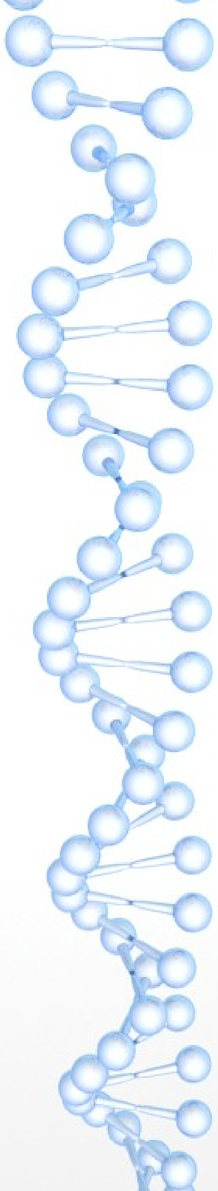


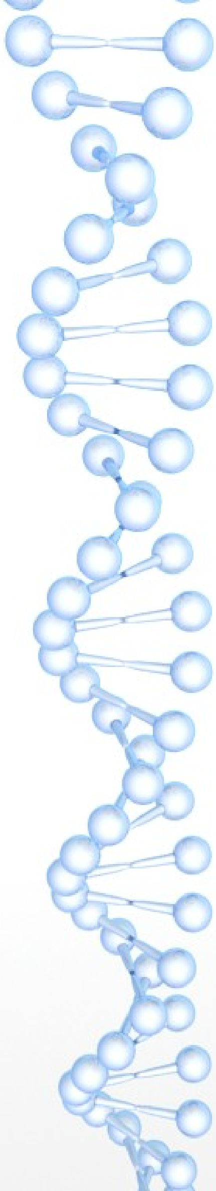
RNA-Seq Analysis of Gene Expression: A Walk-Thru and Tutorial

Helen Nigussie, Michael Mayhew, Dina Machuve

June 4, 2019

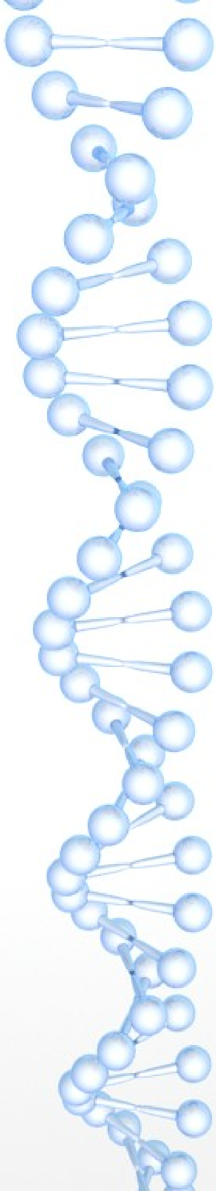
Data Science Africa 2019

Addis Ababa University, Ethiopia

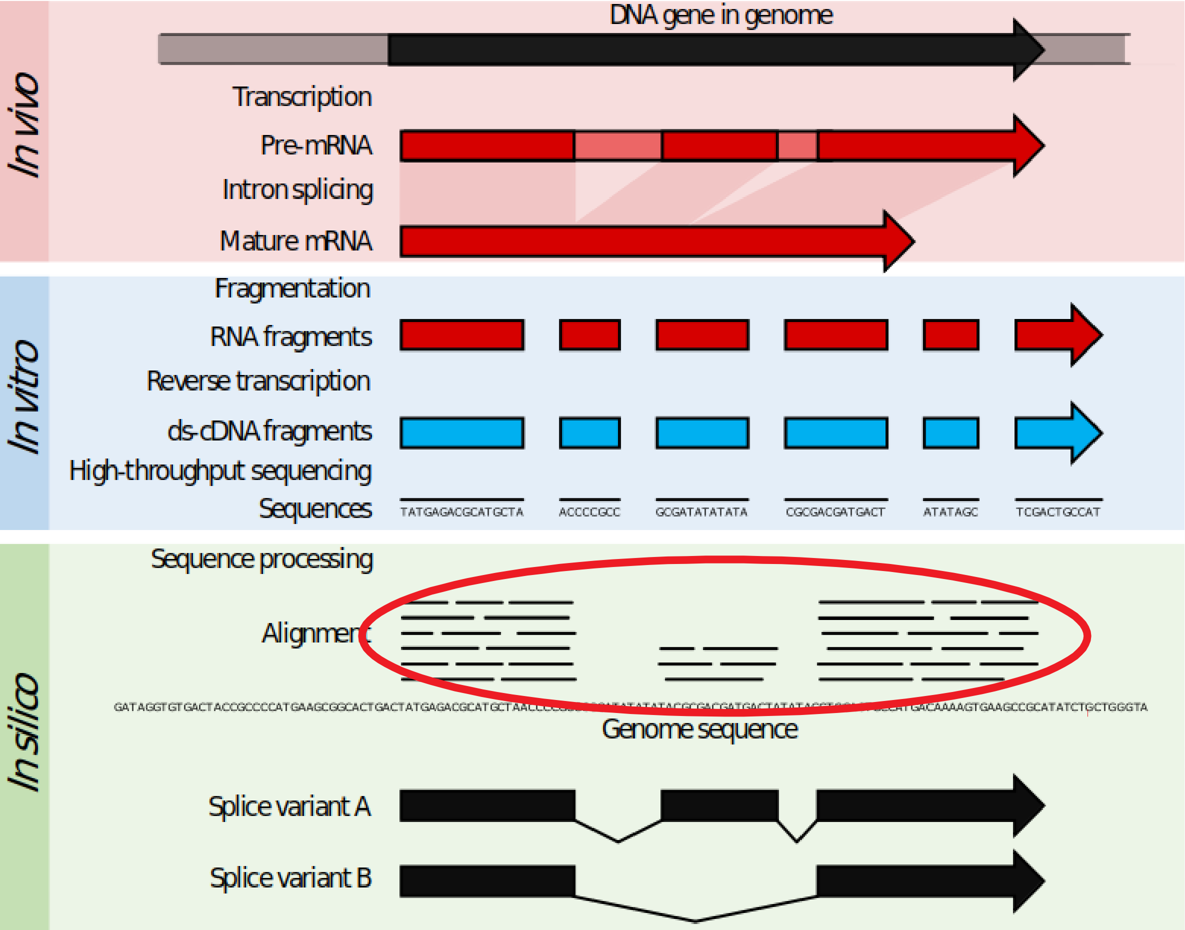


What is RNA-Seq analysis?

- 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

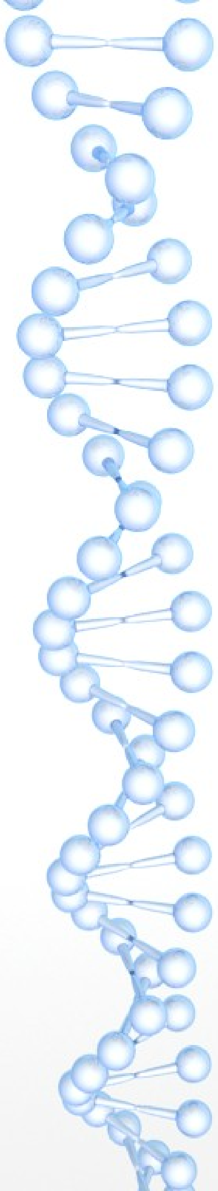


What is RNA-Seq analysis?

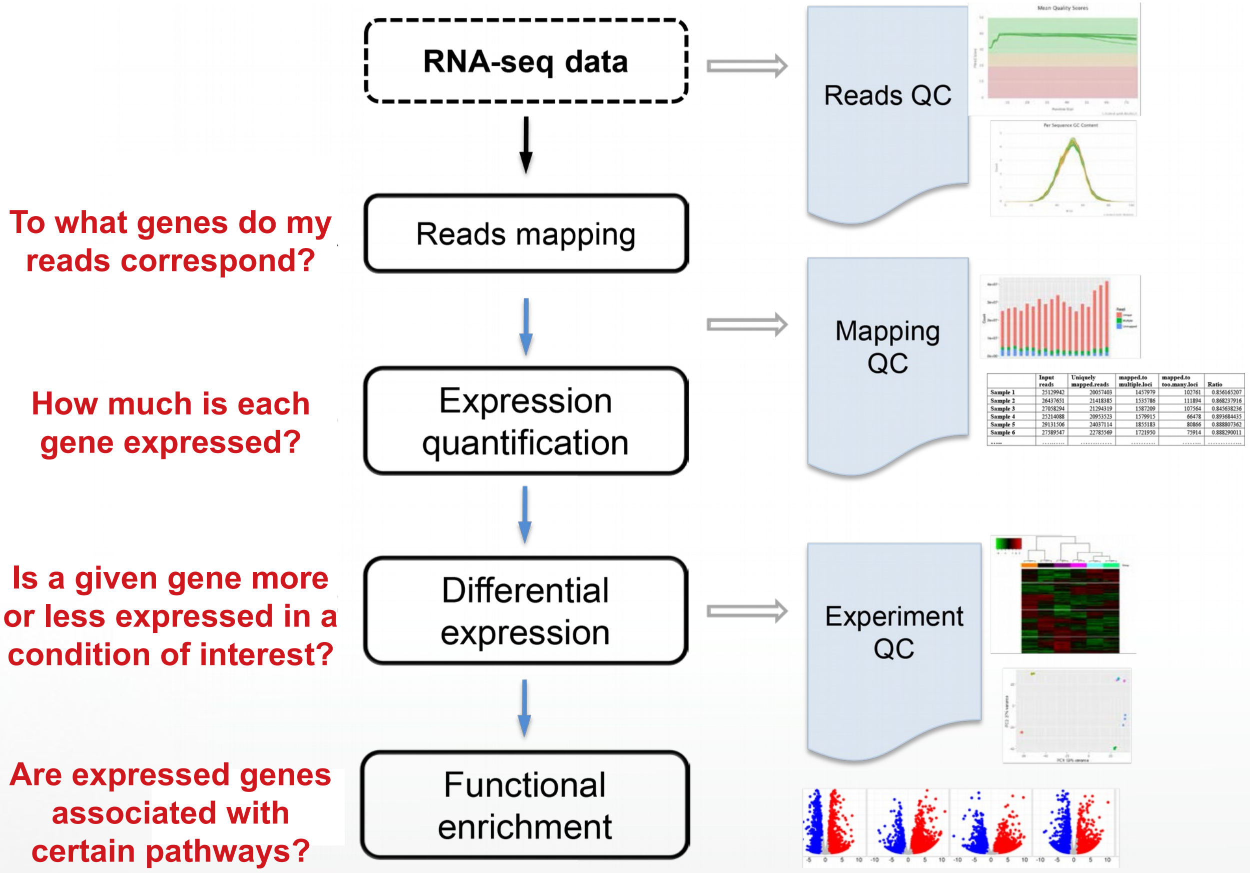


These short strands that result from sequencing are called ‘reads’

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



What are the different stages of RNA-Seq analysis?



RNA-seq data

Reads mapping

Expression quantification

Differential expression

Functional enrichment

Reads QC

Mapping QC

Experiment QC

To what genes do my reads correspond?

How much is each gene expressed?

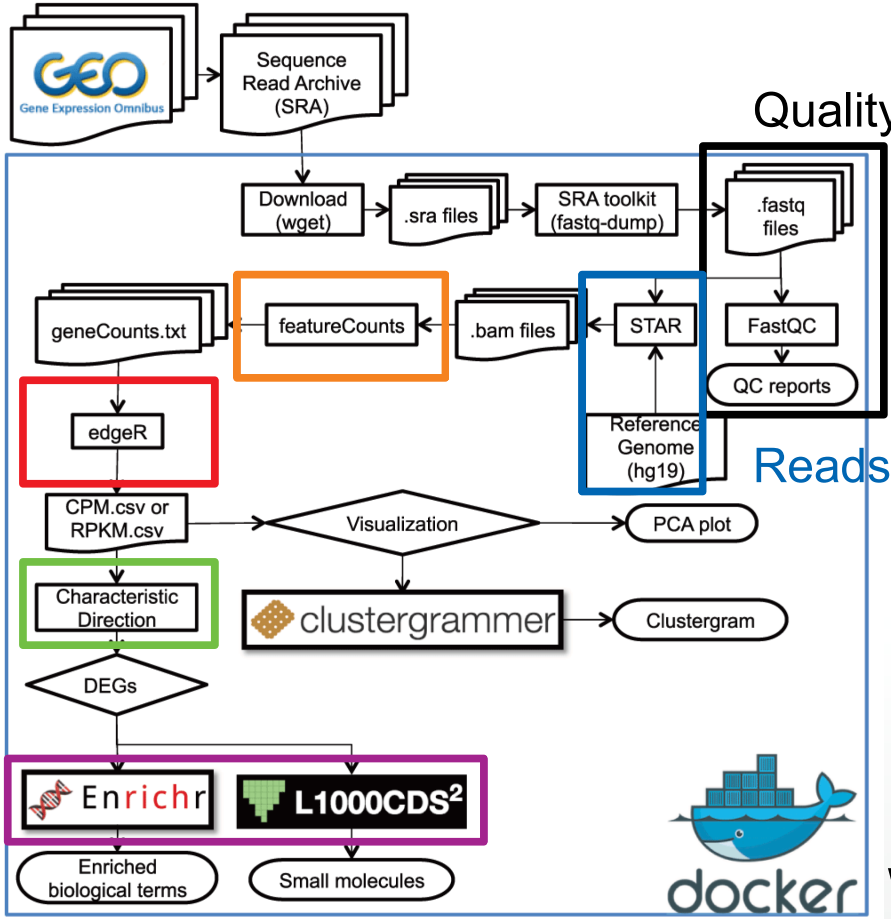
Is a given gene more or less expressed in a condition of interest?

Are expressed genes associated with certain pathways?

	Input reads	Uniquely mapped.reads	mapped.to multiple.loci	mapped.to too.many.loci	Ratio
Sample 1	25129942	20057403	1457979	102761	0.856165207
Sample 2	26437651	21418385	1535786	111894	0.868237916
Sample 3	27058294	21294319	1587209	107564	0.845638236
Sample 4	25214088	20953523	1579915	66478	0.893684435
Sample 5	29131506	24037114	1855183	80866	0.888807362
Sample 6	27589547	22785569	1721950	75914	0.888290011
.....	....	....	....	....	....

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

What are the different stages of RNA-Seq analysis?



Quality control

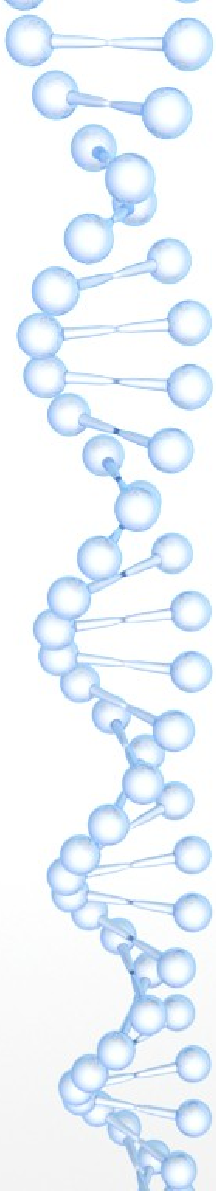
Expression quantification/ normalization

Reads mapping

Differential expression analysis

Functional enrichment

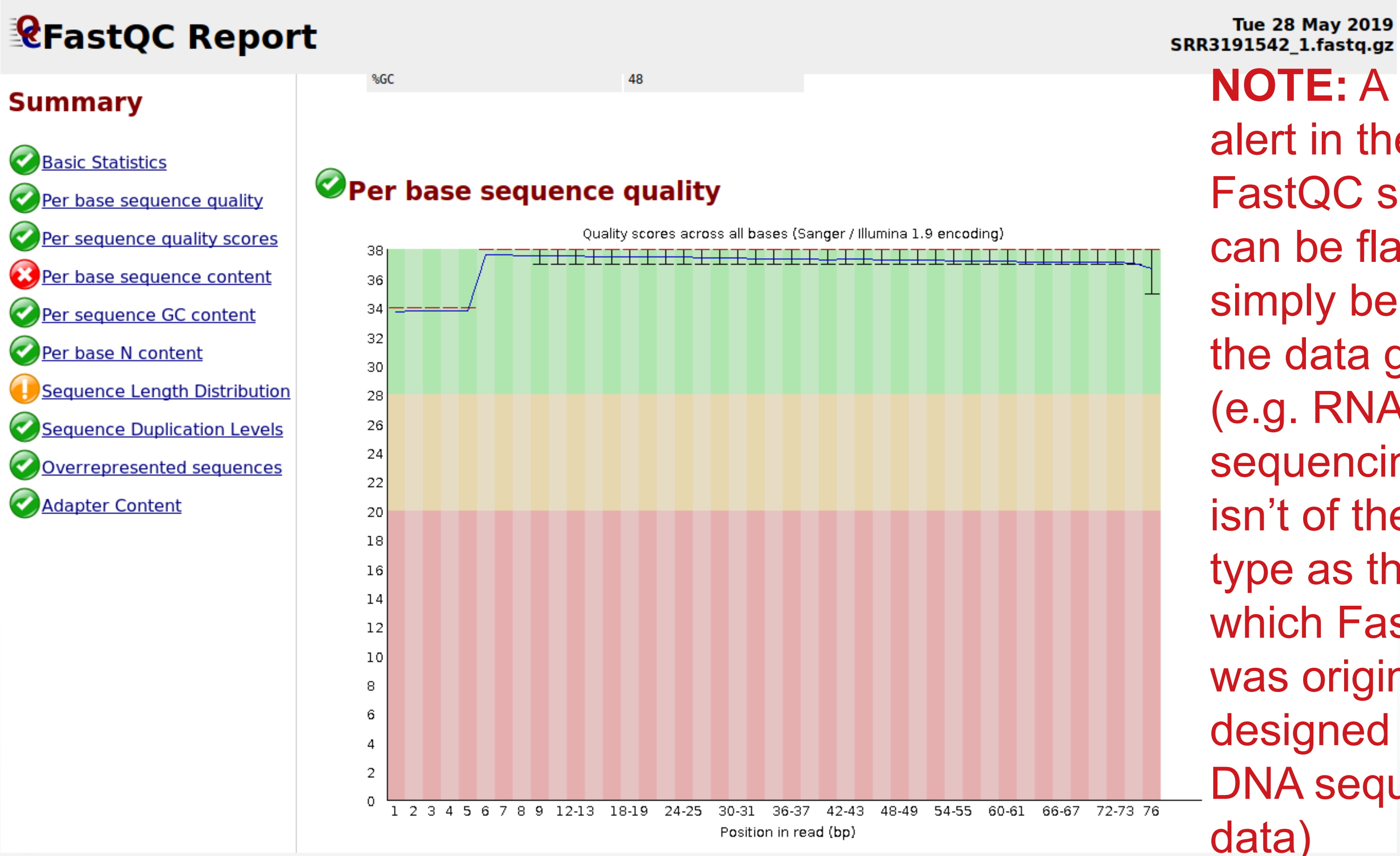
Wang & Ma'ayan, 2016



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)



FastQC Report

Tue 28 May 2019
SRR3191542_1.fastq.gz

%GC	48

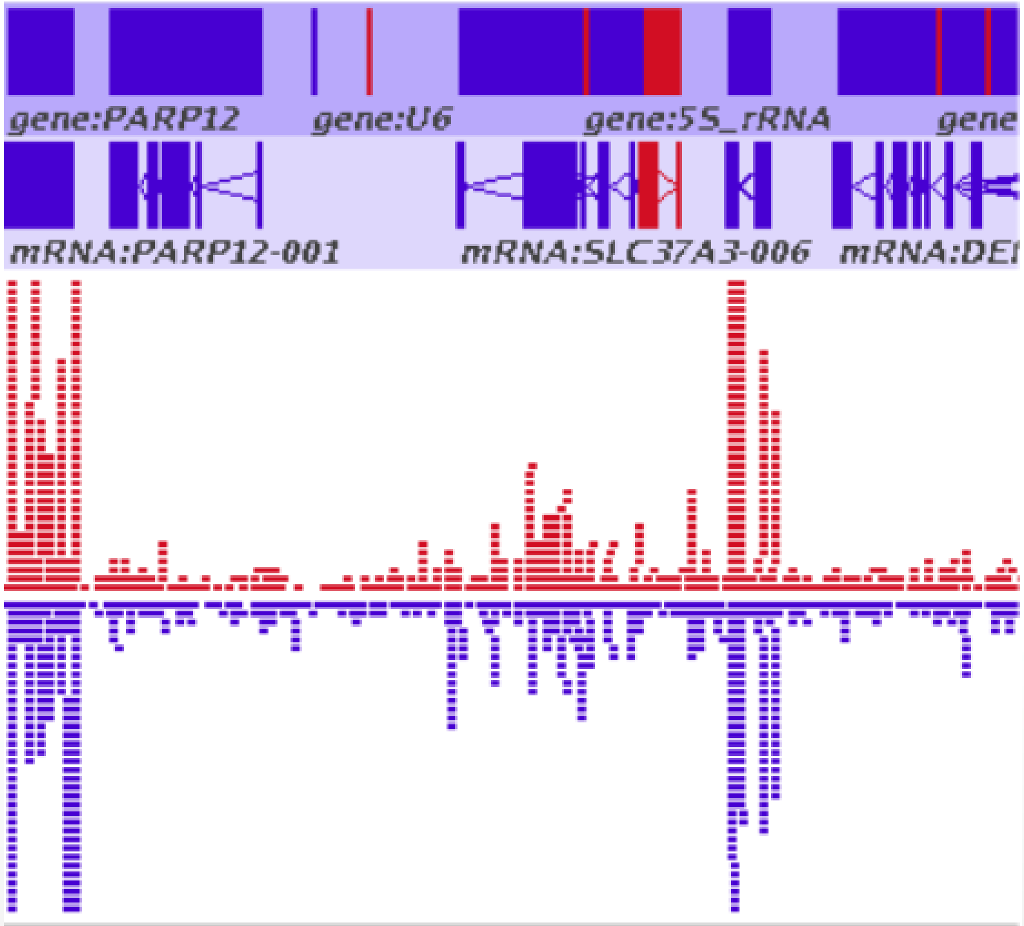
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

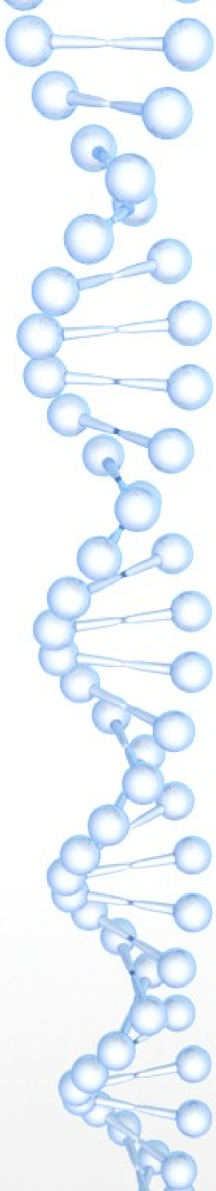
✓ Per base sequence quality

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:

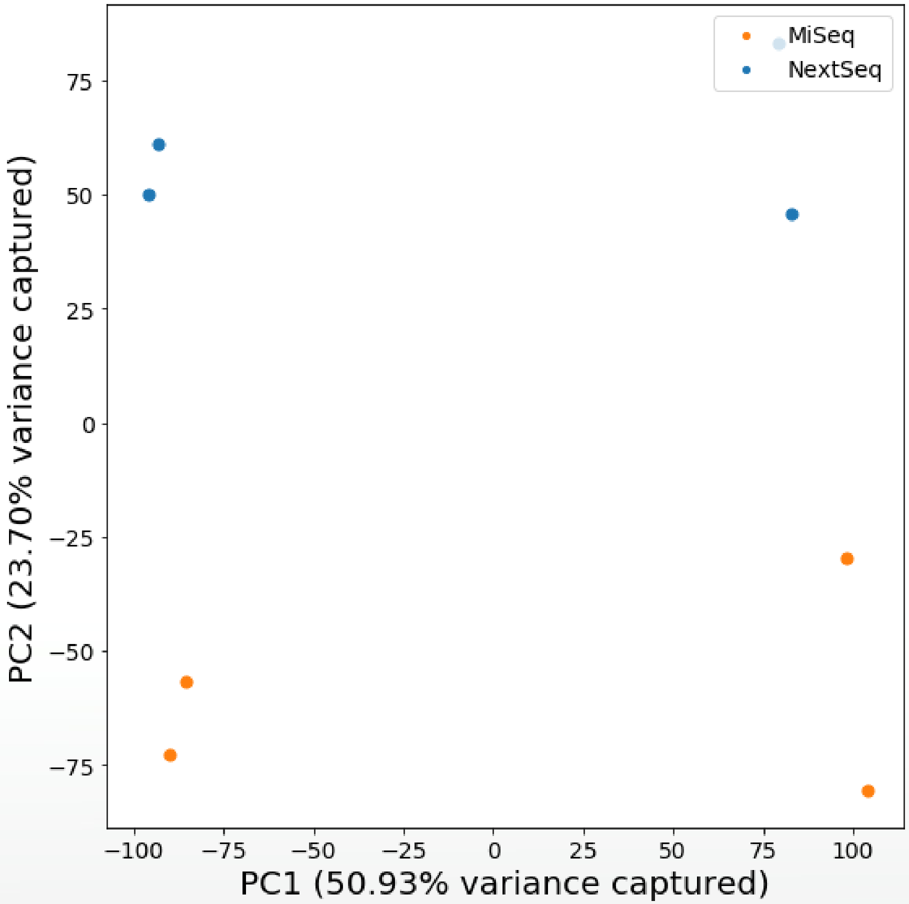
R_i – read counts for gene i

L_i – length in kilobases of gene i

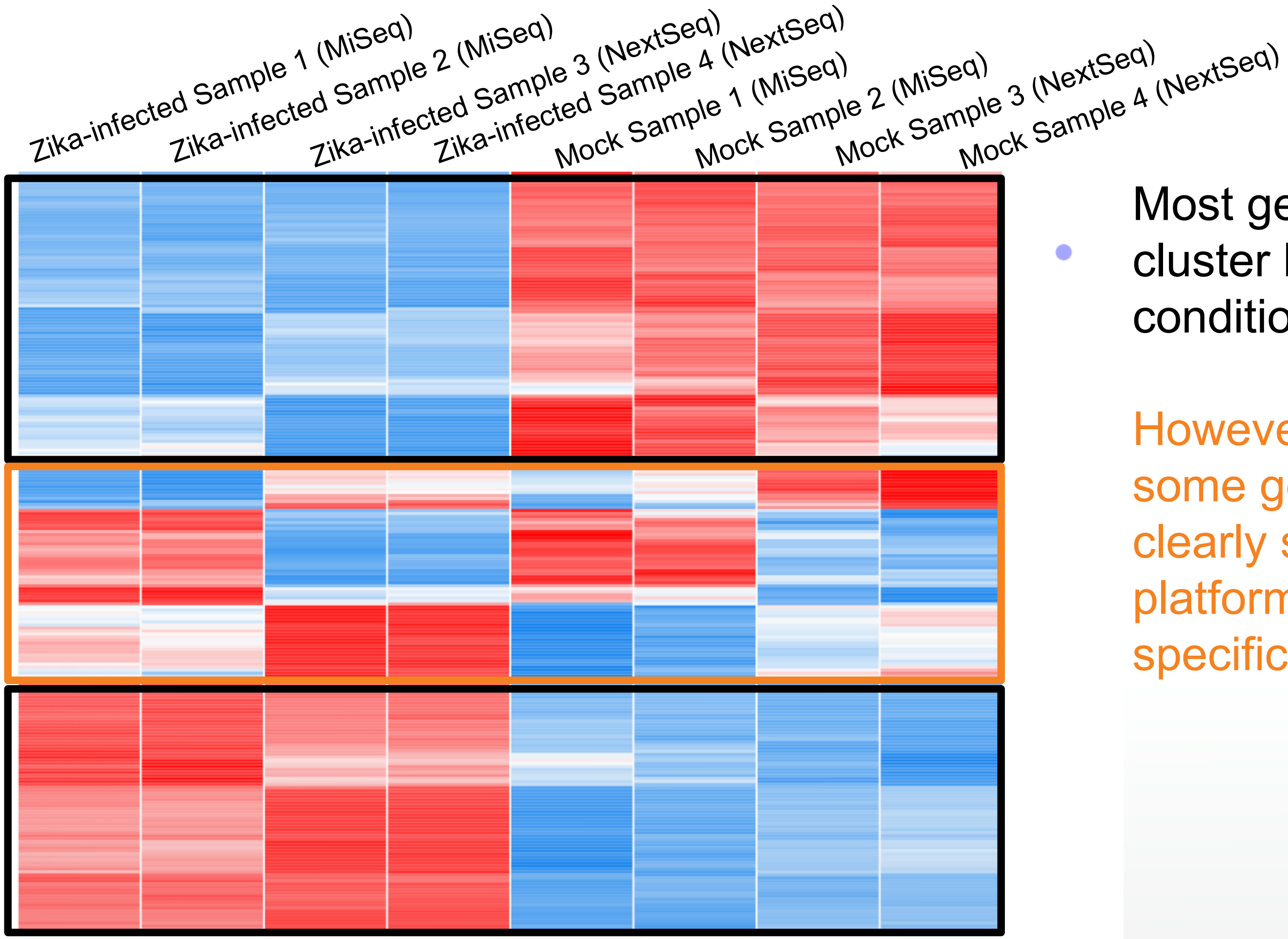
Important considerations when performing an RNA-Seq analysis

- 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?

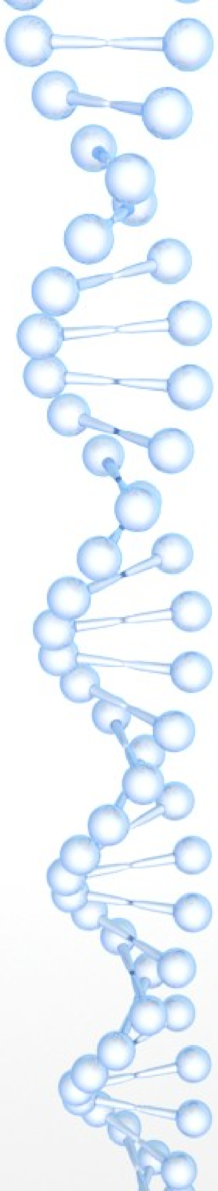


What is the structure in my expression data?



- Most genes cluster by condition.

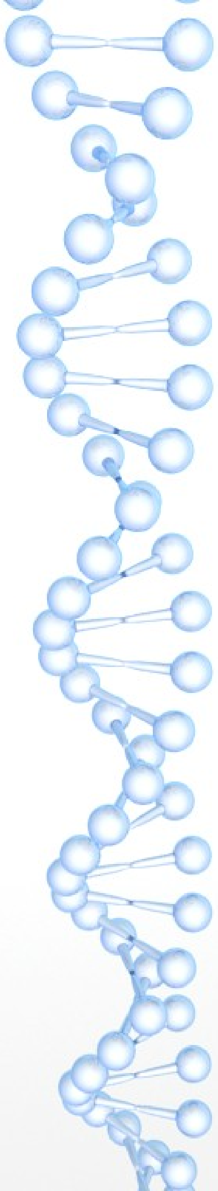
However, some genes clearly show platform-specific effects.

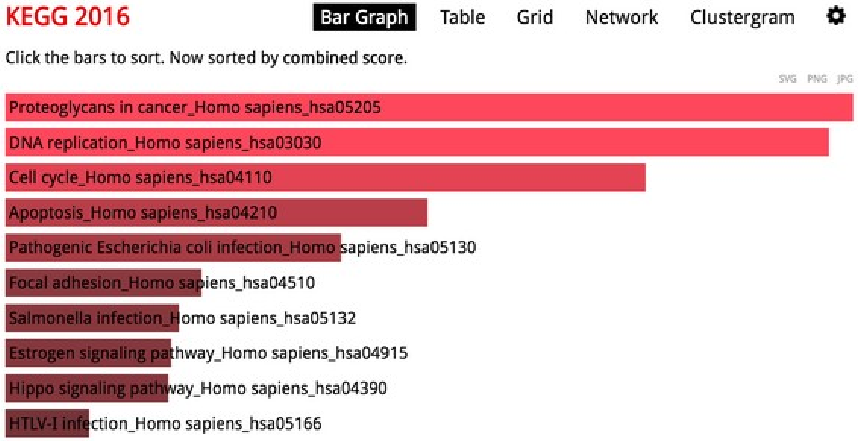


What genes show different expression patterns in my conditions of interest?

	MiSeq	NextSeq 500	Combined
WASH7P	-0.000268	-0.000969	-0.000385
LOC729737	-0.000134	-0.000529	-0.000198
LOC100133331	-0.000755	-0.000849	-0.000701
MIR6723	-0.001514	-0.000954	-0.001068
LOC100288069	-0.000337	-0.000720	-0.000428

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

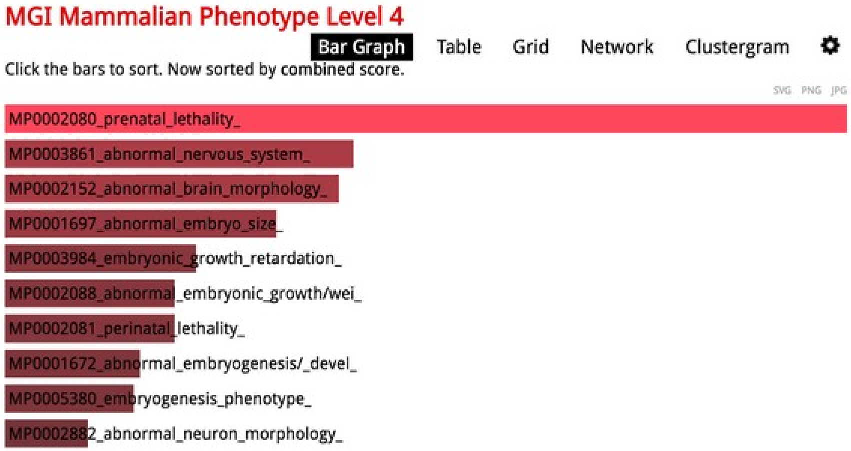




KEGG 2016

Click the bars to sort. Now sorted by combined score.

- Proteoglycans in cancer_Homo sapiens_hsa05205
- DNA replication_Homo sapiens_hsa03030
- Cell cycle_Homo sapiens_hsa04110
- Apoptosis_Homo sapiens_hsa04210
- Pathogenic Escherichia coli infection_Homo sapiens_hsa05130
- Focal adhesion_Homo sapiens_hsa04510
- Salmonella infection_Homo sapiens_hsa05132
- Estrogen signaling pathway_Homo sapiens_hsa04915
- Hippo signaling pathway_Homo sapiens_hsa04390
- HTLV-I infection_Homo sapiens_hsa05166



MGI Mammalian Phenotype Level 4

Click the bars to sort. Now sorted by combined score.

- MP0002080_prenatal_lethality_
- MP0003861_abnormal_nervous_system_
- MP0002152_abnormal_brain_morphology_
- MP0001697_abnormal_embryo_size_
- MP0003984_embryonic_growth_retardation_
- MP0002088_abnormal_embryonic_growth/wei_
- MP0002081_perinatal_lethality_
- MP0001672_abnormal_embryogenesis/_devel_
- MP0005380_embryogenesis_phenotype_
- MP0002882_abnormal_neuron_morphology_

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

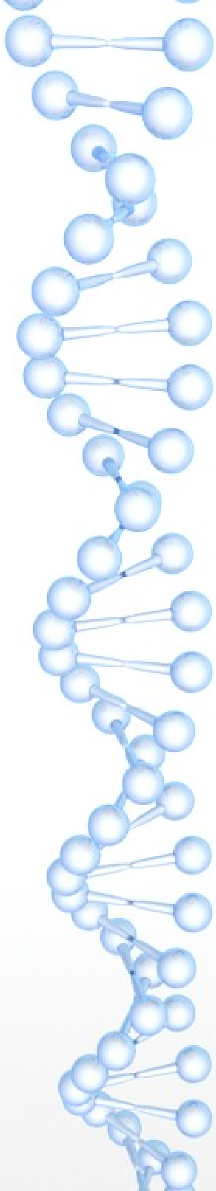
Genes with ***high*** expression in Zika-infected samples are enriched for prenatal lethality phenotypes in mice.

An unsolicited advertisement

ISCB-Africa ASBCB Conference on Bioinformatics
Kumasi, Ghana
November 11-15, 2019
iSCB
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
ASBCB
Official Conference
Kumasi, Ghana
November 11-15, 2019
Mark your calendars!

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

https://www.iscb.org/iscbafrica2019



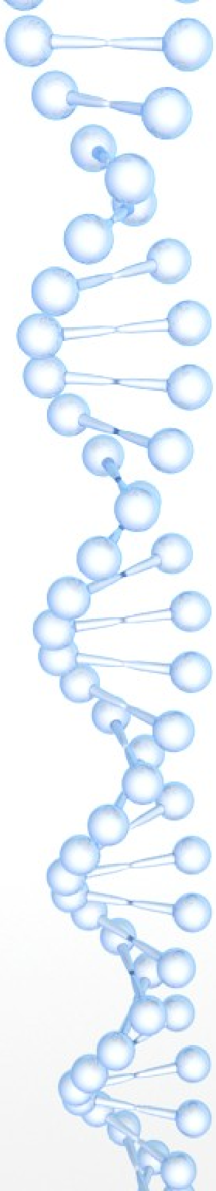
Additional resources

- 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?