Cold Spring Harbor Laboratory
DNA Learning Center

DNALC Live

Intro to RNA-Seq with Jupyter Part I

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 *Live*

- Provide genetics, molecular biology, and bioinformatics learning resources
- Laboratory and computer demos, short online courses for middle school, high school, and the general public
- Interviews with scientists, help for teachers
- At-home activities, social media contests, and more

DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live

DNALC Website and Social Media

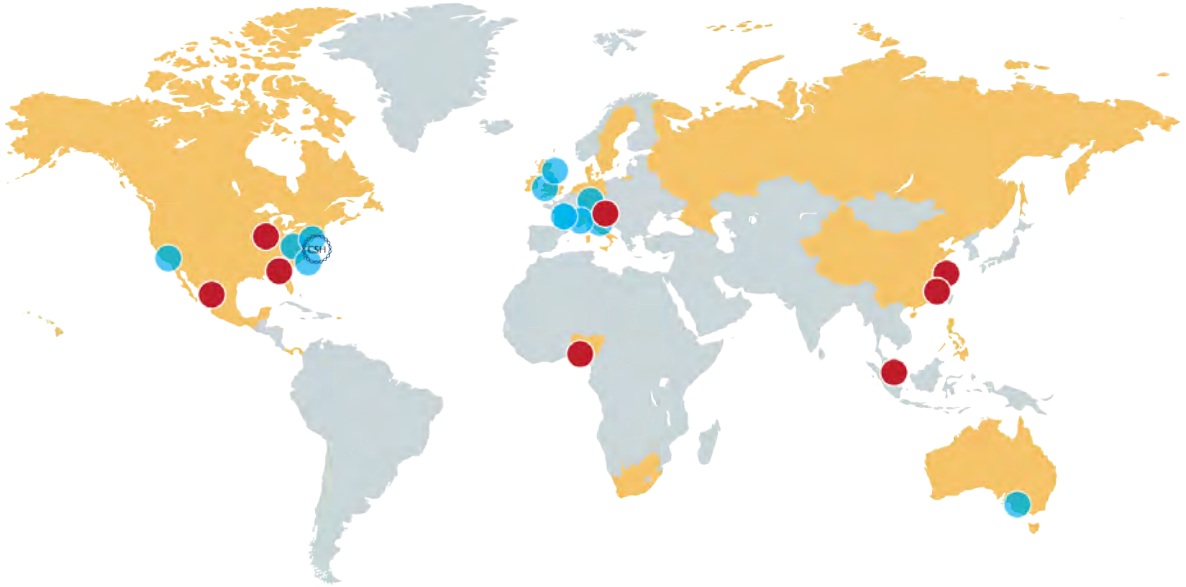
youtube.com/DNALearningCenter

facebook.com/cshldnalc

@dnalc

@dna_learning_center

Cold Spring Harbor Laboratory
DNA Learning Center



- CSHL
- Licensed Centers
- Programs Modeled on the DNALC
- Teacher Training Sites (US States and Countries)

Cold Spring Harbor Laboratory

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 I

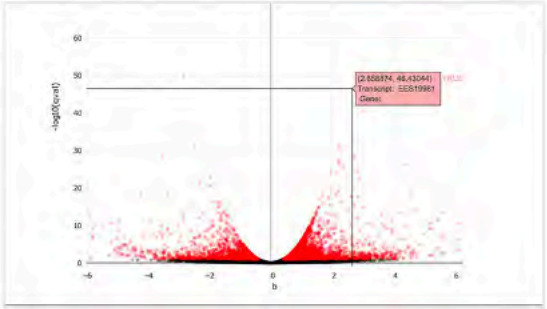
(background and sequence quality)

Steps for today's session

- Introduction to RNA-Seq
- Learn about our example data set
- Learn about high-throughput sequencing and data sources
- Examine DNA sequence quality and QC

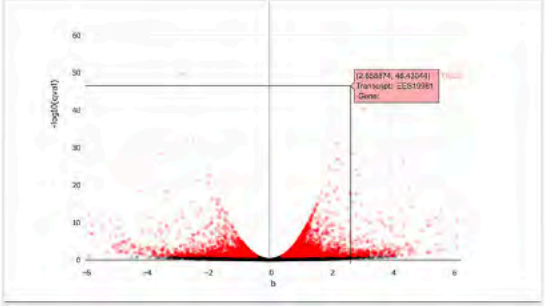
RNA-Seq with *DNA Subway*

dnalc.cshl.edu/dnalc-live



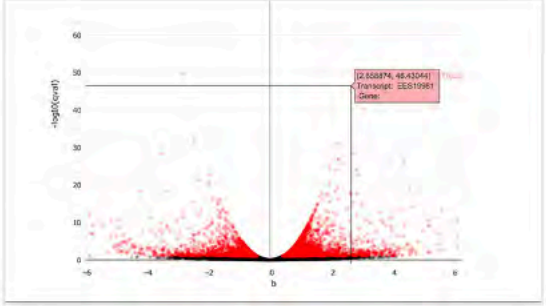
RNA-Seq with DNA Subway, Part I

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

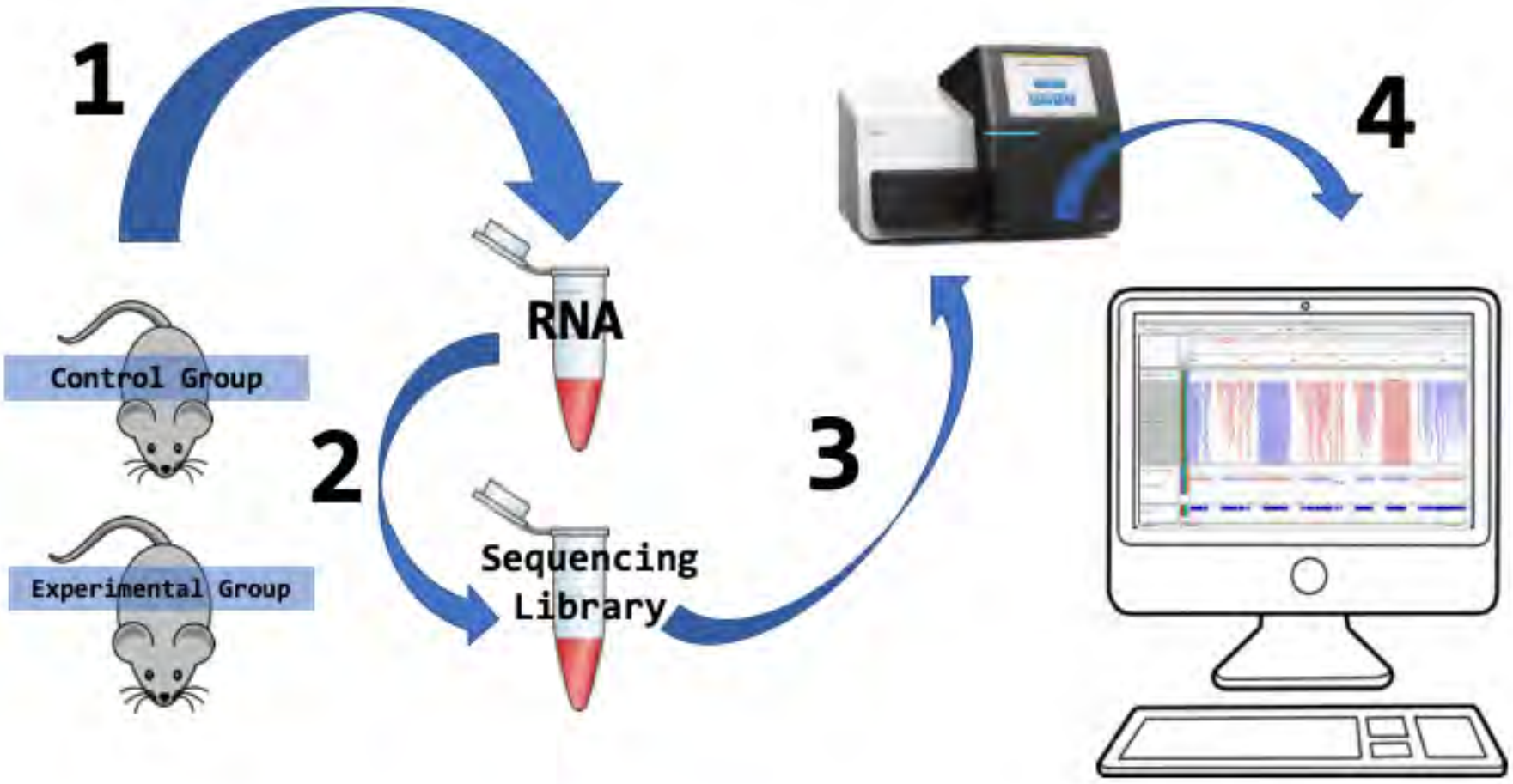
- RNA-Seq allows us to measure the transcriptome – take an account of all transcription occurring in a cell/tissue

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

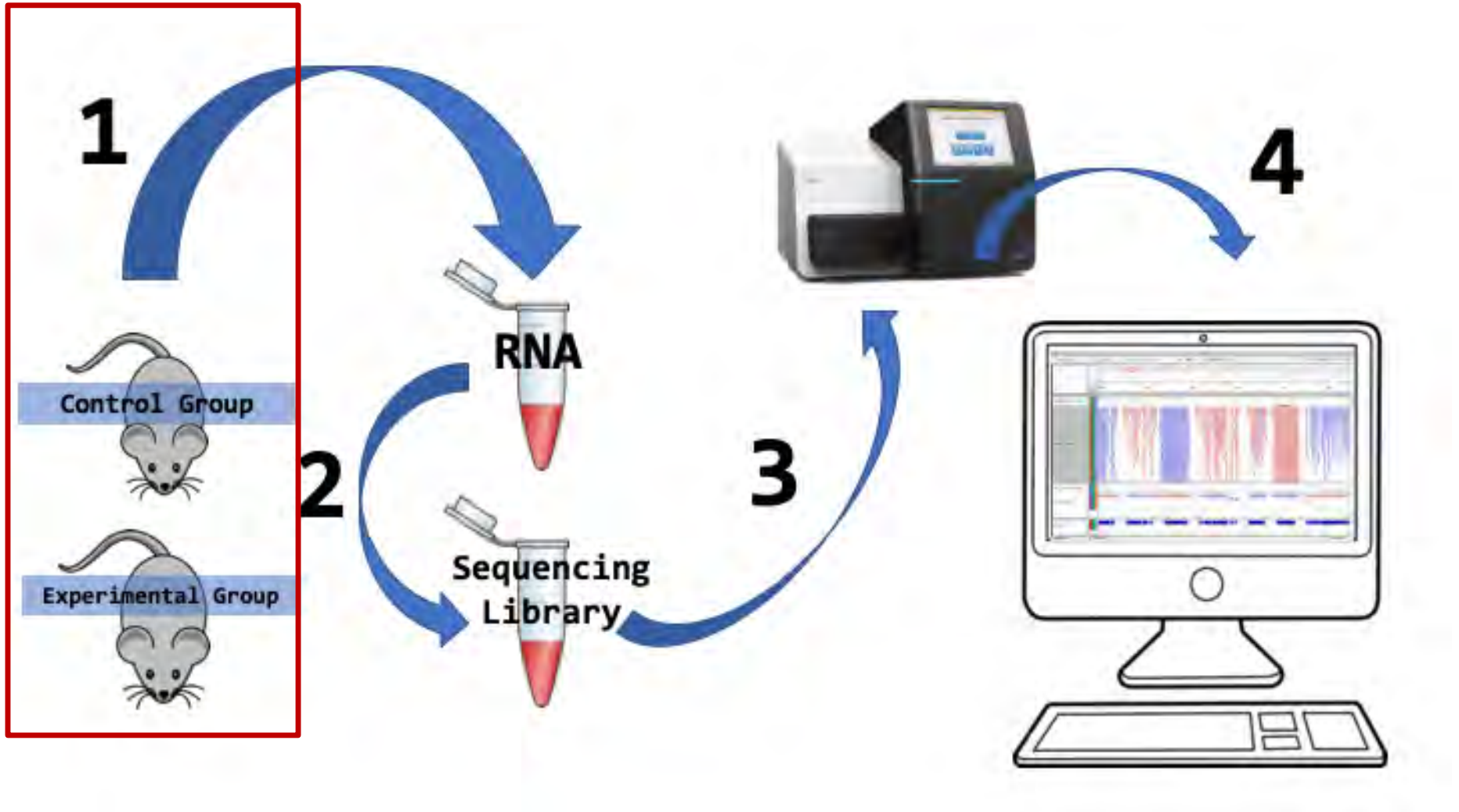
What is RNA-Seq? - measuring the transcriptome

- 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)

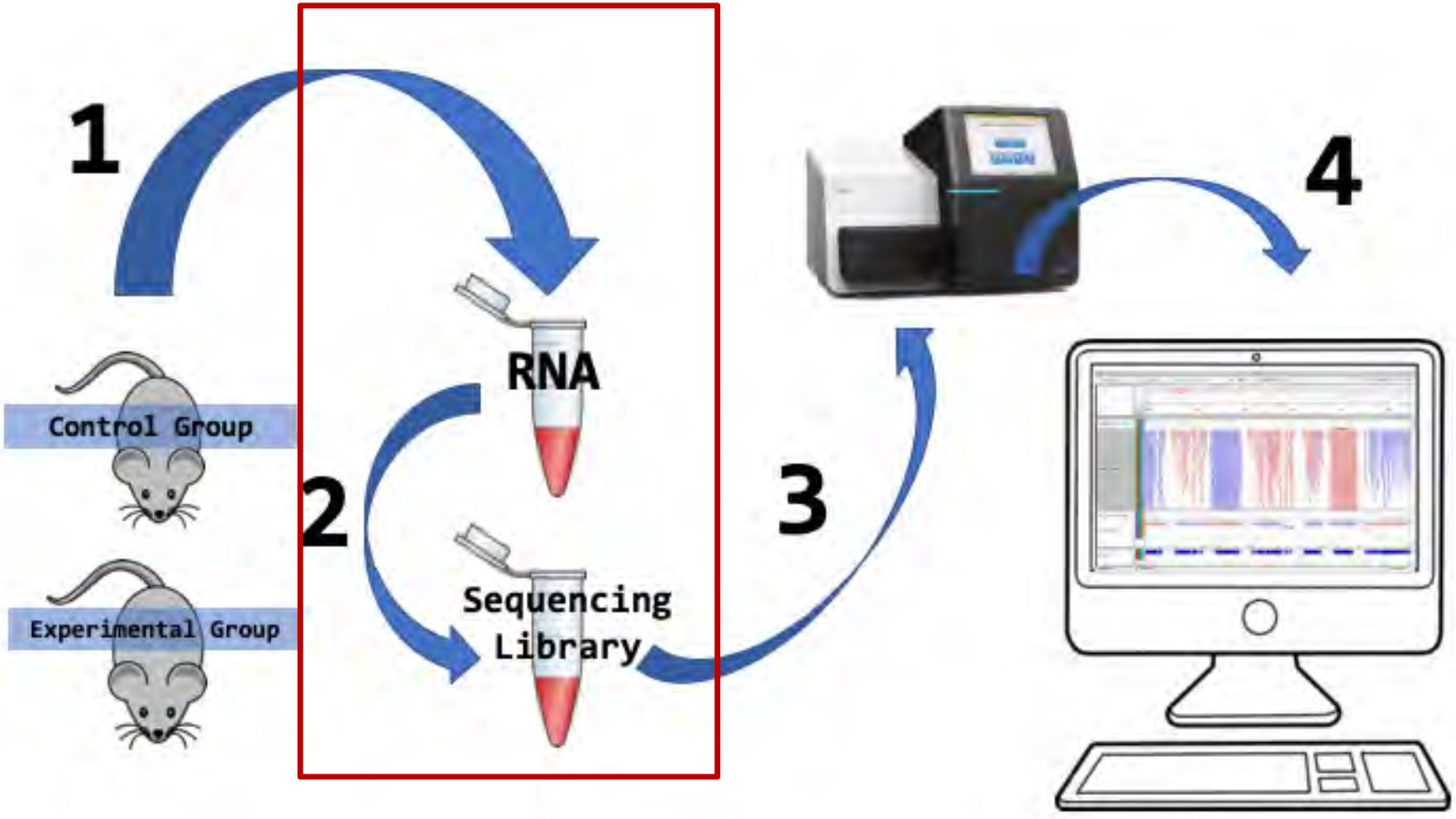
What is RNA-Seq?



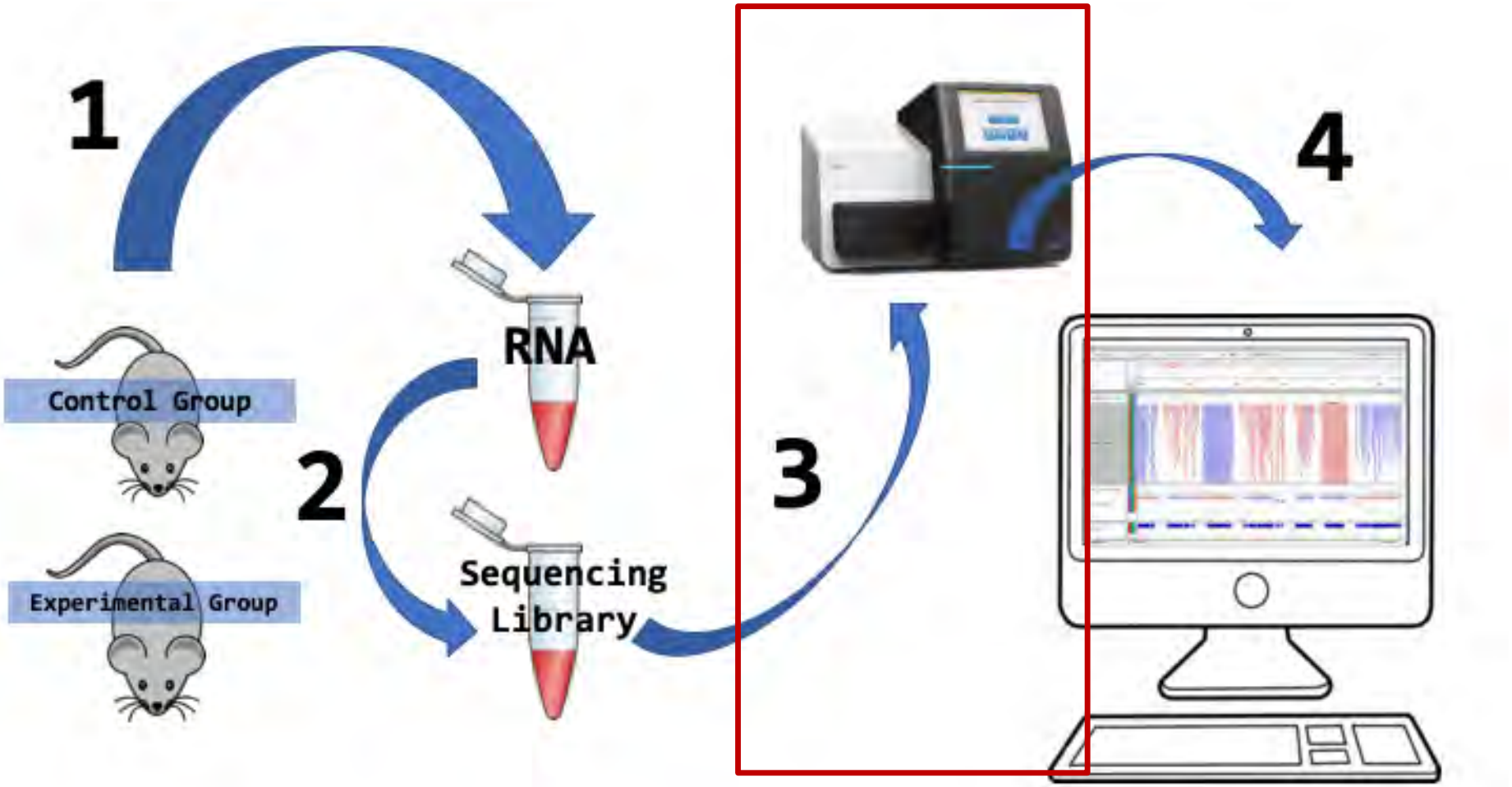
What is RNA-Seq?



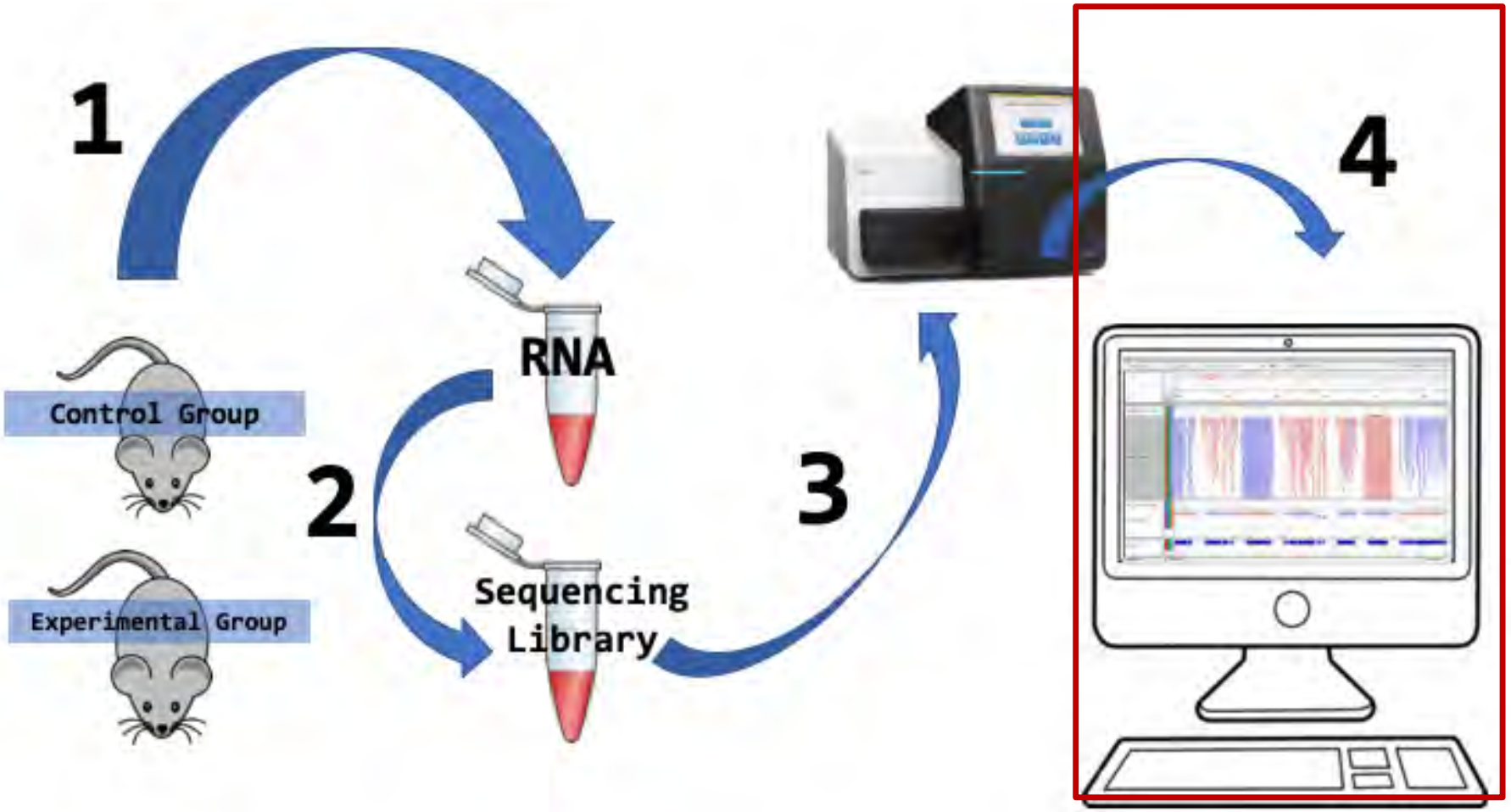
What is RNA-Seq?



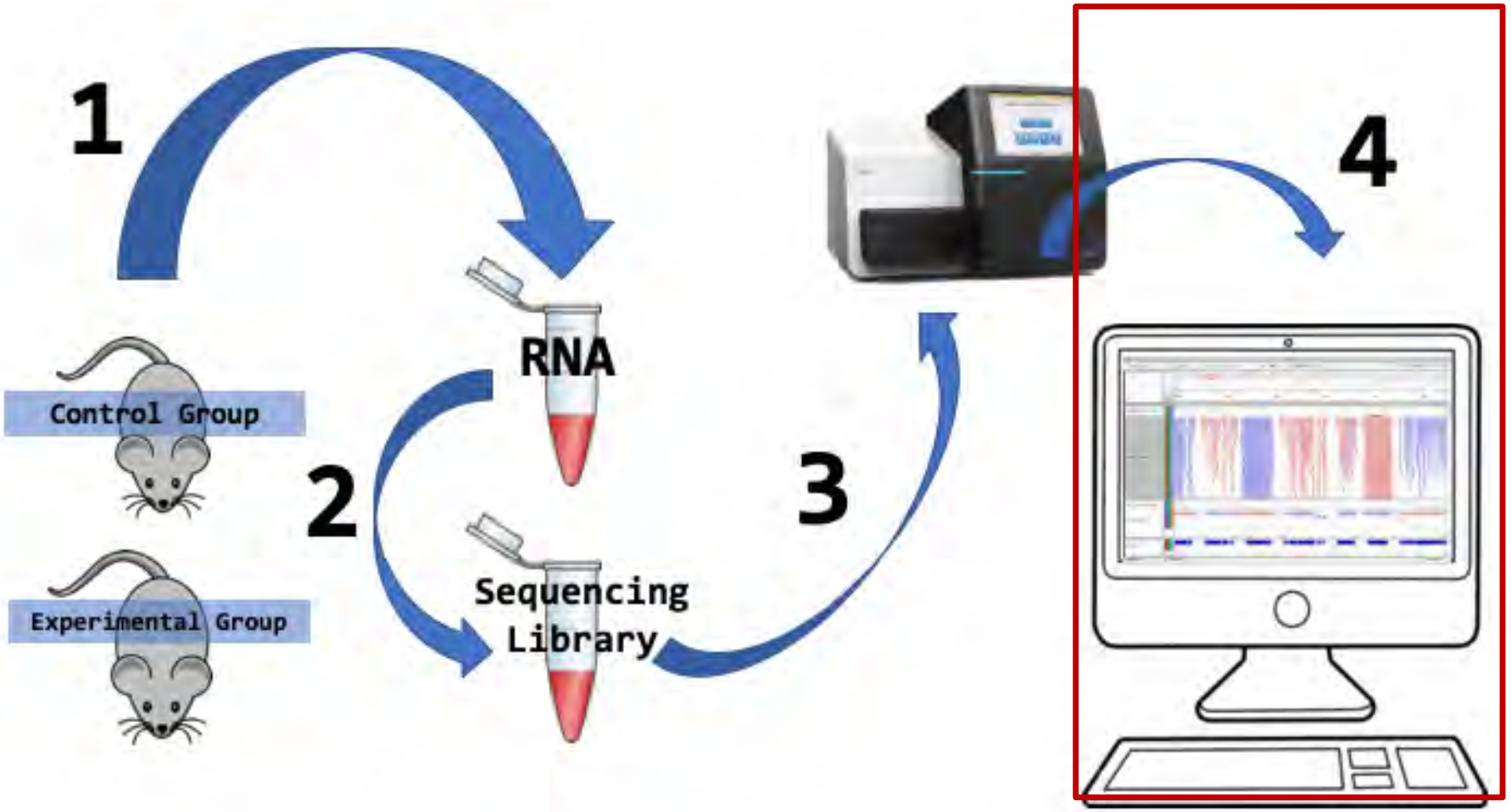
What is RNA-Seq?



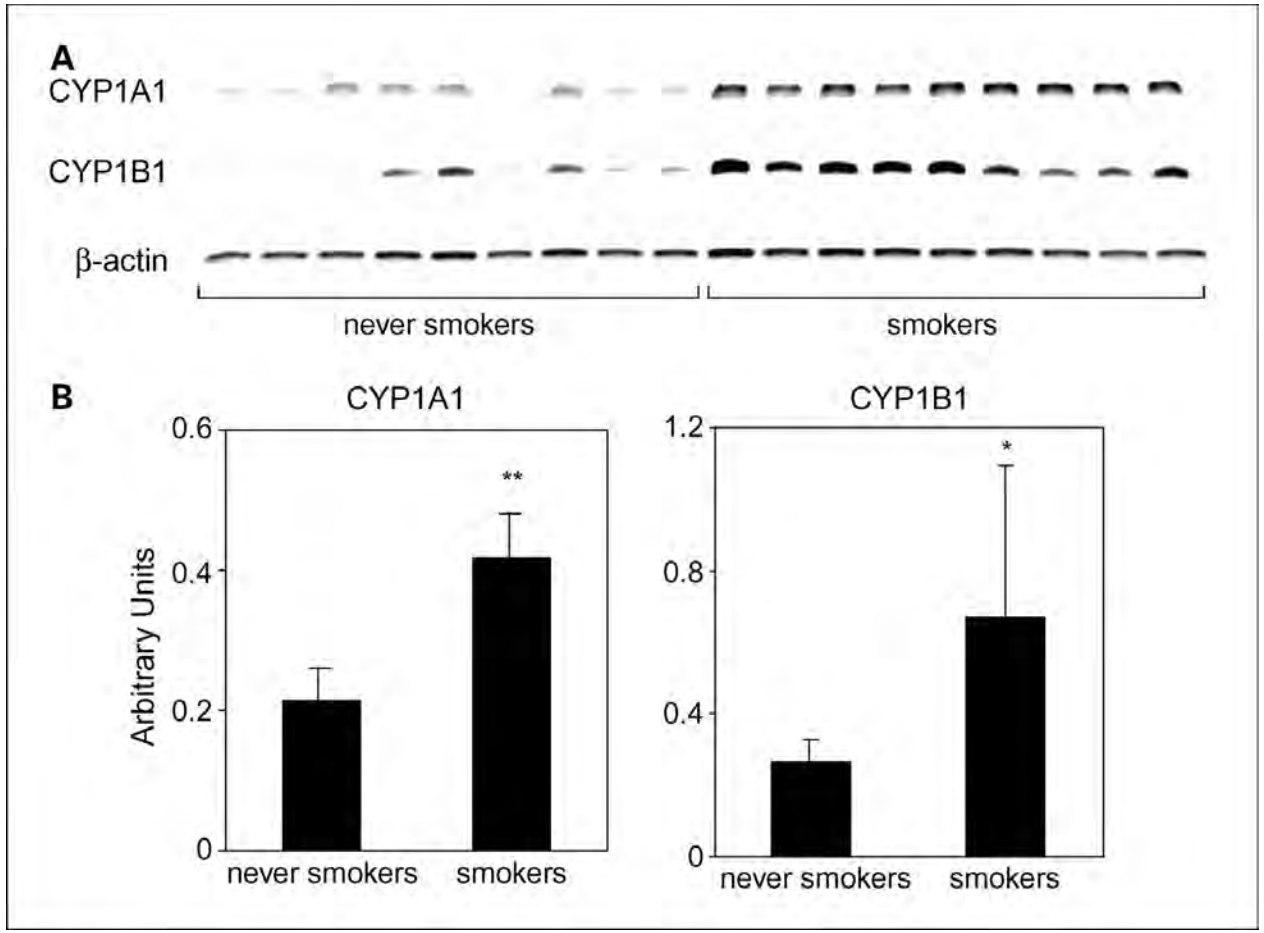
What is RNA-Seq?



What is RNA-Seq?



What can expression tell you?



- CYP1A/1B – Cytochrome p450 family, involved in drug metabolism including processing toxins

Photo Credit:
Effects of Tobacco Smoke on Gene Expression and Cellular Pathways in a Cellular Model of Oral Leukoplakia
Zeynep H. Gümüş, Baoheng Du, Ashutosh Kacker, Jay O. Boyle, Jennifer M. Bocker, Piali Mukherjee, Kotha Subbaramaiah, Andrew J. Dannenberg and Harel Weinstein
Cancer Prev Res July 1 2008 (1) (2) 100-111; DOI: 10.1158/1940-6207.CAPR-08-0007

What can expression tell you?

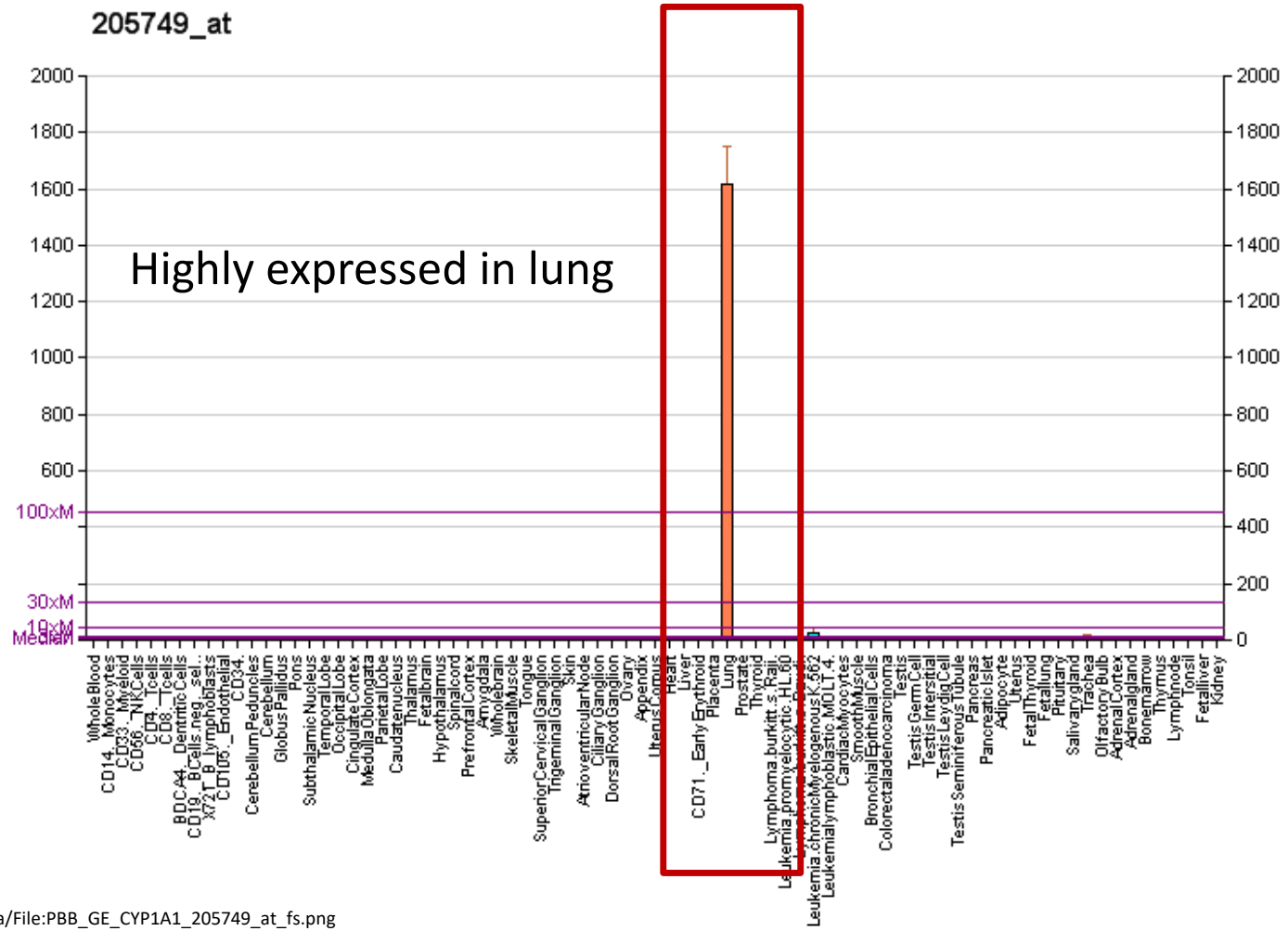


Photo credit:
https://en.wikipedia.org/wiki/CYP1A1#/media/File:PBB_GE_CYP1A1_205749_at_fs.png

Introduction to our data set

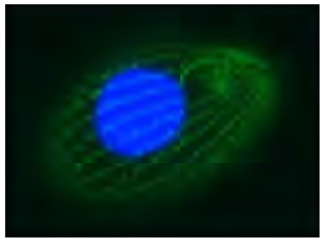
GEA

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

Development of pilot lessons

- **BLAST**
  - Introduce students to the use of the Basic Local Alignment Search Tool (BLAST) to identify related sequences and compare similarity between them
- **Understanding Eukaryotic Genes**
  - Familiarize students with a genome browser (UCSC)
- **RNA-Seq**
  - 3 modules: Use RNA-Seq as a bridge to more advanced bioinformatics and data science (QC, quantification (Kallisto), visualization (UCSC/IGV))

Using Jupyter

Jupyter is a platform that makes it easier to present, organize, and share code/command line tools

Leptin expression vs. diet – RNA-Seq pilot lesson

GEA
Genomics Education Alliance

Try the pilot lessons/join us!

https://gea.qubeshub.org/lessons

RCN-UBE #1827130

Leptin expression vs. diet – RNA-Seq pilot lesson

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

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** ▾

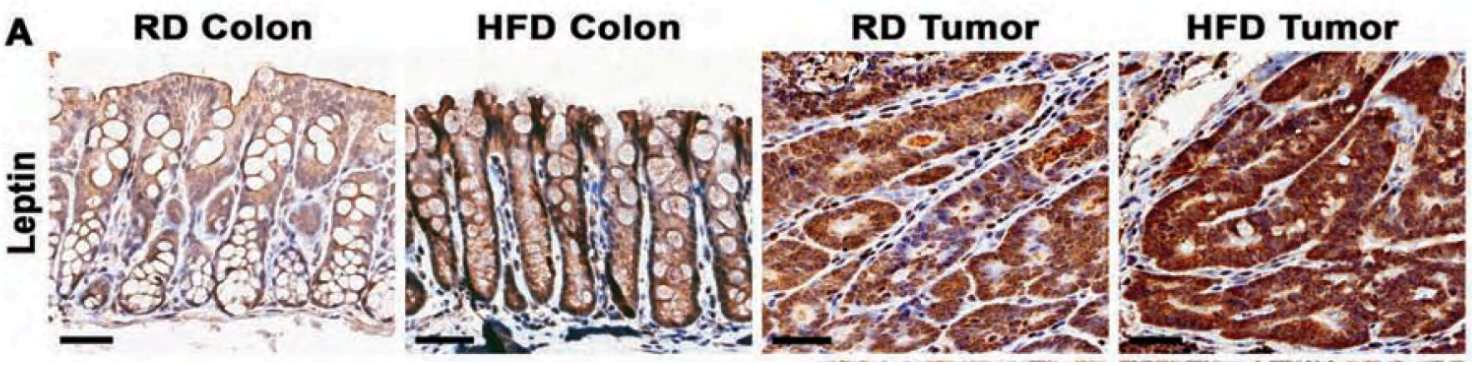
Leptin expression vs. diet – RNA-Seq pilot lesson

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

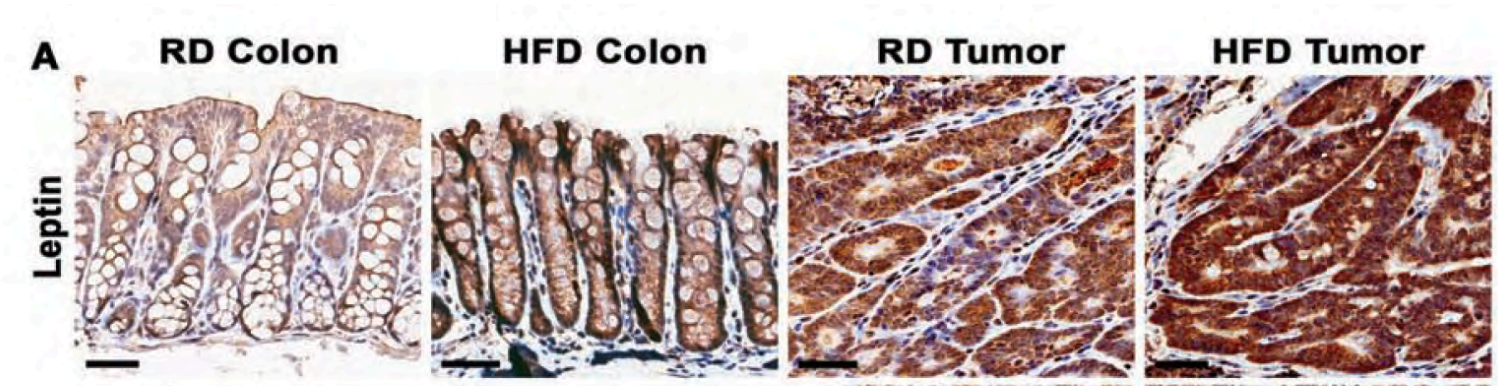
Carcinogenesis, Volume 38, Issue 3, 1 March 2017, Pages 302–311,
https://doi.org/10.1093/carcin/bgx001
Published: 25 January 2017 **Article history** ▾



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

Leptin expression vs. diet – RNA-Seq pilot lesson

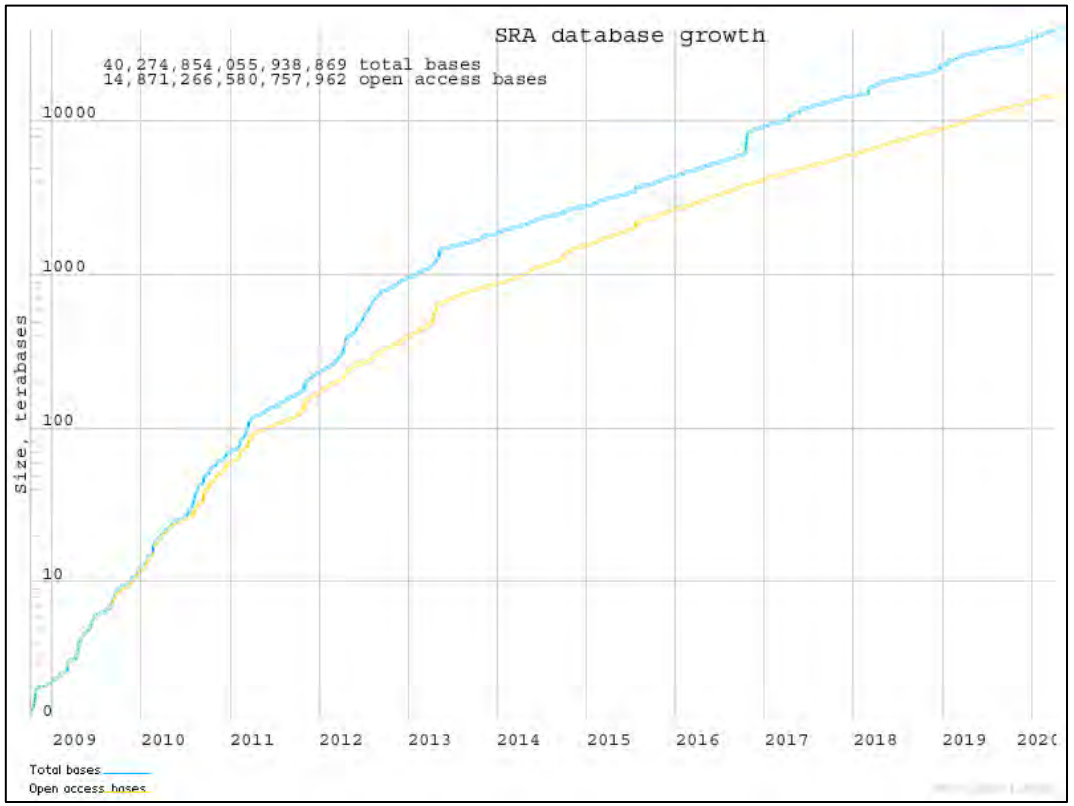
SRA_Sample	Sample_Name
SRS1794108	High-Fat Diet Control 1
SRS1794110	High-Fat Diet Control 2
SRS1794106	High-Fat Diet Control 3
SRS1794105	High-Fat Diet Tumor 1
SRS1794101	High-Fat Diet Tumor 2
SRS1794111	High-Fat Diet Tumor 3



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

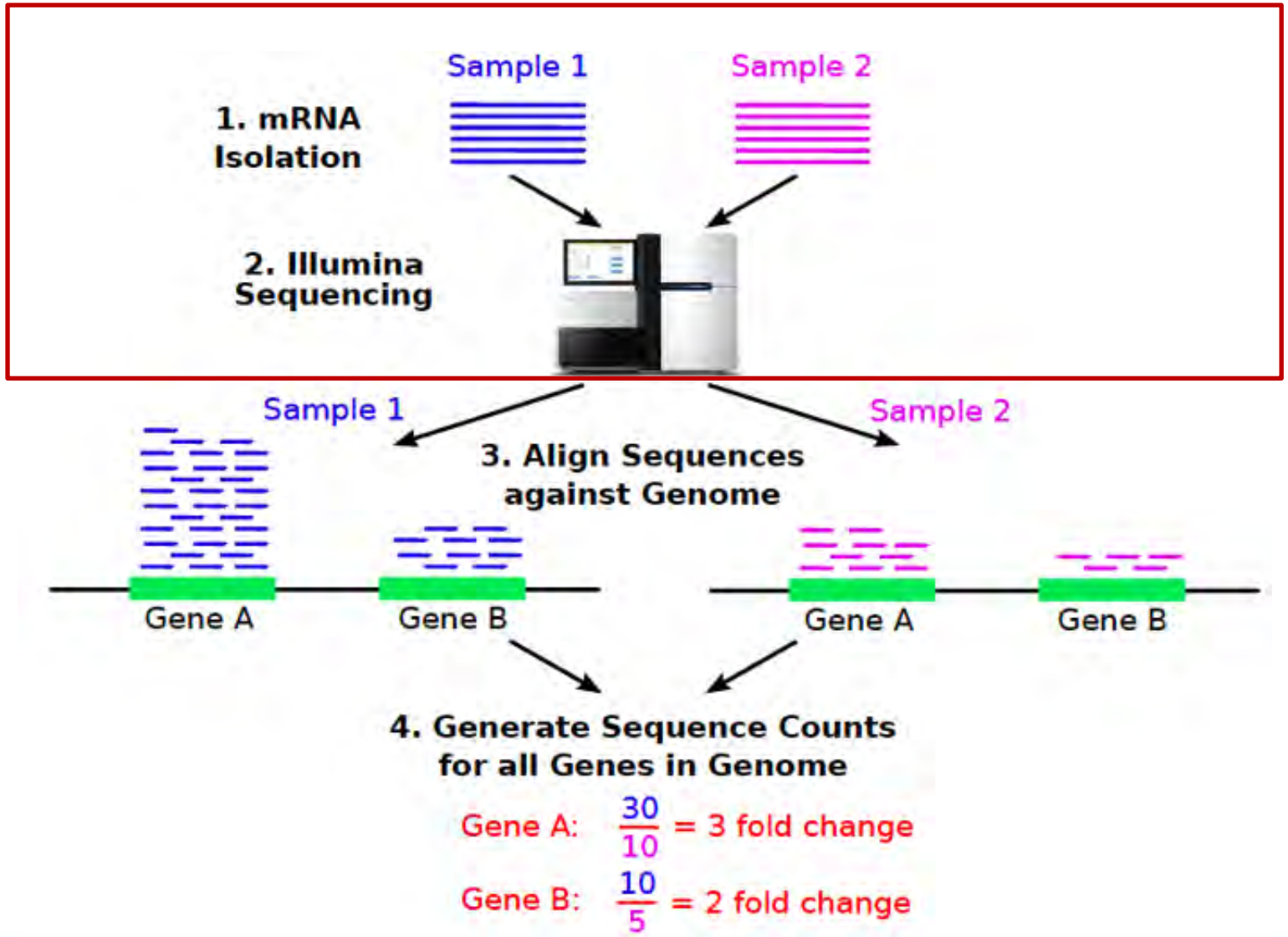
Sequence data from NCBI

National Center for Biotechnology Information
NCBI

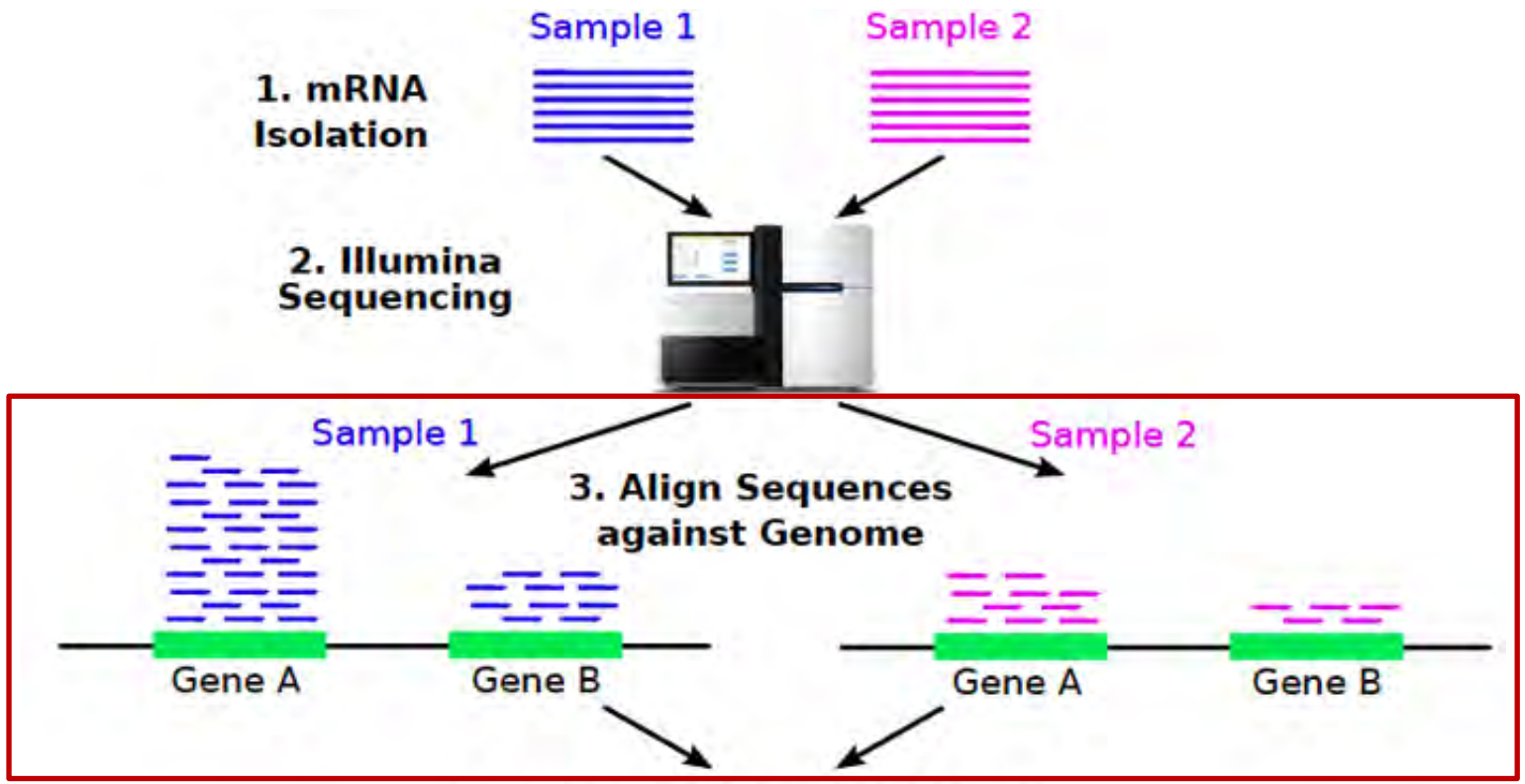


https://www.ncbi.nlm.nih.gov/bioproject/PRJNA353374

Sequencing reads



Sequencing reads

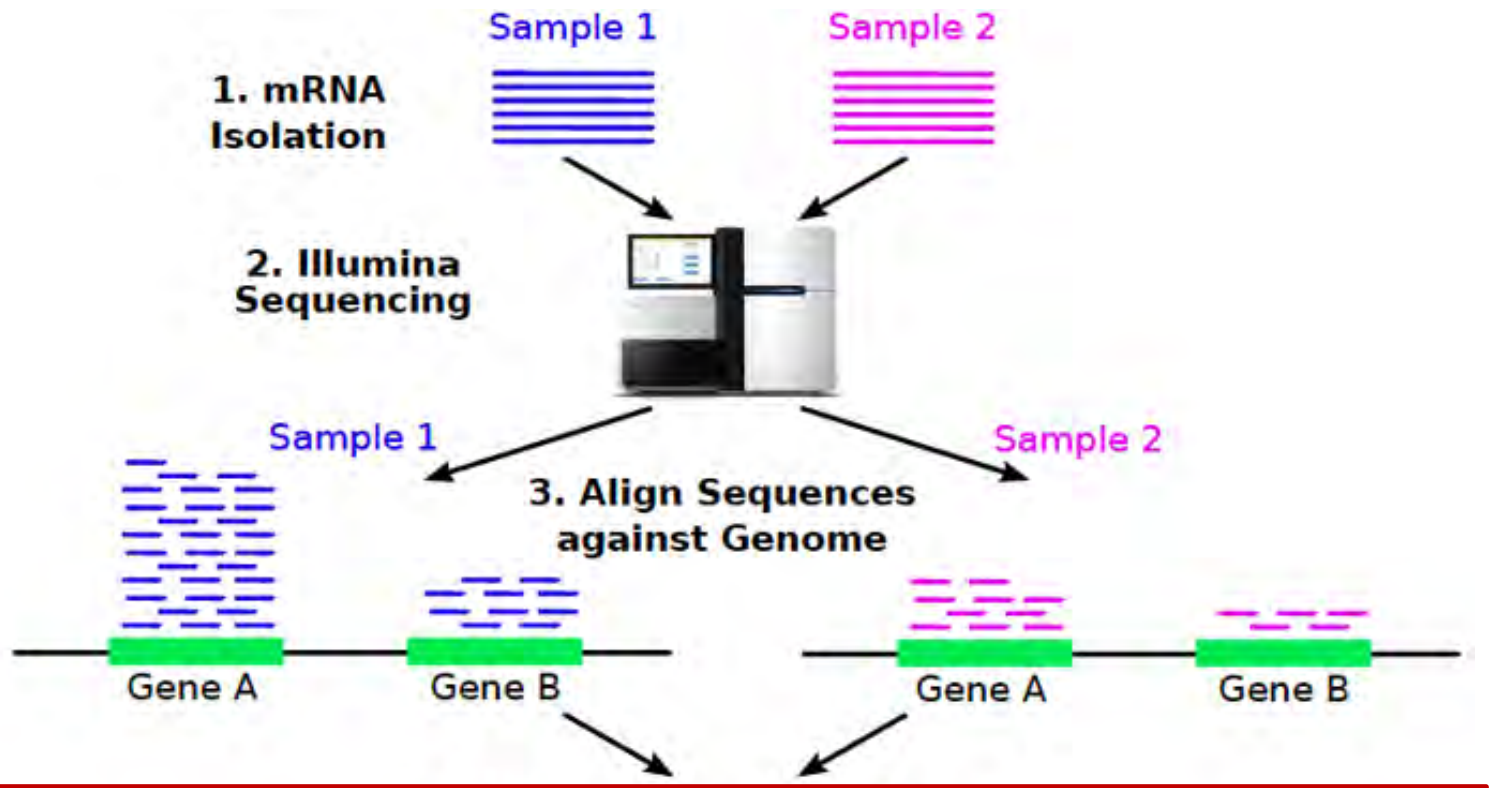


4. Generate Sequence Counts for all Genes in Genome

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

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

Sequencing reads

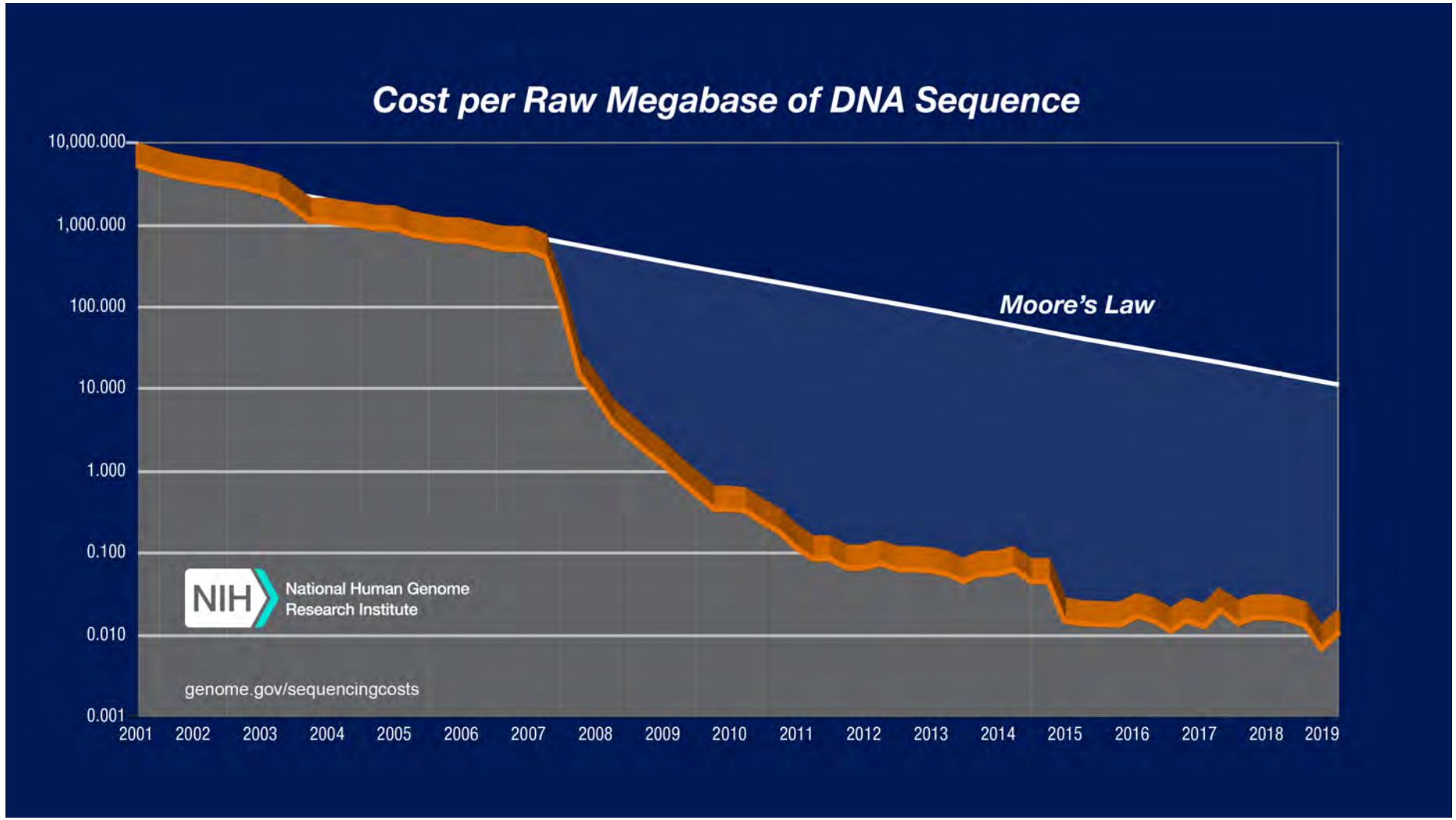


4. Generate Sequence Counts for all Genes in Genome

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

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

Sequencing revolution



Lab – Sequence import from SRA

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

Lab: Sequence QC

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

Key Concept: Sequence quality

Examining quality with FastQC

FastQC Report

Wed 8 Apr 2020
SRR3191542_1.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

Basic Statistics

Measure	Value
Filename	SRR3191542_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	7927777
Sequences flagged as poor quality	0
Sequence length	35-76
%GC	48

Per base sequence quality

Produced by FastQC (version 0.11.5)

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%

If 99% was good enough

If things only work correctly 99.9% of the time...

- 12 newborns will be given to the wrong parents daily.
- 114,500 mismatched pairs of shoes will be shipped/year.
- 18,322 pieces of mail will be mishandled/hour.
- 2,000,000 documents will be lost by the IRS this year.
- 2.5 million books will be shipped with the wrong covers.
- Two planes landed at Chicago's O'Hare airport will be unsafe every day.
- 315 entries in Webster's Dictionary will be misspelled.
- 20,000 incorrect drug prescriptions will be written this year.
- 880,000 credit cards in circulation will turn out to have incorrect
- cardholder information on their magnetic strips.
- 103,260 income tax returns will be processed incorrectly during the year.
- 5.5 million cases of soft drinks produced will be flat.
- 291 pacemaker operations will be performed incorrectly.
- 3056 copies of tomorrow's Wall Street Journal will be missing one of the three sections.

Photo credit
http://www.personal.psu.edu/sxt104/class/99percent.html

Next time:
Sequence alignment to reference and visualization

DNALC Website and Social Media

dnalc.cshl.edu

dnalc.cshl.edu/dnalc-live