



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

Helen Nigussie, Michael Mayhew, Dina Machuve

June 4, 2019

Data Science Africa 2019

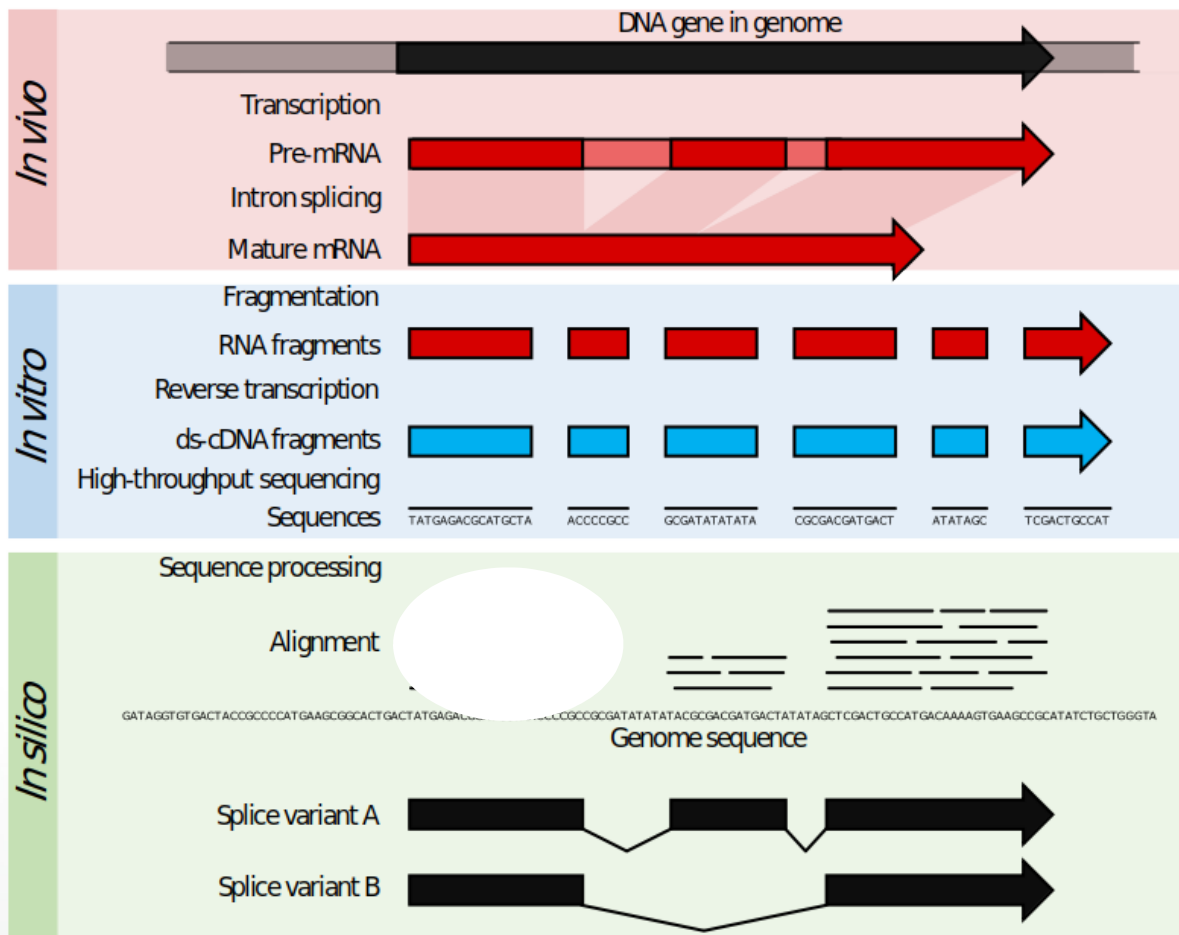
Addis Ababa, Ethiopia



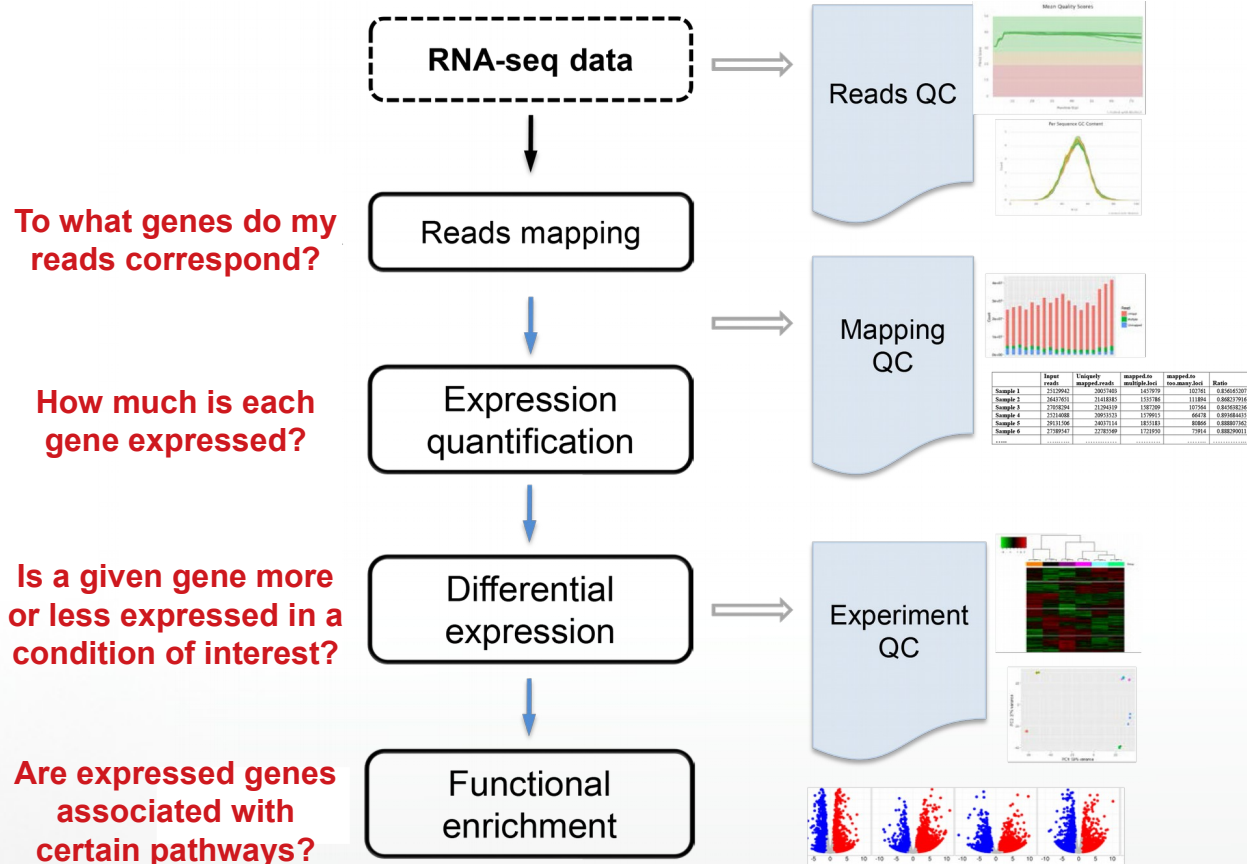
What is RNA-Seq analysis?

- 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

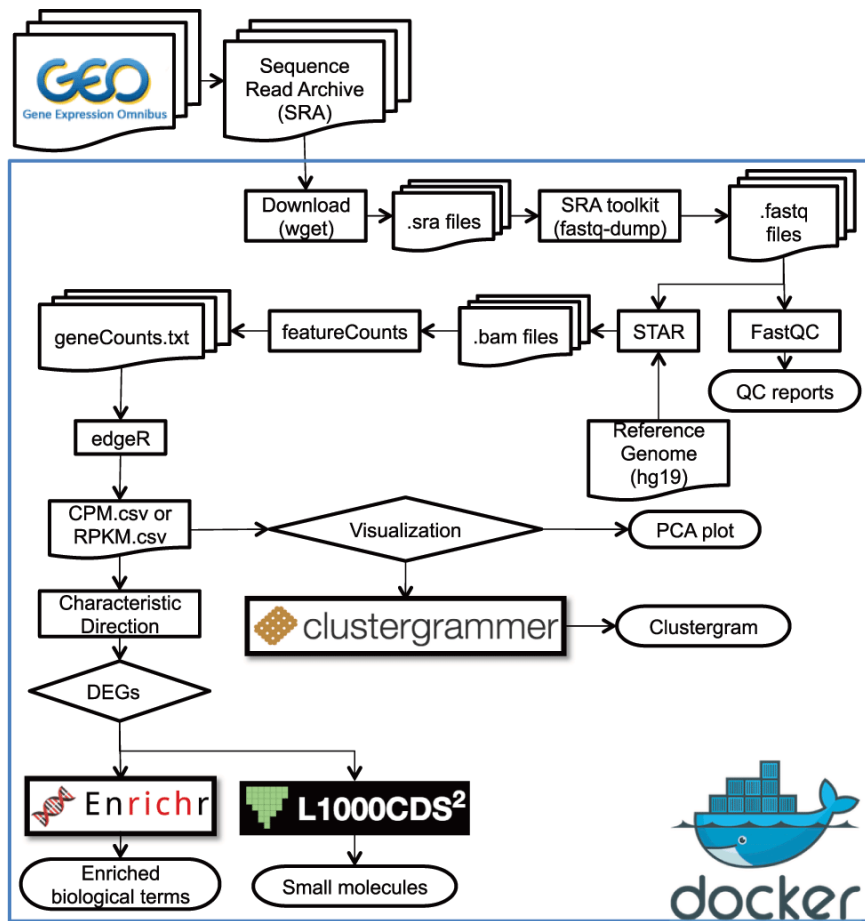
What is RNA-Seq analysis?

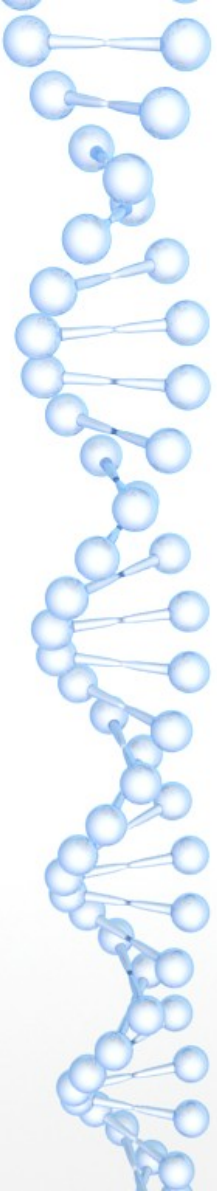


What are the different stages of RNA-Seq analysis?



What are the different stages of RNA-Seq analysis?





Stage 1: Processing and quality control of raw sequencing reads

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

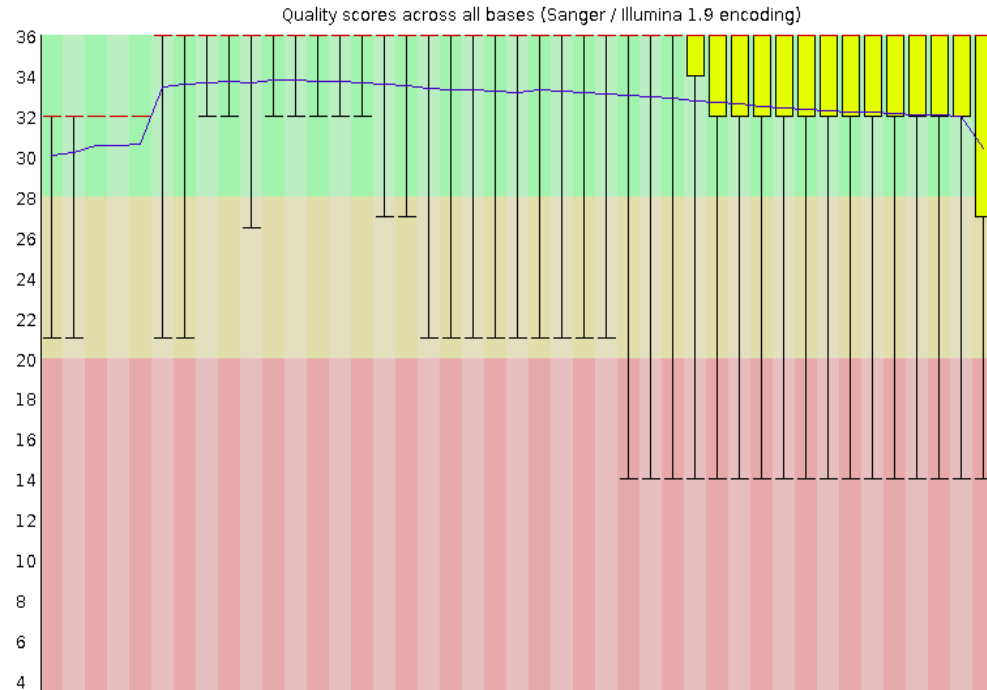
FastQC Report

Tue 28 May 2019
SRR3194429.fastq.gz

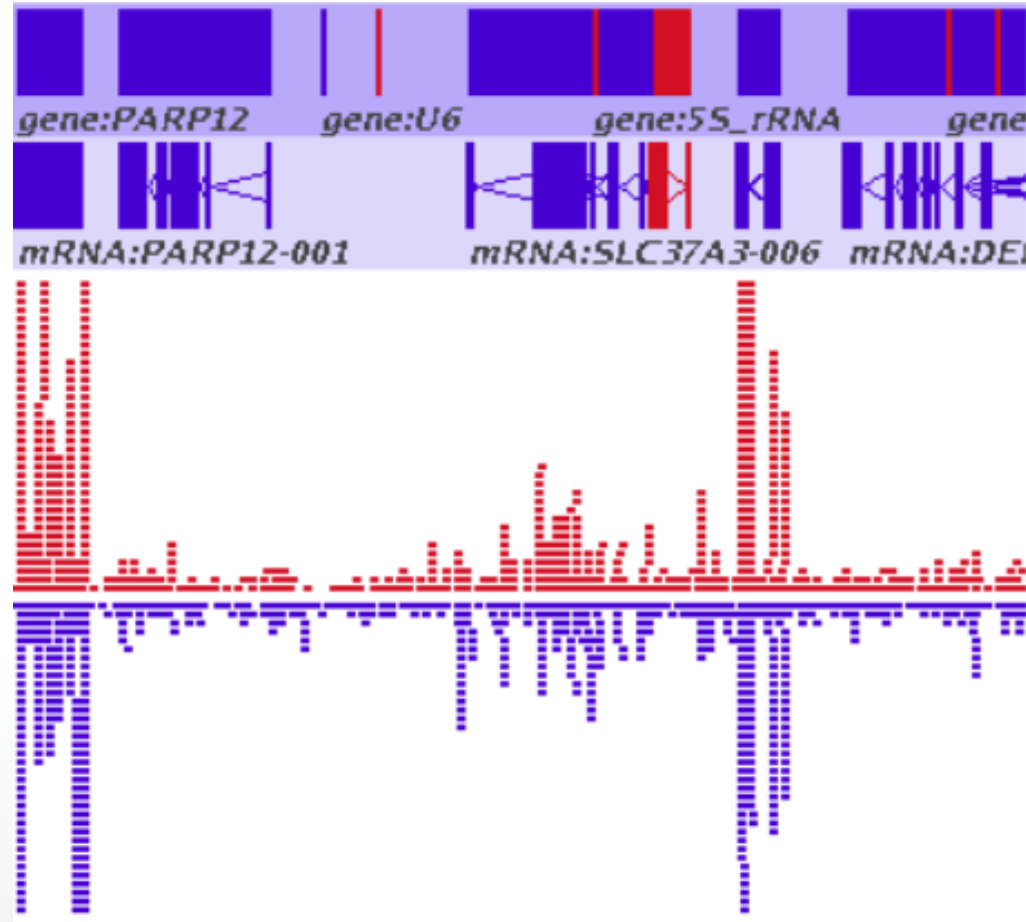
Summary

- ✓ [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



Stage 2: Mapping of sequencing reads to genome



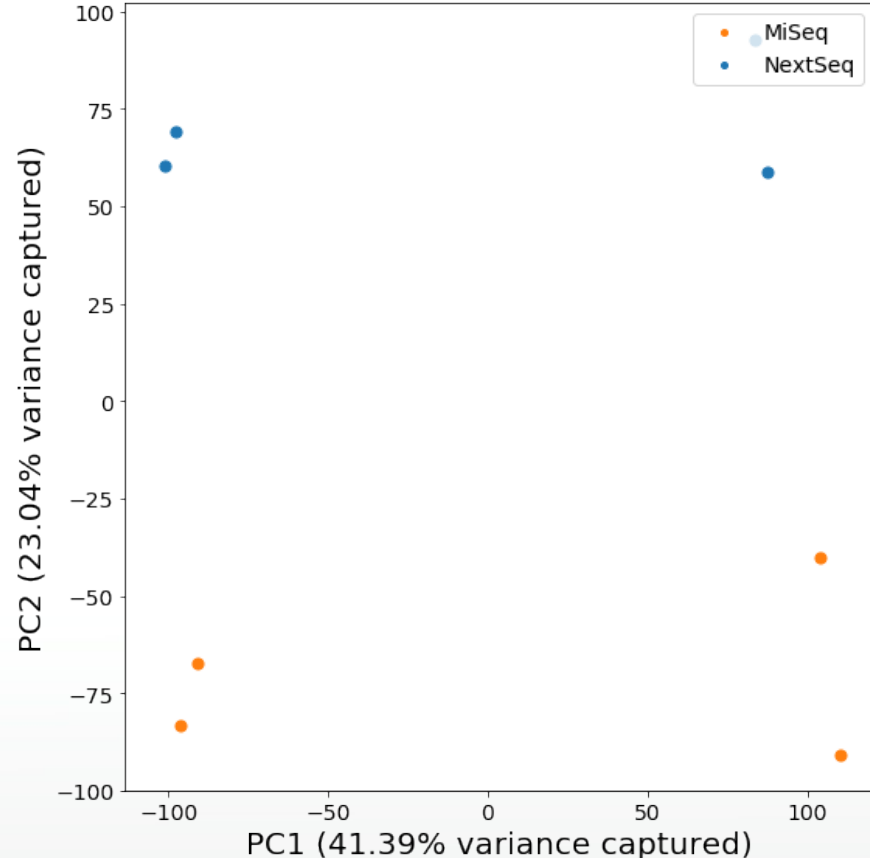


Stage 3: Assignment of reads to individual genes to attain expression measurements

- 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
 - $[(\text{Read count})/(\text{Gene length in kb})] / ($
 - RPKM – reads per kilobase per million

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?

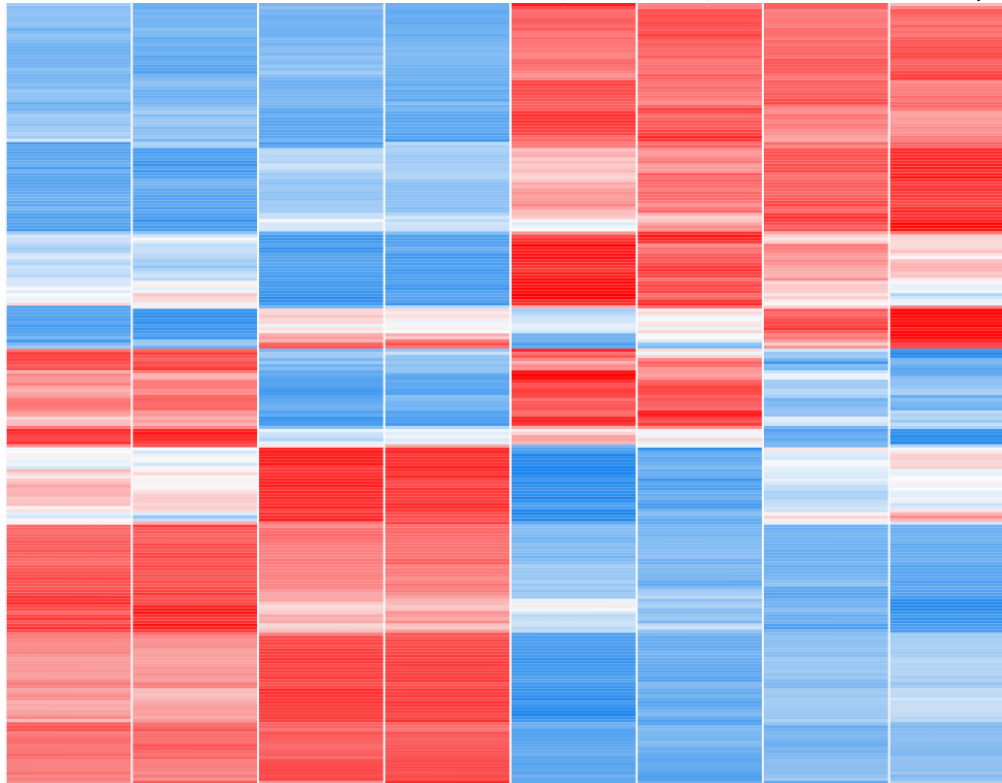


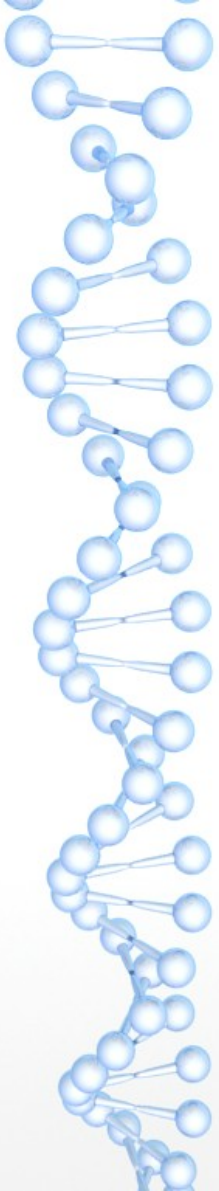
What is the structure in my expression data?

FO
02

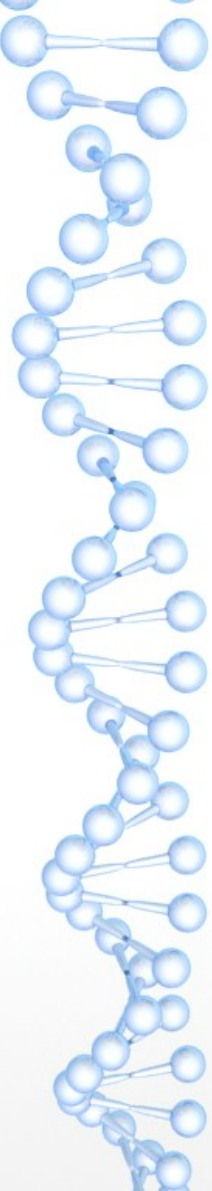


Zika-infected Sample 1 (MiSeq)
Zika-infected Sample 2 (MiSeq)
Zika-infected Sample 3 (NextSeq)
Zika-infected Sample 4 (NextSeq)
Mock Sample 1 (MiSeq)
Mock Sample 2 (MiSeq)
Mock Sample 3 (NextSeq)
Mock Sample 4 (NextSeq)





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



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

An unsolicited advertisement



ISCB-Africa ASBCB Conference on Bioinformatics

Kumasi, Ghana
November 11-15, 2019

iSCB
INTERNATIONAL SOCIETY FOR
COMPUTATIONAL BIOLOGY
Official Conference

ASBCB

Kumasi, Ghana
November 11-15,
2019

Mark your calendars!

Oral Presentation Submission Deadline: September 13, 2019

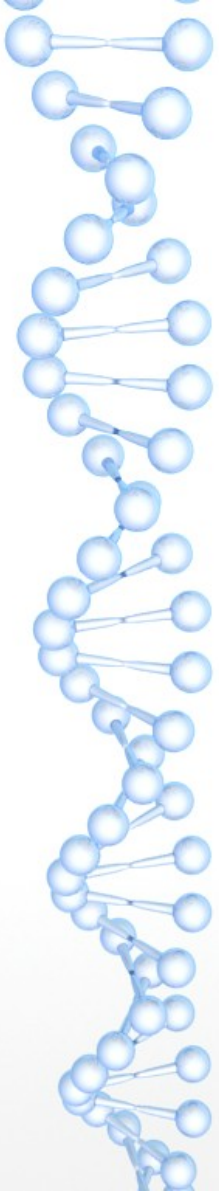
Poster Presentation Submission Deadline: October 15, 2019

<https://www.iscb.org/iscbafrica2019>



Additional resources

- Galaxy Community Hub's RNA-Seq Introduction:
https://galaxyproject.org/tutorials/rb_rnaseq/
- 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?