Bioinformatics Workshop

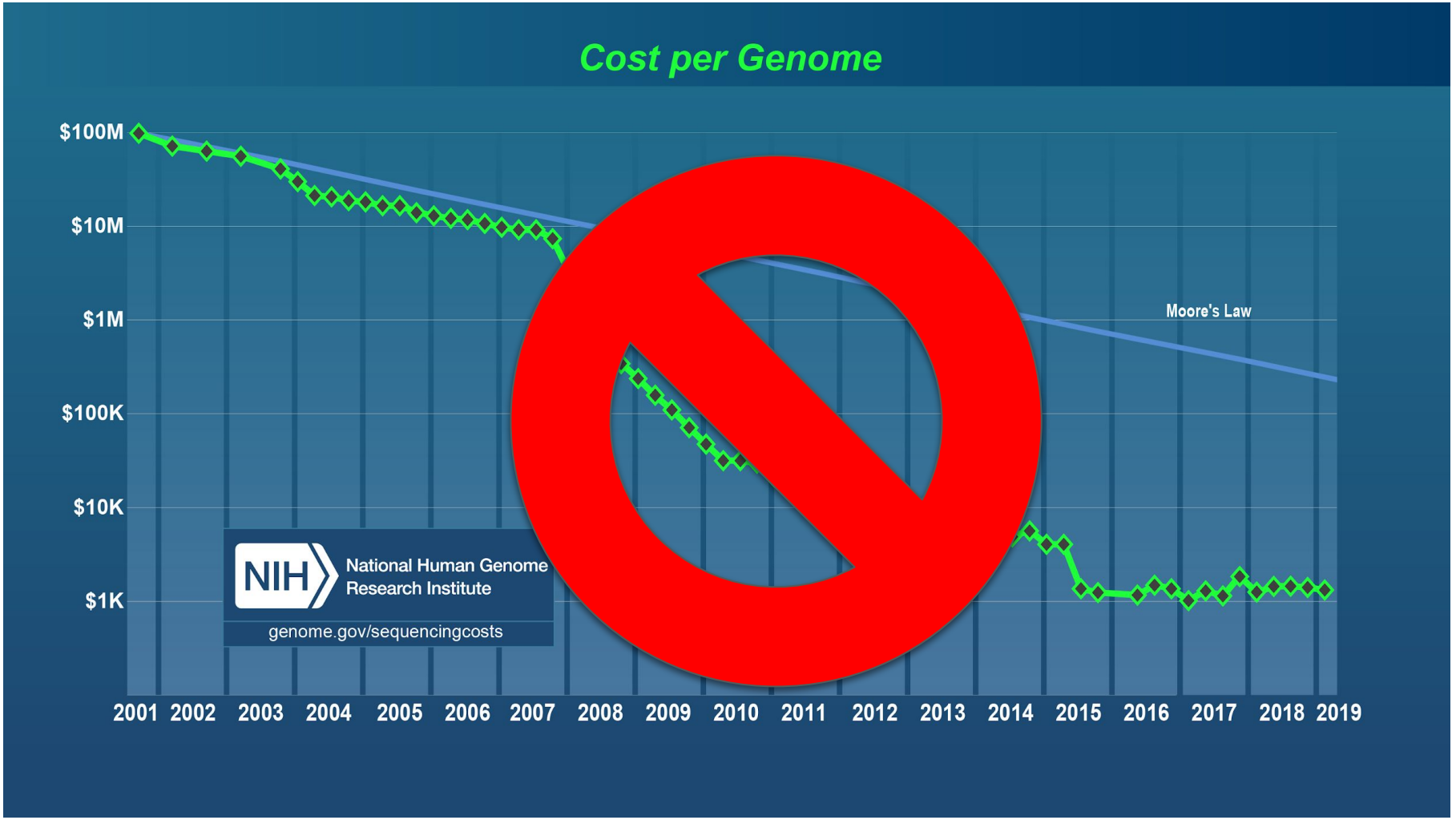
Session #1

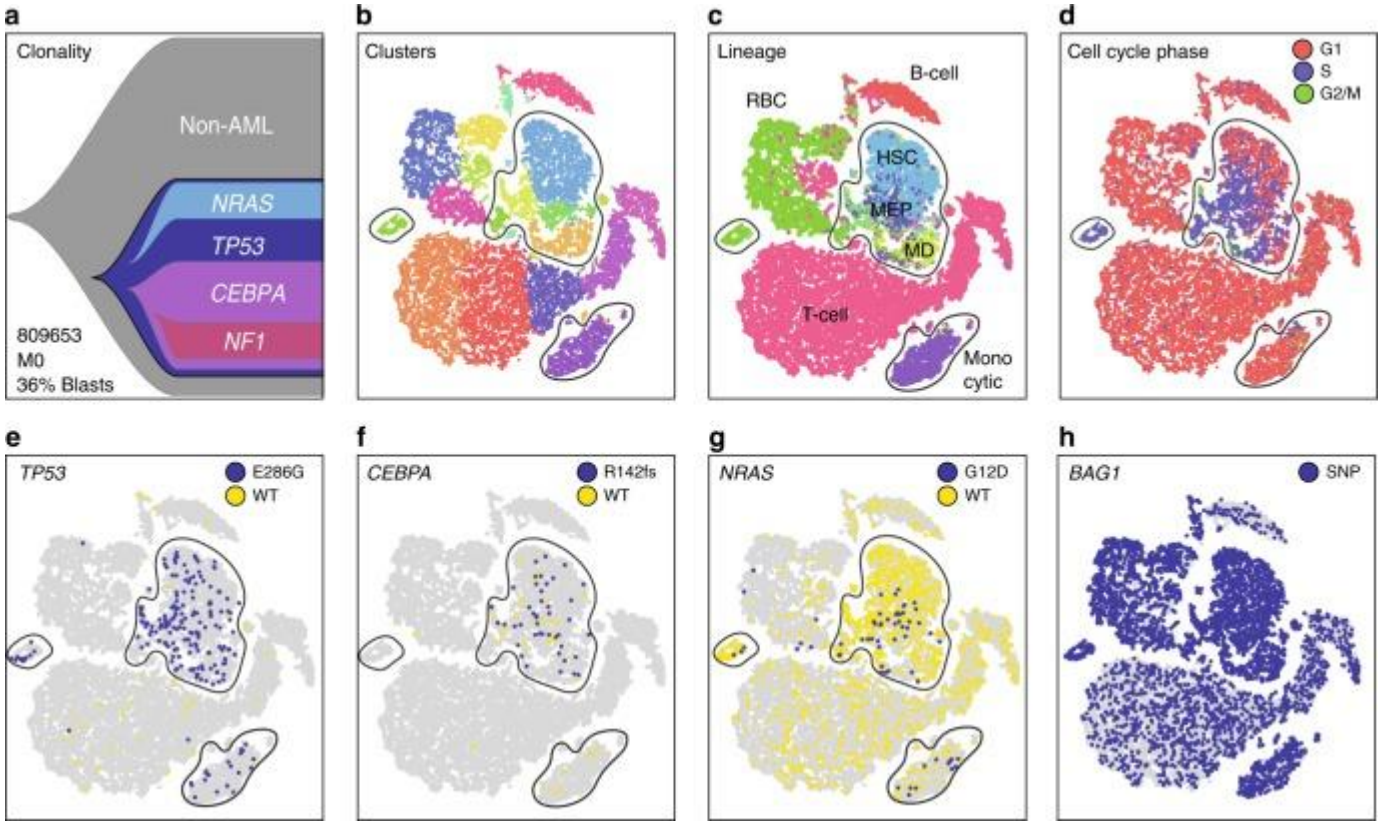
Course Introduction and Prerequisites

Chris Miller and Jason Walker

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019


a
Clonality
Non-AML
NRAS
TP53
CEBPA
NF1
809653
M0
36% Blasts
b
Clusters
c
Lineage
B-cell
RBC
HSC
MEP
MD
T-cell
Mono
cytic
d
Cell cycle phase
G1
S
G2/M
e
TP53
E286G
WT
f
CEBPA
R142fs
WT
g
NRAS
G12D
WT
h
BAG1
SNP

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics
- Skills in programming, statistics, and visualization help you get the most out of your data

People who need complex data analysis

People who know how to do complex data analysis

https://hellogiggles.com/news/how-many-people-attend-coachella/
https://www.nytimes.com/2020/07/02/theater/germany-theater-coronavirus.html

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics
- Skills in programming, statistics, and visualization help you get the most out of your data
- This course aims to teach you the theory and practice of computational biology, with a focus on genomics but lessons that apply broadly

Goals:

- To empower you to improve and expedite your research
- To expose you to new ideas and techniques that may advance your research program

Who we are, and why you should trust us

Jason Walker, M.S.

Informatics group leader,
McDonnell Genome Institute

Chris Miller, Ph.D.

Assistant Professor
Div of Oncology

Over 30 years of combined experience in Bioinformatics and Computational Biology

Other Lecturers/Organizers include:

Malachi Griffith
Zach Skidmore

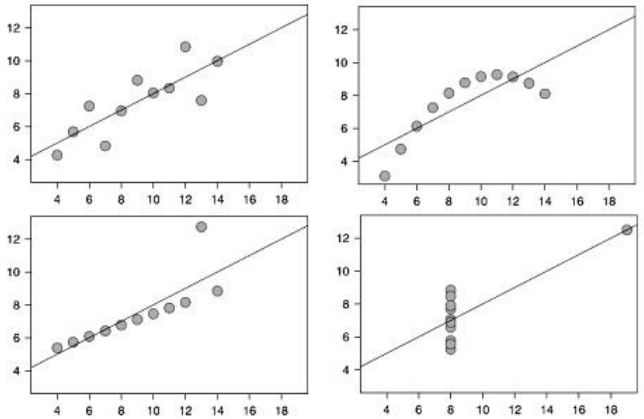
Obi Griffith
Todd Wylie

Allegra Petti
Many others!

Don’t trust your data

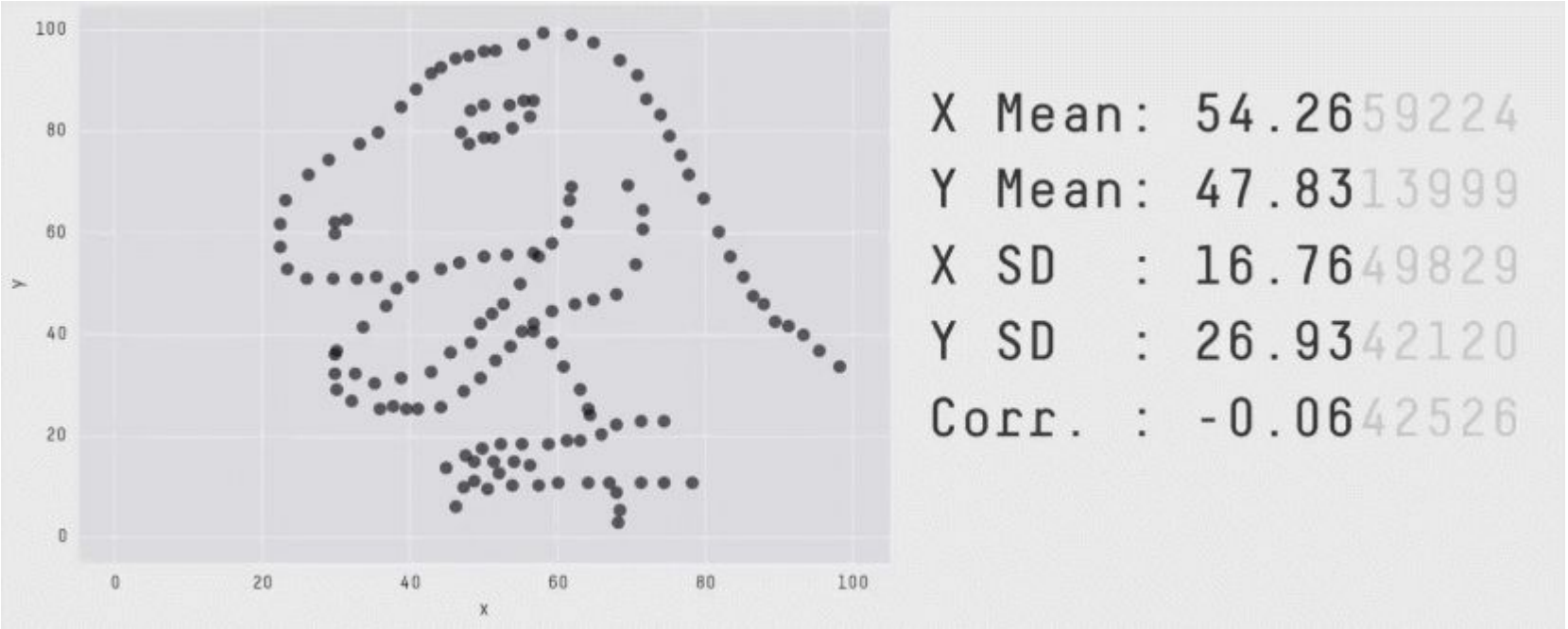
Trusting your data

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : σ^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : σ^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places



https://commons.wikimedia.org/wiki/File:Anscombe%27s_quartet_3_cropped.jpg

Datasaurus Dozen



http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

“Analyzing your data means inherently distrusting your data until you have exhausted yourself into giving up and trusting it.”

-Aaron Quinlan

Course structure

- Pair an introduction to a biological or technical concept with some of the tools needed to analyze it

- Next week:
  - Sequence data generation
  - How to read, manipulate, and run quality control on sequence data

Prerequisites

- You do not need to know all of these things on Day 1
  - Tutorials and documentation in today's notebook

- But you do need to get moving!
  - Hands-on learning is *essential*. If you can't follow along and do the assignments, you will not get the most out of this course

Cluster computing

- Large compute cluster
  - Dozens of servers
  - Thousands of cores
  - Terabytes of RAM

- How do we give everyone access efficiently and fairly?

LSF - Load Sharing Facility

- Binds different machines together into one common pool
- Jobs can be run on any machine

- Need to specify job requirements
  - How many CPUs
  - How much RAM
  - How much tmp (sometimes)
  - What kind of environment to run in

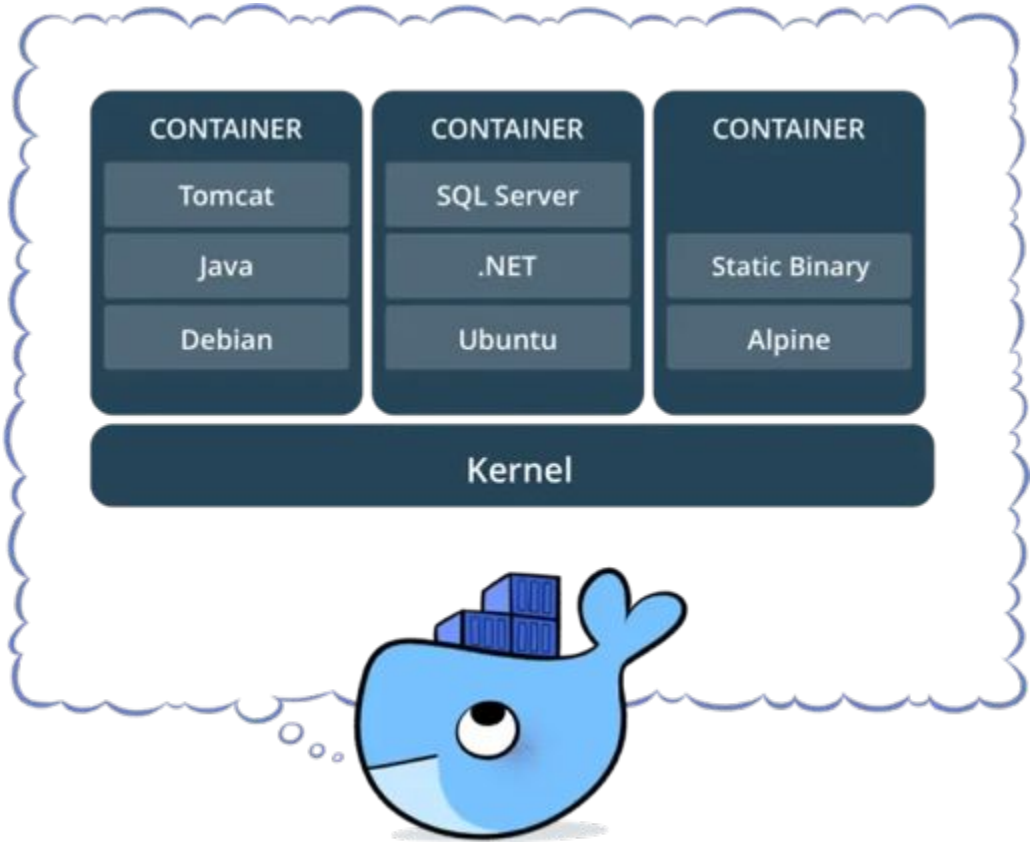
Computing Environments

- Laptop
  - You administer
  - You control completely

Computing Environments

- Laptop
  - You administer
  - You control completely
- Shared compute cluster
  - A sysadmin or group administers it
  - You control very little

Docker containers



Computing Environments

- Laptop
  - You administer
  - You control completely
- Shared compute cluster
  - A sysadmin or group administers it
  - You control very little
- Docker (containers)
  - Sysadmins handle the hardware
  - You control the software almost completely

Finding docker images

- Search engines
  - “docker bedtools”
- Repositories - Bioconda/Quay.io/Dockerhub
  - “docker mosdepth quay”
- Slack - ask around
- Building your own

Homework

- Will not be turned in or graded

- Will be useful for understanding subsequent lectures