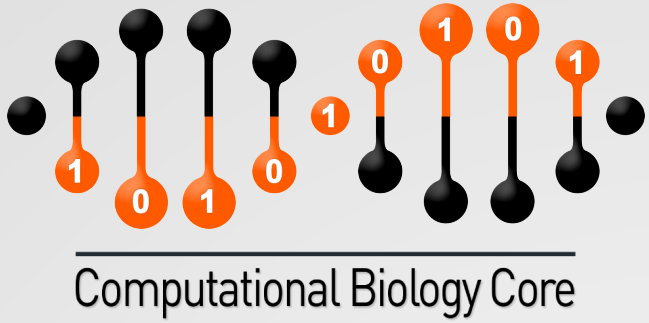
CBC Data Therapy

ChIP-Seq Discussion



UCONN
UNIVERSITY OF CONNECTICUT

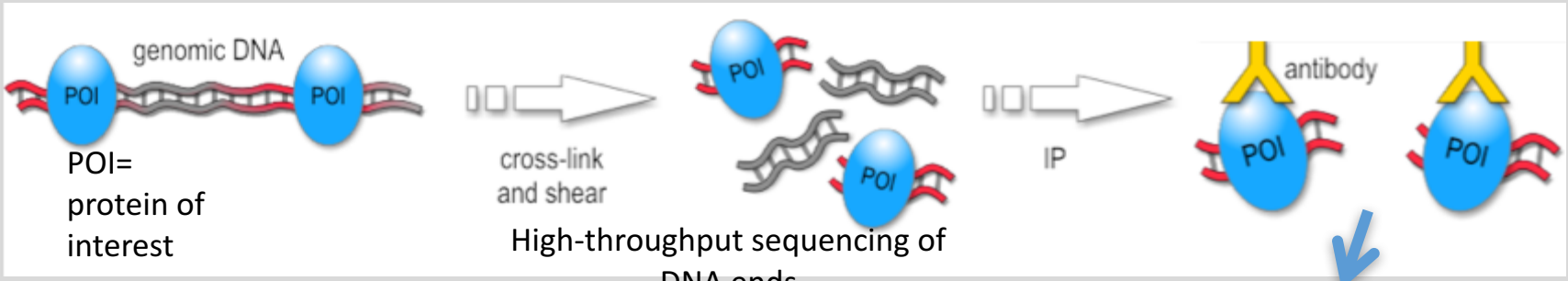
ChIP-seq big picture

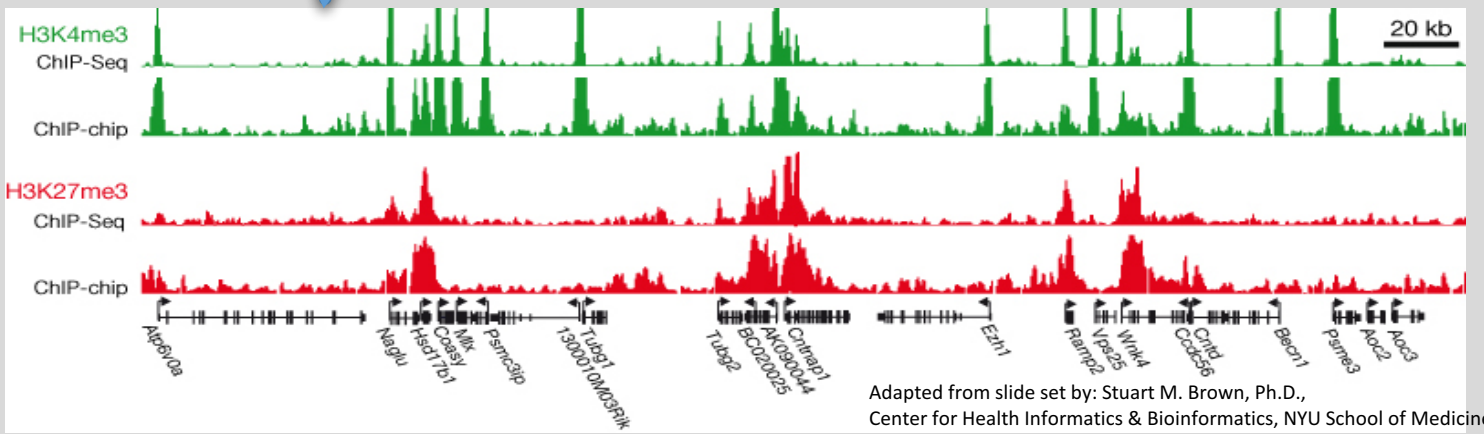
Combine “Next-Generation” sequencing with Chromatin Immunoprecipitation to identify genomewide chromatin binding sites.

Select (and identify) fragments of DNA that interact with specific proteins such as:

- Transcription factors
- Modified histones
- RNA Polymerase (survey actively transcribe portions of the genome)
- DNA polymerase (investigate DNA replication)
- DNA repair enzymes
- Or fragments of DNA that are modified: e.g. CpG methylation

ChIP-seq





Adapted from slide set by: Stuart M. Brown, Ph.D.,
Center for Health Informatics & Bioinformatics, NYU School of Medicine

How many reads do I need?

(a) Number of binding sites

(b) Genome size

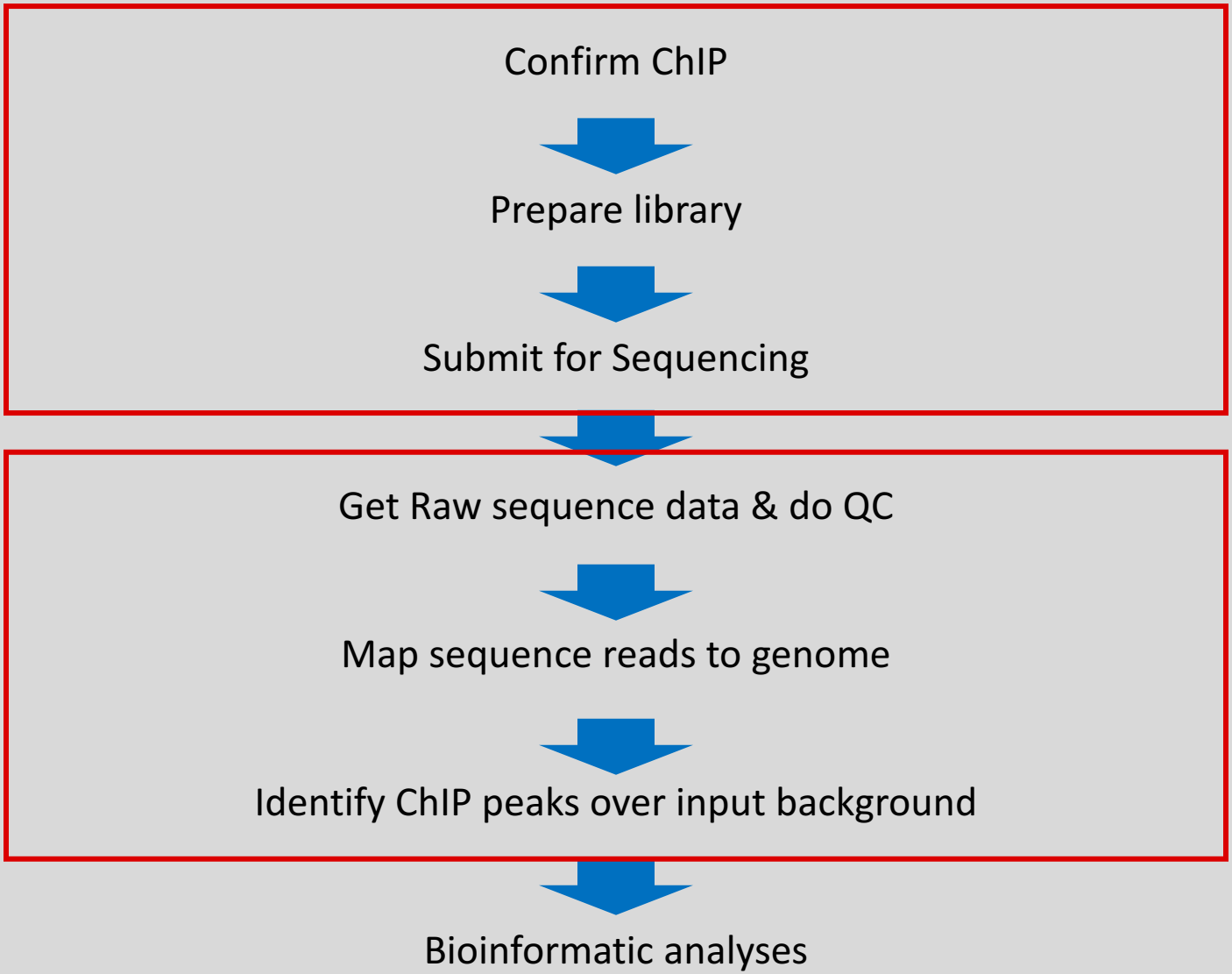
(c) Specificity of antibody/method

Minimum for ChIP-seq of a transcription factor with < ~30,000 binding sites in a mammalian genome:

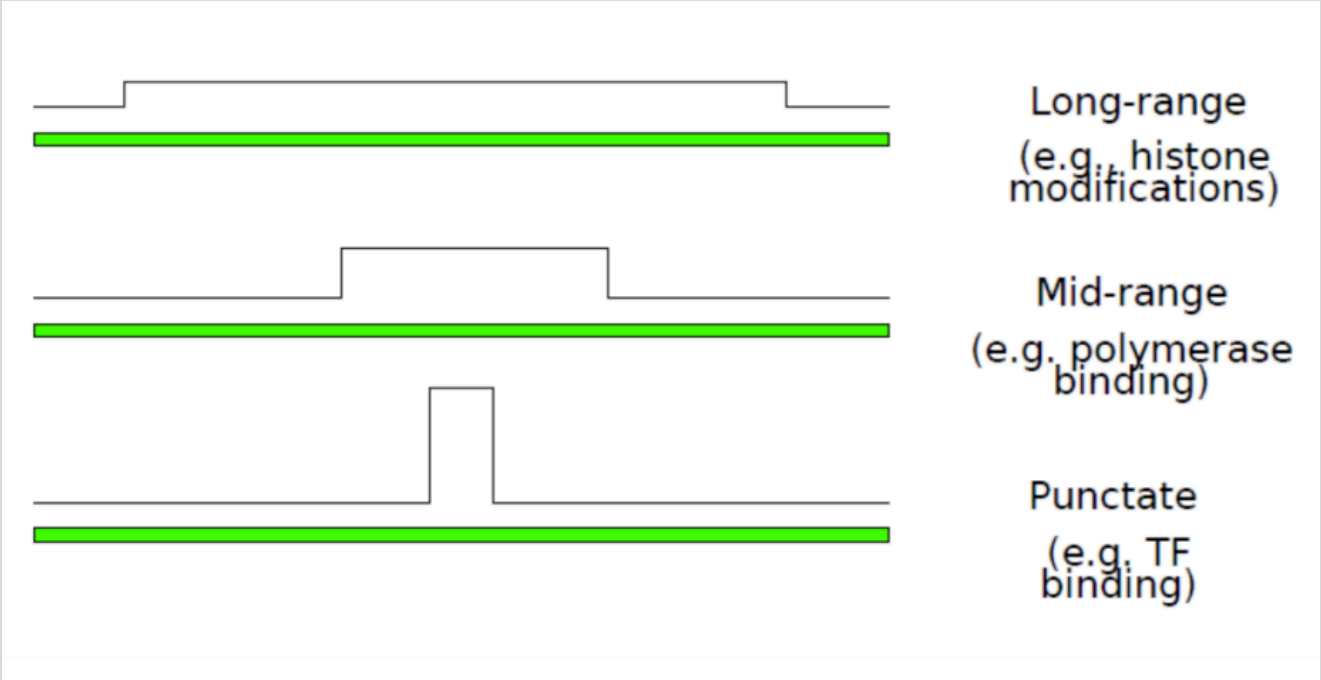
- **3 replicates per condition**
- **20+ million reads per sample** (>10X coverage of binding genome size, proportionately less for smaller genomes & fewer binding peaks)
- **Paired end shorter reads almost always good**

For some applications need many more reads (e.g. mapping nucleosome positions need >400 M). Make your best estimate. If you have too few you can re-sequence the same samples or add additional samples. Reads from all runs can be pooled in the end.

ChIP-seq Workflow



Types of events of interest



Peak calling algorithms/Softwares

1.The most popular peak caller by Tao Liu: MACS2. Now --broad flag supports broad peaks calling as well.
2.TF ChIP-seq peak calling using the Irreproducibility Discovery Rate (IDR) framework and many Software Tools Used to Create the ENCODE Resource
3.SICER for broad histone modification ChIP-seq
4.HOMER can also used to call Transcription factor ChIP-seq peaks and histone modification ChIP-seq peaks.
5.MUSIC
6.permseq R package for mapping protein-DNA interactions in highly repetitive regions of the genomes with prior-enhanced read mapping. Paper on PLos Comp.
7.Ritornello: High fidelity control-free chip-seq peak calling. No input is required!
8.Tumor samples are heterogeneous containing different cell types. MixChIP: a probabilistic method for cell type specific protein-DNA binding analysis
9.Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains tool
10.BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets
11.epic: diffuse domain ChIP-Seq caller based on SICER. It is a re-writen of SICER for faster processing using more CPUs. (Will try it for broad peak for sure).
12.Cistrome: The best place for wet lab scientist to check the binding sites. Developed by Shierly Liu lab in Harvard.
13.Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-Seq peak callers tool in github

MACS2 : Examine the output

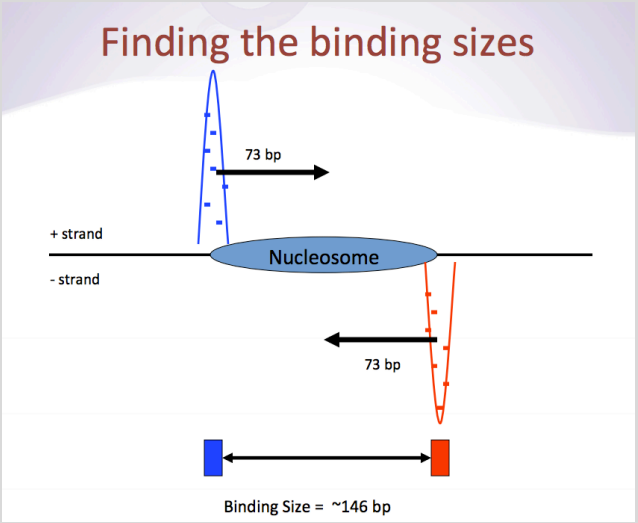
Start with your .macsinfo bsub -oo file.

vi LiE_ERaIPvINPUT_chr19.macsinfo

Use the arrow keys to go to the top, where you'll see all of the parameters you put in to run MACs. After some runtime info (including possible warnings, that you can ignore if there are not millions of them), you'll see:

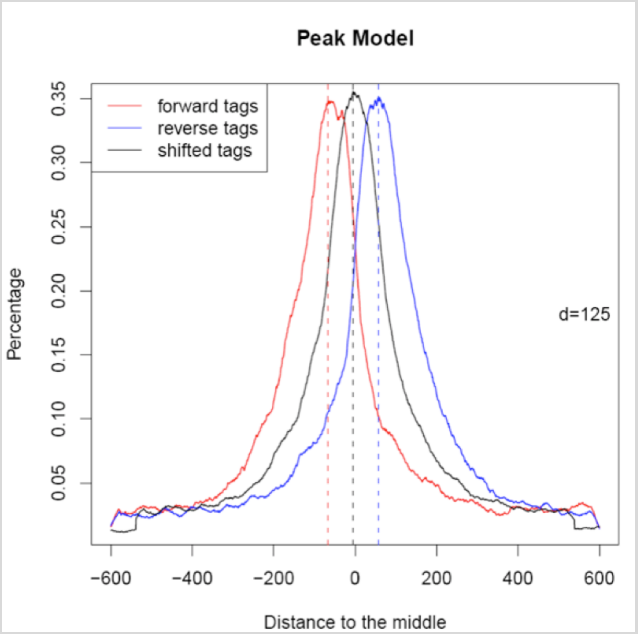
INFO @ Sun, 10 Feb 2013 21:27:51: #1 total tags in treatment: 370513
INFO @ Sun, 10 Feb 2013 21:27:51: #1 user defined the maximum tags...
INFO @ Sun, 10 Feb 2013 21:27:51: #1 filter out redundant tags at the same location and the same strand by allowing at most 1 tag(s)
INFO @ Sun, 10 Feb 2013 21:27:51: #1 tags after filtering in treatment: 275955 (High is is good)
INFO @ Sun, 10 Feb 2013 21:27:51: #1 Redundant rate of treatment: 0.26 (low is good)

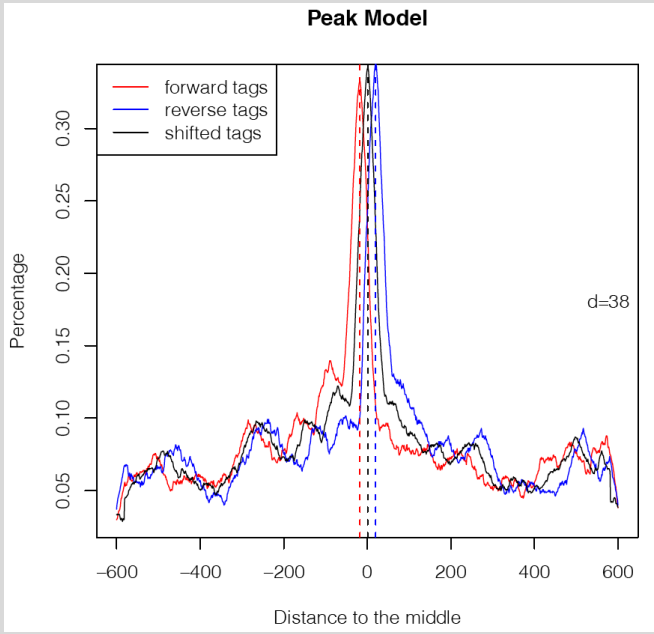
This is useful information. It tells you how many different reads you had (out of all of the reads which mapped to only one place in the mouse genome- from Bowtie). You want this number to be high and the "redundant rate" to be low.



Be on the lookout for MACS building a model from short-separation noise peaks (that may arise from sonication sensitive breakpoints or other things unrelated to your protein binding).

To avoid this, you can decrease the maximum "mfold" so that these strong irrelevant peaks are ignored when the model is built.





Peaks & negative peaks

Keep scrolling down your .macsinfo file until you find…

…

INFO @ Sun, 10 Feb 2013 21:36:47: #3 **Finally, 364 peaks are called!**

INFO @ Sun, 10 Feb 2013 21:36:47: #3 find negative peaks by swapping treat and control

INFO @ Sun, 10 Feb 2013 21:36:52: #3 **Finally, 36 peaks are called!**

INFO @ Sun, 10 Feb 2013 21:36:52: #4 Write output…

This is the pay-off, where MACS identifies your ER alpha peak locations!

364 peaks on chromosome 19 (which is ~1/50th of the genome) suggests ~20,000 peaks for the whole genome, which is not bad!

Equally critical, MACS now swaps treat & control (pretending your INPUT data is your IP & your ChIP data is your input) and looks again for peaks.

The number of "negative" peaks found in this way should be far less than the positive peaks, and the 10:1 ratio here is fine.

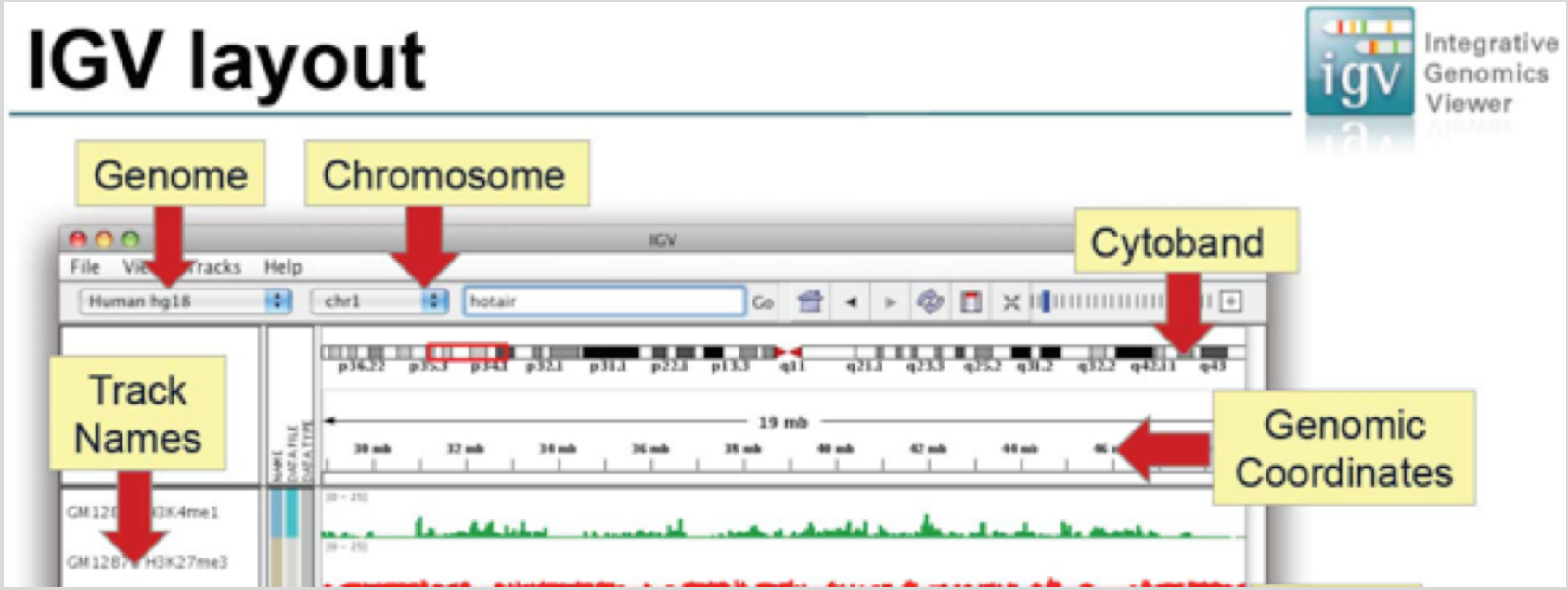
######

Peaks/Negative Peaks ratio is poor or too few peaks are detected:

- Adjust model settings to see if you can improve both. Otherwise, you may have to conclude that 1) your library was no good or 2) the factor just doesn't bind to many places in the genome.

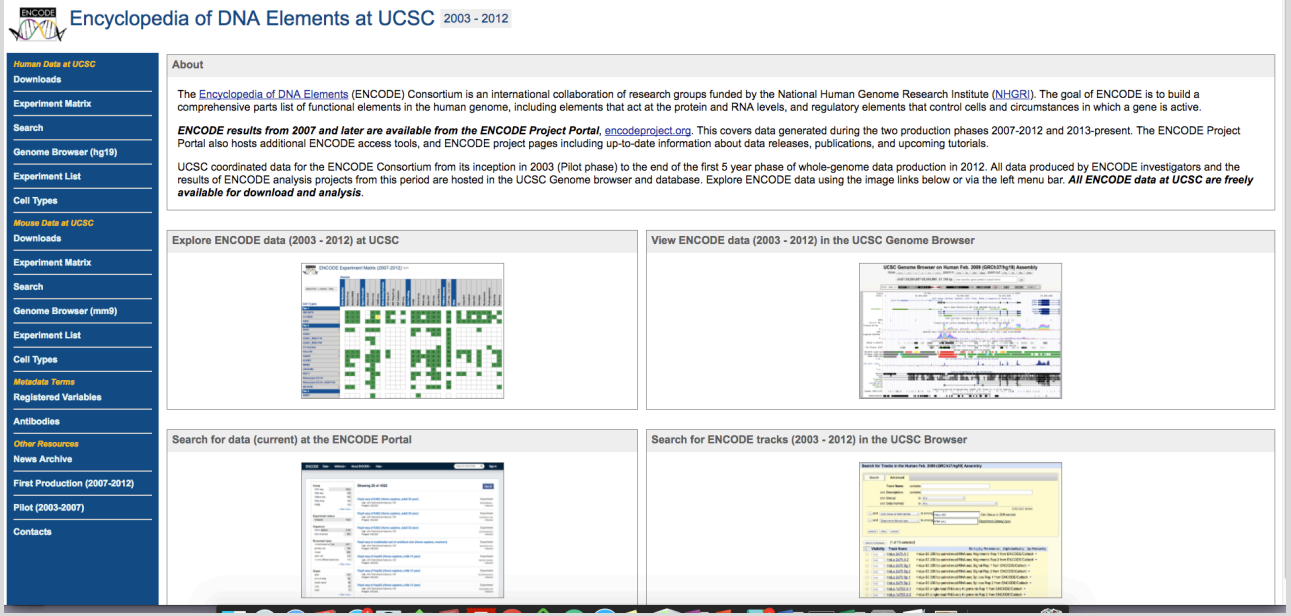
MACS2 output : .bdg files

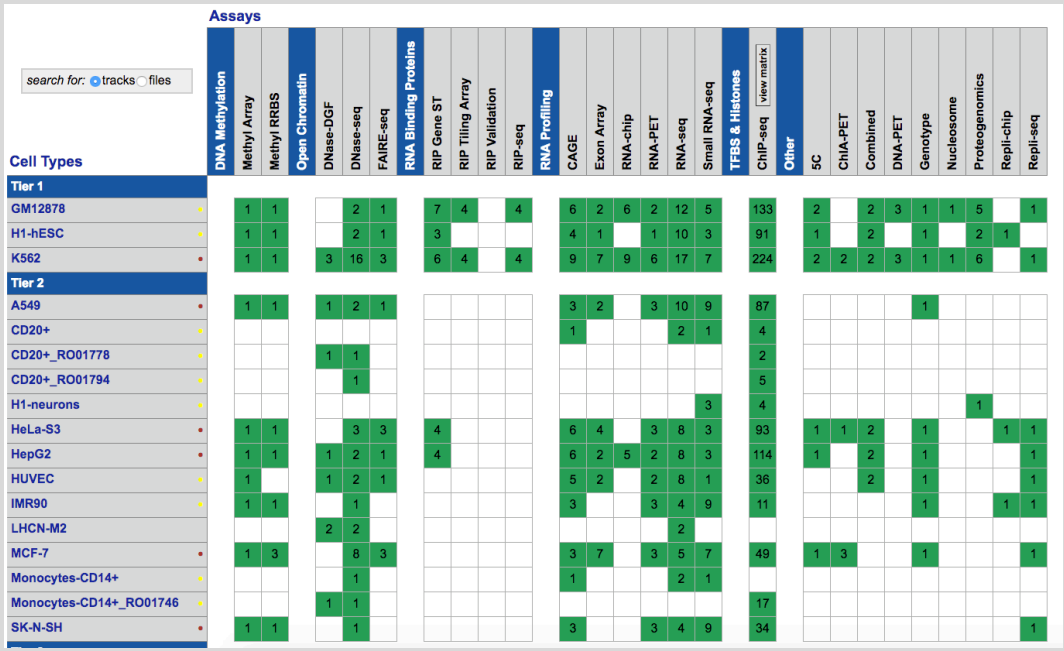
Broad IGV, an alternative to UCSC browser



http://www.broadinstitute.org/igv/
You will need to register, but they don't send you spam.

ENCODE


Encyclopedia of DNA Elements at UCSC 2003 - 2012
Human Data at UCSC
Downloads
Experiment Matrix
Search
Genome Browser (hg19)
Experiment List
Cell Types
Mouse Data at UCSC
Downloads
Experiment Matrix
Search
Genome Browser (mm9)
Experiment List
Cell Types
Metadata Terms
Registered Variables
Antibodies
Other Resources
News Archive
First Production (2007-2012)
Pilot (2003-2007)
Contacts
About
The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.
ENCODE results from 2007 and later are available from the ENCODE Project Portal, encodeproject.org. This covers data generated during the two production phases 2007-2012 and 2013-present. The ENCODE Project Portal also hosts additional ENCODE access tools, and ENCODE project pages including up-to-date information about data releases, publications, and upcoming tutorials.
UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome browser and database. Explore ENCODE data using the image links below or via the left menu bar. All ENCODE data at UCSC are freely available for download and analysis.
Explore ENCODE data (2003 - 2012) at UCSC
View ENCODE data (2003 - 2012) in the UCSC Genome Browser
Search for data (current) at the ENCODE Portal
Search for ENCODE tracks (2003 - 2012) in the UCSC Browser


search for: tracks files
Assays
DNA Methylation
Methyl Array
Methyl RRBS
Open Chromatin
DNase-DGF
DNase-seq
FAIRE-seq
RNA Binding Proteins
RIP Gene ST
RIP Tiling Array
RIP Validation
RIP-seq
RNA Profiling
CAGE
Exon Array
RNA-chip
RNA-PET
RNA-seq
Small RNA-seq
TFBS & Histones
ChIP-seq
view matrix
Other
5C
ChIA-PET
Combined
DNA-PET
Genotype
Nucleosome
Proteogenomics
Repli-chip
Repli-seq
Cell Types
Tier 1
GM12878
H1-hESC
K562
Tier 2
A549
CD20+
CD20+_RO01778
CD20+_RO01794
H1-neurons
HeLa-S3
HepG2
HUVEC
IMR90
LHCN-M2
MCF-7
Monocytes-CD14+
Monocytes-CD14+_RO01746
SK-N-SH

GTRD: Gene Transcription Regulation Database

http://gtrd.biouml.org/

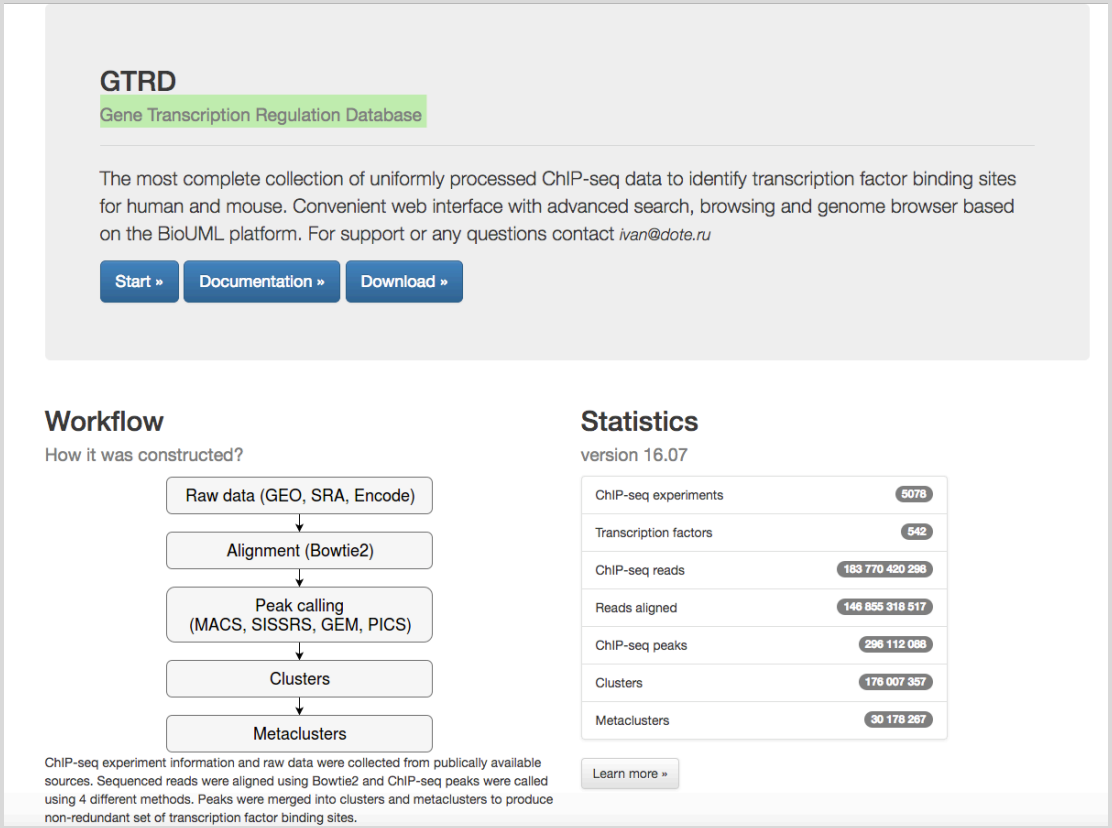


Table 1. Comparison of databases that are based on ChIP-seq data

Database, URL	Source of human and mouse data	Number of samples (TF-related)*	Number of TFs	Number of ChIP-seq peak callers used	Metaclus-ter approach	Uniform data processing	Genome browser
ChIPBase (http://rna.sysu.edu.cn/chipbase)	GEO, ENCODE	total 3549 human 2498 mouse 1036 rat 15	252 TFs and non-TFs for 10 species	>10 in total, but no uniform pipeline, each ChIP-seq is processed by different peak caller	No	No	Self-developed: deepView genomeView
Cistrome DB (http://dc2.cistrome.org/#/)	GEO, SRA, ENA, ENCODE	total 10 276 (TF+non-TF) human 5774 mouse 4502 rat 0	260 TFs and non-TFs	1 (MACS2)	No	Yes	UCSC genome browser
ENCODE (https://www.encodeproject.org)	ENCODE	total 1448 human 1254 mouse 194 rat 0	295 TFs and non-TFs for human, 52 TFs and non-TFs for mouse	5 (SPP, GEM, PeakSeq, MACS, Hotspot/Hotspot2)	No	Yes	Self-developed: UCSC genome browser and WashU epigenome browser
Factorbook (http://www.factorbook.org)	ENCODE	total 1007 human 837 mouse 170 rat 0	167 TFs, co-factors and chromatin remodeling factors for human, 51—for mouse	None	No	No	No
GTRD (http://gtrd.biouml.org)	GEO, SRA, ENCODE	total 5078 human 2955 mouse 2107 rat 16	476 human and 257 mouse sequence specific TFs, corresponding to 542 TFClass classes.	4 (MACS, SISSRs, GEM, PICS)	Yes	Yes	Self-developed
ChIP-Atlas (http://chip-atlas.org)	SRA	total 10 774 human 5914 mouse 4860 rat 0	699 human and 502 mouse TFs and others.	1(MACS2)	No	Yes	IGV
GeneProf (http://www.geneprof.org)	SRA, ENCODE, literature	total 1692 human 693 mouse 999 rat 0	133 human and 131 mouse TFs	1(MACS)	No	Yes	Self-developed: based on GenomeGraphs
NGS-QC (http://www.ngs-qc.org)	GEO	total 6672 human 4234 mouse 2438 rat 0	unknown	None	No	Yes	No

GALAXY/CISTROME

http://cistrome.dfci.harvard.edu/ap/root

Galaxy tools specially designed for ChIP-seq analysis.

Cistrome allows easy access to many analyses that give you some quick insights into your data.

CISTROME TOOLBOX: ASSOCIATION STUDY

GCA: Gene centered annotation **Find the nearest interval in the given intervals set for every annotated coding gene** (e.g. where's the nearest ER binding site for each gene in the genome)

peak2gene: Peak Center Annotation **Input a peak file, and It will search each peak on UCSC GeneTable to get the refGenes near the peak center** (e.g. where's the nearest gene to each ERa binding site in the genome).