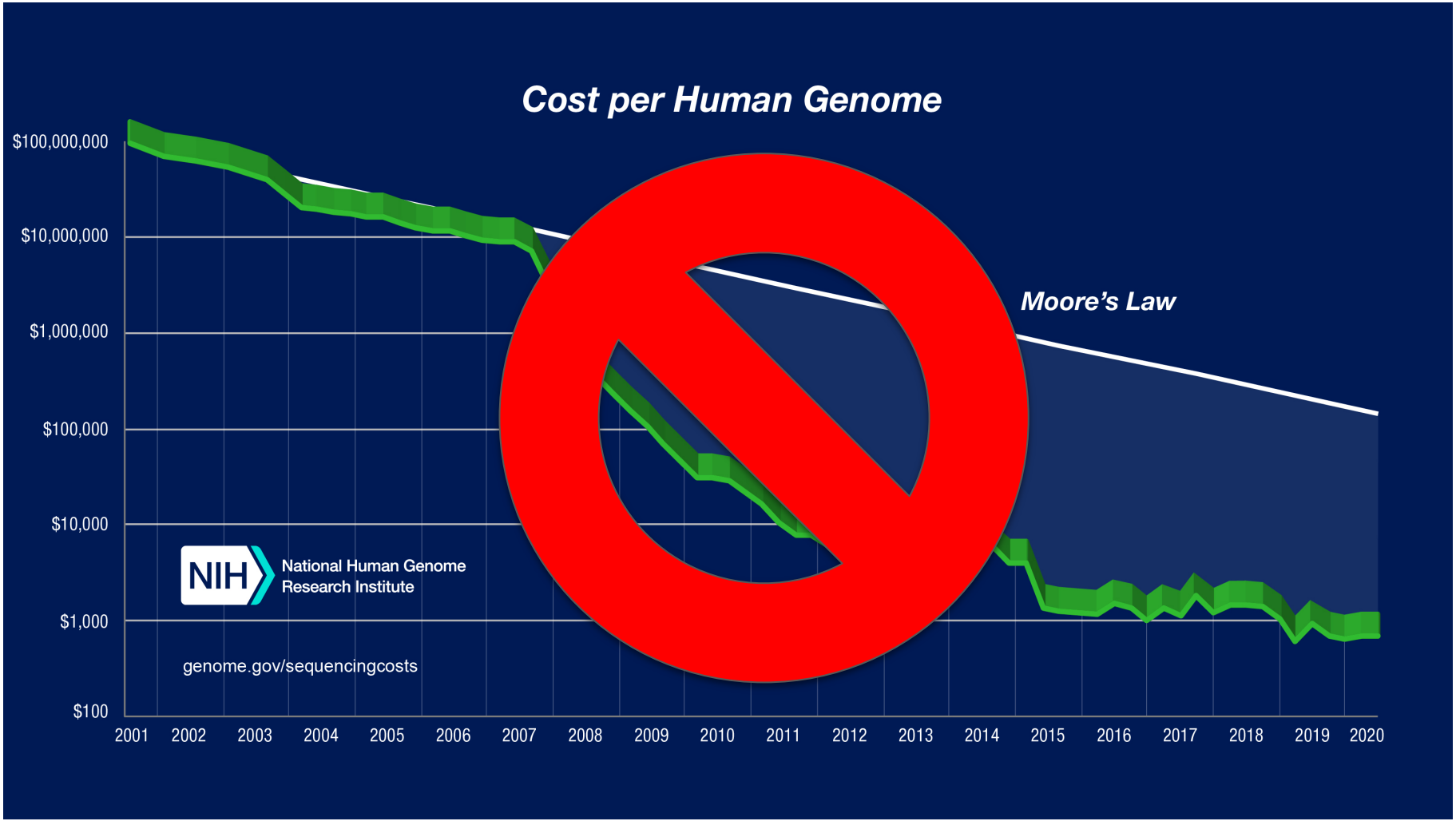
Introduction to Bioinformatics

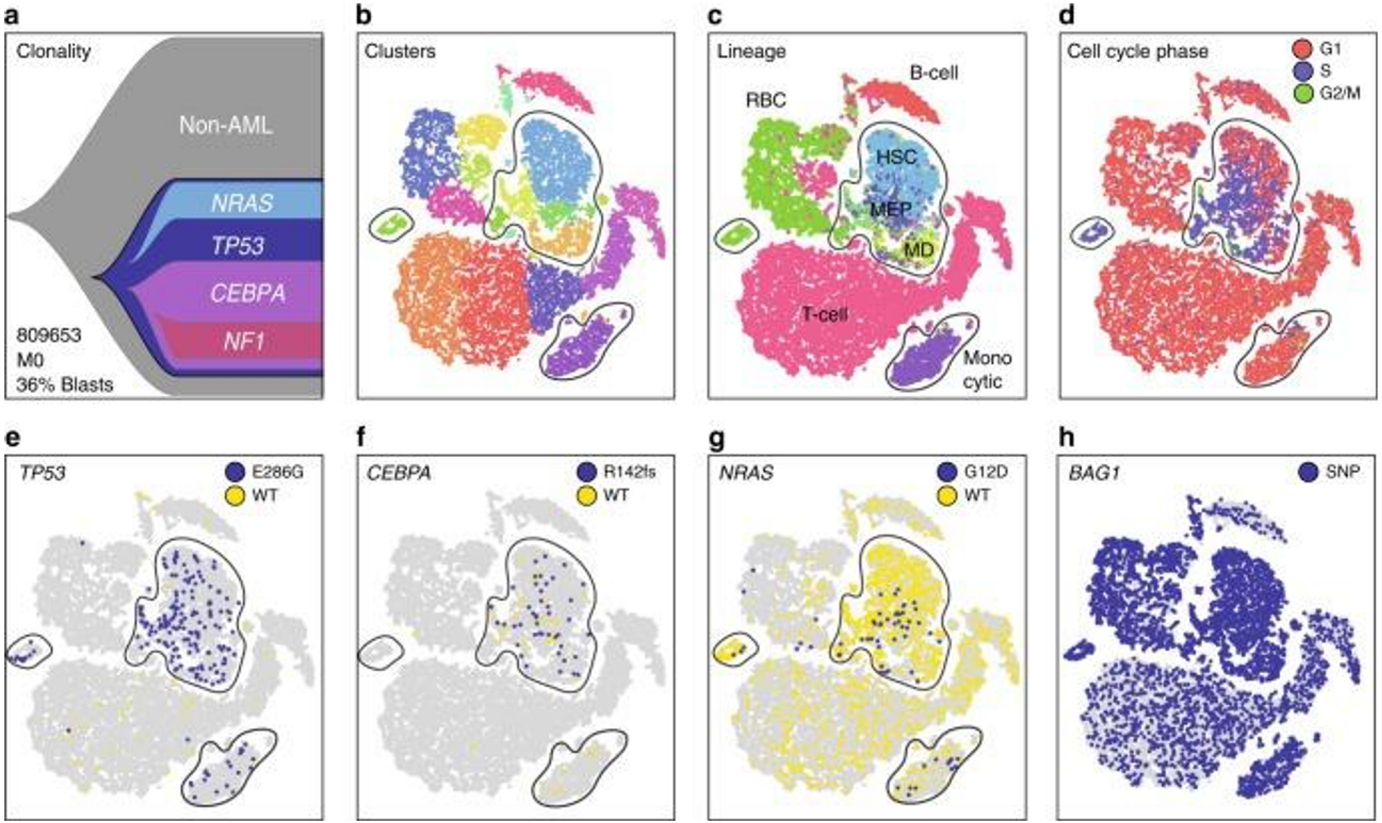
Chris Miller, Ph.D.
Washington University in St. Louis

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
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts


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
- 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

Algorithm design

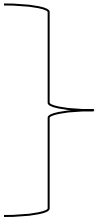
Building Pipelines

Developing Assays

Analysis of my experiment

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

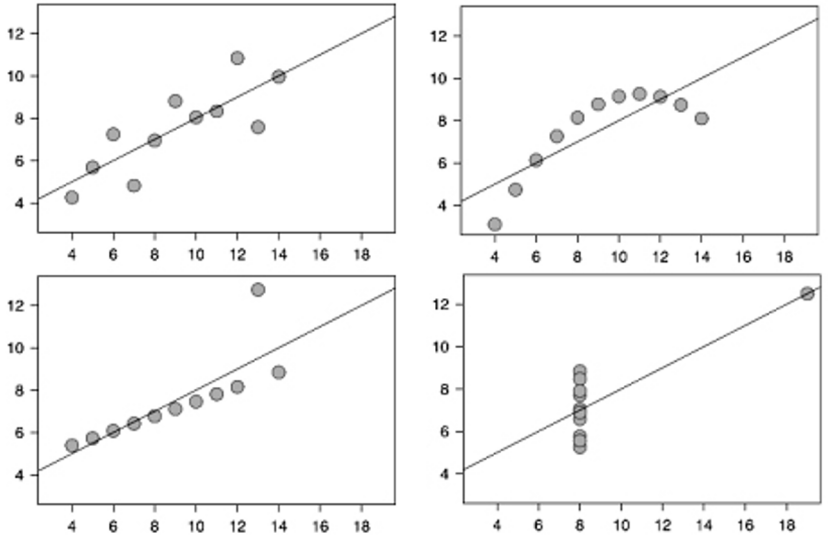
Course structure

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

Don’t trust your data

Trusting your data

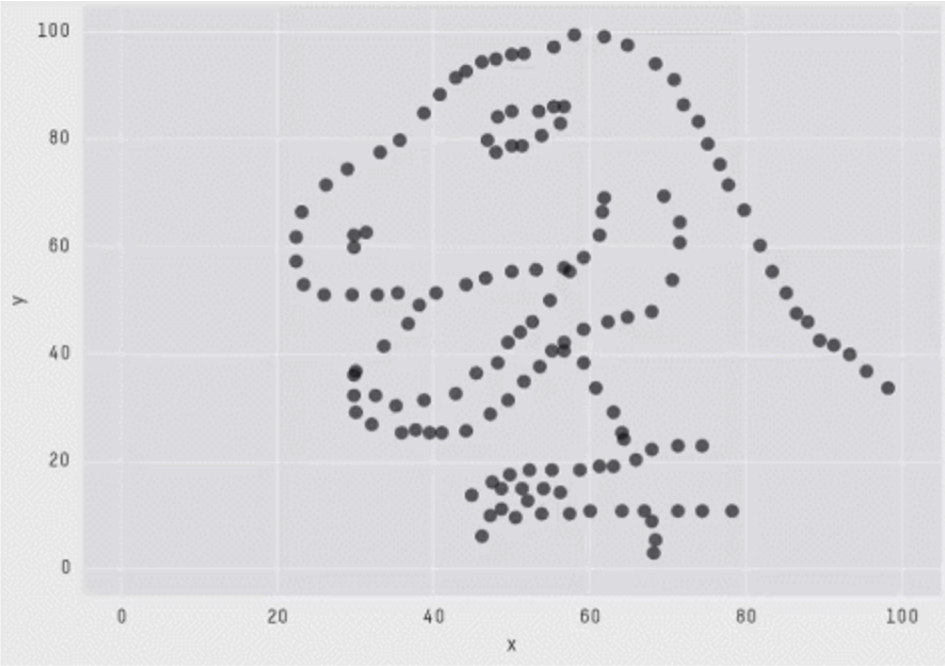
Anscombe's quartet



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

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

Datasaurus Dozen



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

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"

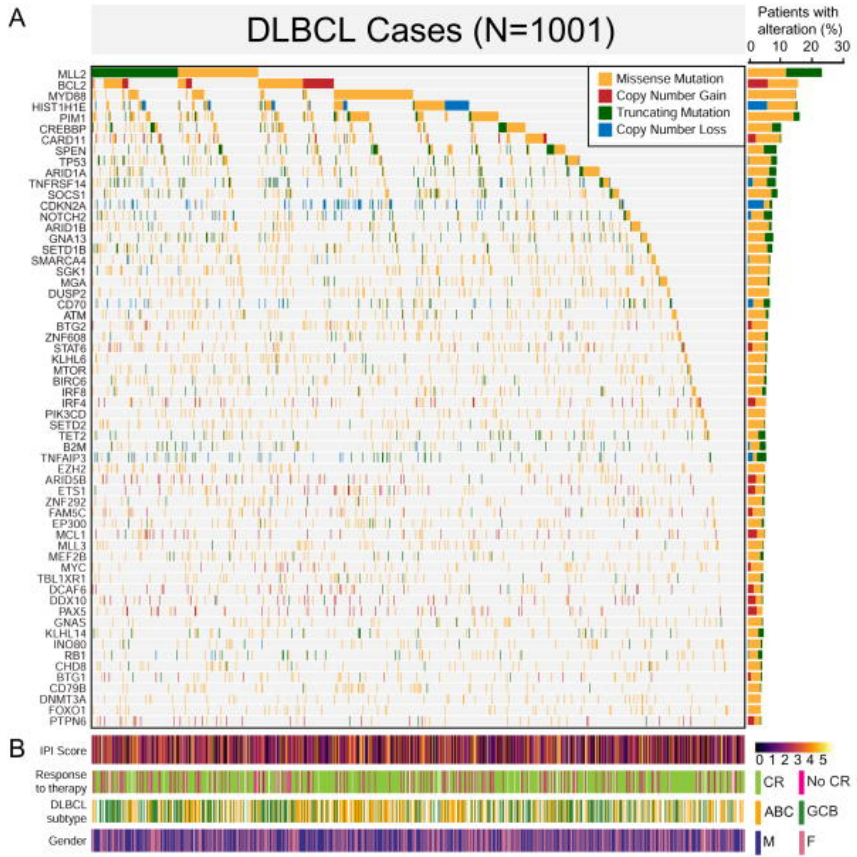
- Ernest Rutherford

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?

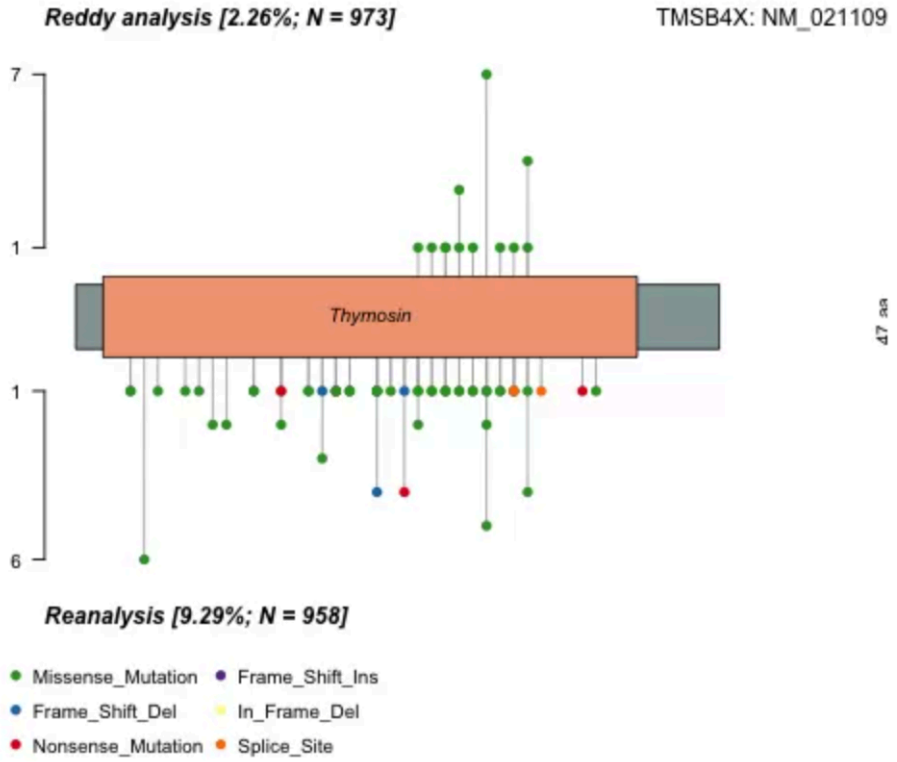
- We ran Mutect to call somatic mutations in this tumor genome. Let’s take it to the bank

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

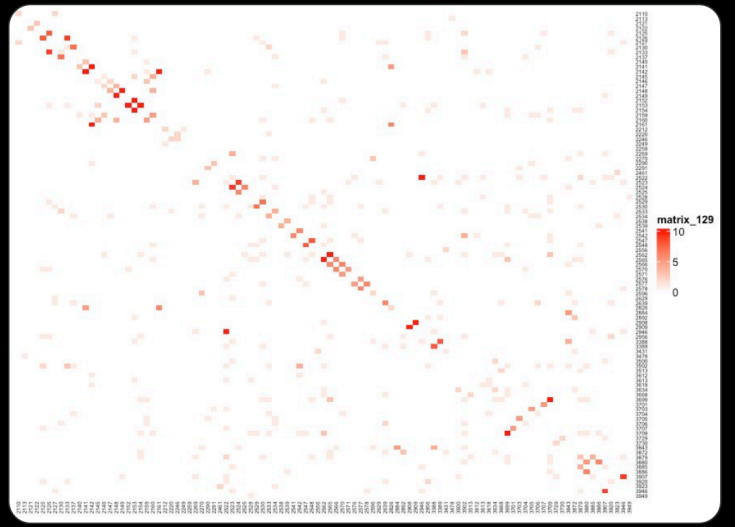
Ryan D Morin @morinryan · Oct 2

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.

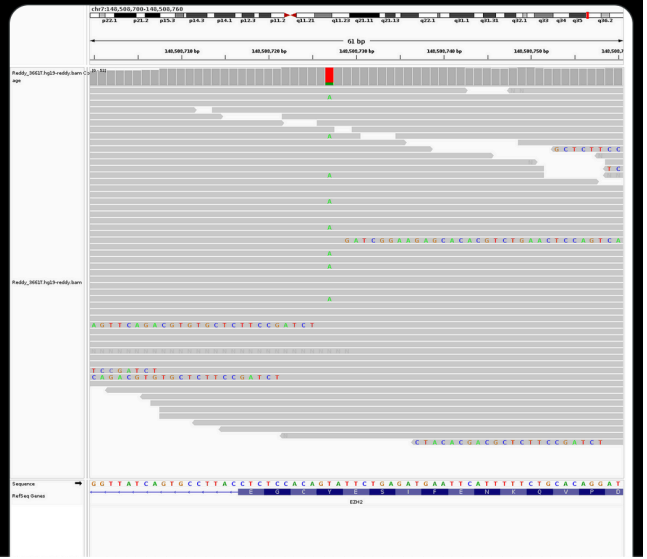


Ryan D Morin @morinryan · Nov 4

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

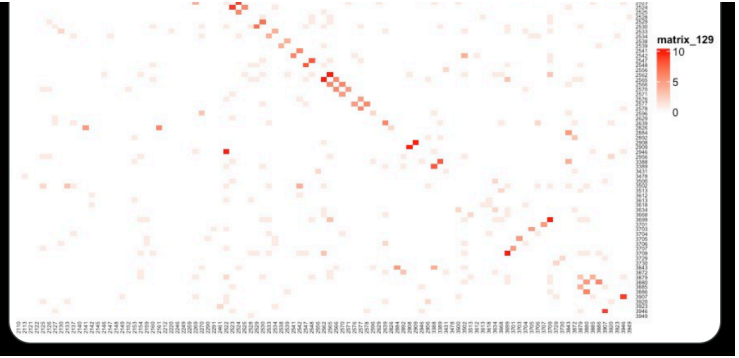
Ryan D Morin @morinryan · Oct 2

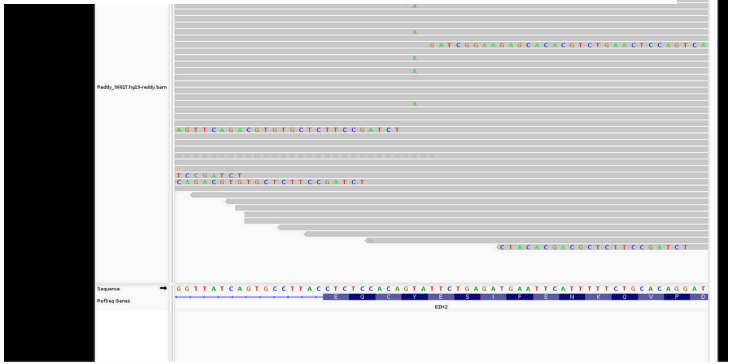
RNA/DNA mismatches (sample swaps) affecting at least 10% of the patients in Reddy et al, a Cell paper with over 700 citations. Same issue was described in a more recent paper from this group. #lymphoma #genomics #goodresearchpractice pubpeer.com/publications/E...

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

Reproducibility

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

sample637.tsv	Mouse_TP53_WT_637.tsv
sample647.tsv	Mouse_TP53_KO_647.tsv
sample983.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

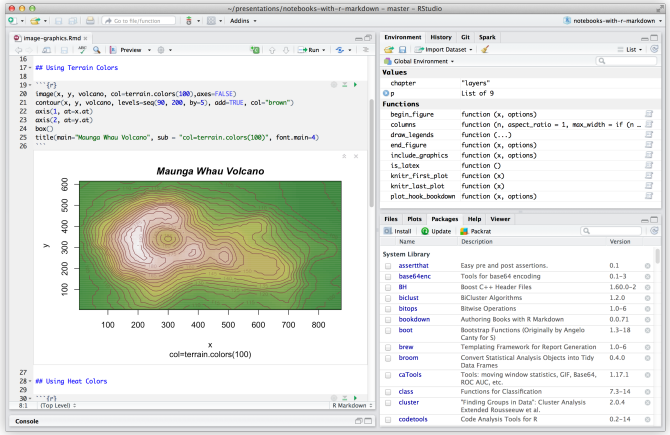
- Hell is other people's data.
  Hell is also your own data 6 months later.

Reproducibility

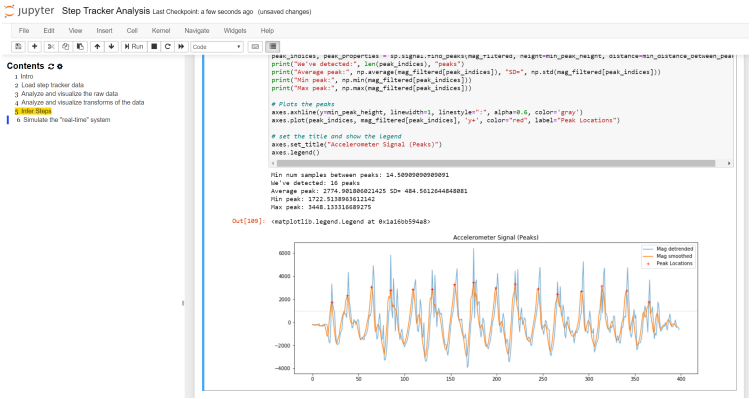
- Your tools/parameters/settings should all be stored in scripts

- 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



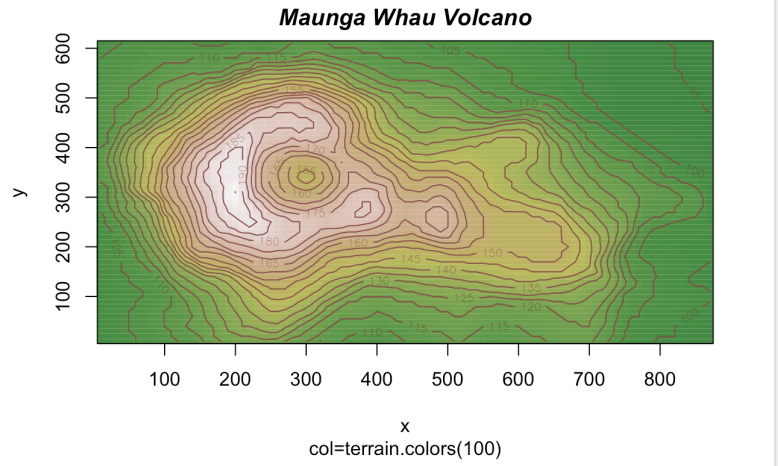
~/presentations/notebooks-with-r-markdown – master – RStudio

Go to file/function | Addins | notebooks-with-r-markdown

image-graphics.Rmd

Preview | Run

```
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) | R Markdown

Console

Environment | History | Git | Spark

Import Dataset | List

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

a and

les

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 plan
- Covers more than just sequence data – gel images, textual qPCR readouts, flow plots/data, etc
- Needs to go into a repository

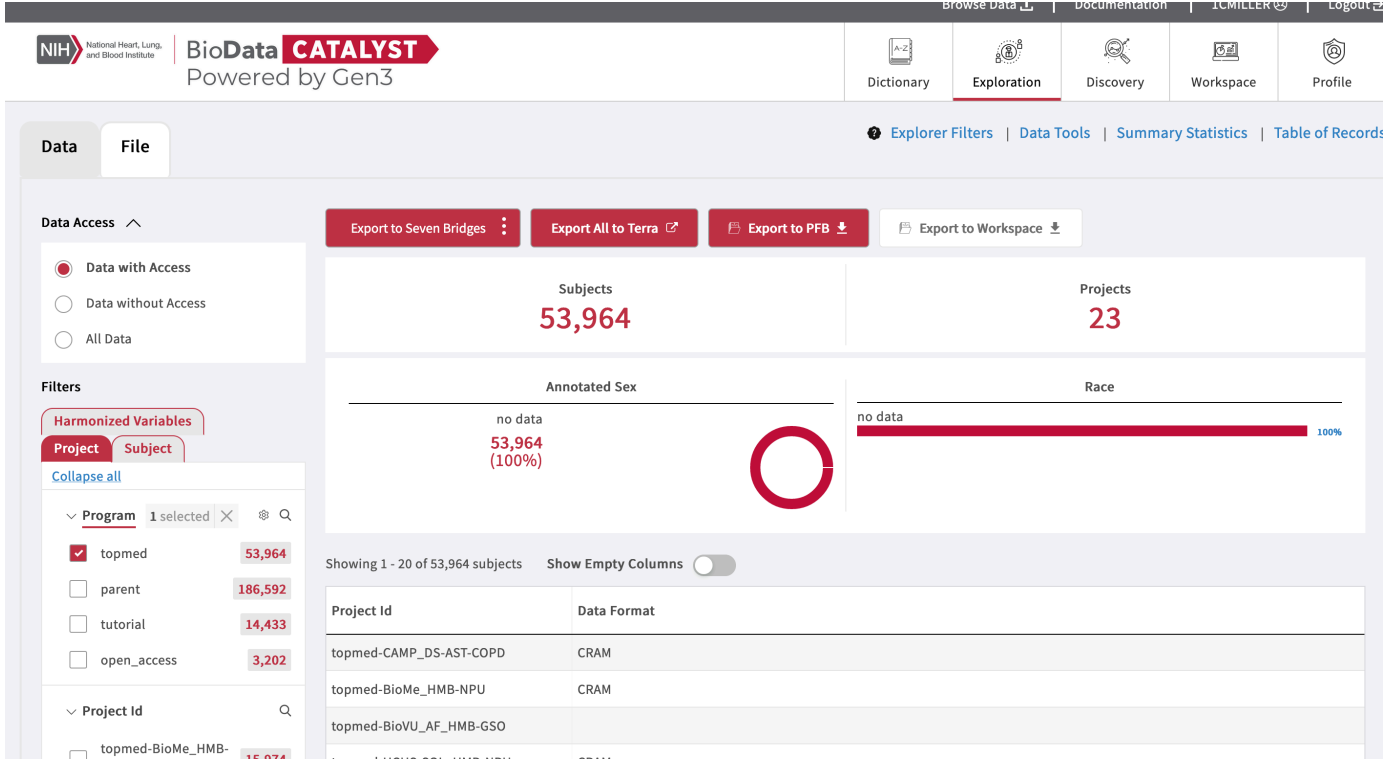
FAIR principles of data

- Findable
- Accessible
- Interoperable
- Reproducible

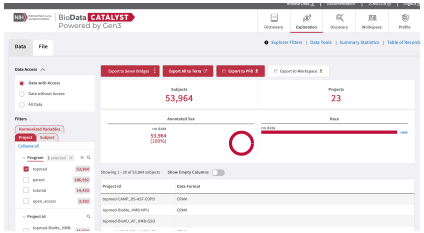
Examples of good places to deposit data

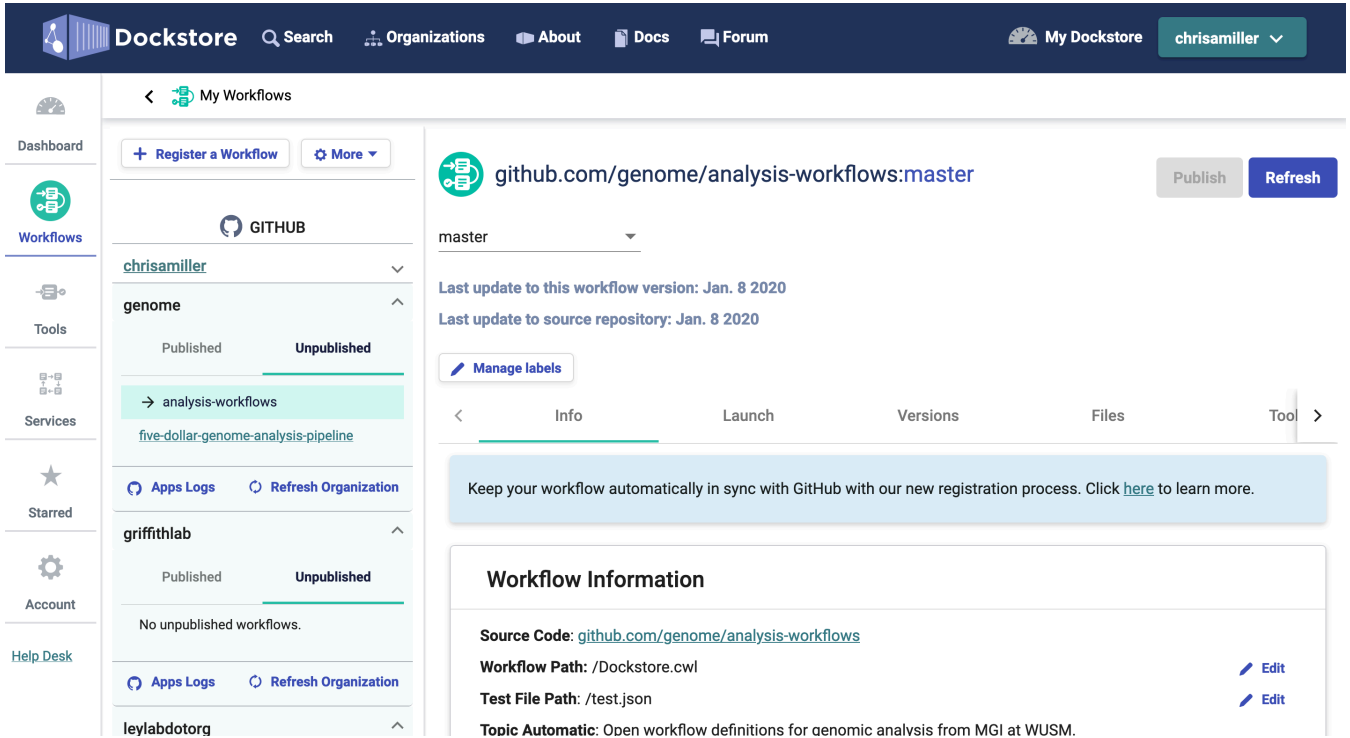
- NCBI repositories
  - SRA – Short Read Archive
  - dbGaP – front-end/access control for human data in SRA
  - GEO – rich metadata associated with experiments (RNA, scRNA, arrays, etc)
- Organism specific repos (Flybase, etc)
- General data repos that assign a doi (zenodo.org)
- Institutional repositories – 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?

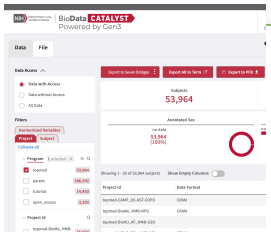


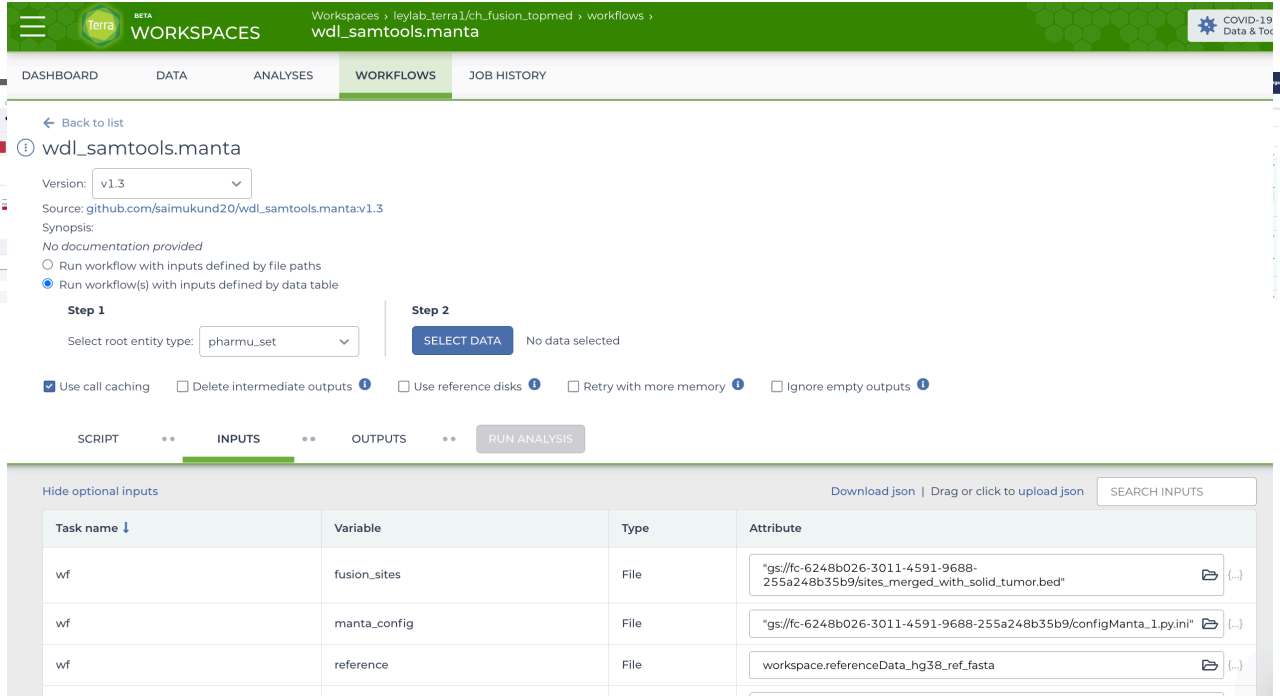
What happens when this works well?

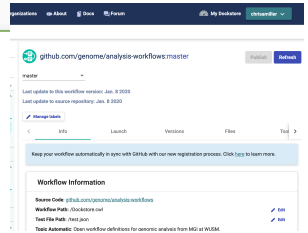




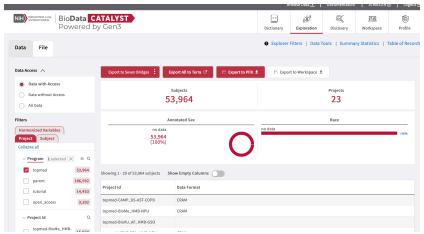
What happens when this works well?

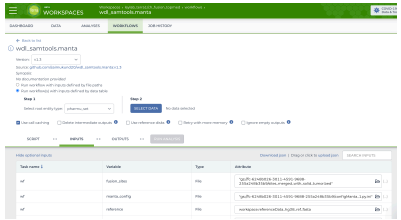


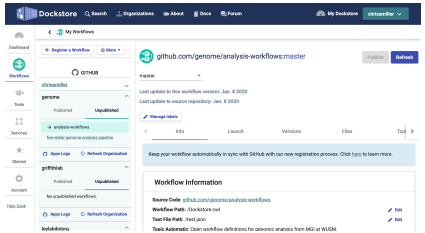




What happens when this works well?







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?)

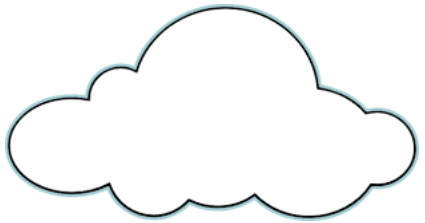
Infrastructure – where do I analyze my data?

- 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?

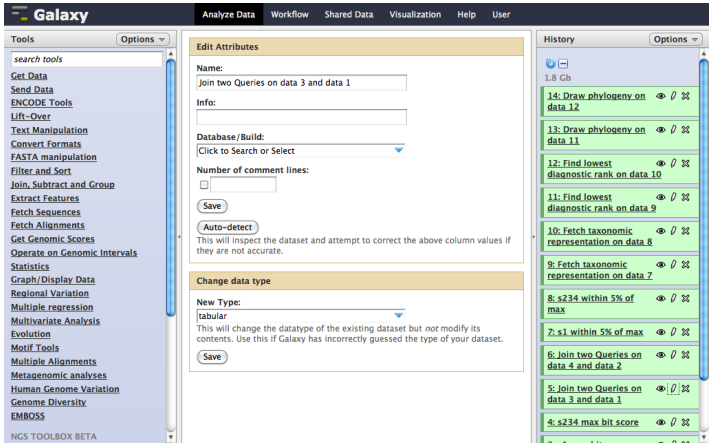
- Cloud (remote compute cluster)
  - Pro: As much CPU/RAM as you can imagine
  - Pro: secure disk, backups, etc
  - Pro: no reasonable limits on access



  - 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"

What is Unix?

Definition 1: Unix is not an acronym; it is a pun on "Multics". Multics was a large multi-user operating system that was being developed at Bell Labs shortly before Unix was created in the early '70s. Brian Kernighan is credited with the name.

Definition 2: Where computational genomics is done.

Definition 3: Your dear friend.

Recommended reading: “The Evolution of the Unix Time-sharing system”, Dennis M. Ritchie
https://pdfs.semanticscholar.org/f64f/6e66da16e93ebf4221fc8915b2420fd56b66.pdf

Unix history

- File system, command interpreter (shell), and process management started by Ken Thompson
- Device files and further dev from Dennis Ritchie, as well as McIlroy and Ossanna (to a lesser degree)
- **Vast array of simple, dependable tools that each do one simple task.**
- By combining these tools, one can conduct rather sophisticated analyses
- Wildly popular platform for high performance computing. Supports parallelism.
- SunOS/Solaris, IBM's AIX, Hewlett-Packard HP-UX, OSX, Linux, Android, etc.

Ken Thompson (sitting) and Dennis Ritchie working together at a PDP-11

Credit:https://en.wikipedia.org/wiki/History_of_Unix

Why Unix?

- **Power!**

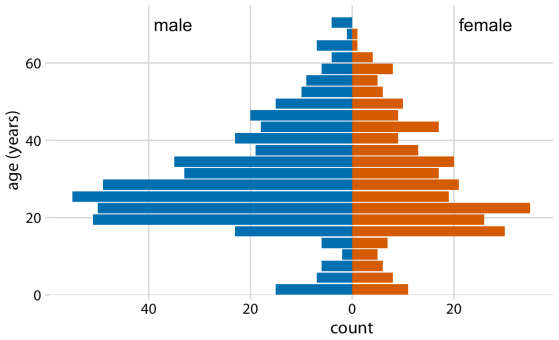
- The big event had to be postponed due to COVID and now we have to change every instance of "Apr 2020" to "Oct 2022". Problem is, there's a huge nested set of directories containing over 10,000 files!

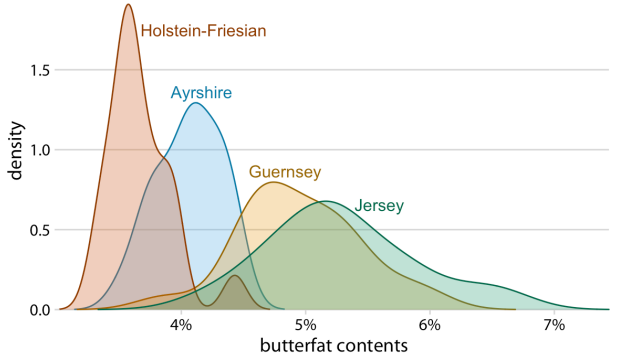
- Clicking around in Windows explorer is not going to get the job done

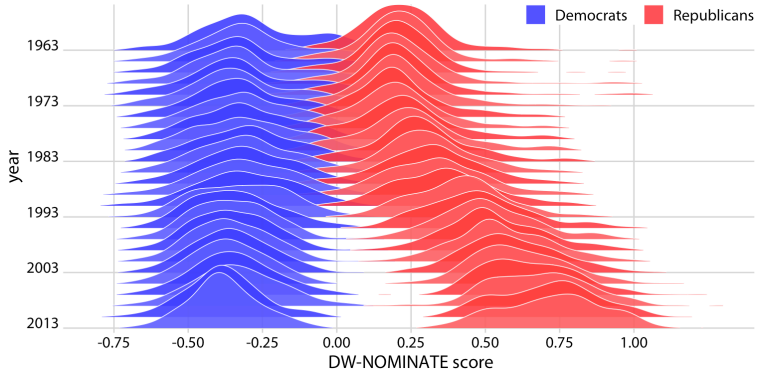
- On a Unix system, that’s just one short line of code:

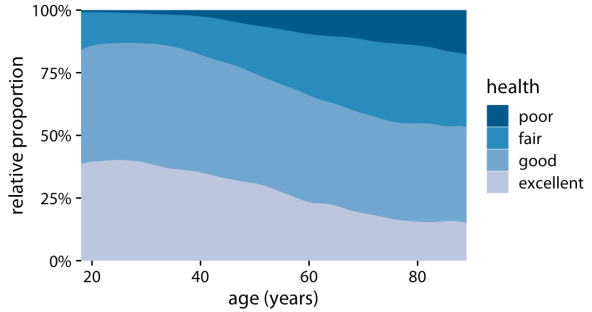
```
find .  -name "*.txt" | xargs –n 1 sed -i.bak 's/Apr 2020/Oct 2022/g'
```

Turning data into insight









Course structure

- Command-line basics to get you up to speed
- Generation of sequencing data, formats, alignment
- 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
- Bedtools/genome arithmetic
- Variant calling and interpretation
- Statistics and probability
- Single-cell epigenomics, ATAC-seq