

Introduction to Bioinformatics

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

Some slides adapted from:

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

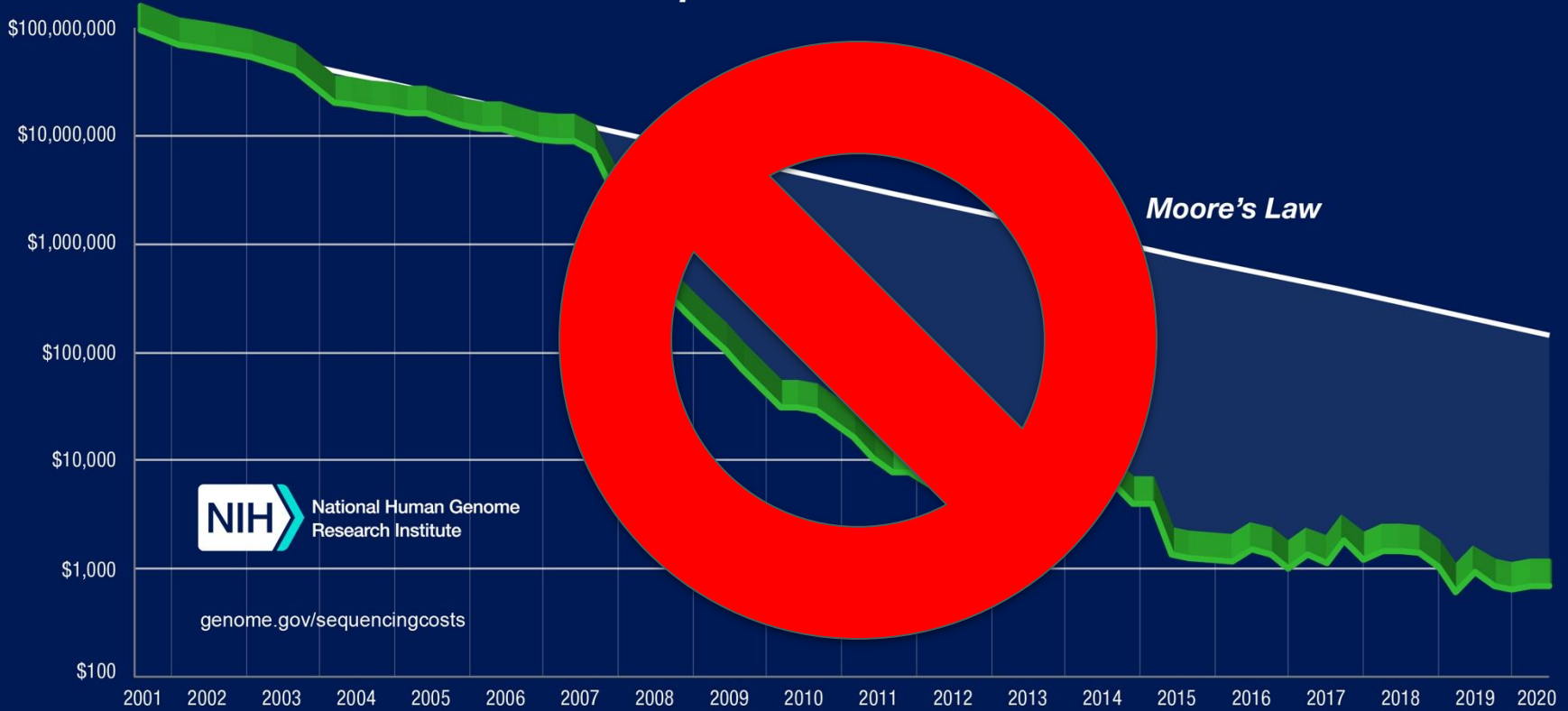
<https://github.com/quinlan-lab/applied-computational-genomics>

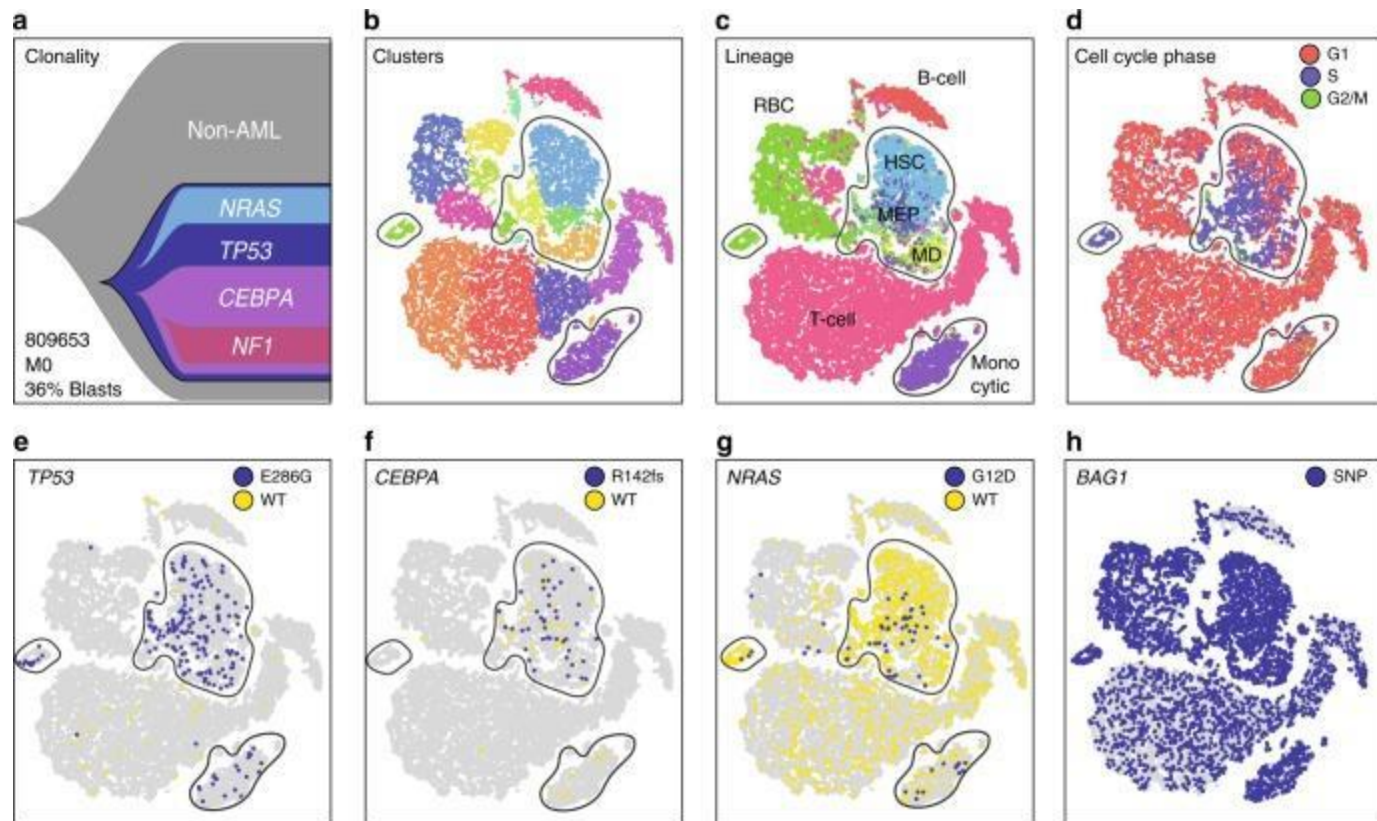


Why learn bioinformatics?

- 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?



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

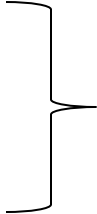
Scientific rigor

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

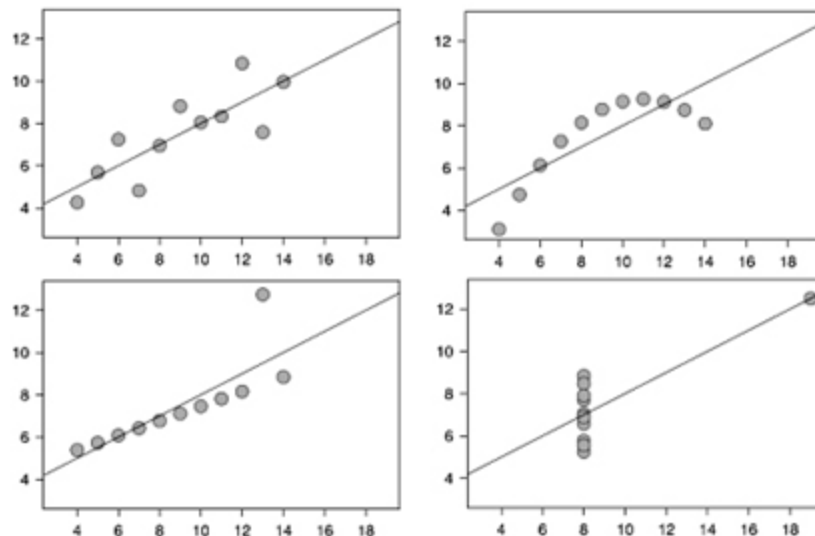
Course structure

- 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

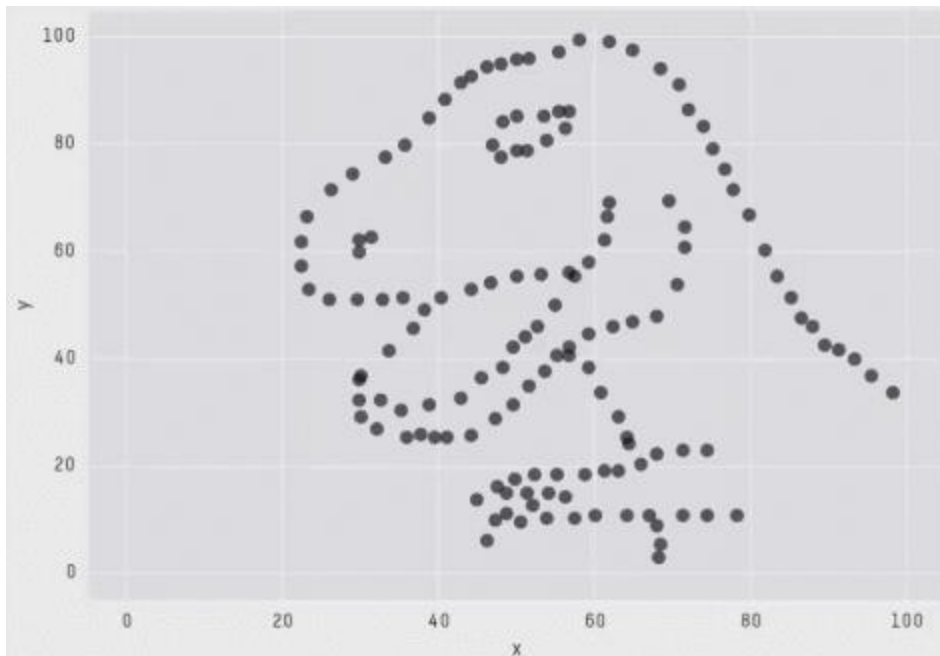
Trusting your data

Anscombe's quartet



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : 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 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

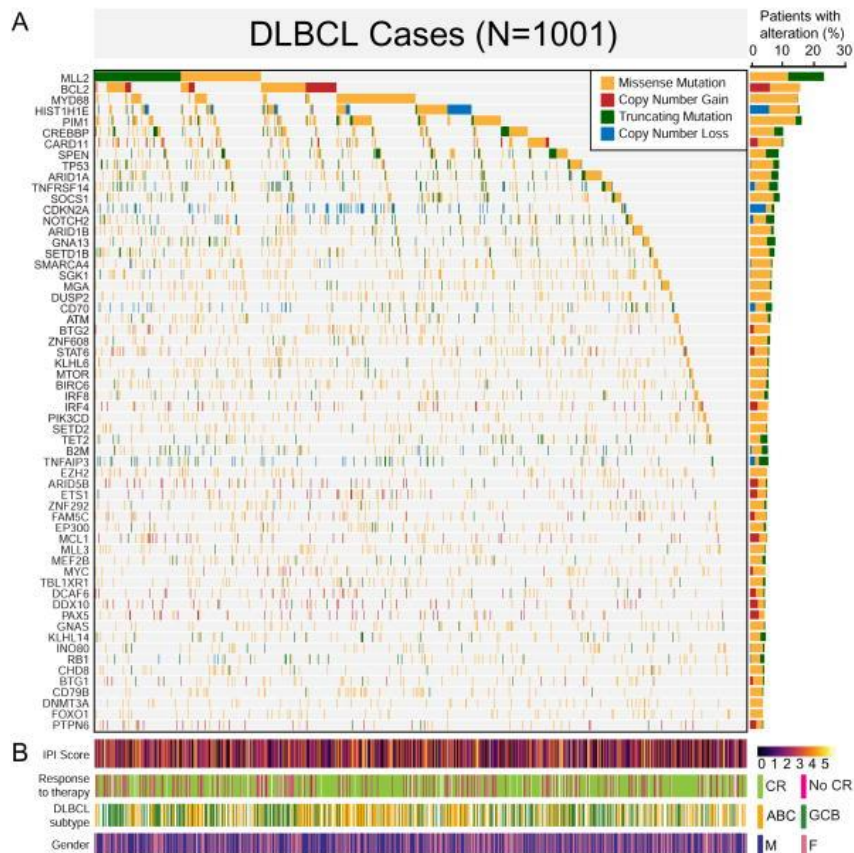
"If your experiment needs statistics, you ought to have done a better experiment"

- Ernest Rutherford

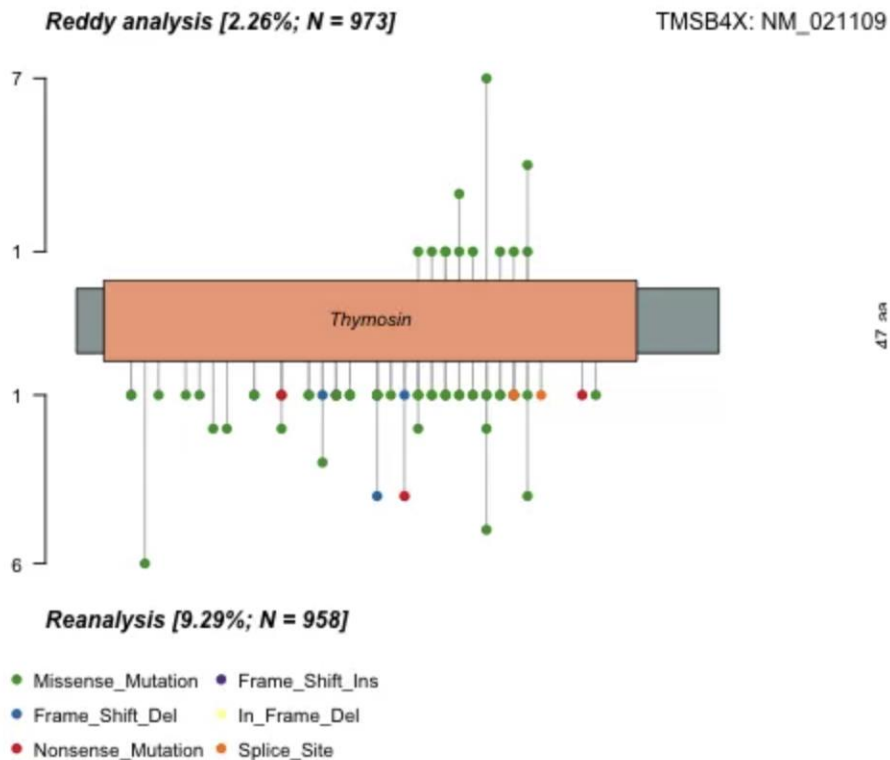
- The bioinformatics core aligned the data and sent me a list of differentially expressed genes. I'm done, right?
- We ran Mutect to call somatic mutations in this tumor genome. Let's take it to the bank



Real world consequences



Real world consequences



Real world consequences

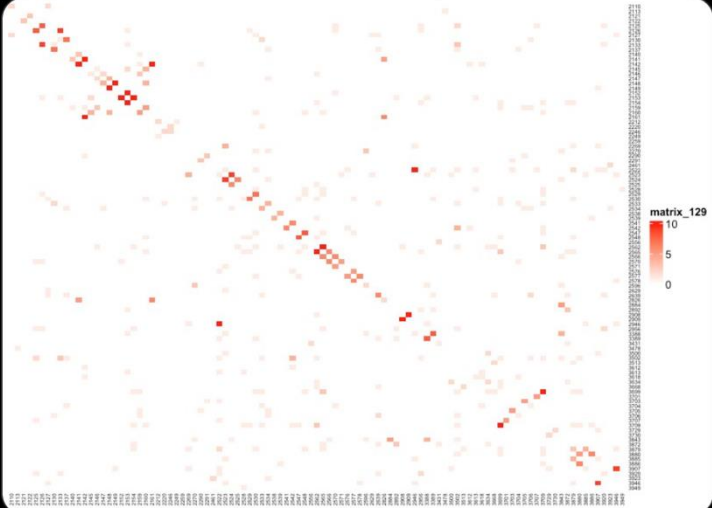
Ryan D Morin @morinryan · Oct 2

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.



matrix_129

10

5

0

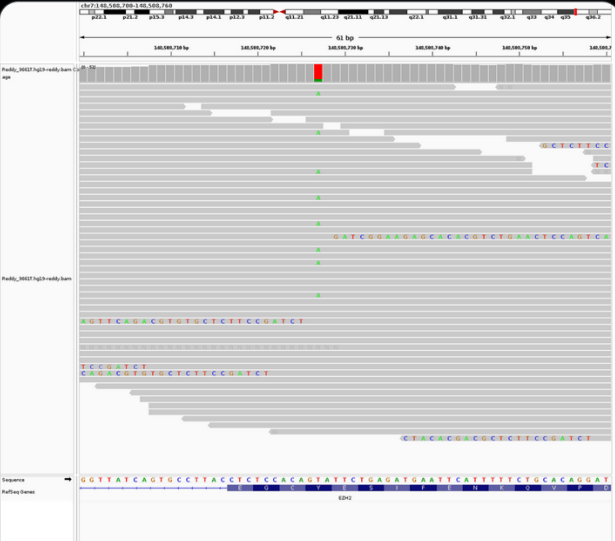
Ryan D Morin @morinryan · Nov 4

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!



chr2:1,545,585,700-1,545,585,740

1,545,585,700 bp 1,545,585,710 bp 1,545,585,720 bp 1,545,585,730 bp 1,545,585,740 bp

62 bp

Protein_36027 IgG2b-weakly bound

Protein_36027 IgG2b-weakly bound

Sequence

RefSeq Gene

IGK1

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...

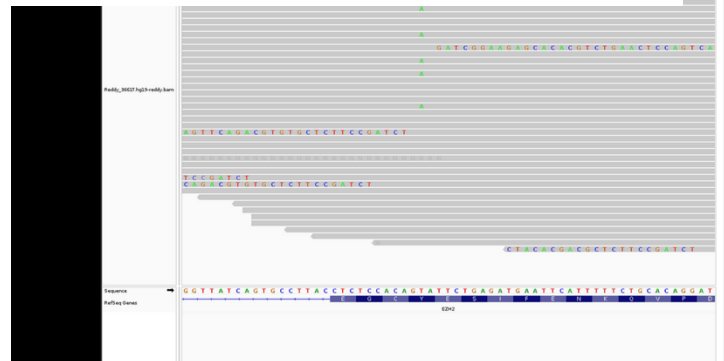
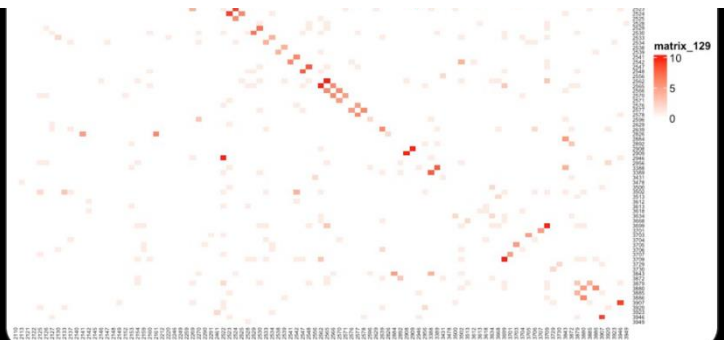


Ryan D Morin @morinryan · Nov 4

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

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)!

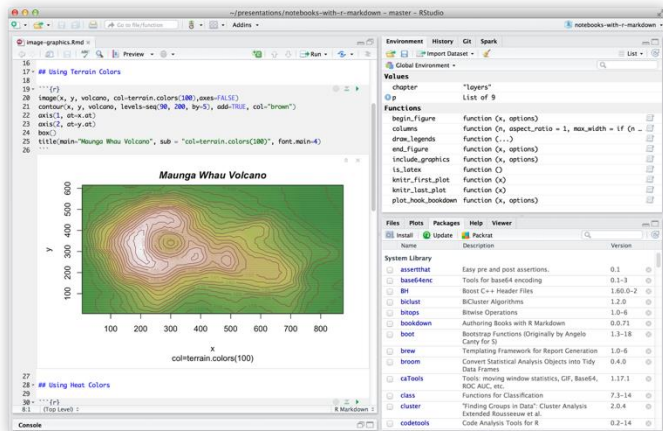
Reproducibility

- 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.

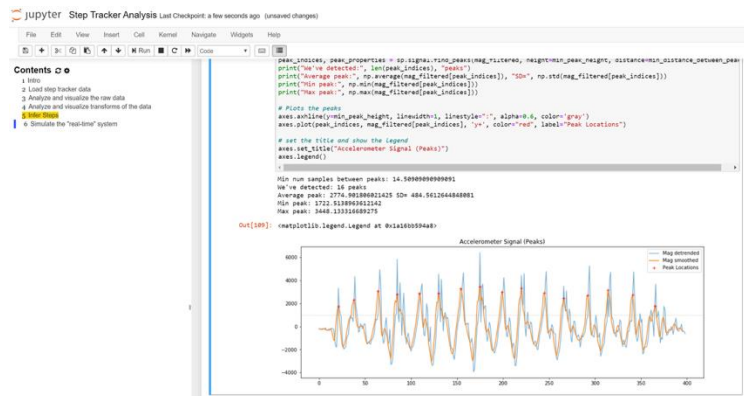
Reproducibility

- 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



Jupyter notebooks

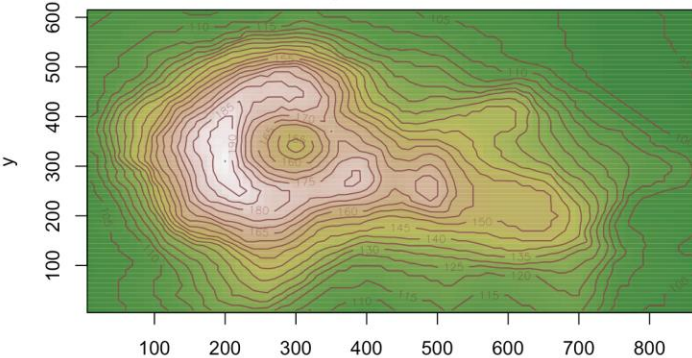


~/presentations/notebooks-with-r-markdown - master - RStudio

image-graphics.Rmd x

```
16
17 ## Using Terrain Colors
18
19 ```{r}
20 image(x, y, volcano, col=terrain.colors(100),axes=FALSE)
21 contour(x, y, volcano, levels=seq(90, 200, by=5), add=TRUE, col="brown")
22 axis(1, at=x.at)
23 axis(2, at=y.at)
24 box()
25 title(main="Maunga Whau Volcano", sub = "col=terrain.colors(100)", font.main=4)
26
27
28 ## Using Heat Colors
29
30 ```{r}
8:1 (Top Level) ▾
```

Maunga Whau Volcano



x
col=terrain.colors(100)

Environment History Git Spark

Global Environment ▾

Values

chapter	"layers"
p	List of 9

Functions

begin_figure	function (x, options)
columns	function (n, aspect_ratio = 1, max_width = if (n ...
draw_legends	function (...)
end_figure	function (x, options)
include_graphics	function (x, options)
is_latex	function ()
knitr_first_plot	function (x)
knitr_last_plot	function (x)
plot_hook_bookdown	function (x, options)

Files Plots Packages Help Viewer

Install Update Packrat

Name	Description	Version
<input type="checkbox"/> assertthat	Easy pre and post assertions.	0.1
<input type="checkbox"/> base64enc	Tools for base64 encoding	0.1-3
<input type="checkbox"/> BH	Boost C++ Header Files	1.60.0-2
<input type="checkbox"/> biclust	BiCluster Algorithms	1.2.0
<input type="checkbox"/> bitops	Bitwise Operations	1.0-6
<input type="checkbox"/> bookdown	Authoring Books with R Markdown	0.0.71
<input type="checkbox"/> boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18
<input type="checkbox"/> brew	Templating Framework for Report Generation	1.0-6
<input type="checkbox"/> broom	Convert Statistical Analysis Objects into Tidy Data Frames	0.4.0
<input type="checkbox"/> caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
<input type="checkbox"/> class	Functions for Classification	7.3-14
<input type="checkbox"/> cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.4
<input type="checkbox"/> codetools	Code Analysis Tools for R	0.2-14

and
files

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?

The screenshot displays the BioData CATALYST web application interface. The top navigation bar includes links for "Browse Data", "Documentation", "TC MILLER", and "Logout". The main header features the NIH logo and the text "BioData CATALYST Powered by Gen3". Below the header, there are tabs for "Dictionary", "Exploration" (which is active), "Discovery", "Workspace", and "Profile".

The "Exploration" tab shows a sidebar with "Data" and "File" sections. Under "Data Access", there are radio buttons for "Data with Access" (selected), "Data without Access", and "All Data". Under "Filters", there are tabs for "Harmonized Variables", "Project", and "Subject". The "Project" filter is expanded, showing a list of projects with checkboxes and counts: "topmed" (53,964), "parent" (186,592), "tutorial" (14,433), and "open_access" (3,202). The "Project Id" filter is also expanded, showing a list of project IDs with checkboxes and counts: "topmed-BioMe_HMB-" (15,074).

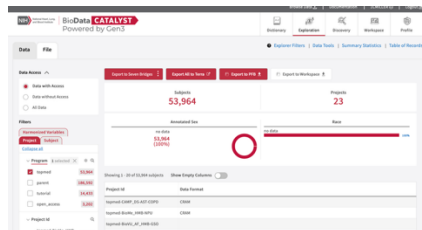
The main content area displays the following statistics:

- Subjects:** 53,964
- Projects:** 23
- Annotated Sex:** no data, 53,964 (100%)
- Race:** no data, 100%

Below the statistics, there is a table showing the first 20 of 53,964 subjects. The table has two columns: "Project Id" and "Data Format".

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?



The screenshot shows the Dockstore 'My Workflows' page for the workflow 'github.com/genome/analysis-workflows:master'. The page includes a sidebar with navigation links (Dashboard, Workflows, Tools, Services, Starred, Account, Help Desk) and a main content area. The main content area displays the workflow details, including the source code (github.com/genome/analysis-workflows:master), the last update to the workflow version (Jan. 8 2020), and the last update to the source repository (Jan. 8 2020). The workflow is currently unpublished. The page also includes a 'Publish' button and a 'Refresh' button. A notification banner at the bottom states: 'Keep your workflow automatically in sync with GitHub with our new registration process. Click [here](#) to learn more.'

Workflow Information

- Source Code: github.com/genome/analysis-workflows
- Workflow Path: /Dockstore.cwl
- Test File Path: /test.json
- Topic Automatic: Open workflow definitions for genomic analysis from MGI at WUSM.

What happens when this works well?

The screenshot displays the Terra WORKSPACES interface for the workflow `wdl_samtools.manta`. The interface is divided into several sections:

- Header:** Shows the Terra logo, "BETA WORKSPACES", and the current workspace path: `Workspaces > leyfab_terra1/ch_fusion_topmed > workflows > wdl_samtools.manta`. A "COVID-19 Data & Test" badge is visible in the top right.
- Navigation:** Tabs for DASHBOARD, DATA, ANALYSES, WORKFLOWS (active), and JOB HISTORY.
- Workflow Details:**
 - Back to list** link.
 - wdl_samtools.manta** title with an information icon.
 - Version:** A dropdown menu showing `v1.3`.
 - Source:** `github.com/saimukund20/wdl_samtools.manta:v1.3`.
 - Synopsis:** "No documentation provided".
 - Run options:** Radio buttons for "Run workflow with inputs defined by file paths" and "Run workflow(s) with inputs defined by data table" (selected).
 - Step 1:** "Select root entity type:" with a dropdown menu showing `pharmu_set`.
 - Step 2:** A "SELECT DATA" button and the text "No data selected".
 - Options:** Checkboxes for "Use call caching" (checked), "Delete intermediate outputs", "Use reference disks", "Retry with more memory", and "Ignore empty outputs".
 - Buttons:** "SCRIPT", "INPUTS" (active), "OUTPUTS", and "RUN ANALYSIS".
- Inputs Table:** A table titled "Hide optional inputs" with columns "Task name", "Variable", "Type", and "Attribute". It lists three inputs for the `wf` task:

Task name	Variable	Type	Attribute
wf	fusion_sites	File	"gs://fc-6248b026-3011-4591-9688-255a248b35b9/sites_merged_with_solid_tumor.bed" [icon] [...]
wf	manta_config	File	"gs://fc-6248b026-3011-4591-9688-255a248b35b9/configManta_1.pyini" [icon] [...]
wf	reference	File	workspace.referenceData_hg38_ref_fasta [icon] [...]

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?

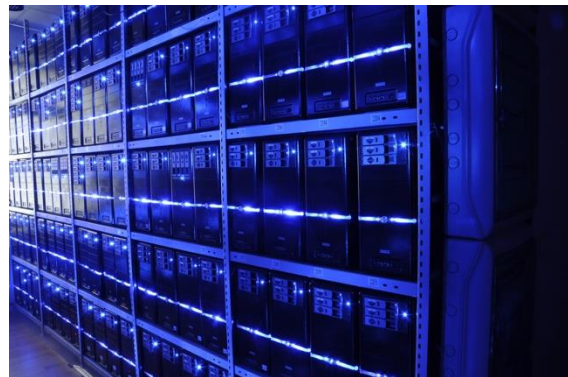
- Big desktop machine or blade in your lab
 - Pro: moderate amounts of CPUs/RAM for big jobs
 - Pro: You have priority access
 - Pro: Can submit jobs and walk away
 - Con: You have to become a sysadmin and take care of it.
(Who applies security updates? What happens if the power supply fails? Backups?)



Infrastructure – where do I analyze my data?

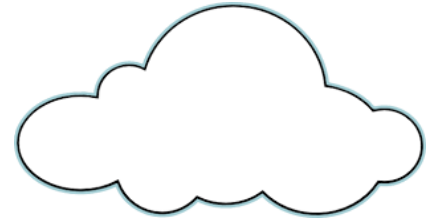
- Local Compute Cluster

- 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!



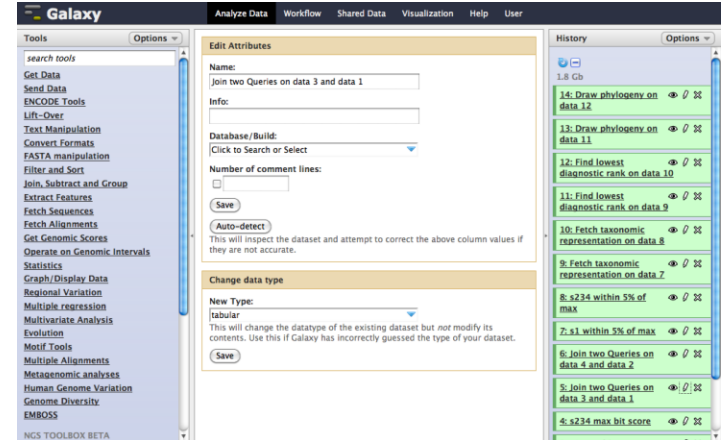
Infrastructure – where do I analyze my data?

- 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



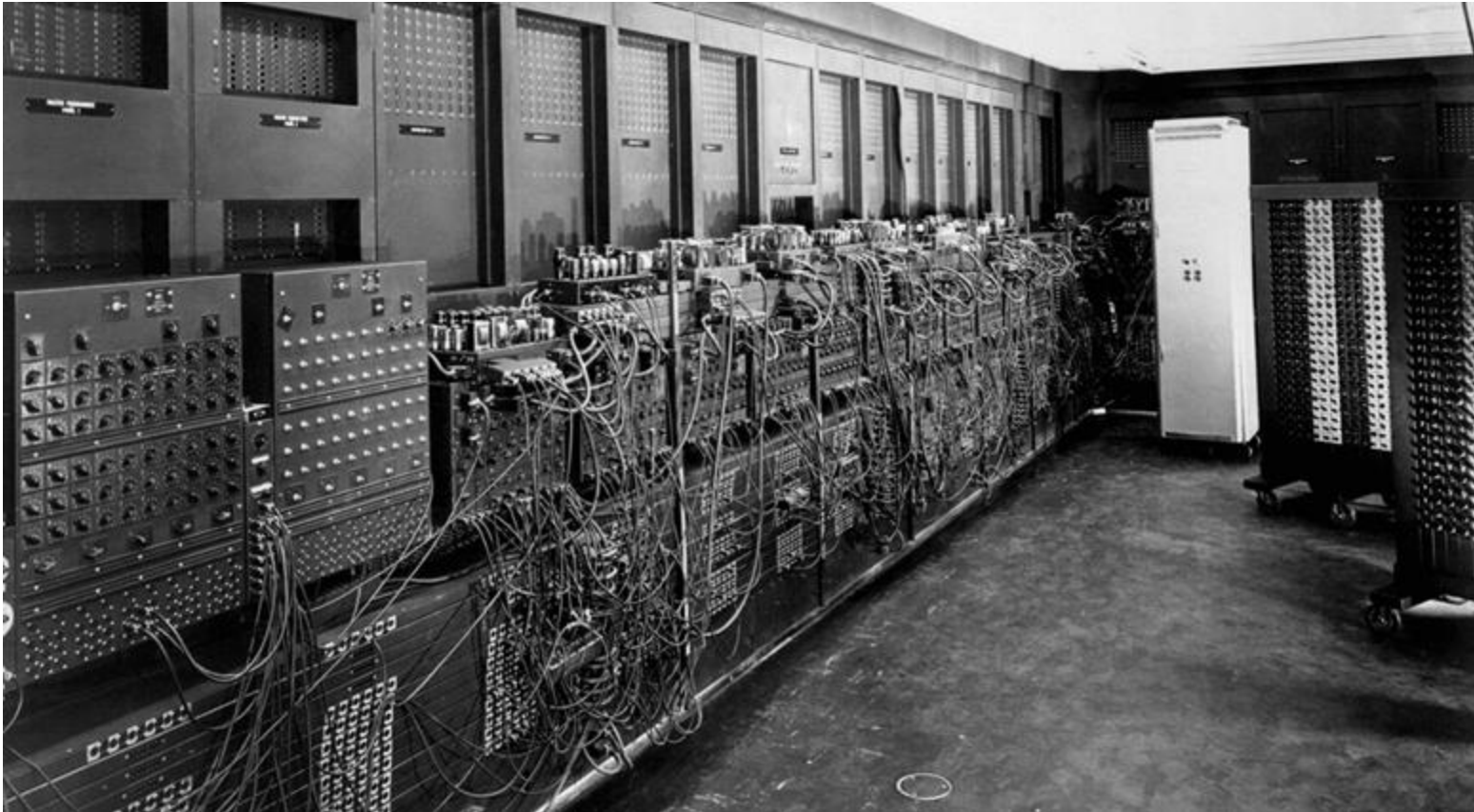
Infrastructure

- 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

9/2

9/9

0800 Anttan started
1000 " stopped - anttan ✓
1300 (032) MP-MC ~~1.30476415~~ 2.130476415
(033) PRO 2 2.130476415
correct 2.130676415

Relays 6-2 in 033 failed special speed test
in relay " 11.000 test.

Relay
2145
Relay 1337

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

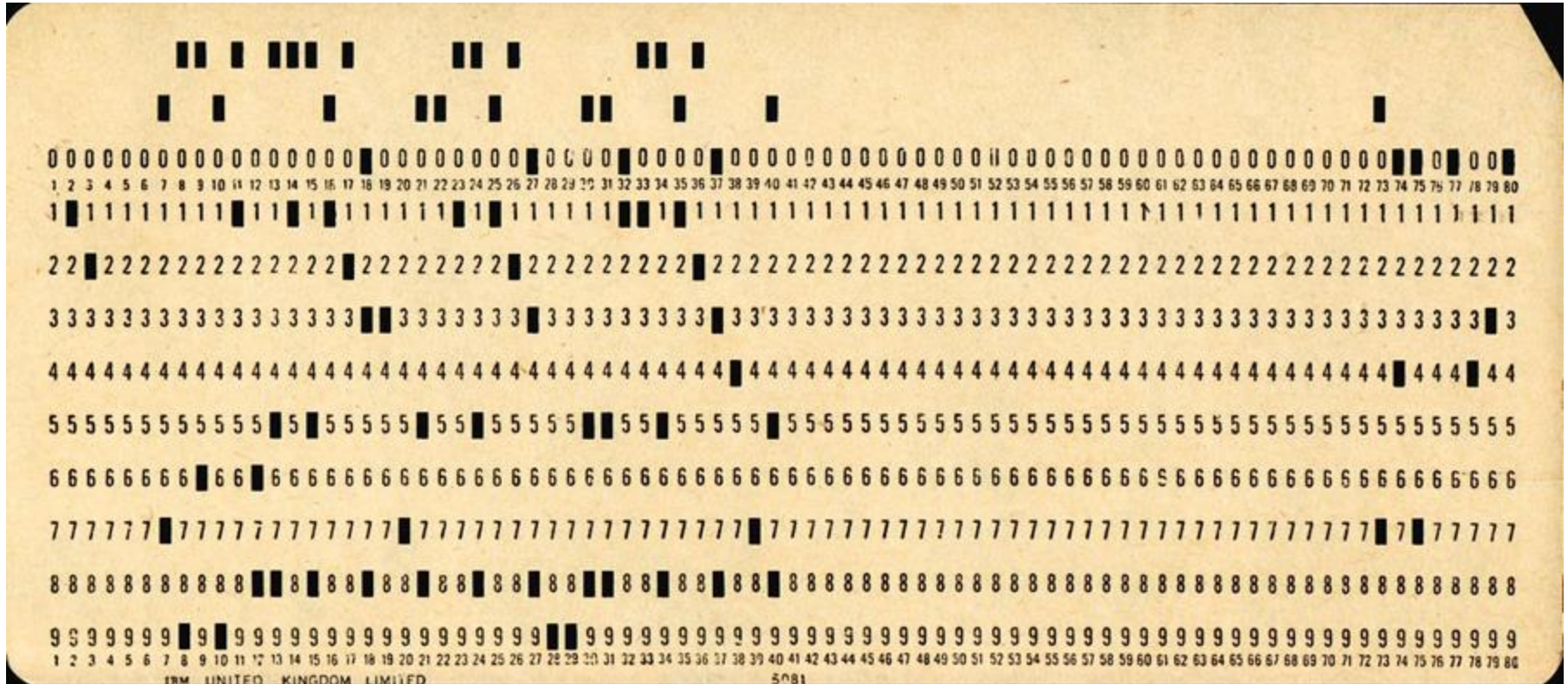
1545



Relay #70 Panel F
(moth) in relay.

First actual case of bug being found.
1630 Anttan started.
1700 closed down.

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

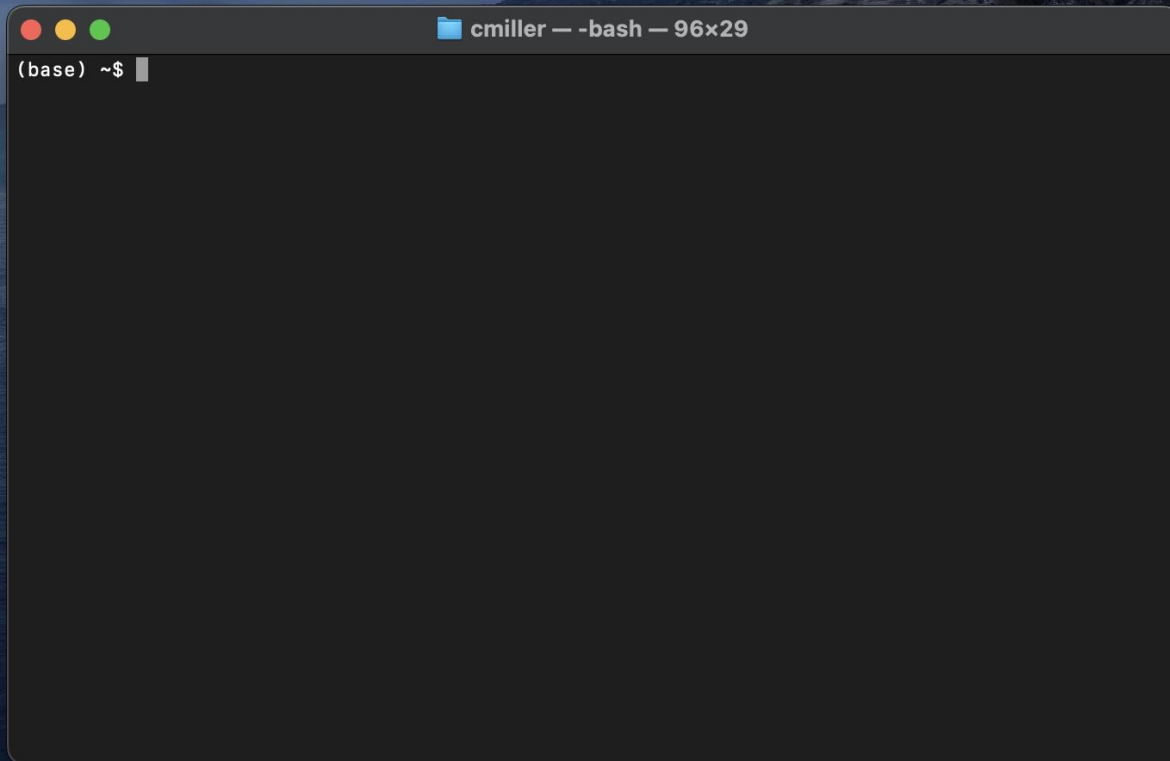
- The big event had to be postponed due to COVID and now we have to change every instance of "Apr 2020" to "Oct 2024". Problem is, there's a huge nested set of directories containing over 10,000 files!
- Clicking around in Windows explorer is not going to get the job done
- On a Unix system, that's just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

