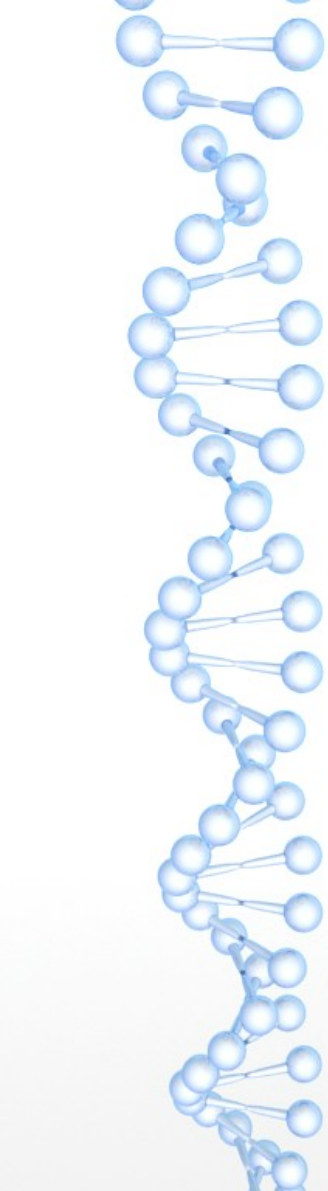


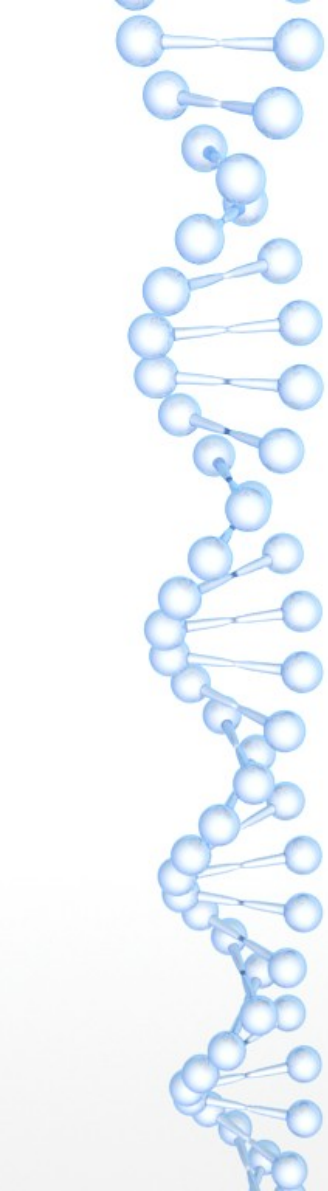
RNA-Seq Analysis of Gene Expression: A Walk-Thru and Tutorial

Helen Nigussie, Michael Mayhew, Dina Machuve

June 4, 2019

Data Science Africa 2019

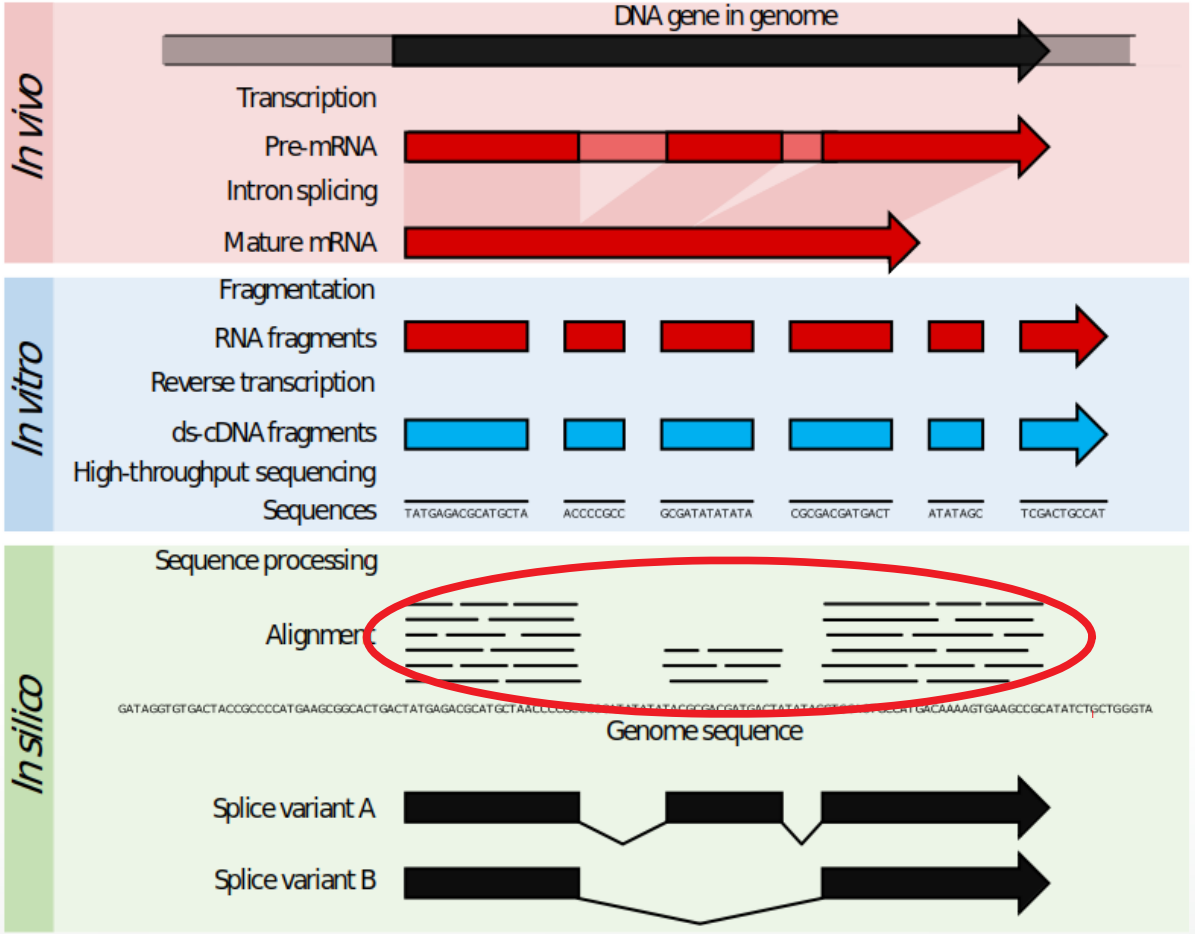
Addis Ababa University, Ethiopia



What is RNA-Seq analysis?

- RNA sequencing *(*RNA-Seq for short) is a process of assessing the ***expression of genes*** across a genome by ***sequencing the RNA transcripts*** from a collection of cells

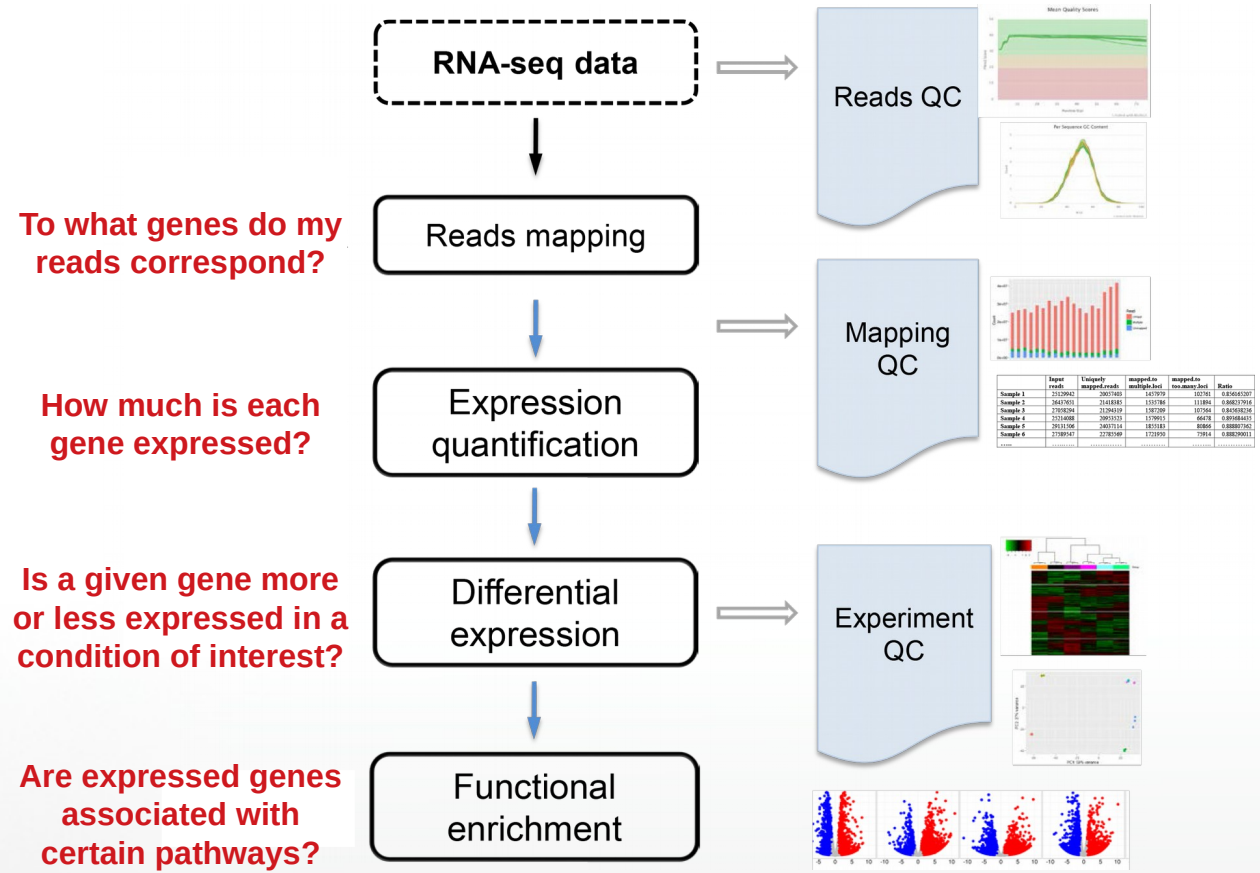
What is RNA-Seq analysis?



These short strands that result from sequencing are called ‘reads’

https://en.wikipedia.org/wiki/RNA-Seq

What are the different stages of RNA-Seq analysis?



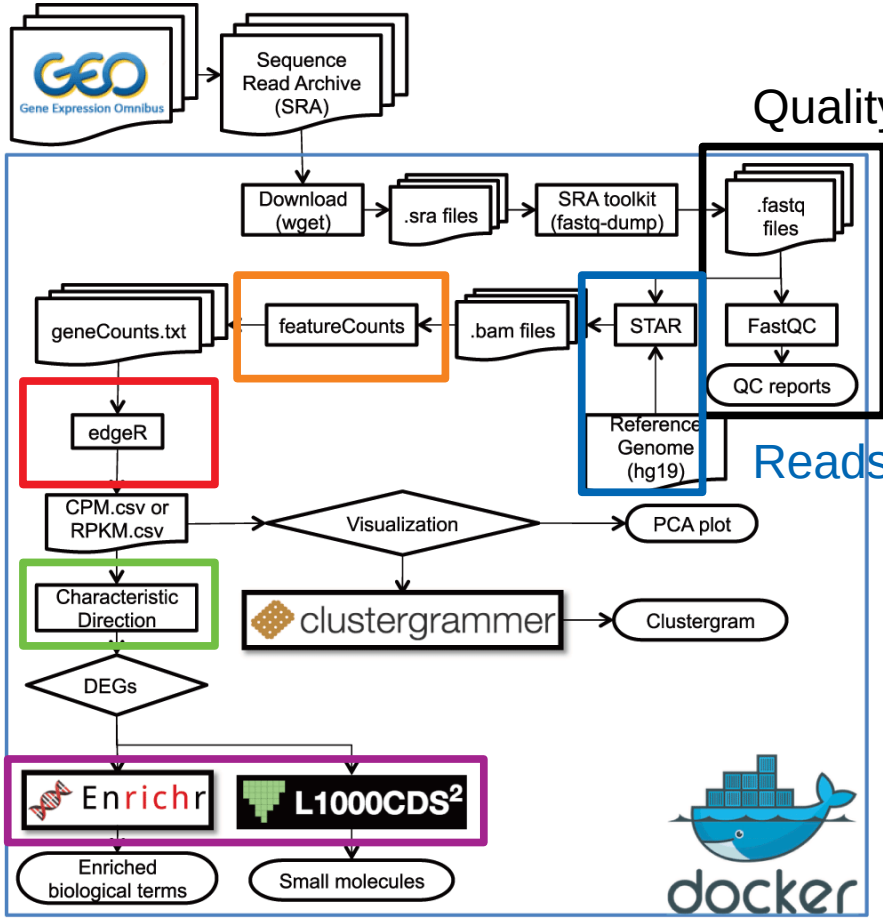
http://bioinfo.vanderbilt.edu/vangard/services-rnaseq.html

What are the different stages of RNA-Seq analysis?

Expression quantification/ normalization

Differential expression analysis

Functional enrichment

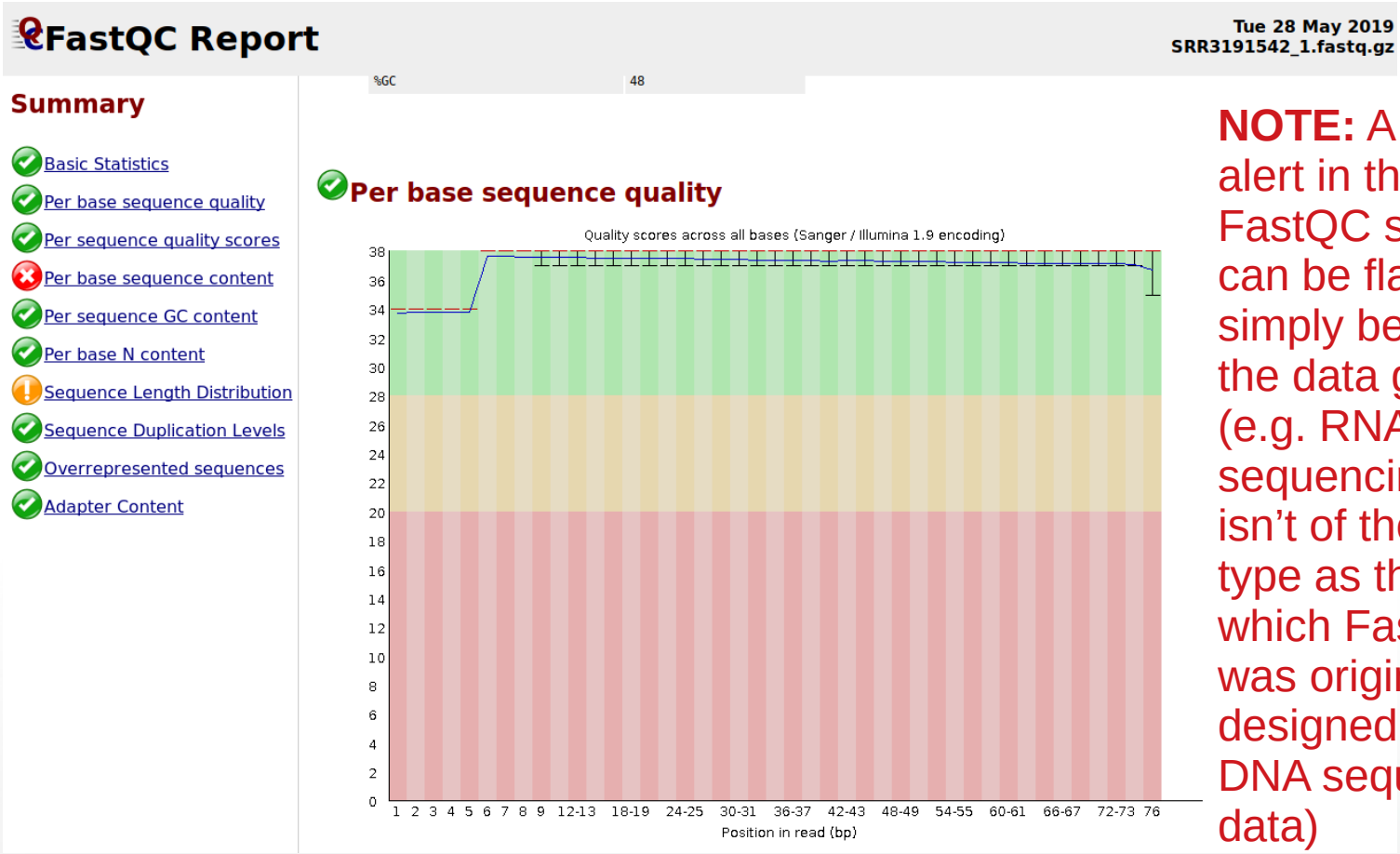


Wang & Ma'ayan, 2016

Stage 1: Processing and quality control of raw sequencing reads

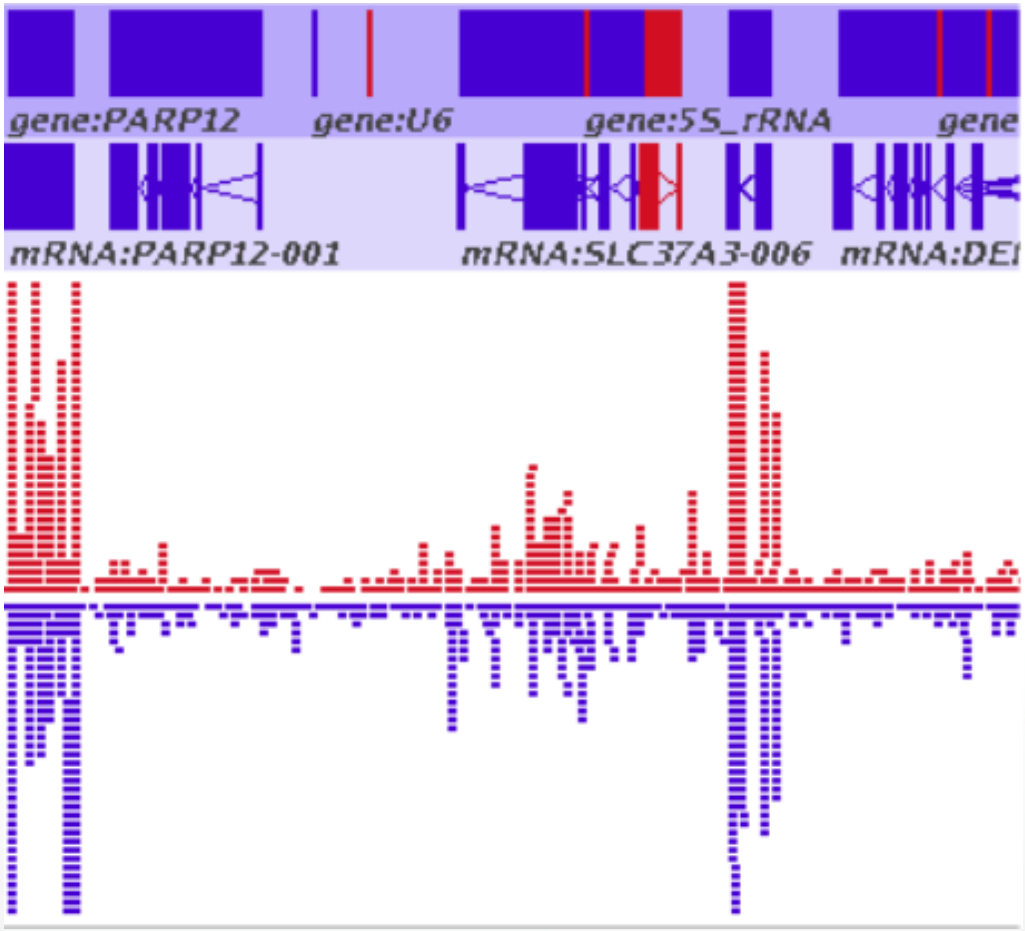
- Reads are often assessed for:
  - Sequencing quality per base
    - We expect generally high quality at all bases
  - Sequencing quality per read
    - We expect high quality for longer reads
  - Sequence content (nucleotide base composition)
    - We expect a roughly uniform base composition across the read (except maybe for the initial bases; depends on how RNA prepared)
  - Per base ‘N’ content (or non-call)
    - Indicates potential instrument failure
  - Other measures

Stage 1: Processing and quality control of raw sequencing reads (cont'd)



NOTE: A 'failure' alert in the FastQC summary can be flagged simply because the data given (e.g. RNA sequencing data) isn't of the same type as that for which FastQC was originally designed (e.g. DNA sequencing data)

Stage 2: Mapping of sequencing reads to genome



The histogram-like plot to the left indicates the cumulative counts of sequencing reads at different positions in the genome.

Stage 3: Assignment of reads to individual genes to attain expression measurements

- Sequencing reads are aligned (‘mapped’) to a reference genome in which locations of genes are known
- Algorithms (like featureCounts) assign the aligned reads to each gene
  - Results in ‘digital’ measures of expression – one unit of expression per mapped read
- Counts are then normalized according to sequencing depth and/or gene length
  - Two common normalized expression measures are:
    - CPM – transcripts or counts per million

1. $RPK_i = \frac{R_i}{L_i}$ 2. $S = \frac{\sum_i RPK_i}{10^6}$ 3. $CPM_i = \frac{RPK_i}{S}$

    - RPKM – reads per kilobase per million

1. $S = \frac{\sum_i R_i}{10^6}$ 2. $RPM_i = \frac{R_i}{S}$ 3. $RPKM_i = \frac{RPM_i}{L_i}$

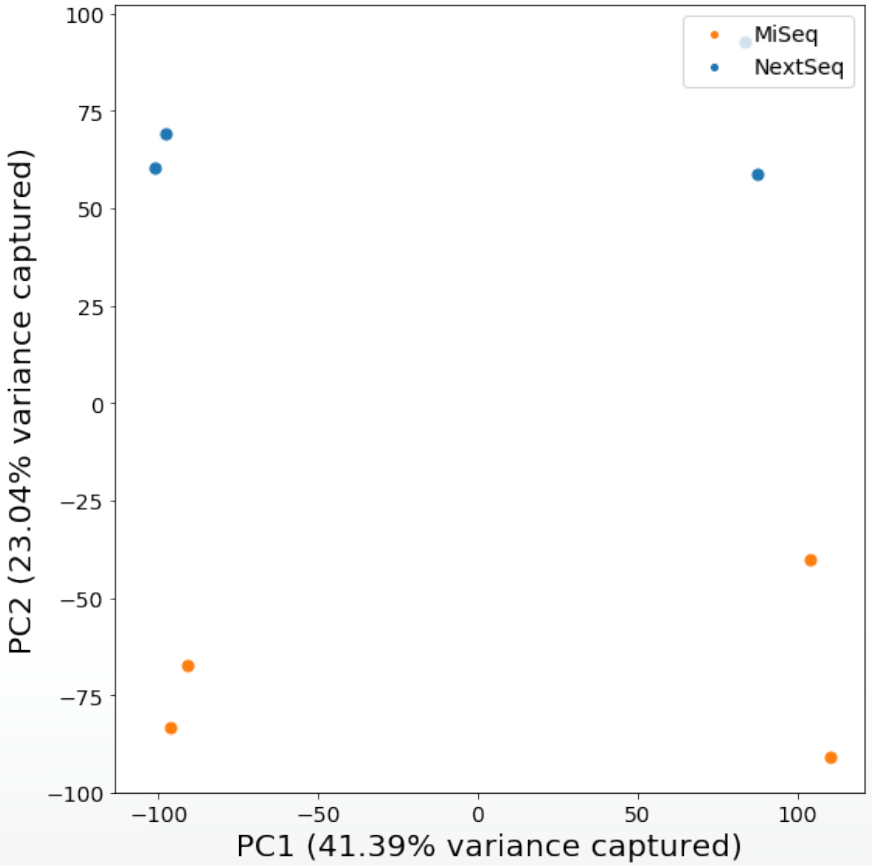
NOTE:

R_i – read counts for gene i

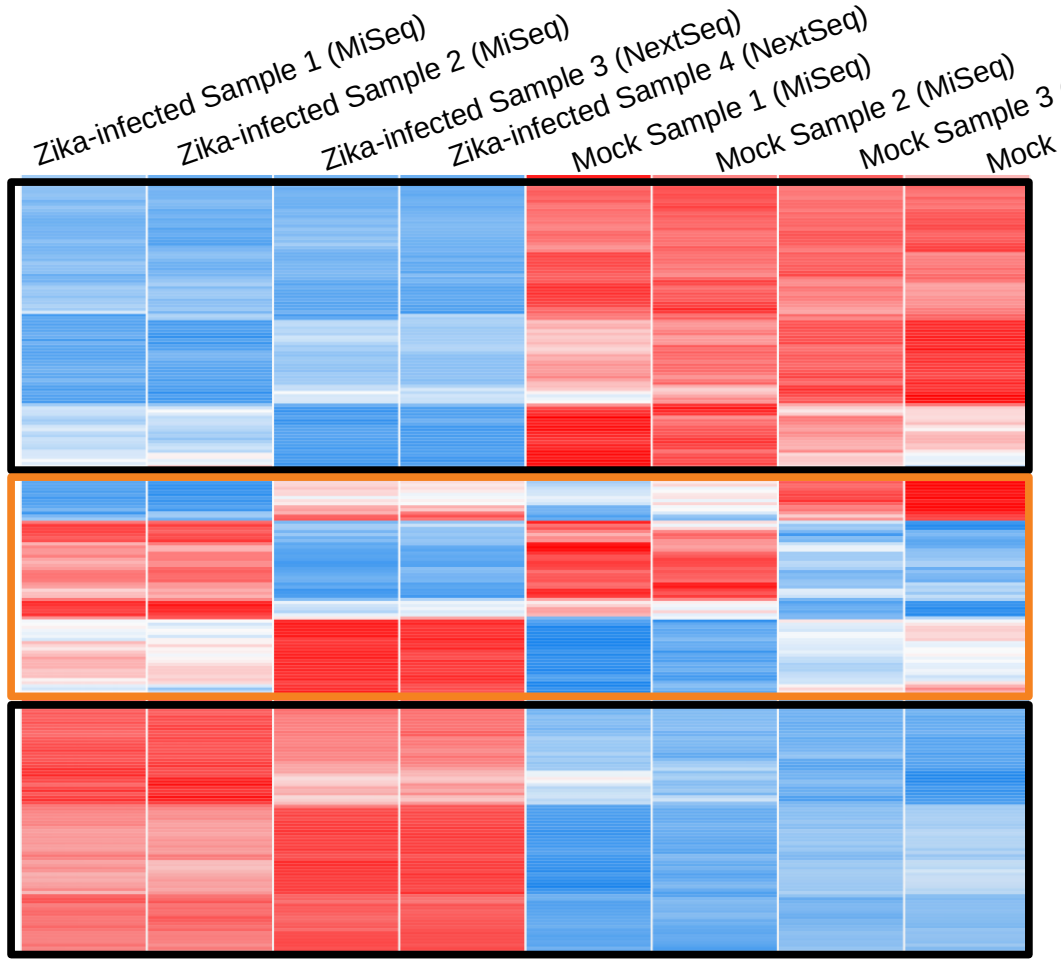
L_i – length in kilobases of gene i

Important considerations when performing an RNA-Seq analysis

- Should I consider all genes in my analysis? What about those with low or no expression across all conditions/platforms?
- Are the expression differences I'm seeing solely due to the condition? Or some other factor?

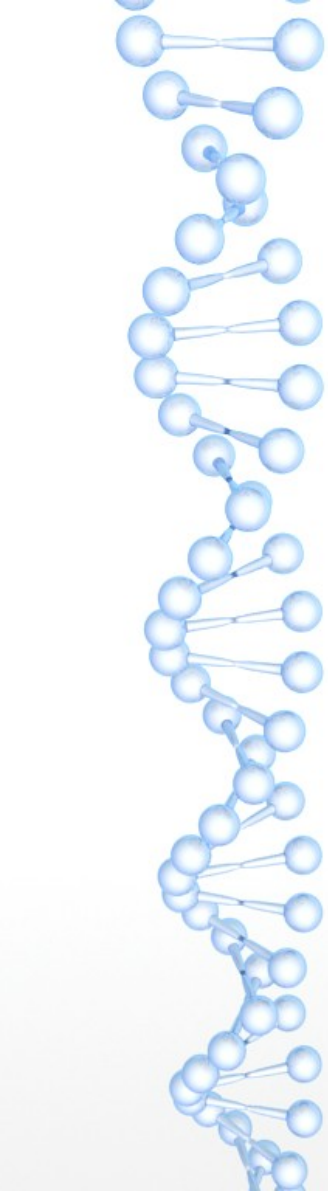


What is the structure in my expression data?



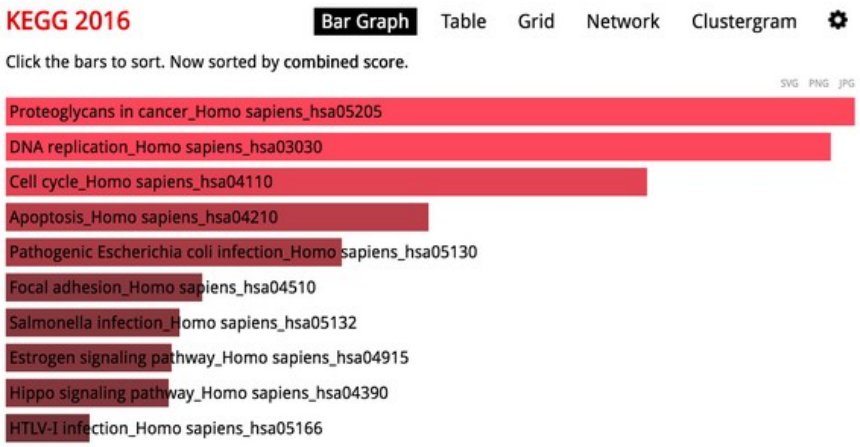
- Most genes cluster by condition.

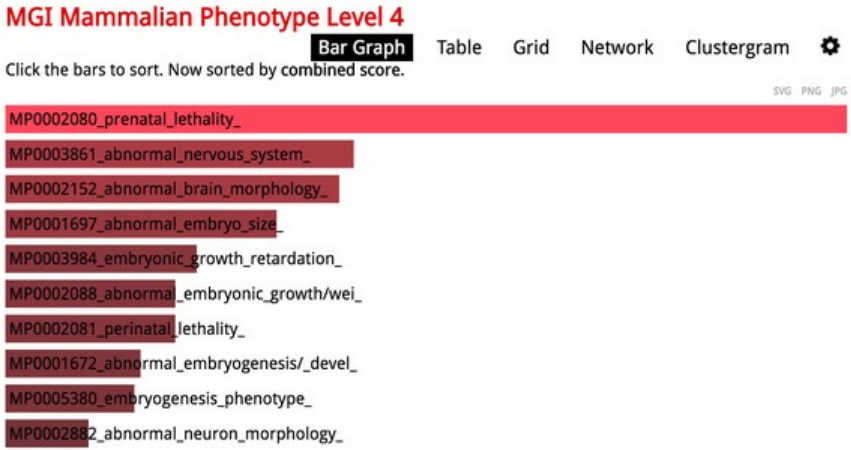
However, some genes clearly show platform-specific effects.



What genes show different expression patterns in my conditions of interest?

Are differentially expressed genes enriched for any biological processes or pharmacological targets?





Genes with ***low*** expression in Zika-infected samples are enriched for cell-cycle and DNA replication processes.

Genes with ***high*** expression in Zika-infected samples are enriched for prenatal lethality phenotypes in mice.

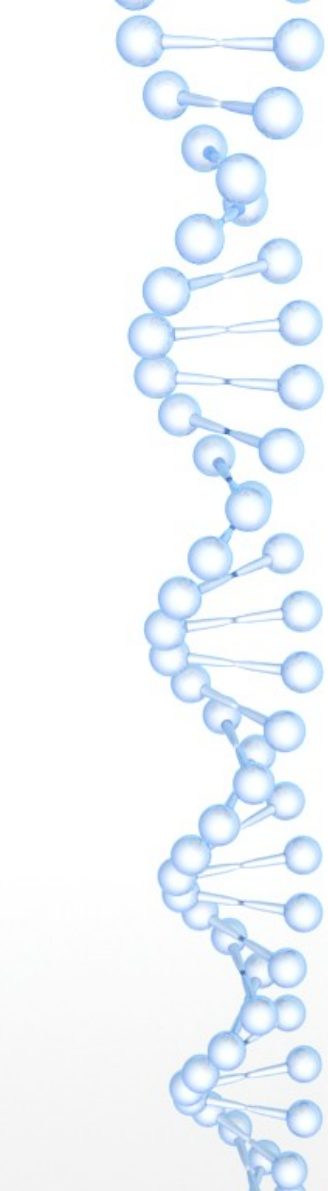
An unsolicited advertisement

Oral Presentation Submission Deadline: September 13, 2019
Poster Presentation Submission Deadline: October 15, 2019

https://www.iscb.org/iscbafrica2019

Additional resources

- Galaxy Community Hub's RNA-Seq Introduction: https://galaxyproject.org/tutorials/rb_rnaseq/
- FastQC Tutorial & FAQ: https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/
- Description of normalized RNA-Seq expression measures: https://statquest.org/2015/07/09/rpkm-fpkm-and-tpm-clearly-explained/
-



Thanks for your attention and see you at the workshop!

Any questions?