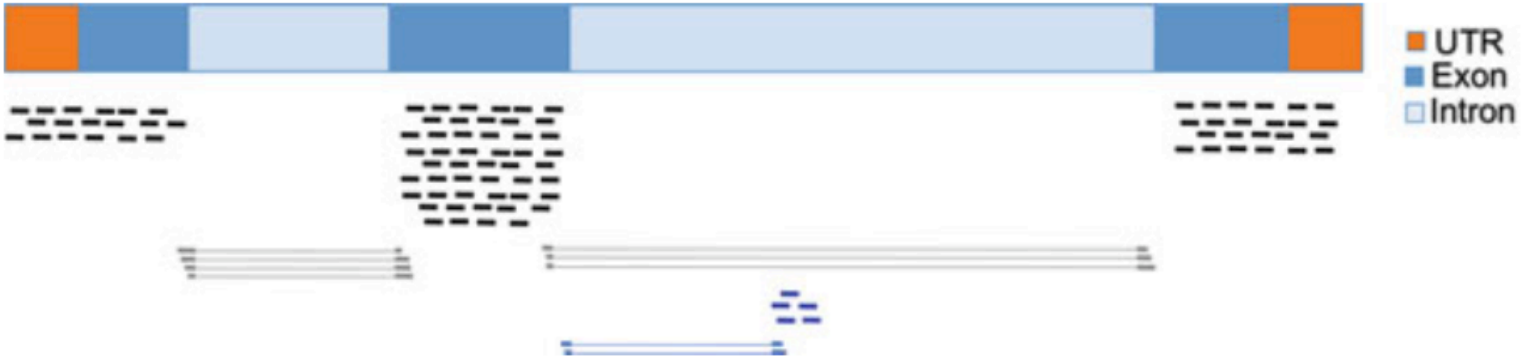
Counting Reads per Gene

- From read alignments, you next need to summarize the reads into a "table of counts"
- From:

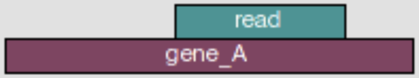
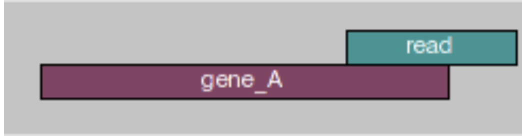
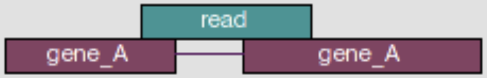
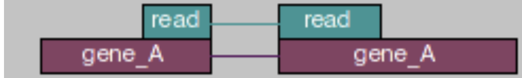
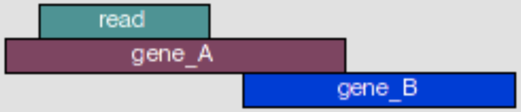
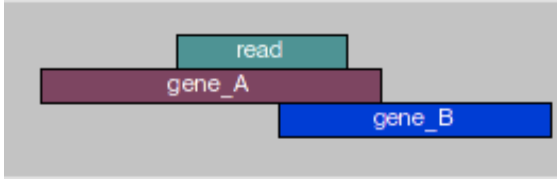
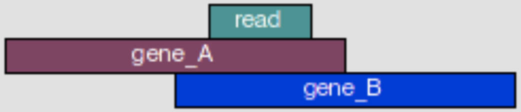


- To:

Gene MLO7 has 100 mapped reads.

HTSeq

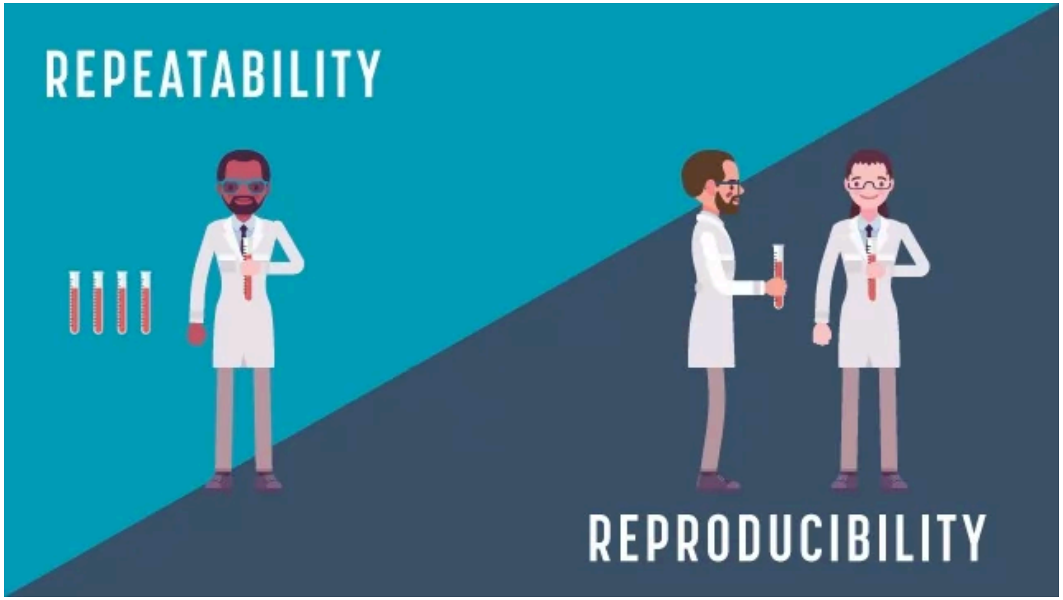
- A Python framework to work with high-throughput sequencing data
- Comes with some very useful scripts, including one to count aligned reads
- Allows users to select from a number of counting strategies
- Can select any feature type from a gff3 file to be the “counting unit”
- Categorizes reads:
  - Maps to a feature
  - Does not map to a feature
  - Is ambiguous (could map to more than one feature)
  - Is too low quality
  - Is not aligned at all
  - Alignment is not unique

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Reproducible Science (and Scripting)

Reproducibility

- Reproducibility is the ability to duplicate an entire experiment or study, either by the same researcher or by someone else working independently.
- Also called replication
- Reproducibility is one of the main principles of the scientific method.
- Scientific validity = Independent replication of experimental results

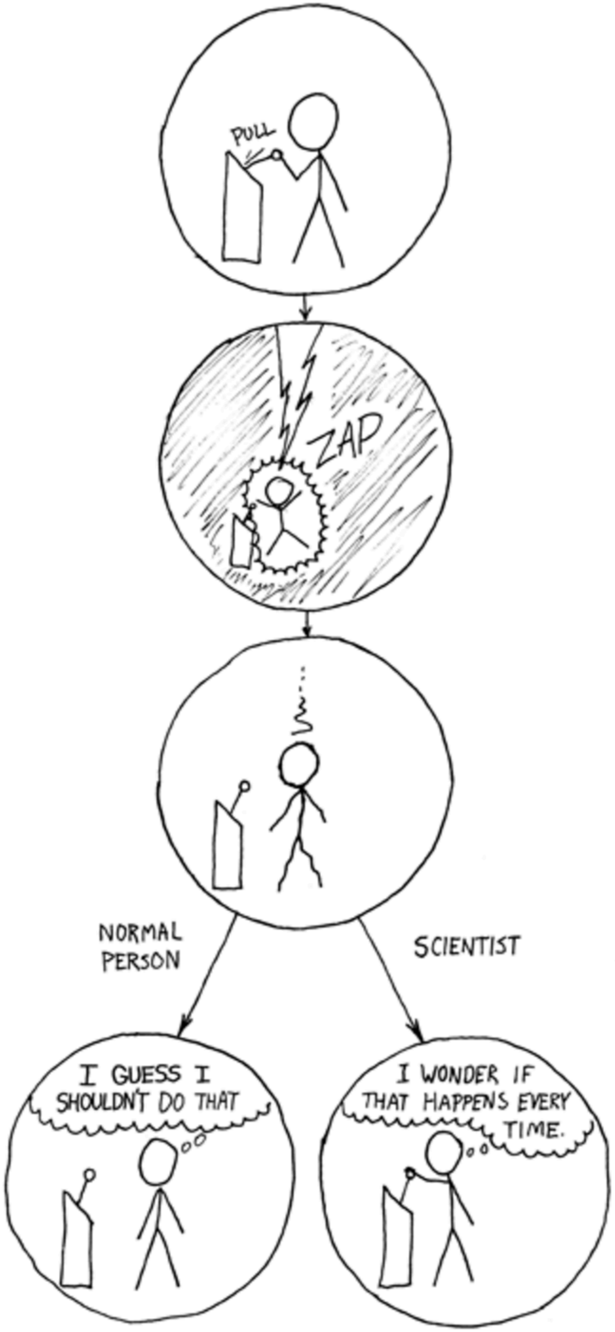


https://www.technologynetworks.com/informatics/articles/repeatability-vs-reproducibility-317157

How do I make my analysis reproducible?

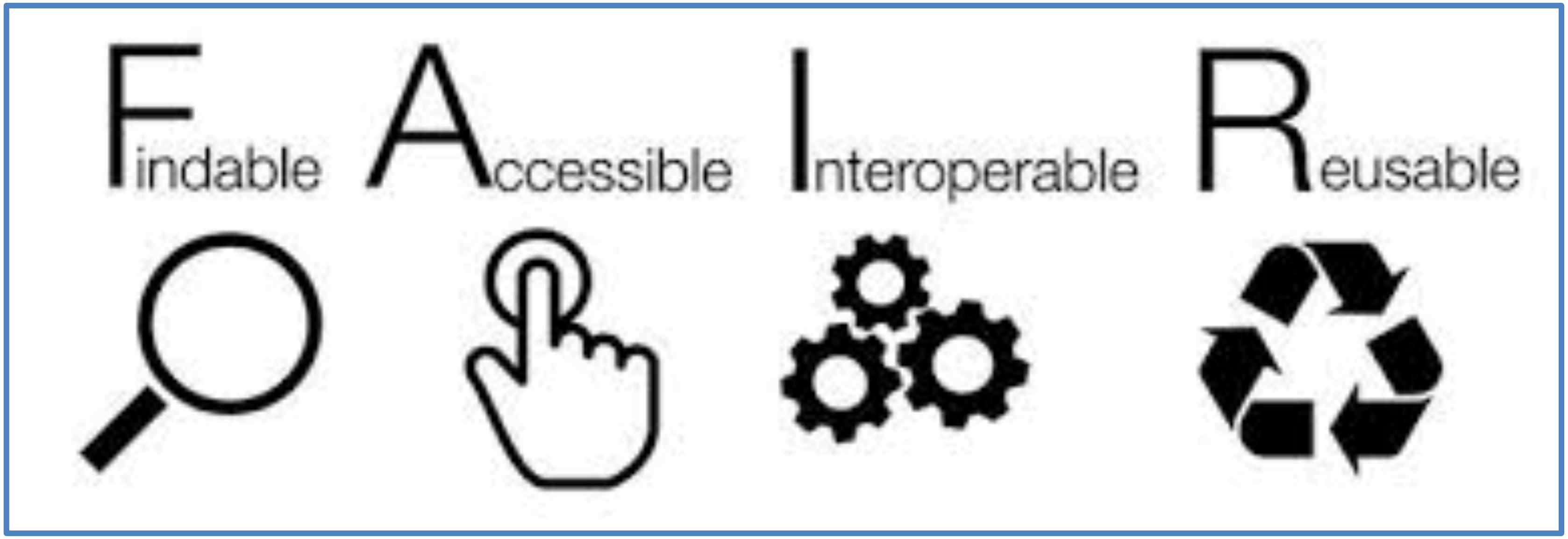
In bioinformatics, this requires sharing code, analysis details and raw data

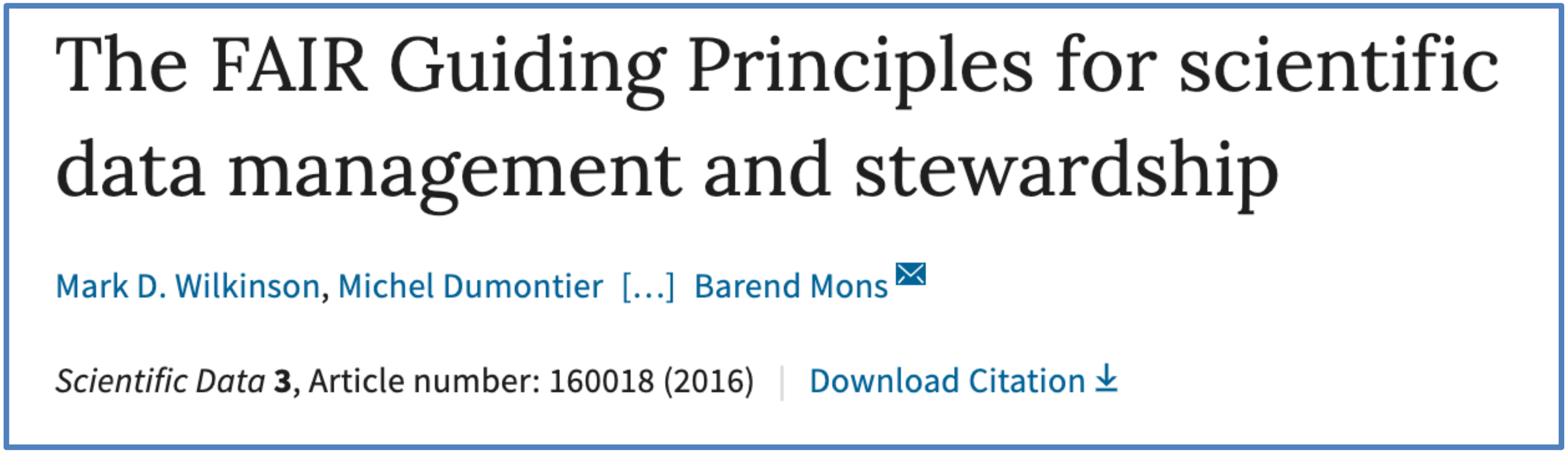
- Data sharing is almost always required by peer-reviewed journals
- Code/analysis sharing is less standardized, but coming soon


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

Randall Munroe, xkcd

FAIR Principles





The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons ✉

Scientific Data **3**, Article number: 160018 (2016) | Download Citation ⤓

Robustness

- In wet lab research, it is often obvious when an experiment fails.
- This is not always true with computational analysis. Software may not print an error, even when things have gone wrong.
- Fewer users = fewer bug reports
- High dimensional data is complex and difficult to have a priori expectations

Robustness

- Never trust your data or tools – always verify in whatever way possible
- Look at results at each intermediate step
- Visualize output in the most meaningful way possible
- Use good controls and examine them in comparison to treatments often

How do I do reproducible and robust research?

- Record keeping – keep a (preferably online) lab notebook for the dry lab
- Scripting to automate tasks
- Write simple, clear scripts that you and others can read later
- Save raw data
- Publicly release raw data and scripts

github
SOCIAL CODING

DRYAD

What is a script?

- A computer program
- Usually somewhat small and dedicated to a single task or just a few tasks
- Automates the execution of tasks
  - In other words…
  - A person could run each of these commands one by one
  - But instead the commands have been written down and given to the computer to run one by one
- **This is also a great form of documentation!!!**

What is a script?

- Can be in any of many different languages:
  - Python
  - Perl
  - R
  - Bash
- Bash???

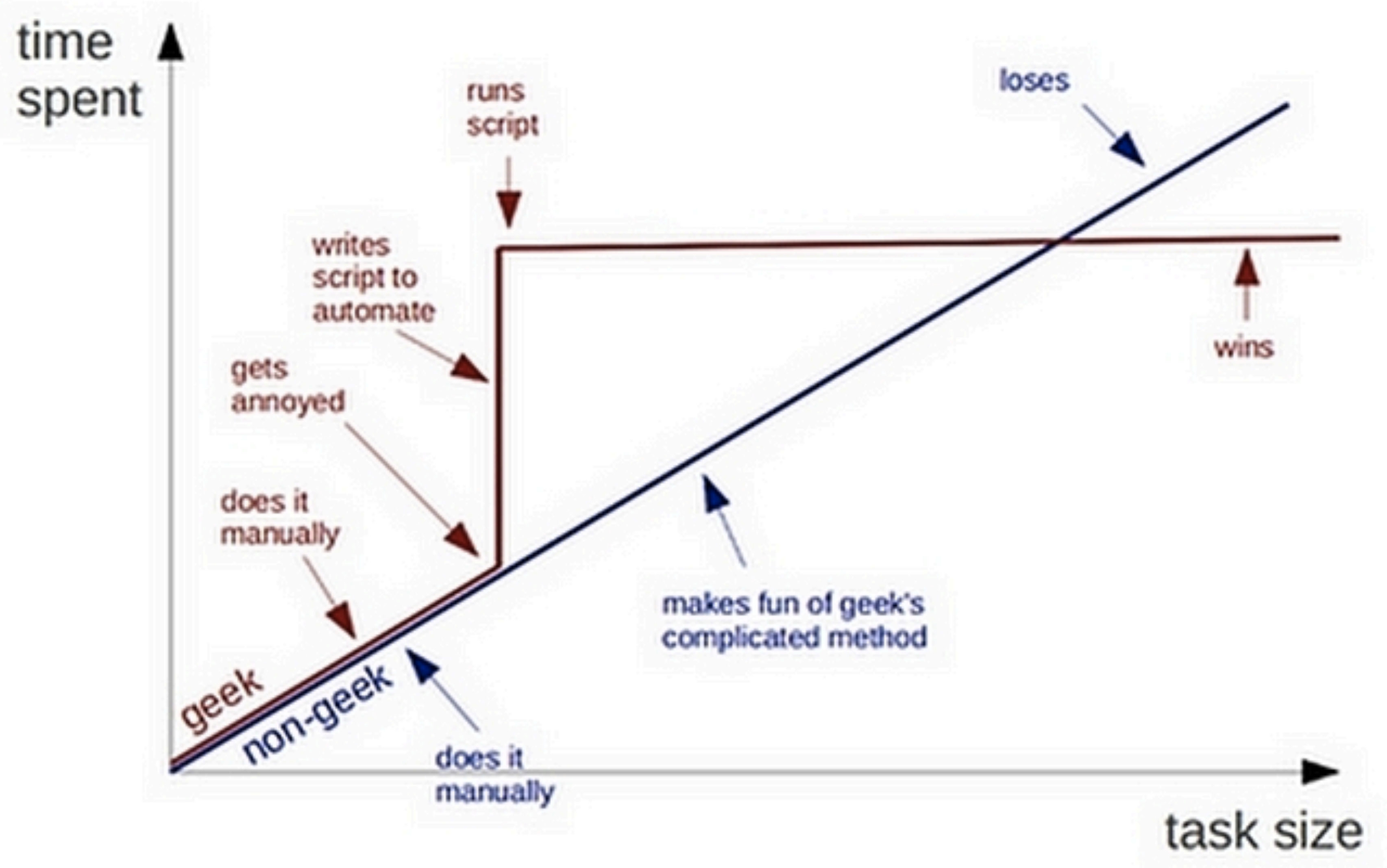
Bash

- Bourne Again SHell
- Bash is the terminal program we are using to talk to the ACF
- It also has a simple language
- You’ve already been learning it! Everything we type on the command line is computer language

More reasons to write scripts!

- Can customize commands and run software with a few key strokes
- Can operate on hundreds to millions of files
- Can have many different jobs running simultaneously across many computers
- Modular workflows and components
  - We can experiment with different pieces of software at each stage of analysis
  - Reuse
  - Examine results at each stage of analysis

Geeks and repetitive tasks


time spent
runs script
loses
writes script to automate
wins
gets annoyed
does it manually
makes fun of geek's complicated method
geek
non-geek
does it manually
task size