From cell line to command line: my journey to bioinformatics

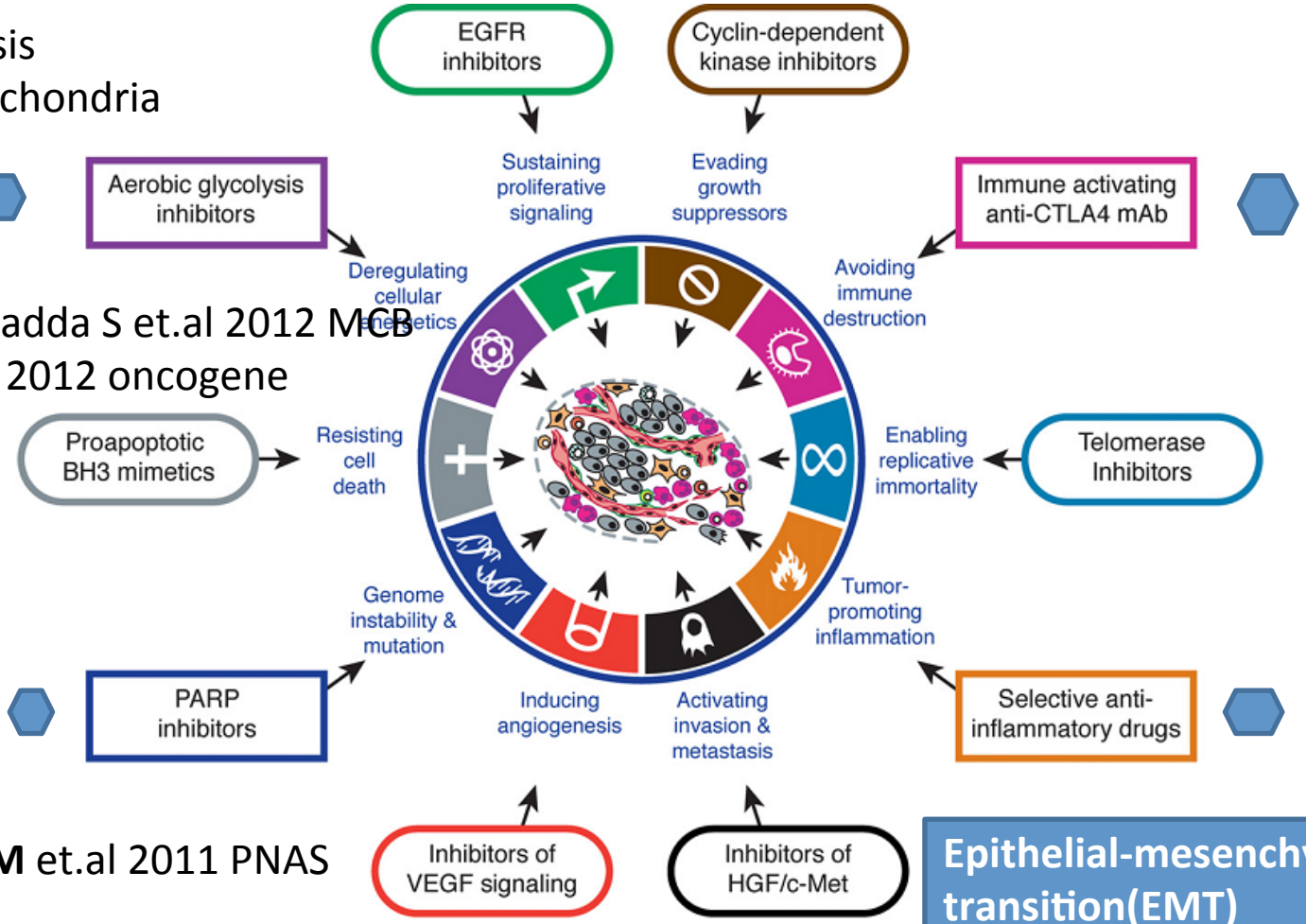
Ming (Tommy) Tang
Research scientist
Twitter @tangming2005
MD Anderson Cancer Center, Houston, TX
UFGI Genetics & Genomics program seminar

Self-introduction: 2008 UF Genetics and Genomics graduate student

Hallmarks of cancer

Glycolysis
VS mitochondria

kamarajugadda S et.al 2012 MCB
Cai Q et.al 2012 oncogene



Tang M et.al 2011 PNAS

Inhibitors of VEGF signaling

Inhibitors of HGF/c-Met

Epithelial-mesenchymal transition(EMT)

vascular endothelial growth factor (VEGF)

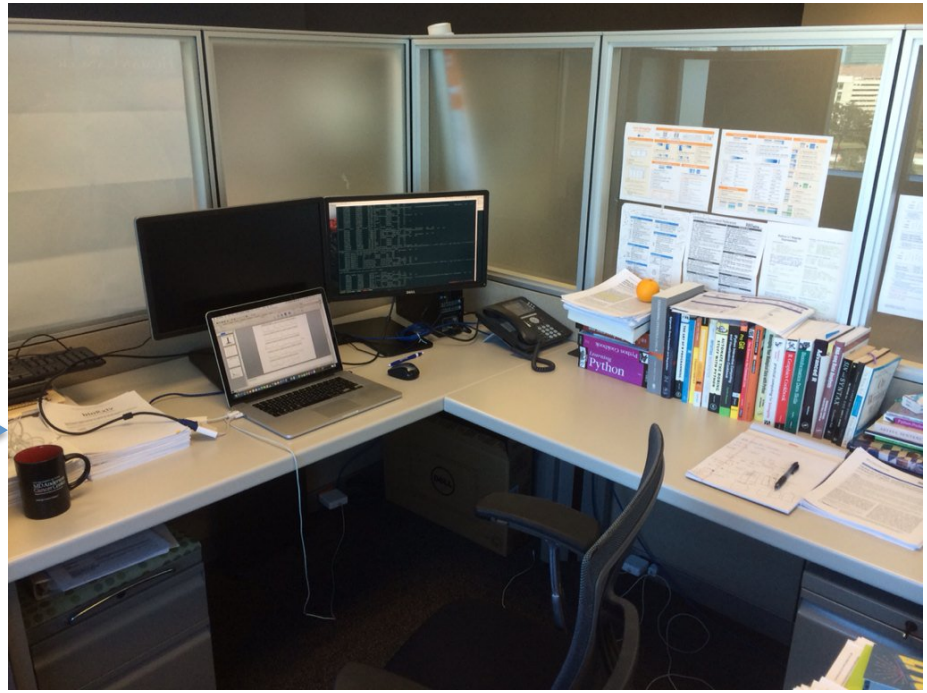
Tang M et.al 2013. JBC

Douglas Hanahan
and Robert A. Weinberg. 2011.Cell

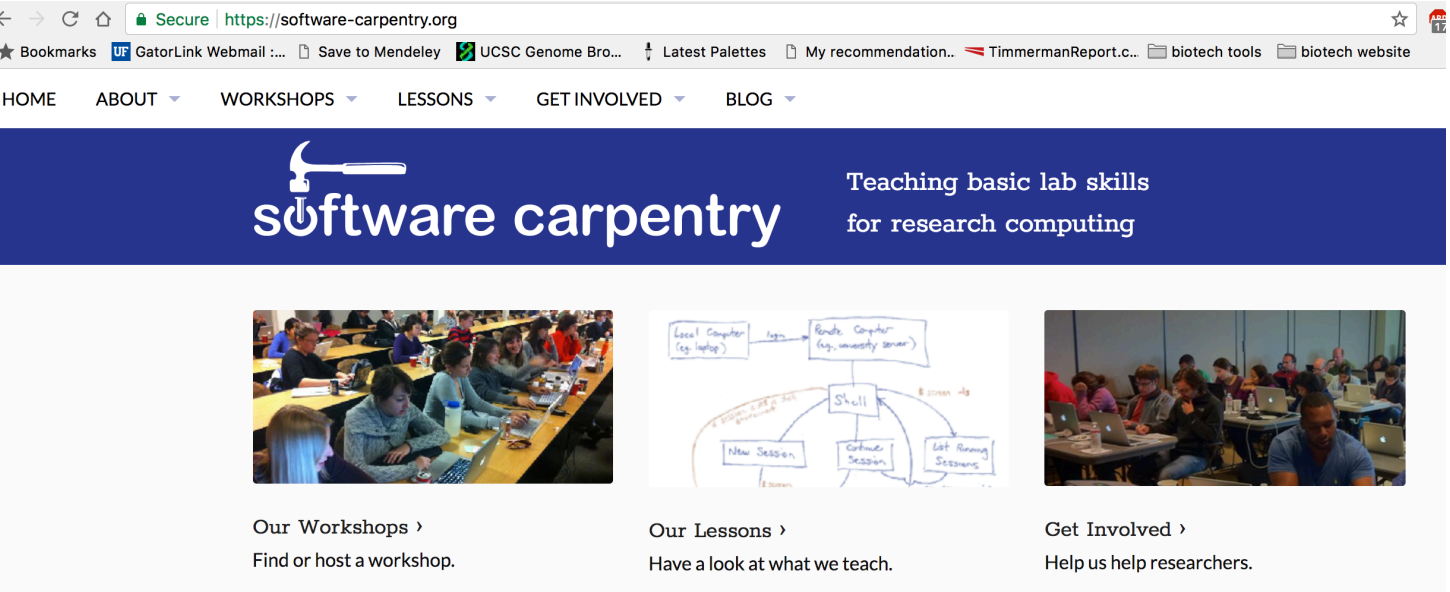
Challenges that I was facing

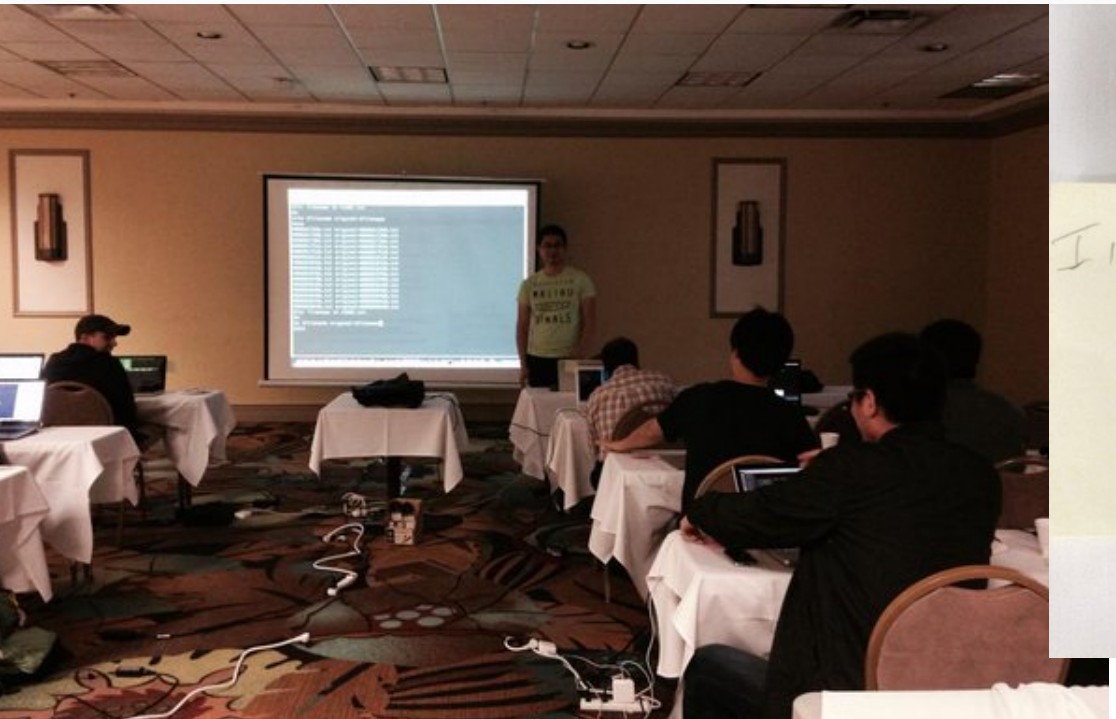
- How do I open this 2G ChIPseq file?
- Excel fails me.
- How do I download the files from GEO and process the raw data?
- That's how I started to teach myself Unix,R and python.

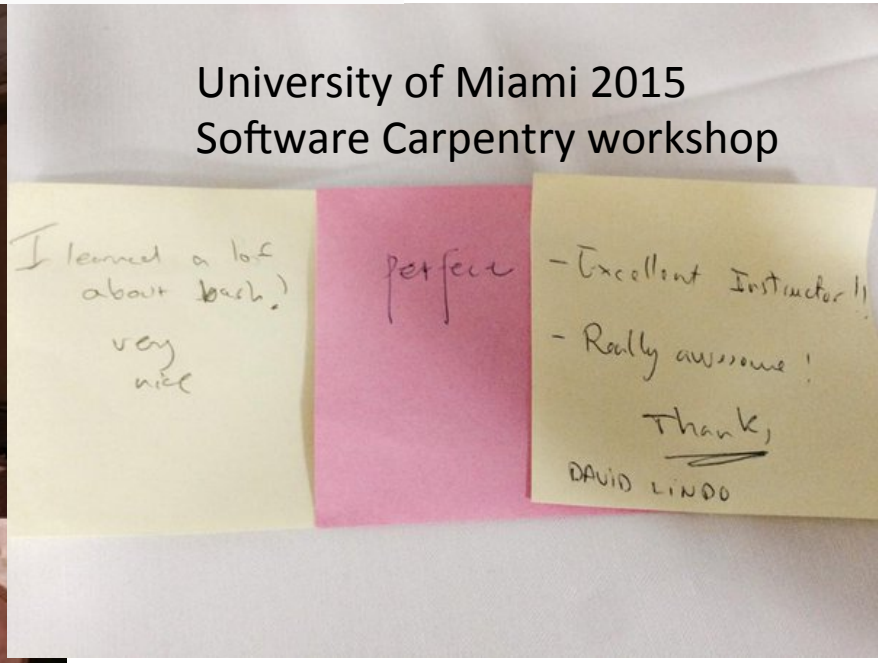
2015.03 joined MD Anderson
With Dr.Roel Verhaak for a
computational
Biology postodoc



Teaching





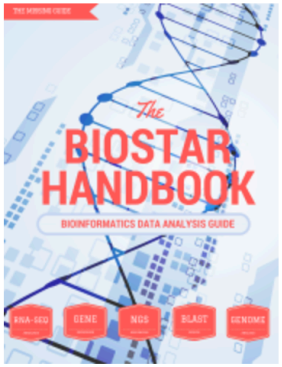


University of Miami 2015
Software Carpentry workshop

A book chapter published in April 2017

The Biostar Handbook: A Beginner's Guide to Bioinformatics

The Biostar Handbook introduces readers to **bioinformatics**, the scientific discipline at the intersection of biology, computer science, and statistical data analytics that is dedicated to the digital processing of genomic information.



The Handbook has been developed, improved and refined over more than a half decade in a research university setting and is used in an accredited PhD level training program. The contents of this book have provided the analytical foundation to hundreds of students, many of whom have become full time bioinformaticians and work at the most innovative companies in the world.

Find out more about the author and the development timeline.

Subscribe to the HandBook News!

email address | Subscribe

More about me

Ming Tang
crazyhottommy

I am a biologist cracking bioinformatics. I am working on cancer (epi)genomics at MD Anderson cancer center. I care reproducible research and open science

Edit bio

MD Anderson Cancer Center

Overview | Repositories 84 | Stars 212 | Followers 396 | Following 23

Pinned repositories — Customize your pinned repositories

ChIP-seq-analysis
ChIP-seq analysis notes from Ming Tang
Python ★ 221 122

RNA-seq-analysis
RNAseq analysis notes from Ming Tang
Python ★ 240 100

getting-started-with-genomics-tools-and-resources
Unix, R and python tools for genomics
Shell ★ 143 65

DNA-seq-analysis
DNA sequencing analysis notes from Ming Tang
Shell ★ 50 33

pyflow-ChIPseq
a snakemake pipeline to process ChIP-seq files from GEO or in-house
Python ★ 15 11

pyflow-ATACseq
ATAC-seq snakemake pipeline
Python ★ 18 9

Diving into Genetics and Genomics

A wet biologist's bioinformatic notes. Mostly is about Linux, R, python, reproducible research, open science and NGS. I am into data science! I am working on cancer genomics and epigenomics at MD Anderson cancer center. Disclaimer: For posts that I copied from other places, credits go to the original authors. Follow the links to the original posts, I mainly put them here for my own future references.

Starting a new job in October

Harvard FAS Informatics About Cores Employment Faq Software Tutorials Archives

HARVARD INFORMATICS

Analysis, training, software, and data management services for Harvard Faculty of Arts and Sciences

Moving to Harvard FAS informatics as a bioinformatics scientist

Challenges and opportunities

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

Data deluge

1.845e+16

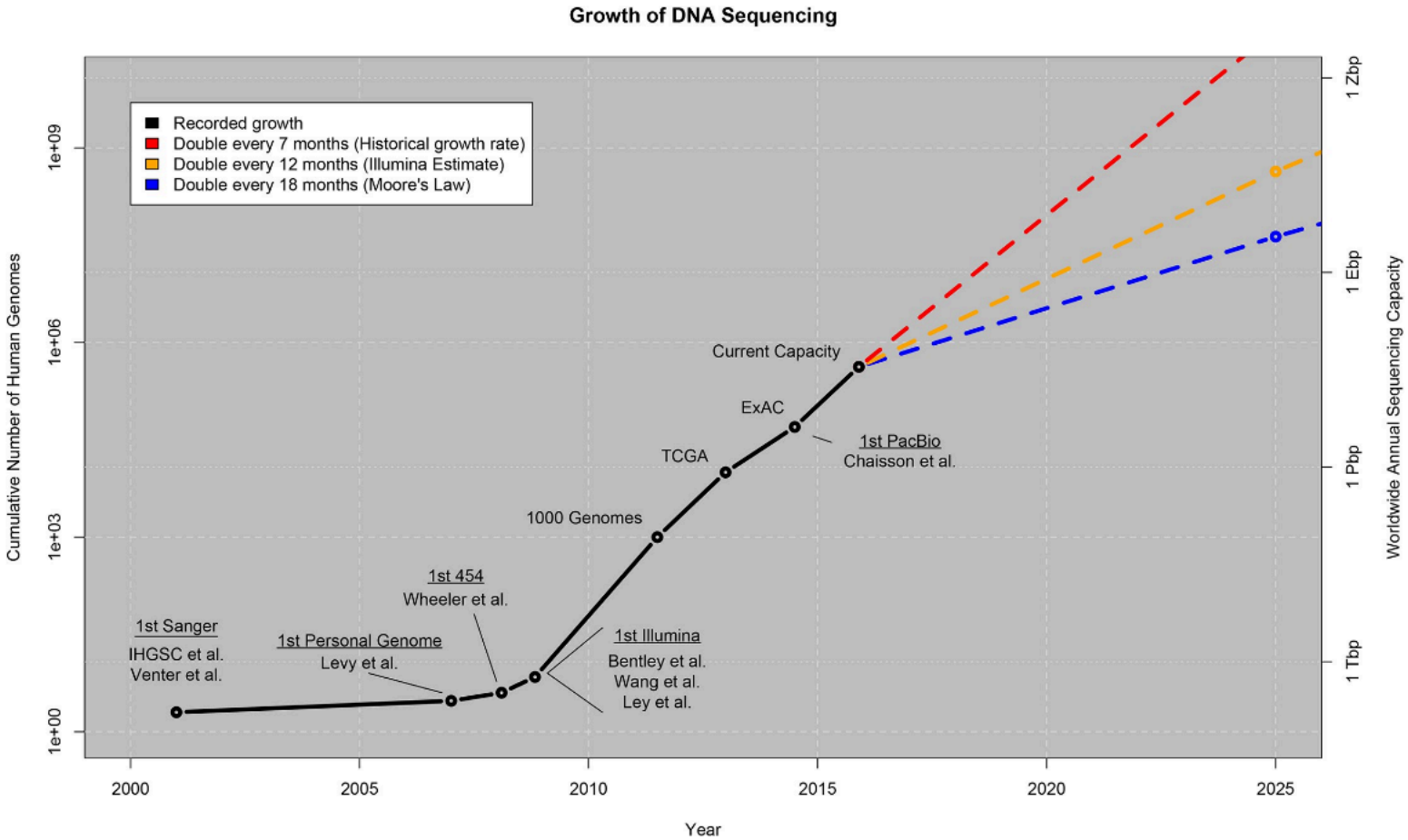
Number of publicly available bases in the NCBI Sequence Read Archive (SRA) as of July 1, 2018. This is the equivalent of 6,153,232 human genomes (which is 3e+9 bases).

30TB

Sean Davis

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

DNA sequencing rates continues to grow.



http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

Very few people are trained in both data analysis and biology.

The *practical* and *pragmatic* issues of data analysis and data interpretation are not something that is taught at undergrad or graduate school level.

Most senior faculty do not know how to do this.

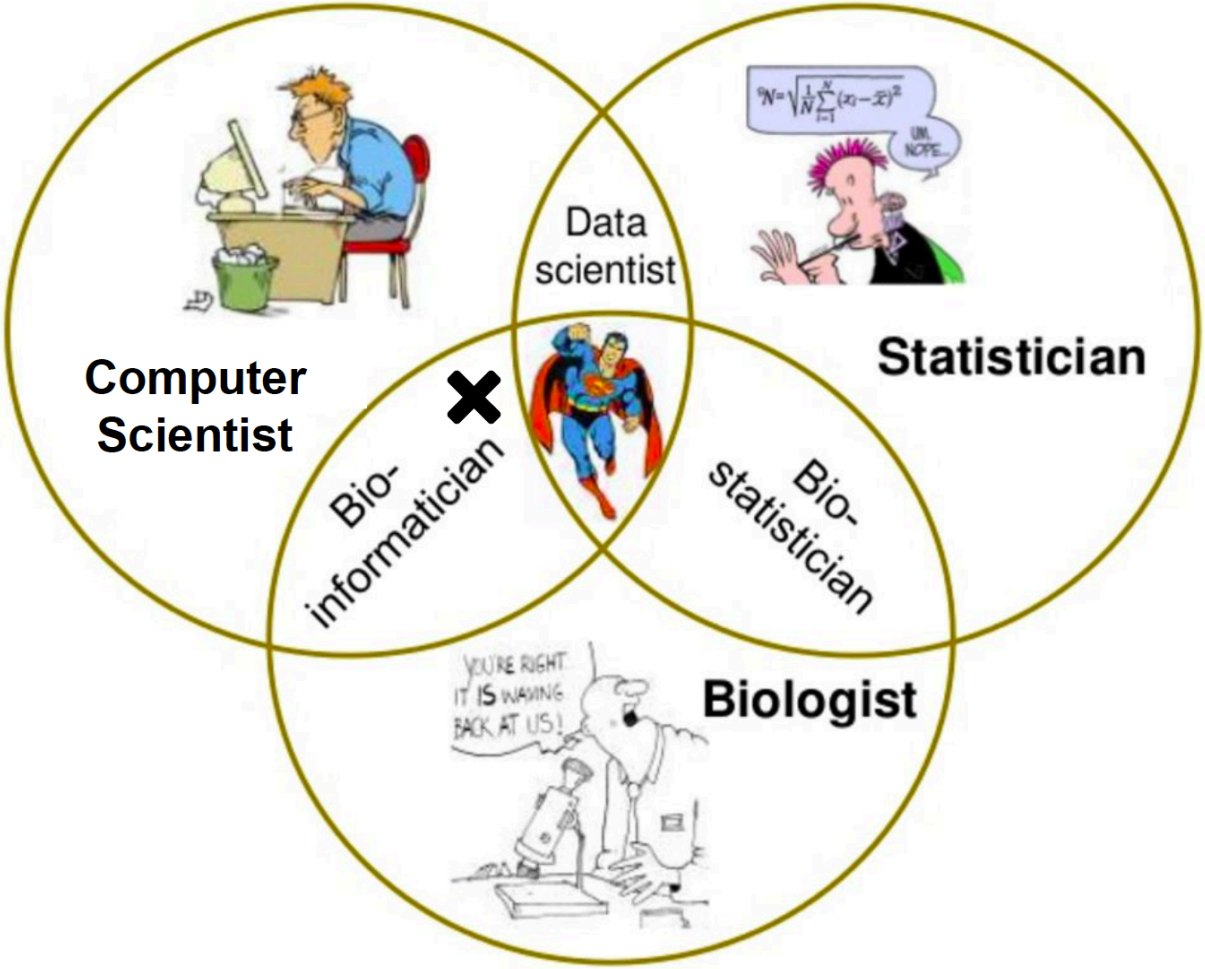
Nor do many junior faculty.

But our field increasingly *depends* on skilled interpretation of private + public data.

So how do we address this? Many layers, many approaches...

Credit: Titus Brown

Superman/Wonder woman



Credit: Torsten Seemann

What is bioinformatics

A brief history of bioinformatics

Jeff Gauthier ✉, Antony T Vincent, Steve J Charette, Nicolas Derome

Briefings in Bioinformatics, bby063, https://doi.org/10.1093/bib/bby063

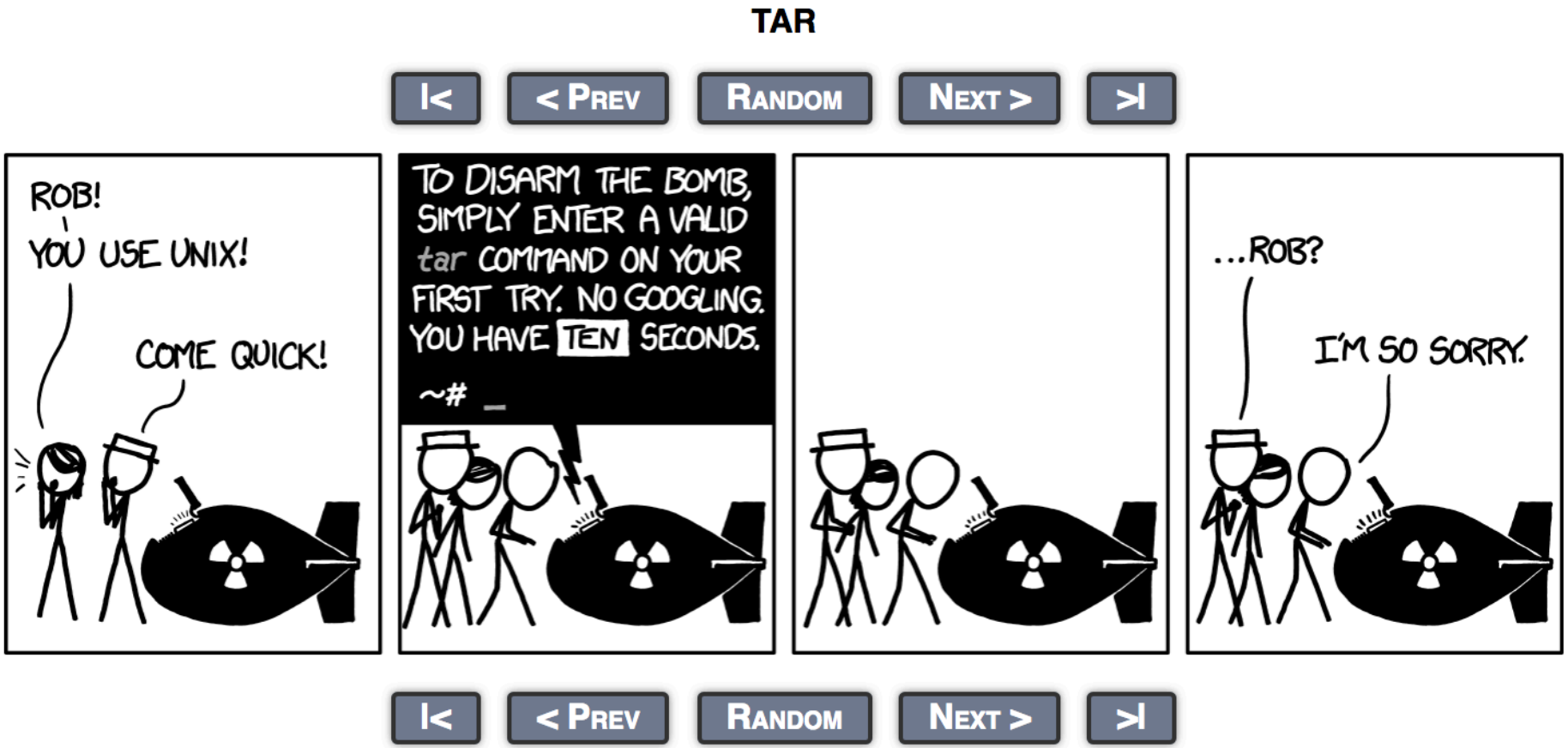
Published: 03 August 2018 **Article history** ▾

https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bby063/5066445

A typical day of my life as a bioinformatics scientist

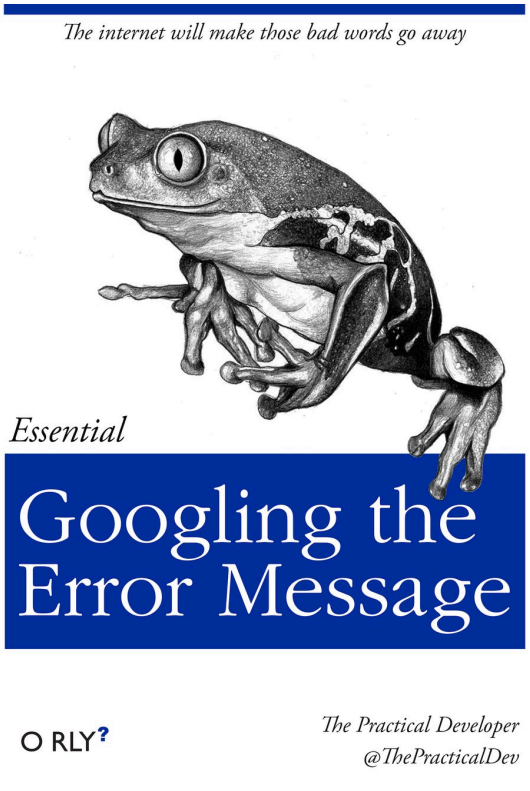
- Googling (error message etc)
- Converting file formats.
- Tidying the data.
- Installing software.
- Real analysis (plotting etc) 20%

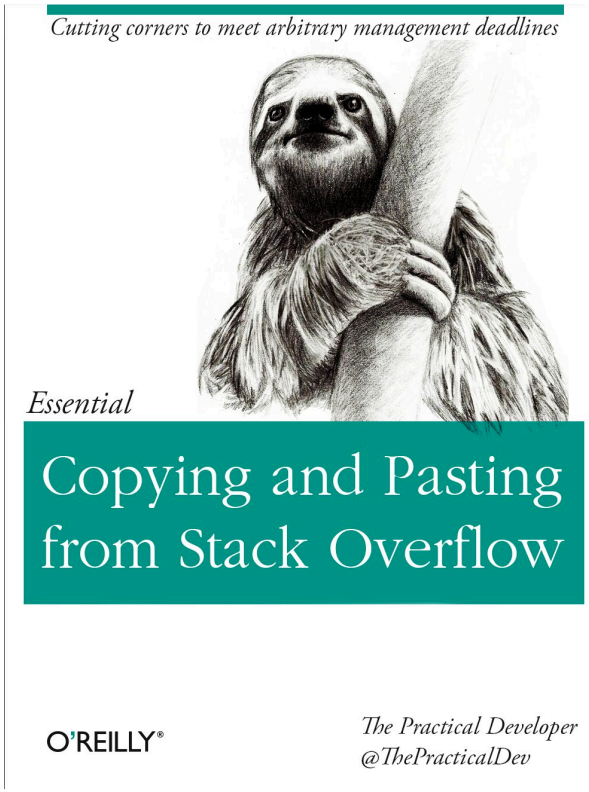
Google is what we do



Ask for help

- google
- SeqAnswer
- Biostars
- Stack overflow





bioiFORMATics

- A real variant calling example:
- Fastq
- sam
- bam
- Vcf
- Bed

conda and biocoda

Conda

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

Learn command line

- Why command line?
- More efficient/powerful
- HPC, cloud computing

Terminal

Welcome!

This is the **command-line bootcamp**, a tutorial that teaches you how to work at the command-line. You'll learn all the basic skills needed to start being productive in the UNIX terminal.

Please post feedback at the issue tracker

Table of Contents


```
learner@:~$
```

a_directory
another_dire…

1

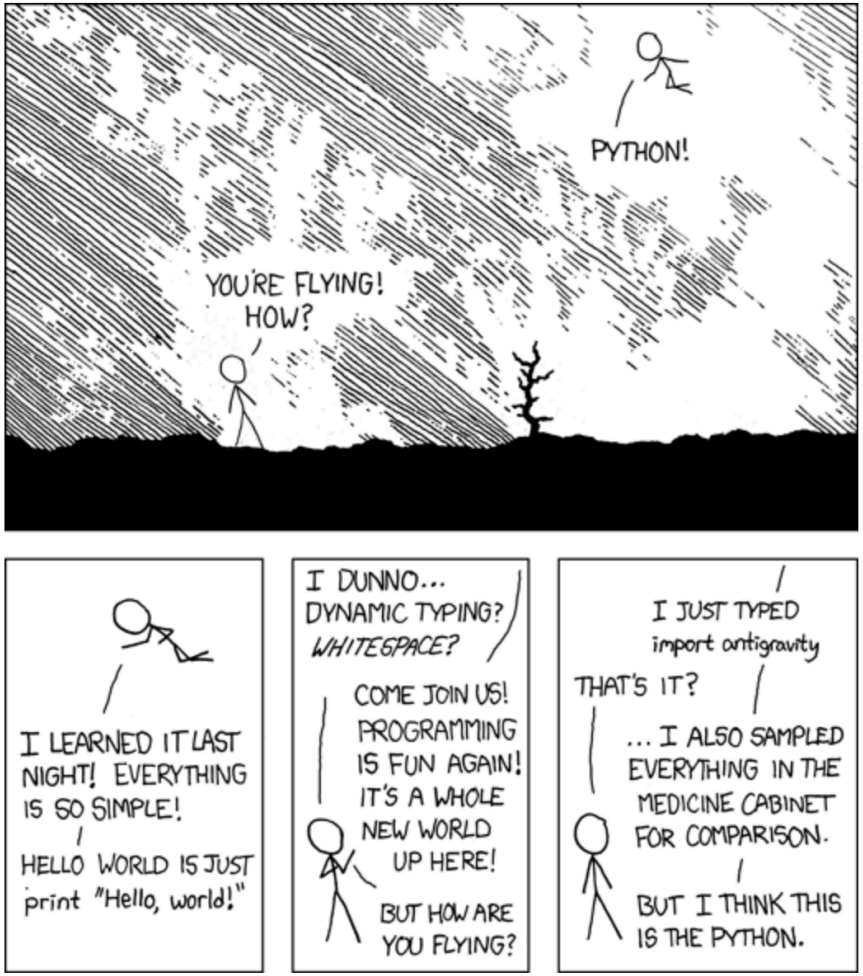
http://rik.smith-unna.com/command_line_bootcamp/

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

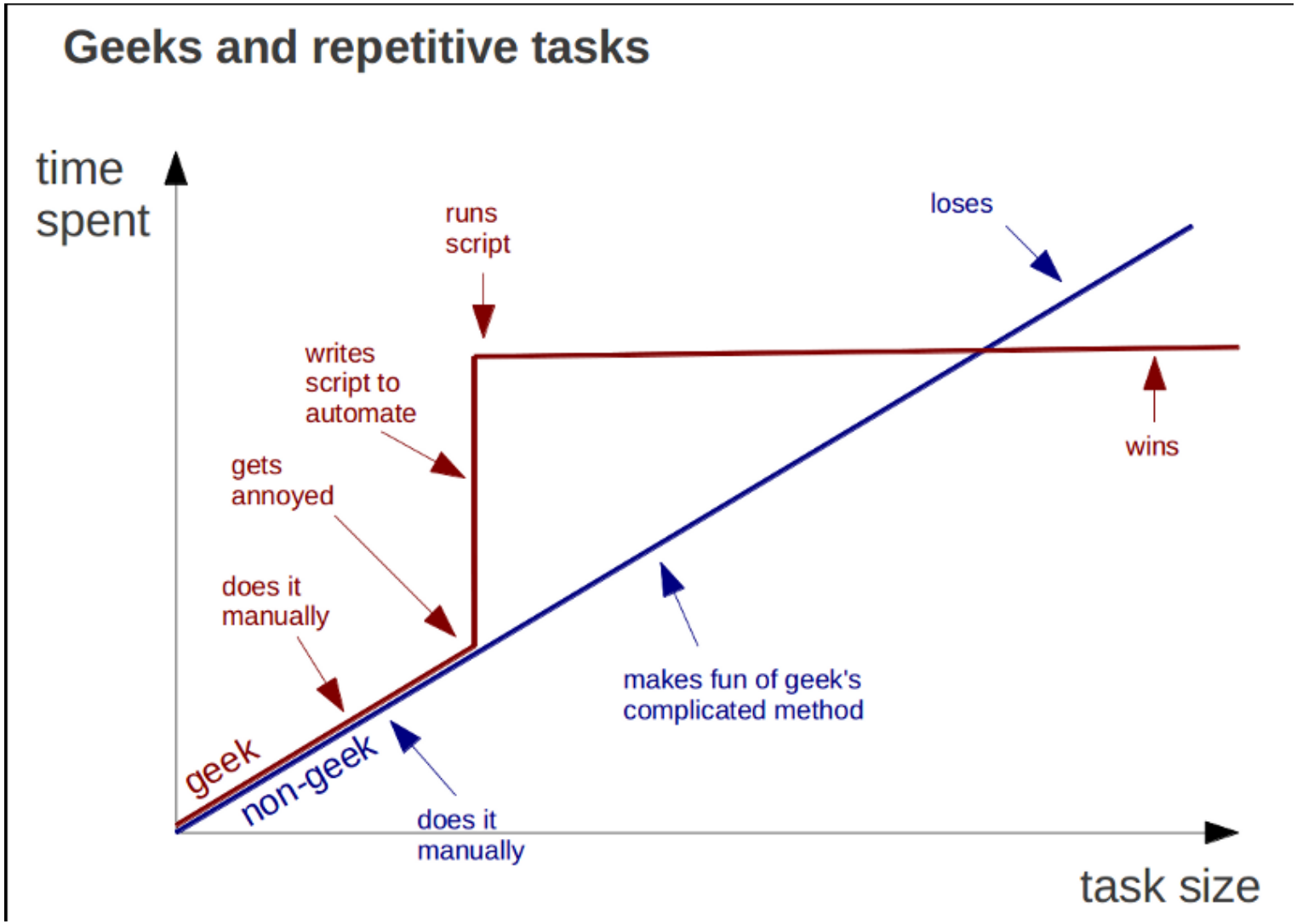
Learn some python

PYTHON

I< < PREV RANDOM NEXT > >I



Automation saves you time in the long run



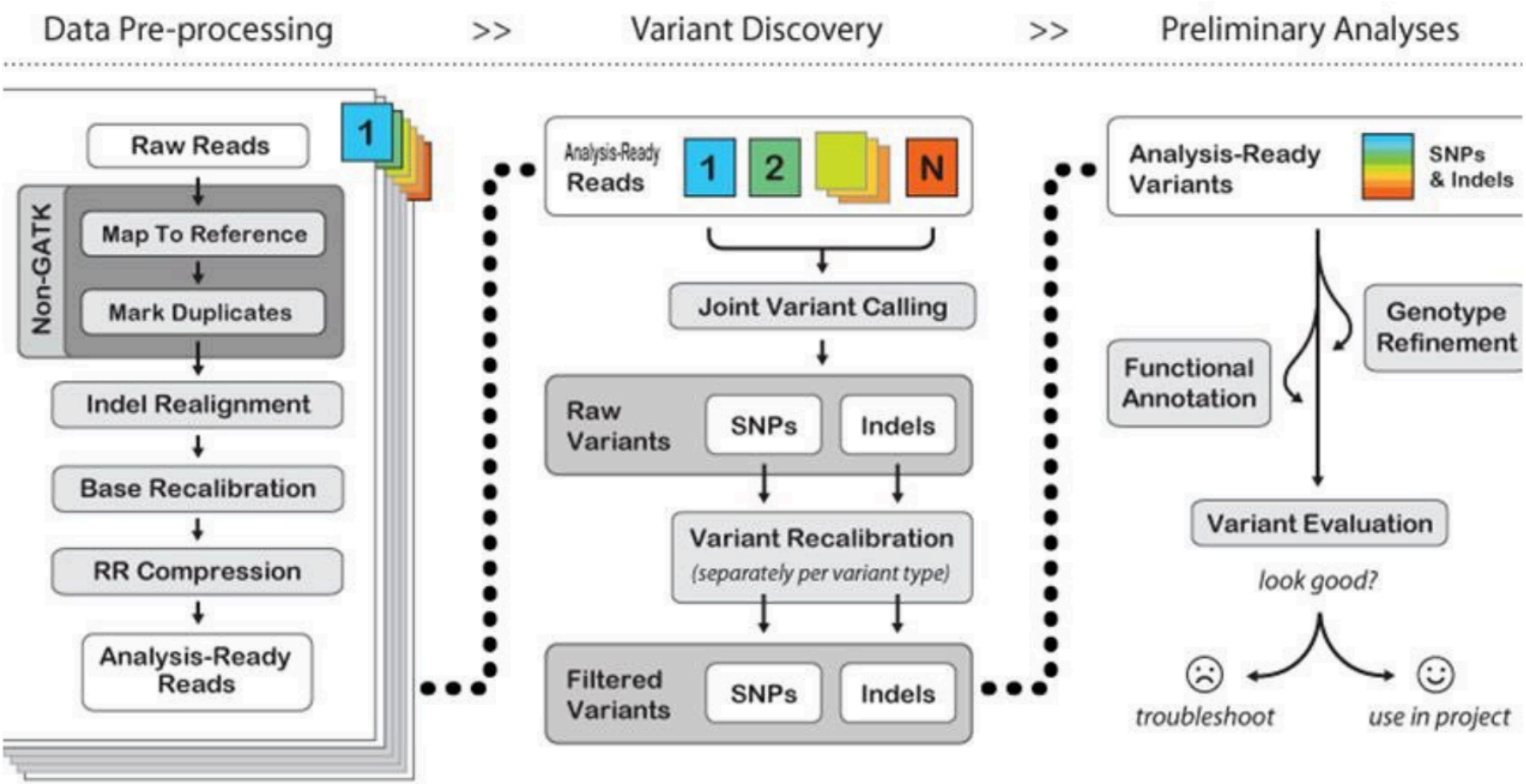
Computers are good at repetitive work

Side effect of automation

- The best documentation is automation
- Write scripts for everything unless it is not possible. (manual editing, document!)

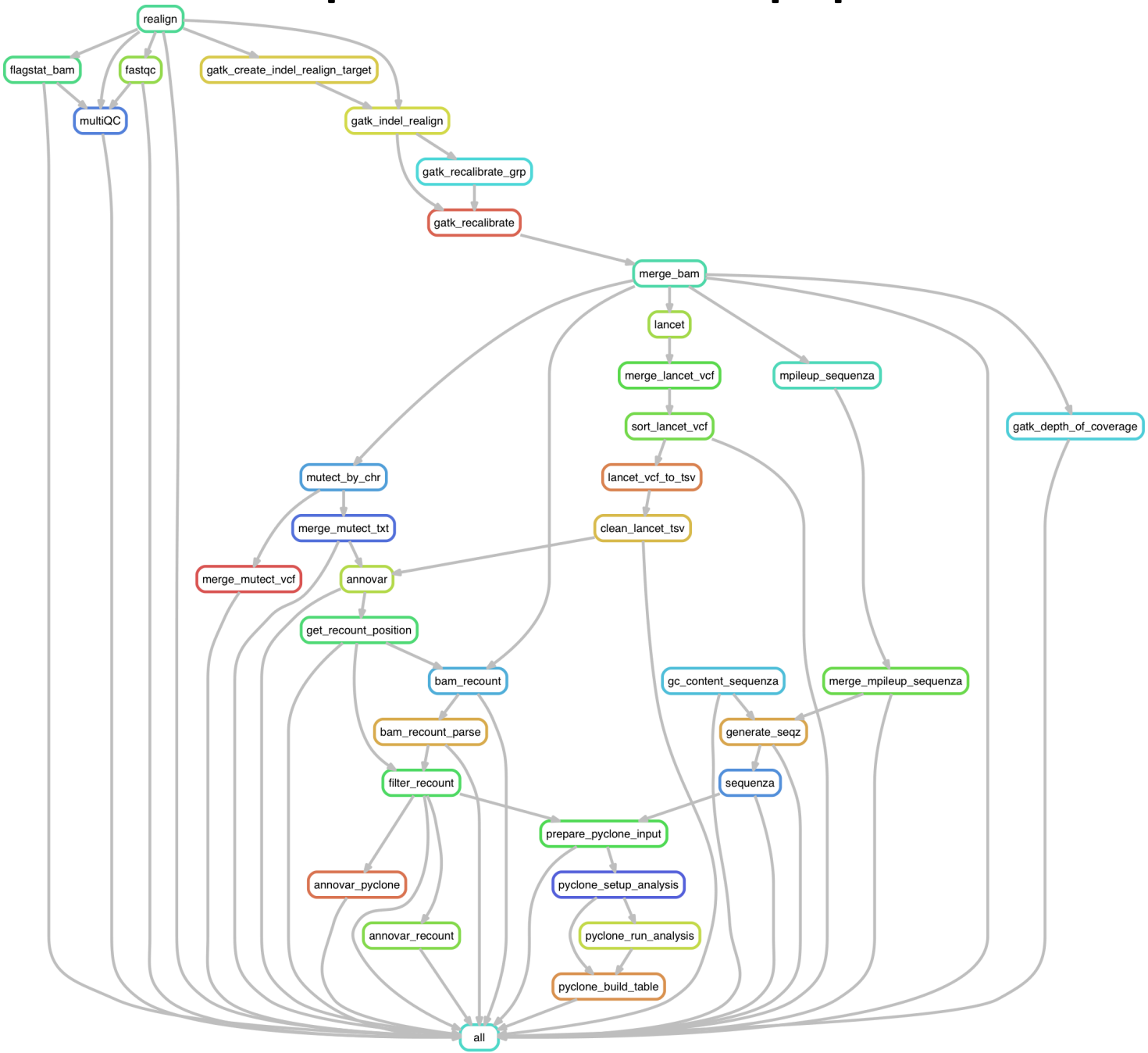
Credit to someone in the twitter-verse ☺

NGS data processing



Taken from: http://www.broadinstitute.org/gatk/guide/best-practices

DNA-seq Snakemake pipeline



A real run

```
[Sat Nov  4 21:01:13 2017] annovar for 06lancet/Pa01_pret_vs_Pa01_wbc_lancet_filtered.txt
[Sat Nov  4 21:01:13 2017]
[Sat Nov  4 21:01:16 2017] Finished job 2186.
[Sat Nov  4 21:01:16 2017] 244 of 599 steps (41%) done
[Sat Nov  4 21:01:33 2017] Finished job 3362.
[Sat Nov  4 21:01:33 2017] 245 of 599 steps (41%) done
[Sat Nov  4 21:01:41 2017] Finished job 2884.
[Sat Nov  4 21:01:41 2017] 246 of 599 steps (41%) done
[Sat Nov  4 21:02:14 2017] Finished job 2359.
[Sat Nov  4 21:02:14 2017] 247 of 599 steps (41%) done
[Sat Nov  4 21:02:24 2017] Finished job 2996.
[Sat Nov  4 21:02:24 2017] 248 of 599 steps (41%) done
[Sat Nov  4 21:02:42 2017] Finished job 2253.
[Sat Nov  4 21:02:42 2017] 249 of 599 steps (42%) done
[Sat Nov  4 21:02:43 2017] Finished job 2193.
[Sat Nov  4 21:02:43 2017] 250 of 599 steps (42%) done
[Sat Nov  4 21:02:46 2017] Finished job 76.
[Sat Nov  4 21:02:46 2017] 251 of 599 steps (42%) done
[Sat Nov  4 21:02:50 2017] Finished job 2267.
[Sat Nov  4 21:02:50 2017] 252 of 599 steps (42%) done
[Sat Nov  4 21:02:51 2017] Finished job 3347.
[Sat Nov  4 21:02:51 2017] 253 of 599 steps (42%) done
[Sat Nov  4 21:02:54 2017] Finished job 2265.
[Sat Nov  4 21:02:54 2017] 254 of 599 steps (42%) done
[Sat Nov  4 21:03:06 2017] Finished job 2346.
[Sat Nov  4 21:03:06 2017] 255 of 599 steps (43%) done
[Sat Nov  4 21:03:12 2017] Finished job 3324.
[Sat Nov  4 21:03:12 2017] 256 of 599 steps (43%) done
[Sat Nov  4 21:03:23 2017] Finished job 3001.
[Sat Nov  4 21:03:23 2017] 257 of 599 steps (43%) done
[Sat Nov  4 21:03:29 2017] Finished job 2636.
```

Output files from the pipeline are organized by folders and uniformly named

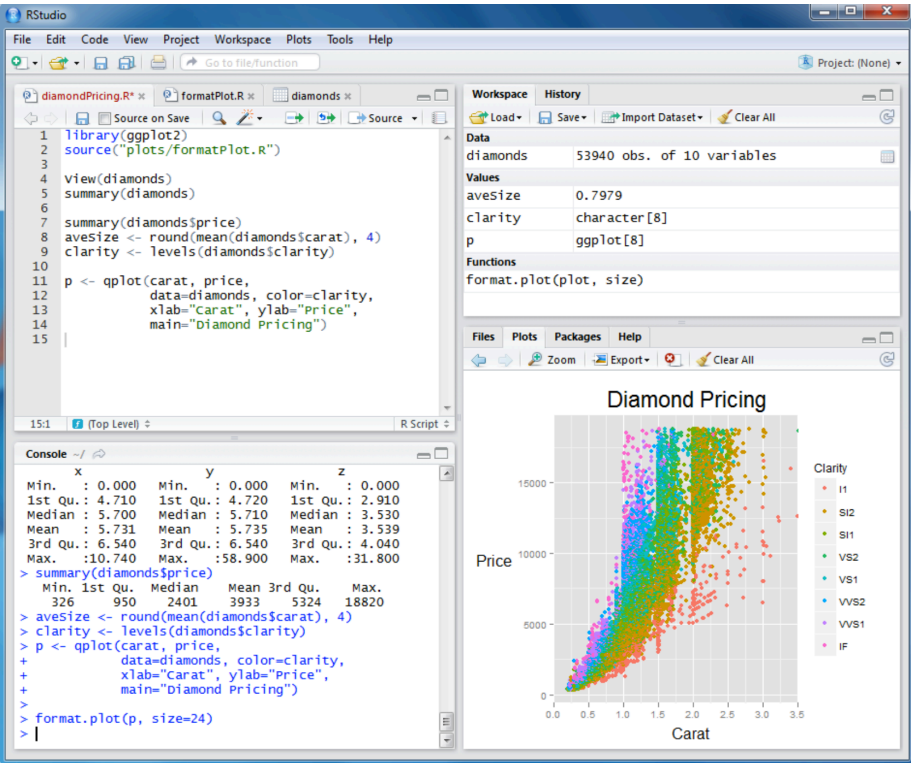
```
krai@chms019:~/mixing_histology_snakemake$ ls
00log                          04mutect_temp        11recount_annovar          bsub_log          pics
01aln                          05annovar            11recount_table_alt_fill   cluster.json      pyflow-DNAseq.sh
01aln_temp                     06lancet             11recount_table_filter     config.txt        README.md
02fqc                          06lancet_temp        12mpileup_sequenza         config.yaml       run_summary.txt
03indel_rln_recal_bam          07annovar_extract    12mpileup_sequenza_temp    custom.lua        sample2json.py
03indel_rln_recal_bam_coverage 08recount_position   13seqz                     files             sample_info.txt
03indel_rln_recal_base_temp    09recount            14sequenza_out             gem.conf          sample_names.json
03indel_rln_recal_grp_temp     10multiQC            15pyclone_input            jobscript.sh      samples.json
03indel_rln_temp               10recount_table      16pyclone_output           meta.txt          scripts
04mutect                       11pyclone_annovar    bsub_cluster.py            mixing_meta.txt   Snakefile
```

```
krai@chms019:~/mixing_histology_snakemake$ ls 03indel_rln_recal_bam
Pa25_N_rln_recal.bai            Pa29_N_rln_recal.bam.bai        Pa31_T2_rln_recal.bai           Pa34_T2_rln_recal.bam.bai
Pa25_N_rln_recal.bam            Pa29_N_rln_recal.bam.flagstat   Pa31_T2_rln_recal.bam           Pa34_T2_rln_recal.bam.flagstat
Pa25_N_rln_recal.bam.bai        Pa29_T1_rln_recal.bai           Pa31_T2_rln_recal.bam.bai       Pa35_N_rln_recal.bai
Pa25_N_rln_recal.bam.flagstat   Pa29_T1_rln_recal.bam           Pa31_T2_rln_recal.bam.flagstat  Pa35_N_rln_recal.bam
Pa25_T1_rln_recal.bai           Pa29_T1_rln_recal.bam.bai       Pa32_N_rln_recal.bai            Pa35_N_rln_recal.bam.bai
Pa25_T1_rln_recal.bam           Pa29_T1_rln_recal.bam.flagstat  Pa32_N_rln_recal.bam            Pa35_N_rln_recal.bam.flagstat
Pa25_T1_rln_recal.bam.bai       Pa29_T2_rln_recal.bai           Pa32_N_rln_recal.bam.bai        Pa35_T1_rln_recal.bai
Pa25_T1_rln_recal.bam.flagstat  Pa29_T2_rln_recal.bam           Pa32_N_rln_recal.bam.flagstat   Pa35_T1_rln_recal.bam
Pa25_T2_rln_recal.bai           Pa29_T2_rln_recal.bam.bai       Pa32_T1_rln_recal.bai           Pa35_T1_rln_recal.bam.bai
Pa25_T2_rln_recal.bam           Pa29_T2_rln_recal.bam.flagstat  Pa32_T1_rln_recal.bam           Pa35_T1_rln_recal.bam.flagstat
Pa25_T2_rln_recal.bam.bai       Pa29_T3_rln_recal.bai           Pa32_T1_rln_recal.bam.bai       Pa35_T2_rln_recal.bai
Pa25_T2_rln_recal.bam.flagstat  Pa29_T3_rln_recal.bam           Pa32_T1_rln_recal.bam.flagstat  Pa35_T2_rln_recal.bam
```

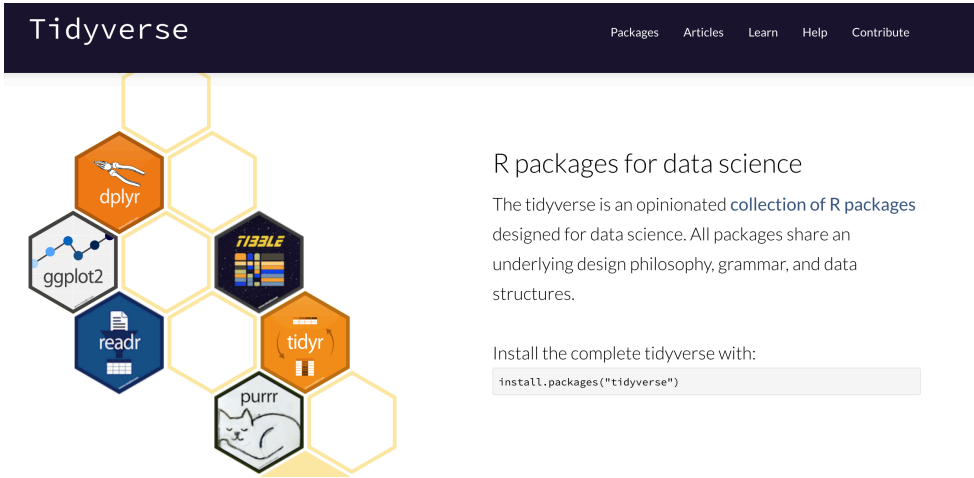
Learn some R

- Rstudio (IDE)
- Bioconductor
- Tidyverse and ggplot2

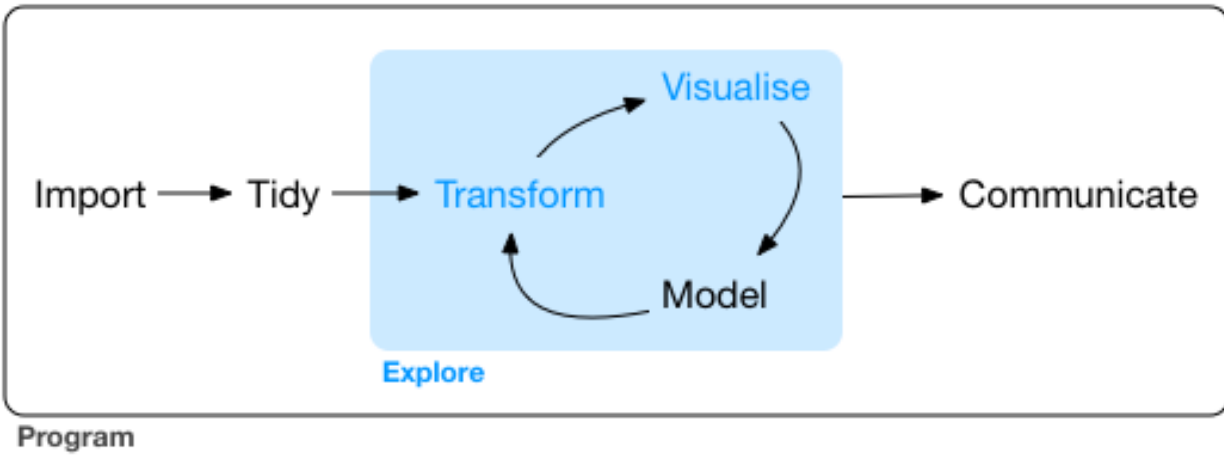
Do not re-invent the wheels

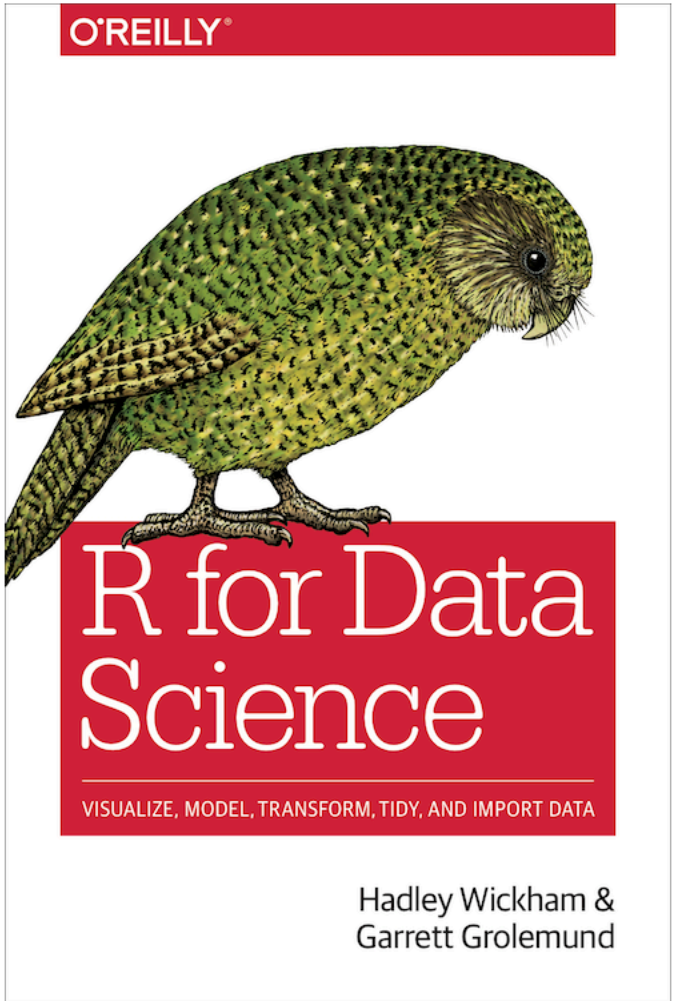


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



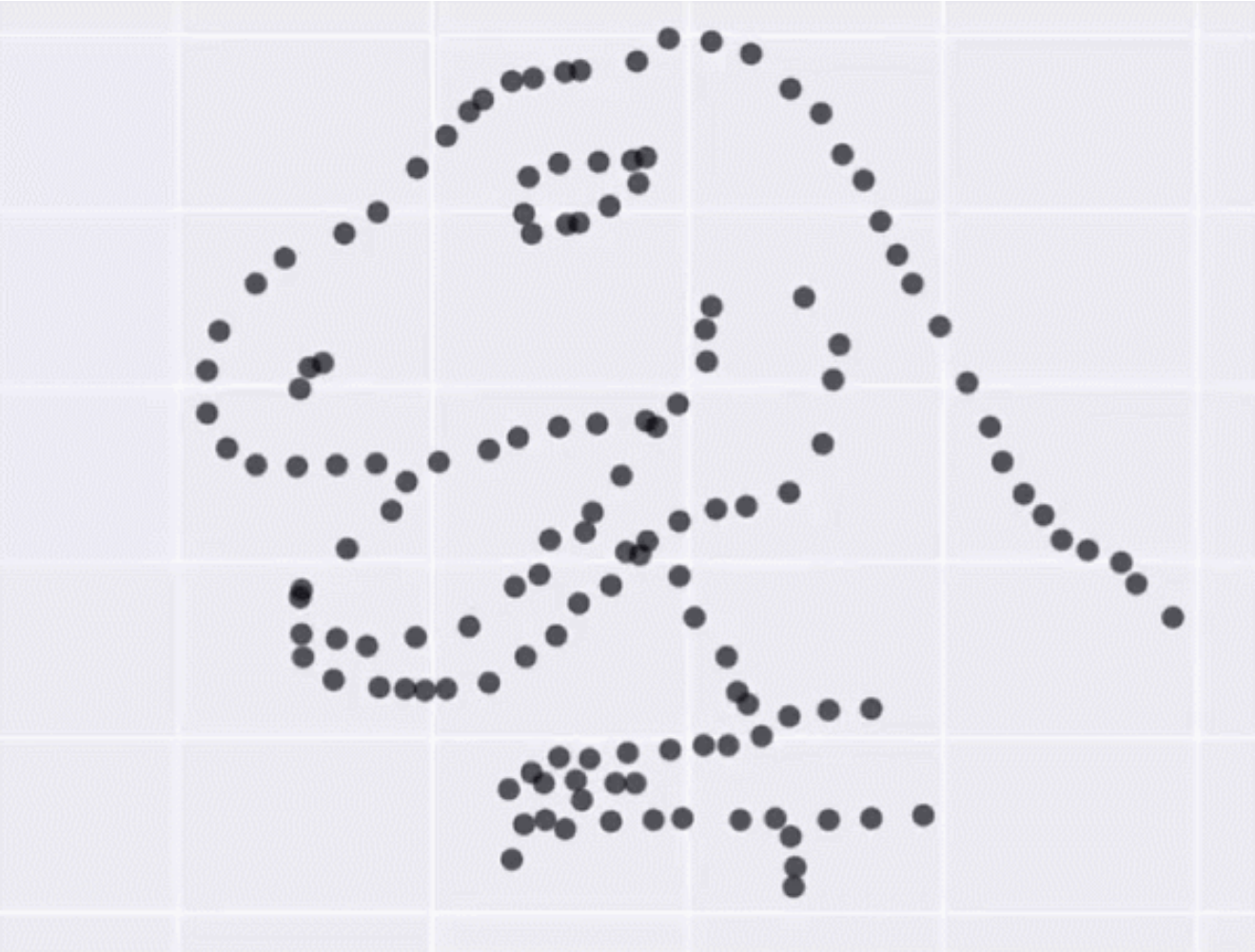
Tidying data





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

Data visualization



Wait, do you really need to learn C and C++?

Living in an Ivory Basement Stochastic thoughts on science, testing, and programming.

misc | personal | python | science | teaching | testing | training

Towards a bioinformatics middle class

Titus Brown

http://ivory.idyll.org/blog/2015-bioinformatics-middle-class.html

Core competencies for each bioinformatics training category.

	Bioinformatics User	Bioinformatics Scientist	Bioinformatics Engineer
(a) An ability to apply knowledge of computing, biology, statistics, and mathematics appropriate to the discipline.		X	X
(b) An ability to analyze a problem and identify and define the computing requirements appropriate to its solution.		X	X
(c) An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs in scientific environments.			X
(d) An ability to use current techniques, skills, and tools necessary for computational biology practice.	X	X	X
(e) An ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modeling and design of computer-based systems in a way that demonstrates comprehension of the tradeoffs involved in design choices.			X
(f) An ability to apply design and development principles in the construction of software systems of varying complexity.			X
(g) An ability to function effectively on teams to accomplish a common goal.	X	X	X
(h) An understanding of professional, ethical, legal, security, and social issues and responsibilities.	X	X	X
(i) An ability to communicate effectively with a range of audiences.	X	X	X
(j) An ability to analyze the local and global impact of bioinformatics and genomics on individuals, organizations, and society.	X	X	X
(k) Recognition of the need for and an ability to engage in continuing professional development.	X	X	X
(l) Detailed understanding of the scientific discovery process and of the role of bioinformatics in it.	X	X	X
(m) An ability to apply statistical research methods in the contexts of molecular biology, genomics, medical, and population genetics research.	X	X	X
(n) Knowledge of general biology, in-depth knowledge of at least one area of biology, and understanding of biological data generation technologies.	X	X	X

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3945096/

Reproducibility crisis

http://biobungalow.weebly.com/bio-bungalow-blog/everybody-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…)

Method matters

RESEARCH ARTICLE

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

Nathaniel D. Anderson[1,2], Richard de Borja[1,*], Matthew D. Young[3,*], Fabio Fuligni[1,*], Andrej Rosic[1], Nicola D. Roberts[3], Simo…
+ See all authors and affiliations

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

Credit: Nicolas Robine

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.

How to ensure reproducibility

- Git version control
- Jupyter/R Notebook, documentation
- Containers (docker, singularity, biocontainers
- https://biocontainers.pro/)

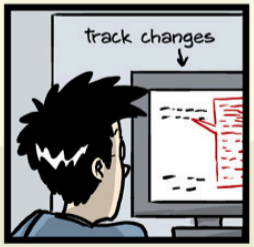
"FINAL".doc

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

WWW.PHDCOMICS.COM

Version control

- Git
- Github
- Gitlab

Notebooks

jupyter nbviewer

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, therefor 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` run the current cell and enters command mode.

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.

https://cyverse-cybercarpentry-container-workshop-2018.readthedocs-hosted.com/en/latest/dock
https://ropenscilabs.github.io/r-docker-tutorial/01-what-and-why.html

Other important untaught skills

- Naming files
- Project organization
- Data organization, backup plans

Naming files

- Three principles for (file) names:
- 1. Machine readable (do not put special characters and space in the name)
- 2. Human readable (Easy to figure out what the heck something is, based on its name, add slug)
- 3. Plays well with default ordering:
- * Put something numeric first
- * Use the ISO 8601 standard for dates (YYYY-MM-DD)
- * Left pad other numbers with zeros

Punctuation

Deliberate use of `"-"` and `"_"` allows recovery of meta-data from the filenames:

- `"_"` underscore used to delimit units of meta-data I want later
- `"-"` hyphen used to delimit words so my eyes don't bleed

```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
```

```
> flist <- list.files(pattern = "Plasmid") %>% head

> stringr::str_split_fixed(flist, "[_\\.]", 5)
     [,1]         [,2]             [,3]                                  [,4]  [,5]
[1,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A01" "csv"
[2,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A02" "csv"
[3,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A03" "csv"
[4,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B01" "csv"
[5,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B02" "csv"
[6,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B03" "csv"
```

date assay sample set well

This happens to be R but also possible in the shell, Python, etc.

Jenny Bryan

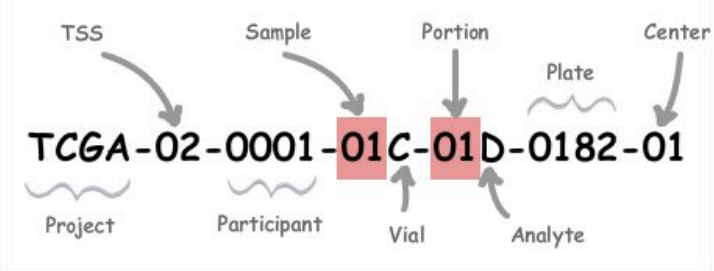
Go forth and use awesome file names :)

```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
```

```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

Jenny Bryan:
https://rawgit.com/Reproducible-Science-Curriculum/rr-organization1/master/organization-01-s

TCGA barcode



Label	Identifier for	Value	Value Description	Possible Values
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	182	The 182nd plate	4-digit alphanumeric value
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	1	The first portion of the sample	01-99
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Project	Project name	TCGA	TCGA project	TCGA
Sample	Sample type	1	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Center	Sequencing or characterization center that will receive the aliquot for analysis	1	The Broad Institute GCC	See Code Tables Report
Participant	Study participant	1	The first participant from MD Anderson for GBM study	Any alpha-numeric value
TSS	Tissue source site	2	GBM (brain tumor) sample from MD Anderson	See Code Tables Report

Organization of each project down-stream analysis

```
→  SKCM_IMT git:(master) ✗ ls
README.md      SKCM_IMT.Rproj data          docs          presentation   results        scripts
→  SKCM_IMT git:(master) ✗ tree -d
.
├── data
│   ├── 12chromHMM_1000bp_15State
│   │   └── MYOUTPUT
│   ├── 12chromHMM_1000bp_15State_24Samples
│   │   └── MYOUTPUT
│   ├── chromHMM
│   ├── copynumber
│   └── metadata
├── docs
├── presentation
├── results
└── scripts

12 directories
```

Rstudio R project

Go to file/function | Addins

pset.R... | 08_pyclone_out_visualization.Rmd | 09_left_join_pyclone_annovar.Rmd | 10_lancet_mutect_venn.Rmd | 11_MAF

Knit | Insert | Run

```
14
15 ```{bash}
16 cd /Users/mtang1/projects/SCLC_chemoRT_resistance/SCLC_chemoRT/data/mutect_lancet_anno_extracted
17 rsync -avhP railab:SCLC_WES_7_patients_snakemake/07annovar_extract/ .
18
19 ```
20
21
22 ```{r}
23 #devtools::install_github("js229/Vennerable")
24 library(Vennerable)
25 library(tidyverse)
26 library(here)
27
28 ann_extracted<- here("data/mutect_lancet_anno_extracted")
29
30 files<- list.files(ann_extracted, pattern = "txt", full.names = T)
31
32 datlist <- lapply(files, function(f) {
33         dat = read.table(f, header =T, sep ="\t", quote = "\"")
34         return(dat)
35 })
36
37 dat <- do.call(rbind, datlist)
```

Environment | History | Git

- 01_QC_bamcoverage.Rmd
- 02_extract_annotated_tsv.Rmd
- 03_venn_diagram_combine_callers.Rmd
- 04_get_recount_positions.Rmd
- 05_clean_gatk_variants_to_table.Rmd
- sequenza.Rmd
- 06_filter_recount_fill_in_alt.Rmd
- 07_generate_pyclone_input.Rmd
- 08_recount_venn_diagram_upset.Rmd
- 08_pyclone_out_visualization.Rmd
- 09_left_join_pyclone_annovar.Rmd
- 10_lancet_mutect_venn.Rmd
- 11_MAFtools_explore.Rmd
- 12_pyclone_cancer_gene_anno.Rmd
- 13_copynumber_cancer_genes.Rmd
- 14_MATH_scores.Rmd
- 15_cancer_evolution.Rmd

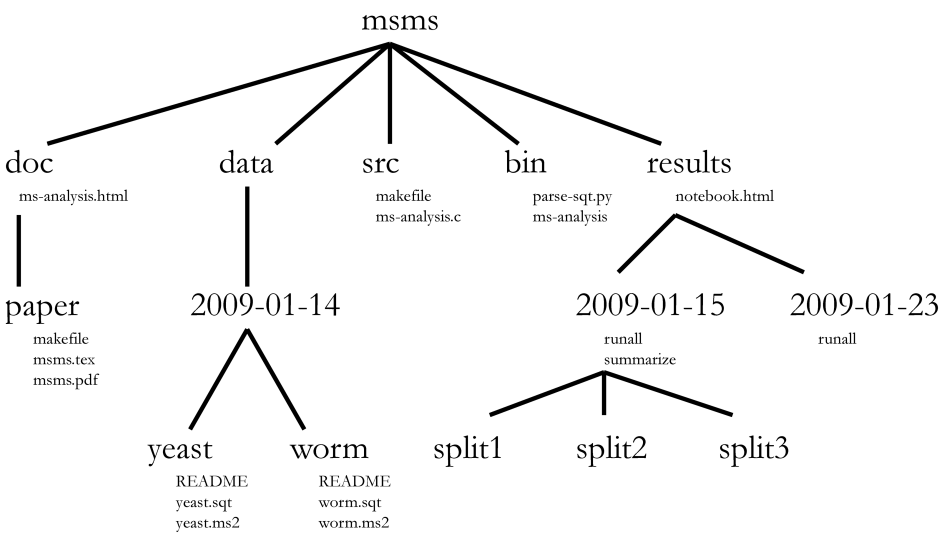
PLOS | COMPUTATIONAL BIOLOGY

EDUCATION

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble

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



PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

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

PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

Browse | Publish | About

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

Stay the current of bioinformatics

- Bioinformatics evolves so fast!
- E.g. sequencing technology: long-read (pacbio, nanopore, single cell) all these require new tools to analyze the associated data.
- I started bioinformatics after reading:

Getting Genetics Done

Getting Things Done in Genetics & Bioinformatics Research

TUESDAY, MAY 29, 2012

How to Stay Current in Bioinformatics/Genomics

http://www.gettinggeneticsdone.com/2012/05/how-to-stay-current-in.html

Social medium network

- Twitter
- Follow papers/tools, jobs, outreaching
- Go to conferences/talk to other people
- Blog posts

Ming Tang

Research Scientist

MD Anderson Cancer Center

About me

I am a computational biologist working on genomics, epigenomics and transcriptomics. I use R primary for data wrangling and visualization in the tidyverse ecosystem; I use python for writing Snakemake workflows; I am a unix geek learning shell tricks almost every day; I care about reproducible research and open science.

Being trained in a wet lab in the University of Florida during my PhD in Dr.Jianrong Lu's lab has established my solid knowledge and skills in experimental molecular cancer biology. Self-teaching and postdoctoral training in Dr.Roel Verhaak's lab has extended my bioinformatics skills in integrating analysis of TB size sequencing data sets. Verhaak lab is well known for studying genomic alterations of brain tumor by analyzing large panels of RNA-seq and DNA-seq data. I gained extensive experience in handling large-scale genomic data and pipelining workflows. I also gained intimate familiarities with public data sets such as ENCODE, TCGA and CCLE. I have put my analysis notes and snakemake pipelines for processing whole-exome, whole-genome DNAseq, RNAseq, single-cell RNAseq, ChIP-seq, ATACseq and RRBS data in my github repos.

https://divingintogeneticsandgenomics.rbind.io/

(III) Advice to the next generation

(or two generations, if you want me to feel *really* old.)

a. Get involved with a broad group of people and ideas (social media FTW!)
b. Learn something about both computing *and* biology.
c. Realize that you have nothing but opportunity, and that there has never been a better time to be in bio research!

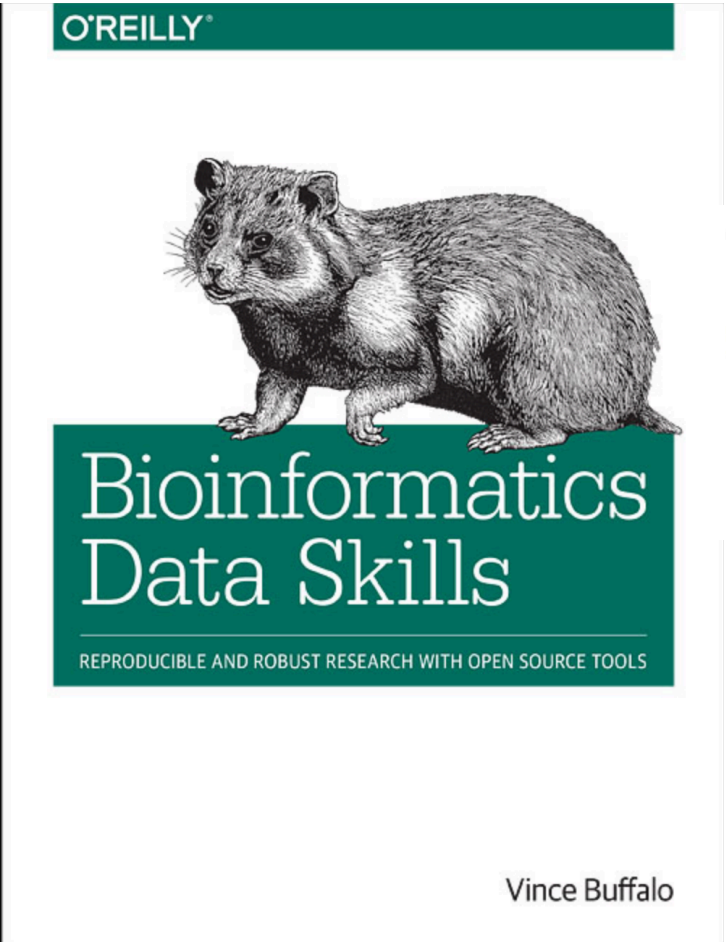
Where to start

- Never too late to start to learn!
- ANGUS workshop
- http://angus.readthedocs.io/en/2018/
- I would love to come back and give workshops!
- Software/Data Carpentries 2-day workshop (Ethan White is heavily involved at UF)
- Many other resources
- https://github.com/crazyhottommy/getting-started-with-genomics-tools-and-resources

Massive Online open Courses(MOOCs) and others

- 1. coursera
- 2. Edx
- 3. Udacity
- 4. Datacamp (not free, interactive lessons)

Learn by doing



What questions do you have?

Acknowledgements

- Thanks Titus Brown, Torsten Seemann and Sean Davis for letting me borrow their slides.
- Thanks Stephen Turner for writing his blog posts.
- Thanks Roel Verhaak lab for giving me the opportunity to learn computational biology.
- Thanks Samir Amin for teaching me so much!

Use excel wisely

Article

Data Organization in Spreadsheets

Karl W. Broman & Kara H. Woo

Pages 2-10 | Received 01 Jun 2017, Accepted author version posted online: 29 Sep 2017, Published online: 29 Sep 2017

Download citation https://doi.org/10.1080/00031305.2017.1375989 Check for updates

Caution with excel

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren and Assam El-Osta

Genome Biology 2016 **17**:177
https://doi.org/10.1186/s13059-016-1044-7
Published: 23 August 2016

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7

Converted to dates

excel.gene2date.xls

	A	B	C	D	E	F	G	H	I	J	K
	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				

Sheet1 Sheet2

Ready Sum=0 SCRL CAPS NUM

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

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

Journal[a]	Number of Excel files screened	Number of gene lists found	Number of papers with gene lists	Number of supplementary files affected	Number of papers affected	Number of gene names converted
PLoS One	7783	2202	994	220	170	4240
BMC Genomics	11464	1650	801	218	158	4932
Genome Res	2607	580	251	114	68	3180
Nucleic Acids Res	2117	540	315	88	67	1661
Genome Biol	2678	664	257	97	63	1878
Genes Dev	932	395	190	75	55	1593
Hum Mol Genet	980	372	168	48	27	1724
Nature	482	150	74	27	23	1375
BMC Bioinformatics	1790	235	152	26	21	534
RNA	569	127	77	20	15	1341
Nat Genet	264	70	37	12	9	178
Bioinformatics	731	112	67	11	6	339
PLoS Comput Biol	177	79	32	6	6	46

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 …

12:27 PM - 6 Aug 2018

17 Retweets **21** Likes

9 17 21

https://www.sciencedirect.com/science/article/pii/S0018506X18302599?via%3Dihub

- https://www.youtube.com/watch?v=s3JldKoA0zw&feature=youtu.be

Regular expression