Introduction to Bioinformatics

Chris Miller, Ph.D.
Washington University in St Louis

Bioinformatics Workshop 2023-2024

Supported by – ICTS Precision Health

- We aim to catalyze genomic research by providing grant review, development services, guidance and resources for genomic researchers and genomics education in the community.

Cite the **NIH CTSA Grant #UL1 TR002345** when research is supported by ICTS/CTSA funding or any ICTS Core Services

BFX Workshop – contact Jenny if you haven't received the following

- Slack access, welcome email, Outlook bfx-workshop-2023 group invite

Register for BFX

https://redcap.link/BFX2023

Institute of Clinical and Translational Sciences

Washington University in St. Louis

icts-precisionhealth.wustl.edu

j.mckenzie@wustl.edu

ICTS Precision Health

Visit

Support Transdisciplinary Research

- Match clinicians and investigators
- Review grants
- Fund Precision Health Innovation awards

Develop Common Workflows

- Genomic consent and return of results
- Enable access to large genomic data sets
- Develop infrastructure to speed discovery translation

Educate the Community

- Expand access to educational research programs
- Educate scientific and clinical community in precision health
- Engage broader STL community in genomic research

Leadership Team

Megan Cooper, MD, PhD

Chris Gurnett, MD, PhD

Institute of Clinical and Translational Sciences

Washington University in St. Louis

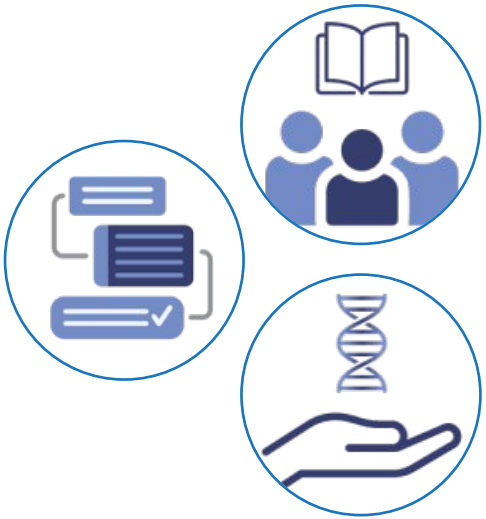
icts-precisionhealth.wustl.edu

j.mckenzie@wustl.edu

Precision Health Led Projects

icts-precisionhealth.wustl.edu
j.mckenzie@wustl.edu

- **Pilot funding & Research reviews**
  - Precision Health Innovation Awards; ICTS Research Development Program
- **Return of Results (ROR) for Research Participants**
  - Genetic counseling, process for returning ACMG secondary results
- **Genomic Database Access and Submission**
  - UK Biobank, All of Us Research Program, dbGaP, AnVIL, SRA
  - Assistance to submit human genomic data to shared repository
- **Institutional Genomic Consent**
  - One Protocol One Consent, BJC-Webb electronic biobank
- **Community Education & Engagement**
  - Precision Health for the Ages Workshop Series



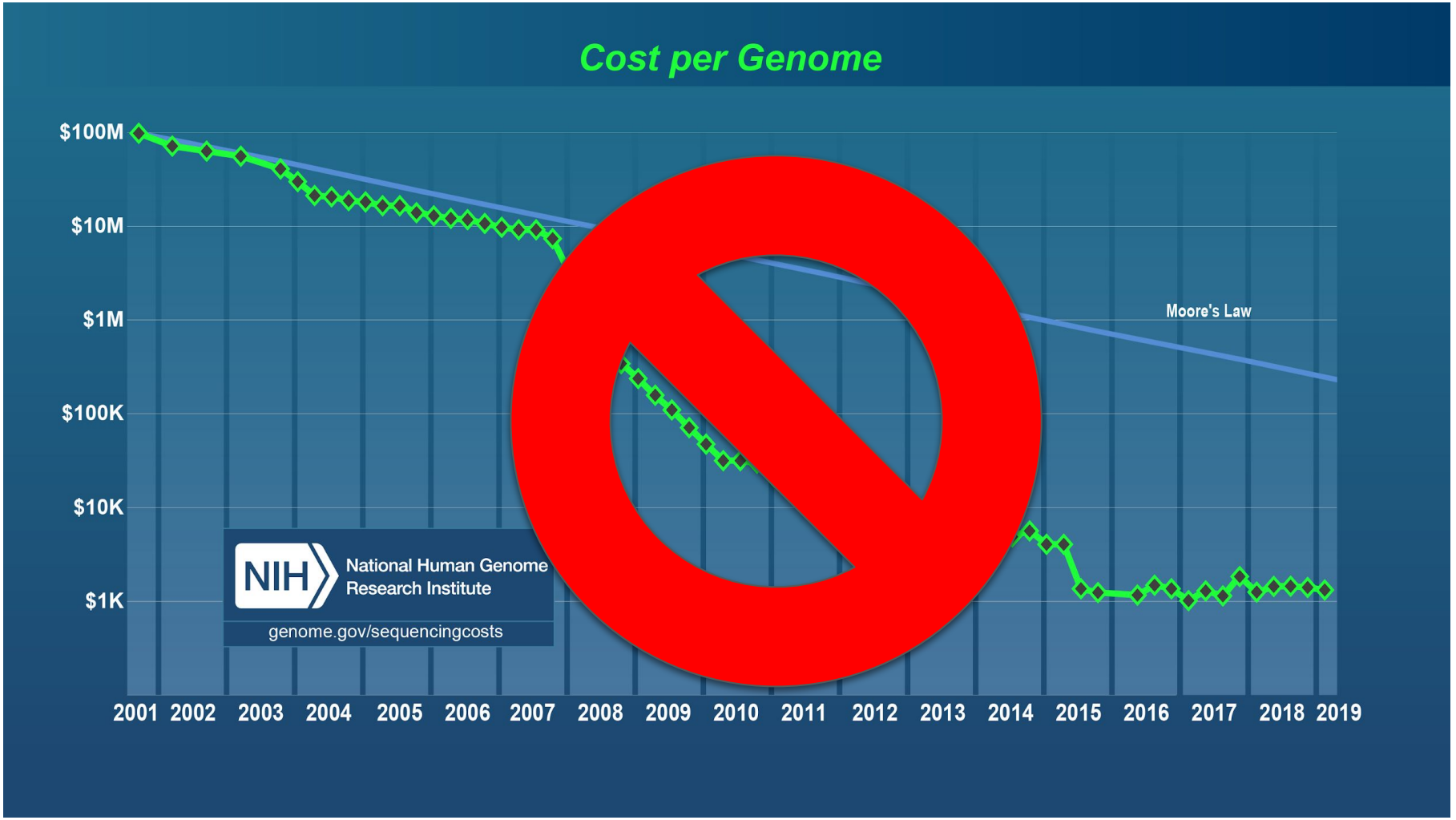
Providing Support For

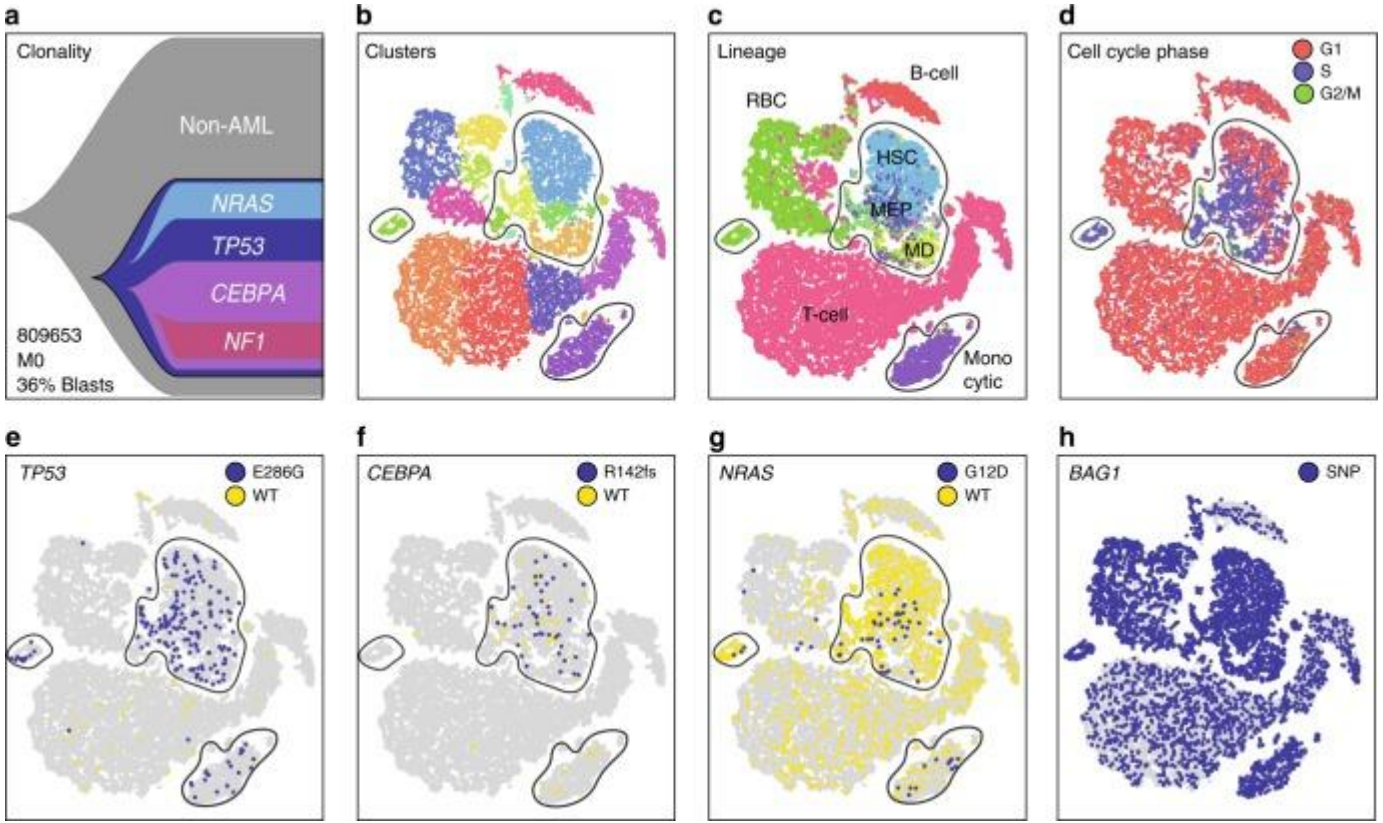
- **Core Services**
  - WU Biological Therapy Core Facility (BTCF), McDonnell Genome Institute (MGI)
- **Informatics Tools for Precision Health**
  - Bioinformatics Workshop (BFX), pVAC, CIViC,
- **Communications and Outreach**
  - Women in Innovation and Technologies (WIT) program, EQUALIZE program through OTM
- **Educational Opportunities**
  - Precision medicine pathway, Bioinformatics Workshop (BFX), Genomics in Medicine

Institute of Clinical and Translational Sciences

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?

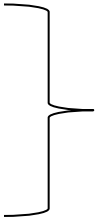
- Biology is now a quantitative discipline - especially genomics
- Skills in programming, statistics, and visualization help you get the most out of your data
- This course aims to teach you the theory and practice of computational biology, with a focus on genomics but lessons that apply broadly

What is bioinformatics?

What is bioinformatics?

Common skills

- Statistics
- Programming
- Visualization



“Data science”

- Deep understanding of the biological system and experiments

What is bioinformatics?

- Application of computational techniques to biological data
- Covers a lot of ground!

- Population genetics
- Cancer genomics
- Microbial genomics
- Proteomics
- Ecology/Evolution
- Medical informatics/EHR mining
- computational behavioral biology
- Epidemiology
- Protein folding
- CryoEM or tomography
- Drug design/molecular dynamics
- Algorithmic design/optimization
- Metabolomics
- Mathematical Biology

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

Chris Miller, Ph.D.

Course Director
Assistant Professor
Division of Oncology

Almost 20 years of experience in Bioinformatics and Computational Biology

Other Lecturers/Organizers include:

Jason Walker	Malachi Griffith	Jennifer Foltz
Susanna Kiwala	Aadel Chaudhuri	Brigida Rusconi

Who we are, and why you should trust us

Chris Miller, Ph.D.

Course Director
Assistant Professor
Division of Oncology

Jenny McKenzie, Ph.D.

Course Coordinator
ICTS Precision Health
Program Scientist

Almost 20 years of experience in Bioinformatics and Computational Biology

Other Lecturers/Organizers include:

Jason Walker
Susanna Kiwala

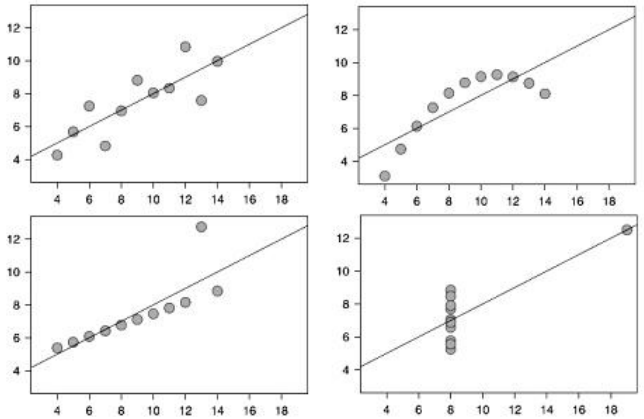
Malachi Griffith
Aadel Chaudhuri

Jennifer Foltz
Brigida Rusconi

Don’t trust your data

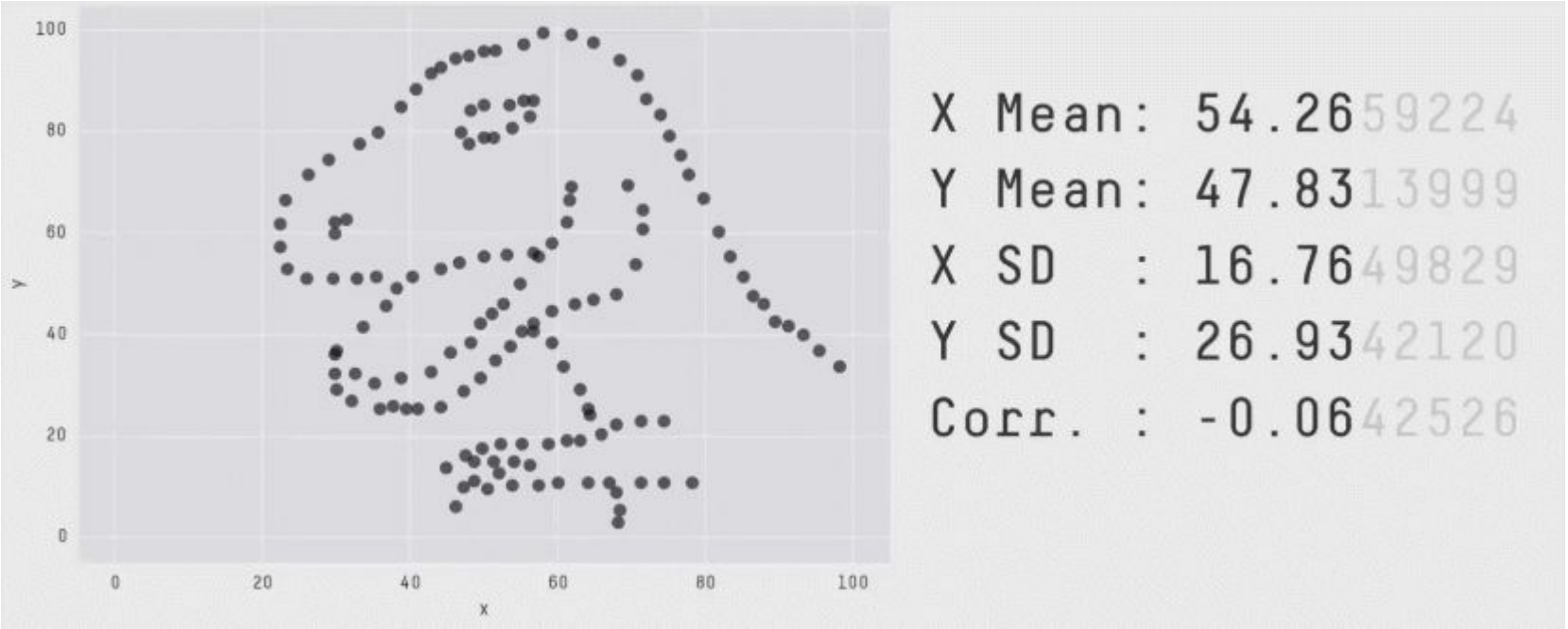
Trusting your data

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

Summary statistics are dangerous

- Visualize your data!
- A picture is worth a thousand p-values

"If your experiment needs statistics, you ought to have done a better experiment"

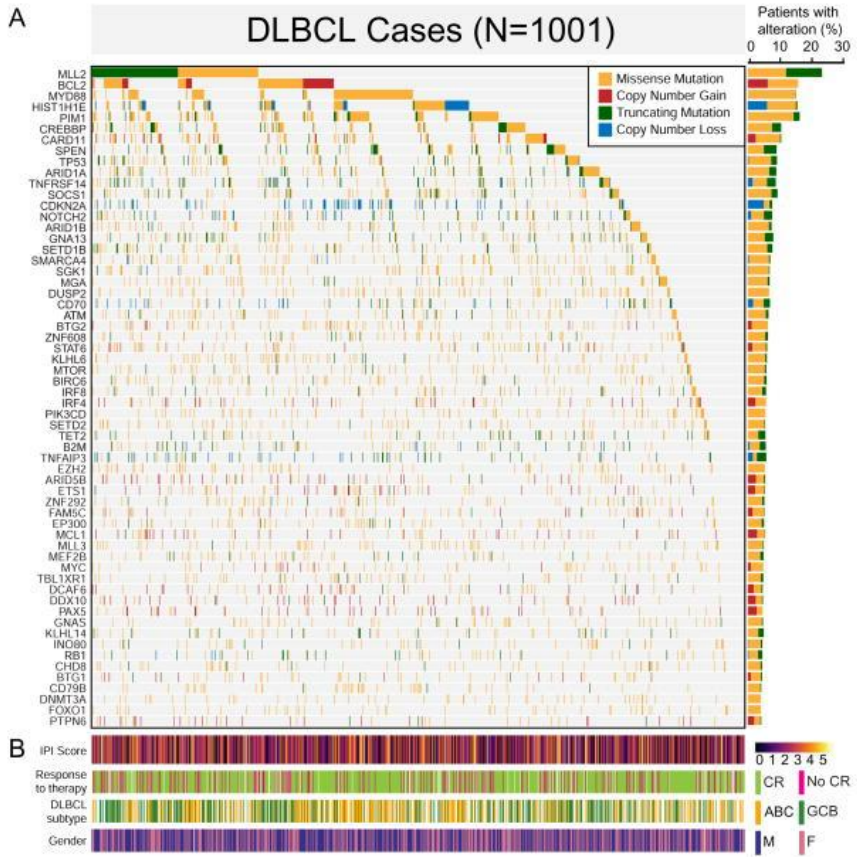
- Ernest Rutherford

I obviously don't subscribe to this, but he's not completely wrong either! If you can't make a plot convincing you that an effect is real, how confident are you, really?

- The bioinformatics core aligned the data and sent me a list of differentially expressed genes. I’m done, right?

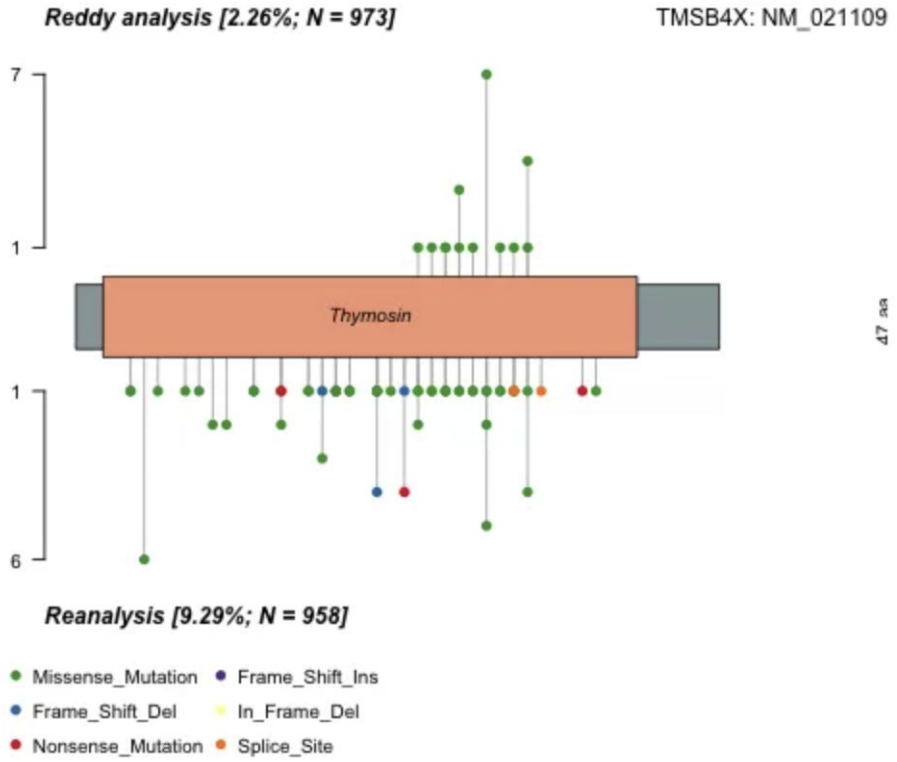
- We ran Mutect to call somatic mutations in this tumor genome. Let’s take it to the bank

Real world consequences



Reddy, et al. 2017 10.1016/j.cell.2017.09.027

Real world consequences



https://pubpeer.com/publications/E61AC72AE0402C6A62A84E36ED2AEA

Real world consequences

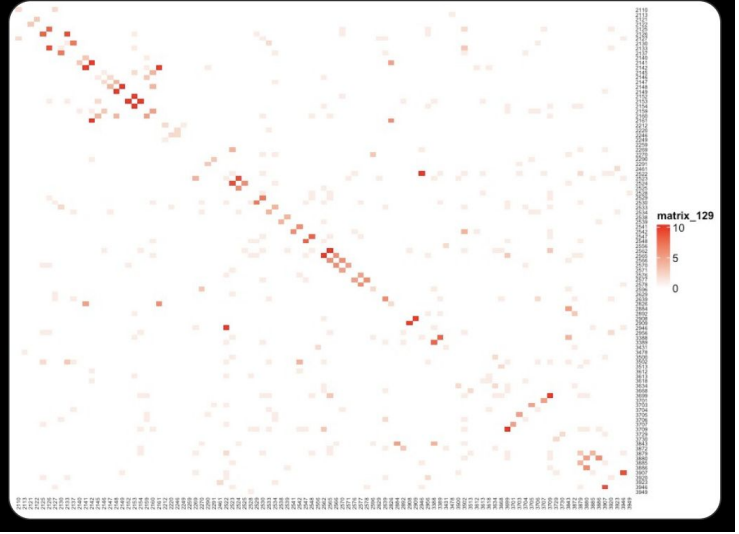
Ryan D Morin @morinryan · Oct 2

RNA/DNA mismatches (sample swaps) affecting at least 10% of the patients in Reddy et al, a Cell paper with over 700 citations. Same issue was described in a more recent paper from this group. #lymphoma #genomics #goodresearchpractice pubpeer.com/publications/E...

1 1 7

Ryan D Morin @morinryan · Oct 2

Sharing of variants between RNA and DNA. Red should be on the diagonal. Most swaps seem to be between adjacent or nearby IDs.

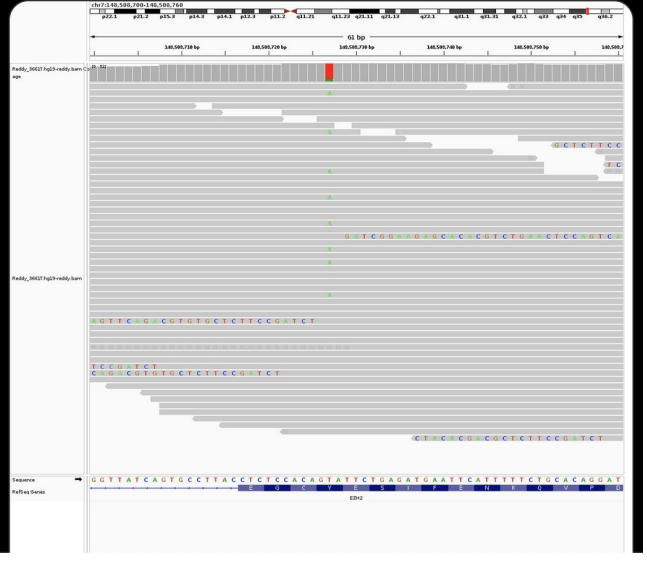


Ryan D Morin @morinryan · Nov 4

There are over 3,600 examples of variants like this, supported by at least 3 somatic variant callers (i.e. by consensus, they're real) and yet Reddy didn't report them. All of these are coding variants in the DLBCL genes described in Reddy but all were absent for some reason.

Ryan D Morin @morinryan · Nov 4

24/33 (this is the limit imposed by Twitter). This is just the first 24. If someone still thinks I'm cherry-picking examples. This is a clinically relevant hot spot that was described 7 years before the Reddy study. Inexcusable to miss this many of them, and yet excuses are made!



Real world consequences

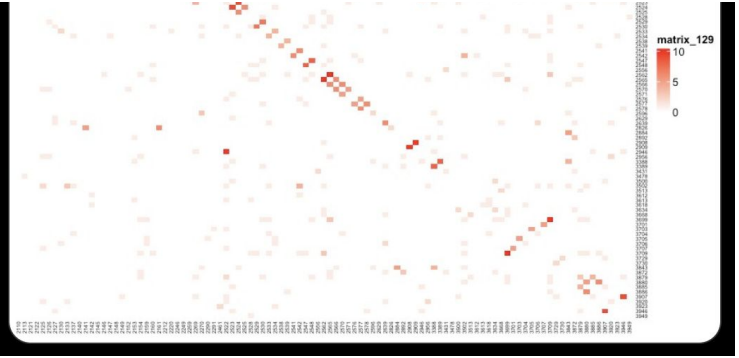
Ryan D Morin @morinryan · Oct 2

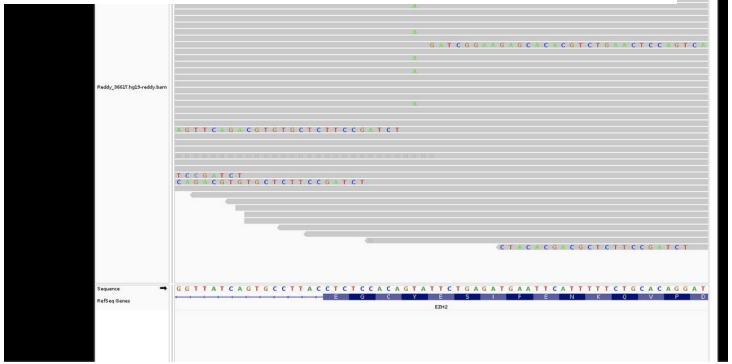
RNA/DNA mismatches (sample swaps) affecting at least 10% of the patients in Reddy et al, a Cell paper with over 700 citations. Same issue was described in a more recent paper from this group. #lymphoma #genomics #goodresearchpractice pubpeer.com/publications/E...

Ryan D Morin @morinryan · Nov 4

There are over 3,600 examples of variants like this, supported by at least 3 somatic variant callers (i.e. by consensus, they're real) and yet Reddy didn't report them. All of these are coding variants in the DLBCL genes described in Reddy but all were absent for some reason.

Although the effects on each conclusion from Panea *et al* has not been evaluated, we demonstrated that ~30% of the reported mutations are not supported by their WGS data, which caused a significant inflation of the mutation prevalence of at least 16 genes and the rate of coding mutations in 9 genes (Supplemental Figure S3). These lead to





Lessons to be learned

- Check and double check and triple check your data and your scripts
- Visualize your data!
- Admit when mistakes are made

Errors

- Will happen!
- Errors of commission vs omission
- Type 1 errors – False positives
- Type 2 errors – False negatives

“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:
  - command line skills, plotting with R

- Following 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

Homework

- Will not be turned in or graded

- Will be useful for understanding subsequent lectures
-
- Posted on the course page

  https://github.com/genome/bfx-workshop

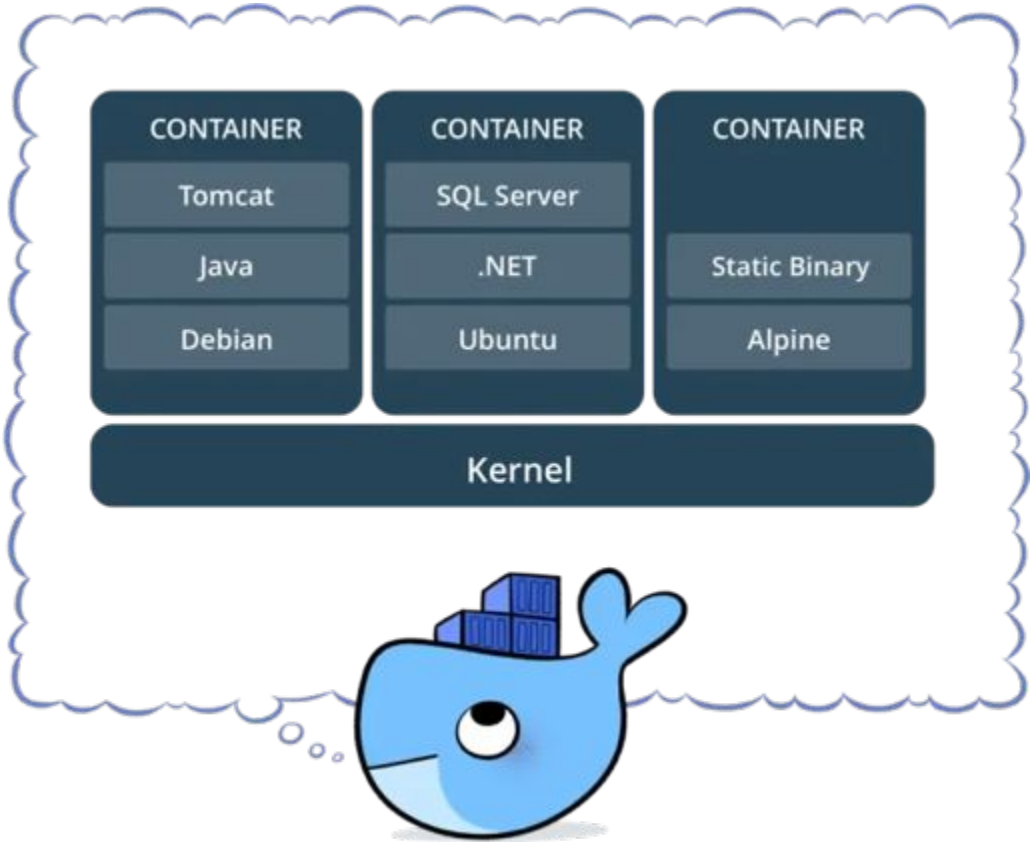
Computing Environments

- 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 in #docker
- Building your own

Turning data into insight