

# 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](mailto: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.

The terminal window shows the following sequence of events:

- The user logs in to the biowulf node.
- The system displays a "WARNING" message about accessing a U.S. Government information system.
- The user runs `module load tmux` and `tmux new -ct 'sinteractive --mem=200g --gres=lscratch:10 --time=12:00:00 --tunnel'` to create a session named "zhu16".
- The user creates a generic SSH tunnel from their workstation to port 40044 on the biowulf node.
- The user runs `ssh -L 40044:localhost:40044 zhu16@biowulf.nih.gov` to establish the tunnel.
- The user runs `module load jupyter R/4.0.5 && jupyter lab` to start the Jupyter Lab server.
- The terminal shows the Jupyter Lab startup logs, including the creation of a notebook named "TUNNEL".
- An arrow points from the terminal window to the bottom right corner where the URL "http://127.0.0.1:40044/lab/tunnels/f4f8479814e83c5459767a7c85d866d5f8fb2bf" is displayed.

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

## Bulk RNA-seq

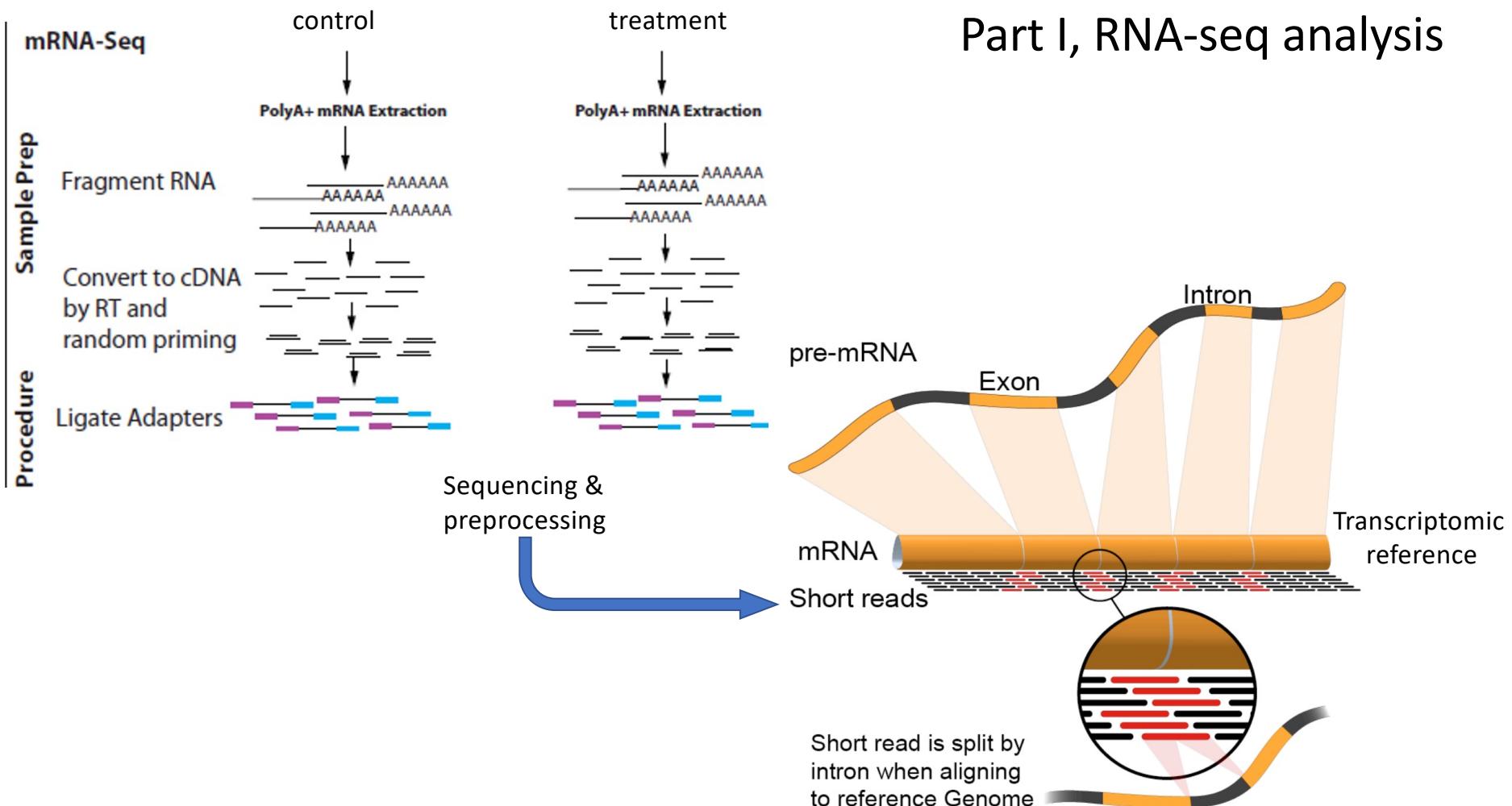
- 8-8:15AM, introduction to bulk RNA-seq
- 8:15-9AM, fastqc and alignment with RSEM, Tophat and visualize the output.
- 9-9:30AM, normalization and differential analysis with DESeq2
- (optional) cellular decomposition using ABIS and cybersort X

## 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.

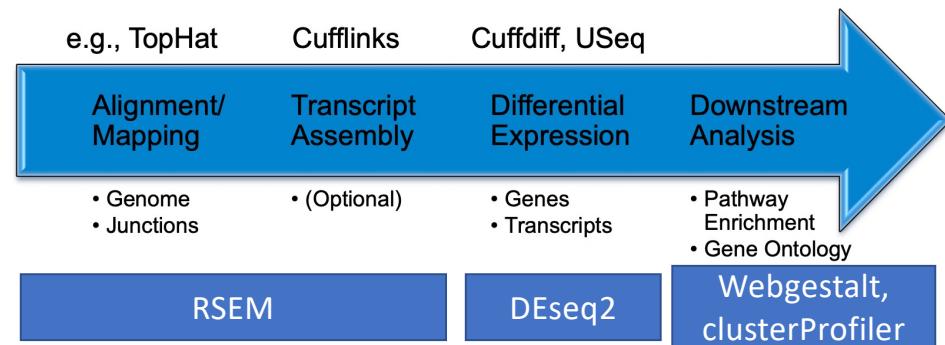
## Review & Conclusion

# Part I, RNA-seq analysis



# 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

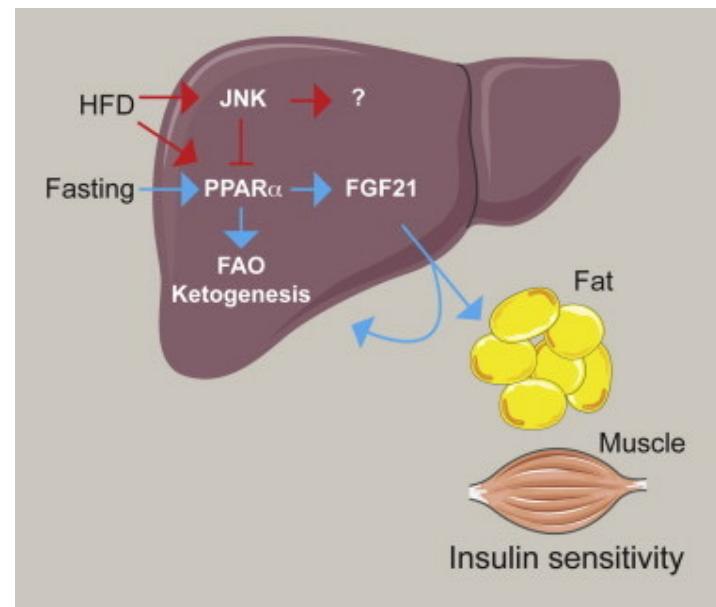
## BWA

Fast

Good for variant analysis, gapped alignment

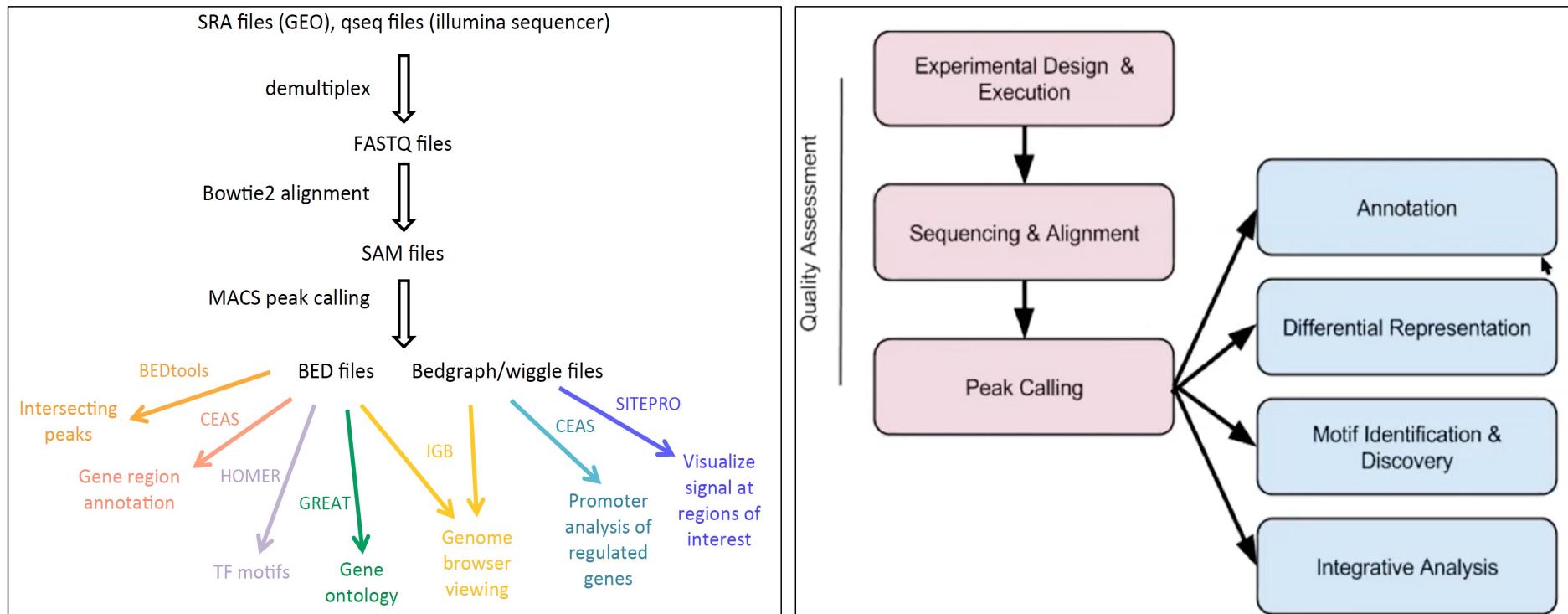
# Demo on RNA-seq

- 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

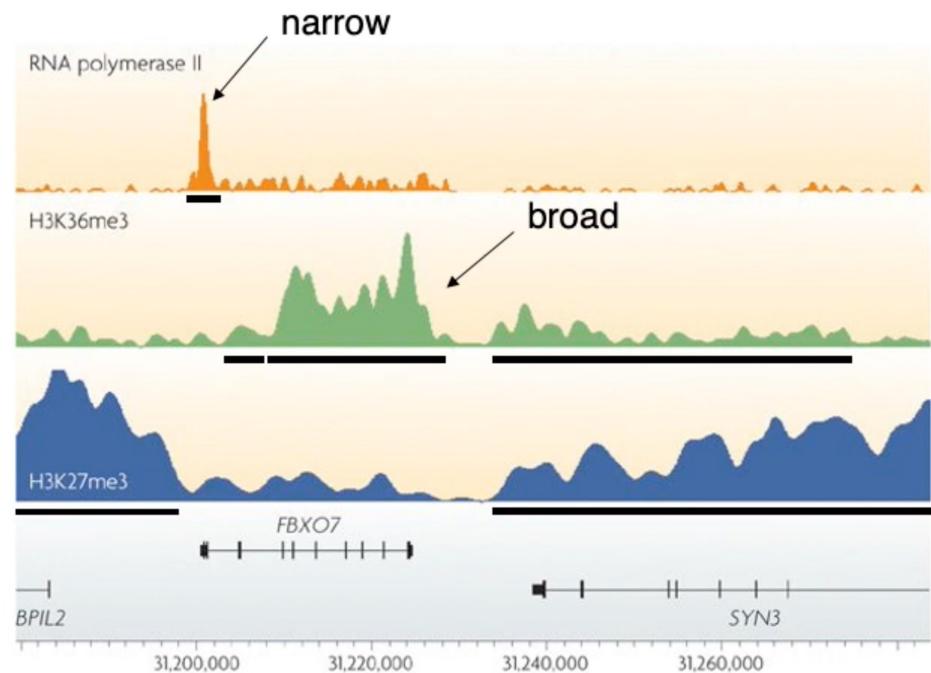
3 Types of peaks:

1. Sharp & narrow (100s bp)

(eg. site-specific TF)

2. Broader but defined (kb)

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



Nature Reviews | Genetics

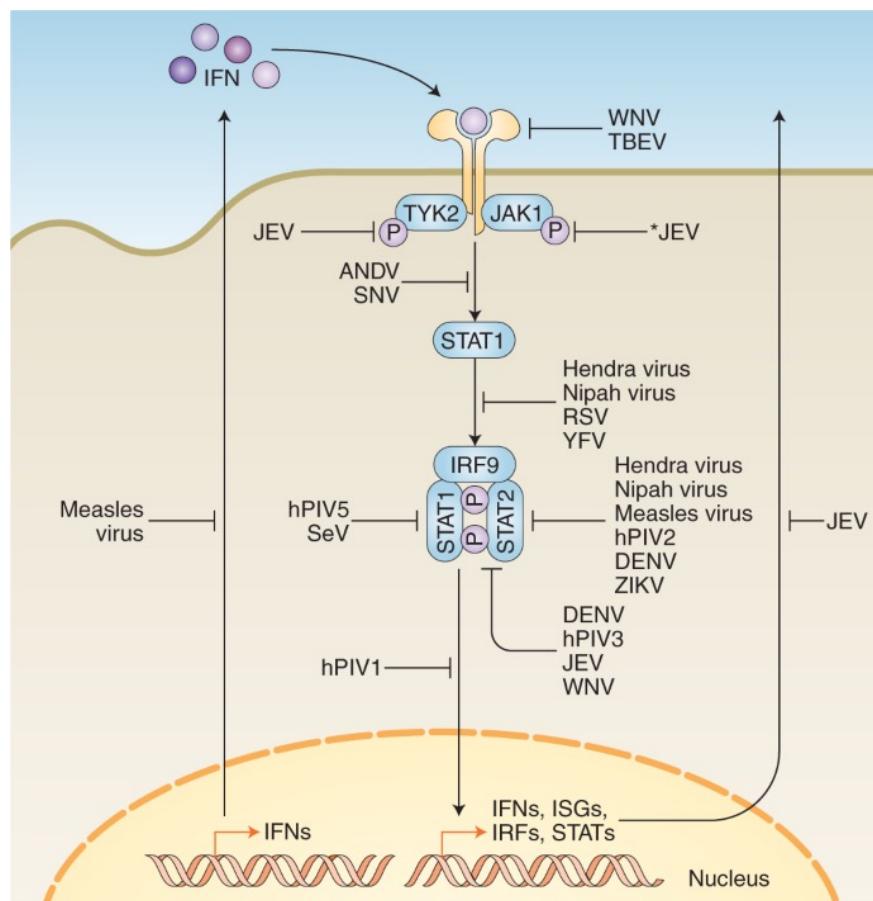
Histone Mark	Type
H3K27me3	broad
H3K36me3	broad
H3K4me3	narrow
H3K27ac	narrow

<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>

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

... 🔍 ⌂ ⌂ Search

NCBI GEO 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 Go

**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).

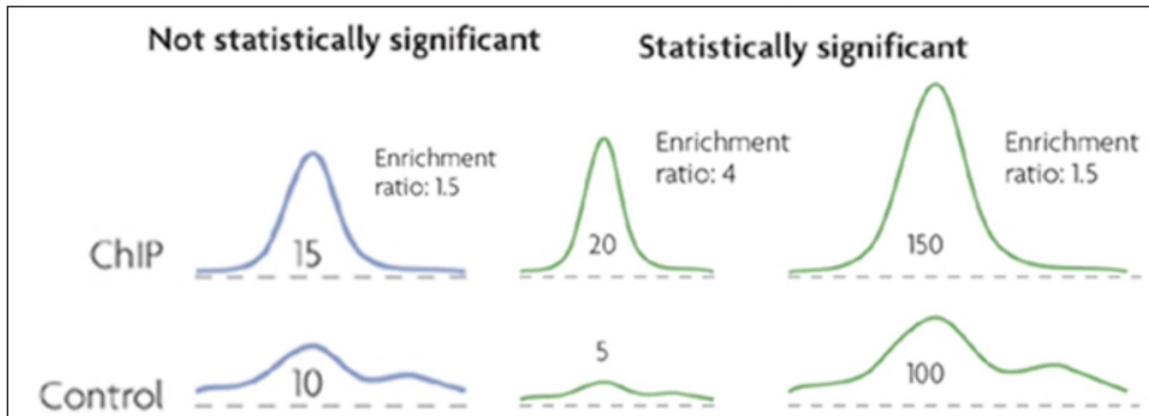
For data usage terms and conditions, please refer to <http://www.genome.gov/27528022> and <http://www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf>

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

# Peak calling: detect regions of enrichment



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

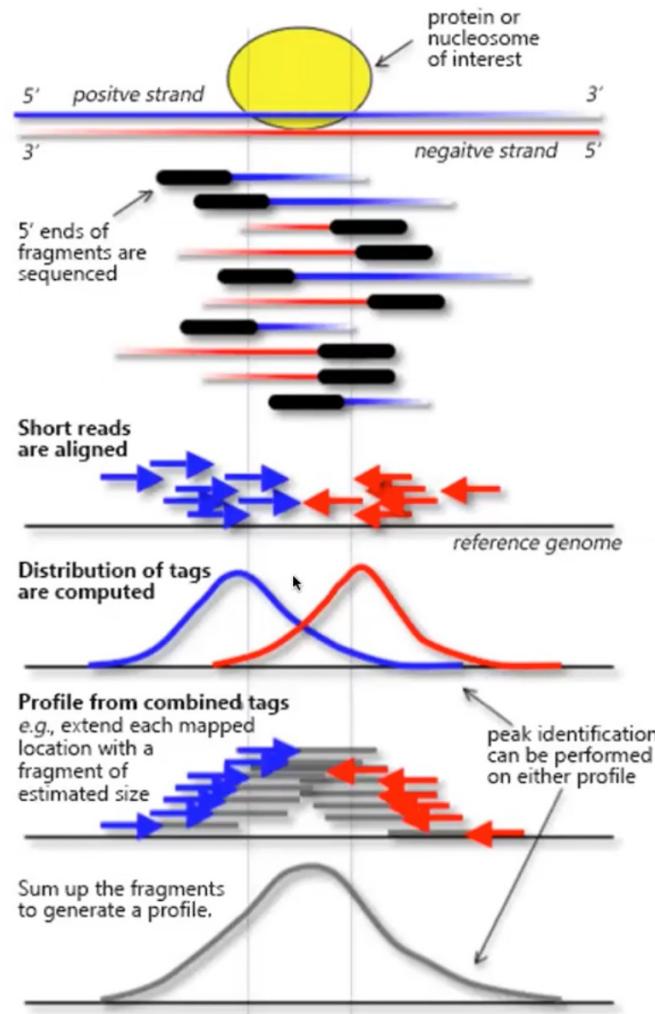
**Steps:**

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

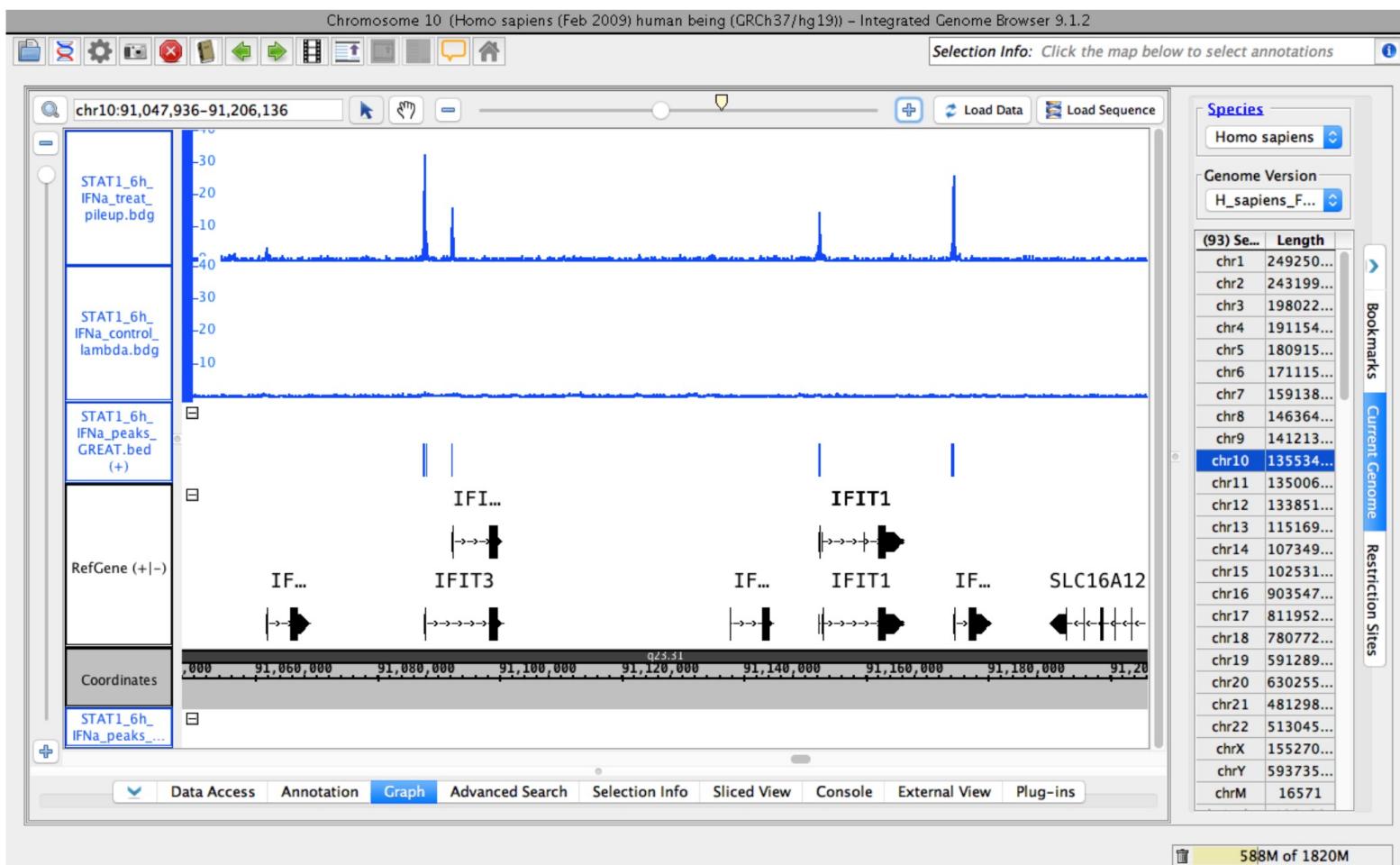
- Model-based Analysis for ChIP-Seq
- Use confident peaks with many read pile up to model shift size



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

STAT115

## Genome Browser Viewing



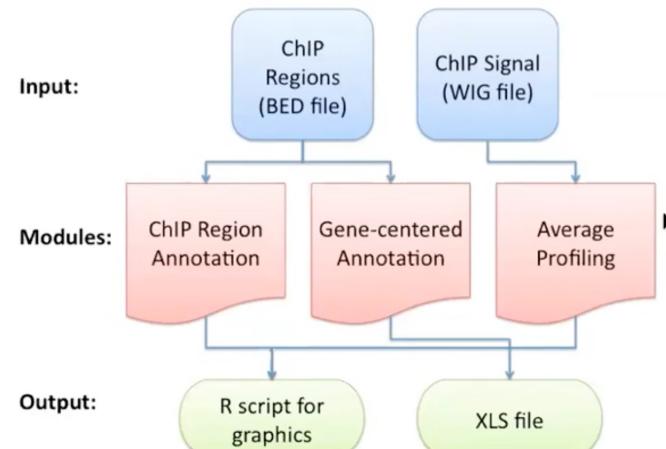
# CEAS peak annotation

- 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>	<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>

STAT1_6h_IFNa_homer				
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](#)

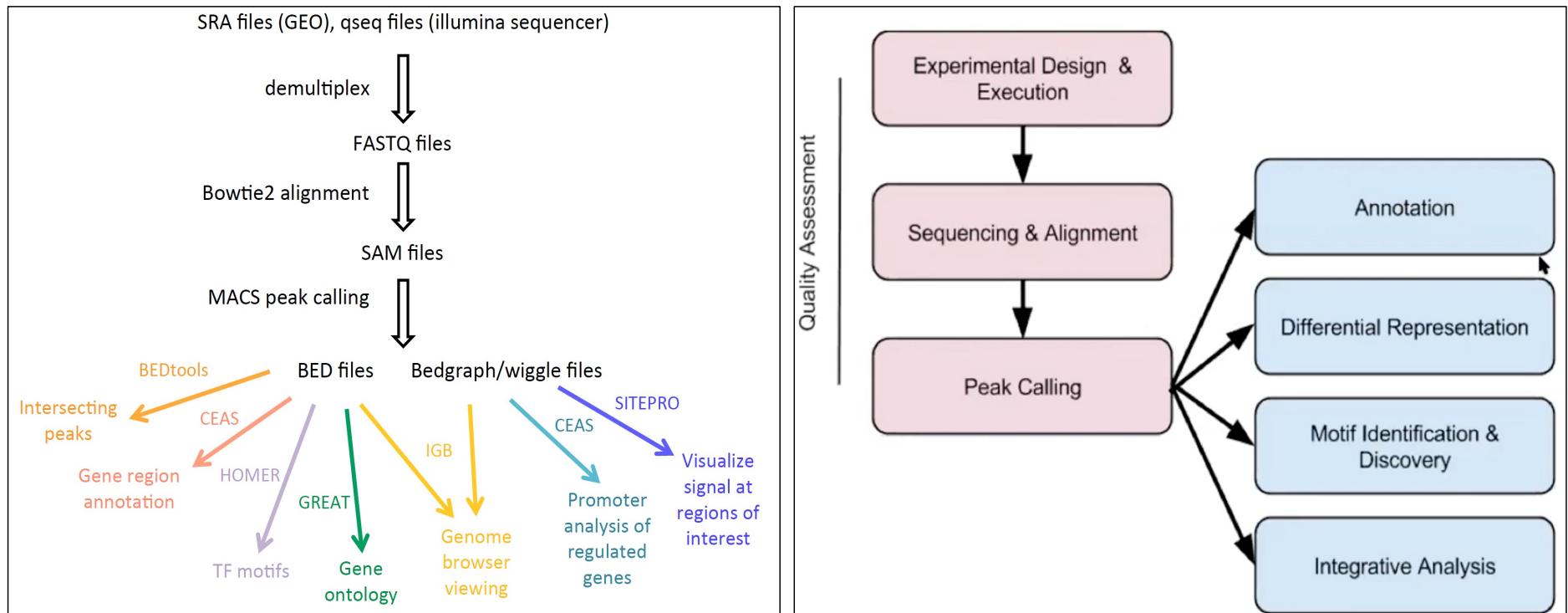
[Gene Ontology Enrichment Results](#)

[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)/Erythoblast-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%

# 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