



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](https://twitter.com/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](https://twitter.com/dnalc)



[@dna_learning_center](https://instagram.com/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)

The screenshot displays the DNA Subway website interface. At the top, it reads "FAST TRACK TO GENE ANNOTATION AND GENOME ANALYSIS" above the "DNA SUBWAY" logo. A login form on the left includes fields for "Username:" and "Password:", with "Log In" and "Enter As Guest" buttons, and links for "Forgot Password?" and "Register".

The main content area features a subway map with five colored lines (red, yellow, blue, green, purple) representing different analysis paths. The red line includes stations: "Annotate a Genomic Sequence", "Find Repeats", "Predict Genes", "Search Databases", and "Build Models". The yellow line includes "Prospect Genomes Using TARGET", "Search Genomes", and "Alignment & Tree Viewer". The blue line includes "Determine Sequence Relationships", "Assemble Sequences", "Add Sequences", and "Analyze Sequences". The green line includes "Next Generation Sequencing", "Manage Data", "Analyze Transcriptome", and "Explore Differential Abundance". The purple line includes "Metabarcoding Analysis", "Metadata + QC", "Clustering Sequences", and "Alpha/Beta Diversity". All lines converge at a station labeled "Browsers & Transfer".

Below the map, a text block explains: "DNA Subway ties together key bioinformatics tools and databases to assemble gene models, investigate genomes, work with phylogenetic trees and analyze DNA barcodes. Roll over the 'stations' on the subway map to find out more about the analysis steps. Analyze your own data or sample data provided. To start a project, select one of the 'lines' (red, yellow, blue, green, purple). Register and login to be able to save and share your results."

The footer contains navigation links: "DNA Subway Training", "DNA Barcoding 101", "Background", "Manual", "Tour", "About", "Credits", "Resources", "Contact Us", and logos for Firefox and Java.

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

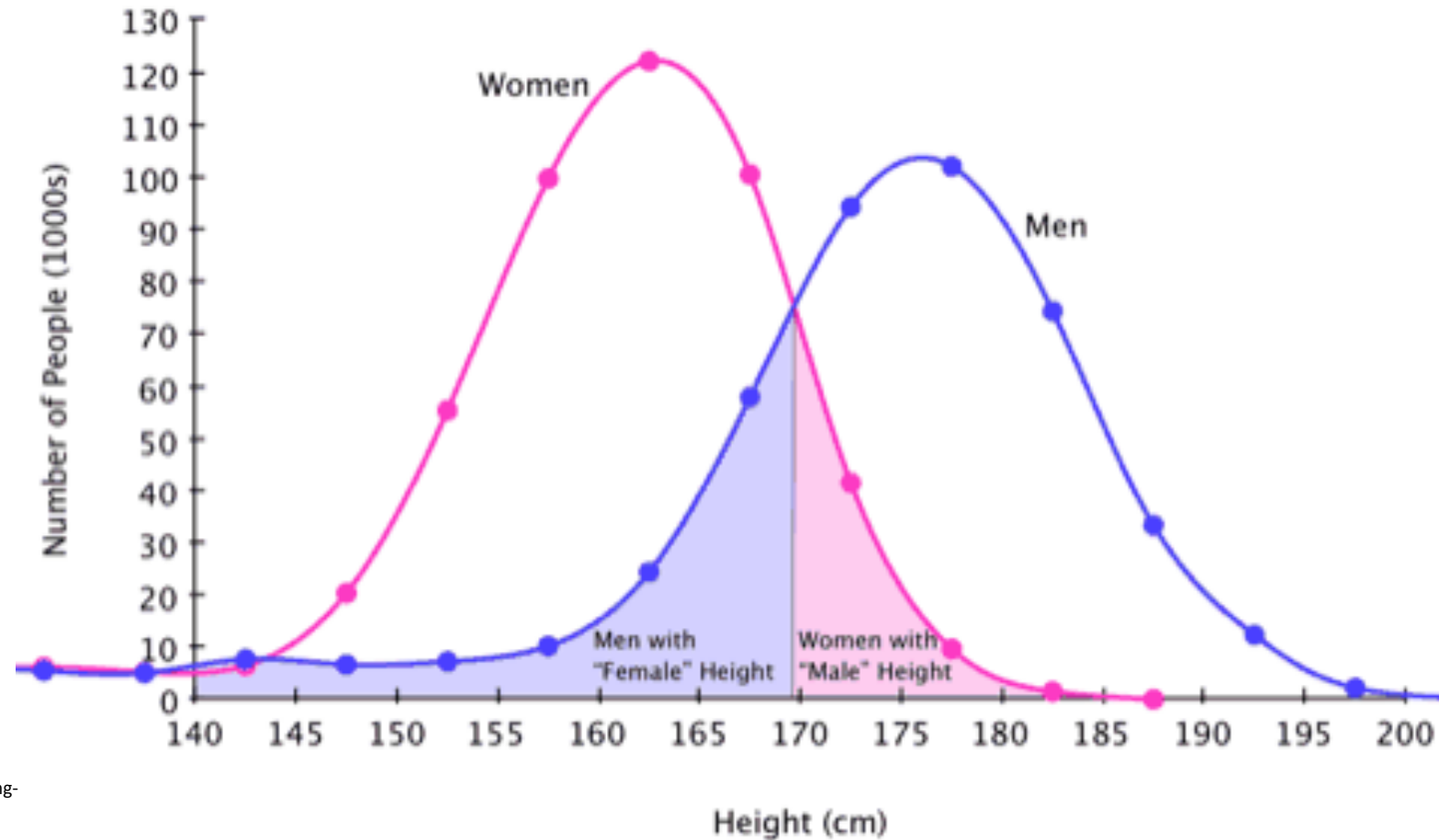


Photo Credit:
<https://www.quora.com/What-are-Overlapping-Bell-Curves-and-how-do-they-affect-Quora-questions-and-answers>

Introduction to our data set

RNA-Seq of hNPC – Zika Virus

PLOS ONE

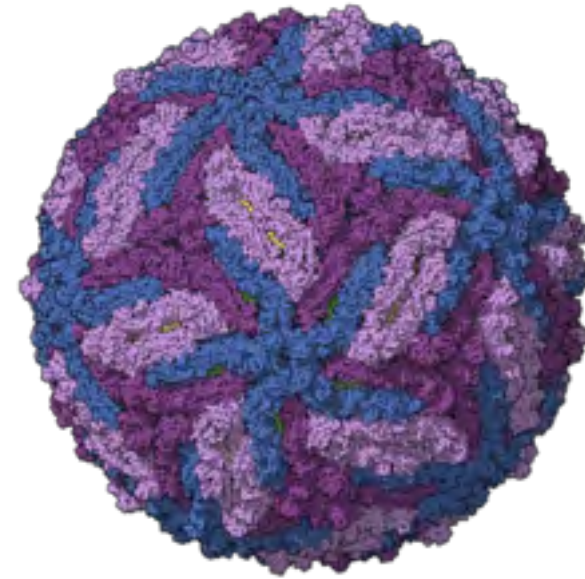
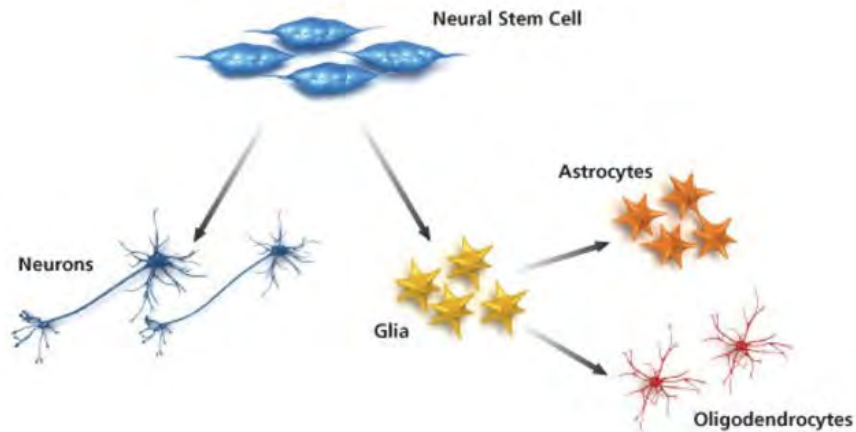
OPEN ACCESS PEER-REVIEWED

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



Zika Virus

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

Working on *DNA Subway* Green Line

FAST TRACK TO GENE ANNOTATION AND GENOME ANALYSIS

Username:
Password:

[Forgot Password?](#) [Register](#)

D **N** **A**

SUBWAY

Annotate a Genomic Sequence (Red line): Find Repeats, Predict Genes, Search Databases, Build Models

Prospect Genomes Using TARGeT (Yellow line): Search Genomes, Alignment & Tree Viewer

Determine Sequence Relationships (Blue line): Assemble Sequences, Add Sequences, Analyze Sequences

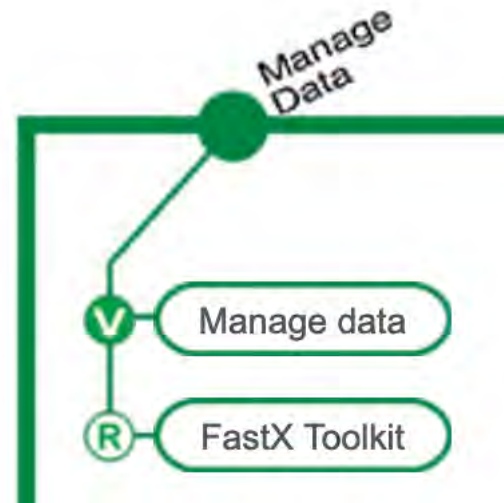
Next Generation Sequencing (Green line): Manage Data, Analyze Transcriptome, Explore Differential Abundance

Metabarcoding Analysis (Purple line): Metadata + QC, Clustering Sequences, Alpha/Beta Diversity

Browsers & Transfer (Central station)

DNA Subway ties together key bioinformatics tools and databases to assemble gene models, investigate genomes, work with phylogenetic trees and analyze DNA barcodes. Roll over the "stations" on the subway map to find out more about the analysis steps. Analyze your own data or sample data provided. To start a project, select one of the "lines" (red, yellow, blue, green, purple). Register and login to be able to save and share your results.

Working on *DNA Subway* Green Line



Key Concept: Sequence Quality

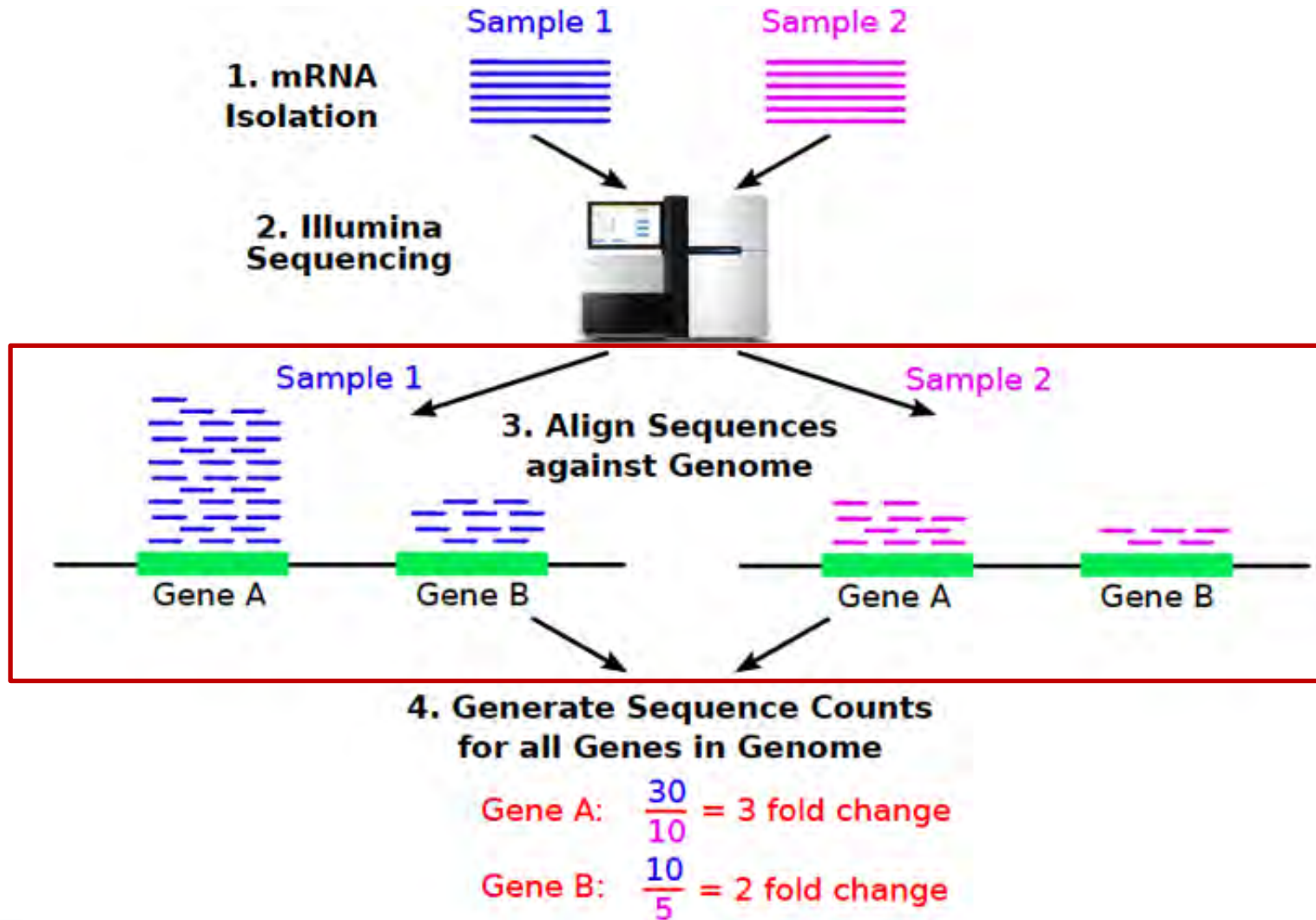
Phred scores...

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%

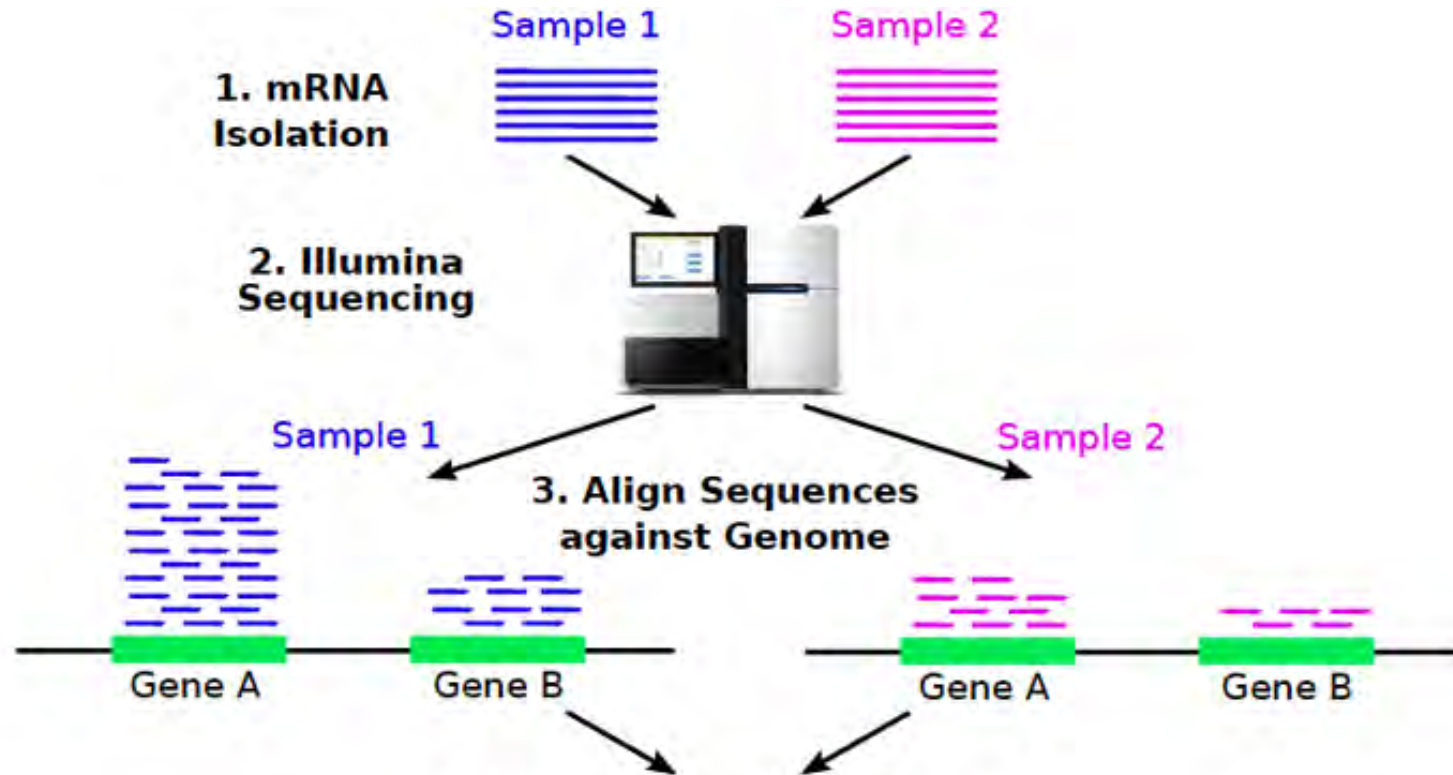
Key Concept: Read Alignment

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

Counting reads



Counting reads



4. Generate Sequence Counts for all Genes in Genome

$$\text{Gene A: } \frac{30}{10} = 3 \text{ fold change}$$

$$\text{Gene B: } \frac{10}{5} = 2 \text{ fold change}$$

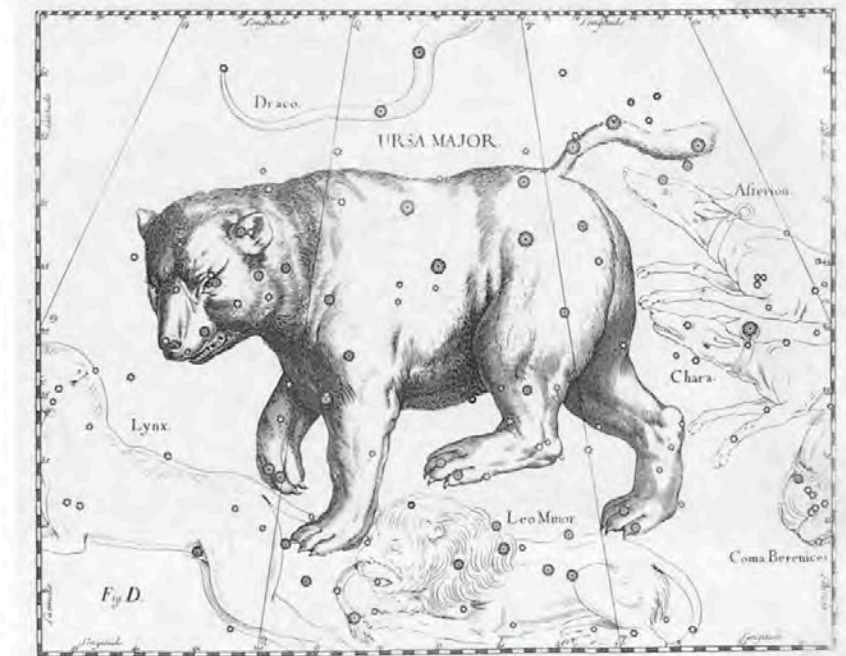
RNA-Seq with Kallisto

nature
biotechnology

NATURE BIOTECHNOLOGY VOLUME 34 NUMBER 5 MAY 2016

Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray¹, Harold Pimentel², Páll Melsted³
& Lior Pachter^{2,4,5}



[Download & Install](#)

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

Kallisto – Pseudoalignment

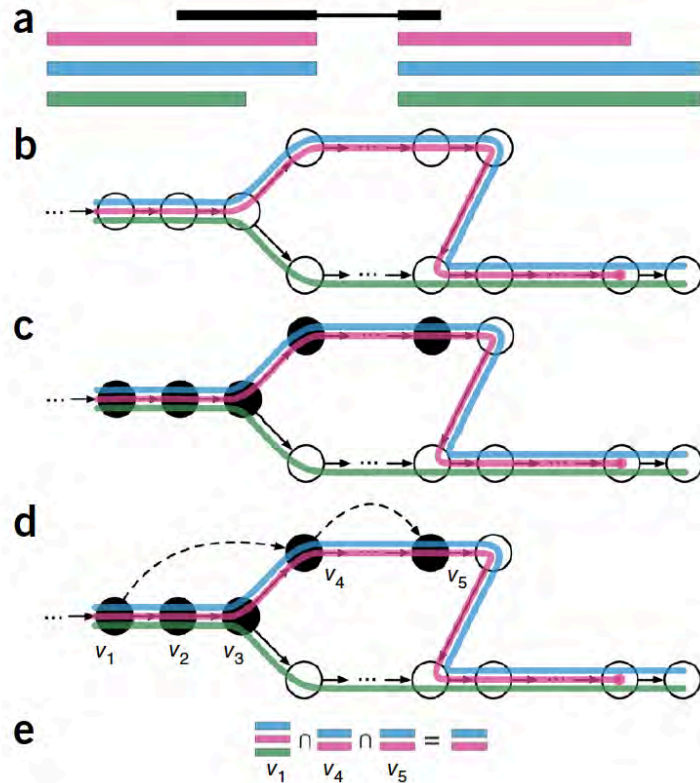
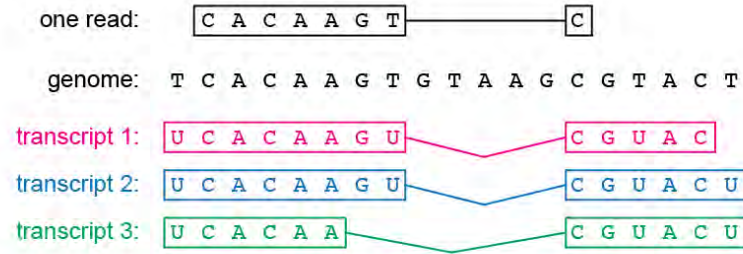


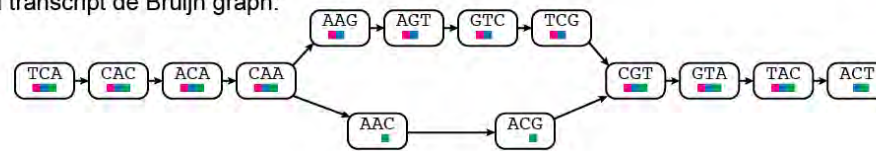
Figure 1 Overview of kallisto. The input consists of a reference transcriptome and reads from an RNA-seq experiment. **(a)** An example of a read (in black) and three overlapping transcripts with exonic regions as shown. **(b)** An index is constructed by creating the transcriptome de Bruijn Graph (T-DBG) where nodes (v_1, v_2, v_3, \dots) 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:

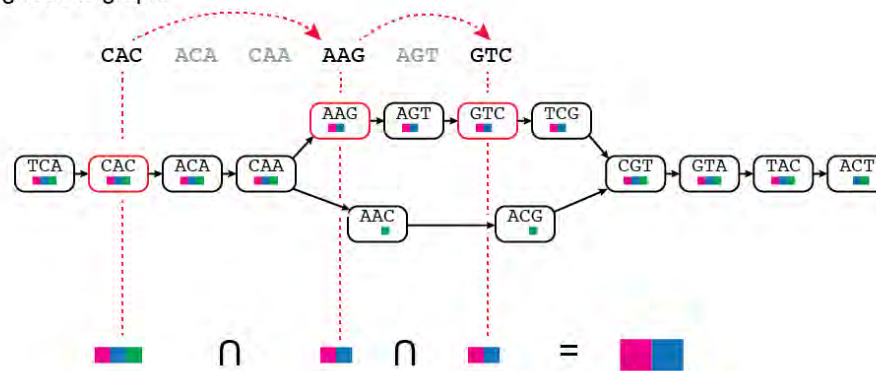


Photo credit:
<http://mcb112.org/w02/w02-lecture.html>

Lab: Pseudoalignment with Kallisto

Lab: Kallisto

Kallisto

New job

Inputs

Left pair	Right pair	Sample name *	Condition *
SRR3191542_1-tophat.fastq.gz	SRR3191542_2-tophat.fastq.gz		
SRR3191543_1-tophat.fastq.gz	SRR3191543_2-tophat.fastq.gz		
SRR3191544_1-tophat.fastq.gz	SRR3191544_2-tophat.fastq.gz		
SRR3191545_1-tophat.fastq.gz	SRR3191545_2-tophat.fastq.gz		

Parameters [show]

[Reset](#) [Submit](#)

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

e!Ensembl east [BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#)

Human (GRCh38.p13) ▼

Search Human (*Homo sapiens*)

Search all categories ▼ Search Human...

e.g. [BRCA2](#) or [17:63992802-64038237](#) or [rs699](#) or [osteoarthritis](#)

Genome assembly: GRCh38.p13 (GCA_000001405.28)

- [More information and statistics](#)
- [Download DNA sequence \(FASTA\)](#)
- [Convert your data to GRCh38 coordinates](#)
- [Display your data in Ensembl](#)

Other assemblies

[View karyotype](#)

[Example region](#)

Gene annotation

What can I find? Prot...
cDNA and protein seq...

- [More about this e...](#)
- [Download FASTA...](#)
- [Download GTF c...](#)
- [Update your old...](#)

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

[Example gene tree](#)

[More about comparative analysis](#)

Variation

What can I find? Sho...
disease and other phe...

[More about varia...](#)

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

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

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

Normalization – gene length

Which is longer (bp)?



Gene A



Gene B

Normalization – gene length

Which has more reads?



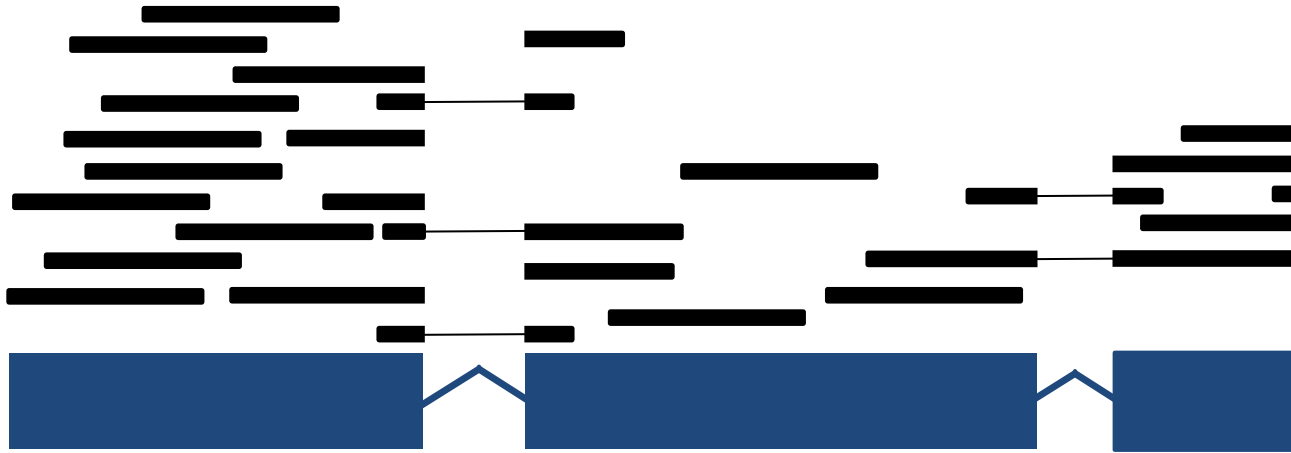
Gene A
(300bp)



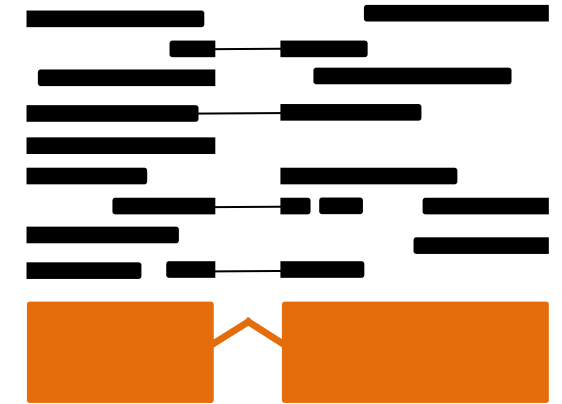
Gene B
(100bp)

Normalization – gene length

Which has more reads?

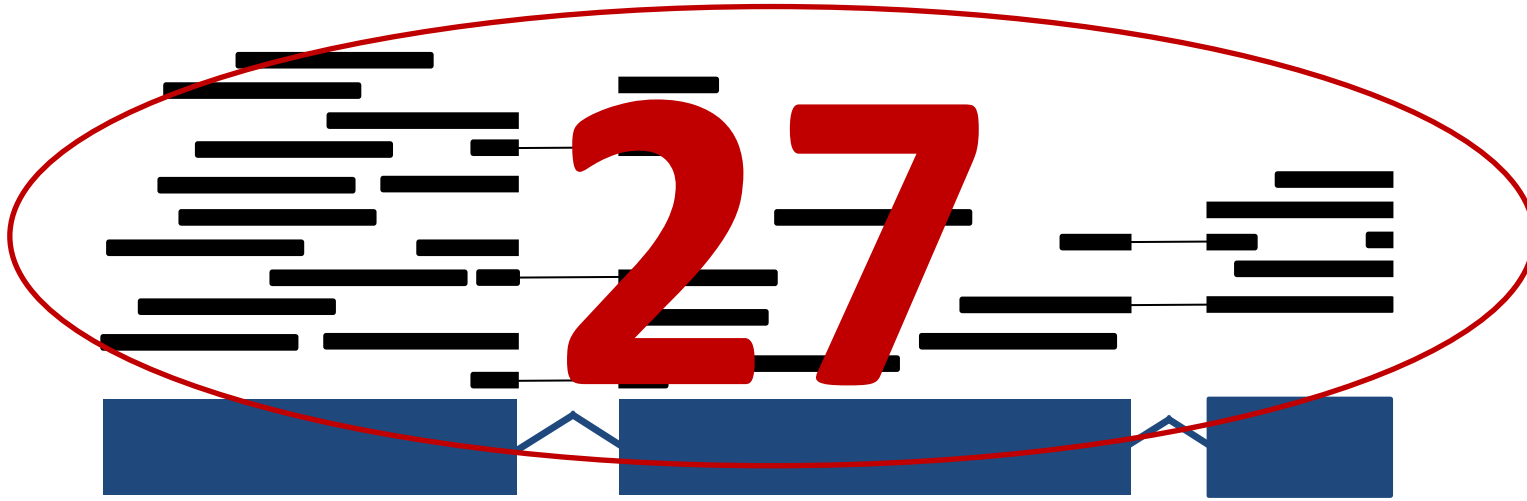


**Gene A
(300bp)**



**Gene B
(100bp)**

Normalization – gene length

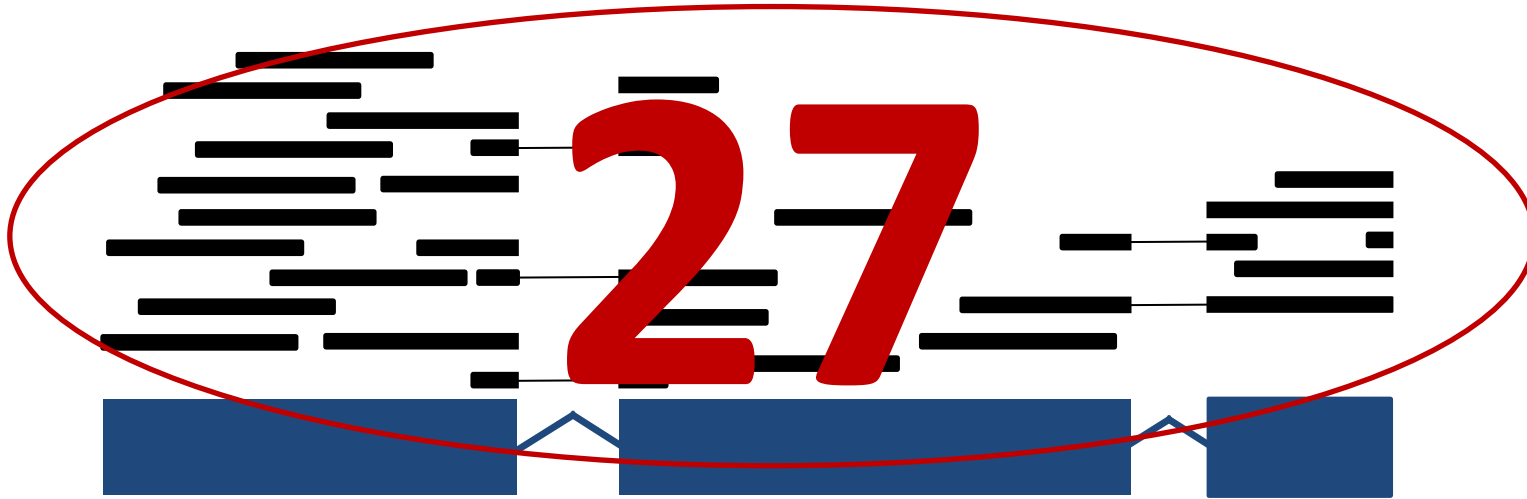


Gene A
(300bp)



Gene B
(100bp)

Normalization – gene length



Gene A
(300bp)

$$27/300 = 0.09$$

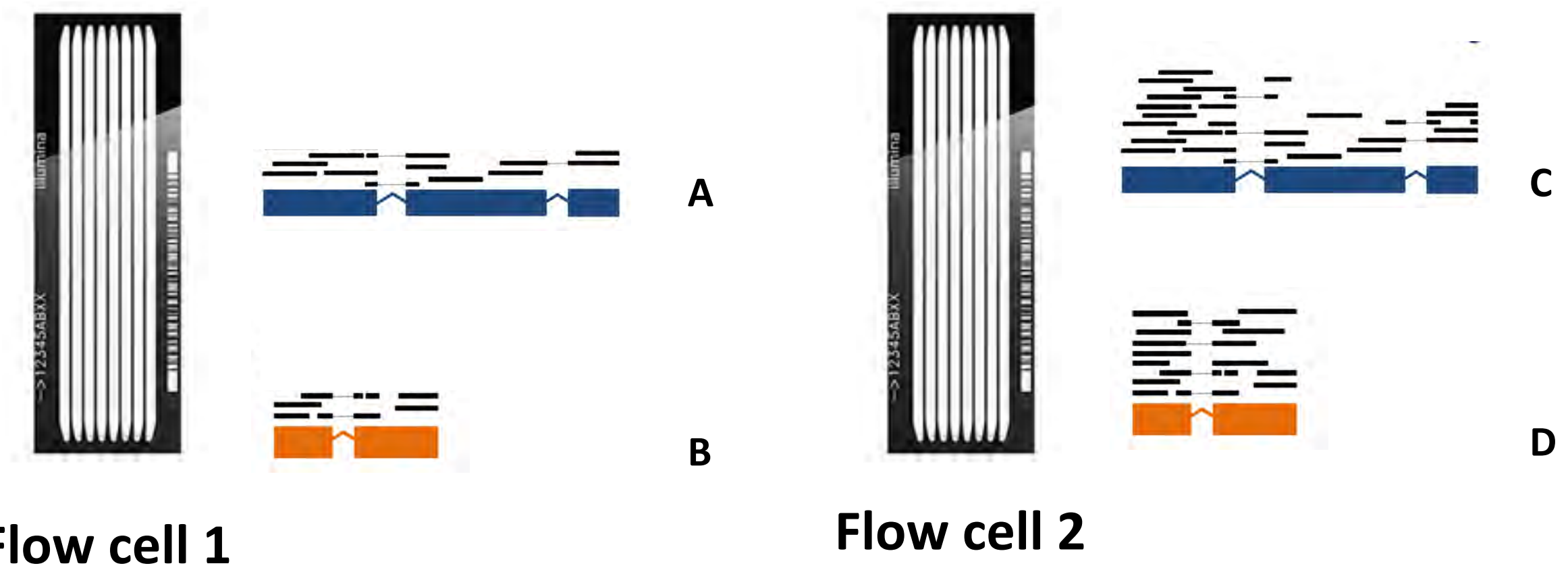


Gene B
(100bp)

$$16/100 = 0.16$$

Normalization – read depth

Which gene is most highly expressed?



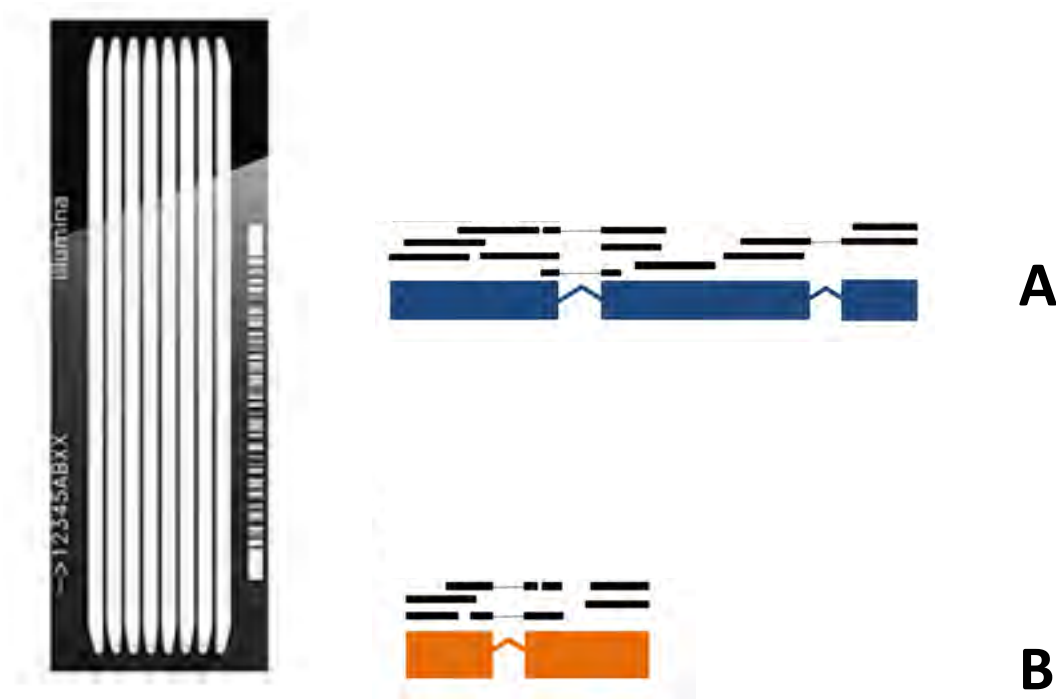
Flow cell 1

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

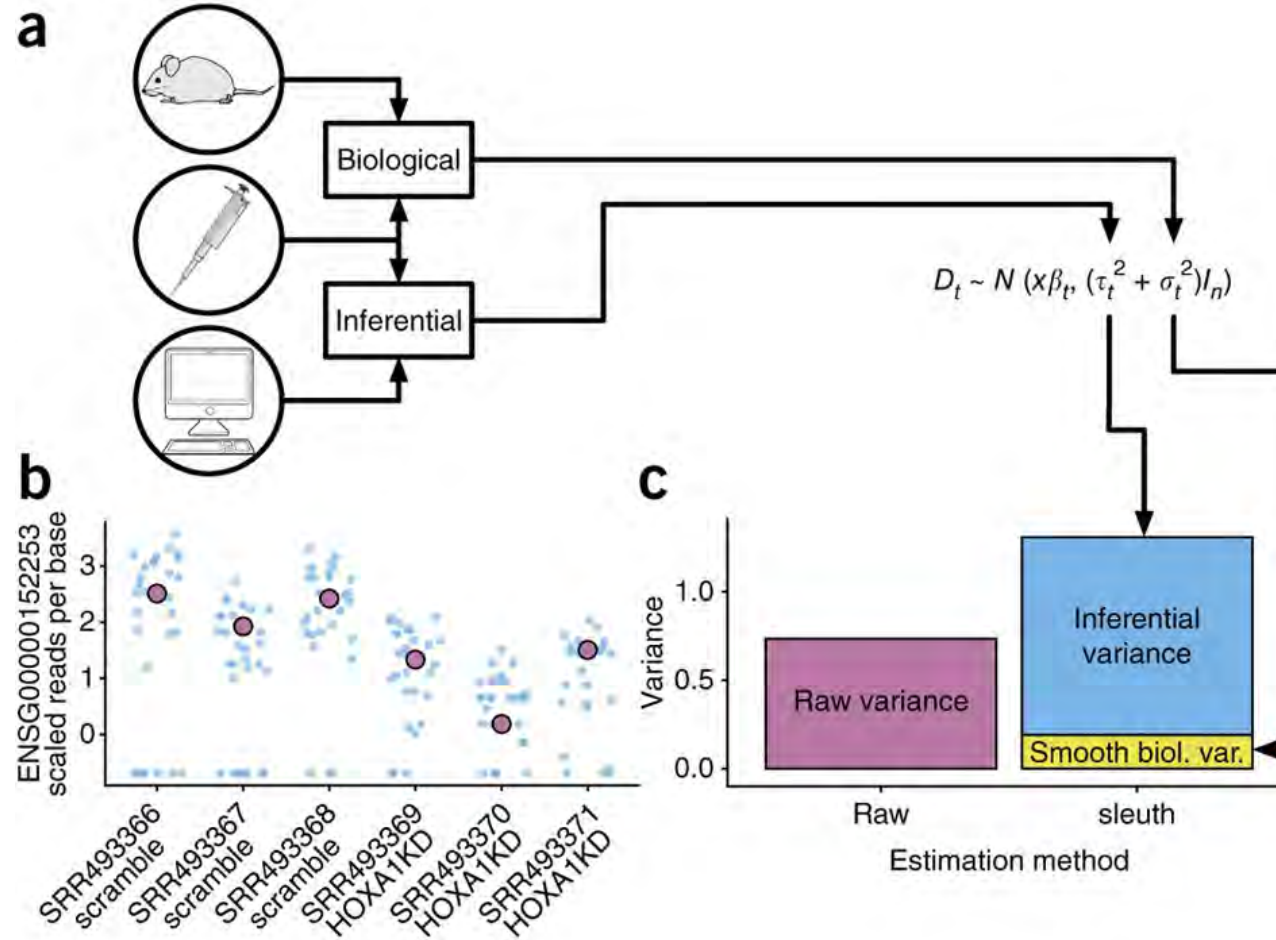


Photo credit

<https://www.nature.com/articles/nmeth.4324/figures/1>

Sleuth linear modeling

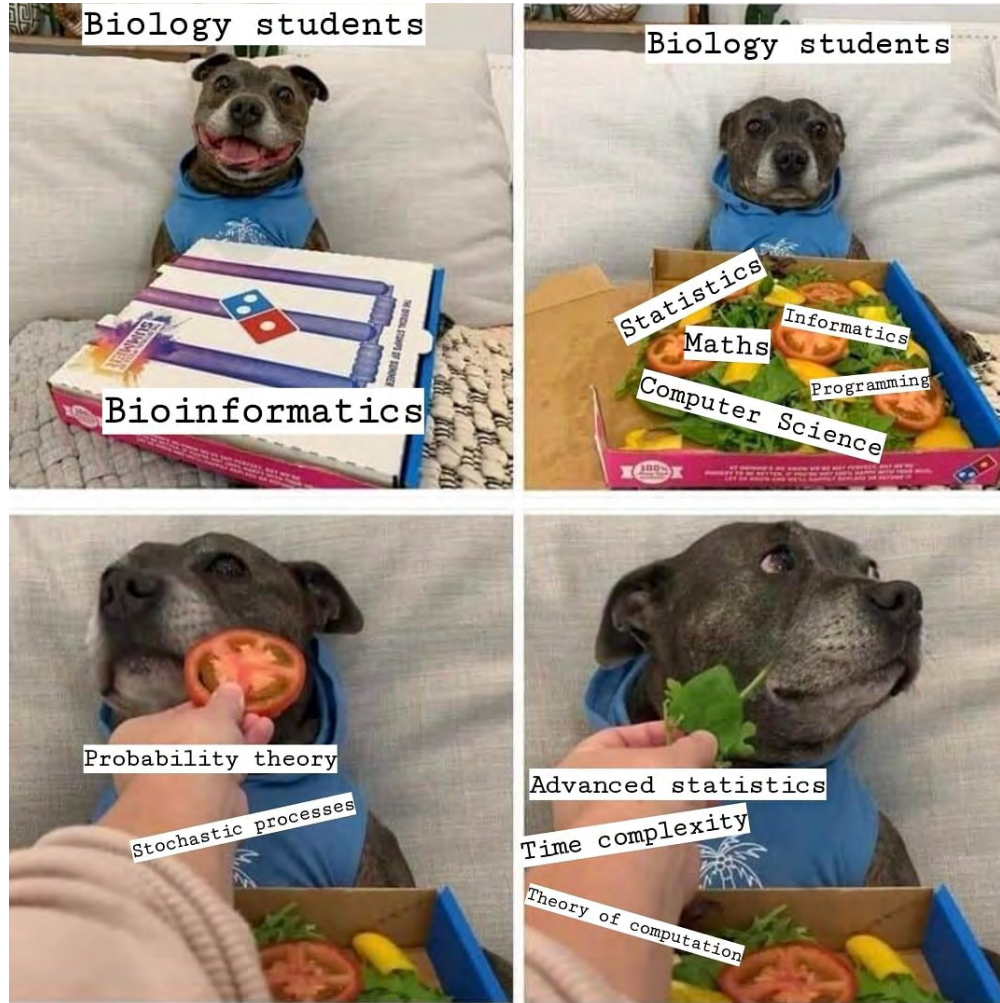


Photo credit
<https://twitter.com/phyloge0>

Lab: Sleuth results

Green Line

DNA SUBWAY

Differential analysis Results Bootstrap PCA Volcano Loadings Sample Heatmap

Differential analysis results

Table columns to display

- target_id: transcript name
- pval: p-value of the chosen model
- qval: false discovery rate adjusted p-value
- b: beta value (effect size). Technically a biased estimator of the fold change
- se_b: standard error of the beta
- mean_obs: mean of natural log counts of observations
- var_obs: variance of observation
- tech_var: technical variance of observation from the bootstraps
- sigma_sq: raw estimator of the variance once the technical variance has been removed
- smooth_sigma_sq: smooth regression fit for the shrinkage estimation
- final_sigma_sq: max(sigma_sq, smooth_sigma_sq); used for covariance estimation of beta
- ens_gene: gene name in ensembl
- ext_gene: external gene name

Show 10 entries

Search:

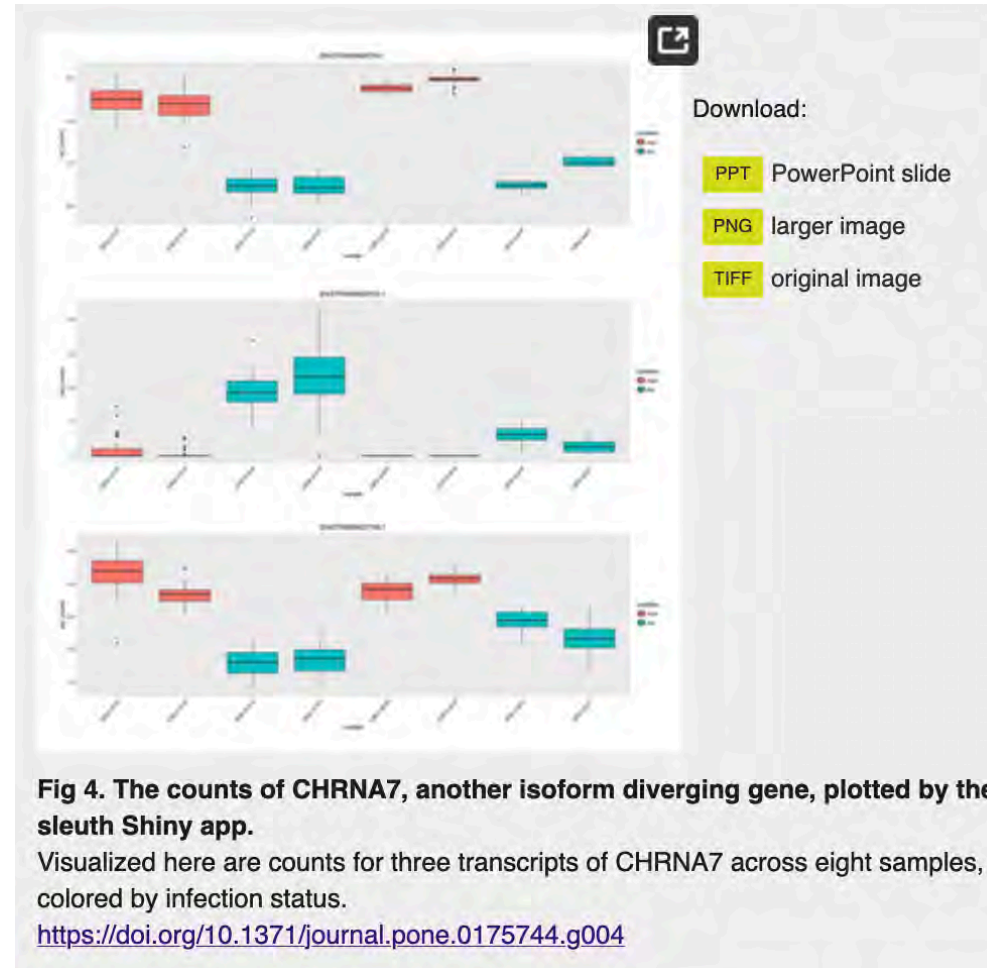
	target_id	qval	b	ext_gene
	All	All	All	All
1	ENST00000617285.4	4.6257e-85	0.90908	HYOU1
2	ENST00000630669.2	4.6257e-85	0.90908	HYOU1
3	ENST00000307365.3	3.941e-82	1.3256	DDIT4
4	ENST00000439211.6	2.3171e-78	-1.7232	DHFR
5	ENST00000280612.9	3.8927e-77	1.7133	SLC7A11
6	ENST00000338663.11	1.9313e-70	1.6309	SLC3A2
7	ENST00000253063.3	7.9561e-66	1.347	SESN2
8	ENST00000361427.5	2.5113e-64	-0.83316	HMG2
9	ENST00000274063.4	5.8317e-64	-1.1304	SFRP2
10	ENST00000332707.9	3.2939e-58	0.94766	XPOT

Showing 1 to 10 of 55,351 entries

Previous 1 2 3 4 5 ... 5536 Next

Download

Comparing to the results of the L. Yi paper



Ontology enrichment: ShinyGO

ShinyGO v0.61: Gene Ontology Enrichment Analysis + more

Reset

1. Select or search for your species.
Best matching species

2. Paste genes Demo genes

Just paste gene lists and click Submit. Most types of gene IDs accepted. Double check the guessed species, and adjust if needed.

3. Submit

P-value cutoff (FDR)
0.05

of most significant terms to show
30

Enrichment Tree Network Genes Groups Plots Genome Promoter STRING ?

2/3/2020: Now published by [Bioinformatics](#).

11/3/2019: V 0.61, Improve graphical visualization (thanks to reviewers). Interactive networks and much more.

5/20/2019: V.0.60, Annotation database updated to Ensembl 96. New bacterial and fungal genomes based on STRING-db!

Just paste your gene list to get enriched GO terms and othe pathways for over 315 plant and animal species, based on annotation from Ensembl (Release 96), Ensembl plants (R. 43) and Ensembl Metazoa (R. 43). An additional 2031 genomes (including bacteria and fungi) are annotated based on STRING-db (v.10). In addition, it also produces KEGG pathway diagrams with your genes highlighted, hierarchical clustering trees and networks summarizing overlapping terms/pathways, protein-protein interaction networks, gene characteristics plots, and enriched promoter motifs. See example outputs below:

Enrichment FDR	Genes in list	Total genes	Functional Category
6.5E-220	86	101	DNA damage checkpoint
5.3E-215	86	108	DNA integrity checkpoint
3.2E-188	86	169	Cell cycle checkpoint
1.1E-131	87	659	Cellular response to DNA damage stimulus
1.2E-113	87	1039	Cell cycle process
3.6E-102	87	1395	Cell cycle
1.1E-100	45	57	Mitotic DNA integrity checkpoint
5.0E-99	87	1517	Cellular response to stress
1.9E-98	43	51	Mitotic DNA damage checkpoint

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

Gene ontology

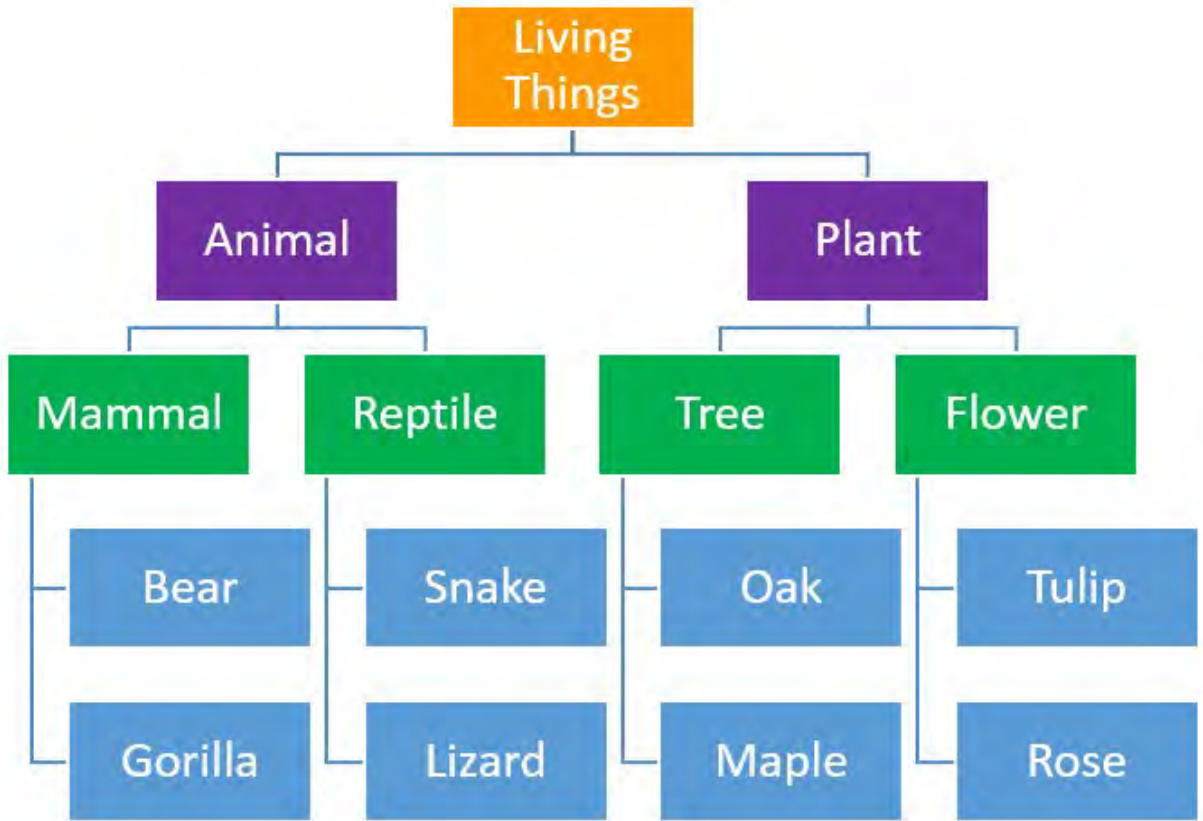


Photo credit:
<https://thepeakperformancecenter.com/educational-learning/learning/memory/stages-of-memory/organization-long-term-memory/>

Gene ontology

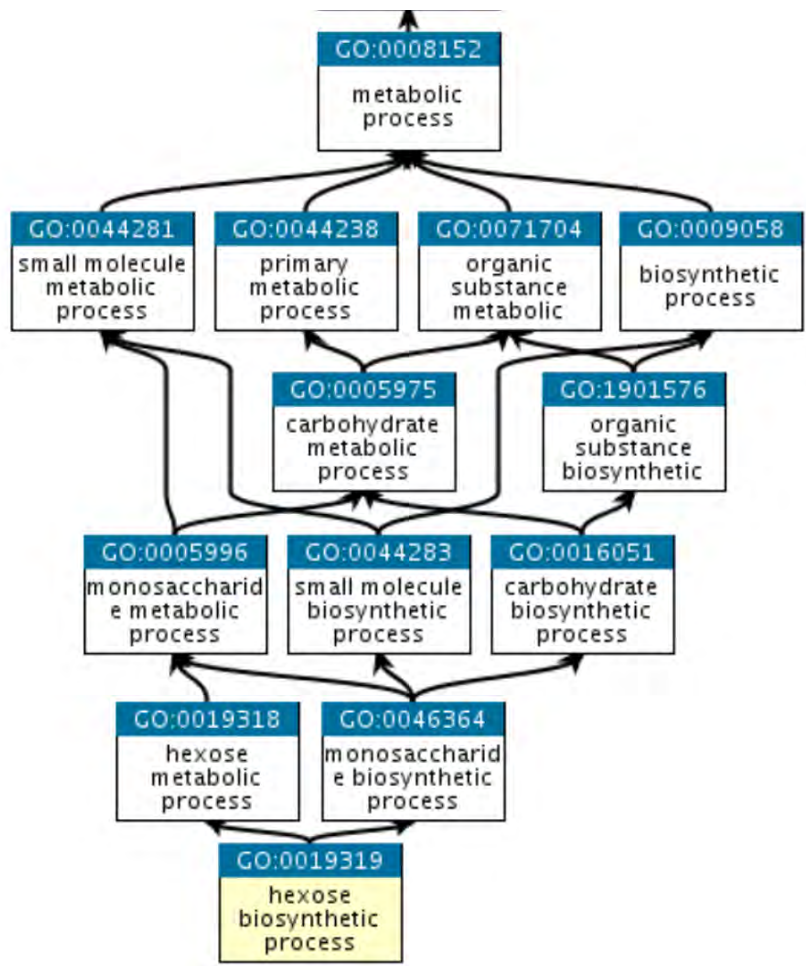


Photo credit:
<http://geneontology.org/docs/ontology-documentation/>

Tang paper ontologies

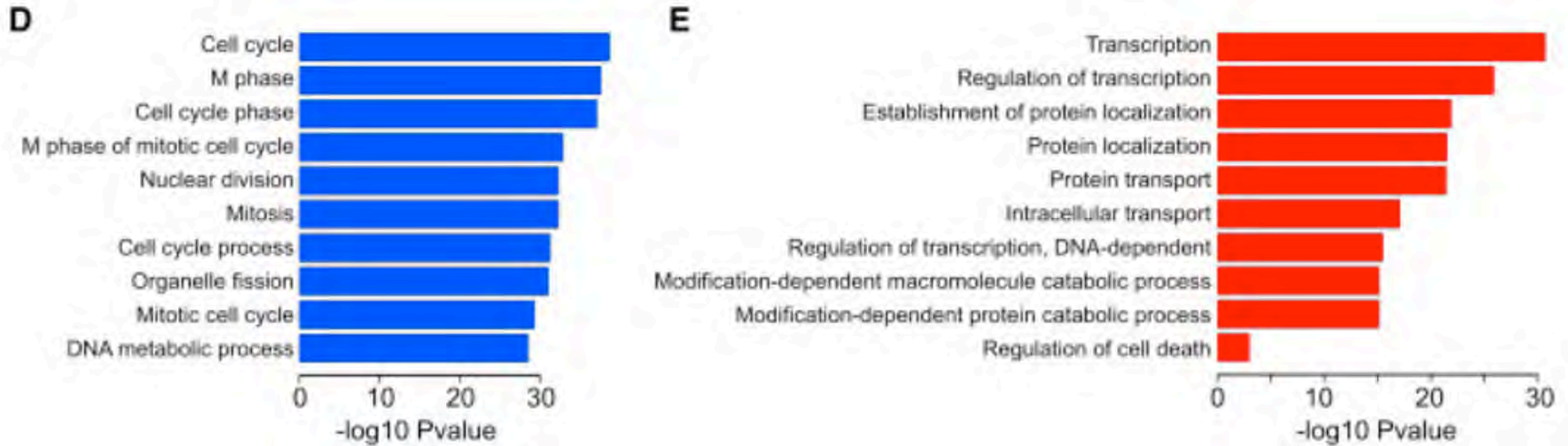


Figure 2 ZIKV-Infected hNPCs Exhibit Increased Cell Death and Dysregulated Cell-Cycle Progression and Gene Expression

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