Day2, Bulk RNA-seq & ChIP-seq

NIEHS Workshop--Analyzing NGS data
https://github.com/zhuy16/NIEHS_NGS_Workshop

July 14, 2021
9am-2pm

Yunhua Zhu, PhD
zzhonghua@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.

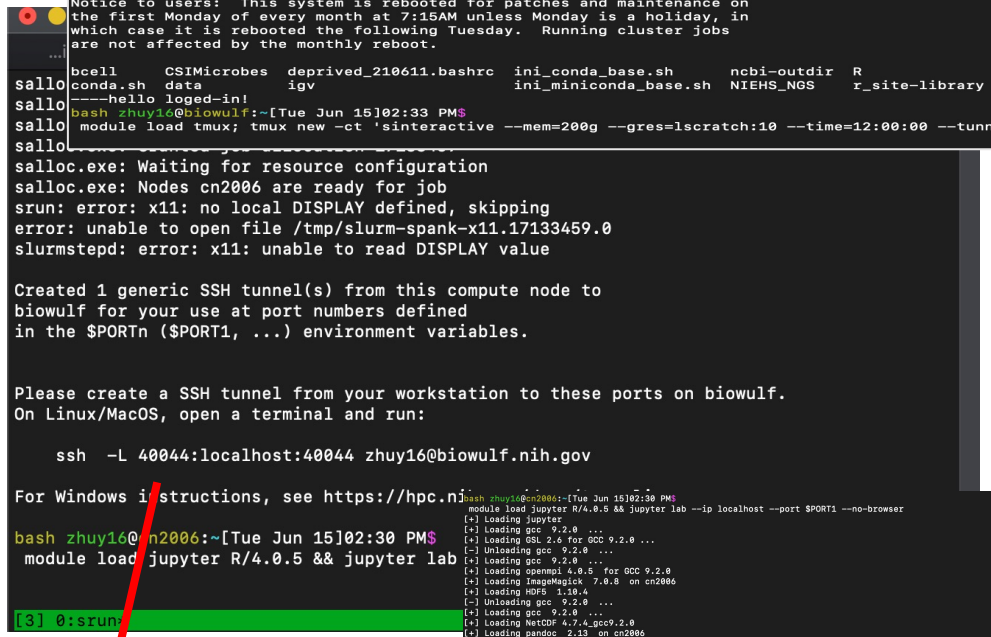
Review & Summary

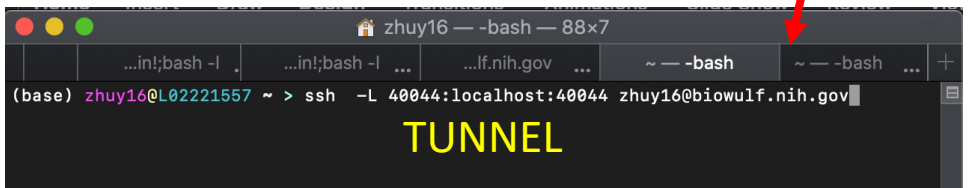
Install software to visualizing alignment results

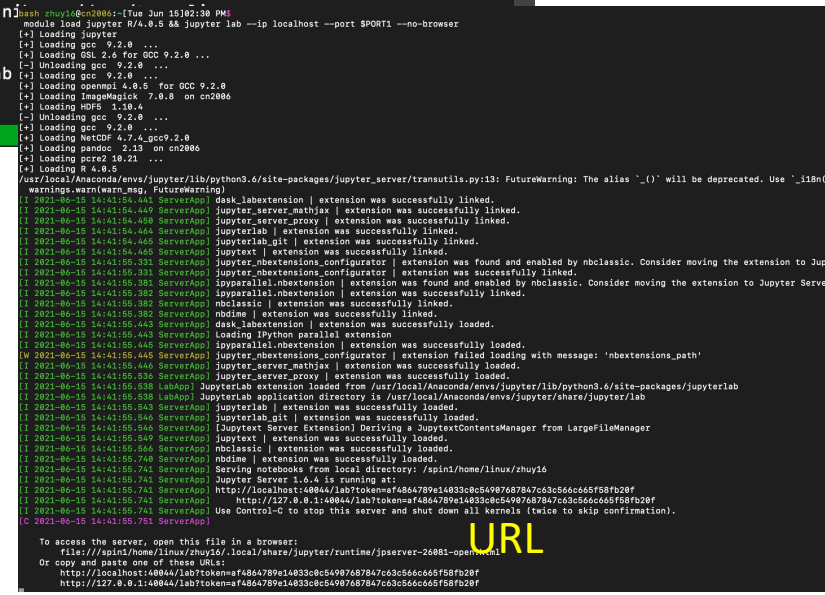
- Install IGV on your local machine
  - https://software.broadinstitute.org/software/igv/download
- Install IGM on your local machine
  - https://www.bioviz.org/

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 (3rd) 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.

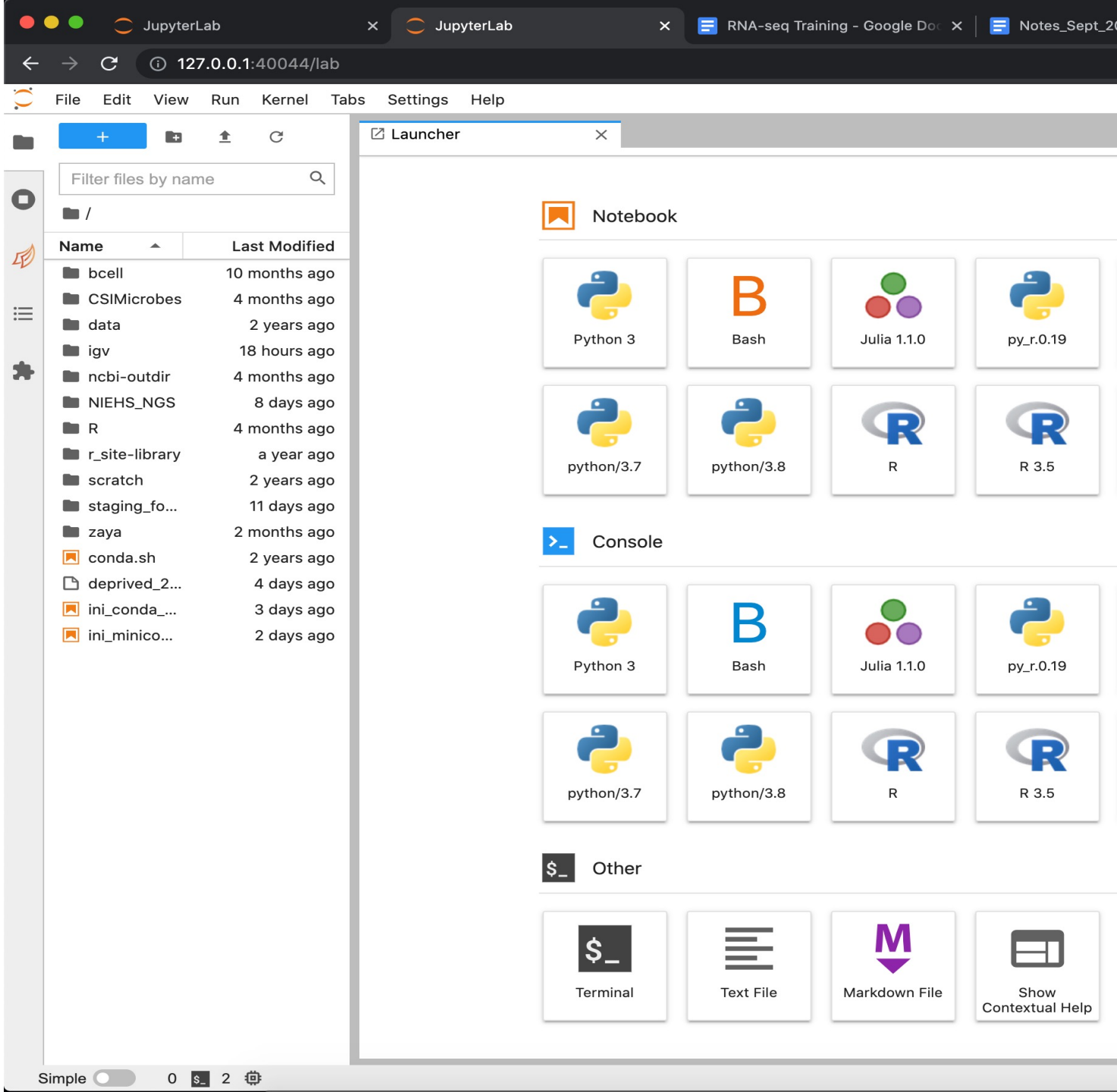






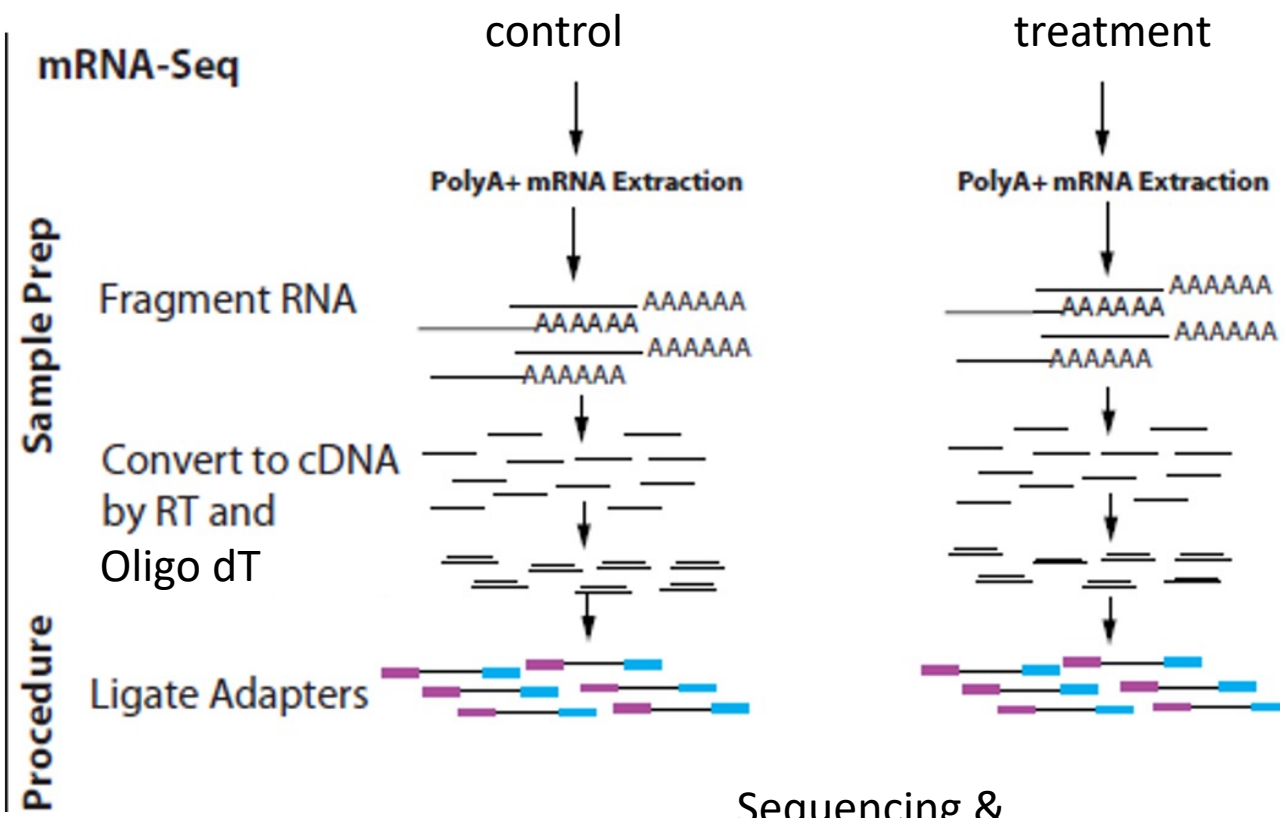
Jupyter lab

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

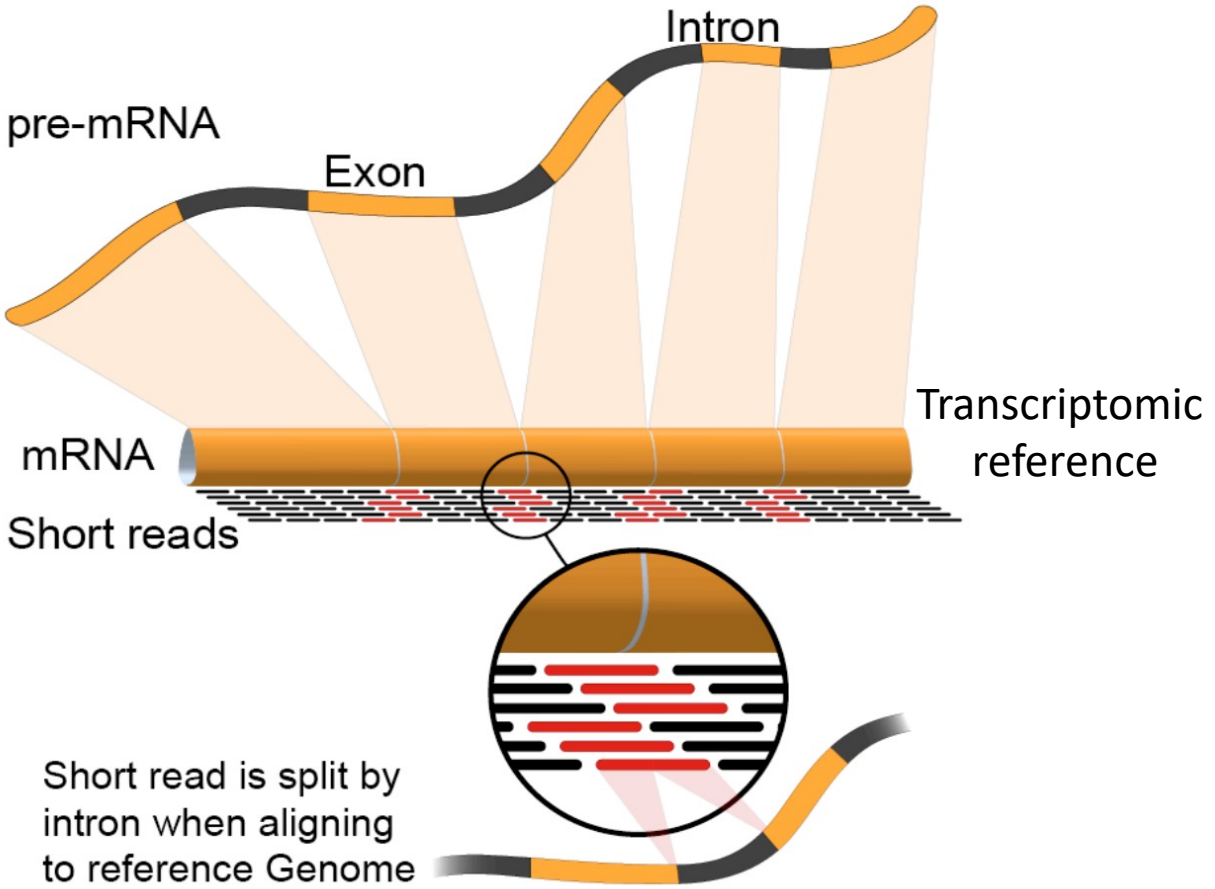


Part II, Bulk RNA-seq

Part II, RNA-seq analysis

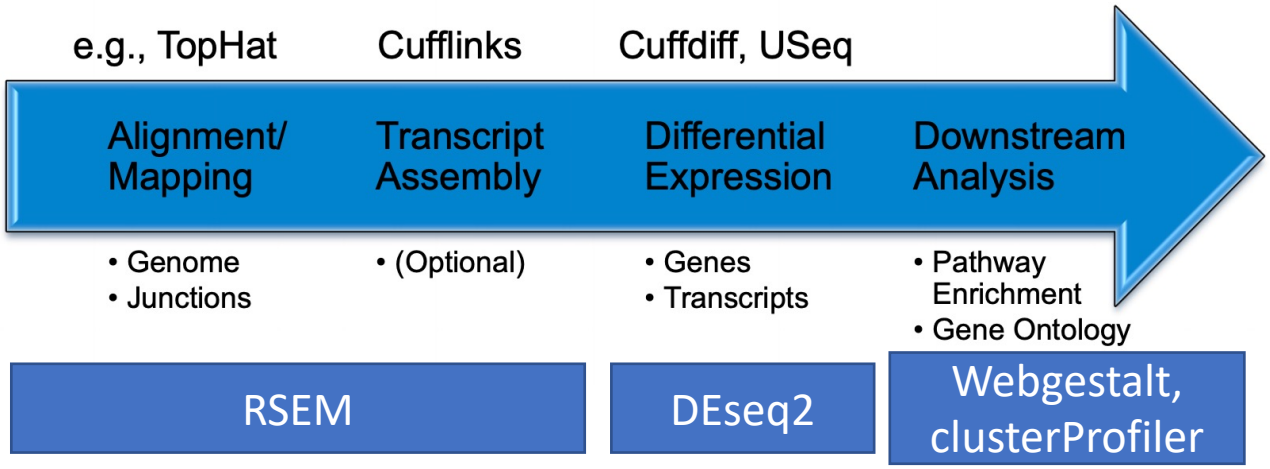


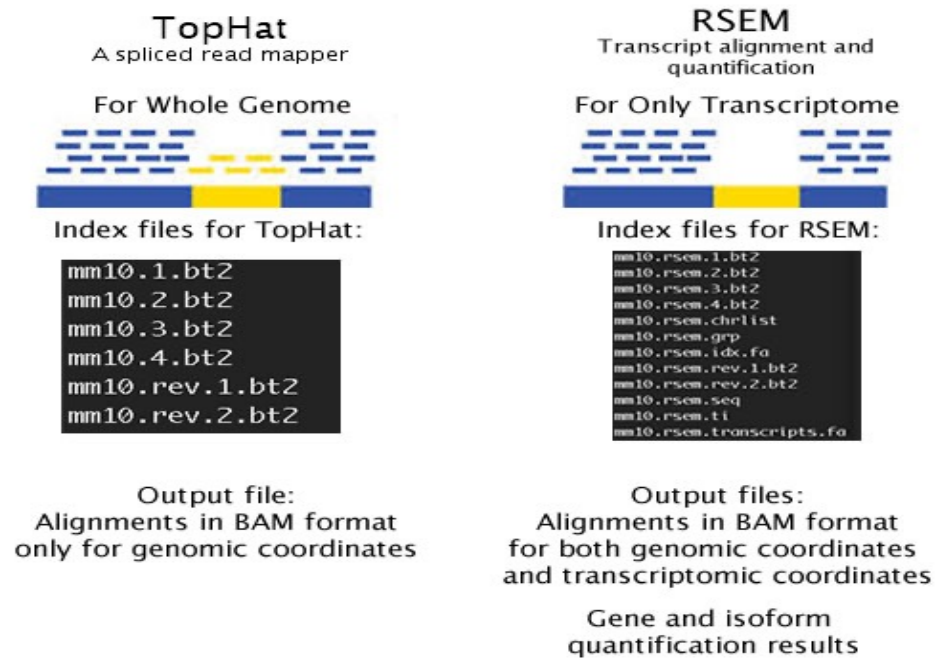
Sequencing &
preprocessing



Steps

- Quality filtering
- Cut adaptors
- Alignment
- Differential expression



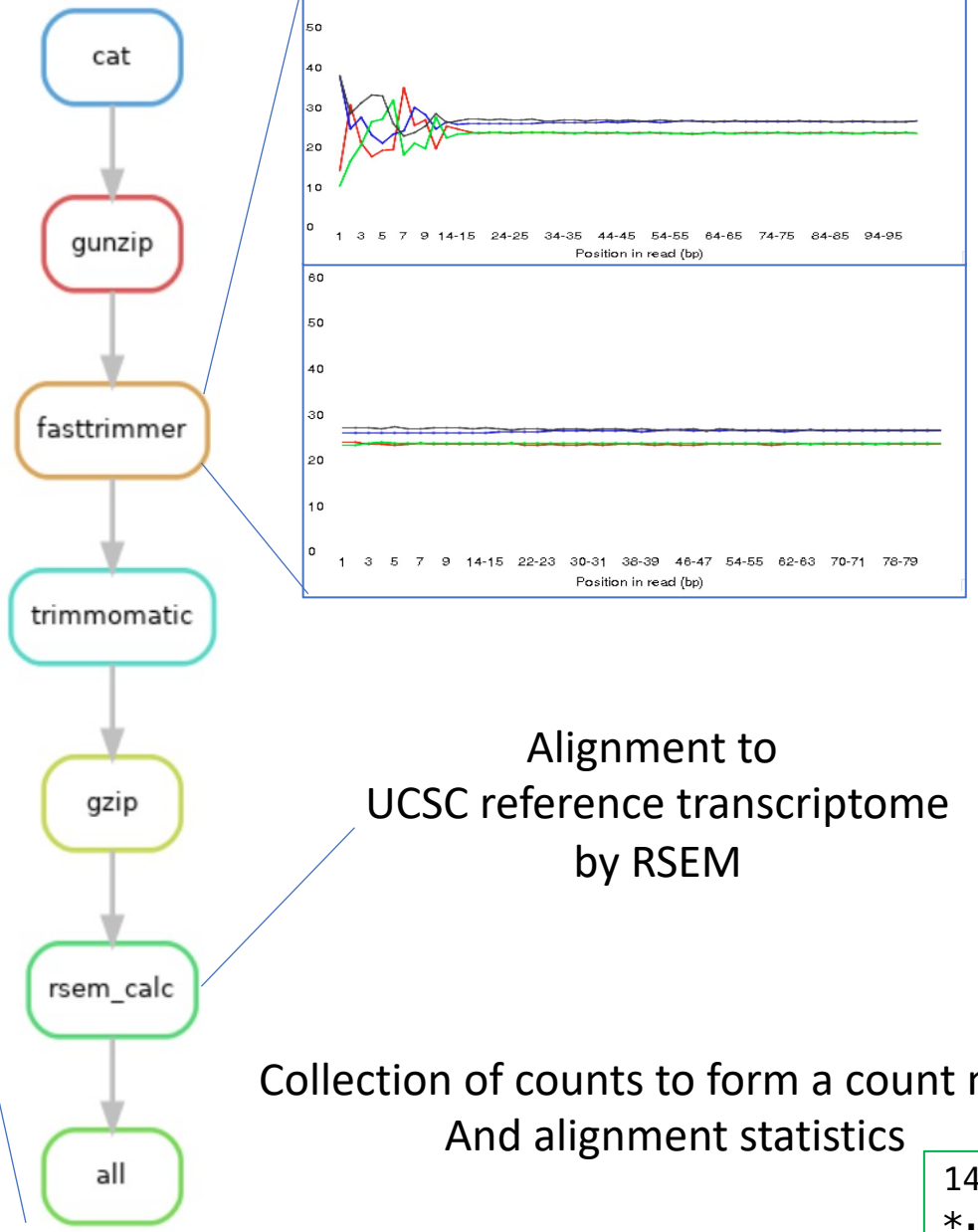


https://bioinfo.umassmed.edu/index.php?p=33

snakemake Pipeline for alignment

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

snakemake pipeline



995Gbytes
2016 files

Snakemake pipeline

Alignment to
UCSC reference transcriptome
by RSEM

Collection of counts to form a count matrix
And alignment statistics

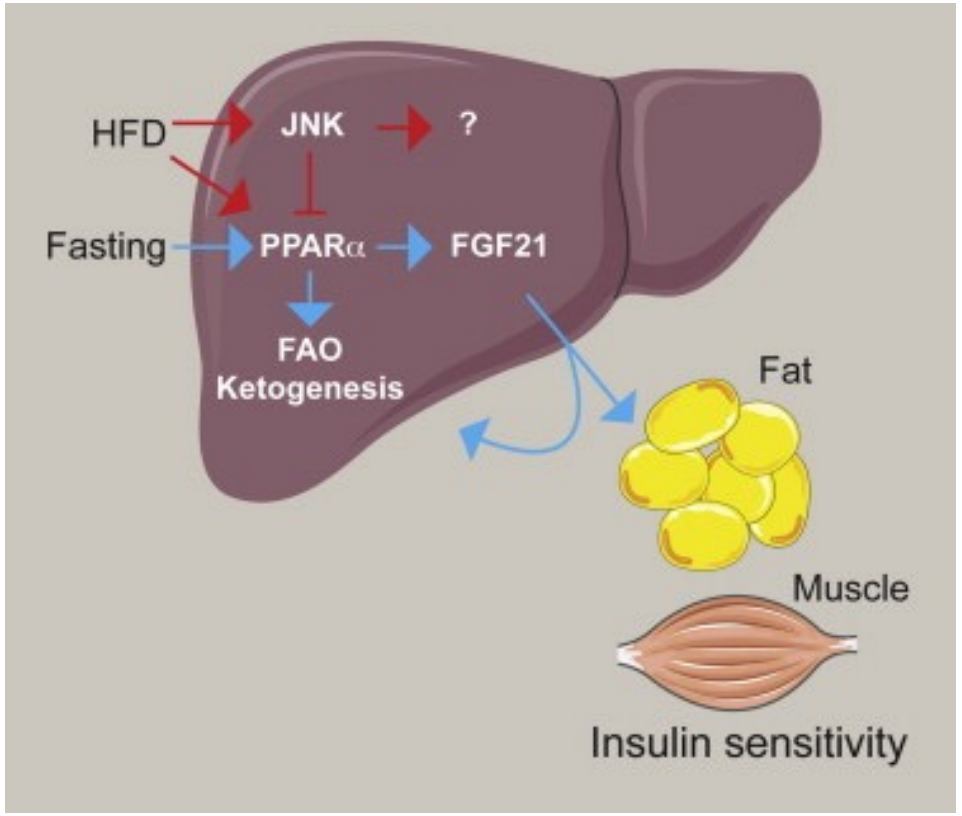
145 files
*.genes.results

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

Jupyter Notebooks on Biowulf

Demo on RNA-seq: a small sample data

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



https://www.sciencedirect.com/science/article/pii/S1550413114002757?via%3Dihub

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

raw count for gene i, sample j

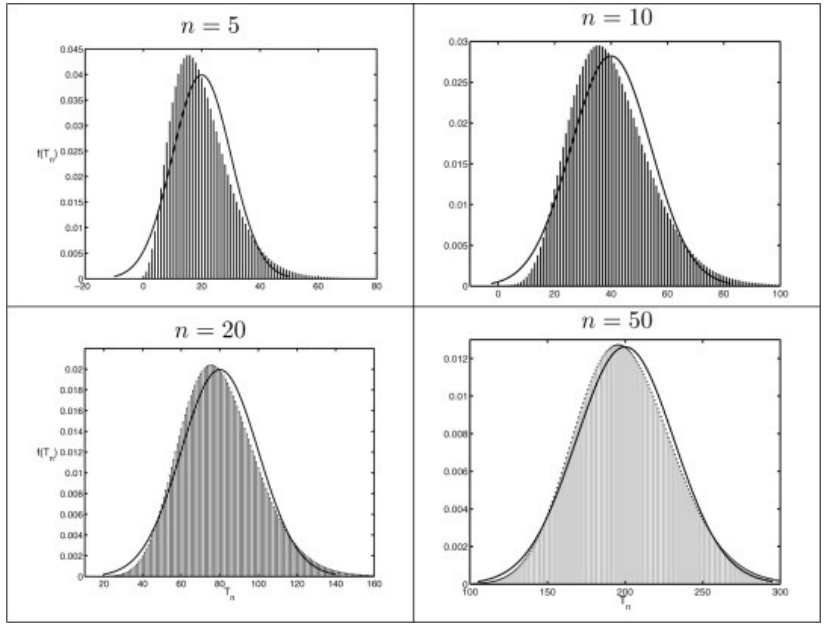
The mean is taken as "normalized counts" scaled by a normalization factor

one dispersion per gene

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

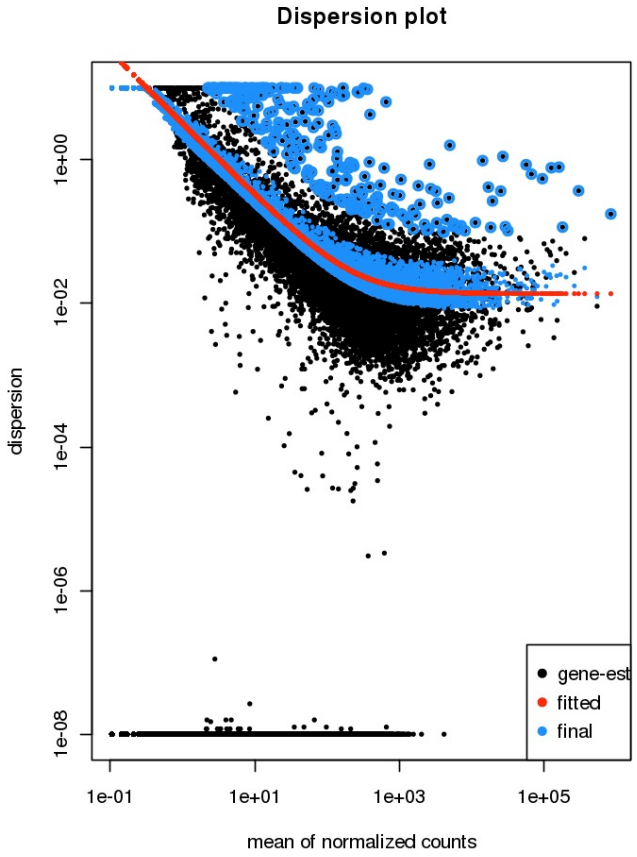
Dispersion vs sd
Wald test vs t test

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

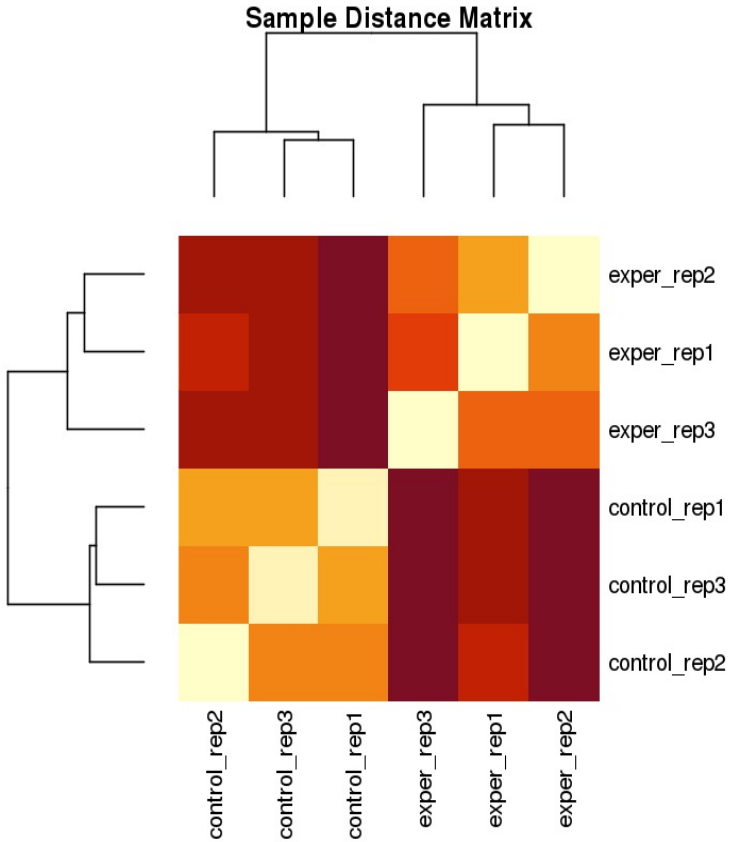


https://hbctraining.github.io/DGE_workshop_salmon/lessons/05_DGE_DESeq2_analysis2.html

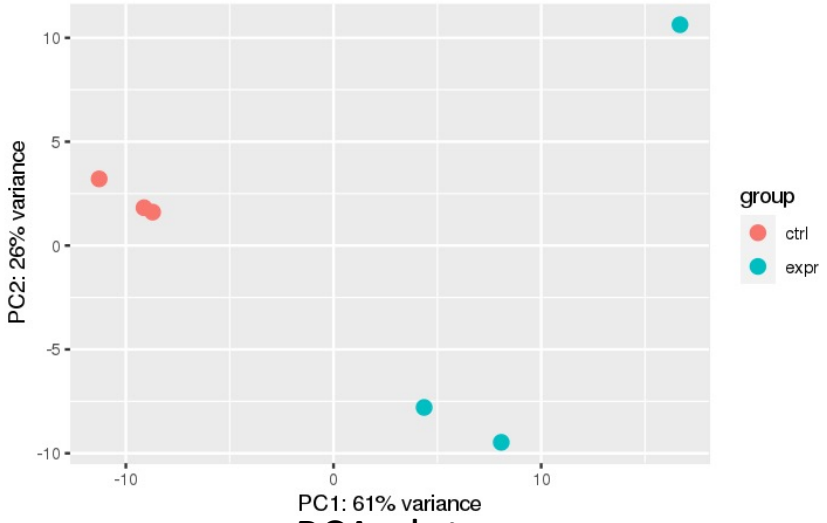
Plots



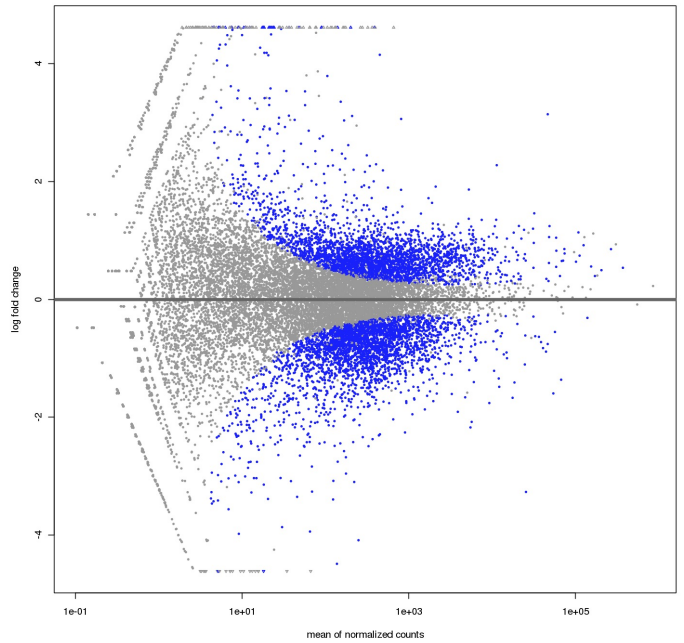
Dispersion plot



Correlation heatmap



PCA plot

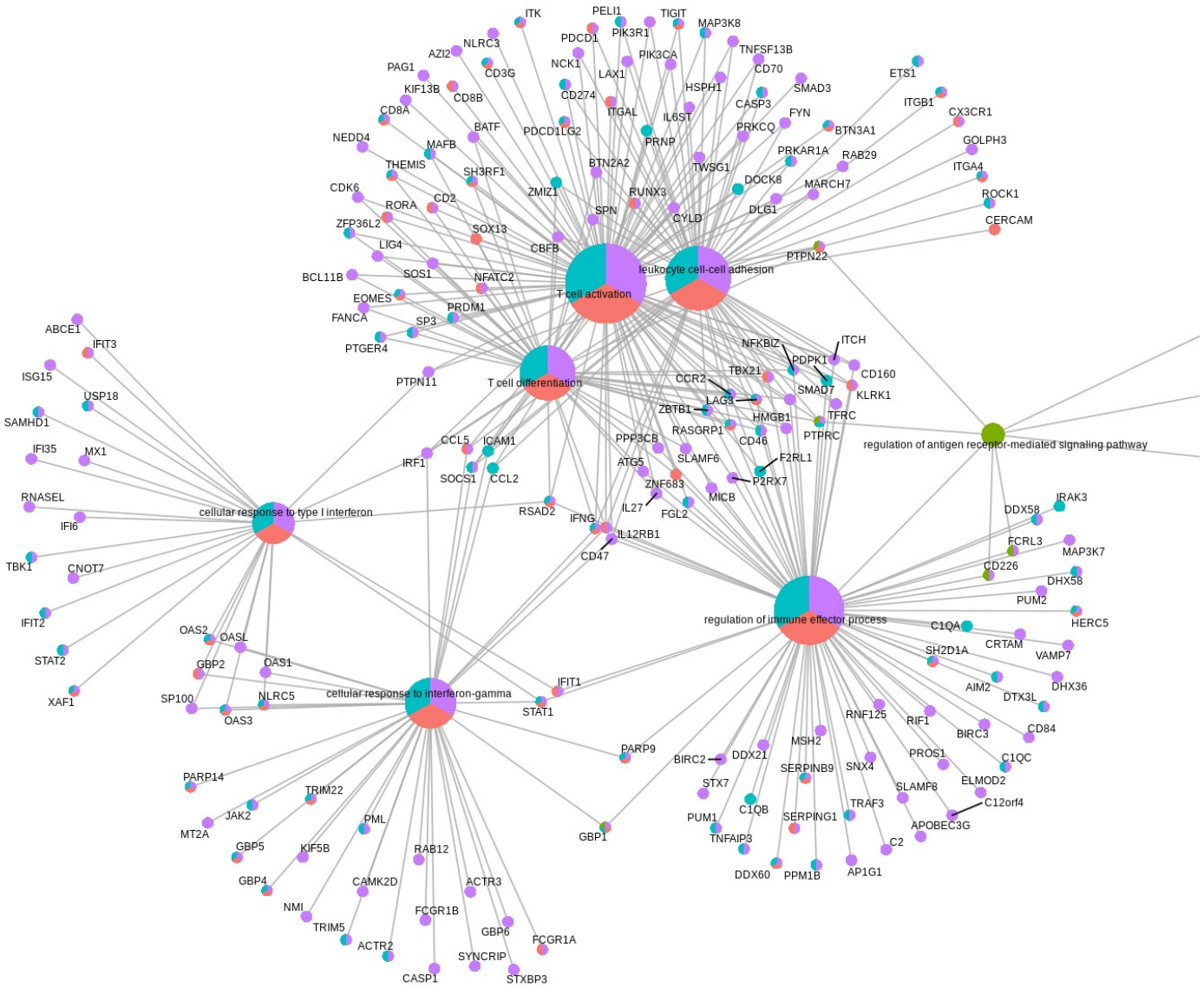


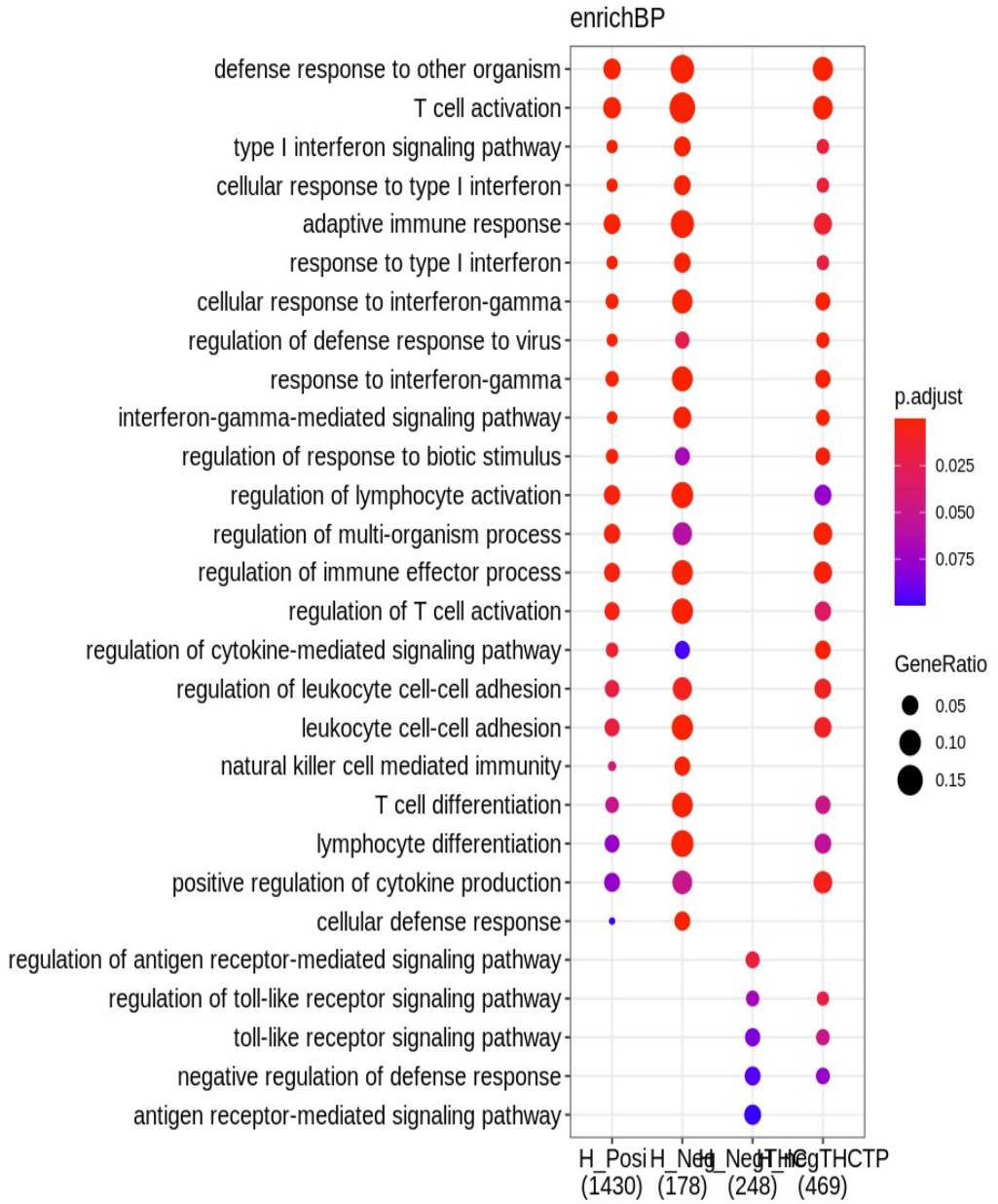
MA plot

Visualizing pathway—gene relationship

Example: selected GO-BP

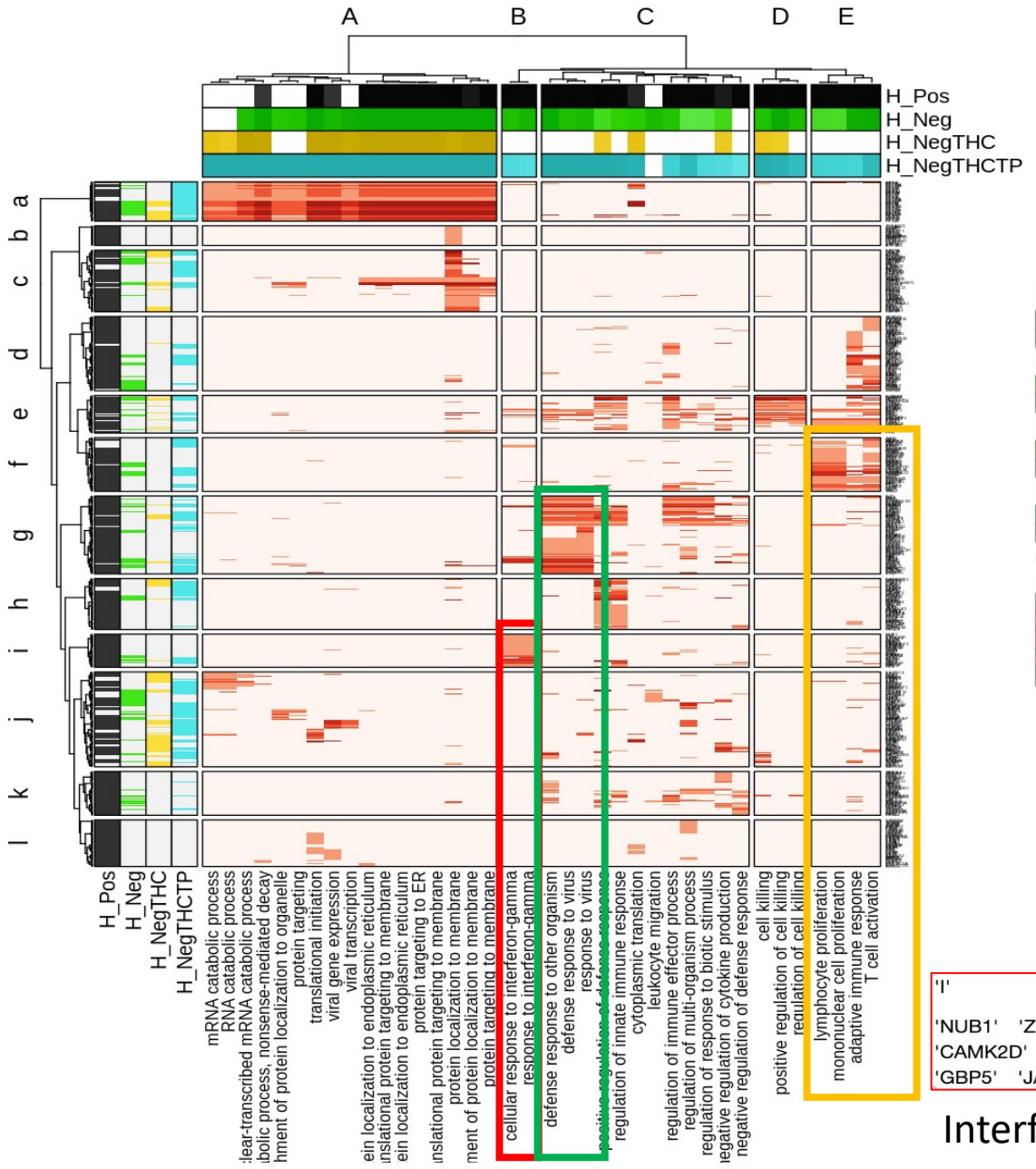
Redundancy of pathway names





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

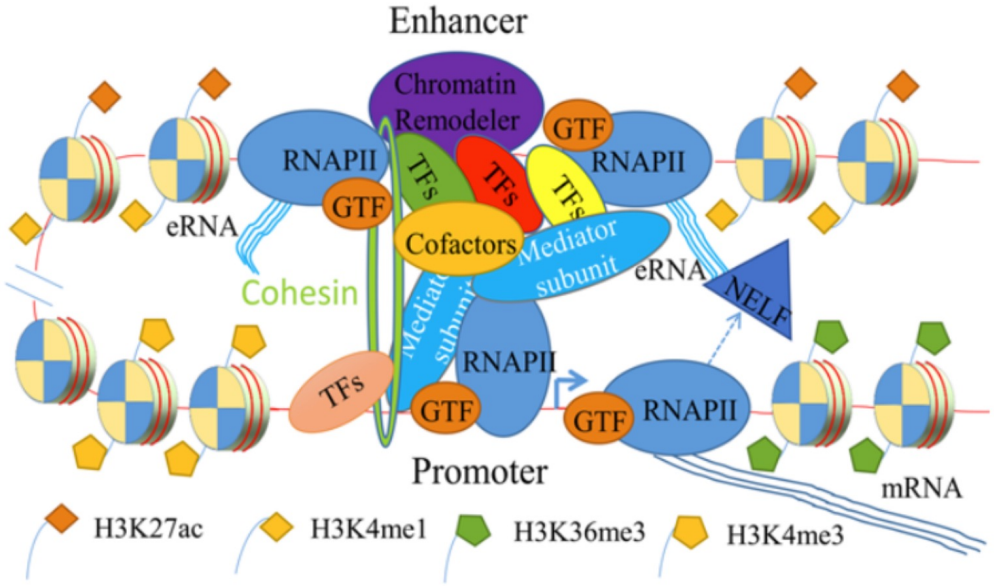
Reference

- 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)

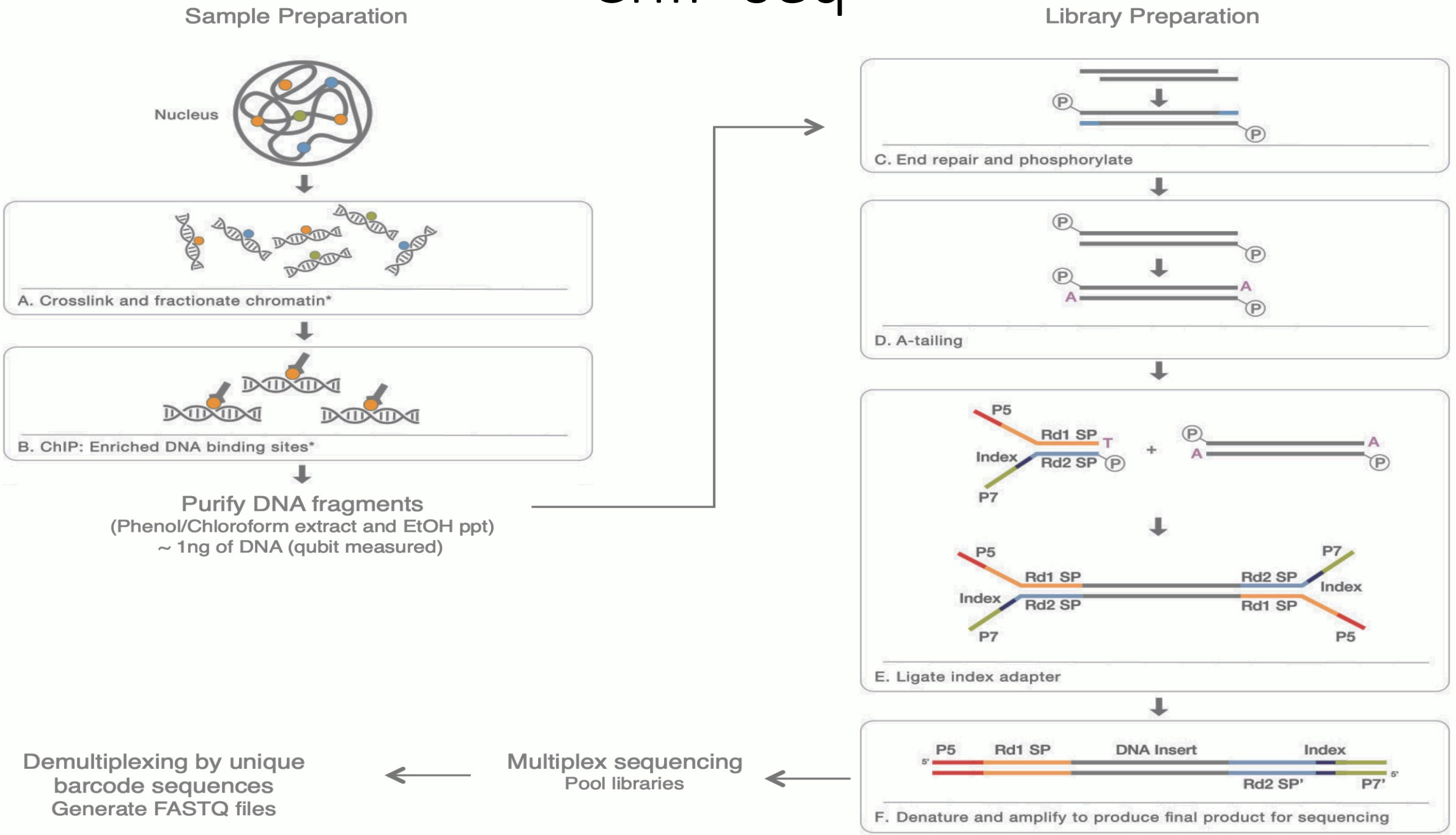
- Background and related chromatin types
- Peaks and input Controls
- Quality controls
- Workflow and sample data
- Demo on BioWulf

Chromatin = DNA and associated proteins



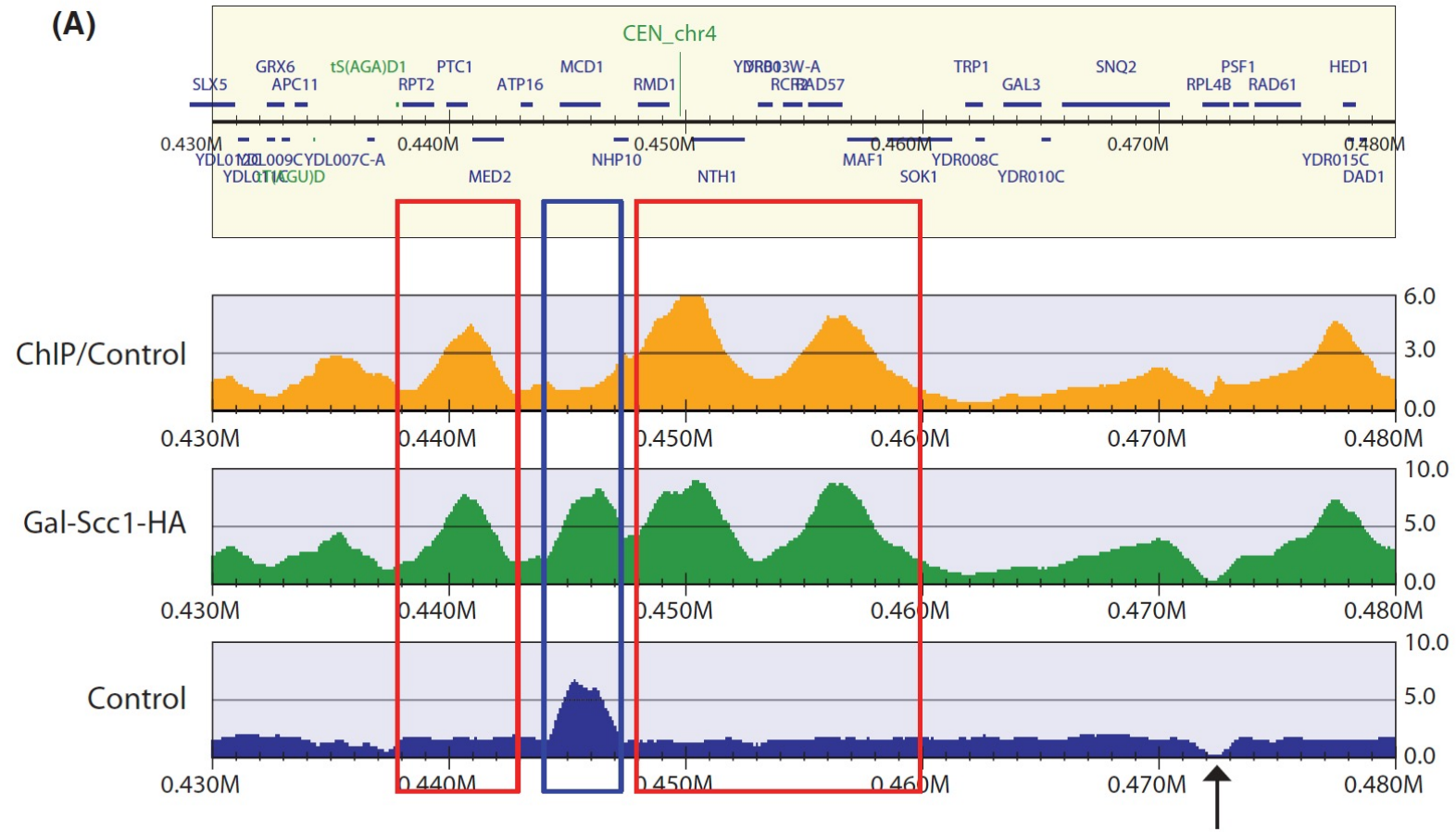
Hu & Tee. Bioscience Reports (2017)

ChIP-seq


Sample Preparation
Nucleus
A. Crosslink and fractionate chromatin*
B. ChIP: Enriched DNA binding sites*
Purify DNA fragments
(Phenol/Chloroform extract and EtOH ppt)
~ 1ng of DNA (qubit measured)
Library Preparation
C. End repair and phosphorylate
D. A-tailing
P5
Rd1 SP
Index
Rd2 SP
P7
E. Ligate index adapter
DNA Insert
F. Denature and amplify to produce final product for sequencing
Multiplex sequencing
Pool libraries
Demultiplexing by unique
barcode sequences
Generate FASTQ files

Peaks are significant enrichments over Input DNA

R Nakato *et al.*

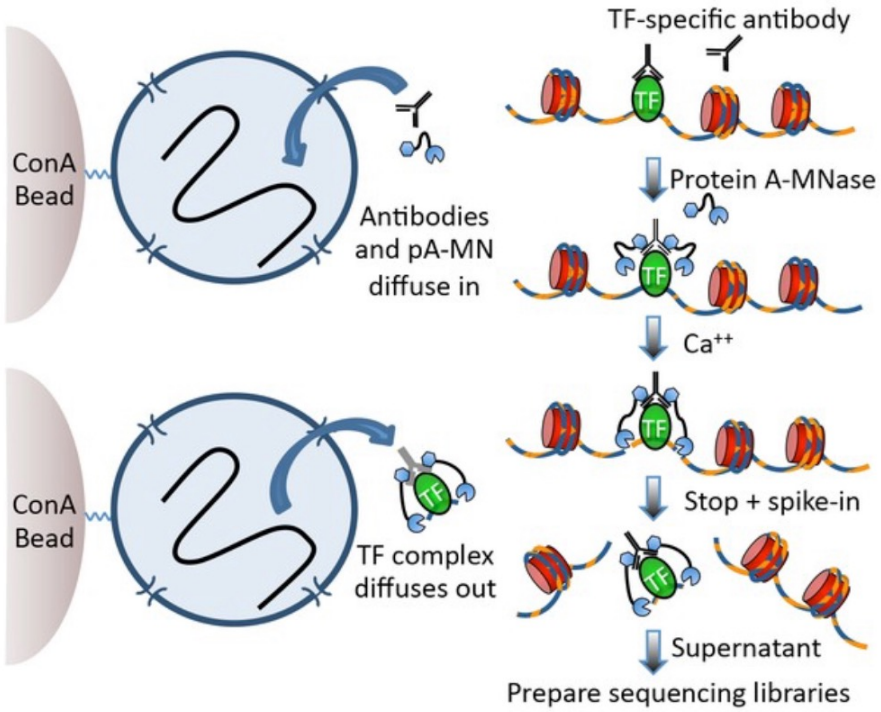


Peaks in greylist
& known blacklist

Control can be used in

MACS2 peak calling
Homer differential peaks
DiffBind greylist removal

Cut&Run - alternative



Skene & Henikoff. eLife (2017)

Benefits:

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

Different IP produce different type of peaks

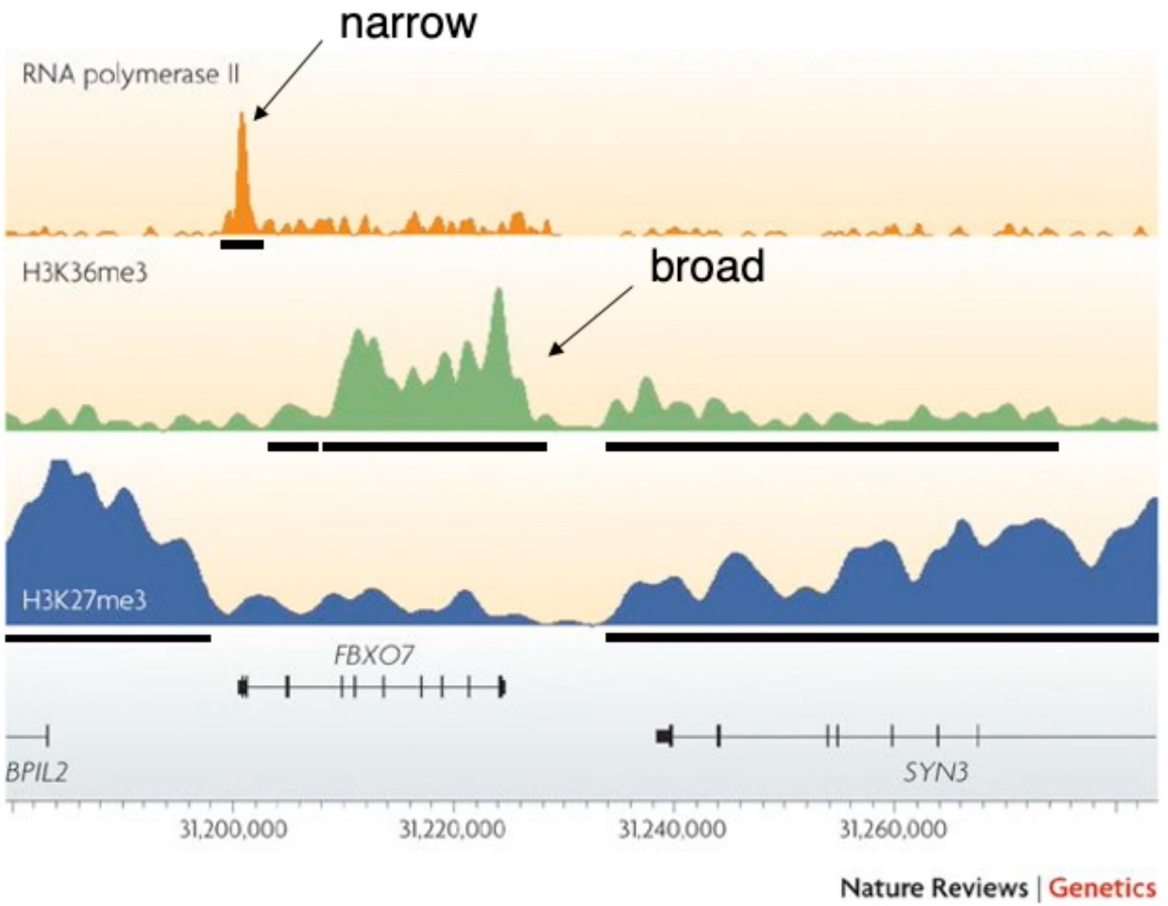
3 Types of peaks:

1. Sharp & narrow (100s bp)
(eg. site-specific TF)

2. Broader but defined (kb)

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

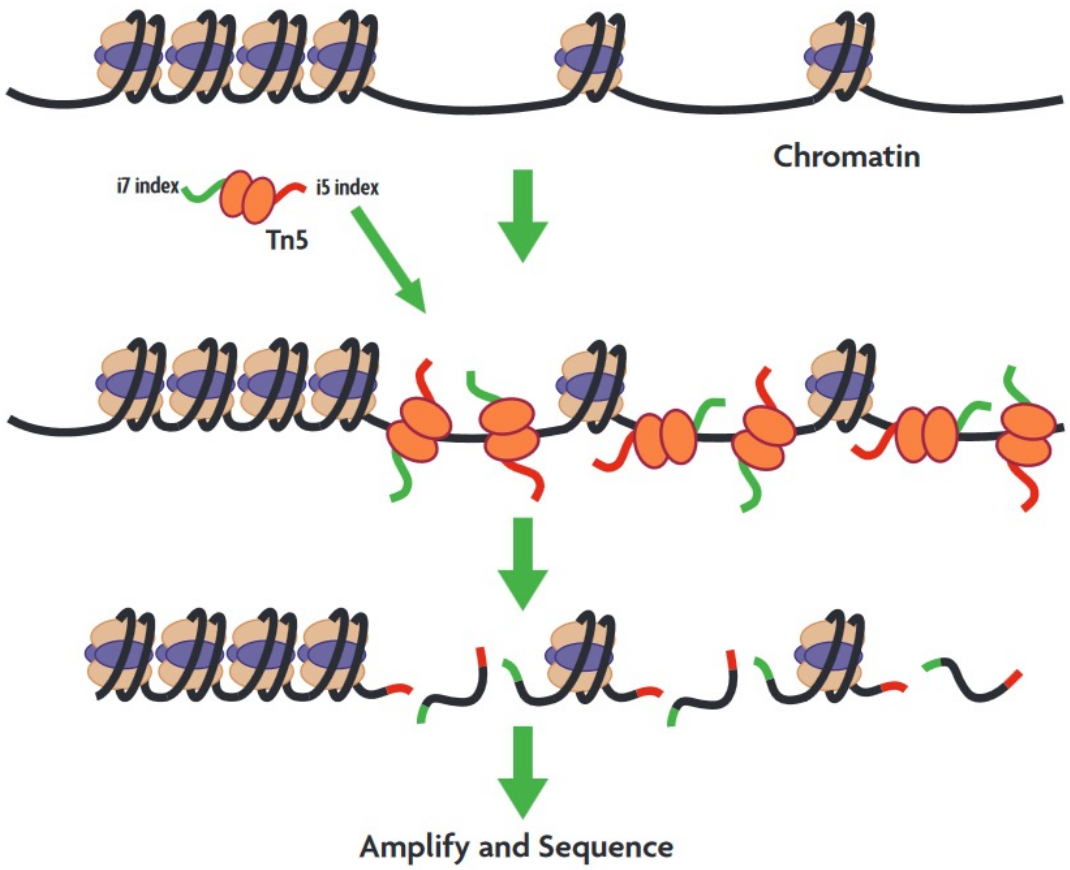
Histone Mark	Type
H3K27me3	broad
H3K36me3	broad
H3K4me3	narrow
H3K27ac	narrow



https://www.nature.com/articles/nrg2641

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.

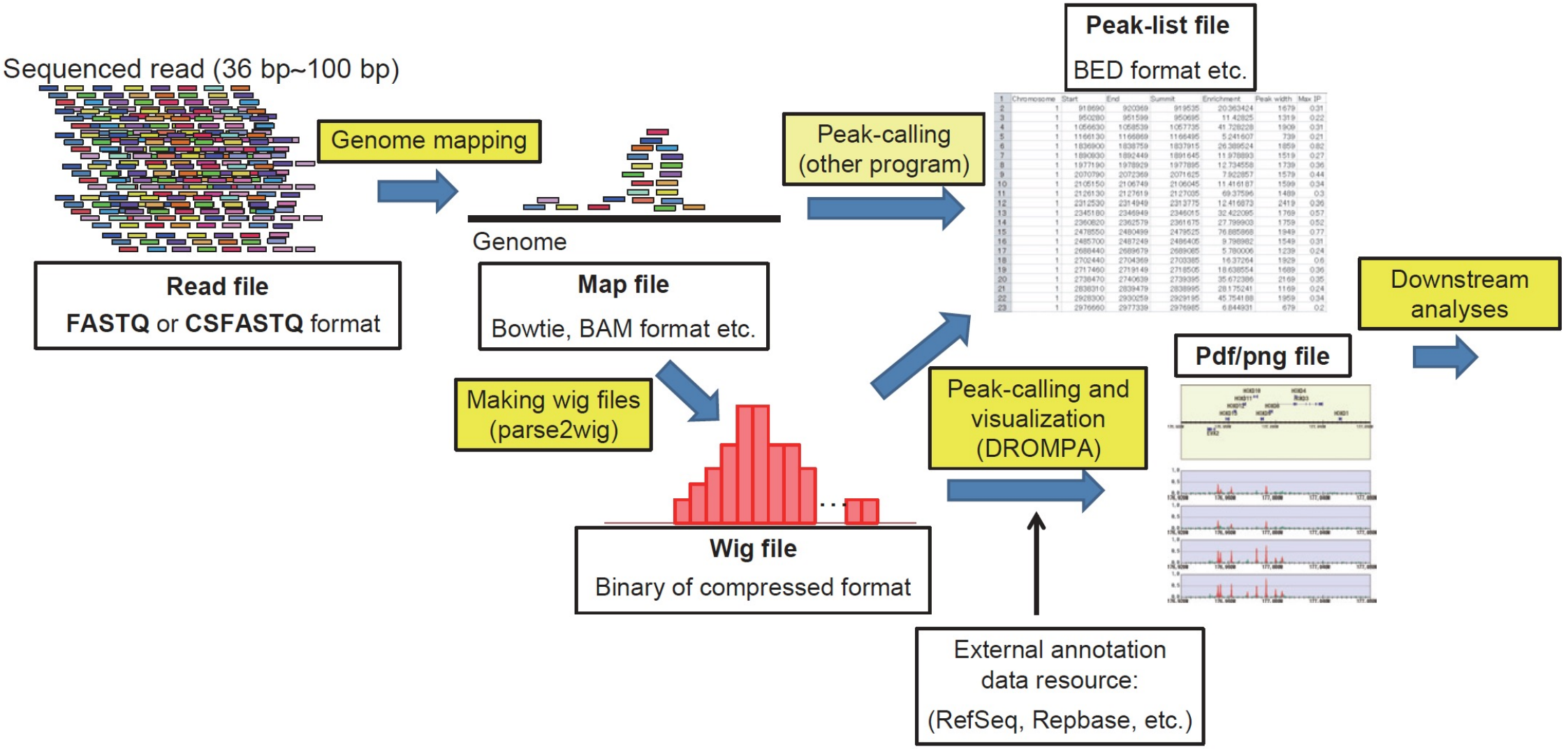


https://www.bioscience.co.uk/cpl/atac-seq-kit

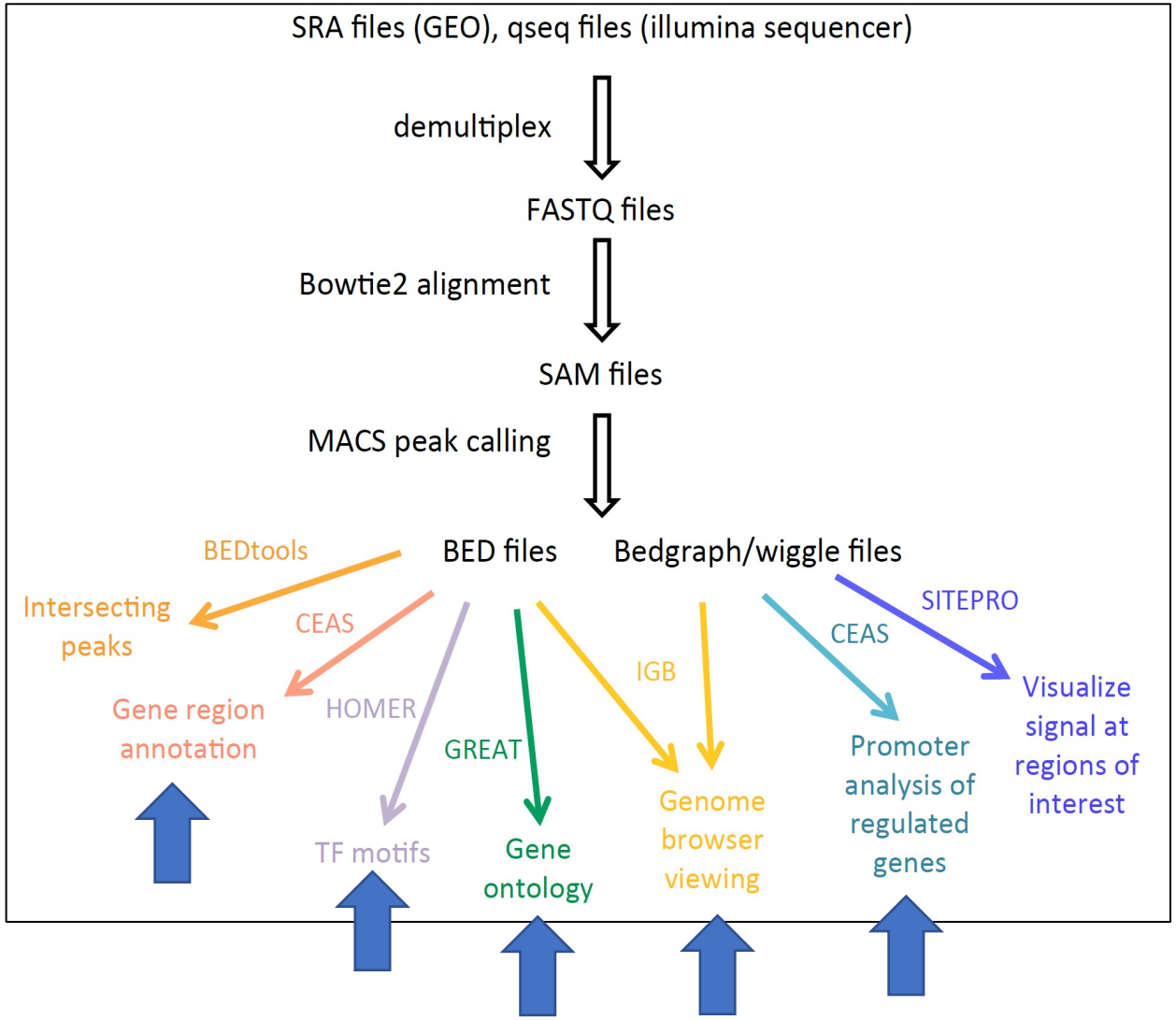
ChIP-Seq analysis pipeline example

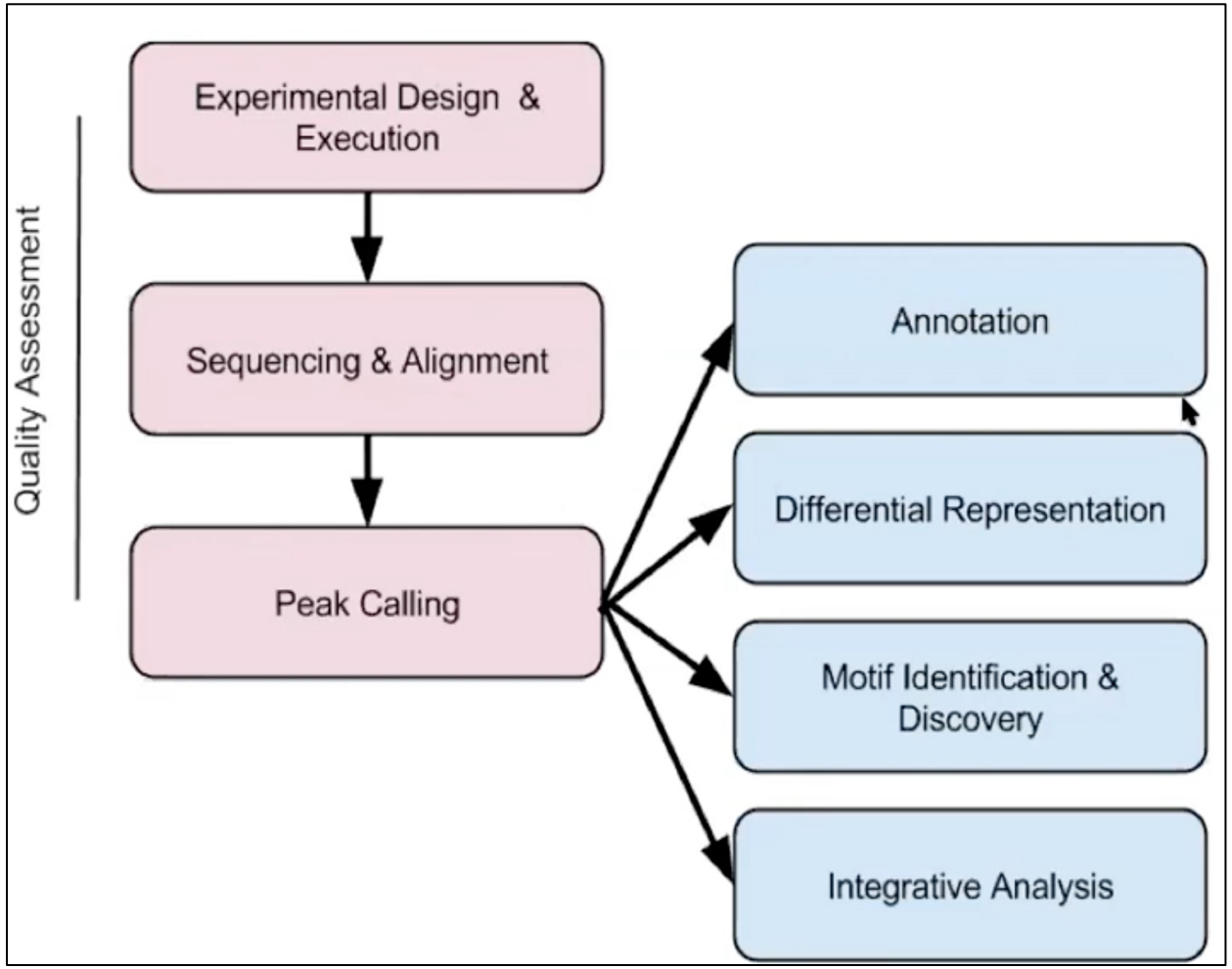
R Nakato *et al.*

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



Part II, ChIP-seq analysis



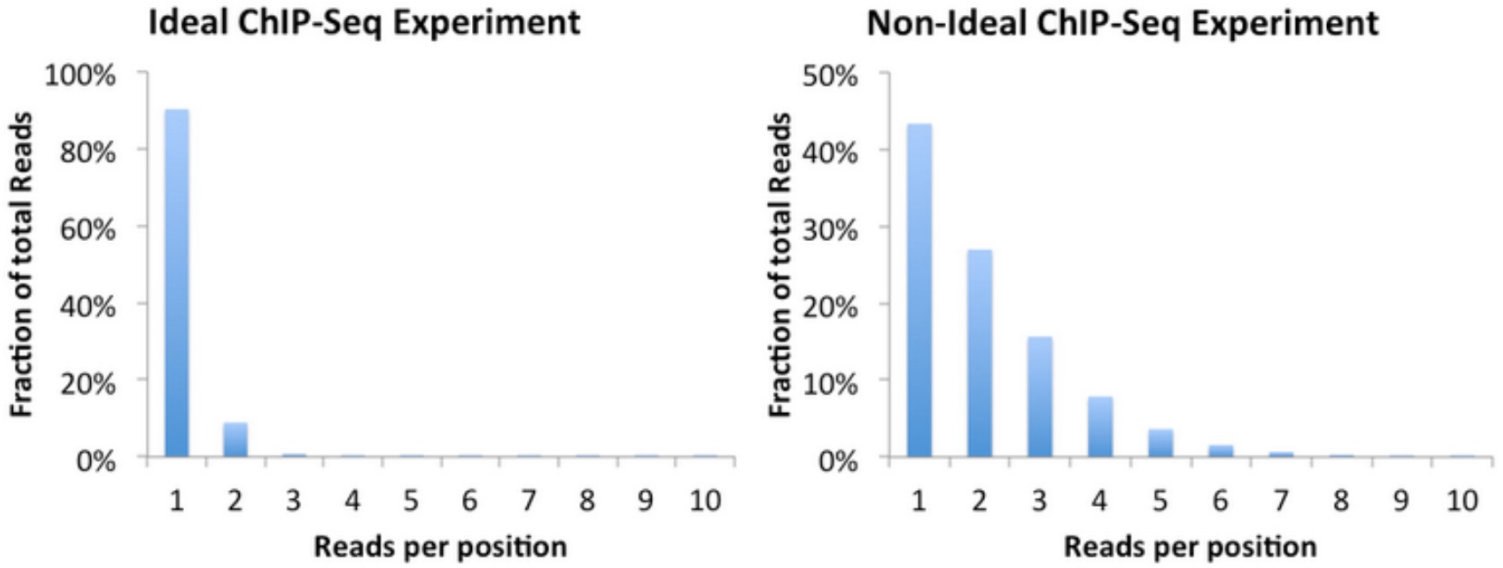


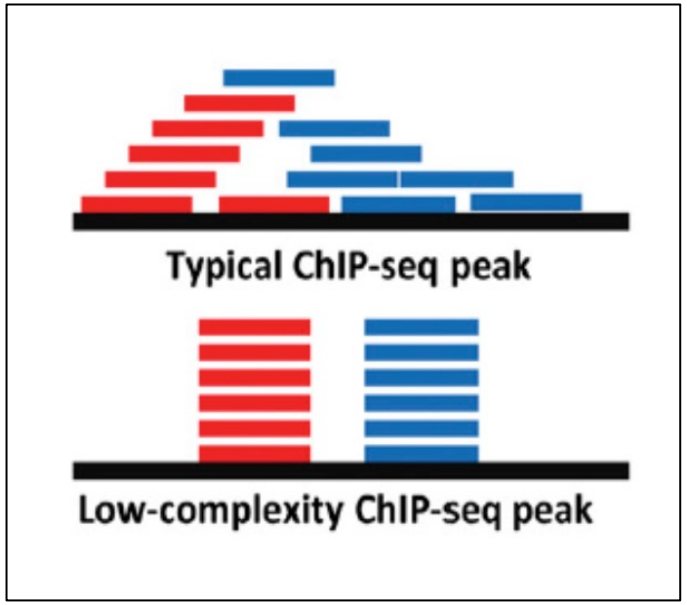
Complexity of library

```
[zemke@n7420 flashscratch]$ bowtie2 -q -p 4 -k 1 --local
mso_k18_chip.fastq > dmso_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
```

Clonal Tag Distribution

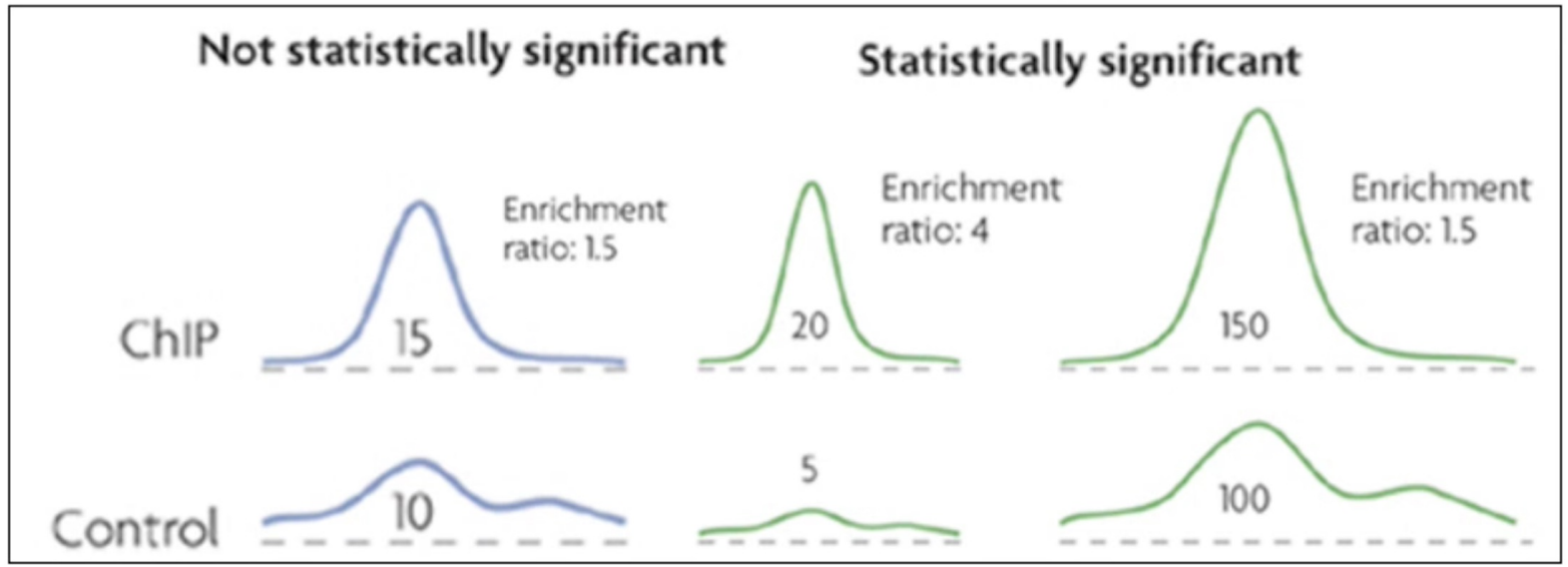
tagCountDistribution.txt - File contains a histogram of clonal read depth, showing the number of reads per unique position. If an experiment is "over-sequenced", you start seeing the same reads over and over instead of unique reads. Sometimes this is a sign there was not enough starting material for 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.





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 maxium tags per bp.

Peak calling: detect regions of enrichment



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

Steps:

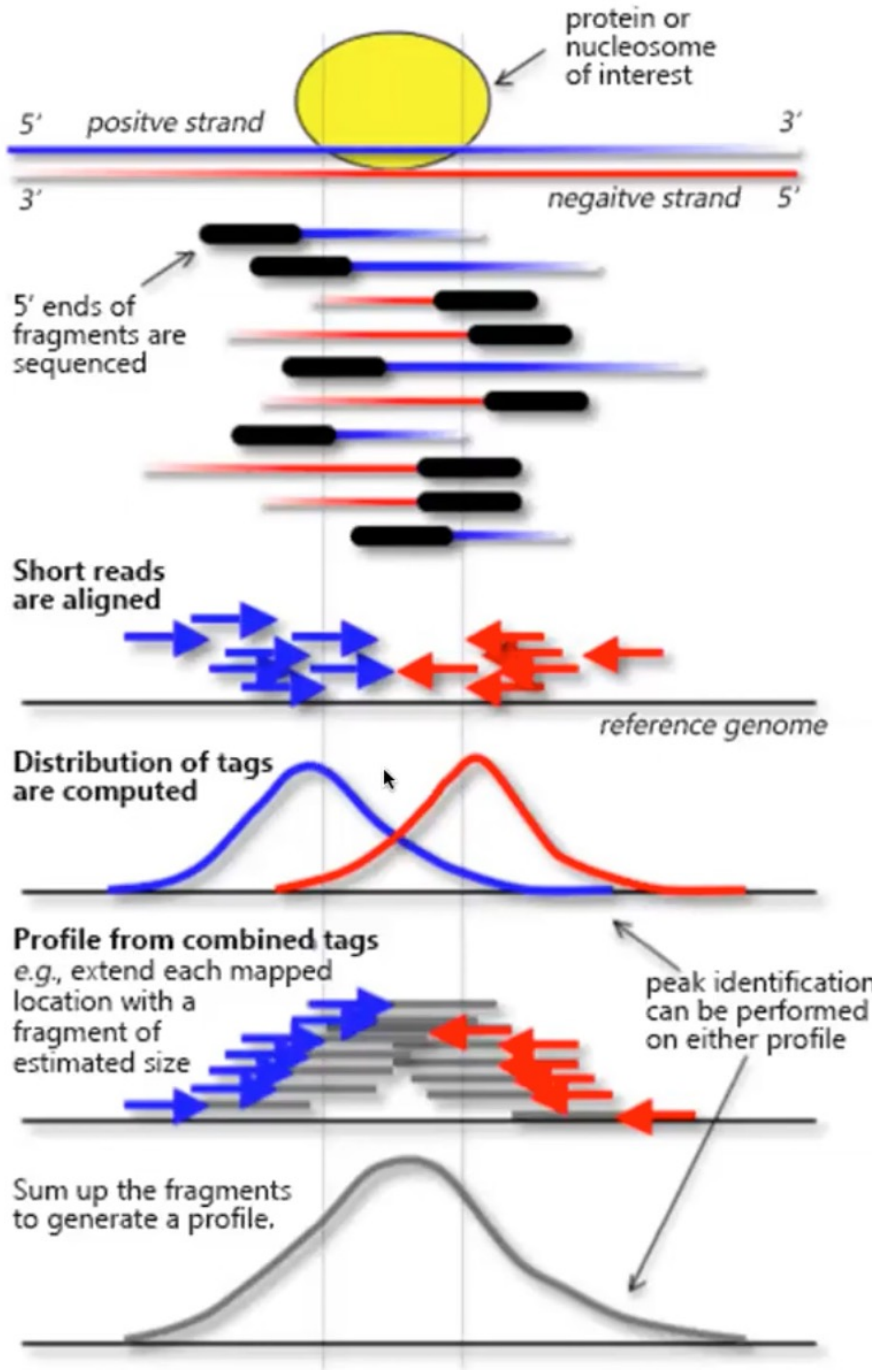
1. Estimate fragment-length *f* using strand cross-correlation analysis
2. Extend each read from 5' to 3' direction to fragment length *f*
3. Sum intensity for each base in 'extended reads' from both strands
4. Perform same operation on input-DNA control data (correct for sequencing depth differences)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

https://www.youtube.com/watch?v=XWcWn8dt4c8

MACS

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

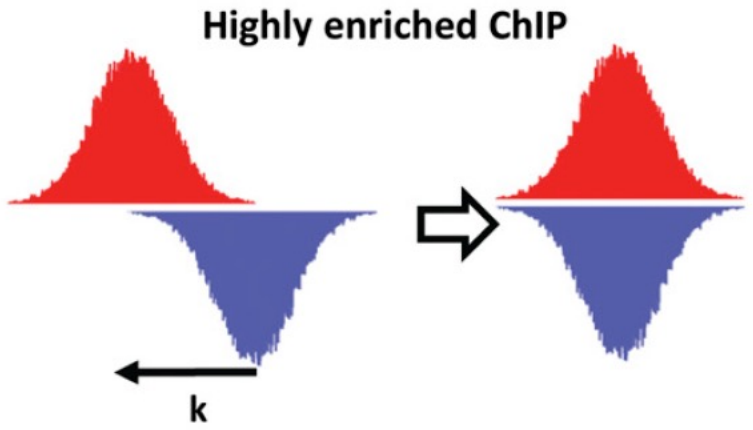
Model is for single end sequencing.
For paired end, modeling step is ignored.

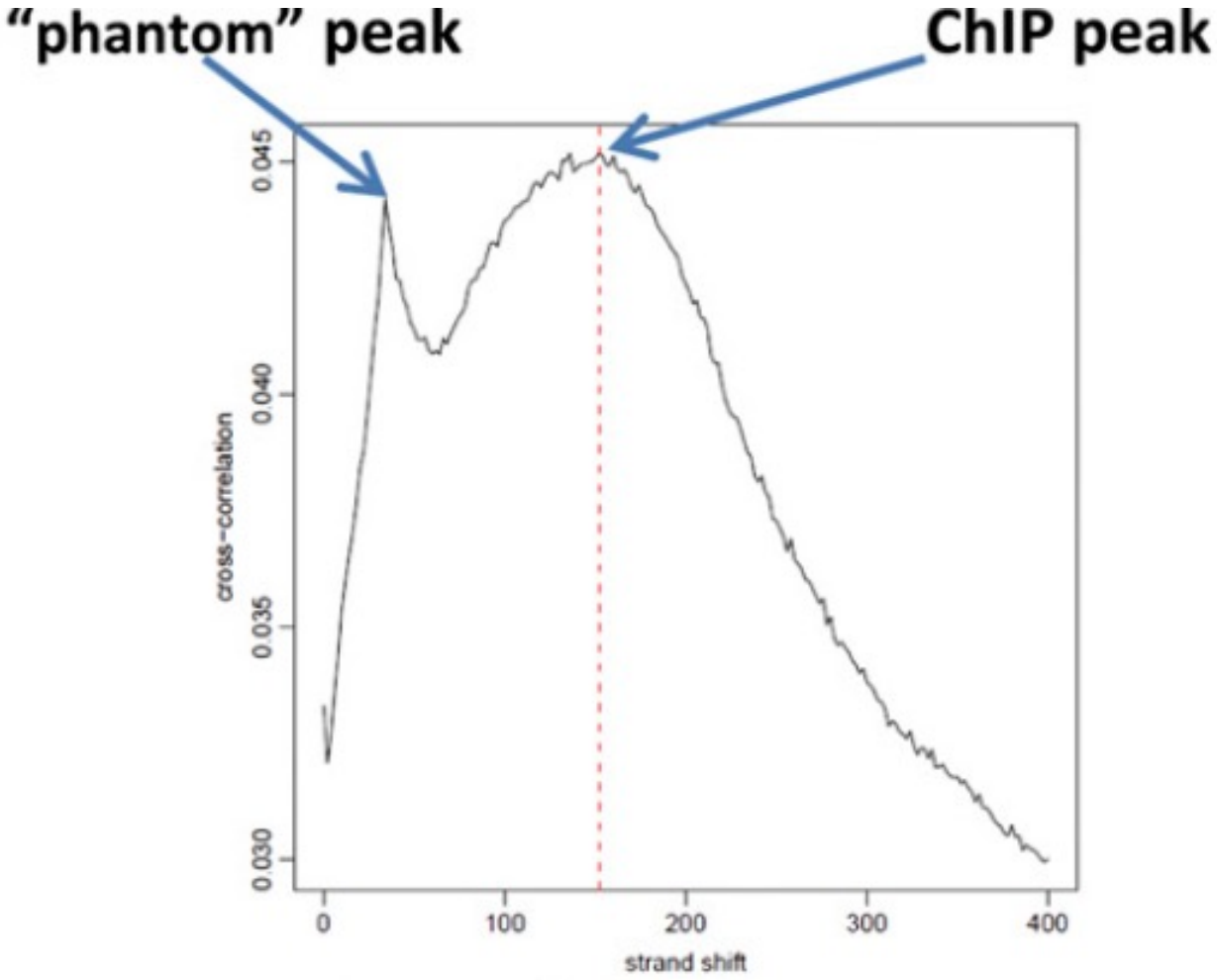


Zhang et al, Genome Biol 2008
Park, Nat Rev Genet 2009

Quality controls

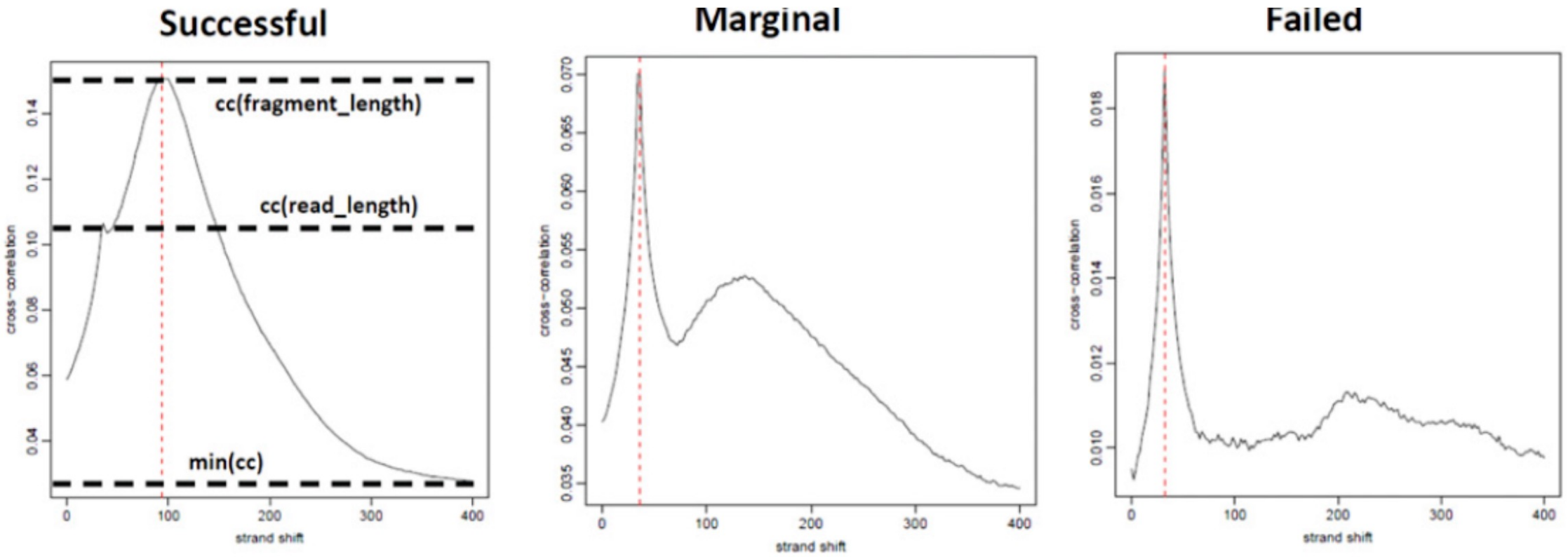
- If DNA is over-fragmented, the phantom peak will be high
- If fragmentation is aproperiate, ChIP peak will dominate the strand shift.





Quality Control

Cross-correlation



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

Normalized strand coefficient

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

Relative strand coefficient

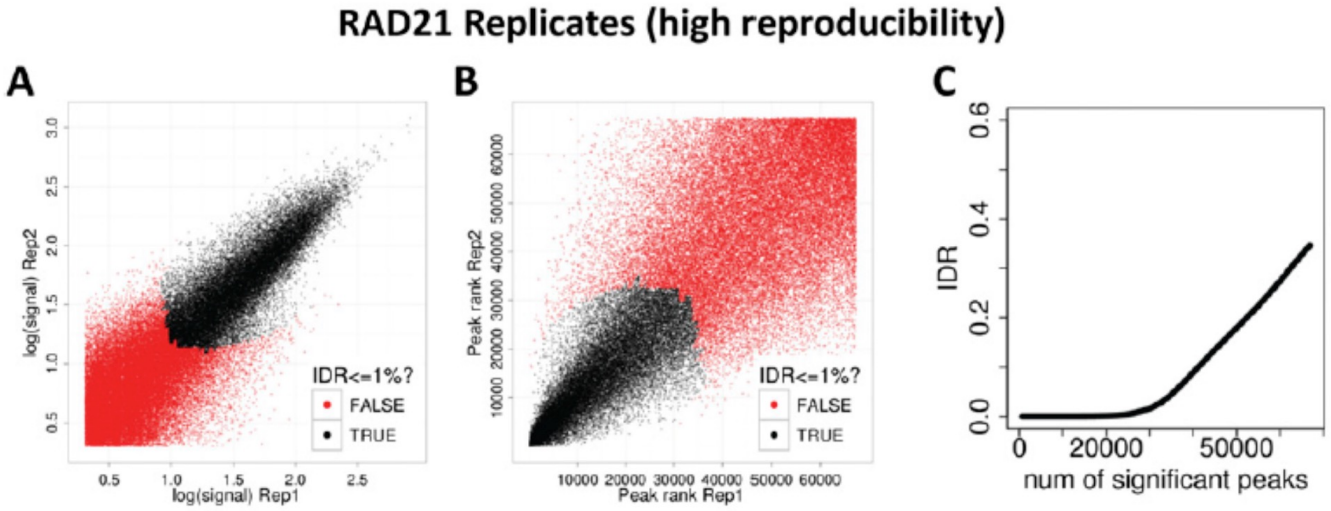
Landt et al. Genome Research (2012)

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

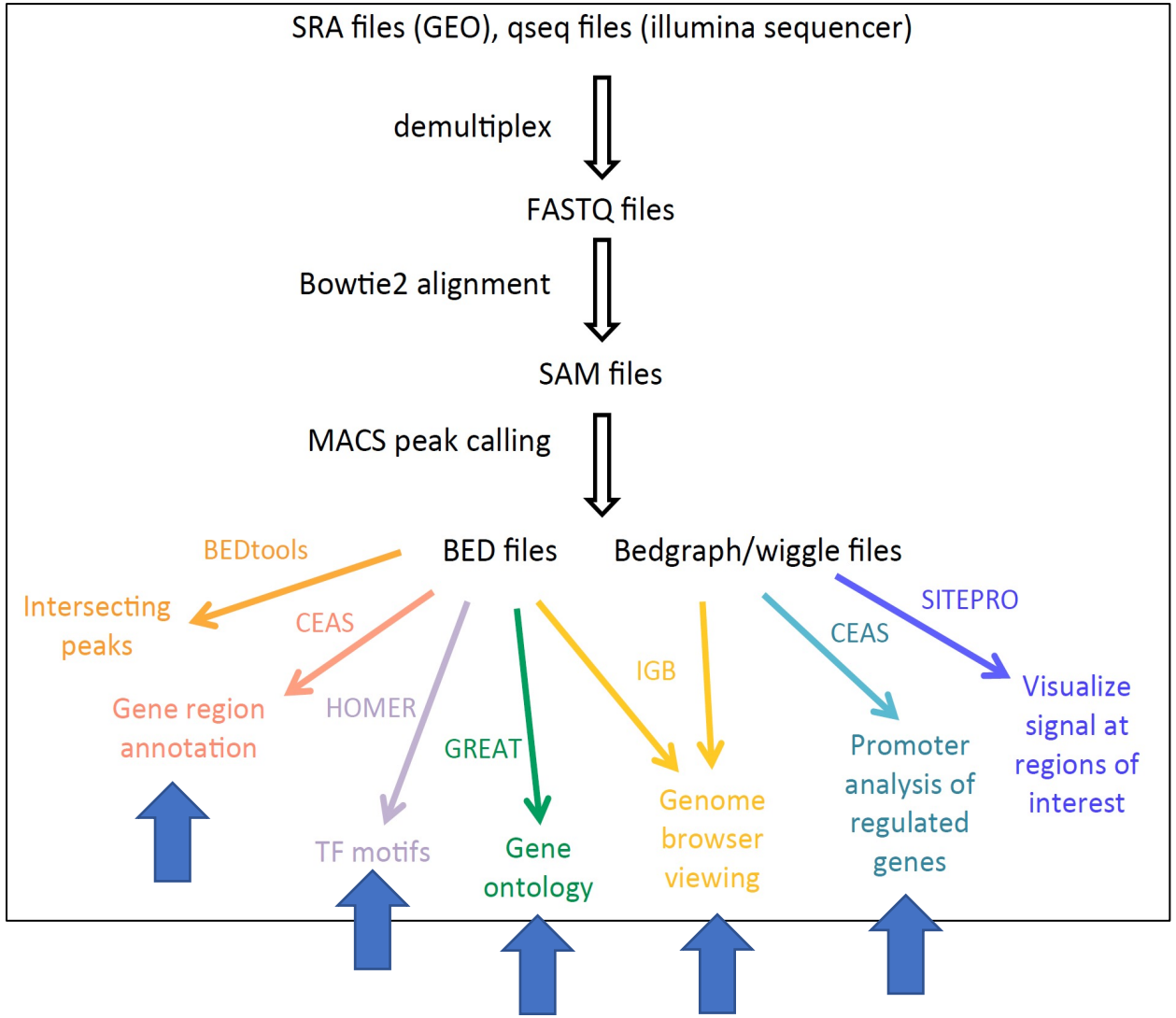


IDR < 1% = high confidence peak list

Landt et al. Genome Research (2012)

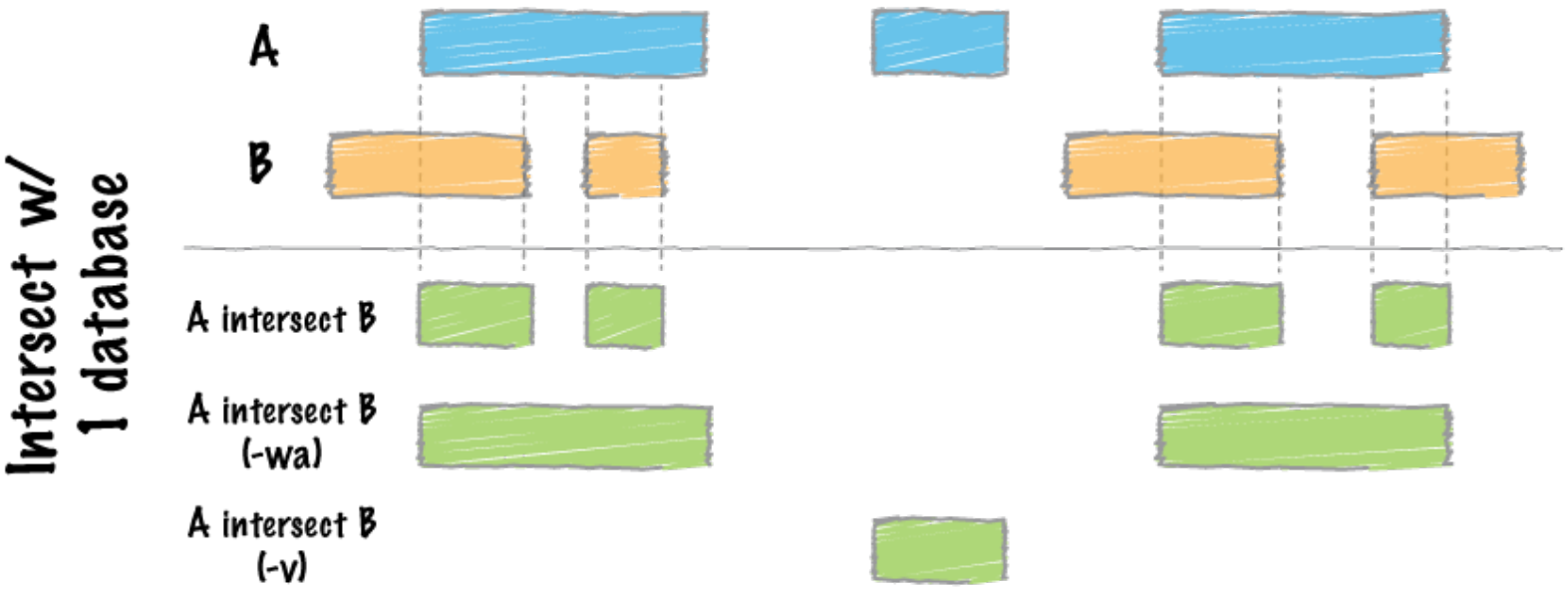
Downstream analysis

- 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
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base
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

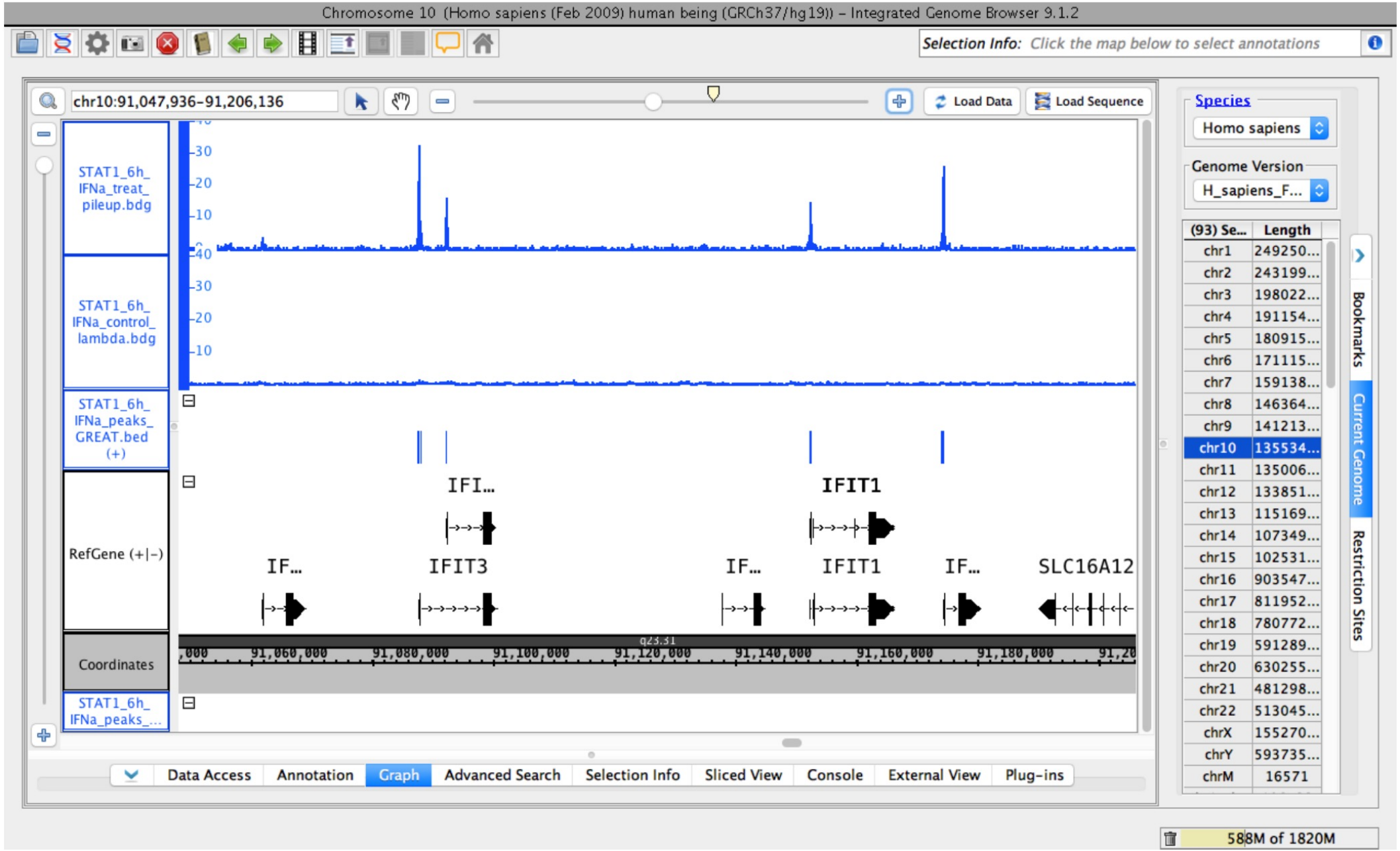
Nine additional fields are optional. Note that columns cannot be empty - lower-numbered fields mu

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



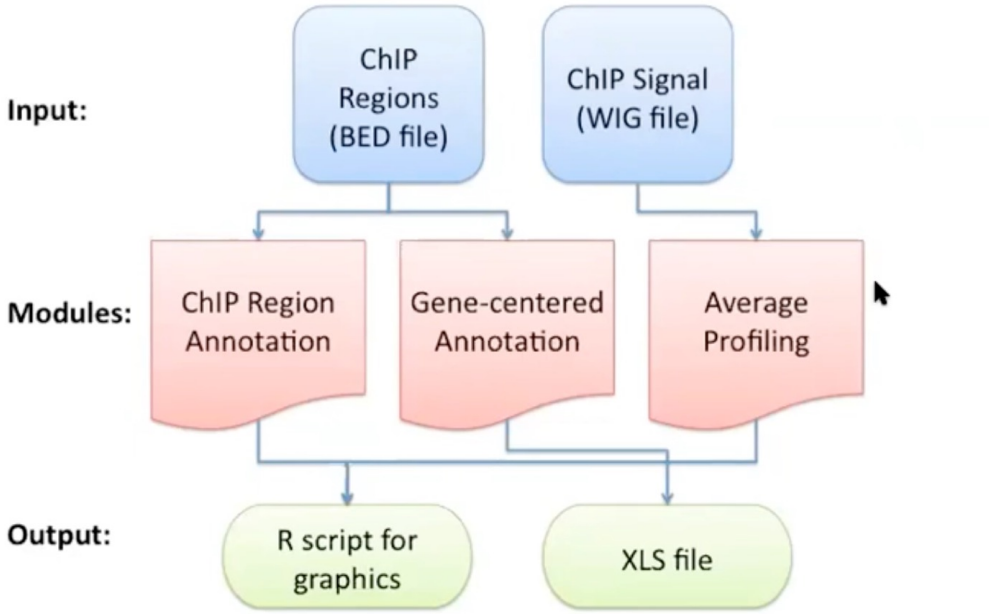
CEAS peak annotation

- Relationship between genes and peaks

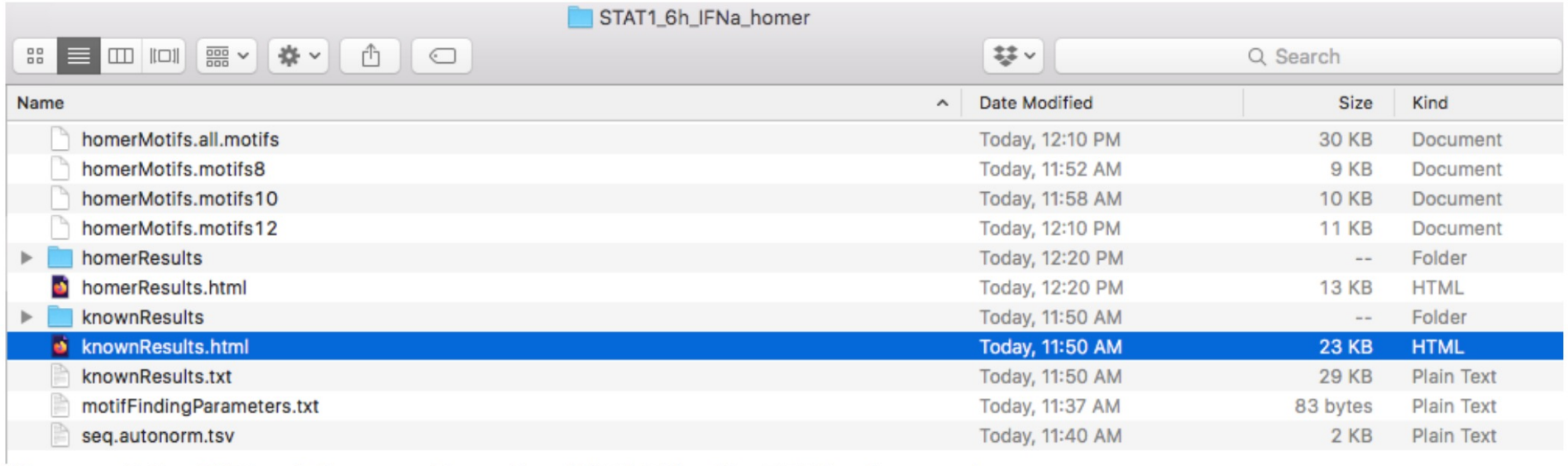
name	chr	txStart	txEnd	strand	dist.u.TSS	dist.d.TSS	dist.u.TTS	dist.d.TTS	3000bp.u.TSS	3000bp.d.TSS	1
<chr>	<chr>	<int>	<int>	<chr>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	
NR_024540	chr1	14362	29370	-	919308	NA	934316	NA	0.00	0.00	
NR_026818	chr1	34611	36081	-	912597	NA	914067	NA	0.00	0.00	
NR_026820	chr1	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



https://www.youtube.com/watch?v=JYBP5BpRfTM



Homer Known Motif Enrichment Results (STAT1_6h_IFNa_homer)

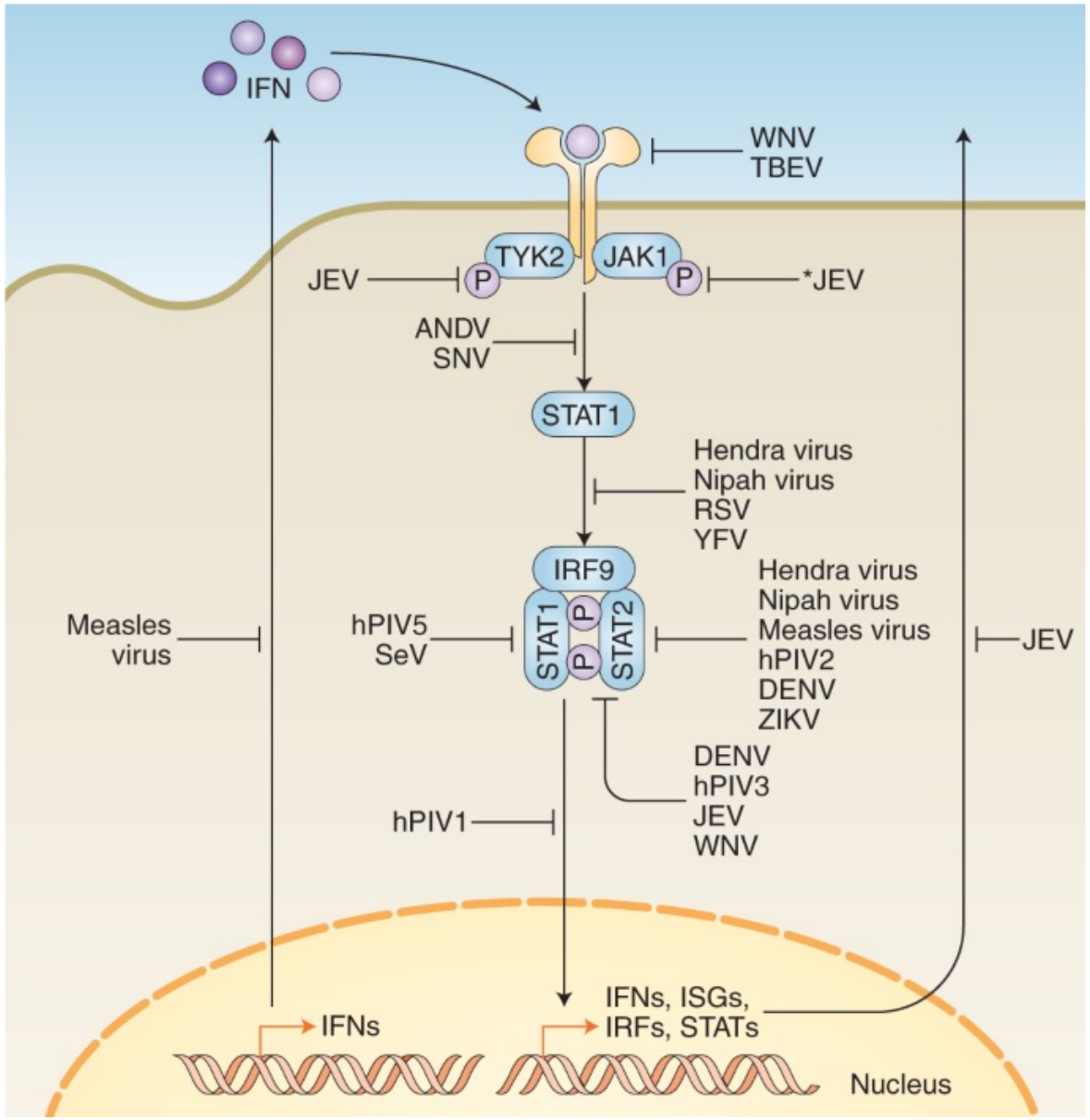
Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 1025, Total Background Sequences = 37194

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)/Erythroblas-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)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	1e-81	-1.865e+02	0.0000	352.0	34.34%	4305.2	11.54%

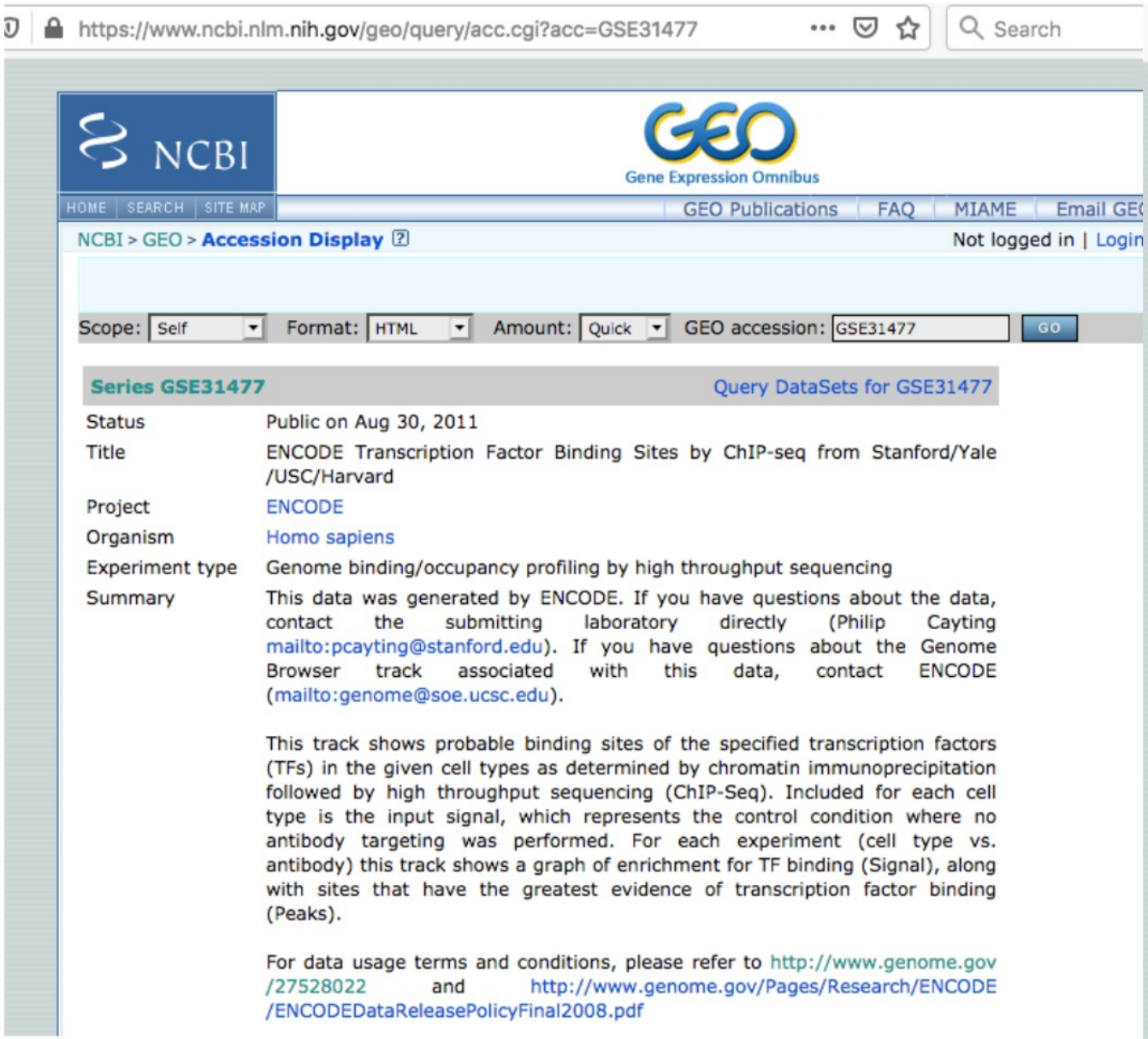
Sample data: the effect of Interferon α on STAT1 binding

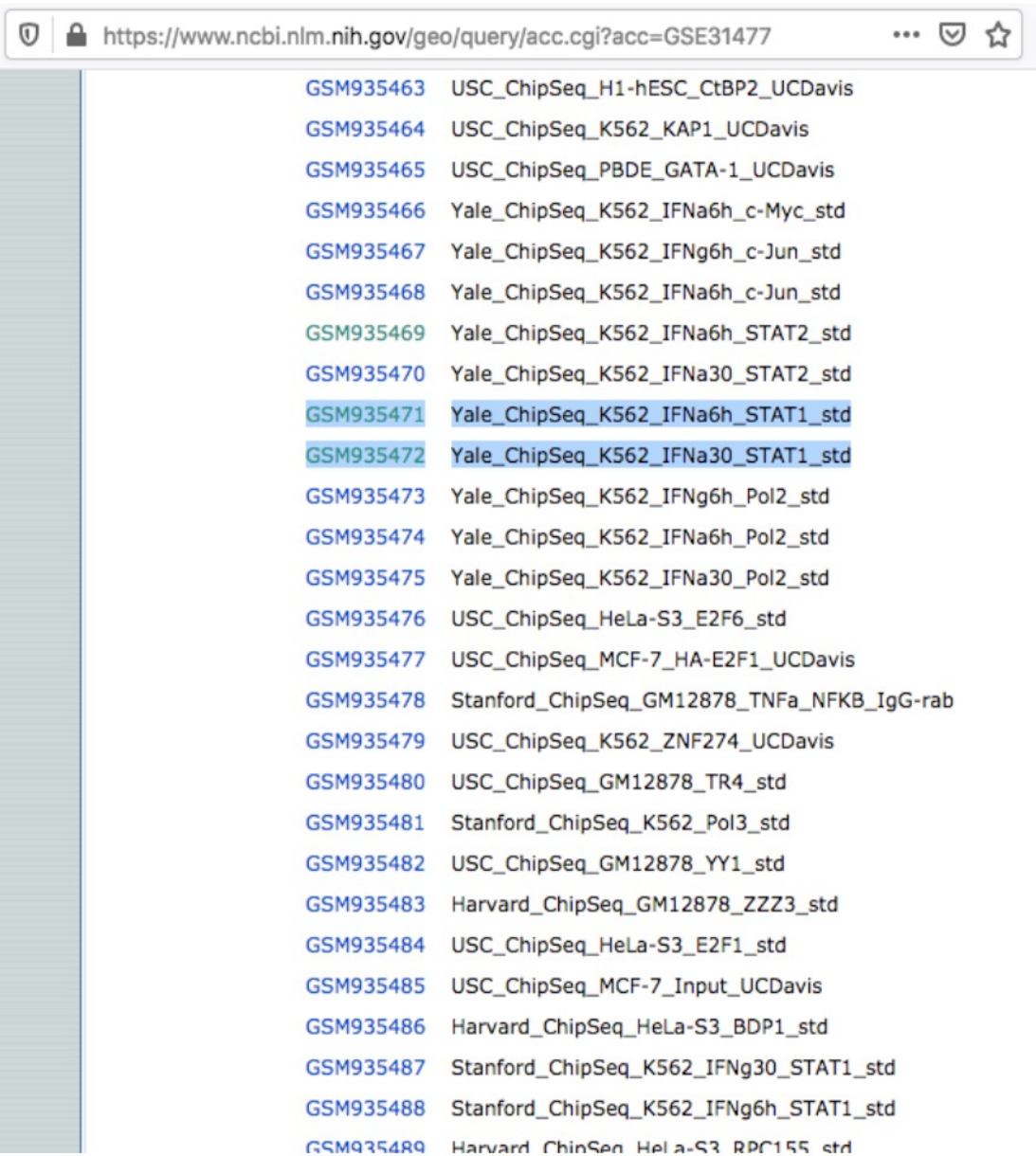
Samples

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

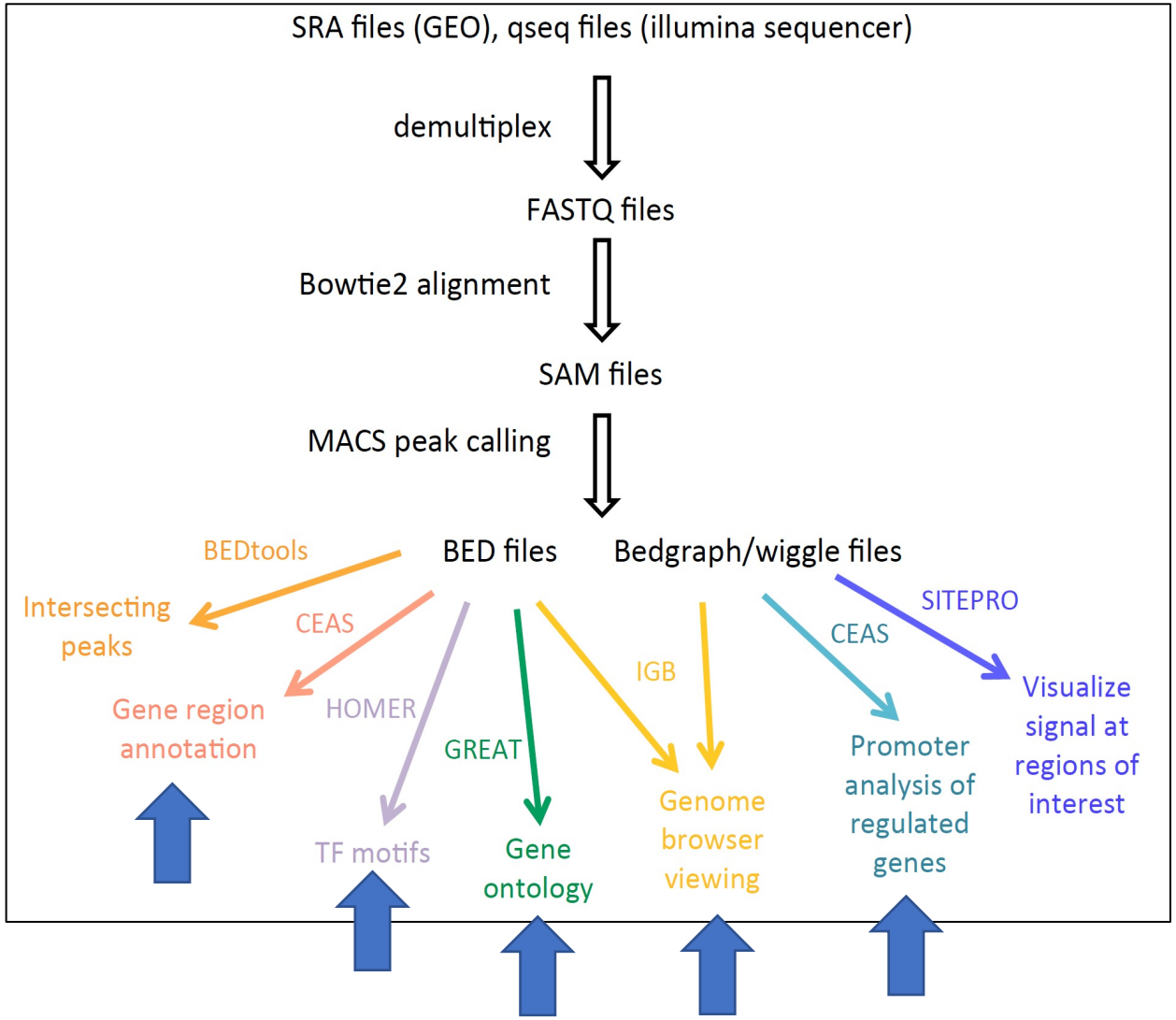


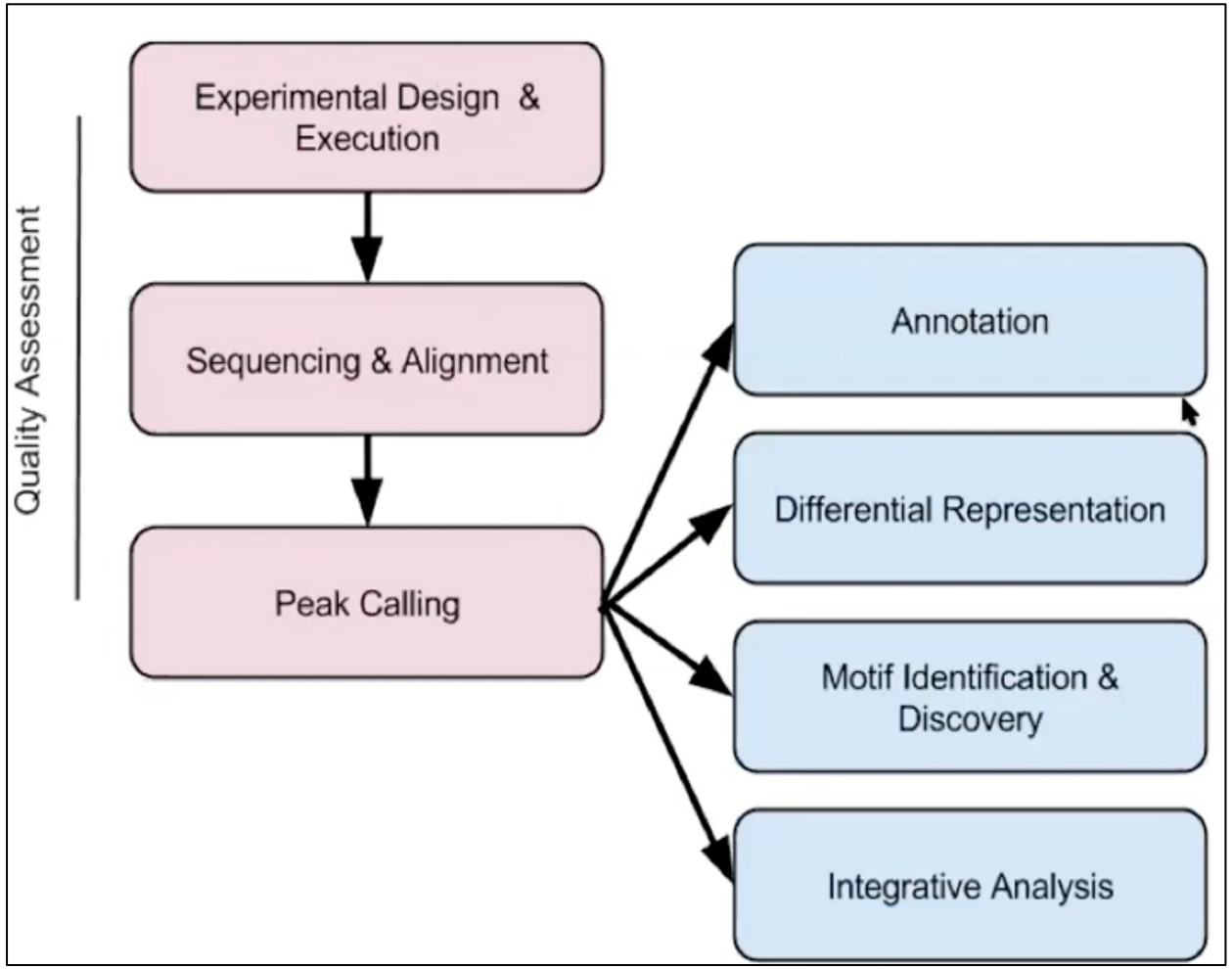
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31477





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
- single cell rna-seq
  - https://github.com/niaid/single-cell-RNA-seq
- Gene Regulatory Networks
  - https://github.com/niaid/Gene_Regulatory_Networks
- Functional annotation with clusterProfiler
  - https://github.com/zhuy16/FunctionalAnnotation_notebooks