Big Data Genomics Education for Undergraduate Professors

Jeffrey Chuang
The Jackson Laboratory for Genomic Medicine

May 2018

The Jackson Laboratory
Leading the search for tomorrow's cures

R25 for Genomics Training

JAX Home > Education & Learning

BIG GENOMIC DATA SKILLS TRAINING FOR GRADUATE PROFESSORS

Location: **The Jackson Laboratory for Genomic Medicine - JAX Genomic Medicine, Farmington CT**

5:00 pm EDT
MAY
14 - 19
2017

The purpose of the course is to offer training to professors who teach graduate students. The goal of this program is to enable professors to launch comprehensive Big Data or Computational Biology training for graduate students at their institution. The course will cover a variety of genomic data analysis topics and allow faculty to jump start their knowledge of experimental design, data analysis, data integration and data visualization.

The program is funded by the NIH Big Data to Knowledge (BD2K) Initiative.

Please contact the event organizer for more information

Introduction of course organizers and staff

Prof. Jeffrey Chuang – JAX; U. Conn Health Department of Genetics and Genome Sciences

Charlie Wray, PhD - JAX
Director of Courses and Conferences

Prof. Reinhard Laubenbacher – JAX; U. Conn. Health: Director of the Department of Quantitative Medicine

Ada Zhan, PhD – JAX
Postdoctoral Fellow

Spencer Glantz - JAX
Postdoctoral Fellow

Amanda Lazarus,
Administrative Assistant
JAX Genomic Education

Sandeep Namburi, Cloud Systems Analyst, JAX IT

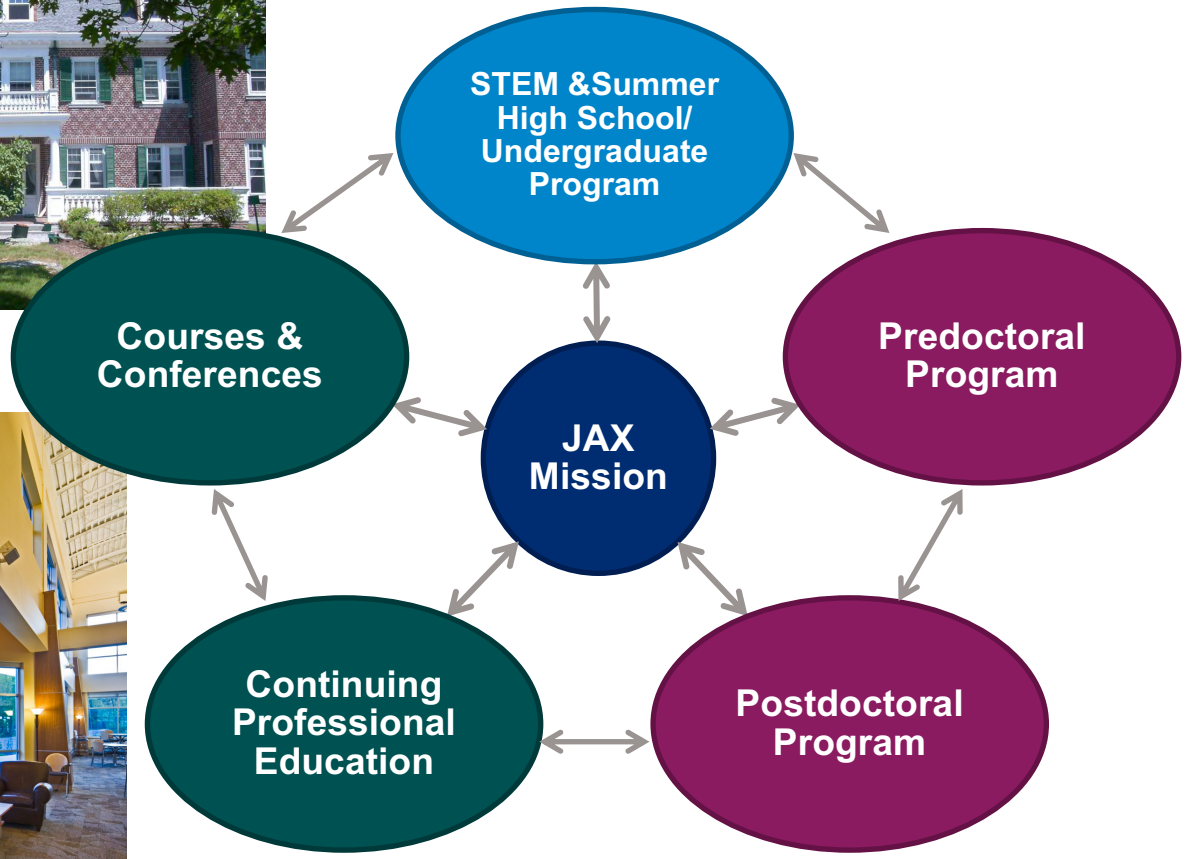
Aditya Kovuri, Programmer Analyst, JAX IT

Overview of The Jackson Laboratory Education

Mission:

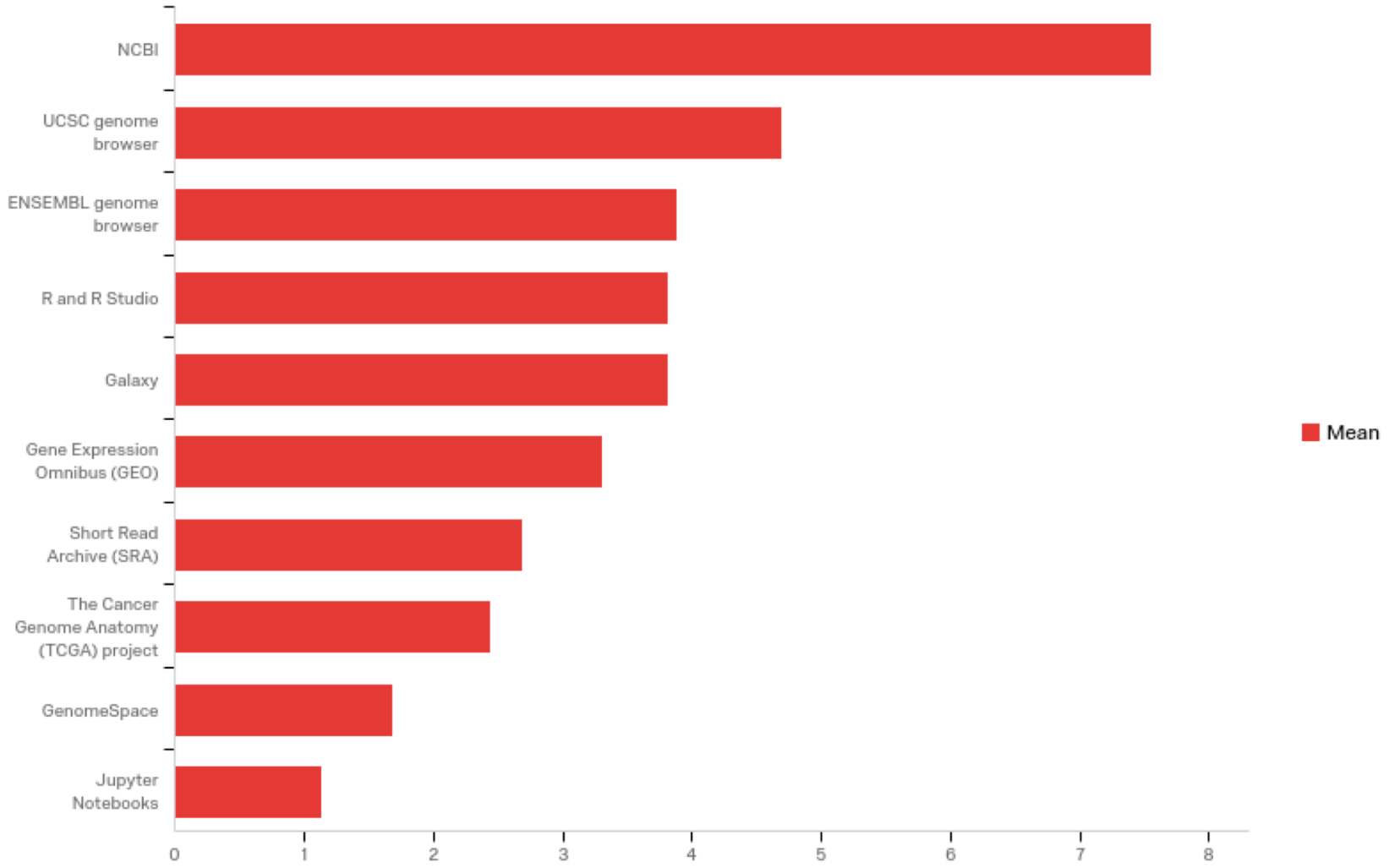
We discover precise genomic solutions for disease and empower the global biomedical community in our shared quest to improve human health.

The JAX Education Ecosystem

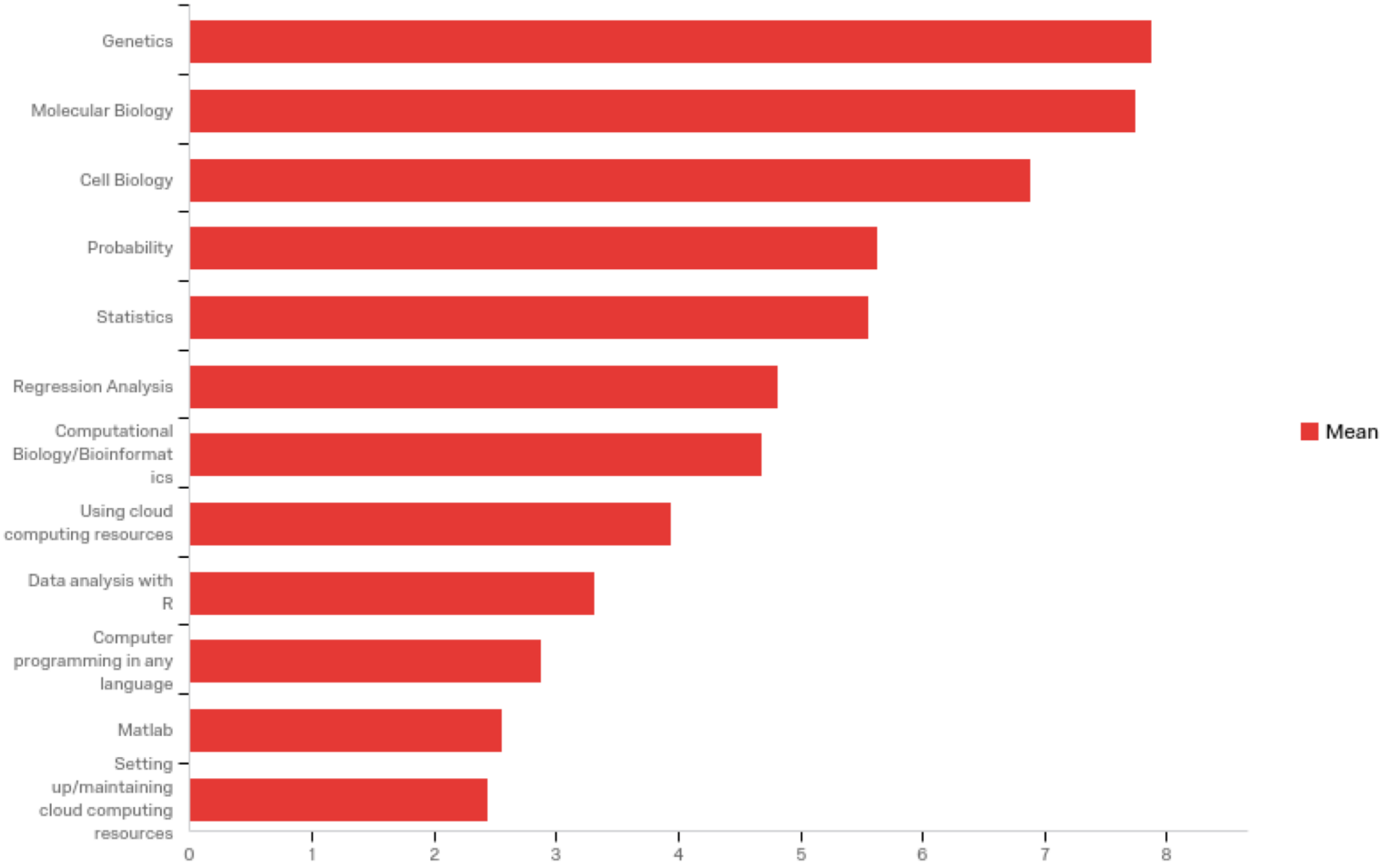


R25 Participant Survey

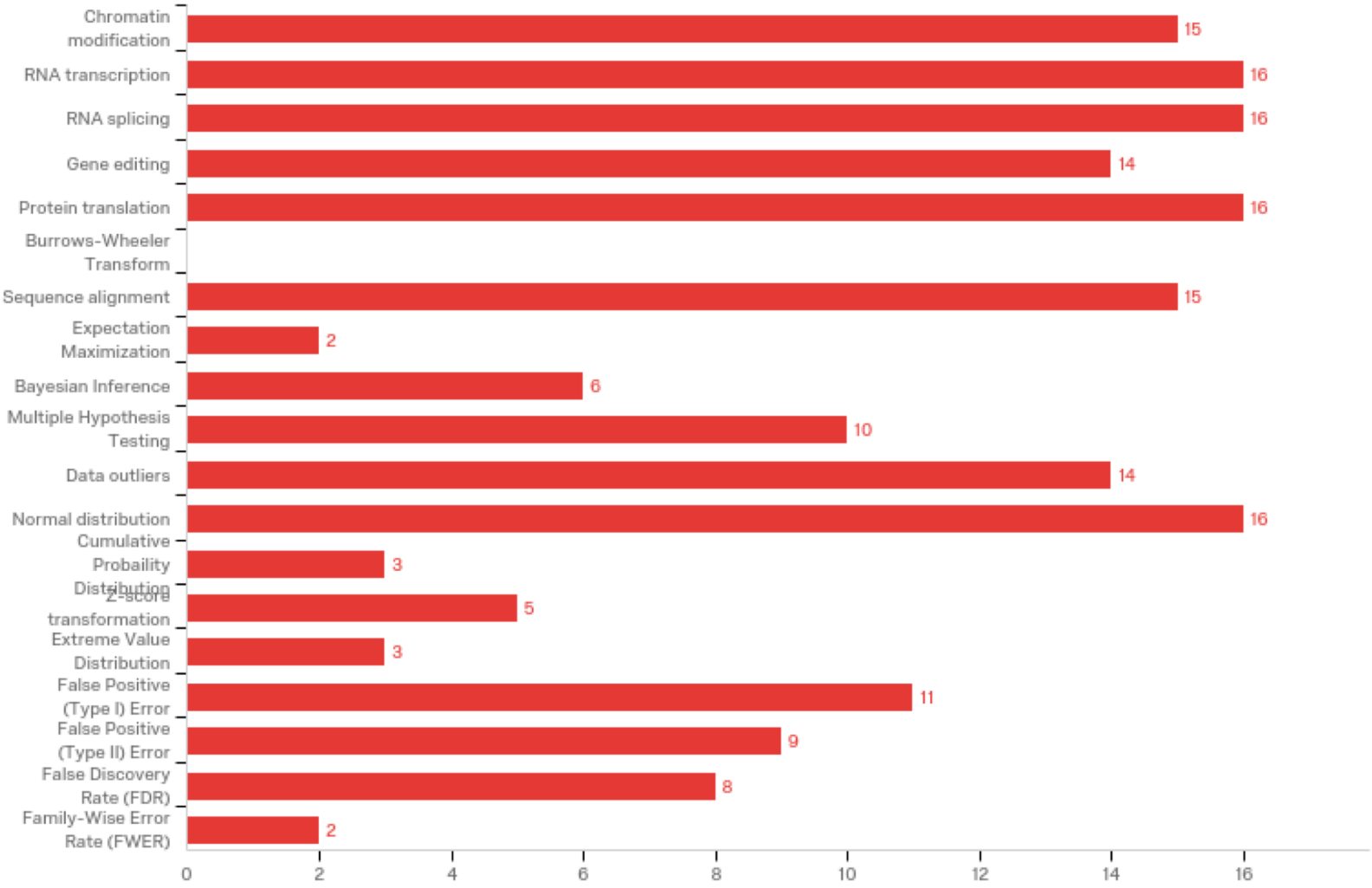
Q1 - From 1 (least) to 10 (highest), rate your knowledge of/familiarity with each of the following:



Q2 - From 1 (no knowledge) to 10 (expert), rate your expertise on each of the following:



Q3 - Which of the following terms could you define without the use of Google or another search engine (select all that apply):



Q3 - Which of the following terms could you define without the use of Google or another search engine (select all that apply):

#	Answer	%	Count
1	Chromatin modification	8.29%	15
2	RNA transcription	8.84%	16
3	RNA splicing	8.84%	16
4	Gene editing	7.73%	14
5	Protein translation	8.84%	16
6	Burrows-Wheeler Transform	0.00%	0
7	Sequence alignment	8.29%	15
8	Expectation Maximization	1.10%	2
9	Bayesian Inference	3.31%	6

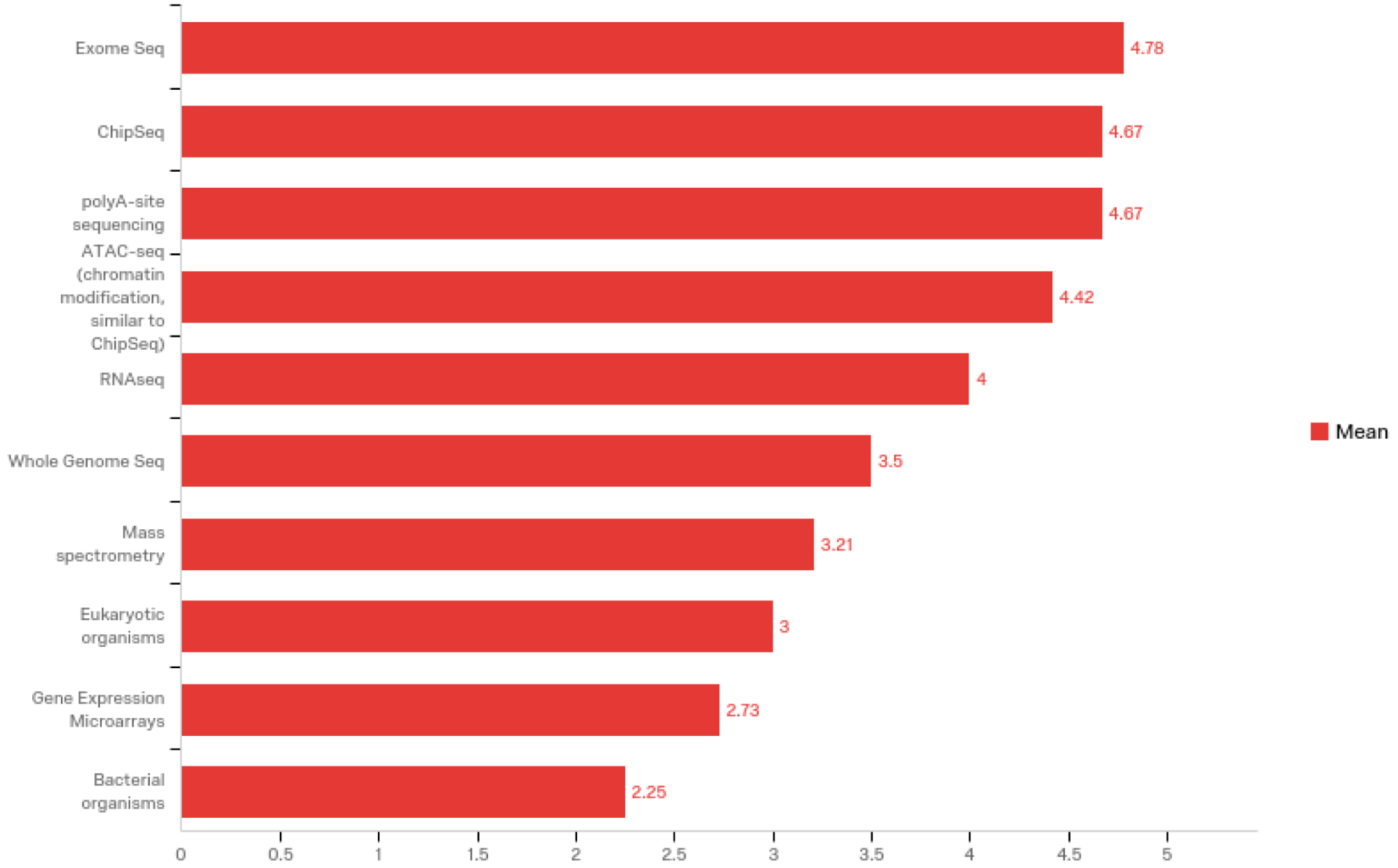
Q3 - Which of the following terms could you define without the use of Google or another search engine (select all that apply):

#	Answer	%	Count
10	Multiple Hypothesis Testing	5.52%	10
11	Data outliers	7.73%	14
12	Normal distribution	8.84%	16
13	Cumulative Probaility Distribution	1.66%	3
14	Z-score transformation	2.76%	5
15	Extreme Value Distribution	1.66%	3
16	False Positive (Type I) Error	6.08%	11
17	False Positive (Type II) Error	4.97%	9

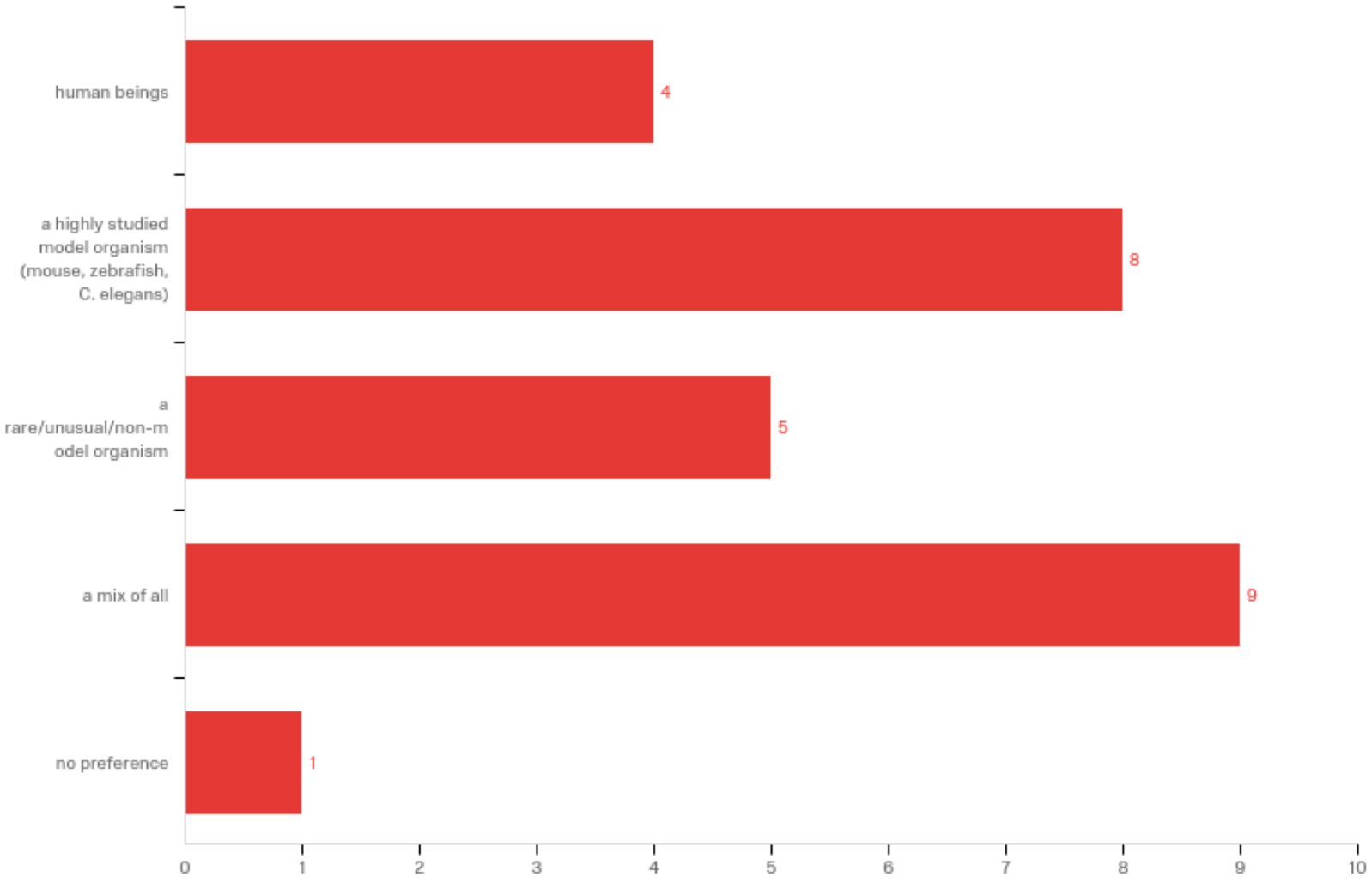
Q3 - Which of the following terms could you define without the use of Google or another search engine (select all that apply):

#	Answer	%	Count
18	False Discovery Rate (FDR)	4.42%	8
19	Family-Wise Error Rate (FWER)	1.10%	2
	Total	100%	181

Q4 - From 1 (not interested) to 5 (very interested), rate your interest in learning to work with the following data types. Please select N/A if you are not familiar with the data type.



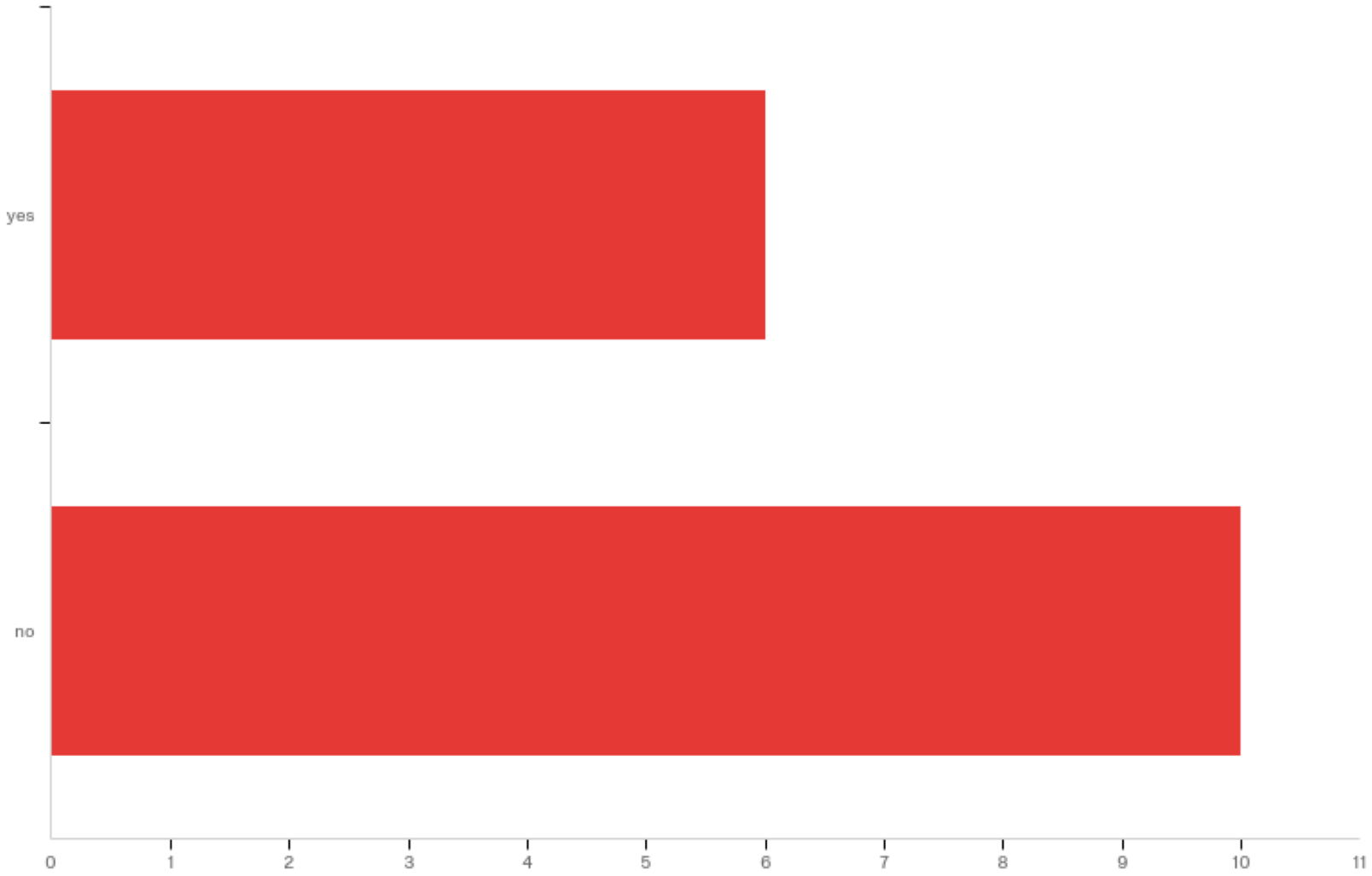
Q5 - Which are you more interested in data generated from:



Q5 - Which are you more interested in data generated from:

#	Answer	%	Count
1	human beings	14.81%	4
2	a highly studied model organism (mouse, zebrafish, C. elegans)	29.63%	8
3	a rare/unusual/non-model organism	18.52%	5
4	a mix of all	33.33%	9
5	no preference	3.70%	1
	Total	100%	27

Q10 - Do you have genomic data you might bring to the course?



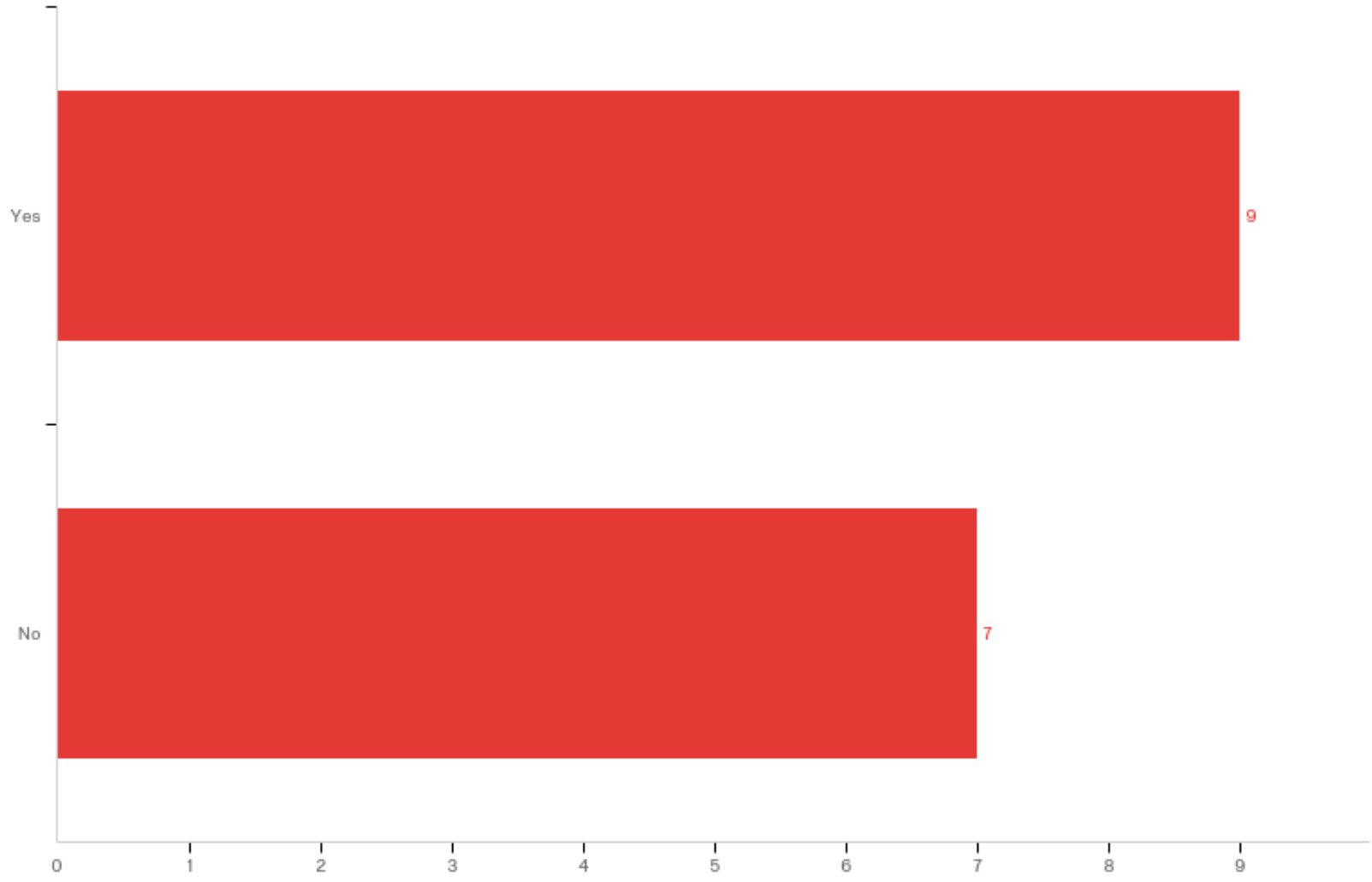
Q10 - Do you have genomic data you might bring to the course?

#	Answer	%	Count
1	yes	37.50%	6
2	no	62.50%	10
	Total	100%	16

Q11 - Please describe data you may bring to the course:

Please describe data you may bring to the course:
I have soft-shell clam genomic data in Galaxy
RNA Seq data from mouse eye treated with NaOH +/- transcription inhibitor
I have several bacterial (Burkholderia) genomes I could bring.
I have proteomic data but not genomic data currently available
Drosophila melanogaster brain RNAseq data
Shotgun Metagenomics and 16S marker gene metagenomic data sets
I have 2 data sets comparing treated and untreated cancer lines and looking at transcript expression using Affymetrix GeneChip Human Transcriptome Arrays 2.0

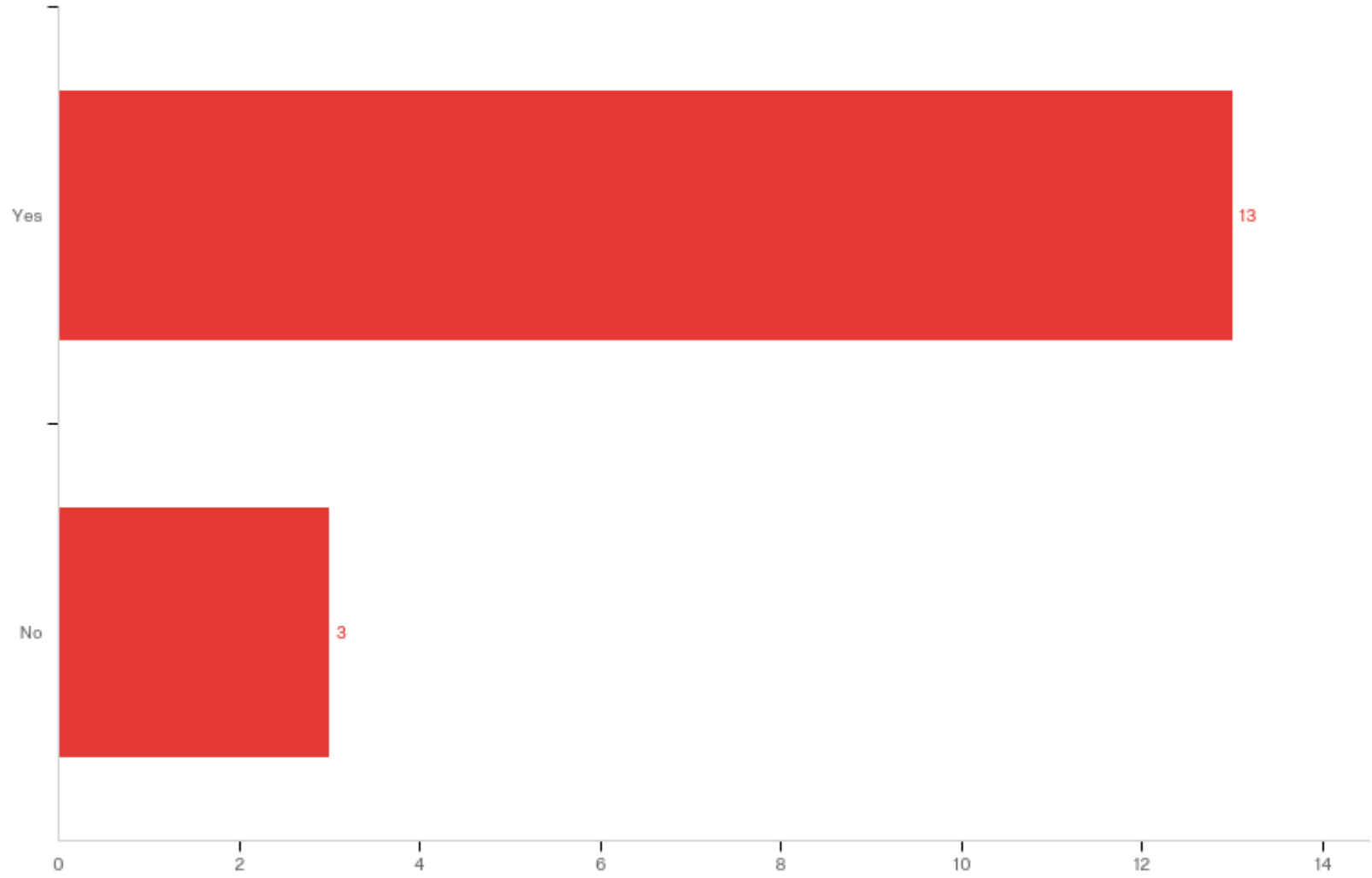
Q6 - Do you have access to a computing cluster that students will be able to login to and perform calculations?



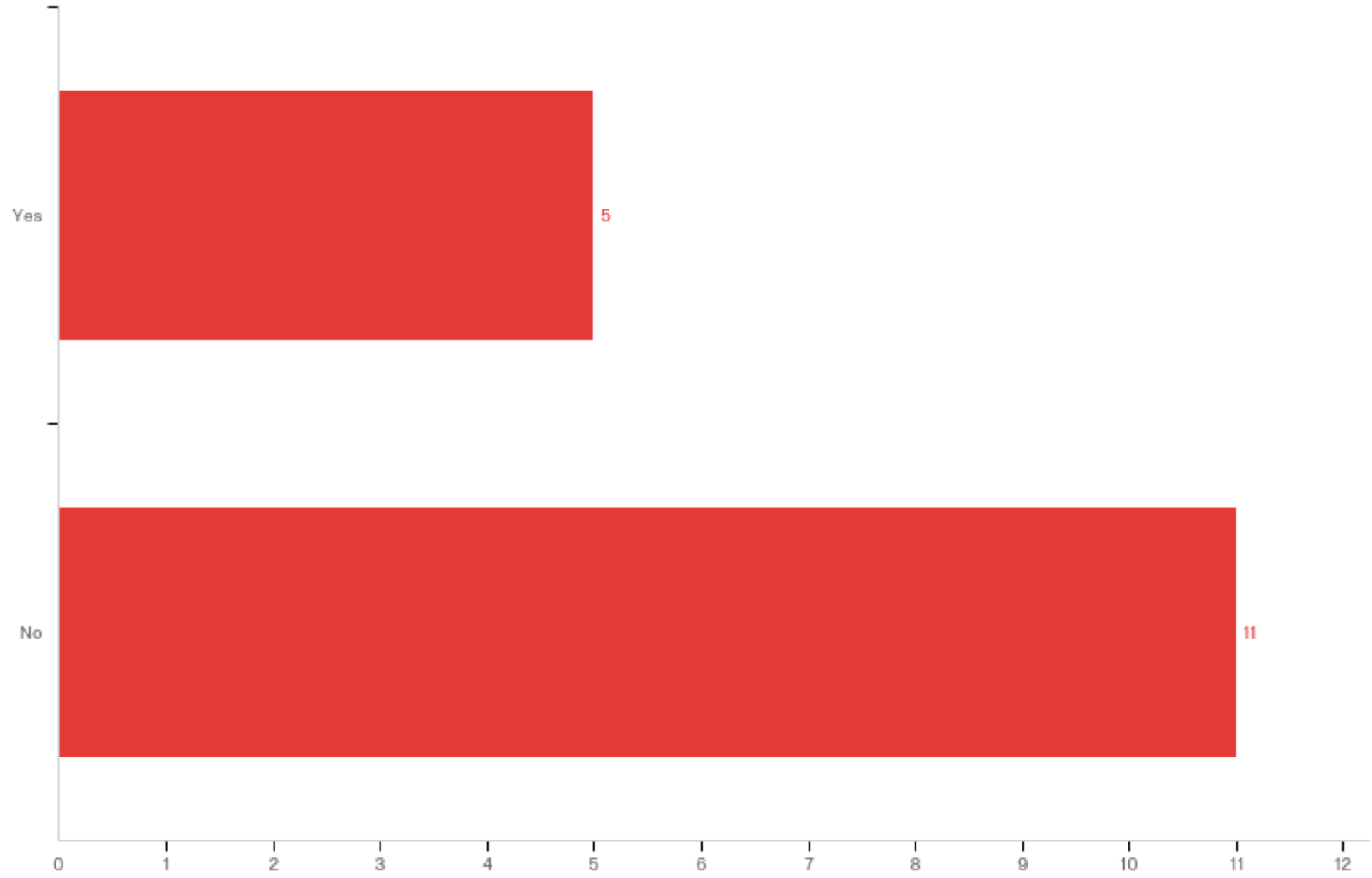
Q6 - Do you have access to a computing cluster that students will be able to login to and perform calculations?

#	Answer	%	Count
1	Yes	56.25%	9
2	No	43.75%	7
	Total	100%	16

Q8 - Does your college or university have dedicated IT support that can assist you with activities such as software installation?



Q9 - Do you have access to teaching assistants to help you mentor students through computational lab activities?



Q9 - Do you have access to teaching assistants to help you mentor students through computational lab activities?

#	Answer	%	Count
1	Yes	31.25%	5
2	No	68.75%	11
	Total	100%	16

Introduction of Participants

Outline of Schedule

Speakers

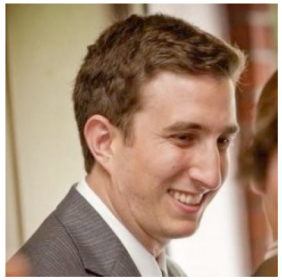
Christine Beck
JAX
Human Genetics

Sheng Li
JAX
RNA-seq

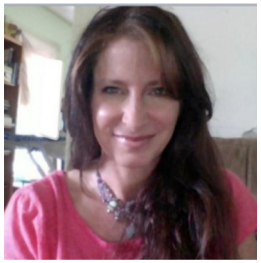
Laura Reinholdt
JAX
Mouse genetic variation

Judy Blake
JAX
Reproducibility

Greg Carter
JAX
Data Integration



Michael Linderman
Middlebury College
Running an UG Course



Andrea Tilden
Colby College
Running an UG Course

Neil Kindlon, JAX, UNIX scripting
Parveen Kumar, JAX, Pathways
Tim Reynolds, JAX, Gene Weaver
Jason Bubier, JAX, Gene Weaver

Speakers

James Taylor
Johns Hopkins
Galaxy

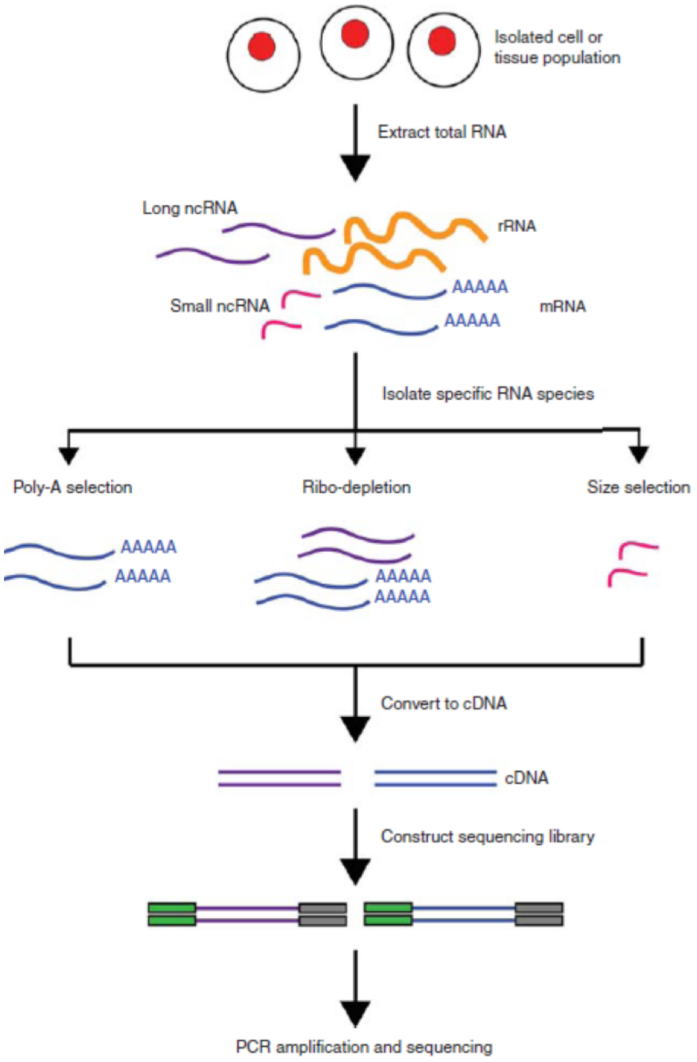
Paola Vera-Licona
UCONN-Health
Network Modeling

Shallee Page
Franklin Pierce University
Running an UG course

Outline Day 1

Monday

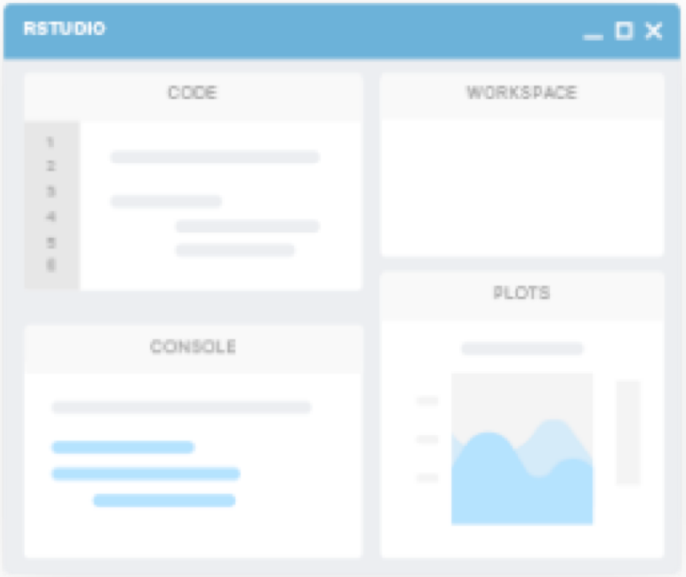
8:30am **Welcome and Introduction**
9:00am **Introduction to Data Science with Biological Emphasis**
Jeffrey Chuang, Ph.D., The Jackson Laboratory for Genomic Medicine
10:00am **Intro to High-Throughput Sequencing Technology**
Charlie Wray, Ph.D., The Jackson Laboratory
11:00am **Hands on Work: Setting up computing infrastructure.** R, Python, Slack and other common Tools,
12:00pm **Lunch**
1:00pm **Introduction to UNIX**
Neil Kindlon, The Jackson Laboratory for Genomic Medicine
2:30pm **Introduction to Statistics**
Jeffrey Chuang
3:30pm **Break**
4:00pm **Curricular Discussion –** Presentation and Discussion of UG courses in Genomics, Reinhard Laubenbacher, Ph.D., JAX-GM / UConn Health; Charlie Wray, Ph.D., The Jackson Laboratory
5:00pm **RNA-seq**
Sheng Li, Ph.D., The Jackson Laboratory for Genomic Medicine
6:00pm **Dinner**
7:00pm **Evening Lecture**
Christine Beck, Ph.D., The Jackson Laboratory for Genomic Medicine



Outline Day 2

Tuesday

8:30am **RNA-seq introduction**
Ada Zhan, Ph.D., JAX-GM
9:00am **RNA-seq module 1**
Ada Zhan
10:00am **Break**
10:30am **RNA-seq module 1, continued**
12:00pm **Lunch**
1:00pm **Module 1 debrief**
1:30pm **Gene Set Enrichment and Pathway Analysis**
Parveen Kumar, Ph.D., JAX-GM
2:30pm **Gene Weaver – Advanced / Derived RNA-seq analysis**
Tim Reynolds and Jason Bubier, PhD,The Jackson Laboratory
4:00pm **Break**
4:30pm **Mutation and Functional Analysis in R. Example lesson.**
Jeffrey Chuang, Ph.D., The Jackson Laboratory for Genomic Medicine
5:30 **End. Dinner on your own.**



Outline Day 3

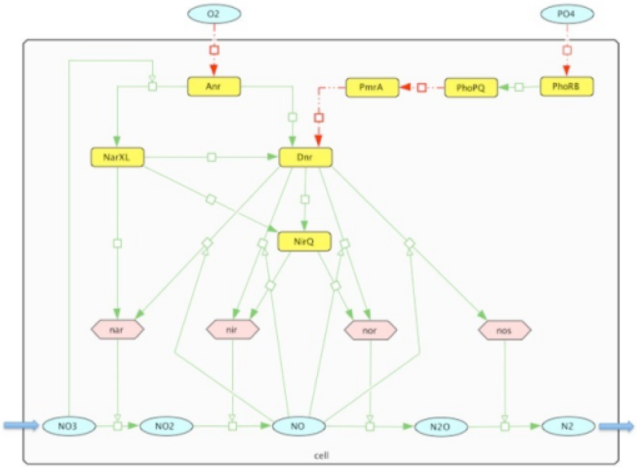
Wednesday

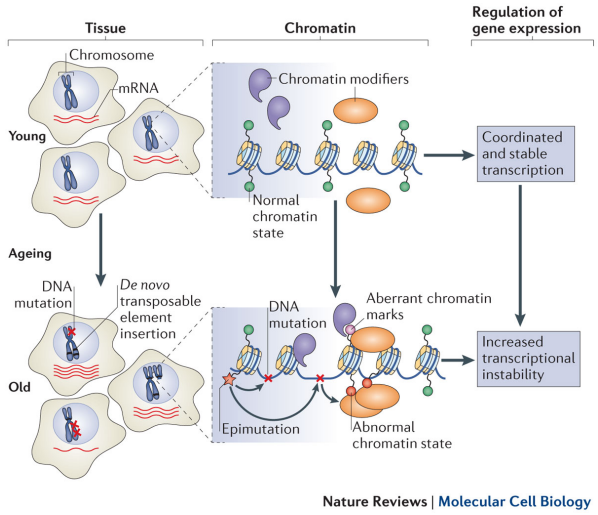
8:30am **Debrief on Module #2 Variant Calling:** Lessons learned, Q & A.
9:00am **Exome Sequencing for Variant Discovery**
Laura Reinholdt, Ph.D., The Jackson Laboratory
10:00m **Module #3: Mouse Exome Variant Discovery.**
Charles Wray, Ph.D. The Jackson Laboratory
12:00pm **Lunch**
1:00pm **Introduction to the Microbiome**
Spencer Glantz, Ph.D., JAX-GM
2:00pm **Hands on Microbiome Work**
Spencer Glantz
3:30pm **Break**
4:00pm **Curricular Discussion #2**
4:30pm **Cloud Resources and Galaxy**
5:30pm **End. Dinner on your own.**

Outline Day 4

Thursday

8:30am **Your Data in Context: Work with Comparative Functional Genomics Information**
Judith Blake, Ph.D., The Jackson Laboratory
9:30am **Network Modeling**
Paola Vera-Licona, Ph.D., UConn Health
10:30am **Network Modeling short exercises**
Reinhard Laubenbacher, Ph.D., The Jackson Laboratory for Genomic Medicine / UConn Health
12:00pm **Lunch**
1:00pm **Data Standards and Best Practices Data Problems: When bad things happen in genomic analysis**
Judith Blake, Ph.D., The Jackson Laboratory
2:00pm **Epigenomics. ChIP-seq module.**
Ada Zhan, Ph.D., The Jackson Laboratory for Genomic Medicine
Shell scripts and online methods.
4:00pm **Break**
4:30pm **Module debrief / Discussion of Participants' Data**
6:00pm **Group Dinner at Butchers and Bakers**





Nature Reviews | Molecular Cell Biology

Outline Day 5

Friday

8:30am Integrated Modeling of Genetic and Genomic Data
Gregory Carter, Ph.D., The Jackson Laboratory
9:30am Running UG Course in Genomics -
Michael Linderman, Ph.D., Middlebury College
Andrea Tilden, Ph.D., Colby College
Shallee Page, Ph.D., Franklin Pierce Univ.
11:00am Cloud environments and web service grants
Sandeep Namburi and Aditya Srikanth Kovuri, The Jackson Laboratory - Information Technology
12:00pm Lunch
1:00pm Collaborative Discussion, Next Steps, and How to Support UG Implementations
Charlie Wray, Jeff Chuang, Reinhard Laubenbacher

pcmag.com

Goals

- Teaching genomics material to participants
- In-class discussions of what approaches would be optimal for different institutes
- Discussion of computational biology core competencies for recommendation to NIH

Intro to Biological Data Science

Sequencing costs over time

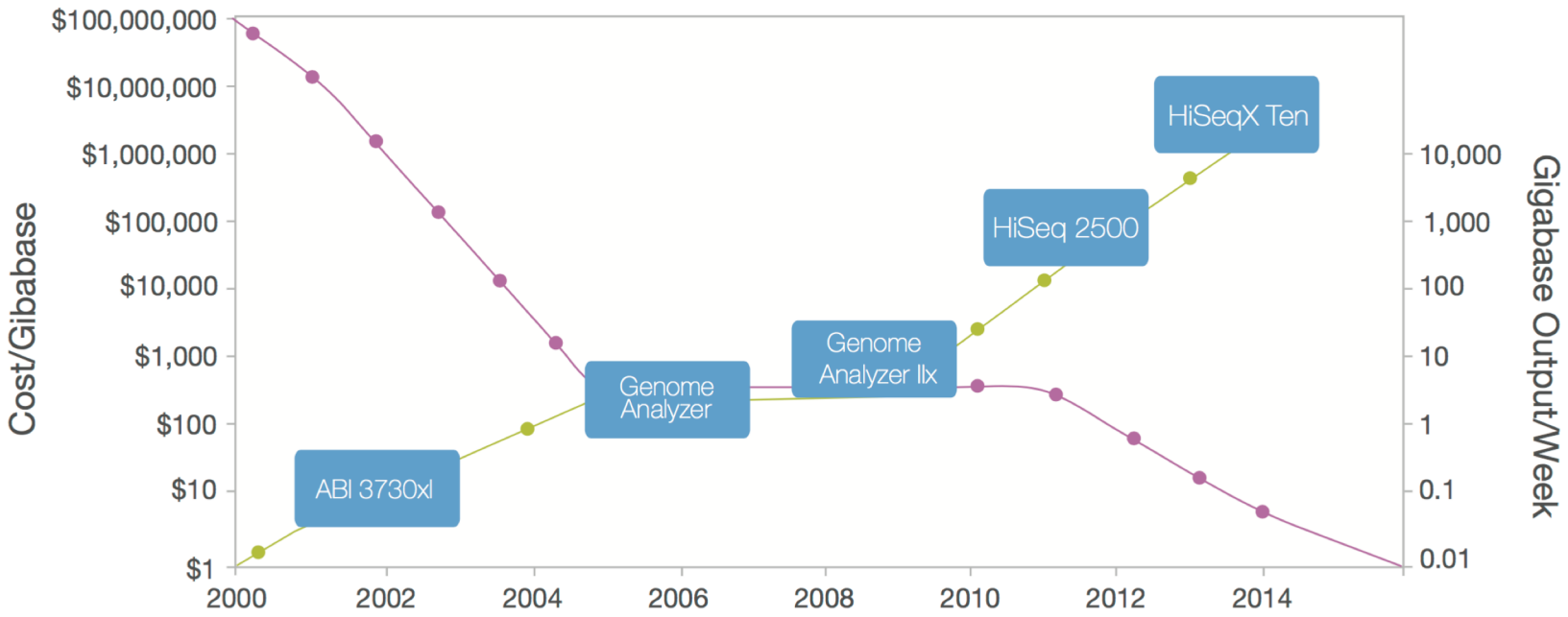
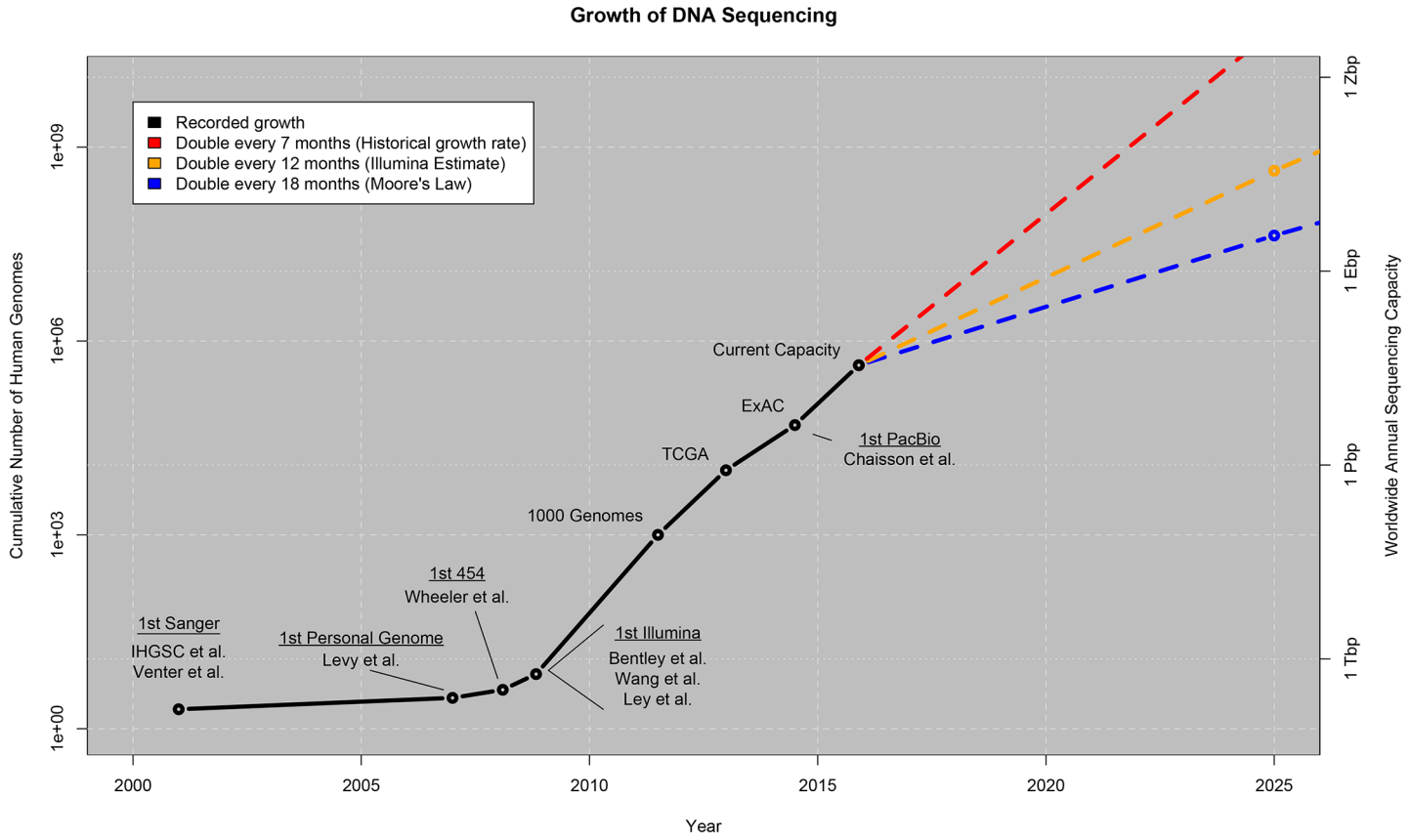


Figure 1: Sequencing Cost and Data Output Since 2000—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

Sequencing costs have fallen dramatically. It now costs ~$1000 to sequence a human genome. Source: Illumina

The scale of biological data



Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

Genomics is among the biggest of big data

10^{21}

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015).
http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

www.vla.nrao.edu

Illumina

Storage at genomics centers

Institute	Num. of Sequencers (omicsmap.com)	Storage (PB)	Citation
BGI (formerly Beijing Genomics Institute)	166	9	[1]
Broad Institute	101	10	[2]
The Genome Center at Washington University	38	10	[3]
Wellcome Trust Sanger Institute	38	22	[4]
Human Genome Sequencing Centre, Baylor College of Medicine	32	3.2	[5]
Macrogen	27	4	[6]
NY Genome Center	27	1	[7]
McGill University and Génome Québec Innovation Centre	22	5	[8]
Yale Center for Genome Analysis	20	2	[9]
DOE Joint Genome Institute	15	2	[10]
Beijing Institute of Genomics	15	1	[11]
CSHL	15	3	[12]
Ontario Institute for Cancer Research	14	3.5	[13]
Canada's Michael Smith Genome Sciences Centre	13	7	[14]
Centro Nacional de Análisis Genómico (CNAG)	12	2	[15]
UCSF	2	7	[16]
St Jude	8	2	[17]
UCSC / CGHUB	6	5	[18]
UMD-IGS	5	1	[19]
EBI	0	1.2	[20]
Sum	***576***	***100.9***	

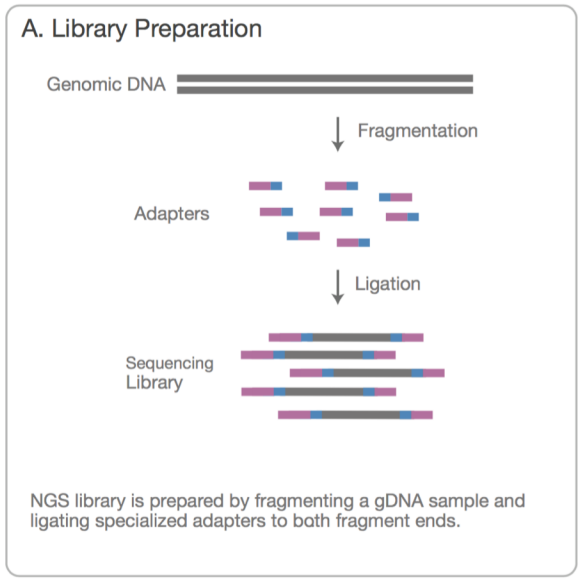
Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. (2015) Big Data: Astronomical or Genomical?. PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195
http://journals.plos.org/plosbiology/article?id=info:doi/10.1371/journal.pbio.1002195

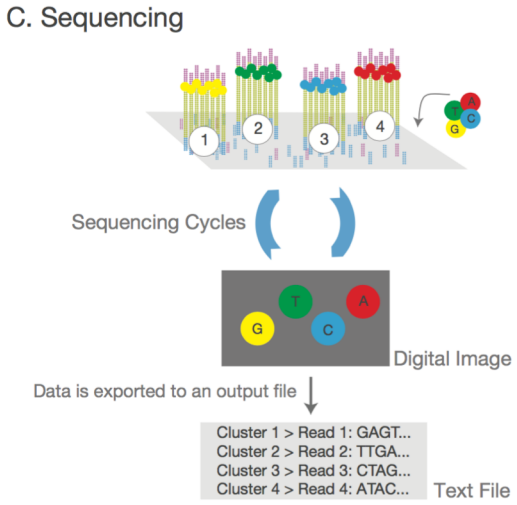
Big biology data require data science training

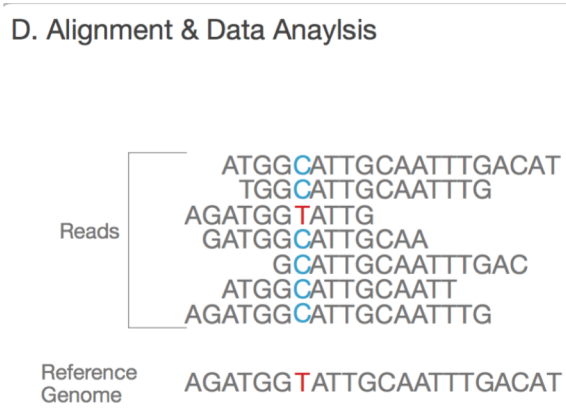
- Data wrangling (raw manipulations such as obtaining, parsing, editing, merging, and filtering data)
- Computational problem solving (making a pipeline run, processing and storage issues, compatibility)
- Statistical tests (clustering, pathway identification, regression)
- Software engineering (optimization, software robustness, visualization)
- Development of new algorithms.

All of these lead up to biological interpretation and generation of new hypotheses.

Sequencing Technology Overview

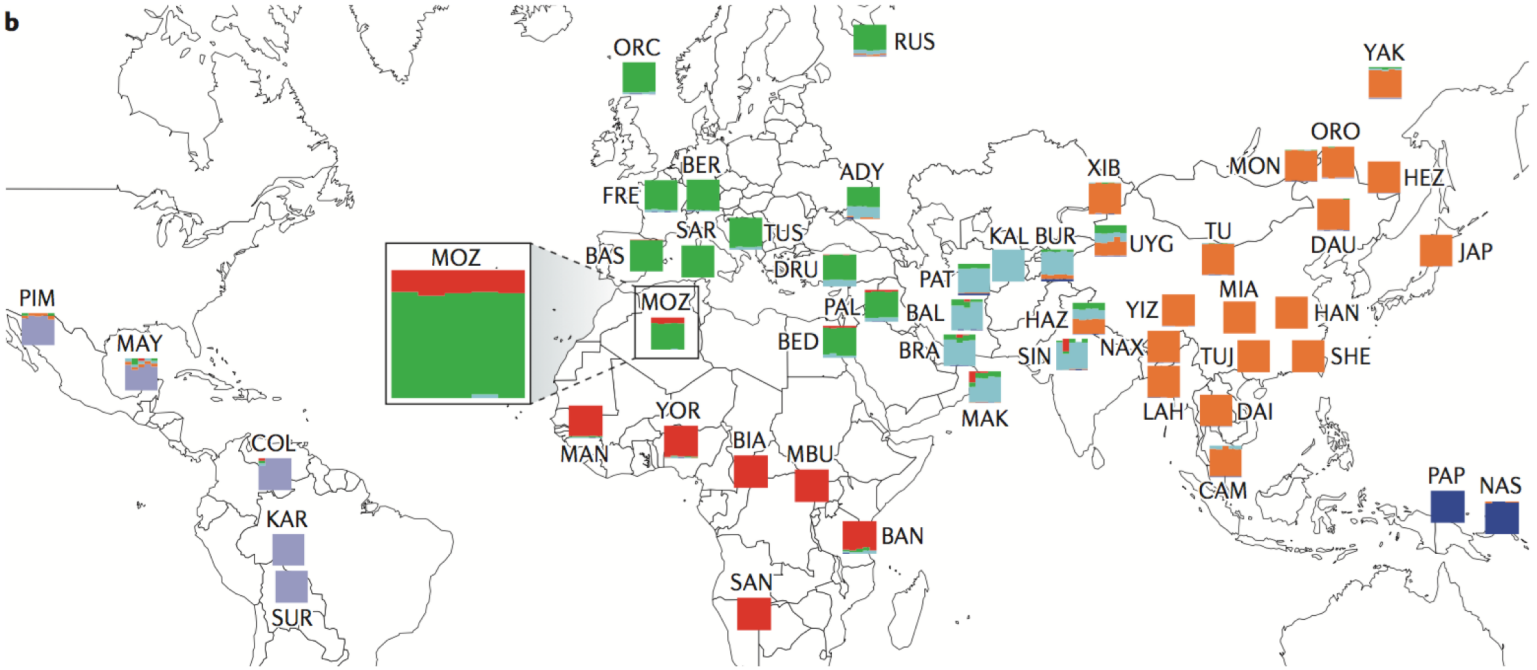






Illumina is currently the dominant sequencing technology company, though other technologies are now being developed. Adapted from Illumina materials.

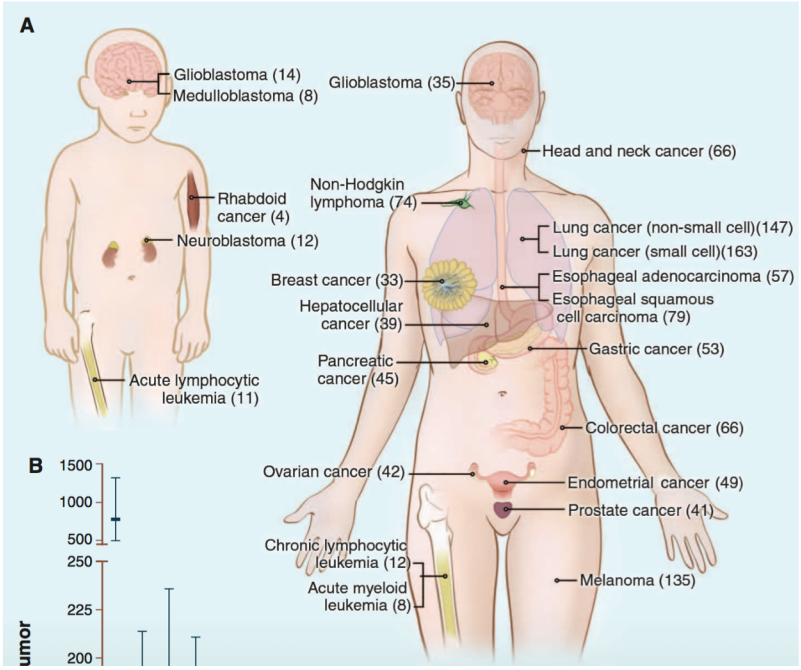
Human sequencing is the major driver of the genomics data explosion



Identification of population groups by their sequence similarity.

Nature Reviews Genetics. 2011. 12:603

Sequencing in medicine: cancer



Drug	Disease	DNA mutation	Action
Imatinib, Dasatinib, Nilotinib, Bosutinib	Chronic myelogenous leukemia	*BCR-ABL1* fusion	Indication for therapy
Ponatinib	Chronic myelogenous leukemia	*BCR-ABL1* fusion	Only indicated for T315I mutations
		T315I resistance mutation	
Erlotinib, Afatinib	Lung adenocarcinoma	*EGFR*	Indication for therapy
		Exon 19 deletions	
		L858R	
Vemurafenib, Dabrafenib	Melanoma	*BRAF* V600E	Indication for therapy
Tramatenib	Melanoma	*BRAF* V600E/K	Indication for therapy
Crizotinib	Lung cancer	*ALK* gene fusions	Indication for therapy
Cetuximab	Colon cancer	*KRAS* codon 12, 13	Contraindication to therapy
Olaparib	Ovarian cancer	*BRCA1* and *BRCA2* mutations	Indication for therapy

Average number of nonsynonymous mutations in different cancers. *Science* 29 Mar 2013: Vol. 339, Issue 6127, pp. 1546-1558

.

Jeffrey Gagan and Eliezer M. Van Allen. Genome Medicine 2015 7:80

More than 100,000 human exomes are now available

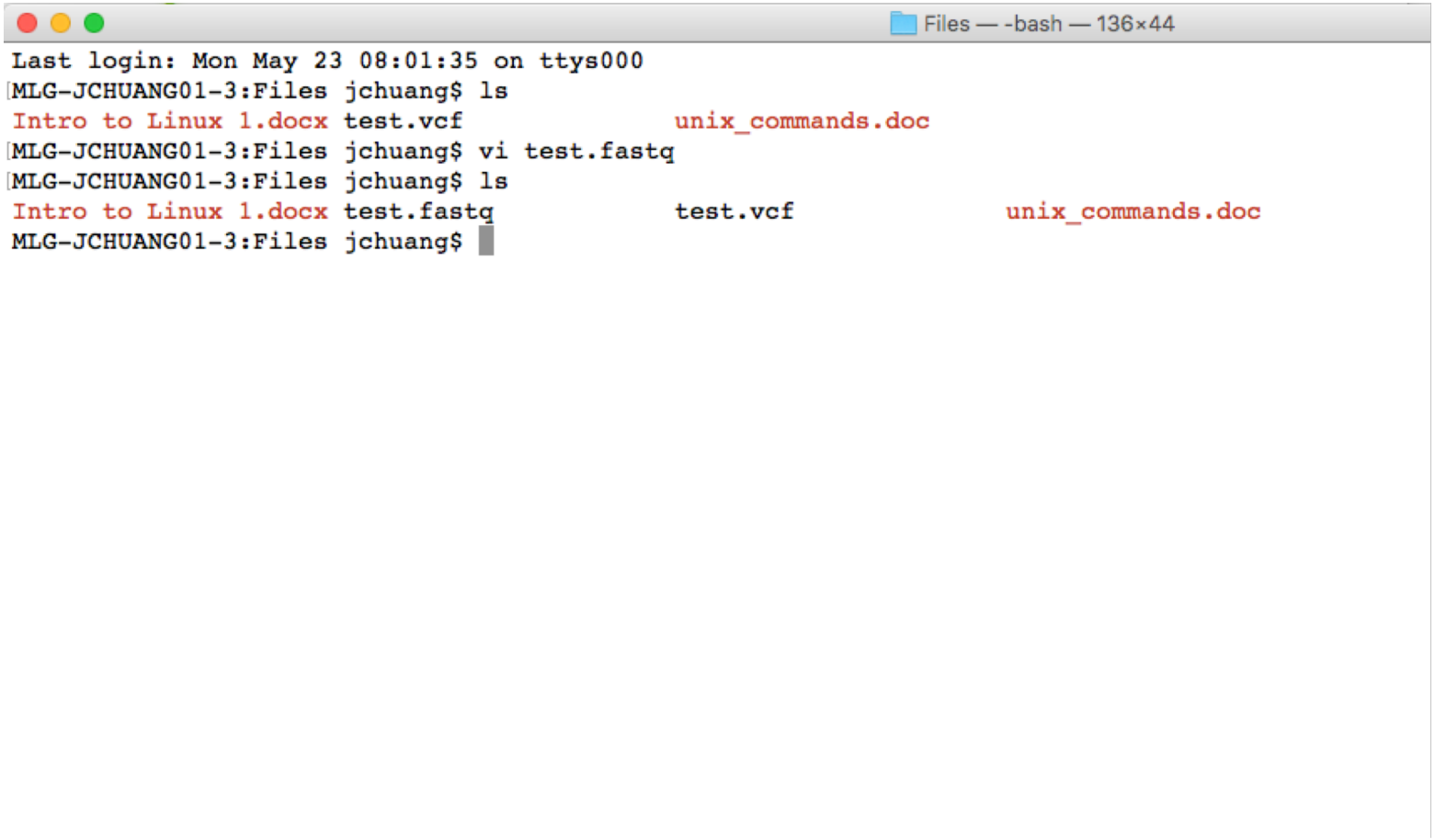
About gnomAD

The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed here.

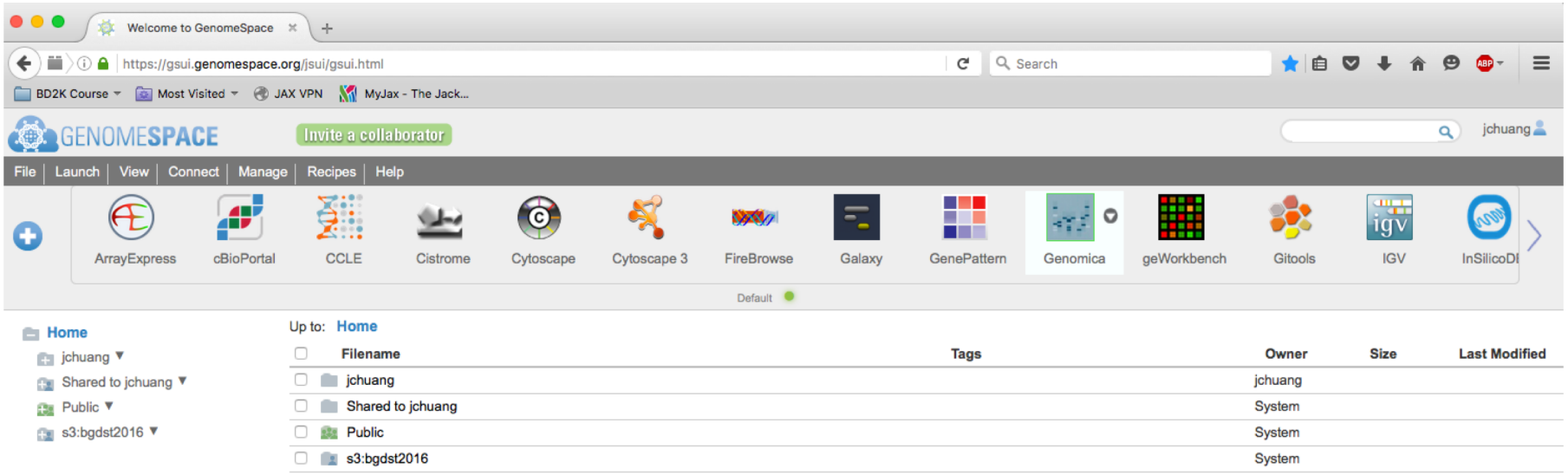
http://gnomad.broadinstitute.org/about

Platforms: command line



Platforms: local packages

Platforms: cloud



Progress in computational biology topics

10 years ago introductory bioinformatics was focused on methods for small numbers of genomes and some resources.
Sequence alignment; NCBI data resources; phylogenetics; mutation calling; differential gene expression analysis; etc.

Current introductory bioinformatics is geared toward mining large datasets and applying tools to easily generated new data.
Pipelines for processing your own sequencing data ; applying existing packages for data analysis; data mining on tens to thousands of genomes; data integration.

Challenges in Computational Biology Education

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Developing a Short Bioinformatics Training Course

Allegra Via[1], Javier De Las Rivas[2], Teresa K. Attwood[3], David Landsman[4], Michelle D. Brazas[5], Jack A. M. Leunissen[6], Anna Tramontano[1], Maria Victoria Schneider[7]*

- Breadth of field makes defining goals critical
- Specialized computational resource needs
- Mixing education and Interactivity
- Necessary computational expertise is project dependent.

Via A, De Las Rivas J, Attwood TK, Landsman D, Brazas MD, et al. (2011) Ten Simple Rules for Developing a Short Bioinformatics Training Course. PLoS Comput Biol 7(10): e1002245. doi:10.1371/ journal.pcbi.1002245

Opportunities

- Although computational biology resources are specialized, many are free or inexpensive.
- Much data is publicly available providing abundant opportunities for new projects. Many new interfacing tools are being developed.
- Rapidly advancing field provides room for a variety of approaches to pedagogy.
- Well-trained students have immediate job opportunities.

Example Paradigm 1: Biologically Driven Approach

- Goals: Teach methods that match a biology curriculum.
- Example topics: expression analysis (RNA-seq data processing), mutations (Exome-seq data processing), heredity (phylogenetic methods) etc.
- Challenges: Difficult to match complexity of computational approaches to biology topics. Requires easy-to-use analysis platform.

Example Paradigm 2: Computational Skills Approach

- Goals: Development of skill in applying computational analysis techniques,
- Example topics: Unix, python, R, data wrangling, web resources, common sequence analysis software, cloud implementations.
- Challenges: Culture difference between programming and biology. Wide variety of choices in what skills to teach.

Example Paradigm 3: Data science approach

- Goal: provide theoretical foundations for biology as a data science.
- Example topics: clustering, regression, other data mining approaches, phylogenetics, evolution
- Challenges: Difficult to identify sufficient students with both mathematical and biological interest.