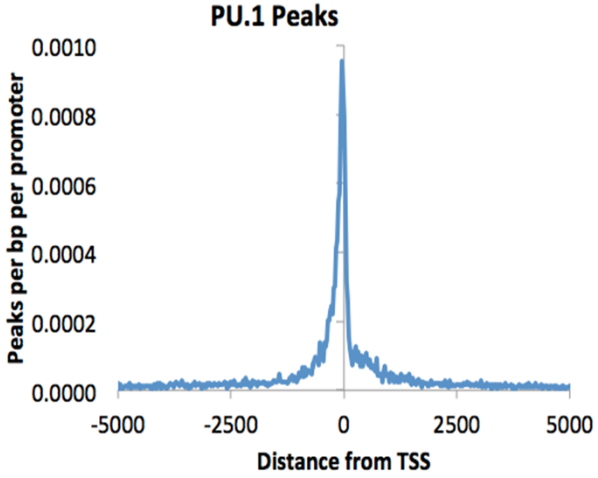
- Need to control for sequencing depth:
  - Normalize to total mapped tags
  - Normalize for IP efficiency – difficult! Need to make assumptions about regions that should have equal coverage
- Differential Peaks:
  - Consider the number of reads at a given genomic locus in each separate experiment
  - 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

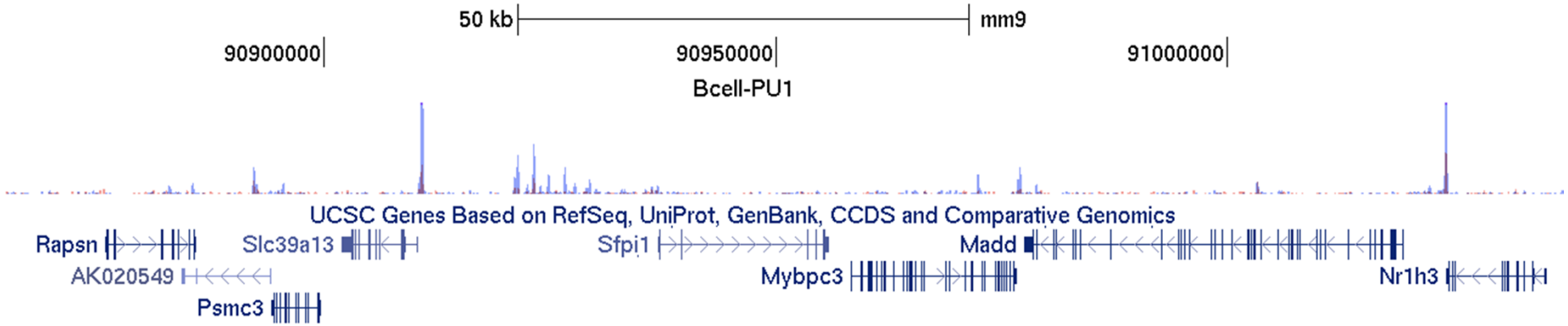
GO Molecular Function

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

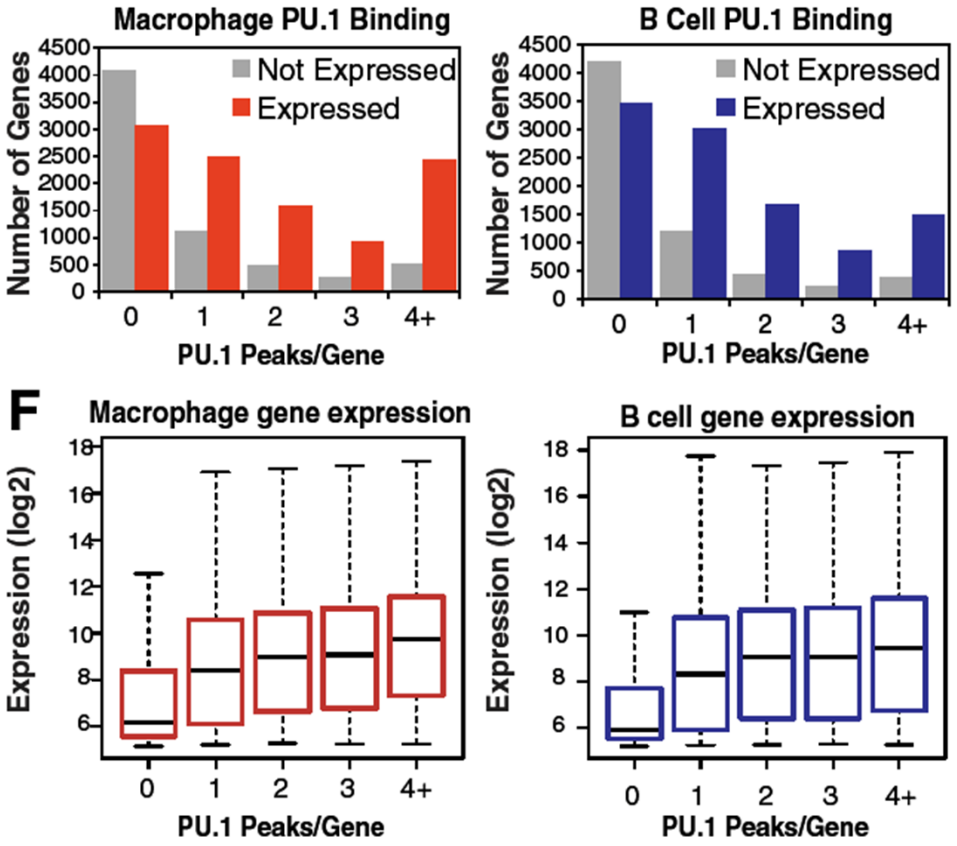
Assigning ChIP-Seq Peaks to Target Genes

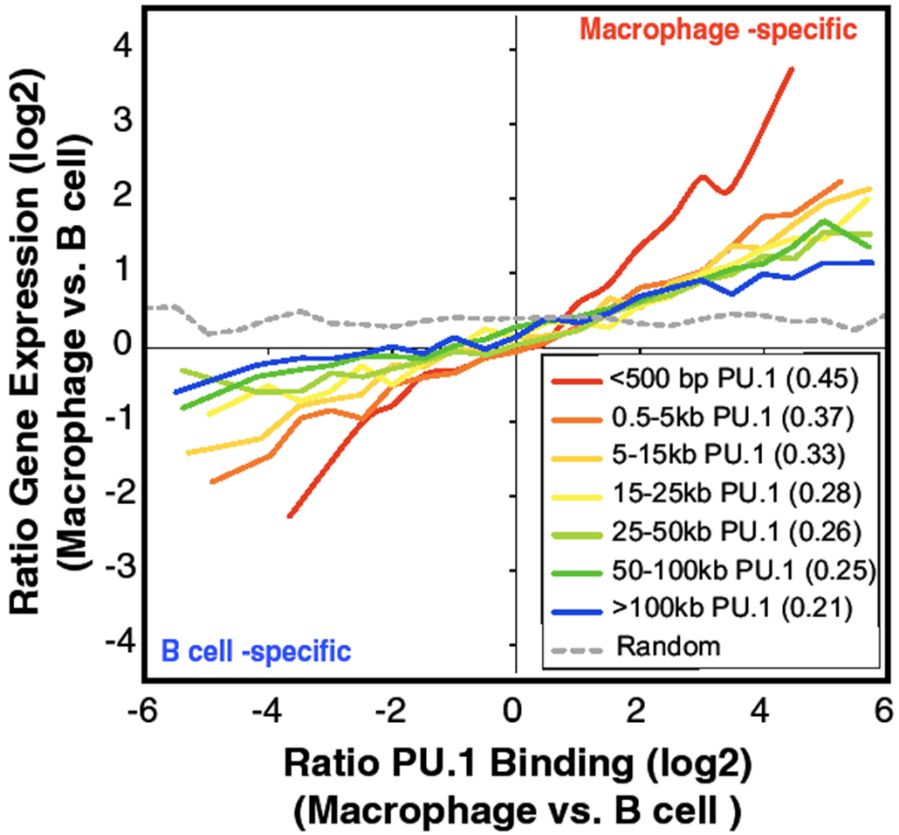
- Assign Peak to nearest TSS
- Assign all peaks within a certain distance to TSS
- Only assign conserved peaks/within same syntenic blocks
- Advanced association (additional information required)
  - eQTL – used genetics to help correlate gene activity with a given genomic region
  - 3D genomic interactions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



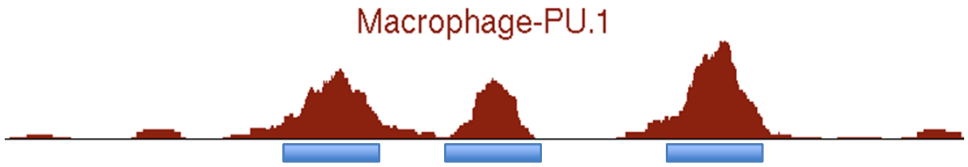
Correlating Gene Expression with ChIP-Seq

- Correlation is usually poor at best…



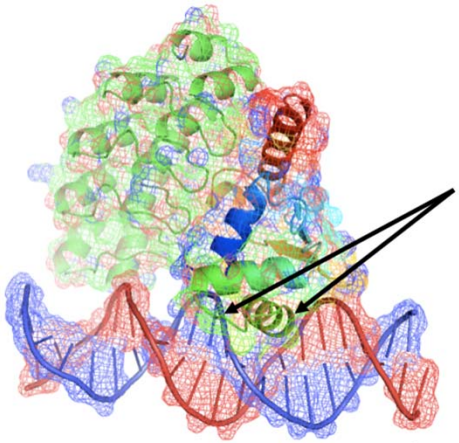


Motif Discovery



- 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)

Pre-processing Phase:

- Remove redundant sequences
- Normalize GC-content

Exhaustive Search Phase:

- Screen all possible oligos for enrichment

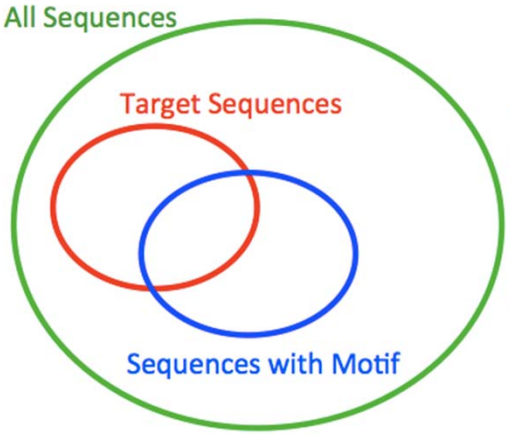
Local Optimization Phase:

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

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…

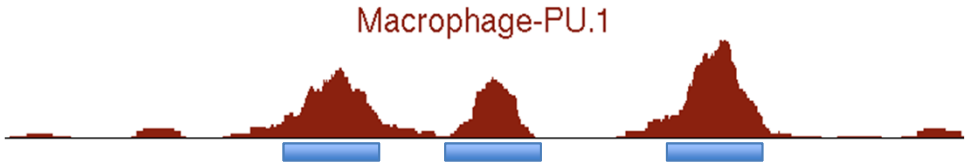
Motif significance calculated using zero or one occurrence per sequence (ZOOPS) coupled with cumulative hypergeometric distribution to calculate enrichment



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

(Maximize the overlap)

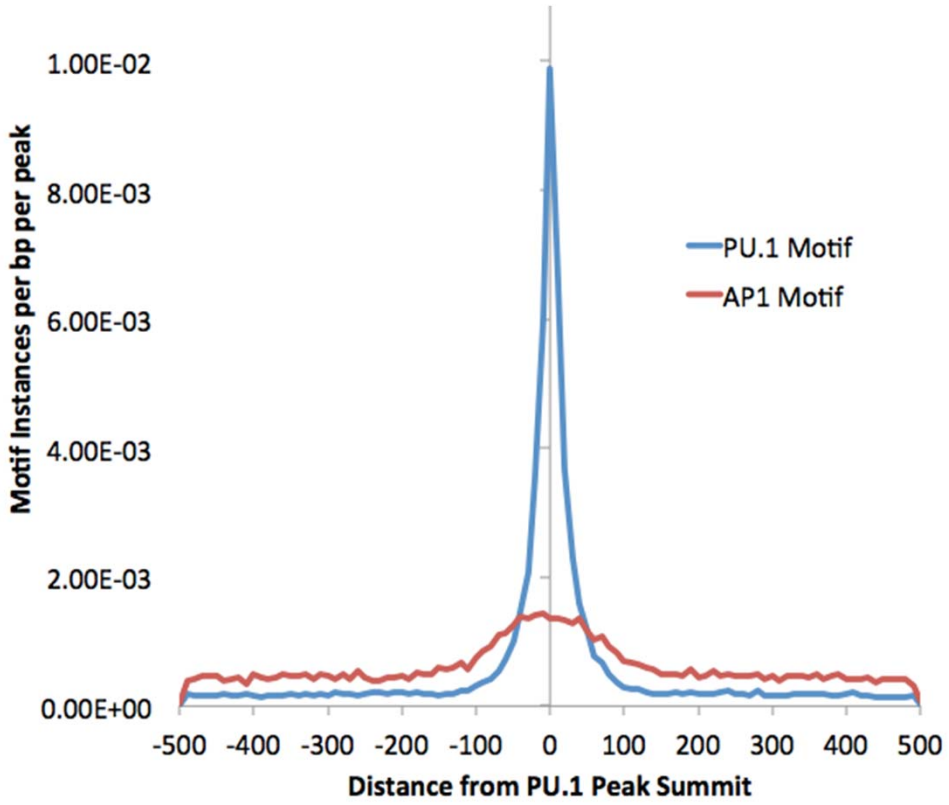
ChIP-Seq Motif Finding



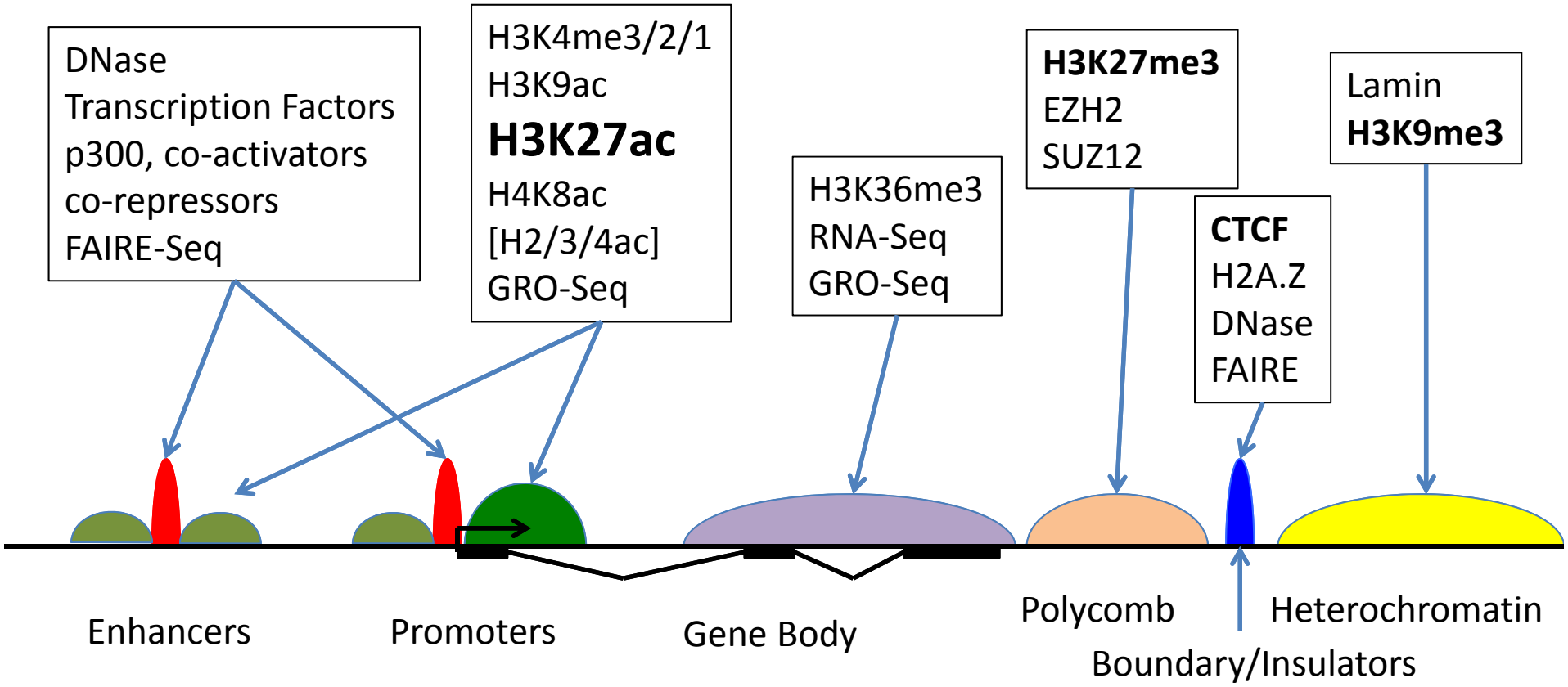
Extract Sequences at peaks, compare to random genomic sequence (matched for GC%)

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



Planning an experiment:
Type of Regions to Interrogate

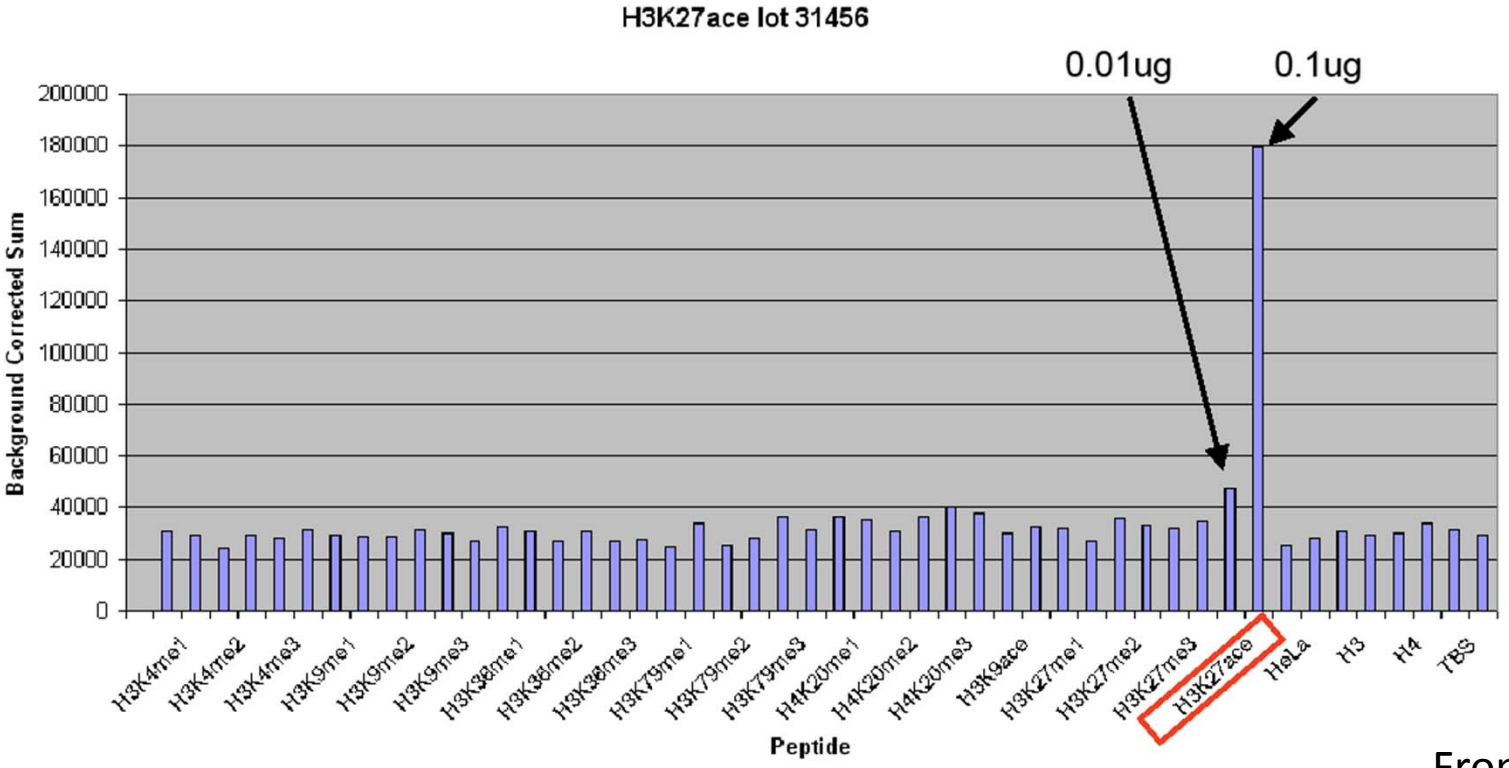


Planning ChIP-Seq Experiments

- Purpose of the Study
  - Study targets of a specific transcription factor
  - Identify enhancers (p300, H3K4me1/2, H3/H4ac, DNase)
  - Check PolII status in transcription Find regions of repressive chromatin
  - Study 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 /CTD-S2/S5 phosphorylation

Choosing the right antibody

- Needs to be ChIP Grade
  - Needs to IP crosslinked antigen (good test if no known target: compare ChIP vs. IP in Western blot)
- Good idea to validate specificity (better: vendor!)



From Encode

Crosslinking

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

- 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, http://www.ncbi.nlm.nih.gov/pubmed/16315372)

Protocol

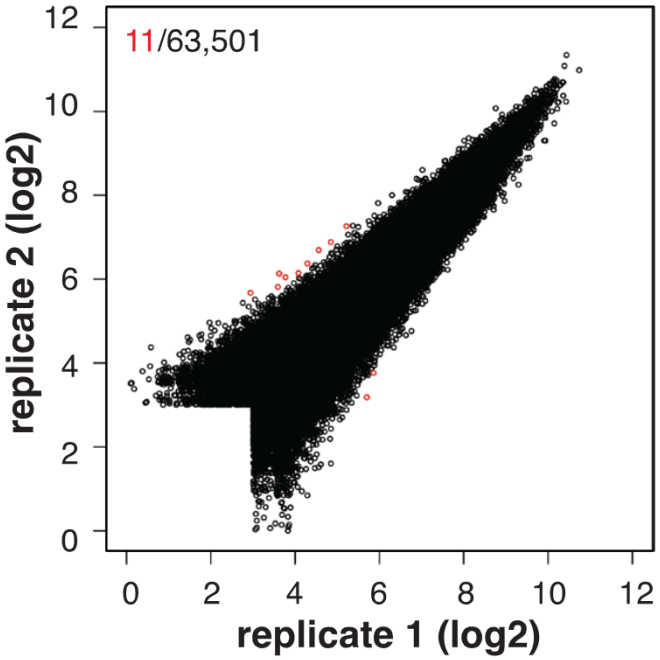
- Many protocols work, and often equally well (if the antibody is “good”)
- Biggest difference: detergent in the lysis buffer
  - Ionic detergents need to be diluted an/or sequestered by non-ionic detergents for the antibody to work. Non-ionic detergents (e.g. RIPA (1% Triton X-100, 0.1% SDS, 0.1% deoxycholate): don’t strip proteins off the chromatin, require more sonication
  - RIPA works better for “weak” antibodies, but less convenient and potentially higher background
- Magnetic beads have somewhat less background than agarose/sepharose, but also less ChIP product for library construction
- 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)
- Too little sonication (especially in RIPA) leaves little product in the size range (100-400 bp) required for sequencing
- Column cleanup has less open chromatin bias than phenol/chloroform (see FAIRE!)
- 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.



Epigenetic Modifications

- Lots of potential Histone modifications to analyze
  - H3K4me3, H4K20me1, H3K9me2, H4K12ac, H3K27me3, … (100+ more)
  - Most are highly correlated – You do not need to do too many histone modifications to have a successful study.
- mCpG ChIPs
- PolII as a surrogate for RNA-Seq/microarrays

Checklist ChIP-Seq

- Good antibody (ChIP-WB, ChIP-PCR)
- Input control

- Map
- QC/Visual Control (UCSC Genome Browser)
- Find Peaks
- Motif Finding
- GO Enrichment (GREAT)

- Archive FASTQ file
- Submit to GEO (Ab lot#, Protocol!)

Links

- UCSC Genome Browser

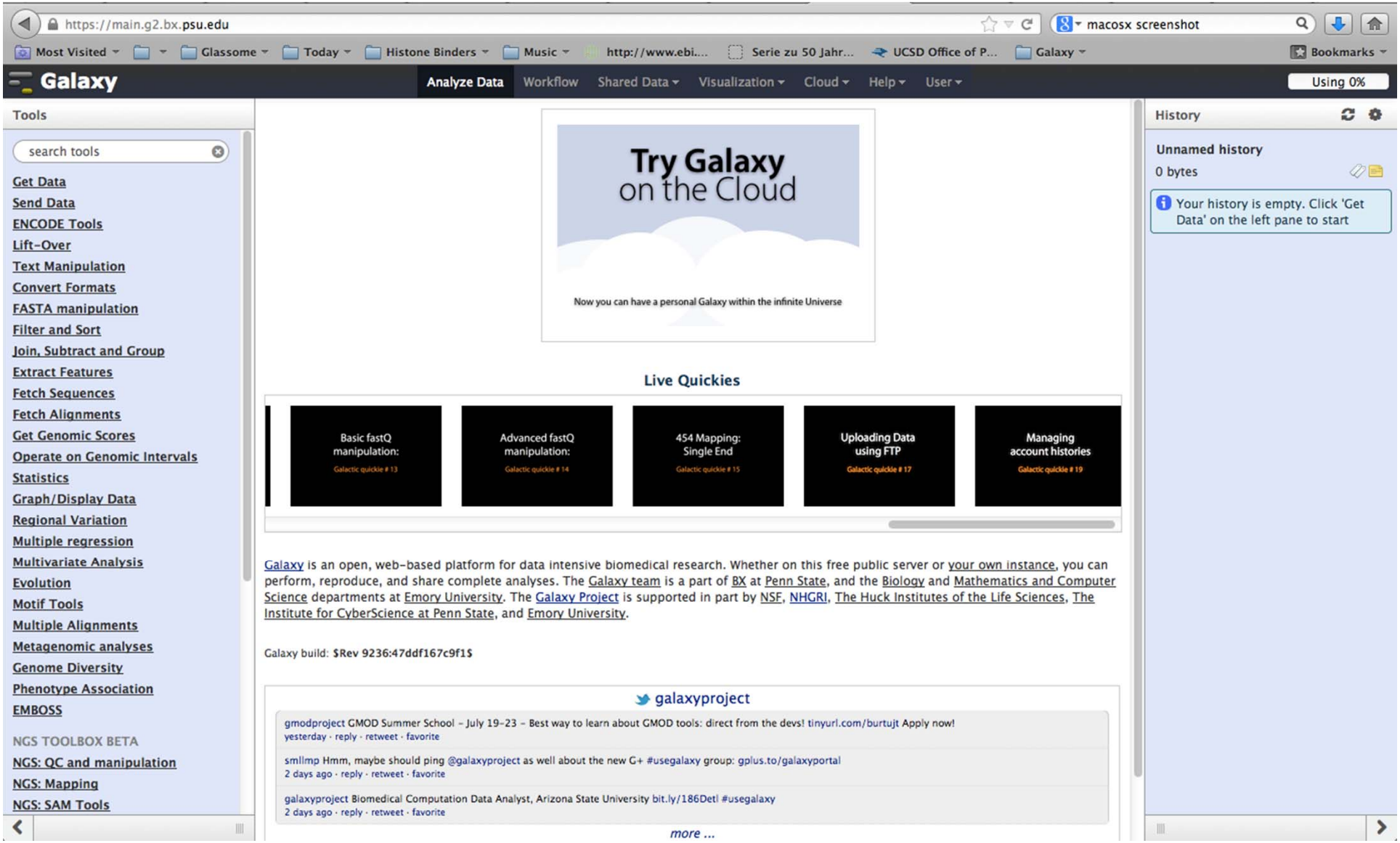
  http://genome.ucsc.edu
- Galaxy

  https://main.g2.bx.psu.edu/
- HOMER

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

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

Galaxy


https://main.g2.bx.psu.edu
macosx screenshot
Most Visited
Glassome
Today
Histone Binders
Music
http://www.ebi....
Serie zu 50 Jahr...
UCSD Office of P...
Galaxy
Bookmarks
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 0%
Tools
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
Genome Diversity
Phenotype Association
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Try Galaxy on the Cloud
Now you can have a personal Galaxy within the infinite Universe
Live Quickies
Basic fastQ manipulation:
Galactic quickie # 13
Advanced fastQ manipulation:
Galactic quickie # 14
454 Mapping: Single End
Galactic quickie # 15
Uploading Data using FTP
Galactic quickie # 17
Managing account histories
Galactic quickie # 19
Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.
Galaxy build: $Rev 9236:47ddf167c9f1$
galaxyproject
gmodproject GMOD Summer School - July 19-23 - Best way to learn about GMOD tools: direct from the devs! tinyurl.com/burtujt Apply now!
yesterday · reply · retweet · favorite
smllmp Hmm, maybe should ping @galaxyproject as well about the new G+ #usegalaxy group: gplus.to/galaxyportal
2 days ago · reply · retweet · favorite
galaxyproject Biomedical Computation Data Analyst, Arizona State University bit.ly/186DetI #usegalaxy
2 days ago · reply · retweet · favorite
more ...
History
Unnamed history
0 bytes
Your history is empty. Click 'Get Data' on the left pane to start

Galaxy: Penn State University