

# Introduction to Bioinformatics

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

Some slides adapted from:

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

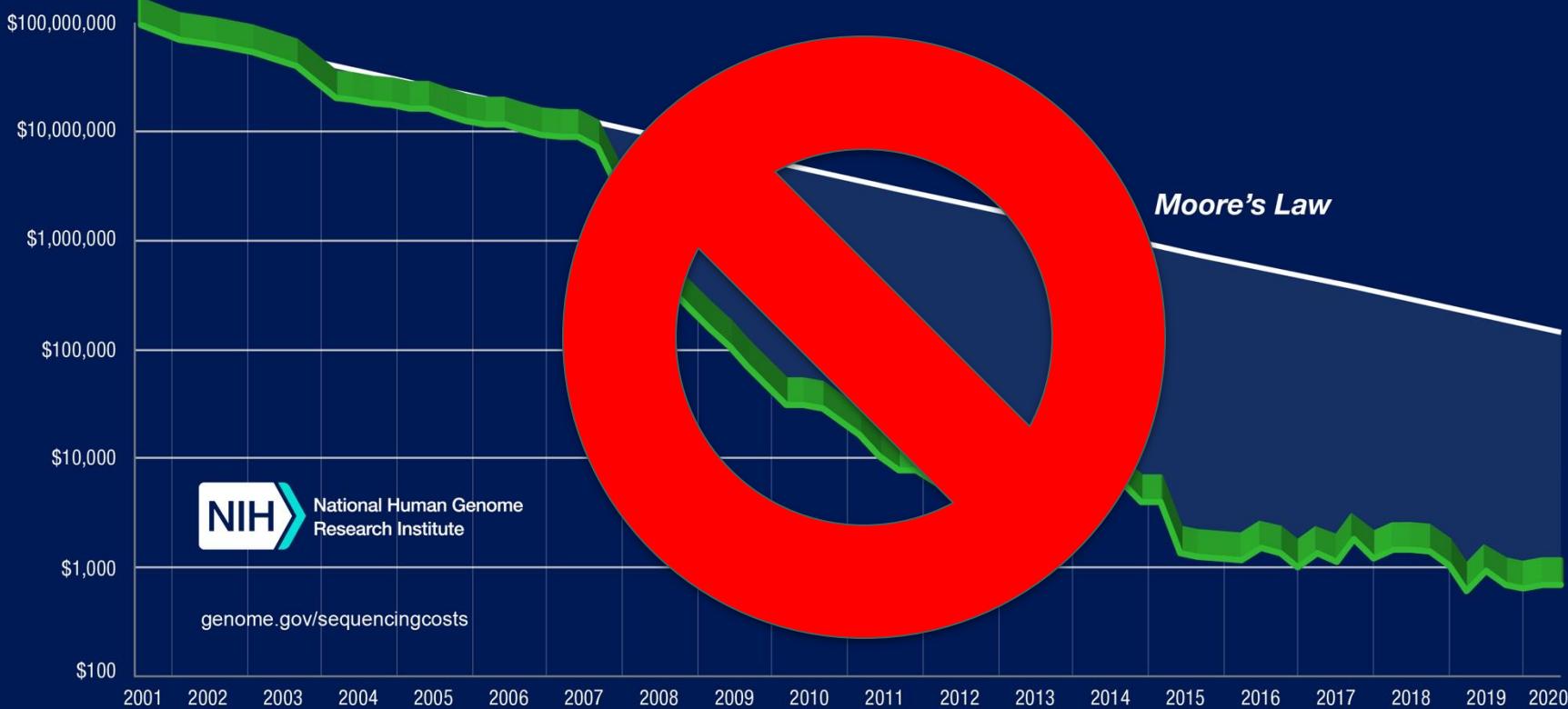
<https://github.com/quinlan-lab/applied-computational-genomics>

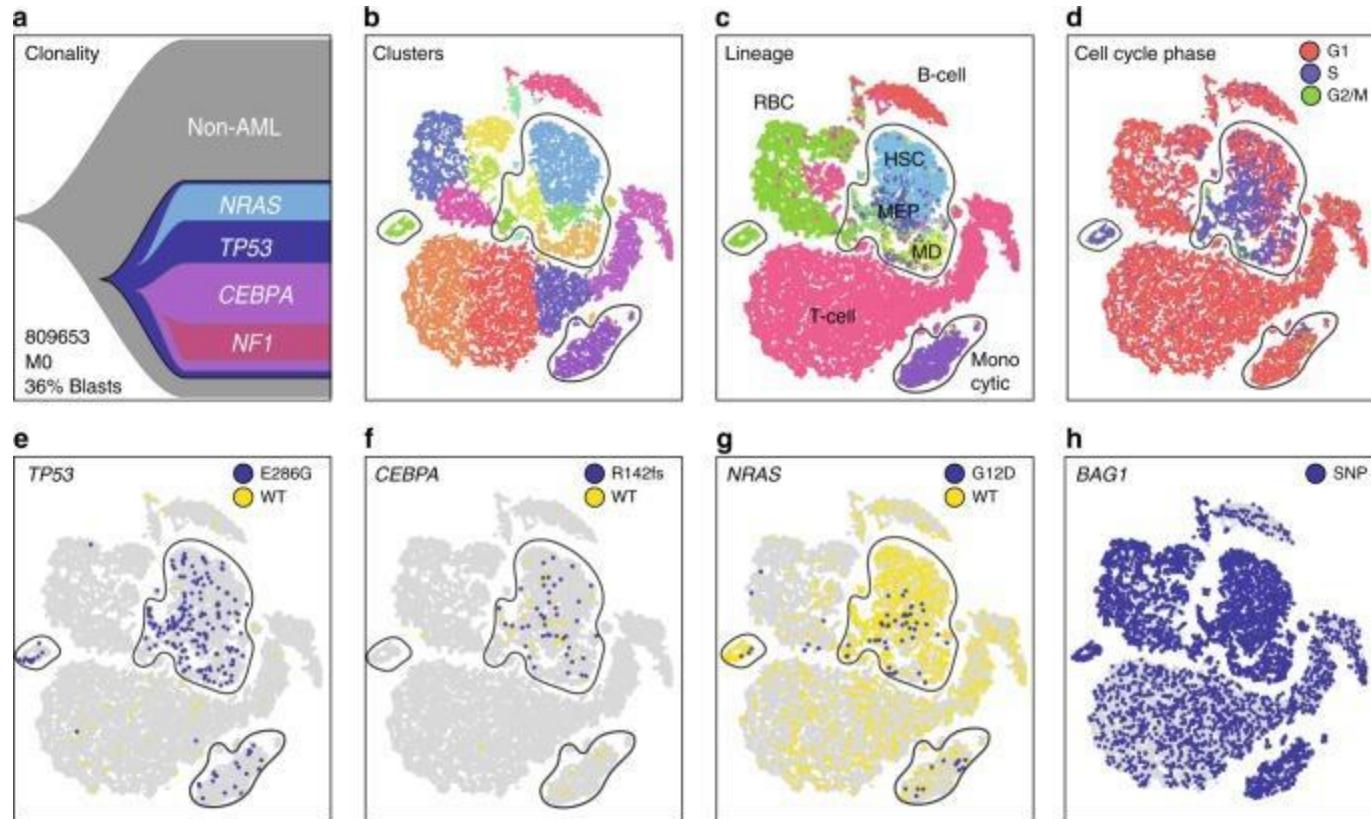


# Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics

## *Cost per Human 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



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
- We're aiming to teach you the theory and practice of computational biology, with a focus on genomics but lessons that apply broadly

# What is bioinformatics?

Chris Miller @chrisamiller · Apr 8

I get that I'm not the arbiter of terminology but it all seems like meaningless distinctions to me. If you're using computers to study biological information, then use any combination of those words as a title. Then more importantly, tell me what you \*actually do\*

Liz Worthey @lizworthey · Apr 8

Bioinformatics vs. Computational Biology: A Comparison  
[medicaltechnologyschools.com/biotechnology/...](http://medicaltechnologyschools.com/biotechnology/)

“Michael” @mikelove · Apr 8

Replying to @chrisamiller

so you’re cool with me saying I study Computer Bioinfology

Rob Patro (@rob@genomic.social) @nomad421 · Apr 8

Replying to @mikelove and @chrisamiller

As long as you sound drunk whenever you say it, I’d be fine with this.

8 3 63 ⬆

2 6 40 ⬆

9 ⬆

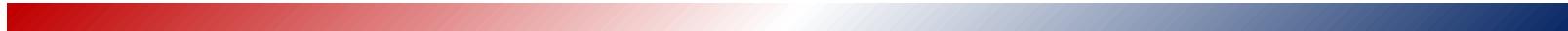
# What is bioinformatics?

- 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

# What is bioinformatics?

More Computational

More biological



Algorithmic Design

Dataviz

Biological expertise

Machine Learning

Statistics

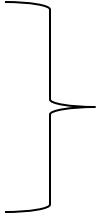
Scientific rigor

High performance/cloud  
computing

Data munging

Communication

# Common skills

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

# Goals of this course

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

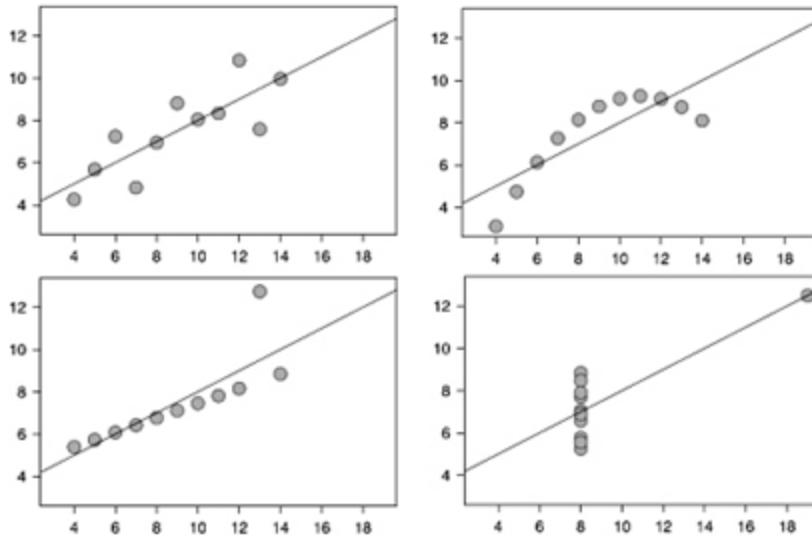
# Course structure

- Command-line basics to get you up to speed
- Generation of sequencing data, formats, alignment
- Variant calling and interpretation
- ChIP-seq and methylation
- Bedtools/genome arithmetic
- Intro to the R programming language
- Bulk RNA-seq, alignment, QC, quantification, diff expression
- Introduction to Python
- Single-cell RNA-seq
- Long read sequencing – RNAseq

**Don't trust your data**

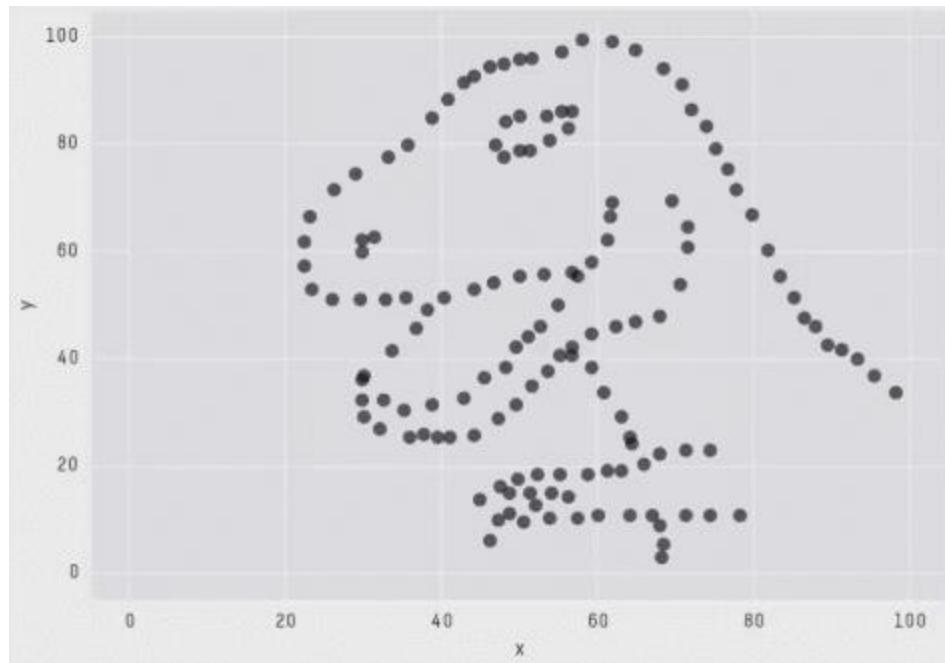
# Trusting your data

Anscombe's quartet



Property	Value	Accuracy
Mean of $x$	9	exact
Sample variance of $x$ : $\sigma^2$	11	exact
Mean of $y$	7.50	to 2 decimal places
Sample variance of $y$ : $\sigma^2$	4.125	$\pm 0.003$
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 are dangerous

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

# 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"

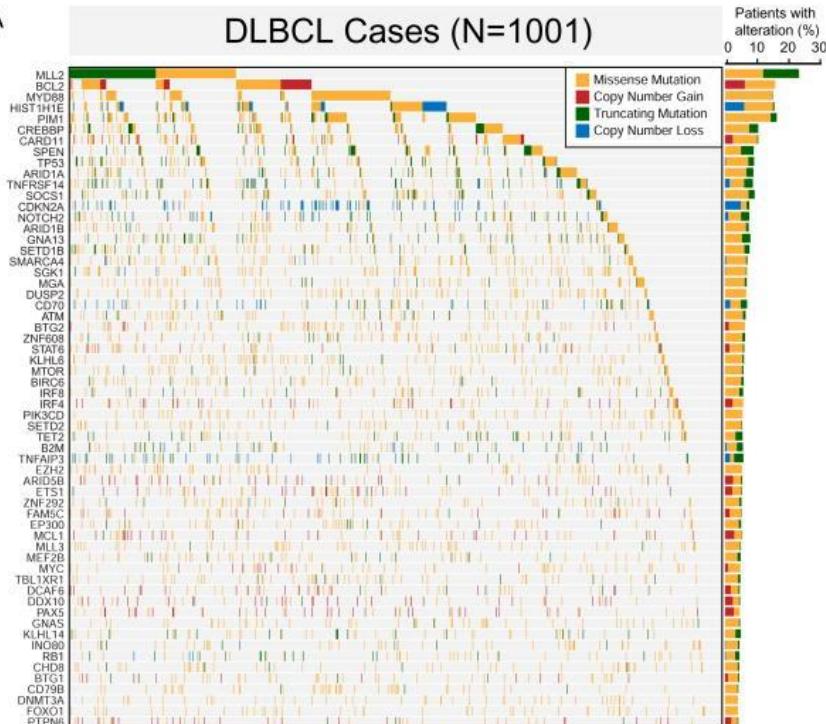
- Ernest Rutherford

- 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

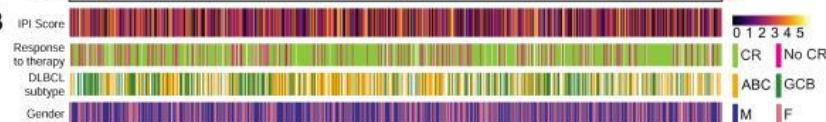


# Real world consequences

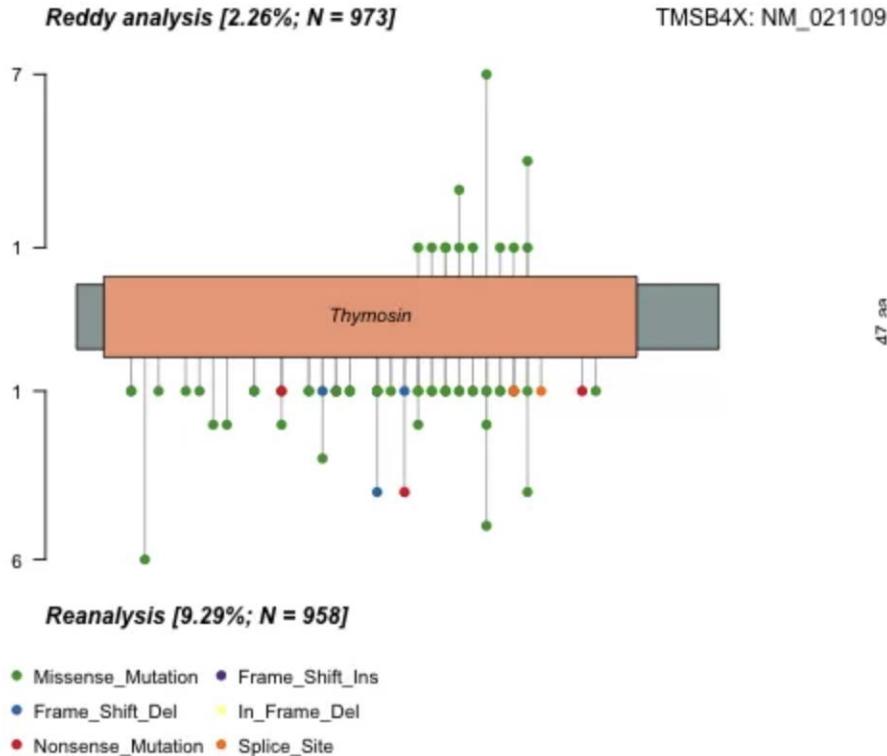
A



B



# Real world consequences



# Real world consequences

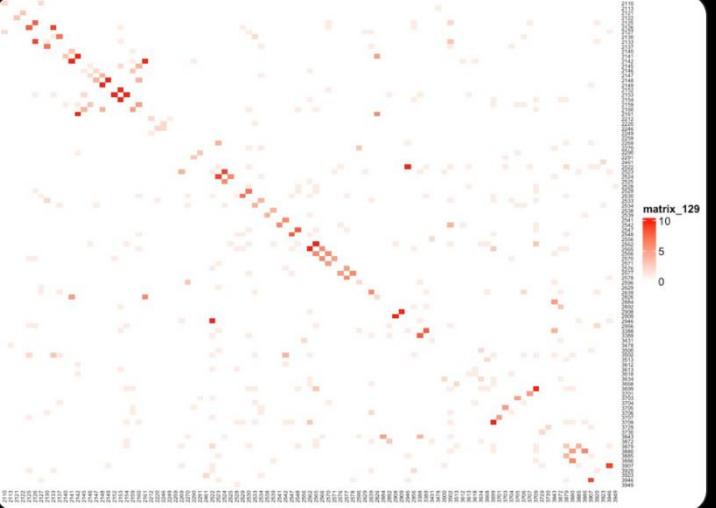
Ryan D Morin @morinryan · Oct 2

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...](http://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.



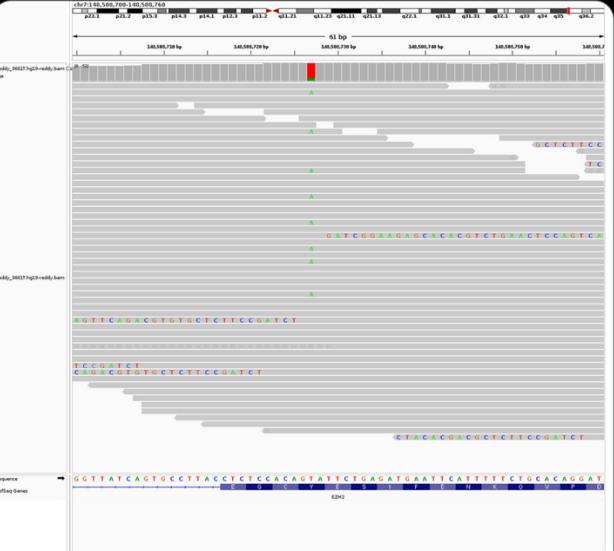
Ryan D Morin @morinryan · Nov 4

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 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

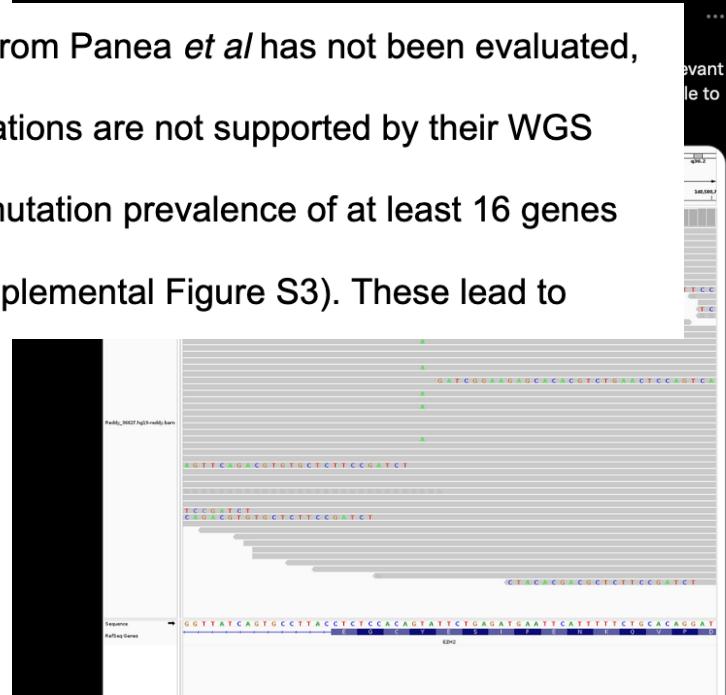
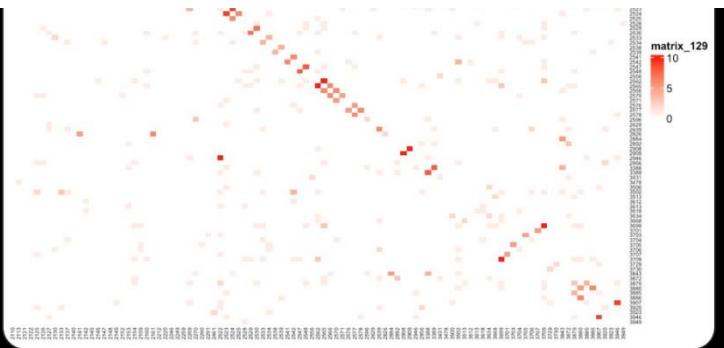
Ryan D Morin @morinryan · Oct 2

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...](http://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

# Bioinformatics is science

- It's iterative. Doing experiments, generating hypothesis, testing hypotheses with new experiments
- It is easy to find programmers who will just feed data through someone else's scripts.
- It is hard to find scientists who have the cross-domain knowledge to do creative science and critically evaluate the results.

# AI / Large language models

- Can be useful
- Are often confidently wrong
- Most dangerous for beginners
- Positive/Negative controls incredibly important

# Reproducibility

- If you're doing bioinformatics right, reproducibility should be "easy"!
- Data will be well organized and stored safely

sample637.tsv  
sample647.tsv  
sample983.tsv

Mouse\_TP53\_WT\_637.tsv  
Mouse\_TP53\_KO\_647.tsv  
Mouse\_TP53\_KO\_983.tsv

Laptop vs compute cluster (both need backups)!

# Reproducibility

- Your tools/parameters/settings should all be stored in scripts
- The ideal to aim for is that you could send someone a link to your data and scripts, and they could sit down, run it, and reproduce your figures/tables
- Hell is other people's data.  
Hell is also your own data 6 months later.

# Reproducibility

- Your tools/parameters/settings should all be stored in scripts
- The ideal to aim for is that you could send someone a link to your data and scripts, and they could sit down, run it, and reproduce your figures/tables

R Markdown

The screenshot shows an RStudio interface with an R Markdown file open. The code in the editor is as follows:

```
16: # Using Terrain Colors
17: 
18: rrr<-raster::raster("MaungaWhauVolcano.tif", col=terrain.colors(100),nres=FALSE)
19: contour(x, y, values, col=terrain.colors(100),lwd=2)
20: contour(x, y, values, levels=seq(90, 200, by=5), add=TRUE, col="brown")
21: axis(1, at=rat)
22: axis(2, at=rat)
23: axis(3, at=rat)
24: axis(4, at=rat)
25: title(main="Maunga Whau Volcano", sub = "col=terrain.colors(100)", font.main=4)
26: 
```

Below the code is a volcano plot titled "Maunga Whau Volcano". The x-axis is labeled "x" and ranges from 100 to 800. The y-axis is labeled "y" and ranges from 100 to 600. The plot shows a central peak with concentric contours.

At the bottom of the RStudio window, the console shows the command "R Markdown" being run.

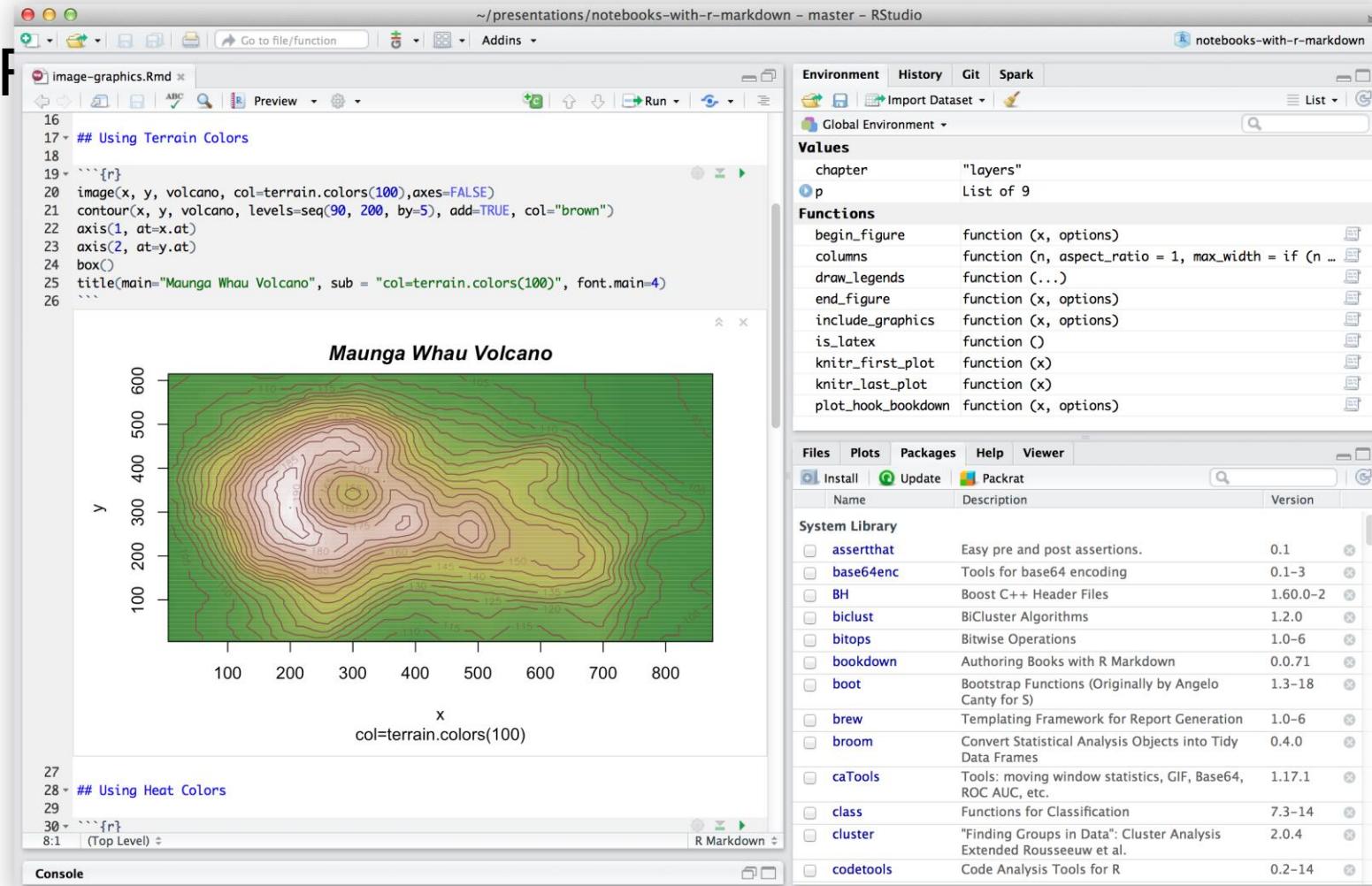
Jupyter notebooks

The screenshot shows a Jupyter notebook interface with a code cell and a figure cell.

The code in the cell is:

```
# Using Step Tracker Data
# Load step tracker data
# Analyze and visualize the raw data
# Analyze and visualize transforms of the data
# Simulate the "resttime" system
```

The figure cell displays a line plot titled "Accelerometer Signal (Peaks)". The x-axis represents time from 0 to 400 seconds, and the y-axis represents signal amplitude from -4000 to 6000. The plot shows a noisy signal with several sharp peaks. A legend indicates three data series: "Mag unfiltered" (blue line), "Mag smoothed" (orange line), and "Peak Locations" (red dots).





R

File Edit View Insert Cell Kernel Help

Not Trusted

Python 3



## Inputs

We are using a toy example data set based on the HCC1395 blood normal cell line. The sequence reads and genome reference are a subset targeting chr6, genes HLA-A and HLA-B-C, and chr17, genes TP53 and BRCA1.

[FASTA Normal Reads Lane 3](#) [Normal Reads Lane 4](#)

All inputs and additional resources can be viewed at: <https://console.cloud.google.com/storage/browser/analysis-workflows-example-data>

In this example, each file was downloaded to ~/Downloads. If you saved the downloaded files in another folder or location, the following paths will need to be updated to account for those differences.

```
In [ ]: !mv ~/Downloads/hla_and_brca_genes.fa $PWD/ref/.  
!mv ~/Downloads/2895499331.bam $PWD/unaligned/normal/.  
!mv ~/Downloads/2895499399.bam $PWD/unaligned/normal/.
```

## Index

Index files of various formats and data structures are used commonly in genomics to provide random access to specific records, locations, or content within much larger domains and coordinate spaces. Ex. entire genome nucleotide sequences, billions of sequence reads, millions of variant records, etc.

## Samtools

Using samtools faidx, we create an index (.fai) of the reference FASTA (.fa) which provides random access to the nucleotides at specific positions within the complete genome reference. The FAIDX index is used both by the samtools faidx command as well as other toolkits, algorithms, and libraries that require random access to specific coordinates in a timely manner.

```
In [ ]: ls $PWD/ref
```

We only have the FASTA sequence, let's make a FAIDX format file using Samtools by running the command with no additional arguments (other than the FASTA file):

# What happens to my data after analysis?

- Journals will not publish data that isn't accessible
  - NOT just "Available upon reasonable request"!
- Every NIH grant now requires a Data Sharing/Management plan
- Covers more than just sequence data – gel images, textual qPCR readouts, flow plots/data, etc
- Needs to go into a repository

# FAIR principles of data

- Findable
- Accessible
- Interoperable
- Reproducible

# Examples of good places to deposit data

- NCBI repositories
  - SRA – Short Read Archive
  - dbGaP – front-end/access control for human data in SRA
  - GEO – rich metadata associated with experiments (RNA, scRNA, arrays, etc)
- Organism specific repos (Flybase, etc)
- General data repos that assign a doi (zenodo.org)
- Institutional repositories – your university library probably runs one
- **NOT** a lab website or a Google bucket (what happens in 5 years?)

# What happens when this works well?

National Heart, Lung, and Blood Institute | **BioData CATALYST** | Powered by Gen3

Browse Data | Documentation | ICMILLER | Logout

**Data** **File**

Explorer Filters | Data Tools | Summary Statistics | Table of Records

**Data Access** ▾

- Data with Access
- Data without Access
- All Data

Export to Seven Bridges ⋮ Export All to Terra ⋮ Export to PFB ⋮ Export to Workspace ⋮

Subjects	Projects
53,964	23

**Annotated Sex**

no data
53,964 (100%)

**Race**

no data
100%

**Filters**

**Harmonized Variables**

**Project** **Subject**

**Collapse all**

**Program** 1 selected X Q

- topmed 53,964
- parent 186,592
- tutorial 14,433
- open\_access 3,202

**Project Id** Q

- topmed-BioMe\_HMB- 15,874

Showing 1 - 20 of 53,964 subjects Show Empty Columns ⋮

Project Id	Data Format
topmed-CAMP_DS-AST-COPD	CRAM
topmed-BioMe_HMB-NPU	CRAM
topmed-BioVU_AF_HMB-GSO	

# What happens when this works well?

This screenshot shows the BioData CATALYST interface. The main search bar at the top contains the query "analysis". Below the search bar, there are several tabs: "Search", "Explore", "Discover", "Manage", and "Profile". The "Search" tab is currently selected. The results table has columns for "Title", "Description", "Type", "Status", and "Actions". There are 53,964 results found. A red circle highlights the "Actions" column for the first result, which is "fifteen".

This screenshot shows the Dockstore interface. The left sidebar includes links for "Dashboard", "Workflows" (which is currently selected), "Tools", "Services", "Starred", "Account", and "Help Desk". The main content area is titled "My Workflows". It shows a "GITHUB" section with a dropdown menu set to "chrisamiller". Under "chrisamiller", there are sections for "genome" (with "Published" and "Unpublished" tabs) and "five-dollar-genome-analysis-pipeline" (under "Published"). Below this, there are sections for "griffithlab" and "leyabdotorg", both showing "No unpublished workflows.". At the top right, there are "Publish" and "Refresh" buttons. The right side of the screen displays information about a workflow from "github.com/genome/analysis-workflows:master". It shows the "Info" tab is active, with the URL "github.com/genome/analysis-workflows:master", the version "master", and the last update date "Jan. 8 2020". It also mentions "Last update to source repository: Jan. 8 2020". A note at the bottom encourages users to "Keep your workflow automatically in sync with GitHub with our new registration process. Click [here](#) to learn more." On the far right, there are "Edit" buttons for "Source Code" and "Workflow Path".

# What happens when this works well?

The screenshot shows the Terra Workflows interface for the workspace 'leylab.terra1/ch\_fusion\_topmed'. The workflow 'wdl\_samtools.manta' is selected. The interface includes a sidebar with 'Data CATALYST' and various filters, a main panel with workflow details, and a preview section at the bottom.

**Workflow Details:**

- Version: v1.3
- Source: [github.com/saimukund20/wdl\\_samtools.manta](https://github.com/saimukund20/wdl_samtools.manta)
- Synopsis: No documentation provided.
- Inputs: Run workflow with inputs defined by data table (selected)

**Step 1:** Select root entity type: pharmu\_set

**Step 2:** SELECT DATA No data selected

**Optional Inputs:**

Task name ↓	Variable	Type	Attribute
wf	fusion_sites	File	"gs://fc-6248b026-3011-4591-9688-255a248b35b9/sites_merged_with_solid_tumor.bed"
wf	manta_config	File	"gs://fc-6248b026-3011-4591-9688-255a248b35b9/configManta_1.py.in"
wf	reference	File	workspace.referenceData_hg38_ref.fasta

# What happens when this works well?

This screenshot shows the BioData CATALYST interface. At the top, there are tabs for Data, File, and various analytical tools. Below this, a main panel displays analysis progress: 53,964 subjects, 13,041 samples, and 23 projects. A red circle highlights the 'Analyze' button. On the left, a sidebar lists 'Data Access' and 'Filters'.



This screenshot shows the WORKSPACES interface. It displays a workflow step titled 'Step 1: align reads to reference genome'. The interface includes tabs for Script, Inputs, and Outputs. The 'Inputs' tab shows two files: 'phred\_33.fastq.gz' and 'reference'. The 'Outputs' tab shows a single file: 'alignments.bam'. A red circle highlights the 'align reads to reference genome' button.



This screenshot shows the Dockstore interface. It displays a GitHub repository named 'github.com/genome/analysis-workflows'. The repository has a 'master' branch and was last updated on Jan. 8, 2020. The interface includes tabs for Info, Launch, Versions, and Files.



Analysis results for >50,000 Whole  
Genome Samples in a few hours

# Infrastructure – where do I analyze my data?

- Laptop
  - Pro: Easy – it's sitting in front of you!
  - Pro: You have root access (can install anything you need)
  - Con: Power is limited – number of cores, amount of RAM
  - Con: amount of disk is limited (a single WGS experiment can be >50 Gb)
  - Con: what if it's stolen? (you do have automatic backups, right?!)
  - Con: what happens when you close the lid?



# Infrastructure – where do I analyze my data?

- Big desktop machine or blade in your lab
  - Pro: moderate amounts of CPUs/RAM for big jobs
  - Pro: You have priority access
  - Pro: Can submit jobs and walk away
  - Con: You have to become a sysadmin and take care of it.  
(Who applies security updates? What happens if the power supply fails? Backups?)



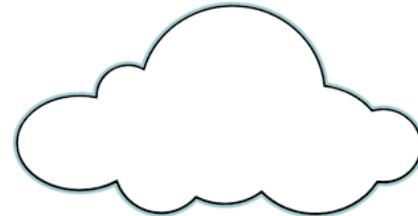
# Infrastructure – where do I analyze my data?

- Local Compute Cluster
  - Pro: Lots of CPUs/RAM for big jobs
  - Pro: probably has dedicated disk with good backups
  - Pro: Can submit jobs and walk away
  - Con: You may have to contact administrators to do installs
  - Con: you have to share resources, and you may not have priority!



# Infrastructure – where do I analyze my data?

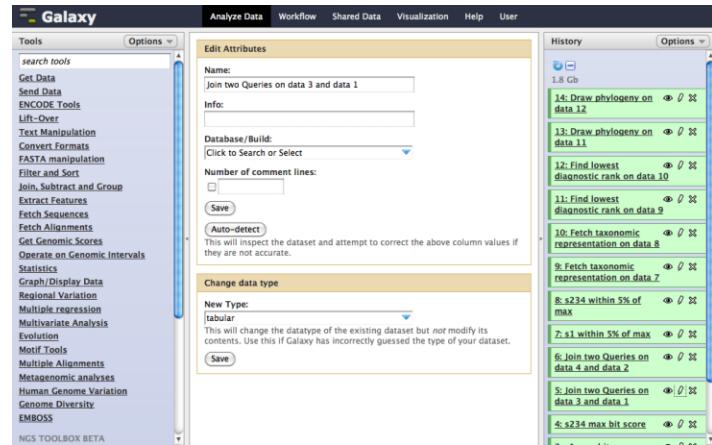
- Cloud (remote compute cluster)
  - Pro: As much CPU/RAM as you can imagine
  - Pro: secure disk, backups, etc
  - Pro: no reasonable limits on access
  - Con: You may have to transfer your data up/down
  - Con: can be pricey (and you have to be so careful!)
  - Con: unless you have institutional support, you have to learn to administer it



# Infrastructure – where do I analyze my data?

- Web analysis portals

- Backed by cloud, more friendly front ends
- Still have to pay for it, learn the system
- Have to transfer your data up/down
- May have GUIs for common tools



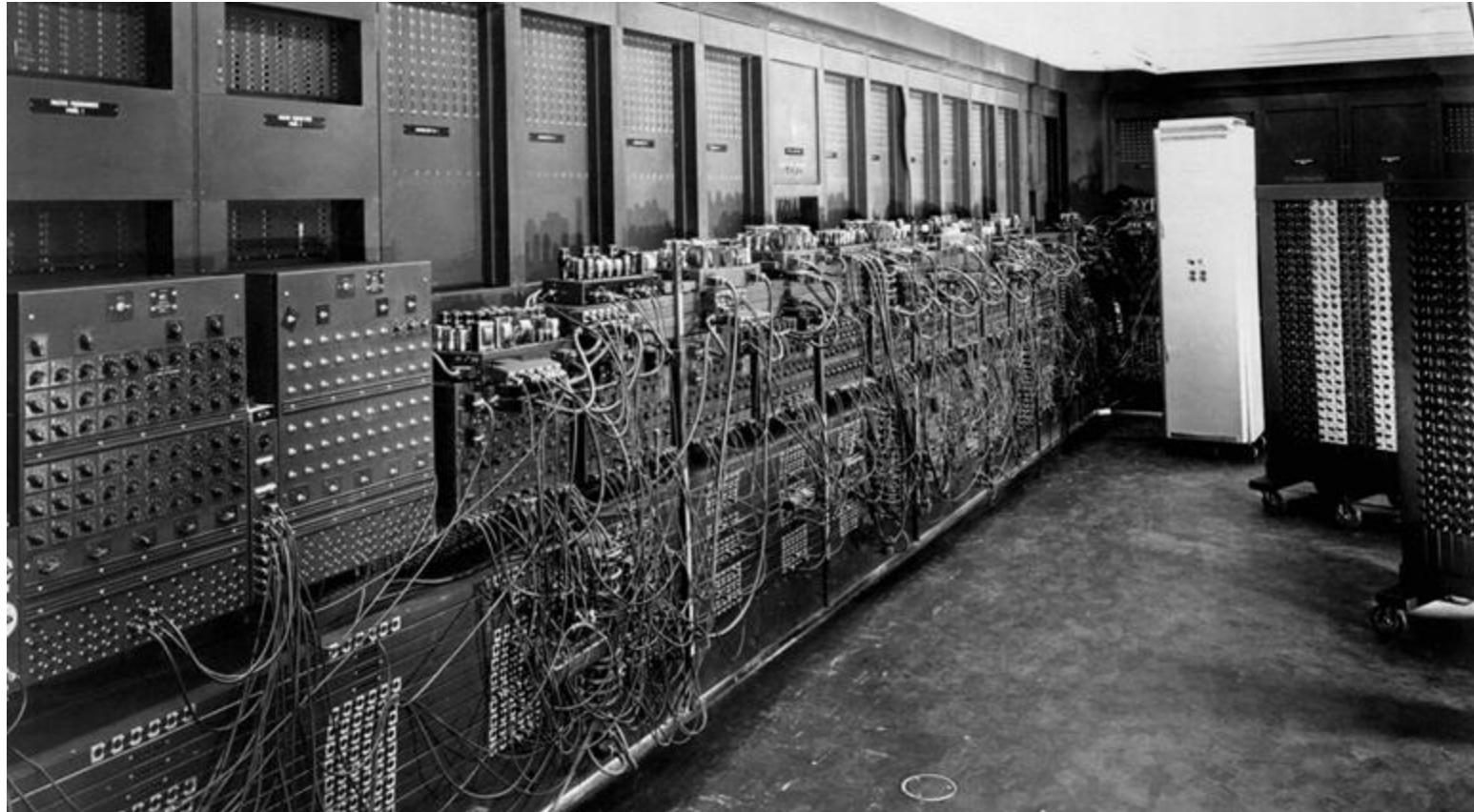
# Infrastructure

- Bioinformatics requires infrastructure just like lab work
  - pipette tips don't appear and cell cultures don't feed themselves
  - servers don't appear and software doesn't install itself

# Funding

- Traditionally has been very hard to get grants for software development
- even harder for maintenance
- large amount of "abandonware"

# How do we interact with our computers?



# Computer "bugs"

Photo # NH 96566-KN (Color) First Computer "Bug", 1947

92

9/9

0800 Antran started  
1000 " stopped - antran ✓ { 1.2700 9.037847025  
13 UC (032) MP - MC 1.2700 9.037846995 const  
(033) PRO 2 2.130476415  
const 2.130476415

Relays 6-2 in 033 failed special speed test  
in relay " 10.00 test .

Relay  
2145  
Relay 3371

1100 Started Cosine Tape (Sine check)  
1525 Started Multi-Adder Test.

1545

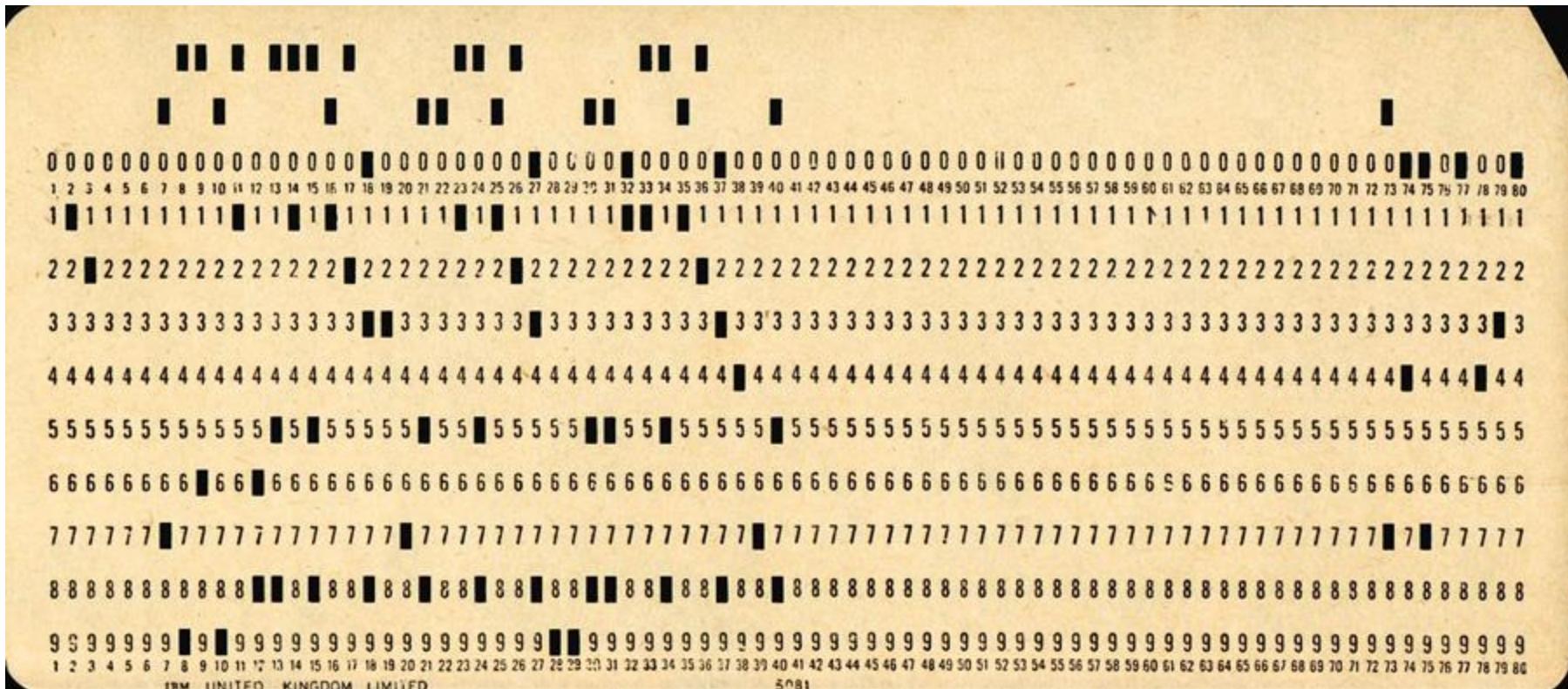


Relay #70 Panel F  
(moth) in relay.

1600 Antran started.  
1700 closed down.

First actual case of bug being found.

# How do we interact with our computers?



# How do we interact with our computers?



Terminals

Read, Evaluate, Print, Loop

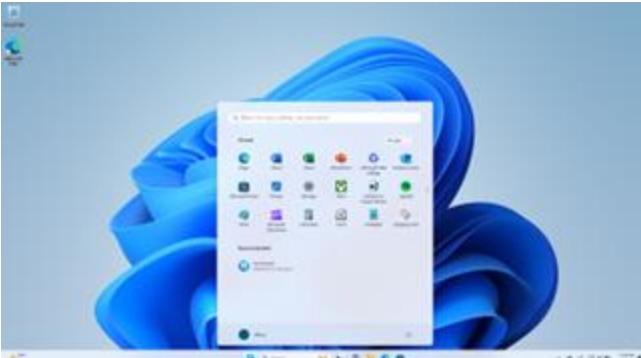
# How do we interact with our computers?



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

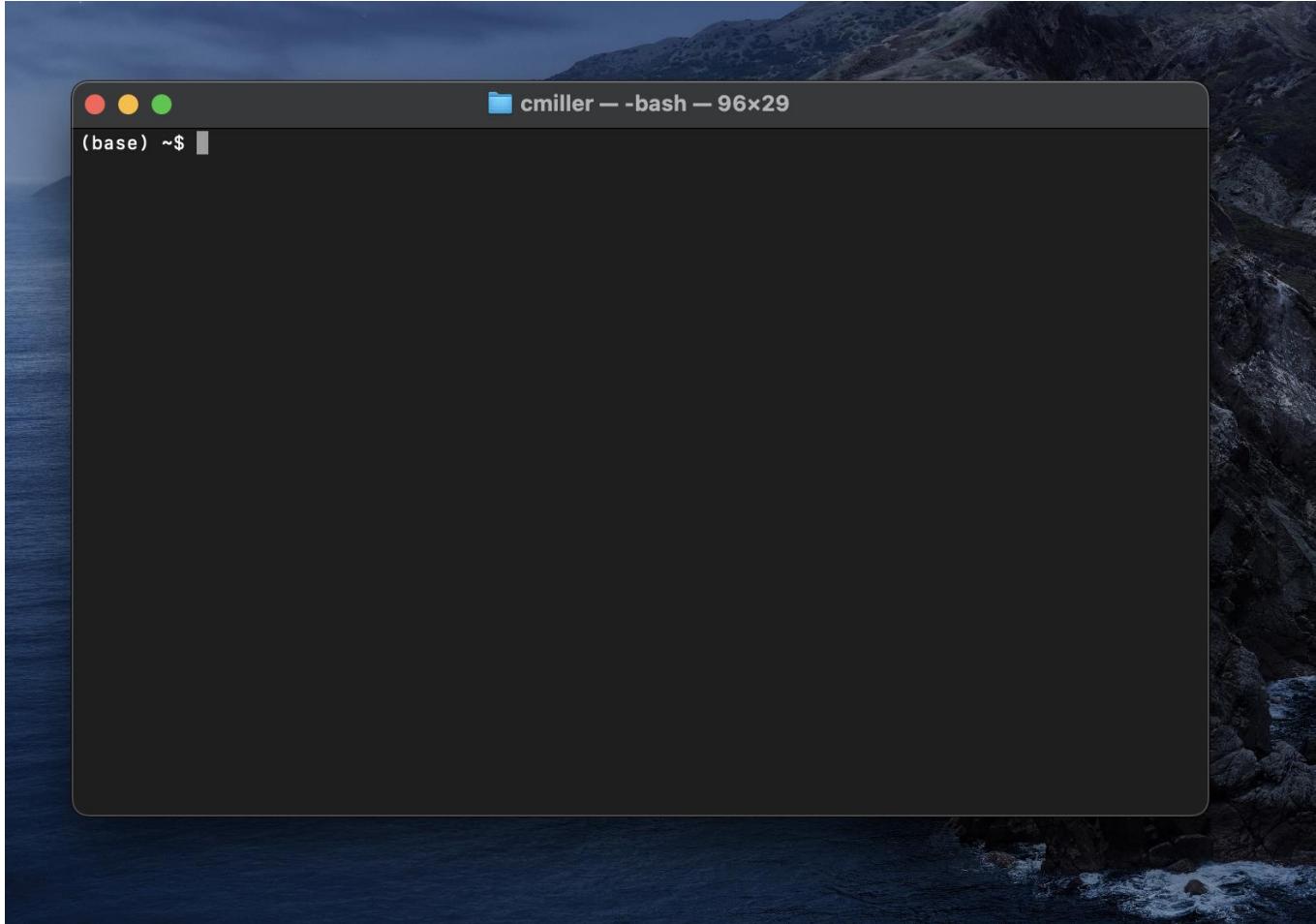
- The big event had to be postponed due to COVID and now we have to change every instance of "Apr 2020" to "Oct 2024". Problem is, there's a huge nested set of directories containing over 10,000 files!
- Clicking around in Windows explorer is not going to get the job done
- On a Unix system, that's just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



# Course structure

- Command-line basics to get you up to speed
- Generation of sequencing data, formats, alignment
- Bedtools/genome arithmetic
- Variant calling and interpretation
- ChIP-seq and methylation, peak calling
- Introduction to the R programming language and data visualization
- Bulk RNA-seq and differential gene expression
- Introduction to Python
- Single-cell RNA-seq
- Statistics and probability
- Long-read RNAseq

# Turning data into insight

