The three R's:

Research data management

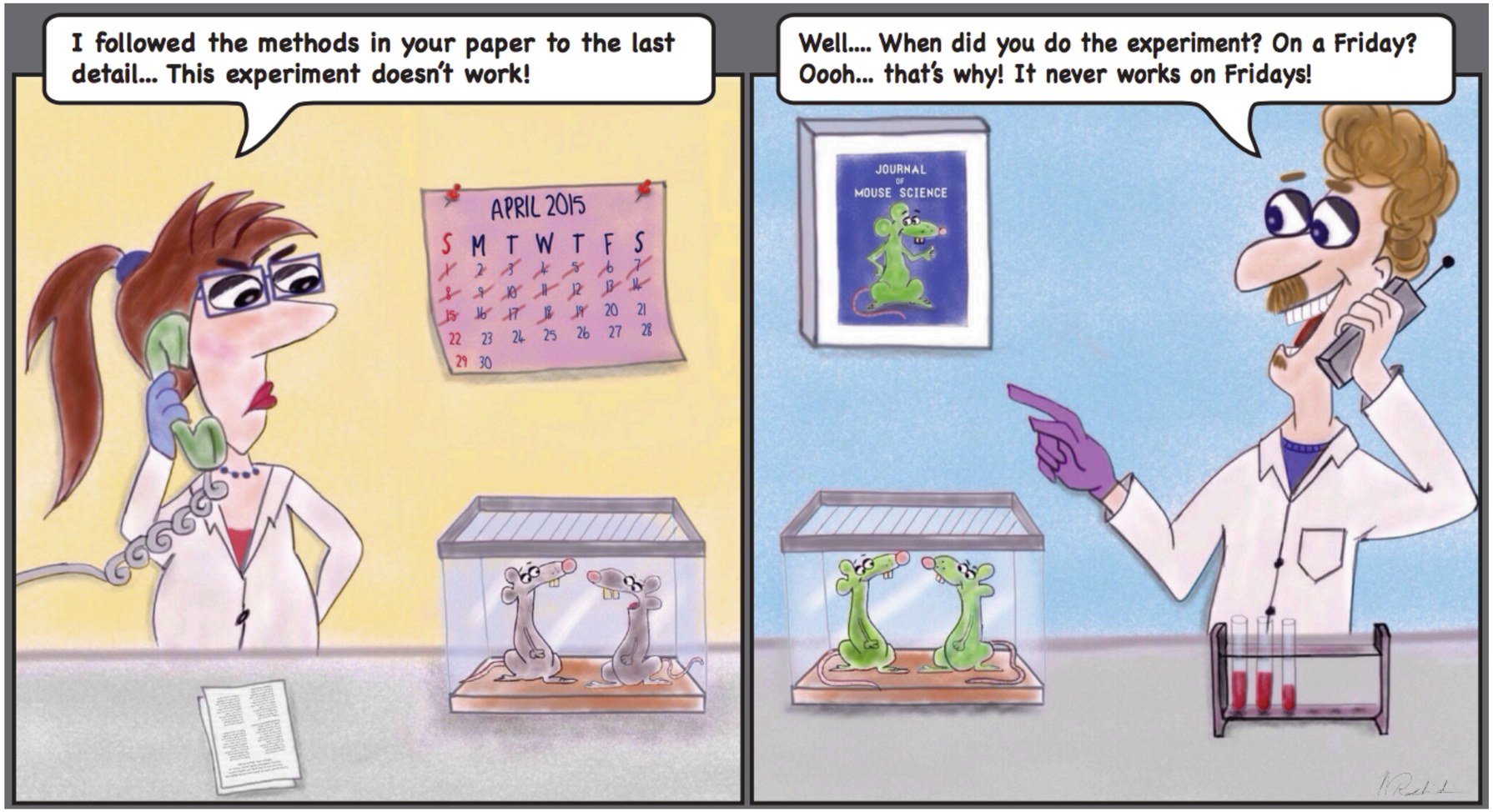
Reproducibility &

Reusability

Dr Jessica Stapley
jessica.stapley@env.ethz.ch

@jessstapley

Reproducibility is a major principle of the scientific method



Reproducibility versus Replication

Reproducibility

An independent researcher can obtain the same results using the original data and original computer code

Replication (Independent verification)

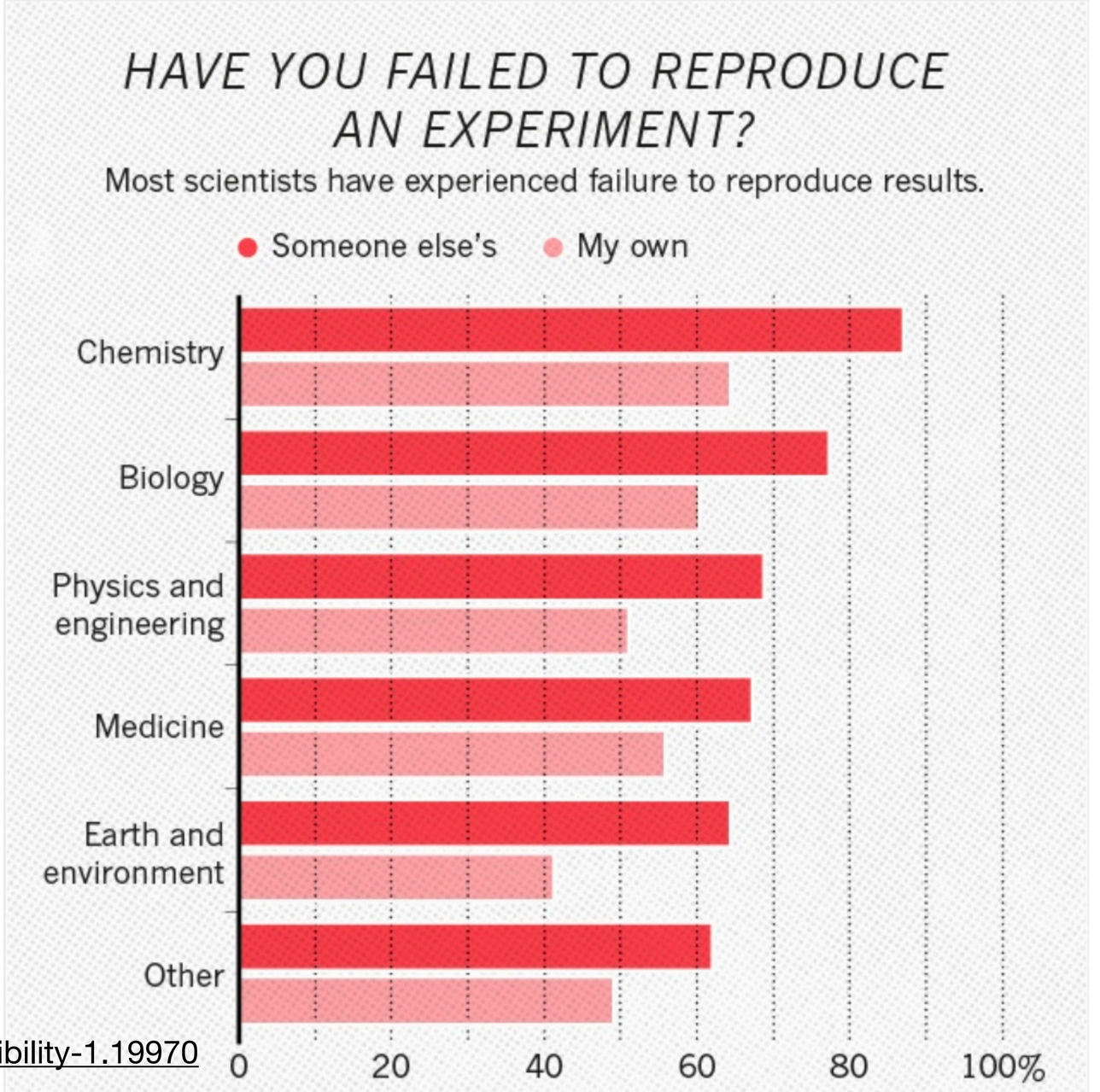
Different researcher/lab conducts an independent study and obtains the same results or comes to the same conclusion as the original study

But how reproducible are most scientific findings?

Not very

Survey of >1500 researchers by Nature

- >70% of researchers have tried and failed to reproduce another lab's experiments
- >50% failed to reproduce their own experiments

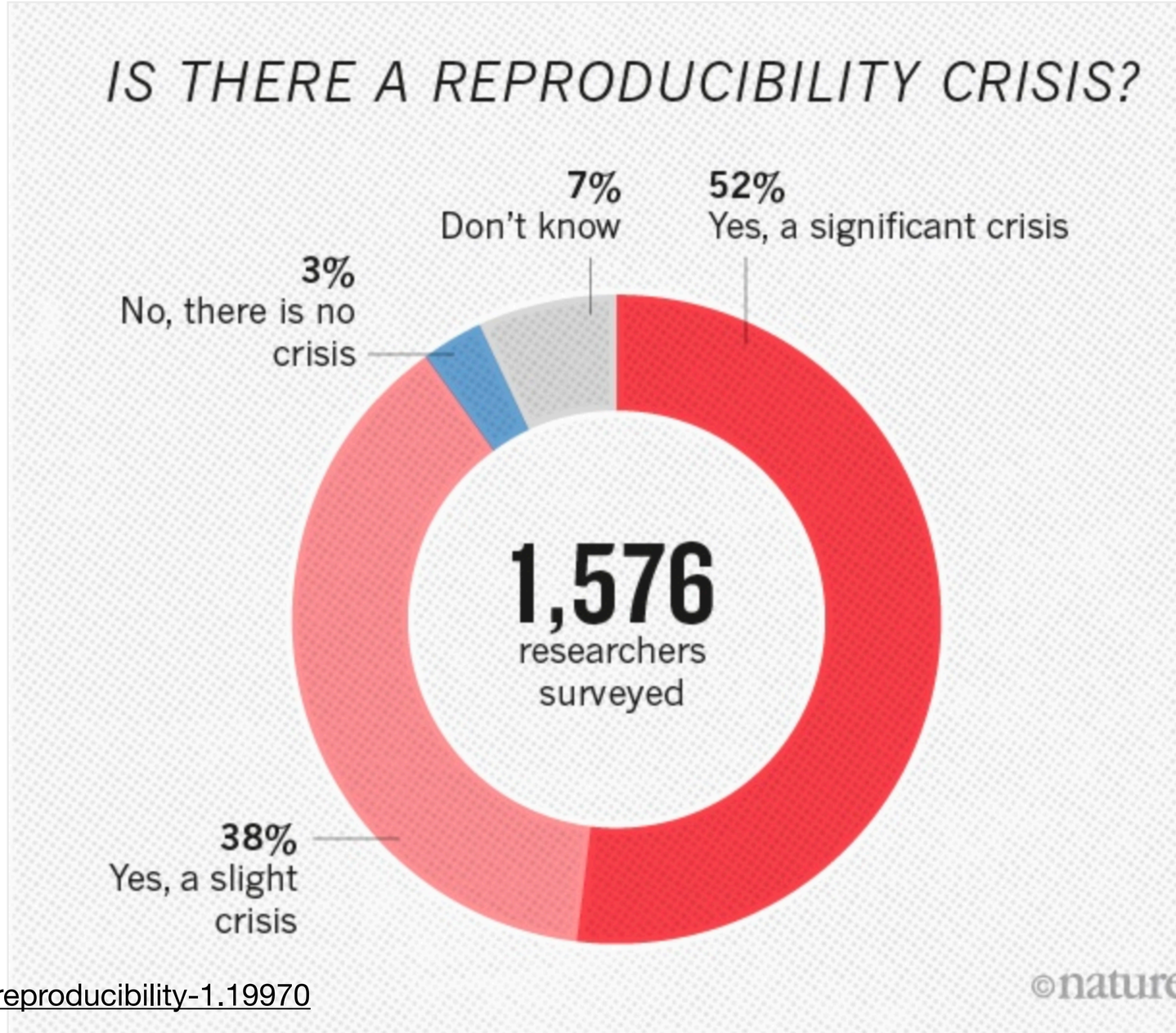


https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

But how reproducible are most scientific findings?

Not very

About half the respondents think there is a significant reproducibility crisis



https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

A similar problem has been found in other surveys and fields - this result has been replicated

AAAS Become a Member

Science Contents News Careers Journals

Read our COVID-19 research and news.

SHARE

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†

*All authors with their affiliations appear at the end of this paper.

†Corresponding author. E-mail: nosek@virginia.edu

– Hide authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

PLOS MEDICINE

BROWSE PUBLISH ABOUT

PDF

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Article | Authors | Metrics | Comments | Media Coverage

TheScientist
EXPLORING LIFE, INSPIRING INNOVATION

Search...

NEWS & OPINION MAGAZINE SUBJ

Home / News & Opinion

Half the Time, Psychology Results Not Reproducible: Study

These failures were not due to differences among sample populations.

Ashley P. Taylor
Nov 20, 2018

A large-scale effort to repeat psychological experiments has failed to confirm the results about half the time, according to a study published yesterday (November 19) on the pre-print server *PsyArXiv* and scheduled for publication in *Advances in Methods and Practices in Psychological Science*. According to the manuscript, the failures were not due to differences in the study populations between the original experiments and the replications, *Nature* reports.

ABOVE: © ISTOCK.COM, BILJANA CVETANOVIC

Comment | Open Access | Published: 28 October 2015

Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum

Shinichi Nakagawa & Timothy H. Parker

BMC Biology **13**, Article number: 88 (2015) | Cite this article

5245 Accesses | **46** Citations | **19** Altmetric | Metrics

Why is research not reproducible?

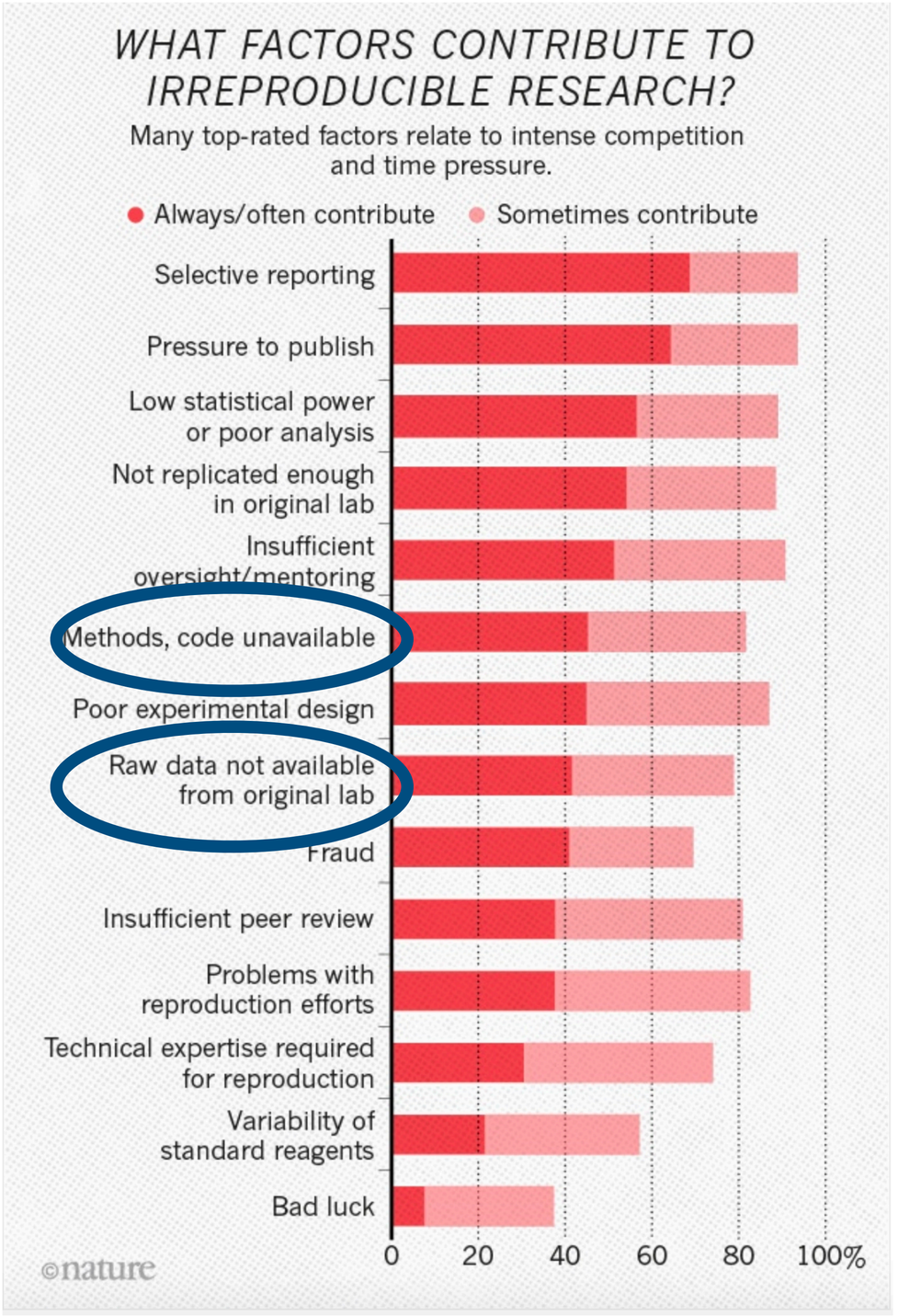
Major factors relate to some questionable activities

- selective reporting - p-hacking
- pressure to publish

Methods and code not available

- poorly described exact procedure can not be repeated

Raw data not available



https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Why make research reproducible and data reusable?

- **Reproducible**
  - Irreproducible research erodes trust in science
  - False findings lead to wasted resources and time
- **Data reuse**
  - Reusing data saves time and resources
  - Publicly funded data belongs to the public

FAIR Guiding principles for scientific data

Need to improve guidance and infrastructure to support data reuse

- multiple stakeholders developed a set of principles: **FAIR** Data principles
- **F**indable, **A**ccessible, **I**nteroperable, **R**eusable

Funding agencies and publishers require that research data be **FAIR**

Open Research Data

FNSNF Swiss National Science Foundation

Research data should be freely accessible to everyone – for scientists as well as for the general public.

The SNSF agrees with this principle. Since October 2017, researchers have to include a data management plan (DMP) in their funding application for most of the funding schemes. At the same time, the SNSF expects that data generated by funded projects are publicly accessible in digital databases provided there are no legal, ethical, copyright or other issues.

Please consult the webpages of the different funding schemes to see whether a DMP is required when submitting an application.

SPRINGER NATURE

Open research | About | Journals & books | Data | Institutional agreements | Funding & support | Publ

ELSEVIER About Elsevier Products & Solutions

Home > Authors > Author resources > Research Data > Open Data

Open Data

We believe that data should be open, accessible and reusable. Data sharing helps speed up the pace of discovery and its benefits to society.

As a proactive partner to the research community, Springer Nature is pioneering new approaches to data sharing and open data. We're committed to supporting researchers in sharing their data, helping to make data sharing the new normal.

What is open data?

Elsevier supports the principle that "Raw research data should be made freely available to all researchers" and authors should be free to publically post their raw research data (see Author Rights for more details).

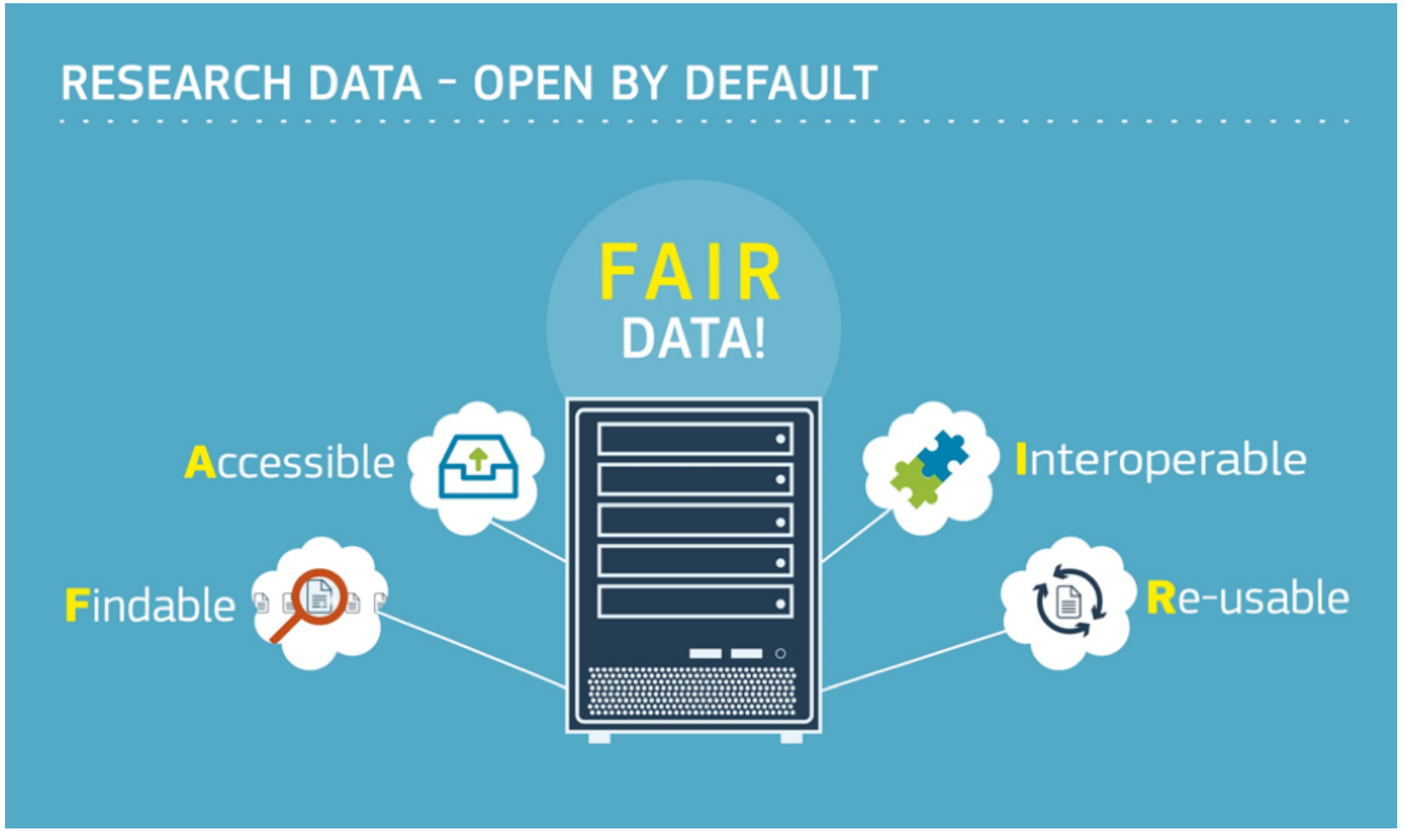
We have developed a simple way for authors to do this, by making their research data available on Mendeley Data and linking it to their article on ScienceDirect.

SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18

Jessica Stapley

ACCESSIBLE
Data and metadata is hosted on a public repository

INTEROPERABLE
Uses standard accessible language and methods. Uses open source software



FINDABLE
Globally unique and eternally persistent identifier, e.g. Digital object identifiers (DOI)

REUSABLE
Data is open, or released under clear and accessible data usage license

F

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

A

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

I

I1. (meta)data use a formal, accessible, shared, and broadly applicable language
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

R

R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18

Direct benefits of reproducible research

- **Benefits you**
  - You will need to reproduce your results

A Beginner's Guide to Conducting Reproducible Research

Jesse M. Alston[1,2] and Jessica A. Rick[1,3]

[1] *Program in Ecology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[2] *Department of Zoology and Physiology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[3] *Department of Botany, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

https://ecoevorxiv.org/h5r6n/

Direct benefits of reproducible research

- **Benefits you**
  - You will need to reproduce your results

A Beginner's Guide to Conducting Reproducible Research

Jesse M. Alston[1,2] and Jessica A. Rick[1,3]

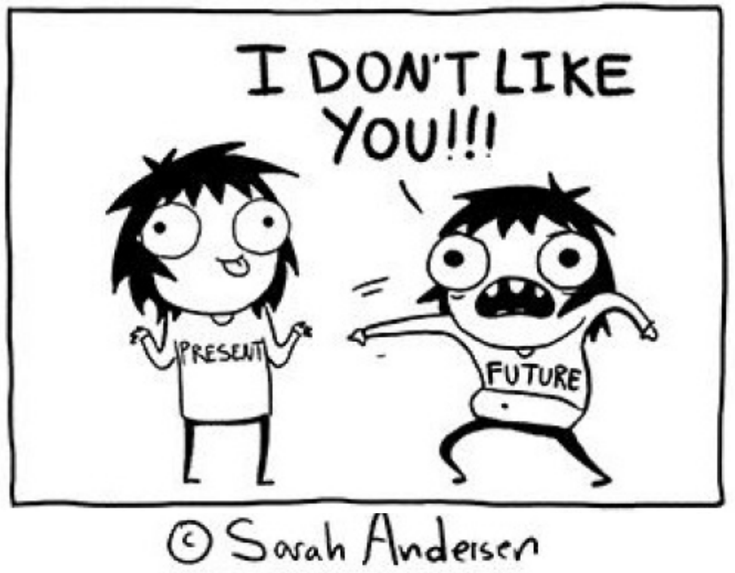
[1] *Program in Ecology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[2] *Department of Zoology and Physiology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[3] *Department of Botany, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

- The first person who will need to reproduce your results will probably be you
  - New data becomes available
  - Return to the project after some time
  - You give the project to student/collaborator
  - A reviewer suggests a change
  - You find an error and you don't know what went wrong

https://ecoevorxiv.org/h5r6n/



Direct benefits of reproducible research

- **Benefits you**
  - You will need to reproduce your results
  - It makes you look good

A Beginner's Guide to Conducting Reproducible Research

Jesse M. Alston[1,2] and Jessica A. Rick[1,3]

[1] *Program in Ecology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[2] *Department of Zoology and Physiology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[3] *Department of Botany, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

https://ecoevorxiv.org/h5r6n/

Direct benefits of reproducible research

- **Benefits you**
  - You will need to reproduce your results
  - It makes you look good
    - Indicator of researcher's rigor, trustworthiness and transparency
      - Reviewers can directly assess the analytical process
      - Catch errors before publication
      - Protects against accusations of research misconduct

A Beginner's Guide to Conducting Reproducible Research

Jesse M. Alston[1,2] and Jessica A. Rick[1,3]

[1] *Program in Ecology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[2] *Department of Zoology and Physiology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[3] *Department of Botany, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

https://ecoevorxiv.org/h5r6n/

Direct benefits of reproducible research

A Beginner's Guide to Conducting Reproducible Research

Jesse M. Alston[1,2] and Jessica A. Rick[1,3]

[1] *Program in Ecology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[2] *Department of Zoology and Physiology, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

[3] *Department of Botany, University of Wyoming, 1000 E University Dr., Laramie, WY 82072 USA*

https://ecoevorxiv.org/h5r6n/

- **Benefits you**
  - You will need to reproduce your results
  - It makes you look good
    - Indicator of researcher's rigor, trustworthiness and transparency
- **Benefits the community**
  - Others learn from you and can repeat the analysis in their own work
  - Researchers can reproduce and build on past work
  - Allows others to find/fix mistakes

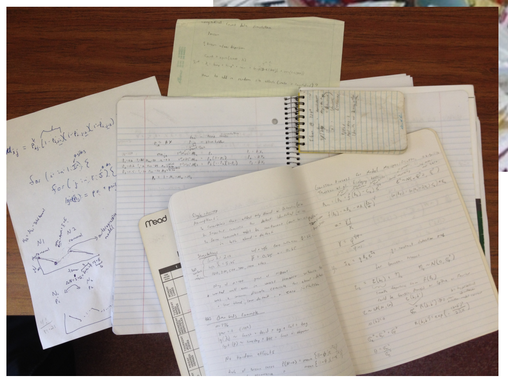
Making your data FAIR and improviong the reproducibility and reuse requires

Good data management

and

Computational reproducibility

Good data good management and organisation



This is better

You need to meticulously record everything, store your data safely and organise the data

“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why” - Bill Noble

Data management and Computational reproducibility

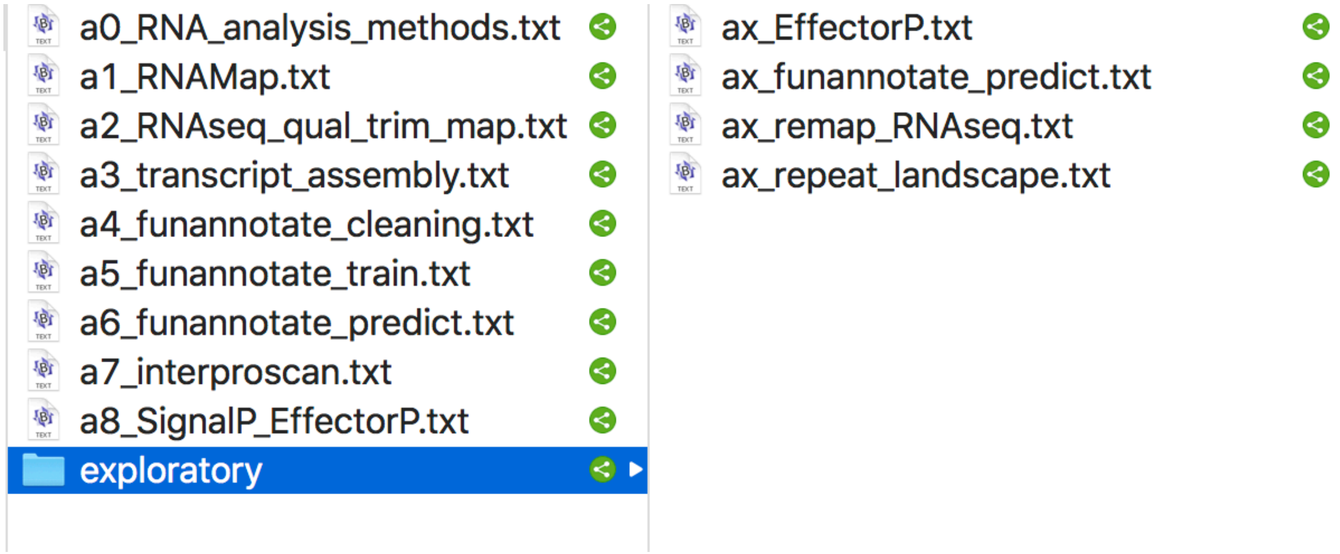
- Organise your data so it can be reproducible/reusable - by you or others

- Key components
  - single well structured directory with meaningful subdirectories

Directory structure

```
project
|   README.md
└──data/
|       └──raw_data/
|           |   data_orig.csv
|       └──processed_data/
|           |   data_clean.csv
|       └──results/
|           |   model_results.csv
└──documents/
|       meeting_notes.md
|       data_dictionary.md
└──code/
|       └──exploration/
|           |   01_data_exploration.Rmd
|           |   02_model_results.Rmd
|       └──scripts/
|           |   01_do_clean_data.R
|           |   02_do_model_data.R
|           └──functions/
|               |   01_funs_clean_data.R
|               |   02_funs_model_data.R
```

Example of my own analysis scripts directory



http://www.rebeccabarter.com/blog/2019-03-07_reproducible_pipeline/

Directory naming

Keep path names short (< 256 characters)

Recommendation for file names:
Unique, reflect content (if possible)

Use only ASCII (American Standard Code for Information Interchange) characters
- NO SPACES
- Be aware of case sensitivity

Bad examples

data%20management%20plan.docx
sup figure 2.png
lab meeting 19.10.2019.pptx

Good examples

Data_management_plan_SNF.docx
sup_figure_02_summary_stats.png
lab_meeting_2019-10-19.pptx

Data management and Computational reproducibility

- Organise your data so it can be reproducible/reusable - by you or others

- Key components
  - single well structured directory with meaningful subdirectories
  - data processing and analysis should carefully documented in interactive notebooks based on open access software (e.g. Jupyter notebook, R markdown)

Data processing and analysis

- Interactive Notebooks - combine documentation, code, input and output generated by the code, e.g. graphs)
- Open access, non proprietary - FREE
- Avoid WORD and EXCEL - not free, version incompatibilities, hidden characters, auto-filling, auto-formatting, scaleability

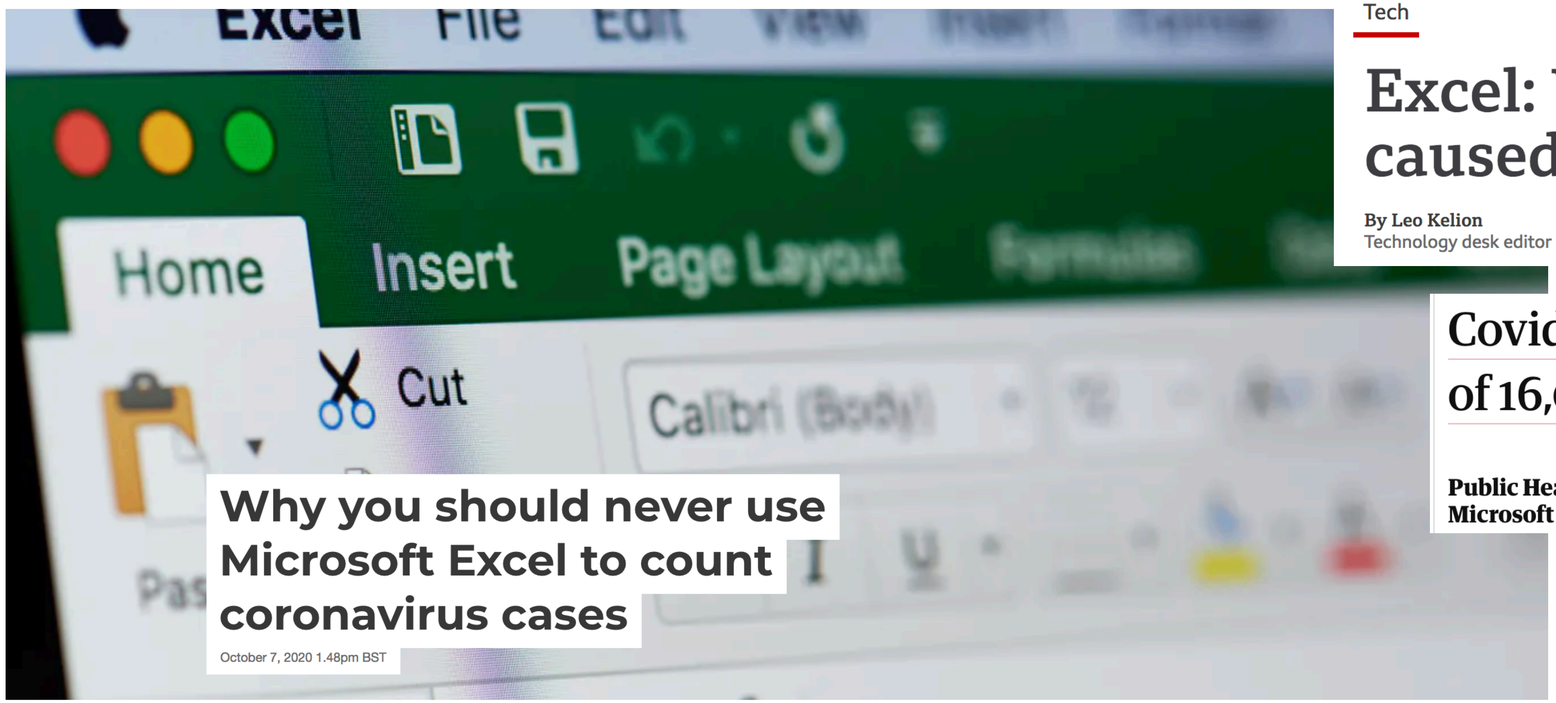
R Markdown

from RStudio

https://rmarkdown.rstudio.com/

https://jupyter.org/

It is OK, I have an excel spreadsheet with all my data



PixieMe/Shutterstock

Tech

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

By Leo Kelion
Technology desk editor

Covid: how Excel may have caused loss of 16,000 test results in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

- 16000 coronavirus cases were missing from daily reports between 25.09-02.10.2020
- Underestimate the scale of infections
- Delayed Track and Trace - possibly increasing community transmission and risk of infection

Data processing and analysis

- Interactive Notebooks - combine documentation, code, input and output generated by the code, e.g. graphs)
- Open access software - accessible
- Avoid WORD and EXCEL - Avoid WORD and EXCEL - not free, version incompatibilities, hidden characters, auto-filling, auto-formatting, scaleability
  - if you use excel - convert xls. files to files that are ‘open format’ not proprietary (e.g. csv, txt). Check the conversion of data, e.g. conversion to dates

R Markdown

from RStudio

https://rmarkdown.rstudio.com/

https://jupyter.org/

Data management and Computational reproducibility

- Organise your data so it can be reproducible/reusable - by you or others

- Key components
  - single well structured directory with meaningful subdirectories
  - data processing and analysis should carefully documented in interactive notebooks based on open access software (e.g. Jupyter notebook, R markdown)
  - accessible and hosted on a repository (e.g. github, DRYAD)

Computational reproducibility

Markdown documents - useful way to present text, code and figures

R projects and markdown - great for sharing

Github - place to store and share the code

Quick definition - Markup and markdown

Markup - system for annotating text documents, e.g. html - NOT EASY ON THE EYE

Markdown - lightweight markup language
- it uses plain-text syntax to be as unobtrusive as possible
- easily read by humans

Markup or HTML

```
<h1>Why <em>you</em> should use Markdown to write your next blog post</h1>
```

Markdown is just so dang legible, it will make your whole life easier. I promise.

Markdown

```
# Why *you* should use Markdown to write your next blog post

[Markdown][1] is just so dang legible, it will make your *whole life* easier. **I promise.**
```

Why markdown?

Markdown documents - useful way to present text, code and figures

Have you ever tried to put a graph or image in MS word?
How did that go?

Moving a picture in Microsoft word

Why markdown?

- Easy way to present text, figures and code
  - R markdown - share code, analysis and results easily
- Plain text, free to see!
  - does not require proprietary software
  - future-proof - standard, no outdated versions, e.g. MS Word has 8 different file types since 2018
- Easy to use (relatively)

Why not markdown?

- No spell checker

R Markdowns - html or pdf

RNASeqMappingStats

Genome size of sequenced genomes available on NCBI

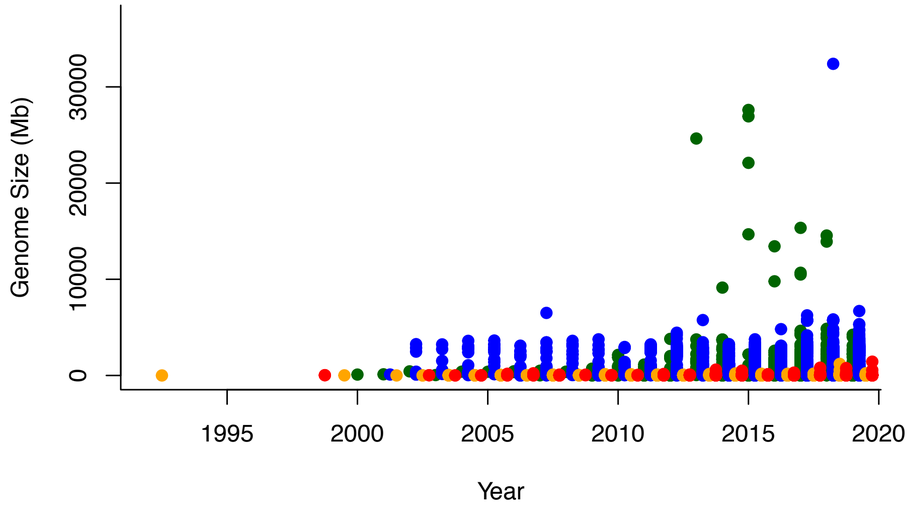
R Markdown

This is an R Markdown document to create plots of data on genome size from NCBI (as of 20.06.19). Here I plot genome size against the year it was released for all the genomes available on NCBI. I also plot the distribution of log genome size using a kernel denisty function.

Sequenced genome size over time for different taxonomic groups

Genomes on NCBI have been getting bigger. The Axolotl genome is the largest. Function to position and size the image from https://scrogster.wordpress.com/2014/06/02/adding-phylopic-org-silhouettes-to-r-plots/

```
par(bty="l")
plot(G.Size.Mb~year, typ="n", dat, ylab="Genome Size (Mb)", xlab="Year", ylim=c(0,37000))
points(G.Size.Mb~year, pch=19, subset(dat, dat$group1=="Plants"), col="dark green")
points(G.Size.Mb~I(year+0.25), pch=19, subset(dat, dat$group1=="Animals"), col="blue")
points(G.Size.Mb~I(year+0.5), pch=19, subset(dat, dat$group1=="Fungi"), col="orange")
points(G.Size.Mb~I(year+0.75), pch=19, subset(dat, dat$group1=="Protists"), col="red")
```



RNA Seq Data 2020

RNA from fungi grown in plant and on plates. In total 25 samples

Few reads mapped in plant compared to in culture, but the proportion of reads properly paired is higher in plant than culture

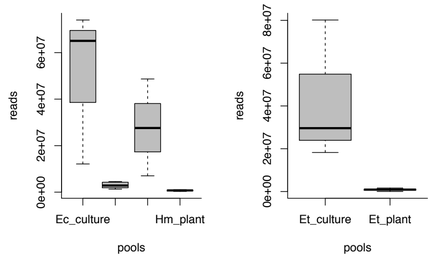
```
ec.id.stats  %>% group_by(pools) %>% summarise(mn.Kread=mean(reads)/1000, mn.pcp = mean(pc.paired))
```

```
## # A tibble: 4 x 3
##   pools      mn.Kread mn.pcp
## * <chr>         <dbl>  <dbl>
## 1 Ec_culture   50395.  0.458
## 2 Ec_plant      2954.  0.845
## 3 Hm_culture   27763.  0.482
## 4 Hm_plant       709.  0.861
```

```
et.id.stats  %>% group_by(pools) %>% summarise(mn.Kread=mean(reads)/1000, mn.pcp = mean(pc.paired))
```

```
## # A tibble: 2 x 3
##   pools      mn.Kread mn.pcp
## * <chr>         <dbl>  <dbl>
## 1 Et_culture   42660.  0.445
## 2 Et_plant       871.  0.706
```

```
par(mfrow=c(1,2), bty="l")
boxplot(reads~pools, ec.id.stats, col="grey")
boxplot(reads~pools, et.id.stats, col="grey")
```



Jessica Stapley

R Markdowns - pdf

```
---
title: "Genome size of sequenced genomes available on NCBI"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document to create plots of data on genome size from NCBI (as of 20.06.19). Here I plot genome size against the year it was released for all the genomes available on NCBI. I also plot the distribution of log genome size using a kernel denisty function.


```{r, echo=FALSE}
dat = read.table("data/NCBI_eukaryotes.txt", header = TRUE)
dat$l.Gsize = log(dat$G.Size.Mb)

```

Sequenced genome size over time for different taxonomic groups

Genomes on NCBI have been getting bigger. The Axolotl genome is the largest.
Function to position and size the image from https://scrogster.wordpress.com/2014/06/02/adding-phylopic-org-silhouettes-to-r-plots/

```{r}
par(bty="l")
plot(G.Size.Mb~year, typ="n", dat, ylab="Genome Size (Mb)", xlab="Year", ylim=c(0,37000))
points(G.Size.Mb~year, pch=19, subset(dat, dat$group1=="Plants"), col="dark green")
points(G.Size.Mb~I(year+0.25), pch=19, subset(dat, dat$group1=="Animals"), col="blue")
points(G.Size.Mb~I(year+0.5), pch=19, subset(dat, dat$group1=="Fungi"), col="orange")
points(G.Size.Mb~I(year+0.75), pch=19, subset(dat, dat$group1=="Protists"), col="red")


```
```

Genome size of sequenced genomes available on NCBI

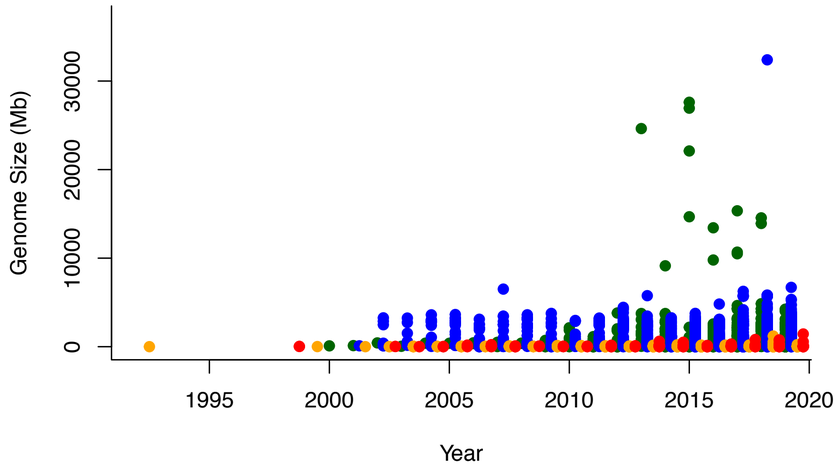
R Markdown

This is an R Markdown document to create plots of data on genome size from NCBI (as of 20.06.19). Here I plot genome size against the year it was released for all the genomes available on NCBI. I also plot the distribution of log genome size using a kernel denisty function.

Sequenced genome size over time for different taxonomic groups

Genomes on NCBI have been getting bigger. The Axolotl genome is the largest. Function to position and size the image from https://scrogster.wordpress.com/2014/06/02/adding-phylopic-org-silhouettes-to-r-plots/

```
par(bty="l")
plot(G.Size.Mb~year, typ="n", dat, ylab="Genome Size (Mb)", xlab="Year", ylim=c(0,37000))
points(G.Size.Mb~year, pch=19, subset(dat, dat$group1=="Plants"), col="dark green")
points(G.Size.Mb~I(year+0.25), pch=19, subset(dat, dat$group1=="Animals"), col="blue")
points(G.Size.Mb~I(year+0.5), pch=19, subset(dat, dat$group1=="Fungi"), col="orange")
points(G.Size.Mb~I(year+0.75), pch=19, subset(dat, dat$group1=="Protists"), col="red")
```



Jessica Stapley

R Markdowns - pdf

```
---
title: "RNASeqMappingStats"
output:
  html_document: default
  pdf_document: default
---
```

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
#setwd("/Users/stapleyj/polybox/Shared_Epichloe/annotation")
library(dplyr)
```
```

```
## RNA Seq Data 2020

RNA from fungi grown in plant and on plates. In total 25 samples
```

```
```{r load files, echo=FALSE}
id <- read.table("sample_info.txt", header=TRUE)
id <- id[,-2]
ec.stats <- read.table("Ec_RNASeqMap.stats")
et.stats <- read.table("Et_RNASeqMap.stats")
names(ec.stats) <- c("Full_ID", "reads", "paired")
names(et.stats) <- c("Full_ID", "reads", "paired")
et.id.stats <- droplevels(merge(et.stats, id))
ec.id.stats <- droplevels(merge(ec.stats, id))
et.id.stats$pc.paired <- et.id.stats$paired/et.id.stats$reads
ec.id.stats$pc.paired <- ec.id.stats$paired/ec.id.stats$reads
# save(list=ls(), file="RNASeqMapStats.RData")
```
```

```
## Few reads mapped in plant compared to in culture, but the proportion of reads properly paired is higher in plant than culture
```

```
```{r mapping, echo=TRUE}
ec.id.stats  %>% group_by(pools) %>% summarise(mn.Kread=mean(reads)/1000, mn.pcp = mean(pc.paired))
et.id.stats  %>% group_by(pools) %>% summarise(mn.Kread=mean(reads)/1000, mn.pcp = mean(pc.paired))

```
```

```
```{r plot mapping, echo=TRUE, out.width="50%"}
par(mfrow=c(1,2), bty="l")
boxplot(reads~pools, ec.id.stats, col="grey")
boxplot(reads~pools, et.id.stats, col="grey")

par(mfrow=c(1,2), bty="l")
boxplot(pc.paired~pools, ec.id.stats, col="grey")
boxplot(pc.paired~pools, et.id.stats, col="grey")
```
```

RNASeqMappingStats

RNA Seq Data 2020

RNA from fungi grown in plant and on plates. In total 25 samples

Few reads mapped in plant compared to in culture, but the proportion of reads properly paired is higher in plant than culture

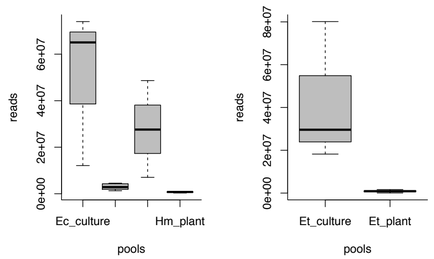
```
ec.id.stats  %>% group_by(pools) %>% summarise(mn.Kread=mean(reads)/1000, mn.pcp = mean(pc.paired))
```

```
## # A tibble: 4 x 3
##   pools      mn.Kread mn.pcp
## * <chr>         <dbl>  <dbl>
## 1 Ec_culture   50395.  0.458
## 2 Ec_plant      2954.  0.845
## 3 Hm_culture   27763.  0.482
## 4 Hm_plant       709.  0.861
```

```
et.id.stats  %>% group_by(pools) %>% summarise(mn.Kread=mean(reads)/1000, mn.pcp = mean(pc.paired))
```

```
## # A tibble: 2 x 3
##   pools      mn.Kread mn.pcp
## * <chr>         <dbl>  <dbl>
## 1 Et_culture   42660.  0.445
## 2 Et_plant       871.  0.706
```

```
par(mfrow=c(1,2), bty="l")
boxplot(reads~pools, ec.id.stats, col="grey")
boxplot(reads~pools, et.id.stats, col="grey")
```



Github Markdown

RNA seq data analysis

This document contains a step-by-step account of how we analyzed the RNA seq data, predicted putative genes using the funannotate pipeline https://funannotate.readthedocs.io/en/latest/ and annotated putative proteins using Inerproscan, signalP and effectorP.

Here we have provided examples of code for each step. The code chunks will not run automatically "as is" it will need to be edited by the user.

Here is a list of the software used (not including all dependencies) funnanotate v1.7.0, fastqc v0.11.4, cufflinks v2.1.1, star v2.5.3a, stringtie v1.3.3b, repeatmasker v4.0.6 , signalP v4.1, effectorP v2.0

Check RNA sequence read quality with `fastqc`

Example code

```
while read p; do
fastqc path_to_seq_data/RNAseq_${name}.fastq.gz -o fastqc_out/
done<RNAseq_sample.list
```

Sequences were not trimmed based on recommendations from this paper https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0956-2

Generate files for `star` and `funannotate`

Convert reference genome gff to gtf using `cufflinks v2.1.1`

```
gffread /path_to_file/Epichloe_clarkii.gff3 -T -F -o Epichloe_clarkii.gtf
```

Github Markdown

Edit file | Preview changes | Tabs | 8 | Soft wrap

```
# RNA seq data analysis
This document contains a step-by-step account of how we analyzed the RNA seq data, predicted putative genes using the funannotate pipeline
https://funannotate.readthedocs.io/en/latest/ and annotated putative proteins using Inerproscan, signalP and effectorP.

Here we have provided examples of code for each step. The code chunks will not run automatically "as is" it will need to be edited by the user.

Here is a list of the software used (not including all dependencies)
funnanotate v1.7.0, fastqc v0.11.4, cufflinks v2.1.1, star v2.5.3a, stringtie v1.3.3b, repeatmasker v4.0.6 , signalP v4.1, effectorP v2.0

## Check RNA sequence read quality with ```fastqc```
Example code

```
while read p; do
fastqc path_to_seq_data/RNAseq_${name}.fastq.gz -o fastqc_out/
done<RNAseq_sample.list

```
Sequences were not trimmed based on recommendations from this paper
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-0956-2

## Generate files for ```star``` and ```funannotate```

Convert reference genome gff to gtf using ```cufflinks v2.1.1```
```
gffread /path_to_file/Epichloe_clarkii.gff3 -T -F -o Epichloe_clarkii.gtf

```
## Clean the genome
```
funannotate clean -i Ety1756_Epichloe_typhina_1756_33930528_v4.fna -o Ety1756_clean.fna
```

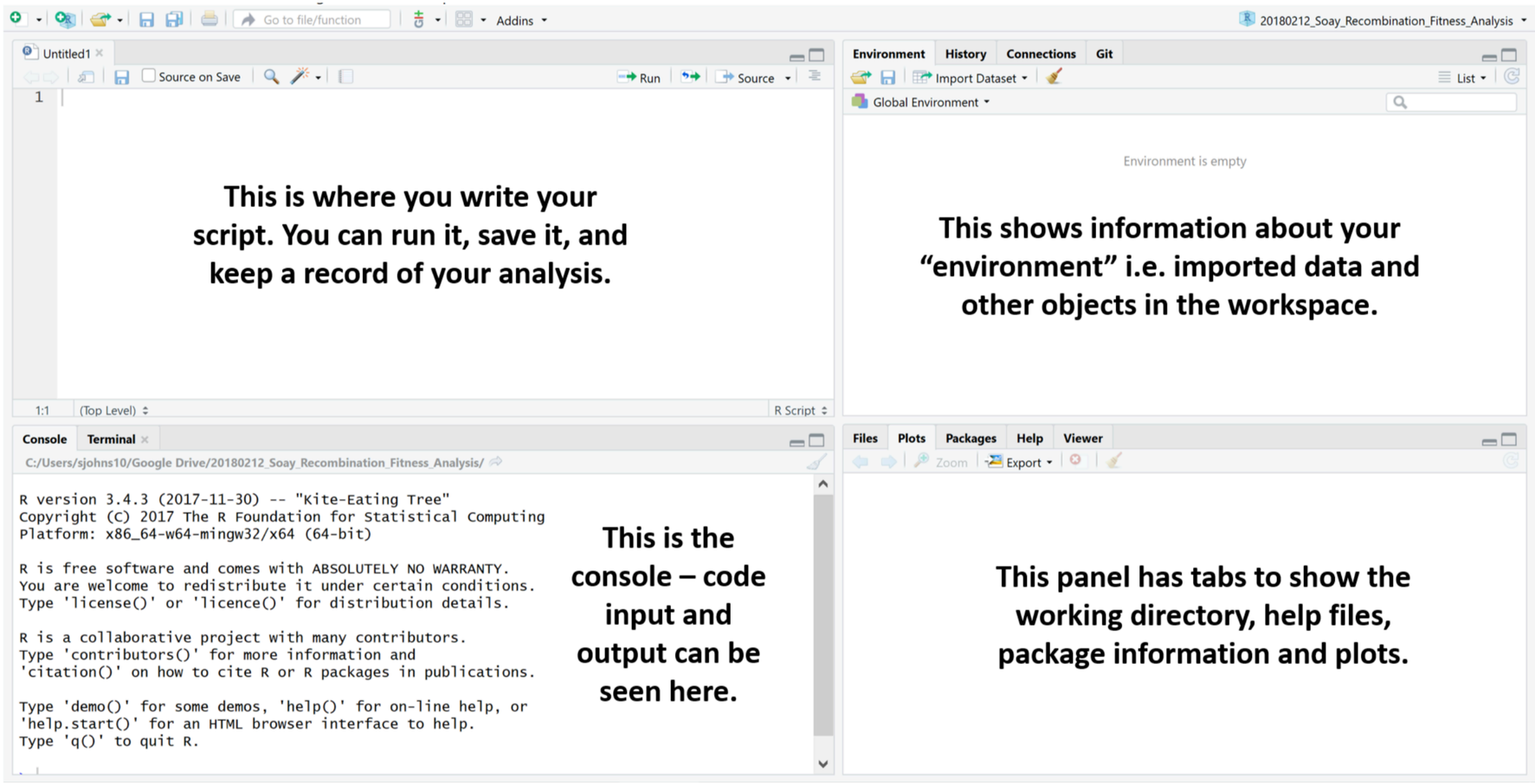
## Generate genome files for ```star```
```

Markdown documents - useful way to present text, code and figures

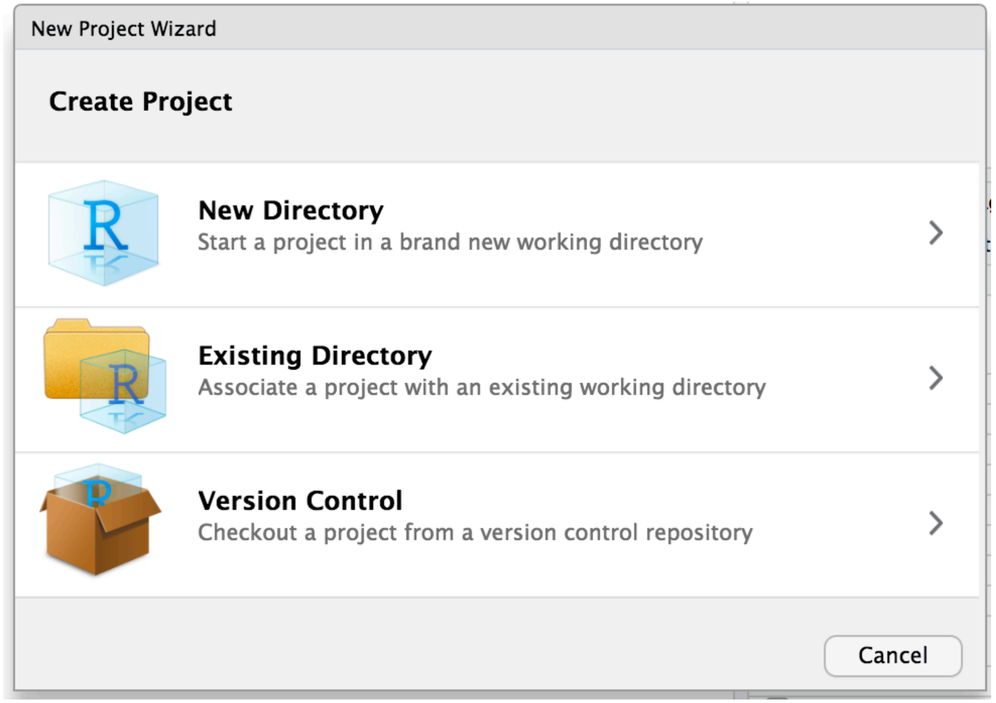
R projects and markdown - great for sharing

Github - place to store and share the markdowns

- RStudio is an integrated development environment (IDE) for R
- makes working with R easier - includes
  - drop down menus
  - syntax highlighting
  - code completion
  - smart indentation
  - workspace browser
  - data viewer
  - plot history
  - … and much more

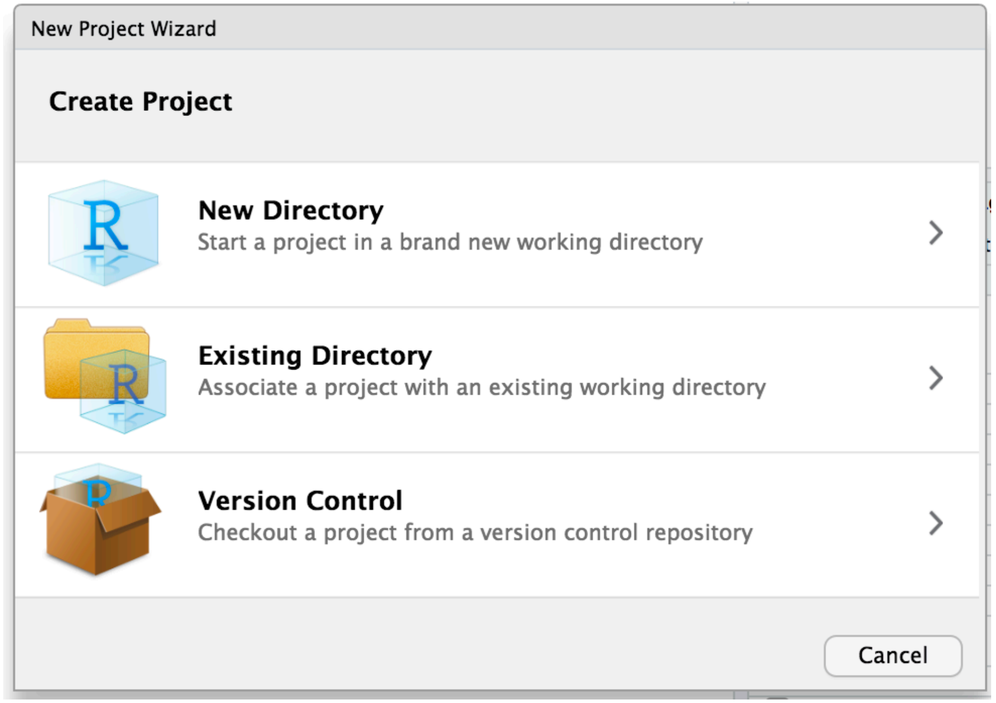


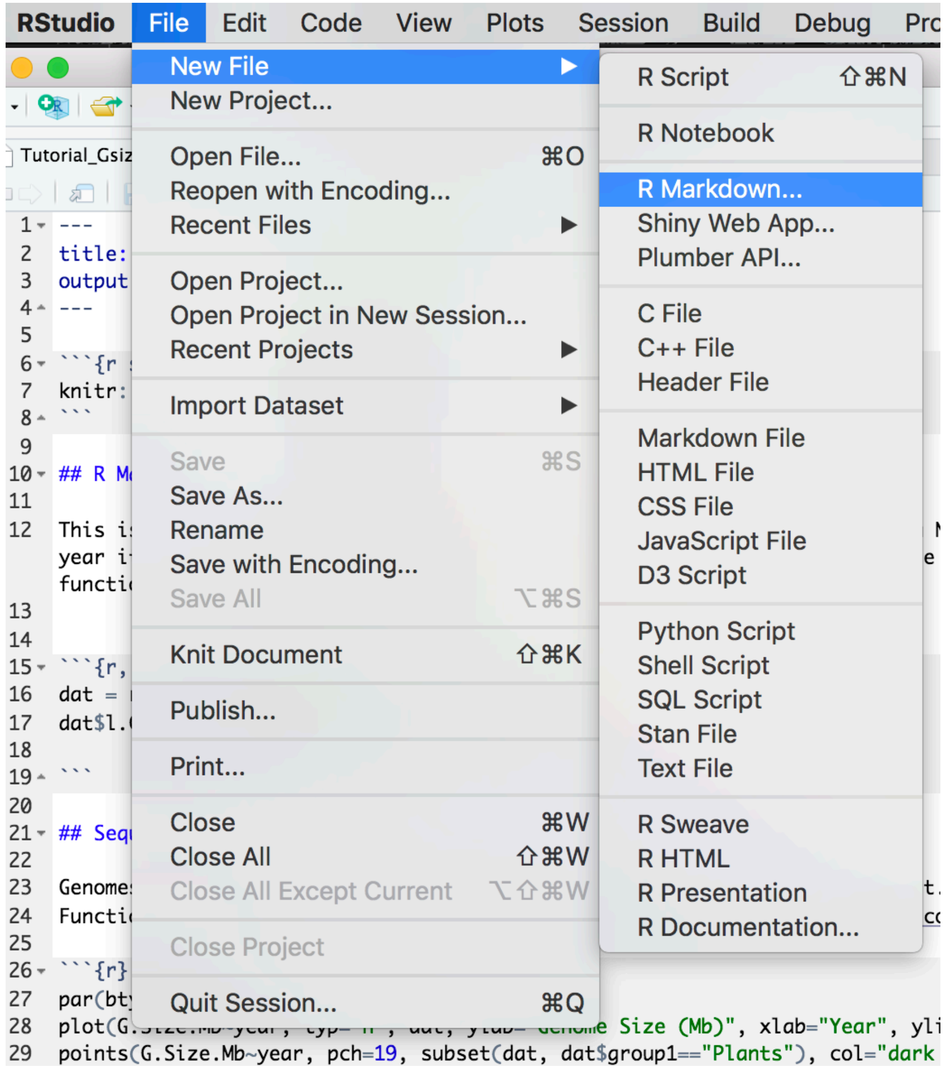
R studio - R projects and R Markdown



Jessica Stapley

R studio - R projects and R Markdown





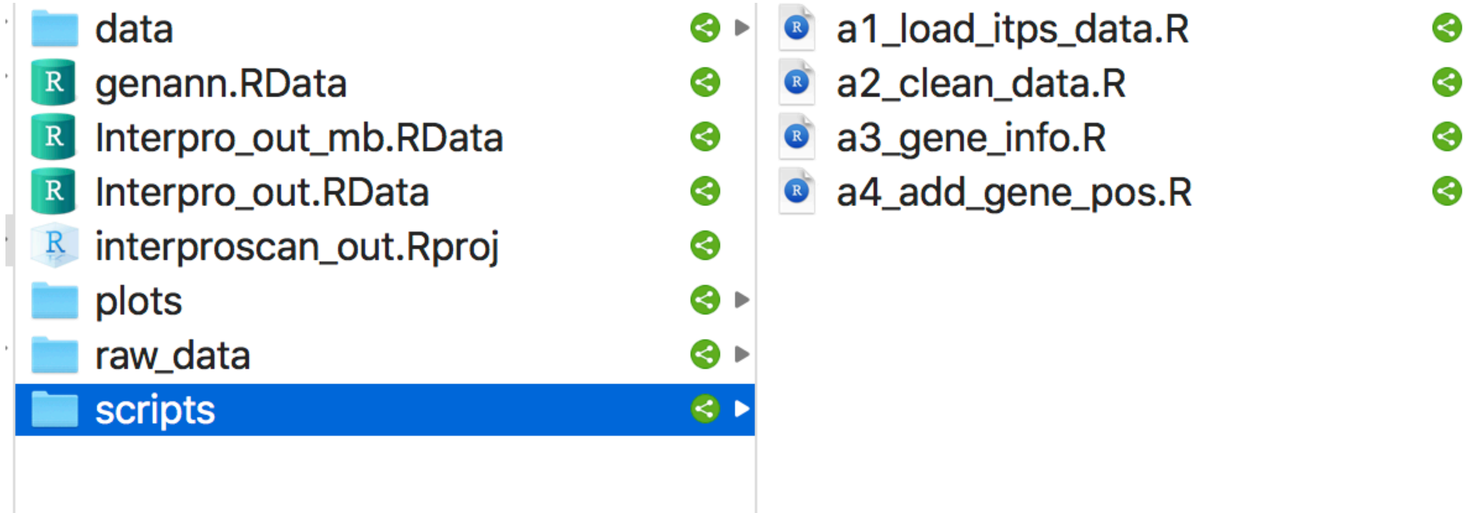
R projects

R projects

- creates a directory where all files and relevant metadata are kept
- keeps data reliable, portable, shareable and reproducible
- sets the environment (e.g.working directory) so the code will run on any machine

No single way to structure the project - as a minimum I have 3 directories

- (raw)data: contains data for project (read only protected)
- scripts: sequentially named so order can be easily seen
- plots: plots of the data



Github for hosting/sharing code

- Git is a version control system
  - software developers working in large teams
  - manages changes in files - rollback to previous versions
- Multiple people can work simultaneously
  - users create branches to work independently and these are merged later
  - all changes have traceable and can be annotated
- Happy git with R (https://happygitwithr.com/)

Hosting data permanently

dbSNP
http://www.ncbi.nlm.nih.gov/snp https://fairsharing.org/biodbcore-000438
Dryad Digital Repository
http://datadryad.org/ https://fairsharing.org/biodbcore-000464
European Nucleotide Archive ENA
http://www.ebi.ac.uk/ena/ https://fairsharing.org/biodbcore-000310
GenBank
http://www.ncbi.nlm.nih.gov/genbank/ https://fairsharing.org/biodbcore-000001
Mendeley Data
https://data.mendeley.com/ https://fairsharing.org/FAIRsharing.3epmpp
NCBI Assembly
http://www.ncbi.nlm.nih.gov/assembly
NCBI Sequence Read Archive SRA
http://www.ncbi.nlm.nih.gov/sra https://fairsharing.org/biodbcore-000444

Not all data is computational

Specimens should adhere to FAIR principles and be
- collected and curated properly
- made publically available

Herbaria/museums should be funded properly

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Biological science practices

Check for updates

Cite this article: Manzano S, Julier ACM. 2021 How FAIR are plant sciences in the twenty-first century? The pressing need for reproducibility in plant ecology and evolution. *Proc. R. Soc. B* **288**: 20202597. https://doi.org/10.1098/rspb.2020.2597

Received: 16 October 2020
Accepted: 13 January 2021

Subject Category:
Ecology

Subject Areas:
ecology, evolution, plant science

Keywords:
FAIR principles, plant ecology, plant evolution,

How FAIR are plant sciences in the twenty-first century? The pressing need for reproducibility in plant ecology and evolution

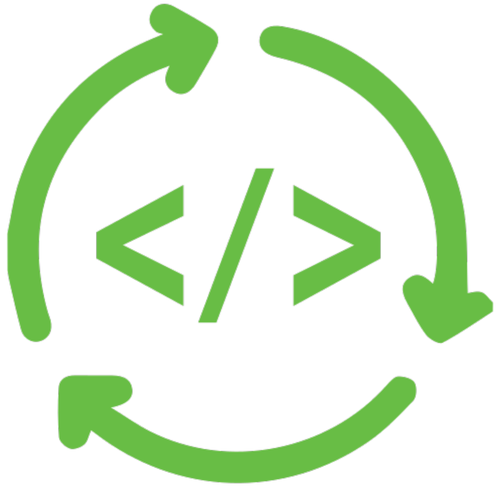
Saúl Manzano and Adele C. M. Julier

Plant Conservation Unit, Department of Biological Sciences, HW Pearson Building, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa

SM, 0000-0002-5720-2768

The need for open, reproducible science is of growing concern in the twenty-first century, with multiple initiatives like the widely supported FAIR principles advocating for data to be Findable, Accessible, Interoperable and Reusable. Plant ecological and evolutionary studies are not exempt from the need to ensure that the data upon which their findings are based are accessible and allow for replication in accordance with the FAIR principles. However, it is common that the collection and curation of herbarium specimens, a foundational aspect of studies involving plants, is neglected by authors. Without publicly available specimens, huge numbers of studies that rely on the field identification of plants are fundamentally not reproducible. We argue that the collection and public availability of herbarium specimens is not only good botanical practice but is also fundamental in ensuring that plant ecological and evolutionary studies are replicable, and thus scientifically sound. Data repositories that adhere to the FAIR principles must make sure that the original data are traceable to and re-examinable at their empirical source. In order to secure replicability, and adherence to the FAIR principles, substantial changes need to be brought about to restore the practice of collecting and curating specimens, to educate students of their importance, and to properly fund the herbaria which house them.

Summary



- We need to ensure our research is reproducible
- FAIR - guide to enhance the reusability of their data
- Recommended (required) by publishers and funding bodies
- Assess the FAIRness of your data (https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/)
- It all starts with good research data management and clear description of what you did and what versions of software you used.
- R studio, R projects, R markdown and integration with Github can help you make your data more open, reusable and reproducible