Introduction II: tools you need to analyse sequencing dataset

Isheng Jason Tsai

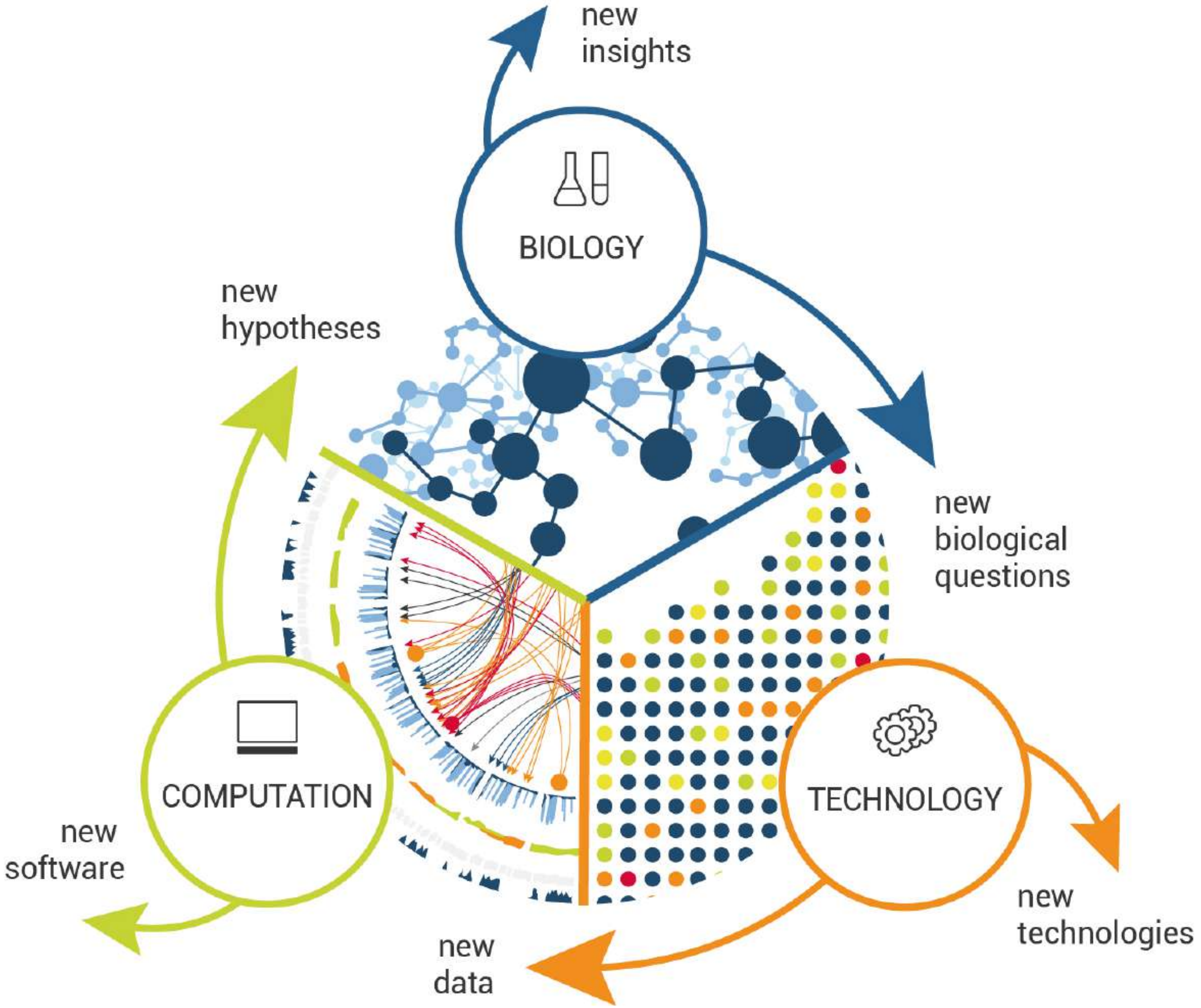
Introduction to NGS Data and Analysis
Lecture 2 ; v2020

This lecture aims to expose you to how computational biologists’ way of thinking (and any small topics that relevant)

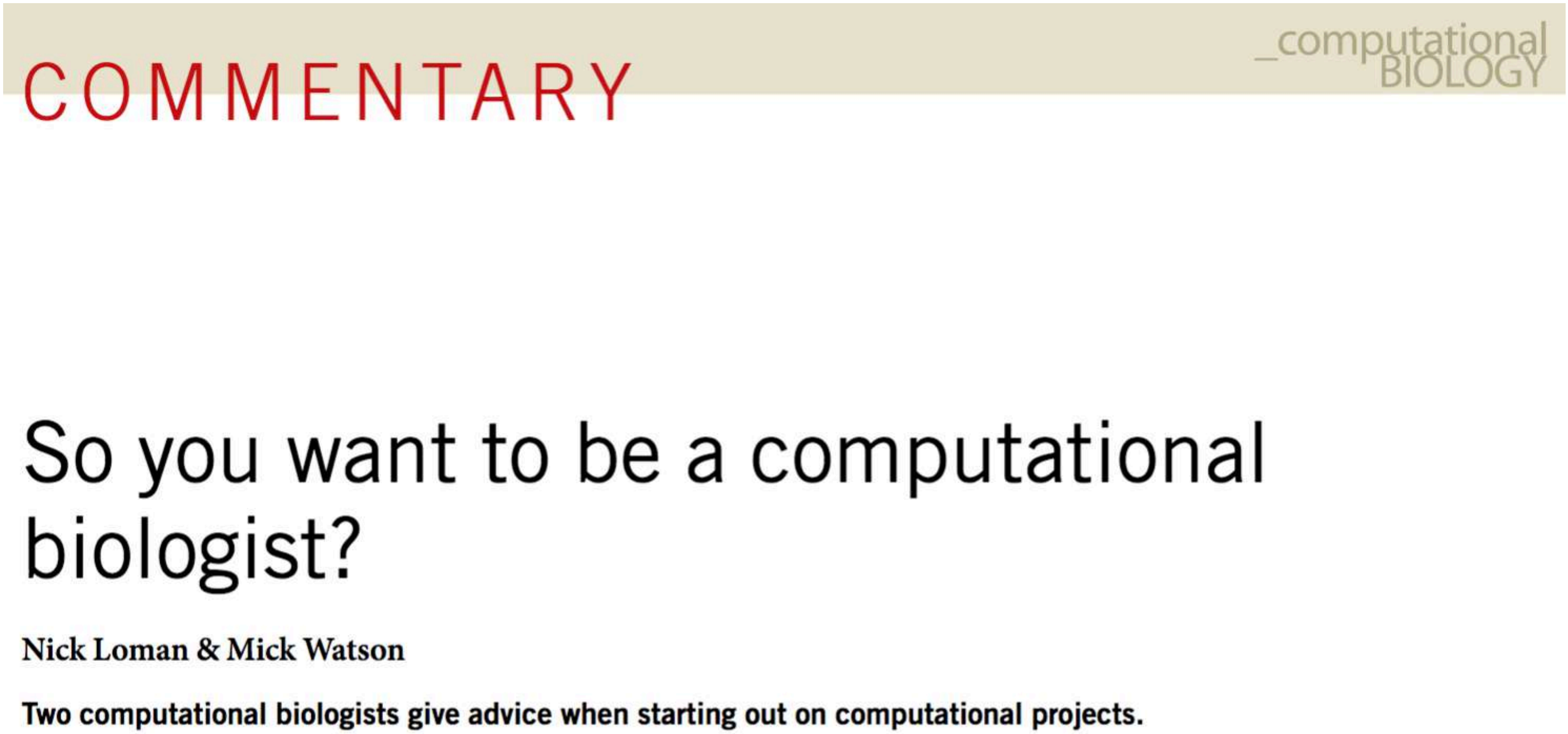
The content of this lecture will not be in the final exam. There is a written homework assignment.

Lecture outline

1. Research Effectively
2. Linux
3. Keep tracking / Reproducible research
4. Data Type / Visualisations
5. R



Credit: http://www.omicscouts.com



COMMENTARY

_computational BIOLOGY

So you want to be a computational biologist?

Nick Loman & Mick Watson

Two computational biologists give advice when starting out on computational projects.

https://www.nature.com/articles/nbt.2740.pdf

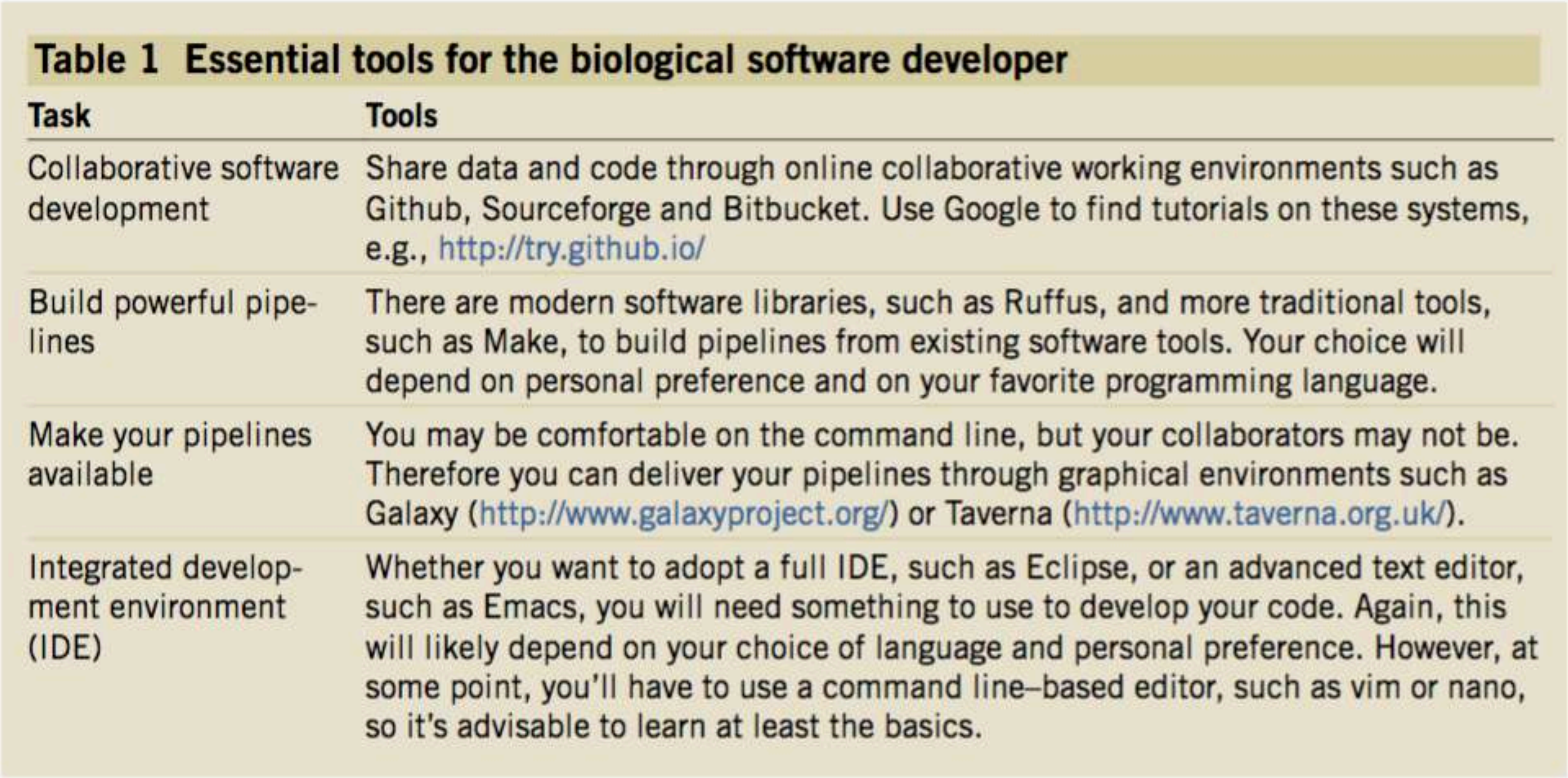


Table 1 Essential tools for the biological software developer

Task	Tools
Collaborative software development	Share data and code through online collaborative working environments such as Github, Sourceforge and Bitbucket. Use Google to find tutorials on these systems, e.g., http://try.github.io/
Build powerful pipelines	There are modern software libraries, such as Ruffus, and more traditional tools, such as Make, to build pipelines from existing software tools. Your choice will depend on personal preference and on your favorite programming language.
Make your pipelines available	You may be comfortable on the command line, but your collaborators may not be. Therefore you can deliver your pipelines through graphical environments such as Galaxy (http://www.galaxyproject.org/) or Taverna (http://www.taverna.org.uk/).
Integrated development environment (IDE)	Whether you want to adopt a full IDE, such as Eclipse, or an advanced text editor, such as Emacs, you will need something to use to develop your code. Again, this will likely depend on your choice of language and personal preference. However, at some point, you'll have to use a command line–based editor, such as vim or nano, so it's advisable to learn at least the basics.

https://www.nature.com/articles/nbt.2740.pdf

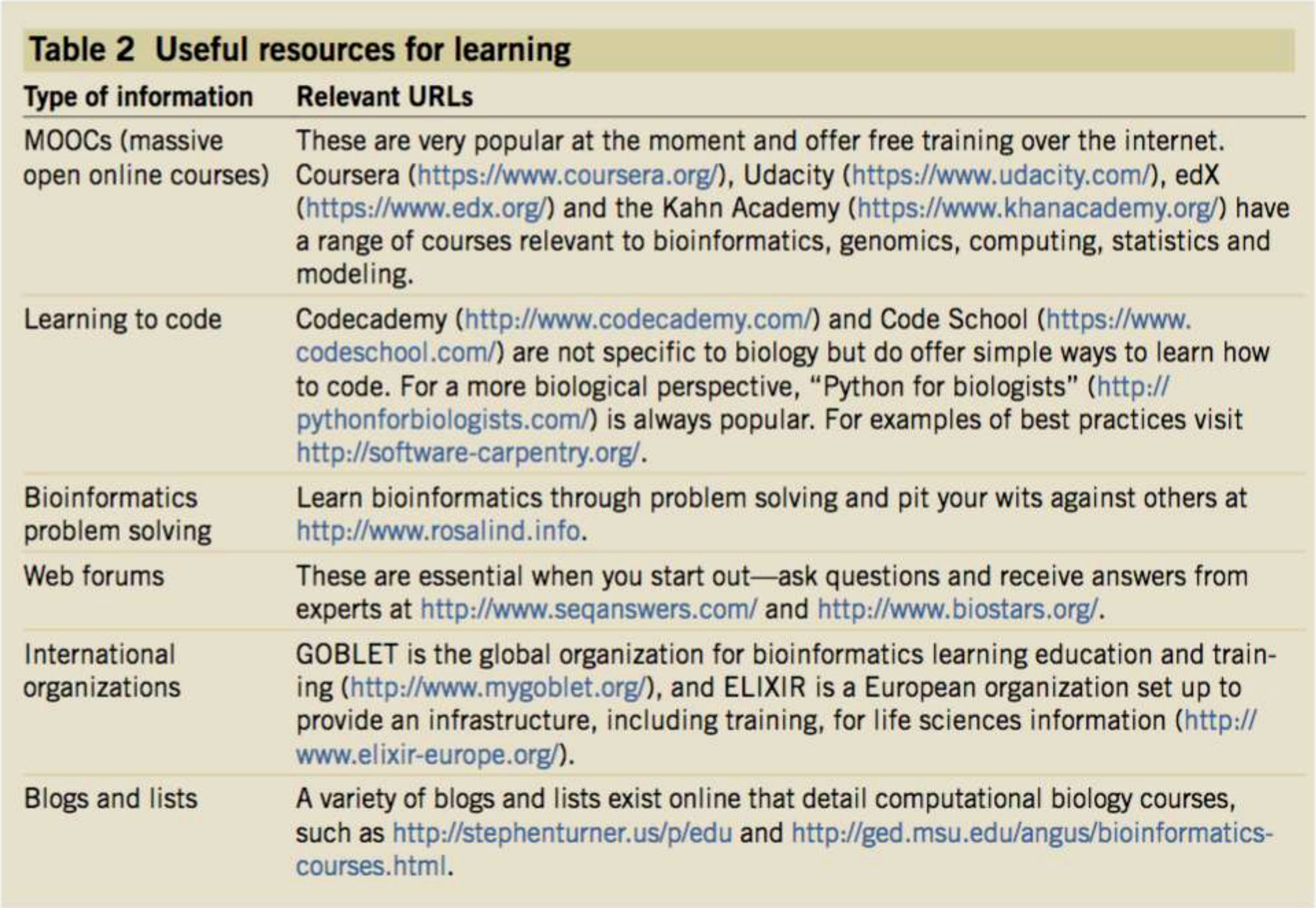


Table 2 Useful resources for learning

Type of information	Relevant URLs
MOOCs (massive open online courses)	These are very popular at the moment and offer free training over the internet. Coursera (https://www.coursera.org/), Udacity (https://www.udacity.com/), edX (https://www.edx.org/) and the Kahn Academy (https://www.khanacademy.org/) have a range of courses relevant to bioinformatics, genomics, computing, statistics and modeling.
Learning to code	Codecademy (http://www.codecademy.com/) and Code School (https://www.codeschool.com/) are not specific to biology but do offer simple ways to learn how to code. For a more biological perspective, "Python for biologists" (http://pythonforbiologists.com/) is always popular. For examples of best practices visit http://software-carpentry.org/.
Bioinformatics problem solving	Learn bioinformatics through problem solving and pit your wits against others at http://www.rosalind.info.
Web forums	These are essential when you start out—ask questions and receive answers from experts at http://www.seqanswers.com/ and http://www.biostars.org/.
International organizations	GOBLET is the global organization for bioinformatics learning education and training (http://www.mygoblet.org/), and ELIXIR is a European organization set up to provide an infrastructure, including training, for life sciences information (http://www.elixir-europe.org/).
Blogs and lists	A variety of blogs and lists exist online that detail computational biology courses, such as http://stephenturner.us/p/edu and http://ged.msu.edu/angus/bioinformatics-courses.html.

https://www.nature.com/articles/nbt.2740.pdf

Ten Simple Rules

"Ten Simple Rules" provide a quick, concentrated guide for mastering some of the professional challenges research scientists face in their careers.

http://collections.plos.org/ten-simple-rules

Research effectively / Data management

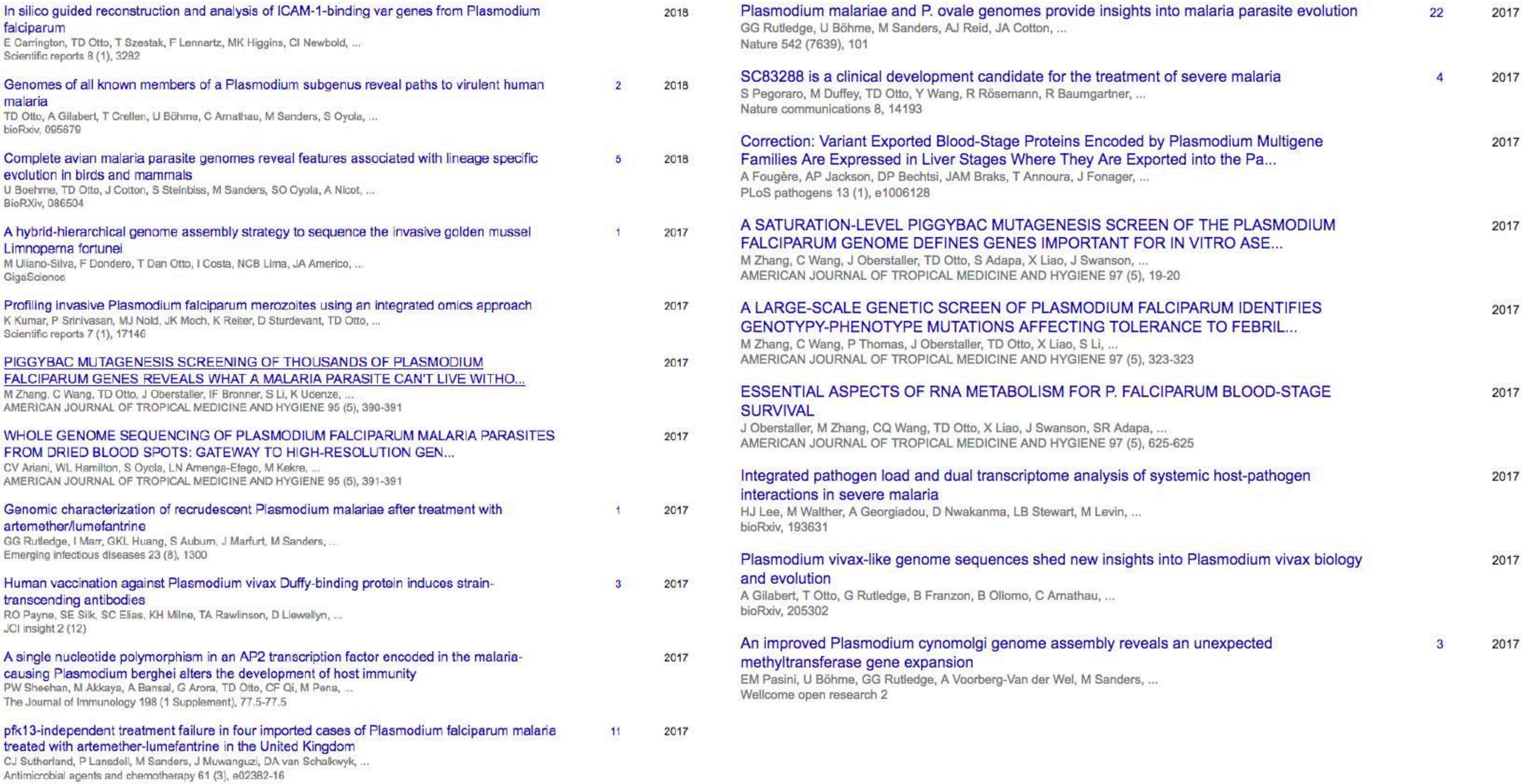
Thomas D. Otto

University of Glasgow

Verified email at glasgow.ac.uk - Homepage

Big Data Algorithms Omics

17 papers in 2017 ; how?
(I know he's doing the work)



In silico guided reconstruction and analysis of ICAM-1-binding var genes from Plasmodium falciparum E Carrington, TD Otto, T Szestak, F Lennartz, MK Higgins, CI Newbold, ... Scientific reports 8 (1), 3282		2018
Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria TD Otto, A Gilabert, T Crellen, U Böhme, C Arnathau, M Sanders, S Oyola, ... bioRxiv, 095679	2	2018
Complete avian malaria parasite genomes reveal features associated with lineage specific evolution in birds and mammals U Boehme, TD Otto, J Cotton, S Steinbiss, M Sanders, SO Oyola, A Nicot, ... BioRXiv, 086504	5	2018
A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel Limnoperna fortunei M Uliano-Silva, F Dondero, T Dan Otto, I Costa, NCB Lima, JA Americo, ... GigaScience	1	2017
Profiling invasive Plasmodium falciparum merozoites using an integrated omics approach K Kumar, P Srinivasan, MJ Nold, JK Moch, K Reiter, D Sturdevant, TD Otto, ... Scientific reports 7 (1), 17146		2017
PIGGYBAC MUTAGENESIS SCREENING OF THOUSANDS OF PLASMODIUM FALCIPARUM GENES REVEALS WHAT A MALARIA PARASITE CAN'T LIVE WITHO... M Zhang, C Wang, TD Otto, J Oberstaller, IF Bronner, S Li, K Udenze, ... AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE 95 (5), 390-391		2017
WHOLE GENOME SEQUENCING OF PLASMODIUM FALCIPARUM MALARIA PARASITES FROM DRIED BLOOD SPOTS: GATEWAY TO HIGH-RESOLUTION GEN... CV Ariani, WL Hamilton, S Oyola, LN Amenga-Etego, M Kekre, ... AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE 95 (5), 391-391		2017
Genomic characterization of recrudescent Plasmodium malariae after treatment with artemether/lumefantrine GG Rutledge, I Marr, GKL Huang, S Auburn, J Marfurt, M Sanders, ... Emerging infectious diseases 23 (8), 1300	1	2017
Human vaccination against Plasmodium vivax Duffy-binding protein induces strain-transcending antibodies RO Payne, SE Silk, SC Elias, KH Milne, TA Rawlinson, D Llewellyn, ... JCI insight 2 (12)	3	2017
A single nucleotide polymorphism in an AP2 transcription factor encoded in the malaria-causing Plasmodium berghei alters the development of host immunity PW Sheehan, M Akkaya, A Bansal, G Arora, TD Otto, CF Qi, M Pena, ... The Journal of Immunology 198 (1 Supplement), 77.5-77.5		2017
pfk13-independent treatment failure in four imported cases of Plasmodium falciparum malaria treated with artemether-lumefantrine in the United Kingdom CJ Sutherland, P Lansdell, M Sanders, J Muwanguzi, DA van Schalkwyk, ... Antimicrobial agents and chemotherapy 61 (3), e02382-16	11	2017
Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution GG Rutledge, U Böhme, M Sanders, AJ Reid, JA Cotton, ... Nature 542 (7639), 101	22	2017
SC83288 is a clinical development candidate for the treatment of severe malaria S Pegoraro, M Duffey, TD Otto, Y Wang, R Rösemann, R Baumgartner, ... Nature communications 8, 14193	4	2017
Correction: Variant Exported Blood-Stage Proteins Encoded by Plasmodium Multigene Families Are Expressed in Liver Stages Where They Are Exported into the Pa... A Fougère, AP Jackson, DP Bechtsi, JAM Braks, T Annoura, J Fonager, ... PLoS pathogens 13 (1), e1006128		2017
A SATURATION-LEVEL PIGGYBAC MUTAGENESIS SCREEN OF THE PLASMODIUM FALCIPARUM GENOME DEFINES GENES IMPORTANT FOR IN VITRO ASE... M Zhang, C Wang, J Oberstaller, TD Otto, S Adapa, X Liao, J Swanson, ... AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE 97 (5), 19-20		2017
A LARGE-SCALE GENETIC SCREEN OF PLASMODIUM FALCIPARUM IDENTIFIES GENOTYPY-PHENOTYPE MUTATIONS AFFECTING TOLERANCE TO FEBRIL... M Zhang, C Wang, P Thomas, J Oberstaller, TD Otto, X Liao, S Li, ... AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE 97 (5), 323-323		2017
ESSENTIAL ASPECTS OF RNA METABOLISM FOR P. FALCIPARUM BLOOD-STAGE SURVIVAL J Oberstaller, M Zhang, CQ Wang, TD Otto, X Liao, J Swanson, SR Adapa, ... AMERICAN JOURNAL OF TROPICAL MEDICINE AND HYGIENE 97 (5), 625-625		2017
Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria HJ Lee, M Walther, A Georgiadou, D Nwakanma, LB Stewart, M Levin, ... bioRxiv, 193631		2017
Plasmodium vivax-like genome sequences shed new insights into Plasmodium vivax biology and evolution A Gilabert, T Otto, G Rutledge, B Franzon, B Ollomo, C Arnathau, ... bioRxiv, 205302		2017
An improved Plasmodium cynomolgi genome assembly reveals an unexpected methyltransferase gene expansion EM Pasini, U Böhme, GG Rutledge, A Voorberg-Van der Wel, M Sanders, ... Wellcome open research 2	3	2017

Don't splash water on my laptop
Now, you're exaggerating
BioinFormaticians: They can work From anywhere
www.biocomicals.com

How are the analysis coming?
Almost ready
www.biocomicals.com

What do we actually do everyday?

- You have got new data!
  - (1) **Need to understand, QC, and analyse the data**. How?

- Once the data has been **explored**, you need to compare against published ones
  - (2) You need to survey, and download the right dataset
  - Move to step (1)

- (3) Then you need to **visualise**
- Does it answer your question? There are times when you need to
  (4) develop new/better algorithms and
  (5) generate more data

Finally to present to an audience

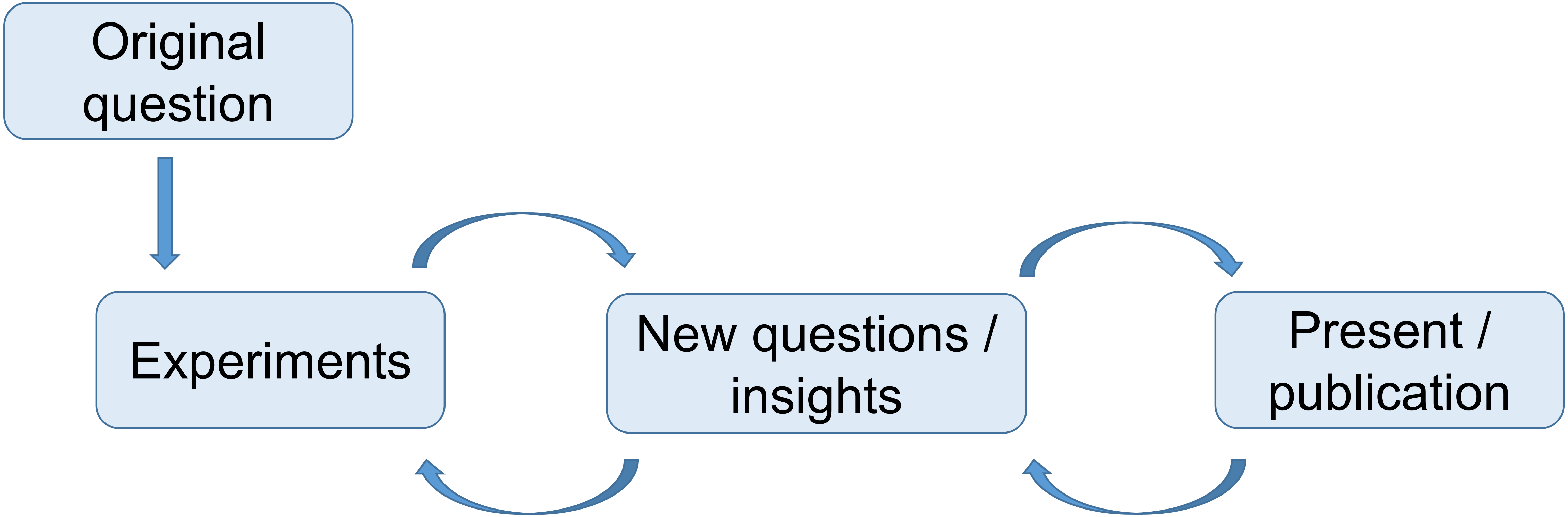
- Remember to save all your work first!
  - Organisation and record-keeping

- Publications? But before that…

- Are the data shared to the public?
  - How?

- Are the results reproducible?
  - How?

Experimental Analysis



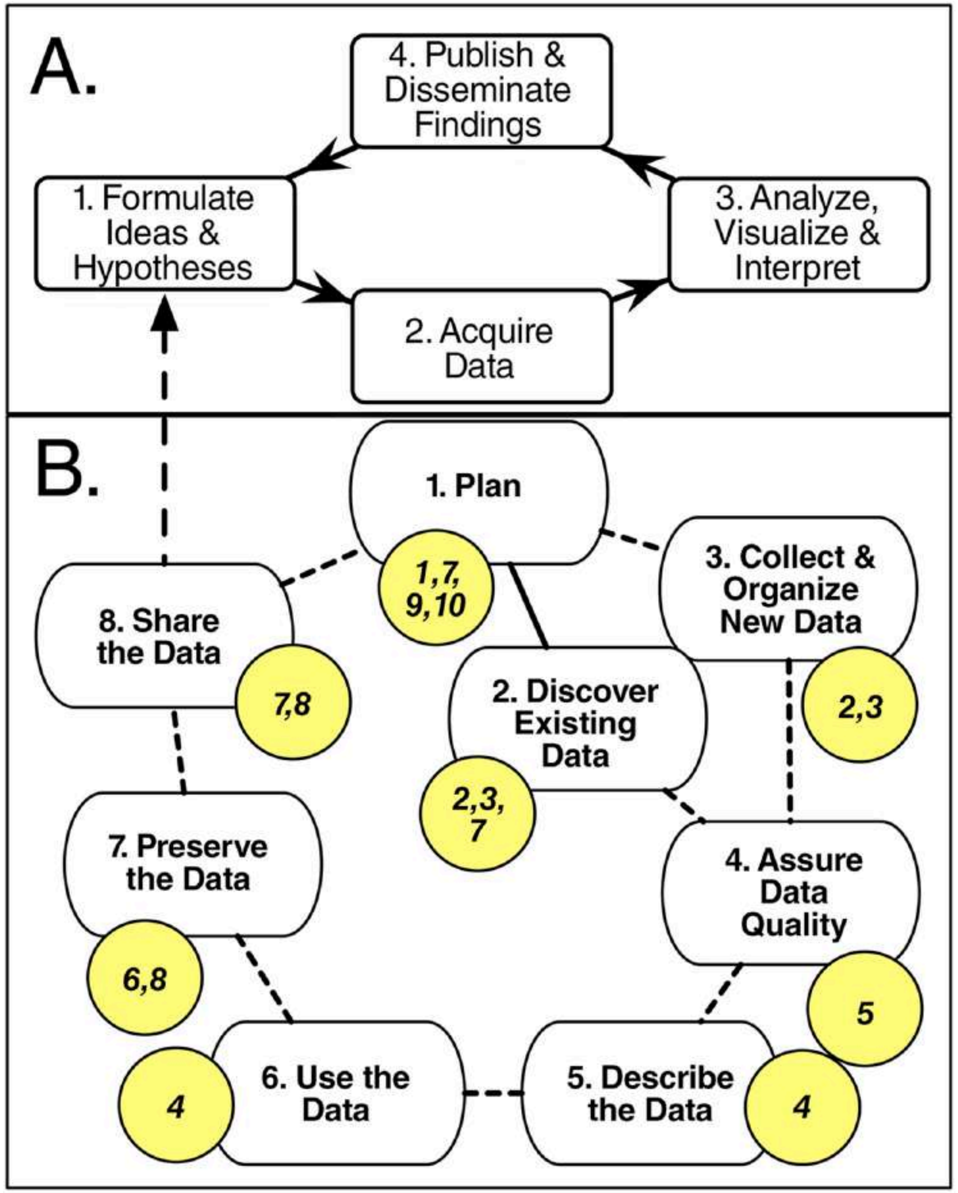
Original question

Experiments

New questions / insights

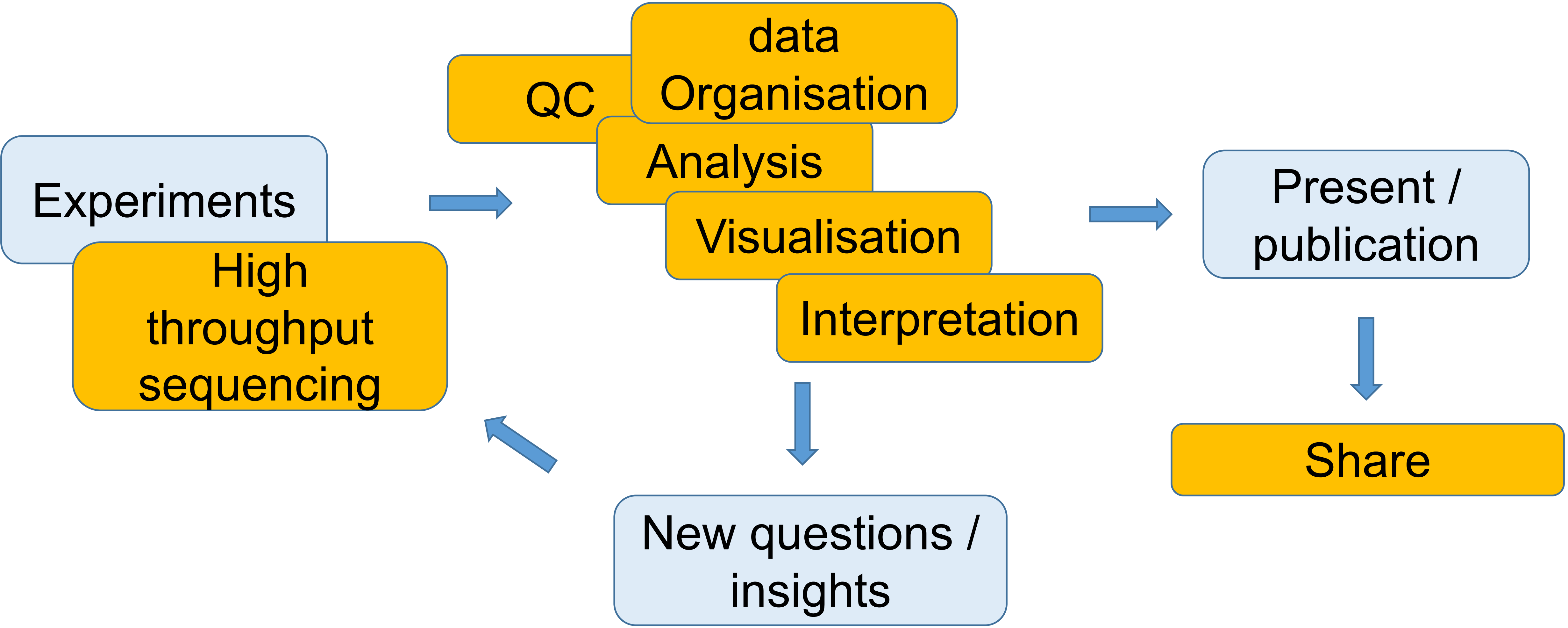
Present / publication

Relationship of the research life cycle (A) to the data life cycle (B)


A.
4. Publish & Disseminate Findings
1. Formulate Ideas & Hypotheses
3. Analyze, Visualize & Interpret
2. Acquire Data
B.
1. Plan
1,7, 9,10
3. Collect & Organize New Data
2,3
8. Share the Data
7,8
2. Discover Existing Data
2,3, 7
7. Preserve the Data
6,8
4. Assure Data Quality
5
6. Use the Data
4
5. Describe the Data
4

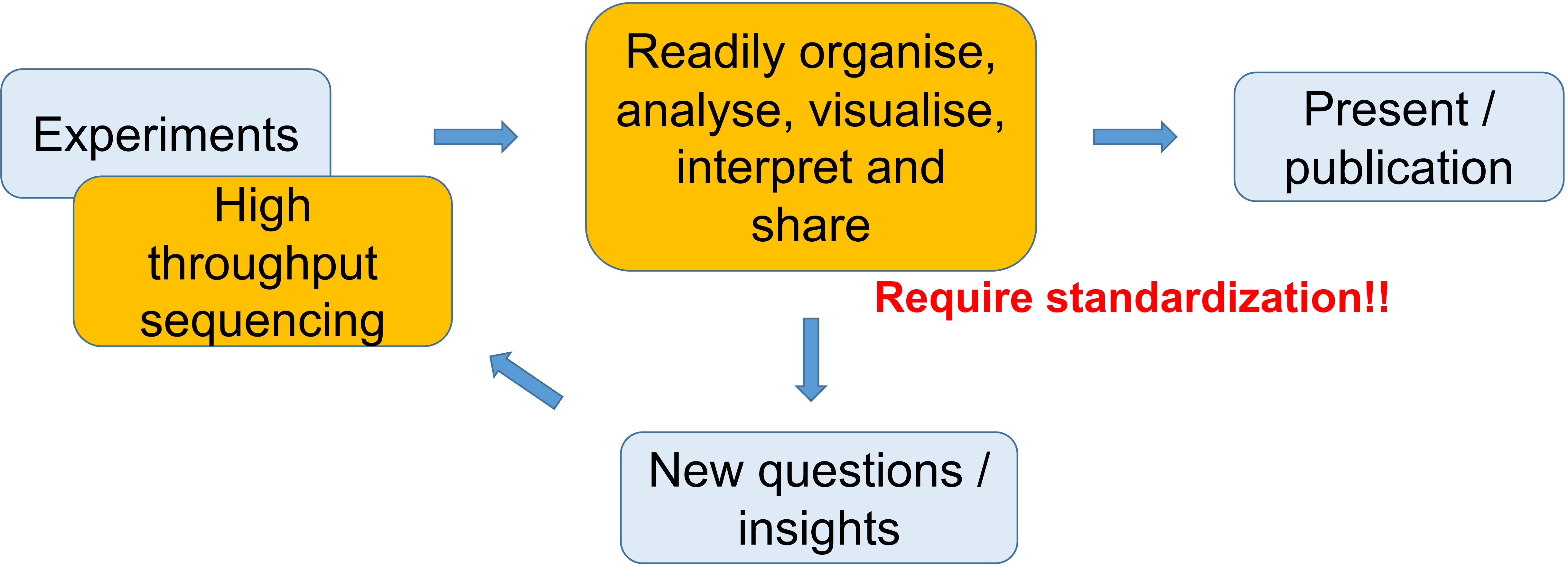
- https://doi.org/10.1371/journal.pcbi.1004525

Analysis in a high throughput world: challenges



x10-30

Analysis in a high throughput world: reorganisation


Experiments

High throughput sequencing

Readily organise, analyse, visualise, interpret and share

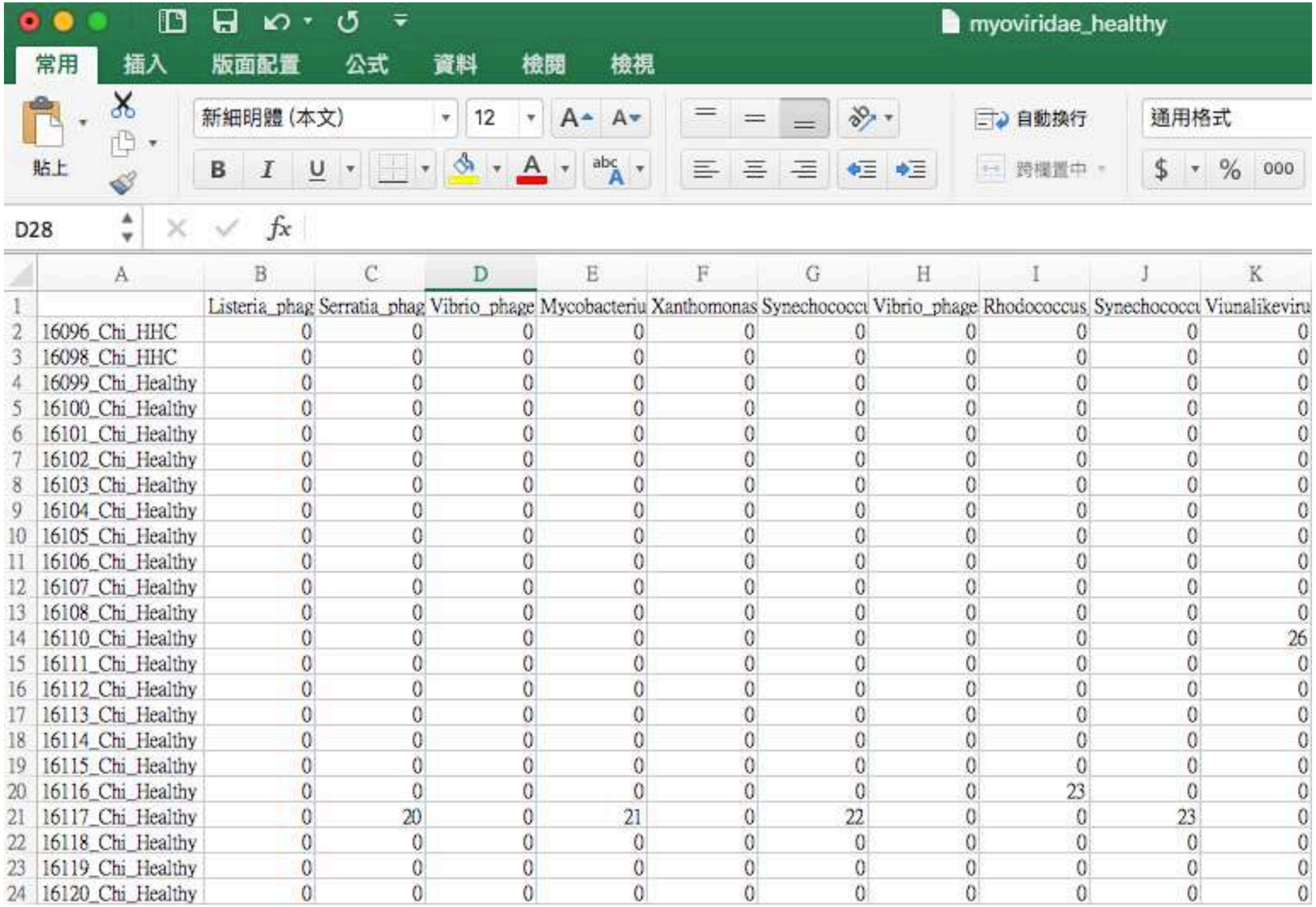
Present / publication

New questions / insights

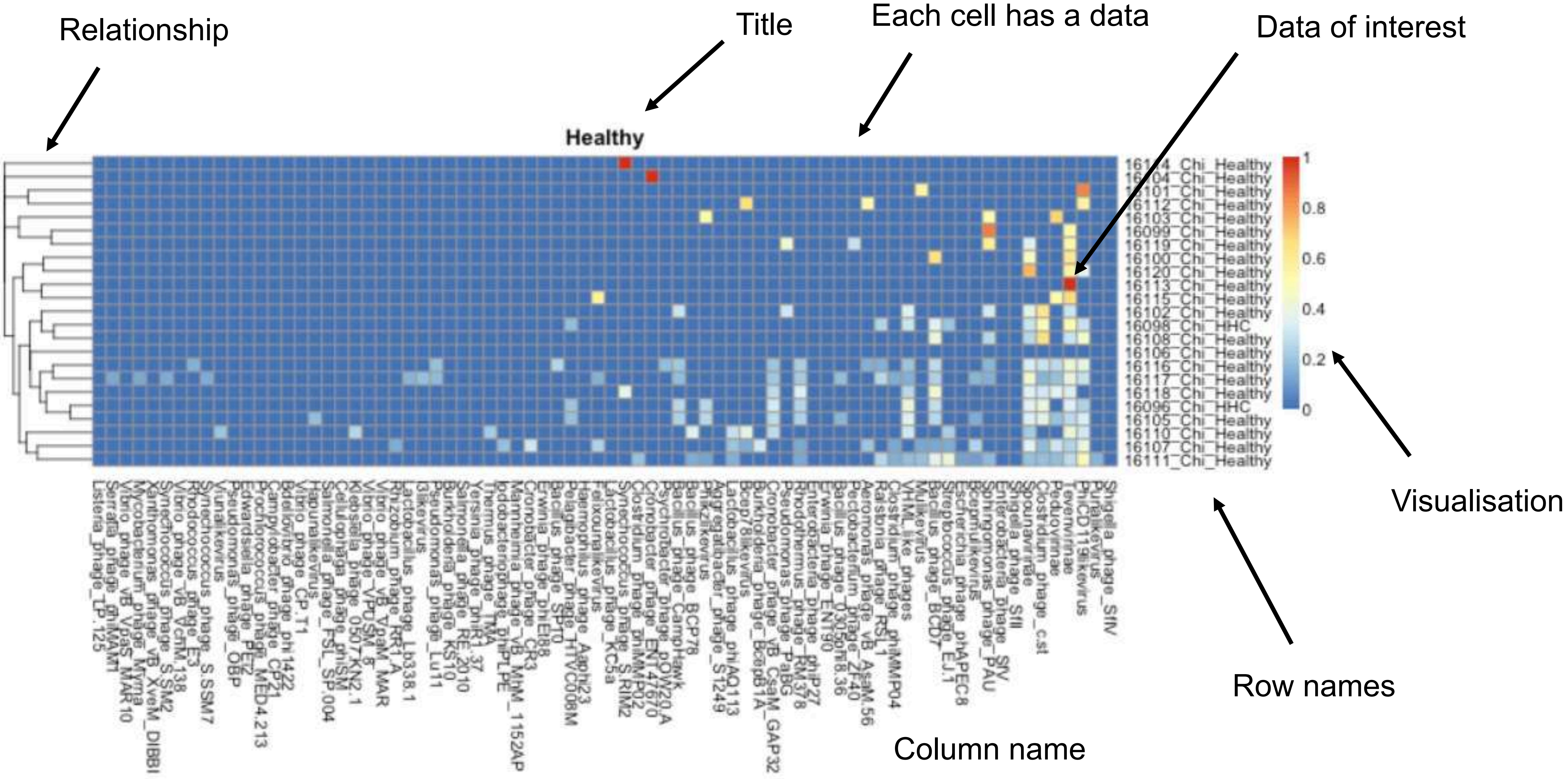
Require standardization!!

x10-30

When most people think of analysis



	A	B	C	D	E	F	G	H	I	J	K
1		Listeria_phag	Serratia_phag	Vibrio_phage	Mycobacteriu	Xanthomonas	Synechococcu	Vibrio_phage	Rhodococcus	Synechococcu	Viunalikeviru
2	16096_Chi_HHC	0	0	0	0	0	0	0	0	0	0
3	16098_Chi_HHC	0	0	0	0	0	0	0	0	0	0
4	16099_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
5	16100_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
6	16101_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
7	16102_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
8	16103_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
9	16104_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
10	16105_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
11	16106_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
12	16107_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
13	16108_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
14	16110_Chi_Healthy	0	0	0	0	0	0	0	0	0	26
15	16111_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
16	16112_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
17	16113_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
18	16114_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
19	16115_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
20	16116_Chi_Healthy	0	0	0	0	0	0	0	23	0	0
21	16117_Chi_Healthy	0	20	0	21	0	22	0	0	23	0
22	16118_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
23	16119_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
24	16120_Chi_Healthy	0	0	0	0	0	0	0	0	0	0



Relationship

Title

Each cell has a data

Data of interest

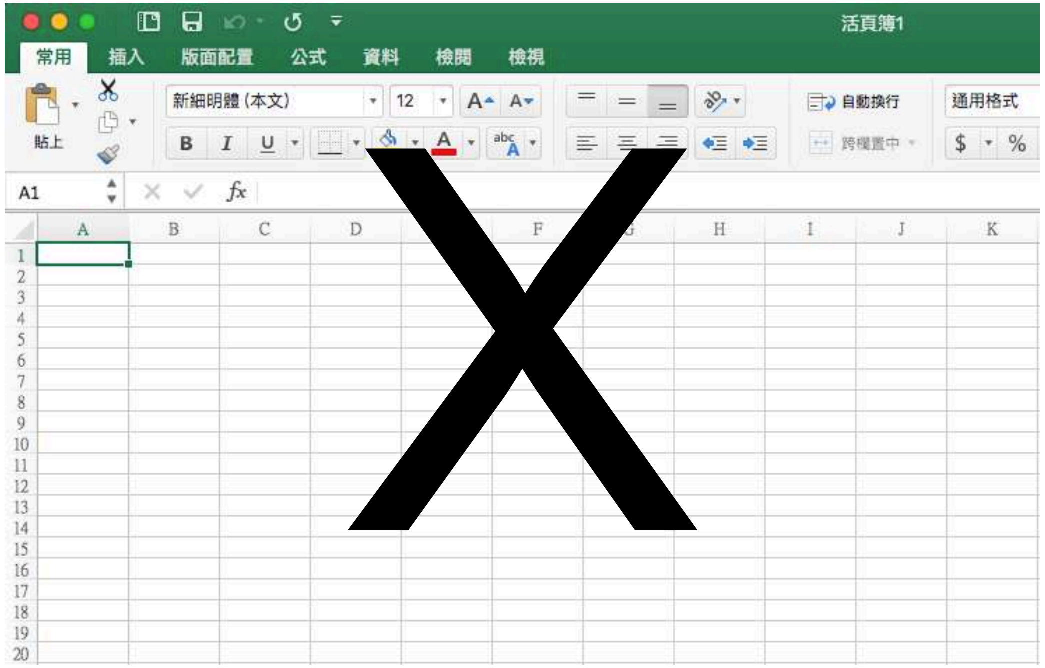
Visualisation

Row names

Column name

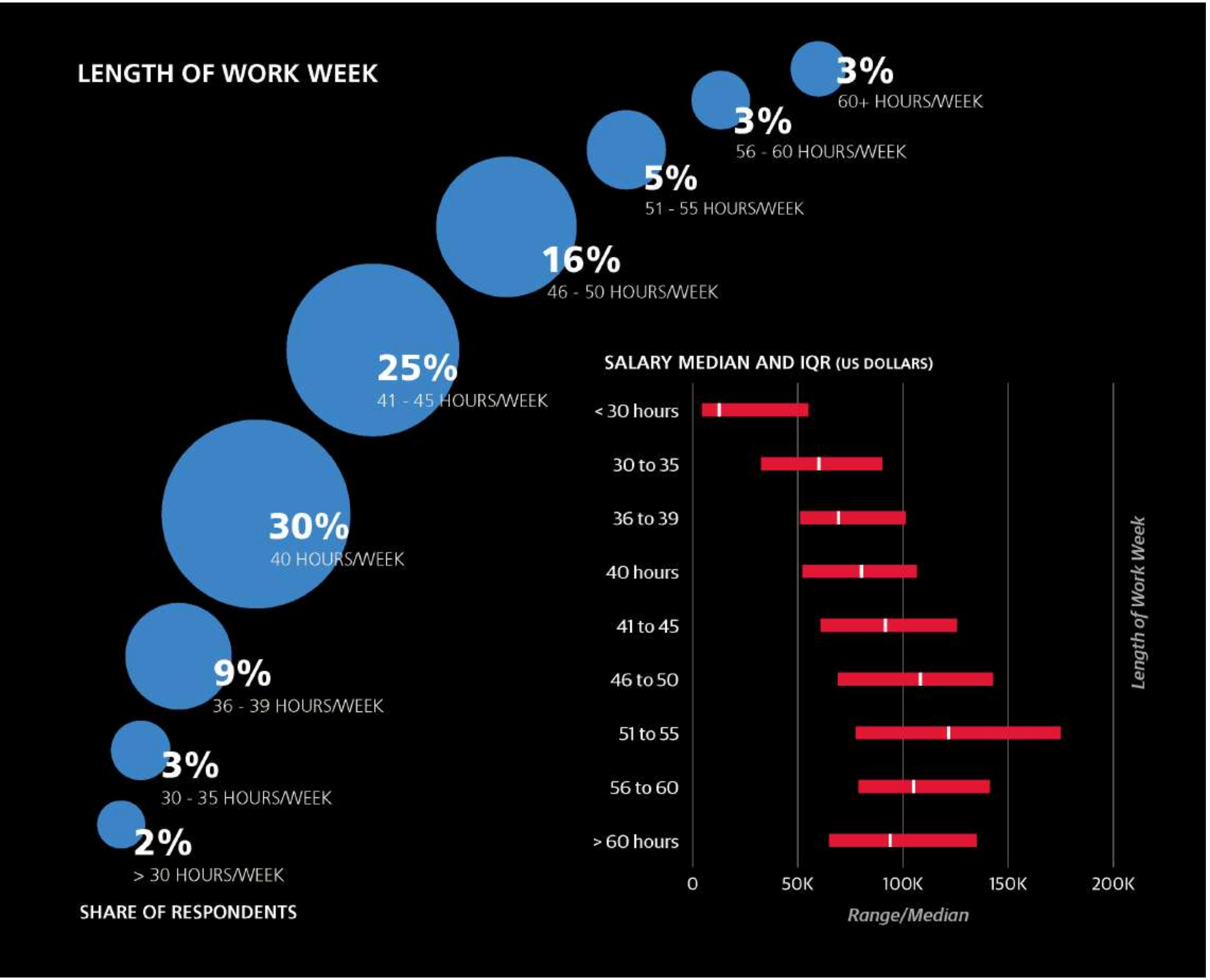
So what do you need?

You need a platform to rearrange, tidy, subset, merge data easily



Recommendation:
R and Python in a linux environment

2016 Data science survey



https://www.oreilly.com/ideas/2016-data-science-salary-survey-results

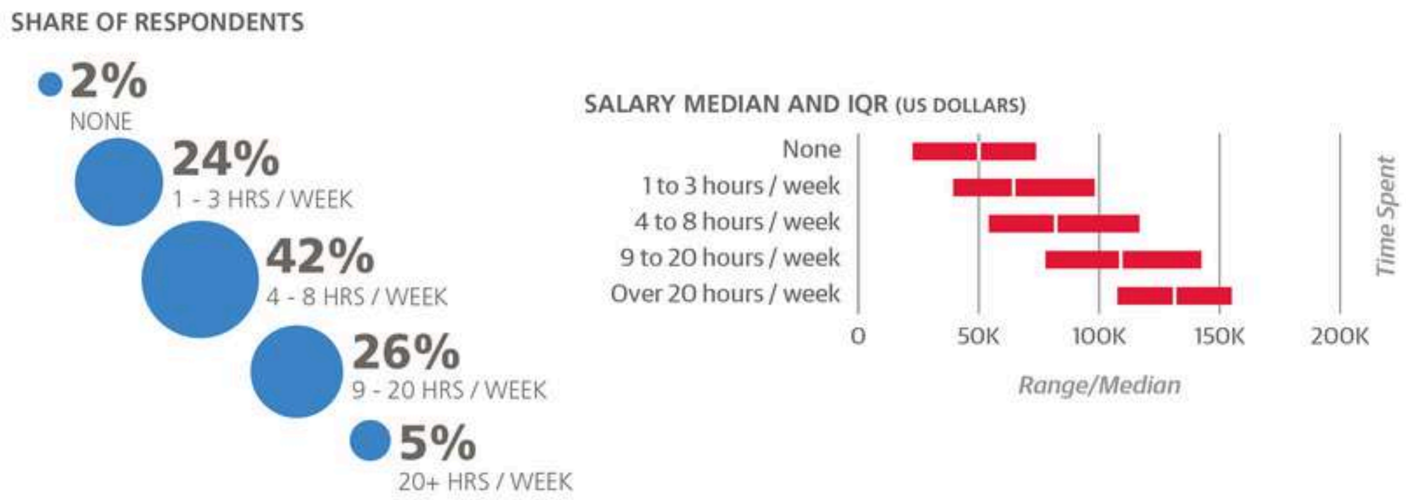
In Taiwan

40 hour = Monday - Friday
9am-6pm
one hour lunch break

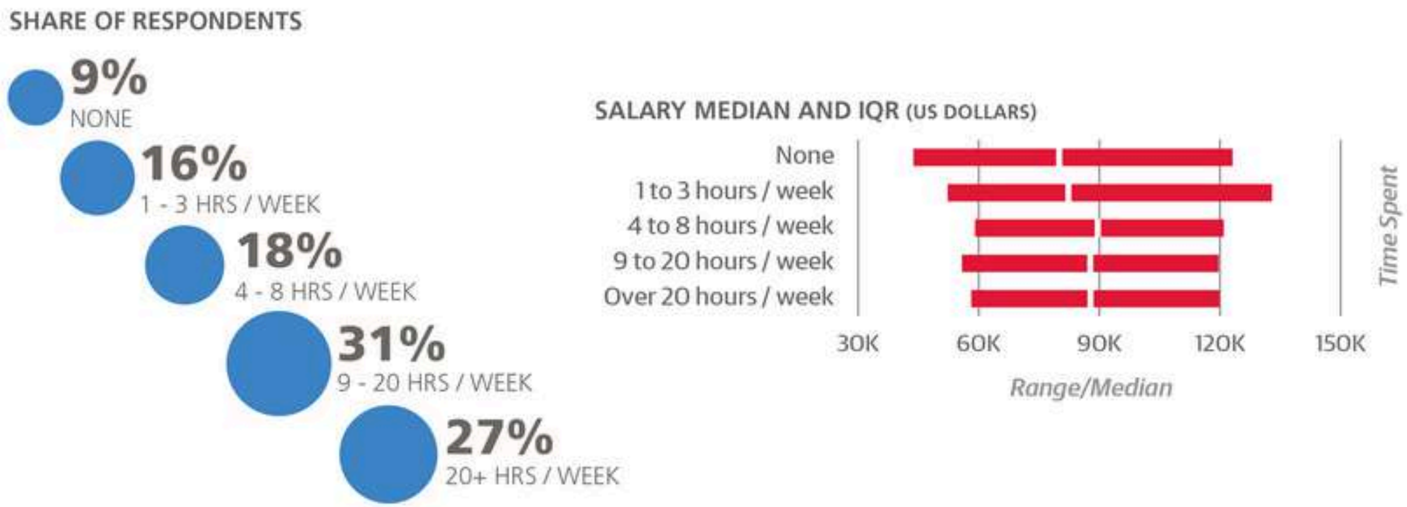
How much do you work a week?

Time spent in meetings and coding

TIME SPENT IN MEETINGS (hours per week)

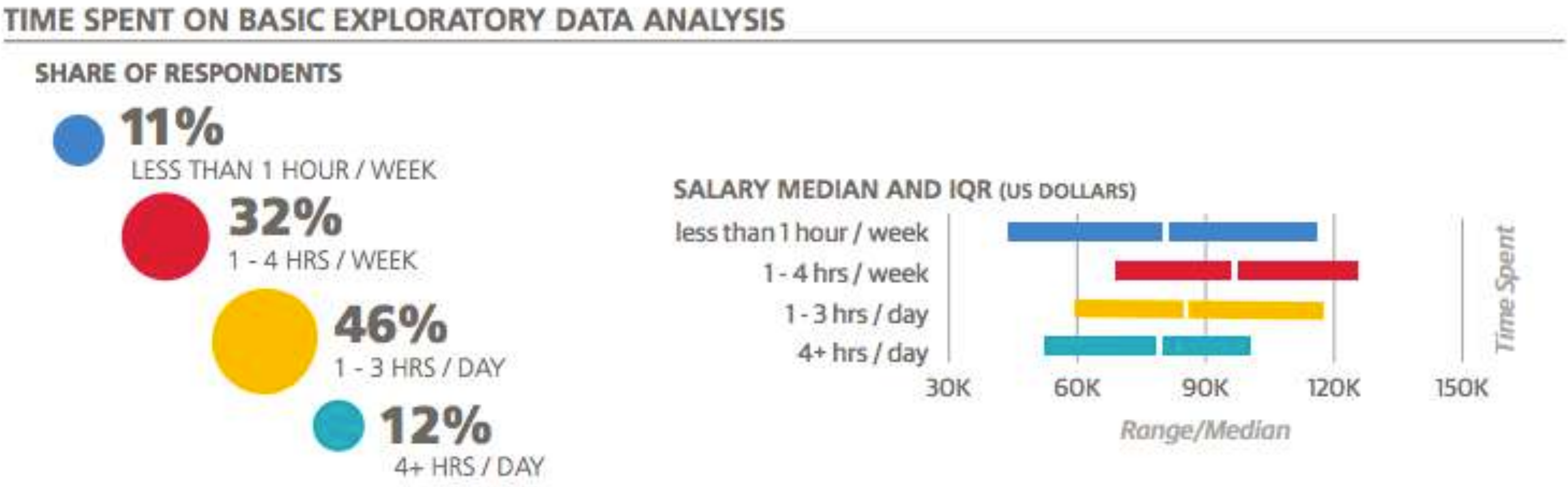


TIME SPENT CODING (hours per week)



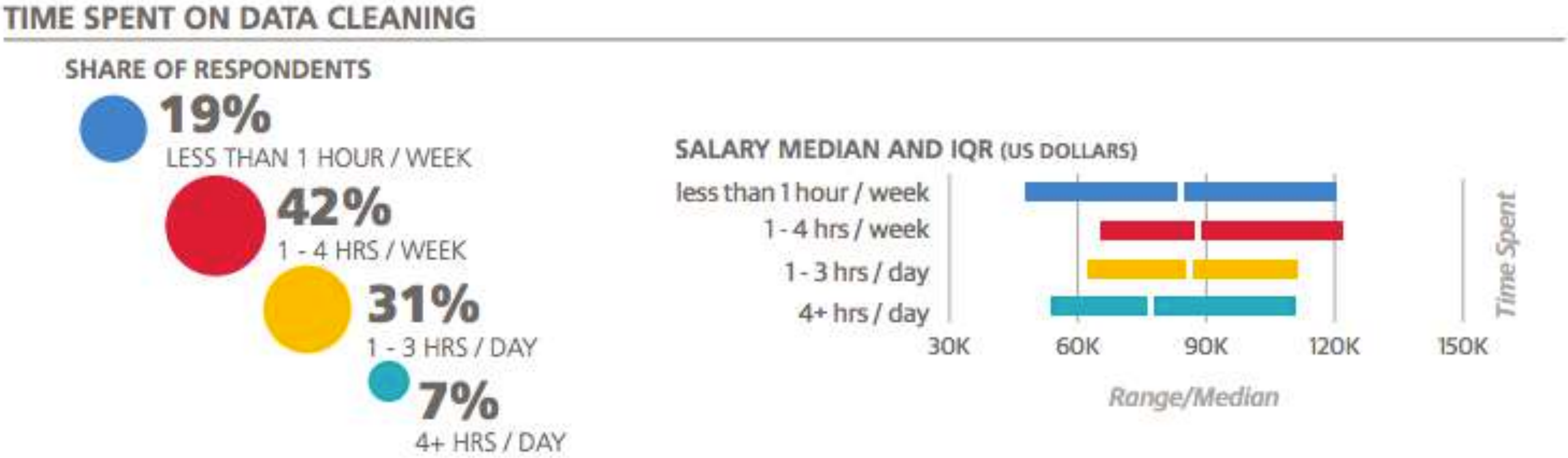
https://www.oreilly.com/ideas/2016-data-science-salary-survey-results

2015 Data science survey



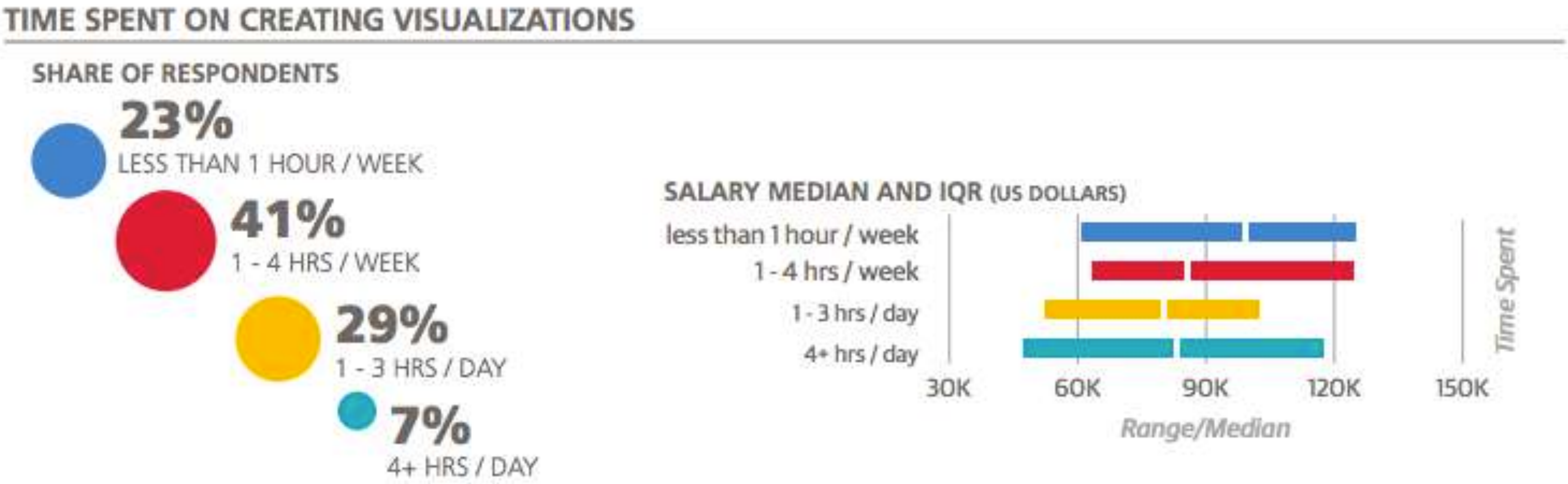
https://www.oreilly.com/ideas/2015-data-science-salary-survey

2015 Data science survey



https://www.oreilly.com/ideas/2015-data-science-salary-survey

2015 Data science survey



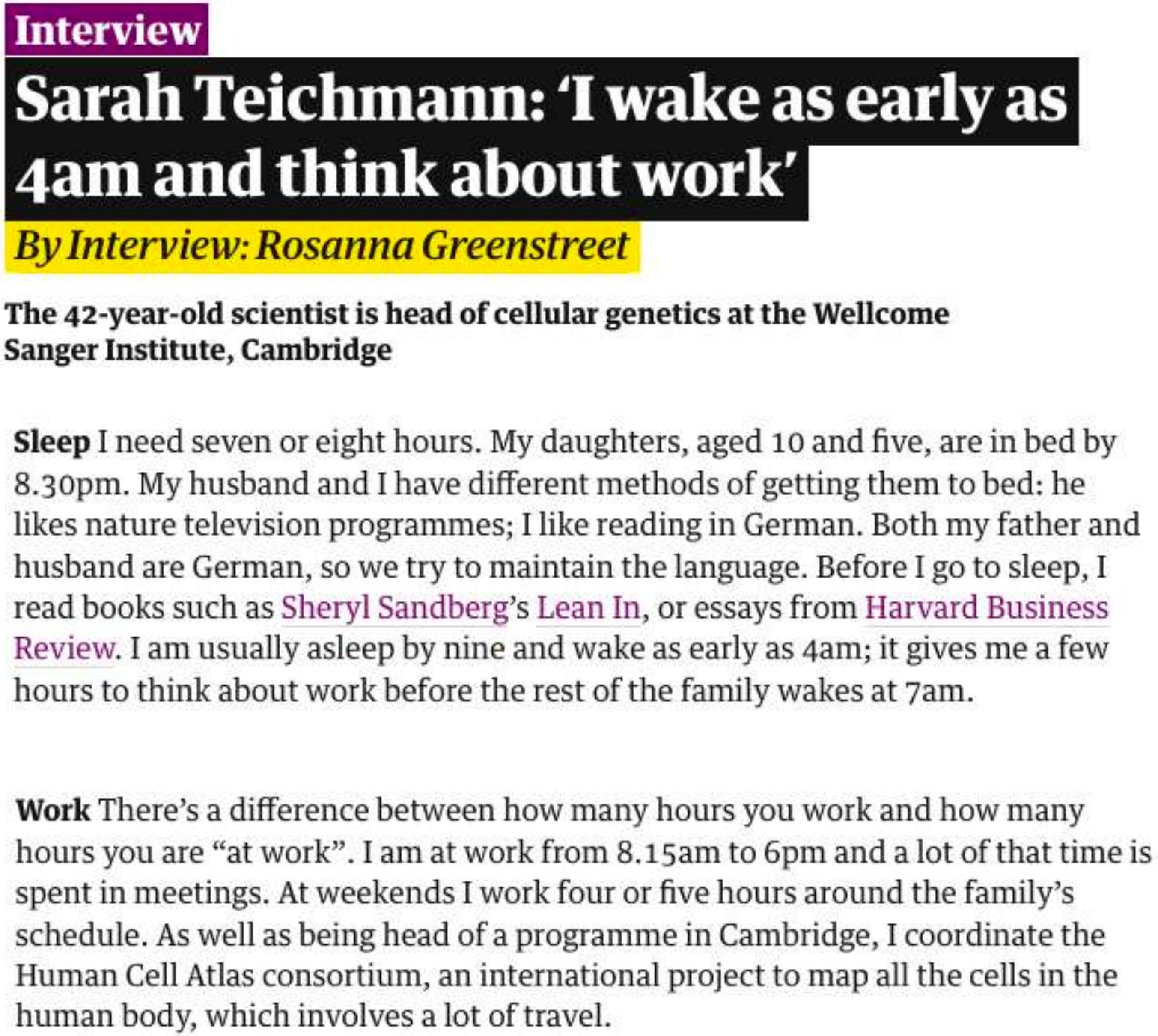
https://www.oreilly.com/ideas/2015-data-science-salary-survey

Some observations

A day of a data scientist /bioinformatician / biologist with lots of data:

- **Less than 1 to 4 hours** to quickly explore data (78%)
- **Less than 1 to 4 hours** to do data cleaning (74%)
- **Less than 1 to 4 hours** to visualise data (70%)
- **Less than 1 to 4+ hours** to present analysis (73%)

= 4 – 16 hours to finish your daily task



Interview

Sarah Teichmann: 'I wake as early as 4am and think about work'

By Interview: Rosanna Greenstreet

The 42-year-old scientist is head of cellular genetics at the Wellcome Sanger Institute, Cambridge

Sleep I need seven or eight hours. My daughters, aged 10 and five, are in bed by 8.30pm. My husband and I have different methods of getting them to bed: he likes nature television programmes; I like reading in German. Both my father and husband are German, so we try to maintain the language. Before I go to sleep, I read books such as Sheryl Sandberg's Lean In, or essays from Harvard Business Review. I am usually asleep by nine and wake as early as 4am; it gives me a few hours to think about work before the rest of the family wakes at 7am.

Work There's a difference between how many hours you work and how many hours you are "at work". I am at work from 8.15am to 6pm and a lot of that time is spent in meetings. At weekends I work four or five hours around the family's schedule. As well as being head of a programme in Cambridge, I coordinate the Human Cell Atlas consortium, an international project to map all the cells in the human body, which involves a lot of travel.

https://www.theguardian.com/money/2018/mar/03/sarah-teichmann-wake-4am-think-about-work

Some observations (my own opinions)

- Data scientist are needed everywhere
- Bioinformatician / data scientist in Biology field are less well-paid in relative to other field,

This will result in

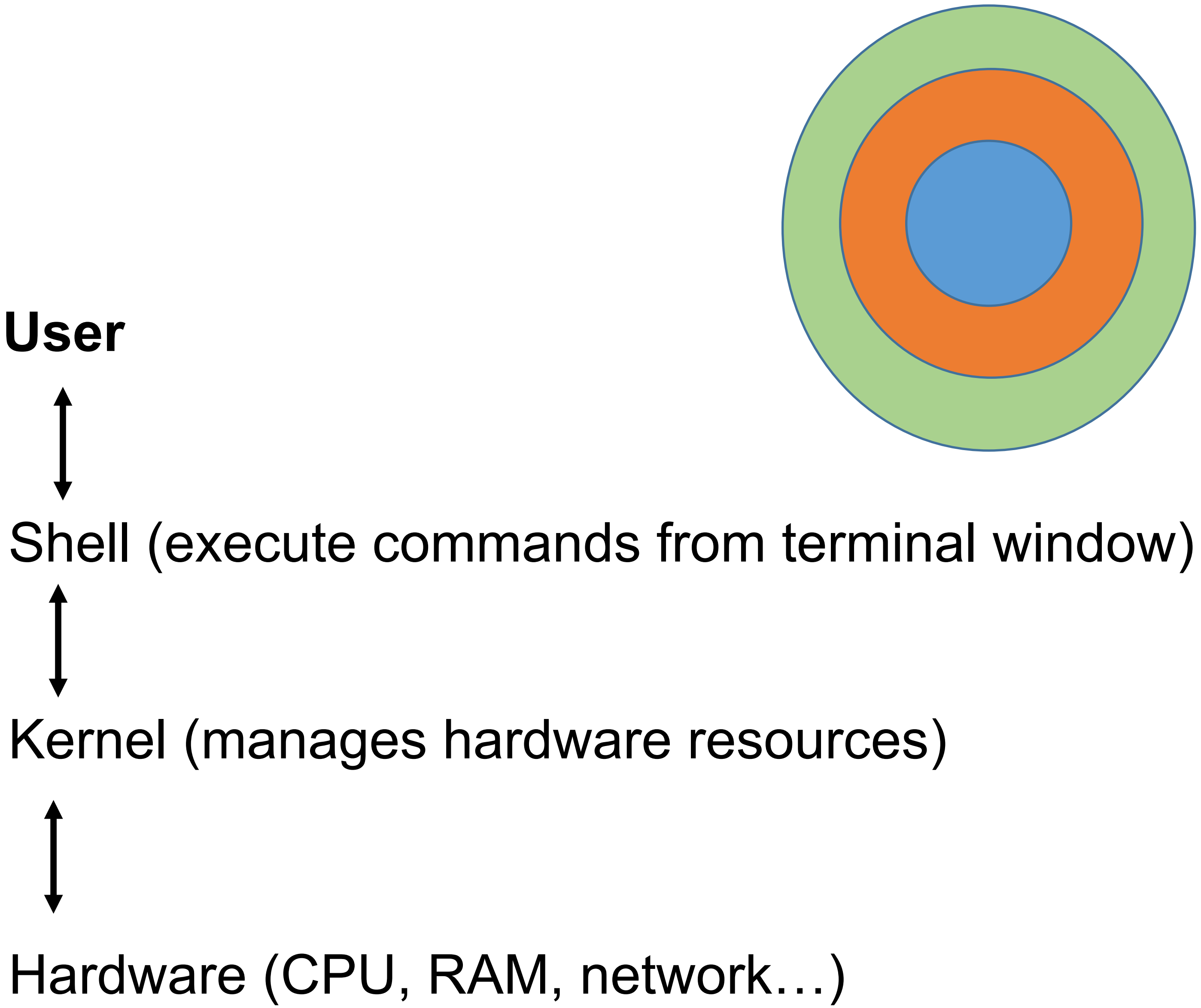
- All high throughput data / analysis are outsourced to companies -> students/labs will not gain the experience
- A few labs can enjoy deal with all the data in Taiwan -> also not good as no energy to initiate novel projects

- Try to be as much hands on as possible early in your training

Linux

History of Unix

Unix is a family of multitasking, multiuser computer operating systems that derive from the original AT&T Unix, development starting in the 1970s at the Bell Labs research center by Ken Thompson, Dennis Ritchie, and others.



User

Shell (execute commands from terminal window)

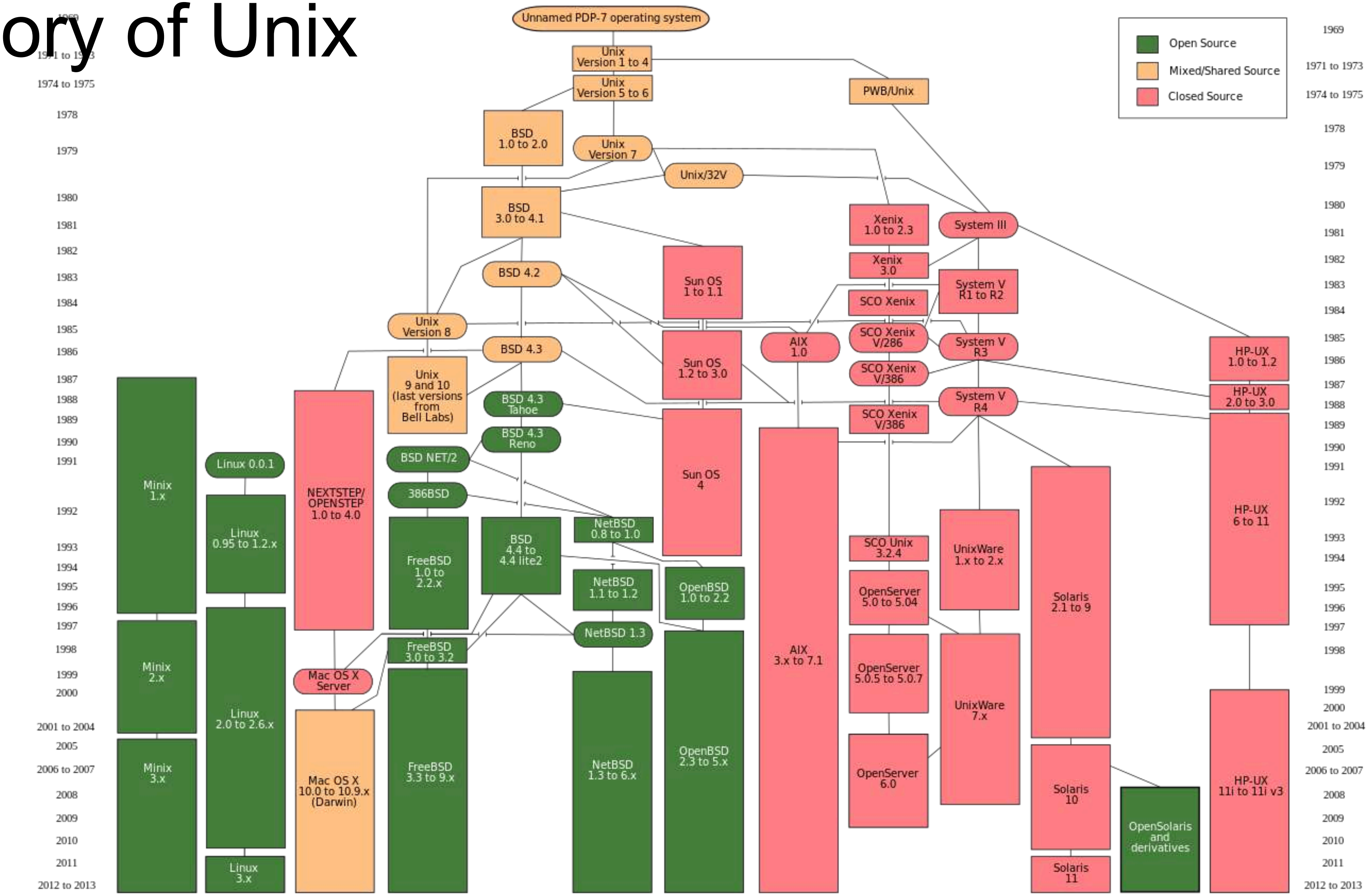
Kernel (manages hardware resources)

Hardware (CPU, RAM, network…)

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

Wiki

History of Unix

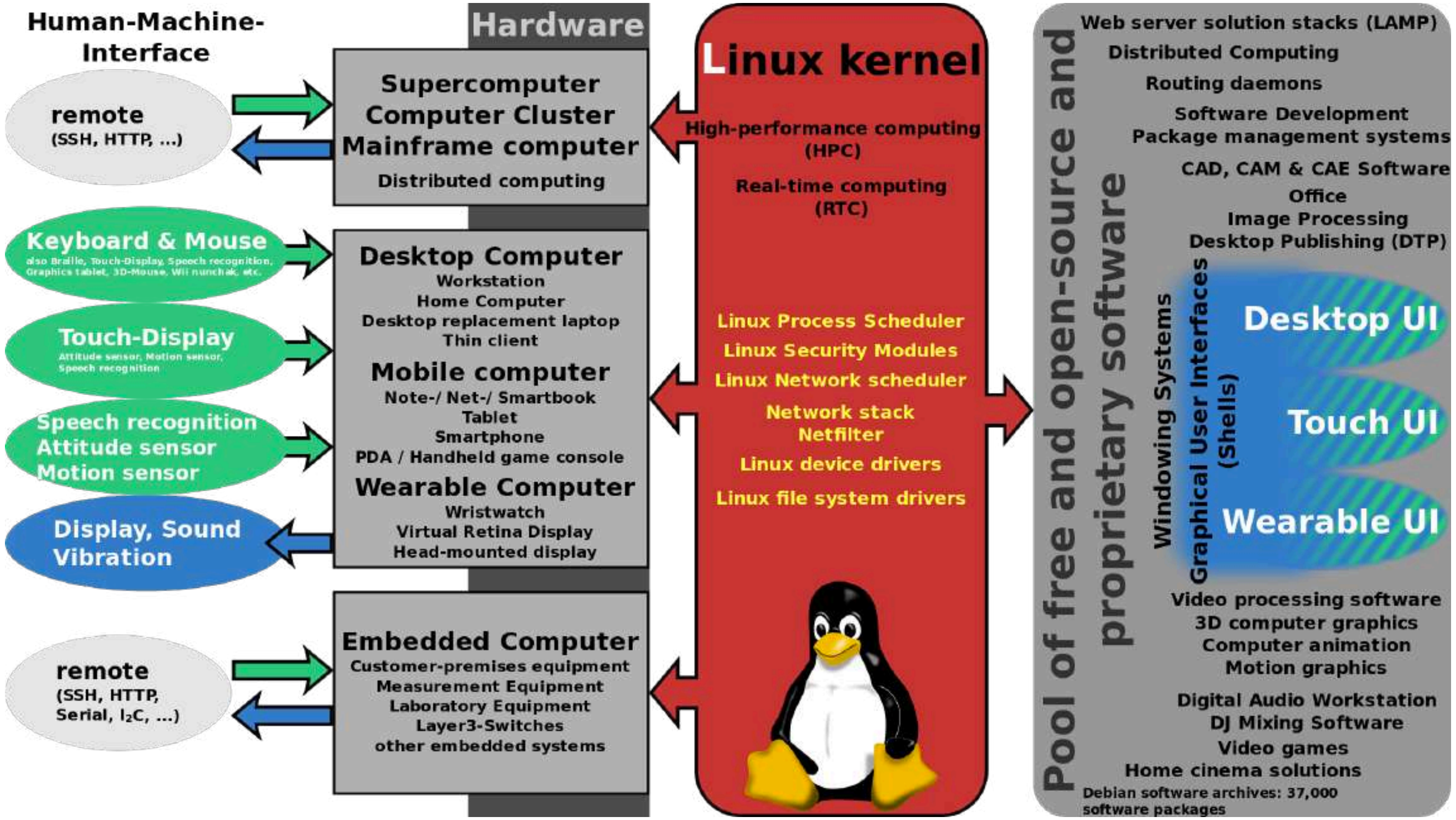


https://en.wikipedia.org/wiki/History_of_Unix

Wiki

What is GNU/Linux?

GNU/Linux is a **Unix-like** computer **operating system (OS)** assembled under the model of **free** and **open-source software** development and distribution.



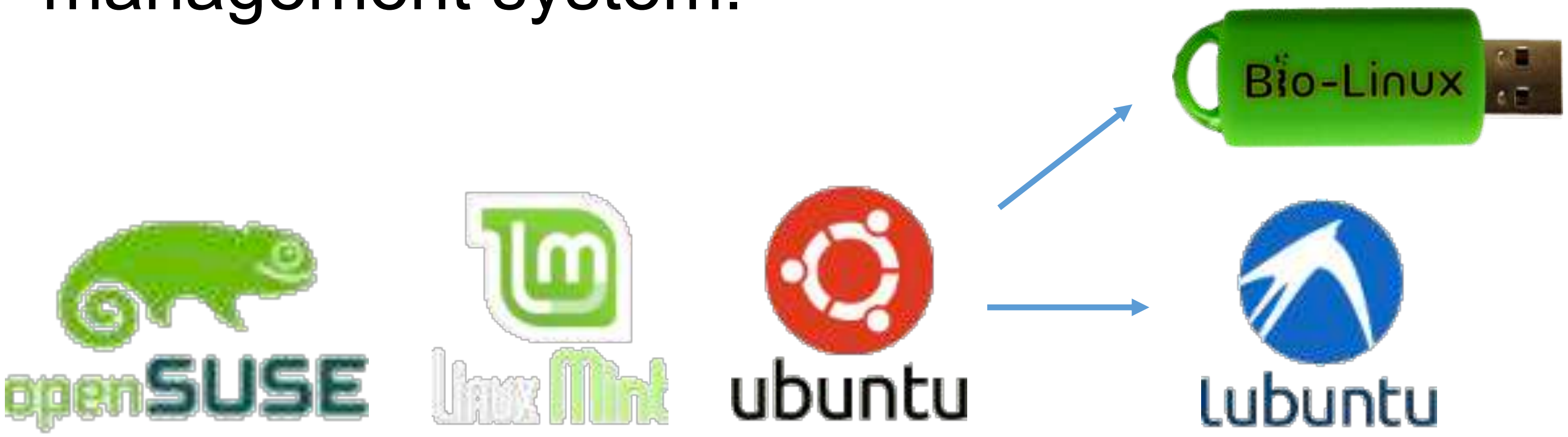
Linux kernel was designed by Linux Torvalds

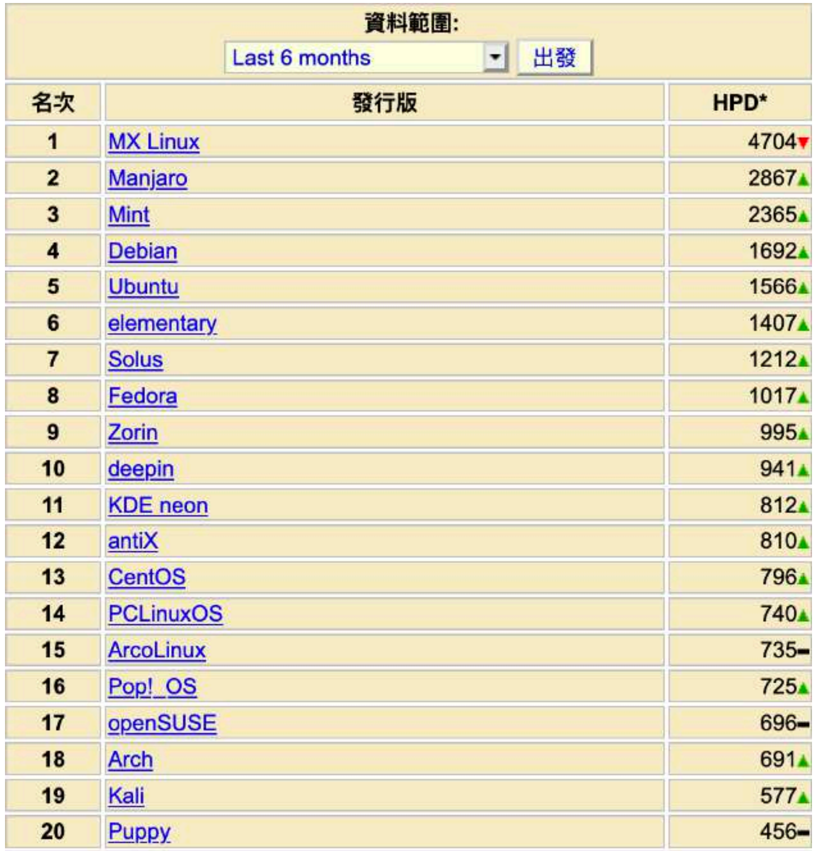
GNU project contains lots of UNIX-like libraries and applications

Wiki

Linux distributions

A **Linux distribution** (often called a distro for short) is an operating system made from **a software collection**, which is based upon the Linux kernel and, often, a package management system.





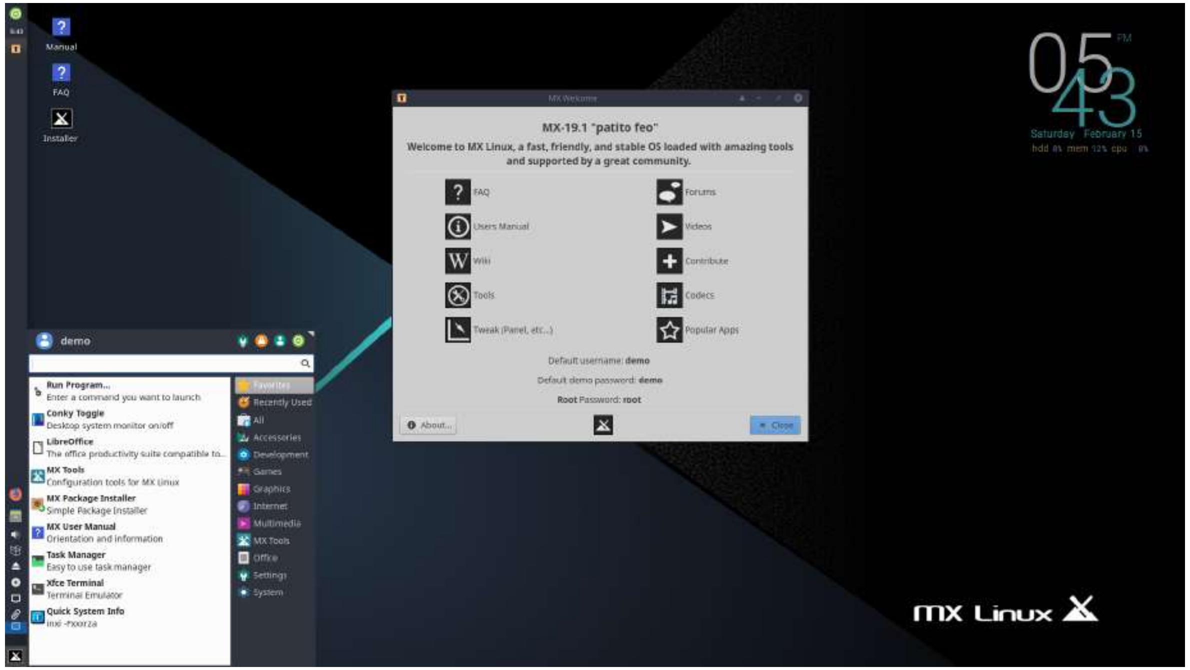
資料範圍:

Last 6 months ▾ 出發

名次	發行版	HPD*
1	MX Linux	4704▼
2	Manjaro	2867▲
3	Mint	2365▲
4	Debian	1692▲
5	Ubuntu	1566▲
6	elementary	1407▲
7	Solus	1212▲
8	Fedora	1017▲
9	Zorin	995▲
10	deepin	941▲
11	KDE neon	812▲
12	antiX	810▲
13	CentOS	796▲
14	PCLinuxOS	740▲
15	ArcoLinux	735–
16	Pop!_OS	725▲
17	openSUSE	696–
18	Arch	691▲
19	Kali	577▲
20	Puppy	456–

Wiki

Linux distributions



MX-19.1 "patito feo"

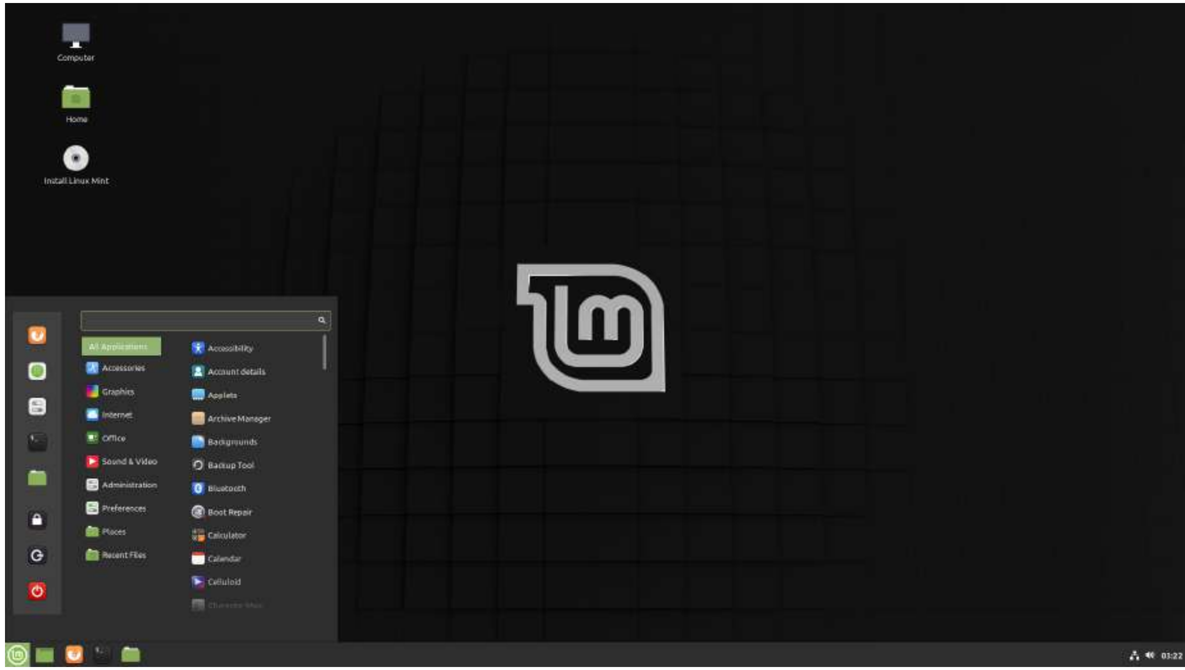
Welcome to MX Linux, a fast, friendly, and stable OS loaded with amazing tools and supported by a great community.

Default username: **demo**

Default demo password: **demo**

Root Password: **root**

Mx Linux



Mint

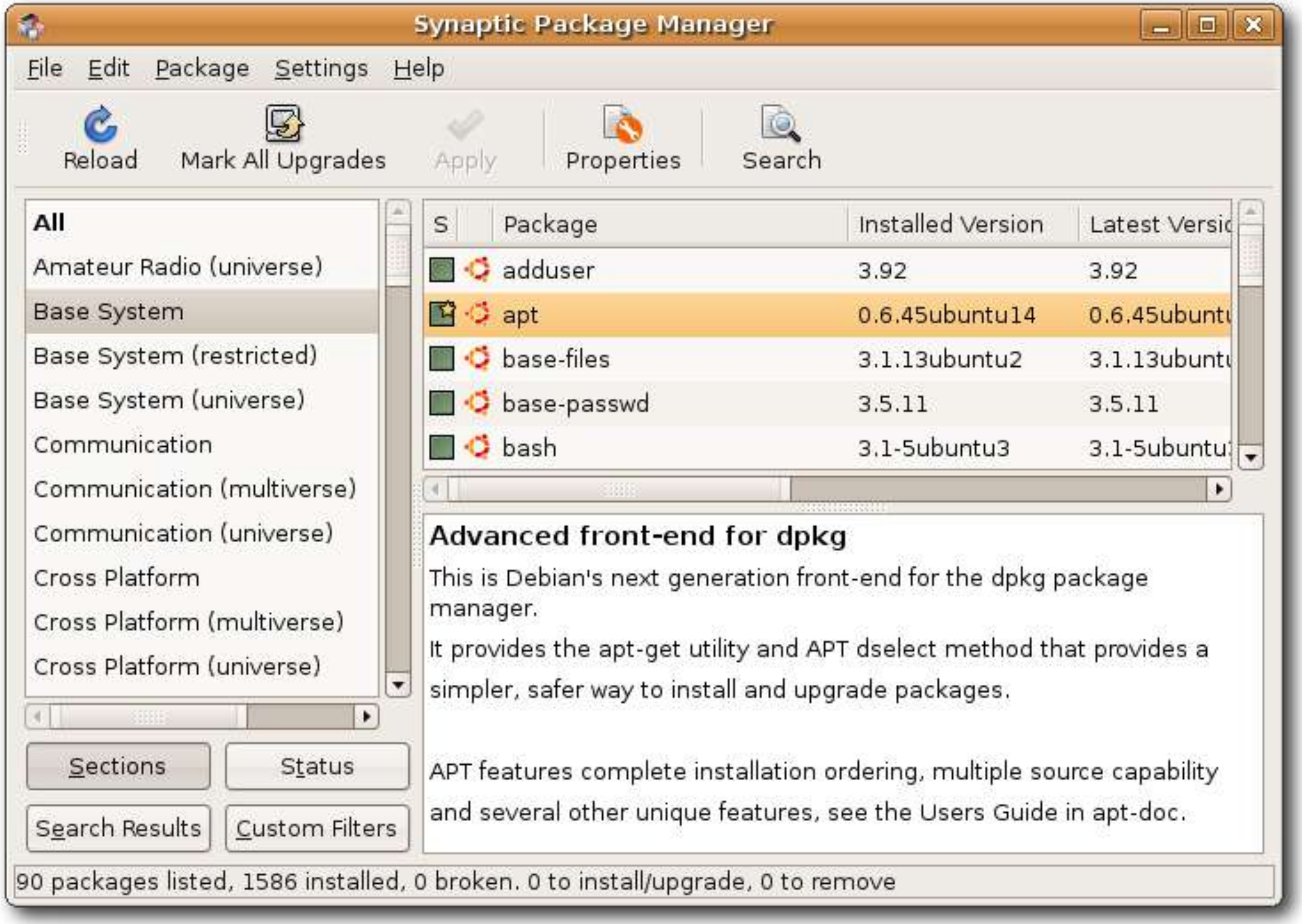
Installing programs in Linux

- Open-source and free
- More than one way of installing
  - From downloaded files
    - Binaries (already executable)
    - Compile from source files
  - From package manager (like App store)
    - Contains official repositories (secure, stable malware-free)
    - New repositories can be added (latest)

- Dependencies
  - A software uses (depends) another software which performs specific tasks

Installing programs in Linux

Desktop type
Software manager



- All
- Amateur Radio (universe)
- Base System
- Base System (restricted)
- Base System (universe)
- Communication
- Communication (multiverse)
- Communication (universe)
- Cross Platform
- Cross Platform (multiverse)
- Cross Platform (universe)

S	Package	Installed Version	Latest Versic
■	adduser	3.92	3.92
■	apt	0.6.45ubuntu14	0.6.45ubunt
■	base-files	3.1.13ubuntu2	3.1.13ubunt
■	base-passwd	3.5.11	3.5.11
■	bash	3.1-5ubuntu3	3.1-5ubuntu

Advanced front-end for dpkg

This is Debian's next generation front-end for the dpkg package manager.

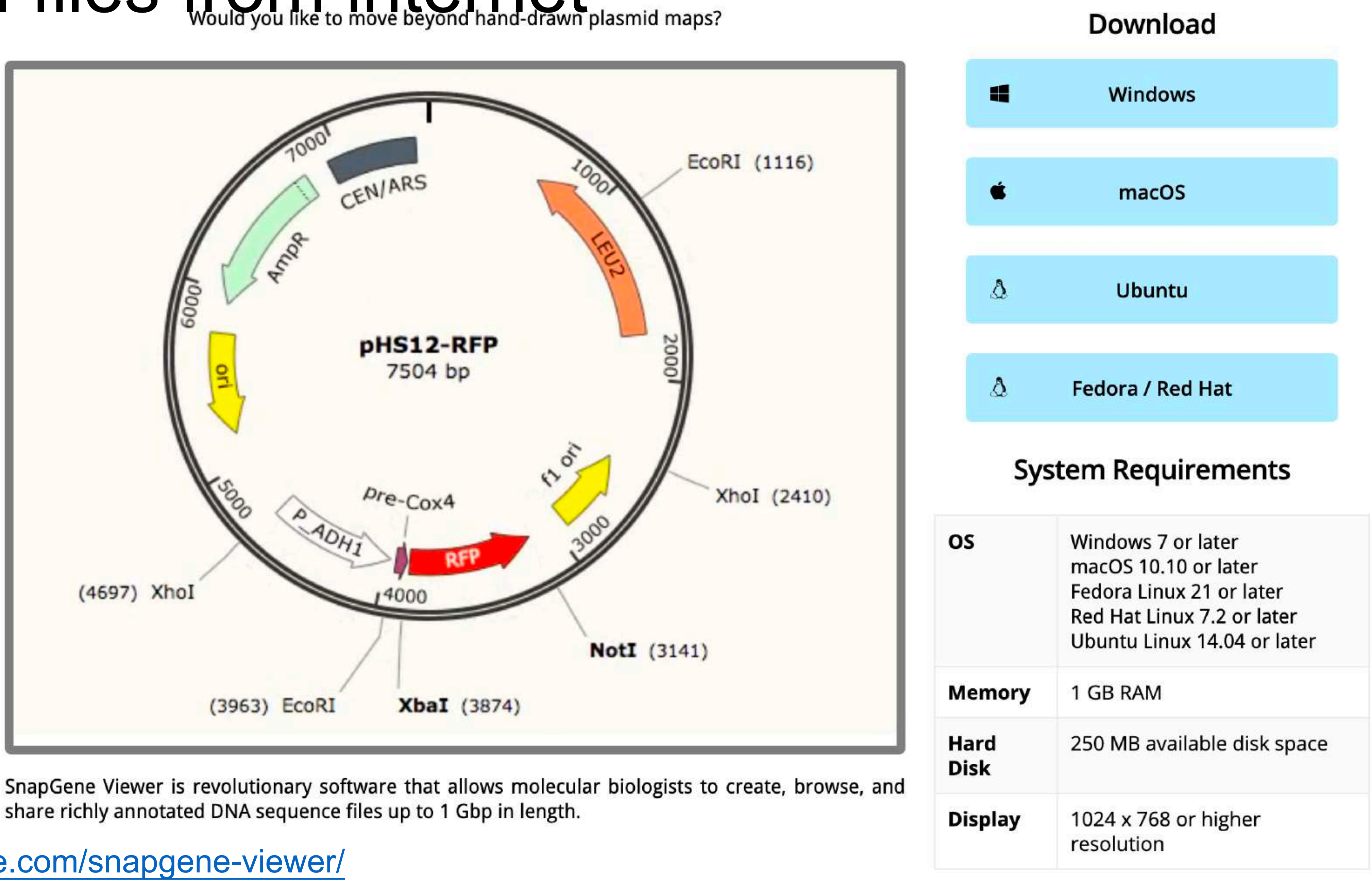
It provides the apt-get utility and APT dselect method that provides a simpler, safer way to install and upgrade packages.

APT features complete installation ordering, multiple source capability and several other unique features, see the Users Guide in apt-doc.

90 packages listed, 1586 installed, 0 broken. 0 to install/upgrade, 0 to remove

https://help.ubuntu.com/community/SynapticHowto

Download files from internet



Would you like to move beyond hand-drawn plasmid maps?

SnapGene Viewer is revolutionary software that allows molecular biologists to create, browse, and share richly annotated DNA sequence files up to 1 Gbp in length.

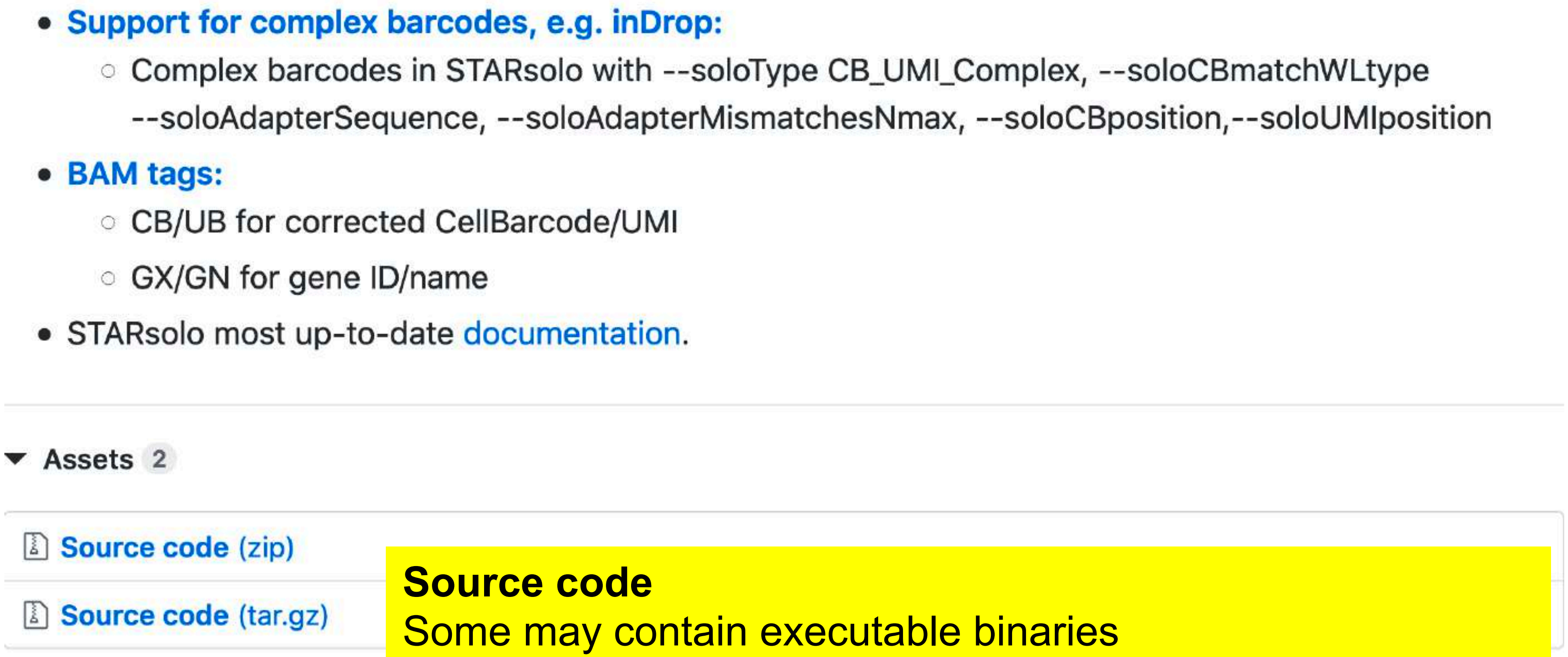
Download

System Requirements

OS	Windows 7 or later macOS 10.10 or later Fedora Linux 21 or later Red Hat Linux 7.2 or later Ubuntu Linux 14.04 or later
Memory	1 GB RAM
Hard Disk	250 MB available disk space
Display	1024 x 768 or higher resolution

https://www.snapgene.com/snapgene-viewer/

Download files from internet (II)



- **Support for complex barcodes, e.g. inDrop:**
  - Complex barcodes in STARsolo with --soloType CB_UMI_Complex, --soloCBmatchWLtype --soloAdapterSequence, --soloAdapterMismatchesNmax, --soloCBposition,--soloUMIposition
- **BAM tags:**
  - CB/UB for corrected CellBarcode/UMI
  - GX/GN for gene ID/name
- STARsolo most up-to-date documentation.

Assets 2

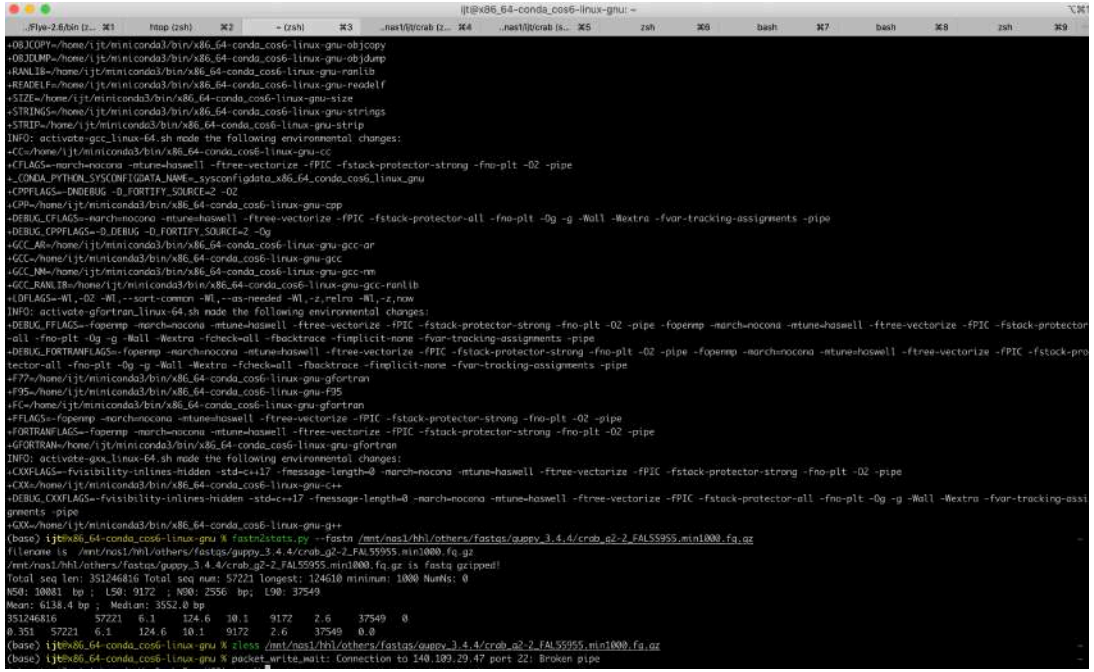
Source code (zip)

Source code (tar.gz)

Source code
Some may contain executable binaries
Some need to be compiled from scratch
Usually come in compressed file (need to decompress them)

https://www.snapgene.com/snapgene-viewer/

Console and Command-line interface

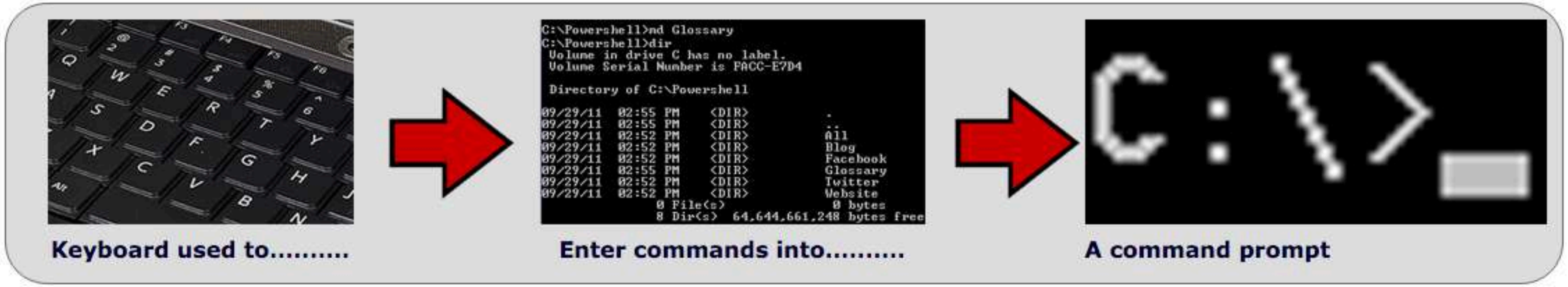


```
+OBJCOPY=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-objcopy
+OBJDUMP=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-objdump
+RANLIB=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-ranlib
+READELF=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-readelf
+SIZE=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-size
+STRINGS=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-strings
+STRIP=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-strip
INFO: activate-gcc_linux-64.sh made the following environmental changes:
+CC=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-cc
+CFLAGS=-march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-strong -fno-plt -O2 -pipe
+_CONDA_PYTHON_SYSCONFIGDATA_NAME=_sysconfigdata_x86_64_conda_cos6_linux_gnu
+CPPFLAGS=-DNDEBUG -D_FORTIFY_SOURCE=2 -O2
+CPP=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-cpp
+DEBUG_CFLAGS=-march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-all -fno-plt -Og -g -Wall -Wextra -fvar-tracking-assignments -pipe
+DEBUG_CPPFLAGS=-D_DEBUG -D_FORTIFY_SOURCE=2 -Og
+GCC_AR=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-gcc-ar
+GCC=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-gcc
+GCC_NM=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-gcc-nm
+GCC_RANLIB=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-gcc-ranlib
+LDFLAGS=-Wl,-O2 -Wl,--sort-common -Wl,--as-needed -Wl,-z,relro -Wl,-z,now
INFO: activate-gfortran_linux-64.sh made the following environmental changes:
+DEBUG_FFLAGS=-fopenmp -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-strong -fno-plt -O2 -pipe -fopenmp -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector
-all -fno-plt -Og -g -Wall -Wextra -fcheck=all -fbacktrace -fimplicit-none -fvar-tracking-assignments -pipe
+DEBUG_FORTRANFLAGS=-fopenmp -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-strong -fno-plt -O2 -pipe -fopenmp -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-pro
tector-all -fno-plt -Og -g -Wall -Wextra -fcheck=all -fbacktrace -fimplicit-none -fvar-tracking-assignments -pipe
+F77=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-gfortran
+F95=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-f95
+FC=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-gfortran
+FFLAGS=-fopenmp -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-strong -fno-plt -O2 -pipe
+FORTRANFLAGS=-fopenmp -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-strong -fno-plt -O2 -pipe
+GFORTRAN=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-gfortran
INFO: activate-gxx_linux-64.sh made the following environmental changes:
+CXXFLAGS=-fvisibility-inlines-hidden -std=c++17 -fmessage-length=0 -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-strong -fno-plt -O2 -pipe
+CXX=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-c++
+DEBUG_CXXFLAGS=-fvisibility-inlines-hidden -std=c++17 -fmessage-length=0 -march=nocona -mtune=haswell -ftree-vectorize -fPIC -fstack-protector-all -fno-plt -Og -g -Wall -Wextra -fvar-tracking-assi
gnments -pipe
+GXX=/home/ijt/miniconda3/bin/x86_64-conda_cos6-linux-gnu-g++
(base) ijt@x86_64-conda_cos6-linux-gnu % fastn2stats.py --fastn /mnt/nas1/hhl/others/fastqs/guppy_3.4.4/crab_g2-2_FAL55955.min1000.fq.gz
filename is  /mnt/nas1/hhl/others/fastqs/guppy_3.4.4/crab_g2-2_FAL55955.min1000.fq.gz
/mnt/nas1/hhl/others/fastqs/guppy_3.4.4/crab_g2-2_FAL55955.min1000.fq.gz is fastq gzipped!
Total seq len: 351246816 Total seq num: 57221 longest: 124610 minimum: 1000 NumNs: 0
N50: 10081  bp ;  L50: 9172  ; N90: 2556  bp;  L90: 37549
Mean: 6138.4 bp ;  Median: 3552.0 bp
351246816       57221   6.1     124.6   10.1    9172    2.6     37549   0
0.351   57221   6.1     124.6   10.1    9172    2.6     37549   0.0
(base) ijt@x86_64-conda_cos6-linux-gnu % zless /mnt/nas1/hhl/others/fastqs/guppy_3.4.4/crab_g2-2_FAL55955.min1000.fq.gz
(base) ijt@x86_64-conda_cos6-linux-gnu % packet_write_wait: Connection to 140.109.29.47 port 22: Broken pipe
```

Console and Command-line interface

Computer terminal or system consoles are the **text entry and display device** for system administration messages, particularly those from the BIOS or boot loader, the kernel, from the init system and from the system logger. It is a **physical device consisting of a keyboard and a screen**.

A **command-line interface** is a means of interacting with a computer program where the **user** issues **commands** to the program (putty, terminal) in the form of successive lines of text (command lines).



```
C:\Powershell>md Glossary
C:\Powershell>dir
 Volume in drive C has no label.
 Volume Serial Number is FACC-E7D4

 Directory of C:\Powershell

09/29/11  02:55 PM    <DIR>          .
09/29/11  02:55 PM    <DIR>          ..
09/29/11  02:52 PM    <DIR>          All
09/29/11  02:52 PM    <DIR>          Blog
09/29/11  02:52 PM    <DIR>          Facebook
09/29/11  02:55 PM    <DIR>          Glossary
09/29/11  02:52 PM    <DIR>          Twitter
09/29/11  02:52 PM    <DIR>          Website
               0 File(s)              0 bytes
               8 Dir(s)  64,644,661,248 bytes free
```

Keyboard used to..........   **Enter commands into..........**   **A command prompt**

http://www.ictlounge.com/html/operating_systems.htm Wiki

Using command line in day-to-day bioinformatics

- Most sequence files are text files
- Text mining easy!

- Features programming functions (e.g., loops, variables)
- Lots of little scripts

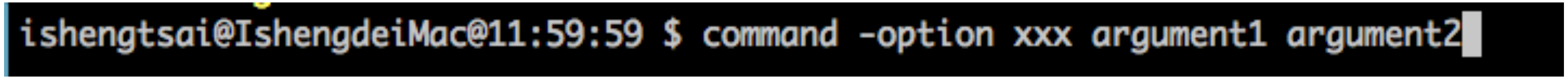
- Package everything (scripts, programs) into working pipelines

- Automation and reproducibility

- Remote access

A typical command

Options always start with ‘-’, and often expect to receive an option (xxx)

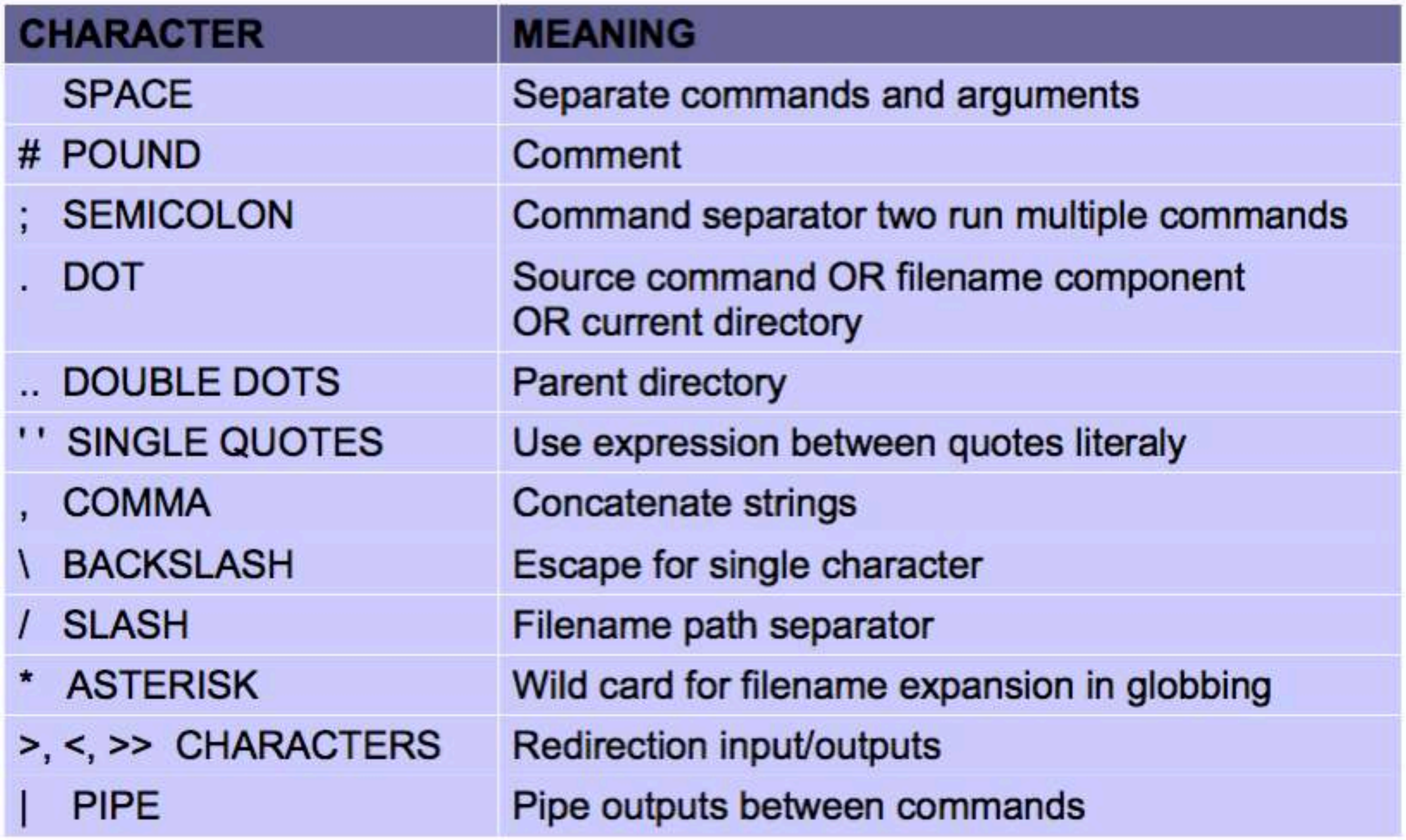


```
ishengtsai@IshengdeiMac@11:59:59 $ command -option xxx argument1 argument2
```

Application or script name

Argument can be passed to programs

Special characters in bash



CHARACTER	MEANING
SPACE	Separate commands and arguments
# POUND	Comment
; SEMICOLON	Command separator two run multiple commands
. DOT	Source command OR filename component OR current directory
.. DOUBLE DOTS	Parent directory
' ' SINGLE QUOTES	Use expression between quotes literaly
, COMMA	Concatenate strings
\ BACKSLASH	Escape for single character
/ SLASH	Filename path separator
* ASTERISK	Wild card for filename expansion in globbing
>, <, >> CHARACTERS	Redirection input/outputs
| PIPE	Pipe outputs between commands

Special characters in bash

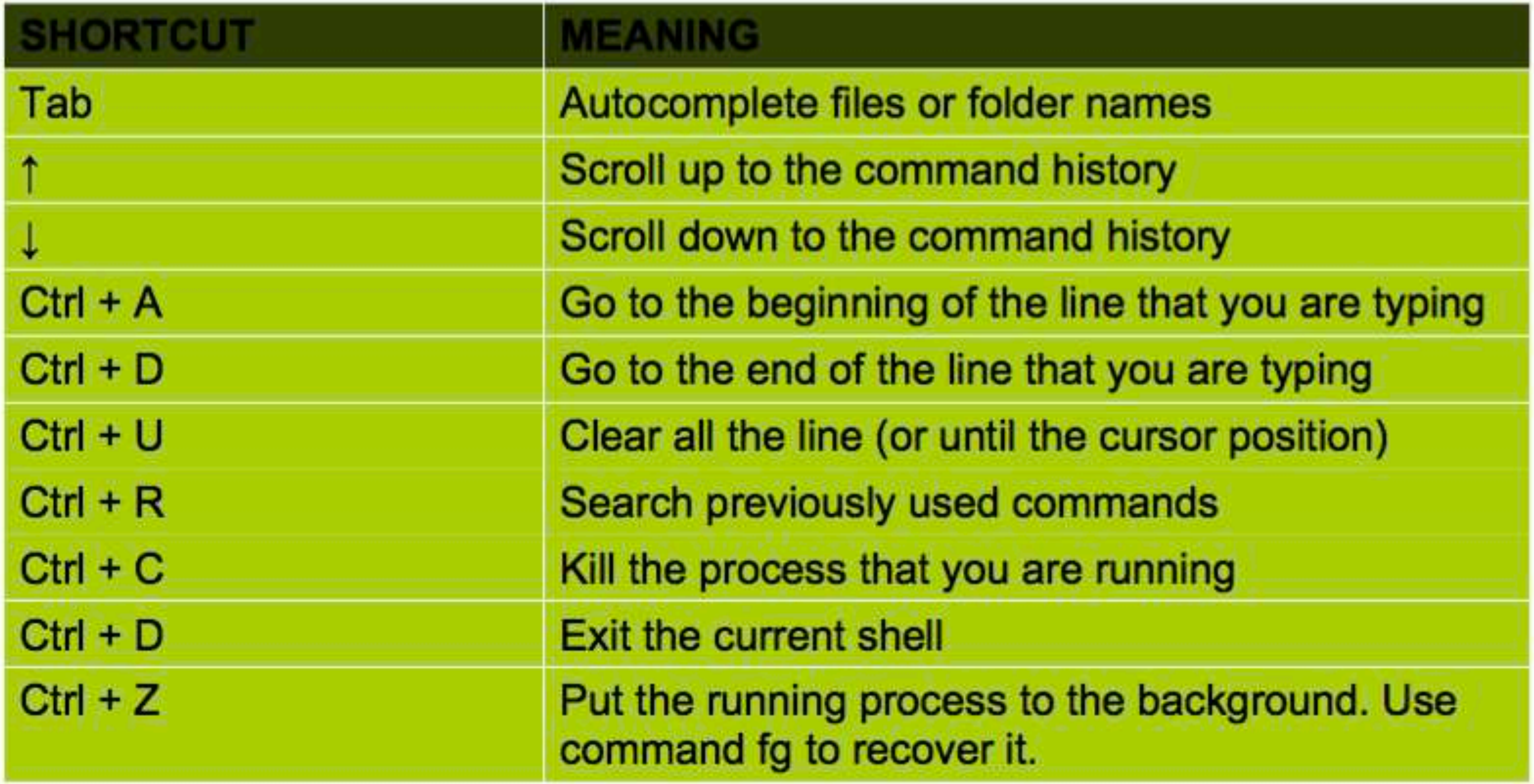
```
$ command xxxx yyyyy
```

Linux treats xxxx and yyyy as two arguments of the command

```
$ command ‘xxxx yyyyy’
$ command xxxx\ yyyy
```

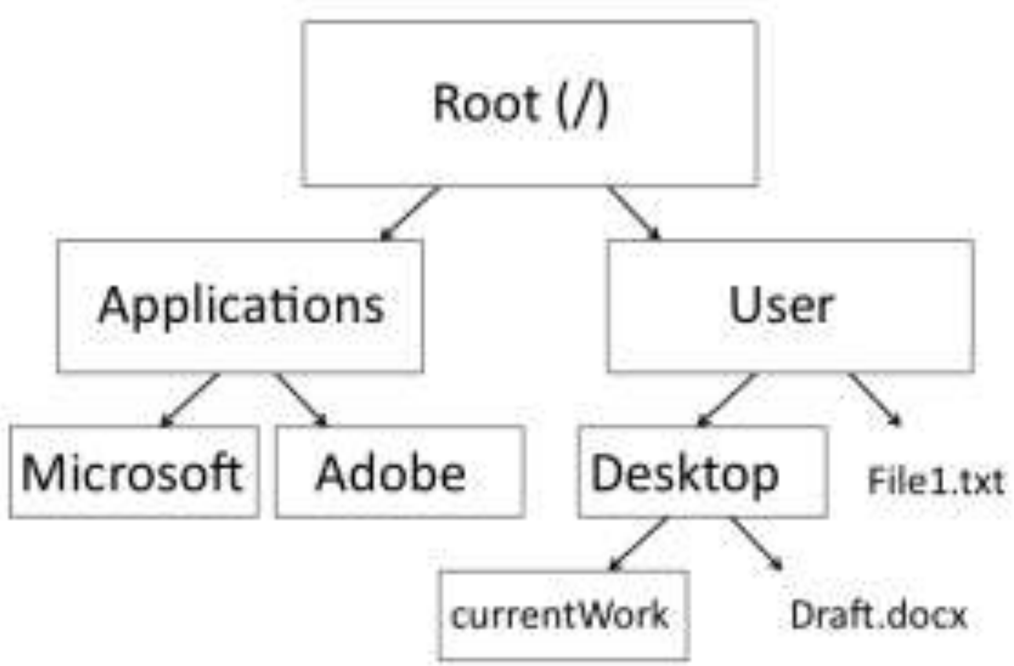
You can uses single quotes or escape to distinguish special characters (in this case: space)

Short cut and emergency command in linux



SHORTCUT	MEANING
Tab	Autocomplete files or folder names
↑	Scroll up to the command history
↓	Scroll down to the command history
Ctrl + A	Go to the beginning of the line that you are typing
Ctrl + D	Go to the end of the line that you are typing
Ctrl + U	Clear all the line (or until the cursor position)
Ctrl + R	Search previously used commands
* Ctrl + C	Kill the process that you are running
Ctrl + D	Exit the current shell
Ctrl + Z	Put the running process to the background. Use command fg to recover it.

Directory structure



Try:
ls (list segment)
cd (change directory)
rm (abbreviation for remove)
mkdir (make directory)
pwd (print working directory)

Directory structure is like a tree

From /home/ishengtsai/

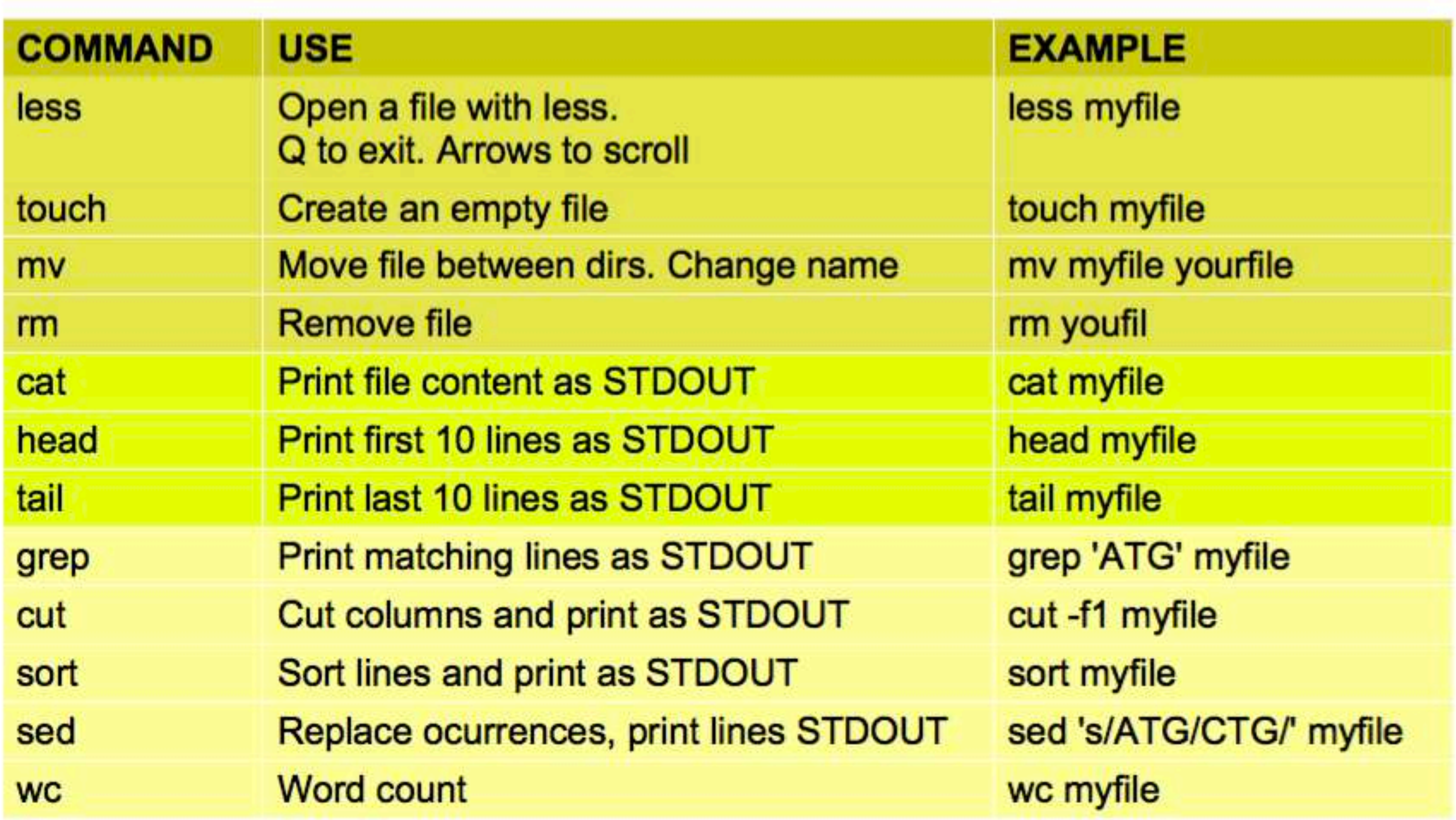
Relative path:

```
cd fungi        # moves into fungi folder
                # now you are in /home/ishengtsai/fungi/
                # you can only do this successfully when you are in /home/ishengtsai/

cd ..           # you go up one directory
                # now you are in /home/
```

Or absolute path:

```
cd /home/ishengtsai/fungi/ ;
```

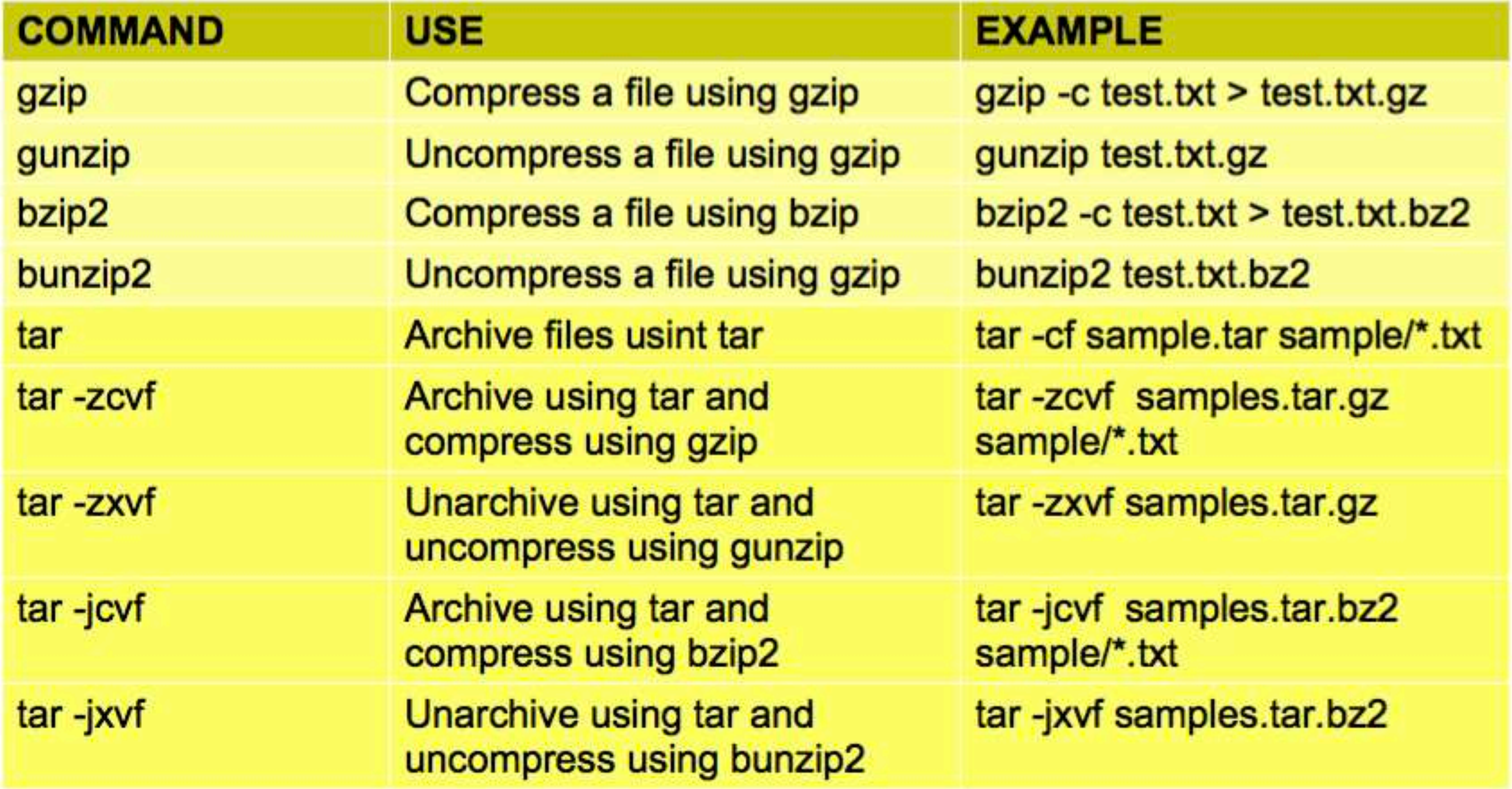
Files commands **



COMMAND	USE	EXAMPLE
less	Open a file with less. Q to exit. Arrows to scroll	less myfile
touch	Create an empty file	touch myfile
mv	Move file between dirs. Change name	mv myfile yourfile
rm	Remove file	rm youfil
cat	Print file content as STDOUT	cat myfile
head	Print first 10 lines as STDOUT	head myfile
tail	Print last 10 lines as STDOUT	tail myfile
grep	Print matching lines as STDOUT	grep 'ATG' myfile
cut	Cut columns and print as STDOUT	cut -f1 myfile
sort	Sort lines and print as STDOUT	sort myfile
sed	Replace ocurrences, print lines STDOUT	sed 's/ATG/CTG/' myfile
wc	Word count	wc myfile

awk https://en.wikipedia.org/wiki/AWK

Compression commands



COMMAND	USE	EXAMPLE
gzip	Compress a file using gzip	gzip -c test.txt > test.txt.gz
gunzip	Uncompress a file using gzip	gunzip test.txt.gz
bzip2	Compress a file using bzip	bzip2 -c test.txt > test.txt.bz2
bunzip2	Uncompress a file using gzip	bunzip2 test.txt.bz2
tar	Archive files usint tar	tar -cf sample.tar sample/*.txt
tar -zcvf	Archive using tar and compress using gzip	tar -zcvf samples.tar.gz sample/*.txt
tar -zxvf	Unarchive using tar and uncompress using gunzip	tar -zxvf samples.tar.gz
tar -jcvf	Archive using tar and compress using bzip2	tar -jcvf samples.tar.bz2 sample/*.txt
tar -jxvf	Unarchive using tar and uncompress using bunzip2	tar -jxvf samples.tar.bz2

Redirection of input / output

```
# The result of the ls command will be output and saved into out.txt
```

```
$ ls > out.txt
```

```
# The result of the ls command will be output and append into out.txt
# If the file out.txt already exists, then the original content will not be replaced, and
# the new information will be added into the file
```

```
$ ls >> out.txt
```

Pipeline

… a **pipeline** is a set of **processes** chained by their **standard streams**, so that the output of each process (stdout) feeds directly as input (stdin) to the next one.

```
program1 | program2 | program3
```

Special character to **pipe** the results

Example:

```
ls -l | grep key | less
```

Demonstration I: daily tasks

1. Login into a terminal
2. Go to a specific directory that contains your data
3. Inspect your **fasta** files

   ```
   $ less ref.fa | grep '>' | less
   $ less ref.fa | grep '>' | wc -l
   ```

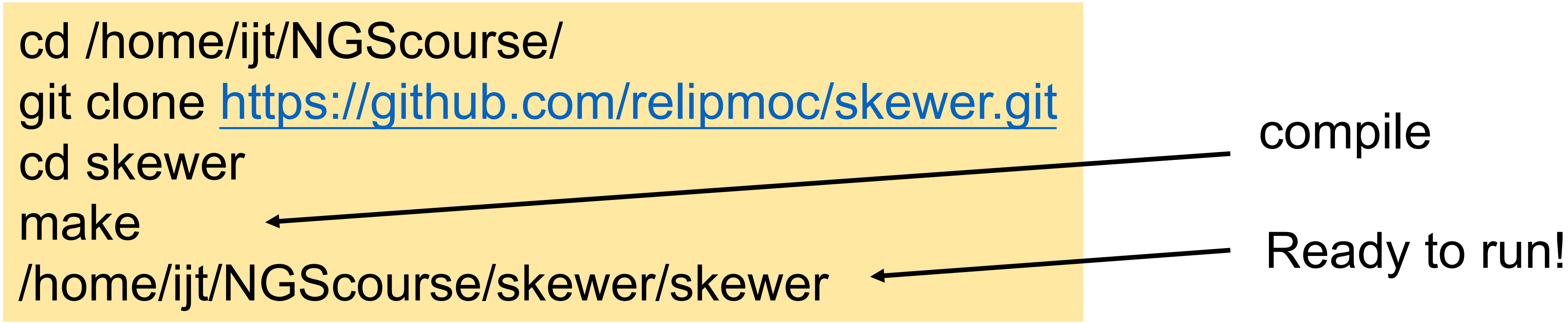
4. How about **fastq** file?
   - how many sequences?

5. How about gff file?
   - how many exons? How many genes?
   - how many genes that are expressed in the forward strand?

6. Check if command is successful

Installation

1. You need a bioinformatics program
   1. Download binaries and it should be ready to execute
   2. Or you have to compile
   3. Most modern program now deposit their program in **github**



```
cd /home/ijt/NGScourse/
git clone https://github.com/relipmoc/skewer.git
cd skewer
make
/home/ijt/NGScourse/skewer/skewer
```

compile

Ready to run!

Jiang *et al. BMC Bioinformatics* 2014, **15**:182
http://www.biomedcentral.com/1471-2105/15/182

BMC Bioinformatics

METHODOLOGY ARTICLE — Open Access

Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads

Hongshan Jiang[1*], Rong Lei[1], Shou-Wei Ding[2] and Shuifang Zhu[1]

Demonstration II: daily tasks

1. Downloaded some sequenced data ; mapped to genome and you want to start looking at it.

2. Look at sam file
   $ samtools view xxx.bam | less

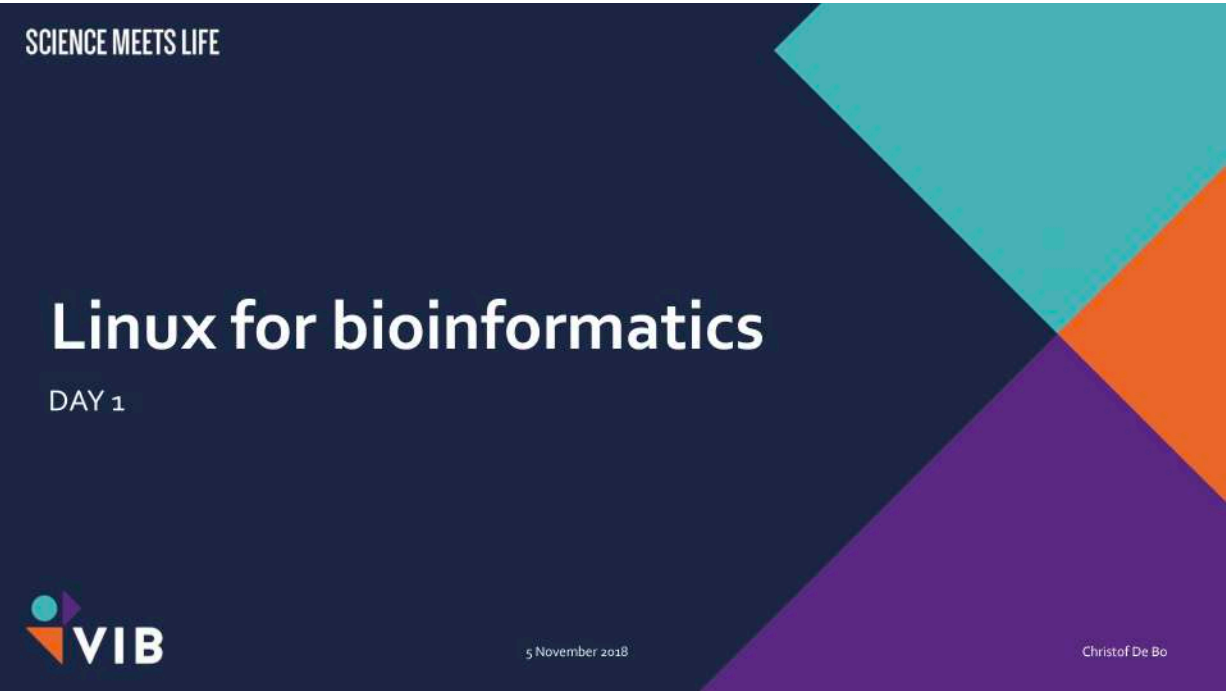
3. Okay, how about if I want to check the insert size of properly mapped reads? What filter to use? (https://broadinstitute.github.io/picard/explain-flags.html)

4. You have a file that you want to visualize, what next?

Demonstration III: daily tasks

1. How many genes are there in a gff file?
2. Length of fasta files
3. Longest sequence in the fasta (if not sorted)
4. Scripts to find out
5. Echo
6. For loop

Good references



SCIENCE MEETS LIFE

Linux for bioinformatics

DAY 1

VIB

5 November 2018

Christof De Bo

SCIENCE MEETS LIFE

Linux for bioinformatics

DAY 2

VIB

12 November 2018

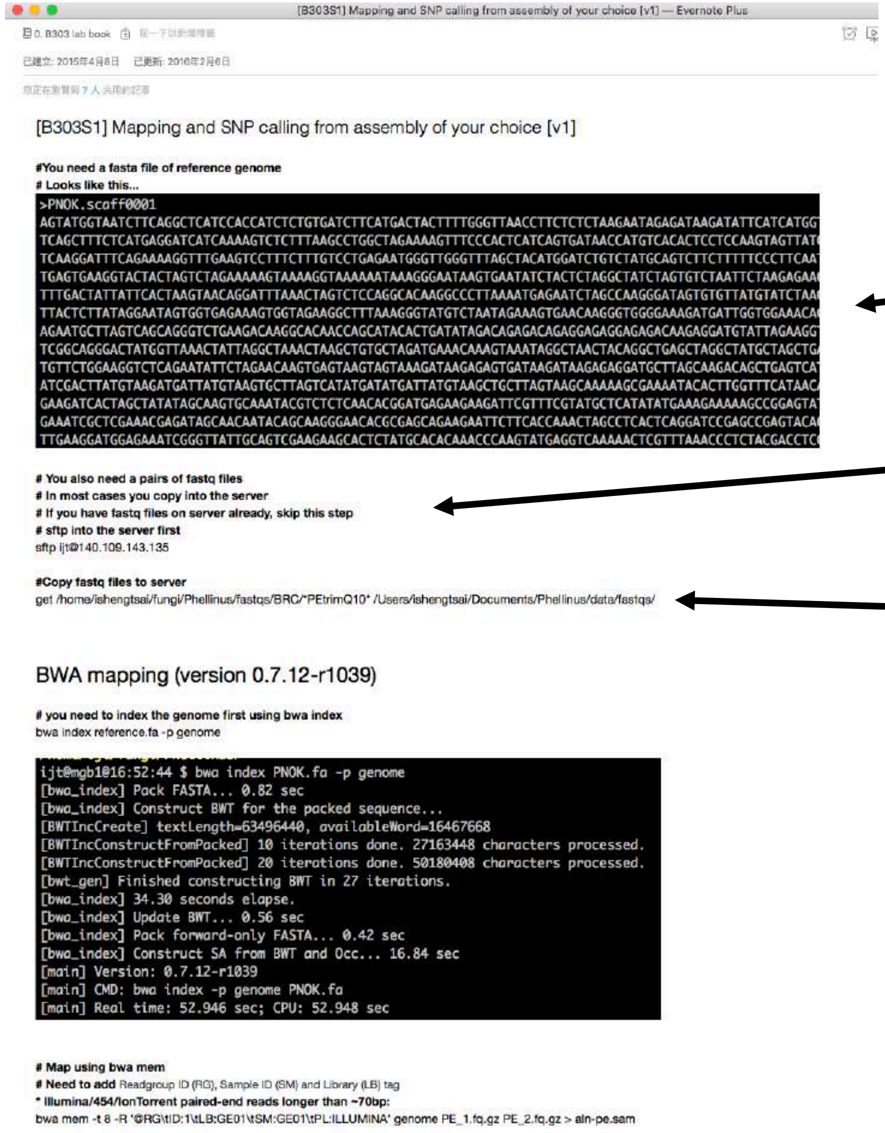
Christof De Bo

https://www.bits.vib.be/training-list/112-bits/training/upcoming-trainings/124-linux-for-bioinformatics

Keep tracking

Keep a track of your science

Evernote; onenote; notion.. Etc?



[B303S1] Mapping and SNP calling from assembly of your choice [v1] — Evernote Plus
0. B303 lab book
按一下以新增標籤
已建立: 2015年4月8日
已更新: 2016年2月6日
您正在與與 7 人 共用的記事

[B303S1] Mapping and SNP calling from assembly of your choice [v1]

#You need a fasta file of reference genome
# Looks like this...

```
>PNOK.scaff0001
AGTATGGTAATCTTCAGGCTCATCCACCATCTCTGTGATCTTCATGACTACTTTTGGGTTAACCTTCTCTCTAAGAATAGAGATAAGATATTCATCATGG
TCAGCTTTCTCATGAGGATCATCAAAAGTCTCTTTAAGCCTGGCTAGAAAAGTTTCCCACTCATCAGTGATAACCATGTCACACTCCTCCAAGTAGTTAT
TCAAGGATTTCAGAAAAGGTTTGAAGTCCTTTCTTTGTCCTGAGAATGGGTTGGGTTTAGCTACATGGATCTGTCTATGCAGTCTTCTTTTTCCCTTCAA
TGAGTGAAGGTACTACTAGTCTAGAAAAAGTAAAAGGTAAAAAATAAAGGGAATAAGTGAATATCTACTCTAGGCTATCTAGTGTCTAATTCTAAGAGAA
TTTGACTATTATTCACTAAGTAACAGGATTTAAACTAGTCTCCAGGCACAAGGCCCTTAAAATGAGAATCTAGCCAAGGGATAGTGTGTTATGTATCTAA
TTACTCTTATAGGAATAGTGGTGAGAAAGTGGTAGAAGGCTTTAAAGGGTATGTCTAATAGAAAGTGAACAAGGGTGGGGAAAGATGATTGGTGGAAACA
AGAATGCTTAGTCAGCAGGGTCTGAAGACAAGGCACAACCAGCATACACTGATATAGACAGAGACAGAGGAGAGGAGAGACAAGAGGATGTATTAGAAGG
TCGGCAGGGACTATGGTTAAACTATTAGGCTAAACTAAGCTGTGCTAGATGAAACAAAGTAAATAGGCTAACTACAGGCTGAGCTAGGCTATGCTAGCTG
TGTTCTGGAAGGTCTCAGAATATTCTAGAACAAGTGAGTAAGTAGTAAAGATAAGAGAGTGATAAGATAAGAGAGGATGCTTAGCAAGACAGCTGAGTCA
ATCGACTTATGTAAGATGATTATGTAAGTGCTTAGTCATATGATATGATTATGTAAGCTGCTTAGTAAGCAAAAAGCGAAAATACACTTGGTTTCATAAC
GAAGATCACTAGCTATATAGCAAGTGCAAATACGTCTCTCAACACGGATGAGAAGAAGATTCGTTTCGTATGCTCATATATGAAAGAAAAAGCCGGAGTA
GAAATCGCTCGAAACGAGATAGCAACAATACAGCAAGGGAACACGCGAGCAGAAGAATTCTTCACCAAACTAGCCTCACTCAGGATCCGAGCCGAGTACA
TTGAAGGATGGAGAAATCGGGTTATTGCAGTCGAAGAAGCACTCTATGCACACAAACCCAAGTATGAGGTCAAAAACTCGTTTAAACCCTCTACGACCTC
```

# You also need a pairs of fastq files
# In most cases you copy into the server
# If you have fastq files on server already, skip this step
# sftp into the server first
sftp ijt@140.109.143.135

#Copy fastq files to server
get /home/ishengtsai/fungi/Phellinus/fastqs/BRC/*PEtrimQ10* /Users/ishengtsai/Documents/Phellinus/data/fastqs/

BWA mapping (version 0.7.12-r1039)

# you need to index the genome first using bwa index
bwa index reference.fa -p genome

```
ijt@mgb1@16:52:44 $ bwa index PNOK.fa -p genome
[bwa_index] Pack FASTA... 0.82 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=63496440, availableWord=16467668
[BWTIncConstructFromPacked] 10 iterations done. 27163448 characters processed.
[BWTIncConstructFromPacked] 20 iterations done. 50180408 characters processed.
[bwt_gen] Finished constructing BWT in 27 iterations.
[bwa_index] 34.30 seconds elapse.
[bwa_index] Update BWT... 0.56 sec
[bwa_index] Pack forward-only FASTA... 0.42 sec
[bwa_index] Construct SA from BWT and Occ... 16.84 sec
[main] Version: 0.7.12-r1039
[main] CMD: bwa index -p genome PNOK.fa
[main] Real time: 52.946 sec; CPU: 52.948 sec
```

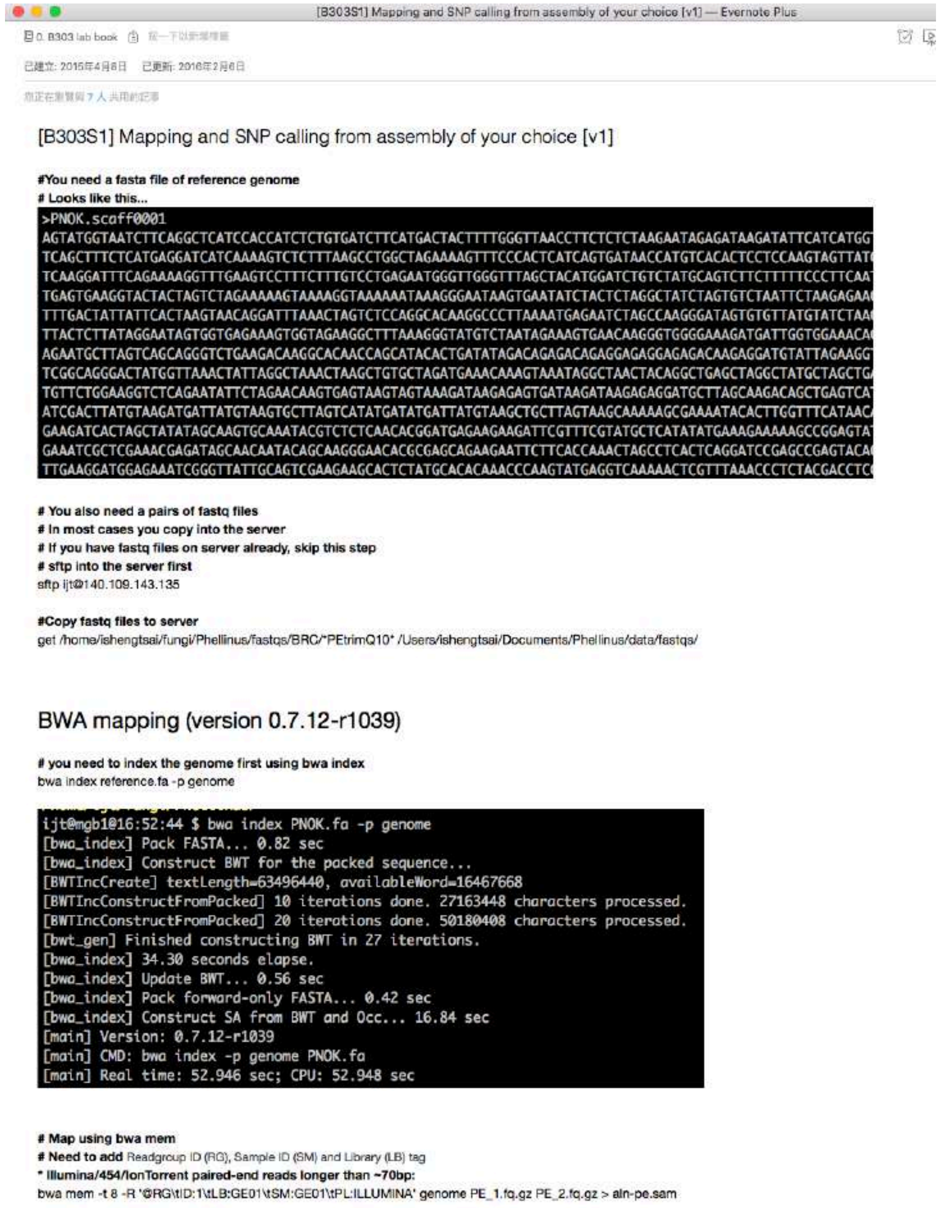
# Map using bwa mem
# Need to add Readgroup ID (RG), Sample ID (SM) and Library (LB) tag
*** Illumina/454/IonTorrent paired-end reads longer than ~70bp:**
bwa mem -t 8 -R '@RG\tID:1\tLB:GE01\tSM:GE01\tPL:ILLUMINA' genome PE_1.fq.gz PE_2.fq.gz > aln-pe.sam

Screenshot to log results

# Comment your code (what was the purpose)

All the command can be reused (copy and paste!)

Evernote / Notion



[B303S1] Mapping and SNP calling from assembly of your choice [v1] — Evernote Plus

0. B303 lab book   按一下以新增標籤

已建立: 2015年4月8日   已更新: 2016年2月6日

您正在與其他 7 人共用的記事

[B303S1] Mapping and SNP calling from assembly of your choice [v1]

#You need a fasta file of reference genome
# Looks like this...

```
>PNOK.scaff0001
AGTATGGTAATCTTCAGGCTCATCCACCATCTCTGTGATCTTCATGACTACTTTTGGGTTAACCTTCTCTCTAAGAATAGAGATAAGATATTCATCATGG
TCAGCTTTCTCATGAGGATCATCAAAAGTCTCTTTAAGCCTGGCTAGAAAAGTTTCCCACTCATCAGTGATAACCATGTCACACTCCTCCAAGTAGTTAT
TCAAGGATTTCAGAAAAGGTTTGAAGTCCTTTCTTTGTCCTGAGAATGGGTTGGGTTTAGCTACATGGATCTGTCTATGCAGTCTTCTTTTTCCCTTCAAT
TGAGTGAAGGTACTACTAGTCTAGAAAAAGTAAAAGGTAAAAAATAAAGGGAATAAGTGAATATCTACTCTAGGCTATCTAGTGTCTAATTCTAAGAGAA
TTTGACTATTATTCACTAAGTAACAGGATTTAAACTAGTCTCCAGGCACAAGGCCCTTAAAATGAGAATCTAGCCAAGGGATAGTGTGTTATGTATCTAA
TTACTCTTATAGGAATAGTGGTGAGAAAGTGGTAGAAGGCTTTAAAGGGTATGTCTAATAGAAAGTGAACAAGGGTGGGGAAAGATGATTGGTGGAAACA
AGAATGCTTAGTCAGCAGGGTCTGAAGACAAGGCACAACCAGCATACACTGATATAGACAGAGACAGAGGAGAGGAGAGACAAGAGGATGTATTAGAAGG
TCGGCAGGGACTATGGTTAAACTATTAGGCTAAACTAAGCTGTGCTAGATGAAACAAAGTAAATAGGCTAACTACAGGCTGAGCTAGGCTATGCTAGCTGA
TGTTCTGGAAGGTCTCAGAATATTCTAGAACAAGTGAGTAAGTAGTAAAGATAAGAGAGTGATAAGATAAGAGAGGATGCTTAGCAAGACAGCTGAGTCAT
ATCGACTTATGTAAGATGATTATGTAAGTGCTTAGTCATATGATATGATTATGTAAGCTGCTTAGTAAGCAAAAAGCGAAAATACACTTGGTTTCATAAC
GAAGATCACTAGCTATATAGCAAGTGCAAATACGTCTCTCAACACGGATGAGAAGAAGATTCGTTTCGTATGCTCATATATGAAAGAAAAAGCCGGAGTAT
GAAATCGCTCGAAACGAGATAGCAACAATACAGCAAGGGAACACGCGAGCAGAAGAATTCTTCACCAAACTAGCCTCACTCAGGATCCGAGCCGAGTACA
TTGAAGGATGGAGAAATCGGGTTATTGCAGTCGAAGAAGCACTCTATGCACACAAACCCAAGTATGAGGTCAAAAACTCGTTTAAACCCTCTACGACCTC
```

# You also need a pairs of fastq files
# In most cases you copy into the server
# If you have fastq files on server already, skip this step
# sftp into the server first
sftp ijt@140.109.143.135

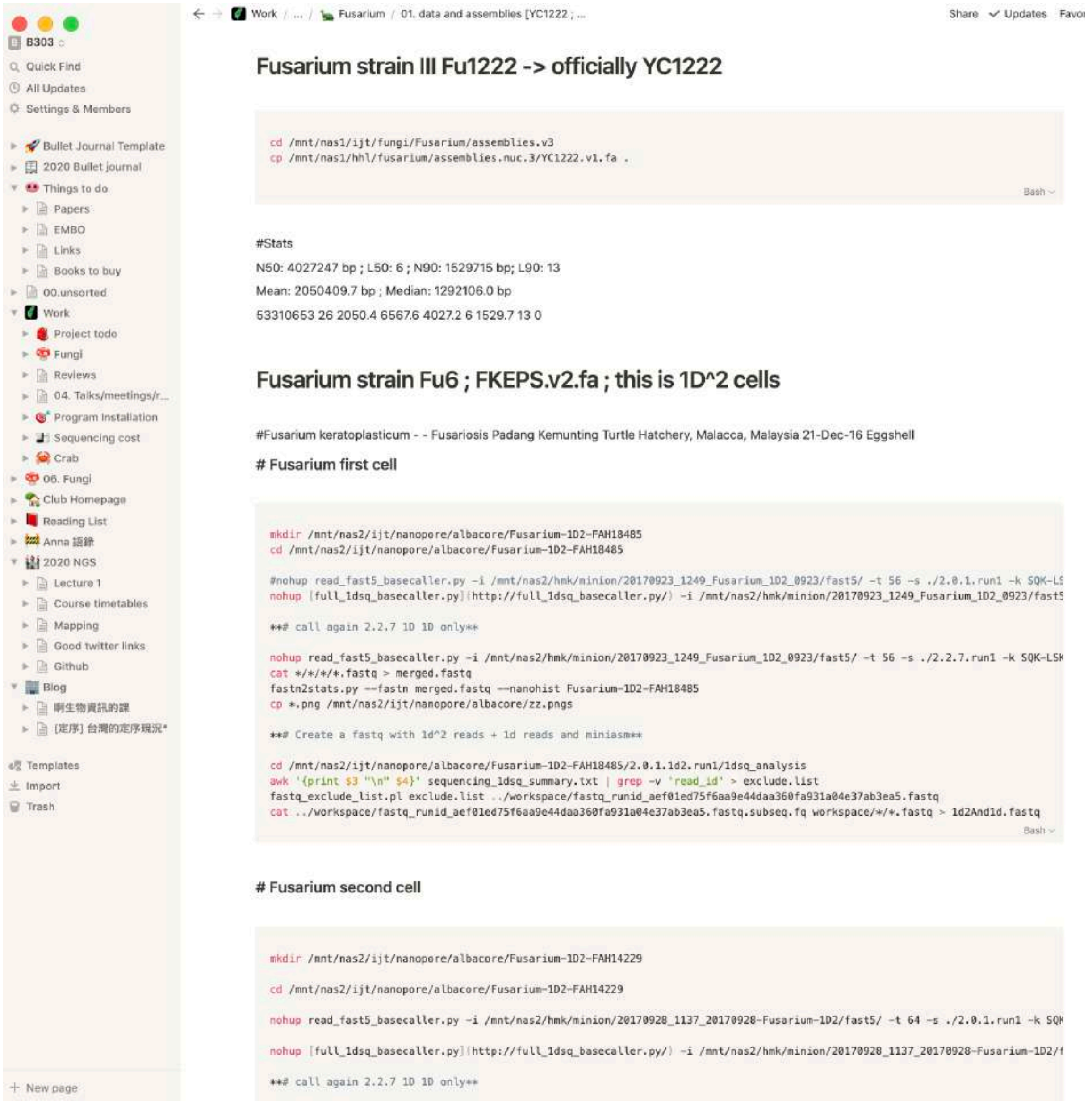
#Copy fastq files to server
get /home/ishengtsai/fungi/Phellinus/fastqs/BRC/*PEtrimQ10* /Users/ishengtsai/Documents/Phellinus/data/fastqs/

BWA mapping (version 0.7.12-r1039)

# you need to index the genome first using bwa index
bwa index reference.fa -p genome

```
ijt@mgb1@16:52:44 $ bwa index PNOK.fa -p genome
[bwa_index] Pack FASTA... 0.82 sec
[bwa_index] Construct BWT for the packed sequence...
[BWTIncCreate] textLength=63496440, availableWord=16467668
[BWTIncConstructFromPacked] 10 iterations done. 27163448 characters processed.
[BWTIncConstructFromPacked] 20 iterations done. 50180408 characters processed.
[bwt_gen] Finished constructing BWT in 27 iterations.
[bwa_index] 34.30 seconds elapse.
[bwa_index] Update BWT... 0.56 sec
[bwa_index] Pack forward-only FASTA... 0.42 sec
[bwa_index] Construct SA from BWT and Occ... 16.84 sec
[main] Version: 0.7.12-r1039
[main] CMD: bwa index -p genome PNOK.fa
[main] Real time: 52.946 sec; CPU: 52.948 sec
```

# Map using bwa mem
# Need to add Readgroup ID (RG), Sample ID (SM) and Library (LB) tag
*** Illumina/454/IonTorrent paired-end reads longer than ~70bp:**
bwa mem -t 8 -R '@RG\tID:1\tLB:GE01\tSM:GE01\tPL:ILLUMINA' genome PE_1.fq.gz PE_2.fq.gz > aln-pe.sam



Fusarium strain III Fu1222 -> officially YC1222

```
cd /mnt/nas1/ijt/fungi/Fusarium/assemblies.v3
cp /mnt/nas1/hhl/fusarium/assemblies.nuc.3/YC1222.v1.fa .
```

#Stats

N50: 4027247 bp ; L50: 6 ; N90: 1529715 bp; L90: 13

Mean: 2050409.7 bp ; Median: 1292106.0 bp

53310653 26 2050.4 6567.6 4027.2 6 1529.7 13 0

Fusarium strain Fu6 ; FKEPS.v2.fa ; this is 1D^2 cells

#Fusarium keratoplasticum - - Fusariosis Padang Kemunting Turtle Hatchery, Malacca, Malaysia 21-Dec-16 Eggshell

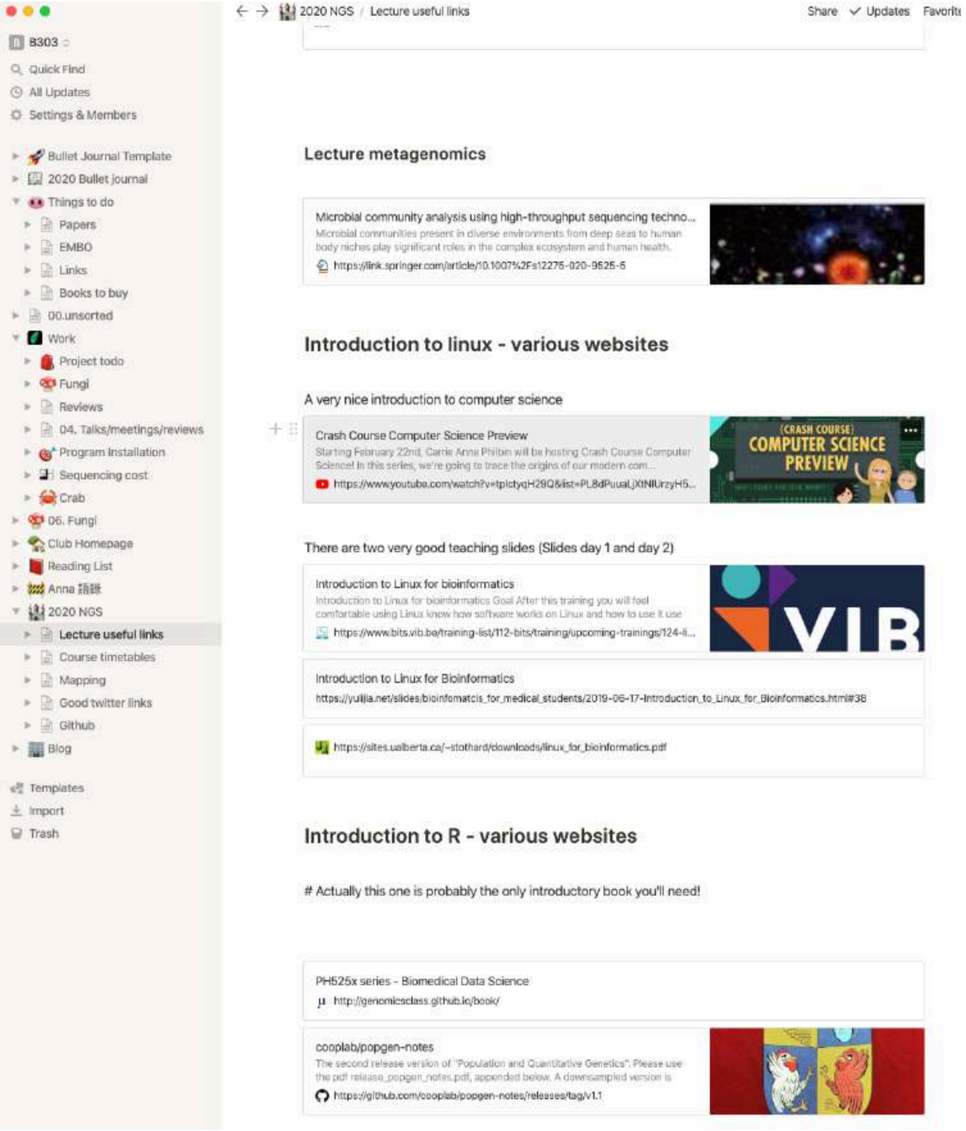
Fusarium first cell

```
mkdir /mnt/nas2/ijt/nanopore/albacore/Fusarium-1D2-FAH18485
cd /mnt/nas2/ijt/nanopore/albacore/Fusarium-1D2-FAH18485

#nohup read_fast5_basecaller.py -i /mnt/nas2/hmk/minion/20170923_1249_Fusarium_1D2_0923/fast5/ -t 56 -s ./2.0.1.run1 -k SQK-LS
nohup [full_1dsq_basecaller.py](http://full_1dsq_basecaller.py/) -i /mnt/nas2/hmk/minion/20170923_1249_Fusarium_1D2_0923/fast5

**# call again 2.2.7 1D 1D only**

nohup read_fast5_basecaller.py -i /mnt/nas2/hmk/minion/20170923_1249_Fusarium_1D2_0923/fast5/ -t 56 -s ./2.2.7.run1 -k SQK-LSK
cat */*/*/*.fastq > merged.fastq
fastn2stats.py --fastn merged.fastq --nanohist Fusarium-1D2-FAH18485
cp *.png /mnt/nas2/ijt/nanopore/albacore/zz.pngs

**# Create a fastq with 1d^2 reads + 1d reads and miniasm**

cd /mnt/nas2/ijt/nanopore/albacore/Fusarium-1D2-FAH18485/2.0.1.1d2.run1/1dsq_analysis
awk '{print $3 "\n" $4}' sequencing_1dsq_summary.txt | grep -v 'read_id' > exclude.list
fastq_exclude_list.pl exclude.list ../workspace/fastq_runid_aef01ed75f6aa9e44daa360fa931a04e37ab3ea5.fastq
cat ../workspace/fastq_runid_aef01ed75f6aa9e44daa360fa931a04e37ab3ea5.fastq.subseq.fq workspace/*/*.fastq > 1d2And1d.fastq
```

Fusarium second cell

```
mkdir /mnt/nas2/ijt/nanopore/albacore/Fusarium-1D2-FAH14229

cd /mnt/nas2/ijt/nanopore/albacore/Fusarium-1D2-FAH14229

nohup read_fast5_basecaller.py -i /mnt/nas2/hmk/minion/20170928_1137_20170928-Fusarium-1D2/fast5/ -t 64 -s ./2.0.1.run1 -k SQK

nohup [full_1dsq_basecaller.py](http://full_1dsq_basecaller.py/) -i /mnt/nas2/hmk/minion/20170928_1137_20170928-Fusarium-1D2/f

**# call again 2.2.7 1D 1D only**
```

Readily share / reproducible



Lecture metagenomics

Microbial community analysis using high-throughput sequencing techno...
Microbial communities present in diverse environments from deep seas to human body niches play significant roles in the complex ecosystem and human health.
https://link.springer.com/article/10.1007%2Fs12275-020-9525-5

Introduction to linux - various websites

A very nice introduction to computer science

Crash Course Computer Science Preview
Starting February 22nd, Carrie Anne Philbin will be hosting Crash Course Computer Science! In this series, we're going to trace the origins of our modern com...
https://www.youtube.com/watch?v=tpIctyqH29Q&list=PL8dPuuaLjXtNlUrzyH5...

There are two very good teaching slides (Slides day 1 and day 2)

Introduction to Linux for bioinformatics
Introduction to Linux for bioinformatics Goal After this training you will feel comfortable using Linux know how software works on Linux and how to use it use
https://www.bits.vib.be/training-list/112-bits/training/upcoming-trainings/124-li...

Introduction to Linux for Bioinformatics
https://yulijia.net/slides/bioinfomatcis_for_medical_students/2019-06-17-Introduction_to_Linux_for_Bioinformatics.html#38

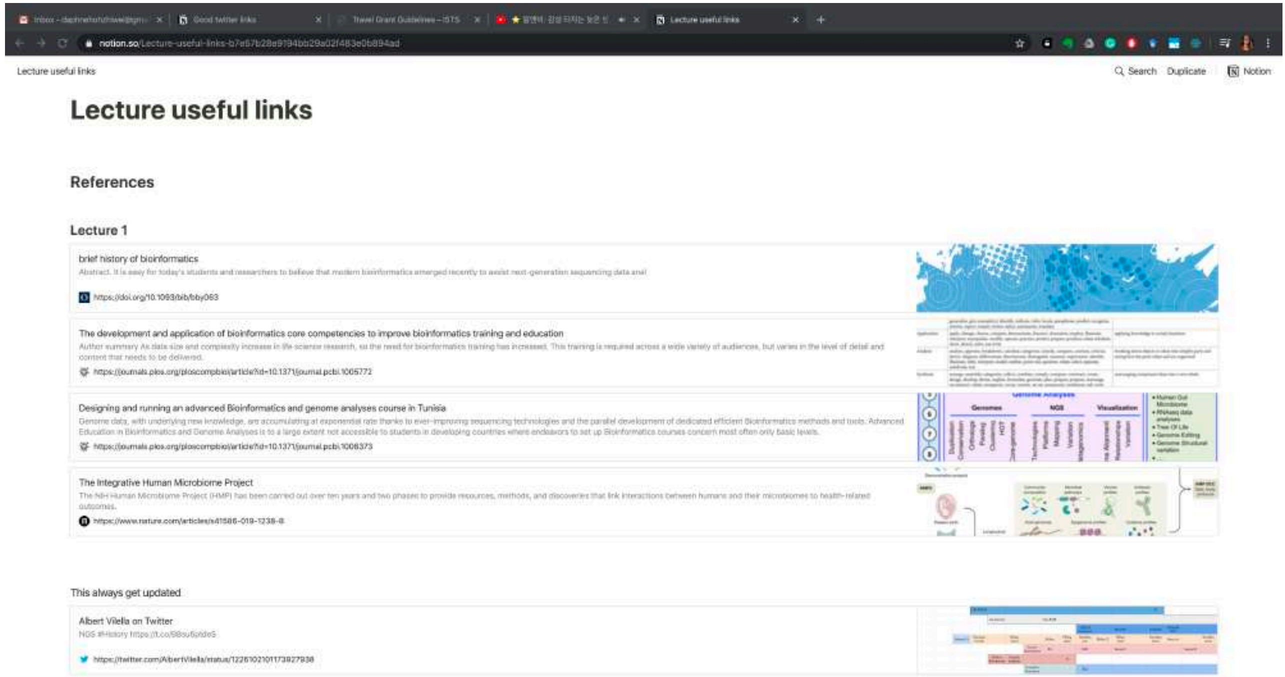
https://sites.ualberta.ca/~stothard/downloads/linux_for_bioinformatics.pdf

Introduction to R - various websites

Actually this one is probably the only introductory book you'll need!

PH525x series - Biomedical Data Science
http://genomicsclass.github.io/book/

cooplab/popgen-notes
The second release version of "Population and Quantitative Genetics". Please use the pdf release_popgen_notes.pdf, appended below. A downsampled version is
https://github.com/cooplab/popgen-notes/releases/tag/v1.1



Lecture useful links

References

Lecture 1

brief history of bioinformatics
Abstract. It is easy for today's students and researchers to believe that modern bioinformatics emerged recently to assist next-generation sequencing data anal
https://doi.org/10.1093/bib/bby063

The development and application of bioinformatics core competencies to improve bioinformatics training and education
Author summary As data size and complexity increase in life science research, so the need for bioinformatics training has increased. This training is required across a wide variety of audiences, but varies in the level of detail and content that needs to be delivered.
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005772

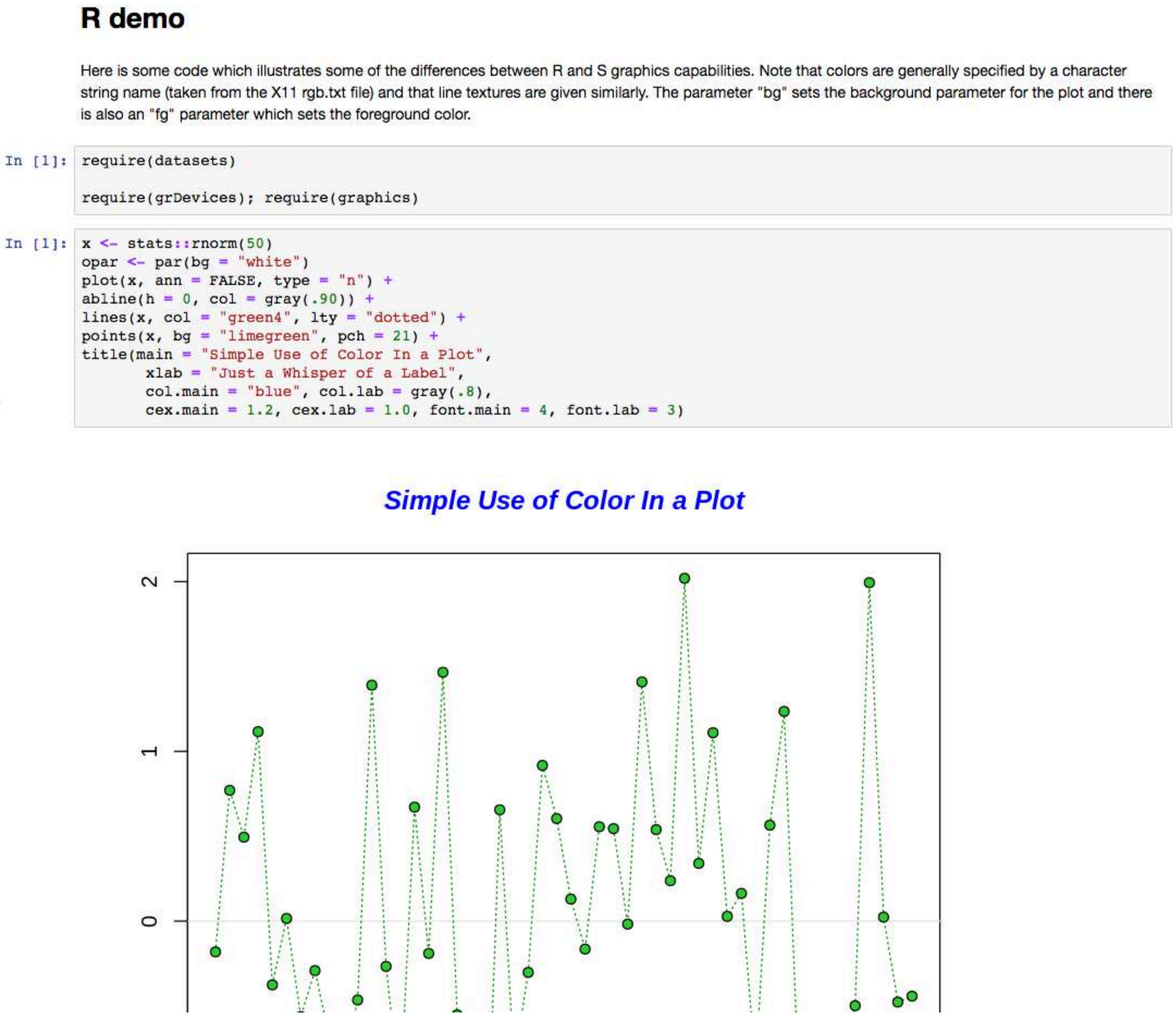
Designing and running an advanced Bioinformatics and genome analyses course in Tunisia
Genome data, with underlying new knowledge, are accumulating at exponential rate thanks to ever-improving sequencing technologies and the parallel development of dedicated efficient Bioinformatics methods and tools. Advanced Education in Bioinformatics and Genome Analyses is to a large extent not accessible to students in developing countries where endeavors to set up Bioinformatics courses concern most often only basic levels.
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006373

The Integrative Human Microbiome Project
The NIH Human Microbiome Project (HMP) has been carried out over ten years and two phases to provide resources, methods, and discoveries that link interactions between humans and their microbiomes to health-related outcomes.
https://www.nature.com/articles/s41586-019-1238-8

This always get updated

Albert Vilella on Twitter
NGS #History https://t.co/S8su6pldeS
https://twitter.com/AlbertVilella/status/1226102101173927938

Markdown and notebook ; Reproducible and redistributable



R demo

Here is some code which illustrates some of the differences between R and S graphics capabilities. Note that colors are generally specified by a character string name (taken from the X11 rgb.txt file) and that line textures are given similarly. The parameter "bg" sets the background parameter for the plot and there is also an "fg" parameter which sets the foreground color.

In [1]:
```
require(datasets)

require(grDevices); require(graphics)
```

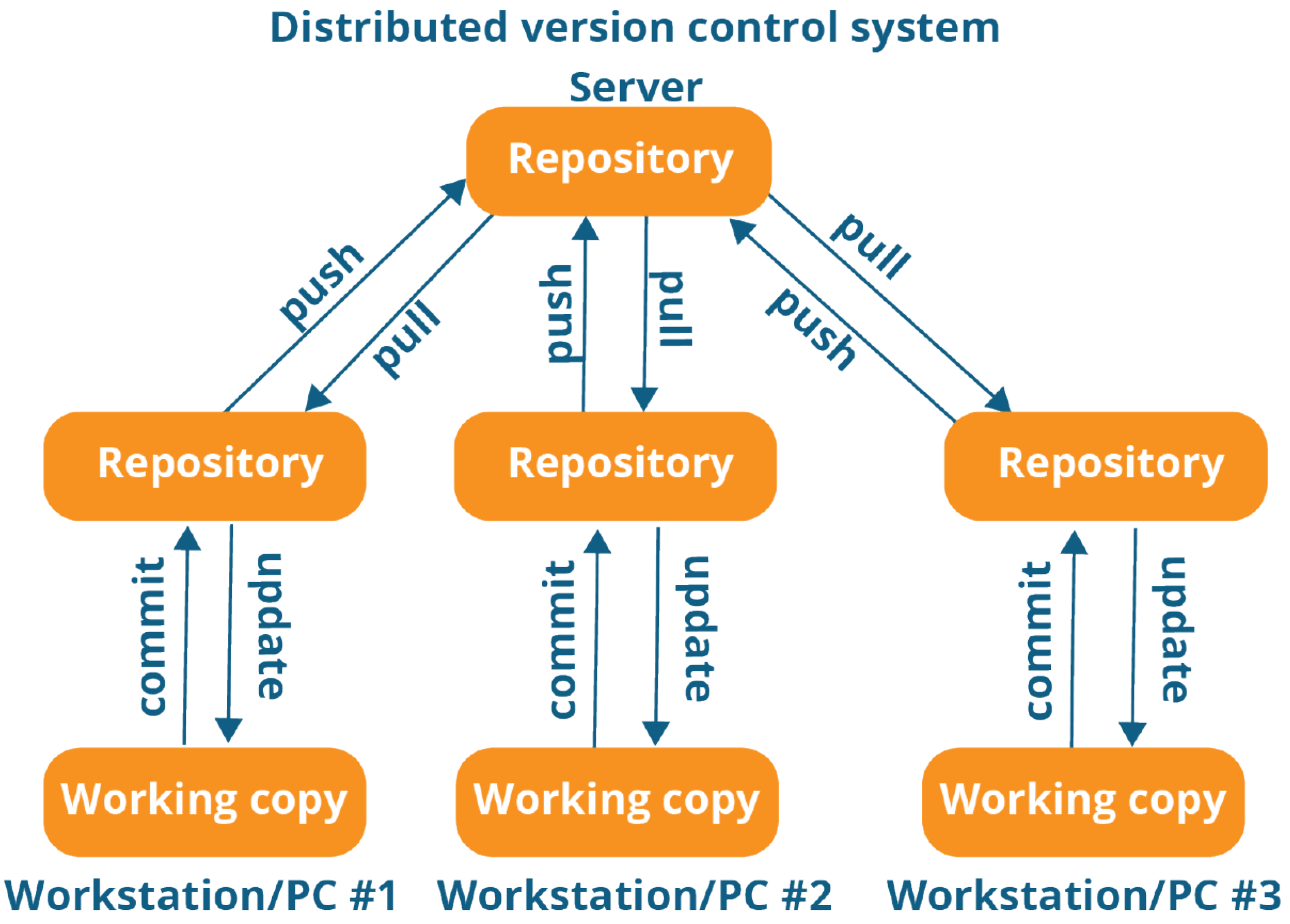
In [1]:
```
x <- stats::rnorm(50)
opar <- par(bg = "white")
plot(x, ann = FALSE, type = "n") +
abline(h = 0, col = gray(.90)) +
lines(x, col = "green4", lty = "dotted") +
points(x, bg = "limegreen", pch = 21) +
title(main = "Simple Use of Color In a Plot",
       xlab = "Just a Whisper of a Label",
       col.main = "blue", col.lab = gray(.8),
       cex.main = 1.2, cex.lab = 1.0, font.main = 4, font.lab = 3)
```

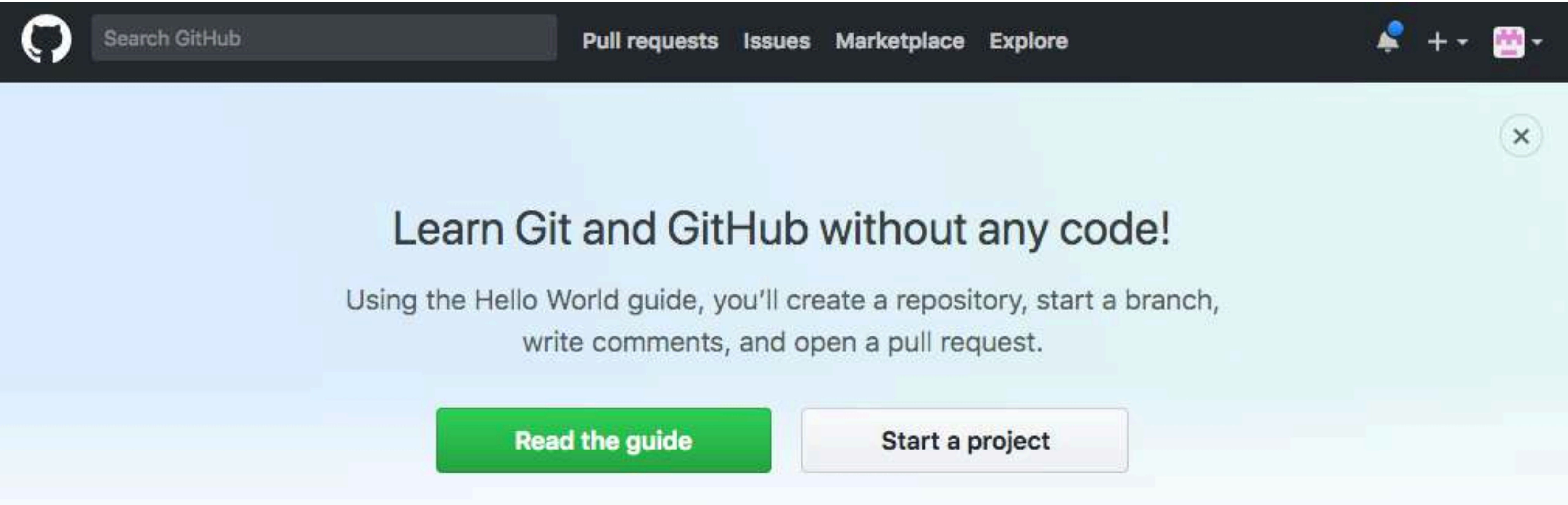
Simple Use of Color In a Plot

https://try.jupyter.org/

Version control: Git

git --fast-version-control





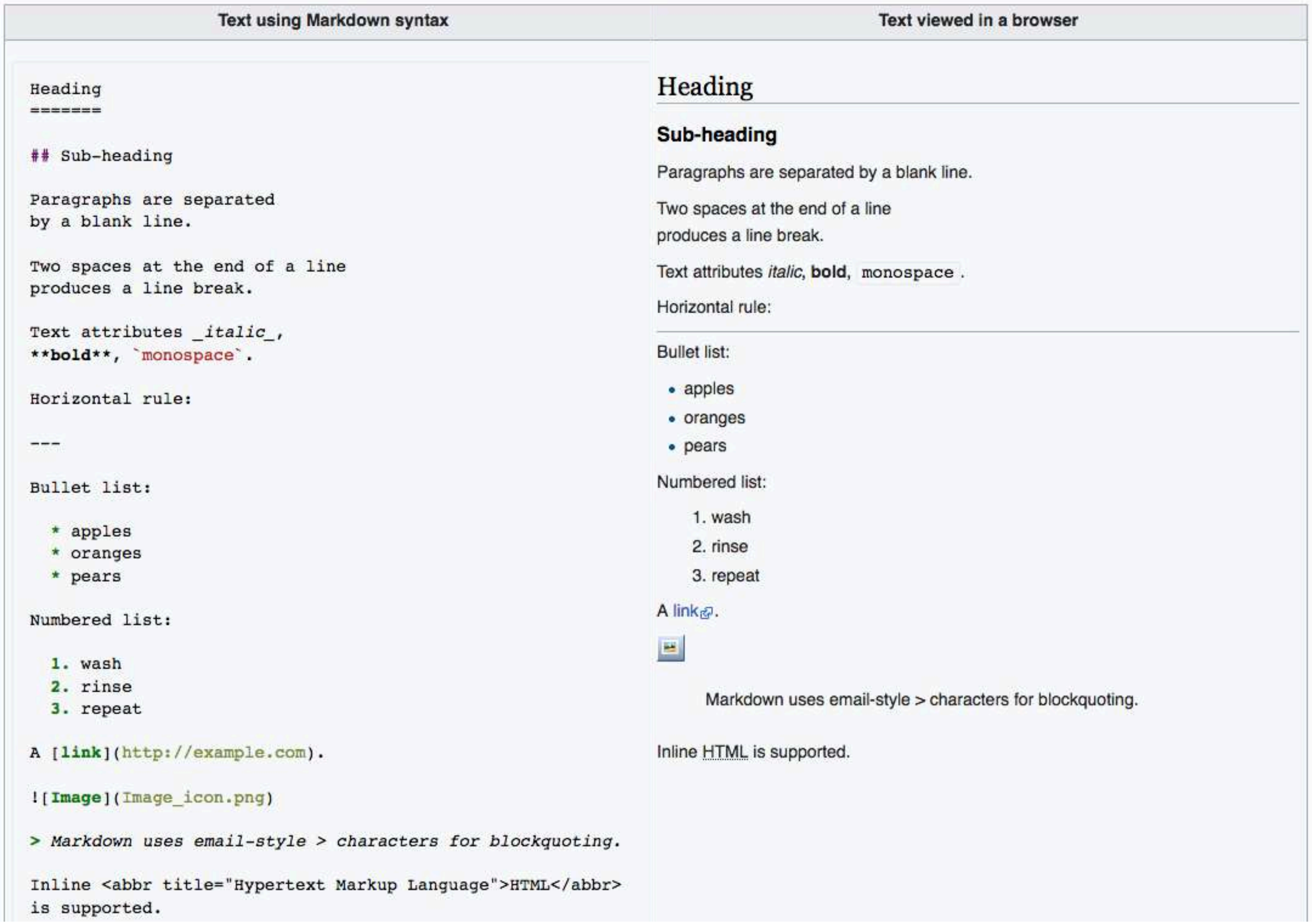
Learn Git and GitHub without any code!

Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request.

https://guides.github.com/activities/hello-world/

Use of markdown

- Created by John Gruber
- Informal plain-text formatting language
- Converts readable text to valid (X)HTML
- **Primary goal - readability**



Text using Markdown syntax

```
Heading
=======

## Sub-heading

Paragraphs are separated
by a blank line.

Two spaces at the end of a line
produces a line break.

Text attributes _italic_,
**bold**, `monospace`.

Horizontal rule:

---

Bullet list:

  * apples
  * oranges
  * pears

Numbered list:

  1. wash
  2. rinse
  3. repeat

A [link](http://example.com).

![Image](Image_icon.png)

> Markdown uses email-style > characters for blockquoting.

Inline <abbr title="Hypertext Markup Language">HTML</abbr>
is supported.
```

Text viewed in a browser

Heading

Sub-heading

Paragraphs are separated by a blank line.

Two spaces at the end of a line
produces a line break.

Text attributes *italic*, **bold**, `monospace`.

Horizontal rule:

Bullet list:

- apples
- oranges
- pears

Numbered list:

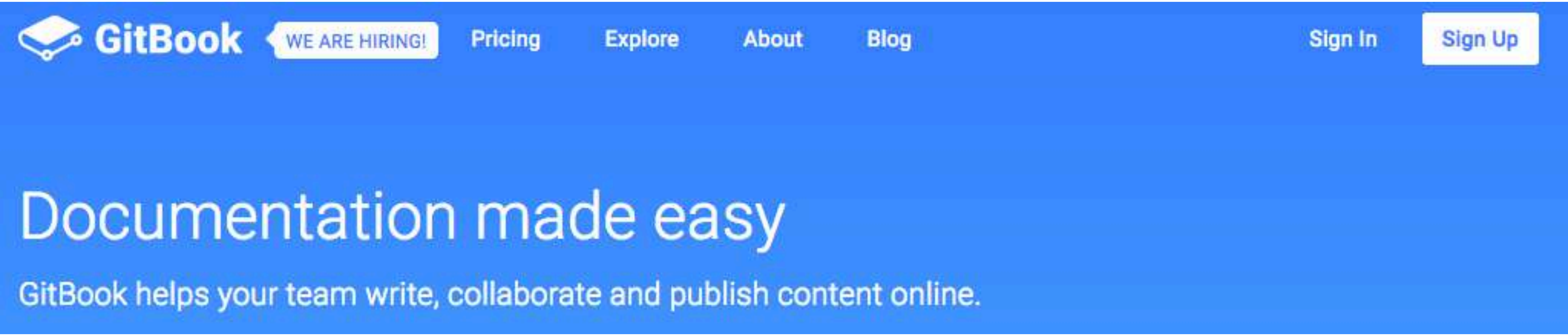
1. wash
2. rinse
3. repeat

A link.

> Markdown uses email-style > characters for blockquoting.

Inline HTML is supported.

Git + Github + markdown



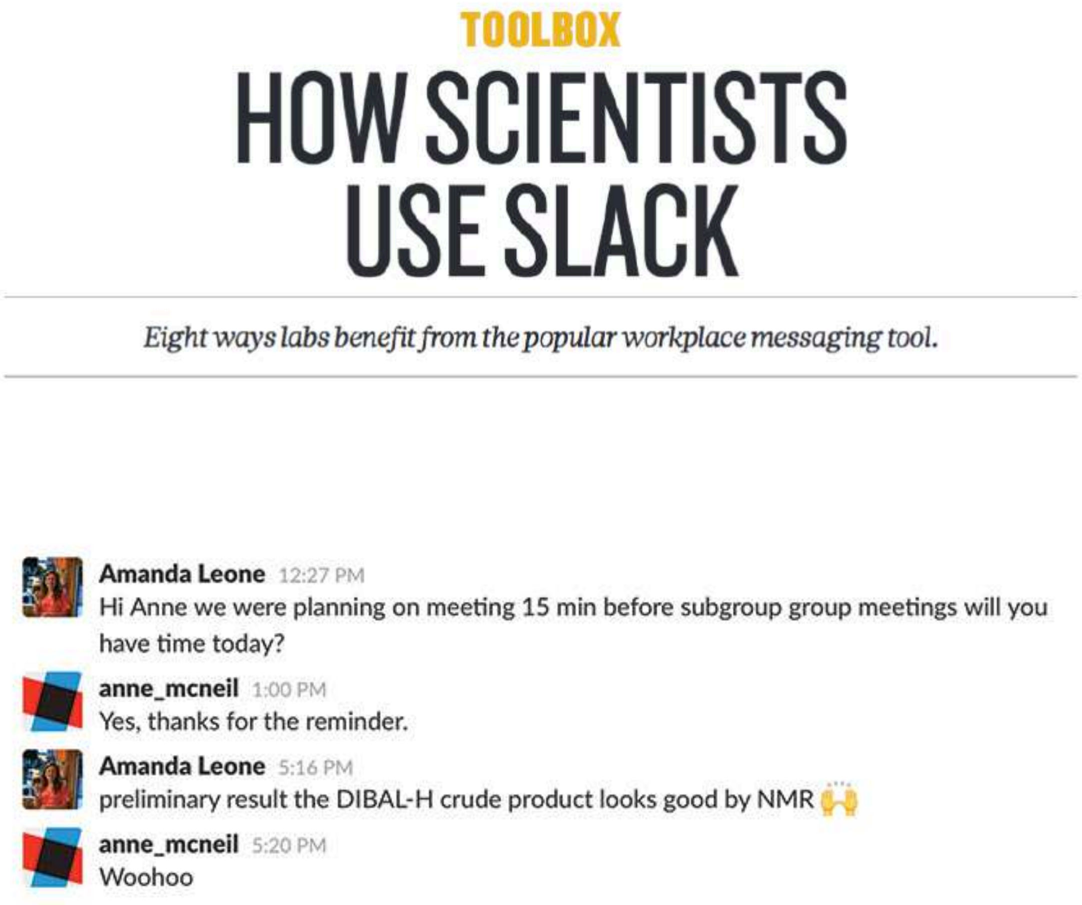
Documentation made easy

GitBook helps your team write, collaborate and publish content online.

Some examples:

- https://cgsb.gitbooks.io/ngs-analysis/content/
- https://pfern.github.io/OSODOS/gitbook/

Lab communication (fb, LINE?; SLACK)



TOOLBOX

HOW SCIENTISTS USE SLACK

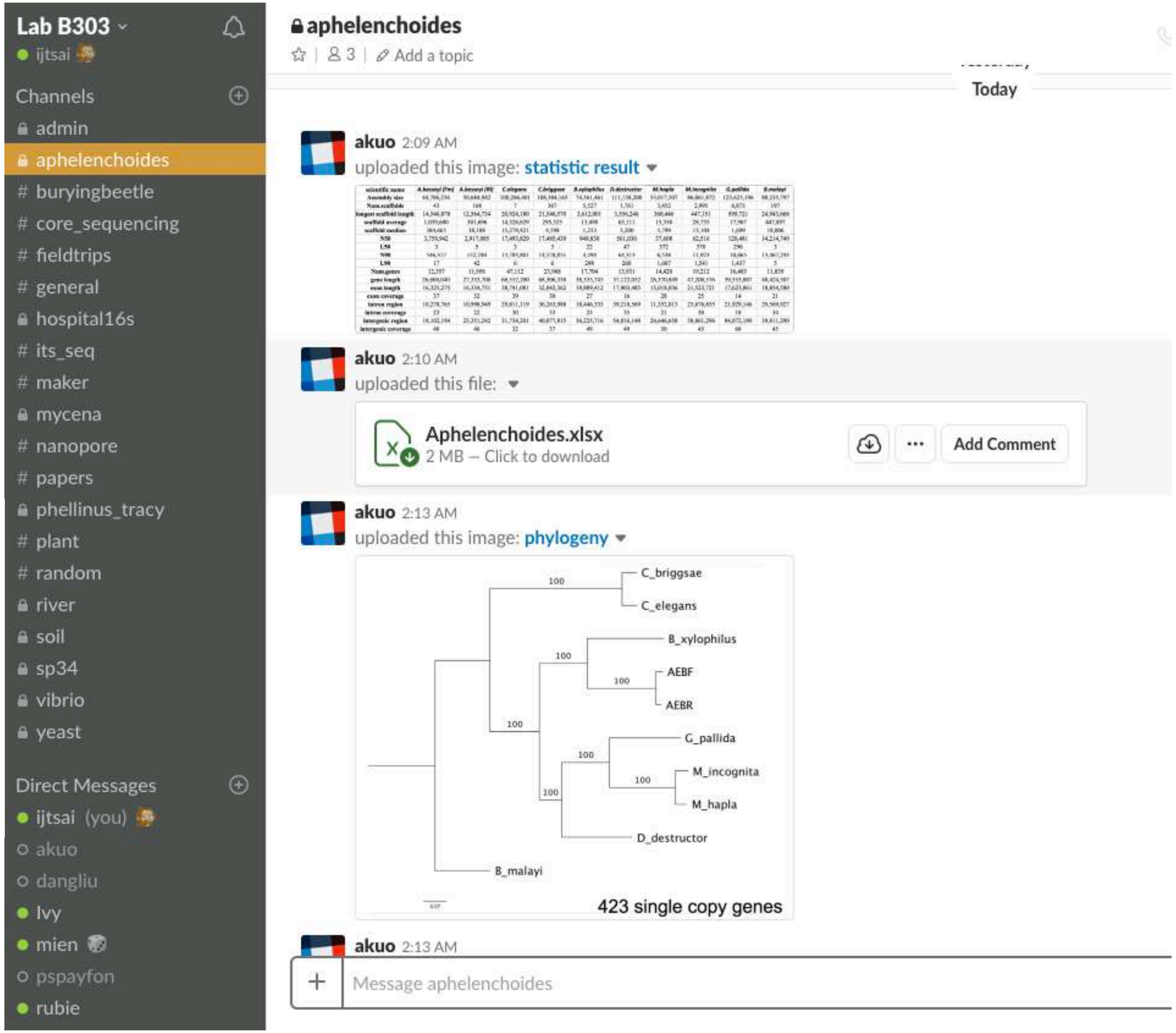
Eight ways labs benefit from the popular workplace messaging tool.

Amanda Leone 12:27 PM
Hi Anne we were planning on meeting 15 min before subgroup group meetings will you have time today?

anne_mcneil 1:00 PM
Yes, thanks for the reminder.

Amanda Leone 5:16 PM
preliminary result the DIBAL-H crude product looks good by NMR 🙌

anne_mcneil 5:20 PM
Woohoo



akuo 2:09 AM
uploaded this image: statistic result

scientific name	A.besseyi (Tm)	A.besseyi (Rl)	C.elegans	C.briggsae	B.xylophilus	D.destructor	M.hapla	M.incognita	G.pallida	B.malayi
Assembly size	44,706,234	50,608,942	100,286,401	108,384,165	74,561,461	111,138,208	53,017,507	86,061,872	123,625,196	88,235,797
Num.scaffolds	43	169	7	367	5,527	1,761	3,452	2,995	6,873	197
longest scaffold length	14,346,878	12,564,734	20,924,180	21,540,578	2,612,001	3,556,246	360,446	447,151	599,721	24,943,668
scaffold average	1,039,680	301,496	14,326,629	295,325	13,490	63,111	15,358	28,735	17,987	447,897
scaffold median	389,663	38,188	15,279,421	4,598	1,253	3,200	5,799	13,108	1,699	18,806
N50	3,759,942	2,917,085	17,493,829	17,485,439	949,830	561,030	37,608	62,516	128,481	14,214,749
L50	5	5	3	5	22	47	372	378	296	3
N90	546,917	112,284	13,783,901	14,578,814	4,298	65,313	6,548	11,929	18,665	13,867,238
L90	17	42	6	6	298	268	1,607	1,541	1,437	5
Num.genes	12,357	11,986	47,112	23,988	17,704	13,931	14,420	19,212	16,465	11,839
gene length	26,604,081	27,153,708	68,512,280	69,396,358	58,535,705	57,272,012	26,370,809	47,208,576	99,553,807	48,424,587
exon length	16,325,275	16,334,731	38,741,081	32,643,362	18,089,612	17,903,483	15,019,036	21,523,721	17,623,861	18,854,580
exon coverage	37	32	39	30	27	16	28	25	14	21
intron region	10,278,765	10,998,949	29,811,119	36,263,988	18,446,333	39,218,569	11,352,813	23,878,855	21,929,146	29,569,927
intron coverage	23	22	30	33	25	35	21	30	18	34
intergenic region	18,102,194	23,351,242	31,734,281	40,677,815	16,225,716	54,016,148	26,646,658	38,861,296	84,072,189	38,811,280
intergenic coverage	40	46	32	37	49	49	50	45	68	45

akuo 2:10 AM
uploaded this file:

Aphelenchoides.xlsx
2 MB – Click to download

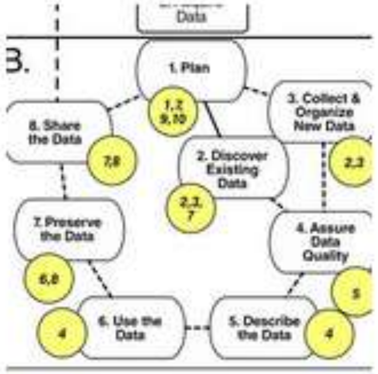
akuo 2:13 AM
uploaded this image: phylogeny

423 single copy genes

akuo 2:13 AM

https://cen.acs.org/articles/94/i29/Slack-ing-helps-chemists-manage.html

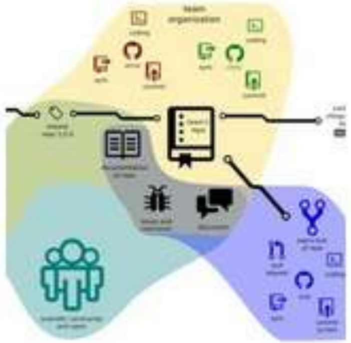
Ten simple rules series



Ten Simple Rules for Creating a Good Data Management Plan

William Michener

PLOS Computational Biology: 22 Oct 2015



Ten Simple Rules for Taking Advantage of Git and GitHub

Yasset Perez-Riverol, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe da Veiga Leprevost, …

PLOS Computational Biology: 14 Jul 2016

Ten Simple Rules for a Computationa Biologist's Laboratory Notebook

Santiago Schnell

PLOS Computational Biology: 10 Sep 2015

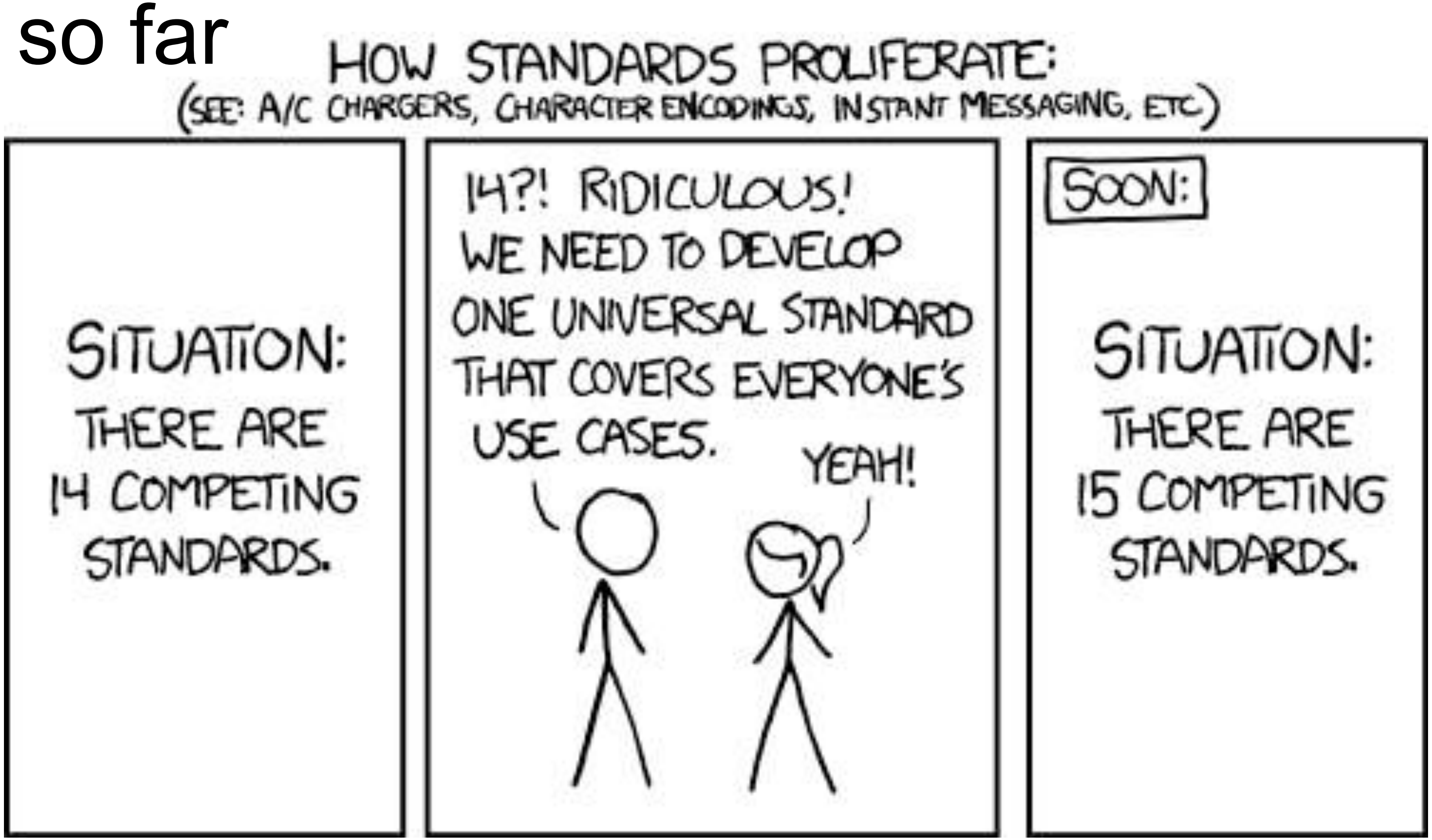
Ten simple rules for biologists learning to program

Maureen A. Carey, Jason A. Papin

PLOS Computational Biology: 04 Jan 2018

http://collections.plos.org/ten-simple-rules

Summary so far


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

https://xkcd.com/927/

- No need to do everything ‘perfect’
- Depending on scale, use something that is most effective

Useful links:

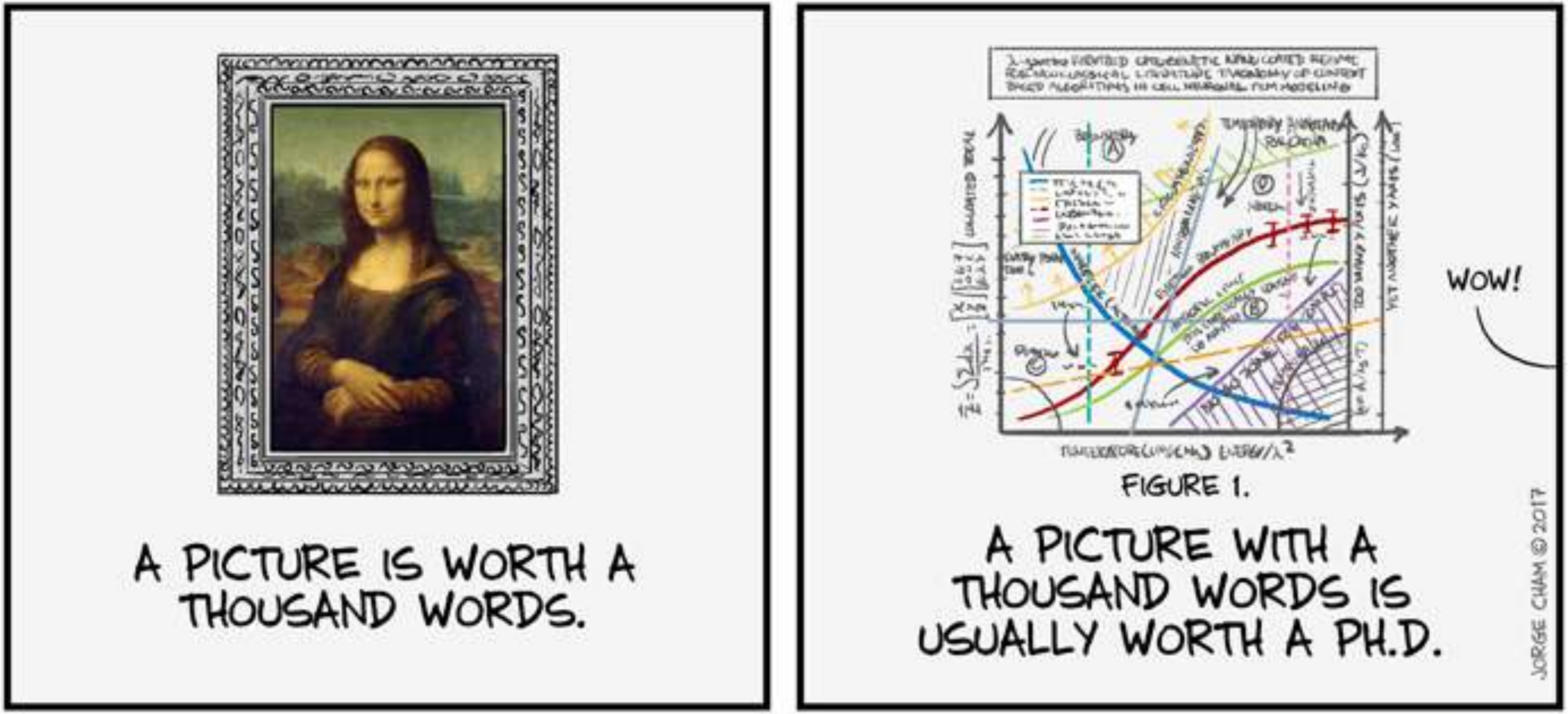
A series of Jypyter notebooks hosted on github

- https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks

Other links

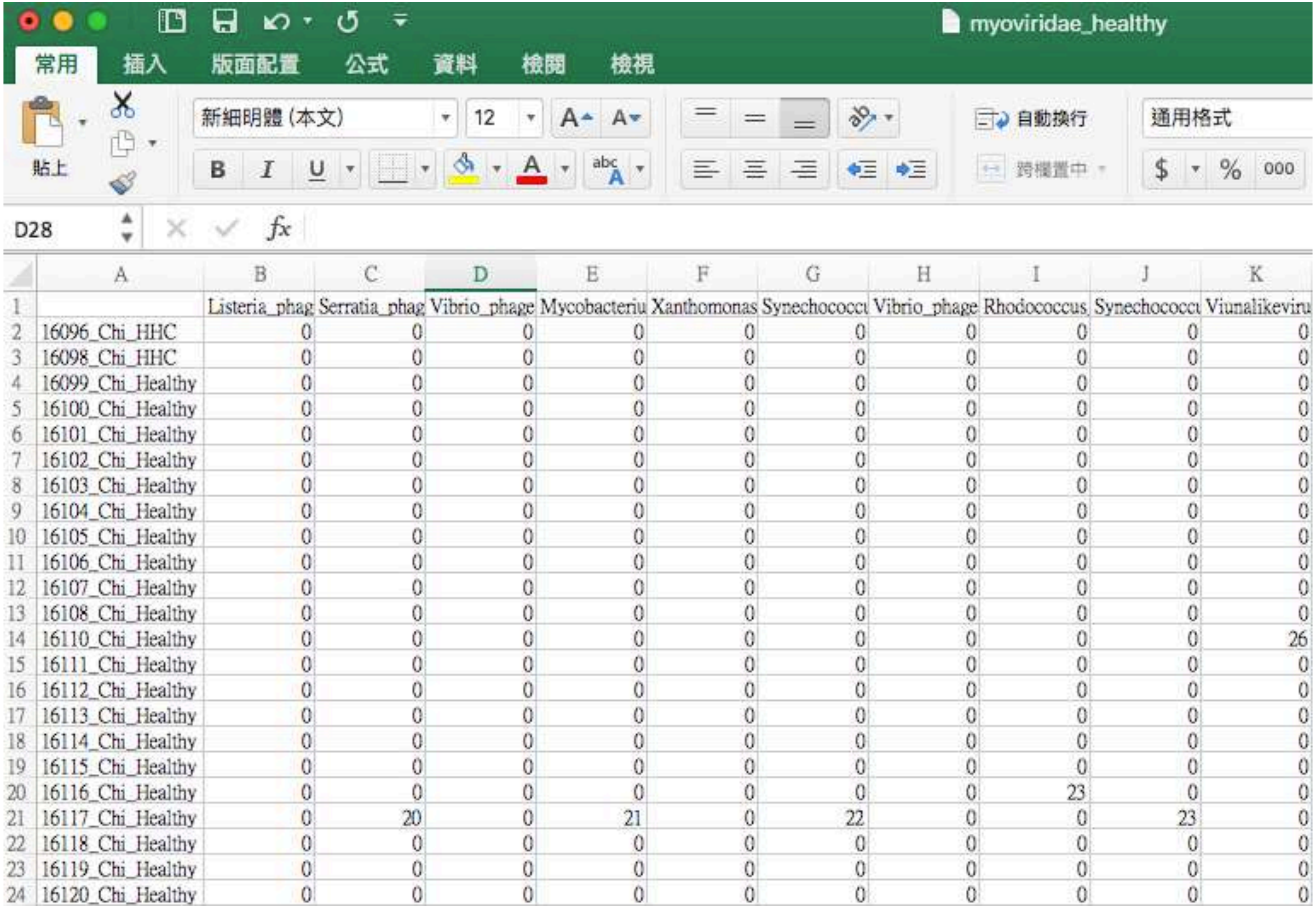
- http://linux.vbird.org/linux_basic/ (Chinese ; extremely useful) ****
- https://evomics.org/learning/unix-tutorial/
- http://www.ark-genomics.org/events-online-training-eu-training-course/introduction-linux
- http://linuxcommand.org/

Data type / Visualisations

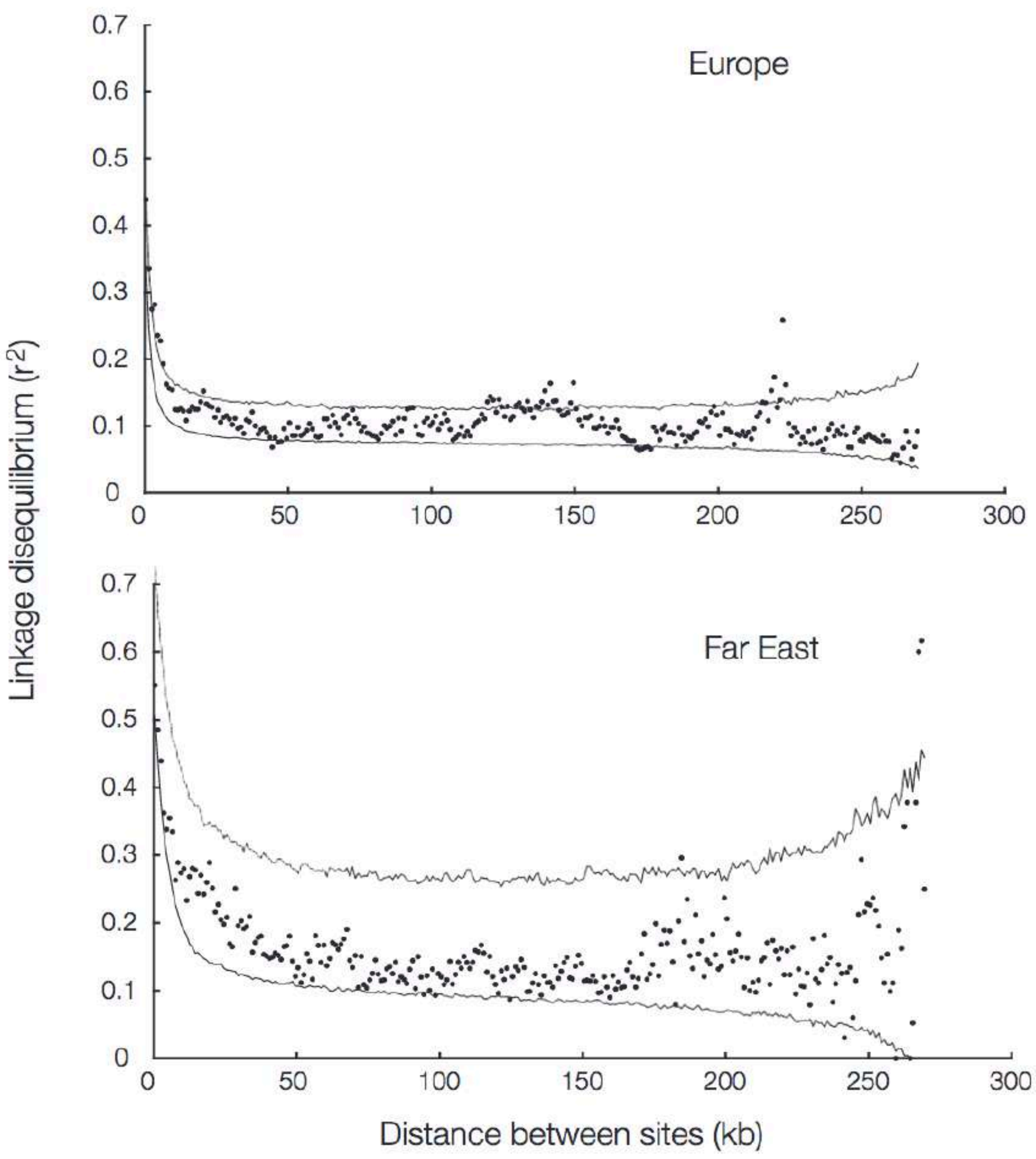


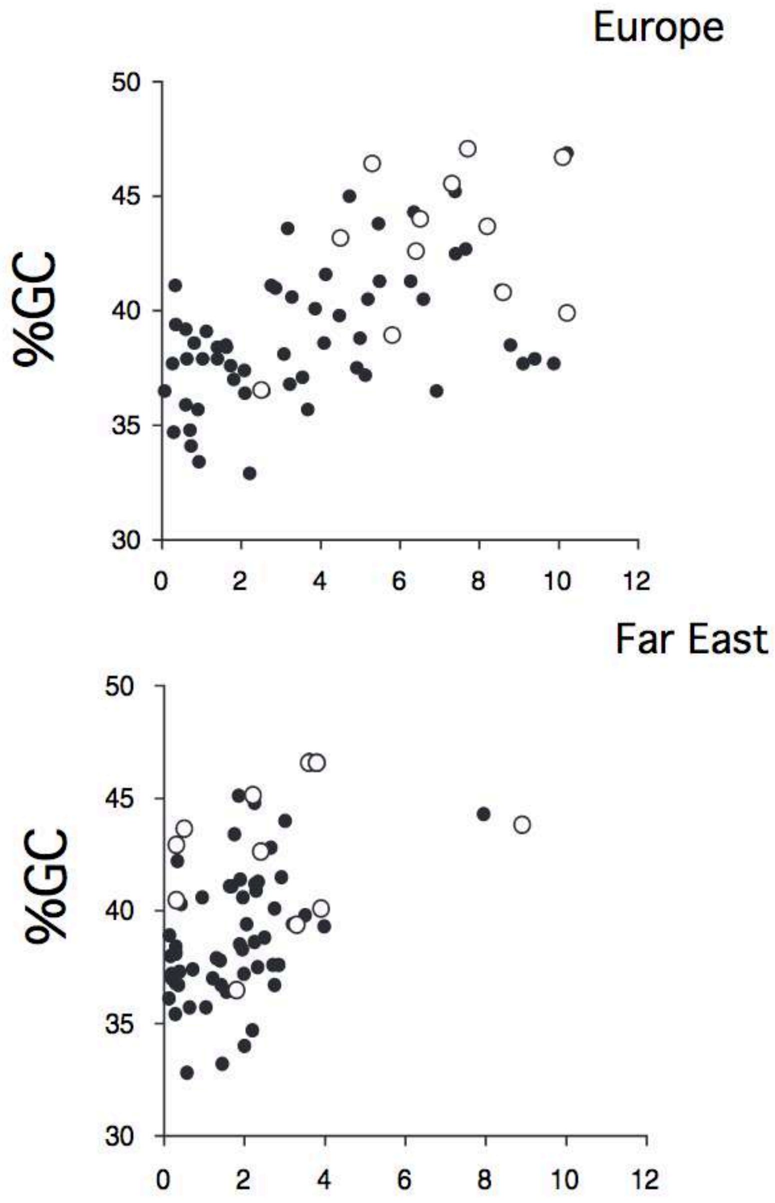
A PICTURE IS WORTH A THOUSAND WORDS.
FIGURE 1.
WOW!
A PICTURE WITH A THOUSAND WORDS IS USUALLY WORTH A PH.D.
JORGE CHAM © 2017

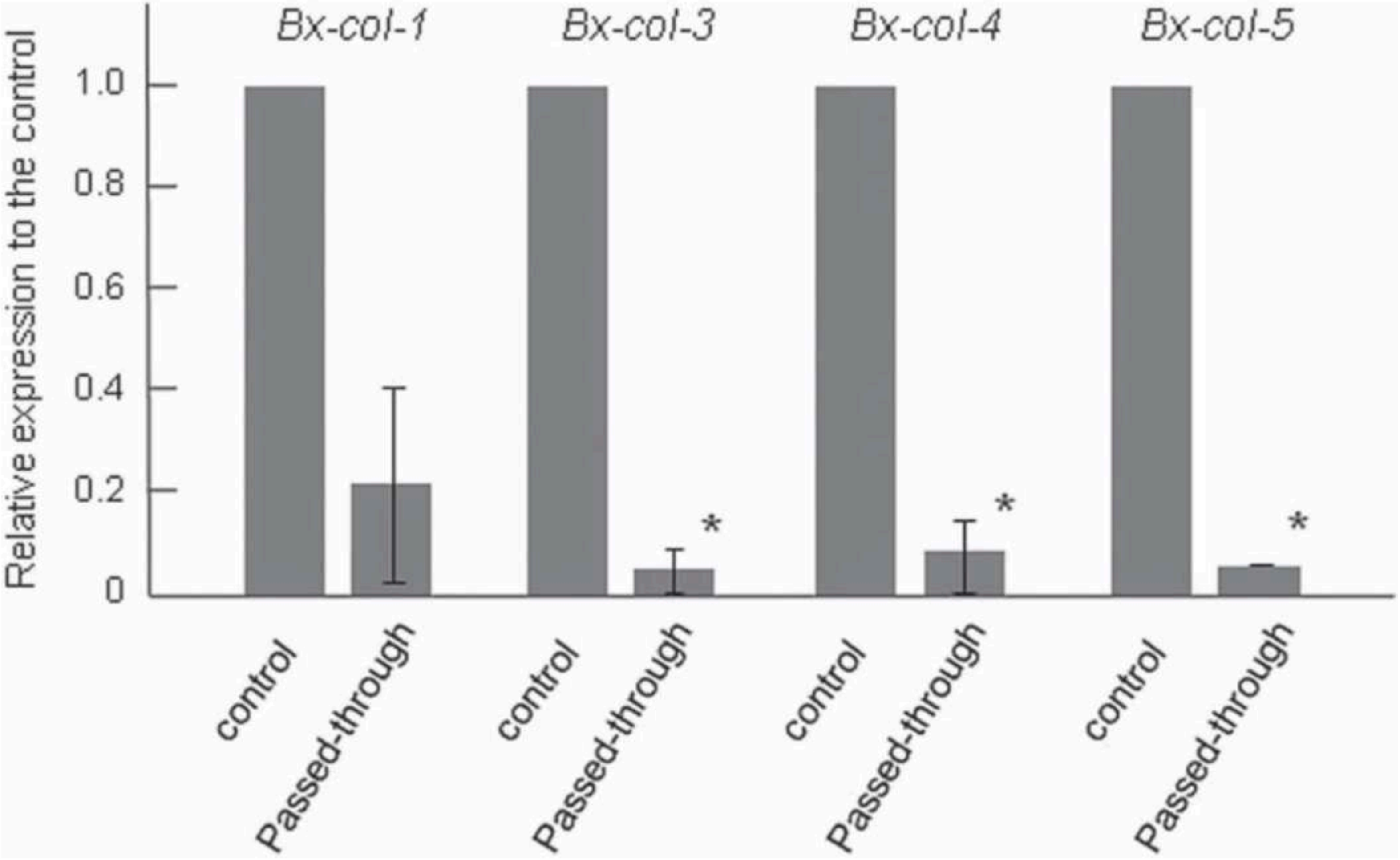
WWW.PHDCOMICS.COM

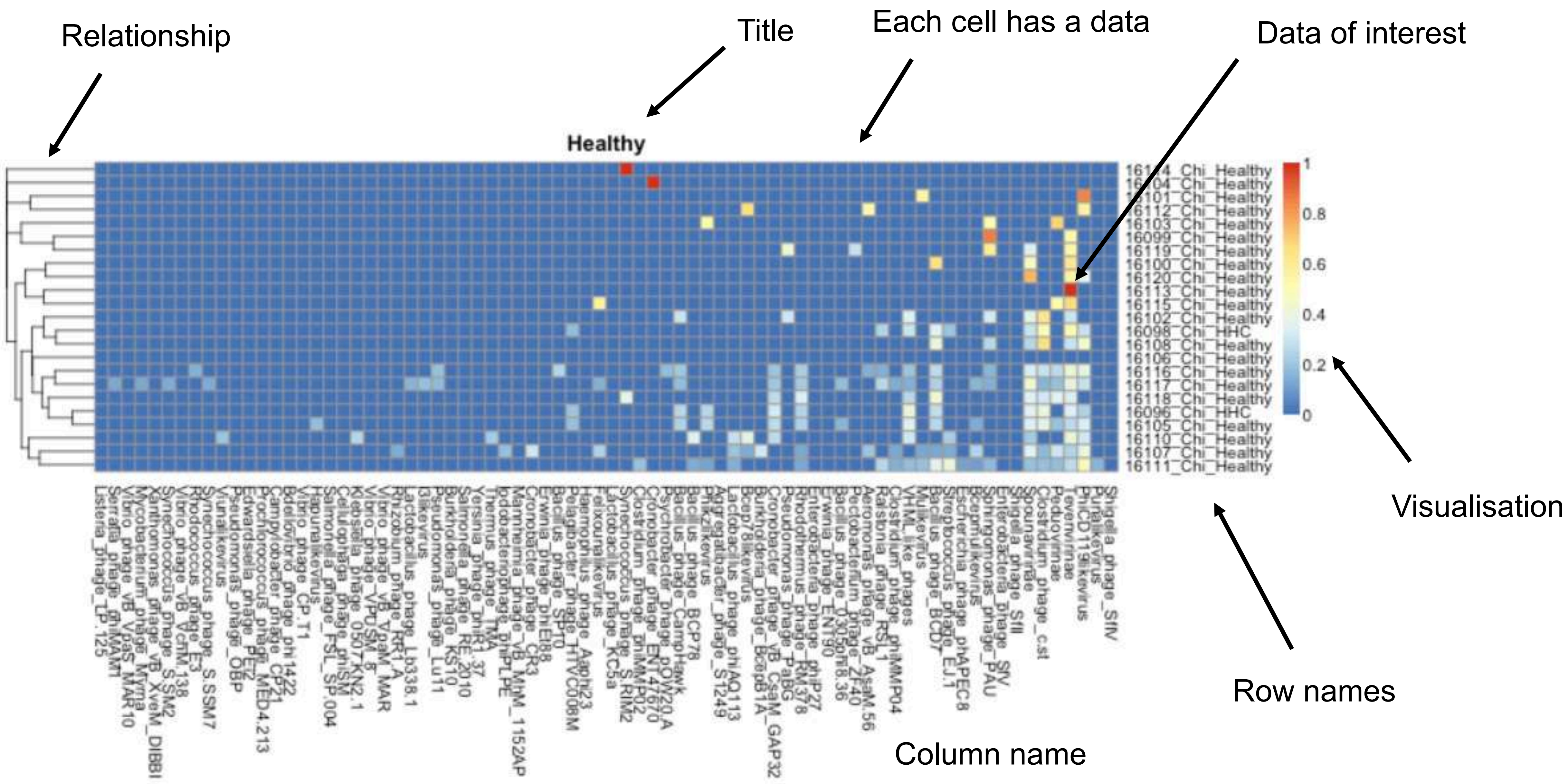


	A	B	C	D	E	F	G	H	I	J	K
1		Listeria_phag	Serratia_phag	Vibrio_phage	Mycobacteriu	Xanthomonas	Synechococci	Vibrio_phage	Rhodococcus	Synechococci	Viunalikeviru
2	16096_Chi_HHC	0	0	0	0	0	0	0	0	0	0
3	16098_Chi_HHC	0	0	0	0	0	0	0	0	0	0
4	16099_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
5	16100_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
6	16101_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
7	16102_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
8	16103_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
9	16104_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
10	16105_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
11	16106_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
12	16107_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
13	16108_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
14	16110_Chi_Healthy	0	0	0	0	0	0	0	0	0	26
15	16111_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
16	16112_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
17	16113_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
18	16114_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
19	16115_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
20	16116_Chi_Healthy	0	0	0	0	0	0	0	23	0	0
21	16117_Chi_Healthy	0	20	0	21	0	22	0	0	23	0
22	16118_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
23	16119_Chi_Healthy	0	0	0	0	0	0	0	0	0	0
24	16120_Chi_Healthy	0	0	0	0	0	0	0	0	0	0


Europe
Far East
Linkage disequilibrium (r^2)
Distance between sites (kb)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300


Europe
Far East
%GC
50
45
40
35
30
0
2
4
6
8
10
12


Bx-col-1
Bx-col-3
Bx-col-4
Bx-col-5
Relative expression to the control
1.0
0.8
0.6
0.4
0.2
0
*
*
*
control
Passed-through
control
Passed-through
control
Passed-through
control
Passed-through



Relationship

Title

Each cell has a data

Data of interest

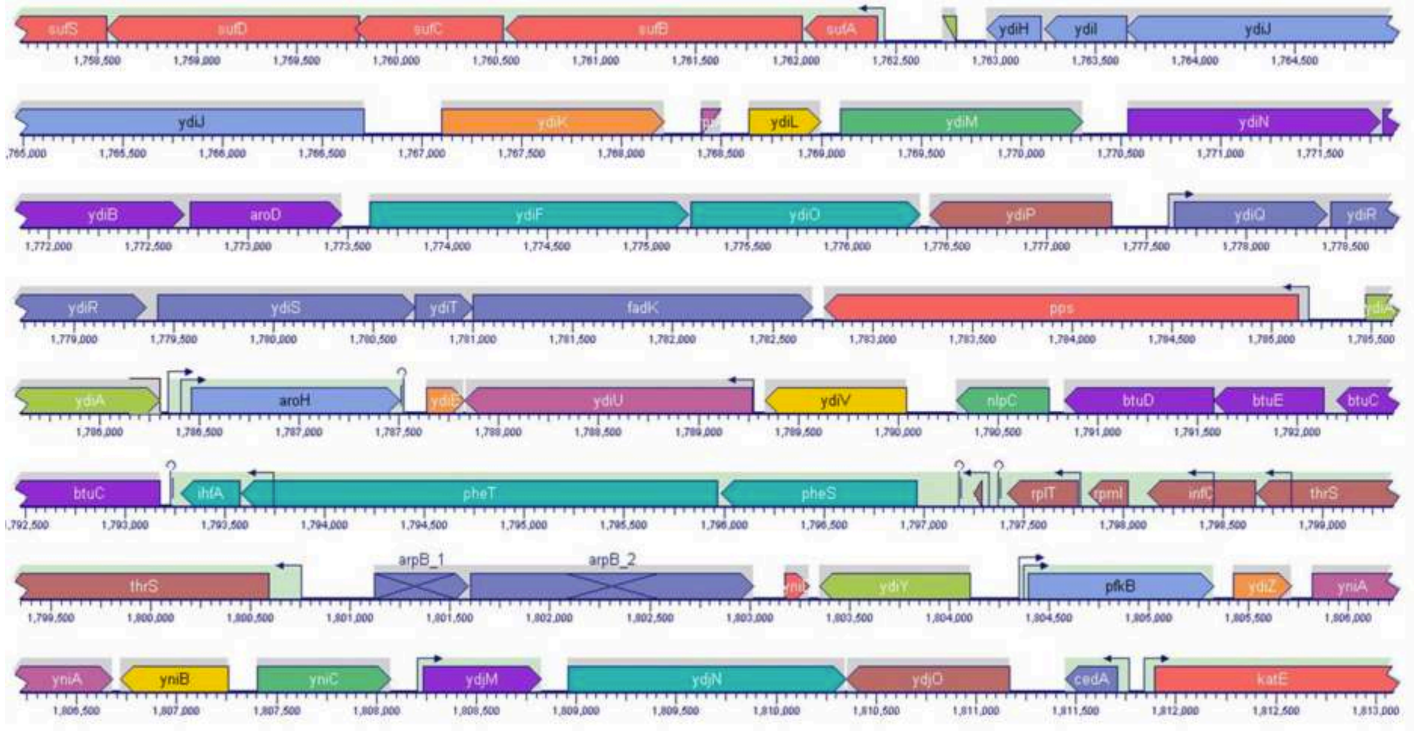
Visualisation

Row names

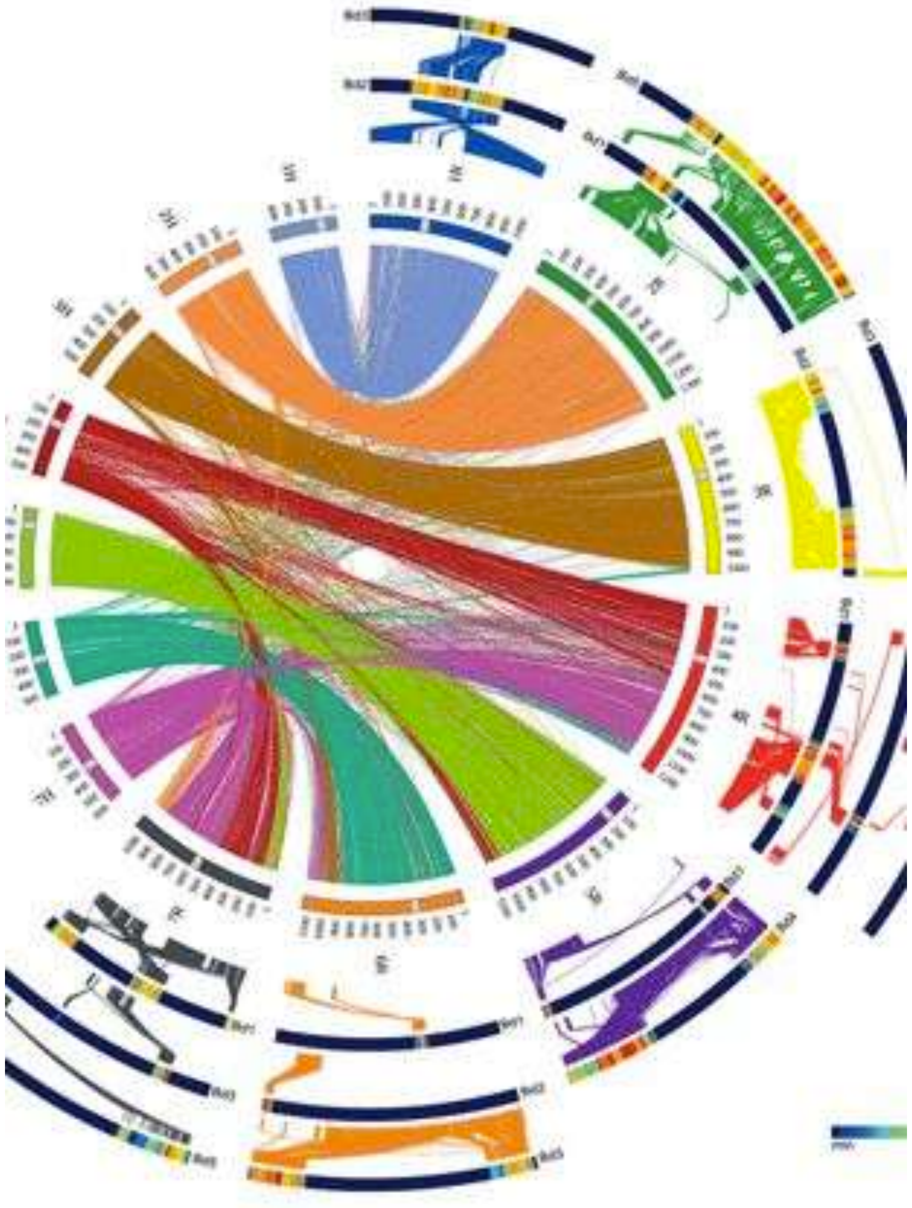
Column name

Locations / maps

- How do we represent/visualise them?

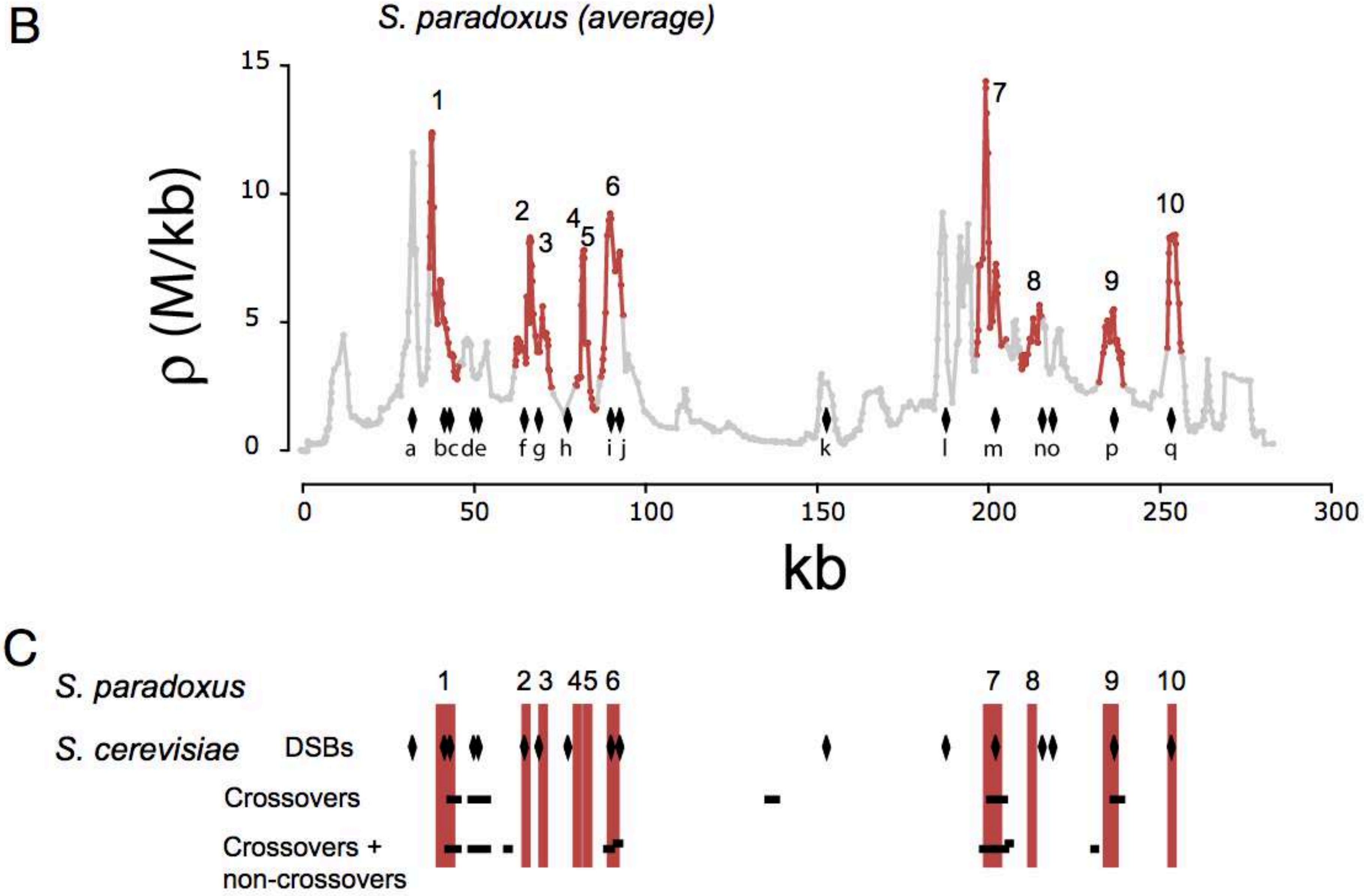


Gene locations / strand

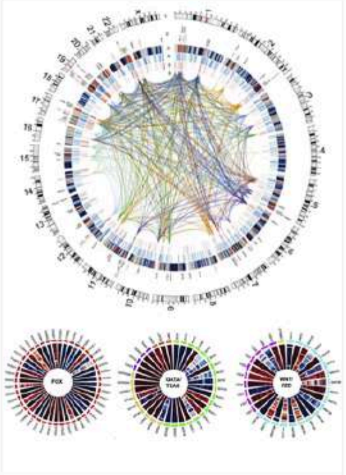


Circos

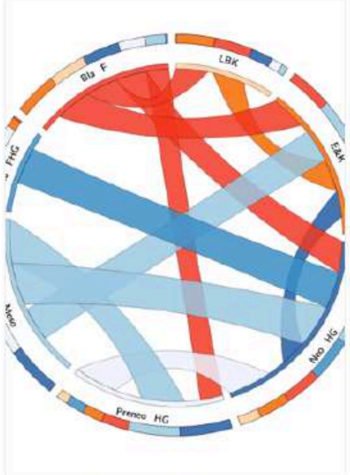
Properties on the genome



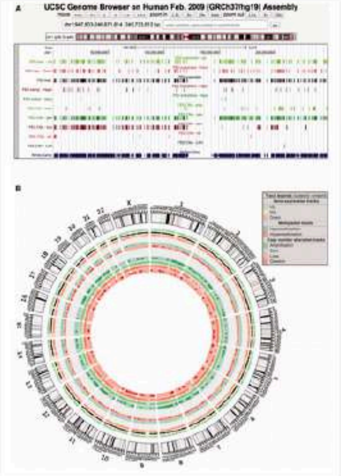
Visualising genomes - Circos



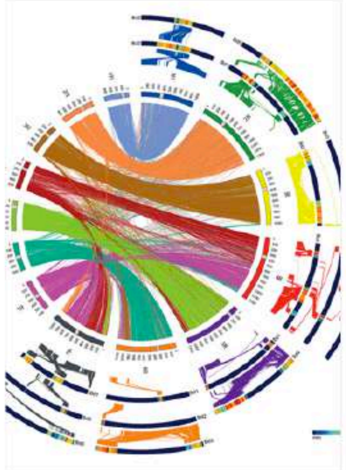
▲ 1 · 1 Dec 2013 | Saben J, Zhong Y, McKelvey S *et al.* (2014) A comprehensive analysis of the human placenta transcriptome *Placenta* **35**:125-131.



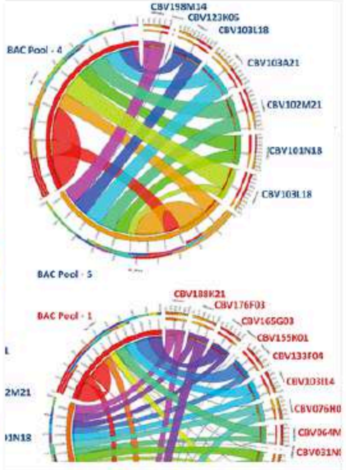
▲ 2 · 25 Oct 2013 | Bollongino R, Nehlich O, Richards MP *et al.* (2013) 2000 years of parallel societies in Stone Age Central Europe *Science* **342**:479-481.


A
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
B

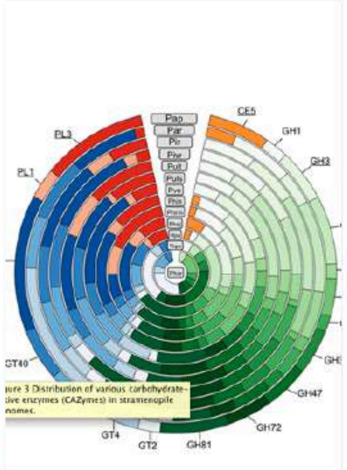
▲ 3 · 25 Oct 2013 | Dayem Ullah AZ, Cutts RJ, Ghetia M *et al.* (2013) The pancreatic expression database: recent extensions and updates *Nucleic Acids Res*



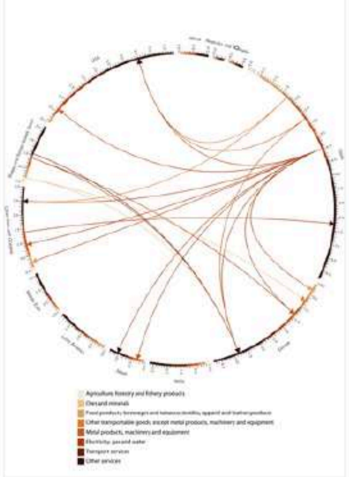
▲ 13 · 8 Oct 2013 | Martis MM, Zhou R, Haseneyer G *et al.* (2013) Reticulate Evolution of the Rye Genome *Plant Cell*



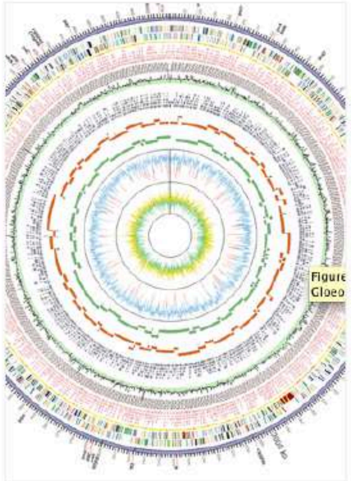
▲ 14 · 8 Oct 2013 | Buyyarapu R, Kantety RV, Yu JZ *et al.* (2013) BAC-Pool Sequencing and Analysis of Large Segments of A12 and D12 Homoeologous Chromosomes in Upland Cotton *PLoS One* 8:e76757.



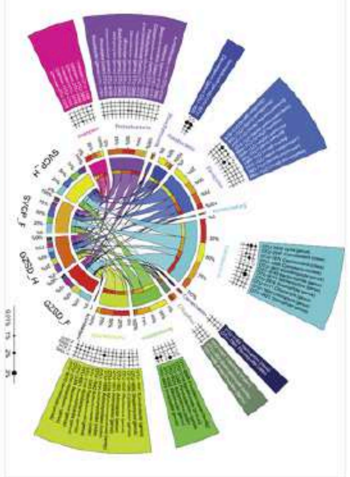
▲ 15 · 4 Oct 2013 | Adhikari BN, Hamilton JP, Zerillo MM *et al.* (2013) Comparative Genomics Reveals Insight into Virulence Strategies of Plant Pathogenic Oomycetes *PLoS One* 8:e75072.



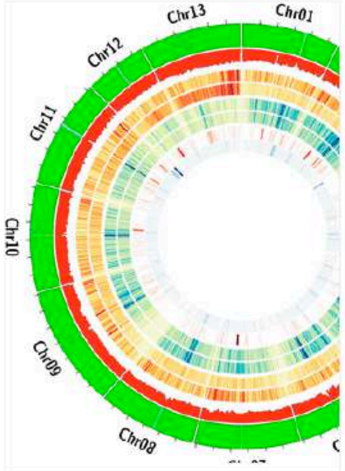
▲ 4 · 23 Oct 2013 | Kanemoto K, Moran D, Lenzen M *et al.* (2013) International trade undermines national emission reduction targets: New evidence from air pollution *Global Environmental Change*



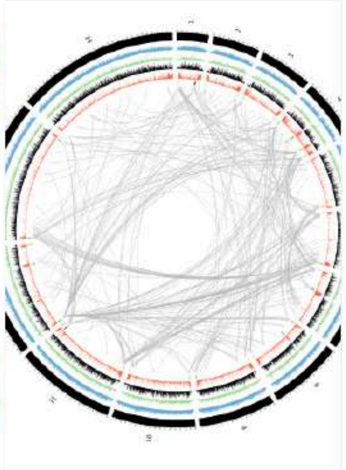
▲ 5 · 23 October 2013 | Saw JHW, Schatz M, Brown MV *et al.* (2013) Cultivation and Complete Genome Sequencing of Gloeobacter kilaueensis sp. nov., from a Lava Cave in Kilauea Caldera, Hawai'i *PLoS One* 8:e76376.



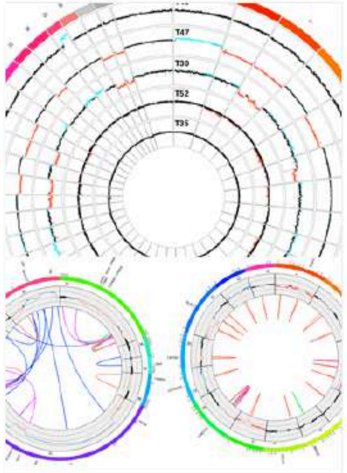
▲ 6 · 17 Oct 2013 | Ye L, Amberg J, Chapman D *et al.* (2013) Fish gut microbiota analysis differentiates physiology and behavior of invasive Asian carp and Indigenous American fish *The ISME Journal*



▲ 16 · 1 Oct 2013 | Page JT, Huynh MD, Liechty ZS *et al.* (2013) Insights into the Evolution of Cotton Diploids and Polyploids from Whole-Genome Re-sequencing *G3: Genes|Genomes|Genetics* **3**:1809-1818.



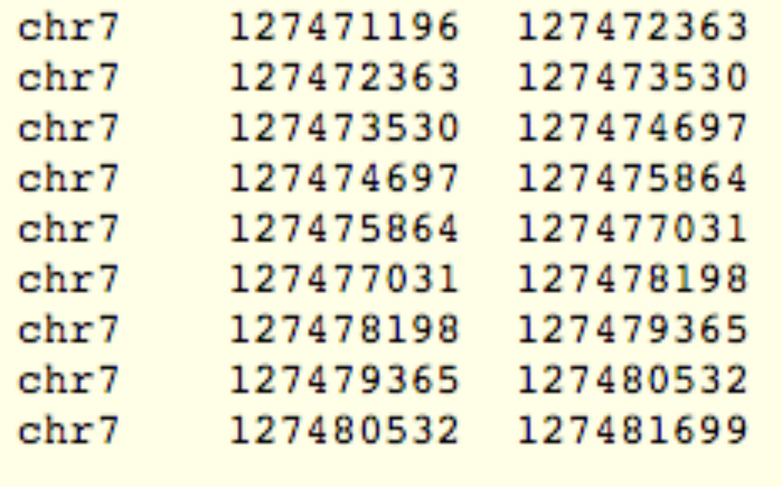
▲ 17 · 30 Sep 2013 | Lemieux JE, Kyes SA, Otto TD *et al.* (2013) Genome-wide profiling of chromosome interactions in Plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic variation *Molecular microbiology*



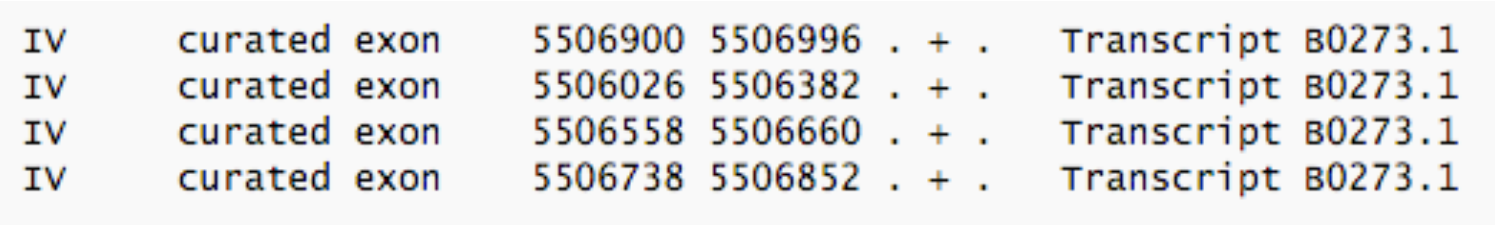
▲ 18 · 30 Sep 2013 | Beck J, Hennecke S, Bornemann-Kolatzki K *et al.* (2013) Genome Aberrations in Canine Mammary Carcinomas and Their Detection in Cell-Free Plasma DNA *PLoS One* 8:e75485.

BED/gff format

- Features on genome use bed / gff files to represent their locations
- “Optional field” can be added for additional information



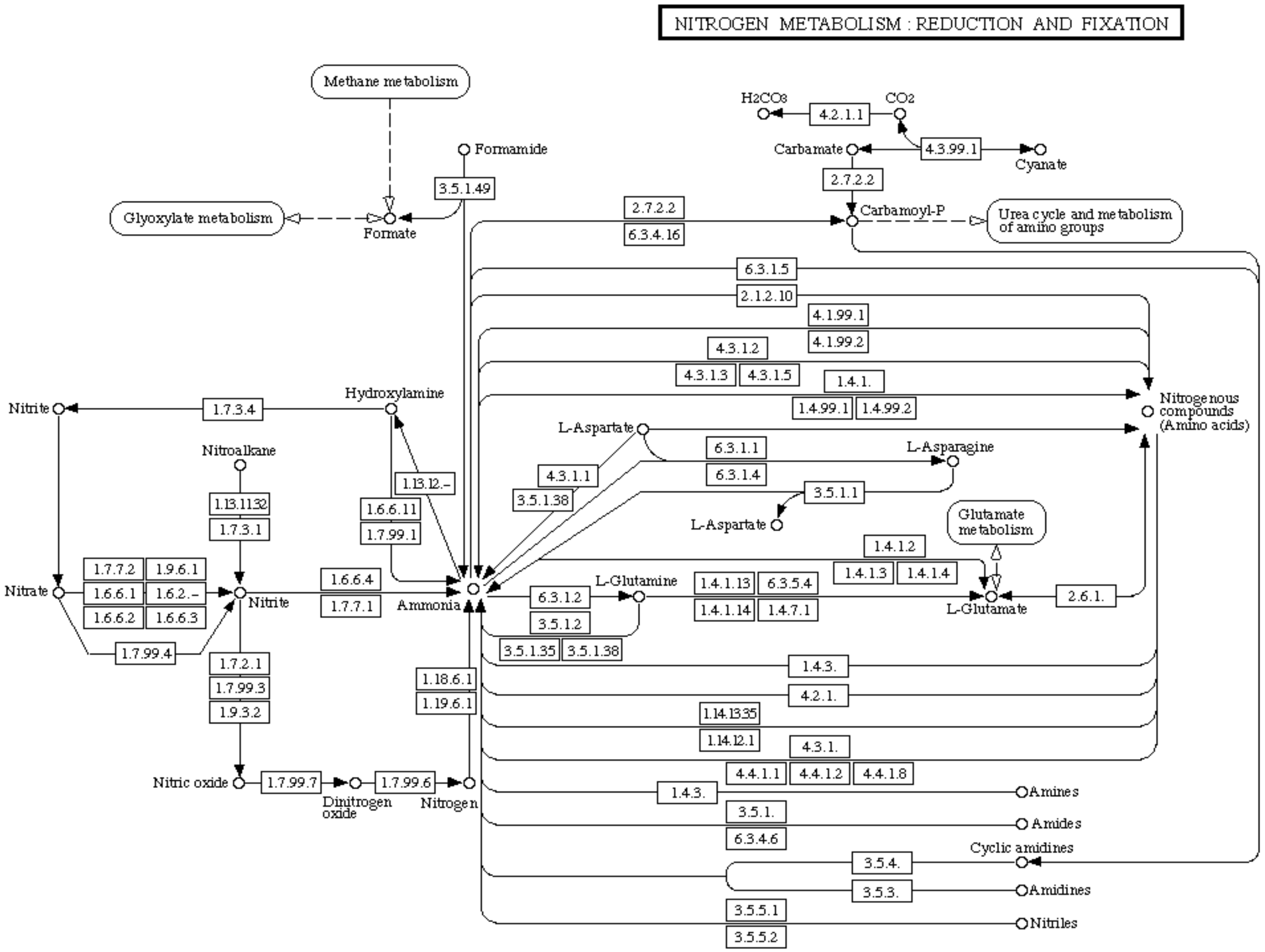
```
chr7    127471196  127472363
chr7    127472363  127473530
chr7    127473530  127474697
chr7    127474697  127475864
chr7    127475864  127477031
chr7    127477031  127478198
chr7    127478198  127479365
chr7    127479365  127480532
chr7    127480532  127481699
```

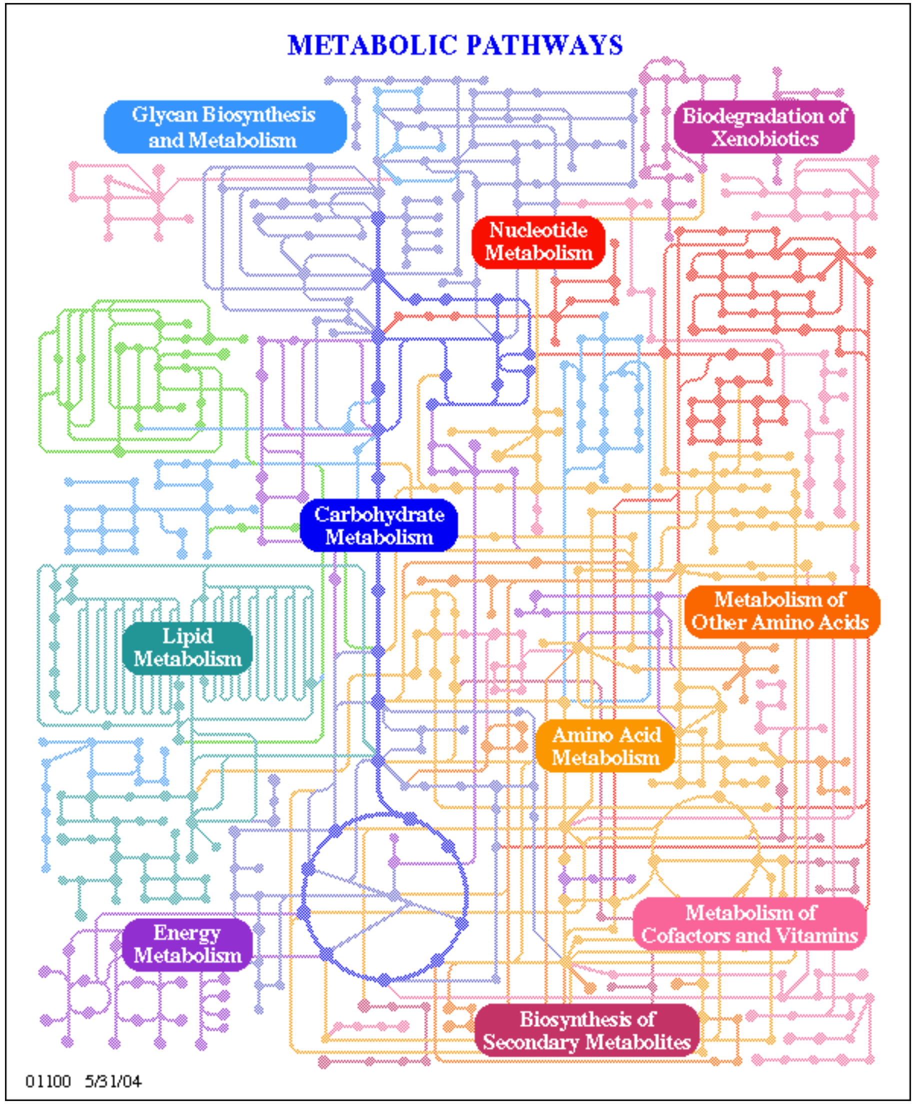


```
IV     curated exon    5506900 5506996 . + .   Transcript B0273.1
IV     curated exon    5506026 5506382 . + .   Transcript B0273.1
IV     curated exon    5506558 5506660 . + .   Transcript B0273.1
IV     curated exon    5506738 5506852 . + .   Transcript B0273.1
```

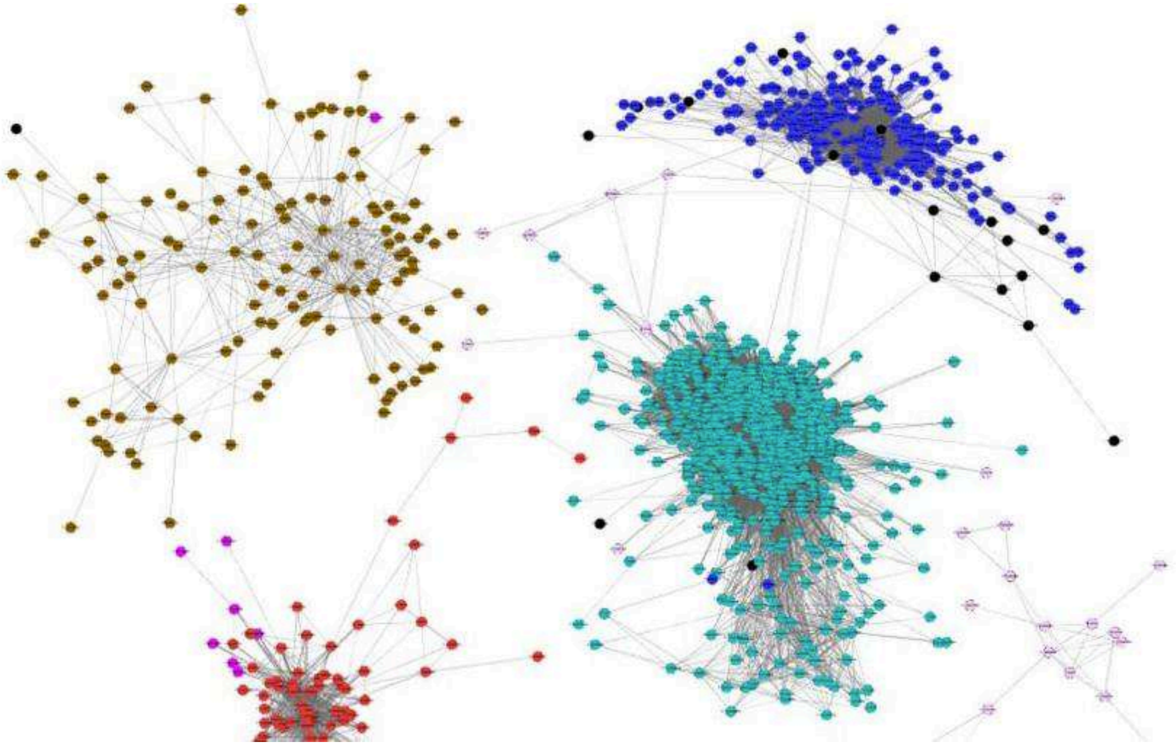
http://genome.ucsc.edu/FAQ/FAQformat#format1
http://gmod.org/wiki/GFF2

Pathways

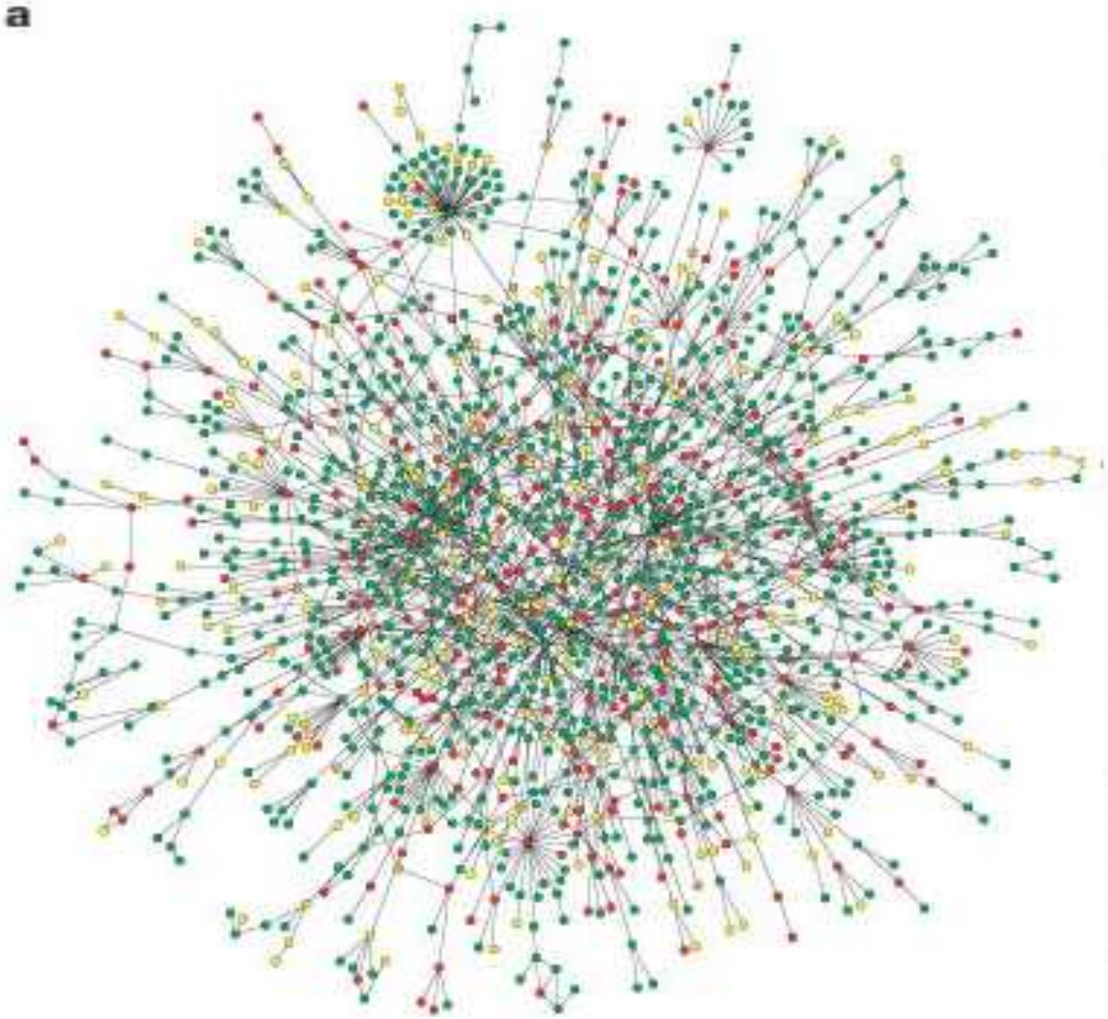

NITROGEN METABOLISM : REDUCTION AND FIXATION
Methane metabolism
Glyoxylate metabolism
Formate
Formamide
3.5.1.49
H2CO3
CO2
4.2.1.1
Carbamate
4.3.99.1
Cyanate
2.7.2.2
Carbamoyl-P
Urea cycle and metabolism of amino groups
2.7.2.2
6.3.4.16
6.3.1.5
2.1.2.10
4.1.99.1
4.1.99.2
4.3.1.2
4.3.1.3
4.3.1.5
1.4.1.
1.4.99.1
1.4.99.2
Nitrogenous compounds (Amino acids)
Nitrite
1.7.3.4
Hydroxylamine
L-Aspartate
6.3.1.1
6.3.1.4
L-Asparagine
4.3.1.1
3.5.1.38
3.5.1.1
L-Aspartate
Nitroalkane
1.13.11.32
1.7.3.1
1.13.12.-
1.6.6.11
1.7.99.1
Glutamate metabolism
1.4.1.2
1.4.1.3
1.4.1.4
Nitrate
1.7.7.2
1.9.6.1
1.6.6.1
1.6.2.-
1.6.6.2
1.6.6.3
Nitrite
1.6.6.4
1.7.7.1
Ammonia
6.3.1.2
L-Glutamine
1.4.1.13
6.3.5.4
1.4.1.14
1.4.7.1
L-Glutamate
2.6.1.
3.5.1.2
3.5.1.35
3.5.1.38
1.7.99.4
1.7.2.1
1.7.99.3
1.9.3.2
1.18.6.1
1.19.6.1
1.4.3.
4.2.1.
1.14.13.35
1.14.12.1
4.3.1.
Nitric oxide
1.7.99.7
Dinitrogen oxide
1.7.99.6
Nitrogen
4.4.1.1
4.4.1.2
4.4.1.8
1.4.3.
Amines
3.5.1.
Amides
6.3.4.6
Cyclic amidines
3.5.4.
3.5.3.
Amidines
3.5.5.1
3.5.5.2
Nitriles


METABOLIC PATHWAYS
Glycan Biosynthesis and Metabolism
Biodegradation of Xenobiotics
Nucleotide Metabolism
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Secondary Metabolites
01100 5/31/04

Importance of networks in biology

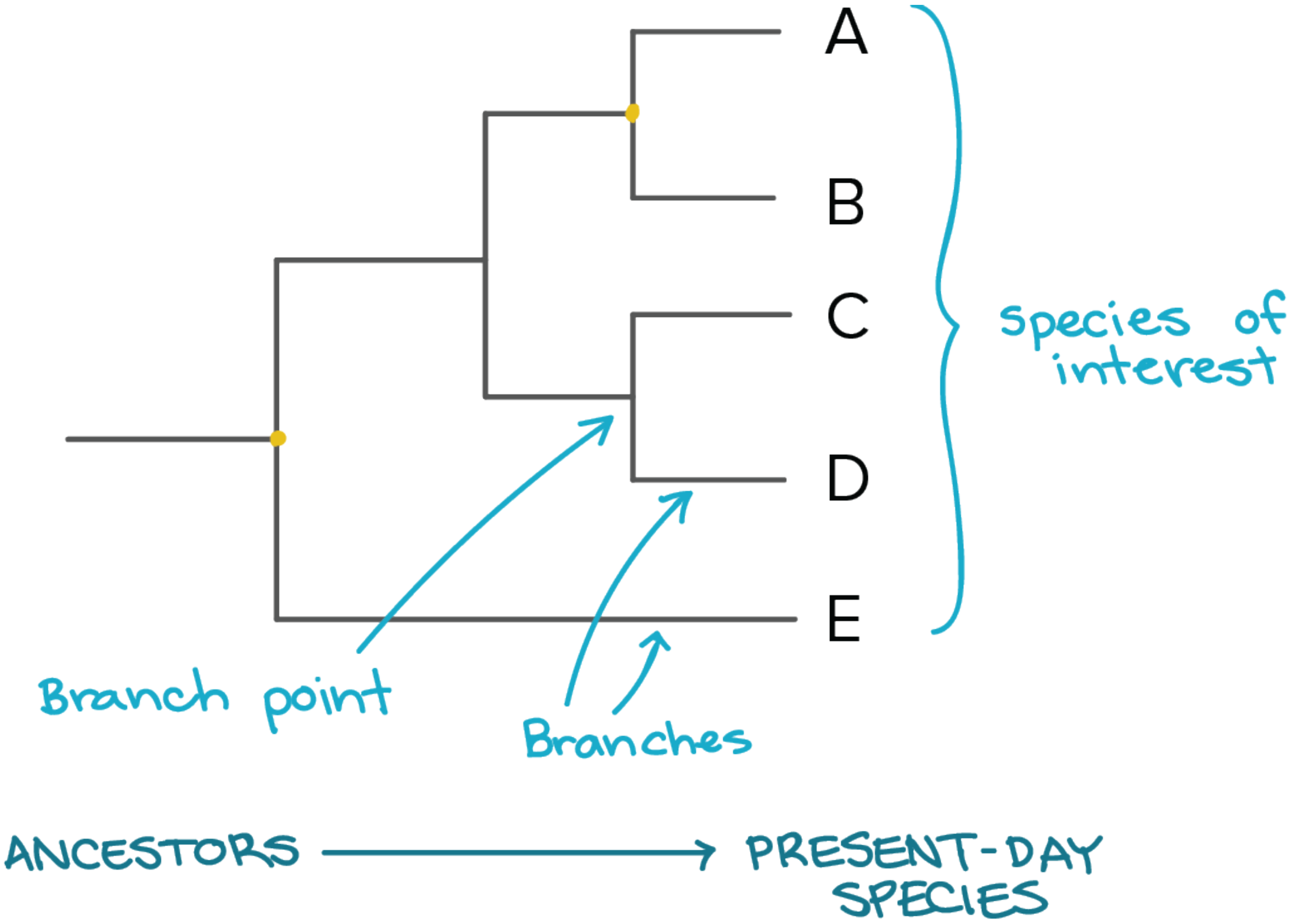


Gene interaction networks



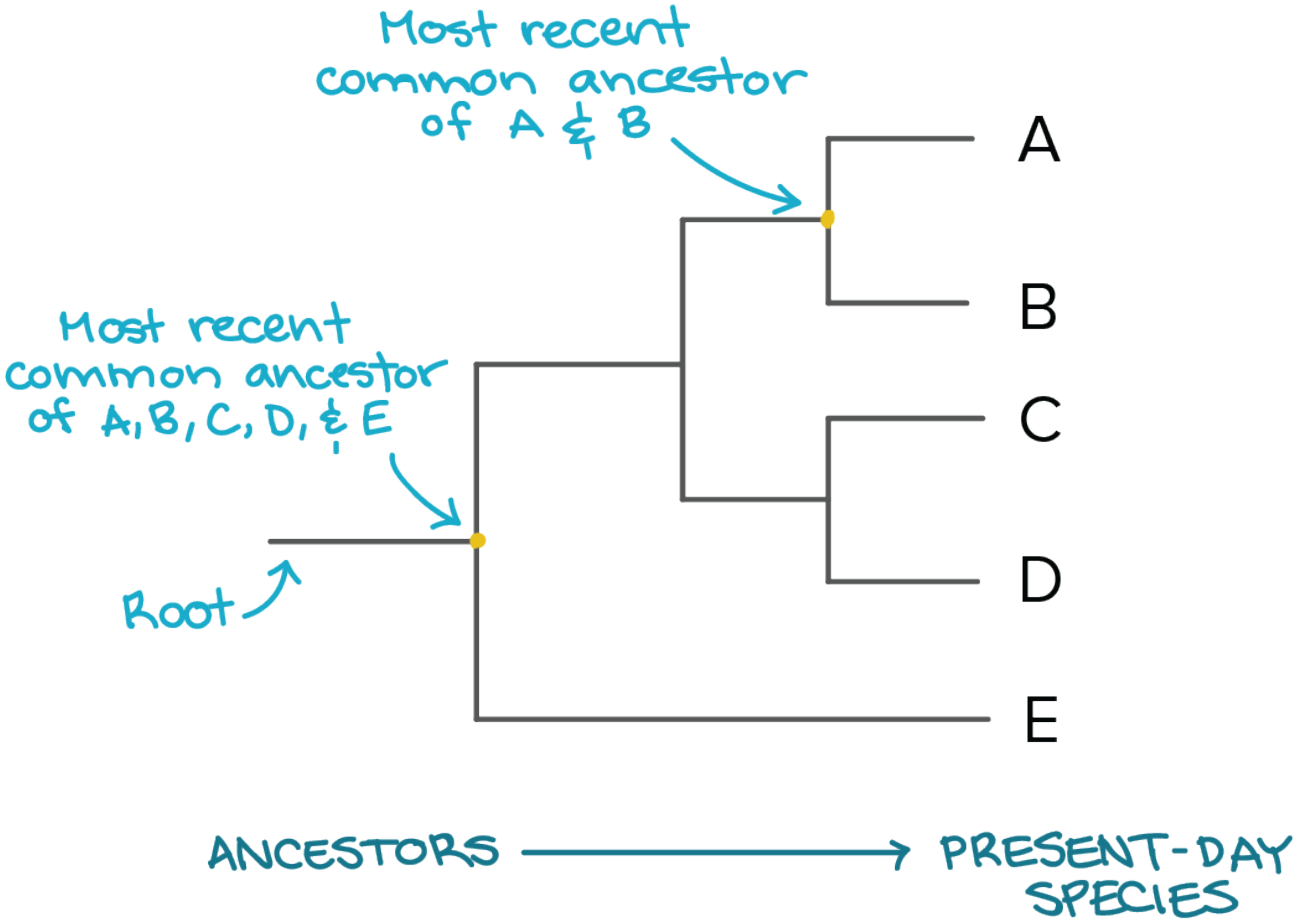
Protein interaction network

Phylogeny



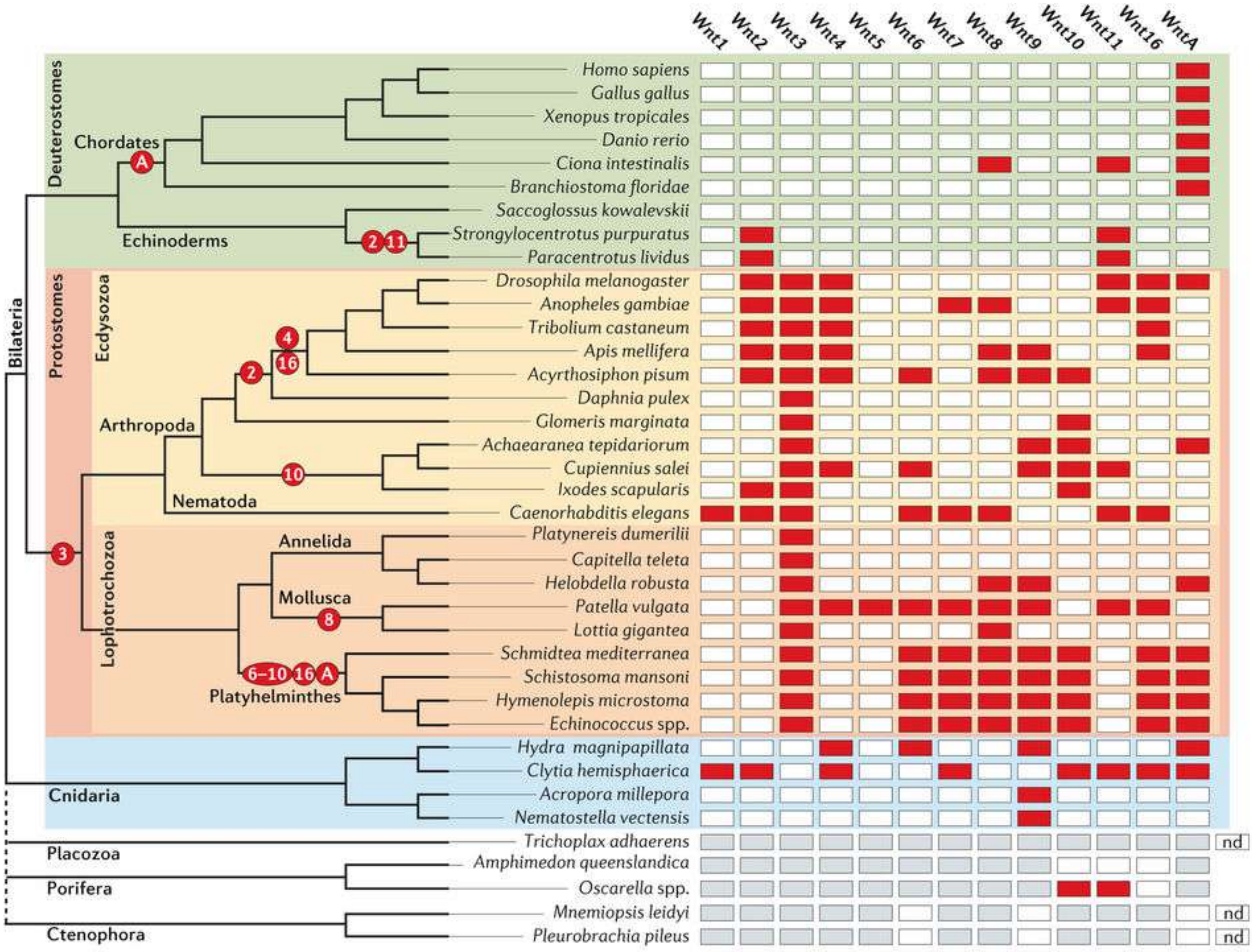
https://www.khanacademy.org/science/biology/her/tree-of-life/a/phylogenetic-trees

Phylogeny


Most recent common ancestor of A & B
A
B
C
D
E
Most recent common ancestor of A, B, C, D, & E
Root
ANCESTORS
PRESENT-DAY SPECIES

https://www.khanacademy.org/science/biology/her/tree-of-life/a/phylogenetic-trees

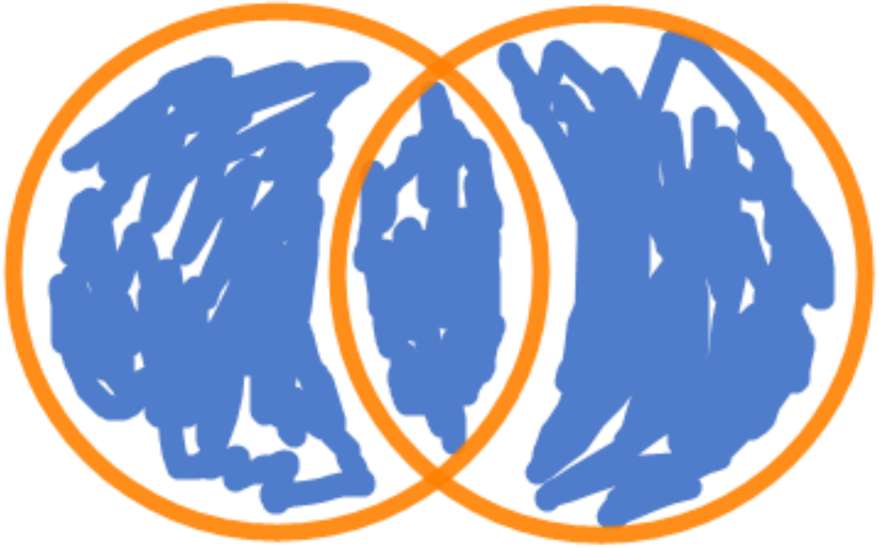
Phylogeny with added features



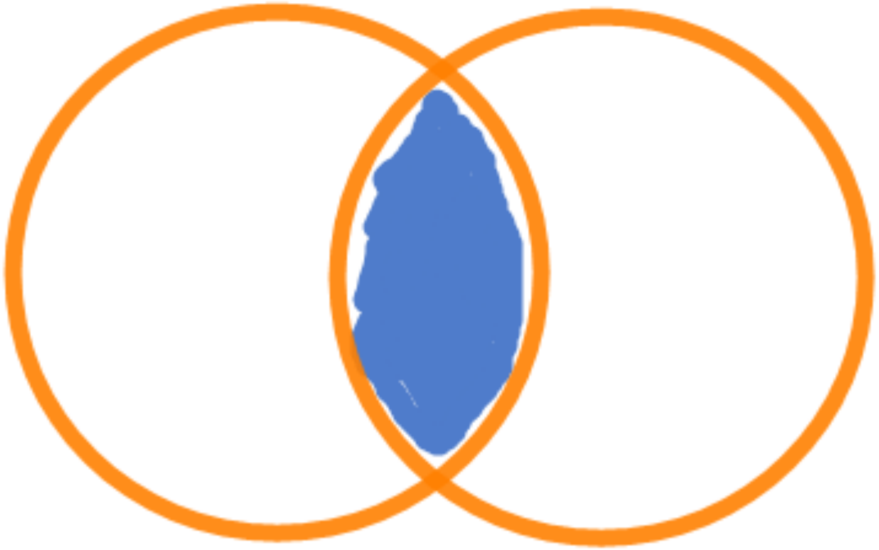
	Wnt1	Wnt2	Wnt3	Wnt4	Wnt5	Wnt6	Wnt7	Wnt8	Wnt9	Wnt10	Wnt11	Wnt16	WntA	
Homo sapiens	□	□	□	□	□	□	□	□	□	□	□	□	🟥	
Gallus gallus	□	□	□	□	□	□	□	□	□	□	□	□	🟥	
Xenopus tropicales	□	□	□	□	□	□	□	□	□	□	□	□	🟥	
Danio rerio	□	□	□	□	□	□	□	□	□	□	□	□	🟥	
Ciona intestinalis	□	□	□	□	□	□	□	🟥	□	□	🟥	□	🟥	
Branchiostoma floridae	□	□	□	□	□	□	□	□	□	□	□	□	🟥	
Saccoglossus kowalevskii	□	□	□	□	□	□	□	□	□	□	□	□	□	
Strongylocentrotus purpuratus	□	🟥	□	□	□	□	□	□	□	□	🟥	□	□	
Paracentrotus lividus	□	🟥	□	□	□	□	□	□	□	□	🟥	□	□	
Drosophila melanogaster	□	🟥	🟥	🟥	□	□	□	□	□	□	🟥	🟥	🟥	
Anopheles gambiae	□	🟥	🟥	🟥	□	□	🟥	🟥	□	□	🟥	🟥	□	
Tribolium castaneum	□	🟥	🟥	🟥	□	□	□	□	□	□	□	🟥	□	
Apis mellifera	□	🟥	🟥	🟥	□	□	□	🟥	🟥	□	□	🟥	□	
Acyrthosiphon pisum	□	🟥	🟥	🟥	□	🟥	□	🟥	🟥	🟥	□	□	□	
Daphnia pulex	□	□	🟥	□	□	□	□	□	□	□	□	□	□	
Glomeris marginata	□	□	🟥	□	□	□	□	□	□	🟥	□	□	□	
Achaearanea tepidariorum	□	□	🟥	□	□	□	□	□	🟥	🟥	□	□	🟥	
Cupiennius salei	□	□	🟥	🟥	□	🟥	□	□	🟥	🟥	🟥	□	□	
Ixodes scapularis	□	🟥	🟥	□	□	□	□	□	□	🟥	□	□	□	
Caenorhabditis elegans	🟥	🟥	🟥	□	□	🟥	🟥	🟥	□	□	🟥	🟥	□	
Platynereis dumerilii	□	□	🟥	□	□	□	□	□	□	□	□	□	□	
Capitella teleta	□	□	🟥	□	□	□	□	□	□	□	□	□	□	
Helobdella robusta	□	□	🟥	□	□	□	□	🟥	🟥	□	□	□	🟥	
Patella vulgata	□	□	🟥	🟥	🟥	🟥	🟥	🟥	🟥	□	🟥	🟥	□	
Lottia gigantea	□	□	🟥	□	□	□	□	🟥	□	□	□	□	□	
Schmidtea mediterranea	□	□	🟥	□	□	🟥	🟥	🟥	🟥	🟥	□	🟥	🟥	
Schistosoma mansoni	□	□	🟥	□	□	🟥	🟥	🟥	🟥	🟥	□	🟥	🟥	
Hymenolepis microstoma	□	□	🟥	□	□	🟥	🟥	🟥	🟥	🟥	□	🟥	🟥	
Echinococcus spp.	□	□	🟥	□	□	🟥	🟥	🟥	🟥	🟥	□	🟥	🟥	
Hydra magnipapillata	□	□	□	🟥	□	🟥	□	□	🟥	□	□	□	🟥	
Clytia hemisphaerica	🟥	🟥	□	🟥	□	□	🟥	□	□	🟥	🟥	🟥	🟥	
Acropora millepora	□	□	□	□	□	□	□	□	🟥	□	□	□	□	
Nematostella vectensis	□	□	□	□	□	□	□	□	🟥	□	□	□	□	
Trichoplax adhaerens	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	nd
Amphimedon queenslandica	▨	▨	▨	▨	▨	▨	▨	▨	▨	□	□	□	▨	
Oscarella spp.	▨	▨	▨	▨	▨	▨	▨	▨	▨	🟥	🟥	□	▨	
Mnemiopsis leidyi	▨	▨	▨	▨	▨	□	▨	▨	□	▨	▨	▨	□	nd
Pleurobrachia pileus	▨	▨	▨	▨	▨	□	▨	▨	□	▨	▨	▨	□	nd

Nature Reviews | Genetics

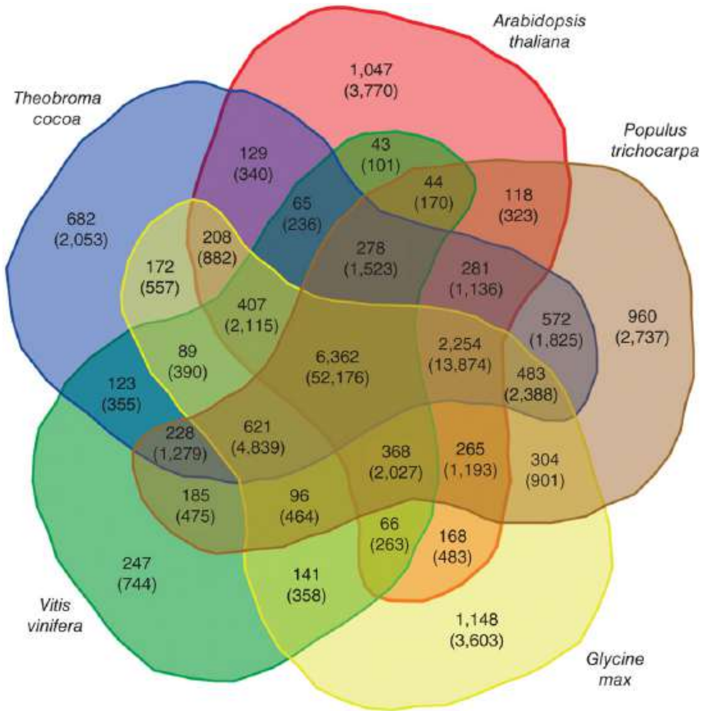
Intersections, unions – Venn diagrams

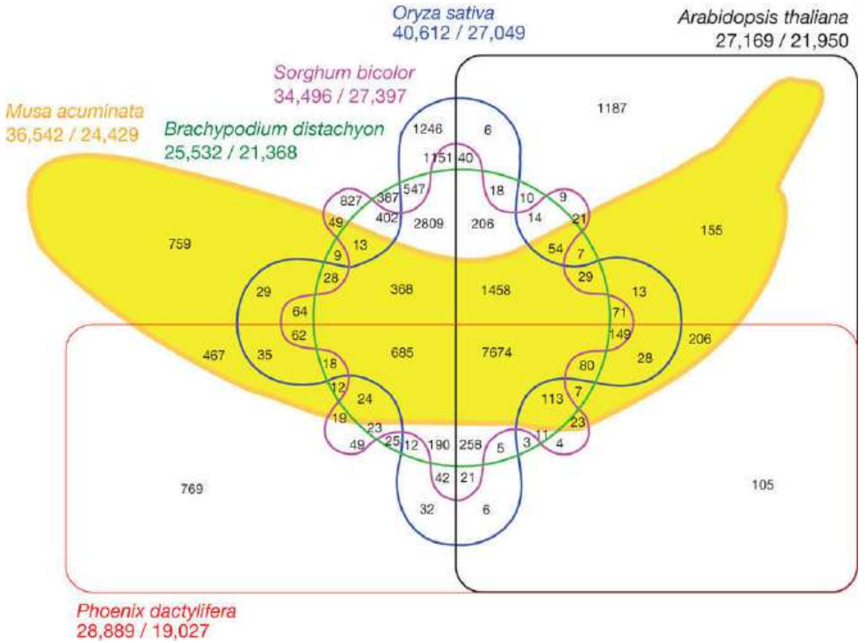


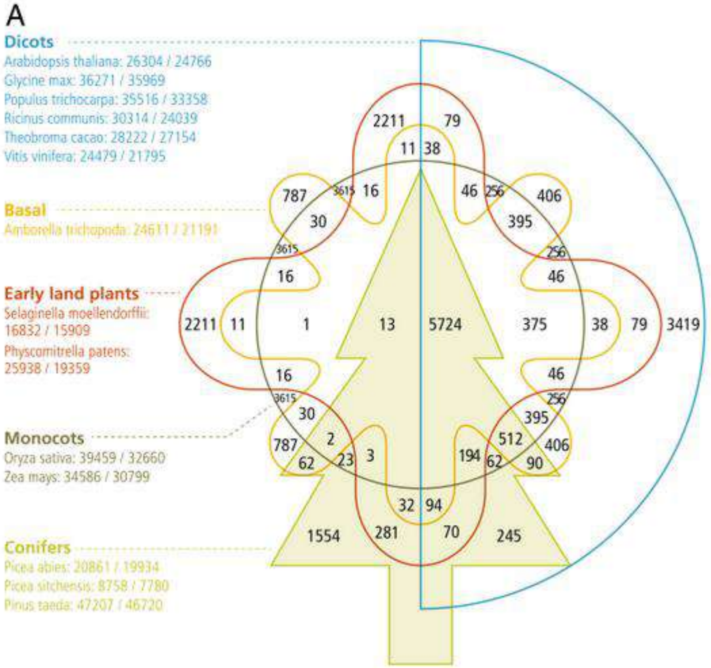
$A \cup B$

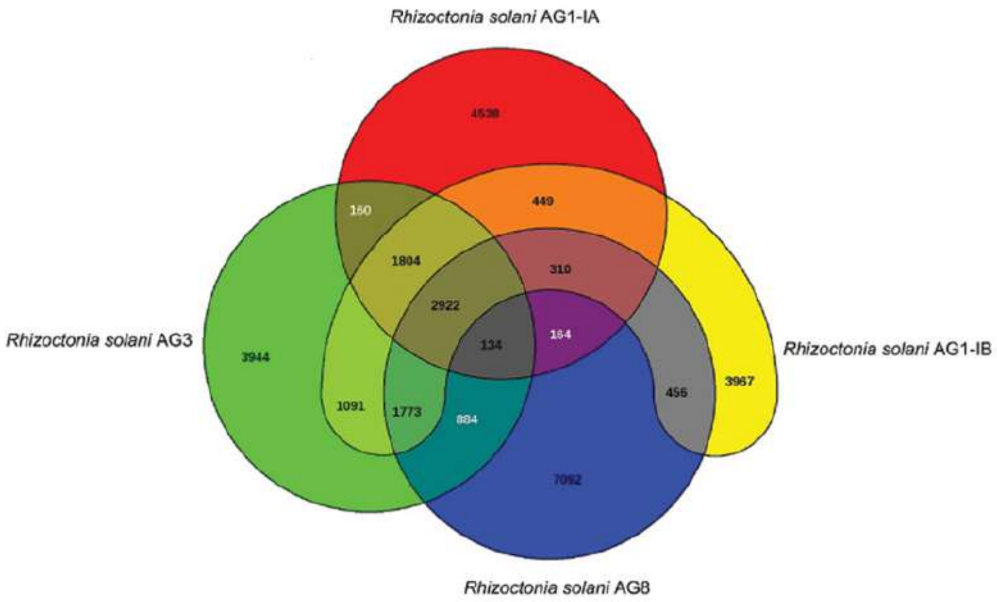


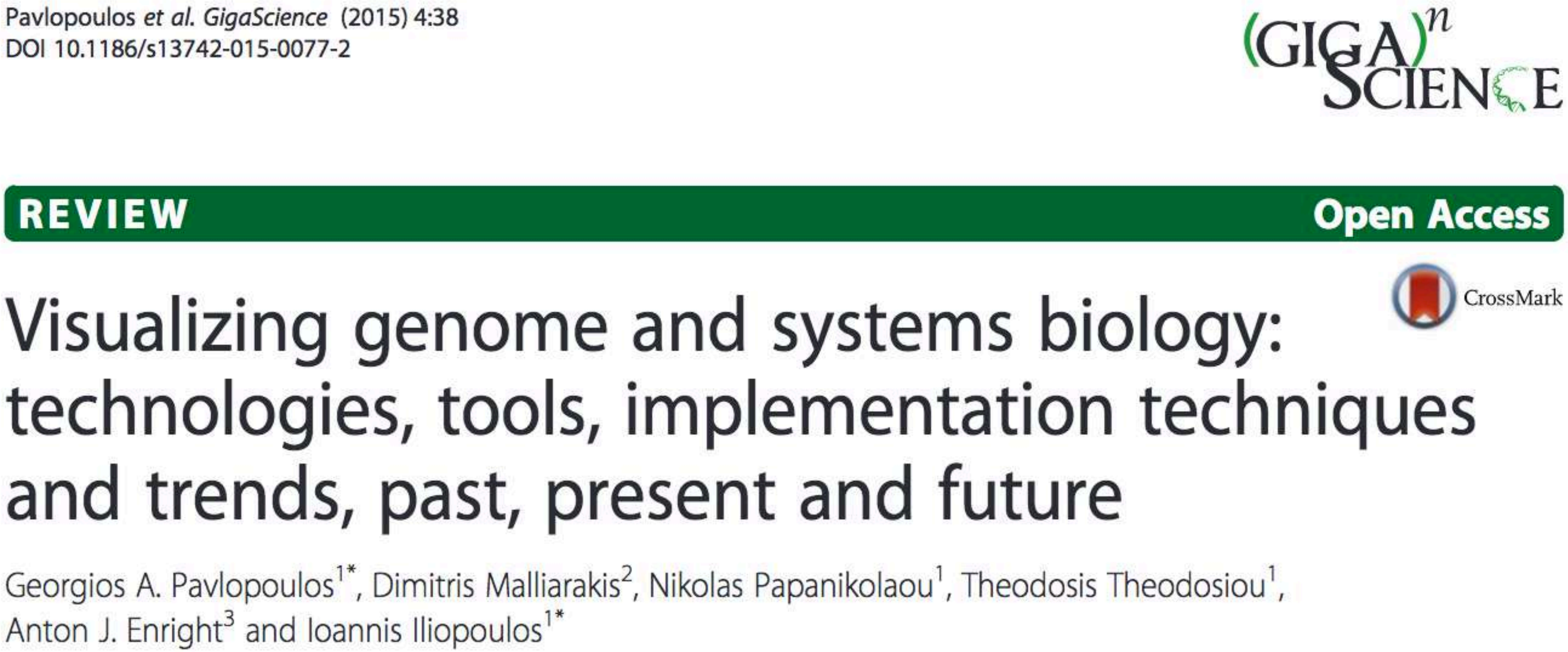
$A \cap B$


Theobroma cocoa
Arabidopsis thaliana
Populus trichocarpa
Vitis vinifera
Glycine max
1,047 (3,770)
129 (340)
43 (101)
44 (170)
118 (323)
682 (2,053)
65 (236)
172 (557)
208 (882)
278 (1,523)
281 (1,136)
960 (2,737)
572 (1,825)
407 (2,115)
89 (390)
6,362 (52,176)
2,254 (13,874)
483 (2,388)
123 (355)
228 (1,279)
621 (4,839)
368 (2,027)
265 (1,193)
304 (901)
185 (475)
96 (464)
66 (263)
168 (483)
247 (744)
141 (358)
1,148 (3,603)


Oryza sativa
40,612 / 27,049
Arabidopsis thaliana
27,169 / 21,950
Sorghum bicolor
34,496 / 27,397
Musa acuminata
36,542 / 24,429
Brachypodium distachyon
25,532 / 21,368
Phoenix dactylifera
28,889 / 19,027
1187
1246
6
1151
40
827
387
547
18
10
9
49
402
2809
206
14
21
155
759
13
9
54
7
28
29
29
368
1458
13
64
71
62
149
467
35
685
7674
28
206
18
80
12
24
113
7
19
23
23
49
25
12
190
258
5
3
11
4
42
21
769
105
32
6


A
Dicots
Arabidopsis thaliana: 26304 / 24766
Glycine max: 36271 / 35969
Populus trichocarpa: 35516 / 33358
Ricinus communis: 30314 / 24039
Theobroma cacao: 28222 / 27154
Vitis vinifera: 24479 / 21795
Basal
Amborella trichopoda: 24611 / 21191
Early land plants
Selaginella moellendorffii: 16832 / 15909
Physcomitrella patens: 25938 / 19359
Monocots
Oryza sativa: 39459 / 32660
Zea mays: 34586 / 30799
Conifers
Picea abies: 20861 / 19934
Picea sitchensis: 8758 / 7780
Pinus taeda: 47207 / 46720
2211
79
11
38
787
3615
16
46
256
406
30
395
3615
256
16
46
2211
11
1
13
5724
375
38
79
3419
16
46
3615
30
256
395
787
2
512
406
62
23
3
194
62
90
32
94
1554
281
70
245


Rhizoctonia solani AG1-IA
Rhizoctonia solani AG3
Rhizoctonia solani AG1-IB
Rhizoctonia solani AG8
4538
160
449
1804
310
2922
134
164
3944
3967
1091
1773
884
456
7092

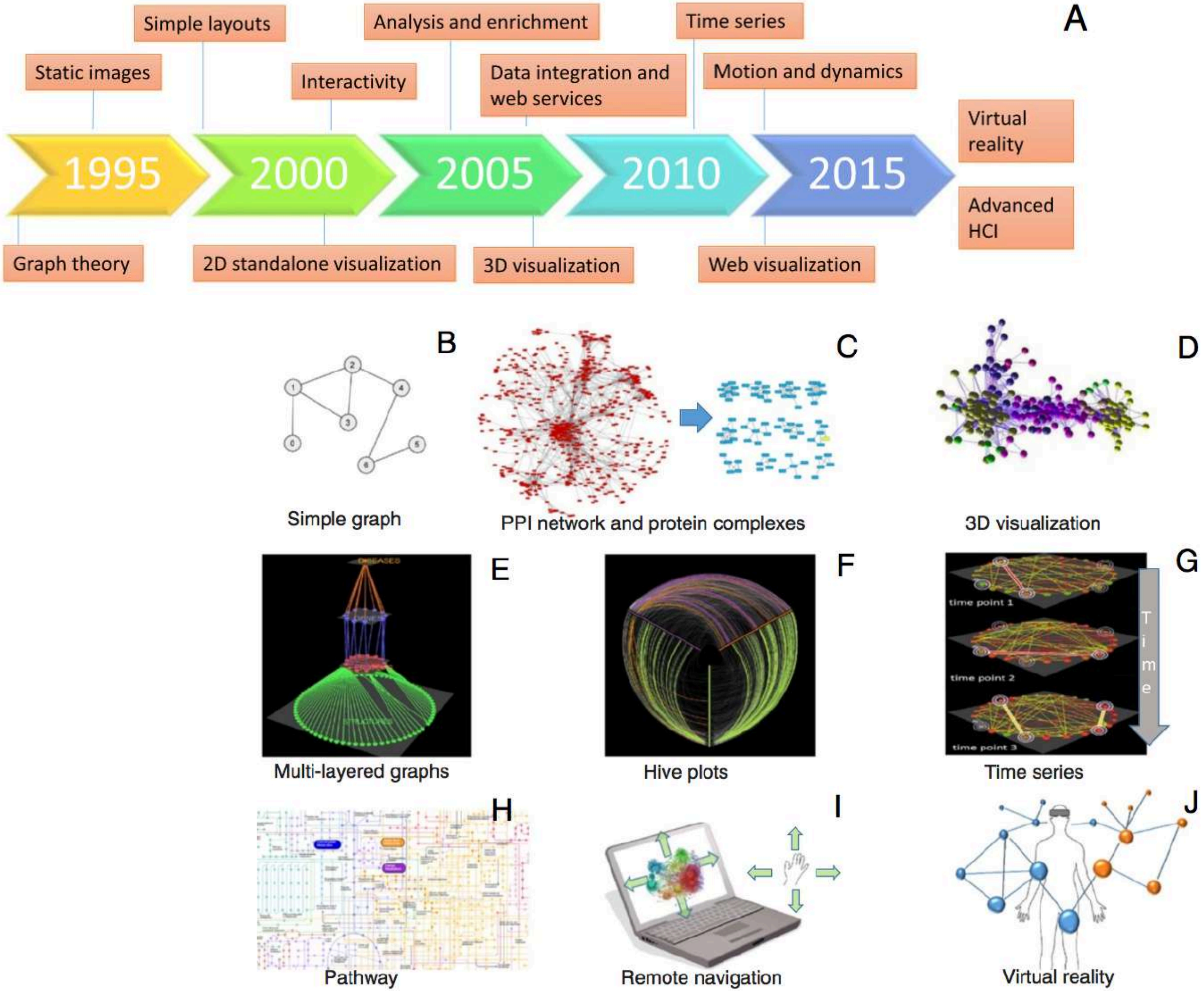


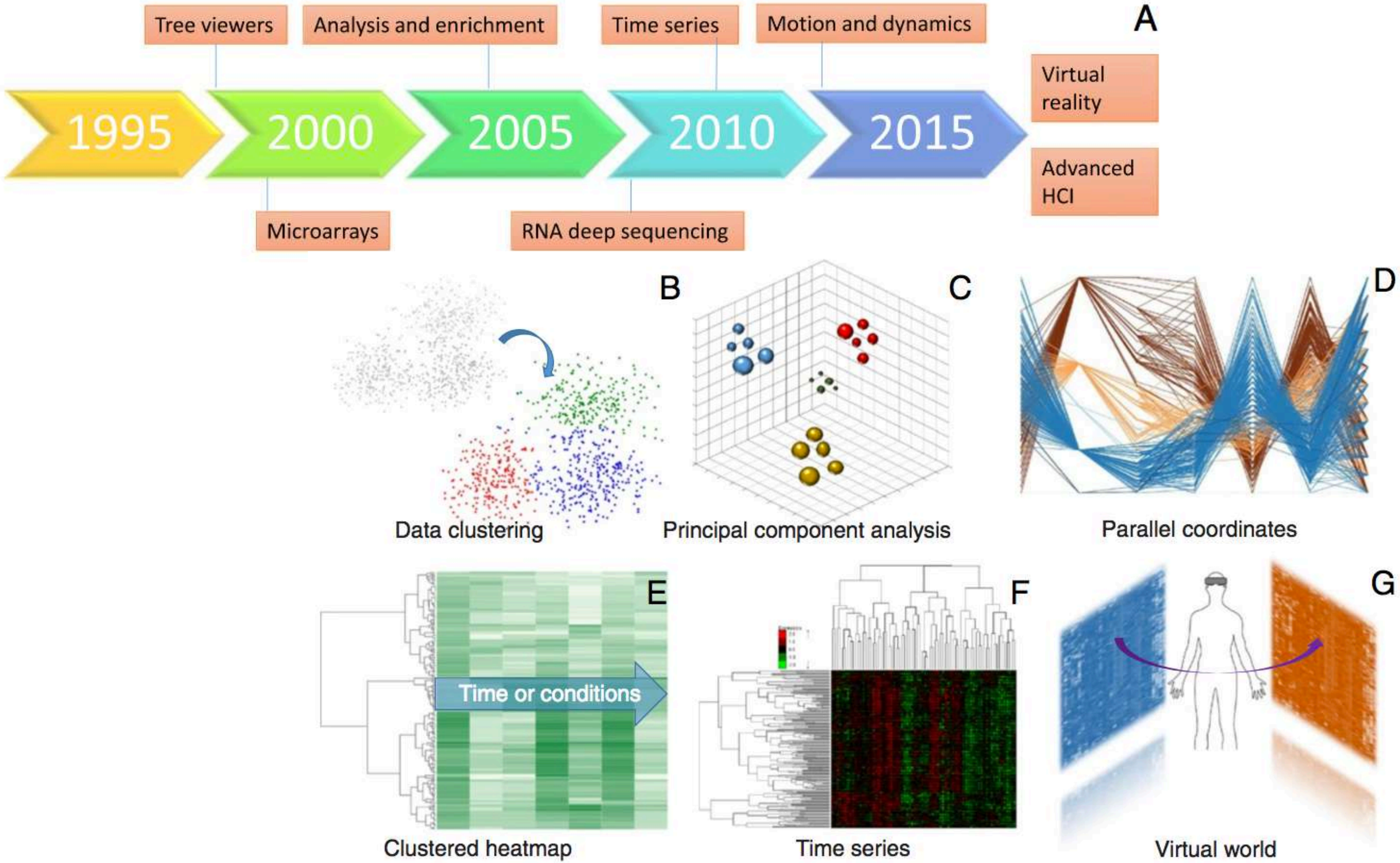
Pavlopoulos *et al. GigaScience* (2015) 4:38
DOI 10.1186/s13742-015-0077-2

REVIEW **Open Access**

Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future

Georgios A. Pavlopoulos[1*], Dimitris Malliarakis[2], Nikolas Papanikolaou[1], Theodosis Theodosiou[1], Anton J. Enright[3] and Ioannis Iliopoulos[1*]


A
Static images
Simple layouts
Interactivity
Analysis and enrichment
Data integration and web services
Time series
Motion and dynamics
Virtual reality
Advanced HCI
1995
2000
2005
2010
2015
Graph theory
2D standalone visualization
3D visualization
Web visualization
B
Simple graph
C
PPI network and protein complexes
D
3D visualization
E
DISEASES
STRUCTURES
Multi-layered graphs
F
Hive plots
G
time point 1
time point 2
time point 3
Time
Time series
H
Pathway
I
Remote navigation
J
Virtual reality

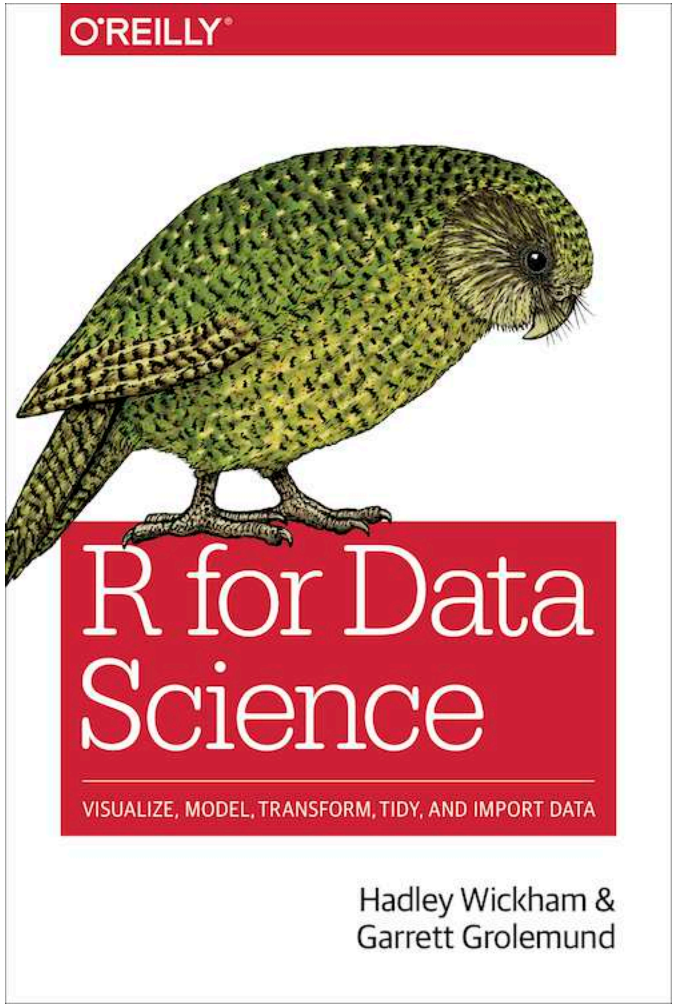

A
Tree viewers
Analysis and enrichment
Time series
Motion and dynamics
1995
2000
2005
2010
2015
Microarrays
RNA deep sequencing
Virtual reality
Advanced HCI
B
Data clustering
C
Principal component analysis
D
Parallel coordinates
E
Time or conditions
Clustered heatmap
F
Time series
G
Virtual world

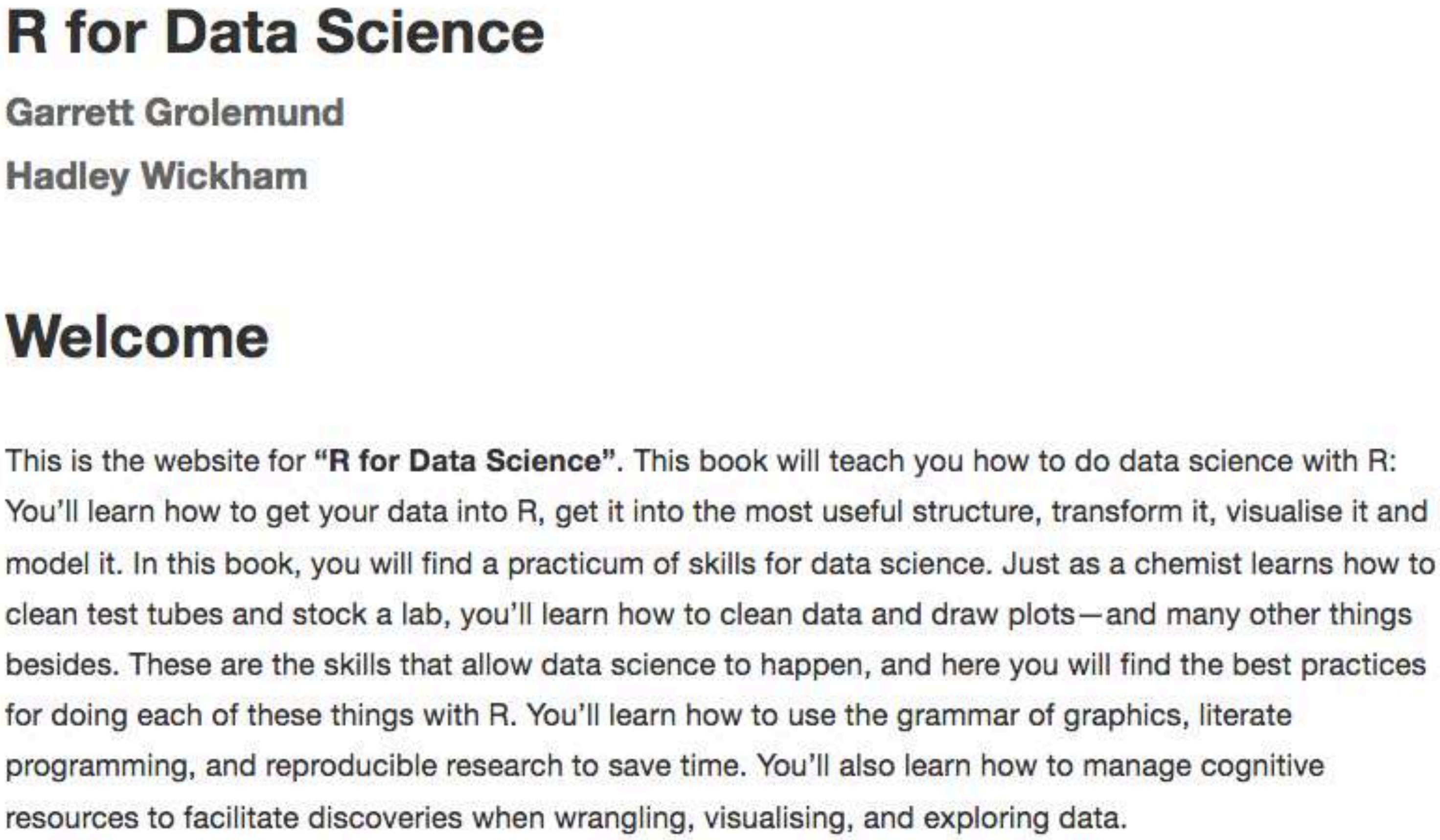
R is a programming environment

- It’s **free**
  - Hence R is supported by a large user network
  - R is open source
- Can be run on Windows, Linux and Mac
- Provides an unparalleled platform for programming new statistical methods in an easy and straightforward manner.
- **Excellent graphics capabilities**
- **Lots and lots of analysis packages**
- It is also **old**, hence you need to know new functions which do things much faster

Suggested textbook (also a gitbook!)

http://r4ds.had.co.nz/





R for Data Science

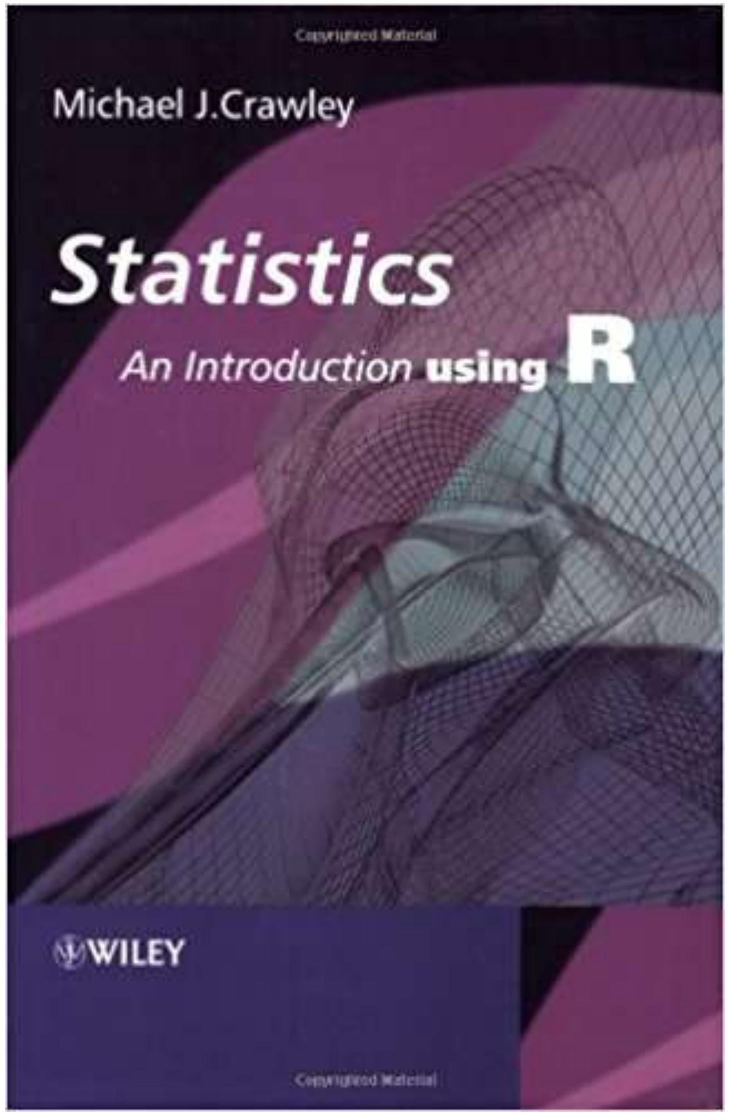
Garrett Grolemund

Hadley Wickham

Welcome

This is the website for **"R for Data Science"**. This book will teach you how to do data science with R: You'll learn how to get your data into R, get it into the most useful structure, transform it, visualise it and model it. In this book, you will find a practicum of skills for data science. Just as a chemist learns how to clean test tubes and stock a lab, you'll learn how to clean data and draw plots—and many other things besides. These are the skills that allow data science to happen, and here you will find the best practices for doing each of these things with R. You'll learn how to use the grammar of graphics, literate programming, and reproducible research to save time. You'll also learn how to manage cognitive resources to facilitate discoveries when wrangling, visualising, and exploring data.

Suggested textbook + learn statistics



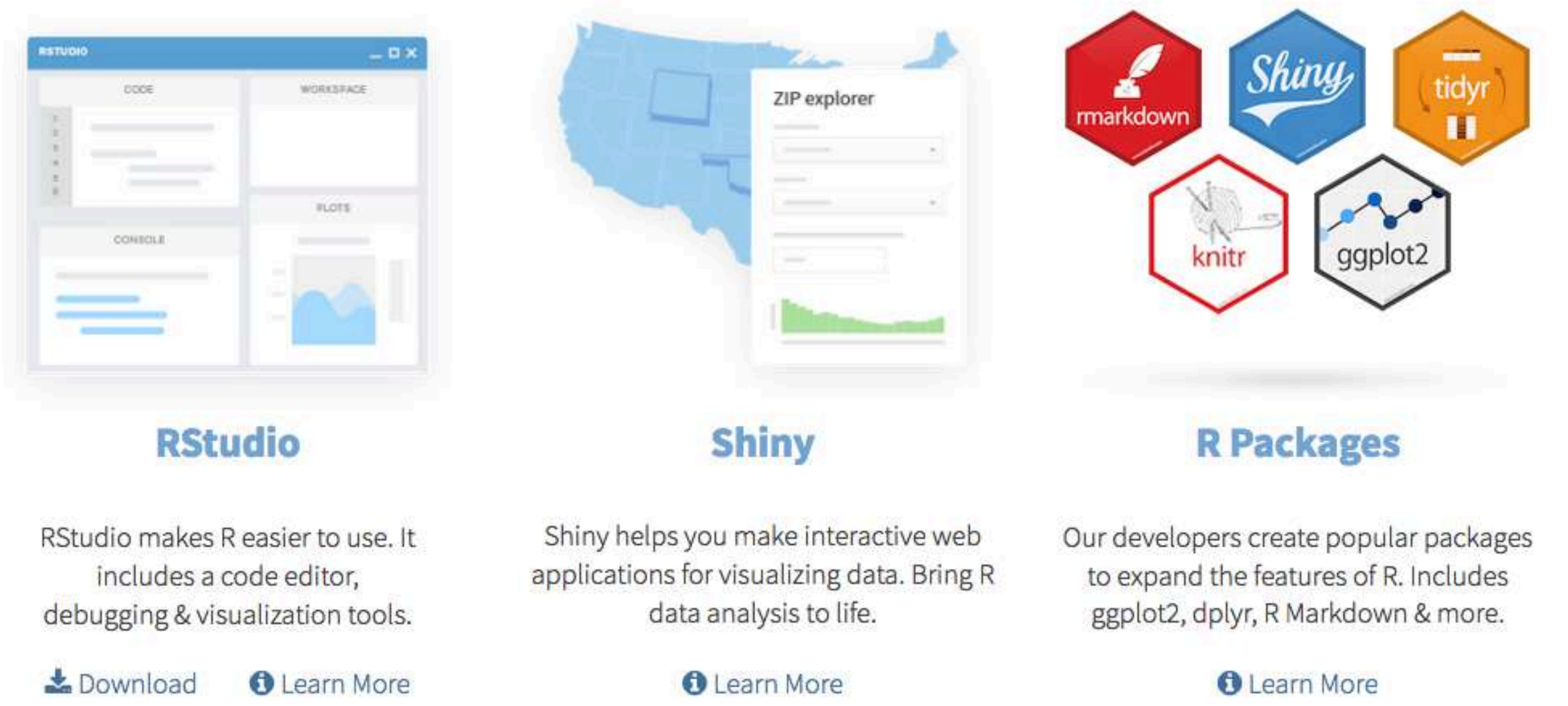
Code is kind of obsoleted but contents about statistics are still outstanding

http://onlinelibrary.wiley.com/book/10.1002/9781119941750

Download R and Rstudio

R Studio

http://www.r-project.org
https://www.rstudio.com/



RStudio

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

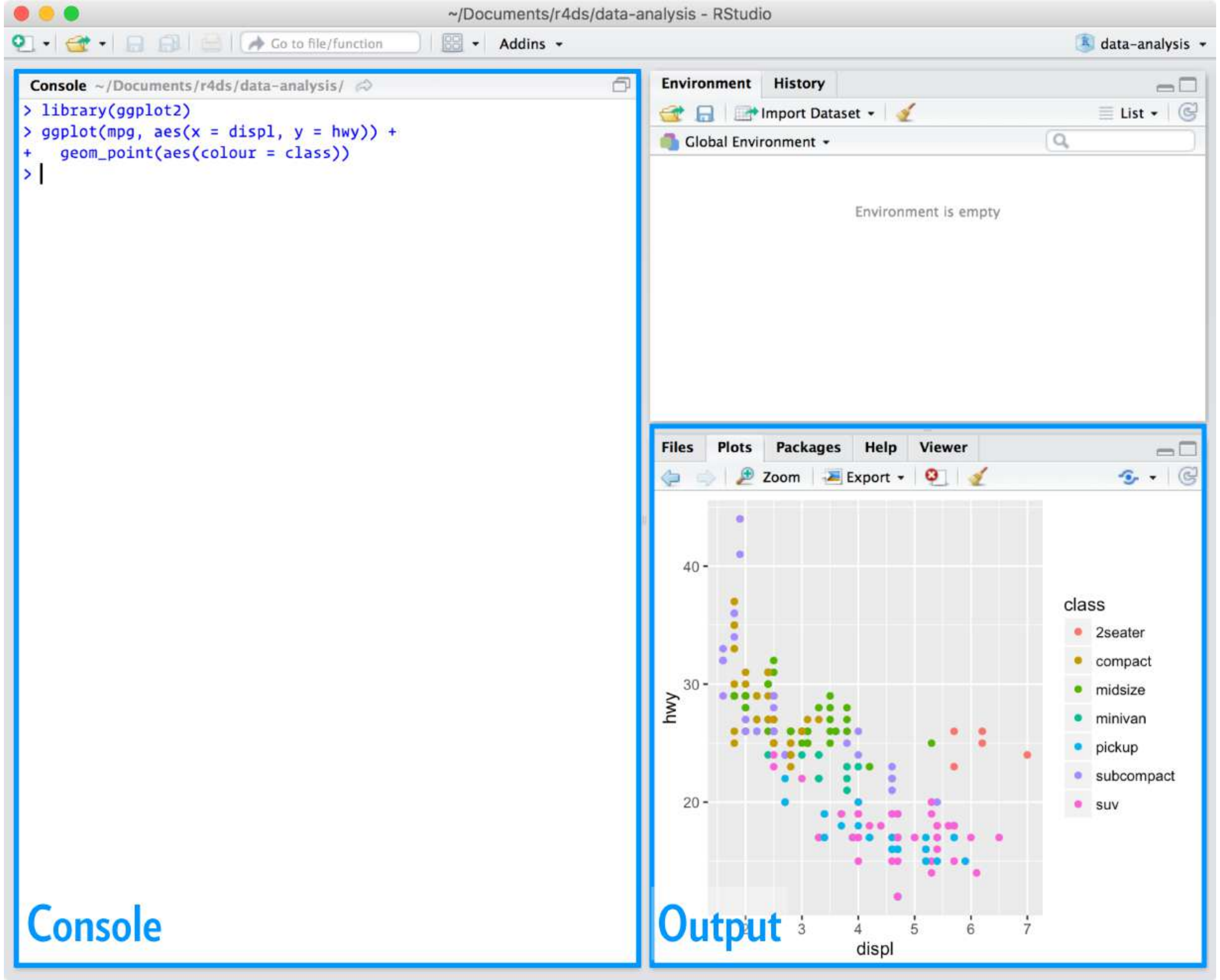
Shiny

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

R Packages

Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

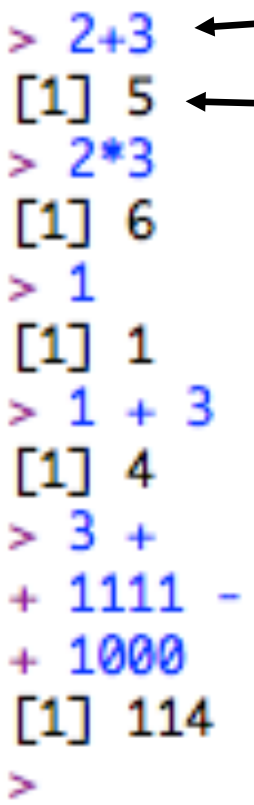
Rstudio interface



```
> library(ggplot2)
> ggplot(mpg, aes(x = displ, y = hwy)) +
+   geom_point(aes(colour = class))
> 
```

Environment is empty

R as a calculator



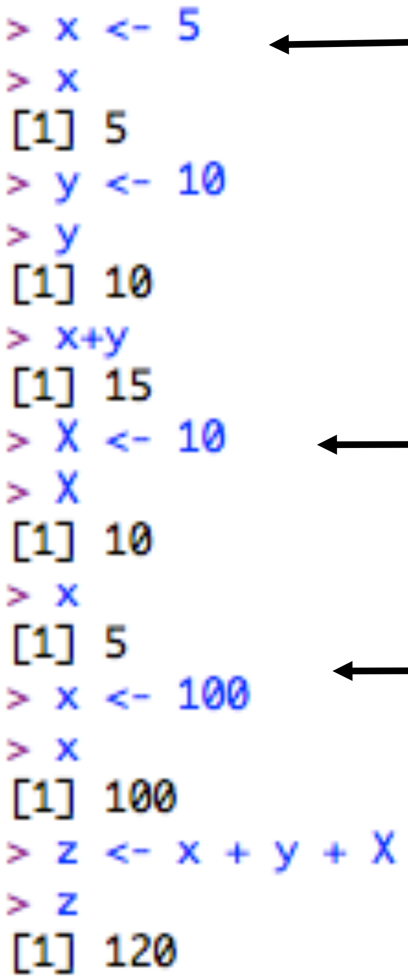
```
> 2+3
[1] 5
> 2*3
[1] 6
> 1
[1] 1
> 1 + 3
[1] 4
> 3 +
+ 1111 -
+ 1000
[1] 114
>
```

Press enter to complete the expression

Completed expression

Incomplete expression will result in continuation prompt **+**

Assignment



```
> x <- 5
> x
[1] 5
> y <- 10
> y
[1] 10
> x+y
[1] 15
> X <- 10
> X
[1] 10
> x
[1] 5
> x <- 100
> x
[1] 100
> z <- x + y + X
> z
[1] 120
```

<- is the assignment operation

R is case sensitive ; **x** does not equal to **X**

Original value is replaced

New value can be assigned as the result of calculation

Boolean assignment

```
student <- 30000
phd <- 56000
```

```
student > phd
```

```
[1] FALSE
```

```
student < phd
```

```
[1] TRUE
```

```
student != phd
```

```
[1] FALSE
```

```
student + student > phd   ←   #Two heads are better than one
```

```
[1] TRUE
```

Vector is the simplest data structure in R

```
x<- c(1,2,3,4,5,6,7,8,9,10)
```

c = combine

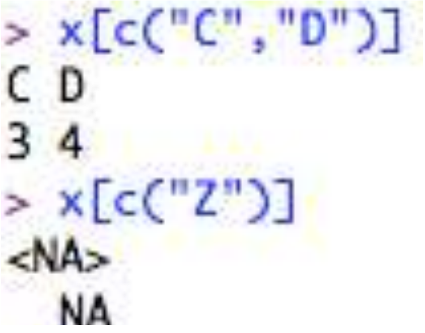
In this case, we assign a **vector** of 10 numbers into x

```
x * 2
x /10 + 1
```

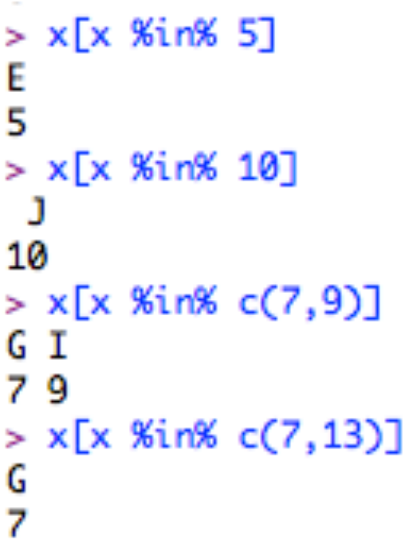
Selection

```
x<- c(1,2,3,4,5,6,7,8,9,10)
names(x)<-c("A","B","C","D","E","F","G","H","I","J")
```

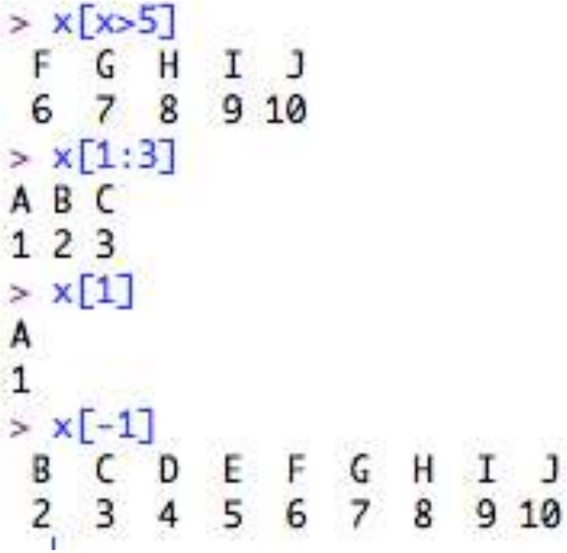
```
x[x>5]
x[1:3]
x[1]
x[-1]
x[c("C","D")]
x[c("Z")]
x[x %in% c(7,9)]
x[x %in% c(7,13)]
```



```
> x[c("C","D")]
C D
3 4
> x[c("Z")]
<NA>
  NA
```



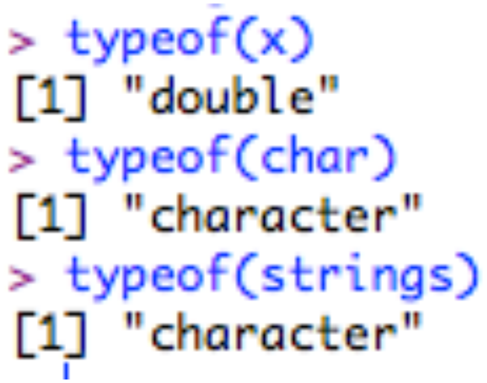
```
> x[x %in% 5]
E
5
> x[x %in% 10]
 J
10
> x[x %in% c(7,9)]
G I
7 9
> x[x %in% c(7,13)]
G
7
```



```
> x[x>5]
 F  G  H  I  J
 6  7  8  9 10
> x[1:3]
A B C
1 2 3
> x[1]
A
1
> x[-1]
 B  C  D  E  F  G  H  I  J
 2  3  4  5  6  7  8  9 10
```

Different types of vectors

```
x<- c(1,2,3,4,5,6,7,8,9,10)
strings <- c("AS","BRC")

typeof(x)
typeof(strings)
```



```
> typeof(x)
[1] "double"
> typeof(char)
[1] "character"
> typeof(strings)
[1] "character"
```

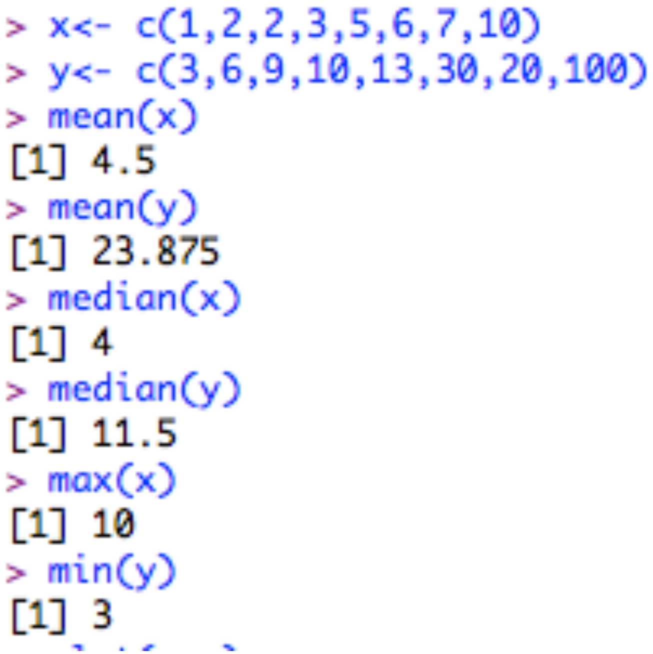
This matters when one data type is numbers, and you want to sort them categorically

Function

function (arg1, arg2, arg3… , option1=,option2=...)

```
x<- c(1,2,3,4,5,6,7,8,9,10)
y<- c(3,6,9,10,13,30,20,100)
```

```
mean(x)
mean(y)
median(x)
max(x)
```

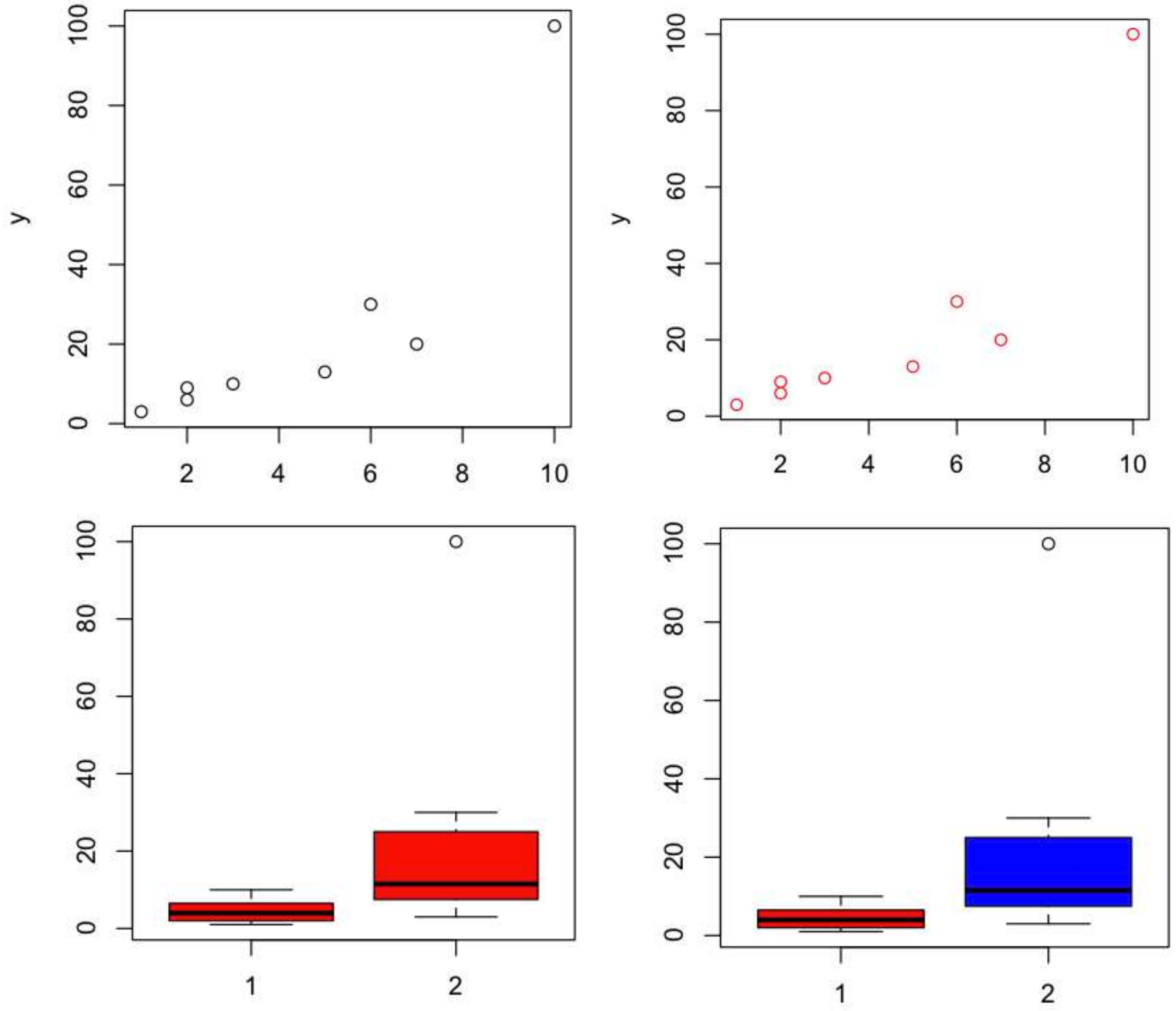


```
> x<- c(1,2,2,3,5,6,7,10)
> y<- c(3,6,9,10,13,30,20,100)
> mean(x)
[1] 4.5
> mean(y)
[1] 23.875
> median(x)
[1] 4
> median(y)
[1] 11.5
> max(x)
[1] 10
> min(y)
[1] 3
```

- Must have **assigned names**
- Applies using **round brackets**
- Takes **argument** and options

R simple plot I

```
x<- c(1,2,3,4,5,6,7,8,9,10)
y<- c(3,6,9,10,13,30,20,100,220,100)

plot(x,y)
plot(x,y,col="red")

boxplot(x,y,col="red")
boxplot(x,y,col=c("hotpink", "yellow"))

boxplot(x,y,col=c("hotpink", "yellow"),main="Lec2")
```

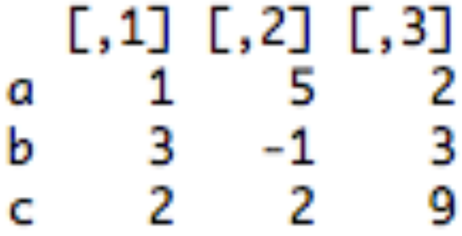


R simple plot II

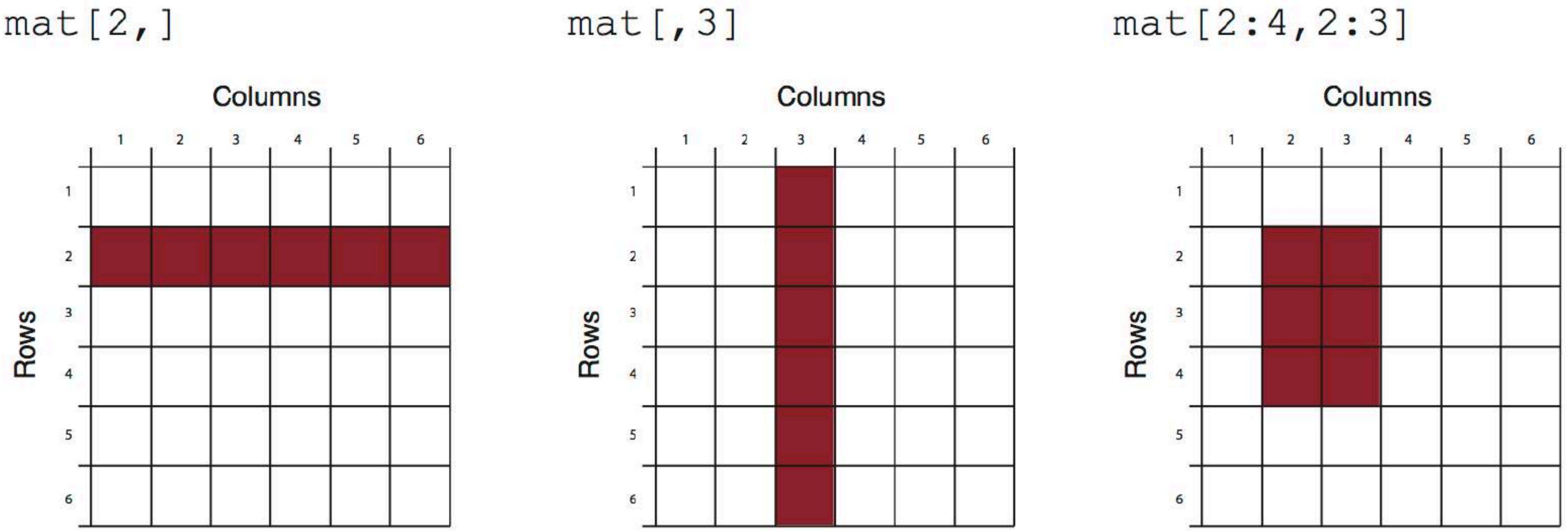
# Follow examples here:
http://al2na.github.io/compgenr/intro_to_r/plotting_in_r.html

Matrices are a collection of vectors of the same type

```
mat <- matrix(c(1, 3, 2, 5, -1, 2, 2, 3, 9), nrow = 3)
rownames(mat) <- c("a","b","c")
colnames(mat) <- c(”x",”y",”z")
```



```
  [,1] [,2] [,3]
a    1    5    2
b    3   -1    3
c    2    2    9
```



mat[2,]

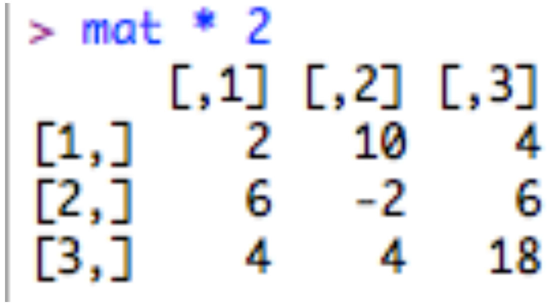
mat[,3]

mat[2:4,2:3]

http://al2na.github.io/compgenr/intro_to_r/data_types.html

Matrices - summary

- Each row and column must have data of the **same** type (numeric, character etc)
- Most useful when do linear algebra (e.g. PCA,)



```
> mat * 2
     [,1] [,2] [,3]
[1,]    2   10    4
[2,]    6   -2    6
[3,]    4    4   18
```

- If you want **different** data types, need to use objects called data.frames

Data frames

- Think of these like Excel spreadsheets
- **All the values of the same variable must go in the same column**
  - E.g., age, sex, RPKM, numbers
- **Rows represent samples**
  - E.g., sample A collected in Taiwan, sample B collected in Japan

- Like matrices but different types of data are allowed

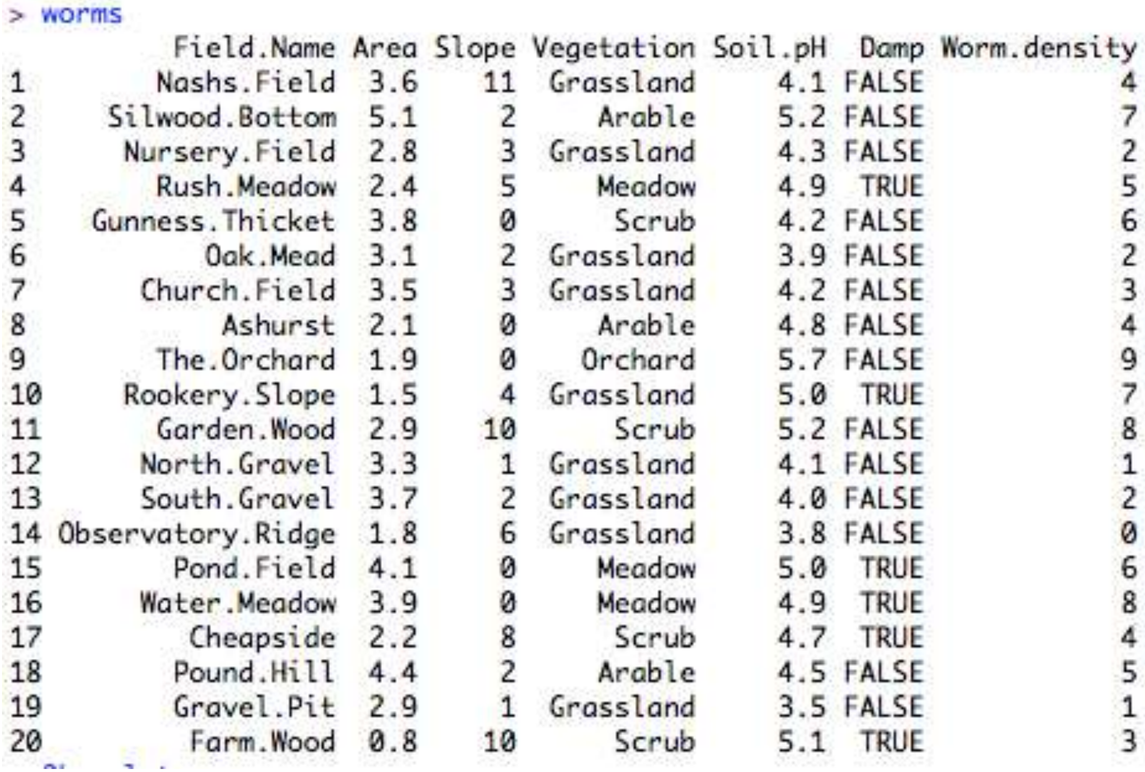
- Tibble from the **dplyr** package ; basically like data frame but much easier to manipulate

R has some pre-installed data frames

```
iris
head(iris)
```

# Or you can read into data

```
worms <- read.table("worms.txt", header=T)
head(worms)
```



```
> worms
          Field.Name Area Slope Vegetation Soil.pH  Damp Worm.density
1        Nashs.Field  3.6    11  Grassland     4.1 FALSE            4
2     Silwood.Bottom  5.1     2     Arable     5.2 FALSE            7
3      Nursery.Field  2.8     3  Grassland     4.3 FALSE            2
4        Rush.Meadow  2.4     5     Meadow     4.9  TRUE            5
5    Gunness.Thicket  3.8     0      Scrub     4.2 FALSE            6
6           Oak.Mead  3.1     2  Grassland     3.9 FALSE            2
7       Church.Field  3.5     3  Grassland     4.2 FALSE            3
8            Ashurst  2.1     0     Arable     4.8 FALSE            4
9        The.Orchard  1.9     0    Orchard     5.7 FALSE            9
10     Rookery.Slope  1.5     4  Grassland     5.0  TRUE            7
11       Garden.Wood  2.9    10      Scrub     5.2 FALSE            8
12      North.Gravel  3.3     1  Grassland     4.1 FALSE            1
13      South.Gravel  3.7     2  Grassland     4.0 FALSE            2
14 Observatory.Ridge  1.8     6  Grassland     3.8 FALSE            0
15        Pond.Field  4.1     0     Meadow     5.0  TRUE            6
16      Water.Meadow  3.9     0     Meadow     4.9  TRUE            8
17         Cheapside  2.2     8      Scrub     4.7  TRUE            4
18        Pound.Hill  4.4     2     Arable     4.5 FALSE            5
19        Gravel.Pit  2.9     1  Grassland     3.5 FALSE            1
20         Farm.Wood  0.8    10      Scrub     5.1  TRUE            3
```

File available here:
https://github.com/shifteight/R/blob/master/TRB/data/worms.txt

Selection in data frames

# Square brackets

- **dat[i ,]** would select the *i*-th row (which is a **vector**)
- **dat[, j]** would select the *j*-th column (which is a **vector**)
- **dat[i, j]** would select the value from the *i*-th row and *j*-th column

```
worms[,1]
worms[1,]
worms[1,1]
```

# dollar ($) operation (for columns only)

```
worms$Area
```

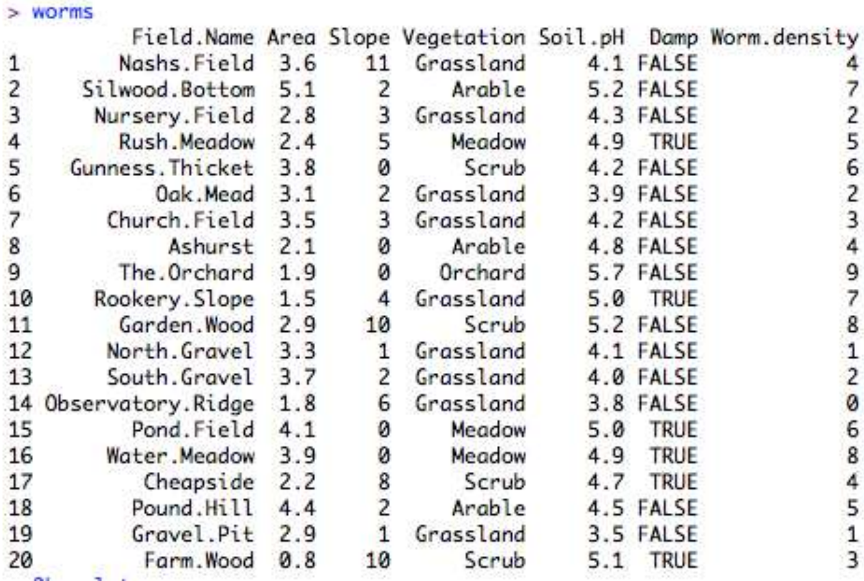
# subset (not discussing today)

Some combinations of it

# Square brackets

- **dat[i ,]** would select the *i*-th row (which is a **vector**)
- **dat[, j]** would select the *j*-th column (which is a **vector**)
- **dat[i, j]** would select the value from the *i*-th row and *j*-th column

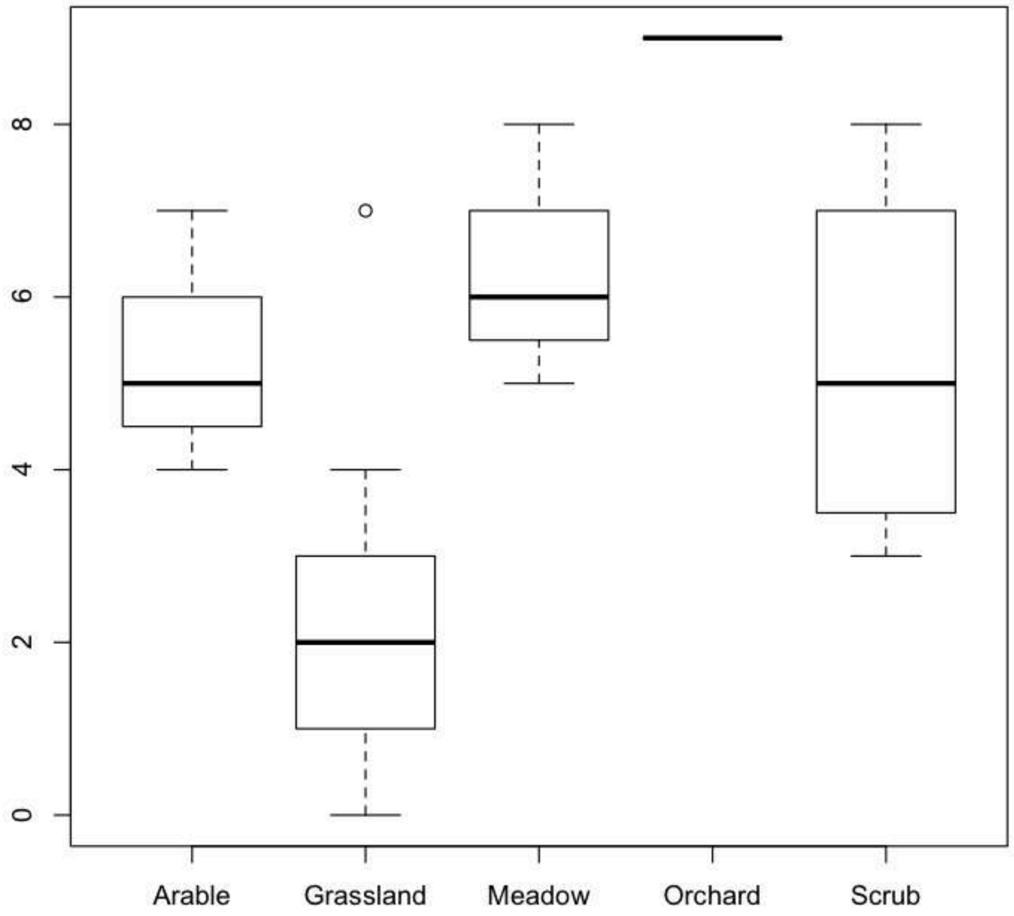
```
worms[worms$Area < 3,]
worms[(worms$Area < 3) & (worms$Worm.density <4),]
worms[(worms$Area < 3) & (worms$Worm.density <4),]$Soil.pH
```

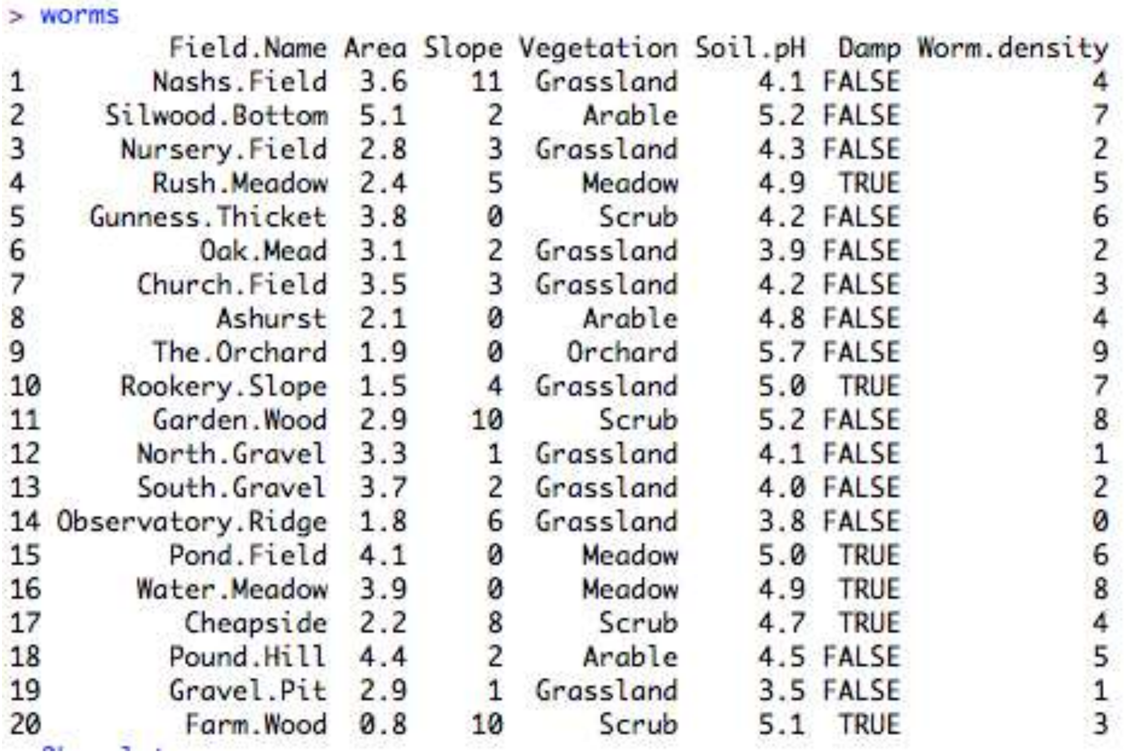


```
> worms
          Field.Name Area Slope Vegetation Soil.pH  Damp Worm.density
1        Nashs.Field  3.6    11  Grassland     4.1 FALSE            4
2     Silwood.Bottom  5.1     2     Arable     5.2 FALSE            7
3      Nursery.Field  2.8     3  Grassland     4.3 FALSE            2
4        Rush.Meadow  2.4     5     Meadow     4.9  TRUE            5
5    Gunness.Thicket  3.8     0      Scrub     4.2 FALSE            6
6           Oak.Mead  3.1     2  Grassland     3.9 FALSE            2
7       Church.Field  3.5     3  Grassland     4.2 FALSE            3
8            Ashurst  2.1     0     Arable     4.8 FALSE            4
9        The.Orchard  1.9     0    Orchard     5.7 FALSE            9
10      Rookery.Slope  1.5     4  Grassland     5.0  TRUE            7
11        Garden.Wood  2.9    10      Scrub     5.2 FALSE            8
12       North.Gravel  3.3     1  Grassland     4.1 FALSE            1
13       South.Gravel  3.7     2  Grassland     4.0 FALSE            2
14 Observatory.Ridge  1.8     6  Grassland     3.8 FALSE            0
15         Pond.Field  4.1     0     Meadow     5.0  TRUE            6
16       Water.Meadow  3.9     0     Meadow     4.9  TRUE            8
17          Cheapside  2.2     8      Scrub     4.7  TRUE            4
18         Pound.Hill  4.4     2     Arable     4.5 FALSE            5
19         Gravel.Pit  2.9     1  Grassland     3.5 FALSE            1
20          Farm.Wood  0.8    10      Scrub     5.1  TRUE            3
```

More plot from dataframes

```
plot(worms$Area,worms$Slope,col=as.numeric(worms$Vegetation))
plot(worms$Area,worms$Slope,col=as.numeric(worms$Vegetation),pch=as.numeric(worms$Vegetation))
boxplot(worms$Worm.density ~ worms$Vegetation)
```





```
> worms
          Field.Name Area Slope Vegetation Soil.pH  Damp Worm.density
1        Nashs.Field  3.6    11  Grassland     4.1 FALSE            4
2     Silwood.Bottom  5.1     2     Arable     5.2 FALSE            7
3      Nursery.Field  2.8     3  Grassland     4.3 FALSE            2
4        Rush.Meadow  2.4     5     Meadow     4.9  TRUE            5
5    Gunness.Thicket  3.8     0      Scrub     4.2 FALSE            6
6           Oak.Mead  3.1     2  Grassland     3.9 FALSE            2
7       Church.Field  3.5     3  Grassland     4.2 FALSE            3
8            Ashurst  2.1     0     Arable     4.8 FALSE            4
9        The.Orchard  1.9     0    Orchard     5.7 FALSE            9
10      Rookery.Slope  1.5     4  Grassland     5.0  TRUE            7
11       Garden.Wood  2.9    10      Scrub     5.2 FALSE            8
12      North.Gravel  3.3     1  Grassland     4.1 FALSE            1
13      South.Gravel  3.7     2  Grassland     4.0 FALSE            2
14 Observatory.Ridge  1.8     6  Grassland     3.8 FALSE            0
15        Pond.Field  4.1     0     Meadow     5.0  TRUE            6
16      Water.Meadow  3.9     0     Meadow     4.9  TRUE            8
17         Cheapside  2.2     8      Scrub     4.7  TRUE            4
18        Pound.Hill  4.4     2     Arable     4.5 FALSE            5
19        Gravel.Pit  2.9     1  Grassland     3.5 FALSE            1
20         Farm.Wood  0.8    10      Scrub     5.1  TRUE            3
```

More useful functions here

```
y<-abs(-20)
x<-Sum(y+5)
Z<-Log(x)
round(x,1)
summary(worms)
head(worms)
tail(worms)
ncol(worms)
nrow(worms)
```

Statistics

```
# Simulate two normal distributions one at mean =4, and another at  6
```

```
x <- rnorm(500,4)                         # mean at 4
y <- rnorm(500,6)                        # mean at 6


# Plot histogram

plot(hist(x), col=rgb(0,0,1,1/4), xlim=c(0,10))
plot(hist(y), col=rgb(1,0,0,1/4), xlim=c(0,10), add=T)

t.test(x,y)
```

```
# Simulate two normal distributions at mean =3
```

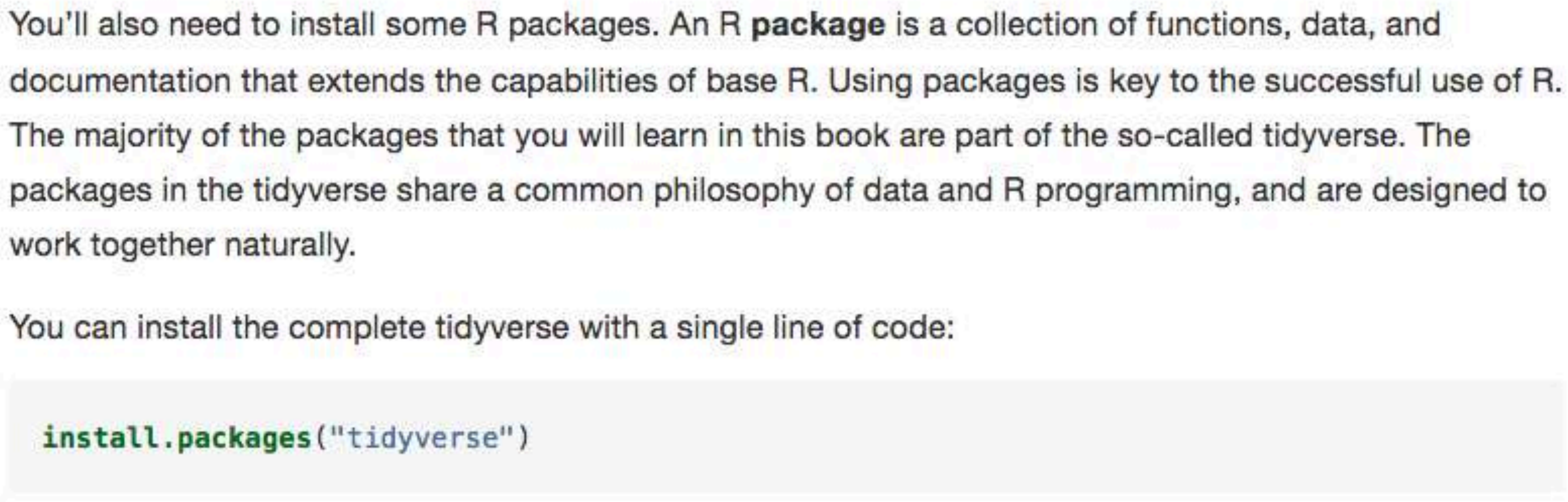
```
x <- rnorm(500,3)
y <- rnorm(500,3)
t.test(x,y)
```

Running out of functions to use?

Use Packages

- R consists of a **core** and **additional packages**.
- Collections of R functions, data, and compiled code
- Well-defined format that ensures easy installation, a basic standard of documentation, and enhances portability and reliability

Install R packages



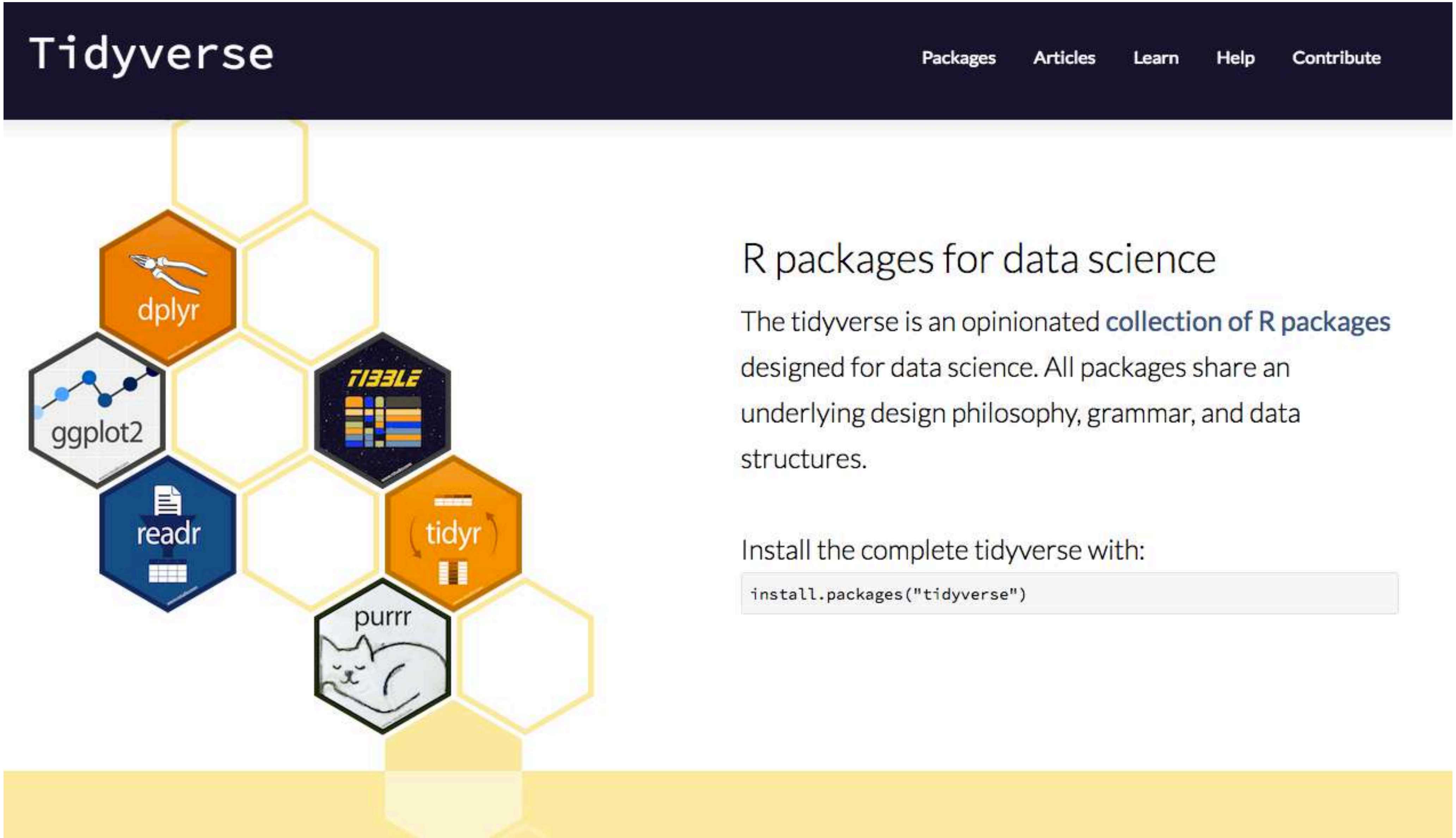
You'll also need to install some R packages. An R **package** is a collection of functions, data, and documentation that extends the capabilities of base R. Using packages is key to the successful use of R. The majority of the packages that you will learn in this book are part of the so-called tidyverse. The packages in the tidyverse share a common philosophy of data and R programming, and are designed to work together naturally.

You can install the complete tidyverse with a single line of code:

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

http://r4ds.had.co.nz/introduction.html#rstudio

Tidyverse package



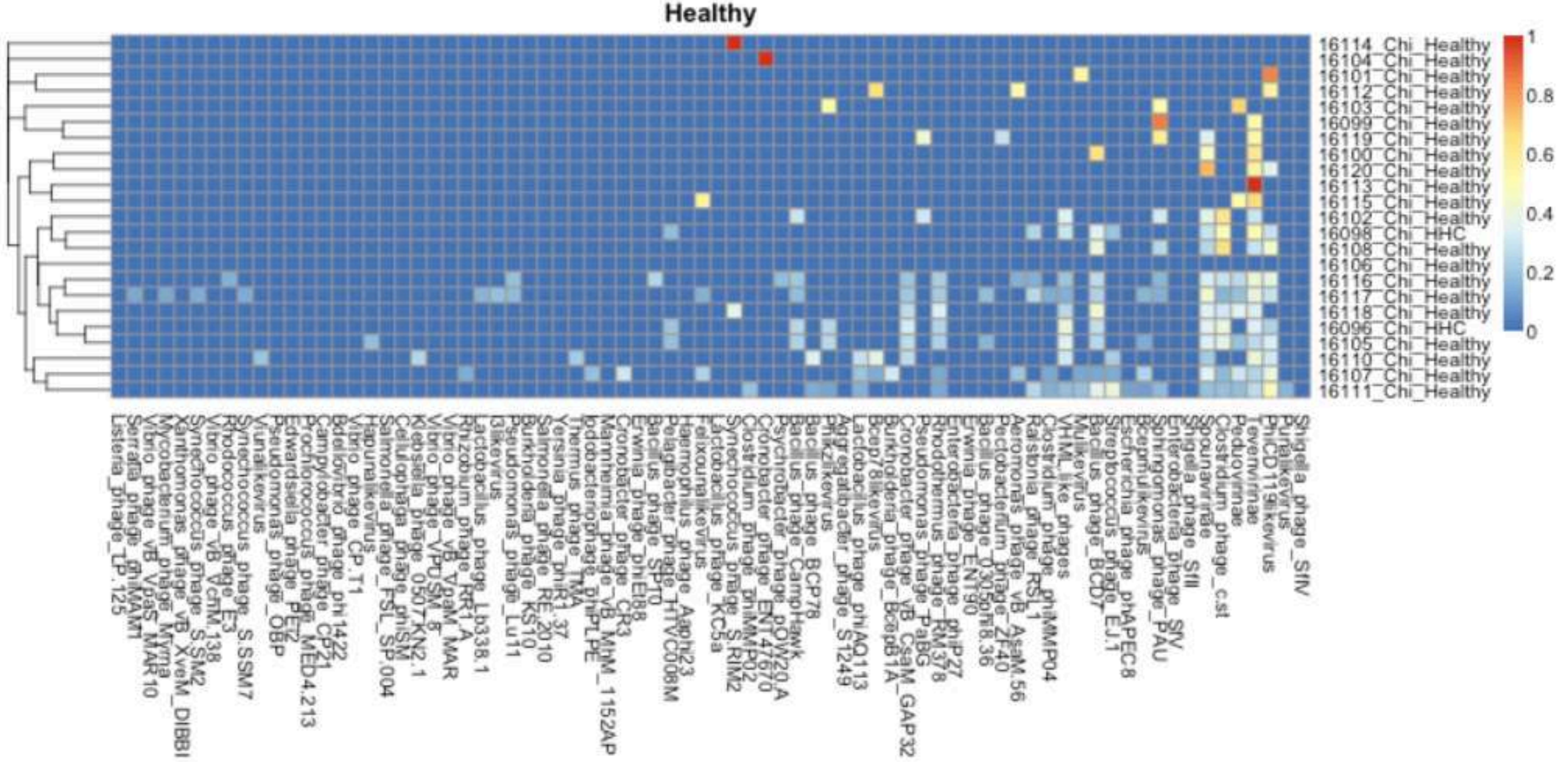
R packages for data science

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")
```

Example I



```
library("pheatmap")
library("vegan")
healthy <- read.table(“myoviridae_healthy.txt")
healthy_hellinger <- decostand(healthy, method="hellinger")
pheatmap(healthy_hellinger, cluster_cols=FALSE, cellwidth=8, cellheight=8, main="Healthy")
```

Case study one (iris)

The data set consists of **50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).** Four features were measured from each sample: the **length and the width of the sepals and petals, in centimetres**. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

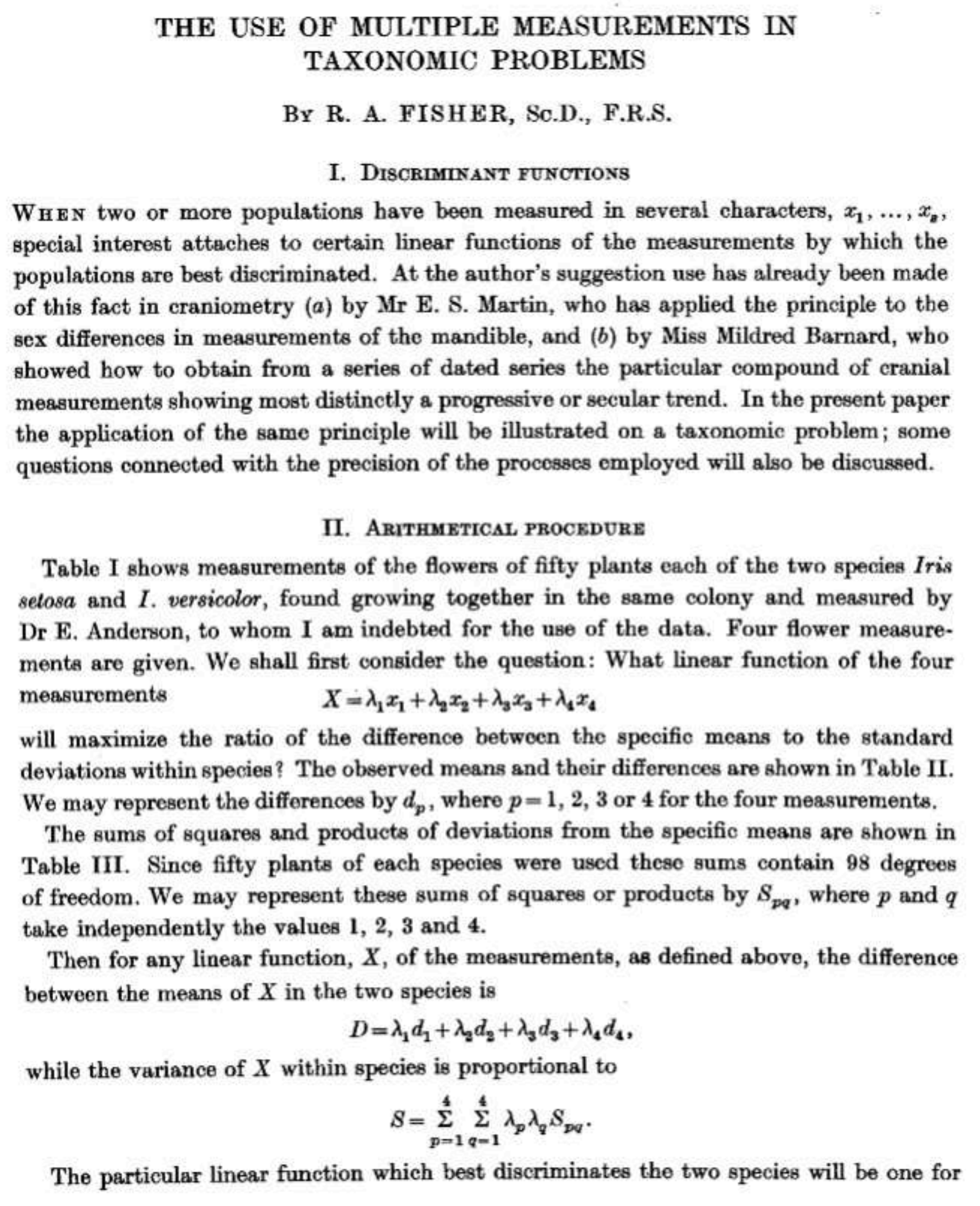
This data set became a typical test case for many statistical classification techniques in machine learning such as support vector machines

https://en.wikipedia.org/wiki/Iris_flower_data_set

Iris setosa

Iris versicolor

Iris virginica



THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

I. Discriminant functions

When two or more populations have been measured in several characters, $x_1, \ldots, x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (*a*) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (*b*) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.

II. Arithmetical procedure

Table I shows measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I. versicolor*, found growing together in the same colony and measured by Dr E. Anderson, to whom I am indebted for the use of the data. Four flower measurements are given. We shall first consider the question: What linear function of the four measurements

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

will maximize the ratio of the difference between the specific means to the standard deviations within species? The observed means and their differences are shown in Table II. We may represent the differences by d_p, where $p = 1, 2, 3$ or 4 for the four measurements.

The sums of squares and products of deviations from the specific means are shown in Table III. Since fifty plants of each species were used these sums contain 98 degrees of freedom. We may represent these sums of squares or products by S_{pq}, where p and q take independently the values 1, 2, 3 and 4.

Then for any linear function, X, of the measurements, as defined above, the difference between the means of X in the two species is

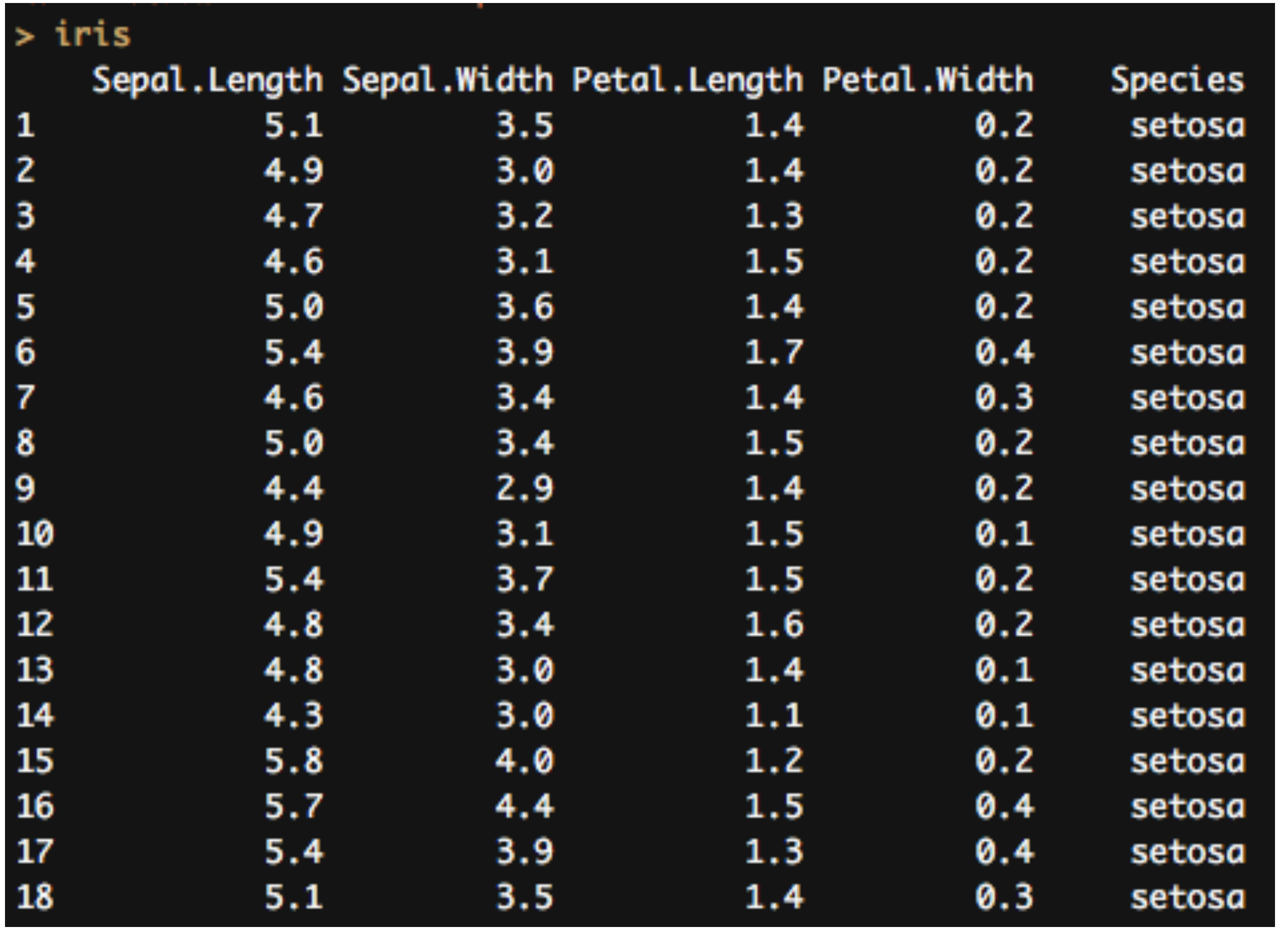
$$D = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 + \lambda_4 d_4,$$

while the variance of X within species is proportional to

$$S = \sum_{p=1}^{4} \sum_{q=1}^{4} \lambda_p \lambda_q S_{pq}.$$

The particular linear function which best discriminates the two species will be one for

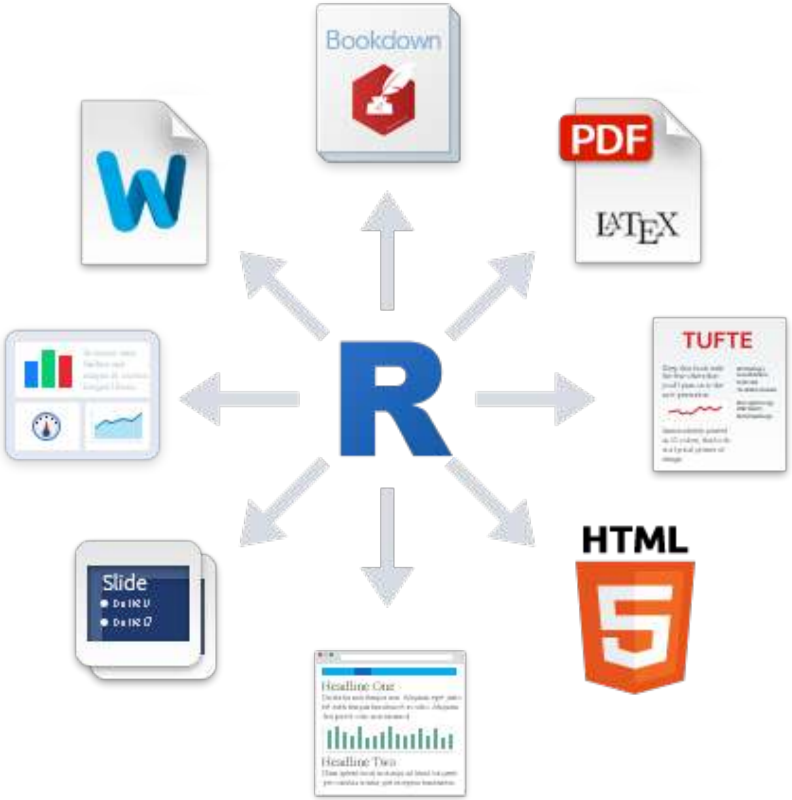
Case study one (iris)

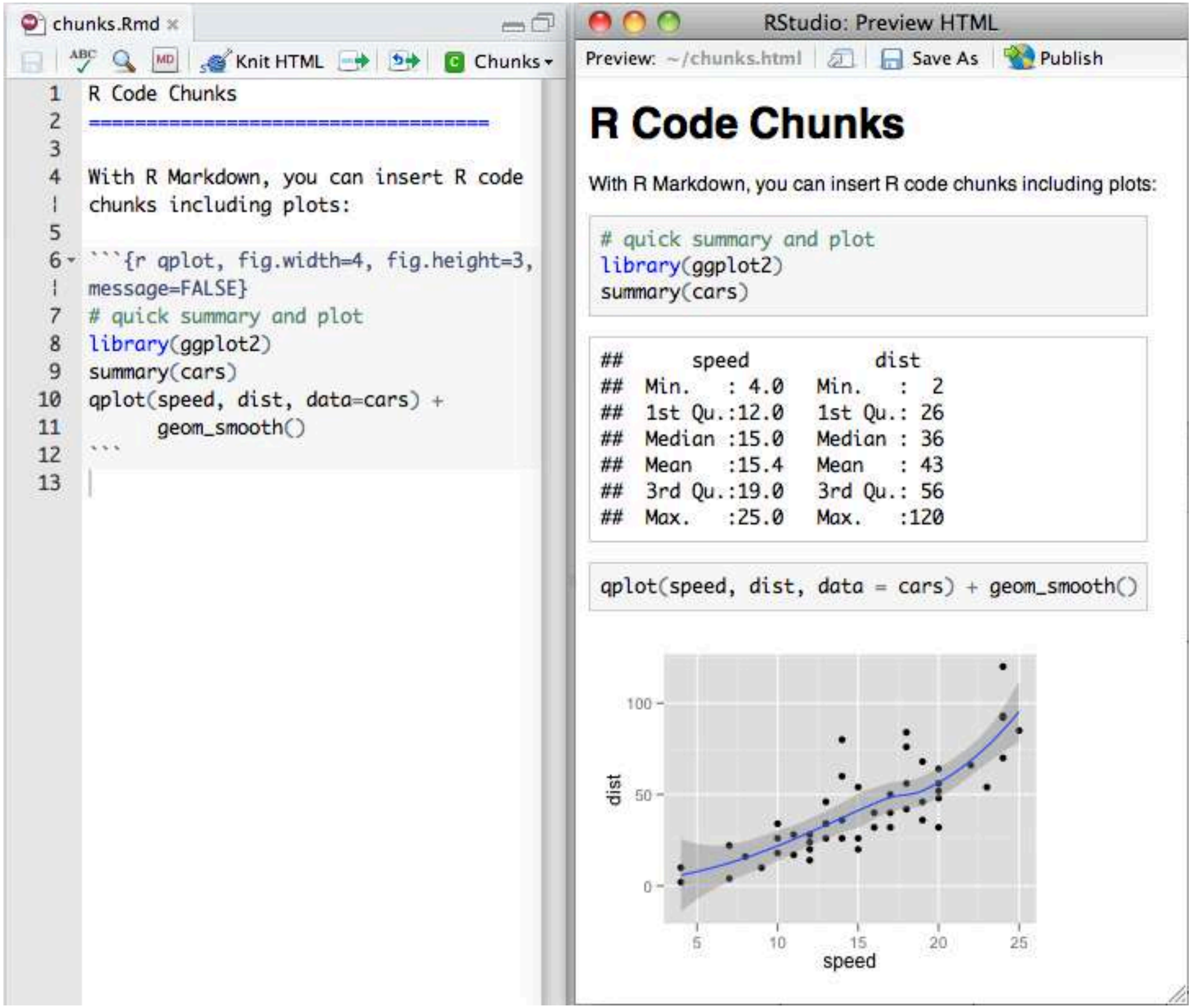


```
> iris
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1           5.1         3.5          1.4         0.2     setosa
2           4.9         3.0          1.4         0.2     setosa
3           4.7         3.2          1.3         0.2     setosa
4           4.6         3.1          1.5         0.2     setosa
5           5.0         3.6          1.4         0.2     setosa
6           5.4         3.9          1.7         0.4     setosa
7           4.6         3.4          1.4         0.3     setosa
8           5.0         3.4          1.5         0.2     setosa
9           4.4         2.9          1.4         0.2     setosa
10          4.9         3.1          1.5         0.1     setosa
11          5.4         3.7          1.5         0.2     setosa
12          4.8         3.4          1.6         0.2     setosa
13          4.8         3.0          1.4         0.1     setosa
14          4.3         3.0          1.1         0.1     setosa
15          5.8         4.0          1.2         0.2     setosa
16          5.7         4.4          1.5         0.4     setosa
17          5.4         3.9          1.3         0.4     setosa
18          5.1         3.5          1.4         0.3     setosa
```

https://rpubs.com/koushikstat/167274

R markdown in Rstudio





~~~
R Code Chunks
=====================================

With R Markdown, you can insert R code
chunks including plots:

```{r qplot, fig.width=4, fig.height=3,
message=FALSE}
# quick summary and plot
library(ggplot2)
summary(cars)
qplot(speed, dist, data=cars) +
      geom_smooth()
```
~~~

R Code Chunks

With R Markdown, you can insert R code chunks including plots:

```
# quick summary and plot
library(ggplot2)
summary(cars)
```

```
##       speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

```
qplot(speed, dist, data = cars) + geom_smooth()
```

https://rmarkdown.rstudio.com/r_notebooks.html

Shiny from R Studio

Interact. Analyze. Communicate.

Take a fresh, interactive approach to telling your data story with Shiny. Let users interact with your data and your analysis. And do it all with R.

https://gallery.shinyapps.io/001-hello/

https://shiny.rstudio.com/gallery/genome-browser.html

https://shiny.rstudio.com/gallery/

Coronavirus examples

An R Package to Explore the Novel Coronavirus

Patrick Tung Follow

Feb 11 · 11 min read ★

https://towardsdatascience.com/an-r-package-to-explore-the-novel-coronavirus-590055738ad6

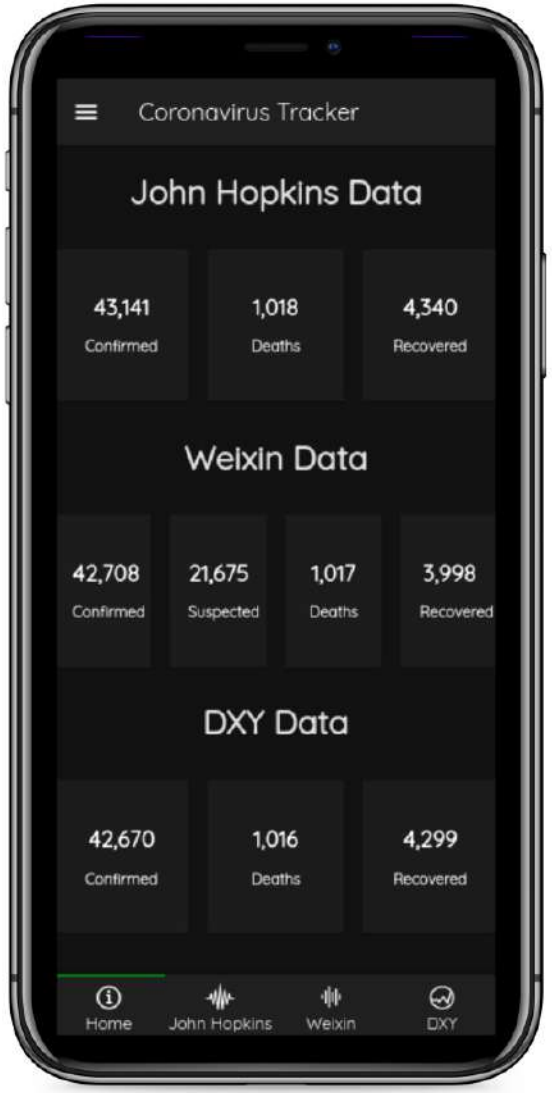
检索疫情数据的R包来了！

原创 Y叔叔 biobabble 2月3日

https://mp.weixin.qq.com/s/bPXdOGFzFK5dWLTEOEJB3g

https://mp.weixin.qq.com/s/_0D8ENb-4IGm4UV16Ok28A

Coronavirus Shiny example



Coronavirus Tracker

John Hopkins Data

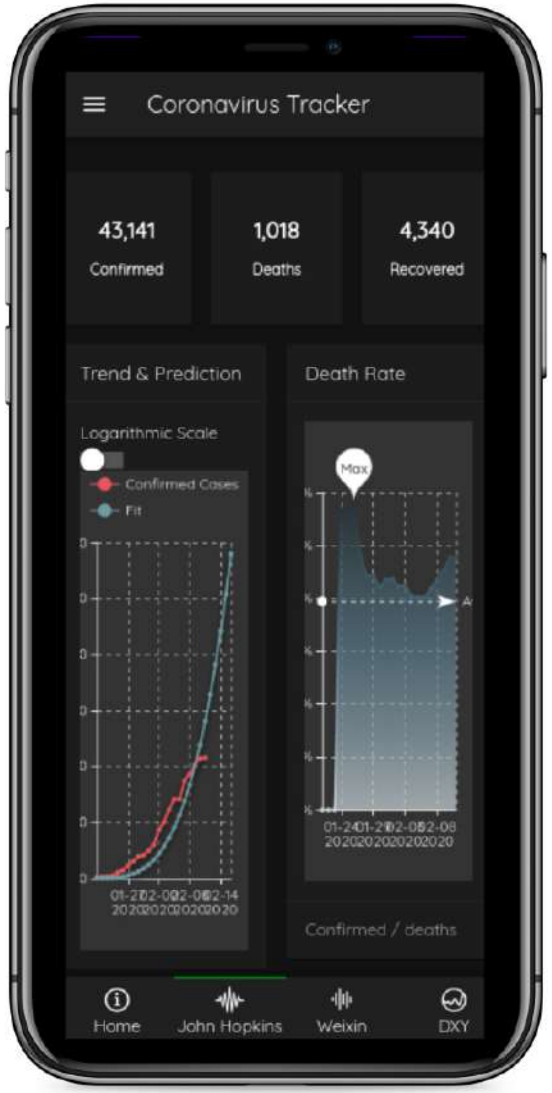
43,141	1,018	4,340
Confirmed	Deaths	Recovered

Weixin Data

42,708	21,675	1,017	3,998
Confirmed	Suspected	Deaths	Recovered

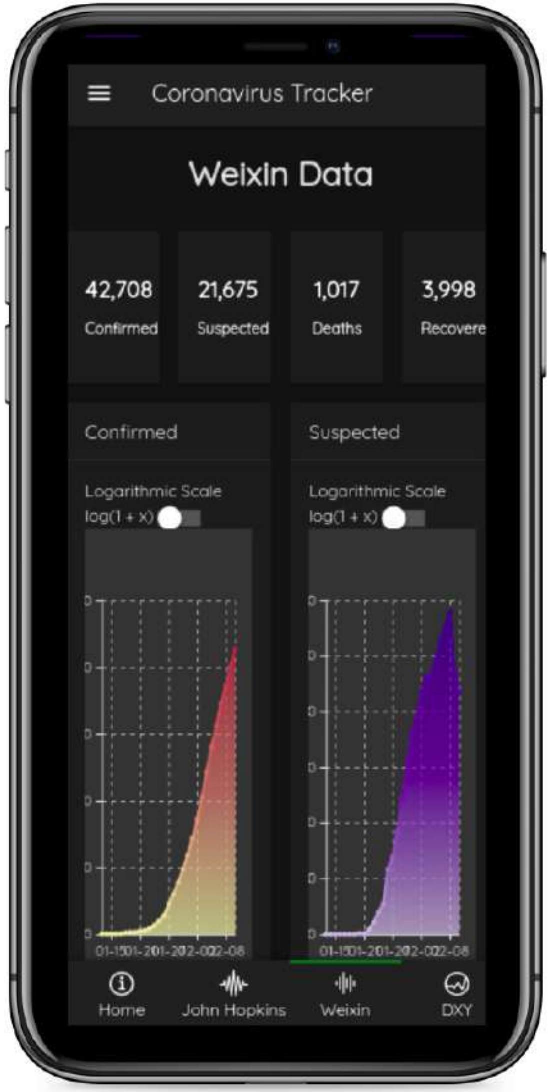
DXY Data

42,670	1,016	4,299
Confirmed	Deaths	Recovered



Coronavirus Tracker

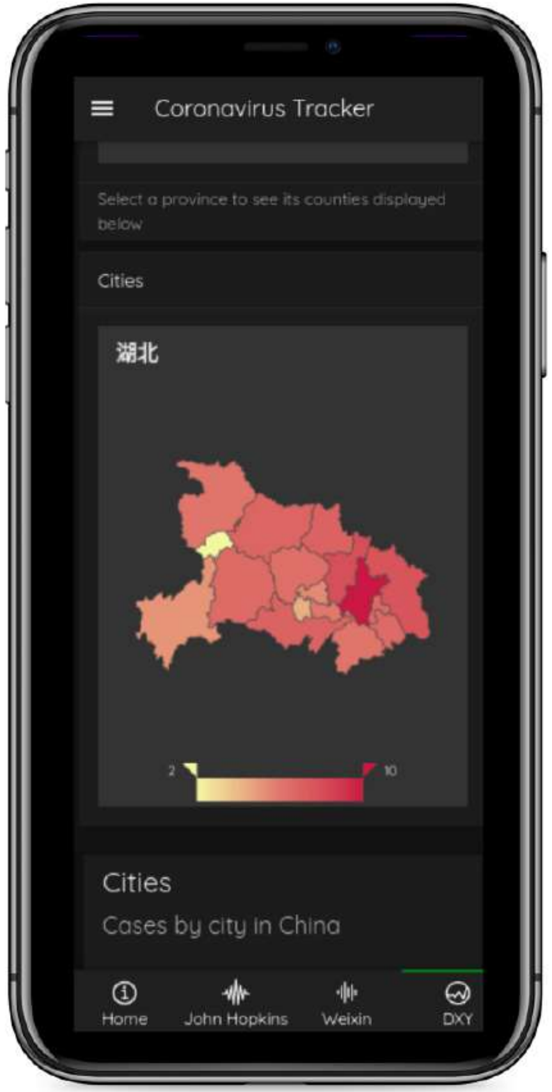
43,141	1,018	4,340
Confirmed	Deaths	Recovered



Coronavirus Tracker

Weixin Data

42,708	21,675	1,017	3,998
Confirmed	Suspected	Deaths	Recovere



Coronavirus Tracker

Select a province to see its counties displayed below

Cities

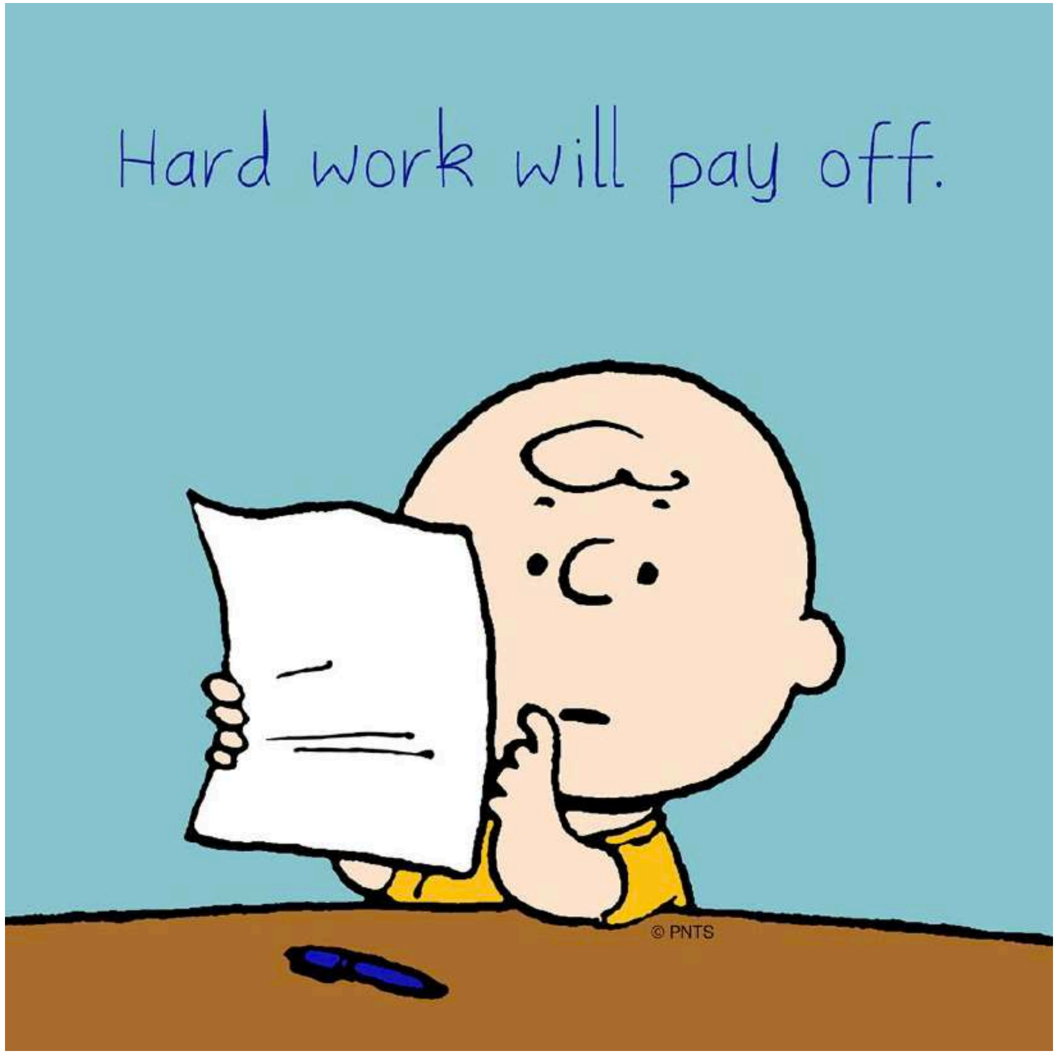
Cases by city in China

https://johncoene.shinyapps.io/contest-coronavirus/_w_5602e0c8/
https://community.rstudio.com/t/coronavirus-2020-shiny-contest-submission/53061

In summary

- Start practicing
- There are so much data out there
- Going through tutorials
- **Learn through real case scenarios**

- Think how to manage your notes and data effectively
- Research fast
- Reproducible research



Data sources

- https://data.gov.tw/
- http://fivethirtyeight.com/
- https://www.kaggle.com/

- All the various R datasets:
  - https://vincentarelbundock.github.io/Rdatasets/datasets.html
  - Iris is part of them