Daria Merkurjev

Chromatin Immunoprecipitation DNA Sequencing (ChIP-seq)

Workshop logistics

Disparate backgrounds

- Some will find this class *easy*
- Others will find it *hard*
- Hints to more advanced topics:
  I can stay longer after class for those interested

Questions

- Of interest to everybody: *anytime*
- Personal curiosity OR requiring long discussion:
  before or after class

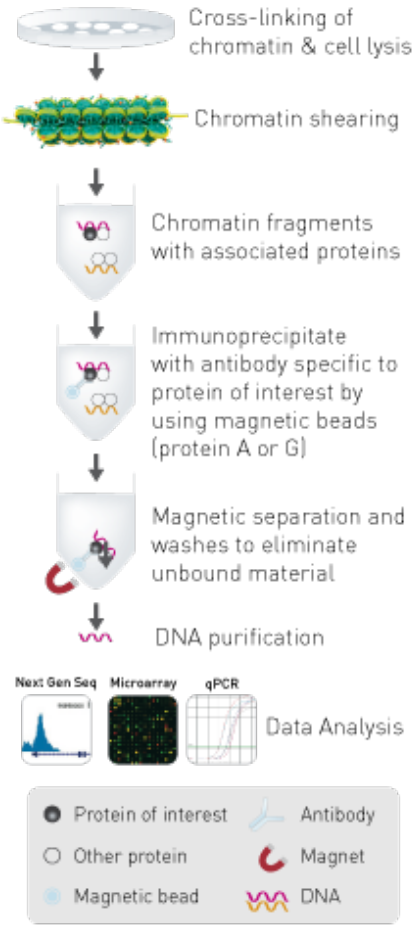
Mission and expectations

- ➢ the collaboratory wants you, the wet lab scientist who did the hard work, to be able to play with the data you generated:
  - understand the vocabulary of your analyst/bioinformatician
  - speak the language of your fellow analyst to provide suggestions based on your wet lab choices
  - analyze the data yourself
- ➢ we want you to learn how to drive a car, not to build one
- ➢ 100% command line interface

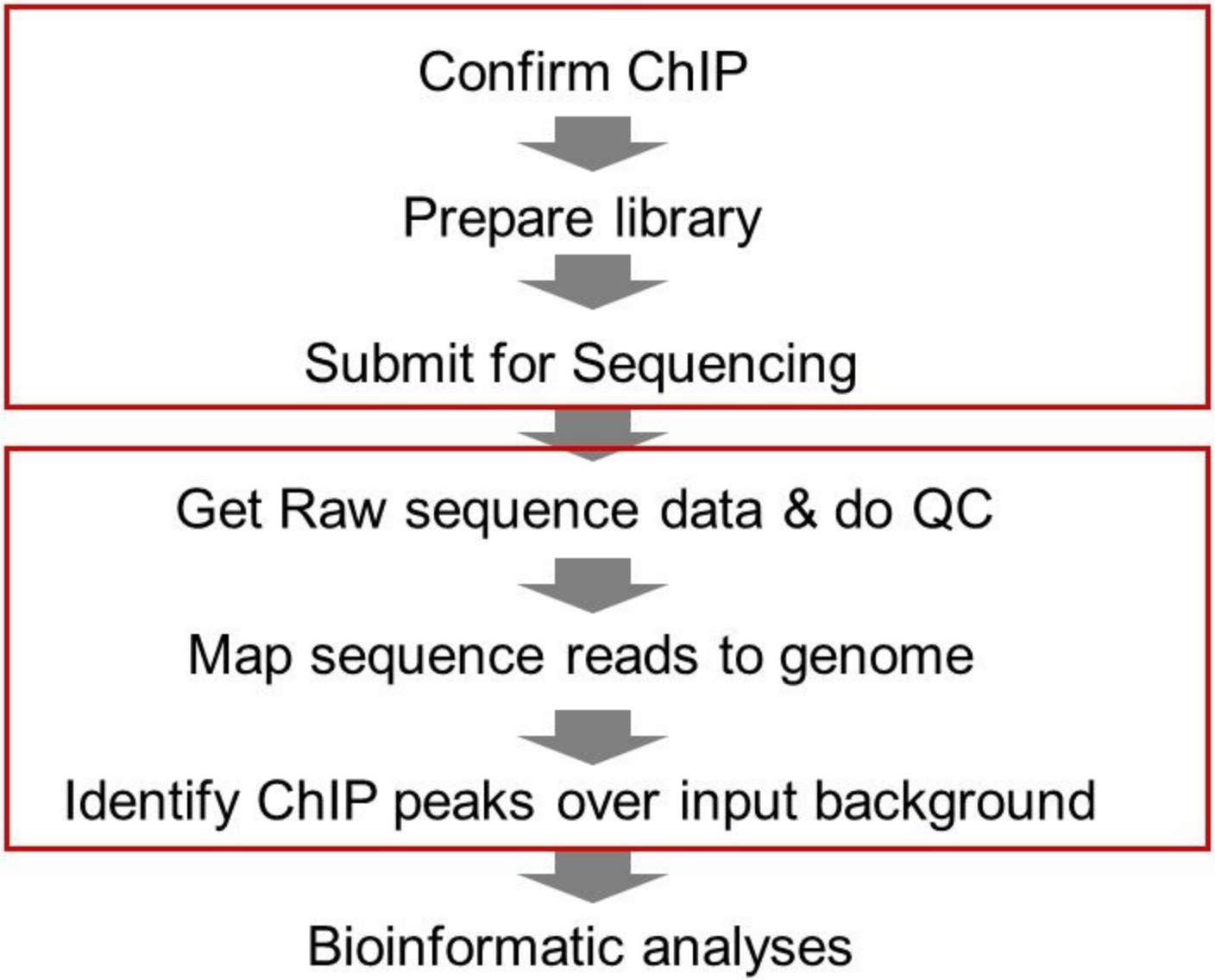
Schedule

- **Day 1**
  - Introduction
  - Cross-correlation analysis and ENCODE QC with SPP
  - BigWig tracks with defined resolution and normalization using Homer/ UCSC tools, and visualization with Genome Browser/IGV
- **Day 2**
  - Peak calling with MACS2
  - QC of replicates with ENCODE's IDR
  - Differential peak calling with MAnorm
- **Day 3**
  - Location annotation with NGSPLOT
  - Motif analysis with HOMER
  - Functional annotation with GREAT
  - (Unix tricks, tools installation)

ChIP workflow



ChIP-seq Workflow



ChIP-Seq: Output

- A list of enriched locations

- Can be used:
  - In combination with RNA-Seq, to determine the biological function of transcription factors
  - Identify genes co-regulated by a common transcription factor
  - Identify common transcription factor binding motifs

ChIP-seq analysis steps in our lab

- Sequencing -> fastq files
- ChIP-seq analyze script (semi-automatic)
  - BAM alignment files
  - Bedgraph files – visualization of peaks
  - Bed files - peak regions
  - Annotation files - (peak positions relative to genes, motif occurrences of two chosen TFBS etc.)
  - GO enrichment analysis
  - denovo motif finding
  - Known motif finding
- Downstream analysis
  - Occupancy analysis
  - Statistics
  - Defining peak subsets, merging, intersecting peaks
  - Comparison of peaksets, subsets
  - Re-doing certain analysis on peaksets, subsets

HOMER

HOMER

Software for motif discovery and next-gen sequencing analysis

---

Next-Generation Sequencing Analysis

ChIP-Seq is the best thing that happened to ChIP since the antibody. It is 100x better than ChIP-Chip since it escapes most of the problems of microarray probe hybridization. Plus it is cheaper, and genome wide. But ChIP-Seq is only the tip of the iceberg - there are many inventive ways to use a sequencer. Below are a list of the the more popular methods that will be covered below:

ChIP-Seq: Isolation and sequencing of genomic DNA "bound" by a specific transcription factor, covalently modified histone, or other nuclear protein. This methodology provides genome-wide maps of factor binding. Most of HOMER's routines cater to the analysis of ChIP-Seq data.

DNase-Seq: Treatment of nuclei with a restriction enzyme such as DNase I will result in cleavage of DNA at accessible regions. Isolation of these regions and their detection by sequencing allows the creation of DNase hypersensitivity maps, providing information about which regulatory elements are accessible in the genome.

MNase-Seq: Micrococcal Nuclease (MNase) is a restriction enzyme that degrades genomic DNA not wrapped around histones. The remaining DNA represents nucleosomal DNA, and can be sequencing to reveal nucleosome positions along the genome. This method can also be combined with ChIP to map nucleosomes that contain specific histone modifications.

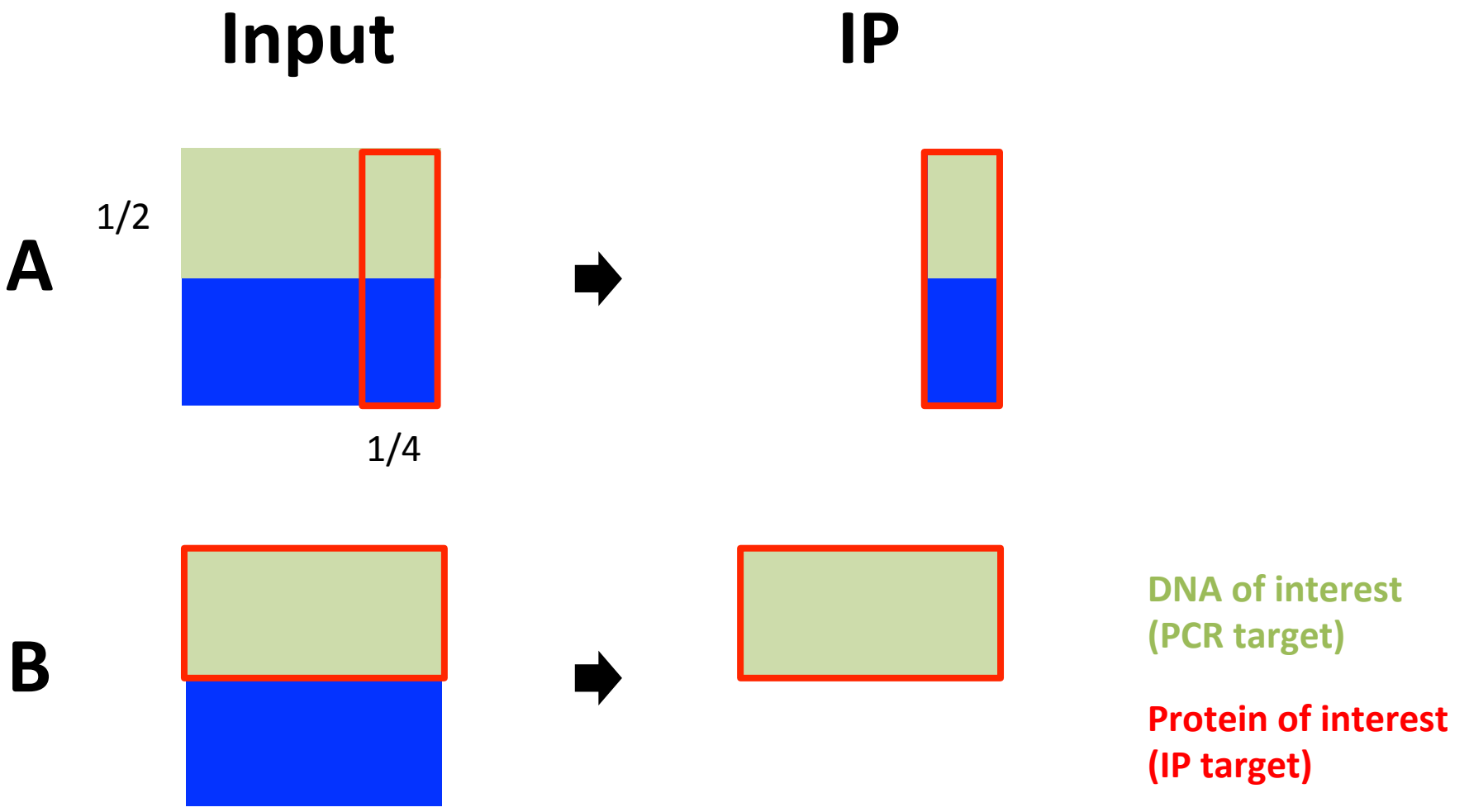
RNA-Seq: Extraction, fragmentation, and sequencing of RNA populations within a sample. The replacement for gene expression measurements by microarray. There are many variants on this, such as Ribo-Seq (isolation of ribosomes translating RNA), small RNA-Seq (to identify miRNAs), etc.

GRO-Seq: RNA-Seq of nascent RNA. Transcription is halted, nuclei are isolated, labeled nucleotides are added back, and transcription briefly restarted resulting in labeled RNA molecules. These newly created, nascent RNAs are isolated and sequenced to reveal "rates of transcription" as opposed to the total number of stable transcripts measured by normal RNA-seq.

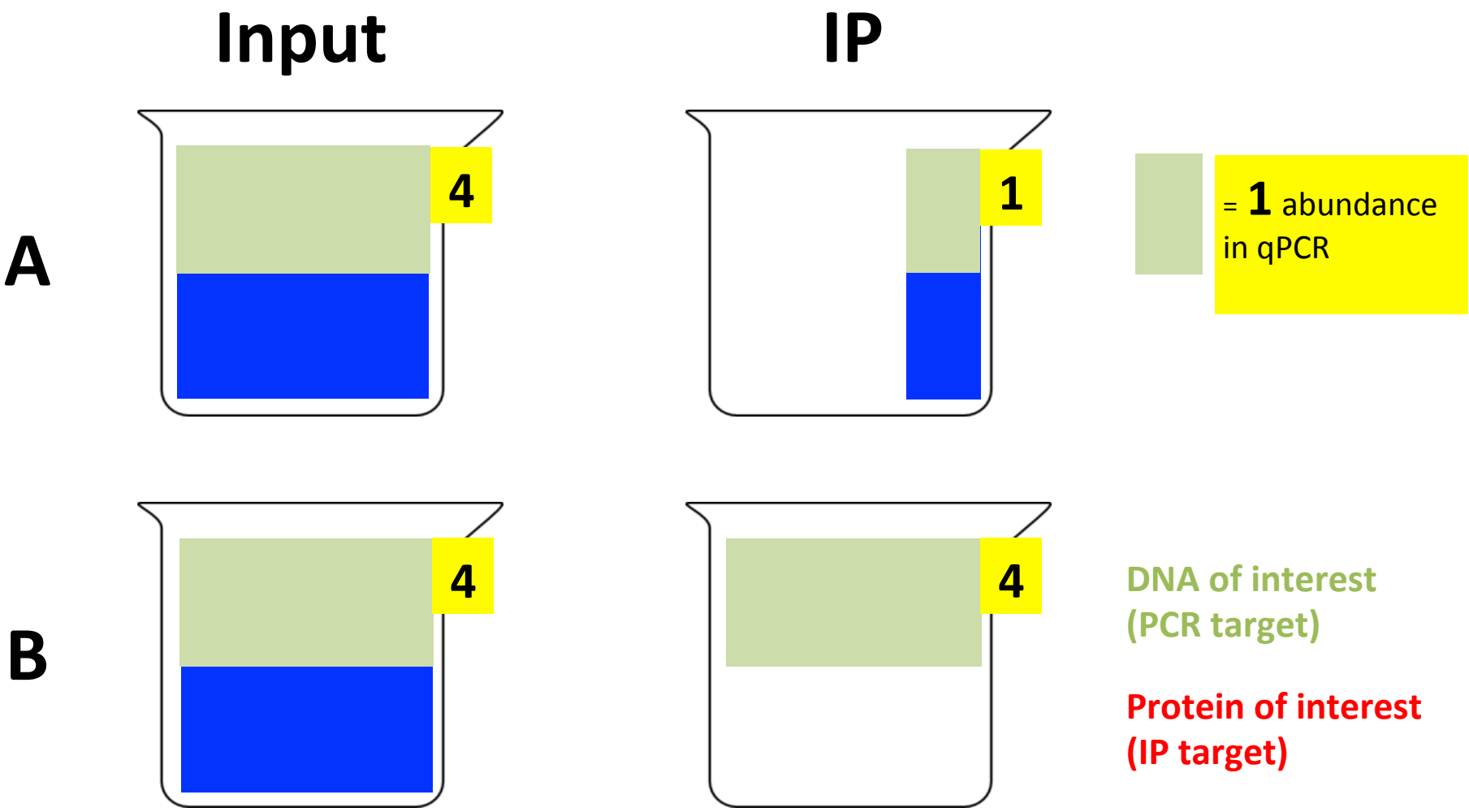
Hi-C: Genomic interaction assay for understanding genome 3D structure. This assay is much more specialized - For more information about how to use HOMER to analyze Hi-C data, check out the Hi-C analysis section.

NORMALIZATION

Icebreaker: ChIP-qPCR normalization

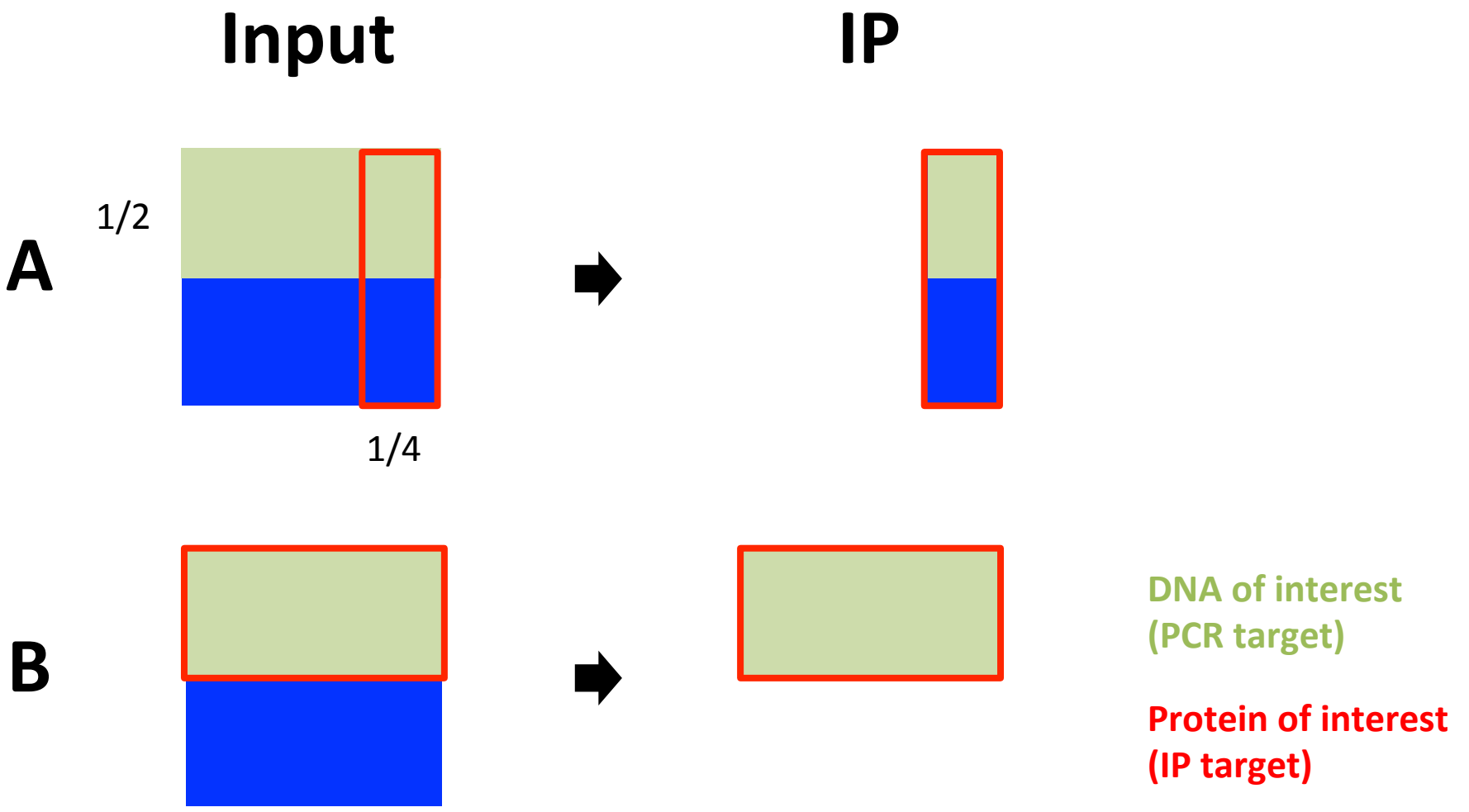


qPCR by equal volumes

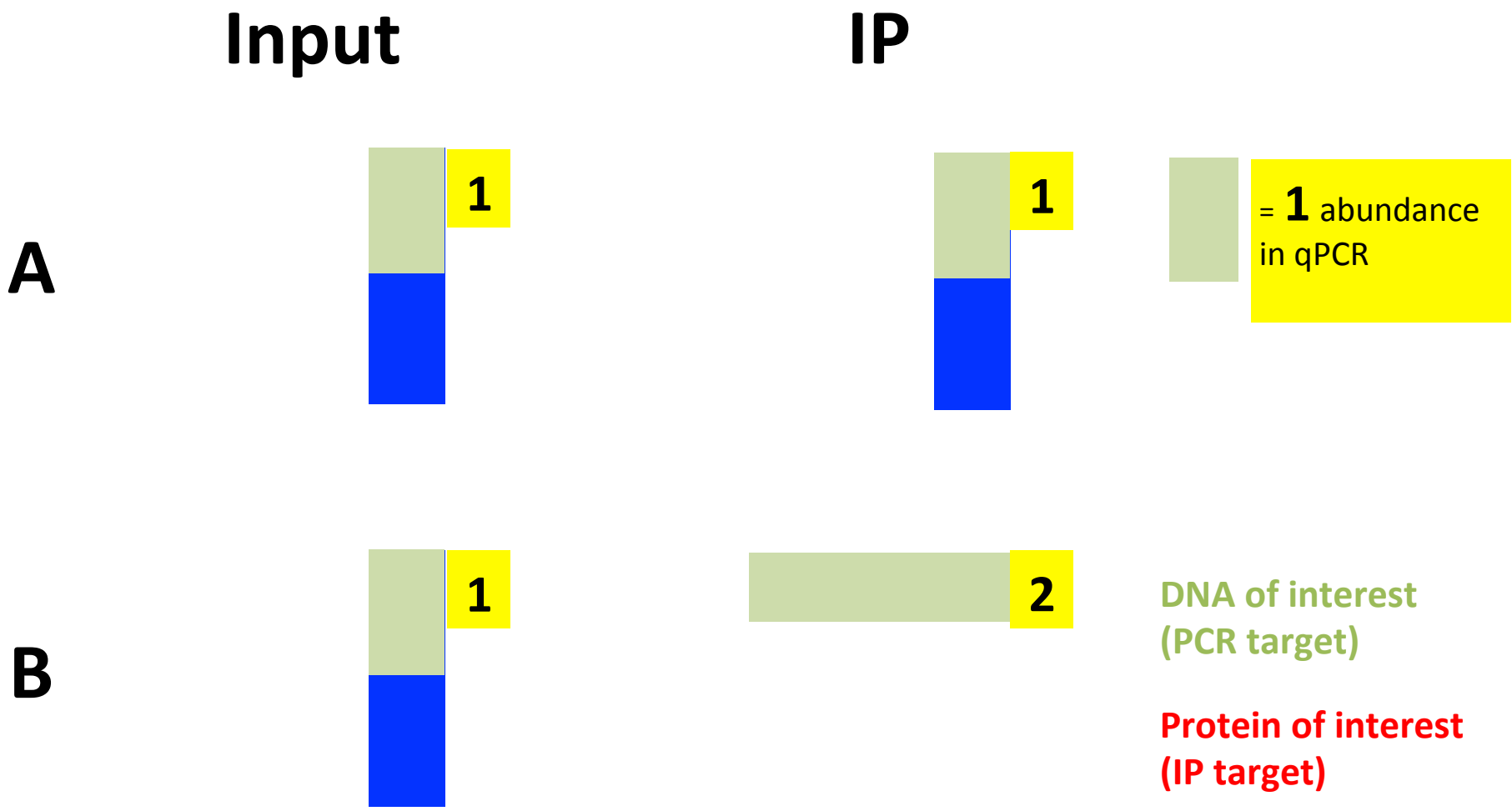


*"DNA of interest is enriched by **4** fold in B compared to A"*

Icebreaker: ChIP-qPCR normalization



qPCR by equal mass



*"DNA of interest is enriched by **2** fold in B compared to A"*

Why should I care for ChIP-seq?

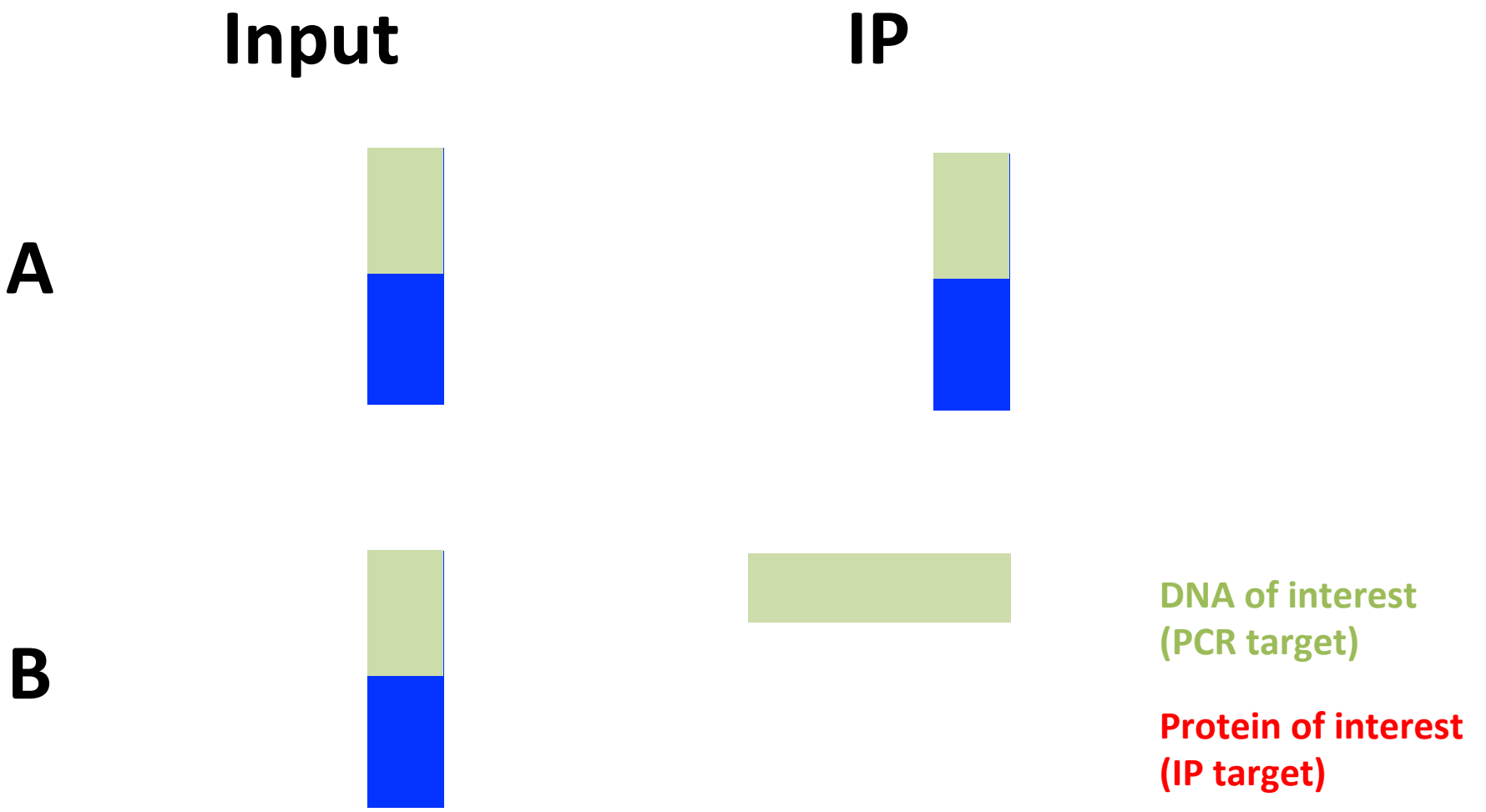
This is you

This is your library

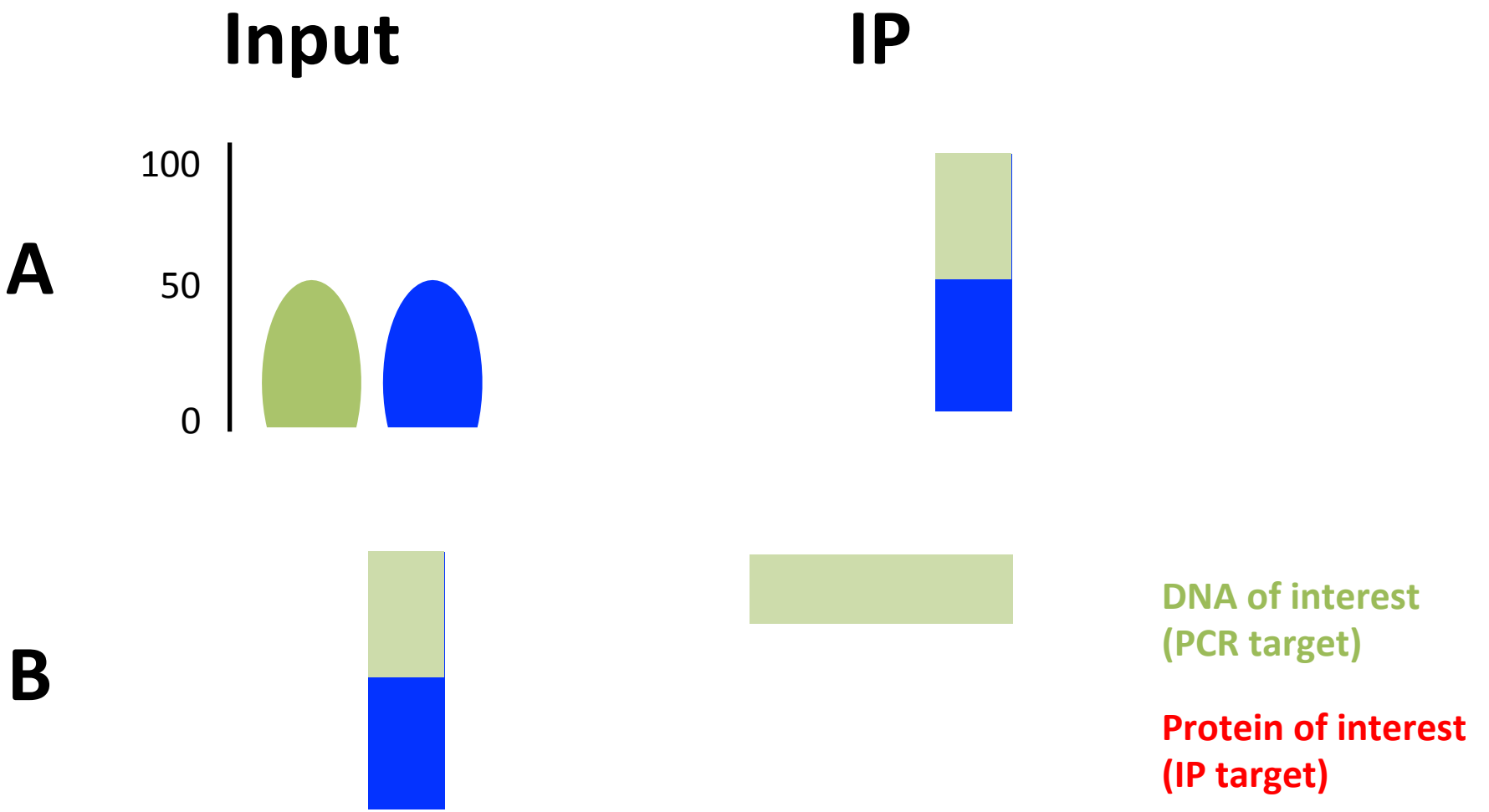
In ChIP-seq, you only normalize by <u>mass</u>:

- 10 nM
- n multiplexed samples
- balanced ratios

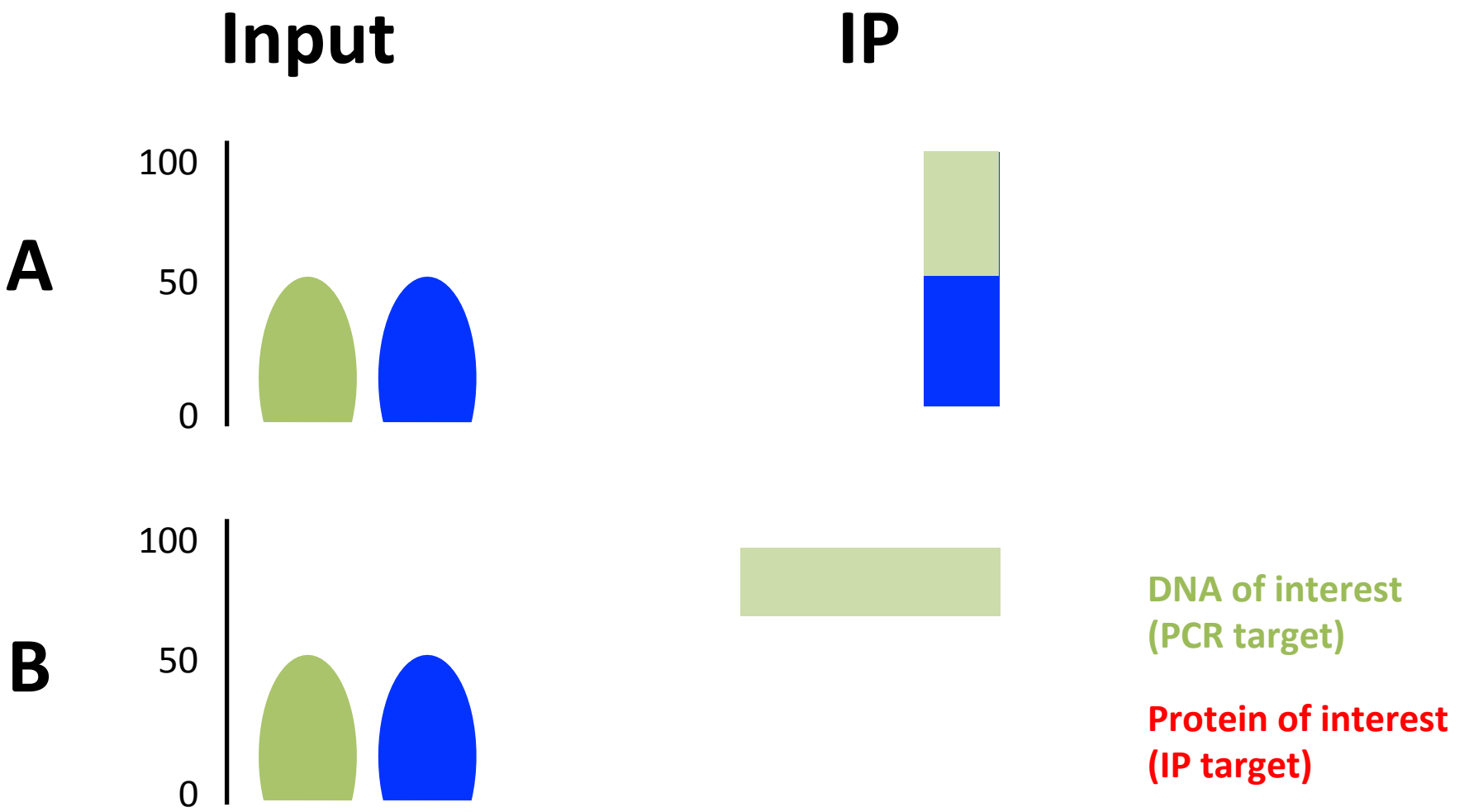
~~qPCR~~ seq by equal mass



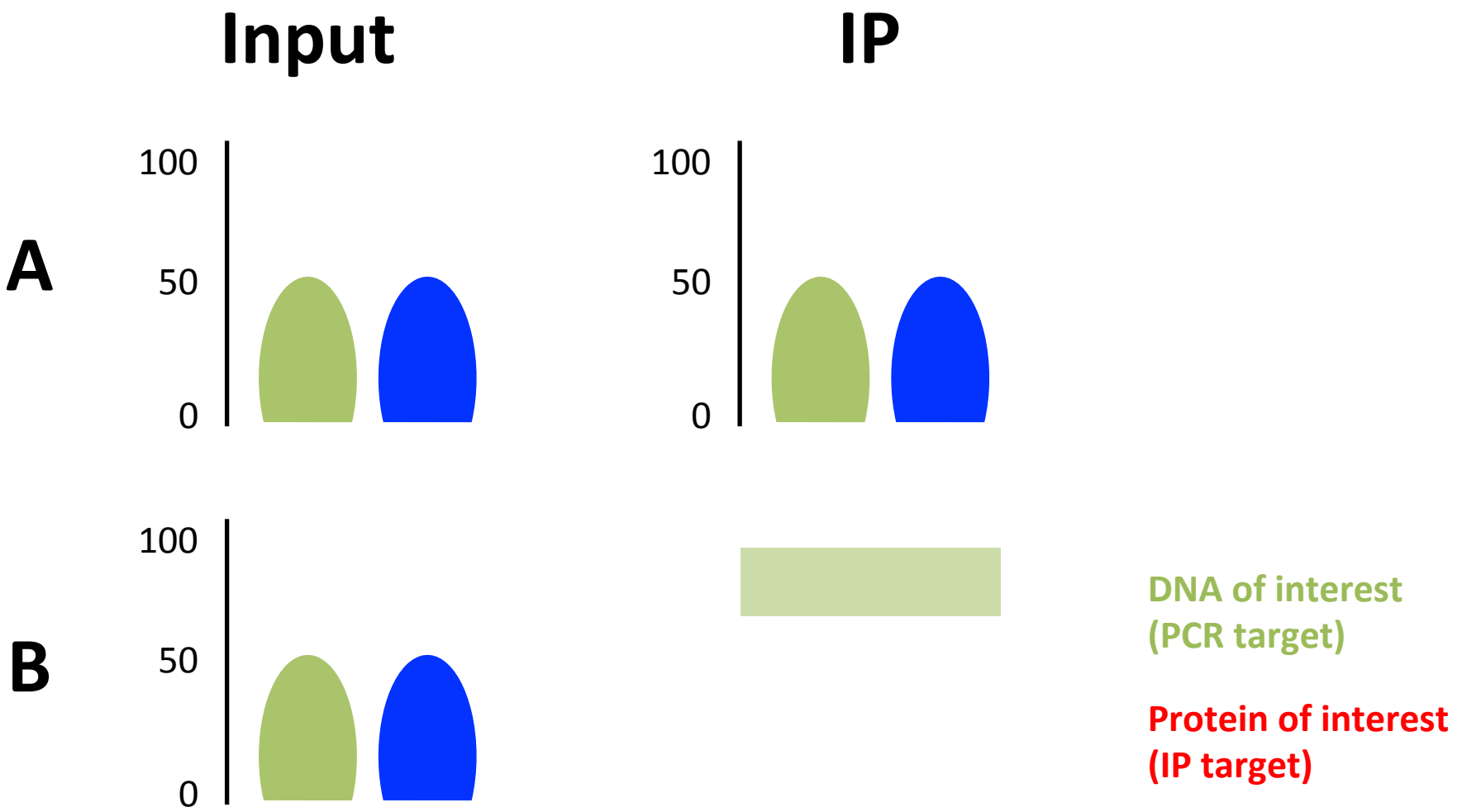
~~qPCR~~ seq by equal mass



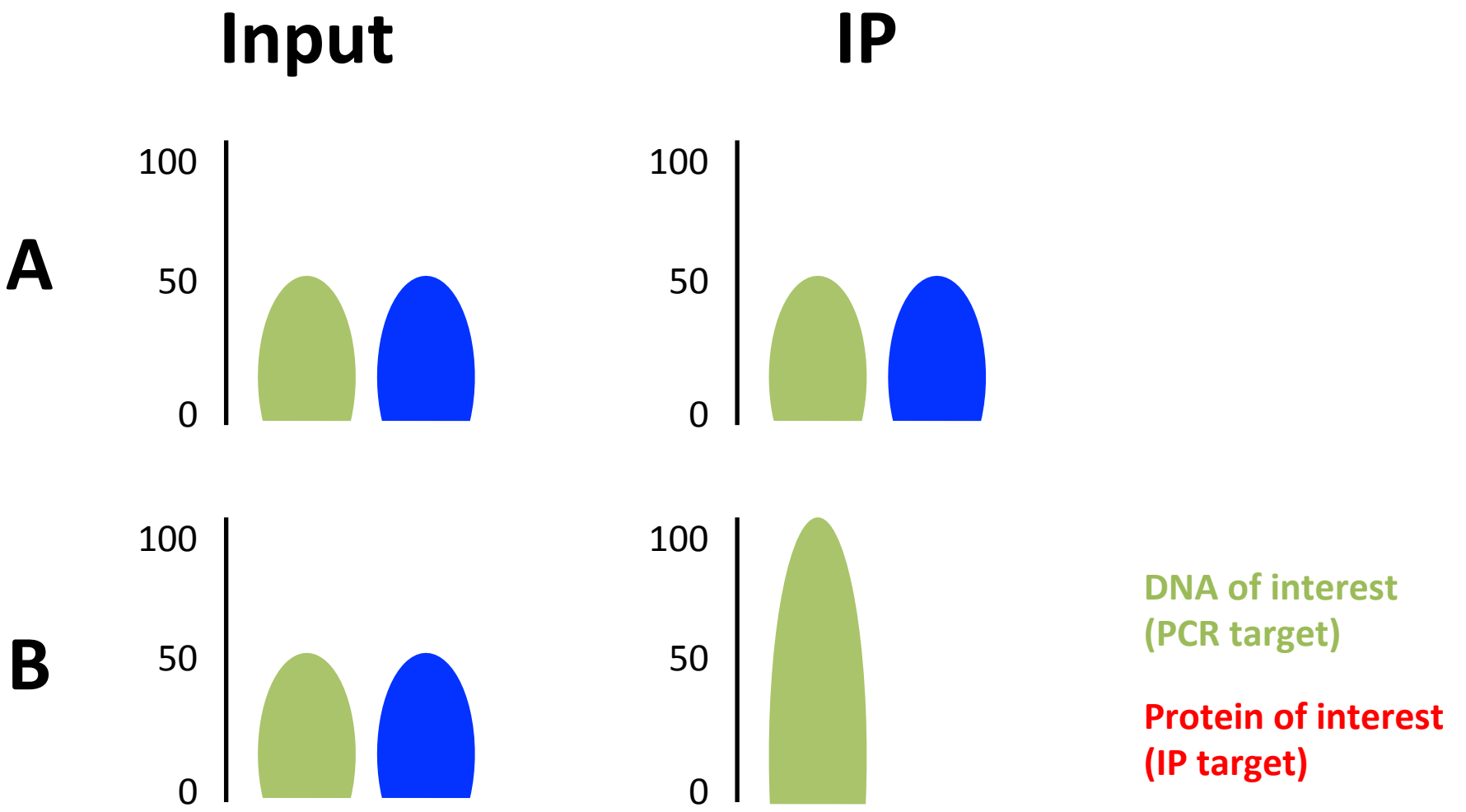
~~qPCR~~ seq by equal mass



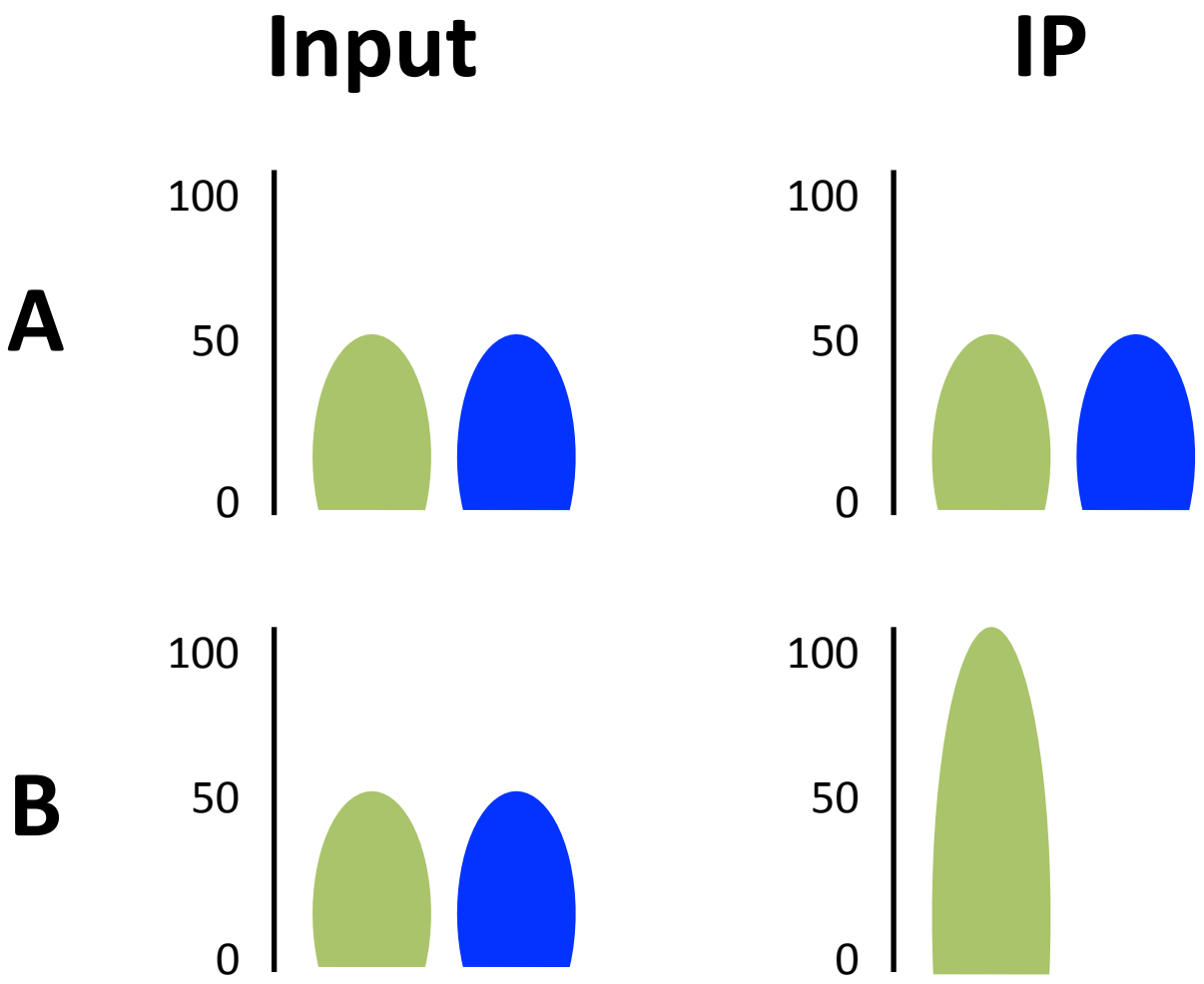
~~qPCR~~ seq by equal mass



~~qPCR~~ seq by equal mass



~~qPCR~~ seq by equal mass



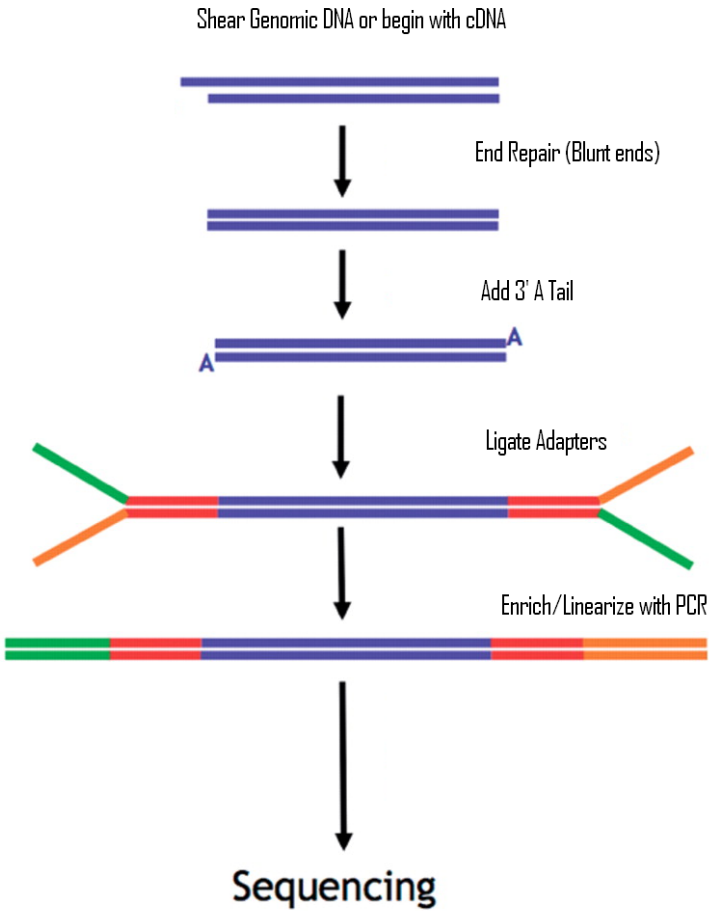
DNA of interest (PCR target)

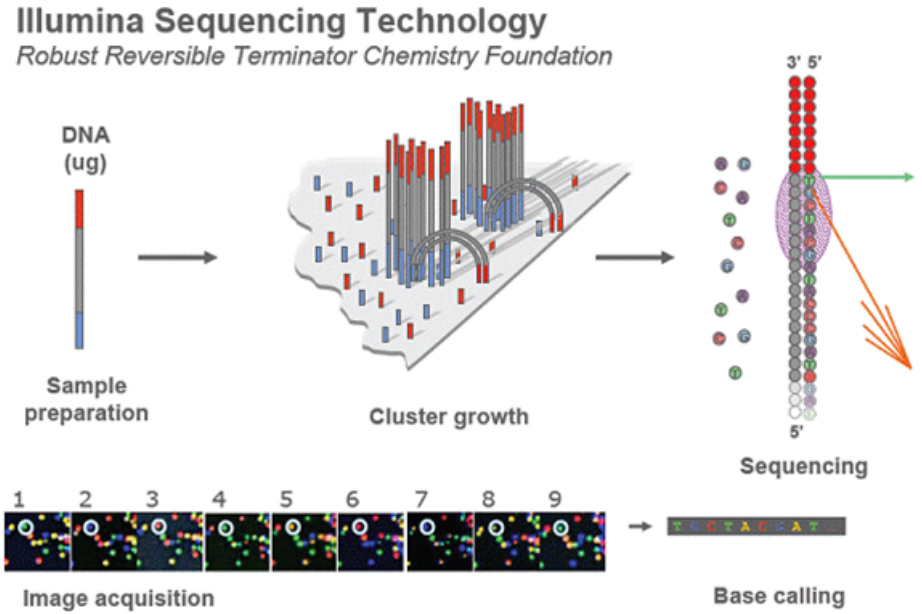
Protein of interest (IP target)

DNA not of interest

*"DNA of interest is enriched by **2** fold in B compared to A"*

Workshops 2: Sequencing





Workshop 2: Pre-processing

Single-end 50 bp best compromise between information and cost in ChIP-seq.

- Convert QSEQ to FASTQ
- Demultiplex > FASTQC
- End trimming > FASTQC
  - Hard trimming
  - Quality-based trimming (e.g. Sickle)
  - Adapter clipping (e.g. Cutadapt, Scythe, Trimmomatic)
- Alignment (e.g. BWA, bowtie, bowtie2)
  - QC (e.g. samtools flagstat, Picard tools, Qualimap)
  - Remove non-mappers and multi-mappers (samtools)
  - Duplicates: usually remove (samtools, Picard tools), but stay tuned

mixed-length reads!
can your tool handle them?

If you need to refresh these concepts, I am happy to stay longer after class

Quality Control

Basic Tag information

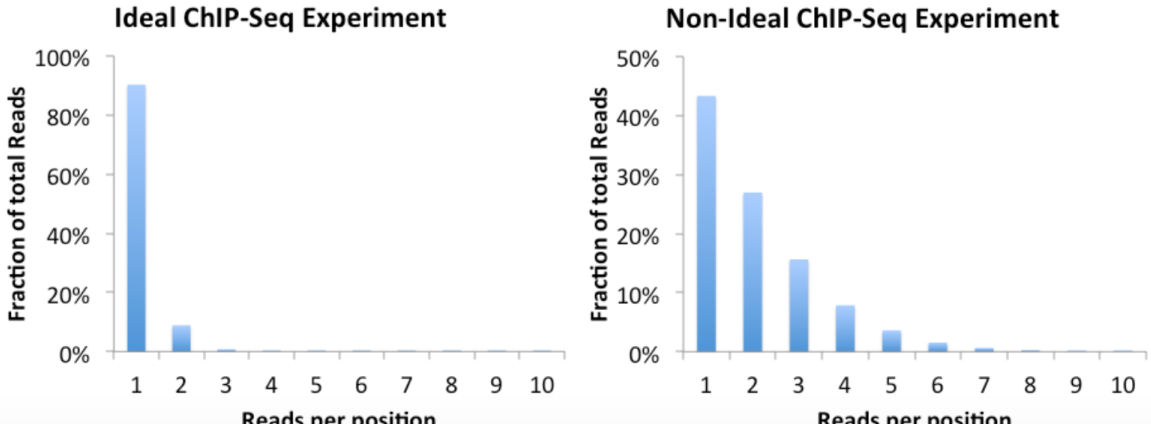
tagInfo.txt - Contains basic configuration information, such as the total number of reads, the total number of unique positions with aligned reads (by genome and chromosome), and various other statistics. One of the more important parameters is "fragmentLengthEstimate=##", which provides an estimate of the length of fragments used for sequencing.

Read Length Distribution

tagLengthDistribution.txt - File contains a histogram of read lengths used for alignment.

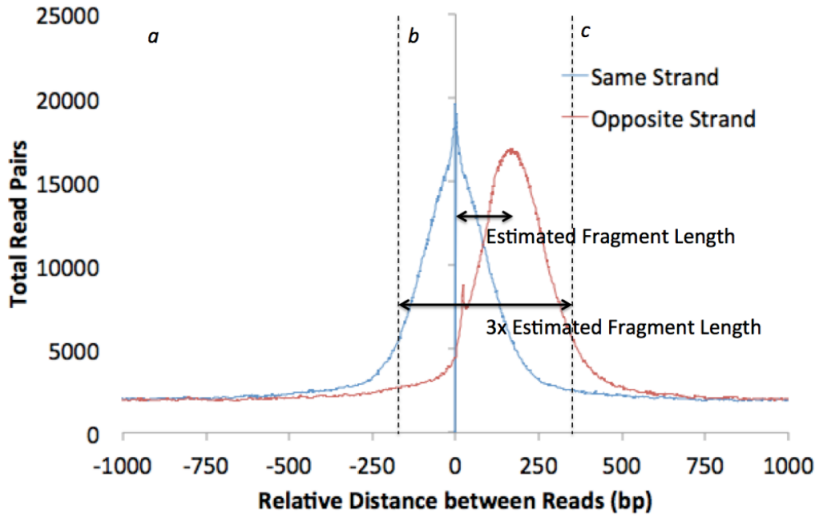
Clonal Tag Distribution

tagCountDistribution.txt - File contains a histogram of clonal read depth, showing the number of reads per unique position. If an experiment is "over-sequenced", you start seeing the same reads over and over instead of unique reads. Sometimes this is a sign there was not enough starting material for sequencing library preparation. Below are examples of ideal and non-ideal results - in the case of the non-ideal experiment, you probably don't want to sequence that library anymore.



Quality Control

Below is a schematic to visualize how these are calculated (keep in mind that the background is calculated out to +/- 2kb):



Same strand and Diff Strand Fold Enrichment: $\frac{\text{Density } b}{\text{Density } a \text{ \& } c}$

Same/Diff Fold Enrichment: $\frac{\text{Density } b \text{ (same strand)}}{\text{Density } b \text{ (Opposite Strand)}}$

Depending on the value of the autocorrelation quality statistics, HOMER will guess what your experiment is:

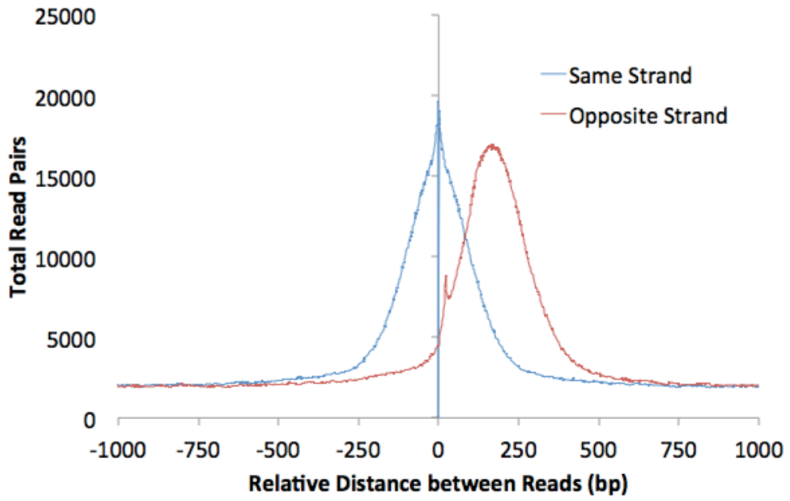
- If the Same/Diff Fold Enrichment is > 8 fold, it's a great chance the sample is strand-specific RNA. If HOMER decides the sample is probably RNA due to the difference between same strand and different strand numbers, it will automatically set the estimated fragment length to 75 bp. This is because it is difficult to estimate the fragment length for RNA/GRO-seq. To manually set the fragment length, use "**-fragLength <#>**"
- If both the "Same Strand Fold Enrichment" and "Diff Strand Fold Enrichment" are both greater than 1.5 fold, there is good chance you're looking at a working ChIP-Seq experiment.

Quality Control

Autocorrelation Analysis

tagAutocorrelation.txt - The autocorrelation routine creates a distribution of distances between adjacent reads in the genome. If reads are mapped to the same strand, they are added to the first column. If adjacent reads map to different strands, they are added to the 2nd column. The results from autocorrelation analysis are very useful for troubleshooting problems with the experiment, and are used to estimate the fragment length for ChIP-Seq and MNase-Seq. The fragment length is estimated by finding the position were the "opposite strand" distribution is maximum. HOMER will use this value as the fragment length unless overridden with the option "**-fragLength <#>**".

Different types of experiments (i.e. ChIP-Seq vs. DNase-Seq) produce different looking autocorrelation plots, and more detailed discussion of these differences can be found in the individual tutorials. Below is an example from a successful ChIP-Seq experiment:

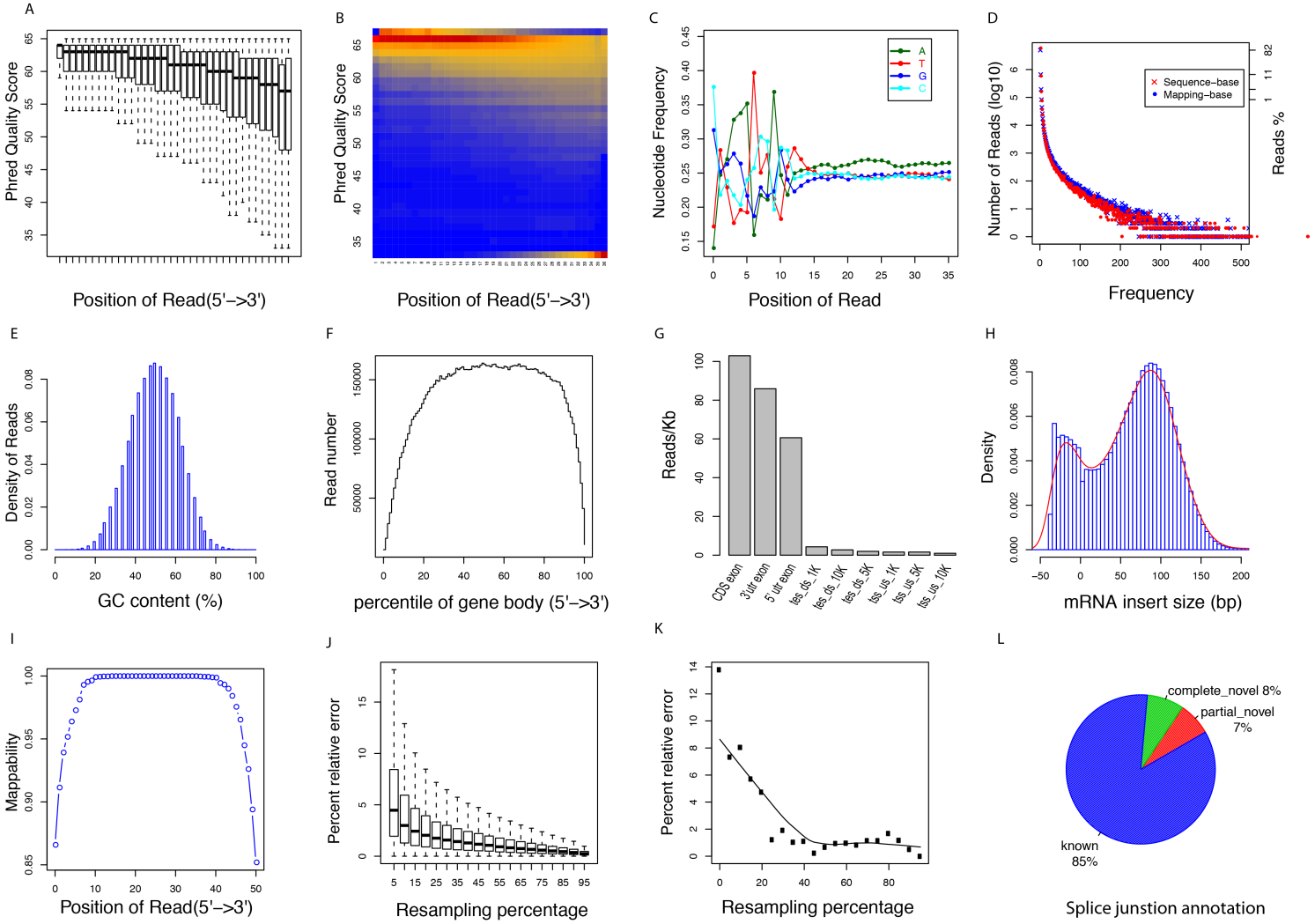


HOMER also uses the autocorrelation results to guess what type of experiment you conducted. It computes 3 statistics:

- Same strand fold enrichment: Enrichment of reads on the same strand within 3x the estimated fragment length
- Diff strand fold enrichment: Enrichment of reads on different strands within 3x the estimated fragment length
- Same / Diff fold enrichment: Difference between enrichment of reads on the same strand or different strands

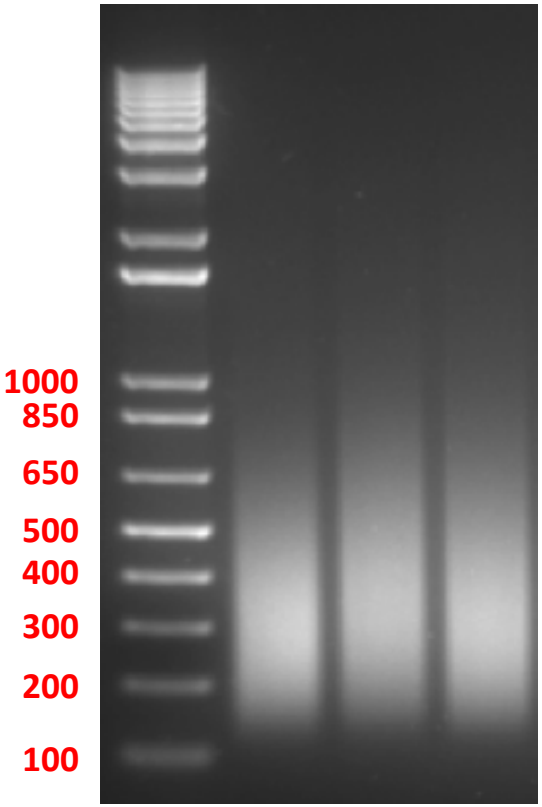
chip class

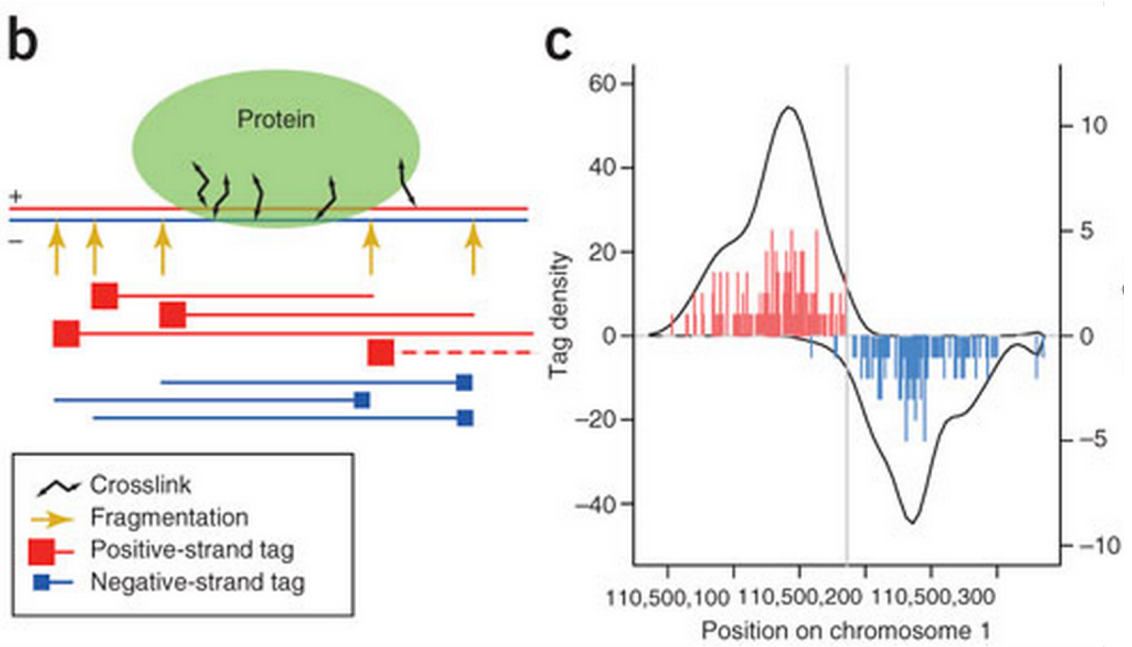
Other Quality Control



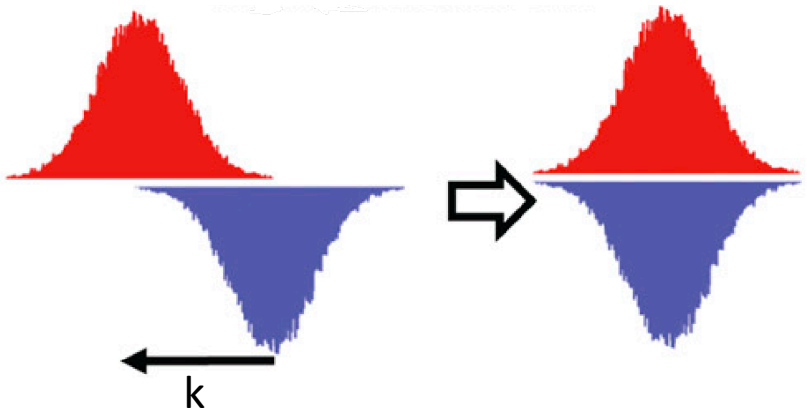
CROSS-CORRELATION QC

Strand cross-correlation



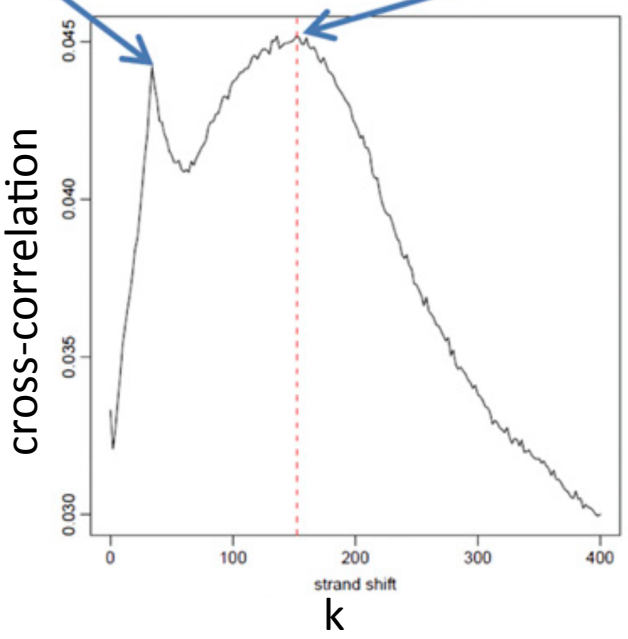


Cross-correlation to estimate fragment length



5’ read end is being shifted

read length = **“phantom” peak** **ChIP peak** = fragment length

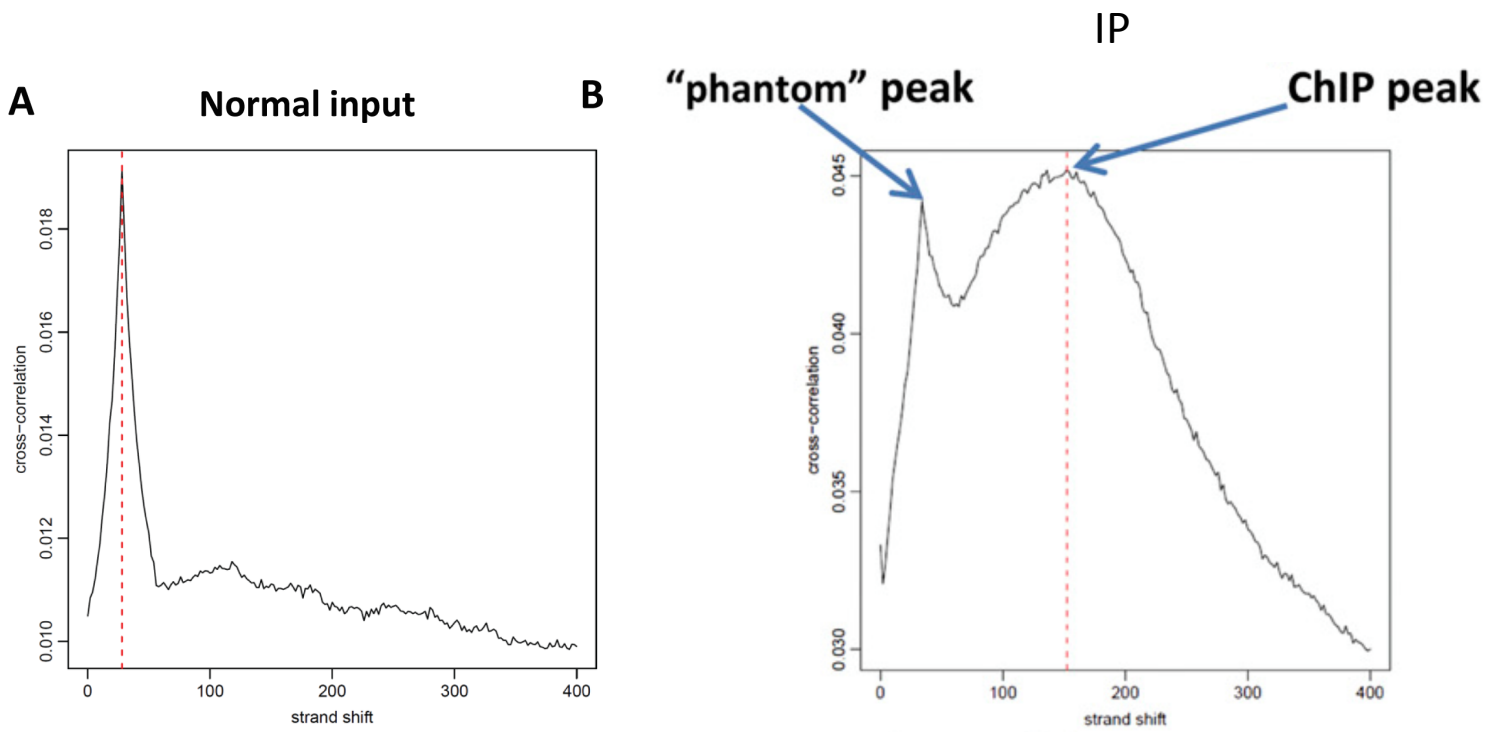


Don’t I know already?

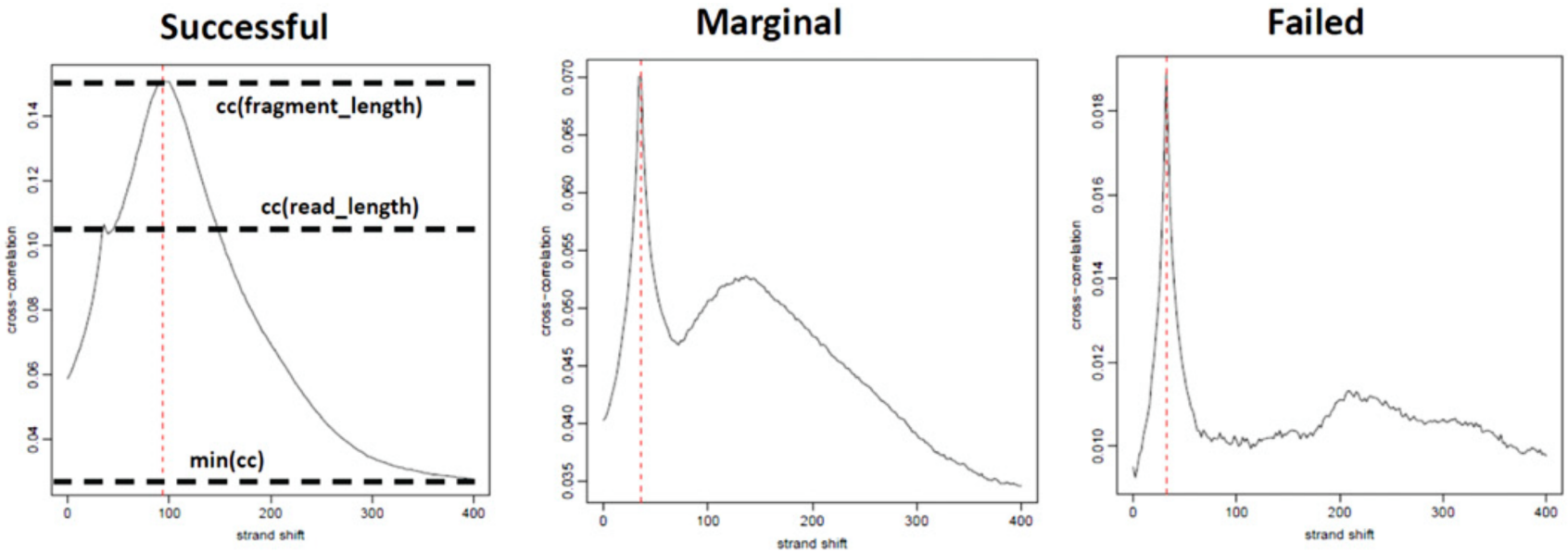
- gel/Bioanalyzer

  But bias may occur during:
  - IP
  - library prep
  - sequencing
- PE reads

Input vs IP


A
Normal input
B
IP
“phantom” peak
ChIP peak
cross−correlation
strand shift
0.010
0.012
0.014
0.016
0.018
0.030
0.035
0.040
0.045
0
100
200
300
400

Cross-correlation for QC



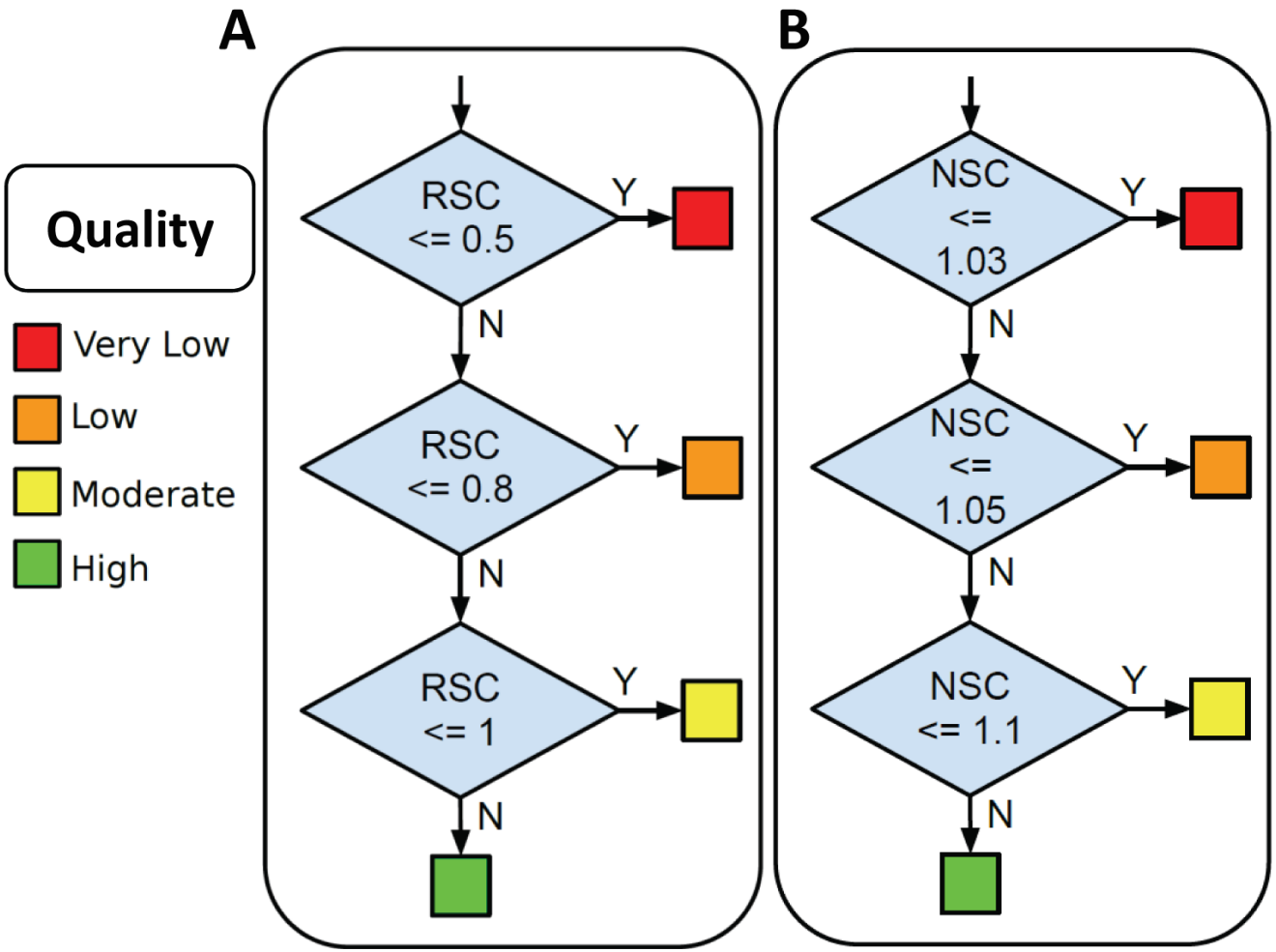
$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

NSC: Normalized Strand cross-correlation Coefficient

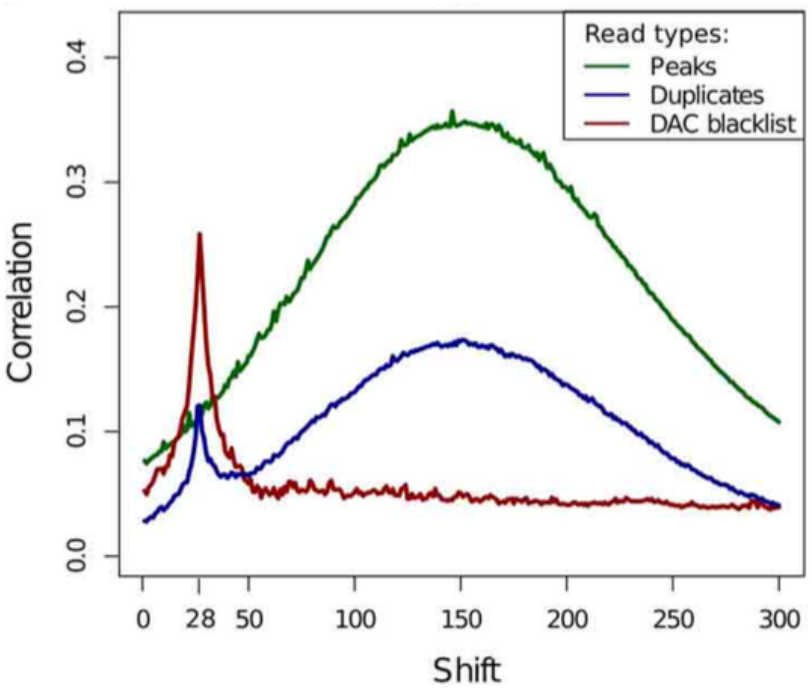
RSC: Relative Strand cross-correlation Coefficient

How good is your sample for ENCODE?



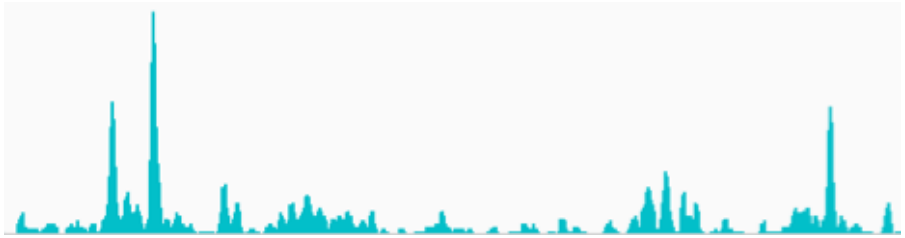
But this is affected by target, antibody, cell type, fragment distribution…

Phantom peak



- Removing blacklisted regions from your alignments makes sense
- However, that invalidates SPP QC: be careful when building workflows
- If anybody is interested, I can show how to remove blacklisted regions using bedtools

VISUALIZATION



Many tools for the job

MACS

HOMER

http://homer.salk.edu/

deepTools bamCoverage

MAX PLANCK INSTITUTE
OF IMMUNOBIOLOGY AND EPIGENETICS

~~bedtools genomecov~~

HOMER

Software for motif discovery and next-gen sequencing analysis

---

Visualizing Experiments with a Genome Browser

The UCSC Genome Browser is quite possibly one of the best computational tools ever developed. Not only does it contain an incredible amount of data in a single application, it allows users to upload custom information such as data from their ChIP-Seq experiments so that they can be easily visualized and compared to other information. There are also other genome browsers that are available, and each has a different strength:

UCSC Genome Browser

: Truly a unique resource, logs of data preloaded and annotations.

WashU Epigenome Browser

: Capable of visualizing long-range interactions (great for data sets like Hi-C), also has a lot of preloaded data.

IGV

: The Integrated Genomics Viewer (IGV), great for looking at reads locally instead of needing to load them to a server/cloud based solution. Great for directly looking at sorted bam/bai files to examine mutations in reads.

Many others...

Most of the tools that are part of HOMER cater to the strengths of the UCSC Genome Browser - however, the bedGraph and other files generated by HOMER can be normally be used in the other genome browsers as well.

Making Genome Browser Files

The basic strategy HOMER uses is to create a bedGraph formatted file that can then be uploaded as a custom track to the genome browser. This is accomplished using the **makeUCSCfile** program. To make a ucsc visualization file, type the following:

makeUCSCfile <tag directory> -o auto

i.e. **makeUCSCfile PU.1-ChIP-Seq/ -o auto**
(output file will be in the PU.1-ChIP-Seq/ folder named PU.1-ChIP-Seq.ucsc.bedGraph.gz)

The "-o auto" with make the program automatically generate an output file name (i.e. TagDirectory.ucsc.bedGraph.gz) and place it in the tag directory which helps with the organization of all these files. The output file can be named differently by specifying "-o outputfilename" or by simply omitting "-o", which will send the output of the program to *stdout* (i.e. add " > outputfile" to capture it in the file outputfile). It is recommended that you zip the file using **gzip** and directly upload the *zipped* file when loading custom tracks at UCSC.

To visualize the experiment in the UCSC Genome Browser, go to Genome Browser page and select the appropriate genome (i.e. the genome that the sequencing tags were mapped to). Then click on the "add custom tracks" button (this will read "manage custom tracks" once at least one custom track is loaded). Enter the file created earlier in the "Paste URLs or data" section and click "Submit".

Problems Loading UCSC Files

The most common problem encountered while loading UCSC files is to see "position exceeds chromosome length" or something to that effect. This is usually caused by one of two problems:

1. You are trying to load the file to the wrong genome assembly. Make sure the assembly is correct!

2. Did you align the genome to a UCSC version? chr1 != chr1 != 1

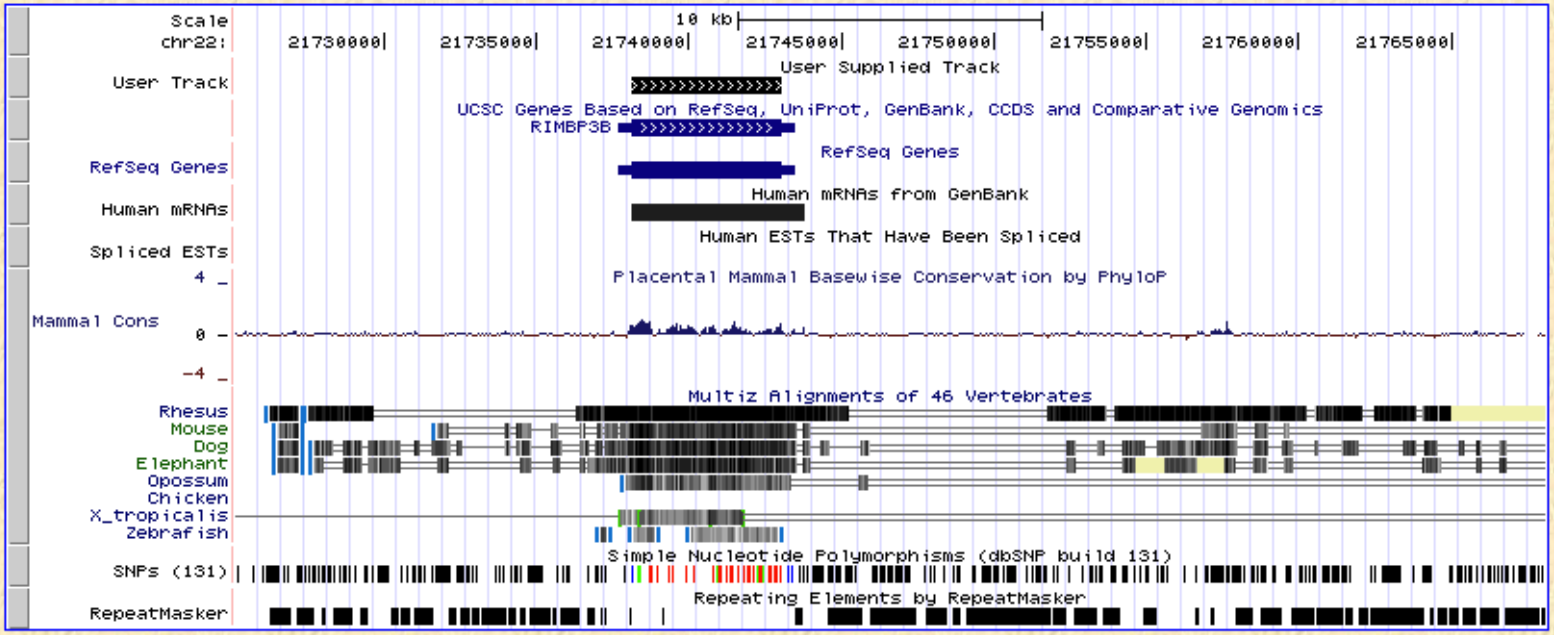
3. Some of your tags are mapping outside the reference chromosome - this can be caused by mapping to non-standard assemblies or by some alignment programs. To remove all reads outside of the UCSC chromosome lengths, you can run the program **removeOutOfBoundsReads.pl**.

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr22:21,725,126-21,768,073 gene jump clear size 42,948 bp. configure



move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks.

move end < 2.0 >

default tracks default order hide all manage custom tracks configure reverse refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

Custom Tracks

refresh

User Track

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal **genome:** Human **assembly:** Dec. 2013 (GRCh38/hg38)

group: Genes and Gene Predictions **track:** GENCODE v24 add custom tracks track hubs

table: knownGene describe table schema

region: genome position chr17:63477071-63498380 lookup define regions

identifiers (names/accessions): paste list upload list

filter: create

intersection: create

correlation: create

output format: all fields from selected table Send output to Galaxy GREAT GenomeSpace

output file: (leave blank to keep output in browser)

file type returned: plain text gzip compressed

get output summary/statistics

To reset **all** user cart settings (including custom tracks), click here.

UCSC Genome Bioinformatics

Genomes | Genome Browser | Tools | Mirrors | Downloads | My Data | Help | About Us

Frequently Asked Questions: Data File Formats

General formats:

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigWig format
- Chain format
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- MAF format
- Microarray format

BED format

BED detail format

Index ▷

This is an extension of BED format. BED detail uses the first 4 to 12 columns of BED format, plus 2 additional fields that are used to enhance the track details pages. The first additional field is an ID, which can be used in place of the name field for creating links from the details pages. The second additional field is a description of the item, which can be a long description and can consist of html, including tables and lists.

Requirements for BED detail custom tracks are: fields must be tab-separated, "type=bedDetail" must be included in the track line, and the name and position fields should uniquely describe items so that the correct ID and description will be displayed on the details pages.

Example:
This example uses the first 4 columns of BED format, but up to 12 may be used. Click here to view this track in the Genome Browser.

```
track name=HbVar type=bedDetail description="HbVar custom track" db=hg19 visibility=3 url="http://globi
chr11   5246919 5246920 Hb_North_York   2619    Hemoglobin variant
chr11   5255660 5255661 HBD c.1 G>A     2659    delta0 thalassemia
chr11   5247945 5247946 Hb Sheffield    2672    Hemoglobin variant
chr11   5255415 5255416 Hb A2-Lyon      2676    Hemoglobin variant
chr11   5248234 5248235 Hb Aix-les-Bains         2677    Hemoglobin variant
```

bedGraph format

Index ▷

The bedGraph format allows display of continuous-valued data in track format. This display type is useful for probability scores and transcriptome data. This track type is similar to the WIG format, but unlike the WIG format, data exported in the bedGraph format are preserved in their original state. This can be seen on export using the table browser. For more information about the bedGraph format, please see the bedGraph details page.

If you have a very large data set and you would like to keep it on your own server, you should use the bigWig format.

BED detail format

BED detail format

Index ▷

This is an extension of BED format. BED detail uses the first 4 to 12 columns of BED format, plus 2 additional fields that are used to enhance the track details pages. The first additional field is an ID, which can be used in place of the name field for creating links from the details pages. The second additional field is a description of the item, which can be a long description and can consist of html, including tables and lists.

Requirements for BED detail custom tracks are: fields must be tab-separated, "type=bedDetail" must be included in the track line, and the name and position fields should uniquely describe items so that the correct ID and description will be displayed on the details pages.

Example:
This example uses the first 4 columns of BED format, but up to 12 may be used. Click here to view this track in the Genome Browser.

```
track name=HbVar type=bedDetail description="HbVar custom track" db=hg19 visibility=3 url="http://globi
chr11   5246919 5246920 Hb_North_York   2619    Hemoglobin variant
chr11   5255660 5255661 HBD c.1 G>A     2659    delta0 thalassemia
chr11   5247945 5247946 Hb Sheffield    2672    Hemoglobin variant
chr11   5255415 5255416 Hb A2-Lyon      2676    Hemoglobin variant
chr11   5248234 5248235 Hb Aix-les-Bains        2677    Hemoglobin variant
```

bedGraph format

Index ▷

The bedGraph format allows display of continuous-valued data in track format. This display type is useful for probability scores and transcriptome data. This track type is similar to the WIG format, but unlike the WIG format, data exported in the bedGraph format are preserved in their original state. This can be seen on export using the table browser. For more information about the bedGraph format, please see the bedGraph details page.

If you have a very large data set and you would like to keep it on your own server, you should use the bigWig format.

Let's practice

Work from your samples folder, e.g. /u/scratch/r/rspreafi/samples

mkdir homer

module load samtools/1.2

module load homer

makeTagDirectory homer/IPr1 -keepAll -fragLength 200 -precision 3 IPr1.bam >homer/IPr1_tag_output.txt 2>&1 &

You can repeat this for all your samples… but not needed now.

BigWig format

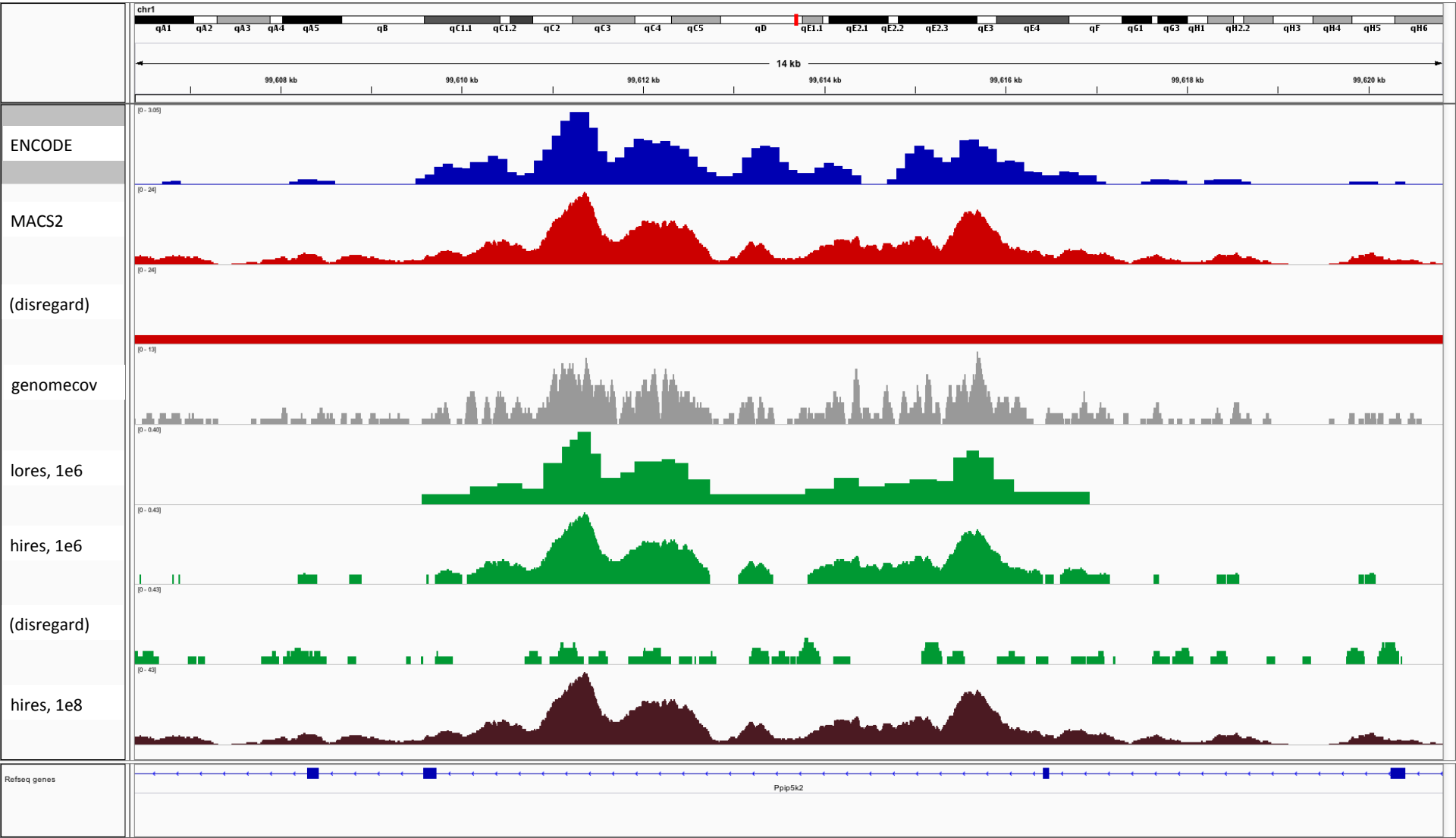
- Used by UCSC Genome Browser
- Standard *de facto*
- hgdownload.cse.ucsc.edu/admin/exe/

tail -n +2 IPr1.bedGraph | sort -k1,1 -k2,2n > IPr1_sorted.bedGraph

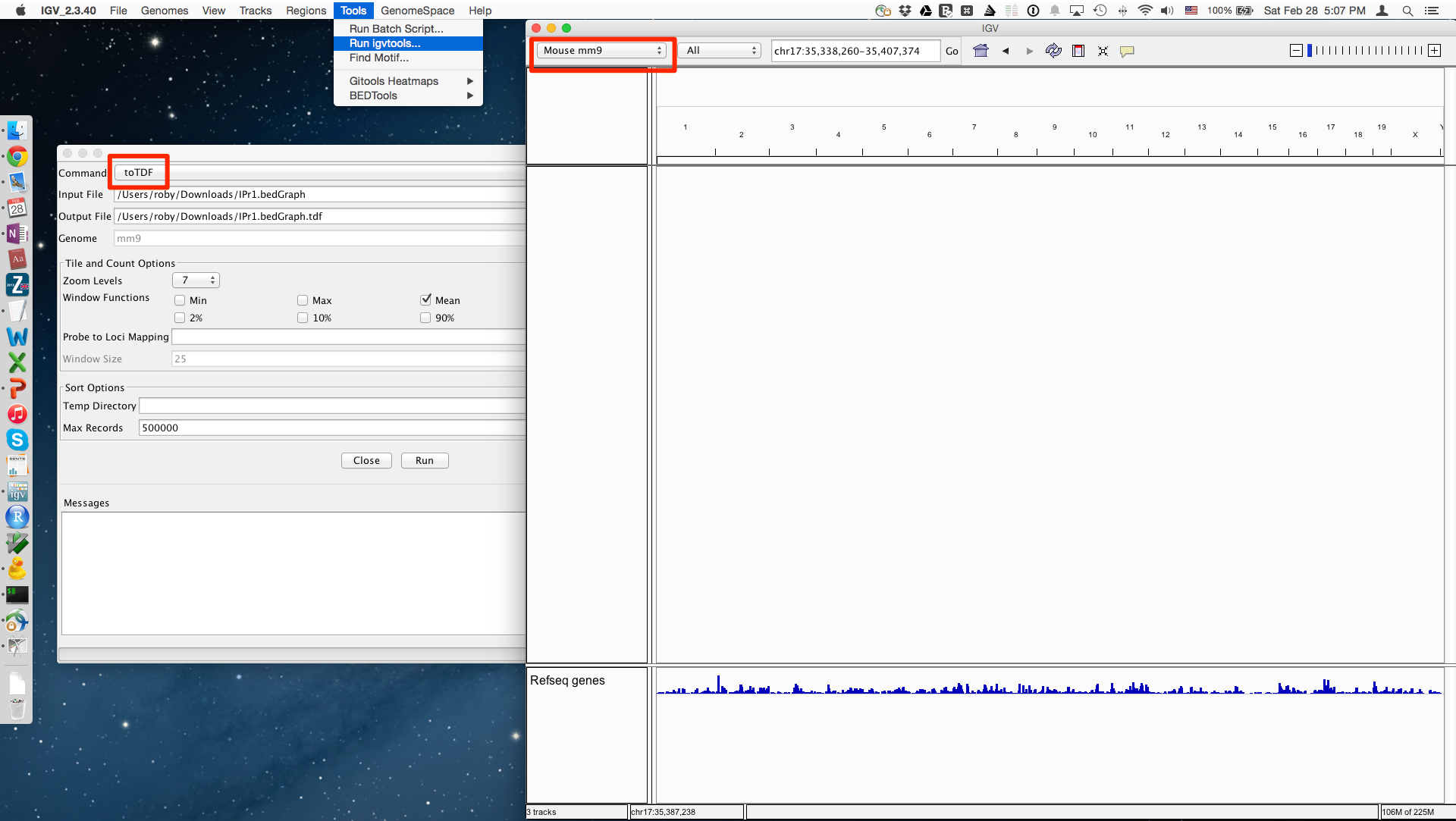
fetchChromSizes.sh mm9 >mm9.chrom.sizes

bedGraphToBigWig IPr1_sorted.bedGraph mm9.chrom.sizes IPr1.bw

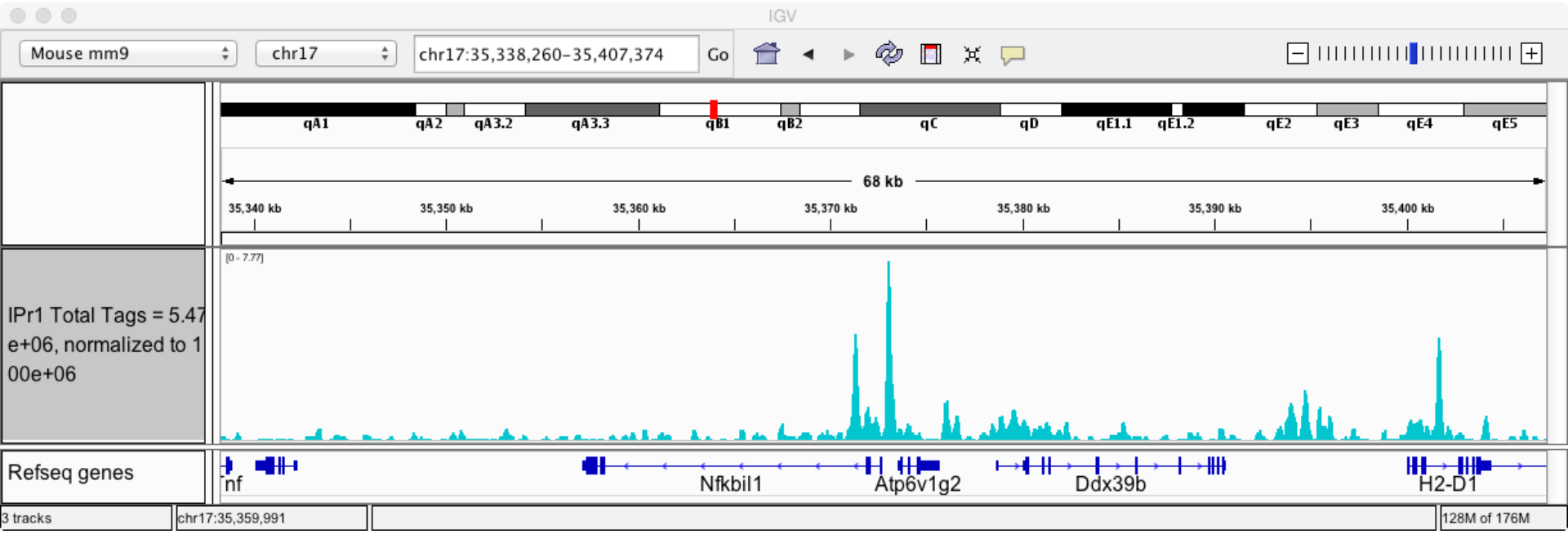
Example of visualization



Retrieve your track and visualize it with IGV

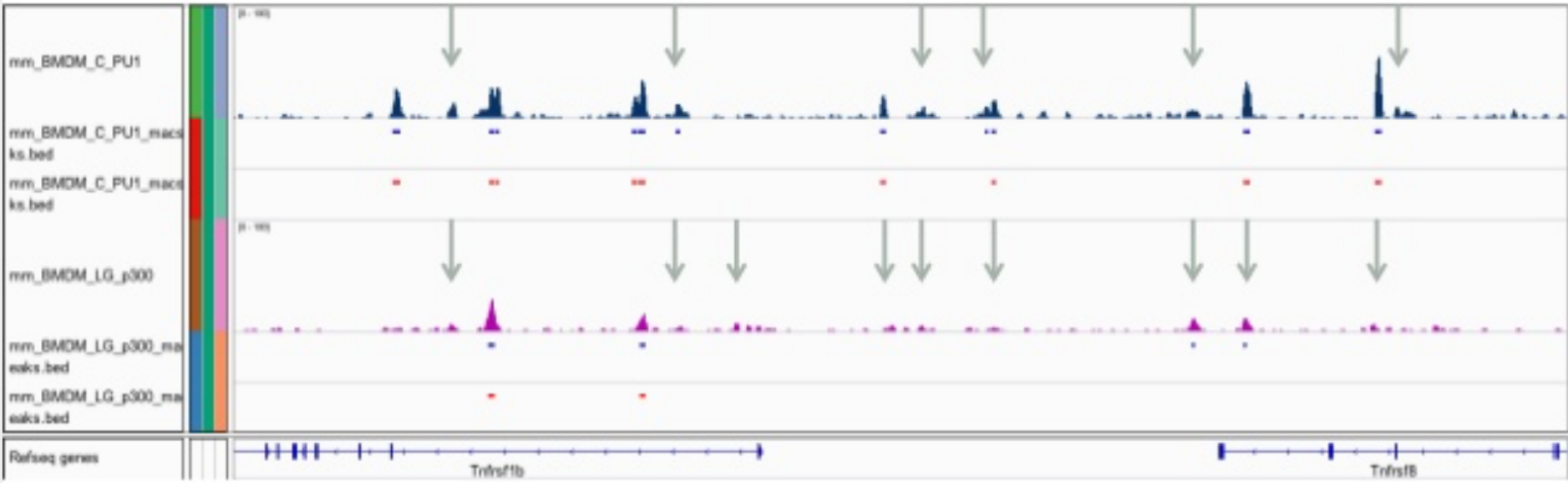


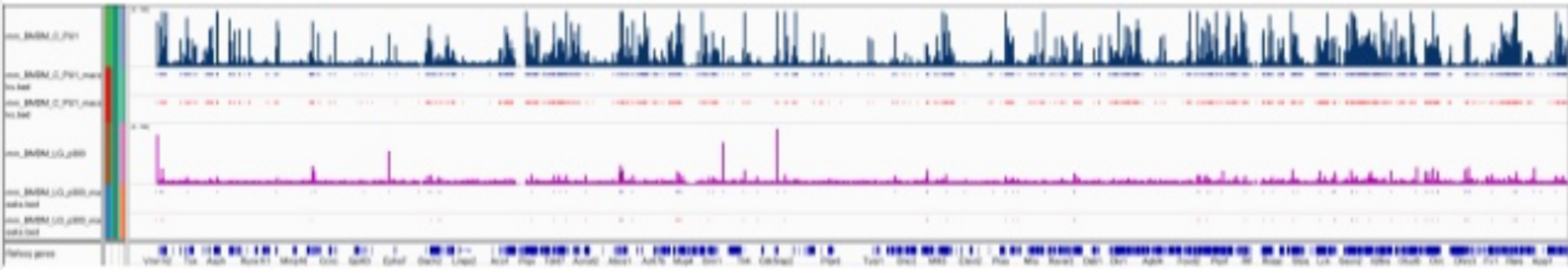
Here we go! Good job!



- When comparing multiple tracks, set same scale
- Remember to match genome assembly (e.g. mm9) between browser and mapping
  - ➢ you will not get any error when you forget!

How to define biologically meaningful peaks?





Peak finding using HOMER (findPeaks)

```
# HOMER Peaks
# Peak finding parameters:
# tag directory = Sox2-ChIP-Seq
#
# total peaks = 10280
# peak size = 137
# peaks found using tags on both strands
# minimum distance between peaks = 342
# fragment length = 132
# genome size = 4000000000
# Total tags = 9806245.0
# Total tags in peaks = 156620.0
# Approximate IP efficiency = 1.56%
# tags per bp = 0.001907
# expected tags per peak = 0.523
# maximum tags considered per bp = 1.0
# effective number of tags used for normalization = 10000000.0
# Peaks have been centered at maximum tag pile-up
# FDR rate threshold = 0.001000
# FDR effective poisson threshold = 0.000000
# FDR tag threshold = 8.0
# number of putative peaks = 10800
#
# size of region used for local filtering = 10000
# Fold over local region required = 4.00
# Poisson p-value over local region required = 1.00e-04
# Putative peaks filtered by local signal = 464
#
# Maximum fold under expected unique positions for tags = 2.00
# Putative peaks filtered for being too clonal = 36
#
# cmd = findPeaks Sox2-ChIP-Seq -style factor -o auto
#
# Column Headers:
```

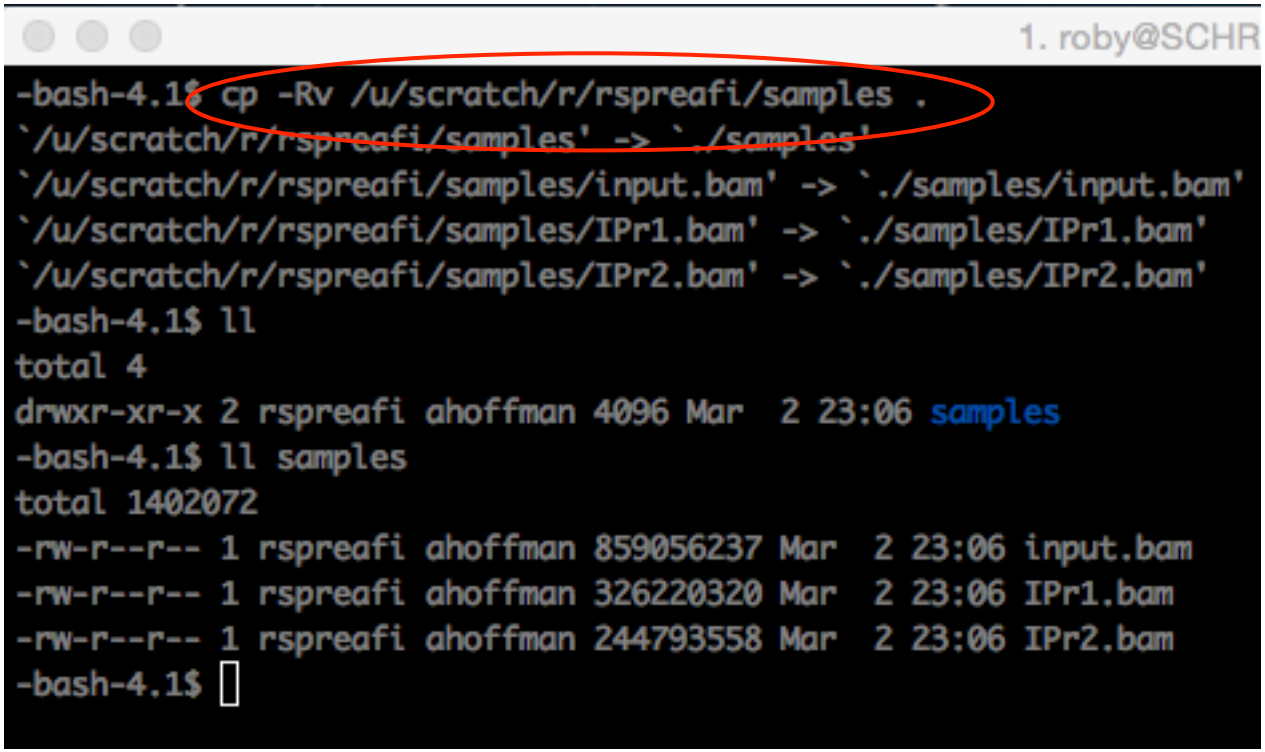
- Column 1: PeakID - a unique name for each peak (very important that peaks have unique names...)
- Column 2: chr - chromosome where peak is located
- Column 3: starting position of peak
- Column 4: ending position of peak
- Column 5: Strand (+/-)
- Column 6: Normalized Tag Counts - number of tags found at the peak, normalized to 10 million total mapped tags (or defined by the user)
- Column 7: (-style factor): Focus Ratio - fraction of tags found appropriately upstream and downstream of the peak center. (see below)
- (-style histone/-style groseq): Region Size - length of enriched region
- Columns 8+: Statistics and Data from filtering

Genome size represents the total effective number of mappable bases in the genome (remember each base could be mapped in each direction)

Approximate IP effeciency describes the fraction of tags found in peaks versus genomic background. This provides an estimate of how well the ChIP worked. Certain antibodies like H3K4me3, ERa, or PU.1 will yield very high IP efficiencies (>20%), while most rand in the 1-20% range. Once this number dips below 1% it's a good sign the ChIP didn't work very well and should probably be optimized.

Copy data into your scratch folder

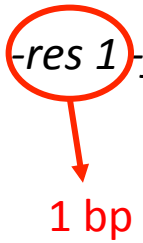
Work from your scratch folder, e.g. /u/scratch/r/rspreafi



Let's practice

Work from your homer folder, e.g. /u/scratch/r/rspreafi/samples/homer

makeUCSCfile IPr1 -o IPr1.bedGraph -fragLength 200 -norm 1000000 -res 1 -fsize 1e20 |& tee IPr1_bedGraph.txt

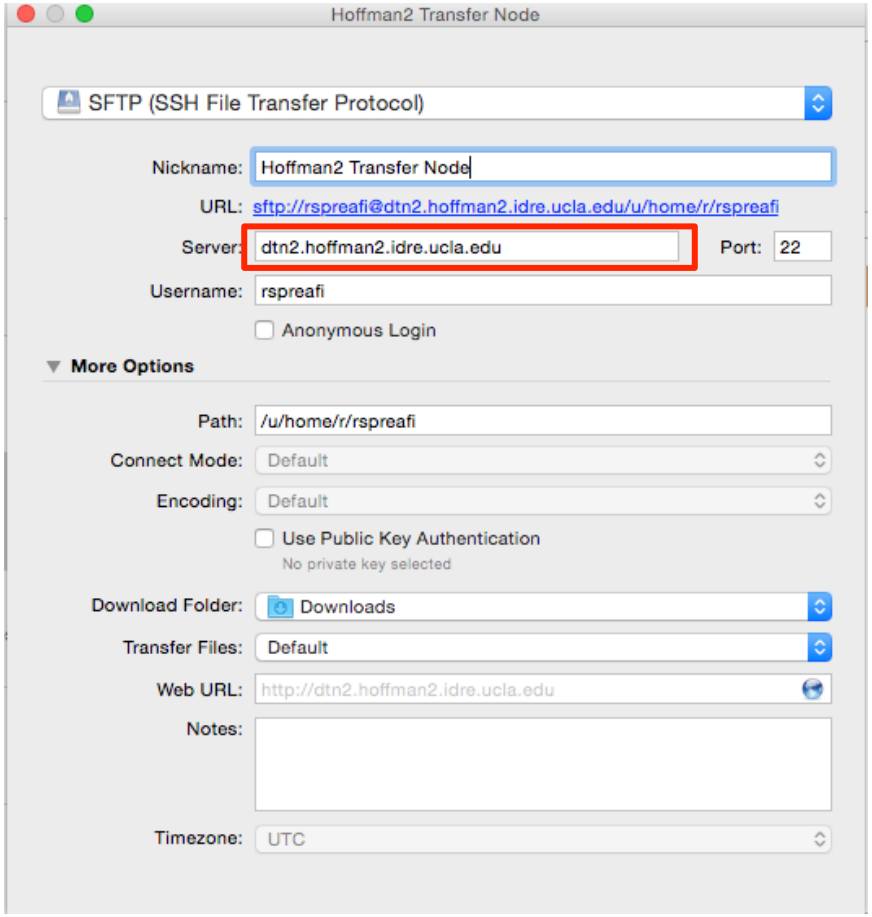


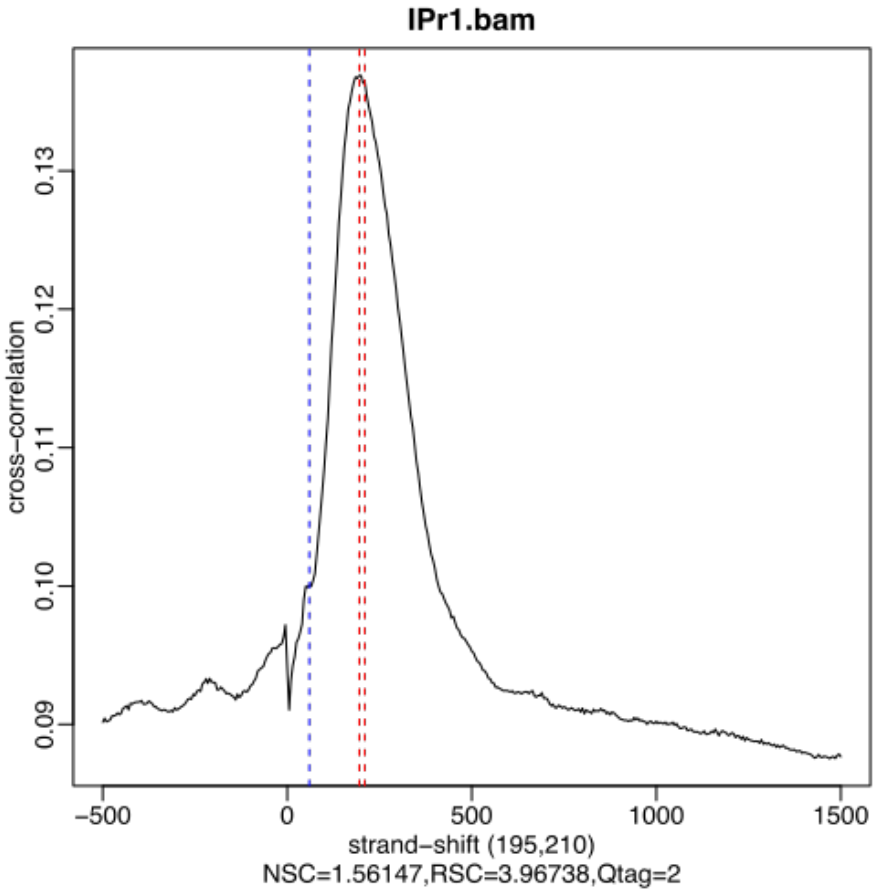
gunzip IPr1.bedGraph.gz

head IPr1.bedGraph

You can repeat this for all your samples… but not needed now.

Retrieve your analysis



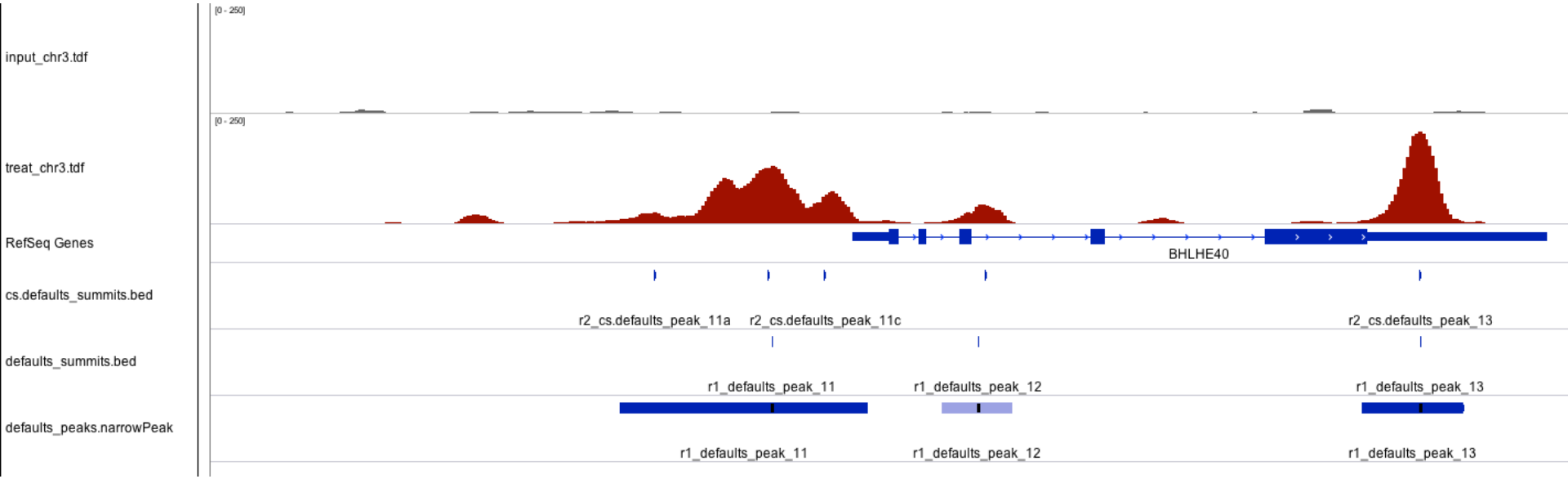


Schedule

- **Day 1**
  - Introduction
  - Cross-correlation analysis and ENCODE QC with SPP
  - BigWig tracks with defined resolution and normalization using Homer/ UCSC tools, and visualization with IGV
- **Day 2**
  - Peak calling with MACS2
  - QC of replicates with ENCODE's IDR
  - Differential peak calling with MAnorm
- **Day 3**
  - Location annotation with NGSPLOT
  - Motif analysis with HOMER
  - Functional annotation with GREAT
  - (Unix tricks, tools installation)

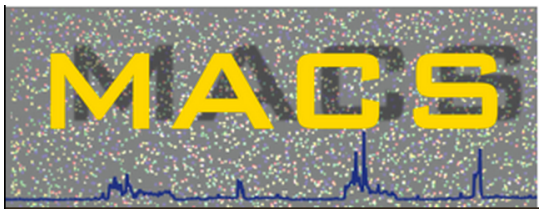
PEAK CALLING

Goal



- Peaks: chr, start, end, score/p-value
- Summits: chr, position

Many tools for the job



github.com/taoliu/MACS/

HOMER

rthurman / **hotspot**

SPP

PeakSeq

Let's practice

Work from your samples folder, e.g. /u/scratch/r/rspreafi/samples

mkdir macs2

module load python/2.7.3

macs2 callpeak -t IPr1.bam -c input.bam -f BAM -n IPr1 -g mm -q 0.01 --keep-dup 1 --call-summits --nomodel --extsize 200 --outdir macs2 >macs2/IPr1_output.txt 2>&1 &

macs2 callpeak -t IPr2.bam -c input.bam -f BAM -n IPr2 -g mm -q 0.01 --keep-dup 1 --call-summits --nomodel --extsize 200 --outdir macs2 >macs2/IPr2_output.txt 2>&1 &

Mouse: mm
Human: hs
Or size of genome in bp

q-value threshold
alternatively, *-p 0.01*

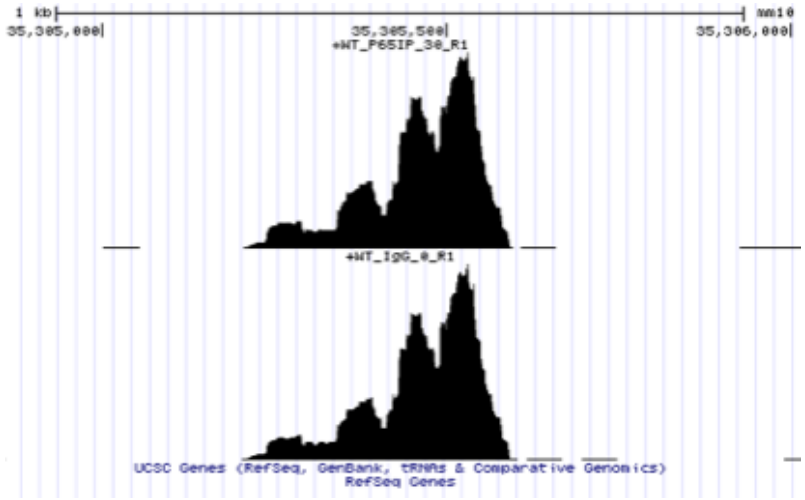
n
auto
all

Call a peak in your head

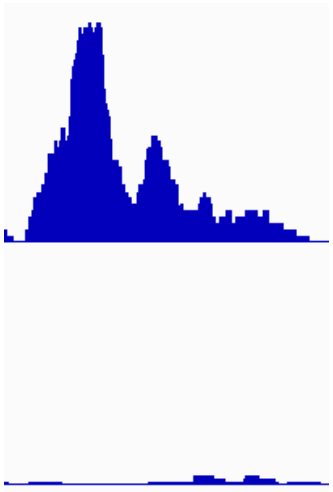
IP

Input

A



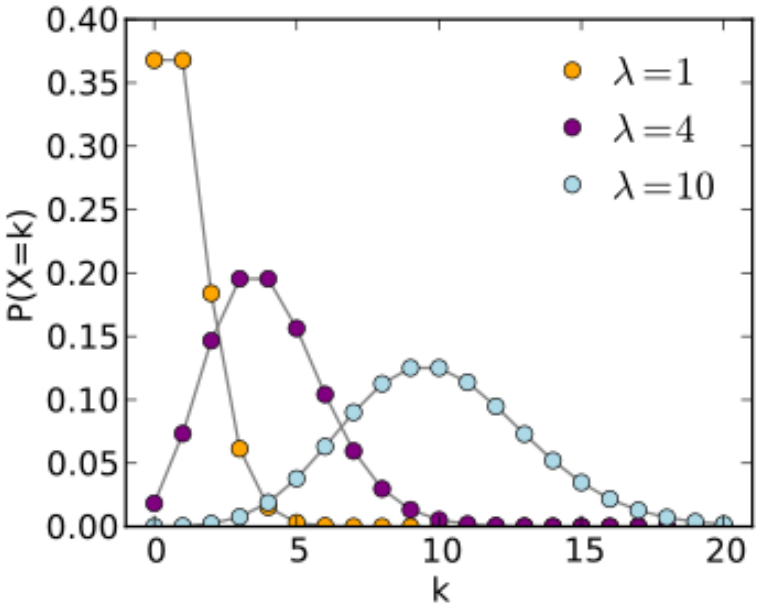
B



Now tell me how you made your decision

How MACS works

- MACS compares IP vs control
- Comparisons are formalized using the Poisson distribution
- MACS computes both global and local Poisson distributions



This is why good ChIP practices recommend sequencing input more deeply than IP samples

MACS Q&A

- Do you support mixed-length reads?
  - Github manual

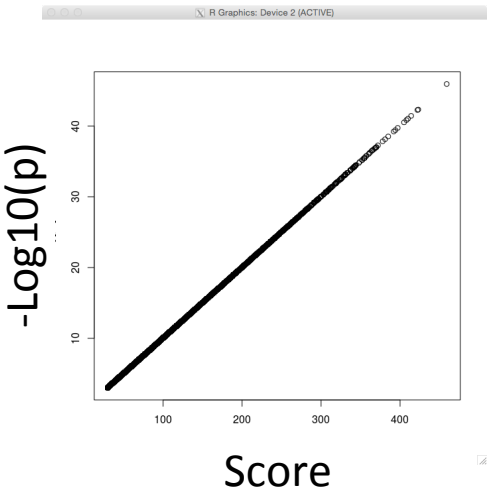
    -s/--tsize The size of sequencing tags. If you don't specify it, MACS will try to use the first 10 sequences from your input treatment file to determine the tag size.

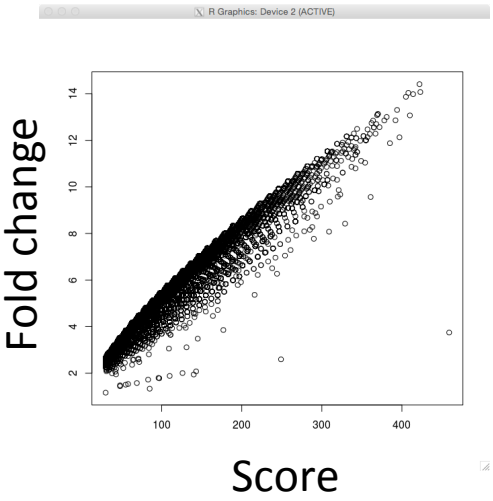
  - Google Groups forum: does not make much difference
- Do you support calling broad peaks (e.g. certain histone marks)?
  - Yes, but only MACS2 (not MACS 1.4)
  - Options: *--broad --broad-cutoff 0.1* (not compatible with *--call-summits*)
  - Other peak callers are specifically designed for broad marks: SICER, RSEG, ZINBA
- Where can I find documentation?
  - The original paper (Genome Biology 2008 9:R137) deals with MACS v1
  - MACS2 is sparsely documented on Github and Google Groups, yet referred to in several publication.
  - Improvements in cross-correlation, peaks called in log space, broad peak detection, removed several biases

narrowPeak output

BED format: 0-based, half closed [start-1, end)

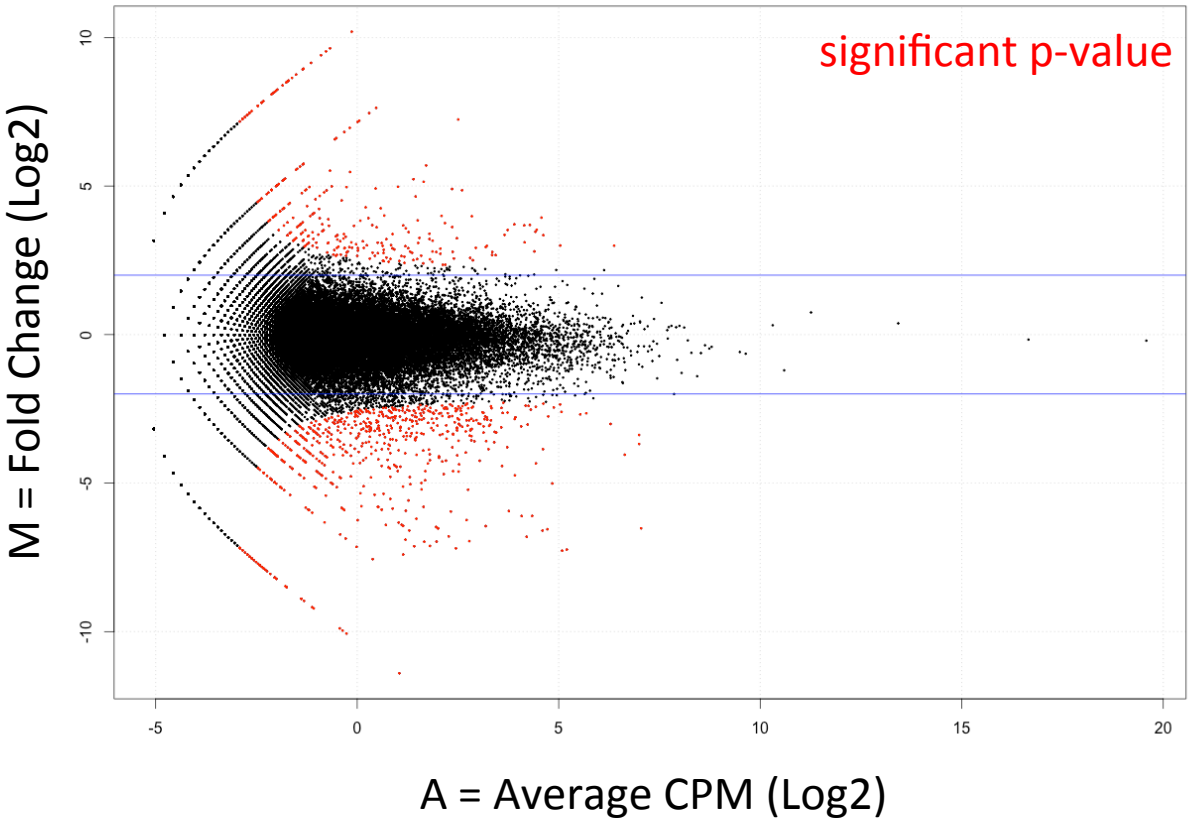
1. Chr
2. Start
3. End
4. ID
5. Score (1-1000)
6. Strand
7. Fold change
8. -Log10(p)
9. -Log10(q)
10. Summit position, offset from start





Don’t forget accuracy!

Fold change vs p-value: example from RNAseq



EVALUATE REPLICATES

(Not) many tools for the job

Irreproducibility Discovery Rate (IDR) from ENCODE

sites.google.com/site/anshulkundaje/projects/idr

New version in progress!
github.com/nboley/idr

Let's practice

Work from your samples folder, e.g. /u/scratch/r/rspreafi/samples

mkdir idr

module load python/2.7.3

macs2 callpeak -t IPr1.bam -c input.bam -f BAM -n IPr1 -g mm -p 1e-3 --keep-dup 1 --call-summits --nomodel --extsize 200 --outdir idr >idr/IPr1_macs.txt 2>&1 &

macs2 callpeak -t IPr2.bam -c input.bam -f BAM -n IPr2 -g mm -p 1e-3 --keep-dup 1 --call-summits --nomodel --extsize 200 --outdir idr >idr/IPr2_macs.txt 2>&1 &

Notice the relaxed threshold: we do want false positives!

The unsophisticated approach

- Call peaks in R1 and R2. Retain those that are significant in both.
  - Variant: call peaks in R1 with FDR correction and retain those also significant in R2 by unadjusted p-value, and viceversa.
- Can we do better?
- Follow you intuition. These peaks come from the same two replicates. Which do you trust more?
  a. R1: $q=10^{-4}$; R2: $q=0.23$
  b. R1: $q=10^{-5}$; R2: $q=10^{-4}$
  c. R1: $q=10^{-9}$; R2: $q=10^{-2}$

Let's practice

Work from your idr folder, e.g. /u/scratch/r/rspreafi/samples/idr

module load R/3.1.1

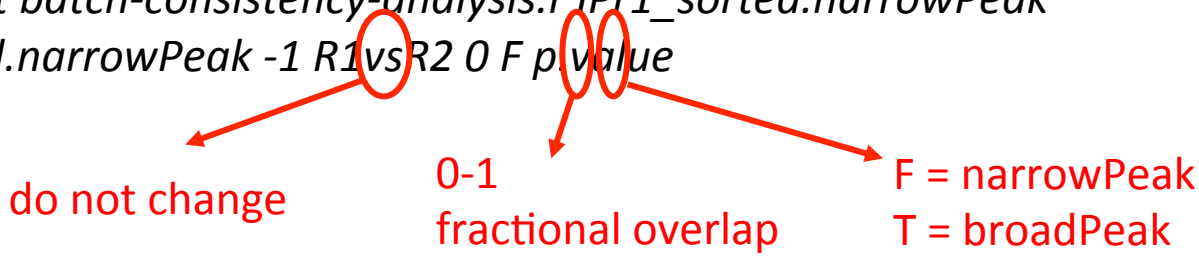
cp -Rv /u/local/apps/idr/2010-10/ .*

cp -v genome_tables/genome_table.mm9.txt genome_table.txt # say yes to overwriting here!

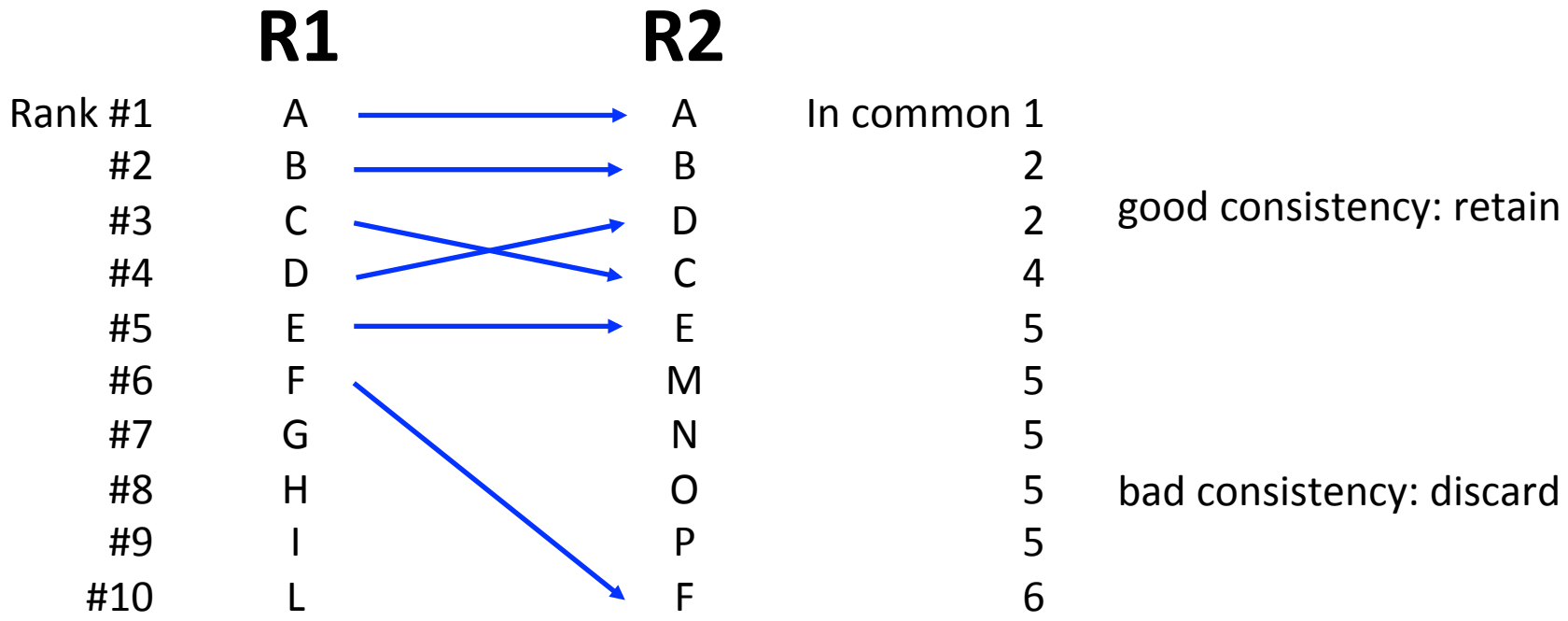
sort -k8,8nr IPr1_peaks.narrowPeak | head -n 100000 >IPr1_sorted.narrowPeak

sort -k8,8nr IPr2_peaks.narrowPeak | head -n 100000 >IPr2_sorted.narrowPeak

Rscript batch-consistency-analysis.r IPr1_sorted.narrowPeak IPr2_sorted.narrowPeak -1 R1vsR2 0 F p.value



How IDR works: rankings

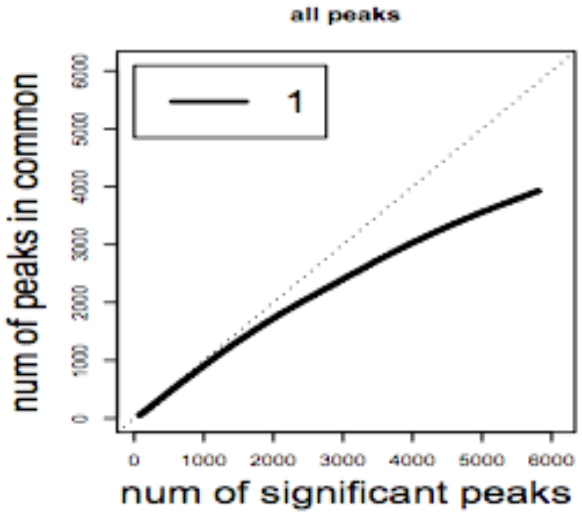


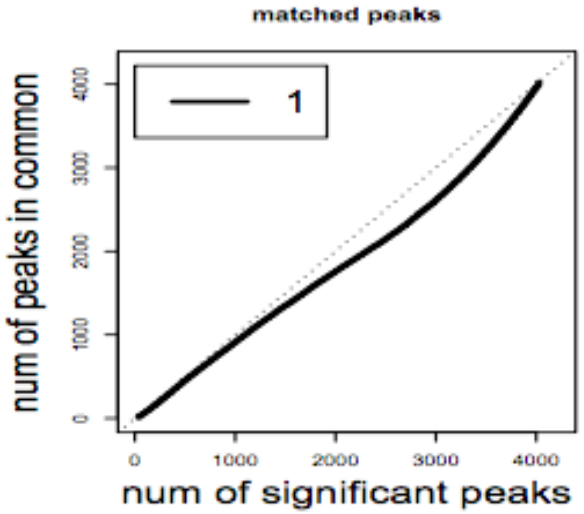
Alternatives?

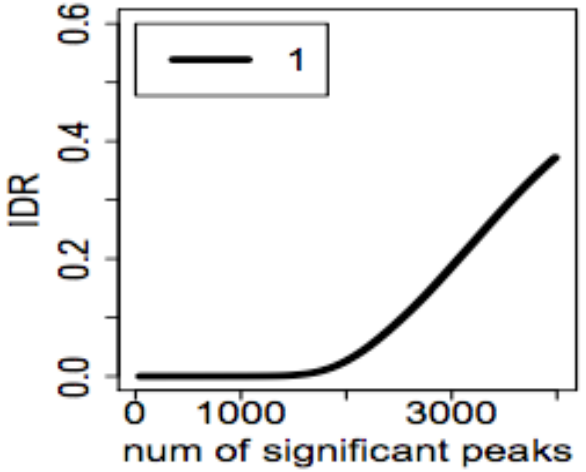
- digital yes/no ("unsophisticated approach"): retain if significant in both replicates. Peak "F" would be retained with this approach.
- absolute p-value: depends on scale, which in turn depends on many factors (such as background, stay tuned).

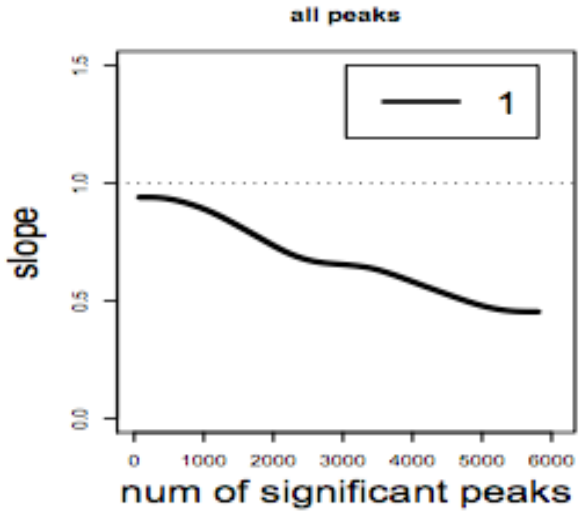
Rankings put p-values on a relative scale. And on relative scales you can use anything: p-value, q-values, fold change… and from different algorithms too!

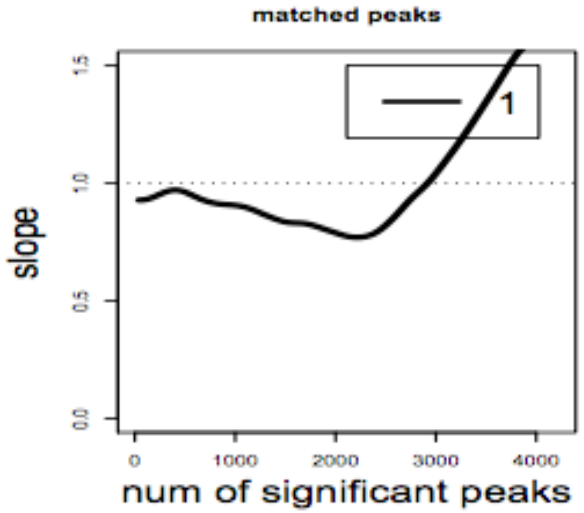
Good replicates



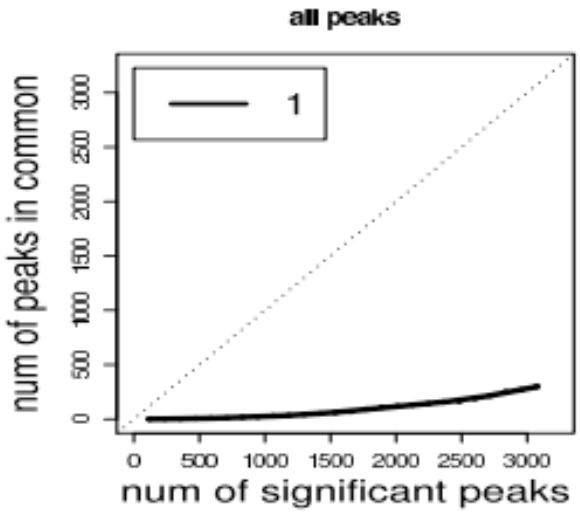


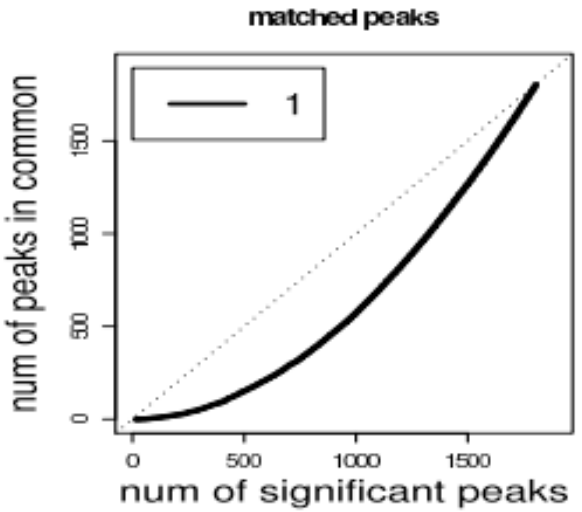


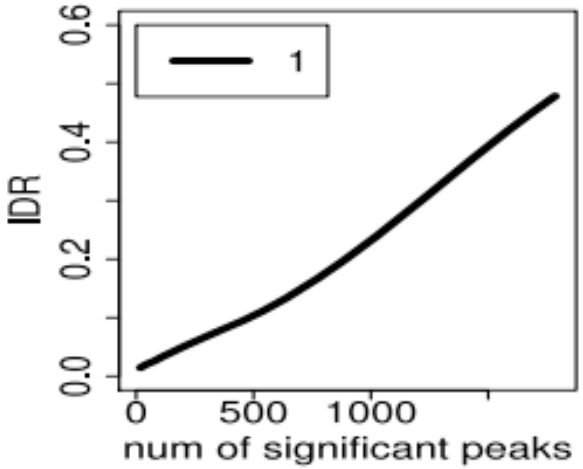


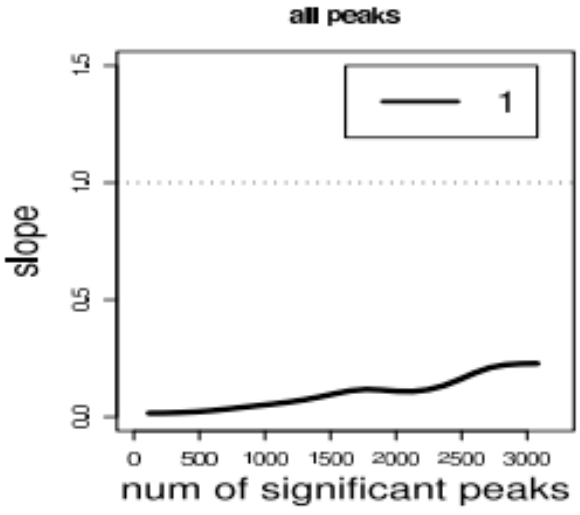


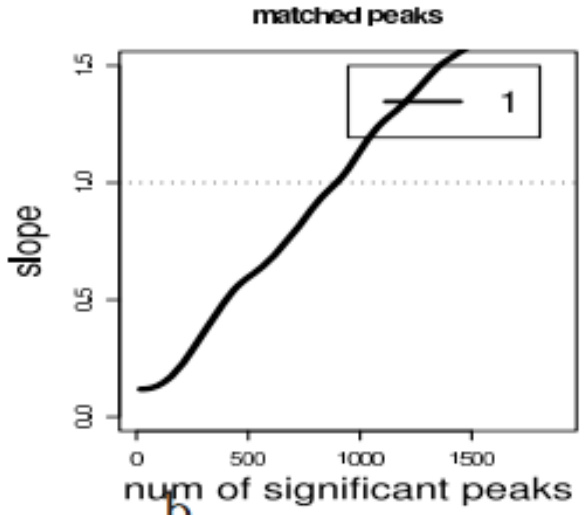
Bad replicates











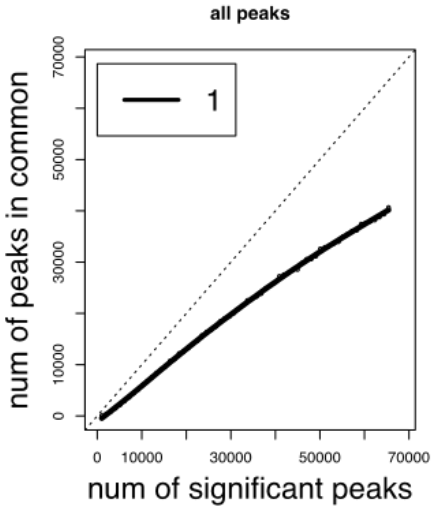
Let's practice

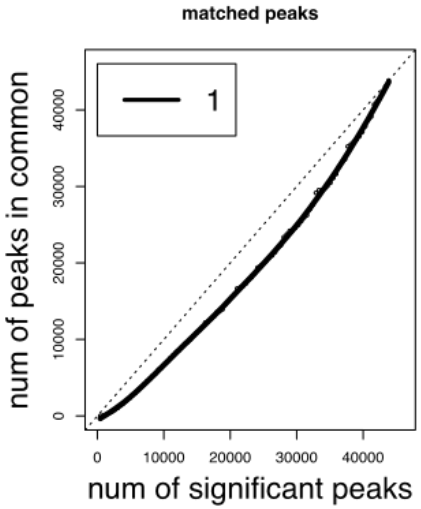
Work from your idr folder, e.g. /u/scratch/r/rspreafi/samples/idr

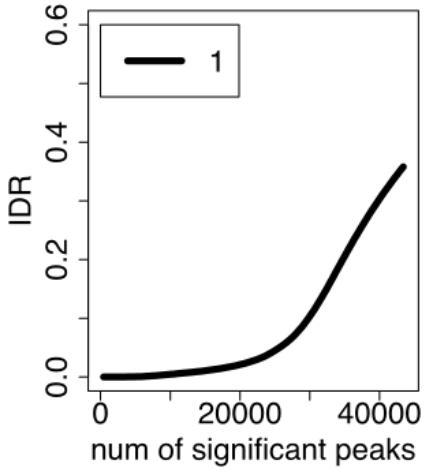
Rscript batch-consistency-plot.r 1 R1vsR2 R1vsR2

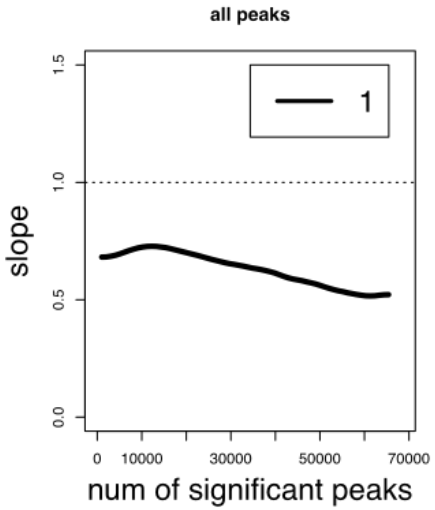
head R1vsR2-overlapped-peaks.txt

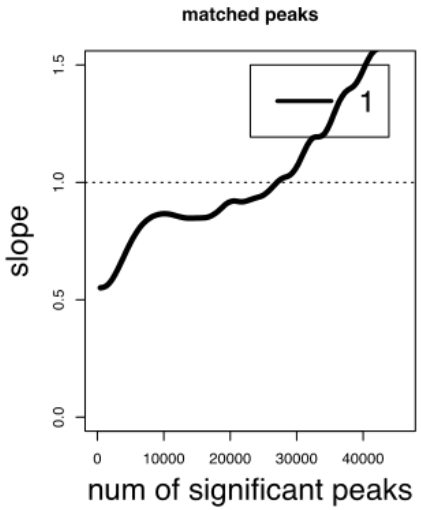
Retrieve your data

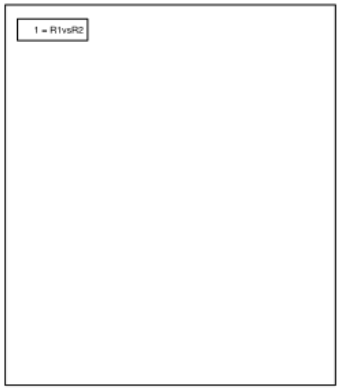




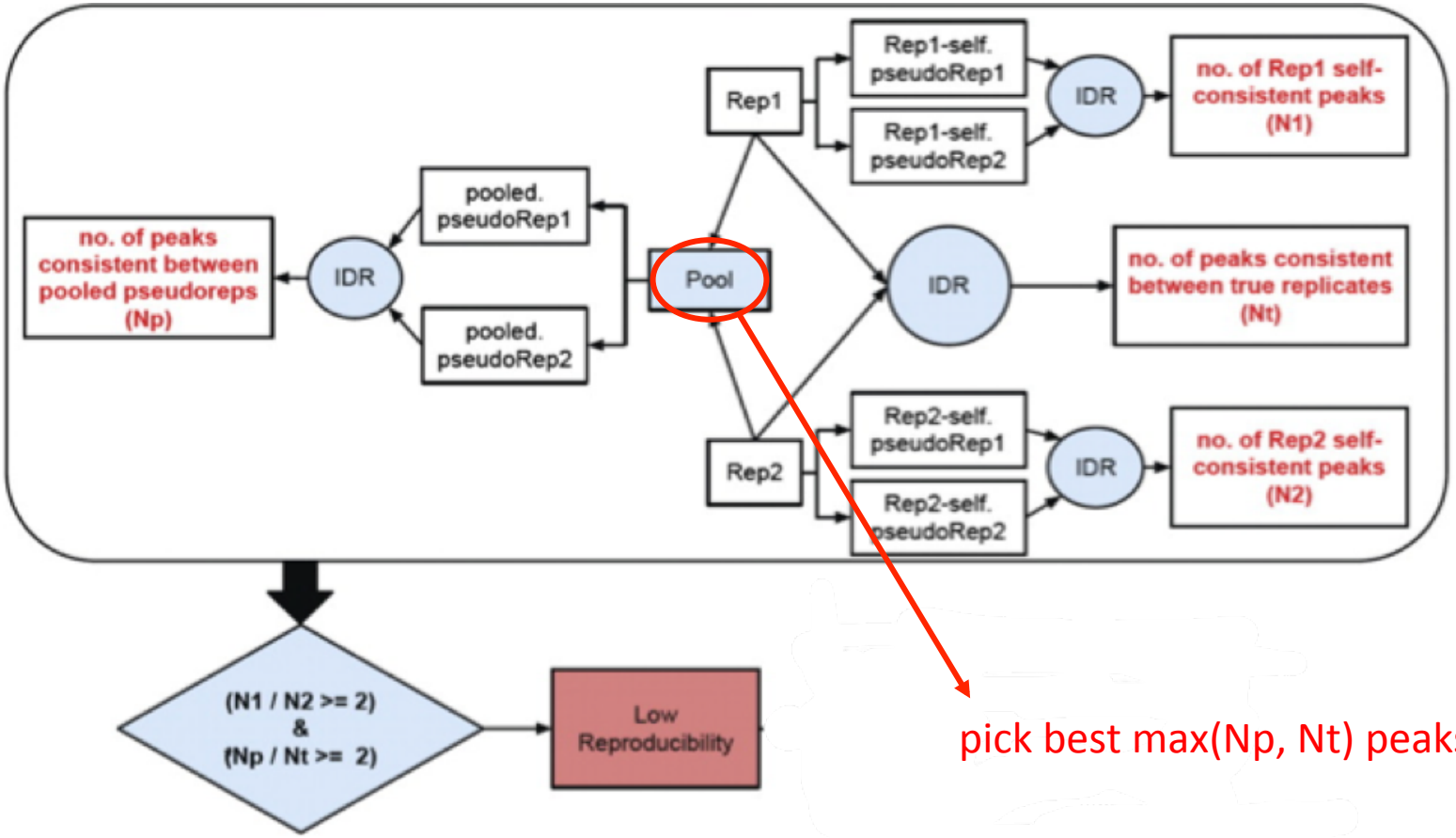








The complete IDR pipeline



sites.google.com/site/anshulkundaje/projects/idr

DIFFERENTIAL PEAK CALLING

Let's practice

➢ Let's pretend that the two replicates are different treatment conditions and let's call differential peaks

Work from your sample folder, e.g. /u/scratch/r/rspreafi/samples/

module load bedtools/2.23.0

awk 'BEGIN{OFS="\t"}{print $1,$2,$3}' macs2/IPr1_peaks.narrowPeak >IPr1_peaks.bed &

awk 'BEGIN{OFS="\t"}{print $1,$2,$3}' macs2/IPr2_peaks.narrowPeak >IPr2_peaks.bed &

bamToBed -i IPr1.bam | awk 'BEGIN{OFS="\t"}{print $1,$2,$3,$6}' >IPr1.bed &

bamToBed -i IPr2.bam | awk 'BEGIN{OFS="\t"}{print $1,$2,$3,$6}' >IPr2.bed &

Let's practice

➢ Let's pretend that the two replicates are different treatment conditions and let's call differential peaks

Work from your sample folder, e.g. /u/scratch/r/rspreafi/samples/

mkdir manorm

cd manorm

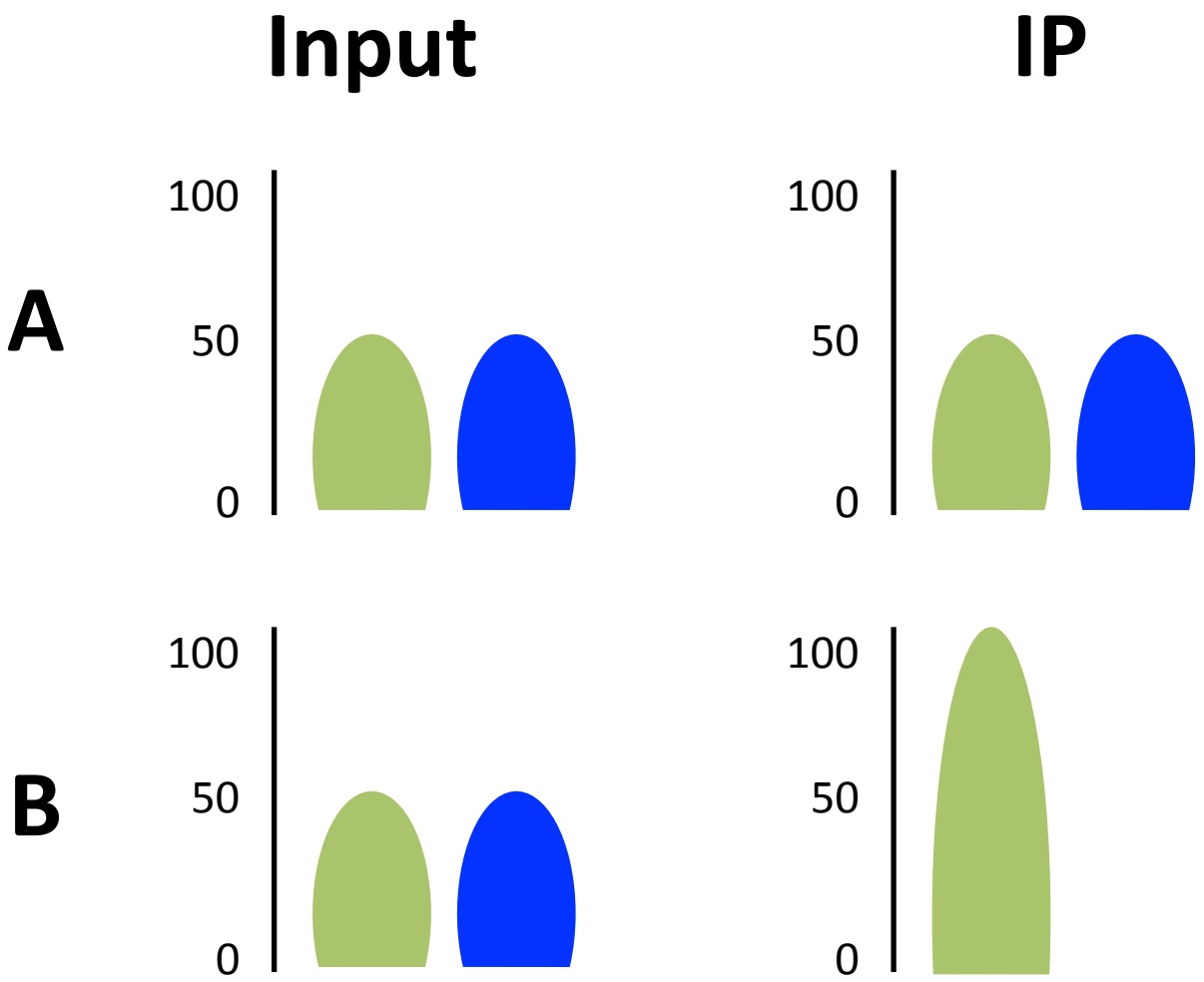
cp /u/local/apps/manorm/2012/MAnorm_Linux_R_Package/MAnorm. .*

chmod 755 MAnorm.sh

./MAnorm.sh ../IPr1_peaks.bed ../IPr2_peaks.bed ../IPr1.bed ../IPr2.bed 200 200 >output.txt 2>&1 &

Break

~~qPCR~~ seq by equal mass



DNA of interest (PCR target)

Protein of interest (IP target)

DNA not of interest

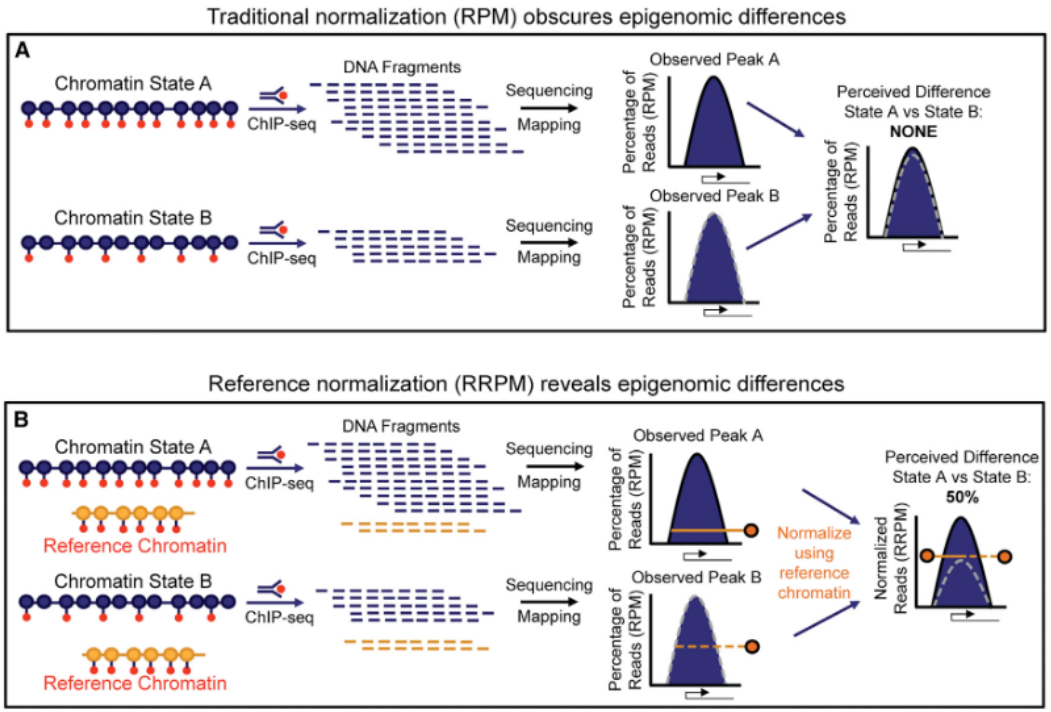
*"DNA of interest is enriched by **2** fold in B compared to A"*

A spike in approach for normalization

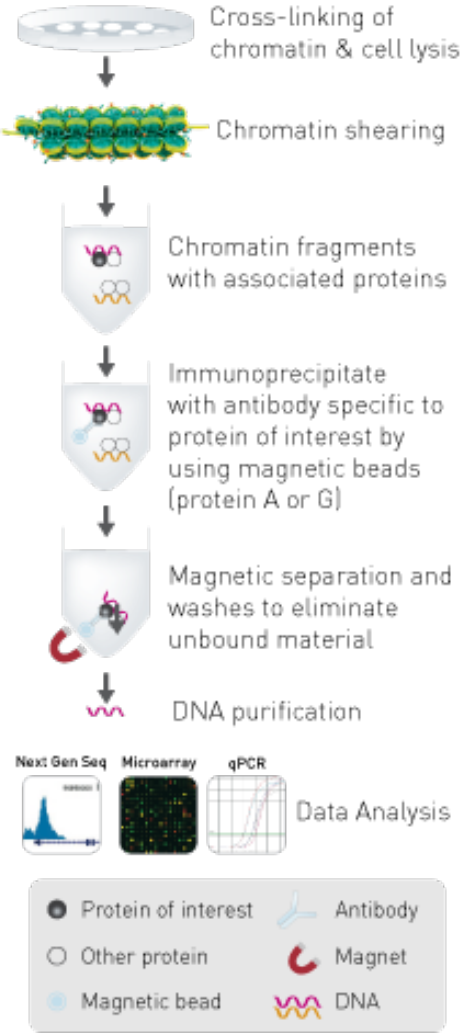
Cell Reports
Resource

Nov 6, 2014

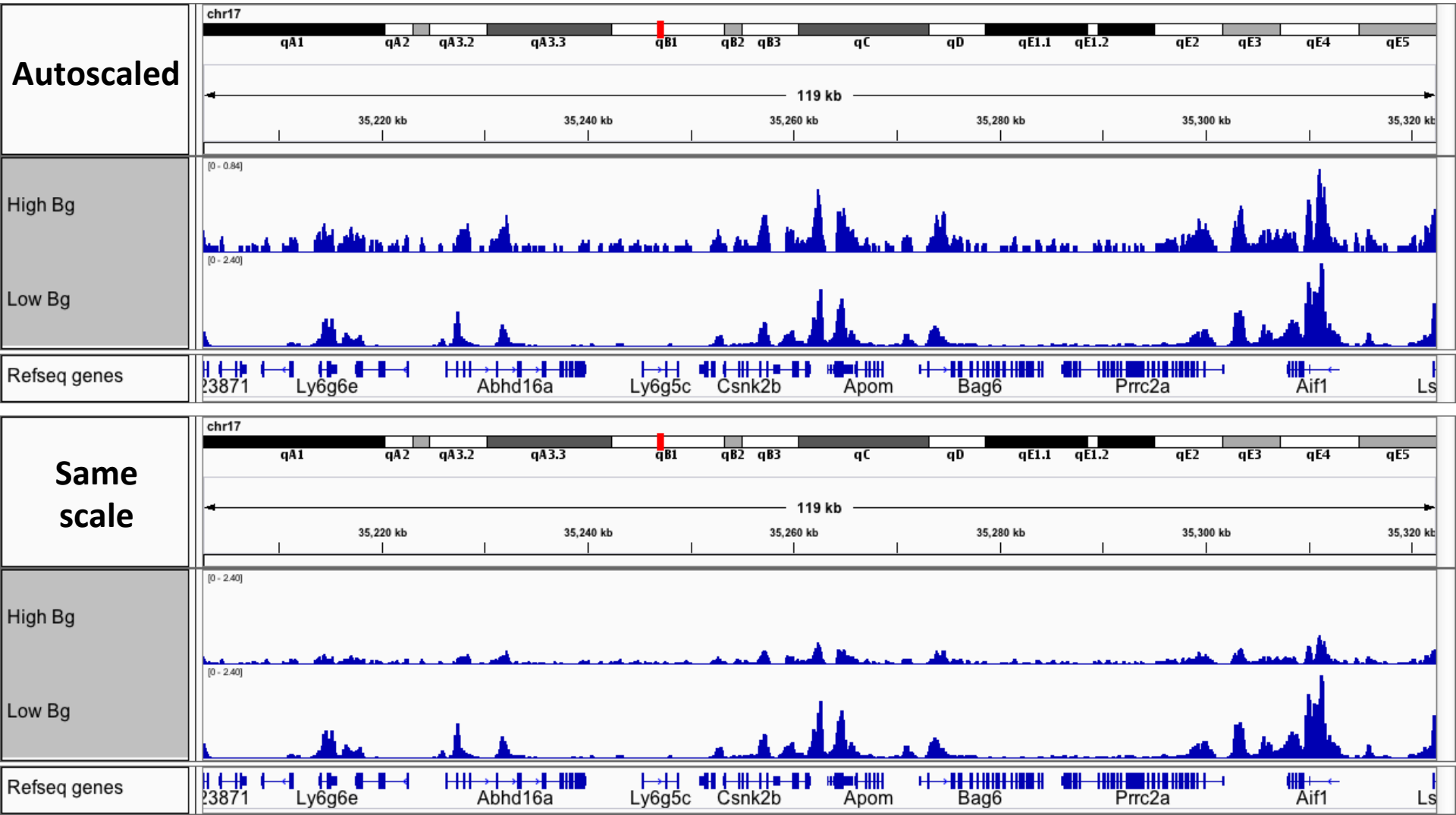
Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome



see also:
www.activemotif.com/catalog/1063/chip-seq-spike-in



An even worse problem: differences in background

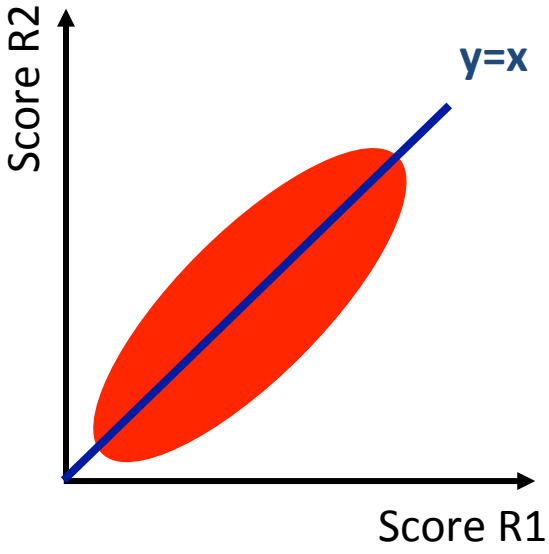


The importance of S/N ratio

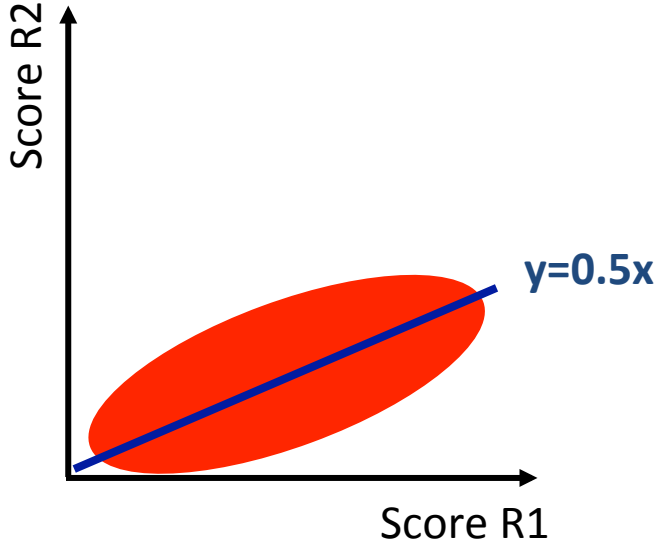
e.g. IP efficiency

e.g. washing steps

Same S/N



Different S/N

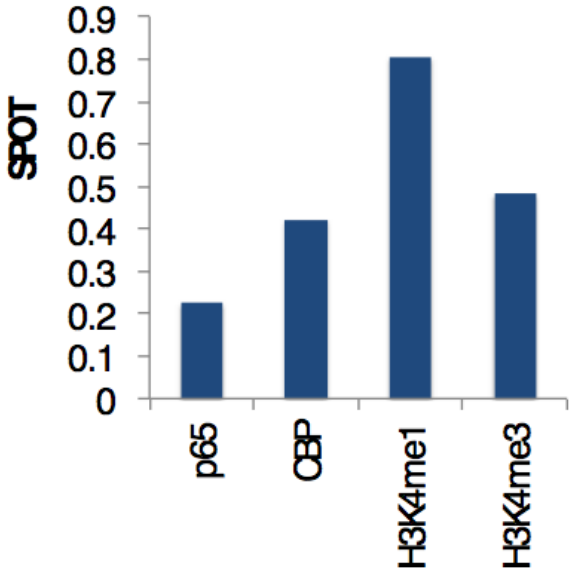


Estimating S/N ratio

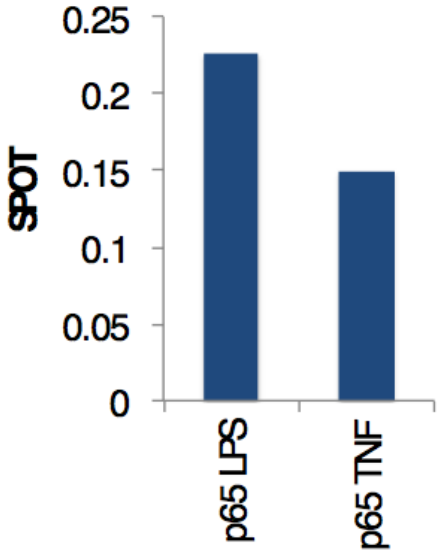
FRiP (Fraction of Reads in Peaks) or

SPOT (from HOTSPOT peak caller)

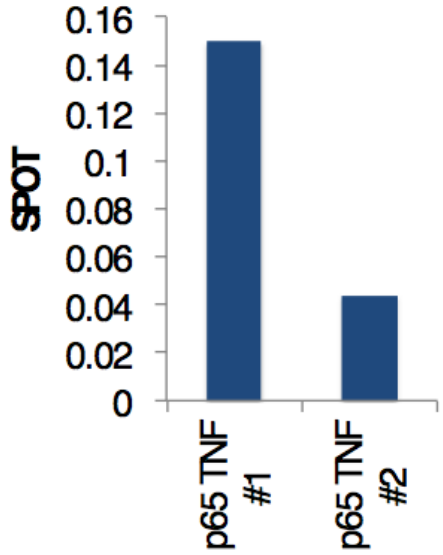
Depends on
- Ab
- background



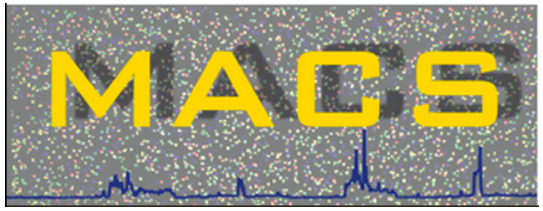
Depends on
- condition
- background



Depends on background only



Many tools for the job



MAnorm

bcb.dfci.harvard.edu/~gcyuan/MAnorm

edgeR

DESeq

MMDiff
(shapes)

DiffBind

(RNA pol II)

POLYPHEMUS

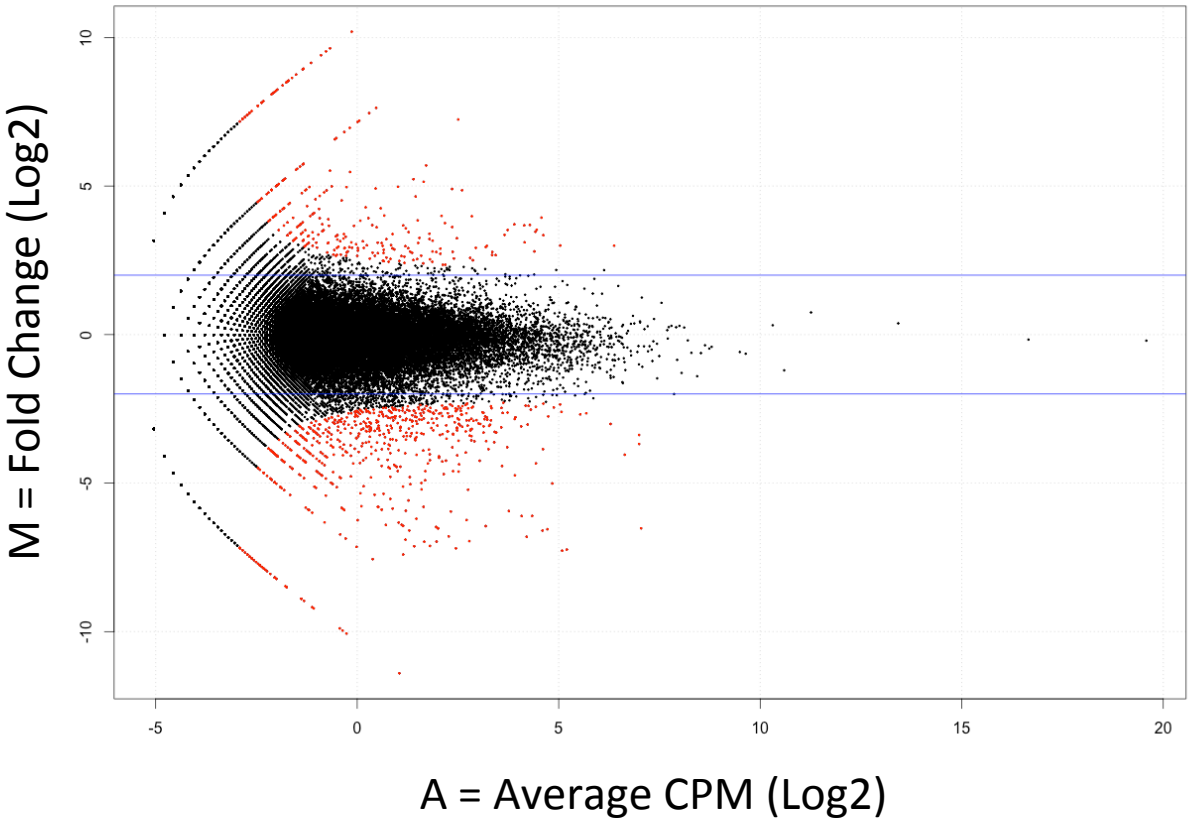
PePr

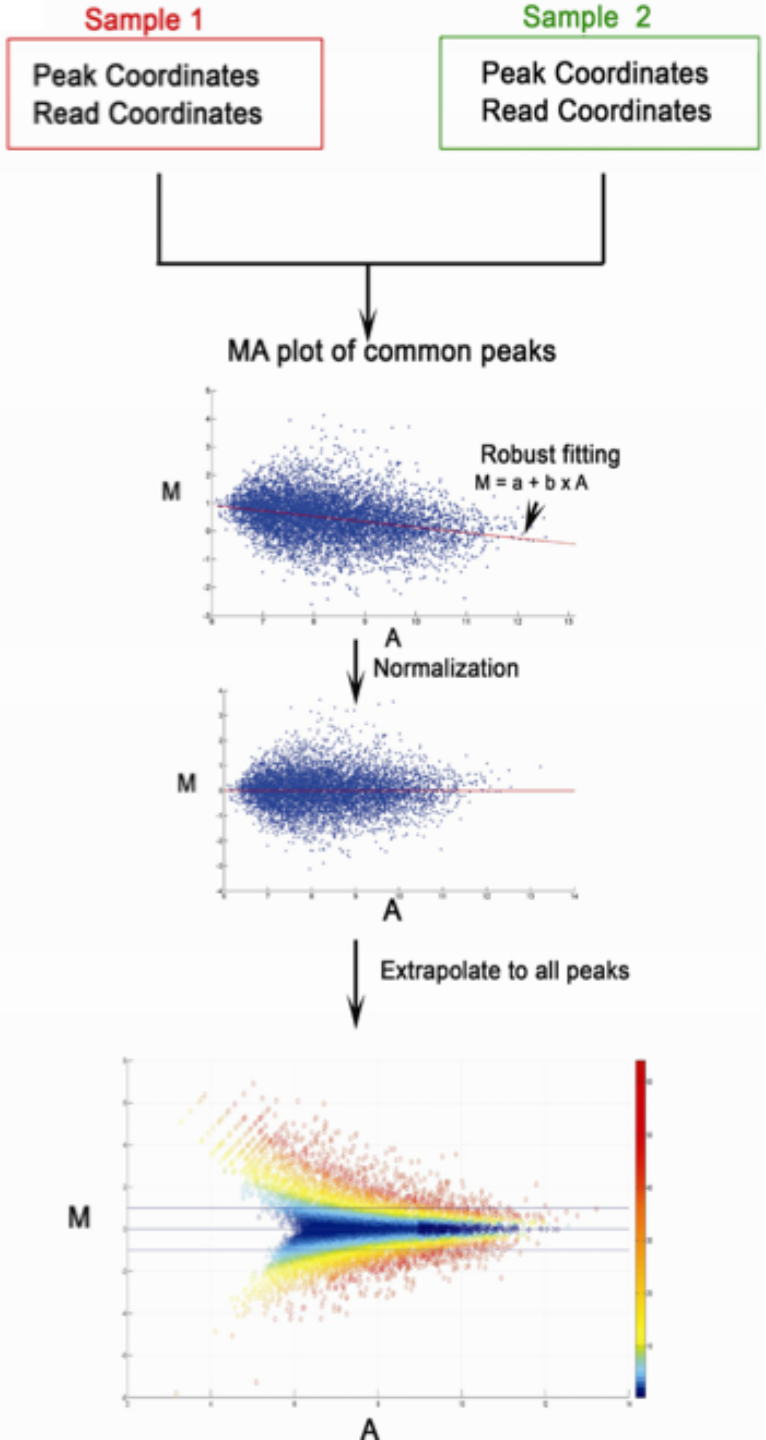
csaw

Two ways

- Compare peaks (e.g. edgeR, MAnorm)
- Discard peaks, start fresh and compare signal window by window (e.g. PePr, csaw)

Remember your friend, the MA plot





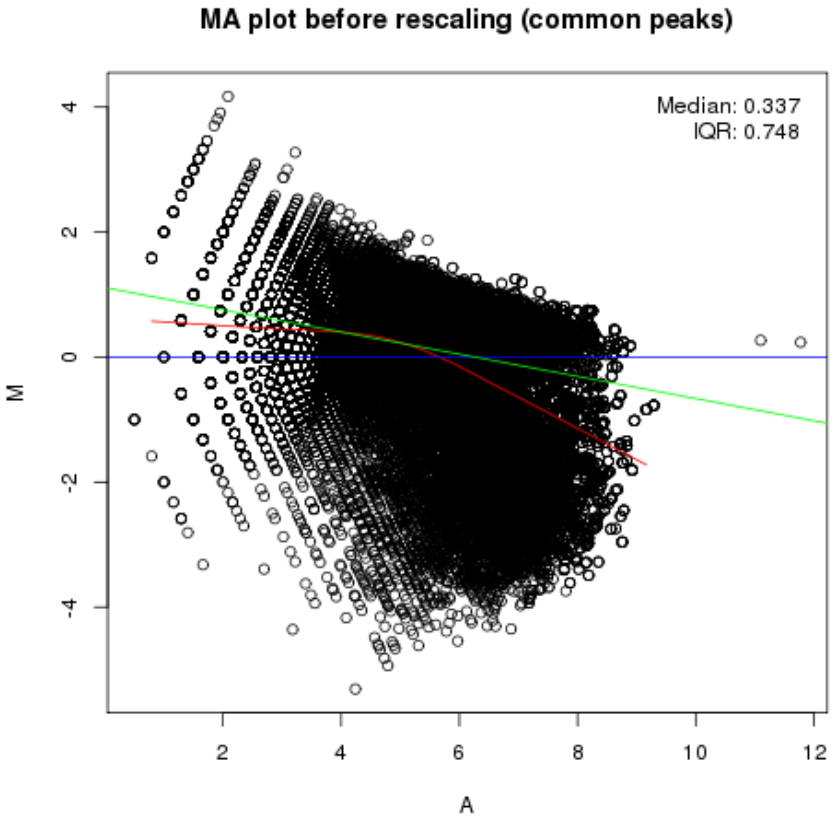
How MAnorm works

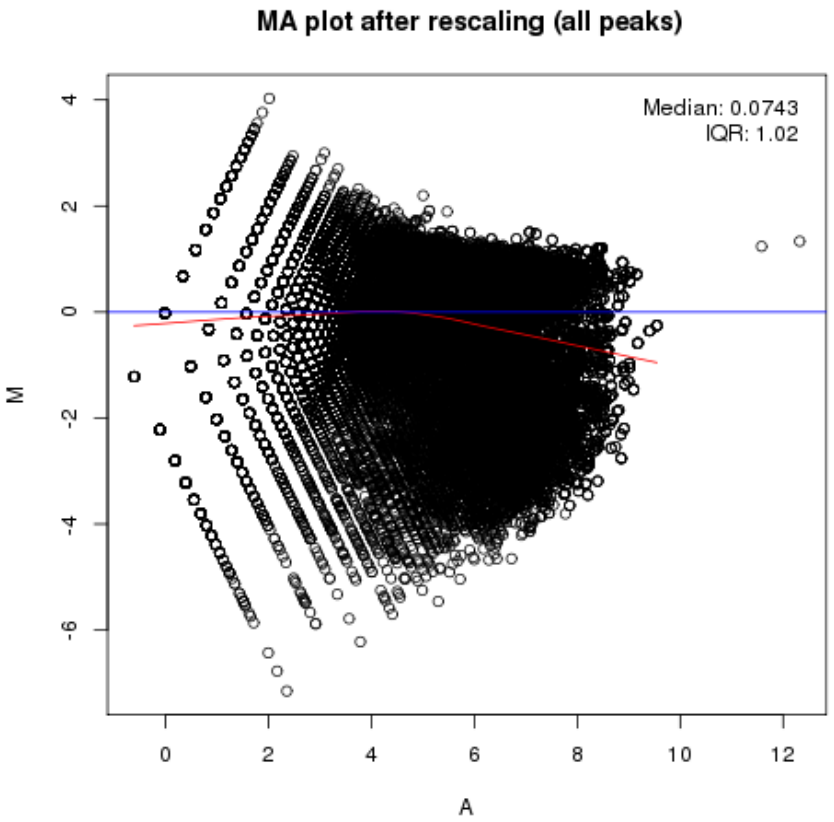
Assumes that bias has linear trend. Note that scaling may not be enough

Assumes that most common peaks have equal intensity

Assumes that the trended bias fitting common peaks applies to unique peaks too

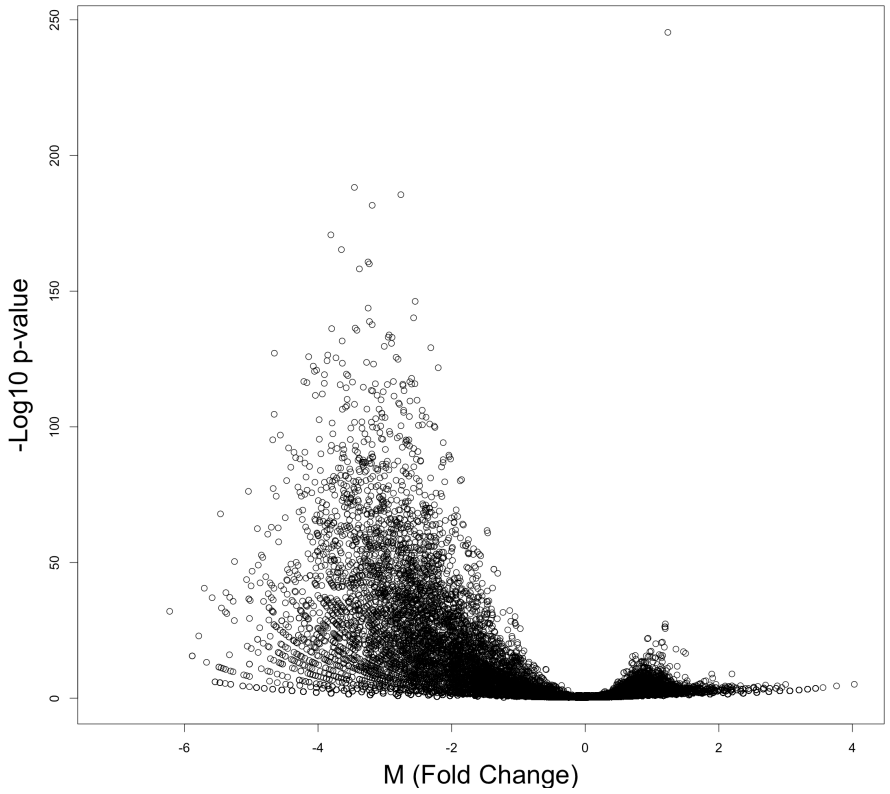
Retrieve your data

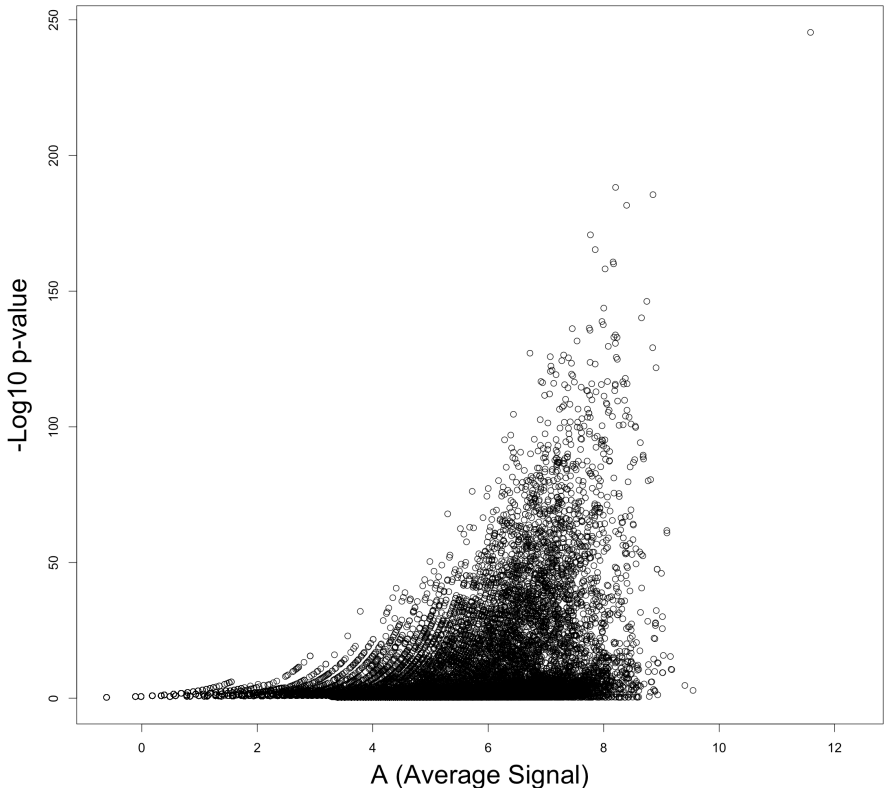


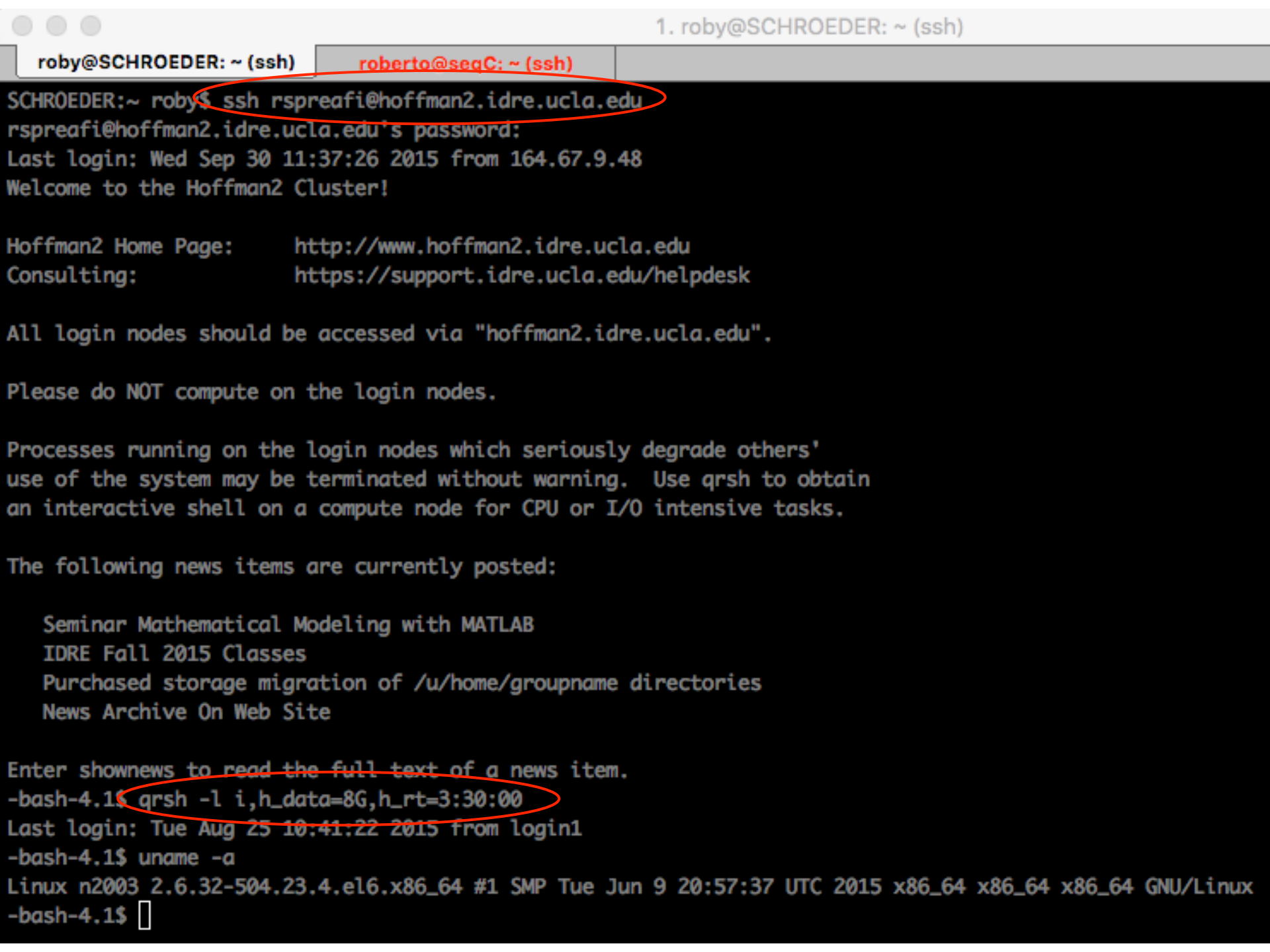


Retrieve your data

chr	start	end	description	#raw_read_1	#raw_read_2	M_rescaled	A_rescaled	-LOG10(p)
chr1	4773028	4773229	unique_peak	4	0	1.55447428	0.77723714	1.20411998
chr1	4776611	4776821	unique_peak	6	0	2.13444762	1.06722381	1.50514998
chr1	4777893	4778093	unique_peak	6	3	0.13444762	2.06722381	0.86417192
chr1	4796483	4796698	unique_peak	9	2	1.16428163	2.16710332	1.45399746
chr1	4797639	4797839	unique_peak	3	2	-0.4151182	1.3774034	0.72699873





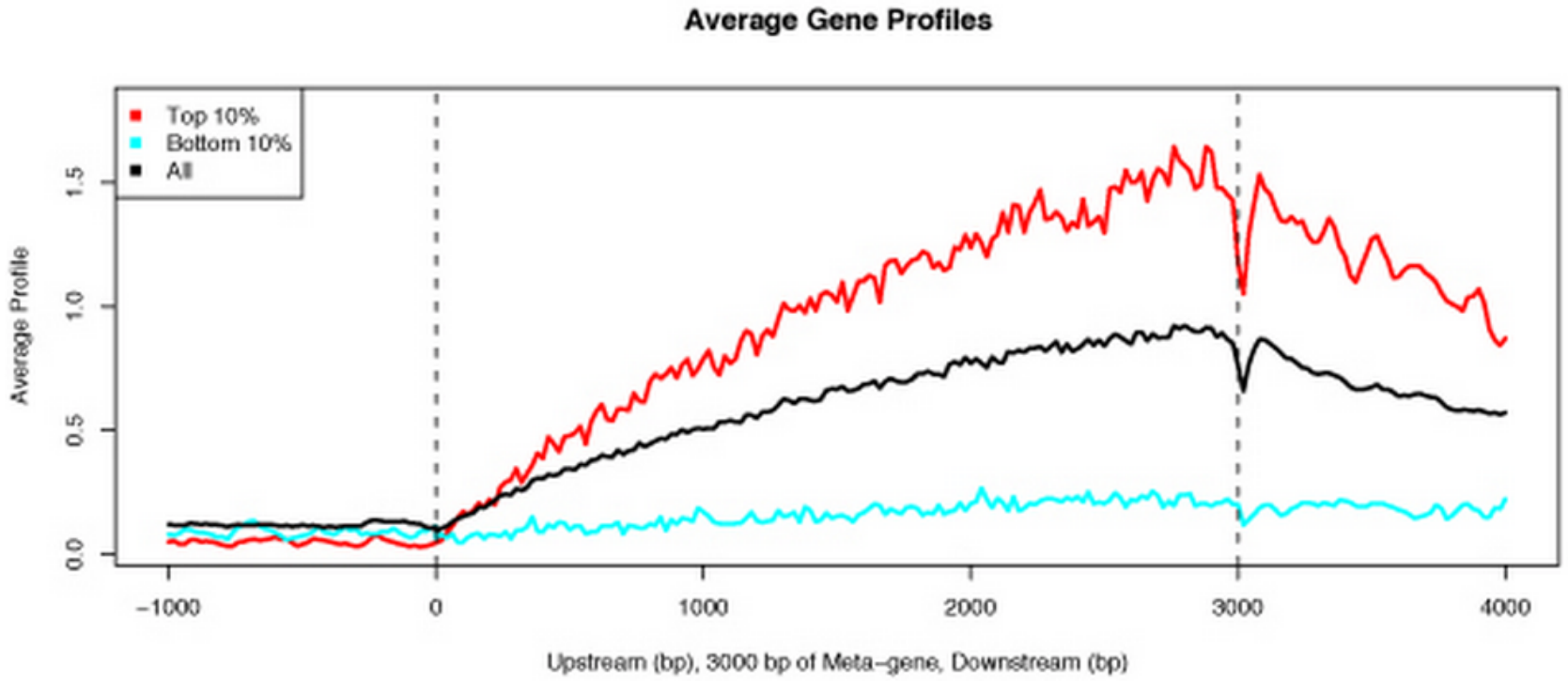

1. roby@SCHROEDER: ~ (ssh)
roby@SCHROEDER: ~ (ssh)
roberto@seqC: ~ (ssh)
SCHROEDER:~ roby$ ssh rspreafi@hoffman2.idre.ucla.edu
rspreafi@hoffman2.idre.ucla.edu's password:
Last login: Wed Sep 30 11:37:26 2015 from 164.67.9.48
Welcome to the Hoffman2 Cluster!
Hoffman2 Home Page: http://www.hoffman2.idre.ucla.edu
Consulting: https://support.idre.ucla.edu/helpdesk
All login nodes should be accessed via "hoffman2.idre.ucla.edu".
Please do NOT compute on the login nodes.
Processes running on the login nodes which seriously degrade others'
use of the system may be terminated without warning. Use qrsh to obtain
an interactive shell on a compute node for CPU or I/O intensive tasks.
The following news items are currently posted:
Seminar Mathematical Modeling with MATLAB
IDRE Fall 2015 Classes
Purchased storage migration of /u/home/groupname directories
News Archive On Web Site
Enter shownews to read the full text of a news item.
-bash-4.1$ qrsh -l i,h_data=8G,h_rt=3:30:00
Last login: Tue Aug 25 10:41:22 2015 from login1
-bash-4.1$ uname -a
Linux n2003 2.6.32-504.23.4.el6.x86_64 #1 SMP Tue Jun 9 20:57:37 UTC 2015 x86_64 x86_64 x86_64 GNU/Linux
-bash-4.1$

Schedule

- **Day 1**
  - Introduction
  - Cross-correlation analysis and ENCODE QC with SPP
  - BigWig tracks with defined resolution and normalization using Homer/ UCSC tools, and visualization with IGV
- **Day 2**
  - Peak calling with MACS2
  - QC of replicates with ENCODE's IDR
  - Differential peak calling with MAnorm
- **Day 3**
  - Location annotation with NGSPLOT
  - Motif analysis with HOMER
  - Functional annotation with GREAT
  - (Unix tricks, tools installation)

LOCATION-BASED ANALYSIS

Example


Average Gene Profiles
Top 10%
Bottom 10%
All
Average Profile
0.0
0.5
1.0
1.5
−1000
0
1000
2000
3000
4000
Upstream (bp), 3000 bp of Meta−gene, Downstream (bp)

Let's practice

Work from your sample folder, e.g. /u/scratch/r/rspreafi/samples/

mkdir ngsplot

module load ngsplot

ngs.plot.r -G mm9 -SS both -FL 200 -R tss -L 5000 -C IPr1.bam -O ngsplot/IPr1_tss -F chipseq -T TSS

both strands

examine 5kb upstream and downstream the TSS

https://github.com/shenlab-sinai/ngsplot/wiki/ProgramArguments101

Many tools for the job

ngsplot

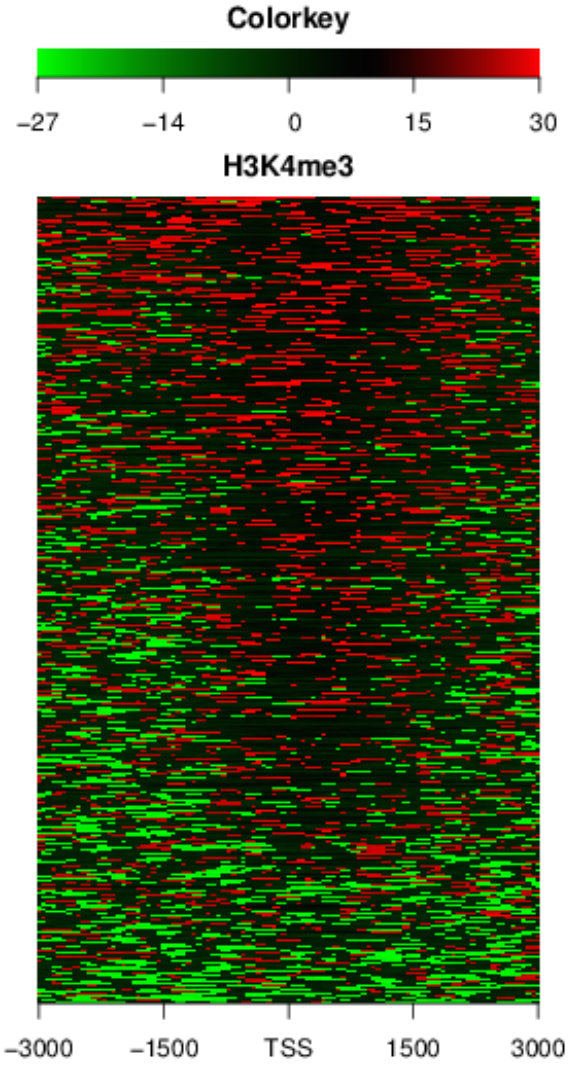
code.google.com/p/ngsplot/

ChIPpeakAnno

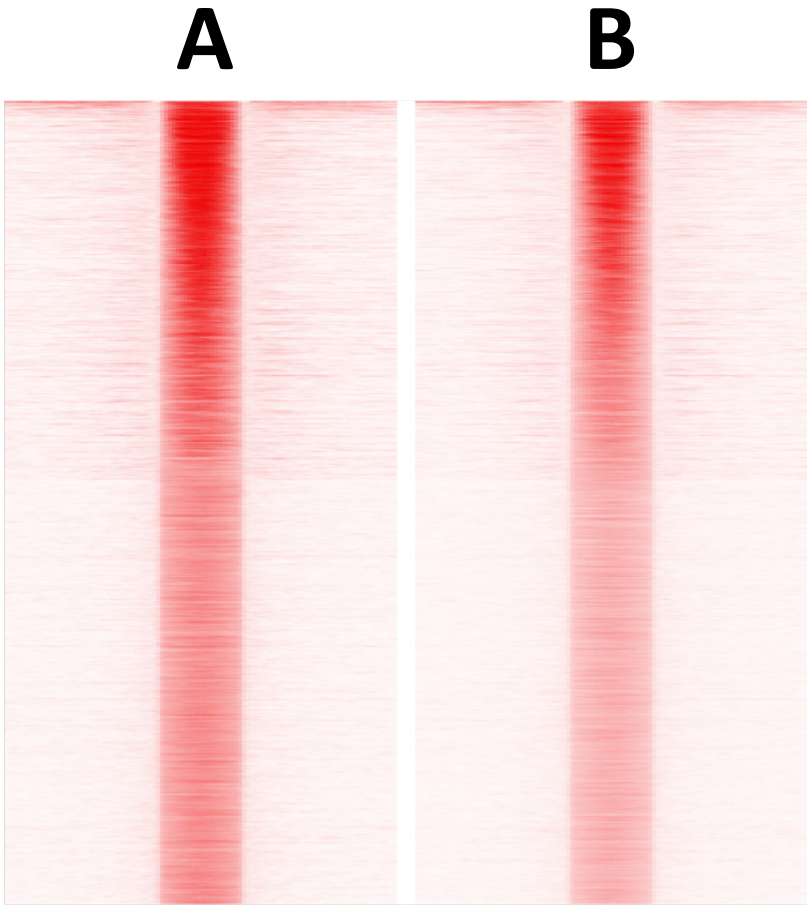
deepTools

CEAS

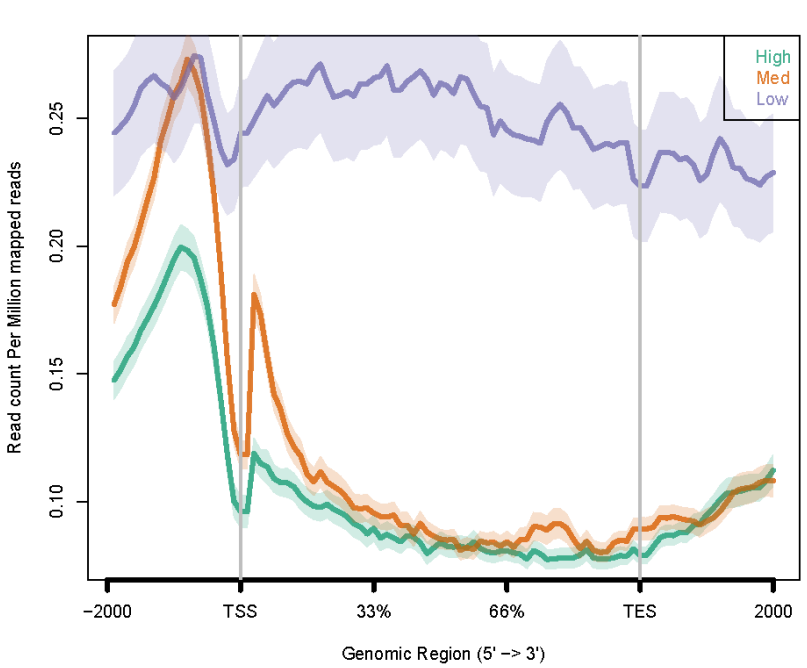
-C sample1.bam:sample2.bam

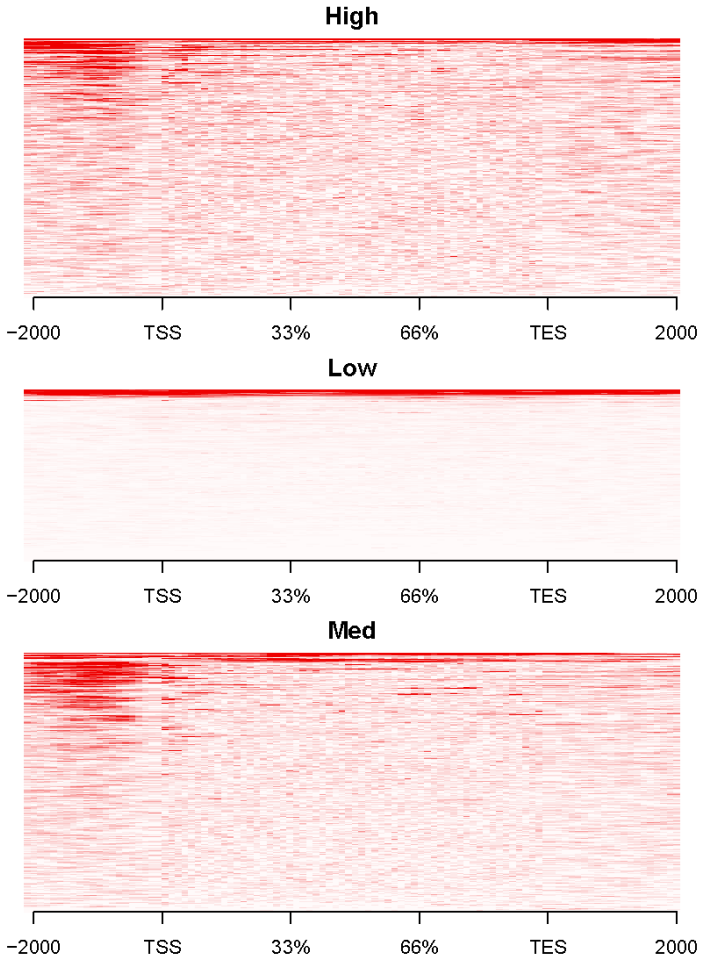


-C config.txt
(two .bam files, same custom bed regions)

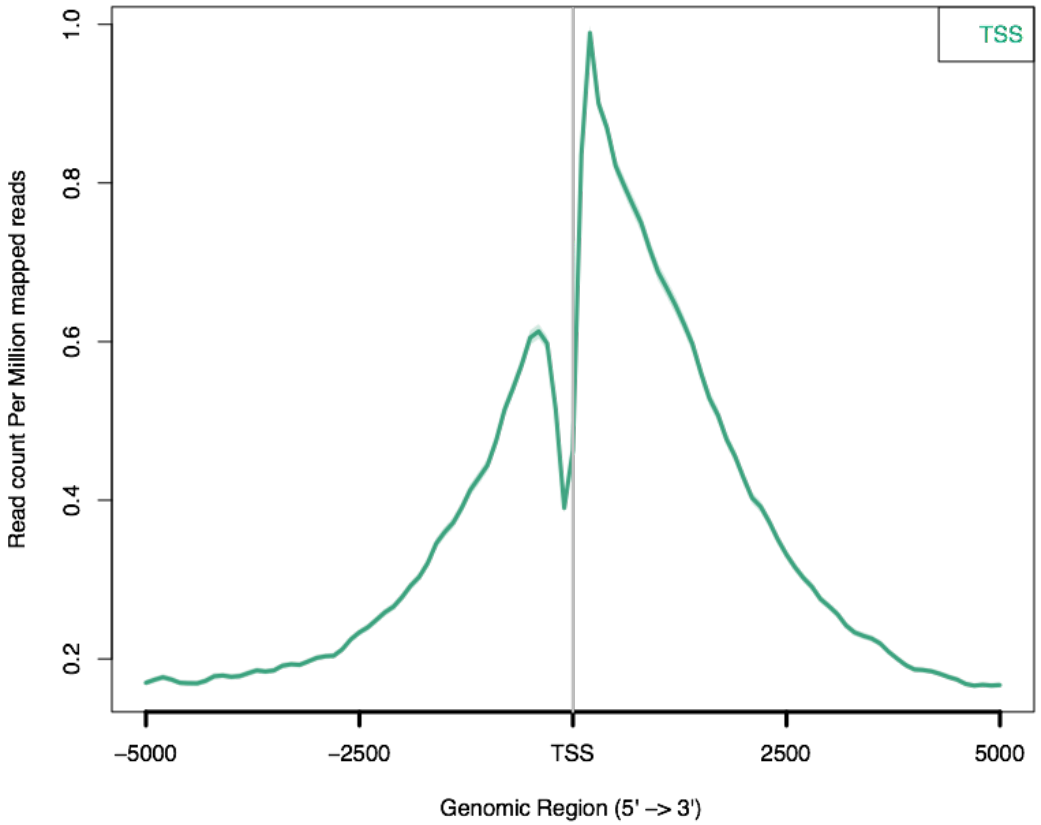


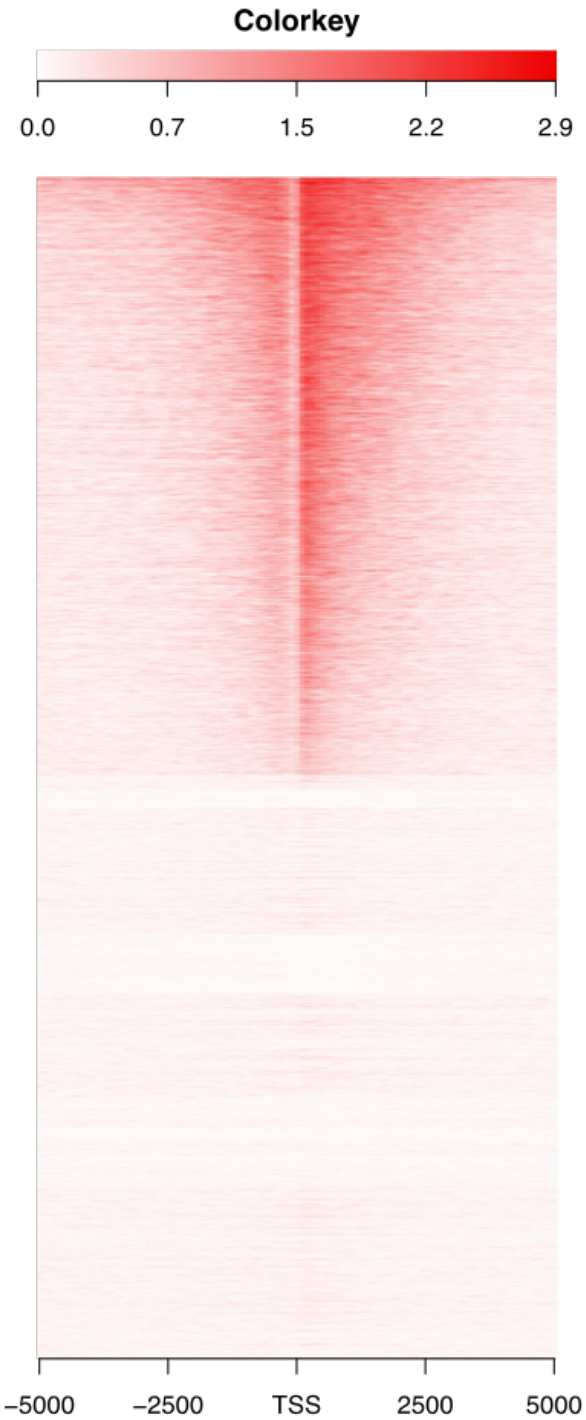
-C config.txt
(same .bam file, different gene lists)





Retrieve your data





MOTIF ANALYSIS

Let's practice

Work from your sample folder, e.g. /u/scratch/r/rspreafi/samples/

cp /u/scratch/r/rspreafi/samples/IP_motif.bam .

module load python/2.7.3

macs2 callpeak -t IP_motif.bam --nolambda -f BAM -n IP_motif -g mm -q 0.01 --keep-dup 1 --call-summits --nomodel --extsize 70 --outdir macs2 >macs2/IP_motif_macs.txt 2>&1 &

Some tools for the job

Bioinformatics Evolution @nd COmparative GeNomics

Pscan-ChIP

Ver. 1.1 (Last update: 27 Sep 2014)

HOMER

http://homer.salk.edu/

The MEME Suite

Motif-based sequence analysis tools

Let's practice

Work from your sample folder, e.g. /u/scratch/r/rspreafi/samples/

module load homer

findMotifsGenome.pl macs2/IP_motif_peaks.narrowPeak mm10 homer/IP_motif -mask -len 8 -S 5 -mis 2 -size 100 -preparsedDir homer/preparsed/mm10 >homer/IP_motif_output.txt 2>&1 &

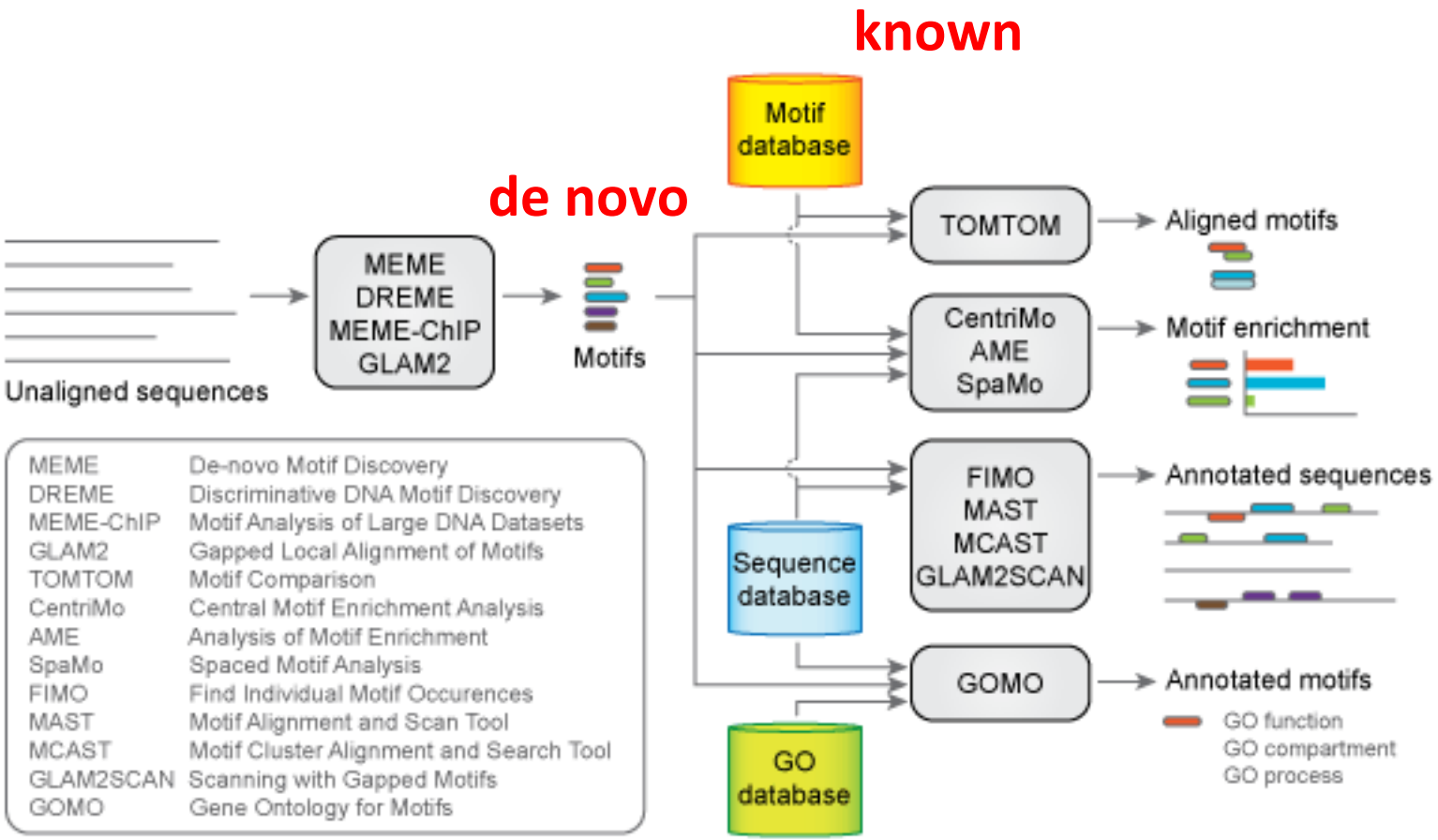
mask repeats

de novo motifs to find

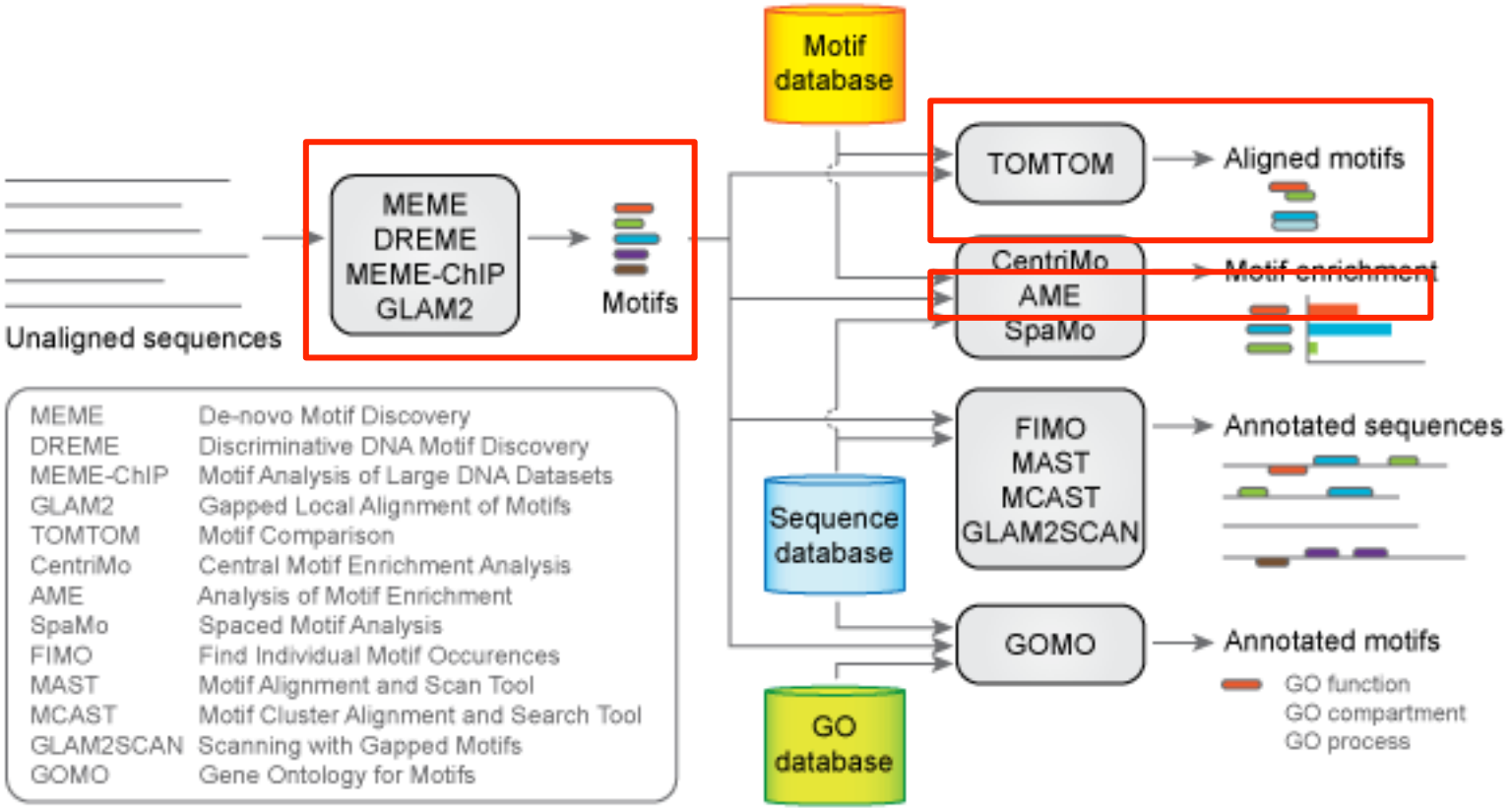
# mismatches
higher = more sensitive, but slower

bases for each peak; or
-size given

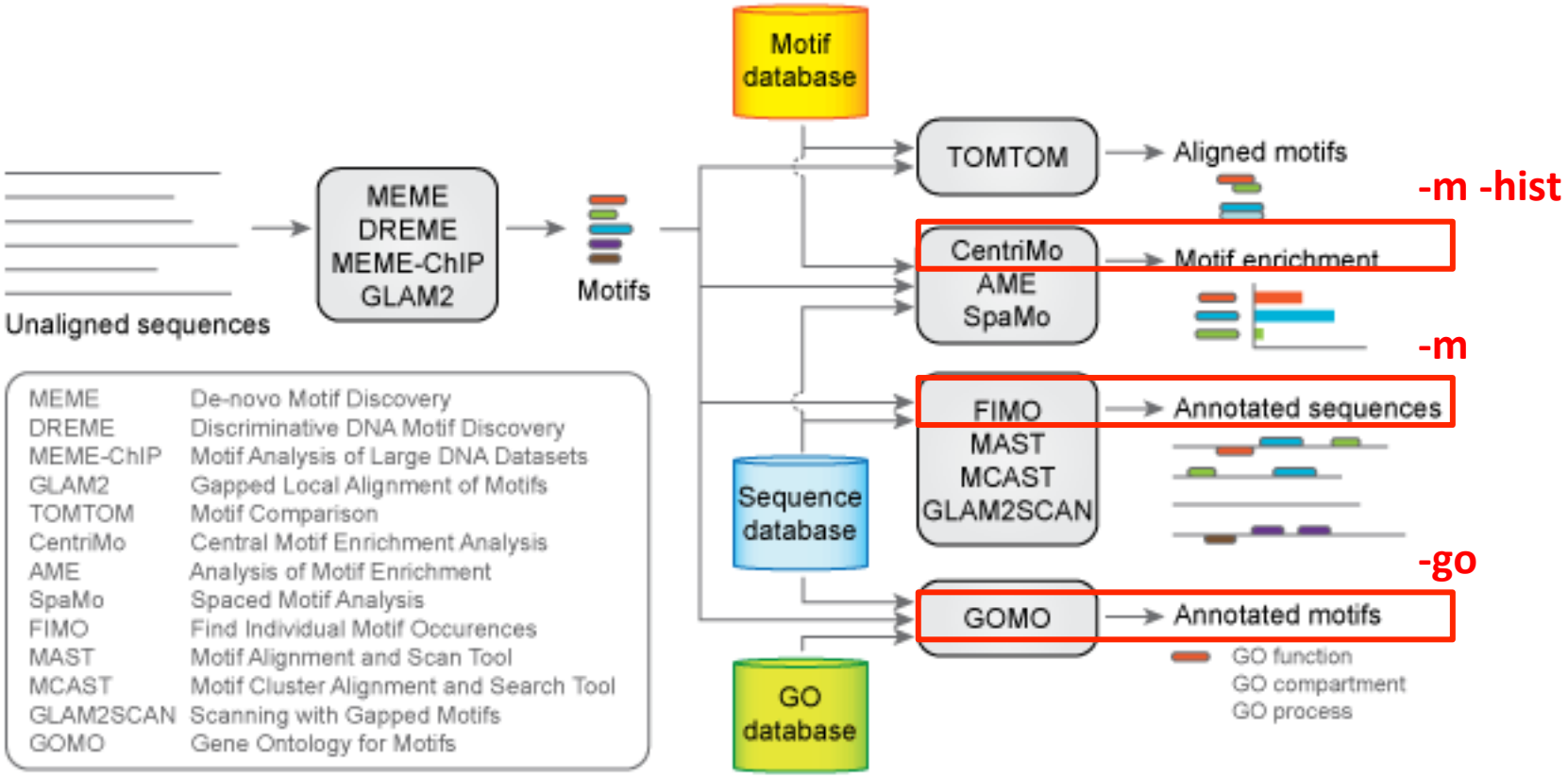
Overview


known
Motif database
de novo
TOMTOM
Aligned motifs
MEME
DREME
MEME-ChIP
GLAM2
Motifs
CentriMo
AME
SpaMo
Motif enrichment
Unaligned sequences
FIMO
MAST
MCAST
GLAM2SCAN
Annotated sequences
Sequence database
GOMO
Annotated motifs
GO function
GO compartment
GO process
GO database
MEME De-novo Motif Discovery
DREME Discriminative DNA Motif Discovery
MEME-ChIP Motif Analysis of Large DNA Datasets
GLAM2 Gapped Local Alignment of Motifs
TOMTOM Motif Comparison
CentriMo Central Motif Enrichment Analysis
AME Analysis of Motif Enrichment
SpaMo Spaced Motif Analysis
FIMO Find Individual Motif Occurences
MAST Motif Alignment and Scan Tool
MCAST Motif Cluster Alignment and Search Tool
GLAM2SCAN Scanning with Gapped Motifs
GOMO Gene Ontology for Motifs

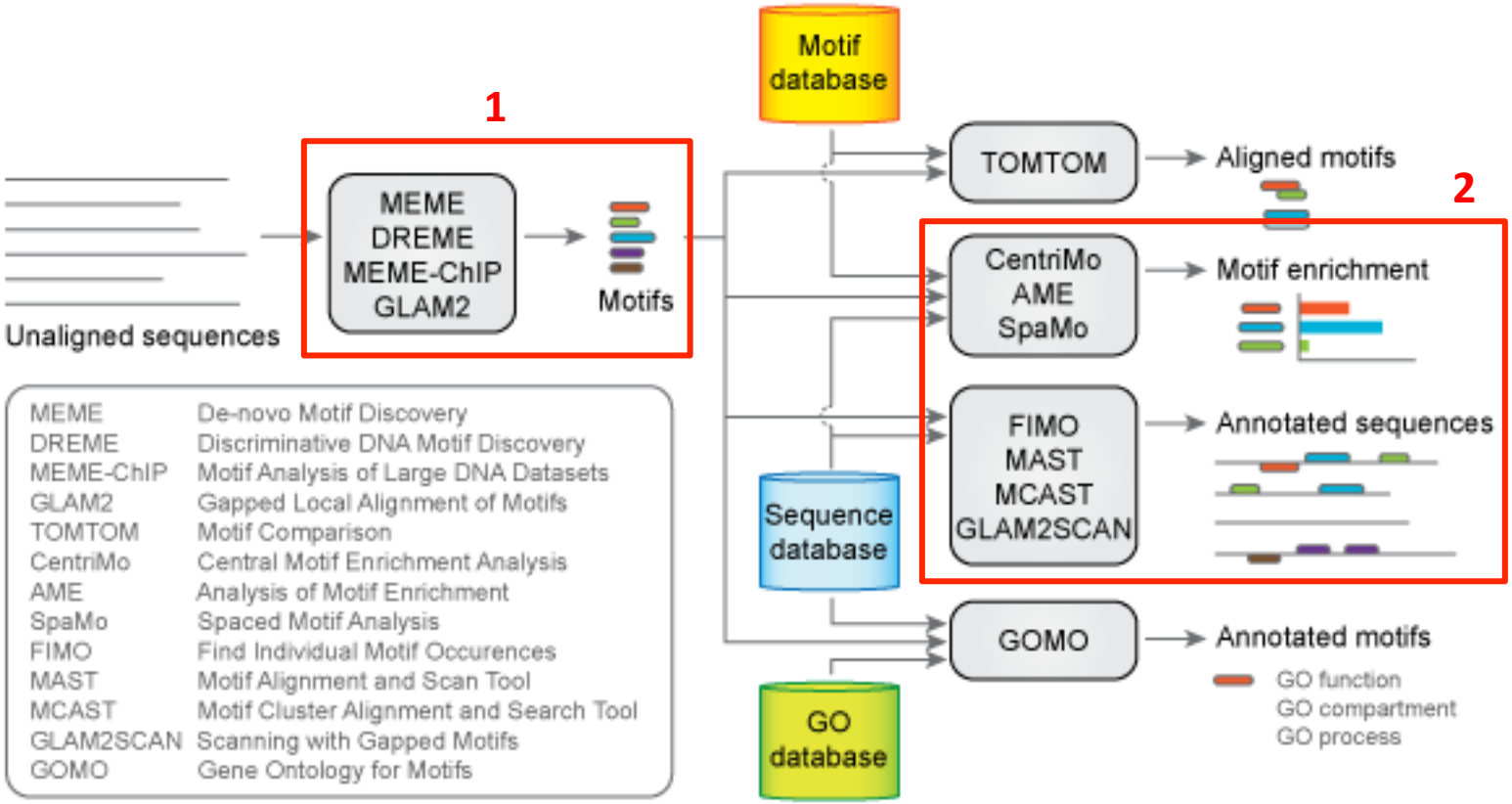
HOMER's findMotifsGenome.pl



HOMER's annotatePeaks.pl



Identifying motifs



Two separate steps: finding a motif in foreground does not mean it is enriched over background!

Defining a motif

```
GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT
```

count

positions

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

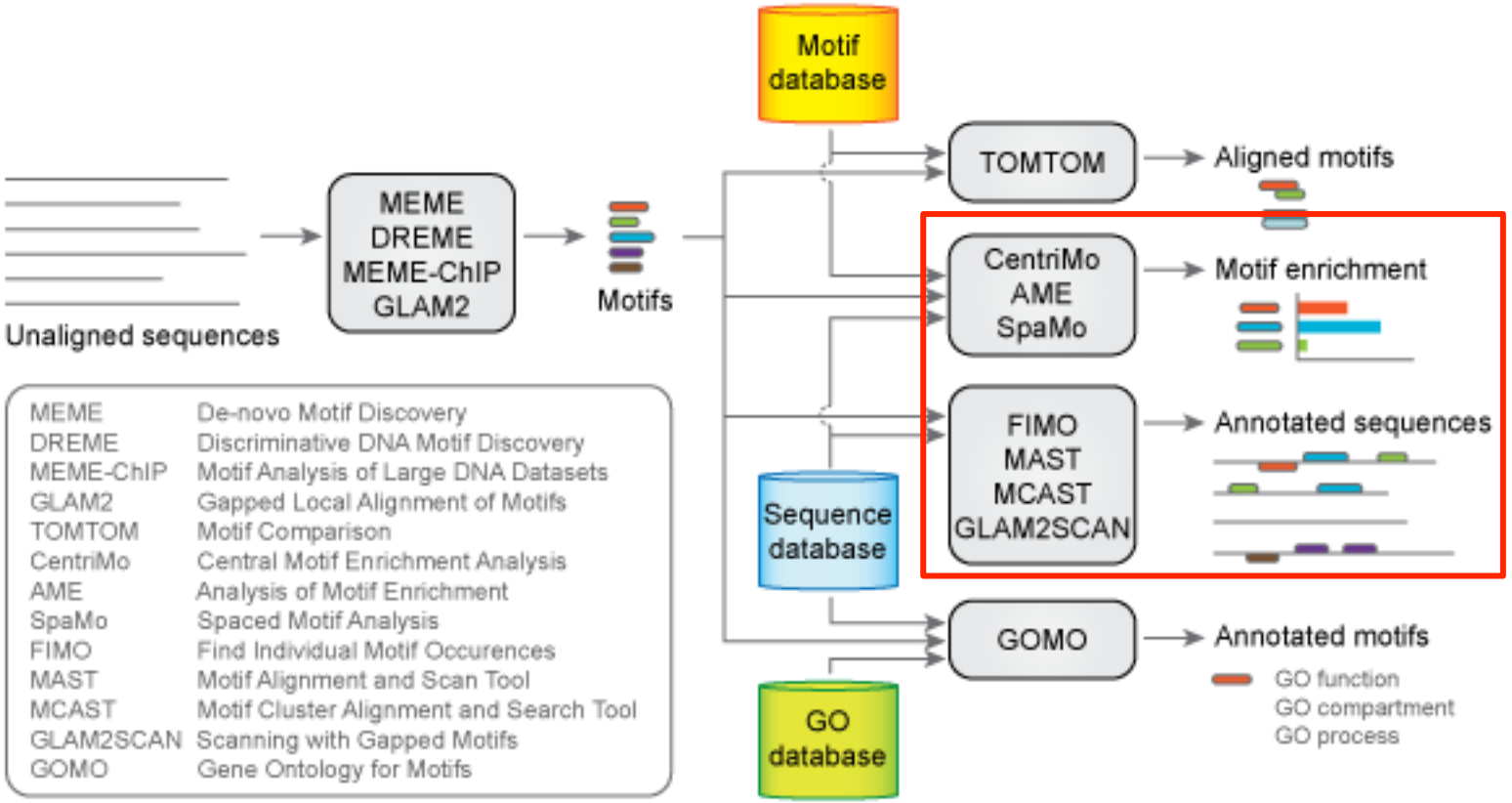
%

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

ln(%/0.25)

$$M = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.18 & 0.87 & -0.91 & -\infty & -\infty & 0.87 & 1.02 & -0.22 & -0.91 \\ -0.22 & -0.22 & -0.91 & -\infty & -\infty & -0.22 & -0.91 & -0.91 & -0.22 \\ -0.91 & -0.91 & 1.02 & 1.38 & -\infty & -0.91 & -0.91 & 0.69 & -0.91 \\ 0.47 & -0.91 & -0.91 & -\infty & 1.38 & -0.91 & -0.91 & -0.22 & 0.87 \end{bmatrix}$$

Spotting motifs in sequences



Is AGGGTACAT related to my motif?

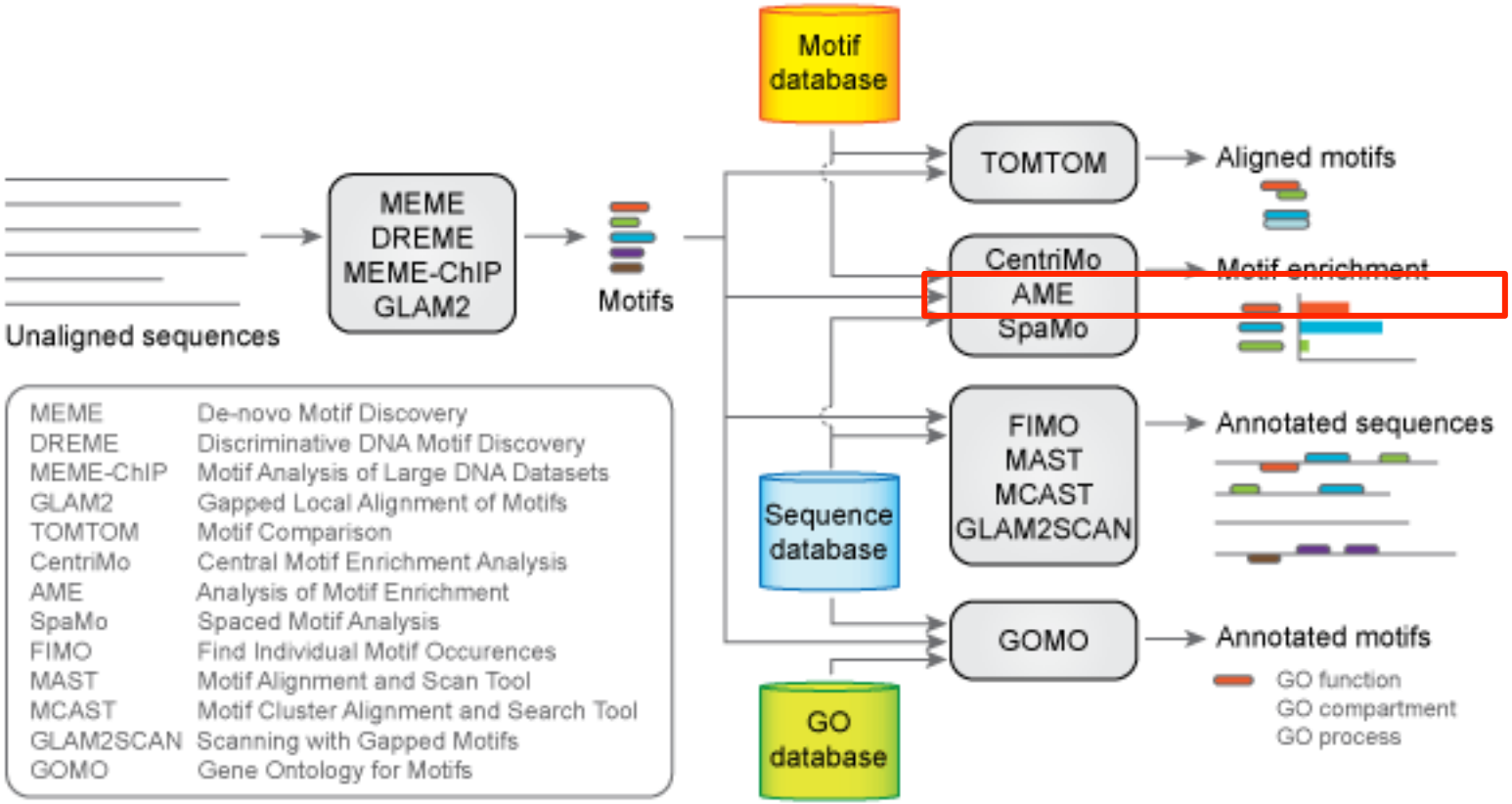
need to correct, e.g. pseudocounts

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.18 & 0.87 & -0.91 & -\infty & -\infty & 0.87 & 1.02 & -0.22 & -0.91 \\ -0.22 & -0.22 & -0.91 & -\infty & -\infty & -0.22 & -0.91 & -0.91 & -0.22 \\ -0.91 & -0.91 & 1.02 & 1.38 & -\infty & -0.91 & -0.91 & 0.69 & -0.91 \\ 0.47 & -0.91 & -0.91 & -\infty & 1.38 & -0.91 & -0.91 & -0.22 & 0.87 \end{bmatrix}$$

0.18 - 0.91 + 1.02 + 1.38 + 1.38 + 0.87 - 0.91 - 0.22 + 0.87 = **3.66**

- > 0 more likely to be functional site (related to motif)
- < 0 more likely to be random site (not related to motif)
- = 0 equal probability of being either functional or random
- thresholds are determined by HOMER to maximize enrichment over background (de novo) or specified in internal DB (known)

Is motif enriched in IP'd regions?



Hypergeometric distribution

	Males	Females
UCLA students	5,000	5,000
World population	3 billion	3 billion

Hypergeometric distribution

	Males	Females
UCLA students	5,000	5,000
World population	3 billion	3 billion

	Males	Females
UCLA students	7,000	3,000
World population	3 billion	3 billion

Hypergeometric distribution

	Males	Females
UCLA students	5,000	5,000
World population	3 billion	3 billion

"not males"

	Males	Females
UCLA students	7,000	3,000
World population	3 billion	3 billion

	Motif present	Motif absent
Peaks	20	100
Genome (bg)	200	20,000

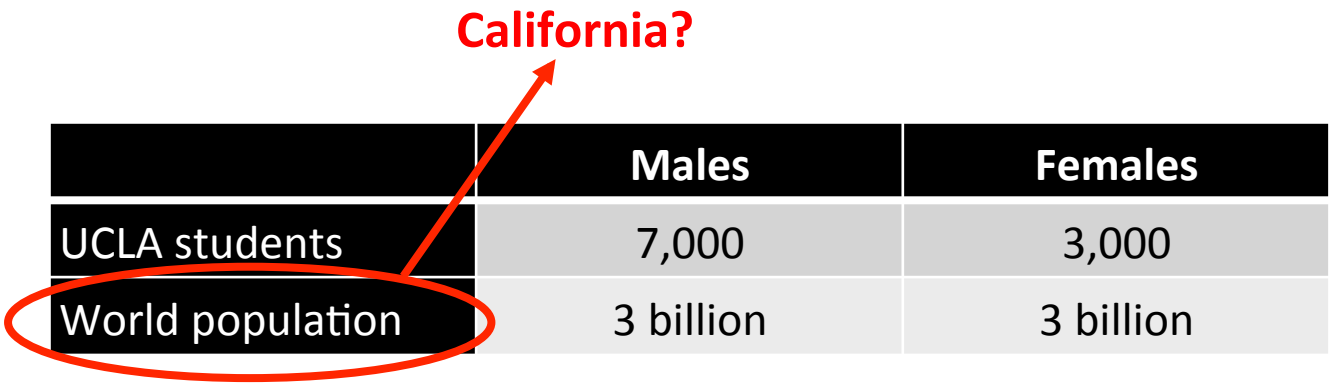
Hypergeometric vs Binomial

	Drawn	Not drawn	
Green marbles	20	100	**120**
Red marbles	180	19,900	**20,080**
	200	**20,000**	

- Note the different (but equivalent) notation: two classes rather than one class and total
- Is sampling with replacement?
  - Yes = Binomial
  - No = Hypergeometric
- HOMER suggests that the hypergeometric distribution may make more sense for motif enrichment applications
- By default, however, it uses the binomial distribution because it is faster to compute, and the error is small if background is large

Background selection

- From genomic regions (TSS +/- 50kb)
  - But you can provide a custom background
- With features similar to query sequences:
  - %GC or %CpG
  - 2-mers, 3-mers



	Males	Females
UCLA students	7,000	3,000
World population	3 billion	3 billion

Retrieve your results: de novo motifs

Homer *de novo* Motif Results (homer/IP_motif/)

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STAMP
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 15827
Total background sequences = 32534
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-3750	-8.636e+03	52.74%	10.59%	22.2bp (34.3bp)	MA0107.1_RELA/Jaspar(0.841) More Information | Similar Motifs Found	motif file (matrix)
2		1e-441	-1.017e+03	9.81%	2.49%	27.0bp (29.3bp)	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer(0.974) More Information | Similar Motifs Found	motif file (matrix)
3		1e-118	-2.725e+02	8.27%	4.13%	26.6bp (31.2bp)	TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer(0.875) More Information | Similar Motifs Found	motif file (matrix)
4		1e-92	-2.130e+02	3.92%	1.53%	26.9bp (26.2bp)	Atf1(bZIP)/K562-ATF1-ChIP-Seq(GSE31477)/Homer(0.953) More Information | Similar Motifs Found	motif file (matrix)

Retrieve your results: known motifs

Homer Known Motif Enrichment Results (homer/IP_motif)

Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 15828, Total Background Sequences = 32543

Rank	Motif	Name	P-value
1		NFkB-p65(RHD)/GM12787-p65-ChIP-Seq(GSE19485)/Homer	1e-5673
2		NFkB-p65-Rel(RHD)/ThioMac-LPS-Expression(GSE23622)/Homer	1e-2961
3		NFkB-p50,p52(RHD)/Monocyte-p50-ChIP-Chip(Schreiber et al.)/Homer	1e-1057
4		BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-461
5		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-457

FUNCTIONAL ENRICHMENT

Many tools for the job

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

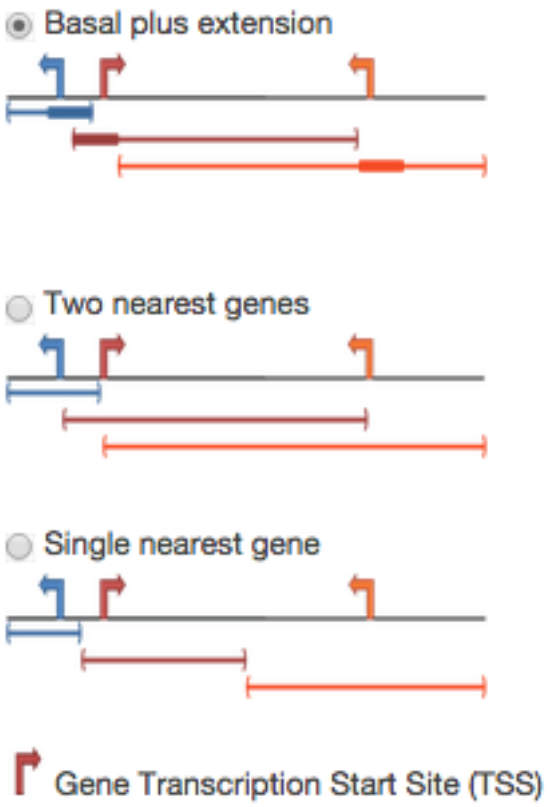
HOMER

The MEME Suite

Motif-based sequence analysis tools

Nothing new under the sun

- This is standard GO analysis, aka gene set analysis
- In transcriptomics, many tools:
  - threshold-based (e.g. DAVID)
  - ranking-based (e.g. GSEA)
- In ChIP-seq, just an extra step: associate peaks to genes by proximity, then use genes to perform standard GO analysis



Why GREAT?

- Easy (web interface)
- Great help
- Actively maintained
- Smarter peak-gene association rules
- Queries many databases (ontologies) simultaneously, not only GO

You can use GREAT to associate genomic regions to genes

What do these tables show?

Genomic region -> gene association table Download table as text.

Region	Gene (distance to TSS)
IP_motif_peak_14210	Chd2 (+3,859), Rgma (+162,367)
IP_motif_peak_13096	Gimap3 (-63,274), Tmem176b (+11,195)
IP_motif_peak_1337a	Tnfaip3 (-83)
IP_motif_peak_2878	Ccl2 (-2,470)
IP_motif_peak_10320b	Mex3a (-22,195), Lmna (-6,865)
IP_motif_peak_13928b	Zfp36 (-2,802)
IP_motif_peak_8014	Ldlrad4 (+125,233), Fam210a (+241,843)
IP_motif_peak_9700	Foxs1 (+1,482), Mylk2 (+20,374)
IP_motif_peak_14278	Pde8a (-58,481), Slc28a1 (+40,265)
IP_motif_peak_15737	Birc3 (-30)
IP_motif_peak_3703a	Nfkbia (-66)
IP_motif_peak_11229	Dmrta1 (-294,579), Cdkn2b (-82,587)
IP_motif_peak_15201	A230052G05Rik (+2,619), Ccdc124 (+7,744)
IP_motif_peak_2650	Stx8 (+125,932), Ntn1 (+308,698)
IP_motif_peak_5746	9130401M01Rik (-102)
IP_motif_peak_14634	Tufm (-126)

Gene -> genomic region association table Download table as text.

Gene	Region (distance to TSS)
0610007P14Rik	IP_motif_peak_3934 (+65,088)
0610009L18Rik	IP_motif_peak_3449 (+6,433)
1100001G20Rik	IP_motif_peak_2906 (-17,644)
1110001J03Rik	IP_motif_peak_13066 (-14)
1110004E09Rik	IP_motif_peak_6720a (+47,798), IP_motif_peak_6720b (+47,798), IP_motif_peak_6718 (+79,186)
1110007C09Rik	IP_motif_peak_4445 (+10,046)
1110017D15Rik	IP_motif_peak_10954 (+3,468)
1110037F02Rik	IP_motif_peak_10851 (-32,016)
1200011I18Rik	IP_motif_peak_5328 (+79,269)
1500009L16Rik	IP_motif_peak_1833 (-10,913), IP_motif_peak_1835b (+186,283), IP_motif_peak_1837 (+248,012), IP_motif_peak_1838 (+423,881), IP_motif_peak_1841 (+667,675), IP_motif_peak_1842 (+679,135), IP_motif_peak_1844 (+710,556), IP_motif_peak_1846 (+714,952)

Déjà vu

	Motif present	Motif absent
Peaks	20	100
Genome (bg)	200	20,000

	In pathway	Not in pathway
Genes with peaks	20	100
All genes	200	20,000

Let's practice

Work from your macs2 folder, e.g. /u/scratch/r/rspreafi/samples/macs2

sort -k8,8nr IP_motif_peaks.narrowPeak | awk 'BEGIN{OFS="\t"}{print $1,$2,$3,$4}' | head -n 5000 >IP_motif_peaks.bed

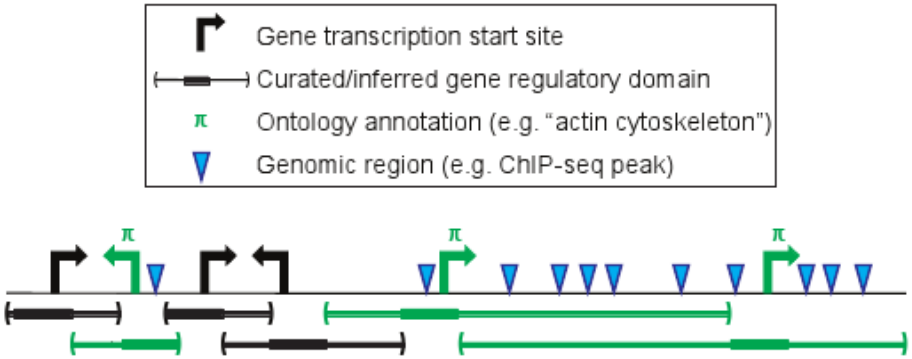
- Go to bejerano.stanford.edu/great/public/html/
- Select:
  - mouse mm10
  - test regions: BED file, upload IP_motif_peak.bed
  - background regions: whole genome
    - be fair when defining background
  - "basal plus extension" association rule

Results

GO Molecular Function (17 terms) — Global controls

Table controls: Export | Shown top rows in this table: 20 Set | Term annotation count: Min: 1 Max: Inf Set | Visualize this table: [select one]

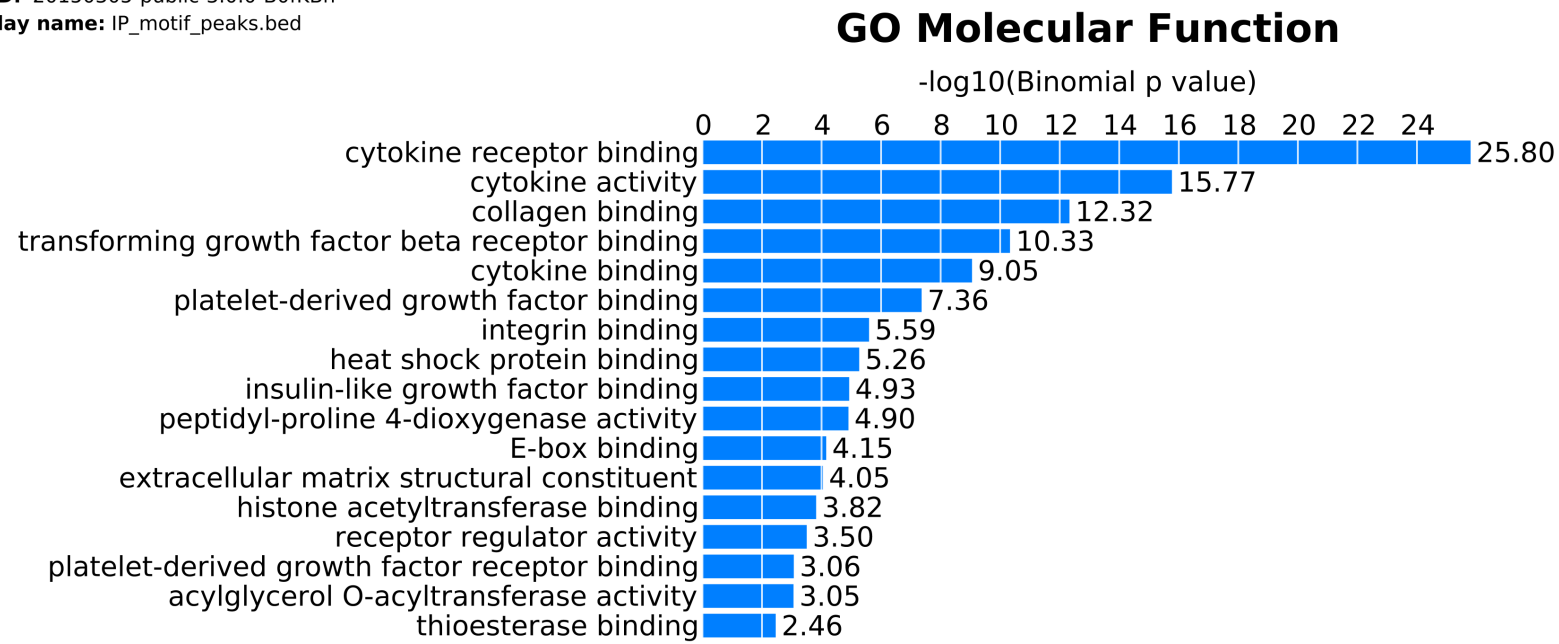
Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment	Hyper Observed Gene Hits	Hyper Total Genes	Hyper Gene Set Coverage
cytokine receptor binding	10	1.5950e-26	5.5650e-24	2.4682	179	3.58%	54	2.2368e-6	1.6953	89	223	1.73%
cytokine activity	19	1.7157e-16	3.1505e-14	2.2125	134	2.68%	141	4.5667e-2	1.3821	68	209	1.32%
collagen binding	30	4.7514e-13	5.5259e-11	3.1594	55	1.10%	80	2.1285e-3	2.3738	19	34	0.37%
transforming growth factor beta receptor binding	48	4.7190e-11	3.4301e-9	3.8539	35	0.70%	71	2.9528e-4	3.2484	13	17	0.25%
cytokine binding	59	8.8772e-10	5.2496e-8	2.5178	57	1.14%	93	5.4946e-3	1.9439	27	59	0.52%



bejerano.stanford.edu/help/display/GREAT/Statistics

Results

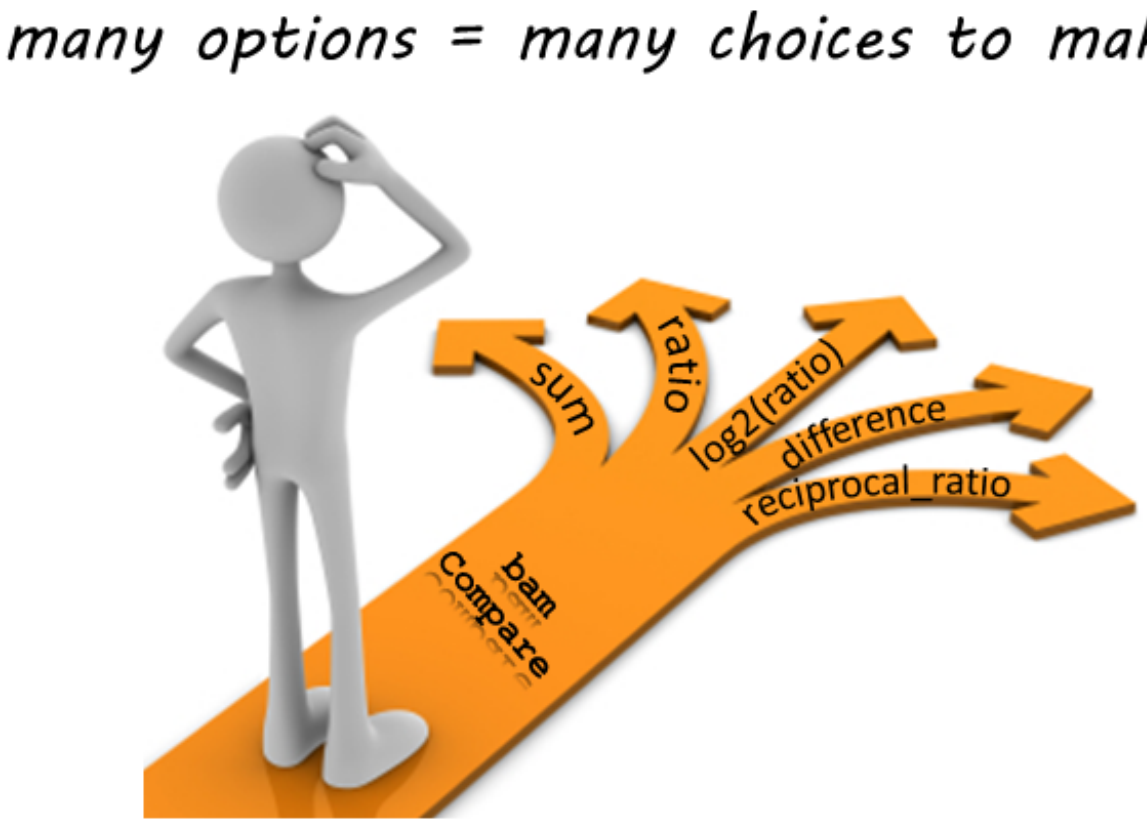
Job ID: 20150305-public-3.0.0-B0fKBh
Display name: IP_motif_peaks.bed



PE reads

- With PE reads you know the fragment length, so you do not need to run cross-correlation
- MACS2 has a *-f BAMPE* option
  - but it then uses only the mean fragment size
- ngsplot automatically extends reads by using fragment size
  - undocumented, as per Google Groups Q&A
- MAnorm does not have dedicated options
- HOMER deals with PE reads
  - but you need to specify *-sspe* (and possibly *-flip*) if the library is strand-specific (usually it is not in ChIP-seq)
- when PE-specific options are not available, you can always approximate by treating PE reads as SE reads
  - Loss of fragment length information
  - Fragments get double counted (1 count for each end)
- look at manuals for more details on these options

The bioinformatics dilemma



http://romebusinessschool.it/en/decision-making-efficace/

Do you feel you have a lot to digest?

No worries, it’s normal!

Leave no trace
(and protect your privacy)

- Delete anything that remembers your password from shared computers
  - e.g. Cyberduck
- Sign out of Gmail if you used it
- Delete personal documents

Please take 5 min for a brief survey

Your feedback is important to us:

www.surveymonkey.com/r/w7april

Thank you for coming! We hope you enjoyed it!