Cold Spring Harbor Laboratory
DNA Learning Center

DNALC Live

Intro to RNA-Seq with Jupyter Part II

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

GEA
Genomics Education Alliance

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: 2 sessions (1 per week); ~ 45 minutes each
- Exercises: Follow along through CyVerse
- Learning resources: Slides and online lesson available

Course Learning Goals

- Understand the rationale of an RNA-Seq experiment and its design
- Learn about the Linux command line
- Use *Jupyter* (*SRA Toolkit*) to import sequence data
- Use *Jupyter* (*FastQC/Trimmomatic*) to quality check/trim sequence data
- Use *Jupyter* (*Kallisto*) to (pseudo)align reads
- Use *Jupyter* (*genomeview/UCSC*) to explore RNA-Seq results

Lab Setup

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

Intro to RNA-Seq with Jupyter Part II

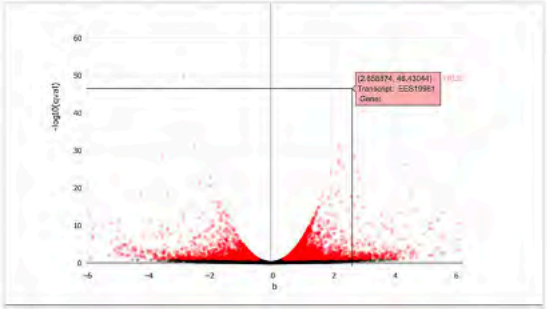
(alignment and visualization)

Steps for today's session

- Review data set and data cleaning
- Learn about alignment of reads
- See how to visualize data in a genome browser

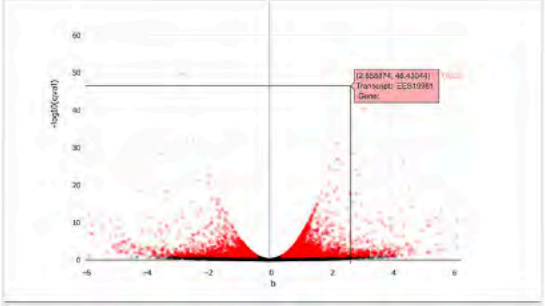
RNA-Seq with *DNA Subway*

dnalc.cshl.edu/dnalc-live



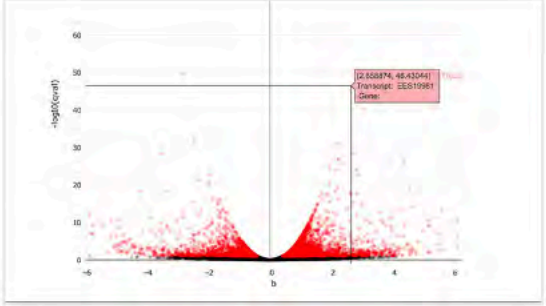
RNA-Seq with DNA Subway, Part I

Undergraduate biology and up
Duration 0:59:27



RNA-Seq with DNA Subway, Part II

Undergraduate biology and up
Duration 0:58:28



RNA-Seq with DNA Subway, Part III

Undergraduate biology and up
Duration 0:50:46

Introduction to RNA-Seq

What is RNA-Seq? - measuring the transcriptome

- To understand what genes are active, and under what circumstances, we must know what genes are being transcribed into messenger RNA

What is RNA-Seq? - measuring the transcriptome

- To understand what genes are active, and under what circumstances, we must know what genes are being transcribed into messenger RNA

- A cell in the liver has the same DNA instructions as a neuron in the brain. However the genes being expressed differ greatly between these cells

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

What is RNA-Seq? - measuring the transcriptome

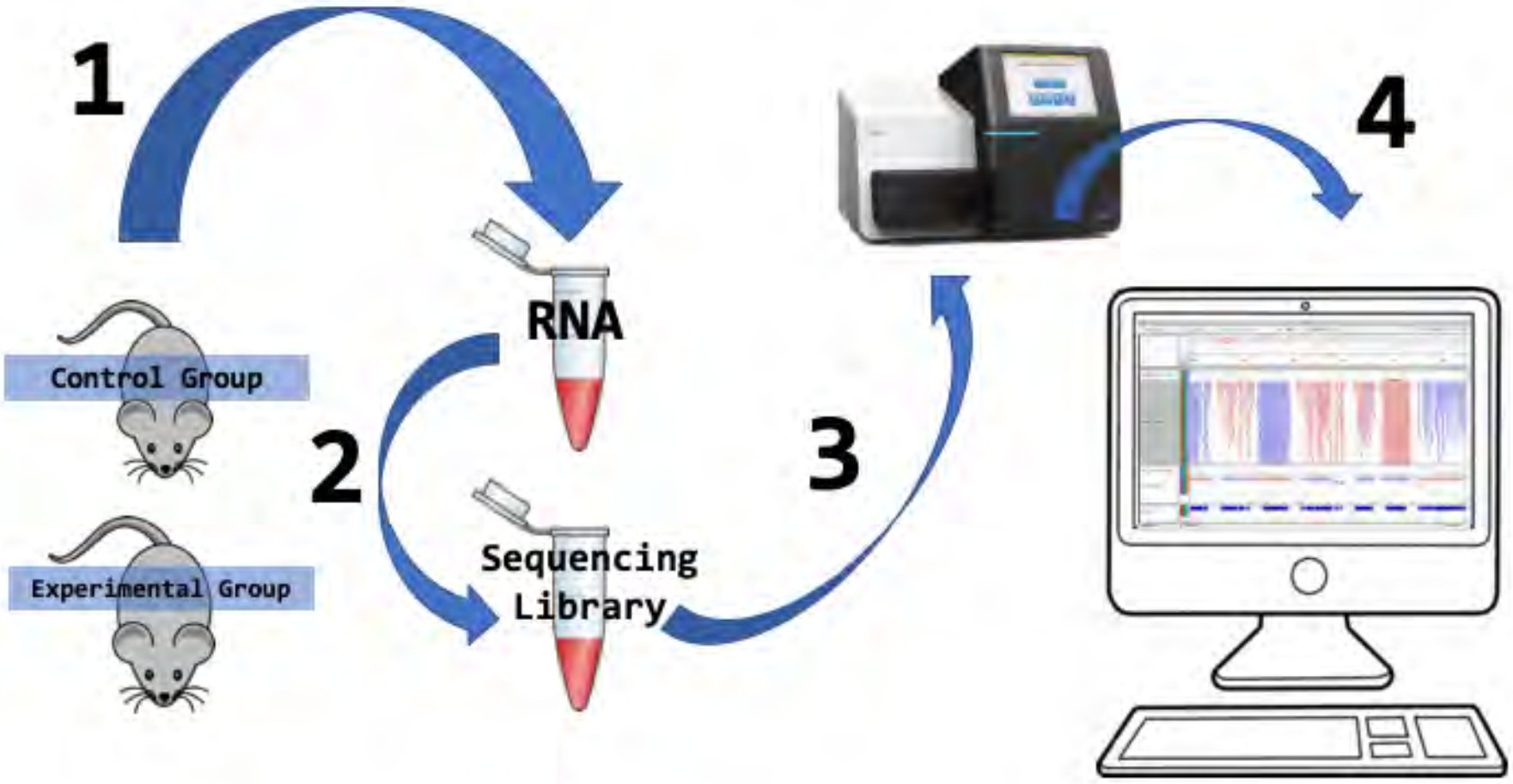
- RNA-Seq allows us to measure the transcriptome – take an account of all transcription occurring in a cell/tissue

- We use the abundance of an RNA transcript as a proxy for the activity of some cellular process (e.g. protein synthesis, regulatory activity)

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)

What is RNA-Seq?



Introduction to our data set

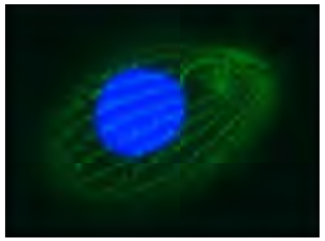
GEA

Genomics Education Alliance

An NSF Research Collaboration Network

Founding Members

Ciliate Genomics Consortium



GCAT-Seek

SEEKquence

Genomics Education Partnership

GEP

Genome Solver

GENOME SOLVER

Network for Integrating Bioinformatics in Life Sciences Education

NIBLSE

Bioinformatics in Life Sciences Education

CyVerse

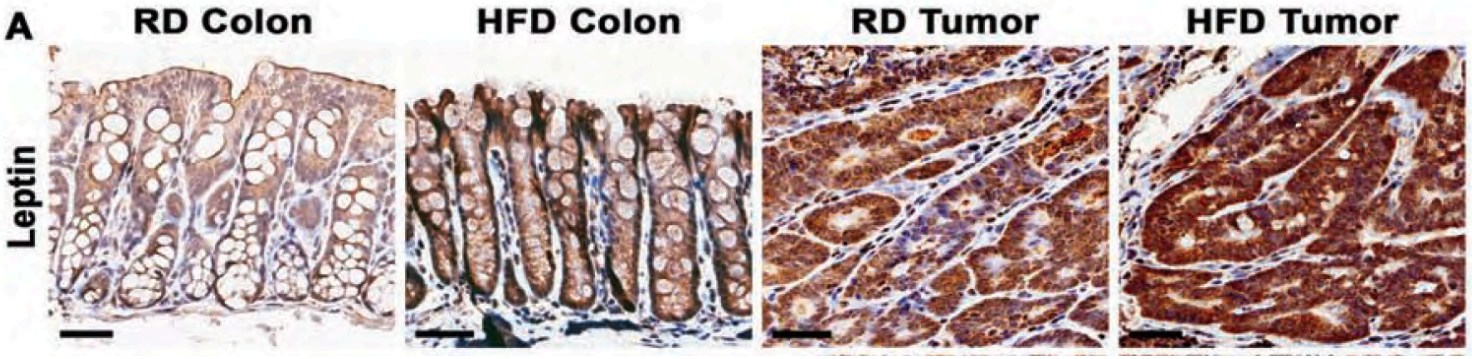
Leptin expression vs. diet – RNA-Seq pilot lesson

Carcinogenesis
Integrative Cancer Research

High-fat diet induced leptin and Wnt expression: RNA-sequencing and pathway analysis of mouse colonic tissue and tumors FREE

Harrison M. Penrose, Sandra Heller, Chloe Cable, Hani Nakhoul, Melody Baddoo, Erik Flemington, Susan E. Crawford, Suzana D. Savkovic
Author Notes

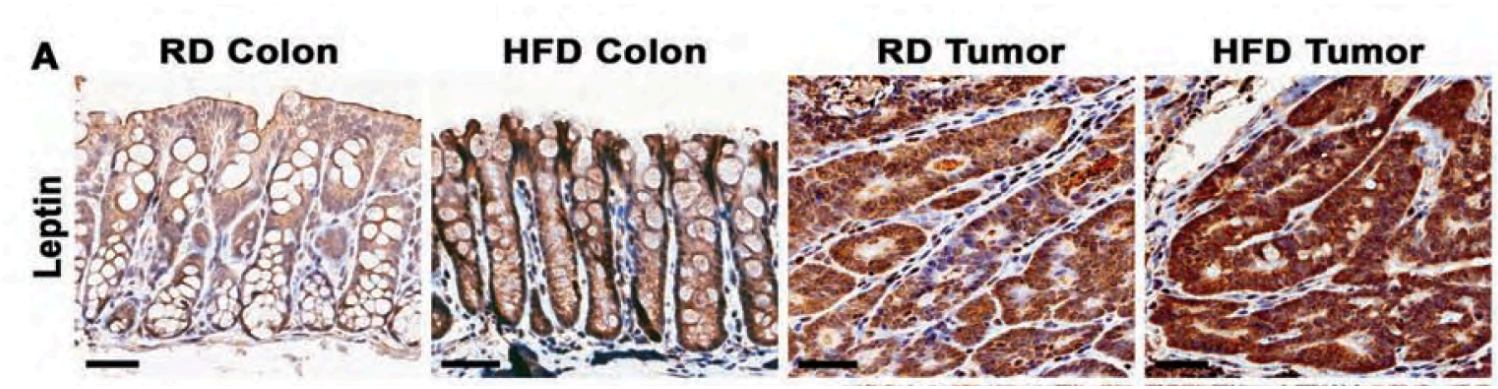
Carcinogenesis, Volume 38, Issue 3, 1 March 2017, Pages 302–311,
https://doi.org/10.1093/carcin/bgx001
Published: 25 January 2017 **Article history** ▾



Colon tissue/tumors in mice raise on Regular (RD) or High-fat (HFD) diet

Leptin expression vs. diet – RNA-Seq pilot lesson

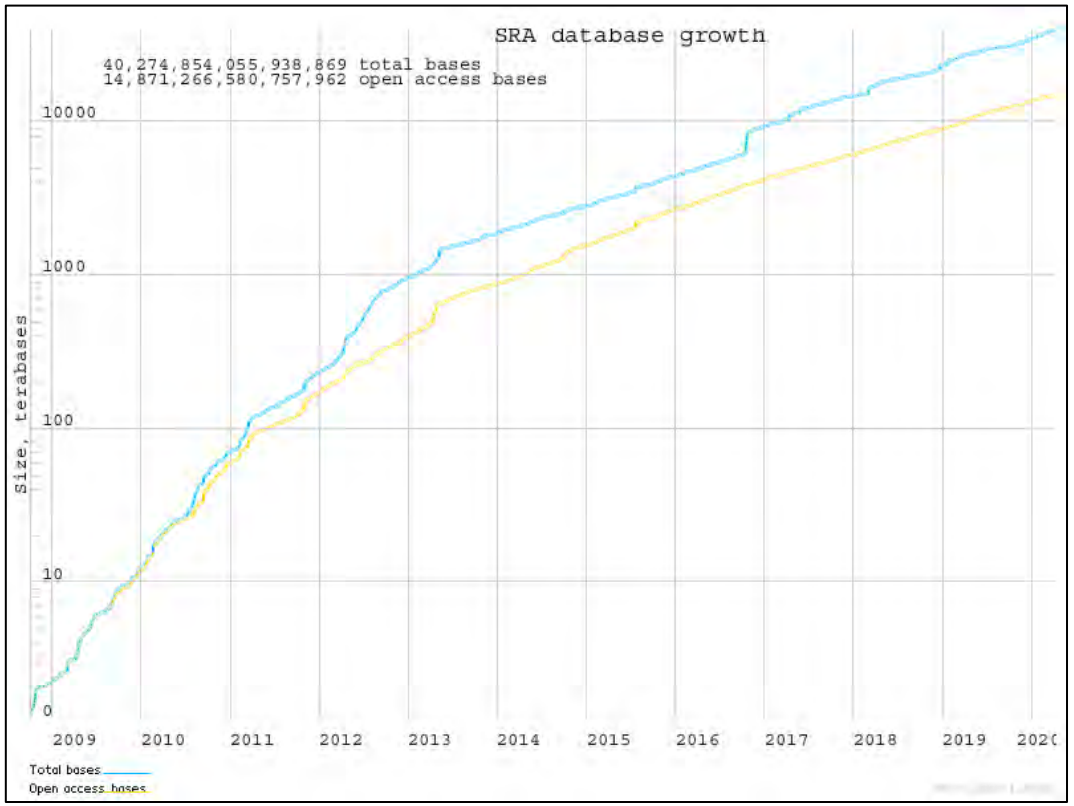
SRA_Sample	Sample_Name
SRS1794108	High-Fat Diet Control 1
SRS1794110	High-Fat Diet Control 2
SRS1794106	High-Fat Diet Control 3
SRS1794105	High-Fat Diet Tumor 1
SRS1794101	High-Fat Diet Tumor 2
SRS1794111	High-Fat Diet Tumor 3



Colon tissue/tumors in mice raise on Regular (RD) or High-fat (HFD) diet

Sequence data from NCBI

National Center for Biotechnology Information
NCBI



https://www.ncbi.nlm.nih.gov/bioproject/PRJNA353374

Access lessons and sign in on CyVerse

RNA-Seq analysis of Mouse Leptin Gene

GEA Genomics Education Alliance

latest

Search docs

- Lesson home
- Launch Lesson on CyVerse
- Jupyter Primer
- Command Line Primer
- Intro to RNA-Seq
- Getting Data from NCBI
- Assessing Data Quality
- Trimming and Filtering Data

Docs » Introduction to RNA-Seq: Leptin expression in mouse

Edit on GitHub

Introduction to RNA-Seq: Leptin expression in mouse

Submission Details

Submission Date	December, 2019
Version	1.0
Authors	• Jason Williams, Cold Spring Harbor Laboratory • Judy Brusslan, California State University Long Beach • Ray Enke, Jame Madison University • Matthew Escobar, California State University San Marcos • Vince Buonaccorsi, Juaniata College

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%

Lab – Sequence alignment

There are many roads to RNA-Seq

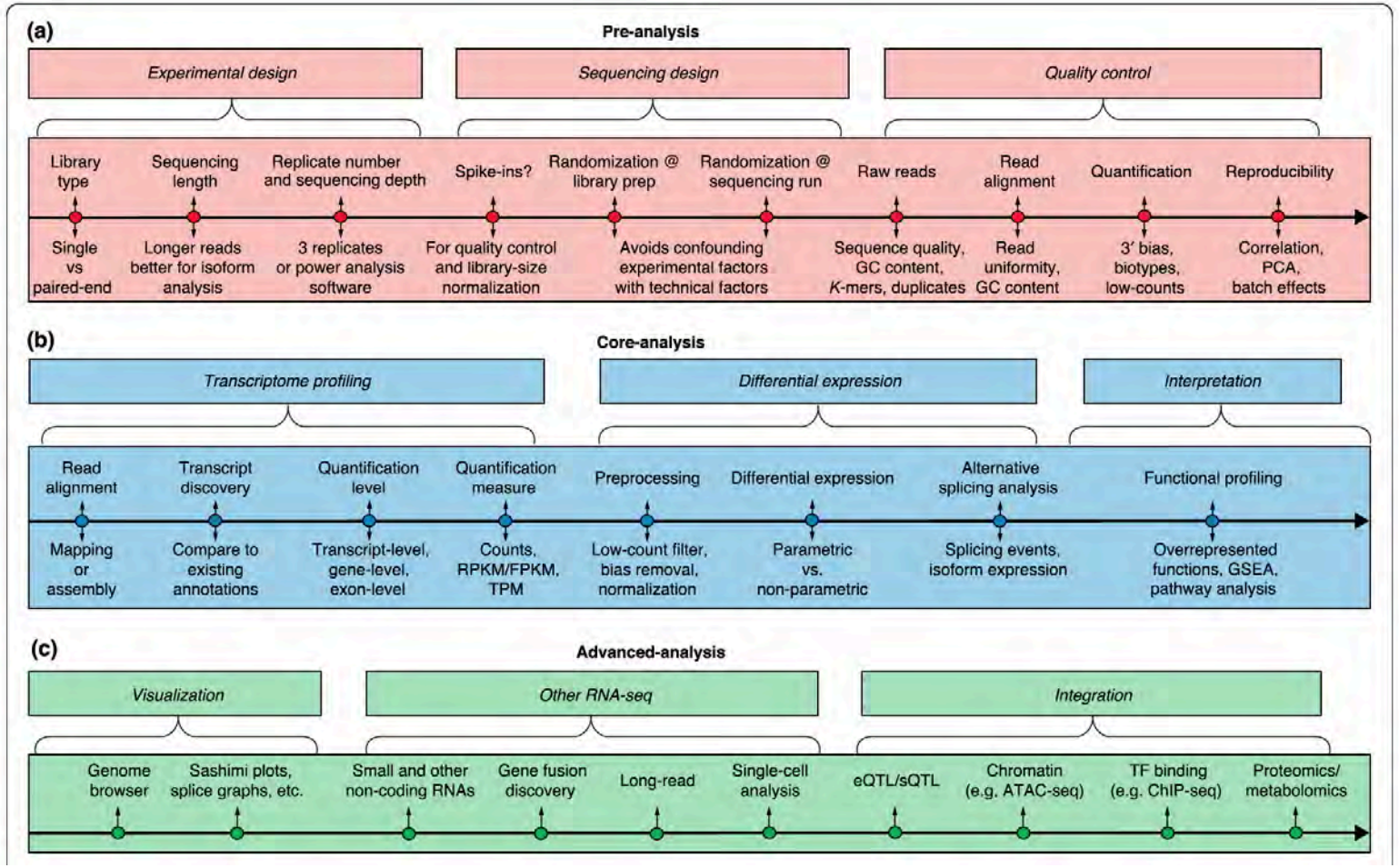


Photo credit:
Conesa et al. Genome Biology (2016) 17:13
DOI 10.1186/s13059-016-0881-8

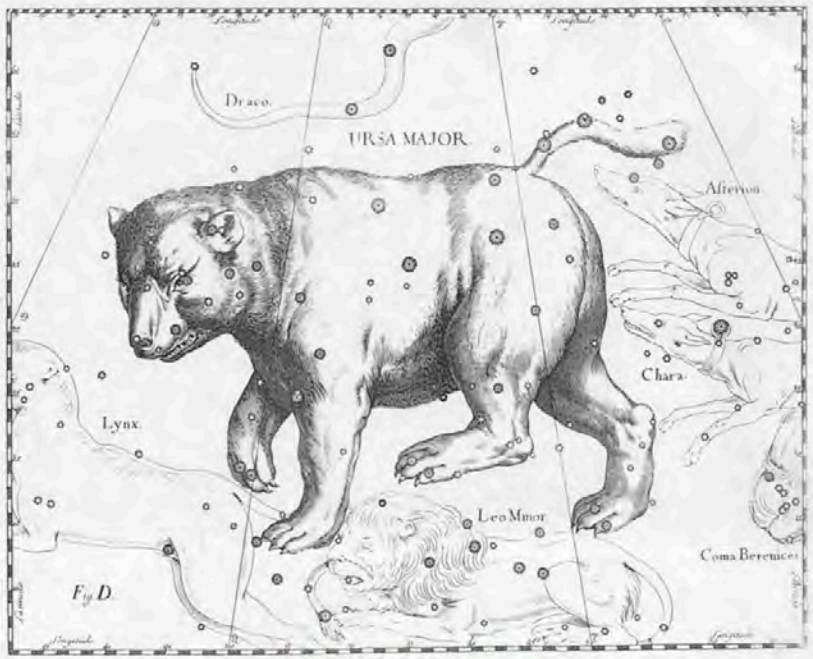
RNA-Seq with Kallisto

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]



Download & Install

RNA-Seq with Kallisto

Kallisto (pseudo)aligns reads to a reference transcriptome

1. An index is built of the reference transcriptome

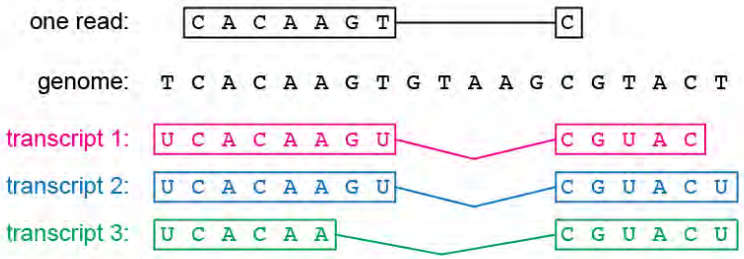
2. Sequence reads are (pseudo)aligned to transcripts

Reference transcriptome

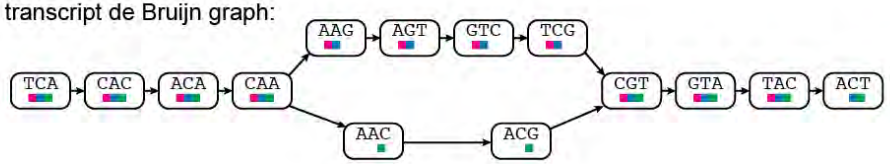
A collection of “all” the transcripts in an organism

Ensembl tour: https://useast.ensembl.org/Homo_sapiens/Info/Index

Kallisto – Pseudoalignment



Colored transcript de Bruijn graph:



Matching read to graph:

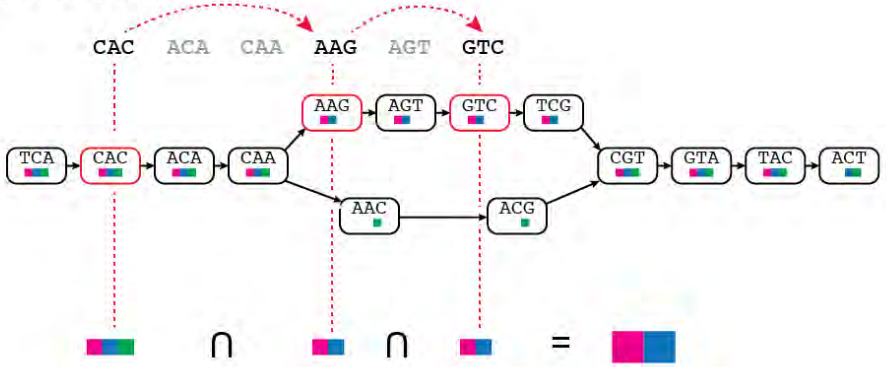


Photo credit:
http://mcb112.org/w02/w02-lecture.html

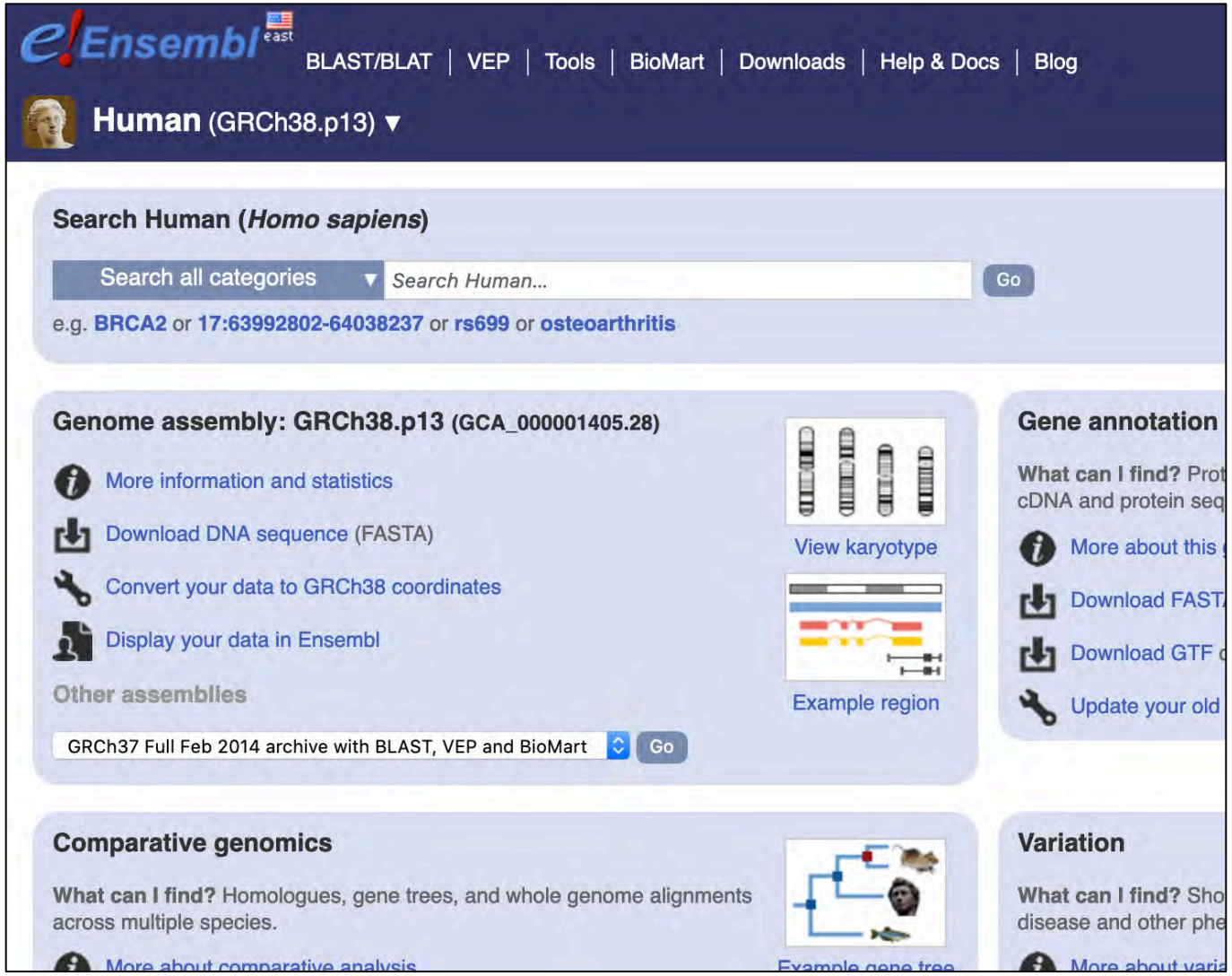
Kallisto results

	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

Normalization – gene length

Which is longer (bp)?

Gene A

Gene B

Normalization – gene length

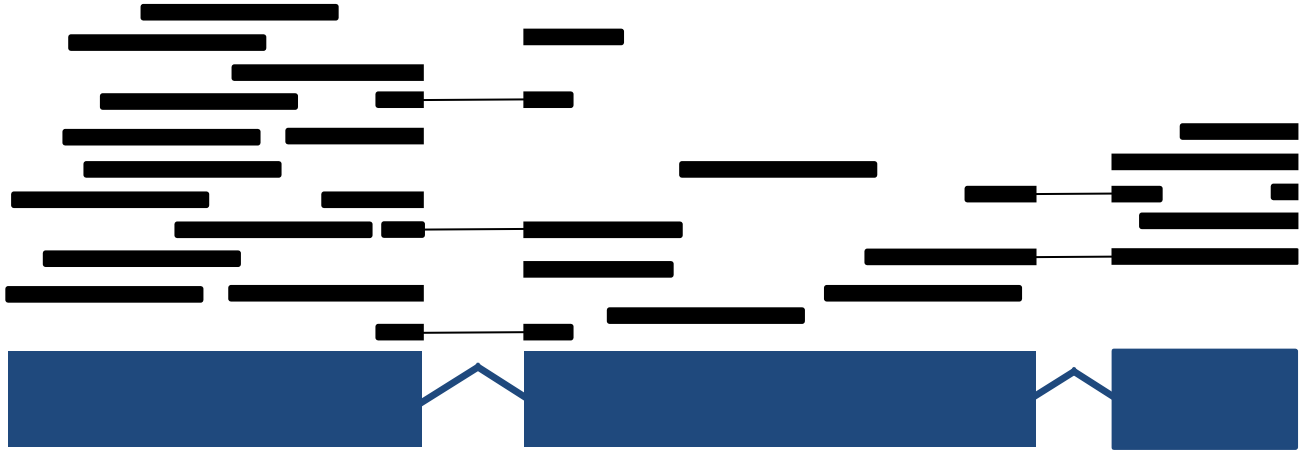
Which has more reads?

Gene A (300bp)

Gene B (100bp)

Normalization – gene length

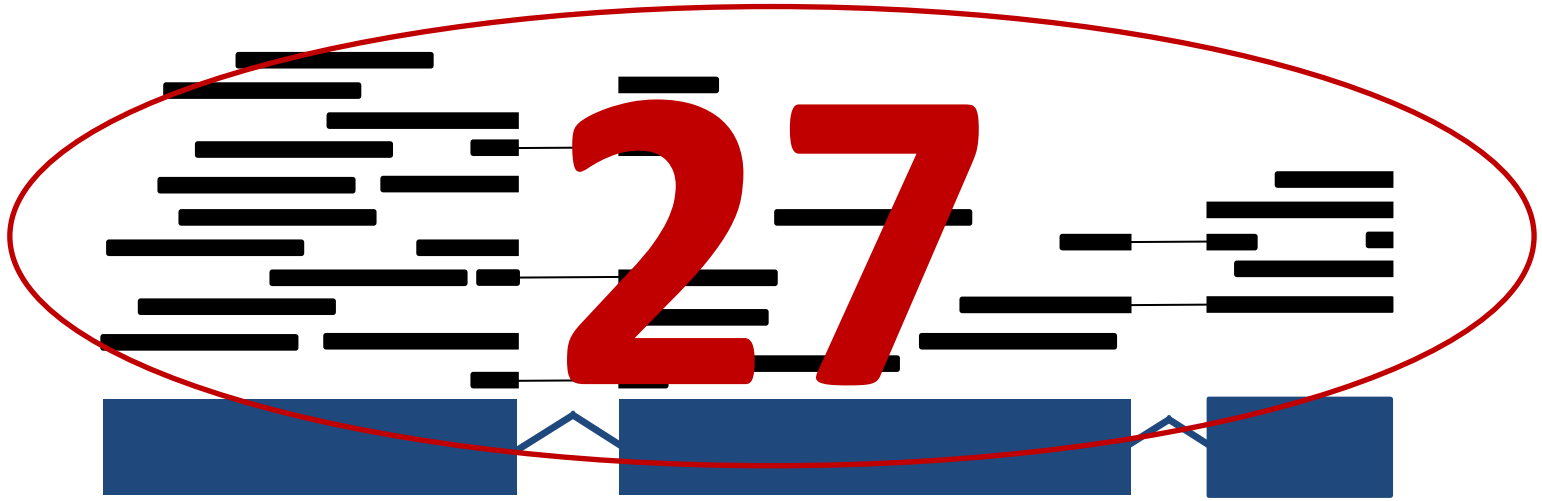
Which has more reads?



Gene A (300bp)

Gene B (100bp)

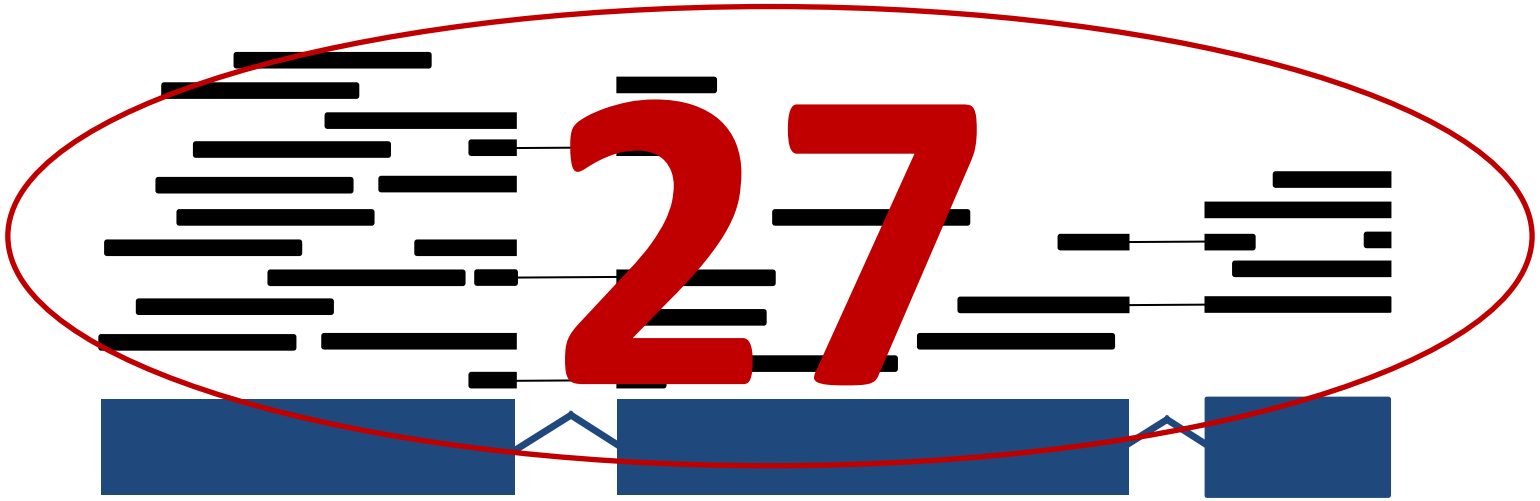
Normalization – gene length



Gene A (300bp)

Gene B (100bp)

Normalization – gene length



Gene A
(300bp)

27/300 = 0.09

Gene B
(100bp)

16/100 = 0.16

Goal recap

- Understand the rationale of an RNA-Seq experiment and its design
- Learn about the Linux command line
- Use *Jupyter* (*SRA Toolkit*) to import sequence data
- Use *Jupyter* (*FastQC/Trimmomatic*) to quality check/trim sequence data
- Use *Jupyter* (*Kallisto*) to (pseudo)align reads
- Use *Jupyter* (*genomeview/UCSC*) to explore RNA-Seq results

DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live