Cold Spring Harbor Laboratory
DNA Learning Center

DNALC Live

RNA-Seq with *DNA Subway* Part III

Jason Williams
Cold Spring Harbor Laboratory, DNA Learning Center
williams@cshl.edu
@JasonWilliamsNY

DNALC *Live*

This is an experiment, give us feedback on what you would like to see!

DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live

DNALC Website and Social Media

youtube.com/DNALearningCenter

facebook.com/cshldnalc

@dnalc

@dna_learning_center

Who is this course for?

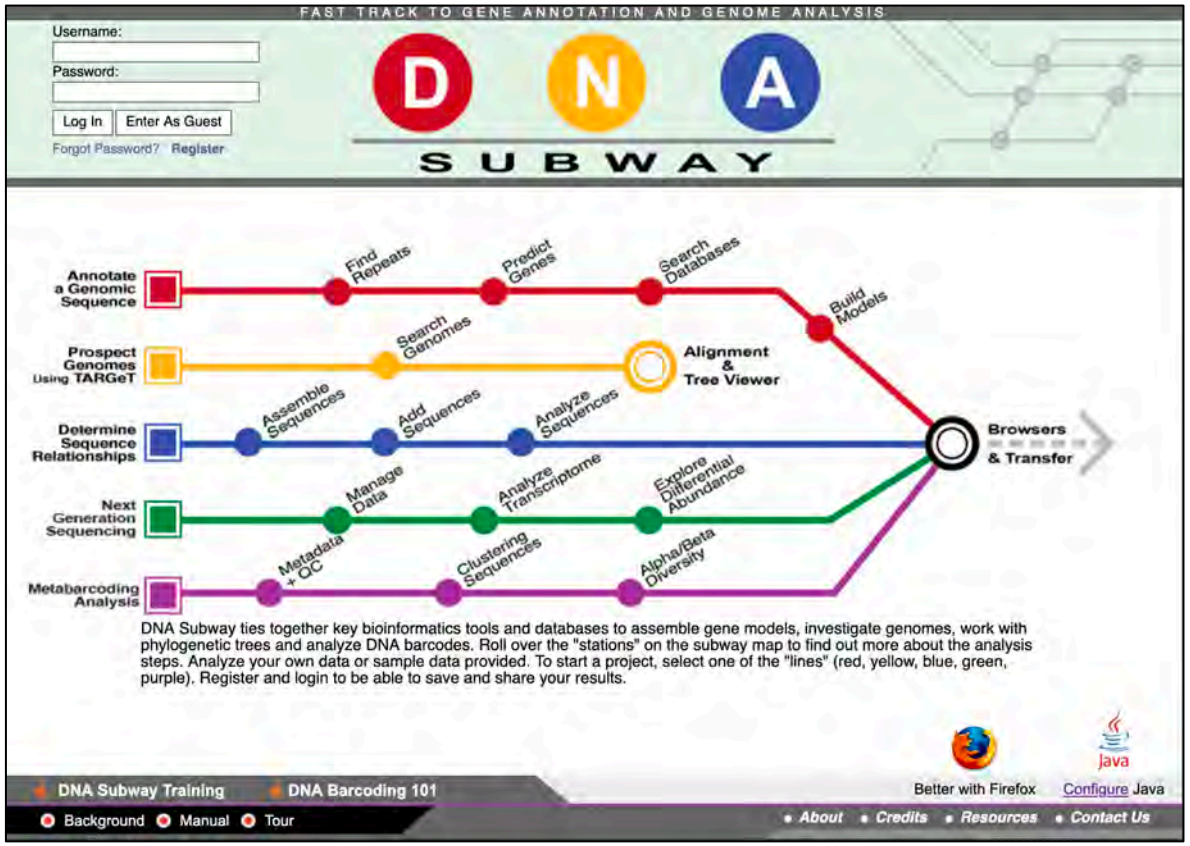
- Audience(s):
  - Undergraduate biology 200 level and up
  - (advanced AP Bio/graduate)
- Format: 3 sessions (1 per week); ~ 45 minutes each
- Exercises: Follow along with our online bioinformatics tool *DNA Subway*
- Learning resources: Slides and resource sheets available

Course Learning Goals

- Understand the rationale of an RNA-Seq experiment and its design
- Understand how we obtain DNA sequence and access its quality
- Use *DNA Subway* (*FastQC*/*FastX*) to QC sequence data
- Use *DNA Subway* (*Kallisto*) to (pseudo)align reads
- Use *DNA Subway* (*Sleuth*) to explore RNA-Seq results

Lab Setup

- We will be using *DNA Subway* – You can get a free account at **cyverse.org** (required)



RNA-Seq with *DNA Subway*

Part III

(differential abundance/expression)

Steps for today's session

- Review our progress so far
- Learn about differential abundance
- Visualize and explore our results

Review of RNA-Seq

What is RNA-Seq? - measuring the transcriptome

- RNA-Seq allows us to measure the transcriptome – take an account of all transcription occurring in a cell/tissue
- We use the abundance of an RNA transcript as a proxy for the activity of some cellular process (e.g. protein synthesis, regulatory activity)
- We analyze these data to compare samples (e.g. cancerous vs. non-cancerous)

Key Concept: Variation vs. Difference

Spot the difference – biological variation

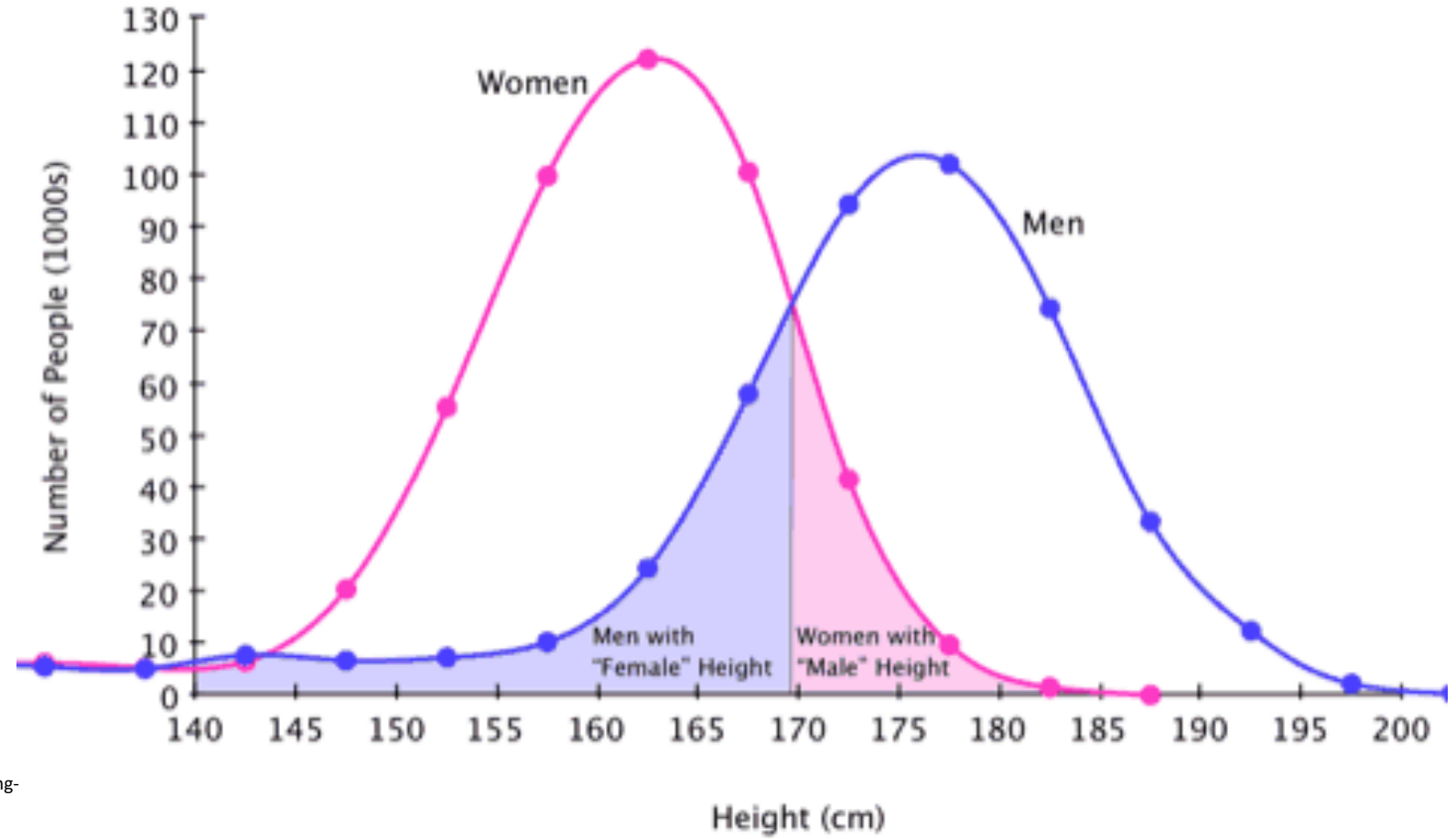


Photo Credit:
https://www.quora.com/What-are-Overlapping-Bell-Curves-and-how-do-they-affect-Quora-questions-and-answers

Introduction to our data set

RNA-Seq of hNPC – Zika Virus

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Zika infection of neural progenitor cells perturbs transcription in neurodevelopmental pathways

Lynn Yi, Harold Pimentel, Lior Pachter

Published: April 27, 2017 • https://doi.org/10.1371/journal.pone.0175744 • >> See the preprint

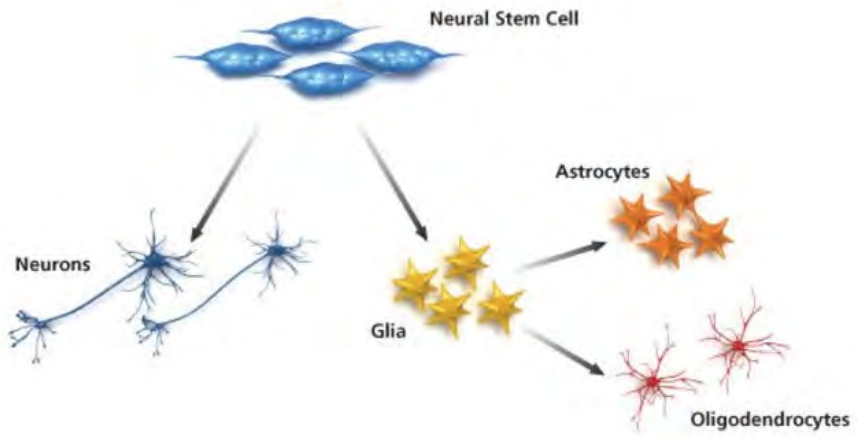
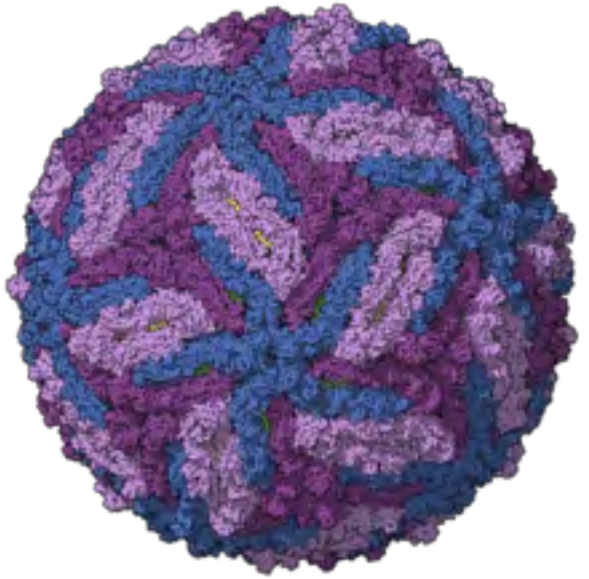
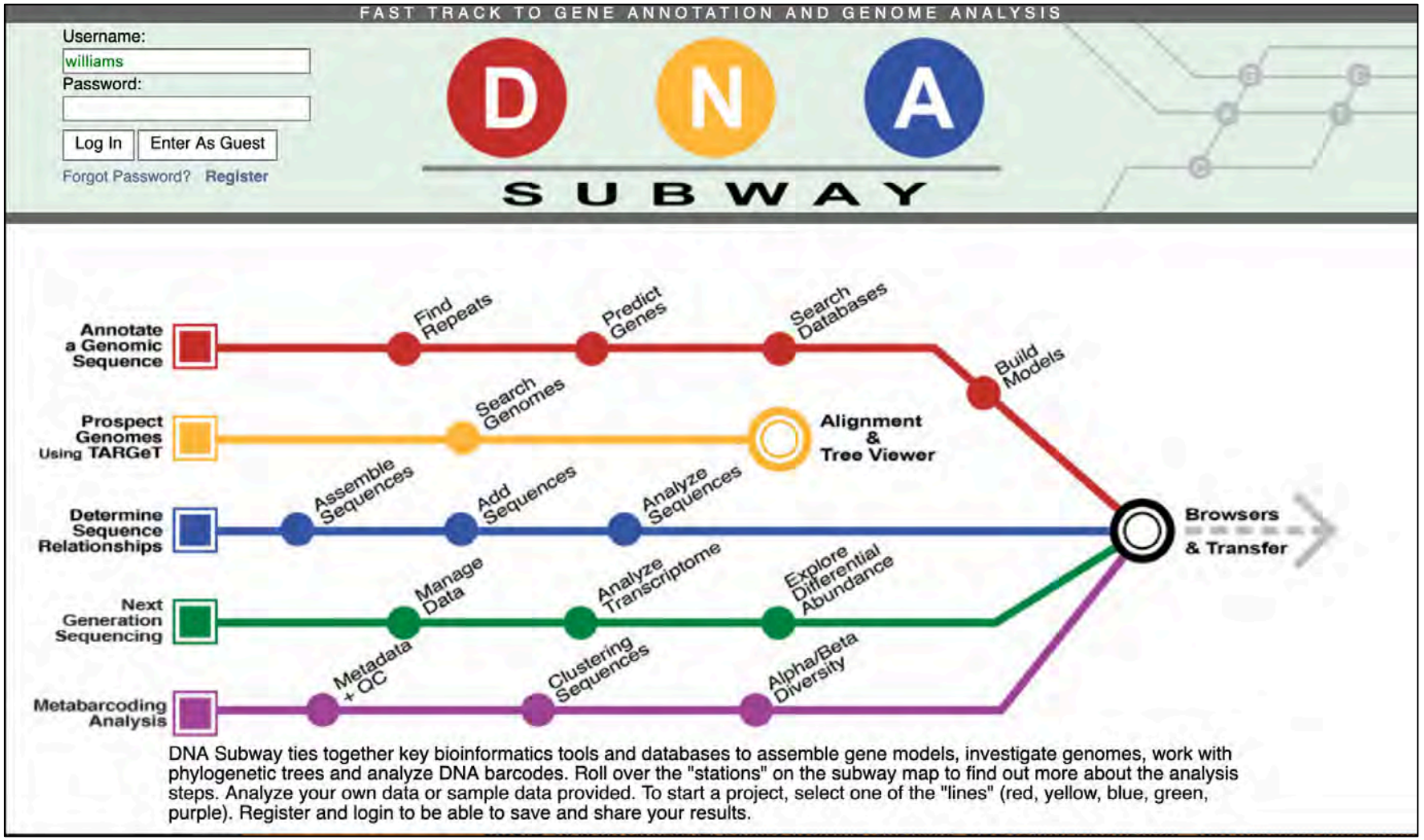


Photo credit:
https://www.sigmaaldrich.com/life-science/cell-biology/neural-stem-cell-biology.html
https://en.wikipedia.org/wiki/Zika_virus#/media/File:Zika-chain-colored.png

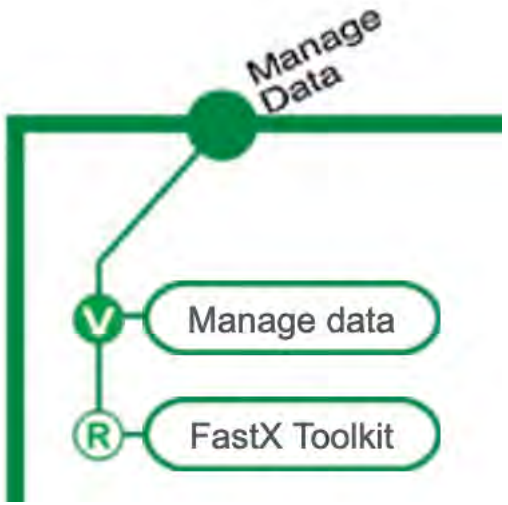


Zika Virus

Working on *DNA Subway* Green Line



Working on *DNA Subway* Green Line



Key Concept: Sequence Quality

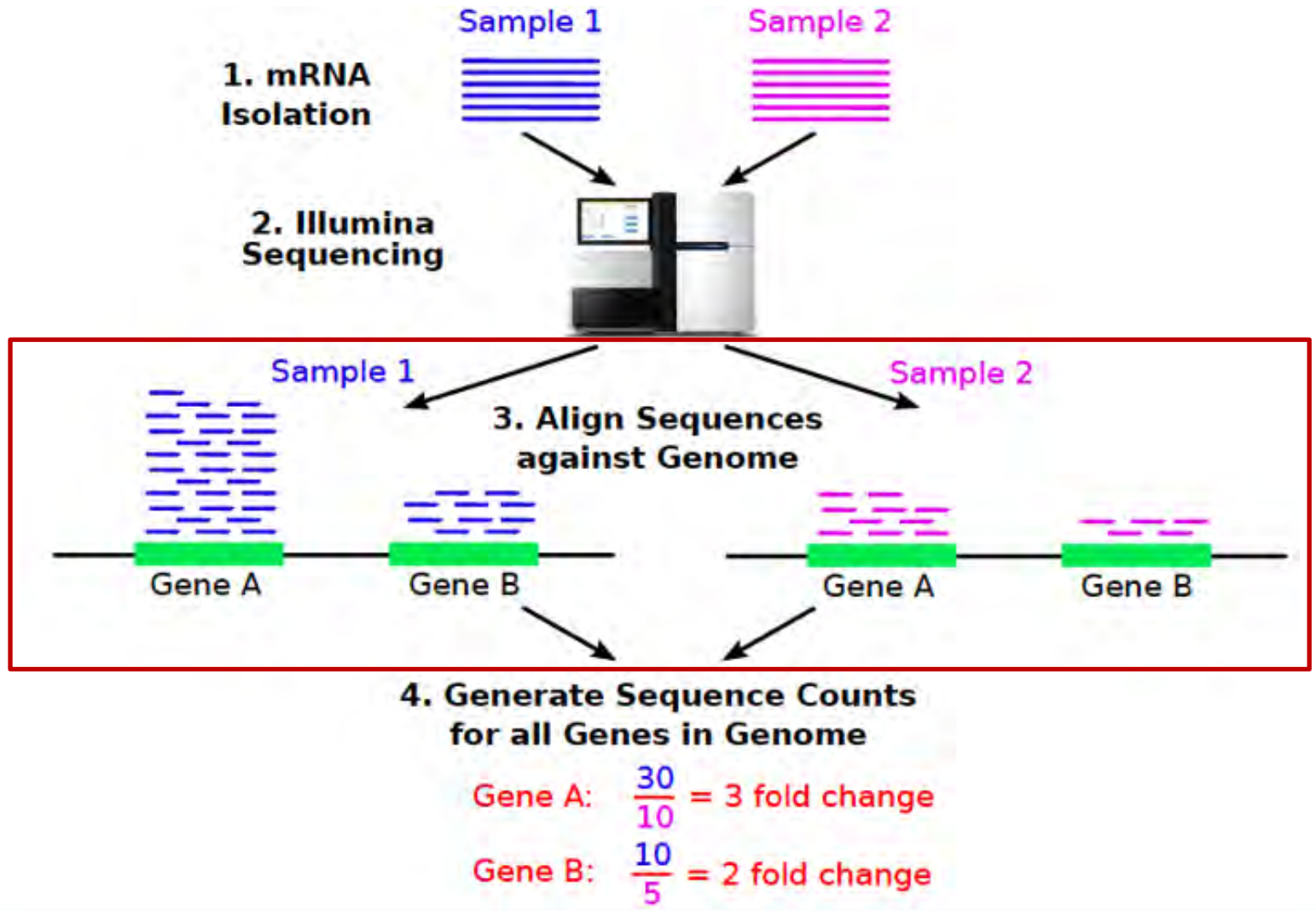
Phred scores…

Phred Score	Error (bases miscalled)	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

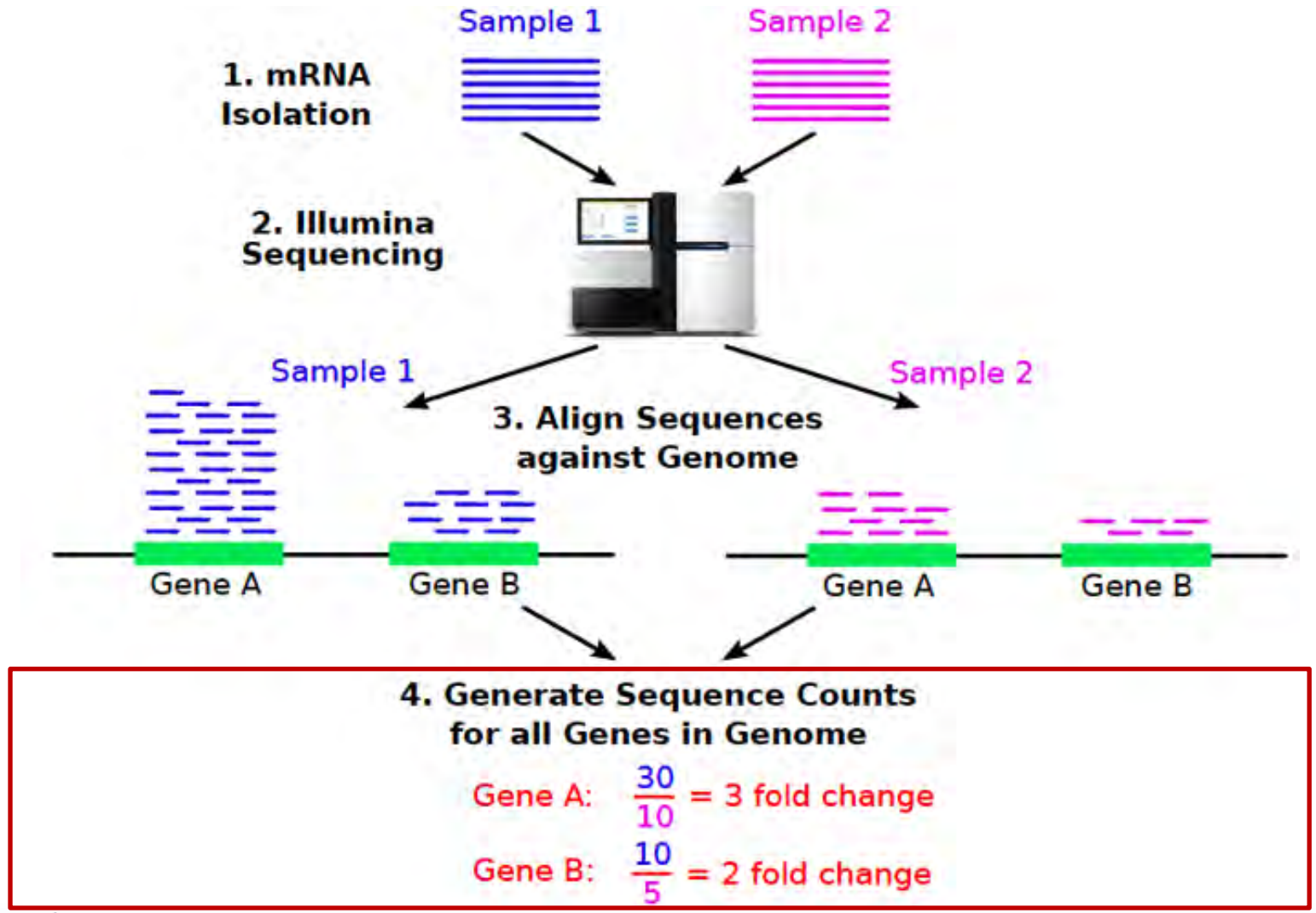
Key Concept: Read Alignment

Intuition: The more reads we observe from a given “gene” the more “active” that gene is

Counting reads



Counting reads



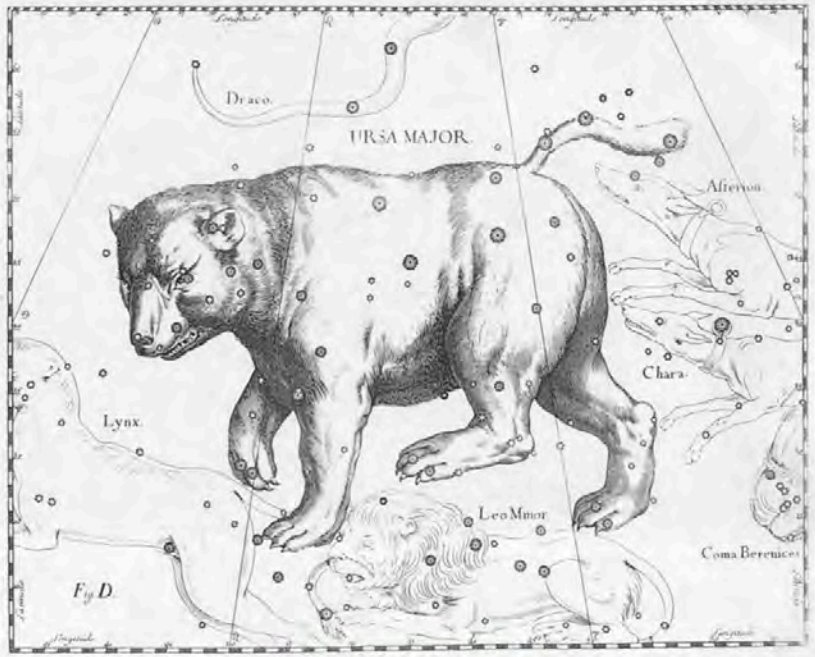
RNA-Seq with Kallisto

nature biotechnology

NATURE BIOTECHNOLOGY VOLUME 34 NUMBER 5 MAY 2016

Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray[1], Harold Pimentel[2], Páll Melsted[3] & Lior Pachter[2,4,5]



Problem: A transcriptome (like a genome) contains thousands of transcripts. How will we match sequence reads with transcripts?

Kallisto – Pseudoalignment

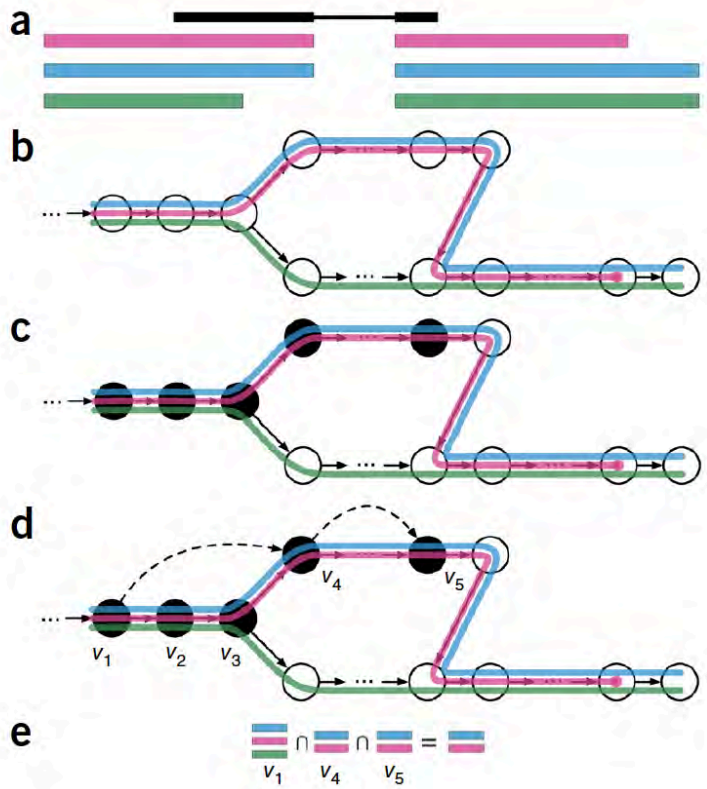
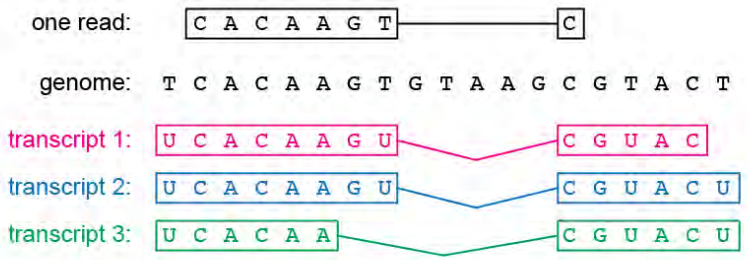


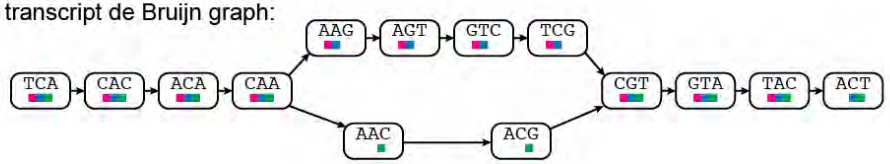
Figure 1 Overview of kallisto. The input consists of a reference transcriptome and reads from an RNA-seq experiment. (**a**) An example of a read (in black) and three overlapping transcripts with exonic regions as shown. (**b**) An index is constructed by creating the transcriptome de Bruijn Graph (T-DBG) where nodes (v_1, v_2, v_3, ...) are *k*-mers, each transcript corresponds to a colored path as shown and the path cover of the transcriptome induces a *k*-compatibility class for each *k*-mer. (**c**) Conceptually, the *k*-mers of a read are hashed (black nodes) to find the *k*-compatibility class of a read. (**d**) Skipping (black dashed lines) uses the information stored in the T-DBG to skip *k*-mers that are redundant because they have the same *k*-compatibility class. (**e**) The *k*-compatibility class of the read is determined by taking the intersection of the *k*-compatibility classes of its constituent *k*-mers.

Photo credit:
https://www.nature.com/articles/nbt.3519

Kallisto – Pseudoalignment



Colored transcript de Bruijn graph:



Matching read to graph:

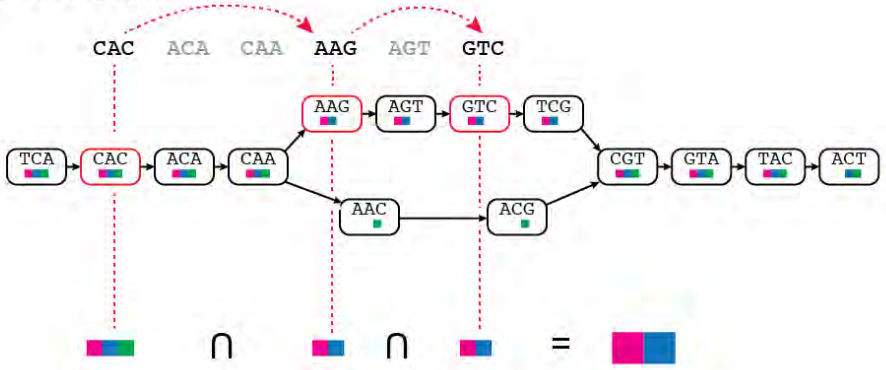
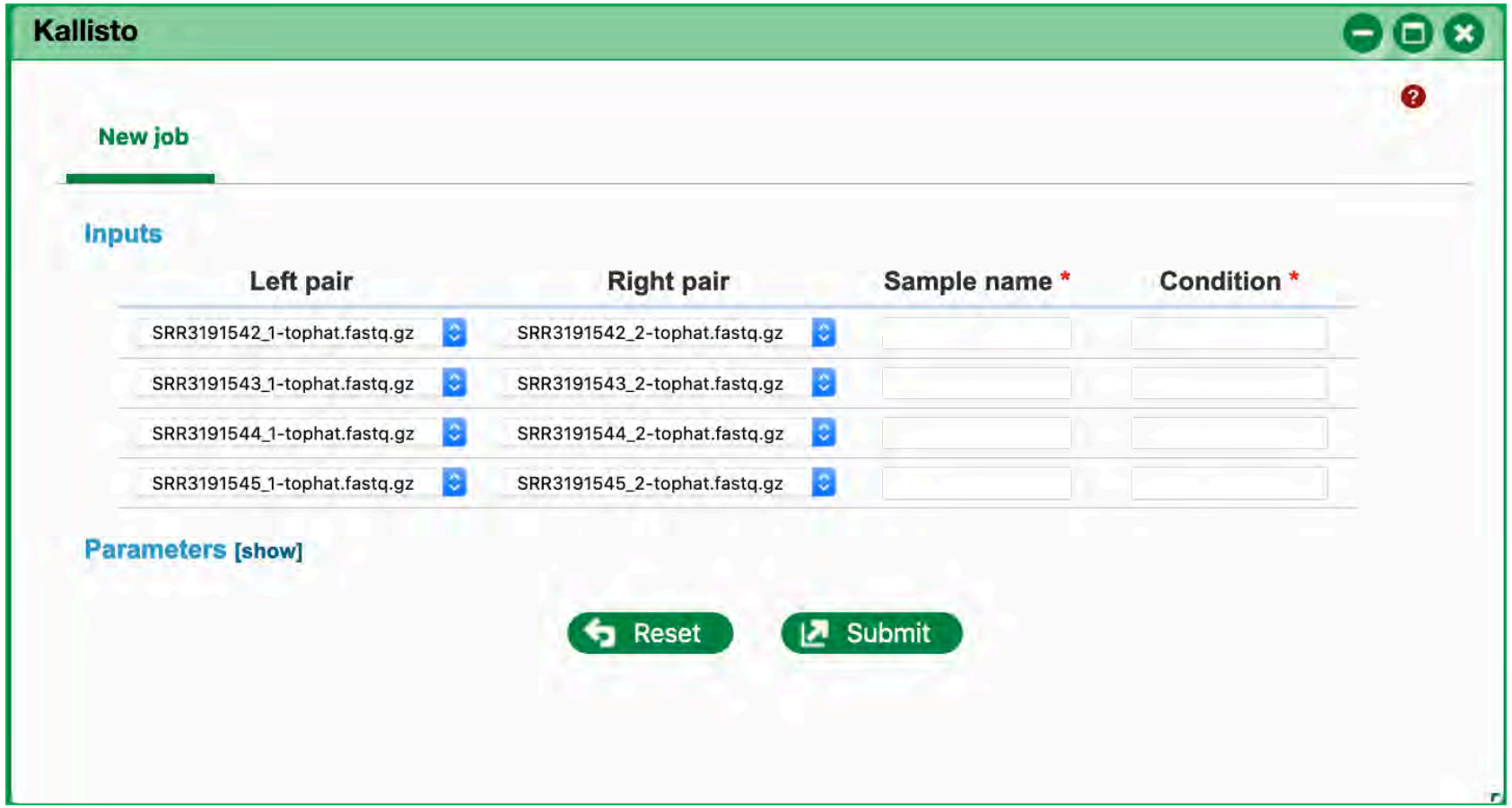


Photo credit:
http://mcb112.org/w02/w02-lecture.html

Lab: Pseudoalignment with Kallisto

Lab: Kallisto



Lab: Kallisto

Left/Right Pair	Sample name	Condition
SRR3191542_1.fastq.gz SRR3191542_2.fastq.gz	Mock1-1	Mock
SRR3191543_1.fastq.gz SRR3191543_1.fastq.gz	Mock2-1	Mock
SRR3191544_1.fastq.gz SRR3191544_2.fastq.gz	ZIKV1-1	Zika
SRR3191545_1.fastq.gz SRR3191545_2.fastq.gz	ZIKV2-1	Zika

See *DNA Subway* Guide (Green Line) on learning.cyverse.org

Kallisto results

	A	B	C	D	E
1	target_id	length	eff_length	est_counts	tpm
2	ENST00000361624.2	1542	1366.02	70979.1	14946.3
3	ENST00000361739.1	684	508.114	25163	14245
4	ENST00000362079.2	784	608.064	18924	8952.1
5	ENST00000361851.1	207	53.9295	1592.53	8494.23
6	ENST00000361899.2	681	505.114	13043.3	7427.79
7	ENST00000361381.2	1378	1202.02	30926	7400.69
8	ENST00000361335.1	297	127.756	3008	6772.67
9	ENST00000331523.6	1923	1747.02	35334.8	5817.9
10	ENST00000361681.2	525	349.373	5789.99	4767.05

Kallisto results

- target_id: Identifier for the transcript (from Ensembl)

Kallisto results

Kallisto results

- target_id: Identifier for the transcript (from Ensembl)
- length: length (nucleotides) of transcript exons

Kallisto results

- target_id: Identifier for the transcript (from Ensembl)
- length: length (nucleotides) of transcript exons
- eff_length: length of transcript that was sampled*

*In the original sequencing library, we rarely sample whole entire transcripts, this number accounts for the fragment length of the library

Kallisto results

- target_id: Identifier for the transcript (from Ensembl)
- length: length (nucleotides) of transcript exons
- eff_length: length of transcript that was sampled*
- est_counts: The estimated number of reads that have mapped to the transcript

*In the original sequencing library, we rarely sample whole entire transcripts, this number accounts for the fragment length of the library

Key Concept: Normalization

(Warning – illustrative “toy” models ahead)

Transcripts per million

- tpm (transcripts per million) normalized counts based on the length of the transcript and total number of sequence reads

Transcripts per million

- tpm (transcripts per million) normalized counts based on the length of the transcript and total number of sequence reads

$$\text{Transcripts per million} \equiv A \cdot \left(\frac{1}{\sum(A)}\right) \cdot 10^6$$

Transcripts per million

- tpm (transcripts per million) normalized counts based on the length of the transcript and total number of sequence reads

$$\text{Transcripts per million} \equiv A \cdot \left(\frac{1}{\sum(A)}\right) \cdot 10^6$$

$$A = \frac{total\ reads\ mapped\ to\ gene\ \cdot\ 10^3}{gene\ length\ (bp)}$$

Normalization – gene length

Which is longer (bp)?

Gene A

Gene B

Normalization – gene length

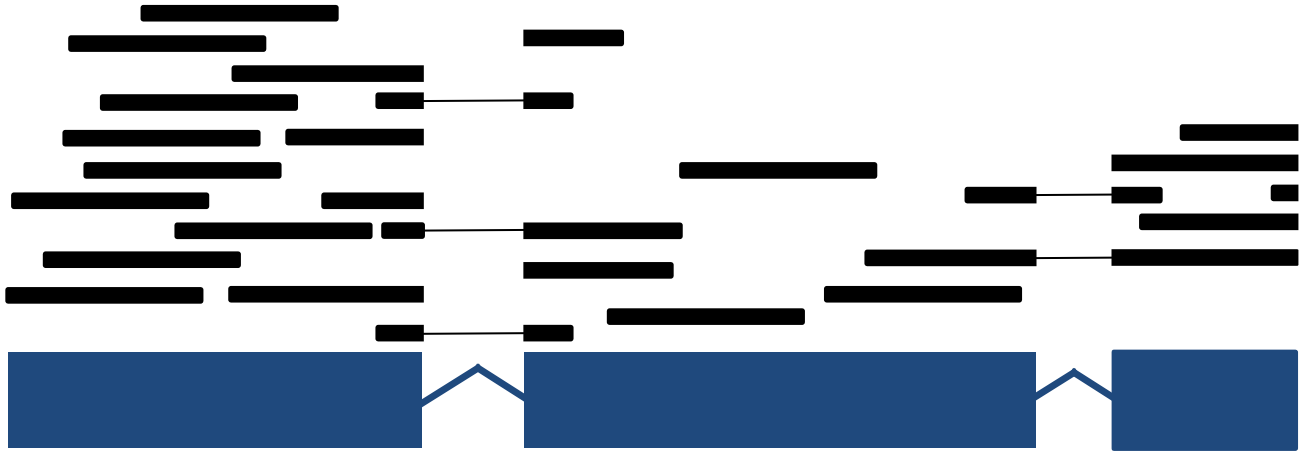
Which has more reads?

Gene A
(300bp)

Gene B
(100bp)

Normalization – gene length

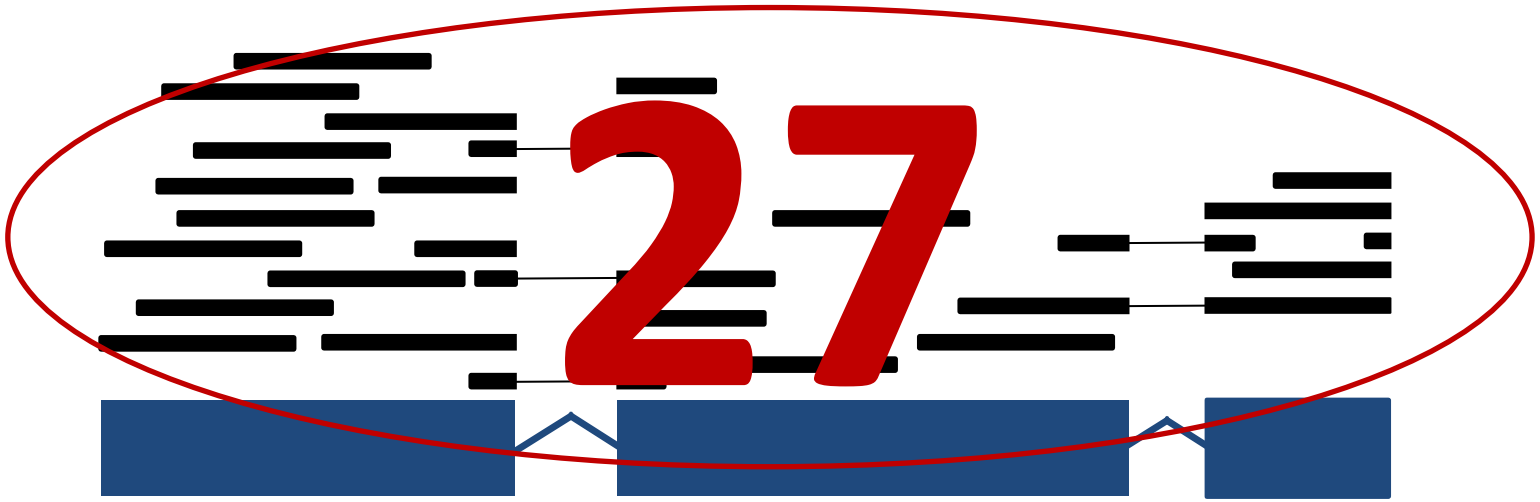
Which has more reads?



Gene A
(300bp)

Gene B
(100bp)

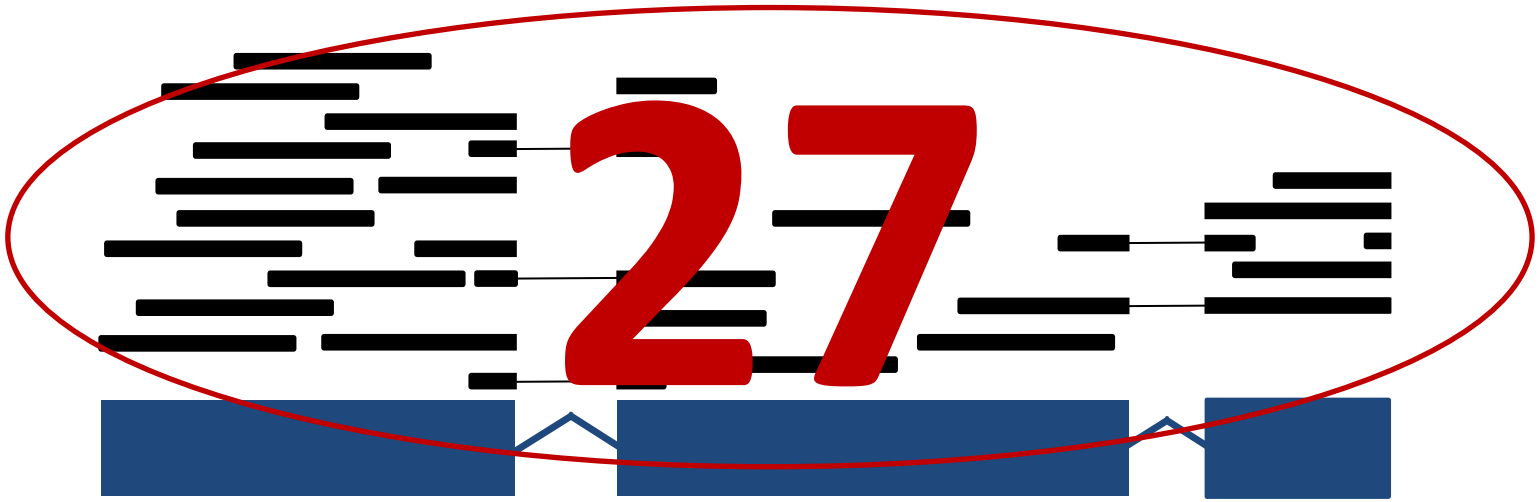
Normalization – gene length



Gene A
(300bp)

Gene B
(100bp)

Normalization – gene length



Gene A
(300bp)

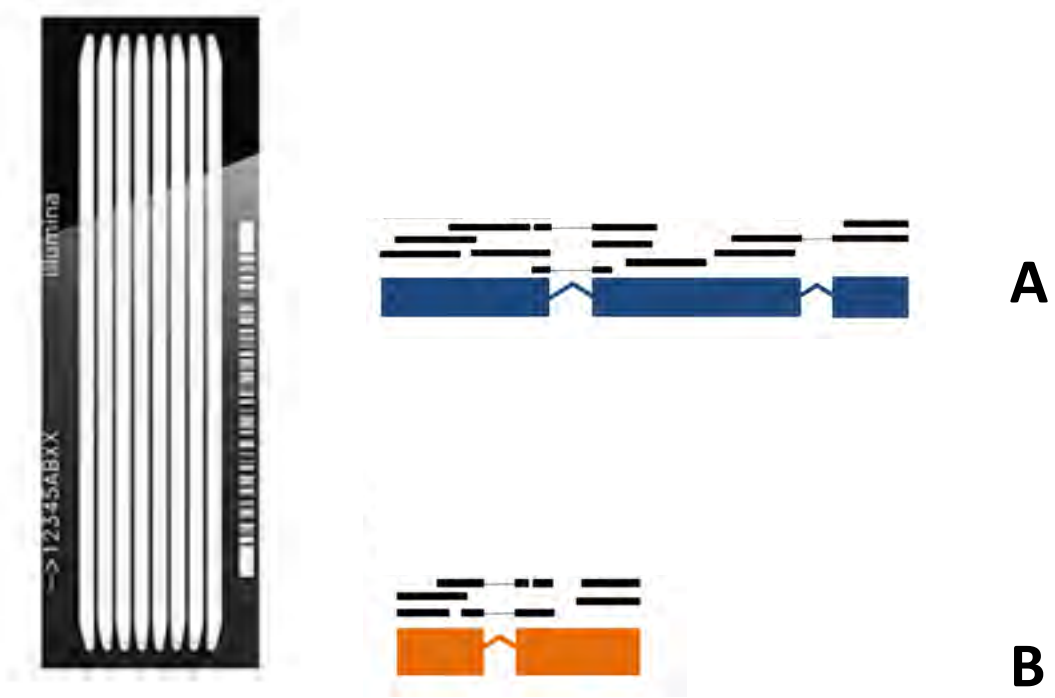
27/300 = 0.09

Gene B
(100bp)

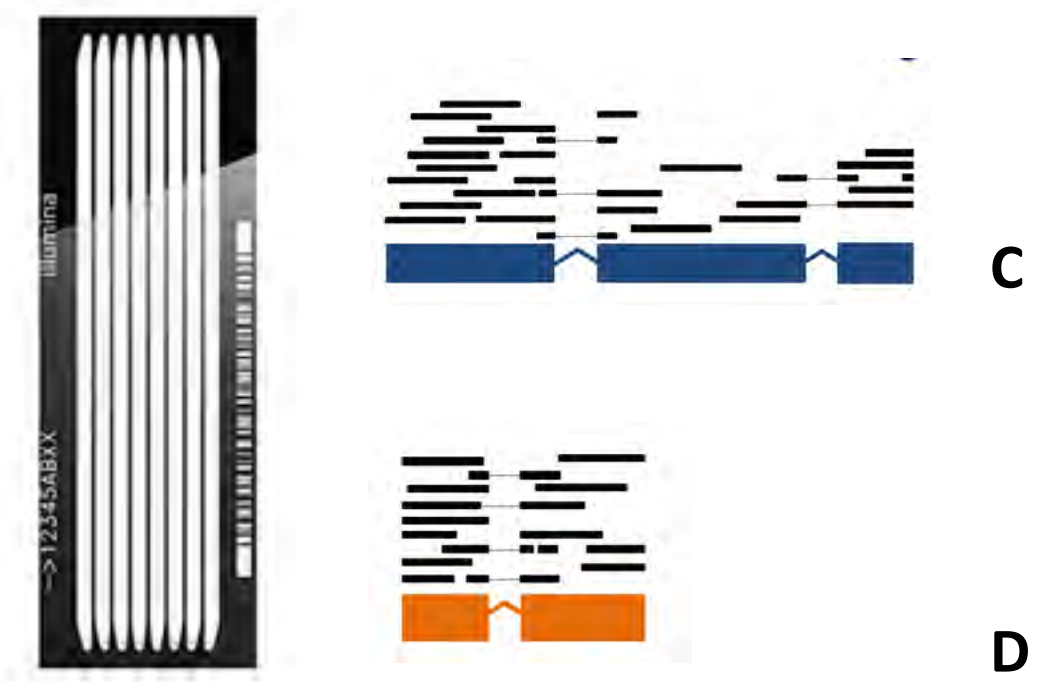
16/100 = 0.16

Normalization – read depth

Which gene is most highly expressed?



Flow cell 1



Flow cell 2

Photo credit
https://www.illumina.com/company/news-center/multimedia-images.html

Normalization – read depth

Which gene is most highly expressed?

Flow cell 1 (10M reads)

Flow cell 2 (20M reads)

Photo credit
https://www.illumina.com/company/news-center/multimedia-images.html

RNA-Seq with Sleuth

nature methods

Brief Communication | Published: 05 June 2017

Differential analysis of RNA-seq incorporating quantification uncertainty

Harold Pimentel, Nicolas L Bray, Suzette Puente, Páll Melsted & Lior Pachter

Nature Methods **14**, 687–690(2017) | Cite this article

6755 Accesses | **264** Citations | **121** Altmetric | Metrics

In an RNA-Seq experiment we need to find "true" differences between samples (while subtracting trivial differences)

Spot the difference

Photo Credit:
https://www.rd.com/culture/spot-10-differences-picture/

Spot the difference

Mock – 1, Replicate 1

	A	B	C	D	E
1	target_id	length	eff_length	est_counts	tpm
2	ENST00000361624.2	1542	1366.02	70979.1	14946.3
3	ENST00000361739.1	684	508.114	25163	14245
4	ENST00000362079.2	784	608.064	18924	8952.1
5	ENST00000361851.1	207	53.9295	1592.53	8494.23
6	ENST00000361899.2	681	505.114	13043.3	7427.79
7	ENST00000361381.2	1378	1202.02	30926	7400.69
8	ENST00000361335.1	297	127.756	3008	6772.67
9	ENST00000331523.6	1923	1747.02	35334.8	5817.9
10	ENST00000361681.2	525	349.373	5789.99	4767.05

Mock – 1, Replicate 2

	A	B	C	D	E
1	target_id	length	eff_length	est_counts	tpm
2	ENST00000361624.2	1542	1366.02	70979.1	14946.3
3	ENST00000361739.1	684	508.114	25163	14245
4	ENST00000362079.2	784	608.064	18924	8952.1
5	ENST00000361851.1	207	53.9295	1592.53	8494.23
6	ENST00000361899.2	681	505.114	13043.3	7427.79
7	ENST00000361381.2	1378	1202.02	30926	7400.69
8	ENST00000361335.1	297	127.756	3008	6772.67
9	ENST00000331523.6	1923	1747.02	35334.8	5817.9
10	ENST00000361681.2	525	349.373	5789.99	4767.05

Zika – 1, Replicate 1

	A	B	C	D	E
1	target_id	length	eff_length	est_counts	tpm
2	ENST00000361624.2	1542	1366.02	70979.1	14946.3
3	ENST00000361739.1	684	508.114	25163	14245
4	ENST00000362079.2	784	608.064	18924	8952.1
5	ENST00000361851.1	207	53.9295	1592.53	8494.23
6	ENST00000361899.2	681	505.114	13043.3	7427.79
7	ENST00000361381.2	1378	1202.02	30926	7400.69
8	ENST00000361335.1	297	127.756	3008	6772.67
9	ENST00000331523.6	1923	1747.02	35334.8	5817.9
10	ENST00000361681.2	525	349.373	5789.99	4767.05

Zika – 1, Replicate 2

	A	B	C	D	E
1	target_id	length	eff_length	est_counts	tpm
2	ENST00000361624.2	1542	1366.02	70979.1	14946.3
3	ENST00000361739.1	684	508.114	25163	14245
4	ENST00000362079.2	784	608.064	18924	8952.1
5	ENST00000361851.1	207	53.9295	1592.53	8494.23
6	ENST00000361899.2	681	505.114	13043.3	7427.79
7	ENST00000361381.2	1378	1202.02	30926	7400.69
8	ENST00000361335.1	297	127.756	3008	6772.67
9	ENST00000331523.6	1923	1747.02	35334.8	5817.9
10	ENST00000361681.2	525	349.373	5789.99	4767.05

Sleuth linear modeling

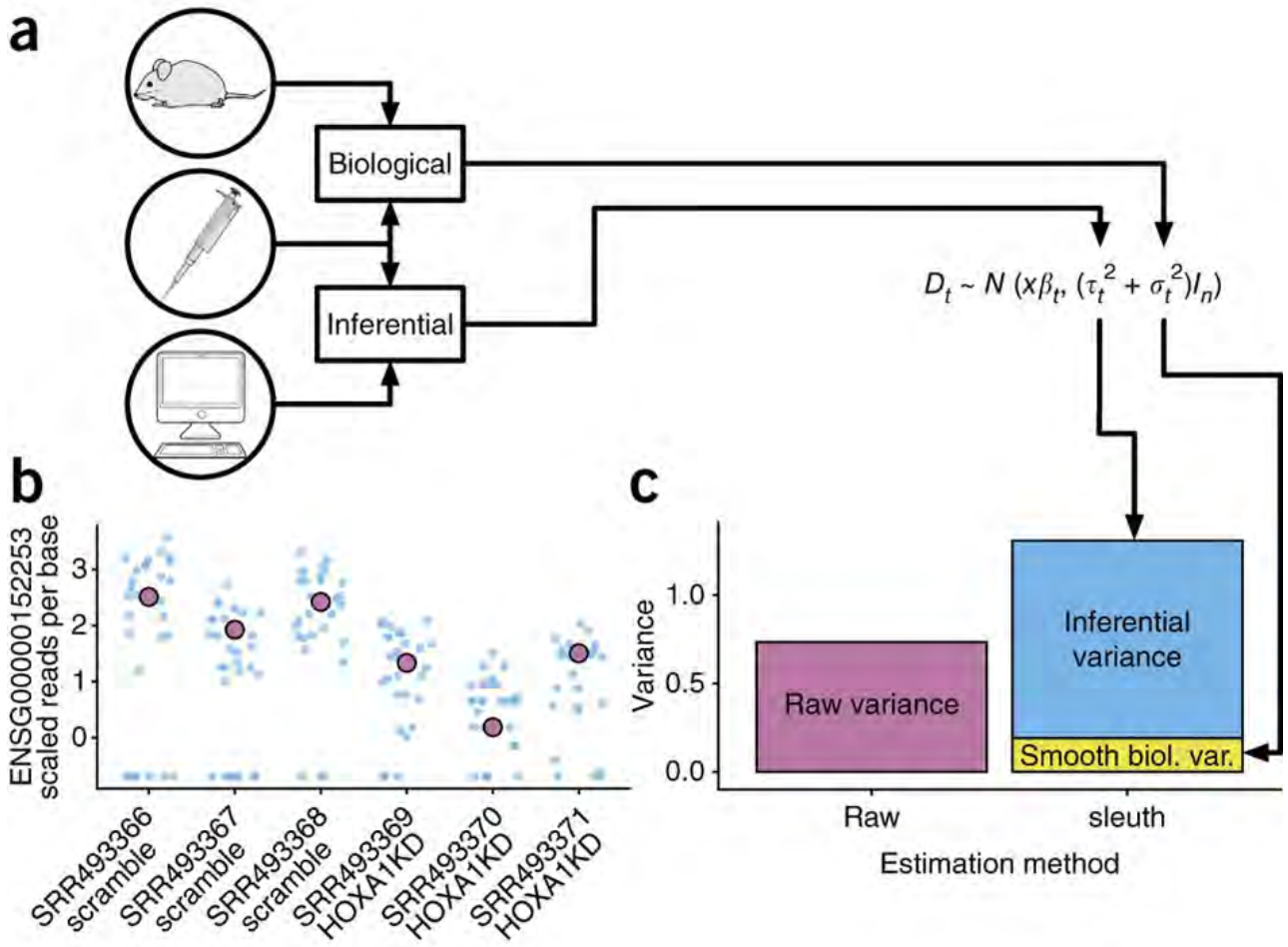


Photo credit
https://www.nature.com/articles/nmeth.4324/figures/1

Sleuth linear modeling

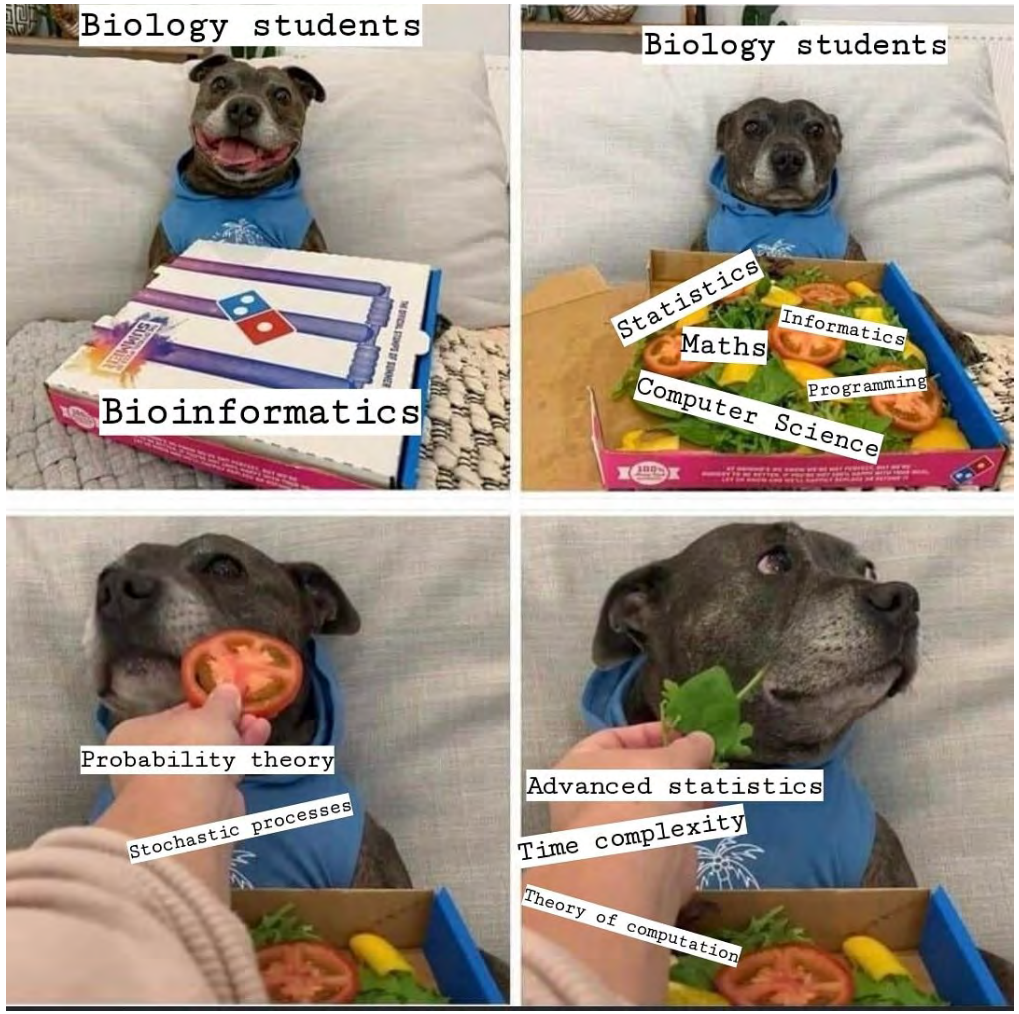
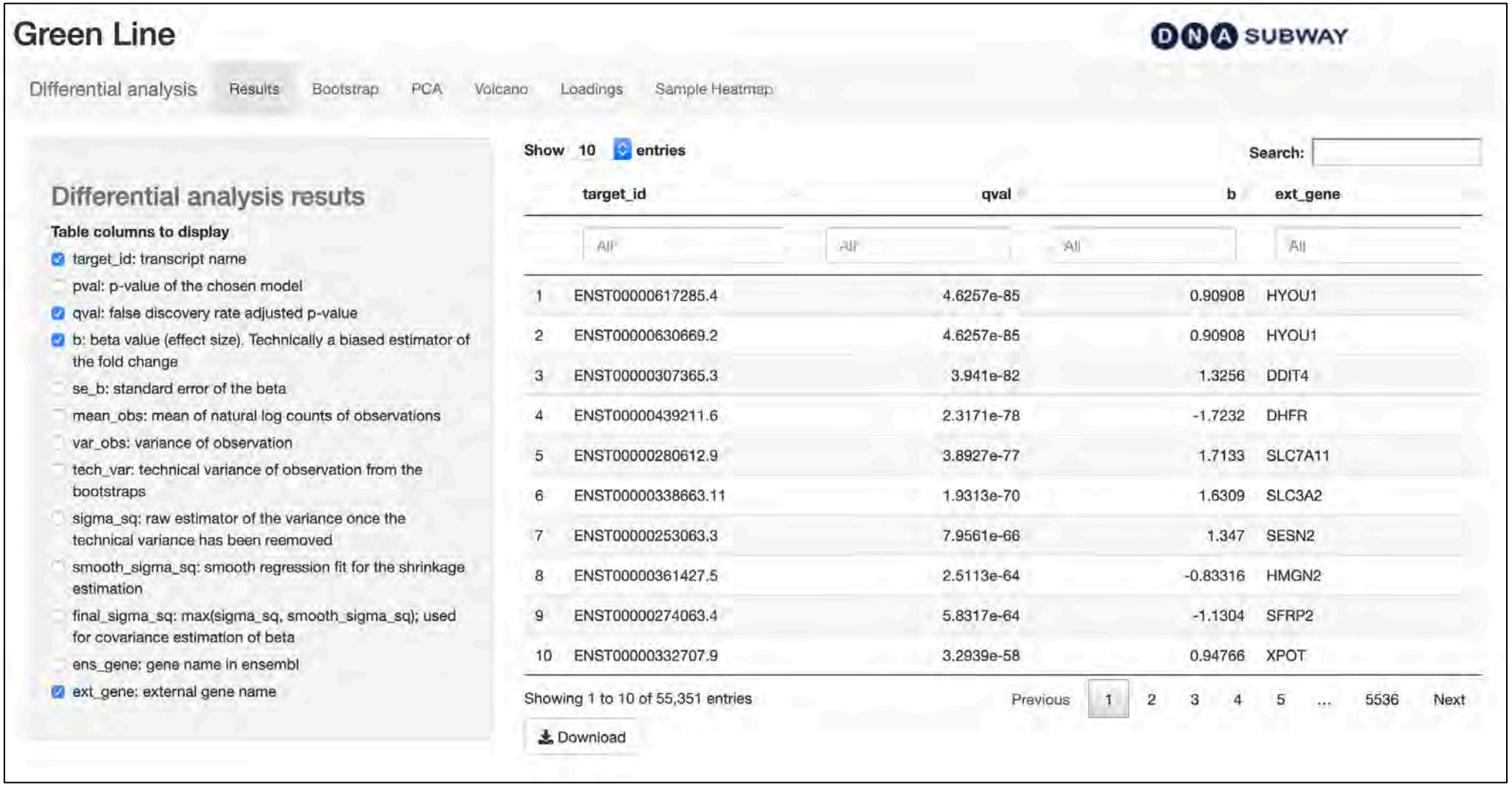
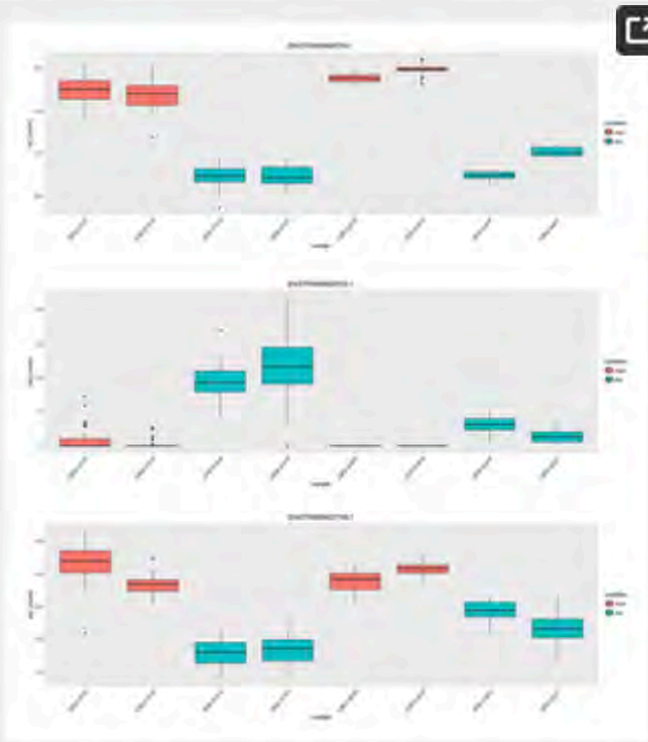


Photo credit
https://twitter.com/phylogeo

Lab: Sleuth results



Comparing to the results of the L. Yi paper



Download:

PPT PowerPoint slide

PNG larger image

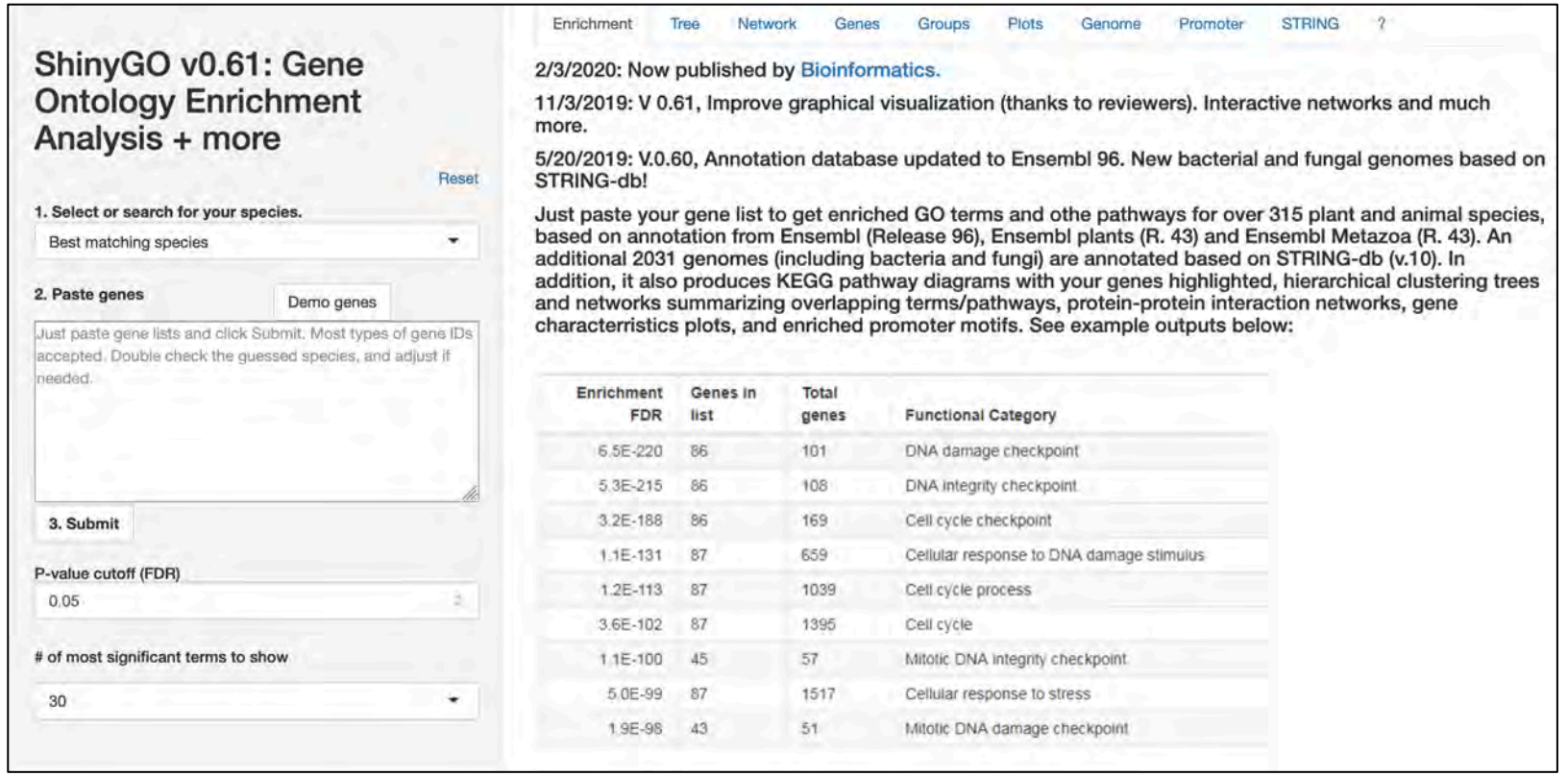
TIFF original image

Fig 4. The counts of CHRNA7, another isoform diverging gene, plotted by the sleuth Shiny app.

Visualized here are counts for three transcripts of CHRNA7 across eight samples, colored by infection status.

https://doi.org/10.1371/journal.pone.0175744.g004

Ontology enrichment: ShinyGO



http://bioinformatics.sdstate.edu/go/

Gene ontology

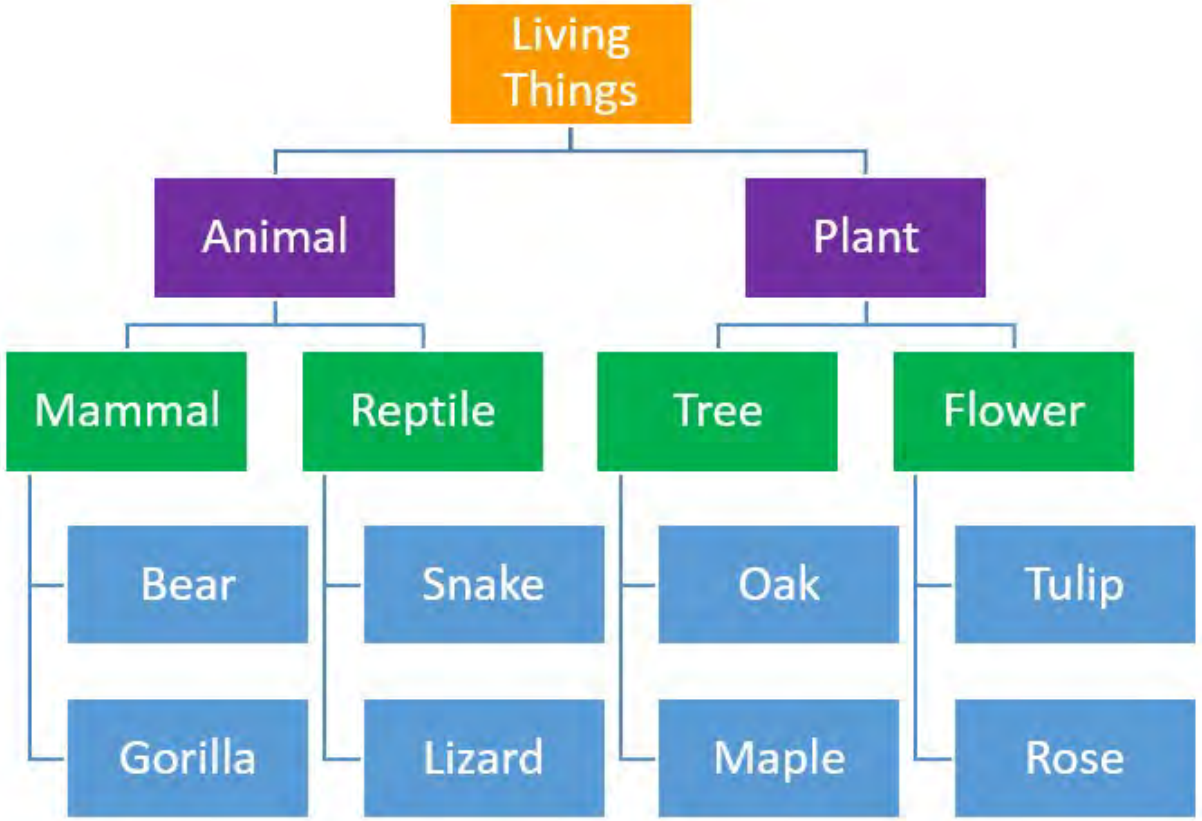


Photo credit:
https://thepeakperformancecenter.com/educational-learning/learning/memory/stages-of-memory/organization-long-term-memory/

Gene ontology

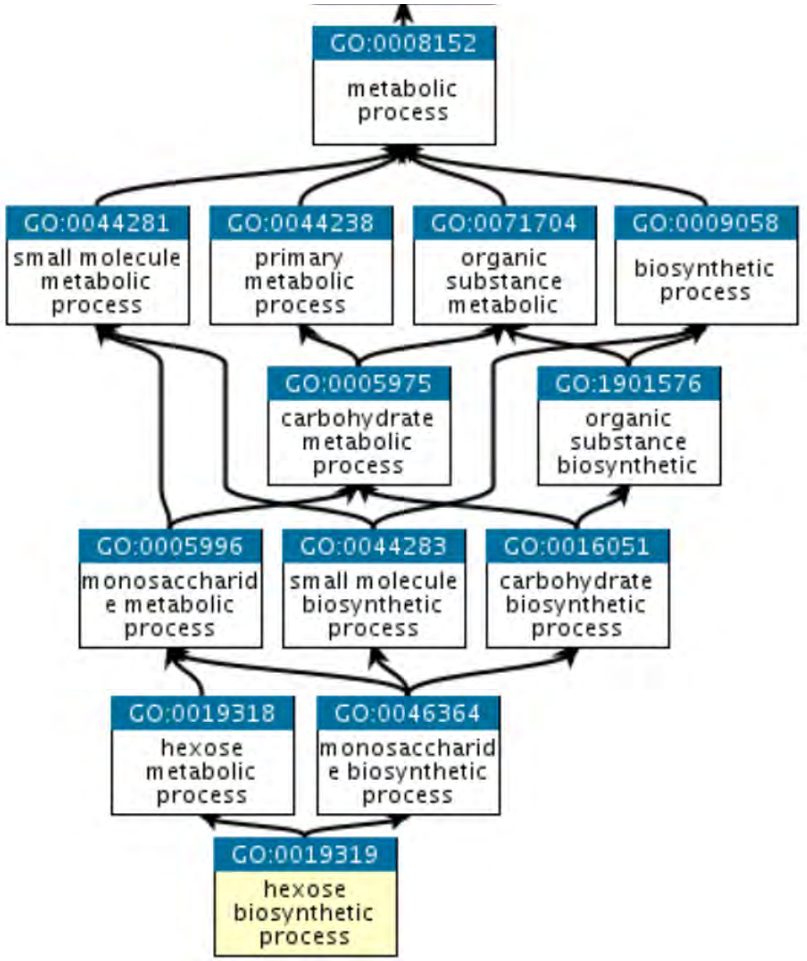


Photo credit:
http://geneontology.org/docs/ontology-documentation/

Tang paper ontologies

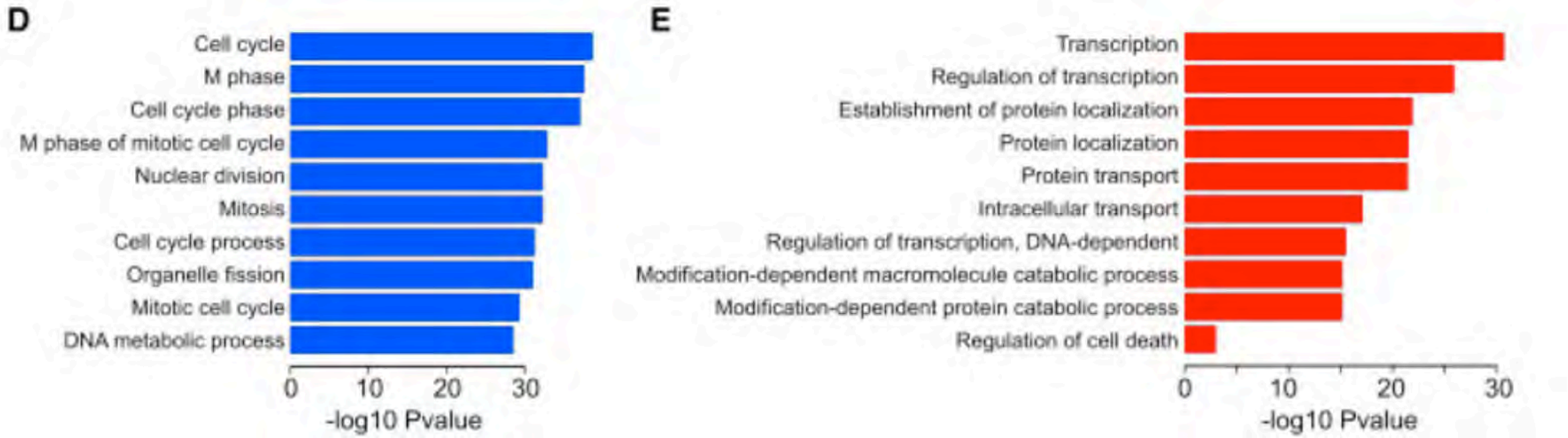


Figure 2 ZIKV-Infected hNPCs Exhibit Increased Cell Death and Dysregulated Cell-Cycle Progression and Gene Expression

Photo credit:
https://www.cell.com/cell-stem-cell/fulltext/S1934-5909(16)00106-5

Goal recap

- Understand the rationale of an RNA-Seq experiment and its design
- Understand how we obtain DNA sequence and access its quality
- Use *DNA Subway* (*FastQC*/*FastX*) to QC sequence data
- Use *DNA Subway* (*Kallisto*) to (pseudo)align reads
- Use *DNA Subway* (*Sleuth*) to explore RNA-Seq results

DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live