

Introduction to Bioinformatics

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

Bioinformatics Workshop 2025-2026

(aka bfx-workshop)



Applied Bioinformatics for Genomics I

(aka BIOL 5624)

Applied Bioinformatics for Genomics II

Course: BIOL 5624

1 Credit Hour DBBS course

- **50% grade:** Attendance
 - 75% (9 lectures) must be in person
 - 3 can be viewed via recordings
- **50% grade:** Assignments
 - choose 8 of the 10 assignments
 - due by the end of the second Friday after the lecture



Register for BFX

<https://redcap.link/BFX2025>

icts-precisionhealth.wustl.edu

johnegarza@wustl.edu

Bioinformatics Workshop 2025-2026

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 John if you haven't received the following

- Slack access, welcome email, Outlook bfx-workshop-2025 group invite



Register for BFX

<https://redcap.link/BFX2025>



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

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





WHO THE HELL ARE YOU? NETFLIX

Who we are, and why you should trust us



Chris Miller, Ph.D.

Course Director
Associate Professor
Division of Oncology



John Garza

Course Coordinator/TA
Bioinformatics/Genome Analytics
Programmer

20 years of experience in
Bioinformatics and Computational Biology

Jenny McKenzie – ICTS
Precision Health Program Scientist

Other Lecturers/Organizers include:

Jason Walker

Juan Macias

Obi Griffith

Brigida Rusconi

Jennifer Foltz

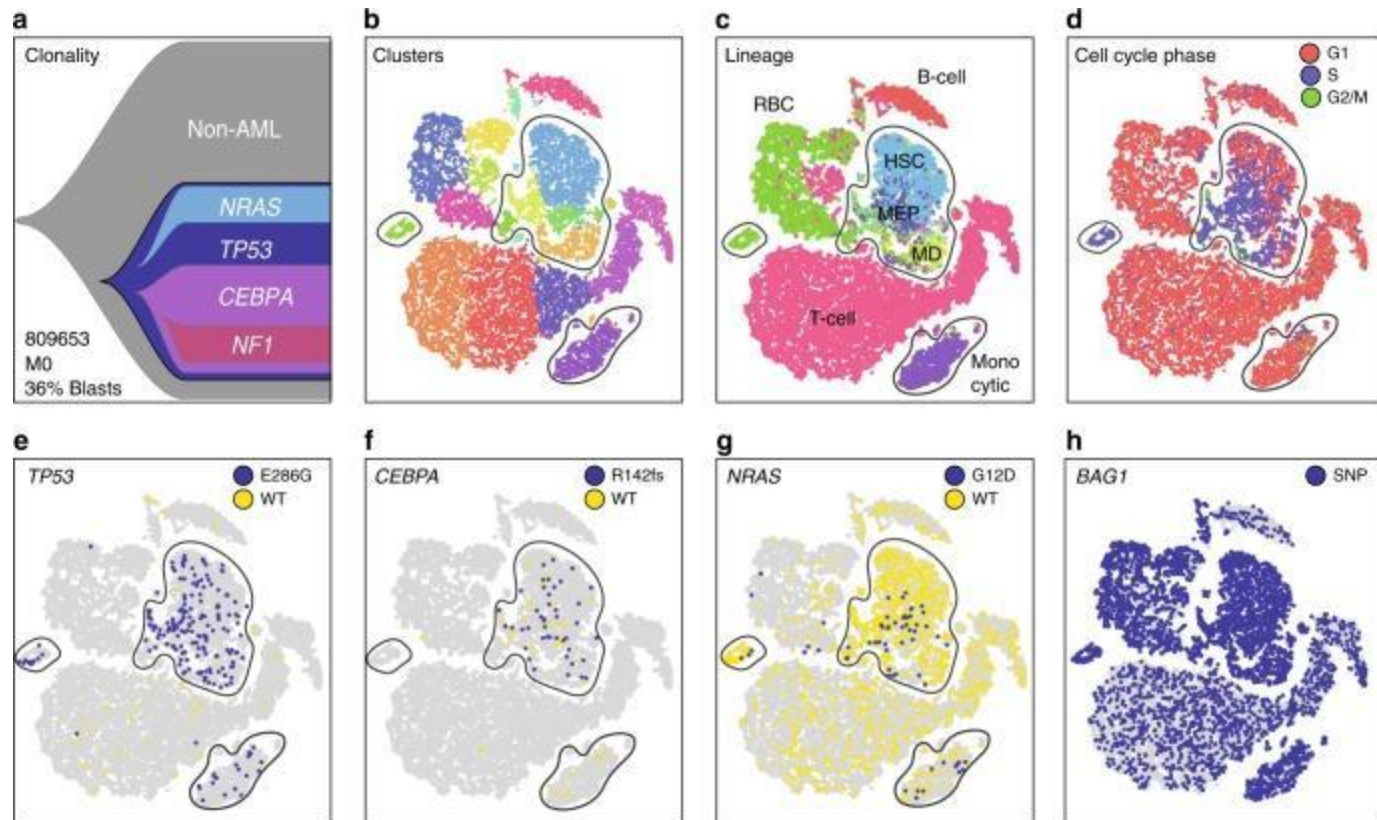
Why learn bioinformatics?

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics

Cost per Genome





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

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



People who know how to do
complex data analysis

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 are the differences between bioinformatics and computational biology?

Bioinformatics vs. Computational Biology

[Computational biology](#), [bioinformatics](#) and [mathematical biology](#) are all interdisciplinary disciplines that combine biology and information science.

Exploring the Differences
Between Computational Biology
and Bioinformatics



Computational Biology vs. Bioinformatics: What's the Difference?

By Samantha Costanzo Carleton

May 28, 2021

Bioinformatics vs. Computational Biology

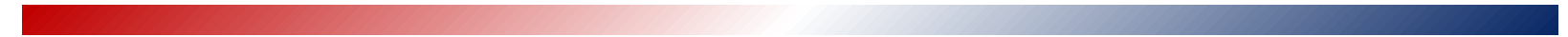
What is bioinformatics?



What is bioinformatics?

More Computational

More biological



Algorithm design

Building Pipelines

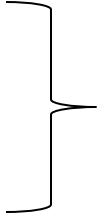
Developing Assays

Analysis of my
experiment

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

Common skills

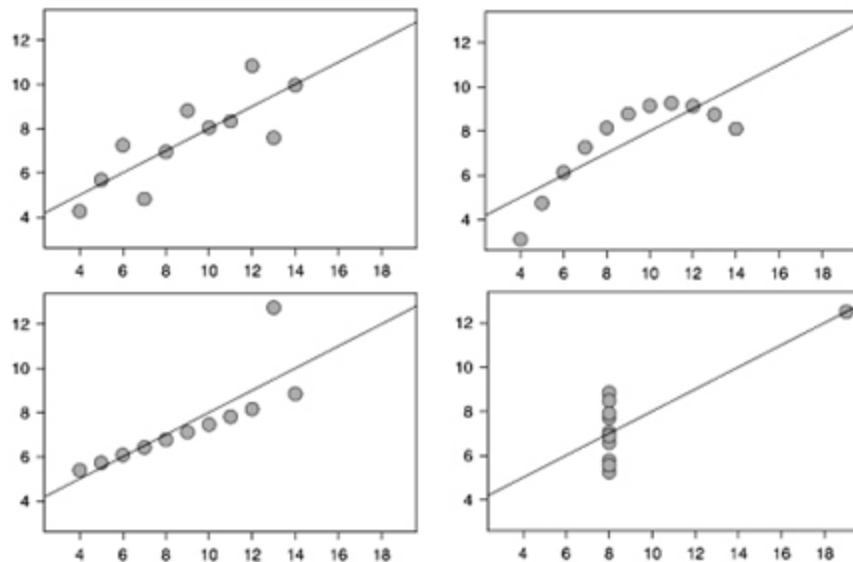
- Statistics
 - Programming
 - Visualization
- 
- “Data science”
- Deep understanding of the biological system and experiments

Goals:

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

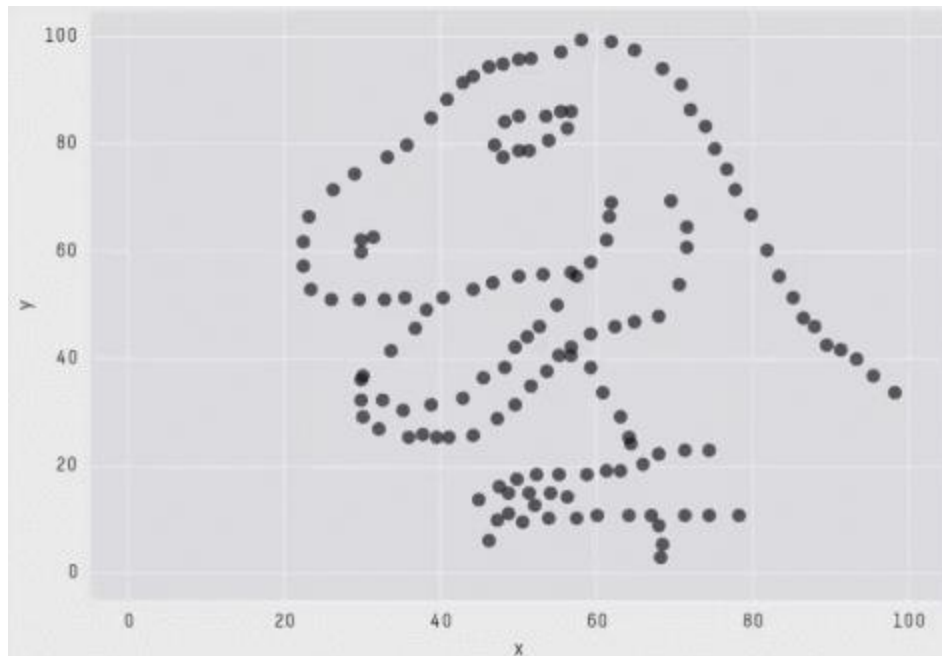
Don't trust your data

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

Datasaurus Dozen



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Summary statistics can be 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

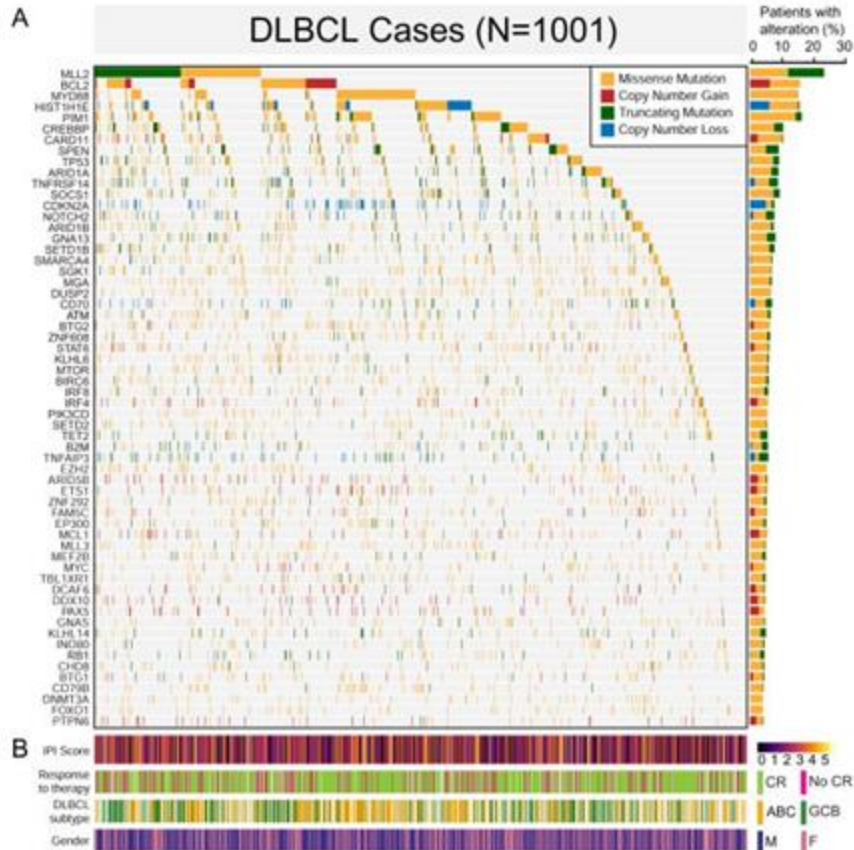
Can't quite co-sign, but he has a point:

If you can't make a plot convincing you that an effect is real, how confident are you, really?

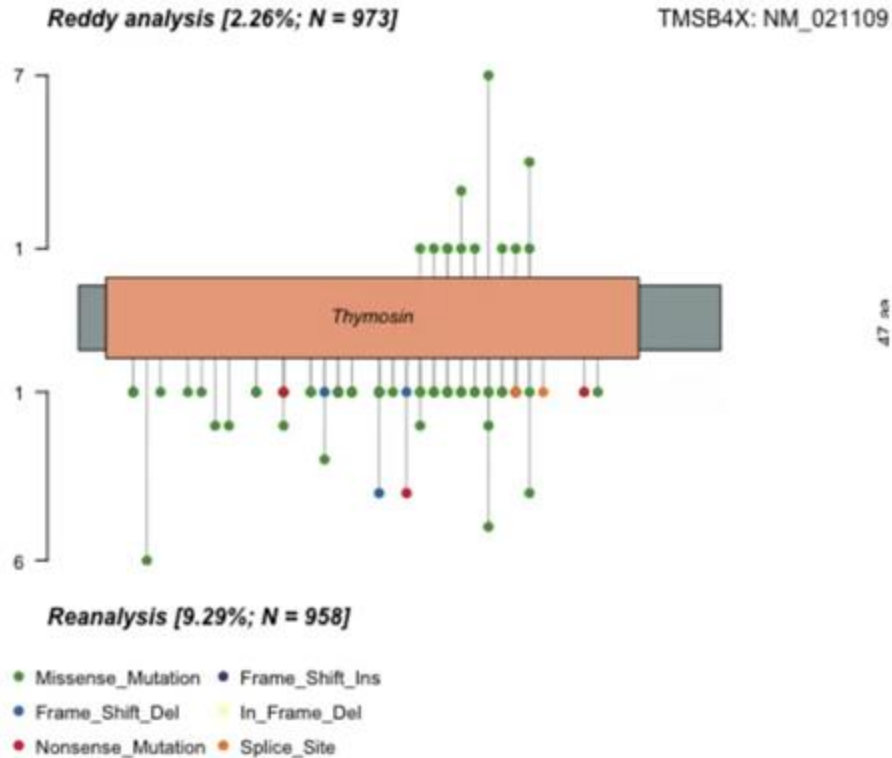
Dangerous situations:

- 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



Real world consequences



Real world consequences

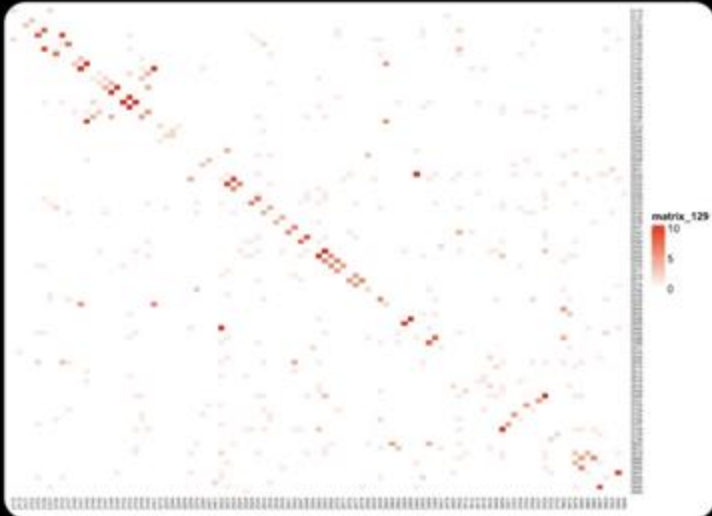
Ryan D Morin @morinryan · Oct 2

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.

1 1 1

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

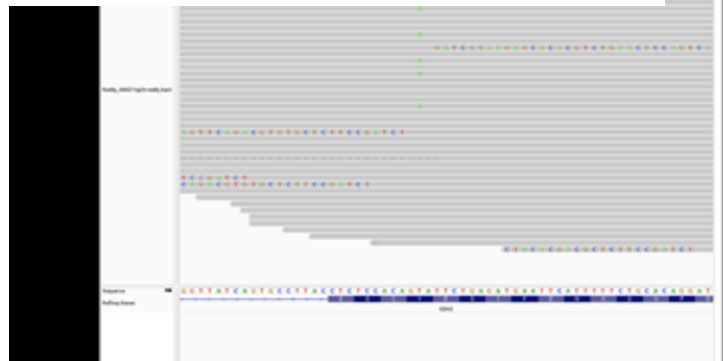
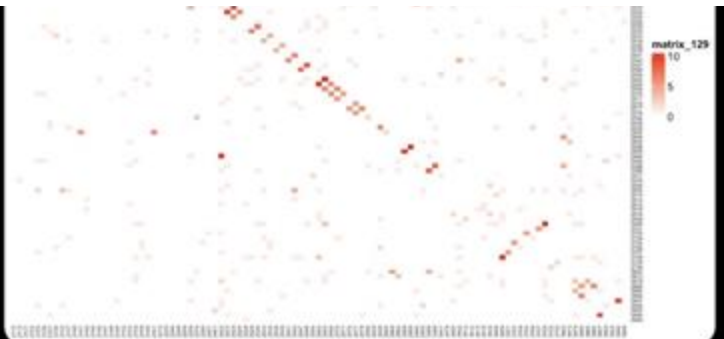
Ryan D Morin @morinryan · Oct 2

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



Errors

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

Lessons to be learned

- Check and double check and triple check your data and your scripts
- Bioinformatic experiments need controls too!
- Sanity checks
- Visualize your data!
- Admit when mistakes are made

“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
- This week
 - setting up your computer
- Next week:
 - Unix command line skills
- Following week:
 - Sequence data generation
 - How to read, manipulate, and run quality control on sequence data

Homework

- Will not be turned in or graded unless you're taking the course for credit
- Will be useful for understanding subsequent lectures
- Posted on the course page

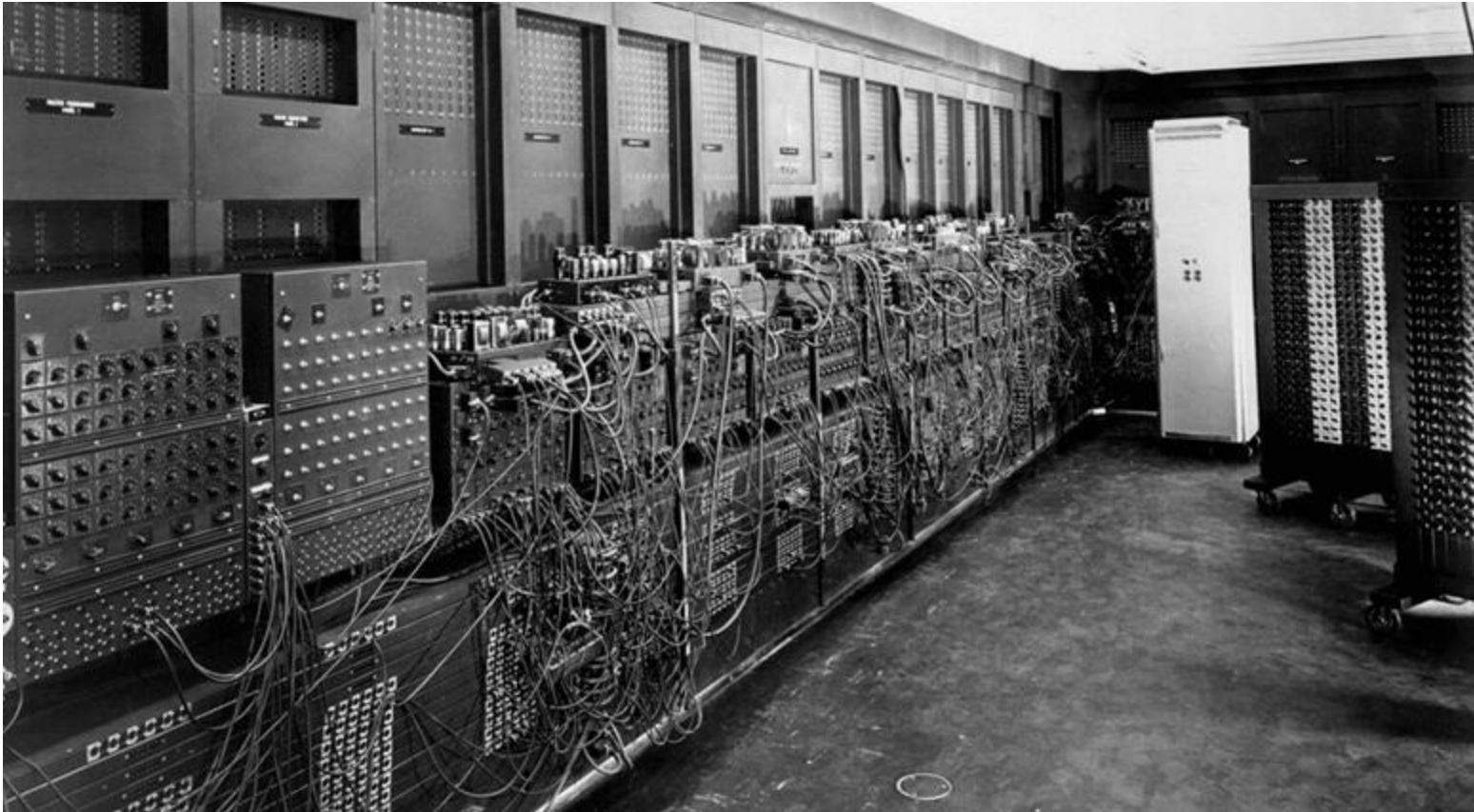
<https://github.com/genome/bfx-workshop>

"Office hours"

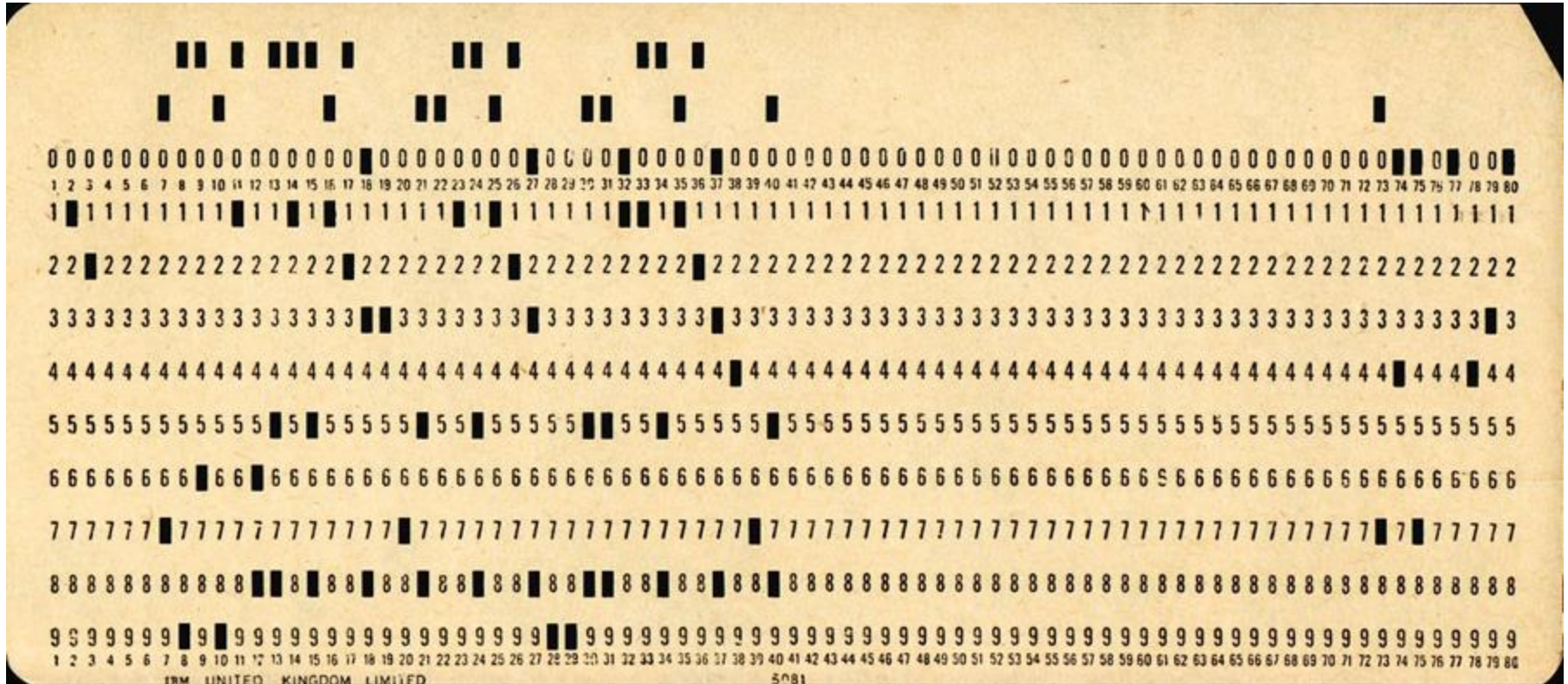
- One hour after class each Monday
 - one extra hour before class next week
- Homework
- Concepts/experimental design
- Your own data/scripts

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

How do we interact with our computers?



How do we interact with our computers?



How do we interact with our computers?



Terminals

Read, Evaluate, Print, Loop

How do we interact with our computers?



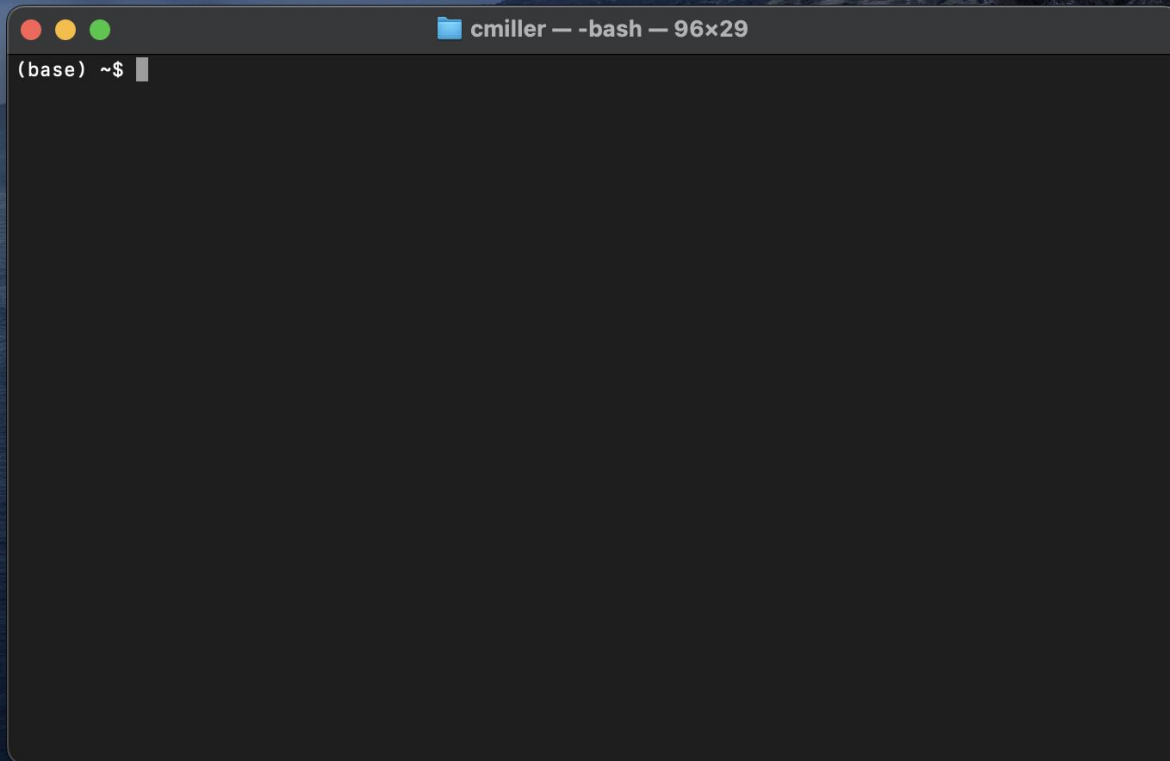
Graphical User Interfaces
(GUIs)

Point and Click

How do we interact with our computers?



GUIs are everywhere, but
terminals aren't dead!



Terminals can do things that GUIs can't

- The big event had to be postponed due to COVID and now we have to change every instance of "Apr 2020" to "Oct 2024". Problem is, there's a huge nested set of directories containing over 10,000 files!
- Clicking around in Windows explorer is not going to get the job done
- On a Unix system, that's just one short line of code:

```
find . -name "*.txt" | xargs -n 1 sed -i.bak 's/Apr 2020/Oct 2024/g'
```

- Seems cryptic at first, but once you learn a little, incredibly powerful!

Unix is the lingua franca of bioinformatics

- high-performance compute clusters run on Unix
- powerful tools for wrangling your data
- writing scripts allows you to do repetitive or error-prone manipulations in a robust and reproducible way
- algorithms for genomics run on the command line

<https://sandbox.bio/tutorials/terminal-basics>

Turning data into insight

