RNA-Seq Overview

Four major steps, semi-independent of each other.

I. Mapping → produce SAM/BAM or counts data.
II. Quantification → produce RPKM for each gene/transcript.
III. Identifying DEG (Differentially expressed genes) → gene list.
IV. Identifying affected biological processes/pathways.

Automate (Streamline) analyses

- Observe the pipeline that combines decompression, mapping, and Cufflinks/Cuffdiff analysis.
- HiPerGator Slurm job file .

How to gain knowledge from HTS data

- Visualization of HTS data in genome context.
- Discovering genomic patterns.
- Identifying novel mechanism – hypothesis generation.

Visualization of HTS data

- UCSC Genome Browser
  - Server-based access to many data.
  - Custom track – upload your own files or many public tracks.
  - Saved session for sharing data
- IGB – sits on your computer.

Visualizing HTS data with UCSC genome browser

Practice & Observe:

1. Register and log in as a user.
2. Load the track file as custom track to the browser by copy/past the URL link or upload the file.
3. View ‘dense’ and then ‘full’ presentation of the track.
4. Save the session with your favorite gene and share it with yourself by sending the link via email.

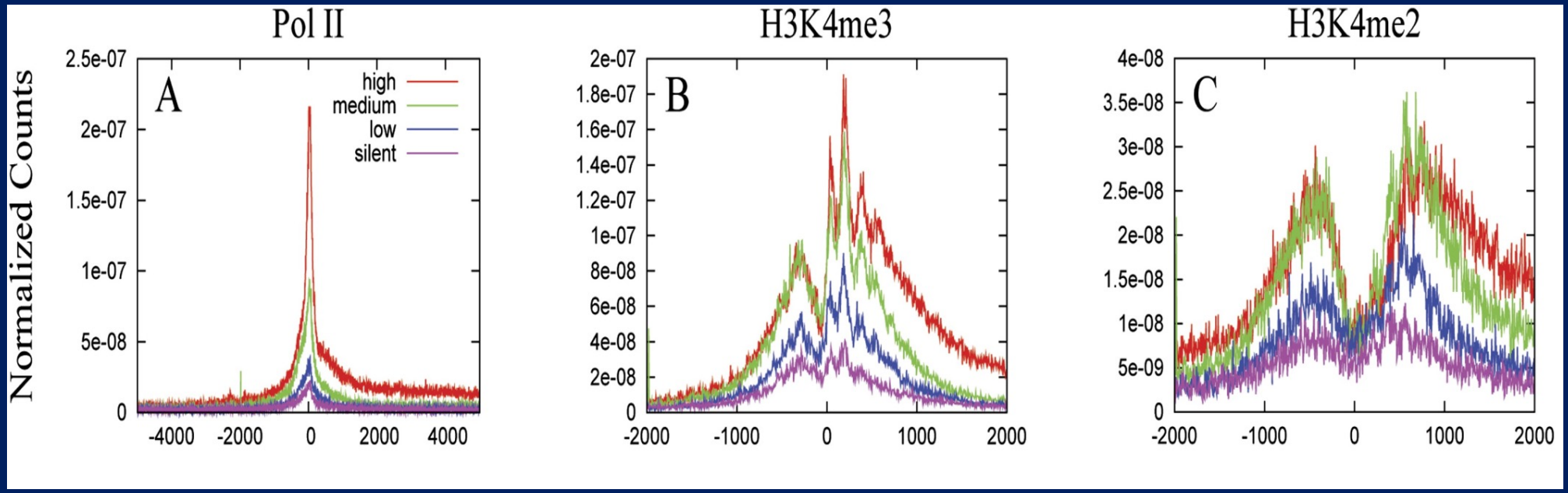
Cautions with genome browser

Coordinates changes with every release/build of genome. – refer to genome release in your work and publication.

Discover genomic pattern

- It may start with anecdotal observation on the browser.
- Aggregated plot to verify the pattern.
  - May need to subset the data.
- Verification.
  - Additional dataset
  - Statistical verification

Discovering genomic patterns



Barski et al. (2007) Cell

Usually requires some programming (scripting).

As a biologist, you need to clearly define your question, and the logic to obtain the data summary.

Practice: Identify TFBSs upstream of the mouse TNF gene.

- Retrieve 2000 bp sequence upstream of the transcription start site and save in a FASTA format file.
- Load the sequence to TFsiteScan or PROMO.

DNA Pattern – Transcription factor binding sites

Con

#	Factor	Model	Beg	Sns	Len	Sequence	L_a	$L_{a/}$	L_q	L_d	L_{pv}	S_c	S_m	S_{pv}	P_{pv}
1	_00000 Bcd	I00231 (Bcd)	1	R	8	CTAATCCC	13.23	1.65	1.000	0.00	5.6e-02	1.00	1.00	0.0e+00	3.2e-01
2	T00063 Bcd	M00140 (mid_c)	1	R	8	CTAATCCC	12.00	1.50	0.884	1.57	2.5e-01	1.00	0.98	1.1e-01	2.2e-01
3	_00000 Prd	I00252 (Prd)	2	R	6	TAATCC	9.04	1.51	0.853	1.56	9.8e-01	1.00	0.93	6.0e-01	6.0e-01
4	_00000 Ftz.2	I00243 (Ftz.2)	2	N	7	TAATCCC	8.49	1.21	1.000	0.00	5.0e-01	1.00	1.00	0.0e+00	7.2e-01
5	T00063 Bcd	Q00016 (-)	2	N	7	TAATCCC	12.73	1.82	1.000	0.00	2.1e-01	1.00	1.00	0.0e+00	4.1e-01
6	T01469 lk-1 T01470 lk-2	R04208 ()	3	R	7	AATCCCA	14.00	2.00	1.000	0.00	nc	?	?	nc	nc
7	T02246 AML1c T02256 AML1a T02457 AML1	R05028 ()	6	R	6	CCCACA	12.00	2.00	1.000	0.00	nc	?	?	nc	nc

Stringency of the matrices

Consensus –10 bp

A	C	G	T	
40	13	23	23	N
20	3	70	5	G
55	3	40	0	R
0	93	0	5	C
53	8	8	30	W
15	0	3	82	T
0	0	100	0	G
0	50	0	50	Y
0	68	0	30	C
12	35	3	48	Y

P53_01

Consensus –20 bp

A	C	G	T	
4	0	13	0	G
5	0	12	0	G
15	0	2	0	A
0	17	0	0	C
17	0	0	0	A
0	0	0	17	T
0	0	17	0	G
0	13	0	4	C
0	17	0	0	C
0	17	0	0	C
0	0	17	0	G
0	0	17	0	G
2	0	15	0	G
0	17	0	0	C
17	0	0	0	A
0	0	0	17	T
0	0	17	0	G
0	2	0	15	T
0	13	0	4	C
0	7	2	7	Y

P53_02

Practice: Identify conserved TFBSs upstream of the human TNF gene.