ChIP-seq Analysis

BaRC Hot Topics - August 27th 2018

Bioinformatics and Research Computing

Whitehead Institute

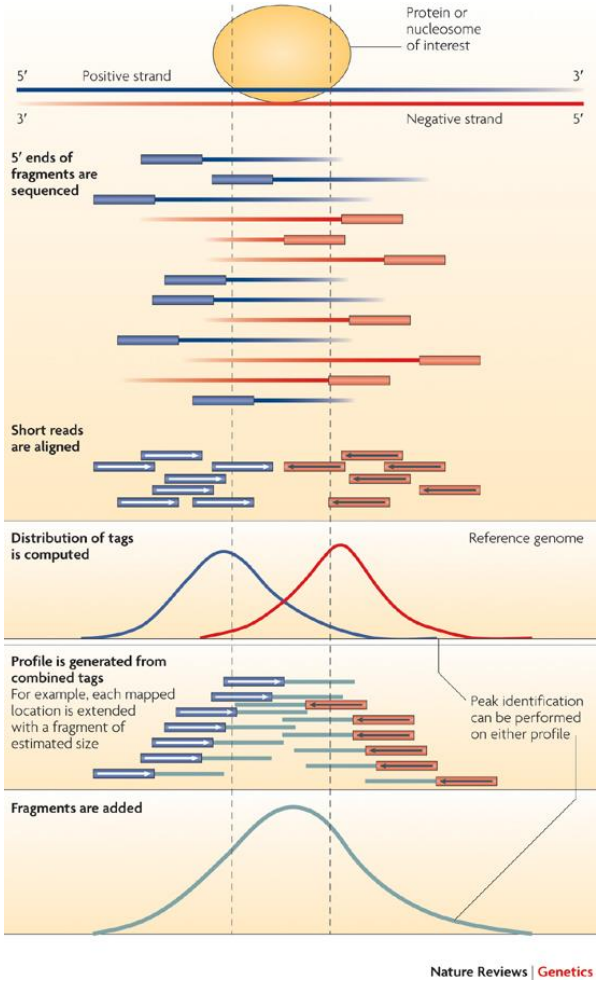
http://barc.wi.mit.edu/hot_topics/

WHITEHEAD INSTITUTE

Outline

- ChIP-seq overview
- Quality control/preprocessing of the reads
- Mapping
  - Map reads
  - Check the profile of the mapped reads (strand cross-correlation analysis)
- Peak calling
- Check ChIP-seq quality
- Looking for potential TFs based on public ChIP-seq depositories

ChIP-Seq overview



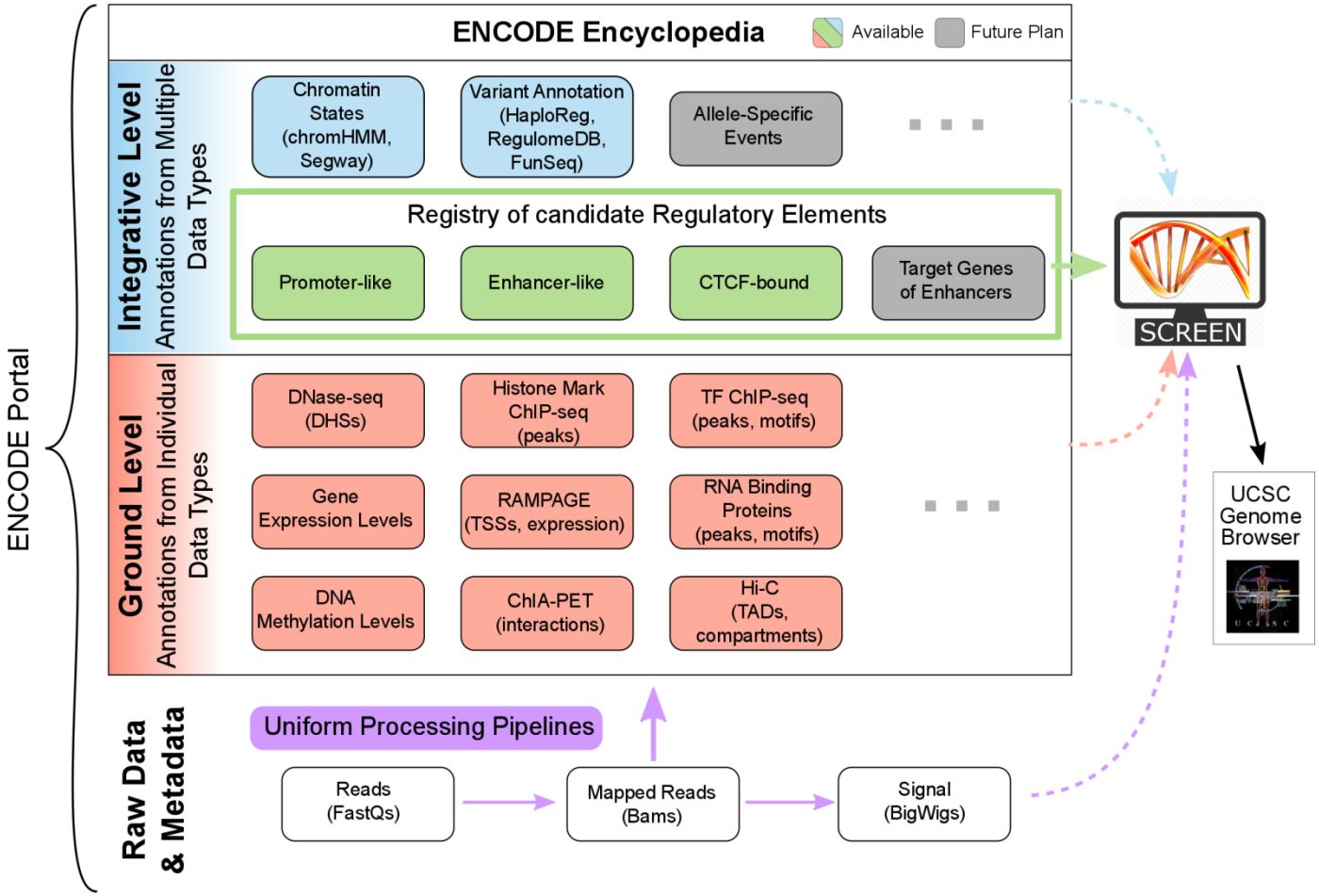
Steps in typical data analysis

1. Quality control

2. Mapping
   Treat IP and control the same way (preprocessing and mapping)
3. Peak calling
   i) Read extension and signal profile generation
   ii) Peak assignment
4. Peak analysis and interpretation
   i) Find genes next to peaks
   ii) Infer possible biological consequences of the binding

Park, P. J., *ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet. Oct;10(10):669-80 (2009)*

ENCODE


ENCODE Portal
ENCODE Encyclopedia
Available
Future Plan
Integrative Level
Annotations from Multiple Data Types
Chromatin States (chromHMM, Segway)
Variant Annotation (HaploReg, RegulomeDB, FunSeq)
Allele-Specific Events
Registry of candidate Regulatory Elements
Promoter-like
Enhancer-like
CTCF-bound
Target Genes of Enhancers
Ground Level
Annotations from Individual Data Types
DNase-seq (DHSs)
Histone Mark ChIP-seq (peaks)
TF ChIP-seq (peaks, motifs)
Gene Expression Levels
RAMPAGE (TSSs, expression)
RNA Binding Proteins (peaks, motifs)
DNA Methylation Levels
ChIA-PET (interactions)
Hi-C (TADs, compartments)
Raw Data & Metadata
Uniform Processing Pipelines
Reads (FastQs)
Mapped Reads (Bams)
Signal (BigWigs)
SCREEN
UCSC Genome Browser

Current ENCODE Guideline

- Transcription Factors:
  - Each replicate should have >=20mil fragments
- Histone:
  - Narrow peak: >= 20mil
  - Broad peak: >= 45 mil

Broad Marks	H3F3A	H3K27me3	H3K36me3	H3K4me1	H3K79me2	H3K79me3	H3K9me1	H3K9me2	H4K20me1
Narrow Marks	H2AFZ	H3ac	H3K27ac	H3K4me2	H3K4me3	H3K9ac			
Exceptions	H3K9me3								

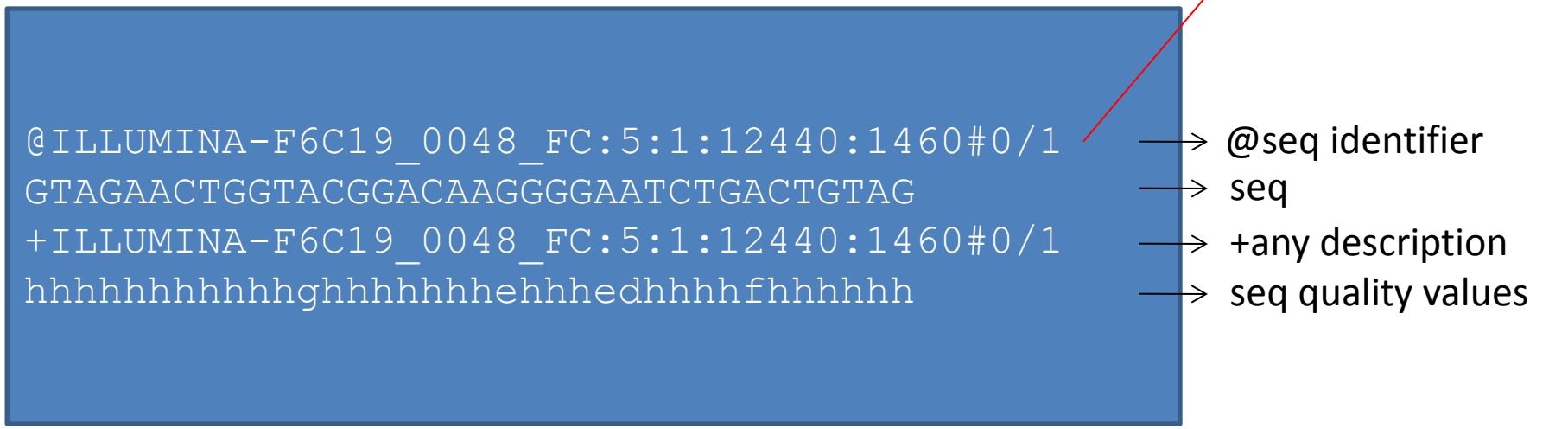
Design

- Read length:
  - Short reads is enough : ~35bp
  - Repeat regions:
    - benefits with pair-end reads, longer reads
- Control samples (Input, IgG):
  - Under the same conditions as the test sample
  - At least the same depth as test sample
- Replicates:
  - Technical replicates: not necessary
  - Biological replicates: at least 2

Illumina data format

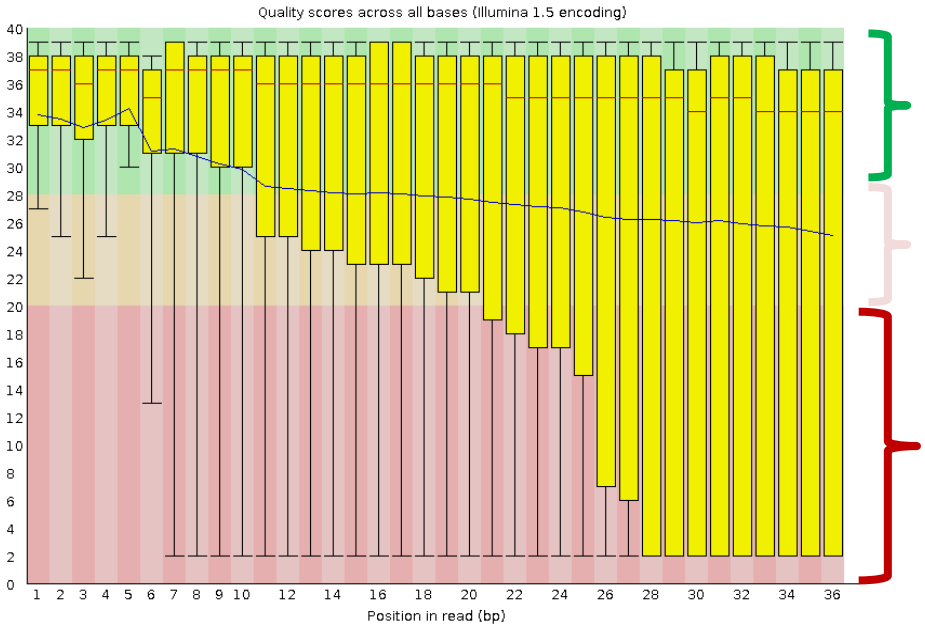
- Fastq format:

  http://en.wikipedia.org/wiki/FASTQ_format



Input qualities	Illumina versions
--solexa-quals	<= 1.2
--phred64	1.3-1.7
--phred33	>= 1.8

FastQC: per base sequence quality



very good quality calls

reasonable quality

poor quality

Red: median
Blue: mean
Yellow: 25%, 75%
Whiskers: 10%, 90%

Quality value = −10 * log10 (error probability)
Quality = 10 => error rate = 10% => base call has 90% confidence
Quality = 20 => error rate = 1% => base call has 99% confidence
Quality = 30 => error rate = 0.1% => base call has 99.9% confidence

Preprocessing tools

- **Fastx Toolkit** (http://hannonlab.cshl.edu/fastx_toolkit/)
  - FASTQ/A Trimmer: Shortening reads (removing barcodes or noise).
  - FASTQ Quality Trimmer: Trims (cuts) sequences based on quality
  - FASTQ Masker: Masks nucleotides with 'N' (or other character) based on quality
- **cutadapt** to remove adapters

  (https://code.google.com/p/cutadapt/)

Recommendation for preprocessing

- Treat IP and control samples the same way during preprocessing and mapping.
- If you have paired-end reads, make sure you still have both reads of the pair after the processing is done.
- Run fastqc on the processed samples to see if the problem has been removed.

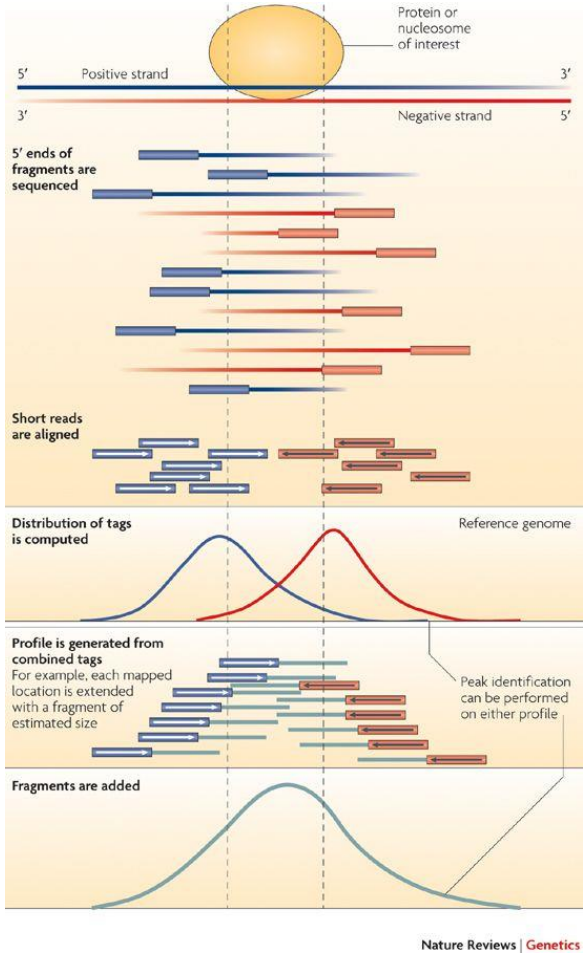
Mapping
Non-spliced alignment software

- Bowtie2:
  - Supports gapped alignment.
  - Use “end-to-end“ (default) alignment mode
  - Previous Hot Topics on how to run the programs
- BWA:
  - refer to the BaRC SOP for detailed information

    http://barcwiki.wi.mit.edu/wiki/SOPs
- Filter reads mapped by quality mapping score
  - 10% mapping position is wrong

  samtools view -bq 10 file.bam > filtered.bam

Strand-specific profile at enriched site



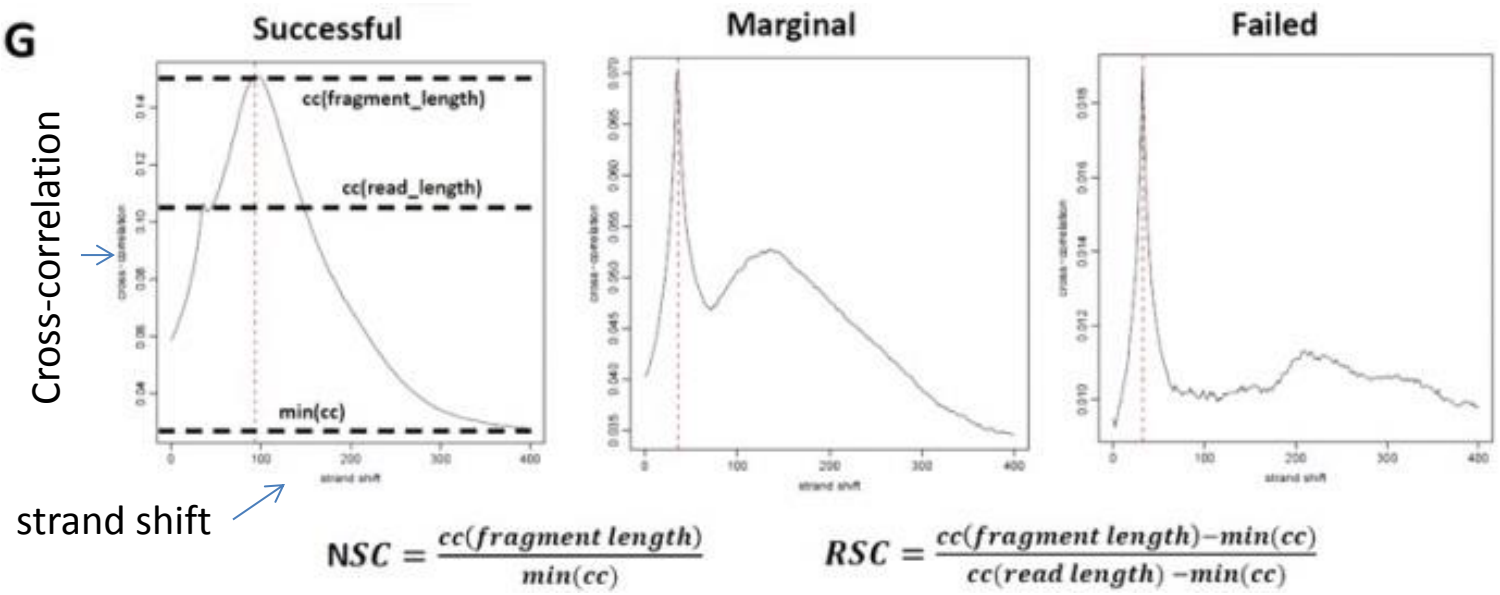
Nature Reviews Genetics 2009

- Fragment size needs to be estimated
- Profile generated by extending each mapped position toward center with estimate fragment size, and adding fragments.

Estimation of the fragment length: Strand cross-correlation analysis

Example command:

run_spp.R -c=H3k4me3_chr1.bam -savp -out=H3k4me3_chr1.run_spp.out



Genome Res. 2012. 22:1813

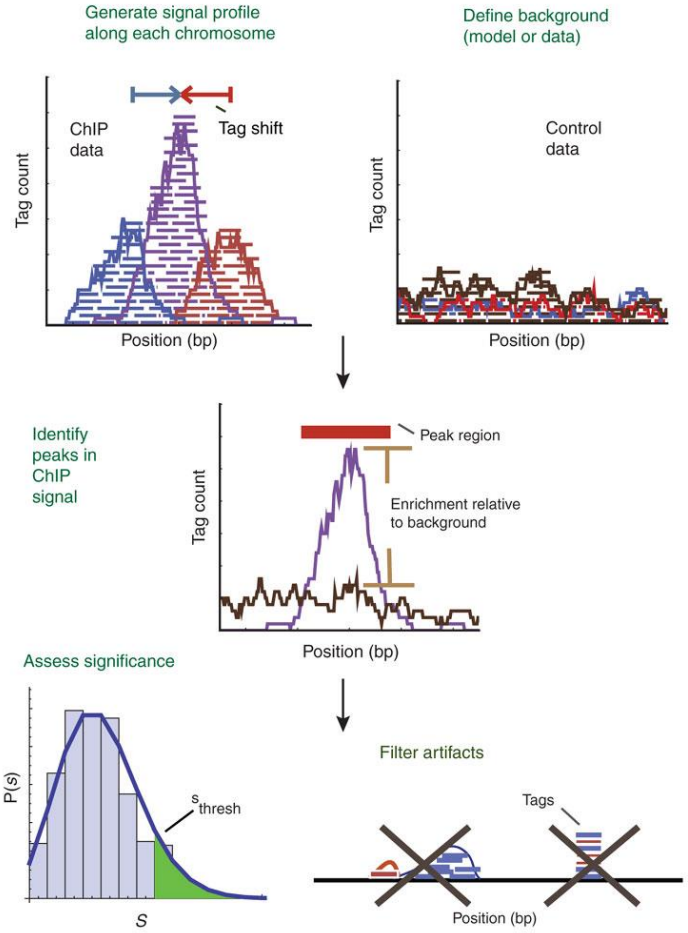
Peak calling

i) Read extension and signal profile generation
- strand cross-correlation can be used to calculate fragment length

ii) Peak evaluation
- Look for fold enrichment of the sample over input or expected background
- Estimate the significance of the fold enrichment using:
  - Poisson distribution
  - background distribution from input DNA
  - model background data to adjust for local variation (MACS)

iii) Broad peaks:

two-level peak calling (MACS):
- embed stronger/narrower calls in weaker/broader calls



Pepke, S. *et al*. Computation for ChIP-seq and RNA-seq studies, *Nat Methods. Nov. 2009*

Peak calling: MACS

- MACS can calculate the fragment length but we will use a different program and give MACS the fragment length as an input parameter.
- It uses a Poisson distribution to assign p-values to peaks. Use local lambda, to capture the influence of local fluctuation and biases.
- MACS default is to filter out redundant tags at the same location and with the same strand by allowing at most 1 tag. This works well for most cases, but not for identifying signals in repeat regions.
- -g: Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts: 'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for *C. elegans* (9e7) and 'dm' for fruit fly (1.2e8), Default:hs
- For broad peaks like some histone modifications it is recommended to use `--nomodel` and if there is no input sample to use `--nolambda`.

Example MACS command

MACS command

bsub macs2 callpeak -t H3k4me3_chr1.bam -c Control_chr1.bam --name H3k4me3_chr1 -f BAM -g hs --nomodel -B --extsize 175

PARAMETERS

- -t TFILE Treatment file
- -c CFILE Control file
- --name NAME Experiment name, which will be used to generate output file names. DEFAULT: "NA"
- -f FORMAT Format of tag file, "BED" or "SAM" or "BAM" or "BOWTIE". DEFAULT: "AUTO"
- --nomodel skips the step of calculating the fragment size.
- -B create a bedgraph
- --extsize EXTSIZE The arbitrary extension size in bp. When nomodel is true, MACS will use this value as fragment size to extend each read towards 3' end, then pile them up. **You can use the value from the strand cross-correlation analysis**

MACS output

Output files:

1. Excel peaks file ("_peaks.xls") contains the following columns

   ```
   Chr, start, end, length, abs_summit, pileup,
   -LOG10(pvalue), fold_enrichment, -
     LOG10(qvalue), name
   ```

2. "_summits.bed": contains the peak summits locations for every peaks. The 5th column in this file is -log10qvalue
3. "_peaks.narrowPeak" is BED6+4 format file. Contains the peak locations together with peak summit, fold-change, pvalue and qvalue.

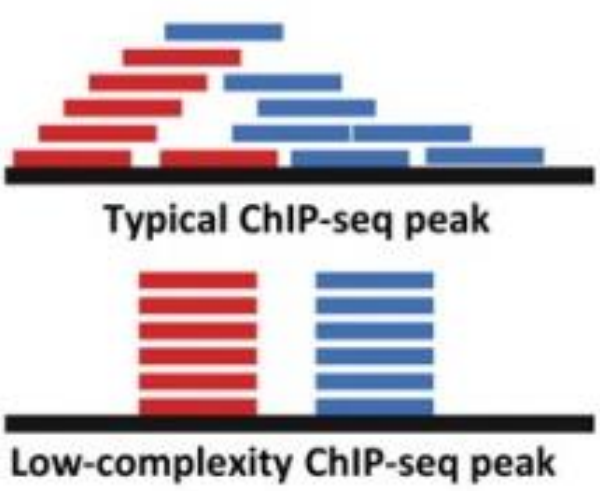
To look at the peaks on a genome browser you can upload one of the output bed files or you can also make a bedgraph file with columns (step 3 of hands on):

```
chr, start, end, fold_enrichment
```

How good is your ChIP-seq?

- Based on mapped reads and peaks:
  - Library Complexity, low duplication level
  - FRiP: Fraction of Reads In Peaks
  - Cross-correlation analysis
  - RiBL: Reads Intercept with Black List
  - Enrichment of reads in genomic features
  - Visualize peaks in genome browser

Library Complexity



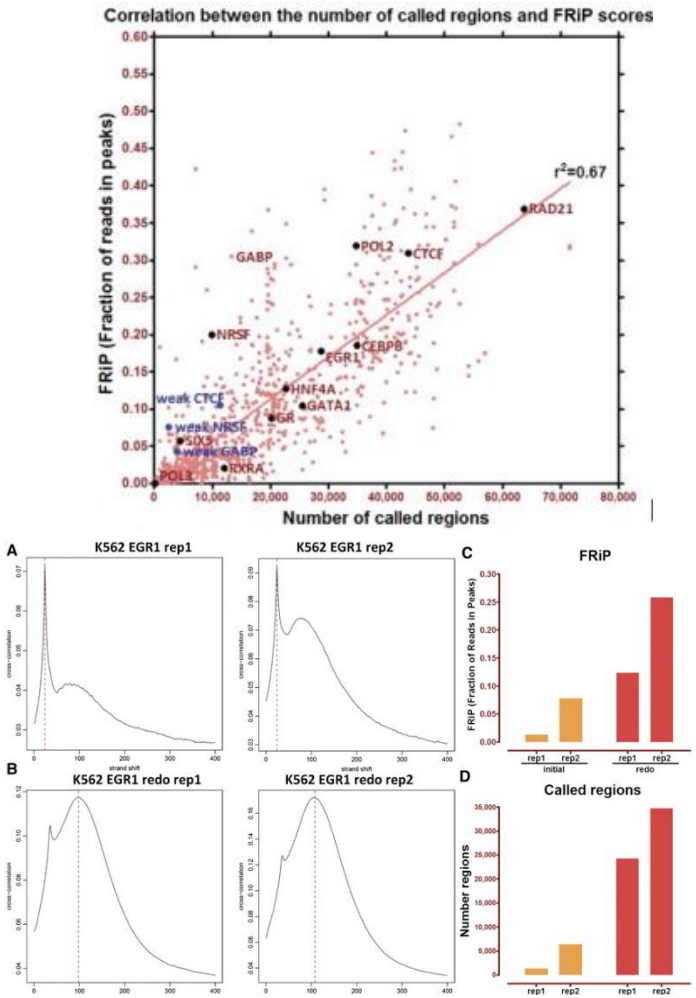
Genome Res. 2012. 22:1813

NRF: nonredundant fraction

- number of positions in the genome with uniquely mappable reads / total number of uniquely mappable reads
- Low NRF:
  - Little DNA, problem with library construction
  - Over PCR saturation:
  - profiling a protein that binds to a small subset of the genome
- ENCODE recommends NRF > 0.9

Can be calculated with QCmetrics from ChIPQC

FRiP: Fraction of Reads In Peaks



Genome Res. 2012. 22:1813

- To check success of the immunoprecipitation
- Often correlates with the number of called regions.
- Different proteins have different values
- Uses:
  - To compare results obtained with the same antibody across cell lines.
  - should be derived from peaks uniformly called by a single algorithm and parameter set.

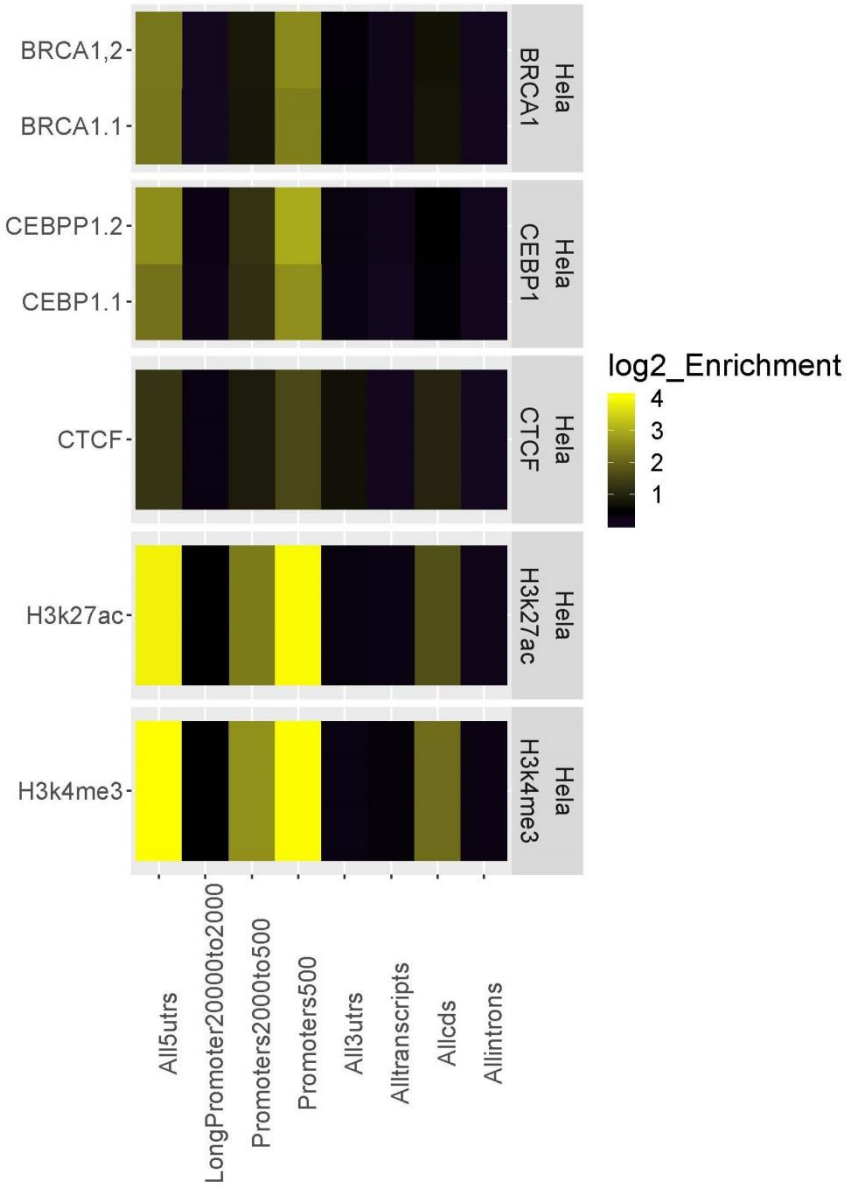
Can be calculated with QCmetrics from ChIPQC

Reads Intercept with Black List (RiBL)

- Artificially high signals :
  - genomic regions with anomalous, unstructured, high signal/read counts in next gen sequencing experiments independent of cell line and type of experiment
- Uniquely mapped, often in repeats (centromeres, telomeres, satellite repeats)
  - 0.5% of genome, but could accounts for >10% total signals
- Influence peak caller, fragment length estimation
- Solution: masking regions prior to analysis
- BaRC_datasets -> ENCODE_blacklists
  - Human, mouse, fly, *C. elegans*

Can be calculated with QCmetrics from ChIPQC

Enrichment of reads in genomic features



plotRegi function from ChIPQC

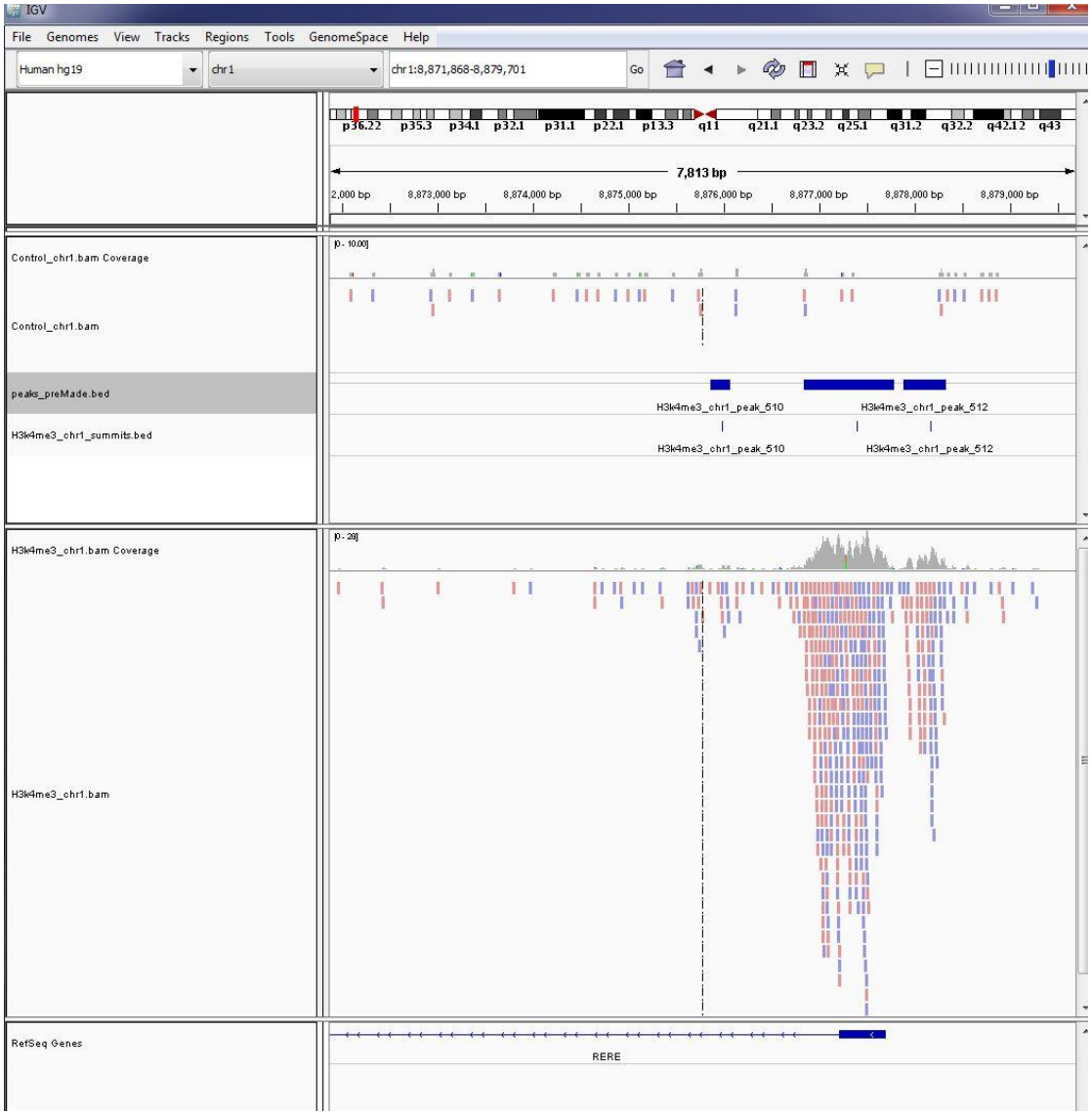
Visualize peaks in IGV

control

peaks

summit

Hek4me3



Comparing ChIP-seq across samples

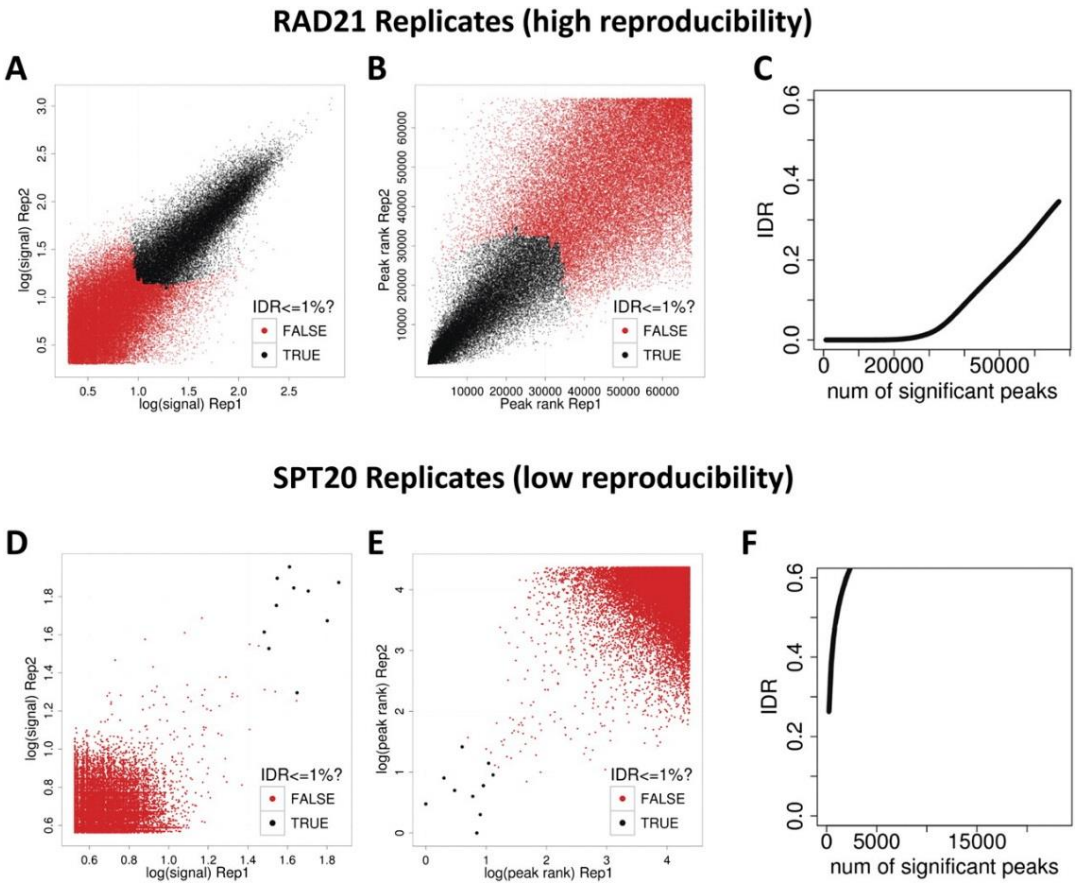
- Bedtools:

  /nfs/BaRC_Public/BaRC_code/Perl/compare_bed_overlaps

- macs2 bdgdiff

- IDR

The Irreproducible Discovery Rate (IDR)

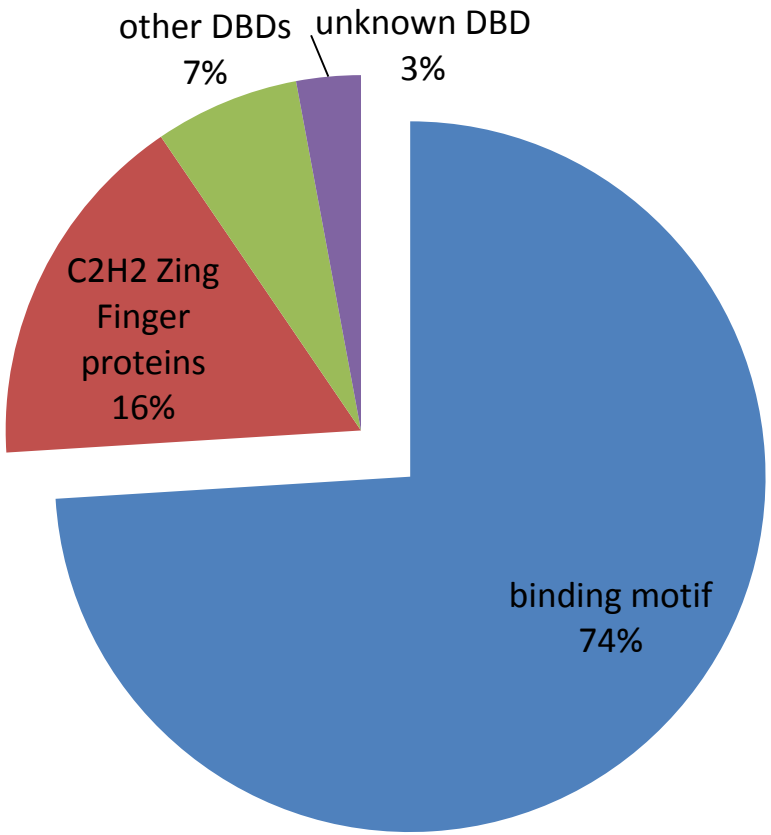


Stephen G. Landt et al. Genome Res. 2012;22:1813-1831

IDR

- No artificial cutoff is needed
- Based on ranking order, can be used to compare peaks in replicates called by different peak callers
- Peak calling should not be very stringent. Noise component should be presented in the data
- Broad peaks: no IDR

Most TFs have a binding motif



Lambert SA, et.al The Human Transcription Factors. Cell 172(4):650 2018

MEME-ChIP and TomTom from MEME Suite are good tools for identifying motifs from sequences in enriched peaks.

Mining public ChIP-seq datasets

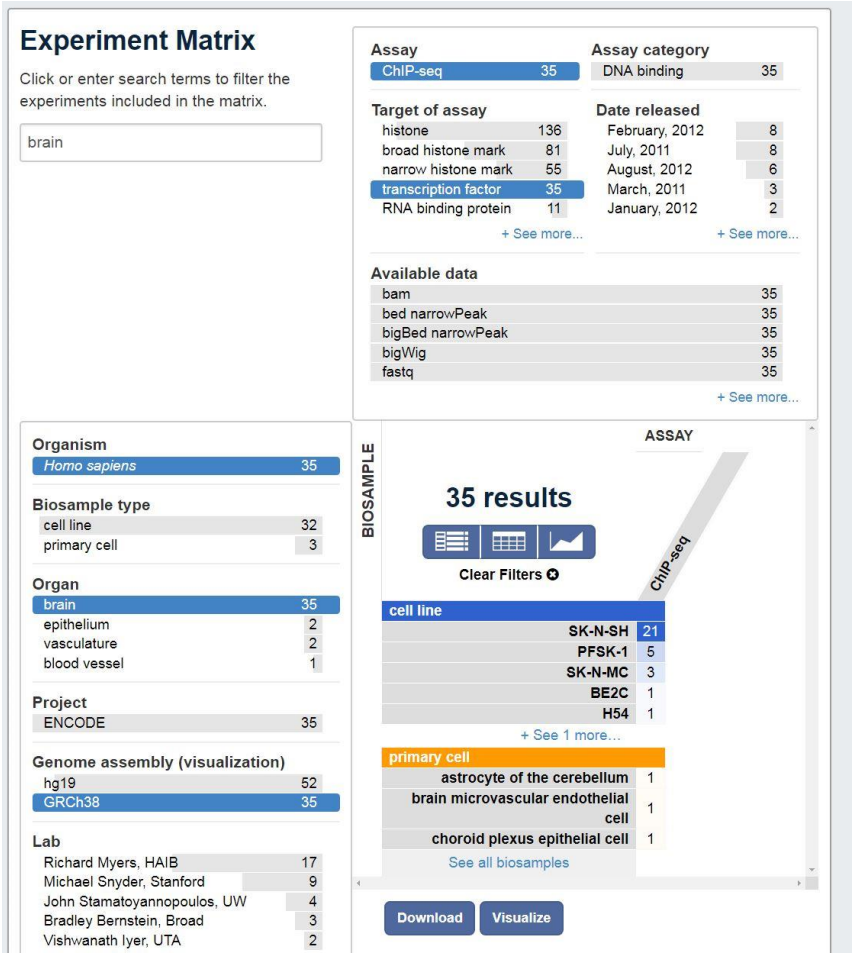
- ENCODE: https://www.encodeproject.org/
  - Factorbook: http://www.factorbook.org/human/chipseq/tf/
    - Motifs of 167 TFs
  - SCREEN: search cis-regulatory elements:
    - http://screen.encodeproject.org/
  - ChIP-Seq Significance Tool: http://encodeqt.simple-encode.org/
    - Identify Enriched TFs from a List of Genes

- ReMap: http://tagc.univ-mrs.fr/remap/
  - human 485 TFs

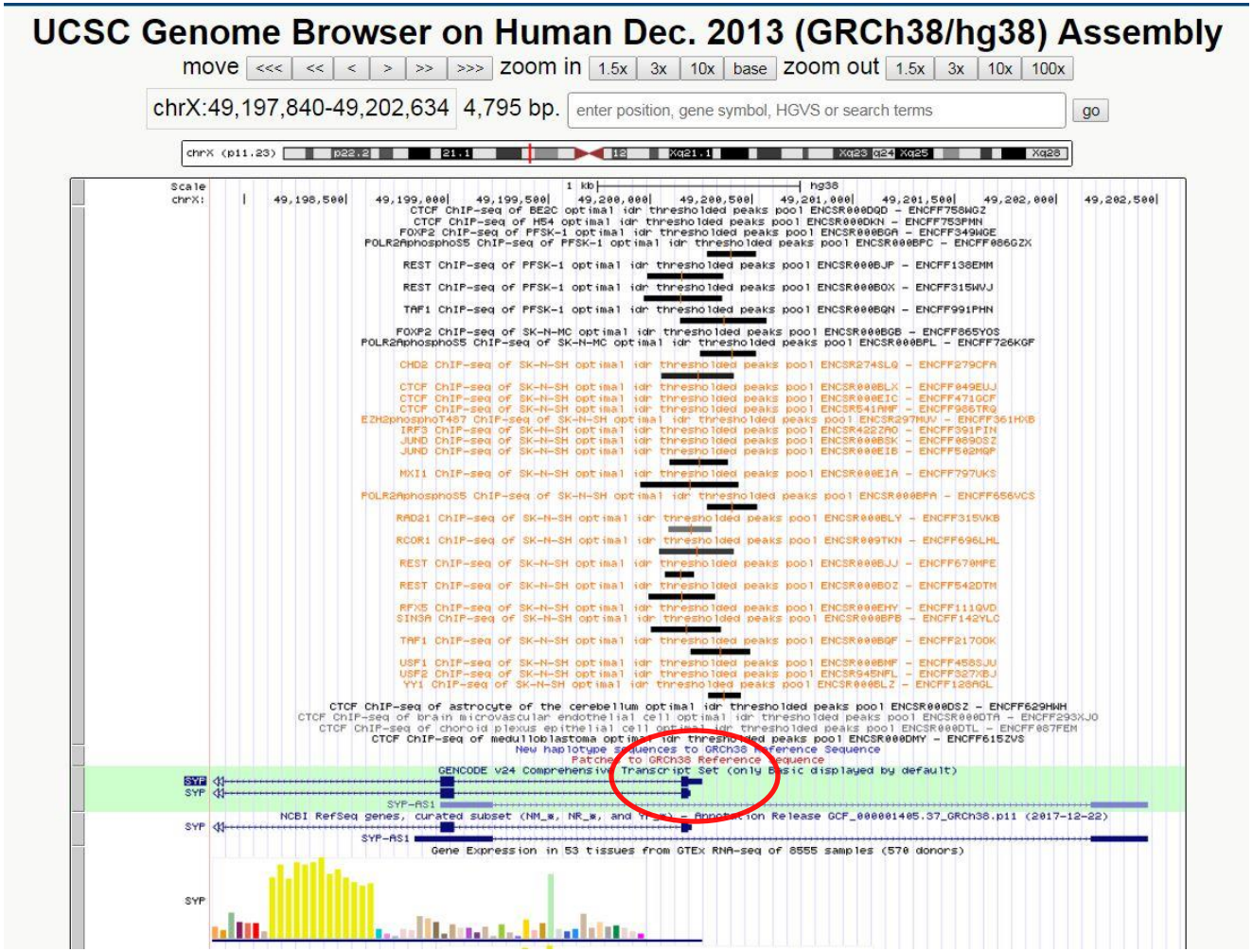
- The Human Transcription Factors (from Lambert et al.)
  - Collection of human 1639 TFs
  - Manually reviewed
  - http://humantfs.ccbr.utoronto.ca/

- Cistrome Project: http://www.cistrome.org/
  - Cistrome Data Browser: collection of ChIP-seq, DNAse-seq datasets
  - Cistrome Cancer: TFs targets and enhancer profiles in cancer

Mining ENCODE data
https://www.encodeproject.org/



Visualize TFs binding sites around TSS



References

- Reviews and benchmark papers:

  ChIP-seq: advantages and challenges of a maturing technology (Oct 09)
  (http://www.nature.com/nrg/journal/v10/n10/full/nrg2641.html)

  Computation for ChIP-seq and RNA-seq studies (Nov 09)
  (http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1371.html)

  Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Comput. Biol*. 2013

  A computational pipeline for comparative ChIP-seq analyses. *Nat. Protoc*. 2011

  ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res. 2012.*

  Identifying and mitigating bias in next-generation sequencing methods for chromatin biology

  Nature Reviews Genetics 15, 709–721 (2014) Meyer and Liu.

- Quality control and strand cross-correlation:

  http://code.google.com/p/phantompeakqualtools/

- MACS:

  Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 2008
  http://liulab.dfci.harvard.edu/MACS/index.html

  Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*. 2011

  http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0214s34/pdf

- Bedtools:

  https://code.google.com/p/bedtools/
  http://bioinformatics.oxfordjournals.org/content/26/6/841.abstract

- ngsplot:

  https://code.google.com/p/ngsplot/ Shen, L.*, Shao, N., Liu, X. and Nestler, E. (2014) BMC Genomics, 15, 284.

Other resources

- Previous Hot Topics
  - Quality Control:

    Jan 2017: : Quality Control
  - Link peaks to genes:

    March 2017 :: Analysis of ChIP-seq data
  - Visualizing ChIP-seq reads with ngsplot :

    May 2014 :: Visualization of Next-Gen Sequencing (NGS) Data: ngsplot

- SOPs

  http://barcwiki.wi.mit.edu/wiki/SOPs/chip_seq_peaks

File location

Unix (tak / Cluster)	Windows	Mac
/nfs/BaRC_training	\\wi-files1\BaRC_training	smb://wi-files1/BaRC_training

UNIX:
tak4.wi.mit.edu
cd /nfs/BaRC_training

Windows:
Start Menu ->
Map Network Drive ->
\\wi-files1\BaRC_training

non-CCDOM computer ->
Connect using different credentials ->
ccdom\username

MAC:
GO ->
Connect to Server ->
smb://wi-files1/BaRC_datasets