

# From Cell Line to Command Line: my unexpected career in Bioinformatics



Ming 'Tommy' Tang

Director of Bioinformatics at AstraZeneca

Twitter/X: tangming2005

<https://divingintogeneticsandgenomics.com/>

YouTube: Chatomics

08/03/2023



# Ming 'Tommy' Tang, PhD

Boston, MA

tangming2005@gmail.com



2008 BS Biotechnology



2014

PhD. Genetics & Genomics

Outstanding International student

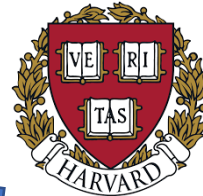
Instructor of Data Carpentries

THE UNIVERSITY OF TEXAS  
**MD Anderson Cancer Center**

Making Cancer History®

2014-2018

Postdoc & Research Scientist



2018-2020

Senior bioinformatician

Top 75 Bioinformatics Blogs by feedspot.com



**Dana-Farber**  
Cancer Institute

2020 – 2021

Lead Scientist

Lead the Bioinformatics effort for **Cancer Immunologic Data Commons**



2021- 2024.08.

Director of computational biology



Director of Bioinformatics



Over 40 publications

**DNA CONFESSES DATA SPEAK**

Blog: <https://divingintogeneticsandgenomics.com>  
6000 views per month



@tangming2005 ~30K followers

Youtube: chatomics ~5000 subscribers

# I am one of you



2008



2018

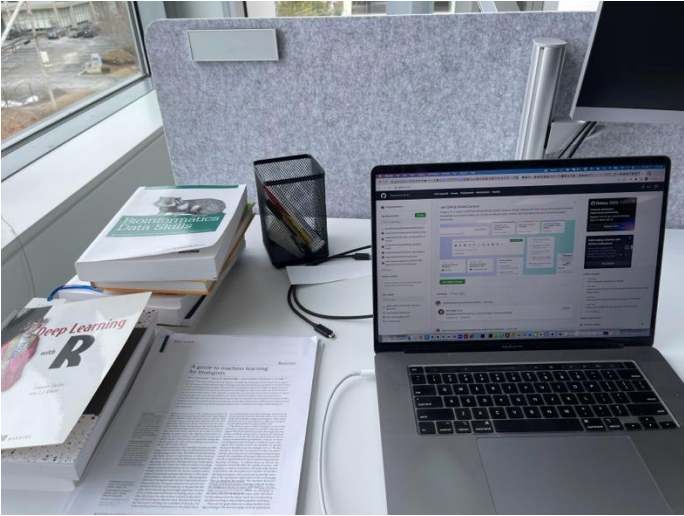
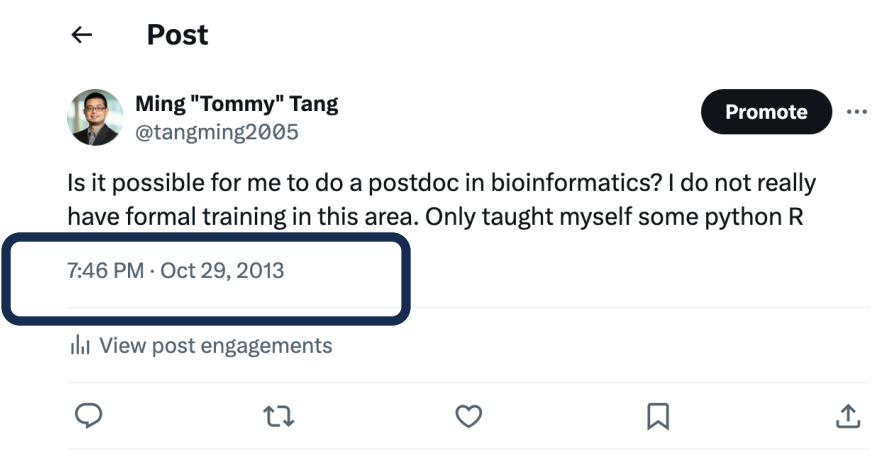
# Two gaps



Dr.Jianrong Lu lab  
University of Florida 2013



GAP



Immunitas Therapeutics  
2023



GAP



# My story in Nature Career column

**nature**

Explore content ▾

About the journal ▾

Publish with us ▾

Subscribe

---

[nature](#) > [career column](#) > article

CAREER COLUMN | 04 October 2023

## **Embracing the command line: my unexpected career in computational biology**

**A crash course in bioinformatics put Ming Tommy Tang on a different path.**

By [Ming Tommy Tang](#)

<https://divingintogeneticsandgenomics.com/publication/2023-10-04-nature-career/>

# Data deluge

# 1.845e+16

Number of publicly available bases in the NCBI Sequence Read Archive (SRA) as of July 1, 2018. This is the equivalent of 6,153,232 human genomes (which is  $3e+9$  bases).

6

# 30TB

Approximate amount of public sequence data received and processed *daily* by the NCBI Sequence Read Archive (SRA).

# All biology is computational biology



RESEARCH MATTERS

## All biology is computational biology

**Florian Markowetz\***

University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom

\* [florian.markowetz@cruk.cam.ac.uk](mailto:florian.markowetz@cruk.cam.ac.uk)

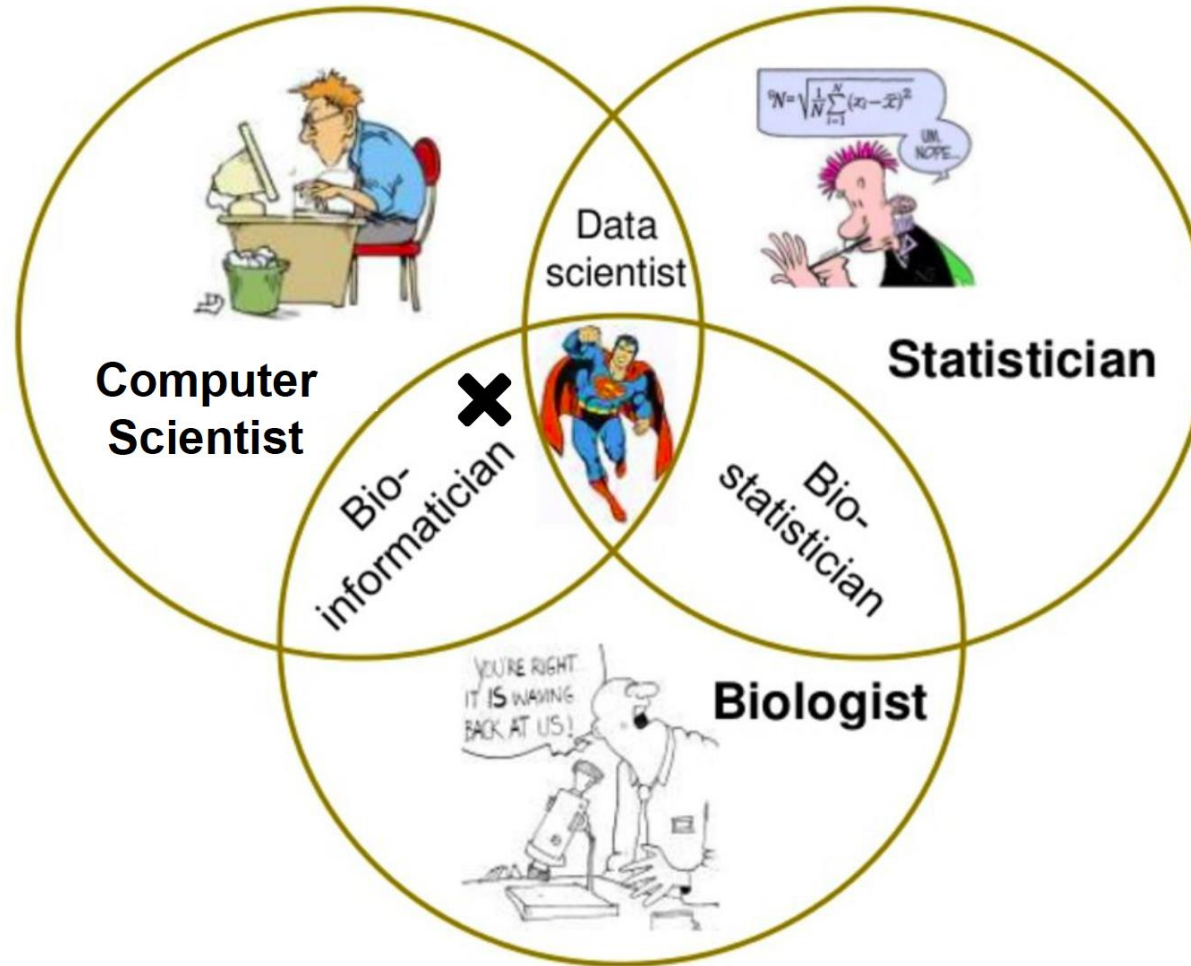
### Abstract

Here, I argue that computational thinking and techniques are so central to the quest of understanding life that today all biology is computational biology. Computational biology brings order into our understanding of life, it makes biological concepts rigorous and testable, and it provides a reference map that holds together individual insights. The next modern synthesis in biology will be driven by mathematical, statistical, and computational methods being absorbed into mainstream biological training, turning biology into a quantitative science.

### Rest in peace, computational biology

Pipette biologist. Microscopy biologist. Cell culture biologist. Have you ever heard any of those job titles? No, of course not. All are biologists, because it is the questions you address that matter, not the tools you use, and computational biologists are just biologists using a different tool.

# Computational biologists are the Superman/Wonder woman



Bioinformatician  
Computational biologist



What should you learn to tame the data?

# Excel is not enough

**BBC** Sign in Home News Sport Reel Worklife Travel

## NEWS

Home | US Election | Coronavirus | Video | World | US & Canada | UK | Business | Tech | Science | Stories

Tech

### Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion  
Technology desk editor

5 October

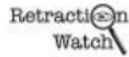
Coronavirus pandemic

The problem is that PHE's own developers picked an old file format to do this - known as XLS.

As a consequence, each template could handle only about 65,000 rows of data rather than the one million-plus rows that Excel is actually capable of.

And since each test result created several rows of data, in practice it meant that each template was limited to about 1,400 cases.

When that total was reached, further cases were simply left off.



**Retraction Watch**

@RetractionWatch

Follow



An Excel screw-up leads to a retraction.  
"This technological issue caused rows to shift and the data from the different groups got mixed up."

[sciencedirect.com/science/articl ...](https://www.sciencedirect.com/science/article/pii/S0018506X18302599)

12:27 PM - 6 Aug 2018

17 Retweets 21 Likes



9



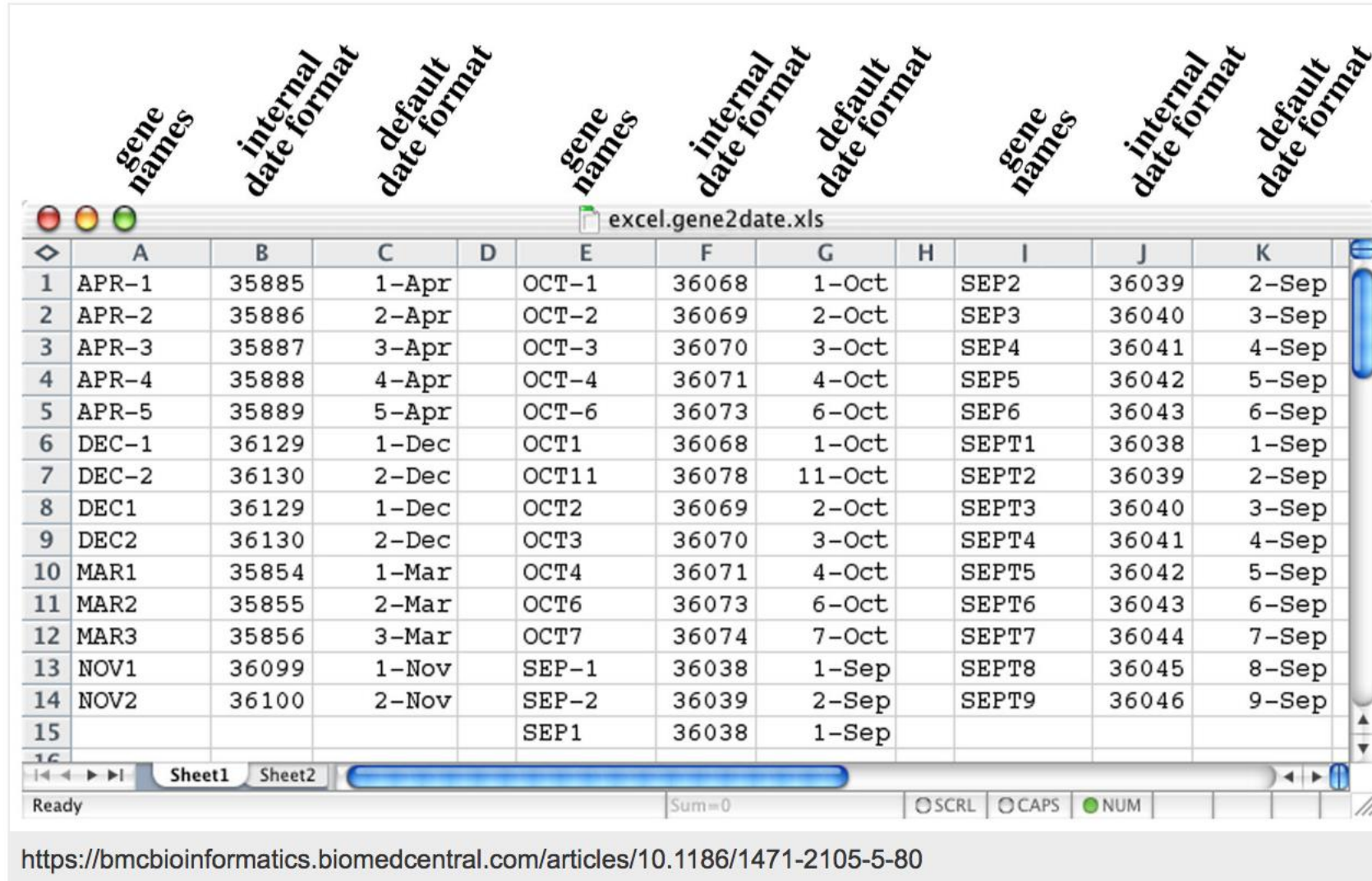
17



21



# Excel converts gene names to dates



excel.gene2date.xls

	gene names	internal date format	default date format		gene names	internal date format	default date format		gene names	internal date format	default date format
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

Ready | Sum=0 | SCRL | CAPS | NUM

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80>

Use R packages:  
Readxl, Janitor  
To work with  
excel sheets

<http://blogs.nature.com/naturejobs/2017/02/27/escape-gene-name-mangling-with-escape-excel/>

# Best practices using spreadsheets



Article

## Data Organization in Spreadsheets

Karl W. Broman & Kara H. Woo

Pages 2-10 | Received 01 Jun 2017, Published online: 24 Apr 2018

Cite this article <https://doi.org/10.1080/00031305.2017.1375989>



**B**

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447



	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

# Learn Unix command line

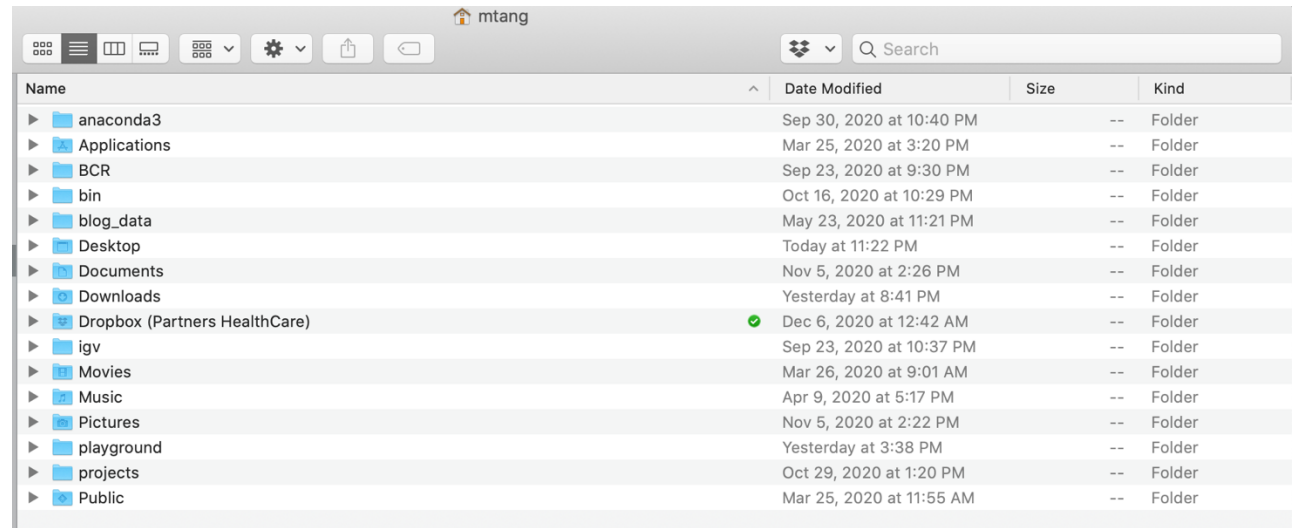
- Why command line?
- The text file is still the "king" format of bioinformatics. Unix commands are perfect to wrangle files.
- Most bioinformatics tools are run by the command line.
- More efficient/powerful: e.g, `cp *png pictures/`
- HPC (high-performance computing cluster), cloud computing

# Terminal

```
(base) → ~ ls
Applications      Pictures
BCR               Public
Desktop           anaconda3
Documents         bin
Downloads         blog_data
Dropbox (Partners HealthCare) igv
Library           playground
Movies            projects
Music
```

```
(base) → ~
```

CLI



GUI

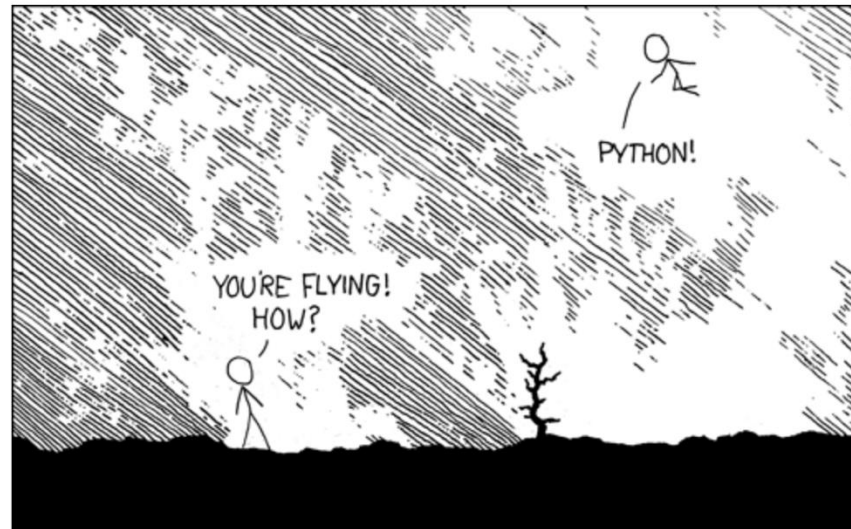
Use a mac/ubuntu or windows10 has a built-in

<http://swcarpentry.github.io/shell-novice/>

# Learn some python

PYTHON

< < PREV RANDOM NEXT > >



I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!  
HELLO WORLD IS JUST  
`print "Hello, world!"`

I DUNNO...  
DYNAMIC TYPING?  
WHITESPACE?

COME JOIN US!  
PROGRAMMING IS FUN AGAIN!  
IT'S A WHOLE NEW WORLD UP HERE!




BUT HOW ARE YOU FLYING?

I JUST TYPED  
`import antigravity`

THAT'S IT?

... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.



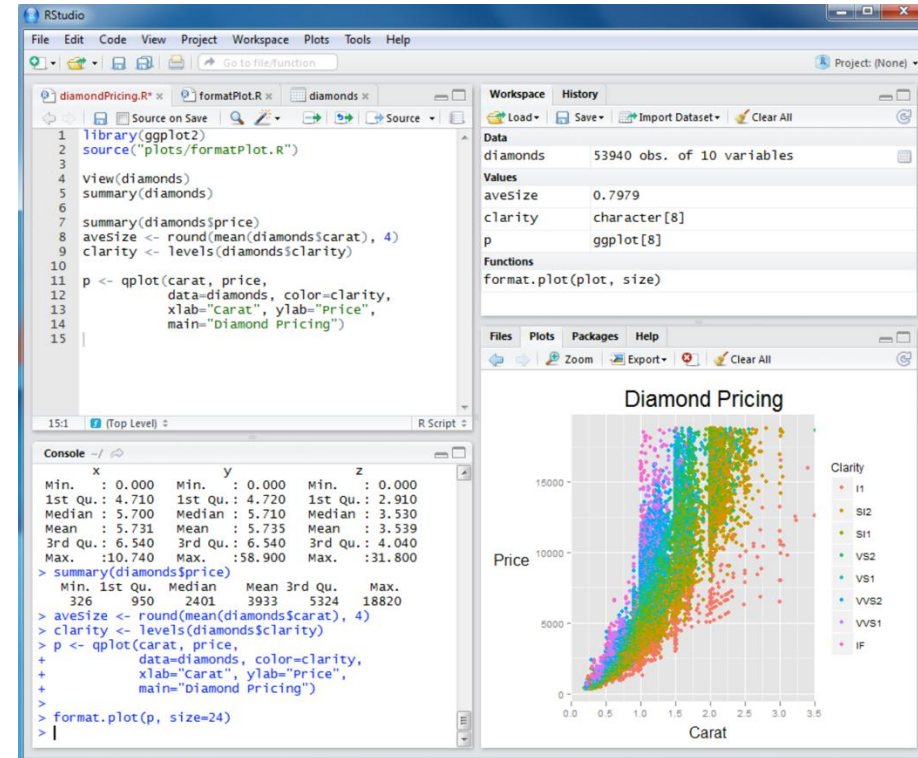
BUT I THINK THIS IS THE PYTHON.

<https://xkcd.com/>



# Learn some R

- Rstudio (IDE)
- Bioconductor
- Tidyverse and ggplot2



<http://adv-r.had.co.nz/> Advanced R  
<https://r4ds.had.co.nz/> R for data science



R packages for data science

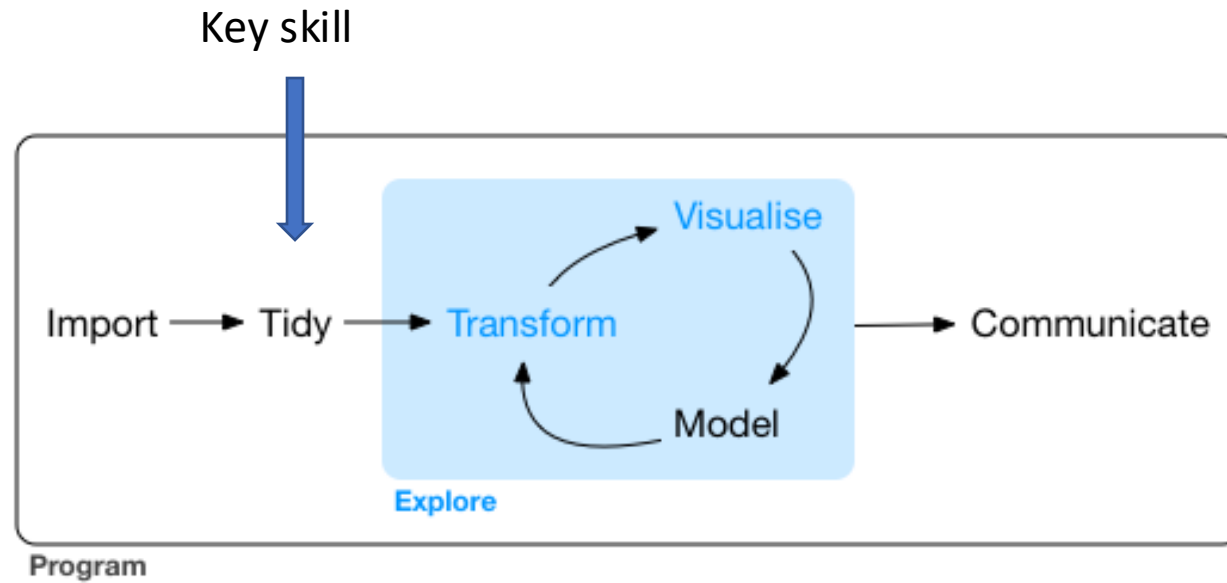
The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

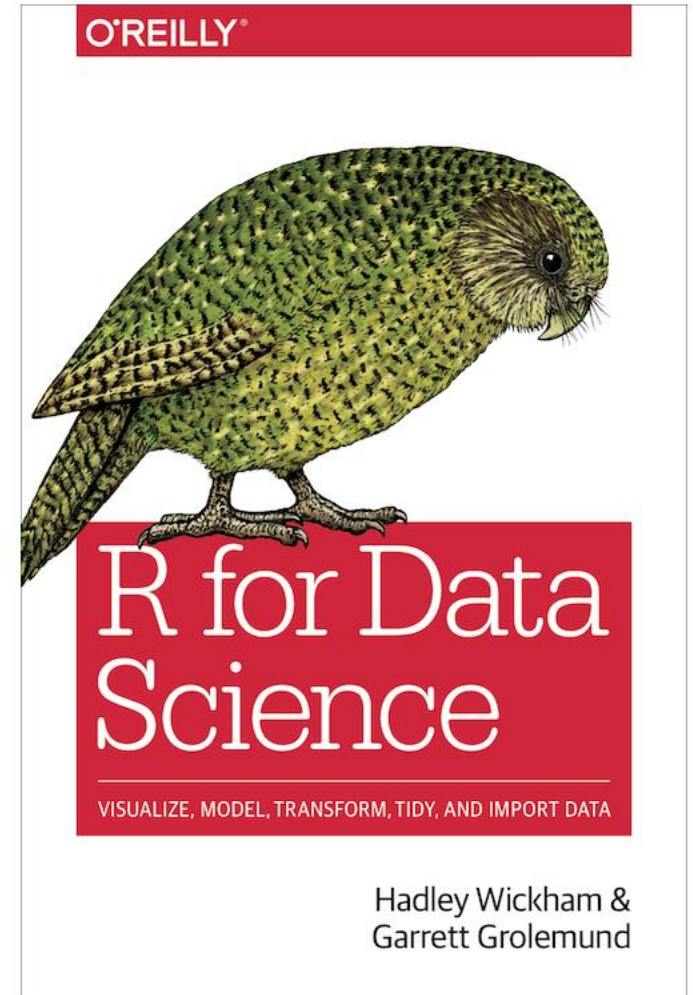
<https://www.tidyverse.org/>

# Data analysis workflow

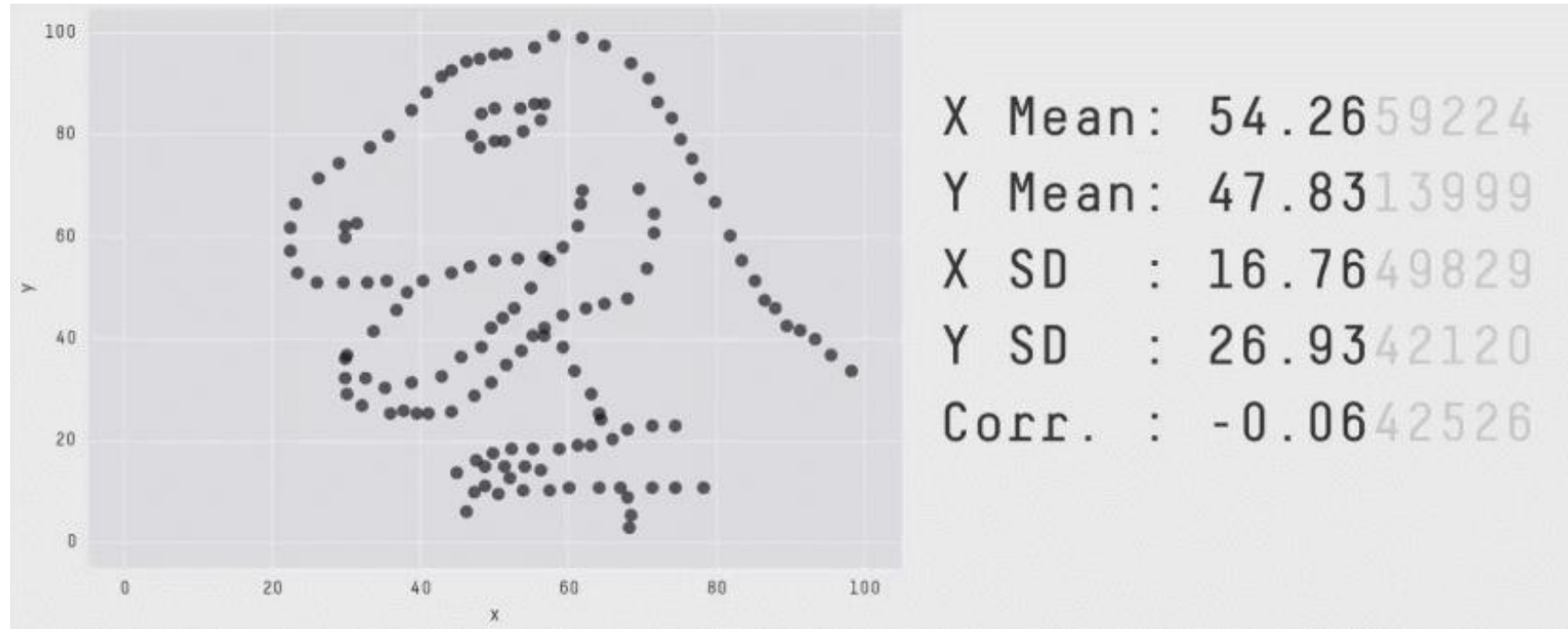


Tidying the data can take 80% of your time

R for data science by Hadley Wickham & Garrett Golemund  
<http://r4ds.had.co.nz/>

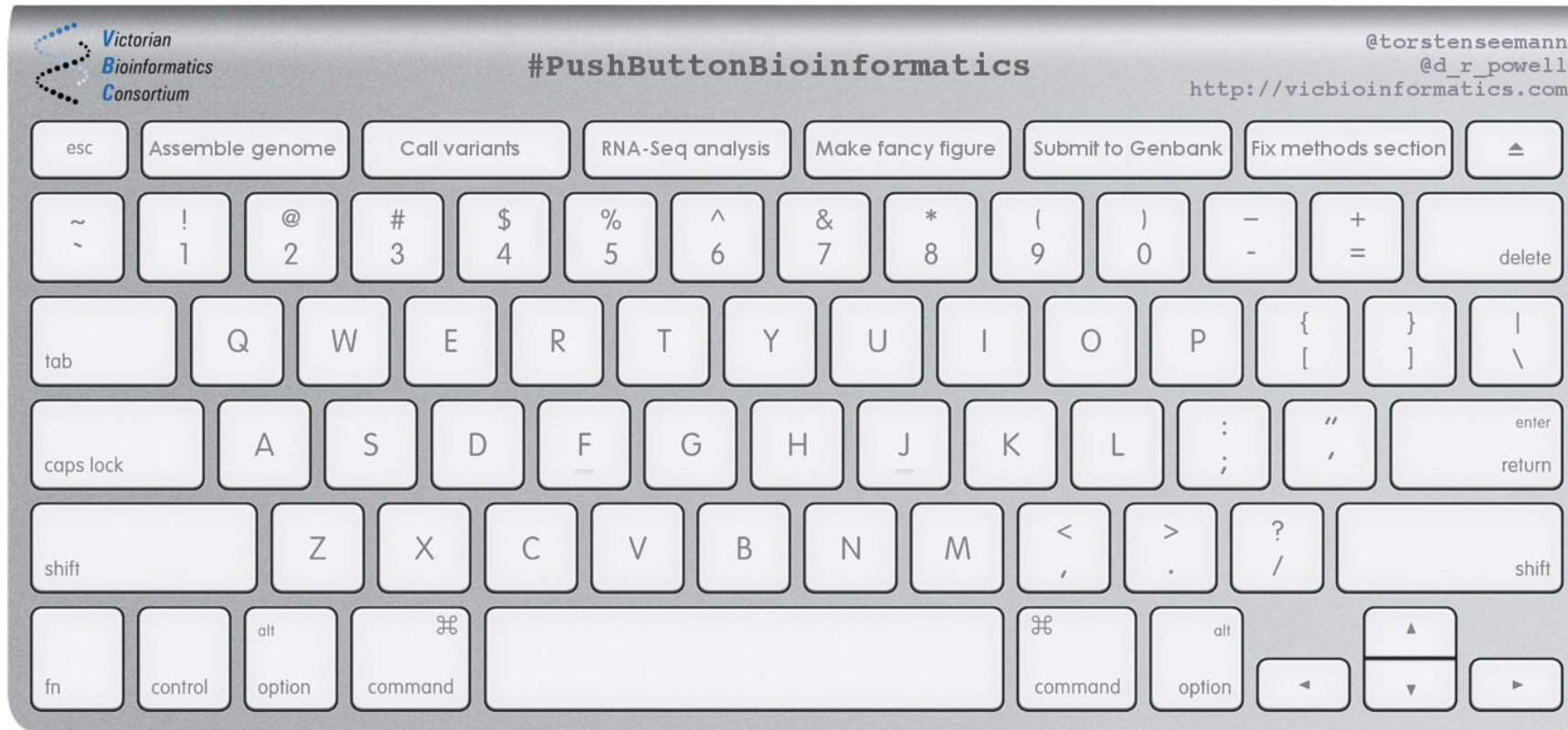


# Data visualization



<https://www.r-bloggers.com/the-datasaurus-dozen/>

# What people think we do



Credit: Torsten Seemann

# A typical day of my life as a computational biologist

- Installing software
- Googling (how to and error message etc).
- Read manuals of bioinformatics tools.
- Converting file formats.
- Tidying the data.
- Real analysis (plotting etc) 20%



**Ming (Tommy) Tang**  
@tangming2005



bioinformatician certificate task #0: install this package without error

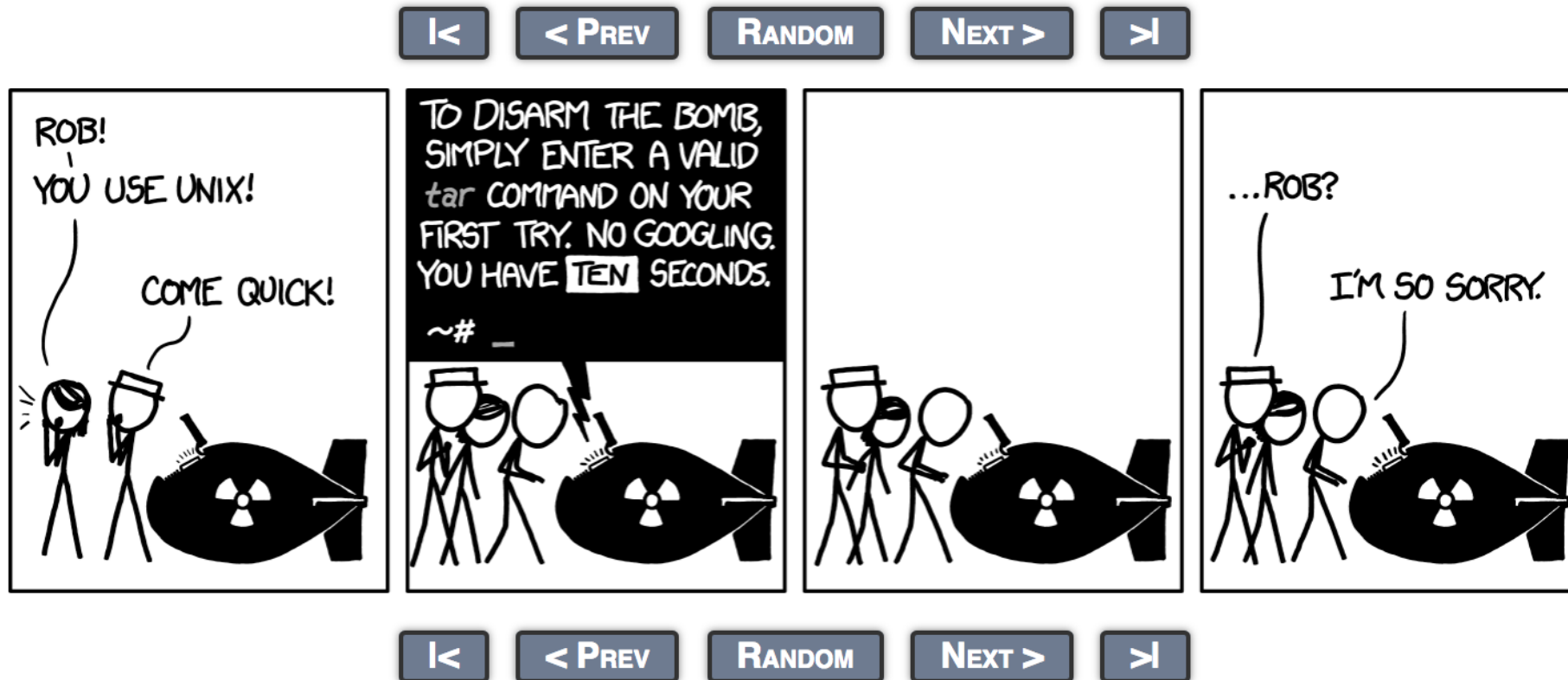
6:53 PM · Dec 7, 2020 · Twitter Web App

||| [View Tweet activity](#)

**3** Quote Tweets **74** Likes

# Google is how we learn and do things

**TAR**: A command to decompress files



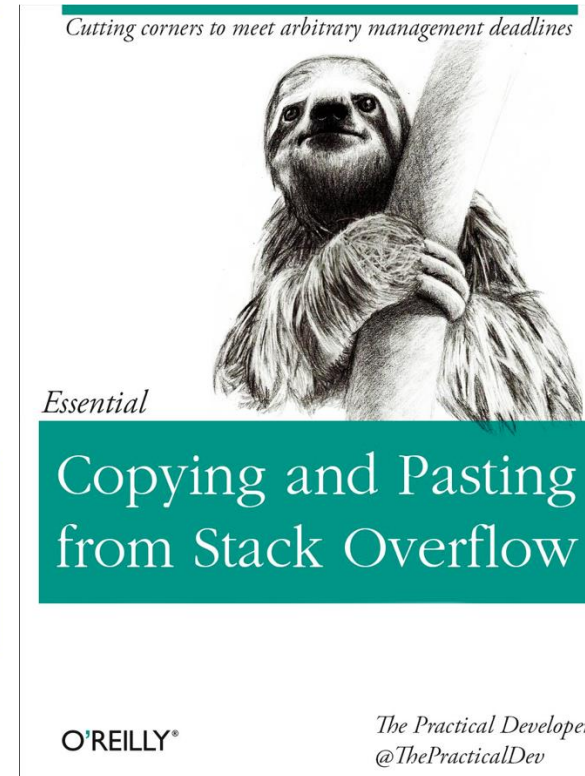
I have to google the tar command everytime...

# Google tricks

- Add `[r]` to search R programming related pages. e.g., “distance measurement [r].” you can do it with other languages too: “rotate x axis labels [python].” Search “patchwork [r]” will find you the R package.
- Use quotations `" "` to search for the exact phrase.
- Add a tilde `~` in front of a word to find synonyms.
- Exclude terms with a minus `-` symbol.
- Search specific sites with `site:` . “heatmap site:<https://support.bioconductor.org>” will search heatmap inside the bioconductor support website.
- Define a filetype by: `heatmap filetype:pdf` it will only give you PDF files in the results.

# Ask for help

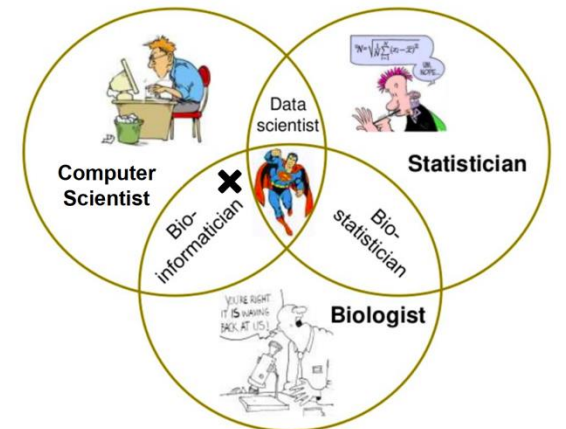
- SeqAnswer
- Biostars
- Stack overflow
- Bioconductor support site





# Key take-aways

- Learn Unix commands, python and R
  - Google is the way. Now, we have Chat-GPT
  - Be cautious with excel
  - Git version control your code
  - Have a consistent folder structure for projects' reproducible computing
  - Learn by doing
  - Focus on your strength: biology domain knowledge
- Computational ***biologist***.



# Tip #1 Get on social medium

- Get on social medium: Twitter/X, Mastodon, bulesky, LinkedIn
- Follow people of the same interest; bioinformatics papers



A screenshot of a Twitter profile for Ming "Tommy" Tang. The profile banner features a dark blue background with a green DNA double helix on the left, the text "Diving into Genetics and Genomics" in white, and "Transform life science with data" in green at the bottom. A small bar chart icon is on the right. The profile picture shows a man with glasses. The name "Ming 'Tommy' Tang" and handle "@tangming2005" are displayed. The bio reads: "Director of computational biology. On my way to helping 1 million people learn bioinformatics. Educator, Biotech, single cell. Also talks about leadership." Location is "Boston, MA" and website is "tommytang.bio.link". It shows "2,196 Following" and "25.6K Followers".

**Ming "Tommy" Tang**  
@tangming2005

Director of computational biology. On my way to helping 1 million people learn bioinformatics. Educator, Biotech, single cell. Also talks about leadership.

Science & Technology Boston, MA [tommytang.bio.link](https://tommytang.bio.link)  
Joined December 2011

2,196 Following 25.6K Followers



I started using twitter after reading Stephen Turner's blog:  
How to stay current on bioinformatics:  
<https://www.r-bloggers.com/2017/02/staying-current-in-bioinformatics-genomics-2017-edition/>

# LinkedIn tips

- 1. take a professional picture
- 2. use a related banner



- 3. headline showing what's your expertise



**Ming "Tommy" Tang**  
Cure Cancer with Data | Computational Biology | Data Science |  
Biotech Executive | Open Science Advocate | Prev. Dana-Farber |  
Harvard | MD Anderson | Join 16K followers on twitter  
@tangming2005 to learn computation

Talks about #teamwork, #leadership, #management, ##medicine, and  
#datascience

Tip #2 Write a blog

**I web, therefore I am**

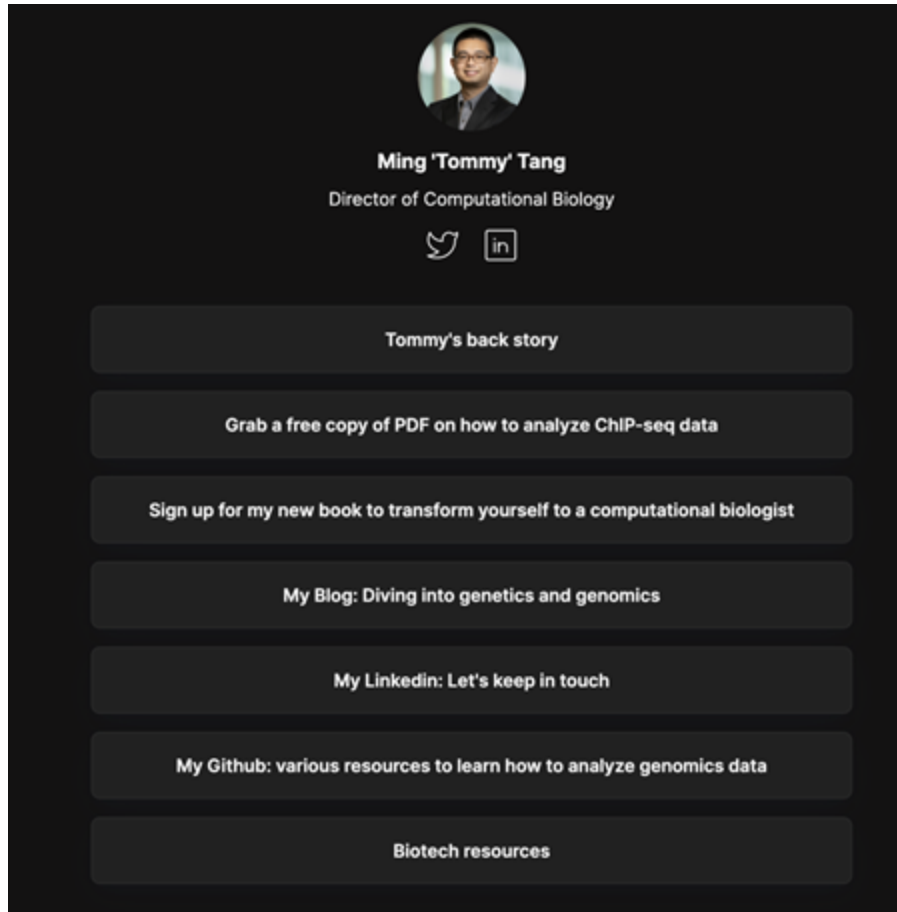
**Yihui Xie**

# Why a blog?

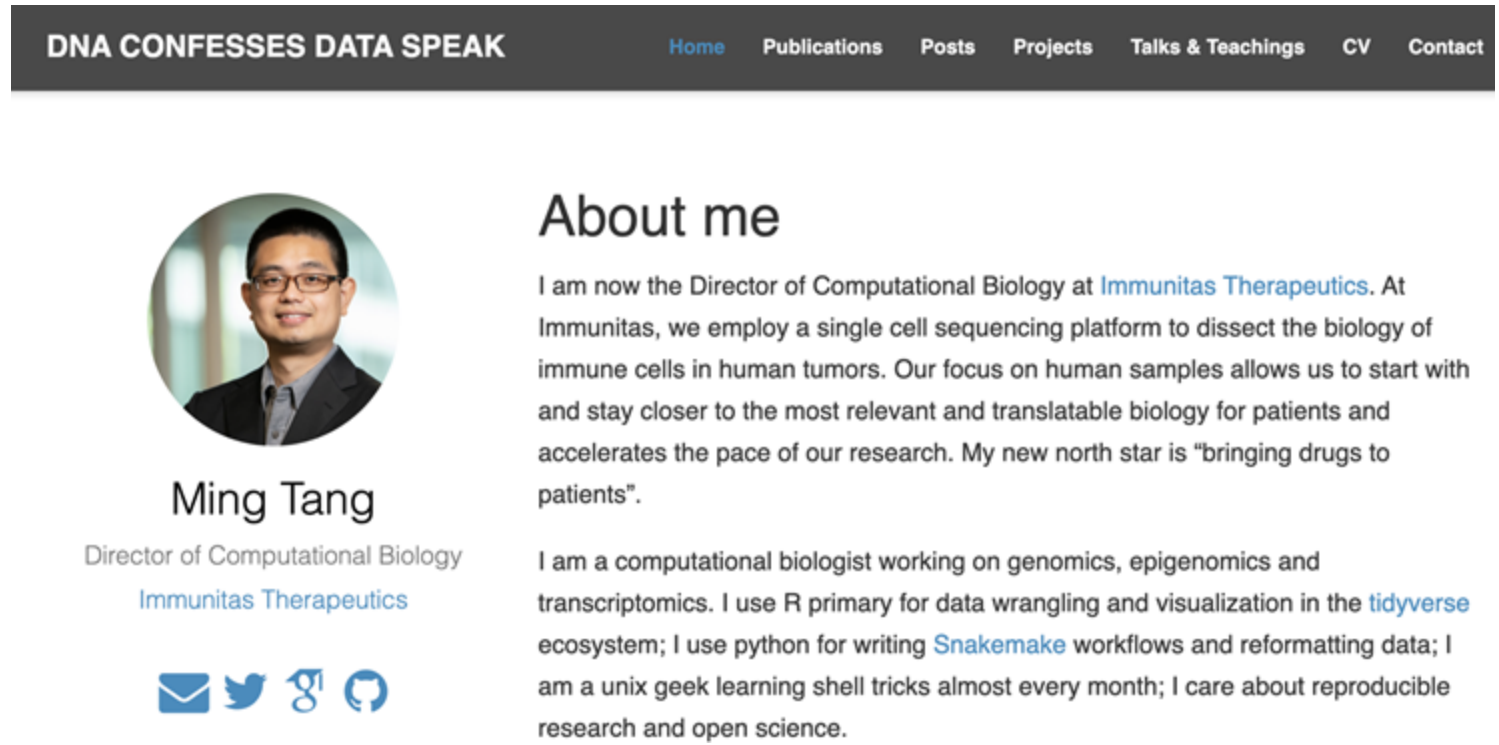
- a blog post is much better than a statement “good at R or Bayesian stats” on your CV
- “spend 30 minutes each day in 5 years building a website” vs “20 hours to write a CV in the last semester”
- there are many things that are more suitable for web pages (see my blog for example)

Credit: Yihui Xie

# Build a website so others can find you



<https://tommytang.bio.link/>



[divingintogeneticsandgenomics.com](http://divingintogeneticsandgenomics.com)

# Start now

- If you do not have a website yet.
- The best time to start one is 10 years ago, the second best time is now.
- Take a weekend to set it up. Be visible: Blogdown or Quarto
- Make a github repository. Put your projects there.

# Tip #3 How to connect with people? on social media or in real life

One core thing to remember: always give value on the table.

Be a giver not a taker

- Complement
- Ask questions. Be genuinely curious
- Offer help



# Tip#4 What to consider in joining a company?

- 1. Are the science cutting edge?
- 2. Are the people you will work with kind and smart?
- 3. is the finance of the company good? Last at least 2-3 years

Credit: Thomas Tan

# Make the transformation you want



University of Florida  
2013



GAP



Immunitas Therapeutics  
2023



GAP



Benefit



# Acknowledgments

Verhaak Lab  
Samir Amin

Immunitas  
Matthew Bernstein  
Shruti Malu

Titus Brown

Data Carpentry <https://datacarpentry.org/>

All the people who share their wisdom on the web  
Thanks!

What questions do you have?

# Reproducibility crisis

Every baby knows the  
**scientific method!**



# Most computational research is not reproducible.

I don't know of a systematic study, but of papers that I read, approximately 95% fail to include details necessary for replication.

**It's very hard to build off of research like this.**

(There's a lot more to say about repeatability, reproducibility and replicability than I can fit in here...)

# An example

- [The Importance of Reproducible Research in High-Throughput Biology.](#)
- <https://www.youtube.com/watch?v=7gYIs7uYbMo>
- By Dr.Keith A. Baggerly from MD Anderson Cancer Center.
- Highly recommend, Keith is very fun.

## Flawed Cancer Trial at Duke Sparks Lawsuit

By [Jennifer Couzin-Frankel](#) | Sep. 9, 2011 , 3:38 PM

---

A dozen plaintiffs have filed a **lawsuit** against Duke University and administrators, researchers, and physicians there, alleging that they engaged in fraudulent and negligent behavior when they enrolled cancer patients in a clinical trial compromised by faulty data. The lawsuit, filed Wednesday in a North Carolina court, comes 14 months after a **scandal erupted at Duke** that finally exposed the extent of the trial's problems: in July 2010, Duke oncologist Anil Potti, whose work was central to the trial, admitted that he had embellished his resume and later **resigned**.

# Method matters

## RESEARCH ARTICLE

# Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors

Nathaniel D. Anderson<sup>1,2</sup>, Richard de Borja<sup>1,\*</sup>, Matthew D. Young<sup>3,\*</sup>, Fabio Fuligni<sup>1,\*</sup>, Andrej Rosic<sup>1</sup>, Nicola D. Roberts<sup>3</sup>, Simo...

+ See all authors and affiliations

*Science* 31 Aug 2018:  
Vol. 361, Issue 6405, eaam8419  
DOI: 10.1126/science.aam8419

## Detection of gene fusions

We detected gene fusions in regions of genomic complexity using an approach that integrates multiple independent fusion algorithms, and then removed those found in normal tissue. Putative fusions were validated by de novo assembly. A total of 1277 normal (nonneoplastic) samples from 43 different tissues were obtained from the NHGRI GTEx consortium (database version 4) and used to remove artifacts. All fusions were visually inspected if one or both genes involved chromoplexy or were adjacent (up to 1 Mbp). Fusions were further filtered by quality of the realigned transcript, breakpoint coverage, and gene expression.



Why reproducibility is hard?

# Why reproducibility is hard?

- no raw data are available.
- scripts/data available upon reasonable request 😊
- lack of method description.
- versions of the tools are different. (e.g. R/python/bioinformatics tools)
- different machines (unix vs windows).

# If it is so hard, should you care?

- Keep this in mind: You are going to do the same analysis for sure in the future yourself!
- This is for your own benefit.
- I want to make sure my analysis is reproducible because I am discovering drug targets for patients!

# How to ensure reproducibility

- Git version control
- Jupyter/R Notebook, documentation
- Containers (docker, singularity, biocontainers <https://biocontainers.pro/>)
- Unit test
- Continuous Integration/development CI/CD (Travis CI, github action)

# "FINAL".doc



FINAL.doc!



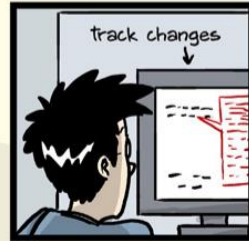
FINAL\_rev.2.doc



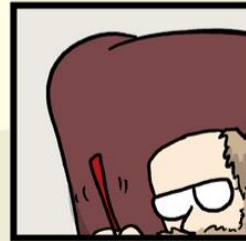
FINAL\_rev.6.COMMENTS.doc



FINAL\_rev.8.comments5.  
CORRECTIONS.doc



FINAL\_rev.18.comments7.  
corrections9.MORE.30.doc



FINAL\_rev.22.comments49.  
corrections.10.#@\$%WHYDID  
ICOMETOGRADSCHOOL?????.doc



# Version control

- Git
- Github
- Gitlab



Five commands can take you very far:

```
git init
```

```
git add
```

```
git commit -m "my first commit"
```

```
git push
```

```
git pull
```

# Jupyter Notebook


[JUPYTER](#)[FAQ](#)

[notebook](#) / [docs](#) / [source](#) / [examples](#) / [Notebook](#)

## Running Code

First and foremost, the Jupyter Notebook is an interactive environment for writing and running code. The notebook is capable of running code in a wide range of languages. However, each notebook is associated with a single kernel. This notebook is associated with the IPython kernel, therefore runs Python code.

## Code cells allow you to enter and run code

Run a code cell using `Shift-Enter` or pressing the  button in the toolbar above:

```
In [2]: a = 10
```

```
In [3]: print(a)
```

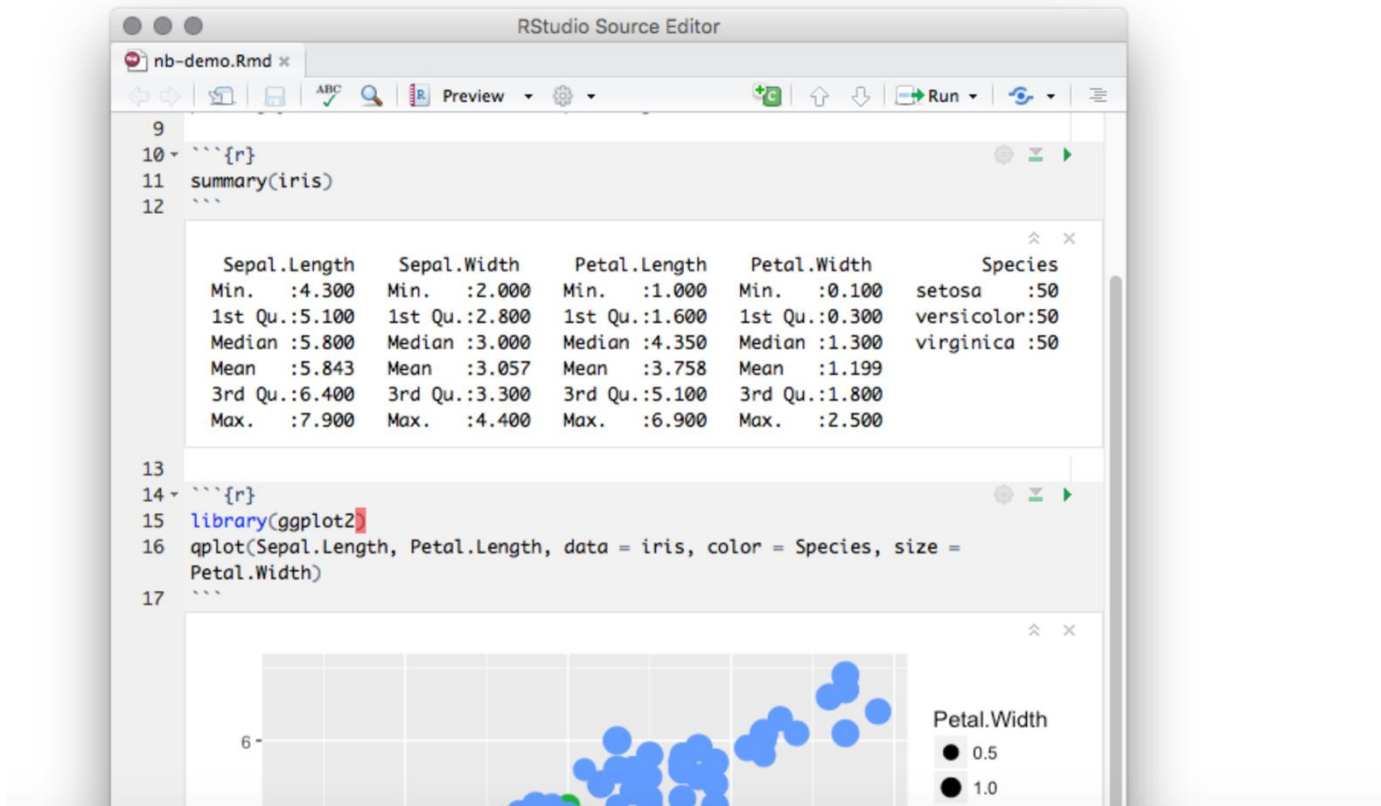
```
10
```

There are two other keyboard shortcuts for running code:

- `Alt-Enter` runs the current cell and inserts a new one below.
- `Ctrl-Enter` runs the current cell and enters command mode.

# R notebook/markdown

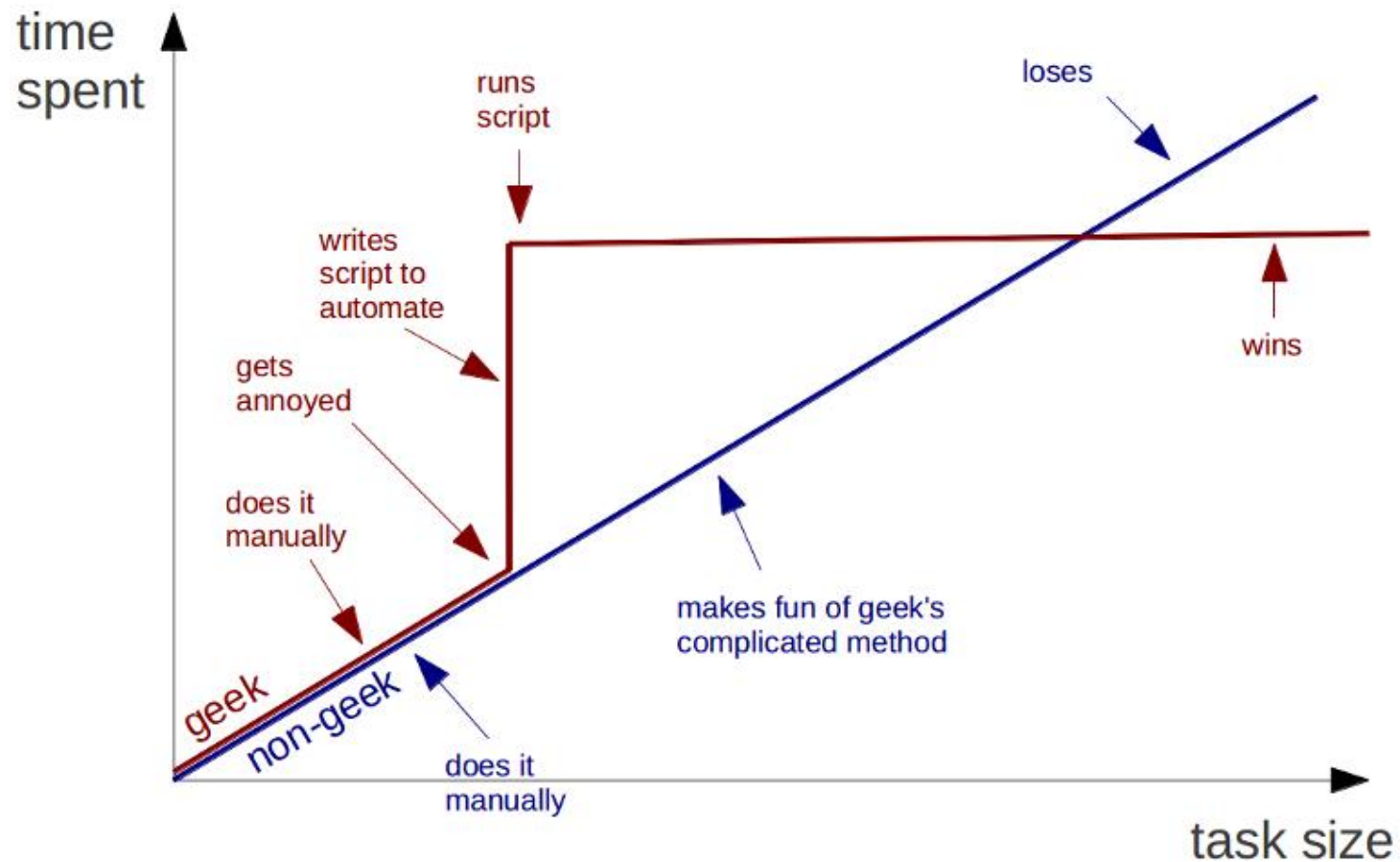
An R Notebook is an R Markdown document with chunks that can be executed independently and interactively, with output visible immediately beneath the input.





# Automation makes your research more reproducible AND saves you time in the long run

## Geeks and repetitive tasks



Computers are good at repetitive work

# Good Side effect of automation

- The best documentation is automation
- Write scripts for everything unless it is not possible. (manual editing, document, document, document!)
- Markdown, MKdocs <https://www.mkdocs.org/>

# Tips for automation

- 1. if you have a repetitive simple task, put them in to a shell script: `my_routine.sh`.
- 2. good old GNU make
- 3. more recent snakemake, nextflow, WDL etc.

## Awesome Pipeline

A curated list of awesome pipeline toolkits inspired by [Awesome Sysadmin](#)

### Pipeline frameworks & libraries

- [ActionChain](#) - A workflow system for simple linear success/failure workflows.
- [Adage](#) - Small package to describe workflows that are not completely known at definition time.
- [Airflow](#) - Python-based workflow system created by Airbnb.
- [Anduril](#) - Component-based workflow framework for scientific data analysis.
- [Anthra](#) - High-level language for biology.
- [AWE](#) - Workflow and resource management system with CWL support
- [Bds](#) - Scripting language for data pipelines.
- [BioMake](#) - GNU-Make-like utility for managing builds and complex workflows.
- [BioQueue](#) - Explicit framework with web monitoring and resource estimation.
- [Bioshake](#) - Haskell DSL built on shake with strong typing and EDAM support
- [Bistro](#) - Library to build and execute typed scientific workflows.



Snakemake—a scalable bioinformatics workflow engine

<b>Publication</b>	Article in <b>Bioinformatics</b> , published October 2012
<b>Authors</b>	Johannes Köster, Sven Rahmann

[↓ More details](#)



<https://github.com/pditommaso/awesome-pipeline>

# conda and biocoda

## Conda



*Package, dependency and environment management for any language—Python, R, Ruby, Lua, Scala, Java, JavaScript, C/ C++, FORTRAN*

MENU ▾

nature|methods

Correspondence | Published: 02 July 2018

## Bioconda: sustainable and comprehensive software distribution for the life sciences

Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris & Johannes Köster ✉ The Bioconda Team

*Nature Methods* **15**, 475–476 (2018) | [Download Citation](#) ↓


# Docker



- Why docker?
- Imagine you are working on an analysis in R and you send your code to a friend. Your friend runs exactly this code on exactly the same data set but gets a slightly different result. This can have various reasons such as a different operating system, a different version of an R package, etc. Docker is trying to solve problems like that.
- Think it as a virtual machine!
- This just happened between me and my colleagues who used a different version of R packages!

<https://cyverse-cybercarpentry-container-workshop-2018.readthedocs-hosted.com/en/latest/docker/dockerintro.html>

<https://ropenscilabs.github.io/r-docker-tutorial/01-what-and-why.html>

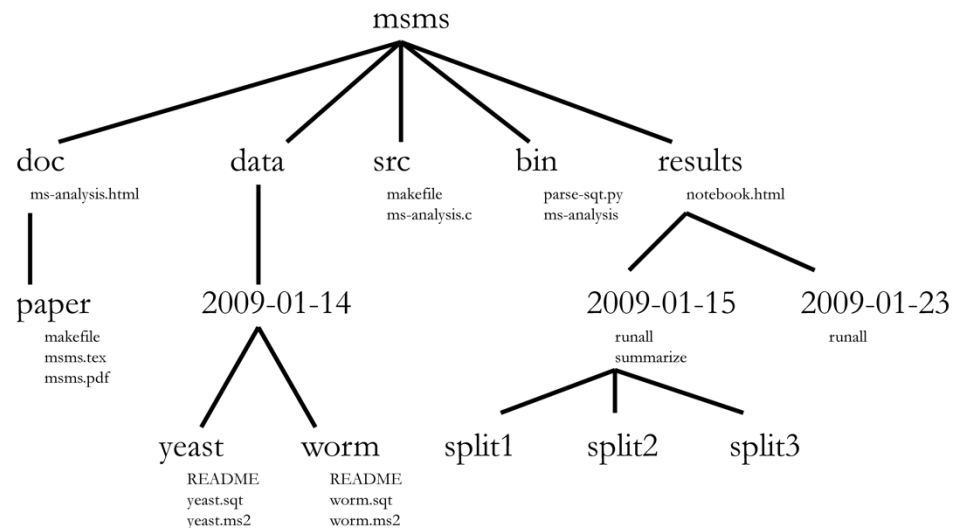
 OPEN ACCESS


EDUCATION

# A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble 

Published: July 31, 2009 • <https://doi.org/10.1371/journal.pcbi.1000424>




 OPEN ACCESS

PERSPECTIVE

## Good enough practices in scientific computing


Greg Wilson  , Jennifer Bryan , Karen Cranston , Justin Kitzes , Lex Nederbragt , Tracy K. Teal 

Published: June 22, 2017 • <https://doi.org/10.1371/journal.pcbi.1005510>

 OPEN ACCESS

COMMUNITY PAGE

## Best Practices for Scientific Computing

Greg Wilson , D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Kathryn D. Huff, Ian M. Mitchell, Mark D. Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

Thursday, August 13, 2015

## 2 cents on coding from a bioinformatics beginner

One needs to be aware that:

1. **Computers make mistakes.** They can give you non-sense results and exit without error, so make extensive tests before running your code.
2. **Share your codes.** Even your codes are correct, you need to share them so that other people can look at them and may improve them.
3. **Make your codes reusable.** Do not hard code your scripts. If it takes a file path as input, make it as an argument in your scripts.
4. **Modulate your scripts.** Data could come in different stage of formats. Take ChIP-sequencing data analysis as an example, if you have a script that starts processing the data from fastq to the final peaks. You may want to modulate your scripts to two modules: one for mapping fastq to bam, and the other for bam to peaks. **Modulate your scripts** so that one can use your script when the data come in a bam format.
5. **Heavily comment your scripts.** It will not only make other people to understand your codes better, but also help the future you to understand what you did.
6. **You need to make your analysis reproducible.** Each step of your analysis should be documented in a markdown file. I say every step, yes, every command that you strike in the terminal getting the intermediate files need to be taken down. Moreover, how, when and where did you download the data need to be documented. This will save the future you! Many experienced programmers overlook this point.