

# Day2, Bulk RNA-seq & ChIP-seq

NIEHS Workshop--Analyzing NGS data  
[https://github.com/zhu16/NIEHS\\_NGS\\_Workshop](https://github.com/zhu16/NIEHS_NGS_Workshop)

July 14, 2021  
9am-2pm

Yunhua Zhu, PhD  
zhanghua@gmail.com  
Computational Genomics Specialist, Transcriptomics  
Medical Science & Computing

# Agenda -- day 2, bulk RNA-seq and ChIP-seq

## Part II, Bulk RNA-seq

- 9:00-9:15AM, connect to biowulf
- 9:15-9:30AM, introduction to bulk RNA-seq
- 9:30-10AM, From Fastq files, alignment with RSEM, Tophat
- 10AM-10:30AM, Visualize the output using IGV.
- 10:30-11:00AM, Normalization and differential analysis with DESeq2
- Questions and break

## Part III, Chip-seq

- 11:30AM-12PM, Introduction to ChIP-seq analysis
- 12:00-12:15AM, Alignment to genome using Bowtie2
- 12:15-12:45AM, Use MACS2 to call peaks, and use Diffbind to call differential peaks
- 12:45-1:15AM, Use CEAS to annotate peaks and summarize statistics.
- 1:15am -1:45AM, Use Homer to study TF binding site analysis. And use GREAT for gene ontology.

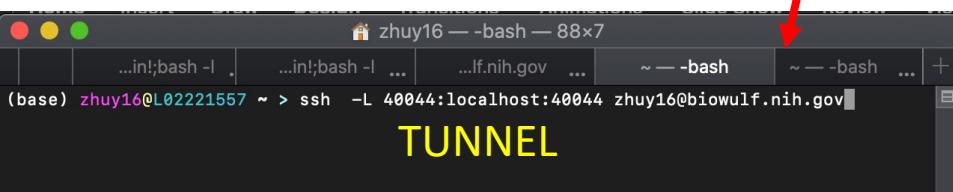
## Review & Summary

# Install software to visualizing alignment results

- Install IGV on your local machine
  - <https://software.broadinstitute.org/software/igv/download>
- Install IGM on your local machine
  - <https://www.bioviz.org/>

# Start up the jupyter lab, --our working interface

- # Open another terminal
- ssh user\_id@biowulf.nih.gov
- Enter password
- module load tmux
- module load tmux; tmux new -ct 'sinteractive --mem=100g --time=12:00:00 --tunnel'
- # Copy the tunnel script to another (3<sup>rd</sup>) terminal, execute and enter password, to establish a ssh tunnel between local computer and the work node.
- module load jupyter R/4.0.5 && jupyter lab --ip localhost --port \$PORT1 --no-browser
- # wait until an URL link appears, copy it to your web browser, to get connected to biowulf through a jupyter lab interface.



TUNNEL

Log in screen

\*\*\*WARNING\*\*\*

You are accessing a U.S. Government information system, which includes (1) this computer, (2) this computer network, (3) all computers connected to this network, and (4) all devices and storage media attached to this network or to a computer on this network. This information system is provided for U.S. Government-authorized use only. Unauthorized or improper use of this system may result in disciplinary action, as well as civil and criminal penalties.

By using this information system, you understand and consent to the following:

- \* You have no reasonable expectation of privacy regarding any communications or data transiting or stored on this information system. At any time, and for any lawful Government purpose, the government may monitor, intercept, record, and search and seize any communication or data transiting or stored on this information system.
- \* Any communication or data transiting or stored on this information system may be disclosed or used for any lawful Government purpose.

Notice to users: This system is rebooted for patches and maintenance on the first Monday of every month at 7:15AM unless Monday is a holiday, in which case it is rebooted the following Tuesday. Running cluster jobs are not affected by the monthly reboot.

bcell CSIMicrobes deprived\_210611.bashrc ini\_conda\_base.sh ncbi-outdir R r\_site-library scratch staging\_for\_deletion zaya  
sallo conda.sh data igv ini\_miniconda\_base.sh NIEHS\_NGS r\_time-library salloc.exe: Waiting for resource configuration  
salloc.exe: Nodes cn2006 are ready for job  
srun: error: x11: no local DISPLAY defined, skipping  
error: unable to open file /tmp/slurm-spawn-x11.17133459.0  
slurmstepd: error: x11: unable to read DISPLAY value  
  
Created 1 generic SSH tunnel(s) from this compute node to biowulf for your use at port numbers defined in the \$PORTn (\$PORT1, ...) environment variables.  
  
Please create a SSH tunnel from your workstation to these ports on biowulf. On Linux/MacOS, open a terminal and run:  
  
ssh -L 40044:localhost:40044 zhuy16@biowulf.nih.gov  
  
For Windows instructions, see <https://hpc.nih.gov/tunnels.html>

```
bash zhuy16@n2006:[Tue Jun 15]02:30 PM$ module load jupyter R/4.0.5 && jupyter lab  
[3] 0:srun
```

Module load jupyter R/4.0.5 && jupyter lab

```
module load jupyter R/4.0.5 && jupyter lab  
[4] 0:srun
```

```
[2] 0:srun
```

```
[3] 0:srun
```

```
[4] 0:srun
```

```
[5] 0:srun
```

```
[6] 0:srun
```

```
[7] 0:srun
```

```
[8] 0:srun
```

```
[9] 0:srun
```

```
[10] 0:srun
```

```
[11] 0:srun
```

```
[12] 0:srun
```

```
[13] 0:srun
```

```
[14] 0:srun
```

```
[15] 0:srun
```

```
[16] 0:srun
```

```
[17] 0:srun
```

```
[18] 0:srun
```

```
[19] 0:srun
```

```
[20] 0:srun
```

```
[21] 0:srun
```

```
[22] 0:srun
```

```
[23] 0:srun
```

```
[24] 0:srun
```

```
[25] 0:srun
```

```
[26] 0:srun
```

```
[27] 0:srun
```

```
[28] 0:srun
```

```
[29] 0:srun
```

```
[30] 0:srun
```

```
[31] 0:srun
```

```
[32] 0:srun
```

```
[33] 0:srun
```

```
[34] 0:srun
```

```
[35] 0:srun
```

```
[36] 0:srun
```

```
[37] 0:srun
```

```
[38] 0:srun
```

```
[39] 0:srun
```

```
[40] 0:srun
```

```
[41] 0:srun
```

```
[42] 0:srun
```

```
[43] 0:srun
```

```
[44] 0:srun
```

```
[45] 0:srun
```

```
[46] 0:srun
```

```
[47] 0:srun
```

```
[48] 0:srun
```

```
[49] 0:srun
```

```
[50] 0:srun
```

```
[51] 0:srun
```

```
[52] 0:srun
```

```
[53] 0:srun
```

```
[54] 0:srun
```

```
[55] 0:srun
```

```
[56] 0:srun
```

```
[57] 0:srun
```

```
[58] 0:srun
```

```
[59] 0:srun
```

```
[60] 0:srun
```

```
[61] 0:srun
```

```
[62] 0:srun
```

```
[63] 0:srun
```

```
[64] 0:srun
```

```
[65] 0:srun
```

```
[66] 0:srun
```

```
[67] 0:srun
```

```
[68] 0:srun
```

```
[69] 0:srun
```

```
[70] 0:srun
```

```
[71] 0:srun
```

```
[72] 0:srun
```

```
[73] 0:srun
```

```
[74] 0:srun
```

```
[75] 0:srun
```

```
[76] 0:srun
```

```
[77] 0:srun
```

```
[78] 0:srun
```

```
[79] 0:srun
```

```
[80] 0:srun
```

```
[81] 0:srun
```

```
[82] 0:srun
```

```
[83] 0:srun
```

```
[84] 0:srun
```

```
[85] 0:srun
```

```
[86] 0:srun
```

```
[87] 0:srun
```

```
[88] 0:srun
```

```
[89] 0:srun
```

```
[90] 0:srun
```

```
[91] 0:srun
```

```
[92] 0:srun
```

```
[93] 0:srun
```

```
[94] 0:srun
```

```
[95] 0:srun
```

```
[96] 0:srun
```

```
[97] 0:srun
```

```
[98] 0:srun
```

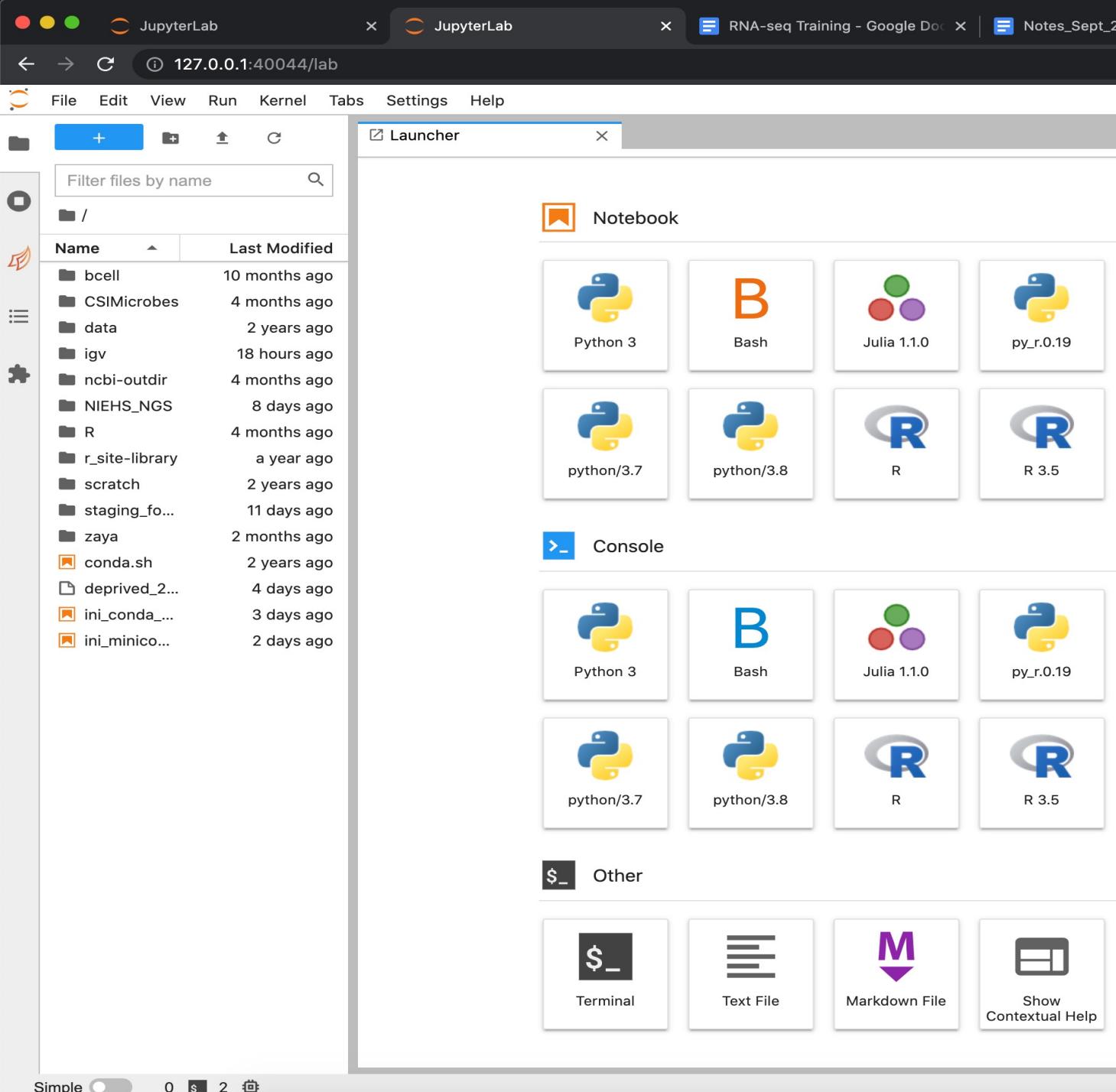
```
[99] 0:srun
```

To access the server, open this file in a browser:  
File:///Users/zhuy16/Desktop/jupyter/tunnel/jupyter-time-jpservr-26081-open.html  
Or connect directly to this URL:  
http://localhost:48844/lab?token=a4864789e1a833dc549976878a7c63566c65f58fb2f  
http://127.0.0.1:48844/lab?token=a4864789e1a833dc549976878a7c63566c65f58fb2f  
Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).

URL

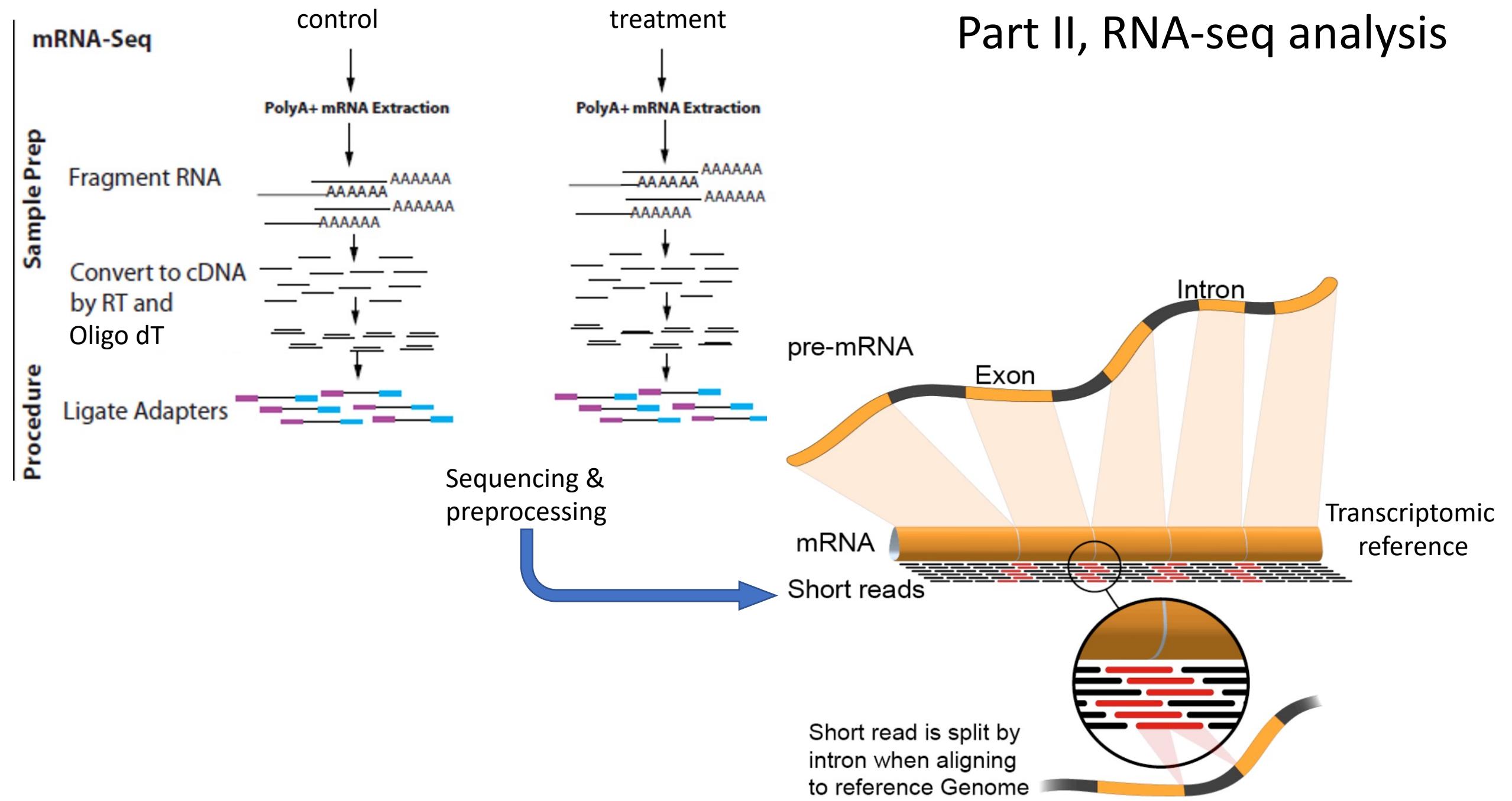
# Jupyter lab

- Navigate files
- Bash terminal
- Control kernels
- Jupyter notebooks for
  - Kernels
    - Bash
    - R
    - Python
    - Julia
  - Document analysis
  - Git version control
  - Report results



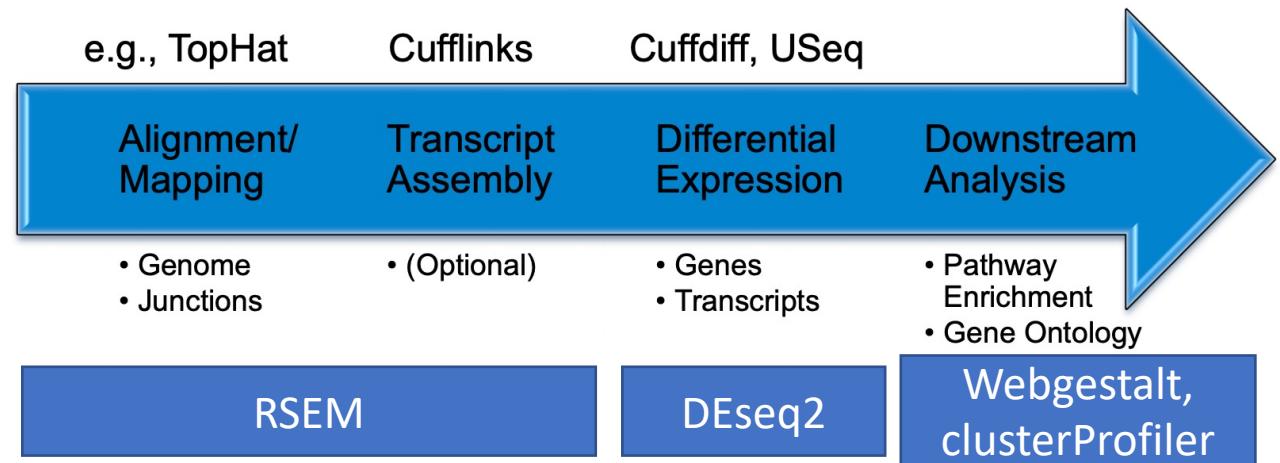
# Part II, Bulk RNA-seq

# Part II, RNA-seq analysis



# Steps

- Quality filtering
- Cut adaptors
- Alignment
- Differential expression



**TopHat**  
A spliced read mapper

For Whole Genome



Index files for TopHat:

```
mm10.1.bt2
mm10.2.bt2
mm10.3.bt2
mm10.4.bt2
mm10.rev.1.bt2
mm10.rev.2.bt2
```

Output file:  
Alignments in BAM format  
only for genomic coordinates

**RSEM**  
Transcript alignment and  
quantification

For Only Transcriptome



Index files for RSEM:

```
mm10.rsem.1.bt2
mm10.rsem.2.bt2
mm10.rsem.3.bt2
mm10.rsem.4.bt2
mm10.rsem.grp
mm10.rsem.idx.fa
mm10.rsem.rev.1.bt2
mm10.rsem.rev.2.bt2
mm10.rsem.seq
mm10.rsem.ti
mm10.rsem.transcripts.fa
```

Output files:  
Alignments in BAM format  
for both genomic coordinates  
and transcriptomic coordinates

Gene and isoform  
quantification results

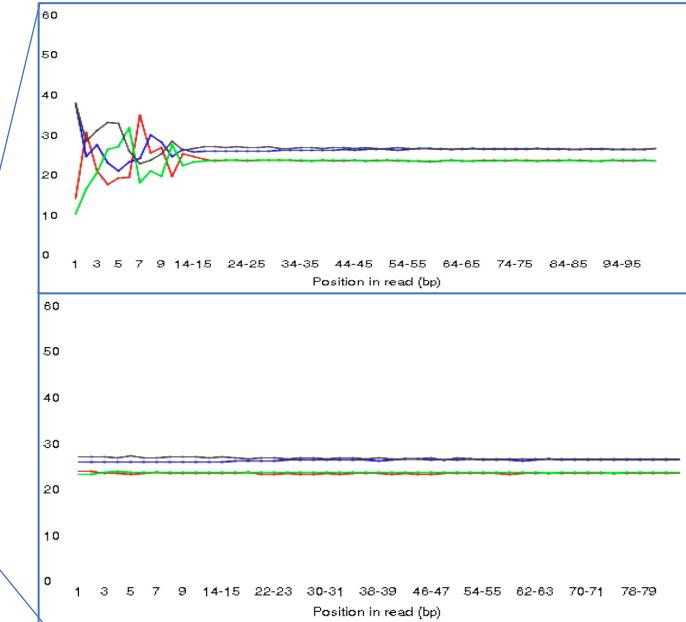
# snakemake Pipeline for alignment

- Pool fastq from different lanes.
- Filtered reads and Trimmed reads
  - fasttrimmer
  - Trimmomatic
- Alignment using RSEM using a UCSC gene annotation reference

snakemake  
pipeline

- Pooling counts
- Formatting metadata
- Differential expression DESeq2
- Functional annotation
  - GSEA
  - Enrichment analysis
  - Cellular decomposition
- Cell type deconvolution

Jupyter  
Notebooks  
on Biowulf



995Gbytes  
2016 files

Snakemake  
pipeline

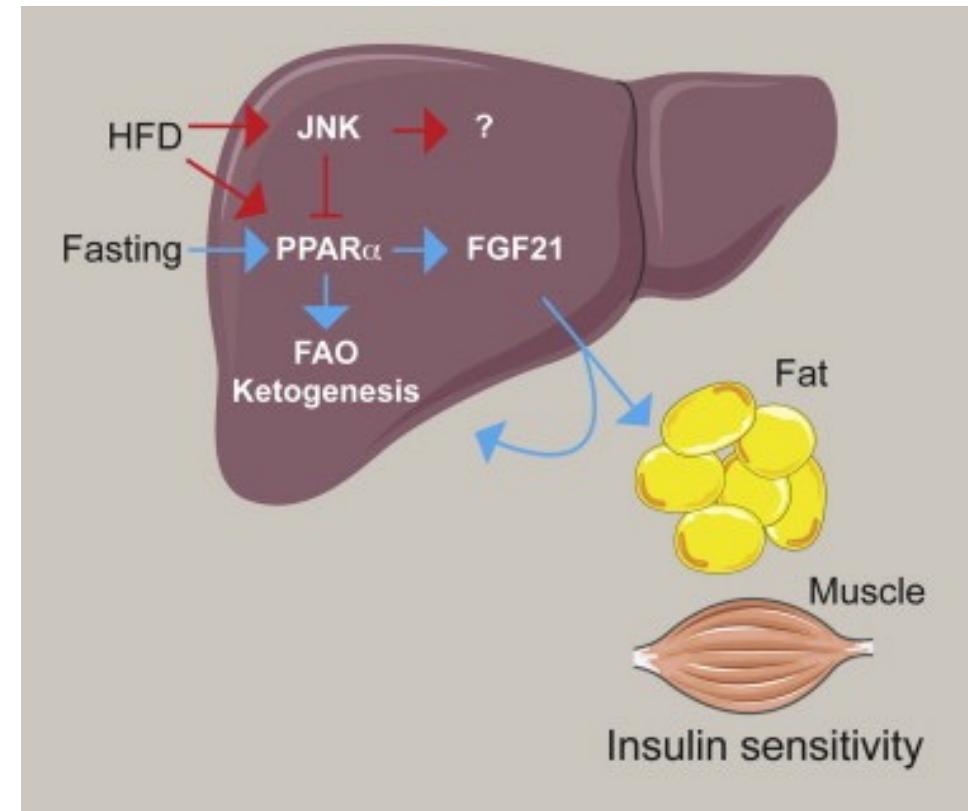
Alignment to  
UCSC reference transcriptome  
by RSEM

Collection of counts to form a count matrix  
And alignment statistics

145 files  
\*.genes.results

# Demo on RNA-seq: a small sample data

- Paired-end sequencing
- 6 samples
  - 3 wild type
  - 3 JNK1/2 double knockout



# DESeq2 models raw count with negative binomial distribution, not with normal/gaussian distribution

raw count for gene i, sample j

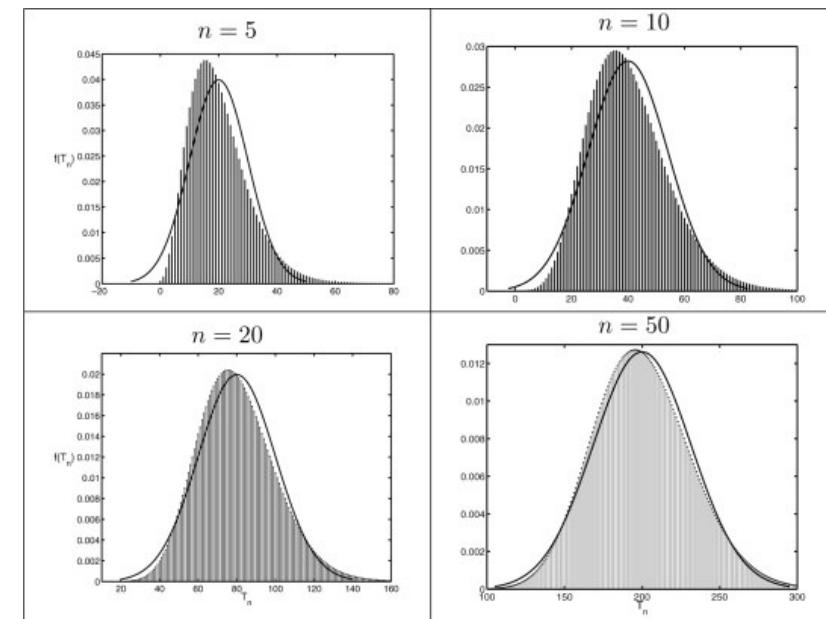
The mean is taken as “normalized counts” scaled by a normalization factor

one dispersion per gene

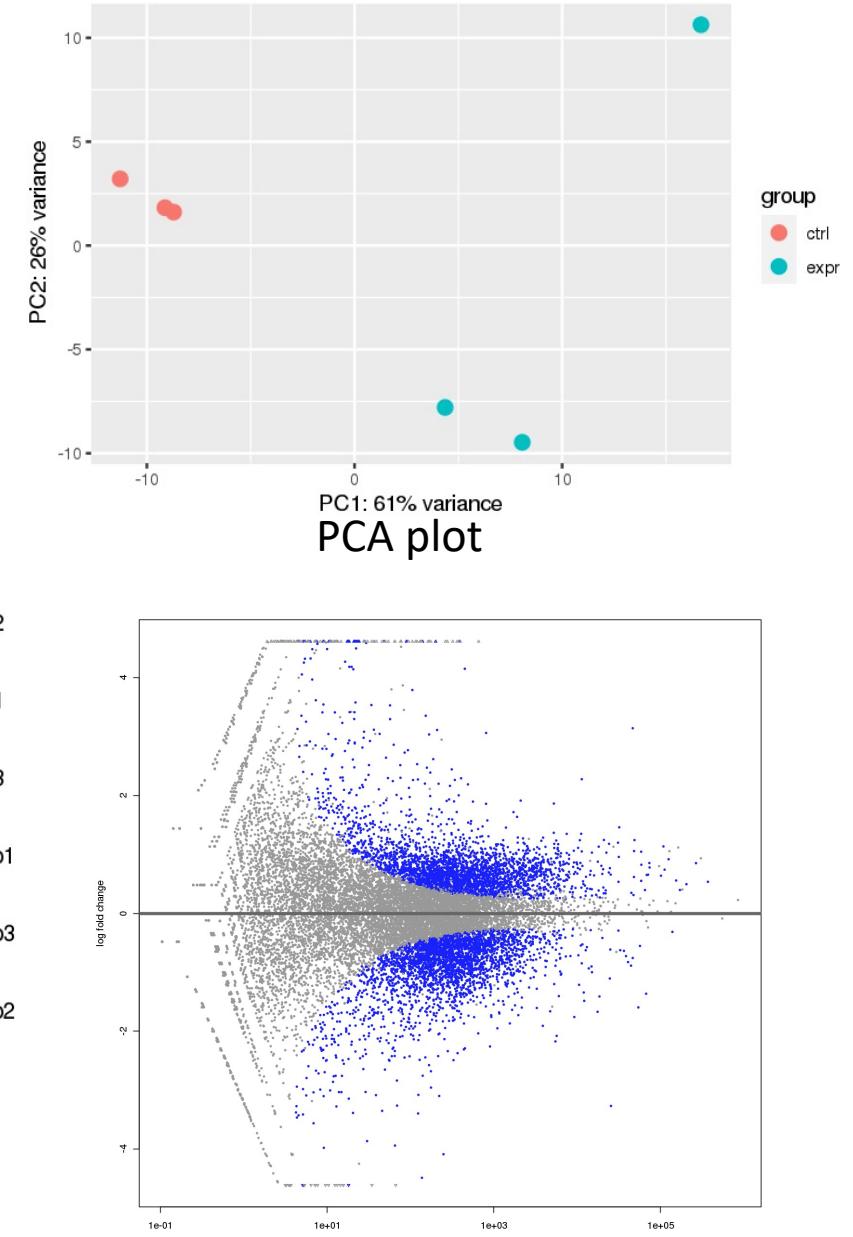
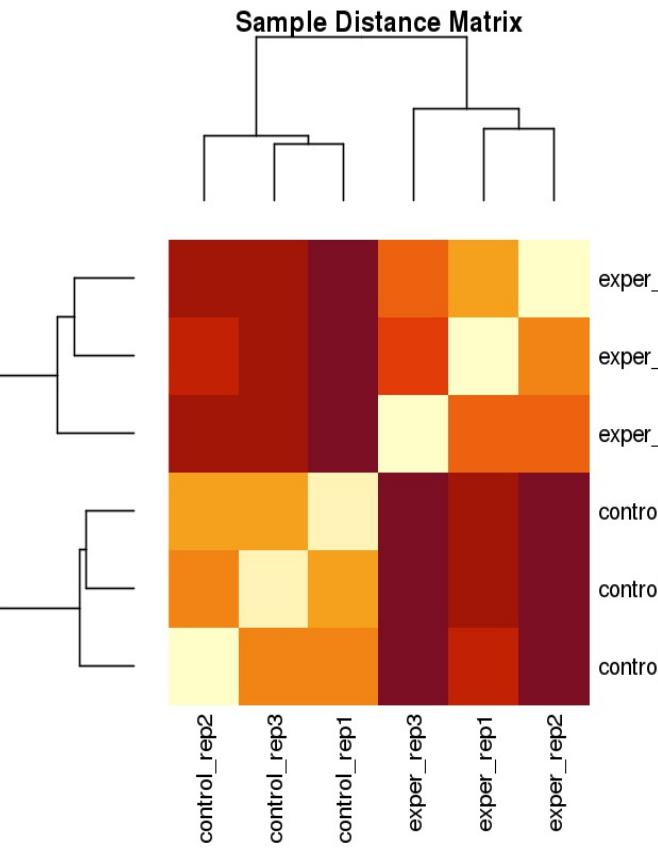
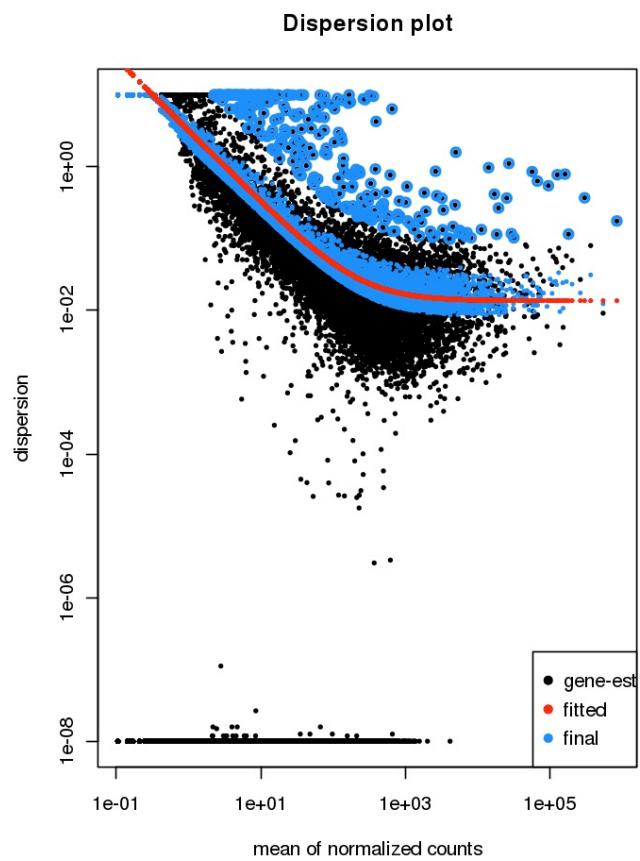
$$K_{ij} \sim NB(s_{ij}q_{ij}, \alpha_i)$$

Dispersion vs sd  
Wald test vs t test

Correction of p values for multiple testing,  
**FDR/Benjamini-Hochberg → Bonferroni**  
(loose → stringent)



# Plots



Dispersion plot

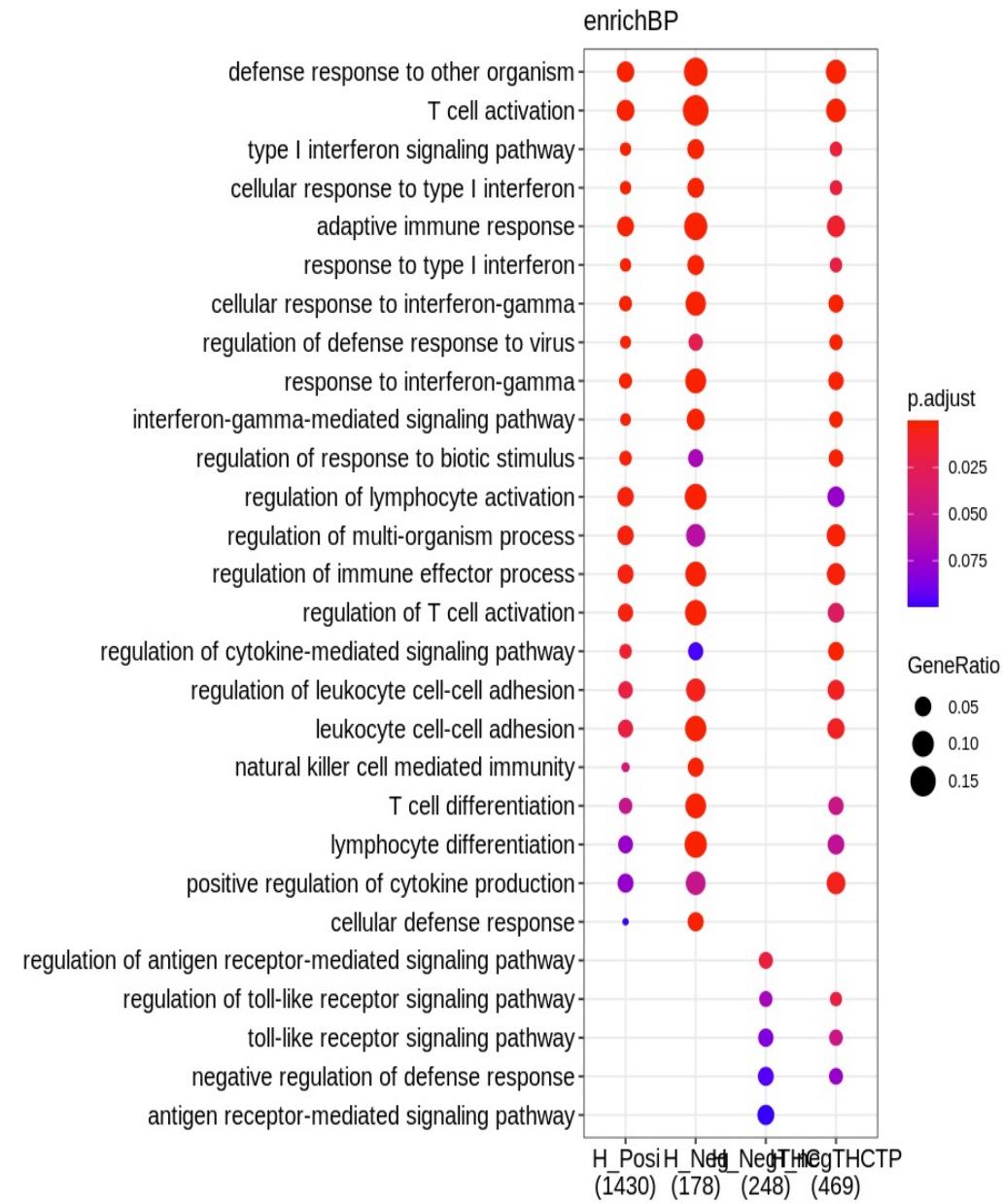
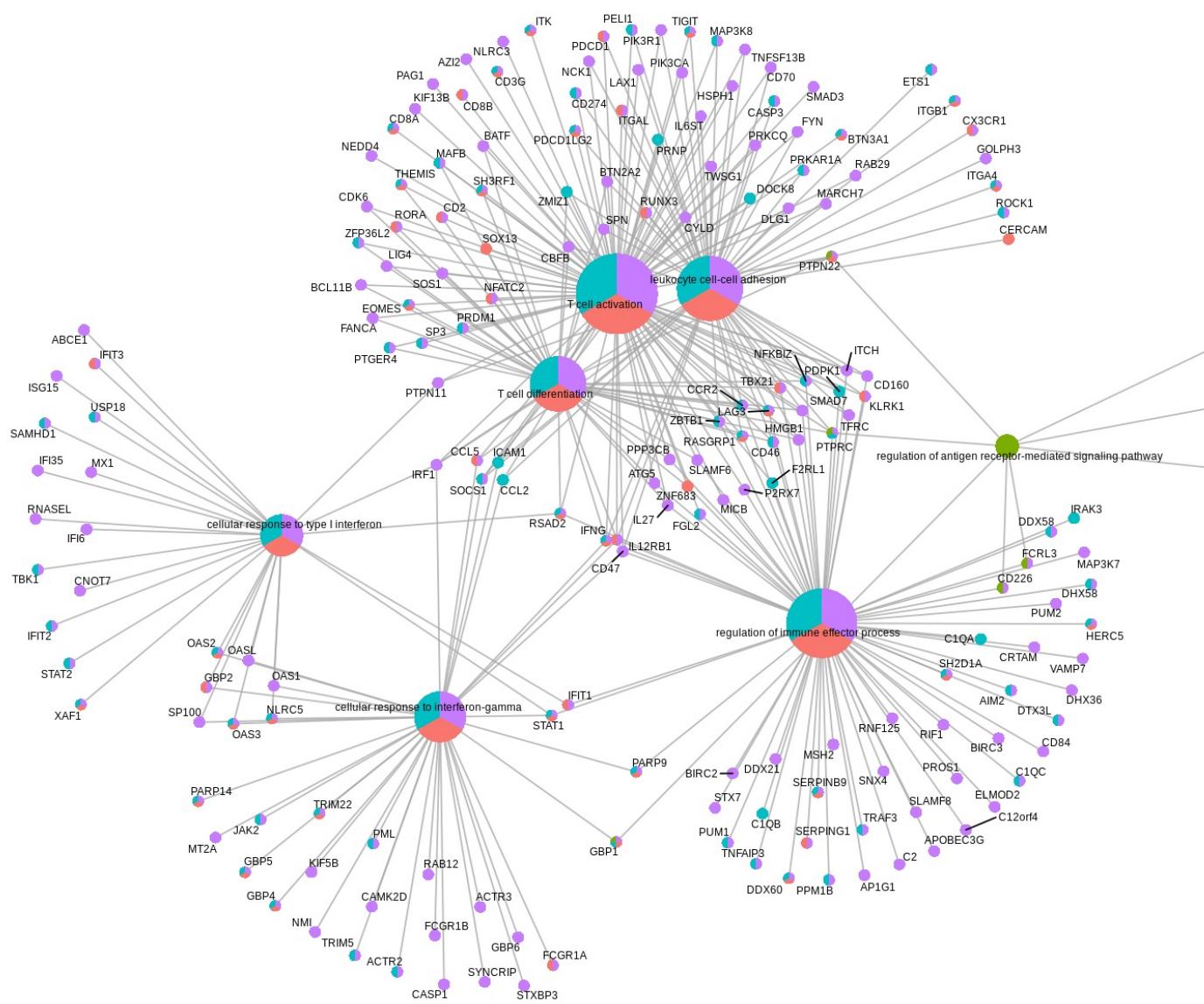
Correlation heatmap

MA plot

# Visualizing pathway—gene relationship

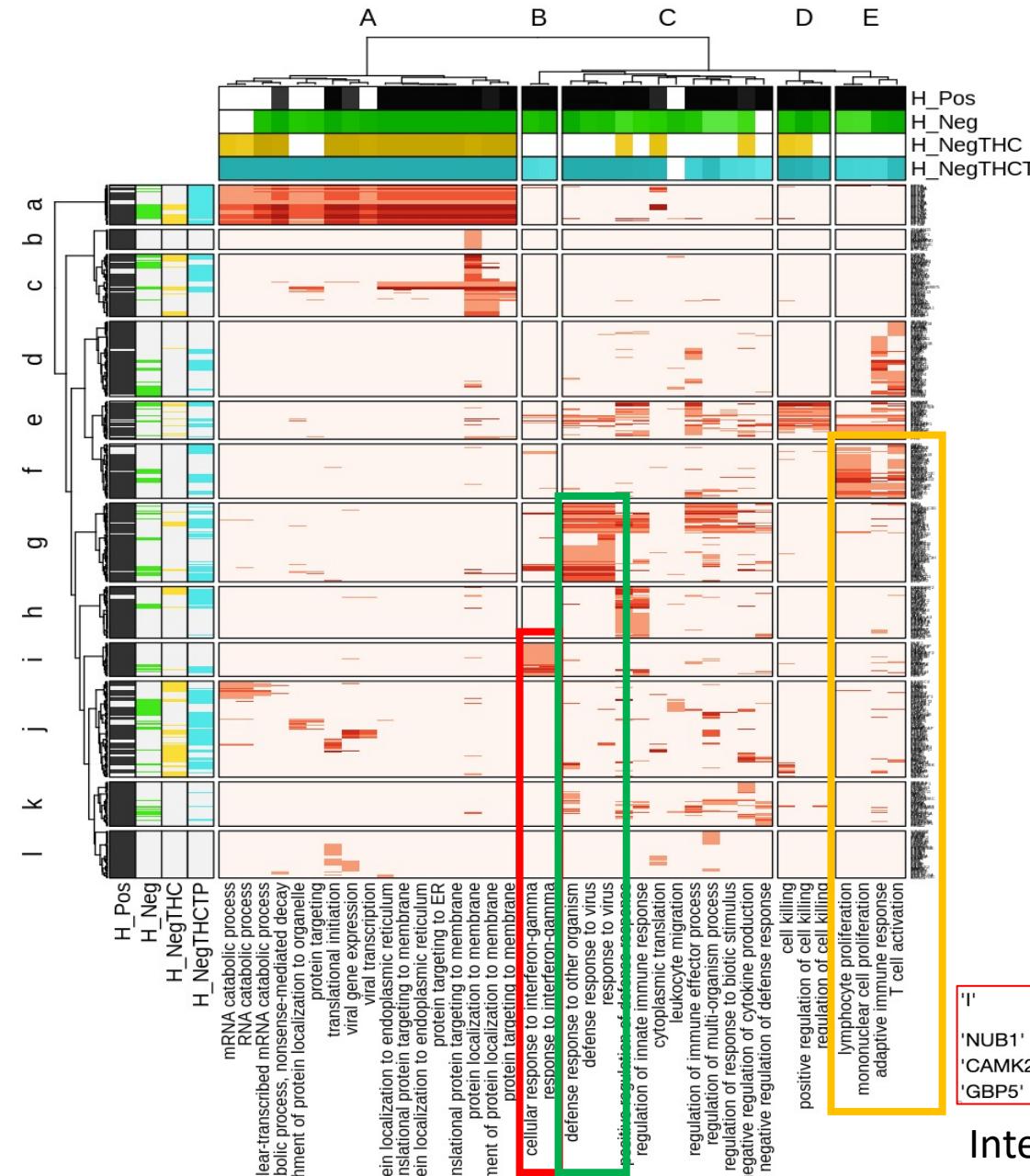
Example: selected GO-BP

Redundancy of pathway names



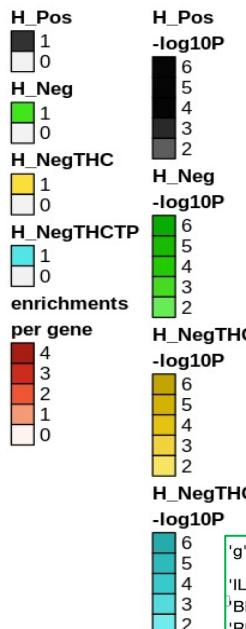
# Gene-Pathway incidence heatmap

Package:  
multienrichjam



## T cell activation, genes

'f'
ZMIZ1' 'SP3' 'MAFB' 'MAP3K8' 'LGALS3' 'PNP' 'DOCK8' 'MAD1L1' 'ICAM1' 'AIF1' 'CCL2' 'SLC39A10' 'IMPDH2' 'IKZF3' 'FLT3LG' 'CD320' 'CD38' 'CORO1A' 'TNFSF14' 'SOS1' 'MARCH7' 'FKBP1B' 'CD70' 'AZI2' 'CCND3' 'NCK1' 'BTN2A2' 'TWSG1' 'CD22' 'IMPDH1' 'IGFBP2' 'NFATC2' 'TNFRSF13C' 'HELLS' 'CASP3' 'PRKAR1A' 'CD274' 'BTN3A1' 'PDCD1LG2' 'SH3RF1' 'TMIGD2' 'CD79A' 'LST1' 'IL6ST' 'DLG1' 'TNFSF13B' 'ICOSLG' 'PRKCQ' 'SLC11A1' 'CD4' 'SPN' 'MIF' 'PRDX2' 'CCR2' 'CD40LG' 'CD46' 'CD74' 'ZP3' 'RAC2' 'TFRC'



## Response to virus, genes

'g'
IL27' 'FGL2' 'MICB' 'IFIT1' 'DTX3L' 'APOBEC3G' 'ELMOD2' 'PCBP2' 'HERC5' 'PPM1B' 'TRIM5' 'PARP9' 'STAT1' 'PUM2' 'PQBP1' 'BIRC2' 'BIRC3' 'RSAD2' 'PUM1' 'TRAF3' 'AIM2' 'DDX60' 'IRAK3' 'IFIH1' 'PLSCR1' 'TNFAIP3' 'DDX58' 'DHX58' 'SEC14L1' 'NLRX1' 'RNF125' 'TBK1' 'ITCH' 'PYCARD' 'ODC1' 'IFIT2' 'CREBZF' 'EEF1G' 'IFITM2' 'IFI44' 'IFIT3' 'URI1' 'HMGA1' 'ACTA2' 'CCDC130' 'BANF1' 'CFL1' 'RNASE6' 'IFI6' 'G3BP1' 'EXOSC5' 'DDX17' 'EXOC1' 'DDX21' 'DHX36' 'LYST' 'UNC13D' 'ISG15' 'ZC3HAV1' 'MX1' 'APOBEC3H' 'FAM111A' 'ABCE1' 'RNASEL' 'SAMHD1' 'CNOT7' 'IFI16' 'PML' 'OAS1' 'OASL' 'OAS3' 'OAS2' 'TRIM22' 'GBP1' 'NLRC5' 'BNIP3L' 'FLNA' 'RTP4' 'CXCL10' 'GBP3' 'IFI44L' 'ZMYND11' 'STAT2' 'NT5C3A' 'SLFN11' 'EIF2AK2' 'IFIT5'

'I'

NUB1' 'ZYX' 'VIM' 'SYNCRIP' 'STXBP3' 'SP100' 'RAB43' 'RAB12' 'MT2A' 'DAPK3' 'CDC42EP2' 'CASP1' 'CAMK2G' 'ACTR3' 'CAMK2D' 'FCGR1B' 'KIF5B' 'TRIM8' 'GBP6' 'IRF8' 'CD47' 'WAS' 'PTAFR' 'STXBP2' 'FCGR1A' 'GBP2' 'GCH1' 'ACTR2' 'GBP4' 'GBP5' 'JAK2' 'PARP14' 'SOCS1' 'NMI' 'NR1H2' 'CDC37' 'HCK'

Interferon response is a distinguishing pathway, and here are the genes

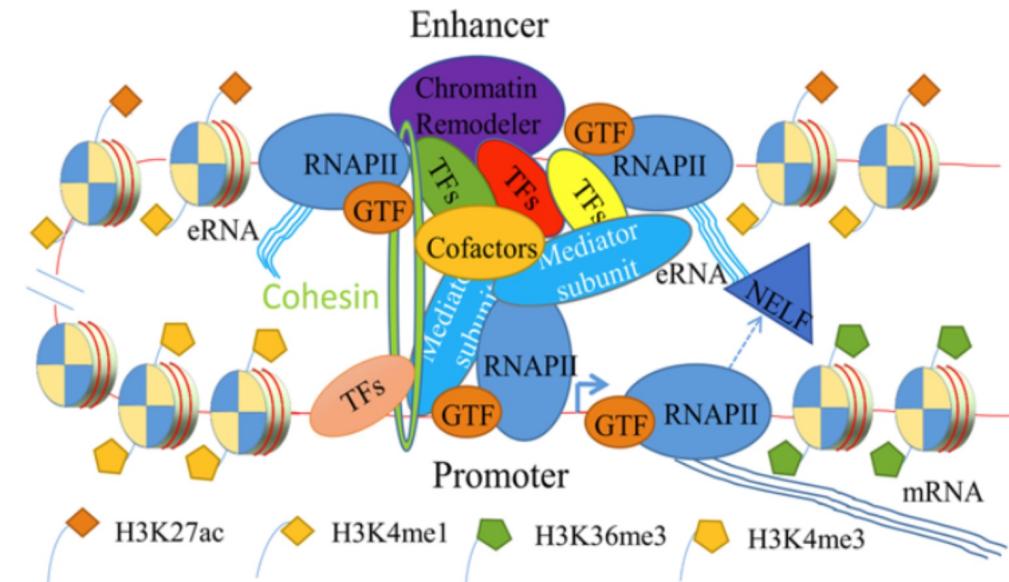
# Reference

- Bulk RNA-seq
  - The pipeline and dataset 1
    - <https://github.com/UMMS-Biocore/RNASeqTutorial/blob/master/RNASeqTutorial.pdf>
    - Dataset 2
      - <https://bioinfo.umassmed.edu/index.php?p=35#p1e3>
  - Functional annotation of gene lists.
    - Webgestalt, <http://www.webgestalt.org/2017/option.php>
    - clusterProfiler
      - Overrepresentation Enrichment analysis (OEA)
      - Geneset Enrichment Analysis (GSEA)
    - Ingenuity Pathway Analysis (**IPA**)
    - multienrichjam, <https://jmw86069.github.io/multienrichjam/articles/importIPA.html>
    - Network analysis
      - Cytoscape and apps <https://cytoscape.org/>

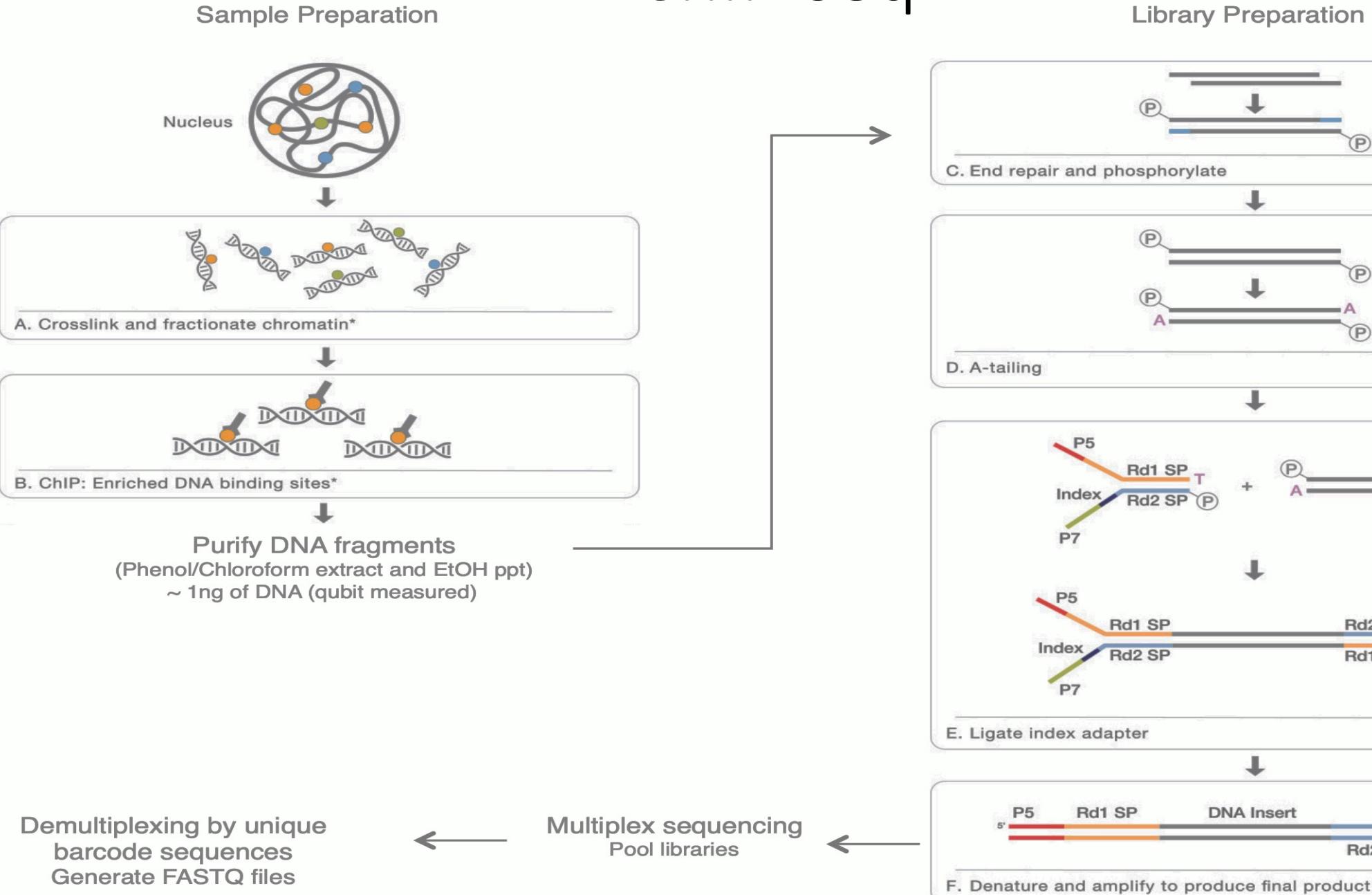
# Part III, Chromatin Immuno Precipitation Sequencing (ChIP-seq)

- Background and related chromatin types
- Peaks and input Controls
- Quality controls
- Workflow and sample data
- Demo on BioWulf

Chromatin = DNA and associated proteins

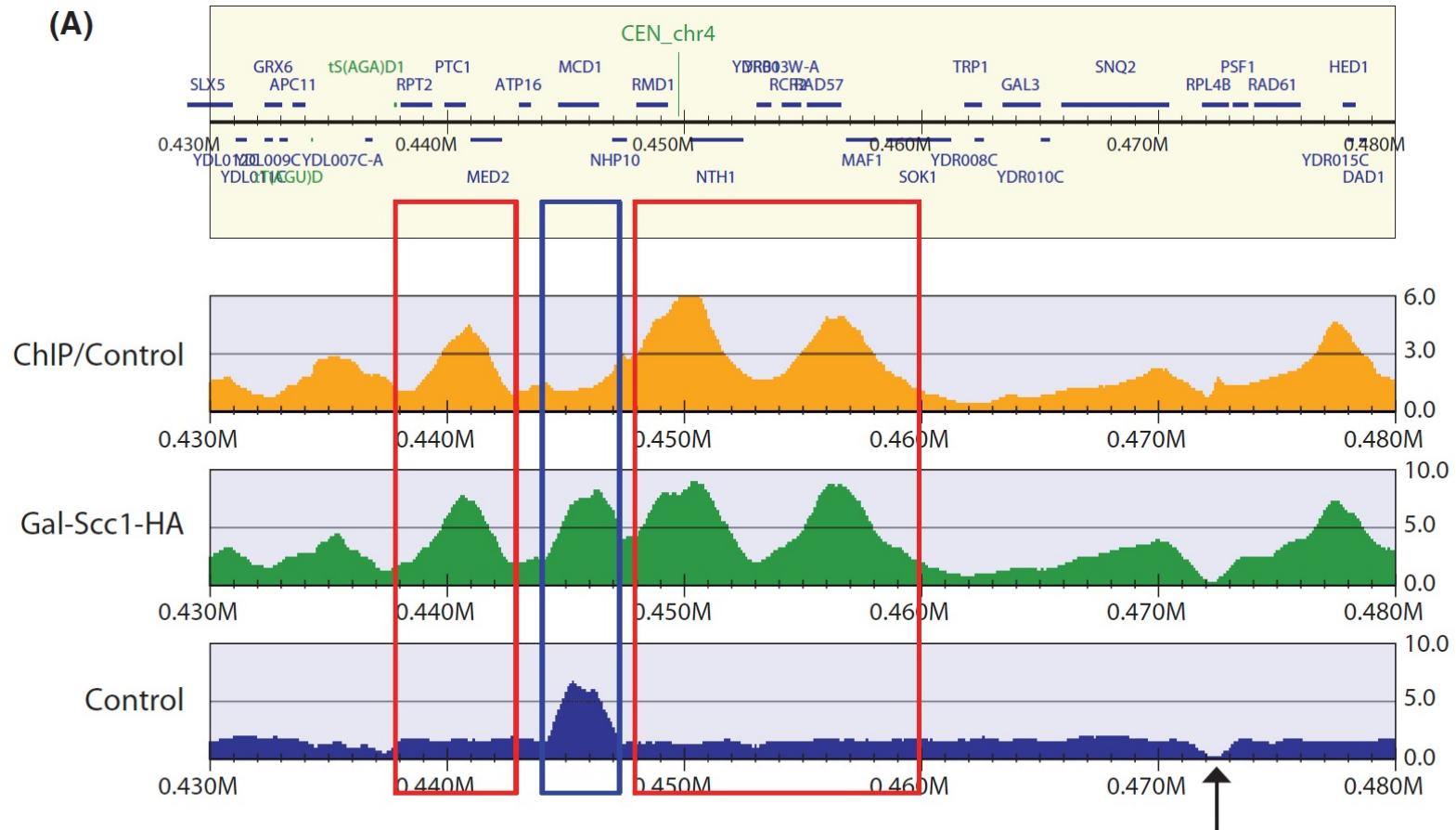


# ChIP-seq

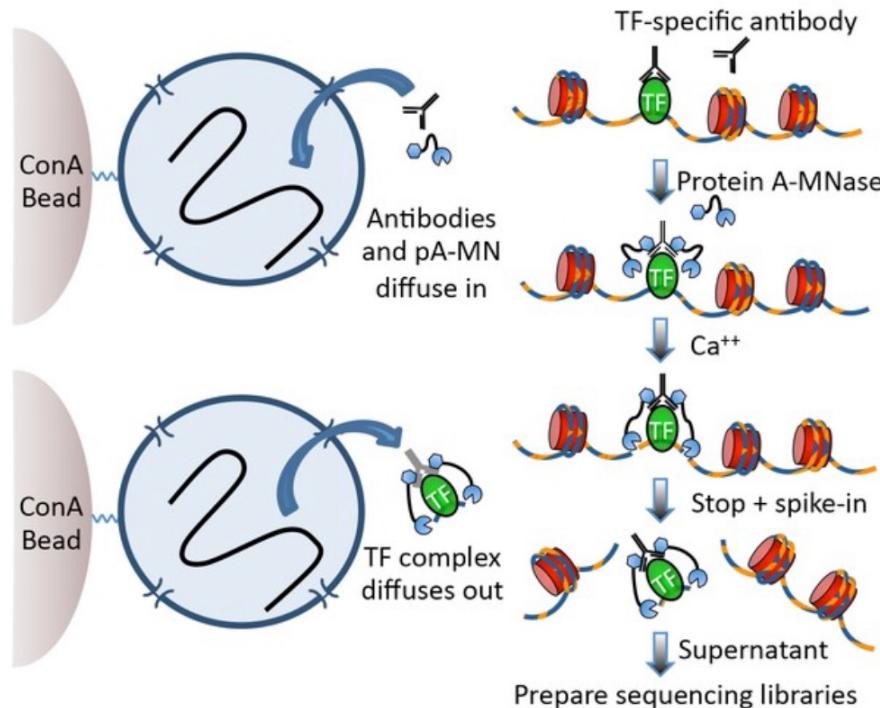


# Peaks are significant enrichments over Input DNA

R Nakato et al.



## Cut&Run - alternative



### Benefits:

- No crosslinking step
- Removes background
- 1/10<sup>th</sup> sequencing depth
- Small numbers of cells  
(100 – 1000s)

Skene & Henikoff. eLife (2017)

# Different IP produce different type of peaks

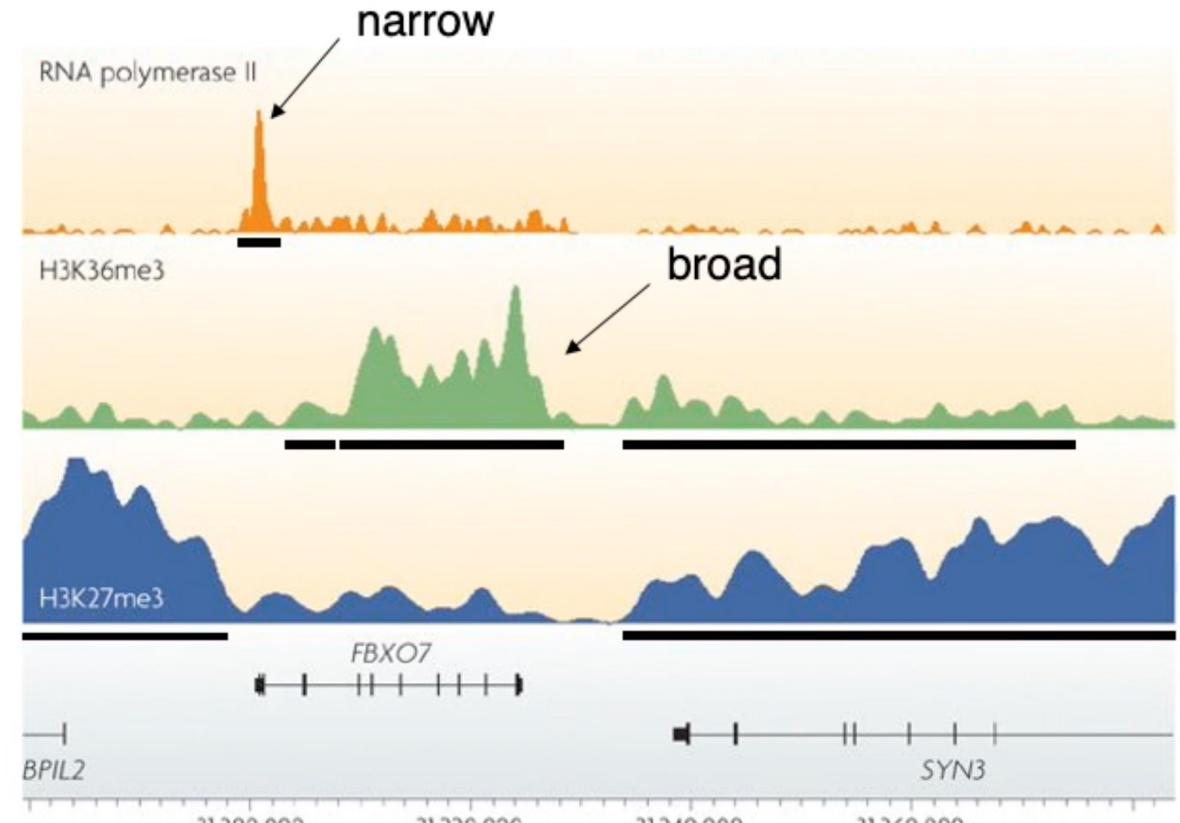
3 Types of peaks:

1. Sharp & narrow (100s bp)

(eg. site-specific TF)

2. Broader but defined (kb)

3. Very broad (regional, 1000s kb)  
(eg. heterochromatin histone marks)

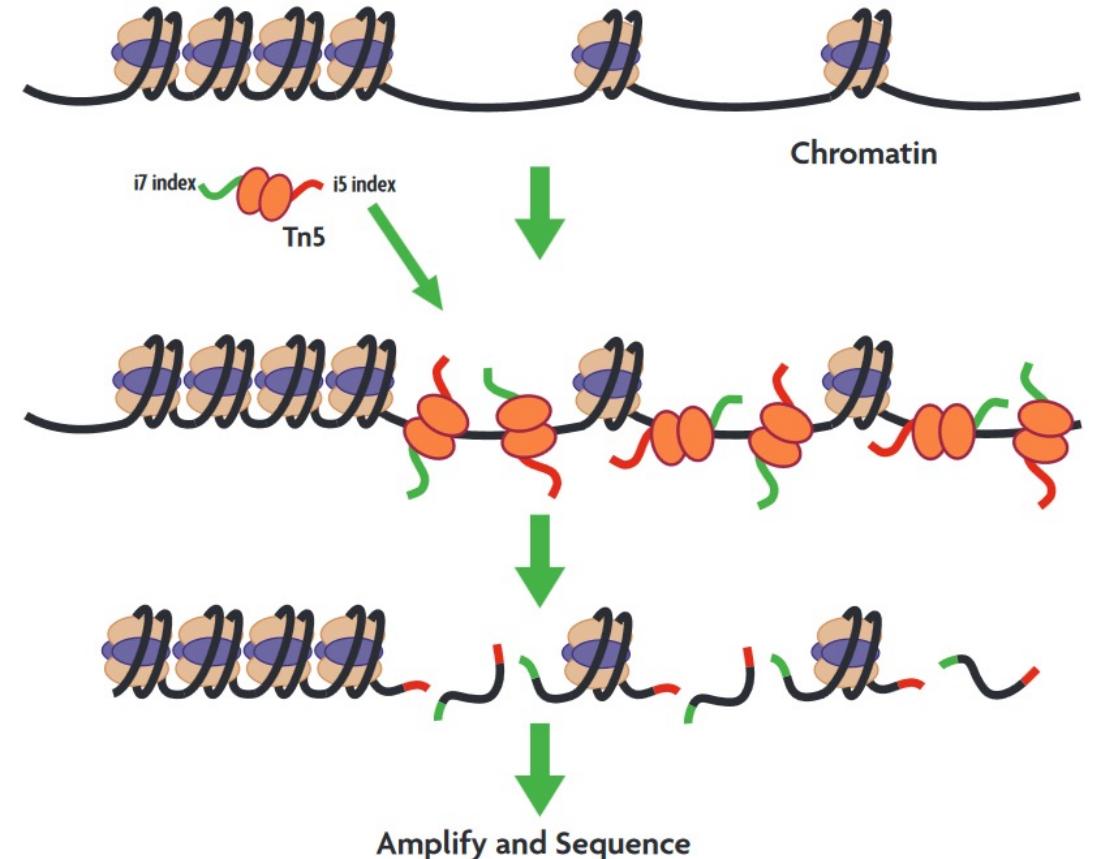


Nature Reviews | Genetics

Histone Mark	Type
H3K27me3	broad
H3K36me3	broad
H3K4me3	narrow
H3K27ac	narrow

# ATAC-seq is very similar but has no input control

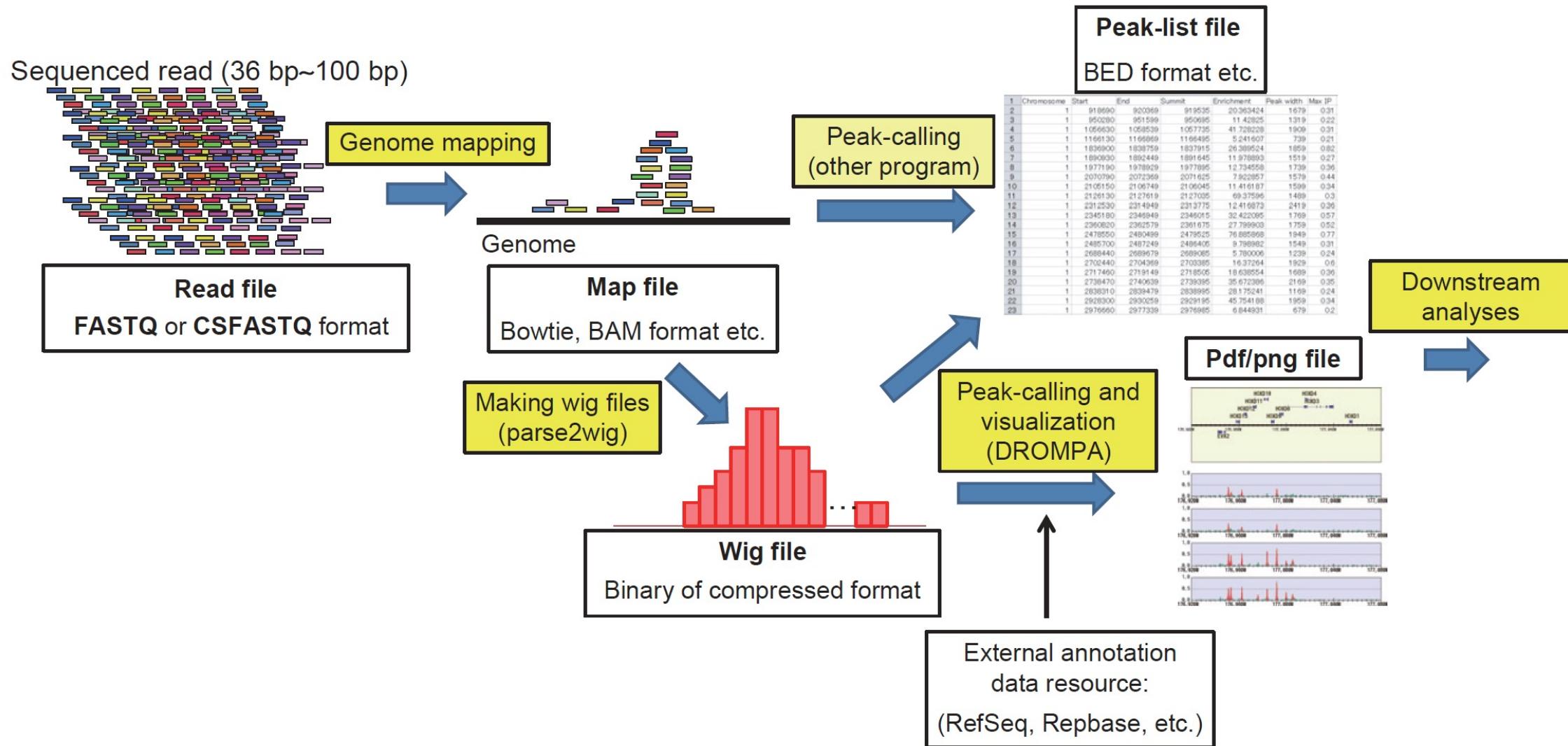
- ATAC-seq assess regions of genome that is not compacted
- Processing of ATAC-seq is very similar to ChIP-seq
- There is no input control for ATAC-seq.



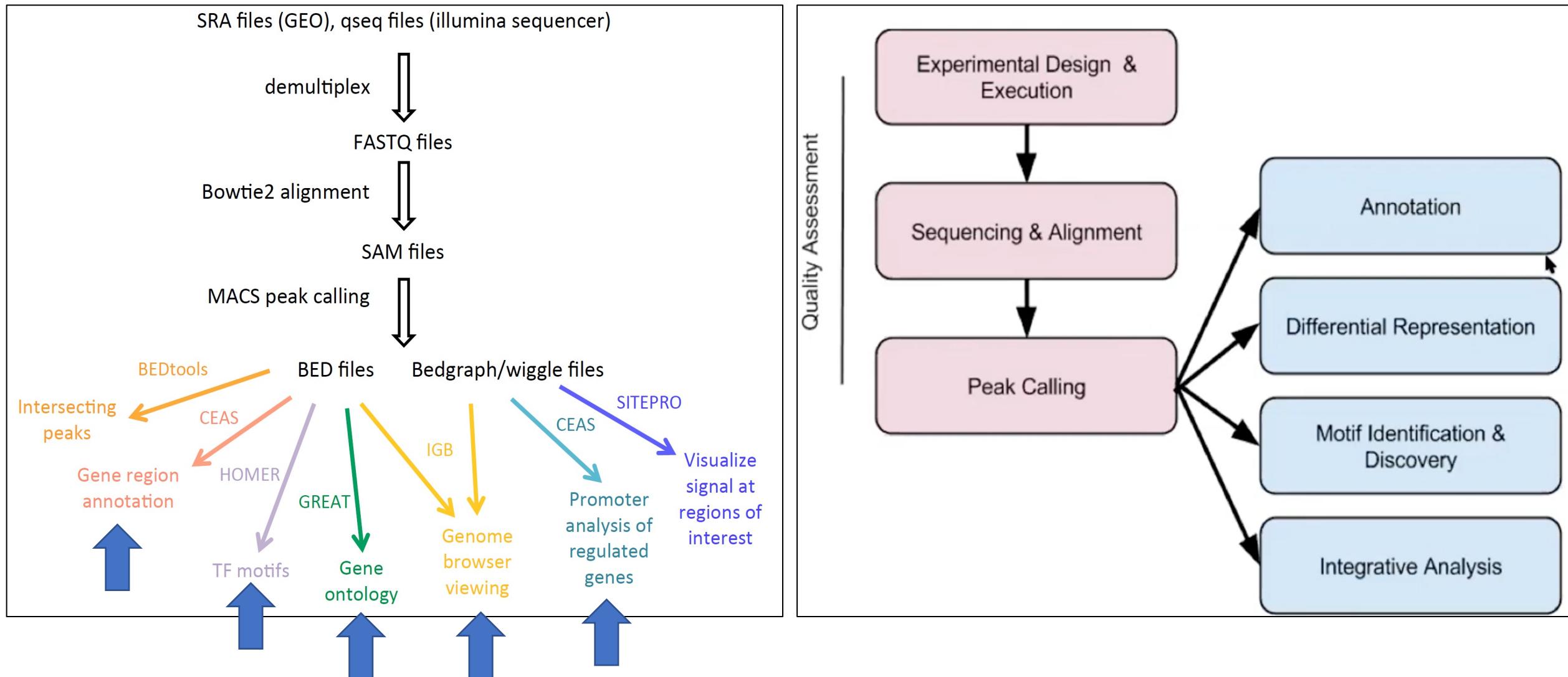
# ChIP-Seq analysis pipeline example

R Nakato *et al.*

<https://m.ensembl.org/info/website/upload/bed.html>



# Part II, ChIP-seq analysis

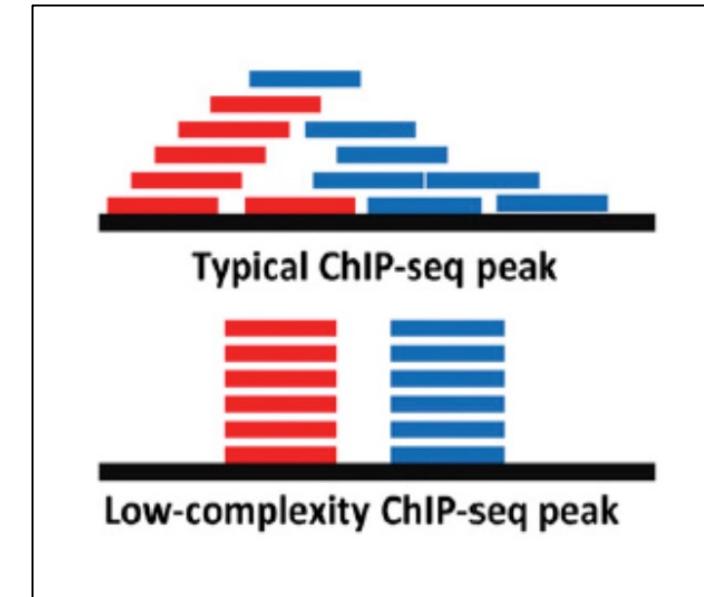
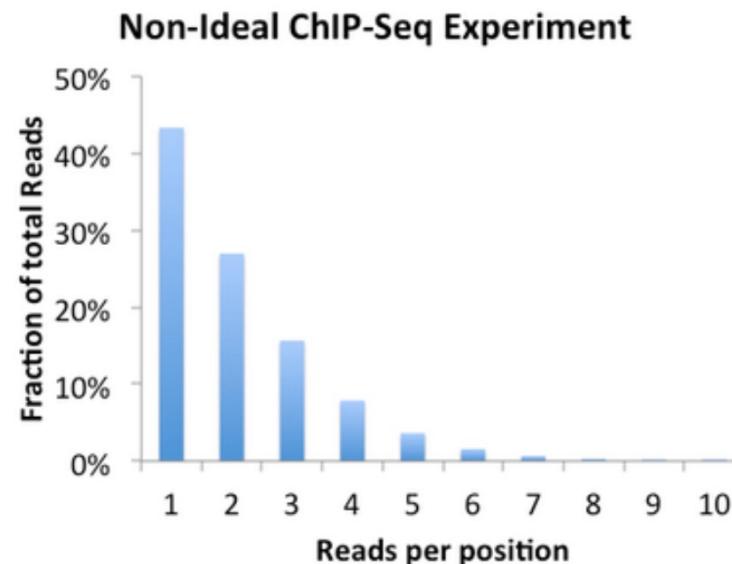
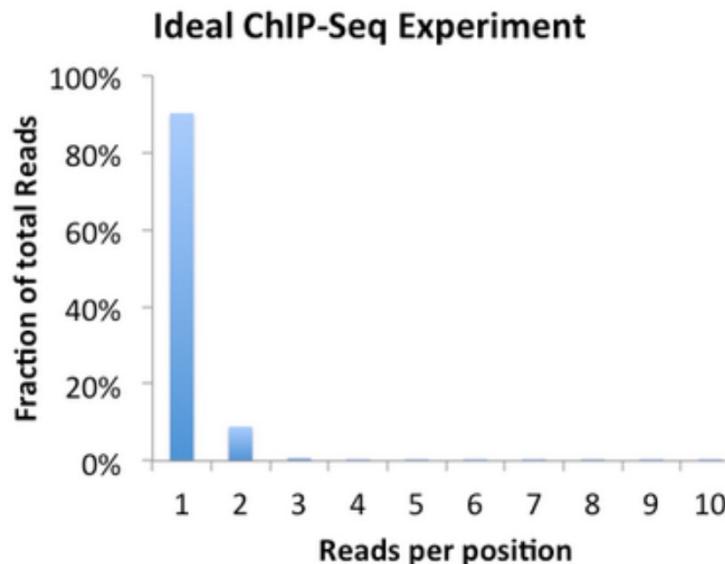


# Complexity of library

## Clonal Tag Distribution

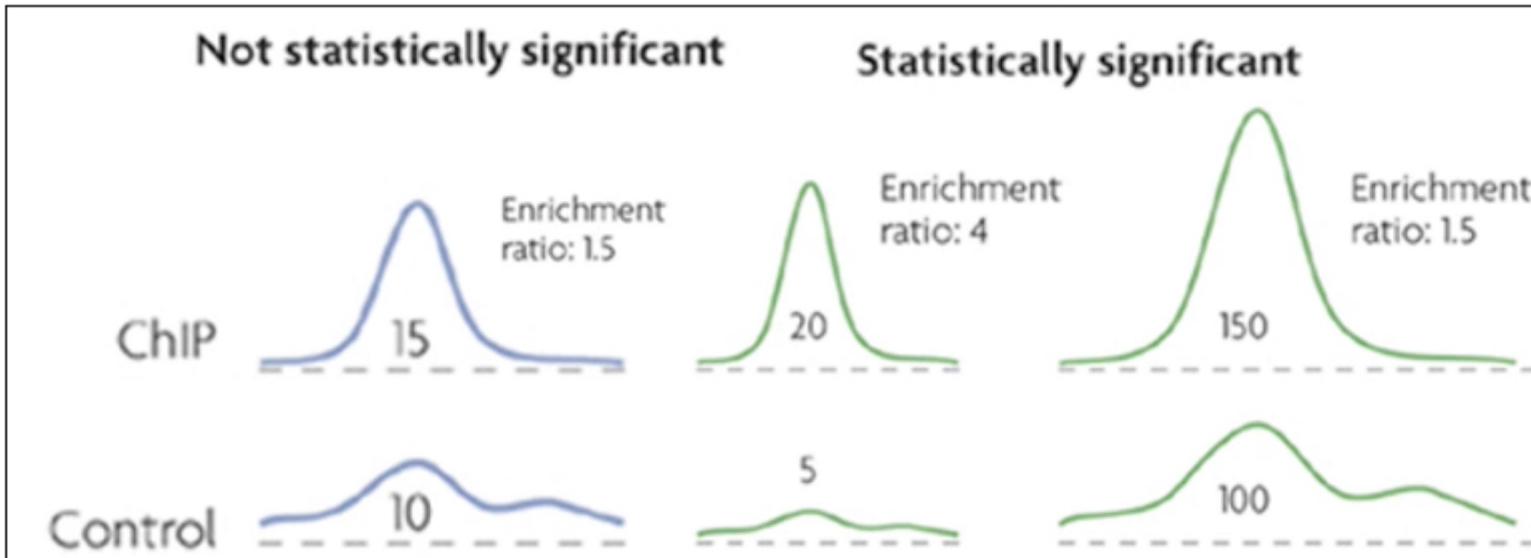
```
[zemke@n7420 flashscratch]$ bowtie2 -q -p 4 -k 1 --local  
mso_k18_chip.fastq > dms0_k18_chip.sam  
42714660 reads; of these:  
    42714660 (100.00%) were unpaired; of these:  
        980458 (2.30%) aligned 0 times  
        41734202 (97.70%) aligned exactly 1 time  
        0 (0.00%) aligned >1 times  
97.70% overall alignment rate
```

**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.



If the experiment is highly clonal and not expected to be, it might help to clean up the downstream analysis by forcing tag counts at each position to be no greater than  $x$ , where  $x$  is usually 1. To do this rerun the makeTagDirectory command and add "-tbp <#>" where # is the maximum tags per bp.

# Peak calling: detect regions of enrichment



**Goal:** Transform read counts into **normalized intensity signal**

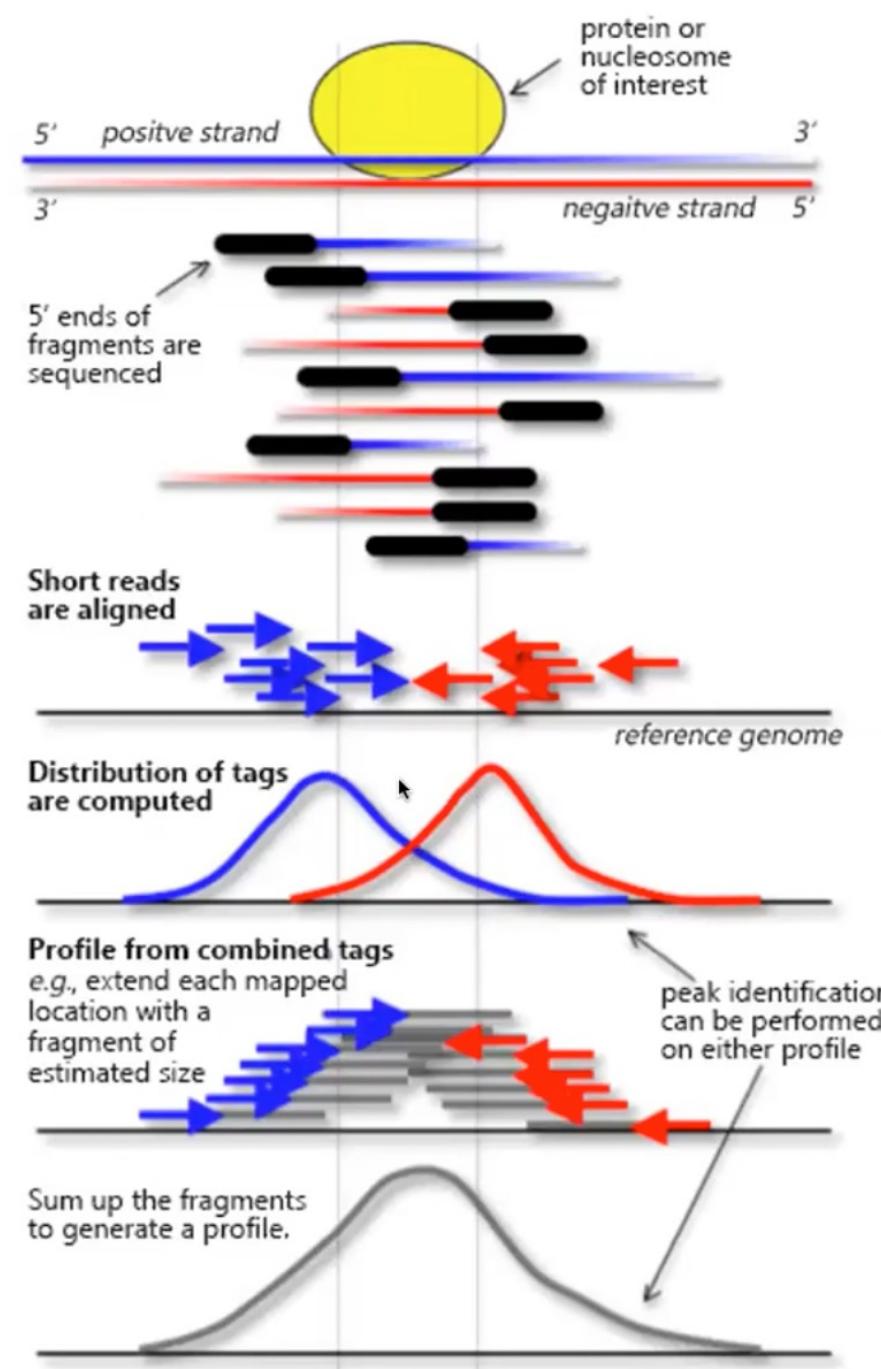
**Steps:**

1. Estimate fragment-length  $f$  using strand cross-correlation analysis
2. Extend each read from 5' to 3' direction to fragment length  $f$
3. Sum intensity for each base in 'extended reads' from both strands
4. Perform same operation on input-DNA control data (correct for sequencing depth differences)
5. Compute p-value based on local-expectation of # reads based on control samples (local-Poisson)

# MACS

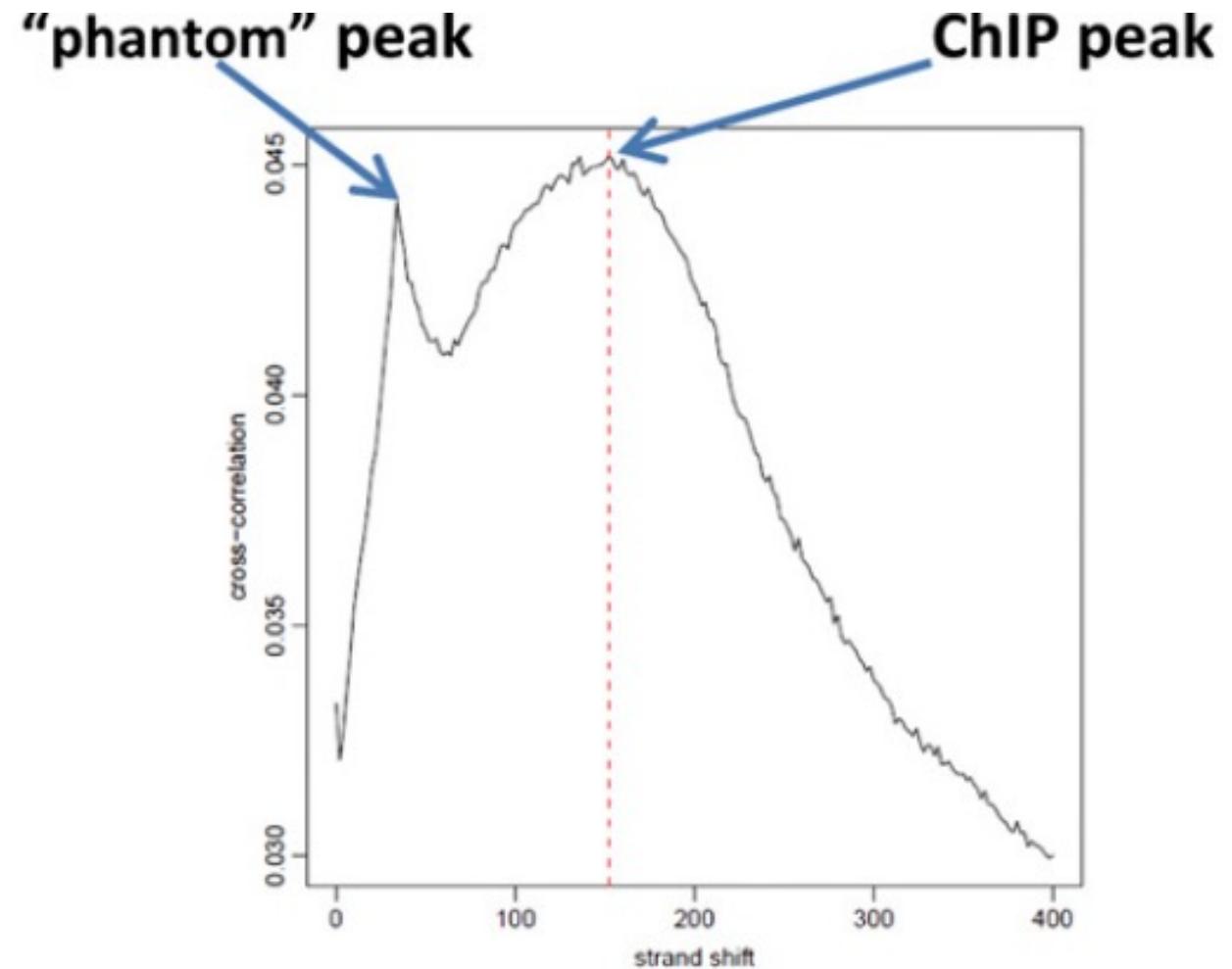
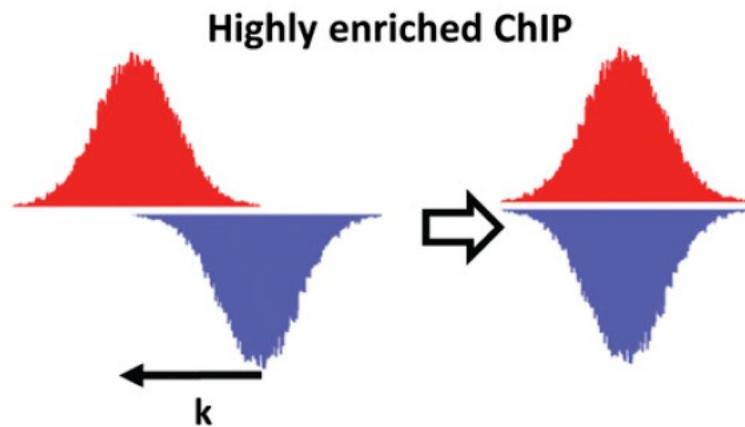
- Model-based Analysis for ChIP-Seq
- Use confident peaks with many read pile up to model shift size

Model is for single end sequencing.  
For paired end, modeling step is ignored.



# Quality controls

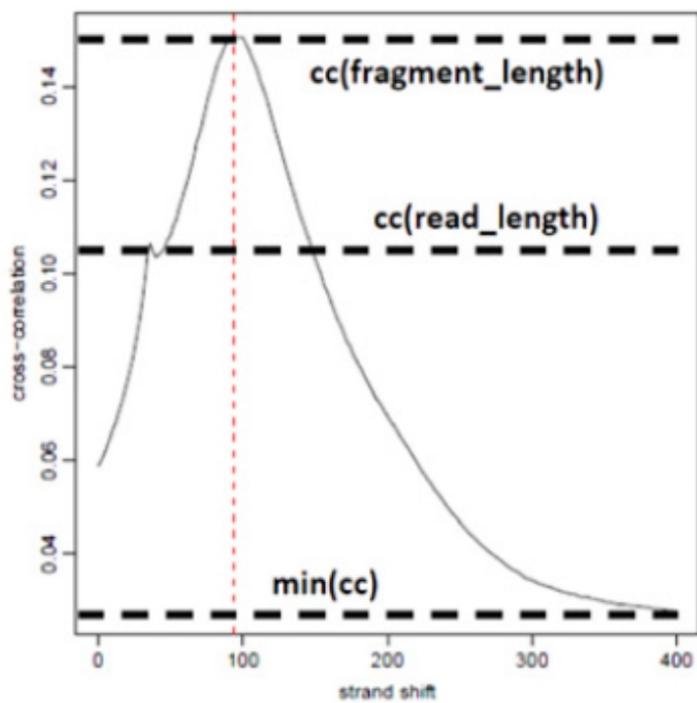
- If DNA is over-fragmented, the phantom peak will be high
- If fragmentation is appropriate, ChIP peak will dominate the strand shift.



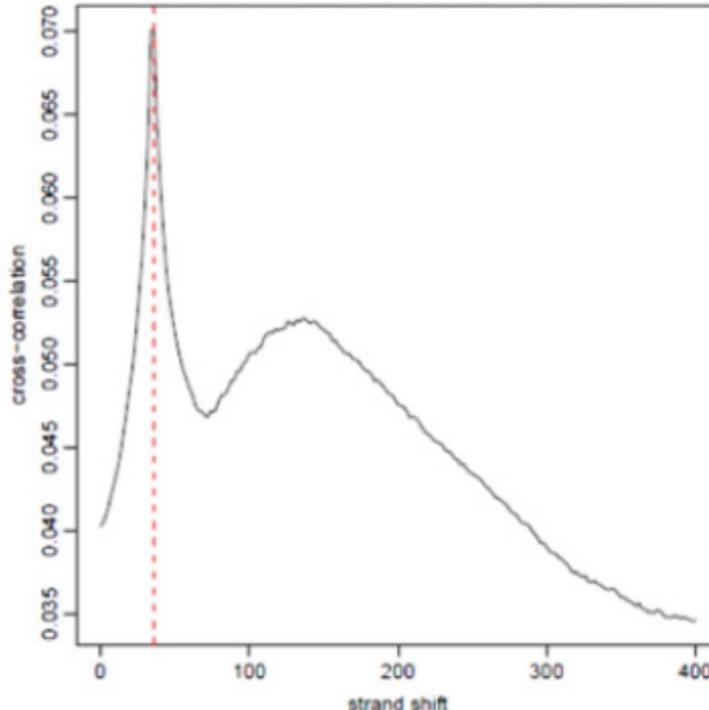
# Quality Control

## Cross-correlation

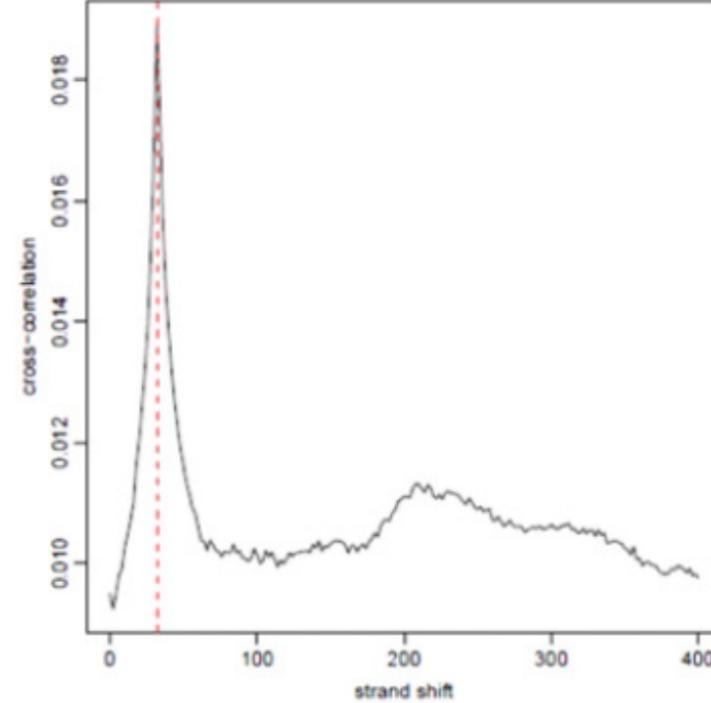
Successful



Marginal



Failed



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

Normalized strand coefficient

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

Relative strand coefficient

Landt et al. Genome Research (2012)

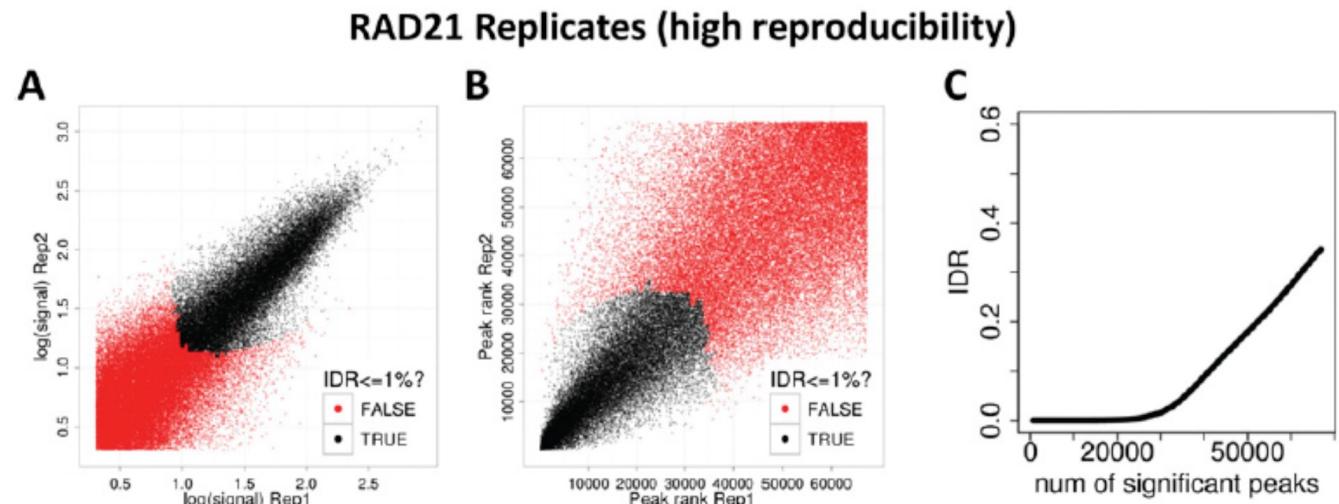
ENCODE repeats experiments with NSC values <1.05 and RSC values <0.8

# Quality control between replicates

- High variability between replicates
- Used correlation between ranking of peaks of each replicate
- More useful for replicates of 2, difficult to implement for bigger numbers of replicates.

Irreproducibility Discovery Rate (**IDR**)

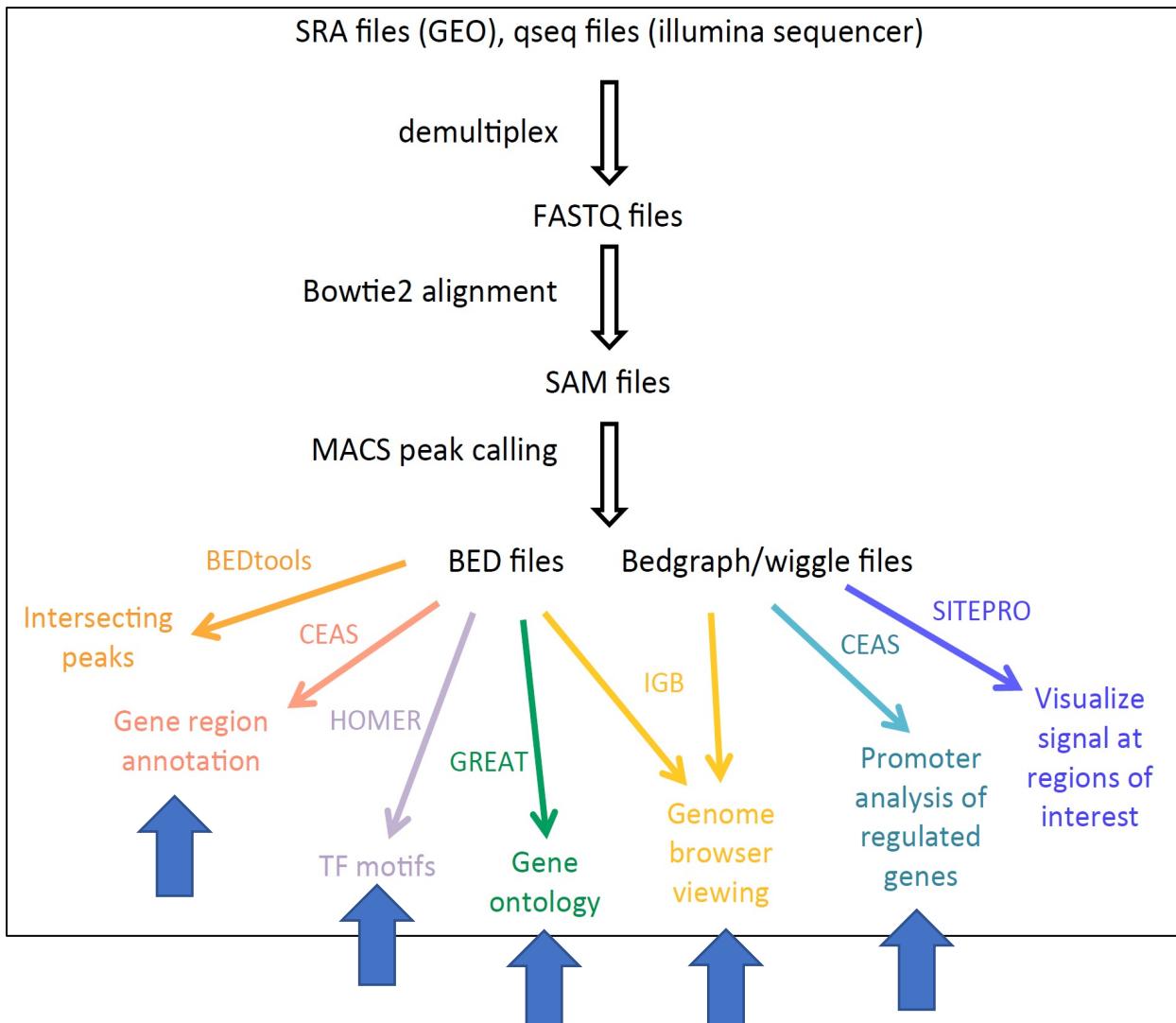
Useful for determining reproducibility of peaks in replicates



IDR < 1% = high confidence peak list

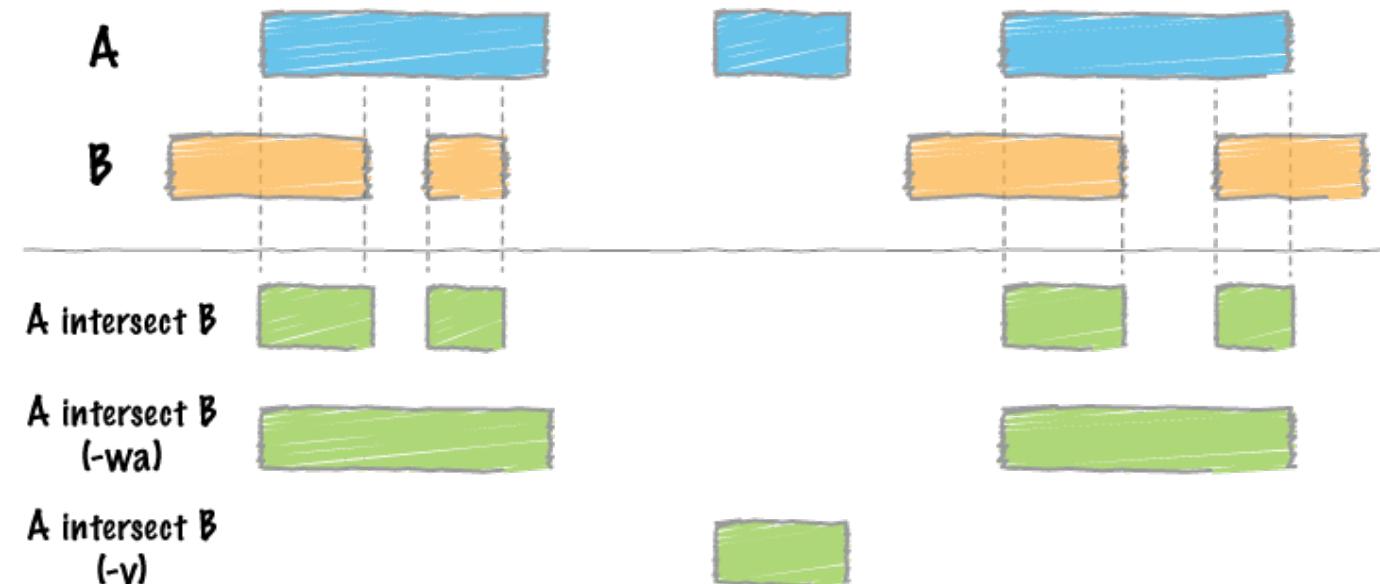
# Downstream analysis

- Peak (bed file) operation using bedtools
- Visualization of peaks using IGM
- Annotate peaks with adjacent genes using CEAS
- Annotate peak locations relative to promoter and Gene ontology using GREAT
- Find enriched motifs from peaks using Homer



# Bed file operations

- Bedtools intersect -wa s1.bed s2.bed
- Bedtools intersect -v s1.bed s2.bed



## Bed file format

### Required fields

The first three fields in each feature line are required:

1. **chrom** - name of the chromosome or scaffold. Any valid seq\_region\_name can be used, and can be a reference genome name (e.g. chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrX, chrY, chrM).
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base pair).
3. **chromEnd** - End position of the feature in standard chromosomal coordinates

```
chr1 213941196 213942363
chr1 213942363 213943530
chr1 213943530 213944697
chr2 158364697 158365864
chr2 158365864 158367031
chr3 127477031 127478198
chr3 127478198 127479365
chr3 127479365 127480532
chr3 127480532 127481699
```

### Optional fields

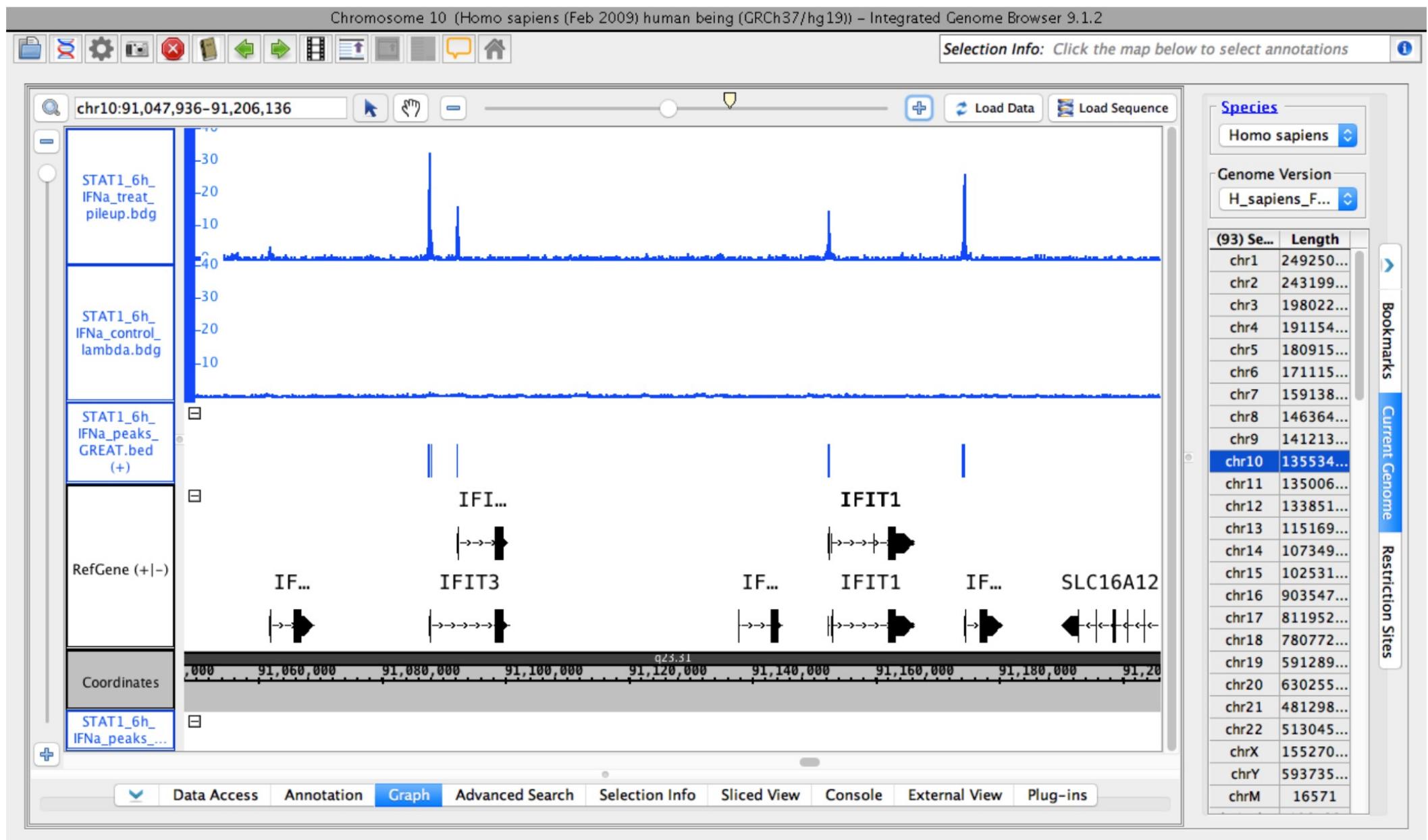
Nine additional fields are optional. Note that columns cannot be empty - lower-numbered fields must have values.

4. **name** - Label to be displayed under the feature, if turned on in "Configure this page".
5. **score** - A score between 0 and 1000. See [track lines](#), below, for ways to configure the display.
6. **strand** - defined as + (forward) or - (reverse).
7. **thickStart** - coordinate at which to start drawing the feature as a solid rectangle
8. **thickEnd** - coordinate at which to stop drawing the feature as a solid rectangle
9. **itemRgb** - an RGB colour value (e.g. 0,0,255). Only used if there is a track line with the value.
10. **blockCount** - the number of sub-elements (e.g. exons) within the feature
11. **blockSizes** - the size of these sub-elements
12. **blockStarts** - the start coordinate of each sub-element

```
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```

# Genome Browser Viewing

## .bdg, .bam, .bigwig, .bed



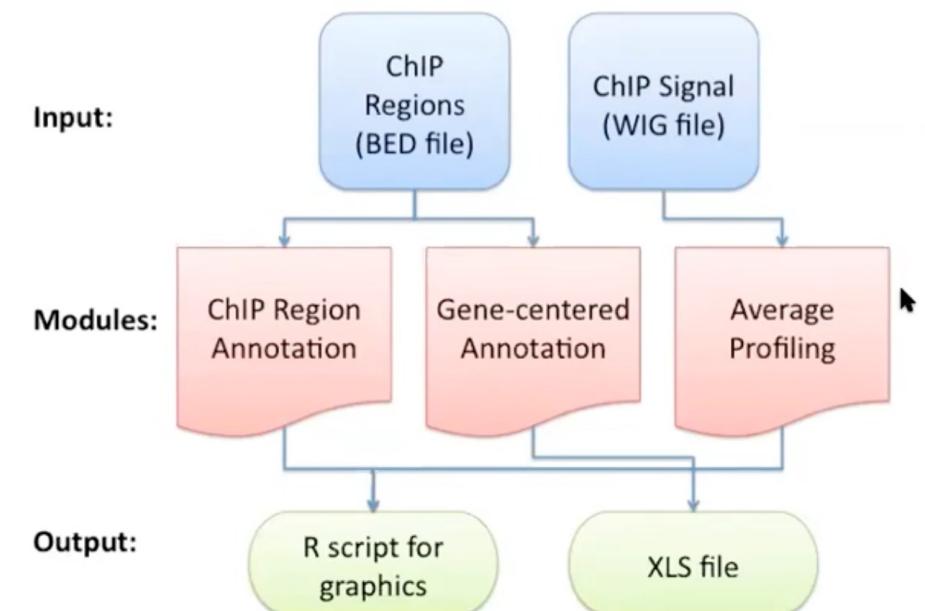
# CEAS peak annotation

- Relationship between genes and peaks

name	chr	txStart	txEnd	strand	dist.u.TSS	dist.d.TSS	dist.u.TTS	dist.d.TTS	3000bp.u.TSS	3000bp.d.TSS	1
<chr>	<chr>	<int>	<int>	<chr>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	
NR_024540	chr1	14362	29370	-	919308	NA	934316	NA	0.00	0.00	
NR_026818	chr1	34611	36081	-	912597	NA	914067	NA	0.00	0.00	
NR_026820	chr1	34611	36081	-	912597	NA	914067	NA	0.00	0.00	
NR_026822	chr1	34611	36081	-	912597	NA	914067	NA	0.00	0.00	
NM_001005484	chr1	69090	70008	+	NA	879588	NA	878670	0.00	0.00	
NR_028322	chr1	323891	328580	+	NA	624787	NA	620098	0.00	0.00	
NR_028327	chr1	323891	328580	+	NA	624787	NA	620098	0.00	0.00	
NR_028325	chr1	323891	328580	+	NA	624787	NA	620098	0.00	0.00	
NM_001005277	chr1	367658	368595	+	NA	581020	NA	580083	0.00	0.00	
NM_001005224	chr1	367658	368595	+	NA	581020	NA	580083	0.00	0.00	
NM_001005221	chr1	367658	368595	+	NA	581020	NA	580083	0.00	0.00	
NR_031741	chr1	566188	566265	-	382413	NA	382490	NA	0.00	0.00	
NM_001005277	chr1	621097	622034	-	326644	NA	327581	NA	0.00	0.00	
NM_001005224	chr1	621097	622034	-	326644	NA	327581	NA	0.00	0.00	

## Get started with CEAS

### CEAS Overview





Name	Date Modified	Size	Kind
homerMotifs.all.motifs	Today, 12:10 PM	30 KB	Document
homerMotifs.motifs8	Today, 11:52 AM	9 KB	Document
homerMotifs.motifs10	Today, 11:58 AM	10 KB	Document
homerMotifs.motifs12	Today, 12:10 PM	11 KB	Document
homerResults	Today, 12:20 PM	--	Folder
homerResults.html	Today, 12:20 PM	13 KB	HTML
knownResults	Today, 11:50 AM	--	Folder
knownResults.html	Today, 11:50 AM	23 KB	HTML
knownResults.txt	Today, 11:50 AM	29 KB	Plain Text
motifFindingParameters.txt	Today, 11:37 AM	83 bytes	Plain Text
seq.autonorm.tsv	Today, 11:40 AM	2 KB	Plain Text

## Homer Known Motif Enrichment Results (STAT1\_6h\_IFNa\_homer)

[Homer \*de novo\* Motif Results](#)

[Gene Ontology Enrichment Results](#)

[Known Motif Enrichment Results \(txt file\)](#)

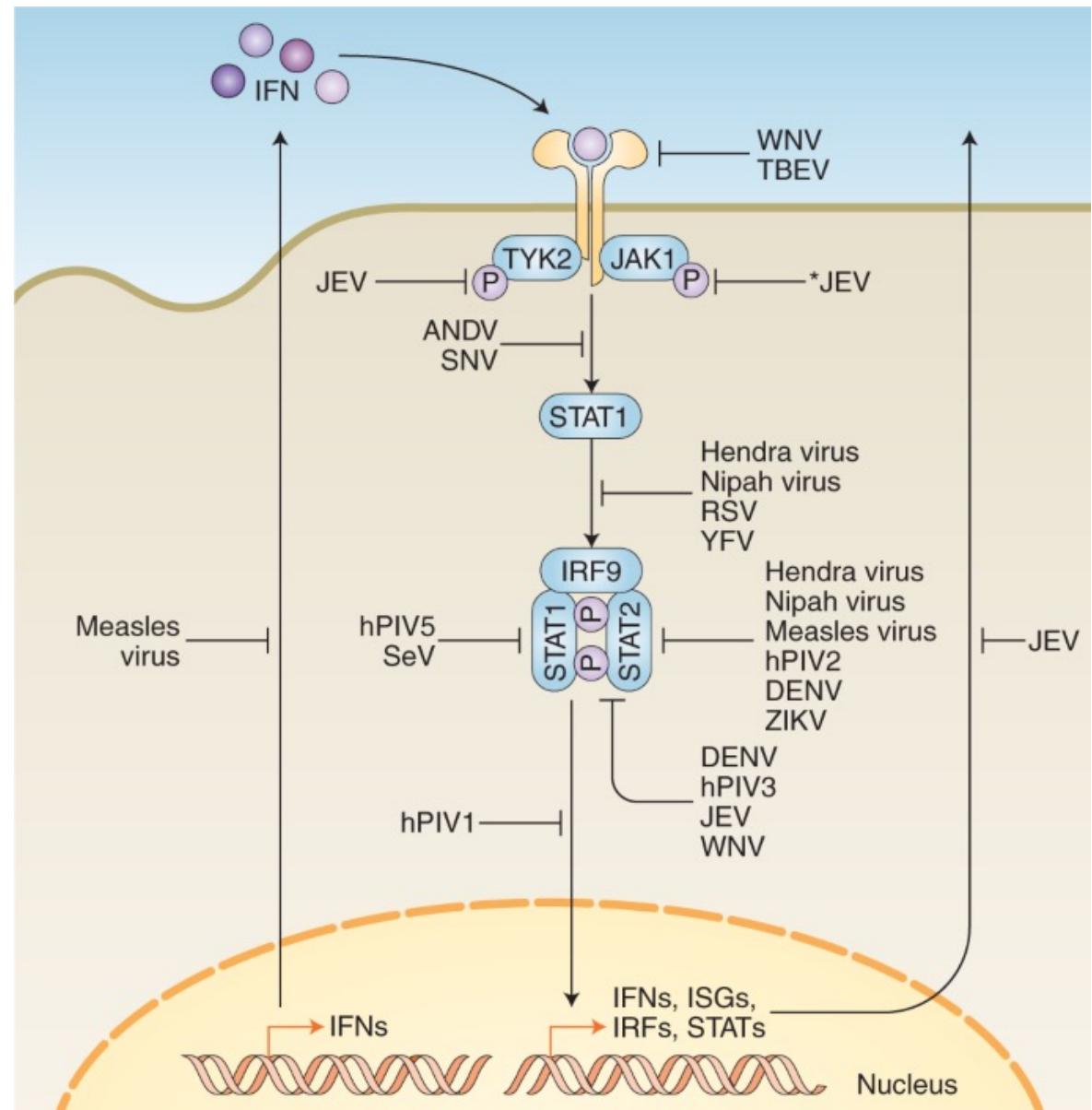
Total Target Sequences = 1025, Total Background Sequences = 37194

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		ISRE(IRF)/ThioMac-LPS-Expression(GSE23622)/Homer	1e-348	-8.033e+02	0.0000	273.0	26.63%	230.6	0.62%
2		IRF2(IRF)/Erythroblasts-IRF2-ChIP-Seq(GSE36985)/Homer	1e-321	-7.414e+02	0.0000	298.0	29.07%	405.3	1.09%
3		IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer	1e-236	-5.450e+02	0.0000	280.0	27.32%	661.8	1.77%
4		TEAD(TEA)/Fibroblast-PU.1-ChIP-Seq(Unpublished)/Homer	1e-84	-1.956e+02	0.0000	348.0	33.95%	4072.1	10.92%
5		TEAD4(TEA)/Trophoblast-Tead4-ChIP-Seq(GSE37350)/Homer	1e-81	-1.865e+02	0.0000	352.0	34.34%	4305.2	11.54%

# Sample data: the effect of Interferon $\alpha$ on STAT1 binding

## Samples

- STAT1 30 mins
- STAT1 6 hours
- Input (INP) 30 mins
- Input (INP) 6 hrs



https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31477

... Search

NCBI Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE31477

**Series GSE31477** Query DataSets for GSE31477

Status Public on Aug 30, 2011

Title ENCODE Transcription Factor Binding Sites by ChIP-seq from Stanford/Yale /USC/Harvard

Project ENCODE

Organism [Homo sapiens](#)

Experiment type Genome binding/occupancy profiling by high throughput sequencing

Summary This data was generated by ENCODE. If you have questions about the data, contact the submitting laboratory directly (Philip Cayting <mailto:pcayting@stanford.edu>). If you have questions about the Genome Browser track associated with this data, contact ENCODE (<mailto:genome@soe.ucsc.edu>).

This track shows probable binding sites of the specified transcription factors (TFs) in the given cell types as determined by chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq). Included for each cell type is the input signal, which represents the control condition where no antibody targeting was performed. For each experiment (cell type vs. antibody) this track shows a graph of enrichment for TF binding (Signal), along with sites that have the greatest evidence of transcription factor binding (Peaks).

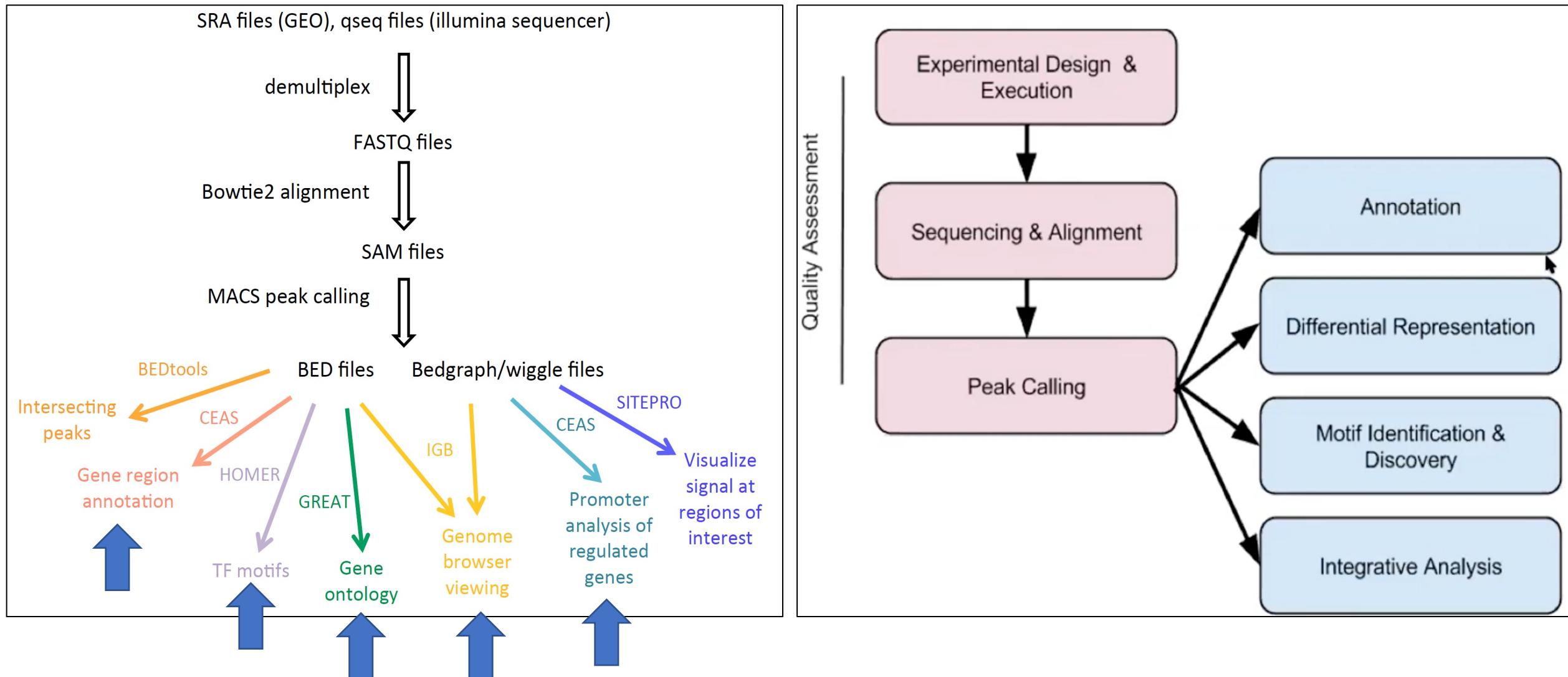
For data usage terms and conditions, please refer to <http://www.genome.gov/27528022> and <http://www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf>

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31477

...

GSM935463	USC_ChipSeq_H1-hESC_CtBP2_UCDavis
GSM935464	USC_ChipSeq_K562_KAP1_UCDavis
GSM935465	USC_ChipSeq_PBDE_GATA-1_UCDavis
GSM935466	Yale_ChipSeq_K562_IFNa6h_c-Myc_std
GSM935467	Yale_ChipSeq_K562_IFNg6h_c-Jun_std
GSM935468	Yale_ChipSeq_K562_IFNa6h_c-Jun_std
GSM935469	Yale_ChipSeq_K562_IFNa6h_STAT2_std
GSM935470	Yale_ChipSeq_K562_IFNa30_STAT2_std
GSM935471	Yale_ChipSeq_K562_IFNa6h_STAT1_std
GSM935472	Yale_ChipSeq_K562_IFNa30_STAT1_std
GSM935473	Yale_ChipSeq_K562_IFNg6h_Pol2_std
GSM935474	Yale_ChipSeq_K562_IFNa6h_Pol2_std
GSM935475	Yale_ChipSeq_K562_IFNa30_Pol2_std
GSM935476	USC_ChipSeq_HeLa-S3_E2F6_std
GSM935477	USC_ChipSeq_MCF-7_HA-E2F1_UCDavis
GSM935478	Stanford_ChipSeq_GM12878_TNFa_NFKB_IgG-rab
GSM935479	USC_ChipSeq_K562_ZNF274_UCDavis
GSM935480	USC_ChipSeq_GM12878_TR4_std
GSM935481	Stanford_ChipSeq_K562_Pol3_std
GSM935482	USC_ChipSeq_GM12878_YY1_std
GSM935483	Harvard_ChipSeq_GM12878_ZZZ3_std
GSM935484	USC_ChipSeq_HeLa-S3_E2F1_std
GSM935485	USC_ChipSeq_MCF-7_Input_UCDavis
GSM935486	Harvard_ChipSeq_HeLa-S3_BDP1_std
GSM935487	Stanford_ChipSeq_K562_IFNg30_STAT1_std
GSM935488	Stanford_ChipSeq_K562_IFNg6h_STAT1_std
GSM935489	Harvard_ChipSeq_HeLa-S3_RPC155_std

# Part II, ChIP-seq analysis



# Reference

- Bulk RNA-seq
  - The pipeline and dataset 1
    - <https://github.com/UMMS-Biocore/RNASeqTutorial/blob/master/RNASeqTutorial.pdf>
  - Dataset 2 <https://bioinfo.umassmed.edu/index.php?p=35#p1e3>
  - Cluster profiler
    - <https://yulab-smu.top/clusterProfiler-book/>
- Chip-seq
  - The pipeline, <https://www.youtube.com/watch?v=JYBP5BpRfTM>
  - DiffBind, <https://www.youtube.com/watch?v=fa9sw3QqfWQ>

# Summary

- Single cell RNA-seq (day 1)
  - Generate FASTQ from bcl files
  - Cellranger count
  - Pipelines for QC and dimension reduction
  - Trajectory inference through RNA velocity
  - **Regulatory network analysis with SCENIC**
- Bulk RNA-seq (day 2)
  - Alignment using RSEM and Tophat
  - Normalization and differential expression via DESeq2
- ChIP-seq
  - Alignment by Bowtie2
  - Peak calling with MACS2
  - Differential Peak calling using Homer
  - Peak annotation with CEAS
  - TF binding motif analysis

# Additional reference

- General tips for learning bioinformatics
  - [https://github.com/zhuy16/learning\\_notes](https://github.com/zhuy16/learning_notes)
- single cell rna-seq
  - <https://github.com/niad/single-cell-RNA-seq>
- Gene Regulatory Networks
  - [https://github.com/niad/Gene\\_Regulatory\\_Networks](https://github.com/niad/Gene_Regulatory_Networks)
- Functional annotation with clusterProfiler
  - [https://github.com/zhuy16/FunctionalAnnotation\\_notebooks](https://github.com/zhuy16/FunctionalAnnotation_notebooks)