Day2, Bulk RNA-seq & ChIP-seq

NIEHS Workshop--Analyzing NGS data

July 14, 2021

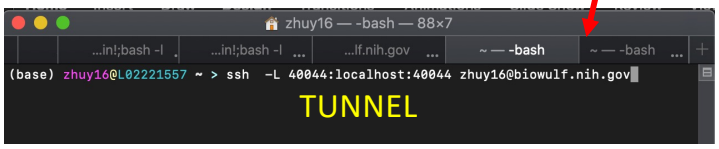
Yunhua Zhu, PhD

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.



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

Bulk RNA-seq

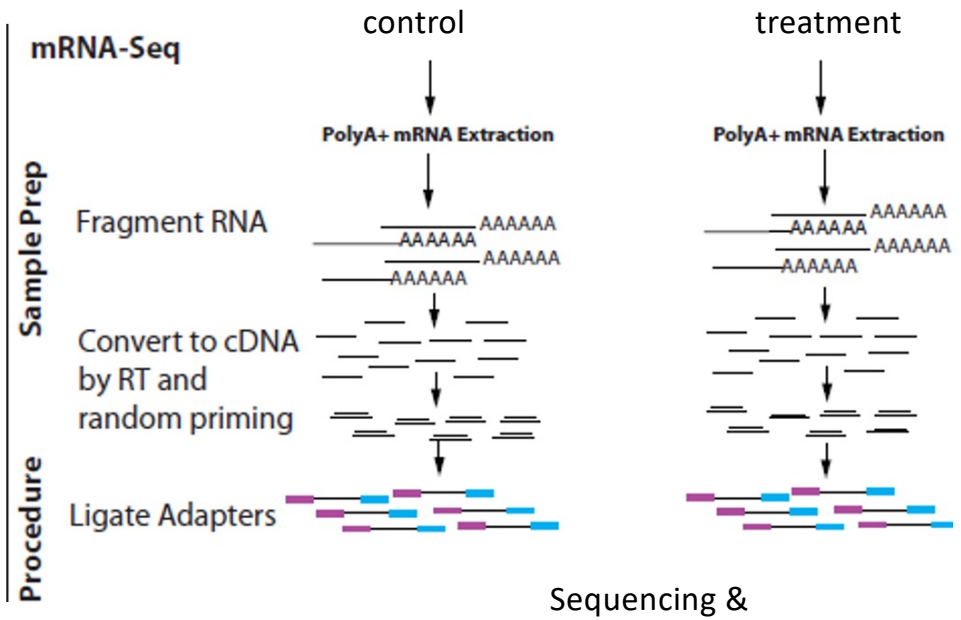
- 8-8:15AM, introduction to bulk RNA-seq
- 8:15-9AM, fastqc and alignment with RSEM, Tophat and visualize the output.
- 9-9:30AM, normalization and differential analysis with DESeq2
- (optional) cellular decomposition using ABIS and cybersort X

Chip-seq

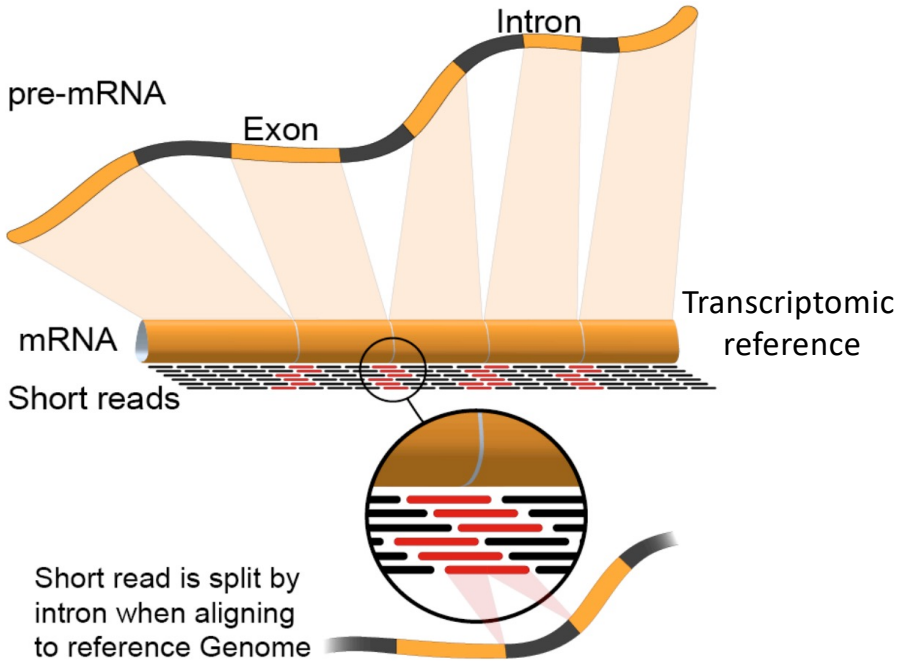
- 9:30-10AM, introduction to ChIP-seq analysis
- 10-10:15AM, download fastq, alignment to genome using Bowtie2
- 10:15-10:45AM, use MACS2 to call peaks, and use Homer to call differential peaks
- 10:45-11:15AM, use CEAS to annotate peaks and summarize statistics.
- 11:15am -11:45AM, use Homer to study TF binding site analysis. And use GREAT for gene ontology.

Review & Conclusion

Part I, RNA-seq analysis

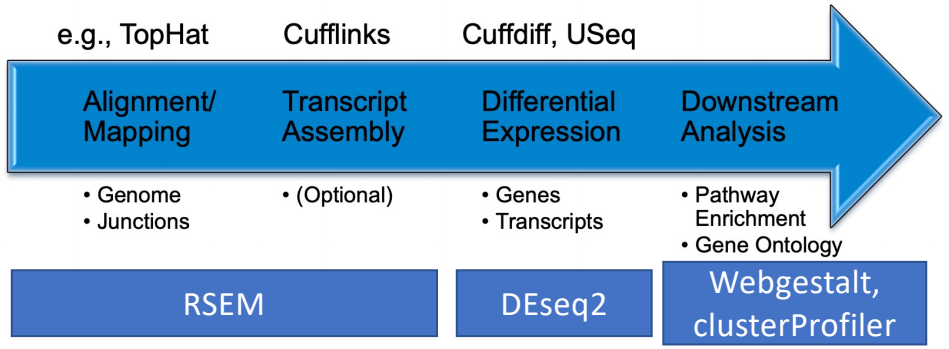


Sequencing & preprocessing



Steps

- Quality filtering
- Cut adaptors
- Alignment
- Differential expression



Bowtie

Fast!

Good for ChIP-seq and other counting-type data

Tophat

Fast (Bowtie-based)

Good for mRNA-seq, mapping novel junctions

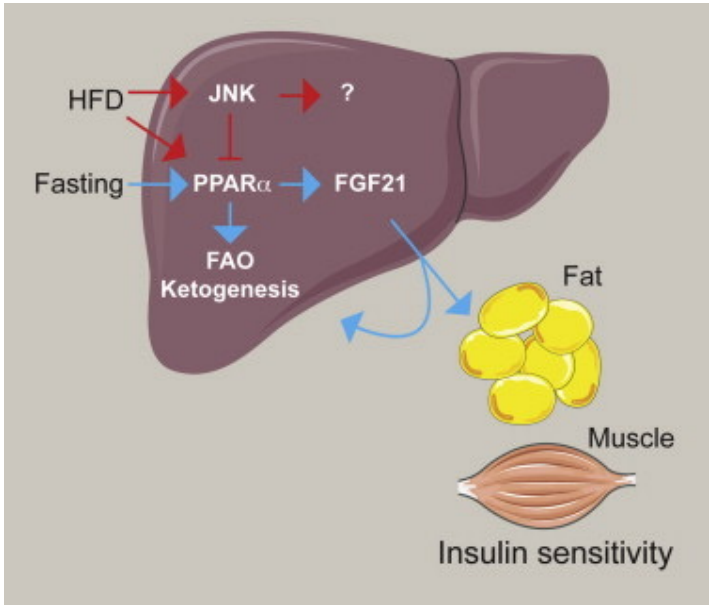
BWA

Fast

Good for variant analysis, gapped alignment

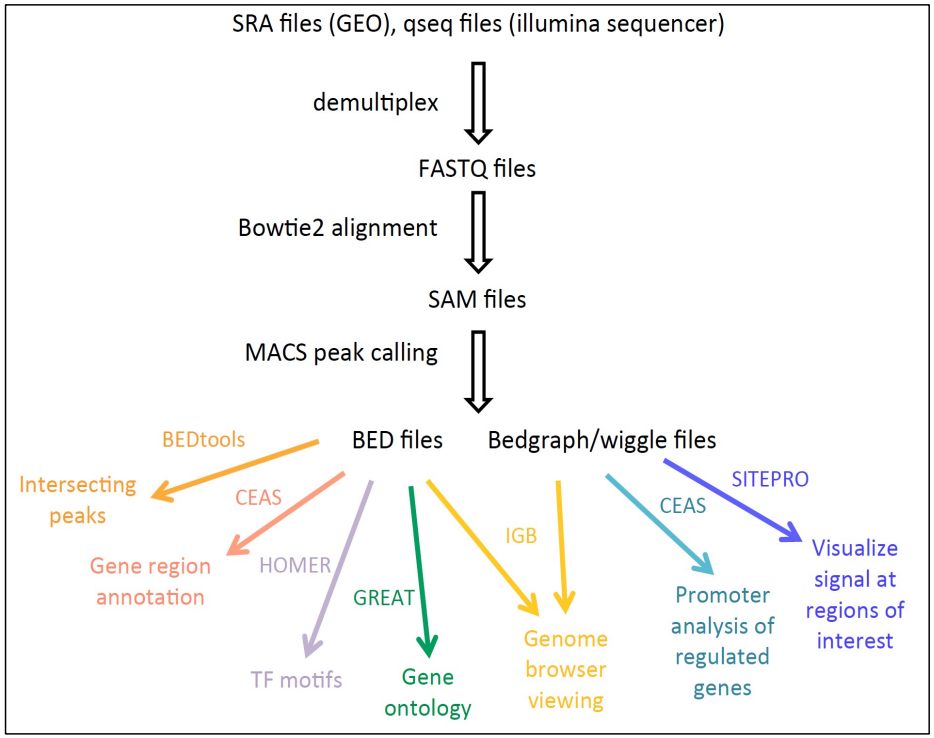
Demo on RNA-seq

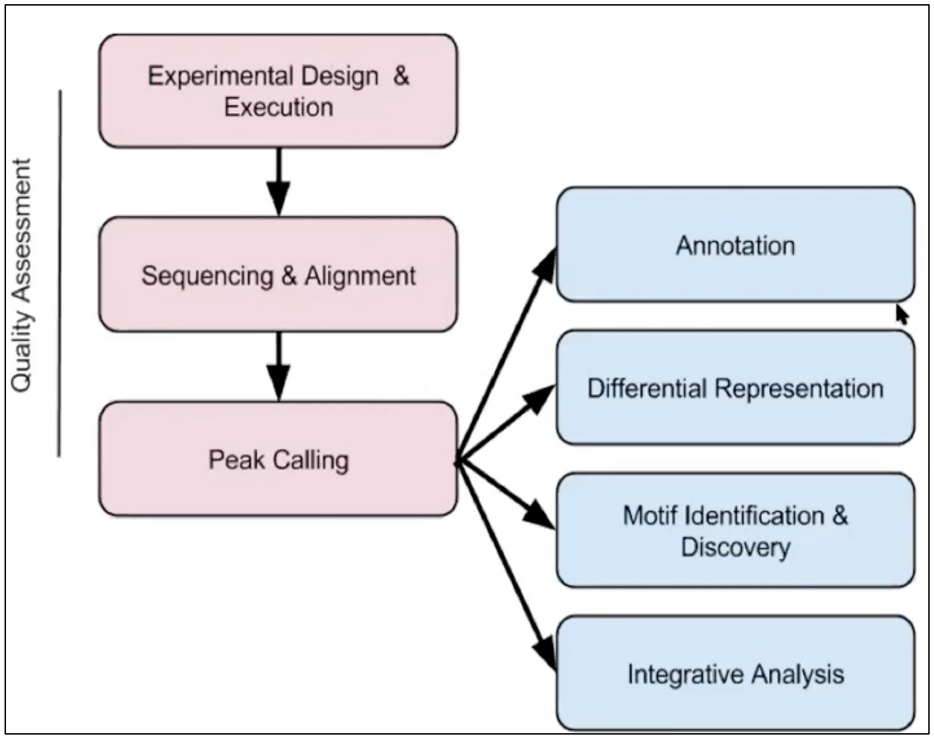
- Paired-end sequencing
- 6 samples
  - 3 wild type
  - 3 JNK1/2 double knockout



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

Part II, ChIP-seq analysis





Different IPs produce different signals

3 Types of peaks:

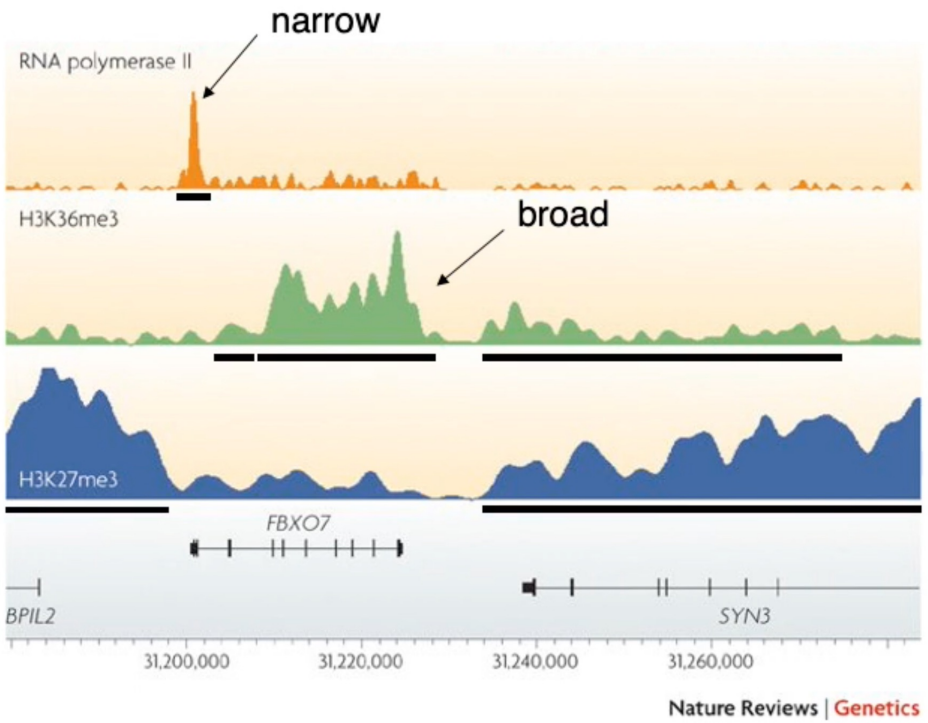
1. Sharp & narrow (100s bp)

(eg. site-specific TF)

2. Broader but defined (kb)

3. Very broad (regional, 1000s kb)

(eg. heterochromatin histone marks)



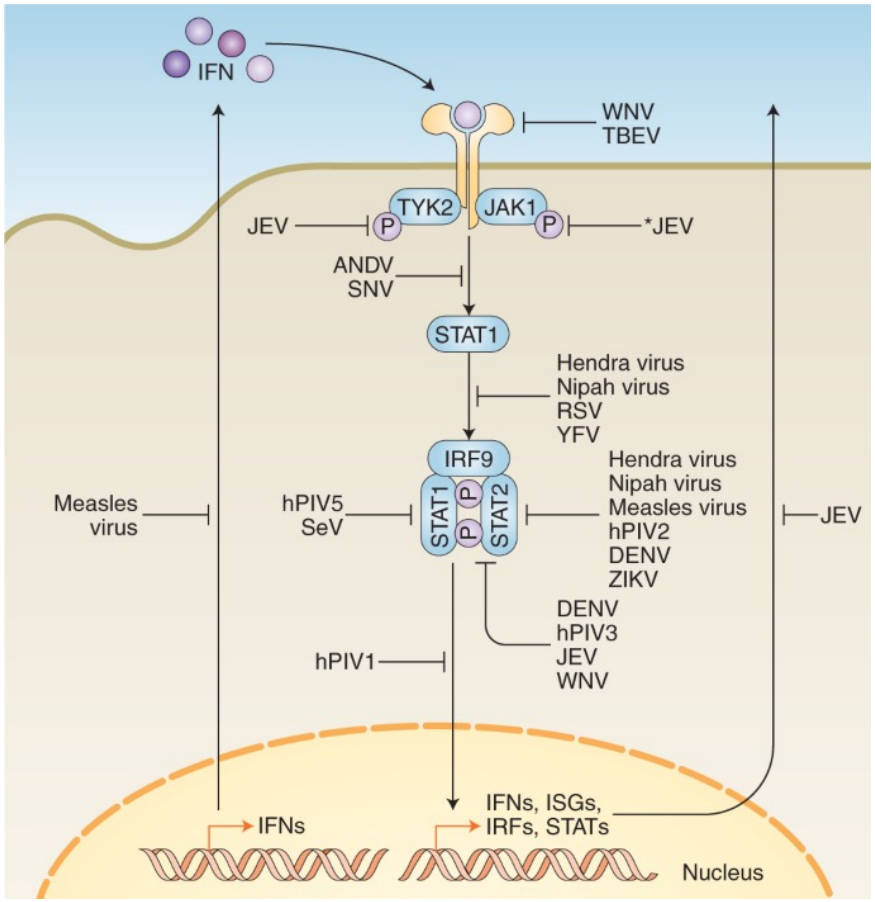
Histone Mark	Type
H3K27me3	broad
H3K36me3	broad
H3K4me3	narrow
H3K27ac	narrow

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

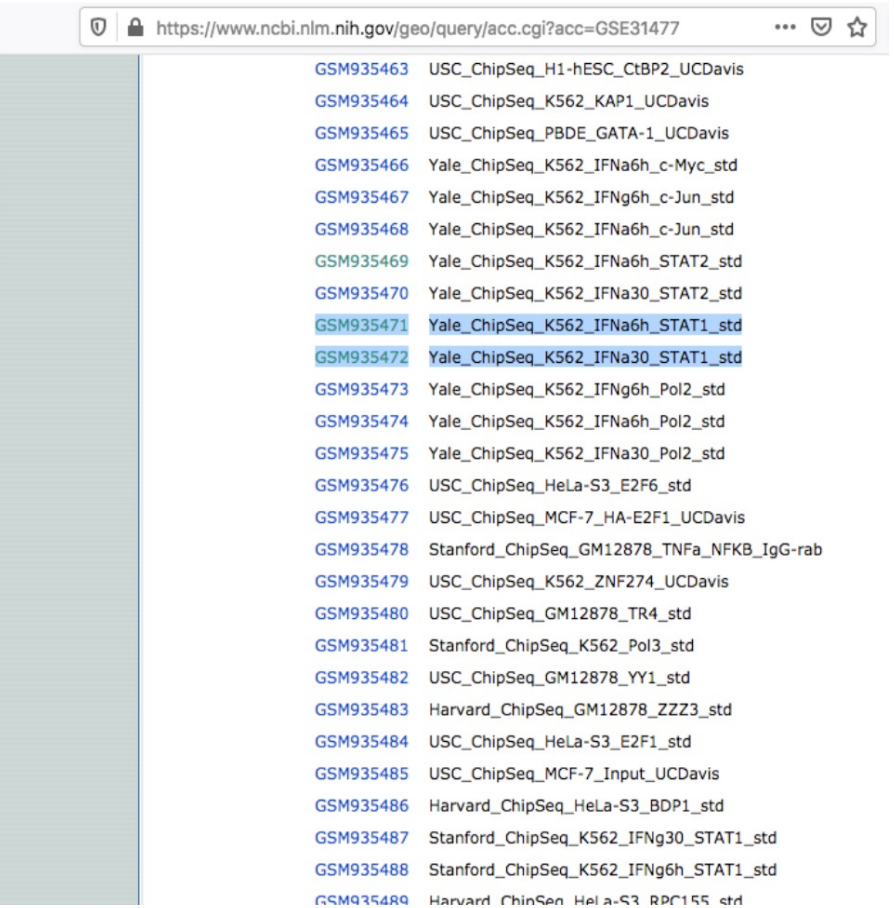
The effect of Interferon alpha on STAT1 ChIP

Samples

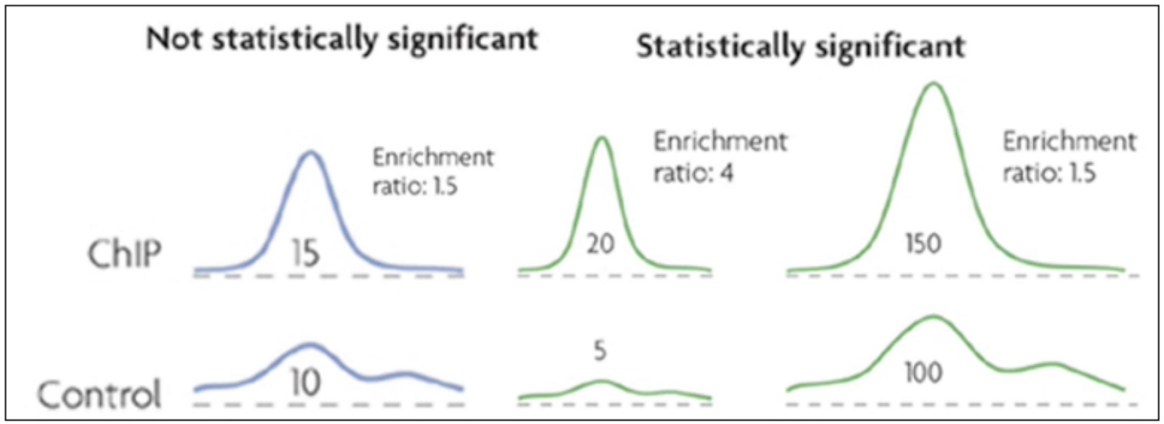
- STAT1 30 mins
- STAT1 6 hours
- Input (INP) 30 mins
- Input (INP) 6 hrs



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



Peak calling: detect regions of enrichment



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

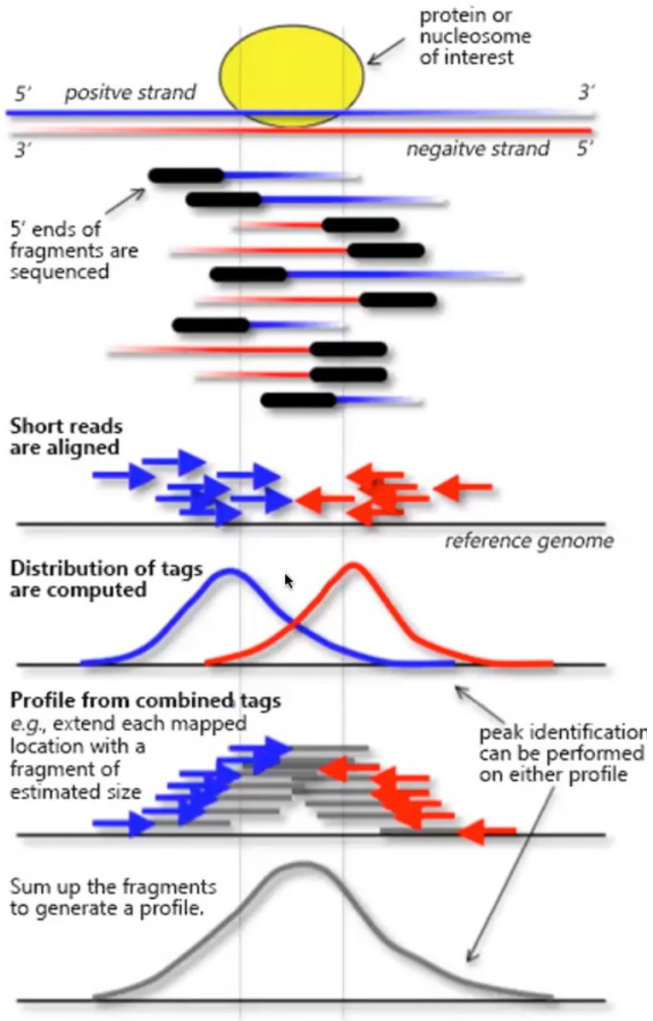
Steps:

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

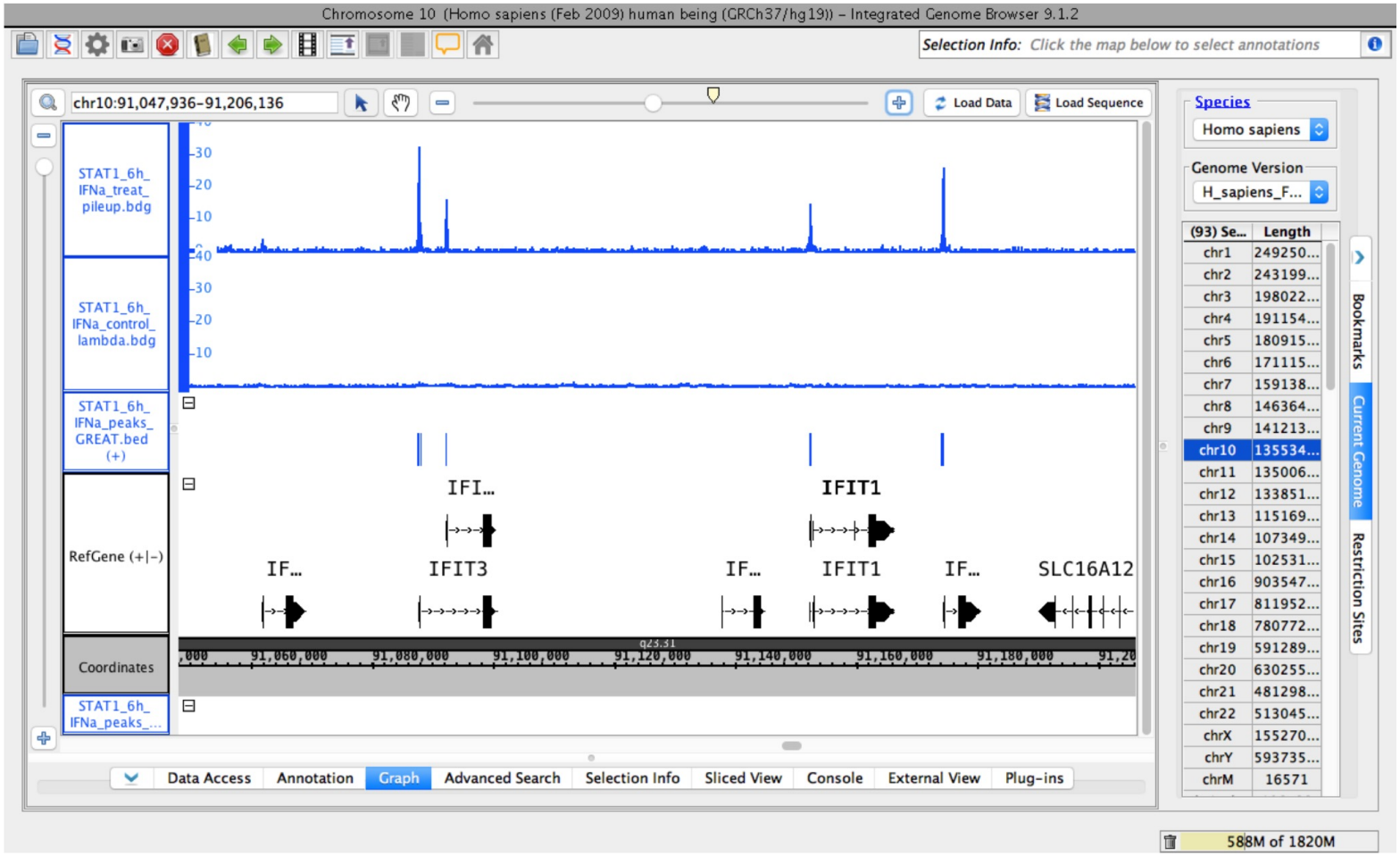
MACS

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



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

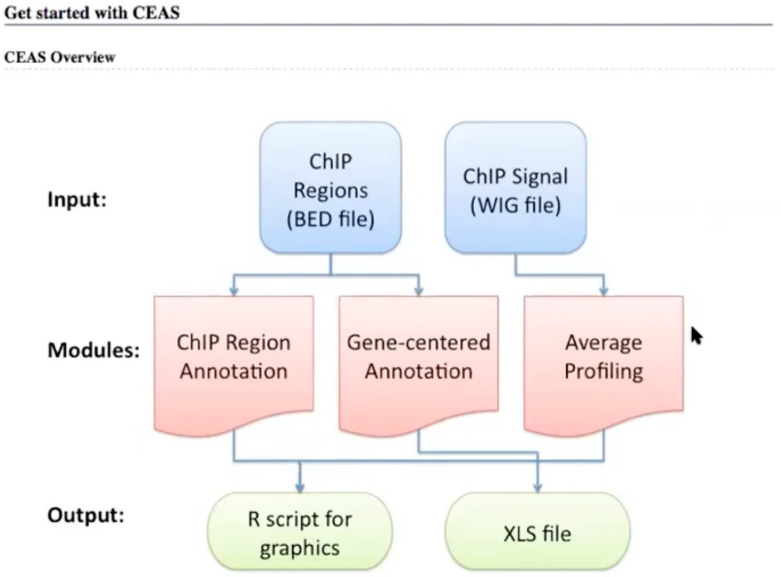
Genome Browser Viewing



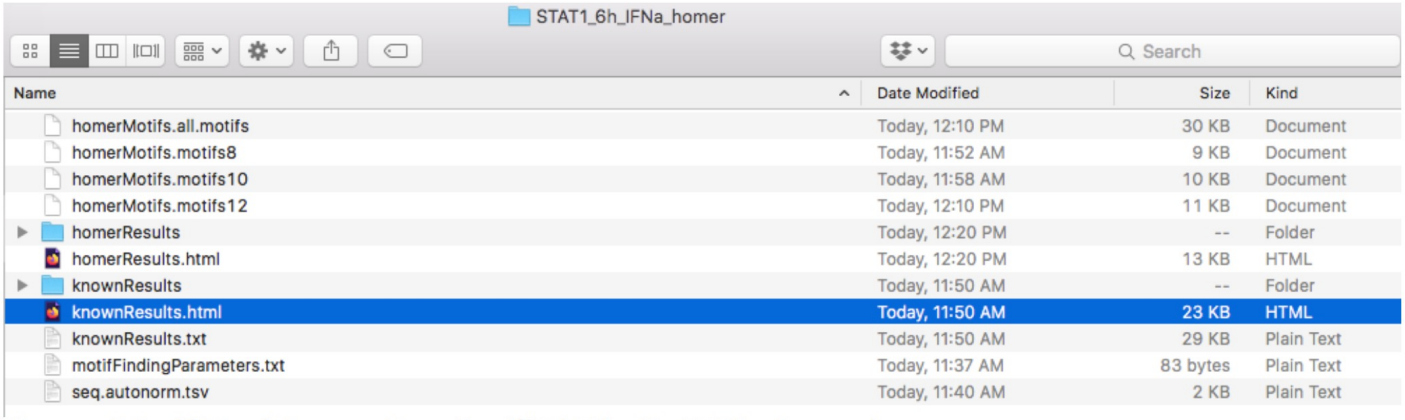
CEAS peak annotation

- 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	



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

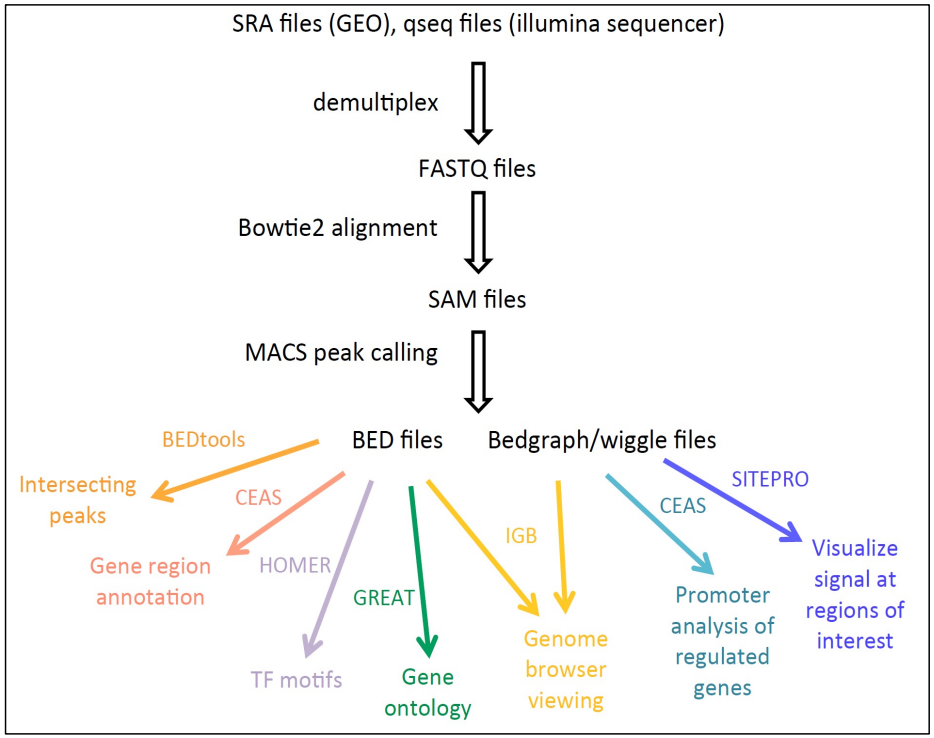


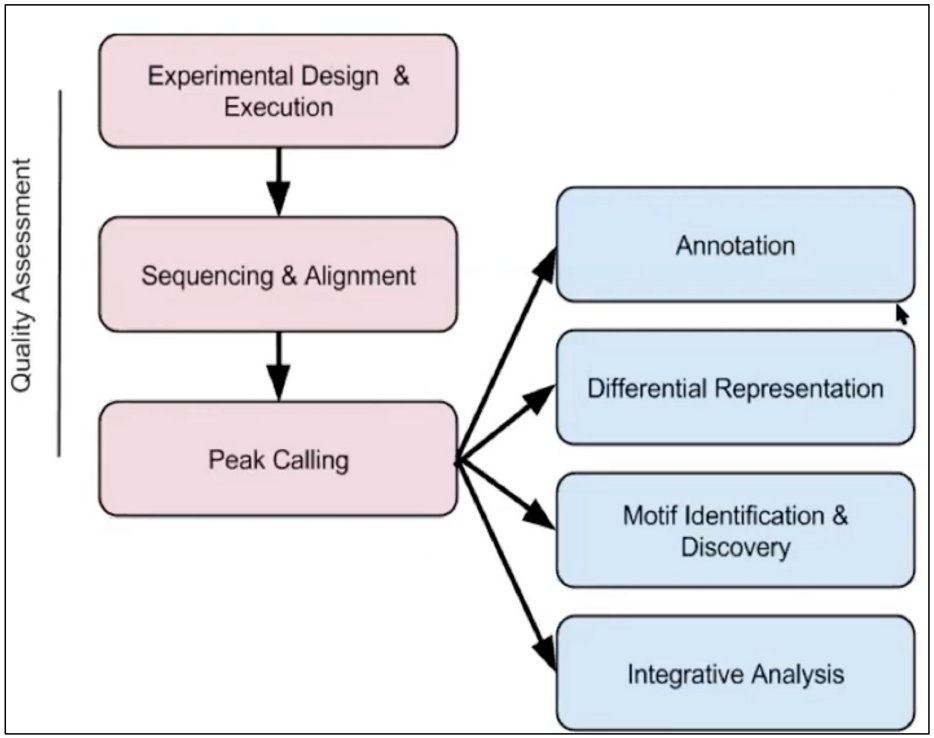
Homer Known Motif Enrichment Results (STAT1_6h_IFNa_homer)

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%

Demo – go to the jupyter lab





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